A packing lemma for VCN_k-dimension and learning high-dimensional data

Leonardo N. Coregliano Maryanthe Malliaris*

May 26, 2025

Abstract

Recently, the authors introduced the theory of high-arity PAC learning, which is well-suited for learning graphs, hypergraphs and relational structures. In the same initial work, the authors proved a high-arity analogue of the Fundamental Theorem of Statistical Learning that almost completely characterizes all notions of high-arity PAC learning in terms of a combinatorial dimension, called the Vapnik–Chervonenkis–Natarajan (VCN_k) k-dimension, leaving as an open problem only the characterization of non-partite, non-agnostic high-arity PAC learnability.

In this work, we complete this characterization by proving that non-partite non-agnostic high-arity PAC learnability implies a high-arity version of the Haussler packing property, which in turn implies finiteness of VCN_k-dimension. This is done by obtaining direct proofs that classic PAC learnability implies classic Haussler packing property, which in turn implies finite Natarajan dimension and noticing that these direct proofs nicely lift to high-arity.

1 Introduction

Consider the following formulation of the question "are there few convex sets?": Given $k \in \mathbb{N}$ and $\varepsilon > 0$, does there exist $m = m(\varepsilon)$ such for every probability measure μ over $\mathbb{R}$, there exist m convex sets $A_1, \dots, A_m \subseteq \mathbb{R}^k$ such that every convex set $B \subseteq \mathbb{R}^k$ is ε-close to one of the A_i in the natural sense of the product measure μ^k, i.e., $\mu^k(A_i \triangle B) < \varepsilon$ (note that m depends only on ε and not on μ)? More generally, we could ask:

Question 1.1. *Which classes $\mathcal{H}$ of subsets of $\mathbb{R}^k$ admit such a "compression property"?*

The main result of this paper provides a complete answer to the general question in terms of a new notion of combinatorial dimension (in particular, the convex sets do have this "compression property"). To continue to illustrate this using the example of convex sets, the reader familiar with the Haussler packing property might want to first consider the case $k = 1$, in which the class of convex sets amounts to the class of intervals; this class has finite VC dimension and the result above is exactly the Haussler packing property. However, when $k \geq 2$, the class of convex sets no longer has finite VC dimension (even in the plane, notice that n points along a circle are easily shattered by convex sets), so usual Haussler theory provably does not apply. In the present work, we will indeed use a general combinatorial dimension (VCN_k-dimension) which specializes to VC dimension in the $k = 1$ case, and we will obtain a characterization of this "compression property", which we call

*Research partially supported by NSF-BSF 2051825.

high-arity Haussler packing property, in terms of finiteness of this dimension. In order to explain what is new and non-trivial about this result, let us explain some ingredients of the argument.

The proof builds on a recent breakthrough technology of high-arity PAC learning [CM24]. We explain this formally below, but briefly, this theory allows for statistical learning to happen in much more complex settings than classical PAC theory. The key ingredient is understanding how to leverage structured-correlation in high-dimensional data to make new kinds of learning possible.

In fact, the results of the present paper answer a major open question of the high-arity PAC paper, by closing an equivalence of the high-arity PAC theory as we will explain below.

2 Technical preliminaries

In the PAC learning theory of Valiant [Val84] (see also [SSBD14] for a more thorough and modern introduction to the topic), an adversary picks a function $F\colon X \to Y$ out from some family $\mathcal{H} \subseteq Y^X$ and a probability measure μ over X and we are tasked to learn F from i.i.d. samples of the form $(\boldsymbol{x}_i, F(\boldsymbol{x}_i))_{i=1}^m$, where each $\boldsymbol{x}_i$ is drawn from μ; our answer is required to be probably approximately correct (PAC) in the sense of having small total loss with high probability over the sample $\boldsymbol{x}$. The Fundamental Theorem of Statistical Learning characterizes PAC learnability of a family $\mathcal{H}$ in terms of finiteness of:

- its Vapnik–Chervonenkis (VC) dimension [VČ71, BEHW89, VC15] when $|Y| = 2$,
- its Natarajan dimension [Nat89] when Y is finite,
- its Daniely–Shalev-Shwartz (DS) dimension [DSS14, BCD$^+$22] in the general case.

See Theorem A below for the case when Y is finite.

In fact, Haussler [Hau92] showed (when Y is finite, but the result was later extended to the general setting [BCD$^+$22]) that the above is also equivalent to agnostic PAC learnability of $\mathcal{H}$, that is, even if we allow our adversary to pick instead a probability measure ν over $X \times Y$ and provide us i.i.d. samples from ν, then we can at least be competitive in the sense that our answer will have total loss close to the best performing member of $\mathcal{H}$ (with high probability over the sample).

In a different work, Haussler [Hau95, Corollary 1, Theorem 2] also showed that finiteness of the Natarajan dimension of a family $\mathcal{H} \subseteq Y^X$ (with Y finite) is equivalent to the following packing property[1]: for every $\varepsilon > 0$, there exists $m = m(\varepsilon)$ such that for every probability measure μ over X, there exist $\mathcal{H}' \subseteq \mathcal{H}$ with $|\mathcal{H}'| \leq m$ such that every $F \in \mathcal{H}$ is ε-close to some $H \in \mathcal{H}'$ in the sense that

$$\mu(\{x \in X \mid F(x) \neq H(x)\}) \leq \varepsilon.$$

Collectively, these works yield the Fundamental Theorem of Statistical Learning:

Theorem A ([VČ71, BEHW89, Nat89, Hau92, Hau95, VC15])**.** *The following are equivalent for a family $\mathcal{H} \subseteq Y^X$ of functions $X \to Y$ with Y finite:*

i. The Natarajan dimension of $\mathcal{H}$ is finite.

ii. $\mathcal{H}$ has the uniform convergence property.

[1]This is a slight generalization and reformulation of Haussler's original work. Instead, Haussler's original work frames this in terms of ε-separated sets and considers only the case $Y = \{0, 1\}$ (which yields a characterization in terms of the VC-dimension).

iii. $\mathcal{H}$ *is agnostically PAC learnable.*

iv. $\mathcal{H}$ *is PAC learnable.*

v. $\mathcal{H}$ *has the Haussler packing property.*

One of the main criticisms of classic PAC learning theory is that it strongly relies on the independence of its samples. While there has been considerable work [HL94, AV95, Gam03, ZZX09, SW10, ZXC12, ZLX12, ZXX14, BGS18, SvS23] in extending PAC theory to allow for some correlation, these works see correlation as an obstacle to learning. More recently, the authors introduced the theory of high-arity PAC learning [CM24], in which correlation is leveraged to increase the learning power.

High-arity PAC learning theory is heavily inspired by the problem of learning a graph over a set X (but more generally, also covers hypergraphs and even relational structures). By encoding such a graph by its adjacency matrix $F\colon X \times X \to \{0,1\}$, one could simply apply classic PAC learning theory; however, this yields a rather unnatural learning framework: the adversary is picking a measure μ over $X \times X$ and revealing several pairs $(\boldsymbol{x}_i, \boldsymbol{x}'_i)$ drawn i.i.d. from μ along with their adjacency $F(\boldsymbol{x}_i, \boldsymbol{x}'_i)$.

Instead, the setup of 2-PAC learning is much more natural: the adversary picks a measure μ over X, draws m vertices $(\boldsymbol{x}_i)_{i=1}^m$ i.i.d. from μ and reveals all the adjacency information between these vertices: $(F(\boldsymbol{x}_i, \boldsymbol{x}_j))_{i,j=1}^m$ (see Section 3 for a formal, but simplified definition and see Appendix A for the full definitions).

In high-arity, there is also a natural partite framework. Namely, if we were trying to learn a bipartite graph $F\colon X_1 \times X_2 \to \{0,1\}$ with a known bipartition (X_1, X_2), we could instead allow the adversary to pick different measures μ_1 and μ_2 over X_1 and X_2, respectively, draw $(\boldsymbol{x}_i^1)_{i=1}^m$ i.i.d. from μ_1 and $(\boldsymbol{x}_i^2)_{i=1}^m$ i.i.d. from μ_2, independently from the previous points and provide us all adjacency information: $(F(\boldsymbol{x}_i^1, \boldsymbol{x}_j^2))_{i,j=1}^m$.

Every (not necessarily bipartite) graph $F\colon X \times X \to \{0,1\}$ can be interpreted as a bipartite graph $F^{2\text{-part}}\colon X_1 \times X_2 \to \{0,1\}$ in which $X_1 = X_2 = X$ (combinatorially, this is doubling the vertices of F), but a priori 2-PAC learnability of a family $\mathcal{H} \subseteq \{0,1\}^{X \times X}$ is not necessarily the same as partite 2-PAC learnability of its partization $\mathcal{H}^{2\text{-part}} \stackrel{\text{def}}{=} \{F^{2\text{-part}} \mid F \in \mathcal{H}\}$.

The interplay between a high-arity hypothesis class $\mathcal{H}$ and its partization $\mathcal{H}^{k\text{-part}}$ plays a crucial role in proving Theorem B below, which is high-arity analogue of Theorem A (we direct the reader again to Section 3 and Appendix A or to the original paper [CM24] for the precise definitions of the concepts in this theorem). To illustrate the non-triviality of the interplay between partite and non-partite, consider the partite version of Question 1.1 in Section 1: if we change the setup to allow for product probability measures of the form $\mu_1 \otimes \cdots \otimes \mu_k$, of not necessarily the same measure, why should we expect the resulting property to be equivalent to the one only for measures of the form μ^k?

Theorem B ([CM24, Theorems 1.1 and 5.1])**.** *Let* $k \in \mathbb{N}_+$*. The following are equivalent for a family* $\mathcal{H} \subseteq Y^{X^k}$ *of functions* $X^k \to Y$ *with* Y *finite and its partization* $\mathcal{H}^{k\text{-part}}$*:*

i. The Vapnik–Chervonenkis–Natarajan k*-dimension of* $\mathcal{H}$ *is finite.*

ii. The partite Vapnik–Chervonenkis–Natarajan k*-dimension of* $\mathcal{H}^{k\text{-part}}$ *is finite.*

iii. $\mathcal{H}^{k\text{-part}}$ *has the uniform convergence property.*

iv. $\mathcal{H}$ *is agnostically* k*-PAC learnable.*

v. $\mathcal{H}^{k\text{-part}}$ *is partite agnostically* k*-PAC learnable.*

vi. $\mathcal{H}^{k\text{-part}}$ *is partite* k*-PAC learnable.*

Furthermore, any of the items above implies the following:

vii. $\mathcal{H}$ *is* k*-PAC learnable.*

(The statement above is the informal version [CM24, Theorem 1.1] that covers only the 0/1-loss function; the formal version [CM24, Theorem 5.1] covers general loss functions (under mild assumptions).)

It turns out that the implication of PAC learnability to finite Natarajan dimension (usually called No Free Lunch Theorem) of classic PAC theory only lifts naturally to the partite setting. The authors were able to partially solve this issue by providing a "departization" operation that allows non-partite agnostic k-PAC learnability to imply its partite analogue. However, this left open the question of whether non-partite k-PAC learnability (item (vii) in Theorem B) can also be included in the equivalence list.

In this paper, we prove that this is indeed the case by developing a high-arity version of the Haussler packing property (which we refer to here as k-ary Haussler packing). Recall, the classic equivalence of the Haussler packing property with finiteness of VC-dimension can be proved by direct implications between the properties[2]. In the present work, we find high-arity proofs of the implications k-PAC learnability $\implies$ k-ary Haussler packing property and k-ary Haussler packing property $\implies$ finite VCN_k-dimension (Theorem C below). The reader should note that when specialized to $k = 1$, these yield the expected direct implications from classic PAC to Haussler packing and from Haussler packing to finite Natarajan dimension.

Theorem C (informal version of Theorems 5.1 and 5.3 of the present paper)**.** *Let* $k \in \mathbb{N}_+$*. Then the following hold for a family* $\mathcal{H} \subseteq Y^{X^k}$ *of functions* $X^k \to Y$ *with* Y *finite:*

i. If $\mathcal{H}$ *is* k*-PAC learnable, then* $\mathcal{H}$ *has the* k*-ary Haussler packing property.*

ii. If $\mathcal{H}$ *has the* k*-ary Haussler packing property, then* $\mathcal{H}$ *has finite* VCN_k*-dimension.*

Note, at the risk of stating the obvious, that two things have happened here: there is a new high-arity notion defined, that of k-ary Haussler packing, and two new direct implications that close the loop of equivalences when put together with the earlier Theorem B. Thus an immediate consequence of the main result in this paper is the following summary theorem (see Figure 1 for a pictorial view of the implications in this work and of [CM24]):

Theorem D. *Let* $k \in \mathbb{N}_+$*. The following are equivalent for a family* $\mathcal{H} \subseteq Y^{X^k}$ *of functions* $X^k \to Y$ *with* Y *finite and its partization* $\mathcal{H}^{k\text{-part}}$*:*

i. The Vapnik–Chervonenkis–Natarajan k*-dimension of* $\mathcal{H}$ *is finite.*

ii. The partite Vapnik–Chervonenkis–Natarajan k*-dimension of* $\mathcal{H}^{k\text{-part}}$ *is finite.*

[2]In fact, this essentially goes back to Haussler [Hau95]; however, he does not prove that the packing property implies finite VC-dimension, he instead shows tightness of an analogous bound when the domain X is finite; but one can derive the implication from his results.

iii. $\mathcal{H}^{k\text{-part}}$ *has the uniform convergence property.*

iv. $\mathcal{H}$ *is agnostically* k*-PAC learnable.*

v. $\mathcal{H}^{k\text{-part}}$ *is partite agnostically* k*-PAC learnable.*

vi. $\mathcal{H}$ *is* k*-PAC learnable.*

vii. $\mathcal{H}^{k\text{-part}}$ *is partite* k*-PAC learnable.*

viii. $\mathcal{H}$ *has the* k*-ary Haussler packing property.*

ix. $\mathcal{H}^{k\text{-part}}$ *has the* k*-ary Haussler packing property.*

In particular, the equivalence of items (i) and (viii) above completely answer Question 1.1.

