Expressiveness as Decomposable Default-VC

Morgan Bryant
mrbryant@stanford.edu
CS 257 / Phil 356C W.2017 Final Project

Introduction In statistical and computational learning theory, there is the general question of assigning complexity to models and datasets, both in terms of what metric to use and in terms of how to tractably do the assigning. Well-known solutions are Information Entropy H from Shannon, Minimum Description Length, Bayesian and Akaike Information Criterions, effective number of parameters θ for parametric models, and Vapnik-Chervonenkis (VC) dimension. Each of these have their effective uses, but each of them are limited in two realms:

- Each is only capable of examining entire systems as indivisible units; this idea is often unintuitive, since models are often designed and understood as complex objects, exemplified by **Task Three** in Figure 1. This signals the primary motivation for this project: to develop a version of complexity that *decomposes* naturally as tasks and models decompose.
- Each assigns a single ordinal number to objects: $H(X) \mapsto [0, 1]$, $VC(X) \in \mathbb{N}$, etc. While this is effective for comparing two objects and each statistic is designed to be used in theoretically-useful settings, these lack an ability to intuitively categorize, in the same setting, the below **Tasks One** and **Two** in Figure 1. Some of the methods listed below, such as the plausibility structure, generalize this capability, yielding more understandable tools for decomposed modelings.

This project makes an attempt to resolve this issue in a restricted domain. Please note that for the sake of brevity and clarity, many inessential formal definitions and proofs are omitted; please consult the references for completions.

1 Overview

This project examines VC dimension as a metric and proposes methods to generalize or augment it for the purpose of (1) decomposable structures and (2) more flexible interpretation. Using logic and notions of probability, I develop several ways to realize this task.

Figure 1: All binary tasks on two atoms, which we call class BIN. Let **Task One** be a linearly seperable dataset in $\mathbb{R}^2$, such as all the tasks below besides those with ID 6 or ID 9. **Task Two** is a binary XOR task in $\mathbb{R}^2$, such as tasks ID 6 or ID 9 below. **Task Three** is a task in $\mathbb{R}^4$ in which all points in the first two dimensions are linearly seperable and the points in the last two dimensions is a binary XOR task, such as direct product (4) $\times$ (6).

A	B	0	$\wedge$	$A \wedge \neg B$	A	$B \wedge \neg A$	B	$\otimes$	$\vee$	$\neg(A \wedge B)$	$\leftrightarrow$	$\neg B$	$\neg A$	$B \rightarrow A$	$A \rightarrow B$	$\neg(A \vee B)$	1
0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
0	1	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
1	0	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
1	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
	ID	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

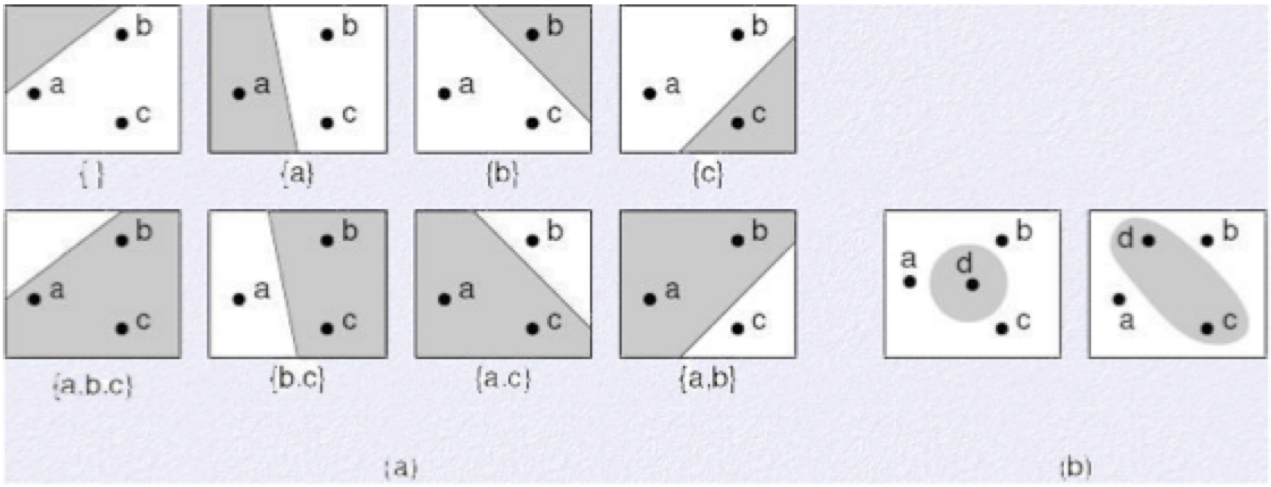
VC Dimension Definition While VC is usually defined on families of functions, it will be more convenient for us to work in a set-theoretic framework, according to [1]. Let us define $\mathcal{N}_{\mathcal{A}}(D) = |\{A \cap D : A \in \mathcal{A}\}|$ be the total number of subsets that can be 'picked out' by elements A of $\mathcal{A}$ on a finite set of x, $\{x\} = D$. Define the n^{th} shatter coefficient of family $\mathcal{A}$ as $r(\mathcal{A}, n) = \max_D \mathcal{N}_{\mathcal{A}}(D)$; note $r(\mathcal{A}, n) \leq 2^n, \forall n$. If $r(\mathcal{A}, n) = 2^n$, then there exists a set of points D such that $\mathcal{N}_{\mathcal{A}}(D) = 2^n$, which we call $\mathcal{A}$ shatters D. Let $VC(\mathcal{A})$ be the largest $k \in \mathbb{N}$ such that $r(\mathcal{A}, k) = 2^k$, which can be interpreted as the largest number of points that can be shattered by $\mathcal{A}$. This all a standard definition of VC.

