High-arity Sample Compression

Leonardo N. Coregliano
Department of Mathematics
University of Chicago
Chicago, IL 60637
lenacore@uchicago.edu

William Opich
Department of Mathematics
University of Chicago
Chicago, IL 60637
wopich@uchicago.edu

Abstract

Recently, a series of works have started studying variations of concepts from learning theory for product spaces, which can be collected under the name high-arity learning theory. In this work, we consider a high-arity variant of sample compression schemes and we prove that the existence of a high-arity sample compression scheme of non-trivial quality implies high-arity PAC learnability.

1 Introduction

Given a family $\mathcal{H}$ of subsets of some fixed X, can we approximately recover a hidden set $F \in \mathcal{H}$ from its behavior on a finite random i.i.d. sample from a distribution on X? Classes that admit such a procedure are called probably approximately correctly (PAC) learnable.

An a priori unrelated question about the family $\mathcal{H}$ is: given a sample $x_1, \dots, x_m \in X$ along with the behavior of some $F \in \mathcal{H}$ on $x_1, \dots, x_m$, is it possible to compress this information by selecting a small subsample $x_{i_1}, \dots, x_{i_s}$ so that it is possible to perfectly retrieve F from its behavior only on the subsample $x_{i_1}, \dots, x_{i_m}$ along with very little extra information? Procedures that achieve this compression and retrieval are called sample compression schemes.

As it turns out [see Littlestone and Warmuth, David, Moran, and Yehudayoff, 2016, Moran and Yehudayoff, 2016], PAC learnability of a class is equivalent to the existence of a sample compression scheme that is non-trivial in the sense that the subsample (along with the extra information) is of size $o(m)$. In turn, sample compression also enjoys a boosting phenomenon in the sense that when a non-trivial sample compression scheme exists, then there is also a sample compression scheme of size $O(\ln m \cdot \ln \ln m)$. This puts sample compression into the list of the many learning theory notions that are equivalent to finiteness of the Vapnik–Červonenkis (VC) dimension [see e.g. Shalev-Shwartz and Ben-David, 2014, Theorem 6.7, for a non-comprehensive list as starting point].

Recently, there has been a series of works [Kobayashi, 2015, Kuriyama and Takeuchi, 2015, Livni and Mansour, 2019a,b, Coregliano and Malliaris, 2024, 2025a,b, Holzman, Moran, and Shlimovich, 2026] studying variations of concepts from learning theory for product spaces, which can be collected under the name "high-arity learning theory" (a term coined by Coregliano and Malliaris [2024], who started its systematic study). The overall philosophy of high-arity learning theory is that when we replace the underlying set X by a product space $X_1 \times \cdots \times X_k$ and require the learning theory concept to respect the product space structure, then more classes satisfy the high-arity learning concept than the classic learning concept. To illustrate this high-arity learning theory principle, introduce some of the high-arity concepts and frame the current work, let us briefly describe some of the results that appear in the works that are more closely related to the current one (see Section 1.1 for a brief description of the others):

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

- Kobayashi [2015], Kuriyama and Takeuchi [2015], motivated by the model-theoretic notion of k-dependence, introduced a high-arity notion of learning called PAC_k-learning in which the learner has oracle access to $(k-1)$-ary slices of a k-ary hypothesis. This is a very weak form of learning (it is implied by all other notions of high-arity learning below) and is not the subject of this work.
- Livni and Mansour [2019a,b] studied discriminating algorithms in terms of IPM distance. For this, they introduced the graph-VC dimension (gVC) of such a class $\mathcal{H}$ as the maximum VC dimension of the classes of slices of $\mathcal{H}$, that is, once we fix values for $k-1$ coordinates, this produces a slice class of sets on the last coordinate, of which one can compute the VC dimension and maximize over the choice of fixed values.

  They also introduced high-arity uniform convergence property of a class $\mathcal{H}$ of subsets of $X_1 \times \cdots \times X_k$ as the property that for all distributions μ_i over X_i, if we sample $m = m(\varepsilon, \delta)$ i.i.d. points $\boldsymbol{x}_1^i, \ldots, \boldsymbol{x}_m^i$ from each μ_i, then with probability at least $1-\delta$, the sample will be ε-representative for $\mathcal{H}$ in the sense that for every $F \in \mathcal{H}$, the proportion of k-tuples of the form $(\boldsymbol{x}_{j_1}^1, \ldots, \boldsymbol{x}_{j_k}^k)$ that belong to F is ε-close to the $\bigotimes_{i=1}^k \mu_i$-measure of F.

  Finally, by leveraging classic PAC learning theory with a clever inductive argument, they proved that high-arity uniform convergence property is equivalent to finiteness of gVC dimension and further equivalent to the existence of discriminating algorithms in terms of IPM distance.
- Coregliano and Malliaris [2024, 2025a] introduced high-arity PAC learnability and was the first work to systematically study several of the high-arity notions. This work also identified a phenomenon exclusive to high-arity theory: the presence of two variants of each high-arity concept called the partite and non-partite; with the former being slightly simpler and being the setting of the works of Livni and Mansour [2019a,b].

  For partite k-PAC learnability, we have a class $\mathcal{H}$ of functions of the form $X_1 \times \cdots \times X_k \to Y$ and we want an algorithm $\mathcal{A}$ such that for all distributions μ_i over X_i, if we sample $m = m(\varepsilon, \delta)$ i.i.d. points $\boldsymbol{x}_1^i, \ldots, \boldsymbol{x}_m^i$ from each μ_i and give $\mathcal{A}$ these $k \cdot m$ points along with the m^k values $F(\boldsymbol{x}_{j_1}^1, \ldots, \boldsymbol{x}_{j_k}^k)$ on every k-tuple $(j_1, \ldots, j_k)$, then with probability at least $1-\delta$, the algorithm $\mathcal{A}$ outputs some H that is ε-close in $\bigotimes_{i=1}^k \mu_i$-measure to F (see Definition 2.1 for a formal statement).

  For non-partite k-PAC learnability, we instead assume that $X_1 = \cdots = X_k$ and we only require $\mathcal{A}$ to handle the case in which $\mu_1 = \cdots = \mu_k$ (see Definition A.2 for a formal statement). For a non-partite class $\mathcal{H}$, we can certainly interpret it as a partite class $\mathcal{H}^{k\text{-part}}$ by simply ignoring the fact that $X_1 = \cdots = X_k$ and it is easy to see that partite k-PAC learnability of $\mathcal{H}^{k\text{-part}}$ implies non-partite k-PAC learnability, but one of the main results of Coregliano and Malliaris [2024] is that, provided the label set Y is finite, these are in fact equivalent, and both characterized by finiteness of VCN_k dimension.

  Coregliano and Malliaris [2024, 2025a] proved that k-PAC learnability (both partite and non-partite) is also equivalent to finiteness of gVC in the binary label case, and in the more general setting of finite labels is equivalent to an analogous dimension, which they called the Vapnik–Červonenkis–Natarajan k-dimension (VCN_k), which is defined analogously to gVC as the maximum Natarajan dimension of the slices.

  This work also introduced a (quite technical) notion of high-arity agnostic learnability based on exchangeability theory, which we will briefly discuss in Section 3.2. This agnostic notion was crucial in proving on general grounds that non-partite agnostic learnability notions are equivalent to their partite counterparts.
- Holzman, Moran, and Shlimovich [2026] showed that partite k-PAC learnability can be upgraded to handle measures that are not product measures, but are in a sense not far from them. More specifically, suppose $\mathcal{P}$ is a collection of distributions on $X_1 \times \cdots \times X_k$ that are uniformly absolutely continuous with respect to their marginals in the sense that for every $\alpha > 0$, there exists $\beta > 0$ such that for every $\mu \in \mathcal{P}$ and every measurable $B \subseteq X_1 \times \cdots \times X_k$ with $\mu(B) \geq \alpha$, then $(\bigotimes_{i=1}^k \mu_i)(B) \geq \beta$, where

μ_i is the marginal of μ on X_i. In this work, finiteness of gVC dimension (which is given the name "linear VC dimension" here) is equivalent to learnability (in the usual classic PAC sense) versus all collections of distributions that are uniformly absolutely continuous with respect to their marginals (but, as expected, the learning guarantee depends on the quality of the modulus of uniform absolute continuity, i.e., the function $\beta = \beta(\alpha)$).

In this work, we introduce the notion of high-arity sample compression (see Definitions 2.2 and A.3) and we prove that both in the partite and non-partite settings, high-arity sample compression implies high-arity PAC learnability (Theorems 2.4 and A.5). As is typical with high-arity concepts, the main challenge is the fact that labels in high-arity no longer are i.i.d., which precludes us from applying concentration estimates such as Chernoff or Bernstein Inequalities. However, we show that both in the partite and non-partite setting, we can appropriately define a martingale that allows us to use Azuma's Inequality instead. We briefly discuss in Section 3.1 the difficulties in proving the converse implication (i.e., that high-arity PAC learnability implies high-arity sample compression).

1.1 Other related work

High-arity learning is much more complex than classical theory. In particular, even for something as simple as PAC learning, the big picture is that there should be different levels of learnability depending on the particular information that is received by the learner. For example, the PAC_k learnability model introduced by Kobayashi [2015], in which the learner also receives oracles for slices of the sample; Kuriyama and Takeuchi [2015] showed that PAC_k learnability is governed by a combinatorial dimension connected to the model-theoretic notion of k-dependence introduced by Shelah [2014, § (H)]. Chernikov, Palacin, and Takeuchi [2019] showed that the same notion of k-dependence is what governs the asymptotic drop via a high-arity analogue of the Sauer–Shelah–Perles Lemma [Sauer, 1972, Shelah, 1972, Perles, 1972] in the number of possible patterns on a grid of size m from m^k to $m^{k-\Omega(1)}$. In turn, such a drop in the number of possible patterns was then used by Coregliano and Malliaris [2025b] to characterize when one can perform matrix-completion in Netflix-like problems.

In a different direction, there is a strong connection between combinatorial dimensions from learning theory to the tamer Szemerédi Regularity Lemmas. In particular, Alon, Fischer, and Newman [2007], Lovász and Szegedy [2010] showed that the VC dimension characterizes classes of graphs that admit regularity partitions that are homogeneous and Fox, Pach, and Suk [2019], Terry [2024b] showed that VC dimension also characterizes classes of hypergraphs that admit a regularity partition into polynomially many parts. On the high-arity side, the gVC dimension (known under other names such as slicewise VC dimension in the combinatorial community) was shown to characterize [Terry and Wolf, 2022, Chernikov and Towsner, 2020] classes of hypergraphs that admit homogeneous (vertex) partitions and classes of hypergraphs that admit weak regularity lemmas into exponentially many parts [Terry, 2024a,b, Gishboliner, Shapira, and Wigderson, 2025]. Finally, Terry and Wolf [2022], Chernikov and Towsner [2020] showed that the aforementioned k-dependence characterizes hypergraph classes that admit a 0/1-valued strong regularity lemma. Most of these connections happen via Haussler packing property [see Haussler, 1995] and a high-arity version of it [see Chernikov and Towsner, 2020, Coregliano and Malliaris, 2025a].

1.2 General notation

The set of non-negative integers is denoted $\mathbb{N}$ and the set of positive integers is denoted $\mathbb{N}_+ \stackrel{\text{def}}{=} \mathbb{N} \setminus \{0\}$. For a finite set V and $k \in \mathbb{N}$, we let $\binom{V}{k} \stackrel{\text{def}}{=} \{U \subseteq V \mid |U| = k\}$ be the set of subsets of V of size k and we let $(V)_k$ be the set of injective functions of the form $[k] \to V$, where $[k] \stackrel{\text{def}}{=} \{1, \dots, k\}$. Given further $n \in \mathbb{N}$, we let $(n)_k \stackrel{\text{def}}{=} n(n-1)\cdots(n-k+1)$ denote the falling factorial (so $|(V)_k| = (|V|)_k$). For the particular case of $([k])_k$, we use the more standard notation S_k (as $([k])_k$ is the set of permutations on $[k]$). For a measurable space $\Omega = (X, \mathcal{B})$, we let $\Pr(\Omega)$ be the set of probability measures on Ω. Countable sets will always be equipped with the discrete σ-algebra.

2 Partite setting

In this section we cover the partite case (the non-partite case is deferred to Appendix A.2 as it uses analogous ideas, but contains small technicalities that are not present in the partite case), that is, the case in which our hypotheses classes are collections of functions of the form $F\colon X_1 \times \cdots \times X_k \to Y$, k-PAC learning notions are with respect to product measures $\bigotimes_{i=1}^k \mu_i$ and samples are thought of as having m points $x_1^j, \ldots, x_m^j$ from each X_j as well as labels $F(x_{\alpha_1}^1, \ldots, x_{\alpha_k}^k) \in Y$ for each k-tuple $\alpha \in [m]^k$.

We use a simplified version of the notation of Coregliano and Malliaris [2024] (the main difference is that since we do not use either agnostic or higher-order variables, which allows us to drastically simplify the notation; however, we briefly comment on how our proofs adapt to these settings in Section 3.2). We also encourage the reader to consult Example 2.3 below as they read the definitions.

Definition 2.1 (Partite setting)**.** Let $k \in \mathbb{N}_+$, let $\Omega = (\Omega_i)_{i=1}^k$ be a k-tuple of non-empty standard Borel spaces $\Omega_i = (X_i, \mathcal{B}_i)$, let $\Lambda = (Y, \mathcal{B}')$ be a non-empty standard Borel space and let $m \in \mathbb{N}$.

1. A *k-partite unlabeled sample of size m* with respect to Ω is an element of $\prod_{i=1}^k X_i^m$. A *k-partite labeled sample of size m* with respect to Ω and Λ is an element (x, y) of $(\prod_{i=1}^k X_i^m) \times Y^{[m]^k}$; we think of the coordinate y_α indexed by some k-tuple $\alpha \in [m]^k$ as the label on $x_{\alpha_1}, \ldots, x_{\alpha_k}$. This concept is formalized as follows: for every $\alpha \in [m]^k$, we contra-variantly define maps $\alpha^*\colon \prod_{i=1}^k X_i^m \to \prod_{i=1}^k X_i$ and $\alpha^*\colon Y^{[m]^k} \to Y$ that select the k-tuple indexed by α and its label respectively via
$$\alpha^*(x)_i \stackrel{\text{def}}{=} x_{\alpha_i} \qquad \left(x \in \prod_{j=1}^k X_j^m, i \in [k]\right),$$
$$\alpha^*(y) \stackrel{\text{def}}{=} y_\alpha, \qquad \left(y \in Y^{[m]^k}\right).$$

2. With a slight abuse of notation, we denote by $\Pr(\Omega)$ the set of k-tuples $\mu = (\mu_i)_{i=1}^k$, where $\mu_i \in \Pr(\Omega_i)$ is a probability measure on Ω_i. Given further $m \in \mathbb{N}$, we let $\mu^m \stackrel{\text{def}}{=} \bigotimes_{i=1}^k \mu_i^m$ denote the product probability measure on the set $\prod_{i=1}^k X_i^m$ of k-partite unlabeled samples of size m that is m copies of each measure μ_i. With a small abuse of notation, we view μ^1 also as a measure over $\prod_{i=1}^k X_i$ via the natural identification with $\prod_{i=1}^k X_i^1$.

3. A *k-partite hypothesis* from Ω to Λ is a measurable function $F\colon \prod_{i=1}^k X_i \to Y$. The set of k-partite hypothesis from Ω to Λ is denoted $\mathcal{F}_k(\Omega, \Lambda)$.

   Given a k-partite hypothesis $F \in \mathcal{F}_k(\Omega, \Lambda)$ and $m \in \mathbb{N}$, we define the function $F_m^*\colon \prod_{i=1}^k X_i^m \to Y^{[m]^k}$ that when given an unlabeled sample of size m, provides the labels of F on the k-tuples of the sample; in a formula:
$$F_m^*(x)_\alpha \stackrel{\text{def}}{=} F\big(\alpha^*(x)\big) \qquad \left(x \in \prod_{i=1}^k X_i^m, \alpha \in [m]^k\right).$$

4. A *k-partite hypothesis class* is a set $\mathcal{H} \subseteq \mathcal{F}_k(\Omega, \Lambda)$ of k-partite hypotheses that is further equipped with a σ-algebra such that:

   - the evaluation map $\mathrm{ev}\colon \mathcal{H} \times \prod_{i=1}^k X_i \to Y$ given by $\mathrm{ev}(H, x) \stackrel{\text{def}}{=} H(x)$ is measurable;
   - for every $H \in \mathcal{H}$, the singleton $\{H\}$ is measurable;
   - for every Borel space Υ and every measurable set $A \subseteq \mathcal{H} \times \Upsilon$, the projection of A onto Υ, i.e., the set $\{\upsilon \in \Upsilon \mid \exists H \in \mathcal{H}, (H, \upsilon) \in A\}$ is universally measurable[1].

[1]The reader unfamiliar with these technical measurability conditions can simply assume that $\mathcal{H}$ is equipped with a σ-algebra that makes it a standard Borel space as it will imply this condition. The reader that wants

5. A *k-partite loss function* over Λ is a measurable function $\ell\colon (\prod_{i=1}^k X_i)\times Y\times Y \to \mathbb{R}_{\geq 0}$ that when given a k-tuple x and two labels y and y' assigns a value $\ell(x,y,y')$ that is the penalty of guessing y on the tuple x when the correct label was y'.

