Theoretical analysis of cross-validation for estimating the risk of the $k$-Nearest Neighbor classifier

Alain Celisse  
Laboratoire de Mathématiques  
Modal Project-Team  
UMR 8524 CNRS-Université Lille 1  
F-59655 Villeneuve d’Ascq Cedex, France  
celisse@math.univ-lille1.fr

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-75231 Paris cedex 05, France  
maryhuar@agroparistech.fr

Abstract

The present work aims at deriving theoretical guaranties on the behavior of some cross-validation procedures applied to the $k$-nearest neighbors ($k$NN) rule in the context of binary $\{0,1\}$-classification. Here we focus on the leave-$p$-out cross-validation ($L^pO$) used to assess the performance of the $k$NN classifier. Remarkably this $L^pO$ estimator can be efficiently computed using closed-form formulas derived by Celisse and Mary-Huard (2011).

We describe a general strategy to derive exponential concentration inequalities for the $L^pO$ estimator applied to the $k$NN classifier. This relies on deriving upper bounds on the polynomial moments of the centered $L^pO$ estimator by first deriving such bounds for the leave-one-out ($L1O$) estimator. Such results are obtained by exploiting the connection between the $L^pO$ estimator and U-statistics as well as by making an intensive use of the generalized Efron-Stein inequality. One other contribution is the extension to the $L^pO$ of the consistency results previously established by Rogers and Wagner (1978) for the $L1O$ as an estimator of the risk and/or the error rate of the $k$NN classifier.

Keywords: Classification, Cross-validation, Risk estimation

1. Introduction

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

The theoretical properties of the \( k \)NN classifier have been already extensively investigated. In the context of binary classification, preliminary theoretical results date back to Cover and Hart (1967); Cover (1968); Györfi (1981). More recently, Psaltis et al. (1994); Kulkarni and Posner (1995) derived an asymptotic equivalent to the performance of the 1NN classification rule, further extended to \( k \)NN by Snapp and Venkatesh (1998). Hall et al. (2008) also derived asymptotic expansions of the risk of the \( k \)NN classifier assuming either a Poisson or a binomial model for the training points, which relates this risk to the parameter \( k \). By contrast to the aforementioned results, the work by Chaudhuri and Dasgupta (2014) focuses on the finite sample framework. They typically provide upper bounds with high probability on the risk of the \( k \)NN classifier where the bounds are not distribution-free. Alternatively in the regression setting, Kulkarni and Posner (1995) provide a finite-sample bound on the performance of 1NN that has been further generalized to the \( k \)NN rule (\( k \geq 1 \)) by Biau et al. (2010a), where a bagged version of the \( k \)NN rule is also analyzed and then applied to functional data Biau et al. (2010b). We refer interested readers to Biau and Devroye (2016) for an almost thorough presentation of known results on the \( k \)NN algorithm in various contexts.

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

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

Although being favored in practice for assessing the risk of the \( k \)NN classifier, the use of CV comes with very few theoretical guarantees regarding its performance. Moreover probably for technical reasons, most existing results apply to Hold-out and leave-one-out (L1O), that is L\( p \)O with \( p = 1 \) (Kearns and Ron, 1999). In this paper we rather consider the general L\( p \)O procedure (for \( 1 \leq p \leq n - 1 \)) used to estimate the risk (alternatively the classification error rate) of the \( k \)NN classifier. Our main purpose is then to provide theoretical guarantees on the behavior of L\( p \)O with respect to \( p \). For instance we aim at
Contributions. The main contribution of the present work is to describe a new general strategy to derive exponential concentration inequalities for the $L^p$ estimator applied to the $k$NN binary classifier.

This strategy relies on several steps. We start by upper bounding the polynomial moments of the centered $L^p$ estimator in terms of those of the $L^1$ estimator. This is first achieved by exploiting the connection between the $L^p$ estimator and U-statistics (Koroljuk and Borovskich, 1994), as well as the Rosenthal inequality (Ibragimov and Sharakhmetov, 2002). Then, we derive upper bounds on the moments of the $L^1$ estimator using the generalized Efron-Stein inequality (Boucheron et al., 2005, 2013, Theorem 15.5) This allows us to infer the influence of the parameter $p$ on the concentration rate of the $L^p$ estimator for any $k$. We finally deduce the new exponential concentration inequalities for the $L^p$ estimator, which gives some insight on the behavior of the $L^p$ estimator whatever the value of the ratio $p/n \in (0, 1)$. In particular while the upper bounds increase with $1 \leq p \leq n/2 + 1$, it is no longer the case if $p > n/2 + 1$.

The remainder of the paper is organized as follows. The connection between the $L^p$ estimator and U-statistics is clarified in Section 2. Order-$q$ moments ($q \geq 2$) of the $L^p$ estimator are then upper bounded in terms of those of the $L^1$ estimator. Section 3 then specifies the previous upper bounds in the case of the $k$NN classifier. This leads, for any $k$, to the main Theorem 3.2 characterizing the concentration behavior of the $L^p$ estimator with respect to $p$. Deriving exponential concentration inequalities for the $L^p$ estimator is the main concern of Section 4. We illustrate the strength of our strategy by first providing concentration inequalities derived with less sophisticated tools, and then provide our main inequality. Finally Section 5 collects new extensions to $L^p$ of previous results originally established for $L^1$. This section ends by assessing the discrepancy between the $L^p$ estimator and the risk of the $k$NN classifier.

2. U-statistics and $L^p$ 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 probability distribution. In what follows no particular distributional assumption is made regarding $X$. To predict the label, one aims at building a classifier $\hat{f} : \mathcal{X} \to \{0, 1\}$ on the basis of a set of random variables $Z_{1:n} = \{Z_1, \ldots, Z_n\}$ called the training sample, where $Z_i = (X_i, Y_i)$, $1 \leq i \leq n$ represent $n$ copies of $(X, Y)$ drawn independently from $P_{(X,Y)}$. Any strategy to build such a classifier is called a classification algorithm or classification rule, and can be formally defined as a function $A : \bigcup_{n \geq 1} \{(X \times \{0, 1\})^n \to \mathcal{F} \text{ that maps a training sample } Z_{1:n} \text{ onto the corresponding classifier } A(Z_{1:n}; \cdot) = \hat{f} \in \mathcal{F}, \text{ where } \mathcal{F} \text{ is the set of all measurable functions from } \mathcal{X} \text{ to } \{0, 1\}. \text{ Numerous 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 Devroye et al. (1996) for many
instances). Here we focus on the $k$-nearest neighbor rule ($k$NN) initially proposed by Fix and Hodges (1951) and further studied for instance by Devroye and Wagner (1977); Rogers and Wagner (1978). For $1 \leq k \leq n$, the $k$NN rule, denoted by $A_k$, consists in classifying any new observation $x$ using a majority vote decision rule based on the label of the $k$ points $X(1)(x), \ldots, X(k)(x)$ closest to $x$ among the training sample $X_1, \ldots, X_n$, according to some distance function:

$$A_k(Z_{1:n}; x) = \tilde{f}_k(Z_{1:n}; x) := \begin{cases} 1 & \text{if } \frac{1}{k} \sum_{i \in V_k(x)} Y_i = \frac{1}{k} \sum_{i=1}^{k} Y(i)(x) > 0.5 \\ 0 & \text{otherwise} \end{cases} \quad (2.1)$$

where $V_k(x) = \{1 \leq i \leq n, \ X_i \in \{X(1)(x), \ldots, X(k)(x)\}\}$ 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 of the $i$-th nearest neighbor of $x$ for $1 \leq i \leq k$. The choice of the distance function will typically depend on the nature of the data to be dealt with, and will not be discussed here. While in some applications adaptive metrics have been considered (see Hastie et al., 2001, Chap. 14 for instance), in what follows we will assume the distance function to be fixed, i.e. that its definition does not depend on the specific training sample at hand. Let us further assume that ties are broken at random, for instance by choosing the smallest index among ties.

For a given sample $Z_{1:n}$, 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

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

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

$$\hat{R}_p(A, Z_{1:n}) = \left( \frac{n}{p} \right)^{-1} \sum_{e \in \mathcal{E}_{n-p}} \mathbb{P} \left( \hat{f}(X) \neq Y \mid Z_{1:n} \right), \quad (2.2)$$

where $A(Z^e; \cdot)$ is the classifier built from $Z^e$. We refer the reader to Arlot and Celisse (2010) for a detailed description of LpO and other cross-validation procedures.

However unlike what arises from (2.2), the LpO estimator can be efficiently computed by use of closed-form formulas with a time complexity linear in $p$ when applied to the $k$NN classification rule Celisse and Mary-Huard (2011). This property remains true in other contexts such as density estimation Celisse and Robin (2008); Celisse (2014), regression Celisse (2008); Arlot and Celisse (2011), 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 $\left( \frac{n}{p} \right)$.

Since no theoretical guarantee does exist on the performance of LpO applied to the $k$NN classifier, one main goal in what follows is to design a general strategy to derive such results.
2.2 U-statistics: General bounds on \( L_p \)O moments

The purpose of the present section is to describe a general strategy allowing to derive new upper bounds on the polynomial moments of the \( L_p \)O estimator. As a first step of this strategy, we establish the connection between the \( L_p \)O risk estimator and U-statistics. Second, we exploit this connection to derive new upper bounds on the order-\( q \) moments of the \( L_p \)O estimator for \( q \geq 2 \). Note that these upper bounds, which relate moments of the \( L_p \)O estimator to those of the \( L_1 \)O estimator, hold true with any classifier.

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

**Definition 2.1 (Koroljuk and Borovskich (1994)).** Let \( h : X^m \to \mathbb{R} \) (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 \( U_n : X^n \to \mathbb{R} \) such that

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

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

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

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

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

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

\[
U_n(x_1, \ldots, x_n) = \frac{1}{n!} \sum_{\sigma} W(x_{\sigma(1)}, \ldots, x_{\sigma(n)}) ,
\]

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

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

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

\[
h_m(Z_1, \ldots, Z_m) = \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_{\{A(Z_{i,m}; X_i) \neq Y_i\}},
\]

where \( Z_{1,m}^{(i)} \) denotes the sample \( (Z_1, \ldots, Z_m) \) with \( Z_i \) withdrawn.
Proof of Theorem 2.1.

From Eq. (2.2), the LpO estimator of the performance of any classification algorithm \( A \) computed from \( Z_{1,n} \) satisfies

\[
\hat{R}_p(A, Z_{1,n}) = \frac{1}{(p)} \sum_{e \in \mathcal{E}_{n-p}} \frac{1}{p} \sum_{i \in e} \mathbb{1}_{\{A(Z^e;X_i) \neq Y_i\}}
\]

\[
= \frac{1}{(p)} \sum_{e \in \mathcal{E}_{n-p}} \frac{1}{p} \sum_{i \in e} \left( \sum_{v \in \mathcal{E}_{n-p+1}} \mathbb{1}_{\{v \cap \{i\}\}} \right) \mathbb{1}_{\{A(Z^e;X_i) \neq Y_i\}},
\]

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

\[
\hat{R}_p(A, Z_{1,n}) = \frac{1}{(n)} \sum_{v \in \mathcal{E}_{n-p+1}} \frac{1}{p} \sum_{i \in v} \left( \sum_{e \in \mathcal{E}_{n-p}} \mathbb{1}_{\{v \cap \{e\}\}} \mathbb{1}_{\{i \in e\}} \right) \mathbb{1}_{\{A(Z^v;X_i) \neq Y_i\}},
\]

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

\[
\hat{R}_p(A, Z_{1,n}) = \frac{1}{(p)} \frac{1}{n} \sum_{v \in \mathcal{E}_{n-p+1}} \sum_{i \in v} \mathbb{1}_{\{i \in v\}} \mathbb{1}_{\{A(Z^{v\setminus\{i\}};X_i) \neq Y_i\}}
\]

\[
= \frac{1}{(n-p+1)} \sum_{v \in \mathcal{E}_{n-p+1}} \frac{1}{n-p+1} \sum_{i \in v} \mathbb{1}_{\{A(Z^{v\setminus\{i\}};X_i) \neq Y_i\}},
\]

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

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

\[
h_m(Z_1,\ldots,Z_m) = \hat{R}_1(A, Z_{1,m}). \quad (2.4)
\]

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

We now derive a general upper bound on the \( q \)-th moment \( (q \geq 1) \) of the LpO estimator that holds true for any classification rule as long as the following quantities remain meaningful.

