Stability revisited: new generalisation bounds for the Leave-one-Out

Alain Celisse, Benjamin Guedj

Stability revisited: new generalisation bounds for the Leave-one-Out

Alain Celisse[*]
Université de Lille & Inria
Lille, France
alain.celisse@math.univ-lille1.fr

Benjamin Guedj[†]
Inria
Lille, France
benjamin.guedj@inria.fr

Abstract

The present paper provides a new generic strategy leading to non-asymptotic theoretical guarantees on the Leave-one-Out procedure applied to a broad class of learning algorithms. This strategy relies on two main ingredients: the new notion of L^q stability, and the strong use of moment inequalities. L^q stability extends the ongoing notion of hypothesis stability while remaining weaker than the uniform stability. It leads to new PAC exponential generalisation bounds for Leave-one-Out under mild assumptions. In the literature, such bounds are available only for uniform stable algorithms under boundedness for instance. Our generic strategy is applied to the Ridge regression algorithm as a first step.

1 Introduction

A massive variety of learning algorithms rely upon unknown parameters that crucially influence the final statistical performance (such as Lasso, Ridge, ...). Cross-validation (CV) procedures are among the most popular data-driven approaches used to assess the performance of estimators, and calibrate their unknown parameters. We refer for instance to Arlot and Celisse [2010] for a survey on CV procedures. Among them, the Leave-one-Out [LoO, Stone, 1974] procedure is fairly intuitive, hence widely used. Yet its popularity contrasts with the few theoretical results often available in specific settings and derived at the price of strong assumptions [such as boundedness in Bousquet and Elisseeff, 2002, Example 3].

The present paper has the ambition to provide non-asymptotic theoretical guarantees on the LoO procedure. We propose a generic strategy to consistently analyse the LoO estimator for learning algorithms. This strategy is based on two ingredients: stability and moment inequalities, which provide concentration results when combined [see for example Boucheron et al., 2013, for an extensive review]. Such concentration results are then precious to derive generalisation bounds, *i.e.*, upper bounds on the discrepancy between the LoO estimator and its prediction error (see for instance McDiarmid, 1989 and Devroye, 1991 for papers in that direction).

The notion of stability has first been introduced by Devroye and Wagner [1979] and further studied for instance by Kearns and Ron [1999] and Bousquet and Elisseeff [2002]. This concept has emerged as an effective measure of the "smoothness" of a learning algorithm with respect to its input data. For an introduction to stability and connections with other topics such as reproducibility, see Yu [2013]. Over the past decades, the use of stability to derive generalisation bounds has received much attention in the statistical and machine learning community. Existing results rely upon stability assumptions such as the *hypothesis* or *uniform stability*. For instance, hypothesis stability is used by Devroye and Wagner [1979, Eq. (7)] to derive an upper bound of order 2 moments of LoO for the k-nearest

[*] http://math.univ-lille1.fr/~celisse/
[†] https://bguedj.github.io

neighbors algorithm. The stronger uniform stability [Bousquet and Elisseeff, 2002, Definition 6] enables to provide a PAC exponential bound for the LoO estimator.

Further insightful analyses of various notions of stability can be found in Kutin and Niyogi [2002], Evgeniou et al. [2004], Elisseeff et al. [2005], Rakhlin et al. [2005], Mukherjee et al. [2006], Shalev-Shwartz et al. [2010], Kale et al. [2011], Kumar et al. [2013] and Villa et al. [2013] to name but a few.

Our main contributions. The present paper introduces a generic strategy to derive new generalisation bounds for the LoO estimator applied to a broad family of learning algorithms. This strategy relies on: (i) a new stability assumption that generalises the existing hypothesis stability [Bousquet and Elisseeff, 2002, Definition 3] while remaining weaker than uniform stability [Bousquet and Elisseeff, 2002, Definition 6], and (ii) moment inequalities. Combining those two ingredients leads to PAC generalisation bounds. For the sake of brevity, we illustrate this strategy by focusing on the Ridge regression algorithm. As part of our contributions, we develop a thorough analysis of the LoO estimator applied to the Ridge regression and obtain generalisation bounds under L^q stability (Theorem 3 and Theorem 4) matching state-of-the-art results earlier established by Bousquet and Elisseeff [2002, Example 3] under the stronger notion of uniform stability. Let us stress though that the proposed strategy is in no way limited to this algorithm and calls for future work to extend it to other algorithms.

The paper is organised as follows: Section 2 contains our notion of L^q stability for learning algorithms. In particular, we provide an upper bound on the L^q stability of the Ridge regression algorithm (Theorem 1). Section 3 establishes generalisation bounds in terms of moment inequalities for LoO (Theorem 2). This allows for PAC exponential generalisation bounds in Section 4, which is the main achievement of the paper. Specific results in Theorem 3 and Theorem 4 for the Ridge regression algorithm are also provided. The paper closes with some perspectives in Section 5, and Appendix A wraps up technical results.

2 Stability of learning algorithms

The main purpose of the present section is to introduce a generalisation of the notion of L^1 stability, also called *hypothesis stability* [Devroye and Wagner, 1979], to the higher order L^q stability with $q \geq 2$. In particular this new notion turns out to be useful to derive PAC generalisation bounds for the LoO estimator of various learning algorithms (see Section 3 and Section 4).

2.1 Framework and notation

In what follows, $\mathcal{A}$ denotes a learning algorithm (see Section 2.3 for examples). From a training sample $\mathcal{D} = (Z_1, \ldots, Z_n) \in (\mathcal{X} \times \mathcal{Y})^n$ of n *independent and identically distributed* random variables with $Z_i = (X_i, Y_i) \sim P$ (unknown), $\mathcal{A}$ outputs an estimator $\mathcal{A}(\mathcal{D}) : \mathcal{X} \subset \mathbb{R}^d \to \mathcal{Y} \subset \mathbb{R}$. Here we only consider *symmetric* algorithms, *i.e.*, $\mathcal{A}$ does not depend on the order of the sample points $Z_1, \ldots, Z_n$. We also assume that for any $1 \leq i \leq n$,

Assumptions. To ease the reading of what follows, we provide simplified results under the following (somewhat restrictive) assumptions:

- *Boundedness of X:*
  Let us assume that there exists $0 < B_X < +\infty$ such that
  $$\forall\, 1 \leq i \leq n, \quad |X_i|_2 \leq B_X, \ a.s. \tag{XBd}$$
- *Boundedness of Y:*
  Let us assume that there exists $0 < B_Y < +\infty$ such that
  $$\forall\, 1 \leq i \leq n, \quad |Y_i| \leq B_Y, \ a.s. \tag{YBd}$$
- *Sub-Gaussianity of Y:*
  Let us assume there exists $v > 0$ such that
  $$\forall\, 1 \leq i \leq n, \quad \|Y_i - \mathbb{E}[Y_i]\|_q \leq 2e\sqrt{v}\sqrt{q}. \tag{SubG}$$

However let us emphasize that most of the forthcoming results can be extended at the price of additional technicalities to the unbounded case (at least for instance to the Gaussian setting for X).