   We define $\|\ell\|_\infty \stackrel{\text{def}}{=} \sup_{x\in\prod_{i=1}^k X_i, y,y'\in Y} \ell(x,y,y')$ and we say that ℓ is *bounded* if $\|\ell\|_\infty < \infty$. Given further $\mu \in \Pr(\Omega)$ and k-partite hypotheses F and H, the *total loss* of H with respect to μ, F and ℓ is $L_{\mu,F,\ell}(H) \stackrel{\text{def}}{=} \mathbb{E}_{\boldsymbol{x}\sim\mu^1}[\ell(\boldsymbol{x}, H(\boldsymbol{x}), F(\boldsymbol{x}))]$.

   We say that F is *realizable* in a k-partite hypothesis class $\mathcal{H} \subseteq \mathcal{F}_k(\Omega,\Lambda)$ with respect to μ and ℓ if $\inf_{H\in\mathcal{H}} L_{\mu,F,\ell}(H) = 0$.

6. For a k-partite hypothesis class $\mathcal{H}' \subseteq \mathcal{F}_k(\Omega,\Lambda)$, a *(k-partite) learning algorithm* with outputs in $\mathcal{H}'$ is a measurable function

$$\mathcal{A}\colon \bigcup_{m\in\mathbb{N}} \left(\left(\prod_{i=1}^k X_i^m \right) \times Y^{[m]^k} \right) \to \mathcal{H}'.$$

   We say that a k-partite hypothesis class $\mathcal{H} \subseteq \mathcal{F}_k(\Omega,\Lambda)$ is *improperly k-PAC learnable* with respect to a k-partite loss function $\ell\colon (\prod_{i=1}^k X_i)\times Y\times Y \to \mathbb{R}_{\geq 0}$ if there exists a learning algorithm $\mathcal{A}$ with outputs in some k-partite hypothesis class $\mathcal{H}'$ and a function $m^{\text{PAC}}_{\mathcal{H},\ell,\mathcal{A}}\colon (0,1)^2 \to \mathbb{R}_{\geq 0}$ such that for every $\varepsilon,\delta\in(0,1)$, every $\mu\in\Pr(\Omega)$, every $F\in\mathcal{F}_k(\Omega,\Lambda)$ that is realizable in $\mathcal{H}$ with respect to μ and ℓ and every integer $m \geq m^{\text{PAC}}_{\mathcal{H},\ell,\mathcal{A}}(\varepsilon,\delta)$, we have

$$\mathbb{P}_{\boldsymbol{x}\sim\mu^m}\left[L_{\mu,F,\ell}\Big(\mathcal{A}\big(\boldsymbol{x}, F^*_m(\boldsymbol{x})\big)\Big) \leq \varepsilon \right] \geq 1-\delta.$$

   In plain English: when given a random sample of size m drawn from μ^m labeled according to some realizable F, the algorithm $\mathcal{A}$ outputs some hypothesis H with total loss at most ε with probability at least $1-\delta$ over the randomness of the sample.

   The function $m^{\text{PAC}}_{\mathcal{H},\ell,\mathcal{A}}$ is called the *learning guarantee* of $\mathcal{A}$ and $\mathcal{A}$ is called an *improper k-PAC learner* for $\mathcal{H}$ with respect to ℓ.

   The notions of *proper k-PAC learnability* and *proper k-PAC learner* are defined analogously but requiring further that $\mathcal{H}' = \mathcal{H}$.

7. Given a k-partite labeled sample $(x,y) \in (\prod_{i=1}^k X_i^m) \times Y^{[m]^k}$ of size m, a k-partite loss function $\ell\colon (\prod_{i=1}^k X_i)\times Y\times Y \to \mathbb{R}_{\geq 0}$ and a hypothesis $H\in\mathcal{F}_k(\Omega,\Lambda)$, the *empirical loss* (or *empirical risk*) of H with respect to (x,y) and ℓ is

$$L_{x,y,\ell}(H) \stackrel{\text{def}}{=} \begin{cases} \dfrac{1}{m^k} \cdot \displaystyle\sum_{\alpha\in[m]^k} \ell\Big(\alpha^*(x), H^*_m(x)_\alpha, y_\alpha\Big), & \text{if } m>0, \\ 0, & \text{if } m=0. \end{cases}$$

   We say that (x,y) is *sample realizable* in a k-partite hypothesis class $\mathcal{H}\subseteq\mathcal{F}_k(\Omega,\Lambda)$ with respect to ℓ if $\inf_{H\in\mathcal{H}} L_{x,y,\ell}(H) = 0$. It is straightforward to check that if $\mu\in\Pr(\Omega)$, $F\in\mathcal{F}_k(\Omega,\Lambda)$ is realizable in $\mathcal{H}$ with respect to μ and ℓ and $\boldsymbol{x}\sim\mu^m$, then with probability 1, $(\boldsymbol{x}, F^*_m(\boldsymbol{x}))$ is sample realizable in $\mathcal{H}$ with respect to ℓ.

8. Given a k-partite labeled sample $(x,y) \in (\prod_{i=1}^k X_i^m) \times Y^{[m]^k}$ of size m, a k-tuple $\alpha = (\alpha_i)_{i=1}^k$ of injective functions $\alpha_i\colon [n]\to[m]$ induces naturally a *subsample* $(\alpha^\#(x), \alpha^\#(y))$ of size n by picking the points and labels corresponding to the coordinates indexed by $\alpha_1,\ldots,\alpha_k$ (in their order). This is formalized as follows: we

to ignore measurability concerns can simply interpret this as "all probabilities computed must make sense"; in fact, this universal measurability condition is required only to make sense of the notion of high-arity uniform convergence to make sense, which is out of the scope of this paper.

contra-variantly define the maps $\alpha^{\#}\colon \prod_{i=1}^{k} X_i^m \to \prod_{i=1}^{k} X_i^n$ and $\alpha^{\#}\colon Y^{[m]^k} \to Y^{[n]^k}$ that select the points and labels according to

$$\left(\alpha^{\#}(x)_i\right)_v \overset{\text{def}}{=} (x_i)_{\alpha_i(v)} \qquad \left(x \in \prod_{j=1}^{k} X_j^m, i \in [k], v \in [n]\right),$$

$$\left(\alpha^{\#}(y)\right)_\beta \overset{\text{def}}{=} y_{\alpha_1(\beta_1),\ldots,\alpha_k(\beta_k)} \qquad \left(y \in Y^{[m]^k}, \beta \in [n]^k\right).$$

With slight abuse of notation, we also contra-variantly define the map $\alpha^{\#}\colon \left(\prod_{i=1}^{k} X_i^m\right) \times Y^{[m]^k} \to \left(\prod_{i=1}^{k} X_i^n\right) \times Y^{[n]^k}$ by letting it act as the previous two maps in the corresponding coordinates. The overload in the notation is justified by the fact that it makes the definition of F_m^* of 2.1.3 equivariant in the sense that the following diagram is commutative:

$$\begin{array}{ccc} \prod_{i=1}^{k} X_i^m & \xrightarrow{F_m^*} & Y^{[m]^k} \\ {\scriptstyle \alpha^{\#}}\downarrow & & \downarrow{\scriptstyle \alpha^{\#}} \\ \prod_{i=1}^{k} X_i^n & \xrightarrow{F_n^*} & Y^{[n]^k} \end{array}$$

The avid reader will note that the maps α^* of 2.1.1 are essentially the particular case of the above when $n = 1$.

The next definition is that of a high-arity selection/compression scheme. Informally, a selection scheme has to have two procedures: a selection procedure κ_m that takes in a sample of size m and outputs a smaller subsample of size s_m along with some extra information (a header) and a reconstruction procedure ρ_m that takes in the output of κ_m and reconstructs the original sample. The selection scheme is then called a compression scheme when it correctly reconstructs the original sample (in the sense of zero loss).

Definition 2.2 (Partite selection/compression schemes)**.** Let $k \in \mathbb{N}_+$, let $\Omega = (\Omega_i)_{i=1}^k$ be a k-tuple of non-empty standard Borel spaces $\Omega_i = (X_i, \mathcal{B}_i)$ and let $\Lambda = (Y, \mathcal{B}')$ be a non-empty standard Borel space.

1. A *k-partite selection scheme* with outputs in a k-partite hypothesis class $\mathcal{H}' \subseteq \mathcal{F}_k(\Omega, \Lambda)$ is a tuple $\mathcal{S} = (\sigma, \eta, \rho)$ such that:

   - $\sigma = (\sigma_m)_{m\in\mathbb{N}}$ is a sequence of maps, called *selection maps*, such that each σ_m when given a sample of size m, outputs k tuples of some length s_m that are the indices of the subsample selected. Formally, each σ_m is a measurable function
   $$\sigma_m\colon \left(\prod_{j=1}^{k} X_j^m\right) \times Y^{[m]^k} \to \prod_{i=1}^{k} ([m])_{s_m},$$
   where $s_m \in \mathbb{N}$ is some integer, which we call the *selection size* (note that we must have $s_m \leq m$ as the domain of σ_m is never empty).
   - $\eta = (\eta_m)_{m\in\mathbb{N}}$ is a sequence of maps, called the *header maps*, that when given a sample of size m, outputs some extra information, encoded as an element of $[h_m]$ for some fixed $h_m \in \mathbb{N}$, that will aid in the reconstruction of the original sample. Formally, each η_m is a measurable function
   $$\eta_m\colon \left(\prod_{j=1}^{k} X_j^m\right) \times Y^{[m]^k} \to [h_m],$$
   where $h_m \in \mathbb{N}_+$ is called the *header size* of η_m.
   - $\rho = (\rho_m)_{m\in\mathbb{N}}$ is a sequence of maps, called the *reconstruction maps*, that when given a labeled sample of size s_m and header of size h_m (intended to have come from σ_m and η_m), reconstruct a labeled sample of size m. Formally, each ρ_m is a measurable function
   $$\rho_m\colon \left(\prod_{i=1}^{k} X_i^{s_m}\right) \times Y^{[s_m]^k} \times [h_m] \to \mathcal{H}'.$$

The compression procedure of $\mathcal{S}$ is formalized by defining a sequence $\kappa = (\kappa_m)_{m\in\mathbb{N}}$ of *compression maps* of the form

$$\kappa_m \colon \left(\prod_{i=1}^{k} X_i^m\right) \times Y^{[m]^k} \to \left(\prod_{i=1}^{k} X_i^{s_m}\right) \times Y^{[s_m]^k} \times [h_m]$$

by

$$\kappa_m(x,y) \stackrel{\text{def}}{=} \big(\sigma_m(x,y)^{\#}(x,y), \eta_m(x,y)\big) \qquad \left(x \in \prod_{i=1}^{k} X_i^m, y \in Y^{[m]^k}\right).$$

(Recall from Definition 2.1.8 that since $\sigma_m(x,y) \in \prod_{i=1}^{k}([m])_{s_m}$ is a k-tuple of injections, it contra-variantly induces a map that selects a subsample $\sigma_m(x,y)^{\#}(x,y)$ of size s_m from the sample (x,y) of size m and the map κ_m then takes this subsample an appends to it the header $\eta_m(x,y)$.)

We will use the notation

$$\sigma^{\mathcal{S}} \stackrel{\text{def}}{=} \sigma = (\sigma_m)_{m\in\mathbb{N}}, \qquad \eta^{\mathcal{S}} \stackrel{\text{def}}{=} \eta = (\eta_m)_{m\in\mathbb{N}}, \qquad \rho^{\mathcal{S}} \stackrel{\text{def}}{=} \rho = (\rho_m)_{m\in\mathbb{N}},$$
$$s_m^{\mathcal{S}} \stackrel{\text{def}}{=} s_m, \qquad h_m^{\mathcal{S}} \stackrel{\text{def}}{=} h_m, \qquad \kappa^{\mathcal{S}} \stackrel{\text{def}}{=} \kappa = (\kappa_m)_{m\in\mathbb{N}}.$$

If Y is finite, the *compression size* and the *compression bitlength* functions of $\mathcal{S}$ are the functions $c_{\mathcal{S}}, b_{\mathcal{S}} \colon \mathbb{N} \to \mathbb{R}_{\geq 0}$ given by

$$c_{\mathcal{S}}(m) \stackrel{\text{def}}{=} h_m \cdot |Y|^{s_m^k}, \qquad b_{\mathcal{S}}(m) \stackrel{\text{def}}{=} \log_2\big(c_{\mathcal{S}}(m)\big).$$

2. A k-partite selection scheme $\mathcal{S}$ with outputs in a k-partite hypothesis class $\mathcal{H}' \subseteq \mathcal{F}_k(\Omega,\Lambda)$ is called a *sample compression scheme* for a k-partite hypothesis class $\mathcal{H} \subseteq \mathcal{F}_k(\Omega,\Lambda)$ with respect to a k-partite loss function $\ell \colon (\prod_{i=1}^{k} X_i) \times Y \times Y \to \mathbb{R}_{\geq 0}$ if for every $m \in \mathbb{N}$ and every labeled sample $(x,y) \in (\prod_{i=1}^{k} X_i^m) \times Y^{[m]^k}$ of size m that is sample realizable with respect to ℓ, we have

$$L_{x,y,\ell}\Big(\rho_m^{\mathcal{S}}\big(\kappa_m^{\mathcal{S}}(x,y)\big)\Big) = 0,$$

or in plain English: if we compress (x,y) via $\kappa^{\mathcal{S}}$ then reconstruct via $\rho^{\mathcal{S}}$, then the result has zero empirical loss.

We say that the sample compression scheme is *proper* if $\mathcal{H} = \mathcal{H}'$.

Example 2.3 (Compact convex sets)**.** To illustrate the definitions above, consider the case in which $k \stackrel{\text{def}}{=} 2$, $X_1 \stackrel{\text{def}}{=} X_2 \stackrel{\text{def}}{=} \mathbb{R}$, $\mathcal{B}_i$ ($i \in [2]$) is the Borel σ-algebra, $Y \stackrel{\text{def}}{=} \{0,1\}$ and $\mathcal{B}'$ is the discrete σ-algebra. We will also use the 0/1-loss function given by $\ell(x,y,y') \stackrel{\text{def}}{=} \mathbb{1}[y \neq y']$, i.e., it gives penalty 1 to incorrect guesses and penalty 0 to correct ones.

Since (Borel-)measurable functions $X_1 \times X_2 \to Y$ are in natural one-to-one correspondence with Borel subsets of $X_1 \times X_2 = \mathbb{R}^2$, we will use the latter representation. We then let $\mathcal{H}$ be the collection of compact convex subsets of $\mathbb{R}^2$; this can be given a σ-algebra that satisfies Definition 2.1.4 [see Coregliano and Malliaris, 2026, Proof of Theorem 4.1, Item (5)].

The VCN_2-dimension of $\mathcal{H}$ is easily seen to be 2: every slice of a compact convex set is a compact interval, which means that the slice families $\mathcal{H}(x)$ are exactly the set of compact intervals of $\mathbb{R}$, which has VC-dimension 2. However, ignoring the product structure, the VC-dimension of the whole class $\mathcal{H}$ is ∞; this is because the circle $C \stackrel{\text{def}}{=} \{(x,y) \in \mathbb{R}^2 \mid x^2+y^2 = 1\}$ is shattered by $\mathcal{H}$ (since the convex hull of any $A \subseteq C$ only intersects C on A).

Coregliano and Malliaris [2024] show that if the VCN_2-dimension is finite, then any 2-partite empirical risk minimizer (ERM) is a k-PAC learner. One particular ERM for our $\mathcal{H}$ is the algorithm that simply returns the convex hull of the points in the sample that have label 1.

Let us now give some evidence that a good sample compression scheme should exist for $\mathcal{H}$. Suppose we are given a grid in the form of m points in the x axis $x_1^1, \ldots, x_m^1$ and m points

in the y axis $x_1^2, \ldots, x_m^2$ along with 0/1-labels $(y_{i,j})_{i,j=1}^m$ corresponding to a convex set. By possibly reordering the points (and labels accordingly), we can assume that $x_1^1 \leq \cdots \leq x_m^1$ and $x_1^2 \leq \cdots \leq x_m^2$. We will assume that there are no repetitions of the coordinates for simplicity. Let C be the convex hull of the points of the grid that are labeled 1. Note that C is completely determined by the points of its perimeter P that lie on the grid. In turn, the perimeter P can be described by two functions f and g, the former determining the upper bound of P and the latter determining the lower bound. If we can encode f and g with few points and a small header, then we have successfully encoded P, hence C.

We focus on f as the argument for g is analogous. Since C is convex, we know that f consists of a strictly increasing part, followed by a constant part, followed by a strictly decreasing part (some of these may be of length 0). To encode f, we start by specifying the x and y coordinates of the points that determine the boundaries of these parts; these amount to at most 4 coordinates x and at most 3 coordinates y. Note that this completely encodes the constant part, but the increasing and decreasing parts need not be straight lines.

Let us now focus on the increasing part and let us make two simplifying assumptions:

1 The grid is equally spaced horizontally and vertically (i.e., there exist c_x, c_y such that $x_{i+1}^1 - x_i^1 = c_x$ and $x_{i+1}^2 - x_i^2 = c_y$ for every $i \in [m-1]$). By rescaling, we may assume without loss of generality, that $c_y = 1$.

