An Exponential Efron-Stein Inequality for L_q Stable Learning Rules

Karim Abou-Moustafa * KARIM.ABOU-MOUSTAFA@SAS.COM
Dept. of Computing Science
University of Alberta
Edmonton, AB T6G 2E8, Canada

Csaba Szepesvári † CSABA.SZEPESVARI@GMAIL.COM
Dept. of Computing Science
University of Alberta
Edmonton, AB T6G 2E8, Canada

Editors: Aurélien Garivier and Satyen Kale

Abstract

There is an accumulating evidence in the literature that *stability of learning algorithms* is a key characteristic that permits a learning algorithm to generalize. Despite various insightful results in this direction, there seems to be an overlooked dichotomy in the type of stability-based generalization bounds we have in the literature. On one hand, the literature seems to suggest that exponential generalization bounds for the estimated risk, which are optimal, can be *only* obtained through *stringent*, *distribution independent* and *computationally intractable* notions of stability such as *uniform stability*. On the other hand, it seems that *weaker* notions of stability such as hypothesis stability, although it is *distribution dependent* and more *amenable* to computation, can *only* yield polynomial generalization bounds for the estimated risk, which are suboptimal.

In this paper, we address the gap between these two regimes of results. In particular, the main question we address here is *whether it is possible to derive exponential generalization bounds for the estimated risk using a notion of stability that is computationally tractable and distribution dependent, but weaker than uniform stability*. Using recent advances in concentration inequalities, and using a notion of stability that is weaker than uniform stability but distribution dependent and amenable to computation, we derive an exponential tail bound for the concentration of the estimated risk of a hypothesis returned by a *general* learning rule, where the estimated risk is expressed in terms of either the resubstitution estimate (empirical error), or the deleted (or, leave-one-out) estimate. As an illustration we derive exponential tail bounds for ridge regression with *unbounded responses* – a setting where uniform stability results of Bousquet and Elisseeff (2002) are not applicable.

Keywords: Generalization bounds, algorithmic stability, the leave one out estimate, resubstitution estimate, empirical error, concentration inequalities, moments bounds, tail bounds, Efron-Stein inequality.

1. Introduction

There is an accumulating evidence in the literature that stability of learning algorithms is a key characteristic that permits a learning algorithm to generalize. The earliest results in this regard are

* Now at SAS Inst. Inc, Cary, North Carolina, USA.
† On leave at DeepMind, London, UK.

due to Devroye and Wagner (1979a,b) where they derive distribution-free *exponential* generalization bounds for the concentration of the *leave-one-out* estimate, or the *deleted* estimate, for k local learning rules. Although the notion of stability was not explicitly mentioned in their work, the exponential bounds of Devroye and Wagner (1979a,b) can be seen as relying on the so called *hypothesis stability*; a concept due to Kearns and Ron (1999).

Various results for different estimates followed the works of Devroye and Wagner (1979b,a). Lugosi and Pawlak (1994) extended the work of Devroye and Wagner (1979b,a) to smooth estimates of the error developed in terms of *a posteriori* distribution for the deleted estimate. Holden (1996) derived *sanity-check bounds* for the deleted estimate and the k *folds cross–validation* (KFCV) estimate using hypothesis stability for few algorithms in the realizable setting.[1] Sanity-check bounds are *assurances* that the worst deleted estimate and the worst KFCV estimate will not be considerably worse than the *training error* or the *resubstitution* estimate (also known as the empirical error, or empirical risk) (Devroye and Wagner, 1979b). Kearns and Ron (1999), using the notion of *error stability*, give sanity-check bounds for the deleted estimate but for more general classes of learning rules (in the unrealizable or agnostic setting). In particular, they show that if a learning algorithm has a finite VC dimension search space, then the algorithm is *error-stable* and its error stability is controlled by the said VC dimension. Hence, using stability as a complexity measure will not yield worse bounds than using the VC dimension. Note that error stability is much weaker than hypothesis stability in the sense that hypothesis stability implies error stability, and this weakness was necessary to obtain more general sanity-check bounds than those obtained by Holden (1996). More recently, Kale et al. (2011) show that, using a weak notion of stability known as *mean-square* stability, the averaging taking place in the KFCV estimation procedure can reduce the variance of the generalization error; i.e. the averaging in the KFCV estimation procedure can improve the concentration of the estimated error around the expected error of the hypothesis returned by the learning rule.

For general learning rules and for *regularized empirical risk minimization* learning rules, Bousquet and Elisseeff (2002) using the notion of *uniform stability*, extended the work of Lugosi and Pawlak (1994) and derived *exponential* generalization bounds for the resubstitution estimate and the deleted estimate. Further generalization results based on uniform stability (or one of its variants) were later obtained in the works of Kutin and Niyogi (2002); Rakhlin et al. (2005); Mukherjee et al. (2006); Shalev-Shwartz et al. (2010), to name but a few. In particular, Shalev-Shwartz et al. (2010) showed that a version of uniform stability is key to learnability in the general learning setting with uniformly bounded losses. These results were reinforced and extended in various directions such as deriving new results for randomized learning algorithms (Elisseeff et al., 2005), transfer and meta learning (Maurer, 2005), adaptive data analysis (Bassily et al., 2016), stochastic gradient descent (Hardt et al., 2016), structured prediction (London et al., 2016), multi-task learning (Zhang, 2015), ranking algorithms (Agarwal and Niyogi, 2009), as well as in understanding the trade-off between sparsity and stability (Xu et al., 2012).

Despite these recent advances, and excluding sanity-check bounds, there seems to be an overlooked dichotomy in the type of stability-based generalization results. In particular, the results on stability and generalization can be grouped into two regimes:

1. In particular, Holden (1996) considered the closure algorithm, and the deterministic 1–inclusion graph prediction strategy.

1. *Polynomial* generalization bounds, which are *sub-optimal* and based on hypothesis stability for instance.

2. *Exponential* generalization bounds, which are *optimal* and based on uniform stability (and its variants).

Comparing *uniform* stability to other notions of stability in the literature, uniform stability is the strongest (most demanding) notion of stability in the sense that it implies all other notions of stability such as hypothesis stability, error stability, and mean-square stability (Bousquet and Elisseeff, 2002). A learning rule is *uniformly stable* if the change in the prediction loss is small, no matter how the input to the learning rule is selected, no matter what value is used as a test example, and no matter which example is removed (or replaced) in the input.

Despite the strength of uniform stability, it is unpleasantly restrictive. First, unlike other notions of stability (e.g. L_2 and L_1 stability), uniform stability is a stringent notion of stability that is insensitive to the data-generating distribution. This is problematic since it removes the possibility of studying large classes of learning rules, or even classes of problems. One particularly striking example is binary classification with the zero-one loss. For this problem, as it was already noted by Bousquet and Elisseeff (2002), *no non-trivial algorithm* can be uniformly β-stable with $\beta < 1$. Another example when uniform stability fails is regression with unbounded losses and response variables. Second, as noted earlier, uniform stability is distribution-free and is thus unsuitable for studying finer details of learning algorithms. Computation is another aspect that distinguishes uniform stability from other notions of stability. While hypothesis, error, and mean-square stability can be estimated using a finite sample, uniform stability is computationally intractable. In other words, although uniform stability yields exponential generalization bounds, these bounds cannot be empirically estimated using a finite sample in the spirit of empirical Bernstein bounds for instance (Audibert et al., 2007; Mnih et al., 2008).

In this research, we are particularly motivated by these previous observations. That is, on the one hand, the literature seems to suggest that exponential generalization bounds for the estimated risk, which are optimal, can be *only* obtained through *stringent*, *distribution independent*, and *computationally intractable* notions of stability such as uniform stability (and its variants). On the other hand, it seems that *weaker* notions of stability such as hypothesis and mean-square stability, although they are *distribution dependent* and potentially more *amenable* to computation, can *only* yield polynomial generalization bounds for the estimated risk, which are sub-optimal.

The chief purpose of this paper is to address the gap between these two regimes of results. In particular, the main question we address here is *whether it is possible to derive exponential generalization bounds for the estimated risk using a notion of stability that is computationally tractable, distribution dependent, but weaker than uniform stability*. Our work here gives a positive answer to this question; we show that using recent advances in exponential concentration inequalities, and using a notion of stability that is distribution dependent, amenable to computation, but weaker than uniform stability, we derive in Theorem 6 an exponential tail bound for the concentration of the estimated risk of a hypothesis returned by a *general* learning rule, where the estimated risk is developed in terms of either the deleted estimate, or the resubstitution estimate (also known as the empirical error).

Two main ingredients that allowed us to bridge the gap between these two regimes of results; (*i*) recent advances in exponential concentration inequalities, in particular the exponential Efron-Stein inequality due to Boucheron et al. (2003) and Boucheron et al. (2013); and (*ii*) the elegant notion of

L_q stability due to Celisse and Guedj (2016) which is distribution dependent, weaker than uniform stability, and generalizes hypothesis stability and mean-square stability to higher order moments.

Exponential Efron-Stein inequalities aim to bound the deviation of a general function f of n independent input random variables (RVs) from its expected value.[2] The seminal works of Boucheron et al. (2003) and Boucheron et al. (2013) bound this deviation by means of variance-like terms that measure the sensitivity of f with respect to the *replacement* of one RV from the n independent input RVs to f, with another independent copy of this RV. This notion of sensitivity with respect to the replacement of RVs is not suitable for our purposes, nor does it fit naturally the empirical estimation of these bounds based on finite datasets. As a byproduct of the results presented here, we derive an extension of the exponential Efron-Stein inequality when the sensitivity of f is measured with respect to the *removal* of one RV from the n independent input RVs to f (see Lemma 14). This notion of sensitivity with respect to the removal of RVs is naturally aligned with the notion of L_q stability, and with error estimates such as the deleted estimate and the KFCV estimate.

Efron-Stein inequalities have long been proposed to study the concentration of error estimates. First, the classic inequality was considered for bounding the variance (e.g., (Bousquet and Elisseeff, 2002)). Soon after Boucheron et al. (2003) introduced the variant for higher moments, Rakhlin et al. (2005) used this for deriving exponential tail bounds for the so-called almost uniformly stable learning algorithms, replicating the results of Kutin and Niyogi (2002), who used an extension of McDiarmid's inequality. More recent use of the higher order moment version is due to Celisse and Guedj (2016), who introduced the distribution-dependent L_q-stability coefficients and used them to derive bounds on the higher moments of the difference between the deleted estimate and the true risk. Celisse and Guedj (2016) used these moment bounds to get exponential tail bounds for the special case of ridge regression.

Our work is closest in spirit to Celisse and Guedj (2016). However, we provide an alternate route for obtaining exponential tail bounds by providing an exponential Efron-Stein inequality of the "removal type" in Lemma 14. This inequality is used to bound the moment-generating function (MGF) of various random variables, such as the deleted estimate, the resubstitution estimate, or the true risk of the random hypothesis returned by the learning rule. In each case, the bound is obtained in terms of the MGF of a random variable that corresponds to an average stability quantity of removing a sample. This latter MGF is bounded by controlling the growth-rate of various L_q stability coefficients, which leads the final exponential tail bounds. We obtain such a tail bound for the deleted estimate (Theorem 6), and also for the resubstitution estimate (Theorem 9). To control the tail of the resubstitution estimate, we observe that it is not sufficient to control the L_q stability coefficients introduced by Celisse and Guedj (2016), but one must also control a related, but distinct quantity, which measures the sensitivity of the algorithm to removing an example from the training set when the algorithm is tested on the example that is removed. We also apply our results to the case of ridge regression with unbounded response variables. In this case, we obtain the first exponential tail bounds for the deleted estimate (the case of resubstitution estimate is similar, but is not given explicitly). Since for unbounded response variables, the ridge regression estimator is not uniformly stable, these tail bounds were out of reach of previous techniques that built on uniform stability.

2. The n independent RVs are not necessarily identically distributed.

2. Setup and Notations

We consider learning in Vapnik's framework for risk minimization with bounded losses (Vapnik, 1995): A learning problem is specified by the triplet $(\mathcal{H}, \mathcal{X}, \ell)$, where $\mathcal{H}, \mathcal{X}$ are sets and $\ell : \mathcal{H} \times \mathcal{X} \to [0, \infty)$. The set $\mathcal{H}$ is called the *hypothesis space*, $\mathcal{X}$ is called the *instance space*, and ℓ is called the *loss function*. The loss $\ell(h, x)$ indicates how well a hypothesis h explains (or fits) an instance $x \in \mathcal{X}$. The learning problem is defined as follows. A learner $\mathtt{A}$ sees a sample in the form of a sequence $\mathcal{S}_n = (X_1, \ldots, X_n) \in \mathcal{X}^n$ where $(X_i)_i$ is sampled in an independent and identically distributed (*i.i.d*) fashion from some unknown distribution $\mathscr{P}$ and returns a hypothesis $\widehat{h}_n = \mathtt{A}(\mathcal{S}_n) \in \mathcal{H}$ based solely on $X_1, \ldots, X_n$.[3] The goal of the learner is to pick hypotheses with a small *risk* (defined shortly).

We assume that a learner is able to process samples (or sequences) of different cardinality. Hence, a learner will be identified with a map $\mathtt{A} : \cup_n \mathcal{X}^n \to \mathcal{H}$. We only consider deterministic learning rules in this work; the extension to randomizing learning rules is left for future work.

Given a distribution $\mathscr{P}$ on $\mathcal{X}$, the risk of a *fixed hypothesis* $h \in \mathcal{H}$ is defined to be $R(h, \mathscr{P}) = \mathbb{E}\left[\ell\left(h, X\right)\right]$, where $X \sim \mathscr{P}$. Since $\mathcal{S}_n$ is a random quantity, so are $\mathtt{A}(\mathcal{S}_n)$ and $R(\mathtt{A}(\mathcal{S}_n), \mathscr{P})$, the latter of which can be also written as $\mathbb{E}[\ell\left(\mathtt{A}(\mathcal{S}_n), X\right) | \mathcal{S}_n]$, where $X \sim \mathscr{P}$ is independent of $\mathcal{S}_n$. Ideal learners keep the risk $R(\mathtt{A}(\mathcal{S}_n), \mathscr{P})$ of the hypothesis returned by $\mathtt{A}$ "small" for a wide range of distributions $\mathscr{P}$.

q-Norm of Random Variables: In the sequel, we will heavily rely on the q-norm for random variables (RVs). For a real RV X, and for $1 \leq q \leq +\infty$, the q-norm of X is defined as: $\|X\|_q \doteq \left(\mathbb{E}\left[|X|^q\right]\right)^{1/q}$, and $\|X\|_\infty$ is the essential supremum of $|X|$. Note that for $1 \leq q \leq p \leq +\infty$, these norms satisfy $\|\cdot\|_q \leq \|\cdot\|_p$.

2.1. Risk Estimators

The generalization bounds on the risk usually center on some point-estimate of the random risk $R(\mathtt{A}(\mathcal{S}_n), \mathscr{P})$. Many estimators are based on calculating the sample mean of losses in one form or another. For any fixed hypothesis $h \in \mathcal{H}$ and dataset $\mathcal{S}_n$, the sample mean of losses of h against $\mathcal{S}_n$, also known as the *empirical risk* of h on $\mathcal{S}_n$, is given by

$$\widehat{R}(h, \mathcal{S}_n) = \frac{1}{n} \sum_{i=1}^{n} \ell\left(h, X_i\right). \tag{1}$$

Plugging $\mathtt{A}(\mathcal{S}_n)$ into $\widehat{R}(\cdot, \mathcal{S}_n)$ we get the *resubstitution (RES) estimate*, or the training error (Devroye and Wagner, 1979b): $\widehat{R}_{\text{RES}}\left(\mathtt{A}, \mathcal{S}_n\right) = \widehat{R}\left(\mathtt{A}\left(\mathcal{S}_n\right), \mathcal{S}_n\right)$. The resubstitution estimate is often overly "optimistic", i.e., it underestimates the actual risk $R(\mathtt{A}(\mathcal{S}_n), \mathscr{P})$. The *deleted (DEL) estimate* defined as

$$\widehat{R}_{\text{DEL}}\left(\mathtt{A}, \mathcal{S}_n\right) = \frac{1}{n} \sum_{i=1}^{n} \ell\left(\mathtt{A}(\mathcal{S}_n^{-i}), X_i\right), \tag{2}$$

is a common alternative to the resubstitution estimate that aims to correct for this optimism. Here, $\mathcal{S}_n^{-i} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$, i.e., it is the sequence $\mathcal{S}_n$ with example X_i removed. Since

3. The set $\mathcal{X}$ is thus measurable. All functions and sets are assumed and/or can be shown to be measurable as needed, saving us from the trouble of mentioning measurability in the rest of the paper.

