Near-tight closure bounds for the Littlestone and threshold dimensions

Badih Ghazi[1] **Noah Golowich**[2] Ravi Kumar[1] Pasin Manurangsi[1]

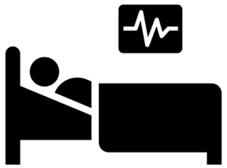
[1]Google Research

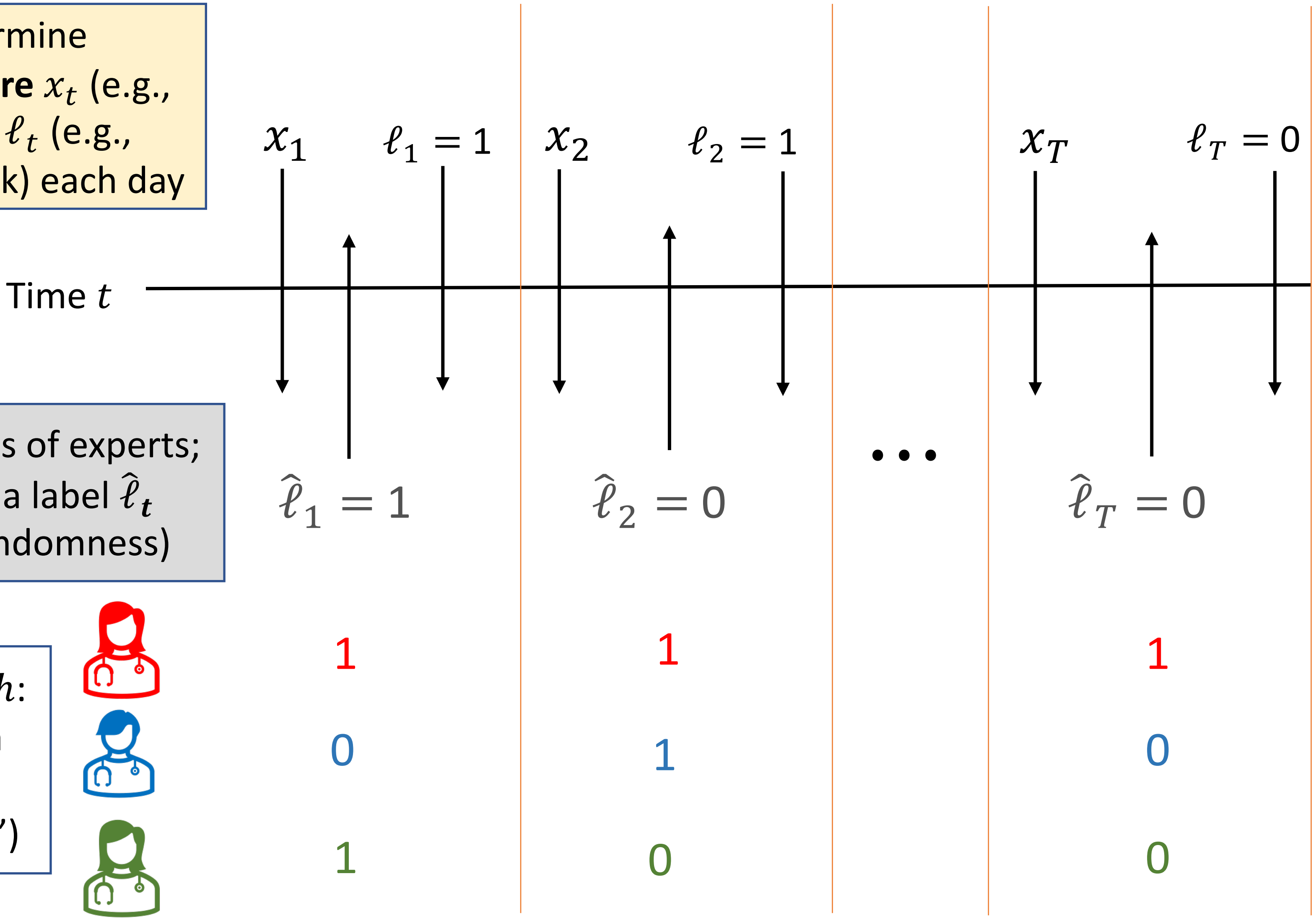
[2]MIT; work done while interning at Google Research