(Note that VC is a unary metric on a model class, implicitly defining which tasks it can solve, namely those with VC(class) number of points. However, since this project is concerned with the class-task *relative fitness* as the standard for expressiveness, I consider the corresponding mapped binary *shatter* metric that takes a class and a task and returns whether the VC-power of the class satisfies the VC-requirement of the task.)

Informally, figure (2) is the canonical VC example: linear classifiers in $\mathbb{R}^2$ with bias terms ($LC2$) can shatter up to three points in any orientation with any classification; hence, model class $LC2$ has VC dimension at least 3. However, there is no classifier in $LC2$ that can completely classify any four points a vector space – that is, versions of $\otimes$ and $\leftrightarrow$ are not shattered; hence, $VC(LC2) < 4$.

One of the simplest motivations for this project is that $LC2$ can shatter *most* of any 4 points but not all – specifically, given a uniform distribution of four points (or two binary atoms) and all logical assignments on those atoms, the probability of the existence of a shattering is $\frac{14}{16} = 0.875$. It seems intuitively harsh to assign $r(LC2, 4) = \bot = 0$. As the number of points and the size of

Figure 2: Linear classifiers in $\mathbb{R}^2$ can shatter any 3 points (a) but not all 4 points (b).



the classifiers grow, the proportion of points and arrangements that linear classifiers can shatter grows exponentially but the shattering dimension is bounded above linearly due to generalized XOR examples. One of the motivating points of this project is to develop a framework in which a notion of partial shattering is founded. Specifically, in the application of fitted models and data-driven classification tasks, it makes sense not to evaluate $\perp$ on $\mathcal{A}$ if the likelihood of its incapacity is exceedingly small on $\mathbb{E}[D]$.

The primary mode of reasoning, which aligns with our coursework, is the usage of *default reasoning* as an in-a-normal-world approach to reimagining VC dimension. Default reasoning makes sense both semantically and procedurally, since it lends to probabilistic and in-limit interpretations. The following sections outline various methods for developing and using this so-called **default-VC**.

- The first section *Extremal Statistics* outlines the overall problem and provides the strictest but most exact encoding of default-VC, using probabilities on logic. It both preserves the idea of typically-shatters and implicitly lends to decomposability. A similar system called ϵ-entailment by Pearl and Adams is presented.

- The next section *PAC-bound* introduces probabilistic approximation to the strict ϵ-entailment. via a reduction of the default-VC to a PAC-satisfying objective. An example shows its versatility.

- Next is a method that measures disjoint covering sets as a different kind of method for determining model sufficiency given a task.

- After that is a plausibility structure, a probability generalization with well-suited proper-

ties.

- And finally, a convergent sequence strategy.

Some definitions: Let us define shatter function $\mathfrak{s}_{\mathcal{A}}(D)$ (analogous to r above) to be a truth valuation indicating whether $\mathcal{A}$ shatters D or not for set of sets (family of functions, class of models) $\mathcal{A}$ and an ordered set (partially or fully observed data sets, a single task, a set of tasks, a vector, a pair (x,y)) called D. $\mathfrak{s}$ maps to [0,1] $\equiv [\top, \bot]$. Further, let us extend $D \to \mathfrak{D}$ be an ordered countably infinite set (infinite vector) and A be a simple set or specific model in $\mathcal{A} : A \in \mathcal{A}$. In this project, we consider the rough terminology: $x \in D \subseteq \mathfrak{D}$, where x is a single point, D is a sample, and $\mathfrak{D}$ is the unknown target task or true distribution; and $A \in \mathcal{A} \subseteq \mathcal{M}_{\underline{\mathcal{A}}}$, where A is a single unparameterized or parameterized model, $\mathcal{A}$ is a 'model class' such as $LC2$, and $\mathcal{M}_{\underline{\mathcal{A}}}$ is a collection of model classes, such as $\{LCn\}$ for $n = 1..N$. Note that a 'task' x can have distinct observation values and target values, such that an appropriate model finds an f such that $f(x_i) = x_o$ and $x = (x_i, x_o)$. However, this structure can be isomorphically mapped to a point in a single simple variable space: for example, the original VC task that considers subsets of points in $\mathbb{R}^n$ can be remapped to $\mathbb{R}^n \times I$ where I is an indicator of set inclusion, $I = \{0, 1\}$. Likewise, while a 'true distribution' $\mathfrak{D}$ might be pairs of measure values assigned to events, the measure can be considered as an appended dimension on the event space; hence, we can consider any task $x \in \mathfrak{D}$ as a simple, 'atomic' value or vector. This is, in part, the reason we can consider a set-theoretic definition of VC dimension instead of a functional one.

2 Extremal Statistics

This is perhaps the most direct method for developing a fuzzy VC. In the words of Pearl, [8], this is a "conservative core".Essentially, using results in default logic, we develop a reasonable interpretation of VC in probabilistic terms derived from logic.

