Generalizing Rademacher Complexity

joint work with (ongoing) Henry Towsner, (recently) Michael Benedikt

Aaron Anderson

Rademacher Instructor, UPenn

Chania, June 4, 2026

Function Classes from Formulas

We will work with a class $\mathcal{H}$ of functions $X \to [0,1]$.

Example

- Fix a structure $\mathcal{M}$ and a formula $\phi(x; y)$ (with x, y possibly tuples).
- Given $b \in M^y$, let $h_b : M^x \to \{0,1\}$ be the characteristic function for $\phi(M; b)$.
- Let $X = M^x$, $\mathcal{H}_\phi = \{h_b : b \in M^y\}$.

More Function Classes from Continuous Logic

In continuous logic, a formula $\phi(x)$ is interpreted over a *metric structure* $\mathcal{M}$ as a function $M^x \to [0,1]$.

Example

In a metric structure, we can let $\mathcal{H}_\phi = \{\phi^{\mathcal{M}}(x;b) : b \in M^y\}$, no extra steps needed.

Uniform Law of Large Numbers

Theorem (Vapnik-Chervonenkis)

The class $\mathcal{H}$ has finite VC-dimension if and only if

- *for every suitable measure μ on X,*
- *for all $\varepsilon > 0, \delta > 0$, there is $n \in \mathbb{N}$ such that*
- *if $\bar{x} = (x_1, \ldots, x_n)$ consists of n i.i.d. random samples from μ,*

$$\mathbb{P}_{\bar{x}}\left[\sup_{h \in \mathcal{H}} \left|\frac{1}{n}\sum_{i=1}^{n} h_b(x_i) - \mathbb{E}_\mu[h_b]\right| > \varepsilon\right] \leq \delta.$$

The original result was for sets/functions $X \to \{0, 1\}$, but the real-valued version is well-known by now.

Symmetrization

Bound the average error by

- introducing another sample,
- swapping pairs x_i, y_i at random,
- decoupling with the triangle inequality.

$$\begin{aligned}
&\mathbb{E}_{\bar{x}}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n} h(x_i) - \mathbb{E}_\mu[h]\right|\right] \\
\leq\, &\mathbb{E}_{\bar{x},\bar{y},\sigma}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n} \sigma_i(h(x_i) - h(y_i))\right|\right] \\
\leq\, &2\mathbb{E}_{\bar{x},\sigma}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n} \sigma_i h(x_i)\right|\right]
\end{aligned}$$

Symmetrization

Bound the average error by

- introducing another sample,
- swapping pairs x_i, y_i at random,
- decoupling with the triangle inequality.

$$
\begin{aligned}
&\mathbb{E}_{\bar{x}}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n} h(x_i) - \mathbb{E}_{\mu}[h]\right|\right] \\
&\leq \mathbb{E}_{\bar{x},\bar{y},\sigma}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n} \sigma_i(h(x_i) - h(y_i))\right|\right] \\
&\leq 2\boxed{\mathbb{E}_{\bar{x},\sigma}\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n} \sigma_i h(x_i)\right|\right]} \leftarrow \mathcal{R}_{\mathcal{H},\mu}(n)
\end{aligned}
$$

Rademacher Complexities

Given $\bar{x}$, define the *empirical Rademacher complexity*:

$$\mathcal{R}_{\mathcal{H}}(\bar{x}) = \mathbb{E}_{\sigma}\left[\sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(x_i) \right|\right].$$

Taking a sup, we get the *worst-case Rademacher complexity*:

$$\mathcal{R}_{\mathcal{H}}(n) = \sup_{\bar{x}} \mathcal{R}_{\mathcal{H}}(\bar{x}),$$

or taking an expectation, just *Rademacher complexity*:

$$\mathcal{R}_{\mathcal{H},\mu}(n) = \mathbb{E}_{\bar{x}}\left[\mathcal{R}_{\mathcal{H}}(\bar{x})\right].$$

