Littlestone’s Dimension and Online Learnability

Shai Shalev-Shwartz

Toyota Technological Institute at Chicago ⇒ The Hebrew University

Talk at UCSD workshop, February, 2009

Joint work with Shai Ben-David and David Pal
For $t = 1, \ldots, T$

- Environment presents input $x_t \in \mathcal{X}$
- 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 $x_t \in \mathcal{X}$
- 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 $x_t \in \mathcal{X}$
- 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 $\mathcal{H}$ ?
### Online Learnability:
Can we be almost as good as the best predictor in a reference class $\mathcal{H}$?

<table>
<thead>
<tr>
<th></th>
<th>Finite $\mathcal{H}$</th>
<th>Infinite $\mathcal{H}$</th>
<th>margin-based $\mathcal{H}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>No noise</td>
<td>Halving</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arbitrary noise</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stochastic noise</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Online Learnability:**
Can we be almost as good as the best predictor in a reference class $\mathcal{H}$?

<table>
<thead>
<tr>
<th></th>
<th>Finite $\mathcal{H}$</th>
<th>Infinite $\mathcal{H}$</th>
<th>margin-based $\mathcal{H}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>No noise</td>
<td>Halving</td>
<td>L’88</td>
<td></td>
</tr>
<tr>
<td>Arbitrary noise</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stochastic noise</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Online Learnability:**
Can we be almost as good as the best predictor in a reference class $\mathcal{H}$?

<table>
<thead>
<tr>
<th></th>
<th>Finite $\mathcal{H}$</th>
<th>Infinite $\mathcal{H}$</th>
<th>margin-based $\mathcal{H}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>No noise</td>
<td>Halving</td>
<td>L’88</td>
<td></td>
</tr>
<tr>
<td>Arbitrary noise</td>
<td>LW’94</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stochastic noise</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Online Learnability:**
Can we be almost as good as the best predictor in a reference class $\mathcal{H}$?

<table>
<thead>
<tr>
<th></th>
<th>Finite $\mathcal{H}$</th>
<th>Infinite $\mathcal{H}$</th>
<th>margin-based $\mathcal{H}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>No noise</td>
<td>Halving</td>
<td>L’88</td>
<td>✓</td>
</tr>
<tr>
<td>Arbitrary noise</td>
<td>LW’94</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Stochastic noise</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
Online Learnability:
Can we be almost as good as the best predictor in a reference class $\mathcal{H}$?

<table>
<thead>
<tr>
<th></th>
<th>Finite $\mathcal{H}$</th>
<th>Infinite $\mathcal{H}$</th>
<th>margin-based $\mathcal{H}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>No noise</td>
<td>Halving</td>
<td>L’88</td>
<td>✓</td>
</tr>
<tr>
<td>Arbitrary noise</td>
<td>LW’94</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Stochastic noise</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

- Upper and (almost) matching lower bounds
- Seamlessly deriving new algorithms/bounds
Realizable Case (no noise)

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

**Theorem (Littlestone’88)**

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

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

But, only in the realizable case ...
Littlestone’s dimension – Motivation

1  2  3  4  5  6  7  8

$h_1$

$h_2$

$h_8$
Littlestone’s dimension – Motivation

\[
\begin{array}{c}
\begin{array}{ccc}
4 & - & + \\
2 & - & + \\
6 & - & + \\
\end{array}
\end{array}
\begin{array}{c}
h_1 \quad h_2 \quad h_5 \quad h_6 \\
\end{array}
\begin{array}{c}
1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 8 \\
\end{array}
\begin{array}{c}
\overset{h_1}{\sim} \hspace{1cm} \overset{h_2}{\sim} \hspace{1cm} h_8 \\
\end{array}
\]
Littlestone’s dimension

**Definition**

$L_{\text{dim}}(\mathcal{H})$ 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 $L_{\text{dim}}(\mathcal{H})$ mistakes*

**Proof.**

Adversarial environment will “walk” on the tree, while on each round setting $y_t = \neg \hat{y}_t$. ✅
**Standard Optimal Algorithm (SOA)**

**initialize:** \(V_1 = \mathcal{H}\)

**for** \(t = 1, 2, \ldots\)
- receive \(x_t\)
  - for \(r \in \{0, 1\}\) let \(V_t^{(r)} = \{h \in V_t : h(x_t) = r\}\)
  - predict \(\hat{y}_t = \arg \max_r \text{Ldim}(V_t^{(r)})\)
  - receive true answer \(y_t\)
  - update \(V_{t+1} = V_t^{(y_t)}\)

**Theorem**

SOA makes at most \(\text{Ldim}(\mathcal{H})\) mistakes.

**Proof.** Whenever SOA errs we have \(\text{Ldim}(V_{t+1}) \leq \text{Ldim}(V_t) - 1\).
Standard Optimal Algorithm (SOA)

initialize: $V_1 = \mathcal{H}$

for $t = 1, 2, \ldots$

receive $x_t$

for $r \in \{0, 1\}$ let $V_t^{(r)} = \{h \in V_t : h(x_t) = r\}$

predict $\hat{y}_t = \arg \max_r \text{Ldim}(V_t^{(r)})$

receive true answer $y_t$

update $V_{t+1} = V_t^{(y_t)}$

Theorem

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

Proof.