2 Every column of the grid has a point in P.

Under these simplifying assumptions, the idea is now to encode the points of P that are on the grid by encoding in the header the steps

$$a_1 \stackrel{\text{def}}{=} x_2^2 - x_1^2, \qquad a_2 \stackrel{\text{def}}{=} x_3^2 - x_2^2, \qquad a_3 \stackrel{\text{def}}{=} x_4^2 - x_3^2, \qquad \ldots, \qquad a_{k-1} \stackrel{\text{def}}{=} x_k^2 - x_{k-1}^2,$$

where k is the length of the increasing part of f. Convexity of C translates into concavity of f, which in turn implies that we must have $a_1 \geq a_2 \geq \cdots \geq a_{k-1}$ and since these must sum to exactly $x_k^2 - x_1^2$ (which was already specified by our y coordinates), $(a_i)_{i \in [k-1]}$ must be a partition of $x_k^2 - x_1^2 \leq m$. By the Hardy–Ramanujan bound, there are at most $\exp(O(\sqrt{m}))$ such partitions, which means that we can encode them in a header of this size.

Under our simplifying assumptions, our compression scheme $\mathcal{S}$ would have

$$s_m^{\mathcal{S}} \leq 8, \qquad\qquad h_m^{\mathcal{S}} \leq \exp(O(\sqrt{m})).$$

While it might seem believable that one might be able to drop the simplifying assumption 2 by analyzing weighted partitions of integers (i.e., k-tuples $(a_i, b_i)_{i \in [k]}$ of positive integers with $\sum_{i=1}^k a_i = c_a$, $\sum_{i=1}^k b_i = c_b$ and $(a_1/b_1) \geq (a_2/b_2) \geq \cdots \geq (a_k/b_k)$), we point out that the simplifying assumption 1 cannot be easily dropped: if $x_{i+1}^2 - x_i^2 = 1$ and $x_{i+1}^1 - x_i^1 = (m+1)^{i-1}$ for every $i \in [k-1]$, then even assuming that every column of the grid has a point in P, it is straightforward to see that as long as the y coordinate of the point of P strictly increases from each column to the next, the resulting function f will be concave. In the worst case, there are $\binom{m-1}{k-1}$ possibilities, which when $k \approx m/2$ is too many possibilities: $\Theta(2^{m/2}/\sqrt{m})$.

The next theorem shows that every k-partite hypothesis class that admits a k-partite compression scheme $\mathcal{S}$ of essentially non-trivial quality must be k-PAC learnable. More specifically, a trivial k-partite compression scheme has $s_m^{\mathcal{S}} = m$ and $h_m^{\mathcal{S}} = 1$, which yields

$$\ln(h_m^{\mathcal{S}}) + k \cdot s_m^{\mathcal{S}} \cdot \ln(m) = k \cdot m \cdot \ln(m)$$

and our assumption is that we have a k-partite compression scheme in which the quantity above is $o(m)$ instead.

Theorem 2.4. *Let $k \in \mathbb{N}_+$, let $\Omega = (\Omega_i)_{i=1}^k$ be a k-tuple of non-empty standard Borel spaces $\Omega_i = (X_i, \mathcal{B}_i)$, let $\Lambda = (Y, \mathcal{B}')$ be a non-empty standard Borel space with $|Y| \geq 2$, let $\ell \colon \prod_{i=1}^k X_i \times Y \times Y \to \mathbb{R}_{\geq 0}$ be a bounded k-partite loss function and let $\mathcal{S}$ be a k-partite compression scheme for a k-partite hypothesis class $\mathcal{H} \subseteq \mathcal{F}_k(\Omega, \Lambda)$ such that*

$$\ln(h_m^{\mathcal{S}}) + k \cdot s_m^{\mathcal{S}} \cdot \ln(m) \leq o(m) \tag{2.1}$$

as $m \to \infty$ (which in particular follows from $b_{\mathcal{S}}(m) \leq o(m/\ln(m))$ when Y is finite), then $\mathcal{H}$ is k-PAC learnable with respect to ℓ. More specifically, letting $\mathcal{A}$ be given by

$$\mathcal{A}(x,y) \stackrel{\text{def}}{=} \rho_m^{\mathcal{S}}\big(\kappa_m^{\mathcal{S}}(x,y)\big) \qquad \left(m \in \mathbb{N}, x \in \prod_{i=1}^{k} X_i^m, y \in Y^{([m])_k}\right)$$

gives a k-PAC learner for $\mathcal{H}$ with respect to ℓ with learning guarantee

$$m^{\text{PAC}}_{\mathcal{H},\ell,\mathcal{A}}(\varepsilon,\delta) \stackrel{\text{def}}{=} \min\Bigg\{ m_0 \in \mathbb{N}_+, \;\Bigg|\; \forall m \geq m_0, \left(1 - \frac{(m - s_m^{\mathcal{S}})^k}{m^k}\right) < \varepsilon \\ \wedge (m)^k_{s_m^{\mathcal{S}}} \cdot h_m^{\mathcal{S}} \cdot \exp\left(-\frac{(\varepsilon - (1-(m-s_m^{\mathcal{S}})^k/m^k)\cdot\|\ell\|_\infty)^2 \cdot (m - s_m^{\mathcal{S}})}{2\cdot k \cdot \|\ell\|_\infty^2}\right) \leq \delta \Bigg\}. \quad (2.2)$$

In particular, if

$$\ln(h_m^{\mathcal{S}}) + k \cdot s_m^{\mathcal{S}} \cdot \ln(m) \leq \big(1 + o(1)\big) \cdot \sqrt{2 \cdot k \cdot m} \cdot \ln(m), \quad (2.3)$$

as $m \to \infty$, then

$$m^{\text{PAC}}_{\mathcal{H},\ell,\mathcal{A}}(\varepsilon,\delta) \leq \big(1 + o(1)\big) \cdot \frac{2 \cdot k \cdot \|\ell\|_\infty^2}{\varepsilon^2} \cdot \max\left\{1, \ln\left(\frac{1}{\delta}\right)\right\},$$

where the $o(1)$ error term is as $\varepsilon \to 0$ with k fixed, is uniform in δ, and has the same uniformity in $\|\ell\|_\infty$ as the estimate $o(1)$ in (2.3) provided $\|\ell\|_\infty$ is bounded away from 0.

3 Discussion and final remarks

Let us conclude the paper by briefly addressing what was not covered in this work.

3.1 What about the other direction?

First and foremost, in the classical setting, David, Moran, and Yehudayoff [2016] proved that sample compression is in fact equivalent to PAC learnability[2], that is, given a PAC learner $\mathcal{A}$, one can produce from it a sample compression scheme. In a high-level, their proof follows by considering a two-player game over a given sample in which a learner attempts to pick a hypothesis H that is the output of $\mathcal{A}$ on a subsample of size $m^{\text{PAC}}(1/3, 1/3)$ while a spoiler tries to pick a point of the sample on which H has large loss. The PAC learning guarantee says that if the spoiler plays a mixed strategy first, then the learner can ensure small loss even with a pure strategy playing second. By von Neumann's Minimax Theorem, there must be a mixed strategy S that the learner can play first that ensures that any pure strategy of the spoiler playing second yields small loss. This allows us to compress the sample by considering a large enough sample from the mixed strategy S and taking a majority vote.

The most natural way to adapt the proof above to high-arity is in the partite setting. However the issue here is that since in high-arity we only allow product distributions over $X_1 \times \cdots \times X_k$, von Neumann's Minimax Theorem no longer applies (there is no reason to believe that optimal mixed strategies must necessarily come from product distributions). An alternative would be to setup the Minimax as a game with one learner and k spoilers, but even though the corresponding cost function is multilinear in the distributions, there is no reason to believe (nor is it a reasonable assumption to make) that we have the correct (quasi-)concavities/convexities. The non-partite setting is even worse: since it requires power distributions, the k spoilers would then have to play the same mixed strategy.

[2]Their proof is when $Y = \{0,1\}$ and with 0/1-loss, but it readily adapts to the case when Y is finite and ℓ is *separated* in the sense that it assigns 0 penalty to correct guesses and is bounded away from 0 in incorrect guesses.

3.2 Approximate/agnostic compression schemes and higher-order variables

David, Moran, and Yehudayoff [2016] also defined the notions of approximate and approximate agnostic sample compression schemes. In an approximate compression scheme, instead of assuming that the reconstruction process yields zero loss, one assumes that it gives small loss, that is, for every realizable sample (x, y) of size m, we want $L_{x,y,\ell}(\rho_m(\kappa_m(x, y))) \leq \varepsilon_m$ where $\varepsilon_m > 0$ (and typically thought to satisfy $\varepsilon_m \to 0$ as $m \to \infty$). For approximate agnostic compression schemes, one drops the realizability assumption but requires instead that the reconstruction is competitive in the sense that

$$L_{x,y,\ell}(\rho_m(\kappa_m(x, y))) \leq \inf_{H \in \mathcal{H}} L_{x,y,\ell}(H) + \varepsilon_m.$$

As was observed by David, Moran, and Yehudayoff [2016], the fact that (agnostic, respectively) PAC learnability implies approximate (agnostic, respectively) sample compressibility is very simple: one simply considers the uniform measure on the sample, uses the PAC guarantee to claim that there exists a small subsample that when fed to the algorithm yields small loss (or is competitive in the agnostic case), then uses a bit of extra information to encode potential repetitions. In the other direction, the argument that sample compressibility implies PAC learnability easily adapts to the approximate (agnostic) case.

Of course, the definitions of approximate and approximate agnostic sample compressibility readily adapt to the high-arity setting (both partite and non-partite). It is straightforward to check that the proof of equivalence between approximate sample compressibility and PAC learnability lifts to the high-arity case (both partite and non-partite). In the remainder of this section, we sketch the equivalence of the agnostic case.

First, what is high-arity agnostic PAC learnability? In the definition of Coregliano and Malliaris [2024] and in analogy with classical PAC theory, there are two changes: the adversary is allowed to play an "agnostic" distribution and the learner is only tasked with being competitive. The key technicality is in what an "agnostic" distribution is. In the non-partite, the adversary picks a distribution ν over $X_1^{\mathbb{N}_+} \times \cdots \times X_k^{\mathbb{N}_+} \times Y^{\mathbb{N}_+^k}$ (we want to think of elements of this set as infinite k-dimensional tensors with names on each of the indices) that satisfies:

Separately exchangeability in the sense that ν is invariant under the natural right-action of $S_{\mathbb{N}_+}^k$ (the ith copy of $S_{\mathbb{N}_+}$ permutes the coordinates of $X_i^{\mathbb{N}_+}$ and $S_{\mathbb{N}_+}^k$ as a whole permutes the coordinates of $Y^{\mathbb{N}_+}$).

Locality in the sense that marginals on $U_1 \times \cdots \times U_k$ and $V_1 \times \cdots \times V_k$ are independent provided $U_i \cap V_i = \varnothing$ for every $i \in [k]$.

A labeled m sample is then generated by taking the marginal of ν on $[m]^k$.

In the non-partite, the adversary picks a distribution ν over $(X^{\mathbb{N}_+})^k \times Y^{(\mathbb{N}_+)^k}$ that satisfies:

(Joint) Exchangeability in that ν is invariant under the natural right-action of $S_{\mathbb{N}_+}$.

Locality in the sense that marginals on U^k and V^k are independent provided $U \cap V = \varnothing$.

A labeled m sample is then generated by taking the marginal of ν on $[m]^k$.

One then gets a handle on these distributions by invoking the Aldous–Hoover Theorem [Hoover, 1979, Aldous, 1985, 1981] from exchangeability theory [see also Kallenberg, 2005, §7, for a more modern presentation] to get a normal representation of such distributions. Without getting into the (highly) technical details, let us mention that the proof presented here readily adapts to this more general setting simply because using this Aldous–Hoover representation, one can check that the martingales defined in the proofs remain martingales (with the same bounds on martingale differences), which allow the key application of (two-sided) Azuma's Inequality. In fact, similar martingales and an application of Azuma's Inequality were used in Coregliano and Malliaris [2024, Lemma 7.3].

References

David J. Aldous. Representations for partially exchangeable arrays of random variables. *J. Multivariate Anal.*, 11(4):581–598, 1981. ISSN 0047-259X. doi: 10.1016/0047-259X(81)90099-3. URL https://doi.org/10.1016/0047-259X(81)90099-3.

David J. Aldous. Exchangeability and related topics. In *École d'été de probabilités de Saint-Flour, XIII—1983*, volume 1117 of *Lecture Notes in Math.*, pages 1–198. Springer, Berlin, 1985. ISBN 3-540-15203-2. doi: 10.1007/BFb0099421. URL https://doi.org/10.1007/BFb0099421.

Noga Alon, Eldar Fischer, and Ilan Newman. Efficient testing of bipartite graphs for forbidden induced subgraphs. *SIAM J. Comput.*, 37(3):959–976, 2007. ISSN 0097-5397,1095-7111. doi: 10.1137/050627915. URL https://doi.org/10.1137/050627915.

Artem Chernikov and Henry Towsner. Hypergraph regularity and higher arity VC-dimension, 2020. URL https://arxiv.org/abs/2010.00726.

Artem Chernikov, Daniel Palacin, and Kota Takeuchi. On n-dependence. *Notre Dame J. Form. Log.*, 60(2):195–214, 2019. ISSN 0029-4527,1939-0726. doi: 10.1215/00294527-2019-0002. URL https://doi.org/10.1215/00294527-2019-0002.

Leonardo N. Coregliano and Maryanthe Malliaris. High-arity PAC learning via exchangeability, 2024. URL https://arxiv.org/abs/2402.14294. To appear in the Journal of the ACM.

Leonardo N. Coregliano and Maryanthe Malliaris. A packing lemma for VCN_k-dimension and learning high-dimensional data, 2025a. URL https://arxiv.org/abs/2505.15688.

Leonardo N. Coregliano and Maryanthe Malliaris. Sample completion, structured correlation, and Netflix problems, 2025b. URL https://arxiv.org/abs/2509.20404.

Leonardo N. Coregliano and Maryanthe Malliaris. Some model-theoretic consequences of high-arity uniform convergence, part I, 2026. URL https://arxiv.org/abs/2605.09911.

Ofir David, Shay Moran, and Amir Yehudayoff. On statistical learning via the lens of compression, 2016. URL https://arxiv.org/abs/1610.03592.

Jacob Fox, János Pach, and Andrew Suk. Erdős-Hajnal conjecture for graphs with bounded VC-dimension. *Discrete Comput. Geom.*, 61(4):809–829, 2019. ISSN 0179-5376,1432-0444. doi: 10.1007/s00454-018-0046-5. URL https://doi.org/10.1007/s00454-018-0046-5.

Lior Gishboliner, Asaf Shapira, and Yuval Wigderson. Is it easy to regularize a hypergraph with easy links?, 2025. URL https://arxiv.org/abs/2506.15582. arXiv:2506.15582.

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. ISSN 0097-3165,1096-0899. doi: 10.1016/0097-3165(95)90052-7. URL https://doi.org/10.1016/0097-3165(95)90052-7.

Ron Holzman, Shay Moran, and Alexander Shlimovich. Uniform laws of large numbers in product spaces, 2026. URL https://arxiv.org/abs/2603.24493.

D. N. Hoover. Relations on probability spaces and arrays of random variables. Preprint, Institute of Advanced Study, Princeton, NJ, 1979.

Olav Kallenberg. *Probabilistic symmetries and invariance principles.* Probability and its Applications (New York). Springer, New York, 2005. ISBN 978-0387-25115-8; 0-387-25115-4.

Munehiro Kobayashi. A generalization of the pac learning in product probability spaces (model theoretic aspects of the notion of independence and dimension). *Model theoretic aspects of the notion of independence and dimension*, 1938:33–37, 04 2015. ISSN 18802818. URL https://cir.nii.ac.jp/crid/1050001335843879296.

Takayuki Kuriyama and Kota Takeuchi. On the PAC_n learning. *Model theoretic aspects of the notion of independence and dimension*, (1938):54–58, 2015. ISSN 1880-2818. URL https://ndlsearch.ndl.go.jp/books/R000000004-I026516592.

Nick Littlestone and Manfred Warmuth. Relating data compression and learnability.

Roi Livni and Yishay Mansour. Graph-based discriminators: Sample complexity and expressiveness, 2019a. URL https://arxiv.org/abs/1906.00264.

Roi Livni and Yishay Mansour. Graph-based discriminators: Sample complexity and expressiveness. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019b. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/e2eabaf96372e20a9e3d4b5f83723a61-Paper.pdf.

László Lovász and Balázs Szegedy. Regularity partitions and the topology of graphons. In *An irregular mind*, volume 21 of *Bolyai Soc. Math. Stud.*, pages 415–446. János Bolyai Math. Soc., Budapest, 2010. ISBN 978-963-9453-14-2; 978-3-642-14443-1. doi: 10.1007/978-3-642-14444-8_12. URL https://doi.org/10.1007/978-3-642-14444-8_12.

Shay Moran and Amir Yehudayoff. Sample compression schemes for VC classes. *J. ACM*, 63(3):Art. 21, 10, 2016. ISSN 0004-5411,1557-735X. doi: 10.1145/2890490. URL https://doi.org/10.1145/2890490.

Micha Perles. Credited by Shelah in 1972 in "A combinatorial problem; stability and order for models and theories in infinitary languages", 1972.

Norbert Sauer. On the density of families of sets. *J. Combinatorial Theory Ser. A*, 13: 145–147, 1972. ISSN 0097-3165. doi: 10.1016/0097-3165(72)90019-2. URL https://doi.org/10.1016/0097-3165(72)90019-2.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014. ISBN 9781107298019. doi: 10.1017/CBO9781107298019.