Preliminaries Define $\models_{\mathfrak{s}}$ as follows. We can say that $\mathbb{P}(\psi|\phi) > 1 - \epsilon$ iff we can know a finite collection Δ of system-P-defined default statements $\{x \ \mid\sim \ y\}_t$ (ie, a sequence of length t of default statements $x_t \ \mid\sim \ y_t$), such that the a collected assertion for all ϵ there is δ for which $\mathbb{P}(y_t|x_t) > 1 - \delta$ implies $\mathbb{P}(\psi|\phi) > 1 - \epsilon$, then $\{x \ \mid\sim \ y\}_t \models_{\mathfrak{s}} \phi \ \mid\sim \ \psi$. (Equivalently, define $\models_{\mathfrak{s}}$ as: given $\mathbb{P}(\psi|\phi) > 1 - \epsilon$, $\{x \ \mid\sim \ y\}_t \models_{\mathfrak{s}} \phi \ \mid\sim \ \psi$ as $t \to \infty$ if $\{x \ \mid\sim \ y\}_t$ can verify $\mathbb{P}(\psi|\phi) > 1 - \epsilon$ almost surely.) That is, if we are assured that any sequence $\{x \ \mid\sim \ y\}_t$ would $\mathfrak{s}$-entail $\phi \ \mid\sim \ \psi$ by an arbitrary bound *for every* $\mathbb{P}$, then $\top \models_{\mathfrak{s}} \phi \ \mid\sim \ \psi$. This we have seen in class. Let $\mathbb{P}(\psi|\phi) \equiv \mathbb{P}(\mathfrak{s}_{\phi}(\psi))$, ie the probability that ϕ shatters ψ.

Now, if we can assert such a claim for all $\mathbb{P}$, then we can demonstrate that a sequence of selected models $\mathcal{A} \subseteq \mathcal{M}_{\underline{\mathcal{A}}}$ of a collection of model classes (such as: $\mathcal{M}_{\underline{\mathcal{A}}}$ = all linear classifiers

in $\mathbb{R}^m$ and $\mathcal{A} = LC2$). We are attempting to find a collection of models that will almost surely shatter $\mathfrak{D}$ a true distribution by considering samples $D_t \subset \mathfrak{D}$: that is, we can talk about $\mathbb{P}(\mathfrak{D}|\mathcal{M}_{\underline{\mathcal{A}}})$ given assertions $\mathcal{A}_t \mid\sim D_t$.

Using probabilities on first-order as in lecture using Gaifman's conditions, we can reduce our problem to the following which will guarantee over all $\mathbb{P}$: we seek:

$$1 - \epsilon < \mathbb{P}(\forall D \exists \mathcal{A} : \mathfrak{s}_{\mathcal{A}}(D)) \text{ for } D \subset \mathfrak{D}, \mathcal{A} \subseteq \mathcal{M}_{\underline{\mathcal{A}}} \tag{1}$$

$$= \inf_{\{D\}} \mathbb{P}(\exists \mathcal{A} : \wedge_D \mathfrak{s}_{\mathcal{A}}(D)) \tag{2}$$

$$= \inf_{\{D\}} \sup_{\{A\}} \mathbb{P}(\vee_{\{A\}} \wedge_{\{D\}} \mathfrak{s}_{\mathcal{A}}(D)) \tag{3}$$

is reached using temporary formula with only one free variable $\xi(D) = (\exists \mathcal{A} : \mathfrak{s}_{\mathcal{A}}(D))$ and quantifier identities, over all finite sequences of $\{D\}$ and $\{\mathcal{A}\}$.

This is a natural method for defining our goal: we have essentially formulated the task as a game in which D attempts to minimize and $\mathcal{A}$ attempts to maximize the shatter valuation. If there is always an $\mathcal{A}$ that can be found that can shatter any sample D, then we can conclude $\mathfrak{s}_{\mathcal{M}_{\underline{\mathcal{A}}}}(\mathfrak{D}) \to \top$. Note that the strictness of this assertion comes with the cost of having to assert infimums over any D for shattering; alternatively, we can approximate $\mathfrak{s}_{\mathcal{M}_{\underline{\mathcal{A}}}}$ tractably using PAC-bounds, as discussed in section *Entropy Methods*. Additionally, this formulation insists on verifying these sequences over every well-defined $\mathbb{P}$; a resolution is presented below as well.

A point of note is that $\Delta \vDash_{\mathfrak{s}} \mathcal{M}_{\underline{\mathcal{A}}} \mid\sim \mathfrak{D}$ with $\Delta = \{\mathcal{A}^t \mid\sim D^t, t \in 0..T\}$ is a fully self-referential system: any $\mathcal{A}^t \mid\sim D^t$ can be itself be $\mathfrak{s}$-entailed as $\Delta^t \vDash_{\mathfrak{s}} \mathcal{A}^t \mid\sim D^t$, with $\Delta^t = \{\mathcal{A}^{tu} \mid\sim D^{tu}$, $u = 0..U\}$ for recursion layer 2; $\{\mathcal{A}^{tuv} \mid\sim D^{tuv}, v = 0..V\} \vDash_{\mathfrak{s}} \mathcal{A}^{tu} \mid\sim D^{tu}$ et cetera via induction. In English (informally), a collection of constituent tasks that are independently modeled and can assert an umbrella task is a fully recursive system. In this setting, standard VC can be interpreted as one of these systems that was recursed all the way down until $D = \{x\}, |D| = 1$ a single datapoint. The shatter valuation of each lone datapoint was determined by the complete set of model classes $\mathcal{M}_{\underline{\mathcal{A}}}$, and the infimum/supremum operation enforced binary returns. For example, $\mathfrak{D} = BIN$ and $\mathcal{M}_{\underline{\mathcal{A}}} = LC2$ would recurse down to $D^{t_1,\ldots,t_r}$ is the unary 'datapoint', one of tasks 0-15 in Figure (1). If $D^{t_1,\ldots,t_r} = \{\wedge\}$, for example, $LC2^{t_1,\ldots,t_r} == LC2$ would shatter $\wedge$ and return 1 up the stack to $\mathfrak{D}$. However, recursion $D^{t_1,\ldots,t_r} = \{\otimes\}$ would not be shattered by $LC2$ and would return 0; up at the root, the infimum would send the entire valuation to 0.

