Open problem: Direct Sums in Learning Theory

Steve Hanneke STEVE.HANNEKE@GMAIL.COM
Department of Computer Science, Purdue University

Shay Moran SMORAN@TECHNION.AC.IL
Departments of Mathematics, Computer Science, and Data and Decision Sciences, Technion, and Google Research

Tom Waknine TOM.WAKNINE@CAMPUS.TECHNION.AC.IL
Department of Mathematics, Technion

Editors: Shipra Agrawal and Aaron Roth

1. Introduction

In computer science, the term 'direct sum' refers to fundamental questions about the scaling of computational or information complexity with respect to multiple task instances. Consider an algorithmic task T and a computational resource C. For instance, T might be the task of computing a polynomial, with C representing the number of arithmetic operations required, or T could be a learning task with its sample complexity as C. The direct sum inquiry focuses on the cost of solving k separate instances of T, particularly how this aggregate cost compares to the resources needed for a single instance. Typically, the cost for multiple instances is at most k times the cost of one, since each can be handled independently.

However, there are intriguing scenarios where the total cost for k instances is less than this linear relationship. These cases suggest more efficient methods for simultaneously handling multiple instances of a task than addressing them one by one. As an example, consider an $n \times n$ matrix A and the objective of calculating its product with an input column vector x, where the computational resource C is the number of arithmetic operations. For a single vector x, it is easy to see that $\Theta(n^2)$ operations are necessary and sufficient. However, if instead of one input vector x, there are n input vectors $x_1, \ldots, x_n$ then one can do better than $n \times \Theta(n^2) = \Theta(n^3)$. Indeed, by arranging these n vectors as columns in an $n \times n$ matrix B, computing the product $A \cdot B$ is equivalent to solving the n products $A \cdot x_i$. This task can be accomplished using roughly $n^\omega \leq n^{2.37}$ arithmetic operations with fast matrix multiplication algorithms. Direct sum questions are well-studied in information theory and complexity theory. For more background we refer the reader to the thesis by Pankratov (2012) or the books by Wigderson (2019) and Rao and Yehudayoff (2020).

Direct Sum in Learning Theory. Natural direct sum questions can also be posed in learning theory. To formalize these, we use the notion of cartesian product of concept classes: consider two concept classes, $\mathcal{C}_1$ and $\mathcal{C}_2$, defined over domains $\mathcal{X}_1$ and $\mathcal{X}_2$, and label spaces $\mathcal{Y}_1$ and $\mathcal{Y}_2$ respectively. Their product, $\mathcal{C}_1 \otimes \mathcal{C}_2$, has domain $\mathcal{X}_1 \otimes \mathcal{X}_2$ and label space $\mathcal{Y}_1 \otimes \mathcal{Y}_2$. Each concept c in $\mathcal{C}_1 \otimes \mathcal{C}_2$ is parameterized by a pair of concepts $c_1 \in \mathcal{C}_1$ and $c_2 \in \mathcal{C}_2$, and is defined as $c((x_1, x_2)) = (c_1(x_1), c_2(x_2))$. Thus, learning c effectively means learning both c_1 and c_2 simultaneously. This definition naturally extends to direct sums of multiple concept classes allowing us to define $\bigotimes_{i=1}^{r} C_i$ where each C_i is a concept class. For a concept class $\mathcal{C}$ we denote $\mathcal{C}^r = \bigotimes_{i=1}^{r} \mathcal{C}$ its direct sum with itself r times.

2. Technical Background

We now present basic definitions from supervised classification in the PAC model. These definitions are stated in the slightly more general context of list learning, which will be useful later on. Let $\mathcal{X}$ denote the domain and $\mathcal{Y}$ denote the label space. A k-list function (or k-list concept) is a function $c : \mathcal{X} \to \binom{\mathcal{Y}}{k}$, where $\binom{\mathcal{Y}}{k}$ denotes the collection of all subsets of $\mathcal{Y}$ of size k. A k-list concept class $\mathcal{C} \subseteq \binom{\mathcal{Y}}{k}^{\mathcal{X}}$ is a set of k-list functions. Note that by identifying sets of size one with their single elements, 1-list concept classes correspond to standard concept classes.

