JMLR: Workshop and Conference Proceedings vol 35:1–28, 2014

Localized Complexities for Transductive Learning

Ilya Tolstikhin ILIYA.TOLSTIKHIN@GMAIL.COM
Computing Centre of Russian Academy of Sciences, Russia

Gilles Blanchard GILLES.BLANCHARD@MATH.UNI-POTSDAM.DE
Department of Mathematics, University of Potsdam, Germany

Marius Kloft * MKLOFT@CS.NYU.EDU
Department of Computer Science, Humboldt University of Berlin, Germany

Abstract

We show two novel concentration inequalities for suprema of empirical processes when sampling *without* replacement, which both take the variance of the functions into account. While these inequalities may potentially have broad applications in learning theory in general, we exemplify their significance by studying the *transductive* setting of learning theory. For which we provide the first excess risk bounds based on the localized complexity of the hypothesis class, which can yield fast rates of convergence also in the transductive learning setting. We give a preliminary analysis of the localized complexities for the prominent case of kernel classes.

Keywords: Statistical Learning, Transductive Learning, Fast Rates, Localized Complexities, Concentration Inequalities, Empirical Processes, Kernel Classes

1. Introduction

The analysis of the stochastic behavior of empirical processes is a key ingredient in learning theory. The supremum of empirical processes is of particular interest, playing a central role in various application areas, including the theory of empirical processes, VC theory, and Rademacher complexity theory, to name only a few. Powerful Bennett-type concentration inequalities on the sup-norm of empirical processes introduced in Talagrand (1996) (see also Bousquet (2002a)) are at the heart of many recent advances in statistical learning theory, including *local Rademacher complexities* (Bartlett et al., 2005; Koltchinskii, 2011b) and related localization strategies (cf. Steinwart and Christmann, 2008, Chapter 4), which can yield *fast rates* of convergence on the excess risk.

These inequalities are based on the assumption of *independent and identically distributed* random variables, commonly assumed in the inductive setting of learning theory and thus implicitly underlying many prevalent machine learning algorithms such as support vector machines (Cortes and Vapnik, 1995; Steinwart and Christmann, 2008). However, in many cases the i.i.d. assumption breaks and substitutes for Talagrand's inequality are required. For instance, the i.i.d. assumption is violated when training and test data come from different distributions or data points exhibit (e.g., temporal) interdependencies (e.g., Steinwart et al., 2009). Both scenarios are typical situations in visual recognition, computational biology, and many other application areas.

* Most of the work was done while MK was with the Courant Institute of Mathematical Sciences and Memorial Sloan-Kettering Cancer Center, New York, NY, USA.

Another example where the i.i.d. assumption is void—in the focus of the present paper—is the *transductive* setting of learning theory, where training examples are sampled independent and *without* replacement from a finite population, instead of being sampled i.i.d. with replacement. The learner in this case is provided with both a labeled training set and an unlabeled test set, and the goal is to predict the label of the test points. This setting naturally appears in almost all popular application areas, including text mining, computational biology, recommender systems, visual recognition, and computer malware detection, as effectively constraints are imposed on the samples, since they are inherently realized within the global system of our world. As an example, consider image categorization, which is an important task in the application area of visual recognition. An object of study here could be the set of all images disseminated in the internet, only some of which are already reliably labeled (e.g., by manual inspection by a human), and the goal is to predict the unknown labels of the unlabeled images, in order to, e.g., make them accessible to search engines for content-based image retrieval.

From a theoretical view, however, the transductive learning setting is yet not fully understood. Several transductive error bounds were presented in series of works (Vapnik, 1982, 1998; Blum and Langford, 2003; Derbeko et al., 2004; Cortes and Mohri, 2006; El-Yaniv and Pechyony, 2009; Cortes et al., 2009), including the first analysis based on *global Rademacher complexities* presented in El-Yaniv and Pechyony (2009). However, the theoretical analysis of the performance of transductive learning algorithms still remains less illuminated than in the classic inductive setting: to the best of our knowledge, existing results do not provide fast rates of convergence in the general transductive setting.[1]

In this paper, we consider the transductive learning setting with arbitrary bounded nonnegative loss functions. The main result is an excess risk bound for transductive learning based on the localized complexity of the hypothesis class. This bound holds under general assumptions on the loss function and hypothesis class and can be viewed as a transductive analogue of Corollary 5.3 in Bartlett et al. (2005). The bound is very generally applicable with loss functions such as the squared loss and common hypothesis classes. By exemplarily applying our bound to kernel classes, we achieve, for the first time in the transductive setup, an excess risk bound in terms of the tailsum of the eigenvalues of the kernel, similar to the best known results in the inductive setting. In addition, we also provide new transductive generalization error bounds that take the variances of the functions into account, and thus can yield sharper estimates.

The localized excess risk bound is achieved by proving two novel *concentration inequalities* for suprema of empirical processes when sampling without replacement. The application of which goes far beyond the transductive learning setting—these concentration inequalities could serve as a fundamental mathematical tool in proving results in various other areas of machine learning and learning theory. For instance, arguably the most prominent example in machine learning and learning theory of an empirical process where sampling without replacement is employed is cross-validation (Stone, 1974), where training and test folds are sampled without replacement from the overall pool of examples, and the new inequalities could help gaining a non-asymptotic understanding of cross-validation procedures. However, the investigation of further applications of the novel concentration inequalities beyond the transductive learning setting is outside of the scope of the present paper.

1. An exception are the works of Blum and Langford (2003); Cortes and Mohri (2006), which consider, however, the case where the Bayes hypothesis has zero error and is contained in the hypothesis class. This is clearly an assumption too restrictive in practice, where the Bayes hypothesis usually cannot be assumed to be contained in the class.

2. The Transductive Learning Setting and State of the Art

From a statistical point of view, the main difference between the transductive and inductive learning settings lies in the protocols used to obtain the training sample S. Inductive learning assumes that the training sample is drawn i.i.d. from some fixed and unknown distribution P on the product space $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the input space and $\mathcal{Y}$ is the output space. The learning algorithm chooses a predictor h from some fixed hypothesis set $\mathcal{H}$ based on the training sample, and the goal is to minimize the true risk $\mathbb{E}_{X\times Y}[\ell\big(h(X), Y\big)] \to \min_{h\in\mathcal{H}}$ for a fixed, bounded, and nonnegative loss function $\ell\colon \mathcal{Y}^2 \to [0, 1]$.

We will use one of the two transductive settings[2] considered in Vapnik (1998), which is also used in Derbeko et al. (2004); El-Yaniv and Pechyony (2009). Assume that a set $\boldsymbol{X}_N$ consisting of N arbitrary input points is given (without any assumptions regarding its underlying source). We then sample $m \leq N$ objects $\boldsymbol{X}_m \subseteq \boldsymbol{X}_N$ uniformly without replacement from $\boldsymbol{X}_N$ (which makes the inputs in $\boldsymbol{X}_m$ dependent). Finally, for the input examples $\boldsymbol{X}_m$ we obtain their outputs $\boldsymbol{Y}_m$ by sampling, for each input $X \in \boldsymbol{X}_m$, the corresponding output Y from some unknown distribution $P(Y|X)$. The resulting *training set* is denoted by $S_m = (\boldsymbol{X}_m, \boldsymbol{Y}_m)$. The remaining unlabeled set $\boldsymbol{X}_u = \boldsymbol{X}_N \setminus \boldsymbol{X}_m$, $u = N - m$ is the *test set*. Note that both Derbeko et al. (2004) and El-Yaniv and Pechyony (2009) consider a special case where the labels are obtained using some unknown but deterministic target function $\phi\colon \mathcal{X} \to \mathcal{Y}$ so that $P\big(\phi(X)|X\big) = 1$. We will adopt the same assumption here. The learner then chooses a predictor h from some fixed hypothesis set $\mathcal{H}$ (not necessarily containing ϕ) based on both the labeled training set S_m and unlabeled test set $\boldsymbol{X}_u$. For convenience let us denote $\ell_h(X) = \ell\big(h(X), \phi(X)\big)$. We define the test and training error, respectively, of hypothesis h as follows: $L_u(h) = \frac{1}{u}\sum_{X\in\boldsymbol{X}_u} \ell_h(X)$, $\hat{L}_m(h) = \frac{1}{m}\sum_{X\in\boldsymbol{X}_m} \ell_h(X)$, where hat emphasizes the fact that the training (empirical) error can be computed from the data. For technical reasons that will become clear later, we also define the overall error of an hypothesis h with regard to the union of the training and test sets as $L_N(h) = \frac{1}{N}\sum_{X\in\boldsymbol{X}_N} \ell_h(X)$ (this quantity will play a crucial role in the upcoming proofs). Note that for a fixed hypothesis h the quantity $L_N(h)$ is not random, as it is invariant under the partition into training and test sets. The main goal of the learner in transductive setting is to select a hypothesis minimizing the test error $L_u(h) \to \inf_{h\in\mathcal{H}}$, which we will denote by h^*_u.

Since the labels of the test set examples are unknown, we cannot compute $L_u(h)$ and need to estimate it based on the training sample S_m. A common choice is to replace the test error minimization by *empirical risk minimization* $\hat{L}_m(h) \to \min_{h\in\mathcal{H}}$ and to use its solution, which we denote by $\hat{h}_m$, as an approximation of h^*_u. For $h \in \mathcal{H}$ let us define an *excess risk* of h:

$$\mathcal{E}_u(h) = L_u(h) - \inf_{g\in\mathcal{H}} L_u(g) = L_u(h) - L_u(h^*_u).$$

A natural question is: how well does the hypothesis $\hat{h}_m$ chosen by the ERM algorithm approximate the theoretical-optimal hypothesis h^*_u?

To this end, we use $\mathcal{E}_u(\hat{h}_m)$ as a measure of the goodness of fit. Obtaining tight upper bounds on $\mathcal{E}_u(\hat{h}_m)$—so-called *excess risk bounds*—is thus the main goal of this paper. Another goal commonly considered in learning literature is the one of obtaining upper bounds on $L_u(\hat{h}_m)$ in terms

2. The second setting assumes that both training and test sets are sampled i. i. d. from the same unknown distribution and the learner is provided with the labeled training and unlabeled test sets. It is pointed out by Vapnik (1998) that any upper bound on $L_u(h) - \hat{L}_m(h)$ in the setting we consider directly implies a bound also for the second setting.

of $\hat{L}_m(\hat{h}_m)$, which measures the generalization performance of empirical risk minimization. Such bounds are known as the *generalization error bounds*. Note that both $\hat{h}_m$ and h_u^* are random, since they depend on the training and test sets, respectively. Note, moreover, that for any fixed $h \in \mathcal{H}$ its excess risk $\mathcal{E}_u(h)$ is also random. Thus both tasks (of obtaining excess risk and generalization bounds, respectively) deal with random quantities and require bounds that hold with high probability.

The most common way to obtain generalization error bounds for $\hat{h}_m$ is to introduce uniform deviations over the class $\mathcal{H}$:

$$L_u(\hat{h}_m) - \hat{L}_m(\hat{h}_m) \leq \sup_{h \in \mathcal{H}} L_u(h) - \hat{L}_m(h). \tag{1}$$

The random variable appearing on the right side is directly related to the sup-norm of the empirical process (Boucheron et al., 2013). It should be clear that, in order to analyze the transductive setting, it is of fundamental importance to obtain high-probability bounds for functions $f(Z_1, \ldots, Z_m)$, where $\{Z_1, \ldots, Z_m\}$ are random variables sampled *without* replacement from some fixed finite set. Of particular interest are concentration inequalities for sup-norms of empirical processes, which we present in Section 3.

2.1. State of the Art and Related Work

Error bounds for transductive learning were considered by several authors in recent years. Here we name only a few of them[3]. The first general bound for binary loss functions, presented in Vapnik (1982), was *implicit* in the sense that the value of the bound was specified as an outcome of a computational procedure. The somewhat refined version of this implicit bound also appears in Blum and Langford (2003). It is well known that generalization error bounds with fast rates of convergence can be obtained under certain restrictive assumptions on the problem at hand. For instance, Blum and Langford (2003) provide a bound that has an order of $\frac{1}{\min(u,m)}$ in the *realizable* case, i.e., when $\phi \in \mathcal{H}$ (meaning that the hypothesis having zero error belongs to $\mathcal{H}$). However, such an assumption is usually unrealistic: in practice it is usually impossible to avoid overfitting when choosing $\mathcal{H}$ so large that it contains the Bayes classifier.