The paper is organized as follows. First, since our main goal is to complete the characterization of (non-partite) k-PAC learnability, we opt to provide in the main text only a simplified version of the high-arity PAC definitions in [CM24] and of our argument that only covers "rank at most 1" hypotheses in the non-partite setting, ignoring the subtlety of "high-order variables" (as these are mostly used in agnostic high-arity PAC, which is not needed for our results); we provide the full version of the high-arity definitions and of our argument in the appendices. Sections 3, 4 and 5 contain the simplified versions of the definitions from [CM24], of our new definitions and of our main results, respectively. Appendix A contains the full version of the same content and also covers a partite version of the Haussler packing property. Appendix B relates the non-partite and partite Haussler packing properties directly through the partization operation.

3 Simplified high-arity PAC definitions

In this section, we lay out the definitions of high-arity PAC theory of [CM24] that we will need in a simplified manner that covers only rank at most 1 hypotheses. We direct the curious reader to Appendix A for the general version of the same definitions.

Definition 3.1 ([CM24, §3] simplified)**.** By a Borel space, we mean a standard Borel space, i.e., a measurable space that is Borel-isomorphic to a Polish space when equipped with the σ-algebra of Borel sets. The space of probability measures on a Borel space Λ is denoted $\Pr(\Lambda)$.

Let $\Omega = (X, \mathcal{B})$ and $\Lambda = (Y, \mathcal{B}')$ be non-empty Borel spaces and $k \in \mathbb{N}_+$.

1. The set of *k-ary hypotheses* from Ω to Λ, denoted $\mathcal{F}_k(\Omega, \Lambda)$, is the set of (Borel) measurable functions from Ω^k to Λ.

2. A *k-ary hypothesis class* is a subset $\mathcal{H}$ of $\mathcal{F}_k(\Omega, \Lambda)$ equipped with a σ-algebra such that

   i. the evaluation map $\mathrm{ev}\colon \mathcal{H} \times \Omega^k \to \Lambda$ given by $\mathrm{ev}(H, x) \stackrel{\mathrm{def}}{=} H(x)$ is measurable;

   ii. for every $H \in \mathcal{H}$, the set $\{H\}$ is measurable;

   iii. 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] (i.e., measurable in every completion of a probability measure on Υ).

3. A *k-ary loss function* over Λ is a measurable function $\ell\colon \Omega^k \times \Lambda^{S_k} \times \Lambda^{S_k} \to \mathbb{R}_{\geq 0}$, where S_k is the symmetric group on $[k] \stackrel{\text{def}}{=} \{1, \dots, k\}$. We further define

$$\|\ell\|_\infty \stackrel{\text{def}}{=} \sup_{\substack{x \in \Omega^k \\ y, y' \in \Lambda^{S_k}}} \ell(x, y, y'), \qquad\qquad s(\ell) \stackrel{\text{def}}{=} \inf_{\substack{x \in \Omega^k \\ y, y' \in \Lambda^{S_k} \\ y \neq y'}} \ell(x, y, y'),$$

and we say that ℓ is:

bounded if $\|\ell\|_\infty < \infty$.
separated if $s(\ell) > 0$ and $\ell(x, y, y) = 0$ for every $x \in \Omega^k$ and every $y \in \Lambda^{S_k}$.

If we are further given k-ary hypotheses $F, H \in \mathcal{F}_k(\Omega, \Lambda)$ and a probability measure $\mu \in \Pr(\Omega)$, then we define the *total loss* of H with respect to μ, F and ℓ by

$$L_{\mu,F,\ell}(H) \stackrel{\text{def}}{=} \mathbb{E}_{\boldsymbol{x} \sim \mu^k}\left[\ell\Big(\boldsymbol{x}, \big(H(\boldsymbol{x}_{\sigma(1)}, \dots, \boldsymbol{x}_{\sigma(k)})\big)_{\sigma \in S_k}, \big(F(\boldsymbol{x}_{\sigma(1)}, \dots, \boldsymbol{x}_{\sigma(k)})\big)_{\sigma \in S_k}\Big)\right].$$

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

5. The *k-ary $0/1$-loss function* over Λ is defined as $\ell_{0/1}(x, y, y') \stackrel{\text{def}}{=} \mathbb{1}[y \neq y']$.

6. A *(k-ary) learning algorithm* for a k-ary hypothesis class $\mathcal{H}$ is a measurable function

$$\mathcal{A}\colon \bigcup_{m \in \mathbb{N}} (\Omega^m \times \Lambda^{([m])_k}) \to \mathcal{H},$$

where $([m])_k$ denotes the set of injections $[k] \to [m]$.

7. We say that a k-ary hypothesis class is *k-PAC learnable* with respect to a k-ary loss function ℓ if there exists a learning algorithm $\mathcal{A}$ for $\mathcal{H}$ 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 $\mu \in \Pr(\Omega)$ and every $F \in \mathcal{F}_k(\Omega, \Lambda)$ that is realizable in $\mathcal{H}$ with respect to ℓ and μ, we have

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

for every integer $m \geq m^{\text{PAC}}_{\mathcal{H},\ell,\mathcal{A}}(\varepsilon, \delta)$. A learning algorithm $\mathcal{A}$ satisfying the above is called a *k-PAC learner* for $\mathcal{H}$ with respect to ℓ.

[3]This assumption about hypothesis classes is not made in [CM24], but it is clearly needed for uniform convergence to make sense; in this document, we will not need this. Also, note that if $\mathcal{H}$ is equipped with a σ-algebra that turns it into a standard Borel space, then this hypothesis is immediately satisfied as Suslin sets are universally measurable.

8. For a k-ary hypothesis $H \in \mathcal{F}_k(\Omega, \Lambda)$ and $x \in \Omega^{k-1}$, we let $H_x \colon \Omega \to \Lambda^{S_k}$ be defined by

$$H_x(x_k)_\sigma \stackrel{\text{def}}{=} H(x_{\sigma(1)}, \dots, x_{\sigma(k)}) \qquad (x_k \in \Omega, \sigma \in S_k).$$

For a k-ary hypothesis class $\mathcal{H}$, the *Vapnik–Chervonenkis–Natarajan k-dimension* of $\mathcal{H}$ (VCN_k-*dimension*) is defined as

$$\mathrm{VCN}_k(\mathcal{H}) \stackrel{\text{def}}{=} \sup_{x \in \Omega^{k-1}} \mathrm{Nat}(\mathcal{H}(x)),$$

where

$$\mathcal{H}(x) \stackrel{\text{def}}{=} \{H_x \mid H \in \mathcal{H}\}$$

and Nat is the Natarajan-dimension (see Definition 3.2 below).

Definition 3.2 (Natarajan dimension [Nat89])**.** Let $\mathcal{F}$ be a collection of functions of the form $X \to Y$ and let $A \subseteq X$.

1. We say that $\mathcal{F}$ *(Natarajan-)shatters* A if there exist functions $f_0, f_1 \colon A \to Y$ such that
   i. for every $a \in A$, we have $f_0(a) \neq f_1(a)$,
   ii. for every $U \subseteq A$, there exists $F_U \in \mathcal{F}$ such that for every $a \in A$ we have

$$F_U(a) = f_{\mathbb{1}[a \in U]}(a) = \begin{cases} f_0(a), & \text{if } a \notin U, \\ f_1(a), & \text{if } a \in U. \end{cases}$$

2. The *Natarajan dimension* of $\mathcal{F}$ is defined as

$$\mathrm{Nat}(\mathcal{F}) \stackrel{\text{def}}{=} \sup\{|A| \mid A \subseteq X \wedge \mathcal{F} \text{ Natarajan-shatters } A\}.$$

4 Simplified versions of new high-arity concepts

In this section, we formalize the high-arity version of the Haussler packing property in the same simplified manner as in Section 3 and the notion of a metric loss (also in simplified manner); again, we direct the curious reader to Appendix A.3 for the general version of the same definitions.

Definition 4.1 (k-ary Haussler packing property)**.** Let $k \in \mathbb{N}_+$, let Ω and Λ be non-empty Borel spaces, let $\mathcal{H} \subseteq \mathcal{F}_k(\Omega, \Lambda)$ be a k-ary hypothesis class and let $\ell \colon \Omega^k \times \Lambda^{S_k} \times \Lambda^{S_k} \to \mathbb{R}_{\geq 0}$ be a k-ary loss function.

We say that $\mathcal{H}$ has the *k-ary Haussler packing property*[4] with respect to ℓ if there exists a function $m^{\mathrm{HP}}_{\mathcal{H},\ell} \colon (0,1) \to \mathbb{R}_{\geq 0}$ such that for every $\varepsilon \in (0,1)$ and every $\mu \in \Pr(\Omega)$, there exists $\mathcal{H}' \subseteq \mathcal{H}$ with $|\mathcal{H}'| \leq m^{\mathrm{HP}}_{\mathcal{H},\ell}(\varepsilon)$ such that for every $F \in \mathcal{H}$, there exists $H \in \mathcal{H}'$ such that $L_{\mu,F,\ell}(H) \leq \varepsilon$. We refer to elements of $\mathcal{H}'$ as *k-ary Haussler centers* of $\mathcal{H}$ at precision ε with respect to μ and ℓ.

We point out that the usual Haussler packing property does a priori make sense in the high-arity setting but would be relative to a probability measure ν on Ω^k and hence would be applicable to very special classes. By contrast, k-ary Haussler considers product measures μ^k for a probability measure μ over Ω. Recall that our goal is to fully characterize which hypothesis classes admit such

[4]It would be perhaps more fitting to call this the Haussler covering property, but we opt retain the historical name of its unary version.

a property. We will see in the course of the proofs the equivalence of k-ary Haussler packing to k-ary PAC learning requires interesting mathematical tools. For simplicity, in the remainder of this paper, for readability we may drop "k-ary" from the terminology.

Definition 4.2 (Metric loss functions)**.** Let $k \in \mathbb{N}_+$ and let Ω and Λ be non-empty Borel spaces.

We say that a k-ary loss function $\ell\colon \Omega^k \times \Lambda^{S_k} \times \Lambda^{S_k} \to \mathbb{R}_{\geq 0}$ is *metric* if for every $x \in \Omega^k$, the function $\ell(x, -, -)$ is a metric on Λ^{S_k} in the usual sense, that is, the following hold for every $x \in \Omega^k$ and every $y, y', y'' \in \Lambda^{S_k}$:

i. We have $\ell(x, y, y') = \ell(x, y', y)$.

ii. We have $\ell(x, y, y') = 0$ if and only if $y = y'$.

iii. We have $\ell(x, y, y'') \leq \ell(x, y, y') + \ell(x, y', y'')$.

5 Main results

In this section we prove that k-PAC learnability implies the Haussler packing property (Theorem 5.1), which in turn implies finite VCN_k-dimension (Theorem 5.3). The former theorem is done under the assumption that the loss function ℓ is metric, but we point out that one could have assumed instead that ℓ is separated and bounded with a slight change in the argument (this is featured in the non-simplified version of the implication, Theorem A.14). Finally, let us point out that the 0/1-loss function $\ell_{0/1}$ satisfies all hypotheses of Theorems 5.1 and 5.3 and [CM24, Theorems 1.1 and 5.1].

Theorem 5.1 (k-PAC learnability implies Haussler packing property, simplified)**.** *Let $k \in \mathbb{N}_+$, let Ω and Λ be non-empty Borel spaces with Λ finite, let $\mathcal{H} \subseteq \mathcal{F}_k(\Omega, \Lambda)$ be a k-ary hypothesis class and let $\ell\colon \Omega^k \times \Lambda^{S_k} \times \Lambda^{S_k} \to \mathbb{R}_{\geq 0}$ be a k-ary loss function.*

Let also

$$\gamma_{\mathcal{H}}(m) \stackrel{\text{def}}{=} \sup_{x \in \Omega^m} \left| \left\{ \left(H(x_{\alpha(1)}, \dots, x_{\alpha(k)})\right)_{\alpha \in ([m])_k} \;\middle|\; H \in \mathcal{H} \right\} \right|$$

be the maximum number of different patterns (in $\Lambda^{([m])_k}$) that can be obtained by considering a fixed $x \in \Omega^m$ and plugging in each injective k-tuple as an input of some $H \in \mathcal{H}$.

If ℓ is metric and $\mathcal{H}$ is k-PAC learnable with a k-PAC learner $\mathcal{A}$, then $\mathcal{H}$ has the Haussler packing property with associated function

$$m^{\mathrm{HP}}_{\mathcal{H},\ell}(\varepsilon) \stackrel{\text{def}}{=} \min_{\delta \in (0,1)} \left\lceil \frac{\gamma_{\mathcal{H}}(\lceil m^{\mathrm{PAC}}_{\mathcal{H},\ell,\mathcal{A}}(\varepsilon/2, \delta) \rceil)}{1 - \delta} \right\rceil - 2 \leq \min_{\delta \in (0,1)} \left\lceil \frac{|\Lambda|^{(\lceil m^{\mathrm{PAC}}_{\mathcal{H},\ell,\mathcal{A}}(\varepsilon/2,\delta) \rceil)_k}}{1 - \delta} \right\rceil - 2, \tag{5.1}$$

where $(m)_k \stackrel{\text{def}}{=} m(m-1)\cdots(m-k+1)$ denotes the falling factorial.

Proof. First note that the inequality in (5.1) follows from the trivial bound

$$\gamma_{\mathcal{H}}(m) \leq |\Lambda|^{(m)_k}.$$

Second, note that due to the ceilings on the expressions in (5.1), the minima are indeed attained as the functions only take values in $\mathbb{N}$.

