An Exponential Tail Bound for L_q Stable Learning Rules Application to k–Folds Cross–Validation

Karim Abou–Moustafa and **Csaba Szepesvári**
Department of Computing Science
University of Alberta
Edmonton, AB T6G 2E8, Canada

Abstract

We consider a priori generalization bounds developed in terms of cross-validation estimates and the stability of learners. In particular, we first derive an exponential Efron-Stein type tail inequality for the concentration of a general function of n independent random variables. Next, under some reasonable notion of stability, we use this exponential tail bound to analyze the concentration of the k-fold cross-validation (KFCV) estimate around the true risk of a hypothesis generated by a general learning rule. While the accumulated literature has often attributed this concentration to the bias and variance of the estimator, our bound attributes this concentration to the stability of the learning rule and the number of folds k. This insight raises valid concerns related to the practical use of KFCV, and suggests research directions to obtain reliable empirical estimates of the actual risk.

k-Folds cross-validation (KFCV) is a widely used procedure to estimate the empirical risk of a hypothesis obtained from a certain learning rule (Stone 1974; Geisser 1975). It is used in practice with the promise of being more accurate than the training error, while not being overly computationally expensive as the deleted (or the leave-one-out) estimate which is considered an unbiased estimate of the actual risk (under some notion of stability of the learning rules) (Devroye, Györfi, and Lugosi 1996; Blum, Kalai, and Langford 1999). As such, it is natural to ask how well does the KFCV estimate concentrate around the risk of the hypothesis returned by the sought learning rule.

Various works have considered different aspects of this question. Blum, Kalai, and Langford (1999) show that the KFCV estimate is more accurate than the training error based on its variance and higher order moments. Kale, Kumar, and Vassilvitskii (2011), under some notion of stability, show that the averaging taking place in the KFCV estimate leads to a tighter concentration of the estimated risk around its expectation. Note that this is different from considering the concentration of the estimated risk around the actual risk of the hypothesis. Cornec (2017), in the spirit of sanity-check bounds (Kearns and Ron 1999), shows that for empirical risk minimizers over VC–classes, the worst case error for the KFCV estimate is not much worse than that of the training error.

In this work we consider the exponential concentration of the KFCV estimate around the actual risk of a hypothesis returned by a stable learning rule under distribution-dependent notions of stability. Our hope is to obtain a high probability generalization bound for the KFCV estimate without being dependent on overly *restrictive notions of stability* such as *uniform stability* (explained below) (Kutin and Niyogi 2002).

Earlier works have derived such concentration results for the *deleted estimate* and learning rules that are *uniformly stable* in the sense that no matter how the input to the learning rule is selected, and no matter what value is used as a test example, replacing (or removing) one example in the input, the prediction loss will change only in a limited fashion (Bousquet and Elisseeff 2002). The stability coefficient of a learning rule is the amount of this change. Bousquet and Elisseeff (2002) considered the concentration of the deleted estimate and resubstitution estimate around the (random) risk of a hypothesis returned by a uniformly stable learning rule. The main observation of Bousquet and Elisseeff (2002) is that uniform stability (a worst-case notion over all training and test examples) allows an elegant use of McDiarmid's inequality, which leads to exponential tail bounds. Kutin and Niyogi (2002) and Rakhlin, Mukherjee, and Poggio (2005) consider a softening of the stringent requirement underlying uniform stability to "almost everywhere" stability. While Kutin and Niyogi (2002) prove their result by extending McDiarmid's inequality, Rakhlin, Mukherjee, and Poggio (2005) used the higher-moment version of the Efron-Stein inequality due to Boucheron, Lugosi, and Massart (2003).

Uniform stability is unpleasantly restrictive: Unlike other notions of stability (e.g., L^2, or L^1 stability), it is insensitive to the data-generating distribution. This is problematic as 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 (Kutin and Niyogi 2002). Another example when uniform stability fails is regression with unbounded response variables and losses. In addition, as noted earlier, uniform stability is distribution-free and is thus unsuitable to studying finer details of learning.

Since we are interested in the tail properties of KFCV and higher moments are sufficient and necessary to characterize the tails of random variables, it is natural to expect that the whole family of L^q-stability coefficients with $q \geq 1$ would

play a role in determining the tail behavior of KFCV. The advantage of using L^q stability coefficients to uniform (which in a way are close to L^∞ coefficients) is that they are distribution dependent and are nontrivial even when the uniform stability coefficient is uncontrolled. Recent, yet unpublished work by Celisse and Guedj (2016) indeed demonstrated that the family of L^q stability coefficients can be successfully used to study the deviation of *deleted estimates*. While we also use the same family of stability coefficients, our work goes beyond the work of Celisse and Guedj (2016) in that we consider distribution dependent concentration bounds for the KFCV estimate. While our techniques resemble those of Celisse and Guedj (2016), we streamline several steps of their proofs. One difference is that we build directly on the elegant Efron-Stein style exponential inequality of Boucheron, Lugosi, and Massart (2003), while Celisse and Guedj (2016) chose a different route.

1 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 $\ell : \mathcal{H} \times \mathcal{X} \to [0, 1]$. 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 $h \in \mathcal{H}$ 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 $X_1, \ldots, X_n$.[1] The goal of the learner is to pick hypotheses with a small *risk* (defined shortly). For readers familiar with learning theory we remark that as opposed to most of statistical learning theory, the only role $\mathcal{H}$ plays is to collect the universe of all choices available to learning rules. In particular, unlike in most of the literature on statistical learning theory, it will not be used to "control the bias of learners".

We assume that a learner is able to process samples of different cardinality. Hence, a learner will be identified with a map $\mathtt{A} : \cup_n \mathcal{X}^n \to \mathcal{H}$. Here, we only consider deterministic learning rules; extension to randomizing learning rules is left for future work. Given a distribution $\mathscr{P}$ on $\mathcal{X}$, and $X \sim \mathscr{P}$, the risk of a *fixed hypothesis* $h \in \mathcal{H}$ is given by $R(h, \mathscr{P}) = \mathbb{E}\left[\ell\left(h, X\right)\right]$. Since $\mathcal{S}_n$ is random, so is $\mathtt{A}(\mathcal{S}_n)$. Therefore, we define the risk of the hypothesis that $\mathtt{A}(\mathcal{S}_n)$ returns by: $R(\mathtt{A}(\mathcal{S}_n), \mathscr{P}) = \mathbb{E}[\ell\left(\mathtt{A}(\mathcal{S}_n), X\right) | \mathcal{S}_n]$. Note that $R(\mathtt{A}(\mathcal{S}_n), \mathscr{P})$ is also a random quantity. 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 RVs In the sequel, we will heavily rely on the q-norm for a random variable (RV). 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 $1 \leq q \leq p \leq +\infty$, the following property holds for the q-norm: $\|\cdot\|_q \leq \|\cdot\|_p$.

[1] The set $\mathcal{X}$ is thus measurable. In general, for the sake of minimizing clutter, we will skip mentioning measurability issues; in particular, all functions are assumed to be measurable as needed.

1.1 Quality Assessment of Learners

Most of statistical learning theory is devoted to answering the following two questions: (*i*) *A posteriori* performance assessment: How well *did* $\mathtt{A}$ work on some data $\mathcal{S}_n$ drawn from some distribution $\mathscr{P}$? (*ii*) *A priori* performance prediction: How well *will* $\mathtt{A}$ perform on data $\mathcal{S}_n$ that will be drawn from some distribution $\mathscr{P}$? For both questions, the answer should be given in terms of the risk $R(\mathtt{A}(\mathcal{S}_n), \mathscr{P})$ of the hypothesis $\mathtt{A}(\mathcal{S}_n)$. Since $\mathcal{S}_n$ and $\mathtt{A}(\mathcal{S}_n)$ are random quantities, in general, the answers to the above questions will be upper bounds, the so-called *generalization bounds*, on the random risk $R(\mathtt{A}(\mathcal{S}_n), \mathscr{P})$ that have a probabilistic nature; i.e. the bounds hold with high probability, or hold for the expected risk $R_n(\mathtt{A}, \mathscr{P}) = \mathbb{E}[\ell\left(\mathtt{A}(\mathcal{S}_n), X\right)]$, or the higher moments of the risk.

