UNIVERSIDAD NACIONAL DE COLOMBIA

Mathematical Logic (Model Theory) meets Statistical Learning Theory

Andrés VILLAVECES - *Universidad Nacional de Colombia - Bogotá*

TMLI (Centre for Topological Machine Learning and Innovation)
The University of Buckingham - September 2024

Our path

Learning: the set-up

What logic (model theory) can say about learning theory

A conversation between three parts of mathematics

In many ways, statistical learning theory has been enacting a conversation between three different areas of mathematics:

A conversation between three parts of mathematics

In many ways, statistical learning theory has been enacting a conversation between three different areas of mathematics:

- Model theory (a part of mathematical logic),

We will explore the natural set-up of learning theory first from the logic perspective (no proofs! just examples and basic notions!) and then contrast it with some of its incarnations in the other two domains.

A conversation between three parts of mathematics

In many ways, statistical learning theory has been enacting a conversation between three different areas of mathematics:

- Model theory (a part of mathematical logic),
- Probability (especially, probabilistic learning theory),

We will explore the natural set-up of learning theory first from the logic perspective (no proofs! just examples and basic notions!) and then contrast it with some of its incarnations in the other two domains.

A conversation between three parts of mathematics

In many ways, statistical learning theory has been enacting a conversation between three different areas of mathematics:

- Model theory (a part of mathematical logic),
- Probability (especially, probabilistic learning theory),
- Combinatorics

We will explore the natural set-up of learning theory first from the logic perspective (no proofs! just examples and basic notions!) and then contrast it with some of its incarnations in the other two domains.

Part 1

Learning: the set-up

The basic set-up (as per Wirth et al.)

Learning problems may be seen as consisting of

- A non-empty set $\mathcal{X}$ called the **instance space** (the class of concepts to learn)
- The **sample space** $\mathcal{Z} = \mathcal{X} \times \{0, 1\}$,
- A non-empty **hypothesis space** $\mathcal{H} \subseteq \{0, 1\}^{\mathcal{X}}$,
- A σ-algebra $\Sigma_{\mathcal{Z}}$ on $\mathcal{Z}$ containing all finite subsets of $\mathcal{Z}$,
- A **set of distributions** $\mathcal{D}$ on $(\mathcal{Z}, \Sigma_{\mathcal{Z}})$.

The basic set-up (as per Wirth et al.)

Learning problems may be seen as consisting of

- A non-empty set $\mathcal{X}$ called the **instance space** (the class of concepts to learn)
- The **sample space** $\mathcal{Z} = \mathcal{X} \times \{0, 1\}$,
- A non-empty **hypothesis space** $\mathcal{H} \subseteq \{0, 1\}^{\mathcal{X}}$,
- A σ-algebra $\Sigma_{\mathcal{Z}}$ on $\mathcal{Z}$ containing all finite subsets of $\mathcal{Z}$,
- A **set of distributions** $\mathcal{D}$ on $(\mathcal{Z}, \Sigma_{\mathcal{Z}})$.

Furthermore, every hypothesis $h \in \mathcal{H}$ determines an element of $\Sigma_{\mathcal{Z}}$:

$$\Gamma(h) = \{(x, y) \in \mathcal{Z} : h(x) = y\}.$$

Trial and error - How?

When we fix an arbitrary distribution $\mathbb{D} \in \mathcal{D}$, we may generate a sequence of samples from $\mathcal{Z}$:

$$\mathsf{z} = \left(\left(x_1, y_1 \right), \ldots, \left(x_m, y_m \right) \right) :$$

the **input data** for a learning function $\mathcal{A}$ that gives as output a hypothesis $h = \mathcal{A}(\mathsf{z})$ in $\mathcal{H}$.