Suppose for a contradiction that the result does not hold, that is, there exists $\varepsilon \in (0,1)$ and $\mu \in \Pr(\Omega)$ such that if m is given by the first minimum in (5.1), then for every $\mathcal{H}' \subseteq \mathcal{H}$ with $|\mathcal{H}'| \leq m$, there exists $F \in \mathcal{H}$ such that $L_{\mu,F,\ell}(H) > \varepsilon$ for every $H \in \mathcal{H}'$. By repeatedly applying this property, it follows that there exist $F_1, \ldots, F_{m+1} \in \mathcal{H}$ such that for every $i, j \in [m+1]$ distinct, we have $L_{\mu,F_i,\ell}(F_j) > \varepsilon$ (recall that ℓ is metric, so $L_{\mu,F_i,\ell}(F_j) = L_{\mu,F_j,\ell}(F_i)$).

Let $\delta \in (0,1)$ attain the first minimum in (5.1) and let

$$\widetilde{m} \stackrel{\text{def}}{=} \left\lceil m^{\text{PAC}}_{\mathcal{H},\ell,\mathcal{A}}\left(\frac{\varepsilon}{2}, \delta\right)\right\rceil.$$

For each $x \in \Omega^{\widetilde{m}}$, let

$$Y(x) \stackrel{\text{def}}{=} \left\{\left(H(x_{\alpha(1)}, \ldots, x_{\alpha(k)})\right)_{\alpha \in ([\widetilde{m}])_k} \;\middle|\; H \in \mathcal{H}\right\} \subseteq \Lambda^{([\widetilde{m}])_k}$$

and note that $|Y(x)| \leq \gamma_{\mathcal{H}}(\widetilde{m})$.

For each $i \in [m+1]$, define the set

$$C_i \stackrel{\text{def}}{=} \left\{x \in \Omega^{\widetilde{m}} \;\middle|\; \forall y \in Y(x), L_{\mu,F_i,\ell}(\mathcal{A}(x,y)) > \frac{\varepsilon}{2}\right\}.$$

Note that by taking

$$y \stackrel{\text{def}}{=} \left(F_i(x_{\alpha(1)}, \ldots, x_{\alpha(k)})\right)_{\alpha \in ([m])_k} \in Y(x)$$

and using the fact that $L_{\mu,F_i,\ell}(F_i) = 0$ (as ℓ is metric) so that F_i is realizable in $\mathcal{H}$ w.r.t. ℓ and μ, PAC learnability implies that $\mu(C_i) \leq \delta$.

Define now the function $G\colon \Omega^{\widetilde{m}} \to \mathbb{R}_{\geq 0}$ by

$$\begin{aligned} G(x) &\stackrel{\text{def}}{=} \sum_{i=1}^{m+1} \mathbb{1}_{C_i}(x) \\ &= \left|\left\{i \in [m+1] \;\middle|\; \forall y \in Y(x), L_{\mu,F_i,\ell}(\mathcal{A}(x,y)) > \frac{\varepsilon}{2}\right\}\right|. \end{aligned}$$

We claim that for every $x \in \Omega^{\widetilde{m}}$ and every $y \in Y(x)$, there exists at most one $i \in [m+1]$ such that $L_{\mu,F_i,\ell}(\mathcal{A}(x,y)) \leq \varepsilon/2$. Indeed, if not, then for some $i, j \in [m+1]$ distinct, we would get

$$L_{\mu,F_i,\ell}(F_j) \leq L_{\mu,F_i,\ell}(\mathcal{A}(x,y)) + L_{\mu,F_j,\ell}(\mathcal{A}(x,y)) \leq \varepsilon,$$

where the first inequality follows since ℓ is metric; the above would then contradict $L_{\mu,F_i,\ell}(F_j) > \varepsilon$.

Thus, we conclude that for every $x \in \Omega^{\widetilde{m}}$, we have

$$G(x) \geq m + 1 - |Y(x)| \geq m + 1 - \gamma_{\mathcal{H}}(\widetilde{m}). \tag{5.2}$$

On the other hand, since $\mu(C_i) \leq \delta$ for every $i \in [m+1]$, we get

$$\int_{\Omega^{\widetilde{m}}} G(x)\; d\mu^{\widetilde{m}}(x) \leq (m+1)\delta,$$

which together with (5.2) implies

$$m \leq \frac{\gamma_{\mathcal{H}}(\widetilde{m})}{1-\delta} - 1,$$

contradicting the definitions of m, δ and $\widetilde{m}$. □

For the next theorem, we will need the following standard combinatorial result about covers of the power set of $[n]$.

Lemma 5.2. *Let $c \in (0, 1/2)$, let $n \in \mathbb{N}$ and let $\mathcal{C} \subseteq 2^{[n]}$ be a collection of subsets of $[n]$. Suppose that for every $U \subseteq [n]$, there exists $V \in \mathcal{C}$ such that $|U \triangle V| \leq cn$. Then*

$$n \leq \frac{\log_2|\mathcal{C}|}{1 - h_2(c)},$$

where

$$h_2(t) \stackrel{\text{def}}{=} t \log_2 \frac{1}{t} + (1-t) \log_2 \frac{1}{1-t}$$

denotes the binary entropy.

Proof. For each $V \in \mathcal{C}$, let

$$B(V) \stackrel{\text{def}}{=} \{U \subseteq [n] \mid |U \triangle V| \leq cn\}$$

and note that

$$|B(V)| = \sum_{i=0}^{\lfloor cn \rfloor} \binom{n}{i} \leq 2^{h_2(c) \cdot n},$$

where the last inequality is the standard upper bound on the volume of a Hamming ball in terms of binary entropy (see e.g. [Ash65, Lemma 4.7.2]).

Since $\bigcup_{V \in \mathcal{C}} B(V) = 2^{[n]}$, we conclude that

$$|\mathcal{C}| \cdot 2^{h_2(c) \cdot n} \geq 2^n,$$

from which the result follows. □

Theorem 5.3 (Haussler packing property implies finite VCN_k-dimension, simplified)**.** *Let $k \in \mathbb{N}_+$, let Ω and Λ be non-empty Borel spaces with Λ finite, let $\mathcal{H} \subseteq \mathcal{F}_k(\Omega, \Lambda)$ be a k-ary hypothesis class and let $\ell \colon \Omega^k \times \Lambda^{S_k} \times \Lambda^{S_k} \to \mathbb{R}_{\geq 0}$ be a k-ary loss function.*

If ℓ is separated and $\mathcal{H}$ has the Haussler packing property, then

$$\mathrm{VCN}_k(\mathcal{H}) \leq \min_{\varepsilon \in (0, \min\{s(\ell) \cdot k!/(2k^k), 1\})} \left\lfloor \frac{\log_2 \lfloor m^{\mathrm{HP}}_{\mathcal{H},\ell}(\varepsilon) \rfloor}{1 - h_2(\varepsilon \cdot k^k / (s(\ell) \cdot k!))} \right\rfloor, \tag{5.3}$$

where

$$h_2(t) \stackrel{\text{def}}{=} t \log_2 \frac{1}{t} + (1-t) \log_2 \frac{1}{1-t}$$

denotes the binary entropy.

Proof. First, note that the minimum in (5.3) is indeed attained as the function only takes values in $\mathbb{N} \cup \{-\infty\}$, so let $\varepsilon \in (0, \min\{s(\ell) \cdot k!/(2k^k), 1\})$ attain the minimum, let d be the value of the minimum and let

$$m \stackrel{\text{def}}{=} \lfloor m^{\mathrm{HP}}_{\mathcal{H},\ell}(\varepsilon) \rfloor$$

so that

$$d = \left\lfloor \frac{\log_2 m}{1 - h_2(\varepsilon \cdot k^k / (s(\ell) \cdot k!))} \right\rfloor.$$

When $\mathcal{H}$ is empty, the result is trivial as $\mathrm{VCN}_k(\mathcal{H}) = -\infty$, so suppose $\mathcal{H}$ is non-empty (hence $m \geq 1$ and $d \geq 0$).

By the definition of VCN_k-dimension, we have to show that if $x \in \Omega^{k-1}$, then $\mathrm{Nat}(\mathcal{H}(x)) \leq d$. In turn, it suffices to show that if $V \subseteq \Omega$ is a (finite) set that is Natarajan-shattered by $\mathcal{H}(x)$ and $n \stackrel{\text{def}}{=} |V|$, then $n \leq d$.

Let $\mu \in \Pr(\Omega)$ be given by

$$\frac{1}{k}\left(\nu_V + \sum_{j=1}^{k-1} \delta_{x_j}\right),$$

where ν_V is the uniform probability measure on V and δ_t is the Dirac delta concentrated on t.

Since V is Natarajan-shattered by $\mathcal{H}(x)$, there exist functions $f_0, f_1 \colon V \to \Lambda^{S_k}$ with $f_0(v) \neq f_1(v)$ for every $v \in V$ and there exists a family $\{F^U \mid U \subseteq V\} \subseteq \mathcal{H}$ such that for every $U \subseteq V$ and every $v \in V$, we have $F^U_x(v) = f_{\mathbb{1}[v \in U]}(v)$.

For $F, F' \in \mathcal{H}$, let us define the set

$$D(F, F') \stackrel{\text{def}}{=} \{v \in V \mid F_x(v) \neq F'_x(v)\}.$$

Clearly, for every $U, U' \subseteq V$, we have $D(F_U, F_{U'}) = U \triangle U'$.

Note also that by the definition of μ, for $F, F' \in \mathcal{H}$, we have

$$\begin{aligned} L_{\mu,F,\ell}(F') &\geq \mathbb{P}_{\boldsymbol{z}\sim\mu^k}\left[\exists \sigma \in S_k, \forall j \in [k-1], \boldsymbol{z}_{\sigma(j)} = x_j \wedge \boldsymbol{z}_{\sigma(k)} \in D(F, F')\right] \cdot s(\ell) \\ &\geq \frac{s(\ell) \cdot k!}{k^k \cdot n} \cdot |D(F, F')|. \end{aligned} \tag{5.4}$$

Since m is defined via Haussler packing property, we know that there exists $\mathcal{H}' \subseteq \mathcal{H}$ such that $|\mathcal{H}'| \leq m$ and for every $U \subseteq V$, there exists $H \in \mathcal{H}'$ such that $L_{\mu,F_U,\ell}(H) \leq \varepsilon$.

For each $H \in \mathcal{H}'$, let

$$\begin{aligned} U_H &\stackrel{\text{def}}{=} \{v \in V \mid H_x(v) = f_1(v)\}, \\ B(H) &\stackrel{\text{def}}{=} \{U \subseteq V \mid L_{\mu,F_U,\ell}(H) \leq \varepsilon\}, \\ B'(H) &\stackrel{\text{def}}{=} \left\{U \subseteq V \;\middle|\; |U \triangle U_H| \leq \frac{\varepsilon \cdot k^k \cdot n}{s(\ell) \cdot k!}\right\}. \end{aligned}$$

The Haussler packing property assumption implies

$$\bigcup_{H \in \mathcal{H}'} B(H) = 2^V. \tag{5.5}$$

On the other hand, by (5.4), for every $U \subseteq V$ and every $H \in \mathcal{H}'$, we have

$$L_{\mu,F_U,\ell}(H) \geq \frac{s(\ell) \cdot k!}{k^k \cdot n} \cdot |D(F_U, H)| \geq \frac{s(\ell) \cdot k!}{k^k \cdot n} \cdot |D(F_U, F_{U_H})| = \frac{s(\ell) \cdot k!}{k^k \cdot n} \cdot |U \triangle U_H|,$$

which along with (5.5) implies $\bigcup_{H \in \mathcal{H}'} B'(H) = 2^V$.

Since $\varepsilon \cdot k^k/(s(\ell) \cdot k!) < 1/2$, by Lemma 5.2, we get

$$n \leq \frac{\log_2 |\mathcal{H}'|}{1 - h_2(\varepsilon \cdot k^k \cdot n/(s(\ell) \cdot k!))} \leq \frac{\log_2 m}{1 - h_2(\varepsilon \cdot k^k \cdot n/(s(\ell) \cdot k!))},$$

which yields $n \leq d$ as n is an integer and d is the floor of the right-hand side of the above. □

References

[Ash65] Robert Ash. *Information theory*, volume No. 19 of *Interscience Tracts in Pure and Applied Mathematics*. Interscience Publishers John Wiley & Sons, New York-London-Sydney, 1965.

[AV95] David Aldous and Umesh Vazirani. A Markovian extension of Valiant's learning model. *Inform. and Comput.*, 117(2):181–186, 1995.

[BCD+22] Nataly Brukhim, Daniel Carmon, Irit Dinur, Shay Moran, and Amir Yehudayoff. A characterization of multiclass learnability. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science—FOCS 2022*, pages 943–955. IEEE Computer Soc., Los Alamitos, CA, [2022] ©2022.

[BEHW89] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *J. Assoc. Comput. Mach.*, 36(4):929–965, 1989.

[BGS18] Guy Bresler, David Gamarnik, and Devavrat Shah. Learning graphical models from the Glauber dynamics. *IEEE Trans. Inform. Theory*, 64(6):4072–4080, 2018.

[CM24] Leonardo N. Coregliano and Maryanthe Malliaris. High-arity PAC learning via exchangeability, 2024.

[DSS14] Amit Daniely and Shai Shalev-Shwartz. Optimal learners for multiclass problems. In Maria Florina Balcan, Vitaly Feldman, and Csaba Szepesvári, editors, *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 287–316, Barcelona, Spain, Jun 2014. PMLR.

[Gam03] David Gamarnik. Extension of the PAC framework to finite and countable Markov chains. *IEEE Trans. Inform. Theory*, 49(1):338–345, 2003.

