Learnability Beyond Uniform Convergence

Shai Shalev-Shwartz

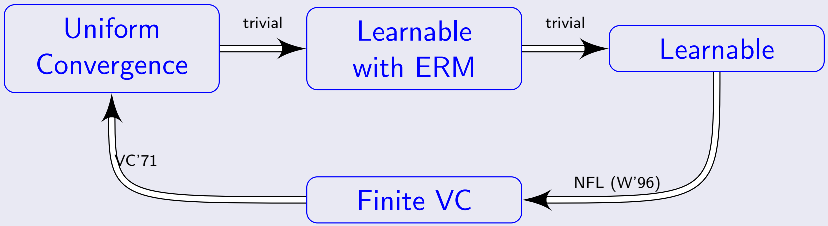
School of CS and Engineering,
The Hebrew University of Jerusalem

"Algorithmic Learning Theory",
Lyon 2012

Joint work with:
N. Srebro, O. Shamir, K. Sridharan (COLT'09,JMLR'11)
A. Daniely, S. Sabato, S. Ben-David (COLT'11)
A. Daniely, S. Sabato (NIPS'12)

The Fundamental Theorem of Learning Theory

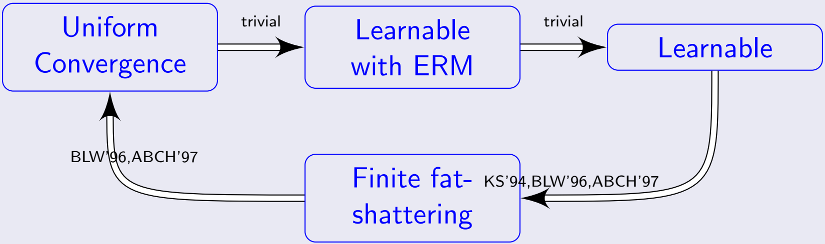
For Binary Classification



VC = Vapnik and Chervonenkis, W = Wolpert

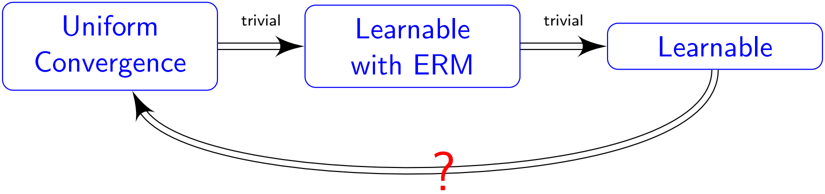
The Fundamental Theorem of Learning Theory

For Regression

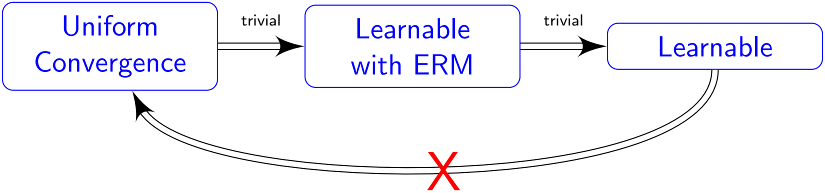


BLW = Bartlett, Long, Williamson. ABCH = Alon, Ben-David, Cesa-Bianchi, Hausler. KS = Kearns and Schapire

For general learning problems?

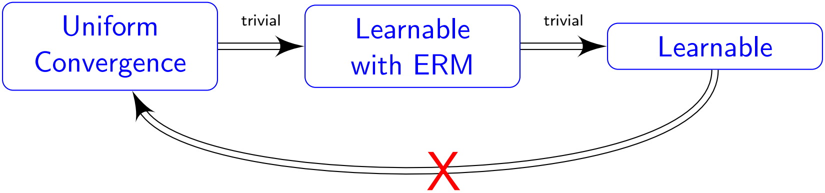


For general learning problems?



- Not true

For general learning problems?