Saharon Shelah. A combinatorial problem; stability and order for models and theories in infinitary languages. *Pacific J. Math.*, 41:247–261, 1972. ISSN 0030-8730,1945-5844. URL http://projecteuclid.org/euclid.pjm/1102968432.

Saharon Shelah. Strongly dependent theories. *Israel J. Math.*, 204(1):1–83, 2014. ISSN 0021-2172,1565-8511. doi: 10.1007/s11856-014-1111-2. URL https://doi.org/10.1007/s11856-014-1111-2.

Caroline Terry. Growth of regular partitions 1: Improved bounds for small slicewise VC-dimension, 2024a. URL https://arxiv.org/abs/2404.01274. arXiv:2404.01274.

Caroline Terry. Growth of regular partitions 2: Weak regularity, 2024b. URL https://arxiv.org/abs/2404.01293.

Caroline Terry and Julia Wolf. Irregular triads in 3-uniform hypergraphs, 2022. URL https://arxiv.org/abs/2111.01737.

A Proofs

A.1 Partite setting

In this subsection, we prove that the existence of a k-partite compression scheme (of essentially non-trivial quality) implies k-PAC learnability. We start with a concentration lemma based on Azuma's Inequality.

Lemma A.1. *Let $k \in \mathbb{N}_+$, let $\Omega = (\Omega_i)_{i=1}^k$ be a k-tuple of non-empty standard Borel spaces $\Omega_i = (X_i, \mathcal{B}_i)$, let $\Lambda = (Y, \mathcal{B}')$ be a non-empty standard Borel space, let $\mathcal{S}$ be a k-partite selection scheme and let $\ell\colon \prod_{i=1}^k X_i \times Y \times Y \to \mathbb{R}_{\geq 0}$ be a bounded k-partite loss function. Then for every $\varepsilon > 0$, every $\mu \in \Pr(\Omega)$, every integer $m \in \mathbb{N}$ such that*

$$\left(1 - \frac{(m - s_m^{\mathcal{S}})^k}{m^k}\right) \cdot \|\ell\|_\infty < \varepsilon, \tag{A.1}$$

every $\eta \in [h_m^{\mathcal{S}}]$, every k-tuple $\sigma = (\sigma_i)_{i=1}^k$ of injections $\sigma_i\colon [s_m^{\mathcal{S}}] \to [m]$ and every $F \in \mathcal{F}_k(\Omega, \Lambda)$, we have

$$\mathbb{P}_{\boldsymbol{x}\sim\mu^m}\left[L_{\mu,F,\ell}\left(\rho^{\mathcal{S}}\left(\sigma^{\#}\left(\boldsymbol{x}, F_m^*(\boldsymbol{x})\right), \eta\right)\right) - L_{\boldsymbol{x},F_m^*(\boldsymbol{x}),\ell}\left(\rho^{\mathcal{S}}\left(\sigma^{\#}\left(\boldsymbol{x}, F_m^*(\boldsymbol{x})\right), \eta\right)\right) \geq \varepsilon\right] \leq \exp\left(-\frac{\widetilde{\varepsilon}\cdot(m - s_m^{\mathcal{S}})}{2\cdot k\cdot\|\ell\|_\infty^2}\right),$$

where

$$\widetilde{\varepsilon} \stackrel{\text{def}}{=} (\varepsilon - (1 - (m - s_m^{\mathcal{S}})^k/m^k)\cdot\|\ell\|_\infty)^2 > 0.$$

In particular, if $s_m^{\mathcal{S}} \leq o(m)$ as $m \to \infty$, then

$$\mathbb{P}_{\boldsymbol{x}\sim\mu^m}\left[L_{\mu,F,\ell}\left(\rho^{\mathcal{S}}\left(\sigma^{\#}\left(\boldsymbol{x}, F_m^*(\boldsymbol{x})\right), \eta\right)\right) - L_{\boldsymbol{x},F_m^*(\boldsymbol{x}),\ell}\left(\rho^{\mathcal{S}}\left(\sigma^{\#}\left(\boldsymbol{x}, F_m^*(\boldsymbol{x})\right), \eta\right)\right) \geq \varepsilon\right] \leq \exp\left(-\frac{\varepsilon^2}{2\cdot k\cdot\|\ell\|_\infty^2}\cdot m + \frac{\varepsilon\cdot(\varepsilon + \|\ell\|_\infty)}{\|\ell\|_\infty^2}\cdot o(m)\right),$$

where the $o(m)$ error term is as $m \to \infty$ with k fixed, is uniform in ε, and has the same uniformity in $\|\ell\|_\infty$ as the estimate $s_m^{\mathcal{S}} \leq o(m)$.

Proof. The result is trivial if $m = 0$, so suppose $m > 0$. By symmetry, we may suppose without loss of generality that for each $i \in [k]$, $\sigma_i\colon [s_m^{\mathcal{S}}] \to [m]$ is such that $\operatorname{im}(\sigma_i) = \{m - s_m^{\mathcal{S}} + 1, m - s_m^{\mathcal{S}} + 2, \ldots, m\}$. Pick $\boldsymbol{x}$ at random according to μ^m and for ease of notation, let

$$n \stackrel{\text{def}}{=} m - s_m^{\mathcal{S}}, \qquad \boldsymbol{H} \stackrel{\text{def}}{=} \rho_m^{\mathcal{S}}\left(\sigma^{\#}\left(\boldsymbol{x}, F_m^*(\boldsymbol{x})\right), \eta\right), \qquad \boldsymbol{x'} \stackrel{\text{def}}{=} \iota^{\#}(\boldsymbol{x}),$$

where $\iota = (\iota_i)_{i=1}^k$ is the k-tuple of inclulsion maps $\iota_i\colon [n] \to [m]$ (so $\boldsymbol{x'}$ is the random sample corresponding to the first n points of $\boldsymbol{x}$ in each of the k coordinates, in their natural order).

Note that

$$\begin{aligned}
L_{\boldsymbol{x},F_m^*(\boldsymbol{x}),\ell}(\boldsymbol{H}) - L_{\boldsymbol{x'},F_n^*(\boldsymbol{x'}),\ell}(\boldsymbol{H}) &= \frac{1}{m^k}\cdot\sum_{\alpha\in[m]^k}\ell\Big(\alpha^*(\boldsymbol{x}), \boldsymbol{H}_m^*(\boldsymbol{x})_\alpha, F_m^*(\boldsymbol{x})\Big) \\
&\quad - \frac{1}{n^k}\cdot\sum_{\alpha\in[n]^k}\ell\Big(\alpha^*(\boldsymbol{x}), \boldsymbol{H}_m^*(\boldsymbol{x})_\alpha, F_m^*(\boldsymbol{x})\Big) \\
&= \left(\frac{1}{m^k} - \frac{1}{n^k}\right)\cdot\sum_{\alpha\in[n]^k}\ell\Big(\alpha^*(\boldsymbol{x}), \boldsymbol{H}_m^*(\boldsymbol{x})_\alpha, F_m^*(\boldsymbol{x})\Big) \\
&\quad + \frac{1}{m^k}\cdot\sum_{\alpha\in[m]^k\setminus[n]^k}\ell\Big(\alpha^*(\boldsymbol{x}), \boldsymbol{H}_m^*(\boldsymbol{x})_\alpha, F_m^*(\boldsymbol{x})\Big),
\end{aligned}$$

and since ℓ is bounded and noting that the first term in the above is non-positive (as $n \le m$) and the second term is non-negative, we conclude that

$$\begin{aligned}\left|L_{\boldsymbol{x},F_m^*(\boldsymbol{x}),\ell}(\boldsymbol{H}) - L_{\boldsymbol{x'},F_n^*(\boldsymbol{x'}),\ell}(\boldsymbol{H})\right| &\le \|\ell\|_\infty \cdot \max\left\{\left(\frac{1}{n^k} - \frac{1}{m^k}\right)\cdot n^k, \frac{1}{m^k}\cdot(m^k - n^k)\right\} \\ &= \left(1 - \frac{n^k}{m^k}\right)\cdot \|\ell\|_\infty.\end{aligned} \tag{A.2}$$

Since we are mostly concerned with the setting in which $s_m^{\mathcal{S}} \le o(m)$ (i.e., $n = (1-o(1))\cdot m$), the difference above is small, so we will compare $L_{\mu,F,\ell}(\boldsymbol{H})$ with $L_{\boldsymbol{x'},F_n^*(\boldsymbol{x'}),\ell}$ first.

Let $\mathcal{F}_0$ be the σ-algebra generated by $((\boldsymbol{x}_i)_v)_{i\in[k],v\in[m]\setminus[n]}$ (i.e., generated by $\sigma^{\#}(\boldsymbol{x})$), so $\boldsymbol{H}$ is $\mathcal{F}_0$-measurable. For each $t \in [k\cdot n]$, we let $\mathcal{F}_i$ be the σ-algebra generated by $\mathcal{F}_0$ and all $(\boldsymbol{x}_i)_v$ with $(i,v) \in [k]\times[n]$ satisfying

$$(i-1)\cdot n + v - 1 < t,$$

that is, from $\mathcal{F}_{t-1}$ to $\mathcal{F}_t$, we add the variable $(\boldsymbol{x}_{i_t})_{v_t}$, where

$$i_t \stackrel{\text{def}}{=} \left\lfloor \frac{t-1}{n} \right\rfloor + 1, \qquad\qquad v_t \stackrel{\text{def}}{=} \big((t-1) \bmod n\big) + 1. \tag{A.3}$$

We also let $\boldsymbol{Z}_t \stackrel{\text{def}}{=} \mathbb{E}[L_{\boldsymbol{x'},F_n^*(\boldsymbol{x'}),\ell}(\boldsymbol{H}) \mid \mathcal{F}_t]$ and we note that $(\boldsymbol{Z}_t)_{t=0}^{k\cdot n}$ forms a (Doob) martingale with respect to $(\mathcal{F}_t)_{t=0}^{k\cdot n}$:

$$\mathbb{E}[\boldsymbol{Z}_{t+1} \mid \mathcal{F}_t] = \mathbb{E}\Big[\mathbb{E}\big[L_{\boldsymbol{x'},F_n^*(\boldsymbol{x'}),\ell}(\boldsymbol{H}) \;\big|\; \mathcal{F}_{t+1}\big] \;\Big|\; \mathcal{F}_t\Big] = \mathbb{E}\big[L_{\boldsymbol{x'},F_n^*(\boldsymbol{x'}),\ell}(\boldsymbol{H}) \;\big|\; \mathcal{F}_t\big] = \boldsymbol{Z}_t.$$

Note further that

$$\begin{aligned}\boldsymbol{Z}_0 &= \mathbb{E}\big[L_{\boldsymbol{x'},F_n^*(\boldsymbol{x'}),\ell}(\boldsymbol{H}) \;\big|\; \mathcal{F}_0\big] = L_{\mu,F,\ell}(\boldsymbol{H}), \\ \boldsymbol{Z}_{k\cdot n} &= \mathbb{E}\big[L_{\boldsymbol{x'},F_n^*(\boldsymbol{x'}),\ell}(\boldsymbol{H}) \;\big|\; \mathcal{F}_{k\cdot n}\big] = L_{\boldsymbol{x'},F_n^*(\boldsymbol{x'}),\ell}(\boldsymbol{H}),\end{aligned}$$

where the former is by linearity of conditional expectation (and the fact that $\boldsymbol{H}$ is $\mathcal{F}_0$-measurable) and the latter is simply because the underlying random variable is $\mathcal{F}_{k\cdot n}$-measurable.

We will show that $\boldsymbol{Z}_0$ and $\boldsymbol{Z}_{k\cdot n}$ are close with high (conditional) probability using Azuma's Inequality. Toward that, we note that for each $t \in [k\cdot n]$ we have

$$\begin{aligned}|\boldsymbol{Z}_t - \boldsymbol{Z}_{t-1}| = \Bigg|\frac{1}{n^k}\cdot \sum_{\alpha\in[n]^k} \Bigg(&\mathbb{E}\Big[\ell\big(\alpha^*(\boldsymbol{x'}), \boldsymbol{H}_n^*(\boldsymbol{x'}), \boldsymbol{F}_n^*(\boldsymbol{x'})\big) \;\Big|\; \mathcal{F}_t\Big] \\ &- \mathbb{E}\Big[\ell\big(\alpha^*(\boldsymbol{x'}), \boldsymbol{H}_n^*(\boldsymbol{x'}), \boldsymbol{F}_n^*(\boldsymbol{x'})\big) \;\Big|\; \mathcal{F}_{t-1}\Big]\Bigg)\Bigg|.\end{aligned} \tag{A.4}$$

Let us now argue that most of the terms in the sum above are zero. Recall that what changes from $\mathcal{F}_{t-1}$ to $\mathcal{F}_t$ is the addition of the variable $(\boldsymbol{x}_{i_t})_{v_t}$, where (i_t, v_t) are given by (A.3). Consider a term $\alpha \in [n]^k$ such that $\alpha_{i_t} \ne v_t$, then

$$\ell\big(\alpha^*(\boldsymbol{x'}), \boldsymbol{H}_n^*(\boldsymbol{x'}), \boldsymbol{F}_n^*(\boldsymbol{x'})\big)$$

is independent from $(\boldsymbol{x}_{i_t})_{v_t}$; in fact, since $((\boldsymbol{x}_i)_v)_{i\in[k],v\in[m]}$ is mutually independent, the above is conditionally independent from $(\boldsymbol{x}_{i_t})_{v_t}$ given $\mathcal{F}_{t-1}$, so the above gives the same value when conditioned on $\mathcal{F}_{t-1}$ as it does when conditioned on $\mathcal{F}_t$. Thus, whenever $\alpha_{i_t} \ne v_t$, the corresponding term in (A.4) is zero.

As there are exactly n^{k-1} many $\alpha \in [n]^k$ with $\alpha_{i_t} = v_t$, using triangle inequality and the fact that ℓ is bounded, we conclude that

$$|\boldsymbol{Z}_t - \boldsymbol{Z}_{t-1}| \le \frac{n^{k-1}}{n^k}\cdot\|\ell\|_\infty = \frac{\|\ell\|_\infty}{n}.$$

Recalling that

$$\widetilde{\varepsilon} \stackrel{\text{def}}{=} \varepsilon - \left(1 - \frac{n^k}{m^k}\right) \cdot \|\ell\|_\infty > 0$$

(where the inequality follows from our assumption (A.1)), by Azuma's Inequality (conditioned on $\mathcal{F}_0$), we get

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{x}\sim\mu^m}\left[L_{\mu,F,\ell}(\boldsymbol{H}) - L_{\boldsymbol{x'},F_n^*(\boldsymbol{x'}),\ell}(\boldsymbol{H}) \geq \widetilde{\varepsilon}\right] &= \mathbb{P}[\boldsymbol{Z}_{k\cdot n} - \boldsymbol{Z}_0 \leq -\widetilde{\varepsilon}] \\
&\leq \exp\left(-\frac{\widetilde{\varepsilon}^2}{2\cdot\sum_{t=1}^{k\cdot n}(\|\ell\|_\infty/n)^2}\right) \\
&= \exp\left(-\frac{\widetilde{\varepsilon}^2\cdot n}{2\cdot k\cdot\|\ell\|_\infty^2}\right) \\
&= \exp\left(-\frac{\widetilde{\varepsilon}^2\cdot(m-s_m^{\mathcal{S}})}{2\cdot k\cdot\|\ell\|_\infty^2}\right).
\end{aligned}$$

Putting this together with (A.2), we conclude that

$$\mathbb{P}_{\boldsymbol{x}\sim\mu^m}\left[L_{\mu,F,\ell}(\boldsymbol{H}) - L_{\boldsymbol{x},F_m^*(\boldsymbol{x}),\ell}(\boldsymbol{H}) \geq \varepsilon\right] \leq \exp\left(-\frac{\widetilde{\varepsilon}\cdot(m-s_m^{\mathcal{S}})}{2\cdot k\cdot\|\ell\|_\infty^2}\right),$$

as desired.

Let us now estimate the above asymptotically when $s_m^{\mathcal{S}} \leq o(m)$. First note that $1-(m-s^{\mathcal{S}})^k/m^k = o(1)$ as $m\to\infty$ with k fixed as $m - s_m^{\mathcal{S}} = (1-o(1))\cdot m$. Then we get

$$\begin{aligned}
\exp\left(-\frac{\widetilde{\varepsilon}^2\cdot(m-s_m^{\mathcal{S}})}{2\cdot k\cdot\|\ell\|_\infty^2}\right), &= \exp\left(-\frac{(\varepsilon - \|\ell\|_\infty\cdot o(1))^2\cdot(1-o(1))\cdot m}{2\cdot k\cdot\|\ell\|_\infty^2}\right) \\
&\leq \exp\left(-\frac{\varepsilon^2}{2\cdot k\cdot\|\ell\|_\infty^2}\cdot m + \frac{\varepsilon\cdot(\varepsilon+\|\ell\|_\infty)}{\|\ell\|_\infty^2}\cdot o(m)\right),
\end{aligned}$$

where the $o(m)$ error term is as $m\to\infty$ with k fixed, is uniform in ε, and has the same uniformity in $\|\ell\|_\infty$ as the estimate $s_m^{\mathcal{S}} \leq o(m)$. □

We now prove Theorem 2.4 (restated below).