Whenever SOA errs we have $\text{Ldim}(V_{t+1}) \leq \text{Ldim}(V_t) - 1$. 

\[ \text{Ldim}(V_{t+1}) \leq \text{Ldim}(V_t) - 1. \]
Intermediate Summary

- Littlestone’s dimension characterizes online learnability
- **Example:**
  \[ \mathcal{H} = \{ \text{all 100 characters long C++ functions} \} \]
  \[ \Rightarrow \text{Ldim}(\mathcal{H}) \leq 500 \]
Intermediate Summary

- Littlestone’s dimension characterizes online learnability
  
  **Example:**
  \[ \mathcal{H} = \{ \text{all 100 characters long C++ functions} \} \]
  \[ \Rightarrow \quad \text{Ldim}(\mathcal{H}) \leq 500 \]

- 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
  \[ \Rightarrow \text{new algorithms/bounds} \]
- Make no assumptions on origin of labels
- Analyze regret of not following best predictor in $\mathcal{H}$:
  $$\sum_{t=1}^{T} |\hat{y}_t - y_t| - \min_{h \in \mathcal{H}} \sum_{t=1}^{T} |h(x_t) - y_t|$$
- When can we guarantee low regret?
Cover’s impossibility result

- \( \mathcal{H} = \{ h(x) = 1, h(x) = 0 \} \)
- \( \text{Ldim}(\mathcal{H}) = 1 \)
- 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 ?!?
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:

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

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 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 $\sqrt{\text{Ldim}(\mathcal{H})T \log(T)}$
- No learner can have expected regret smaller than $\sqrt{\text{Ldim}(\mathcal{H})T}$

Therefore: $\mathcal{H}$ is agnostic online learnable $\iff$ $\text{Ldim}(\mathcal{H}) < \infty$
Proof idea

Expert\((i_1, \ldots, i_L)\)

**initialize:** \(V_1 = \mathcal{H}\)

**for** \(t = 1, 2, \ldots\)

receive \(x_t\)

for \(r \in \{0, 1\}\) let \(V_t^{(r)} = \{h \in V_t : h(x_t) = r\}\)

define \(\hat{y}_t = \arg \max_r \text{Ldim}(V_t^{(r)})\)

**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* \(\text{Ldim}(\mathcal{H}) < \infty\), *then for any* \(h \in \mathcal{H}\) *exists* \(i_1, \ldots, i_L, L < \text{Ldim}(\mathcal{H})\), *s.t.*

\(\text{Expert}(i_1, \ldots, i_L)\) *agrees with* \(h\) *on the entire sequence.*
- 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:

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

- **Learner is better than teacher:** Learner makes \( O(\ln(T)) \) mistakes while teacher makes \( \gamma T \) mistakes
Fat Littlestone’s dimension

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

**Theorem**

Let $M$ be expected number of mistakes of online learner

Let $M_\gamma(\mathcal{H})$ be number of margin-mistakes of optimal $h \in \mathcal{H}$

\[
M \leq M_\gamma(\mathcal{H}) + \sqrt{\text{Ldim}_\gamma(\mathcal{H}) \ln(T)} T
\]
Fat Littlestone’s dimension of linear separators

Linear predictors: \( \mathcal{H} = \{ x \mapsto \langle w, x \rangle : \| w \| \leq 1 \} \)
Fat Littlestone’s dimension of linear separators

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

Lemma

If $\mathcal{X}$ is the unit ball of a $\sigma$-regular Banach space $(B, \| \cdot \|_*)$, then

\[ \text{Ldim}_\gamma(\mathcal{H}) \leq \frac{\sigma}{\gamma^2} \]
Fat Littlestone’s dimension of linear separators

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

**Lemma**

If \( \mathcal{X} \) is the unit ball of a \( \sigma \)-regular Banach space \( (B, \| \cdot \|_*) \), then

\[
\text{Ldim}_\gamma(\mathcal{H}) \leq \frac{\sigma}{\gamma^2}
\]

**Examples:**

<table>
<thead>
<tr>
<th>( \mathcal{X} )</th>
<th>( \mathcal{H} )</th>
<th>( \text{Ldim}_\gamma(\mathcal{H}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( { x : | x |_2 \leq 1 } )</td>
<td>( { x \mapsto \langle w, x \rangle : | w |_2 \leq 1 } )</td>
<td>( \frac{1}{\gamma^2} )</td>
</tr>
<tr>
<td>( { x : | x |_\infty \leq 1 } )</td>
<td>( { x \mapsto \langle w, x \rangle : | w |_1 \leq 1 } )</td>
<td>( \frac{\log(n)}{\gamma^2} )</td>
</tr>
</tbody>
</table>
(Surprising) Corollary: Regret with non-convex loss

\[ M \leq M_\gamma(\mathcal{H}) + \frac{1}{\gamma} \sqrt{\ln(T) T} \]

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

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Online Learning</th>
<th>PAC Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Realizable case:</td>
<td>$\frac{\dim}{T}$</td>
<td>✓</td>
</tr>
<tr>
<td>Agnostic case:</td>
<td>$\sqrt{\frac{\dim}{T}}$</td>
<td>✓</td>
</tr>
<tr>
<td>Low noise:</td>
<td>$\frac{\dim}{T}$</td>
<td>✓</td>
</tr>
<tr>
<td>Margin:</td>
<td></td>
<td>✓</td>
</tr>
</tbody>
</table>
Some Open Problems

- $L_{dim}$ and fat-$L_{dim}$ 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 $L_{dim} \Rightarrow$ Compression scheme $\Rightarrow$ Low VCdim
- Low $L_{dim} \Leftrightarrow$ Compression scheme $\Leftrightarrow$ Low VCdim