A k-list learning rule is a map $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^* \to \binom{\mathcal{Y}}{k}^{\mathcal{X}}$, i.e. it gets a finite sequence of labeled examples as input and outputs a k-list function. A learning problem $\mathcal{D}$ is a distribution over $\mathcal{X} \times \mathcal{Y}$. The population loss of a k-list function c with respect to $\mathcal{D}$ is defined by $\mathtt{L}_{\mathcal{D}}(c) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[1_{y\notin c(x)}]$.

We quantify the learning rate of a given learning rule on a given learning problem using learning curves: The learning curve of a learner $\mathcal{A}$ with respect to a learning problem $\mathcal{D}$ is the sequence $\{\varepsilon_n(\mathcal{D}|\mathcal{A})\}_{n=1}^{\infty}$, where

$$\varepsilon_n(\mathcal{D}|\mathcal{A}) = \mathop{\mathbb{E}}_{S\sim\mathcal{D}^n}\Big[\mathtt{L}_{\mathcal{D}}\big(\mathcal{A}(S)\big)\Big].$$

In words, $\varepsilon_n(\mathcal{D}|\mathcal{A})$ is the expected error of the learner $\mathcal{A}$ on samples of size n drawn from the distribution $\mathcal{D}$. For a sequence S of labeled examples, the empirical loss of a k-list function c with respect to S is $\mathtt{L}_S(c) = \frac{1}{|S|}\sum_{(x,y)\in S} 1_{y\notin c(x)}$. A sequence $S \in (\mathcal{X} \times \mathcal{Y})^n$ is realizable by a k-list function c if $y \in c(x)$ for every $(x, y) \in S$. It is realizable by a concept class $\mathcal{C}$ if it is realizable by some concept $c \in \mathcal{C}$. A learning problem $\mathcal{D}$ is realizable by a concept class $\mathcal{C}$ if for any $n \in \mathbb{N}$, a random sample $S \sim \mathcal{D}^n$ is realizable by $\mathcal{C}$ with probability 1. We say that a concept class $\mathcal{C}$ is agnostically k-list learnable if there exists a k-list learning rule $\mathcal{A}$ and a sequence $\varepsilon_n \xrightarrow{n\to\infty} 0$ such that for every learning problem $\mathcal{D}$, $(\forall n) : \varepsilon_n(\mathcal{D}|A) \leq \inf_{c\in\mathcal{C}} \mathtt{L}_{\mathcal{D}}(c)+\varepsilon_n$. If the latter only holds for $\mathcal{C}$-realizable distributions then we say that $\mathcal{C}$ is k-list learnable in the realizable case. The k-list realizable PAC learning curve of a concept class $\mathcal{C}$ is defined as follows:

$$\varepsilon(n|\mathcal{C}) = \inf_{\mathcal{A}} \sup_{\mathcal{D}} \varepsilon_n(\mathcal{D}|A),$$

where the infimum is taken over all k-list learning rules $\mathcal{A}$ and the supremum over all distributions $\mathcal{D}$ that are realizable by $\mathcal{C}$.

3. Direct Sum Questions

3.1. Direct Sum of Learning Rates

One of the most natural questions regarding direct sums of learning problems is the following question: given two learning tasks, can we learn both of them in a faster way than learning each individuality? Perhaps the simplest case is of multiple instances of the same learning task. Let $\mathcal{C}$ be a concept class and recall that for $r \in \mathbb{N}$, the r'th power of $\mathcal{C}$ is denoted by $\mathcal{C}^r = \underbrace{\mathcal{C} \otimes \mathcal{C} \cdots \otimes \mathcal{C}}_{r \text{ times}}$.

How does the learning rate of $\mathcal{C}^r$ scale in terms of the learning rate of $\mathcal{C}$?

This problem can be investigated with respect to various formulations of 'learning rate', for example:

Open Question *Let $\mathcal{C} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a concept class,with a realizable PAC learning curve $\varepsilon(n|\mathcal{C})$, By a union bound, learning each component independently we have the following bound*

$$\varepsilon(n|\mathcal{C}^r) \leq r \cdot \varepsilon(n|\mathcal{C}).$$

Can this upper bound be asymptotically improved for some classes $\mathcal{C}$?