The authors of Cortes and Mohri (2006) consider a transductive regression problem with bounded squared loss and obtain a generalization error bound of the order $\sqrt{\hat{L}_m(\hat{h}_m)\frac{\log N}{\min(m,u)}}$, which also does not attain a fast rate. Several PAC-Bayesian bounds were presented in Blum and Langford (2003); Derbeko et al. (2004) and others. However their tightness critically depends on the *prior distribution* over the hypothesis class, which should be fixed by the learner prior to observing the training sample. Transductive bounds based on algorithmic stability were presented for classification in El-Yaniv and Pechyony (2006) and for regression in Cortes et al. (2009). However both are roughly of the order $\min(u,m)^{-1/2}$. Finally, we mention the results of El-Yaniv and Pechyony (2009) based on transductive Rademacher complexities. However, the analysis was based on the *global* Rademacher complexity combined with a McDiarmid-style concentration inequality for sampling without replacement and thus does not yield fast convergence rates.

3. For an extensive review of transductive error bounds we refer to Pechyony (2008).

3. Novel Concentration Inequalities for Sampling Without replacement

In this section, we present two new *concentration inequalities* for suprema of empirical processes when sampling without replacement. The first one is a sub-Gaussian inequality that is based on a result by Bobkov (2004) and closely related to the *entropy method* (Boucheron et al., 2013). The second inequality is an analogue of Bousquet's version of Talagrand's concentration inequality (Bousquet, 2002b,a; Talagrand, 1996) and is based on the reduction method first suggested in Hoeffding (1963).

Next we state the setting and introduce the necessary notation. Let $\mathcal{C} = \{c_1, \dots, c_N\}$ be some finite set. For $m \leq N$ let $\{Z_1, \dots, Z_m\}$ and $\{X_1, \dots, X_m\}$ be sequences of random variables sampled uniformly without and with replacement from $\mathcal{C}$, respectively. Let $\mathcal{F}$ be a (countable[4]) class of functions $f : \mathcal{C} \to \mathbb{R}$, such that $\mathbb{E}[f(Z_1)] = 0$ and $f(x) \in [-1, 1]$ for all $f \in \mathcal{F}$ and $x \in \mathcal{C}$. It follows that $\mathbb{E}[f(X_1)] = 0$ since Z_1 and X_1 are identically distributed. Define the variance $\sigma^2 = \sup_{f\in\mathcal{F}} \mathbb{V}[f(Z_1)]$. Note that $\sigma^2 = \sup_{f\in\mathcal{F}} \mathbb{E}[f^2(Z_1)] = \sup_{f\in\mathcal{F}} \mathbb{V}[f(X_1)]$. Let us define the *supremum of the empirical process* for sampling with and without replacement, respectively:[5]

$$Q_m = \sup_{f\in\mathcal{F}} \sum_{i=1}^{m} f(X_i), \quad Q'_m = \sup_{f\in\mathcal{F}} \sum_{i=1}^{m} f(Z_i).$$

The random variable Q_m is well studied in the literature and there are remarkable Bennett-type concentration inequalities for Q_m, including Talagrand's inequality (Talagrand, 1996) and its versions due to Bousquet (2002a,b) and others.[6] The random variable Q'_m, on the other hand, is much less understood, and no Bennett-style concentration inequalities are known for it up to date.

3.1. The New Concentration Inequalities

In this section, we address the lack of Bennett-type concentration inequalities for Q'_m by presenting two novel inequalities for suprema of empirical processes when sampling without replacement.

Theorem 1 (Sub-Gaussian concentration inequality for sampling *without* replacement) *For any* $\epsilon \geq 0$,

$$\mathbb{P}\left\{Q'_m - \mathbb{E}[Q'_m] \geq \epsilon\right\} \leq \exp\left\{-\frac{(N+2)\epsilon^2}{8N^2\sigma^2}\right\} < \exp\left\{-\frac{\epsilon^2}{8N\sigma^2}\right\}. \tag{2}$$

The same bound also holds for $\mathbb{P}\left\{\mathbb{E}[Q'_m] - Q'_m \geq \epsilon\right\}$. *Also for all* $t \geq 0$ *the following holds with probability greater than* $1 - e^{-t}$*:*

$$Q'_m \leq \mathbb{E}[Q'_m] + 2\sqrt{2N\sigma^2 t}. \tag{3}$$

Theorem 2 (Talagrand-type concentration inequality for sampling *without* replacement) *Define* $v = m\sigma^2 + 2\mathbb{E}[Q_m]$. *For* $u > -1$ *define* $\phi(u) = e^u - u - 1$, $h(u) = (1+u)\log(1+u) - u$. *Then for*

4. Note that all results can be translated to the uncountable classes, for instance, if the empirical process is *separable*, meaning that $\mathcal{F}$ contains a dense countable subset. We refer to page 314 of Boucheron et al. (2013) or page 72 of Bousquet (2002b).
5. The results presented in this section can be also generalized to $\sup_{f\in\mathcal{F}} \left|\sum_{i=1}^{m} f(Z_i)\right|$ using the same techniques.
6. For completeness we present one such inequality for Q_m as Theorem 17 in Appendix A. For the detailed review of concentration inequalities for Q_m we refer to Section 12 of Boucheron et al. (2013).

any $\epsilon \geq 0$*:*

$$\mathbb{P}\left\{Q'_m - \mathbb{E}[Q_m] \geq \epsilon\right\} \leq \exp\left(-vh\left(\frac{\epsilon}{v}\right)\right) \leq \exp\left(-\frac{\epsilon^2}{2v + \frac{2}{3}\epsilon}\right).$$

Also for any $t \geq 0$ *following holds with probability greater than* $1 - e^{-t}$*:*

$$Q'_m \leq \mathbb{E}[Q_m] + \sqrt{2vt} + \frac{t}{3}.$$

The appearance of $\mathbb{E}[Q_m]$ in the last theorem might seem unexpected on the first view. Indeed, one usually wants to control the concentration of a random variable around its expectation. However, it is shown in the lemma below that in many cases $\mathbb{E}[Q_m]$ will be close to $\mathbb{E}[Q'_m]$:

Lemma 3

$$0 \leq \mathbb{E}[Q_m] - \mathbb{E}[Q'_m] \leq 2\frac{m^3}{N}.$$

The above lemma is proved in Appendix B. It shows that for $m = o(N^{2/5})$ the order of $\mathbb{E}[Q_m] - \mathbb{E}[Q'_m]$ does not exceed $\sqrt{m}$, and thus Theorem 2 can be used to control the deviations of Q'_m above its expectation $\mathbb{E}[Q'_m]$ at a fast rate. However, generally $\mathbb{E}[Q'_m]$ could be smaller than $\mathbb{E}[Q_m]$, which may potentially lead to significant gap, in which case Theorem 1 is the preferred choice to control the deviations of Q'_m around $\mathbb{E}[Q'_m]$.

3.2. Discussion

It is worth comparing the two novel inequalities for Q'_m to the best known results in the literature. To this end, we compare our inequalities with the McDiarmid-style inequality recently obtained in El-Yaniv and Pechyony (2009) (and slightly improved in Cortes et al. (2009)):

Theorem 4 (El-Yaniv and Pechyony (2009)[7]) *For all* $\epsilon \geq 0$*:*

$$\mathbb{P}\left\{Q'_m - \mathbb{E}[Q'_m] \geq \epsilon\right\} \leq \exp\left\{-\frac{\epsilon^2}{2m}\left(\frac{N - 1/2}{N - m}\right)\left(1 - \frac{1}{2\max(m, N-m)}\right)\right\}. \tag{4}$$

The same bound also holds for $\mathbb{P}\left\{\mathbb{E}[Q'_m] - Q'_m \geq \epsilon\right\}$.

To begin with, let us notice that Theorem 4 does not account for the variance $\sup_{f\in\mathcal{F}} \mathbb{V}[f(X_1)]$, while Theorems 1 and Theorem 2 do. As it will turn out in Section 4, this refined treatment of the variance σ^2 allows us to use localization techniques, facilitating to obtain sharp estimates (and potentially, fast rates) also in the transductive learning setup. The comparison between concentration inequalities (2) of Theorem 2 and (4) of Theorem 4 is as follows: note that the term $1 - \frac{1}{2\max(m,N-m)}$ is negligible for large N, so that slightly re-writing the inequalities boils down to comparing $-\frac{\epsilon^2}{8m\sigma^2}\frac{m}{N}$ and $-\frac{\epsilon^2}{2m}\left(\frac{N-1/2}{N-m}\right)$. For $m = o(N)$ (which in a way transforms sampling without replacement to sampling with replacement), the second inequality clearly outperforms the first one. However, for the case when $m = \Omega(N)$ (frequently used in the transductive setting),

7. This bound does not appear explicitly in El-Yaniv and Pechyony (2009); Cortes et al. (2009), but can be immediately obtained using Lemma 2 of El-Yaniv and Pechyony (2009) for Q'_m with $\beta = 2$.

say $N = 2m$, the comparison depends on the relation between $-\epsilon^2/(16\, m\sigma^2)$ and $-\epsilon^2/m$ and the result of Theorem 1 outperforms the one of El-Yaniv and Pechyony for $\sigma^2 \leq 1/16$. The comparison between Theorems 2 and 4 for both cases ($m = o(N)$ and $m = \Omega(N)$) depends on the value of σ^2.

Theorem 2 is a direct analogue of Bousquet's version of Talagrand's inequality (see Theorem 17 in Appendix A of the supplementary material), frequently used in the learning literature. It states that the upper bound on Q_m, provided by Bousquet's inequality, also holds for Q'_m. Now we compare Theorems 1 and 2. First of all note that the deviation bound (3) does not have the term $2\mathbb{E}[Q_m] \geq 0$ under the square root in contrast to Theorem 2. As will be shown later, in some cases this fact can result in improved constants when applying Theorem 1. Another nice thing about Theorem 1 is that it provides upper bounds for both $Q'_m - \mathbb{E}[Q'_m]$ and $\mathbb{E}[Q'_m] - Q'_m$, while Theorem 2 upper bounds only $Q'_m - \mathbb{E}[Q'_m]$. The main drawback of Theorem 1 is the factor N appearing in the exponent. Later we will see that in some cases it is more preferable to use Theorem 2 because of this fact.

We also note that, if $m = \Omega(N)$ or $m = o(N^{2/5})$, we can control the deviations of Q'_m around $\mathbb{E}[Q'_m]$ with inequalities that are similar to i.i.d. case. It is an open question, however, whether this can be done also for other regimes of m and N. It should be clear though that we can obtain at least as good rates as in the inductive setting using Theorem 2. To summarize the discussion, when N is large and $m = o(N)$, Theorems 2 or 4 (depending on σ^2 and the order of $\mathbb{E}[Q_m] - \mathbb{E}[Q'_m]$) can be significantly tighter than Theorem 1. However, if $m = \Omega(N)$, Theorem 1 is more preferable. Further discussions are presented in Appendix C.

3.3. Proof Sketch

Here we briefly outline the proofs of Theorems 1 and 2. Detailed proofs are given in Appendix B of the supplementary material.

Theorem 2 is a direct consequence of Bousquet's inequality and Hoefding's reduction method. It was shown in Theorem 4 of Hoeffding (1963) that, for any convex function f, the following inequality holds:

$$\mathbb{E}\left[f\left(\sum_{i=1}^{m} Z_i\right)\right] \leq \mathbb{E}\left[f\left(\sum_{i=1}^{m} X_i\right)\right].$$

Although not stated explicitly in Hoeffding (1963), the same result also holds if we sample from finite set of vectors instead of real numbers (Gross and Nesme, 2010). This reduction to the i.i.d. setting together with some minor technical results is enough to bound the moment generating function of Q'_m and obtain a concentration inequality using Chernoff's method (for which we refer to the Section 2.2 of Boucheron et al. (2013)).

