Learnability Is a Compact Property

Julian Asilis Siddartha Devic Shaddin Dughmi
Vatsal Sharan Shang-Hua Teng

USC

Warm-up: binary classification

Known

Domain $\mathcal{X}$

Label set $\mathcal{Y} = \{0, 1\}$

Class $\mathcal{H} \subseteq \{0, 1\}^{\mathcal{X}}$

Unknown

Distribution $\mathcal{D}$ on $\mathcal{X}$

(Realizable learning: $\mathcal{D}$ arbitrary)

Ground truth $h^* \in \mathcal{H}$

Warm-up: binary classification

Known

Domain $\mathcal{X}$

Label set $\mathcal{Y} = \{0, 1\}$

Class $\mathcal{H} \subseteq \{0, 1\}^{\mathcal{X}}$

Unknown

Distribution $\mathcal{D}$ on $\mathcal{X}$

(Realizable learning: $\mathcal{D}$ arbitrary)

Ground truth $h^* \in \mathcal{H}$

Goal

Given iid draws from $\mathcal{D}$ (labeled by h^*),

guess h^*!

Judged by error,

$$\mathbb{P}_{x \sim \mathcal{D}}\left(f(x) \neq h^*(x)\right)$$

Warm-up: binary classification

Known

Domain $\mathcal{X}$

Label set $\mathcal{Y} = \{0, 1\}$

Class $\mathcal{H} \subseteq \{0, 1\}^{\mathcal{X}}$

Unknown

Distribution $\mathcal{D}$ on $\mathcal{X}$

(Realizable learning: $\mathcal{D}$ arbitrary)

Ground truth $h^* \in \mathcal{H}$

Goal

Given iid draws from $\mathcal{D}$ (labeled by h^*),

guess h^*!

Judged by error,

$$\mathbb{P}_{x \sim \mathcal{D}}\big(f(x) \neq h^*(x)\big)$$

Can $\mathcal{H}$ be learned with error $\to 0$ as # samples $\to \infty$?

VC dimension is all you need

VC dimension

$\mathcal{H}$ *shatters* $\mathrm{S} = (x_1, \ldots, x_n)$ when
$\mathcal{H}|_S = \{0, 1\}^n$

VC$(\mathcal{H}) =$ size of largest shattered set

Fundamental theorem:

$\mathcal{H}$ is learnable $\Leftrightarrow$ VC $(\mathcal{H}) < \infty$.

Attaining error $\leq \varepsilon$ w.h.p. requires $\tilde{\Theta}(\frac{VC(\mathcal{H})}{\varepsilon})$ points.

VC dimension is all you need

VC dimension

$\mathcal{H}$ *shatters* $S = (x_1, \dots, x_n)$ when
$\mathcal{H}|_S = \{0,1\}^n$

VC$(\mathcal{H}) =$ size of largest shattered set

Fundamental theorem:

$\mathcal{H}$ is learnable $\Leftrightarrow$ VC $(\mathcal{H}) < \infty$.

Attaining error $\leq \varepsilon$ w.h.p. requires $\tilde{\Theta}(\frac{VC(\mathcal{H})}{\varepsilon})$ points.

Binary classification solved 🙂

An observation: VC dimension only "knows" about finite projections of $\mathcal{H}$…

Why is that enough?

Binary classification is “compact”

VC theory reveals compactness:

- If $\mathcal{H}$’s finite projections look good, then $\mathcal{H}$ is learnable

- Equiv: if $\mathcal{H}$ is not learnable, it has arbitrarily bad finite projections

Binary classification is "compact"

VC theory reveals compactness:

– If $\mathcal{H}$'s finite projections look good, then $\mathcal{H}$ is learnable

– Equiv: if $\mathcal{H}$ is not learnable, it has arbitrarily bad finite projections

Why does this work?

When learning $\mathcal{H}$, distribution $\mathcal{D}$ can have infinite support, even be continuous!

Considering finite projections $\mathcal{H}|_S$ doesn't pick up on hardness of learning these distributions…

Beyond binary classification

Much of learning theory follows the skeleton of VC dimension

1. Say $\mathcal{H}$ shatters $\mathrm{S} = (x_1, \ldots, x_n)$ if $\mathcal{H}|_S$ has a finite subset such that…
2. Let $\mathrm{d} = \mathrm{d}(\mathcal{H})$ be the size of the largest shattered set
3. Prove $\mathcal{H}$ is learnable $\Leftrightarrow d < \infty$

Beyond binary classification

Much of learning theory follows the skeleton of VC dimension

1. Say $\mathcal{H}$ shatters $\mathrm{S} = (x_1, \dots, x_n)$ if $\mathcal{H}|_S$ has a finite subset such that…
2. Let $\mathrm{d} = \mathrm{d}(\mathcal{H})$ be the size of the largest shattered set
3. Prove $\mathcal{H}$ is learnable $\Leftrightarrow d < \infty$

Examples

– Fat shattering dimension

– Graph dimension

– Natarajan dimension

– DS dimension

– Littlestone dimension

Beyond binary classification

Much of learning theory follows the skeleton of VC dimension

1. Say $\mathcal{H}$ shatters $S = (x_1, \dots, x_n)$ if $\mathcal{H}|_S$ has a finite subset such that…
2. Let $\mathrm{d} = \mathrm{d}(\mathcal{H})$ be the size of the largest shattered set
3. Prove $\mathcal{H}$ is learnable $\Leftrightarrow d < \infty$

Examples

– Fat shattering dimension

– Graph dimension

– Natarajan dimension

– DS dimension

– Littlestone dimension

Why is this happening? Will we eventually describe all kinds of learning in this way?

