The Learning Complexity of Subsequence Detection

Mason DiCicco, Daniel Reichman
Worcester Polytechnic Institute

June 26, 2023

Subsequences

A **subsequence** of a **binary string** x is obtained by *deleting* bits from x.

100101110010000000000010101111111111111100101000

$\downarrow$

100110000000000101011111111 01000

Subsequences

A **subsequence** of a **binary string** x is obtained by *deleting* bits from x.

$$1001011100100000000000101011111111111111100101000$$
$$\downarrow$$
$$10011000000000001010111111111101000$$

> In this talk, we will **lower bound** the **number of samples** required to **learn a binary string** from subsequences/supersequences.

Classifiers

Let $\mathcal{H}_k^n$ denote the **hypothesis class** of **length-k subsequence classifiers:**

$$\mathcal{H}_k^n := \{h_y : \{0,1\}^n \to \{0,1\} \mid y \in \{0,1\}^k\},$$

where

$$h_y(x) = \begin{cases} 1 & y \text{ is a subsequence of } x \\ 0 & \text{otherwise} \end{cases}$$

Classifiers

Let $\mathcal{H}_k^n$ denote the **hypothesis class** of **length-k subsequence classifiers:**

$$\mathcal{H}_k^n := \{h_y : \{0,1\}^n \to \{0,1\} \mid y \in \{0,1\}^k\},$$

where

$$h_y(x) = \begin{cases} 1 & y \text{ is a subsequence of } x \\ 0 & \text{otherwise} \end{cases}$$

Example

$$h_0(101) = 1 \quad h_{10}(0011) = 0$$
$$h_{000}(000) = 1 \quad h_{111}(11) = 0$$

Learning Complexity

The PAC learning framework ([Shalev-Shwartz and Ben-David, 2014])

Learning Complexity

The PAC learning framework ([Shalev-Shwartz and Ben-David, 2014])

Fix a secret $h_y \in \mathcal{H}_k^n$ and a distribution $\mathcal{D}$ over $\{0, 1\}^n$.

The PAC learning framework ([Shalev-Shwartz and Ben-David, 2014])

Fix a secret $h_y \in \mathcal{H}_k^n$ and a distribution $\mathcal{D}$ over $\{0,1\}^n$.

Given a **sample** $x^{(1)}, \cdots, x^{(m)} \sim \mathcal{D}$ with **labels** $h_y(x^{(1)}), \cdots, h_y(x^{(m)})$,

The PAC learning framework ([Shalev-Shwartz and Ben-David, 2014])

Fix a secret $h_y \in \mathcal{H}_k^n$ and a distribution $\mathcal{D}$ over $\{0,1\}^n$.

Given a **sample** $x^{(1)}, \cdots, x^{(m)} \sim \mathcal{D}$ with **labels** $h_y(x^{(1)}), \cdots, h_y(x^{(m)})$,

return an $h_z \in \mathcal{H}_k^n$ approximating h_y:

$$\mathbb{P}_{x\sim\mathcal{D}}(h_z(x) \neq h_y(x)) \leq \epsilon$$

Learning Complexity

The PAC learning framework ([Shalev-Shwartz and Ben-David, 2014])

Fix a secret $h_y \in \mathcal{H}_k^n$ and a distribution $\mathcal{D}$ over $\{0,1\}^n$.

Given a **sample** $x^{(1)}, \cdots, x^{(m)} \sim \mathcal{D}$ with **labels** $h_y(x^{(1)}), \cdots, h_y(x^{(m)})$,

return an $h_z \in \mathcal{H}_k^n$ approximating h_y:

$$\mathbb{P}_{x \sim \mathcal{D}}(h_z(x) \neq h_y(x)) \leq \epsilon$$

Sample complexity: How many samples are needed to learn h_z (w.h.p.)?

The PAC learning framework ([Shalev-Shwartz and Ben-David, 2014])

Fix a secret $h_y \in \mathcal{H}_k^n$ and a distribution $\mathcal{D}$ over $\{0,1\}^n$.

Given a **sample** $x^{(1)}, \cdots, x^{(m)} \sim \mathcal{D}$ with **labels** $h_y(x^{(1)}), \cdots, h_y(x^{(m)})$,

return an $h_z \in \mathcal{H}_k^n$ approximating h_y:

$$\mathbb{P}_{x \sim \mathcal{D}}(h_z(x) \neq h_y(x)) \leq \epsilon$$

Sample complexity: How many samples are needed to learn h_z (w.h.p.)?