The proof of Theorem 1 is more involved. It is based on the sub-Gaussian inequality presented in Theorem 2.1 of Bobkov (2004), which is related to the *entropy method* introduced by M. Ledoux (see Boucheron et al. (2013) for references). Consider a function g defined on the partitions $X^N = (X^m \cup X^u)$ of a fixed finite set X^N of cardinality N into two disjoint subsets X^m and X^u of cardinalities m and u, respectively, where $N = m + u$. Bobkov's inequality states that, roughly speaking, if g is such that the Euclidean length of its discrete gradient $|\nabla g(X^m \cup X^u)|^2$ is bounded by a constant Σ^2, and if the partitions $(X^m \cup X^u)$ are sampled uniformly from the set of all such partitions, then $g(X^m \cup X^u)$ is sub-Gaussian with parameter Σ^2.

3.4. Applications of the New Inequalities

The novel concentration inequalities presented above can be generally used as a mathematical tool in various areas of machine learning and learning theory where suprema of empirical processes over sampling without replacement are of interest, including the analysis of cross-validation and low-rank matrix factorization procedures as well as the transductive learning setting. Exemplifying their applications, we show in the next section—for the first time in the transductive setting of learning theory—excess risk bounds in terms of localized complexity measures, which can yield sharper estimates than global complexities.

4. Excess Risk Bounds for Transductive Learning via Localized Complexities

We start with some preliminary generalization error bounds that show how to apply the concentration inequalities of Section 3 to obtain risk bounds in the transductive learning setting. Note that (1) can be written in the following way:

$$L_u(\hat{h}_m) - \hat{L}_m(\hat{h}_m) \leq \sup_{h\in\mathcal{H}} L_u(h) - \hat{L}_m(h) = \frac{N}{u} \cdot \sup_{h\in\mathcal{H}} L_N(h) - \hat{L}_m(h),$$

where we used the fact that $N \cdot L_N(h) = m \cdot \hat{L}_m(h) + u \cdot L_u(h)$. Note that for any fixed $h \in \mathcal{H}$, we have $L_N(h) - \hat{L}_m(h) = \frac{1}{m}\sum_{X\in\boldsymbol{X}_m}\big(L_N(h) - \ell_h(X)\big)$, where $\boldsymbol{X}_m$ is sampled uniformly without replacement from $\boldsymbol{X}_N$. Note that we clearly have $L_N(h) - \ell_h(X) \in [-1, 1]$ and $\mathbb{E}[L_N(h) - \ell_h(X)] = L_N(h) - \mathbb{E}[\ell_h(X)] = 0$. Thus we can use the setting described in Section 3, with $\boldsymbol{X}_N$ playing the role of $\mathcal{C}$ and considering the function class $\mathcal{F}_\mathcal{H} = \{f_h \colon f_h(X) = L_N(h) - \ell_h(X), h \in \mathcal{H}\}$ associated with $\mathcal{H}$, to obtain high-probability bounds for $\sup_{f_h\in\mathcal{F}_\mathcal{H}}\sum_{X\in\boldsymbol{X}_m} f_h(X) = m \cdot \sup_{h\in\mathcal{H}}\big(L_N(h) - \hat{L}_m(h)\big)$. Note that in Section 3 we considered unnormalized sums, hence we obtain a factor of m in the above equation. As already noted, for fixed h, $L_N(h)$ is not random; also centering random variable does not change its variance. Keeping this in mind, we define

$$\sigma^2_\mathcal{H} = \sup_{f_h\in\mathcal{F}_\mathcal{H}} \mathbb{V}[f_h(X)] = \sup_{h\in\mathcal{H}} \mathbb{V}[\ell_h(X)] = \sup_{h\in\mathcal{H}} \left(\frac{1}{N}\sum_{X\in\boldsymbol{X}_N}\big(\ell_h(X) - L_N(h)\big)^2\right). \tag{5}$$

Using Theorems 1 and 2, we can obtain the following results that hold *without any assumptions* on the learning problem at hand, except for the boundedness of the loss function in the interval $[0, 1]$. Our first result of this section follows immediately from the new concentration inequality of Theorem 1:

Theorem 5 *For any $t \geq 0$ with probability greater than $1 - e^{-t}$ the following holds:*

$$\forall h \in \mathcal{H}: \quad L_N(h) - \hat{L}_m(h) \leq \mathbb{E}\left[\sup_{h\in\mathcal{H}}\big(L_N(h) - \hat{L}_m(h)\big)\right] + 2\sqrt{2\left(\frac{N}{m^2}\right)\sigma^2_\mathcal{H} t},$$

where $\sigma^2_\mathcal{H}$ was defined in (5).

Let $\{\xi_1, \ldots, \xi_m\}$ be random variables sampled *with replacement* from $\boldsymbol{X}_N$ and denote

$$E_m = \mathbb{E}\left[\sup_{h\in\mathcal{H}}\left(L_N(h) - \frac{1}{m}\sum_{i=1}^m \ell_h(\xi_i)\right)\right].$$

The following result follows from Theorem 2 by simple calculus. We provide the detailed proof in the supplementary material.

Theorem 6 *For any $t \geq 0$ with probability greater than $1 - e^{-t}$ the following holds:*

$$\forall h \in \mathcal{H}: \quad L_N(h) - \hat{L}_m(h) \leq 2E_m + \sqrt{\frac{2\sigma_{\mathcal{H}}^2 t}{m}} + \frac{4t}{3m},$$

where $\sigma_{\mathcal{H}}^2$ was defined in (5).

Remark 7 *Note that E_m is an expected sup-norm of the empirical process naturally appearing in inductive learning. Using the well-known* symmetrization inequality *(see Section 11.3 of Boucheron et al. (2013)), we can upper bound it by twice the expected value of the supremum of the Rademacher process. In this case, the last theorem thus gives exactly the same upper bound on the quantity $\sup_{h\in\mathcal{H}}\big(L_N(h) - \hat{L}_m(h)\big)$ as the one of Theorem 2.1 of Bartlett et al. (2005) (with $\alpha = 1$ and $(b-a) = 1$).*

Here we provide some discussion on the two generalization error bounds presented above. Note that $\sigma_{\mathcal{H}}^2 \leq 1/4$, since $\sigma_{\mathcal{H}}^2$ is the variance of a random variable bounded in the interval $[0, 1]$. We conclude that the bound of Theorem 6 is of the order $m^{-1/2}$, since the typical order[8] of E_m is also $m^{-1/2}$. Note that repeating the proof of Lemma 3 we immediately obtain the following corollary:

Corollary 8 *Let $\{\xi_1, \ldots, \xi_m\}$ be random variables sampled* with replacement *from $\boldsymbol{X}_N$. For any countable set of functions $\mathcal{F}$ defined on $\boldsymbol{X}_N$ the following holds:*

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \mathbb{E}[f(X)] - \frac{1}{m}\sum_{X\in\boldsymbol{X}_m} f(X)\right] \leq \mathbb{E}\left[\sup_{f\in\mathcal{F}} \mathbb{E}[f(X)] - \frac{1}{m}\sum_{i=1}^{m} f(\xi_i)\right].$$

The corollary shows that for $m = \Omega(N)$ the bound of Theorem 5 also has the order $m^{-1/2}$. However, if $m = o(N)$, the convergence becomes slower and it can even diverge for $m = o(N^{1/2})$.

Remark 9 *The last corollary enables us to use also in the transductive setting all the established techniques related to the inductive Rademacher process, including symmetrization and contraction inequalities. Later in this section, we will employ this result to derive excess risk bounds for kernel classes in terms of the tailsum of the eigenvalues of the kernel, which can yield a fast rate of convergence. However, we should keep in mind that there might be a significant gap between $\mathbb{E}[Q_m]$ and $\mathbb{E}[Q'_m]$, in which case such a reduction can be loose.*

4.1. Excess Risk Bounds

The main goal of this section is to analyze to what extent the known results on localized risk bounds presented in series of works (Koltchinskii and Panchenko, 1999; Massart, 2000; Bartlett et al., 2005; Koltchinskii, 2006) can be generalized to the transductive learning setting. These results essentially show that the rate of convergence of the excess risk is related to the fixed point of the modulus of continuity of the empirical process associated with the hypothesis class. Our main tools to this end

8. For instance if $\mathcal{F}$ is finite it follows from Theorems 2.1 and 3.5 of Koltchinskii (2011a).

will be the sub-Gaussian and Bennett-style concentration inequalities of Theorems 1 and 2 presented in the previous section.

From now on it will be convenient to introduce the following operators, mapping functions f defined on $\boldsymbol{X}_N$ to $\mathbb{R}$:

$$Ef = \frac{1}{N} \sum_{X \in \boldsymbol{X}_N} f(X), \quad \hat{E}_m f = \frac{1}{m} \sum_{X \in \boldsymbol{X}_m} f(X).$$

Using this notation we have: $L_N(h) = E\ell_h$ and $\hat{L}_m(h) = \hat{E}_m \ell_h$.

Define the *excess loss class* $\mathcal{F}^* = \{\ell_h - \ell_{h_N^*}, h \in \mathcal{H}\}$. Throughout this section, we will assume that the loss function ℓ and hypothesis class $\mathcal{H}$ satisfy the following assumptions:

Assumptions 10

1. *There is a function $h_N^* \in \mathcal{H}$ satisfying $L_N(h_N^*) = \inf_{h \in \mathcal{H}} L_N(h)$.*
2. *There is a constant $B > 0$ such that for every $f \in \mathcal{F}^*$ we have $Ef^2 \leq B \cdot Ef$.*
3. *As before the loss function ℓ is bounded in the interval $[0, 1]$.*

Here we shortly discuss these assumptions. Assumption 10.1 is quite common and not restrictive. Assumption 10.2 can be satisfied, for instance, when the loss function ℓ is Lipschitz and there is a constant $T > 1$ such that $\frac{1}{N} \sum_{X \in \boldsymbol{X}_N} \big(h(X) - h_N^*(X)\big)^2 \leq T\big(L_N(h) - L_N(h_N^*)\big)$ for all $h \in \mathcal{H}$. These conditions are satisfied for example for the quadratic loss $\ell(y, y') = (y - y')^2$ with uniformly bounded convex classes $\mathcal{H}$ (for other examples we refer to Section 5.2 of Bartlett et al. (2005) and Section 2.1 of Bartlett et al. (2010)). Assumption 10.3 could be possibly relaxed using some analogues of Theorems 1 and 2 that hold for classes $\mathcal{F}$ with unbounded functions[9].

Next we present the main results of this section, which can be considered as an analogues of Corollary 5.3 of Bartlett et al. (2005). The results come in pairs, depending on whether Theorem 1 or 2 is used in the proof. We will need the notion of a *sub-root* function, which is a nondecreasing and nonnegative function $\psi \colon [0, \infty) \to [0, \infty)$, such that $r \to \psi(r)/\sqrt{r}$ is nonincreasing for $r > 0$. It can be shown that any sub-root function has a unique and positive fixed point.

Theorem 11 *Let $\mathcal{H}$ and ℓ be such that Assumptions 10 are satisfied. Assume there is a sub-root function $\psi_m(r)$ such that*

$$B\,\mathbb{E}\left[\sup_{f \in \mathcal{F}^* : Ef^2 \leq r} (Ef - \hat{E}_m f)\right] \leq \psi_m(r). \tag{6}$$

Let r_m^* be a fixed point of $\psi_m(r)$. Then for any $t > 0$ with probability greater than $1 - e^{-t}$ we have:

$$L_N(\hat{h}_m) - L_N(h_N^*) \leq 51 \frac{r_m^*}{B} + 17Bt\left(\frac{N}{m^2}\right).$$

We emphasize that the constants appearing in Theorem 11 are slightly better than the ones appearing in Corollary 5.3 of Bartlett et al. (2005). This result is based on Theorem 1 and thus shares the disadvantages of Theorem 5 discussed above: the bound does not converge for $m = o(N^{-1/2})$. However, by using Theorem 2 instead of Theorem 1 in the proof, we can replace the factor of N/m^2 appearing in the bound by a factor of $1/m$ at a price of slightly worse constants:

9. Adamczak (2008) show a version of Talagrand's inequality for unbounded functions in the i.i.d. case.

Theorem 12 *Let $\mathcal{H}$ and ℓ be such that Assumptions 10 are satisfied. Let $\{\xi_1, \ldots, \xi_m\}$ be random variables sampled* with replacement *from $\boldsymbol{X}_N$. Assume there is a sub-root function $\psi_m(r)$ such that*

$$B \cdot \mathbb{E}\left[\sup_{f \in \mathcal{F}^* : Ef^2 \leq r} \left(Ef - \frac{1}{m}\sum_{i=1}^{m} f(\xi_i)\right)\right] \leq \psi_m(r). \tag{7}$$

Let r_m^* be a fixed point of $\psi_m(r)$. Then for any $t > 0$, with probability greater than $1 - e^{-t}$, we have:

$$L_N(\hat{h}_m) - L_N(h_N^*) \leq 901\frac{r_m^*}{B} + \frac{t(16 + 25B)}{3m}.$$

We also note that in Theorem 12 the modulus of continuity of the empirical process over sampling without replacement appearing in the left-hand side of (6) is replaced with its inductive analogue. As follows from Corollary 8, the fixed point r_m^* of Theorem 11 can be smaller than that of Theorem 12 and thus, for large N and $m = \Omega(N)$ the first bound can be tighter. Otherwise, if $m = o(N)$, Theorem 12 can be more preferable.