Theorem 2.2. For any classification rule \( A \), let \( A(Z_{1,n};\cdot) \) and \( A(Z_{1,m};\cdot) \) 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, and any \( q \geq 1 \),

\[
\mathbb{E} \left[ \left| \hat{R}_p(A, Z_{1,n}) - \mathbb{E} \left[ \hat{R}_p(A, Z_{1,n}) \right] \right|^q \right] \leq \mathbb{E} \left[ \left| \hat{R}_1(A, Z_{1,m}) - \mathbb{E} \left[ \hat{R}_1(A, Z_{1,m}) \right] \right|^q \right]. \quad (2.5)
\]

Furthermore as long as \( p > n/2 + 1 \), one also gets
• for \( q = 2 \)

\[
\mathbb{E} \left[ \left( \bar{R}_p(A, Z_{1,n}) - \mathbb{E} \left[ \bar{R}_p(A, Z_{1,n}) \right] \right)^2 \right] \leq \mathbb{E} \left[ \left( \bar{R}_1(A, Z_{1,m}) - \mathbb{E} \left[ \bar{R}_1(A, Z_{1,m}) \right] \right)^2 \right] \left( \frac{n}{m} \right) \cdot \left( 2.6 \right)
\]

• for every \( q > 2 \)

\[
\mathbb{E} \left[ \left| \bar{R}_p(A, Z_{1,n}) - \mathbb{E} \left[ \bar{R}_p(A, Z_{1,n}) \right] \right|^q \right] \leq B(q, \gamma) \times \left\{ 2^{q\gamma} \frac{n}{m} \mathbb{E} \left[ \left( \bar{R}_1(A, Z_{1,m}) - \mathbb{E} \left[ \bar{R}_1(A, Z_{1,m}) \right] \right)^q \right] \right\} \left( 2.7 \right)
\]

where \( \gamma > 0 \) is a numeric constant and \( B(q, \gamma) \) denotes the optimal constant defined in the Rosenthal inequality (Proposition D.2), and \( a \lor b = \max(a, b) \) for every \( a, b \in \mathbb{R} \).

The proof is given in Appendix A.1. Eq. (2.5) and Eq. (2.6) straightforwardly result from the Jensen inequality applied to the average over all permutations provided in Proposition 2.1. If \( p > n/2 + 1 \), the integer part \( \lfloor n/m \rfloor \) becomes larger than 1 and Eq. (2.6) becomes better than Eq. (2.5) for \( q = 2 \). As a consequence of our strategy of proof, the right-hand side of Eq. (2.6) is equal to the classical upper bound on the variance of U-statistics.

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

3. New bounds on \( L_p \)O moments for the \( k \)NN classifier

Our goal is now to specify the general upper bounds provided by Theorem 2.2 in the case of the \( k \)NN classification rule \( A_k \) (1 \( \leq k \leq n \)) introduced by (2.1).

Since Theorem 2.2 expresses the moments of the \( L_p \)O estimator in terms of those of the L1O estimator, the next step consists in focusing on the L1O moments. Deriving tight upper bounds on the moments of the L1O is achieved using a generalization of the well-known Efron-Stein inequality (see Theorem D.1 for Efron-Stein’s inequality and Theorem 15.5 in Boucheron et al. (2013) for its generalization). For the sake of completeness, we first recall a corollary of this generalization that is proved in Section D.1.4 (see Corollary D.1).

**Proposition 3.1.** Let \( X_1, \ldots, X_n \) 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 \( Z'_i = \ldots \)
f(\(X_1, \ldots, X'_i, \ldots, X_n\)), where \(X'_1, \ldots, X'_n\) are independent copies of the \(X_i\)s, there exists a universal constant \(\kappa \leq 1.271\) such that for any \(q \geq 2\),

\[
\|Z - EZ\|_q \leq \sqrt{2\kappa q} \left\| \sum_{i=1}^{n} (Z - Z'_i)^2 \right\|_{q/2}.
\]

Then applying Proposition 3.1 with \(Z = \hat{R}_1(A_k(Z_{1,m};\cdot))\) leads to the following Theorem 3.1, which finally allows to control the order-\(q\) moments of the L1O estimator applied to the \(k\)NN classifier.

**Theorem 3.1.** For every \(1 \leq k \leq n - 1\), let \(A_k(Z_{1,m};\cdot)\) \((m = n - p + 1)\) denote the \(k\)NN classifier learnt from \(Z_{1,m}\) and \(\hat{R}_1(A_k(Z_{1,m};\cdot))\) be the corresponding L1O estimator given by Eq. (2.2). Then

- for \(q = 2\),
  \[
  \mathbb{E} \left[ \left( \hat{R}_1(A_k, Z_{1,m}) - \mathbb{E} \left[ \hat{R}_1(A_k, Z_{1,m}) \right] \right)^2 \right] \leq C_1 \sqrt{k} \left( \frac{\sqrt{k}}{\sqrt{m}} \right)^2;
  \]

- for every \(q > 2\),
  \[
  \mathbb{E} \left[ \left| \hat{R}_1(A_k, Z_{1,m}) - \mathbb{E} \left[ \hat{R}_1(A_k, Z_{1,m}) \right] \right|^q \right] \leq (C_2 \sqrt{q})^q \left( \frac{\sqrt{k}}{\sqrt{m}} \right)^k,
  \]

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

Its proof (detailed in Section A.2) involves the use of Stone’s lemma (Lemma D.5), which enables to upper bound, for a given \(X_i\), the number of points \(\{X_j\}_{j \neq i}\) having \(X_i\) among their \(k\) nearest neighbors by \(k\gamma_d\).

First, the maths in these upper bounds have not been simplified on purpose to facilitate their comparison and emphasize the difference in the dependence with respect to \(k\). In particular the larger dependence on \(k\) for \(q > 2\) results from the difficulty to derive a tight upper bound for the expectation of \((\sum_{i=1}^{n} \mathbb{1}\{A_k(Z^{(i)}_{1,m};X_i) \neq A_k(Z_{1,m};X_i)\})^q\) in this case, where \(Z^{(i)}_{1,m}\) (resp. \(Z^{(i,j)}_{1,m}\)) denotes the sample \(Z_{1,m}\) where \(Z_i\) has been (resp. \(Z_i\) and \(Z_j\) have been) removed.

Second, the easier case \(q = 2\) enables to exploit exact calculations (rather than upper bounds) of the variance of the L1O. Note that this \(k^{3/2}/m\) is better than than the ongoing upper bound we would obtain from using the sub-Gaussian exponential concentration inequality provided by Theorem 24.4 in Devroye et al. (1996), which only leads to a \(k^2/m\).

We are now in position to state the main result of this section. It follows from the combination of Theorem 2.2 (connecting moments of the LpO estimator to those of the L1O) and Theorem 3.1 (providing an upper bound on the order-\(q\) moments of the L1O).
Theorem 3.2. For every $p, k \geq 1$ such that $p + k \leq n$, let $\hat{R}_p(A_k, Z_{1,n})$ denote the $L_p$ risk estimator (see (2.2)) of the $k$NN classifier $A_k(Z_1, n \cdot )$ defined by (2.1). Then there exist (known) constants $C_1, C_2 > 0$ such that for every $1 \leq p \leq n - k$,

- for $q = 2$,
  \[ \mathbb{E} \left[ \left( \hat{R}_p(A_k, Z_{1,n}) - \mathbb{E} \left[ \hat{R}_p(A_k, Z_{1,n}) \right] \right)^2 \right] \leq C_1 \frac{k^{3/2}}{(n - p + 1)} ; \quad (3.1) \]

- for every $q > 2$,
  \[ \mathbb{E} \left[ \left| \hat{R}_p(A_k, Z_{1,n}) - \mathbb{E} \left[ \hat{R}_p(A_k, Z_{1,n}) \right] \right|^q \right] \leq (C_2 k)^q \left( \frac{q}{n-p+1} \right)^{q/2} , \quad (3.2) \]

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

- for $q = 2$,
  \[ \mathbb{E} \left[ \left( \hat{R}_p(A_k, Z_{1,n}) - \mathbb{E} \left[ \hat{R}_p(A_k, Z_{1,n}) \right] \right)^2 \right] \leq C_1 \frac{k^{3/2}}{(n - p + 1) \left[ \frac{n}{n-p+1} \right]} , \quad (3.3) \]

- for every $q > 2$,
  \[ \mathbb{E} \left[ \left| \hat{R}_p(A_k, Z_{1,n}) - \mathbb{E} \left[ \hat{R}_p(A_k, Z_{1,n}) \right] \right|^q \right] \leq \left\lfloor \frac{n}{n - p + 1} \right\rfloor \Gamma^q \left( \frac{k^{3/2}}{(n - p + 1) \left[ \frac{n}{n-p+1} \right]} \right)^{q/2} \left( \frac{k^2}{(n - p + 1) \left[ \frac{n}{n-p+1} \right]^2} \right)^{q/2} \left( \frac{q}{n-p+1} \right)^{q/2} , \quad (3.4) \]

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

The straightforward proof is detailed in Section A.3.

Let us start by noticing that both Eq. (3.1) and Eq. (3.2) provide upper bounds which deteriorate as $p$ grows. This is no longer the case for Eq. (3.3) and Eq. (3.4), which are specifically designed to cover the setup where $p > n/2 + 1$, that is where $\lfloor n/m \rfloor$ is no longer equal to 1. The interest of these last two inequalities is also illustrated by considering the case where $n - p$ remains fixed, that is independent of $n$. This is a particular instance of the setup where $p/n \to 1$, as $n \to +\infty$, which has been investigated in different frameworks by Shao (1993); Yang (2006, 2007); Celisse (2014). 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 $k^{3/2}/n$ (for $q = 2$) and $k^q/n^{q-1}$ (for $q > 2$).

One can also emphasize that, as a U-statistic of order $m = n - p + 1$, the $L_p$O estimator has a known limiting distribution when its order $m$ remains constant with respect to $n$, \[ \text{Performance of CV to estimate the risk of } k\text{NN} \]
which amounts to require \( n - p \) is equal to a constant. In this setup, it is known (see Theorem A, Section 5.5.1 Serfling, 1980) that
\[
\sqrt{\frac{n}{m}} \left( \hat{R}_p(A_k, Z_{1,n}) - \mathbb{E} \left[ \hat{R}_p(A_k, Z_{1,n}) \right] \right) \xrightarrow{n++\infty} \mathcal{N}(0, \zeta_1),
\]
where \( \zeta_1 = \text{Var} \left[ g(Z_1) \right] \), with \( g(z) = \mathbb{E} \left[ h_m(z, Z_2, \ldots, Z_m) \right] \). Therefore the upper bound given by Eq. (3.3) has the right magnitude with respect to \( n \) as long as \( m = n - p + 1 \) is assumed to be constant.

Finally Eq. (3.4) has been derived using a specific version of the Rosenthal inequality (Ibragimov and Sharakhmetov, 2002) stated with the optimal constant and involving a “balancing factor”. In particular this balancing factor has allowed us to optimize the relative weight of the two terms between brackets in Eq. (3.4). This leads us to claim that the dependence of the upper bound with respect to \( q \) cannot be improved with this line of proof. However we do not conclude that the term in \( q^3 \) cannot be improved using other technical arguments.

4. Exponential concentration inequalities

In this section, we provide exponential concentration inequalities for the \( L_p \) estimator applied to the \( k \)NN classifier. The main inequalities we provide at the end of this section heavily rely on the moments inequalities previously derived in Section 3, that is Theorem 3.2. In order to emphasize the interest of our approach, we start this section by proving two exponential inequalities obtained with less sophisticated tools. For each of them, we discuss its strength and weakness to justify the additional refinements we further explore step by step.

A first exponential concentration inequality for \( \hat{R}_p(A_k, Z_{1,n}) \) can be derived by use of the bounded difference inequality following the line of proof of Devroye et al. (1996, Theorem 24.4) originally developed for the \( L_1 \) estimator.

**Proposition 4.1.** For any integers \( p, k \geq 1 \) such that \( p + k \leq n \), let \( \hat{R}_p(A_k, Z_{1,n}) \) denote the \( L_p \) estimator \((2.2)\) of the classification error of the \( k \)NN classifier \( A_k(Z_{1,n}; \cdot) \) defined by \((2.1)\). Then for every \( t > 0 \),
\[
P \left( \left| \hat{R}_p(A_k, Z_{1,n}) - \mathbb{E} \left( \hat{R}_p(A_k, Z_{1,n}) \right) \right| > t \right) \leq 2e^{-n \frac{t^2}{8(k + p - 1)^2\gamma_d}} .
\]
where \( \gamma_d \) denotes the constant introduced in Stone’s lemma (Lemma D.5).

The proof is given in Appendix B.1 and relies on the McDiarmid inequality (Theorem D.3).

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

This reasoning results in a rough upper bound since the dominator in the exponent exhibits a \( (k + p - 1)^2 \) factor. This indicates we do not distinguish between points for
which $X_j$ is among the $k$ nearest neighbors of $X_i$ in the whole sample, or above the $k$-th one. However these two setups lead to strongly different probabilities of being among the $k$ nearest neighbors in the training sample in practice. Consequently the dependence of the convergence rate on $k$ and $p$ in Proposition 4.1 is not optimal, as confirmed by forthcoming Theorems 4.1 and 4.2.

Based on the previous comments, a sharper quantification of the influence of each nearest 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 $\hat{R}_p(A_k, Z_{1,n})$ denote the $L^p$ estimator (2.2) of the classification error of the $k$NN classifier $A_k(Z_{1,n}; \cdot)$ defined by (2.1). Then there exists a numeric constant $\Box > 0$ such that for every $t > 0$,

$$
P \left( \hat{R}_p(A_k, Z_{1,n}) - \mathbb{E} \left( \hat{R}_p(A_k, Z_{1,n}) \right) > t \right) \lor \mathbb{P} \left( \mathbb{E} \left[ \hat{R}_p(A_k, Z_{1,n}) \right] - \hat{R}_p(A_k, Z_{1,n}) > t \right) \leq \exp \left( -\frac{nt^2}{\Box k^2 \left[ 1 + (k + p) \frac{p-1}{n-p} \right]} \right),$$

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

The proof is given in Section B.2.

