A Sharp Lower Bound for Agnostic Learning with Sample Compression Schemes

Steve Hanneke steve.hanneke@gmail.com
Toyota Technological Institute at Chicago

Aryeh Kontorovich karyeh@bgu.sc.il
Ben-Gurion University

Editor: Satyen Kale and Aurélien Garivier

Abstract

We establish a tight characterization of the worst-case rates for the excess risk of agnostic learning with sample compression schemes and for uniform convergence for agnostic sample compression schemes. In particular, we find that the optimal rates of convergence for size-k sample compression schemes are of the form $\sqrt{\frac{k\log(n/k)}{n}}$, which contrasts with agnostic learning with classes of VC dimension k, where the optimal rates are of the form $\sqrt{\frac{k}{n}}$.

1. Introduction

Compression-based arguments provide some of the simplest and tightest generalization bounds in the literature. These are known as *Occam learning* in the most general setting (Blumer et al., 1989), and the special case of *sample compression* (Littlestone and Warmuth, 1986; Devroye et al., 1996; Graepel et al., 2005; Floyd and Warmuth, 1995) has been receiving a fair amount of recent attention (Moran and Yehudayoff, 2016; David et al., 2016; Zhivotovskiy, 2017; Hanneke et al., 2018).

As the present paper deals with lower bounds, we stress up-front that these are *statistical* lower bounds (rather than, say, computational (Gottlieb et al., 2014) or communication-based (Kane et al., 2017)). In the realizable case, Littlestone and Warmuth (1986); Floyd and Warmuth (1995) showed that a k-compression scheme on a sample of size $n \geq ek$ achieves an expected generalization error bound of order

$$\frac{k\log(n/k)}{n}. \tag{1}$$

As the compression size k is a rough analogue of the VC-dimension, and the factor $\log(n/k)$ is known to be removable from the analogous realizable-case generalization bound for classes of VC dimension k (Haussler et al., 1994; Hanneke, 2016), one is immediately led to inquire into the necessity of the $\log(n/k)$ factor in the bound for sample compression schemes. If it were found not to be necessary, it would immediately imply improved generalization guarantees for many learning algorithms. Floyd and Warmuth (1995) take up this question in a brief but insightful remark, where they establish that the $\log(n/k)$ factor in (1) is

actually tight, in that there exist compression schemes for which an $\Omega((k/n)\log(n/k))$ lower bound also holds.

Turning to the agnostic case, the corresponding compression result from Graepel et al. (2005) implies an upper bound on the expected excess generalization error of a certain k-compression scheme on a sample of size $n \geq ek$ by a bound of order

$$\sqrt{\frac{k\log(n/k)}{n}}. \tag{2}$$

Here again, by the analogy to bounds based on the VC dimension, we are led to wonder whether the factor $\log(n/k)$ is necessary, since it is known to be removable in the analogous excess risk guarantees for agnostic learning with classes of VC dimension k (Anthony and Bartlett, 1999, Theorem 4.10). Though it is a simpler matter to give an $\Omega(\sqrt{k/n})$ lower bound, it proves significantly more challenging to determine whether the factor of $\log(n/k)$ is required for this general bound. Again, since the above compression-based generalization bound is a widely-used technique for establishing generalization guarantees for certain types of learning algorithms, if one could show that the factor $\log(n/k)$ is superfluous, it would immediately have a wide range of substantial implications by improving the known generalization guarantees for many learning algorithms in the literature. However, as our main result in this work (Section 2), we prove that this $\log(n/k)$ factor in (2) generally *cannot* be removed. Specifically, we argue that, for a certain family of reconstruction functions, regardless of the choice of compression function, a lower bound of the form (2) holds.

We stress that this fact is not at all obvious from the necessity arguments offered by Floyd and Warmuth (1995) for the realizable case. The argument used there is essentially a reduction to the known fact that, in the realizable case with certain classes of VC dimension k, there are empirical risk minimization (ERM) learning rules whose expected error rate can be as high as $\Omega((k/n)\log(n/k))$. They establish their lower bound for realizable-case compression schemes of size k by arguing that, in a certain scenario of this type, there is a compression scheme that emulates one of these high-error ERM learners. In contrast, in the case of *agnostic* learning, it is known that *all* ERM learners guarantee expected excess error $O(\sqrt{k/n})$ (Anthony and Bartlett, 1999, Theorem 4.10). Thus, the approach of Floyd and Warmuth (1995) will not suffice for establishing a lower bound proportional to $\sqrt{(k/n)\log(n/k)}$ for agnostic learning with compression schemes. Indeed, the construction we arrive at in our proof below is necessarily significantly more-involved.

In addition to this result for the basic order-invariant compression schemes, we also prove an analogous lower bound for *order-dependent* compression schemes (Section 3), where the factor becomes $\log(n)$, which again is tight.

2. Order-Independent Compression Schemes

Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is any nonempty set and $\mathcal{Y} = \{0,1\}$, and suppose $\mathcal{X}$ is equipped with a σ-algebra defining the measurable sets. An agnostic sample compression scheme is specified by a *size* $k \in \mathbb{N}$ and a *reconstruction function* ρ, which maps any (multi)set $\{z_1, \ldots, z_{k'}\} \subseteq \mathcal{Z}$ with $0 \leq k' \leq k$ to a measurable function $h : \mathcal{X} \to \mathcal{Y}$. For any $n \in \mathbb{N}$ and any sequence $z_1, \ldots, z_n$, define

$$\mathcal{H}_{k,\rho}(z_1, \ldots, z_n) = \{\rho(\{z_{i_1}, \ldots, z_{i_{k'}}\}) : k' \leq k, 1 \leq i_1 < \cdots < i_{k'} \leq n\}.$$

Now for any probability measure P on $\mathcal{Z}$ and any $n \in \mathbb{N}$, let $Z_{[n]} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ be independent P-distributed random variables, and for any classifier $h : \mathcal{X} \to \mathcal{Y}$, define $R(h; P) = P(\{(x, y) : h(x) \neq y\})$ the *error rate* of h, and define $\hat{R}(h; Z_{[n]}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[h(X_i) \neq Y_i]$ the *empirical error rate* of h.

Now there are essentially two types of results for agnostic compression schemes in the literature: namely, *uniform convergence* rates and *agnostic learning* excess risk guarantees. We begin with the first of these. For any fixed agnostic sample compression scheme (k, ρ), denote

$$\mathcal{E}_{\mathrm{uc}}(n, k, \rho, P) = \mathbf{E} \sup_{h \in \mathcal{H}_{k,\rho}(Z_{[n]})} |\hat{R}(h; Z_{[n]}) - R(h; P)|.$$

Then, for any $n, k \in \mathbb{N}$, define

$$\mathcal{E}_{\mathrm{uc}}(n, k) = \sup_{P, \rho} \mathcal{E}_{\mathrm{uc}}(n, k, \rho, P),$$

where P ranges over all probability measures on $\mathcal{Z}$, and ρ ranges over all reconstruction functions (for the given size k). For results on uniform convergence for agnostic compression schemes, this is the object of primary interest to this work.

It is known (essentially from the arguments of Graepel et al. (2005, Theorem 2)) that for any $n, k \in \mathbb{N}$ with $n \geq ek$,

$$\mathcal{E}_{\mathrm{uc}}(n, k) \lesssim \sqrt{\frac{k \log(n/k)}{n}}.$$

This upper bound is similar in form to the original bound of Vapnik and Chervonenkis (1971) for uniform convergence rates for VC classes of VC dimension k. However, that bound was later refined[1] to the form $\sqrt{\frac{k}{n}}$, removing the factor $\log(n/k)$. It is therefore natural to wonder whether this same refinement might be achieved by size-k agnostic sample compression schemes. To our knowledge, this question has not previously been addressed in the literature.