Proof sketch: Now we briefly outline the proof of Theorem 11. It is based on the *peeling technique* and consists of the steps described below (similar to the proof of the first part of Theorem 3.3 in Bartlett et al. (2005)). The proof of Theorem 12 repeats the same steps with the only difference being that Theorem 2 is used on Step 2 instead of Theorem 1. The detailed proofs are presented in Section D of the supplementary material.

STEP 1 First we fix an arbitrary $r > 0$ and introduce the rescaled version of the centered loss class: $\mathcal{G}_r = \left\{\frac{r}{\Delta(r,f)} f,\ f \in \mathcal{F}^*\right\}$, where $\Delta(r, f)$ is chosen such that the variances of the functions contained in $\mathcal{G}_r$ do not exceed r.

STEP 2 We can use Theorem 1 to obtain the following upper bound on $V_r = \sup_{g \in \mathcal{G}_r}(Eg - \hat{E}_m g)$ which holds with probability greater than $1 - e^{-t}$: $V_r \leq \mathbb{E}[V_r] + 2\sqrt{2t\left(\frac{N}{m^2}\right) r}$.

STEP 3 Using the *peeling technique* (which consists in dividing the class $\mathcal{F}^*$ into slices of functions having variances within a certain range), we are able to show that $\mathbb{E}[V_r] \leq 5\psi_m(r)/B$. Also, using the definition of sub-root functions, we conclude that $\psi_m(r) \leq \sqrt{r r_m^*}$ for any $r \geq r_m^*$, which gives us $V_r \leq \sqrt{r}\left(5\frac{\sqrt{r_m^*}}{B} + 2\sqrt{2t\left(\frac{N}{m^2}\right)}\right)$.

STEP 4 Now we can show that by properly choosing $r_0 > r_m^*$ we can get that, for any $K > 1$, it holds $V_{r_0} \leq \frac{r_0}{KB}$. Using the definition of V_r, we obtain that the following holds, with probability greater than $1 - e^{-t}$:

$$\forall f \in \mathcal{F}^*, \forall K > 1 : \quad Ef - \hat{E}_m f \leq \frac{\max\left(r_0, Ef^2\right)}{r_0}\frac{r_0}{KB} = \frac{\max\left(r_0, Ef^2\right)}{KB}.$$

STEP 5 Finally it remains to upper bound Ef for the two cases $Ef^2 > r_0$ and $Ef^2 \leq r_0$ (which can be done using Assumption 10.2), to combine those two results, and to notice that $\hat{E}_m f \leq 0$ for $f(X) = \ell_{\hat{h}_m}(X) - \ell_{h_N^*}(X)$. ■

We finally present excess risk bounds for $\mathcal{E}_u(\hat{h}_m)$. The first one is based on Theorem 11:

Corollary 13 *Under the assumptions of Theorem 11, for any $t > 0$, with probability greater than $1 - 2e^{-t}$, we have:*

$$\mathcal{E}_u(\hat{h}_m) \leq \frac{N}{u}\left(51\frac{r_m^*}{B} + 17Bt\frac{N}{m^2}\right) + \frac{N}{m}\left(51\frac{r_u^*}{B} + 17Bt\frac{N}{u^2}\right).$$

The following version is based on Theorem 12 and replaces the factors N/m^2 and N/u^2 appearing in the previous excess risk bound by $1/m$ and $1/u$, respectively:

Corollary 14 *Under the assumptions of Theorem 12, for any $t > 0$, with probability greater than $1 - 2e^{-t}$, we have:*

$$\mathcal{E}_u(\hat{h}_m) \leq \frac{N}{u}\left(901\frac{K}{B}r_m^* + \frac{t(16+25B)}{3m}\right) + \frac{N}{m}\left(901\frac{K}{B}r_u^* + \frac{t(16+25B)}{3u}\right).$$

Proof sketch Corollary 13 can be proved by noticing that h_u^* is an empirical risk minimizer (similar to $\hat{h}_m$, but computed on the test set). Thus, repeating the proof of Theorem 11, we immediately obtain the same bound for h_u^* as in Theorem 11 with r_m^* and N/m^2 replaced by r_u^* and N/u^2, respectively. This shows that the overall errors $L_N(\hat{h}_m)$ and $L_N(h_u^*)$ are close to each other. It remains to apply an intermediate step, obtained in the proof of Theorem 11. Corollary 14 is proved in a similar way. The detailed proofs are presented in Appendix D. ■

In order to get a more concrete grasp of the key quantities r_m^* and r_u^* in Corollary 14, we can directly apply the machinery developed in the inductive case by Bartlett et al. (2005) to get an upper bound. For concreteness, we consider below the case of a kernel class. Observe that, by an application of Corollary 8 to the left-hand side of (6), the bounds below for the inductive r_m^*, r_u^* of Corollary 14 are valid as well for their transductive siblings of Corollary 13; though by doing so we lose essentially any potential advantage (apart from tighter multiplicative constants) of using Theorem 11/Corollary 13 over Theorem 12/Corollary 14. As pointed out in Remark 9, the regime of sampling without replacement could lead potentially to an improved bound (at least when $m = \Omega(N)$). Whether it is possible to take advantage of this fact and develop tighter bounds specifically for the fixed point of (6) is an open question and left for future work.

Corollary 15 *Let k be a positive semidefinite kernel on $\mathcal{X}$ with $\sup_{x\in\mathcal{X}} k(x,x) \leq 1$, and $\mathcal{C}_k$ the associated reproducing kernel Hilbert space. Let $\mathcal{H} := \{f \in \mathcal{C}_k : \|f\| \leq 1\}$, and $\mathcal{F}^*$ the associated excess loss class. Suppose that Assumptions 10 are satisfied and assume moreover that the loss function ℓ is L-Lipschitz in its first variable and also that $E\big(h(X) - h^*(X)\big)^2 \leq B\big(L(h) - L(h^*)\big)$ for all $h \in \mathcal{H}$. Let K_N be the $N \times N$ normalized kernel Gram matrix with entries $(K_N)_{ij} := \frac{1}{N}k(X_i, X_j)$, where $\mathbf{X}_N = (X_1, \ldots, X_N)$; denote $\lambda_{1,N} \geq \ldots \geq \lambda_{N,N}$ its ordered eigenvalues. Then, for $k = u$ or $k = m$:*

$$r_k^* \leq c_L \min_{0\leq\theta\leq k}\left(\frac{\theta}{k} + \sqrt{\frac{1}{k}\sum_{i\geq\theta}\lambda_{i,N}}\right),$$

where c_L is a constant depending only on L.

This result is obtained as a direct application of the results of Bartlett et al., 2005, Section 6.3; Mendelson, 2003, the only important point being that the generating distribution is the uniform distribution on $\mathbf{X}_N$. Similar to the discussion there, we note that r_m^* and r_u^* are at most of order $1/\sqrt{m}$ and $1/\sqrt{u}$, respectively, and possibly much smaller if the eigenvalues have a fast decay .

Remark 16 *The question of* transductive convergence rates *is somewhat delicate, since all results stated here assume a fixed set $\mathbf{X}_N$, as reflected for instance in the bound of Corollary 15 depending on the eigenvalues of the kernel Gram matrix of the set $\mathbf{X}_N$. In order to give a precise meaning to* rates*, one has to specify how $\mathbf{X}_N$ evolves as N grows. A natural setting for this is Vapnik (1998)'s second transductive setting where $\mathbf{X}_N$ is i.i.d. from some generating distribution. In that case we think it is possible to adapt once again the results of Bartlett et al. (2005) in order to relate the quantities $r_m^*(N)$ to asymptotic counterparts as $N \to \infty$, though we do not pursue this avenue in the present work.*

5. Conclusion

In this paper, we have considered the setting of transductive learning over a broad class of bounded and nonnegative loss functions. We provide excess risk bounds for the transductive learning setting based on the localized complexity of the hypothesis class, which hold under general assumptions on the loss function and the hypothesis class. When applied to kernel classes, the transductive excess risk bound can be formulated in terms of the tailsum of the eigenvalues of the kernels, similar to the best known estimates in inductive learning. The localized excess risk bound is achieved by proving two novel and very general *concentration inequalities* for suprema of empirical processes when sampling without replacement, which are of potential interest also in various other application areas in machine learning and learning theory, where they may serve as a fundamental mathematical tool.

For instance, sampling without replacement is commonly employed in the Nyström method (Kumar et al., 2012), which is an efficient technique to generate low-rank matrix approximations in large-scale machine learning. Another potential application area of our novel concentration inequalities could be the analysis of randomized sequential algorithms such as stochastic gradient descent and randomized coordinate descent, practical implementations of which often deploy sampling without replacement (Recht and Re, 2012). Very interesting also would be to explore whether the proposed techniques could be used to generalize matrix Bernstein inequalities (Tropp, 2012) to the case of sampling without replacement, which could be used to analyze matrix completion problems (Koltchinskii et al., 2011). The investigation of application areas beyond the transductive learning setting is, however, outside of the scope of the present paper.

Acknowledgments

The authors are thankful to Sergey Bobkov, Stanislav Minsker, and Mehryar Mohri for stimulating discussions and to the anonymous reviewers for their helpful comments. Marius Kloft acknowledges a postdoctoral fellowship by the German Research Foundation (DFG).

References

R. Adamczak. A tail inequality for suprema of unbounded empirical processes with applications to markov chains. *Electronic Journal of Probability*, 34(13), 2008.

R. Bardenet and O.-A. Maillard. Concentration inequalities for sampling without replacement. http://arxiv.org/abs/1309.4029, 2013.

P. Bartlett, O. Bousquet, and S. Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):14971537, 2005.

P. L. Bartlett, S. Mendelson, and P. Phillips. On the optimality of sample-based estimates of the expectation of the empirical minimizer. *ESAIM: Probability and Statistics*, 2010.

A. Blum and J. Langford. PAC-MDL bounds. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 2003.

S. Bobkov. Concentration of normalized sums and a central limit theorem for noncorrelated random variables. *Annals of Probability*, 32, 2004.

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

O. Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Acad. Sci. Paris, Ser. I*, 334:495–500, 2002a.

O. Bousquet. *Concentration Inequalities and Empirical Processes Theory Applied to the Analysis of Learning Algorithms*. PhD thesis, Ecole Polytechnique, 2002b.

C. Cortes and M. Mohri. On transductive regression. In *Advances in Neural Information Processing Systems (NIPS)*, 2006.

C. Cortes and V. Vapnik. Support-vector networks. *Mach. Learn.*, 20:273–297, September 1995. ISSN 0885-6125.

C. Cortes, M. Mohri, D. Pechyony, and A. Rastogi. Stability analysis and learning bounds for transductive regression algorithms. http://arxiv.org/abs/0904.0814, 2009.

P. Derbeko, R. El-Yaniv, and R. Meir. Explicit learning curves for transduction and application to clustering and compression algorithms. *Journal of Artificial Intelligence Research*, 22, 2004.

R. El-Yaniv and D. Pechyony. Stable transductive learning. In *Proceedings of the International Conference on Computational Learning Theory (COLT)*, 2006.