Beyond binary classification

Much of learning theory follows the skeleton of VC dimension

1. Say $\mathcal{H}$ shatters $\mathrm{S} = (x_1, \dots, x_n)$ if $\mathcal{H}|_S$ has a finite subset such that…
2. Let $\mathrm{d} = \mathrm{d}(\mathcal{H})$ be the size of the largest shattered set
3. Prove $\mathcal{H}$ is learnable $\Leftrightarrow d < \infty$

Examples

– Fat shattering dimension

– Graph dimension

– Natarajan dimension

– DS dimension

– Littlestone dimension

Why is this happening? Will we eventually describe all kinds of learning in this way? **No.**

EMX learning: noncompact

EMX Learning

$\mathcal{X} = \mathbb{R}, \mathcal{Y} = \{0,1\}$
$\mathcal{H} = \{h : |h^{-1}(1)| < \infty\}$

Given h^*, $\mathcal{D}$ must be supported on $(h^*)^{-1}(1)$. I.e., only see the label 1

Learner must be *proper*, emit an $h \in \mathcal{H}$

EMX learning: noncompact

<u>EMX Learning</u>

$\mathcal{X} = \mathbb{R}, \mathcal{Y} = \{0,1\}$
$\mathcal{H} = \{h \ : \ |h^{-1}(1)| < \infty\}$

Given h^*, $\mathcal{D}$ must be supported on $(h^*)^{-1}(1)$. I.e., only see the label 1

Learner must be *proper*, emit an $h \in \mathcal{H}$

In English:

– Ground set $\mathcal{X}$

– Distribution $\mathcal{D}$ over $\mathcal{X}$ (finite support)

– Given iid samples from $\mathcal{D}$, pick finite $S \subseteq \mathcal{X}$ with maximum $\mathcal{D}$-measure

When $\mathcal{X}$ is finite, trivial.

Pick $S = \mathcal{X}$!

What about $\mathcal{X} = \mathbb{R}$?

EMX learning: noncompact

For infinite $\mathcal{X}$, learnability depends on $|\mathcal{X}|$

(Such that $\mathcal{H}$ is learnable $\Leftrightarrow |\mathcal{X}| < \aleph_\omega$. Thus, undecidable when $\mathcal{X} = \mathbb{R}$.)

If $\mathcal{X}$ too large, $\mathcal{H}$ is not learnable. Even though all its finite restrictions are easy!

Failure of compactness!
(When learners are required to be proper)

In English:

– Ground set $\mathcal{X}$

– Distribution $\mathcal{D}$ over $\mathcal{X}$ (finite support)

– Given iid samples from $\mathcal{D}$, pick finite $S \subseteq \mathcal{X}$ with maximum $\mathcal{D}$-measure

When $\mathcal{X}$ is finite, trivial.

Pick $S = \mathcal{X}$!

What about $\mathcal{X} = \mathbb{R}$?

EMX learning: noncompact

For infinite $\mathcal{X}$, learnability depends on $|\mathcal{X}|$

(Such that $\mathcal{H}$ is learnable $\Leftrightarrow |\mathcal{X}| < \aleph_\omega$. Thus, undecidable when $\mathcal{X} = \mathbb{R}$.)

If $\mathcal{X}$ too large, $\mathcal{H}$ is not learnable. Even though all its finite restrictions are easy!

Failure of compactness!
(When learners are required to be proper)

Where and why does compactness appear in improper supervised learning?

In light of EMX learning, why do standard learning paradigms <u>happen</u> to be compact?

Our Results

Let:
- $\mathcal{X}$ = arbitrary set
- $\mathcal{Y}$ = *proper* metric space
- $\mathcal{H}$ = hypothesis class
- "Finite projection" of $\mathcal{H}$ = finite subset of $\mathcal{H}|_S$ for finite $S \subseteq \mathcal{X}$

Our Results

Let:

- $\mathcal{X}$ = arbitrary set
- $\mathcal{Y}$ = *proper* metric space
- $\mathcal{H}$ = hypothesis class
- "Finite projection" of $\mathcal{H}$ = finite subset of $\mathcal{H}|_S$ for finite $S \subseteq \mathcal{X}$

Compact $\Longleftrightarrow$ closed & bounded

- $\mathbb{R}^n$ and its closed subsets (any norm)
- Any finite space
- Any compact space

Our Results

Let:

- $\mathcal{X}$ = arbitrary set
- $\mathcal{Y}$ = proper metric space
- $\mathcal{H}$ = hypothesis class
- "Finite projection" of $\mathcal{H}$ = finite subset of $\mathcal{H}|_S$ for finite $S \subseteq \mathcal{X}$

Theorem: For realizable learning, the following are equivalent,

1. $\mathcal{H}$ can be learned with transductive sample complexity m
2. Any finite projection of $\mathcal{H}$ can be learned with complexity m

Our Results

Let:
- $\mathcal{X}$ = arbitrary set
- $\mathcal{Y}$ = proper metric space
- $\mathcal{H}$ = hypothesis class
- "Finite projection" of $\mathcal{H}$ = finite subset of $\mathcal{H}|_S$ for finite $S \subseteq \mathcal{X}$

Theorem: For realizable learning, the following are equivalent,
1. $\mathcal{H}$ can be learned with transductive sample complexity m
2. Any finite projection of $\mathcal{H}$ can be learned with complexity m

Very general and **exact** form of compactness!