In default-VC, the desire for decomposability is pleased by natural recursion and division of tasks, but it often makes sense to stop recursion early at some point if a value beyond standard binary VC shatter would be helpful.

For clarity, The following is an example of an approximate default-VC. Consider $\mathfrak{D} = BIN$ and $\mathcal{M}_{\underline{\mathcal{A}}} = LC2$. To determine the *probability* that $LC2$ would shatter an arbitrary binary task

on 2 variables, sample a x task (one of the 16 binary tasks) M times, determining if $LC2$ shatters x, and taking the average of M runs. In the limit as $M \to \infty$, the probability of $\mathfrak{s}$ would approach 0.875 indeed. This process could include one recursion: first, sort the task by the value of atom A (which has $\mathbb{P}(A = \bot) = \mathbb{P}(A = \top) = 0.5$ as a prior), then multiply that by the probability of $\mathfrak{s}$ given the rest of the sample, the value of atom B. This is loosely: $\mathfrak{s}_{\mathcal{M}_{\underline{\mathcal{A}}}}(\mathfrak{s}(\mathfrak{D}_B|A) * \mathbb{P}(A))$.

An important point is that the above definition of probability on first-order logic statements is only valid if it can be asserted for *all* $\mathbb{P}$. This, however, is easier to satisfy that would be immediately apparent; see the section *measures of disjoint covering sets*.

3 PAC-bound

Another useful tool that utilizes entropy methods is the probably approximately correct (PAC) bound [9,11]. This comes particularly in handy when our candidate models $\hat{A}$ as above are fit to surpass a fixed but substantive threshold of error. Consider:

$$\mathbb{P}(\{D : KL(\mathfrak{s}_{\mathcal{M}_{\underline{\mathcal{A}}}}(\mathfrak{D})||\mathfrak{s}_{\mathcal{M}_{\underline{\mathcal{A}}}}(D)) \leq \eta\}) \geq 1 - \epsilon$$

using approximation parameter $\eta > 0$, $\epsilon > 0$ as a confidence parameter, KL as the Kullback-Leibler divergence or relative entropy, and $\mathfrak{s}$ as defined in section *Extremal Statistics*. If $\mathbb{P}$ satisfies this bound, it is called PAC and there is a theory of results that support it as a qualification for statistical learnability [11]. If this $\mathbb{P}$ can be verified over a finite sample of $\{D\}$ of size n, using arbitrary n, then our original task is asserted. For modeling tasks, this can often be asserted if the error individual models or the expected error over models can be found.

Applied to our problem, let us maintain $\mathfrak{s} \mapsto \{\bot, \top\}$ and define $\bar{\mathfrak{s}}(D) := \frac{1}{d}\sum_D \mathfrak{s}(D)$ is the sample mean of shatters in [0,1]. Let $d := |\{D\}| < \mathbb{N}$. Then, using Chernoff's bound:

$$\hat{\mathbb{P}}_{\{D\}\sim\mathfrak{D}}(\mathfrak{s}(D) \neq \top) = \mathbb{P}_{\mathfrak{D}}(\bar{\mathfrak{s}}(\mathfrak{D}) \leq 1 - \epsilon) \leq e^{-dO(\epsilon^2)} \tag{4}$$

under fixed $\epsilon > 0$, $\hat{\mathbb{P}}$ an unbiased estimate or an estimate as $d \to \infty$, and all samples $D \sim \mathfrak{D}$ iid. This bound can be interpreted as: the probability that a collection of learned models (with known bounded shatterings) would overestimate the overall shattering of the dataset $\mathfrak{D}$ decreases exponentially fast in the number of samples modeled d. That is, the shatter valuation $\mathfrak{s}$ built for our question $\mathcal{M}_{\underline{\mathcal{A}}} \mid\sim \mathfrak{D}$ using only finite samples satisfies PAC learnability, around which there is substantial literature. A particularly satisfying result from [10] explains that tasks that can be modeled with decision trees over k-ary CNF clauses on n atoms ($n > k$) (which we call the model class kDT) are PAC-bound preserving. Additionally, kDT is distribution-free, hence, approximately effective for all $\mathfrak{D}$. That is, kDT is a class of models that can preserve our $\mathfrak{s}$ relationship on independently-modeled data samples $D \sim \mathfrak{D}$; further, the fact that kDT is easily boostable [2] lends to implementational tractability.

What this section shows is that adding the probabilistic statistic of (relative) entropy allows us to operate uncertainly. Extremal statistics gave us a universal 'game' in which we reason about first order formulas on $\mathfrak{s}$, and it used infimums and supremums on sampled fittings of models to derive universal declarations about the true task. The strategy in this section instead uses the average $\mathfrak{s}$ over the sampled fits instead of inf/sup, and such an interpretation lends to a system in which we can assert global statistical bounds – but without the strict assurance that extremal statistics yields. Additionally, PAC works under the assumptions of iid, which extremal statistics does not: that is, when PAC is implemented, the target task is no longer further decomposable.

