ERRATUM TO "SHIFTING: ONE-INCLUSION MISTAKE BOUNDS AND SAMPLE COMPRESSION"

Benjamin I. P. Rubinstein[1], **Peter L. Bartlett**[1,2], **J. Hyam Rubinstein**[3]
[1]CS Division, UC Berkeley [2]Department of Statistics, UC Berkeley
[3]Department of Mathematics & Statistics, University of Melbourne

Abstract. H. Simon and B. Szörényi have found an error in the proof of Theorem 52 of "Shifting: One-Inclusion Mistake Bounds and Sample Compression" [3]. In this note we provide a corrected proof of a slightly weakened version of this theorem. Our new bound on the density of one-inclusion hypergraphs is again in terms of the capacity of the multilabel concept class. Simon and Szörényi have recently proved an alternate result in [4].

1. Multiclass Hypergraph Density Bound & Mistake Bound

This note is devoted to proving an upper-bound on one-inclusion hypergraph density, correcting a result stated in [3, Theorem 52]. We refer the reader to [3] for definitions related to one-inclusion hypergraphs and mistake bounds. The following definition, due to Ben-David *et al.* [1], did not appear in [3].

Definition 1. *Let $n > 0$ and $q > 1$ be integers, and let set $\mathcal{X}$ be an instance domain. Following the notation of* [3, Definition 5]*, let $\Psi_B = \{0, 1, \star\}^{\{0,\ldots,q-1\}}$ be a family of translation mappings from $\{0, \ldots, q-1\}$ to $\{0, 1, \star\}$. Then for any concept class $C \subseteq \{0, \ldots, q-1\}^{\mathcal{X}}$, the dimension*

$$\Psi_B\text{-dim}(C) \quad = \quad \sup\left\{n \mid \exists \mathbf{x} \in \mathcal{X}^n, \psi \in \Psi_B^n \ s.t.\ \{0,1\}^n \subseteq \psi(\Pi_{\mathbf{x}}(C)\right\} \ ,$$

is the largest number of points whose images under a vector of translation mappings are shattered.

The Ψ_B-dim is one out of many possible measures of multiclass capacity explored in [1]. The result stated in [3, Theorem 52] claimed to upper-bound one-inclusion hypergraph density by the Pollard pseudo-dimension Ψ_P-dim(C), which lower-bounds Ψ_B-dim(C). The following theorem is the main result of this note.

Theorem 2. *For any $C \subseteq \{0, \ldots, q-1\}^n$ the density of the one-inclusion hypergraph induced by C is upper-bounded by $\left(2^k - 1\right) \Psi_B\text{-dim}(C)$ where $k = \lceil \log_2 q \rceil$. In particular the density is always bounded by $(2q - 3)\Psi_B\text{-dim}(C)$.*

The recent result due to Simon and Szörényi upper-bounds one-inclusion hypergraph density by $\Psi_{GP}\text{-dim}'(C)$ which lower-bounds the pseudo-dimension and is thus stronger than Theorem 2 (see [4] for details). The proof of Theorem 2 is a careful reduction to the binary one-inclusion graph case. To simplify notation we initially assume that $q = 2^k$ for some $k \in \mathbb{N}$. The reduction proceeds by encoding the multiclass label set $\{0, \ldots, q-1\}$ as bit strings in $\{0,1\}^k$. We will construct this bijection $\phi : \{0, \ldots, q-1\} \to \{0,1\}^k$ and then apply it coordinate-wise to the class $C \subseteq \{0, \ldots, q-1\}^n$ to get binary class $\phi(C) \subseteq \{0,1\}^{nk}$ (abusing notation slightly). We will find it useful to introduce notation for the multiclass coordinate that a binary coordinate is mapped from: $part(i) = (i - 1 \bmod(k)) + 1$, which maps elements from $[nk]$ to $[n]$. An immediate consequence of Theorem 2 is the corresponding mistake bound for multiclass classification.