R. El-Yaniv and D. Pechyony. Transductive rademacher complexity and its applications. *Journal of Artificial Intelligence Research*, 2009.

D. Gross and V. Nesme. Note on sampling without replacing from a fnite collection of matrices. http://arxiv.org/abs/1001.2738v2, 2010.

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

T. Klein and E. Rio. Concentration around the mean for maxima of empirical processes. *The Annals of Probability*, 33(3):10601077, 2005.

V. Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):25932656, 2006.

V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: École DÉté de Probabilités de Saint-Flour XXXVIII-2008*. Ecole d'été de probabilités de Saint-Flour. Springer, 2011a.

V. Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems.* École d'été de probabilités de Saint-Flour XXXVIII-2008. Springer Verlag, 2011b.

V. Koltchinskii and D. Panchenko. Rademacher processes and bounding the risk of function learning. In D. E. Gine and J.Wellner, editors, *High Dimensional Probability, II*, pages 443–457. Birkhauser, 1999.

V. Koltchinskii, K. Lounici, and A. B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Annals of Statistics*, 39:2302–2329, 2011. doi: 10.1214/11-AOS894.

S. Kumar, M. Mohri, and A. Talwalkar. Sampling methods for the nyström method. *Journal of Machine Learning Research*, 13(1):981–1006, Apr. 2012.

P. Massart. Some applications of concentration inequalities to statistics. *Ann. Fac. Sci. Toulouse Math.*, 9(6):245–303, 2000.

S. Mendelson. On the performance of kernel classes. *J. Mach. Learn. Res.*, 4:759–771, December 2003.

D. Pechyony. *Theory and Practice of Transductive Learning*. PhD thesis, Technion, 2008.

B. Recht and C. Re. Toward a noncommutative arithmetic-geometric mean inequality: Conjectures, case-studies, and consequences. In *COLT*, 2012.

R. J. Serfling. Probability inequalities for the sum in sampling without replacement. *The Annuals of Statistics*, 2(1):39–48, 1974.

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387772413.

I. Steinwart, D. R. Hush, and C. Scovel. Learning from dependent observations. *J. Multivariate Analysis*, 100(1):175–194, 2009.

M. Stone. Cross-validatory choice and assessment of statistical predictors (with discussion). *Journal of the Royal Statistical Society*, B36:111–147, 1974.

M. Talagrand. New concentration inequalities in product spaces. *Inventiones Mathematicae*, 126, 1996.

J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

V. N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer-Verlag New York, Inc., 1982.

V. N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998.

Appendix A. Bousquet's version of Talagrand's concentration inequality

Here we use the setting and notations of Section 3.

Theorem 17 (Bousquet (2002b)) *Let $v = m\sigma^2 + 2\mathbb{E}[Q_m]$ and for $u > -1$ let $\phi(u) = e^u - u - 1$, $h(u) = (1+u)\log(1+u) - u$. Then for any $\lambda \geq 0$ the following upper bound on the moment generating function holds:*

$$\mathbb{E}\left[e^{\lambda(Q_m - \mathbb{E}[Q_m])}\right] \leq e^{v\phi(\lambda)}. \tag{8}$$

We also have for any $\epsilon \geq 0$:

$$\mathbb{P}\left\{Q_m - \mathbb{E}[Q_m] \geq \epsilon\right\} \leq \exp\left\{-vh\left(\frac{\epsilon}{v}\right)\right\}. \tag{9}$$

Noting that $h(u) \geq \frac{u^2}{2(1+u/3)}$ for $u > 0$, one can derive the following more illustrative version:

$$\mathbb{P}\left\{Q_m - \mathbb{E}[Q_m] \geq \epsilon\right\} \leq \exp\left\{-\frac{\epsilon^2}{2(v+\epsilon/3)}\right\}. \tag{10}$$

Also for all $t \geq 0$ the following holds with probability greater than $1 - e^{-t}$:

$$Q_m \leq \mathbb{E}[Q_m] + \sqrt{2vt} + \frac{t}{3}. \tag{11}$$

Note that Bousquet (2002a) provides similar bounds for $Q_m = \sup_{f\in\mathcal{F}} |\sum_{i=1}^m f(X_i)|$.

Appendix B. Proofs from Section 3

First we are going to prove Theorem 2, which is a direct consequence of Bousquet's inequality of Theorem 17. It is based on the following *reduction theorem* due to Hoeffding (1963):

Theorem 18 (Hoeffding (1963)) [10] *Let $\{U_1, \ldots, U_m\}$ and $\{W_1, \ldots, W_m\}$ be sampled uniformly from a finite set of d-dimensional vectors $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_N\} \subset \mathbb{R}^d$ with and without replacement, respectively. Then, for any continuous and convex function $F\colon \mathbb{R}^d \to \mathbb{R}$, the following holds:*

$$\mathbb{E}\left[F\left(\sum_{i=1}^m W_i\right)\right] \leq \mathbb{E}\left[F\left(\sum_{i=1}^m U_i\right)\right].$$

Also we will need the following technical lemma:

Lemma 19 *Let $\boldsymbol{x} = (x_1, \ldots, x_d)^{\mathrm{T}} \in \mathbb{R}^d$. Then the following function is convex for all $\lambda > 0$*

$$F(\boldsymbol{x}) = \exp\left(\lambda \sup_{i=1,\ldots,d} x_i\right).$$

10. Hoeffding initially stated this result only for real valued random variables. However all the steps of proof hold also for vector-valued random variables. For the reference see Section D of Gross and Nesme (2010).

Proof Let us show that, if $g\colon \mathbb{R} \to \mathbb{R}$ is a convex and nondecreasing function and $f\colon \mathbb{R}^d \to \mathbb{R}$ is convex, then $g\big(f(\boldsymbol{x})\big)$ is also convex. Indeed, for $\alpha \in [0,1]$ and $\boldsymbol{x}', \boldsymbol{x}'' \in \mathbb{R}^d$:

$$g\Big(f\big(\alpha\boldsymbol{x}' + (1-\alpha)\boldsymbol{x}''\big)\Big) \leq g\big(\alpha f(\boldsymbol{x}') + (1-\alpha)f(\boldsymbol{x}'')\big) \leq \alpha g\big(f(\boldsymbol{x}')\big) + (1-\alpha)\alpha g\big(f(\boldsymbol{x}'')\big).$$

Considering the fact that $g(y) = e^{\lambda y}$ is convex and increasing for $\lambda > 0$, it remains to show that $f(\boldsymbol{x}) = \sup_{i=1,\dots,d}(x_i)$ is convex. For all $\alpha \in [0,1]$ and $\boldsymbol{x}', \boldsymbol{x}'' \in \mathbb{R}^d$, the following holds:

$$\sup_{i=1,\dots,d}\big(\alpha x_i' + (1-\alpha)x_i''\big) \leq \alpha \sup_{i=1,\dots,d} x_i' + (1-\alpha)\sup_{i=1,\dots,d} x_i'',$$

which concludes the proof. ■

We will proove Theorem 2 for a finite class of functions $\mathcal{F} = \{f_1, \dots, f_M\}$. The result for countable case follows by taking a limit of a sequence of finite sets.

Proof of Theorem 2: Let $\{U_1, \dots, U_m\}$ and $\{W_1, \dots, W_m\}$ be sampled uniformly from a finite set of M-dimensional vectors $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_N\} \subset \mathbb{R}^M$ with and without replacement respectively, where $\boldsymbol{v}_j = \big(f_1(c_j), \dots, f_M(c_j)\big)^{\mathrm{T}}$. Using Lemma 19 and Theorem 18, we get that for all $\lambda > 0$:

$$\mathbb{E}\left[e^{\lambda Q_m'}\right] = \mathbb{E}\left[\exp\left(\lambda \sup_{j=1,\dots,M}\left(\sum_{i=1}^m W_i\right)_j\right)\right] \leq \mathbb{E}\left[\exp\left(\lambda \sup_{j=1,\dots,M}\left(\sum_{i=1}^m U_i\right)_j\right)\right] = \mathbb{E}\left[e^{\lambda Q_m}\right], \tag{12}$$

where the lower index j indicates the j-th coordinate of a vector. Using the upper bound (8) on the moment generating function of Q_m provided by Theorem 17, we proceed as follows:

$$\mathbb{E}\left[e^{\lambda Q_m'}\right] \leq \mathbb{E}\left[e^{\lambda Q_m}\right] \leq e^{\lambda\mathbb{E}[Q_m] + v\phi(\lambda)},$$

or, equivalently,

$$\mathbb{E}\left[e^{\lambda(Q_m' - \mathbb{E}[Q_m'])}\right] \leq e^{\lambda(\mathbb{E}[Q_m] - \mathbb{E}[Q_m']) + v\phi(\lambda)}.$$

Using Chernoff's method, we obtain for all $\epsilon \geq 0$ and $\lambda > 0$:

$$\mathbb{P}\left\{Q_m' - \mathbb{E}[Q_m'] \geq \epsilon\right\} \leq \frac{\mathbb{E}\left[e^{\lambda(Q_m' - \mathbb{E}[Q_m'])}\right]}{e^{\lambda\epsilon}} \leq \exp\big(\lambda(\mathbb{E}[Q_m] - \mathbb{E}[Q_m']) + v\phi(\lambda) - \lambda\epsilon\big). \tag{13}$$

The term on the right-hand side of the last inequality achieves its minimum for

$$\lambda = \log\left(\frac{v + \epsilon - \mathbb{E}[Q_m] + \mathbb{E}[Q_m']}{v}\right). \tag{14}$$

Thus we have the technical condition $\epsilon \geq \mathbb{E}[Q_m] - \mathbb{E}[Q_m']$. Otherwise we set $\lambda = 0$ and obtain a trivial bound equal to 1. The inequality $\mathbb{E}[Q_m] \geq \mathbb{E}[Q_m']$ also follows from Theorem 18 by exploiting the fact that the supremum is a convex function (which we showed in the proof of Lemma 19). Inserting (14) into (13), we obtain the first inequality of Theorem 2. The second inequality follows from observing that $h(u) \geq \frac{u^2}{2(1+u/3)}$ for $u > 0$. The deviation inequality can then be obtained using standard calculus. For details we refer to Section 2.7.2 of Bousquet (2002b). ■.

Remark 20 *It should be noted that Klein and Rio (2005) derive an upper bound on* $\mathbb{E}\left[e^{-\lambda(Q_m-\mathbb{E}[Q_m])}\right]$ *for* $\lambda \geq 0$*. This upper bound on the moment generating function together with Chernoff's method leads to an upper bound on* $\mathbb{P}\left\{\mathbb{E}[Q_m] - Q_m \geq \epsilon\right\}$*. However, the proof technique used in Theorem 2 cannot be used in this case, since Lemma 19 does not hold for negative* λ*.*

Proof of Lemma 3: We have already proved the first inequality of the lemma. Regarding the second one, using the definitions of $\mathbb{E}[Q_m]$ and $\mathbb{E}[Q'_m]$, we have:

$$\mathbb{E}[Q_m] - \mathbb{E}[Q'_m] = \frac{1}{N^m} \sum_{x_1,\ldots,x_m} \sup_{f\in\mathcal{F}} \sum_{i=1}^{m} f(x_i) + \left(\frac{1}{N^m} - \frac{(N-m)!}{N!}\right) \sum_{z_1,\ldots,z_m} \sup_{f\in\mathcal{F}} \sum_{i=1}^{m} f(z_i),$$

where the first sum is over all ordered sequences $\{x_1,\ldots,x_m\} \subset \mathcal{C}$ containing duplicate elements, while the second sum is over all ordered sequences $\{z_1,\ldots,z_m\} \subset \mathcal{C}$ with no duplicates. Note that the second sum has exactly $m! \cdot C_N^m$ summands, which means that the first one has $N^m - m! \cdot C_N^m$ summands. Considering the fact that $\frac{1}{N^m} \leq \frac{(N-m)!}{N!}$ and $f(x) \in [-1,1]$ for all $x \in \mathcal{C}$, we obtain:

$$\begin{aligned}
\mathbb{E}[Q_m] - \mathbb{E}[Q'_m] &\leq m\left(\frac{N^m - m!\cdot C_N^m}{N^m}\right) + m\left(\frac{(N-m)!}{N!} - \frac{1}{N^m}\right) m!\cdot C_N^m \\
&= 2m\left(\frac{N^m - m!\cdot C_N^m}{N^m}\right) \\
&= 2m - 2m\left(1\cdot\left(1-\frac{1}{N}\right)\cdots\left(1-\frac{m-1}{N}\right)\right) \\
&\leq 2m - 2m\left(1-\frac{m-1}{N}\right)^m \\
&= 2m\left(\frac{m-1}{N}\right)\left(1+\left(1-\frac{m-1}{N}\right)+\cdots+\left(1-\frac{m-1}{N}\right)^{m-1}\right) \\
&\leq 2m\left(\frac{m-1}{N}\right) m \\
&\leq 2\frac{m^3}{N},
\end{aligned}$$

which was to show. ■

In order to prove Theorem 1, we need to state the result presented in Theorem 2.1 of Bobkov (2004) and derive its slightly modified version. From now on we will follow the presentation in Bobkov (2004).