Motivation: closure properties for online learning

- Online binary prediction in the adversarial setting: [Ben-David-Pal-Shalev-Shwartz, '09] [Littlestone, '88]

Nature: determine adversarially **feature** x_t (e.g., vitals) and **label** ℓ_t (e.g., whether patient sick) each day





Learner: knows class of experts; given x_t, predict a label $\hat{\ell}_t$ (perhaps using randomness)

Experts: functions h: given x_t, output a label $h(x_t)$ ("sick" or "healthy")

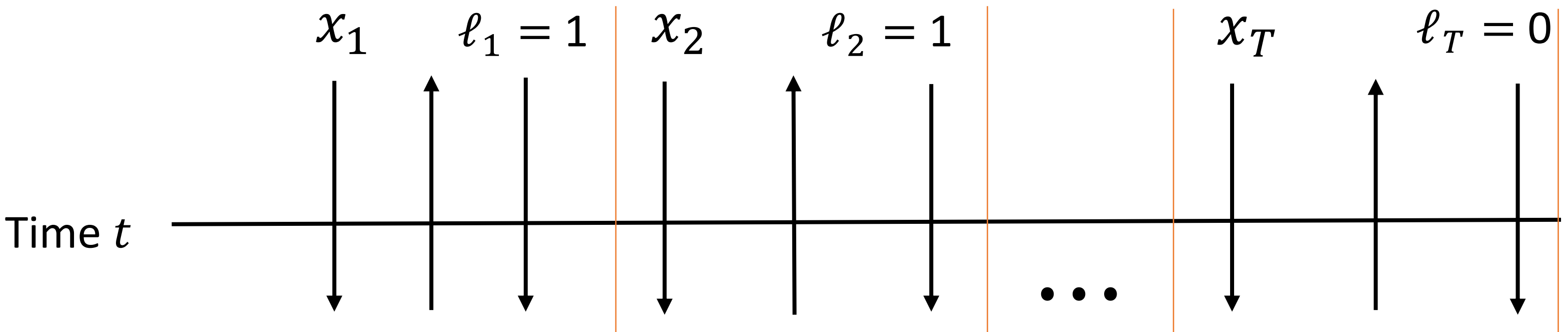
Learner's goal: minimize **regret R_T**:

$$R_T := \frac{1}{T}\sum_{t\le T}\mathbb{E}|\hat{\ell}_t - \ell_t| - \min_h \frac{1}{T}\sum_{t\le T}|\ell_t - h(x_t)|$$

Motivation: closure properties for online learning

- What happens when we combine predictions of experts?

Nature:



Learner:

	$t=1$	$t=2$	…	$t=T$
Learner	$\hat{\ell}_1 = 1$	$\hat{\ell}_2 = 0$		$\hat{\ell}_T = 0$
green ∨ blue	1	1		0
red ∨ green	1	1		1
red ∨ blue	1	1		1
red	1	1		1
blue	0	1		0
green	1	0		0

Add all pairwise-ORs of experts to the class

Experts: functions h

Learner's goal:
minimize **regret R_T:**

$$R_T := \frac{1}{T}\sum_{t\le T}\mathbb{E}|\hat{\ell}_t - \ell_t| - \min_h \frac{1}{T}\sum_{t\le T}|\ell_t - h(x_t)|$$

Regret R_T now **harder to bound** since minimizing over **larger** class of experts h (but small R_T means more)

Informal overview of results: tight closure bounds

- Fix any k-wise aggregation rule for experts: function $\{0,1\}^k \to \{0,1\}$
  - e.g., k-wise OR, k-wise AND, majority
- **What is the best regret bound for the class consisting of all possible k-wise aggregations of experts, in terms of that for the original class?**

Theorem (informal): regret blows up by at most factor $k \log k$ (& this is tight).