Let us first remark that, in accordance with the previous comments on the deficiencies of Proposition 4.1, taking into account the rank of each neighbor in the whole sample enables to considerably reduce the weight of the denominator in the exponent. In particular, one observes that letting $p/n \to 0$ as $n \to +\infty$ (with $k$ assumed to be fixed for instance) makes the influence of the $k+p$ factor asymptotically negligible. This would allow to recover (up to numeric constants) a similar upper bound to that of Devroye et al. (1996, Theorem 24.4), which is achieved by ours in the particular case where $p = 1$.

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 the upper bound seems to strictly deteriorate as $p$ increases, which contrasts with the upper bounds derived for $p > n/2 + 1$ in Theorem 3.2. This drawback is overcome by the following result, which is our main contribution in the present section.

**Theorem 4.2.** For every $p, k \geq 1$ such that $p + k \leq n$, let $\hat{R}_p(A_k, Z_{1,n})$ denote the $L^p$ estimator of the classification error of the $k$NN classifier $\hat{f}_k = A_k(Z_{1,n}; \cdot)$ defined by (2.1). Then for every $t > 0$,

$$
P \left( \hat{R}_p(A_k, Z_{1,n}) - \mathbb{E} \left[ \hat{R}_p(A_k, Z_{1,n}) \right] > t \right) \lor \mathbb{P} \left( \mathbb{E} \left[ \hat{R}_p(A_k, Z_{1,n}) \right] - \hat{R}_p(A_k, Z_{1,n}) > t \right) \leq \exp \left( -(n-p+1) \frac{t^2}{\Delta^2 k^2} \right),$$

where $\Delta = 4 e \max \left( C_2, \sqrt{C_1} \right)$ with $C_1, C_2 > 0$ defined in Theorem 3.1.
Furthermore in the particular setting where \( p > n/2 + 1 \), it comes

\[
\Pr \left( \hat{R}_p(A_k, Z_{1,n}) - \mathbb{E} \left[ \hat{R}_p(A_k, Z_{1,n}) \right] > t \right) \lor \Pr \left( \mathbb{E} \left[ \hat{R}_p(A_k, Z_{1,n}) \right] - \hat{R}_p(A_k, Z_{1,n}) > t \right) 
\leq e \left( \frac{n}{n-p+1} \right) \times 
\exp \left[ -\frac{1}{2e} \min \left\{ (n-p+1) \left( \frac{n}{n-p+1} \right), \left( \frac{n}{n-p+1} \right)^2 \right\} \left( (n-p+1) \left( \frac{n}{n-p+1} \right)^2 \right)^{1/3} \right],
\]

where \( \Gamma \) arises in Eq. (3.4) and \( \gamma_d \) denotes the constant introduced in Stone’s lemma (Lemma D.5).

The proof has been postponed to Appendix B.3. It involves different arguments for the two inequalities (4.2) and (4.3) depending on the range of \( p \). Firstly, for \( p \leq n/2 + 1 \), the two corresponding inequalities of Theorem 3.2 on the moments of the \( L_p \) estimator allow to characterize its sub-Gaussian behavior in terms of its even moments. Ineq. (4.2) then straightforwardly results from Lemma D.2. Secondly, for \( p > n/2 + 1 \), we rather exploit: (i) the appropriate upper bounds on the moments of the \( L_p \) estimator given by Theorem 3.2, and (ii) a dedicated Proposition D.1 which relates moment upper bounds to exponential concentration inequalities.

In accordance with the conclusions drawn about Theorem 3.2, one observes that the upper bound of Eq. (4.2) increases as \( p \) grows, unlike that of Eq. (4.3) which improves as \( p \) increases. In particular the best concentration rate in Eq. (4.3) is achieved for \( p = n - 1 \), whereas Eq. (4.2) turns out to be useless in that setting. Let us also notice that Eq. (4.2) is strictly better than Theorem 4.1 as long as \( p/n \to \delta \in [0,1) \), as \( n \to +\infty \).

In order to allow an easier interpretation of the last Ineq. (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.** With the same notation as Theorem 4.2, for any \( p, k \geq 1 \) such that \( p+k \leq n \), \( p > n/2 + 1 \), and for every \( t > 0 \)

\[
\Pr \left[ \hat{R}_p(A_k, Z_{1,n}) - \mathbb{E} \left[ \hat{R}_p(A_k, Z_{1,n}) \right] \right] > \frac{\sqrt{2e} \Gamma}{\sqrt{n-p+1}} \left( \frac{k^{3/2}}{n-p+1} t + 2e \frac{k}{n-p+1} t^{3/2} \right) 
\leq \left( \frac{n}{n-p+1} \right) e^{-e^{-t}},
\]

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

By comparison with the well-known Bernstein inequality (Boucheron et al., 2013, Theorem 2.10), let us remark the present inequality is very similar, except the second deviation term in \( t^{3/2} \) instead of \( t \) (for the Bernstein inequality). The first deviation term is of order \( \approx k^{3/4}/\sqrt{n} \), which is the same order with respect to \( n \) as what we would get in the
Bernstein inequality. The second deviation term is of a somewhat different order, that is \( \approx k\sqrt{n} - p + 1/n \) as compared with the usual \( 1/n \) in the Bernstein inequality. This means that varying \( p \) allows to interpolate between the \( k/\sqrt{n} \) rate and the \( k/n \) rate achieved for instance with \( p = n - 1 \).

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 \( k^2 \) which would result from Ineq. (4.2) in Theorem 4.2. However since the dependence of the two deviation terms is inherited from the upper bound on the \( L^1 \) stability established by Devroye and Wagner (1979, Eq. (14)), any improvement of the latter would lead to enhance the present concentration inequality.

5. Assessing the gap between \( L_p \)O and prediction error

In the present section, we derive new upper bounds on different measures of the discrepancy between \( \hat{R}_p(A_k, Z_1, n) \) and \( L(\hat{f}_k) \) or \( R(\hat{f}_k) = \mathbb{E} \left[ L(\hat{f}_k) \right] \). These bounds on the \( L_p \)O estimator are completely new for \( 1 < p < n - k \). Some of them are extensions of former ones specifically derived for the \( L_1 \)O estimator applied to the \( kNN \) classifier.

**Theorem 5.1.** For every \( p, k \geq 1 \) such that \( p + k < n \) let \( \hat{R}_p(A_k, Z_1, n) \) denote the \( L_p \)O risk estimator (see (2.2)) of the \( kNN \) classifier \( \hat{f}_k = A_k(Z_1, n; \cdot) \) defined by (2.1). Then,

\[
\left| \mathbb{E} \left[ \hat{R}_p(A_k, Z_1, n) \right] - R(\hat{f}_k) \right| \leq \frac{4}{\sqrt{2\pi}} \frac{p\sqrt{k}}{n} , \tag{5.1}
\]

and

\[
\mathbb{E} \left[ \left( \hat{R}_p(A_k, Z_1, n) - R(\hat{f}_k) \right)^2 \right] \leq 128 \kappa \gamma d \frac{k^{3/2}}{\sqrt{2\pi}} \frac{k^3}{n - p + 1} + \frac{16 p^2 k}{2\pi n^2} . \tag{5.2}
\]

Moreover,

\[
\mathbb{E} \left[ \left( \hat{R}_p(A_k, Z_1, n) - L(\hat{f}_k) \right)^2 \right] \leq \frac{2\sqrt{2} (2p + 3)\sqrt{k}}{\sqrt{\pi}} \frac{1}{n} + \frac{1}{n} . \tag{5.3}
\]

**Proof of Theorem 5.1.**

**Proof of (5.1):** Lemma D.6 immediately provides

\[
\left| \mathbb{E} \left[ \hat{R}_p(A_k, Z_1, n) - L(\hat{f}_k) \right] \right| = \left| \mathbb{E} \left[ L(\hat{f}_k) \right] - \mathbb{E} \left[ L(\hat{f}_k) \right] \right| \leq \mathbb{E} \left[ \mathbb{E} \left[ \mathbb{E} \left[ L(\hat{f}_k) \right] \right] - \mathbb{E} \left[ L(\hat{f}_k) \right] \right] \leq \mathbb{P} \left( A_k(Z_1, n; X) \neq A_k(Z_1, n; X) \right) \leq \frac{4}{\sqrt{2\pi}} \frac{p\sqrt{k}}{n} .
\]
Proof of (5.2): The proof combines the previous upper bound with the one established for the variance of the LpO estimator, that is Eq. (3.1).

\[
\begin{align*}
E & \left[ \left( \hat{R}_p(\mathcal{A}_k, Z_1, n) - E \left[ L(\hat{f}_k) \right] \right)^2 \right] \\
& = E \left[ \left( \hat{R}_p(\mathcal{A}_k, Z_1, n) - E \left[ \hat{R}_p(\mathcal{A}_k, Z_1, n) \right] \right)^2 \right] + \left( E \left[ \hat{R}_p(\mathcal{A}_k, Z_1, n) \right] - E \left[ L(\hat{f}_k) \right] \right)^2 \\
& \leq \frac{128\kappa\gamma d}{\sqrt{2\pi}} \frac{k^{3/2}}{n - p + 1} + \left( \frac{4}{\sqrt{2\pi}} \frac{p\sqrt{k}}{n} \right)^2,
\end{align*}
\]

which concludes the proof.

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

Keeping in mind that \( E \left[ \hat{R}_p(\mathcal{A}_k, Z_1, n) \right] = R(\mathcal{A}_k(Z_1, m)) \) (with \( m = n - p + 1 \)), the right-hand side of Ineq. (5.1) is an upper bound on the bias of the LpO estimator, that is on the difference between the risks of the classifiers built from respectively \( n - p \) and \( n \) points. Therefore, the fact that this upper bound increases with \( p \) is reliable since the classifiers \( \mathcal{A}_k(Z_1, m; \cdot) \) and \( \mathcal{A}_k(Z_1, n; \cdot) \) can become more and more different from one another as \( p \) increases.

More precisely, the upper bound in Ineq. (5.1) goes to 0 provided \( p\sqrt{k}/n \) does. It seems somewhat more restrictive than the usual condition on \( k \) and \( n \), that is \( k/n \to 0 \) as \( n \to +\infty \) (see Devroye et al., 1996, Chap. 6.6 for instance). However if one rather assumes \( p/n \to \delta \in (0, 1] \), then this upper bound does no longer decrease to 0. Here again this seemingly restriction is straightforwardly inherited from the bound on the \( L^1 \) stability of the \( k \)NN classifier (Devroye and Wagner, 1979, Eq. (14)). Therefore any improvement of this \( L^1 \) stability upper bound would enhance Ineq. (5.1).

Let us further notice that this restriction to values of \( p \) such that \( p/n \to 0 \) justifies the use of Eq. (3.1) along the derivation of Ineq. (5.2), which is relevant with \( p \leq n/2 + 1 \). Note that an upper bound similar to that of Ineq. (5.2) can be easily derived for any order-\( q \) moment \( (q \geq 2) \) at the price of increasing the constants by using \( (a + b)^q \leq 2^{q-1}(a^q + b^q) \), for every \( a, b \geq 0 \). We also emphasize that Ineq. (5.2) allows to control the discrepancy between the LpO estimator and the risk of the \( k \)NN classifier, that is the expectation of its classification error. Ideally we would have liked to replace the risk \( R(\hat{f}_k) \) by the prediction error \( L(\hat{f}_k) \). But using our strategy of proof, this would require an additional distribution-free concentration inequality on the prediction error of the \( k \)NN classifier. To the best of our knowledge, such a concentration inequality is not available up to now.

Finally upper bounding the squared difference between the LpO estimator and the prediction error is precisely the purpose of Ineq. (5.3). Proving the latter inequality requires a completely different strategy of proof which can be traced back to an earlier proof by Rogers and Wagner (1978, see the proof of Theorem 2.1) applying to the LIO estimator. It is also noticeable that Ineq. (5.3) combined with the Jensen inequality lead to a less accurate upper bound than Ineq. (5.1).

Let us conclude this section with a corollary, which provides a finite-sample bound on the gap between \( \hat{R}_p(\mathcal{A}_k, Z_1, n) \) and \( R(\hat{f}_k) = E \left[ L(\hat{f}_k) \right] \) with high probability. It is stated
under the same restriction on $p$ as the previous Theorem 5.1 it is based on, that is mainly for small values of $p$.

**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

$$\left| R(\hat{f}_k) - \hat{R}_p(A_k, Z_{1,n}) \right| \leq \sqrt{\frac{\Delta^2 k^2}{n (1 - p - 1/n)}} x + \frac{4 \sqrt{2\pi} p \sqrt{k}}{n} + 4 \sqrt{\frac{2\pi}{p \sqrt{k}}} n,$$

(5.4)

where $\hat{f}_k = A_k(Z_{1,n}; \cdot)$.

**Proof of Corollary 5.1.** Ineq. (5.4) results from combining Ineq. (4.2) (from Theorem 4.2) and Ineq. (5.1).

$$\left| R(\hat{f}_k) - \hat{R}_p(A_k, Z_{1,n}) \right| \leq \left| R(\hat{f}_k) - \mathbb{E} \left[ \hat{R}_p(A_k, Z_{1,n}) \right] \right| + \mathbb{E} \left[ \hat{R}_p(A_k, Z_{1,n}) - \hat{R}_p(A_k, Z_{1,n}) \right]$$

$$\leq \frac{4 \sqrt{\Delta^2 k^2}}{n (1 - p - 1/n)} x + \frac{4 \sqrt{2\pi} p \sqrt{k}}{n} + 4 \sqrt{\frac{2\pi}{p \sqrt{k}}} n.$$

It relies on the combination of the exponential concentration result derived for $\hat{R}_p(A_k, Z_{1,n})$ (Theorem 4.2) with the upper bound on the bias, that is Ineq. (5.1).