Let us consider the following subset of discrete cube, which we call *the slice*:

$$\mathcal{D}_{n,k} = \left\{x = (x_1,\ldots,x_n) \in \{0,1\}^n \colon x_1 + \cdots + x_n = k\right\}.$$

Neighbors are points that differ exactly in two coordinates. Thus every point $x \in \mathcal{D}_{n,k}$ has exactly $k(n-k)$ neighbours $\{s_{ij}x\}_{i\in I(x), j\in J(x)}$, where

$$I(x) = \{i \leq n \colon x_i = 1\}, \quad J(x) = \{j \leq n \colon x_j = 0\},$$

and $(s_{ij}x)_r = x_r$ for $r \neq i, j$, $(s_{ij}x)_i = x_j$, $(s_{ij}x)_j = x_i$. For any function g defined on $\mathcal{D}_{n,k}$ and $x \in \mathcal{D}_{n,k}$, let us introduce the following quantity:

$$V^g(x) = \sum_{i \in I(x)} \sum_{j \in J(x)} \big(g(x) - g(s_{ij}x)\big)^2,$$

which can be viewed as the Euclidean length of the discrete gradient $|\nabla g(x)|^2$ of the function g.

The following result can be found in Theorem 2.1 of Bobkov (2004):

Theorem 21 (Bobkov (2004)) *Consider the real-valued function g defined on $\mathcal{D}_{n,k}$ and the uniform distribution μ over the set $\mathcal{D}_{n,k}$. Assume there is a constant $\Sigma^2 \geq 0$ such that $V^g(x) \leq \Sigma^2$ for all x. Then for all $\epsilon \geq 0$:*

$$\mu\{g(x) - \mathbb{E}[g(x)] \geq \epsilon\} \leq \exp\left\{-\frac{(n+2)\epsilon^2}{4\Sigma^2}\right\}.$$

The same upper bound also holds for $\mu\{\mathbb{E}[g(x)] - g(x) \geq \epsilon\}$.

Using the notations of Section 3, we define the following function $g\colon \mathcal{D}_{N,m} \to \mathbb{R}$:

$$g(x) = \sup_{f \in \mathcal{F}} \sum_{i \in I(x)} f(c_i). \tag{15}$$

Note that, if x is distributed uniformly over the set $\mathcal{D}_{N,m}$, the random variables Q'_m and $g(x)$ are identically distributed. Thus we thus can use Theorem 21 to derive concentration inequalities for Q'_m. However, it is not trivial to bound the quantity $V^g(x)$. Instead we define the following quantity, related to $V^g(x)$:

$$V^g_+(x) = \sum_{i \in I(x)} \sum_{j \in J(x)} \big(g(x) - g(s_{ij}x)\big)^2 \mathbb{1}\{g(x) \geq g(s_{ij}x)\},$$

where $\mathbb{1}\{A\}$ is an indicator function. Now we state the following modified version of Theorem 21:

Theorem 22 *Consider a real-valued function g defined on $\mathcal{D}_{n,k}$ and the uniform distribution μ over the set $\mathcal{D}_{n,k}$. Assume there is a constant $\Sigma^2 \geq 0$ such that $V^g_+(x) \leq \Sigma^2$ for all x. Then for all $\epsilon \geq 0$:*

$$\mu\{g(x) - \mathbb{E}[g(x)] \geq \epsilon\} \leq \exp\left\{-\frac{(n+2)\epsilon^2}{8\Sigma^2}\right\}.$$

The same upper bound also holds for $\mu\{\mathbb{E}[g(x)] - g(x) \geq \epsilon\}$.

Proof We are going to follow the steps of the proof of Theorem 21, presented in Bobkov (2004). The author shows that, for any real-valued function g defined on $\mathcal{D}_{n,k}$, the following holds:

$$\begin{aligned}
&(n+2)\big(\mathbb{E}\big[e^{g(x)} \log e^{g(x)}\big] - \mathbb{E}\big[e^{g(x)}\big]\mathbb{E}\big[\log e^{g(x)}\big]\big) \\
&\leq \mathbb{E}\left[\sum_{i \in I(x)} \sum_{j \in J(x)} \big(g(x) - g(s_{ij}x)\big)\big(e^{g(x)} - e^{g(s_{ij}x)}\big)\right].
\end{aligned} \tag{16}$$

Note that for any $a, b \in \mathbb{R}$:

$$(a-b)(e^a - e^b) \leq \frac{e^a + e^b}{2}(a-b)^2. \tag{17}$$

We can re-write the right-hand side of inequality (16) in the following way:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i\in I(x)}\sum_{j\in J(x)}\big(g(x)-g(s_{ij}x)\big)\big(e^{g(x)}-e^{g(s_{ij}x)}\big)\right] \\
&= 2\cdot\mathbb{E}\left[\sum_{i\in I(x)}\sum_{j\in J(x)}\big(g(x)-g(s_{ij}x)\big)\big(e^{g(x)}-e^{g(s_{ij}x)}\big)\mathbb{1}\{g(x)\geq g(s_{ij}x)\}\right].
\end{aligned}$$

Using (17), we get:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i\in I(x)}\sum_{j\in J(x)}\big(g(x)-g(s_{ij}x)\big)\big(e^{g(x)}-e^{g(s_{ij}x)}\big)\right] \\
&\leq 2\cdot\mathbb{E}\left[\sum_{i\in I(x)}\sum_{j\in J(x)}\frac{\big(e^{g(x)}+e^{g(s_{ij}x)}\big)}{2}\big(g(x)-g(s_{ij}x)\big)^2\mathbb{1}\{g(x)\geq g(s_{ij}x)\}\right] \\
&\leq 2\cdot\mathbb{E}\left[\sum_{i\in I(x)}\sum_{j\in J(x)}e^{g(x)}\big(g(x)-g(s_{ij}x)\big)^2\mathbb{1}\{g(x)\geq g(s_{ij}x)\}\right] \\
&= 2\mathbb{E}\left[V_+^g(x)e^{g(x)}\right].
\end{aligned}$$

Thus we obtain the following inequality:

$$(n+2)\big(\mathbb{E}\big[e^{g(x)}\log e^{g(x)}\big]-\mathbb{E}\big[e^{g(x)}\big]\mathbb{E}\big[\log e^{g(x)}\big]\big)\leq 2\mathbb{E}\left[V_+^g(x)e^{g(x)}\right].$$

Applying this inequality to λg, where $\lambda \in \mathbb{R}$, we get:

$$(n+2)\Big(\mathbb{E}\big[e^{\lambda g(x)}\log e^{\lambda g(x)}\big]-\mathbb{E}\big[e^{\lambda g(x)}\big]\mathbb{E}\big[\log e^{\lambda g(x)}\big]\Big)\leq 2\mathbb{E}\left[V_+^{\lambda g}(x)e^{\lambda g(x)}\right]\leq 2\Sigma^2\lambda^2\mathbb{E}\left[e^{\lambda g(x)}\right]. \tag{18}$$

As mentioned in the proof of Theorem 21 in Bobkov (2004), inequality (18) implies the following upper bound on the moment generating function:

$$\mathbb{E}\left[e^{\lambda(g(x)-\mathbb{E}[g(x)])}\right]\leq e^{\frac{2\Sigma^2\lambda^2}{n+2}}. \tag{19}$$

This fact is known as the *Herbst argument* and plays an important role in the entropy method (Boucheron et al., 2013). Now we apply Chernoff's method, which gives us for all $\lambda, \epsilon \geq 0$:

$$\mu\left\{g(x)-\mathbb{E}[g(x)]\geq\epsilon\right\}\leq\frac{\mathbb{E}\left[e^{\lambda(g(x)-\mathbb{E}[g(x)])}\right]}{e^{\lambda\epsilon}}\leq e^{\frac{2\Sigma^2\lambda^2}{n+2}-\lambda\epsilon}.$$

We conclude the proof by choosing $\lambda = \frac{\epsilon(n+2)}{4\Sigma^2}$.

An upper bound for $\mu\{\mathbb{E}[g(x)] - g(x) \geq \epsilon\}$ can be obtained using (19) for $\lambda < 0$:

$$\mu\left\{\mathbb{E}[g(x)] - g(x) \geq \epsilon\right\} = \mu\left\{\lambda\big(g(x) - \mathbb{E}[g(x)]\big) \geq -\lambda\epsilon\right\} \leq \frac{\mathbb{E}\left[e^{\lambda(g(x)-\mathbb{E}[g(x)])}\right]}{e^{-\lambda\epsilon}} \leq e^{\frac{2\Sigma^2\lambda^2}{n+2}+\lambda\epsilon}.$$

Now it remains to choose $\lambda = -\frac{\epsilon(n+2)}{4\Sigma^2}$. ■

We will need the following technical result:

Lemma 23 *For any sequence of real numbers $\{x_1, \ldots, x_n\}$ the following holds:*

$$\frac{1}{n^2}\sum_{1\leq i<j\leq n}(x_i - x_j)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \frac{1}{n}\sum_{j=1}^{n}x_j\right)^2.$$

Proof Notice that it holds:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \frac{1}{n}\sum_{j=1}^{n}x_j\right)^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(x_i^2 - \frac{2}{n}x_i\sum_{j=1}^{n}x_j + \frac{1}{n^2}\left(\sum_{j=1}^{n}x_j\right)^2\right)\\
&= \frac{1}{n}\left(\sum_{i=1}^{n}x_i^2 - \frac{2}{n}\sum_{i=1}^{n}x_i\sum_{j=1}^{n}x_j + \frac{1}{n^2}\sum_{i=1}^{n}\left(\sum_{j=1}^{n}x_j\right)^2\right)\\
&= \frac{1}{n}\left(\sum_{i=1}^{n}x_i^2 - \frac{1}{n}\left(\sum_{j=1}^{n}x_j\right)^2\right)\\
&= \frac{1}{n^2}\left((n-1)\sum_{i=1}^{n}x_i^2 - 2\sum_{1\leq i<j\leq n}x_ix_j\right)\\
&= \frac{1}{n^2}\sum_{1\leq i<j\leq n}(x_i - x_j)^2.
\end{aligned}$$

■

Now we are ready to prove Theorem 1.

Proof of Theorem 1: We will apply Theorem 22 for the function $g(x)$ defined in (15), where x is distributed uniformly over $\mathcal{D}_{N,m}$. As noted above this will lead to a concentration inequality for Q'_m, since Q'_m and $g(x)$ are distributed identically. Hence, all we need is to obtain an upper bound on $V_+^g(x)$.

