Journal of Machine Learning Research 18 (2018) 1-54 Submitted 09/15; Revised 07/18; Published 11/18

Theoretical Analysis of Cross-Validation for Estimating the Risk of the k-Nearest Neighbor Classifier

Alain Celisse CELISSE@MATH.UNIV-LILLE1.FR
Laboratoire de Mathématiques
UMR 8524 CNRS-Université de Lille
Inria – MODAL Project-team, Lille
F-59 655 Villeneuve d'Ascq Cedex, France

Tristan Mary-Huard MARYHUAR@AGROPARISTECH.FR
INRA, UMR 0320 / UMR 8120 Génétique Quantitative et Évolution
Le Moulon, F-91190 Gif-sur-Yvette, France
UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay
F-75005, Paris, France

Editor: Hui Zou

Abstract

The present work aims at deriving theoretical guaranties on the behavior of some cross-validation procedures applied to the k-nearest neighbors (kNN) rule in the context of binary classification. Here we focus on the leave-p-out cross-validation (LpO) used to assess the performance of the kNN classifier. Remarkably this LpO estimator can be efficiently computed in this context using closed-form formulas derived by Celisse and Mary-Huard (2011).

We describe a general strategy to derive moment and exponential concentration inequalities for the LpO estimator applied to the kNN classifier. Such results are obtained first by exploiting the connection between the LpO estimator and U-statistics, and second by making an intensive use of the generalized Efron-Stein inequality applied to the L1O estimator. One other important contribution is made by deriving new quantifications of the discrepancy between the LpO estimator and the classification error/risk of the kNN classifier. The optimality of these bounds is discussed by means of several lower bounds as well as simulation experiments.

Keywords: Classification, Cross-validation, Risk estimation

1. Introduction

The k-nearest neighbor (kNN) algorithm (Fix and Hodges, 1951) in binary classification is a popular prediction algorithm based on the idea that the predicted value at a new point is based on a majority vote from the k nearest labeled neighbors of this point. Although quite simple, the kNN classifier has been successfully applied to many difficult classification tasks (Li et al., 2004; Simard et al., 1998; Scheirer and Slaney, 2003). Efficient implementations have been also developed to allow dealing with large datasets (Indyk and Motwani, 1998; Andoni and Indyk, 2006).

The theoretical performances of the kNN classifier have been already extensively investigated. In the context of binary classification preliminary theoretical results date back to

Cover and Hart (1967); Cover (1968); Györfi (1981). The kNN classifier has been proved to be (weakly) universally consistent by Stone (1977) as long as $k = k_n \to +\infty$ and $k/n \to 0$ as $n \to +\infty$. For the 1NN classifier, an asymptotic expansion of the error rate has been derived by Psaltis et al. (1994). The same strategy has been successfully applied to the kNN classifier by Snapp and Venkatesh (1998). Hall et al. (2008) study the influence of the parameter k on the risk of the kNN classifier by means of an asymptotic expansion derived from a Poisson or binomial model for the training points. More recently, Cannings et al. (2017) pointed out some limitations suffered by the "classical" kNN classifier and deduced an improved version based on a local choice of k in the semi-supervised context. In contrast to the aforementioned results, the work by Chaudhuri and Dasgupta (2014) focuses on the finite-sample framework. They typically provide upper bounds with high probability on the risk of the kNN classifier where the bounds are not distribution-free. Alternatively in the regression setting, Kulkarni and Posner (1995) derived a strategy leading to a finite-sample bound on the performance of 1NN, which has been extended to the (weighted) kNN rule ($k \geq 1$) by Biau et al. (2010a,b) (see also Berrett et al., 2016, where a weighted kNN estimator is designed for estimating the entropy). We refer interested readers to Biau and Devroye (2016) for an almost thorough presentation of known results on the kNN algorithm in various contexts.

In numerous (if not all) practical applications, computing the cross-validation (CV) estimator (Stone, 1974, 1982) has been among the most popular strategies to evaluate the performance of the kNN classifier (Devroye et al., 1996, Section 24.3). All CV procedures share a common principle which consists in splitting a sample of n points into two disjoint subsets called *training* and *test* sets with respective cardinalities $n - p$ and p, for any $1 \leq p \leq n-1$. The $n-p$ training set data serve to compute a classifier, while its performance is evaluated from the p *left out* data of the test set. For a complete and comprehensive review on cross-validation procedures, we refer the interested reader to Arlot and Celisse (2010).

In the present work, we focus on the leave-p-out (LpO) cross-validation. Among CV procedures, it belongs to exhaustive strategies since it considers (and averages over) all the $\binom{n}{p}$ possible such splittings of $\{1, \ldots, n\}$ into training and test sets. Usually the induced computation time of the LpO is prohibitive, which gives rise to its surrogate called $V-$fold cross-validation (V-FCV) with $V \approx n/p$ (Geisser, 1975). However, Steele (2009); Celisse and Mary-Huard (2011) recently derived closed-form formulas respectively for the bootstrap and the LpO procedures applied to the kNN classifier. Such formulas allow for an efficient computation of the LpO estimator. Moreover since the V-FCV estimator suffers the same bias but a larger variance than the LpO one (Celisse and Robin, 2008; Arlot and Celisse, 2010), LpO (with $p = \lfloor n/V \rfloor$) strictly improves upon V-FCV in the present context.

Although being favored in practice for assessing the risk of the kNN classifier, the use of CV comes with very few theoretical guarantees regarding its performance. Moreover probably for technical reasons, most existing results apply to Hold-out and leave-one-out (L1O), that is LpO with $p = 1$ (Kearns and Ron, 1999). In this paper we rather consider the general LpO procedure (for $1 \leq p \leq n - 1$) used to estimate the risk (alternatively the classification error rate) of the kNN classifier. Our main purpose is then to provide distribution-free theoretical guarantees on the behavior of LpO with respect to influential parameters such as p, n, and k. For instance we aim at answering questions such as: "Does there exist any regime of $p = p(n)$ (with $p(n)$ some function of n) where the LpO estimator is a consistent estimate of the risk of the kNN classifier?", or "Is it possible to describe the convergence rate of the LpO estimator depending on p?"

Contributions. The main contribution of the present work is two-fold: (*i*) we describe a new general strategy to derive moment and exponential concentration inequalities for the LpO estimator applied to the kNN binary classifier, and (*ii*) these inequalities serve to derive the convergence rate of the LpO estimator towards the risk of the kNN classifier.

This new strategy relies on several steps. First exploiting the connection between the LpO estimator and U-statistics (Koroljuk and Borovskich, 1994) and the Rosenthal inequality (Ibragimov and Sharakhmetov, 2002), we prove that upper bounding the polynomial moments of the centered LpO estimator reduces to deriving such bounds for the simpler L1O estimator. Second, we derive new upper bounds on the moments of the L1O estimator using the generalized Efron-Stein inequality (Boucheron et al., 2005, 2013, Theorem 15.5). Third, combining the two previous steps provides some insight on the interplay between p/n and k in the concentration rates measured in terms of moments. This finally results in new exponential concentration inequalities for the LpO estimator applying whatever the value of the ratio $p/n \in (0, 1)$. In particular while the upper bounds increase with $1 \leq p \leq n/2 + 1$, it is no longer the case if $p > n/2 + 1$. We also provide several lower bounds suggesting our upper bounds cannot be improved in some sense in a distribution-free setting.

The remainder of the paper is organized as follows. The connection between the LpO estimator and U-statistics is clarified in Section 2, where we also recall the closed-form formula of the LpO estimator applied to the kNN classifier (Celisse and Mary-Huard, 2011). Order-q moments ($q \geq 2$) of the LpO estimator are then upper bounded in terms of those of the L1O estimator. This step can be applied to any classification algorithm. Section 3 then specifies the previous upper bounds in the case of the kNN classifier, which leads to the main Theorem 3.2 characterizing the concentration behavior of the LpO estimator with respect to p, n, and k in terms of polynomial moments. Deriving exponential concentration inequalities for the LpO estimator is the main concern of Section 4 where we highlight the strength of our strategy by comparing our main inequalities with concentration inequalities derived with less sophisticated tools. Finally Section 5 exploits the previous results to bound the gap between the LpO estimator and the classification error of the kNN classifier. The optimality of these upper bounds is first proved in our distribution-free framework by establishing several new lower bounds matching the upper ones in some specific settings. Second, empirical experiments are also reported which support the above conclusions.

2. U-statistics and LpO estimator

2.1. Statistical framework

Classification We tackle the binary classification problem where the goal is to predict the unknown label $Y \in \{0, 1\}$ of an observation $X \in \mathcal{X} \subset \mathbb{R}^d$. The random variable (X, Y) has an *unknown* joint distribution $P_{(X,Y)}$ defined by $P_{(X,Y)}(B) = \mathbb{P}\left[\,(X, Y) \in B\,\right]$ for any Borelian set $B \in \mathcal{X} \times \{0, 1\}$, where $\mathbb{P}$ denotes a reference probability distribution. In what follows no particular distributional assumption is made regarding X. To predict the label, one aims at building a classifier $\hat{f} : \mathcal{X} \to \{0, 1\}$ on the basis of a set of random variables

$\mathcal{D}_n = \{Z_1, \ldots, Z_n\}$ called the training sample, where $Z_i = (X_i, Y_i),\ 1 \leq i \leq n$ represent n copies of (X, Y) drawn independently from $P_{(X,Y)}$. In settings where no confusion is possible, we will replace $\mathcal{D}_n$ by $\mathcal{D}$.

Any strategy to build such a classifier is called a *classification algorithm*, and can be formally defined as a function $\mathcal{A} :\ \cup_{n\geq 1} \{\mathcal{X} \times \{0,1\}\}^n \to \mathcal{F}$ that maps a training sample $\mathcal{D}_n$ onto the corresponding classifier $\mathcal{A}^{\mathcal{D}_n}(\cdot) = \hat{f} \in \mathcal{F}$, where $\mathcal{F}$ is the set of all measurable functions from $\mathcal{X}$ to $\{0, 1\}$. Numerous classifiers have been considered in the literature and it is out of the scope of the present paper to review all of them (see Devroye et al. (1996) for many instances). Here we focus on the k-nearest neighbor rule (kNN) initially proposed by Fix and Hodges (1951) and further studied for instance by Devroye and Wagner (1977); Rogers and Wagner (1978).

The kNN algorithm For $1 \leq k \leq n$, the kNN classification algorithm, denoted by $\mathcal{A}_k$, consists in classifying any new observation x using a *majority vote* decision rule based on the labels of the k closest points to x, denoted by $X_{(1)}(x), \ldots, X_{(k)}(x)$, among the training sample $X_1, \ldots, X_n$. In what follows these k *nearest neighbors* are chosen according to the distance associated with the usual Euclidean norm in $\mathbb{R}^d$. Note that other *adaptive metrics* have been also considered in the literature (see for instance Hastie et al., 2001, Chap. 14). But such examples are out of the scope of the present work, that is our reference distance does not depend on the training sample at hand. Let us also emphasize that possible ties are broken by using the *smallest index* among ties, which is one possible choice for the Stone lemma to hold true (Biau and Devroye, 2016, Lemma 10.6, p.125).

Formally, given $V_k(x) = \left\{1 \leq i \leq n,\ X_i \in \left\{X_{(1)}(x), \ldots, X_{(k)}(x)\right\}\right\}$ the set of indices of the k nearest neighbors of x among $X_1, \ldots, X_n$, the kNN classifier is defined by

$$\mathcal{A}_k(\mathcal{D}_n; x) = \widehat{f}_k(\mathcal{D}_n; x) := \begin{cases} 1 & \text{, if } \frac{1}{k}\sum_{i\in V_k(x)} Y_i = \frac{1}{k}\sum_{i=1}^{k} Y_{(i)}(x) > 0.5 \\ 0 & \text{, if } \frac{1}{k}\sum_{i=1}^{k} Y_{(i)}(x) < 0.5 \\ \mathcal{B}(0.5) & \text{, otherwise} \end{cases} , \quad (2.1)$$

where $Y_{(i)}(x)$ is the label of the i-th nearest neighbor of x for $1 \leq i \leq k$, and $\mathcal{B}(0.5)$ denotes a Bernoulli random variable with parameter 1/2.

Leave-p-out cross-validation For a given sample $\mathcal{D}_n$, the performance of any classifier $\hat{f} = \mathcal{A}^{\mathcal{D}_n}(\cdot)$ (respectively of any classification algorithm $\mathcal{A}$) is assessed by the classification error $L(\hat{f})$ (respectively the risk $R(\hat{f})$) defined by

$$L(\hat{f}) = \mathbb{P}\left(\hat{f}(X) \neq Y \mid \mathcal{D}_n\right) \ , \quad \text{and} \qquad R(\hat{f}) = \mathbb{E}\left[\mathbb{P}\left(\hat{f}(X) \neq Y \mid \mathcal{D}_n\right)\right] .$$

In this paper we focus on the estimation of $L(\hat{f})$ (and its expectation $R(\hat{f})$) by use of the *Leave-p-Out* (LpO) cross-validation for $1 \leq p \leq n-1$ (Zhang, 1993; Celisse and Robin, 2008). LpO successively considers all possible splits of $\mathcal{D}_n$ into a training set of cardinality $n-p$ and a test set of cardinality p. Denoting by $\mathcal{E}_{n-p}$ the set of all possible subsets of $\{1, \ldots, n\}$ with cardinality $n-p$, any $e \in \mathcal{E}_{n-p}$ defines a split of $\mathcal{D}_n$ into a training sample $\mathcal{D}^e = \{Z_i \mid i \in e\}$ and a test sample $\mathcal{D}^{\bar{e}}$, where $\bar{e} = \{1, \ldots, n\} \setminus e$. For a given classification algorithm $\mathcal{A}$, the final LpO estimator of the performance of $\mathcal{A}^{\mathcal{D}_n}(\cdot) = \widehat{f}$ is the average (over all possible splits) of the classification error estimated on each test set, that is

$$\widehat{R}_p(\mathcal{A}, \mathcal{D}_n) = \binom{n}{p}^{-1} \sum_{e\in\mathcal{E}_{n-p}} \left(\frac{1}{p} \sum_{i\in\bar{e}} \mathbb{1}_{\left\{\mathcal{A}^{\mathcal{D}^e}(X_i)\neq Y_i\right\}} \right), \qquad (2.2)$$

where $\mathcal{A}^{\mathcal{D}^e}(\cdot)$ is the classifier built from $\mathcal{D}^e$. We refer the reader to Arlot and Celisse (2010) for a detailed description of LpO and other cross-validation procedures. In the sequel, the lengthy notation $\widehat{R}_p(\mathcal{A}, \mathcal{D}_n)$ is replaced by $\widehat{R}_{p,n}$ in settings where no confusion can arise about the algorithm $\mathcal{A}$ or the training sample $\mathcal{D}_n$, and by $\widehat{R}_p(\mathcal{D}_n)$ if the training sample has to be kept in mind.

Exact LpO for the kNN classification algorithm Usually due to its seemingly prohibitive computational cost, LpO is not applied except with $p = 1$ where it reduces to the well known leave-one-out. However in several contexts such as density estimation (Celisse and Robin, 2008; Celisse, 2014) or regression (Celisse, 2008), closed-form formulas have been derived for the LpO estimator when applied with projection and kernel estimators. The kNN classifier is another instance of such estimators for which efficiently computing the LpO estimator is possible. Its computation requires a time complexity that is linear in p as previously established by Celisse and Mary-Huard (2011). Let us briefly recall the main steps leading to the closed-form formula.

1. From Eq. (2.2) the LpO estimator can be expressed as a sum (over the n observations of the complete sample) of probabilities:

$$\binom{n}{p}^{-1} \sum_{e \in \mathcal{E}_{n-p}} \frac{1}{p} \left(\sum_{i \notin e} \mathbb{1}_{\left\{\mathcal{A}^{\mathcal{D}^e}(X_i) \neq Y_i\right\}} \right) = \frac{1}{p} \sum_{i=1}^{n} \left[\binom{n}{p}^{-1} \sum_{e \in \mathcal{E}_{n-p}} \mathbb{1}_{\left\{\mathcal{A}^{\mathcal{D}^e}(X_i) \neq Y_i\right\}} \mathbb{1}_{\{i \notin e\}} \right]$$
$$= \frac{1}{p} \sum_{i=1}^{n} \mathbb{P}_e(\mathcal{A}^{\mathcal{D}^e}(X_i) \neq Y_i \mid i \notin e) \mathbb{P}_e(i \notin e).$$

   Here $\mathbb{P}_e$ means that the integration is made with respect to the random variable $e \in \mathcal{E}_{n-p}$, which follows the uniform distribution over the $\binom{n}{p}$ possible subsets in $\mathcal{E}_{n-p}$ with cardinality $n-p$. For instance $\mathbb{P}_e(i \notin e) = p/n$ since it is the proportion of subsamples with cardinality $n-p$ which do not contain a given prescribed index i, which equals $\binom{n-1}{n-p}/\binom{n}{p}$. (See also Lemma D.4 for further examples of such calculations.)

2. For any X_i, let $X_{(1)}, ..., X_{(k+p-1)}, X_{(k+p)}, ..., X_{(n-1)}$ be the ordered sequence of neighbors of X_i. This list depends on X_i, that is $X_{(1)}$ should be noted $X_{(i,1)}$. But this dependency is skipped here for the sake of readability.

   The key in the derivation is to condition with respect to the random variable R_k^i which denotes the rank (in the whole sample $\mathcal{D}_n$) of the k-th neighbor of X_i in the $\mathcal{D}^e$. For instance $R_k^i = j$ means that $X_{(j)}$ is the k-th neighbor of X_i in $\mathcal{D}^e$. Then

$$\mathbb{P}_e(\mathcal{A}^{\mathcal{D}^e}(X_i) \neq Y_i | i \notin e) = \sum_{j=k}^{k+p-1} \mathbb{P}_e(\mathcal{A}^{\mathcal{D}^e}(X_i) \neq Y_i \mid R_k^i = j,\ i \notin e) \mathbb{P}_e(R_k^i = j \mid i \notin e),$$

   where the sum involves p terms since only $X_{(k)}, \ldots, X_{(k+p-1)}$ are candidates for being the k-th neighbor of X_i in at least one training subset e.

3. Observe that the resulting probabilities can be easily computed (see Lemma D.4):

   ⋆ $\mathbb{P}_e(i \notin e) = \frac{p}{n}$
   ⋆ $\mathbb{P}_e(R_k^i = j | i \notin e) = \frac{k}{j} P\left(U = j - k\right)$
   ⋆ $\mathbb{P}_e(\mathcal{A}^{\mathcal{D}^e}(X_i) \neq Y_i | V_k^i = j,\ i \notin e) = (1 - Y_j)\left[1 - F_H\left(\frac{k+1}{2}\right)\right] + Y_j\left[1 - F_{H'}\left(\frac{k-1}{2}\right)\right],$

with $U \sim \mathcal{H}(j, n-j-1, p-1)$, $H \sim \mathcal{H}(N_i^j, j-N_i^j-1, k-1)$, and $H' \sim \mathcal{H}(N_i^j - 1, j-N_i^j, k-1)$, where F_H and $F_{H'}$ respectively denote the cumulative distribution functions of H and H', $\mathcal{H}$ denotes the hypergeometric distribution, and N_i^j is the number of 1's among the j nearest neighbors of X_i in $\mathcal{D}_n$.

The computational cost of LpO for the kNN classifier is the same as that of L1O for the $(k+p-1)$NN classifier whatever p, that is $O(pn)$. This contrasts with the usual $\binom{n}{p}$ prohibitive computational complexity seemingly suffered by LpO.

2.2. U-statistics: General bounds on LpO moments

The purpose of the present section is to describe a general strategy allowing to derive new upper bounds on the polynomial moments of the LpO estimator. As a first step of this strategy, we establish the connection between the LpO risk estimator and U-statistics. Second, we exploit this connection to derive new upper bounds on the order-q moments of the LpO estimator for $q \geq 2$. Note that these upper bounds, which relate moments of the LpO estimator to those of the L1O estimator, hold true with any classifier.

Let us start by introducing U-statistics and recalling some of their basic properties that will serve our purposes. For a thorough presentation, we refer to the books by Serfling (1980); Koroljuk and Borovskich (1994). The first step is the definition of a U-statistic of order $m \in \mathbb{N}^*$ as an average over all m-tuples of distinct indices in $\{1, \ldots, n\}$.

Definition 2.1 (Koroljuk and Borovskich (1994)). *Let $h: \mathcal{X}^m \longrightarrow \mathbb{R}$ denote any measurable function where $m \geq 1$ is an integer. Let us further assume h is a symmetric function of its arguments. Then any function $U_n: \mathcal{X}^n \longrightarrow \mathbb{R}$ such that*

$$U_n(x_1, \ldots, x_n) = U_n(h)(x_1, \ldots, x_n) = \binom{n}{m}^{-1} \sum_{1 \leq i_1 < \ldots < i_m \leq n} h\left(x_{i_1}, \ldots, x_{i_m}\right)$$

where $m \leq n$, is a U-statistic of order m and kernel h.

Before clarifying the connection between LpO and U-statistics, let us introduce the main property of U-statistics our strategy relies on. It consists in representing any U-statistic as an average, over all permutations, of sums of independent variables.

Proposition 2.1 (Eq. (5.5) in Hoeffding (1963)). *With the notation of Definition 2.1, let us define $W: \mathcal{X}^n \longrightarrow \mathbb{R}$ by*

$$W(x_1, \ldots, x_n) = \frac{1}{r} \sum_{j=1}^{r} h\left(x_{(j-1)m+1}, \ldots, x_{jm}\right), \tag{2.3}$$

where $r = \lfloor n/m \rfloor$ denotes the integer part of n/m. Then

$$U_n(x_1, \ldots, x_n) = \frac{1}{n!} \sum_{\sigma} W\left(x_{\sigma(1)}, \ldots, x_{\sigma(n)}\right),$$

where $\sum_\sigma$ denotes the summation over all permutations σ of $\{1, \ldots, n\}$.

We are now in position to state the key remark of the paper. All the developments further exposed in the following result from this connection between the LpO estimator defined by Eq. (2.2) and U-statistics.

Theorem 2.1. *For any classification algorithm $\mathcal{A}$ and any $1 \leq p \leq n-1$ such that a classifier can be computed from $\mathcal{A}$ on $n-p$ training points, the LpO estimator $\widehat{R}_{p,n}$ is a U-statistic of order $m = n-p+1$ with kernel $h_m : \mathcal{X}^m \longrightarrow \mathbb{R}$ defined by*

$$h_m(Z_1, \ldots, Z_m) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{\left\{\mathcal{A}^{\mathcal{D}_m^{(i)}}(X_i) \neq Y_i\right\}},$$

where $\mathcal{D}_m^{(i)}$ denotes the sample $\mathcal{D}_m = (Z_1, \ldots, Z_m)$ with Z_i withdrawn.

Note for instance that when $\mathcal{A} = \mathcal{A}_k$ denotes the kNN algorithm, the cardinality of $\mathcal{D}_m^{(i)}$ has to satisfy $n - p \geq k$, which implies that $1 \leq p \leq n-k \leq n-1$.

Proof of Theorem 2.1.

From Eq. (2.2), the LpO estimator of the performance of any classification algorithm $\mathcal{A}$ computed from $\mathcal{D}_n$ satisfies

$$\begin{aligned}
\widehat{R}_p(\mathcal{A}, \mathcal{D}_n) = \widehat{R}_{p,n} &= \frac{1}{\binom{n}{p}} \sum_{e \in \mathcal{E}_{n-p}} \frac{1}{p} \sum_{i \in \bar{e}} \mathbb{1}_{\left\{\mathcal{A}^{\mathcal{D}^e}(X_i) \neq Y_i\right\}} \\
&= \frac{1}{\binom{n}{p}} \sum_{e \in \mathcal{E}_{n-p}} \frac{1}{p} \sum_{i \in \bar{e}} \left(\sum_{v \in \mathcal{E}_{n-p+1}} \mathbb{1}_{\{v = e \cup \{i\}\}} \right) \mathbb{1}_{\left\{\mathcal{A}^{\mathcal{D}^e}(X_i) \neq Y_i\right\}},
\end{aligned}$$

since there is a unique set of indices v with cardinality $n-p+1$ such that $v = e \cup \{i\}$. Then

$$\widehat{R}_{p,n} = \frac{1}{\binom{n}{p}} \sum_{v \in \mathcal{E}_{n-p+1}} \frac{1}{p} \sum_{i=1}^{n} \left(\sum_{e \in \mathcal{E}_{n-p}} \mathbb{1}_{\{v = e \cup \{i\}\}} \mathbb{1}_{\{i \in \bar{e}\}} \right) \mathbb{1}_{\left\{\mathcal{A}^{\mathcal{D}^{v \setminus \{i\}}}(X_i) \neq Y_i\right\}}.$$

Furthermore for v and i fixed, $\sum_{e \in \mathcal{E}_{n-p}} \mathbb{1}_{\{v = e \cup \{i\}\}} \mathbb{1}_{\{i \in \bar{e}\}} = \mathbb{1}_{\{i \in v\}}$ since there is a unique set of indices e such that $e = v \setminus i$. One gets

$$\begin{aligned}
\widehat{R}_{p,n} &= \frac{1}{p} \frac{1}{\binom{n}{p}} \sum_{v \in \mathcal{E}_{n-p+1}} \sum_{i=1}^{n} \mathbb{1}_{\{i \in v\}} \mathbb{1}_{\left\{\mathcal{A}^{\mathcal{D}^{v \setminus \{i\}}}(X_i) \neq Y_i\right\}} \\
&= \frac{1}{\binom{n}{n-p+1}} \sum_{v \in \mathcal{E}_{n-p+1}} \frac{1}{n-p+1} \sum_{i \in v} \mathbb{1}_{\left\{\mathcal{A}^{\mathcal{D}^{v \setminus \{i\}}}(X_i) \neq Y_i\right\}},
\end{aligned}$$

by noticing $p\binom{n}{p} = \frac{pn!}{p!n-p!} = \frac{n!}{p-1!n-p!} = (n-p+1)\binom{n}{n-p+1}$.

