New upper bounds on cross-validation for the k-Nearest Neighbor classification rule

Alain Celisse, Tristan Mary-Huard

New upper bounds on cross-validation for the k-Nearest Neighbor classification rule

Alain Celisse
Laboratoire de Mathématiques, UMR 8524 CNRS-Université Lille 1
Modal Project-Team
F-59 655 Villeneuve d'Ascq Cedex, France
celisse@math.univ-lille1.fr

and

Tristan Mary-Huard
INRA, UMR 0320 / UMR 8120 Génétique Végétale et Évolution
Le Moulon, F-91190 Gif-sur-Yvette, France
UMR AgroParisTech INRA MIA 518, Paris, France
16 rue Claude Bernard
F-75 231 Paris cedex 05, France
maryhuar@agroparistech.fr

Abstract

The present work addresses binary classification by use of the k-nearest neighbors (kNN) classifier. Among several assets, it belongs to intuitive majority vote classification rules and also adapts to spatial inhomogeneity, which is particularly relevant in high dimensional settings where no *a priori* partitioning of the space seems realistic.

However the performance of the kNN classifier crucially depends on the number k of neighbors that will be considered. To calibrate the parameter k, cross-validation procedures such as V-fold or leave-one-out are usually used. But on the one hand these procedures can become highly time-consuming. On the other hand, not that much theoretical guaranties do exist on the performance of such procedures. Recently [11] have derived closed-form formulas for the leave-p-out estimator of the kNN classifier performance. Such formulas now allow to efficiently perform cross-validation.

The main purpose of the present article is twofold: First, we provide a new strategy to derive bounds on moments of the leave-p-out estimator used to assess the performance of the kNN classifier. This new strategy exploits the link between leave-p-out and U-statistics as well as the generalized Efron-Stein inequality. Second, these moment upper bounds are used to settle a new exponential concentration inequality for

the LpO risk estimator, which characterizes its behavior with respect to the influential parameter k.

1 Introduction

Cross-validation (CV) refers to a set of widely used procedures introduced and analyzed by [45, 24, 44] to assess the performance of an estimator and choose the best one among a given collection, that is to perform *model selection* (or parameter calibration). For a given $1 \leq p \leq n-1$, all CV procedures rely on splitting a sample of cardinality n into two disjoint subsets called *training* and *test* sets with respective cardinality $n-p$ and p. The $n-p$ data in the training set serve to compute an estimator, while its performance is evaluated from the p left out data of the test set. This splitting scheme avoids overly optimistic performance estimations such as that of the re-substitution error which evaluates the performance of an estimator on the data used to compute it. For a complete and comprehensive review on cross-validation procedures, we refer the interested reader to [1].

Among CV procedures, two types can be distinguished: exhaustive and non-exhaustive ones. For $1 \leq p \leq n-1$, the *leave-p-out* cross-validation (LpO) is an instance of exhaustive procedure since it considers all possible splits of the sample into a training set of cardinality $n-p$ and a test set of cardinality p. Therefore LpO enjoys a minimal variance property among CV procedures with a test set of cardinality p. However due to the high induced computational cost, LpO cannot be computed in general, except for instance if $p = 1$ where it reduces to the celebrated *leave-one-out* (L1O). To avoid this high computation cost, approximations to the LpO procedure has been proposed such as the Hold-Out (HO) and the V-fold cross-validation (V-FCV) (with $p = \lfloor n/V \rfloor$). V-FCV is less computationally demanding than LpO but more variable since it depends on a preliminary random partition of the data into V disjoint subsets [1]. For instance in density estimation, [12] have quantified the additional variance of V-FCV in comparison to that of LpO. Recently a new interest has been given to LpO since it has been shown closed-form formulas can be derived for the LpO estimator in a wide range of settings, which makes LpO attractive from a statistical and also a computational point of view. For instance such closed-form formulas have been settled in density estimation [14, 10], in nonparametric regression with projection or kernel estimators [13], and change-point detection [2].

Despite the practical success of CV, only very few things are known on its theoretical properties in terms of model selection with respect to p. And also probably for technical reasons, existing results are mainly settled for HO and L10. For instance, the asymptotic equivalence between L1O and penalized criteria such as AIC or Mallow's C_p is settled by [35]. In the density estimation framework, [12, 14, 13] describe the behavior of LpO in terms of bias and variance, while [4, 10] respectively analyze the performance of $V-$FCV

and LpO for model selection. With a focus on regression, [9] considers V-FCV for which he settles the asymptotic dependence of the bias and variance with respect to V, and [40, 41] characterize the asymptotic behavior of CV for model selection in the linear regression framework. Note that in a general regression model, [46] recently studied the performance of CV to recover the best predictor among a finite number of candidates depending on p. The binary classification framework has also attracted some attention. First [30] addresses the HO model selection performance under assumptions on the VC dimension of the models, and then provides sanity-check bounds for the L1O [31]. Second, [47] proves consistency in selection for some CV procedures for identifying the best classifier among a finite number of candidates. A noticeable aspect in the CV literature is that unlike a common idea about the way p should be chosen "in practice", numerous settings have been reported where the ratio $p/n \to 1$ is required as n tends to $+\infty$ to be able to recover the best model in a collection for instance. This phenomenon, called "paradox of CV" by [47], had already been noticed by [40] in linear regression. More recently [46] (regression) and [10] (density estimation) have made this observation more precise by relating the optimal value of the ratio p/n to the convergence rate of the best candidate estimator. For more references, we refer to [1] and also to the recent paper by [49] for ongoing misleading ideas about CV.

The present work mainly focuses on the popular k-nearest neighbor rule (kNN) introduced by [22]. It is based on a simple idea that the predicted value at a new point is based on a majority vote among the k nearest neighbors of that point. Since it automatically adapts the scaling to low-density regions of the space, the kNN rule is particularly relevant in high dimensional settings where no preliminary partitioning of the space seems realistic. However from a theoretical point of view, there is no existing guideline that could help to choose the influential number k of neighbors in practice. In regression, [33] provide a bound on the performance of 1NN that has been further generalized to the kNN rule ($k \geq 1$) by [5], where a bagged version of the kNN rule is also analyzed and then applied to functional data [6]. In the present work, the main focus is given to the binary classification framework where preliminary theoretical results date back to [17, 16, 25]. More recently, [37, 34] derived an asymptotic equivalent to the performance of the 1NN classification rule, further extended to kNN by [42]. Under various distributional assumptions, [26] derived asymptotic expansions of the risk of the kNN classifier, which relates this risk to parameter k and other unknown distributional parameters. By contrast to previous results, [18, 15] settled finite sample upper bounds on the risk of the kNN classifier under mild assumptions.

Since these results cannot serve to provide a data-driven choice of k at this stage, this choice is usually made by HO or V-FCV in practice without any theoretical validation of the resulting choice [19, 27]. However, [43, 11] have recently derived closed-form formulas respectively for the bootstrap and the LpO estimator of the performance achieved by the

kNN classification rule. In particular such formulas for the LpO estimator allow to improve on the more variable V-fold estimator (with $p = \lfloor n/V \rfloor$), while the traditional question "Which p leads to the best k?" still remains an open problem.

The first step toward an answer is to understand the link between the moments of the LpO estimator and parameters p and k. Deriving upper bounds on these moments would then provide concentration inequalities [8, 7]. Some preliminary results in this direction are only available for the L1O ($p = 1$) estimator of the kNN classifier [20, 38, 21].

The connection between the LpO risk estimator and U-statistics is stated in Section 2. A first general result is provided for order q moments ($q \geq 2$) of the LpO estimator that are related to moments of the L1O estimator. This result applies to any classifier as long as the considered quantities remain well defined. Section 3 then specifies the upper bounds stated by the previous result in the case of the kNN classifier. This leads to the main Theorem 3.2 that characterizes the behavior of the LpO estimator with respect to p and k. In particular while the upper bounds increase with $1 \leq p \leq n/2 + 1$, it is no longer the case if $p > n/2 + 1$. Deriving exponential concentration inequalities for the LpO estimator is the main concern of Section 4. We illustrate the strength of our strategy based on $U-$statistics and moment inequalities by first providing concentration inequalities derived with less sophisticated tools. We then state our main inequality as a consequence of Theorem 3.2 and highlight the improvements it allows by comparison to the previous ones. Finally Section 5 briefly collects some new results that are extensions to LpO of previous ones originally stated for L1O. This section ends with a corollary assessing the magnitude of the gap between the LpO estimator and the risk of the kNN classifier with high probability.

2 U-statistics and LpO estimator

2.1 Statistical framework

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 $P_{(X,Y)}(A) = \mathbb{P}[(X, Y) \in A]$ for any Borelian set in $\mathcal{X} \times \{0, 1\}$, where $\mathbb{P}$ denotes a reference probability. To this end, one aims at building a classifier $f : \mathcal{X} \to \{0, 1\}$ that predicts $f(x)$ given $x \in \mathcal{X}$ with the best possible *classification error*

$$L(f) = \mathbb{P}(f(X) \neq Y).$$

The minimizer of the classification error over the set $\mathcal{F}$ all mesurable functions from $\mathcal{X}$ to $\{0, 1\}$ is known to be the Bayes classifier f^* defined for every $x \in \mathcal{X}$ by

$$f^*(x) = \mathbb{1}_{\{\eta(x) \geq 1/2\}}, \quad \text{with } \eta(x) = \mathbb{P}(Y = 1 \mid X = x), \tag{2.1}$$

where $\mathbb{1}_U(x)$ denotes the indicator function of the set U that is equals to 1 if $x \in U$ and 0 otherwise, and $\eta(\cdot)$ is the regression function of Y given $X = \cdot$.