To this end, let us consider two functions $g_1, g_2 \colon \mathcal{A} \to \mathbb{R}$, defined on some set $\mathcal{A}$. Assume that $\sup_{a\in\mathcal{A}} g_1(a) = g_1(\bar{a})$ for some $\bar{a} \in \mathcal{A}$. Then the following holds:

$$\left(\sup_{a\in\mathcal{A}} g_1(a) - \sup_{a\in\mathcal{A}} g_2(a)\right)^2 \mathbb{1}\left\{\sup_{a\in\mathcal{A}} g_1(a) \geq \sup_{a\in\mathcal{A}} g_2(a)\right\} \leq \big(g_1(\bar{a}) - g_2(\bar{a})\big)^2. \tag{20}$$

Let us assume that, for $x \in \mathcal{D}_{N,m}$, the supremum in the definition (15) of $g(x)$ is achieved for $\bar{f} \in \mathcal{F}$. Then we have:

$$\begin{aligned}
V_+^g(x) &= \sum_{i \in I(x)} \sum_{j \in J(x)} \left(g(x) - g(s_{ij}x)\right)^2 \mathbb{1}\left\{g(x) \geq g(s_{ij}x)\right\} \\
&\leq \sum_{i \in I(x)} \sum_{j \in J(x)} \left(\sum_{k \in I(x)} \bar{f}(c_k) - \sum_{k \in I(s_{ij}x)} \bar{f}(c_k) \right)^2 \\
&= \sum_{i \in I(x)} \sum_{j \in J(x)} \left(\bar{f}(c_i) - \bar{f}(c_j) \right)^2 \\
&\leq \sum_{1 \leq i < j \leq N} \left(\bar{f}(c_i) - \bar{f}(c_j) \right)^2 \\
&= N^2 \mathbb{V}\left[\bar{f}(X_1)\right],
\end{aligned}$$

where the first inequality follows from (20) and the second inequality follows from Lemma 23. Now note that, since the function $\bar{f}$ depends on the choice of x, the following holds for all $x \in \mathcal{D}_{N,m}$:

$$V_+^g(x) \leq N^2 \sup_{f \in \mathcal{F}} \mathbb{V}[f(X_1)] = N^2 \sigma^2.$$

We conclude the proof by an application of Theorem 22. ■

Appendix C. Further discussions on Section 3

First we note that the result of Theorem 4 is uniformly sharper than what could have been obtained for Q_m using McDiarmid's inequality, by a factor of $\frac{N-1/2}{N-m}$ (fraction of the training sample) in the exponent. This suggests that when sampling without replacement things are more concentrated than when sampling with replacement. This general phenomenon is pointed out by several authors: Serfling (1974) obtains a refinement of Hoeffding's inequality, El-Yaniv and Pechyony (2009) improves McDiarmid's inequality, and Bardenet and Maillard (2013) improve Bennet's and Bernstein's inequalities in the same way for sampling without replacement—opposed to the fact that the results of Theorems 1 and 2 unfortunately do not improve the known analogues for Q_m. This drawback can possibly be overcome by a more detailed analysis. This direction is left for the future work.

Appendix D. Proofs for Section 4

Proof of Theorem 6: Applying Theorem 2, we get that with probability greater than $1 - e^{-t}$:

$$\sup_{h \in \mathcal{H}} \left(L_N(h) - \hat{L}_m(h) \right) \leq E_m + \sqrt{2\left(\sigma_{\mathcal{H}}^2 + 2E_m\right)\frac{t}{m}} + \frac{t}{3m},$$

which can be further simplified using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ and then $\sqrt{ab} \leq \frac{a+b}{2}$ in the following way:

$$\sup_{h \in \mathcal{H}} \left(L_N(h) - \hat{L}_m(h) \right) \leq E_m + \sqrt{\frac{2\sigma_{\mathcal{H}}^2 t}{m}} + 2\sqrt{\frac{E_m t}{m}} + \frac{t}{3m}$$
$$\leq 2E_m + \sqrt{\frac{2\sigma_{\mathcal{H}}^2 t}{m}} + \frac{4t}{3m}.$$

■

The proof of Theorem 11 is based on the following intermediate result appearing in the proof of Theorem 3.3 in Bartlett et al. (2005). We state it as a lemma:

Lemma 24 *(Peeling Lemma using Theorem 1) Assume the conditions of Theorem 11 hold. Fix some $\lambda > 1$. For $w(r, f) = \min\{r\lambda^k \colon k \in \mathbb{N}, r\lambda^k \geq Ef^2\}$, define the following rescaled version of excess loss class:*

$$\mathcal{G}_r = \left\{ \frac{r}{w(r, f)} f \colon f \in \mathcal{F}^* \right\}.$$

Then for any $r > r_m^*$ and $t > 0$, with probability greater than $1 - e^{-t}$, we have:

$$\sup_{g \in \mathcal{G}_r} Eg - \hat{E}_m g \leq \sqrt{r} \left(5\frac{\sqrt{r_m^*}}{B} + 2\sqrt{2t\frac{N}{m^2}} \right). \tag{21}$$

Proof We can repeat exactly the same steps presented in the proof of the first part of Theorem 3.3 of Bartlett et al. (2005) (see pages 15–16), but using Theorem 1 in place of Talagrand's inequality.

Clearly, for any $f \in \mathcal{F}^*$, we have

$$\mathbb{V}[f(X)] = Ef^2 - (Ef)^2 \leq Ef^2. \tag{22}$$

Let us now fix some $\lambda > 1$ and $r > 0$ and introduce the following rescaled version of excess loss class:

$$\mathcal{G}_r = \left\{ \frac{r}{w(r, f)} f \colon f \in \mathcal{F}^* \right\},$$

where $w(r, f) = \min\{r\lambda^k \colon k \in \mathbb{N}, r\lambda^k \geq Ef^2\}$. Let us consider functions $f \in \mathcal{F}^*$ such that $Ef^2 \leq r$, meaning $w(r, f) = r$. The functions $g \in \mathcal{G}_r$ corresponding to those functions satisfy $g = f$ and thus $\mathbb{V}[g(X)] = \mathbb{V}[f(X)] \leq Ef^2 \leq r$. Otherwise, if $Ef^2 > r$, then $w(r, f) = \lambda^k r$, and thus the functions $g \in \mathcal{G}_r$ corresponding to them satisfy $g = f/\lambda^k$ and $Ef^2 \in (r\lambda^{k-1}, r\lambda^k]$. Thus we have $\mathbb{V}[g] = \mathbb{V}[f]/\lambda^{2k} \leq Ef^2/\lambda^{2k} \leq r$. We conclude that, for any $g \in \mathcal{G}_r$, it holds $\mathbb{V}[g(x)] \leq r$.

Now we want to upper bound the following quantity:

$$V_r = \sup_{g \in \mathcal{G}_r} Eg - \hat{E}_m g.$$

Note that any $f \in \mathcal{F}^*$ satisfies $f(X) \in [-1, 1]$, and, consequently, all $g \in \mathcal{G}_r$ satisfy $g(X) \in [-1, 1]$. Notice that

$$\frac{1}{2}\left(Eg - \hat{E}_m g \right) = \frac{1}{m} \sum_{X \in \boldsymbol{X}_m} \frac{Eg - g(X)}{2}.$$

Note that $(Eg - g(X))/2 \in [-1, 1]$ and also $\mathbb{E}\left[Eg - g(X)\right] = 0$. Since Eg is not random, using (22), we also have

$$\mathbb{V}[(Eg - g(X))/2] = \mathbb{V}[g(x)]/4 \leq r/4$$

for all $g \in \mathcal{G}_r$. We can now apply either Theorem 1 or Theorem 2 for the following function class: $\{(Eg - g(X))/2, g \in \mathcal{G}_r\}$. Here we present the proof based on Theorem 1. Applying it we get that for all $t > 0$ with probability greater than $1 - e^{-t}$, we have:

$$\begin{aligned}\frac{1}{2}\sup_{g\in\mathcal{G}_r} Eg - \hat{E}_m g &\leq \frac{1}{2}\mathbb{E}\left[\sup_{g\in\mathcal{G}_r} Eg - \hat{E}_m g\right] + 2\sqrt{2t\left(\frac{N}{m^2}\right)\frac{1}{4}\sup_{g\in\mathcal{G}_r}\mathbb{V}[g(X)]} \\ &\leq \frac{1}{2}\mathbb{E}\left[\sup_{g\in\mathcal{G}_r} Eg - \hat{E}_m g\right] + \sqrt{2t\left(\frac{N}{m^2}\right)r}\end{aligned}$$

or, rewriting,

$$V_r \leq \mathbb{E}[V_r] + 2\sqrt{2t\left(\frac{N}{m^2}\right)r}. \tag{23}$$

Now we set $\mathcal{F}^*(x, y) = \{f \in \mathcal{F}^* \colon x \leq Ef^2 \leq y\}$. Note that by the assumptions of the theorem, for $f \in \mathcal{F}^*$, we have $\mathbb{V}[f(X)] \leq Ef^2 \leq B \cdot Ef \leq B$. Define k to be the smallest integer such that $r\lambda^{k+1} \geq B$. Notice that, for any sets A and B, we have:

$$\mathbb{E}\left[\sup_{g\in A\cup B} Eg - \hat{E}_m g\right] \leq \mathbb{E}\left[\sup_{g\in B} Eg - \hat{E}_m g\right] + \mathbb{E}\left[\sup_{g\in A} Eg - \hat{E}_m g\right].$$

Indeed, since supremum is a convex function, we can use Jensen's inequality to show that each of the terms is positive. Then we have:

$$\begin{aligned}\mathbb{E}[V_r] &= \mathbb{E}\left[\sup_{g\in\mathcal{G}_r} Eg - \hat{E}_m g\right] \\ &\leq \mathbb{E}\left[\sup_{f\in\mathcal{F}^*(0,r)} Ef - \hat{E}_m f\right] + \mathbb{E}\left[\sup_{f\in\mathcal{F}^*(r,B)} \frac{r}{w(r,f)}(Ef - \hat{E}_m f)\right] \\ &\leq \mathbb{E}\left[\sup_{f\in\mathcal{F}^*(0,r)} Ef - \hat{E}_m f\right] + \sum_{j=0}^{k}\mathbb{E}\left[\sup_{f\in\mathcal{F}^*(r\lambda^j, r\lambda^{j+1})} \frac{r}{w(r,f)}(Ef - \hat{E}_m f)\right] \\ &\leq \mathbb{E}\left[\sup_{f\in\mathcal{F}^*(0,r)} Ef - \hat{E}_m f\right] + \sum_{j=0}^{k}\lambda^{-j}\mathbb{E}\left[\sup_{f\in\mathcal{F}^*(r\lambda^j, r\lambda^{j+1})} (Ef - \hat{E}_m f)\right] \\ &\leq \frac{\psi_m(r)}{B} + \frac{1}{B}\sum_{j=0}^{k}\lambda^{-j}\psi(r\lambda^{j+1})\end{aligned}$$

where in the last step we used the assumptions of the theorem. Now since ψ_m is sub-root, for any $\beta \geq 1$, we have $\psi_m(\beta r) \leq \sqrt{\beta}\psi_m(r)$. Thus

$$\mathbb{E}[V_r] \leq \frac{\psi_m(r)}{B}\left(1 + \sqrt{\lambda}\sum_{j=0}^{k}\lambda^{-j/2}\right).$$

Taking $\lambda = 4$ the r.h.s. is upper bounded by $5\psi_m(r)/B$. Finally we note that for $r \geq r_m^*$ we have that, for all $r \geq r_m^*$, it holds $\psi_m(r) \leq \sqrt{r/r_m^*}\psi_m(r_m^*) = \sqrt{rr_m^*}$ and thus

$$\mathbb{E}[V_r] \leq \frac{5}{B}\sqrt{rr_m^*}.$$

Inserting this upper bound into (23), we conclude the proof. ■

Proof of Theorem 11: Using Lemma 24, we obtain that, for any $r > r_m^*$, $t > 0$, and $\lambda > 1$, with probability greater than $1 - e^{-t}$, we have:

$$\sup_{g \in \mathcal{G}_r} Eg - \hat{E}_m g \leq \sqrt{r}\left(5\frac{\sqrt{r_m^*}}{B} + 2\sqrt{2t\frac{N}{m^2}}\right), \tag{24}$$

where we introduced the rescaled excess loss class:

$$\mathcal{G}_r = \left\{\frac{r}{w(r,f)}f \colon f \in \mathcal{F}^*\right\},$$

and $w(r,f) = \min\{r\lambda^k \colon k \in \mathbb{N}, r\lambda^k \geq Ef^2\}$. Now we want to chose $r_0 > r_m^*$ in such a way that the upper bound of (24) becomes of a form $r_0/(\lambda BK)$. We achieve this by setting:

$$r_0 = K^2\lambda^2\left(5\sqrt{r_m^*} + 2B\sqrt{2t\frac{N}{m^2}}\right)^2 > r_m^*.$$

