Dense Subsets of Pseudorandom Sets

O. Reingold\textsuperscript{1} \hspace{1cm} L. Trevisan\textsuperscript{2} \hspace{1cm} M. Tulsiani\textsuperscript{2} \hspace{1cm} S. Vadhan\textsuperscript{3}

\textsuperscript{1}Weizmann Institute

\textsuperscript{2}UC Berkeley

\textsuperscript{3}Harvard University
Theorem (Szemerédi 1975)

Any set of $A$ of $\delta N$ integers in $\{1, \ldots, N\}$ contains a length $k$-AP if $N$ is large enough.
Theorem (Szemerédi 1975)

Any set of $A$ of $\delta N$ integers in $\{1, \ldots, N\}$ contains a length $k$-AP if $N$ is large enough.

Theorem (Green-Tao 2004)

The set of primes in $\{1, \ldots, N\}$ contains a length $k$-AP if $N$ is large enough.
Progressions in Subsets of Integers

Theorem (Szemeredi 1975)

Any set of $A$ of $\delta N$ integers in $\{1, \ldots, N\}$ contains a length $k$-AP if $N$ is large enough.

Theorem (Green-Tao 2004)

The set of primes in $\{1, \ldots, N\}$ contains a length $k$-AP if $N$ is large enough.

Green-Tao showed that a property of dense subsets of the integers (having progressions) also holds for the primes.
Thm 1  There is a pseudorandom set $R \subseteq \{1, \ldots, N\}$ such that primes have constant density in $R$. 
Thm 1 There is a pseudorandom set $R \subseteq \{1, \ldots, N\}$ such that primes have constant density in $R$. 
**The Green-Tao Proof**

**Thm 1** There is a pseudorandom set $R \subseteq \{1, \ldots, N\}$ such that primes have constant density in $R$.

**Thm 2** If $R$ is a pseudorandom subset of $\{1, \ldots, N\}$ and if $D$ is a dense subset i.e. $|D| \geq \delta R$, then $D$ contains a length $k$-AP.

![Diagram showing $R \subseteq \{1, \ldots, N\}$ with $2, 3, 5, \ldots$ contained within $R$.]
The Green-Tao Proof

Thm 1 There is a pseudorandom set $R \subseteq \{1, \ldots, N\}$ such that primes have constant density in $R$.

Thm 2 If $R$ is a pseudorandom subset of $\{1, \ldots, N\}$ and if $D$ is a dense subset i.e. $|D| \geq \delta R$, then $D$ contains a length $k$-AP.
Proof of Theorem 2

If $D$ is a dense in a pseudorandom set $R$ ($|D| \geq \delta |R|$), then there is a dense model set $M$ ($|M| \geq \delta N$) indistinguishable from $D$. 

“A dense subset of a pseudorandom set has a dense model.” Can we prove this in general?
Proof of Theorem 2

- If $D$ is a dense in a pseudorandom set $R$ ($|D| \geq \delta |R|$), then there is a dense model set $M$ ($|M| \geq \delta N$) indistinguishable from $D$.

- $M$ must contain length $k$-APs (Szemeredi). So does $D$.
If $D$ is a dense in a pseudorandom set $R$ ($|D| \geq \delta|R|$), then there is a dense model set $M$ ($|M| \geq \delta N$) indistinguishable from $D$.

$M$ must contain length $k$-APs (Szemeredi). So does $D$.

“A dense subset of a pseudorandom set has a dense model.” Can we prove this in general?
A finite universe $X$ (e.g. $\{1, \ldots, N\}$, $\{0, 1\}^n$).

A family of distinguishers $\mathcal{F} = \{f : X \rightarrow \{0, 1\}\}$ (e.g. Circuits of size $s$).
Abstracting out...