(Note: this resembles the **trace reconstruction** problem of [Batu et al., 2004])

Learning Complexity

Theorem ([Ehrenfeucht et al., 1989])

The *sample complexity* of PAC learning any family $\mathcal{H}$ (with failure probability δ) is

$$\Theta_{\epsilon,\delta}(\mathrm{VCdim}(\mathcal{H})),$$

where $\mathrm{VCdim}(\mathcal{H})$ denotes the *Vapnik–Chervonenkis dimension* of $\mathcal{H}$.

VC Dimension

Definition

Let $\mathcal{H}$ be a family of classifiers for some domain $\mathcal{X}$ (i.e., functions $f : \mathcal{X} \to \{0, 1\}$.)

VC Dimension

Definition

Let $\mathcal{H}$ be a family of classifiers for some domain $\mathcal{X}$ (i.e., functions $f : \mathcal{X} \to \{0, 1\}$.)

A subset $S \subseteq \mathcal{X}$ is **shattered** if every $A \subseteq S$ is *realized* by a unique $f_A \in \mathcal{H}$:

$$f_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \in S \setminus A, \end{cases}$$

VC Dimension

Definition

Let $\mathcal{H}$ be a family of classifiers for some domain $\mathcal{X}$ (i.e., functions $f : \mathcal{X} \to \{0, 1\}$.)

A subset $S \subseteq \mathcal{X}$ is **shattered** if every $A \subseteq S$ is *realized* by a unique $f_A \in \mathcal{H}$:

$$f_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \in S \setminus A, \end{cases}$$

The **VC dimension** of $\mathcal{H}$ is equal to the size of the *largest shattered subset of* $\mathcal{X}$.

VC Dimension

Definition

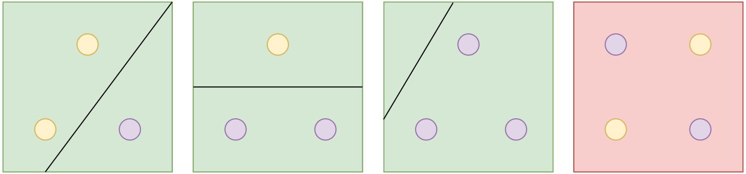
Let $\mathcal{H}$ be a family of classifiers for some domain $\mathcal{X}$ (i.e., functions $f : \mathcal{X} \to \{0, 1\}$.)

A subset $S \subseteq \mathcal{X}$ is **shattered** if every $A \subseteq S$ is *realized* by a unique $f_A \in \mathcal{H}$:

$$f_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \in S \setminus A, \end{cases}$$

The **VC dimension** of $\mathcal{H}$ is equal to the size of the *largest shattered subset of* $\mathcal{X}$.

Example (Linear separators on $\mathbb{R}^2$ have VC dimension 3)



VC Dimension

Main Theorem

$\text{VCdim}(\mathcal{H}_k^n) = \Theta(k)$

Main Theorem

$\mathrm{VCdim}(\mathcal{H}_k^n) = \Theta(k)$

Proof (Outline)

- Define *disjointness classifiers* $\mathcal{J}_k^n$

VC Dimension

Main Theorem

$\text{VCdim}(\mathcal{H}_k^n) = \Theta(k)$

Proof (Outline)

- Define *disjointness classifiers* $\mathcal{J}_k^n$
- Prove $\mathcal{J}_k^n \geq k$ for $k \leq n/2$.

VC Dimension

Main Theorem

$\text{VCdim}(\mathcal{H}_k^n) = \Theta(k)$

Proof (Outline)

- Define *disjointness classifiers* $\mathcal{J}_k^n$
- Prove $\mathcal{J}_k^n \geq k$ for $k \leq n/2$.
- Use a **reduction** to show $\text{VCdim}(\mathcal{H}_{2n+k}^{3n}) \geq \text{VCdim}(\mathcal{J}_k^n)$

VC Dimension

Main Theorem

$\text{VCdim}(\mathcal{H}_k^n) = \Theta(k)$

Proof (Outline)