The other type of results of interest for agnostic compression schemes are agnostic learning excess risk guaratnees. Specifically, a *compression function* κ is a mapping from any sequence $z_1, \ldots, z_n$ in $\mathcal{Z}$ to an unordered sub(multi)set[2] $S \subseteq \{z_1, \ldots, z_n\}$ of size at most k. Then, denoting $\hat{h}_n = \rho(\kappa(Z_{[n]}))$, define

$$\mathcal{E}_{\mathrm{ag}}(n, k, \rho, \kappa, P) = \mathbf{E}\left[R(\hat{h}_n; P) - \min_{h \in \mathcal{H}_{k,\rho}(Z_{[n]})} R(h; P)\right]$$

and then define

$$\mathcal{E}_{\mathrm{ag}}(n, k) = \sup_{\rho} \inf_{\kappa} \sup_{P} \mathcal{E}_{\mathrm{ag}}(n, k, \rho, \kappa, P),$$

1. A detailed account of the intermediate steps leading to this seminal result is presented in Anthony and Bartlett (1999); significant milestones include Pollard (1982); Koltchinskii (1981); Talagrand (1994); Haussler (1995).
2. An element in S may repeat up to as many times as it occurs in the sequence $z_1, \ldots, z_n$, so that S effectively corresponds to picking a set of up to k distinct *indices* in $\{1, \ldots, n\}$ to include the corresponding z_i points.

where again P ranges over all probability measures on $\mathcal{Z}$ and ρ ranges over all reconstruction functions (for the given size k), and where κ ranges over all compression functions (for the given size k).

By a standard argument, if we specify κ so as to always minimize the empirical error rate $\hat{R}(\rho(\kappa(Z_{[n]})))$, then the excess error rate can be bounded by twice the uniform convergence bound, which immediately implies

$$\mathcal{E}_{\mathrm{ag}}(n,k) \leq 2\mathcal{E}_{\mathrm{uc}}(n,k). \tag{3}$$

An immediate implication from above is then that any n, k with $n \geq ek$ has

$$\mathcal{E}_{\mathrm{ag}}(n,k) \lesssim \sqrt{\frac{k\log(n/k)}{n}}.$$

Here again, this bound is of the same form originally proven by Vapnik and Chervonenkis (1971) for empirical risk minimization in classes of VC dimension k, which was later refined to a sharp bound of order $\sqrt{k/n}$ (Anthony and Bartlett, 1999, Theorem 4.10). As such, it is again natural to ask whether the $\log(n/k)$ factor in the above bound for agnostic sample compression can be reduced to a constant, or is in fact necessary. Our main contribution in this work is a construction showing that this log factor is indeed necessary, as stated in the following results. In all of the results below, c represents a numerical constant, whose value must be set sufficiently large (as discussed in the proofs) for the results to hold.

Theorem 1 *For any $n, k \in \mathbb{N}$ with $|\mathcal{X}| \geq n \geq ck$,*

$$\mathcal{E}_{\mathrm{ag}}(n,k) \gtrsim \sqrt{\frac{k\log(n/k)}{n}}.$$

By the relation (3) discussed above, between uniform convergence and agnostic learning by empirical risk minimization over $\mathcal{H}_{k,\rho}(Z_{[n]})$, this also has the following immediate implication.

Theorem 2 *For any $n, k \in \mathbb{N}$ with $|\mathcal{X}| \geq n \geq ck$,*

$$\mathcal{E}_{\mathrm{uc}}(n,k) \gtrsim \sqrt{\frac{k\log(n/k)}{n}}.$$

Together with the known upper bounds mentioned above, this provides a tight characterization of the worst-case rate of uniform convergence for agnostic sample compression schemes.

Corollary 3 *For any $n, k \in \mathbb{N}$ with $|\mathcal{X}| \geq n \geq ck$,*

$$\mathcal{E}_{\mathrm{ag}}(n,k) \asymp \sqrt{\frac{k\log(n/k)}{n}}$$

and

$$\mathcal{E}_{\mathrm{uc}}(n,k) \asymp \sqrt{\frac{k\log(n/k)}{n}}.$$

We now present the proof of Theorem 1.

Proof [Proof of Theorem 1] Fix any $n, k \in \mathbb{N}$ with $|\mathcal{X}| \geq n \geq ck$ for a sufficiently large numerical constant $c \geq 4$ (discussed below), denote $m = 2^{\lfloor \log_2(n/k) \rfloor}$, and let $x_0, \ldots, x_{km-1}$ denote any km distinct elements of $\mathcal{X}$. For simplicity, suppose $m/\log_2(m) \in \mathbb{N}$ (the argument easily extends to the general case by introducing floor functions, with only the numerical constants changing in the final result). The essential strategy behind our construction is to create an embedded instance of a construction for proving the lower bound for agnostic learning in VC classes, where here the VC dimension of the embedded scenario will be $k \log_2(m)$. The construction of this embedded scenario is our starting point. From there we also need to argue that there is a function contained in $\mathcal{H}_{k,\rho}(Z_{[n]})$ with risk not too much larger than the best classifier in the embedded VC class, which allows us to extend the lower bound argument for the embedded VC class to compression schemes. For any $0 \leq i \leq m-1$, let $b_j(i)$ denote the $(j+1)^{\text{th}}$ bit of i in the binary representation of i: that is, $i = \sum_{j=0}^{\log_2(m)-1} b_j(i) 2^j$, with $b_0(i), \ldots, b_{\log_2(m)-1}(i) \in \{0,1\}$.

We construct the reconstruction function based on k "blocks", each with $m/\log_2(m)$ "sub-blocks". Specifically, for each $t \in \{1, \ldots, k\}$, define a block $B_t = \{(t-1)m, \ldots, tm-1\}$, and for each $s \in \{1, \ldots, m/\log_2(m)\}$, define a sub-block

$$B_{ts} = \{(t-1)m + (s-1)\log_2(m), \ldots, (t-1)m + s\log_2(m) - 1\}.$$

Then for any $i \in B_t$ and $t \in \{1, \ldots, k\}$, define $h_{t,i} : \mathcal{X} \to \mathcal{Y}$ as any function satisfying the property that, for $j = (t-1)m + (s-1)\log_2(m) + r \in B_{ts}$ (for any $s \in \{1, \ldots, m/\log_2(m)\}$ and $r \in \{0, \ldots, \log_2(m) - 1\}$),

$$h_{t,i}(x_j) = b_r(i - (t-1)m).$$

Thus, the subsequence of x_j points corresponding to the indices j within each sub-block B_{ts} have $h_{t,i}(x_j)$ values corresponding to the bits of the integer $i - (t-1)m$, and this repeats identically for every sub-block B_{ts} in the block B_t.

Now we construct a reconstruction function ρ that outputs functions which correspond to some such $h_{t,i}$ function within each block B_t, but potentially using a different bit pattern $i - (t-1)m$ for each t. Formally, for any $i_1, \ldots, i_k \in \mathbb{N} \cup \{0\}$ with $i_t \in B_t$ (for each $t \in \{1, \ldots, k\}$), and any $y_1, \ldots, y_k \in \mathcal{Y}$, define $\rho(\{(x_{i_1}, y_1), \ldots, (x_{i_k}, y_k)\}) = \tilde{h}_{i_1,\ldots,i_k}$, where $\tilde{h}_{i_1,\ldots,i_k} : \mathcal{X} \to \mathcal{Y}$ is any function satisfying the property that each $t \in \{1, \ldots, k\}$ and $j \in \{(t-1)m, \ldots, tm-1\}$ has $\tilde{h}_{i_1,\ldots,i_k}(x_j) = h_{t,i_t}(x_j)$: that is, the points x_{i_t} in the compression set are interpreted by the compression scheme as encoding the desired *label sequence* for sub-blocks B_{ts} in the *bits* of $i_t - (t-1)m$. For our purposes, $\tilde{h}_{i_1,\ldots,i_k}(x)$ may be defined arbitrarily for $x \in \mathcal{X} \setminus \{x_0, \ldots, x_{km-1}\}$. Note that $\rho(\{(x_{i_1}, y_1), \ldots, (x_{i_k}, y_k)\})$ is invariant to the $y_1, \ldots, y_k$ values, so for brevity we will drop the y_i arguments and simply write $\rho(\{x_{i_1}, \ldots, x_{i_k}\})$ (this is often referred to as an *unlabeled* compression scheme in the literature). For completeness, $\rho(S)$ should also be defined for sets $S \subseteq \mathcal{X}$ of size at most k that do not have exactly one element x_i with $i \in B_t$ for every t; for our purposes, let us suppose that in these cases, for every t with $S \cap \{x_i : i \in B_t\} \neq \emptyset$, let $i_t = \min\{i \in B_t : x_i \in S\}$, and for every t with $S \cap \{x_i : i \in B_t\} = \emptyset$, let $i_t = (t-1)m$; then define $\rho(S) = \tilde{h}_{i_1,\ldots,i_k}$. In this way, $\rho(S)$ is defined for all $S \subseteq \mathcal{X}$ with $|S| \leq k$.