Trial and error - Goal

Idea: minimizing the true error of h:

$$\mathrm{er}_{\mathbb{D}}(h) = \mathbb{D}\left(\{(x,y) \in \mathcal{Z} : h(x) \neq y\}\right) = \mathbb{D}\left(\mathcal{Z} \setminus \Gamma(h)\right).$$

This captures the goal: getting an error close to

$$\mathrm{opt}_{\mathbb{D}}\left(\mathcal{H}\right) = \inf_{h \in \mathcal{H}} \mathrm{er}_{\mathbb{D}}(h).$$

PAC learning

A learning function

$$\mathcal{A} : \bigcup_{m \in \mathbb{N}} \mathcal{Z}^m \to \mathcal{H}$$

for $\mathcal{H}$ is probably approximately correct (PAC) with respect to $\mathcal{D}$ if $\forall \epsilon, \delta \in (0,1) \exists m_0 \in \mathbb{N} \forall m \geq m_0$ and $\forall \mathbb{D} \in \mathcal{D}$

$$\mathbb{D}^m \left(\{ z \in \mathcal{Z}^m : \mathrm{er}_{\mathbb{R}} \left(\mathcal{A}(z) - \mathrm{opt}_{\mathbb{D}}(\mathcal{H}) \leq \epsilon \right) \} \right) \geq 1 - \delta.$$

This may be refined with extra probability theory information on the guessing set: $C \in \Sigma_{\mathcal{Z}} \ldots$

PAC learning

A learning function

$$\mathcal{A} : \bigcup_{m \in \mathbb{N}} \mathcal{Z}^m \to \mathcal{H}$$

for $\mathcal{H}$ is probably approximately correct (PAC) with respect to $\mathcal{D}$ if $\forall \epsilon, \delta \in (0,1) \exists m_0 \in \mathbb{N} \forall m \geq m_0$ and $\forall \mathbb{D} \in \mathcal{D}$

$$\mathbb{D}^m \left(\{ z \in \mathcal{Z}^m : \mathrm{er}_{\mathbb{R}} \left(\mathcal{A}(z) - \mathrm{opt}_{\mathbb{D}}(\mathcal{H}) \leq \epsilon \right) \} \right) \geq 1 - \delta.$$

$\mathcal{H}$ is PAC learnable if $\exists \mathcal{A}$ a PAC learning function for $\mathcal{H}$.

PAC learning

A learning function

$$\mathcal{A} : \bigcup_{m \in \mathbb{N}} \mathcal{Z}^m \to \mathcal{H}$$

for $\mathcal{H}$ is probably approximately correct (PAC) with respect to $\mathcal{D}$ if $\forall \epsilon, \delta \in (0,1) \exists m_0 \in \mathbb{N} \forall m \geq m_0$ and $\forall \mathbb{D} \in \mathcal{D}$

$$\mathbb{D}^m \left(\{ z \in \mathcal{Z}^m : \mathrm{er}_{\mathbb{R}} \left(\mathcal{A}(z) - \mathrm{opt}_{\mathbb{D}}(\mathcal{H}) \leq \epsilon \right) \} \right) \geq 1 - \delta.$$

$\mathcal{H}$ is PAC learnable if $\exists \mathcal{A}$ a PAC learning function for $\mathcal{H}$.

This may be refined with extra probability theory information on the guessing set: $C \in \Sigma_{\mathcal{Z}} \ldots$

Sampling

To estimate the error in guessing one may use the sample error of h on a sample $\mathsf{z} = (z_1, \ldots, z_m) \in \mathcal{Z}^m$:

$$\hat{\mathrm{er}}_{\mathsf{z}} = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{\mathcal{Z} \setminus \Gamma(h)}(z_i),$$

a measurable map.

Minimizing Sample Error (SEM)

A goal is then to choose a learning function $\mathcal{A}$ such that

$$\hat{\mathrm{er}}\left(\mathcal{A}\left(\mathrm{z}\right)\right) = \min_{h \in \mathcal{H}} \hat{\mathrm{er}}_{\mathrm{z}}(h).$$

This is sample error minimization (SEM).