Theorem 2.4. *Let $k\in\mathbb{N}_+$, let $\Omega = (\Omega_i)_{i=1}^k$ be a k-tuple of non-empty standard Borel spaces $\Omega_i = (X_i,\mathcal{B}_i)$, let $\Lambda = (Y,\mathcal{B}')$ be a non-empty standard Borel space with $|Y|\geq 2$, let $\ell\colon \prod_{i=1}^k X_i\times Y\times Y\to\mathbb{R}_{\geq 0}$ be a bounded k-partite loss function and let $\mathcal{S}$ be a k-partite compression scheme for a k-partite hypothesis class $\mathcal{H}\subseteq\mathcal{F}_k(\Omega,\Lambda)$ such that*

$$\ln(h_m^{\mathcal{S}}) + k\cdot s_m^{\mathcal{S}}\cdot\ln(m) \leq o(m) \tag{2.1}$$

as $m\to\infty$ (which in particular follows from $b_{\mathcal{S}}(m)\leq o(m/\ln(m))$ when Y is finite), then $\mathcal{H}$ is k-PAC learnable with respect to ℓ. More specifically, letting $\mathcal{A}$ be given by

$$\mathcal{A}(x,y) \stackrel{\text{def}}{=} \rho_m^{\mathcal{S}}\big(\kappa_m^{\mathcal{S}}(x,y)\big) \qquad \left(m\in\mathbb{N}, x\in\prod_{i=1}^k X_i^m, y\in Y^{([m])_k}\right)$$

gives a k-PAC learner for $\mathcal{H}$ with respect to ℓ with learning guarantee

$$\begin{aligned}
m_{\mathcal{H},\ell,\mathcal{A}}^{\text{PAC}}(\varepsilon,\delta) \stackrel{\text{def}}{=} \min\Bigg\{ m_0\in\mathbb{N}_+, \;\Bigg|\; \forall m\geq m_0, \left(1-\frac{(m-s_m^{\mathcal{S}})^k}{m^k}\right) < \varepsilon \\
\wedge\,(m)_{s_m^{\mathcal{S}}}^k\cdot h_m^{\mathcal{S}}\cdot\exp\left(-\frac{(\varepsilon-(1-(m-s_m^{\mathcal{S}})^k/m^k)\cdot\|\ell\|_\infty)^2\cdot(m-s_m^{\mathcal{S}})}{2\cdot k\cdot\|\ell\|_\infty^2}\right)\leq\delta\Bigg\}.
\end{aligned} \tag{2.2}$$

In particular, if

$$\ln(h_m^{\mathcal{S}}) + k \cdot s_m^{\mathcal{S}} \cdot \ln(m) \leq \big(1 + o(1)\big) \cdot \sqrt{2 \cdot k \cdot m} \cdot \ln(m), \tag{2.3}$$

as $m \to \infty$*, then*

$$m^{\mathrm{PAC}}_{\mathcal{H},\ell,\mathcal{A}}(\varepsilon, \delta) \leq \big(1 + o(1)\big) \cdot \frac{2 \cdot k \cdot \|\ell\|_\infty^2}{\varepsilon^2} \cdot \max\left\{1, \ln\left(\frac{1}{\delta}\right)\right\},$$

where the $o(1)$ *error term is as* $\varepsilon \to 0$ *with* k *fixed, is uniform in* δ*, and has the same uniformity in* $\|\ell\|_\infty$ *as the estimate* $o(1)$ *in* (2.3) *provided* $\|\ell\|_\infty$ *is bounded away from* 0*.*

Proof. We start by showing that $b_{\mathcal{S}}(m) \leq o(m/\ln(m))$ (when Y is finite) implies the condition (2.1) (i.e., $\ln(h_m^{\mathcal{S}}) + k \cdot s_m^{\mathcal{S}} \cdot \ln(m) \leq o(m)$). Indeed, for $m \geq e$, we have

$$\ln(h_m^{\mathcal{S}}) + k \cdot s_m^{\mathcal{S}} \cdot \ln(m) \leq \ln(m) \cdot \big(\log_2(h_m^{\mathcal{S}}) + k \cdot s_m^{\mathcal{S}} \cdot \log_2|Y|\big) \leq \ln(m) \cdot b_{\mathcal{S}}(m) \leq o(m).$$

Let us now argue that the minimum in (2.2) indeed exists. For this, it suffices to argue that there exists $m_0 \in \mathbb{N}$ satisfying both conditions in the definition of the minimum. But indeed, since we are assuming that (2.1) (i.e., $\ln(h_m^{\mathcal{S}}) + k \cdot s_m^{\mathcal{S}} \cdot \ln(m) \leq o(m)$) holds, it follows that both

$$\left(1 - \frac{(m - s_m^{\mathcal{S}})^k}{m^k}\right) \cdot \|\ell\|_\infty,$$

$$(m)^k_{s_m^{\mathcal{S}}} \cdot h_m^{\mathcal{S}} \cdot \exp\left(-\frac{(\varepsilon - (1 - (m - s_m^{\mathcal{S}})^k/m^k) \cdot \|\ell\|_\infty)^2 \cdot (m - s_m^{\mathcal{S}})}{2 \cdot k \cdot \|\ell\|_\infty^2}\right)$$

converge to 0 as $m \to \infty$ (with all other parameters fixed). So there must exist $m_0 \in \mathbb{N}$ large enough so that for every $m \geq m_0$, the first expression above is less than ε and the second expression above is at most δ.

We now prove the result. Let $\mu \in \Pr(\Omega)$ and $F \in \mathcal{F}_k(\Omega, \Lambda)$ be a k-partite hypothesis that is realizable with respect to μ and ℓ. Given $\varepsilon, \delta \in (0,1)$, let $m \geq m^{\mathrm{PAC}}_{\mathcal{H},\ell,\mathcal{A}}(\varepsilon, \delta)$ be an integer and let $\widetilde{\varepsilon}$ be defined in terms of m, $s_m^{\mathcal{S}}$, k and ℓ as in Lemma A.1 and note that since $m \geq m^{\mathrm{PAC}}_{\mathcal{H},\ell,\mathcal{A}}(\varepsilon, \delta)$, the first condition in the definition of the latter implies that the condition (A.1) of Lemma A.1 is satisfied.

We now pick $\boldsymbol{x}$ at random according to μ^m and for each k-tuple $\sigma = (\sigma_i)_{i=1}^k$ of injections $[s_m^{\mathcal{S}}] \to [m]$ and each $\eta \in [h_m^{\mathcal{S}}]$, we let $E_{\sigma,\eta}(\boldsymbol{x})$ be the event

$$L_{\mu,F,\ell}\left(\rho_m^{\mathcal{S}}\left(\sigma^{\#}\big(\boldsymbol{x}, F_m^*(\boldsymbol{x})\big), \eta\right)\right) - L_{\boldsymbol{x},F_m^*(\boldsymbol{x}),\ell}\left(\rho_m^{\mathcal{S}}\left(\sigma^{\#}\big(\boldsymbol{x}, F_m^*(\boldsymbol{x})\big), \eta\right)\right) \geq \varepsilon$$

so that Lemma A.1 says that

$$\mathbb{P}_{\boldsymbol{x}\sim\mu^m}\big[E_{\sigma,\eta}(\boldsymbol{x})\big] \leq \exp\left(-\frac{\widetilde{\varepsilon}^2 \cdot (m - s_m^{\mathcal{S}})}{2 \cdot k \cdot \|\ell\|_\infty^2}\right).$$

If we let $E(\boldsymbol{x})$ be the disjunction of the events $E_{\sigma,\eta}(\boldsymbol{x})$, since there are $(m)^k_{s_m^{\mathcal{S}}}$ many such σ and $h_m^{\mathcal{S}}$ many such η, by the union bound, we get

$$\mathbb{P}_{\boldsymbol{x}\sim\mu^m}\big[E(\boldsymbol{x})\big] \leq (m)^k_{s_m^{\mathcal{S}}} \cdot h_m^{\mathcal{S}} \cdot \exp\left(-\frac{\widetilde{\varepsilon}^2 \cdot (m - s_m^{\mathcal{S}})}{2 \cdot k \cdot \|\ell\|_\infty^2}\right).$$

We now note that we have the compression scheme guarantee that

$$L_{\boldsymbol{x},F_m^*(\boldsymbol{x}),\ell}\left(\rho_m^{\mathcal{S}}\left(\kappa_m^{\mathcal{S}}\big(\boldsymbol{x}, F_m^*(\boldsymbol{x})\big)\right)\right) = 0$$

whenever $(\boldsymbol{x}, F_m^*(\boldsymbol{x}))$ is sample realizable in $\mathcal{H}$ with respect to ℓ, which happens with probability 1 as F is realizable in $\mathcal{H}$ with respect to ℓ.

The above along with the fact that the (random) injection $\sigma_m^{\mathcal{S}}(\boldsymbol{x}, F_m^*(\boldsymbol{x}))$ and the (random) header $\eta_m^{\mathcal{S}}(\boldsymbol{x}, F_m^*(\boldsymbol{x}))$ are elements of $([m])_{s^{\mathcal{S}}}^k$ and $[h_m^{\mathcal{S}}]$, respectively, implies that

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{x}\sim\mu^m}\left[L_{\mu,F,\ell}\Big(\mathcal{A}\big(\boldsymbol{x}, F_m^*(\boldsymbol{x})\big)\Big) > \varepsilon\right] \\
&\quad= \mathbb{P}_{\boldsymbol{x}\sim\mu^m}\left[L_{\mu,F,\ell}\left(\rho_m^{\mathcal{S}}\Big(\kappa_m^{\mathcal{S}}\big(\boldsymbol{x}, F_m^*(\boldsymbol{x})\big)\Big)\right) > \varepsilon\right] \\
&\quad\le \mathbb{P}_{\boldsymbol{x}\sim\mu^m}\big[E(\boldsymbol{x})\big] \\
&\quad\le (m)_{s_m^{\mathcal{S}}}^k \cdot h_m^{\mathcal{S}} \cdot \exp\left(-\frac{\widetilde{\varepsilon}^2 \cdot (m - s_m^{\mathcal{S}})}{2\cdot k \cdot \|\ell\|_\infty^2}\right). \\
&\quad= (m)_{s_m^{\mathcal{S}}}^k \cdot h_m^{\mathcal{S}} \cdot \exp\left(-\frac{(\varepsilon - (1 - (m - s_m^{\mathcal{S}})^k/m^k)\cdot\|\ell\|_\infty)^2 \cdot (m - s_m^{\mathcal{S}})}{2\cdot k \cdot \|\ell\|_\infty^2}\right) \\
&\quad\le \delta,
\end{aligned}$$

where the last inequality follows from $m \ge m_{\mathcal{H},\ell,\mathcal{A}}^{\mathrm{PAC}}(\varepsilon,\delta)$ and using the second condition in the definition of the latter. Thus $\mathcal{A}$ is a k-PAC learner with respect to ℓ.

It remains to make the asymptotic estimate of $m_{\mathcal{H},\ell,\mathcal{A}}^{\mathrm{PAC}}(\varepsilon,\delta)$ when the stronger assumption (2.3) (i.e., $\ln(h_m^{\mathcal{S}}) + k\cdot s_m^{\mathcal{S}}\cdot\ln(m) \le (1+o(1))\cdot\sqrt{2\cdot k\cdot m}\cdot\ln(m)$) holds. (It is clear that (2.3) implies the condition (2.1) that $\ln(h_m^{\mathcal{S}}) + k\cdot s_m^{\mathcal{S}}\cdot\ln(m) \le o(m)$.)

Let $m_1(\varepsilon) \in \mathbb{N}_+$ be the least integer such that for every $m \ge m_1(\varepsilon)$, we have

$$\left(1 - \frac{(m - s^{\mathcal{S}})^k}{m^k}\right)\cdot\|\ell\|_\infty < \varepsilon \tag{A.5}$$

and let $m_2(\varepsilon,\delta)$ be the least integer such that for every $m \ge m_2(\varepsilon,\delta)$, we have

$$(m)_{s_m^{\mathcal{S}}}^k \cdot h_m^{\mathcal{S}} \cdot \exp\left(-\frac{(\varepsilon - (1 - (m - s_m^{\mathcal{S}})^k/m^k)\cdot\|\ell\|_\infty)^2 \cdot (m - s_m^{\mathcal{S}})}{2\cdot k \cdot \|\ell\|_\infty^2}\right) \le \delta, \tag{A.6}$$

Since $m_{\mathcal{H},\ell,\mathcal{A}}^{\mathrm{PAC}}(\varepsilon,\delta) = \max\{m_1(\varepsilon), m_2(\varepsilon,\delta)\}$, it suffices to give asymptotic estimates for $m_1(\varepsilon)$ and $m_2(\varepsilon,\delta)$.

To estimate $m_1(\varepsilon)$, note that if $\|\ell\|_\infty < \varepsilon$, then $m_1(\varepsilon) = 1$, so suppose that $\|\ell\|_\infty \ge \varepsilon$ and note that

$$\begin{aligned}
0 \le 1 - \frac{(m - s_m^{\mathcal{S}})^k}{m^k} &= 1 - \left(1 - \frac{s_m^{\mathcal{S}}}{m}\right)^k \\
&\le 1 - \left(1 - \big(1+o(1)\big)\cdot\sqrt{\frac{2}{k\cdot m}}\right)^k = \big(1+o(1)\big)\cdot\sqrt{\frac{2\cdot k}{m}}
\end{aligned} \tag{A.7}$$

as $m \to \infty$ with k fixed.

Plugging this asymptotic estimate in the definition (A.5) of $m_1(\varepsilon)$, we conclude that

$$m_1(\varepsilon) \le \big(1+o(1)\big)\frac{2\cdot k\cdot\|\ell\|_\infty^2}{\varepsilon^2} \tag{A.8}$$

where the $o(1)$ error term is as $\varepsilon \to 0$ with k fixed, is uniform in δ (as $m_1(\varepsilon)$ does not depend on δ at all), and has the same uniformity in $\|\ell\|_\infty$ as the estimate $o(1)$ in (2.3) provided $\|\ell\|_\infty$ is bounded away from 0 (this is to cover the case $\|\ell\|_\infty < \varepsilon$ when $m_1(\varepsilon) = 1$).

To estimate $m_2(\varepsilon,\delta)$, we use (A.7) along with the weaker condition (2.1) (i.e., $\ln(h_m^{\mathcal{S}}) + k\cdot s_m^{\mathcal{S}}\cdot\ln(m) \le o(m)$) to get

$$\begin{aligned}
&(m)_{s_m^{\mathcal{S}}}^k \cdot h_m^{\mathcal{S}} \cdot \exp\left(-\frac{(\varepsilon - (1 - (m - s_m^{\mathcal{S}})^k/m^k)\cdot\|\ell\|_\infty)^2 \cdot (m - s_m^{\mathcal{S}})}{2\cdot k \cdot \|\ell\|_\infty^2}\right) \\
&\quad\le \exp\left(-\frac{(\varepsilon - o(1)\cdot\|\ell\|_\infty)^2\cdot(1+o(1))\cdot m}{2\cdot k\cdot\|\ell\|_\infty^2} + o(m)\right) \\
&\quad\le \exp\left(-\left(\frac{\varepsilon^2}{2\cdot k\cdot\|\ell\|_\infty^2} - o(1)\right)\cdot m\right),
\end{aligned}$$

where the error terms are as $m \to \infty$ with k fixed, are uniform in ε, and have the same uniformity in $\|\ell\|_\infty$ as the estimate $o(1)$ in (2.3).

Plugging this asymptotic estimate in the definition (A.6) of $m_2(\varepsilon, \delta)$, we conclude that

$$m_2(\varepsilon, \delta) \leq \left(\frac{2 \cdot k \cdot \|\ell\|^2}{\varepsilon^2} + o(1) \right) \cdot \ln\left(\frac{1}{\delta}\right),$$

where the $o(1)$ error term is as $\varepsilon \to 0$, is uniform in δ, and has the same uniformity in $\|\ell\|_\infty$ as the estimate $o(1)$ in (A.14).

Putting the above together with our estimate (A.8) for $m_1(\varepsilon)$, we get

$$m^{\mathrm{PAC}}_{\mathcal{H},\ell,\mathcal{A}}(\varepsilon, \delta) \leq \big(1 + o(1)\big) \cdot \frac{2 \cdot k \cdot \|\ell\|^2_\infty}{\varepsilon^2} \cdot \max\left\{1, \ln\left(\frac{1}{\delta}\right)\right\},$$

as desired since $m^{\mathrm{PAC}}_{\mathcal{H},\ell,\mathcal{A}} = \max\{m_1(\varepsilon), m_2(\varepsilon, \delta)\}$. □

A.2 Non-partite setting

In this section, we cover the non-partite case, that is, the case in which our hypotheses classes are collections of functions of the form $F\colon X^k \to Y$, k-PAC learning notions are with respect to power measures μ^k of the same probability measure on X and sample are thought of as having m points $x_1, \ldots, x_m$ from X as well as labels $F(x_{\alpha_1}, \ldots, x_{\alpha_k}) \in Y$ for each injective k-tuple $\alpha \in ([m])_k$.

