Applications of NIP in Statistical Learning Theory:

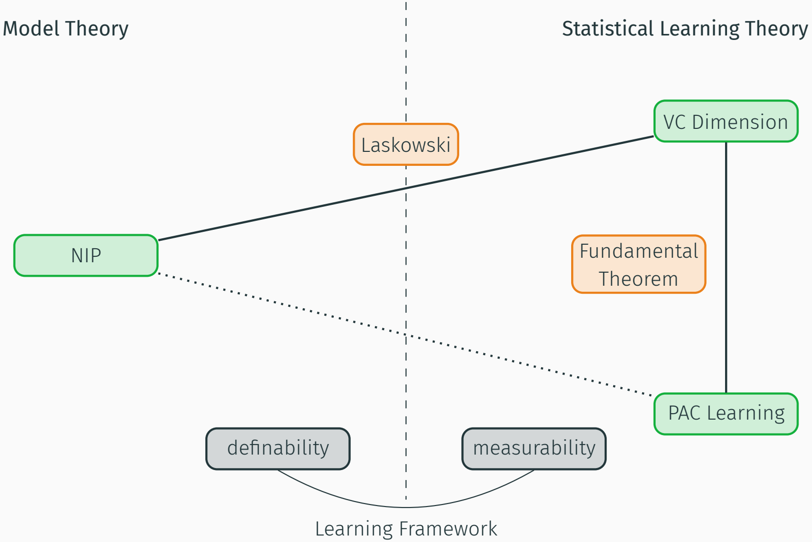
Measurability Aspects

Laura Wirth
University of Konstanz

Z75: Geometry from the model theorist's point of view
Mathematical Institute and St. Hilda's College, University of Oxford, September 2024

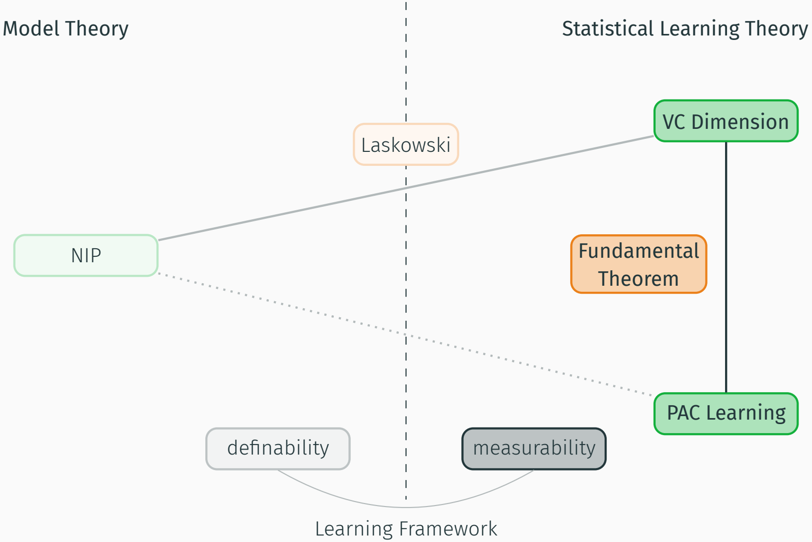
Bridge of Sighs, Oxford

Overview



Model Theory
Statistical Learning Theory
VC Dimension
Laskowski
NIP
Fundamental Theorem
PAC Learning
definability
measurability
Learning Framework

Statistical Learning Theory



Learning Problem

Ingredients of a Learning Problem.

- $\emptyset \neq \mathcal{X}$ — instance space
- $\mathcal{Z} = \mathcal{X} \times \{0,1\}$ — sample space
- $\emptyset \neq \mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ — hypothesis space
- $\Sigma_{\mathcal{Z}}$ — σ–algebra on $\mathcal{Z}$ with $\mathcal{P}_{\mathrm{fin}}(\mathcal{Z}) \subseteq \Sigma_{\mathcal{Z}}$
- $\mathcal{D}$ — set of distributions on $(\mathcal{Z}, \Sigma_{\mathcal{Z}})$

Assumption.