What if $\mathcal{Y}$ isn't proper?

Our Results

Let:

- $\mathcal{X}$ = arbitrary set
- $\mathcal{Y}$ = arbitrary metric space
- $\mathcal{H}$ = hypothesis class
- "Finite projection" of $\mathcal{H}$ = finite subset of $\mathcal{H}|_S$ for finite $S \subseteq \mathcal{X}$

Our Results

Let:
- $\mathcal{X}$ = arbitrary set
- $\mathcal{Y}$ = arbitrary metric space
- $\mathcal{H}$ = hypothesis class
- "Finite projection" of $\mathcal{H}$ = finite subset of $\mathcal{H}|_S$ for finite $S \subseteq \mathcal{X}$

Theorem: For realizable learning, there exists an (improper) $\mathcal{Y}$ s.t.
1. Any finite projection of $\mathcal{H}$ can be learned with complexity m
2. Learning $\mathcal{H}$ requires $m_{\mathcal{H}} > m$ samples, with $m_{\mathcal{H}}(\varepsilon) \geq m(\varepsilon/2)$ for some ε

Improper $\mathcal{Y}$: compactness can fail by at least a factor of 2

Our Results

Let:
- $\mathcal{X}$ = arbitrary set
- $\mathcal{Y}$ = arbitrary metric space
- $\mathcal{H}$ = hypothesis class
- "Finite projection" of $\mathcal{H}$ = finite subset of $\mathcal{H}|_S$ for finite $S \subseteq \mathcal{X}$

Theorem: Suppose any finite projection of $\mathcal{H}$ can be learned with realizable complexity m. Then $\mathcal{H}$ is learnable with at most $m(\varepsilon/2)$ samples.

Improper $\mathcal{Y}$: compactness can fail by at ~~least~~ **most** a factor of 2.

Complete characterization of compactness for realizable learning with metric losses!

Beyond the realizable case

Agnostic learning

$\mathcal{D}$ can be any distribution on $\mathcal{X} \times \mathcal{Y}$

Proper $\mathcal{Y}$: **exact** compactness of sample complexity! $\mathcal{H}$ learnable with m samples $\Leftrightarrow$ finite projections learnable with m samples

Improper $\mathcal{Y}$: compactness can fail by at least a factor of 2. Maybe more?

Beyond the realizable case

Agnostic learning

$\mathcal{D}$ can be any distribution on $\mathcal{X} \times \mathcal{Y}$

Proper $\mathcal{Y}$: **exact** compactness of sample complexity! $\mathcal{H}$ learnable with m samples $\Longleftrightarrow$ finite projections learnable with m samples

Improper $\mathcal{Y}$: compactness can fail by at least a factor of 2. Maybe more?

Distribution-family learning

$\mathcal{D}$ constrained to certain distributions on $\mathcal{X} \times \mathcal{Y}$, i.e., $\mathcal{D} \in \mathbb{D}$

Call $\mathbb{D}$ **well-behaved** if it is closed under empirical distributions

($\forall\, \mathcal{D} \in \mathbb{D}$ and $S \sim \mathcal{D}^n$, Unif(S) $\in \mathbb{D}$. E.g., partial, EMX, etc.)

Proper $\mathcal{Y}$, well-behaved $\mathbb{D}$: **exact** compactness of sample complexity

Beyond the realizable case

Agnostic learning

$\mathcal{D}$ can be any distribution on $\mathcal{X} \times \mathcal{Y}$

Proper $\mathcal{Y}$: **exact** compactness of sample complexity! $\mathcal{H}$ learnable with m samples $\Longleftrightarrow$ finite projections learnable with m samples

Improper $\mathcal{Y}$: compactness can fail by at least a factor of 2. Maybe more?

Distribution-family learning

$\mathcal{D}$ constrained to certain distributions on $\mathcal{X} \times \mathcal{Y}$, i.e., $\mathcal{D} \in \mathbb{D}$

Call $\mathbb{D}$ **well-behaved** if it is closed under empirical distributions

($\forall\ \mathcal{D} \in \mathbb{D}$ and $S \sim \mathcal{D}^n$, Unif(S) $\in \mathbb{D}$. E.g., partial, EMX, etc.)

Proper $\mathcal{Y}$, well-behaved $\mathbb{D}$: **exact** compactness of sample complexity

EMX pathology relies on constraining to proper learners!

Transductive learning

Transductive learning model

1. Adversary selects *n* datapoints
2. One label removed uniformly at random
3. Fill in the blank

Error = average loss over uniformly random "?"

Transductive learning

Transductive learning model

1. Adversary selects *n* datapoints
2. One label removed uniformly at random
3. Fill in the blank

Cat

Dog

?

Dog

Error = average loss over uniformly random "?"