ϵ-entailment Pearl in [8] describes our *Extremal Statistics* scheme as the 'conservative core' of probabilistic default logic, called ϵ-entailment using the standard system-P default $\mid\!\sim$ (and attributes the original conception to Adams). [15] shows that ϵ-entailment is equivalent to system-P as a default logic system. Notably, he formalizes this version under fixed distributions that preserve $\mathbb{P}(y|x) \geq 1 - \epsilon$ for $x \mid\!\sim y \in \Delta$: only these $\mathbb{P}$ are the ones which must satisfy $\mathbb{P}(\psi|\phi) = 1 - O(\epsilon)$. Because this aligns with our above PAC results, we can confidently say that PAC well-represents the system-P default declarations. What's significant is that, as in CS 257 lecture slides on 2-14-17, slide 21 [3], the system-P semantics on default statements can be used to derive exact relationships between ϵ and δ. Generalized, this means that given confidence requirements for $\mathbb{P}(\mathfrak{s}_{\mathcal{M}_{\underline{\mathcal{A}}}}(\mathfrak{D})) > 1 - \epsilon$, we can determine exact requisite probabilities on the decomposed partitions, $\mathbb{P}(\mathfrak{s}_{\mathcal{A}}(D)) > 1 - \delta$. According to [17], we can often conclude that $\delta = \frac{\epsilon}{|D|}$ is sufficient.

This supports how our above *extremal statistics* can be demonstrated as consistent with default tenets: given $\delta := \epsilon/|D|$, the system-P rules of AND and OR deduces, for arbitrary formula:

$$\frac{X^* \mid\!\sim Y^*}{(\underset{x \in X}{\vee} x) \mid\!\sim (\underset{y \in Y}{\wedge} y)}$$

which is sensible if we consider all x_1, x_2, y_1, y_2 independent from each other as not intersecting. This observation assists in understanding why ordinal $\mathbb{P}$ is somewhat limited where y are logical covers of Y^* (ie, $\overline{Y} \supseteq \overline{Y^*}$) and $X \supset X^*$. The ability to apply system-P semantics to default-VC suggests an abstraction of $\mathfrak{s}$ from distinct models and tasks to a generalized $\mathbb{P}$-space mapped from arbitrary objects, a point that is expanded on starting in section *Plausibility*.

Finally, the ϵ-entailment definition lets us present Δ as a directed network of default statements involving possibly intersecting sets of atoms, a result that preserves natural intuitions of default statements [8]. Such a representation is natural for both decomposition of the parent task $\mathfrak{D}$ into $\{D\}$ and lends to calculating the relationship between ϵ and δ.

4 Disjoint Cover

The Disjoint Cover strategy of asserting $\mathcal{M}_{\underline{\mathcal{A}}} \mid\sim \mathfrak{D}$ considers $\mathcal{M}_{\underline{\mathcal{A}}}$ and $\mathfrak{D}$ as sets whose properties may be intractable to understand directly but have subsets that are tractably understandable. One approach begins by considering a target set S (eg, either $\mathcal{M}_{\underline{\mathcal{A}}}$ or $\mathfrak{D}$) as a bounded set. Let the Haussdorf measure μ have $\mu(S) = m$, $\mu(\emptyset) = 0$, and $0 \leq_\mu \mu(s) \leq_\mu m$ for any $s \subseteq S$. That is, $\emptyset \subseteq s \subseteq S$ are formally *measurable* and S is a measure space iff $\mu \mapsto [0, m] \subset \mathbb{R}$ is totally ordered. Otherwise, μ defines a weakened finite 'measure' on a partially ordered space. We will use this weakened measure in this section and also later in section *Plausibility*.

If we can assert that S is covered by collection $\{s\}$, then $\mu(\cup s) = \mu(S)$. (If we enforce $|\{s\}|$ is countable, or equivalently if countable additivity holds, then a $\mu \mapsto [0, 1]$ is trivially a true probability measure.) Note that $\mu(\cup s) = \mu(S)$ is valid because $\{s\}$ covers S implies we can develop a disjoint subcover as well, and s need not be open because we specified nothing about the size of $\{s\}$. Under additivity, if $\{s\}$ is disjoint, then also $\mu(S) = \sum_s \mu(s)$.

Let us consider a μ defined over $S = \mathfrak{s}$ as a true arbitrary probability measure: $\mu \mapsto [0, m{:=}1]$, given with respect to $\mathfrak{D}$ and $\mathcal{M}_{\underline{\mathcal{A}}}$. With this, we can develop two ways to procedurally verify $\mathcal{M}_{\underline{\mathcal{A}}} \mid\sim \mathfrak{D}$. **Bottom-up** If we can take a collection of disjoint $\{D\}$ that cover $\mathfrak{D}$, then $\Delta = \{\mathcal{M}_{\underline{\mathcal{A}}} \mid\sim D\}$ can validate $\mathcal{M}_{\underline{\mathcal{A}}} \mid\sim \mathfrak{D}$ if $1 = \mu(\mathfrak{D}) = \sum_D \mu(D)$ for some discovered $\{\mathcal{A}\}$. **Top-down** If we take a sequence of covers $\{D\}^t$ such that as $t \to \infty$, $\mu(D_i \cap D_j) \to 0$ for all $i \neq j$ and $1 \leq \sum_{D \in \{D\}^t} \mu(D)$, then we can conclude $\mathcal{M}_{\underline{\mathcal{A}}} \mid\sim \mathfrak{D}$.

Both of these strategies use our measure μ as a mapping of subsections of our data space $\mathfrak{D}$ to metrics μ that determine *how much* of the dataspace is covered by any given D. In a sense, we de-condition our original shatter probability $\mathfrak{s}_{\mathcal{A}}(D)$ from D by multiplying by a $\mu(D) \leq 1$ on our measure defined on $\mathcal{M}_{\underline{\mathcal{A}}}$. The primary reason for this construction is that it permits a weighting of importance to various portions of our dataspace. Furthermore, while the weighting can be selected, the original ϵ-entailment conception can remain constant. That is:

$$\Delta \models_{\mathfrak{s}} \mu(\mathfrak{D}) > 1 - \epsilon \text{ if}$$
$$\sum_{D \in \Delta} \mu(D) > 1 - \epsilon, \ \Delta \text{ disjoint.}$$

This μ can be considered our first example of a system that uses a kind of *shatter mapping* of entailment: by sending $\mathcal{M}_{\underline{\mathcal{A}}}$ and $\mathfrak{D}$ or subsets to measures besides straightforward shatter valuations, we can construct helpful machinery for entailment.