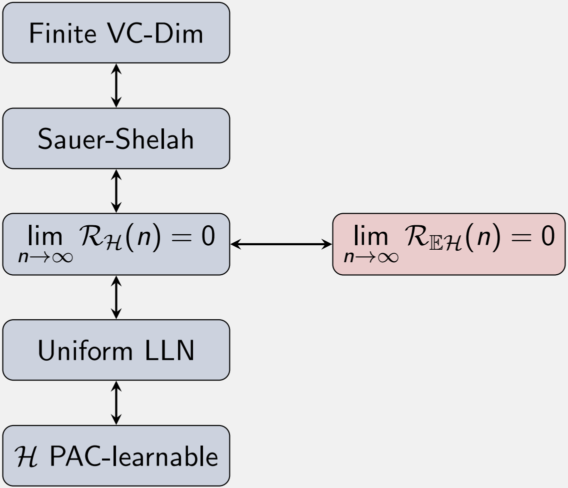
Rademacher Complexity as Discrepancy

- Let $\bar{x}$ be a specific, worst-case, or random sample.
- Flip a coin n times to split $\bar{x}$ into two random subsamples.
- How much (in expectation) does $\frac{1}{n}\sum_{i \in E} h(x_i)$ differ between the two samples, for worst-case $h \in \mathcal{H}$?

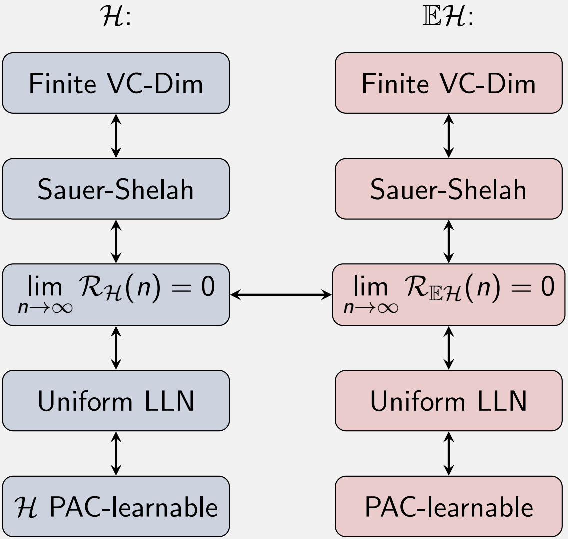
Bridge to Learnability

$$
\begin{array}{c}
\boxed{\text{Finite VC-Dim}} \\
\Updownarrow \\
\boxed{\text{Sauer-Shelah}} \\
\Updownarrow \\
\boxed{\lim_{n\to\infty} \mathcal{R}_{\mathcal{H}}(n) = 0} \\
\Updownarrow \\
\boxed{\text{Uniform LLN}} \\
\Updownarrow \\
\boxed{\mathcal{H}\ \text{PAC-learnable}}
\end{array}
$$

Bridge to Learnability and Randomization



Bridge to Learnability and Randomization



Goals (including some work in progress)

- Link VC-dimension, and even VC_k-dimension, to Rademacher complexity.
- Connect *sequential* Rademacher complexity to stability and NFOP_2.
- Derive some statistical results.
- Show all of these are preserved under randomization.

Real-Valued NIP

Definition

For $\gamma > 0$, say $\mathcal{H}$ *γ-shatters* $\bar{x} \in X^n$ when there is a witness $v \in [0,1]^n$ such that for all $\sigma \in \{-1,1\}^n$, there is $h_\sigma \in \mathcal{H}$ with

$$\forall i \in [n], \sigma_i(h_\sigma(x_i) - v_i) \geq \frac{\gamma}{2}.$$

Definition

Say $\mathcal{H}$ has

- *γ-VC-dimension* $\geq d$ when it shatters $\bar{x} \in X^d$.
- *finite VC-dimension* when $\forall \gamma > 0$, γ-VC-dim is finite.

Higher-Arity NIP

Notation

Suppose $X = \prod_{i=1}^{k} X_i$.
Given tuples $\bar{x}^i \in X_i$ for $1 \leq i \leq k$, define $\vec{x} = \bigotimes_{i=1}^{k} \bar{x}^i \in X^d$ by

$$\vec{x}_{\bar{\imath}} = (x^1_{i_1}, \ldots, x^k_{i_k}).$$

Definition

Say $\mathcal{H}$ has

- *γ-VC_k-dimension* $\geq d$ when it shatters $\bigotimes_{i=1}^{k} \bar{x}^i \in X^d$.
- *finite VC_k-dimension* when $\forall \gamma > 0$, γ-VC-dim is finite.