Any classifier $\hat{f} \in \mathcal{F}$ results from a strategy $\mathcal{A}$, called *classification algorithm* or *classification rule*, applied from a set of training random variables $Z_{1,n} = \{Z_1, \ldots, Z_n\}$, where $Z_i = (X_i, Y_i)$ for every $1 \leq i \leq n$. In what follows, let us further define $Z^v = \{Z_i \mid i \in v\}$ for any subset $v \subset \{1, \ldots, n\}$ such that if $v = \{1, \ldots, n\}$, $Z^v = Z_{1,n}$. When applied to different samples, any classification rule $\mathcal{A} : \cup_{n \geq 1} \{\mathcal{X} \times \{0,1\}\}^n \to \mathcal{F}$, which maps a training sample $Z_{1,n}$ onto the corresponding classifier $\mathcal{A}(Z_{1,n}; \cdot) = \hat{f} \in \mathcal{F}$, leads to different classifiers. When no confusion is possible, the notation $\mathcal{A}$ will be used as a shortcut for $\mathcal{A}(Z_{1,n}; \cdot) \in \mathcal{F}$. Many classification rules have been considered in the literature and it is out of the scope of the present paper to review all of them (see [19, 27] for various instances). Here we mainly focus on the k-nearest neighbor rule (kNN) initially proposed by [22] and further studied for instance by [20, 38]. For $1 \leq k \leq n$, the kNN rule, denoted by $\mathcal{A}_k$, consists in classifying any new observation x using a majority vote decision rule based on the label of the k closest points $X_{(1)}(x), \ldots, X_{(k)}(x)$ to x among the training sample $X_1, \ldots, X_n$:

$$\mathcal{A}_k(Z_{1,n}; x) = \widehat{f}_k(Z_{1,n}; x) := \begin{cases} 1 & \text{if } \frac{1}{k}\sum_{i \in V_k(x)} Y_i = \frac{1}{k}\sum_{i=1}^{k} Y_{(i)}(x) > 0.5 \\ 0 & \text{otherwise} \end{cases} , \qquad (2.2)$$

where $V_k(x) = \left\{1 \leq i \leq n,\ X_i \in \left\{X_{(1)}(x), \ldots, X_{(k)}(x)\right\}\right\}$ denotes the set of indices of the k nearest neighbors of x among $X_1, \ldots, X_n$, and $Y_{(i)}(x)$ is the label i-th neighbor of x for $1 \leq i \leq k$.

For a given sample $Z_{1,n}$ (respectively for a given sample size $n \geq 1$), the performance of any classifier $\hat{f} = \hat{f}(Z_{1,n}; \cdot)$ is assessed by the classification error $L(\hat{f})$ (respectively the risk $R(\hat{f})$) defined by

$$L(\hat{f}) = \mathbb{P}\left(\hat{f}(X) \neq Y \mid Z_{1,n}\right) , \quad \text{and} \qquad R(\hat{f}) = \mathbb{E}\left[\mathbb{P}\left(\hat{f}(X) \neq Y \mid Z_{1,n}\right)\right].$$

In this paper we focus on the estimation of $L(\hat{f})$ (and $R(\hat{f})$) by use of the *Leave-p-Out* (LpO) cross-validation [48, 12]. Let us briefly recall what it consists in. LpO successively considers all possible splits of $Z_{1,n}$ into a training set of cardinality $n - p$ and a test set of cardinality p. The final LpO estimator is the average (over all these splits) of the classification error estimated on each test set:

$$\widehat{R}_p(\widehat{f}) = \binom{n}{p}^{-1} \sum_{e \in \mathcal{E}_{n-p}} \left(\frac{1}{p}\sum_{i \in \bar{e}} \mathbb{1}_{\{\hat{f}(Z^e; X_i) \neq Y_i\}}\right) , \qquad (2.3)$$

where $\mathcal{E}_{n-p}$ denotes the set of indices of all possible training samples of cardinality $n - p$, for every $e \in \mathcal{E}_{n-p}$, $\bar{e} = \mathcal{E}_{n-p} \setminus e$, and $\hat{f}(Z^e; \cdot)$ is the classifier built from Z^e. We refer the reader to [1] for a detailed description of LpO and other cross-validation procedures.

However unlike what arises from (2.3), the LpO estimator can be efficiently computed by use of closed-form formulas with a time complexity linear in p when applied to the kNN classification rule [11]. This property remains true in other contexts such as density estimation [12, 10], regression [13, 2], and so on. In particular this property contrasts with the usual prohibitive computational complexity suffered by LpO due to the high cardinality of $\mathcal{E}_{n-p}$ that is equal to $\binom{n}{p}$. LpO can be therefore used to assess the performance of any kNN classifier and to choose the best value of the parameter k.

2.2 U-statistics: General bounds on LpO moments

The purpose of the present section is describe a general strategy allowing to derive new upper bounds on the moments of the LpO estimator of the risk. As a first step of this strategy, we settle the connection between the LpO risk estimator and U-statistics. Second, we exploit this connection to derive new upper bounds on the moments of order $q > 1$ of the LpO estimator. In particular 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. For an extensive review, we refer to books by [39, 32]. 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 ([32])**.** *Let $h : \mathcal{X}^m \longrightarrow \mathbb{R}$ (or $\mathbb{R}^k$) denote any Borelian 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 [28])**.** *With the notation of Definition 2.1, let us define $W : \mathcal{X}^n \longrightarrow \mathbb{R}$ by*

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

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 of the paper result from this connection between the LpO estimator defined by Eq. (2.3) and U-statistics.

Theorem 2.1. *For any classification rule $\mathcal{A}$ (leading to the classifier $\mathcal{A}(Z_{1,n}:,\cdot) = \hat{f}$) and any $1 \leq p \leq n-1$ such that the following quantities are well defined, the LpO estimator $\widehat{R}_p(\hat{f})$ is a U-statistic of order $m = n-p+1$ with kernel $h_m : \mathcal{X}^m \longrightarrow \mathbb{R}$ defined by*

$$h_m(Z_1, \ldots, Z_m) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{\left\{\hat{f}(Z_{1,m}^{(i)}; X_i) \neq Y_i\right\}},$$

where $Z_{1,m}^{(i)}$ denotes the sample $(Z_1, \ldots, Z_m)$ with Z_i withdrawn.

Proof of Theorem 2.1.
In what follows, let $\mathcal{E}_t$ denote the set of all possible training sample indices e of cardinality $t \geq 1$. Let us start from Eq. (2.3) and notice that for every $e \in \mathbb{E}_{n-p}$, $\hat{f}(Z^e; X_i) = \hat{f}^{-i}(Z^e \cup Z_i; X_i)$, where $\hat{f}^{-i}(Z^e \cup Z_i; \cdot)$ is the classifier computed from $\{Z^e \cup Z_i\} \setminus Z_i$. Then, it comes

$$\begin{aligned}
\widehat{R}_p(\widehat{f}) &= \frac{1}{\binom{n}{p}} \sum_{e \in \mathcal{E}_{n-p}} \frac{1}{p} \sum_{i \in \bar{e}} \mathbb{1}_{\left\{\widehat{f}^{-i}(Z^e \cup Z_i; X_i) \neq Y_i\right\}} \\
&= \frac{1}{(n-p+1)\binom{n}{n-p+1}} \sum_{e' \in \mathcal{E}_{n-p+1}} \sum_{i \in e'} \mathbb{1}_{\left\{\widehat{f}^{-i}(Z^{e'}; X_i) \neq Y_i\right\}} \\
&= \frac{1}{\binom{n}{m}} \sum_{e' \in \mathcal{E}_m} \left(\frac{1}{m} \sum_{i \in e'} \mathbb{1}_{\left\{\widehat{f}^{-i}(Z^{e'}; X_i) \neq Y_i\right\}} \right) \\
&= \frac{1}{\binom{n}{m}} \sum_{1 \leq i_1 < \ldots < i_m \leq n} h_m\left(Z_{i_1}, \ldots, Z_{i_m}\right),
\end{aligned}$$

with

$$h_m\left(Z_{i_1}, \ldots, Z_{i_m}\right) = \frac{1}{m} \sum_{i \in e'} \mathbb{1}_{\left\{\widehat{f}^{-i}(Z^{e'}; X_i) \neq Y_i\right\}}.$$

□

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

$$h_m\left(Z_1, \ldots, Z_m\right) = \widehat{R}_1\left(\mathcal{A}\left(Z_{1,m}; \cdot\right)\right). \tag{2.4}$$

In the context of testing whether two binary classifiers have different error rates, this fact has already been pointed out by [23].

We now derive a general upper bound on the q-th moment ($q > 1$) of the LpO estimator that holds true for any classification rule. This results from the combination of Proposition 2.1 and Theorem 2.1.