Let us point out some of the peculiarities of the non-partite case:

- Labels on k-ary samples only take into account injective tuples, that is, once we are given m points $x_1, \ldots, x_m \in X$, to form a labeled sample, we only collect labels from a hypothesis $F\colon X^k \to Y$ of the form $F(x_{\alpha_1}, \ldots, x^k_{\alpha_k}) \in Y$ with $\alpha\colon [k] \to [m]$ injective. The reason why we do not collect labels of non-injective tuples is technical and comes from the agnostic setup via exchangeability theory. Nevertheless, we point out that even if we included non-injective tuples, since these account for a negligible proportion of all tuples (for m large), this would only change bounds on lower order terms by discounting non-injective tuples from calculations.
- Loss functions ℓ collect labels from all orientations of a k-set, that is, given a k-tuple $(x_1, \ldots, x_k) \in X^k$ and hypotheses $F, H\colon X^k \to Y$, we collect the labels of the hypotheses on all possible orientations of the k-tuple $(x_1, \ldots, x_k)$ into an element of Y^{S_k} (recall that $S_k = ([k])_k$ is the symmetric group on $[k]$) before feeding them into the loss function. This accounts for the fact that X does not come with an inherent ordering and allows loss functions to take into account other orientations of the k-tuples before assigning a penalty.
- When computing the empirical loss on a sample of m points $x_1, \ldots, x_m$, the k-tuples of it do not come with an inherent order, so there must be an order choice that picks one particular orientation for each k-subset of $[m]$ as the "standard direction" in which the loss function is going to be computed.

We again follow a simplified notation of Coregliano and Malliaris [2024].

Definition A.2 (Non-partite setting)**.** Let $k \in \mathbb{N}_+$, let $\Omega = (X, \mathcal{B})$ and $\Lambda = (Y, \mathcal{B}')$ be non-empty standard Borel spaces and let $m \in \mathbb{N}$.

1. A *k-ary unlabeled sample of size m* with respect to Ω is an element of X^m. A *k-ary labeled sample* of size m with respect to Ω and Λ is an element of $X^m \times Y^{([m])_k}$; we think of the coordinate y_α indexed by some injective k-tuple $\alpha \in ([m])_k$ as the label on $x_{\alpha_1}, \ldots, x_{\alpha_k}$.

   More generally, every injection $\alpha\colon [n] \to [m]$ induces naturally a *subsample* $\alpha^*(x)$ of size n by picking the points and labels corresponding to the coordinates indexed by α (in the order of α). This is formalized as follows: we contra-variantly define the maps

   $$\alpha^*\colon X^m \to X^n, \qquad \alpha^*\colon Y^{([m])_k} \to Y^{([n])_k}$$

that select the points and labels according to

$$\alpha^*(x)_i \stackrel{\text{def}}{=} x_{\alpha(i)} \qquad \big(x \in X^m, i \in [n]\big),$$
$$\alpha^*(y)_\beta \stackrel{\text{def}}{=} y_{\alpha\circ\beta} \qquad \big(y \in Y^{([m])_k}, \beta \in ([m])_k\big).$$

With slight abuse of notation, we also contra-variantly define the map

$$\alpha^*\colon X^m \times Y^{([m])_k} \to X^n \times Y^{([n])_k}$$

by letting it act as the previous two maps in the corresponding coordinates.

2. A *k-ary hypothesis* from Ω to Λ is a measurable function $F\colon X^k \to Y$. The set of k-ary hypothesis from Ω to Λ is denoted $\mathcal{F}_k(\Omega, \Lambda)$.

   Given a k-ary hypothesis $F \in \mathcal{F}_k(\Omega, \Lambda)$ and $m \in \mathbb{N}$, we define the function $F^*_m\colon X^m \to Y^{([m])_k}$ that when given an unlabeled sample of size m, provides the labels of F on the injective k-tuples of the sample; in a formula:

   $$F^*_m(x)_\alpha \stackrel{\text{def}}{=} F\big(\alpha^*(x)\big) \qquad \big(x \in X^m, \alpha \in ([m])_k\big).$$

   Similarly to the partite case, this definition is equivariant in the sense that the following diagram is commutative:

   $$\begin{array}{ccc} X^m & \xrightarrow{F^*_m} & Y^{([m])_k} \\ {\scriptstyle\alpha^*}\downarrow & & \downarrow{\scriptstyle\alpha^*} \\ X^n & \xrightarrow{F^*_n} & Y^{([n])_k} \end{array}$$

3. A *k-ary hypothesis class* is a set $\mathcal{H} \subseteq \mathcal{F}_k(\Omega, \Lambda)$ of k-ary hypotheses that is further equipped with a σ-algebra such that:

   - the evaluation map $\mathrm{ev}\colon \mathcal{H} \times X^k \to Y$ given by $\mathrm{ev}(H, x) \stackrel{\text{def}}{=} H(x)$ is measurable;
   - for every $H \in \mathcal{H}$, the singleton $\{H\}$ is measurable;
   - for every Borel space Υ and every measurable set $A \subseteq \mathcal{H} \times \Upsilon$, the projection of A onto Υ, i.e., the set $\{\upsilon \in \Upsilon \mid \exists H \in \mathcal{H}, (H, \upsilon) \in A\}$ is universally measurable[3].

4. A *k-ary loss function* over Λ is a measurable function $\ell\colon X^k \times Y^{S_k} \times Y^{S_k} \to \mathbb{R}_{\geq 0}$ that when given a k-tuple x and two labels y and y' of all possible orientations of the k-tuple x, assigns a value $\ell(x, y, y')$ that is the penalty of guessing y on the tuple x when the correct label was y'.

   We define $\|\ell\|_\infty \stackrel{\text{def}}{=} \sup_{x\in X^k, y, y' \in Y^{S_k}} \ell(x, y, y')$ and we say that ℓ is *bounded* if $\|\ell\|_\infty < \infty$. Given further $\mu \in \Pr(\Omega)$ and k-ary hypotheses F and H, the *total loss* of H with respect to μ, F and ℓ is $L_{\mu,F,\ell}(H) \stackrel{\text{def}}{=} \mathbb{E}_{\boldsymbol{x}\sim\mu^k}[\ell(\boldsymbol{x}, H^*_k(\boldsymbol{x}), F^*_k(\boldsymbol{x}))]$.

   We say that F is *realizable* in a k-ary hypothesis class $\mathcal{H} \subseteq \mathcal{F}_k(\Omega, \Lambda)$ with respect to μ and ℓ if $\inf_{H\in\mathcal{H}} L_{\mu,F,\ell}(H) = 0$.

5. For a k-ary hypothesis class $\mathcal{H}' \subseteq \mathcal{F}_k(\Omega, \Lambda)$, a *(k-ary) learning algorithm* with outputs in $\mathcal{H}'$ is a measurable function

   $$\mathcal{A}\colon \bigcup_{m\in\mathbb{N}} (X^m \times Y^{([m])_k}) \to \mathcal{H}'.$$

   We say that a k-ary hypothesis class $\mathcal{H} \subseteq \mathcal{F}_k(\Omega, \Lambda)$ is *improperly k-PAC learnable* with respect to a k-ary loss function $\ell\colon X^k \times Y^{S_k} \times Y^{S_k} \to \mathbb{R}_{\geq 0}$ if there exists a learning algorithm $\mathcal{A}$ with outputs in some k-ary hypothesis class $\mathcal{H}'$ and a function

[3]Similarly to the partite case, the reader unfamiliar with these technical measurability conditions can simply assume that $\mathcal{H}$ is equipped with a σ-algebra that makes it a standard Borel space as it will imply this condition; or they can simply interpret this as "all probabilities computed must make sense".

$m^{\mathrm{PAC}}_{\mathcal{H},\ell,\mathcal{A}}\colon (0,1)^2 \to \mathbb{R}_{\geq 0}$ such that for every $\varepsilon, \delta \in (0,1)$, every $\mu \in \Pr(\Omega)$, every $F \in \mathcal{F}_k(\Omega,\Lambda)$ that is realizable in $\mathcal{H}$ with respect to μ and ℓ and every integer $m \geq m^{\mathrm{PAC}}_{\mathcal{H},\ell,\mathcal{A}}(\varepsilon,\delta)$, we have

$$\mathbb{P}_{\boldsymbol{x}\sim\mu^m}\left[L_{\mu,F,\ell}\Big(\mathcal{A}\big(\boldsymbol{x}, F^*_m(\boldsymbol{x})\big)\Big) \leq \varepsilon\right] \geq 1-\delta.$$

In plain English: when given a random sample of size m drawn from μ^m labeled according to some realizable F, the algorithm $\mathcal{A}$ outputs some hypothesis H with total loss at most ε with probability at least $1-\delta$ over the randomness of the sample.

The function $m^{\mathrm{PAC}}_{\mathcal{H},\ell,\mathcal{A}}$ is called the *learning guarantee* of $\mathcal{A}$ and $\mathcal{A}$ is called an *improper k-PAC learner* for $\mathcal{H}$ with respect to ℓ.

The notions of *proper k-PAC learnability* and *proper k-PAC learner* are defined analogously but requiring further that $\mathcal{H}' = \mathcal{H}$.

6. A *(k-ary) order choice* for $[m]$ is a sequence $\alpha = (\alpha_U)_{U\in\binom{V}{k}}$ of injections of the form $\alpha_U\colon [k] \to [m]$ with $\mathrm{im}(\alpha_U) = U$.

   Any such order choice α naturally induces a map $b_\alpha\colon Y^{([m])_k} \to (Y^{S_k})^{\binom{[m]}{k}}$ that bundles up all labels corresponding to orientations of a k-set U into the coordinate indexed by U, where the standard orientation is considered to be in the direction of α_U; formally, the map b_α is given by

$$\big(b_\alpha(y)_U\big)_\pi \stackrel{\mathrm{def}}{=} y_{\alpha_U\circ\pi} \qquad \left(y \in Y^{([m])_k}, U \in \binom{[m]}{k}, \pi \in S_k\right).$$

   Given further k-ary labeled sample $(x,y) \in (X^m \times Y^{([m])_k}$ of size m, a k-ary loss function $\ell\colon X^k \times Y^{S_k} \times Y^{S_k} \to \mathbb{R}_{\geq 0}$ and a hypothesis $H \in \mathcal{F}_k(\Omega,\Lambda)$, the *empirical loss* (or *empirical risk*) of H with respect to (x,y), ℓ and α is

$$L^\alpha_{x,y,\ell}(H) \stackrel{\mathrm{def}}{=} \begin{cases} \dfrac{1}{\binom{m}{k}} \cdot \displaystyle\sum_{U\in\binom{[m]}{k}} \ell\Big(\alpha^*_U(x), b_\alpha\big(H^*_m(x)\big)_U, b_\alpha(y)_U\Big), & \text{if } m \geq k, \\ 0, & \text{if } m < k. \end{cases}$$

   We say that (x,y) is *sample realizable* in a k-ary hypothesis class $\mathcal{H} \subseteq \mathcal{F}_k(\Omega,\Lambda)$ with respect to ℓ and α if $\inf_{H\in\mathcal{H}} L^\alpha_{x,y,\ell}(H) = 0$. It is straightforward to check that if $\mu \in \Pr(\Omega)$, $F \in \mathcal{F}_k(\Omega,\Lambda)$ is realizable in $\mathcal{H}$ with respect to μ and ℓ and $\boldsymbol{x} \sim \mu^m$, then with probability 1, we have that $(\boldsymbol{x}, F^*_m(\boldsymbol{x}))$ is sample realizable in $\mathcal{H}$ with respect to ℓ and α.

The next definition is that of a high-arity selection/compression scheme in the non-partite case. Note that since the definition of a compression schemes rely on empirical loss, in the non-partite case, we will also need to specify an order choice.

Definition A.3 (Non-partite setting)**.** Let $k \in \mathbb{N}_+$ and let $\Omega = (X,\mathcal{B})$ and $\Lambda = (Y,\mathcal{B}')$ be non-empty standard Borel spaces.

1. A *k-ary selection scheme* with outputs in a k-ary hypothesis class $\mathcal{H}' \subseteq \mathcal{F}_k(\Omega,\Lambda)$ is a tuple $\mathcal{S} = (\sigma,\eta,\rho)$ such that:

   - $\sigma = (\sigma_m)_{m\in\mathbb{N}}$ is a sequence of maps, called *selection maps*, such that each σ_m when given a sample of size m, outputs a tuple of some length s_m that contains the indices of the subsample selected. Formally, each σ_m is a measurable function

$$\sigma_m\colon X^m \times Y^{([m])_k} \to ([m])_{s_m}$$

     where $s_m \in \mathbb{N}$ ($i \in [k]$) is some integer, which we call the *selection size* (note that we must have $s_m \leq m$ as the domain of σ_m is never empty).

- $\eta = (\eta_m)_{m\in\mathbb{N}}$ is a sequence of maps, called the *header maps*, that when given a sample of size m, outputs some extra information, encoded as an element of $[h_m]$ for some fixed $h_m \in \mathbb{N}$, that will aid in the reconstruction of the original sample. Formally, each η_m is a measurable function
$$\eta_m \colon X^m \times Y^{([m])_k} \to [h_m],$$
where $h_m \in \mathbb{N}_+$ is called the *header size* of η_m.
- $\rho = (\rho_m)_{m\in\mathbb{N}}$ is a sequence of maps, called the *reconstruction maps*, that when given a labeled sample of size s_m and header of size h_m (intended to have come from σ_m and η_m), reconstruct a labeled sample of size m. Formally, each ρ_m is a measurable function
$$\rho_m \colon X^{s_m} \times Y^{([s_m])_k} \times [h_m] \to \mathcal{H}'.$$

The compression procedure of $\mathcal{S}$ is formalized by defining a sequence $\kappa = (\kappa_m)_{m\in\mathbb{N}}$ of *compression maps* of the form
$$\kappa_m \colon X^m \times Y^{([m])_k} \to X^{s_m} \times Y^{([s_m])_k} \times [h_m]$$
by
$$\kappa_m(x,y) \stackrel{\text{def}}{=} \big(\sigma_m(x,y)^*(x,y), \eta_m(x,y)\big) \qquad \big(x \in X^m, y \in Y^{([m])_k}\big).$$
(Recall from Definition A.2.1 that since $\sigma_m(x,y) \in ([m])_{s_m}$ is a injection, it contravariantly induces a map that selects a subsample $\sigma_m(x,y)^*(x,y)$ of size s_m from the sample (x,y) of size m and the map κ_m then takes this subsample an appends to it the header $\eta_m(x,y)$.)

We will use the notation
$$\sigma^{\mathcal{S}} \stackrel{\text{def}}{=} \sigma = (\sigma_m)_{m\in\mathbb{N}}, \qquad \eta^{\mathcal{S}} \stackrel{\text{def}}{=} \eta = (\eta_m)_{m\in\mathbb{N}}, \qquad \rho^{\mathcal{S}} \stackrel{\text{def}}{=} \rho = (\rho_m)_{m\in\mathbb{N}},$$
$$s_m^{\mathcal{S}} \stackrel{\text{def}}{=} s_m, \qquad h_m^{\mathcal{S}} \stackrel{\text{def}}{=} h_m, \qquad \kappa^{\mathcal{S}} \stackrel{\text{def}}{=} \kappa = (\kappa_m)_{m\in\mathbb{N}}.$$

The *compression size* and the *compression bitlength* functions of $\mathcal{S}$ are the functions $c_{\mathcal{S}}, b_{\mathcal{S}} \colon \mathbb{N} \to \mathbb{R}_{\geq 0}$ given by
$$c_{\mathcal{S}}(m) \stackrel{\text{def}}{=} h_m \cdot |Y|^{(s_m)_k}, \qquad b_{\mathcal{S}}(m) \stackrel{\text{def}}{=} \log_2\big(c_{\mathcal{S}}(m)\big).$$

2. A k-ary selection scheme $\mathcal{S}$ with outputs in a k-ary hypothesis class $\mathcal{H}' \subseteq \mathcal{F}_k(\Omega,\Lambda)$ is called a *sample compression scheme* for a k-ary hypothesis class $\mathcal{H} \subseteq \mathcal{F}_k(\Omega,\Lambda)$ with respect to a k-ary loss function $\ell \colon X^k \times Y^{S_k} \times Y^{S_k} \to \mathbb{R}_{\geq 0}$ if for every $m \in \mathbb{N}$ and every labeled sample $(x,y) \in X^m \times Y^{([m])_k}$ of size m that is sample realizable with respect to ℓ and every order choice α for $[m]$, we have
$$L^{\alpha}_{x,y,\ell}\Big(\rho^{\mathcal{S}}_m\big(\kappa^{\mathcal{S}}_m(x,y)\big)\Big) = 0,$$
or in plain English: if we compress (x,y) via $\kappa^{\mathcal{S}}$ then reconstruct via $\rho^{\mathcal{S}}$, then the result has zero empirical loss.

We now prove a concentration lemma based on Azuma's Inequality that is the non-partite analogue of Lemma A.1; the main differences of this lemma to its partite counterpart is that we need to keep track of the order choice α and the bounds are slightly different due to normalization by $\binom{m}{k}$ as opposed to m^k and the fact that the martingale has n steps instead of $k \cdot n$ steps.