- Not true
  - Not true in "Convex learning problems" !
  - Not true even in "multiclass categorization" !
- What is learnable ? How to learn ?

Outline

The General Learning Setting (Vapnik)

- Hypothesis class $\mathcal{H}$
- Examples domain $\mathcal{Z}$ with unknown distribution $\mathcal{D}$
- Loss function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$

Given: Training set $S \sim \mathcal{D}^m$
Goal: Solve:

$$\min_{h \in \mathcal{H}} L(h) \quad \text{where} \quad L(h) = \mathop{\mathbb{E}}_{z \sim \mathcal{D}} [\ell(h, z)]$$

in the **P**robably (w.p. $\geq 1 - \delta$) **A**pproximately **C**orrect (up to ϵ) sense

The General Learning Setting (Vapnik)

- Hypothesis class $\mathcal{H}$
- Examples domain $\mathcal{Z}$ with unknown distribution $\mathcal{D}$
- Loss function $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$

Given: Training set $S \sim \mathcal{D}^m$
Goal: Solve:

$$\min_{h \in \mathcal{H}} L(h) \quad \text{where} \quad L(h) = \mathop{\mathbb{E}}_{z \sim \mathcal{D}}[\ell(h, z)]$$

in the **P**robably (w.p. $\geq 1 - \delta$) **A**pproximately **C**orrect (up to ϵ) sense

Training loss: $L_S(h) = \dfrac{1}{m} \displaystyle\sum_{i=1}^{m} \ell(h, z_i)$

Examples

- Binary classification:
  - $\mathcal{Z} = \mathcal{X} \times \{0,1\}$
  - $h \in \mathcal{H}$ is a predictor $h : \mathcal{X} \to \{0,1\}$
  - $\ell(h,(x,y)) = \mathbf{1}[h(x) \neq y]$
- Multiclass categorization:
  - $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$
  - $h \in \mathcal{H}$ is a predictor $h : \mathcal{X} \to \mathcal{Y}$
  - $\ell(h,(x,y)) = \mathbf{1}[h(x) \neq y]$
- k-means clustering:
  - $\mathcal{Z} = \mathbb{R}^d$
  - $\mathcal{H} \subset (\mathbb{R}^d)^k$ specifies k cluster centers
  - $\ell((\mu_1,\ldots,\mu_k),z) = \min_j \|\mu_j - z\|$
- Density Estimation:
  - h is a parameter of a density $p_h(z)$
  - $\ell(h,z) = -\log p_h(z)$

Learnability, ERM, Uniform convergence

- **Uniform Convergence**: For $m \geq m_{\mathrm{UC}}(\epsilon, \delta)$

$$\mathop{\mathbb{P}}_{S \sim \mathcal{D}^m} \left[\forall h \in \mathcal{H},\ |L_S(h) - L(h)| \leq \epsilon \right] \geq 1 - \delta$$

Learnability, ERM, Uniform convergence

- **Uniform Convergence**: For $m \geq m_{\mathrm{UC}}(\epsilon, \delta)$

$$\mathop{\mathbb{P}}_{S \sim \mathcal{D}^m} \left[\forall h \in \mathcal{H},\ |L_S(h) - L(h)| \leq \epsilon\right] \geq 1 - \delta$$

- **Learnable**: $\exists \mathcal{A}$ s.t. for $m \geq m_{\mathrm{PAC}}(\epsilon, \delta)$,

$$\mathop{\mathbb{P}}_{S \sim \mathcal{D}^m} \left[L(\mathcal{A}(S)) \leq \min_{h \in \mathcal{H}} L(h) + \epsilon\right] \geq 1 - \delta$$

Learnability, ERM, Uniform convergence

- **Uniform Convergence**: For $m \geq m_{\mathrm{UC}}(\epsilon, \delta)$

$$\underset{S \sim \mathcal{D}^m}{\mathbb{P}} \left[\forall h \in \mathcal{H},\ |L_S(h) - L(h)| \leq \epsilon\right] \geq 1 - \delta$$