□

The kernel h_m is a deterministic and symmetric function of its arguments that does only depend on m. Let us also notice that $h_m(Z_1,\ldots,Z_m)$ reduces to the L1O estimator of the risk of the classifier $\mathcal{A}$ computed from $Z_1,\ldots,Z_m$, that is

$$h_m(Z_1,\ldots,Z_m) = \widehat{R}_1(\mathcal{A},\mathcal{D}_m) = \widehat{R}_{1,n-p+1}. \tag{2.4}$$

In the context of testing whether two binary classifiers have different error rates, this fact has already been pointed out by Fuchs et al. (2013).

We now derive a general upper bound on the q-th moment ($q \geq 1$) of the LpO estimator that holds true for any classifier (as long as the following expectations are well defined).

Theorem 2.2. *For any classifier $\mathcal{A}$, let $\mathcal{A}^{\mathcal{D}_n}(\cdot)$ and $\mathcal{A}^{\mathcal{D}_m}(\cdot)$ be the corresponding classifiers built from respectively $\mathcal{D}_n$ and $\mathcal{D}_m$, where $m = n-p+1$. Then for every $1 \leq p \leq n-1$ such that a classifier can be computed from $\mathcal{A}$ on $n-p$ training points, and for any $q \geq 1$,*

$$\mathbb{E}\left[\left|\widehat{R}_{p,n} - \mathbb{E}\left[\widehat{R}_{p,n}\right]\right|^q\right] \leq \mathbb{E}\left[\left|\widehat{R}_{1,m} - \mathbb{E}\left[\widehat{R}_{1,m}\right]\right|^q\right]. \tag{2.5}$$

Furthermore as long as $p > n/2+1$, one also gets

- *for $q = 2$*

$$\mathbb{E}\left[\left|\widehat{R}_{p,n} - \mathbb{E}\left[\widehat{R}_{p,n}\right]\right|^2\right] \leq \frac{\mathbb{E}\left[\left|\widehat{R}_{1,m} - \mathbb{E}\left[\widehat{R}_{1,m}\right]\right|^2\right]}{\left\lfloor \frac{n}{m}\right\rfloor} \cdot \tag{2.6}$$

- *for every $q > 2$*

$$\mathbb{E}\left[\left|\widehat{R}_{p,n} - \mathbb{E}\left[\widehat{R}_{p,n}\right]\right|^q\right] \leq B(q,\gamma)\times$$
$$\max\left\{2^q\gamma\left\lfloor\frac{n}{m}\right\rfloor\mathbb{E}\left[\left|\frac{\widehat{R}_{1,m} - \mathbb{E}\left[\widehat{R}_{1,m}\right]}{\left\lfloor\frac{n}{m}\right\rfloor}\right|^q\right], \left(\sqrt{\frac{2\mathrm{Var}\left(\widehat{R}_{1,m}\right)}{\left\lfloor\frac{n}{m}\right\rfloor}}\right)^q\right\}, \tag{2.7}$$

where $\gamma > 0$ is a numeric constant and $B(q,\gamma)$ denotes the optimal constant defined in the Rosenthal inequality (Proposition D.2).

The proof is given in Appendix A.1. Eq. (2.5) and (2.6) straightforwardly result from the Jensen inequality applied to the average over all permutations provided in Proposition 2.1. If $p > n/2+1$, the integer part $\lfloor n/m\rfloor$ becomes larger than 1 and Eq. (2.6) becomes better than Eq. (2.5) for $q = 2$. As a consequence of our strategy of proof, the right-hand side of Eq. (2.6) is equal to the classical upper bound on the variance of U-statistics which suggests it cannot be improved without adding further assumptions.

Unlike the above ones, Eq. (2.7) is derived from the Rosenthal inequality, which enables us to upper bound a sum $\left\|\sum_{i=1}^r \xi_i\right\|_q$ of independent and identically centered random variables in terms of $\sum_{i=1}^r \left\|\xi_i\right\|_q$ and $\sum_{i=1}^r \mathrm{Var}(\xi_i)$. Let us remark that, for $q = 2$, both terms of the right-hand side of Eq. (2.7) are of the same order as Eq. (2.6) up to constants. Furthermore using the Rosenthal inequality allows taking advantage of the integer part $\lfloor n/m\rfloor$ when $p > n/2+1$ (unlike what we get by using Eq.(2.5) for $q > 2$). In particular it provides a new understanding of the behavior of the LpO estimator when $p/n \to 1$ as highlighted later by Proposition 4.2.

3. New bounds on LpO moments for the kNN classifier

Our goal is now to specify the general upper bounds provided by Theorem 2.2 in the case of the kNN algorithm $\mathcal{A}_k$ ($1 \leq k \leq n$) introduced by (2.1).

Since Theorem 2.2 expresses the moments of the LpO estimator in terms of those of the L1O estimator computed from $\mathcal{D}_m$ (with $m = n - p + 1$), the next step consists in focusing on the L1O moments. Deriving upper bounds on the moments of the L1O is achieved using a generalization of the well-known Efron-Stein inequality (see Theorem D.1 for Efron-Stein's inequality and Theorem 15.5 in Boucheron et al. (2013) for its generalization). For the sake of completeness, we first recall a corollary of this generalization that is proved in Section D.1.4 (see Corollary D.1).

Proposition 3.1. *Let $\xi_1, \ldots, \xi_n$ denote n independent Ξ-valued random variables and $\zeta = f(\xi_1, \ldots, \xi_n)$, where $f : \Xi^n \to \mathbb{R}$ is any measurable function. With $\xi_1', \ldots, \xi_n'$ independent copies of the ξ_is, there exists a universal constant $\kappa \leq 1.271$ such that for any $q \geq 2$,*

$$\|\zeta - \mathbb{E}\zeta\|_q \leq \sqrt{2\kappa q} \sqrt{\left\| \sum_{i=1}^{n} \left(f(\xi_1, \ldots, \xi_i, \ldots, \xi_n) - f(\xi_1, \ldots, \xi_i', \ldots, \xi_n) \right)^2 \right\|_{q/2}} .$$

Then applying Proposition 3.1 with $\zeta = \widehat{R}_1(\mathcal{A}_k, \mathcal{D}_m) = \widehat{R}_{1,m}$ (L1O estimator computed from $\mathcal{D}_m$ with $m = n - p + 1$) and $\Xi = \mathbb{R}^d \times \{0, 1\}$ leads to the following Theorem 3.1. It controls the order-q moments of the L1O estimator applied to the kNN classifier.

Theorem 3.1. *For every $1 \leq k \leq n-1$, let $A_k^{\mathcal{D}_m}$ ($m = n-p+1$) denote the kNN classifier learnt from $\mathcal{D}_m$ and $\widehat{R}_{1,m}$ be the corresponding L1O estimator given by Eq. (2.2). Then*

- *for $q = 2$,*

$$\mathbb{E}\left[\left(\widehat{R}_{1,m} - \mathbb{E}\left[\widehat{R}_{1,m} \right] \right)^2 \right] \leq C_1 \frac{k^{3/2}}{m} \quad ; \tag{3.1}$$

- *for every $q > 2$,*

$$\mathbb{E}\left[\left| \widehat{R}_{1,m} - \mathbb{E}\left[\widehat{R}_{1,m} \right] \right|^q \right] \quad \leq (C_2 \cdot k)^q \left(\tfrac{q}{m} \right)^{q/2} , \tag{3.2}$$

with $C_1 = 2 + 16\gamma_d$ and $C_2 = 4\gamma_d \sqrt{2\kappa}$, where γ_d is a constant (arising from Stone's lemma, see Lemma D.5) that grows exponentially with dimension d, and κ is defined in Proposition 3.1.

Its proof (detailed in Section A.2) relies on Stone's lemma (Lemma D.5). For a given X_i, it proves that the number of points in $\mathcal{D}_n^{(i)}$ having X_i among their k nearest neighbors is not larger than $k\gamma_d$. The dependence of our upper bounds with respect to γ_d (see explicit constants C_1 and C_2) induces their strong deterioration as the dimension d grows since $\gamma_d \approx 4.8^d - 1$. Therefore the larger the dimension d, the larger the required sample size n for the upper bound to be small (at least smaller than 1). Note also that the tie breaking strategy (based on the smallest index in the present work) is chosen so that it ensures Stone's lemma to hold true.

In Eq. (3.1), the easier case $q = 2$ enables to exploit exact calculations of (rather than upper bounds on) the variance of the L1O estimator. Since $\mathbb{E}\left[\widehat{R}_{1,m}\right] = R\left(A_k^{\mathcal{D}_{n-p}}\right)$ (risk of the kNN classifier computed from $\mathcal{D}_{n-p}$), the resulting $k^{3/2}/m$ rate is a strict improvement upon the usual k^2/m that is derived from using the sub-Gaussian exponential concentration inequality proved by Theorem 24.4 in Devroye et al. (1996).

By contrast the larger k^q arising in Eq. (3.2) results from the difficulty to derive a tight upper bound for the expectation of $(\sum_{i=1}^n \mathbb{1}_{\left\{A_k^{\mathcal{D}_m^{(i)}}(X_i)\neq A_k^{\mathcal{D}_m^{(i,j)}}(X_i)\right\}})^q$ with $q > 2$, where $\mathcal{D}_m^{(i)}$ (resp. $\mathcal{D}_m^{(i,j)}$) denotes the sample $\mathcal{D}_m$ where Z_i has been (resp. Z_i and Z_j have been) removed.

We are now in position to state the main result of this section. It follows from the combination of Theorem 2.2 (connecting moments of the LpO and L1O estimators) and Theorem 3.1 (providing an upper bound on the order-q moments of the L1O).

Theorem 3.2. *For every $p, k \geq 1$ such that $p+k \leq n$, let $\widehat{R}_{p,n}$ denote the LpO risk estimator (see* (2.2)*) of the kNN classifier $\mathcal{A}_k^{\mathcal{D}_n}(\cdot)$ defined by* (2.1)*. Then there exist (known) constants $C_1, C_2 > 0$ such that for every $1 \leq p \leq n-k$,*

- *for $q = 2$,*

$$\mathbb{E}\left[\left(\widehat{R}_{p,n} - \mathbb{E}\left[\widehat{R}_{p,n}\right]\right)^2\right] \leq C_1 \frac{k^{3/2}}{(n-p+1)} \quad ; \tag{3.3}$$

- *for every $q > 2$,*

$$\mathbb{E}\left[\left|\widehat{R}_{p,n} - \mathbb{E}\left[\widehat{R}_{p,n}\right]\right|^q\right] \leq (C_2 k)^q \left(\tfrac{q}{n-p+1}\right)^{q/2} , \tag{3.4}$$

with $C_1 = \frac{128\kappa\gamma_d}{\sqrt{2\pi}}$ and $C_2 = 4\gamma_d\sqrt{2\kappa}$, where γ_d denotes the constant arising from Stone's lemma (Lemma D.5). Furthermore in the particular setting where $n/2+1 < p \leq n-k$, then

- *for $q = 2$,*

$$\mathbb{E}\left[\left(\widehat{R}_{p,n} - \mathbb{E}\left[\widehat{R}_{p,n}\right]\right)^2\right] \leq C_1 \frac{k^{3/2}}{(n-p+1)\left\lfloor \frac{n}{n-p+1}\right\rfloor} \quad , \tag{3.5}$$

- *for every $q > 2$,*

$$\begin{aligned}&\mathbb{E}\left[\left|\widehat{R}_{p,n} - \mathbb{E}\left[\widehat{R}_{p,n}\right]\right|^q\right]\\ \leq\ & \left\lfloor \frac{n}{n-p+1}\right\rfloor \Gamma^q \max\left(\frac{k^{3/2}}{(n-p+1)\left\lfloor \frac{n}{n-p+1}\right\rfloor}\,q, \frac{k^2}{(n-p+1)\left\lfloor \frac{n}{n-p+1}\right\rfloor^2}\,q^3\right)^{q/2}\end{aligned} \tag{3,6}$$

where $\Gamma = 2\sqrt{2e}\max\left(\sqrt{2C_1}, 2C_2\right)$.

The straightforward proof is detailed in Section A.3. Let us start by noticing that both upper bounds in Eq. (3.3) and (3.4) deteriorate as p grows. This is no longer the case for Eq. (3.5) and (3.6), which are specifically designed to cover the setup where $p > n/2 + 1$, that is where $\lfloor n/m \rfloor$ is no longer equal to 1. Therefore unlike Eq. (3.3) and (3.4), these last two inequalities are particularly relevant in the setup where $p/n \to 1$, as $n \to +\infty$, which has been investigated by Shao (1993); Yang (2006, 2007); Celisse (2014). Eq. (3.5) and (3.6) lead to respective convergence rates at worse $k^{3/2}/n$ (for $q = 2$) and k^q/n^{q-1} (for $q > 2$). In particular this last rate becomes approximately equal to $(k/n)^q$ as q gets large.

One can also emphasize that, as a U-statistic of fixed order $m = n - p + 1$, the LpO estimator has a known Gaussian limiting distribution, that is (see Theorem A, Section 5.5.1 Serfling, 1980)

$$\frac{\sqrt{n}}{m}\left(\widehat{R}_{p,n} - \mathbb{E}\left[\widehat{R}_{p,n}\right]\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \sigma_1^2\right),$$

where $\sigma_1^2 = \mathrm{Var}\left[\, g(Z_1)\,\right]$, with $g(z) = E\left[\, h_m(z, Z_2, \ldots, Z_m)\,\right]$. Therefore the upper bound given by Eq. (3.5) is non-improvable in some sense with respect to the interplay between n and p since one recovers the right magnitude for the variance term as long as $m = n-p+1$ is assumed to be constant.

Finally Eq. (3.6) has been derived using a specific version of the Rosenthal inequality (Ibragimov and Sharakhmetov, 2002) stated with the optimal constant and involving a "balancing factor". In particular this balancing factor has allowed us to optimize the relative weight of the two terms between brackets in Eq. (3.6). This leads us to claim that the dependence of the upper bound with respect to q cannot be improved with this line of proof. However we cannot conclude that the term in q^3 cannot be improved using other technical arguments.

4. Exponential concentration inequalities

This section provides exponential concentration inequalities for the LpO estimator applied to the kNN classifier. Our main results heavily rely on the moment inequalities previously derived in Section 3, namely Theorem 3.2. In order to emphasize the gain allowed by this strategy of proof, we start this section by successively proving two exponential inequalities obtained with less sophisticated tools. We then discuss the strength and weakness of each of them to justify the additional refinements we introduce step by step along the section.

A first exponential concentration inequality for $\widehat{R}_p(\mathcal{A}_k, \mathcal{D}_n) = \widehat{R}_{p,n}$ can be derived by use of the bounded difference inequality following the line of proof of Devroye et al. (1996, Theorem 24.4) originally developed for the L1O estimator.

Proposition 4.1. *For any integers $p, k \geq 1$ such that $p + k \leq n$, let $\widehat{R}_{p,n}$ denote the LpO estimator (2.2) of the classification error of the kNN classifier $\mathcal{A}_k^{\mathcal{D}_n}(\cdot)$ defined by (2.1). Then for every $t > 0$,*

$$\mathbb{P}\left(\left|\widehat{R}_{p,n} - \mathbb{E}\left(\widehat{R}_{p,n}\right)\right| > t\right) \leq 2e^{-n\frac{t^2}{8(k+p-1)^2\gamma_d^2}}. \tag{4.1}$$

where γ_d denotes the constant introduced in Stone's lemma (Lemma D.5).

The proof is given in Appendix B.1.

The upper bound of Eq. (4.1) strongly exploits the facts that: (i) for X_j to be one of the k nearest neighbors of X_i in at least one subsample X^e, it requires X_j to be one of the $k+p-1$ nearest neighbors of X_i in the complete sample, and (ii) the number of points for which X_j may be one of the $k+p-1$ nearest neighbors cannot be larger than $(k+p-1)\gamma_d$ by Stone's Lemma (see Lemma D.5).

This reasoning results in a rough upper bound since the denominator in the exponent exhibits a $(k+p-1)^2$ factor where k and p play the same role. The reason is that we do not distinguish between points for which X_j is among or above the k nearest neighbors of X_i in the whole sample (although these two setups lead to highly different probabilities of being among the k nearest neighbors in the training sample). Consequently the dependence of the convergence rate on k and p in Proposition 4.1 can be improved, as confirmed by forthcoming Theorems 4.1 and 4.2.

Based on the previous comments, a sharper quantification of the influence of each neighbor among the $k+p-1$ ones leads to the next result.

Theorem 4.1. *For every $p, k \geq 1$ such that $p+k \leq n$, let $\widehat{R}_{p,n}$ denote the LpO estimator (2.2) of the classification error of the kNN classifier $\mathcal{A}_k^{\mathcal{D}_n}(\cdot)$ defined by (2.1). Then there exists a numeric constant $\Box > 0$ such that for every $t > 0$,*

$$\max\left(\mathbb{P}\left(\widehat{R}_{p,n} - \mathbb{E}\left(\widehat{R}_{p,n}\right) > t\right), \mathbb{P}\left(\mathbb{E}\left(\widehat{R}_{p,n}\right) - \widehat{R}_{p,n} > t\right)\right) \leq \exp\left(-n\frac{t^2}{\Box k^2\left[1+(k+p)\frac{p-1}{n-1}\right]}\right),$$

with $\Box = 1024e\kappa(1+\gamma_d)$, where γ_d is introduced in Lemma D.5 and $\kappa \leq 1.271$ is a universal constant.

The proof is given in Section B.2.

Unlike Proposition 4.1, taking into account the rank of each neighbor in the whole sample enables us to considerably reduce the weight of p (compared to that of k) in the denominator of the exponent. In particular, letting $p/n \to 0$ as $n \to +\infty$ (with k assumed to be fixed for instance) makes the influence of the $k+p$ factor asymptotically negligible. This would allow for recovering (up to numeric constants) a similar upper bound to that of Devroye et al. (1996, Theorem 24.4), achieved with $p = 1$.

However the upper bound of Theorem 4.1 does not reflect the right dependencies with respect to k and p compared with what has been proved for polynomial moments in Theorem 3.2. In particular it deteriorates as p increases unlike the upper bounds derived for $p > n/2+1$ in Theorem 3.2. This drawback is overcome by the following result, which is our main contribution in the present section.

Theorem 4.2. *For every $p, k \geq 1$ such that $p+k \leq n$, let $\widehat{R}_{p,n}$ denote the LpO estimator of the classification error of the kNN classifier $\hat{f}_k = \mathcal{A}_k^{\mathcal{D}_n}(\cdot)$ defined by (2.1). Then for every $t > 0$,*

$$\max\left(\mathbb{P}\left(\widehat{R}_{p,n} - \mathbb{E}\left[\widehat{R}_{p,n}\right] > t\right), \mathbb{P}\left(\mathbb{E}\left[\widehat{R}_{p,n}\right] - \widehat{R}_{p,n} > t\right)\right) \leq \exp\left(-(n-p+1)\frac{t^2}{\Delta^2 k^2}\right), \tag{4.2}$$

where $\Delta = 4\sqrt{e}\max\left(C_2, \sqrt{C_1}\right)$ *with* $C_1, C_2 > 0$ *defined in Theorem 3.1.*

Furthermore in the particular setting where $p > n/2+1$*, it comes*

$$\max\left(\mathbb{P}\left(\widehat{R}_{p,n} - \mathbb{E}\left[\widehat{R}_{p,n}\right] > t\right), \mathbb{P}\left(\mathbb{E}\left[\widehat{R}_{p,n}\right] - \widehat{R}_{p,n} > t\right)\right) \leq e \left\lfloor \frac{n}{n-p+1} \right\rfloor \times$$
$$\exp\left[-\frac{1}{2e}\min\left\{(n-p+1)\left\lfloor \frac{n}{n-p+1} \right\rfloor \frac{t^2}{4\Gamma^2 k^{3/2}}, \left((n-p+1)\left\lfloor \frac{n}{n-p+1} \right\rfloor^2 \frac{t^2}{4\Gamma^2 k^2}\right)^{1/3}\right\}\right], \tag{4.3}$$

where Γ *arises in Eq.* (3.6) *and* γ_d *denotes the constant introduced in Stone's lemma (Lemma D.5).*

The proof has been postponed to Appendix B.3. It involves different arguments for deriving the two inequalities (4.2) and (4.3) depending on the range of values of p. Firstly for $p \leq n/2+1$, a simple argument is applied to derive Ineq. (4.2) from the two corresponding moment inequalities of Theorem 3.2 characterizing the sub-Gaussian behavior of the LpO estimator in terms of its even moments (see Lemma D.2). Secondly for $p > n/2+1$, we rather exploit: (i) the appropriate upper bounds on the moments of the LpO estimator given by Theorem 3.2, combined with (ii) Proposition D.1 which establishes exponential concentration inequalities from general moment upper bounds.

In accordance with the conclusions drawn about Theorem 3.2, the upper bound of Eq. (4.2) increases as p grows unlike that of Eq. (4.3). The best concentration rate in Eq. (4.3) is achieved as $p/n \to 1$, whereas Eq. (4.2) turns out to be useless in that setting. However Eq. (4.2) remains strictly better than Theorem 4.1 as long as $p/n \to \delta \in [0,1[$ as $n \to +\infty$. Note also that the constants Γ and γ_d are the same as in Theorem 3.1. Therefore the same comments regarding their dependence with respect to the dimension d apply here.

In order to facilitate the interpretation of the last Ineq. (4.3), we also derive the following proposition (proved in Appendix B.3) which focuses on the description of each deviation term in the particular case where $p > n/2+1$.

Proposition 4.2. *With the same notation as Theorem 4.2, for any* $p, k \geq 1$ *such that* $p+k \leq n$*,* $p > n/2+1$*, and for every* $t > 0$

$$\mathbb{P}\left[\left|\widehat{R}_{p,n} - \mathbb{E}\left[\widehat{R}_{p,n}\right]\right| > \frac{\sqrt{2e}\Gamma}{\sqrt{n-p+1}}\left(\sqrt{\frac{k^{3/2}}{\left\lfloor\frac{n}{n-p+1}\right\rfloor}t} + 2e\frac{k}{\left\lfloor\frac{n}{n-p+1}\right\rfloor}t^{3/2}\right)\right] \leq \left\lfloor \frac{n}{n-p+1} \right\rfloor e \cdot e^{-t},$$

where $\Gamma > 0$ *is the constant arising from* (3.6).

The present inequality is very similar to the well-known Bernstein inequality (Boucheron et al., 2013, Theorem 2.10) except the second deviation term of order $t^{3/2}$ instead of t (for the Bernstein inequality).

With respect to n, the first deviation term is of order $\approx k^{3/2}/\sqrt{n}$, which is the same as with the Bernstein inequality. The second deviation term is of a somewhat different order, that is $\approx k\sqrt{n-p+1}/n$, as compared with the usual $1/n$ in the Bernstein inequality.

Nevertheless we almost recover the k/n rate by choosing for instance $p \approx n(1 - \log n/n)$, which leads to $k\sqrt{\log n}/n$. Therefore varying p allows to interpolate between the $k/\sqrt{n}$ and the k/n rates.

Note also that the dependence of the first (sub-Gaussian) deviation term with respect to k is only $k^{3/2}$, which improves upon the usual k^2 resulting from Ineq. (4.2) in Theorem 4.2 for instance. However this $k^{3/2}$ remains certainly too large for being optimal even if this question remains widely open at this stage in the literature.

More generally one strength of our approach is its versatility. Indeed the two above deviation terms directly result from the two upper bounds on the moments of the L1O established in Theorem 3.1. Therefore any improvement of the latter upper bounds would immediately lead to enhance the present concentration inequality (without changing the proof).

5. Assessing the gap between LpO and classification error

5.1. Upper bounds

First, we derive new upper bounds on different measures of the discrepancy between $\widehat{R}_{p,n} = \widehat{R}_p\left(\mathcal{A}_k, \mathcal{D}_n\right)$ and the classification error $L(\hat{f}_k)$ or the risk $R(\hat{f}_k) = \mathbb{E}\left[\, L(\hat{f}_k) \,\right]$. These bounds on the L$p$O estimator are completely new for $p > 1$, some of them being extensions of former ones specifically derived for the L1O estimator applied to the kNN classifier.