The two questions are similar in that both of them concern performance on unseen data (since the definition of the risk involves future unseen data). As a result, often the questions are answered using similar tools. The two questions are also fundamentally different: in the case of the first question the data $\mathcal{S}_n$ that produces the hypothesis $\mathtt{A}(\mathcal{S}_n)$ is already given, while in the second case the data is yet unknown at the time when the question is asked. Correspondingly, we call bounds answering the first question *a posteriori* ("after the fact") bounds, while we call bounds answering the second question *a priori* bounds. Ideal a posteriori bounds depend on both $\mathtt{A}$ and $\mathcal{S}_n$ (i.e., these bounds should be learner- and data-dependent), while in the case of a priori bounds, the bound can at best depend on $\mathtt{A}$ and $\mathscr{P}$ (i.e., they can be learner- and distribution-dependent).

In this paper we consider the second question, i.e., a priori generalization bounds. In particular, we consider a priori generalization bounds developed in terms of cross-validation estimates and the stability of learners.

2 Efron-Stein Concentration Inequalities

The main tool for our work is an extension of the Efron-Stein inequality (Efron and Stein 1981; Steele 1986), to a stronger version known as the exponential Efron-Stein inequality (Boucheron, Lugosi, and Massart 2003). The Efron-Stein inequality is a strong tool itself to bound the variance $\mathbb{V}[Z] \doteq \mathbb{E}[(Z - \mathbb{E}Z)^2]$ of a random variable Z which is a function (call this f) of a number of independent RVs. The idea of the Efron-Stein inequality is to "decompose" the variance into the sum V of variance-like terms that express the sensitivity of the function f to its individual variables in an appropriate manner. Oftentimes, these individual sensitivities are easier to control than the variance directly. The crucial feature of the inequality is that it avoids pessimistic worst-case bounds like those that underly McDiarmid's inequality (McDiarmid 1989). While bounding the variance itself is crucial, we will need exponential concentration bounds on the tails of Z. Such bounds were derived in the work of Boucheron, Lugosi, and Massart (2003), and Boucheron, Lugosi, and Massart (2013). Here, based on the techniques

developed in this groundbreaking work, we derive a new tail inequality which will better suit our purposes.

We start by introducing the Efron-Stein inequality and some variations. The inequalities shown here will be useful in our derivations on their own. Let $f : \mathcal{X}^n \longmapsto \mathbb{R}$ be a real-valued function of n variables, where $\mathcal{X}$ is a measurable space (not necessarily the same as in the previous section). If $X_1, \ldots, X_n$ are independent (not necessarily identically distributed) RVs taking values in $\mathcal{X}$, 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 $\mathcal{S}_n^{-i} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$; i.e., it is the sequence $\mathcal{S}_n$ with example X_i removed. Informally, $\mathbb{E}_{-i}Z$ "integrates" Z over X_i and *also over any other source of randomness* in Z except $\mathcal{S}_n^{-i}$. The celebrated Efron-Stein inequality bounds the variance of Z as shown in the following theorem:

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

The proof of Theorem 1 can be found in (Boucheron, Lugosi, and Massart 2004). Another variant of the Efron-Stein inequality which will turn 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, Lugosi, and Massart 2004, Theorem 6)

Corollary 1 (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*

$$\mathbb{V}[Z] \leq \mathbb{E}V \leq \mathbb{E}V_{DEL} \,. \tag{1}$$

It may be surprising at a first sight that $\mathbb{V}[Z]$ can be bounded in terms of V_{DEL} which relies on the arbitrary functions f_i *unrelated* to f. The proof in (Boucheron, Lugosi, and Massart 2004) reveals that there is no mistake here.

2.1 An Exponential Efron-Stein Inequality

The work of Boucheron, Lugosi, and Massart (2003) has focused on controlling the tail of general functions of independent RVs in terms of the tail behavior of the Efron-Stein variance terms such as V and V_{DEL}, as well as other variance terms known as V^+ and V^- (Boucheron, Lugosi, and Massart 2013). The later variance terms will not be presented here since they do not serve our purpose. 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: $\psi_Z(\lambda) \doteq \log \mathbb{E}\left[\exp(\lambda(Z - \mathbb{E}Z))\right]$, where $\lambda \in \text{dom}(\psi_Z) \subset \mathbb{R}$, and belongs to a suitable neighborhood of zero. The main result of Boucheron, Lugosi, and Massart (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}. The proof is given in the Appendix.

Theorem 2. *Let* $Z = f(X_1, \ldots, X_n)$ *be a real valued function of* n *independent RVs. For all* $\theta > 0$, $\lambda \in (0, 1]$, $\theta\lambda < 1$, *and* $\mathbb{E}e^{\lambda V_{DEL}} < \infty$, *the following holds*

$$\begin{aligned} &\log \mathbb{E}\left[\exp\left(-\lambda(Z - \mathbb{E}Z)\right)\right] \\ &\leq \lambda\theta(1 - \lambda\theta)^{-1} \log \mathbb{E}\left[\exp\left(\lambda\theta^{-1}V_{DEL}\right)\right] \,. \end{aligned} \tag{2}$$

Theorem 2 states that the CGF of the centered RV Z is upper bounded by the CGF of the RV V_{DEL}. Hence, when V_{DEL} behaves "nicely", the tail of Z can be controlled. The value of θ in the upper bound is a free parameter that can be optimized. For Theorem 2 to be useful in our context, further control is required to upper bound the tail of V_{DEL}. Our approach to control the tail of V_{DEL} will be, again, through its CGF. In particular, we will 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 KFCV estimate.

Sub-Gamma RVs: 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

$$\psi_X(\lambda) \leq \tfrac{1}{2}\lambda^2 v(1 - c\lambda)^{-1}. \tag{3}$$

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 $X \in \Gamma_+(v, c)$ 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 tail behavior or moment conditions as follows from Theorem 2.3 stated in (Boucheron, Lugosi, and Massart 2013):

Theorem 3. *Let* X *be a centered RV. If for some* $v > 0$ *and* $c \geq 0$

$$\mathbb{P}\left[X > \sqrt{2vt} + ct\right] \vee \mathbb{P}\left[-X > \sqrt{2vt} + ct\right] \leq e^{-t} \,, \tag{4}$$

for every $t > 0$, *then for every integer* $q \geq 1$

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

where $A = 8v$, $B = 4c$, *and* $x \vee y = \max(x, y)$. *Conversely, if for some positive constants* u *and* w, *for any integer* $q \geq 1$,

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

then (4) *holds with* $v = 4(1.1u + 0.73^2w^2)$ *and* $c = 1.46w$.

The reader may notice that Theorem 3 is slightly different than the original version in (Boucheron, Lugosi, and Massart 2013). Our extension to the main result of Boucheron, Lugosi, and Massart (2013) is based on simple calculations that are merely for convenience with respect to our purpose.

2.2 An Exponential Tail Bound for Z

In this section we assume that V_{DEL} is a sub-gamma RV with variance factor $v > 0$, scale parameter $c \geq 0$, and $c\lambda < 1$.

Hence, from (3) it holds that

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

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. Therefore, it is possible now to use the sub-gamma property of V_{DEL} to extend the result of the exponential Efron-Stein inequality in Theorem 2. 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 for the KFCV estimate. The proof is given in the Appendix.

Lemma 1. *Let the RVs Z, Z_{-i}, and V_{DEL} be defined as above. 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 $\delta \in (0,1)$, $a > 0$, with probability $1-\delta$*

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

Parameter a in the upper bound is a free parameter that can be optimized to provide the tightest possible bound. A typical choice of a would be the inverse standard deviation of Z. Lemma 1 is our first contribution in this work: recalling the definition of the RV Z – a function of n independent RVs – Lemma 1 gives an exponential tail bound on the deviation of Z from its expectation by controlling the tails of its variance-like components Z_{-i} and hence V_{DEL}, which in turn is a sub-gamma RV with bounded higher order moments. In the second contribution, we will use Lemma 1 to develop a high probability generalization bound for the KFCV estimate (which will replace the RV Z) in terms of the "*stability*" of the learning rule. Due to the definition of V_{DEL}, stability of the learning rule will turn to be instrumental in bounding the higher order moments of V_{DEL}, and hence for upper bounding the KFCV estimate. However, to derive the desired bound, it remains to formally define the KFCV estimate, and the notion of stability that will permit us to derive such a high probability bound. We pursue this in the following two sections.