- **Learnable**: $\exists \mathcal{A}$ s.t. for $m \geq m_{\mathrm{PAC}}(\epsilon, \delta)$,

$$\underset{S \sim \mathcal{D}^m}{\mathbb{P}} \left[L(\mathcal{A}(S)) \leq \min_{h \in \mathcal{H}} L(h) + \epsilon\right] \geq 1 - \delta$$

- **ERM**:
  An algorithm that returns $\mathcal{A}(S) \in \mathrm{argmin}_{h \in \mathcal{H}} L_S(h)$

Learnability, ERM, Uniform convergence

- **Uniform Convergence**: For $m \geq m_{\mathrm{UC}}(\epsilon, \delta)$

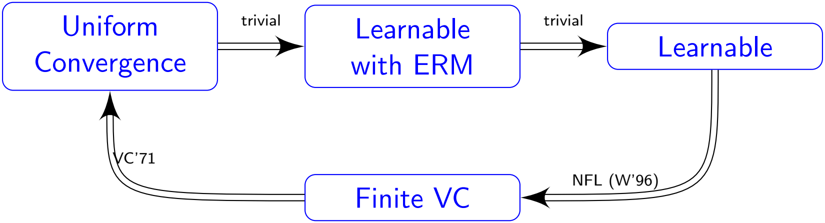
$$\mathop{\mathbb{P}}_{S \sim \mathcal{D}^m} \left[\forall h \in \mathcal{H},\ |L_S(h) - L(h)| \leq \epsilon\right] \geq 1 - \delta$$

- **Learnable**: $\exists \mathcal{A}$ s.t. for $m \geq m_{\mathrm{PAC}}(\epsilon, \delta)$,

$$\mathop{\mathbb{P}}_{S \sim \mathcal{D}^m} \left[L(\mathcal{A}(S)) \leq \min_{h \in \mathcal{H}} L(h) + \epsilon\right] \geq 1 - \delta$$

- **ERM**:
  An algorithm that returns $\mathcal{A}(S) \in \mathrm{argmin}_{h \in \mathcal{H}} L_S(h)$
- **Learnable by arbitrary ERM** (with rate $m_{\mathrm{ERM}}(\epsilon, \delta)$)
  Like "Learnable" but $\mathcal{A}$ should be an ERM.

For Binary Classification



$$m_{\mathrm{UC}}(\epsilon,\delta) \;\approx\; m_{\mathrm{ERM}}(\epsilon,\delta) \;\approx\; m_{\mathrm{PAC}}(\epsilon,\delta) \;\approx\; \frac{\mathrm{VC}(\mathcal{H})\log(1/\delta)}{\epsilon^2}$$

Outline

Counter Example — Stochastic Convex Optimization

Consider the family of problems:

- $\mathcal{H}$ is a convex set with $\max_{h \in \mathcal{H}} \|h\| \leq 1$
- For all z, $\ell(h, z)$ is convex and Lipschitz w.r.t. h

Counter Example — Stochastic Convex Optimization

Consider the family of problems:

- $\mathcal{H}$ is a convex set with $\max_{h \in \mathcal{H}} \|h\| \leq 1$
- For all z, $\ell(h, z)$ is convex and Lipschitz w.r.t. h

Claim:

- Problem is learnable by the rule:

$$\underset{h \in \mathcal{H}}{\operatorname{argmin}} \frac{\lambda_m}{2} \|h\|^2 + \frac{1}{m} \sum_{i=1}^{m} \ell(h, z_i)$$

- No uniform convergence
- Not learnable by ERM

Counter Example — Stochastic Convex Optimization

Proof (of "not learnable by arbitrary ERM")

- 1-Mean + missing features

Counter Example — Stochastic Convex Optimization

Proof (of "not learnable by arbitrary ERM")