Theorem 5.1. *For every $p, k \geq 1$ such that $p \leq \sqrt{k}$ and $\sqrt{k} + k \leq n$, let $\widehat{R}_{p,n}$ denote the LpO risk estimator (see (2.2)) of the kNN classifier $\hat{f}_k = \mathcal{A}_k^{\mathcal{D}_n}(\cdot)$ defined by (2.1). Then,*

$$\left|\mathbb{E}\left[\, \widehat{R}_{p,n} \,\right] - R(\hat{f}_k)\right| \leq \frac{4}{\sqrt{2\pi}} \frac{p\sqrt{k}}{n} \quad , \tag{5.1}$$

and

$$\mathbb{E}\left[\left(\widehat{R}_{p,n} - R(\hat{f}_k)\right)^2\right] \leq \frac{128\kappa\gamma_d}{\sqrt{2\pi}} \frac{k^{3/2}}{n-p+1} + \frac{16}{2\pi} \frac{p^2 k}{n^2} \quad \cdot \tag{5.2}$$

Moreover,

$$\mathbb{E}\left[\left(\widehat{R}_{p,n} - L(\hat{f}_k)\right)^2\right] \leq \frac{2\sqrt{2}}{\sqrt{\pi}} \frac{(2p+3)\sqrt{k}}{n} + \frac{1}{n} \quad \cdot \tag{5.3}$$

In contrast to the results in the previous sections, a new restriction on p arises in Theorem 5.1, that is $p \leq \sqrt{k}$. This comes from the use of Lemma D.6 (proved by Devroye and Wagner (1979b)), which gives an upper bound on the L^1 stability of the kNN classifier when p observations are removed from the training sample $\mathcal{D}_n$. Actually this upper bound only remains meaningful as long as $1 \leq p \leq \sqrt{k}$.

Proof of Theorem 5.1.
Proof of (5.1)**:** With $\hat{f}_k^e = \mathcal{A}_k^{\mathcal{D}^e}$, Lemma D.6 immediately provides

$$\begin{aligned}\left|\mathbb{E}\left[\widehat{R}_{p,n} - L(\hat{f}_k)\right]\right| &= \left|\mathbb{E}\left[L(\hat{f}_k^e)\right] - \mathbb{E}\left[L(\hat{f}_k)\right]\right| \\ &\leq \mathbb{E}\left[\left|\mathbb{1}_{\{\mathcal{A}_k^{\mathcal{D}^e}(X)\neq Y\}} - \mathbb{1}_{\{\mathcal{A}_k^{\mathcal{D}_n}(X)\neq Y\}}\right|\right] \\ &= \mathbb{P}\left(\mathcal{A}_k^{\mathcal{D}^e}(X) \neq \mathcal{A}_k^{\mathcal{D}_n}(X)\right) \leq \frac{4}{\sqrt{2\pi}}\frac{p\sqrt{k}}{n} \ .\end{aligned}$$

Proof of (5.2)**:** The proof combines the previous upper bound with the one established for the variance of the LpO estimator, that is Eq. (3.3).

$$\begin{aligned}\mathbb{E}\left[\left(\widehat{R}_{p,n} - \mathbb{E}\left[L(\hat{f}_k)\right]\right)^2\right] &= \mathbb{E}\left[\left(\widehat{R}_{p,n} - \mathbb{E}\left[\widehat{R}_{p,n}\right]\right)^2\right] + \left(\mathbb{E}\left[\widehat{R}_{p,n}\right] - \mathbb{E}\left[L(\hat{f}_k)\right]\right)^2 \\ &\leq \frac{128\kappa\gamma_d}{\sqrt{2\pi}}\frac{k^{3/2}}{n-p+1} + \left(\frac{4}{\sqrt{2\pi}}\frac{p\sqrt{k}}{n}\right)^2 \quad ,\end{aligned}$$

which concludes the proof.

The proof of Ineq. (5.3) is more intricate and has been postponed to Appendix C.1. □

Keeping in mind that $\mathbb{E}\left[\widehat{R}_{p,n}\right] = R(\mathcal{A}_k^{\mathcal{D}_{n-p}})$, the right-hand side of Ineq. (5.1) is an upper bound on the bias of the LpO estimator, that is on the difference between the risks of the classifiers built from respectively $n-p$ and n points. Therefore, the fact that this upper bound increases with p is reliable since the classifiers $\mathcal{A}_k^{\mathcal{D}_{n-p+1}}(\cdot)$ and $\mathcal{A}_k^{\mathcal{D}_n}(\cdot)$ can become more and more different from one another as p increases. More precisely, the upper bound in Ineq. (5.1) goes to 0 provided $p\sqrt{k}/n$ does. With the additional restriction $p \leq \sqrt{k}$, this reduces to the usual condition $k/n \to 0$ as $n \to +\infty$ (see Devroye et al., 1996, Chap. 6.6 for instance), which is used to prove the universal consistency of the kNN classifier (Stone, 1977). The monotonicity of this upper upper bound with respect to k can seem somewhat unexpected. One could think that the two classifiers would become more and more “similar” to each other as k increases enough. However it can be proved that, in some sense, this dependence cannot be improved in the present distribution-free framework (see Proposition 5.1 and Figure 1).

Note that an upper bound similar to that of Ineq. (5.2) can be easily derived for any order-q moment ($q \geq 2$) at the price of increasing the constants by using $(a+b)^q \leq 2^{q-1}(a^q + b^q)$, for every $a, b \geq 0$. We also emphasize that Ineq. (5.2) allows us to control the discrepancy between the LpO estimator and the risk of the kNN classifier, that is the expectation of its classification error. Ideally we would have liked to replace the risk $R(\hat{f}_k)$ by the prediction error $L(\hat{f}_k)$. But with our strategy of proof, this would require an additional distribution-free concentration inequality on the prediction error of the kNN classifier. To the best of our knowledge, such a concentration inequality is not available up to now.

Upper bounding the squared difference between the LpO estimator and the prediction error is precisely the purpose of Ineq. (5.3). Proving the latter inequality requires a completely different strategy which can be traced back to an earlier proof by Rogers and Wagner

(1978, see the proof of Theorem 2.1) applying to the L1O estimator. Let us mention that Ineq. (5.3) combined with the Jensen inequality lead to a less accurate upper bound than Ineq. (5.1).

Finally the apparent difference between the upper bounds in Ineq. (5.2) and (5.3) results from the completely different schemes of proof. The first one allows us to derive general upper bounds for all centered moments of the LpO estimator, but exhibits a worse dependence with respect to k. By contrast the second one is exclusively dedicated to upper bounding the mean squared difference between the prediction error and the LpO estimator and leads to a smaller $\sqrt{k}$. However (even if probably not optimal), the upper bound used in Ineq. (5.2) still enables to achieve minimax rates over some Hölder balls as proved by Proposition 5.3.

5.2. Lower bounds

5.2.1. Bias of the L1O estimator

The purpose of the next result is to provide a counter-example highlighting that the upper bound of Eq. (5.1) cannot be improved in some sense. We consider the following discrete setting where $\mathcal{X} = \{0, 1\}$ with $\pi_0 = P[X = 0]$, and we define $\eta_0 = \mathbb{P}[Y = 1 \mid X = 0]$ and $\eta_1 = \mathbb{P}[Y = 1 \mid X = 1]$. In what follows this two-class generative model will be referred to as the discrete setting **DS**.

Note that (i) the 3 parameters π_0, η_0 and η_1 fully describe the joint distribution $P_{(X,Y)}$, and (ii) the distribution of **DS** satisfies the strong margin assumption of Massart and Nédélec (2006) if both η_0 and η_1 are chosen away from $1/2$. However this favourable setting has no particular effect on the forthcoming lower bound except a few simplifications along the calculations.

Proposition 5.1. *Let us consider the **DS** setting with $\pi_0 = 1/2$, $\eta_0 = 0$ and $\eta_1 = 1$, and assume that k is odd. Then there exists a numeric constant $C > 1$ independent of n and k such that, for all $n/2 \le k \le n-1$, the kNN classifiers $\mathcal{A}_k^{\mathcal{D}_n}$ and $\mathcal{A}_k^{\mathcal{D}_{n-1}}$ satisfy*

$$\mathbb{E}\left[L\left(\mathcal{A}_k^{\mathcal{D}_n}\right) - L\left(\mathcal{A}_k^{\mathcal{D}_{n-1}}\right) \right] \ge C\frac{\sqrt{k}}{n} \cdot$$

The proof of Proposition 5.1 is provided in Appendix C.2. The rate $\sqrt{k}/n$ in the right-hand side of Eq. (5.1) is then achieved under the generative model **DS** for any $k \ge n/2$. As a consequence this rate cannot be improved without any additional assumption, for instance on the distribution of the X_is. See also Figure 1 below and related comments.

Empirical illustration

To further illustrate the result of Proposition 5.1, we simulated data according to **DS**, for different values of n ranging from 100 to 500 and different values of k ranging from 5 to $n-1$.

Figure 1 (a) displays the evolution of the absolute bias $\left|\mathbb{E}\left[L\left(\mathcal{A}_k^{\mathcal{D}_n}\right) - L\left(\mathcal{A}_k^{\mathcal{D}_{n-1}}\right) \right]\right|$ as a function of k, for several values of n (plain curves). The absolute bias is a nondecreasing function of k, as suggested by the upper bound provided in Eq. (5.1) which is also plotted (dashed lines) to ease the comparison. The non-decreasing behavior of the absolute bias is not always restricted to high values of k (w.r.t. n), as illustrated in Figure 1 (b) which corresponds to **DS** with parameter values $(\pi_0, \eta_0, \eta_1) = (0.2, 0.2, 0.9)$. In particular the non-decreasing behavior of the absolute bias now appears for a range of values of k that are smaller than $n/2$.

Note that a rough idea about the location of the peak, denoted by k_{peak}, can be deduced as follows in the simple case where $\eta_0 = 0$ and $\eta_1 = 1$.

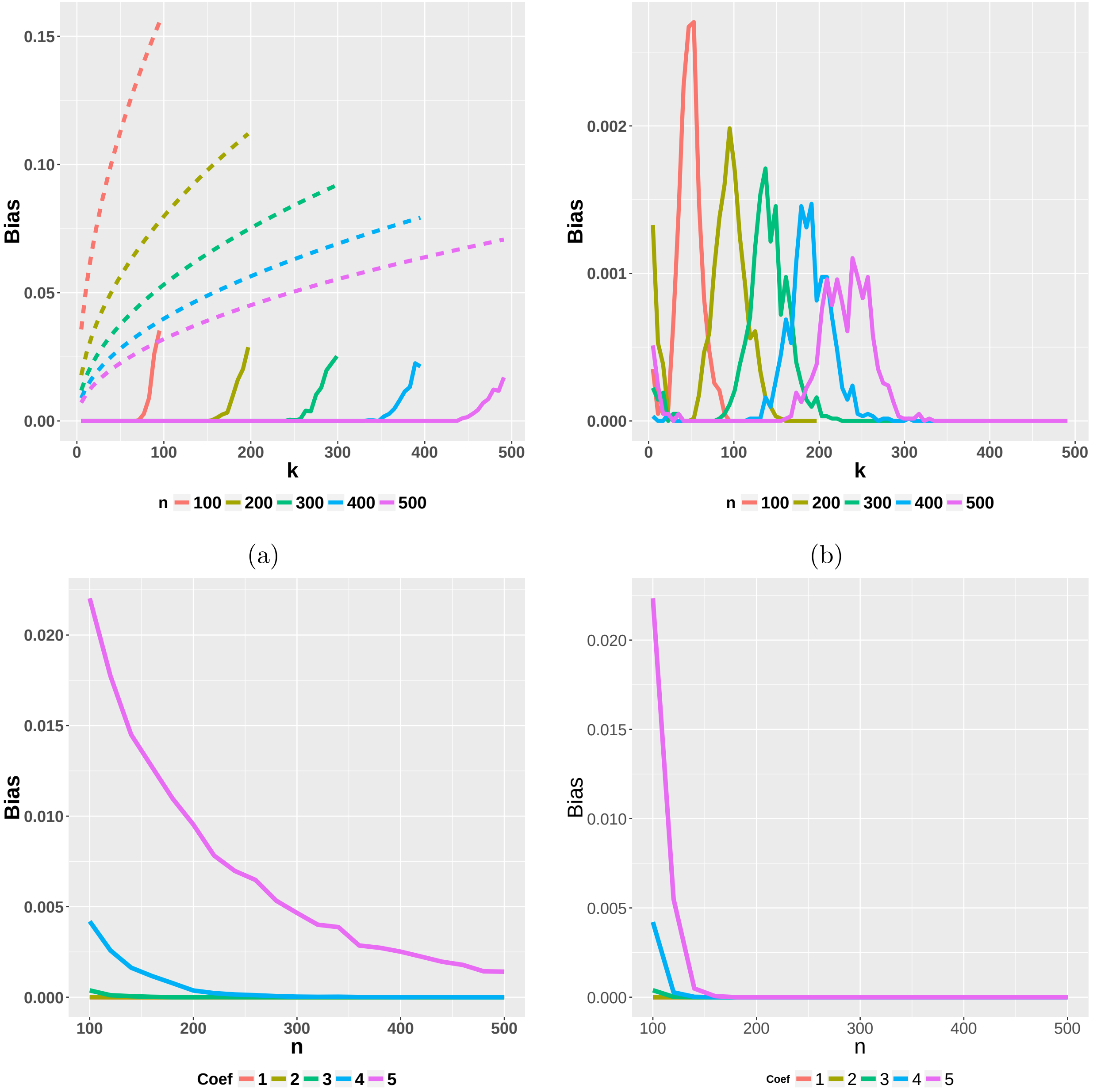
- For the peak to arise, the two classifiers (based on n and respectively $n-1$ observations) have to disagree the most strongly.
- This requires one of the two classifiers – say the first one – to have ties among the k nearest neighbors of each label in at least one of the two cases $X = 0$ or $X = 1$.
- With $\pi_0 < 0.5$, then ties will most likely occur for the case $X = 0$. Therefore the discrepancy between the two classifiers will be the highest at any new observation $x_0 = 0$.
- For the tie situation to arise at x_0, half of its neighbors have to be 1. This only occurs if *(i)* $k > n_0$ (with n_0 the number of observations such that $X = 0$ in the training set), and *(ii)* $k_0\eta_0 + k_1\eta_1 = k/2$, where k_0 (resp. k_1) is the number of neighbors of x_0 such that $X = 0$ (resp. $X = 1$).
- Since $k > n_0$, one has $k_0 = n_0$ and the last expression boils down to $k = \dfrac{n_0(\eta_1 - \eta_0)}{\eta_1 - 1/2}$ ·
- For large values of n, one should have $n_0 \approx n\pi_0$, that is the peak should appear at $k_{peak} \approx \dfrac{n\pi_0(\eta_1 - \eta_0)}{\eta_1 - 1/2}$ ·

In the setting of Proposition 5.1, this reasoning remarkably yields $k_{peak} \approx n$, while it leads to $k_{peak} \approx 0.4n$ in the setting of Figure 1 (b), which is close to the location of the observed peaks. This also suggests that even smaller values of k_{peak} can arise by tuning the parameter π_0 close to 0. Let us mention that very similar curves have been obtained for a Gaussian mixture model with two disjoint classes (not reported here). On the one hand this empirically illustrates that the $\sqrt{k}/n$ rate is not limited to **DS** (discrete setting). On the other hand, all of this confirms that this rate cannot be improved in the present distribution-free framework.

Let us finally consider Figure 1 (c), which displays the absolute bias as a function of n where $k = \lfloor \text{Coef} \times n \rfloor$ for different values of Coef, where $\lfloor\cdot\rfloor$ denotes the integer part. With this choice of k, Proposition 5.1 implies that the absolute bias should decrease at a $1/\sqrt{n}$ rate, which is supported by the plotted curves. By contrast, panel (d) of Figure 1 illustrates that choosing smaller values of k, that is $k = \lfloor \text{Coef} \times \sqrt{n} \rfloor$, leads to a faster decreasing rate.

5.2.2. Mean squared error

Following an example described by Devroye and Wagner (1979a), we now provide a lower bound on the minimal convergence rate of the mean squared error (see also Devroye et al., 1996, Chap. 24.4, p.415 for a similar argument).



(a)

(b)

(c)

(d)

Figure 1: **(a)** Evolution of the absolute value of the bias as a function of k, for different values of n (plain lines). The dashed lines correspond to the upper bound obtained in (5.1). **(b)** Same as previous, except that data were generated according to the **DS** setting with parameters $(\pi_0, \eta_0, \eta_1) = (0.2, 0.2, 0.9)$. Upper bounds are not displayed in order to fit the scale of the absolute bias. **(c)** Evolution of the absolute value of the bias with respect to n, when k is chosen such that $k = \lfloor \text{Coef} \times n \rfloor$ ($\lfloor \cdot \rfloor$ denotes the integer part). The different colors correspond to different values of Coef. **(d)** Same as previous, except that k is chosen such that $k = \lfloor \text{Coef} \times \sqrt{n} \rfloor$.

Proposition 5.2. *Let us assume n is even, and that $P(Y=1 \mid X) = P(Y=1) = 1/2$ is independent of X. Then for $k = n-1$ (k odd), it results*

$$\mathbb{E}\left[\left(\widehat{R}_{1,n} - L(\hat{f}_k) \right)^2 \right] = \int_0^1 2t \cdot \mathbb{P}\left[\left| \widehat{R}_{1,n} - L(\hat{f}_k) \right| > t \right] dt \geq \frac{1}{8\sqrt{\pi}} \cdot \frac{1}{\sqrt{n}} \cdot$$

From the upper bound of order $\sqrt{k}/n$ provided by Ineq. (5.3) (with $p=1$), choosing $k = n-1$ leads to the same $1/\sqrt{n}$ rate as that of Proposition 5.2. This suggests that, at least for very large values of k, the $\sqrt{k}/n$ rate is of the right order and cannot be improved in the distribution-free framework.

5.3. Minimax rates

Let us conclude this section with a corollary, which provides a finite-sample bound on the gap between $\widehat{R}_{p,n}$ and $R(\hat{f}_k) = \mathbb{E}\left[L(\hat{f}_k) \right]$ with high probability. It is stated under the same restriction on p as the previous Theorem 5.1 it is based on, that is for $p \leq \sqrt{k}$.

Corollary 5.1. *With the notation of Theorems 4.2 and 5.1, let us assume $p, k \geq 1$ with $p \leq \sqrt{k}$, $\sqrt{k} + k \leq n$, and $p \leq n/2 + 1$. Then for every $x > 0$, there exists an event with probability at least $1 - 2e^{-x}$ such that*

$$\left| R(\hat{f}_k) - \widehat{R}_{p,n} \right| \leq \sqrt{\frac{\Delta^2 k^2}{n\left(1 - \frac{p-1}{n}\right)} x} + \frac{4}{\sqrt{2\pi}} \frac{p\sqrt{k}}{n} \ , \tag{5.4}$$

where $\hat{f}_k = \mathcal{A}_k^{\mathcal{D}_n}(\cdot)$.

Proof of Corollary 5.1. Ineq. (5.4) results from combining the exponential concentration result derived for $\widehat{R}_{p,n}$, namely Ineq. (4.2) (from Theorem 4.2) and the upper bound on the bias, that is Ineq. (5.1).

$$\begin{aligned} \left| R(\hat{f}_k) - \widehat{R}_{p,n} \right| &\leq \left| R(\hat{f}_k) - \mathbb{E}\left[\widehat{R}_{p,n} \right] \right| + \left| \mathbb{E}\left[\widehat{R}_{p,n} \right] - \widehat{R}_{p,n} \right| \\ &\leq \frac{4}{\sqrt{2\pi}} \frac{p\sqrt{k}}{n} + \sqrt{\frac{\Delta^2 k^2}{n - p + 1} x} \ \cdot \end{aligned}$$

□

Note that the right-hand side of Ineq. (5.4) could be used to derive bounds on $R(\hat{f}_k)$ that seem similar to confidence bounds. However we do not recommend doing this in practice for several reasons. On the one hand, Ineq. (5.4) results from the repeated use of concentration inequalities where numeric constants are not optimized at all. This would lead to require a large sample size n for the deviation terms to be small in practice. On the other hand, explicit numeric constants such as Δ^2 in Corollary 5.1 exhibit a dependence on $\gamma_d \approx 4.8^d - 1$, which becomes exponentially large as d increases. Proving that this dependence can be weakened or not remains a completely open question at this stage. Nevertheless one can highlight that, for a given n, increasing d will quickly make the deviation term larger than 1, whereas both $R(\hat{f}_k)$ and $\widehat{R}_{p,n}$ belong to $[0, 1]$.

The right-most term of order $\sqrt{k}/n$ in Ineq. (5.4) results from the bias. This is a necessary price to pay which cannot be improved in the present distribution-free framework according to Proposition 5.1. Besides combining the restriction $p \leq \sqrt{k}$ with the usual consistency constraint $k/n = o(1)$ leads to the conclusion that small values of p (w.r.t. n) have almost no effect on the convergence rate of the LpO estimator. Weakening the key restriction $p \leq \sqrt{k}$ would be necessary to potentially nuance this conclusion.

In order to highlight the interest of the above deviation inequality, let us deduce an optimality result in terms of minimax rate over Hölder balls $\mathcal{H}(\tau, \alpha)$ defined by

$$\mathcal{H}(\tau,\alpha) = \left\{ g : \; \mathbb{R}^d \mapsto \mathbb{R}, \quad |g(x) - g(y)| \leq \tau \left\| x - y \right\|^{\alpha} \right\},$$

with $\alpha \in]0,1[$ and $\tau > 0$. In the following statement, Corollary 5.1 is used to prove that, uniformly with respect to k, the LpO estimator $\widehat{R}_{p,n}$ and the risk $R(\hat{f}_k)$ of the kNN classifier remain close to each other with high probability.

Proposition 5.3. *With the same notation as Corollary 5.1, for every $C > 1$ and $\theta > 0$, there exists an event of probability at least $1 - 2 \cdot n^{-(C-1)}$ on which, for any $p, k \geq 1$ such that $p \leq \sqrt{k}$, $k + \sqrt{k} \leq n$, and $p \leq n/2+1$, the LpO estimator of the kNN classifier satisfies*

$$\begin{aligned}(1-\theta)\left[R(\hat{f}_k) - L^\star \right] - \frac{\theta^{-1}\Delta^2 C}{4} \frac{k^2 \log(n)}{n\left(R(\hat{f}_k) - L^\star\right)} - \frac{4}{\sqrt{2\pi}} \frac{p\sqrt{k}}{n} \\ \leq \widehat{R}_p(\mathcal{A}_k, \mathcal{D}_n) - L^\star \leq (1+\theta)\left[R(\hat{f}_k) - L^\star \right] + \frac{\theta^{-1}\Delta^2 C}{4} \frac{k^2 \log(n)}{n\left(R(\hat{f}_k) - L^\star\right)} + \frac{4}{\sqrt{2\pi}} \frac{p\sqrt{k}}{n},\end{aligned} \tag{5.5}$$

where $L^\star$ denotes the classification error of the Bayes classifier.

Furthermore if one assumes the regression function η belongs to a Hölder ball $\mathcal{H}(\tau, \alpha)$ for some $\alpha \in]0, \min(d/4, 1)[$ (recall that $X_i \in \mathbb{R}^d$) and $\tau > 0$, then choosing $k = k^\star = k_0 \cdot n^{\frac{2\alpha}{2\alpha+d}}$ leads to

$$\widehat{R}_p(\mathcal{A}_{k^\star}, \mathcal{D}_n) - L^\star \sim_{n \to +\infty} R(\hat{f}_{k^\star}) - L^\star, \qquad a.s. \; . \tag{5.6}$$

Ineq. (5.5) gives a uniform control (over k) of the gap between the excess risk $R(\hat{f}_k) - L^\star$ and the corresponding LpO estimator $\widehat{R}_p(\hat{f}_k) - L^\star$ with high probability. The decreasing rate (in $n^{-(C-1)}$) of this probability is directly related to the $\log(n)$ factor in the lower and upper bounds. This decreasing rate could be made faster at the price of increasing the exponent of the $\log(n)$ factor. In a similar way the numeric constant θ has no precise meaning and can be chosen as close to 0 as we want, leading to increase one of the other deviation terms by a numeric factor θ^{-1}. For instance one could choose $\theta = 1/\log(n)$, which would replace the $\log(n)$ by a $(\log n)^2$.

The equivalence established by Eq. (5.6) results from knowing that this choice $k = k^\star$ makes the kNN classifier achieve the minimax rate $n^{-\frac{\alpha}{2\alpha+d}}$ over Hölder balls (Yang, 1999). This holds true for $\alpha \in]0,1[$ as long as $d \geq 4$. However if $d < 4$ the minimax rate is only achieved over $]0, d/4[$. This limitation results from the dependence of the deviation terms with respect to k^2 in Eq. (5.5), which is not optimal and should be further improved.

