Proceedings of Machine Learning Research vol 132:1–25, 2021 32nd International Conference on Algorithmic Learning Theory

Stable Sample Compression Schemes: New Applications and an Optimal SVM Margin Bound

Steve Hanneke STEVE.HANNEKE@GMAIL.COM
Toyota Technological Institute at Chicago

Aryeh Kontorovich KARYEH@CS.BGU.AC.IL
Ben-Gurion University

Editors: Vitaly Feldman, Katrina Ligett and Sivan Sabato

Abstract

We analyze a family of supervised learning algorithms based on sample compression schemes that are *stable*, in the sense that removing points from the training set which were not selected for the compression set does not alter the resulting classifier. We use this technique to derive a variety of novel or improved data-dependent generalization bounds for several learning algorithms. In particular, we prove a new margin bound for SVM, removing a log factor. The new bound is provably optimal. This resolves a long-standing open question about the PAC margin bounds achievable by SVM.

Keywords: sample compression; support vector machines; margin

1. Introduction

The recent work of Bousquet, Hanneke, Moran, and Zhivotovskiy (2020a,b) introduced a new technique for proving PAC generalization guarantees for a special type of compression scheme, called a *stable* compression scheme: namely, a compression scheme (κ, ρ) for which any $S, S' \in (\mathcal{X} \times \mathcal{Y})^*$ with $\kappa(S) \subseteq S' \subseteq S$ has $\rho(\kappa(S')) = \rho(\kappa(S))$. They proved that for any stable compression scheme of size k, it holds with probability at least $1 - \delta$ over the draw of n i.i.d. samples S, that the risk $R(\rho(\kappa(S))) = O\left(\frac{k}{n} + \frac{1}{n}\log\left(\frac{1}{\delta}\right)\right)$, provided that the sample S is guaranteed to be separable by the compression scheme: that is, the empircal risk $\hat{R}_S(\rho(\kappa(S))) = 0$. This presents an improvement over the traditional analysis of general compression schemes, which has an additional $\log n$ factor on the first term (Littlestone and Warmuth, 1986; Floyd and Warmuth, 1995). They used this result to provide new PAC generalization guarantees for a variety of learning algorithms and techniques, including establishing that the *Support Vector Machine* (SVM) algorithm for learning linear separators in $\mathbb{R}^d$ obtains the minimax optimal bound $O\left(\frac{d}{n} + \frac{1}{n}\log\left(\frac{1}{\delta}\right)\right)$, which resolved a long-standing open question dating back to the early work of Vapnik and Chervonenkis (1974).

In the present work, we explore further implications of this technique for analyzing stable compression schemes. Our contributions are of two types. First, while the applications discussed by Bousquet, Hanneke, Moran, and Zhivotovskiy (2020a,b) focused on data-independent generalization bounds, in the present work we investigate further implications of this technique toward providing improved *data-dependent* PAC generalization bounds for several learning algorithms. In particular, one of the results we prove in this vein is a sharp *margin* bound, which holds for both

SVM and Perceptron: namely, a bound (holding with probability at least $1-\delta$) of the form

$$R_P(\hat{h}) = O\left(\frac{r^2}{\gamma^2}\frac{1}{n} + \frac{1}{n}\log\left(\frac{1}{\delta}\right)\right),$$

where γ is the geometric margin and r is the radius of the data. This refines all previous existing margin bounds for SVM by a log factor, and is provably *optimal*. Establishing a bound of this form has been a known open question in the literature for many years (see Hanneke and Kontorovich, 2019a, for background). Our results on data-dependent PAC generalization bounds also include a new tighter data-dependent bound for all empirical risk minimization algorithms in general VC classes. We also prove a new tighter bound on the probability in the region of disagreement of version spaces, which has implications for the analysis of disagreement-based active learning algorithms.

A second type of extension we provide is to establish new PAC generalization bounds for stable compression schemes in the *agnostic* setting: that is, where $\hat{R}_S(\rho(\kappa(S)))$ may be non-zero. We specifically prove that, with probability at least $1-\delta$, $|\hat{R}_S(\rho(\kappa(S))) - R(\rho(\kappa(S)))| = O\left(\sqrt{\frac{1}{n}\left(|\kappa(S)| + \log\left(\frac{1}{\delta}\right)\right)}\right)$, which is provably better than is achievable by general (non-stable) sample compression schemes in the agnostic setting. We additionally establish a new Bernstein-type bound for stable sample compression schemes, of the form $|\hat{R}_S(\rho(\kappa(S))) - R(\rho(\kappa(S)))| = O\left(\sqrt{\hat{R}_S(\rho(\kappa(S)))\frac{1}{n}\left(|\kappa(S)| + \log\left(\frac{1}{\delta}\right)\right)} + \frac{1}{n}\left(|\kappa(S)| + \log\left(\frac{1}{\delta}\right)\right)\right)$. As a concrete implication of these general results, we provide a sharper generalization bound for the technique of *compressed nearest neighbor* prediction studied by Hanneke, Kontorovich, Sabato, and Weiss (2019).

2. Main Results

This section provides formal statements of the main results of this work.

2.1. Notation

Before stating our results more formally, we introduce some basic notation to be used throughout. We denote by $\mathcal{X}$ an *instance space*: a non-empty set equipped with a σ-algebra specifying the measurable subsets. Denote by $\mathcal{Y}$ a *label space*. For our general results, $\mathcal{Y}$ may be any non-empty set (equipped with a σ-algebra specifying the measurable subsets), though our applications will focus on the case $\mathcal{Y} = \{-1, 1\}$, corresponding to *binary classification*. We refer to any measurable function $h : \mathcal{X} \to \mathcal{Y}$ as a *classifier*. For any distribution P on $\mathcal{X} \times \mathcal{Y}$, define the *risk* of a classifier h by $R_P(h) = P((x, y) : h(x) \neq y)$. For any finite *data set* $S \in (\mathcal{X} \times \mathcal{Y})^*$, also define the *empirical risk*: $\hat{R}_S(h) = \frac{1}{|S|}\sum_{(x,y)\in S} \mathbb{1}[h(x) \neq y]$. Also, in the results below, we use the convention $\log(x) = \max\{\ln(x), 1\}$ for any $x \geq 0$.

2.2. Optimal Margin Bounds for SVM and Perceptron

Linear classifiers lie at the very foundations of modern machine learning theory (Aizerman et al., 1964) — and as such, their risk rates have been an active topic of research. The agnostic case is well-understood: if all but a few of n labeled data points residing on the d-dimensional unit sphere are linearly separated with margin at least γ (the few exceptions being treated as sample errors), then the expected excess risk decays as (Devroye et al., 1996; Mohri et al., 2012) $\Theta\left(\sqrt{\min\{d, 1/\gamma^2\}/n}\right)$. For the separable case, in which there exists a hyperplane in $\mathbb{R}^d$ consistent with the n sample points and having margin at least γ, the best guarantee on the expected risk by any learning algorithm is lower-bounded by $\Omega(\min\{d, 1/\gamma^2\}/n)$. Similarly, any generalization bound that holds with probability $1-\delta$ is lower bounded by $\Omega\left(\left(\min\{d, 1/\gamma^2\} + \log(1/\delta)\right)/n\right)$. Nearly matching upper bounds are readily available via standard VC theory and Rademacher analysis, but these have additional $\log n$ factors. For weaker in-expectation bounds, the upper and lower bounds match up to constants. Since Vapnik and Chervonenkis (1974), where the margin-based in-expectation bound was first stated, it has been an open problem to extend these to tight high-probability bounds. (We refer the reader to Hanneke and Kontorovich (2019c) for proofs of the aforementioned claims as well as comprehensive background.) Recently, Bousquet et al. (2020a) obtained the sharp high-probability upper bound in terms of the dimension d.

Our contribution. We prove an optimal PAC margin bound for SVM. This matches minimax lower bounds, and is therefore the first proof that SVM achieves the optimal margin bound (previously only known to be achievable by certain online-to-batch conversion techniques applied to Perceptron).

For any $d \in \mathbb{N}$ and data set $S \in (\mathbb{R}^d \times \{-1, 1\})^*$, let $r(S) := \max_{(\boldsymbol{x},y)\in S} \|\boldsymbol{x}\|$ and let

$$\gamma(S) := \max_{\boldsymbol{w}\in\mathbb{R}^d, b\in\mathbb{R}: \|\boldsymbol{w}\|=1} \min_{(\boldsymbol{x},y)\in S} y(\boldsymbol{w}\cdot\boldsymbol{x} + b).$$

Define $\hat{h}_{\mathrm{SVM}} = \mathrm{SVM}(S)$ as the function $\hat{h}_{\mathrm{SVM}}(x) = \mathrm{sign}(\hat{\boldsymbol{w}}\cdot\boldsymbol{x} + \hat{b})$ where $\hat{\boldsymbol{w}}, \hat{b}$ realize the max in the above definition.

Theorem 1 *For any distribution P, any $n \in \mathbb{N}$, and any $\delta \in (0,1)$, for $S \sim P^n$, with probability at least $1-\delta$, if S is linearly separable, then for $\hat{h}_{\mathrm{SVM}} = \mathrm{SVM}(S)$, letting $r = r(S)$ and $\gamma = \gamma(S)$,*

$$R_P(\hat{h}_{\mathrm{SVM}}) = O\left(\frac{r^2}{\gamma^2}\frac{1}{n} + \frac{1}{n}\log\left(\frac{1}{\delta}\right)\right).$$

This matches (up to numerical constants) a known minimax lower bound, and therefore establishes that SVM achieves the optimal PAC margin bound (up to numerical constants). This optimal margin bound was previously only known for a certain (more-involved) online-to-batch conversion technique of Littlestone (1989) applied to the Perceptron algorithm.

Additionally, we extend a result on online-to-batch conversion to prove that the *Perceptron* online learning algorithm also achieves the *same* optimal PAC margin bound. Again, this bound was previously only known to be achievable for certain modified variants of Perceptron (e.g., which record all the intermediate classifiers and select on based on a hold-out set). Here we show that, if we simply take $\hat{h}_p$ as the *final* predictor from cycling Perceptron through the data set until it makes a complete pass without making any mistakes, then $\hat{h}_p$ also achieves the optimal margin bound. Formally, the algorithm is defined as follows.

Definition 2 (Perceptron) *If S is a linearly separable data set, then the* Perceptron *algorithm (Rosenblatt, 1958), denoted by $\mathcal{A}_p$, initializes $(\boldsymbol{w}, b) \in \mathbb{R}^{d+1}$ to $\mathbf{0}$ and cycles through S (in order), evaluating $\mathrm{sign}(\boldsymbol{w}\cdot\boldsymbol{x}_i + b)$ on each data point. On each mistake (i.e., $\mathrm{sign}(\boldsymbol{w}\cdot\boldsymbol{x}_i + b) \neq y_i$), $(\boldsymbol{w}, b)$ is updated via the rule $\boldsymbol{w} \leftarrow \boldsymbol{w} + y_i\boldsymbol{x}_i$ and $b \leftarrow b + y_i$.*

The key property of Perceptron that enables us to obtain margin bounds is the following. For data with x components lying in a ball of radius r, and linearly separable with margin γ, Perceptron was shown by Novikoff (1963) to make at most $\frac{r^2+1}{\gamma^2}$ mistakes ($\mathcal{A}_p$ does not terminate on non-separable inputs). Here the "+1" in the numerator is accounting for applying the rule with a bias term b; for homogeneous separators, the number of mistakes is at most $\frac{r^2}{\gamma^2}$. The claim for the non-homogeneous case is obtained by reduction to the homogeneous case, increasing the dimension by 1 and letting each x have coordinate $d+1$ fixed to 1, in which case the radius of the data in $d+1$ dimensions is at most $\sqrt{r^2+1}$, and the margin of the data with respect to non-homogeneous separators is precisely the margin of the $d+1$ dimensional augmented data with respect to homogeneous separators.

We prove the following result for this algorithm (where $r(S)$ and $\gamma(S)$ are as defined above). The proof is included in Section 4.3.

Theorem 3 *For any distribution P, any $n \in \mathbb{N}$, and any $\delta \in (0,1)$, for $S \sim P^n$, with probability at least $1-\delta$, if S is linearly separable, then letting $r = r(S)$ and $\gamma = \gamma(S)$, the classifier $\hat{h}_p = \mathcal{A}_p(S)$ satisfies*

$$R_P(\hat{h}_p) = O\left(\frac{r^2}{\gamma^2}\frac{1}{n} + \frac{1}{n}\log\left(\frac{1}{\delta}\right)\right).$$

We note that this result is stronger than the analogous PAC bounds known for using Perceptron with other well-known online-to-batch conversion techniques, such as the "longest survivor" technique (Kearns, Li, Pitt, and Valiant, 1987; Gallant, 1990) or voted Perceptron (Freund and Schapire, 1999). It also matches (up to constants) the result of Littlestone (1989), which was for a considerably more-involved online-to-batch conversion technique, which keeps all of the intermediate hypotheses and in the end selects one using a held-out portion of the data.