The performance of algorithm $\mathcal{A}$ trained from $\mathcal{D}$ and evaluated at point X is $c(\mathcal{A}(\mathcal{D}, X), Y)$, where $c(\cdot, \cdot) : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ is a *cost function*. The *prediction error* of the estimator $\mathcal{A}(D)$ is the random variable depending on $\mathcal{D}$ given by

$$\mathcal{L}_P\left(\mathcal{A}(\mathcal{D})\right) = \mathbb{E}_{(X,Y)\sim P}\left[\, c\left(\mathcal{A}(\mathcal{D}, X), Y\right)\right]. \tag{1}$$

In the sequel we let $|\cdot|$ denote the absolute value in $\mathbb{R}$, $|\cdot|_2$ the Euclidean norm in $\mathbb{R}^d$, $\|\cdot\|_{op}$ the operator norm over the set of $d \times d$ matrices $\mathcal{M}_d(\mathbb{R})$, and $\|\cdot\|_q$ the $L^q(\mathbb{P})$-norm for any $q \geq 1$ where $\mathbb{P}$ is a reference probability, *i.e.*, $\|U\|_q = \left(\int |U|^q \, d\mathbb{P}\right)^{1/q}$ for any real-valued random variable U.

2.2 A new notion of stability: L^q stability

Our purpose is to bridge the gap between the weak notion of L^1 stability [Bousquet and Elisseeff, 2002, Definition 3], which only provides PAC polynomial generalisation bounds [Bousquet and Elisseeff, 2002, Section 4.1], and the stronger notion of uniform stability [Bousquet and Elisseeff, 2002, Definition 6], which leads to PAC exponential bounds yet may appear too restrictive [Kutin and Niyogi, 2002, Section 3.1].

To this end the following definition generalises the L^1 stability to higher order moments.

Definition 1 (L^q stability)**.** *Let $\mathcal{A}$ denote any symmetric learning algorithm, and $c(\cdot, \cdot)$ be any cost function. Then for every $q \geq 1$, $\mathcal{A}$ is said γ_q-L^q stable if there exists $\gamma_q > 0$ such that*

$$\forall\, 1 \leq j \leq n, \quad \mathcal{S}_q(\mathcal{A}, n)^q = \mathbb{E}\left[\, |c\left(\mathcal{A}(\mathcal{D}, X), Y\right) - c\left(\mathcal{A}(\tau_j(\mathcal{D}), X), Y\right)|^q \,\right] \leq \gamma_q^q,$$

where the expectation is computed over $\mathcal{D}$ and $(X, Y) \sim P$, with (X, Y) independent of $\mathcal{D}$, and $\tau_j(\mathcal{D}) = (Z_1, \ldots, Z_{j-1}, Z_{j+1}, \ldots, Z_n)$ is the sample $\mathcal{D}$ where $Z_j = (X_j, Y_j)$ has been removed.

The above Definition 1 requires to bound the variation of $\mathcal{A}$ induced by removing one training point. This is in accordance with earlier definitions [Devroye and Wagner, 1979, Bousquet and Elisseeff, 2002 and Evgeniou et al., 2004]. However, controlling high order moments provides more information on the distribution of $c\left(\mathcal{A}(\mathcal{D}, X), Y\right)$ than simply considering hypothesis stability, that is L^q stability with $q = 1$. Let us also mention that other notions of stability have been introduced, which replace one training point by an independent copy [Kutin and Niyogi, 2002, Kale et al., 2011 and Kumar et al., 2013]. Finally let us emphasise that uniform stability obviously implies L^q stability for every $q \geq 1$.

2.3 Instances of stable learning algorithms

We now illustrate how the L^q stability notion translates onto two learning algorithms: the k-nearest neighbors and the Ridge regression algorithms [Friedman et al., 2009, Section 13.3 and Section 3.4].

The k-nearest neighbors algorithm. For $1 \leq k \leq n-1$, let $V_k(x)$ be the set of indices of the k nearest neighbors (kNN) of x among $X_1, \ldots, X_n$. For binary classification, the kNN classifier is

$$\mathcal{A}_k(\mathcal{D}; x) = \begin{cases} 1 & \text{if } \sum_{j=1}^n Y_j \mathbb{1}_{\{j \in V_k(x)\}} \geq k/2, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Proposition 1 (Devroye and Wagner [1979], Eq. (14))**.** *With the above notation, for every $1 \leq k \leq n-1$, $\mathcal{A}_k$ is γ_1-L^1 stable for the $0-1$ cost function $c(y, y') = \mathbb{1}_{\{y \neq y'\}}$ with*

$$\gamma_1 = \frac{4}{\sqrt{2\pi}} \frac{\sqrt{k}}{n} \cdot$$

Proof of Proposition 1. For every $1 \leq j \leq n$, and using Lemma 5 for the last inequality,

$$\begin{aligned} \mathbb{E}\left[\, |c\left(\mathcal{A}(\mathcal{D})(X), Y\right) - c\left(\mathcal{A}(\tau_j(\mathcal{D}))(X), Y\right)|\,\right] &= \mathbb{E}\left[\, \left|\mathbb{1}_{\{\mathcal{A}_k(\mathcal{D})(X) \neq Y\}} - \mathbb{1}_{\{\mathcal{A}_k(\tau_j(\mathcal{D}))(X) \neq Y\}}\right|\,\right] \\ &= \mathbb{P}\left[\, \mathcal{A}_k(\mathcal{D})(X) \neq \mathcal{A}_k(\tau_j(\mathcal{D}))(X)\,\right] \leq \frac{4}{\sqrt{2\pi}} \frac{\sqrt{k}}{n} \cdot \end{aligned}$$

□

The Ridge regression algorithm. Let us recall that for any $\lambda > 0$, the Ridge estimator is given by

$$\mathcal{A}_\lambda(\mathcal{D}) = \underset{\beta \in \mathbb{R}^d}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \langle X_i, \beta \rangle_{\mathbb{R}^d} \right)^2 + \lambda |\beta|_2^2 \right\} = \frac{1}{n} \left(\widehat{\Sigma} + \lambda I_d \right)^{-1} X^T Y, \tag{3}$$

where $\widehat{\Sigma} = X^T X / n = 1/n \sum_{i=1}^{n} X_i X_i^T$ denotes the empirical covariance matrix. Then,

Theorem 1. *For any sample size $n > 1$, $\eta \in (0,1)$, and $\lambda > [\eta(n-1)]^{-1}$, let $\mathcal{A}_\lambda$ be given by Eq. (3) and set $c(y,y') = (y-y')^2$. Then, assuming (**XBd**), $\mathcal{A}_\lambda$ is γ_q-L^q stable for any $q \geq 1$ with*

$$\gamma_q = 2 \|Y\|_{2q}^2 \frac{B_X^2}{n\lambda} \left(1 + \frac{B_X^2 + \lambda}{\lambda(1-\eta)} \right) \left(1 + \frac{B_X^2}{\lambda} \right),$$

where $\gamma_q = +\infty$, if $\|Y\|_{2q} = +\infty$.

The L^q stability holds with the Ridge algorithm under the very mild assumption that Y admits finite moments of order q for some $q \geq 1$. Unsurprisingly, the stronger assumption of Bousquet and Elisseeff [2002, Example 3] leads to a similar upper bound in terms of L^q stability.

The proof of Theorem 1 relies on the two following technical lemmas.