Note that the right-hand side of Ineq. (5.4) could be used to derive bounds on $R(\hat{f}_k)$ that are similar to confidence bounds. However we emphasize these confidence bounds have not been optimized with respect to the constants, which somewhat limits their practical applicability in finite-sample size settings.

Let us also recall that the last Ineq. (5.1) essentially applies to small values of $p$, which justifies the restriction in Corollary 5.1 to $p \leq n/2 + 1$. Indeed choosing $p$ such that $n - p$ remains constant would make the right-hand side of Ineq. (5.1) nondecreasing. However let us assume a tighter upper bound than that of Ineq. (5.1) can be derived in the setting where $p > n/2 + 1$. Then one would get an inequality similar to Ineq. (5.4) by solely replacing Theorem 4.2 by Proposition 4.2, which applies to $p > n/2 + 1$.

**6. Discussion**

The present work provides theoretical guarantees on the performance of $L_p$O used for estimating the risk of the $k$NN classifier. First the results derived in Section 4 give some new insight on the concentration of the $L_p$O estimator around its expectation for different rates of $p/n$. Furthermore the upper bounds in Ineq. (5.2) and (5.3) of Section 5 straightforwardly imply the consistency of the $L_p$O estimator towards the risk (or the classification error rate) of the $k$NN classifier.

It is worth mentioning that the upper-bounds derived in Sections 4 and 5 — see for instance Theorem 5.1 — can be minimized by choosing $p = 1$, suggesting that the L1O estimator is optimal in terms of risk estimation when applied to the $k$NN classification.
algorithm. This observation corroborates the results of the simulation study presented in Celisse and Mary-Huard (2011), where it is empirically shown that small values of $p$ (and in particular $p = 1$) lead to the best estimation of the risk, whatever the value of parameter $k$ or the level of noise in the data. The suggested optimality of L1O (for risk estimation) is also consistent with results by Burman (1989) and Celisse (2014), where it is proved that L1O is the best cross-validation procedure to perform risk estimation in the context of regression and density estimation respectively. However, note that Theorem 5.1 only provides an upper-bound of the risk, whereas a thorough analysis would rather require (at least) an asymptotic equivalent of the measure of the discrepancy between $\hat{R}_p(A_k, Z_{1,n})$ and $L(\hat{f}_k)$.

Alternatively, the LpO estimator can be also used as a data-dependent calibration procedure to choose $k$: the value $\hat{k}_p$ corresponding to the minimum LpO estimate will be selected. Although the focus of the present paper is different, it is worth mentioning that the concentration results established in Section 4 are a significant early step towards deriving theoretical guarantees on LpO as a model selection procedure. Indeed, exponential concentration inequalities have been a key ingredient to assess model selection consistency or model selection efficiency in various contexts (see for instance Celisse (2014) or Arlot and Lerasle (2012) in the density estimation framework). Still risk estimation and model selection are different objectives, and it is well known that the best estimator in terms of risk estimation can be different from the best one in terms of model selection. For instance in the regression context, L1O is known to provide the best estimator of the risk (Burman, 1989). But it leads to an inconsistent model selection procedure (Shao, 1997).

Investigating the behavior of $\hat{k}_p$ requires some further dedicated theoretical developments. One first step towards such results is to derive a tighter upper bound on the bias between the LpO estimator and the risk. The best known upper bound currently available is derived from Devroye and Wagner (1980, see Lemma D.6 in the present paper). Unfortunately it does not fully capture the true behavior of the LpO estimator with respect to $p$ (at least as $p$ becomes large) and could be improved as emphasized in the comments at the end of Section 5. Another important direction for studying the model selection behavior of the LpO procedure is to prove a concentration inequality for the classification error rate of the $k$NN classifier around its expectation. While such concentration results have been established for the $k$NN algorithm in the (fixed-design) regression framework (Arlot and Bach, 2009), deriving similar results in the classification context remains a widely open problem to the best of our knowledge.

References


S. Arlot and F. Bach. Data-driven calibration of linear estimators with minimal penalties. 


S. Arlot and A. Celisse. Segmentation of the mean of heteroscedastic data via cross-


G. Biau, F. Cérou, and A. Guyader. On the rate of convergence of the bagged nearest 

G. Biau, F. Cérou, and A. Guyader. Rates of convergence of the functional-nearest neighbor 

S. Boucheron, O. Bousquet, G. Lugosi, and P. Massart. Moment inequalities for functions 

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic 

P. Burman. Comparative study of Ordinary Cross-Validation, v-Fold Cross-Validation and 

A. Celisse. *Model selection via cross-validation in density estimation, regression 

A. Celisse. Optimal cross-validation in density estimation with the $l^2$-loss. *The Annals of 

A. Celisse and T. Mary-Huard. Exact cross-validation for knn: applications to passive and 

A. Celisse and S. Robin. Nonparametric density estimation by exact leave-p-out cross-

K. Chaudhuri and S. Dasgupta. Rates of convergence for nearest neighbor classification. In 

T. M. Cover. Rates of convergence for nearest neighbor procedures. In *Proceedings of the 

T. M. Cover and P. E. Hart. Nearest neighbor pattern classification. *Information Theory, 


Appendix A. Proofs of polynomial moment upper bounds

A.1 Proof of Theorem 2.2

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

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

A.1.1 Exploiting Proposition 2.1

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

$$
\hat{R}_p(A, Z_{1:n}) = \frac{1}{n!} \sum_{\sigma} W(Z_{\sigma(1)}, \ldots, Z_{\sigma(n)})
$$

with

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

and

$$
h_m(Z_1, \ldots, Z_m) = \frac{1}{m} \sum_{i=1}^{m} I\{A(Z_{1:m-1}^{(i)}, X_i) \neq Y_i\} = \hat{R}_1(A, Z_{1:n-p+1})
$$

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

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

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

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

$$
\mathbb{E}\left[\left|\hat{R}_p(A, Z_{1:n}) - \mathbb{E}\left[\hat{R}_p(A, Z_{1:n})\right]\right|^q\right] \leq \mathbb{E}\left[\left|\bar{W}(Z_1, \ldots, Z_n)\right|^q\right] \quad \text{(Jensen’s inequality)}
$$

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

$$
= \left[\frac{n}{m}\right]^{-q} \mathbb{E}\left[\left(\sum_{i=1}^{\left\lfloor \frac{n}{m} \right\rfloor} \bar{h}_m(Z_{(i-1)m+1}, \ldots, Z_{im})\right)^q\right].
$$
A.1.2 The setting $q = 2$

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

$$E \left[ |\hat{R}_p(A, Z_{1, n}) - E[\hat{R}_p(A, Z_{1, n})]|^q \right] \leq \left( \frac{n}{m} \right)^{-2} \text{Var} \left( \sum_{i=1}^{\lfloor \frac{n}{m} \rfloor} h_m(Z_{(i-1)m+1}, \ldots, Z_{im}) \right)$$

$$= \left( \frac{n}{m} \right)^{-2} \sum_{i=1}^{\lfloor \frac{n}{m} \rfloor} \text{Var} \left( h_m(Z_{(i-1)m+1}, \ldots, Z_{im}) \right)$$

$$= \left( \frac{n}{m} \right)^{-1} \text{Var} \left( \hat{R}_1(A, Z_{1, n-p+1}) \right),$$

which leads to the result.

A.1.3 The setting $q > 2$

If $p \leq n/2 + 1$: A straightforward use of Jensen’s inequality from (A.1) provides

$$E \left[ |\hat{R}_p(A, Z_{1, n}) - E[\hat{R}_p(A, Z_{1, n})]|^q \right] \leq \left( \frac{n}{m} \right)^{-1} \text{Var} \left( \sum_{i=1}^{\lfloor \frac{n}{m} \rfloor} h_m(Z_{(i-1)m+1}, \ldots, Z_{im}) \right)$$

$$= \left( \frac{n}{m} \right)^{-1} \text{Var} \left( \hat{R}_1(A, Z_{1, n-p+1}) - E[\hat{R}_1(A, Z_{1, n-p+1})] \right)^q.$$

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

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

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

$$E \left[ \left( \sum_{i=1}^{\lfloor n/m \rfloor} h_m(Z_{(i-1)m+1}, \ldots, Z_{im}) \right)^q \right] \leq E \left[ \left( \sum_{i=1}^{\lfloor n/m \rfloor} \zeta_i \right)^q \right],$$

which implies

$$E \left[ \left( \sum_{i=1}^{\lfloor n/m \rfloor} h_m(Z_{(i-1)m+1}, \ldots, Z_{im}) \right)^q \right] \leq B(q, \gamma) \left\{ \gamma \sum_{i=1}^{\lfloor n/m \rfloor} E[|\zeta_i|^q] \vee \left( \sum_{i=1}^{\lfloor n/m \rfloor} E[|\zeta_i^2|] \right)^q \right\}.$$
it comes
\[
\begin{align*}
\mathbb{E} \left[ \left| \sum_{i=1}^{\frac{n}{m}} \hat{h}_m \left( Z_{(i-1)m+1}, \ldots, Z_{im} \right) \right|^q \right] \\
\leq B(q, \gamma) \left( 2^q \gamma \left[ \frac{n}{m} \right] - q+1 \right) \mathbb{E} \left[ \left| \hat{R}_1(A, Z_{1,m}) - \mathbb{E} \left[ \hat{R}_1(A, Z_{1,m}) \right] \right|^q \right] \vee \\
\left( \sqrt{\left[ \frac{n}{m} \right]} \cdot \sqrt{2\text{Var} \left( \hat{R}_1(A, Z_{1,m}) \right)} \right)^q \right).
\end{align*}
\]

Hence, it results for every \( q > 2 \)
\[
\begin{align*}
\mathbb{E} \left[ \left| \hat{R}_p(A, Z_{1,n}) - \mathbb{E} \left[ \hat{R}_p(A, Z_{1,n}) \right] \right|^q \right] \\
\leq B(q, \gamma) \left( 2^q \gamma \left[ \frac{n}{m} \right] - q+1 \right) \mathbb{E} \left[ \left| \hat{R}_1(A, Z_{1,m}) - \mathbb{E} \left[ \hat{R}_1(A, Z_{1,m}) \right] \right|^q \right] \vee \\
\left( \sqrt{\left[ \frac{n}{m} \right]} \cdot \sqrt{2\text{Var} \left( \hat{R}_1(A, Z_{1,m}) \right)} \right)^q \right),
\end{align*}
\]

which concludes the proof.

A.2 Proof of Theorem 3.1

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

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

where \( h_m(Z_1, \ldots, Z_m) = \hat{R}_1(A_k(Z_{1:m}::)) \) by Eq. (2.4), and \( \hat{h}_m(Z_1, \ldots, Z_m) = h_m(Z_1, \ldots, Z_m) - \mathbb{E} \left[ h_m(Z_1, \ldots, Z_m) \right] \). The second idea consists in deriving upper bounds of

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

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

A.2.1 Upper bounding \( \Delta^j h_m \)

Let us now introduce the notation \( Z^{(i)} = \hat{Z}_{1:m}^{(i)} \) (see Theorem 2.1), and \( Z_j^{(i)} = \left( Z_{1:m}^{(i)} \right)^{(i)} \) with \( Z_{1:m}^{(i)} = \left( Z_1, \ldots, Z_j', \ldots, Z_n \right) \). Then, \( \Delta^j h_m = h_m(Z_1, \ldots, Z_m) - h_m(Z_1, \ldots, Z_j', \ldots, Z_m) \)
is now upper bounded by
\[
|\Delta^j h_m| \leq \frac{1}{m} + \frac{1}{m} \sum_{i \neq j} \left| \mathbb{1}_{\{A_k(Z^{(i)};X_i) \neq Y_i\}} - \mathbb{1}_{\{A_k(Z^{(j)};X_j) \neq Y_j\}} \right|
\]
\[
\leq \frac{1}{m} + \frac{1}{m} \sum_{i \neq j} \left| \mathbb{1}_{\{A_k(Z^{(i)};X_i) \neq A_k(Z^{(j)};X_j)\}} \right| .
\] (A.2)

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

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

\[
|\Delta^j h_m| \leq \frac{1}{m} + \frac{1}{m} \sum_{i \in B_j} \left| \mathbb{1}_{\{A_k(Z^{(i)};X_i) \neq A_k(Z^{(j)};X_j)\}} \right| .
\] (A.3)

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

**A.2.2 Case \(q > 2\)**

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

\[
|\Delta^j h_m| \leq \frac{1}{m} + \frac{1}{m} \sum_{i \in B_j} \left| \mathbb{1}_{\{A_k(Z^{(i)};X_i) \neq A_k(Z^{(j)};X_j)\}} \right| \leq \frac{1}{m} + \frac{2k\gamma_d}{m} .
\]

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

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

hence

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

This leads for every \(q > 2\) to

\[
\|h_m(Z_1, \ldots, Z_m)\|_q \leq \sqrt[2q]{2k \frac{4k\gamma_d}{\sqrt{m}}} ,
\]
which enables to conclude.
A.2.3 Case $q = 2$

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

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

using Jensen’s inequality. Lemma D.5 implies $|B_j| \leq 2k\gamma_d$, which allows to conclude

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

Summing over $j$, one derives

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

(A.4)

where $Z_0$ is an independent copy of $Z_1$, and the last but one inequality results from Lemma D.6.