2.3. The Probability in the Region of Disagreement of a Version Space

Fix any measurable class of functions: $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$. For any $n \in \mathbb{N}$ and $S \in (\mathcal{X} \times \mathcal{Y})^n$, define $\mathcal{H}[S] = \{h \in \mathcal{H} : \hat{R}_S(h) = 0\}$. Fix any distribution P on $\mathcal{X} \times \mathcal{Y}$, let $(X_1, Y_1), (X_2, Y_2), \ldots$ be i.i.d. with distribution P, and for any $n \in \mathbb{N}$ let $S_n = \{(X_i, Y_i)\}_{i=1}^n$. Define the *version space* $V_n = \mathcal{H}[S_n]$ (Mitchell, 1977), and define its *region of disagreement* $\mathrm{DIS}(V_n) = \{x \in \mathcal{X} : \exists h, h' \in V_n, h(x) \neq h'(x)\}$ (Cohn, Atlas, and Ladner, 1994; Balcan, Beygelzimer, and Langford, 2006; Hanneke, 2014). The set $\mathrm{DIS}(V_n)$ plays an important role in certain disagreement-based active learning algorithms, and the analysis thereof: most-notably, the *CAL* active learning algorithm introduced by Cohn, Atlas, and Ladner (1994) and studied theoretically in great detail by a large number of works (e.g., Hanneke, 2007b, 2009, 2011, 2012, 2014, 2016; Balcan, Even-Dar, Hanneke, Kearns, Mansour, and Wortman, 2007; Balcan, Hanneke, and Vaughan, 2010; Hsu, 2010; El-Yaniv and Wiener, 2010, 2012; Wiener, Hanneke, and El-Yaniv, 2015; Hanneke and Yang, 2015). Of particular interest in quantifying the *label complexity* of CAL is bounding $P_X(\mathrm{DIS}(V_n))$ as a function of n. Classic well-known bounds on this quantity were established by Hanneke (2007b, 2009, 2011) based on a quantity known as the *disagreement coefficient* (see also Hanneke, 2014). However, more-recently Wiener, Hanneke, and El-Yaniv (2015) and Hanneke (2016) established bounds that are sometimes tighter, based on a quantity $\hat{t}_n$ called the *version space compression set size* (El-Yaniv and Wiener, 2010) (also known as the *empirical teaching dimension* in earlier work of Hanneke, 2007a). Specifically, define

$$\hat{t}_n = \min\{|S'| : S' \subseteq S_n, \mathcal{H}[S'] = V_n\}.$$

That is, $\hat{t}_n$ is the size of the smallest subset of S_n that induces the same version space.

Based on stable compression arguments, Bousquet, Hanneke, Moran, and Zhivotovskiy (2020b) prove a data-independent bound on $P_X(\mathrm{DIS}(V_n))$ in terms of a combinatorial quantity called the *star number* from Hanneke and Yang (2015), which improved the numerical constant factors compared to a result of the same form established by Hanneke (2016). However, they did not discuss data-dependent or distribution-dependent bounds on $P_X(\mathrm{DIS}(V_n))$. In particular, since $\hat{t}_n$ is never larger than the star number, and is often strictly smaller (e.g., the example in Remark 26 below), bounds on $P_X(\mathrm{DIS}(V_n))$ based on $\hat{t}_n$ may be considered stronger than bounds based on the star number. While prior works by Wiener, Hanneke, and El-Yaniv (2015) and Hanneke (2016) have established bounds on $P_X(\mathrm{DIS}(V_n))$ in terms of $\hat{t}_n$, both of these results have a suboptimal form (as we discuss in detail in Section 4.4). In the present work, we prove the following result, which improves over all of these previously known bounds on $P_X(\mathrm{DIS}(V_n))$ based on $\hat{t}_n$. The proof is presented in Section 4.4.

Theorem 4 *For any $n \in \mathbb{N}$ and $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$P_X(\mathrm{DIS}(V_n)) = O\left(\frac{1}{n}\left(\hat{t}_n + \log\left(\frac{1}{\delta}\right)\right)\right).$$

2.4. A New Data-dependent Bound for All ERM Algorithms

Continuing the notation from Section 2.3, we can also derive a new data-dependent PAC bound on the risk of any ERM learning algorithm expressed in terms of $\hat{t}_n$, which improves over the previous best known such bound from Hanneke (2016) (as we discuss in detail in Section 4.5). Specifically, we have the following result. The arguments underlying the result, and its relation to existing results in the literature, are discussed in Section 4.5.

Theorem 5 *Denote by d the VC dimension of $\mathcal{H}$ (Vapnik and Chervonenkis, 1971). Let P be any distribution such that $\inf_{h\in\mathcal{H}} R_P(h) = 0$. For any $n \in \mathbb{N}$ and $\delta \in (0,1)$, with probability at least $1-\delta$, every $h \in V_n$ satisfies*

$$R_P(h) = O\left(\frac{1}{n}\left(\mathrm{d}\log\left(\frac{\hat{t}_{\lfloor n/2\rfloor}}{\mathrm{d}}\right) + \log\left(\frac{1}{\delta}\right)\right)\right).$$

2.5. New Generalization Bounds for Compressed Nearest Neighbor Predictors

In nearest-neighbor classification, it is a classic fact that the 1-NN rule is not Bayes-consistent while k-NN is, where k grows appropriately with n (see Hanneke et al. (2019) for a detailed background). A recent line of work (Gottlieb et al., 2018; Kontorovich and Weiss, 2015) has shown that a margin-regularized 1-NN can be made Bayes-consistent, provided the margin is chosen via an SRM principle. Furthermore, the generalization bounds provided by this technique are compression-based and fully empirical. This line of research culminated in Hanneke et al. (2019), where an algorithm called OptiNet was presented and shown to be strongly universally Bayes-consistent in any metric space where *any* learner enjoys this property.

The OptiNet algorithm is described in detail in Hanneke et al. (2019). Briefly, for any choice of $\gamma > 0$, one constructs a γ-net on the sample — that is, a γ-separated set that is also a γ-cover. The following greedy algorithm constructs a γ-net in time $O(n^2)$ in any metric space; more efficient algorithms are known in doubling spaces (Krauthgamer and Lee, 2004; Gottlieb et al., 2014). Initialize the γ-net N_γ as the empty set. Traverse the datapoints in order, and if the ith point is not covered by the current partial net N_γ, it is appended to N_γ. It is easily verified that this construction indeed yields a γ-net, which will serve as the compression set. Furthermore, this compression scheme is stable: if any point not included in N_γ is omitted from the sample, the construction will yield the same N_γ.

Given the γ-net N_γ constructed as above, based on a given data set S, define a classifier $\hat{h}_\gamma = \mathcal{A}_\gamma(S)$ as follows: For any point $x \in \mathcal{X}$ let $\hat{x}$ be the element of N_γ nearest to x in the metric (breaking ties according to a measurable total order of the space $\mathcal{X}$; see Hanneke, Kontorovich, Sabato, and Weiss, 2019), and predict $\hat{h}_\gamma(x) = \hat{y}$, where $\hat{y}$ is the majority label among all $(x', y') \in S$ for which $\hat{x}$ is also the nearest element to x' in the γ-net N_γ. For this classifier, we have the following result (whose proof is deferred to Section 4.6):

Theorem 6 *Fix any $\gamma > 0$. For any distribution P, any $n \in \mathbb{N}$, and any $\delta \in (0,1)$, for $S \sim P^n$, with probability at least $1-\delta$, the classifier $\hat{h}_\gamma = \mathcal{A}_\gamma(S)$ satisfies*

$$\left| R_P(\hat{h}_\gamma) - \hat{R}_S(\hat{h}_\gamma) \right| = O\left(\sqrt{\hat{R}_S(\hat{h}_\gamma) \frac{1}{n} \left(|N_\gamma| + \log\left(\frac{1}{\delta}\right) \right)} + \frac{1}{n} \left(|N_\gamma| + \log\left(\frac{1}{\delta}\right) \right) \right).$$

This refines a data-dependent bound used by Hanneke, Kontorovich, Sabato, and Weiss (2019), which contained an additional log factor, and was based on a significantly more-involved argument needed in order to maintain permutation-invariance of certain subsets of the arguments to the reconstruction function used there (since otherwise the log factor would be of the form $\log(n)$, rather than $\log(n/|N_\gamma|)$, which was important for the proof of universal consistency in that work). In contrast, since the above bound is proven via establishing that $\hat{h}_n$ agrees with a *stable* compression scheme, we do not need to worry about the fact that the corresponding reconstruction function may be order-dependent, since there is no log factor to be concerned with. Hanneke, Kontorovich, Sabato, and Weiss (2019) used their bound on $R_P(\hat{h}_\gamma)$ in a procedure, called OptiNet, which optimizes the bound over the choice of γ; they show that doing so yields a universally strongly Bayes-consistent learning algorithm, in all metric spaces where Bayes-consistent learning is possible. Based on the above refinement, we could instead substitute this new tighter bound

$$\hat{R}_S(\hat{h}_\gamma) + O\left(\sqrt{\hat{R}_S(\hat{h}_\gamma) \frac{1}{n} \left(|N_\gamma| + \log\left(\frac{1}{\delta}\right) \right)} + \frac{1}{n} \left(|N_\gamma| + \log\left(\frac{1}{\delta}\right) \right) \right)$$

in the optimization in OptiNet, and the universal consistency result would still hold.

3. Definitions and Theorems for Stable Compression Schemes

The following notion was introduced by Littlestone and Warmuth (1986); Floyd and Warmuth (1995).

Definition 7 *A* compression scheme (κ, ρ) *consists of a* compression function κ*, which maps any $S \in (\mathcal{X} \times \mathcal{Y})^*$ to a subsequence $\kappa(S) \subseteq S$, and a* reconstruction function $\rho : (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{Y}^{\mathcal{X}}$ *mapping any $S \in (\mathcal{X} \times \mathcal{Y})^*$ to a function $\rho(S) : \mathcal{X} \to \mathcal{Y}$.*

For our purposes below, to be a valid compression scheme (κ, ρ), we also require that the function $(S, x) \mapsto \rho(\kappa(S))(x)$ mapping $(\mathcal{X} \times \mathcal{Y})^n \times \mathcal{X} \to \mathcal{Y}$ be a measurable function, for every $n \in \mathbb{N} \cup \{0\}$. For $k \in \mathbb{N} \cup \{0\}$, we say that a compression scheme (κ, ρ) has *size* k if every $S \in (\mathcal{X} \times \mathcal{Y})^*$ has $|\kappa(S)| \leq k$. The following definition is from (Bousquet, Hanneke, Moran, and Zhivotovskiy, 2020a) (previously also studied by Vapnik and Chervonenkis, 1974; Zhivotovskiy, 2017).

Definition 8 *A compression scheme (κ, ρ) is called* stable *if, for any $S \in (\mathcal{X} \times \mathcal{Y})^*$, $\forall S' \subseteq S \setminus \kappa(S)$, it holds that $\rho(\kappa(S \setminus S')) = \rho(\kappa(S))$.*

3.1. Results for Sample-Consistent Stable Compression Schemes

In the literature on sample compression schemes, considerable attention has been given to the special case when the compression scheme is *sample-consistent*, meaning that for a data set S, it holds that $\hat{R}_S(\rho(\kappa(S))) = 0$. For general sample compression schemes of size k, the best general result for the sample-consistent case is due to Littlestone and Warmuth (1986); Floyd and Warmuth (1995), stating that for $S \sim P^n$, with probability at least $1 - \delta$, if $\hat{R}_S(\rho(\kappa(S))) = 0$ then

$$R_P(\rho(\kappa(S))) = O\left(\frac{1}{n}\left(k \log(n) + \log\left(\frac{1}{\delta}\right)\right)\right),$$

with a slight improvement to

$$R_P(\rho(\kappa(S))) = O\left(\frac{1}{n}\left(k \log\left(\frac{n}{k}\right) + \log\left(\frac{1}{\delta}\right)\right)\right)$$

in the case that the reconstruction function ρ is permutation-invariant. Floyd and Warmuth (1995) also showed that the above bounds are sharp, in that there exist compression schemes for which, for certain distributions, one can prove lower bounds matching the above inequalities up to numerical constant factors.