Sauer-Shelah

To bound Rademacher complexity, start with Sauer-Shelah.

Definition

Given $\bar{x} \in X^n$, let $\mathcal{H}(\bar{x}) = \{(h(x_1), \dots, h(x_n)) : h \in \mathcal{H}\}$.

Lemma (Sauer-Shelah)

If $\mathcal{H}$ is $\{0,1\}$-valued and has VC-dim $\leq d$, $|\mathcal{H}(\bar{x})| = O(n^d)$.

Lemma (Chernikov-Palacin-Takeuchi)

If $\mathcal{H}$ is $\{0,1\}$-valued and has VC$_k$-dim $\leq d$, there is $\varepsilon = \varepsilon(k,d)$ with

$$\left|\mathcal{H}\left(\bigotimes_{i=1}^{k} \bar{x}^i\right)\right| = O(2^{n^{k-\varepsilon}}).$$

Sauer-Shelah as Covering Number Bound

If $\mathcal{H}$ is $[0,1]$-valued, $\mathcal{H}\mid_{\bar{x}}$ is probably infinite. Bound covering numbers instead.

Definition

For $A \subseteq [0,1]^n$, let $\mathcal{N}_\gamma(A)$ be the minimum size of a γ-*cover* C:

$$\forall a \in A, \exists c \in C, |a - c|_\infty \leq \gamma.$$

Lemma (Alon, Ben-David, Cesa-Bianchi, Haussler)

If $\mathcal{H}$ has $\frac{\gamma}{4}$-VC-dimension $\leq d$,

$$\log\left(\mathcal{N}_\gamma(\mathcal{H}(\bar{x}))\right) = O\left(d(\log n)^2\right)$$

Higher Arity Covering Number Bound

Either $\mathcal{N}_\gamma(\mathcal{H}(\bar{x}))$ is small, or $\mathcal{H}$ shatters enough subtuples of $\bar{x}$ that one of them must have size d.

Theorem (Erdős)

If $I \subseteq [n]^k$ has $|I| \geq z_k(n, d+1)$, where $z_k(n, d+1) = O(n^{k-\varepsilon})$ for some $\varepsilon = \varepsilon(k, d)$, then I contains some $\prod_{i=1}^k B_i$ with $|B_i| \geq d$.

If we ask for a shattered set of size $z_k(n, d+1)$, we must shatter a *grid* of size d^k.

Lemma (A., Towsner)

If $\mathcal{H}$ has γ-VC_k-dimension $\leq d$, then for some $\varepsilon = \varepsilon(k, \gamma, d)$,

$$\log\left(\mathcal{N}_\gamma(\mathcal{H}(\otimes_{i=1}^k \bar{x}^i))\right) = O\left(n^{k-\varepsilon}\right).$$

Back to Rademacher

Definition

Given $A \subseteq [0,1]^n$, define *Rademacher mean width*

$$w_{\mathcal{R}}(A) = \mathbb{E}_\sigma \left[\sup_{a \in A} \sigma \cdot a \right].$$

Then $\mathcal{R}_{\mathcal{H}}(\bar{x}) = \frac{1}{n} w_{\mathcal{R}}(\mathcal{H}(\bar{x}))$.

Theorem (variation on Massart's Lemma)

$$w_{\mathcal{R}}(A) \leq \inf_{\gamma \geq 0} \left(\gamma n + \sqrt{\frac{n}{2} \log(\mathcal{N}_\gamma(A))} \right).$$

Rademacher Upper Bounds

Corollary

If $\mathcal{H}$ has VC-dim $\leq d$, then

$$\mathcal{R}_{\mathcal{H}}(n) = O\left(\gamma + n^{-1/2}\log n\right) = o(1).$$

Definition

If $X = \prod_{i=1}^{k} X_i$, let $\mathcal{R}_{\mathcal{H}}^{k}(n) = \sup_{\bar{x}^i} \mathcal{R}_{\mathcal{H}}\left(\bigotimes_{i=1}^{k} \bar{x}^i\right)$.