Lemma A.4. *Let $\Omega = (X,\mathcal{B})$ and $\Lambda = (Y,\mathcal{B}')$ be non-empty standard Borel spaces, let $k \in \mathbb{N}_+$, let $\mathcal{S}$ be a k-ary selection scheme and let $\ell \colon X^k \times Y^{S_k} \times Y^{S_k} \to \mathbb{R}_{\geq 0}$ be a bounded k-ary loss function.*

Then for every $\varepsilon > 0$, every probability measure $\mu \in \Pr(\Omega)$, every integer $m \in \mathbb{N}$ such that
$$\left(1 - \frac{(m - s_m^{\mathcal{S}})_k}{(m)_k}\right) \cdot \|\ell\|_\infty < \varepsilon, \tag{A.9}$$

every order choice α for $[m]$, every $\eta \in [h_m^{\mathcal{S}}]$, every injection $\sigma\colon [s_m^{\mathcal{S}}] \to [m]$ and every $F \in \mathcal{F}_k(\Omega, \Lambda)$, we have

$$\mathbb{P}_{\boldsymbol{x}\sim\mu^m}\left[L_{\mu,F,\ell}\Big(\rho^{\mathcal{S}}\Big(\sigma^*\big(\boldsymbol{x}, F_m^*(\boldsymbol{x})\big), \eta\Big)\Big) - L^{\alpha}_{\boldsymbol{x},F_m^*(\boldsymbol{x}),\ell}\Big(\rho^{\mathcal{S}}\Big(\sigma^*\big(\boldsymbol{x}, F_m^*(\boldsymbol{x})\big), \eta\Big)\Big) \geq \varepsilon\right] \leq \exp\left(-\frac{\widetilde{\varepsilon}^2 \cdot (m - s_m^{\mathcal{S}})}{2 \cdot k^2 \cdot \|\ell\|_\infty^2}\right),$$

where

$$\widetilde{\varepsilon} \stackrel{\text{def}}{=} \varepsilon - \left(1 - \frac{(m - s_m^{\mathcal{S}})_k}{(m)_k}\right) \cdot \|\ell\|_\infty > 0.$$

In particular, if $s_m^{\mathcal{S}} \leq o(m)$ as $m \to \infty$, then

$$\mathbb{P}_{\boldsymbol{x}\sim\mu^m}\left[L_{\mu,F,\ell}\Big(\rho^{\mathcal{S}}\Big(\sigma^*\big(\boldsymbol{x}, F_m^*(\boldsymbol{x})\big), \eta\Big)\Big) - L^{\alpha}_{\boldsymbol{x},F_m^*(\boldsymbol{x}),\ell}\Big(\rho^{\mathcal{S}}\Big(\sigma^*\big(\boldsymbol{x}, F_m^*(\boldsymbol{x})\big), \eta\Big)\Big) \geq \varepsilon\right] \leq \exp\left(-\frac{\varepsilon^2}{2 \cdot k^2 \cdot \|\ell\|_\infty^2} \cdot m + \frac{\varepsilon \cdot (\varepsilon + \|\ell\|_\infty)}{\|\ell\|_\infty^2} \cdot o(m)\right),$$

where the $o(m)$ error term is as $m \to \infty$ with k fixed, is uniform in ε, and has the same uniformity in $\|\ell\|_\infty$ as the estimate $s_m^{\mathcal{S}} \leq o(m)$.

Proof. The result is trivial if $m < k$, so suppose $m \geq k$. By symmetry, we may also suppose without loss of generality that $\sigma\colon [s_m^{\mathcal{S}}] \to [m]$ is such that $\operatorname{im}(\sigma) = \{m - s_m^{\mathcal{S}} + 1, m - s_m^{\mathcal{S}} + 2, \ldots, m\}$. Let us pick $\boldsymbol{x}$ at random according to μ^m and for ease of notation, we let

$$n \stackrel{\text{def}}{=} m - s_m^{\mathcal{S}}, \qquad \boldsymbol{H} \stackrel{\text{def}}{=} \rho_m^{\mathcal{S}}\Big(\sigma^*\big(\boldsymbol{x}, F_m^*(\boldsymbol{x})\big), \eta\Big), \qquad \boldsymbol{x'} \stackrel{\text{def}}{=} \iota^*(\boldsymbol{x}),$$

where $\iota\colon [n] \to [m]$ is the inclusion map (so $\boldsymbol{x'}$ is the random sample corresponding to the first n coordinates of $\boldsymbol{x}$, in their natural order). We also let $\alpha' \stackrel{\text{def}}{=} \alpha|_{\binom{[n]}{k}}$ be the order choice for $[n]$ obtained by restricting α to the k-subsets of $[n]$.

Note that

$$\begin{aligned}
&L^{\alpha}_{\boldsymbol{x},F_m^*(\boldsymbol{x}),\ell}(\boldsymbol{H}) - L^{\alpha'}_{\boldsymbol{x'},F_n^*(\boldsymbol{x'}),\ell}(\boldsymbol{H}) \\
&\quad= \frac{1}{\binom{m}{k}} \cdot \sum_{U \in \binom{[m]}{k}} \ell\Big(\alpha_U^*(\boldsymbol{x}), b_\alpha\big(\boldsymbol{H}_m^*(\boldsymbol{x})\big)_U, b_\alpha\big(F_m^*(\boldsymbol{x})\big)_U\Big) \\
&\qquad\quad - \frac{1}{\binom{n}{k}} \cdot \sum_{U \in \binom{[n]}{k}} \ell\Big(\alpha_U^*(\boldsymbol{x}), b_\alpha\big(\boldsymbol{H}_m^*(\boldsymbol{x})\big)_U, b_\alpha\big(F_m^*(\boldsymbol{x})\big)_U\Big) \\
&\quad= \left(\frac{1}{\binom{m}{k}} - \frac{1}{\binom{n}{k}}\right) \cdot \sum_{U \in \binom{[n]}{k}} \ell\Big(\alpha_U^*(\boldsymbol{x}), b_\alpha\big(\boldsymbol{H}_m^*(\boldsymbol{x})\big)_U, b_\alpha\big(F_m^*(\boldsymbol{x})\big)_U\Big) \\
&\qquad\quad + \frac{1}{\binom{m}{k}} \cdot \sum_{U \in \binom{[m]}{k} \setminus \binom{[n]}{k}} \ell\Big(\alpha_U^*(\boldsymbol{x}), b_\alpha\big(\boldsymbol{H}_m^*(\boldsymbol{x})\big)_U, b_\alpha\big(F_m^*(\boldsymbol{x})\big)_U\Big),
\end{aligned}$$

and since ℓ is bounded and noting that the first term in the above is non-positive (as $n \leq m$) while the second is non-negative, we conclude that

$$\begin{aligned}
&\big|L^{\alpha}_{\boldsymbol{x},F_m^*(\boldsymbol{x}),\ell}(\boldsymbol{H}) - L^{\alpha'}_{\boldsymbol{x'},F_n^*(\boldsymbol{x'}),\ell}(\boldsymbol{H})\big| \\
&\qquad\leq \|\ell\|_\infty \cdot \max\left\{\left(\frac{1}{\binom{n}{k}} - \frac{1}{\binom{m}{k}}\right) \cdot \binom{n}{k}, \frac{1}{\binom{m}{k}} \cdot \left(\binom{m}{k} - \binom{n}{k}\right)\right\} \qquad \text{(A.10)} \\
&\qquad= \left(1 - \frac{(n)_k}{(m)_k}\right) \cdot \|\ell\|_\infty.
\end{aligned}$$

Since we are mostly concerned with the setting in which $s_m^{\mathcal{S}} \leq o(m)$ (or equivalently, $n = (1 - o(1)) \cdot m$), the difference above is small, it will suffice to compare $L_{\mu,F,\ell}(\boldsymbol{H})$ with $L^{\alpha'}_{\boldsymbol{x'},F_n^*(\boldsymbol{x'}),\ell}(\boldsymbol{H})$ (instead of comparing with $L^{\alpha}_{\boldsymbol{x},F_m^*(\boldsymbol{x}),\ell}(\boldsymbol{H})$ directly).

Let $\mathcal{F}_0$ be the σ-algebra generated by $\boldsymbol{x}_{m-s_m^{\mathcal{S}}+1}, \ldots, \boldsymbol{x}_m$ (i.e., generated by $\sigma^*(\boldsymbol{x})$), so $\boldsymbol{H}$ is $\mathcal{F}_0$-measurable. For each $i \in [m]$, let $\mathcal{F}_i$ be the σ-algebra generated by $\mathcal{F}_0$ and $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_i)$ and let $\boldsymbol{Z}_i \stackrel{\text{def}}{=} \mathbb{E}[L^{\alpha'}_{\boldsymbol{x'},F_n^*(\boldsymbol{x'}),\ell}(\boldsymbol{H}) \mid \mathcal{F}_i]$.

We note that $(\boldsymbol{Z}_i)_{i=0}^n$ forms a (Doob) martingale with respect to $(\mathcal{F}_i)_{i=0}^n$:

$$\mathbb{E}[\boldsymbol{Z}_{i+1} \mid \mathcal{F}_i] = \mathbb{E}\Big[\mathbb{E}\big[L^{\alpha'}_{\boldsymbol{x'},F_n^*(\boldsymbol{x'}),\ell}(\boldsymbol{H}) \;\big|\; \mathcal{F}_{i+1}\big] \;\Big|\; \mathcal{F}_i\Big] = \mathbb{E}\big[L^{\alpha'}_{\boldsymbol{x'},F_n^*(\boldsymbol{x'}),\ell}(\boldsymbol{H}) \;\big|\; \mathcal{F}_i\big] = \boldsymbol{Z}_i.$$

We also note that

$$\begin{aligned}
\boldsymbol{Z}_0 &= \mathbb{E}\big[L^{\alpha'}_{\boldsymbol{x'},F_n^*(\boldsymbol{x'}),\ell}(\boldsymbol{H}) \;\big|\; \mathcal{F}_0\big] = L_{\mu,F,\ell}(\boldsymbol{H}),\\
\boldsymbol{Z}_n &= \mathbb{E}\big[L^{\alpha'}_{\boldsymbol{x'},F_n^*(\boldsymbol{x'}),\ell}(\boldsymbol{H}) \;\big|\; \mathcal{F}_n\big] = L^{\alpha'}_{\boldsymbol{x'},F_n^*(\boldsymbol{x'}),\ell}(\boldsymbol{H}),
\end{aligned}$$

where the former is by linearity of conditional expectation (and the fact that $\boldsymbol{H}$ is $\mathcal{F}_0$-measurable) and the latter is simply because the underlying random variable is $\mathcal{F}_n$-measurable.

We will show that $\boldsymbol{Z}_0$ and $\boldsymbol{Z}_n$ are close with high (conditional) probability using Azuma's Inequality. Toward that, let us bound $|\boldsymbol{Z}_i - \boldsymbol{Z}_{i-1}|$ for each $i \in [n]$. By linearity of conditional expectation, we have

$$\begin{aligned}
|\boldsymbol{Z}_i - \boldsymbol{Z}_{i-1}| = \Bigg| \frac{1}{\binom{n}{k}} \cdot \sum_{U \in \binom{[n]}{k}} \Bigg(&\mathbb{E}\Big[\ell\Big(\alpha_U^*(\boldsymbol{x}), b_\alpha\big(\boldsymbol{H}_m^*(x)\big)_U, b_\alpha\big(F_m^*(x)\big)_U\Big) \;\Big|\; \mathcal{F}_i\Big] \\
&- \mathbb{E}\Big[\ell\Big(\alpha_U^*(\boldsymbol{x}), b_\alpha\big(\boldsymbol{H}_m^*(x)\big)_U, b_\alpha\big(F_m^*(x)\big)_U\Big) \;\Big|\; \mathcal{F}_{i+1}\Big]\Bigg)\Bigg|. \quad \text{(A.11)}
\end{aligned}$$

We now argue that most of the terms in the above cancel out. Namely, consider a term corresponding to $U \in \binom{[n]}{k}$ such that $i \notin U$, then

$$\ell\Big(\alpha_U^*(\boldsymbol{x}), b_\alpha\big(\boldsymbol{H}_m^*(x)\big)_U, b_\alpha\big(F_m^*(x)\big)_U\Big)$$

is independent from $\boldsymbol{x}_i$; in fact, since $(\boldsymbol{x}_j)_{j=1}^m$ is mutually independent, the above is conditionally independent from $\boldsymbol{x}_i$ given $\mathcal{F}_{i-1}$, so the above gives the same value when conditioned on $\mathcal{F}_{i-1}$ as it does when conditioned on $\mathcal{F}_i$. Thus, whenever $i \notin U$, the corresponding term in (A.11) cancels out.

As there are exactly $\binom{n-1}{k-1}$ many $U \in \binom{[n]}{k}$ with $i \in U$, using triangle inequality and the fact that ℓ is bounded, we conclude that

$$|\boldsymbol{Z}_i - \boldsymbol{Z}_{i-1}| \leq \frac{\binom{n-1}{k-1}}{\binom{n}{k}} \cdot \|\ell\|_\infty = \frac{k}{n} \cdot \|\ell\|_\infty.$$

We now note that since

$$\widetilde{\varepsilon} \stackrel{\text{def}}{=} \varepsilon - \left(1 - \frac{(m - s_m^{\mathcal{S}})_k}{(m)_k}\right) \cdot \|\ell\|_\infty = \varepsilon - \left(1 - \frac{(n)_k}{(m)_k}\right) \cdot \|\ell\|_\infty,$$

the condition (A.9) implies $\widetilde{\varepsilon} > 0$, so by Azuma's Inequality (conditioned on $\mathcal{F}_0$), we get

$$\begin{aligned}\mathbb{P}_{\boldsymbol{x}\sim\mu^m}\left[L_{\mu,F,\ell}(\boldsymbol{H}) - L^{\alpha'}_{\boldsymbol{x}',F^*_n(\boldsymbol{x}'),\ell}(\boldsymbol{H}) \geq \widetilde{\varepsilon}\right] &= \mathbb{P}[\boldsymbol{Z}_n - \boldsymbol{Z}_0 \leq -\widetilde{\varepsilon}] \\ &\leq \exp\left(-\frac{\widetilde{\varepsilon}^2}{2\cdot\sum_{i=1}^n (k\cdot\|\ell\|_\infty/n)^2}\right) \\ &= \exp\left(-\frac{\widetilde{\varepsilon}^2\cdot n}{2\cdot k^2\cdot\|\ell\|_\infty^2}\right) \\ &= \exp\left(-\frac{\widetilde{\varepsilon}^2\cdot(m-s^{\mathcal{S}}_m)}{2\cdot k^2\cdot\|\ell\|_\infty^2}\right).\end{aligned}$$

Putting this together with (A.10), we conclude that

$$\mathbb{P}_{\boldsymbol{x}\sim\mu^m}\left[L_{\mu,F,\ell}(\boldsymbol{H}) - L^{\alpha}_{\boldsymbol{x},F^*_n(\boldsymbol{x}),\ell}(\boldsymbol{H}) \geq \varepsilon\right] \leq \exp\left(-\frac{\widetilde{\varepsilon}^2\cdot(m-s^{\mathcal{S}}_m)}{2\cdot k^2\cdot\|\ell\|_\infty^2}\right),$$

as desired.

Let us now estimate the above asymptotically when $s^{\mathcal{S}}_m \leq o(m)$. First note that

$$0 \leq 1 - \frac{(m-s^{\mathcal{S}}_m)_k}{(m)_k} \leq 1 - \left(1-\frac{s^{\mathcal{S}}_m}{m}\right)^k \leq o(1)$$

as $m\to\infty$ with k fixed. Since we also have $m - s^{\mathcal{S}}_m = (1-o(1))\cdot m$, we conclude that

$$\begin{aligned}\exp\left(-\frac{\widetilde{\varepsilon}^2\cdot(m-s^{\mathcal{S}}_m)}{2\cdot k^2\cdot\|\ell\|_\infty^2}\right) &= \exp\left(-\frac{(\varepsilon - \|\ell\|_\infty\cdot o(1))^2\cdot(1-o(1))\cdot m}{2\cdot k^2\cdot\|\ell\|_\infty^2}\right) \\ &\leq \exp\left(-\frac{\varepsilon^2}{2\cdot k^2\cdot\|\ell\|_\infty^2}\cdot m + \frac{\varepsilon\cdot(\varepsilon+\|\ell\|_\infty)}{\|\ell\|_\infty^2}\cdot o(m)\right),\end{aligned}$$

where the $o(m)$ error term is as $m\to\infty$ with k fixed, is uniform in ε, and has the same uniformity in $\|\ell\|_\infty$ as the estimate $s^{\mathcal{S}}_m \leq o(m)$. □

The next theorem shows that every k-ary hypothesis class that admits a k-ary compression scheme $\mathcal{S}$ of essentially non-trivial quality must be k-PAC learnable. More specifically, a trivial k-ary compression scheme has $s^{\mathcal{S}}_m = m$ and $h^{\mathcal{S}}_m = 1$, which yields

$$\ln(h^{\mathcal{S}}_m) + s^{\mathcal{S}}_m\cdot\ln(m) = m\cdot\ln(m)$$

and our assumption is that we have a k-ary compression scheme in which the quantity above is $o(m)$ instead. The main differences of the next theorem to its partite counterpart, Theorem 2.4, stem from the fact that the bounds of Lemma A.4 are different from those in Lemma A.1 and the union bound involves a different number of events.