- Define *disjointness classifiers* $\mathcal{J}_k^n$
- Prove $\mathcal{J}_k^n \geq k$ for $k \leq n/2$.
- Use a **reduction** to show $\text{VCdim}(\mathcal{H}_{2n+k}^{3n}) \geq \text{VCdim}(\mathcal{J}_k^n)$
- Padding argument: $\text{VCdim}(\mathcal{H}_{2n+k}^{3n}) \longrightarrow \text{VCdim}(\mathcal{H}_k^n)$

□

Disjointness Classifiers

Let $\mathcal{J}_k^n$ denote the hypothesis class of **size-k disjointness classifiers**:

$$\mathcal{J}_k^n := \left\{ d_A : 2^{[n]} \to \{0,1\} \;\middle|\; A \in \binom{[n]}{k} \right\}$$

where

$$d_A(B) = \begin{cases} 1 & A \cap B = \emptyset \\ 0 & \text{otherwise} \end{cases}$$

Disjointness Classifiers

Let $\mathcal{J}_k^n$ denote the hypothesis class of **size-k disjointness classifiers**:

$$\mathcal{J}_k^n := \left\{ d_A : 2^{[n]} \to \{0,1\} \;\middle|\; A \in \binom{[n]}{k} \right\}$$

where

$$d_A(B) = \begin{cases} 1 & A \cap B = \emptyset \\ 0 & \text{otherwise} \end{cases}$$

[Kremer et al., 1999] looked at this set **without the restriction** that $|A| = k$:

$$\mathcal{J}^n = \mathcal{J}_1^n \cup \cdots \cup \mathcal{J}_n^n$$

to prove that $R^{A \to B}(\mathrm{DISJ}^n) \geq \Omega(\mathrm{VCdim}(\mathcal{J}^n)) = \Omega(n)$.

Disjointness Classifiers

Lemma ([Kremer et al., 1999])

$\mathrm{VCdim}(\mathcal{J}^n) = n$

Proof.

The set of singletons $S = \{\{1\}, \cdots, \{n\}\} \subset 2^{[n]}$ is shattered.

Disjointness Classifiers

Lemma ([Kremer et al., 1999])

$\text{VCdim}(\mathcal{J}^n) = n$

Proof.

The set of singletons $S = \{\{1\}, \cdots, \{n\}\} \subset 2^{[n]}$ is shattered.

Indeed, any $B = \{\{i_1\}, \cdots, \{i_m\}\} \subseteq S$ is realized by $A = [n] \setminus \{i_1, \cdots, i_m\} \subseteq [n]$.

Disjointness Classifiers

Lemma ([Kremer et al., 1999])

$\text{VCdim}(\mathcal{J}^n) = n$

Proof.

The set of singletons $S = \{\{1\}, \cdots, \{n\}\} \subset 2^{[n]}$ is shattered.

Indeed, any $B = \{\{i_1\}, \cdots, \{i_m\}\} \subseteq S$ is realized by $A = [n] \setminus \{i_1, \cdots, i_m\} \subseteq [n]$.

A	B
$\{1, 2\}$	$\emptyset$
$\{2\}$	$\{\{1\}\}$
$\{1\}$	$\{\{2\}\}$
$\{\ \}$	$\{\{1\}, \{2\}\}$

VC Dimension of $\mathcal{J}_k^n$ (Example)

Example ($n = 4,\ k = 2$)

If we want every A to have the same size, we can do this:

A	B		A	B
$\{1, 2\}$	$\emptyset$		$\{1, 2\}$	$\emptyset$
$\{2\}$	$\{\{1\}\}$	$\longrightarrow$	$\{2, 3\}$	$\{\{1\}\}$
$\{1\}$	$\{\{2\}\}$		$\{1, 3\}$	$\{\{2\}\}$
$\{\ \}$	$\{\{1\}, \{2\}\}$		$\{3, 4\}$	$\{\{1\}, \{2\}\}$

VC Dimension of $\mathcal{J}_k^n$

Lemma

$\text{VCdim}(\mathcal{J}_k^n) \geq k$ for $k \leq n/2$.

VC Dimension of $\mathcal{J}_k^n$

Lemma
$\text{VCdim}(\mathcal{J}_k^n) \geq k$ for $k \leq n/2$.

Proof.
We can still shatter a set of k singletons: $S = \{\{1\}, \{2\}, \cdots, \{k\}\}$.