3 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 $h \in \mathcal{H}$ we define the *empirical risk* of h on $\mathcal{S}_n$ as $\widehat{R}(h, \mathcal{S}_n) = \frac{1}{n}\sum_{i=1}^n \ell(h, X_i)$. Plugging $\mathtt{A}(\mathcal{S}_n)$ into $\widehat{R}(\cdot, \mathcal{S}_n)$ we get the *training error* or *resubstitution (RES) estimate* (Devroye and Wagner 1979): $\widehat{R}_{\text{RES}}(\mathtt{A}, \mathcal{S}_n) = \widehat{R}(\mathtt{A}(\mathcal{S}_n), \mathcal{S}_n)$. The resubstitution estimate is often overly "optimistic", i.e., it underestimates the actual risk $R(\mathtt{A}(\mathcal{S}_n), \mathscr{P})$.

The *leave–one–out* or *deleted (DEL) estimate* (Devroye and Wagner 1979) is a common alternative to the resubstitution estimate that aims to correct for this: $\widehat{R}_{\text{DEL}}(\mathtt{A}, \mathcal{S}_n) = \frac{1}{n}\sum_{i=1}^n \ell\left(\mathtt{A}(\mathcal{S}_n^{-i}), X_i\right)$, where $\mathcal{S}_n^{-i}$ is defined as in the previous section. Since $\mathbb{E}[\ell\left(\mathtt{A}(\mathcal{S}_n^{-i}), X_i\right)] = R_{n-1}(\mathtt{A}, \mathscr{P})$, 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. 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 is expected to be higher than that of the resubstitution estimate (there is much redundancy in the information content of $\ell(\mathtt{A}(\mathcal{S}_n^{-i}), X_i)$ and $\ell(\mathtt{A}(\mathcal{S}_n^{-j}), X_j)$ for $i \neq j$). Another downside of the deleted estimate is its high computational cost. That is, to evaluate $\widehat{R}_{\text{DEL}}(\mathtt{A}, \mathcal{S}_n)$ for $\mathcal{S}_n$, one has to execute the learner $\mathtt{A}$ on $\mathcal{S}_n^{-i}$ to obtain hypothesis $\widehat{h}_i$, for $i = 1, \ldots, n$; i.e. execute $\mathtt{A}$ for n times. For large n, this is indeed prohibitive.

The KFCV estimate provides a way of naturally interpolating between the resubstitution and the deleted estimate (Stone 1974; Geisser 1975). For simplicity, assume that the sequence $\mathcal{S}_n$ can be partitioned into k equal folds $\mathcal{F}_{1,\ldots,k} \doteq (\mathcal{F}_1 \ \ldots \ \mathcal{F}_k)$, where each fold $\mathcal{F}_j$ is a sequence that has exactly m examples from $\mathcal{S}_n$; i.e. $\mathcal{S}_n = (\mathcal{F}_1\, \mathcal{F}_2 \ldots \mathcal{F}_k)$. In particular, we assume that $n = mk$. This assumption is merely for convenience: all of our results extend to the general case with some extra effort. KFCV proceeds by learning k hypotheses $\widehat{h}_1, \ldots, \widehat{h}_k$, where $\widehat{h}_j = \mathtt{A}(\mathcal{S}_n^{-[\mathcal{F}_j]})$, and $\mathcal{S}_n^{-[\mathcal{F}_j]}$ is the sequence $(\mathcal{F}_1 \ \ldots \ \mathcal{F}_{j-1}\, \mathcal{F}_{j+1} \ \ldots \ \mathcal{F}_k)$. The empirical risk of $\widehat{h}_j$ is obtained by evaluating $\widehat{h}_j$ on $\mathcal{F}_j$ which was "held out" while running $\mathtt{A}$ on $\mathcal{S}_n^{-[\mathcal{F}_j]}$. The KFCV estimate for the risk is the average of the empirical risks of the k hypotheses $\widehat{h}_1, \ldots, \widehat{h}_k$:

$$\widehat{R}_{\text{CV}}(\mathtt{A}, \mathcal{F}_{1,\ldots,k}) = \frac{1}{k}\sum_{j=1}^k \widehat{R}\left(\mathtt{A}(\mathcal{S}_n^{-[\mathcal{F}_j]}), \mathcal{F}_j\right) = \frac{1}{km}\sum_{j=1}^k \sum_{x \in \mathcal{F}_j} \ell\left(\mathtt{A}(\mathcal{S}_n^{-[\mathcal{F}_j]}), x\right) \quad . \quad (5)$$

In the last expression of this display we are abusing the notation by using the membership operator '$\in$' with the sequence $\mathcal{F}_j$. In particular, in the sum every element of the set formed of the members of $\mathcal{F}_j$ appears with its multiplicity in $\mathcal{F}_j$.

Note that we obtain the deleted estimate as a special case of the KFCV estimate when $k = n$ and $m = 1$. The main goal of this paper is to develop a high probability upper bound on the absolute deviation $|\widehat{R}_{\text{CV}}(\mathtt{A}(\mathcal{S}_n), \mathcal{F}_{1,\ldots,k}) - R(\mathtt{A}(\mathcal{S}_n), \mathscr{P})|$ in terms of the "stability" of $\mathtt{A}$, which is defined next.

4 Stability of Learning Rules

We start with the definition of L_q-stability, which specializes to Definition 1 by Celisse and Guedj (2016) when $m = 1$. For $m \in \mathbb{N}$, let $[m] \doteq \{1, \ldots, m\}$. Fix $1 \leq m < n$, and let $\mathcal{S}_n^{-[m]}$ denote the sequence $\mathcal{S}_n$ after removing the first m

examples from it.[2]

Definition 1 (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 1. For $1 \leq m < n$, and $q \geq 1$, the L_q-*stability *coefficient of $\mathtt{A}$ with respect to ℓ, $\mathscr{P}$, and n, m is denoted by $\beta_q(\mathtt{A}, \ell, \mathscr{P}, n, m)$ and is defined as*

$$\beta_q^q(\mathtt{A}, \ell, \mathscr{P}, n, m) = \mathbb{E}[|\widehat{R}(\mathtt{A}(\mathcal{S}_n), \mathcal{F}') - \widehat{R}(\mathtt{A}(\mathcal{S}_n^{-[m]}), \mathcal{F}')|^q],$$

where $\mathcal{F}' = (X_1', \ldots, X_m') \sim \mathscr{P}^m$ is independent of $\mathcal{S}_n$.

Since the examples in $\mathcal{S}_n$ are *i.i.d*, the joint distribution of $(\mathtt{A}(\mathcal{S}_n), \mathtt{A}(\mathcal{S}_n^{-[m]}), X_1', \ldots, X_m')$ does not depend on which (fixed) m examples are removed from $\mathcal{S}_n$, hence, for simplicity, in this definition we simply assume that it is always the first m examples that are removed. Note that quite a few previous works restrict notions of algorithmic stability to learning rules that are permutation invariant, or *"symmetric"*; i.e. learning rules that yield identical output under different permutations of the examples presented to them (Rogers and Wagner 1978; Devroye and Wagner 1979; Kearns and Ron 1999; Bousquet and Elisseeff 2002; Shalev-Shwartz et al. 2010). For the same reason of why it does not matter which examples are removed, it does not matter whether the learning rule is symmetric or not.

Since often $\mathtt{A}$, ℓ, $\mathscr{P}$ are fixed, we will drop them from the notation and will just use $\beta_q^q(n, m)$. However, 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. By avoiding a worst-case approach in the definitions, we will be able to get a finer picture than if we took a worst-case approach.

The L_q-stability coefficient quantifies the variation of the random risk of $\mathtt{A}$ induced by removing m samples from the training set. Often, this is known as a *removal type* notion of stability which is different from (but related to) the *replacement type* notion of stability where the example X_i is replaced with the example X_i' s.t. $X_i' \sim \mathscr{P}$ and X_i' is independent of $\mathcal{S}_n$. This definition of stability is therefore in accordance with previous notions of stability (Rogers and Wagner 1978; Devroye and Wagner 1979; Kearns and Ron 1999; Bousquet and Elisseeff 2002).