Looks more fine grained than iid model, i.e., sample by sample

However, essentially equivalent to PAC
(Sample complexities equivalent up to log factors)

Key point: one-inclusion graphs (OIGs) perfect to study transductive model

One-inclusion graphs

Realizable **one-inclusion graph** of $\mathcal{H}$ on $S \in \mathcal{X}^n$:

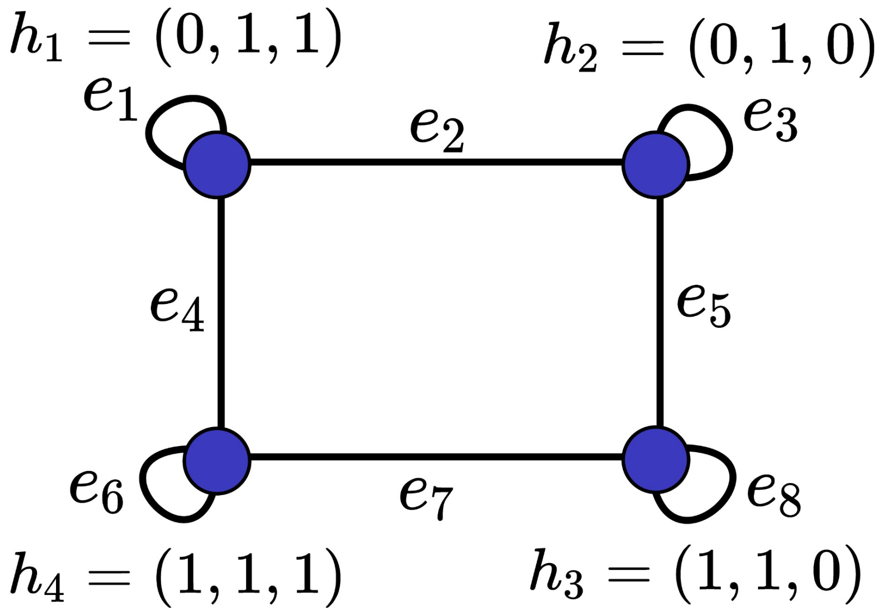
- Vertex set: $\mathcal{H}|_S$
- Edge set: group hypotheses that agree on $n - 1$ points

One-inclusion graphs

Realizable **one-inclusion graph** of $\mathcal{H}$ on $S \in \mathcal{X}^n$:

– Vertex set: $\mathcal{H}|_S$

– Edge set: group hypotheses that agree on $n - 1$ points



One-inclusion graphs

Realizable **one-inclusion graph** of $\mathcal{H}$ on $S \in \mathcal{X}^n$:

– Vertex set: $\mathcal{H}|_S$

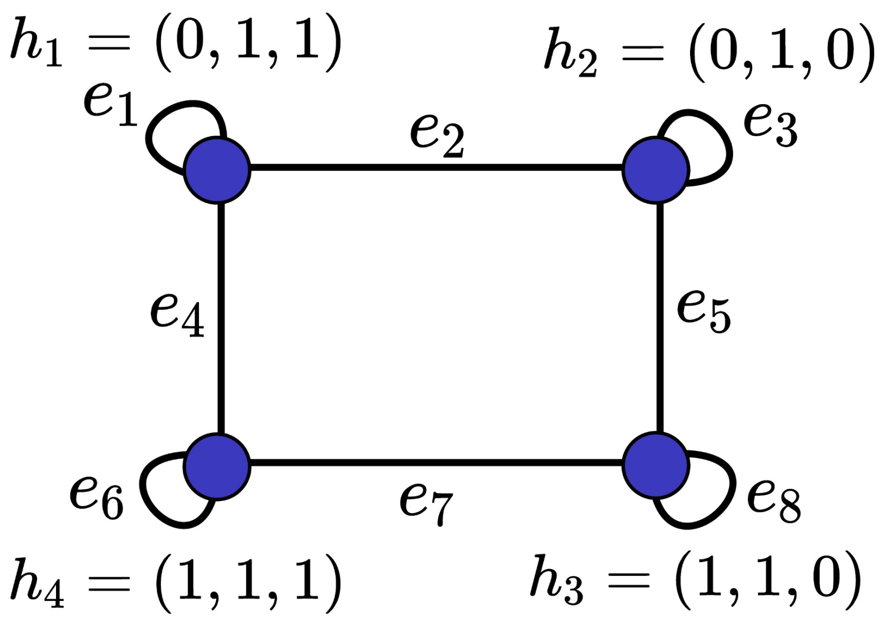
– Edge set: group hypotheses that agree on $n-1$ points

Learner for $\mathcal{H}$ = orientation of OIGs

– edge = training set + unlabeled test point

 – e.g., $e_2 = (0, 1, ?\,)$

– Completing “?” = choice of incident node



One-inclusion graphs

Realizable **one-inclusion graph** of $\mathcal{H}$ on $S \in \mathcal{X}^n$:

– Vertex set: $\mathcal{H}|_S$

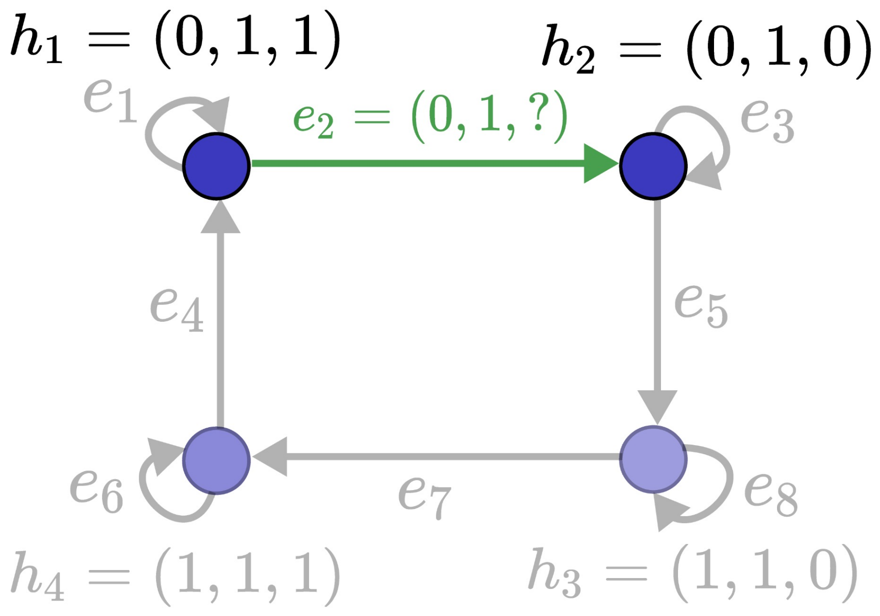
– Edge set: group hypotheses that agree on $n-1$ points

Learner for $\mathcal{H}$ = orientation of OIGs

– edge = training set + unlabeled test point

  – e.g., $e_2 = (0, 1, ?)$

– Completing "?" = choice of incident node



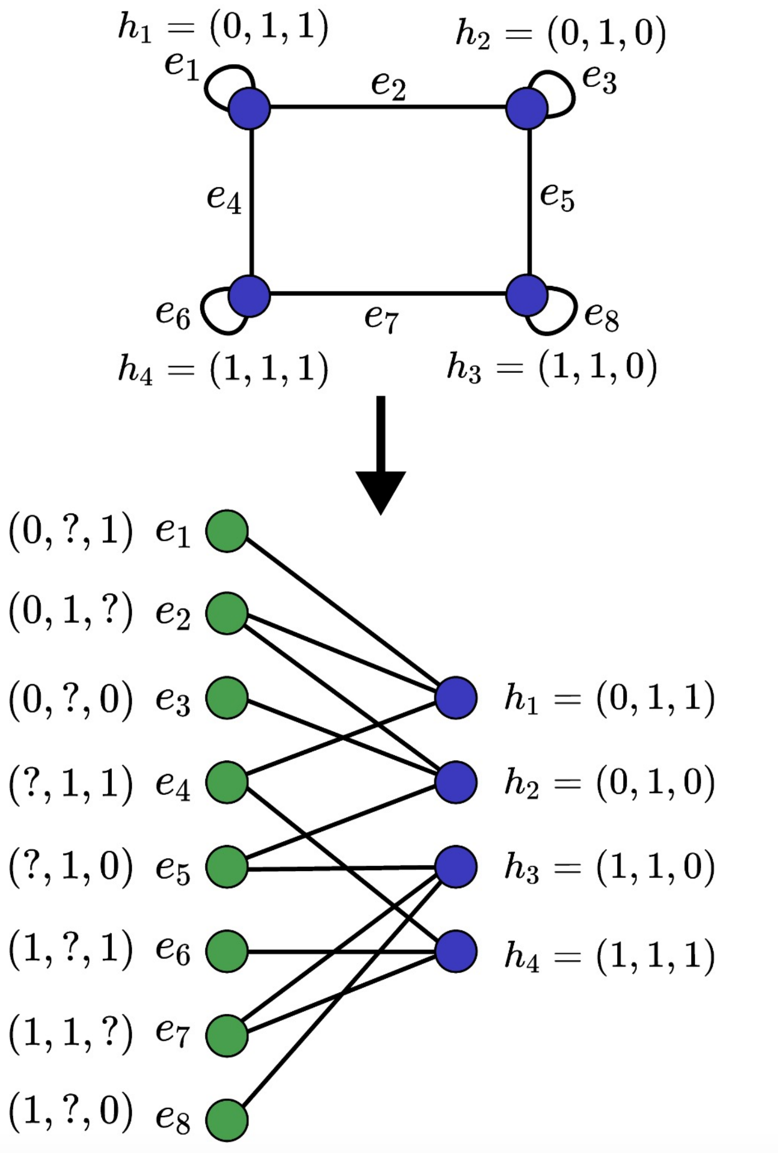
One-inclusion graphs

Bipartite view:

– LHS = *variables* valued in $\mathcal{Y}$

– RHS = *functions* tracking error of ground truth

  – E.g., $h_4(e_4, e_6, e_7) = \ell(1, e_4) + \ell(1, e_6) + \ell(1, e_7)$



One-inclusion graphs

Bipartite view:

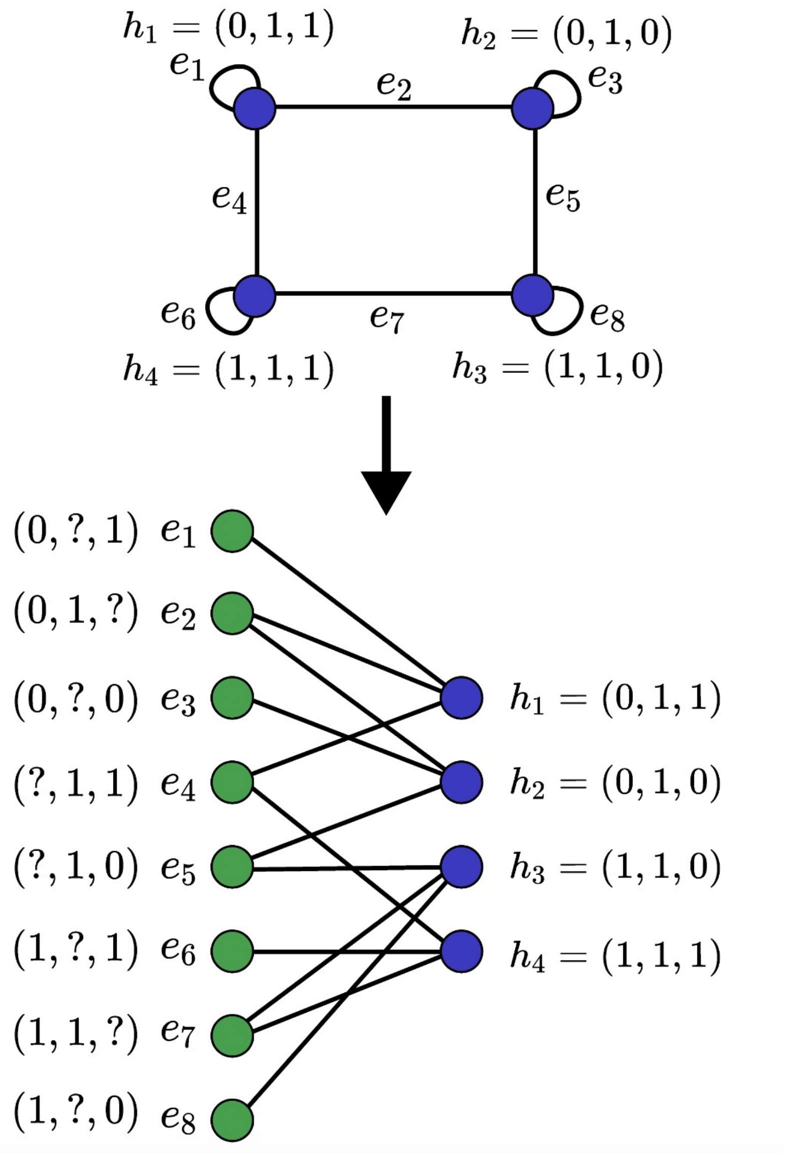
– LHS = *variables* valued in $\mathcal{Y}$

– RHS = *functions* tracking error of ground truth

  – E.g., $h_4(e_4, e_6, e_7) = \ell(1, e_4) + \ell(1, e_6) + \ell(1, e_7)$

Now, learner = assignment of variables

Goal: assign variables to keep all functions below ϵ



Realizable compactness

Theorem: Let

- L = set of variables, valued in metric space
- R = set of *proper* functions, each of form $\prod_{i=1}^{n} \ell_i \to \mathbb{R}_{\geq 0}$

Pre-image of compact is compact

Realizable compactness

Theorem: Let
- L = set of variables, valued in metric space
- R = set of *proper* functions, each of form $\prod_{i=1}^{n} \ell_i \to \mathbb{R}_{\geq 0}$

Then the following are equivalent:
1. Can assign variables to keep all functions $\leq \epsilon$
2. For each finite $S \subseteq R$, can assign variables to keep those functions $\leq \epsilon$

Realizable compactness

Theorem: Let
- L = set of variables, valued in metric space
- R = set of *proper* functions, each of form $\prod_{i=1}^{n} \ell_i \to \mathbb{R}_{\geq 0}$

Then the following are equivalent:
1. Can assign variables to keep all functions $\leq \epsilon$
2. For each finite $S \subseteq R$, can assign variables to keep those functions $\leq \epsilon$

Proof sketch:
1 $\Longrightarrow$ 2: immediate

2 $\Longrightarrow$ 1: Zorn's lemma

Realizable compactness

Theorem: Let
- L = set of variables, valued in metric space
- R = set of *proper* functions, each of form $\prod_{i=1}^{n} \ell_i \to \mathbb{R}_{\geq 0}$

Then the following are equivalent:
1. Can assign variables to keep all functions $\leq \epsilon$
2. For each finite $S \subseteq R$, can assign variables to keep those functions $\leq \epsilon$

Proof sketch:
1 $\Longrightarrow$ 2: immediate

2 $\Longrightarrow$ 1: Zorn's lemma
- $\mathcal{P}$ = partial assignments of variables that can be completed to satisfy any finite $S \subseteq R$
- Any $\mathrm{P} \in \mathcal{P}$ can have one free variable assigned

  (Use finite intersection property of compact sets)
- Chains in $\mathcal{P}$ have upper bounds

  (Use fact that each $\mathrm{r} \in R$ depends upon finitely many variables)
- Thus maximal element = total assignment