Proof of Proposition 5.3. Let us define $K \leq n$ as the maximum value of k and assume $x_k = C \cdot \log(n)$ (for some constant $C > 1$) for any $1 \leq k \leq K$. Let us also introduce the event

$$\Omega_n = \left\{ \forall 1 \leq k \leq K, \ \left| R(\hat{f}_k) - \widehat{R}_p(\mathcal{A}_k, \mathcal{D}_n) \right| \leq \sqrt{\Delta^2 \frac{k^2}{n} x_k} + \frac{4}{\sqrt{2\pi}} \frac{p\sqrt{k}}{n} \right\}.$$

Then $\mathbb{P}\left[\Omega_n^c \right] \leq \frac{1}{n^{C-1}} \to 0$, as $n \to +\infty$, since a union bound leads to

$$\sum_{k=1}^{K} e^{-x_k} = \sum_{k=1}^{K} e^{-C \cdot \log(n)} = K \cdot e^{-C \cdot \log(n)} \leq e^{-(C-1) \cdot \log(n)} = \frac{1}{n^{C-1}} \ \cdot$$

Furthermore combining (for $a, b > 0$) the inequality $ab \leq a^2\theta^2 + b^2\theta^{-2}/4$ for every $\theta > 0$ with $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, it results that

$$\begin{aligned} \sqrt{\Delta^2 \frac{k^2}{n} x_k} &\leq \theta \left(R(\hat{f}_k) - L^\star \right) + \frac{\theta^{-1}}{4} \Delta^2 \frac{k^2}{n \left(R(\hat{f}_k) - L^\star \right)} x_k \\ &\leq \theta \left(R(\hat{f}_k) - L^\star \right) + \frac{\theta^{-1}}{4} \Delta^2 \frac{k^2}{n \left(R(\hat{f}_k) - L^\star \right)} C \cdot \log(n), \end{aligned}$$

hence Ineq. (5.5).

Let us now prove the next equivalence, namely (5.6), by means of the Borel-Cantelli lemma.

First Yang (1999) combined with Theorem 7 in Chaudhuri and Dasgupta (2014) provide that the minimax rate over the Hölder ball $\mathcal{H}(\tau, \alpha)$ is achieved by the kNN classifier with $k = k^\star$), that is

$$\left(R(\hat{f}_{k^\star}) - L^\star \right) \asymp n^{-\frac{\alpha}{2\alpha+d}},$$

where $a \asymp b$ means there exist numeric constants $l, u > 0$ such that $l \cdot b \leq a \leq u \cdot b$. Moreover it is then easy to check that

- $D_1 := \frac{\theta^{-1}}{4} \Delta^2 \frac{k^{\star 2}}{n\left(R(\hat{f}_{k^\star}) - L^\star\right)} C \cdot \log(n) \asymp \frac{C\theta^{-1}\Delta^2 k_0}{4} \cdot \left(n^{-\frac{d-3\alpha}{2\alpha+d}} \log(n) \right) = o_{n\to+\infty} \left(R(\hat{f}_{k^\star}) - L^\star \right),$
- $D_2 := \frac{p\sqrt{k^\star}}{n} \leq \frac{k^\star}{n} = k_0 \cdot n^{-\frac{d}{2\alpha+d}} = o_{n\to+\infty} \left(R(\hat{f}_{k^\star}) - L^\star \right).$

Besides

$$\begin{aligned} \frac{1}{n^{C-1}} \geq \mathbb{P}\left[\Omega_n^c \right] &\geq \mathbb{P}\left[\left| \frac{\left(\widehat{R}_p(\mathcal{A}_{k^\star}, \mathcal{D}_n) - L^\star \right) - \left(R(\hat{f}_{k^\star}) - L^\star \right)}{R(\hat{f}_{k^\star}) - L^\star} \right| > \theta + \frac{D_1 + D_2}{R(\hat{f}_{k^\star}) - L^\star} \right] \\ &= \mathbb{P}\left[\left| \frac{\left(\widehat{R}_p(\mathcal{A}_{k^\star}, \mathcal{D}_n) - L^\star \right)}{R(\hat{f}_{k^\star}) - L^\star} - 1 \right| > \theta + \frac{D_1 + D_2}{R(\hat{f}_{k^\star}) - L^\star} \right]. \end{aligned}$$

Let us now choose any $\epsilon > 0$ and introduce the sequence of events $\{A_n(\epsilon)\}_{n\geq 1}$ such that

$$A_n(\epsilon) = \left\{ \left| \frac{\left(\widehat{R}_p(\mathcal{A}_{k^\star}, \mathcal{D}_n) - L^\star\right)}{R(\hat{f}_{k^\star}) - L^\star} - 1 \right| > \epsilon \right\} .$$

Using that D_1 and D_2 are negligible with respect to $R(\hat{f}_{k^\star}) - L^\star$ as $n \to +\infty$, there exists an integer $n_0 = n_0(\epsilon) > 0$ such that, for all $n \geq n_0$ and with $\theta = \epsilon/2$,

$$\theta + \frac{D_1 + D_2}{R(\hat{f}_{k^\star}) - L^\star} \leq \epsilon.$$

Hence

$$\mathbb{P}\left[\, A_n(\epsilon) \,\right] \leq \mathbb{P}\left[\left\{ \left| \frac{\left(\widehat{R}_p(\mathcal{A}_{k^\star}, \mathcal{D}_n) - L^\star\right)}{R(\hat{f}_{k^\star}) - L^\star} - 1 \right| > \theta + \frac{D_1 + D_2}{R(\hat{f}_{k^\star}) - L^\star} \right\} \right] \leq \frac{1}{n^{C-1}} \ \cdot$$

Finally, choosing any $C > 2$ leads to $\sum_{n=1}^{+\infty} \mathbb{P}\left[\, A_n(\epsilon) \,\right] < +\infty$, which provides the expected conclusion by means of the Borel-Cantelli lemma. □

6. Discussion

The present work provides several new results quantifying the performance of the L*p*O estimator applied to the *k*NN classifier. By exploiting the connexion between L*p*O and U-statistics (Section 2), the polynomial and exponential inequalities derived in Sections 3 and 4 give some new insight on the concentration of the L*p*O estimator around its expectation for different regimes of p/n. In Section 5, these results serve for instance to conclude to the consistency of the L*p*O estimator towards the risk (or the classification error rate) of the *k*NN classifier (Theorem 5.1). They also allow us to establish the asymptotic equivalence between the L*p*O estimator (shifted by the Bayes risk $L^\star$) and the excess risk over some Hölder class of regression functions (Proposition 5.3).

It is worth mentioning that the upper-bounds derived in Sections 4 and 5 — see for instance Theorem 5.1 — can be minimized by choosing $p = 1$, suggesting that the L1O estimator is optimal in terms of risk estimation when applied to the *k*NN classification algorithm. This observation corroborates the results of the simulation study presented in Celisse and Mary-Huard (2011), where it is empirically shown that small values of p (and in particular $p = 1$) lead to the best estimation of the risk for any fixed k, whatever the level of noise in the data. The suggested optimality of L1O (for risk estimation) is also consistent with results by Burman (1989) and Celisse (2014), where it is proved that L1O is asymptotically the best cross-validation procedure to perform risk estimation in the context of low-dimensional regression and density estimation respectively.

Alternatively, the L*p*O estimator can also be used as a data-dependent calibration procedure to tune k, by choosing the value $\hat{k}_p$ which minimizes the L*p*O estimate. For instance in classification, LpO can be used to get the value of k leading to the best prediction performance. In this context the value of p (the splitting ratio) leading to the best kNN classifier can be very different from $p = 1$. This is illustrated by the simulation results summarized by Figure 2 in Celisse and Mary-Huard (2011) where p has to be larger than 1 as the noise level becomes strong. This phenomenon is not limited to the kNN classifier, but extends to various estimation/prediction problems (Breiman and Spector, 1992; Arlot and Lerasle, 2012; Celisse, 2014). If we turn now to the question of identifying the best predictor among several candidates, choosing $p = 1$ also leads to poor selection performances as proved by Shao (1993, Eq. (3.8)) with the linear regression model. For the LpO, Shao (1997, Theorem 5) proves the model selection consistency if $p/n \to 1$ and $n - p \to +\infty$ as $n \to +\infty$. For recovering the best predictor among two candidates, Yang (2006, 2007) proved the consistency of CV under conditions relating the optimal splitting ratio p to the convergence rates of the predictors to be compared, and further requiring that $\min(p, n - p) \to +\infty$ as $n \to +\infty$.

Although the focus of the present paper is different, it is worth mentioning that the concentration results established in Section 4 are a significant early step towards deriving theoretical guarantees on LpO as a model selection procedure. Indeed, exponential concentration inequalities have been a key ingredient to assess model selection consistency or model selection efficiency in various contexts (see for instance Celisse (2014) or Arlot and Lerasle (2012) in the density estimation framework). Still theoretically investigating the behavior of $\hat{k}_p$ requires some further dedicated developments. One first step towards such results is to derive a tighter upper bound on the bias between the LpO estimator and the risk. The best known upper bound currently available is derived from Devroye and Wagner (1980, see Lemma D.6 in the present paper). Unfortunately it does not fully capture the true behavior of the LpO estimator with respect to p (at least as p becomes large) and could be improved in particular for $p > \sqrt{k}$ as emphasized in the comments following Theorem 5.1. Another important direction for studying the model selection behavior of the LpO procedure is to prove a concentration inequality for the classification error rate of the kNN classifier around its expectation. While such concentration results have been established for the kNN algorithm in the (fixed-design) regression framework (Arlot and Bach, 2009), deriving similar results in the classification context remains a challenging problem to the best of our knowledge.

Acknowledgments

The authors would like to thank the associate editor and reviewers for their highlightful comments which greatly helped to improve the presentation of the paper. This work was partially funded by the “BeFast” PEPS CNRS.

Appendix A. Proofs of polynomial moment upper bounds

A.1. Proof of Theorem 2.2

The proof relies on Proposition 2.1 that allows to relate the LpO estimator to a sum of independent random variables. In the following, we distinguish between the two settings $q = 2$ (where exact calculations can be carried out), and $q > 2$ where only upper bounds can be derived.

When $q > 2$, our proof deals separately with the cases $p \leq n/2 + 1$ and $p > n/2 + 1$. In the first one, a straightforward use of Jensen's inequality leads to the result. In the second setting, one has to be more cautious when deriving upper bounds. This is done by using the more sophisticated Rosenthal's inequality, namely Proposition D.2.

A.1.1. Exploiting Proposition 2.1

According to the proof of Proposition 2.1, it arises that the LpO estimator can be expressed as a U-statistic since

$$\widehat{R}_{p,n} = \frac{1}{n!} \sum_{\sigma} W\left(Z_{\sigma(1)}, \ldots, Z_{\sigma(n)}\right) \quad ,$$

with

$$\begin{aligned} W\left(Z_1, \ldots, Z_n\right) &= \left\lfloor \frac{n}{m} \right\rfloor^{-1} \sum_{a=1}^{\lfloor \frac{n}{m} \rfloor} h_m\left(Z_{(a-1)m+1}, \ldots, Z_{am}\right) \qquad \text{(with } m = n-p+1\text{)} \\ \text{and} \quad h_m\left(Z_1, \ldots, Z_m\right) &= \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{\left\{\mathcal{A}^{\mathcal{D}_m^{(i)}}(X_i) \neq Y_i\right\}} = \widehat{R}_{1,n-p+1} \quad , \end{aligned}$$

where $\mathcal{A}^{\mathcal{D}_m^{(i)}}(.)$ denotes the classifier based on sample $\mathcal{D}_m^{(i)} = (Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_m)$. Further centering the LpO estimator, it comes

$$\widehat{R}_{p,n} - \mathbb{E}\left[\widehat{R}_{p,n}\right] = \frac{1}{n!} \sum_{\sigma} \bar{W}\left(Z_{\sigma(1)}, \ldots, Z_{\sigma(n)}\right),$$

where $\bar{W}(Z_1, \ldots, Z_n) = W(Z_1, \ldots, Z_n) - \mathbb{E}\left[W(Z_1, \ldots, Z_n)\right]$.

Then with $\bar{h}_m(Z_1, \ldots, Z_m) = h_m(Z_1, \ldots, Z_m) - \mathbb{E}\left[h_m(Z_1, \ldots, Z_m)\right]$, one gets

$$\begin{aligned} \mathbb{E}\left[\left|\widehat{R}_{p,n} - \mathbb{E}\left[\widehat{R}_{p,n}\right]\right|^q\right] &\leq \mathbb{E}\left[\left|\bar{W}\left(Z_1, \ldots, Z_n\right)\right|^q\right] \quad \text{(Jensen's inequality)} \\ &= \mathbb{E}\left[\left|\left\lfloor \frac{n}{m} \right\rfloor^{-1} \sum_{i=1}^{\lfloor \frac{n}{m} \rfloor} \bar{h}_m\left(Z_{(i-1)m+1}, \ldots, Z_{im}\right)\right|^q\right] \qquad \text{(A.1)} \\ &= \left\lfloor \frac{n}{m} \right\rfloor^{-q} \mathbb{E}\left[\left|\sum_{i=1}^{\lfloor \frac{n}{m} \rfloor} \bar{h}_m\left(Z_{(i-1)m+1}, \ldots, Z_{im}\right)\right|^q\right] . \end{aligned}$$

A.1.2. The setting $q = 2$

If $q = 2$, then by independence it comes

$$\begin{aligned}\mathbb{E}\left[\left|\widehat{R}_{p,n} - \mathbb{E}\left[\widehat{R}_{p,n}\right]\right|^q\right] &\leq \left\lfloor \frac{n}{m} \right\rfloor^{-2} \mathrm{Var}\left(\sum_{i=1}^{\lfloor \frac{n}{m} \rfloor} h_m\left(Z_{(i-1)m+1}, \ldots, Z_{im}\right)\right) \\ &= \left\lfloor \frac{n}{m} \right\rfloor^{-2} \sum_{i=1}^{\lfloor \frac{n}{m} \rfloor} \mathrm{Var}\left[\, h_m\left(Z_{(i-1)m+1}, \ldots, Z_{im}\right)\right] \\ &= \left\lfloor \frac{n}{m} \right\rfloor^{-1} \mathrm{Var}\left(\widehat{R}_1(\mathcal{A}, Z_{1,n-p+1})\right),\end{aligned}$$

which leads to the result.

A.1.3. The setting $q > 2$

If $p \leq n/2 + 1$:

A straightforward use of Jensen's inequality from (A.1) provides

$$\begin{aligned}\mathbb{E}\left[\left|\widehat{R}_{p,n} - \mathbb{E}\left[\widehat{R}_{p,n}\right]\right|^q\right] &\leq \left\lfloor \frac{n}{m} \right\rfloor^{-1} \sum_{i=1}^{\lfloor \frac{n}{m} \rfloor} \mathbb{E}\left[\left|\bar{h}_m\left(Z_{(i-1)m+1}, \ldots, Z_{im}\right)\right|^q\right] \\ &= \mathbb{E}\left[\left|\widehat{R}_{1,n-p+1} - \mathbb{E}\left[\,\widehat{R}_{1,n-p+1}\,\right]\right|^q\right].\end{aligned}$$

If $p > n/2 + 1$: Let us now use Rosenthal's inequality (Proposition D.2) by introducing symmetric random variables $\zeta_1, \ldots, \zeta_{\lfloor n/m \rfloor}$ such that

$$\forall 1 \leq i \leq \lfloor n/m \rfloor, \quad \zeta_i = h_m\left(Z_{(i-1)m+1}, \ldots, Z_{im}\right) - h_m\left(Z'_{(i-1)m+1}, \ldots, Z'_{im}\right),$$

where $Z'_1, \ldots, Z'_n$ are *i.i.d.* copies of $Z_1, \ldots, Z_n$. Then it comes for every $\gamma > 0$

$$\mathbb{E}\left[\left|\sum_{i=1}^{\lfloor \frac{n}{m} \rfloor} \bar{h}_m\left(Z_{(i-1)m+1}, \ldots, Z_{im}\right)\right|^q\right] \leq \mathbb{E}\left[\left|\sum_{i=1}^{\lfloor \frac{n}{m} \rfloor} \zeta_i\right|^q\right],$$

which implies

$$\mathbb{E}\left[\left|\sum_{i=1}^{\lfloor \frac{n}{m} \rfloor} \bar{h}_m\left(Z_{(i-1)m+1}, \ldots, Z_{im}\right)\right|^q\right] \leq B(q, \gamma) \max\left\{\gamma \sum_{i=1}^{\lfloor \frac{n}{m} \rfloor} \mathbb{E}\left[\,|\zeta_i|^q\,\right], \left(\sqrt{\sum_{i=1}^{\lfloor \frac{n}{m} \rfloor} \mathbb{E}\left[\,\zeta_i^2\,\right]}\right)^q\right\}.$$

Then using for every i that

$$\mathbb{E}\left[\,|\zeta_i|^q\,\right] \leq 2^q \mathbb{E}\left[\,\left|\bar{h}_m\left(Z_{(i-1)m+1}, \ldots, Z_{im}\right)\right|^q\,\right],$$

it comes

$$\mathbb{E}\left[\left|\sum_{i=1}^{\lfloor \frac{n}{m}\rfloor} \bar{h}_m\left(Z_{(i-1)m+1},\ldots,Z_{im}\right)\right|^q\right]$$
$$\leq B(q,\gamma)\max\left(2^q\gamma\left\lfloor\frac{n}{m}\right\rfloor\mathbb{E}\left[\left|\widehat{R}_{1,m}-\mathbb{E}\left[\widehat{R}_{1,m}\right]\right|^q\right],\left(\sqrt{\left\lfloor\frac{n}{m}\right\rfloor 2\mathrm{Var}\left(\widehat{R}_{1,m}\right)}\right)^q\right).$$

Hence, it results for every $q > 2$

$$\mathbb{E}\left[\left|\widehat{R}_{p,n}-\mathbb{E}\left[\widehat{R}_{p,n}\right]\right|^q\right]$$
$$\leq B(q,\gamma)\max\left(2^q\gamma\left\lfloor\frac{n}{m}\right\rfloor^{-q+1}\mathbb{E}\left[\left|\widehat{R}_{1,m}-\mathbb{E}\left[\widehat{R}_{1,m}\right]\right|^q\right],\left\lfloor\frac{n}{m}\right\rfloor^{-q/2}\left(\sqrt{2\mathrm{Var}\left(\widehat{R}_{1,m}\right)}\right)^q\right),$$

which concludes the proof.

A.2. Proof of Theorem 3.1

Our strategy of proof follows several ideas. The first one consists in using Proposition 3.1 which says that, for every $q \geq 2$,

$$\left\|\bar{h}_m(Z_1,\ldots,Z_m)\right\|_q \leq \sqrt{2\kappa q}\sqrt{\left\|\sum_{j=1}^m\left(h_m(Z_1,\ldots,Z_m)-h_m(Z_1,\ldots,Z_j',\ldots,Z_m)\right)^2\right\|_{q/2}},$$

where $h_m(Z_1,\ldots,Z_m)=\widehat{R}_{1,m}$ by Eq. (2.4), and $\bar{h}_m(Z_1,\ldots,Z_m)=h_m(Z_1,\ldots,Z_m)-\mathbb{E}\left[h_m(Z_1,\ldots,Z_m)\right]$. The second idea consists in deriving upper bounds of

$$\Delta^j h_m = h_m(Z_1,\ldots,Z_j,\ldots,Z_m)-h_m(Z_1,\ldots,Z_j',\ldots,Z_m)$$

by repeated uses of Stone's lemma, that is Lemma D.5 which upper bounds by $k\gamma_d$ the maximum number of X_is that can have a given X_j among their k nearest neighbors. Finally, for technical reasons we have to distinguish the case $q = 2$ (where we get tighter bounds) and $q > 2$.

A.2.1. Upper bounding $\Delta^j h_m$

For the sake of readability let us now use the notation $\mathcal{D}^{(i)} = \mathcal{D}_m^{(i)}$ (see Theorem 2.1), and let $\mathcal{D}_j^{(i)}$ denote the set$\left(Z_1,\ldots,Z_j',\ldots,Z_n\right)$ where the i-th coordinate has been removed. Then, $\Delta^j h_m = h_m(Z_1,\ldots,Z_m)-h_m(Z_1,\ldots,Z_j',\ldots,Z_m)$ is now upper bounded by

$$\left|\Delta^j h_m\right| \leq \frac{1}{m}+\frac{1}{m}\sum_{i\neq j}\left|\mathbb{1}_{\left\{\mathcal{A}_k^{\mathcal{D}^{(i)}}(X_i)\neq Y_i\right\}}-\mathbb{1}_{\left\{\mathcal{A}_k^{\mathcal{D}_j^{(i)}}(X_i)\neq Y_i\right\}}\right|$$
$$\leq \frac{1}{m}+\frac{1}{m}\sum_{i\neq j}\left|\mathbb{1}_{\left\{\mathcal{A}_k^{\mathcal{D}^{(i)}}(X_i)\neq\mathcal{A}_k^{\mathcal{D}_j^{(i)}}(X_i)\right\}}\right|. \quad \text{(A.2)}$$

Furthermore, let us introduce for every $1 \leq j \leq n$,

$$A_j = \{1 \leq i \leq m,\ i \neq j,\ j \in V_k(X_i)\} \text{ and } A'_j = \left\{1 \leq i \leq m,\ i \neq j,\ j \in V'_k(X_i)\right\}$$

where $V_k(X_i)$ and $V'_k(X_i)$ denote the indices of the k nearest neighbors of X_i respectively among $X_1, \ldots, X_{j-1}, X_j, X_{j+1}, \ldots, X_m$ and $X_1, ..., X_{j-1}, X'_j, X_{j+1}, \ldots, X_m$. Setting $B_j = A_j \cup A'_j$, one obtains

$$\left|\Delta^j h_m\right| \leq \frac{1}{m} + \frac{1}{m} \sum_{i \in B_j} \left| \mathbb{1}_{\left\{\mathcal{A}_k^{\mathcal{D}^{(i)}}(X_i) \neq \mathcal{A}_k^{\mathcal{D}_j^{(i)}}(X_i)\right\}} \right| . \tag{A.3}$$

From now on, we distinguish between $q = 2$ and $q > 2$ because we will be able to derive a tighter bound for $q = 2$ than for $q > 2$.

A.2.2. Case $q > 2$

From (A.3), Stone's lemma (Lemma D.5) provides

$$\left|\Delta^j h_m\right| \leq \frac{1}{m} + \frac{1}{m} \sum_{i \in B_j} \mathbb{1}_{\left\{\mathcal{A}_k^{\mathcal{D}^{(i)}}(X_i) \neq \mathcal{A}_k^{\mathcal{D}_j^{(i)}}(X_i)\right\}} \leq \frac{1}{m} + \frac{2k\gamma_d}{m} \ \cdot$$

Summing over $1 \leq j \leq n$ and applying $(a+b)^q \leq 2^{q-1}\left(a^q + b^q\right)$ ($a, b \geq 0$ and $q \geq 1$), it comes

$$\sum_j \left(\Delta^j h_m\right)^2 \leq \frac{2}{m}\left(1 + (2k\gamma_d)^2\right) \leq \frac{4}{m}(2k\gamma_d)^2 \ ,$$

hence

$$\left\| \sum_{j=1}^m \left(h_m(Z_1, \ldots, Z_m) - h_m(Z_1, \ldots, Z'_j, \ldots, Z_m)\right)^2 \right\|_{q/2} \leq \frac{4}{m}(2k\gamma_d)^2.$$

This leads for every $q > 2$ to

$$\left\|\bar{h}_m(Z_1, \ldots, Z_m)\right\|_q \leq q^{1/2}\sqrt{2\kappa}\frac{4k\gamma_d}{\sqrt{m}} \ ,$$

which enables to conclude.

A.2.3. Case $q = 2$

It is possible to obtain a slightly better upper bound in the case $q = 2$ with the following reasoning. With the same notation as above and from (A.3), one has

$$\begin{aligned} \mathbb{E}\left[\left(\Delta^j h_m\right)^2\right] &= \frac{2}{m^2} + \frac{2}{m^2}\mathbb{E}\left[\left(\sum_{i \in B_j} \mathbb{1}_{\left\{\mathcal{A}_k^{\mathcal{D}^{(i)}}(X_i) \neq \mathcal{A}_k^{\mathcal{D}_j^{(i)}}(X_i)\right\}}\right)^2\right] \qquad \text{(using } \mathbb{1}_{\{\cdot\}} \leq 1\text{)} \\ &\leq \frac{2}{m^2} + \frac{2}{m^2}\mathbb{E}\left[|B_j| \sum_{i \in B_j} \mathbb{1}_{\left\{\mathcal{A}_k^{\mathcal{D}^{(i)}}(X_i) \neq \mathcal{A}_k^{\mathcal{D}_j^{(i)}}(X_i)\right\}}\right] . \end{aligned}$$

Lemma D.5 implies $|B_j| \leq 2k\gamma_d$, which allows to conclude

$$\mathbb{E}\left[\left(\Delta^j h_m\right)^2\right] \leq \frac{2}{m^2} + \frac{4k\gamma_d}{m^2}\mathbb{E}\left[\sum_{i\in B_j} \mathbb{1}_{\left\{\mathcal{A}_k^{\mathcal{D}^{(i)}}(X_i)\neq \mathcal{A}_k^{\mathcal{D}_j^{(i)}}(X_i)\right\}}\right] .$$

Summing over j and introducing an independent copy of Z_1 denoted by Z_0, one derives

$$\begin{aligned}
&\sum_{j=1}^m \mathbb{E}\left[\left(h_m(Z_1,\ldots,Z_m) - h_m(Z_1,\ldots,Z_j',\ldots,Z_m)\right)^2\right] \\
&\leq \frac{2}{m} + \frac{4k\gamma_d}{m}\sum_{i=1}^m \mathbb{E}\left[\mathbb{1}_{\left\{\mathcal{A}_k^{\mathcal{D}^{(i)}}(X_i)\neq \mathcal{A}_k^{\mathcal{D}^{(i)}\cup Z_0}(X_i)\right\}} + \mathbb{1}_{\left\{\mathcal{A}_k^{\mathcal{D}^{(i)}\cup Z_0}(X_i)\neq \mathcal{A}_k^{\mathcal{D}_j^{(i)}}(X_i)\right\}}\right] \\
&\leq \frac{2}{m} + 4k\gamma_d \times 2\frac{4\sqrt{k}}{\sqrt{2\pi}m} = \frac{2}{m} + \frac{32\gamma_d}{\sqrt{2\pi}}\frac{k\sqrt{k}}{m} \leq (2+16\gamma_d)\frac{k\sqrt{k}}{m} \quad , \qquad \text{(A.4)}
\end{aligned}$$

where the last but one inequality results from Lemma D.6.