For any hypothesis $h \in \mathcal{H}$ we have

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

Learning from Examples – Basic Procedure

Using an arbitrary distribution $\mathbb{D} \in \mathcal{D}$, a sequence of iid samples from $\mathcal{Z}$ is generated:

$$\boldsymbol{z} = ((x_1, y_1), \dots, (x_m, y_m)).$$

These samples provide the input data for a learning function $\mathcal{A}$ that determines a hypothesis $h = \mathcal{A}(\boldsymbol{z})$ in $\mathcal{H}$.

S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*, (Cambridge University Press, Cambridge, 2014).

Learning from Examples – Goal

The goal is to minimize the (true) error of h given by

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

More precisely, we want to achieve an error that is close to

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

S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*, (Cambridge University Press, Cambridge, 2014).

PAC Learning

Definition.

A learning function

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

for $\mathcal{H}$ is said to be probably approximately correct (PAC) (with respect to $\mathcal{D}$) if it satisfies the following condition:

$$\forall \varepsilon, \delta \in (0,1)\ \exists m_0 \in \mathbb{N}\ \forall m \geq m_0\ \forall \mathbb{D} \in \mathcal{D}:$$
$$\mathbb{D}^m(\{\mathbf{z} \in \mathcal{Z}^m \mid \mathrm{er}_{\mathbb{D}}(\mathcal{A}(\mathbf{z})) - \mathrm{opt}_{\mathbb{D}}(\mathcal{H}) \leq \varepsilon\}) \geq 1 - \delta.$$

L. G. Valiant, ‘A Theory of the Learnable’, *Comm. ACM* **27** (1984) 1134–1142.

PAC Learning

Definition.

A learning function

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

for $\mathcal{H}$ is said to be probably approximately correct (PAC) (with respect to $\mathcal{D}$) if it satisfies the following condition:

$$\forall \varepsilon, \delta \in (0,1)\ \exists m_0 \in \mathbb{N}\ \forall m \geq m_0\ \forall \mathbb{D} \in \mathcal{D}:$$
$$\mathbb{D}^m(\{\mathbf{z} \in \mathcal{Z}^m \mid \mathrm{er}_{\mathbb{D}}(\mathcal{A}(\mathbf{z})) - \mathrm{opt}_{\mathbb{D}}(\mathcal{H}) \leq \varepsilon\}) \geq 1 - \delta.$$

The hypothesis space $\mathcal{H}$ is said to be PAC learnable if there exists a learning function for $\mathcal{H}$ that is PAC.

PAC Learning – refined

Definition.
A learning function

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

for $\mathcal{H}$ is said to be probably approximately correct (PAC) (with respect to $\mathcal{D}$) if it satisfies the following condition:

$$\forall \varepsilon, \delta \in (0,1)\ \exists m_0 \in \mathbb{N}\ \forall m \geq m_0\ \forall \mathbb{D} \in \mathcal{D}\ \exists C \in \mathbf{\Sigma}_{\mathcal{Z}}^m :$$
$$C \subseteq \{\mathbf{z} \in \mathcal{Z}^m \mid \mathrm{er}_{\mathbb{D}}(\mathcal{A}(\mathbf{z})) - \mathrm{opt}_{\mathbb{D}}(\mathcal{H}) \leq \varepsilon\}$$
$$\text{and } \mathbb{D}^m(C) \geq 1 - \delta.$$

The hypothesis space $\mathcal{H}$ is said to be PAC learnable if there exists a learning function for $\mathcal{H}$ that is PAC.

Sample Error

The sample error of h on a multi-sample $\boldsymbol{z} = (z_1, \ldots, z_m) \in \mathcal{Z}^m$ given by

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

provides a useful estimate for the true error.

Remark.

The map

$$\mathcal{Z}^m \to \left\{ \tfrac{k}{m} \;\middle|\; k \in \{0, 1, \ldots, m\} \right\}, \; \boldsymbol{z} \mapsto \hat{\mathrm{er}}_{\boldsymbol{z}}(h)$$

is $\Sigma_{\mathcal{Z}}^{m}$–measurable.