$\mathbb{E}[\ell\left(\mathtt{A}(\mathcal{S}_n^{-i}), X_i\right)] = R_{n-1}(\mathtt{A}, \mathscr{P})$, then $\mathbb{E}[\widehat{R}_{\text{DEL}}(\mathtt{A}, \mathcal{S}_n)] = R_{n-1}(\mathtt{A}, \mathscr{P})$. When the latter is close to $R_n(\mathtt{A}, \mathscr{P})$, i.e., $\mathtt{A}$ is "stable", the deleted estimate may be a good alternative to the resubstitution estimate (Devroye et al., 1996). However, due to the potentially strong correlations between elements of $(\ell(\mathtt{A}(\mathcal{S}_n^{-i}), X_i))_i$, the variance of the deleted estimate *may be* significantly higher than that of the resubstitution estimate due to the overly redundant information content between $\ell(\mathtt{A}(\mathcal{S}_n^{-i}), X_i)$ and $\ell(\mathtt{A}(\mathcal{S}_n^{-j}), X_j)$ for $i \neq j$. The main goal of this work is to develop a high probability upper bound on the absolute deviation $|\widehat{R}_{\text{DEL}}\left(\mathtt{A}, \mathcal{S}_n\right) - R\left(\mathtt{A}(\mathcal{S}_n), \mathscr{P}\right)|$ in terms of the "stability" of $\mathtt{A}$, which is defined next.

3. Notions of Stability for Learning Rules

In the following, we go briefly over some well-known notions of algorithmic stability, introduce the notion of L_q stability coefficients and finally discuss its properties.

The first known notion of *algorithmic stability* is the so-called *hypothesis stability*, or L_1-*stability*, which is due to Devroye and Wagner (1979b).[4]

Definition 1 (Hypothesis Stability) *Algorithm $\mathtt{A}$ has hypothesis (or L_1) stability*[5] *β_h w.r.t to the loss function ℓ if the following holds*

$$\forall i \in \{1, \ldots, n\}\,, \ \mathbb{E}\left[|\ell\left(\mathtt{A}(\mathcal{S}_n), X\right) - \ell\left(\mathtt{A}(\mathcal{S}_n^{-i}), X\right)|\right] \ \leq \ \beta_h\,,$$

where randomness is over $\mathcal{S}_n$ and $X \sim \mathscr{P}$, and X is independent of $\mathcal{S}_n$.

Kearns and Ron (1999) proposed a weaker notion of stability known as *error stability* which measures the absolute change in the expected loss of a learning algorithm instead of the average absolute pointwise change in the loss:

Definition 2 (Error Stability) *Algorithm $\mathtt{A}$ has error stability β_e w.r.t the loss function ℓ if the following holds*

$$\forall i \in \{1, \ldots, n\}\,, \ |\mathbb{E}\left[\ell\left(\mathtt{A}(\mathcal{S}_n), X\right)\right] - \mathbb{E}\left[\ell\left(\mathtt{A}(\mathcal{S}_n^{-i}), X\right)\right]| \ \leq \ \beta_e\,,$$

where randomness is over $\mathcal{S}_n$ and $X \sim \mathscr{P}$, and X is independent of $\mathcal{S}_n$.

As noted by Kutin and Niyogi (2002), error stability is weaker than hypothesis stability (in the sense that if $\mathtt{A}$ has β hypothesis stability then it also has β error stability). Furthermore, this notion is not sufficiently strong to "guarantee generalization" in the sense that there are algorithms $\mathtt{A}$ such that their generalization gap, $\mathbb{E}[\widehat{R}_{\text{RES}}(\mathtt{A}, \mathcal{S}_n) - R(\mathtt{A}(\mathcal{S}_n), \mathscr{P})]$ stays positive, while the algorithm's error stability coefficient converges to zero as $n \to \infty$.

Kale et al. (2011) proposed another weak notion of stability known as *mean-square (MS) stability*, or L_2-*stability*.

4. The definitions are sometimes stated with their "high probability" variants. We prefer the expectation-versions as they fit our purposes better.
5. We believe that in all these definitions the word "sensitivity" should be used rather "stability". To be in line with the literature, we kept the terminology, though with much doubt about whether this is the correct decision.

Definition 3 (Mean-Square Stability) *Algorithm* $\mathtt{A}$ *has mean-square (or* L_2*) stability* β_{ms} *w.r.t the loss function* ℓ *if the following holds*

$$\forall i \in \{1, \ldots, n\}\,,\ \mathbb{E}\left[\left(\ell\left(\mathtt{A}(\mathcal{S}_n), X\right) - \ell\left(\mathtt{A}(\mathcal{S}_n^{-i}), X\right)\right)^2\right] \ \leq\ \beta_{ms}\,,$$

where randomness is over $\mathcal{S}_n$ *and* $X \sim \mathscr{P}$*, and* X *is independent of* $\mathcal{S}_n$*.*

We comment on the relationship between mean-square stability and the other notions of stability shortly.[6] Bousquet and Elisseeff (2002) proposed the strongest and most strict notion of stability, known as *uniform stability*, which implies all previous notions of stability. This notion of stability, together with McDiarmid inequality, permitted the derivation of the first exponential generalization error bound for the deleted estimate and the resubstitution estimate.

Definition 4 (Uniform Stability) *Algorithm* $\mathtt{A}$ *has uniform stability* β_u *w.r.t the loss function* ℓ *if the following holds*

$$\forall\, \mathcal{S}_n \in \mathcal{X}^n,\ \forall i \in \{1, \ldots, n\}\,,\ \forall x \in \mathcal{X},\ \left|\ell\left(\mathtt{A}(\mathcal{S}_n), x\right) - \ell\left(\mathtt{A}(\mathcal{S}_n^{-i}), x\right)\right| \ \leq\ \beta_u\,.$$

Finally, we arrive at our notion of L_q stability coefficients:

Definition 5 (L_q Stability Coefficient) *Let* $\mathcal{S}_n$ *be a sequence of* n *i.i.d random variables (RVs) drawn from* $\mathcal{X}$ *according to* $\mathscr{P}$*. Let* $\mathtt{A}$ *be a deterministic learning rule, and* ℓ *be a loss function as defined in Section 2. For* $q \geq 1$*, the* L_q stability *coefficient of* $\mathtt{A}$ *w.r.t* ℓ*,* $\mathscr{P}$*, and* n *is denoted by* $\beta_q(\mathtt{A}, \ell, \mathscr{P}, n)$ *and is defined as*

$$\beta_q^2(\mathtt{A}, \ell, \mathscr{P}, n) \ =\ \frac{1}{n}\sum_{i=1}^{n}\left\|\ell\left(\mathtt{A}(\mathcal{S}_n), X\right) - \ell\left(\mathtt{A}(\mathcal{S}_n^{-i}), X\right)\right\|_q^2\,,$$

where $X \sim \mathscr{P}$ *is independent of* $\mathcal{S}_n$*.*

Recall that a learning algorithm is symmetric, if $\mathtt{A}(\mathcal{S}_n) = \mathtt{A}(\mathcal{S}_n')$ for any two $\mathcal{S}_n$, $\mathcal{S}_n'$ which are reorderings of each other. For symmetric learning algorithms, the above definition simplifies to

$$\beta_q(\mathtt{A}, \ell, \mathscr{P}, n) = \left\|\ell\left(\mathtt{A}(\mathcal{S}_n), X\right) - \ell\left(\mathtt{A}(\mathcal{S}_n^{-1}), X\right)\right\|_q = \max_i \left\|\ell\left(\mathtt{A}(\mathcal{S}_n), X\right) - \ell\left(\mathtt{A}(\mathcal{S}_n^{-i}), X\right)\right\|_q\,,$$

the latter expression coinciding with the definition given by Celisse and Guedj (2016). Thus, for a symmetric algorithm, the stability coefficient $\beta_q(n) \doteq \beta_q(\mathtt{A}, \ell, \mathscr{P}, n)$ is in fact a q–norm for the RV $\Delta_n(\mathtt{A}) := \ell\left(\mathtt{A}(\mathcal{S}_n), X\right) - \ell\left(\mathtt{A}(\mathcal{S}_n^{-1}), X\right)$. The reason we chose the particular averaging in our definition is because this definition gives the best fit to the derivations we use.

We note in passing that the other stability notions could also consider averaging over index i, instead of taking the worst-case sensitivity as shown above. For symmetric rules, the difference, again, does not matter. However, for nonsymmetric algorithms this difference is nontrivial: While uniform stability is trivial for binary classification with the current definition, it is a useful notion with averaging as shown by the results of Shalev-Shwartz et al. (2010).

Since often $\mathtt{A}$, ℓ, $\mathscr{P}$, n are fixed, we will drop them (or any of them) from the notation and will just use, for example, $\beta_q, \beta_q(n)$, etc.[7] Note that for $1 \leq q \leq q'$, it holds that $\beta_q \leq \beta_{q'}$, which

6. Note that in the above definition, β_{ms} is not squared to keep our definitions in sync with the literature.
7. This should not be mistaken to taking a supremum over any subset of the dropped quantities: The stability coefficients are meant to be algorithm, loss and distribution dependent.

follows from the definition of q-norms. Now, the relationship between the various stability concepts becomes clear. Taking $\beta_e, \beta_h, \beta_{ms}, \beta_u$ as the smallest values that are possible (i.e., changing the inequalities in their definitions to equalities), assuming symmetric learning rules, we have $\beta_e \leq \beta_1 = \beta_h \leq \beta_2 = \beta_{ms}^{1/2} \leq \beta_u$, where the last inequality follows because the L^∞-norm is the largest of all of the q-norms.

The L_q stability coefficient quantifies the variation of the loss of $\mathtt{A}$ induced by removing one example from the training set. This is known as a *removal type* notion of stability and is in accordance with the previous notions of stability introduced earlier. The difference between L_q stability and earlier notions of stability is that L_q stability is in terms of the higher order moments of the RVs $|\ell\left(\mathtt{A}(\mathcal{S}_n), X\right) - \ell\left(\mathtt{A}(\mathcal{S}_n^{-i}), X\right)|$. The reason we care about higher moments is because we are interested in controlling the tail behavior of the deleted estimate. As will be shown, the tail behavior of the deleted estimate is also dependent on the tail behavior of RVs characterizing stability. As is well-known, knowledge of the higher moments of a RV is equivalent to knowledge of the tail behavior of the RV. As such, controlling the higher order moments provides more information on the distribution of this RV than simply considering first order (L_1) and second order (L_2) moments. As it will turn out, the L_q stability coefficients alone are insufficient to control either the bias, or the the tail behavior of the resubstitution estimate. To control these, we will introduce further stability coefficients, but we prefer to do this just before we need them.

4. Main Results

We give here the main results of our work, namely an exponential tail bound for the concentration of the estimated risk, expressed in terms of the deleted, or the resubstitution estimate. We start with the deleted estimate.

Before stating the result for the deleted estimate, we first state our two assumptions, both of which concern the behavior of the stability coefficients. While the first assumption is concerned with their dependence on n, the second assumption is concerned with their behavior as a function of q.

Assumption 1 *For a fixed $q > 0$, $\beta_q(n)$ is a nonincreasing function of n.*

Now note that our results remain valid if $\beta_q(n)$ is replaced with an upper bound on it (such as $\bar{\beta}_q(n)$), provided that the upper bound satisfies our assumptions. Defining $\bar{\beta}_q(n) = \max_{m \geq n} \beta_q(n)$, we find that the map $n \mapsto \bar{\beta}_q(n)$ is nonincreasing. This provides us with a general approach to meet Assumption 1, although we would often expect this assumption to be met anyways.

Assumption 2 $\exists\, u_1, w_1 \geq 0$ *s.t. for any integer $q \geq 1$, it holds that*

$$2n\beta_{4q}^2(n-1) + \tfrac{2}{n^2} \sum_{i=1}^{n} \left\| \ell\left(\mathtt{A}(\mathcal{S}_n^{-i}), X_i\right) \right\|_{4q}^2 \leq \sqrt{qu_1} \vee qw_1 \,, \tag{3}$$

where $a \vee b = \max(a, b)$.

This assumption will be clarified once we introduce our main tool (the exponential Efron-Stein inequality) and the notion of sub-gamma random variables in the following sections. The reader wondering about whether this assumption can be met will be pleased to find a positive answer

presented in Section 7 for the case of unbounded response ridge regression, where the assumption translates into conditions for the tail behavior of the response variable. Note that here u_1, v_1 will be distribution and sample size dependent constants, generally decreasing with the sample size.

With this, our main result for the deleted estimate is as follows:

Theorem 6 (Deleted estimate tail bound) *Let $\mathcal{X}$, $\mathcal{H}$ and ℓ be as previously defined. Let $\mathcal{S}_n$ be the dataset defined in Section 2.1, where $n \geq 2$. Let $\widehat{R}_{\text{DEL}}(\mathtt{A}, \mathcal{S}_n)$ be the deleted estimate defined in Eq. (2), and $R(\mathtt{A}(\mathcal{S}_n), \mathscr{P})$ be the risk for hypothesis $\mathtt{A}(\mathcal{S}_n)$. Then, under Assumptions 1 and 2, for $\delta \in (0,1)$ and $a > 0$, with probability $1 - 2\delta$ the following holds*

$$|\widehat{R}_{\text{DEL}}(\mathtt{A}, \mathcal{S}_n) - R(\mathtt{A}(\mathcal{S}_n), \mathscr{P})| \leq \beta_1(n) + 4\sqrt{(n\beta_2^2(n-1) + C_1)\log\left(\tfrac{2}{\delta}\right)} + C_2 \log\left(\tfrac{2}{\delta}\right) , \quad (4)$$

where $C_1 = 2.2a^2u_1 + 1.07a^2w_1^2$, and $C_2 = \frac{4}{3}(1.46aw_1 + \frac{1}{a})$.

While the above result bounds both sides of the tail, a one side version with $\log\left(\frac{2}{\delta}\right)$ replaced by $\log\left(\frac{1}{\delta}\right)$, also holds for both the upper and lower tails. We will soon explain the various terms in this bound, but first let us explain how the result is proven.

Once we establish our exponential Efron-Stein inequality (Lemma 14), the proof of Theorem 6 is relatively straightforward. The essence of the proof can be summarized as follows: In order to control the concentration of the random quantity $\widehat{R}_{\text{DEL}}(\mathtt{A}, \mathcal{S}_n)$ around the true risk, we study the concentration of $\widehat{R}_{\text{DEL}}(\mathtt{A}, \mathcal{S}_n)$ around its mean, the concentration of the true risk of $\mathtt{A}(\mathcal{S}_n)$ around its own mean, and the difference between the mentioned means. The latter is bounded by the $\beta_1(n)$ stability coefficient, using elementary arguments. To control the tails (or the higher order moments) of $\widehat{R}_{\text{DEL}}(\mathtt{A}, \mathcal{S}_n)$, we use our exponential Efron-Stein inequality, which tells us that we need to control the tails of another intermediary random variable, V_{DEL} (see Corollary 11 and its use in Section 6.1), a variance-type measure of the sensitivity of $\widehat{R}_{\text{DEL}}$ to the removal of one of the training examples at a time. The moments of V_{DEL} are shown be controlled by the expression on the LHS of Eq. (3) of Assumption 2. The assumption then helps to turn these bounds into a bound on the MGF of V_{DEL}, leading to tail bounds. The concentration of the true risk of $\mathtt{A}(\mathcal{S}_n)$ around its mean is controlled similarly. The full proof is presented in Section 6.