However, in the special case of *stable* compression schemes, Bousquet, Hanneke, Moran, and Zhivotovskiy (2020a) proved that the log factor in the above inequalities is superfluous. Specifically, they established the following theorem.[1]

Theorem 9 *Let $k \in \mathbb{N} \cup \{0\}$ and let (κ, ρ) be any stable compression scheme of size k. For any distribution P, any integer $n > 2k$, and any $\delta \in (0, 1)$, letting $S \sim P^n$, with probability at least $1 - \delta$, if $\hat{R}_S(\rho(\kappa(S))) = 0$, then*

$$R_P(\rho(\kappa(S)) \leq \frac{2}{n - 2k}\left(k \ln(4) + \ln\left(\frac{1}{\delta}\right)\right).$$

Bousquet, Hanneke, Moran, and Zhivotovskiy (2020a) stated several implications of this result for compression schemes of known bounded size k. For instance, their result established, for the first time, that the support vector machine achieves a (minimax-optimal) generalization bound $R_P(\hat{h}_{\mathrm{SVM}}) = O\left(\frac{1}{n}\left(d + \log\left(\frac{1}{\delta}\right)\right)\right)$ for learning linear separators on $\mathbb{R}^d$ in the realizable case, based on the fact that it can be expressed as a stable compression scheme of size $d + 1$.

1. The version stated here is slightly more general than the original result of Bousquet, Hanneke, Moran, and Zhivotovskiy (2020a), as required for our proofs, so for completeness we include a proof in Appendix A.

As one part of the present work, we explore further implications of this result, focusing on *data-dependent* generalization bounds. For this purpose, we will use the following easy extension of Theorem 9 holding for data-dependent compression set sizes.

Theorem 10 *Let (κ, ρ) be any stable compression scheme. For any distribution P, any $n \in \mathbb{N}$, and any $\delta \in (0,1)$, for $S \sim P^n$, with probability at least $1-\delta$, if $\hat{R}_S(\rho(\kappa(S))) = 0$ and $|\kappa(S)| < n/2$, then*

$$R_P(\rho(\kappa(S))) \leq \frac{2}{n - 2|\kappa(S)|}\left(|\kappa(S)|\ln(4) + \ln\left(\frac{(|\kappa(S)|+1)(|\kappa(S)|+2)}{\delta}\right)\right).$$

Proof For each $k \in \mathbb{N} \cup \{0\}$, let (κ_k, ρ_k) be a compression scheme such that, for any S, if $|\kappa(S)| \leq k$, then $\kappa_k(S) = \kappa(S)$, and otherwise $\kappa_k(S) = \emptyset$; in any case, $\rho_k = \rho$. In particular, note that $|\kappa_k(S)| \leq k$ always. For each k, Theorem 9 implies that, with probability at least $1 - \frac{\delta}{(k+1)(k+2)}$, if $\hat{R}_S(\rho_k(\kappa_k(S))) = 0$, then

$$R_P(\rho_k(\kappa_k(S))) \leq \frac{2}{n-2k}\left(k\ln(4) + \ln\left(\frac{(k+1)(k+2)}{\delta}\right)\right).$$

By the union bound, the above claim holds simultaneously for all $k \in \mathbb{N} \cup \{0\}$ with probability at least $1 - \sum_k \frac{\delta}{(k+1)(k+2)} = 1 - \delta$. Finally, note that there necessarily exists *some* $k \in \mathbb{N} \cup \{0\}$ for which $|\kappa(S)| = k$, in which case $\rho(\kappa(S)) = \rho_k(\kappa_k(S))$ for this k. The theorem follows immediately from this. ■

In particular, the following corollary is immediate, by relaxing the expression in the above theorem into a simpler form.

Corollary 11 *Let (κ, ρ) be any stable compression scheme. For any distribution P, any $n \in \mathbb{N}$, and any $\delta \in (0,1)$, for $S \sim P^n$, with probability at least $1-\delta$, if $\hat{R}_S(\rho(\kappa(S))) = 0$ then*

$$R_P(\rho(\kappa(S))) \leq \frac{4}{n}\left(6|\kappa(S)| + \ln\left(\frac{e}{\delta}\right)\right).$$

Proof Since $\ln(x) < \sqrt{x}/2$ for $x \geq 3$, we have $\ln\big((|\kappa(S)|+2)^2\big) < (1/2)|\kappa(S)| + 1$, so that the right hand side of the inequality in Theorem 10 is at most $\frac{2}{n-2|\kappa(S)|}\left(|\kappa(S)|((1/2)+\ln(4)) + \ln\left(\frac{e}{\delta}\right)\right)$. Furthermore, since this is greater than 1 if $|\kappa(S)| > n/(3+2\ln(4))$, and $R_P(\rho(\kappa(S))) \leq 1$ always, any upper bound on this expression nondecreasing in $|\kappa(S)|$ and holding for all $|\kappa(S)| \leq n/(3 + 2\ln(4))$ is a valid bound on $R_P(\rho(\kappa(S)))$. In particular, for $|\kappa(S)| \leq n/(3+2\ln(4))$, it holds that $n - 2|\kappa(S)| \geq \left(1 - \frac{2}{3+2\ln(4)}\right)n$, which implies a bound of $\frac{1}{n}\left(4\ln(16e^3)|\kappa(S)| + \frac{2\ln(16e^3)}{\ln(16e)}\ln\left(\frac{e}{\delta}\right)\right)$. The stated bound follows by noting $\frac{2\ln(16e^3)}{\ln(16e)} < 4$ and $4\ln(16e^3) < 24$. ■

3.2. Results for Agnostic Stable Compression Schemes

The best known bounds for the general agnostic setting, holding for any compression scheme, are from Graepel, Herbrich, and Shawe-Taylor (2005). Specifically, for any compression scheme of size k, they show a bound (holding with probability at least $1-\delta$)

$$\left|R_P(\rho(\kappa(S))) - \hat{R}_S(\rho(\kappa(S)))\right| = O\left(\sqrt{\frac{1}{n}\left(k\log(n) + \log\left(\frac{1}{\delta}\right)\right)}\right),$$

or a slightly tighter bound

$$\left|R_P(\rho(\kappa(S))) - \hat{R}_S(\rho(\kappa(S)))\right| = O\left(\sqrt{\frac{1}{n}\left(k\log\left(\frac{n}{k}\right) + \log\left(\frac{1}{\delta}\right)\right)}\right)$$

in the case of permutation-invariant reconstruction function ρ. It was shown by Hanneke and Kontorovich (2019b) that both of these bounds are generally sharp: that is, for any k, n, there exist compression schemes of size k, and distributions P, such that a lower bound holds which matches the above up to numerical constants.

Here we show that if the compression scheme is *stable*, the log factor in the above bounds can be removed, analogous to the result of Bousquet, Hanneke, Moran, and Zhivotovskiy (2020a) for the realizable case.

Theorem 12 *For any $k \in \mathbb{N} \cup \{0\}$, let (κ, ρ) be any stable compression scheme of size k. For any distribution P, any $n \in \mathbb{N}$ with $n > 2k$, and any $\delta \in (0,1)$, for $S \sim P^n$, with probability at least $1-\delta$,*

$$\left|R_P(\rho(\kappa(S))) - \hat{R}_S(\rho(\kappa(S)))\right| \leq \sqrt{\frac{4}{n-2k}\left(k\ln(4) + \ln\left(\frac{4}{\delta}\right)\right)}.$$

Proof For brevity, define $[m] = \{1, \ldots, m\}$ for any $m \in \mathbb{N}$. The proof partly follows an argument from the original proof of Theorem 9 by Bousquet et al. (2020a), but with some important modifications to account for the fact that $\hat{R}_S(\rho(\kappa(S)))$ may be nonzero.

If $k = 0$, the result trivially follows from Hoeffding's inequality, so let us suppose $k \geq 1$. As in the proof of Bousquet et al. (2020a), fix any $T_n \in [n-1]$ and let $\mathcal{I}_n$ be any family of subsets of $[n]$ with the properties that each $I \in \mathcal{I}_n$ has $|I| \leq n - T_n$, and for every $i_1, \ldots, i_k \in [n]$ there exists $I \in \mathcal{I}_n$ with $\{i_1, \ldots, i_k\} \subseteq I$.

In particular, Bousquet et al. (2020a) construct a family $\mathcal{I}_n$ satisfying the properties above with $T_n = k\lfloor n/(2k)\rfloor$, and with $|\mathcal{I}_n| = \binom{2k}{k} < 4^k$: namely, let $D_1, \ldots, D_{2k}$ be any partition of $[n]$ with each $|D_i| \in \{\lfloor n/(2k)\rfloor, \lceil n/(2k)\rceil\}$, and define $\mathcal{I}_n = \{\bigcup\{D_j : j \in \mathcal{J}\} : \mathcal{J} \subseteq [2k], |\mathcal{J}| = k\}$; that is, $\mathcal{I}_n$ contains all unions of exactly k of the $2k$ sets D_j.

Let $S = \{(X_i, Y_i)\}_{i=1}^n \sim P^n$, and for any $I \subseteq [n]$ define $S_I = \{(X_i, Y_i) : i \in I\}$. As a new component needed in the present proof, let $\sigma : [n] \to [n]$ be a uniform random permutation of $[n]$; this will only become important in the second half of the proof below.

For any $I \subset [n]$, since $S_{[n]\setminus I}$ is independent of S_I, Hoeffding's inequality (applied under the conditional distribution given S_I) and the law of total probability imply that, with probability at least $1 - \frac{\delta}{2|\mathcal{I}_n|}$,

$$\left|R_P(\rho(\kappa(S_I))) - \hat{R}_{S_{[n]\setminus I}}(\rho(\kappa(S_I)))\right| \leq \sqrt{\frac{\ln(4|\mathcal{I}_n|/\delta)}{2(n-|I|)}}.$$

Applying this under the conditional distribution given σ, together with the union bound and the law of total probability, we have that with probability at least $1-\frac{\delta}{2}$, every $I \in \mathcal{I}_n$ has

$$\left| R_P(\rho(\kappa(S_{\sigma^{-1}(I)}))) - \hat{R}_{S_{[n]\setminus\sigma^{-1}(I)}}(\rho(\kappa(S_{\sigma^{-1}(I)}))) \right| \leq \sqrt{\frac{\ln(4|\mathcal{I}_n|/\delta)}{2(n-|I|)}}.$$

In particular, letting $i_1^*, \ldots, i_{|\kappa(S)|}^*$ be the $|\kappa(S)|$ indices such that $\kappa(S) = \{(X_{i_j^*}, Y_{i_j^*})\}_{j=1}^{|\kappa(S)|}$, the defining properties of $\mathcal{I}_n$ imply that there exists $I^* \in \mathcal{I}_n$ with $\{\sigma(i_1^*), \ldots, \sigma(i_{|\kappa(S)|}^*)\} \subseteq I^*$. Since (κ, ρ) is a *stable* compression scheme, this also implies $\rho(\kappa(S_{\sigma^{-1}(I^*)})) = \rho(\kappa(S))$. Furthermore, by the defining properties of $\mathcal{I}_n$, we have $n-|I^*| \geq T_n$. Therefore, on the above event of probability at least $1-\frac{\delta}{2}$,

$$\left| R_P(\rho(\kappa(S))) - \hat{R}_{S_{[n]\setminus\sigma^{-1}(I^*)}}(\rho(\kappa(S))) \right| \leq \sqrt{\frac{\ln(4|\mathcal{I}_n|/\delta)}{2T_n}}. \tag{1}$$

Now, unlike the original proof of Bousquet et al. (2020a), to complete the present proof we must still relate $\hat{R}_{S_{[n]\setminus\sigma^{-1}(I^*)}}(\rho(\kappa(S)))$ to $\hat{R}_S(\rho(\kappa(S)))$. This is where the random permutation σ becomes important, as it enables us to introduce a concentration argument which accounts for the possibility that $\rho(\kappa(\cdot))$ may be order-dependent in its argument. Let $\hat{h} = \rho(\kappa(S))$. For each $i \in [n]$, let $\ell_i = \mathbb{1}[\hat{h}(X_i) \neq Y_i]$. For any $I \in \mathcal{I}_n$, by Hoeffding's inequality (for sampling without replacement; see Hoeffding, 1963) applied under the conditional distribution given S, together with the law of total probability, with probability at least $1-\frac{\delta}{2|\mathcal{I}_n|}$, it holds that

$$\left| \frac{1}{n-|I|} \sum_{i \in [n]\setminus\sigma^{-1}(I)} \ell_i - \hat{R}_S(\rho(\kappa(S))) \right| \leq \sqrt{\frac{\ln(4|\mathcal{I}_n|/\delta)}{2(n-|I|)}}.$$

By the union bound, this holds simultaneously for all $I \in \mathcal{I}_n$ with probability at least $1-\frac{\delta}{2}$. In particular, taking $I = I^*$, and recalling that $n-|I^*| \geq T_n$, on this event we have that

$$\left| \hat{R}_{S_{[n]\setminus\sigma^{-1}(I^*)}}(\rho(\kappa(S))) - \hat{R}_S(\rho(\kappa(S))) \right| \leq \sqrt{\frac{\ln(4|\mathcal{I}_n|/\delta)}{2T_n}}. \tag{2}$$

By the union bound, the above two events (each of probability at least $1-\frac{\delta}{2}$) hold simultaneously with probability at least $1-\delta$, in which case (1) and (2) together imply

$$\begin{aligned}
&\left| R_P(\rho(\kappa(S))) - \hat{R}_S(\rho(\kappa(S))) \right| \\
&\leq \left| R_P(\rho(\kappa(S))) - \hat{R}_{S_{[n]\setminus\sigma^{-1}(I^*)}}(\rho(\kappa(S))) \right| + \left| \hat{R}_{S_{[n]\setminus\sigma^{-1}(I^*)}}(\rho(\kappa(S))) - \hat{R}_S(\rho(\kappa(S))) \right| \\
&\leq \sqrt{\frac{2\ln(4|\mathcal{I}_n|/\delta)}{T_n}}.
\end{aligned}$$

The theorem now immediately follows by plugging in the aforementioned family $\mathcal{I}_n$ from Bousquet et al. (2020a), having $|\mathcal{I}_n| = \binom{2k}{k} < 4^k$ and $T_n = k\lfloor n/(2k) \rfloor > \frac{n-2k}{2}$. ■

As above, this easily extends to data-dependent compression sizes, as stated in the following theorem. The proof follows the same argument as in the proof of Theorem 10, and so we omit the details.