Inserting $r = r_0$ into (24) we obtain:

$$\sup_{g \in \mathcal{G}_{r_0}} Eg - \hat{E}_m g \leq \frac{r_0}{\lambda BK}. \tag{25}$$

Moreover, using $(u+v)^2 \leq 2(u^2+v^2)$, we have

$$r_0 \leq 50K^2\lambda^2 r_m^* + 16K^2\lambda^2 B^2 t\left(\frac{N}{m^2}\right). \tag{26}$$

Recall that, for any $r > 0$ and all $g \in \mathcal{G}_r$, the following holds with probability 1:

$$Eg - \hat{E}_m g \leq \sup_{g \in \mathcal{G}_r} Eg - \hat{E}_m g.$$

Using definition of $\mathcal{G}_r$, we get that, for all $f \in \mathcal{F}^*$, the following holds with probability 1:

$$E\left(\frac{r}{w(r,f)}f\right) - \hat{E}_m\left(\frac{r}{w(r,f)}f\right) \leq \sup_{g \in \mathcal{G}_r} Eg - \hat{E}_m g,$$

or, equivalently,

$$Ef - \hat{E}_m f \leq \frac{w(r,f)}{r}\sup_{g \in \mathcal{G}_r} Eg - \hat{E}_m g.$$

Setting $r = r_0$ and using (25), we obtain that with probability greater than $1 - e^{-t}$

$$\forall f \in \mathcal{F}^*, \forall K > 1: \quad Ef - \hat{E}_m f \leq \frac{w(r_0, f)}{r_0} \frac{r_0}{\lambda K B} = \frac{w(r_0, f)}{\lambda K B}.$$

Now we will use Assumption 10.2. If, for $f \in \mathcal{F}^*$, $Ef^2 \leq r_0$ then $w(r_0, f) = r_0$ and using (26) we obtain:

$$Ef - \hat{E}_m f \leq \frac{w(r_0, f)}{\lambda K B} = \frac{r_0}{\lambda K B} \leq 50 \frac{K}{B} \lambda r_m^* + 16 \lambda K B t \left(\frac{N}{m^2} \right)$$

or, rewriting,

$$Ef \leq \hat{E}_m f + 50 \frac{K}{B} \lambda r_m^* + 16 \lambda K B t \left(\frac{N}{m^2} \right). \tag{27}$$

If otherwise $Ef^2 > r_0$, then $w(r_0, f) = \lambda^i r_0$ for certain value of $i > 0$ and also $Ef^2 \in (r_0 \lambda^{i-1}, r_0 \lambda^i]$. Then we have:

$$Ef - \hat{E}_m f \leq \frac{w(r_0, f)}{\lambda K B} = \frac{\lambda^i r_0}{\lambda K B} = \frac{\lambda \cdot (\lambda^{i-1} r_0)}{\lambda K B} \leq \frac{Ef^2}{KB} \leq \frac{Ef}{K}.$$

Thus

$$Ef \leq \frac{K}{K-1} \hat{E}_m f. \tag{28}$$

Combining (27) and (28), we finally get that with probability greater than $1 - e^{-t}$

$$\forall f \in \mathcal{F}^*, \forall K > 1: \quad Ef \leq \inf_{K>1} \frac{K}{K-1} \hat{E}_m f + 50 \frac{K}{B} \lambda r_m^* + 16 \lambda K B t \left(\frac{N}{m^2} \right). \tag{29}$$

In the very last step, we recall the definition of $\mathcal{F}^*$ and put $\hat{f}_m = \ell_{\hat{h}_m} - \ell_{h_N^*}$. Notice that

$$\begin{aligned} \hat{E}_m \hat{f}_m &= \hat{E}_m \ell_{\hat{h}_m} - \hat{E}_m \ell_{h_N^*} \\ &= \hat{L}_m(\hat{h}_m) - \hat{L}_m(h_N^*) \leq 0, \end{aligned}$$

while

$$E\hat{f}_m = L_N(\hat{h}_m) - L_N(h_N^*),$$

which concludes the proof. ■

Let $\{\xi_1, \ldots, \xi_m\}$ be random variables sampled *with replacement* from $\boldsymbol{X}_N$. Denote

$$E_{r,m} = \mathbb{E}\left[\sup_{f \in \mathcal{F}^*: Ef^2 \leq r} \left(Ef - \frac{1}{m} \sum_{i=1}^{m} f(\xi_i) \right) \right]. \tag{30}$$

Repeating the proof of peeling Lemma 24 and using Theorem 2 instead of Theorem 1 we immediately obtain the following result:

Lemma 25 *(Peeling Lemma using Theorem 2) Let $\mathcal{H}$ and ℓ be such that Assumptions 10 are satisfied. Assume there is a sub-root function $\psi_m(r)$ such that*

$$B \cdot E_{r,m} \leq \psi_m(r),$$

where $E_{r,m}$ was defined in (30). Let r_m^* be a fixed point of $\psi_m(r)$.

Fix some $\lambda > 1$. Then, for any $r > r_m^*$ and $t > 0$, with probability greater than $1 - e^{-t}$, we have:

$$\sup_{g \in \mathcal{G}_r} Eg - \hat{E}_m g \leq \sqrt{r}\left(15\frac{\sqrt{r_m^*}}{B} + \sqrt{\frac{2t}{m}}\right) + \frac{8t}{3m}, \tag{31}$$

where for $w(r,f) = \min\{r\lambda^k \colon k \in \mathbb{N}, r\lambda^k \geq Ef^2\}$ we define the following rescaled version of excess loss class:

$$\mathcal{G}_r = \left\{\frac{r}{w(r,f)} f \colon f \in \mathcal{F}^*\right\}.$$

Proof of Theorem 12 is based on the peeling Lemma 25 and is similar to the one of Theorem 11.

Proof of Corollary 13: Note that, since h_u^* is also an empirical risk minimizer (computed on the test set), the results of Theorems 11 and 12 also hold for h_u^* with every m in the statement replaced by u. Also note that the following holds almost surely:

$$\begin{aligned}
0 &\leq L_N(\hat{h}_m) - L_N(h_N^*) \\
&= L_N(\hat{h}_m) - L_N(h_N^*) - \hat{L}_m(\hat{h}_m) + \hat{L}_m(h_N^*) + \hat{L}_m(\hat{h}_m) - \hat{L}_m(h_N^*) \\
&\leq L_N(\hat{h}_m) - L_N(h_N^*) - \hat{L}_m(\hat{h}_m) + \hat{L}_m(h_N^*) \\
&= \frac{u}{N}\left(L_u(\hat{h}_m) - L_u(h_N^*) - \hat{L}_m(\hat{h}_m) + \hat{L}_m(h_N^*)\right)
\end{aligned} \tag{32}$$

and

$$\begin{aligned}
0 &\leq L_N(h_u^*) - L_N(h_N^*) \\
&= L_N(h_u^*) - L_N(h_N^*) - L_u(h_u^*) + L_u(h_N^*) + L_u(h_u^*) - L_u(h_N^*) \\
&\leq L_N(h_u^*) - L_N(h_N^*) - L_u(h_u^*) + L_u(h_N^*) \\
&= \frac{m}{N}\left(\hat{L}_m(h_u^*) - \hat{L}_m(h_N^*) - L_u(h_u^*) + L_u(h_N^*)\right),
\end{aligned}$$

where last equations in both cases use the following:

$$N \cdot L_N(h) = m \cdot \hat{L}_m(h) + u \cdot L_u(h).$$

Now we are going to use inequality (29) obtained in the proof of Theorem 11. Using the last equation in (32) and, subsequently, employing (29) for $f = \ell_{\hat{h}_m} - \ell_{h_N^*}$, where we subtract $\hat{E}_m f$ from both sides of (29), we obtain:

$$\begin{aligned}
0 &\leq L_u(\hat{h}_m) - L_u(h_N^*) - \hat{L}_m(\hat{h}_m) + \hat{L}_m(h_N^*) \\
&\leq \frac{N}{u}\left(\inf_{K>1} \frac{1}{K-1} \underbrace{\hat{L}_m(\hat{h}_m - h_N^*)}_{\leq 0} + 50\frac{K}{B}\lambda r_m^* + 16\lambda K B t \frac{N}{m^2}\right),
\end{aligned}$$

which holds with probability greater than $1 - e^{-t}$. As noted above the same argument can be used for h_u^*, which gives that the following holds:

$$\begin{aligned} 0 &\leq \hat{L}_m(h_u^*) - \hat{L}_m(h_N^*) - L_u(h_u^*) + L_u(h_N^*) \\ &\leq \frac{N}{m}\left(\inf_{K>1} \frac{1}{K-1} \underbrace{L_u(h_u^* - h_N^*)}_{\leq 0} + 50\frac{K}{B}\lambda r_u^* + 16\lambda K B t \frac{N}{u^2} \right), \end{aligned}$$

with probability greater than $1 - e^{-t}$. The union bound gives us that both inequalities hold simultaneously with probability greater than $1 - 2e^{-t}$. Or, equivalently,

$$0 \leq L_u(\hat{h}_m) - L_u(h_N^*) - \hat{L}_m(\hat{h}_m) + \hat{L}_m(h_N^*) \leq \frac{N}{u}\left(50K\lambda\frac{r_m^*}{B} + 16\lambda K B t \frac{N}{m^2} \right)$$

and

$$0 \leq \hat{L}_m(h_u^*) - \hat{L}_m(h_N^*) - L_u(h_u^*) + L_u(h_N^*) \leq \frac{N}{m}\left(50K\lambda\frac{r_u^*}{B} + 16\lambda K B t \frac{N}{u^2} \right).$$

Summing these two inequalities we obtain

$$\begin{aligned} 0 &\leq L_u(\hat{h}_m) - L_u(h_u^*) - \hat{L}_m(\hat{h}_m) + \hat{L}_m(h_u^*) \\ &\leq \frac{N}{u}\left(50\lambda K\frac{r_m^*}{B} + 16\lambda K B t \frac{N}{m^2} \right) + \frac{N}{m}\left(50\lambda K\frac{r_u^*}{B} + 16\lambda K B t \frac{N}{u^2} \right). \end{aligned}$$

Using the fact that $\hat{h}_m$ and h_u^* are the empirical risk minimizers on the training and test sets, respectively, we finally get:

$$\begin{aligned} 0 &\leq L_u(\hat{h}_m) - L_u(h_u^*) \\ &\leq \hat{L}_m(\hat{h}_m) - \hat{L}_m(h_u^*) + \frac{N}{u}\left(50\lambda K\frac{r_m^*}{B} + 16\lambda K B t \frac{N}{m^2} \right) + \frac{N}{m}\left(50\lambda K\frac{r_u^*}{B} + 16\lambda K B t \frac{N}{u^2} \right) \\ &\leq \frac{N}{u}\left(50\lambda K\frac{r_m^*}{B} + 16\lambda K B t \frac{N}{m^2} \right) + \frac{N}{m}\left(50\lambda K\frac{r_u^*}{B} + 16\lambda K B t \frac{N}{u^2} \right). \end{aligned}$$

■

Proof of Corollary 14 repeats the same steps using Theorem 12 instead of Theorem 11.

We also provide the following auxiliary result:

Corollary 26 *Under the assumptions of Theorem 11, for any $t > 0$ and any $K > 1$, with probability greater than $1 - 2e^{-t}$, we have:*

$$\begin{aligned} |L_N(\hat{h}_m) - L_N(h_u^*)| &\leq \max\left(2K\frac{r_m^*}{B} + 16KBt\left(\frac{N}{m^2}\right), 2K\frac{r_u^*}{B} + 16KBt\left(\frac{N}{u^2}\right) \right) \\ &\leq 2K\frac{r_m^* + r_u^*}{B} + 16KBtN\left(\frac{1}{m^2} + \frac{1}{u^2}\right). \end{aligned} \tag{33}$$

Proof Notice that $L_N(h_u^*) - L_N(h_N^*) \geq 0$ as well as $L_N(\hat{h}_m) - L_N(h_N^*) \geq 0$ and then combine Theorem 11 with its analogue for h_u^* in a union bound. ■