□

VC Dimension of $\mathcal{J}_k^n$

Lemma

$\text{VCdim}(\mathcal{J}_k^n) \geq k$ for $k \leq n/2$.

Proof.

We can still shatter a set of k singletons: $S = \{\{1\}, \{2\}, \cdots, \{k\}\}$.

Indeed, every $B = \{\{i_1\}, \cdots, \{i_m\}\} \subseteq S$ is realized by $A = [k] \setminus \{i_1, \cdots, i_m\} \subseteq [k]$.

□

VC Dimension of $\mathcal{J}_k^n$

Lemma

$\text{VCdim}(\mathcal{J}_k^n) \geq k$ for $k \leq n/2$.

Proof.

We can still shatter a set of k singletons: $S = \{\{1\}, \{2\}, \cdots, \{k\}\}$.

Indeed, every $B = \{\{i_1\}, \cdots, \{i_m\}\} \subseteq S$ is realized by $A = [k] \setminus \{i_1, \cdots, i_m\} \subseteq [k]$.

> **But,** A must contain *exactly* k elements.

□

VC Dimension of $\mathcal{J}_k^n$

Lemma

$\text{VCdim}(\mathcal{J}_k^n) \geq k$ for $k \leq n/2$.

Proof.

We can still shatter a set of k singletons: $S = \{\{1\}, \{2\}, \cdots, \{k\}\}$.

Indeed, every $B = \{\{i_1\}, \cdots, \{i_m\}\} \subseteq S$ is realized by $A = [k] \setminus \{i_1, \cdots, i_m\} \subseteq [k]$.

But, A must contain *exactly* k elements.

$\Rightarrow$ Add some unused elements from $\{\{k+1\}, \cdots, \{n\}\}$ to A as "padding".

□

VC Dimension of $\mathcal{J}_k^n$

Lemma

$\text{VCdim}(\mathcal{J}_k^n) \geq k$ for $k \leq n/2$.

Proof.

We can still shatter a set of k singletons: $S = \{\{1\}, \{2\}, \cdots, \{k\}\}$.

Indeed, every $B = \{\{i_1\}, \cdots, \{i_m\}\} \subseteq S$ is realized by $A = [k] \setminus \{i_1, \cdots, i_m\} \subseteq [k]$.

But, A must contain *exactly* k elements.

$\Rightarrow$ Add some unused elements from $\{\{k+1\}, \cdots, \{n\}\}$ to A as "padding".

In the worst case, when $B = S$, we need k padding elements for $A = \emptyset$.

□

VC Dimension of $\mathcal{J}_k^n$

Lemma

$\text{VCdim}(\mathcal{J}_k^n) \geq k$ for $k \leq n/2$.

Proof.

We can still shatter a set of k singletons: $S = \{\{1\}, \{2\}, \cdots, \{k\}\}$.

Indeed, every $B = \{\{i_1\}, \cdots, \{i_m\}\} \subseteq S$ is realized by $A = [k] \setminus \{i_1, \cdots, i_m\} \subseteq [k]$.

But, A must contain *exactly* k elements.

$\Rightarrow$ Add some unused elements from $\{\{k+1\}, \cdots, \{n\}\}$ to A as "padding".

In the worst case, when $B = S$, we need k padding elements for $A = \emptyset$.

This is possible when $|\{\{k+1\}, \cdots, \{n\}\}| \geq k$ (i.e., $k \leq n/2$). □

Reduction

Theorem
$\mathrm{VCdim}(\mathcal{H}^{3n}_{2n+k}) \geq \mathrm{VCdim}(\mathcal{J}^{n}_{k}) \geq k$

Proof Idea:
There exist maps $\rho : \binom{[n]}{k} \to \{0,1\}^{2n+k}$ and $\phi : 2^{[n]} \to \{0,1\}^{3n}$ such that

$$d_A(B) = h_{\rho(A)}(\phi(B))$$

for all A, B.

Reduction

Theorem

$\mathrm{VCdim}(\mathcal{H}^{3n}_{2n+k}) \geq \mathrm{VCdim}(\mathcal{J}^n_k) \geq k$

Proof Idea:

There exist maps $\rho : \binom{[n]}{k} \to \{0,1\}^{2n+k}$ and $\phi : 2^{[n]} \to \{0,1\}^{3n}$ such that

$$d_A(B) = h_{\rho(A)}(\phi(B))$$

for all A, B.