S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*, (Cambridge University Press, Cambridge, 2014).

A Simple Learning Principle

The sample error of h on a multi-sample $\boldsymbol{z} = (z_1, \ldots, z_m) \in \mathcal{Z}^m$ given by

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

provides a useful estimate for the true error.

Sample Error Minimization (SEM).

Choose a learning function $\mathcal{A}$ such that

$$\hat{\mathrm{er}}_{\mathbf{z}}(\mathcal{A}(\boldsymbol{z})) = \min_{h \in \mathcal{H}} \hat{\mathrm{er}}_{\mathbf{z}}(h)$$

for any multi-sample $\boldsymbol{z}$.

VC Dimension

Definition.

Given $A \subseteq \mathcal{X}$, we say that $\mathcal{H}$ shatters A if

$$\{h\restriction_A \mid h \in \mathcal{H}\} = \{0,1\}^A.$$

If $\mathcal{H}$ cannot shatter sets of arbitrarily large size, then we say that $\mathcal{H}$ has finite VC dimension.

V. N. Vapnik and A. Ja. Červonenkis, 'Uniform Convergence of Frequencies of Occurrence of Events to Their Probabilities', *Dokl. Akad. Nauk SSSR* **181** (1968) 781–783 (Russian), *Sov. Math., Dokl.* **9** (1968) 915–918 (English).

Fundamental Theorem of Statistical Learning

The following result is due to Blumer, Ehrenfeucht, Haussler and Warmuth 1989.

Theorem.

Under certain measurability conditions, a hypothesis space $\mathcal{H}$ is PAC learnable with respect to a set $\mathcal{D}$ of distributions if and only if its VC dimension is finite.

A. Blumer, A. Ehrenfeucht, D. Haussler and M. K. Warmuth, 'Learnability and the Vapnik-Chervonenkis dimension', *J. Assoc. Comput. Mach.* **36** (1989) 929–965.

Definition.

A hypothesis space $\mathcal{H}$ is called well-behaved (with respect to $\mathcal{D}$) if it satisfies the following conditions:

- $\Gamma(h) \in \Sigma_{\mathcal{Z}}$ for any $h \in \mathcal{H}$.
- The map

$$U \colon \mathcal{Z}^m \to [0,1], \; \mathbf{z} \mapsto \sup_{h \in \mathcal{H}} \left| \mathrm{er}_{\mathbb{D}}(h) - \hat{\mathrm{er}}_{\mathbf{z}}(h) \right|$$

  is $\Sigma_{\mathcal{Z}}^m$–measurable for any $m \geq m_{\mathcal{H}}$ and any $\mathbb{D} \in \mathcal{D}$.
- The map

$$V \colon \mathcal{Z}^{2m} \to [0,1], \; (\mathbf{z}, \mathbf{z}') \mapsto \sup_{h \in \mathcal{H}} \left| \hat{\mathrm{er}}_{\mathbf{z}'}(h) - \hat{\mathrm{er}}_{\mathbf{z}}(h) \right|$$

  is $\Sigma_{\mathcal{Z}}^{2m}$–measurable for any $m \geq m_{\mathcal{H}}$.

Pathological Hypothesis Spaces

Fundamental Theorem.