[Hau92] David Haussler. Decision-theoretic generalizations of the PAC model for neural net and other learning applications. *Inform. and Comput.*, 100(1):78–150, 1992.

[Hau95] 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.

[HL94] D. P. Helmbold and P. M. Long. Tracking drifting concepts by minimizing disagreements. *Machine Learning*, 14:27–45, 1994.

[Nat89] Balas K. Natarajan. On learning sets and functions. *Machine Learning*, 4(1):67–97, 1989.

[SSBD14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[SvS23] Nikola Sandrić and Stjepan ˇ Sebek. Learning from non-irreducible Markov chains. *J. Math. Anal. Appl.*, 523(2):Paper No. 127049, 14, 2023.

[SW10] Hongwei Sun and Qiang Wu. Regularized least square regression with dependent samples. *Adv. Comput. Math.*, 32(2):175–189, 2010.

[Val84] L. G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, nov 1984.

[VČ71] V. N. Vapnik and A. Ja. Červonenkis. The uniform convergence of frequencies of the appearance of events to their probabilities. *Teor. Verojatnost. i Primenen.*, 16:264–279, 1971.

[VC15] V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity*, pages 11–30. Springer, Cham, 2015. Reprint of Theor. Probability Appl. **16** (1971), 264–280.

[ZLX12] Bin Zou, Luoqing Li, and Zongben Xu. Generalization performance of least-square regularized regression algorithm with Markov chain samples. *J. Math. Anal. Appl.*, 388(1):333–343, 2012.

[ZXC12] Bin Zou, Zongben Xu, and Xiangyu Chang. Generalization bounds of ERM algorithm with V-geometrically ergodic Markov chains. *Adv. Comput. Math.*, 36(1):99–114, 2012.

[ZXX14] Bin Zou, Zong-ben Xu, and Jie Xu. Generalization bounds of ERM algorithm with Markov chain samples. *Acta Math. Appl. Sin. Engl. Ser.*, 30(1):223–238, 2014.

[ZZX09] Bin Zou, Hai Zhang, and Zongben Xu. Learning from uniformly ergodic Markov chains. *J. Complexity*, 25(2):188–200, 2009.

A High-arity PAC with high-order variables

In this section we lay out the definitions from high-arity PAC learning [CM24] in their full generality including high-order variables and the new definitions, theorems and proofs from this paper in the same full generality. The definitions from [CM24] are done in a streamlined manner, so we refer the reader to the original paper for a more thorough treatment accompanied by intuition; to facilitate the process, numbers in square brackets below refer to the exact location of the corresponding definition in [CM24].

A.1 Definitions in the non-partite

Definition A.1 (Borel templates [3.1])**.** By a Borel space, we mean a standard Borel space, i.e., a measurable space that is Borel-isomorphic to a Polish space when equipped with the σ-algebra of Borel sets. The space of probability measures on a Borel space Λ is denoted $\Pr(\Lambda)$.

1. [3.1.1] A *Borel template* is a sequence $\Omega = (\Omega_i)_{i\in\mathbb{N}_+}$, where $\Omega_i = (X_i, \mathcal{B}_i)$ is a non-empty Borel space.
2. [3.1.2] A *probability template* on a Borel template Ω is a sequence $\mu = (\mu_i)_{i\in\mathbb{N}_+}$, where $\mu_i \in \Pr(\Omega_i)$ is a probability measure on Ω_i. We denote the space of probability templates on a Borel template Ω by $\Pr(\Omega)$.
3. [3.1.4] For a (finite or) countable set V and a Borel template Ω, we define
$$\mathcal{E}_V(\Omega) \stackrel{\text{def}}{=} \prod_{A\in r(V)} X_{|A|}$$
equipping it with the product σ-algebra, where
$$r(V) \stackrel{\text{def}}{=} \{A \subseteq V \mid A \text{ finite non-empty}\}. \tag{A.1}$$
If $\mu \in \Pr(\Omega)$ is a probability template on Ω, then we let $\mu^V \stackrel{\text{def}}{=} \bigotimes_{A\in r(V)} \mu_{|A|}$ be the product measure. We use the shorthands $r(m) \stackrel{\text{def}}{=} r([m])$, $\mathcal{E}_m(\Omega) \stackrel{\text{def}}{=} \mathcal{E}_{[m]}(\Omega)$ and $\mu^m \stackrel{\text{def}}{=} \mu^{[m]}$ when $m \in \mathbb{N}$ (where $[m] \stackrel{\text{def}}{=} \{1, \dots, m\}$).
4. [3.1.5] For an injective function $\alpha\colon U \to V$ between countable sets, we contra-variantly define the map $\alpha^*\colon \mathcal{E}_V(\Omega) \to \mathcal{E}_U(\Omega)$ by
$$\alpha^*(x)_A \stackrel{\text{def}}{=} x_{\alpha(A)} \qquad \big(x \in \mathcal{E}_V(\Omega), A \in r(U)\big).$$

Definition A.2 (Hypotheses [3.2, 3.5])**.** Let Ω be a Borel template, $\Lambda = (Y, \mathcal{B}')$ be a non-empty Borel space and $k \in \mathbb{N}_+$.

1. [3.2.1] The set of *k-ary hypotheses* from Ω to Λ, denoted $\mathcal{F}_k(\Omega, \Lambda)$, is the set of (Borel) measurable functions from $\mathcal{E}_k(\Omega)$ to Λ.
2. [3.2.2] A *k-ary hypothesis class* is a subset $\mathcal{H}$ of $\mathcal{F}_k(\Omega, \Lambda)$ equipped with a σ-algebra such that:
   i. the evaluation map $\mathrm{ev}\colon \mathcal{H} \times \mathcal{E}_k(\Omega) \to \Lambda$ given by $\mathrm{ev}(H, x) \stackrel{\text{def}}{=} H(x)$ is measurable;

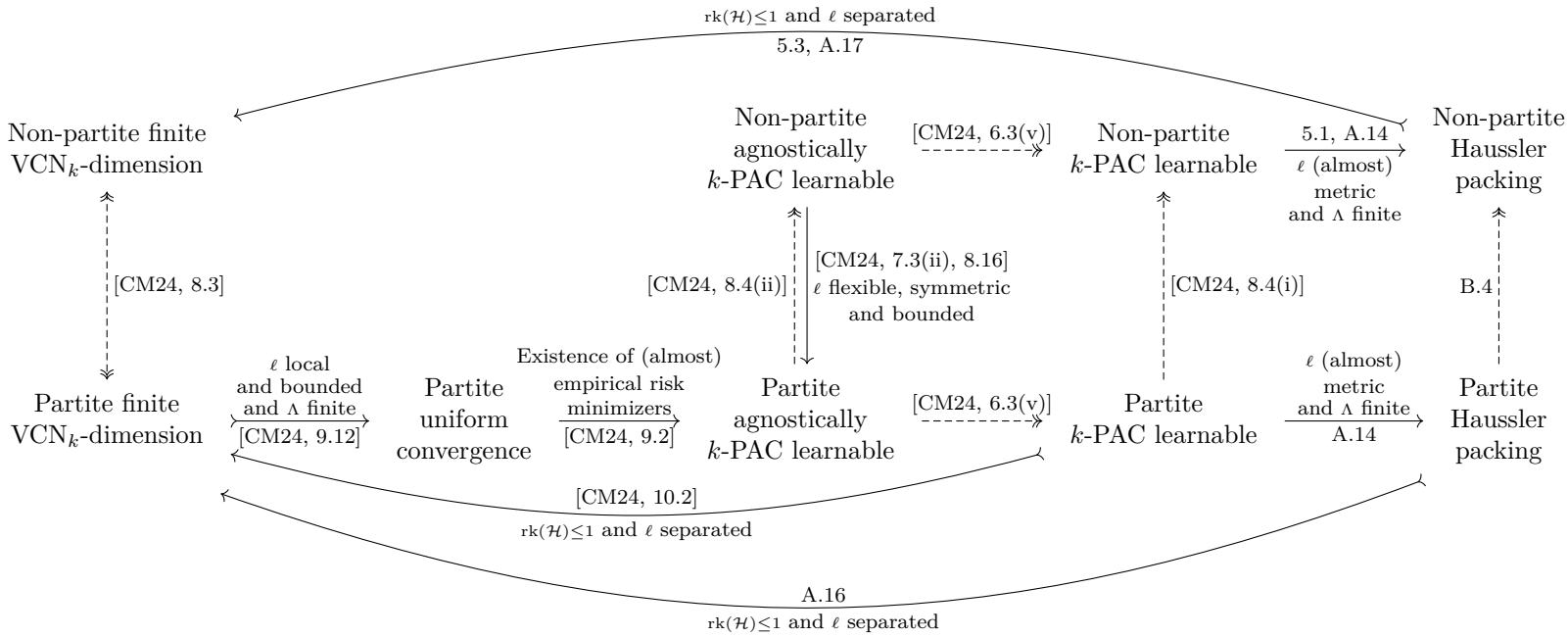

Figure 1: Diagram of implications between different high-arity PAC learning notions. Labels on arrows contain the number of the theorem (or the specific proposition in [CM24]) that contains the proof of the implication and extra hypotheses needed. In the above "ℓ (almost) metric" means that either ℓ is metric or ℓ is separated and bounded. Arrows with two heads ($\twoheadrightarrow$) are tight in some sense with an obvious proof of tightness. Dashed arrows involve a construction (meaning that either the hypothesis class changes and/or the loss function changes) due to being in different settings; this also means that objects in one of the sides of the implication might not be completely general (as they are required to be in the image of the construction). Arrows with tails ($\rightarrowtail$) mean that exactly one of the sides involves a loss function. Under appropriate hypotheses, all items are proved equivalent (for example, if Λ is finite and the loss is the 0/1-loss $\ell_{0/1}$).

ii. for every $H \in \mathcal{H}$, the set $\{H\}$ is measurable;

iii. 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[5] (i.e., measurable in every completion of a probability measure on Υ).

3. [3.2.3] Given $F \in \mathcal{F}_k(\Omega, \Lambda)$ and $m \in \mathbb{N}$, we define the function $F^*_m \colon \mathcal{E}_m(\Omega) \to \Lambda^{([m])_k}$ by

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

where $([m])_k$ is the set of injections $[k] \to [m]$; when $k = m$, we have $F^*_k \colon \mathcal{E}_k(\Omega) \to \Lambda^{S_k}$, where $S_k = ([k])_k$ is the symmetric group on $[k]$.

4. [3.5.1] The *rank* of a k-ary hypothesis $F \in \mathcal{F}_k(\Omega, \Lambda)$, denoted $\mathrm{rk}(F)$ is the minimum $r \in \mathbb{N}$ such that F factors as

$$F(x) \stackrel{\text{def}}{=} F'\big((x_A)_{A \in r(k), |A| \le r}\big) \qquad (x \in \mathcal{E}_k(\Omega))$$

for some function $F' \colon \prod_{A \in r(k), |A| \le r} X_A \to \Lambda$.

5. [3.5.2] The *rank* of a k-ary hypothesis class $\mathcal{H} \subseteq \mathcal{F}_k(\Omega, \Lambda)$ is defined as

$$\mathrm{rk}(\mathcal{H}) \stackrel{\text{def}}{=} \sup_{F \in \mathcal{H}} \mathrm{rk}(F).$$

Definition A.3 (Loss functions [3.7])**.** Let Ω be a Borel template, Λ be a non-empty Borel space and $k \in \mathbb{N}_+$.

1. [3.7.1] A *k-ary loss function* over Λ is a measurable function $\ell \colon \mathcal{E}_k(\Omega) \times \Lambda^{S_k} \times \Lambda^{S_k} \to \mathbb{R}_{\ge 0}$.

2. [3.7.2] For a k-ary loss function ℓ, we define

$$\|\ell\|_\infty \stackrel{\text{def}}{=} \sup_{\substack{x \in \mathcal{E}_k(\Omega) \\ y, y' \in \Lambda^{S_k}}} \ell(x, y, y'), \qquad\qquad s(\ell) \stackrel{\text{def}}{=} \inf_{\substack{x \in \mathcal{E}_k(\Omega) \\ y, y' \in \Lambda^{S_k} \\ y \neq y'}} \ell(x, y, y').$$

3. [3.7.3 simplified] A k-ary loss function is:

bounded if $\|\ell\|_\infty < \infty$.

separated if $s(\ell) > 0$ and $\ell(x, y, y) = 0$ for every $x \in \mathcal{E}_k(\Omega)$ and every $y \in \Lambda^{S_k}$.

4. [3.7.4] For a k-ary loss function ℓ, hypotheses $F, H \in \mathcal{F}_k(\Omega, \Lambda)$ and a probability template $\mu \in \Pr(\Omega)$, the *total loss* of H with respect to μ, F and ℓ is

$$L_{\mu, F, \ell}(H) \stackrel{\text{def}}{=} \mathbb{E}_{\boldsymbol{x} \sim \mu^k}\Big[\ell\big(\boldsymbol{x}, H^*_k(\boldsymbol{x}), F^*_k(\boldsymbol{x})\big)\Big].$$

[5] Footnote 3 also applies here.

5. [3.7.5] We say that $F \in \mathcal{F}_k(\Omega, \Lambda)$ is *realizable* in $\mathcal{H} \subseteq \mathcal{F}_k(\Omega, \Lambda)$ with respect to a k-ary loss function ℓ and $\mu \in \Pr(\Omega)$ if $\inf_{H \in \mathcal{H}} L_{\mu,F,\ell}(H) = 0$.

Definition A.4 (k-PAC learnability [3.8])**.** Let Ω be a Borel template, Λ be a non-empty Borel space and $\mathcal{H} \subseteq \mathcal{F}_k(\Omega, \Lambda)$ be a k-ary hypothesis class.