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 RV $|\widehat{R}(\mathtt{A}(\mathcal{S}_n), \mathcal{F}') - \widehat{R}(\mathtt{A}(\mathcal{S}_n^{-[m]}), \mathcal{F}')|$. The reason we care about higher moments is because we are interested in controlling the tail behavior of the KFCV estimate. It is then quite expected that the tail behavior of the KFCV 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.

[2] The notation $\mathcal{S}_n^{-i}$, $\mathcal{S}_n^{-\mathcal{F}_j}$, and $\mathcal{S}^{-[m]}$ might be overwhelming at first glance. Indeed, these are all related, and our objective is to simplify the notation. Besides these, we only need $\mathcal{S}_n^{-\{\mathcal{F}_i, \mathcal{F}_j\}}$, which denotes that both folds indexed by $i \neq j$ are removed from $\mathcal{S}_n$.

From the q-Norm properties of RVs, it holds that $\beta_q \leq \beta_p$ for $1 \leq q \leq p \leq +\infty$. As a result, for a fixed ℓ, $\mathtt{A}$ and $\mathscr{P}$, $\beta_q(n, m) \doteq \beta_q(\mathtt{A}, \ell, \mathscr{P}, n, m)$ is an increasing function of q. Furthermore, we also expect that $\beta_q(n, m)$ will be a decreasing function of n and an increasing function of m.

5 Application: An Exponential Tail Bound for The KFCV Estimate

We finally arrive to the main goal of this paper: develop a high probability upper bound on the absolute deviation of $\widehat{R}_{\text{CV}}(\mathtt{A}, \mathcal{F}_{1,\ldots,k})$ from the risk $R(\mathtt{A}(\mathcal{S}_n), \mathscr{P}) \doteq \mathbb{E}[\ell(\mathtt{A}(\mathcal{S}_n), X) | \mathcal{S}_n]$ using the tools developed earlier. To do so, we decompose $|\widehat{R}_{\text{CV}}(\mathtt{A}, \mathcal{F}_{1,\ldots,k}) - R(\mathtt{A}(\mathcal{S}_n), \mathscr{P})|$ into three terms

$$|\widehat{R}_{\text{CV}}(\mathtt{A}, \mathcal{F}_{1,\ldots,k}) - R(\mathtt{A}(\mathcal{S}_n), \mathscr{P})| \leq \text{I} + \text{II} + \text{III}\,, \qquad (6)$$

where

$$\begin{aligned} \text{I} &= |\mathbb{E}\widehat{R}_{\text{CV}}(\mathtt{A}, \mathcal{F}_{1,\ldots,k}) - \widehat{R}_{\text{CV}}(\mathtt{A}, \mathcal{F}_{1,\ldots,k})|, \\ \text{II} &= |R(\mathtt{A}(\mathcal{S}_n), \mathscr{P}) - \mathbb{E}R(\mathtt{A}(\mathcal{S}_n), \mathscr{P})|, \text{ and} \\ \text{III} &= |\mathbb{E}R(\mathtt{A}(\mathcal{S}_n), \mathscr{P}) - \mathbb{E}\widehat{R}_{\text{CV}}(\mathtt{A}, \mathcal{F}_{1,\ldots,k})|. \end{aligned}$$

If the three terms in the RHS of (6) are properly upper bounded, we will have the desired final upper bound. Terms I and II shall be bounded using the exponential Efron-Stein inequality from Lemma 1. Further, we hope that the final upper bounds can be in terms of the L_q stability of $\mathtt{A}$. Term III, however, always holds since it does not involve random quantities, and it shall be directly bounded using the notion of L_q stability.

For terms I and II, the key quantity for using the exponential Efron-Stein inequality is the RV V_{DEL}. In particular, the requirement for using V_{DEL} is two-fold. First, since $V_{\text{DEL}} = \sum_{i=1}^n (Z - Z_{-i})^2$, recall from Corollary 1 that $Z_{-i} = f_i(\mathcal{S}_n^{-i})$, where f_i is an arbitrary function of $n-1$ independent RVs; i.e. $f_i : \mathcal{X}^{n-1} \longmapsto \mathbb{R}$. As such, the first requirement of using V_{DEL} is to choose an appropriate function f_i given our knowledge of the RV Z. Second, once Z_{-i} is defined, we need to show that V_{DEL} is a sub-gamma RV using the characterization of sub-gamma RVs from Theorem 3. For this, from Theorem 3 we know that it suffices to show that for all integers $q \geq 1$,

$$\|V_{\text{DEL}}\|_{2q} \;\leq\; \sqrt{qu} \vee qw\,, \qquad (7)$$

for some positive constants u and w, and $a \vee b = \max(a, b)$. 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 (7).

5.1 Upper Bounding Terms I, II, and III

In this section we derive the desired upper bounds for Terms I, II, and III. Unless otherwise stated, all proofs for the results in this section can be found in the Appendix. First, we consider term I in the RHS of inequality (6).

An Upper Bound for Term I This is the deviation $|\mathbb{E}\widehat{R}_{\text{CV}}(\mathtt{A}, \mathcal{F}_{1,\ldots,k}) - \widehat{R}_{\text{CV}}(\mathtt{A}, \mathcal{F}_{1,\ldots,k})|$. Note that $\widehat{R}_{\text{CV}}(\mathtt{A}, \mathcal{F}_{1,\ldots,k}) \equiv \widehat{R}_{\text{CV}}(\mathtt{A}, \mathcal{F}_1, \ldots, \mathcal{F}_k)$ is a function of k independent random sequences, and each sequence $\mathcal{F}_j$ has m *i.i.d* examples drawn from $\mathcal{X}$ according to $\mathscr{P}$. Hence, the exponential Efron-Stein inequality in Lemma 1 seems to be an appropriate tool to bound this deviation. To use Lemma 1, we need to (*i*) define the RV V_{DEL}, and (*ii*) show that V_{DEL} is a sub-gamma RV. Let the RVs Z and Z_{-i} be defined as follows:

$$\begin{aligned} Z &= \widehat{R}_{\text{CV}}(\mathtt{A}, \mathcal{F}_{1,\ldots,k}) \\ Z_{-i} &= \frac{1}{k}\sum_{j=1,\ j\neq i}^{k} \widehat{R}\left(\mathtt{A}(\mathcal{S}_n^{-\{\mathcal{F}_i,\mathcal{F}_j\}}, \mathcal{F}_j)\right), \end{aligned} \tag{8}$$

where $-\{\mathcal{F}_i, \mathcal{F}_j\}$ indicates the removal of folds $\mathcal{F}_i$ and $\mathcal{F}_j$ from $\mathcal{S}_n = (\mathcal{F}_1\ \mathcal{F}_2\ \cdots\ \mathcal{F}_k)$ and in particular $\mathcal{S}_n^{-\{\mathcal{F}_i,\mathcal{F}_j\}}$ indicates the same sequence regardless of whether $i < j$ or $j < i$. Recall that $V_{\text{DEL}} = \sum_i (Z - Z_{-i})^2$. The first result here is an upper bound on $\mathbb{E}V_{\text{DEL}}$ in terms of the L_2 stability of $\mathtt{A}$.

Lemma 2. *Using the previous setup and definitions, let Z and Z_{-i} be defined as in* (8)*, and let $V_{DEL} = \sum_{i=1}^{k}(Z - Z_{-i})^2$. Then, for $k \geq 1$, and $n > m \geq 1$, the following holds*

$$\mathbb{E}V_{DEL} \ \leq\ k\beta_2^2(n-m, m)\ . \tag{9}$$

The second requirement to use Lemma 1 is to show that V_{DEL} is a sub-gamma RV. To do so, first we need the following lemma to bound the q-norm of V_{DEL} in terms of the L_q stability of $\mathtt{A}$.