Lemma 1. *With the above notation, let us define $\eta \in (0,1)$ and n satisfy $n\eta > 1$. If (**XBd**) holds true, then for every $\lambda > B_X^2 (n\eta - 1)^{-1}$, it results*

$$|\mathcal{A}_\lambda(\mathcal{D}) - \mathcal{A}_\lambda(\tau_j(\mathcal{D}))|_2 \leq \frac{B_X}{n\lambda} \left(|Y_j| + \frac{B_X^2 + \lambda}{\lambda(1-\eta)} \left[\frac{1}{n-1} \sum_{i \neq j} |Y_i| \right] \right).$$

Proof of Lemma 1. Set $\widehat{\Sigma}^{(j)} = 1/(n-1) \sum_{i \neq j} X_i X_i^T$.

$$\begin{aligned} |\mathcal{A}_\lambda(\mathcal{D}) - \mathcal{A}_\lambda(\tau_j(\mathcal{D}))|_2 &\leq \left| \frac{1}{n} \left(\widehat{\Sigma} + \lambda I_d \right)^{-1} \left[X^T Y - (X^{(j)})^T Y^{(j)} \right] \right|_2 \\ &\quad + \left| \left[\frac{1}{n} \left(\widehat{\Sigma} + \lambda I_d \right)^{-1} - \frac{1}{n-1} \left(\widehat{\Sigma}^{(j)} + \lambda I_d \right)^{-1} \right] (X^{(j)})^T Y^{(j)} \right|_2 \\ &= T_1 + T_2. \end{aligned}$$

First, Lemma 7 provides

$$T_1 = \left| \frac{1}{n} \left(\widehat{\Sigma} + \lambda I_d \right)^{-1} \left[X^T Y - (X^{(j)})^T Y^{(j)} \right] \right|_2 \leq \frac{1}{n} \left\| \left(\widehat{\Sigma} + \lambda I_d \right)^{-1} \right\|_{op} |X_j Y_j|_2 \leq \frac{1}{n\lambda} |Y_j| \, |X_j|_2 .$$

Then, (**XBd**) yields

$$T_1 \leq \frac{1}{n\lambda} |Y_j| \, B_X. \tag{4}$$

Second, it is straightforward to observe that

$$T_2 \leq \frac{1}{n} \left\| \left(\widehat{\Sigma} + \lambda I_d \right)^{-1} - \left(\frac{n-1}{n} \widehat{\Sigma}^{(j)} + \lambda I_d - \frac{1}{n} \lambda I_d \right)^{-1} \right\|_{op} \left| \sum_{i \neq j} X_i Y_i \right|_2 .$$

Further, writing $\frac{n-1}{n} \widehat{\Sigma}^{(j)} + \lambda I_d = \widehat{\Sigma} + \lambda I_d + \frac{n-1}{n} \widehat{\Sigma}^{(j)} - \widehat{\Sigma}$, we obtain

$$\left(\widehat{\Sigma} + \lambda I_d \right)^{-1} - \left(\frac{n-1}{n} \widehat{\Sigma}^{(j)} + \lambda I_d - \frac{1}{n} \lambda I_d \right)^{-1} = \left(\widehat{\Sigma} + \lambda I_d \right)^{-1} - \left(\widehat{\Sigma} + \lambda I_d + B_j \right)^{-1},$$

with $B_j = \frac{n-1}{n} \widehat{\Sigma}^{(j)} - \widehat{\Sigma} - \frac{1}{n} \lambda I_d = - \left(X_j X_j^T + \lambda I_d \right) / n$. Then, Lemma 6 and Lemma 7 yield

$$\left\| \left(\widehat{\Sigma} + \lambda I_d \right)^{-1} - \left(\frac{n-1}{n} \widehat{\Sigma}^{(j)} + \lambda I_d - \frac{1}{n} \lambda I_d \right)^{-1} \right\|_{op}$$

$$\leq \left\| \left(\widehat{\Sigma} + \lambda I_d \right)^{-1} \right\|_{op}^2 \left\| \left(X_j X_j^T + \lambda I_d \right) / n \right\|_{op} \left\| \left(I_d + B_j \left(\widehat{\Sigma} + \lambda I_d \right)^{-1} \right)^{-1} \right\|_{op}$$
$$\leq \frac{|X_j|_2^2 + \lambda}{n\lambda^2} \left\| \left(I_d + B_j \left(\widehat{\Sigma} + \lambda I_d \right)^{-1} \right)^{-1} \right\|_{op}.$$

Further assuming that for every $\eta \in (0,1)$, $n > \eta^{-1}$ and $\lambda > \frac{B_X^2}{n\eta - 1}$, then Lemma 8 and (**XBd**) lead to

$$\left\| \left(I_d + B_j \left(\widehat{\Sigma} + \lambda I_d \right)^{-1} \right)^{-1} \right\|_{op} \leq \left(1 - \frac{\|B_j\|_{op}}{\lambda} \right)^{-1} \leq \left(1 - \frac{B_X^2 + \lambda}{n\lambda} \right)^{-1} \leq (1-\eta)^{-1}.$$

Using (**XBd**), this allows to conclude that

$$T_2 \leq \frac{B_X^2 + \lambda}{n\lambda^2} \frac{B_X}{1-\eta} \left[\frac{1}{n-1} \sum_{i \neq j} |Y_i| \right]. \tag{5}$$

Finally gathering Eq. (4) and (5), one obtains

$$|\mathcal{A}_\lambda(\mathcal{D}) - \mathcal{A}_\lambda(\tau_j(\mathcal{D}))|_2 \leq \frac{B_X}{n\lambda} \left(|Y_j| + \frac{B_X^2 + \lambda}{\lambda(1-\eta)} \left[\frac{1}{n-1} \sum_{i \neq j} |Y_i| \right] \right).$$

□

Lemma 2. *With the above notation, let us define $\eta \in (0,1)$ and n satisfy $n\eta > 1$. If (**XBd**) holds true, then for every $\lambda > B_X^2 (n\eta - 1)^{-1}$, one has*

$$|2Y - \mathcal{A}_\lambda(\mathcal{D}; X) - \mathcal{A}_\lambda(\tau_j(\mathcal{D}); X)| \leq 2|Y| + B_X \left(\frac{1}{n\lambda} \sum_{i=1}^n |Y_i| + \frac{1}{(n-1)\lambda} \sum_{i \neq j} |Y_i| \right).$$

Proof of Lemma 2. Combining the Cauchy-Schwarz inequality with (**XBd**) yields

$$\begin{aligned} |2Y - \mathcal{A}_\lambda(\mathcal{D}; X) - \mathcal{A}_\lambda(\tau_j(\mathcal{D}); X)| &\leq 2|Y| + |< X, \mathcal{A}_\lambda(\mathcal{D}) >_{\mathbb{R}^d}| + |< X, \mathcal{A}_\lambda(\tau_j(\mathcal{D})) >_{\mathbb{R}^d}| \\ &\leq 2|Y| + B_X \left(|\mathcal{A}_\lambda(\mathcal{D})|_2 + |\mathcal{A}_\lambda(\tau_j(\mathcal{D}))|_2 \right) \\ &\leq 2|Y| + B_X \left(\frac{1}{n\lambda} \sum_{i=1}^n |Y_i| + \frac{1}{(n-1)\lambda} \sum_{i \neq j} |Y_i| \right), \end{aligned}$$

since Eq. (3) and (**XBd**) imply $|\mathcal{A}_\lambda(\mathcal{D})|_2 \leq \frac{1}{n\lambda} \left| \sum_{i=1}^n X_i Y_i \right|_2 \leq B_X / (n\lambda) \sum_{i=1}^n |Y_i|$.