Theorem 13 *Let (κ, ρ) be any stable compression scheme. For any distribution P, any $n \in \mathbb{N}$, and any $\delta \in (0,1)$, for $S \sim P^n$, with probability at least $1-\delta$, if $|\kappa(S)| < n/2$, then*

$$\left|R_P(\rho(\kappa(S))) - \hat{R}_S(\rho(\kappa(S)))\right| \leq \sqrt{\frac{4}{n-2|\kappa(S)|}\left(|\kappa(S)|\ln(4) + \ln\left(\frac{4(|\kappa(S)|+1)(|\kappa(S)|+2)}{\delta}\right)\right)}.$$

Also analogous to the results for the realizable case, the above bound can be further relaxed into a simple expression, as follows. The proof is nearly identical to that of Corollary 11, and so we omit it for brevity.

Corollary 14 *Let (κ, ρ) be any stable compression scheme. For any distribution P, any $n \in \mathbb{N}$, and any $\delta \in (0,1)$, for $S \sim P^n$, with probability at least $1-\delta$, then*

$$\left|R_P(\rho(\kappa(S))) - \hat{R}_S(\rho(\kappa(S)))\right| \leq \sqrt{\frac{8}{n}\left(6|\kappa(S)| + \ln\left(\frac{4e}{\delta}\right)\right)}.$$

While the bound of Theorem 12 holds for $1-\delta$ fraction of data sets from any distribution, and is therefore more general than Theorem 9 (which restricts to the sample-consistent case), the bound is not as tight in the specific case where Theorem 9 applies. As such, it is desirable to also state a bound which interpolates between the two: that is, which does not require the compression scheme to be sample-consistent to provide a non-trivial bound, but yet is able to recover the form of the bound in Theorem 9 in the case where it happens to be sample-consistent. We provide such a result in the following theorem.

Theorem 15 *For any $k \in \mathbb{N} \cup \{0\}$, let (κ, ρ) be any stable compression scheme of size k. For any distribution P, any $n \in \mathbb{N}$ with $n > 4k$, and any $\delta \in (0,1)$, for $S \sim P^n$, with probability at least $1-\delta$,*

$$\begin{aligned}&\left|R_P(\rho(\kappa(S))) - \hat{R}_S(\rho(\kappa(S)))\right| \\ &\leq \sqrt{\hat{R}_S(\rho(\kappa(S)))\frac{72}{n}\left(k\ln(4) + \ln\left(\frac{4}{\delta}\right)\right)} + \frac{32}{n}\left(k\ln(4) + \ln\left(\frac{4}{\delta}\right)\right).\end{aligned}$$

Before stating the proof, we first recall the following so-called "ratio-type" inequality, based on Bernstein's inequality.

Lemma 16 *For any $n \in \mathbb{N}$, we consider two cases simultaneously: (i) let $p \in [0,1]$ and let $Z_1, \ldots, Z_n$ be i.i.d.* Bernoulli(p) *random variables, (ii) let $t \geq n$, $\{B_1, \ldots, B_t\} \in \{0,1\}^t$, $p = \frac{1}{t}\sum_{i=1}^t B_i$, and let $Z_1, \ldots, Z_n$ be random variables sampled uniformly without replacement from $\{B_1, \ldots, B_t\}$. In either case, for any $\delta \in (0,1)$, defining $\bar{Z} = \frac{1}{n}\sum_{i=1}^n Z_i$, with probability at least $1-\delta$,*

$$\left|\bar{Z} - p\right| \leq \sqrt{\min\{2\bar{Z}, p\}\frac{2}{n}\ln\left(\frac{2}{\delta}\right)} + \frac{4}{n}\ln\left(\frac{2}{\delta}\right).$$

Proof In both cases covered by the claim, Bernstein's inequality implies that

$$\Pr\left(\left|\bar{Z} - p\right| > \varepsilon\right) \le 2\exp\left\{-\frac{(1/2)\varepsilon^2 n}{p + (\varepsilon/3)}\right\}.$$

Setting the right hand side equal δ and solving for ε yields that, with probability at least $1-\delta$,

$$\left|\bar{Z} - p\right| \le \sqrt{p\frac{2}{n}\ln\left(\frac{2}{\delta}\right) + \frac{1}{9n^2}\ln^2\left(\frac{2}{\delta}\right)} + \frac{1}{3n}\ln\left(\frac{2}{\delta}\right).$$

In particular, relaxing the right hand side above implies that, on this event,

$$\left|\bar{Z} - p\right| \le \sqrt{p\frac{2}{n}\ln\left(\frac{2}{\delta}\right)} + \frac{2}{3n}\ln\left(\frac{2}{\delta}\right). \tag{3}$$

Furthermore, for any non-negative values A, B, C, it holds that $A \le B + C\sqrt{A} \Rightarrow A \le B + C^2 + \sqrt{B}C$. Therefore, on the above event,

$$p \le \bar{Z} + \frac{8}{3n}\ln\left(\frac{2}{\delta}\right) + \sqrt{\bar{Z} + \frac{2}{3n}\ln\left(\frac{2}{\delta}\right)}\sqrt{\frac{2}{n}\ln\left(\frac{2}{\delta}\right)} \le 2\bar{Z} + \frac{16}{3n}\ln\left(\frac{2}{\delta}\right).$$

Plugging back into (3) yields that, on this same event,

$$\left|\bar{Z} - p\right| \le \sqrt{2\bar{Z}\frac{2}{n}\ln\left(\frac{2}{\delta}\right)} + \left(\sqrt{\frac{8}{3}} + \frac{1}{3}\right)\frac{2}{n}\ln\left(\frac{2}{\delta}\right) \le \sqrt{2\bar{Z}\frac{2}{n}\ln\left(\frac{2}{\delta}\right)} + \frac{4}{n}\ln\left(\frac{2}{\delta}\right).$$

This inequality and (3) together imply the claimed bound. ■

We are now ready for the proof of Theorem 15.

Proof of Theorem 15 This proof follows essentially similar arguments to the proof of Theorem 12, except using Lemma 16 in place of Hoeffding's inequality in both places in the proof where such inequalities are used. Let $\mathcal{I}_n$ and T_n be as in the proof of Theorem 12, and let $[m] = \{1, \ldots, m\}$ for any $m \in \mathbb{N}$.

If $k = 0$, the result trivially follows from Lemma 16, so let us suppose $k \ge 1$. Let $S = \{(X_i, Y_i)\}_{i=1}^n \sim P^n$, and for any $I \subseteq [n]$ define $S_I = \{(X_i, Y_i) : i \in I\}$. As in Theorem 12, let $\sigma : [n] \to [n]$ be a uniform random permutation of $[n]$.

For any $I \subset [n]$, since $S_{[n]\setminus I}$ is independent of S_I, Lemma 16 (applied under the conditional distribution given S_I) and the law of total probability imply that, with probability at least $1 - \frac{\delta}{2|\mathcal{I}_n|}$,

$$\left|R_P(\rho(\kappa(S_I))) - \hat{R}_{S_{[n]\setminus I}}(\rho(\kappa(S_I)))\right| \le \sqrt{\hat{R}_{S_{[n]\setminus I}}(\rho(\kappa(S_I)))\frac{4}{n-|I|}\ln\left(\frac{4|\mathcal{I}_n|}{\delta}\right)} + \frac{4}{n-|I|}\ln\left(\frac{4|\mathcal{I}_n|}{\delta}\right).$$

Applying this under the conditional distribution given σ, together with the union bound and the law of total probability, we have that with probability at least $1 - \frac{\delta}{2}$, every $I \in \mathcal{I}_n$ has

$$\begin{aligned}&\left|R_P(\rho(\kappa(S_{\sigma^{-1}(I)}))) - \hat{R}_{S_{[n]\setminus\sigma^{-1}(I)}}(\rho(\kappa(S_{\sigma^{-1}(I)})))\right| \\ &\le \sqrt{\hat{R}_{S_{[n]\setminus\sigma^{-1}(I)}}(\rho(\kappa(S_{\sigma^{-1}(I)})))\frac{4}{n-|I|}\ln\left(\frac{4|\mathcal{I}_n|}{\delta}\right)} + \frac{4}{n-|I|}\ln\left(\frac{4|\mathcal{I}_n|}{\delta}\right).\end{aligned}$$

In particular, letting $i_1^*, \ldots, i_{|\kappa(S)|}^*$ be the $|\kappa(S)|$ indices such that $\kappa(S) = \{(X_{i_j^*}, Y_{i_j^*})\}_{j=1}^{|\kappa(S)|}$, the defining properties of $\mathcal{I}_n$ imply that there exists $I^* \in \mathcal{I}_n$ with $\{\sigma(i_1^*), \ldots, \sigma(i_{|\kappa(S)|}^*)\} \subseteq I^*$. Since (κ, ρ) is a *stable* compression scheme, this also implies $\rho(\kappa(S_{\sigma^{-1}(I^*)})) = \rho(\kappa(S))$. Furthermore, by the defining properties of $\mathcal{I}_n$, we have $n - |I^*| \geq T_n$. Also note that $\hat{R}_{S_{[n]\setminus\sigma^{-1}(I^*)}}(\rho(\kappa(S))) \leq \frac{n}{T_n}\hat{R}_S(\rho(\kappa(S)))$. Therefore, on the above event of probability at least $1 - \frac{\delta}{2}$,

$$\left| R_P(\rho(\kappa(S))) - \hat{R}_{S_{[n]\setminus\sigma^{-1}(I^*)}}(\rho(\kappa(S))) \right| \leq \sqrt{\frac{n}{T_n}\hat{R}_S(\rho(\kappa(S)))\frac{4}{T_n}\ln\left(\frac{4|\mathcal{I}_n|}{\delta}\right)} + \frac{4}{T_n}\ln\left(\frac{4|\mathcal{I}_n|}{\delta}\right). \tag{4}$$

Let $\hat{h} = \rho(\kappa(S))$. For each $i \in [n]$, let $\ell_i = \mathbb{1}[\hat{h}(X_i) \neq Y_i]$. For any $I \in \mathcal{I}_n$, by Lemma 16 (the case holding for sampling without replacement) applied under the conditional distribution given S, together with the law of total probability, with probability at least $1 - \frac{\delta}{2|\mathcal{I}_n|}$, it holds that

$$\left| \frac{1}{n-|I|} \sum_{i \in [n]\setminus\sigma^{-1}(I)} \ell_i - \hat{R}_S(\rho(\kappa(S))) \right| \leq \sqrt{\hat{R}_S(\rho(\kappa(S)))\frac{2}{n-|I|}\ln\left(\frac{4|\mathcal{I}_n|}{\delta}\right)} + \frac{4}{n-|I|}\ln\left(\frac{4|\mathcal{I}_n|}{\delta}\right).$$

By the union bound, this holds simultaneously for all $I \in \mathcal{I}_n$ with probability at least $1 - \frac{\delta}{2}$. In particular, taking $I = I^*$, and recalling that $n - |I^*| \geq T_n$, on this event we have that

$$\left| \hat{R}_{S_{[n]\setminus\sigma^{-1}(I^*)}}(\rho(\kappa(S))) - \hat{R}_S(\rho(\kappa(S))) \right| \leq \sqrt{\hat{R}_S(\rho(\kappa(S)))\frac{2}{T_n}\ln\left(\frac{4|\mathcal{I}_n|}{\delta}\right)} + \frac{4}{T_n}\ln\left(\frac{4|\mathcal{I}_n|}{\delta}\right). \tag{5}$$

By the union bound, the above two events (each of probability at least $1-\frac{\delta}{2}$) hold simultaneously with probability at least $1 - \delta$, in which case (4) and (5) together imply

$$\begin{aligned}
&\left| R_P(\rho(\kappa(S))) - \hat{R}_S(\rho(\kappa(S))) \right| \\
&\leq \left| R_P(\rho(\kappa(S))) - \hat{R}_{S_{[n]\setminus\sigma^{-1}(I^*)}}(\rho(\kappa(S))) \right| + \left| \hat{R}_{S_{[n]\setminus\sigma^{-1}(I^*)}}(\rho(\kappa(S))) - \hat{R}_S(\rho(\kappa(S))) \right| \\
&\leq \left(1 + \sqrt{\frac{2n}{T_n}}\right)\sqrt{\hat{R}_S(\rho(\kappa(S)))\frac{2}{T_n}\ln\left(\frac{4|\mathcal{I}_n|}{\delta}\right)} + \frac{8}{T_n}\ln\left(\frac{4|\mathcal{I}_n|}{\delta}\right).
\end{aligned}$$

The theorem now immediately follows by plugging in the family $\mathcal{I}_n$ from Bousquet, Hanneke, Moran, and Zhivotovskiy (2020a) (described in the proof of Theorem 12 above), having $|\mathcal{I}_n| = \binom{2k}{k} < 4^k$ and $T_n = k\lfloor n/(2k) \rfloor > \frac{n-2k}{2} > \frac{n}{4}$, and relaxing the numerical constants to simplify the expression. ■

As above, we can also easily extend this to data-dependent compression sizes, stated in the following theorem. The proof is nearly identical to the proof of Theorem 10 (except using Theorem 15 in place of Theorem 9) and so we omit the proof for brevity.