Now define a family of distributions $P^{(\sigma)}$, $\sigma = \{\sigma_{t,r}\}$, with $\sigma_{t,r} \in \{-1, 1\}$ for $t \in \{1, \ldots, k\}$ and $r \in \{0, \ldots, \log_2(m) - 1\}$, as follows. Every $P^{(\sigma)}$ has marginal P_X on $\mathcal{X}$

uniform on $x_0, \ldots, x_{km-1}$, and for each $j = (t-1)m + (s-1)\log_2(m) + r \in B_{ts}$ (for $t \in \{1, \ldots, k\}$, $s \in \{1, \ldots, m/\log_2(m)\}$, and $r \in \{0, \ldots, \log_2(m) - 1\}$) set $P^{(\sigma)}(Y = 1 | X = x_j) = \frac{1}{2} + \frac{\epsilon}{2}\sigma_{t,r}$, where

$$\epsilon = \sqrt{\frac{k \log_2(m)}{n}}.$$

Now let us suppose σ is chosen *randomly*, with $\sigma_{t,r}$ independent Uniform($\{-1, 1\}$). Then (since max $\geq$ average) note that choosing $P = P^{(\sigma)}$ now results in

$$\mathcal{E}_{\mathrm{ag}}(n, k) \geq \mathbf{E}\left[\inf_{\kappa} \mathcal{E}_{\mathrm{ag}}(n, k, \rho, \kappa, P^{(\sigma)})\right],$$

so that it suffices to study the expectation on the right hand side.

As mentioned, the purpose of this construction is to create an embedded instance of a scenario that witnesses the lower bound for agnostic learning in VC classes, where the VC dimension of the embedded scenario here is $k \log_2(m)$. Specifically, in our construction, for any $t \in \{1, \ldots, k\}$ and $r \in \{0, \ldots, \log_2(m) - 1\}$, denoting by

$$C_{t,r} = \{(t-1)m + (s-1)\log_2(m) + r : s \in \{1, \ldots, m/\log_2(m)\}\},$$

the locations $\{x_j : j \in C_{t,r}\}$ together essentially represent a single location in the embedded problem: that is, their $h_{t,i}(x_j)$ values are bound together, as are their $P(Y = 1 | X = x_j)$ values. However, this itself is not sufficient to supply a lower bound, since the constructed scenario exists only in the *complete* space of possible reconstructions $\mathcal{H}^*_{k,\rho} = \{\rho(\{x_{i_1}, \ldots, x_{i_k}\}) : i_1, \ldots, i_k \in \{0, \ldots, km - 1\}\}$, and it is entirely possible that $\min_{h \in \mathcal{H}_{k,\rho}(Z_{[n]})} R(h; P) > \min_{h \in \mathcal{H}^*_{k,\rho}} R(h; P)$: that is, the smallest error rate achievable in $\mathcal{H}_{k,\rho}(Z_{[n]})$ can conceivably be significantly larger than the smallest error rate achievable in the embedded VC class, so that compression schemes in this scenario do not automatically inherit the lower bounds for the constructed VC class. To account for this, we will study a decomposition of the construction into k subproblems, corresponding to the k blocks B_t in the construction, and we will argue that within these subproblems there remains in $\mathcal{H}_{k,\rho}(Z_{[n]})$ a function with optimal predictions on *most* of the points, and then stitch these functions together to argue that there do exist functions in $\mathcal{H}_{k,\rho}(Z_{[n]})$ having near-optimal error rates relative to the best in $\mathcal{H}^*_{k,\rho}$.

Specifically, fix any $t \in \{1, \ldots, k\}$ and let $P^{(\sigma)}_t$ denote the conditional distribution of $(X, Y) \sim P^{(\sigma)}$ given σ and the event that $X \in \{x_j : j \in B_t\}$. Also denote $\mathcal{H}^*_t = \{h_{t,i} : i \in B_t\}$, $i^*_t = \mathrm{argmin}_{i \in B_t} R(h_{t,i}; P^{(\sigma)}_t)$, $h^*_t = h_{t,i^*_t}$, and

$$\mathcal{H}_t(Z_{[n]}) = \{h_{t,i} : i \in B_t, x_i \in \{x_{(t-1)m}, X_1, \ldots, X_n\}\}.$$

These correspond to the classifications of block t realizable by classifiers in $\mathcal{H}_{k,\rho}(Z_{[n]})$ (where the addition of the $x_{(t-1)m}$ point to the data set is due to our specification of $\rho(S)$ for sets S that contain no elements x_i with $i \in B_t$, so that classifying block t according to $h_{t,(t-1)m}$ is always possible). There are now two components at this stage in the argument: first, that any compression function κ results in $\hat{h} = \rho(\kappa(Z_{[n]}))$ with $\mathbf{E}[R(\hat{h}; P^{(\sigma)}_t) - R(h^*_t; P^{(\sigma)}_t)] \geq \epsilon/(8e^4)$, and second, that $\mathbf{E}[\min_{h \in \mathcal{H}_t(Z_{[n]})} R(h; P^{(\sigma)}_t) - R(h^*_t; P^{(\sigma)}_t)] \leq \epsilon/(16e^4)$.

For the first part, note that for any $r \in \{0, \dots, \log_2(m) - 1\}$, for any $j \in C_{t,r}$, $h_t^*(x_j) = \frac{\sigma_{t,r}+1}{2}$. Furthermore, for any compression function κ, note that any $\hat{h}$ that $\rho(\kappa(Z_{[n]}))$ is capable of producing has $\hat{h}(x_j) = \hat{h}(x_{j'})$ for every $j, j' \in C_{t,r}$. In particular, if we let $\hat{i}_t \in B_t$ be the index with $b_r(\hat{i}_t - (t-1)m) = \hat{h}(x_{(t-1)m+r})$ for every $r \in \{0, \dots, \log_2(m) - 1\}$, then $\hat{h}$ and $h_{t,\hat{i}_t}$ agree on every element of $\{x_j : j \in B_t\}$. This also implies

$$\begin{aligned} R(\hat{h}; P_t^{(\sigma)}) - R(h_t^*; P_t^{(\sigma)}) &= R(h_{t,\hat{i}_t}; P_t^{(\sigma)}) - R(h_t^*; P_t^{(\sigma)}) \\ &= \frac{1}{\log_2(m)} \sum_{r=0}^{\log_2(m)-1} \epsilon \mathbb{I}\left[b_r(\hat{i}_t - (t-1)m) \neq \frac{\sigma_{t,r}+1}{2} \right]. \end{aligned}$$

Therefore, denoting by $n_{t,r} = |\{i \leq n : X_i \in \{x_j : j \in C_{t,r}\}\}|$, we have

$$\mathbf{E}[R(\hat{h}; P_t^{(\sigma)}) - R(h_t^*; P_t^{(\sigma)})] = \frac{\epsilon}{\log_2(m)} \sum_{r=0}^{\log_2(m)-1} \mathbf{E}\left[\mathbf{P}\left(b_r(\hat{i}_t - (t-1)m) \neq \frac{\sigma_{t,r}+1}{2} \middle| n_{t,r} \right) \right].$$

For any given $r \in \{0, \dots, \log_2(m) - 1\}$, enumerate the $n_{t,r}$ random variables (X_i, Y_i) with $X_i \in \{x_j : j \in C_{t,r}\}$ as $(X_{i(r,1)}, Y_{i(r,1)}), \dots, (X_{i(r,n_{t,r})}, Y_{i(r,n_{t,r})})$, and note that given $n_{t,r}$, the values $(Y_{i(r,1)}, \dots, Y_{i(r,n_{t,r})})$ are a *sufficient statistic* for $\sigma_{t,r}$ (see Definition 2.4 of Schervish (1995)), and therefore (see Theorem 3.18 of Schervish (1995)) there exists a (randomized) decision rule $\hat{f}_{t,r}(Y_{i(r,1)}, \dots, Y_{i(r,n_{t,r})})$ depending only on these variables and independent random bits such that

$$\mathbf{P}\left(b_r(\hat{i}_t - (t-1)m) \neq \frac{\sigma_{t,r}+1}{2} \middle| n_{t,r} \right) = \mathbf{P}\left(\hat{f}_{t,r}(Y_{i(r,1)}, \dots, Y_{i(r,n_{t,r})}) \neq \frac{\sigma_{t,r}+1}{2} \middle| n_{t,r} \right).$$