Realizable compactness

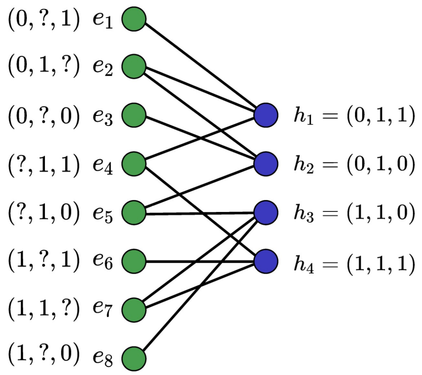
Theorem: Let
- L = set of variables, valued in metric space
- R = set of *proper* functions, each of form $\prod_{i=1}^{n} \ell_i \to \mathbb{R}_{\geq 0}$

Then the following are equivalent:
1. Can assign variables to keep all functions $\leq \epsilon$
2. For each finite $S \subseteq R$, can assign variables to keep those functions $\leq \epsilon$

For learning:
- L = LHS nodes, thought of as variables in $\mathcal{Y}$
- R = RHS nodes, tracking transductive error
  - E.g., $h_4(e_4, e_6, e_7) = \ell(1, e_4) + \ell(1, e_6) + \ell(1, e_7)$
  - When $\mathcal{Y}$ is proper, these functions are proper, b/c continuous & reflect bounded sets

1. = learning $\mathcal{H}$

2. = learning $\mathcal{H}$'s finite projections



Realizable *non*compactness

Build a pathological $\mathcal{Y}$:

– $\mathcal{Y} = \mathcal{A} \cup \mathcal{B}$

– $\mathcal{A} =$ infinite set, points all distance 2 apart

Realizable *non*compactness

Build a pathological $\mathcal{Y}$:

- $\mathcal{Y} = \mathcal{A} \cup \mathcal{B}$
- $\mathcal{A} =$ infinite set, points all distance 2 apart
- $\mathcal{B} =$ points indexed by finite subsets of $\mathcal{A}$, e.g., b_A for $A \subseteq \mathcal{A}$.
  - b_A is distance 1 from points in A and $\mathcal{B}$, distance 2 from $\mathcal{A} \setminus A$

Realizable *non*compactness

Build a pathological $\mathcal{Y}$:

- $\mathcal{Y} = \mathcal{A} \cup \mathcal{B}$
- $\mathcal{A} =$ infinite set, points all distance 2 apart
- $\mathcal{B} =$ points indexed by finite subsets of $\mathcal{A}$,
  e.g., b_A for $A \subseteq \mathcal{A}$.
  - b_A is distance 1 from points in A and $\mathcal{B}$,
    distance 2 from $\mathcal{A} \setminus A$

In English:

- $\mathcal{Y}$ = infinite set of points all distance 2 apart
- *But* each finite $\mathrm{Y} \subseteq \mathcal{Y}$ has a "center"
  distance 1 from all points in Y

Realizable *non*compactness

Build a pathological $\mathcal{Y}$:

- $\mathcal{Y} = \mathcal{A} \cup \mathcal{B}$
- $\mathcal{A} =$ infinite set, points all distance 2 apart
- $\mathcal{B} =$ points indexed by finite subsets of $\mathcal{A}$,
  e.g., b_A for $A \subseteq \mathcal{A}$.
  - b_A is distance 1 from points in A and $\mathcal{B}$, distance 2 from $\mathcal{A} \setminus A$

In English:

- $\mathcal{Y}$ = infinite set of points all distance 2 apart
- *But* each finite $\mathrm{Y} \subseteq \mathcal{Y}$ has a "center" distance 1 from all points in Y

Let $\mathcal{H}$ be very complex class (e.g., $\mathcal{Y}^{\mathcal{X}}$)

- Learning $\mathcal{H}$: pay distance 2 in worst case
- Learning finite projection: promised to only see labels from $\mathrm{Y} \subseteq \mathcal{Y}$
  - Predict Y's "center" to lock in loss ≤ 1

Realizable *non*compactness

Build a pathological $\mathcal{Y}$:

- $\mathcal{Y} = \mathcal{A} \cup \mathcal{B}$
- $\mathcal{A} =$ infinite set, points all distance 2 apart
- $\mathcal{B} =$ points indexed by finite subsets of $\mathcal{A}$, e.g., b_A for $A \subseteq \mathcal{A}$.
  - b_A is distance 1 from points in A and $\mathcal{B}$, distance 2 from $\mathcal{A} \setminus A$

In English:

- $\mathcal{Y}$ = infinite set of points all distance 2 apart
- *But* each finite $\mathrm{Y} \subseteq \mathcal{Y}$ has a "center" distance 1 from all points in Y

Let $\mathcal{H}$ be very complex class (e.g., $\mathcal{Y}^{\mathcal{X}}$)

- Learning $\mathcal{H}$: pay distance 2 in worst case
- Learning finite projection: promised to only see labels from $\mathrm{Y} \subseteq \mathcal{Y}$
  - Predict Y's "center" to lock in loss ≤ 1

Hence failure of compactness by factor 2

- *But* this is tight: similar(ish) use of Zorn's lemma
- Factor 2 arises from triangle inequality

Beyond realizable

Agnostic and distribution-family: use abstract compactness result, black-box

Theorem: Let

- L = set of variables, valued in metric space
- R = set of *proper* functions, each of form $\prod_{i=1}^{n} \ell_i \to \mathbb{R}_{\geq 0}$