Theorem 17 *Let (κ, ρ) be any stable compression scheme. For any distribution P, any $n \in \mathbb{N}$, and any $\delta \in (0, 1)$, for $S \sim P^n$, with probability at least $1 - \delta$, if $|\kappa(S)| < n/4$, then*

$$\begin{aligned}
&\left| R_P(\rho(\kappa(S))) - \hat{R}_S(\rho(\kappa(S))) \right| \\
&\leq \sqrt{\hat{R}_S(\rho(\kappa(S))) \frac{72}{n} \left(|\kappa(S)| \ln(4) + \ln\!\left(\frac{4(|\kappa(S)|+1)(|\kappa(S)|+2)}{\delta} \right) \right)} \\
&\quad + \frac{32}{n} \left(|\kappa(S)| \ln(4) + \ln\!\left(\frac{4(|\kappa(S)|+1)(|\kappa(S)|+2)}{\delta} \right) \right).
\end{aligned}$$

Also as above, we can state a bound in a simpler form by relaxing the above inequality, as stated in the following corollary. The proof follows similar arguments as in the proof of Corollary 11, so we omit the proof for brevity.

Corollary 18 *Let (κ, ρ) be any stable compression scheme. For any distribution P, any $n \in \mathbb{N}$, and any $\delta \in (0, 1)$, for $S \sim P^n$, with probability at least $1 - \delta$,*

$$\begin{aligned}
&\left| R_P(\rho(\kappa(S))) - \hat{R}_S(\rho(\kappa(S))) \right| \\
&\leq \sqrt{\hat{R}_S(\rho(\kappa(S))) \frac{72}{n} \left(2|\kappa(S)| + \ln\!\left(\frac{4e}{\delta} \right) \right)} + \frac{32}{n} \left(2|\kappa(S)| + \ln\!\left(\frac{4e}{\delta} \right) \right).
\end{aligned}$$

4. Details of the Applications

This section provides the proofs and discussions related to the various main results from Section 2.

4.1. Proof of the Optimal PAC Margin Bound for SVM

For the SVM algorithm, under linearly separable distributions P, an in-expectation margin bound was established very early by Vapnik and Chervonenkis (1974); Vapnik and Chapelle (2000): namely, $\mathbb{E}[R_P(\mathrm{SVM}(S_n))] \leq \mathbb{E}\left[\frac{r(S_{n+1})^2}{\gamma(S_{n+1})^2} \frac{1}{n+1}\right]$, for $S_n \sim P^n$ and $S_{n+1} \sim P^{n+1}$. However, determining whether SVM obtains the optimal data-dependent *PAC* margin bound has remained a challenging open problem, with several sub-optimal bounds appearing in prior works in the literature, which include extra log factors (Shawe-Taylor, Bartlett, Williamson, and Anthony, 1998; Hanneke and Kontorovich, 2019b). We resolve this question here. Specifically, we prove the following result, from which Theorem 1 immediately follows.

Theorem 19 *For any distribution P, any $n \in \mathbb{N}$, and any $\delta \in (0, 1)$, for $S \sim P^n$, with probability at least $1 - \delta$, if S is linearly separable, then letting $\hat{h}_{\mathrm{SVM}} = \mathrm{SVM}(S)$, $r = r(S)$, and $\gamma = \gamma(S)$, we have*

$$R_P(\hat{h}_{\mathrm{SVM}}) \leq \frac{4}{n} \left(6 \frac{r^2}{\gamma^2} + \ln\!\left(\frac{e}{\delta} \right) \right)$$

and if $\frac{r^2}{\gamma^2} < n/2$, then

$$R_P(\hat{h}_{\mathrm{SVM}}) \leq \frac{2}{n - 2r^2/\gamma^2} \left(\frac{r^2}{\gamma^2} \ln(4) + 2 \ln\!\left(\frac{r^2}{\gamma^2} + 2 \right) + \ln\!\left(\frac{1}{\delta} \right) \right).$$

Proof It has been known since the initial work of Vapnik and Chervonenkis (1974) that SVM can be expressed as a compression scheme, where the compression points are the *support vectors*: that is, the samples receiving non-zero weight in the solution to the dual formulation of the optimization problem. However, the support vectors are not always uniquely defined, so that the size of the compression scheme depends on which solution is used. However, since the actual classifier $\hat{h}_{\rm SVM}$ does *not* depend on which solution we choose, we can analyze $R_P(\hat{h}_{\rm SVM})$ by identifying *any* complete set of support vectors of some desired number.

Following Vapnik and Chervonenkis (1974), in a given data set S, define an *essential support vector* as any $(x, y) \in S$ such that ${\rm SVM}(S \setminus \{(x, y)\}) \neq {\rm SVM}(S)$. The essential support vectors do not necessarily form a complete set of support vectors (indeed, there may be *no* essential support vectors in some cases). However, we can use a universal bound on the number of essential support vectors to identify a particular compression scheme of a desirable size, corresponding to ${\rm SVM}(S)$. Specifically, Vapnik and Chervonenkis (1974) showed that for any linearly separable data set S, there are at most $\frac{r(S)^2}{\gamma(S)^2}$ essential support vectors (see also Hanneke and Kontorovich, 2019a).

Now we describe a compression scheme (κ, ρ) with $|\kappa(S)| \leq \frac{r(S)^2}{\gamma(S)^2}$ for linearly separable data sets S, such that $\rho(\kappa(S)) = {\rm SVM}(S)$. Fix any $r, \gamma > 0$; we inductively construct a compression scheme $(\kappa_{r,\gamma}, \rho_{r,\gamma})$ that, for any data set S with $r(S) \leq r$ and $\gamma(S) \geq \gamma$, it holds that $|\kappa_{r,\gamma}(S)| \leq \frac{r^2}{\gamma^2}$ and $\rho_{r,\gamma}(\kappa_{r,\gamma}(S)) = {\rm SVM}(S)$. In particular, we will always define $\rho_{r,\gamma}(S) = {\rm SVM}(S)$, so that it remains only to define $\kappa_{r,\gamma}$. First, if $|S| \leq \frac{r^2}{\gamma^2}$, simply define $\kappa_{r,\gamma}(S) = S$, so that $\rho_{r,\gamma}(\kappa_{r,\gamma}(S)) = {\rm SVM}(S)$ trivially. This is our base case in the inductive construction. Next, take as an inductive hypothesis that S is a linearly separable set with $r(S) \leq r$, $\gamma(S) \geq \gamma$, and $|S| > \frac{r^2}{\gamma^2}$, and that every strict subset $S' \subset S$ has $\kappa_{r,\gamma}(S')$ defined, with $|\kappa_{r,\gamma}(S')| \leq \frac{r^2}{\gamma^2}$ and $\rho_{r,\gamma}(\kappa_{r,\gamma}(S')) = {\rm SVM}(S')$ (noting that $r(S') \leq r(S) \leq r$, and $\gamma(S') \geq \gamma(S) \geq \gamma$). Since $|S| > \frac{r^2}{\gamma^2} \geq \frac{r(S)^2}{\gamma(S)^2}$, the result above implies that S necessarily contains at least one point that is *not* an essential support vector (with respect to applying SVM to S). Let (x, y) be the first element of S (by their order in the sequence S) that is not an essential support vector, and define $\kappa_{r,\gamma}(S) = \kappa_{r,\gamma}(S \setminus \{(x, y)\})$. By the inductive hypothesis, $|\kappa_{r,\gamma}(S)| = |\kappa_{r,\gamma}(S \setminus \{(x, y)\})| \leq \frac{r^2}{\gamma^2}$, and $\rho_{r,\gamma}(\kappa_{r,\gamma}(S)) = \rho_{r,\gamma}(\kappa_{r,\gamma}(S \setminus \{(x, y)\})) = {\rm SVM}(S \setminus \{(x, y)\})$; moreover, since (x, y) is not an essential support vector, ${\rm SVM}(S \setminus \{(x, y)\}) = {\rm SVM}(S)$, so that we have confirmed that $\rho_{r,\gamma}(\kappa_{r,\gamma}(S)) = {\rm SVM}(S)$. By the principle of induction, we have constructed $(\kappa_{r,\gamma}, \rho_{r,\gamma})$ satisfying the claim for all linearly separable S with $r(S) \leq r$ and $\gamma(S) \geq \gamma$.

Now, for any linearly separable data set S, define $\kappa(S) = \kappa_{r(S),\gamma(S)}(S)$, and generally define $\rho(\kappa(S)) = {\rm SVM}(\kappa(S))$. By the above argument, every linearly separable set S has $|\kappa(S)| \leq \frac{r(S)^2}{\gamma(S)^2}$ and $\rho(\kappa(S)) = {\rm SVM}(S)$. Moreover, since ${\rm SVM}(\kappa(S)) = {\rm SVM}(S)$, any subset $S' \subset S$ with $\kappa(S) \subseteq S'$ must also have ${\rm SVM}(S') = {\rm SVM}(S)$ (since S' contains a complete set of support vectors with respect to applying SVM to S), so that the property of the construction of κ above implies $\rho(\kappa(S')) = {\rm SVM}(\kappa(S')) = {\rm SVM}(S') = {\rm SVM}(S) = {\rm SVM}(\kappa(S)) = \rho(\kappa(S))$. Thus, (κ, ρ) is also a *stable* compression scheme. The theorem now follows immediately from Theorem 10 and Corollary 11. ■

Remark 20 *We also note that this bound can be further refined by replacing $r(S)$ with the* span *of the data, defined by Vapnik and Chapelle (2000), as that work also established a bound on the* *number of essential support vectors in terms of the span of S, and the span is nonincreasing as we inductively remove data from S in the argument used in the above proof.*

4.2. A Data-dependent Online-to-Batch Conversion Bound

A result in Bousquet, Hanneke, Moran, and Zhivotovskiy (2020b) establishes a bound for online-to-batch conversion for conservative online learners with an *a priori* mistake bound. Specifically, from (Littlestone, 1988), an online learning algorithm $\mathcal{A}$ is a (measurable) map $(\mathcal{X} \times \mathcal{Y})^* \times \mathcal{X} \to \mathcal{Y}$. For a given concept class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ of functions, the *mistake bound* of $\mathcal{A}$ is defined as

$$M(\mathcal{A}, \mathcal{H}) = \sup_{x_1, x_2, \ldots \in \mathcal{X}} \sup_{h \in \mathcal{H}} \sum_{t=1}^{\infty} \mathbb{1}[\mathcal{A}(\{(x_i, h(x_i))\}_{i=1}^{t-1}, x_t) \neq h(x_t)].$$

In other words, $M(\mathcal{A}, \mathcal{H})$ is the largest number of mistakes the algorithm $\mathcal{A}$ will make on any sequence labeled according to some target concept in $\mathcal{H}$. It is known that the minimum possible value of $M(\mathcal{A}, \mathcal{H})$ is equal to the *Littlestone dimension* of $\mathcal{H}$, defined by Littlestone (1988). As a special type of algorithm of considerable interest, an online learning algorithm $\mathcal{A}$ is called *conservative* if the consecutive predictors $\mathcal{A}(\{(x_i, y_i)\}_{i=1}^{t-1}, \cdot)$ and $\mathcal{A}(\{(x_i, y_i)\}_{i=1}^{t}, \cdot)$ only differ when $\mathcal{A}(\{(x_i, y_i)\}_{i=1}^{t-1}, x_t) \neq y_t$: that is, the algorithm's hypothesis is only updated after each *mistake*. Formally, $\mathcal{A}$ is conservative if, for any n and $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \mathcal{Y}$, letting $\hat{m} = \sum_{t=1}^{n} \mathbb{1}[\mathcal{A}(\{(x_i, y_i)\}_{i=1}^{t-1}, x_t) \neq y_t]$ and denoting by $i_1, \ldots, i_{\hat{m}}$ the subsequence of $\{1, \ldots, n\}$ with $\mathcal{A}(\{(x_i, y_i)\}_{i=1}^{i_j - 1}, x_{i_j}) \neq y_{i_j}$, and letting $i_0 = 0$ and $i_{\hat{m}+1} = n+1$, for every $j \in \{0, \ldots, \hat{m}\}$ and every $t \in \{i_j + 1, \ldots, \min\{i_{j+1}, n\}\}$, $\mathcal{A}(\{(x_i, y_i)\}_{i=1}^{t-1}, x_t) = \mathcal{A}(\{(x_{i_{j'}}, y_{i_{j'}})\}_{j'=1}^{j}, x_t)$.