- 1-Mean + missing features
- $z = (\alpha, x)$, $\alpha \in \{0,1\}^d$, $x \in \mathbb{R}^d$, $\|x\| \le 1$
- $\ell(h, (\alpha, x)) = \sqrt{\sum_i \alpha_i (h_i - x_i)^2}$
- Take $\mathbb{P}[\alpha_i = 1] = 1/2$, $\mathbb{P}[x = \mu] = 1$
- Let $h^{(i)}$ be s.t.

$$h_j^{(i)} = \begin{cases} 1 - \mu_j & \text{if } j = i \\ \mu_j & \text{o.w.} \end{cases}$$

- If d is large enough, exists i such that $h^{(i)}$ is an ERM
- But $L(h^{(i)}) \ge 1/\sqrt{2}$

Counter Example — Stochastic Convex Optimization

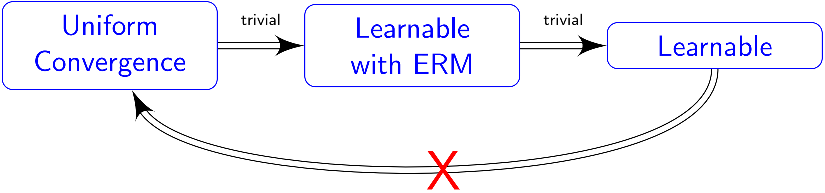
Proof (of "not even learnable by a unique ERM")

Perturb the loss a little bit:

$$\ell(h, (\alpha, x)) = \sqrt{\sum_i \alpha_i (h_i - x_i)^2} + \epsilon \sum_i 2^{-i} (h_i - 1)^2$$

- Now loss is strictly convex — unique ERM
- But the unique ERM does not generalize (as before)

For general learning problems?



- Not true
  - Not true in "Convex learning problems" ! ✓
  - **Not true even in "multiclass categorization" !**

Counter Example — Multiclass

- $\mathcal{X}$ – a set, $\mathcal{Y} = \{0, 1, 2, \ldots, 2^{|\mathcal{X}|} - 1\}$
- Let $n : 2^{\mathcal{X}} \to \mathcal{Y}$ be defined by binary encoding
- $\mathcal{H} = \{h_T : T \subset \mathcal{X}\}$ where

$$h_T(x) = \begin{cases} 0 & x \notin T \\ n(T) & x \in T \end{cases}$$

Counter Example — Multiclass

- $\mathcal{X}$ – a set, $\mathcal{Y} = \{0, 1, 2, \ldots, 2^{|\mathcal{X}|} - 1\}$
- Let $n : 2^{\mathcal{X}} \to \mathcal{Y}$ be defined by binary encoding
- $\mathcal{H} = \{h_T : T \subset \mathcal{X}\}$ where

$$h_T(x) = \begin{cases} 0 & x \notin T \\ n(T) & x \in T \end{cases}$$

- Claim: No uniform convergence: $m_{\mathrm{UC}} \geq |\mathcal{X}|/\epsilon$
  - Target function is $h_\emptyset$
  - For any training set S, take $T = \mathcal{X} \setminus S$
  - $L_S(h_T) = 0$ but $L(h_T) = \mathbb{P}[T]$

Counter Example — Multiclass

- $\mathcal{X}$ – a set, $\mathcal{Y} = \{0, 1, 2, \ldots, 2^{|\mathcal{X}|} - 1\}$
- Let $n : 2^{\mathcal{X}} \to \mathcal{Y}$ be defined by binary encoding
- $\mathcal{H} = \{h_T : T \subset \mathcal{X}\}$ where

$$h_T(x) = \begin{cases} 0 & x \notin T \\ n(T) & x \in T \end{cases}$$

- Claim: $\mathcal{H}$ is Learnable: $m_{\mathrm{PAC}} \leq \frac{1}{\epsilon}$
  - Let T be the target
  - $\mathcal{A}(S) = h_T$ if $(x, n(T)) \in S$
  - $\mathcal{A}(S) = h_\emptyset$ if $S = \{(x_1, 0), \ldots, (x_m, 0)\}$
  - In the 1st case, $L(\mathcal{A}(S)) = 0$.
  - In the 2nd case, $L(\mathcal{A}(S)) = \mathbb{P}[T]$
  - With high probability, if $\mathbb{P}[T] > \epsilon$ then we'll be in the 1st case