Part 2

What logic (model theory) can say about learning theory

Logic and learning

Regarding ML and model theory, is the connection with definable sets the fact that in ML we are learning a "definable" function from the data? If so, I think it's important to keep in mind that in ML we are trying to approximate the "true", unknown function with an element from a (pre-selected) set of approximating functions. Is there a notion in model theory of an "approximately definable set"?

Correspondence with M. Ettinger

Definable Sets

In Model Theory, a set D is definable in a structure $\mathfrak{A}$ if there exists a formula $\varphi(x)$ such that $D = \varphi(\mathfrak{A}) = \{a \in A | \mathfrak{A} \models \varphi[a]\}$.

Definable Sets

In Model Theory, a set D is definable in a structure $\mathfrak{A}$ if there exists a formula $\varphi(x)$ such that $D = \varphi(\mathfrak{A}) = \{a \in A | \mathfrak{A} \models \varphi[a]\}$.
D is then the **locus** of a formula φ, in $\mathfrak{A}$.

Definable Sets

In Model Theory, a set D is definable in a structure $\mathfrak{A}$ if there exists a formula $\varphi(x)$ such that $D = \varphi(\mathfrak{A}) = \{a \in A | \mathfrak{A} \models \varphi[a]\}$.
D is then the **locus** of a formula φ, in $\mathfrak{A}$.
Classical examples of definable sets in a field include affine varieties: the elliptic curve given by

$$y^2 = x^3 + ax + bc + c$$

Definable Sets

In Model Theory, a set D is definable in a structure $\mathfrak{A}$ if there exists a formula $\varphi(x)$ such that $D = \varphi(\mathfrak{A}) = \{a \in A | \mathfrak{A} \models \varphi[a]\}$.
D is then the **locus** of a formula φ, in $\mathfrak{A}$.
Classical examples of definable sets in a field include affine varieties: the elliptic curve given by

$$y^2 = x^3 + ax + bc + c$$

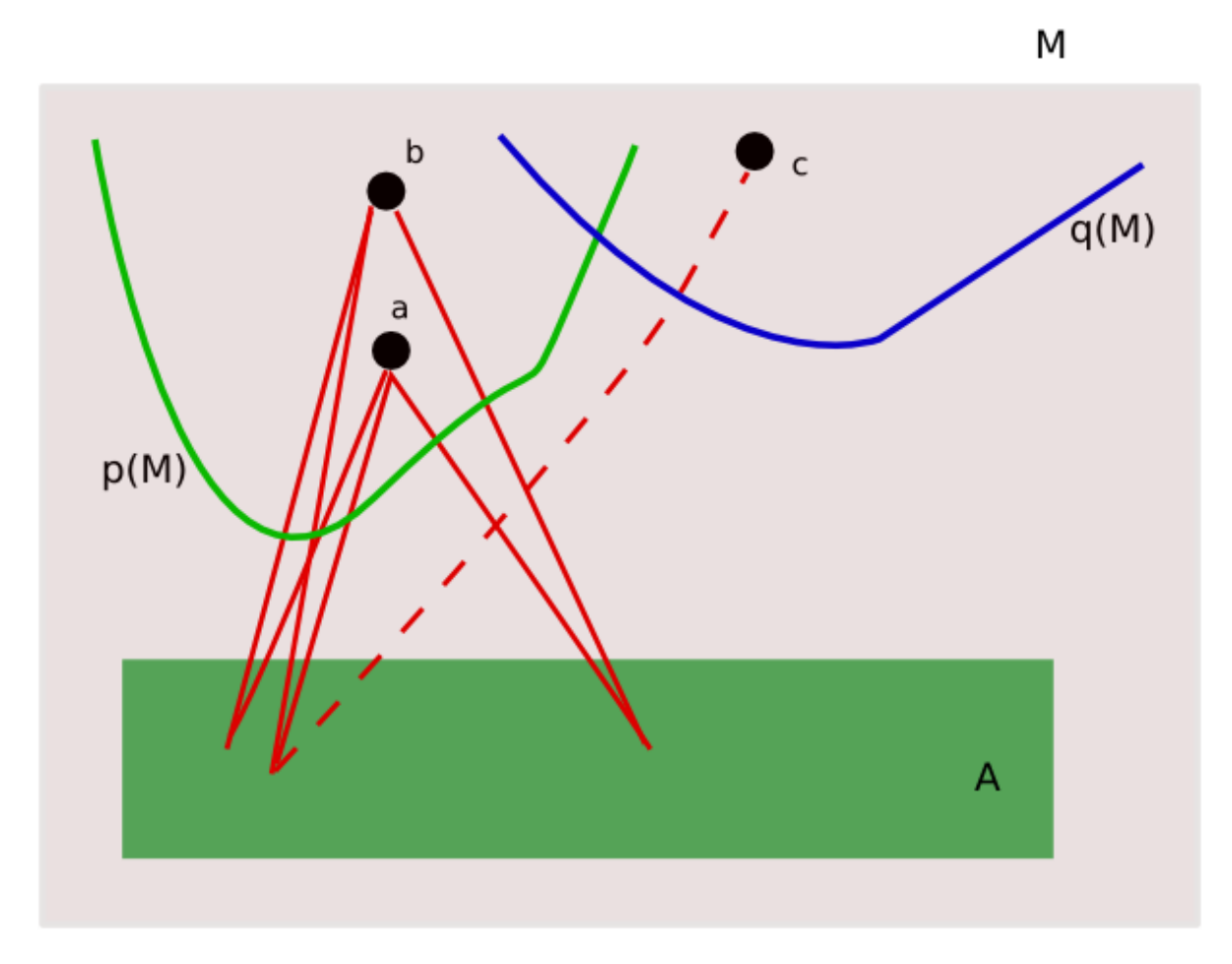
can be understood as the definable set D_C over $\langle \mathbb{C}, +, \cdot, a, b, c \rangle$,