1. [3.8.2] A *(k-ary) learning algorithm* for $\mathcal{H}$ is a measurable function
$$\mathcal{A} \colon \bigcup_{m \in \mathbb{N}} (\mathcal{E}_m(\Omega) \times \Lambda^{([m])_k}) \to \mathcal{H}.$$

2. [3.8.3] We say that $\mathcal{H}$ is *k-PAC learnable* with respect to a k-ary loss function ℓ if there exist a learning algorithm $\mathcal{A}$ for $\mathcal{H}$ and a function $m^{\mathrm{PAC}}_{\mathcal{H},\ell,\mathcal{A}} \colon (0,1)^2 \to \mathbb{R}_{\geq 0}$ such that for every $\varepsilon, \delta \in (0,1)$, every $\mu \in \Pr(\Omega)$ and every $F \in \mathcal{F}_k(\Omega, \Lambda)$ that is realizable in $\mathcal{H}$ with respect to ℓ and μ, we have
$$\mathbb{P}_{\boldsymbol{x} \sim \mu^m}\left[L_{\mu,F,\ell}\Big(\mathcal{A}\big(\boldsymbol{x}, F^*_m(\boldsymbol{x})\big)\Big) \leq \varepsilon\right] \geq 1 - \delta$$
for every integer $m \geq m^{\mathrm{PAC}}_{\mathcal{H},\ell,\mathcal{A}}(\varepsilon, \delta)$. A learning algorithm $\mathcal{A}$ satisfying the above is called a *k-PAC learner* for $\mathcal{H}$ with respect to ℓ.

Definition A.5 (VCN_k-dimension [3.14])**.** Let Ω be a Borel template, Λ be a non-empty Borel space, $k \in \mathbb{N}_+$ and $\mathcal{H} \subseteq \mathcal{F}_k(\Omega, \Lambda)$ be a k-ary hypothesis class.

1. [3.14.1] For $H \in \mathcal{F}_k(\Omega, \Lambda)$ and $x \in \mathcal{E}_{k-1}(\Omega)$, let
$$H^*_k(x, -) \colon \prod_{A \in r(k) \setminus r(k-1)} X_{|A|} \to \Lambda^{S_k}$$
be the function obtained from H^*_k by fixing its $\mathcal{E}_{k-1}(\Omega)$ arguments to be x and let
$$\mathcal{H}(x) \stackrel{\mathrm{def}}{=} \{H^*_k(x, -) \mid H \in \mathcal{H}\}.$$

2. [3.14.2] The *Vapnik–Chervonenkis–Natarajan k-dimension* of $\mathcal{H}$ (VCN_k*-dimension*) is defined as
$$\mathrm{VCN}_k(\mathcal{H}) \stackrel{\mathrm{def}}{=} \sup_{x \in \mathcal{E}_{k-1}(\Omega)} \mathrm{Nat}(\mathcal{H}(x)).$$

A.2 Definitions in the partite

Definition A.6 (Borel k-partite templates [4.1])**.** Let $k \in \mathbb{N}_+$.

1. [4.1.1] A *Borel k-partite template* is a sequence $\Omega = (\Omega_A)_{A \in r(k)}$, where $\Omega_A = (X_A, \mathcal{B}_A)$ is a non-empty (standard) Borel space and $r(k) = r([k])$ is given by (A.1).

2. [4.1.2] A *probability k-partite template* on a Borel k-partite template Ω is a sequence $\mu = (\mu_A)_{A \in r(k)}$, where μ_A is a probability measure on Ω_A. The space of probability k-partite templates on Ω is denoted $\Pr(\Omega)$.

3. [4.1.4 simplified] For a Borel k-partite template Ω, a non-empty Borel space Λ and $m \in \mathbb{N}_+$, we define

$$\mathcal{E}_m(\Omega) \stackrel{\text{def}}{=} \prod_{f \in r_k(m)} X_{\text{dom}(f)},$$

equipping it with the product σ-algebra, where

$$r_k(m) \stackrel{\text{def}}{=} \{f \colon A \to [m] \mid A \in r(k)\} = \bigcup_{A \in r(k)} [m]^A.$$

If $\mu \in \Pr(\Omega)$ is a probability k-partite template on Ω, we let $\mu^m \stackrel{\text{def}}{=} \bigotimes_{f \in r_k(m)} \mu_{\text{dom}(f)}$ be the product measure.

4. [4.1.5 simplified] For $\alpha \in [m]^k$, we define the map $\alpha^* \colon \mathcal{E}_m(\Omega) \to \mathcal{E}_1(\Omega)$ by

$$\alpha^*(x)_f \stackrel{\text{def}}{=} x_{\alpha|_{\text{dom}(f)}} \qquad (x \in \mathcal{E}_m(\Omega), f \in r_k(1)).$$

Definition A.7 (k-Partite hypotheses [4.2, 4.5])**.** Let $k \in \mathbb{N}_+$, let Ω be a Borel k-partite template and let $\Lambda = (Y, \mathcal{B}')$ be a non-empty Borel space.

1. [4.2.1] The set of *k-partite hypotheses* from Ω to Λ, denoted $\mathcal{F}_k(\Omega, \Lambda)$, is the set f (Borel) measurable functions from $\mathcal{E}_1(\Omega)$ to Λ.

2. [4.2.2] A *k-partite hypothesis class* is a subset $\mathcal{H}$ of $\mathcal{F}_k(\Omega, \Lambda)$ equipped with a σ-algebra such that:

   i. the evaluation map $\text{ev} \colon \mathcal{H} \times \mathcal{E}_1(\Omega) \to \Lambda$ given by $\text{ev}(H, x) \stackrel{\text{def}}{=} H(x)$ is measurable;

   ii. for every $H \in \mathcal{H}$, the set $\{H\}$ is measurable;

   iii. 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[6] (i.e., measurable in every completion of a probability measure on Υ).

3. [4.2.3 simplified] For a k-partite hypothesis $F \in \mathcal{F}_k(\Omega, \Lambda)$, we let $F^*_m \colon \mathcal{E}_m(\Omega) \to \Lambda^{[m]^k}$ be given by

$$F^*_m(x)_\alpha \stackrel{\text{def}}{=} F(\alpha^*(x)) \qquad (x \in \mathcal{E}_m(\Omega), \alpha \in [m]^k).$$

4. [4.5.1] The *rank* of a k-partite hypothesis $F \in \mathcal{F}_k(\Omega, \Lambda)$, denote $\text{rk}(F)$ is the minimum $r \in \mathbb{N}$ such that F factors as

$$F(x) = F'((x_f)_{f \in r_k(1), |\text{dom}(f)| \le r}) \qquad (x \in \mathcal{E}_1(\Omega))$$

for some function $F' \colon \prod_{f \in r_k(1), |\text{dom}(f)| \le r} X_{\text{dom}(f)} \to \Lambda$.

[6] Footnote 3 also applies here.

5. [4.5.2] The *rank* of a k-partite hypothesis class $\mathcal{H} \subseteq \mathcal{F}_k(\Omega, \Lambda)$ is defined as

$$\mathrm{rk}(\mathcal{H}) \stackrel{\mathrm{def}}{=} \sup_{F \in \mathcal{H}} \mathrm{rk}(F).$$

Definition A.8 (k-Partite loss functions [4.7])**.** Let $k \in \mathbb{N}_+$, let Ω be a Borel k-partite template and let Λ be a non-empty Borel space.

1. [4.7.1] A *k-partite loss function* over Λ is a measurable function $\ell\colon \mathcal{E}_1(\Omega) \times \Lambda \times \Lambda \to \mathbb{R}_{\geq 0}$.

2. [4.7.2] For a k-partite loss function ℓ, we define

$$\|\ell\|_\infty \stackrel{\mathrm{def}}{=} \sup_{\substack{x \in \mathcal{E}_1(\Omega) \\ y, y' \in \Lambda}} \ell(x, y, y'), \qquad\qquad s(\ell) \stackrel{\mathrm{def}}{=} \inf_{\substack{x \in \mathcal{E}_1(\Omega) \\ y, y' \in \Lambda \\ y \neq y'}} \ell(x, y, y').$$

3. [4.7.3] A k-partite loss function is:

   bounded if $\|\ell\|_\infty < \infty$.

   separated if $s(\ell) > 0$ and $\ell(x, y, y)$ for every $x \in \mathcal{E}_1(\Omega)$ and every $y \in \Lambda$.

4. [4.7.4] For a k-partite loss function ℓ, k-partite hypotheses $F, H \in \mathcal{F}_k(\Omega, \Lambda)$ and a probability k-partite template $\mu \in \Pr(\Omega)$, the *total loss* of H with respect to μ, F and ℓ is

$$L_{\mu,F,\ell}(H) \stackrel{\mathrm{def}}{=} \mathbb{E}_{\boldsymbol{x} \sim \mu^1}\Big[\ell\big(\boldsymbol{x}, H(\boldsymbol{x}), F(\boldsymbol{x})\big)\Big].$$

5. [4.7.5] We say that $F \in \mathcal{F}_k(\Omega, \Lambda)$ is *realizable* in $\mathcal{H} \subseteq \mathcal{F}_k(\Omega, \Lambda)$ with respect to a k-partite loss function ℓ and $\mu \in \Pr(\Omega)$ if $\inf_{H \in \mathcal{H}} L_{\mu,F,\ell}(H) = 0$.

6. [4.7.6] The *k-partite $0/1$-loss function* over Λ is defined as $\ell_{0/1}(x, y, y') \stackrel{\mathrm{def}}{=} \mathbb{1}[y \neq y']$.

Definition A.9 (Partite k-PAC learnability [4.8])**.** Let $k \in \mathbb{N}_+$, let Ω be a Borel k-partite template, let Λ be a non-empty Borel space and let $\mathcal{H} \subseteq \mathcal{F}_k(\Omega, \Lambda)$ be a k-partite hypothesis class.

1. [4.8.2] A *(k-partite) learning algorithm* for $\mathcal{H}$ is a measurable function

$$\mathcal{A}\colon \bigcup_{m \in \mathbb{N}} (\mathcal{E}_m(\Omega) \times \Lambda^{[m]^k}) \to \mathcal{H}.$$

2. [4.8.3] We say that $\mathcal{H}$ is *k-PAC learnable* with respect to a k-partite loss function ℓ if there exist a learning algorithm $\mathcal{A}$ for $\mathcal{H}$ and a function $m^{\mathrm{PAC}}_{\mathcal{H},\ell,\mathcal{A}}\colon (0,1)^2 \to \mathbb{R}_{\geq 0}$ such that for every $\varepsilon, \delta \in (0,1)$, every $\mu \in \Pr(\Omega)$ and every $F \in \mathcal{F}_k(\Omega, \Lambda)$ that is realizable in $\mathcal{H}$ with respect to ℓ and μ, we have

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

   for every integer $m \geq m^{\mathrm{PAC}}_{\mathcal{H},\ell,\mathcal{A}}(\varepsilon, \delta)$. A learning algorithm $\mathcal{A}$ satisfying the above is called a *k-PAC learner* for $\mathcal{H}$ with respect to ℓ.

Definition A.10 (Partite VCN_k-dimension [4.13])**.** Let $k \in \mathbb{N}_+$, let Ω be a Borel k-partite template, let Λ be a non-empty Borel space and let $\mathcal{H} \subseteq \mathcal{F}_k(\Omega, \Lambda)$ be a k-partite hypothesis class.

1. [4.13.1] For $A \in \binom{[k]}{k-1}$, let

$$r_{k,A} \stackrel{\text{def}}{=} \{f \in r_k(1) \mid \mathrm{dom}(f) \subseteq A\},$$

for $x \in \prod_{f \in r_{k,A}} X_{\mathrm{dom}(f)}$ and $H \in \mathcal{F}_k(\Omega, \Lambda)$, let

$$H(x, -)\colon \prod_{f \in r_k(1) \setminus r_{k,A}} X_{\mathrm{dom}(f)} \to \Lambda$$

be the function obtained from H by fixing its arguments in $\prod_{f \in r_{k,A}} X_{\mathrm{dom}(f)}$ to be x and let

$$\mathcal{H}(x) \stackrel{\text{def}}{=} \{H(x, -) \mid H \in \mathcal{H}\}.$$

2. [4.13.2] The *Vapnik–Chervonenkis–Natarajan k-dimension* of $\mathcal{H}$ (VCN_k*-dimension*) is defined as

$$\mathrm{VCN}_k(\mathcal{H}) \stackrel{\text{def}}{=} \sup_{\substack{A \in \binom{[k]}{k-1} \\ x \in \prod_{f \in r_{k,A}} X_{\mathrm{dom}(f)}}} \mathrm{Nat}\big(\mathcal{H}(x)\big).$$

A.3 New high-arity PAC definitions

In this subsection, we lay out the new high-arity definitions of this paper in full generality.

Definition A.11 (k-ary Haussler packing property)**.** Let $k \in \mathbb{N}_+$, let Ω be a Borel (k-partite, respectively) template, let Λ be a non-empty Borel space, let $\mathcal{H} \subseteq \mathcal{F}_k(\Omega, \Lambda)$ be a k-ary (k-partite, respectively) hypothesis class and let $\ell\colon \mathcal{E}_k(\Omega) \times \Lambda^{S_k} \times \Lambda^{S_k} \to \mathbb{R}_{\geq 0}$ ($\ell\colon \mathcal{E}_1(\Omega) \times \Lambda \times \Lambda \to \mathbb{R}_{\geq 0}$, respectively) be a k-ary (k-partite, respectively) loss function.