A.3 Proof of Theorem 3.2

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

Proof of Ineq. (3.1), (3.2), and (3.3): These inequalities straightforwardly result from the combination of Theorem 2.2 and Ineq. (2.5) and (2.6) from Theorem 3.1.
Proof of Ineq. (3.4): It results from the upper bounds proved in Theorem 3.1 and plugged in Ineq. (2.7) (derived from Rosenthal’s inequality with optimized constant $\gamma$, namely Proposition D.3).

Then it comes

$$
\mathbb{E} \left[ \left| \hat{R}_p(A_k, Z_{1,n}) - \mathbb{E} \left[ \hat{R}_p(A_k, Z_{1,n}) \right] \right|^q \right] \leq \left( 2\sqrt{2e} \right)^q \times
$$

$$
\left\{ \left( \sqrt{q} \right)^q \left( \left[ \frac{n}{n-p+1} \right]^{-1} 2C_1 \sqrt{k} \left( \frac{\sqrt{k}}{\sqrt{n-p+1}} \right)^2 \right)^q \sqrt{q^q} \left| \frac{n}{n-p+1} \right|^{-q+1} \left( 2C_2 \sqrt{q} \right)^q \left( \frac{k}{\sqrt{n-p+1}} \right)^q \right\}
$$

$$
= \left( 2\sqrt{2e} \right)^q \times
$$

$$
\left\{ \left( \sqrt{q} \right)^q \left( \sqrt{2C_1} \sqrt{k} \left( \frac{k}{(n-p+1)} \left[ \frac{n}{n-p+1} \right] \right)^q \right) \left( q^{3/2} \right)^q \left| \frac{n}{n-p+1} \right| \left( 2C_2 \left[ \frac{n}{n-p+1} \right] \sqrt{n-p+1} \right)^q \right\}
$$

$$
\leq \left| \frac{n}{n-p+1} \right|^q \{ \left( \lambda_1 q^{1/2} \right)^q \sqrt{q^{3/2}} \} \cdot \left( \lambda_2 q^{3/2} \right)^q \}
$$

with

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

Finally introducing $\Gamma = 2\sqrt{2e} \max \left( 2C_2, \sqrt{2C_1} \right)$ provides the result.
Appendix B. Proofs of exponential concentration inequalities

B.1 Proof of Proposition 4.1

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

First, let us start by upper bounding \( |\hat{R}_p(A_k(Z_{1,n};\cdot)) - \hat{R}_p(A_k(Z_{1,n}';\cdot))| \) for every \( 1 \leq j \leq n \), where \( Z_{1,n}^j = (Z_1, \ldots, Z_{j-1}, Z_j', Z_{j+1}, \ldots, Z_n) \).

Using Eq. (2.2), one has

\[
|\hat{R}_p(A_k(Z_{1,n};\cdot)) - \hat{R}_p(A_k(Z_{1,n}';\cdot))| \\
\leq \frac{1}{p} \sum_{i=1}^{n} \sum_{e} (\frac{n}{p})^{-1} \sum_{e} |I\{A_k(Z^{e};X_i) \neq Y_i\} - I\{A_k(Z^{e';e};X_i) \neq Y_i\}| I\{i\not\in e\}
\]

\[
\leq \frac{1}{p} \sum_{i=1}^{n} \sum_{e} (\frac{n}{p})^{-1} \sum_{e} I\{A_k(Z^{e};X_i) \neq A_k(Z^{e';e};X_i)\} I\{i\not\in e\}
\]

\[
\leq \frac{1}{p} \sum_{i \neq j} \sum_{e} (\frac{n}{p})^{-1} \sum_{e} \left[ I\{j\not\in V_k(X_i)\} + I\{j\in V_k^{e';e}(X_i)\} \right] I\{i\not\in e\} + \frac{1}{p} \left(\frac{n}{p}\right)^{-1} \sum_{e} I\{j\not\in e\},
\]

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

Second, let us now introduce

\[
B_{j}^{e,n-p} = \bigcup_{e \in E_{n-p}} \left\{ 1 \leq i \leq n, \ i \not\in e \cup \{j\}, \ V_k^{i,j,e}(X_i) \ni j \text{ or } V_k^e(X_i) \ni j \right\}.
\]

Then Lemma D.5 implies \( \text{Card}(B_{j}^{e,n-p}) \leq 2(k + p - 1) \gamma_d \), hence

\[
|\hat{R}_p(A_k(Z_{1,n};\cdot)) - \hat{R}_p(A_k(Z_{1,n}';\cdot))| \leq \frac{1}{p} \sum_{e \in E_{n-p}} (\frac{n}{p})^{-1} \sum_{e} 2 \cdot I\{i\not\in e\} + \frac{1}{n}
\]

\[
\leq \frac{4(k + p - 1) \gamma_d}{n} + \frac{1}{n}.
\]

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

B.2 Proof of Theorem 4.1

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

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

To do so, our strategy is to prove a sub-Gaussian concentration inequality by use of Lemma D.2, which requires the control of the even moments of the \( L^p \) estimator \( \hat{R}_p \).

Such upper bounds are derived
• First, by using Ineq. (D.5) (generalized Efron-Stein inequality), which amounts to control the q-th moments of the differences
\[ \Delta_n^j = \hat{R}_p(A_k(Z_{1,n};\cdot)) - \hat{R}_p(A_k(Z^j_{1,n};\cdot)). \]
• Second, by precisely evaluating the contribution of each neighbor \( X_i \) of a given \( X_j \), that is by computing quantities such as \( \mathbb{P}_e \{ j \in e, \ i \in \bar{e}, \ j \in V_k^e(X_i) \} \), where \( \mathbb{P}_e[\cdot] \) denotes the probability measure with respect to the uniform random variable \( e \) over \( E_{n-p} \), and \( V_k^e(X_i) \) denotes the indices of the \( k \) nearest neighbors of \( X_i \) among \( X^e = \{ X_\ell, \ \ell \in e \} \).

**B.2.1 Upper bounding \( \Delta_n^j \)**

For every \( 1 \leq j \leq n \), one gets
\[
\Delta_n^j = \left( \frac{n}{p} \right)^{-1} \sum_e \left\{ \mathbb{1}_{\{j \in e\}} \frac{1}{p} \left( \mathbb{1}_{\{A_k(Z^e;X_j)\neq Y_j\}} - \mathbb{1}_{\{A_k(Z^e;X_j')\neq Y_j'\}} \right) + \mathbb{1}_{\{j \in e\}} \frac{1}{p} \sum_{i \in \bar{e}} \mathbb{1}_{\{A_k(Z^e;X_i)\neq A_k(Z^{e,j};X_i)\}} \right\}.
\]

Absolute values and Jensen’s inequality then provide
\[
\left| \hat{R}_p(A_k(Z_{1,n};\cdot)) - \hat{R}_p(A_k(Z^j_{1,n};\cdot)) \right|
\leq \left( \frac{n}{p} \right)^{-1} \sum_e \left\{ \mathbb{1}_{\{j \in e\}} \frac{1}{p} + \mathbb{1}_{\{j \in e\}} \frac{1}{p} \sum_{i \in \bar{e}} \mathbb{1}_{\{A_k(Z^e;X_i)\neq A_k(Z^{e,j};X_i)\}} \right\}
\leq \frac{1}{n} + \left( \frac{n}{p} \right)^{-1} \sum_e \mathbb{1}_{\{j \in e\}} \frac{1}{p} \sum_{i \in \bar{e}} \mathbb{1}_{\{A_k(Z^e;X_i)\neq A_k(Z^{e,j};X_i)\}}
\leq \frac{1}{n} + \frac{1}{p} \sum_{i=1}^n \mathbb{P}_e \left\{ j \in e, \ i \in \bar{e}, \ A_k(Z^e;X_i) \neq A_k(Z^{e,j};X_i) \right\},
\]

where \( \mathbb{P}_e \) denotes the discrete uniform probability over the set \( E_{n-p} \) of all \( n - p \) distinct indices among \( \{1, \ldots, n\} \).

Let us further notice that \( \{A_k(Z^e;X_i) \neq A_k(Z^{e,j};X_i)\} \subset \{ j \in V_k^e(X_i) \cup V_k^{e,j}(X_i) \} \), where \( V_k^{e,j}(X_i) \) denotes the set of indices of the \( k \) nearest neighbors of \( X_i \) among \( Z^{e,j} \) with the notation of the proof of Proposition 4.1. Then it results
\[
\sum_{i=1}^n \mathbb{P}_e \left\{ j \in e, \ i \in \bar{e}, \ A_k(Z^e;X_i) \neq A_k(Z^{e,j};X_i) \right\}
\leq \sum_{i=1}^n \mathbb{P}_e \left\{ j \in e, \ i \in \bar{e}, \ j \in V_k^e(X_i) \cup V_k^{e,j}(X_i) \right\}
\leq \sum_{i=1}^n \left( \mathbb{P}_e \left\{ j \in e, \ i \in \bar{e}, \ j \in V_k^e(X_i) \right\} + \mathbb{P}_e \left\{ j \in e, \ i \in \bar{e}, \ j \in V_k^e(X_i) \cup V_k^{e,j}(X_i) \right\} \right)
\leq 2 \sum_{i=1}^n \mathbb{P}_e \left\{ j \in e, \ i \in \bar{e}, \ j \in V_k^e(X_i) \right\},
\]

27
which leads to
\[ |\Delta_k| \leq \frac{1}{n} + \frac{2}{p} \sum_{i=1}^{n} \mathbb{P}_e [j \in e, \ i \in \bar{e}, \ j \in V_k^e(X_i)] . \]

Summing over \( 1 \leq j \leq n \) the square of the above quantity, it results
\[
\sum_{j=1}^{n} (\Delta_k)^2 \leq \sum_{j=1}^{n} \left\{ \frac{1}{n} + \frac{2}{p} \sum_{i=1}^{n} \mathbb{P}_e [j \in e, \ i \in \bar{e}, \ j \in V_k^e(X_i)] \right\}^2 \\
\leq 2 \sum_{j=1}^{n} \frac{1}{n^2} + 2 \left\{ \frac{2}{p} \sum_{i=1}^{n} \mathbb{P}_e [j \in e, \ i \in \bar{e}, \ j \in V_k^e(X_i)] \right\}^2 \\
\leq \frac{2}{n} + 8 \sum_{j=1}^{n} \left\{ \frac{1}{p} \sum_{i=1}^{n} \mathbb{P}_e [j \in e, \ i \in \bar{e}, \ j \in V_k^e(X_i)] \right\}^2 .
\]

### B.2.2 Evaluating the influence of each neighbor

Further using that
\[
\sum_{j=1}^{n} \left( \frac{1}{p} \sum_{i=1}^{n} \mathbb{P}_e [j \in e, \ i \in \bar{e}, \ j \in V_k^e(X_i)] \right)^2 \\
= \sum_{j=1}^{n} \frac{1}{p^2} \sum_{i=1}^{n} (\mathbb{P}_e [j \in e, \ i \in \bar{e}, \ j \in V_k^e(X_i)])^2 + \\
\sum_{j=1}^{n} \frac{1}{p^2} \sum_{1 \leq i \neq j \leq n} \mathbb{P}_e [j \in e, \ i \in \bar{e}, \ j \in V_k^e(X_i)] \mathbb{P}_e [j \in e, \ i \in \bar{e}, \ j \in V_k^e(X_i)] \\
= T1 + T2 ,
\]

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

Then Lemma D.4 leads to
\[
\sum_{j \in V_k(X_i)} \left\{ \mathbb{P}_e [j \in e, \ i \in \bar{e}, \ j \in V_k^e(X_i)] \right\}^2 + \sum_{j \in V_{k+\rho}(X_i) \setminus V_k(X_i)} \left\{ \mathbb{P}_e [j \in e, \ i \in \bar{e}, \ j \in V_k^e(X_i)] \right\}^2 \\
\leq \sum_{j \in V_k(X_i)} \left( \frac{p \ n - p}{n \ n - 1} \right)^2 + \sum_{j \in V_{k+\rho}(X_i) \setminus V_k(X_i)} \mathbb{P}_e [j \in e, \ i \in \bar{e}, \ j \in V_k^e(X_i)] \frac{p \ n - p}{n \ n - 1} \\
= k \left( \frac{p \ n - p}{n \ n - 1} \right)^2 + \frac{kp \ p - 1}{p \ n - p} \frac{p \ n - p}{n \ n - 1} \frac{p \ n - p}{n \ n - 1} = k \left( \frac{p}{n} \right)^2 \frac{n - p}{n - 1} ,
\]

28
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

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

**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 \( \ell \) in the whole sample. Then,

\[
T2 = \frac{1}{p^2} \sum_{j=1}^{n} \sum_{\ell \neq i \ j \in V_k(X_\ell)} \mathbb{P}_e [ j \in e, \ i \in \bar{e}, \ j \in V_k^e(X_i) ] \mathbb{P}_e [ j \in e, \ \ell \in \bar{e}, \ j \in V_k^e(X_\ell) ]
\leq \frac{1}{p^2} \sum_{i=1}^{n} \sum_{\ell \neq i} \sum_{j \in V_k(X_\ell)} \mathbb{P}_e [ j \in e, \ i \in \bar{e}, \ j \in V_k^e(X_i) ] \frac{p n - p}{n n - 1}
\]

\[
+ \frac{1}{p^2} \sum_{i=1}^{n} \sum_{\ell \neq i} \sum_{j \in \bigcap_{k=p}^{p-1} V_k(X_\ell) \setminus V_k(X_\ell)} \mathbb{P}_e [ j \in e, \ i \in \bar{e}, \ j \in V_k^e(X_i) ] \frac{k p - 1}{n n - 1}.
\]