To interpret the bound, it is worthwhile to simplify it at the price of losing a bit on its tightness. In particular, further upper bounding the RHS using $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ and then choosing a optimally, yields the following simplified bound

$$|\widehat{R}_{\text{DEL}}(\mathtt{A}, \mathcal{S}_n) - R(\mathtt{A}(\mathcal{S}_n), \mathscr{P})| \leq \beta_1(n) + 4\sqrt{n\beta_2^2(n-1)\log\left(\tfrac{2}{\delta}\right)} + \\ + 8\sqrt{\tfrac{1}{3}\left(\sqrt{(2.2u_1 + 1.07w_1^2)} + \tfrac{1}{3}1.46w_1\right)\log\left(\tfrac{2}{\delta}\right)} , \quad (5)$$

where for the final form, we assumed that $\delta \leq 1/e$. What can be noticed is that the tail bound has the form that we expect to see for sub-gamma RVs; note the presence of the $\sqrt{\log(2/\delta)}$ and $\log(2/\delta)$ terms. Note also that, by assumption, $u_1^{1/2}$ and w_1 are both at least of size $\Omega(1/n)$, regardless the stability of $\mathtt{A}$. As a result, the coefficient of the $\log(2/\delta)$ term is at least of order $\Omega(\sqrt{1/n})$ even for algorithms where $\beta_q = 0$. This is expected because the deleted estimate is a "noisy" estimate of the true risk no matter the algorithm – one expects an $\Omega(\sqrt{1/n})$ lower bound to hold in general. Finally, we note in passing that with a bit more care, in the case of $w_1 = 0$ it is possible to slightly reduce the exponent of $\log(2/\delta)$ to $\log^{3/4}(2/\delta)$.

We can gain further insight by qualitatively comparing our bound in Theorem 6 with the exponential bound for the deleted estimate obtained by Bousquet and Elisseeff (2002, Theorem 12). To make the comparison easier, we first state their result using our notation. We also give the two sided version.

Theorem 7 (Deleted estimate tail bound through uniform stability) *Let* $\mathtt{A}$ *be a learning rule with uniform stability* β_u *(see Section 3) with respect to the loss function* ℓ *and assume that this loss function is in addition bounded:* $0 \leq \ell\left(\mathtt{A}(\mathcal{S}_n), X\right) \leq M$ *holds almost surely. Then, for any* $n \geq 1$*, and any* $\delta \in (0,1)$*, with probability* $1-\delta$*, the following holds*

$$\left| R(\mathtt{A}(\mathcal{S}_n), \mathscr{P}) - \widehat{R}_{\scriptscriptstyle DEL}\left(\mathtt{A}, \mathcal{S}_n\right) \right| \leq \beta_u(n) + 4n\beta_u(n)\sqrt{\tfrac{\log(2/\delta)}{2n}} + M\sqrt{\tfrac{\log(2/\delta)}{2n}}\,. \tag{6}$$

The bound in Theorem 7 has three terms; the first two terms are dependent on the uniform stability of the learning rule, and a third term that only depends on the loss function ℓ and the sample size n. When β_u scales as $1/n$ the bound becomes tight in a worst-case sense. As with our bound, even when $\beta_u = 0$, the third term stays positive (as it should).

Let us now compare the RHS of Eq. (6) to our simplified bound presented in Eq. (5). Both bounds have three terms, corresponding to different power of $\log\left(\frac{1}{\delta}\right)$. The first two terms in Eq. (6) have the same form as the first two terms in Eq. (5) except that in (6) β_u is used, while in (5) $\beta_1 (\leq \beta_2)$ is used. As discussed earlier, $\beta_1 \leq \beta_2 \leq \beta_u$, and in particular the gap between these quantities can be large; even β_u may be unbounded while the others bounded (as the example in Section 7 will show). At the same time, the constant coefficient of the second term (5) is larger than the corresponding coefficient in the second term of (6). Other than these differences, the terms are analogous.

As discussed earlier, our last term scales with $\log\left(\frac{1}{\delta}\right)$ rather than with with its square root. This is the price we pay partly because we allow *unbounded losses*. Further, the coefficients of this term, as discussed earlier, depend on the stability of the algorithm but the magnitude of the multiplier of $\log\left(\frac{1}{\delta}\right)$ will be at least of order $\Omega(\sqrt{1/n})$.

Note that Bousquet and Elisseeff (2002) state an identical result (to that shown above in (6)) for the generalization gap, $\widehat{R}_{\text{RES}} - R(\mathtt{A}(\mathcal{S}_n), \mathscr{P})$; i.e. the gap for the resubstitution estimate (or the training error). As we shall see soon, our bound also extends to this case with some modifications.

While preparing the final version of this manuscript, we noted the recent work of Feldman and Vondrak (2018) who improve this result of Bousquet and Elisseeff (2002) by replacing the second term in (6) by $\sqrt{\beta_u(n)\log\left(\frac{2}{\delta}\right)}$. This is an improvement whenever $\beta_u(n) \geq 1/n$ (i.e., for "not too stable" algorithms). One may hope that a similar improvement may be possible with non-uniform (distribution-dependent) notions of stability, but this is left for future work for now.

Notice that the above results are for the gap between the true risk and the deleted estimate, whereas oftentimes one wishes to control the gap between the true risk and the resubstitution (or *empirical*) estimate (or the training error); i.e., the well-known *generalization gap*. Indeed, one can follow the same path for our proof technique and derive an exponential tail bound for the concentration of the empirical estimate.

As it turns out, this result requires the introduction of a new type of stability coefficients: The reason is that there are stable algorithms that can overfit the training data in the sense that their training error is small. An example of such an algorithm for the binary classification setting is the "short-range nearest neighbor" rule which recalls the label of the closest training example to the

input when their distance is $o(1/n)$ and outputs a fixed label (say, 1) otherwise. As n increases, this algorithm will converge to output the a priori chosen label always. As such, the algorithm will also be very stable. Yet its training error is always zero, which can be far from its true risk. The situation is summarized in the following result (for details, see Appendix H):

Proposition 8 *There exist a distribution $\mathscr{P}$ and a learning algorithm* A *such that, everywhere,*

$$\lim_{n\to\infty} R(\mathtt{A}(\mathcal{S}_n), \mathscr{P}) - \widehat{R}_{\mathit{RES}}(\mathtt{A}, \mathcal{S}_n) > 0 \,, \tag{7}$$

while $\sup_{q\geq 1} \beta_q(\mathtt{A}, \ell, \mathscr{P}, n)/q \to 0$ *as* $n \to \infty$.

Note that by Theorem 6, $\widehat{R}_{\text{DEL}}(\mathtt{A}, \mathcal{S}_n) - R(\mathtt{A}(\mathcal{S}_n), \mathscr{P}) \xrightarrow{P} 0$ as long as $\sup_{q\geq 1} \frac{\beta_q(n)}{q} \to 0$ as $n \to \infty$. It follows that the deleted estimate is consistently estimating the risk of the short-range nearest neighbor rule, while the resubstitution estimate fails to be consistent. While it is common wisdom that the resubstitution estimate is often overly "optimistic", the example is a very clear demonstration of this weakness and shows that one has to be quite careful when using the training error, e.g., for model selection; as noted already in Devroye and Wagner (1979c).

One may then think that we should never use the training error, but this is easier said than done for most algorithms will in some form minimize the training error. Thus, the question remains, if the L_q stability in the previous sense is insufficient to guarantee the concentration of the training error around the true loss, what other property should an algorithm posses to control this concentration? We know that uniform stability provides a positive answer, but is there an analogue to the L_q stability coefficients that is sufficient for this purpose? The answer to this question can be obtained by repeating the derivations done in the proof of Theorem 6 and discovering the modifications necessary to control all the terms. This results in the definition of what we call the L_q resubstitution stability coefficients, which, given an algorithm A, are defined as follows:

$$\gamma_q^2(n) = \frac{1}{n}\sum_{i=1}^{n} \left\| \ell\left(\mathtt{A}(\mathcal{S}_n), X_i\right) - \ell\left(\mathtt{A}(\mathcal{S}_n^{-i}), X_i\right) \right\|_q^2 .$$

This is a direct analogue of the L_q stability coefficients: The main difference is that here, the algorithm is evaluated on training examples, with and without the example being removed, while for the L_q stability coefficients, the algorithm was evaluated on an example independent of the training data. This should make sense for already if we want to control the bias $\mathbb{E}[\widehat{R}_{\text{RES}}(\mathtt{A}, \mathcal{S}_n) - R(\mathtt{A}(\mathcal{S}_n), \mathscr{P})]$ we see the need to control the deviation between the loss measured at training examples and the loss measured outside – which is exactly what is captured by the γ_q coefficients.

We also replace Assumption 2 with the following assumption:

Assumption 3 $\exists\, u_1, w_1 \geq 0$ *s.t. for any integer* $q \geq 1$*, it holds that*

$$6n\left(\gamma_{4q}^2(n) + \gamma_{4q}^2(n-1) + \beta_{4q}^2(n-1)\right) + \frac{2\gamma_{4q}^2(n)}{n} + \frac{2}{n^2}\sum_{i=1}^{n} \left\| \ell\left(\mathtt{A}(\mathcal{S}_n^{-i}), X_i\right) \right\|_{4q}^2 \leq \sqrt{qu_1} \vee qw_1 \,.$$

Note that this assumption implies Assumption 2. Thus, any algorithm that satisfies this assumption, will also necessarily satisfy Assumption 2. With this, we are ready to state our result for tail of the resubstitution estimator:

Theorem 9 (Resubstitution estimate tail bound) *Using the setup of Theorem 6, but using Assumption 3 in place of Assumption 2, for $\delta \in (0,1)$ and $a > 0$, with probability $1-2\delta$ the following holds:*

$$|\widehat{R}_{RES}(\mathtt{A}, \mathcal{S}_n) - R(\mathtt{A}(\mathcal{S}_n), \mathscr{P})| \leq \beta_1(n) + \gamma_1(n) + 4\sqrt{(n\beta_2^2(n-1) + C_1)\log\left(\tfrac{2}{\delta}\right)} + C_2 \log\left(\tfrac{2}{\delta}\right) ,$$

where $C_1 = C_1(a)$ and $C_2 = C_2(a)$ are as in Theorem 6.

By and large, the proof follows the same line as the proof Theorem 6 with some necessary modifications. A proof of this result is provided in Appendix I. As can be noticed, the only difference to the bound available for the deleted estimate is the presence of the $(\gamma_q(n))$ terms, both in the assumption, and the result. As our previous example shows, these cannot be removed (in the case of the short-range nearest neighbor rule, these coefficients will be large). This suggests that the *deleted estimate* is in a way a *much better* behaving estimator of the true risk than the resubstitution estimate.

5. Main Tool

The main tool for our work is an extension of the celebrated Efron-Stein inequality (Efron and Stein, 1981; Steele, 1986), to a stronger version known as the exponential Efron-Stein inequality (Boucheron et al., 2003). We start by introducing the Efron-Stein inequality and some variations. Let $f : \mathcal{X}^n \longmapsto \mathbb{R}$ be a real-valued function of n variables, where $\mathcal{X}$ is a measurable space. Let $X_1, \ldots, X_n$ be independent (not necessarily identically distributed) RVs taking values in $\mathcal{X}$ and define the RV $Z = f(X_1, \ldots, X_n) \equiv f(\mathcal{S}_n)$. Define the shorthand for the conditional expectation $\mathbb{E}_{-i}Z \doteq \mathbb{E}\left[Z|\mathcal{S}_n^{-i}\right]$, where $\mathcal{S}_n^{-i}$ is defined as in the previous section. Informally, $\mathbb{E}_{-i}Z$ "integrates" Z over X_i and *also over any other source of randomness* in Z except $\mathcal{S}_n^{-i}$. For every $i = 1, \ldots, n$, let X_i' be an independent copy from X_i, and let $Z_i' = f(X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n)$. The Efron-Stein inequality bounds the variance of Z as shown in the following theorem.

Theorem 10 (Efron-Stein Inequality – Replacement Case) *Let $V = \sum_{i=1}^n (Z - \mathbb{E}_{-i}Z)^2$. Under the settings described in this section, it holds that $\mathbb{V}[Z] \leq \mathbb{E}V = \frac{1}{2}\sum_{i=1}^n \mathbb{E}[(Z - Z_i')^2]$.*

The proof of Theorem 10 can be found in (Boucheron et al., 2004). Another variant of the Efron-Stein inequality that is more useful for our context, is concerned with the removal of one example from $\mathcal{S}_n$. To state the result, let $f_i : \mathcal{X}^{n-1} \longmapsto \mathbb{R}$, for $1 \leq i \leq n$, be an arbitrary measurable function, and define the RV $Z_{-i} = f_i(\mathcal{S}_n^{-i})$. Then, the Efron-Stein inequality can be also stated in the following interesting form (Boucheron et al., 2004).

Corollary 11 (Efron-Stein Inequality – Removal Case) *Assume that $\mathbb{E}_{-i}[Z_{-i}]$ exists for all $1 \leq i \leq n$, and let $V_{DEL} = \sum_{i=1}^n (Z - Z_{-i})^2$. Then it holds that*

$$\mathbb{V}[Z] \leq \mathbb{E}V \leq \mathbb{E}V_{DEL} . \tag{8}$$

The proof of Corollary 11 is given in Appendix A. This is a standard proof and it is replicated here for the reader's benefit.

5.1. An Exponential Efron-Stein Inequality

The work of Boucheron et al. (2003) has focused on controlling the tail of general functions of independent RVs in terms of the tail behavior of Efron-Stein variance-like terms such as V and V_{DEL}, as well as other terms known as V^+ and V^-. The variance-like terms V, V^+ and V^- measure the sensitivity of a function of n independent RVs w.r.t the *replacement* of one RV from the n independent RVs. The term V_{DEL} on the other hand, measures the sensitivity of a function of n independent RVs w.r.t the *removal* of one RV from the n independent RVs. In this work, we favor V_{DEL} over the other terms since it is more suitable for our choice of stability coefficient (the L_q stability), which is also a removal version. The removal version of stability is preferred as it is more natural in the learning context where one is given a fixed sample. In particular, the removal version seems to be a better fit when it comes to empirically estimating stability (which is an interesting future direction), where working with the replacement version will need extra data, or extra assumptions.

The tail of a RV is often controlled through bounding the logarithm of the moment generating function (MGF) of the RV. This is known as the *cumulant generating function* (CGF) of the RV and is defined as

$$\psi_Z(\lambda) \doteq \log \mathbb{E}\left[\exp(\lambda Z)\right], \tag{9}$$

where $\lambda \in \text{dom}(\psi_Z) \subset \mathbb{R}$ and belongs to a suitable neighborhood of zero. The main result of Boucheron et al. (2003) bounds ψ_Z in terms of the MGF for V, V^+ and V^-, but not in terms of the MGF for V_{DEL}. Since we are particularly interested in the RV V_{DEL}, the following theorem bounds the tail of ψ_Z in terms of the MGF for V_{DEL}.

Theorem 12 *Let Z, V_{DEL} be defined as in Corollary 11 and assume that $|Z - Z_{-i}| \leq 1$ almost surely for all i. For all $\theta > 0$, s.t. $\lambda \in (0, 1]$, $\theta\lambda < 1$, and $\mathbb{E}e^{\lambda V_{DEL}} < \infty$, the following holds*

$$\log \mathbb{E}\left[\exp\left(\lambda(Z - \mathbb{E}Z)\right)\right] \leq \tfrac{\lambda\theta}{(1-\lambda\theta)} \log \mathbb{E}\left[\exp\left(\tfrac{\lambda V_{DEL}}{\theta}\right)\right]. \tag{10}$$

The proof of Theorem 12 is given in Appendix B. Theorem 12 states that the CGF of the centered RV $Z - \mathbb{E}Z$ is upper bounded by the CGF of the RV V_{DEL}. Hence, when V_{DEL} behaves "nicely", the (upper) tail of Z can be controlled. The value of θ in the upper bound is a free parameter that can be optimized to give the tightest bound. Because $\lambda > 0$, the bound is Eq. (10) is only for the upper tail of the RV Z. A similar bound for the lower tail can be obtained by replacing Z with $-Z$ and applying the result. Note also that for both sides, upper tail and lower tail, the same requirements for λ and θ in Theorem 12 apply.