Theorem 2.2. *For any classification rule $\mathcal{A}$, let $\widehat{g}_n = \mathcal{A}(Z_{1,n}; \cdot)$ and $\widehat{g}_m = \mathcal{A}(Z_{1,m}; \cdot)$ be the corresponding classifiers built from respectively $Z_1, \ldots, Z_n$ and $Z_1, \ldots, Z_m$, where $m = n - p + 1$. Then for every $1 \leq p \leq n-1$ such that the following quantities are well defined,*

$$\mathbb{E}\left[\left|\widehat{R}_p(\widehat{g}_n) - \mathbb{E}\left[\widehat{R}_p(\widehat{g}_n)\right]\right|^q\right] \leq \mathbb{E}\left[\left|\widehat{R}_1(\widehat{g}_m) - \mathbb{E}\left[\widehat{R}_1(\widehat{g}_m)\right]\right|^q\right]. \tag{2.5}$$

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

- *for $q = 2$*

$$\mathbb{E}\left[\left|\widehat{R}_p(\widehat{g}_n) - \mathbb{E}\left[\widehat{R}_p(\widehat{g}_n)\right]\right|^2\right] \leq \left\lfloor \frac{n}{m} \right\rfloor^{-1} \mathbb{E}\left[\left|\widehat{R}_1(\widehat{g}_m) - \mathbb{E}\left[\widehat{R}_1(\widehat{g}_m)\right]\right|^2\right]. \tag{2.6}$$

- *for every $q > 2$*

$$\mathbb{E}\left[\left|\widehat{R}_p(\widehat{g}_n) - \mathbb{E}\left[\widehat{R}_p(\widehat{g}_n)\right]\right|^q\right] \leq B(q,\gamma) \times \left\{ 2^q \gamma \left\lfloor \frac{n}{m} \right\rfloor \mathbb{E}\left[\left| \frac{\widehat{R}_1(\widehat{g}_m) - \mathbb{E}\left[\widehat{R}_1(\widehat{g}_m)\right]}{\left\lfloor \frac{n}{m} \right\rfloor} \right|^q\right] \vee \left(\sqrt{\frac{2\mathrm{Var}\left(\widehat{R}_1(\widehat{g}_m)\right)}{\left\lfloor \frac{n}{m} \right\rfloor}} \right)^q \right\} \tag{2.7}$$

  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. Theorem 2.2 relates the upper bounds on the moments of the LpO estimator to those of the L1O estimator. Eq. (2.6) and (2.7) emphasize different convergence rates for the moments of the LpO estimator that can be achieved in the particular setting where $p > n/2+1$. This point will be further discussed in Remark 2 (following Theorem 3.2) and illustrated 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 classification rule $\mathcal{A}_k$ ($1 \leq k \leq n$) introduced by (2.2).

Since Theorem 2.2 expresses moments of the LpO estimator in terms of those of the L1O estimator, the next step consists in focusing on the L1O moments. Deriving tight upper on the moments of the L1O is made by use of a generalization of the well-known Efron-Stein inequality (see Theorem D.1 for Efron-Stein's inequality and Theorem 15.5 in [7] for its generalization). For the sake of completeness, we first recall a corollary of this generalization that is proved in Section D.1.5 (see Corollary D.1).

Proposition 3.1. *Let $X_1, \ldots, X_n$ denote n independent random variables and $Z = f(X_1, \ldots, X_n)$, where $f : \mathbb{R}^n \to \mathbb{R}$ is any Borelian function. With $Z_i' = f(X_1, \ldots, X_i', \ldots, X_n)$, where $X_1', \ldots, X_n'$ are independent copies of the X_is, there exists a universal constant $\kappa \leq 1.271$ such that for any $q \geq 2$,*

$$\|Z - \mathbb{E}Z\|_q \leq \sqrt{2\kappa q} \sqrt{\left\| \sum_{i=1}^{n} (Z - Z_i')^2 \right\|_{q/2}}.$$

Then applying Proposition 3.1 to $Z = \widehat{R}_1(A_k(Z_{1,m}; \cdot))$ provides the following Theorem 3.1, which specifies the upper bound on the q-th moment of the L1O estimator. Its proof is detailed in Section A.2.

Theorem 3.1. *For every $1 \leq k \leq n-1$, let $\hat{f}_{k,m} = A_k(Z_{1,m}; \cdot)$ denote the kNN classifier learnt from $Z_{1,m}$ and $\widehat{R}_1(\hat{f}_{k,m})$ be the corresponding L1O risk estimator (see (2.3)) ($m = n - p + 1$). Then*

- *for $q = 2$,*

$$\mathbb{E}\left[\left(\widehat{R}_1(\hat{f}_{k,m}) - \mathbb{E}\left[\widehat{R}_1(\hat{f}_{k,m}) \right] \right)^2 \right] \leq C_1 \sqrt{k} \left(\frac{\sqrt{k}}{\sqrt{m}} \right)^2 ;$$

- *for every $q > 2$,*

$$\mathbb{E}\left[\left| \widehat{R}_1(\hat{f}_{k,m}) - \mathbb{E}\left[\widehat{R}_1(\hat{f}_{k,m}) \right] \right|^q \right] \leq (C_2 \sqrt{q})^q \left(\frac{k}{\sqrt{m}} \right)^q,$$

with $C_1 = 2 + 16\gamma_d$ and $C_2 = 4\gamma_d\sqrt{2\kappa}$, where γ_d denotes the constant arising from Stone's lemma (Lemma D.6) and κ is defined in Proposition 3.1.

We are now in position to state the main result of this section. It follows from the combination of Theorem 2.2 and Theorem 3.1 in a straightforward way.

Theorem 3.2. *For every $p, k \geq 1$ such that $p + k \leq n$, let $\widehat{R}_p(\hat{f}_k)$ denote the LpO risk estimator (see (2.3)) of the kNN classifier $\hat{f}_k = \mathcal{A}_k(Z_{1,n}; \cdot)$ defined by (2.2). Then there exist (known) constants $C_1, C_2 > 0$ such that for every $1 \leq p \leq n - k$,*

- *for* $q = 2$,

$$\mathbb{E}\left[\left(\widehat{R}_p(\hat{f}_k) - \mathbb{E}\left[\widehat{R}_p(\hat{f}_k)\right]\right)^2\right] \leq C_1 \left(\sqrt{\frac{k\sqrt{k}}{(n-p+1)}}\right)^2 \quad ; \tag{3.1}$$

- *for every* $q > 2$,

$$\mathbb{E}\left[\left|\widehat{R}_p(\hat{f}_k) - \mathbb{E}\left[\widehat{R}_p(\hat{f}_k)\right]\right|^q\right] \leq \left(C_2\sqrt{\tfrac{k^2}{n-p+1}}q^{1/2}\right)^q, \tag{3.2}$$

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.6). Furthermore in the particular setting where* $n/2 + 1 < p \leq n - k$, *then*

- *for* $q = 2$,

$$\mathbb{E}\left[\left(\widehat{R}_p(\hat{f}_k) - \mathbb{E}\left[\widehat{R}_p(\hat{f}_k)\right]\right)^2\right] \leq C_1 \left(\sqrt{\frac{k\sqrt{k}}{(n-p+1)\left\lfloor\frac{n}{n-p+1}\right\rfloor}}\right)^2, \tag{3.3}$$

- *for every* $q > 2$,

$$\begin{aligned}&\mathbb{E}\left[\left|\widehat{R}_p(\hat{f}_k) - \mathbb{E}\left[\widehat{R}_p(\hat{f}_k)\right]\right|^q\right] \\ \leq\ & \left\lfloor\frac{n}{n-p+1}\right\rfloor \Gamma^q \left(\sqrt{\frac{k\sqrt{k}}{(n-p+1)\left\lfloor\frac{n}{n-p+1}\right\rfloor}}q^{1/2} \vee \sqrt{\frac{k^2}{(n-p+1)\left\lfloor\frac{n}{n-p+1}\right\rfloor^2}}q^{3/2}\right)^q\end{aligned} \tag{3.4}$$

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

The proof is detailed in Section A.3.

Remark 1. *Eq. (3.4) results from the version of Rosenthal's inequality derived for symmetric random variables by [29]. In this inequality the optimal constants depend on a parameter* $\gamma > 0$ *to be tuned. It has been calibrated to provide tight upper bounds in our setting (see Propositions D.2 and D.3). Note that the dependence of (3.4) with respect to* q *has been proved to be non improvable in [36].*

Remark 2. *Eq. (3.3) and (3.4) focus on the setting where* $n/2 + 1 < p \leq n - k$. *To stress the interest of these bounds, let us consider the case where* k *and* $n - p$ *are kept fixed as* n *increases. In this context Eq. (3.1) and (3.2) provide non informative upper bounds, whereas Eq. (3.3) and (3.4) lead to respective convergence rates at worse* $1/n$ *and* $(1/n)^{q/2-1}$, *for* $q > 2$. *The previous example is a particular instance of the more general setting where* $p/n \to 1$ *as* n *tends to* $+\infty$, *that has been investigated in various contexts by [40, 47, 46, 10].*

4 Exponential concentration inequalities

In this section, we provide exponential concentration inequalities for the LpO estimator of the risk when the kNN classification rule is used. The main inequalities we provide at the end of this section heavily rely on the moments inequalities previously derived in Section 3 (namely Theorem 3.2). In order to highlight the interest of our approach based on moment inequalities, we start this section by stating two exponential inequalities obtained with less sophisticated tools. For each of them, we discuss its strength and weakness to justify the refinements we further explore step by step.

A first exponential concentration inequality for $\widehat{R}_p(\hat{f}_k)$ can be derived by use of the bounded difference inequality following the line of proof of [19] originally developed for the L1O estimator. Its proof is given in Appendix B.1.

Proposition 4.1. *For any integers $p, k \geq 1$ such that $p+k \leq n$, let $\widehat{R}_p(\hat{f}_k)$ denote the LpO estimator (2.3) of the classification error of the kNN classifier $\hat{f}_k = \mathcal{A}_k(Z_{1,n}; \cdot)$ defined by (2.2). Then for every $t > 0$,*

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

where γ_d denotes the constant introduced in Stone's lemma (Lemma D.6).

The upper bound (4.1) obtained for the difference strongly relies on the facts that: (i) for X_j to be one of the k neighbors of X_i in at least one subsample X^e, it requires X_j to be one of the $k+p-1$ neighbors of X_i in the complete sample, and (ii) the number of points for which X_j may be one of the $k+p-1$ neighbors is bounded by Lemma D.6. This reasoning results in a rough upper bound since one does not distinguish between points for which X_j is among the k first neighbors or above the k-th one, whereas these are strongly different situations in practice. Subsequently the dependence of the convergence rate on k and p in Proposition 4.1 is not optimal, as confirmed by 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 of a given point in the complete sample leads to the next result.