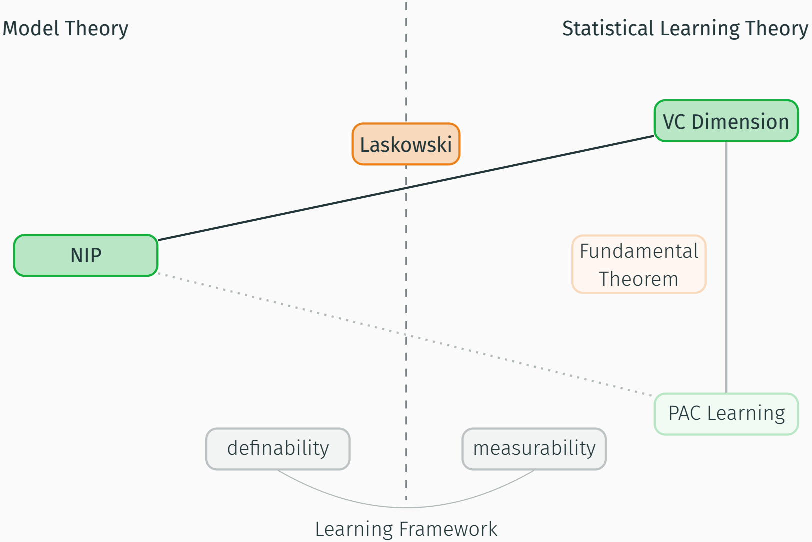
Under certain measurability conditions, a hypothesis space $\mathcal{H}$ is PAC learnable with respect to a set $\mathcal{D}$ of distributions if and only if its VC dimension is finite.

Open Question.

Are there a hypothesis space $\mathcal{H}$ with finite VC dimension and a set $\mathcal{D}$ of distributions such that $\mathcal{H}$ is not PAC learnable with respect to $\mathcal{D}$?

Note: Such a hypothesis space would not be well-behaved.

NIP and VC Dimension



Definable Hypothesis Spaces

Definiton.

Let $\mathcal{L}$ be a language, let $\mathcal{M}$ be an $\mathcal{L}$–structure and let $\varphi(x_1, \ldots, x_n; p_1, \ldots, p_\ell)$ be an $\mathcal{L}$–formula. For any $\mathbf{w} \in M^\ell$, set

$$\varphi(\mathcal{M}, \mathbf{w}) = \{\mathbf{a} \in M^n \mid \mathcal{M} \models \varphi(\mathbf{a}; \mathbf{w})\}.$$

Then the hypothesis space $\mathcal{H}^\varphi \subseteq \{0,1\}^{M^n}$ is given by

$$\mathcal{H}^\varphi := \left\{ \mathbb{1}_{\varphi(\mathcal{M};\mathbf{w})} \,\middle|\, \mathbf{w} \in M^\ell \right\}.$$

Further, given a non-empty set $\mathcal{X} \subseteq M^n$ that is definable over $\mathcal{M}$, the hypothesis space $\mathcal{H}^\varphi_{\mathcal{X}} \subseteq \{0,1\}^{\mathcal{X}}$ is given by

$$\mathcal{H}^\varphi_{\mathcal{X}} := \{h{\restriction}_{\mathcal{X}} \mid h \in \mathcal{H}^\varphi\}.$$

NIP and VC Dimension

The following result is due to Laskowski 1992.

Proposition.

Let $\mathcal{L}$ be a language and let $\mathcal{M}$ be an $\mathcal{L}$–structure. Then the following conditions are equivalent:

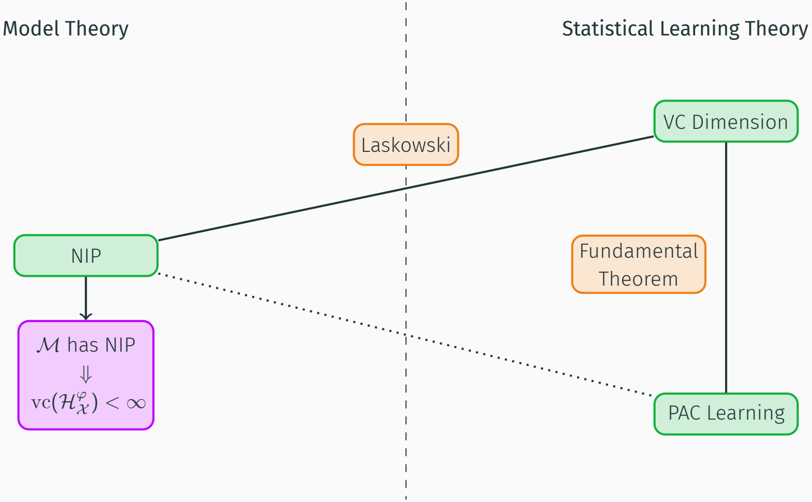
(1) $\mathcal{M}$ has NIP.