Disjoint set schemes can be realized by the two variants above of bottom-up or top-down via a partitioning of $\mathfrak{D}$ (explicitly or implicitly). The simplest partition is to simply split $\mathfrak{D}$ into d areas and model each area piecewise with a known size of Δ. Alternatively, if $D = \{x\}$ are interpreted

as samples of a distribution $\mathfrak{D}$, then the strategy holds as the number of samples grows and there is confidence in the continuity or openness of the D as a true cover of $\mathfrak{D}$. Other strategies work, so long as $\{D\}$ is a cover of $\mathfrak{D}$ in the limit. In this context, the usage of asymptotic equipartion sets of information theory could be helpful in determining the statistical covering capabilities of a $\{D\}$ or to

Another method related to *plausibility* is to use a μ based on $\mathcal{M}_{\underline{\mathcal{A}}}$ and $\mathfrak{D}$ but where $S \neq \mathfrak{s}$: μ could, say, be the number of data points x shattered and $\mathfrak{D}$ is the total number of data points; μ could, say, map to other mathematical objects whose metrics are more easily manipulable, such as a composed mapping of $\mathcal{M}_{\underline{\mathcal{A}}}$ to the space of tasks that it can solve in x-space, then a measure of whether $\mathcal{M}_{\underline{\mathcal{A}}} \supset D$.

Even more generally, two mappings $\$(\mathfrak{D})$ and $\$(\mathcal{M}_{\underline{\mathcal{A}}})$ can be defined to map both models and tasks to the same shatter space codomain Y, in which the two comappings are compared using μ such as a $\leq$ or overlap metric, etc. An example would be to use a hidden network representation that takes two of either kind of set and returns one of $(\leq, =, \geq, \neq)$ to indicate which of two objects has a higher hidden shatter.

5 Plausibility

The last mapping defined, \$, indicates a final general strategy that involves mappings of models and tasks to new spaces. First, there are two main motivating points that the above methods are not entirely effective at:

- From [14] we have the lemma: $\mathcal{M}_{\underline{\mathcal{A}}} \mathrel{|\!\sim} \mathcal{M}_{\underline{\mathcal{B}}}$ yields the same principles as $\mathcal{M}_{\underline{\mathcal{A}}} \wedge \mathcal{M}_{\underline{\mathcal{B}}}$. This informally means that there is a *set space* that categorizes what models do, what tasks operate in, via cautious monotonicity; the ability to compare two models should be as natural as comparing a model to a task. This means that default-VC systems should naturally be able to compare arbitrary model classes and data tasks. However, the above methods still lack a way to resolve the differences between, for example, the two model classes *LC2* and $BINH$ in terms of what they can shatter. ($BINH \triangleq$ the contrived model class on two binary atoms with nonzero entropy, namely, all tasks except 0 and 15.) Both $BINH$ and *LC2* shatter with probability 0.875, but the fact that they have different domains should be modeled.

- Cumulative models as in [14] suggest a hierarchy of sufficient models that is reminiscient of SRM from Vapnik. That is, there is an 'optimal', minimal model that is most reasonable in a normal world for a task. This aspect of system-P follows with the intuitive idea of nonmonotonicity is insensible for generalized VC. For example, in VC, the function

$\sin(\alpha x)$ has infinite VC dimension because α can always be set arbitrarily low so as to perfectly capture any binary set with range with codomain [-1,1]. But for reasons such as generalizability or continuity of parameters given a small change in data $\sin(\alpha x)$ is often rather not 'optimal' and a model with smaller VC dimension may be stronger.

These two situations signal that a metric as simple as $\mathfrak{s} \mapsto \{\top, \bot\}$ or $\mathbb{P} \mapsto [0, 1]$ may be too simple to well-compare arbitrary models and tasks. Generally, a totally ordered metric may be too strict. As such, we introduce here Plausibility Structures as in [15] which have an optional generalization to partial ordering, as well as a number of other properties such as a rather weak set of constraints on the codomain.

Rather than $\mathbb{P} \mapsto [0, 1]$, define a similar $\$ \mapsto C$ for special a set C that satisfies partial ordered relationship $\leq$ and for any $c \subseteq C$, $\bot \leq c \leq \top$ for special symbols $\bot, \top$: $\$(\emptyset) = \bot$ and $\$(C) = \top$. It can be shown that $\$ is equivalent to system-P if:

- For pairwise disjoint X,Y,Z, $[\$(X \cup Y) > \$(Z)] \wedge [\$(X \cup Z) > \$(Y)] \rightarrow [\$(X) > \$(Y \cup Z)]$
- For any X,Y,Z, $[\$(X \cap Y) > \$(X \cap \bar{Y})] \wedge [\$(X \cap Z) > \$(X \cap \bar{Z})] \rightarrow [\$(X \cap Y \cap Z) > \$(X \cap \overline{(Y \cap Z)})]$
- For any X,Y, $\$(X) = \bot \wedge \$(Y) = \bot \rightarrow \$(X \cup Y) = \bot$

Under these (mild) conditions, we can insert our default shatter entailment constructs with $\$ in place of $\mathbb{P}$; these conditions even give guidance as to how to decompose or deconstruct $\mathfrak{D}$ or $\mathcal{M}_{\underline{\mathcal{A}}}$ into smaller pieces for independent computation.