Corollary (A., Towsner)

If $\mathcal{H}$ has VC_k-dim $\leq d$, then

$$\mathcal{R}_{\mathcal{H}}^{k}(n) = O\left(\gamma + n^{-\varepsilon}\log n\right) = o(1).$$

Rademacher Lower Bound

Lemma (A., Towsner)

If $\mathcal{H}$ γ-shatters $\bar{x}$, then $\mathcal{R}_{\mathcal{H}}(\bar{x}) \geq \gamma$.

Corollary (A., Towsner)

If $\mathcal{H}$ has VC_k-dim $\geq d$, then $\mathcal{R}^k_{\mathcal{H}}(d) \geq \gamma$.

Theorem (A., Towsner)

$\mathcal{H}$ has finite VC_k-dim $\iff \lim_{n\to\infty} \mathcal{R}^k_{\mathcal{H}}(n) = 0$.

Shattering Trees

Definition

For $\gamma > 0$, say $\mathcal{H}$ *γ-shatters* a tree $\bar{x} \in X^{\{-1,1\}^{<n}}$ when there is a witness $v \in [0,1]^{\{-1,1\}^{<n}}$ such that for all $\sigma \in \{-1,1\}^n$, there is $h_\sigma \in \mathcal{H}$ with

$$\forall i \in [n], \sigma_i(h_\sigma(x_{\sigma|_{<i}}) - v_{\sigma|_{<i}}) \geq \frac{\gamma}{2}.$$

Definition

Say $\mathcal{H}$ has

- *γ-Littlestone dimension* $\geq d$ when it shatters $\bar{x} \in X^{\{-1,1\}^{<d}}$. (Shelah 2-rank.)
- *finite Littlestone dimension* when $\forall \gamma > 0$, γ-Littlestone $< \infty$. (Stability for continuous logic!)

Sequential Covering Numbers

Definition

Given a branch $\beta \in \{-1,1\}^n$, let $\pi_\beta : X^{\{-1,1\}^{<n}} \to X^n$ be the projection onto substrings of β.

Definition

Say $C \subseteq [0,1]^{\{-1,1\}^{<n}}$ is a *sequential* γ*-cover* for $A \subseteq [0,1]^{\{-1,1\}^{<n}}$ when for all $\beta \in \{-1,1\}^n$, $\pi_\beta(C)$ is a γ-cover for $\pi_\beta(A)$.
Call the *sequential* γ*-covering number* $\mathcal{N}^{\mathrm{seq}}_\gamma(A) = \min |C|$.

Theorem (Rakhlin, Sridharan, Tewari)

If $\mathcal{H}$ has γ-Littlestone dimension $\leq d$, then $\mathcal{N}^{\mathrm{seq}}_\gamma(A) = O(n^d)$.

Sequential Rademacher Complexity

Definition

If $\bar{x} \in X^{\{-1,1\}^{<n}}$, define the *sequential Rademacher complexity*

$$\mathcal{R}^{\mathrm{seq}}_{\mathcal{H}}(\bar{x}) = \frac{1}{n}\mathbb{E}\left[\sup_{h\in\mathcal{H}} \sum_{i=1}^{n} \sigma_i h(x_{\sigma|_{<i}})\right].$$

Theorem (Rakhlin, Sridharan, Tewari)

$$\mathcal{R}^{\mathrm{seq}}_{\mathcal{H}}(\bar{x}) \leq \inf_{\gamma\geq 0}\left(\gamma + \sqrt{\frac{2}{n}\log(\mathcal{N}_\gamma(A))}\right).$$

Adversarial Samples

Think of $\bar{x} \in X^{\{-1,1\}^{<n}}$ as a strategy tree.

- I reveal $x_\emptyset$.
- You flip a coin for σ_1, deciding which sample $x_\emptyset$ is in.
- Depending on that, I reveal x_{σ_1}.
- You flip a coin for σ_2, deciding which sample that is in.
- I reveal $x_{\sigma_1 \sigma_2}$.
- ...
- We compare h on the two samples.

Functional Littlestone Dimension (towards NFOP_2)

Say $\mathcal{H}$ is defined on $X \times Y$.

Definition

Given a tree $\bar{x} \in X^{\{-1,1\}^{<n}}$ and a tuple $\bar{y} \in Y^n$, define a "product tree" $\bar{x} \rtimes \bar{y} \in (X \times Y)^{\{-1,1\}^{<n^2}}$, with levels corresponding to $[n]^2$ in lex order.