A.3. Proof of Theorem 3.2

The idea is to plug the upper bounds previously derived for the L1O estimator, namely Ineq. (2.5) and (2.6) from Theorem 2.2, in the inequalities proved for the moments of the LpO estimator in Theorem 2.2.

Proof of Ineq. (3.3)**,** (3.4)**, and** (3.5)**:** These inequalities straightforwardly result from the combination of Theorem 2.2 and Ineq. (2.5) and (2.6) from Theorem 3.1.

Proof of Ineq. (3.6)**:** It results from the upper bounds proved in Theorem 3.1 and plugged in Ineq. (2.7) (derived from Rosenthal's inequality with optimized constant γ, namely Proposition D.3).

Then it comes

$$\begin{aligned}
&\mathbb{E}\left[\left|\widehat{R}_{p,n} - \mathbb{E}\left[\widehat{R}_{p,n}\right]\right|^q\right] \leq \left(2\sqrt{2e}\right)^q \times \\
&\max\left\{(\sqrt{q})^q\left(\sqrt{\left\lfloor\frac{n}{n-p+1}\right\rfloor^{-1} 2C_1\sqrt{k}\left(\frac{\sqrt{k}}{\sqrt{n-p+1}}\right)^2}\right)^q, q^q\left\lfloor\frac{n}{n-p+1}\right\rfloor^{-q+1}(2C_2\sqrt{q})^q\left(\frac{k}{\sqrt{n-p+1}}\right)^q\right\} \\
&= \left(2\sqrt{2e}\right)^q \times \\
&\max\left\{(\sqrt{q})^q\left(\sqrt{2C_1\sqrt{k}}\sqrt{\frac{k}{(n-p+1)\left\lfloor\frac{n}{n-p+1}\right\rfloor}}\right)^q, \left(q^{3/2}\right)^q\left\lfloor\frac{n}{n-p+1}\right\rfloor\left(2C_2\frac{k}{\left\lfloor\frac{n}{n-p+1}\right\rfloor\sqrt{n-p+1}}\right)^q\right\} \\
&\leq \left\lfloor\frac{n}{n-p+1}\right\rfloor\max\left\{\left(\lambda_1 q^{1/2}\right)^q, \left(\lambda_2 q^{3/2}\right)^q\right\},
\end{aligned}$$

with

$$\lambda_1 = 2\sqrt{2e}\sqrt{2C_1\sqrt{k}}\sqrt{\frac{k}{(n-p+1)\left\lfloor\frac{n}{n-p+1}\right\rfloor}}, \quad \lambda_2 = 2\sqrt{2e}2C_2\frac{k}{\left\lfloor\frac{n}{n-p+1}\right\rfloor\sqrt{n-p+1}} \quad .$$

Finally introducing $\Gamma = 2\sqrt{2e}\max\left(2C_2, \sqrt{2C1}\right)$ provides the result.

Appendix B. Proofs of exponential concentration inequalities

B.1. Proof of Proposition 4.1

The proof relies on two successive ingredients: McDiarmid's inequality (Theorem D.3), and Stone's lemma (Lemma D.5).

First with $\mathcal{D}_n = \mathcal{D}$ and $\mathcal{D}_j = (Z_1, \ldots, Z_{j-1}, Z'_j, Z_{j+1}, \ldots, Z_n)$, let us start by upper bounding $\left| \widehat{R}_p(\mathcal{D}_n) - \widehat{R}_p(\mathcal{D}_j) \right|$ for every $1 \leq j \leq n$.

Using Eq. (2.2), one has

$$\begin{aligned}
&\left| \widehat{R}_p(\mathcal{D}) - \widehat{R}_p(\mathcal{D}_j) \right| \\
&\leq \frac{1}{p} \sum_{i=1}^{n} \binom{n}{p}^{-1} \sum_{e} \left| \mathbb{1}_{\left\{ \mathcal{A}_k^{\mathcal{D}^e}(X_i) \neq Y_i \right\}} - \mathbb{1}_{\left\{ \mathcal{A}_k^{\mathcal{D}^e_j}(X_i) \neq Y_i \right\}} \right| \mathbb{1}_{\{i \notin e\}} \\
&\leq \frac{1}{p} \sum_{i=1}^{n} \binom{n}{p}^{-1} \sum_{e} \mathbb{1}_{\left\{ \mathcal{A}_k^{\mathcal{D}^e}(X_i) \neq \mathcal{A}_k^{\mathcal{D}^e_j}(X_i) \right\}} \mathbb{1}_{\{i \notin e\}} \\
&\leq \frac{1}{p} \sum_{i \neq j}^{n} \binom{n}{p}^{-1} \sum_{e} \left[\mathbb{1}_{\left\{ j \in V_k^{\mathcal{D}^e}(X_i) \right\}} + \mathbb{1}_{\left\{ j \in V_k^{\mathcal{D}^e_j}(X_i) \right\}} \right] \mathbb{1}_{\{i \notin e\}} + \frac{1}{p} \binom{n}{p}^{-1} \sum_{e} \mathbb{1}_{\{j \notin e\}},
\end{aligned}$$

where $\mathcal{D}^e_j$ denotes the set of random variables among $\mathcal{D}_j$ having indices in e, and $V_k^{\mathcal{D}^e}(X_i)$ (resp. $V_k^{\mathcal{D}^e_j}(X_i)$) denotes the set of indices of the k nearest neighbors of X_i among $\mathcal{D}^e$ (resp. $\mathcal{D}^e_j$).

Second, let us now introduce

$$B_j^{\mathcal{E}_{n-p}} = \underset{e \in \mathcal{E}_{n-p}}{\cup} \left\{ 1 \leq i \leq n,\ i \notin e \cup \{j\}\,,\ V_k^{\mathcal{D}^e_j}(X_i) \ni j \text{ or } V_k^{\mathcal{D}^e}(X_i) \ni j \right\}.$$

Then Lemma D.5 implies $\mathrm{Card}(B_j^{\mathcal{E}_{n-p}}) \leq 2(k+p-1)\gamma_d$, hence

$$\left| \widehat{R}_p(\mathcal{D}_n) - \widehat{R}_p(\mathcal{D}_j) \right| \leq \frac{1}{p} \sum_{i \in B_j^{\mathcal{E}_{n-p}}} \binom{n}{p}^{-1} \sum_{e} 2 \cdot \mathbb{1}_{\{i \notin e\}} + \frac{1}{n} \leq \frac{4(k+p-1)\gamma_d}{n} + \frac{1}{n} \ \cdot$$

The conclusion results from McDiarmid's inequality (Section D.1.5).

B.2. Proof of Theorem 4.1

In this proof, we use the same notation as in that of Proposition 4.1.

The goal of the proof is to provide a refined version of previous Proposition 4.1 by taking into account the status of each X_j as one of the k nearest neighbors of a given X_i (or not).

To do so, our strategy is to prove a sub-Gaussian concentration inequality by use of Lemma D.2, which requires the control of the even moments of the LpO estimator $\widehat{R}_p$.

Such upper bounds are derived

- First, by using Ineq. (D.4) (generalized Efron-Stein inequality), which amounts to control the q-th moments of the differences

$$\widehat{R}_p(\mathcal{D}) - \widehat{R}_p(\mathcal{D}_j)\,.$$

- Second, by precisely evaluating the contribution of each neighbor X_i of a given X_j, that is by computing quantities such as $\mathbb{P}_e\left[j \in e,\ i \in \bar{e},\ j \in V_k^{\mathcal{D}^e}(X_i)\right]$, where $\mathbb{P}_e\left[\,\cdot\,\right]$ denotes the probability measure with respect to the uniform random variable e over $\mathcal{E}_{n-p}$, and $V_k^{\mathcal{D}^e}(X_i)$ denotes the indices of the k nearest neighbors of X_i among $X^e = \{X_\ell, \ell \in e\}$.

B.2.1. Upper bounding $\widehat{R}_p(\mathcal{D}) - \widehat{R}_p\left(\mathcal{D}_j\right)$

For every $1 \leq j \leq n$, one gets

$$\begin{aligned}\widehat{R}_p(\mathcal{D}) - \widehat{R}_p\left(\mathcal{D}_j\right) = \binom{n}{p}^{-1} \sum_e \Bigg\{ & \mathbb{1}_{\{j\in\bar{e}\}} \frac{1}{p}\left(\mathbb{1}_{\left\{\mathcal{A}_k^{\mathcal{D}^e}(X_j)\neq Y_j\right\}} - \mathbb{1}_{\left\{\mathcal{A}_k^{\mathcal{D}^e}(X'_j)\neq Y'_j\right\}} \right) \\ & + \mathbb{1}_{\{j\in e\}} \frac{1}{p} \sum_{i\in\bar{e}} \left(\mathbb{1}_{\left\{\mathcal{A}_k^{\mathcal{D}^e}(X_i)\neq Y_i\right\}} - \mathbb{1}_{\left\{\mathcal{A}_k^{\mathcal{D}^e_j}(X_i)\neq Y_i\right\}} \right) \Bigg\} .\end{aligned}$$

Absolute values and Jensen's inequality then provide

$$\begin{aligned}\left|\widehat{R}_p(\mathcal{D}) - \widehat{R}_p\left(\mathcal{D}_j\right)\right| & \leq \binom{n}{p}^{-1} \sum_e \left\{ \mathbb{1}_{\{j\in\bar{e}\}} \frac{1}{p} + \mathbb{1}_{\{j\in e\}} \frac{1}{p} \sum_{i\in\bar{e}} \mathbb{1}_{\left\{\mathcal{A}_k^{\mathcal{D}^e}(X_i)\neq \mathcal{A}_k^{\mathcal{D}^e_j}(X_i)\right\}} \right\} \\ & \leq \frac{1}{n} + \binom{n}{p}^{-1} \sum_e \mathbb{1}_{\{j\in e\}} \frac{1}{p} \sum_{i\in\bar{e}} \mathbb{1}_{\left\{\mathcal{A}_k^{\mathcal{D}^e}(X_i)\neq \mathcal{A}_k^{\mathcal{D}^e_j}(X_i)\right\}} \\ & = \frac{1}{n} + \frac{1}{p} \sum_{i=1}^n \mathbb{P}_e\left[j\in e,\ i\in\bar{e},\ \mathcal{A}_k^{\mathcal{D}^e}(X_i) \neq \mathcal{A}_k^{\mathcal{D}^e_j}(X_i) \right] .\end{aligned}$$

where the notation $\mathbb{P}_e$ means the integration is carried out with respect to the random variable $e \in \mathcal{E}_{n-p}$, which follows a discrete uniform distribution over the set $\mathcal{E}_{n-p}$ of all $n-p$ distinct indices among $\{1, \ldots, n\}$.

Let us further notice that $\left\{\mathcal{A}_k^{\mathcal{D}^e}(X_i) \neq \mathcal{A}_k^{\mathcal{D}^e_j}(X_i)\right\} \subset \left\{ j \in V_k^{\mathcal{D}^e}(X_i) \cup V_k^{\mathcal{D}^e_j}(X_i) \right\}$, where $V_k^{\mathcal{D}^e_j}(X_i)$ denotes the set of indices of the k nearest neighbors of X_i among $\mathcal{D}^e_j$ with the notation of the proof of Proposition 4.1. Then it results

$$\begin{aligned}& \sum_{i=1}^n \mathbb{P}_e\left[j\in e,\ i\in\bar{e},\ \mathcal{A}_k^{\mathcal{D}^e}(X_i) \neq \mathcal{A}_k^{\mathcal{D}^e_j}(X_i) \right] \\ & \leq \sum_{i=1}^n \mathbb{P}_e\left[j\in e,\ i\in\bar{e},\ j \in V_k^{\mathcal{D}^e}(X_i) \cup V_k^{\mathcal{D}^e_j}(X_i) \right] \\ & \leq \sum_{i=1}^n \left(\mathbb{P}_e\left[j\in e,\ i\in\bar{e},\ j \in V_k^{\mathcal{D}^e}(X_i) \right] + \mathbb{P}_e\left[j\in e,\ i\in\bar{e},\ j \in V_k^{\mathcal{D}^e}(X_i) \cup V_k^{\mathcal{D}^e_j}(X_i) \right] \right) \\ & \leq 2 \sum_{i=1}^n \mathbb{P}_e\left[j\in e,\ i\in\bar{e},\ j \in V_k^{\mathcal{D}^e}(X_i) \right] ,\end{aligned}$$

which leads to

$$\left|\widehat{R}_p(\mathcal{D}) - \widehat{R}_p\left(\mathcal{D}_j\right)\right| \leq \frac{1}{n} + \frac{2}{p} \sum_{i=1}^n \mathbb{P}_e\left[j\in e,\ i\in\bar{e},\ j \in V_k^{\mathcal{D}^e}(X_i) \right] .$$

Summing over $1 \leq j \leq n$ the square of the above quantity, it results

$$\begin{aligned}\sum_{j=1}^{n}\left(\widehat{R}_p(\mathcal{D})-\widehat{R}_p\left(\mathcal{D}_j\right)\right)^2 &\leq \sum_{j=1}^{n}\left\{\frac{1}{n}+\frac{2}{p}\sum_{i=1}^{n}\mathbb{P}_e\left[\, j\in e,\ i\in\bar{e},\ j\in V_k^{\mathcal{D}^e}(X_i)\,\right]\right\}^2\\ &\leq 2\sum_{j=1}^{n}\frac{1}{n^2}+2\left\{\frac{2}{p}\sum_{i=1}^{n}\mathbb{P}_e\left[\, j\in e,\ i\in\bar{e},\ j\in V_k^{\mathcal{D}^e}(X_i)\,\right]\right\}^2\\ &\leq \frac{2}{n}+8\sum_{j=1}^{n}\left\{\frac{1}{p}\sum_{i=1}^{n}\mathbb{P}_e\left[\, j\in e,\ i\in\bar{e},\ j\in V_k^{\mathcal{D}^e}(X_i)\,\right]\right\}^2 .\end{aligned}$$

B.2.2. Evaluating the influence of each neighbor

Further using that

$$\begin{aligned}&\sum_{j=1}^{n}\left(\frac{1}{p}\sum_{i=1}^{n}\mathbb{P}_e\left[\, j\in e,\ i\in\bar{e},\ j\in V_k^{\mathcal{D}^e}(X_i)\,\right]\right)^2\\ &=\sum_{j=1}^{n}\frac{1}{p^2}\sum_{i=1}^{n}\left(\mathbb{P}_e\left[\, j\in e,\ i\in\bar{e},\ j\in V_k^{\mathcal{D}^e}(X_i)\,\right]\right)^2+\\ &\qquad\sum_{j=1}^{n}\frac{1}{p^2}\sum_{1\leq i\neq \ell\leq n}\mathbb{P}_e\left[\, j\in e,\ i\in\bar{e},\ j\in V_k^{\mathcal{D}^e}(X_i)\,\right]\mathbb{P}_e\left[\, j\in e,\ i\in\bar{e},\ j\in V_k^{\mathcal{D}^e}(X_\ell)\,\right]\\ &=\quad T1 \qquad + \qquad T2\ ,\end{aligned}$$

let us now successively deal with each of these two terms.

Upper bound on $T1$

First, we start by partitioning the sum over j depending on the rank of X_j as a neighbor of X_i in the whole sample $(X_1,\ldots,X_n)$. It comes

$$\begin{aligned}&=\sum_{j=1}^{n}\sum_{i=1}^{n}\left\{\mathbb{P}_e\left[\, j\in e,\ i\in\bar{e},\ j\in V_k^{\mathcal{D}^e}(X_i)\right]\right\}^2\\ &=\sum_{i=1}^{n}\left(\sum_{j\in V_k(X_i)}\left\{\mathbb{P}_e\left[\, j\in e,\ i\in\bar{e},\ j\in V_k^{\mathcal{D}^e}(X_i)\right]\right\}^2+\sum_{j\in V_{k+p}(X_i)\setminus V_k(X_i)}\left\{\mathbb{P}_e\left[\, j\in e,\ i\in\bar{e},\ j\in V_k^{\mathcal{D}^e}(X_i)\right]\right\}^2\right).\end{aligned}$$

Then Lemma D.4 leads to

$$\begin{aligned}&\sum_{j\in V_k(X_i)}\left\{\mathbb{P}_e\left[\, j\in e,\ i\in\bar{e},\ j\in V_k^{\mathcal{D}^e}(X_i)\,\right]\right\}^2+\sum_{j\in V_{k+p}(X_i)\setminus V_k(X_i)}\left\{\mathbb{P}_e\left[\, j\in e,\ i\in\bar{e},\ j\in V_k^{\mathcal{D}^e}(X_i)\,\right]\right\}^2\\ &\leq\sum_{j\in V_k(X_i)}\left(\frac{p}{n}\frac{n-p}{n-1}\right)^2+\sum_{j\in V_{k+p}(X_i)\setminus V_k(X_i)}\mathbb{P}_e\left[\, j\in e,\ i\in\bar{e},\ j\in V_k^{\mathcal{D}^e}(X_i)\,\right]\frac{p}{n}\frac{n-p}{n-1}\\ &=k\left(\frac{p}{n}\frac{n-p}{n-1}\right)^2+\frac{kp}{n}\frac{p-1}{n-1}\frac{p}{n}\frac{n-p}{n-1}=k\left(\frac{p}{n}\right)^2\frac{n-p}{n-1}\ ,\end{aligned}$$

where the upper bound results from $\sum_j a_j^2 \leq (\max_j a_j) \sum_j a_j$, for $a_j \geq 0$. It results

$$T1 = \frac{1}{p^2} \sum_{j=1}^{n} \sum_{i=1}^{n} \left\{ \mathbb{P}_e \left[j \in e,\ i \in \bar{e},\ j \in V_k^{\mathcal{D}^e}(X_i) \right] \right\}^2 \leq \frac{1}{p^2} n \left[k \left(\frac{p}{n} \right)^2 \frac{n-p}{n-1} \right] = \frac{k}{n} \frac{n-p}{n-1} \ \cdot$$

Upper bound on $T2$

Let us now apply the same idea to the second sum, partitioning the sum over j depending on the rank of j as a neighbor of ℓ in the whole sample. Then,

$$\begin{aligned} T2 &= \frac{1}{p^2} \sum_{j=1}^{n} \sum_{1 \leq i \neq \ell \leq n} \mathbb{P}_e \left[j \in e,\ i \in \bar{e},\ j \in V_k^{\mathcal{D}^e}(X_i) \right] \mathbb{P}_e \left[j \in e,\ \ell \in \bar{e},\ j \in V_k^{\mathcal{D}^e}(X_\ell) \right] \\ &\leq \frac{1}{p^2} \sum_{i=1}^{n} \sum_{\ell \neq i} \sum_{j \in V_k(X_\ell)} \mathbb{P}_e \left[j \in e,\ i \in \bar{e},\ j \in V_k^{\mathcal{D}^e}(X_i) \right] \frac{p}{n} \frac{n-p}{n-1} \\ &\quad + \frac{1}{p^2} \sum_{i=1}^{n} \sum_{\ell \neq i} \sum_{j \in V_{k+p}(X_\ell) \setminus V_k(X_\ell} \mathbb{P}_e \left[j \in e,\ i \in \bar{e},\ j \in V_k^{\mathcal{D}^e}(X_i) \right] \frac{kp}{n} \frac{p-1}{n-1} \ \cdot \end{aligned}$$

We then apply Stone's lemma (Lemma D.5) to get

$$\begin{aligned} &T2 \\ &= \frac{1}{p^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{P}_e \left[j \in e,\ i \in \bar{e},\ j \in V_k^{\mathcal{D}^e}(X_i) \right] \left(\sum_{\ell \neq i} \mathbb{1}_{j \in V_k(X_\ell)} \frac{p}{n} \frac{n-p}{n-1} + \sum_{\ell \neq i} \mathbb{1}_{j \in V_{k+p}(X_\ell) \setminus V_k(X_\ell} \frac{kp}{n} \frac{p-1}{n-1} \right) \\ &\leq \frac{1}{p^2} \sum_{i=1}^{n} \frac{kp}{n} \left(k \gamma_d \frac{p}{n} \frac{n-p}{n-1} + (k+p) \gamma_d \frac{kp}{n} \frac{p-1}{n-1} \right) = \gamma_d \frac{k^2}{n} \left(\frac{n-p}{n-1} + (k+p) \frac{p-1}{n-1} \right) \\ &= \gamma_d \frac{k^2}{n} \left(1 + (k+p-1) \frac{p-1}{n-1} \right). \end{aligned}$$

Gathering the upper bounds

The two previous bounds provide

$$\begin{aligned} \sum_{j=1}^{n} \left\{ \frac{1}{p} \sum_{i=1}^{n} \mathbb{P}_e \left[j \in e,\ i \in \bar{e},\ j \in V_k^{\mathcal{D}^e}(X_i) \right] \right\}^2 &= T1 + T2 \\ &\leq \frac{k}{n} \frac{n-p}{n-1} + \gamma_d \frac{k^2}{n} \left(1 + (k+p-1) \frac{p-1}{n-1} \right), \end{aligned}$$

which enables to conclude

$$\begin{aligned} &\sum_{j=1}^{n} \left(\widehat{R}_p(\mathcal{D}) - \widehat{R}_p\left(\mathcal{D}_j\right) \right)^2 \\ &\leq \frac{2}{n} \left(1 + 4k + 4k^2 \gamma_d \left[1 + (k+p) \frac{p-1}{n-1} \right] \right) \leq \frac{8k^2(1+\gamma_d)}{n} \left[1 + (k+p) \frac{p-1}{n-1} \right]. \end{aligned}$$

B.2.3. Generalized Efron-Stein inequality

Then (D.4) provides for every $q \geq 1$

$$\left\| \widehat{R}_{p,n} - \mathbb{E}\left[\widehat{R}_{p,n} \right] \right\|_{2q} \leq 4\sqrt{\kappa q} \sqrt{\frac{8(1+\gamma_d)k^2}{n} \left[1 + (k+p)\frac{p-1}{n-1} \right]}.$$

Hence combined with $q! \geq q^q e^{-q}\sqrt{2\pi q}$, it comes

$$\begin{aligned} \mathbb{E}\left[\left(\widehat{R}_{p,n} - \mathbb{E}\left[\widehat{R}_{p,n} \right] \right)^{2q} \right] &\leq (16\kappa q)^q \left(\frac{8(1+\gamma_d)k^2}{n} \left[1 + (k+p)\frac{p-1}{n-1} \right] \right)^q \\ &\leq q! \left(16e\kappa \frac{8(1+\gamma_d)k^2}{n} \left[1 + (k+p)\frac{p-1}{n-1} \right] \right)^q . \end{aligned}$$

The conclusion follows from Lemma D.2 with $C = 16e\kappa\frac{8(1+\gamma_d)k^2}{n}\left[1 + (k+p)\frac{p-1}{n-1} \right]$. Then for every $t > 0$,

$$\mathbb{P}\left(\widehat{R}_{p,n} - \mathbb{E}\left(\widehat{R}_{p,n} \right) > t \right) \vee \mathbb{P}\left(\mathbb{E}\left(\widehat{R}_{p,n} \right) - \widehat{R}_{p,n} > t \right) \leq \exp\left(-\frac{nt^2}{1024e\kappa k^2(1+\gamma_d)\left[1 + (k+p)\frac{p-1}{n-1} \right]} \right).$$

B.3. Proof of Theorem 4.2 and Proposition 4.2

B.3.1. Proof of Theorem 4.2

If $p < n/2 + 1$:
In what follows, we exploit a characterization of sub-Gaussian random variables by their $2q$-th moments (Lemma D.2).

From (3.3) and (3.4) applied with $2q$, and further introducing a constant $\Delta = 4\sqrt{e}\max\left(\sqrt{C_1/2}, C_2 \right) > 0$, it comes for every $q \geq 1$

$$\mathbb{E}\left[\left| \widehat{R}_{p,n} - \mathbb{E}\left[\widehat{R}_{p,n} \right] \right|^{2q} \right] \leq \left(\frac{\Delta^2}{16e}\frac{k^2}{n-p+1} \right)^q (2q)^q \leq \left(\frac{\Delta^2}{8}\frac{k^2}{n-p+1} \right)^q q! \ , \qquad \text{(B.1)}$$

with $q^q \leq q!e^q/\sqrt{2\pi q}$. Then Lemma D.2 provides for every $t > 0$

$$\mathbb{P}\left(\widehat{R}_{p,n} - \mathbb{E}\left[\widehat{R}_{p,n} \right] > t \right) \vee \mathbb{P}\left(\mathbb{E}\left[\widehat{R}_{p,n} \right] - \widehat{R}_{p,n} > t \right) \leq \exp\left(-(n-p+1)\frac{t^2}{\Delta^2 k^2} \right).$$

If $p \geq n/2 + 1$:
This part of the proof relies on Proposition D.1 which provides an exponential concentration inequality from upper bounds on the moments of a random variable.