Lemma 3. *Using the previous setup and definitions, let Z, Z_{-i}, and V_{DEL} be defined as above. Then for any integer $q \geq 1$, $k \geq 1$, and $n > m \geq 1$, the following holds*

$$\|V_{DEL}\|_{2q} \ \leq\ k\beta_{4q}^2(n-m, m)\ . \tag{10}$$

Next, to show that V_{DEL} is a sub-gamma RV, we need to make the following reasonable assumption.

Assumption 1. *$\exists\ u_1, w_1 \geq 0$ s.t. for any integer $q \geq 1$, it holds that $k\beta_{4q}^2(n-m, m) \leq \sqrt{qu_1} \vee qw_1$.*

This assumption is needed since our results are in terms of the stability of a *generic* learning rule $\mathtt{A}$ with minimal knowledge about it and about its stability. However, once $\mathtt{A}$ is specified, this assumption will not be needed since an upper bound can be realized for β_{4q}^2. For instance, as shown in (Celisse and Guedj 2016), and for the ridge regression case, β_q is upper bounded by the q-norm of the response variable Y.

Corollary 2. *Using the previous definitions, and under Assumption 1, $V_{DEL} \in \Gamma(v_1, c_1)$, where $v_1 = 4(1.1u_1 + 0.73^2 w_1^2)$ and $c_1 = 1.46w_1$.*

The statement of Corollary 2 follows from Lemma 3, and using Assumption 1 and Theorem 3. Plugging the result of Lemma 2 and Corollary 2 into Lemma 1 gives the desired final upper bound.

Lemma 4. *Under Assumption 1, and for $k \geq 1$, and $n > m \geq 1$, let $r_1 = k\beta_2^2(n-m, m)$. Then for $\delta \in (0,1)$ and $a > 0$, with probability $1-\delta$ the following holds*

$$\begin{aligned} &\left|\mathbb{E}\widehat{R}_{CV}(\mathtt{A}, \mathcal{F}_{1,\ldots,k}) - \widehat{R}_{CV}(\mathtt{A}, \mathcal{F}_{1,\ldots,k})\right| \\ &\leq \tfrac{4}{3}(1.46aw_1 + \tfrac{1}{a})\log\left(\tfrac{2}{\delta}\right) \\ &+ 2\sqrt{(r_1 + 2.2a^2u_1 + 1.07(aw_1)^2)\log\left(\tfrac{2}{\delta}\right)}. \end{aligned}$$

Note that u_1 and w_1 are controlled by $k\beta_{4q}^2(n-m, m)$, and that $r_1 = k\beta_2^2(n-m, m)$. Recall that $m = n/k$; i.e. for k fixed, the L_q stability coefficients depend on n. Recall from Section 4 that for fixed $\mathtt{A}$, ℓ, and $\mathscr{P}$, we assume that $\beta_2^2(n-m, m)$ is a decreasing function of n and increasing in m. In particular, similar to Bousquet and Elisseeff (2002) and Celisse and Guedj (2016), for this bound to be useful, $\beta_{4q}^2(n-m, m)$ has to be decreasing as n is increasing, hopefully as fast as possible. If $\beta_{4q}^2(n-m, m) = o(1/\sqrt{n})$, then $\widehat{R}_{\text{CV}}(\mathtt{A}, \mathcal{F}_{1,\ldots,k})$ is a consistent estimator of $\mathbb{E}\widehat{R}_{\text{CV}}(\mathtt{A}, \mathcal{F}_{1,\ldots,k})$. The terms that depend on a from the bound scale as $ak^{-1}n^{-p} + \frac{1}{a}$ with n and k. Choosing $a = k^{1/2}n^{p/2}$ makes both the a-dependent part, as well as the whole of the bound scale with $k^{-1/2}n^{-p/2}$ as a function of n and k. In general, we expect $w_1 \approx \sqrt{u_1}$, in which case choosing $a = w_1^{-1/2}$ makes the bound scale with $w_1^{1/2}$. Then, $w_1^{1/2} = o(1)$ (as $n \to \infty$) will be sufficient for consistency, translating to $\beta_{4q}(n - n/k, n/k) = o(1)$ with k fixed.

An Upper Bound for Term II Second, we consider term II in inequality (6). 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 1 will be our tool to bound this deviation. To do so, 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:

$$\begin{aligned} Z &= R\left(\mathtt{A}(\mathcal{S}_n), \mathscr{P}\right) \\ Z_{-i} &= R\left(\mathtt{A}(\mathcal{S}_n^{-i}), \mathscr{P}\right). \end{aligned} \tag{11}$$

Similarly to Lemma (3) we have the following result:

Lemma 5. *Let Z and Z_{-i} be defined as in* (11) *and let $V_{DEL} = \sum_{i=1}^{n}(Z - Z_{-i})^2$. Then for any real $q \geq 1/2$, and $n \geq 2$, the following holds:*

$$\|V_{DEL}\|_{2q} \ \leq\ n\beta_{4q}^2(n, 1)\ . \tag{12}$$

Lemma 1 also requires a bound on $\mathbb{E}V_{\text{DEL}}$. As before, we obtain this from (12) directly by noticing that $V_{\text{DEL}} \geq 0$, and for $q = 1/2$, $\|V_{\text{DEL}}\|_{2q} = \mathbb{E}V_{\text{DEL}}$:

$$\mathbb{E}V_{\text{DEL}} \ \leq\ n\beta_2^2(n, 1)\ . \tag{13}$$

Then, for the same reason we made Assumption 1, we need to make the following assumption.

Assumption 2. *$\exists\ u_2, w_2 \geq 0$ s.t. for any integer $q \geq 1$, it holds that $n\beta_{4q}^2(n, 1) \leq \sqrt{qu_2} \vee qw_2$.*

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 the results of Lemma (12) and (13) into Lemma 1.

Lemma 6. *Under Assumption 2, and for $n \geq 2$, let $r_2 = n\beta_2^2(n,1)$. Then for $\delta \in (0,1)$ and $a > 0$, with probability $1-\delta$ the following holds*

$$|\mathbb{E}R\left(\mathtt{A}(\mathcal{S}_n),\mathscr{P}\right) - R\left(\mathtt{A}(\mathcal{S}_n),\mathscr{P}\right)| \leq$$
$$\tfrac{4}{3}(1.46aw_2 + \tfrac{1}{a})\log\left(\tfrac{2}{\delta}\right) + 2\sqrt{(r_2 + 2.2a^2u_2 + 1.07a^2w_2^2)\log\left(\tfrac{2}{\delta}\right)}$$

Again, note that u_2 and w_2 are controlled by $\beta_{4q}^2(n,1)$, and that $r_2 = n\beta_2^2(n,1)$. Recall that $m = n/k$; i.e. in particular, for k fixed, the L_q stability coefficients depend on n. From Section 4, for fixed $\mathtt{A}$, ℓ, and $\mathscr{P}$, we assume that $\beta_2^2(n-m,m)$ is a decreasing function of n and increasing in m. As such, for this bound to be useful, $\beta_{4q}^2(n,1)$ has to be decreasing as n is increasing, hopefully as fast as possible. If $\beta_{4q}^2(n,1) = o(1/\sqrt{n})$, then $R\left(\mathtt{A}(\mathcal{S}_n),\mathscr{P}\right)$ is a consistent estimator of $\mathbb{E}R\left(\mathtt{A}(\mathcal{S}_n),\mathscr{P}\right)$. Concerning the choice of a, the discussion after Lemma (4) applies.

An Upper Bound for Term III For term III in inequality (6) there are no random quantities to account for since both terms in the absolute value are expectations of RVs. Hence, an upper bound on this deviation will always hold.