Theorem 4.1. *For every $p, k \geq 1$ such that $p + k \leq n$, let $\widehat{R}_p(\hat{f}_k)$ denote the LpO estimator (2.3) of the classification error of the kNN classifier $\hat{f}_k = \mathcal{A}_k(Z_{1,n}; \cdot)$ defined by (2.2). Then there exists a numeric constant $\square > 0$ such that for every $t > 0$,*

$$\mathbb{P}\left(\widehat{R}_p(\hat{f}_k) - \mathbb{E}\left(\widehat{R}_p(\hat{f}_k)\right) > t\right) \vee \mathbb{P}\left(\mathbb{E}\left(\widehat{R}_p(\hat{f}_k)\right) - \widehat{R}_p(\hat{f}_k) > t\right) \leq \exp\left(-\frac{nt^2}{\square k^2\left[1 + (k+p)\frac{p-1}{n-1}\right]}\right),$$

with $\square = 1024e\kappa(1+\gamma_d)$, where γ_d is introduced in Lemma D.6 and $\kappa \leq 1.271$ is a universal constant.

The proof is given in Section B.2. Note that with $p = 1$ one recovers (up to constants) the bound derived by [19] for L1O. Compared with Proposition 4.1, we still have a k^2 in the denominator but also a $(k+p) \times (p-1)/(n-1)$ term. The latter vanishes as long as $(k \vee p)p/n = o(1)$, which makes this bound tighter than the previous one.

However the upper bound of Theorem 4.1 does not reflect the same dependencies with respect to k and p as what has been proved for polynomial moments in Theorem 3.2. In particular, we do not observe the concentration improvement allowed with high order moments as p is chosen large enough with $p > n/2 + 1$ (see also the remarks following Theorem 3.2). This drawback is overcome by the following upper bounds.

Theorem 4.2. *For every $p, k \geq 1$ such that $p+k \leq n$, let $\widehat{R}_p(\hat{f}_k)$ denote the LpO estimator of the classification error of the kNN classifier $\hat{f}_k = \mathcal{A}_k(Z_{1,n}; \cdot)$ defined by (2.2). Then for every $t > 0$,*

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

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

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

where Γ arises in Eq. (3.4) and γ_d denotes the constant introduced in Stone's lemma (Lemma D.6).

The proof has been postponed to Appendix B.3. It is based on the combination of Theorem 3.2 and Lemma D.3 and Proposition D.1 respectively to derive Eq. (4.2) and (4.3). Note that Theorem 4.2 provides a strict improvement upon Theorem 4.1. Unlike the upper bound provided in the latter theorem, Eq. (4.2) remains meaningful as long as $p/n \to \delta \in [0, 1[$ as n tends to $+\infty$, while (4.3) deals with the case where $\delta = 1$.

In order to allow a better interpretation of the last inequality (4.3), we also provide 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. *For any $p, k \geq 1$ such that $p + k \leq n$, $p > n/2 + 1$, and for every $t > 0$*

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

where $\Gamma > 0$ is the constant arising from (3.4).

Let us notice that for every fixed k, both deviation terms are of order $1/\sqrt{n}$ as long as $p/n \to \delta \in [0, 1[$ since

$$(n-p+1)\left\lfloor\frac{n}{n-p+1}\right\rfloor = n(1+o(1)) = (n-p+1)\left\lfloor\frac{n}{n-p+1}\right\rfloor^2$$

as n tends to $+\infty$. However if $p/n \to 1$ as n tends to $+\infty$, by setting $r_n = 1 - p/n = o(1)$, it results the first deviation term remains of order $1/\sqrt{n}$ while the order of the second one becomes $\sqrt{r_n/n} = o(1/\sqrt{n})$ as n tends to $+\infty$. Note also that the dependence of the first (sub-Gaussian) deviation term with respect to k is only $k\sqrt{k}$, which improves upon the k^2 provided by usual results such as Ineq. (4.2) in Theorem 4.2.

5 Assessing the gap between LpO and prediction error

In the present section, we derive new upper bounds on different measures of the discrepancy between $\widehat{R}_p(\hat{f}_k)$ and $L(\hat{f}_k)$. These bounds on the LpO estimator are completely new for $1 < p \leq n - 1$. Some of them are extensions of former ones specifically derived for the L1O risk estimator of the kNN classifier.

Following the same line of proof as Theorem 2.1 in [38] originally developed for the L1O estimator, we were able to upper bound the mean of the (squared) difference between $\widehat{R}_p(\hat{f}_k)$ and $L(\hat{f}_k)$. These bounds reflect the reliable dependence of this error in terms of influential quantities such as p, k, and n.

Theorem 5.1. *For every $p, k \geq 1$ such that $p + k \leq n$, let $\widehat{R}_p(\hat{f}_k)$ denote the LpO risk estimator (see (2.3)) of the kNN classifier $\hat{f}_k$ defined by (2.2). Then,*

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

Moreover,

$$\mathbb{E}\left[\left(\widehat{R}_p(\hat{f}_k) - L(\hat{f}_k)\right)^2\right] \leq \frac{2\sqrt{2}}{\sqrt{\pi}}\frac{(2p+3)\sqrt{k}}{n} + \frac{1}{n} \cdot \tag{5.2}$$

Proof of Ineq. (5.1).
Lemma D.7 immediately provides

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

□

The proof of Ineq. (5.2) is more intricate and has been postponed to Appendix C.1. First, Ineq. (5.1) can be interpreted as an upper bound on the bias between the classification error of the kNN classifier computed from respectively n and $n-p$ observations. Therefore, the fact that this upper bound increases with p seems reliable since the two classification errors becomes more and more different from one another as p grows. Note also that Ineq. (5.1) combined with Jensen's inequality leads to a less accurate upper bound than (5.2). Second, let us notice we recover (up to constants) the original bound derived for the L1O estimator with $p = 1$. However we have no idea whether the precise dependence on p and k is optimal or not. Finally Ineq. (5.1) entails the LpO estimator $\widehat{R}_p(\hat{f}_k)$ of $L(\hat{f}_k)$ is consistent as long as $p\sqrt{k}/n = o(1)$, which is in accordance with the traditional consistency assumption on the kNN classification rule, that is $k/n \to 0$ as n tends to $+\infty$ (see [19], Chap. 6.6 for instance).

Let us conclude this section with a corollary, which provides a finite-sample bound on the gap between $\widehat{R}_p(\hat{f}_k)$ and $R(\hat{f}_k) = \mathbb{E}\left[L(\hat{f}_k)\right]$ with high probability. It relies on the combination of the exponential concentration result we derived for $\widehat{R}_p(\hat{f}_k)$ (Theorem 4.2) with our upper bound on the bias (5.1).

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

$$\left|R(\hat{f}_k)) - \widehat{R}_p(\hat{f}_k)\right| \leq \sqrt{\frac{\Delta^2 k^2}{n\left(1 - \frac{p-1}{n}\right)}x} + \frac{4}{\sqrt{2\pi}}\frac{p\sqrt{k}}{n} \ . \tag{5.3}$$

Let us observe that if k is kept fixed (independent of n), making this gap decrease to 0 is possible following our upper bound by requiring $p/n \to 0$ as n tends to $+\infty$. More precisely it allows to achieve a convergence rate of $1/\sqrt{n}$ as long as $p = O(\sqrt{n})$.

Proof of Corollary 5.1. Ineq. (5.3) results from combining Ineq. (4.2) (from Theorem 4.2) and Ineq. (5.1).

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

□

References

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

[2] S. Arlot and A. Celisse. Segmentation of the mean of heteroscedastic data via cross-validation. *Statistics and Computing*, 21(4):613–632, 2011.

[3] Sylvain Arlot. *Resampling and Model Selection*. PhD thesis, University Paris-Sud 11, December 2007. oai:tel.archives-ouvertes.fr:tel-00198803_v1.

[4] Sylvain Arlot and Matthieu Lerasle. Why v= 5 is enough in v-fold cross-validation. *arXiv preprint arXiv:1210.5830*, 2012.

[5] Gérard Biau, Frédéric Cérou, and Arnaud Guyader. On the rate of convergence of the bagged nearest neighbor estimate. *The Journal of Machine Learning Research*, 11:687–712, 2010.

[6] Gérard Biau, Frédéric Cérou, and Arnaud Guyader. Rates of convergence of the functional-nearest neighbor estimate. *Information Theory, IEEE Transactions on*, 56(4):2034–2040, 2010.

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

[8] Stéphane Boucheron, Olivier Bousquet, Gábor Lugosi, and Pascal Massart. Moment inequalities for functions of independent random variables. *Ann. Probab.*, 33(2):514–560, 2005.

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

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

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

[12] 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.

[13] Alain 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.

[14] Alain Celisse and Stéphane Robin. A cross-validation based estimation of the proportion of true null hypotheses. *Journal of Statistical Planning and Inference*, 140(11):3132–3147, 2010.

[15] Kamalika Chaudhuri and Sanjoy Dasgupta. Rates of convergence for nearest neighbor classification. In *Advances in Neural Information Processing Systems*, pages 3437–3445, 2014.

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

[17] Thomas M Cover and Peter E Hart. Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13(1):21–27, 1967.

[18] Sanjoy Dasgupta. Consistency of nearest neighbor classification under selective sampling. In *COLT*, pages 18–1, 2012.

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

[20] Luc P. Devroye and Terry J. Wagner. The strong uniform consistency of nearest neighbor density estimates. *Ann. Statist.*, 5(3):536–540, 1977.