For Theorem 12 to be useful in our context, further control is required to upper bound the tail of the RV V_{DEL}. Our approach to control the tail of V_{DEL} will be, again, through its CGF. In particular, we aim to show that when V_{DEL} is a sub-gamma RV (defined shortly) we can obtain a high probability tail bound on the deviation of the RV Z. The obtained tail bound will be instrumental in deriving the exponential tail bound for the deleted estimate.

5.2. Sub-Gamma Random Variables

We follow here the notation of Boucheron et al. (2013). A real valued centered RV X is said to be *sub-gamma* on the right tail with variance factor v and scale parameter c if for every λ such that

$0 < \lambda < 1/c$, the following holds

$$\psi_X(\lambda) \quad \leq \quad \frac{\lambda^2 v}{2(1-c\lambda)} \,. \tag{11}$$

This is denoted by $X \in \Gamma_+(v,c)$. Similarly, X is said to be a sub-gamma RV on the left tail with variance factor v and scale parameter c if $-X \in \Gamma_+(v,c)$. This is denoted as $X \in \Gamma_-(v,c)$. Finally, X is simply a sub-gamma RV with variance factor v and scale parameter c if both $X \in \Gamma_+(v,c)$ and $X \in \Gamma_-(v,c)$. This is denoted by $X \in \Gamma(v,c)$.

The sub-gamma property can be characterized in terms of moments conditions or tail bounds. In particular, if a centered RV $X \in \Gamma(v,c)$, then for every $t > 0$,

$$\mathbb{P}\left[X > \sqrt{2vt} + ct\right] \vee \mathbb{P}\left[-X > \sqrt{2vt} + ct\right] \leq e^{-t} \,, \tag{12}$$

where $a \vee b = \max(a,b)$. The following theorem from (Boucheron et al., 2013) characterizes this notion more precisely:

Theorem 13 *Let X be a centered RV. If for some $v > 0$ and $c \geq 0$*

$$\mathbb{P}\left[X > \sqrt{2vt} + ct\right] \vee \mathbb{P}\left[-X > \sqrt{2vt} + ct\right] \leq e^{-t} \,, \text{ for every } t > 0 \,, \tag{13}$$

then for every integer $q \geq 1$

$$\|X\|_{2q} \;\leq\; (q!A^q + (2q)!B^{2q})^{1/2q} \;\leq\; \sqrt{16.8qv} \vee 9.6qc \;\leq\; 10(\sqrt{qv} \vee qc) \,,$$

where $A = 8v$, $B = 4c$. Conversely, if for some positive constants u and w, for any integer $q \geq 1$,

$$\|X\|_{2q} \;\leq\; \sqrt{qu} \vee qw \,,$$

then $X \in \Gamma(v,c)$ with $v = 4(1.1u + 0.53w^2)$ and $c = 1.46w$, and therefore (13) also holds.

The reader may notice that Theorem 13 is slightly different than the version in the book of Boucheron et al. (2013). Our extension is based on simple (and standard) calculations that are merely for convenience with respect to our purpose.

5.3. An Exponential Tail Bound for Z

In this section we assume that the centered RV $V_{\text{DEL}} - \mathbb{E}V_{\text{DEL}} \in \Gamma(v,c)$ with variance factor $v > 0$, scale parameter $c \geq 0$, $0 < c|\lambda| < 1$. Hence, from inequality (11) it holds that

$$\psi_{V_{\text{DEL}} - \mathbb{E}V_{\text{DEL}}}(\lambda) = \log \mathbb{E}\left[\exp(\lambda(V_{\text{DEL}} - \mathbb{E}V_{\text{DEL}}))\right] \leq \tfrac{1}{2}\lambda^2 v(1-c|\lambda|)^{-1} \,.$$

The sub-gamma property of V_{DEL} provides the desired control on its tail. That is, after arranging the terms of the above inequality, the CGF of V_{DEL} which controls the tail of V_{DEL}, is upper bounded by the deterministic quantities: $\mathbb{E}V_{\text{DEL}}$, the variance v, and the scale parameter c.

It is possible now to use the sub-gamma property of V_{DEL} in the result of the exponential Efron-Stein inequality in Theorem 12. In particular, the following lemma gives an exponential tail bound on the deviation of a function of independent RVs, i.e. $Z = f(X_1, \ldots, X_n)$, in terms of $\mathbb{E}V_{\text{DEL}}$, the variance factor v, and the scale parameter c. This lemma will be our main tool to derive the exponential tail bound on the DEL estimate.

Lemma 14 *Let Z, Z_{-i}, V_{DEL} be as in Corollary 11. If $V_{DEL} - \mathbb{E}V_{DEL}$ is a sub-gamma RV with variance parameter $v > 0$ and scale parameter $c \geq 0$, then for any $\delta \in (0,1)$, $a > 0$, with probability $1-\delta$,*

$$|Z - \mathbb{E}Z| \leq \tfrac{2}{3}(ac + 1/a)\log\left(\tfrac{2}{\delta}\right) + 2\sqrt{(\mathbb{E}V_{DEL} + a^2 v/2)\log\left(\tfrac{2}{\delta}\right)}. \tag{14}$$

The proof of Lemma 14 is given in Appendix C. Parameter a is a free parameter that can be optimized to give the tightest possible bound. In particular, a can be chosen to provide the appropriate scaling for the RV Z such that the bound goes to zero as fast as possible. A typical choice of a would be the inverse standard deviation of Z.

6. Proof of Theorem 6

In this section we derive our main result given in Theorem 6, namely the concentration of the following random quantity

$$|\widehat{R}_{\text{DEL}}(\mathtt{A}, \mathcal{S}_n) - R(\mathtt{A}(\mathcal{S}_n), \mathscr{P})| \,.$$

To bound this RV, we decompose it into three terms

$$\left|\widehat{R}_{\text{DEL}}(\mathtt{A}, \mathcal{S}_n) - R(\mathtt{A}(\mathcal{S}_n), \mathscr{P})\right| \;\leq\; \text{I} + \text{II} + \text{III}\,, \tag{15}$$

where

$$\begin{aligned}
\text{I} &= |\mathbb{E}\widehat{R}_{\text{DEL}}(\mathtt{A}, \mathcal{S}_n) - \widehat{R}_{\text{DEL}}(\mathtt{A}, \mathcal{S}_n)|\,, \\
\text{II} &= |\mathbb{E}R(\mathtt{A}(\mathcal{S}_n), \mathscr{P}) - R(\mathtt{A}(\mathcal{S}_n), \mathscr{P})|\,, \quad \text{and} \\
\text{III} &= |\mathbb{E}R(\mathtt{A}(\mathcal{S}_n), \mathscr{P}) - \mathbb{E}\widehat{R}_{\text{DEL}}(\mathtt{A}, \mathcal{S}_n)|\,.
\end{aligned}$$

If the three terms in the RHS of (15) are properly upper bounded, we will have the desired final high probability bound. Terms I and II shall be bounded using the exponential Efron-Stein inequality in Lemma 14. Further, we hope that the final upper bounds can be in terms of the L_q stability coefficient of $\mathtt{A}$. Term III, however, is non-random and thus shall be directly bounded using some L_q stability coefficient.

For terms I and II, the key quantity for using the exponential Efron-Stein inequality in Lemma 14 is the RV V_{DEL}. In particular, the requirement for using V_{DEL} is two-fold. First, since $V_{\text{DEL}} = \sum_{i=1}^n (Z - Z_{-i})^2$, where $Z_{-i} = f_i(\mathcal{S}_n^{-i})$ for some function f_i, we need to choose f_i appropriately. Second, once Z_{-i} is defined, to be able to use Lemma 14 we need to show that V_{DEL} is a sub-gamma RV. For this, using Theorem 13, it will be sufficient to show that for all integers $q \geq 1$,

$$\|V_{\text{DEL}}\|_{2q} \leq \sqrt{qu} \vee qw\,, \tag{16}$$

for some positive constants u and w. Here, we will relate $\|V_{\text{DEL}}\|_{2q}$ to L^q stability coefficients and then we "reverse engineer" appropriate assumptions on the L^q-stability coefficients that imply (16).

6.1. Upper Bounding Term I

We begin by deriving an upper bound for term I in the RHS of (15). This is the deviation $|\mathbb{E}\widehat{R}_{\text{DEL}}(\mathtt{A},\mathcal{S}_n) - \widehat{R}_{\text{DEL}}(\mathtt{A},\mathcal{S}_n)|$. Note that $\widehat{R}_{\text{DEL}}(\mathtt{A},\mathcal{S}_n)$ is a function of n independent random variables, and hence the Exponential Efron-Stein inequality from Lemma 14 can be applied to bound this deviation. Following our two-steps plan to use Lemma 14, we define the random variables Z and Z_{-i} as follows

$$Z \;=\; \widehat{R}_{\text{DEL}}(\mathtt{A},\mathcal{S}_n) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(\mathtt{A}(\mathcal{S}_n^{-i}), X_i\right), \qquad Z_{-i} \;=\; \frac{1}{n}\sum_{\substack{j=1\\ j\neq i}}^{n} \ell\left(\mathtt{A}(\mathcal{S}_n^{-i,-j}), X_j\right), \tag{17}$$

where $\mathcal{S}_n^{-i,-j}$ indicates the removal of examples X_i and X_j from $\mathcal{S}_n$. Note that $Z_{-i} = \frac{n-1}{n}\widehat{R}_{\text{DEL}}(\mathtt{A},\mathcal{S}_n^{-i})$ – the scaling factor is chosen to minimize the bound soon to be presented. Recall that $V_{\text{DEL}} = \sum_i (Z - Z_{-i})^2$, and given the definition of Z and Z_{-i} in (17), we need to show that V_{DEL} is a sub-gamma RV and derive a bound on $\mathbb{E}V_{\text{DEL}}$. This can be done by bounding the higher order moments of V_{DEL} as stated in the following lemma.

Lemma 15 *Let Z, Z_{-i} be defined as in (17), and let $V_{DEL} = \sum_{i=1}^{n}(Z - Z_{-i})^2$. Then for any real $q \geq 1/2$ and integer $n \geq 2$, the following holds*

$$\|V_{DEL}\|_{2q} \;\leq\; \tfrac{2}{n^2}\sum_{i=1}^{n} \left\|\ell\left(\mathtt{A}(\mathcal{S}_n^{-i}), X_i\right)\right\|_{4q}^2 + 2n\beta_{4q}^2(n-1), \tag{18}$$

and, in particular, $\mathbb{E}V_{DEL} \;\leq\; \frac{2}{n^2}\sum_{i=1}^{n} \left\|\ell\left(\mathtt{A}(\mathcal{S}_n^{-i}), X_i\right)\right\|_2^2 + 2n\beta_2^2(n-1)$.

The proof is given in Appendix D. Lemma 15 gives the desired upper bound for the higher order moments of V_{DEL} including the upper bound for $\mathbb{E}V_{\text{DEL}}$. To use Lemma 14, it remains to show that V_{DEL} is a sub-gamma RV according to the characterization in Theorem 13. As happens, Assumption 2, stated earlier, is sufficient to achieve this.

Corollary 16 *Using the previous definitions, and under Assumption 2, $V_{DEL} \in \Gamma(v_1, c_1)$, where $v_1 = 4(1.1u_1 + 0.53w_1^2)$ and $c_1 = 1.46w_1$.*

The statement of Corollary 16 follows from Lemma 15, and using Assumption 2 and Theorem 13. Plugging the result of Corollary 16 into Lemma 14 gives the desired final upper bound for Term I in the RHS of (15).

Lemma 17 *Suppose that Assumption 2 holds and $n \geq 2$. Then for any $\delta \in (0,1)$ and $a > 0$, with probability $1-\delta$ the following holds*

$$|\mathbb{E}\widehat{R}_{DEL}(\mathtt{A},\mathcal{S}_n) - \widehat{R}_{DEL}(\mathtt{A},\mathcal{S}_n)| \leq \tfrac{2}{3}(1.46aw_1 + \tfrac{1}{a})\log\left(\tfrac{2}{\delta}\right) + 2\sqrt{\left(n\beta_2^2(n-1) + \rho_1(u_1,w_1)\right)\log\left(\tfrac{2}{\delta}\right)},$$

where $\rho_1(u_1,w_1) = 2.2a^2u_1 + 1.07a^2w_1^2$.

Consider now the choice of a in the context of how it may scale with n and its impact on the behavior of this bound. First, note that u_1 and w_1 are controlled by $n\beta_{4q}^2(n-1)$, and from Assumption 1, we assume that $\beta_{4q}^2(n-1)$ is a nonincreasing function of n. If, for example, $n\beta_2^2(n-$

$1) \sim \frac{1}{n^p}$ for some $p > 0$, then $u_1 \sim n^{-2p}$, $w_1 \sim n^{-p}$, and $w_1 \approx \sqrt{u_1}$. The terms in the bound that depend on a scale as $\frac{a}{n^p} + \frac{1}{a}$ with n. Hence, choosing $a = n^{p/2}$, or $a = w_1^{-1/2}$, makes both, the a dependent term, as well as the whole bound, scale with $n^{-p/2}$ as a function of n; i.e. the bound scales as $w_1^{1/2}$, and $w_1^{1/2} = o(1)$ as $n \to \infty$. This translates to $n\beta_{4q}^2(n-1) = o(1)$ as $n \to \infty$; (and in particular, $\beta_2(n-1) = o(n^{-1/2})$) which is sufficient for the consistency of $\widehat{R}_{\text{DEL}}(\mathtt{A}, \mathcal{S}_n)$. A similar condition for consistency was also identified by Bousquet and Elisseeff (2002) and Celisse and Guedj (2016).

6.2. Upper Bounding Term II

Consider now term II in inequality (15). This is the deviation $|\mathbb{E}R\left(\mathtt{A}(\mathcal{S}_n), \mathscr{P}\right) - R\left(\mathtt{A}(\mathcal{S}_n), \mathscr{P}\right)|$. Note that $R\left(\mathtt{A}(\mathcal{S}_n), \mathscr{P}\right)$ is a function of n independent RVs, and therefore, Lemma 14 will be our tool to bound this deviation. Following the steps for upper bounding Term I in the previous section, we need to define the RVs Z and Z_{-i}, and show that V_{DEL} is a sub-gamma RV. Let the RVs Z and Z_{-i} be defined as follows

$$Z \;=\; R\left(\mathtt{A}(\mathcal{S}_n), \mathscr{P}\right), \qquad Z_{-i} \;=\; R\left(\mathtt{A}(\mathcal{S}_n^{-i}), \mathscr{P}\right). \tag{19}$$

Similar to Lemma 15 we have the following result:

Lemma 18 *Let Z and Z_{-i} be defined as in (19) and let $V_{\text{DEL}} = \sum_{i=1}^n (Z - Z_{-i})^2$. Then for any real $q \geq 1/2$, and $n \geq 2$, the following holds*

$$\|V_{\text{DEL}}\|_{2q} \;\leq\; n\beta_{4q}^2(n)\,, \tag{20}$$

and, in particular, $\mathbb{E}V_{\text{DEL}} \;\leq\; n\beta_2^2(n)$.

By Assumption 1, $n \mapsto \beta_q(n)$ is nonincreasing. This, combined with Assumption 2 gives the following result, which parallels Corollary 16:

Corollary 19 *Using the previous definitions, and under Assumptions 1 and 2, $V_{\text{DEL}} \in \Gamma(v_1, c_1)$, where $v_1 = 4(1.1u_1 + 0.53w_1^2)$ and $c_1 = 1.46w_1$.*

The steps to derive the final bound for Term II are exactly the same derivation steps for the previous bound. The final bound is given by the following lemma which simply plugs in the results of Lemma 18 and Corollary 19 into Lemma 14.

Lemma 20 *Suppose that Assumptions 1 and 2 hold and $n \geq 2$. Then, for any $\delta \in (0,1)$ and $a > 0$, with probability $1 - \delta$ the following holds*

$$|\mathbb{E}R\left(\mathtt{A}(\mathcal{S}_n), \mathscr{P}\right) - R\left(\mathtt{A}(\mathcal{S}_n), \mathscr{P}\right)| \leq \tfrac{2}{3}(1.46aw_1 + \tfrac{1}{a})\log\left(\tfrac{2}{\delta}\right) + 2\sqrt{(n\beta_2^2(n) + \rho_1(u_1, w_1))\log\left(\tfrac{2}{\delta}\right)}\,,$$

where, as before, $\rho_1(u_1, w_1) = 2.2a^2u_1 + 1.07a^2w_1^2$.