□

Proof of Theorem 1. With $c(t(x), y) = (t(x) - y)^2$, any $q \geq 1$, the Cauchy-Schwarz inequality provides

$$\begin{aligned} \mathcal{S}_q(\mathcal{A}_\lambda(\mathcal{D})) &= \|c(\mathcal{A}_\lambda(\mathcal{D}; X), Y) - c(\mathcal{A}_\lambda(\tau_j(\mathcal{D}); X), Y)\|_q \\ &\leq \|\mathcal{A}_\lambda(\mathcal{D}; X) - \mathcal{A}_\lambda(\tau_j(\mathcal{D}); X)\|_{2q} \|2Y - \mathcal{A}_\lambda(\mathcal{D}; X) - \mathcal{A}_\lambda(\tau_j(\mathcal{D}); X)\|_{2q}, \end{aligned}$$

since $c(a, y) - c(b, y) = (a - b)(a + b - 2y)$.

Another use of the Cauchy-Schwarz inequality combined with $\mathcal{A}_\lambda(\mathcal{D}; X) = \langle \mathcal{A}_\lambda(\mathcal{D}), X \rangle_2$ and the independence between (X, Y) and $\mathcal{D}$ leads to

$$\begin{aligned} \|\mathcal{A}_\lambda(\mathcal{D}; X) - \mathcal{A}_\lambda(\tau_j(\mathcal{D}); X)\|_{2q} &\leq \||X|_2\|_{2q} \times \left\| |\mathcal{A}_\lambda(\mathcal{D}) - \mathcal{A}_\lambda(\tau_j(\mathcal{D}))|_2 \right\|_{2q} \\ &\leq B_X \times \left\| |\mathcal{A}_\lambda(\mathcal{D}) - \mathcal{A}_\lambda(\tau_j(\mathcal{D}))|_2 \right\|_{2q}. \end{aligned} \tag{6}$$

Let us notice that Lemma 1 implies

$$\left\| |\mathcal{A}_\lambda(\mathcal{D}) - \mathcal{A}_\lambda(\tau_j(\mathcal{D}))|_2 \right\|_{2q} \leq \frac{B_X}{n\lambda}\left(\|Y_j\|_{2q} + \frac{B_X^2+\lambda}{\lambda(1-\eta)} \left\| \frac{1}{n-1}\sum_{i\neq j} |Y_i| \right\|_{2q} \right)$$
$$\leq \frac{B_X}{n\lambda}\left(1 + \frac{B_X^2+\lambda}{\lambda(1-\eta)}\right)\|Y\|_{2q}, \tag{7}$$

where Y denotes an independent copy of the Y_is.

Likewise, Lemma 2 results in

$$\|2Y - \mathcal{A}_\lambda(\mathcal{D};X) - \mathcal{A}_\lambda(\tau_j(\mathcal{D});X)\|_{2q}$$
$$\leq 2\|Y\|_{2q} + B_X^2\left(\left\| \frac{1}{n\lambda}\sum_{i=1}^n |Y_i| \right\|_{2q} + \left\| \frac{1}{(n-1)\lambda}\sum_{i\neq j} |Y_i| \right\|_{2q} \right).$$

Since the Y_is are identically distributed, the triangular inequality gives

$$\|2Y - \mathcal{A}_\lambda(\mathcal{D};X) - \mathcal{A}_\lambda(\tau_j(\mathcal{D});X)\|_{2q} \leq 2\|Y\|_{2q}\left(1 + \frac{B_X^2}{\lambda}\right). \tag{8}$$

By combining Eq. (6), (7) and (**XBd**), we obtain

$$\|c(\mathcal{A}_\lambda(\mathcal{D};X),Y) - c(\mathcal{A}_\lambda(\tau_j(\mathcal{D});X),Y)\|_q$$
$$\leq \frac{B_X^2}{n\lambda}\left(1 + \frac{B_X^2+\lambda}{\lambda(1-\eta)}\right)\|Y\|_{2q} \times 2\|Y\|_{2q}\left(1+\frac{B_X^2}{\lambda}\right)$$
$$\leq 2\|Y\|_{2q}^2 \frac{B_X^2}{n\lambda}\left(1 + \frac{B_X^2+\lambda}{\lambda(1-\eta)}\right)\left(1+\frac{B_X^2}{\lambda}\right).$$

□

3 Deriving moment generalisation bounds from L^q stability

Let us recall that the goal is to upper bound with high probability the discrepancy between the LoO estimator and the prediction error $\widehat{R}_1(\mathcal{A},\mathcal{D}) - \mathcal{L}_P(\mathcal{A}(\mathcal{D}))$. We now derive generalisation bounds in terms of moments for the LoO estimator of some learning algorithm $\mathcal{A}$. Recalling that the LoO estimator associated with $\mathcal{A}$ is

$$\widehat{R}_1(\mathcal{A},\mathcal{D}) = \frac{1}{n}\sum_{i=1}^n c(\mathcal{A}(\tau_i(\mathcal{D}),X_i),Y_i),$$

where $\tau_i(\mathcal{D}) = (Z_1,\ldots,Z_{i-1},Z_{i+1},\ldots,Z_n)$, we now provide the main result of this section. It upper bounds the moments of $\widehat{R}_1(\mathcal{A},\mathcal{D}) - \mathcal{L}_P(\mathcal{A}(\mathcal{D}))$. The main steps of its proof follow Corollary 1 below.

Theorem 2. *For any sample size $n \geq 2$ and any symmetric learning algorithm $\mathcal{A}$, let $\widehat{R}_1(\mathcal{A},\mathcal{D})$ and $\mathcal{L}_P(\mathcal{A}(\mathcal{D}))$ respectively denote the LoO estimator and the prediction error (see Eq. (1)). Then there exists a numerical constant $0 < \kappa \leq 1.271$ such that, for any $q \geq 2$,*

$$\left\| \widehat{R}_1(\mathcal{A},\mathcal{D}) - \mathcal{L}_P(\mathcal{A}(\mathcal{D})) - \left(\mathbb{E}\left[\mathcal{L}_P(\mathcal{A}(\mathcal{D}))\right] - \mathbb{E}\left[\widehat{R}_1(A,\mathcal{D})\right] \right) \right\|_q$$
$$\leq \sqrt{\kappa q n}\left[\sqrt{2}\,\mathcal{S}_q(\mathcal{A},n) + 4\,\mathcal{S}_q(\mathcal{A},n-1) \right] + \frac{2\sqrt{\kappa q}}{\sqrt{n}} \left\| c(\mathcal{A}(\tau_j(\mathcal{D}),X_j),Y_j) - c(\mathcal{A}(\tau_j(\mathcal{D}),X_j'),Y_j') \right\|_q,$$

where $\mathcal{S}_q(\mathcal{A},n)$ is given by Definition 1.

This result implies that $\widehat{R}_1(\mathcal{A},\mathcal{D})$ is a consistent estimator of $\mathcal{L}_P(\mathcal{A}(\mathcal{D}))$ in $L^q(\mathbb{P})$ provided that $\mathcal{S}_q(\mathcal{A},n) = o(1/\sqrt{n})$ as $n \to +\infty$. For instance this holds true for the Ridge estimator with any $q \geq 2$ (Theorem 1).

The proof of Theorem 2 heavily relies on the following generalisation of the Efron-Stein inequality [see Bousquet and Elisseeff, 2002, Theorem 1 for Efron-Stein inequality].