- Prior work [Alon-Beimel-Moran-Stemmer, '20]: blowup of $\leq 2^{2k} k^2$
- We also show: nearly tight upper bound on **threshold dimension** of class of k-wise aggregations of experts
  - Exponential improvement (in k) from [Alon-Beimel-Moran-Stemmer, '20]

Characterization of optimal regret

Learner's goal:
minimize **regret R_T**:

$$R_T := \frac{1}{T}\sum_{t \le T} \mathbb{E}|\hat{\ell}_t - \ell_t| - \min_h \frac{1}{T}\sum_{t \le T} |\ell_t - h(x_t)|$$

- Recall we are given a known set of experts (**hypotheses**) h
  - Call this set of all experts H
- *Given arbitrary H, what is the optimal regret bound R_T for any learner?*

$$\Omega(\sqrt{\mathrm{Ldim}(H)\,/\,T}) \le R_T \le O(\sqrt{\mathrm{Ldim}(H)\,/\,T})$$

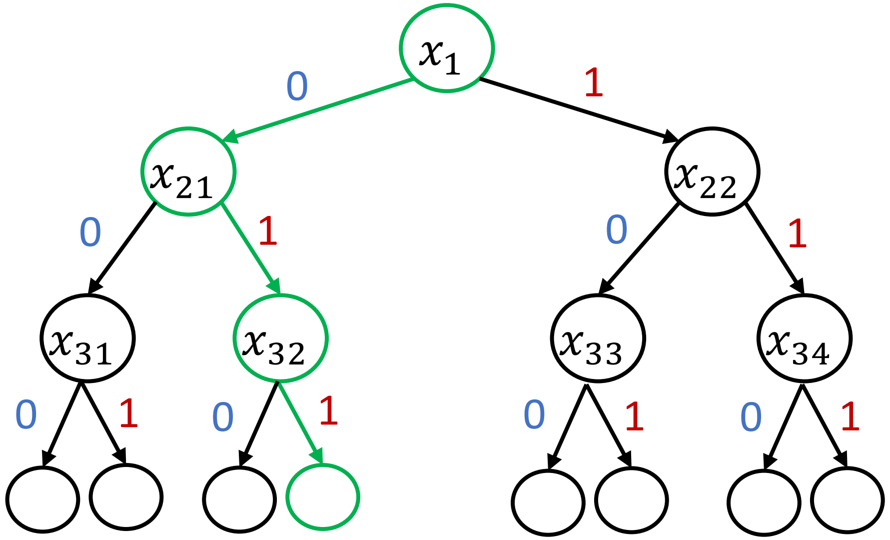
[Ben-David-Pal-Shalev-Shwartz, '09] (lower bound)

[Alon-Ben-Eliezer-Dagan-Moran-Naor-Yogev, '21] (upper bound)

[Ben-David-Pal-Shalev-Shwartz, '09]: $O(\sqrt{\mathrm{Ldim}(H)\log T\,/\,T})$

- $\mathrm{Ldim}(H)$ represents **Littlestone dimension**: a combinatorial parameter

Littlestone dimension: definition



Defn: For a binary tree with all internal nodes labeled by elements of X, edges labeled by $\{0,1\}$:

- It is **shattered** by H if for each leaf ℓ there is some $h_\ell \in H$ which labels all nodes on the root-to-leaf path for ℓ according to the labels on the edges.
- E.g., for the green leaf: need $h_\ell(x_1) = 0, h_\ell(x_{21}) = 1, h_\ell(x_{32}) = 1$.