Another natural version of the above is assuming a fixed marginal distribution $\mathcal{D}$: Let $\mathcal{D}$ be a fixed distribution over the domain $\mathcal{X}$ and let $\mathcal{C}$ be a concept class. For any $c : \mathcal{X} \to \mathcal{Y}$ let $\mathcal{D}_c$ be the distribution in which $(x, y) \sim \mathcal{D}_c$ satisfies $x \sim \mathcal{D}$ and $y = c(x)$. Define the fixed-marginal learning curve $\varepsilon(n|\mathcal{D}, \mathcal{C})$ by

$$\varepsilon(n|\mathcal{D}, \mathcal{C}) = \inf_{\mathcal{A}} \sup_{c \in \mathcal{C}} \varepsilon_n(\mathcal{D}_c|\mathcal{A})$$

where the infimum is taken over all learning rules $\mathcal{A}$. Note that for any $c \in \mathcal{C}$ we have that $\mathcal{D}_c$ is a realizable distribution, hence $\varepsilon(n|\mathcal{D}, \mathcal{C}) \leq \varepsilon(n|\mathcal{C})$ always holds. For any $r > 0$, let $\mathcal{D}^r$ be the product measure over $\mathcal{X}^r$.

Open Question 1 *Similarly to the case of the PAC learning curve, a simple union bound will give the upper bound $\varepsilon(n|\mathcal{D}^r, \mathcal{C}^r) \leq r \cdot \varepsilon(n|\mathcal{D}, \mathcal{C})$. Can the upper bound be asymptotically improved for some concept classes $\mathcal{C}$ and marginal distributions $\mathcal{D}$?*

One can ask similar questions about agnostic learning curves and uniform convergence. However, in these cases the baseline additive upper bound does not apply. The reason is because these curves concern relative quantities (indeed, the agnostic learning curve measures the excess loss and uniform convergence curve measures the maximum difference between the empirical and population losses).

For instance, given a distribution $\mathcal{D}$ over the product space $(\mathcal{X}^2 \times \mathcal{Y}^2)$ defined with marginal distributions $\mathcal{D}_1, \mathcal{D}_2$ over $(\mathcal{X} \times \mathcal{Y})$ we have by the union bound that $\mathtt{L}_{\mathcal{D}}(h_1 \otimes h_2) \leq \mathtt{L}_{\mathcal{D}_1}(h_1) + \mathtt{L}_{\mathcal{D}_2}(h_2)$. Similarly if $S = \{\big((x_{i,1}, x_{i,2}), (y_{i,1}, y_{i,2})\big)\}_{i=1}^n$, letting $S_b = \{(x_{i,b}, y_{i,b})\}_{i=1}^n$, we have $\mathtt{L}_S(h_1 \otimes h_2) \leq \mathtt{L}_{S_1}(h_1) + \mathtt{L}_{S_2}(h_2)$. These bounds, however, do not allow us to bound the *difference* $|\mathtt{L}_D(h_1 \otimes h_2) - \mathtt{L}_S(h_1 \otimes h_2)|$ as needed to bound the uniform convergence rate. Define the uniform convergence rate of $\mathcal{C}$ by

$$\varepsilon_{\mathtt{UC}}(n|\mathcal{C}) = \sup_{\mathcal{D}} \mathop{\mathbb{E}}_{S \sim \mathcal{D}^n} [\sup_{h \in \mathcal{C}} |\mathtt{L}_{\mathcal{D}}(h) - \mathtt{L}_S(h)|]$$

where the supremum is over all distributions $\mathcal{D}$.

Open Question 2 *How does $\varepsilon_{\mathtt{UC}}(n|\mathcal{C}^r)$ scale as a function of $\varepsilon_{\mathtt{UC}}(n|\mathcal{C})$ and r?*

A similar phenomenon happens in the case of agnostic learning: define the agnostic learning curve of a concept class $\mathcal{C}$ by

$$\varepsilon_{\mathtt{agn}}(n|\mathcal{C}) = \inf_{\mathcal{A}} \sup_{\mathcal{D}} (\mathtt{L}_{\mathcal{D},n}(\mathcal{A}) - \mathtt{L}_{\mathcal{D}}(\mathcal{C})),$$

where the infimum is taken over all learning rules $\mathcal{A}$, the supremum is taken over all distributions, $\mathtt{L}_{\mathcal{D},n}(\mathcal{A}) = \mathbb{E}_{S \sim \mathcal{D}^n}[\mathtt{L}_{\mathcal{D}}(\mathcal{A}(S))]$, and $\mathtt{L}_{\mathcal{D}}(\mathcal{C}) = \inf_{c \in \mathcal{C}} \mathtt{L}_{\mathcal{D}}(c)$. Here again we do not have simple bounds to the agnostic learning curve of $\mathcal{C}^r$ in terms of the agnostic learning curve of $\mathcal{C}$.