Proposition 2 (Boucheron et al. [2013], Celisse and Mary-Huard [2015]). *Let $X_1, \dots, X_n$ denote n independent random variables and $Z = f(X_1, \dots, X_n)$, where $f : \mathbb{R}^n \to \mathbb{R}$ is any Borel function. With $Z'_j = f(X_1, \dots, X'_j, \dots, X_n)$, where $X'_1, \dots, X'_n$ are independent copies of the X_is, there exists a universal constant $\kappa \leq 1.271$ such that for any $q \geq 2$,*

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

Proposition 2 is a concentration result of Z around its expectation. This suggests a strategy to prove Theorem 2 that is based on the triangular inequality and the successive control of $\widehat{R}_1(\mathcal{A}, \mathcal{D})$ and $\mathcal{L}_P\left(\mathcal{A}(\mathcal{D})\right)$ around their expectations. This is done in the following two lemmas.

Lemma 3. *With the above notation, for any $q \geq 2$,*

$$\begin{aligned}
&\left\| \widehat{R}_1(\mathcal{A}, \mathcal{D}) - \mathbb{E}\left[\widehat{R}_1(\mathcal{A}, \mathcal{D}) \right] \right\|_q \\
&\leq 2\sqrt{\kappa q} \left[2\sqrt{n}\, \mathcal{S}_q(\mathcal{A}, n-1) + \frac{1}{\sqrt{n}} \left\| c(\mathcal{A}(\tau_j(\mathcal{D}), X_j), Y_j) - c(\mathcal{A}(\tau_j(\mathcal{D}), X'_j), Y'_j) \right\|_q \right].
\end{aligned}$$

Proof of Lemma 3. First, note that Proposition 2 gives

$$\left\| \widehat{R}_1(\mathcal{A}, \mathcal{D}) - \mathbb{E}\left[\widehat{R}_1(\mathcal{A}, \mathcal{D}) \right] \right\|_q \leq \sqrt{2\kappa q} \sqrt{\left\| \sum_{j=1}^{n} \left(\widehat{R}_1(\mathcal{A}, \mathcal{D}) - \widehat{R}_1(\mathcal{A}, \mathcal{D}^j) \right)^2 \right\|_{q/2}},$$

where $\mathcal{D}^j = (Z_1, \dots, Z'_j, \dots, Z_n)$ and Z'_j is an independent copy of Z_j for any j. Moreover,

$$\begin{aligned}
&\left\| \sum_{j=1}^{n} \left(\widehat{R}_1(\mathcal{A}, \mathcal{D}) - \widehat{R}_1(\mathcal{A}, \mathcal{D}^j) \right)^2 \right\|_{q/2} \\
&\leq \sum_{j=1}^{n} 2 \left(\frac{1}{n} \sum_{i \neq j} \left\| c(\mathcal{A}(\tau_i(\mathcal{D}), X_i), Y_i) - c(\mathcal{A}(\tau_i(\mathcal{D}^j), X_i), Y_i) \right\|_q \right)^2 \\
&\quad + \frac{2}{n} \left\| c(\mathcal{A}(\tau_j(\mathcal{D}), X_j), Y_j) - c(\mathcal{A}(\tau_j(\mathcal{D}), X'_j), Y'_j) \right\|_q^2 \\
&= 2n \left(\frac{n-1}{n} \right)^2 4\, \mathcal{S}_q^2(\mathcal{A}, n-1) + \frac{2}{n} \left\| c(\mathcal{A}(\tau_j(\mathcal{D}), X_j), Y_j) - c(\mathcal{A}(\tau_j(\mathcal{D}), X'_j), Y'_j) \right\|_q^2 \\
&\leq 8n\, \mathcal{S}_q^2(\mathcal{A}, n-1) + \frac{2}{n} \left\| c(\mathcal{A}(\tau_j(\mathcal{D}), X_j), Y_j) - c(\mathcal{A}(\tau_j(\mathcal{D}), X'_j), Y'_j) \right\|_q^2,
\end{aligned}$$

and the proof ends by taking the square root on both sides of the inequality and using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, for every $a, b \geq 0$. □

Lemma 4. *Let $\mathcal{L}_P\left(\mathcal{A}(\mathcal{D})\right)$ denote the prediction error given by Eq. (1). Then for any $q \geq 2$,*

$$\left\| \mathcal{L}_P\left(\mathcal{A}(\mathcal{D})\right) - \mathbb{E}\left[\mathcal{L}_P\left(\mathcal{A}(\mathcal{D})\right) \right] \right\|_q \leq \sqrt{2\kappa q} \sqrt{n}\, \mathcal{S}_q(\mathcal{A}, n).$$

Proof. The proof is similar to that of Lemma 3.

$$\begin{aligned}
\left\| \mathcal{L}_P\left(\mathcal{A}(\mathcal{D})\right) - \mathbb{E}\left[\mathcal{L}_P\left(\mathcal{A}(\mathcal{D})\right) \right] \right\|_q &\leq \sqrt{2\kappa q} \sqrt{\left\| \sum_{j=1}^{n} \left[\mathcal{L}_P\left(\mathcal{A}(\mathcal{D})\right) - \mathcal{L}_P\left(\mathcal{A}(\tau_j(\mathcal{D}))\right) \right]^2 \right\|_{q/2}} \\
&\leq \sqrt{2\kappa q} \sqrt{n \left\| \left[\mathcal{L}_P\left(\mathcal{A}(\mathcal{D})\right) - \mathcal{L}_P\left(\mathcal{A}(\tau_j(\mathcal{D}))\right) \right]^2 \right\|_{q/2}}
\end{aligned}$$

$$\leq \sqrt{2\kappa q}\sqrt{n\left\|\left[\,c\left(\mathcal{A}(\mathcal{D},X),Y\right)-c\left(\mathcal{A}(\tau_j(\mathcal{D}),X),Y\right)\right]^2\right\|_{q/2}}$$
$$= \sqrt{2\kappa q}\sqrt{n}\left\|\left[\,c\left(\mathcal{A}(\mathcal{D},X),Y\right)-c\left(\mathcal{A}(\tau_j(\mathcal{D}),X),Y\right)\right]^2\right\|_{q}$$
$$= \sqrt{2\kappa q}\sqrt{n}\,\mathcal{S}_q(\mathcal{A},n).$$

□

Proof of Theorem 2. The triangular inequality and Lemmas 3 and 4 lead to

$$\left\|\mathcal{L}_P\left(\mathcal{A}(\mathcal{D})\right)-\widehat{R}_1(A,\mathcal{D})-\left(\mathbb{E}\left[\,\mathcal{L}_P\left(\mathcal{A}(\mathcal{D})\right)\right]-\mathbb{E}\left[\,\widehat{R}_1(A,\mathcal{D})\,\right]\right)\right\|_q$$
$$\leq \left\|\mathcal{L}_P\left(\mathcal{A}(\mathcal{D})\right)-\mathbb{E}\left[\,\mathcal{L}_P\left(\mathcal{A}(\mathcal{D})\right)\right]\right\|_q+\left\|\mathbb{E}\left[\,\widehat{R}_1(A,\mathcal{D})\,\right]-\widehat{R}_1(A,\mathcal{D})\right\|_q$$
$$\leq \sqrt{\kappa q n}\left[\,\sqrt{2}\,\mathcal{S}_q(\mathcal{A},n)+4\,\mathcal{S}_q(\mathcal{A},n-1)\,\right]+\frac{2\sqrt{\kappa q}}{\sqrt{n}}\left\|c(\mathcal{A}(\tau_j(\mathcal{D}),X_j),Y_j)-c(\mathcal{A}(\tau_j(\mathcal{D}),X'_j),Y'_j)\right\|_q.$$

□

3.1 Application to Ridge regression

We now extend the result of Theorem 2 to the case of the Ridge regression algorithm.