Plausibility is compelling partly because it has numerous convenient properties. This general structure exactly characterizes our μ measure and is a concrete way to frame the two mappings $\$ presented in section *Disjoint Cover*. From [15], this structure can handle statistical first order formulations in our context with ease due to its generalized set formulation: so, our query $\forall D \subset \mathfrak{D}, \exists \mathcal{A} \subset \mathcal{M}_{\underline{\mathcal{A}}} : \mathfrak{s}_{\mathcal{A}}(D)$ fits quite naturally into plausibility without need of Gaifman's conditions due to a property that if a formula can be satisfied, in can be done with a finite structure [15]. The partial ordering allows the $\sin(\alpha x)$ situation to be unordered compared to more 'reasonable' models yet still made less minimal. If $\mathcal{M}_{\underline{\mathcal{A}}}$ and $\mathfrak{D}$ are defined to not occupy the same space, then there are $\mathfrak{s}_{\mathcal{M}_{\underline{\mathcal{A}}}}(\mathfrak{D}) = \$(\mathcal{M}_{\underline{\mathcal{A}}}) > \$(\mathfrak{D}) \rightarrow \mathbb{P}(\mathfrak{D}|\mathcal{M}_{\underline{\mathcal{A}}}, \mathfrak{D}) \geq 1 - O(\epsilon)$. As in probability, there are analogues to conditioning, independence, and Markovity via notions of set addition and multiplication [16].

Other properties are: a plausibility structure is fully decomposable under addition (disjoint union). Plausibility is more intuitively effective at handling irrelevant formulae (due in part to the capacity for determining optimality structures.

Generally speaking, plausibility provides a framework that is both more powerful than standard probability while also preserving properties of probability that are useful for developing more normal default conceptions of VC shattering dimension. Besides providing a generalized space for shatter valuations and comparisons, it alternatively provides an entire nearly-arbitrary mapping space in which to compare sets that preserves probabilistic concepts (and, therefore, roughly default concepts) which were demonstrated as useful for VC.

6 Sequences

Finally, a very different method for resolving shattering is a method from analysis.

Consider the liar's paradox, $A = [A = \bot]$; this statement is a contradiction. Likewise, while a $\mathcal{M}_{\underline{\mathcal{A}}}$ might almost always shatter a D, it might not shatter $\mathfrak{D}$, which is unintuitive and not default-rational. A resolution to this is to consider the following iterative approach, both generally and then as applied to default-VC, with infinite time.

Take a random initial assignment to atoms. Repeatedly determine if the statement is true; if false, randomly flip one value. If true, keep the assignment for the next iteration. If the sequence of flips reaches $\top$ in the limit and the assignments converge, then the sequence is valid.

Analagously, select a random initial model $\{\mathcal{A}\} = \underline{\mathcal{A}}$, and for some sample D, test if some $\mathcal{A} \in \underline{\mathcal{A}}$ shatters D, ie $\mathfrak{s}_{\mathcal{A}}(D) = \top$; if so, keep $\underline{\mathcal{A}}$ the same; else, add some $\mathcal{A} \in \mathcal{M}_{\underline{\mathcal{A}}}$ to $\underline{\mathcal{A}}$. If $\mathfrak{s}_{\underline{\mathcal{A}}}(D)$ converges to $\top$ in the limit, then $\mathcal{M}_{\underline{\mathcal{A}}} \mathrel{|\!\sim} \mathfrak{D}$. That is: if a sequence of $\mathfrak{s}_{\underline{\mathcal{A}}}(D)$ is Cauchy, it has a valuation (ie, is not a contradiction) and if the limit is $\top$, then $\mathcal{M}_{\underline{\mathcal{A}}} \mathrel{|\!\sim} \mathfrak{D}$.

If the model can VC-shatter the task, then the sequence will have limit $\top$; else, it will approach its limit as close it can and then not converge in the variables it cannot solve. If the value is near $\top$, then the model approximately solved the task; otherwise, it converged only on the solved variables and the modeler can focus on the variables that did not converge.

This formulation provides a simple way to evaluate valuations possibly contradictory on undecidable formulae: the liar's paradox, for example, will never converge to an assignment, and the formula $\bot$ does not converge to $\top$. If attempting to determine $\mathcal{M}_{\underline{\mathcal{A}}} \mathrel{|\!\sim} \mathfrak{D}$, one could attempt to see if a sequence of samples Cauchy-converges. *LC2* will solve most of the tasks but will fluctuate on the two tasks $\leftrightarrow, \otimes$ that it cannot solve. This method also naturally lends to a decomposable structure. Furthermore, a version of the PAC-bound can be used to determine a stopping criterion for a sequence. In contrast, in [17] Adams demonstrates that a reasonable consequence $\mathrel{|\!\sim}$ makes sense iff $\lim_{n\to\infty} P_n(A) = 1$ but says nothing if the sequence does not converge. Similarly, if a $\mathcal{M}_{\underline{\mathcal{A}}}$ and a $\mathfrak{D}$ are set up such that they cannot be well-evaluated for a shatter, they will not converge and randomly fluctuate.