(2) The hypothesis space $\mathcal{H}^{\varphi}$ has finite VC dimension for any $\mathcal{L}$–formula $\varphi(x; p)$.

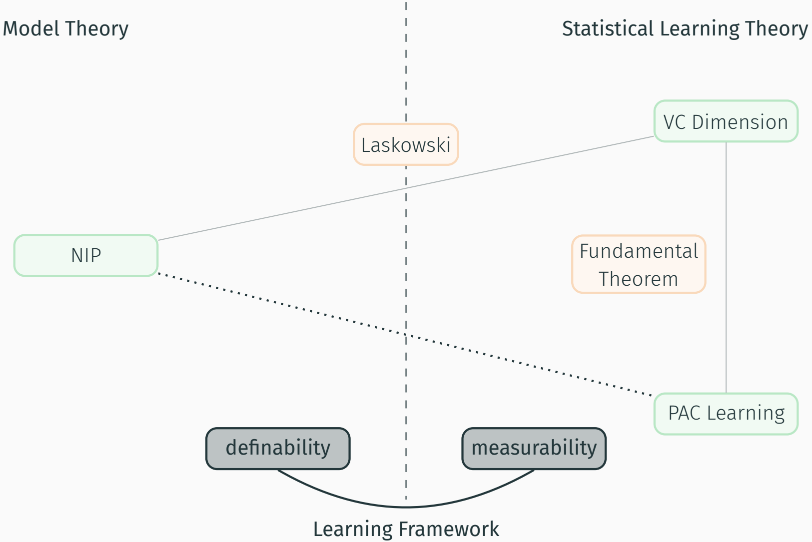
(3) The hypothesis space $\mathcal{H}^{\varphi}_{\mathcal{X}}$ has finite VC dimension for any $\mathcal{L}$–formula $\varphi(x; p)$ and any non-empty set $\mathcal{X}$ definable over $\mathcal{M}$.

M. C. Laskowski, 'Vapnik–Chervonenkis classes of definable sets', *J. Lond. Math. Soc.* **45** (1992) 377–384.

Jumping to Conclusions



Developing a Model-Theoretic Learning Framework



Borel σ–Algebra

Given $k \in \mathbb{N}$, the Borel σ–algebra $\mathcal{B}(\mathbb{R}^k)$ of $\mathbb{R}^k$ is the smallest σ–algebra containing all open sets in $\mathbb{R}^k$.
For $\mathcal{Y} \subseteq \mathbb{R}^k$, we consider the trace σ–algebra given by

$$\mathcal{B}(\mathcal{Y}) := \{B \cap \mathcal{Y} \mid B \in \mathcal{B}(\mathbb{R}^k)\}.$$

Given $k \in \mathbb{N}$, the Borel σ–algebra $\mathcal{B}(\mathbb{R}^k)$ of $\mathbb{R}^k$ is the smallest σ–algebra containing all open sets in $\mathbb{R}^k$.
For $\mathcal{Y} \subseteq \mathbb{R}^k$, we consider the trace σ–algebra given by

$$\mathcal{B}(\mathcal{Y}) := \{B \cap \mathcal{Y} \mid B \in \mathcal{B}(\mathbb{R}^k)\}.$$

Set $\mathcal{L}_{\mathrm{or}} := \{+, \cdot, -, 0, 1, <\}$ and $\mathbb{R}_{\mathrm{or}} := (\mathbb{R}, +, \cdot, -, 0, 1, <)$.

Lemma.

Let $\mathcal{L}$ be a language expanding $\mathcal{L}_{\mathrm{or}}$, let $\mathcal{R}$ be an o-minimal $\mathcal{L}$–expansion of $\mathbb{R}_{\mathrm{or}}$, let $\varphi(x_1, \ldots, x_m; p_1, \ldots, p_\ell)$ be an $\mathcal{L}$–formula and let $w \in \mathbb{R}^\ell$. Then $\varphi(\mathcal{R}; w) \in \mathcal{B}(\mathbb{R}^n)$.