Theorem A.5. *Let $\Omega = (X,\mathcal{B})$ and $\Lambda = (Y,\mathcal{B}')$ be non-empty standard Borel spaces with $|Y|\geq 2$, let $k\in\mathbb{N}_+$, let $\ell\colon X^k\times Y^{S_k}\times Y^{S_k}\to\mathbb{R}_{\geq 0}$ be a bounded k-ary loss function and let $\mathcal{S}$ be a k-ary compression scheme for a k-ary hypothesis class $\mathcal{H}\subseteq\mathcal{F}_k(\Omega,\Lambda)$ such that*

$$\ln(h^{\mathcal{S}}_m) + s^{\mathcal{S}}_m\cdot\ln(m) \leq o(m) \tag{A.12}$$

as $m\to\infty$ (which in particular follows from $b_{\mathcal{S}}(m)\leq o(m/\ln(m))$ when Y is finite), then $\mathcal{H}$ is k-PAC learnable with respect to ℓ. More specifically, letting $\mathcal{A}$ be given by

$$\mathcal{A}(x,y) \overset{\text{def}}{=} \rho^{\mathcal{S}}_m\big(\kappa^{\mathcal{S}}_m(x,y)\big) \qquad (m\in\mathbb{N}, x\in X^m, y\in Y^{([m])_k})$$

gives a k-PAC learner for $\mathcal{H}$ with respect to ℓ with learning guarantee

$$\begin{aligned}m^{\mathrm{PAC}}_{\mathcal{H},\ell,\mathcal{A}}(\varepsilon,\delta) \overset{\text{def}}{=} \min\Bigg\{ m_0\in\mathbb{N}_+ \;\Bigg|\; &\forall m\geq m_0, \left(1-\frac{(m-s^{\mathcal{S}}_m)_k}{(m)_k}\right)\cdot\|\ell\|_\infty < \varepsilon \\ &\wedge (m)_{s^{\mathcal{S}}_m}\cdot h^{\mathcal{S}}_m\cdot\exp\left(-\frac{(\varepsilon - (1-(m-s^{\mathcal{S}}_m)_k/(m)_k)\cdot\|\ell\|_\infty)^2\cdot(m-s^{\mathcal{S}}_m)}{2\cdot k^2\cdot\|\ell\|_\infty^2}\right)\leq\delta\Bigg\}.\end{aligned} \tag{A.13}$$

In particular, if

$$\ln(h_m^{\mathcal{S}}) + s_m^{\mathcal{S}} \cdot \ln(m) \leq \big(1 + o(1)\big) \cdot \sqrt{2 \cdot m} \cdot \ln(m), \tag{A.14}$$

as $m \to \infty$*, then*

$$m^{\mathrm{PAC}}_{\mathcal{H},\ell,\mathcal{A}}(\varepsilon, \delta) \leq \big(1 + o(1)\big) \cdot \frac{2 \cdot k^2 \cdot \|\ell\|_\infty^2}{\varepsilon^2} \cdot \max\left\{1, \ln\left(\frac{1}{\delta}\right)\right\},$$

where the $o(1)$ *error term is as* $\varepsilon \to 0$ *with* k *fixed, is uniform in* δ*, and has the same uniformity in* $\|\ell\|_\infty$ *as the estimate* $o(1)$ *in* (A.14) *provided* $\|\ell\|_\infty$ *is bounded away from* 0*.*

Proof. Let us start by showing that $b_{\mathcal{S}}(m) \leq o(m/\ln(m))$ (when Y is finite) implies the condition (A.12) (i.e., $\ln(h_m^{\mathcal{S}}) + s_m^{\mathcal{S}} \cdot \ln(m) \leq o(m)$). Indeed, for $m \geq e$, we have

$$\ln(h_m^{\mathcal{S}}) + s_m^{\mathcal{S}} \cdot \ln(m) \leq \ln(m) \cdot \big(\log_2(h_m^{\mathcal{S}}) + s_m^{\mathcal{S}} \cdot \log_2|Y|\big) \leq \ln(m) \cdot b_{\mathcal{S}}(m) \leq o(m).$$

Let us now argue that the minimum in (A.13) indeed exists. For this, we simply need to argue that there exists $m_0 \in \mathbb{N}_+$ satisfying both conditions in the definition of the minimum. But indeed, since we are assuming that (A.12) (i.e., $\ln(h_m^{\mathcal{S}}) + s_m^{\mathcal{S}} \cdot \ln(m) \leq o(m)$) holds, it follows that both

$$\left(1 - \frac{(m - s_m^{\mathcal{S}})_k}{(m)_k}\right) \cdot \|\ell\|_\infty,$$

$$(m)_{s_m^{\mathcal{S}}} \cdot h_m^{\mathcal{S}} \cdot \exp\left(-\frac{(\varepsilon - (1 - (m - s_m^{\mathcal{S}})_k/(m)_k) \cdot \|\ell\|_\infty)^2 \cdot (m - s_m^{\mathcal{S}})}{2 \cdot k^2 \cdot \|\ell\|_\infty^2}\right)$$

tend to 0 as $m \to \infty$ (with all other parameters fixed). So there must exist some $m_0 \in \mathbb{N}$ large enough so that for every $m \geq m_0$, the first expression above is less than ε and the second expression above is at most δ.

We now prove the result. Let $\mu \in \Pr(\Omega)$ be a probability measure and $F \in \mathcal{F}_k(\Omega, \Lambda)$ be a k-ary hypothesis that is realizable with respect to μ and ℓ. Given $\varepsilon, \delta \in (0,1)$, let $m \geq m^{\mathrm{PAC}}_{\mathcal{H},\ell,\mathcal{A}}(\varepsilon, \delta)$ be an integer, let α be an order choice for $[m]$ and let $\widetilde{\varepsilon}$ be defined in terms of m, $s_m^{\mathcal{S}}$, k and ℓ as in Lemma A.4 and note that since $m \geq m^{\mathrm{PAC}}_{\mathcal{H},\ell,\mathcal{A}}(\varepsilon, \delta)$, the first condition in the definition of the latter implies that the condition (A.9) of Lemma A.4 is satisfied.

We now pick $\boldsymbol{x}$ at random according to μ^m and for each $\sigma \in ([m])_{s_m^{\mathcal{S}}}$ and each $\eta \in [h_m^{\mathcal{S}}]$, we let $E_{\sigma,\eta}(\boldsymbol{x})$ be the event

$$L_{\mu,F,\ell}\left(\rho_m^{\mathcal{S}}\left(\sigma^*\big(\boldsymbol{x}, F_m^*(\boldsymbol{x})\big), \eta\right)\right) - L^{\alpha}_{\boldsymbol{x},F_m^*(\boldsymbol{x}),\ell}\left(\rho_m^{\mathcal{S}}\left(\sigma^*\big(\boldsymbol{x}, F_m^*(\boldsymbol{x})\big), \eta\right)\right) \geq \varepsilon$$

so that Lemma A.4 says

$$\mathbb{P}_{\boldsymbol{x}\sim\mu^m}\big[E_{\sigma,\eta}(\boldsymbol{x})\big] \leq \exp\left(-\frac{\widetilde{\varepsilon}^2 \cdot (m - s_m^{\mathcal{S}})}{2 \cdot k^2 \cdot \|\ell\|_\infty^2}\right).$$

If we let $E(\boldsymbol{x})$ be the disjunction of the events $E_{\sigma,\eta}(\boldsymbol{x})$, since there are $(m)_{s_m^{\mathcal{S}}}$ many such σ and $h_m^{\mathcal{S}}$ many such η, by the union bound, we get

$$\mathbb{P}_{\boldsymbol{x}\sim\mu^m}\big[E(\boldsymbol{x})\big] \leq (m)_{s_m^{\mathcal{S}}} \cdot h_m^{\mathcal{S}} \cdot \exp\left(-\frac{\widetilde{\varepsilon}^2 \cdot (m - s_m^{\mathcal{S}})}{2 \cdot k^2 \cdot \|\ell\|_\infty^2}\right).$$

Note also that we have the compression scheme guarantee that

$$L^{\alpha}_{\boldsymbol{x},F_m^*(\boldsymbol{x}),\ell}\left(\rho_m^{\mathcal{S}}\left(\kappa_m^{\mathcal{S}}\big(\boldsymbol{x}, F_m^*(\boldsymbol{x})\big)\right)\right) = 0$$

whenever $(\boldsymbol{x}, F_m^*(\boldsymbol{x}))$ is sample realizable in $\mathcal{H}$ with respect to ℓ and α, which happens with probability 1 as F is realizable in $\mathcal{H}$ with respect to ℓ.

The above along with the fact that the (random) injection $\sigma_m^{\mathcal{S}}(\boldsymbol{x}, F_m^*(\boldsymbol{x}))$ and the (random) header $\eta_m^{\mathcal{S}}(\boldsymbol{x}, F_m^*(\boldsymbol{x}))$ are elements of $([m])_{s_m^{\mathcal{S}}}$ and $[h_m^{\mathcal{S}}]$, respectively, implies that

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{x}\sim\mu^m}\left[L_{\mu,F,\ell}\Big(\mathcal{A}\big(\boldsymbol{x}, F_m^*(\boldsymbol{x})\big)\Big) > \varepsilon\right] \\
&\quad = \mathbb{P}_{\boldsymbol{x}\sim\mu^m}\left[L_{\mu,F,\ell}\left(\rho_m^{\mathcal{S}}\Big(\kappa_m^{\mathcal{S}}\big(\boldsymbol{x}, F_m^*(\boldsymbol{x})\big)\Big)\right) > \varepsilon\right] \\
&\quad \le \mathbb{P}_{\boldsymbol{x}\sim\mu^m}\big[E(\boldsymbol{x})\big] \\
&\quad \le (m)_{s_m^{\mathcal{S}}} \cdot h_m^{\mathcal{S}} \cdot \exp\left(-\frac{\widetilde{\varepsilon}^2 \cdot (m - s_m^{\mathcal{S}})}{2 \cdot k^2 \cdot \|\ell\|_\infty^2}\right) \\
&\quad = (m)_{s_m^{\mathcal{S}}} \cdot h_m^{\mathcal{S}} \cdot \exp\left(-\frac{(\varepsilon - (1 - (m - s_m^{\mathcal{S}})_k/(m)_k) \cdot \|\ell\|_\infty)^2 \cdot (m - s_m^{\mathcal{S}})}{2 \cdot k^2 \cdot \|\ell\|_\infty^2}\right) \\
&\quad \le \delta,
\end{aligned}$$

where the last inequality follows since $m \ge m_{\mathcal{H},\ell,\mathcal{A}}^{\mathrm{PAC}}(\varepsilon, \delta)$ and using the second condition in the definition of the latter. Thus $\mathcal{A}$ is a k-PAC learner for $\mathcal{H}$ with respect to ℓ.

It remains to make the asymptotic estimate of $m_{\mathcal{H},\ell,\mathcal{A}}^{\mathrm{PAC}}(\varepsilon, \delta)$ when the stronger condition (A.14) (i.e., $\ln(h_m^{\mathcal{S}}) + s_m^{\mathcal{S}} \cdot \ln(m) \le (1 + o(1)) \cdot \sqrt{2 \cdot m} \cdot \ln(m)$) holds. (It is clear that the condition (A.14) implies the condition (A.12) that $\ln(h_m^{\mathcal{S}}) + s_m^{\mathcal{S}} \cdot \ln(m) \le o(m)$.)

Let $m_1(\varepsilon) \in \mathbb{N}_+$ be the least integer such that for every $m \ge m_1(\varepsilon)$, we have

$$\left(1 - \frac{(m - s_m^{\mathcal{S}})_k}{(m)_k}\right) \cdot \|\ell\|_\infty < \varepsilon \tag{A.15}$$

and let $m_2(\varepsilon, \delta) \in \mathbb{N}_+$ be the least integer such that for every $m \ge m_2(\varepsilon, \delta)$, we have

$$(m)_{s_m^{\mathcal{S}}} \cdot h_m^{\mathcal{S}} \cdot \exp\left(-\frac{(\varepsilon - (1 - (m - s_m^{\mathcal{S}})_k/(m)_k) \cdot \|\ell\|_\infty)^2 \cdot (m - s_m^{\mathcal{S}})}{2 \cdot k^2 \cdot \|\ell\|_\infty^2}\right) \le \delta. \tag{A.16}$$

Since $m_{\mathcal{H},\ell,\mathcal{A}}^{\mathrm{PAC}}(\varepsilon, \delta) = \max\{m_1(\varepsilon), m_2(\varepsilon, \delta)\}$, it suffices to give asymptotic estimates for $m_1(\varepsilon)$ and $m_2(\varepsilon, \delta)$.

To estimate $m_1(\varepsilon)$, note that if $\|\ell\|_\infty < \varepsilon$, then $m_1(\varepsilon) = k$, so suppose that $\|\ell\|_\infty \ge \varepsilon$ and note that

$$\begin{aligned}
0 \le 1 - \frac{(m - s_m^{\mathcal{S}})_k}{(m)_k} &\le 1 - \left(1 - \frac{s_m^{\mathcal{S}} + k}{m}\right)^k \\
&\le 1 - \left(1 - \big(1 + o(1)\big) \cdot \sqrt{\frac{2}{m}}\right)^k = \big(1 + o(1)\big) \cdot k \cdot \sqrt{\frac{2}{m}}
\end{aligned} \tag{A.17}$$

as $m \to \infty$ with k fixed.

Plugging this asymptotic estimate in the definition (A.15) of $m_1(\varepsilon)$, we conclude that

$$m_1(\varepsilon) \le \big(1 + o(1)\big) \cdot \frac{2 \cdot k^2 \cdot \|\ell\|_\infty^2}{\varepsilon^2}, \tag{A.18}$$

where the $o(1)$ error term is as $\varepsilon \to 0$ with k fixed, is uniform in δ (as $m_1(\varepsilon)$ does not depend on δ at all), and has the same uniformity in $\|\ell\|_\infty$ as the estimate $o(1)$ in (A.14) provided $\|\ell\|_\infty$ is bounded away from 0 (this is to cover the case $\|\ell\|_\infty < \varepsilon$ when $m_1(\varepsilon) = k$).

To estimate $m_2(\varepsilon, \delta)$, we use (A.17) along with the condition (A.12) (i.e., $\ln(h_m^{\mathcal{S}}) + s_m^{\mathcal{S}} \cdot \ln(m) \le o(m)$) to obtain

$$\begin{aligned}
&(m)_{s_m^{\mathcal{S}}} \cdot h_m^{\mathcal{S}} \cdot \exp\left(-\frac{(\varepsilon - (1 - (m - s_m^{\mathcal{S}})_k/(m)_k) \cdot \|\ell\|_\infty)^2 \cdot (m - s_m^{\mathcal{S}})}{2 \cdot k^2 \cdot \|\ell\|_\infty^2}\right) \\
&\quad \le \exp\left(-\frac{(\varepsilon - o(1) \cdot \|\ell\|_\infty)^2 \cdot (1 + o(1)) \cdot m}{2 \cdot k^2 \cdot \|\ell\|_\infty^2} + o(m)\right) \\
&\quad \le \exp\left(-\left(\frac{\varepsilon^2}{2 \cdot k^2 \cdot \|\ell\|_\infty^2} - o(1)\right) \cdot m\right),
\end{aligned}$$

where the error terms are as $m \to \infty$ with k fixed, are uniform in ε, and have the same uniformity in $\|\ell\|_\infty$ as the estimate $o(1)$ in (A.14).

Plugging this asymptotic estimate in the definition (A.16) of $m_2(\varepsilon, \delta)$, we conclude that

$$m_2(\varepsilon, \delta) \leq \left(\frac{2 \cdot k^2 \cdot \|\ell\|^2}{\varepsilon^2} + o(1) \right) \cdot \ln\left(\frac{1}{\delta}\right),$$

where the $o(1)$ error term is as $\varepsilon \to 0$, is uniform in δ, and has the same uniformity in $\|\ell\|_\infty$ as the estimate $o(1)$ in (A.14).

Putting the estimate above with our estimate for $m_1(\varepsilon)$ in (A.18), we get

$$m^{\mathrm{PAC}}_{\mathcal{H},\ell,\mathcal{A}}(\varepsilon, \delta) \leq \big(1 + o(1)\big) \cdot \frac{2 \cdot k^2 \cdot \|\ell\|^2_\infty}{\varepsilon^2} \cdot \max\left\{1, \ln\left(\frac{1}{\delta}\right)\right\}$$

as desired since $m^{\mathrm{PAC}}_{\mathcal{H},\ell,\mathcal{A}} = \max\{m_1(\varepsilon), m_2(\varepsilon, \delta)\}$. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main theorem of the paper, i.e., that the existence of a high-arity sample compression scheme of essentially non-trivial quality implies high-arity PAC learnability, is included in the abstract.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [N/A]

   Justification: The results of this paper are theoretical and as such have no limitations. Nevertheless, we also discuss obstacles for conjectured theoretical results in the last section.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve

the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The theorems are fully stated in the main body of the article and the appendix contains fully formal proofs.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A]

   Justification: This paper is theoretical and as such does not involve any experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     i. If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     ii. If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

iii. If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

iv. We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [N/A]

   Justification: This paper is theoretical and as such does not involve data or code.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: This paper is theoretical and as such does not involve experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [N/A]

Justification: This paper is theoretical and as such does not involve experiments.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: This paper is theoretical and as such does not involve experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research follows the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This paper is theoretical and as such has no immediate direct societal impact.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper is theoretical and as such does not involve data or concrete models.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [N/A]

Justification: This paper is theoretical and as such does not use any code, data or concrete models.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: This paper is theoretical and as such does not produce any assets requiring documentation (such as code or data).

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: This paper is theoretical and as such does not involve any crowdsource nor experiments.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper is theoretical and as such does not involve any experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: No LLMs were involved in any aspect of the research conducted for this paper.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.