We say that $\mathcal{H}$ has the *k-ary Haussler packing property* with respect to ℓ if there exists a function $m^{\mathrm{HP}}_{\mathcal{H},\ell}\colon (0,1) \to \mathbb{R}_{\geq 0}$ such that for every $\varepsilon \in (0,1)$ and every $\mu \in \Pr(\Omega)$, there exists $\mathcal{H}' \subseteq \mathcal{H}$ with $|\mathcal{H}'| \leq m^{\mathrm{HP}}_{\mathcal{H},\ell}(\varepsilon)$ such that for every $F \in \mathcal{H}$, there exists $H \in \mathcal{H}'$ such that $L_{\mu,F,\ell}(H) \leq \varepsilon$. We refer to elements of $\mathcal{H}'$ as *k-ary Haussler centers* of $\mathcal{H}$ at precision ε with respect to μ and ℓ.

Definition A.12 (Metric loss functions)**.** Let $k \in \mathbb{N}_+$, let Ω be a Borel (k-partite, respectively) template, let Λ be a non-empty Borel space.

We say that a k-ary (k-partite, respectively) loss function $\ell\colon \mathcal{E}_k(\Omega) \times \Lambda^{S_k} \times \Lambda^{S_k} \to \mathbb{R}_{\geq 0}$ ($\ell\colon \mathcal{E}_1(\Omega) \times \Lambda \times \Lambda \to \mathbb{R}_{\geq 0}$, respectively) is *metric* if for every $x \in \mathcal{E}_k(\Omega)$ ($x \in \mathcal{E}_1(\Omega)$, respectively), the function $\ell(x, -, -)$ is a metric on Λ^{S_k} (Λ, respectively) in the usual sense, that is, the following hold for every $x \in \mathcal{E}_k(\Omega)$ and $y, y', y'' \in \Lambda^{S_k}$ ($x \in \mathcal{E}_1(\Omega)$ and $y, y', y'' \in \Lambda$, respectively):

i. We have $\ell(x, y, y') = \ell(x, y', y)$.

ii. We have $\ell(x, y, y') = 0$ if and only if $y = y'$.

iii. We have $\ell(x, y, y'') \leq \ell(x, y, y') + \ell(x, y', y'')$.

A.4 Main results

In this section, we prove the main results in full generality including the high-order variables. For the particular case of the counterpart of Theorem 5.1, Theorem A.14, we will also show that instead of assuming that the loss function ℓ is metric, one could assume that ℓ is separated and bounded; for this, we will use the lemma below that says that separated and bounded loss functions make the total loss satisfy a weak version of triangle inequality with a rescaling factor. This justifies the usage of the name "(almost) metric" for losses that are either metric or both separated and bounded in Figure 1. Finally, we point out that the 0/1-loss function $\ell_{0/1}$ satisfies all hypotheses of Theorems A.14, A.16 and A.17.

Lemma A.13. *Let Ω be a Borel (k-partite, respectively) template, let Λ be a non-empty Borel space, let $\mathcal{H} \subseteq \mathcal{F}_k(\Omega, \Lambda)$ be a k-ary (k-partite, respectively) hypothesis class, let ℓ be a k-ary (k-partite, respectively) loss function and let $\mu \in \Pr(\Omega)$ be a probability (k-partite, respectively) template.*

For each $F, H \in \mathcal{H}$, let

$$D(F,H) \stackrel{\text{def}}{=} \begin{cases} \{x \in \mathcal{E}_k(\Omega) \mid F_k^*(x) \neq H_k^*(x)\}, & \text{in the non-partite case,} \\ \{x \in \mathcal{E}_1(\Omega) \mid F(x) \neq H(x)\}, & \text{in the partite case.} \end{cases}$$

Then the following hold:

i. We have

$$s(\ell) \cdot M(F,H) \leq L_{F,\mu,\ell}(H) \leq \|\ell\|_\infty \cdot M(F,H),$$

where

$$M(F,H) \stackrel{\text{def}}{=} \begin{cases} \mu^k(D(F,H)), & \text{in the non-partite case,} \\ \mu^1(D(F,H)), & \text{in the partite case.} \end{cases}$$

ii. If ℓ is separated and $F, F', H \in \mathcal{H}$, then

$$L_{\mu,F,\ell}(F') \leq \frac{\|\ell\|_\infty}{s(\ell)} \big(L_{\mu,F,\ell}(H) + L_{\mu,F',\ell}(H)\big)$$

Proof. Item (i) follows since

$$L_{\mu,F,\ell}(H) = \begin{cases} \mathbb{E}_{\boldsymbol{x}\sim\mu^k}\Big[\ell\big(\boldsymbol{x}, H_k^*(\boldsymbol{x}), F_k^*(\boldsymbol{x})\big)\Big], & \text{in the non-partite case,} \\ \mathbb{E}_{\boldsymbol{x}\sim\mu^1}\ell\Big[\big(\boldsymbol{x}, H(\boldsymbol{x}), F(\boldsymbol{x})\big)\Big], & \text{in the partite case.} \end{cases}$$

For item (ii), by item (i), we have

$$\begin{aligned} L_{\mu,F,\ell}(F') &\leq \|\ell\|_\infty \cdot M(F,F') \leq \|\ell\|_\infty \cdot \big(M(F,H) + M(F',H)\big) \\ &\leq \frac{\|\ell\|_\infty}{s(\ell)} \cdot \big(L_{\mu,F,\ell}(H) + L_{\mu,F',\ell}(H)\big) \end{aligned}$$

where the second inequality is the (usual) triangle inequality. □

We now prove the counterpart of Theorem 5.1 both in the non-partite and partite settings.

Theorem A.14 (k-PAC learnability implies Haussler packing property)**.** *Let $k \in \mathbb{N}_+$, let Ω be a Borel (k-partite, respectively) template, let Λ be a finite non-empty Borel space, let $\mathcal{H} \subseteq \mathcal{F}_k(\Omega, \Lambda)$ be a k-ary (k-partite, respectively) hypothesis class and let ℓ be a k-ary (k-partite, respectively) loss function. Suppose that $\mathcal{H}$ is k-PAC learnable with a k-PAC learner $\mathcal{A}$.*

Let also

$$\gamma_{\mathcal{H}}(m) \stackrel{def}{=} \sup_{x \in \mathcal{E}_m(\Omega)} |\{H_m^*(x) \mid H \in \mathcal{H}\}|$$

be the maximum number of different patterns (in $\Lambda^{([m])_k}$ in the non-partite case or in $\Lambda^{[m]^k}$ in the partite case) that can be obtained by evaluating all elements of $\mathcal{H}$ in a fixed $x \in \mathcal{E}_m(\Omega)$. Then the following hold:

i. If ℓ is separated and bounded, then $\mathcal{H}$ has the Haussler packing property with associated function

$$\begin{aligned} m_{\mathcal{H},\ell}^{\mathrm{HP}}(\varepsilon) &\stackrel{def}{=} \min_{\delta \in (0,1)} \left\lceil \frac{\gamma_{\mathcal{H}}(\lceil m_{\mathcal{H},\ell,\mathcal{A}}^{\mathrm{PAC}}(s(\ell)\varepsilon/(2\|\ell\|_\infty), \delta)\rceil)}{1-\delta} \right\rceil - 2 \\ &\leq \begin{cases} \min\limits_{\delta \in (0,1)} \left\lceil \dfrac{|\Lambda|^{(\lceil m_{\mathcal{H},\ell,\mathcal{A}}^{\mathrm{PAC}}(s(\ell)\varepsilon/(2\|\ell\|_\infty),\delta)\rceil)_k}}{1-\delta} \right\rceil - 2, & \text{in the non-partite case,} \\ \min\limits_{\delta \in (0,1)} \left\lceil \dfrac{|\Lambda|^{\lceil m_{\mathcal{H},\ell,\mathcal{A}}^{\mathrm{PAC}}(s(\ell)\varepsilon/(2\|\ell\|_\infty),\delta)\rceil^k}}{1-\delta} \right\rceil - 2, & \text{in the partite case.} \end{cases} \end{aligned} \tag{A.2}$$

ii. If ℓ is metric, then $\mathcal{H}$ has the Haussler packing property with associated function

$$\begin{aligned} m_{\mathcal{H},\ell}^{\mathrm{HP}}(\varepsilon) &\stackrel{def}{=} \min_{\delta \in (0,1)} \left\lceil \frac{\gamma_{\mathcal{H}}(\lceil m_{\mathcal{H},\ell,\mathcal{A}}^{\mathrm{PAC}}(\varepsilon/2, \delta)\rceil)}{1-\delta} \right\rceil - 2 \\ &\leq \begin{cases} \min\limits_{\delta \in (0,1)} \left\lceil \dfrac{|\Lambda|^{(\lceil m_{\mathcal{H},\ell,\mathcal{A}}^{\mathrm{PAC}}(\varepsilon/2,\delta)\rceil)_k}}{1-\delta} \right\rceil - 2, & \text{in the non-partite case,} \\ \min\limits_{\delta \in (0,1)} \left\lceil \dfrac{|\Lambda|^{\lceil m_{\mathcal{H},\ell,\mathcal{A}}^{\mathrm{PAC}}(\varepsilon/2,\delta)\rceil^k}}{1-\delta} \right\rceil - 2, & \text{in the partite case.} \end{cases} \end{aligned} \tag{A.3}$$

Proof. For item (i), note that due to the ceilings on the right-hand sides of (A.2), the minima are indeed attained as the functions only take values in $\mathbb{N}$.

The inequalities also clearly follow from the trivial bound:

$$\gamma_{\mathcal{H}}(m) \leq \begin{cases} |\Lambda|^{(m)_k}, & \text{in the non-partite case,} \\ |\Lambda|^{m^k}, & \text{in the partite case.} \end{cases}$$

Suppose for a contradiction that the result does not hold, that is, there exist $\varepsilon \in (0,1)$ and $\mu \in \Pr(\Omega)$ such that if m is given by the right-hand side of (A.2), then for every $\mathcal{H}' \subseteq \mathcal{H}$ with $|\mathcal{H}'| \leq m$, there exists $F \in \mathcal{H}$ such that $L_{\mu,F,\ell}(H) > \varepsilon$ for every $H \in \mathcal{H}'$. By starting with $\mathcal{H}' \stackrel{\text{def}}{=} \varnothing$ and inductively applying this property with $\mathcal{H}' \stackrel{\text{def}}{=} \{F_1, \ldots, F_t\}$ to produce F_{t+1}, it follows that there exist $F_1, \ldots, F_{m+1} \in \mathcal{H}$ such that for every $i, j \in [m+1]$ with $i < j$, we have $L_{\mu,F_i,\ell}(F_j) > \varepsilon$.

Let $\delta \in (0,1)$ attain the first minimum in (A.2) and let

$$\widetilde{m} \stackrel{\text{def}}{=} \left\lceil m^{\text{PAC}}_{\mathcal{H},\ell,\mathcal{A}}\left(\frac{s(\ell)\cdot\varepsilon}{2\cdot\|\ell\|_\infty},\delta\right)\right\rceil.$$

For each $x \in \mathcal{E}_{\widetilde{m}}(\Omega)$, let

$$Y(x) \stackrel{\text{def}}{=} \{H^*_m(x) \mid H \in \mathcal{H}\}$$

and note that $|Y(x)| \leq \gamma_{\mathcal{H}}(\widetilde{m})$.

For each $i \in [m+1]$, define the set

$$C_i \stackrel{\text{def}}{=} \left\{x \in \mathcal{E}_{\widetilde{m}}(\Omega) \;\middle|\; \forall y \in Y(x), L_{\mu,F_i,\ell}(\mathcal{A}(x,y)) > \frac{s(\ell)\cdot\varepsilon}{2\cdot\|\ell\|_\infty}\right\}.$$

Note that by taking $y \stackrel{\text{def}}{=} (F_i)^*_m(x) \in Y(x)$ and using the fact that ℓ is separated so that F_i is realizable in $\mathcal{H}$ w.r.t. ℓ and μ, PAC learnability implies that $\mu(C_i) \leq \delta$.

Define now the function $G\colon \mathcal{E}_{\widetilde{m}}(\Omega) \to \mathbb{R}_{\geq 0}$ by

$$\begin{aligned} G(x) &\stackrel{\text{def}}{=} \sum_{i=1}^{m+1} \mathbb{1}_{C_i}(x) \\ &= \left|\left\{i \in [m+1] \;\middle|\; \forall y \in Y(x), L_{\mu,F_i,\ell}(\mathcal{A}(x,y)) > \frac{s(\ell)\cdot\varepsilon}{2\cdot\|\ell\|_\infty}\right\}\right|. \end{aligned}$$

We claim that for every $x \in \mathcal{E}_{\widetilde{m}}(\Omega)$ and every $y \in Y(x)$, there exists at most one $i \in [m+1]$ such that $L_{\mu,F_i,\ell}(\mathcal{A}(x,y)) \leq s(\ell)\varepsilon/(2\|\ell\|_\infty)$. Indeed, if not, then for some $i,j \in [m+1]$ with $i<j$, we would get

$$\frac{s(\ell)\cdot\varepsilon}{\|\ell\|_\infty} \geq L_{\mu,F_i,\ell}(\mathcal{A}(x,y)) + L_{\mu,F_j,\ell}(\mathcal{A}(x,y)) \geq \frac{s(\ell)}{\|\ell\|_\infty}\cdot L_{\mu,F_i,\ell}(F_j),$$

where the last inequality follows from Lemma A.13(ii); the above would then contradict $L_{\mu,F_i,\ell}(F_j) > \varepsilon$.

Thus, we conclude that

$$G(x) \geq m+1-|Y(x)| \geq m+1-\gamma_{\mathcal{H}}(\widetilde{m}) \tag{A.4}$$

for every $x \in \mathcal{E}_{\widetilde{m}}(\Omega)$.