M. Karpinski and A. Macintyre, 'Approximating Volumes and Integrals in o-Minimal and p-Minimal Theories', *Connections between model theory and algebraic and analytic geometry* (ed. Macintyre), Quad. Mat. 6 (2000) 149–177.

Theorem.

Let

- $\mathcal{L}$ be a language expanding $\mathcal{L}_{\mathrm{or}}$,
- $\mathcal{R}$ be an o-minimal $\mathcal{L}$–expansion of $\mathbb{R}_{\mathrm{or}}$,
- $\mathcal{X} \subseteq \mathbb{R}^n$ be a non-empty set that is definable over $\mathcal{R}$,
- $\varphi(x_1, \ldots, x_n; p_1, \ldots, p_\ell)$ be an $\mathcal{L}$–formula,
- $\Sigma_{\mathcal{Z}}$ be a σ–algebra on $\mathcal{Z} = \mathcal{X} \times \{0, 1\}$ with $\mathcal{B}(\mathcal{Z}) \subseteq \Sigma_{\mathcal{Z}}$, and
- $\mathcal{D}$ be a set of distributions on $(\mathcal{Z}, \Sigma_{\mathcal{Z}})$ such that $(\mathcal{Z}^m, \Sigma_{\mathcal{Z}}^m, \mathbb{D}^m)$ is a complete probability space for any $\mathbb{D} \in \mathcal{D}$ and any $m \in \mathbb{N}$.

Then $\mathcal{H}_{\mathcal{X}}^{\varphi}$ is PAC learnable with respect to $\mathcal{D}$.

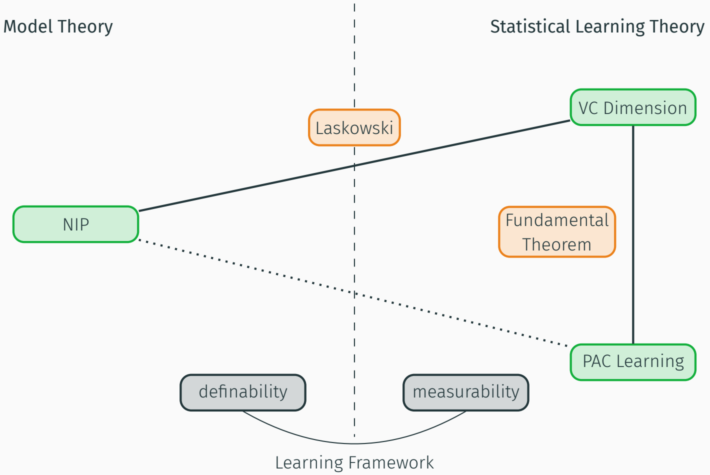
M. Karpinski and A. Macintyre, 'Approximating Volumes and Integrals in o-Minimal and p-Minimal Theories', *Connections between model theory and algebraic and analytic geometry* (ed. Macintyre), Quad. Mat. 6 (2000) 149–177.

Proof Sketch.

- O-minimality implies NIP.
- Thus, $\mathcal{H}_{\mathcal{X}}^{\varphi}$ has finite VC dimension.
- Aim: Apply Fundamental Theorem.
- To this end: Verify well-behavedness.
- $\Gamma(h) \in \Sigma_{\mathcal{Z}}$ for any $h \in \mathcal{H}_{\mathcal{X}}^{\varphi}$.
- Technical analysis and application of Pollard's arguments regarding measurability of suprema establish measurability of the maps U and V.

D. Pollard, *Convergence of Stochastic Processes*, Springer Ser. Stat. (Springer, New York, 1984).

Summary



Advertisement: Lothar Sebastian Krapp and Laura Wirth, 'Measurability in the Fundamental Theorem of Statistical Learning', in preparation.

Appendix

NIP

Notation.

$[m] := \{1, \ldots, m\}$ for $m \in \mathbb{N}$.

Definition.