$\implies$ Shattered sets under $\mathcal{J}^n_k$ map to shattered sets under $\mathcal{H}^{3n}_{2n+k}$.

Reduction from Disjointness

Proof.

Let $a, b \in \{0,1\}^n$ denote the characteristic vectors of $A, B \subseteq [n]$.

Reduction from Disjointness

Proof.

Let $a, b \in \{0,1\}^n$ denote the characteristic vectors of $A, B \subseteq [n]$.

- ϕ independently maps $0 \mapsto 010$ and $1 \mapsto 100$.

b	0	1	0	1	$\cdots$
$\phi(b)$	010	100	010	100	$\cdots$

- ρ independently maps $0 \mapsto 00$ and $1 \mapsto 010$

a	0	0	1	1	$\cdots$
$\rho(a)$	00	00	010	010	$\cdots$

Reduction from Disjointness

Proof.

Let $a, b \in \{0,1\}^n$ denote the characteristic vectors of $A, B \subseteq [n]$.

- ϕ independently maps $0 \mapsto 010$ and $1 \mapsto 100$.

b	0	1	0	1	$\cdots$
$\phi(b)$	010	100	010	100	$\cdots$

- ρ independently maps $0 \mapsto 00$ and $1 \mapsto 010$

a	0	0	1	1	$\cdots$
$\rho(a)$	00	00	010	010	$\cdots$

Note that every cell has **exactly two zeros**.

Reduction from Disjointness (Example)

Let $a = 0011$ and $b = 0101$.

Reduction from Disjointness (Example)

Let $a = 0011$ and $b = 0101$.

- $\phi(b) = 010\ 100\ 010\ 100$,

Reduction from Disjointness (Example)

Let $a = 0011$ and $b = 0101$.

- $\phi(b) = 010\ 100\ 010\ 100$,
- $\rho(a) = \ 00 \quad 00 \ \ 010\ 010$

Reduction from Disjointness (Example)

Let $a = 0011$ and $b = 0101$.

- $\phi(b) = 010\ 100\ 010\ 100$,
- $\rho(a) = \ 00 \quad 00 \ \ 010\ 010$

$\rho(a)$ is not a subsequence of $\phi(b)$.

Reduction from Disjointness (Example)

Let $a = 0011$ and $b = 0101$.

- $\phi(b) = 010\ 100\ 010\ 100$,
- $\rho(a) = \ 00 \quad 00 \ \ 010\ 010$

$\rho(a)$ is not a subsequence of $\phi(b)$.

Let $a = 0011$ and $b = 1100$.

- $\phi(b) = 100\ 100\ 010\ 010$,
- $\rho(a) = \ 00 \quad 00 \ \ 010\ 010$

$\rho(a)$ is a subsequence of $\phi(b)$.

Reduction from Disjointness

b	0	1	0	1	$\cdots$
$\phi(b)$	010	100	010	100	$\cdots$

a	0	0	1	1	$\cdots$
$\rho(a)$	00	00	010	010	$\cdots$

Proof (cont.)

- If a and b are disjoint, then $\rho(a)$ is a subsequence of $\phi(b)$ (column-wise).

Reduction from Disjointness

b	0	1	0	1	$\cdots$
$\phi(b)$	010	100	010	100	$\cdots$

a	0	0	1	1	$\cdots$
$\rho(a)$	00	00	010	010	$\cdots$

Proof (cont.)

- If a and b are disjoint, then $\rho(a)$ is a subsequence of $\phi(b)$ (column-wise).
- Otherwise, partition $\phi(b)$ and $\rho(a)$ around the offending index i with $a_i = b_i = 1$.

$$\phi(b) = \beta \cdot 100 \cdot \delta$$

$$\rho(a) = \alpha \cdot 010 \cdot \gamma$$

Reduction from Disjointness

b	0	1	0	1	$\cdots$
$\phi(b)$	010	100	010	100	$\cdots$

a	0	0	1	1	$\cdots$
$\rho(a)$	00	00	010	010	$\cdots$

Proof (cont.)