$$D_C = \varphi_C(\mathbb{C}, a, b, c) = \{(x, y) \in \mathbb{C}^2 | \varphi_C(x, y, a, b, c)\},$$

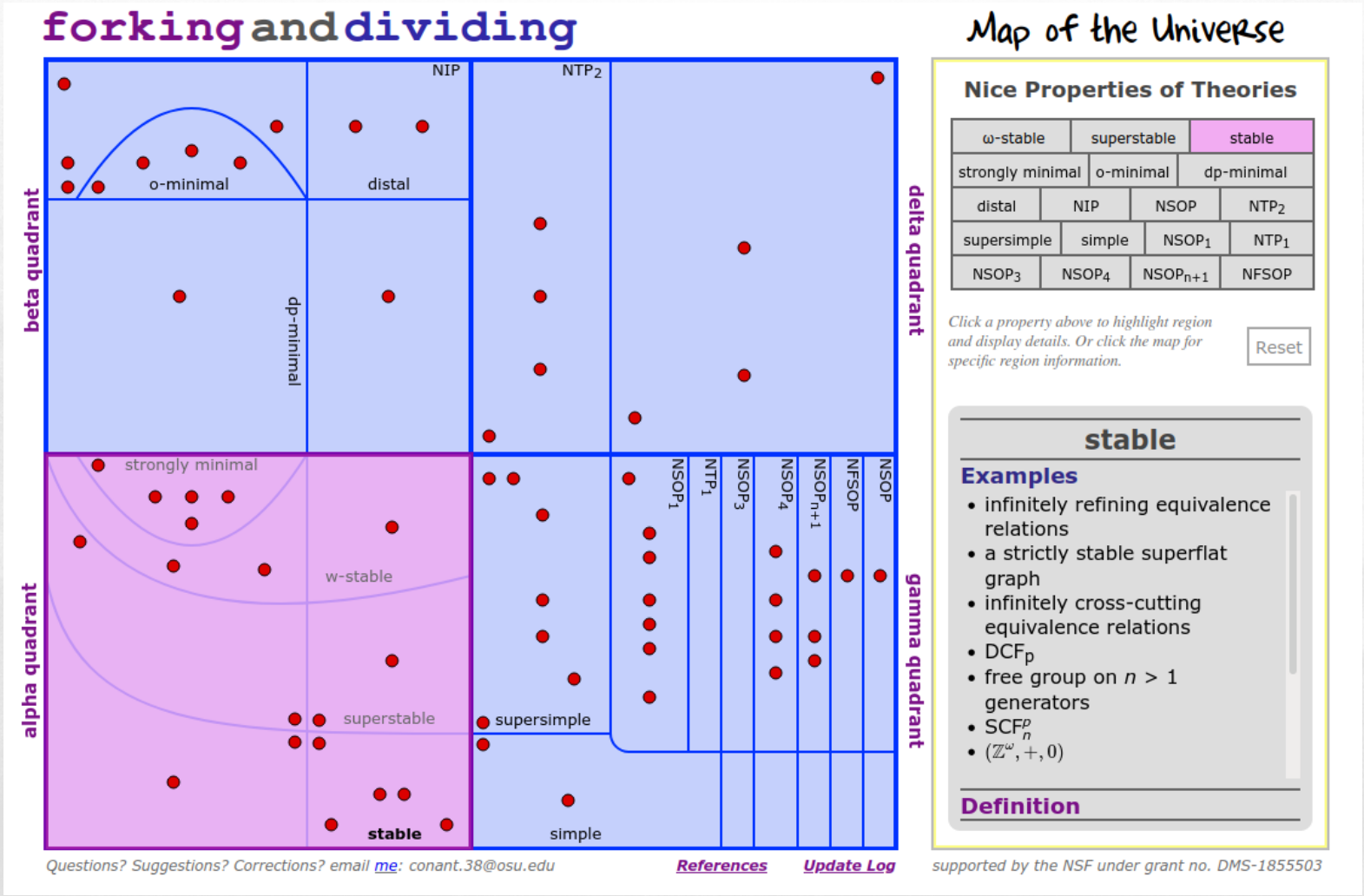
where $\varphi_C(x, y, a, b, c)$ is the formula $y^2 = x^3 + ax + bc + c$.

More complex. . .

Also, infinitely many formulas (forming "types") - like "solving simultaneously many equations":



A map of the universe: forkinganddividing.com



Theories - Dividing lines - Areas of similarity

Shattering - VC dimension

Given $\mathcal{C} \subseteq \mathcal{X}$, the hypothesis space $\mathcal{H}$ shatters $\mathcal{C}$ if

$$\{h \restriction A : h \in \mathcal{H}\} = \{0,1\}^{\mathcal{C}}.$$

Shattering - VC dimension

Given $\mathcal{C} \subseteq \mathcal{X}$, the hypothesis space $\mathcal{H}$ shatters $\mathcal{C}$ if

$$\{h \restriction A : h \in \mathcal{H}\} = \{0,1\}^{\mathcal{C}}.$$

This roughly means that the complexity of hypotheses is exponential. If $\mathcal{H}$ can**not** shatter sets of arbitrarily large size, we say that $\mathcal{H}$ has finite VC dimension.

(Vapnik-Červonenkis, 1968)

Warmuth et al.

A crucial theorem of statistical learning:

(Blumer, Ehrenfeucht, Haussler, Warmuth - 1989)