Definition

Say $\mathcal{H}$ has *functional γ-Littlestone dimension* $\geq d$ when it γ-shatters a product tree $\bar{x} \rtimes \bar{y}$ with $\bar{x} \in X^{\{-1,1\}^{<d}}$, $\bar{y} \in Y^d$.

Theorem (A., Towsner)

There is $\varepsilon = \varepsilon(\gamma, d)$ (from z_2) such that either $\mathcal{H}$ has functional γ-Littlestone dimension $\geq d$ or

$$\log\left(\mathcal{N}_\gamma^{\mathrm{seq}}(\mathcal{H}(\bar{x} \rtimes \bar{y}))\right) = O(n^{2-\varepsilon}).$$

Characterizing $NFOP_2$

Definition

Define

$$\mathcal{R}^{\mathrm{fseq}}_{\mathcal{H}}(\bar{x}, \bar{y}) = \frac{1}{n^2} w^{\mathrm{seq}}_{\mathcal{R}}(\bar{x} \rtimes \bar{y})$$

and so on.

Corollary

$\mathcal{H}$ has finite functional Littlestone dimension

$$\iff \lim_{n \to \infty} \mathcal{R}^{\mathrm{fseq}}_{\mathcal{H}}(n) = 0.$$

In Progress

$\mathcal{H}$ has finite functional Littlestone dimension iff it is $NFOP_2$.

Sample Completion PAC Learning

- Start with a grid-shaped sample $\vec{x} = \bigotimes_{i=1}^{k} \bar{x}^i$, $\bar{x}^i \in X_i^n$.
- We choose $E \subseteq [n]^k$ at random.
- I pick $h \in \mathcal{H}$, tell you $h(\vec{x}_{\bar{\imath}})$ *only when* $\bar{\imath} \in E$.
- For big n, can you guess the average $\frac{1}{n}\sum_{i=1}^{n} h(\vec{x}_{\bar{\imath}})$ and be
    - **P**robably $(\mathbb{P} > 1 - \delta)$,
    - **A**pproximately $\left(\left|\text{Guess} - \frac{1}{n^k}\sum_{\bar{\imath}\in[n]^k} h(\vec{x}_{\bar{\imath}})\right| \leq \varepsilon\right)$,
    - **C**orrect?

Theorem (Coregliano, Malliaris)

$\mathcal{H}$ is sample-completion PAC learnable when it has finite VC_k-dimension. (Assuming $h : \prod_{i=1}^{k} X_i \to \{0,1\}$.)

Expected Error

Alternate Proof (A., Towsner).

Making the sensible guess $\frac{2}{n^k}\sum_{\bar{\imath}\in E} h(\vec{x}_{\bar{\imath}})$, assuming worst-case h, your expected error is

$$\begin{aligned}
&\mathbb{E}_E\left[\sup_{h\in\mathcal{H}}\left|\frac{2}{n^k}\sum_{i\in E} h(x_i) - \frac{1}{n^k}\sum_{\bar{\imath}\in[n]^k} h(\vec{x}_{\bar{\imath}})\right|\right] \\
&= \mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{\bar{\imath}\in[n]^k}\sigma_{\bar{\imath}} h(\vec{x}_{\bar{\imath}})\right|\right] \\
&= \mathcal{R}_{\mathcal{H}}(\vec{x}) \\
&\to 0.
\end{aligned}$$

$\square$

General Mean Widths

Definition

Given $A \subseteq [0,1]^n$, a random variable $\sigma \in \mathbb{R}^n$, let

$$w(A, \sigma) = \mathbb{E}\left[\sup_{a \in A} \sum_{i=1}^{n} \sigma_i a_i\right].$$

Can define $\mathcal{R}_{\mathcal{H}}(n), \mathcal{R}^k_{\mathcal{H}}(n), \mathcal{R}^{\mathrm{seq}}_{\mathcal{H}}(n), \mathcal{R}^{\mathrm{fseq}}_{\mathcal{H}}(n)$ with various A and σ.