Let $\mathcal{L}$ be a language and let $\mathcal{M}$ be an $\mathcal{L}$–structure.
A (partitioned) $\mathcal{L}$–formula $\varphi(x_1, \ldots, x_n; p_1, \ldots, p_\ell)$ has NIP over $\mathcal{M}$ if there is $m \in \mathbb{N}$ such that for any object set $\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m\} \subseteq M^n$ and any parameter set $\{\boldsymbol{w}_I \mid I \subseteq [m]\} \subseteq M^\ell$, there is some $J \subseteq [m]$ such that

$$\mathcal{M} \not\models \underbrace{\bigwedge_{i \in J} \varphi(\boldsymbol{a}_i; \boldsymbol{w}_J) \wedge \bigwedge_{i \in [m] \setminus J} \neg\varphi(\boldsymbol{a}_i; \boldsymbol{w}_J)}_{\varphi(\boldsymbol{a}_i; \boldsymbol{w}_J) \text{ is true iff } i \in J}.$$

S. Shelah, ‘Stability, the f.c.p., and superstability; model theoretic properties of formulas in first order theory’, *Ann. Math. Logic* **3** (1971) 271–362.

NIP

Definition.

Let $\mathcal{L}$ be a language and let $\mathcal{M}$ be an $\mathcal{L}$–structure.
A (partitioned) $\mathcal{L}$–formula $\varphi(x_1, \ldots, x_n; p_1, \ldots, p_\ell)$ has NIP over $\mathcal{M}$ if there is $m \in \mathbb{N}$ such that for any object set $\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m\} \subseteq M^n$ and any parameter set $\{\boldsymbol{w}_I \mid I \subseteq [m]\} \subseteq M^\ell$, there is some $J \subseteq [m]$ such that

$$\mathcal{M} \not\models \underbrace{\bigwedge_{i \in J} \varphi(\boldsymbol{a}_i; \boldsymbol{w}_J) \wedge \bigwedge_{i \in [m] \setminus J} \neg\varphi(\boldsymbol{a}_i; \boldsymbol{w}_J)}_{\varphi(\boldsymbol{a}_i; \boldsymbol{w}_J) \text{ is true iff } i \in J}.$$

The $\mathcal{L}$–structure $\mathcal{M}$ has NIP if every $\mathcal{L}$–formula has NIP over $\mathcal{M}$.

S. Shelah, 'Stability, the f.c.p., and superstability; model theoretic properties of formulas in first order theory', *Ann. Math. Logic* **3** (1971) 271–362.

NIP Formulas and VC Dimension

The following result is due to Laskowski 1992.

Lemma.

Let $\mathcal{L}$ be a language, let $\mathcal{M}$ be an $\mathcal{L}$–structure and let $\varphi(x_1, \ldots, x_n; p_1, \ldots, p_\ell)$ be an $\mathcal{L}$–formula. Then φ has NIP over $\mathcal{M}$ if and only if the hypothesis space $\mathcal{H}^\varphi$ has finite VC dimension.

M. C. Laskowski, 'Vapnik–Chervonenkis classes of definable sets', *J. Lond. Math. Soc.* **45** (1992) 377–384.

Sufficient Conditions for Well-Behavedness

Remark.

Sufficient conditions for the measurability of the maps U and V:

- $\mathcal{X}$ is countable.
- $\mathcal{H}$ is countable.
- $\mathcal{H}$ is universally separable.

Definition.

The hypothesis space $\mathcal{H}$ is called universally separable if there exists a countable subset $\mathcal{H}_0 \subseteq \mathcal{H}$ such that for any $h \in \mathcal{H}$ there exists a sequence $\{h_n\}_{n\in\mathbb{N}} \subseteq \mathcal{H}_0$ converging pointwise to h.

Book Recommendations

- M. Anthony and P. L. Bartlett, *Neural Network Learning: Theoretical Foundations*, (Cambridge University Press, Cambridge, 1999).
- S. Ben-David and S. Shalev-Shwartz, *Understanding Machine Learning: From Theory to Algorithms*, (Cambridge University Press, Cambridge, 2014).
- M. Vidyasagar, *Learning and Generalisation: With Applications to Neural Networks*, Commun. Control Eng. (Springer, London, 2003).