Let us now use (3.3) and (3.6) combined with (D.1), where $C = \left\lfloor \frac{n}{n-p+1} \right\rfloor$, $q_0 = 2$, and $\min_j \alpha_j = 1/2$. This provides for every $t > 0$

$$\mathbb{P}\left[\left| \widehat{R}_{p,n} - \mathbb{E}\left[\widehat{R}_{p,n} \right] \right| > t \right] \leq \left\lfloor \frac{n}{n-p+1} \right\rfloor e \times$$

$$\exp\left[-\frac{1}{2e}\min\left\{ (n-p+1)\left\lfloor \frac{n}{n-p+1} \right\rfloor \frac{t^2}{4\Gamma^2 k\sqrt{k}},\ \left((n-p+1)\left\lfloor \frac{n}{n-p+1} \right\rfloor^2 \frac{t^2}{4\Gamma^2 k^2} \right)^{1/3} \right\} \right],$$

where Γ arises from Eq. (3.6).

B.3.2. Proof of Proposition 4.2

As in the previous proof, the derivation of the deviation terms results from Proposition D.1.

With the same notation and reasoning as in the previous proof, let us combine (3.3) and (3.6). From (D.2) of Proposition D.1 where $C = \left\lfloor \frac{n}{n-p+1} \right\rfloor$, $q_0 = 2$, and $\min_j \alpha_j = 1/2$, it results for every $t > 0$

$$\mathbb{P}\left[\left| \widehat{R}_{p,n} - \mathbb{E}\left[\widehat{R}_{p,n} \right] \right| > \Gamma \sqrt{\frac{2e}{(n-p+1)}} \left(\sqrt{\frac{k^{3/2}}{\left\lfloor \frac{n}{n-p+1} \right\rfloor} t} + 2e \frac{k}{\left\lfloor \frac{n}{n-p+1} \right\rfloor} t^{3/2} \right) \right] \leq \left\lfloor \frac{n}{n-p+1} \right\rfloor e \cdot e^{-t},$$

where $\Gamma > 0$ is given by Eq. (3.6).

Appendix C. Proofs of deviation upper bounds

C.1. Proof of Ineq. (5.3) in Theorem 5.1

The proof follows the same strategy as that of Theorem 2.1 in Rogers and Wagner (1978).

Along the proof, we will repeatedly use some notation that we briefly introduce here. First, let us define $Z_0 = (X_0, Y_0)$ and $Z_{n+1} = (X_{n+1}, Y_{n+1})$ that are independent copies of Z_1. Second to ease the reading of the proof, we also use several shortcuts: $\widehat{f}_k(X_0) = \mathcal{A}_k^{\mathcal{D}_n}(X_0)$, and $\widehat{f}_k^e(X_0) = \mathcal{A}_k^{\mathcal{D}^e}(X_0)$ for every set of indices $e \in \mathcal{E}_{n-p}$ (with cardinality $n-p$).

Finally along the proof, $e, e' \in \mathcal{E}_{n-p}$ denote two *random variables* which are sets of distinct indices *with discrete uniform distribution over* $\mathcal{E}_{n-p}$. The notation $\mathbb{P}_e$ (resp. $\mathbb{P}_{e,e'}$) means the integration is made with respect to the sample $\mathcal{D}$ and also the random variable e (resp. $\mathcal{D}$ and also the random variables e, e'). $\mathbb{E}_e[\cdot]$ and $\mathbb{E}_{e,e'}[\cdot]$ are teh corresponding expectations. Note that the sample $\mathcal{D}$ and the random variables e, e' are independent from each other, so that computing for instance $\mathbb{P}_e(i \notin e)$ amounts to integrating with respect to the random variable e only.

C.1.1. Main part of the proof

With the notation $L_n = L(\mathcal{A}_k^{\mathcal{D}_n})$, let us start from

$$\mathbb{E}\left[(\widehat{R}_{p,n} - L_n)^2\right] = \mathbb{E}\left[\widehat{R}_p^2(\mathcal{A}_k^{\mathcal{D}_n})\right] + \mathbb{E}\left[L_n^2\right] - 2\mathbb{E}\left[\widehat{R}_{p,n} L_n\right],$$

let us notice that

$$\mathbb{E}\left[L_n^2\right] = \mathbb{P}\left(\widehat{f}_k(X_0) \neq Y_0, \widehat{f}_k(X_{n+1}) \neq Y_{n+1}\right),$$

and

$$\mathbb{E}\left[\widehat{R}_{p,n} L_n\right] = \mathbb{P}_e\left(\widehat{f}_k(X_0) \neq Y_0, \widehat{f}_k^e(X_i) \neq Y_i |\ i \notin e\right)\mathbb{P}_e\left(i \notin e\right).$$

It immediately comes

$$\mathbb{E}\left[(\widehat{R}_{p,n} - L_n)^2\right]$$
$$= \mathbb{E}\left[\widehat{R}_p^2(\mathcal{A}_k^{\mathcal{D}_n})\right] - \mathbb{P}_e\left(\widehat{f}_k(X_0) \neq Y_0, \widehat{f}_k^e(X_i) \neq Y_i \mid\ i \notin e\right)\mathbb{P}_e\left(i \notin e\right) \tag{C.1}$$
$$+ \left[\mathbb{P}\left(\widehat{f}_k(X_0) \neq Y_0, \widehat{f}_k(X_{n+1}) \neq Y_{n+1}\right) - \mathbb{P}_e\left(\widehat{f}_k(X_0) \neq Y_0, \widehat{f}_k^e(X_i) \neq Y_i |\ i \notin e\right)\mathbb{P}_e\left(i \notin e\right)\right]. \tag{C.2}$$

The proof then consists in successively upper bounding the two terms (C.1) and (C.2) of the last equality.

Upper bound of (C.1)

First, we have

$$\begin{aligned} p^2\mathbb{E}\left[\widehat{R}_p^2(\mathcal{A}_k^{\mathcal{D}_n})\right] &= \sum_{i,j} \mathbb{E}_{e,e'}\left[\mathbb{1}_{\left\{\widehat{f}_k^e(X_i) \neq Y_i\right\}}\mathbb{1}_{\{i\notin e\}}\mathbb{1}_{\left\{\widehat{f}_k^{e'}(X_j) \neq Y_j\right\}}\mathbb{1}_{\{j\notin e'\}}\right] \\ &= \sum_{i} \mathbb{E}_{e,e'}\left[\mathbb{1}_{\left\{\widehat{f}_k^e(X_i) \neq Y_i\right\}}\mathbb{1}_{\{i\notin e\}}\mathbb{1}_{\left\{\widehat{f}_k^{e'}(X_i) \neq Y_i\right\}}\mathbb{1}_{\{i\notin e'\}}\right] \\ &\quad + \sum_{i \neq j} \mathbb{E}_{e,e'}\left[\mathbb{1}_{\left\{\widehat{f}_k^e(X_i) \neq Y_i\right\}}\mathbb{1}_{\{i\notin e\}}\mathbb{1}_{\left\{\widehat{f}_k^{e'}(X_j) \neq Y_j\right\}}\mathbb{1}_{\{j\notin e'\}}\right]. \end{aligned}$$

Let us now introduce the five following events where we emphasize e and e' are random variables with the discrete uniform distribution over $\mathcal{E}_{n-p}$:

$$\begin{aligned}
S_i^0 &= \{i \notin e,\ i \notin e'\}, \\
S_{i,j}^1 &= \{i \notin e,\ j \notin e',\ i \notin e',\ j \notin e\}, \qquad S_{i,j}^2 = \{i \notin e,\ j \notin e',\ i \notin e',\ j \in e\}, \\
S_{i,j}^3 &= \{i \notin e,\ j \notin e',\ i \in e',\ j \notin e\}, \qquad S_{i,j}^4 = \{i \notin e,\ j \notin e',\ i \in e',\ j \in e\}.
\end{aligned}$$

Then,

$$\begin{aligned}
p^2 \mathbb{E}\left[\widehat{R}_p^2(\mathcal{A}_k^{\mathcal{D}_n})\right] &= \sum_i \mathbb{P}_{e,e'}\left(\widehat{f}_k^e(X_i) \neq Y_i,\ \widehat{f}_k^{e'}(X_i) \neq Y_i | S_i^0\right) \mathbb{P}_{e,e'}\left(S_i^0\right) \\
&\quad + \sum_{i \neq j} \sum_{\ell=1}^4 \mathbb{P}_{e,e'}\left(\widehat{f}_k^e(X_i) \neq Y_i,\ \widehat{f}_k^{e'}(X_i) \neq Y_i | S_{i,j}^\ell\right) \mathbb{P}_{e,e'}\left(S_{i,j}^\ell\right) \\
&= n \mathbb{P}_{e,e'}\left(\widehat{f}_k^e(X_1) \neq Y_1,\ \widehat{f}_k^{e'}(X_1) \neq Y_1 | S_1^0\right) \mathbb{P}_{e,e'}\left(S_1^0\right) \\
&\quad + n(n-1) \sum_{\ell=1}^4 \mathbb{P}_{e,e'}\left(\widehat{f}_k^e(X_1) \neq Y_1,\ \widehat{f}_k^{e'}(X_2) \neq Y_2 \mid S_{1,2}^\ell\right) \mathbb{P}_{e,e'}\left(S_{1,2}^\ell\right).
\end{aligned}$$

Furthermore since

$$\frac{1}{p^2}\left[n \mathbb{P}_{e,e'}\left(S_1^0\right) + n(n-1) \sum_{\ell=1}^4 \mathbb{P}_{e,e'}\left(S_{1,2}^\ell\right)\right] = \frac{1}{p^2} \sum_{i,j} \mathbb{P}_{e,e'}\left(i \notin e,\ j \notin e'\right) = 1,$$

it comes

$$\mathbb{E}\left[\widehat{R}_p^2(\mathcal{A}_k^{\mathcal{D}_n})\right] - \mathbb{P}_{e,e'}\left(\widehat{f}_k(X_0) \neq Y_0, \widehat{f}_k^e(X_1) \neq Y_1\right) \quad = \quad \frac{n}{p^2} A + \frac{n(n-1)}{p^2} B, \qquad \text{(C.3)}$$

where

$$\begin{aligned}
A = &\left[\mathbb{P}_{e,e'}\left(\widehat{f}_k^e(X_1) \neq Y_1,\ \widehat{f}_k^{e'}(X_1) \neq Y_1 \mid S_1^0\right) - \mathbb{P}_{e,e'}\left(\widehat{f}_k(X_0) \neq Y_0, \widehat{f}_k^e(X_1) \neq Y_1 \mid S_1^0\right)\right] \\
&\times \mathbb{P}_{e,e'}\left(S_1^0\right),
\end{aligned}$$

$$\begin{aligned}
\text{and} \quad B = &\sum_{\ell=1}^4 \left[\mathbb{P}_{e,e'}\left(\widehat{f}_k^e(X_1) \neq Y_1,\ \widehat{f}_k^{e'}(X_2) \neq Y_2 \mid S_{1,2}^\ell\right) - \mathbb{P}_{e,e'}\left(\widehat{f}_k(X_0) \neq Y_0, \widehat{f}_k^e(X_1) \neq Y_1 \mid S_{1,2}^\ell\right)\right] \\
&\times \mathbb{P}_{e,e'}\left(S_{1,2}^\ell\right).
\end{aligned}$$

• **Upper bound for A:**
To upper bound A, simply notice that:

$$A \quad \leq \quad \mathbb{P}_{e,e'}\left(S_i^0\right) \leq \mathbb{P}_{e,e'}\left(i \notin e,\ i \notin e'\right) \leq \left(\frac{p}{n}\right)^2.$$

• **Upper bound for B:**
To obtain an upper bound for B, one needs to upper bound

$$\mathbb{P}_{e,e'}\left(\widehat{f}_k^e(X_1) \neq Y_1,\ \widehat{f}_k^{e'}(X_2) \neq Y_2 \mid S_{1,2}^\ell\right) - \mathbb{P}_{e,e'}\left(\widehat{f}_k(X_0) \neq Y_0, \widehat{f}_k^e(X_1) \neq Y_1 \mid S_{1,2}^\ell\right), \text{ (C.4)}$$

which depends on ℓ, i.e. on the fact that index 2 belongs or not to the training set e.

- If $2 \notin e$ (i.e. $\ell = 1$ or 3): Then, Lemma C.2 proves

$$(\text{C.4}) \leq \frac{4p\sqrt{k}}{\sqrt{2\pi}n} \ \cdot$$

- If $2 \in e$ (i.e. $\ell = 2$ or 4): Then, Lemma C.3 settles

$$(\text{C.4}) \leq \frac{8\sqrt{k}}{\sqrt{2\pi}(n-p)} + \frac{4p\sqrt{k}}{\sqrt{2\pi}n} \ \cdot$$

Combining the previous bounds and Lemma C.1 leads to

$$\begin{aligned}
B &\leq \left(\frac{4p\sqrt{k}}{\sqrt{2\pi}n}\right)\left[\mathbb{P}_{e,e'}\left(S^1_{1,2}\right) + \mathbb{P}_{e,e'}\left(S^3_{1,2}\right)\right] + \left(\frac{8\sqrt{k}}{\sqrt{2\pi}(n-p)} + \frac{4p\sqrt{k}}{\sqrt{2\pi}n}\right)\left[\mathbb{P}_{e,e'}\left(S^2_{1,2}\right) + \mathbb{P}_{e,e'}\left(S^4_{1,2}\right)\right] \\
&\leq \frac{2\sqrt{2}}{\sqrt{\pi}}\sqrt{k}\left[\frac{p}{n}\left[\mathbb{P}_{e,e'}\left(S^1_{1,2}\right) + \mathbb{P}_{e,e'}\left(S^3_{1,2}\right)\right] + \left(\frac{2}{n-p} + \frac{p}{n}\right)\left[\mathbb{P}_{e,e'}\left(S^2_{1,2}\right) + \mathbb{P}_{e,e'}\left(S^4_{1,2}\right)\right]\right] \\
&\leq \frac{2\sqrt{2}}{\sqrt{\pi}}\sqrt{k}\left[\frac{p}{n}\mathbb{P}_{e,e'}\left(i \notin e,\ j \notin e'\right) + \frac{2}{n-p}\left(\mathbb{P}_{e,e'}\left(S^2_{1,2}\right) + \mathbb{P}_{e,e'}\left(S^4_{1,2}\right)\right)\right] \\
&\leq \frac{2\sqrt{2}}{\sqrt{\pi}}\sqrt{k}\left[\frac{p}{n}\left(\frac{p}{n}\right)^2 + \frac{2}{n-p}\left(\frac{(n-p)p^2(p-1)}{n^2(n-1)^2} + \frac{(n-p)^2p^2}{n^2(n-1)^2}\right)\right] \\
&\leq \frac{2\sqrt{2}}{\sqrt{\pi}}\sqrt{k}\left(\frac{p}{n}\right)^2\left[\frac{p}{n} + \frac{2}{n-1}\right].
\end{aligned}$$

Back to Eq. (C.3), one deduces

$$\begin{aligned}
\mathbb{E}\left[\widehat{R}^2_p(\mathcal{A}^{\mathcal{D}_n}_k)\right] - \mathbb{P}_{e,e'}\left(\widehat{f}_k(X_0) \neq Y_0, \widehat{f}^e_k(X_1) \neq Y_1\right) &= \frac{n}{p^2}A + \frac{n(n-1)}{p^2}B \\
&\leq \frac{1}{n} + \frac{2\sqrt{2}}{\sqrt{\pi}}\frac{(p+2)\sqrt{k}}{n} \ \cdot
\end{aligned}$$

Upper bound of (C.2) First observe that

$$\mathbb{P}_{e,e'}\left(\widehat{f}_k(X_0) \neq Y_0, \widehat{f}^e_k(X_i) \neq Y_i \mid i \notin e\right) = \mathbb{P}_{e,e'}\left(\widehat{f}^{(-1)}_k(X_0) \neq Y_0, \widehat{f}^e_k(X_{n+1}) \neq Y_{n+1}\right)$$

where $\widehat{f_k}^{(-1)}$ is built on sample $(X_2, Y_2), ..., (X_{n+1}, Y_{n+1})$. One has

$$\begin{aligned}
& \mathbb{P}\left(\widehat{f}_k(X_0) \neq Y_0, \widehat{f}_k(X_{n+1}) \neq Y_{n+1}\right) - \mathbb{P}_{e,e'}\left(\widehat{f}_k(X_0) \neq Y_0, \widehat{f}^e_k(X_i) \neq Y_i \mid i \notin e\right) \\
= \ & \mathbb{P}\left(\widehat{f}_k(X_0) \neq Y_0, \widehat{f}_k(X_{n+1}) \neq Y_{n+1}\right) - \mathbb{P}_{e,e'}\left(\widehat{f_k}^{(-1)}(X_0) \neq Y_0, \widehat{f}^e_k(X_{n+1}) \neq Y_{n+1}\right) \\
\leq \ & \mathbb{P}\left(\widehat{f}_k(X_0) \neq \widehat{f_k}^{(-1)}(X_0)\right) + \mathbb{P}_{e,e'}\left(\widehat{f}^e_k(X_{n+1}) \neq \widehat{f}_k(X_{n+1})\right) \\
\leq \ & \frac{4\sqrt{k}}{\sqrt{2\pi}n} + \frac{4p\sqrt{k}}{\sqrt{2\pi}n} \ ,
\end{aligned}$$

where we used Lemma D.6 again to obtain the last inequality.

Conclusion:

The conclusion simply results from combining bonds (C.1) and (C.2), which leads to

$$\mathbb{E}\left[\left(\widehat{R}_{p,n} - L_n\right)^2\right] \leq \frac{2\sqrt{2}}{\sqrt{\pi}} \frac{(2p+3)\sqrt{k}}{n} + \frac{1}{n} \quad \cdot$$

C.1.2. Combinatorial lemmas

All the lemmas of the present section are proved with the notation introduced at the beginning of Section C.1.

Lemma C.1. *For any* $1 \leq i \neq j \leq n$*,*

$$\mathbb{P}_{e,e'}\left(S^1_{i,j}\right) = \frac{\binom{n-2}{n-p}}{\binom{n}{n-p}} \times \frac{\binom{n-2}{n-p}}{\binom{n}{n-p}}, \qquad \mathbb{P}_{e,e'}\left(S^2_{i,j}\right) = \frac{\binom{n-p-1}{n-2}}{\binom{n}{n-p}} \times \frac{\binom{n-p}{n-2}}{\binom{n}{n-p}} \ ,$$

$$\mathbb{P}_{e,e'}\left(S^3_{i,j}\right) = \frac{\binom{n-p}{n-2}}{\binom{n}{n-p}} \frac{\binom{n-p-1}{n-2}}{\binom{n}{n-p}}, \qquad \mathbb{P}_{e,e'}\left(S^4_{i,j}\right) = \frac{\binom{n-p-1}{n-2}}{\binom{n}{n-p}} \times \frac{\binom{n-p-1}{n-2}}{\binom{n}{n-p}} \ \cdot$$

Proof of Lemma C.1. Along the proof, we repeatedly exploit the independence of the random variables e and e', which are set of $n-p$ distinct indices with the discrete uniform distribution over $\mathcal{E}_{n-p}$.

Note also that an important ingredient is that the probability of each one of the following events does not depend on the particular choice of the indices (i,j), but only on the fact that $i \neq j$.

$$\begin{aligned}
\mathbb{P}_{e,e'}\left(S^1_{i,j}\right) &= \mathbb{P}_{e,e'}\left(i \notin e,\ j \notin e',\ i \notin e',\ j \notin e\right) \\
&= \mathbb{P}_{e}\left(i \notin e,\ j \notin e\right) \mathbb{P}_{e'}\left(j \notin e',\ i \notin e'\right) = \frac{\binom{n-2}{n-p}}{\binom{n}{n-p}} \times \frac{\binom{n-2}{n-p}}{\binom{n}{n-p}} \quad \cdot \\
\mathbb{P}_{e,e'}\left(S^2_{i,j}\right) &= \mathbb{P}_{e,e'}\left(i \notin e,\ j \notin e',\ i \notin e',\ j \in e\right) \\
&= \mathbb{P}_{e}\left(i \notin e,\ j \in e\right) \mathbb{P}_{e'}\left(j \notin e',\ i \notin e'\right) = \frac{\binom{n-p-1}{n-2}}{\binom{n}{n-p}} \times \frac{\binom{n-p}{n-2}}{\binom{n}{n-p}} \quad \cdot \\
\mathbb{P}_{e,e'}\left(S^3_{i,j}\right) &= \mathbb{P}_{e,e'}\left(i \notin e,\ j \notin e',\ i \in e',\ j \notin e\right) \\
&= \mathbb{P}_{e}\left(i \notin e,\ j \notin e\right) \mathbb{P}_{e'}\left(j \notin e',\ i \in e'\right) = \frac{\binom{n-p}{n-2}}{\binom{n}{n-p}} \frac{\binom{n-p-1}{n-2}}{\binom{n}{n-p}} \quad \cdot \\
\mathbb{P}_{e,e'}\left(S^4_{i,j}\right) &= \mathbb{P}_{e,e'}\left(i \notin e,\ j \notin e',\ i \in e',\ j \in e\right) \\
&= \mathbb{P}_{e}\left(i \notin e,\ j \in e\right) \mathbb{P}_{e'}\left(j \notin e',\ i \in e'\right) = \frac{\binom{n-p-1}{n-2}}{\binom{n}{n-p}} \times \frac{\binom{n-p-1}{n-2}}{\binom{n}{n-p}} \quad \cdot
\end{aligned}$$

□

Lemma C.2. *With the above notation, for $\ell \in \{1,3\}$, it comes*

$$\mathbb{P}_e\left(\widehat{f}_k^e(X_1) \neq Y_1,\ \widehat{f}_k^{e'}(X_2) \neq Y_2 \mid S_{1,2}^\ell\right) - \mathbb{P}_e\left(\widehat{f}_k(X_0) \neq Y_0, \widehat{f}_k^e(X_1) \neq Y_1 \mid S_{1,2}^\ell\right) \leq \frac{4p\sqrt{k}}{\sqrt{2\pi}n} \cdot$$

Proof of Lemma C.2. First remind that as a test sample element Z_0 cannot belong to either e or e'. Consequently, an exhaustive formulation of

$$\mathbb{P}_e\left(\widehat{f}_k(X_0) \neq Y_0, \widehat{f}_k^e(X_1) \neq Y_1 \mid S_{1,2}^\ell\right) = \mathbb{P}_e\left(\widehat{f}_k(X_0) \neq Y_0, \widehat{f}_k^e(X_1) \neq Y_1 \mid S_{1,2}^\ell\right) .$$

Then it results

$$\mathbb{P}_e\left(\widehat{f}_k(X_0) \neq Y_0, \widehat{f}_k^e(X_1) \neq Y_1 \mid S_{1,2}^\ell\right) \quad = \quad \mathbb{P}_e\left(\widehat{f_k}^{(2)}(X_2) \neq Y_2,\ \widehat{f}_k^e(X_1) \neq Y_1 \mid S_{1,2}^\ell\right),$$

where $\widehat{f_k}^{(2)}$ is built on sample $(X_0,Y_0),(X_1,Y_1),(X_3,Y_3),...,(X_n,Y_n)$.