Counter Example — Multiclass

Corollary

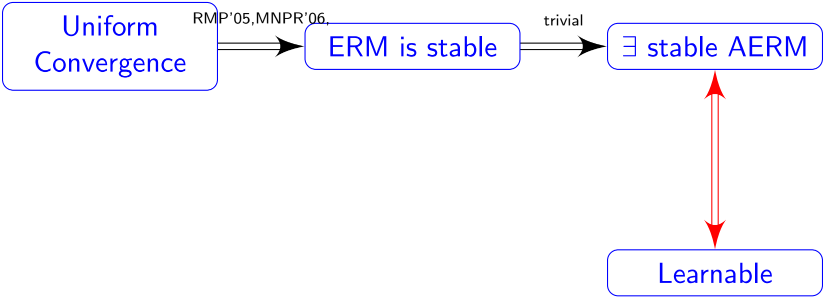
- $\frac{m_{UC}}{m_{PAC}} \approx |\mathcal{X}|$.
- *If $|\mathcal{X}| \to \infty$ then the problem is learnable but there is no uniform convergence!*

Outline

Characterizing Learnability using Stability

Theorem

A sufficient and necessary condition for learnability is the existence of Asymptotic ERM (AERM) which is stable.



More formally

Definition (Stability)

We say that A is $\epsilon_{\text{stable}}(m)$-replace-one stable if for all $\mathcal{D}$,

$$\mathop{\mathbb{E}}_{S,z',i} |\ell(\mathcal{A}(S^{(i)}); z') - \ell(\mathcal{A}(S); z')| \leq \epsilon_{\text{stable}}(m).$$

More formally

Definition (Stability)

We say that A is $\epsilon_{\text{stable}}(m)$-replace-one stable if for all $\mathcal{D}$,

$$\mathop{\mathbb{E}}_{S,z',i} |\ell(\mathcal{A}(S^{(i)}); z') - \ell(\mathcal{A}(S); z')| \leq \epsilon_{\text{stable}}(m).$$

Definition (AERM)

We say that $\mathcal{A}$ is an *AERM (Asymptotic Empirical Risk Minimizer)* with rate $\epsilon_{\text{erm}}(m)$ if for all $\mathcal{D}$:

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} [L_S(\mathcal{A}(S)) - \min_{h\in\mathcal{H}} L_S(h)] \leq \epsilon_{\text{erm}}(m)$$

Proof sketch: (Stable AERM is sufficient and necessary for Learnability)

Sufficient:

- For AERM: stability $\Rightarrow$ generalization
- AERM+generalization $\Rightarrow$ consistency

Necessary:

- $\exists$ consistent $\mathcal{A} \Rightarrow$
  $\exists$ consistent and generalizing A' (using subsampling)
- Consistent+generalizing $\Rightarrow$ AERM
- AERM+generalizing $\Rightarrow$ stable

Intermediate Summary

- Learnability $\iff$ $\exists$ stable AERM
- But, how do we find one?
- And, is there a combinatorial notion of learnability (like VC dimension) ?