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

\[
T2 = \frac{1}{p^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{P}_e [ j \in e, \ i \in \bar{e}, \ j \in V_k^e(X_i) ] \left( \sum_{\ell \neq i} \mathbb{1}_{1 \in V_k(X_\ell)} \frac{p n - p}{n n - 1} + \sum_{\ell \neq i} \mathbb{1}_{1 \in V_k+p(X_i) \setminus V_k(X_i)} \frac{k p - 1}{n n - 1} \right)
\]

\[
\leq \frac{1}{p^2} \sum_{i=1}^{n} \frac{k p}{n} \left( k \gamma_d \frac{p n - p}{n n - 1} + (k + p) \gamma_d \frac{k p - 1}{n n - 1} \right) = \gamma_d \frac{k^2}{n} \left( \frac{n - p}{n - 1} + (k + p) \frac{p - 1}{n - 1} \right)
\]

\[
= \gamma_d \frac{k^2}{n} \left( 1 + (k + p - 1) \frac{p - 1}{n - 1} \right).
\]

**Gathering the upper bounds** The two previous bounds provide

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

\[
\leq \frac{k n - p}{n n - 1} + \gamma_d \frac{k^2}{n} \left( 1 + (k + p - 1) \frac{p - 1}{n - 1} \right),
\]

which enables to conclude

\[
\sum_{j=1}^{n} \left( \hat{R}_p(A_k(Z_{1,n}; \cdot)) - \hat{R}_p(A_k(Z_{1,n}^j; \cdot)) \right)^2
\]

\[
\leq \frac{2}{n} \left( 1 + 4k + 4k^2 \gamma_d \left[ 1 + (k + p) \frac{p - 1}{n - 1} \right] \right) \leq \frac{8k^2 (1 + \gamma_d)}{n} \left[ 1 + (k + p) \frac{p - 1}{n - 1} \right].
\]
B.2.3 Generalized Efron-Stein inequality

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

\[
\left\| \hat{R}_p(A_k, Z_{1,n}) - \mathbb{E} \left[ \hat{R}_p(A_k, Z_{1,n}) \right] \right\|_2^q \leq 4\sqrt{\kappa q} \sqrt{\frac{8(1 + \gamma d)k^2}{n} \left[ 1 + \frac{(k + p)\frac{p - 1}{n - 1}}{n} \right]}.
\]

Hence combined with \( q! \geq q^qe^{-q/2\pi q} \), it comes

\[
\mathbb{E} \left[ \left( \hat{R}_p(A_k, Z_{1,n}) - \mathbb{E} \left[ \hat{R}_p(A_k, Z_{1,n}) \right] \right)^{2q} \right] \leq (16\kappa q)^q \left( \frac{8(1 + \gamma d)k^2}{n} \left[ 1 + \frac{(k + p)\frac{p - 1}{n - 1}}{n} \right] \right)^q \\
\leq q^q \left( 16\kappa \frac{8(1 + \gamma d)k^2}{n} \left[ 1 + \frac{(k + p)\frac{p - 1}{n - 1}}{n} \right] \right)^q.
\]

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

\[
P \left( \hat{R}_p(A_k, Z_{1,n}) - \mathbb{E} \left[ \hat{R}_p(A_k, Z_{1,n}) \right] > t \right) \vee P \left( \mathbb{E} \left[ \hat{R}_p(A_k, Z_{1,n}) \right] - \hat{R}_p(A_k, Z_{1,n}) > t \right)
\leq \exp \left( -\frac{nt^2}{1024\kappa e^{8(1 + \gamma d)k^2(1 + (k + p)\frac{p - 1}{n - 1})}} \right).
\]

B.3 Proof of Theorem 4.2 and Proposition 4.2

B.3.1 Proof of Theorem 4.2

If \( p < n/2 + 1 \):

In what follows, we exploit a characterization of sub-Gaussian random variables by their \( 2q \)-th moments (Lemma D.2).

From (3.1) 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 \( q \geq 1 \)

\[
\mathbb{E} \left[ \left| \hat{R}_p(A_k, Z_{1,n}) - \mathbb{E} \left[ \hat{R}_p(A_k, Z_{1,n}) \right] \right|^{2q} \right] \leq \left( \frac{\Delta^2}{16e} \frac{k^2}{n - p + 1} \right)^q \leq \left( \frac{\Delta^2}{8} \frac{k^2}{n - p + 1} \right)^q q^q,
\]

with \( q^q \leq q^qe^{q/\sqrt{2\pi q}}. \) Then Lemma D.2 provides for every \( t > 0 \)

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

If \( p \geq n/2 + 1 \):

This part of the proof relies on Proposition D.1 which provides an exponential concentration inequality from upper bounds on the moments of a random variable.
Let us now use (3.1) and (3.4) combined with (D.1), where $C = \left\lceil \frac{n}{n-p+1} \right\rceil$, $q_0 = 2$, and $\min_j \alpha_j = 1/2$. This provides for every $t > 0$

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

where $\Gamma$ arises from Eq. (3.4).

**B.3.2 Proof of Proposition 4.2**

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

With the same notation and reasoning as in the previous proof, let us combine (3.1) and (3.4). From (D.2) of Proposition D.1 where $C = \left\lceil \frac{n}{n-p+1} \right\rceil$, $q_0 = 2$, and $\min_j \alpha_j = 1/2$, it results for every $t > 0$

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

where $\Gamma > 0$ is given by Eq. (3.4).
Appendix C. Proofs of deviation upper bounds

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

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

Along the proof, we will repeatedly use some notation that we briefly introduce here.

First, let us 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: $\hat{f}_k(X_0) = A_k(Z_{1,n}; X_0)$, and $\hat{f}_k(e, X_0) = A_k(Z_{1,n}; X_0)$ 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 $P_e$ (resp. $P_{e,e'}$) 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

$$\mathbb{E} \left[ (\hat{R}_p(A_k, Z_{1,n}) - L(A_k, Z_{1,n}))^2 \right] = \mathbb{E} \left[ \hat{R}_p^2(A_k, Z_{1,n}) \right] + \mathbb{E} \left[ L_n^2 \right] - 2 \mathbb{E} \left[ \hat{R}_p(A_k, Z_{1,n}) L(A_k, Z_{1,n}) \right],$$

let us notice that

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

and

$$\mathbb{E} \left[ \hat{R}_p(A_k, Z_{1,n}) L(A_k, Z_{1,n}) \right] = P \left( \hat{f}_k(X_0) \neq Y_0, \hat{f}_k(e, X_i) \neq Y_i | i \notin e \right) P_e (i \notin e).$$

It immediately comes

$$\mathbb{E} \left[ (\hat{R}_p(A_k, Z_{1,n}) - L(A_k, Z_{1,n}))^2 \right]$$

$$= \mathbb{E} \left[ \hat{R}_p^2(A_k, Z_{1,n}) \right] - P \left( \hat{f}_k(X_0) \neq Y_0, \hat{f}_k(e, X_i) \neq Y_i | i \notin e \right) P_e (i \notin e)$$

$$+ \left[ P \left( \hat{f}_k(X_0) \neq Y_0, \hat{f}_k(X_{n+1}) \neq Y_{n+1} \right) - P \left( \hat{f}_k(X_0) \neq Y_0, \hat{f}_k(e, X_i) \neq Y_i | i \notin e \right) P_e (i \notin e) \right].$$

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

$$p^2 \mathbb{E} \left[ \hat{R}_p^2(A_k, Z_{1,n}) \right] = \sum_{i,j} \mathbb{E} \left[ \mathbb{I}_{\hat{f}_k(e, X_i) \neq Y_i} \mathbb{I}_{i \notin e} \mathbb{I}_{\hat{f}_k(e', X_j) \neq Y_j} \mathbb{I}_{j \notin e} \right]$$

$$= \sum_i \mathbb{E} \left[ \mathbb{I}_{\hat{f}_k(e, X_i) \neq Y_i} \mathbb{I}_{i \notin e} \mathbb{I}_{\hat{f}_k(e', X_j) \neq Y_j} \mathbb{I}_{j \notin e} \right]$$

$$+ \sum_{i \neq j} \mathbb{E} \left[ \mathbb{I}_{\hat{f}_k(e, X_i) \neq Y_i} \mathbb{I}_{i \notin e} \mathbb{I}_{\hat{f}_k(e', X_j) \neq Y_j} \mathbb{I}_{j \notin e} \right].$$
Let us now introduce the following events.

\[ A_{e,e',i} = \{ i \not= e, j \not= e' \}, \]
\[ A_{e,e',i,j}^1 = \{ i \not= e, j \not= e', i \not= e', j \not= e \}, \quad A_{e,e',i,j}^2 = \{ i \not= e, j \not= e', i \not= e', j \in e \}, \]
\[ A_{e,e',i,j}^3 = \{ i \not= e, j \not= e', i \in e', j \not= e \}, \quad A_{e,e',i,j}^4 = \{ i \not= e, j \not= e', i \in e', j \in e \}. \]

Then,

\[
p^2 \mathbb{E} \left[ \hat{R}_p^2(A_k, Z_{1,n}) \right] = \sum_i \mathbb{P} ( \hat{f}_k(e, X_i) \neq Y_i, \hat{f}_k(e', X_i) \neq Y_i | \hat{A}_{e,e'}(i) ) \mathbb{P}_{e,e'} ( \hat{A}_{e,e'}(i) ) \\
+ \sum_{i \neq j} 4 \sum_{\ell=1} \mathbb{P} ( \hat{f}_k(e, X_i) \neq Y_i, \hat{f}_k(e', X_i) \neq Y_i | \hat{A}_{e,e',i,j}^\ell ) \mathbb{P}_{e,e'} ( \hat{A}_{e,e',i,j}^\ell ) \\
= n \mathbb{P} ( \hat{f}_k(e, X_1) \neq Y_1, \hat{f}_k(e', X_1) \neq Y_1 | \hat{A}_{e,e',1} ) \mathbb{P}_{e,e'} ( \hat{A}_{e,e',1} ) \\
+ n(n-1) \sum_{\ell=1} 4 \mathbb{P} ( \hat{f}_k(e, X_1) \neq Y_1, \hat{f}_k(e', X_2) \neq Y_2 | \hat{A}_{e,e',1,2}^\ell ) \mathbb{P}_{e,e'} ( \hat{A}_{e,e',1,2}^\ell ).
\]

Furthermore since

\[
\frac{1}{p^2} \left[ n \mathbb{P}_{e,e'} ( \hat{A}_{e,e',1} ) + n(n-1) \sum_{\ell=1} 4 \mathbb{P}_{e,e'} ( \hat{A}_{e,e',1,2}^\ell ) \right] = \frac{1}{p^2} \sum_{i,j} \mathbb{P}_{e,e'} ( i \not= e, j \not= e' ) = 1,
\]

it comes

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

where

\[
A = \left[ \mathbb{P} ( \hat{f}_k(e, X_1) \neq Y_1, \hat{f}_k(e', X_1) \neq Y_1 | \hat{A}_{e,e',1} ) \\
- \mathbb{P} ( \hat{f}_k(X_0) \neq Y_0, \hat{f}_k(e, X_1) \neq Y_1 | \hat{A}_{e,e',1} ) \right] \mathbb{P}_{e,e'} ( \hat{A}_{e,e',1} ) ,
\]

and

\[
B = \sum_{\ell=1} 4 \left[ \mathbb{P} ( \hat{f}_k(e, X_1) \neq Y_1, \hat{f}_k(e', X_2) \neq Y_2 | \hat{A}_{e,e',1,2}^\ell ) \\
- \mathbb{P} ( \hat{f}_k(X_0) \neq Y_0, \hat{f}_k(e, X_1) \neq Y_1 | \hat{A}_{e,e',1,2}^\ell ) \right] \mathbb{P}_{e,e'} ( \hat{A}_{e,e',1,2}^\ell ).
\]

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

\[
A \leq \mathbb{P}_{e,e'} ( \hat{A}_{e,e',i} ) \leq \mathbb{P}_{e,e'} ( i \not= e, i \not= e' ) \leq \left( \frac{p}{n} \right)^2
\]

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

\[
\mathbb{P} ( \hat{f}_k(e, X_1) \neq Y_1, \hat{f}_k(e', X_2) \neq Y_2 | \hat{A}_{e,e',1,2}^\ell ) - \mathbb{P} ( \hat{f}_k(X_0) \neq Y_0, \hat{f}_k(e, X_1) \neq Y_1 | \hat{A}_{e,e',1,2}^\ell ) \quad \text{(C.4)}
\]
which depends on \( \ell \), i.e. on the fact that index 2 belongs or not to the training set indices \( e \).