Hence Lemma D.6 implies

$$\begin{aligned}
&\mathbb{P}_{e,e'}\left(\widehat{f}_k^e(X_1) \neq Y_1,\ \widehat{f}_k^{e'}(X_2) \neq Y_2 \mid S_{1,2}^\ell\right) - \mathbb{P}_{e,e'}\left(\widehat{f}_k(X_0) \neq Y_0, \widehat{f}_k^e(X_1) \neq Y_1 \mid S_{1,2}^\ell\right)\\
&= \mathbb{P}_{e,e'}\left(\widehat{f}_k^e(X_1) \neq Y_1,\ \widehat{f}_k^{e'}(X_2) \neq Y_2 \mid S_{1,2}^\ell\right) - \mathbb{P}_{e,e'}\left(\widehat{f_k}^{(2)}(X_2) \neq Y_2,\ \widehat{f}_k^e(X_1) \neq Y_1 \mid S_{1,2}^\ell\right)\\
&\leq \mathbb{P}_{e,e'}\left(\left\{\widehat{f}_k^e(X_1) \neq Y_1\right\} \triangle \left\{\widehat{f}_k^e(X_1) \neq Y_1\right\} \mid S_{1,2}^\ell\right) + \mathbb{P}_{e,e'}\left(\left\{\widehat{f_k}^{(2)}(X_2) \neq Y_2\right\} \triangle \left\{\widehat{f}_k^{e'}(X_2) \neq Y_2\right\} \mid S_{1,2}^\ell\right)\\
&= \mathbb{P}_{e,e'}\left(\widehat{f_k}^{(2)}(X_2) \neq \widehat{f}_k^{e'}(X_2) \mid S_{1,2}^\ell\right) \leq \frac{4p\sqrt{k}}{\sqrt{2\pi}n} \cdot
\end{aligned}$$

□

Lemma C.3. *With the above notation, for $\ell \in \{2,4\}$, it comes*

$$\begin{aligned}
&\mathbb{P}_{e,e'}\left(\widehat{f}_k^e(X_1) \neq Y_1,\ \widehat{f}_k^{e'}(X_2) \neq Y_2 \mid S_{1,2}^\ell\right) - \mathbb{P}_{e,e'}\left(\widehat{f}_k(X_0) \neq Y_0, \widehat{f}_k^e(X_1) \neq Y_1 \mid S_{1,2}^\ell\right)\\
&\leq \frac{8\sqrt{k}}{\sqrt{2\pi}(n-p)} + \frac{4p\sqrt{k}}{\sqrt{2\pi}n} \cdot
\end{aligned}$$

Proof of Lemma C.3. As for the previous lemma, first notice that

$$\mathbb{P}_{e,e'}\left(\widehat{f}_k(X_0) \neq Y_0, \widehat{f}_k^e(X_1) \neq Y_1 \mid S_{1,2}^\ell\right) = \mathbb{P}_{e,e'}\left(\widehat{f_k}^{(2)}(X_2) \neq Y_2,\ \widehat{f_k}^{e_0}(X_1) \neq Y_1 \mid S_{1,2}^\ell\right),$$

where $\widehat{f_k}^{e_0}$ is built on sample e with observation (X_2,Y_2) replaced with (X_0,Y_0). Then

$$\begin{aligned}
&\mathbb{P}_{e,e'}\left(\widehat{f}_k^e(X_1) \neq Y_1,\ \widehat{f}_k^{e'}(X_2) \neq Y_2 \mid S_{1,2}^\ell\right) - \mathbb{P}_{e,e'}\left(\widehat{f}_k(X_0) \neq Y_0, \widehat{f}_k^e(X_1) \neq Y_1 \mid S_{1,2}^\ell\right)\\
&= \mathbb{P}_{e,e'}\left(\widehat{f}_k^e(X_1) \neq Y_1,\ \widehat{f}_k^{e'}(X_2) \neq Y_2 \mid S_{1,2}^\ell\right) - \mathbb{P}_{e,e'}\left(\widehat{f_k}^{(2)}(X_2) \neq Y_2,\ \widehat{f_k}^{e_0}(X_1) \neq Y_1 \mid S_{1,2}^\ell\right)\\
&\leq \mathbb{P}_{e,e'}\left(\left\{\widehat{f}_k^e(X_1) \neq Y_1\right\} \triangle \left\{\widehat{f_k}^{e_0}(X_1) \neq Y_1\right\} \mid S_{1,2}^\ell\right) + \mathbb{P}_{e,e'}\left(\left\{\widehat{f_k}^{(2)}(X_2) \neq Y_2\right\} \triangle \left\{\widehat{f}_k^{e'}(X_2) \neq Y_2\right\} \mid S_{1,2}^\ell\right)\\
&= \mathbb{P}_{e,e'}\left(\widehat{f}_k^e(X_1) \neq \widehat{f_k}^{e_0}(X_1) \mid S_{1,2}^\ell\right) + \mathbb{P}_{e,e'}\left(\widehat{f_k}^{(2)}(X_2) \neq \widehat{f}_k^{e'}(X_2) \mid S_{1,2}^\ell\right) \leq \frac{8\sqrt{k}}{\sqrt{2\pi}(n-p)} + \frac{4p\sqrt{k}}{\sqrt{2\pi}n} \cdot
\end{aligned}$$

□

C.2. Proof of Proposition 5.1

The bias of the L1O estimator is equal to

$$\begin{aligned}
&\mathbb{E}\left[L\left(\mathcal{A}_k^{\mathcal{D}_n}\right) - L\left(\mathcal{A}_k^{\mathcal{D}_{n-1}}\right) \right] \\
&= -2\mathbb{E}\left[(\eta(X) - 1/2) \left(\mathbb{E}\left[\mathbb{E}\left[\mathcal{A}_k^{\mathcal{D}_n}(X) - \mathcal{A}_k^{\mathcal{D}_{n-1}}(X) \mid X_{(k+1)}(X), X\right] \mid X\right]\right)\right] \\
&= -2\mathbb{E}\left[(\eta(X) - 1/2) \left(\mathbb{E}\left[\mathbb{E}\left[\mathcal{A}_k^{\mathcal{D}_n}(X) - \mathcal{A}_k^{\mathcal{D}_{n-1}}(X) \mid X_{(k+1)}(X), X\right] \mid X\right]\right)\right] \\
&= 1/2 \left\{ \mathbb{E}\left[\mathcal{A}_k^{\mathcal{D}_n}(0) - \mathcal{A}_k^{\mathcal{D}_{n-1}}(0) \mid X_{(k+1)}(0) = 0, X = 0 \right] \mathbb{P}\left[X_{(k+1)}(0) = 0 \mid X = 0\right] \right. \\
&\qquad\qquad \left. + \mathbb{E}\left[\mathcal{A}_k^{\mathcal{D}_n}(0) - \mathcal{A}_k^{\mathcal{D}_{n-1}}(0) \mid X_{(k+1)}(0) = 1, X = 0 \right] \mathbb{P}\left[X_{(k+1)}(0) = 1 \mid X = 0\right] \right\} \\
&\quad - 1/2 \left\{ \mathbb{E}\left[\mathcal{A}_k^{\mathcal{D}_n}(1) - \mathcal{A}_k^{\mathcal{D}_{n-1}}(1) \mid X_{(k+1)}(1) = 0, X = 1 \right] \mathbb{P}\left[X_{(k+1)}(1) = 0 \mid X = 1\right] \right. \\
&\qquad\qquad \left. + \mathbb{E}\left[\mathcal{A}_k^{\mathcal{D}_n}(1) - \mathcal{A}_k^{\mathcal{D}_{n-1}}(1) \mid X_{(k+1)}(1) = 1, X = 1 \right] \mathbb{P}\left[X_{(k+1)}(1) = 1 \mid X = 1\right] \right\},
\end{aligned}$$

where $X_{(k+1)}(x)$ denotes the $k+1$-th neighbor of x.

Then, a few remarks lead to simplify the above expression.

- On the one hand it is easy to check that

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{A}_k^{\mathcal{D}_n}(0) - \mathcal{A}_k^{\mathcal{D}_{n-1}}(0) \mid X_{(k+1)}(0) = 0, X = 0 \right] \\
&= \mathbb{E}\left[\mathcal{A}_k^{\mathcal{D}_n}(1) - \mathcal{A}_k^{\mathcal{D}_{n-1}}(1) \mid X_{(k+1)}(1) = 1, X = 1 \right] = 0,
\end{aligned}$$

  since all of the $k+1$ nearest neighbors share the same label.

- On the other hand, let us notice

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{A}_k^{\mathcal{D}_n}(0) - \mathcal{A}_k^{\mathcal{D}_{n-1}}(0) \mid X_{(k+1)}(0) = 1, X = 0 \right] \\
&= \mathbb{P}\left[\mathcal{A}_k^{\mathcal{D}_n}(0) = 1, \mathcal{A}_k^{\mathcal{D}_{n-1}}(0) = 0 \mid X_{(k+1)}(0) = 1, X = 0 \right] \\
&- \mathbb{P}\left[\mathcal{A}_k^{\mathcal{D}_n}(0) = 0, \mathcal{A}_k^{\mathcal{D}_{n-1}}(0) = 1 \mid X_{(k+1)}(0) = 1, X = 0 \right].
\end{aligned}$$

  Then knowing $X_{(k+1)}(X)$ and X are not equal implies the only way for $\mathcal{A}_k^{\mathcal{D}_n}$ and $\mathcal{A}_k^{\mathcal{D}_{n-1}}$ to differ is that the numbers of k nearest neighbors of each label are almost equal, that is either equal to $(k-1)/2$ or to $(k+1)/2$ (k is odd by assumption).

  With N_0^1 (respectively $\tilde{N}_0^1$) denoting the number of 1s among th k nearest neighbors of $X = 0$ among $X_1, \ldots, X_n$ (resp. $X_1, \ldots, X_{n-1}$), the proof of Theorem 3 in Chaudhuri and Dasgupta (2014) leads to

$$\begin{aligned}
&\mathbb{P}\left[\, \mathcal{A}_k^{\mathcal{D}_n}(0) = 1, \mathcal{A}_k^{\mathcal{D}_{n-1}}(0) = 0 \mid X_{(k+1)}(0) = 1, X = 0 \,\right] \\
&= \mathbb{P}\left[\, n \in V_k(0), N_0^1 = (k+1)/2, \tilde{N}_0^1 = (k-1)/2 \mid X_{(k+1)}(0) = 1, X = 0 \,\right] \\
&= \frac{k}{n} \times \mathbb{P}\left[\, \tilde{N}_0^1 = (k-1)/2 \mid N_0^1 = (k+1)/2, X_{(k+1)}(0) = 1, X = 0 \,\right] \\
&\times \mathbb{P}\left[\, N_0^1 = (k+1)/2 \mid X_{(k+1)}(0) = 1, X = 0 \,\right] \\
&= \frac{k}{n} \times \mathbb{P}\left[\, \mathcal{H}\left(\frac{k+1}{2}, \frac{k-1}{2}; 1\right) = 1 \,\right] \cdot \eta_1 \times \binom{k}{(k+1)/2} \bar{\eta}^{(k+1)/2} \left(1 - \bar{\eta}\right)^{(k-1)/2} \\
&= \frac{k+1}{2n} \times \eta_1 \times \binom{k}{(k+1)/2} \bar{\eta}^{(k+1)/2} \left(1 - \bar{\eta}\right)^{(k-1)/2},
\end{aligned}$$

where $\mathcal{H}(a, b; c)$ denotes a hypergeometric random variable with a successes in a population of cardinality $a + b$, and c draws, and $\bar{\eta} = \pi_0\eta_0 + (1 - \pi_0)\eta_1 = 1/2$.

Following the same reasoning for $\mathbb{P}\left[\, \mathcal{A}_k^{\mathcal{D}_n}(0) = 0, \mathcal{A}_k^{\mathcal{D}_{n-1}}(0) = 1 \mid X_{(k+1)}(0) = 1, X = 0 \,\right]$ and recalling that $\eta_0 = 0$ and $\eta_1 = 1$ by assumption, it results

$$\mathbb{E}\left[\, \mathcal{A}_k^{\mathcal{D}_n}(0) - \mathcal{A}_k^{\mathcal{D}_{n-1}}(0) \mid X_{(k+1)}(0) = 1, X = 0 \,\right] = -\frac{k+1}{2n} \times \binom{k}{(k+1)/2} (1/2)^k .$$

- Similar calculations applied to $X = 1$ finally lead to

$$\begin{aligned}
\mathbb{E}\left[\, L\left(\mathcal{A}_k^{\mathcal{D}_n}\right) - L\left(\mathcal{A}_k^{\mathcal{D}_{n-1}}\right) \,\right] &= \frac{k+1}{2n} \times \binom{k}{(k+1)/2} (1/2)^k \times \mathbb{P}\left[\, X_{(k+1)}(0) = 1 \mid X = 0 \,\right] \\
&= \frac{k+1}{2n} \times \binom{k}{(k+1)/2} (1/2)^k \times \mathbb{P}\left[\, \mathcal{B}(n, 1/2) \le k \,\right].
\end{aligned}$$

- The conclusion then follows from considering $k \ge n/2$ which entails that $\mathbb{P}\left[\, \mathcal{B}(n, 1/2) \le k \,\right] \ge 1/2$, and also by noticing that

$$\frac{k+1}{2n} \times \binom{k}{(k+1)/2} (1/2)^k \ge C \frac{\sqrt{k}}{n},$$

where denotes a numeric constant independent of n and k.

Appendix D. Technical results

D.1. Main inequalities

D.1.1. From moment to exponential inequalities

Proposition D.1 (see also Arlot (2007), Lemma 8.10)**.** *Let X denote a real valued random variable, and assume there exist $C \geq 1$, $\lambda_1, \ldots, \lambda_N > 0$, and $\alpha_1, \ldots, \alpha_N > 0$ ($N \in \mathbb{N}^*$) such that for every $q \geq q_0$,*

$$\mathbb{E}\left[\,|X|^q\,\right] \leq C \left(\sum_{i=1}^N \lambda_i q^{\alpha_i} \right)^q .$$

Then for every $t > 0$,

$$\mathbb{P}\left[\,|X| > t\,\right] \quad \leq C e^{q_0 \min_j \alpha_j} e^{-(\min_i \alpha_i) e^{-1} \min_j \left\{ \left(\frac{t}{N\lambda_j} \right)^{\frac{1}{\alpha_j}} \right\}}, \tag{D.1}$$

Furthermore for every $x > 0$, it results

$$\mathbb{P}\left[|X| > \sum_{i=1}^N \lambda_i \left(\frac{ex}{\min_j \alpha_j} \right)^{\alpha_i} \right] \leq C e^{q_0 \min_j \alpha_j} \cdot e^{-x}. \tag{D.2}$$

Proof of Proposition D.1. By use of Markov's inequality applied to $|X|^q$ ($q > 0$), it comes for every $t > 0$

$$\mathbb{P}\left[\,|X| > t\,\right] \leq \mathbb{1}_{q \geq q_0} \frac{\mathbb{E}\left[\,|X|^q\,\right]}{t^q} + \mathbb{1}_{q < q_0} \leq \mathbb{1}_{q \geq q_0} C \left(\frac{\sum_{i=1}^N \lambda_i q^{\alpha_i}}{t} \right)^q + \mathbb{1}_{q < q_0}.$$

Now using the upper bound $\sum_{i=1}^N \lambda_i q^{\alpha_i} \leq N \max_i \left\{ \lambda_i q^{\alpha_i} \right\}$ and choosing the particular value $\tilde{q} = \tilde{q}(t) = e^{-1} \min_j \left\{ \left(\frac{t}{N\lambda_j} \right)^{\frac{1}{\alpha_j}} \right\}$, one gets

$$\begin{aligned} \mathbb{P}\left[\,|X| > t\,\right] &\leq \mathbb{1}_{\tilde{q} \geq q_0} C \left(\frac{\max_i \left\{ N\lambda_i \left(e^{-\alpha_i} \min_j \left\{ \left(\frac{t}{N\lambda_j} \right)^{\frac{1}{\alpha_j}} \right\} \right)^{\alpha_i} \right\}}{t} \right)^{\tilde{q}} + \mathbb{1}_{\tilde{q} < q_0} \\ &\leq \mathbb{1}_{\tilde{q} \geq q_0} C e^{-(\min_i \alpha_i) \left[e^{-1} \min_j \left\{ \left(\frac{t}{N\lambda_j} \right)^{\frac{1}{\alpha_j}} \right\} \right]} + \mathbb{1}_{\tilde{q} < q_0}, \end{aligned}$$

which provides (D.1).

Let us now turn to the proof of (D.2). From $t^* = \sum_{i=1}^N \lambda_i \left(\frac{ex}{\min_j \alpha_j} \right)^{\alpha_i}$ combined with $q^* = \frac{x}{\min_j \alpha_j}$, it arises for every $x > 0$

$$\frac{\sum_{i=1}^N \lambda_i (q^*)^{\alpha_i}}{t^*} = \frac{\sum_{i=1}^N \lambda_i \left(e^{-1} \frac{ex}{\min_j \alpha_j} \right)^{\alpha_i}}{\sum_{i=1}^N \lambda_i \left(\frac{ex}{\min_j \alpha_j} \right)^{\alpha_i}} \leq \left(\max_k e^{-\alpha_k} \right) \frac{\sum_{i=1}^N \lambda_i \left(\frac{ex}{\min_j \alpha_j} \right)^{\alpha_i}}{\sum_{i=1}^N \lambda_i \left(\frac{ex}{\min_j \alpha_j} \right)^{\alpha_i}} = e^{-\min_k \alpha_k}.$$

Then,

$$C\left(\frac{\sum_{i=1}^N \lambda_i (q^*)^{\alpha_i}}{t^*}\right)^{q^*} \leq Ce^{-(\min_k \alpha_k)\frac{x}{\min_j \alpha_j}} = Ce^{-x}.$$

Hence,

$$\mathbb{P}\left[|X| > \sum_{i=1}^N \lambda_i \left(\frac{ex}{\min_j \alpha_j}\right)^{\alpha_i}\right] \leq Ce^{-x}\mathbb{1}_{q^* \geq q_0} + \mathbb{1}_{q^* < q_0} \leq Ce^{q_0 \min_j \alpha_j} \cdot e^{-x},$$

since $e^{q_0 \min_j \alpha_j} \geq 1$ and $-x + q_0 \min_j \alpha_j \geq 0$ if $q < q_0$.

□

D.1.2. Sub-Gaussian random variables

Lemma D.1 (Theorem 2.1 in Boucheron et al. (2013) first part)**.** *Any centered random variable X such that $\mathbb{P}\left(X > t\right) \vee \mathbb{P}\left(-X > t\right) \leq e^{-t^2/(2\nu)}$ satisfies*

$$\mathbb{E}\left[X^{2q}\right] \leq q!\left(4\nu\right)^q.$$

for all q in $\mathbb{N}_+$.

Lemma D.2 (Theorem 2.1 in Boucheron et al. (2013) second part)**.** *Any centered random variable X such that*

$$\mathbb{E}\left[X^{2q}\right] \leq q!C^q.$$

for some $C > 0$ and q in $\mathbb{N}_+$ satisfies $\mathbb{P}\left(X > t\right) \vee \mathbb{P}\left(-X > t\right) \leq e^{-t^2/(2\nu)}$ with $\nu = 4C$.

D.1.3. The Efron-Stein inequality

Theorem D.1 (Efron-Stein's inequality Boucheron et al. (2013), Theorem 3.1)**.** *Let $X_1, \ldots, X_n$ be independent random variables and let $Z = f\left(X_1, \ldots, X_n\right)$ be a square-integrable function. Then*

$$\mathrm{Var}(Z) \leq \sum_{i=1}^n \mathbb{E}\left[\left(Z - \mathbb{E}\left[Z \mid (X_j)_{j \neq i}\right]\right)^2\right] = \nu.$$

Moreover if $X_1', \ldots, X_n'$ denote independent copies of $X_1, \ldots, X_n$ and if we define for every $1 \leq i \leq n$

$$Z_i' = f\left(X_1, \ldots, X_i', \ldots, X_n\right),$$

then

$$\nu = \frac{1}{2}\sum_{i=1}^n \mathbb{E}\left[\left(Z - Z_i'\right)^2\right].$$

D.1.4. Generalized Efron-Stein's inequality

Theorem D.2 (Theorem 15.5 in Boucheron et al. (2013))**.** *Let $\xi_1, \ldots, \xi_n$ be n independent Ξ-valued random variables, $f : \Xi^n \to \mathbb{R}$ denote a measurable function, and define $\zeta = f(\xi_1, \ldots, \xi_n)$ and $\zeta'_i = f(\xi_1, \ldots, \xi'_i, \ldots, \xi_n)$, with $\xi'_1, \ldots, \xi'_n$ independent copies of ξ_i. Furthermore let $V_+ = \mathbb{E}\left[\sum_{i=1}^n \left[(\zeta - \zeta'_i)_+\right]^2 \mid \xi_1^n\right]$ and $V_- = \mathbb{E}\left[\sum_{i=1}^n \left[(\zeta - \zeta'_i)_-\right]^2 \mid \xi_1^n\right]$. Then there exists a constant $\kappa \leq 1,271$ such that for all q in $[2, +\infty[$,*

$$\left\|(\zeta - \mathbb{E}\zeta)_+\right\|_q \leq \sqrt{2\kappa q \left\|V_+\right\|_{q/2}} \ , \qquad \text{and} \qquad \left\|(\zeta - \mathbb{E}\zeta)_-\right\|_q \leq \sqrt{2\kappa q \left\|V_-\right\|_{q/2}} \ .$$

Corollary D.1. *With the same notation, it comes*

$$\left\|\zeta - \mathbb{E}\zeta\right\|_q \leq \sqrt{2\kappa q}\sqrt{\left\|\sum_{i=1}^n (\zeta - \zeta'_i)^2\right\|_{q/2}} \leq 2\sqrt{\kappa q}\sqrt{\left\|\sum_{i=1}^n \left(\zeta - \mathbb{E}\left[\zeta \mid (\xi_j)_{j\neq i}\right]\right)^2\right\|_{q/2}} \ . \tag{D.3}$$

Moreover considering $\zeta^{(j)} = f(\xi_1, \ldots, \xi_{j-1}, \xi_{j+1}, \ldots, \xi_n)$ for every $1 \leq j \leq n$, it results

$$\left\|\zeta - \mathbb{E}\zeta\right\|_q \ \leq 2\sqrt{2\kappa q}\sqrt{\left\|\sum_{i=1}^n \left(\zeta - \zeta^{(j)}\right)^2\right\|_{q/2}} \ . \tag{D.4}$$

D.1.5. McDiarmid's inequality

Theorem D.3. *Let $X_1, ..., X_n$ be independent random variables taking values in a set A, and assume that $f : A^n \to \mathbb{R}$ satisfies*

$$\sup_{x_1,...,x_n,x'_i} \left|f(x_1, ..., x_i, ..., x_n) - f(x_1, ..., x'_i, ..., x_n)\right| \leq c_i, \ 1 \leq i \leq n \ .$$

Then for all $\varepsilon > 0$, one has

$$\begin{aligned} \mathbb{P}\left(f(X_1, ..., X_n) - E\left[f(X_1, ..., X_n)\right] \geq \varepsilon\right) &\leq e^{-2\varepsilon^2/\sum_{i=1}^n c_i^2} \\ \mathbb{P}\left(E\left[f(X_1, ..., X_n)\right] - f(X_1, ..., X_n) \geq \varepsilon\right) &\leq e^{-2\varepsilon^2/\sum_{i=1}^n c_i^2} \end{aligned}$$

A proof can be found in Devroye et al. (1996) (see Theorem 9.2).

D.1.6. Rosenthal's inequality

Proposition D.2 (Eq. (20) in Ibragimov and Sharakhmetov (2002))**.** *Let $X_1, \ldots, X_n$ denote independent real random variables with symmetric distributions. Then for every $q > 2$ and $\gamma > 0$,*

$$E\left[\left|\sum_{i=1}^n X_i\right|^q\right] \leq B(q, \gamma)\left\{\gamma \sum_{i=1}^n E\left[\,|X_i|^q\,\right] \vee \left(\sqrt{\sum_{i=1}^n E\left[\,X_i^2\,\right]}\right)^q\right\},$$

where $a \vee b = \max(a, b)$ $(a, b \in \mathbb{R})$, and $B(q, \gamma)$ denotes a positive constant only depending on q and γ. Furthermore, the optimal value of $B(q, \gamma)$ is given by

$$\begin{aligned} B^*(q, \gamma) &= 1 + \tfrac{E[\,|N|^q\,]}{\gamma} && , \ \textit{if} \quad 2 < q \leq 4, \\ &= \gamma^{-q/(q-1)} E\left[\,|Z - Z'|^q\,\right] && , \ \textit{if} \quad 4 < q, \end{aligned}$$

where N denotes a standard Gaussian variable, and Z, Z' are i.i.d. random variables with Poisson distribution $\mathcal{P}\left(\frac{\gamma^{1/(q-1)}}{2}\right)$.

Proposition D.3. *Let $X_1, \ldots, X_n$ denote independent real random variables with symmetric distributions. Then for every $q > 2$,*

$$E\left[\left|\sum_{i=1}^n X_i\right|^q\right] \le \left(2\sqrt{2e}\right)^q \max\left\{q^q \sum_{i=1}^n E\left[\,|X_i|^q\,\right], (\sqrt{q})^q \left(\sqrt{\sum_{i=1}^n E\left[\,X_i^2\,\right]}\right)^q\right\}.$$

Proof of Proposition D.3. From Lemma D.3, let us observe

- if $2 < q \le 4$, choosing $\gamma = 1$ provides
$$B^*(q,\gamma) \le \left(2\sqrt{2e}\sqrt{q}\right)^q.$$
- if $4 < q$, $\gamma = q^{(q-1)/2}$ leads to
$$B^*(q,\gamma) \le q^{-q/2}\left(\sqrt{4eq\left(q^{1/2}+q\right)}\right)^q \le q^{-q/2}\left(\sqrt{8eq}\right)^q = \left(2\sqrt{2e}\sqrt{q}\right)^q.$$

Plugging the previous upper bounds in Rosenthal's inequality (Proposition D.2), it results for every $q > 2$

$$E\left[\left|\sum_{i=1}^n X_i\right|^q\right] \le \left(2\sqrt{2e}\sqrt{q}\right)^q \max\left\{(\sqrt{q})^q \sum_{i=1}^n E\left[\,|X_i|^q\,\right], \left(\sqrt{\sum_{i=1}^n E\left[\,X_i^2\,\right]}\right)^q\right\}.$$

□

Lemma D.3. *With the same notation as Proposition D.2 and for every $\gamma > 0$, it comes*

- *for every $2 < q \le 4$,*
$$B^*(q,\gamma) \le 1 + \frac{\left(\sqrt{2e}\sqrt{q}\right)^q}{\gamma},$$
- *for every $4 < q$,*
$$B^*(q,\gamma) \le \gamma^{-q/(q-1)}\left(\sqrt{4eq\left(\gamma^{1/(q-1)}+q\right)}\right)^q.$$

Proof of Lemma D.3. If $2 < q \le 4$,

$$B^*(q,\gamma) = 1 + \frac{E\left[\,|N|^q\,\right]}{\gamma} \le 1 + \frac{\sqrt{2}e\sqrt{q}\left(\frac{q}{e}\right)^{\frac{q}{2}}}{\gamma} \le 1 + \frac{\sqrt{2e}^q\sqrt{e}^q\left(\frac{q}{e}\right)^{\frac{q}{2}}}{\gamma} = 1 + \frac{\left(\sqrt{2e}\sqrt{q}\right)^q}{\gamma}\ ,$$

by use of Lemma D.9 and $\sqrt{q}^{1/q} \leq \sqrt{e}$ for every $q > 2$.