Bousquet, Hanneke, Moran, and Zhivotovskiy (2020b) propose a new PAC bound for conservative online learning algorithms. Specifically, for any given data set $S = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, and any conservative online learning algorithm, they consider running $\mathcal{A}$ through the data set S in order, and cycling through repeatedly until it makes a full pass through S without making any mistakes. Formally, letting $t_i = i - n\lfloor i/n \rfloor$ for each $i \in \mathbb{N}$, define $\hat{h}_n(\cdot) = \mathcal{A}(\{(X_{t_i}, Y_{t_i})\}_{i=1}^{T}, \cdot)$ for T the smallest positive integer multiple of n for which $\sum_{j=T-n}^{T-1} \mathbb{1}[\mathcal{A}(\{(X_{t_i}, Y_{t_i})\}_{i=1}^{j}, X_{t_{j+1}}) \neq Y_{t_{j+1}}] = 0$. If no such T exists, we will say $\hat{h}_n$ is undefined. They prove the following result for this $\hat{h}_n$, by noting that it can be viewed as a stable compression scheme.

Theorem 21 *(Bousquet, Hanneke, Moran, and Zhivotovskiy, 2020b) Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be any nonempty concept class of measurable functions, let $\mathcal{A}$ be any conservative online learning algorithm with $M(\mathcal{A}, \mathcal{H}) < \infty$, let P be any distribution on $\mathcal{X} \times \mathcal{Y}$ such that $\exists h^* \in \mathcal{H}$ with $R_P(h^*) = 0$, let $n \in \mathbb{N}$ with $n > 2M(\mathcal{A}, \mathcal{H})$, and let $S = \{(X_i, Y_i)\}_{i=1}^{n} \sim P^n$. For $\hat{h}_n$ as defined above, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$R_P(\hat{h}_n) \leq \frac{2}{n - 2M(\mathcal{A}, \mathcal{H})} \left(M(\mathcal{A}, \mathcal{H}) \ln(4) + \ln\left(\frac{1}{\delta}\right) \right).$$

The above result matches (up to constants) an online-to-batch conversion technique of Littlestone (1989), which was considerably more involved (requiring the learner to keep track of all intermediate hypotheses, and in the end select one of these using a held-out portion of the data). Also, the form of the bound in Theorem 21 is better than analogous PAC bounds known for other well-known online-to-batch conversion techniques, such as the "longest survivor" technique (Kearns, Li, Pitt, and Valiant, 1987; Gallant, 1990) or the voting technique (Freund and Schapire, 1999).

While this result is very useful for analyzing certain algorithms, there are some online learning algorithms for which there are provably bounds on the number of mistakes, but only as a function of a property of the data sequence. Such scenarios require an extension of this result to allow data-dependent mistake bounds. An important instance of this is the Perceptron algorithm, where a bound on the number of mistakes is known, but is quantified in terms of the *margin* of the data set; we discuss this in detail below. To extend the online-to-batch conversion result to cover these scenarios as well, we may apply our Theorem 10, which allows for data-dependent compression sizes.

Specifically, for any $n \in \mathbb{N}$ and $S = \{(X_i, Y_i)\}_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$, continuing the notation from above, define $M(\mathcal{A}, S) = \sum_{j=1}^{\infty} \mathbb{1}[\mathcal{A}(\{(X_{t_i}, Y_{t_i})\}_{i=1}^j, X_{t_{j+1}}) \neq Y_{t_{j+1}}]$. In other words, $M(\mathcal{A}, S)$ is the number of mistakes $\mathcal{A}$ would make if we cycle it through the data set S indefinitely. In particular, note that $\hat{h}_n$ is well-defined as long as $M(\mathcal{A}, S) < \infty$. We have the following result.

Theorem 22 *Let $\mathcal{A}$ be any conservative online learning algorithm, let P be any distribution on $\mathcal{X} \times \mathcal{Y}$, let $n \in \mathbb{N}$ and $\delta \in (0,1)$, and let $S = \{(X_i, Y_i)\}_{i=1}^n \sim P^n$. With probability at least $1 - \delta$, if $M(\mathcal{A}, S) < n/2$, then for $\hat{h}_n$ as defined above,*

$$R_P(\hat{h}_n) \leq \frac{2}{n - 2M(\mathcal{A}, S)} \left(M(\mathcal{A}, S) \ln(4) + \ln\left(\frac{(M(\mathcal{A}, S) + 1)(M(\mathcal{A}, S) + 2)}{\delta} \right) \right).$$

Proof For completeness, we briefly outline here an argument of Bousquet, Hanneke, Moran, and Zhivotovskiy (2020b) showing that $\mathcal{A}$ may be expressed as a stable compression scheme. Fix any $n \in \mathbb{N}$ and any data set $S = \{(X_i, Y_i)\}_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$. From the definition of $M(\mathcal{A}, S)$, let $j_1, \ldots, j_{M(\mathcal{A},S)}$ be the subsequence of $\mathbb{N}$ with $\mathbb{1}[\mathcal{A}(\{(X_{t_i}, Y_{t_i})\}_{i=1}^{j_s}, X_{t_{j_s+1}}) \neq Y_{t_{j_s+1}}]$, and define $j_0 = 0$ and $j_{M(\mathcal{A},S)+1} = \infty$. Note that, since $\mathcal{A}$ is conservative, for any $s \in \{0, \ldots, M(\mathcal{A}, S)\}$, any $j \in \mathbb{N}$ with $j_s + 1 \leq j \leq j_{s+1}$ has $\mathcal{A}(\{(X_{t_i}, Y_{t_i})\}_{i=1}^{j-1}, \cdot) = \mathcal{A}(\{(X_{t_{j_{s'}}}, Y_{t_{j_{s'}}})\}_{s'=1}^{s}, \cdot)$, so that removing any points from S that are not among $\{(X_{t_j}, Y_{t_j})\}_{j=1}^{M(\mathcal{A},S)}$ does not change the final predictor $\hat{h}_n$. Thus, $\hat{h}_n$ can be expressed as the output of a stable compression scheme: namely, $\kappa(S) = \{(X_{t_j}, Y_{t_j})\}_{j=1}^{M(\mathcal{A},S)}$, and for any $S' \in (\mathcal{X} \times \mathcal{Y})^*$, the function $h_{S'}(\cdot)$ produced by $\rho(S')$ is $\mathcal{A}(S', \cdot)$.

Therefore, for $S \sim P^n$, since $M(\mathcal{A}, S) < n/2$ also implies $\hat{R}_S(\hat{h}_n) = 0$, and $|\kappa(S)| = M(\mathcal{A}, S)$, the theorem follows immediately from Theorem 10. ■

The bound can also be relaxed to a simpler form, with slightly worse numerical constants.

Corollary 23 *Let $\mathcal{A}$ be any conservative online learning algorithm, let P be any distribution on $\mathcal{X} \times \mathcal{Y}$, let $n \in \mathbb{N}$ and $\delta \in (0,1)$, and let $S = \{(X_i, Y_i)\}_{i=1}^n \sim P^n$. With probability at least $1 - \delta$, if $M(\mathcal{A}, S) < \infty$ (so that $\hat{h}_n$ is well-defined), then for $\hat{h}_n$ as defined above,*

$$R_P(\hat{h}_n) \leq \frac{4}{n} \left(6M(\mathcal{A}, S) + \ln\left(\frac{e}{\delta} \right) \right).$$

4.3. Proof of the Optimal PAC Margin Bound for Perceptron

As was the case for SVM, an in-expectation form of the bound was established relatively early (Vapnik and Chervonenkis, 1974; Freund and Schapire, 1999), stating that $\mathbb{E}[R_P(\mathcal{A}_p(S_n))] =$

$O\left(\mathbb{E}\left[\frac{r(S_{n+1})^2}{\gamma(S_{n+1})^2}\frac{1}{n}\right]\right)$, where $S_n \sim P^n$ and $S_{n+1} \sim P^{n+1}$. However, extending the result to an optimal data-dependent *PAC* margin bound has remained open, since the naïve approach based on sample compression-based generalization bounds from Littlestone and Warmuth (1986); Floyd and Warmuth (1995) include an extra log factor. Instead, alternative more-involved online-to-batch conversion techniques have been needed to obtain the optimal form of the PAC margin bound, such as a technique by Littlestone (1989) whereby we retain all of the intermediate hypotheses produced by the algorithm as it passes through the data, and also hold out a portion of the data, using it to select which of these intermediate hypotheses to return by choosing the one making the fewest mistakes on the held-out data.

We prove the following result, from which Theorem 3 immediately follows.

Theorem 24 *For any distribution P, any $n \in \mathbb{N}$, and any $\delta \in (0,1)$, for $S \sim P^n$, with probability at least $1-\delta$, if S is linearly separable, then letting $r = r(S)$ and $\gamma = \gamma(S)$, the classifier $\hat{h}_p = \mathcal{A}_p(S)$ satisfies*

$$R_P(\hat{h}_p) \le \frac{4}{n}\left(6\frac{r^2+1}{\gamma^2} + \ln\left(\frac{e}{\delta}\right)\right)$$

and if $\frac{r^2+1}{\gamma^2} < n/2$ then

$$R_P(\hat{h}_p) \le \frac{2}{n - 2(r^2+1)/\gamma^2}\left(\frac{r^2+1}{\gamma^2}\ln(4) + 2\ln\left(\frac{r^2+1}{\gamma^2}+2\right) + \ln\left(\frac{1}{\delta}\right)\right).$$

Proof For data lying in a ball of radius r and separable with margin γ, a result of Novikoff (1963) implies that the conservative online learning algorithm $\mathcal{A}_p$ makes at most $\frac{r^2+1}{\gamma^2}$ mistakes (where the "+1" is due to the increased radius when adding an additional constant-1 feature to reduce the non-homogeneous case to the homogeneous case). The theorem now immediately follows from Theorem 22 and Corollary 23. ■

4.4. An Improved Bound on the Probability in the Region of Disagreement

Consider now the definitions from Section 2.3.

To relate $P_X(\mathrm{DIS}(V_n))$ to $\hat{t}_n$, Hanneke (2016) proved that

$$\mathbb{E}[P_X(\mathrm{DIS}(V_n))] \le \frac{\mathbb{E}[\hat{t}_{n+1}]}{n+1},$$

based on a leave-one-out argument. While this bound on the expectation appears fairly tight, in contrast the analogous known bounds on $P_X(\mathrm{DIS}(V_n))$ holding with high probability $1-\delta$ each seem to involve some slack. Specifically, Wiener, Hanneke, and El-Yaniv (2015) proved that, for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$P_X(\mathrm{DIS}(V_n)) \le \frac{1}{n}\left(10\hat{t}_n \ln\left(\frac{en}{\hat{t}_n}\right) + 4\ln\left(\frac{2}{\delta}\right)\right). \tag{6}$$

This bound was refined by Hanneke (2016) to remove the factor $\ln\left(\frac{en}{\hat{t}_n}\right)$, but at the expense of larger numerical constants and a more-involved dependence on version space compression sizes.

Specifically, Hanneke (2016) proved that, with probability at least $1-\delta$,

$$P_X(\mathrm{DIS}(V_n)) \leq \frac{16}{n}\left(2\max_{i\leq n}\hat{t}_i + \ln\left(\frac{3}{\delta}\right)\right). \tag{7}$$

The above bound features importantly in obtaining sharp distribution-dependent bounds on the label complexity of the CAL active learning algorithm (Hanneke, 2016). It is also an important component of the analysis of the risk of general *empirical risk minimization* learning algorithms in traditional (passive) supervised learning, established by Hanneke (2016).