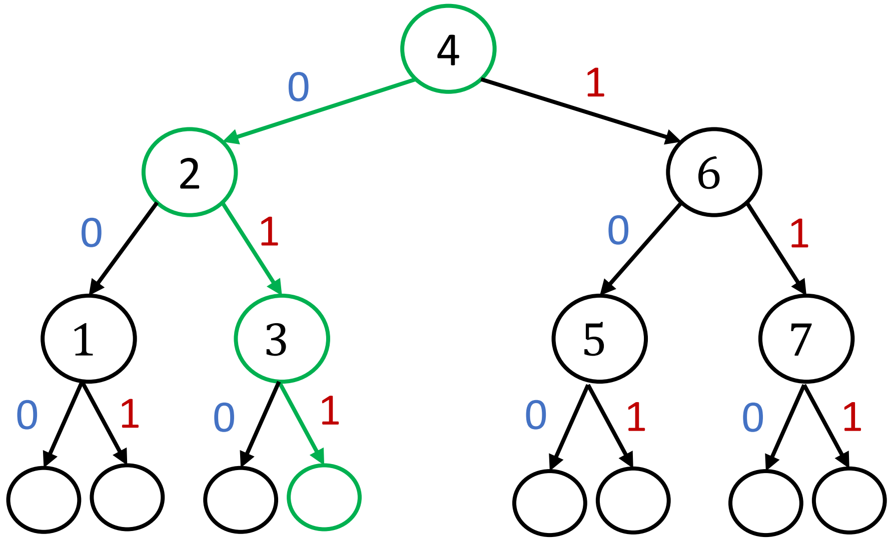
Defn: **Littlestone dimension** of hypothesis class H, denoted $\mathrm{Ldim}(H)$, is largest d so that there exists tree of depth d shattered by H.

- **Other applications beyond online learning**: Hypotheses classes H with a **private PAC learning algorithm** achieving error $o(1)$ are exactly those with finite Littlestone dimension *[Alon-Livni-Malliaris-Moran '19] [Bun-Livni-Moran '20]*

Examples: finite Littlestone dimension classes

- Any finite class H has Littlestone dimension $\mathrm{Ldim}(H) \leq \log(|H|)$
- Class of threshold functions $H_{\mathrm{thr},d}$ on $X = \{1, 2, \dots, 2^d\}$ has $\mathrm{Ldim}(H) = d$
  - 2^d such thresholds; threshold i evaluates to 1 on $j \in X$ iff $i \leq j$

Example of shattered tree for $d = 3$:



Green leaf corresponds to threshold which evaluates to 1 on x iff $x \leq 3$

- For general d: the range query (binary search) tree on $\{1, \dots, 2^d\}$ shows $\mathrm{Ldim}(H_{\mathrm{thr},d}) \geq d$

Results: closure properties for Littlestone dimension

- Data space X, $k \in \mathbb{N}$
- Binary hypothesis classes $H_1, \dots, H_k$ (i.e., consisting of $h : X \to \{0,1\}$)
- Aggregation rule $G : \{0,1\}^k \to \{0,1\}$
- **Defn**: $G(H_1, \dots, H_k) := \{x \mapsto G(h_1(x), \dots, h_k(x)) : h_1 \in H_1, \dots, h_k \in H_k\}$

k-wise aggregation via G

Theorem (closure property for Littlestone dimension): Suppose $\mathrm{Ldim}(H_i) \leq d$ for all $1 \leq i \leq k$. Then $\mathrm{Ldim}(G(H_1, \dots, H_k)) \leq O(d \cdot k \log k)$.

- Previous work: $\tilde{O}(2^{2^k} k^2 d)$ [Alon-Beimel-Moran-Stemmer, '20]
- Proof: 0-covering number for trees (similar to closure bound for VCdim)
- Let $G_{\mathrm{OR},k} : \{0,1\}^k \to \{0,1\}$ be the k-wise OR function:

Theorem (lower bound; tightness of above): There is a class H with:

1. $\mathrm{Ldim}(H) \leq d$.
2. $\mathrm{Ldim}\left(G_{\mathrm{OR},k}(H, \dots, H)\right) \geq \Omega(d \cdot k \log k)$.