Furthermore, by Lemma 5.1 of Anthony and Bartlett (1999)[3], we have

$$\mathbf{P}\left(\hat{f}_{t,r}(Y_{i(r,1)}, \dots, Y_{i(r,n_{t,r})}) \neq \frac{\sigma_{t,r}+1}{2} \middle| n_{t,r} \right) > \frac{1}{8e} \exp\{-(8/3) n_{t,r} \epsilon^2\}.$$

Altogether, and combined with Jensen's inequality, we have that

$$\begin{aligned} &\mathbf{E}[R(\hat{h}; P_t^{(\sigma)}) - R(h_t^*; P_t^{(\sigma)})] \\ &\geq \frac{\epsilon}{8e \log_2(m)} \sum_{r=0}^{\log_2(m)-1} \mathbf{E}\left[\exp\{-(8/3) n_{t,r} \epsilon^2\}\right] \geq \frac{\epsilon}{8e \log_2(m)} \sum_{r=0}^{\log_2(m)-1} \exp\{-(8/3) \mathbf{E}[n_{t,r}] \epsilon^2\} \\ &= \frac{\epsilon}{8e \log_2(m)} \sum_{r=0}^{\log_2(m)-1} \exp\left\{ -(8/3) \frac{n}{k \log_2(m)} \epsilon^2 \right\} \geq \frac{\epsilon}{8e \log_2(m)} \sum_{r=0}^{\log_2(m)-1} e^{-(8/3)} \geq \frac{\epsilon}{8e^4}. \end{aligned}$$

Now for the second part, for any $x \in \{x_i : i \in \{0, \dots, km-1\}\}$, denote by $I(x)$ the index i such that $x = x_i$. Note that an i for which $h_{t,i}$ has minimal $R(h_{t,i}; P_t^{(\sigma)})$ among all $h_{t,i'} \in \mathcal{H}_t(Z_{[n]})$ can equivalently be defined as an i with minimal $\sum_{j=0}^{\log_2(m)-1} \mathbb{I}[b_j(i - (t -$

3. The lower bound in (Anthony and Bartlett, 1999, Lemma 5.1) relied on Slud's lemma; the analysis has since been tightened to yield asymptotically optimal lower bounds (Kontorovich and Pinelis, 2016).

$1)m) \neq b_j(i_t^* - (t-1)m)]$ among all $i' \in B_t \cap \{I(X_1), \ldots, I(X_n), (t-1)m\}$, and furthermore, for such an i,

$$R(h_{t,i}; P_t^{(\sigma)}) - R(h_t^*; P_t^{(\sigma)}) = \frac{\epsilon}{\log_2(m)} \sum_{j=0}^{\log_2(m)-1} \mathbb{I}[b_j(i-(t-1)m) \neq b_j(i_t^* - (t-1)m)].$$

For any $i \in B_t$, denote

$$\Delta_t(i) = \sum_{j=0}^{\log_2(m)-1} \mathbb{I}[b_j(i-(t-1)m) \neq b_j(i_t^* - (t-1)m)].$$

Thus, it suffices to establish the stated upper bound for the quantity

$$\frac{\epsilon}{\log_2(m)} \mathbf{E}\left[\min_{i \in B_t \cap \{I(X_1), \ldots, I(X_n), (t-1)m\}} \Delta_t(i)\right].$$

Now consider a random variable $X \sim P_X(\cdot | \{x_i : i \in B_t)$: that is, X has distribution the same as the marginal of $P_t^{(\sigma)}$ on $\mathcal{X}$. Then note that the conditional distribution of $\Delta_t(I(X))$ given σ is Binomial$(\log_2(m), \frac{1}{2})$. Let $q = 16e^4$, and suppose the numerical constant c is sufficiently large so that $q \leq (1/2)\log_2(m)$. Then we have

$$\mathbf{P}\left(\Delta_t(I(X)) \leq \frac{1}{2q}\log_2(m) \middle| \sigma\right) = \sum_{\ell=0}^{\lfloor (1/2q)\log_2(m) \rfloor} \binom{\log_2(m)}{\ell} \frac{1}{m}$$
$$\geq \frac{1}{m}\left(\frac{\log_2(m)}{\lfloor (1/2q)\log_2(m) \rfloor}\right)^{\lfloor (1/2q)\log_2(m) \rfloor} \geq \frac{1}{m}(4q)^{(1/2q)\log_2(m)} = m^{(1/2q)\log_2(4q)-1}.$$

Thus, by independence of the samples $X_1, \ldots, X_n$, denoting $n_t = |\{i \leq n : X_i \in \{x_j : j \in B_t\}\}|$, we have

$$\mathbf{P}\left(\min_{i \in B_t \cap \{I(X_1), \ldots, I(X_n), (t-1)m\}} \Delta_t(i) > \frac{1}{2q}\log_2(m) \middle| \sigma, n_t\right)$$
$$\leq \mathbf{P}\left(\forall i \in B_t \cap \{I(X_1), \ldots, I(X_n)\}, \Delta_t(i) > \frac{1}{2q}\log_2(m) \middle| \sigma, n_t\right)$$
$$= \mathbf{P}\left(\Delta_t(I(X)) > \frac{1}{2q}\log_2(m) \middle| \sigma\right)^{n_t}$$
$$\leq \left(1 - m^{(1/2q)\log_2(4q)-1}\right)^{n_t} \leq \exp\left\{-m^{(1/2q)\log_2(4q)-1} n_t\right\}.$$

Altogether, by the law of total expectation, and using the fact that $R(h; P_t^{(\sigma)}) \leq 1$, we have established that

$$\mathbf{E}\left[\min_{h \in \mathcal{H}_t(Z_{[n]})} R(h; P_t^{(\sigma)}) - R(h_t^*; P_t^{(\sigma)})\right] \leq \frac{\epsilon}{2q} + \mathbf{E}\left[\exp\left\{-m^{(1/2q)\log_2(4q)-1} n_t\right\}\right].$$

Since n_t is a Binomial$(n, 1/k)$ random variable, the rightmost term evaluates to the moment generating function of this distribution: that is,

$$\begin{aligned}
&\mathbf{E}\Big[\exp\Big\{-m^{(1/2q)\log_2(4q)-1}n_t\Big\}\Big] = \left(1 - \frac{1}{k} + \frac{1}{k}\exp\Big\{-m^{(1/2q)\log_2(4q)-1}\Big\}\right)^n \\
&\leq \max\left\{2\left(1-\frac{1}{k}\right)^n, 2\left(\frac{1}{k}\right)^n \exp\Big\{-m^{(1/2q)\log_2(4q)-1}n\Big\}\right\} \\
&\leq \max\left\{2e^{-n/k}, 2\exp\Big\{-m^{(1/2q)\log_2(4q)}\Big\}\right\} \\
&= \max\left\{2e^{-n/k}, 2\left(\exp\Big\{-(1/2q)\log_2(4q)m^{(1/2q)\log_2(4q)}\Big\}\right)^{\frac{2q}{\log_2(4q)}}\right\} \\
&\leq \max\left\{2e^{-n/k}, 2\left(\frac{2q}{\log_2(4q)}\right)^{\frac{2q}{\log_2(4q)}}\frac{1}{m}\right\}.
\end{aligned}$$

Since both of these terms shrink strictly faster than the above specification of ϵ as a function of n/k, and therefore, for a sufficiently large choice of the numerical constant c, both of these terms are smaller than $\frac{\epsilon}{32e^4}$. Therefore, we conclude that

$$\mathbf{E}\left[\min_{h\in\mathcal{H}_t(Z_{[n]})} R(h; P_t^{(\sigma)}) - R(h_t^*; P_t^{(\sigma)})\right] \leq \frac{\epsilon}{16e^4},$$

as claimed.