7 Conclusion

In this project, we considered the VC dimension and its fundamental problem of being rigid both in what it is defined on and in its valuation, in addition to an incapacity to decompose its problems. Here, we successfully developed several theoretical strategies grounded under default logic: the first strategy, *Extremal Statistics*, made a game-theoretic interpretation through a probabilistic vehicle. This method had a structure implicity decomposable and that could make partial conclusions on finite, partitioned objects. The second strategy, *PAC-bound*, actually considered the task as a probabilistic one for which bounds could be produced for asymptotics could be evaluated over samples. ϵ-entailment related this strategy back to our default logic foundation. The third strategy, *Disjoint Cover*, engaged both kinds of previously seen task decompositions as covering problems and developed two set-theoretic strategies for achieving these; furthermore, it built a new measure function that was more general than probability and that permits useful reweightings of the data space. The fourth strategy, *Plausibility*, used the theory of plausibility to resolve two lasting issues of incompatible models (or tasks) and codify notions of ordering. Plausibility has a collection of highly useful properties – including default logic equivalence – that make it a reasonable general-purpose probability generalization for use as a way to formulate shatter depending on the tasks and models at hand. Finally, a short strategy of *Sequences* shows an alternative strategy absent default logic that can handle contradictions. Future work would certainly include implementations.

8 References

(1) Appendix B of: Error Estimation for Pattern Recognition, First Edition. Ulisses M. Braga-Neto and Edward R. Dougherty. 2015 The Institute of Electrical and Electronics Engineers, Inc. Published 2015 by John Wiley & Sons, Inc.

(2) Elements of Statistical Learning, Friedman, Hastie, Tibshirani, Springer, 2nd edition, 2009

(3) CS 257 lecture slides from 2-14-17, 2-16-17

(4) VC Dimensions and ϵ-net. Rudolf Fleischer, Fudan University Computer Science, Web. Accessed 3-14-17, modified December 2007. `http://www.tcs.fudan.edu.cn/rudolf/Courses/Algorithms/Alg_ss_07w/Webprojects/Qinbo_diameter/e_net.htm`

(5) Kraus, Sarit, Daniel Lehmann, and Menachem Magidor. "Nonmonotonic reasoning, preferential models and cumulative logics." Artificial intelligence 44.1-2 (1990): 167-207.

(6) Friedman, Nir, Joseph Y. Halpern, and Daphne Koller. "First-order conditional logic for default reasoning revisited." ACM Transactions on Computational Logic (TOCL) 1.2 (2000): 175-207.

(7) Friedman, Nir, and Joseph Y. Halpern. "Plausibility measures and default reasoning." Journal of the ACM 48.4 (2001): 648-685.

(8) Pearl, Judea. Probabilistic semantics for nonmonotonic reasoning: A survey. University of California (Los Angeles). Computer Science Department, 1989.

(9) Koller, Daphne, and Nir Friedman. Probabilistic graphical models: principles and techniques. MIT press, 2009.

(10) Rivest, Ronald L. "Learning decision lists." Machine learning 2.3 (1987): 229-246.

(11) Valiant, Leslie G. "A theory of the learnable." Communications of the ACM 27.11 (1984): 1134-1142.

(12) Lebesgue measure. (2017, February 6). In Wikipedia, The Free Encyclopedia. Retrieved 11:20, March 24, 2017, from `https://en.wikipedia.org/w/index.php?title=Lebesgue_measure&oldid=764082424`

(13) Dembo, Amir. Stochastic Processes. Lecture notes for Stats 219. Department of Statistics, Stanford University. August 2013.

(14) Arieli, Ofer, and Arnon Avron. "General patterns for nonmonotonic reasoning: From basic entailments to plausible relations." Logic Journal of IGPL 8.2 (2000): 119-148.

(15) Friedman, N. and J. Y. Halpern (1995a). Plausibility measures and default reasoning. Technical Report 9959, IBM. Available by anonymous ftp from `starry.stanford.edu/pub/nir` or via WWW at `http://robotics.stanford.edu/users/nir`.

(16) Friedman, Nir, and Joseph Y. Halpern. "Plausibility measures: a user's guide." Proceedings of the Eleventh conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., 1995.

(17) Adams, Ernest W. "Probability and the logic of conditionals." Studies in Logic and the Foundations of Mathematics 43 (1966): 265-316.

(18) Algorithms, Sanjoy Dasgupta, Christos Papadimitriou, and Umesh Vazirani. Chapter 8: NP-Complete Problems. "McGraw-Hill, 2008." 978-007.

(19) Koiran, Pascal, and Eduardo D. Sontag. "Neural networks with quadratic VC dimension." Journal of Computer and System Sciences 54.1 (1997): 190-198.

(20) Valiant, Leslie. Probably Approximately Correct: Natures Algorithms for Learning and Prospering in a Complex World. Basic Books, 2013.

(21) Batu, Tugkan, et al. "The complexity of approximating the entropy." SIAM Journal on Computing 35.1 (2005): 132-150.

(22) Vapnik, Vladimir Naumovich, and Vlamimir Vapnik. Statistical learning theory. Vol. 1. New York: Wiley, 1998.

(23) Miao, Xu, and Lin Liao. "VC-Dimension and its Applications in Machine Learning."

(24) Shalev-Shwartz, Shai, and Shai Ben-David. Understanding machine learning: From theory to algorithms. Cambridge university press, 2014.

(25) Ebbinghaus, Heinz-Dieter, and Jorg Flum. Finite model theory. Springer Science & Business Media, 2005.

(26) Paris, J.B. & Vencovska, A., Pure Inductive Logic, to appear in the ASL Perspectives in Logic Series, Cambridge University Press, 2013.

(27) de Salvo Braz, Rodrigo, Eyal Amir, and Dan Roth. "A survey of first-order probabilistic models." Innovations in Bayesian Networks. Springer Berlin Heidelberg, 2008. 289-317.

(28) Nakazawa, Makoto, et al. "On the complexity of hypothesis space and the sample complexity for machine learning." Systems, Man, and Cybernetics, 1994. Humans, Information and Technology., 1994 IEEE International Conference on. Vol. 1. IEEE, 1994.