Corollary 1 (Ridge: bounding the moments)**.** *With the notation of Theorem 2, for any sample size $n>2$, $\eta\in(0,1)$, and $\lambda>\left[\,\eta(n-2)\,\right]^{-1}$, let $\mathcal{A}_\lambda(\mathcal{D})$ denote the Ridge estimator from Eq. (3). Then, assuming (**XBd**), for any $q\geq 2$,*

(i)

$$\left\|\widehat{R}_1(\mathcal{A}_\lambda,\mathcal{D})-\mathcal{L}_P\left(\mathcal{A}_\lambda(\mathcal{D})\right)-\mathbb{E}\left[\,\widehat{R}_1(\mathcal{A}_\lambda,\mathcal{D})-\mathcal{L}_P\left(\mathcal{A}_\lambda(\mathcal{D})\right)\right]\right\|_q$$
$$\leq \frac{\sqrt{q}}{\sqrt{n}}\left(\Gamma_{\lambda,1}\left\|Y\right\|_q^2+\Gamma_{\lambda,2}\left\|Y\right\|_{2q}^2\right), \tag{9}$$

(ii)

$$\left\|\widehat{R}_1(\mathcal{A}_\lambda,\mathcal{D})-\mathcal{L}_P\left(\mathcal{A}_\lambda(\mathcal{D})\right)\right\|_q\leq\frac{\sqrt{q}}{\sqrt{n}}\left(\Gamma_{\lambda,1}\left\|Y\right\|_q^2+\Gamma_{\lambda,2}\left\|Y\right\|_{2q}^2\right)+\frac{\Gamma_{\lambda,3}}{n}\left\|Y\right\|_{2q}^2, \tag{10}$$

where $\Gamma_{\lambda,1}=8\sqrt{\kappa}B_X^2/\lambda$, $\Gamma_{\lambda,2}=2\sqrt{\kappa}B_X^2/\lambda\left[\,(8+\sqrt{2})\left(1+\frac{B_X^2+\lambda}{\lambda(1-\eta)}\right)\left(1+\frac{B_X^2}{\lambda}\right)+\frac{4B_X^2}{\lambda}\,\right]$, and $\Gamma_{\lambda,3}=\frac{2B_X^2}{\lambda}\left(1+\frac{B_X^2+\lambda}{\lambda(1-\eta)}\right)\left(1+\frac{B_X^2}{\lambda}\right)$.

The first Eq. (9) is a direct consequence of Theorem 2. We recall that our goal is to quantify the discrepancy between the LoO estimator and the prediction error $\widehat{R}_1(\mathcal{A}_\lambda,\mathcal{D})-\mathcal{L}_P\left(\mathcal{A}_\lambda(\mathcal{D})\right)$. This justifies introducing Eq. (10).

With the Ridge estimator, the convergence rate of $\widehat{R}_1(\mathcal{A}_\lambda,\mathcal{D})$ towards $\mathcal{L}_P\left(\mathcal{A}_\lambda(\mathcal{D})\right)$ is of order $1/\sqrt{n}$ for any $q\geq 2$. The upper bound highlights that the distribution of Y influences the convergence rate, which deteriorates as $\left\|Y\right\|_q$ and $\left\|Y\right\|_{2q}$ grow. In particular, the dependence of the rate with respect to q is strictly $\sqrt{q}$ whenever Y is almost surely bounded as in (**YBd**). Under the weaker assumption that Y is sub-Gaussian, an additional terms will emerge Note also that this bound allows to derive PAC polynomial bounds by use of Markov-type inequalities.

The proof of Corollary 1 relies on the following proposition.

Proposition 3. *Let $\mathcal{A}_\lambda(\mathcal{D})$ denote the Ridge estimator from Eq. (3) for every $\lambda>0$, and assume (**XBd**) holds true. Then, we have*

$$\left\|c(\mathcal{A}_\lambda(\tau_j(\mathcal{D}),X_j),Y_j)-c(\mathcal{A}_\lambda(\tau_j(\mathcal{D}),X'_j),Y'_j)\right\|_q\leq\frac{4B_X^2}{\lambda}\left\|Y\right\|_q^2+\frac{4B_X^4}{\lambda^2}\left\|Y\right\|_{2q}^2.$$

Proof of Proposition 3. From (**XBd**), it comes

$$\left|\left\langle \mathcal{A}_\lambda(\tau_j(\mathcal{D}), X_j - X_j'\right\rangle_2\right| \leq 2B_X \left|\mathcal{A}_\lambda(\tau_j(\mathcal{D})\right|_2,$$

and

$$\left|Y_j - \left\langle \mathcal{A}_\lambda(\tau_j(\mathcal{D}), X_j\right\rangle_2 + Y_j' - \left\langle \mathcal{A}_\lambda(\tau_j(\mathcal{D}), X_j'\right\rangle_2\right| \leq |Y_j| + \left|Y_j'\right| + 2B \left|\mathcal{A}_\lambda(\tau_j(\mathcal{D})\right|_2.$$

Since Eq. (3) and (**XBd**) imply $\left|\mathcal{A}_\lambda(\tau_j(\mathcal{D})\right|_2 \leq \frac{B_X}{(n-1)\lambda} \sum_{i\neq j} |Y_i|$, the independence between Y_j, Y_j' and $\{Y_i\}_{i\neq j}$ provides

$$\begin{aligned}
&\left\|c\left(\mathcal{A}_\lambda(\tau_j(\mathcal{D}), X_j), Y_j\right) - c\left(\mathcal{A}_\lambda(\tau_j(\mathcal{D}), X_j'), Y_j'\right)\right\|_q \\
&\leq 2B_X^2 \left\|\frac{1}{(n-1)\lambda}\sum_{i\neq j}|Y_i|\right\|_q \left\|\,|Y_j| + |Y_j'|\,\right\|_q + 4B_X^4 \left\|\left(\frac{1}{(n-1)\lambda}\sum_{i\neq j}|Y_i|\right)^2\right\|_q \\
&\leq \frac{4B_X^2}{\lambda}\|Y\|_q^2 + \frac{4B_X^4}{\lambda^2}\|Y\|_{2q}^2,
\end{aligned}$$

where Y denotes an independent copy of the Y_is and Y_j'.