The original proof of (7) by Hanneke (2016) used the fact that the indicator function for $\mathrm{DIS}(V_n)$ can be expressed as a sample compression scheme (with the compression set being the subset of S_n of size $\hat{t}_n$ from the definition of $\hat{t}_n$), and moreover that $\mathrm{DIS}(V_n)$ is monotonic in n. However, Bousquet, Hanneke, Moran, and Zhivotovskiy (2020b) make the observation that this compression scheme is in fact *stable*. They use this fact to refine numerical constants in a particular *distribution-free* bound on $P_X(\mathrm{DIS}(V_n))$ from Hanneke (2016) based on a combinatorial complexity measure called the *star number* from Hanneke and Yang (2015). However, they did not explore the implications of this observation for refining the data-dependent bounds on $P_X(\mathrm{DIS}(V_n))$ based on the version space compression set size $\hat{t}_n$. Here we show that our Theorem 10 and Corollary 11 apply directly to this scenario, and offer an immediate improvement to the bounds (6) and (7) in two respects: namely, we can replace $\max_{i\leq n}\hat{t}_i$ with simply $\hat{t}_n$, and we can sharpen the numerical constant factors in the bound, yielding the result claimed in Theorem 4. Specifically, we have the following slightly more-detailed result, which immediately implies Theorem 4.

Theorem 25 *For any $n \in \mathbb{N}$ and $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$P_X(\mathrm{DIS}(V_n)) \leq \frac{4}{n}\left(6\hat{t}_n + \ln\left(\frac{e}{\delta}\right)\right)$$

and if $\hat{t}_n < n/2$,

$$P_X(\mathrm{DIS}(V_n)) \leq \frac{2}{n-2\hat{t}_n}\left(\hat{t}_n\ln(4) + \ln\left(\frac{(\hat{t}_n+1)(\hat{t}_n+2)}{\delta}\right)\right).$$

Proof Define a function $\hat{h}$ on $\mathcal{X}\times\mathcal{Y}$ as $\hat{h}(x,y) = 2\mathbb{1}_{\mathrm{DIS}(V_n)}(x) - 1$, and define a distribution $\tilde{P}$ on $\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}$ such that, for $(X,Y,Z)\sim\tilde{P}$, it holds that $(X,Y)\sim P$, and that $Z=-1$ with probability one. In particular, note that $R_{\tilde{P}}(\hat{h}) = P_X(\mathrm{DIS}(V_n))$. Also, for each n, define $\tilde{S}_n = \{(X,Y,-1) : (X,Y)\in S_n\}$ (retaining the original order from S_n). Note that, by the definition of V_n, we have $\hat{R}_{\tilde{S}_n}(\hat{h}) = 0$. Furthermore, by the definition of $\hat{t}_n$, there exists a subset $S'\subseteq\tilde{S}_n$ of size at most $\hat{t}_n$ such that $\hat{h}(x,y) = 2\mathbb{1}_{\mathrm{DIS}(\mathcal{H}[\{(x,y):(x,y,-1)\in S'\}])}(x) - 1$, so that $\hat{h}$ may be viewed as a compression scheme (κ,ρ) with $|\kappa(\tilde{S}_n)| = \hat{t}_n$: that is, for any $\tilde{S}$, $\kappa(\tilde{S})$ selects any subset S' of $\tilde{S}$ of minimum size such that $\mathcal{H}[\{(x,y):(x,y,-1)\in S'\}] = \mathcal{H}[\{(x,y):(x,y,-1)\in\tilde{S}\}]$, and $\rho(S')(x,y) = 2\mathbb{1}_{\mathrm{DIS}(\mathcal{H}[\{(x,y):(x,y,-1)\in S'\}])}(x) - 1$. Furthermore, clearly any $S''\subseteq\tilde{S}$ with $S'\subseteq S''$ has $\mathcal{H}[\{(x,y):(x,y,-1)\in S''\}] = \mathcal{H}[\{(x,y):(x,y,-1)\in\tilde{S}\}]$, so that $\rho(\kappa(S'')) = \rho(\kappa(\tilde{S}))$: that is, (κ,ρ) is a *stable* compression scheme. Therefore, the theorem follows directly from Theorem 10 and Corollary 11. ■

Remark 26 *To illustrate that this can sometimes be a significant improvement over the previous results* (6) *and* (7) *of Wiener, Hanneke, and El-Yaniv (2015) and Hanneke (2016), respectively, consider a scenario where* $\mathcal{X} = [0,1]$*,* P *has* P_X *uniform on* $\mathcal{X}$ *and for* $(X,Y) \sim P$ *we have* $Y = 2\mathbb{1}_{[a,b]}(X) - 1$*, for some* $a, b \in (0,1)$ *with* $a < b$*. For* $n > \frac{1}{b-a}\ln\left(\frac{1}{\delta}\right)$*, with probability greater than* $1-\delta$ *we have* $\hat{t}_n \leq 4$*. However, for* $n < \frac{1}{b-a}$ *there is a nonzero constant probability that all samples have negative labels, in which case* $\hat{t}_n = n$*. Thus, for all large values of* n*, the bound in Theorem 25 is* $O\left(\frac{1}{n}\log\left(\frac{1}{\delta}\right)\right)$*, whereas the bound* (6) *is* $\Omega\left(\frac{1}{n}\log\left(\frac{n}{\delta}\right)\right)$ *and the bound* (7) *is* $\Omega\left(\frac{1}{n}\left(\frac{1}{b-a} + \log\left(\frac{1}{\delta}\right)\right)\right)$*. Thus, the improvement in Theorem 25 can be quite significant when* $b-a$ *is relatively small, and* n *is large relative to* $\frac{1}{b-a}$*.*

4.5. Proof of the Improved Data-dependent Bound for All ERM Algorithms

Here we present the details related to Theorem 5: data-dependent risk bounds holding for all ERM learning algorithms.

The result in Theorem 5 improves over a previous result of Hanneke (2016), which states that, with probability at least $1-\delta$, every $h \in V_n$ satisfies

$$R_P(h) = O\left(\frac{1}{n}\left(\mathrm{d}\log\left(\frac{\max_{i\leq n}\hat{t}_i}{\mathrm{d}}\right) + \log\left(\frac{1}{\delta}\right)\right)\right). \tag{8}$$

Comparing the two bounds reveals that the improvement in Theorem 5 is in replacing $\max_{i\leq n}\hat{t}_i$ with $\hat{t}_{\lfloor n/2 \rfloor}$. Recalling Remark 26 above, this change can sometimes be significant. In particular, in the example discussed in that remark, for large n the bound in Theorem 5 would be $O\left(\frac{1}{n}\log\left(\frac{1}{\delta}\right)\right)$, whereas the previously known bound from (8) would be $\Omega\left(\frac{1}{n}\log\left(\frac{1}{(b-a)\delta}\right)\right)$. Thus, in this example, Theorem 5 reflects a significant improvement when $b-a$ is small, and n is large relative to $\frac{1}{b-a}$.

The proof of Theorem 5 follows identical arguments to those used by Hanneke (2016) to prove (8), aside from substituting the bound on $P_X(\mathrm{DIS}(V_{\lfloor n/2 \rfloor}))$ from Theorem 25 in place of the bound (7) also established by Hanneke (2016). As such we omit the details, and merely sketch the main ideas underlying the argument.

The proof follows a "conditioning" argument common to this literature on refining log factors in risk bounds, originating in a proof from Hanneke (2009) of a related distribution-dependent bound for all ERM learners (which itself is slightly looser than (8); see Hanneke and Yang, 2015 and Hanneke, 2016 for relations between the relevant complexity measures). The high-level idea is to note that for $S_n = \{(X_i, Y_i)\}_{i=1}^n$, the set $D_n := \{(X_i, Y_i) : i > n/2, X_i \in \mathrm{DIS}(V_{\lfloor n/2 \rfloor})\}$ contains roughly $(n/2)P_X(\mathrm{DIS}(V_{\lfloor n/2 \rfloor}))$ elements (with high probability), which are conditionally i.i.d. with distribution $Q := P_X(\cdot | X \in \mathrm{DIS}(V_{\lfloor n/2 \rfloor}))$ given $S_{\lfloor n/2 \rfloor}$ and $|D_n|$. Furthermore, any $h \in V_n$ has $\hat{R}_{D_n}(h) = 0$. Thus, applying classic generalization bounds for ERM from Vapnik and Chervonenkis (1974); Blumer, Ehrenfeucht, Haussler, and Warmuth (1989) implies that (with high probability) every $h \in V_n$ has

$$\begin{aligned} R_Q(h) &= O\left(\frac{1}{|D_n|}\left(\mathrm{d}\log\left(\frac{|D_n|}{\mathrm{d}}\right) + \log\left(\frac{1}{\delta}\right)\right)\right) \\ &= O\left(\frac{1}{nP_X(\mathrm{DIS}(V_{\lfloor n/2 \rfloor}))}\left(\mathrm{d}\log\left(\frac{nP_X(\mathrm{DIS}(V_{\lfloor n/2 \rfloor}))}{\mathrm{d}}\right) + \log\left(\frac{1}{\delta}\right)\right)\right). \end{aligned}$$

Since every $h \in V_n$ agrees with the *best* classifier $h^* \in \mathcal{H}$ on $\mathcal{X} \setminus \mathrm{DIS}(V_{\lfloor n/2 \rfloor})$, we have $R_P(h) = R_Q(h) P_X(\mathrm{DIS}(V_{\lfloor n/2 \rfloor}))$, so that

$$R_P(h) = O\left(\frac{1}{n}\left(\mathrm{d}\log\left(\frac{n P_X(\mathrm{DIS}(V_{\lfloor n/2 \rfloor}))}{\mathrm{d}}\right) + \log\left(\frac{1}{\delta}\right)\right)\right).$$

Finally, plugging in the bound on $P_X(\mathrm{DIS}(V_{\lfloor n/2 \rfloor}))$ from Theorem 25, together with a union bound over the above high-probability events, and simplifying the expression, we get that (with high probability) each $h \in V_n$ satisfies

$$R_P(h) = O\left(\frac{1}{n}\left(\mathrm{d}\log\left(\frac{\hat{t}_{\lfloor n/2 \rfloor}}{\mathrm{d}}\right) + \log\left(\frac{1}{\delta}\right)\right)\right),$$

which is the bound claimed by Theorem 5. Readers interested in the details (and handling corner cases, numerical constants, and such things) are refered to the detailed proof of (8) by Hanneke (2016).

4.6. Proof of the Improved Bound for Compressed 1-Nearest Neighbor

We present here the details regarding Theorem 6: the a generalization bound for the compression-based nearest neighbor predictor. We have the following result, from which Theorem 6 immediately follows.

Theorem 27 *Fix any $\gamma > 0$. For any distribution P, any $n \in \mathbb{N}$, and any $\delta \in (0,1)$, for $S \sim P^n$, with probability at least $1-\delta$, the classifier $\hat{h}_\gamma = \mathcal{A}_\gamma(S)$ satisfies*

$$\left|R_P(\hat{h}_\gamma) - \hat{R}_S(\hat{h}_\gamma)\right| \le \sqrt{\hat{R}_S(\hat{h}_\gamma)\frac{72}{n}\left(4|N_\gamma| + \ln\left(\frac{4e}{\delta}\right)\right)} + \frac{32}{n}\left(4|N_\gamma| + \ln\left(\frac{4e}{\delta}\right)\right).$$

Proof We prove this by arguing that $\hat{h}_\gamma$ is the output of $\rho(\kappa(S))$ for some stable compression scheme (κ, ρ) with $|\kappa(S)| = |N_\gamma|$ contained in a particular family of such compression schemes. Specifically, for each $k \in \mathbb{N}$ and $\mathbf{b} = (b_1, \ldots, b_k) \in \mathcal{Y}^k$, define a compression scheme $(\kappa, \rho_{\mathbf{b}})$ such that $\kappa(S) = N_\gamma$ (the γ-net corresponding to S), and $\rho_{\mathbf{b}}(\kappa(S))$ is a function such that, for any x, if the nearest neighbor of x among N_γ is the i^{th} element of N_γ, then if $i \le k$ then $\rho_{\mathbf{b}}(\kappa(S)) = b_i$, and otherwise if $i > k$ then $\rho_{\mathbf{b}}(\kappa(S)) = -1$. Since $\kappa(S)$ stable, in the sense that any $S' \subset S$ with $\kappa(S) \subseteq S'$ has $\kappa(S') = \kappa(S)$, it follows that $(\kappa, \rho_{\mathbf{b}})$ is a stable compression scheme. Thus, for each $k \in \mathbb{N}$ and $\mathbf{b} \in \mathcal{Y}^k$, Corollary 18 implies that, with probability at least $1 - \frac{\delta}{2^{2k}}$, it holds that

$$\begin{aligned}&\left|R_P(\rho_{\mathbf{b}}(\kappa(S))) - \hat{R}_S(\rho_{\mathbf{b}}(\kappa(S)))\right| \\ &\le \sqrt{\hat{R}_S(\rho_{\mathbf{b}}(\kappa(S)))\frac{72}{n}\left(2|N_\gamma| + 2k + \ln\left(\frac{4e}{\delta}\right)\right)} + \frac{32}{n}\left(2|N_\gamma| + 2k + \ln\left(\frac{4e}{\delta}\right)\right).\end{aligned}$$

By the union bound, this holds simultaneously for all $k \in \mathbb{N}$ and $\mathbf{b} \in \mathcal{Y}^k$, with probability at least $1-\delta$. In particular, note that we are guarnateed that $\hat{h}_\gamma = \rho_{\mathbf{b}}(\kappa(S))$ for some $\mathbf{b} \in \mathcal{Y}^{|N_\gamma|}$. Thus, with probability at least $1-\delta$,

$$\left|R_P(\hat{h}_\gamma) - \hat{R}_S(\hat{h}_\gamma)\right| = \left|R_P(\rho_{\mathbf{b}}(\kappa(S))) - \hat{R}_S(\rho_{\mathbf{b}}(\kappa(S)))\right|$$
$$\leq \sqrt{\hat{R}_S(\hat{h}_\gamma)\frac{72}{n}\left(4|N_\gamma| + \ln\left(\frac{4e}{\delta}\right)\right)} + \frac{32}{n}\left(4|N_\gamma| + \ln\left(\frac{4e}{\delta}\right)\right).$$

■

References

M. A. Aizerman, E. A. Braverman, and L. Rozonoer. Theoretical foundations of the potential function method in pattern recognition learning. In *Automation and Remote Control*, number 25, pages 821–837, 1964.