Concerning the choice of a, the discussion after Lemma 17 applies.

6.3. Upper Bounding Term III

For term III in inequality (15) there are no random quantities to account for since both terms in the modulus are expectations of RVs. Hence, an upper bound on this deviation will always hold.

Lemma 21 *Using the previous setup and definitions, let* $\mathtt{A}$ *be a learning rule with* L_2 *stability coefficient* $\beta_2(n)$*. Then for* $n \geq 2$*, the following holds*

$$|\mathbb{E}R(\mathtt{A}(\mathcal{S}_n), \mathscr{P}) - \mathbb{E}\widehat{R}_{\scriptscriptstyle DEL}(\mathtt{A}, \mathcal{S}_n)| \;\leq\; \beta_1(n) \leq \beta_2(n)\;. \tag{21}$$

6.3.1. PROOF OF THEOREM 6

At this point, we have obtained the three desired upper bounds for each term in the RHS of inequality (15). The proof of Theorem 6 starts by plugging the results of Lemma 17, Lemma 20, and Lemma 21 into inequality (15) and then simplifying the expression to improve the presentation of the final result.

7. Example (Application to Unbounded Ridge Regression)

In this section we apply the exponential tail bound in Theorem 6 to the ridge regression rule with bounded covariates and *unbounded response variables*. Note that in the presence of unbounded response variables, ridge regression is *not* uniformly stable. In particular, the bound of Bousquet and Elisseeff (2002) is not applicable in this setting. We follow the setup of Celisse and Guedj (2016) (except that we allow unbounded response variables) and we will borrow some results from their work. Let the data be $(\mathbf{x}_1, Y_1), \ldots, (\mathbf{x}_n, Y_n)$, $\mathbf{x}_i \in \mathbb{R}^d$, $Y_i \in \mathbb{R}$ $(1 \leq i \leq n)$, and fix $\lambda > 0$. The ridge regression estimator $\mathtt{A}_\lambda$ is defined via

$$\begin{aligned}\mathtt{A}_\lambda(\mathcal{S}_n) &= \operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^d} \left\{ \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathbf{w}^\top \mathbf{x}_i)^2 + \lambda \left\|\mathbf{w}\right\|_2^2 \right\} \\ &= (\widehat{\mathbf{\Sigma}} + n\lambda \mathbf{I}_d)^{-1}\mathbf{X}^\top \mathbf{y}\;,\end{aligned} \tag{22}$$

where $\mathbf{X}$ is the $n \times d$ matrix obtained by stacking the d-dimensional vectors $\mathbf{x}_1^\top, \ldots, \mathbf{x}_n^\top$, $\widehat{\mathbf{\Sigma}} = \mathbf{X}^\top\mathbf{X}$ is the (unnormalized) sample covariance matrix, and $\mathbf{y} = [Y_1, \ldots, Y_n]^\top$ is the vector of response variables. The loss $\ell(\cdot)$ is the quadratic loss: $\ell(\mathbf{w}, (\mathbf{x}, y)) = (\mathbf{w}^\top x - y)^2$. As usual, we assume that the data is i.i.d. from some common distribution. For the purpose of this example, we have the following two assumptions on this distribution:

Assumption 4 $\exists\, 0 < B_X < +\infty$ *s.t.* $\|\mathbf{x}_1\| \leq B_X$ *a.s.*

Assumption 5 $\exists\, u_Y, w_Y \geq 0$ *s.t.* $\forall q \geq 1$, $\left\|Y_1^4\right\|_{2q} \leq \sqrt{qu_Y} \vee qw_Y$.

Note that this last assumption allows unbounded responses, as long as their 4th moment is sub-gamma. For example, $Y_1 = \sqrt{|Z|}\,\mathrm{sgn}(Z)$ with a gaussian Z satisfies this condition.

To use Theorem 6, the L_q stability coefficient for the ridge estimator, or an upper bound on it, needs to be calculated. This is given in the next theorem taken from the paper of Celisse and Guedj (2016). Their result is applicable because the ridge regression estimator is symmetric, and hence, our definition for the L_q stability coefficients then coincides with theirs, as it was noted earlier.[8]

8. The result is streamlined by choosing the value of η in their result to minimize the upper bound on the stability coefficient.

Theorem 22 *Let $\mathtt{A}_\lambda$ be the ridge estimator in Eq. (22) and let Assumption 4 hold. Then, for any sample size $n > 1$, as long as $s_{\lambda,n} = \lambda - \frac{1}{n-1} > 0$, for any $q \geq 1$, $\mathtt{A}_\lambda$ is L_q stable with the following bound on its stability:*

$$\beta_q(\mathtt{A}_\lambda, \ell, \mathscr{P}, n) \leq 2 \left\|Y_1\right\|_{2q}^2 \frac{B_X^2}{n\lambda} \left(1 + \frac{B_X^2 + \lambda}{s_{\lambda,n}}\right) \left(1 + \frac{B_X^2}{\lambda}\right) . \tag{23}$$

To simplify the expression for the upper bound in Eq. (23), let

$$\kappa = 2\frac{B_X^2}{\lambda} \left(1 + \frac{B_X^2 + \lambda}{s_{\lambda,n-1}}\right) \left(1 + \frac{B_X^2}{\lambda}\right) . \tag{24}$$

Then, $\beta_2(n) \leq \frac{\kappa}{n} \left\|Y_1^2\right\|_2$, $n\beta_2^2(n-1) \leq n\frac{\kappa^2}{(n-1)^2} \left\|Y_1^2\right\|_2^2$, Furthermore, $\beta_q(n-1) \leq \frac{\kappa}{n-1} \left\|Y_1\right\|_{2q}^2$ and, hence

$$n\beta_{4q}^2(n-1) \leq \frac{\kappa^2}{n-1} \left\|Y_1\right\|_{8q}^4 = \frac{\kappa^2}{n-1} \left\|Y_1^4\right\|_{2q} .$$

Some calculations gives (cf. Appendix G)

$$\frac{2}{n} \left\|\ell\left(\mathtt{A}_\lambda(\mathcal{S}_n^{-1}), (x_1, Y_1)\right)\right\|_{4q}^2 \leq \frac{4\left\|Y_1^4\right\|_{2q}}{n} \left(1 + \frac{B_X^4}{\lambda^2}\right)^2 . \tag{25}$$

Thus,

$$n\beta_{4q}^2(n-1) + \frac{2}{n} \left\|\ell\left(\mathtt{A}_\lambda(\mathcal{S}_n^{-1}), (x_1, Y_1)\right)\right\|_{4q}^2 \leq \frac{\left\|Y_1^4\right\|_{2q}}{n-1}\hat{\kappa}$$

where

$$\hat{\kappa} = \left(\kappa^2 + 4\left(1 + \frac{B_X^4}{\lambda^2}\right)^2\right) .$$

Thus, to meet Assumption 2, we can choose $u_1 = \frac{\hat{\kappa}^2}{(n-1)^2} u_Y$ and $w_1 = \frac{\hat{\kappa}}{(n-1)} w_Y$. Note that $\hat{\kappa}$ only depends on B_X and λ, but is independent of n. In particular $\hat{\kappa}$ scales with $1/\lambda^6$ (κ scales with $1/\lambda^3$). We can now plug into the simplified version (5) of the bound of Theorem 6 to obtain an exponential tail bound for the deleted estimate for ridge regression.

Corollary 23 *Given all definitions above, let $\widehat{R}_{DEL}(\mathtt{A}_\lambda, \mathcal{S}_n)$ be the deleted estimate for the ridge regression rule, $R(\mathtt{A}_\lambda(\mathcal{S}_n), \mathscr{P})$ be its risk, and assume that Assumption 4 and Assumption 5 hold. Further, let $\mu = \left\|Y_1^2\right\|_2$. Then, for $\delta \in (0,1)$, with probability $1 - \delta$ the following holds*

$$\begin{aligned} |R(\mathtt{A}_\lambda(\mathcal{S}_n), \mathscr{P}) - \widehat{R}_{DEL}(\mathtt{A}_\lambda, \mathcal{S}_n)| &\leq \frac{\kappa\mu}{n} + 4\kappa\mu\sqrt{\tfrac{n}{(n-1)^2}\log\left(\tfrac{2}{\delta}\right)} + \\ &+ 8\sqrt{\tfrac{\hat{\kappa}}{3(n-1)}\left(\sqrt{(2.2u_Y + 1.07w_Y^2)} + \tfrac{1}{3}1.46w_Y\right)\log\left(\tfrac{2}{\delta}\right)} . \end{aligned} \tag{26}$$

Note that as far as we know this is the first bound for the deleted estimate for ridge regression which allows unbounded response variables.The proof of Corollary 23 is straightforward and is

hence omitted. As we see the bound scales with $1/\sqrt{n}$ regardless the value of λ. However, the bound scales quite poorly with $1/\lambda$. This poor scaling is not inherent to ridge regression but follows from the (oversimplified) analysis. However, for now, we leave it to future work to address this defect of our bound. Finally, let us note that while not shown here, a similar bound is available for the resubstitution estimate: The γ_q coefficients show a behavior similar to the β_q coefficients.

The reader may also be wondering about how the presented bound compares with that presented by Celisse and Guedj (2016) in their Theorem 4. Unfortunately, this comparison is meaningless as the bounds here are incorrect. The problem originates in Proposition 3 where on the right-hand side some terms (corresponding to (25)) are missing: In the proof, the authors incorrectly use $(w^\top x - y)^2 - (w^\top x' - y')^2 = (w^\top (x - x') + y' - y)(w^\top (x + x') - y - y')$: It appears that in their calculations, Celisse and Guedj have accidentally dropped the $y' - y$ term from the first term on the RHS. After correcting for this, the L_q norm of Y will appear on the right-hand side in the inequality stated in this proposition, corresponding to the bound (25). We believe that after the mistakes are corrected, one will arrive at a bound that will near identical to ours.

8. Concluding Remarks

In this work we consider the gap between two regimes of stability-based generalization results; (*i*) exponential generalization bounds based on strong notions of stability which are distribution independent and computationally intractable, such as uniform stability, and (*ii*) polynomial generalization bounds based on weaker notions of stability but are distribution dependent and computationally tractable such as hypothesis stability and L_q stability. Using the exponential Efron-Stein inequality we were able to bridge this gap by deriving an exponential concentration bound for L_q stable learning rules, where the loss of the learning rules is expressed in terms of the deleted estimate.

We believe that our result is one step forward on two fronts; (*i*) computing empirical tight confidence intervals for the expected loss of a learning rule where the confidence interval holds with high probability; and (*ii*) understanding the role of stability in the concentration of different empirical loss estimates around their expectations (in supervised and unsupervised learning). For instance, it will be interesting to understand how the stability of a learning rule can guide our choice for k, and hence the fold size, for the KFCV estimate, such that the estimate concentrates well around the expected risk. Last, we second on the question posed by Bousquet and Elisseeff (2002), of whether it is possible to design algorithms that can maximize their own stability while gaining also on performance.

Acknowledgments

We would like to thank our ALT Reviewers and AC for their thoughtful comments which helped us improve the presentation of our manuscript.

References

Shivani Agarwal and Partha Niyogi. Generalization bounds for ranking algorithms via algorithmic stability. *Journal of Machine Learning Research (JMLR)*, 10:441–474, Jun 2009.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Tuning bandit algorithms in stochastic environments. In Marcus Hutter, Rocco A. Servedio, and Eiji Takimoto, editors, *Algorithmic Learning Theory*, pages 150–165. Springer Berlin Heidelberg, 2007.

Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *Proceedings of the Forty-eighth Annual ACM Symposium on Theory of Computing*, STOC'16, pages 1046–1059. ACM, 2016.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. Concentration inequalities using the entropy method. *The Annals of Probability*, 31(3):1583–1614, 2003.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. Concentration inequalities. In Olivier Bousquet, Ulrike von Luxburg, and Gunnar Rätsch, editors, *Advanced Lectures in Machine Learning*, pages 208–240. Springer, 2004.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: a nonasymptotic theory of independence*. Oxford University Press, 2013.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research (JMLR)*, 2:499–526, 2002.

Alain Celisse and Benjamin Guedj. Stability revisited: new generalisation bounds for the leave-one-out. *ArXiv e-prints*, 1608.06412v1, Aug 2016.

Luc Devroye and Terry Wagner. Distribution-free inequalities for the deleted and holdout error estimates. *IEEE Trans. on Information Theory*, 25(2):202–207, 1979a.

Luc Devroye and Terry Wagner. Distribution-free performance bounds for potential function rules. *IEEE Trans. on Information Theory*, 25(5):601–604, 1979b.

Luc Devroye and Terry J. Wagner. Distribution-free performance bounds with the resubstitution error estimate. *IEEE Trans. on Information Theory*, 25(2):208–210, 1979c.

Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

Bradely Efron and Charles M. Stein. The jackknife estimate of variance. *The Annals of Statistics*, 9 (3):586–596, 05 1981.

Andre Elisseeff, Theodoros Evgeniou, and Massimiliano Pontil. Stability of randomized learning algorithms. *Journal of Machine Learning Research (JMLR)*, 6:55–79, Dec 2005.

Vitaly Feldman and Jan Vondrak. Generalization bounds for uniformly stable algorithms. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9747–9757. Curran Associates, Inc., 2018.

Moritz Hardt, Benjamin Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of the 33rd International Conference on Machine Learning*, ICML'16, pages 1225–1234. JMLR.org, 2016.

Sean B. Holden. PAC-like upper bounds for the sample complexity of leave-one-out cross-validation. In *Proc. of the Ninth Annual Conference on Computational Learning Theory*, COLT '96, pages 41–50. ACM, 1996.

Satyen Kale, Ravi Kumar, and Sergi Vassilvitskii. Cross validation and mean-square stability. In *In Proceedings of the Second Symposium on Innovations in Computer Science ICS'2011*, 2011.

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

Samuel Kutin and Partha Niyogi. Almost-everywhere algorithmic stability and generalization error. In *Proc. of Uncertainty in Artificial Intelligence (UAI)*, pages 275–282, 2002.

Ben London, Bert Huang, and Lise Getoor. Stability and generalization in structured prediction. *Journal of Machine Learning Research (JMLR)*, 17(222):1–52, 2016.

Gabor Lugosi and Miroslaw Pawlak. On the posterior-probability estimate of the error rate of nonparametric classification rules. *IEEE Trans. on Information Theory*, 40(2):475–481, Mar 1994.

Pascal Massart. About the constants in Talagrand's concentration inequalities for empirical processes. *The Annals of Probability*, 28(2):863–884, 04 2000.

Andreas Maurer. Algorithmic stability and meta-learning. *Journal of Machine Learning Research (JMLR)*, 6:967–994, Dec 2005.

Volodymyr Mnih, Csaba Szepesvári, and Jean-Yves Audibert. Empirical Bernstein stopping. In *Proceedings of the 25th International Conference on Machine Learning*, ICML'08, pages 672–679. ACM, 2008.

Sayan Mukherjee, Partha Niyogi, Tomaso Poggio, and Ryan Rifkin. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 25:161–193, 2006.

Alexander Rakhlin, Sayan Mukherjee, and Tomaso Poggio. Stability result in learning theory. *Analysis and Applications*, 3(4):397–417, 2005.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research (JMLR)*, 11:2635–2670, 2010.

John Michael Steele. An Efron-Stein inequality for nonsymmetric statistics. *The Annals of Statistics*, 14(2):753–758, June 1986.

Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.

Huan Xu, Constantine Caramanis, and Shie Mannor. Sparse algorithms are not stable: A no-free-lunch theorem. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 34(1):187–193, Jan 2012.

Yu Zhang. Multi-task learning and algorithmic stability. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 3181–3187. AAAI Press, 2015.