Open Question 3 *Let $\varepsilon_{\mathtt{agn}}(n|\mathcal{C})$ be the agnostic PAC learning curve of $\mathcal{C}$. How does $\varepsilon_{\mathtt{agn}}(n|\mathcal{C}^r)$ scale as a function of $\varepsilon_{\mathtt{agn}}(n|\mathcal{C})$ and r?*

3.2. Direct Sum of Learnability Parameters

Another important resource in the context of list learning is the minimal list size k for which a given class $\mathcal{C}$ is k-list PAC learnable. This raises the following questions:

Open Question 4 *Let $\mathcal{C}_1, \mathcal{C}_2$ be concept classes and assume k_1 and k_2 are the minimal integers such that $\mathcal{C}_1$ is k_1-list PAC learnable and $\mathcal{C}_2$ is k_2-list PAC learnable. What is the minimal integer k such that $\mathcal{C}_1 \otimes \mathcal{C}_2$ is k-list PAC learnable? How does it scale as a function of k_1 and k_2.*

It is not hard to see that $k \leq k_1 \cdot k_2$ by just learning each component independently and taking all possible pairs of labels in the marginal lists. We also show that $k \geq (k_1 - 1) \cdot (k_2 - 1)$ (see Equation (1) below). However, it remains open to determine tight bounds on k.

We raise the parallel question regarding compressibility:

Open Question 5 *Let $\mathcal{C}_1, \mathcal{C}_2$ be concept classes and assume k_1 and k_2 are the minimal integers such that $\mathcal{C}_1$ is k_1-list compressible and $\mathcal{C}_2$ is k_2-list compressible. What is the minimal integer k such that $\mathcal{C}_1 \otimes \mathcal{C}_2$ is k-list PAC learnable? How does it scale as a function of k_1 and k_2.*

A natural way to approach questions such as the ones above and in Section 3 is by analyzing combinatorial parameters that capture the corresponding resources.

Open Question 6 *Let $\mathcal{F}, \mathcal{G}$ be concept classes, and let $\mathtt{dim}(\cdot)$ be either the Graph dimension, the Natarajan dimension, the Littlestone dimension, or the Daniely-Shwartz dimension. How does $\mathtt{dim}(\mathcal{F} \otimes \mathcal{G})$ scale in terms of $\mathtt{dim}(\mathcal{F})$ and $\mathtt{dim}(\mathcal{G})$?*

We next state some preliminary results whose proofs can be found in Hanneke, Moran, and Waknine (2024).

Proposition 1 *Let $\mathtt{d_N}(\cdot)$ be the Natarajan dimension, and let $\mathcal{F}$ and $\mathcal{G}$ be concept classes. Then,*

$$\mathtt{d_N}(\mathcal{F}) + \mathtt{d_N}(\mathcal{G}) - 2 \leq \mathtt{d_N}(\mathcal{F} \otimes \mathcal{G}) \leq \mathtt{d_N}(\mathcal{F}) + \mathtt{d_N}(\mathcal{G}).$$

Proposition 2 *Let $\mathtt{LS}(\cdot)$ be the Littlestone dimension, then for any $\mathcal{F}, \mathcal{G}$ concept classes we have*

$$\mathtt{LS}(\mathcal{F} \otimes \mathcal{G}) = \mathtt{LS}(\mathcal{F}) + \mathtt{LS}(\mathcal{G})$$

The following lemma utilizes the concept of the Daniely-Shwartz (DS) dimension, which characterizes PAC learnability in the multiclass setting. This DS dimension-based characterization extends the VC dimension-based characterization used in the binary case, see Brukhim et al. (2022); Charikar and Pabbaraju (2022).

