Littlestone's Dimension and Online Learnability

Shai Shalev-Shwartz

Toyota Technological Institute at Chicago $\Rightarrow$ The Hebrew University

Talk at UCSD workshop, February, 2009

Joint work with Shai Ben-David and David Pal

Online Learning

For $t = 1, \ldots, T$

- Environment presents input $\mathbf{x}_t \in \mathcal{X}$
- Learner predicts label $\hat{y}_t \in \{0, 1\}$
- Environment reveals true label $y_t \in \{0, 1\}$
- Learner pays 1 if $\hat{y}_t \neq y_t$ and 0 otherwise

Goal: Make few mistakes

Online Learning

For $t = 1, \ldots, T$

- Environment presents input $\mathbf{x}_t \in \mathcal{X}$
- Learner predicts label $\hat{y}_t \in \{0, 1\}$
- Environment reveals true label $y_t \in \{0, 1\}$
- Learner pays 1 if $\hat{y}_t \neq y_t$ and 0 otherwise

Goal: Make few mistakes

Online Learnability: When can we guarantee to make few mistakes ?

Online Learning

For $t = 1, \ldots, T$

- Environment presents input $\mathbf{x}_t \in \mathcal{X}$
- Learner predicts label $\hat{y}_t \in \{0, 1\}$
- Environment reveals true label $y_t \in \{0, 1\}$
- Learner pays 1 if $\hat{y}_t \neq y_t$ and 0 otherwise

Goal: Make few mistakes

Online Learnability: When can we guarantee to make few mistakes ?

PAC Learnability: well understood (VC theory)

Outline

Online Learnability:

Can we be almost as good as the best predictor in a reference class $\mathcal{H}$?

Outline

Online Learnability:

Can we be almost as good as the best predictor in a reference class $\mathcal{H}$?

	Finite $\mathcal{H}$	Infinite $\mathcal{H}$	margin-based $\mathcal{H}$
No noise	Halving		
Arbitrary noise			
Stochastic noise			

Outline

Online Learnability:

Can we be almost as good as the best predictor in a reference class $\mathcal{H}$?

	Finite $\mathcal{H}$	Infinite $\mathcal{H}$	margin-based $\mathcal{H}$
No noise	Halving	L'88	
Arbitrary noise			
Stochastic noise			

Outline

Online Learnability:

Can we be almost as good as the best predictor in a reference class $\mathcal{H}$?

	Finite $\mathcal{H}$	Infinite $\mathcal{H}$	margin-based $\mathcal{H}$
No noise	Halving	L'88	
Arbitrary noise	LW'94		
Stochastic noise			

Outline

Online Learnability:

Can we be almost as good as the best predictor in a reference class $\mathcal{H}$?

	Finite $\mathcal{H}$	Infinite $\mathcal{H}$	margin-based $\mathcal{H}$
No noise	Halving	L'88	✓
Arbitrary noise	LW'94	✓	✓
Stochastic noise	✓	✓	✓

Outline

Online Learnability:

Can we be almost as good as the best predictor in a reference class $\mathcal{H}$?

	Finite $\mathcal{H}$	Infinite $\mathcal{H}$	margin-based $\mathcal{H}$
No noise	Halving	L'88	✓
Arbitrary noise	LW'94	✓	✓
Stochastic noise	✓	✓	✓

- Upper and (almost) matching lower bounds
- Seamlessly deriving new algorithms/bounds

Realizable Case (no noise)

Realizable Assumption: Environment answers $y_t = h(\mathbf{x}_t)$, where $h \in \mathcal{H}$ and the hypothesis class, $\mathcal{H}$, is known to the learner

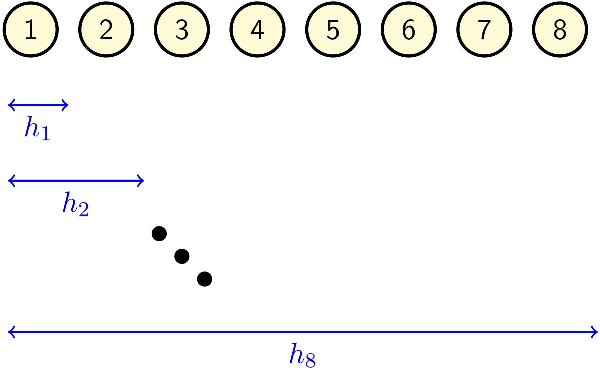
Theorem (Littlestone'88)