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

[22] 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.

[23] 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.

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

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

[26] Peter Hall, Byeong U Park, and Richard J Samworth. Choice of neighbor order in nearest-neighbor classification. *The Annals of Statistics*, pages 2135–2152, 2008.

[27] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*. Springer Series in Statistics. Springer-Verlag, New York, 2001. Data mining, inference, and prediction.

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

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

[30] M. Kearns. A Bound on the Error of Cross Validation Using the Approximation and Estimation Rates, with Consequences for the Training-Test Split. *Neural Comput.*, 9(5):1143–1161, 1997.

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

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

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

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

[35] K.-C. Li. Asymptotic Optimality for C_p, C_L, Cross-Validation and Generalized Cross-Validation: Discrete Index Set. *The Annals of Statistics*, 15(3):958–975, 1987.

[36] IF Pinelis and SA Utev. Estimates of the moments of sums of independent random variables. *Theory of Probability & Its Applications*, 29(3):574–577, 1985.

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

[38] 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.

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

[40] Jun Shao. Linear model selection by cross-validation. *J. Amer. Statist. Assoc.*, 88(422):486–494, 1993.

[41] Jun Shao. An asymptotic theory for linear model selection. *Statist. Sinica*, 7(2):221–264, 1997. With comments and a rejoinder by the author.

[42] Robert R Snapp, Santosh S Venkatesh, et al. Asymptotic expansions of the k nearest neighbor risk. *The Annals of Statistics*, 26(3):850–878, 1998.

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

[44] Charles J. Stone. Optimal global rates of convergence for nonparametric regression. *Ann. Statist.*, 10(4):1040–1053, 1982.

[45] M. Stone. Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B*, 36:111–147, 1974. 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.

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

[47] Yuhong Yang. Comparing learning methods for classification. *Statist. Sinica*, 16(2):635–657, 2006.

[48] Ping Zhang. Model selection via multifold cross validation. *Ann. Statist.*, 21(1):299–313, 1993.

[49] Yongli Zhang and Yuhong Yang. Cross-validation for selecting a model selection procedure. *Journal of Econometrics*, 187(1):95–112, 2015.

A Proofs of polynomial moment upper bounds

A.1 Proof of Theorem 2.2

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

$$\widehat{R}_p(\hat{g}_n) = \frac{1}{n!}\sum_{\sigma} W\left(Z_{\sigma(1)},\ldots,Z_{\sigma(n)}\right) \ ,$$

with

$$W\left(Z_1,\ldots,Z_n\right) = \left\lfloor \frac{n}{m} \right\rfloor^{-1} \sum_{a=1}^{\lfloor \frac{n}{m} \rfloor} h_m\left(Z_{(a-1)m+1},\ldots,Z_{am}\right) \qquad (\text{with } m = n-p+1)$$

$$\text{and} \quad h_m\left(Z_1,\ldots,Z_m\right) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{\left\{\mathcal{A}(Z_{1,m}^{(i)};X_i)\neq Y_i\right\}} = \widehat{R}_1(\widehat{g}_{n-p+1}) \ ,$$

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

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

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

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

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

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

which leads to the result.

If $q > 2$ and $p \leq n/2 + 1$, then a straightforward use of Jensen's inequality from (A.1) provides

$$\begin{aligned}\mathbb{E}\left[\left|\widehat{R}_p(\hat{g}_n) - \mathbb{E}\left[\widehat{R}_p(\hat{g}_n)\right]\right|^q\right] &\leq \left\lfloor \frac{n}{m} \right\rfloor^{-1} \sum_{i=1}^{\lfloor \frac{n}{m} \rfloor} \mathbb{E}\left[\left|\bar{h}_m\left(Z_{(i-1)m+1}, \ldots, Z_{im}\right)\right|^q\right] \\ &= \mathbb{E}\left[\left|\widehat{R}_1(\widehat{g}_{n-p+1}) - \mathbb{E}\left[\,\widehat{R}_1(\widehat{g}_{n-p+1})\,\right]\right|^q\right].\end{aligned}$$

If $q > 2$ and $p > n/2 + 1$, let us 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

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

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

$$\mathbb{E}\left[\left|\sum_{i=1}^{\lfloor \frac{n}{m} \rfloor} \bar{h}_m\left(Z_{(i-1)m+1},\ldots,Z_{im}\right)\right|^q\right]$$
$$\leq B(q,\gamma)\left\{2^q\gamma\left\lfloor\frac{n}{m}\right\rfloor \mathbb{E}\left[\left|\widehat{R}_1(\widehat{g}_{n-p+1}) - \mathbb{E}\left[\,\widehat{R}_1(\widehat{g}_{n-p+1})\,\right]\right|^q\right] \vee \left(\sqrt{\left\lfloor\frac{n}{m}\right\rfloor 2\mathrm{Var}\left(\widehat{R}_1(\widehat{g}_{n-p+1})\right)}\right)^q\right\}.$$

Hence, it results for every $q > 2$

$$\mathbb{E}\left[\left|\widehat{R}_p(\hat{g}_n) - \mathbb{E}\left[\widehat{R}_p(\hat{g}_n)\right]\right|^q\right]$$
$$\leq B(q,\gamma)\left\{2^q\gamma\left\lfloor\frac{n}{m}\right\rfloor^{-q+1} \mathbb{E}\left[\left|\widehat{R}_1(\widehat{g}_{n-p+1}) - \mathbb{E}\left[\,\widehat{R}_1(\widehat{g}_{n-p+1})\,\right]\right|^q\right] \vee \left\lfloor\frac{n}{m}\right\rfloor^{-q/2}\left(\sqrt{2\mathrm{Var}\left(\widehat{R}_1(\widehat{g}_{n-p+1})\right)}\right)^q\right\},$$

which concludes the proof.

A.2 Proof of Theorem 3.1

From Theorem 2.2 and Eq. (2.4), let us observe it is enough to upper bound $\mathbb{E}\left[\left|\bar{h}_m\left(Z_1,\ldots,Z_m\right)\right|^q\right]$. Then, Proposition 3.1 provides for every $q \geq 2$

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

The j-th term in the above sum is now upper bounded by

$$\begin{aligned}\left|h_m(Z_1,\ldots,Z_m) - h_m(Z_1,\ldots,Z'_j,\ldots,Z_m)\right| &\leq \frac{1}{m}\sum_{i=1}^m \left|\mathbb{1}_{\left\{f(Z^{(i)};X_i)\neq Y_i\right\}} - \mathbb{1}_{\left\{f(Z'^{,(i)};X'_i)\neq Y'_i\right\}}\right| \\ &\leq \frac{1}{m} + \frac{1}{m}\sum_{i\neq j}\left|\mathbb{1}_{\left\{f(Z^{(i)};X_i)\neq Y_i\right\}} - \mathbb{1}_{\left\{f(Z'^{,(i)};X'_i)\neq Y'_i\right\}}\right| \\ &\leq \frac{1}{m} + \frac{1}{m}\sum_{i\neq j}\left|\mathbb{1}_{\left\{f(Z^{(i)};X_i)\neq f(Z'^{,(i)};X'_i)\right\}}\right|,\end{aligned} \tag{A.2}$$

where $Z^{(i)} = Z_{1,n}^{(i)}$ and $Z'^{,(i)} = Z_{1,n}'^{,(i)}$ with the notation of Theorem 2.1.

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

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

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

$$\left|h_m(Z_1, \ldots, Z_m) - h_m(Z_1, \ldots, Z'_j, \ldots, Z_m)\right| \leq \frac{1}{m} + \frac{1}{m} \sum_{i \in B_j} \left| \mathbb{1}_{\left\{f(Z^{(i)}; X_i) \neq f(Z'^{,(i)}; X'_i)\right\}} \right|. \tag{A.3}$$

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.2.1 Case $q > 2$

From (A.3), Stone's lemma (Lemma D.6) provides

$$\begin{aligned}\left|h_m(Z_{1,m}) - h_m(Z'^{,j}_{1,m})\right| &\leq \frac{1}{m} + \frac{1}{m} \sum_{i \in B_j} \mathbb{1}_{\left\{f(Z^{(i)}; X_i) \neq f(Z'^{,(i)}; X_i)\right\}} \\ &\leq \frac{1}{m} + \frac{2k\gamma_d}{m} \ ,\end{aligned}$$

where $Z'^{,j}_{1,m} = (Z_1, \ldots, Z_{j-1}, Z'_j, Z_{j+1}, \ldots, Z_m)$.