- If a and b are disjoint, then $\rho(a)$ is a subsequence of $\phi(b)$ (column-wise).
- Otherwise, partition $\phi(b)$ and $\rho(a)$ around the offending index i with $a_i = b_i = 1$.

$$\phi(b) = \beta \cdot 100 \cdot \delta$$

$$\rho(a) = \alpha \cdot 010 \cdot \gamma$$

Two zeros in every cell $\implies$ $\rho(a)$ cannot be a subsequence of $\phi(b)$ $\square$

Reduction from Disjointness

b	0	1	0	1	$\cdots$
$\phi(b)$	010	100	010	100	$\cdots$

a	0	0	1	1	$\cdots$
$\rho(a)$	00	00	010	010	$\cdots$

Proof (cont.)

Thus, $\rho(a)$ is a subsequence of $\phi(b)$ **if and only if** a is disjoint from b.

Reduction from Disjointness

b	0	1	0	1	$\cdots$
$\phi(b)$	010	100	010	100	$\cdots$
a	0	0	1	1	$\cdots$
$\rho(a)$	00	00	010	010	$\cdots$

Proof (cont.)

Thus, $\rho(a)$ is a subsequence of $\phi(b)$ **if and only if** a is disjoint from b.

Note that $|\phi(b)| = 3n$ and $|\rho(a)| = 2n + k$.

Reduction from Disjointness

b	0	1	0	1	$\cdots$
$\phi(b)$	010	100	010	100	$\cdots$
a	0	0	1	1	$\cdots$
$\rho(a)$	00	00	010	010	$\cdots$

Proof (cont.)

Thus, $\rho(a)$ is a subsequence of $\phi(b)$ **if and only if** a is disjoint from b.

Note that $|\phi(b)| = 3n$ and $|\rho(a)| = 2n + k$.

$\implies$ $S \subseteq 2^{[n]}$ is shattered by $\mathcal{J}_k^n$ **if and only if** $\phi(S) \in \{0,1\}^{3n}$ is shattered by $\mathcal{H}_{2n+k}^{3n}$.

□

Padding

Note that we may “pad” with ones

Padding

Note that we may "pad" with ones

- ϕ independently maps $0 \mapsto 010$ and $1 \mapsto 100$.

b	0	1	0	1	$\cdots$	
$\phi(b)$	010	100	010	100	$\cdots$	1^N

- ρ independently maps $0 \mapsto 00$ and $1 \mapsto 010$

a	0	0	1	1	$\cdots$	
$\rho(a)$	00	00	010	010	$\cdots$	

Conclusion

Theorem

For all $n \geq \frac{6}{5}k \geq 0$,

$$\frac{k}{5} \leq \mathrm{VCdim}(\mathcal{H}_k^n) \leq k$$

Conclusion

Theorem

For all $n \geq \frac{6}{5}k \geq 0$,

$$\frac{k}{5} \leq \text{VCdim}(\mathcal{H}_k^n) \leq k$$

Proof.

Recall $\text{VCdim}(\mathcal{H}_{2n+k}^{3n+N}) \geq \text{VCdim}(\mathcal{J}_k^n) \geq k$ **only when** $k \leq n/2$.

So, we substitute $n = 2k$ to get

$$k \leq \text{VCdim}(\mathcal{H}_{5k}^{6k+N}) \leq 5k$$

□

Final Comments

Other results

- The VC dimension of **supersequence** classifiers is also $\Omega(k)$.
- The **communication complexity** of subsequence detection is $\Theta(k)$.
- The **threshold circuit complexity** of subsequence detection is $\Omega(k)$.

Open Questions

- What happens with **larger alphabets?**
- How is **contiguity** related to learning complexity?
  - **Contiguous** string-matching is $\tilde{O}(\log|\Sigma| \log k)$ [Golovnev et al., 2019].

Thank You

Questions?

References I

Batu, T., Kannan, S., Khanna, S., and McGregor, A. (2004).
Reconstructing strings from random traces.
Departmental Papers (CIS), page 173.

Ehrenfeucht, A., Haussler, D., Kearns, M., and Valiant, L. (1989).
A general lower bound on the number of examples needed for learning.
Information and Computation, 82(3):247–261.

Golovnev, A., Göös, M., Reichman, D., and Shinkar, I. (2019).
String matching: Communication, circuits, and learning.
In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

References II

Kremer, I., Nisan, N., and Ron, D. (1999).
On randomized one-round communication complexity.
Computational Complexity, 8(1):21–49.

Shalev-Shwartz, S. and Ben-David, S. (2014).
Understanding machine learning: From theory to algorithms.
Cambridge university press.