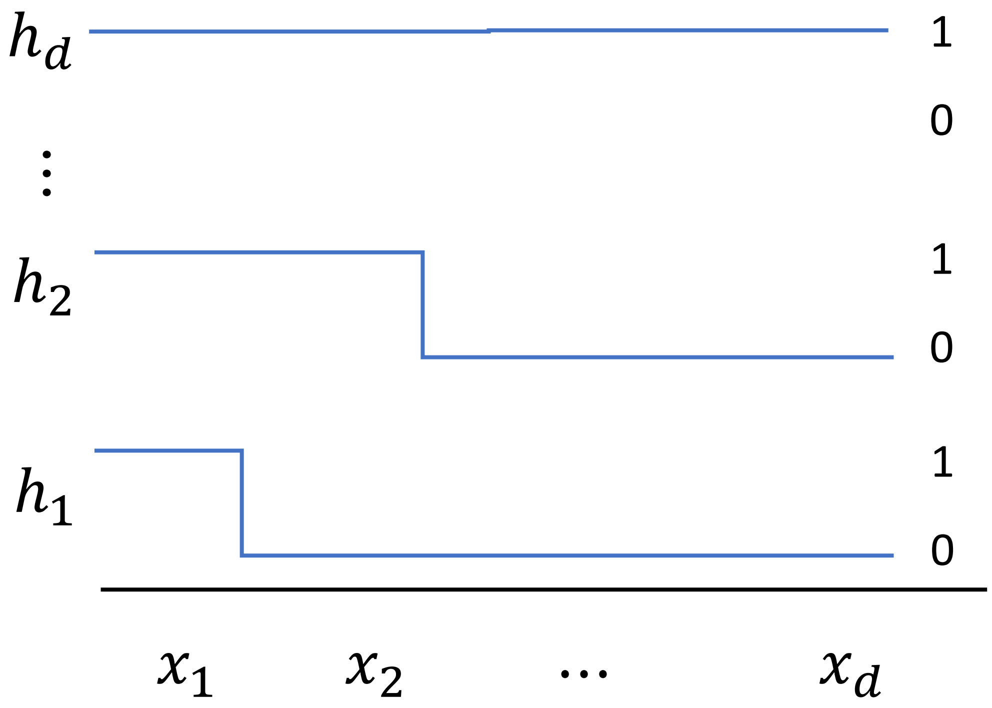
Threshold dimension: definition

- Fix X, and H consisting of $h : X \to \{0,1\}$.

Defn: **Threshold dimension** of hypothesis class H, denoted $\mathrm{Tdim}(H)$, is largest d so that there exists:

- $x_1, \dots, x_d \in X$;
- $h_1, \dots, h_d \in H$;

so that $h_i(x_j) = \mathbf{1}[j \le i]$ for all $1 \le i, j \le d$.



Motivation:

- Threshold dimension used to show finiteness of $\mathrm{Ldim}(H)$ *is* necessary for H to be privately PAC learnable
- In particular, following is used [Shelah, '78]: for any H,

$$\mathrm{Ldim}(H) \ge \lfloor \log \mathrm{Tdim}(H) \rfloor, \qquad \mathrm{Tdim}(H) \ge \lfloor \log \mathrm{Ldim}(H) \rfloor$$

Tight: class of thresholds on $\{1, \dots, 2^d\}$ has $\mathrm{Ldim} = d$, $\mathrm{Tdim} = 2^d$

Unknown if tight

Results: closure properties for threshold dimension

Recall: for binary hypothesis classes $H_1, \dots, H_k$:

$$G(H_1, \dots, H_k) \coloneqq \{x \mapsto G\big(h_1(x), \dots, h_k(x)\big): h_1 \in H_1, \dots, h_k \in H_k\}$$

Theorem (closure property for threshold dimension): Suppose $\mathrm{Tdim}(H_i) \le d$ for all $1 \le i \le k$. Then $\mathrm{Tdim}\big(G(H_1, \dots, H_k)\big) \le 2^{O(d \cdot k \log k)}$.