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

$$\sum_j \left(h_m(Z_{1,m}) - h_m(Z'^{,j}_{1,m})\right)^2 \leq \frac{2}{m}\left(1 + (2k\gamma_d)^2\right) \leq \frac{4}{m}(2k\gamma_d)^2,$$

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.2.2 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

$$\mathbb{E}\left[\left(h_m(Z_{1,m}) - h_m(Z'^{,j}_{1,m})\right)^2\right] = \frac{2}{m^2} + \frac{2}{m^2}\mathbb{E}\left[\left(\sum_{i\in B_j} \mathbb{1}_{\left\{f(Z^{(i)};X_i)\neq f(Z'^{,(i)};X_i)\right\}}\right)^2\right]$$
$$\leq \frac{2}{m^2} + \frac{2}{m^2}\mathbb{E}\left[|B_j| \sum_{i\in B_j} \mathbb{1}_{\left\{f(Z^{(i)};X_i)\neq f(Z'^{,(i)};X_i)\right\}}\right]$$

using Jensen's inequality. Lemma D.6 implies $|B_j| \leq 2k\gamma_d$, which allows to conclude

$$\mathbb{E}\left[\left(h_m(Z_{1,m}) - h_m(Z'^{,j}_{1,m})\right)^2\right] \leq \frac{2}{m^2} + \frac{4k\gamma_d}{m^2}\mathbb{E}\left[\sum_{i\in B_j} \mathbb{1}_{\left\{f(Z^{(i)};X_i)\neq f(Z'^{,(i)};X_i)\right\}}\right].$$

Summing over j, one derives

$$\sum_{j=1}^{m} \mathbb{E}\left[\left(h_m(Z_1,\ldots,Z_m) - h_m(Z_1,\ldots,Z'_j,\ldots,Z_m)\right)^2\right]$$
$$\leq \frac{2}{m} + \frac{4k\gamma_d}{m^2}\sum_{j=1}^{m}\mathbb{E}\left[\sum_{i\in B_j} \mathbb{1}_{\left\{f(Z^{(i)};X_i)\neq f(Z'^{(i)};X_i)\right\}}\right] = \frac{2}{m} + \frac{4k\gamma_d}{m}\mathbb{E}\left[\sum_{i\in B_j} \mathbb{1}_{\left\{f(Z^{(i)};X_i)\neq f(Z'^{(i)};X_i)\right\}}\right]$$
$$\leq \frac{2}{m} + \frac{4k\gamma_d}{m}\sum_{i=1}^{m}\mathbb{E}\left[\mathbb{1}_{\left\{f(Z^{(i)};X_i)\neq f(Z^{(i)},Z_0;X_i)\right\}} + \mathbb{1}_{\left\{f(Z'^{(i)},Z_0;X_i)\neq f(Z'^{(i)};X_i)\right\}}\right]$$
$$\leq \frac{2}{m} + 4k\gamma_d \times 2\frac{4\sqrt{k}}{\sqrt{2\pi}m} = \frac{2}{m} + \frac{32\gamma_d}{\sqrt{2\pi}}\frac{k\sqrt{k}}{m} \leq (2 + 16\gamma_d)\frac{k\sqrt{k}}{m}\ , \tag{A.4}$$

where Z_0 is an independent copy of Z_1, and the last but one inequality results from Lemma D.7.

A.3 Proof of Theorem 3.2

Proofs of Ineq. (3.1), (3.2), and (3.3) straightforwardly result from the combination of Theorem 3.1 and Ineq. (2.5) and (2.6) from Theorem 2.2.

The proof of Ineq. (3.4) results from the upper bounds settled 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

$$\mathbb{E}\left[\left|\widehat{R}_p(\hat{f}_k) - \mathbb{E}\left[\widehat{R}_p(\hat{f}_k)\right]\right|^q\right] \leq \left(2\sqrt{2e}\right)^q \times$$
$$\left\{(\sqrt{q})^q \left(\sqrt{\left\lfloor \frac{n}{n-p+1}\right\rfloor^{-1} 2C_1\sqrt{k}\left(\frac{\sqrt{k}}{\sqrt{n-p+1}}\right)^2}\right)^q \vee q^q \left\lfloor \frac{n}{n-p+1}\right\rfloor^{-q+1} (2C_2\sqrt{q})^q \left(\frac{k}{\sqrt{n-p+1}}\right)^q\right\}$$
$$= \left(2\sqrt{2e}\right)^q \times$$
$$\left\{(\sqrt{q})^q \left(\sqrt{2C_1\sqrt{k}}\sqrt{\frac{k}{(n-p+1)\left\lfloor \frac{n}{n-p+1}\right\rfloor}}\right)^q \vee \left(q^{3/2}\right)^q \left\lfloor \frac{n}{n-p+1}\right\rfloor \left(2C_2 \frac{k}{\left\lfloor \frac{n}{n-p+1}\right\rfloor \sqrt{n-p+1}}\right)^q\right\}$$
$$\leq \left\lfloor \frac{n}{n-p+1}\right\rfloor \left\{\left(\lambda_1 q^{1/2}\right)^q \vee \left(\lambda_2 q^{3/2}\right)^q\right\},$$

with

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

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

B Proofs of exponential concentration inequalities

B.1 Proof of Proposition 4.1

Here the strategy is to use McDiarmid's inequality (Theorem D.3). This justifies to start by upper bounding the following difference $\left|\widehat{R}_p(\mathcal{A}_k\left(Z_{1,n};\cdot\right)) - \widehat{R}_p(\mathcal{A}_k(Z'^{,j}_{1,n};\cdot))\right|$ for every $1 \leq j \leq n$, where $Z'^{,j}_{1,n} = (Z_1, \ldots, Z_{j-1}, Z'_j, Z_{j+1}, \ldots, Z_n)$.

Then using Eq. (2.3), one has

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

where $Z'^{,j,e}$ denotes the set of random variables among $Z'^{,j}_{1,n}$ having indices in e, and $V_k^e(X_i)$ (resp. $V_k'^{,j,e}(X_i)$) denotes the set of indices of the k nearest neighbors of X_i among Z^e (resp. $Z'^{,j,e}$).

Setting $\mathcal{E}_{n-p} = \mathcal{E}$, let is now introduce

$$
B_j^{\mathcal{E}} = \bigcup_{e\in\mathcal{E}}\left\{1 \leq i \leq n,\ i \notin e \cup \{j\},\ V_k'^{,j,e}(X_i) \ni j \text{ or } V_k^e(X_i) \ni j\right\}.
$$

Then Lemma D.6 implies $\mathrm{Card}(B_j^{\mathcal{E}}) \leq 2(k+p-1)\gamma_d$, hence

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

Applying McDiarmid's inequality (Section D.1.6) then completes the proof.

B.2 Proof of Theorem 4.1

The first step of the proof consists in using Ineq. (D.5) (generalized Efron-Stein inequality) to upper bound the $2q$-th moments of

$$
\widehat{R}_p(\hat{f}_k) = \frac{1}{\binom{n}{p}}\sum_{e\in\mathcal{E}}\frac{1}{p}\sum_{i\in e}\mathbb{1}_{\{\hat{f}_k(Z^e;X_i)\neq Y_i\}}.
$$

With the same notation as in the proof of Proposition 4.1, one gets for every $1 \leq j \leq n$

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

Absolute values and Jensen's inequality then provide

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

where $\mathbb{P}_e$ denotes the discrete uniform probability over the set $\mathcal{E}_{n-p}$ of all $n-p$ distinct indices among $\{1,\ldots,n\}$.

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

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

which leads to

$$\left|\widehat{R}_p(\mathcal{A}_k(Z_{1,n};\cdot)) - \widehat{R}_p(\mathcal{A}_k(Z'^{,j}_{1,n};\cdot))\right| \leq \frac{1}{n} + \frac{2}{p} \sum_{i=1}^n \mathbb{P}_e \left[j\in e,\ i\in\bar{e},\ j\in V_k^e(X_i) \right].$$

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

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

Further using that

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

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

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

Then Lemma D.5 leads to

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

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

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

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,

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

We then apply Stone's lemma (Lemma D.6) to get

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

Gathering the upper bounds The two previous bounds provide

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

which enables to conclude

$$\begin{aligned}&\textstyle\sum_{j=1}^{n}\left(\widehat{R}_p(\mathcal{A}_k(Z_{1,n};\cdot))-\widehat{R}_p(\mathcal{A}_k(Z'^{,j}_{1,n};\cdot))\right)^2\\ &\le \tfrac{2}{n}\left(1+4k+4k^2\gamma_d\left[1+(k+p)\tfrac{p-1}{n-1}\right]\right)\le \tfrac{8k^2(1+\gamma_d)}{n}\left[1+(k+p)\tfrac{p-1}{n-1}\right].\end{aligned}$$

Then (D.5) provides for every $q\ge 1$

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

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

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

The conclusion follows from Lemma D.3 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 $t>0$,

$$\begin{aligned}&\mathbb{P}\left(\widehat{R}_p(\hat{f}_k)-\mathbb{E}\left(\widehat{R}_p(\hat{f}_k)\right)>t\right)\vee\mathbb{P}\left(\mathbb{E}\left(\widehat{R}_p(\hat{f}_k)\right)-\widehat{R}_p(\hat{f}_k)>t\right)\\ &\le\exp\left(-\frac{nt^2}{1024e\kappa k^2(1+\gamma_d)\left[1+(k+p)\frac{p-1}{n-1}\right]}\right).\end{aligned}$$

B.3 Proof of Theorem 4.2 and Proposition 4.2

Proof of Theorem 4.2.

If $p<n/2+1$:
From (3.1) and (3.2) 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 $q\ge 1$

$$\mathbb{E}\left[\left|\widehat{R}_p(\hat{f}_k)-\mathbb{E}\left[\widehat{R}_p(\hat{f}_k)\right]\right|^{2q}\right]\le\left(\frac{\Delta^2}{16e}\frac{k^2}{n-p+1}\right)^q(2q)^q\le\left(\frac{\Delta^2}{8}\frac{k^2}{n-p+1}\right)^q q!\ , \quad \text{(B.1)}$$

with $q^q \leq q! e^q/\sqrt{2\pi q}$. Then Lemma D.3 provides for every $t > 0$

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

If $p \geq n/2 + 1$:
Let us now use (3.1) and (3.4) 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 $t > 0$

$$\mathbb{P}\left[\left|\widehat{R}_p(\hat{f}_k) - \mathbb{E}\left[\widehat{R}_p(\hat{f}_k)\right]\right| > t\right] \leq \left\lfloor \frac{n}{n-p+1} \right\rfloor e \times$$
$$\exp\left[-\frac{1}{2e}\min\left\{(n-p+1)\left\lfloor \frac{n}{n-p+1}\right\rfloor \frac{t^2}{4\Gamma^2 k\sqrt{k}},\ \left((n-p+1)\left\lfloor \frac{n}{n-p+1}\right\rfloor^2 \frac{t^2}{4\Gamma^2 k^2}\right)^{1/3}\right\}\right],$$

where Γ arises in Eq. (3.4).