Appendix A. Proof of Corollary 11

Corollary 11 (Efron-Stein Inequality – Removal Case) *Assume that $\mathbb{E}_{-i}[Z_{-i}]$ exists for all $1 \leq i \leq n$, and let $V_{DEL} = \sum_{i=1}^{n} (Z - Z_{-i})^2$. Then it holds that*

$$\mathbb{V}[Z] \;\leq\; \mathbb{E}V \;\leq\; \mathbb{E}V_{DEL}\,. \tag{8}$$

Proof For any RV X, we have the following fact: $\mathbb{V}[X] \leq \mathbb{E}[(X-a)^2]$, for any $a \in \mathbb{R}$. Assume that $\mathbb{E}_{-i}[Z_{-i}]$ exists for all $1 \leq i \leq n$, and applying the previous fact conditionally for $\mathcal{S}_n^{-i}$, then $\mathbb{E}_{-i}[(Z - \mathbb{E}_{-i}Z)^2] \leq \mathbb{E}_{-i}[(Z - Z_{-i})^2]$. Taking expectations and summing over all i we get that $\mathbb{E}V \leq \mathbb{E}V_{\text{DEL}}$. Combining the Efron-Stein inequality for RV Z with the previous inequality, we get the desired result. ■

Appendix B. Proof of Theorem 12

Theorem 12 *Let Z, V_{DEL} be defined as in Corollary 11 and assume that $|Z - Z_{-i}| \leq 1$ almost surely for all i. For all $\theta > 0$, s.t. $\lambda \in (0,1]$, $\theta\lambda < 1$, and $\mathbb{E}e^{\lambda V_{DEL}} < \infty$, the following holds*

$$\log \mathbb{E}\left[\exp\left(\lambda(Z - \mathbb{E}Z)\right)\right] \leq \tfrac{\lambda\theta}{(1-\lambda\theta)} \log \mathbb{E}\left[\exp\left(\tfrac{\lambda V_{DEL}}{\theta}\right)\right]. \tag{10}$$

Proof The proof of this theorem relies on the result of Theorem 6.6 in (Boucheron et al., 2013) which we state here for convenience as a proposition without proof.

Proposition 24 *Let $\phi(u) = e^u - u - 1$. Then for all $\lambda \in \mathbb{R}$,*

$$\lambda\mathbb{E}\left[Z\exp(\lambda Z)\right] - \mathbb{E}\left[\exp(\lambda Z)\right]\log\mathbb{E}\left[\exp(\lambda Z)\right] \leq \sum_{i=1}^{n} \mathbb{E}\left[\exp(\lambda Z)\phi\left(-\lambda(Z - Z_{-i})\right)\right]\,. \tag{27}$$

To make use of inequality (27), we need to establish an appropriate upper bound for the RHS of (27). Note that for $u \leq 1$, $\phi(u) \leq u^2$. By assumption $|Z - Z_{-i}| \leq 1$ holds almost surely. Since $0 < \lambda \leq 1$, we get

$$\begin{aligned}\sum_{i=1}^{n} \mathbb{E}\left[\exp(\lambda Z)\phi\left(-\lambda(Z - Z_{-i})\right)\right] &\leq \lambda^2 \sum_{i=1}^{n} \mathbb{E}\left[\exp(\lambda Z)\left(Z - Z_{-i}\right)^2\right] \\ &= \lambda^2 \mathbb{E}\left[V_{\text{DEL}}\exp(\lambda Z)\right]\,.\end{aligned}$$

It follows that (27) can be written as

$$\lambda\mathbb{E}\left[Z\exp(\lambda Z)\right] - \mathbb{E}\left[\exp(\lambda Z)\right]\log\mathbb{E}\left[\exp(\lambda Z)\right] \leq \lambda^2\mathbb{E}\left[\exp(\lambda Z)V_{\text{DEL}}\right]\,. \tag{28}$$

The RHS of the previous inequality has two coupled random variables; $\exp(\lambda Z)$ and V_{DEL}. To make use of (27), we decouple the two random variables using the following useful tool from (Massart, 2000) which we state as a proposition without a proof.

Proposition 25 *For random variable W, and for any $\lambda \in \mathbb{R}$, if $\mathbb{E}\left[\exp(\lambda W)\right] < \infty$, then the following holds*

$$\frac{\mathbb{E}\lambda W\exp(\lambda Z)}{\mathbb{E}\exp(\lambda Z)} \leq \frac{\mathbb{E}\lambda Z\exp(\lambda Z)}{\mathbb{E}\exp(\lambda Z)} - \log\mathbb{E}\exp(\lambda Z) + \log\mathbb{E}\exp(\lambda W)\,. \tag{29}$$

Multiplying both sides of (29) by $\mathbb{E}\exp(\lambda Z)$ and replacing W with V_{DEL}/θ we get that:

$$\mathbb{E}\exp(\lambda Z)V_{\text{DEL}} \leq \theta\left[\mathbb{E}Z\exp(\lambda Z) - \frac{1}{\lambda}\mathbb{E}\exp(\lambda Z)\log\mathbb{E}\exp(\lambda Z) + \frac{1}{\lambda}\mathbb{E}\exp(\lambda Z)\log\mathbb{E}\exp\left(\lambda\frac{V_{\text{DEL}}}{\theta}\right)\right]. \tag{30}$$

Introduce $F(\lambda) = \mathbb{E}\exp(\lambda Z)$, and $G(\lambda) = \log\mathbb{E}\exp(\lambda V_{\text{DEL}})$. Note that $F'(\lambda) = \mathbb{E}Z\exp(\lambda Z)$. Plugging (30) into (28) and using the compact notation $F(\lambda)$, $F'(\lambda)$, and $G(\lambda/\theta)$ we get that:

$$\lambda F'(\lambda) - F(\lambda)\log F(\lambda) \leq \lambda^2\theta\left(F'(\lambda) - \frac{1}{\lambda}F(\lambda)\log F(\lambda) + \frac{1}{\lambda}F(\lambda)G(\lambda/\theta)\right). \tag{31}$$

Dividing both sides by $\lambda^2 F(\lambda)$ and rearranging the terms:

$$\frac{1}{\lambda}\frac{F'(\lambda)}{F(\lambda)} - \frac{1}{\lambda^2}\log F(\lambda) \leq \frac{\theta G(\lambda/\theta)}{\lambda(1-\lambda\theta)}. \tag{32}$$

The rest of the proof continues exactly as the proof of Theorem 2 from (Boucheron et al., 2003): As the left-hand side of the above display is just the derivative of $H(\lambda) = \frac{1}{\lambda}\log F(\lambda)$, (32) is equivalent to $H'(\lambda) \leq \frac{\theta G(\lambda/\theta)}{\lambda(1-\lambda\theta)}$. Recalling that $\lim_{\lambda\to 0+} H(\lambda) = \mathbb{E}[Z]$, the integration of the differential inequality gives $H(\lambda) \leq \mathbb{E}[Z] + \theta\int_0^\lambda \frac{G(s/\theta)}{s(1-s\theta)}\,ds$. Notice that G is convex. This implies that the integrand is a nondecreasing function of s and therefore $\log F(\lambda) \leq \lambda\mathbb{E}[Z] + \frac{\lambda\theta G(\lambda/\theta)}{1-\lambda\theta}$. ■

Appendix C. Proof of Lemma 14

Lemma 14 *Let Z, Z_{-i}, V_{DEL} be as in Corollary 11. If $V_{DEL} - \mathbb{E}V_{DEL}$ is a sub-gamma RV with variance parameter $v > 0$ and scale parameter $c \geq 0$, then for any $\delta \in (0,1)$, $a > 0$, with probability $1-\delta$,*

$$|Z - \mathbb{E}Z| \leq \tfrac{2}{3}(ac + 1/a)\log\left(\tfrac{2}{\delta}\right) + 2\sqrt{(\mathbb{E}V_{DEL} + a^2 v/2)\log\left(\tfrac{2}{\delta}\right)}. \tag{14}$$

Proof Since $V_{\text{DEL}} - \mathbb{E}V_{\text{DEL}} \in \Gamma_+(v,c)$, for any $\lambda \in (0, 1/c)$ we have

$$\psi_{V_{\text{DEL}}-\mathbb{E}V_{\text{DEL}}}(\lambda) = \log\mathbb{E}\left[\exp(\lambda(V_{\text{DEL}} - \mathbb{E}V_{\text{DEL}}))\right] \leq \frac{\lambda^2 v}{2(1-c\lambda)}.$$

Rearranging the terms we get

$$\log\mathbb{E}\left[\exp(\lambda V_{\text{DEL}})\right] \leq \lambda\mathbb{E}V_{\text{DEL}} + \frac{\lambda^2(v/2)}{1-c\lambda}. \tag{33}$$

Combining this with the result of Theorem 12 where we choose $\theta = 1$, we get

$$\psi_{Z-\mathbb{E}Z}(\lambda) \leq \frac{\lambda}{1-\lambda}\left(\lambda\mathbb{E}V_{\text{DEL}} + \frac{\lambda^2(v/2)}{1-c\lambda}\right). \tag{34}$$

We upper bound the term on the right-hand side as follows

$$\begin{aligned}
\frac{\lambda}{1-\lambda}\left(\lambda \mathbb{E}V_{\text{DEL}} + \frac{\lambda^2(v/2)}{1-c\lambda}\right) &= \frac{\lambda}{1-\lambda}\left(\frac{\lambda\mathbb{E}V_{\text{DEL}} - c\lambda^2\mathbb{E}V_{\text{DEL}} + \lambda^2 v/2}{(1-c\lambda)}\right) \\
&\leq \frac{\lambda}{1-\lambda}\left(\frac{\lambda\mathbb{E}V_{\text{DEL}} + \lambda^2(v/2)}{(1-c\lambda)}\right) \\
&= \frac{\lambda^2\mathbb{E}V_{\text{DEL}} + \lambda^3(v/2)}{(1-\lambda)(1-c\lambda)} \\
&\leq \frac{\lambda^2\mathbb{E}V_{\text{DEL}} + \lambda^2(v/2)}{(1-\lambda)(1-c\lambda)} \\
&= \frac{\lambda^2(\mathbb{E}V_{\text{DEL}} + v/2)}{(1-\lambda)(1-c\lambda)} \\
&\leq \frac{\lambda^2(\mathbb{E}V_{\text{DEL}} + v/2)}{(1-(c+1)\lambda)}\,,
\end{aligned}$$

where the last inequality holds provided that $0 < \lambda < 1/(c+1)$. Thus we finally get that

$$\psi_{Z-\mathbb{E}Z}(\lambda) \leq \frac{\lambda^2(\mathbb{E}V_{\text{DEL}} + v/2)}{(1-(c+1)\lambda)}\,. \tag{35}$$

Recall that the Cramer-Chernoff method gives that for any $\lambda > 0$,

$$\mathbb{P}\left[Z > \mathbb{E}Z + t\right] \leq \exp(-(\lambda t - \psi_{Z-\mathbb{E}Z}(\lambda)))\,.$$

This combined with (35), we see that we need to lower bound

$$\lambda t - \psi_{Z-\mathbb{E}Z}(\lambda) \geq \lambda t - \frac{\lambda^2(\mathbb{E}V_{\text{DEL}} + v/2)}{(1-(c+1)\lambda)}\,,$$

where $\lambda \in (0,1] \cap (0, 1/(c+1)) = (0, 1/(c+1))$ can be chosen so that the lower bound is the largest. From Lemma 11 of Boucheron et al. (2003), we have that for any $p, q > 0$,

$$\sup_{\lambda \in [0,1/q)}\left(\lambda t - \frac{\lambda^2 p}{1-q\lambda}\right) \geq \frac{t^2}{4p + 2q(t/3)}\,,$$

and the supremum is attained at

$$\lambda = \frac{1}{q}\left(1 - \left(1 + \frac{qt}{p}\right)^{-1/2}\right)\,.$$

Setting $p = \mathbb{E}V_{\text{DEL}} + v/2$, $q = c+1$, we see that the optimizing λ belongs to $(0, 1/(c+1))$. Hence,

$$\mathbb{P}\left[Z > \mathbb{E}Z + t\right] \;\leq\; \exp\left(\frac{-t^2}{4(\mathbb{E}V_{\text{DEL}} + v/2) + 2(c+1)t/3}\right)\,.$$

The previous inequality gives an exponential bound on the upper tail for the deviation of the RV Z from its expectation.

Finally, letting the right hand side of the previous inequality to equal δ and solving for t then after some further upper bounding to simplify the resulting expression (in particular, using $\sqrt{|a|+|b|} \leq \sqrt{|a|}+\sqrt{|b|}$) and using a union bound to obtain a two-sided tail inequality, we get

$$|Z - \mathbb{E}Z| \;\leq\; \tfrac{2}{3}(c+1)\log\left(\tfrac{2}{\delta}\right) + 2\sqrt{(\mathbb{E}V_{\text{DEL}} + v/2)\log\left(\tfrac{2}{\delta}\right)}\,. \tag{36}$$

The result now follows by applying (36) to $Z' = aZ$, $Z'_{-i} = aZ_{-i}$ and $V'_{\text{DEL}} = \sum_i (Z' - Z'_{-i})^2$. Noting that $V'_{\text{DEL}} = a^2 V_{\text{DEL}} \in \Gamma(a^4 v, a^2 c)$, we get

$$a\,|Z - \mathbb{E}Z| \;\leq\; \tfrac{2}{3}(a^2 c+1)\log\left(\tfrac{2}{\delta}\right) + 2\sqrt{(a^2\mathbb{E}V_{\text{DEL}} + a^4 v/2)\log\left(\tfrac{2}{\delta}\right)}\,.$$

Dividing both sides by a gives the desired inequality. ■

Appendix D. Proof of Lemma 15

Lemma 15 *Let Z, Z_{-i} be defined as in (17), and let $V_{DEL} = \sum_{i=1}^n (Z - Z_{-i})^2$. Then for any real $q \geq 1/2$ and integer $n \geq 2$, the following holds*

$$\|V_{DEL}\|_{2q} \;\leq\; \tfrac{2}{n^2}\sum_{i=1}^n \left\|\ell\left(\mathtt{A}(\mathcal{S}_n^{-i}), X_i\right)\right\|_{4q}^2 + 2n\beta_{4q}^2(n-1)\,, \tag{18}$$

and, in particular, $\mathbb{E}V_{DEL} \;\leq\; \tfrac{2}{n^2}\sum_{i=1}^n \left\|\ell\left(\mathtt{A}(\mathcal{S}_n^{-i}), X_i\right)\right\|_2^2 + 2n\beta_2^2(n-1)$.