On the other hand, since $\mu(C_i) \leq \delta$ for every $i \in [m+1]$, we get

$$\int_{\mathcal{E}_{\widetilde{m}}(\Omega)} G(x)\; d\mu^{\widetilde{m}}(x) \leq (m+1)\delta,$$

which together with (A.4) implies

$$m \leq \frac{\gamma_{\mathcal{H}}(\widetilde{m})}{1-\delta} - 1,$$

contradicting the definitions of m, δ and $\widetilde{m}$.

The proof of item (ii) is completely analogous (see the proof of Theorem 5.1), but instead of using Lemma A.13, we use triangle inequality since ℓ is metric, which allows us to improve the bound to (A.3) instead of (A.2) (also note that the fact that ℓ is metric implies that $L_{\mu,F_i,\ell}(F_i) = 0$). □

Remark A.15. The final bound provided on Theorem A.14 is provably not tight when $\mathrm{rk}(\mathcal{H}) \leq 1$ and ℓ is bounded. This is because a posteriori, we know that all results of high-arity PAC in [CM24] along with Theorem A.17 prove that the notions considered are also equivalent to finiteness of VCN_k-dimension, which in turn implies that

$$\gamma_{\mathcal{H}}(m) \leq \begin{cases} (m+1)^{\mathrm{VCN}_k(\mathcal{H})\cdot\binom{m}{k-1}} \cdot \binom{|\Lambda|}{2}^{\mathrm{VCN}_k(\mathcal{H})\cdot\binom{m}{k-1}}, & \text{in the non-partite case,} \\ (m+1)^{\mathrm{VCN}_k(\mathcal{H})\cdot m^{k-1}} \cdot \binom{|\Lambda|}{2}^{\mathrm{VCN}_k(\mathcal{H})\cdot m^{k-1}}, & \text{in the partite case.} \end{cases} \tag{A.5}$$
$$\leq \left(\frac{|\Lambda|^2 \cdot (m+1)}{2}\right)^{\mathrm{VCN}_k(\mathcal{H})\cdot m^{k-1}},$$

instead of the trivial bounds for $\gamma_{\mathcal{H}}(m)$ used on Theorem A.14.

For the counterpart of Theorem 5.3, it will be more convenient to split it into the partite version (Theorem A.16) and the non-partite version (Theorem A.17). We start with the easier one: the partite.

Theorem A.16 (Haussler packing property implies finite VCN_k-dimension, partite version)**.** *Let $k \in \mathbb{N}_+$, let Ω be a Borel k-partite template, let Λ be a finite non-empty Borel space, let $\mathcal{H} \subseteq \mathcal{F}_k(\Omega, \Lambda)$ be a k-partite hypothesis class and let $\ell\colon \mathcal{E}_1(\Omega) \times \Lambda \times \Lambda \to \mathbb{R}_{\geq 0}$ be a k-partite loss function. Suppose ℓ is separated and $\mathrm{rk}(\mathcal{H}) \leq 1$.*

If $\mathcal{H}$ has the Haussler packing property, then

$$\mathrm{VCN}_k(\mathcal{H}) \leq \min_{\varepsilon \in (0, \min\{s(\ell)/2, 1\})} \left\lfloor \frac{\log_2 \lfloor m^{\mathrm{HP}}_{\mathcal{H},\ell}(\varepsilon) \rfloor}{1 - h_2(\varepsilon / s(\ell))} \right\rfloor, \tag{A.6}$$

where

$$h_2(t) \stackrel{\mathrm{def}}{=} t \log_2 \frac{1}{t} + (1-t) \log_2 \frac{1}{1-t}$$

denotes the binary entropy.

Proof. Note that the minimum in (A.6) is indeed attained as the function only takes values in $\mathbb{N} \cup \{-\infty\}$, so let $\varepsilon \in (0, \min\{s(\ell)/2, 1\})$ attain the minimum, let d be the value of the minimum and let

$$m \stackrel{\mathrm{def}}{=} \lfloor m^{\mathrm{HP}}_{\mathcal{H},\ell}(\varepsilon) \rfloor$$

so that

$$d = \left\lfloor \frac{\log_2 m}{1 - h_2(\varepsilon / s(\ell))} \right\rfloor.$$

When $\mathcal{H}$ is empty, the result is trivial as $\mathrm{VCN}_k(\mathcal{H}) = -\infty$ so suppose $\mathcal{H}$ is non-empty (hence $m \geq 1$ and $d \geq 0$).

By the definition of VCN_k-dimension, we have to show that if $A \in \binom{[k]}{k-1}$ and $x \in \prod_{f \in r_{k,A}} X_{\mathrm{dom}(f)}$, then $\mathrm{Nat}(\mathcal{H}(x)) \leq d$. In turn, it suffices to show that if $V \subseteq \prod_{f \in r_k(1) \setminus r_{k,A}} X_{\mathrm{dom}(f)}$ is a (finite) set that is Natarajan-shattered by $\mathcal{H}(x)$, then $|V| \leq d$.

Let

$$R \stackrel{\mathrm{def}}{=} r_k(1) \setminus (r_{k,A} \cup \{1^{\{a\}}\})$$

where $1^{\{a\}}$ is the unique function $\{a\} \to [1]$ and let $x' \in \prod_{f \in R} X_{\mathrm{dom}(f)}$ be any fixed point.

Since $\mathrm{rk}(\mathcal{H}) \leq 1$, it follows that the projection V' of V onto the coordinate indexed by $1^{\{a\}}$ is Natarajan-shattered by

$$\{H(x, x', -) \mid H \in \mathcal{H}\}.$$

Let $n \stackrel{\mathrm{def}}{=} |V|$ (which is equal to $|V'|$ as $\mathrm{rk}(\mathcal{H}) \leq 1$ and V is Natarajan-shattered by $\mathcal{H}(x)$) and let $\mu \in \Pr(\Omega)$ be the probability k-partite template given by:

- For each $f \in r_{k,A}$, μ_f is the Dirac delta concentrated on x_f.
- For each $f \in R$, μ_f is the Dirac delta concentrated on x'_f.
- $\mu_{1^{\{a\}}}$ is the uniform measure on V'.

Since V' is Natarajan-shattered by $\{H(x, x', -) \mid H \in \mathcal{H}\}$, there exist functions $f_0, f_1 \colon V' \to \Lambda$ with $f_0(v) \neq f_1(v)$ for every $v \in V'$ and there exists a family $\{F_U \mid U \subseteq V'\} \subseteq \mathcal{H}$ such that for every $U \subseteq V'$ and every $v \in V'$, we have $F_U(x, x', v) = f_{\mathbb{1}[v \in U]}(v)$.

Using now our definition of m via Haussler packing property, we know that there exists $\mathcal{H}' \subseteq \mathcal{H}$ such that $|\mathcal{H}'| \leq m$ and for every $U \subseteq V'$, there exists $H \in \mathcal{H}'$ such that $L_{\mu,F_U,\ell}(H) \leq \varepsilon$.

For each $H \in \mathcal{H}'$, let

$$\begin{aligned} U_H &\stackrel{\mathrm{def}}{=} \{v \in V' \mid H(x, x', v) = f_1(v)\}, \\ B(H) &\stackrel{\mathrm{def}}{=} \{U \subseteq V' \mid L_{\mu,F_U,\ell}(H) \leq \varepsilon\}, \\ B'(H) &\stackrel{\mathrm{def}}{=} \left\{U \subseteq V' \;\middle|\; |U \mathbin{\triangle} U_H| \leq \frac{\varepsilon \cdot n}{s(\ell)}\right\}. \end{aligned}$$

The Haussler packing property assumption implies

$$\bigcup_{H \in \mathcal{H}'} B(H) = 2^{V'}. \tag{A.7}$$

Since ℓ is separated, by Lemma A.13(i), for every $H \in \mathcal{H}'$ and every $U \subseteq V'$, we have

$$L_{\mu,F_U,\ell}(H) \geq s(\ell) \cdot \mu^k(D(F_U, H)) \geq s(\ell) \cdot \mu^k(D(F_U, F_{U_H})) \geq \frac{s(\ell)}{n} \cdot |U \mathbin{\triangle} U_H|,$$

hence $B(H) \subseteq B'(H)$, which together with (A.7) implies $\bigcup_{H \in \mathcal{H}'} B'(H) = 2^{V'}$.

Since $\varepsilon / s(\ell) < 1/2$, by Lemma 5.2, we get

$$n \leq \frac{\log_2 |\mathcal{H}'|}{1 - h_2(\varepsilon / s(\ell))} \leq \frac{\log_2 m}{1 - h_2(\varepsilon / s(\ell))},$$

which yields $n \leq d$ as n is an integer. □

Theorem A.17 (Haussler packing property implies finite VCN_k-dimension, non-partite version)**.** *Let $k \in \mathbb{N}_+$, let Ω be a Borel template, let Λ be a finite non-empty Borel space, let $\mathcal{H} \subseteq \mathcal{F}_k(\Omega, \Lambda)$ be a k-ary hypothesis class and let $\ell \colon \mathcal{E}_k(\Omega) \times \Lambda^{S_k} \times \Lambda^{S_k} \to \mathbb{R}_{\geq 0}$ be a k-ary loss function. Suppose that ℓ is separated and* $\mathrm{rk}(\mathcal{H}) \leq 1$.

If $\mathcal{H}$ has the Haussler packing property, then

$$\mathrm{VCN}_k(\mathcal{H}) \leq \min_{\varepsilon \in (0, \min\{s(\ell)\cdot k!/(2k^k), 1\})} \left\lfloor \frac{\log_2 \lfloor m^{\mathrm{HP}}_{\mathcal{H},\ell}(\varepsilon) \rfloor}{1 - h_2(\varepsilon \cdot k^k/(s(\ell)\cdot k!))} \right\rfloor, \tag{A.8}$$

where

$$h_2(t) \stackrel{\mathrm{def}}{=} t \log_2 \frac{1}{t} + (1-t) \log_2 \frac{1}{1-t}$$

denotes the binary entropy.

Proof. Note that the minimum in (A.8) is indeed attained as the function only takes values in $\mathbb{N} \cup \{-\infty\}$, so let $\varepsilon \in (0, \min\{s(\ell)k!/(2k^k), 1\})$ attain the minimum, let d be the value of the minimum and let

$$m \stackrel{\mathrm{def}}{=} \lfloor m^{\mathrm{HP}}_{\mathcal{H},\ell}(\varepsilon) \rfloor$$

so that

$$d = \left\lfloor \frac{\log_2 m}{1 - h_2(\varepsilon \cdot k^k/(s(\ell)\cdot k!))} \right\rfloor.$$

When $\mathcal{H}$ is empty, the result is trivial as $\mathrm{VCN}_k(\mathcal{H}) = -\infty$, so suppose $\mathcal{H}$ is non-empty (hence $m \geq 1$ and $d \geq 0$).

By the definition of VCN_k-dimension, we have to show that if $x \in \mathcal{E}_{k-1}(\Omega)$, then $\mathrm{Nat}(\mathcal{H}(x)) \leq d$. In turn, it suffices to show that if $V \subseteq \prod_{A \in r(k)\setminus r(k-1)} X_{|A|}$ is a (finite) set that is Natarajan-shattered by $\mathcal{H}(x)$, then $|V| \leq d$.

Let

$$R \stackrel{\mathrm{def}}{=} r(k) \setminus (r(k-1) \cup \{\{k\}\})$$

and let $x' \in \prod_{A \in R} X_A$ be any fixed point.

Since $\mathrm{rk}(\mathcal{H}) \leq 1$, it follows that the projection V' of V onto the coordinate indexed by $\{k\}$ is Natarajan-shattered by

$$\{H^*_k(x, x', -) \mid H \in \mathcal{H}\}.$$

Note that $|V'| = |V|$ as $\mathrm{rk}(\mathcal{H}) \leq 1$ and V is Natarajan-shattered by $\mathcal{H}(x)$; thus our goal is to show $|V'| \leq d$.

Let $n \stackrel{\mathrm{def}}{=} |V|$ (which is equal to $|V'|$ as $\mathrm{rk}(\mathcal{H}) \leq 1$ and V is Natarajan-shattered by $\mathcal{H}(x)$) and let $\mu \in \Pr(\Omega)$ be the probability template given by letting μ_i be any probability measure on Ω_i for each $i \in \mathbb{N}$ with $i \geq 2$ and letting

$$\mu_1 \stackrel{\mathrm{def}}{=} \frac{1}{k} \left(\nu_{V'} + \sum_{j=1}^{k-1} \delta_{x_{\{j\}}} \right),$$

where $\nu_{V'}$ is the uniform probability measure on V' and δ_t is the Dirac delta concentrated on t.

Since V' is Natarajan-shattered by $\{H^*_k(x, x', -) \mid H \in \mathcal{H}\}$, there exist functions $f_0, f_1 \colon V' \to \Lambda^{S_k}$ with $f_0(v) \neq f_1(v)$ for every $v \in V'$ and there exists a family $\{F_U \mid U \subseteq V'\} \subseteq \mathcal{H}$ such that for every $U \subseteq V'$ and every $v \in V'$, we have $(F_U)^*_k(x, x', v) = f_{\mathbb{1}[v \in U]}(v)$.

Recall that for $H_1, H_2 \in \mathcal{H}$, Lemma A.13 defines $D(H_1, H_2) \stackrel{\mathrm{def}}{=} \{x \in \mathcal{E}_k(\Omega) \mid (H_1)^*_k(x) \neq (H_2)^*_k(x)\}$. Let us also define

$$D'(H_1, H_2) \stackrel{\mathrm{def}}{=} \{v \in V' \mid (H_1)^*_k(x, x', v) \neq (H_2)^*_k(x, x', v)\}.$$

Clearly, for every $U, U' \subseteq V'$, we have $D'(F_U, F_{U'}) = U \triangle U'$.