Together, these two components imply that

$$\begin{aligned}
&\mathbf{E}\left[R(\hat{h}; P_t^{(\sigma)}) - \min_{h\in\mathcal{H}_t(Z_{[n]})} R(h; P_t^{(\sigma)})\right] \\
&= \mathbf{E}\Big[R(\hat{h}; P_t^{(\sigma)}) - R(h_t^*; P_t^{(\sigma)})\Big] - \mathbf{E}\left[\min_{h\in\mathcal{H}_t(Z_{[n]})} R(h; P_t^{(\sigma)}) - R(h_t^*; P_t^{(\sigma)})\right] \geq \frac{\epsilon}{16e^4}.
\end{aligned}$$

Finally, it is time to combine these results for the individual B_t blocks into a global statement about $P^{(\sigma)}$. In particular, note that any h has $R(h; P^{(\sigma)}) = \frac{1}{k}\sum_{t=1}^k R(h; P_t^{(\sigma)})$. Also note that any h that ρ is capable of producing from arguments that are subsets of $\{X_1, \ldots, X_n\}$ can be represented as $h = \tilde{h}_{i_1,\ldots,i_k}$ for some $i_1, \ldots, i_k$ where every $t \in \{1, \ldots, k\}$ has $i_t \in B_t$ and $x_{i_t} \in \{X_1, \ldots, X_n, x_{(t-1)m}\}$ (where the addition of the $x_{(t-1)m}$ covers the case that the set does not include any x_i with $i \in B_t$, as we defined that case above). Furthermore, every function $\tilde{h}_{i_1,\ldots,i_k}$ with i_t values satisfying these conditions *can* be realized by ρ using an argument S that is a subset of $\{X_1, \ldots, X_n\}$ of size at most k: namely, the

set $\{x_{i_t} : t \in \{1,\ldots,k\}, i_t \neq (t-1)m\} \subseteq \{X_1,\ldots,X_n\}$. Therefore,

$$\min_{h\in\mathcal{H}_{k,\rho}(Z_{[n]})} R(h;P^{(\sigma)}) = \min_{\substack{(i_1,\ldots,i_k)\in B_1\times\cdots\times B_k:\\ \{x_{i_1},\ldots,x_{i_k}\}\subseteq\{X_1,\ldots,X_n\}\cup\{x_{(t-1)m}:t\le k\}}} R(\tilde{h}_{i_1,\ldots,i_k};P^{(\sigma)})$$

$$= \min_{\substack{(i_1,\ldots,i_k)\in B_1\times\cdots\times B_k:\\ \{x_{i_1},\ldots,x_{i_k}\}\subseteq\{X_1,\ldots,X_n\}\cup\{x_{(t-1)m}:t\le k\}}} \frac{1}{k}\sum_{t=1}^{k} R(h_{t,i_t};P_t^{(\sigma)})$$

$$= \frac{1}{k}\sum_{t=1}^{k} \min_{\substack{i_t\in B_t:\\ x_{i_t}\in\{X_1,\ldots,X_n,x_{(t-1)m}\}}} R(h_{t,i_t};P_t^{(\sigma)}) = \frac{1}{k}\sum_{t=1}^{k} \min_{h\in\mathcal{H}_t(Z_{[n]})} R(h;P_t^{(\sigma)}).$$

Thus, for any compression function κ, denoting $\hat{h} = \rho(\kappa(Z_{[n]}))$,

$$\mathbf{E}\left[R(\hat{h};P^{(\sigma)}) - \min_{h\in\mathcal{H}_{k,\rho}(Z_{[n]})} R(h;P^{(\sigma)})\right]$$

$$\ge \frac{1}{k}\sum_{t=1}^{k}\mathbf{E}\left[R(\hat{h};P_t^{(\sigma)}) - \min_{h\in\mathcal{H}_t(Z_{[n]})} R(h;P_t^{(\sigma)})\right] \ge \frac{1}{16e^4}\epsilon \gtrsim \sqrt{\frac{k\log(n/k)}{n}}.$$

■

3. Order-Dependent Compression Schemes

The above construction shows that the well-known $\sqrt{\frac{k\log(n/k)}{n}}$ upper bound for agnostic compression schemes is sometimes tight. Note that, in the definition of agnostic compression schemes, we required that the reconstruction function ρ take as input a (multi)set. This type of compression scheme is often referred to as being *permutation invariant*, since the compression set argument is unordered (or equivalently ρ does not depend on the order of elements in its argument).

We can also show a related result for the case of *order-dependent* compression schemes. An order-dependent agnostic sample compression scheme is specified by a *size* $k \in \mathbb{N}$ and an *order-dependent reconstruction function* ρ, which maps any ordered *sequence* $(z_1,\ldots,z_{k'}) \in \mathcal{Z}^{k'}$ with $0 \le k' \le k$ to a measurable function $h : \mathcal{X} \to \mathcal{Y}$. For any $n \in \mathbb{N}$ and any sequence $z_1,\ldots,z_n$, define

$$\mathcal{H}_{k,\rho}(z_1,\ldots,z_n) = \{\rho((z_{i_1},\ldots,z_{i_{k'}})) : k' \le k, i_1,\ldots,i_{k'} \in \{1,\ldots,n\}\}.$$

Now for any probability measure P on $\mathcal{Z}$ and any $n \in \mathbb{N}$, continuing the notation from above, for any fixed order-dependent agnostic sample compression scheme (k,ρ), as above denote

$$\mathcal{E}^o_{\mathrm{uc}}(n,k,\rho,P) = \mathbf{E}\sup_{h\in\mathcal{H}_{k,\rho}(Z_1,\ldots,Z_n)} |\hat{R}(h;Z_{[n]}) - R(h;P)|,$$

and for any $n,k \in \mathbb{N}$, define

$$\mathcal{E}^o_{\mathrm{uc}}(n,k) = \sup_{P,\rho} \mathcal{E}^o_{\mathrm{uc}}(n,k,\rho,P),$$

where P ranges over all probability measures on $\mathcal{Z}$, and ρ ranges over all order-dependent reconstruction functions (for the given size k).

It is known (Graepel et al., 2005) that for any $n, k \in \mathbb{N}$,

$$\mathcal{E}^o_{\mathrm{uc}}(n,k) \lesssim \sqrt{\frac{k\log(n)}{n}}.$$

In comparison with the above upper bound for permutation-invariant compression schemes, this bound has a factor $\log(n)$ in place of $\log(n/k)$.

Similarly, we can also define analogous quantities for agnostic learning excess risk guarantees. Specifically, in this context, an *ordered* compression function κ is a mapping from any sequence $z_1, \ldots, z_n$ in $\mathcal{Z}$ to an *ordered sequence* $S = (z_{i_1}, \ldots, z_{i_{k'}})$ for some $k' \leq k$ and $i_1, \ldots, i_{k'} \in \{1, \ldots, n\}$. Then, denoting $\hat{h}_n = \rho(\kappa(Z_{[n]}))$, define

$$\mathcal{E}^o_{\mathrm{ag}}(n,k,\rho,\kappa,P) = \mathbf{E}\left[R(\hat{h}_n;P) - \min_{h \in \mathcal{H}_{k,\rho}(Z_{[n]})} R(h;P)\right]$$

and then define

$$\mathcal{E}^o_{\mathrm{ag}}(n,k) = \sup_{\rho} \inf_{\kappa} \sup_{P} \mathcal{E}_{\mathrm{ag}}(n,k,\rho,\kappa,P),$$

where again P ranges over all probability measures on $\mathcal{Z}$ and ρ ranges over all order-dependent reconstruction functions (for the given size k), and where κ ranges over all ordered compression functions (for the given size k).

By the same standard argument involving empirical risk minimization, it remains true in this context that

$$\mathcal{E}^o_{\mathrm{ag}}(n,k) \leq 2\mathcal{E}^o_{\mathrm{uc}}(n,k) \tag{4}$$

and an immediate implication is then that any n, k has

$$\mathcal{E}^o_{\mathrm{ag}}(n,k) \lesssim \sqrt{\frac{k\log(n)}{n}}.$$

As above, it is interesting to ask whether the $\log(n)$ factor is required is necessary. Analogously to the order-invariant compression schemes above, we find that it is indeed necessary, as stated in the following theorem. Note that this lower bound for order-dependent compression schemes is slightly larger than that established above for order-independent compression schemes.

Theorem 4 *For any $n, k \in \mathbb{N}$ with $|\mathcal{X}| \geq n \geq ck\log(n)$,*

$$\mathcal{E}^o_{\mathrm{ag}}(n,k) \gtrsim \sqrt{\frac{k\log(n)}{n}}.$$

Together with (4), this has the following immediate implication.

Theorem 5 *For any $n, k \in \mathbb{N}$ with $|\mathcal{X}| \geq n \geq ck\log(n)$,*

$$\mathcal{E}^o_{\mathrm{uc}}(n,k) \gtrsim \sqrt{\frac{k\log(n)}{n}}.$$

As above, combining this with the known upper bound, this provides a tight characterization of the worst-case rate of uniform convergence for order-dependent agnostic sample compression schemes.