Proof Let $q \geq 1$. Then,

$$\|V_{\text{DEL}}\|_q = \left\|\sum_{i=1}^n (Z - Z_{-i})^2\right\|_q \leq \sum_{i=1}^n \left\|(Z - Z_{-i})^2\right\|_q\,, \tag{37}$$

where the inequality is by the triangle inequality. Now, using the definitions of Z and Z_{-1} (cf. Eq. (17)),

$$\begin{aligned}
&(Z - Z_{-1})^2 \\
&\quad= \left(\frac{1}{n}\ell\left(\mathtt{A}(\mathcal{S}_n^{-1}), X_1\right) + \frac{1}{n}\sum_{i=2}^n \left(\ell\left(\mathtt{A}(\mathcal{S}_n^{-i}), X_i\right) - \ell\left(\mathtt{A}(\mathcal{S}_n^{-1,-i}), X_i\right)\right)\right)^2 \\
&\quad\leq \frac{2}{n^2}\ell^2\left(\mathtt{A}(\mathcal{S}_n^{-1}), X_1\right) + 2\left(\frac{1}{n}\sum_{i=2}^n \left(\ell\left(\mathtt{A}(\mathcal{S}_n^{-i}), X_i\right) - \ell\left(\mathtt{A}(\mathcal{S}_n^{-1,-i}), X_i\right)\right)\right)^2 \qquad ((a+b)^2 \leq 2a^2 + 2b^2) \\
&\quad\leq \frac{2}{n^2}\ell^2\left(\mathtt{A}(\mathcal{S}_n^{-1}), X_1\right) + 2\frac{1}{n}\sum_{i=2}^n \left(\ell\left(\mathtt{A}(\mathcal{S}_n^{-i}), X_i\right) - \ell\left(\mathtt{A}(\mathcal{S}_n^{-1,-i}), X_i\right)\right)^2\,. \qquad \text{(Jensen's inequality)}
\end{aligned}$$

Taking the q-norm of both sides, using the triangle inequality and that for any U RV, $\left\|U^2\right\|_q = \|U\|_{2q}^2$, and using the definition of the stability coefficients we get

$$\left\|(Z - Z_{-1})^2\right\|_q \leq \frac{2}{n^2}\left\|\ell\left(\mathtt{A}(\mathcal{S}_n^{-1}), X_1\right)\right\|_{2q}^2 + 2\frac{1}{n}\sum_{i=2}^n \left\|\ell\left(\mathtt{A}(\mathcal{S}_n^{-i}), X_i\right) - \ell\left(\mathtt{A}(\mathcal{S}_n^{-1,-i}), X_i\right)\right\|_{2q}^2\,.$$

An analogous inequality holds for $\left\|(Z-Z_{-j})^2\right\|_q$ with $j>1$. Summing up all these, using that $(\mathtt{A}(\mathcal{S}_n^{-j}),X_j)_j$ share the same distribution and combining with Eq. (37), we get

$$\begin{aligned}\|V_{\text{DEL}}\|_q &\le \frac{2}{n^2}\sum_{j=1}^n \left\|\ell\left(\mathtt{A}(\mathcal{S}_n^{-j}),X_j\right)\right\|_{2q}^2 + 2\frac{1}{n}\sum_{j=1}^n\sum_{i\neq j}\left\|\ell\left(\mathtt{A}(\mathcal{S}_n^{-i}),X_i\right)-\ell\left(\mathtt{A}(\mathcal{S}_n^{-j,-i}),X_i\right)\right\|_{2q}^2 \\ &= \frac{2}{n^2}\sum_{j=1}^n \left\|\ell\left(\mathtt{A}(\mathcal{S}_n^{-j}),X_j\right)\right\|_{2q}^2 + 2\sum_{i=1}^n \underbrace{\frac{1}{n}\sum_{j\neq i}\left\|\ell\left(\mathtt{A}(\mathcal{S}_n^{-i}),X_i\right)-\ell\left(\mathtt{A}(\mathcal{S}_n^{-i,-j}),X_i\right)\right\|_{2q}^2}_{\beta_{2q}^2(n-1)} \\ &= \frac{2}{n^2}\sum_{i=1}^n \left\|\ell\left(\mathtt{A}(\mathcal{S}_n^{-i}),X_i\right)\right\|_{2q}^2 + 2n\beta_{2q}^2(n-1)\,.\end{aligned}$$

Replacing q with $2q$ gives the desired result. ■

Appendix E. Proof of Lemma 18

Lemma 18 *Let Z and Z_{-i} be defined as in (19) and let $V_{DEL}=\sum_{i=1}^n (Z-Z_{-i})^2$. Then for any real $q\ge 1/2$, and $n\ge 2$, the following holds*

$$\|V_{DEL}\|_{2q} \;\le\; n\beta_{4q}^2(n)\,, \tag{20}$$

and, in particular, $\mathbb{E}V_{DEL} \;\le\; n\beta_2^2(n)$.

Proof Let $q\ge 1$. Then, similar to the previous proof,

$$\|V_{\text{DEL}}\|_q = \left\|\sum_{i=1}^n (Z-Z_{-i})^2\right\|_q \le \sum_{i=1}^n \left\|(Z-Z_{-i})^2\right\|_q = \sum_{i=1}^n \|(Z-Z_{-i})\|_{2q}^2 \tag{38}$$

where the last inequality is because for any RV U, $\left\|U^2\right\|_q = \|U\|_{2q}^2$. Then, using the definitions of Z and Z_{-1},

$$\begin{aligned}\|Z-Z_{-1}\|_{2q}^2 &\le \left\|R(\mathtt{A}(\mathcal{S}_n),\mathscr{P})-R(\mathtt{A}(\mathcal{S}_n^{-1}),\mathscr{P})\right\|_{2q}^2 \\ &= \left\|\mathbb{E}\left[\ell\left(\mathtt{A}(\mathcal{S}_n),X\right)-\ell\left(\mathtt{A}(\mathcal{S}_n^{-1}),X\right)|\mathcal{S}_n\right]\right\|_{2q}^2 && \text{(tower rule)} \\ &= \mathbb{E}\left[\left|\mathbb{E}\left[\ell\left(\mathtt{A}(\mathcal{S}_n),X\right)-\ell\left(\mathtt{A}(\mathcal{S}_n^{-1}),X\right)|\mathcal{S}_n\right]\right|^{2q}\right]^{2/(2q)} \\ &\le \mathbb{E}\left[\mathbb{E}\left[\left|\ell\left(\mathtt{A}(\mathcal{S}_n),X\right)-\ell\left(\mathtt{A}(\mathcal{S}_n^{-1}),X\right)\right|^{2q}|\mathcal{S}_n\right]\right]^{2/(2q)} && \text{(Jensen's inequality)} \\ &= \mathbb{E}\left[\left|\ell\left(\mathtt{A}(\mathcal{S}_n),X\right)-\ell\left(\mathtt{A}(\mathcal{S}_n^{-1}),X\right)\right|^{2q}\right]^{2/(2q)} && \text{(tower rule)} \\ &= \left\|\ell\left(\mathtt{A}(\mathcal{S}_n),X\right)-\ell\left(\mathtt{A}(\mathcal{S}_n^{-1}),X\right)\right\|_{2q}^2\end{aligned}$$

An analogous inequality holds for $\|Z-Z_{-i}\|_{2q}^2$ with $i>1$. Summing up all these, combining with (38) we get

$$\|V_{\text{DEL}}\|_q \le n\,\frac{1}{n}\sum_{i=1}^n \left\|\ell\left(\mathtt{A}(\mathcal{S}_n),X\right)-\ell\left(\mathtt{A}(\mathcal{S}_n^{-i}),X\right)\right\|_{2q}^2 = n\beta_{2q}^2(n)\,.$$

Replacing q with $2q$ yields that

$$\|V_{\text{DEL}}\|_{2q} \;\leq\; n\beta_{4q}^2(n)\,.$$

■

Appendix F. Proof of Lemma 21

Lemma 21 *Using the previous setup and definitions, let* $\mathtt{A}$ *be a learning rule with* L_2 *stability coefficient* $\beta_2(n)$*. Then for* $n \geq 2$*, the following holds*

$$\left|\mathbb{E}R(\mathtt{A}(\mathcal{S}_n),\mathscr{P}) - \mathbb{E}\widehat{R}_{\text{DEL}}(\mathtt{A},\mathcal{S}_n)\right| \;\leq\; \beta_1(n) \leq \beta_2(n)\,. \tag{21}$$

Proof To derive a bound on $\left|\mathbb{E}R(\mathtt{A}(\mathcal{S}_n),\mathscr{P}) - \mathbb{E}\widehat{R}_{\text{DEL}}(\mathtt{A},\mathcal{S}_n)\right|$ in terms of L_q-stability, we proceed as follows. First, note that $\mathbb{E}R\left(\mathtt{A}(\mathcal{S}_n),\mathscr{P}\right) = \mathbb{E}\left[\ell\left(\mathtt{A}(\mathcal{S}_n),X\right)\right]$. Second, for $\mathbb{E}\widehat{R}_{\text{DEL}}(\mathtt{A},\mathcal{S}_n)$, we have

$$\begin{aligned}
\mathbb{E}\widehat{R}_{\text{DEL}}(\mathtt{A},\mathcal{S}_n) &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\ell\left(\mathtt{A}\left(\mathcal{S}_n^{-i}\right),X_i\right)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\ell\left(\mathtt{A}\left(\mathcal{S}_n^{-i}\right),X_i\right)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\ell\left(\mathtt{A}\left(\mathcal{S}_n^{-i}\right),X\right)\right] && \text{(by } \textit{i.i.d} \text{ of the examples)}
\end{aligned}$$

where $X \sim \mathscr{P}$ is independent of $\mathcal{S}_n$.[9]

It follows that

$$\begin{aligned}
\left|\mathbb{E}R\left(\mathtt{A}(\mathcal{S}_n),\mathscr{P}\right) - \mathbb{E}\widehat{R}_{\text{DEL}}(\mathtt{A},\mathcal{S}_n)\right| &= \left|\mathbb{E}[\ell\left(\mathtt{A}(\mathcal{S}_n),X\right)] - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\ell(\mathtt{A}(\mathcal{S}_n^{-i}),X)]\right| \\
&= \left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\ell\left(\mathtt{A}(\mathcal{S}_n),X\right) - \ell(\mathtt{A}(\mathcal{S}_n^{-i}),X)\right]\right| \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left|\ell\left(\mathtt{A}(\mathcal{S}_n),X\right) - \ell(\mathtt{A}(\mathcal{S}_n^{-i}),X)\right|\right] && \text{(Jensen's inequality)} \\
&\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\|\ell\left(\mathtt{A}(\mathcal{S}_n),X\right) - \ell(\mathtt{A}(\mathcal{S}_n^{-i}),X)\right\|_1^2} && \text{(Cauchy-Schwartz)} \\
&= \beta_1(n) \leq \beta_2(n)\,, && (39)
\end{aligned}$$

where the last equality uses the definition of β_1, and the last inequality uses that $\beta_q \leq \beta_{q'}$ for $q \leq q'$. ■

9. Note that, of course, as is well known, $\mathbb{E}\left[\ell\left(\mathtt{A}\left(\mathcal{S}_n^{-i}\right),X\right)\right] = \mathbb{E}\left[\ell\left(\mathtt{A}\left(\mathcal{S}_n^{-1}\right),X\right)\right]$ also holds for any $i > 1$, but we will not need this identity here.

Appendix G. Ridge regression: Proving Eq. (25)

For the convenience of the reader, let us restate Eq. (25):

$$\frac{2}{n}\left\|\ell\left(\mathtt{A}_\lambda(\mathcal{S}_n^{-1}),(x_1,Y_1)\right)\right\|_{4q}^2 \leq \frac{4\left\|Y_1^4\right\|_{2q}}{n}\left(1+\frac{B_X^4}{\lambda^2}\right)^2 .$$

Introduce the shorthand $\tilde{w} = \mathtt{A}_\lambda(\mathcal{S}_n^{-1})$. We have

$$\ell\left(\mathtt{A}_\lambda(\mathcal{S}_n^{-1}),(x_1,Y_1)\right) = (x_1^\top \tilde{w} - Y_1)^2 \leq 2(x_1 x^\top \tilde{w})^2 + 2Y_1^2$$

Then, $|x_1^\top \tilde{w}| \leq \|x_1\|\,\|\tilde{w}\| \leq B_X\,\|\tilde{w}\|$ ($\|\cdot\|$ denotes the 2-norm). Introduce the abbreviation $\|Z\|_{2,q} = \left\|\|Z\|\right\|_q$. Hence,

$$\left\||x_1^\top \tilde{w}|^2\right\|_q \leq B_X^2\left\|\|\tilde{w}\|^2\right\|_q = B_X^2\,\|\tilde{w}\|_{2,2q}^2 .$$

Let $\tilde{X} = [x_2 \ldots x_n]^\top$ (thus, $\tilde{X} \in \mathbb{R}^{(n-1)\times d}$, with x_1 left out), $\tilde{Y} = [Y_2,\ldots,Y_n]^\top$ and $\tilde{\Sigma}_\lambda = \tilde{X}^\top\tilde{X} + n\lambda I$ so that $\tilde{w} = \tilde{\Sigma}_\lambda^{-1}\tilde{X}^\top\tilde{Y}$.

We calculate $\|\tilde{w}\| \leq \left\|\tilde{\Sigma}_\lambda^{-1}\right\| \sum_{i=2}^n |Y_i|\,\|x_i\| \leq \frac{B_X}{n\lambda}\sum_{i=2}^n |Y_i|$ and so

$$\|\tilde{w}\|_{2,2q} = \left\|\tilde{\Sigma}_\lambda^{-1}\tilde{X}^\top\tilde{Y}\right\|_{2,2q} \leq \frac{B_X}{\lambda}\,\|Y_1\|_{2q} ,$$

where the second inequality used that Y_1 has the same distribution as Y_i with $i > 1$. Putting things together,

$$\left\|\ell\left(\mathtt{A}_\lambda(\mathcal{S}_n^{-1}),(x_1,Y_n)\right)\right\|_q \leq 2\frac{B_X^4}{\lambda^2}\,\|Y_1\|_{2q}^2 + 2\,\|Y_1\|_{2q}^2 = 2\,\|Y_1\|_{2q}^2\left(1+\frac{B_X^4}{\lambda^2}\right) \tag{40}$$

and thus

$$\frac{2}{n}\left\|\ell\left(\mathtt{A}_\lambda(\mathcal{S}_n^{-1}),(x_1,Y_n)\right)\right\|_{4q}^2 \leq \frac{4}{n}\,\|Y_1\|_{8q}^4\left(1+\frac{B_X^4}{\lambda^2}\right)^2 = \frac{4}{n}\left\|Y_1^4\right\|_{2q}\left(1+\frac{B_X^4}{\lambda^2}\right)^2 ,$$

finishing the proof.

Appendix H. Proof of Proposition 8

By reusing Example 3.11 of Kutin and Niyogi (2002), we show that the following holds:

Proposition 8 *There exist a distribution $\mathscr{P}$ and a learning algorithm $\mathtt{A}$ such that, everywhere,*

$$\lim_{n\to\infty} R(\mathtt{A}(\mathcal{S}_n),\mathscr{P}) - \widehat{R}_{\mathit{RES}}(\mathtt{A},\mathcal{S}_n) > 0 , \tag{7}$$

while $\sup_{q\geq 1}\beta_q(\mathtt{A},\ell,\mathscr{P},n)/q \to 0$ *as* $n\to\infty$.

Proof We consider classification of points of the $[0,1]$ interval with the zero-one loss. Let the response be ± 1. The loss is given by $\ell(g,(x,y)) = \mathbf{I}\left(g(x) \neq y\right)$, where $(x,y) \in [0,1] \times \{\pm 1\}$ and $g : [0,1] \to \{\pm 1\}$. Choose the distribution $\mathscr{P}$ so that the marginal on the input is the uniform distribution: If $(X,Y) \sim \mathscr{P}$, $\mathbb{P}\left[X \in [a,b]\right] = |b-a|$ for $0 \leq a \leq b \leq 1$. Let $\mathscr{P}$ be such that $\eta := \mathbb{P}\left[Y = 1\right] > 0$.

The learning algorithm is what one may want to call a "short-range nearest neighbor classifier". For $n \geq 1$, let $d_n \geq 0$ be the "range" parameter and assume that $d_n = o(1/n)$. For data $\mathcal{S}_n = ((X_1,Y_1),\ldots,(X_n,Y_n))$, let $\mathrm{NN}(\mathcal{S}_n, x) = \mathrm{argmin}_{1\leq i\leq n} |X_i - x|$ be the index of the nearest neighbor of $x \in [0,1]$ in $\mathcal{S}_n$ (with arbitrary tie-breaking). Now, for input x the learning algorithm $\mathtt{A}$ predicts label $Y_{\mathrm{NN}(\mathcal{S}_n,x)}$ if $|\mathrm{NN}(\mathcal{S}_n,x) - x| \leq d_n$ and it predicts label $+1$ otherwise. Clearly,

$$\widehat{R}_{\mathrm{RES}}\left(\mathtt{A}, \mathcal{S}_n\right) = 0\,. \tag{41}$$

Further,

$$\begin{aligned}
\mathbb{P}\left[\mathtt{A}(\mathcal{S}_n)(X) = -1 \,|\, \mathcal{S}_n\right] &\leq \mathbb{P}\left[\,|\mathrm{NN}(\mathcal{S}_n, X) - X| \leq d_n \,|\, \mathcal{S}_n\right] \\
&\leq \sum_{i=1}^{n} \mathbb{P}\left[\,|X - X_i| \leq d_n \,|\, \mathcal{S}_n\right] \\
&\leq 2nd_n\,,
\end{aligned}$$

where the last inequality follows because X is uniformly distributed.