Corollary 3. *Consider any integer $q > 1$, set $\mathcal{X}$ and family of multiclass classifiers $\mathcal{F} \subseteq \{0, \ldots, q-1\}^{\mathcal{X}}$ on domain $\mathcal{X}$ with $\Psi_B\text{-dim}(\mathcal{F}) < \infty$. The multiclass one-inclusion prediction strategy* [3, Algorithm 1] *has worst-case expected risk bounded by $\hat{M}_{Q_{\mathcal{G},\mathcal{F}},\mathcal{F}}(n) \le (2^k - 1)\Psi_B\text{-dim}(\mathcal{F})/n$ for all sample sizes $n \in \mathbb{N}$, where $k = \lceil \log_2 q \rceil$.*

Submitted 9/18/2009.

1.1. **Proof of Density Bound.** We will now establish that the one-inclusion hypergraph density of C is bounded by the graph density of $\phi(C)$ (Lemma 4) which in turn (by the classic shifting proof) is bounded by $VC(C)$ (Lemma 5) which we will show is bounded by Ψ_B-dim(C) (Lemma 6).

Lemma 4. *Consider any $k \in \mathbb{N}$ and $q = 2^k$. For any class $C \subseteq \{0, \ldots, q-1\}^n$ there exists a bijective binary encoding $\phi : \{0, \ldots, q-1\} \to \{0,1\}^k$ such that the one-inclusion hypergraph density of C is bounded by the one-inclusion graph density of ϕ applied coordinate-wise to C: $dens(C) \leq \frac{2^k-1}{k} dens(\phi(C))$.*

Proof. Trivially the vertex-set cardinality is invariant under a one-to-one encoding. Let W denote the number of hyperedges induced by C, and let E denote the number of edges induced by $\phi^\star(C)$, where $\phi^\star$ is the selected encoding. We will show that $W \leq \frac{2^k-1}{k} E$ for some $\phi^\star$.

Note that any hyperedge in C must involve vertices with at least two labels in $\{0, \ldots, q-1\}$. The set of such label pairs $\mathcal{L}$ is clearly of size $q(q-1)/2$. Let $\mathcal{H}$ denote the set of hyperedges induced by C. We can cover $\mathcal{H}$ by subsets $\mathcal{H}_{ij} \subseteq \mathcal{H}$ consisting of all hyperedges containing labels i, j, for $0 \leq i \neq j \leq q-1$. Let h_{ij} denote the number of hyperedges in $\mathcal{H}_{ij}$. Under any encoding ϕ we say that a pair $\{i, j\} \subset \{0, \ldots, q-1\}$ is *encoding connected* if $\phi(i)$ and $\phi(j)$ are hamming-1 apart. We denote the set of label pairs that are ϕ-encoding connected by $sample(\phi) = \{\{i,j\} \in \mathcal{L} : \|\phi(i) - \phi(j)\|_1 = 1\}$. We claim that

$$\max_{\phi} \frac{\sum_{i,j \in sample(\phi)} h_{ij}}{\sum_{i,j \in \mathcal{L}} h_{ij}} \;\geq\; \mathbb{E}_{\phi \sim Unif}\left[\frac{\sum_{i,j \in sample(\phi)} h_{ij}}{\sum_{i,j \in \mathcal{L}} h_{ij}}\right] \;=\; \frac{k}{2^k - 1} . \tag{1.1}$$

The first inequality is true trivially. The equality follows by computing expectation, using the observation that the number of ordered sets of 2^k binary k-strings with the first two strings hamming-1 apart is $k2^k\left(2^{k-2}!\right)$.

$$\begin{aligned}
\frac{1}{2^k!}\sum_{\phi} \frac{\sum_{i,j \in sample(\phi)} h_{ij}}{\sum_{i,j \in \mathcal{L}} h_{ij}} &= \frac{1}{2^k!}\sum_{\phi} \frac{\sum_{i,j \in \mathcal{L}} \mathbf{1}\left[i,j \in sample(\phi)\right] h_{ij}}{\sum_{i,j \in \mathcal{L}} h_{ij}} \\
&= \frac{1}{2^k!} \frac{\sum_{i,j \in \mathcal{L}} h_{ij} \sum_{\phi} \mathbf{1}\left[i,j \in sample(\phi)\right]}{\sum_{i,j \in \mathcal{L}} h_{ij}} \\
&= \frac{1}{2^k!} \frac{\sum_{i,j \in \mathcal{L}} h_{ij} k 2^k (2^{k-2}!)}{\sum_{i,j \in \mathcal{L}} h_{ij}} \\
&= \frac{k}{2^k - 1} .
\end{aligned}$$