Corollary 6 *For any $n, k \in \mathbb{N}$ with $|\mathcal{X}| \geq n \geq ck\log(n)$,*

$$\mathcal{E}^o_{\text{ag}}(n,k) \asymp \sqrt{\frac{k\log(n)}{n}}$$

and

$$\mathcal{E}^o_{\text{uc}}(n,k) \asymp \sqrt{\frac{k\log(n)}{n}}.$$

We now present the proof of Theorem 4.

Proof [Proof of Theorem 4] The construction used in this proof is analogous to that from the proof of Theorem 1, and in fact is slightly simpler. Fix any $n, k \in \mathbb{N}$ with $|\mathcal{X}| \geq n \geq ck\log_2(n)$ for a sufficiently large numerical constant $c \geq 4$ (discussed below). The essential strategy here is the same as in the permutation-invariant compression schemes, in that we are constructing an embedded agnostic learning problem for a constructed VC class, but in this case the VC dimension will be larger: $k\log_2(m)$, with $m \approx n$. Specifically, let $m = 2^{\lfloor \log_2(n) \rfloor}$, and let $x_0, \ldots, x_{m-1}$ denote any m distinct elements of $\mathcal{X}$. For simplicity, suppose $\frac{m}{k\log_2(m)} \in \mathbb{N}$ (as before, the argument easily extends to the general case by introducing floor functions, and only the numerical constants change).

We break the space up into *blocks* as before, but now for each $t \in \{1, \ldots, k\}$ we let $B_t = \left\{(t-1)\frac{m}{k}, \ldots, t\frac{m}{k} - 1\right\}$, and for each $s \in \{1, \ldots, m/(k\log_2(m))\}$ we define a *sub-block*

$$B_{ts} = \left\{(t-1)\frac{m}{k} + (s-1)\log_2(m), \ldots, (t-1)\frac{m}{k} + s\log_2(m) - 1\right\}.$$

Thus, as before, a sub-block consists of $\log_2(m)$ indices, but now a block only contains m/k indices, and hence $\frac{m}{k\log_2(m)}$ sub-blocks. Now for $t \in \{1, \ldots, k\}$ and $i \in \{0, \ldots, m-1\}$, define a classifier $h_{t,i} : \mathcal{X} \to \mathcal{Y}$ with the property that, $\forall s \in \{1, \ldots, m/(k\log_2(m))\}$, $\forall r \in \{0, \ldots, \log_2(m) - 1\}$, for $j = (t-1)\frac{m}{k} + (s-1)\log_2(m) + r$,

$$h_{t,i}(x_j) = b_r(i),$$

where as above, $b_r(i)$ is the $(r+1)^{\text{th}}$ bit in the binary representation of i: i.e., $i = \sum_{\ell=0}^{\log_2(m)-1} b_\ell(i)2^\ell$, with $b_0(i), \ldots, b_{\log_2(m)-1}(i) \in \{0,1\}$. Thus, the index i encodes the prediction values for the points $\{x_\ell : \ell \in B_{ts}\}$ as the bits of i; this is slightly different from the $h_{t,i}$ functions we defined above, since i is already in $\{0, \ldots, m-1\}$ here, so there is no need to subtract anything from it.

Now we construct a reconstruction function ρ that outputs functions which again correspond to some such $h_{t,i}$ function within each block B_t, and which potentially uses a different bit pattern i for each t. Formally, for any $i_1, \ldots, i_k \in \{0, \ldots, m-1\}$ and any $y_1, \ldots, y_k \in \mathcal{Y}$, define $\rho(((x_{i_1}, y_1), \ldots, (x_{i_k}, y_k))) = \tilde{h}_{i_1,\ldots,i_k}$, where here $\tilde{h}_{i_1,\ldots,i_k} : \mathcal{X} \to \mathcal{Y}$ is any function satisfying the property that each $t \in \{1, \ldots, k\}$ and $j \in \{(t-1)m, \ldots, tm-1\}$ has $\tilde{h}_{i_1,\ldots,i_k}(x_j) = h_{t,i_t}(x_j)$: that is, the points x_{i_t} in the compression set are interpreted by the compression scheme as encoding the desired label sequence for sub-blocks B_{ts} in

the *bits* of i_t. Note that unlike the order-independent compression scheme construction, we do not require i_t to be in block B_t. Instead, we are able to distinguish which i_t to use to specify the h_{t,i_t} sub-predictor for block B_t simply using the *order* of the sequence $((x_{i_1}, y_1), \ldots, (x_{i_k}, y_k))$. For our purposes, $\tilde{h}_{i_1,\ldots,i_k}(x)$ may be defined arbitrarily for $x \in \mathcal{X} \setminus \{x_0, \ldots, x_{m-1}\}$. Again, since $\rho(((x_{i_1}, y_1), \ldots, (x_{i_k}, y_k)))$ is invariant to the $y_1, \ldots, y_k$ values, for brevity we will drop the y_i arguments and simply write $\rho((x_{i_1}, \ldots, x_{i_k}))$. For completeness, $\rho(S)$ should also be defined for sequences S of length strictly less than k, or sequences containing elements not in $\{x_0, \ldots, x_{m-1}\}$; for our purposes, in these cases, if k' of the elements in S *are* contained in $\{x_0, \ldots, x_{m-1}\}$, then enumerate them as $x_{i'_1}, \ldots, x_{i'_{k'}}$; then if $k' < k$, let $i'_{k'+1} = \cdots = i'_k = 0$, and finally define the output of $\rho(S)$ as $\tilde{h}_{i'_1,\ldots,i'_k}$: that is, it interprets the sub-sequence of points in S contained in $\{x_0, \ldots, x_{\log_2(m)-1}\}$ as the initial indices i_t, and fills in the rest of the indices up to i_k using 0's.

Now define a family of distributions $P^{(\sigma)}$, $\sigma = \{\sigma_{t,r}\}$, with $\sigma_{t,r} \in \{-1, 1\}$ for $t \in \{1, \ldots, k\}$ and $r \in \{0, \ldots, \log_2(m) - 1\}$, as follows. Every $P^{(\sigma)}$ has marginal P_X on $\mathcal{X}$ uniform on $x_0, \ldots, x_{m-1}$, and for each $j = (t-1)\frac{m}{k} + (s-1)\log_2(m) + r \in B_{ts}$ (for $t \in \{1, \ldots, k\}$, $s \in \{1, \ldots, m/(k\log_2(m))\}$, and $r \in \{0, \ldots, \log_2(m) - 1\}$) set $P^{(\sigma)}(Y = 1 | X = x_j) = \frac{1}{2} + \frac{\epsilon}{2}\sigma_{t,r}$, where

$$\epsilon = \sqrt{\frac{k \log_2(m)}{n}}.$$

Now let us suppose σ is chosen *randomly*, with $\sigma_{t,r}$ independent Uniform($\{-1, 1\}$). Then

$$\mathcal{E}^o_{\mathrm{ag}}(n, k) \geq \mathbf{E}\left[\inf_{\kappa} \mathcal{E}^o_{\mathrm{ag}}(n, k, \rho, \kappa, P^{(\sigma)})\right],$$

so that it suffices to lower-bound the expression on the right hand side.

For any $t \in \{1, \ldots, k\}$ and $r \in \{0, \ldots, \log_2(m) - 1\}$, denote

$$C_{t,r} = \{(t-1)\frac{m}{k} + (s-1)\log_2(m) + r : s \in \{1, \ldots, m/(k\log_2(m))\}\}.$$

Also define $\mathcal{H}^*_{k,\rho} = \{\rho((x_{i_1}, \ldots, x_{i_k})) : i_1, \ldots, i_k \in \{0, \ldots, m-1\}\}$, the space of all possible classifiers ρ can produce. As before, we are concerned both with constructing a lower bound on the excess risk of $\hat{h} = \rho(\kappa(Z_{[n]}))$ relative to $\min_{h \in \mathcal{H}^*_{k,\rho}} R(h; P^{(\sigma)})$ via a traditional VC lower bound argument, and also with upper-bounding $\min_{h \in \mathcal{H}_{k,\rho}(Z_{[n]})} R(h; P^{(\sigma)}) - \min_{h \in \mathcal{H}^*_{k,\rho}} R(h; P^{(\sigma)})$, so that the lower bound remains nearly valid for the excess risk of $\hat{h}$ relative to classifiers ρ can actually produce given sequences within this data set $Z_{[n]}$.