Note that

$$\begin{aligned}\mu^k(D(H_1,H_2)) &\geq \mathbb{P}_{\boldsymbol{z}\sim\mu^k}\left[\exists\sigma\in S_k, \forall j\in[k-1], \boldsymbol{z}_{\{\sigma(j)\}} = x_{\{j\}} \wedge \boldsymbol{z}_{\{\sigma(k)\}} \in D'(H_1,H_2)\right]\cdot s(\ell)\\ &\geq \frac{k!}{k^k\cdot n}\cdot|D'(H_1,H_2)|.\end{aligned} \tag{A.9}$$

(This is true even if there are repetitions among $x_{\{1\}},\dots,x_{\{k-1\}}$ and even if V' contains some of these elements.)

Using now our definition of m via Haussler packing property, we know that there exists $\mathcal{H}' \subseteq \mathcal{H}$ such that $|\mathcal{H}'| \leq m$ and for every $U \subseteq V'$, there exists $H \in \mathcal{H}'$ such that $L_{\mu,F_U,\ell}(H) \leq \varepsilon$.

For each $H \in \mathcal{H}'$, let

$$\begin{aligned}U_H &\stackrel{\text{def}}{=} \{v \in V' \mid H(x,x',v) = f_1(v)\},\\ B(H) &\stackrel{\text{def}}{=} \{U \subseteq V' \mid L_{\mu,F_U,\ell}(H) \leq \varepsilon\},\\ B'(H) &\stackrel{\text{def}}{=} \left\{U \subseteq V' \;\middle|\; |U \triangle U_H| \leq \frac{\varepsilon\cdot k^k\cdot n}{s(\ell)\cdot k!}\right\}.\end{aligned}$$

The Haussler packing property assumption implies

$$\bigcup_{H\in\mathcal{H}'} B(H) = 2^{V'}. \tag{A.10}$$

Since ℓ is separated, by Lemma A.13(i), for every $H \in \mathcal{H}'$ and every $U \subseteq V'$, we have

$$\begin{aligned}L_{\mu,F_U,\ell}(H) &\geq s(\ell)\cdot\mu^k(D(F_U,H)) \geq \frac{s(\ell)\cdot k!}{k^k\cdot n}\cdot|D'(F_U,H)|\\ &\geq \frac{s(\ell)\cdot k!}{k^k\cdot n}\cdot|D'(F_U,F_{U_H})| = \frac{s(\ell)\cdot k!}{k^k\cdot n}\cdot|U\triangle U_H|,\end{aligned}$$

where the second inequality follows from (A.9), hence $B(H) \subseteq B'(H)$, which together with (A.10) implies $\bigcup_{H\in\mathcal{H}'} B'(H) = 2^{V'}$.

Since $\varepsilon\cdot k^k/(s(\ell)\cdot k!) < 1/2$, by Lemma 5.2, we get

$$n \leq \frac{\log_2|\mathcal{H}'|}{1-h_2(\varepsilon\cdot k^k\cdot n/(s(\ell)\cdot k!))} \leq \frac{\log_2 m}{1-h_2(\varepsilon\cdot k^k\cdot n/(s(\ell)\cdot k!))},$$

which yields $n \leq d$ as n is an integer. □

B The partization operation

In this section, we recall the definition of the partization operation from [CM24, Definition 4.20] and we prove that the partite Haussler packing property of $\mathcal{H}^{k\text{-part}}$ implies the non-partite Haussler packing property of $\mathcal{H}$. Similarly to Section A, numbers in square brackets in the definition below refer to the exact location of the concept in [CM24].

Definition B.1 (Partization [4.20])**.** Let $k \in \mathbb{N}_+$, let Ω be a Borel template and let Λ be a non-empty Borel space.

1. [4.20.1] The *k-partite version* of Ω is the Borel k-partite template $\Omega^{k\text{-part}}$ given by $\Omega^{k\text{-part}}_A \stackrel{\text{def}}{=} \Omega_{|A|}$ $(A \in r(k))$.

2. [4.20.2] For $\mu \in \Pr(\Omega)$, the *k-partite version* of μ is $\mu^{k\text{-part}} \in \Pr(\Omega^{k\text{-part}})$ given by $\mu^{k\text{-part}}_A \stackrel{\text{def}}{=} \mu_{|A|}$ $(A \in r(k))$.

3. [4.20.3] For a hypothesis $F \in \mathcal{F}_k(\Omega, \Lambda)$, the *k-partite version* of F is the k-partite hypothesis $F^{k\text{-part}} \in \mathcal{F}_k(\Omega^{k\text{-part}}, \Lambda^{S_k})$ given by

$$F^{k\text{-part}}(x) \stackrel{\text{def}}{=} F^*_k\big(\iota_{k\text{-part}}(x)\big) \qquad \big(x \in \mathcal{E}_1(\Omega^{k\text{-part}})\big),$$

where $\iota_{k\text{-part}} \colon \mathcal{E}_1(\Omega^{k\text{-part}}) \to \mathcal{E}_k(\Omega)$ is given by

$$\iota_{k\text{-part}}(x)_A \stackrel{\text{def}}{=} x_{1^A} \qquad \big(x \in \mathcal{E}_1(\Omega^{k\text{-part}}), A \in r(k)\big) \tag{B.1}$$

and $1^A \in r_k(1)$ is the unique function $A \to [1]$.

4. [4.20.4] For a hypothesis class $\mathcal{H} \subseteq \mathcal{F}_k(\Omega, \Lambda)$, the *k-partite version* of $\mathcal{H}$ is $\mathcal{H}^{k\text{-part}} \stackrel{\text{def}}{=} \{H^{k\text{-part}} \mid H \in \mathcal{H}\}$, equipped with the pushforward σ-algebra of the one of $\mathcal{H}$. In [CM24, Lemma 8.1] (see Lemma B.2 below), it is shown that $\iota_{k\text{-part}}$ is a Borel-isomorphism, which in turn implies that $\mathcal{H} \ni F \mapsto F^{k\text{-part}} \in \mathcal{H}^{k\text{-part}}$ is a bijection (so singletons of $\mathcal{H}^{k\text{-part}}$ are indeed measurable) and that $\mathcal{H} \mapsto \mathcal{H}^{k\text{-part}}$ is an injection. We denote by $\mathcal{H}^{k\text{-part}} \ni G \mapsto G^{k\text{-part},-1} \in \mathcal{H}$ the inverse of $\mathcal{H} \ni F \mapsto F^{k\text{-part}} \in \mathcal{H}^{k\text{-part}}$.

5. [4.20.5] For a k-ary loss function ℓ over Λ, the *k-partite version* of ℓ is $\ell^{k\text{-part}} \colon \mathcal{E}_1(\Omega^{k\text{-part}}) \times \Lambda^{S_k} \times \Lambda^{S_k} \to \mathbb{R}_{\geq 0}$ given by

$$\ell^{k\text{-part}}(x, y, y') \stackrel{\text{def}}{=} \ell\big(\iota_{kpart}(x), y, y'\big) \qquad \big(\mathcal{E}_1(\Omega^{k\text{-part}}), y, y' \in \Lambda^{S_k}\big).$$

Lemma B.2 (Partization basics [CM24, Lemma 8.1])**.** *Let Ω be a Borel template, let $k \in \mathbb{N}_+$, let Λ be a non-empty Borel space. Then the following hold:*

i. For $\mu \in \Pr(\Omega)$ and $m \in \mathbb{N}$ the function $\phi_m \colon \mathcal{E}_m(\Omega) \to \mathcal{E}_{\lfloor m/k \rfloor}(\Omega^{k\text{-part}})$ given by

$$\phi_m(x)_f \stackrel{\text{def}}{=} x_{\{(i-1)\lfloor m/k \rfloor + f(i) \mid i \in \mathrm{dom}(f)\}} \qquad \big(f \in r_k\left(\lfloor m/k \rfloor\right)\big). \tag{B.2}$$

is measure-preserving with respect to μ^m and $(\mu^{k\text{-part}})^{\lfloor m/k \rfloor}$. Furthermore, if m is divisible by k, then ϕ_m is a measure-isomorphism.

Moreover, we have $\phi_k^{-1} = \iota_{k\text{-part}}$, where $\iota_{k\text{-part}}$ is given by (B.1).

ii. For $m \in \mathbb{N}$, $F \in \mathcal{F}_k(\Omega, \Lambda)$ and $\Phi_m \colon \Lambda^{([m])_k} \to (\Lambda^{S_k})^{[\lfloor m/k \rfloor]^k}$ given by

$$(\Phi_m(y)_\alpha)_\tau \stackrel{\text{def}}{=} y_{\beta_\alpha \circ \tau} \qquad \Big(\alpha \in [\lfloor m/k \rfloor]^k, \tau \in S_k\Big), \tag{B.3}$$

where $\beta_\alpha \in ([m])_k$ is given by

$$\beta_\alpha(i) \stackrel{\text{def}}{=} (i-1)\left\lfloor \frac{m}{k} \right\rfloor + \alpha(i) \qquad \left(\alpha \in \left[\left\lfloor \frac{m}{k} \right\rfloor\right]^k, i \in [k]\right), \tag{B.4}$$

the diagram

$$\begin{array}{ccc} \mathcal{E}_m(\Omega) & \xrightarrow{\quad F_m^* \quad} & \Lambda^{([m])_k} \\ \Big\downarrow{\scriptstyle \phi_m} & & \Big\downarrow{\scriptstyle \Phi_m} \\ \mathcal{E}_{\lfloor m/k \rfloor}(\Omega^{k\text{-part}}) & \xrightarrow{(F^{k\text{-part}})^*_{\lfloor m/k \rfloor}} & (\Lambda^{S_k})^{[\lfloor m/k \rfloor]^k} \end{array}$$

commutes, where ϕ_m is given by (B.2)*.*

The following lemma is equation (8.7) within the proof of [CM24, Proposition 8.4]. For completeness purposes, we restate this below with a self-contained proof.

Lemma B.3. *Let Ω be a Borel template, let $k \in \mathbb{N}_+$, let Λ be a non-empty Borel space, let $\ell\colon \mathcal{E}_k(\Omega) \times \Lambda^{S_k} \times \Lambda^{S_k} \to \mathbb{R}_{\geq 0}$ be a k-ary loss function and let $F, H \in \mathcal{F}_k(\Omega, \Lambda)$ be hypotheses. Then*

$$L_{\mu,F,\ell}(H) = L_{\mu^{k\text{-part}},F^{k\text{-part}},\ell^{k\text{-part}}}(H^{k\text{-part}}).$$

Proof. This follows directly from

$$\begin{aligned} L_{\mu,F,\ell}(H) &= \mathbb{E}_{\boldsymbol{x}\sim\mu^k}[\ell(\boldsymbol{x}, H_k^*(\boldsymbol{x}), F_k^*(\boldsymbol{x}))] \\ &= \mathbb{E}_{\boldsymbol{x}\sim\mu^k}[\ell^{k\text{-part}}(\phi_k(\boldsymbol{x}), H^{k\text{-part}}(\phi_k(\boldsymbol{x})), F^{k\text{-part}}(\phi_k(\boldsymbol{x})))] \\ &= \mathbb{E}_{\boldsymbol{z}\sim(\mu^{k\text{-part}})^1}[\ell^{k\text{-part}}(\boldsymbol{z}, H^{k\text{-part}}(\boldsymbol{z}), F^{k\text{-part}}(\boldsymbol{z}))] \\ &= L_{\mu^{k\text{-part}},F^{k\text{-part}},\ell^{k\text{-part}}}(H^{k\text{-part}}), \end{aligned}$$

where the second equality follows from the definition of $\ell^{k\text{-part}}$ and Lemma B.2(ii) and the third equality follows since ϕ_m is measure-preserving by Lemma B.2(i). □

Theorem B.4 (Haussler property: partite to non-partite)**.** *Let Ω be a Borel template, let $k \in \mathbb{N}_+$, let Λ be a non-empty Borel space, let $\mathcal{H} \subseteq \mathcal{F}_k(\Omega, \Lambda)$ be a k-ary hypothesis class, let $\ell\colon \mathcal{E}_k(\Omega) \times \Lambda^{S_k} \times \Lambda^{S_k} \to \mathbb{R}_{\geq 0}$ be a k-ary loss function.*

If $\mathcal{H}^{k\text{-part}}$ has the Haussler packing property with respect to $\ell^{k\text{-part}}$, then $\mathcal{H}$ has the Haussler packing property with respect to ℓ with associated function $m^{\mathrm{HP}}_{\mathcal{H},\ell} \stackrel{\mathrm{def}}{=} m^{\mathrm{HP}}_{\mathcal{H}^{k\text{-part}},\ell^{k\text{-part}}}$.

Proof. For $\varepsilon \in (0,1)$ and $\mu \in \Pr(\Omega)$, we know that there exists $\widetilde{\mathcal{H}} \subseteq \mathcal{H}^{k\text{-part}}$ with $|\widetilde{\mathcal{H}}| \leq m^{\mathrm{HP}}_{\mathcal{H}^{k\text{-part}},\ell^{k\text{-part}}}(\varepsilon)$ such that

$$\forall F \in \mathcal{H}^{k\text{-part}}, \exists H \in \widetilde{\mathcal{H}}, L_{\mu^{k\text{-part}},F,\ell^{k\text{-part}}}(H) \leq \varepsilon$$

Letting $\mathcal{H}' \stackrel{\mathrm{def}}{=} \widetilde{\mathcal{H}}^{k\text{-part},-1}$, the above implies

$$\forall F \in \mathcal{H}, \exists H \in \mathcal{H}', L_{\mu^{k\text{-part}},F^{k\text{-part}},\ell^{k\text{-part}}}(H^{k\text{-part}}) \leq \varepsilon,$$

which by Lemma B.3 yields

$$\forall F \in \mathcal{H}, \exists H \in \mathcal{H}', L_{\mu,F,\ell}(H) \leq \varepsilon,$$

as desired. □