Proposition 3 *Let $\mathcal{F}, \mathcal{G}$ be partial function classes, and let $k, k' \geq 1$. Denote $\mathtt{DS}_k(\mathcal{C})$ the k-Daniely-Shwartz (DS) dimension of a concept class $\mathcal{C}$. Then,*

1. $\mathtt{DS}_{k \cdot k'}(\mathcal{F} \otimes \mathcal{G}) \geq \min\left(\mathtt{DS}_k(\mathcal{F}), \mathtt{DS}_{k'}(\mathcal{G})\right)$.
2. $\mathtt{DS}_{\min(k,k')}(\mathcal{F} \otimes \mathcal{G}) \geq \mathtt{DS}_k(\mathcal{F}) + \mathtt{DS}_{k'}(\mathcal{G}) - 1$.

Note that Lemma 3 has direct implications relevant to Open Question 4. Specifically, it implies that if $\mathcal{F}$ is not k-list learnable and $\mathcal{G}$ is not k'-list learnable then $\mathcal{F} \otimes \mathcal{G}$ is not $k \cdot k'$-list learnable. Conversely we know that if $\mathcal{F}$ is k-list learnable and $\mathcal{G}$ is k' list learnable then $\mathcal{F} \otimes \mathcal{G}$ is $k \cdot k'$-list learnable. Thus, letting $K(\mathcal{C})$ denote the minimal k such that a concept class $\mathcal{C}$ is k-list learnable (or infinity if there is no such k) we can summarize the above as

$$(K(\mathcal{F}) - 1)(K(\mathcal{G}) - 1) \leq K(\mathcal{F} \otimes \mathcal{G}) \leq K(\mathcal{F}) \cdot K(\mathcal{G}) \tag{1}$$

We may also ask similar questions about compressibility, an answer to which would be relevant to Open Question 5.

Lemma 3 also has implications to Open Question 3.1 for (1-list) PAC learnable classes. Indeed, let $\mathcal{C}$ be a PAC learnable class. Then, by Item 2 in Lemma 3, it follows that $\mathtt{DS}_1(\mathcal{C}^r) \geq r\mathtt{DS}_1(\mathcal{C})$. Hence, since the Daniely-Shwartz dimension lower bounds the PAC learning curve (Charikar and Pabbaraju, 2022) we get

$$\varepsilon(n|\mathcal{C}^r) \geq \frac{\mathtt{DS}_1(\mathcal{C}^r)}{n} \geq \frac{r \cdot \mathtt{DS}_1(\mathcal{C}) - r}{n}.$$

Thus, if it turns out that the realizable PAC learning curve is $\Theta\left(\frac{\mathtt{DS}_1}{n}\right)$ then the above in combination with the naive union bound argument mentioned in Open Question 3.1 would answer this question up to universal multiplicative constants.

References

Nataly Brukhim, Daniel Carmon, Irit Dinur, Shay Moran, and Amir Yehudayoff. A characterization of multiclass learnability. In *63rd IEEE Annual Symposium on Foundations of Computer Science, FOCS 2022, Denver, CO, USA, October 31 - November 3, 2022*, pages 943–955. IEEE, 2022. doi: 10.1109/FOCS54457.2022.00093. URL https://doi.org/10.1109/FOCS54457.2022.00093.

Moses Charikar and Chirag Pabbaraju. A characterization of list learnability. *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, 2022. URL https://api.semanticscholar.org/CorpusID:253420830.

Steve Hanneke, Shay Moran, and Tom Waknine. List sample compression and uniform convergence. *CoRR*, abs/2403.10889, 2024. doi: 10.48550/ARXIV.2403.10889. URL https://doi.org/10.48550/arXiv.2403.10889.

Denis Pankratov. Direct sum questions in classical communication complexity. *Master's thesis, University of Chicago*, 2012.

Anup Rao and Amir Yehudayoff. *Bibliography*, page 244–249. Cambridge University Press, 2020.

A. Wigderson. *Mathematics and Computation: A Theory Revolutionizing Technology and Science*. Princeton University Press, 2019. ISBN 9780691189130. URL https://books.google.co.il/books?id=-WCqDwAAQBAJ.