A combinatorial dimension, $\mathrm{Ldim}(\mathcal{H})$*, characterizes online learnability:*

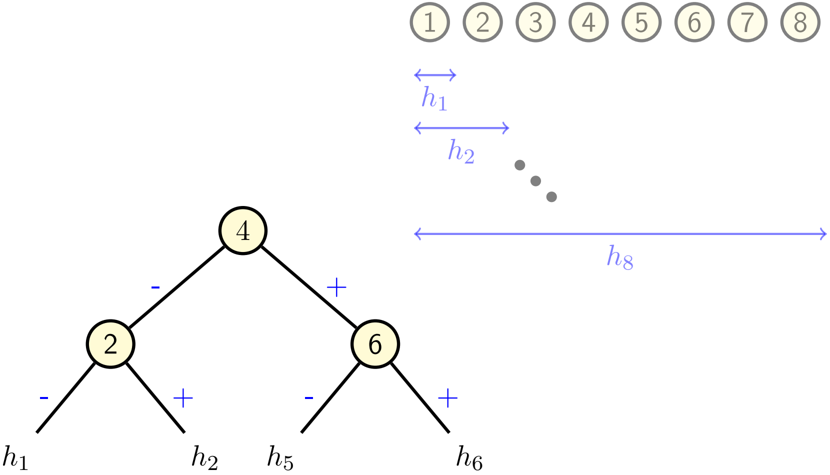
- *Any algorithm might make at least* $\mathrm{Ldim}(\mathcal{H})$ *mistakes*
- *Exists algorithm that makes at most* $\mathrm{Ldim}(\mathcal{H})$ *mistakes*

But, only in the realizable case ...

Littlestone's dimension – Motivation



Littlestone's dimension – Motivation



Littlestone's dimension

Definition

$\mathrm{Ldim}(\mathcal{H})$ is the maximal depth of a full binary tree such that each path is "explained" by some $h \in \mathcal{H}$

Lemma

Any learner can be forced to make at least $\mathrm{Ldim}(\mathcal{H})$ *mistakes*

Proof.

Adversarial environment will "walk" on the tree, while on each round setting $y_t = \neg\hat{y}_t$. $\square$

Standard Optimal Algorithm (SOA)

initialize: $V_1 = \mathcal{H}$
for $t = 1, 2, \ldots$
  receive $\mathbf{x}_t$
  for $r \in \{0, 1\}$ let $V_t^{(r)} = \{h \in V_t : h(\mathbf{x}_t) = r\}$
  predict $\hat{y}_t = \arg\max_r \mathrm{Ldim}(V_t^{(r)})$
  receive true answer y_t
  update $V_{t+1} = V_t^{(y_t)}$

Standard Optimal Algorithm (SOA)

initialize: $V_1 = \mathcal{H}$
for $t = 1, 2, \ldots$
  receive $\mathbf{x}_t$
  for $r \in \{0,1\}$ let $V_t^{(r)} = \{h \in V_t : h(\mathbf{x}_t) = r\}$
  predict $\hat{y}_t = \arg\max_r \mathrm{Ldim}(V_t^{(r)})$
  receive true answer y_t
  update $V_{t+1} = V_t^{(y_t)}$

Theorem

SOA makes at most $\mathrm{Ldim}(\mathcal{H})$ *mistakes.*

Proof.

Whenever SOA errs we have $\mathrm{Ldim}(V_{t+1}) \le \mathrm{Ldim}(V_t) - 1$. □

Intermediate Summary

- Littlestone's dimension characterizes online learnability
- Example:
  $\mathcal{H} = \{$ all 100 characters long C++ functions $\}$
  $\Rightarrow \quad \mathrm{Ldim}(\mathcal{H}) \leq 500$

Intermediate Summary

- Littlestone's dimension characterizes online learnability
- Example:
  $\mathcal{H} = \{$ all 100 characters long C++ functions $\}$
  $\Rightarrow \quad \mathrm{Ldim}(\mathcal{H}) \leq 500$
- Received relatively little attention by researchers
- Maybe due to:
  - Non-realistic realizable assumption
  - Lack of interesting examples
  - Lack of margin-based theory
- Coming Next – Generalizing to:
  - Agnostic case (noise is allowed)
  - Fat dimension and margin-based bounds
  - Linear separators
  - $\Rightarrow$ new algorithms/bounds

Agnostic Online Learning and Regret Analysis

- Make no assumptions on origin of labels
- Analyze regret of not following best predictor in $\mathcal{H}$:

$$\sum_{t=1}^{T} |\hat{y}_t - y_t| \; - \; \min_{h \in \mathcal{H}} \sum_{t=1}^{T} |h(\mathbf{x}_t) - y_t|$$

- When can we guarantee low regret ?