If $q > 4$,

$$\begin{aligned}
B^*(q,\gamma) &= \gamma^{-q/(q-1)} E\left[\left|Z - Z'\right|^q \right] \\
&\leq \gamma^{-q/(q-1)} 2^{q/2+1} e\sqrt{q} \left[\frac{q}{e} \left(\gamma^{1/(q-1)} + q \right) \right]^{q/2} \\
&\leq \gamma^{-q/(q-1)} 2^{q/2} \sqrt{2e}^q \sqrt{e}^q \left[\frac{q}{e} \left(\gamma^{1/(q-1)} + q \right) \right]^{q/2} \\
&\leq \gamma^{-q/(q-1)} \left[4eq \left(\gamma^{1/(q-1)} + q \right) \right]^{q/2} = \gamma^{-q/(q-1)} \left(\sqrt{4eq \left(\gamma^{1/(q-1)} + q \right)} \right)^q ,
\end{aligned}$$

applying Lemma D.11 with $\lambda = 1/2\gamma^{1/(q-1)}$.

□

D.2. Technical lemmas

D.2.1. Basic computations for resampling applied to the kNN algorithm

Lemma D.4. *For every $1 \leq i \leq n$ and $1 \leq p \leq n$, one has*

$$\mathbb{P}_e\left(i \in \bar{e}\right) = \frac{p}{n}, \tag{D.5}$$

$$\sum_{j=1}^{n} \mathbb{P}_e\left[\, i \in \bar{e},\ j \in V_k^e(X_i) \,\right] = \frac{kp}{n} \ , \tag{D.6}$$

$$\sum_{k<\sigma_i(j)\leq k+p} \mathbb{P}_e\left[\, i \in \bar{e},\ j \in V_k^e(X_i) \,\right] = \frac{kp}{n}\frac{p-1}{n-1} \ \cdot \tag{D.7}$$

Proof of Lemma D.4. The first equality is straightforward. The second one results from simple calculations as follows.

$$\begin{aligned}
\sum_{j=1}^{n} \mathbb{P}_e\left[\, i \in \bar{e},\ j \in V_k^e(X_i) \,\right] &= \sum_{j=1}^{n} \binom{n}{p}^{-1} \sum_{e} \mathbb{1}_{i\in\bar{e}} \mathbb{1}_{j \in V_k^e(X_i)} = \binom{n}{p}^{-1} \sum_{e} \mathbb{1}_{i\in\bar{e}} \left(\sum_{j=1}^{n} \mathbb{1}_{j \in V_k^e(X_i)} \right) \\
&= \left(\binom{n}{p}^{-1} \sum_{e} \mathbb{1}_{i\in\bar{e}} \right) k = \frac{p}{n} k \ .
\end{aligned}$$

For the last equality, let us notice every $j \in V_i$ satisfies

$$\mathbb{P}_e\left[\, i \in \bar{e},\ j \in V_k^e(X_i) \,\right] = \mathbb{P}_e\left[\, j \in V_k^e(X_i) \mid i \in \bar{e} \,\right] \mathbb{P}_e\left[\, i \in \bar{e} \,\right] = \frac{n-1}{n-p}\frac{p}{n} \ ,$$

hence

$$\begin{aligned}
\sum_{k<\sigma_i(j)\leq k+p} \mathbb{P}_e\left[\, i \in \bar{e},\ j \in V_k^e(X_i) \,\right] &= \sum_{j=1}^{n} \mathbb{P}_e\left[\, i \in \bar{e},\ j \in V_k^e(X_i) \,\right] - \sum_{\sigma_i(j)\leq k} \mathbb{P}_e\left[\, i \in \bar{e},\ j \in V_k^e(X_i) \,\right] \\
&= k\frac{p}{n} - k\frac{n-1}{n-p}\frac{p}{n} = k\frac{p}{n}\frac{p-1}{n-1} \ \cdot
\end{aligned}$$

□

D.2.2. Stone's lemma

Lemma D.5 (Devroye et al. (1996), Corollary 11.1, p. 171)**.** *Given n points $(x_1, ..., x_n)$ in $\mathbb{R}^d$, any of these points belongs to the k nearest neighbors of at most $k\gamma_d$ of the other points, where γ_d increases on d.*

D.2.3. Stability of the kNN classifier when removing p observations

Lemma D.6 (Devroye and Wagner (1979b), Eq. (14))**.** *For every $1 \leq k \leq n$, let $\mathcal{A}_k$ denote k-NN classification algorithm defined by Eq. (2.1), and let $Z_1, \ldots, Z_n$ denote n i.i.d. random variables such that for every $1 \leq i \leq n$, $Z_i = (X_i, Y_i) \sim P$. Then for every $1 \leq p \leq n - k$,*

$$\mathbb{P}\left[\, \mathcal{A}_k(Z_{1,n}; X) \neq \mathcal{A}_k(Z_{1,n-p}; X) \,\right] \leq \frac{4}{\sqrt{2\pi}} \frac{p\sqrt{k}}{n} \quad ,$$

where $Z_{1,i} = (Z_1, \ldots, Z_i)$ for every $1 \leq i \leq n$, and $(X, Y) \sim P$ is independent of $Z_{1,n}$.

D.2.4. Exponential concentration inequality for the L1O estimator

Lemma D.7 (Devroye et al. (1996), Theorem 24.4)**.** *For every $1 \leq k \leq n$, let $\mathcal{A}_k$ denote k-NN classification algorithm defined by Eq. (2.1). Let also $\widehat{R}_1(\cdot)$ denote the L1O estimator defined by Eq. (2.2) with $p = 1$. Then for every $\varepsilon > 0$,*

$$\mathbb{P}\left(\left| \widehat{R}_1(\mathcal{A}_k, Z_{1,n}) - \mathbb{E}\left[\widehat{R}_1(\mathcal{A}_k, Z_{1,n}) \right] \right| > \varepsilon \right) \leq 2 \exp\left\{ -n \frac{\varepsilon^2}{\gamma_d^2 k^2} \right\} .$$

D.2.5. Moment upper bounds for the L1O estimator

Lemma D.8. *For every $1 \leq k \leq n$, let $\mathcal{A}_k$ denote k-NN classification algorithm defined by Eq. (2.1). Let also $\widehat{R}_1(\cdot)$ denote the L1O estimator defined by Eq. (2.2) with $p = 1$. Then for every $q \geq 1$,*

$$\mathbb{E}\left[\left| \widehat{R}_1\left(\mathcal{A}_k, Z_{1,n}\right) - \mathbb{E}\left[\widehat{R}_1\left(\mathcal{A}_k, Z_{1,n}\right) \right] \right|^{2q} \right] \leq q! \left(2 \frac{(k\gamma_d)^2}{n} \right)^q . \tag{D.8}$$

The proof is straightforward from the combination of Lemmas D.1 and D.7.

D.2.6. Upper bound on the optimal constant in the Rosenthal's inequality

Lemma D.9. *Let N denote a real-valued standard Gaussian random variable. Then for every $q > 2$, one has*

$$\mathbb{E}\left[\, |N|^q \,\right] \leq \sqrt{2} e \sqrt{q} \left(\frac{q}{e} \right)^{\frac{q}{2}} .$$

Proof of Lemma D.9. If q is even ($q = 2k > 2$), then

$$\begin{aligned} \mathbb{E}\left[\, |N|^q \,\right] &= 2 \int_0^{+\infty} x^q \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \sqrt{\frac{2}{\pi}} (q-1) \int_0^{+\infty} x^{q-2} e^{-\frac{x^2}{2}} dx \\ &= \sqrt{\frac{2}{\pi}} \frac{(q-1)!}{2^{k-1}(k-1)!} = \sqrt{\frac{2}{\pi}} \frac{q!}{2^{q/2}(q/2)!} \quad . \end{aligned}$$

Then using for any positive integer a

$$\sqrt{2\pi a}\left(\frac{a}{e}\right)^a < a! < \sqrt{2e\pi a}\left(\frac{a}{e}\right)^a ,$$

it results

$$\frac{q!}{2^{q/2}(q/2)!} < \sqrt{2e}\, e^{-q/2} q^{q/2},$$

which implies

$$\mathbb{E}\left[\, |N|^q \,\right] \le 2\sqrt{\frac{e}{\pi}}\left(\frac{q}{e}\right)^{q/2} < \sqrt{2}e\sqrt{q}\left(\frac{q}{e}\right)^{\frac{q}{2}} \quad .$$

If q is odd ($q = 2k+1 > 2$), then

$$\mathbb{E}\left[\, |N|^q \,\right] = \sqrt{\frac{2}{\pi}}\int_0^{+\infty} x^q e^{-\frac{x^2}{2}}\, dx = \sqrt{\frac{2}{\pi}}\int_0^{+\infty} \sqrt{2t}^q e^{-t}\frac{dt}{\sqrt{2t}},$$

by setting $x = \sqrt{2t}$. In particular, this implies

$$\mathbb{E}\left[\, |N|^q \,\right] \le \sqrt{\frac{2}{\pi}}\int_0^{+\infty} (2t)^k\, e^{-t} dt = \sqrt{\frac{2}{\pi}} 2^k k! = \sqrt{\frac{2}{\pi}} 2^{\frac{q-1}{2}}\left(\frac{q-1}{2}\right)! < \sqrt{2}e\sqrt{q}\left(\frac{q}{e}\right)^{\frac{q}{2}} .$$

□

Lemma D.10. *Let S denote a binomial random variable such that $S \sim \mathcal{B}(k, 1/2)$ ($k \in \mathbb{N}^*$). Then for every $q > 3$, it comes*

$$\mathbb{E}\left[\, |S - \mathbb{E}\left[\, S \,\right]|^q \,\right] \le 4\sqrt{e}\sqrt{q}\sqrt{\frac{qk}{2e}}^q \quad .$$

Proof of Lemma D.10. Since $S - \mathbb{E}(S)$ is symmetric, it comes

$$\mathbb{E}\left[\, |S - \mathbb{E}\left[\, S \,\right]|^q \,\right] = 2\int_0^{+\infty} \mathbb{P}\left[\, S < \mathbb{E}\left[\, S \,\right] - t^{1/q} \,\right] dt = 2q\int_0^{+\infty} \mathbb{P}\left[\, S < \mathbb{E}\left[\, S \,\right] - u \,\right] u^{q-1}\, du.$$

Using Chernoff's inequality and setting $u = \sqrt{k/2}v$, it results

$$\mathbb{E}\left[\, |S - \mathbb{E}\left[\, S \,\right]|^q \,\right] \le 2q\int_0^{+\infty} u^{q-1} e^{-\frac{u^2}{k}}\, du = 2q\sqrt{\frac{k}{2}}^q \int_0^{+\infty} v^{q-1} e^{-\frac{v^2}{2}}\, dv.$$

If q is even, then $q-1 > 2$ is odd and the same calculations as in the proof of Lemma D.9 apply, which leads to

$$\mathbb{E}\left[\, |S - \mathbb{E}\left[\, S \,\right]|^q \,\right] \le 2\sqrt{\frac{k}{2}}^q 2^{q/2}\left(\frac{q}{2}\right)! \le 2\sqrt{\frac{k}{2}}^q 2^{q/2}\sqrt{\pi e q}\left(\frac{q}{2e}\right)^{q/2} = 2\sqrt{\pi e}\sqrt{q}\sqrt{\frac{qk}{2e}}^q < 4\sqrt{e}\sqrt{q}\sqrt{\frac{qk}{2e}}^q \quad .$$

If q is odd, then $q-1>2$ is even and another use of the calculations in the proof of Lemma D.9 provides

$$\mathbb{E}\left[\,|S-\mathbb{E}\left[\,S\,\right]|^q\,\right] \leq 2q\sqrt{\frac{k}{2}}^q \frac{(q-1)!}{2^{(q-1)/2}\frac{q-1}{2}!} = 2\sqrt{\frac{k}{2}}^q \frac{q!}{2^{(q-1)/2}\frac{q-1}{2}!}.$$

Let us notice

$$\frac{q!}{2^{(q-1)/2}\frac{q-1}{2}!} \leq \frac{\sqrt{2\pi e q}\left(\frac{q}{e}\right)^q}{2^{(q-1)/2}\sqrt{\pi(q-1)}\left(\frac{q-1}{2e}\right)^{(q-1)/2}} = \sqrt{2e}\sqrt{\frac{q}{q-1}}\frac{\left(\frac{q}{e}\right)^q}{\left(\frac{q-1}{e}\right)^{(q-1)/2}}$$
$$= \sqrt{2e}\sqrt{\frac{q}{q-1}}\left(\frac{q}{e}\right)^{(q+1)/2}\left(\frac{q}{q-1}\right)^{(q-1)/2}$$

and also that

$$\sqrt{\frac{q}{q-1}}\left(\frac{q}{q-1}\right)^{(q-1)/2} \leq \sqrt{2e}.$$

This implies

$$\frac{q!}{2^{(q-1)/2}\frac{q-1}{2}!} \leq 2e\left(\frac{q}{e}\right)^{(q+1)/2} = 2\sqrt{e}\sqrt{q}\left(\frac{q}{e}\right)^{q/2},$$

hence

$$\mathbb{E}\left[\,|S-\mathbb{E}\left[\,S\,\right]|^q\,\right] \leq 2\sqrt{\frac{k}{2}}^q 2\sqrt{e}\sqrt{q}\left(\frac{q}{e}\right)^{q/2} = 4\sqrt{e}\sqrt{q}\sqrt{\frac{qk}{2e}}^q \ .$$

□

Lemma D.11. *Let X, Y be two i.i.d. random variables with Poisson distribution $\mathcal{P}(\lambda)$ ($\lambda > 0$). Then for every $q > 3$, it comes*

$$\mathbb{E}\left[\,|X-Y|^q\,\right] \leq 2^{q/2+1}e\sqrt{q}\left[\frac{q}{e}\left(2\lambda+q\right)\right]^{q/2}.$$

Proof of Lemma D.11. Let us first remark that

$$\mathbb{E}\left[\,|X-Y|^q\,\right] = \mathbb{E}_N\left[\mathbb{E}\left[\,|X-Y|^q \mid N\,\right]\right] = 2^q\mathbb{E}_N\left[\mathbb{E}\left[\,|X-N/2|^q \mid N\,\right]\right],$$

where $N = X+Y$. Furthermore, the conditional distribution of X given $N = X+Y$ is a binomial distribution $\mathcal{B}(N, 1/2)$. Then Lemma D.10 provides that

$$\mathbb{E}\left[\,|X-N/2|^q \mid N\,\right] \leq 4\sqrt{e}\sqrt{q}\sqrt{\frac{qN}{2e}}^q \qquad a.s.,$$

which entails that

$$\mathbb{E}\left[\,|X-Y|^q\,\right] \leq 2^q\mathbb{E}_N\left[4\sqrt{e}\sqrt{q}\sqrt{\frac{qN}{2e}}^q\right] = 2^{q/2+2}\sqrt{e}\sqrt{q}\sqrt{\frac{q}{e}}^q\,\mathbb{E}_N\left[N^{q/2}\right].$$

It only remains to upper bound the last expectation where N is a Poisson random variable $\mathcal{P}(2\lambda)$ (since X, Y are *i.i.d.*):

$$\mathbb{E}_N\left[N^{q/2}\right] \leq \sqrt{\mathbb{E}_N\left[N^q\right]}$$

by Jensen's inequality. Further introducing Touchard polynomials and using a classical upper bound, it comes

$$\begin{aligned}\mathbb{E}_N\left[N^{q/2}\right] &\leq \sqrt{\sum_{i=1}^{q}(2\lambda)^i \frac{1}{2}\binom{q}{i} i^{q-i}} \leq \sqrt{\sum_{i=0}^{q}(2\lambda)^i \frac{1}{2}\binom{q}{i} q^{q-i}} \\ &= \sqrt{\frac{1}{2}\sum_{i=0}^{q}\binom{q}{i}(2\lambda)^i q^{q-i}} = \sqrt{\frac{1}{2}\left(2\lambda+q\right)^q} = 2^{\frac{-1}{2}}\left(2\lambda+q\right)^{q/2}.\end{aligned}$$

Finally, one concludes

$$\mathbb{E}\left[\,|X-Y|^q\,\right] \leq 2^{q/2+2}\sqrt{e}\sqrt{q}\sqrt{\frac{q}{e}}^{q} 2^{\frac{-1}{2}}\left(2\lambda+q\right)^{q/2} < 2^{q/2+1}e\sqrt{q}\left[\frac{q}{e}\left(2\lambda+q\right)\right]^{q/2}.$$

□

References

A. Andoni and P. Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*, pages 459–468. IEEE, 2006.

S. Arlot. *Resampling and Model Selection.* PhD thesis, University Paris-Sud 11, December 2007. URL `http://tel.archives-ouvertes.fr/tel-00198803/en/`. oai:tel.archives-ouvertes.fr:tel-00198803_v1.

S. Arlot and F. Bach. Data-driven calibration of linear estimators with minimal penalties. *Advances in Neural Information Processing Systems (NIPS)*, 2:46–54, 2009.

S. Arlot and A. Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79, 2010.

S. Arlot and M. Lerasle. Why v= 5 is enough in v-fold cross-validation. *arXiv preprint arXiv:1210.5830*, 2012.

T. B. Berrett, R. J. Samworth, and M. Yuan. Efficient multivariate entropy estimation via k-nearest neighbour distances. *arXiv preprint arXiv:1606.00304*, 2016.

G. Biau and L. Devroye. *Lectures on the nearest neighbor method.* Springer, 2016.

G. Biau, F. Cérou, and A. Guyader. On the rate of convergence of the bagged nearest neighbor estimate. *The Journal of Machine Learning Research*, 11:687–712, 2010a.

G. Biau, F. Cérou, and A. Guyader. Rates of convergence of the functional-nearest neighbor estimate. *Information Theory, IEEE Transactions on*, 56(4):2034–2040, 2010b.

S. Boucheron, O. Bousquet, G. Lugosi, and P. Massart. Moment inequalities for functions of independent random variables. *Ann. Probab.*, 33(2):514–560, 2005. ISSN 0091-1798.

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

L. Breiman and Ph. Spector. Submodel selection and evaluation in regression. the x-random case. *International Statistical Review*, 60(3):291–319, 1992.

P. Burman. Comparative study of Ordinary Cross-Validation, v-Fold Cross-Validation and the repeated Learning-Testing Methods. *Biometrika*, 76(3):503–514, 1989.

T. I. Cannings, T. B. Berrett, and R. J. Samworth. Local nearest neighbour classification with applications to semi-supervised learning. *arXiv preprint arXiv:1704.00642*, 2017.

A. Celisse. *Model selection via cross-validation in density estimation, regression and change-points detection. (In English)*. PhD thesis, University Paris-Sud 11. `http://tel.archives-ouvertes.fr/tel-00346320/en/`., December 2008. URL `http://tel.archives-ouvertes.fr/tel-00346320/en/`.

A. Celisse. Optimal cross-validation in density estimation with the l^2-loss. *The Annals of Statistics*, 42(5):1879–1910, 2014.

A. Celisse and T. Mary-Huard. Exact cross-validation for knn: applications to passive and active learning in classification. *JSFdS*, 152(3), 2011.

A. Celisse and S. Robin. Nonparametric density estimation by exact leave-p-out cross-validation. *Computational Statistics and Data Analysis*, 52(5):2350–2368, 2008.

K. Chaudhuri and S. Dasgupta. Rates of convergence for nearest neighbor classification. In *Advances in Neural Information Processing Systems*, pages 3437–3445, 2014.

T. M. Cover. Rates of convergence for nearest neighbor procedures. In *Proceedings of the Hawaii International Conference on Systems Sciences*, pages 413–415, 1968.

T. M. Cover and P. E. Hart. Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13(1):21–27, 1967.

L. Devroye and T. Wagner. Distribution-free performance bounds for potential function rules. *IEEE Transactions on Information Theory*, 25:601–604, 1979a.

L. Devroye, L. Györfi, and G. Lugosi. *A Probilistic Theory of Pattern Recognition*. Springer Verlag, 1996.

L. P. Devroye and T. J. Wagner. The strong uniform consistency of nearest neighbor density estimates. *Ann. Statist.*, 5(3):536–540, 1977. ISSN 0090-5364.

L. P. Devroye and T. J. Wagner. Distribution-free inequalities for the deleted and holdout error estimates. *Information Theory, IEEE Transactions on*, 25(2):202–207, 1979b.

L.P. Devroye and T.J. Wagner. Distribution-free consistency results in nonparametric discrimination and regression function estimation. *Ann. Statist.*, 8(2):231–239, 1980.

E. Fix and J. Hodges. *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*, chapter Discriminatory analysis- nonparametric discrimination: Consistency principles. IEEE Computer Society Press, Los Alamitos, CA, 1951. Reprint of original work from 1952.

M. Fuchs, R. Hornung, R. De Bin, and A.-L. Boulesteix. A u-statistic estimator for the variance of resampling-based error estimators. Technical report, arXiv, 2013.

S. Geisser. The predictive sample reuse method with applications. *J. Amer. Statist. Assoc.*, 70:320–328, 1975.

L. Györfi. The rate of convergence of k_n-nn regression estimates and classification rules. *IEEE Trans. Commun*, 27(3):362–364, 1981.

P. Hall, B. U. Park, and R. J. Samworth. Choice of neighbor order in nearest-neighbor classification. *The Annals of Statistics*, pages 2135–2152, 2008.

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning.* Springer Series in Statistics. Springer-Verlag, New York, 2001. ISBN 0-387-95284-5. Data mining, inference, and prediction.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journ. of the American Statistical Association*, 58(301):13–30, 1963.

R. Ibragimov and S. Sharakhmetov. On extremal problems and best constants in moment inequalities. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 42–56, 2002.

P. Indyk and R. Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pages 604–613. ACM, 1998.

M. Kearns and D. Ron. Algorithmic Stability and Sanity-Check Bounds for Leave-One-Out Cross-Validation. *Neural Computation*, 11:1427–1453, 1999.

V. S. Koroljuk and Y. V. Borovskich. *Theory of U-statistics.* Springer, 1994.

S. R. Kulkarni and S. E. Posner. Rates of convergence of nearest neighbor estimation under arbitrary sampling. *Information Theory, IEEE Transactions on*, 41(4):1028–1039, 1995.

L. Li, D. M. Umbach, P. Terry, and J. A. Taylor. Application of the ga/knn method to seldi proteomics data. *Bioinformatics*, 20(10):1638–1640, 2004.

P. Massart and E. Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, pages 2326–2366, 2006.

D. Psaltis, R. R. Snapp, and S. S. Venkatesh. On the finite sample performance of the nearest neighbor classifier. *Information Theory, IEEE Transactions on*, 40(3):820–837, 1994.

W. H. Rogers and T. J. Wagner. A finite sample distribution-free performance bound for local discrimination rules. *Annals of Statistics*, 6(3):506–514, 1978.

E. D. Scheirer and M. Slaney. Multi-feature speech/music discrimination system, May 27 2003. US Patent 6,570,991.

R. J. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons Inc., 1980.

J. Shao. Linear model selection by cross-validation. *J. Amer. Statist. Assoc.*, 88(422): 486–494, 1993. ISSN 0162-1459.

J. Shao. An asymptotic theory for linear model selection. *Statistica Sinica*, 7:221–264, 1997.

P. Y. Simard, Y. A. LeCun, J. S. Denker, and B. Victorri. Transformation invariance in pattern recognition tangent distance and tangent propagation. In *Neural networks: tricks of the trade*, pages 239–274. Springer, 1998.

R. R Snapp and S. S. Venkatesh. Asymptotic expansions of the k nearest neighbor risk. *The Annals of Statistics*, 26(3):850–878, 1998.

B. M. Steele. Exact bootstrap k-nearest neighbor learners. *Machine Learning*, 74(3):235–255, 2009.

C. J. Stone. Consistent nonparametric regression. *The annals of statistics*, pages 595–620, 1977.

C. J. Stone. Optimal global rates of convergence for nonparametric regression. *Ann. Statist.*, 10(4):1040–1053, 1982. ISSN 0090-5364.

M. Stone. Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B*, 36:111–147, 1974. ISSN 0035-9246. With discussion by G. A. Barnard, A. C. Atkinson, L. K. Chan, A. P. Dawid, F. Downton, J. Dickey, A. G. Baker, O. Barndorff-Nielsen, D. R. Cox, S. Giesser, D. Hinkley, R. R. Hocking, and A. S. Young, and with a reply by the authors.

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

Y. Yang. Comparing learning methods for classification. *Statistica Sinica*, 16(2):635–657, 2006. ISSN 1017-0405.

Y. Yang. Consistency of cross-validation for comparing regression procedures. *The Annals of Statistics*, 35(6):2450–2473, 2007.

P. Zhang. Model selection via multifold cross validation. *Ann. Statist.*, 21(1):299–313, 1993. ISSN 0090-5364.