- If \( 2 \not\in e \) (i.e. \( \ell = 1 \) or \( 3 \)): Then, Lemma C.2 proves

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

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

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

Combining the previous bounds and Lemma C.1 leads to

\[
B \leq \left( \frac{4p\sqrt{k}}{\sqrt{2\pi n}} \right) \left[ \mathbb{P}_{e,e'}(A_{e,e',1,2}^1) + \mathbb{P}_{e,e'}(A_{e,e',1,2}^3) \right]
+ \left( \frac{8\sqrt{k}}{\sqrt{2\pi(n-p)}} + \frac{4p\sqrt{k}}{\sqrt{2\pi n}} \right) \left[ \mathbb{P}_{e,e'}(A_{e,e',1,2}^2) + \mathbb{P}_{e,e'}(A_{e,e',1,2}^4) \right]
\leq \frac{2\sqrt{2}}{\sqrt{\pi}} \mathbb{P}_{e,e'} \left( \frac{p}{n} \right) \left[ \mathbb{P}_{e,e'}(A_{e,e',1,2}^1) + \mathbb{P}_{e,e'}(A_{e,e',1,2}^3) \right]
+ \left( \frac{2}{n-p} + \frac{p}{n} \right) \left[ \mathbb{P}_{e,e'}(A_{e,e',1,2}^2) + \mathbb{P}_{e,e'}(A_{e,e',1,2}^4) \right]
\leq \frac{2\sqrt{2}}{\sqrt{\pi}} \mathbb{P}_{e,e'} \left( \frac{p}{n} \right) \left[ \frac{p}{n} \mathbb{P}_{e,e'}(i \not\in e, j \not\in e') + \frac{2}{n-p} \mathbb{P}_{e,e'}(A_{e,e',1,2}^2) + \mathbb{P}_{e,e'}(A_{e,e',1,2}^4) \right]
\leq \frac{2\sqrt{2}}{\sqrt{\pi}} \mathbb{P}_{e,e'} \left( \frac{p}{n} \right)^2 \left[ \frac{2}{n-p} \left( \frac{(n-p)p^2(p-1)}{n^2(n-1)^2} + \frac{(n-p)^2p^2}{n^2(n-1)^2} \right) \right]
\leq \frac{2\sqrt{2}}{\sqrt{\pi}} \mathbb{P}_{e,e'} \left( \frac{p}{n} \right)^2 \left[ \frac{2}{n-p} + \frac{2}{n-1} \right].
\]

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

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

**Upper bound of (C.2)** First observe that

\[
\mathbb{P} \left( \hat{f}_k(X_0) \neq Y_0, \hat{f}_k(e, X_i) \neq Y_i \mid i \notin e \right) = \mathbb{P} \left( \hat{f}_k^{-1}(X_0) \neq Y_0, \hat{f}_k(e, X_{n+1}) \neq Y_{n+1} \right)
\]
where $\hat{f}_k^{(-1)}$ is built on sample $(X_2, Y_2), \ldots, (X_{n+1}, Y_{n+1})$. One has

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

$$
= \mathbb{P}\left(\hat{f}_k(X_0) \neq Y_0, \hat{f}_k(X_{n+1}) \neq Y_{n+1}\right) - \mathbb{P}\left(\hat{f}_k^{(-1)}(X_0) \neq Y_0, \hat{f}_k(e, X_{n+1}) \neq Y_{n+1}\right)
$$

$$
\leq \mathbb{P}\left(\hat{f}_k(X_0) \neq \hat{f}_k^{(-1)}(X_0)\right) + \mathbb{P}\left(\hat{f}_k(e, X_{n+1}) \neq \hat{f}_k(X_{n+1})\right)
$$

$$
\leq \frac{4\sqrt{k}}{\sqrt{2\pi n}} + \frac{4p\sqrt{k}}{\sqrt{2\pi n}},
$$

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

**Conclusion:**

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

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

**C.1.2 Combinatorial lemmas**

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

**Lemma C.1.**

$$
P_{e,e'}^{A_1^{1}}(A_{e,e',1,2}) = \frac{n-2}{n-p} \frac{n-2}{n-p} \frac{n}{n-p}, \quad P_{e,e'}^{A_2^{1}}(A_{e,e',i,j}) = \frac{n-p-1}{n-2} \frac{n-p-1}{n-2} \frac{n}{n-p},
$$

$$
P_{e,e'}^{A_3^{3}}(A_{e,e',i,j}) = \frac{n-p}{n-p} \frac{n-p-1}{n-p} \frac{n}{n-p}, \quad P_{e,e'}^{A_4^{4}}(A_{e,e',i,j}) = \frac{n-p-1}{n-p} \frac{n-p-1}{n-p} \frac{n}{n-p}.
$$
Proof of Lemma C.1.

\[
\begin{align*}
\mathbb{P}_{e,e'} (A_{e,e',1,2}^1) &= \mathbb{P}_{e,e'} (i \notin e, j \notin e', i \notin e', j \notin e) \\
&= \mathbb{P}_{e,e'} (i \notin e, j \notin e) \mathbb{P}_{e,e'} (j \notin e', i \notin e') \\
&= \frac{\binom{n-2}{n-p}}{\binom{n}{n-p}} \times \frac{\binom{n-2}{n-p}}{\binom{n}{n-p}}.
\end{align*}
\]

\[
\begin{align*}
\mathbb{P}_{e,e'} (A_{e,e',i,j}^2) &= \mathbb{P}_{e,e'} (i \notin e, j \notin e', i \notin e', j \notin e) \\
&= \mathbb{P}_{e,e'} (i \notin e, j \notin e) \mathbb{P}_{e,e'} (j \notin e', i \notin e') \\
&= \frac{\binom{n-p-1}{n-2}}{\binom{n}{n-p}} \times \frac{\binom{n-p-1}{n-2}}{\binom{n}{n-p}}.
\end{align*}
\]

\[
\begin{align*}
\mathbb{P}_{e,e'} (A_{e,e',i,j}^3) &= \mathbb{P}_{e,e'} (i \notin e, j \notin e', i \notin e', j \notin e) \\
&= \mathbb{P}_{e,e'} (i \notin e, j \notin e) \mathbb{P}_{e,e'} (j \notin e', i \notin e') \\
&= \frac{\binom{n-p-1}{n-2}}{\binom{n}{n-p}} \times \frac{\binom{n-p-1}{n-2}}{\binom{n}{n-p}}.
\end{align*}
\]

\[
\begin{align*}
\mathbb{P}_{e,e'} (A_{e,e',i,j}^4) &= \mathbb{P}_{e,e'} (i \notin e, j \notin e', i \notin e', j \notin e) \\
&= \mathbb{P}_{e,e'} (i \notin e, j \notin e) \mathbb{P}_{e,e'} (j \notin e', i \notin e') \\
&= \frac{\binom{n-p-1}{n-2}}{\binom{n}{n-p}} \times \frac{\binom{n-p-1}{n-2}}{\binom{n}{n-p}}.
\end{align*}
\]

□

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

\[
\mathbb{P} \left( \hat{f}_k (e, X_1) \neq Y_1, \hat{f}_k (e', X_2) \neq Y_2 \mid A_{e,e',1,2} \right) - \mathbb{P} \left( \hat{f}_k (X_0) \neq Y_0, \hat{f}_k (e, X_1) \neq Y_1 \mid A_{e,e',1,2} \right) \leq \frac{4p\sqrt{K}}{\sqrt{2\pi n}}.
\]

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

\[
\mathbb{P} \left( \hat{f}_k (X_0) \neq Y_0, \hat{f}_k (e, X_1) \neq Y_1 \mid A_{e,e',1,2} \right)
\]

is

\[
\mathbb{P} \left( \hat{f}_k (X_0) \neq Y_0, \hat{f}_k (e, X_1) \neq Y_1 \mid A_{e,e',1,2}, 0 \notin e, 0 \notin e' \right).
\]

Then one has

\[
\begin{align*}
\mathbb{P} \left( \hat{f}_k (X_0) \neq Y_0, \hat{f}_k (e, X_1) \neq Y_1 \mid A_{e,e',1,2} \right) &= \mathbb{P} \left( \hat{f}_k^{(-2)} (X_2) \neq Y_2, \hat{f}_k (e, X_1) \neq Y_1 \mid A_{e,e',1,2}, 0 \notin e, 0 \notin e' \right)
\end{align*}
\]

36
where \( \hat{f}_k^{(-2)} \) is built on sample \((X_0, Y_0), (X_1, Y_1), (X_3, Y_3), \ldots, (X_n, Y_n)\). Hence

\[
P\left(\hat{f}_k(e, X_1) \neq Y_1, \hat{f}_k(e', X_2) \neq Y_2 \mid A_{e,e',1,2}^{\ell} \right) - P\left(\hat{f}_k(X_0) \neq Y_0, \hat{f}_k(e, X_1) \neq Y_1 \mid A_{e,e',1,2}^{\ell} \right)
= P\left(\hat{f}_k(e, X_1) \neq Y_1, \hat{f}_k(e', X_2) \neq Y_2 \mid A_{e,e',1,2}^{\ell}, 0 \notin e, 0 \notin e' \right)
- P\left(\hat{f}_k^{(-2)}(X_2) \neq Y_2, \hat{f}_k(e, X_1) \neq Y_1 \mid A_{e,e',1,2}^{\ell}, 0 \notin e, 0 \notin e' \right)
\leq P\left(\{\hat{f}_k(e, X_1) \neq Y_1\} \triangle \{\hat{f}_k(e, X_1) \neq Y_1\} \mid A_{e,e',1,2}^{\ell}, 0 \notin e, 0 \notin e' \right)
+ P\left(\{\hat{f}_k^{(-2)}(X_2) \neq Y_2\} \triangle \{\hat{f}_k^{(-2)}(X_2) \neq Y_2\} \mid A_{e,e',1,2}^{\ell}, 0 \notin e, 0 \notin e' \right)
= P\left(\hat{f}_k^{(-2)}(X_2) \neq \hat{f}_k(e', X_2) \mid A_{e,e',1,2}^{\ell} \right) \leq \frac{4p\sqrt{k}}{\sqrt{2\pi n}},
\]
by Lemma D.6.

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

\[
P\left(\hat{f}_k(e, X_1) \neq Y_1, \hat{f}_k(e', X_2) \neq Y_2 \mid A_{e,e',1,2}^{\ell} \right) - P\left(\hat{f}_k(X_0) \neq Y_0, \hat{f}_k(e, X_1) \neq Y_1 \mid A_{e,e',1,2}^{\ell} \right)
\leq \frac{8\sqrt{k}}{\sqrt{2\pi(n - p)}} + \frac{4p\sqrt{k}}{\sqrt{2\pi n}}.
\]

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

\[
P\left(\hat{f}_k(X_0) \neq Y_0, \hat{f}_k(e, X_1) \neq Y_1 \mid A_{e,e',1,2}^{\ell} \right) = P\left(\hat{f}_k^{(-2)}(X_2) \neq Y_2, \hat{f}_k^{e_0}(X_1) \neq Y_1 \mid A_{e,e',1,2}^{\ell} \right),
\]
where \( \hat{f}_k^{e_0} \) is built on sample \( e \) with observation \((X_2, Y_2)\) replaced with \((X_0, Y_0)\). Then

\[
P\left(\hat{f}_k(e, X_1) \neq Y_1, \hat{f}_k(e', X_2) \neq Y_2 \mid A_{e,e',1,2}^{\ell} \right) - P\left(\hat{f}_k(X_0) \neq Y_0, \hat{f}_k(e, X_1) \neq Y_1 \mid A_{e,e',1,2}^{\ell} \right)
= P\left(\hat{f}_k(e, X_1) \neq Y_1, \hat{f}_k(e', X_2) \neq Y_2 \mid A_{e,e',1,2}^{\ell} \right) - P\left(\hat{f}_k^{(-2)}(X_2) \neq Y_2, \hat{f}_k^{e_0}(X_1) \neq Y_1 \mid A_{e,e',1,2}^{\ell} \right)
\leq P\left(\{\hat{f}_k(e, X_1) \neq Y_1\} \triangle \{\hat{f}_k^{e_0}(X_1) \neq Y_1\} \mid A_{e,e',1,2}^{\ell} \right)
+ P\left(\{\hat{f}_k^{(-2)}(X_2) \neq Y_2\} \triangle \{\hat{f}_k^{(-2)}(X_2) \neq Y_2\} \mid A_{e,e',1,2}^{\ell} \right)
= P\left(\hat{f}_k(e, X_1) \neq \hat{f}_k^{e_0}(X_1) \mid A_{e,e',1,2}^{\ell} \right) + P\left(\hat{f}_k^{(-2)}(X_2) \neq \hat{f}_k(e', X_2) \mid A_{e,e',1,2}^{\ell} \right)
\leq \frac{8\sqrt{k}}{\sqrt{2\pi(n - p)}} + \frac{4p\sqrt{k}}{\sqrt{2\pi n}}.
\]
Appendix D. Technical results

D.1 Main inequalities

D.1.1 From moment to exponential inequalities

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

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

Then for every $t > 0$,

$$
P[|X| > t] \leq Ce^{\theta_0 \min_j \alpha_j} e^{-(\min_i \alpha_i) e^{-1} \min_j \left\{ \left( \frac{t}{N\lambda_j} \right)^{\frac{1}{\alpha_j}} \right\}},
$$

(D.1)

Furthermore for every $x > 0$, it results

$$
P\left[ |X| > \sum_{i=1}^{N} \lambda_i \left( \frac{e x}{\min_j \alpha_j} \right)^{\alpha_i} \right] \leq Ce^{\theta_0 \min_j \alpha_j} \cdot e^{-x}.
$$

(D.2)

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

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

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

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

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

which provides (D.1).