□

Proof of Corollary 1. To prove claim (i), note that Theorem 2 and Proposition 3 lead to

$$\begin{aligned}
&\left\|\widehat{R}_1(\mathcal{A}_\lambda, \mathcal{D}) - \mathcal{L}_P\left(\mathcal{A}_\lambda(\mathcal{D})\right) - \mathbb{E}\left[\,\widehat{R}_1(\mathcal{A}_\lambda, \mathcal{D}) - \mathcal{L}_P\left(\mathcal{A}_\lambda(\mathcal{D})\right)\right]\right\|_q \\
&\leq \sqrt{2\kappa q}\sqrt{n}\,\mathcal{S}_q(\mathcal{A}_\lambda, n) + 4\sqrt{\kappa q}\sqrt{n}\,\mathcal{S}_q(\mathcal{A}_\lambda, n-1) \\
&\quad + \frac{2\sqrt{\kappa q}}{\sqrt{n}}\left\|c(\mathcal{A}_\lambda(\tau_j(\mathcal{D}), X_j), Y_j) - c(\mathcal{A}_\lambda(\tau_j(\mathcal{D}), X_j'), Y_j')\right\|_q \\
&\leq 2(8+\sqrt{2})\sqrt{\kappa q}\sqrt{n}\,\|Y\|_{2q}^2 \frac{B_X^2}{n\lambda}\left(1 + \frac{B_X^2+\lambda}{\lambda(1-\eta)}\right)\left(1+\frac{B_X^2}{\lambda}\right) \\
&\quad + \frac{2\sqrt{\kappa q}}{\sqrt{n}}\left(\frac{4B_X^2}{\lambda}\|Y\|_q^2 + \frac{4B_X^4}{\lambda^2}\|Y\|_{2q}^2\right) \\
&= \frac{2\sqrt{\kappa q}}{\sqrt{n}}\frac{B_X^2}{\lambda}\left[(8+\sqrt{2})\|Y\|_{2q}^2\left(1+\frac{B_X^2+\lambda}{\lambda(1-\eta)}\right)\left(1+\frac{B_X^2}{\lambda}\right) + 4\|Y\|_q^2 + \frac{4B_X^2}{\lambda}\|Y\|_{2q}^2\right] \\
&= \frac{2\sqrt{\kappa q}}{\sqrt{n}}\frac{B_X^2}{\lambda}\left(4\|Y\|_q^2 + \left[(8+\sqrt{2})\left(1+\frac{B_X^2+\lambda}{\lambda(1-\eta)}\right)\left(1+\frac{B_X^2}{\lambda}\right) + \frac{4B_X^2}{\lambda}\right]\|Y\|_{2q}^2\right).
\end{aligned}$$

To prove claim (ii), it only remains to notice that $\left|\mathbb{E}\left[\,\mathcal{L}_P\left(\mathcal{A}_\lambda(\mathcal{D})\right)\right] - \mathbb{E}\left[\,\widehat{R}_1(A, \mathcal{D})\right]\right| \leq \mathcal{S}_q(\mathcal{A}_\lambda, n)$ by Jensen's inequality. It results

$$\begin{aligned}
&\left\|\widehat{R}_1(\mathcal{A}_\lambda, \mathcal{D}) - \mathcal{L}_P\left(\mathcal{A}_\lambda(\mathcal{D})\right)\right\|_q \\
&\leq \frac{2\sqrt{\kappa q}}{\sqrt{n}}\frac{B_X^2}{\lambda}\left(4\|Y\|_q^2 + \left[(8+\sqrt{2})\left(1+\frac{B_X^2+\lambda}{\lambda(1-\eta)}\right)\left(1+\frac{B_X^2}{\lambda}\right) + \frac{4B_X^2}{\lambda}\right]\|Y\|_{2q}^2\right) \\
&\quad + \frac{2B_X^2}{n\lambda}\left(1+\frac{B_X^2+\lambda}{\lambda(1-\eta)}\right)\left(1+\frac{B_X^2}{\lambda}\right)\|Y\|_{2q}^2.
\end{aligned}$$

□

4 PAC exponential inequalities for the Leave-one-Out estimator

We now state our final results for the LoO estimator. The key ingredient is the following proposition, which establishes a link between moment inequalities and PAC exponential inequalities.

Proposition 4 (Celisse and Mary-Huard [2015], Proposition D.1)**.** *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 any $q \geq q_0$, $\mathbb{E}\left[|X|^q\right] \leq C\left(\sum_{i=1}^N \lambda_i q^{\alpha_i}\right)^q$. Then for every $x > 0$,*

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

The following two final results are our most refined PAC exponential inequalities for the LoO estimator, and follow from Corollary 1 (derived for the Ridge algorithm) combined with Proposition 4 where $q_0 = 2$, $C = 1$ and $\min_j \alpha_j = 1/2$.

Theorem 3. *With the setting of Corollary 1 and assuming (**YBd**), we have for every $x > 0$, with probability at least $1 - e \cdot e^{-x}$,*

$$\left|\widehat{R}_1(\mathcal{A}_\lambda, \mathcal{D}) - \mathcal{L}_P\left(\mathcal{A}_\lambda(\mathcal{D})\right)\right| \leq \sqrt{\frac{2ex}{n}} B_Y^2 \left(\Gamma_{\lambda,1} + \Gamma_{\lambda,2} + \Gamma_{\lambda,3}\right), \tag{11}$$

where

$$\Gamma_{\lambda,1} = 8\sqrt{\kappa} B_X^2/\lambda,$$
$$\Gamma_{\lambda,2} = 2\sqrt{\kappa} B_X^2/\lambda \left[(8+\sqrt{2})\left(1 + \frac{B_X^2 + \lambda}{\lambda(1-\eta)}\right)\left(1 + \frac{B_X^2}{\lambda}\right) + \frac{4B_X^2}{\lambda}\right],$$
$$\Gamma_{\lambda,3} = \frac{2B_X^2}{\lambda}\left(1 + \frac{B_X^2 + \lambda}{\lambda(1-\eta)}\right)\left(1 + \frac{B_X^2}{\lambda}\right).$$

This result establishes with high probability that the LoO estimator is $1/\sqrt{n}$ close to the prediction error when applied to the Ridge regression estimator.

This rate of convergence is preserved when (**YBd**) is relaxed to (**SubG**) as shown below.

Theorem 4. *With the setting of Corollary 1 and assuming (**SubG**), we have for every $x > 0$, with probability at least $1 - e \cdot e^{-x}$*

$$\left|\widehat{R}_1(\mathcal{A}_\lambda, \mathcal{D}) - \mathcal{L}_P\left(\mathcal{A}_\lambda(\mathcal{D})\right)\right| \leq \frac{M_1 (\mathbb{E}[Y])^2 \sqrt{x} + M_2\, v\, x^{3/2}}{\sqrt{n}}, \tag{12}$$

where $M_1 = 2\sqrt{2e}\left(\Gamma_{\lambda,1} + \Gamma_{\lambda,2} + \Gamma_{\lambda,3}\right)$, and $M_2 = 16e^2(2e)^{3/2}\left(\Gamma_{\lambda,1} + \Gamma_{\lambda,2} + \Gamma_{\lambda,3}\right)$.

Both Eqs. (11) and (12) lead to deviations for the LoO estimator of order $1/\sqrt{n}$, which is similar to that of Bousquet and Elisseeff [2002, Theorem 18, Eq. 18], but derived under weaker assumptions (in particular L^q stability instead of uniform stability). Note also that relaxing (**YBd**) to (**SubG**) results in an additional term of magnitude $x^{3/2}$.