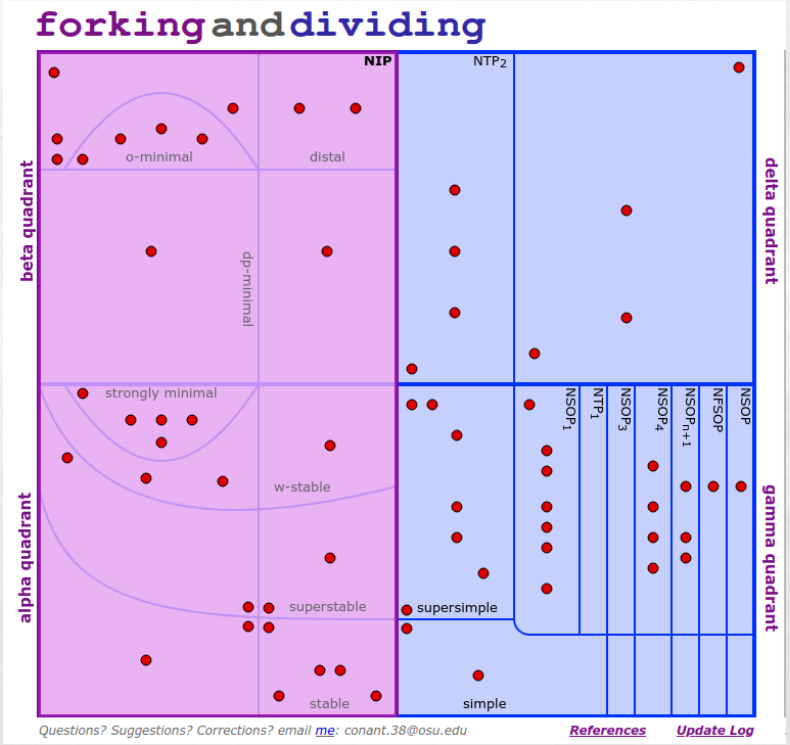
A hypothesis space $\mathcal{H}$ is PAC learnable with respect to a set of distributions $\mathcal{D}$ **iff** $\mathcal{H}$ has finite VC dimension.

This is fundamental. It relates a statistical/probabilistic definition with a purely discrete combinatorial one.

Laskowski

Moreover, Laskowski proved in 1992 that the following are equivalent (for a language $\mathcal{L}$ and an $\mathcal{L}$-structure $\mathcal{M}$):

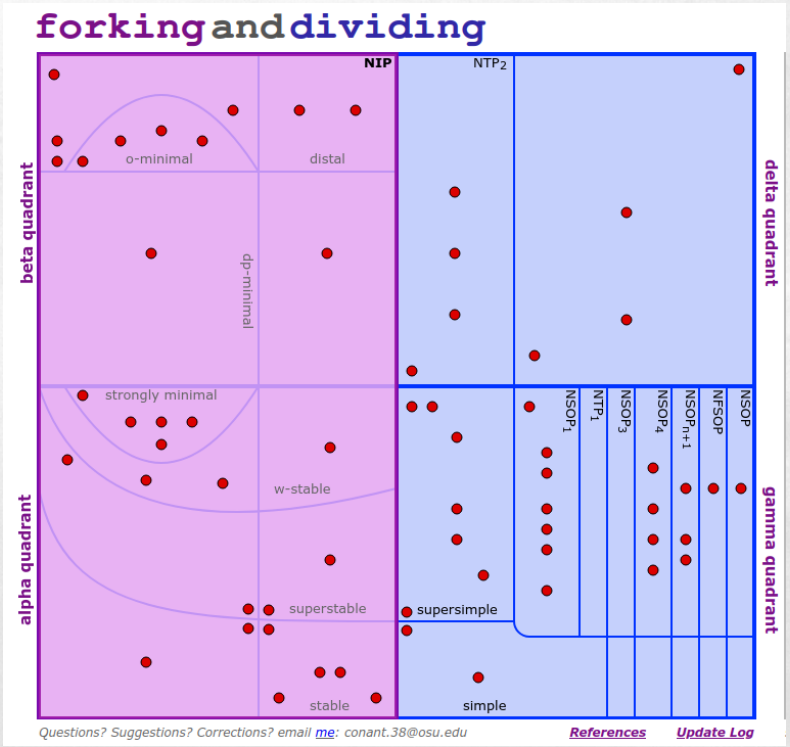
- $\mathcal{M}$ is in the NIP zone of the map,
- The hypothesis space $\mathcal{H}^{\varphi}$ has finite VC dimension, for any formula $\varphi(x, p)$.