□

Proof of Proposition 4.2. With the same notation and reasoning as in the previous proof, let us combine (3.1) and (3.4). 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 $t > 0$

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

where $\Gamma > 0$ is given by Eq. (3.4).

□

C Proofs of deviation upper bounds

C.1 Proof of Ineq. (5.2) in Theorem 5.1

The proof follows the same strategy as that of Theorem 2.1 in [38].

Along the proof, we will repeatedly use some notation that we briefly introduce here. First, let us introduce $Z_0 = (X_0, Y_0)$ 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: $\widehat{f}_k(X_0) = \widehat{f}_k(Z_{1,n}; X_0)$, and $\widehat{f}_k(e, X_0) = \widehat{f}_k(Z^e_{1,n}; X_0)$ 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 random sets of distinct indices with discrete uniform distribution over $\mathcal{E}_{n-p}$. Therefore the notation $\mathbb{P}_e$ (resp. $\mathbb{P}_{e,e'}$) is used to emphasize that integration is made with respect to e (resp. to e, e').

C.1.1 Main part of the proof

Starting from

$$\mathbb{E}\left[(\widehat{R}_p(\hat{f}_k) - L(\hat{f}_k))^2\right] = \mathbb{E}\left[\widehat{R}_p^2(\hat{f}_k)\right] + \mathbb{E}\left[L_n^2\right] - 2\mathbb{E}\left[\widehat{R}_p(\hat{f}_k)L(\hat{f}_k)\right],$$

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

$$\mathbb{E}\left[\widehat{R}_p(\hat{f}_k)L(\hat{f}_k)\right] = \mathbb{P}\left(\widehat{f}_k(X_0) \neq Y_0, \widehat{f}_k(e, X_i) \neq Y_i |\ i \notin e\right)\mathbb{P}_e\left(i \notin e\right).$$

It immediately comes

$$\begin{aligned}
&\mathbb{E}\left[(\widehat{R}_p(\hat{f}_k) - L(\hat{f}_k))^2\right] \\
&= \left\{\mathbb{E}\left[\widehat{R}_p^2(\hat{f}_k)\right] - \mathbb{P}\left(\widehat{f}_k(X_0) \neq Y_0, \widehat{f}_k(e, X_i) \neq Y_i \mid\ i \notin e\right)\mathbb{P}_e\left(i \notin e\right)\right\} \qquad \text{(C.1)} \\
&+ \left\{\mathbb{P}\left(\widehat{f}_k(X_0) \neq Y_0, \widehat{f}_k(X_{n+1}) \neq Y_{n+1}\right) - \mathbb{P}\left(\widehat{f}_k(X_0) \neq Y_0, \widehat{f}_k(e, X_i) \neq Y_i |\ i \notin e\right)\mathbb{P}_e\left(i \notin e\right)\right\}. \qquad \text{(C.2)}
\end{aligned}$$

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

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

Let us now introduce the following events.

$$\begin{aligned} A_{e,e',i} &= \{i\notin e,\ i\notin e'\}, \\ A^1_{e,e',i,j} &= \{i\notin e,\ j\notin e',\ i\notin e',\ j\notin e\}, \qquad A^2_{e,e',i,j} = \{i\notin e,\ j\notin e',\ i\notin e',\ j\in e\}, \\ A^3_{e,e',i,j} &= \{i\notin e,\ j\notin e',\ i\in e',\ j\notin e\}, \qquad A^4_{e,e',i,j} = \{i\notin e,\ j\notin e',\ i\in e',\ j\in e\}. \end{aligned}$$

Then,

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

Furthermore since

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

it comes

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

where

$$A = \Big[\mathbb{P}\left(\widehat{f}_k(e, X_1) \neq Y_1,\ \widehat{f}_k(e', X_1) \neq Y_1 \mid A_{e,e',1}\right) - \mathbb{P}\left(\widehat{f}_k(X_0) \neq Y_0, \widehat{f}_k(e, X_1) \neq Y_1 \mid A_{e,e',1}\right)\Big] \mathbb{P}_{e,e'}\left(A_{e,e',1}\right),$$

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

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

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

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

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

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

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

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

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

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

Combining the previous bounds and Lemma C.1 leads to

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

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

$$
\mathbb{E}\left[\widehat{R}^2_p(\hat{f}_k)\right]-\mathbb{P}\left(\widehat{f}_k(X_0)\neq Y_0,\widehat{f}_k(e,X_1)\neq Y_1\right)=\frac{n}{p^2}A+\frac{n(n-1)}{p^2}B\leq\frac{1}{n}+\frac{2\sqrt{2}}{\sqrt{\pi}}\frac{(p+2)\sqrt{k}}{n}\ .
$$

Upper bound of (C.2) First observe that

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

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

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

where we used Lemma D.7 again to obtain the last inequality.

Conclusion:

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

$$\mathbb{E}\left[\left(\widehat{R}_p(\hat{f}_k) - L(\hat{f}_k)\right)^2\right] \leq \frac{2\sqrt{2}}{\sqrt{\pi}} \frac{(2p+3)\sqrt{k}}{n} + \frac{1}{n} \ .$$

C.1.2 Combinatorial lemmas

All the lemmas of the present section are settled with the same notation as in the proof of Theorem 5.1 (see Section C.1.1).

Lemma C.1.

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

Proof of Lemma C.1.

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

□

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

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

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

$$\mathbb{P}\left(\widehat{f}_k(X_0) \neq Y_0, \widehat{f}_k(e,X_1) \neq Y_1 \mid A^\ell_{e,e',1,2}\right)$$

is

$$\mathbb{P}\left(\widehat{f}_k(X_0) \neq Y_0, \widehat{f}_k(e,X_1) \neq Y_1 \mid A^\ell_{e,e',1,2}, 0 \notin e, 0 \notin e'\right)\ .$$

Then one has

$$\begin{aligned}
&\mathbb{P}\left(\widehat{f}_k(X_0) \neq Y_0, \widehat{f}_k(e,X_1) \neq Y_1 \mid A^\ell_{e,e',1,2}\right) \\
=\ &\mathbb{P}\left(\widehat{f}_k^{(-2)}(X_2) \neq Y_2,\ \widehat{f}_k(e,X_1) \neq Y_1 \mid A^\ell_{e,e',1,2}, 0 \notin e, 0 \notin e'\right)
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{P}\left(\widehat{f}_k(e,X_1)\neq Y_1,\ \widehat{f}_k(e',X_2)\neq Y_2 \mid A^{\ell}_{e,e',1,2}\right)-\mathbb{P}\left(\widehat{f}_k(X_0)\neq Y_0,\widehat{f}_k(e,X_1)\neq Y_1 \mid A^{\ell}_{e,e',1,2}\right)\\
=\ &\mathbb{P}\left(\widehat{f}_k(e,X_1)\neq Y_1,\ \widehat{f}_k(e',X_2)\neq Y_2 \mid A^{\ell}_{e,e',1,2},0\notin e,0\notin e'\right)\\
&-\mathbb{P}\left(\widehat{f}_k^{(-2)}(X_2)\neq Y_2,\ \widehat{f}_k(e,X_1)\neq Y_1 \mid A^{\ell}_{e,e',1,2},0\notin e,0\notin e'\right)\\
\leq\ &\mathbb{P}\left(\left\{\widehat{f}_k(e,X_1)\neq Y_1\right\}\triangle\left\{\widehat{f}_k(e,X_1)\neq Y_1\right\} \mid A^{\ell}_{e,e',1,2},0\notin e,0\notin e'\right)\\
&+\mathbb{P}\left(\left\{\widehat{f}_k^{(-2)}(X_2)\neq Y_2\right\}\triangle\left\{\widehat{f}_k(e',X_2)\neq Y_2\right\} \mid A^{\ell}_{e,e',1,2},0\notin e,0\notin e'\right)\\
=\ &\mathbb{P}\left(\widehat{f}_k^{(-2)}(X_2)\neq \widehat{f}_k(e',X_2) \mid A^{\ell}_{e,e',1,2}\right)\leq \frac{4p\sqrt{k}}{\sqrt{2\pi}n}\ ,
\end{aligned}$$

by Lemma D.7.

□

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

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

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

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

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

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

□

D Technical results

D.1 Main inequalities

D.1.1 Hoeffding's lemma for finite populations

Lemma D.1. *Let $X_1, ..., X_n$ denote a random sample without replacement in a finite population of N elements with values $c_1, ..., c_N$. If $a < c_i < b$ for $i = 1, ..., N$ then*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq t\right) \leq 2\exp\left[-\frac{2nt^2}{(b-a)^2}\right]$$

where $\mu = \frac{1}{N}\sum_{i=1}^{N} c_i$.

This lemma is proved in [28].

D.1.2 From moment to exponential inequalities

Proposition D.1 (see also [3], Lemma 8.10)**.** *Let X denote a real valued random variable, and assume there exist $C > 0$, $\lambda_1, \ldots, \lambda_N > 0$, 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 $t > 0$,

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

Furthermore for every $x > 0$, it results

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

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 $\sum_{i=1}^N \lambda_i q^{\alpha_i} \leq N \max_i \{\lambda_i q^{\alpha_i}\}$ 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

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

$$\leq \mathbb{1}_{\tilde{q} \geq q_0} C e^{-(\min_i \alpha_i) \left[e^{-1} \min_j \left\{ \left(\frac{t}{N\lambda_j} \right)^{\frac{1}{\alpha_j}} \right\} \right]} + \mathbb{1}_{\tilde{q} < q_0},$$

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,

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

Hence,

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

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.1.3 Sub-Gaussian random variables

Lemma D.2 (Theorem 2.1 in [7] first part)**.** *Any centered random variable X such that $\mathbb{P}(X > t) \vee \mathbb{P}(-X > t) \leq e^{-t^2/(2\nu)}$ satisfies*

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

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

Lemma D.3 (Theorem 2.1 in [7] 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 $\mathbb{P}(X > t) \vee \mathbb{P}(-X > t) \leq e^{-t^2/(2\nu)}$ with $\nu = 4C$.