- Previous work: upper bound of $2^{d \cdot 4k \cdot 4^k}$ [Alon-Beimel-Moran-Stemmer, '20]

Theorem (lower bound; near-tightness of above): For any $k \in \mathbb{N}$, there are classes $H_1, \dots, H_k$ and a function $G : \{0,1\}^k \to \{0,1\}$ so that:

1. $\mathrm{Tdim}(H_i) \le d$ for all $1 \le i \le k$.
2. $\mathrm{Tdim}\big(G(H_1, \dots, H_k)\big) \ge 2^{\Omega(dk)}$.

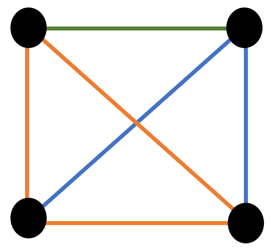
- Previous work: lower bound of $2^{\Omega(d)}$ [Alon-Beimel-Moran-Stemmer, '20]

Proof of upper bound (& improving the lower bound)

- For $N \in \mathbb{N}$, let K_N be complete graph on N vertices

Defn: For $r, c \in \mathbb{N}$, define **Ramsey number** $R_c(r)$ as minimum $N \in \mathbb{N}$ so that for any coloring of edges of K_N with c colors, there exists a monochromatic (1-colored) clique of size r.

"Clique" is complete graph (subgraph of K_N)
e.g., orange triangle is monochromatic clique of size 3



- **Ramsey's theorem**: $R_c(r) \leq 2^{r \cdot c \log c}$
- **Proof of closure upper bound**: upper bound on $R_c(r)$ implies upper bound on $\mathrm{Tdim}(G(H_1, \dots, H_k))$
- **Contrapositive**: if we have $H_1, \dots, H_k$ with, $\forall i$, $\mathrm{Tdim}(H_i) \leq d$ but $\mathrm{Tdim}(G(H_1, \dots, H_k)) \geq 2^{\boldsymbol{k \cdot \alpha(k)}}$ for some $\alpha(k) \to \infty$, then:

$$\limsup_{c \to \infty} R_c(2d+1)^{1/c} = \infty$$

Thm (upper bound): Suppose $\mathrm{Tdim}(H_i) \leq d$ for $i \leq k$. Then $\mathrm{Tdim}(G(H_1, \dots, H_k)) \leq 2^{O(d \cdot k \log k)}$.

Thm (lower bound): There are classes $H_1, \dots, H_k$ and $G : \{0,1\}^k \to \{0,1\}$ so that:

1. $\mathrm{Tdim}(H_i) \leq d$ for all $1 \leq i \leq k$.
2. $\mathrm{Tdim}(G(H_1, \dots, H_k)) \geq 2^{\boldsymbol{\Omega(dk)}}$.

(Would resolve long-standing open problem in Ramsey theory) ⇒ lower bound above (probably) hard to improve

Summary: overview of results

Throughout: d defined as upper bound on Littlestone/threshold dimension on $H_1, \dots, H_k$

	Upper bound (for any G and $H_1, \dots, H_k$, upper bound on $\dim(G(H_1, \dots, H_k))$)	**Lower bound** (there exist G and $H_1, \dots, H_k$, so that we can lower bound $\dim(G(H_1, \dots, H_k))$)
Littlestone dimension	$O(d \cdot k \log k)$	$\Omega(d \cdot k \log k)$
Threshold dimension	$2^{O(d \cdot k \log k)}$	$2^{\Omega(d \cdot k)}$

Thank you for listening!