Slicewise NIP/Stability

Definition

If $X = \prod_{i=1}^{k} X_i$, say $\mathcal{H}$ has *slicewise VC-dimension* $\leq d$ when for all $1 \leq i \leq k$, $\bar{y} \in \prod_{j \neq i} X_i$, $\mathcal{H}(x_i, \bar{y})$ has VC-dimension $\leq d$.
Similar for Littlestone dimension and real-valued versions.

Lemma (A., Towsner)

There are sequences of random variables σ defining $\mathcal{R}^{\mathrm{sl}}_{\mathcal{H}}(n), \mathcal{R}^{\mathrm{seq,sl}}_{\mathcal{H}}(n)$ which characterize finite slicewise VC-dimension, slicewise Littlestone dimension respectively.

Slicewise NIP/Stability

Corollary (Originally Coregliano, Malliaris)

$\mathcal{H}$ having slicewise VC-dimension $\leq d$ implies a uniform law of large numbers over product measures.

In Progress

$\mathcal{H}$ having slicewise Littlestone dimension $\leq d$ implies an adversarial uniform law of large numbers over product measures.

Randomization

- Let Ω be a probability space, and let $\mathcal{M}$ be an $\mathcal{L}$-structure.
- Let M^R be the set of random variables $\Omega \to M$.
- (Skipping measurability details.)

Definition

Make $\mathcal{M}^R$ a metric structure, the *randomization* of $\mathcal{M}$, by equipping it with relations $\mathbb{E}[\phi(\bar{x})]$ for all $\mathcal{L}$-formulas ϕ.

Theorem (Ben Yaacov, Keisler)

$\mathcal{M}^R$ has QE in this language.

Preservation Under Randomization

Theorem (Ben Yaacov, Keisler)

If $\mathcal{M}$ is stable, so is $\mathcal{M}^R$.

Theorem (Ben Yaacov)

If $\mathcal{M}$ is NIP, so is $\mathcal{M}^R$.

Theorem (Chernikov, Towsner)

If $\mathcal{M}$ is NIP_k, so is $\mathcal{M}^R$.

Three totally different proofs!

Preservation Under Randomization

Theorem (Ben Yaacov, Keisler)

If $\mathcal{M}$ is stable, so is $\mathcal{M}^R$.

Theorem (Ben Yaacov)

If $\mathcal{M}$ is NIP, so is $\mathcal{M}^R$.
Uses Gaussian complexity, similar to Rademacher complexity.

Theorem (Chernikov, Towsner)

If $\mathcal{M}$ is NIP_k, so is $\mathcal{M}^R$.

Three totally different proofs!

Preservation Under Randomization

Lemma (A., Benedikt)

Let $\bar{\mathbf{x}}$ *be a tuple of random variables in* X, *and* σ *random in* $[-1,1]^n$.

$$w(\mathcal{H}_{\mathbb{E}[\phi]}(\bar{\mathbf{x}}), \sigma) \leq \sup_{\omega \in \Omega} w(\mathcal{H}(\bar{x}_\omega), \sigma).$$

Corollary

- $\mathcal{R}_{\mathcal{H}_{\mathbb{E}[\phi]}}(n) \leq \mathcal{R}_{\mathcal{H}_\phi}(n)$
- $\mathcal{R}^k_{\mathcal{H}_{\mathbb{E}[\phi]}}(n) \leq \mathcal{R}^k_{\mathcal{H}_\phi}(n)$
- $\mathcal{R}^{\mathrm{seq}}_{\mathcal{H}_{\mathbb{E}[\phi]}}(n) \leq \mathcal{R}^{\mathrm{seq}}_{\mathcal{H}_\phi}(n)$
- $\mathcal{R}^{\mathrm{fseq}}_{\mathcal{H}_{\mathbb{E}[\phi]}}(n) \leq \mathcal{R}^{\mathrm{fseq}}_{\mathcal{H}_\phi}(n)$

Corollary

- *NIP*
- *NIP*$_k$
- *Stability*
- *NFOP*$_2$

preserved under randomization.

Future Questions

When can trace-definability dividing lines be characterized in terms of Rademacher complexity variants?
(Main missing step seems to be Sauer-Shelah.)

Connect VC_k-dimension to something closer to standard PAC learning.

These complexities are all about the *expected discrepancy* between two random subsamples. Distality/strong Erdös-Hajnal seems to be connected to other measures of discrepancy.

Thank you!