D.1.4 The Efron-Stein inequality

Theorem D.1 (Efron-Stein's inequality [7], Theorem 3.1)**.** *Let $X_1, \ldots, X_n$ be independent random variables and let $Z = f(X_1, \ldots, X_n)$ 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.1.5 Generalized Efron-Stein's inequality

Theorem D.2 (Theorem 15.5 in [7])**.** *Let $X_1, \ldots, X_n$ n independent random variables, $f : \mathbb{R}^n \to \mathbb{R}$ a measurable function, and define $Z = f(X_1, \ldots, X_n)$ and $Z_i' = f(X_1, \ldots, X_i', \ldots, X_n)$, with $X_1', \ldots, X_n'$ independent copies of X_i. Furthermore let $V_+ = \mathbb{E}\left[\sum_i^n \left[(Z - Z_i')_+\right]^2 \mid X_1^n\right]$ and $V_- = \mathbb{E}\left[\sum_i^n \left[(Z - Z_i')_-\right]^2 \mid X_1^n\right]$. Then there exists a constant $\kappa \leq 1,271$ such that for all q in $[2, +\infty[$,*

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

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

$$\left\|Z - \mathbb{E}Z\right\|_q \leq \sqrt{2\kappa q} \sqrt{\left\|\sum_{i=1}^{n} \left(Z - Z_i'\right)^2\right\|_{q/2}} \tag{D.3}$$

$$\leq \sqrt{4\kappa q} \sqrt{\left\|\sum_{i=1}^{n} \left(Z - \mathbb{E}\left[Z \mid (X_j)_{j \neq i}\right]\right)^2\right\|_{q/2}}. \tag{D.4}$$

Moreover considering $Z^j = f(X_1, \dots, X_{j-1}, X_{j+1}, \dots, X_n)$ *for every* $1 \le j \le n$, *it results*

$$\|Z - \mathbb{E}Z\|_q \le 2\sqrt{2\kappa q}\sqrt{\left\|\sum_{i=1}^n (Z - Z^j)^2\right\|_{q/2}} \,. \tag{D.5}$$

Proof of Corollary D.1.
First note that

$$\left\|(Z - \mathbb{E}Z)_+\right\|_q^q + \left\|(Z - \mathbb{E}Z)_-\right\|_q^q = \|Z - \mathbb{E}Z\|_q^q \,.$$

Consequently,

$$\begin{aligned}
\|Z - \mathbb{E}Z\|_q^q &\le \sqrt{2\kappa q}^q \left(\sqrt{\|V_+\|_{q/2}}^q + \sqrt{\|V_-\|_{q/2}}^q\right) \\
&\le \sqrt{2\kappa q}^q \left(\|V_+\|_{q/2}^{q/2} + \|V_-\|_{q/2}^{q/2}\right) \\
&\le \sqrt{2\kappa q}^q \left\|\sum_{i=1}^n \mathbb{E}\left[\left(Z - Z_i'\right)^2 \mid X_1^n\right]\right\|_{q/2}^{q/2} \,.
\end{aligned}$$

Besides,

$$\begin{aligned}
\mathbb{E}\left[\left(Z - Z_i'\right)^2 \mid X_1^n\right] &= \mathbb{E}\left[\left(Z - \mathbb{E}\left[Z \mid (X_j)_{j\neq i}\right] + \mathbb{E}\left[Z \mid (X_j)_{j\neq i}\right] - Z_i'\right)^2 \mid X_1^n\right] \\
&= \mathbb{E}\left[\left(Z - \mathbb{E}\left[Z \mid (X_j)_{j\neq i}\right]\right)^2 + \left(\mathbb{E}\left[Z \mid (X_j)_{j\neq i}\right] - Z_i'\right)^2 \mid X_1^n\right] \\
&= \mathbb{E}\left[\left(Z - \mathbb{E}\left[Z \mid (X_j)_{j\neq i}\right]\right)^2 \mid X_1^n\right] + \mathbb{E}\left[\left(\mathbb{E}\left[Z_i' \mid (X_j)_{j\neq i}\right] - Z_i'\right)^2 \mid X_1^n\right] \,.
\end{aligned}$$

Combining the two previous results leads to

$$\begin{aligned}
&\|Z - \mathbb{E}Z\|_q \\
\le&\sqrt{2\kappa q}\sqrt{\left\|\sum_{i=1}^n \left(Z - \mathbb{E}\left[Z \mid (X_j)_{j\neq i}\right]\right)^2\right\|_{q/2} + \left\|\sum_{i=1}^n \mathbb{E}\left[\left(\mathbb{E}\left[Z_i' \mid (X_j)_{j\neq i}\right] - Z_i'\right)^2 \mid X_1^n\right]\right\|_{q/2}} \\
=&\sqrt{4\kappa q}\sqrt{\left\|\sum_{i=1}^n \left(Z - \mathbb{E}\left[Z \mid (X_j)_{j\neq i}\right]\right)^2\right\|_{q/2}} \,.
\end{aligned}$$

□

D.1.6 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 [19] (see Theorem 9.2).

D.1.7 Rosenthal's inequality

Proposition D.2 (Eq. (20) in [29])**.** *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 $q > 2$,*

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

Proof of Proposition D.3. From Lemma D.4, let us observe

- if $2 < q \leq 4$,

$$B^*(q,\gamma) \leq \left(2\sqrt{2e}\sqrt{q}\right)^q$$

by choosing $\gamma = 1$.

- if $4 < q$,

$$B^*(q,\gamma) \leq q^{-q/2}\left(\sqrt{4eq\left(q^{1/2}+q\right)}\right)^q \leq q^{-q/2}\left(\sqrt{8eq}\right)^q = \left(2\sqrt{2e}\sqrt{q}\right)^q,$$

with $\gamma = q^{(q-1)/2}$.

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

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

which leads to the conclusion.

□

Lemma D.4. *With the same notation as Proposition D.2 and for every $\gamma > 0$, it comes*

- *for every $2 < q \leq 4$,*

$$B^*(q,\gamma) \leq 1 + \frac{\left(\sqrt{2e}\sqrt{q}\right)^q}{\gamma},$$

- *for every $4 < q$,*

$$B^*(q,\gamma) \leq \gamma^{-q/(q-1)}\left(\sqrt{4eq\left(\gamma^{1/(q-1)}+q\right)}\right)^q.$$

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

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

by use of Lemma D.10 and $\sqrt{q}^{1/q} \leq \sqrt{e}$ for every $q > 2$.

If $q > 4$,

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

applying Lemma D.12 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.5. *For every $1 \le i \le n$ and $1 \le p \le n$, one has*

$$\mathbb{P}_e\left(i \in \bar{e}\right) = \frac{p}{n} \tag{D.6}$$

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

In the same way,

$$\sum_{k<\sigma_i(j)\le k+p} \mathbb{P}_e\left[\,i \in \bar{e},\ j \in V_k^e(X_i)\,\right] = \frac{kp}{n}\frac{p-1}{n-1}\ . \tag{D.8}$$

Proof of Lemma D.5. 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

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

□

D.2.2 Stone's lemma

Lemma D.6. *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 only depends on d.*

A proof of this lemma can be found in [19] (see Corollary 11.1).

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

Lemma D.7. *Let $\hat{f}_k$ and $\widehat{g}$ denote k-NN classifiers built respectively from $(X_1, Y_1), \ldots, (X_n, Y_n)$ and $(X_1, Y_1), \ldots, (X_{n-p}, Y_{n-p})$, for $1 \leq p \leq n-1$. Then for a new random variable (X, Y) with the same distribution as (X_i, Y_i), it comes*

$$\mathbb{P}\left(\hat{f}_k(X) \neq \widehat{g}(X)\right) \leq \frac{4}{\sqrt{2\pi}}\frac{p\sqrt{k}}{n} \ .$$

This lemma is proved in [21, Formula 14].

D.2.4 Upper bound for the L1O estimator

Lemma D.8. *One has*

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

where $\widehat{f}_k = \widehat{f}_k(Z_{1,n-1}; .)$ is the kNN classifier built from a sample of cardinality $n-1$.

This lemma corresponds to Theorem 24.4 in [19] where the proof can be found.

D.2.5 Moment upper bounds for the L1O estimator

Lemma D.9.

$$\mathbb{E}\left[\left(\bar{h}\left(Z_1,\ldots,Z_m\right)\right)^{2q}\right] \leq q!\left(2\frac{\left(k\gamma_d\right)^2}{m}\right)^q. \tag{D.9}$$

The proof is straightforward from the combination of Lemmas D.2 and D.8.

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

Lemma D.10. *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.10. If q is even ($q = 2k > 2$), then

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

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

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

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

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

□

Lemma D.11. *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*

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

Proof of Lemma D.11. 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

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

If q is even, then $q - 1 > 2$ is odd and the same calculations as in the proof of Lemma D.10 apply, which leads to

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

If q is odd, then $q - 1 > 2$ is even and another use of the calculations in the proof of Lemma D.10 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

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

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

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

□

Lemma D.12. *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.12. 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.11 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

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

□