Laskowski

Moreover, Laskowski proved in 1992 that the following are equivalent (for a language $\mathcal{L}$ and an $\mathcal{L}$-structure $\mathcal{M}$):

- $\mathcal{M}$ is in the NIP zone of the map,
- The hypothesis space $\mathcal{H}^{\varphi}$ has finite VC dimension, for any formula $\varphi(x, p)$.

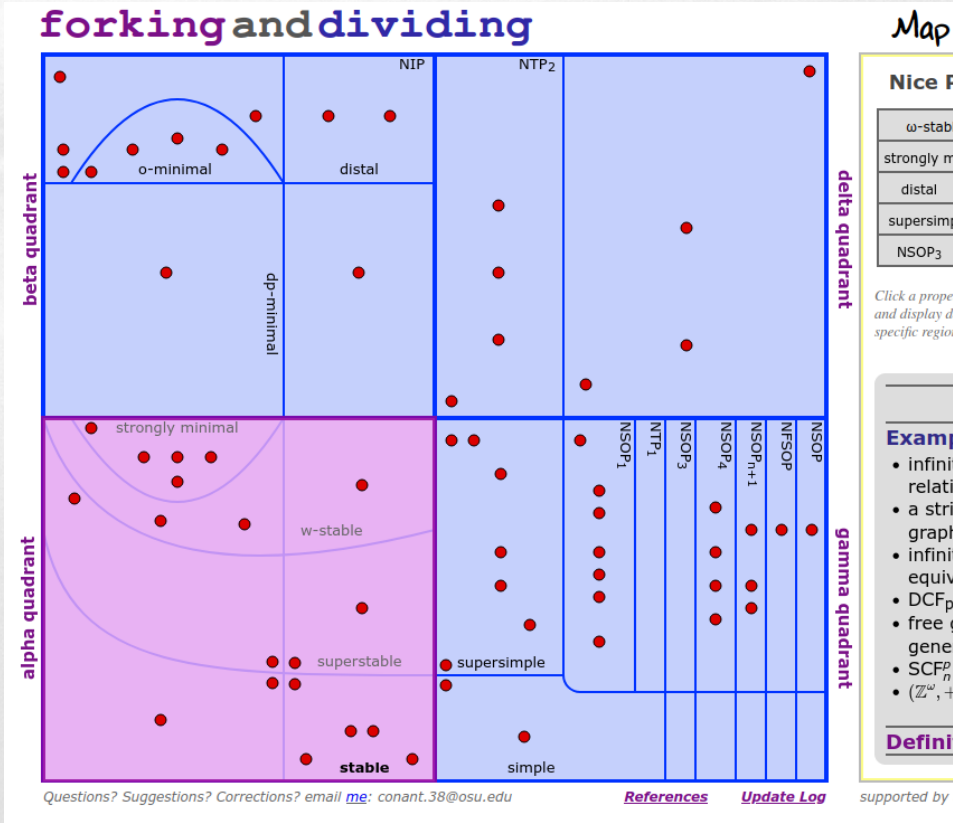


$$\mathcal{H}^{\varphi} = \{\mathbb{1}_{\varphi(\mathcal{M},w):w\in M^{\ell}}\},\ \varphi(\mathcal{M}, w) = \{a \in M^{n} : \mathcal{M} \models \varphi(a, w)\}.$$

Chase-Freitag

The following are equivalent:

- $\mathcal{M}$ is in the **stable** zone of the map,
- The hypothesis space $\mathcal{H}$ is **online learnable.**



Malliaris-Moran

In a series of papers, Malliaris and Moran have established the equivalence

$$\text{Online Learnability} = \text{Private Learnability}$$

Work by Kaplan using model theory has also helped understand better the notion of compression in learnability theory.

Thank you! (But let us try to start a conversation on these and hopefully other topics relating ML and Model Theory!)