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

$$
\sum_{i=1}^{N} \frac{\lambda_i (q^*)^{\alpha_i}}{t^*} = \sum_{i=1}^{N} \lambda_i \left( \frac{e^{-1} \frac{ex}{\min_j \alpha_j}}{\min_j \alpha_j} \right)^{\alpha_i} \leq \min_k e^{-\alpha_k} \sum_{i=1}^{N} \lambda_i \left( \frac{ex}{\min_j \alpha_j} \right)^{\alpha_i} = e^{-\min_k \alpha_k}.
$$
Performance of CV to estimate the risk of \(k\)NN

Then,
\[
C \left( \sum_{i=1}^{N} \lambda_i (q^*)^{\alpha_i} \right)^{q^*} \leq Ce^{-(\min_k \alpha_k) \min_j \alpha_j} = Ce^{-x}.
\]

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

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

\[\square\]

D.1.2 Sub-Gaussian random variables

**Lemma D.1** (Theorem 2.1 in Boucheron et al. (2013) first part). Any centered random variable \(X\) such that \(\mathbb{P}(X > t) \vee \mathbb{P}(-X > t) \leq e^{-t^2/(2\nu)}\) satisfies
\[
\mathbb{E} \left[ X^{2q} \right] \leq q! (4\nu)^q.
\]

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

**Lemma D.2** (Theorem 2.1 in Boucheron et al. (2013) second part). Any centered random variable \(X\) such that
\[
\mathbb{E} \left[ X^{2q} \right] \leq q! C^q.
\]

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

D.1.3 The Efron-Stein inequality

**Theorem D.1** (Efron-Stein’s inequality Boucheron et al. (2013), Theorem 3.1). Let \(X_1, \ldots, X_n\) be independent random variables and let \(Z = f(X_1, \ldots, X_n)\) be a square-integrable function. Then
\[
\text{Var}(Z) \leq \sum_{i=1}^{n} \mathbb{E} \left[ (Z - \mathbb{E}[Z \mid (X_j)_{j \neq i}])^2 \right] = \nu.
\]

Moreover if \(X'_1, \ldots, X'_n\) denote independent copies of \(X_1, \ldots, X_n\) and if we define for every \(1 \leq i \leq n\)
\[
Z_i' = f(X_1, \ldots, X'_i, \ldots, X_n),
\]

then
\[
\nu = \frac{1}{2} \sum_{i=1}^{n} \mathbb{E} \left[ (Z - Z_i')^2 \right].
\]
D.1.4 Generalized Efron-Stein’s inequality

**Theorem D.2** (Theorem 15.5 in Boucheron et al. (2013)). 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 $V_+ = \mathbb{E} \left[ \sum_{i=1}^{n} [(Z - Z'_i)^+]^2 \mid X^n_1 \right]$ and $V_- = \mathbb{E} \left[ \sum_{i=1}^{n} [(Z - Z'_i)^-]^2 \mid X^n_1 \right]$. Then there exists a constant $\kappa \leq 1,271$ such that for all $q \in [2, +\infty]$,

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

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

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

$$
\|Z - \mathbb{E} Z\|_q \leq \sqrt{2\kappa q \sqrt{\| \sum_{i=1}^{n} (Z - Z'_i)^2 \|_{q/2}} \leq \sqrt{4\kappa q \sqrt{\| \sum_{i=1}^{n} (Z - \mathbb{E} Z) (|X_j \mid j \neq i) \|_{q/2}}.}
$$

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

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

**Proof of Corollary D.1.**

First note that

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

Consequently,

$$
\|Z - \mathbb{E} Z\|_q^q \leq \sqrt{2\kappa q} \left( \sqrt{\| V_+ \|_{q/2}^q} + \sqrt{\| V_- \|_{q/2}^q} \right)
\leq \sqrt{2\kappa q} \left( \| V_+ \|_{q/2}^{q/2} + \| V_- \|_{q/2}^{q/2} \right)
\leq \sqrt{2\kappa q} \left\| \sum_{i=1}^{n} \mathbb{E} \left[ (Z - Z'_i)^2 \mid X^n_1 \right] \right\|_{q/2}^{q/2}.
$$

Besides,

$$
\mathbb{E} \left[ (Z - Z'_i)^2 \mid X^n_1 \right] = \mathbb{E} \left[ (Z - \mathbb{E} [Z \mid (X_j)_{j \neq i}]) + \mathbb{E} [Z \mid (X_j)_{j \neq i}] - Z'_i)^2 \mid X^n_1 \right]
= \mathbb{E} \left[ (Z - \mathbb{E} [Z \mid (X_j)_{j \neq i}])^2 + \mathbb{E} [Z \mid (X_j)_{j \neq i}] - Z'_i)^2 \mid X^n_1 \right]
= \mathbb{E} \left[ (Z - \mathbb{E} [Z \mid (X_j)_{j \neq i}])^2 \mid X^n_1 \right] + \mathbb{E} \left[ (\mathbb{E} [Z \mid (X_j)_{j \neq i}] - Z'_i)^2 \mid X^n_1 \right].
$$
Performance of CV to estimate the risk of kNN

Combining the two previous results leads to

\[ \|Z - \mathbb{E}Z\|_q \leq \sqrt{2\kappa q} \left( \sum_{i=1}^n (Z - \mathbb{E}[Z | (X_j)_{j \neq i}])^2 \right)^{q/2} + \sqrt{2\kappa q} \left( \sum_{i=1}^n \mathbb{E}[\mathbb{E}[Z' | (X_j)_{j \neq i}] - Z'_i]^2 | X_i^q] \right)^{q/2} \]

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

\[ \Box \]

D.1.5 McDiarmid’s inequality

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

\[ \sup_{x_1, \ldots, x_n, x_i} |f(x_1, \ldots, x_i, \ldots, x_n) - f(x_1, \ldots, x'_i, \ldots, x_n)| \leq c_i, \ 1 \leq i \leq n . \]

Then for all \( \varepsilon > 0 \), one has

\[ P(f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)] \geq \varepsilon) \leq e^{\frac{-2\varepsilon^2}{\sum_{i=1}^n c_i^2}} \]

\[ P(\mathbb{E}[f(X_1, \ldots, X_n)] - f(X_1, \ldots, X_n) \geq \varepsilon) \leq e^{\frac{-2\varepsilon^2}{\sum_{i=1}^n c_i^2}} \]

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

D.1.6 Rosenthal’s inequality

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

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

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

\[ B^*(q, \gamma) = 1 + \frac{\mathbb{E}[|N|^q]}{\mathbb{E}[|Z-Z'|^q]} \]

\[ = \gamma^{-q/(q-1)} \mathbb{E}[|Z-Z'|^q] , \text{ if } 4 < q, \]

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

41
Proposition D.3. Let $X_1, \ldots, X_n$ denote independent real random variables with symmetric distributions. Then for every $q > 2$,

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

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

- if $2 < q \leq 4$,
  
  $$
  B^*(q, \gamma) \leq \left( 2\sqrt[2]{2e\sqrt[q]{q}} \right)^q,
  $$
  
  by choosing $\gamma = 1$.

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

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

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

which leads to the conclusion.

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

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

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

$$
B^*(q, \gamma) = 1 + \frac{E \left[ |N|^q \right]}{\gamma} \leq 1 + \frac{\sqrt[2]{2e\sqrt[q]{q}} \left( \frac{q}{2} \right)^{\frac{q}{2}}}{\gamma} \leq 1 + \frac{\sqrt[2]{2e\sqrt[q]{q}} \left( \frac{q}{2} \right)^{\frac{q}{2}}}{\gamma} = 1 + \frac{\left( \sqrt[2]{2e\sqrt[q]{q}} \right)^q}{\gamma},
$$
Performance of CV to estimate the risk of \( k \)NN by use of Lemma D.9 and \( \sqrt{q^{1/q}} \leq \sqrt{e} \) for every \( q > 2 \).

If \( q > 4 \),

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

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

\[ \square \]

D.2 Technical lemmas

D.2.1 Basic computations for resampling applied to the \( k \)NN algorithm

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

\[
\mathbb{P}_e (i \in \bar{e}) = \frac{p}{n} \quad \text{(D.6)} \\
\sum_{j=1}^{n} \mathbb{P}_e \left[ i \in \bar{e}, j \in V_k^e(X_i) \right] = \frac{kp}{n} \quad \text{(D.7)}
\]

In the same way,

\[
\sum_{k < \sigma_1(j) \leq k+p} \mathbb{P}_e \left[ i \in \bar{e}, j \in V_k^e(X_i) \right] = \frac{kp}{n} \frac{p-1}{n-1} \quad \text{(D.8)}
\]

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

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

For the last equality, let us notice every \( j \in V_k \) 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
\[
\sum_{k<\sigma_i(j)\leq k+p} \mathbb{P}_e[i \in \bar{e}, j \in V_k^e(X_i)] = \sum_{j=1}^n \mathbb{P}_e[i \in \bar{e}, j \in V_k^e(X_i)] - \sum_{\sigma_i(j)\leq k} \mathbb{P}_e[i \in \bar{e}, j \in V_k^e(X_i)] \\
= kp - k \frac{n-1}{n-p} = k \frac{p-1}{n-1}.
\]

\[\square\]

D.2.2 Stone’s lemma

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

D.2.3 Stability of the \(k\)NN classifier when removing \(p\) observations

Lemma D.6 (Devroye and Wagner (1979), Eq. (14)). For every \(1 \leq k \leq n\), let \(A_k\) denote \(k\)-NN classification algorithm defined by Eq. (2.1), and let \(Z_1, \ldots, Z_n\) denote \(n\) i.i.d. random variables such that for every \(1 \leq i \leq n\), \(Z_i = (X_i, Y_i) \sim P\). Then for every \(1 \leq p \leq n - k\),
\[
\mathbb{P}[A_k(Z_{1:n}; X) \neq A_k(Z_{1:n-p}; X)] \leq \frac{4p\sqrt{k}}{\sqrt{2\pi n}},
\]
where \(Z_{1:i} = (Z_1, \ldots, Z_i)\) for every \(1 \leq i \leq n\), and \((X, Y) \sim P\) is independent of \(Z_{1:n}\).

D.2.4 Exponential concentration inequality for the L1O estimator

Lemma D.7 (Devroye et al. (1996), Theorem 24.4). For every \(1 \leq k \leq n\), let \(A_k\) denote \(k\)-NN classification algorithm defined by Eq. (2.1). Let also \(\hat{R}(\cdot)\) denote the L1O estimator defined by Eq. (2.2) with \(p = 1\). Then for every \(\varepsilon > 0\),
\[
\mathbb{P}\left(\left|\hat{R}(A_k, Z_{1:n}) - \mathbb{E}\left[\hat{R}(A_k, Z_{1:n})\right]\right| > \varepsilon\right) \leq 2 \exp\left\{-n\frac{\varepsilon^2}{2(k\gamma_d^2)}\right\}.
\]

D.2.5 Moment upper bounds for the L1O estimator

Lemma D.8. For every \(1 \leq k \leq n\), let \(A_k\) denote \(k\)-NN classification algorithm defined by Eq. (2.1). Let also \(\hat{R}(\cdot)\) denote the L1O estimator defined by Eq. (2.2) with \(p = 1\). Then for every \(q \geq 1\),
\[
\mathbb{E}\left[\left(\hat{R}(A_k, Z_{1:n}) - \mathbb{E}\left[\hat{R}(A_k, Z_{1:n})\right]\right)^2\right]^{q/2} \leq q! \left(\frac{2(k\gamma_d^2)}{n}\right)^q.
\]

The proof is straightforward from the combination of Lemmas D.1 and D.7.
D.2.6 Upper bound on the optimal constant in the Rosenthal’s inequality

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

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

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

$$
\mathbb{E}[|N|^q] = 2 \int_0^{+\infty} x^q e^{-\frac{x^2}{2}} \, dx = \sqrt{\frac{2}{\pi}}(q-1) \int_0^{+\infty} x^{q-2} e^{-\frac{x^2}{2}} \, dx
$$

$$
= \sqrt{\frac{2}{\pi}} \frac{(q-1)!}{2^{k-1}(k-1)!} = \sqrt{\frac{2}{\pi}} \frac{q!}{2^{k-1}(k-1)!}.
$$

Then using for any positive integer $a$

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

it results

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

which implies

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

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

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

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

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

\[\square\]

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

$$
\mathbb{E}[|S - \mathbb{E}[S]|^q] \leq 4\sqrt{e}\sqrt{q} \sqrt{\frac{kq^q}{2e}}.
$$
**Proof of Lemma D.10.** Since $S - \mathbb{E}(S)$ is symmetric, it comes

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

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

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

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

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

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

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

Let us notice

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

and also that

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

This implies

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

hence

$$\mathbb{E}[|S - \mathbb{E}[S]|^q] \leq 2\sqrt{\frac{k}{q}} 2\sqrt{\pi e} \sqrt{q} \left(\frac{q}{e}\right)^{q/2} = 4\sqrt{\pi e} \sqrt{\frac{qk^q}{2e}}.$$
Lemma D.11. Let $X, Y$ be two i.i.d. random variables with Poisson distribution $\mathcal{P}(\lambda)$ ($\lambda > 0$). Then for every $q > 3$, it comes

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

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

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

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

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

which entails that

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

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

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

Finally, one concludes

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

$\blacksquare$