Cover's impossibility result

- $\mathcal{H} = \{h(x) = 1, h(x) = 0\}$
- $\mathrm{Ldim}(\mathcal{H}) = 1$
- Environment will output $y_t = \neg \hat{y}_t$
- Learner makes T mistakes
- Best in $\mathcal{H}$ makes at most $T/2$ mistakes
- Regret is at least $T/2$

Corollary: Online learning in the non-realizable case is impossible ?!?

Randomized Prediction and Expected Regret

- Let's weaken the environment – it should decide on y_t before seeing $\hat{y}_t$
- For deterministic learner, environment can simulate learner so there's no difference
- For learner that randomizes his predictions – big difference
- We analyze expected regret

$$\sum_{t=1}^{T} \mathbb{E}[|\hat{y}_t - y_t|] \; - \; \min_{h \in \mathcal{H}} \sum_{t=1}^{T} |h(\mathbf{x}_t) - y_t|$$

- This enables to sidestep Cover's impossibility result
- Online learning in the non-realizable case becomes possible !

Weighed Majority

WM for learning with d experts

initialize: assign weight $w_i = 1$ for each expert

for $t = 1, 2, \ldots, T$

each expert predicts $f_i \in \{0, 1\}$

environment determines y_t without revealing it to the learner

predict $\hat{y}_t = 1$ w.p. $\propto \sum_{i: f_i = 1} w_i$

receive label y_t

foreach wrong expert: $w_i \leftarrow \eta\, w_i$

Weighed Majority

WM for learning with d experts

initialize: assign weight $w_i = 1$ for each expert

for $t = 1, 2, \ldots, T$

each expert predicts $f_i \in \{0, 1\}$

environment determines y_t without revealing it to the learner

predict $\hat{y}_t = 1$ w.p. $\propto \sum_{i: f_i = 1} w_i$

receive label y_t

foreach wrong expert: $w_i \leftarrow \eta\, w_i$

Theorem

WM achieves expected regret of at most: $\sqrt{\ln(d)\, T}$

WM and Online Learnability

- WM regret bound $\Rightarrow$ a finite $\mathcal{H}$ is learnable with regret $\sqrt{\ln(|\mathcal{H}|)\,T}$
- Is this the best we can do ? And, what if $\mathcal{H}$ is infinite ?
- Solution: Combing WM with SOA

WM and Online Learnability

- WM regret bound $\Rightarrow$ a finite $\mathcal{H}$ is learnable with regret $\sqrt{\ln(|\mathcal{H}|)\,T}$
- Is this the best we can do ? And, what if $\mathcal{H}$ is infinite ?
- Solution: Combing WM with SOA

Theorem

- *Exists learner with expected regret* $\sqrt{\mathrm{Ldim}(\mathcal{H})\,T\,\log(T)}$
- *No learner can have expected regret smaller than* $\sqrt{\mathrm{Ldim}(\mathcal{H})\,T}$

Therefore: $\mathcal{H}$ *is agnostic online learnable* $\iff \mathrm{Ldim}(\mathcal{H}) < \infty$

Proof idea

Expert($i_1, \ldots, i_L$)

initialize: $V_1 = \mathcal{H}$
for $t = 1, 2, \ldots$
  receive $\mathbf{x}_t$
  for $r \in \{0, 1\}$ let $V_t^{(r)} = \{h \in V_t : h(\mathbf{x}_t) = r\}$
  define $\hat{y}_t = \arg\max_r \mathrm{Ldim}(V_t^{(r)})$
  if $t \in \{i_1, \ldots, i_L\}$ flip prediction: $\hat{y}_t \leftarrow \neg \hat{y}_t$
  update $V_{t+1} = V_t^{(\hat{y}_t)}$

Lemma

If $\mathrm{Ldim}(\mathcal{H}) < \infty$, *then for any* $h \in \mathcal{H}$ *exists* $i_1, \ldots, i_L$, $L < \mathrm{Ldim}(\mathcal{H})$, *s.t. Expert(*$i_1, \ldots, i_L$*) agrees with* h *on the entire sequence.*