Lemma 7. *Using the previous setup and definitions, let $\mathtt{A}$ be a learning rule with L_2 stability coefficient $\beta_2(n,m)$. Then, for $k \geq 1$, and $n > m \geq 1$, the following holds*

$$|\mathbb{E}R(\mathtt{A}(\mathcal{S}_n),\mathscr{P}) - \mathbb{E}\widehat{R}_{CV}(\mathtt{A},\mathcal{F}_{1,\ldots,k})| \;\leq\; \beta_2(n,m)\;.$$

5.2 Main Result

Theorem 4. *Let $\mathcal{X}$, $\mathcal{H}$ and ℓ be as previously defined. Let $\mathcal{S}_n \doteq (\mathcal{F}_1,\ldots,\mathcal{F}_k)$ be the dataset defined in Section 3, where $k \geq 1$, $n > m \geq 1$, and $n = km$. Let $\widehat{R}_{CV}\left(\mathtt{A},\mathcal{F}_{1,\ldots,k}\right)$ be the cross validation estimate defined in* (5)*, and $R(\mathtt{A}(\mathcal{S}_n),\mathscr{P})$ be the risk for hypothesis $\mathtt{A}(\mathcal{S}_n)$. Then, under Assumption 1 and Assumption 2, for $\delta \in (0,1)$ and $a > 0$, with probability $1-\delta$ the following holds*

$$|\widehat{R}_{CV}\left(\mathtt{A},\mathcal{F}_{1,\ldots,k}\right) - R(\mathtt{A}(\mathcal{S}_n),\mathscr{P})|$$
$$\leq 2(aw_1 + aw_2 + 2)\log\left(\tfrac{4}{\delta}\right) + \beta_2(n,m)$$
$$+ 4\left(\sqrt{\pi_1} + \sqrt{\pi_2}\right)\log\left(\tfrac{4}{\delta}\right)^{\frac{1}{2}},$$

where

$$\pi_1 = 2k\beta_2^2(n-m,m) + 2.2au_1 + 1.07a^2w_1^2, \text{ and}$$
$$\pi_2 = n\beta_2^2(n,1) + 2.2au_2 + 1.07a^2w_2^2.$$

The proof of Theorem 4 simply plugs in the results of Lemma 4, Lemma 6, and Lemma 7 into inequality (6). Concerning the choice of a, the discussions after Lemma (4) and Lemma (6) apply. For consistency, under assumptions stated there, we need $\beta_{4q}(n,1) = o(1/n^{1/2})$ and $\beta_{4q}(n-n/k,n/k) = o(1)$ with k fixed, while we need $\beta_{4q}(n,1) = o(1/n^{1/2})$ only when $k = k_n = n$ (deleted estimate).

Discussion: First, the discussion that follows Lemma 4 and Lemma 6 applies here as well. Second, consider the four terms that constitute the final bound. The first term is due to the higher order moments of the L_q stability, in particular β_{4q}^2 which are the tails for the RV $|\ell\left(\mathtt{A}(\mathcal{S}_n),X\right) - \ell(\mathtt{A}(\mathcal{S}_n^{-1}),X)|^{4q}$. Note that from Theorem 3, w_1 and w_2 are in fact controlled by the stability β_{4q}^2. Therefore, as the stability is improving, w_1 and w_2 will be small, thereby making the bound tighter. Note that the same applies to u_1 and u_2 in π_1 and π_2, respectively.

The four terms in the bound show that the concentration of the KFCV estimate around the true risk depends on the stability of the learning rule, and the number of folds k (and consequently m). For a stable learning rule with small higher order moments, the bound is tightest for the deleted estimate ($k = n$). In a general sense, this agrees with earlier results that, for a stable learning rule, the deleted estimate is *almost an unbiased* estimate of the true risk (Devroye, Györfi, and Lugosi 1996). At a more specific level, our results show that in order to control the concentration of the KFCV estimate around the true risk, one has to control the tail of the RV $\widehat{R}_{\text{CV}}\left(\mathtt{A},\mathcal{S}_n\right)$. And to control the tail of $\widehat{R}_{\text{CV}}\left(\mathtt{A},\mathcal{S}_n\right)$, one has to control the tails, or the higher order moments, for the RV V_{DEL}, or $|\ell\left(\mathtt{A}(\mathcal{S}_n),X\right) - \ell(\mathtt{A}(\mathcal{S}_n^{-1}),X)|$, which turns to be the stability of the learning rule – in terms of the loss function $\ell\left(\mathtt{A}(\mathcal{S}_n),X\right)$ – w.r.t the removal of one example (or m examples) from $\mathcal{S}_n$.

At a higher level, for the KFCV estimate to concentrate around the true risk, and assuming a highly stable learning rule, large values of k will give a better estimate for the generalization error. By contrast, when $k = 1$ (i.e. no cross validation), $\widehat{R}_{\text{CV}}$ turns to be the resubstitution estimate $\widehat{R}_{\text{RES}}\left(\mathtt{A},\mathcal{S}_n\right)$, and the bound becomes loose which agrees with the classical result that $\widehat{R}_{\text{RES}}\left(\mathtt{A},\mathcal{S}_n\right)$ *overly underestimates* the true risk (Devroye and Wagner 1979).

6 Concluding Remarks

As stated earlier, our concentration bound for the KFCV estimate shows that in order for the estimate to concentrate around the true risk, the learning rule $\mathtt{A}$ has to be stable. More specifically, to control the tail of the KFCV estimate, one has to control its higher order moments which, in turn, is controlled by the stability of the learning rule. Depending on the degree of stability, one may then need to increase k as a function of the sample size. This insight was only possible through the interplay between the exponential Efron-Stein inequality and the notion of L_q stability.

According to our results, and considering the practical side of using machine learning algorithms, one has to question the widely used practice of setting k into a predefined value to report the empirical generalization error for a learning rule, regardless of the sample size n; for instance setting $k = 10$ or $k = 5$. Note that for a fixed k, as the sample size is increasing, the size of each fold, $m = n/k$ is also increasing. This implicitly assumes that the learning rule, and in terms of the hypothesis loss, is stable w.r.t the removal of

m examples from the training set. If there is no justification for this stability assumption, one should wonder the faithfulness of such empirical results. This practice is even more alarming in the absence of any empirical measure for the stability of a learning rule. Nevertheless, this also suggests two promising research directions; (*i*) a computationally efficient *mechanism* for choosing the value of k to improve the reliability of the KFCV estimate; and (*ii*) a computationally efficient (meta)algorithm (hopefully with some guarantees) to estimate the stability of a learning rule.

On the theoretical side, our result, so far, is in terms of the stability of the learning rule $\mathtt{A}$, k, n, and m, but did not consider any particular learning rule in specific. As such, further insight and refined results can be obtained if our bound is applied to well known classes of algorithms such as potential function rules (which include the k-NN rule) (Devroye and Wagner 1979), and empirical risk minimizers for instance.

References

[Blum, Kalai, and Langford 1999] Blum, A.; Kalai, A.; and Langford, J. 1999. Beating the hold-out: Bounds for k-fold and progressive cross-validation. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, COLT '99, 203–208. New York, NY, USA: ACM.

[Boucheron, Lugosi, and Massart 2003] Boucheron, S.; Lugosi, G.; and Massart, P. 2003. Concentration inequalities using the entropy method. *The Annals of Probability* 31(3):1583–1614.

[Boucheron, Lugosi, and Massart 2004] Boucheron, S.; Lugosi, G.; and Massart, P. 2004. Concentration inequalities. In Bousquet, O.; von Luxburg, U.; and Rätsch, G., eds., *Advanced Lectures in Machine Learning*. Springer. 208–240.

[Boucheron, Lugosi, and Massart 2013] Boucheron, S.; Lugosi, G.; and Massart, P. 2013. *Concentration inequalities: a nonasymptotic theory of independence*. Oxford University Press.

[Bousquet and Elisseeff 2002] Bousquet, O., and Elisseeff, A. 2002. Stability and generalization. *JMLR* 2:499–526.

[Celisse and Guedj 2016] Celisse, A., and Guedj, B. 2016. Stability revisited: new generalisation bounds for the leave-one-out. *ArXiv e-prints* (1608.06412).

[Cornec 2017] Cornec, M. 2017. Concentration inequalities of the cross-validation estimator for empirical risk minimizer. *Statistics* 51(1):43–60.

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

[Devroye, Györfi, and Lugosi 1996] Devroye, L.; Györfi, L.; and Lugosi, G. 1996. *A Probabilistic Theory of Pattern Recognition*. New York: Springer.

[Efron and Stein 1981] Efron, B., and Stein, C. M. 1981. The jackknife estimate of variance. *Ann. Statist.* 9(3):586–596.

[Geisser 1975] Geisser, S. 1975. The predictive sample reuse method with applications. *Journal of the American Statistical Association* 70(350):320–328.

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

[Kearns and Ron 1999] Kearns, M., and Ron, D. 1999. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. In *COLT*, 152–162. ACM.