M.-F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. In *Proceedings of the* 23^{rd} *International Conference on Machine Learning*, 2006.

M.-F. Balcan, E. Even-Dar, S. Hanneke, M. Kearns, Y. Mansour, and J. Wortman. Asymptotic active learning. In *NIPS Workshop on Principles of Learning Problem Design*, 2007.

M.-F. Balcan, S. Hanneke, and J. Wortman Vaughan. The true sample complexity of active learning. *Machine Learning*, 80(2–3):111–139, 2010.

A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the Association for Computing Machinery*, 36(4):929–965, 1989.

O. Bousquet, S. Hanneke, S. Moran, and N. Zhivotovskiy. Proper learning, Helly number, and an optimal SVM bound. In *Proceedings of the* 33^{rd} *Annual Conference on Learning Theory*, 2020a.

O. Bousquet, S. Hanneke, S. Moran, and N. Zhivotovskiy. Proper learning, Helly number, and an optimal SVM bound. *arXiv:2005.11818*, 2020b.

D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. *Machine Learning*, 15(2):201–221, 1994.

Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1996. ISBN 0-387-94618-7.

R. El-Yaniv and Y. Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(5):1605–1641, 2010.

R. El-Yaniv and Y. Wiener. Active learning via perfect selective classification. *Journal of Machine Learning Research*, 13(2):255–279, 2012.

S. Floyd and M. Warmuth. Sample compression, learnability, and the Vapnik-Chervonenkis dimension. *Machine Learning*, 21(3):269–304, 1995.

Y. Freund and R. Schapire. Large margin classification using the Perceptron algorithm. *Machine Learning*, 37:277–296, 1999.

S. Gallant. Perceptron-based learning algorithms. *IEEE Transactions on Neural Networks*, 1(2): 179–191, 1990.

Lee-Ad Gottlieb, Aryeh Kontorovich, and Robert Krauthgamer. Efficient classification for metric data (extended abstract: COLT 2010). *IEEE Transactions on Information Theory*, 60(9):5750–5759, 2014. doi: 10.1109/TIT.2014.2339840. URL http://dx.doi.org/10.1109/TIT.2014.2339840.

Lee-Ad Gottlieb, Aryeh Kontorovich, and Pinhas Nisnevitch. Near-optimal sample compression for nearest neighbors (extended abstract: NIPS 2014). *IEEE Trans. Information Theory*, 64(6): 4120–4128, 2018. doi: 10.1109/TIT.2018.2822267. URL https://doi.org/10.1109/TIT.2018.2822267.

T. Graepel, R. Herbrich, and J. Shawe-Taylor. PAC-Bayesian compression bounds on the prediction error of learning algorithms for classification. *Machine Learning*, 59(1-2):55–76, 2005.

S. Hanneke. Teaching dimension and the complexity of active learning. In *Proceedings of the* $20^{\rm th}$ *Conference on Learning Theory*, 2007a.

S. Hanneke. A bound on the label complexity of agnostic active learning. In *Proceedings of the* $24^{\rm th}$ *International Conference on Machine Learning*, 2007b.

S. Hanneke. *Theoretical Foundations of Active Learning*. PhD thesis, Machine Learning Department, School of Computer Science, Carnegie Mellon University, 2009.

S. Hanneke. Rates of convergence in active learning. *The Annals of Statistics*, 39(1):333–361, 2011.

S. Hanneke. Activized learning: Transforming passive to active with improved label complexity. *Journal of Machine Learning Research*, 13(5):1469–1587, 2012.

S. Hanneke. Theory of disagreement-based active learning. *Foundations and Trends in Machine Learning*, 7(2–3):131–309, 2014.

S. Hanneke. Refined error bounds for several learning algorithms. *Journal of Machine Learning Research*, 17(135):1–55, 2016.

S. Hanneke and A. Kontorovich. Optimality of SVM: Novel proofs and tighter bounds. *Theoretical Computer Science*, 796:99–113, 2019a.

S. Hanneke and A. Kontorovich. A sharp lower bound for agnostic learning with sample compression schemes. In *Proceedings of the* $30^{\rm th}$ *International Conference on Algorithmic Learning Theory*, 2019b.

S. Hanneke and L. Yang. Minimax analysis of active learning. *Journal of Machine Learning Research*, 16(12):3487–3602, 2015.

S. Hanneke, A. Kontorovich, S. Sabato, and R. Weiss. Universal Bayes consistency in metric spaces. *arXiv:1906.09855*, 2019.

Steve Hanneke and Aryeh Kontorovich. Optimality of SVM: novel proofs and tighter bounds. *Theor. Comput. Sci.*, 796:99–113, 2019c. doi: 10.1016/j.tcs.2019.08.030. URL https://doi.org/10.1016/j.tcs.2019.08.030.

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

D. Hsu. *Algorithms for Active Learning*. PhD thesis, Department of Computer Science and Engineering, School of Engineering, University of California, San Diego, 2010.

M. Kearns, M. Li, L. Pitt, and L. G. Valiant. Recent results on boolean concept learning. In *Proceedings of the* 4^{th} *International Workshop on Machine Learning*, 1987.

Aryeh Kontorovich and Roi Weiss. A Bayes consistent 1-NN classifier. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, AISTATS*, 2015. URL http://jmlr.org/proceedings/papers/v38/kontorovich15.html.

Robert Krauthgamer and James R. Lee. Navigating nets: Simple algorithms for proximity search. In *15th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 791–801, January 2004.

N. Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2:285–318, 1988.

N. Littlestone. From on-line to batch learning. In *Proceedings of the* 2^{nd} *Annual Workshop on Computational Learning Theory*, 1989.

N. Littlestone and M. Warmuth. Relating data compression and learnability. *Unpublished manuscript*, 1986.

T. Mitchell. Version spaces: A candidate elimination approach to rule learning. In *Proceedings of the* 5^{th} *International Joint Conference on Artificial Intelligence*, 1977.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations Of Machine Learning*. The MIT Press, 2012.

Albert Novikoff. On convergence proofs for perceptrons. In *Proc. Sympos. Math. Theory of Automata (New York, 1962)*, pages 615–622. Polytechnic Press of Polytechnic Inst. of Brooklyn, Brooklyn, N.Y., 1963.

F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958. doi: 10.1037/h0042519. URL https://doi.org/10.1037/h0042519.

J. Shawe-Taylor, P. L. Bartlett, R. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 44:1926–1940, 1998.

V. Vapnik and O. Chapelle. Bounds on error expectation for support vector machines. *Neural Computation*, 12:2013–2036, 2000.

V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.

V. Vapnik and A. Chervonenkis. *Theory of Pattern Recognition*. Nauka, Moscow, 1974.

Y. Wiener, S. Hanneke, and R. El-Yaniv. A compression technique for analyzing disagreement-based active learning. *Journal of Machine Learning Research*, 16(4):713–745, 2015.

N. Zhivotovskiy. Optimal learning via local entropies and sample compression. In *Proceedings of the* 30^{th} *Conference on Learning Theory*, pages 2023–2065, 2017.

Appendix A. Proof of Theorem 9

The version of Theorem 9 stated above is slightly more general than the original result of Bousquet, Hanneke, Moran, and Zhivotovskiy (2020a), as is required for deriving our implications. Specifically, the original result required (κ, ρ) to be a stable compression scheme guaranteed to *always* satisfy $\hat{R}_S(\rho(\kappa(S))) = 0$, whereas Theorem 9 allows *any* stable compression scheme, but only requires the bound on $R_P(\rho(\kappa(S)))$ to hold on the *event* that $\hat{R}_S(\rho(\kappa(S))) = 0$. For completeness, we include a proof of this slightly more-general claim below. It is essentially identical to the original proof of Bousquet, Hanneke, Moran, and Zhivotovskiy (2020a), aside from a few small modifications to accommodate this change.

Proof of Theorem 9 For brevity, define $[m] = \{1, \ldots, m\}$ for any $m \in \mathbb{N}$. If $k = 0$, the result trivially follows from noting that $\rho(\kappa(S))$ is a fixed pre-defined function, so that if $R_P(\rho(\kappa(S))) > \varepsilon$, then the probability of $\hat{R}_S(\rho(\kappa(S))) = 0$ is less than $(1-\varepsilon)^n \leq e^{-\varepsilon n}$, and taking $\varepsilon = \frac{1}{n}\ln\left(\frac{1}{\delta}\right)$ makes $e^{-\varepsilon n} = \delta$.

To address the remaining nontrivial cases, let us suppose $k \geq 1$. Fix any $T_n \in [n-1]$ and let $\mathcal{I}_n$ be any family of subsets of $[n]$ with the properties that each $I \in \mathcal{I}_n$ has $|I| \leq n - T_n$, and for every $i_1, \ldots, i_k \in [n]$ there exists $I \in \mathcal{I}_n$ with $\{i_1, \ldots, i_k\} \subseteq I$.

In particular, Bousquet et al. (2020a) construct a family $\mathcal{I}_n$ satisfying the properties above with $T_n = k\lfloor n/(2k) \rfloor$, and with $|\mathcal{I}_n| = \binom{2k}{k} < 4^k$: namely, let $D_1, \ldots, D_{2k}$ be any partition of $[n]$ with each $|D_i| \in \{\lfloor n/(2k) \rfloor, \lceil n/(2k) \rceil\}$, and define $\mathcal{I}_n = \{\bigcup\{D_j : j \in \mathcal{J}\} : \mathcal{J} \subseteq [2k], |\mathcal{J}| = k\}$; that is, $\mathcal{I}_n$ contains all unions of exactly k of the $2k$ sets D_j.

Let $S = \{(X_i, Y_i)\}_{i=1}^n \sim P^n$, and for any $I \subseteq [n]$ define $S_I = \{(X_i, Y_i) : i \in I\}$. For any $I \subset [n]$, since $S_{[n]\setminus I}$ is independent of S_I, for any $\varepsilon > 0$, the probability that $R_P(\rho(\kappa(S_I))) > \varepsilon$ and $\hat{R}_{S_{[n]\setminus I}}(\rho(\kappa(S_I))) = 0$ is at most $(1-\varepsilon)^{n-|I|} \leq e^{-\varepsilon(n-|I|)}$. In particular, for $I \in \mathcal{I}_n$, taking $\varepsilon = \frac{1}{T_n}\ln\left(\frac{|\mathcal{I}_n|}{\delta}\right)$, we have $e^{-\varepsilon(n-|I|)} \leq \frac{\delta}{|\mathcal{I}_n|}$. Thus, by the union bound, with probability at least $1-\delta$, every $I \in \mathcal{I}_n$ with $\hat{R}_{S_{[n]\setminus I}}(\rho(\kappa(S_I))) = 0$ has $R_P(\rho(\kappa(S_I))) \leq \frac{1}{T_n}\ln\left(\frac{|\mathcal{I}_n|}{\delta}\right)$.

In particular, note that there must exist some $I^* \in \mathcal{I}_n$ with $\kappa(S) \subseteq S_{I^*}$. Since (κ, ρ) is stable, this implies $\rho(\kappa(S)) = \rho(\kappa(S_{I^*}))$. Thus, on the above event of probability at least $1-\delta$, if $\hat{R}_S(\rho(\kappa(S))) = 0$, then also $\hat{R}_{S_{[n]\setminus I^*}}(\rho(\kappa(S_{I^*}))) = 0$, so that $R_P(\rho(\kappa(S))) = R_P(\rho(\kappa(S_{I^*}))) \leq \frac{1}{T_n}\ln\left(\frac{|\mathcal{I}_n|}{\delta}\right)$. The claimed result then follows by plugging in the set $\mathcal{I}_n$ described above, for which $T_n = k\lfloor n/(2k) \rfloor > \frac{n-2k}{2}$ and $|\mathcal{I}_n| = \binom{2k}{k} < 4^k$. ■