Fix any $t \in \{1, \ldots, k\}$ and let $P^{(\sigma)}_t$ denote the conditional distribution of $(X, Y) \sim P^{(\sigma)}$ given σ and the event that $X \in \{x_j : j \in B_t\}$. Also denote $\mathcal{H}^*_t = \{h_{t,i} : i \in \{0, \ldots, m-1\}\}$, $i^*_t = \mathrm{argmin}_{i \in \{0,\ldots,m-1\}} R(h_{t,i}; P^{(\sigma)}_t)$ $h^*_t = h_{t,i^*_t}$, and

$$\mathcal{H}_t(Z_{[n]}) = \{h_{t,i} : i \in \{0, \ldots, m-1\}, x_i \in \{X_1, \ldots, X_n\}\}.$$

As before, we are now interested in proving that any compression function κ results in $\hat{h} = \rho(\kappa(Z_{[n]}))$ with $\mathbf{E}[R(\hat{h}; P^{(\sigma)}_t) - R(h^*_t; P^{(\sigma)}_t)] \geq \epsilon/(8e^4)$, and also that $\mathbf{E}[\min_{h \in \mathcal{H}_t(Z_{[n]})} R(h; P^{(\sigma)}_t) - R(h^*_t; P^{(\sigma)}_t)] \leq \epsilon/(16e^4)$.

The first part proceeds nearly identically to the corresponding part in the proof of Theorem 1, with a few changes needed to convert to this scenario. For any $r \in \{0, \ldots, \log_2(m) - 1\}$, for any $j \in C_{t,r}$, note that $h_t^*(x_j) = \frac{\sigma_{t,r}+1}{2}$. Also, for any compression function κ, any $\hat{h}$ that $\rho(\kappa(Z_{[n]}))$ is capable of producing has $\hat{h}(x_j) = \hat{h}(x_{j'})$ for every $j, j' \in C_{t,r}$. In particular, if we let $\hat{i}_t \in \{0, \ldots, m-1\}$ be the index with $b_r(\hat{i}_t) = \hat{h}(x_{(t-1)(m/2)+r})$ for every $r \in \{0, \ldots, \log_2(m) - 1\}$, then $\hat{h}$ and $h_{t,\hat{i}_t}$ agree on every element of $\{x_j : j \in B_t\}$. This also implies

$$\begin{aligned} R(\hat{h}; P_t^{(\sigma)}) - R(h_t^*; P_t^{(\sigma)}) &= R(h_{t,\hat{i}_t}; P_t^{(\sigma)}) - R(h_t^*; P_t^{(\sigma)}) \\ &= \frac{1}{\log_2(m)} \sum_{r=0}^{\log_2(m)-1} \epsilon \mathbb{I}\left[b_r(\hat{i}_t) \neq \frac{\sigma_{t,r}+1}{2} \right]. \end{aligned}$$

Therefore, denoting by $n_{t,r} = |\{i \leq n : X_i \in \{x_j : j \in C_{t,r}\}\}|$, we have

$$\mathbf{E}[R(\hat{h}; P_t^{(\sigma)}) - R(h_t^*; P_t^{(\sigma)})] = \frac{\epsilon}{\log_2(m)} \sum_{r=0}^{\log_2(m)-1} \mathbf{E}\left[\mathbf{P}\left(b_r(\hat{i}_t) \neq \frac{\sigma_{t,r}+1}{2} \middle| n_{t,r} \right) \right].$$

For any $r \in \{0, \ldots, \log_2(m) - 1\}$, enumerate the $n_{t,r}$ random variables (X_i, Y_i) with $X_i \in \{x_j : j \in C_{t,r}\}$ as $(X_{i(r,1)}, Y_{i(r,1)}), \ldots, (X_{i(r,n_{t,r})}, Y_{i(r,n_{t,r})})$, and note that given $n_{t,r}$, the values $(Y_{i(r,1)}, \ldots, Y_{i(r,n_{t,r})})$ are a *sufficient statistic* for $\sigma_{t,r}$ (see Definition 2.4 of Schervish (1995)), and therefore (see Theorem 3.18 of Schervish (1995)) there exists a (randomized) decision rule $\hat{f}_{t,r}(Y_{i(r,1)}, \ldots, Y_{i(r,n_{t,r})})$ depending only on these variables and independent random bits such that

$$\mathbf{P}\left(b_r(\hat{i}_t) \neq \frac{\sigma_{t,r}+1}{2} \middle| n_{t,r} \right) = \mathbf{P}\left(\hat{f}_{t,r}(Y_{i(r,1)}, \ldots, Y_{i(r,n_{t,r})}) \neq \frac{\sigma_{t,r}+1}{2} \middle| n_{t,r} \right).$$

Furthermore, by Lemma 5.1 of Anthony and Bartlett (1999), we have

$$\mathbf{P}\left(\hat{f}_{t,r}(Y_{i(r,1)}, \ldots, Y_{i(r,n_{t,r})}) \neq \frac{\sigma_{t,r}+1}{2} \middle| n_{t,r} \right) > \frac{1}{8e} \exp\{-(8/3) n_{t,r} \epsilon^2\}.$$

Altogether, and combined with Jensen's inequality, we have that

$$\begin{aligned} &\mathbf{E}[R(\hat{h}; P_t^{(\sigma)}) - R(h_t^*; P_t^{(\sigma)})] \\ &\geq \frac{\epsilon}{8e \log_2(m)} \sum_{r=0}^{\log_2(m)-1} \mathbf{E}\left[\exp\{-(8/3) n_{t,r} \epsilon^2\}\right] \geq \frac{\epsilon}{8e \log_2(m)} \sum_{r=0}^{\log_2(m)-1} \exp\{-(8/3) \mathbf{E}[n_{t,r}] \epsilon^2\} \\ &= \frac{\epsilon}{8e \log_2(m)} \sum_{r=0}^{\log_2(m)-1} \exp\left\{ -(8/3) \frac{n}{k \log_2(m)} \epsilon^2 \right\} \geq \frac{\epsilon}{8e \log_2(m)} \sum_{r=0}^{\log_2(m)-1} e^{-(8/3)} \geq \frac{\epsilon}{8e^4}. \end{aligned}$$

Next, continuing on to the second part, for any $x \in \{x_i : i \in \{0, \ldots, m-1\}\}$, denote by $I(x)$ the index i such that $x = x_i$. Similarly to before, an i for which $h_{t,i}$ has minimal $R(h_{t,i}; P_t^{(\sigma)})$ among all $h_{t,i'} \in \mathcal{H}_t(Z_{[n]})$ can equivalently be defined as an i with minimal

$\sum_{j=0}^{\log_2(m)-1} \mathbb{I}[b_j(i) \neq b_j(i_t^*)]$ among all $i' \in \{I(X_1), \ldots, I(X_n)\}$, and furthermore, for such an i,

$$R(h_{t,i}; P_t^{(\sigma)}) - R(h_t^*; P_t^{(\sigma)}) = \frac{\epsilon}{\log_2(m)} \sum_{j=0}^{\log_2(m)-1} \mathbb{I}[b_j(i) \neq b_j(i_t^*)].$$

For any $i \in \{0, \ldots, m-1\}$, denote $\Delta_t(i) = \sum_{j=0}^{\log_2(m)-1} \mathbb{I}[b_j(i) \neq b_j(i_t^*)]$. It therefore suffices to prove an upper bound for the quantity

$$\frac{\epsilon}{\log_2(m)} \mathbf{E}\left[\min_{i \in \{I(X_1),\ldots,I(X_n)\}} \Delta_t(i)\right].$$

Define a random variable X with distribution P_X (recalling that this is uniform on $\{x_0, \ldots, x_{m-1}\}$). Then the conditional distribution of $\Delta_t(I(X))$ given σ is Binomial($\log_2(m), \frac{1}{2}$). Letting $q = 16e^4$, and supposing c is sufficiently large so that $q \leq (1/2)\log_2(m)$, following the argument from the analogous step in the proof of Theorem 1 (where an analysis is given that would apply to *any* Binomial($\log_2(m), \frac{1}{2}$) random variable) we have

$$\mathbf{P}\left(\Delta_t(I(X)) \leq \frac{1}{2q}\log_2(m) \middle| \sigma\right) \geq m^{(1/2q)\log_2(4q)-1},$$

which implies (still following similar derivations as in the proof of Theorem 1, except with n_t replaced by n)

$$\mathbf{P}\left(\min_{i \in \{I(X_1),\ldots,I(X_n)\}} \Delta_t(i) > \frac{1}{2q}\log_2(m) \middle| \sigma\right) \leq \exp\left\{-m^{(1/2q)\log_2(4q)-1} n\right\}.$$