Outline

Why multiclass learning

- Practical relevance
- A simple twist of binary classification

The Natarajan Dimension

Natarajan dimension: Maximal size of N-shattered set where:

C is N-shattered by $\mathcal{H}$ if $\exists f_1, f_2 \in \mathcal{H}$ s.t. $\forall x \in C,\ f_1(x) \neq f_2(x)$, and for every $T \subseteq C$ exists $h \in \mathcal{H}$ with

$$h(x) = \begin{cases} f_1(x) & \text{if } x \in T \\ f_2(x) & \text{if } x \in C \setminus T \end{cases}$$

The Natarajan Dimension

Natarajan dimension: Maximal size of N-shattered set where:

C is N-shattered by $\mathcal{H}$ if $\exists f_1, f_2 \in \mathcal{H}$ s.t. $\forall x \in C,\ f_1(x) \neq f_2(x)$, and for every $T \subseteq C$ exists $h \in \mathcal{H}$ with

$$h(x) = \begin{cases} f_1(x) & \text{if } x \in T \\ f_2(x) & \text{if } x \in C \setminus T \end{cases}$$

- When $|\mathcal{Y}| = 2$, Natarajan dimension equals to VC dimension

Does Natarajan dimension characterize multiclass learnability ?

Theorem (Natarajan'89, Ben-David et al 95)

If $\mathcal{H}$ is a class of functions with Natarajan dimension d then

$$\frac{d + \ln(1/\delta)}{\epsilon} \;\le\; m_{PAC}(\epsilon, \delta) \;\le\; \frac{d \;\mathbf{ln}(|\boldsymbol{\mathcal{Y}}|)\;\; \ln(1/\epsilon) + \ln(1/\delta)}{\epsilon} \;.$$

Does Natarajan dimension characterize multiclass learnability ?

Theorem (Natarajan'89, Ben-David et al 95)

If $\mathcal{H}$ is a class of functions with Natarajan dimension d then

$$\frac{d + \ln(1/\delta)}{\epsilon} \leq m_{PAC}(\epsilon, \delta) \leq \frac{d\ \mathbf{ln}(|\boldsymbol{\mathcal{Y}}|)\ \ln(1/\epsilon) + \ln(1/\delta)}{\epsilon} .$$

Remark:

- A large gap when $\mathcal{Y}$ is large
- Uniform convergence rate does depend on $\mathcal{Y}$

How to design good ERM algorithm?

- Consider again our counter example: $\mathcal{Y} = \{0, \ldots, 2^{|\mathcal{X}|} - 1\}$ and $\mathcal{H} = \{h_T : T \subset \mathcal{X}\}$ with

$$h_T(x) = \begin{cases} 0 & x \notin T \\ n(T) & x \in T \end{cases}$$

How to design good ERM algorithm?

- Consider again our counter example: $\mathcal{Y} = \{0, \ldots, 2^{|\mathcal{X}|} - 1\}$ and $\mathcal{H} = \{h_T : T \subset \mathcal{X}\}$ with

$$h_T(x) = \begin{cases} 0 & x \notin T \\ n(T) & x \in T \end{cases}$$

- Bad ERM:
  - If $S = (x_1, 0), \ldots, (x_m, 0)$ return h_T with $T = \mathcal{X} \setminus \{x_1, \ldots, x_m\}$
- Good ERM
  - If $S = (x_1, 0), \ldots, (x_m, 0)$ return $h_\emptyset$

How to design a good ERM algorithm?

Definition

A has an essential range r if $\forall h \in \mathcal{H}, \exists \mathcal{Y}'(h)$ with $|\mathcal{Y}'(h)| \leq r$ s.t. for all S labeled by h we have $A(S) \in \mathcal{Y}'(h)$

A Principle for Designing Good ERMs

A good ERM is an ERM that has a small essential range

Theorem

If a learner has an "essential" range r then

$$m_{\mathcal{A}}(\epsilon, \delta) \leq \frac{d \ln(r/\epsilon) + \ln(1/\delta)}{\epsilon}$$

Characterizing Multiclass Learnability

Conjecture

For any $\mathcal{H}$ of Natarajan dimension d,

$$\frac{d + \ln(1/\delta)}{\epsilon} \;\leq\; m_{PAC}(\epsilon, \delta) \;\leq\; \frac{d \ln(d/\epsilon) + \ln(1/\delta)}{\epsilon} \; .$$