In particular we have inequality (1.1) holding for some maximizing encoding $\phi^\star$. Since $h_{ij} > 0$ for some $i, j \in \mathcal{L}$ (for non-empty hyperedge-set; the density bound is trivially true otherwise), we can invert the obtained inequality to get

$$W \;\leq\; \sum_{i,j \in \mathcal{L}} h_{ij} \;\leq\; \frac{2^k - 1}{k} \sum_{i,j \in sample(\phi^\star)} h_{ij} \;\leq\; \frac{2^k-1}{k} E .$$

□

Lemma 5 (Lemma 2 [2])**.** *For any concept class $V \subseteq \{0,1\}^n$, the one-inclusion graph density of V is no greater than $VC(V)$.*

Lemma 6. *Consider any $k \in \mathbb{N}$ and $q = 2^k$. For any binary encoding ϕ and class $C \subseteq \{0, \ldots, 2^k - 1\}^n$, $VC(\phi(C)) \leq k\Psi_B$-dim(C).*

Proof. Suppose $\phi(C)$ shatters index set $I \subseteq [nk]$. Let $I' = \{i \in I \mid \nexists j \in I, part(j) = part(i), j < i\}$. Notice that $|I| \leq k|I'|$. Construct a vector of multilabel to $\{0, 1, \star\}$ translations for the n coordinates of the multilabel class C. For each $i \in [n]$, define the i^{th} coordinate translation

$$\psi_i(x) \quad = \quad \begin{cases} \phi(x)_j\,, & \text{if } \exists j \in I',\, i = part(j) \\ \star\,, & \text{otherwise} \end{cases} \quad .$$

That is, if $i = part(j)$ for some $j \in I'$ (there can be only one by definition of I') then we set the i^{th} coordinate translation to be the j^{th} bit of the encoding: all labels in $\{0,..,2^k-1\}$ that have j^{th} bit as 0 translate to 0 and all labels that have j^{th} bit as 1 translate to 1. If the i^{th} coordinate does not encode to any coordinates that are shattered, then we arbitrarily translate all multiclass labels to $\star$.

Under the translations $(\psi_1,\ldots,\psi_n)$, C's image shatters $\{part(j) \mid j \in I'\}$. Thus $\Psi_B\text{-dim}(C) \geq |I'|$. □

Proof of Theorem 2. We first consider the case of $q = 2^k$. For any $C \subseteq \{0,\ldots,2^k-1\}^n$, there exists an encoding ϕ such that $dens(C) \leq \frac{2^k-1}{k} dens(\phi(C))$ by Lemma 4. Lemma 5 then establishes that $dens(\phi(C)) \leq VC(\phi(C))$ by shifting. Finally Lemma 6 states that $VC(\phi(C)) \leq k\Psi_B\text{-dim}(C)$. These inequalities combine to prove the claim. In the general case where $q \neq 2^k$ for all k, it suffices to embed C in $\{0,\ldots,2^{\lceil \log_2 q \rceil}\}^n$ and apply the previous argument. □

References

[1] Shai Ben-David, Nicolò Cesa-Bianchi, David Haussler, and Philip M. Long. Characterizations of learnability for classes of $\{0,\ldots,n\}$-valued functions. *Journal of Computer and System Sciences*, 50(1):74–86, 1995.

[2] David Haussler. Sphere packing numbers for subsets of the boolean n-cube with bounded vapnik-chervonenkis dimension. *Journal of Combinatorial Theory (A)*, 69(2):217–232, 1995.

[3] Benjamin I. P. Rubinstein, Peter L. Bartlett, and J. Hyam Rubinstein. Shifting: One-inclusion mistake bounds and sample compression. *Journal of Computer and System Sciences*, 75(1):37–59, 2009.

[4] Hans Ulrich Simon and Balázs Szörényi. One-inclusion hypergraph density revisited, 2009. submitted for publication.