Note that $\{\mathtt{A}(\mathcal{S}_n)(X) \neq Y\} \supset \{Y = -1\} \setminus \{\mathtt{A}(\mathcal{S}_n)(X) = -1\}$, hence, using (41) and that by the independence of (X,Y) and $\mathcal{S}_n$, $\mathbb{P}\left[Y = -1 | \mathcal{S}_n\right] = \mathbb{P}\left[Y = -1\right]$,

$$\begin{aligned}
R\left(\mathtt{A}(\mathcal{S}_n), \mathscr{P}\right) - \widehat{R}_{\mathrm{RES}}\left(\mathtt{A}, \mathcal{S}_n\right) &\geq \mathbb{P}\left[Y = -1\right] - \mathbb{P}\left[\mathtt{A}(\mathcal{S}_n)(X) = -1 | \mathcal{S}_n\right] \\
&\geq \eta - 2nd_n \to \eta \text{ as } n \to \infty\,,
\end{aligned}$$

which establishes Eq. (7). It remains to show that $\beta_q(n) \to 0$. From the definition, since X has a density, the rule is almost surely symmetric, we need to evaluate

$$\beta_q^q(n) = \mathbb{E}\left[\left|\ell\left(\mathtt{A}(\mathcal{S}_n),(X,Y)\right) - \ell\left(\mathtt{A}(\mathcal{S}_n^{-1}),(X,Y)\right)\right|^q\right]\,,$$

where $(X,Y) \sim \mathscr{P}$, independently of $\mathcal{S}_n$. Note that $|\ell\left(\mathtt{A}(\mathcal{S}_n),(X,Y)\right) - \ell\left(\mathtt{A}(\mathcal{S}_n^{-1}),(X,Y)\right)|$ is either zero or one. Let us focus on the case when this difference is nonzero, i.e., the two losses are different. Clearly, if $\mathtt{A}(\mathcal{S}_n)(X) = \mathtt{A}(\mathcal{S}_n^{-1})(X)$ then the two losses were the same (both predictions are compared to the same Y). Hence, if the loss is nonzero, the predictions must be different. It follows that

$$\begin{aligned}
\beta_q^q(n) &= \mathbb{P}\left[\ell\left(\mathtt{A}(\mathcal{S}_n),(X,Y)\right) \neq \ell\left(\mathtt{A}(\mathcal{S}_n^{-1}),(X,Y)\right)\right] \leq \mathbb{P}\left[\mathtt{A}(\mathcal{S}_n)(X) \neq \mathtt{A}(\mathcal{S}_n^{-1})(X)\right] \\
&= \mathbb{E}\left[\,\mathbb{P}\left[\mathtt{A}(\mathcal{S}_n)(X) \neq \mathtt{A}(\mathcal{S}_n^{-1})(X) \,|\, \mathcal{S}_n\right]\,\right]\,.
\end{aligned}$$

Now, $\mathtt{A}(\mathcal{S}_n)(x) \neq \mathtt{A}(\mathcal{S}_n^{-1})(x)$ implies that $|x - X_1| \leq d_n$. Since X is uniformly distributed, independently of $\mathcal{S}_n$, $\mathbb{P}\left[\mathtt{A}(\mathcal{S}_n)(X) \neq \mathtt{A}(\mathcal{S}_n^{-1})(X) \,|\, \mathcal{S}_n\right] \leq \mathbb{P}\left[|X - X_1| \leq d_n \,|\, \mathcal{S}_n\right] \leq 2d_n$. Hence,

$$\beta_q^q(n) \leq 2d_n$$

and as such, $\beta_q(n) \leq \sup_{q\geq 1}(2d_n)^q/q \to 0$ as $n \to \infty$, as a simple calculation shows. ■

Note that $\beta_u = 1$ in this example: For $\mathcal{S}_n$ such that $|X_1 - X_2| < d_{n-1}$, $Y_1 \neq Y_2$, $\mathtt{A}(\mathcal{S}_n)(X_1) \neq \mathtt{A}(\mathcal{S}_n^{-1})(X_1)$, and

$$|\ell\left(\mathtt{A}(\mathcal{S}_n), (X_1, y)\right) - \ell\left(\mathtt{A}(\mathcal{S}_n^{-1}), (X_1, y)\right)| = 1\,.$$

Also, short-range nearest neighbor is of course a terrible algorithm. Nevertheless if one is looking for ways of figuring out whether an algorithm is terrible or not, this example is important in that it shows that one should better look beyond the empirical error. While this seems obvious in retrospect, we have not seen this mentioned in previous literature on comparing error estimation techniques.

Note also that the 1-nearest neighbor rule is also subject to the same phenomenon as the "short-range nearest neighbor rule": Its training error is always zero, while its risk converges to twice the Bayes risk. At the same time, it is stable in the above sense, and the deleted estimate concentrates around its true risk (Devroye et al., 1996).

Appendix I. Proof for Theorem 9

Here, we provide a proof for the exponential tail bound for the generalization gap defined using the empirical error:

Theorem 9 (Resubstitution estimate tail bound) *Using the setup of Theorem 6, but using Assumption 3 in place of Assumption 2, for $\delta \in (0,1)$ and $a > 0$, with probability $1 - 2\delta$ the following holds:*

$$|\widehat{R}_{RES}\left(\mathtt{A}, \mathcal{S}_n\right) - R(\mathtt{A}(\mathcal{S}_n), \mathscr{P})| \leq \beta_1(n) + \gamma_1(n) + 4\sqrt{(n\beta_2^2(n-1) + C_1)\log\left(\tfrac{2}{\delta}\right)} + C_2 \log\left(\tfrac{2}{\delta}\right)\,,$$

where $C_1 = C_1(a)$ and $C_2 = C_2(a)$ are as in Theorem 6.

Proof As before,

$$\begin{aligned}|\widehat{R}_{\text{RES}}\left(\mathtt{A}, \mathcal{S}_n\right) - R\left(\mathtt{A}(\mathcal{S}_n), \mathscr{P}\right)| \leq\ & |\widehat{R}_{\text{RES}}\left(\mathtt{A}, \mathcal{S}_n\right) - \mathbb{E}\widehat{R}_{\text{RES}}\left(\mathtt{A}, \mathcal{S}_n\right)| \\ & + |R\left(\mathtt{A}(\mathcal{S}_n), \mathscr{P}\right) - \mathbb{E}R\left(\mathtt{A}(\mathcal{S}_n), \mathscr{P}\right)| \\ & + |\mathbb{E}\widehat{R}_{\text{RES}}\left(\mathtt{A}, \mathcal{S}_n\right) - \mathbb{E}R\left(\mathtt{A}(\mathcal{S}_n), \mathscr{P}\right)|\,. \end{aligned} \tag{42}$$

To control the first term, one can use the same argument as in Section 6.1 with the difference that we should use

$$Z \;=\; \widehat{R}_{\text{RES}}(\mathtt{A}, \mathcal{S}_n) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(\mathtt{A}(\mathcal{S}_n), X_i\right), \qquad Z_{-i} \;=\; \frac{1}{n}\sum_{\substack{j=1\\ j\neq i}}^{n} \ell\left(\mathtt{A}(\mathcal{S}_n^{-i}), X_j\right), \tag{43}$$

As required, Z_{-i} does not depend on X_i. The proof of Lemma 15 presented in Appendix D goes through verbatim with the necessary adjustments to account for the differences in the definitions of Z and Z_{-i}. To show the differences encountered, note that

$$Z - Z_{-1} = \frac{1}{n}\ell\left(\mathtt{A}(\mathcal{S}_n), X_1\right) + \frac{1}{n}\sum_{i=2}^{n}\left(\ell\left(\mathtt{A}(\mathcal{S}_n), X_i\right) - \ell\left(\mathtt{A}(\mathcal{S}_n^{-1}), X_i\right)\right)\,.$$

Then, following the steps of the proof of Appendix D, we get

$$\|V_{\text{DEL}}\|_q \leq \frac{2}{n^2}\sum_{i=1}^{n}\|\ell\left(\mathtt{A}(\mathcal{S}_n),X_i\right)\|_{2q}^2+\frac{2}{n}\sum_{\substack{i,j=1\\ i\neq j}}^{n}\left\|\ell\left(\mathtt{A}(\mathcal{S}_n),X_i\right)-\ell\left(\mathtt{A}(\mathcal{S}_n^{-j}),X_i\right)\right\|_{2q}^2$$

Now, notice that

$$\frac{1}{n}\sum_i\|\ell\left(\mathtt{A}(\mathcal{S}_n),X_i\right)\|_{2q}^2\leq\underbrace{\frac{1}{n}\sum_i\left\|\ell\left(\mathtt{A}(\mathcal{S}_n),X_i\right)-\ell\left(\mathtt{A}(\mathcal{S}_n^{-i}),X_i\right)\right\|_{2q}^2}_{\gamma_{2q}^2(n)}+\frac{1}{n}\sum_i\left\|\ell\left(\mathtt{A}(\mathcal{S}_n^{-i}),X_i\right)\right\|_{2q}^2.$$

Furthermore,

$$\begin{aligned}&\left\|\ell\left(\mathtt{A}(\mathcal{S}_n),X_i\right)-\ell\left(\mathtt{A}(\mathcal{S}_n^{-j}),X_i\right)\right\|_{2q}\\&\quad\leq\left\|\ell\left(\mathtt{A}(\mathcal{S}_n),X_i\right)-\ell\left(\mathtt{A}(\mathcal{S}_n^{-i}),X_i\right)\right\|_{2q}+\left\|\ell\left(\mathtt{A}(\mathcal{S}_n^{-j}),X_i\right)-\ell\left(\mathtt{A}(\mathcal{S}_n^{-i}),X_i\right)\right\|_{2q}.\end{aligned}$$

For the second term in the RHS we have

$$\begin{aligned}&\left\|\ell\left(\mathtt{A}(\mathcal{S}_n^{-j}),X_i\right)-\ell\left(\mathtt{A}(\mathcal{S}_n^{-i}),X_i\right)\right\|_{2q}\\&\quad\leq\left\|\ell\left(\mathtt{A}(\mathcal{S}_n^{-j}),X_i\right)-\ell\left(\mathtt{A}(\mathcal{S}_n^{-i,-j}),X_i\right)\right\|_{2q}+\left\|\ell\left(\mathtt{A}(\mathcal{S}_n^{-i}),X_i\right)-\ell\left(\mathtt{A}(\mathcal{S}_n^{-i,-j}),X_i\right)\right\|_{2q}.\end{aligned}$$

Combining these inequalities and using $(a+b+c)^2\leq 3(a^2+b^2+c^2)$, we get

$$\begin{aligned}\sum_{i\neq j}\left\|\ell\left(\mathtt{A}(\mathcal{S}_n),X_i\right)-\ell\left(\mathtt{A}(\mathcal{S}_n^{-j}),X_i\right)\right\|_{2q}^2&\leq 3\sum_{i\neq j}\left\|\ell\left(\mathtt{A}(\mathcal{S}_n),X_i\right)-\ell\left(\mathtt{A}(\mathcal{S}_n^{-i}),X_i\right)\right\|_{2q}^2\\&\quad+3\sum_{i\neq j}\left\|\ell\left(\mathtt{A}(\mathcal{S}_n^{-j}),X_i\right)-\ell\left(\mathtt{A}(\mathcal{S}_n^{-i,-j}),X_i\right)\right\|_{2q}^2\\&\quad+3\sum_{\substack{i,j=1\\ i\neq j}}^{n}\left\|\ell\left(\mathtt{A}(\mathcal{S}_n^{-i}),X_i\right)-\ell\left(\mathtt{A}(\mathcal{S}_n^{-i,-j}),X_i\right)\right\|_{2q}^2\\&=3n^2\gamma_{2q}^2(n)+3n^2\gamma_{2q}^2(n-1)+3n^2\beta_{2q}^2(n-1).\end{aligned}$$

Putting things together,

$$\|V_{\text{DEL}}\|_q\leq\frac{2}{n^2}\sum_{i=1}^{n}\left\|\ell\left(\mathtt{A}(\mathcal{S}_n^{-i}),X_i\right)\right\|_{2q}^2+6n(\gamma_{2q}^2(n)+\gamma_{2q}^2(n-1)+\beta_{2q}^2(n-1))+\frac{2}{n}\gamma_{2q}^2(n).$$

Replacing q with $2q$ we get the analogue of Eq. (18). It follows that under Assumption 3 (in place of Assumption 2), Corollary 16 and Lemma 17 will hold with $\widehat{R}_{\text{DEL}}$ replaced by $\widehat{R}_{\text{RES}}$, and β_q replaced with γ_q, but with no other changes.

The second term of the RHS of (42) is controlled in and the derivations here are applicable given that Assumption 3 implies Assumption 2. Previously, the third term was controlled in Section 6.3. Here, we need to change the reasoning a bit. We start by noting that

$$\mathbb{E}\widehat{R}_{\text{RES}}\left(\mathtt{A}, \mathcal{S}_n\right) - \mathbb{E}R\left(\mathtt{A}(\mathcal{S}_n), \mathscr{P}\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\ell\left(\mathtt{A}(\mathcal{S}_n), X_i\right) - \ell\left(\mathtt{A}(\mathcal{S}_n), X\right)\right] ,$$

where $X \sim \mathscr{P}$ is independent of $\mathcal{S}_n$. Define $S_n^{i\backslash x} = (X_1, \ldots, X_{i-1}, x, X_{i+1}, \ldots, X_n)$. Then, $\mathbb{E}\ell\left(\mathtt{A}(\mathcal{S}_n), X_i\right) = \mathbb{E}\ell\left(\mathtt{A}(S_n^{i\backslash X}), X\right)$ and hence

$$\mathbb{E}\left[\ell\left(\mathtt{A}(\mathcal{S}_n), X_i\right) - \ell\left(\mathtt{A}(\mathcal{S}_n), X\right)\right] = \mathbb{E}\left[\ell\left(\mathtt{A}(S_n^{i\backslash X}), X\right) - \ell\left(\mathtt{A}(\mathcal{S}_n), X\right)\right] .$$

Now, taking absolute values, using $|\mathbb{E}[V]| \leq \mathbb{E}[|V|]$, subtracting and adding $\ell\left(\mathtt{A}(\mathcal{S}_n^{-i}), X\right)$, and using the triangle inequality,

$$\begin{aligned}
&\mathbb{E}\left[\left|\ell\left(\mathtt{A}(S_n^{i\backslash X}), X\right) - \ell\left(\mathtt{A}(\mathcal{S}_n), X\right)\right|\right] \\
&\quad \leq \mathbb{E}\left[\left|\ell\left(\mathtt{A}(S_n^{i\backslash X}), X\right) - \ell\left(\mathtt{A}(\mathcal{S}_n^{-i}), X\right)\right|\right] + \mathbb{E}\left[\left|\ell\left(\mathtt{A}(\mathcal{S}_n), X\right) - \ell\left(\mathtt{A}(\mathcal{S}_n^{-i}), X\right)\right|\right] \\
&\quad = \mathbb{E}\left[\left|\ell\left(\mathtt{A}(S_n), X_i\right) - \ell\left(\mathtt{A}(\mathcal{S}_n^{-i}), X_i\right)\right|\right] + \mathbb{E}\left[\left|\ell\left(\mathtt{A}(\mathcal{S}_n), X\right) - \ell\left(\mathtt{A}(\mathcal{S}_n^{-i}), X\right)\right|\right] .
\end{aligned}$$

where the equality follows because the joint distribution of $(S_n^{i\backslash X}, \mathcal{S}_n^{-i}, X)$ is the same as that of (S_n, S_n^{-i}, X_i). Putting things together, we get

$$\begin{aligned}
&\left|\mathbb{E}\widehat{R}_{\text{RES}}\left(\mathtt{A}, \mathcal{S}_n\right) - \mathbb{E}R\left(\mathtt{A}(\mathcal{S}_n), \mathscr{P}\right)\right| \\
&\quad \leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\left|\ell\left(\mathtt{A}(S_n), X_i\right) - \ell\left(\mathtt{A}(\mathcal{S}_n^{-i}), X_i\right)\right|\right] + \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\left|\ell\left(\mathtt{A}(\mathcal{S}_n), X\right) - \ell\left(\mathtt{A}(\mathcal{S}_n^{-i}), X\right)\right|\right] \\
&\quad \leq \gamma_1(n) + \beta_1(n) ,
\end{aligned}$$

where the last inequality is by Cauchy-Schwartz. Combining this with the bounds on the other two terms in (42) gives the desired result. ■