Agnostic Online Learning – Bounded Stochastic Noise

- Previous theorem holds for any noise
- For stochastic noise – better results
- Assume: $y_t = h(x_t) +_2 \nu_t$, where $\mathbb{P}[\nu_t = 1] \leq \gamma < \frac{1}{2}$
- Then, there exists learner with:

$$\mathbb{E}\left[\sum_{t=1}^{T} |\hat{y}_t - h(x_t)|\right] \;\leq\; \frac{1}{1-2\sqrt{\gamma(1-\gamma)}} \,\mathrm{Ldim}(\mathcal{H})\,\ln(T)$$

- Learner is better than teacher: Learner makes $O(\ln(T))$ mistakes while teacher makes γT mistakes

Fat Littlestone's dimension

- Consider hypotheses of the form $h : \mathcal{X} \to \mathbb{R}$, where actual prediction is $\mathrm{sign}(h(\mathbf{x}))$
- Fat Littlestone's dimension: Maximal depth of tree such that each path is explained by some $h \in \mathcal{H}$ with margin γ
- Importance: Can apply analysis tools for bounding a combinatorial object

Theorem

- *Let M be expected #mistakes of online learner*
- *Let $M_\gamma(\mathcal{H})$ be #margin-mistakes of optimal $h \in \mathcal{H}$*

$$M \;\leq\; M_\gamma(\mathcal{H}) + \sqrt{\mathrm{Ldim}_\gamma(\mathcal{H})\,\ln(T)\,T}$$

Fat Littlestone's dimension of linear separators

Linear predictors: $\mathcal{H} = \{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle : \|\mathbf{w}\| \le 1\}$

Fat Littlestone's dimension of linear separators

Linear predictors: $\mathcal{H} = \{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle : \|\mathbf{w}\| \leq 1\}$

Lemma

If $\mathcal{X}$ is the unit ball of a σ-regular Banach space $(B, \|\cdot\|_\star)$, then $\mathrm{Ldim}_\gamma(\mathcal{H}) \leq \frac{\sigma}{\gamma^2}$

Fat Littlestone's dimension of linear separators

Linear predictors: $\mathcal{H} = \{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle : \|\mathbf{w}\| \le 1\}$

Lemma

If $\mathcal{X}$ is the unit ball of a σ-regular Banach space $(B, \|\cdot\|_\star)$, then $\mathrm{Ldim}_\gamma(\mathcal{H}) \le \frac{\sigma}{\gamma^2}$

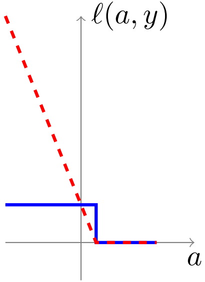
Examples:

$\mathcal{X}$	$\mathcal{H}$	$\mathrm{Ldim}_\gamma(\mathcal{H})$
$\{\mathbf{x} : \|\mathbf{x}\|_2 \le 1\}$	$\{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle : \|\mathbf{w}\|_2 \le 1\}$	$\frac{1}{\gamma^2}$
$\{\mathbf{x} : \|\mathbf{x}\|_\infty \le 1\}$	$\{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle : \|\mathbf{w}\|_1 \le 1\}$	$\frac{\log(n)}{\gamma^2}$

(Surprising) Corollary: Regret with non-convex loss

$$M \;\leq\; M_\gamma(\mathcal{H}) + \tfrac{1}{\gamma}\sqrt{\ln(T)\,T}$$

- Freund and Schapire'99 – Quadratic loss
- Gentile 02 – hinge loss
- No result with non-convex loss



Summary

	Online Learning	PAC Learning
Dimension	$\mathrm{Ldim}(\mathcal{H})$	$\mathrm{VCdim}(\mathcal{H})$
Realizable case: $\frac{\dim}{T}$	✓	✓
Agnostic case: $\sqrt{\frac{\dim}{T}}$	✓	✓
Low noise: $\frac{\dim}{T}$	✓	✓
Margin:	✓	✓

Some Open Problems

- Ldim and fat-Ldim calculus
- Bridging the $\log(T)$ gap between lower and upper bounds
- Other noise conditions (Tsybakov, Steinwart)
- Multiclass prediction with *bandit* feedback: Efficient algorithms? Lower bounds ?
- Low $\text{Ldim} \Rightarrow$ Compression scheme $\Rightarrow$ Low VCdim
- Low $\text{Ldim} \stackrel{?}{\Leftarrow}$ Compression scheme $\stackrel{?}{\Leftarrow}$ Low VCdim