By the law of total expectation and the fact that $R(h; P_t^{(\sigma)}) \leq 1$, we have

$$\mathbf{E}\left[\min_{h \in \mathcal{H}_t(Z_{[n]})} R(h; P_t^{(\sigma)}) - R(h_t^*; P_t^{(\sigma)})\right] \leq \frac{\epsilon}{2q} + \exp\left\{-m^{(1/2q)\log_2(4q)-1} n\right\}.$$

Then note that

$$\begin{aligned}
&\exp\left\{-m^{(1/2q)\log_2(4q)-1} n\right\} \leq \exp\left\{-m^{(1/2q)\log_2(4q)}\right\} \\
&= \left(\exp\left\{-(1/2q)\log_2(4q) m^{(1/2q)\log_2(4q)}\right\}\right)^{\frac{2q}{\log_2(4q)}} \\
&\leq \left(\frac{2q}{\log_2(4q)} \frac{1}{m^{(1/2q)\log_2(4q)}}\right)^{\frac{2q}{\log_2(4q)}} = \left(\frac{2q}{\log_2(4q)}\right)^{\frac{2q}{\log_2(4q)}} \frac{1}{m}.
\end{aligned}$$

Since this last expression shrinks strictly faster than the above specification of ϵ as a function of $n/(k\log(n))$, we may conclude that for a sufficiently large choice of the numerical constant c, this expression is smaller than $\frac{\epsilon}{32e^4}$. Therefore, we conclude that

$$\mathbf{E}\left[\min_{h \in \mathcal{H}_t(Z_{[n]})} R(h; P_t^{(\sigma)}) - R(h_t^*; P_t^{(\sigma)})\right] \leq \frac{\epsilon}{16e^4}.$$

These two parts combine to imply that

$$\mathbf{E}\left[R(\hat{h};P_t^{(\sigma)}) - \min_{h\in\mathcal{H}_t(Z_{[n]})} R(h;P_t^{(\sigma)})\right]$$
$$= \mathbf{E}\Big[R(\hat{h};P_t^{(\sigma)}) - R(h_t^*;P_t^{(\sigma)})\Big] - \mathbf{E}\left[\min_{h\in\mathcal{H}_t(Z_{[n]})} R(h;P_t^{(\sigma)}) - R(h_t^*;P_t^{(\sigma)})\right] \geq \frac{\epsilon}{16e^4}.$$

As a final step, we stitch together these lower bounds for the blocks to create a lower bound under the full distribution $P^{(\sigma)}$. Toward this end, note that any h has $R(h;P^{(\sigma)}) = \frac{1}{k}\sum_{t=1}^k R(h;P_t^{(\sigma)})$. Also note that, for this reconstruction function ρ, every $\tilde{h}_{i_1,\ldots,i_k}$ function with $i_1,\ldots,i_k \in \{I(X_1),\ldots,I(X_n)\}$ can be produced by ρ using an argument sequence S of at most k elements of $\{X_1,\ldots,X_n\}$: namely, $S=(x_{i_1},\ldots,x_{i_k})$, since each of these x_{i_t} are in $\{X_1,\ldots,X_n\}$ due to $i_t \in \{I(X_1),\ldots,I(X_n)\}$. Also note that $R(\tilde{h}_{i_1,\ldots,i_k};P_t^{(\sigma)}) = R(h_{t,i_t};P_t^{(\sigma)})$. Therefore,

$$\min_{h\in\mathcal{H}_{k,\rho}(Z_{[n]})} R(h;P^{(\sigma)}) \leq \min_{i_1,\ldots,i_k\in\{I(X_1),\ldots,I(X_n)\}} R(\tilde{h}_{i_1,\ldots,i_k};P^{(\sigma)})$$
$$= \min_{i_1,\ldots,i_k\in\{I(X_1),\ldots,I(X_n)\}} \frac{1}{k}\sum_{t=1}^k R(h_{t,i_t};P_t^{(\sigma)})$$
$$= \frac{1}{k}\sum_{t=1}^k \min_{i_t\in\{I(X_1),\ldots,I(X_n)\}} R(h_{t,i_t};P_t^{(\sigma)}) = \frac{1}{k}\sum_{t=1}^k \min_{h\in\mathcal{H}_t(Z_{[n]})} R(h;P_t^{(\sigma)}).$$

Thus, for any compression function κ, denoting $\hat{h}=\rho(\kappa(Z_{[n]}))$,

$$\mathbf{E}\left[R(\hat{h};P^{(\sigma)}) - \min_{h\in\mathcal{H}_{k,\rho}(Z_{[n]})} R(h;P^{(\sigma)})\right]$$
$$\geq \frac{1}{k}\sum_{t=1}^k \mathbf{E}\left[R(\hat{h};P_t^{(\sigma)}) - \min_{h\in\mathcal{H}_t(Z_{[n]})} R(h;P_t^{(\sigma)})\right] \geq \frac{1}{16e^4}\epsilon \gtrsim \sqrt{\frac{k\log(n)}{n}}.$$

■

References

Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge, 1999. ISBN 0-521-57353-X. doi: 10.1017/CBO9780511624216. URL http://dx.doi.org/10.1017/CBO9780511624216.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *J. Assoc. Comput. Mach.*, 36(4):929–965, 1989. ISSN 0004-5411.

Ofir David, Shay Moran, and Amir Yehudayoff. Supervised learning through the lens of compression. In *Advances in Neural Information Processing Systems*, pages 2784–2792, 2016.

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.

Sally Floyd and Manfred K. Warmuth. Sample compression, learnability, and the vapnik-chervonenkis dimension. *Machine Learning*, 21(3):269–304, 1995.

Lee-Ad Gottlieb, Aryeh Kontorovich, and Pinhas Nisnevitch. Near-optimal sample compression for nearest neighbors. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 370–378, 2014.

Thore Graepel, Ralf Herbrich, and John Shawe-Taylor. Pac-bayesian compression bounds on the prediction error of learning algorithms for classification. *Machine Learning*, 59 (1-2):55–76, 2005.

Steve Hanneke. The optimal sample complexity of PAC learning. *Journal of Machine Learning Research*, 17:38:1–38:15, 2016. URL `http://jmlr.org/papers/v17/15-389.html`.

Steve Hanneke, Aryeh Kontorovich, and Menachem Sadigurschi. Efficient conversion of learners to bounded sample compressors. 2018.

David Haussler. Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension. *J. Combin. Theory Ser. A*, 69(2):217–232, 1995.

David Haussler, Nick Littlestone, and Manfred K. Warmuth. Predicting {0,1}-functions on randomly drawn points. *Inf. Comput.*, 115(2):248–292, 1994. doi: 10.1006/inco.1994.1097.

Daniel M. Kane, Roi Livni, Shay Moran, and Amir Yehudayoff. On communication complexity of classification problems. *CoRR*, abs/1711.05893, 2017. URL `http://arxiv.org/abs/1711.05893`.

Vladimir I. Koltchinskii. On the central limit theorem for empirical measures. *Theory of Probability and Mathematical Statistics*, 24:71–82, 1981.

Aryeh Kontorovich and Iosif Pinelis. Exact lower bounds for the agnostic probably-approximately-correct (PAC) machine learning model. *CoRR*, abs/1606.08920, 2016. URL `http://arxiv.org/abs/1606.08920`.

Nick Littlestone and Manfred K. Warmuth. Relating data compression and learnability, unpublished. 1986.

Shay Moran and Amir Yehudayoff. Sample compression schemes for VC classes. *J. ACM*, 63(3):21:1–21:10, 2016. doi: 10.1145/2890490. URL `http://doi.acm.org/10.1145/2890490`.

David Pollard. A central limit theorem for empirical processes. *Journal of the Australian Mathematical Society*, 33(2):235–248, 1982.

Mark J. Schervish. *Theory of Statistics*. Springer-Verlag, 1995.

Michel Talagrand. Sharper bounds for gaussian and empirical processes. *Ann. Probab.*, 22(1):28–76, 01 1994. doi: 10.1214/aop/1176988847. URL http://dx.doi.org/10.1214/aop/1176988847.

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.

Nikita Zhivotovskiy. Optimal learning via local entropies and sample compression. In *COLT Conference on Learning Theory*, 2017.