[Kutin and Niyogi 2002] Kutin, S., and Niyogi, P. 2002. Almost-everywhere algorithmic stability and generalization error. In *Proc. of Uncertainty in Artificial Intelligence (UAI)*, 275–282.

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

[McDiarmid 1989] McDiarmid, C. 1989. On the method of bounded differences. In *Surveys in Combinatorics*, number 141 in London Mathematical Society Lecture Note Series, 148–188. Cambridge University Press.

[Rakhlin, Mukherjee, and Poggio 2005] Rakhlin, A.; Mukherjee, S.; and Poggio, T. 2005. Stability results in learning theory. *Analysis and Applications* 03(04):397–417.

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

[Shalev-Shwartz et al. 2010] Shalev-Shwartz, S.; Shamir, O.; Srebro, N.; and Sridharan, K. 2010. Learnability, stability and uniform convergence. *JMLR* 11:2635–2670.

[Steele 1986] Steele, J. M. 1986. An Efron-Stein inequality for nonsymmetric statistics. *Ann. Statist.* 14(2):753–758.

[Stone 1974] Stone, M. 1974. Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society. Series B (Methodological)* 36(2):111–147.

[Vapnik 1995] Vapnik, V. N. 1995. *The Nature of Statistical Learning Theory*. Springer-Verlag.

Appendix

Proof of Theorem 2

Theorem. 2 *Let $Z = f(X_1, \ldots, X_n)$ be a real valued function of n independent random variables. For all $\theta > 0$, $\lambda \in (0, 1]$, $\theta\lambda < 1$, and $\mathbb{E}e^{\lambda V_{DEL}} < \infty$:*

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

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

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

$$\lambda\mathbb{E}\left[Z\exp(\lambda Z)\right] - \mathbb{E}\left[\exp(\lambda Z)\right]\log\mathbb{E}\left[\exp(\lambda Z)\right] \leq \sum_{i=1}^{n} \mathbb{E}\left[\exp(\lambda Z)\phi\left(-\lambda(Z - Z_{-i})\right)\right] . \tag{14}$$

To make use of inequality (14), we need to establish an appropriate upper bound for the RHS of (14). Note that for $u \leq 1$, $\phi(u) \leq u^2$. Assume that $|Z - Z_{-i}| \leq 1$ and since $0 < \lambda \leq 1$, then

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

It follows that (14) can be written as

$$\lambda\mathbb{E}\left[Z\exp(\lambda Z)\right] - \mathbb{E}\left[\exp(\lambda Z)\right]\log\mathbb{E}\left[\exp(\lambda Z)\right] \leq \lambda^2\mathbb{E}\left[\exp(\lambda Z)V_{\text{DEL}}\right] . \tag{15}$$

The RHS of the previous inequality has two coupled random variables; $\exp(\lambda Z)$ and V_{DEL}. To make use of (14), we decouple the two random variables using the following useful tool from (Massart 2000) which we state as a proposition without a proof.

Proposition 2. *For random variable W, and for any $\lambda \in \mathbb{R}$, if $\mathbb{E}\left[\exp(\lambda W)\right] < \infty$, then the following holds*

$$\frac{\mathbb{E}\lambda W\exp(\lambda Z)}{\mathbb{E}\exp(\lambda Z)} \leq \frac{\mathbb{E}\lambda Z\exp(\lambda Z)}{\mathbb{E}\exp(\lambda Z)} - \log\mathbb{E}\exp(\lambda Z) + \log\mathbb{E}\exp(\lambda W) . \tag{16}$$

Multiplying both sides of (16) by $\mathbb{E}\exp(\lambda Z)$ and replacing W with V_{DEL}/θ we get that:

$$\mathbb{E}\exp(\lambda Z)V_{\text{DEL}} \leq \theta\left[\mathbb{E}Z\exp(\lambda Z) - \frac{1}{\lambda}\mathbb{E}\exp(\lambda Z)\log\mathbb{E}\exp(\lambda Z) + \frac{1}{\lambda}\mathbb{E}\exp(\lambda Z)\log\mathbb{E}\exp\left(\lambda\frac{V_{\text{DEL}}}{\theta}\right)\right] . \tag{17}$$

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 (17) into (15) and using the compact notation $F(\lambda)$, $F'(\lambda)$, and $G(\lambda/\theta)$ we get that:

$$\lambda F'(\lambda) - F(\lambda)\log F(\lambda) \leq \lambda^2\theta\left(F'(\lambda) - \frac{1}{\lambda}F(\lambda)\log F(\lambda) + \frac{1}{\lambda}F(\lambda)G(\lambda/\theta)\right) . \tag{18}$$

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

$$\frac{1}{\lambda}\frac{F'(\lambda)}{F(\lambda)} - \frac{1}{\lambda^2}\log F(\lambda) \leq \frac{\theta G(\lambda/\theta)}{\lambda(1-\lambda\theta)} . \tag{19}$$

The rest of the proof continues exactly as the proof of Theorem 2 in (Boucheron, Lugosi, and Massart 2003). A slightly different version of this proof was given for Theorem 6.16 in (Boucheron, Lugosi, and Massart 2013). □

Proof of Lemma 1

Lemma 1. *Let the RVs Z, Z_{-i}, and V_{DEL} be defined as above. 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 $\delta \in (0, 1)$, $a > 0$, with probability $1 - \delta$*

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

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

$$\log \mathbb{E}\left[\exp(\lambda V_{\text{DEL}})\right] \leq \lambda \mathbb{E}V_{\text{DEL}} + \frac{\lambda^2(v/2)}{1-c\lambda} . \tag{20}$$

Combining this with the result of 2 where we choose $\theta = 1$, we get

$$\psi_{Z-\mathbb{E}Z}(\lambda) \leq \frac{\lambda}{1-\lambda}\left(\lambda \mathbb{E}V_{\text{DEL}} + \frac{\lambda^2(v/2)}{1-c\lambda}\right) . \tag{21}$$

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

$$\psi_{Z-\mathbb{E}Z}(\lambda) \leq \frac{\lambda^2(\mathbb{E}V_{\text{DEL}} + v/2)}{(1-(c+1)\lambda)} . \tag{22}$$

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 (22), 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 in Boucheron, Lugosi, and Massart (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,

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

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 $\sqrt{|a| + |b|} \leq \sqrt{|a|} + \sqrt{|b|}$), we get

$$|Z - \mathbb{E}Z| \;\leq\; \tfrac{4}{3}(c+1)\log\left(\tfrac{2}{\delta}\right) + 2\sqrt{(\mathbb{E}V_{\text{DEL}} + v/2)\log\left(\tfrac{2}{\delta}\right)} . \tag{23}$$

The result now follows by applying (23) 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

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

Dividing both sides by a gives the desired inequality. □

Proof of Lemma 3

Lemma 3. *Using the previous setup and definitions, let Z, Z_{-i}, and V_{DEL} be defined as above. Then for any integer $q \geq 1$, $k \geq 1$, and $n > m \geq 1$, the following holds*