Characterizing Multiclass Learnability

Conjecture

For any $\mathcal{H}$ of Natarajan dimension d,

$$\frac{d + \ln(1/\delta)}{\epsilon} \;\leq\; m_{PAC}(\epsilon, \delta) \;\leq\; \frac{d\ln(d/\epsilon) + \ln(1/\delta)}{\epsilon} \; .$$

- Cannot rely on uniform convergence / arbitrary ERM
- Maybe there's always an ERM with a small essential range ?
- Holds for symmetric classes

Outline

Sample Complexity of Specific classes

- Enables a rigorous comparison of known multiclass algorithms
  - Previous analyses (e.g. ASS'01,BL'07): how the binary error translates to multiclass error
- Multiclass predictors:
  - One-vs-All (OvA)
  - Multiclass SVM (MSVM): $\arg\max_i (Wx)_i$
  - Tree Classifiers (TC), with $\tilde{O}(|\mathcal{Y}|)$ nodes
  - Error Correcting Output Codes (ECOC), with code-length $\tilde{O}(|\mathcal{Y}|)$
- Use linear predictors in $\mathbb{R}^d$ as the binary classifiers

Sample Complexity of Specific classes

- Enables a rigorous comparison of known multiclass algorithms
  - Previous analyses (e.g. ASS'01,BL'07): how the binary error translates to multiclass error
- Multiclass predictors:
  - One-vs-All (OvA)
  - Multiclass SVM (MSVM): $\arg\max_i (Wx)_i$
  - Tree Classifiers (TC), with $\tilde{O}(|\mathcal{Y}|)$ nodes
  - Error Correcting Output Codes (ECOC), with code-length $\tilde{O}(|\mathcal{Y}|)$
- Use linear predictors in $\mathbb{R}^d$ as the binary classifiers

Theorem

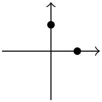
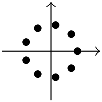
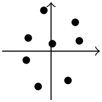
The sample complexity of all the above classes is $\tilde{\Theta}(d\,|\mathcal{Y}|)$.

Comparing Approximation Error

Definition

- We say that $\mathcal{H}$ *essentially contains* $\mathcal{H}'$ if for any distribution, the approximation error of $\mathcal{H}$ is at most the approximation error of $\mathcal{H}'$.
- $\mathcal{H}$ *strictly contains* $\mathcal{H}'$ if, in addition, there is a distribution for which the approximation error of $\mathcal{H}$ is strictly smaller than that of $\mathcal{H}'$.

Comparing Approximation Error

			
MSVM	✓	✓	✓
OvA	✓	✓	✗
TC/ECOC	✓	✗	✗

* Assuming tree structure and ECOC code are chosen randomly

Comparing Approximation Error

	TC	**OvA**	**MSVM**	**random ECOC**
Est.	$d\|\mathcal{Y}\|$	$d\|\mathcal{Y}\|$	$d\|\mathcal{Y}\|$	$d\|\mathcal{Y}\|$
Approx. error	$\geq$ MSVM $\approx 1/2$ if $d \ll \|\mathcal{Y}\|$	$\geq$ MSVM	best	incomparable $\approx 1/2$ if $d \ll \|\mathcal{Y}\|$

Open Questions

- Equivalence between uniform convergence and learnability breaks even in multiclass problems
- What characterizes multiclass learnability ?
- What is the corresponding learning rule ?
- What characterizes learnability in the general learning setting ?
- What is the corresponding learning rule ?

Open Questions

- Equivalence between uniform convergence and learnability breaks even in multiclass problems
- What characterizes multiclass learnability ?
- What is the corresponding learning rule ?
- What characterizes learnability in the general learning setting ?
- What is the corresponding learning rule ?

THANKS