Then the following are equivalent:

1. Can assign variables to keep all functions $\leq \epsilon$
2. For each finite $S \subseteq R$, can assign variables to keep those functions $\leq \epsilon$

Beyond realizable

Agnostic and distribution-family: use abstract compactness result, black-box

- L = transductive learning instances, with "?"
  - E.g., $(y_1, y_2, ?\,, y_4)$
  - Thought of as variable valued in $\mathcal{Y}$
- R = excess transductive error of ground truths
  - Subtract error of best $h \in \mathcal{H}$

Theorem: Let
- L = set of variables, valued in metric space
- R = set of *proper* functions, each of form $\prod_{i=1}^{n} \ell_i \to \mathbb{R}_{\geq 0}$

Then the following are equivalent:
1. Can assign variables to keep all functions $\leq \epsilon$
2. For each finite $S \subseteq R$, can assign variables to keep those functions $\leq \epsilon$

Beyond realizable

Agnostic and distribution-family: use abstract compactness result, black-box

- L = transductive learning instances, with "?"
  - E.g., $(y_1, y_2, ?, y_4)$
  - Thought of as variable valued in $\mathcal{Y}$
- R = excess transductive error of ground truths
  - Subtract error of best $h \in \mathcal{H}$

Exact compactness for proper $\mathcal{Y}$

By same counterexample, fails by factor of 2 for improper $\mathcal{Y}$. Maybe more?

Theorem: Let

- L = set of variables, valued in metric space
- R = set of *proper* functions, each of form $\prod_{i=1}^{n} \ell_i \to \mathbb{R}_{\geq 0}$

Then the following are equivalent:

1. Can assign variables to keep all functions $\leq \epsilon$
2. For each finite $S \subseteq R$, can assign variables to keep those functions $\leq \epsilon$

Bonus result: Hall's theorem

Proper $\mathcal{Y}$: covers almost everything

– $\mathbb{R}^n$ and its closed subsets (any norm)

– Finite metric spaces

– Compact metric spaces

But doesn't cover multiclass classification
with arbitrary # labels.

Bonus result: Hall's theorem

Proper $\mathcal{Y}$: covers almost everything

– $\mathbb{R}^n$ and its closed subsets (any norm)

– Finite metric spaces

– Compact metric spaces

But doesn't cover multiclass classification with arbitrary # labels. Nevertheless, it's compact!

Theorem: Classification enjoys *exact* compactness, in both the realizable and agnostic cases.

Bonus result: Hall's theorem

Proper $\mathcal{Y}$: covers almost everything

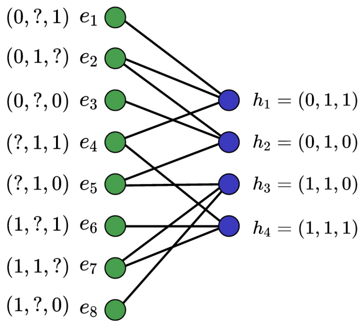
- $\mathbb{R}^n$ and its closed subsets (any norm)
- Finite metric spaces
- Compact metric spaces

But doesn't cover multiclass classification with arbitrary # labels. Nevertheless, it's compact!

Theorem: Classification enjoys *exact* compactness, in both the realizable and agnostic cases.

Proof sketch:

- Under 0-1 loss, transductive error equals the **indegree** of a RHS node
  - Complete "?" by picking desired ground truth



- Learning becomes a matching problem

Bonus result: Hall's theorem

Proper $\mathcal{Y}$: covers almost everything

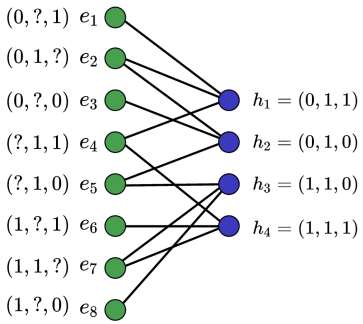
- $\mathbb{R}^n$ and its closed subsets (any norm)
- Finite metric spaces
- Compact metric spaces

But doesn't cover multiclass classification with arbitrary # labels. Nevertheless, it's compact!

Theorem: Classification enjoys *exact* compactness, in both the realizable and agnostic cases.

Proof sketch:

- Under 0-1 loss, transductive error equals the **indegree** of a RHS node
  - Complete "?" by picking desired ground truth

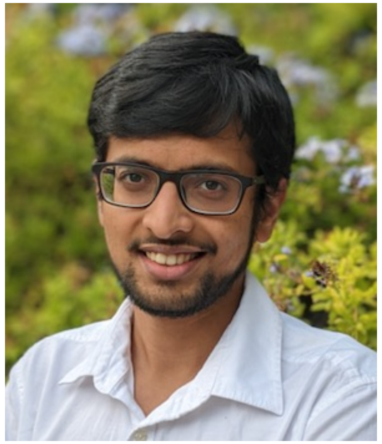


- Learning becomes a matching problem
- Key step: our compactness result implies M. Hall's theorem for infinite graphs
  - Uses fact that RHS degrees are all finite
- Thus matchability ≡ Hall's criterion. Done!

Thank you

Sid Devic

Vatsal Sharan



Shaddin Dughmi

Shang-Hua Teng