$$\|V_{DEL}\|_{2q} \;\leq\; k\beta_{4q}^2(n-m,m)\;. \tag{10}$$

Proof. Let $q \geq 1$. Then,

$$\begin{aligned}
\|V_{\text{DEL}}\|_q &= \left\| \sum_{i=1}^{k} (Z - Z_{-i})^2 \right\|_q \\
&\leq \sum_{i=1}^{k} \left\| (Z - Z_{-i})^2 \right\|_q \quad \text{(by triangle inequality)} \\
&= \sum_{i=1}^{k} \left\| \left(\frac{1}{k} \sum_{j=1}^{k} \widehat{R}\left(\mathtt{A}(\mathcal{S}_n^{-\mathcal{F}_j}), \mathcal{F}_j\right) - \frac{1}{k-1} \sum_{\substack{q=1 \\ q\neq i}}^{k} \widehat{R}\left(\mathtt{A}(\mathcal{S}_n^{-\{\mathcal{F}_i,\mathcal{F}_q\}}), \mathcal{F}_q\right) \right)^2 \right\|_q \\
&= \sum_{i=1}^{k} \left\| \left(\frac{1}{k(k-1)} \sum_{j=1}^{k} \sum_{\substack{q=1 \\ q\neq i}}^{k} \left(\widehat{R}\left(\mathtt{A}(\mathcal{S}_n^{-\mathcal{F}_j}), \mathcal{F}_j\right) - \widehat{R}\left(\mathtt{A}(\mathcal{S}_n^{-\{\mathcal{F}_i,\mathcal{F}_q\}}), \mathcal{F}_q\right) \right) \right)^2 \right\|_q \\
&\leq \sum_{i=1}^{k} \left\| \frac{1}{k(k-1)} \sum_{j=1}^{k} \sum_{\substack{q=1 \\ q\neq i}}^{k} \left(\widehat{R}\left(\mathtt{A}(\mathcal{S}_n^{-\mathcal{F}_j}), \mathcal{F}_j\right) - \widehat{R}\left(\mathtt{A}(\mathcal{S}_n^{-\{\mathcal{F}_i,\mathcal{F}_q\}}), \mathcal{F}_q\right) \right)^2 \right\|_q \quad \text{(by Jensen inequality)} \\
&\leq \frac{1}{k(k-1)} \sum_{i=1}^{k} \sum_{j=1}^{k} \sum_{\substack{q=1 \\ q\neq i}}^{k} \left\| \left(\widehat{R}\left(\mathtt{A}(\mathcal{S}_n^{-\mathcal{F}_j}), \mathcal{F}_j\right) - \widehat{R}\left(\mathtt{A}(\mathcal{S}_n^{-\{\mathcal{F}_i,\mathcal{F}_q\}}), \mathcal{F}_q\right) \right)^2 \right\|_q \quad \text{(by triangle inequality)} \\
&= \frac{1}{k(k-1)} \sum_{i=1}^{k} \sum_{j=1}^{k} \sum_{\substack{q=1 \\ q\neq i}}^{k} \beta_{2q}^2(n-m,m) \\
&= k\beta_{2q}^2(n-m,m)\;.
\end{aligned} \tag{24}$$

Now observe that replacing q with $2q$ yields that

$$\|V_{\text{DEL}}\|_{2q} \leq k\beta_{4q}^2(n-m,m)\;, \tag{25}$$

which completes the proof. □

Proof of Lemma 5

Lemma 5. *Let Z and Z_{-i} be defined as in* (11) *and let $V_{DEL} = \sum_{i=1}^{n}(Z - Z_{-i})^2$. Then for any real $q \geq 1/2$, and $n \geq 2$, the following holds:*

$$\|V_{DEL}\|_{2q} \;\leq\; n\beta_{4q}^2(n,1)\;. \tag{12}$$

Proof. Let $q \geq 1$. Then,

$$\begin{aligned}
\|V_{\text{DEL}}\|_q &= \left\| \sum_{i=1}^{n} (Z - Z_{-i})^2 \right\|_q \\
&\leq \sum_{i=1}^{n} \left\| (Z - Z_{-i})^2 \right\|_q \quad \text{(by triangle inequality)} \\
&= \sum_{i=1}^{n} \|(Z - Z_{-i})\|_{2q}^2 \quad \text{(since } \left\|X^2\right\|_q = \|X\|_{2q}^2\text{)} \\
&= \sum_{i=1}^{n} \left\| R(\mathtt{A}(\mathcal{S}_n), \mathscr{P}) - R(\mathtt{A}(\mathcal{S}_n^{-i}), \mathscr{P}) \right\|_{2q}^2 \\
&= \sum_{i=1}^{n} \left\| \mathbb{E}\left[\ell\left(\mathtt{A}(\mathcal{S}_n), X\right) - \ell\left(\mathtt{A}(\mathcal{S}_n^{-i}), X\right) | \mathcal{S}_n \right] \right\|_{2q}^2 \\
&= \sum_{i=1}^{n} \left\| \mathbb{E}\left[\ell\left(\mathtt{A}(\mathcal{S}_n), X\right) - \ell\left(\mathtt{A}(\mathcal{S}_n^{-1}), X\right) | \mathcal{S}_n \right] \right\|_{2q}^2 \quad \text{(by } \textit{i.i.d} \text{ of the examples)} \\
&= n\beta_{2q}^2(n, 1) \,. \qquad (26)
\end{aligned}$$

Replacing q with $2q$ yields that

$$\|V_{\text{DEL}}\|_{2q} \;\leq\; n\beta_{4q}^2(n, 1) \,. \qquad (27)$$

□

Proof of Lemma 7

Lemma 7. *Using the previous setup and definitions, let* $\mathtt{A}$ *be a learning rule with* L_2 *stability coefficient* $\beta_2(n, m)$*. Then, for* $k \geq 1$*, and* $n > m \geq 1$*, the following holds*

$$\left| \mathbb{E}R(\mathtt{A}(\mathcal{S}_n), \mathscr{P}) - \mathbb{E}\widehat{R}_{CV}(\mathtt{A}, \mathcal{F}_{1,\ldots,k}) \right| \;\leq\; \beta_2(n, m) \,.$$

Proof. To derive a bound on $\left| \mathbb{E}R(\mathtt{A}(\mathcal{S}_n), \mathscr{P}) - \mathbb{E}\widehat{R}_{\text{CV}}(\mathtt{A}, \mathcal{F}_{1,\ldots,k}) \right|$ 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 $\mathbb{E}\widehat{R}_{\text{CV}}\left(\mathtt{A}, \mathcal{F}_{1,\ldots,k}\right)$, we have

$$\begin{aligned}
\mathbb{E}\widehat{R}_{\text{CV}}\left(\mathtt{A}, \mathcal{F}_{1,\ldots,k}\right) &= \mathbb{E}\left[\frac{1}{km} \sum_{j=1}^{k} \sum_{x_i \in \mathcal{F}_j} \ell\left(\mathtt{A}\left(\mathcal{S}_n^{-\mathcal{F}_j}, x_i\right)\right) \right] \\
&= \frac{1}{km} \sum_{j=1}^{k} \sum_{x_i \in \mathcal{F}_j} \mathbb{E}\left[\ell\left(\mathtt{A}\left(\mathcal{S}_n^{-\mathcal{F}_j}\right), x_i\right) \right] \\
&= \frac{1}{km} \sum_{j=1}^{k} \sum_{i=1}^{m} \mathbb{E}\left[\ell\left(\mathtt{A}\left(\mathcal{S}_n^{-[m]}\right), x_i'\right) \right] \quad \text{(by } \textit{i.i.d} \text{ of the examples)} \\
&= \mathbb{E}\left[\ell\left(\mathtt{A}\left(\mathcal{S}_n^{-[m]}\right), X\right) \right] \,,
\end{aligned}$$

where $(X_1', \ldots, X_m')$ are *i.i.d* examples drawn from $\mathcal{X}$ according to $\mathscr{P}$.

It follows that

$$\begin{aligned}
\left| \mathbb{E}R\left(\mathtt{A}(\mathcal{S}_n), \mathscr{P}\right) - \mathbb{E}\widehat{R}_{\text{CV}}\left(\mathtt{A}, \mathcal{F}_{1,\ldots,k}\right) \right| &= \left| \mathbb{E}[\ell\left(\mathtt{A}(\mathcal{S}_n), X\right)] - \mathbb{E}[\ell(\mathtt{A}(\mathcal{S}_n^{-[m]}), X)] \right| \\
&= \left| \mathbb{E}\left[\ell\left(\mathtt{A}(\mathcal{S}_n), X\right) - \ell(\mathtt{A}(\mathcal{S}_n^{-[m]}), X) \right] \right| \\
&\leq \mathbb{E}\left[\left| \ell\left(\mathtt{A}(\mathcal{S}_n), X\right) - \ell(\mathtt{A}(\mathcal{S}_n^{-[m]}), X) \right| \right] \\
&\leq \sqrt{\mathbb{E}\left[\left(\ell\left(\mathtt{A}(\mathcal{S}_n), X\right) - \ell(\mathtt{A}(\mathcal{S}_n^{-[m]}), X) \right)^2 \right]} \\
&= \beta_2(n, m). \quad \text{(by definition of } L_2 \text{ Stability)} \qquad (28)
\end{aligned}$$

□