Proof of Theorem 3. Using Assumption (**YBd**) and Corollary 1, apply Proposition 4 with $N = 1$, $\alpha_1 = 1/2$ and $\lambda_1 = B_Y^2\left(\Gamma_{\lambda,1} + \Gamma_{\lambda,2} + \Gamma_{\lambda,3}\right)/\sqrt{n}$ to obtain Eq. (11). □

Proof of Theorem 4. From Assumption (**SubG**), the triangular inequality yields

$$\|Y\|_q^2 \leq 2\left(\mathbb{E}[Y]\right)^2 + 2\|Y - \mathbb{E}[Y]\|_q^2 \leq 2\left(\mathbb{E}[Y]\right)^2 + 2(2e)^2 vq$$
$$\|Y\|_{2q}^2 \leq 2\left(\mathbb{E}[Y]\right)^2 + 2\|Y - \mathbb{E}[Y]\|_{2q}^2 \leq 2\left(\mathbb{E}[Y]\right)^2 + 4(2e)^2 vq.$$

To simplify the derivation, we use $\|Y\|_q^2 \leq 2\left(\mathbb{E}[Y]\right)^2 + 4(2e)^2 vq$ and $1/n \leq 1/\sqrt{n}$. From Corollary 1,

$$\left\|\widehat{R}_1(\mathcal{A}, \mathcal{D}) - \mathcal{L}_P\left(\mathcal{A}(\mathcal{D})\right)\right\|_q \leq \frac{\sqrt{q}}{\sqrt{n}}\left(\Gamma_{\lambda,1} + \Gamma_{\lambda,2} + \Gamma_{\lambda,3}\right)\left[2\left(\mathbb{E}[Y]\right)^2 + 4(2e)^2 vq\right]$$
$$\leq \frac{\sqrt{q}}{\sqrt{n}}\left(\Gamma_{\lambda,1} + \Gamma_{\lambda,2} + \Gamma_{\lambda,3}\right)\left[2\left(\mathbb{E}[Y]\right)^2 + 4(2e)^2 vq\right]$$

To achieve the proof, it only remains to apply Proposition 4 with $N = 2$, $(\alpha_1, \alpha_2) = (1/2, 3/2)$, and

$$\lambda_1 = 2\sqrt{2e}\left(\Gamma_{\lambda,1} + \Gamma_{\lambda,2} + \Gamma_{\lambda,3}\right)\left(\mathbb{E}[Y]\right)^2,$$
$$\lambda_2 = 16e^2(2e)^{3/2}\left(\Gamma_{\lambda,1} + \Gamma_{\lambda,2} + \Gamma_{\lambda,3}\right) v.$$

□

5 Perspectives

We introduced a new stability notion—L^q stability—which generalises existing hypothesis stability while remaining weaker than uniform stability. It provides PAC exponential generalisation bounds similar to those originally derived under the stronger uniform stability assumption. This has been achieved using a new generic strategy relying on moment inequalities.

In the present paper, this strategy has been successfully applied to the Ridge regression algorithm. A natural next step is to explore the collection of learning algorithms falling into the scope of our L^q stability notion.

From both practical and theoretical perspectives, it is crucial to provide a thorough analysis of CV. On the basis of our generic analysis of LoO, we intend to investigate other CV procedures such as V-fold CV, Leave-v-Out, and so on.

A Technical results

Lemma 5 (Devroye and Wagner [1979], Eq. (14)). *For any $1 \leq k \leq n$, let $\mathcal{A}_k$ defined by Eq. (2), and let $Z_1, \ldots, Z_n$ denote n independent and identically distributed random variables such that for any $1 \leq i \leq n$, $Z_i = (X_i, Y_i) \sim P$. Then for any $1 \leq j \leq n$,*

$$\mathbb{P}\left[\, \mathcal{A}_k(\mathcal{D}; X) \neq \mathcal{A}_k(\tau_j(\mathcal{D}); X) \,\right] \leq \frac{4}{\sqrt{2\pi}} \frac{\sqrt{k}}{n} \ .$$

Lemma 6 (Henderson and Searle [1981], Eq. (18)). *Let A and B denote two symmetric matrices in $\mathcal{M}_d(\mathbb{R})$ for some integer $d > 0$ such that A and $A + B$ are invertible. Then,*

$$A^{-1} - (A + B)^{-1} = A^{-1} B A^{-1} \left(I_d + B A^{-1}\right)^{-1} .$$

Lemma 7. *Let $M \in \mathcal{M}_d(\mathbb{R})$ be any symmetric positive semidefinite matrix. Then for every $\lambda > 0$,*

$$\left\| (M + \lambda I_d)^{-1} \right\|_{op} \leq \frac{1}{\lambda} \ .$$

Lemma 8 (Bhatia [2013], Theorem VIII.3.1). *Let $D \in \mathcal{M}_d(\mathbb{R})$ be a diagonal matrix and $M \in \mathcal{M}_d(\mathbb{R})$ denote any matrix. Then,*

$$\max_{1 \leq j \leq d} \min_{1 \leq i \leq d} |\sigma_i(D) - \sigma_j(M)| \leq \|D - M\|_{op} \,,$$

where $\sigma_1(N) \geq \ldots \geq \sigma_d(N)$ denote the d eigenvalues of matrix N in nonincreasing order, and $\|\cdot\|_{op}$ is the operator norm.

References

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

R. Bhatia. *Matrix analysis*, volume 169. Springer Science & Business Media, 2013.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. OUP Oxford, 2013.

O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2002.

A. Celisse and T. Mary-Huard. New upper bounds on cross-validation for the k-Nearest Neighbor classification rule. arXiv preprint, 2015.

L. Devroye. Exponential inequalities in nonparametric estimation. In *Nonparametric functional estimation and related topics*, pages 31–44. Springer, 1991.

L. Devroye and T. J. Wagner. Distribution-Free performance Bounds for Potential Function Rules. *IEEE Transaction in Information Theory*, 25(5):601–604, 1979.

A. Elisseeff, T. Evgeniou, and M. Pontil. Stability of randomized learning algorithms. In *Journal of Machine Learning Research*, pages 55–79, 2005.

T. Evgeniou, M. Pontil, and A. Elisseeff. Leave-one-out error, stability, and generalization of voting combinations of classifiers. *Machine Learning*, 55(1):71–97, 2004.

J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*. Springer, Berlin, 2009.

H. V. Henderson and S. R. Searle. On deriving the inverse of a sum of matrices. *Siam Review*, 23(1): 53–60, 1981.

S. Kale, R. Kumar, and S. Vassilvitskii. Cross-validation and mean-square stability. In *Proceedings of the Second Symposium on Innovations in Computer Science (ICS2011)*, 2011.

M. Kearns and D. Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Computation*, 11(6):1427–1453, 1999.

R. Kumar, D. Lokshtanov, S. Vassilvitskii, and A. Vattani. Near-optimal bounds for cross-validation via loss stability. In *Proceedings of The 30th International Conference on Machine Learning*, pages 27–35, 2013.

S. Kutin and P. Niyogi. Almost-everywhere algorithmic stability and generalization error. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pages 275–282, 2002.

C. McDiarmid. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.

S. Mukherjee, P. Niyogi, T. Poggio, and R. Rifkin. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 25(1-3):161–193, 2006.

A. Rakhlin, S. Mukherjee, and T. Poggio. Stability results in learning theory. *Analysis and Applications*, 3(04):397–417, 2005.

S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.

M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B*, 36:111–147, 1974.

S. Villa, L. Rosasco, and T. Poggio. On learnability, complexity and stability. In *Empirical Inference*, pages 59–69. Springer, 2013.

B. Yu. Stability. *Bernoulli*, 19(4):1484–1500, 2013.