- A finite universe $X$ (e.g. $\{1, \ldots, N\}, \{0, 1\}^n$).
- A family of distinguishers $\mathcal{F} = \{ f : X \rightarrow \{0, 1\} \}$ (e.g. Circuits of size $s$).
- Distributions $A$ and $B$ are $\epsilon$-indistinguishable by $\mathcal{F}$ if
  \[ \forall f \in \mathcal{F} \left| \mathbb{E}_f(A) - \mathbb{E}_f(B) \right| \leq \epsilon \]
  
  $R$ is $\epsilon$-pseudorandom if $R$ is $\epsilon$-indistinguishable from $U_X$ (uniform on $X$).
A finite universe $X$ (e.g. $\{1, \ldots, N\}, \{0, 1\}^n$).

A family of distinguishers $\mathcal{F} = \{f : X \rightarrow \{0, 1\}\}$ (e.g. Circuits of size $s$).

Distributions $A$ and $B$ are $\epsilon$-indistinguishable by $\mathcal{F}$ if

$$\forall f \in \mathcal{F} \, |\mathbb{E}f(A) - \mathbb{E}f(B)| \leq \epsilon$$

$R$ is $\epsilon$-pseudorandom if $R$ is $\epsilon$-indistinguishable from $U_X$ (uniform on $X$).

$A$ is $\delta$-dense in $B$ if

$$\mathbb{P}(A = x) \leq \frac{1}{\delta} \mathbb{P}(B = x)$$

(e.g. $B = U_X$, $A$ uniform on $\delta|X|$ elements $\Rightarrow \mathbb{P}(A = x) = \frac{1}{\delta|X|}$).
What should a “Dense Model Theorem” be?

\[ D \text{ is } \delta\text{-dense in } R, \ R \text{ is } \epsilon\text{-pseudorandom w.r.t } \mathcal{F}. \]

\[ \Downarrow \]

There is \( M \) \( \delta\text{-dense in } U_X \), \( \epsilon\text{-indistinguishable from } D \) by \( \mathcal{F} \).
What should a “Dense Model Theorem” be?

\[ D \text{ is } \delta\text{-dense in } R, \text{ } R \text{ is } \epsilon\text{-pseudorandom w.r.t } \mathcal{F}. \]
\[ \Downarrow \]
\[ \text{There is } M \text{ } \delta\text{-dense in } U_X, \text{ } \epsilon\text{-indistinguishable from } D \text{ by } \mathcal{F}. \]

equivalently,

\[ \text{Every } M \text{ } \delta\text{-dense in } U_X \text{ is } \epsilon\text{-distinguishable from } D \text{ by } \mathcal{F} \]
\[ \Downarrow \]
\[ R \text{ is } \epsilon\text{-distinguishable from } U_X \text{ by } \mathcal{F}. \]
What should a “Dense Model Theorem” be?

- $D$ is $\delta$-dense in $R$, $R$ is $\epsilon'$-pseudorandom w.r.t $\mathcal{F}'$.
- There is $M$ $\delta$-dense in $U_X$, $\epsilon$-indistinguishable from $D$ by $\mathcal{F}$.

Equivalently,

- Every $M$ $\delta$-dense in $U_X$ is $\epsilon$-distinguishable from $D$ by $\mathcal{F}$

  $\Downarrow$

- $R$ is $\epsilon'$-distinguishable from $U_X$ by $\mathcal{F}'$.

Relation between $(\epsilon, \epsilon')$ and $(\mathcal{F}, \mathcal{F}')$ depends on the reduction.
The Results

Theorem (Tao-Ziegler 2006)

Suppose for all \( M \) \( \delta \)-dense in \( U_X \), some function in \( \mathcal{F} \) \( \epsilon \)-distinguishes \( M \) and \( D \). Then, there is a function \( h : X \rightarrow \{0, 1\}^n \) of the form

\[
h(x) = g(f_1(x), \ldots, f_k(x)) \quad f_i \in \mathcal{F}, \quad k = \text{poly}(1/\epsilon, 1/\delta)
\]

s.t.

\[
|\mathbb{E} h(R) - \mathbb{E} h(U_X)| \geq \text{poly}(\epsilon, \delta)
\]
The Results

**Theorem (Tao-Ziegler 2006)**

Suppose for all $M_\delta$-dense in $U_X$, some function in $\mathcal{F}_\epsilon$-distinguishes $M$ and $D$. Then, there is a function $h : X \to \{0, 1\}^n$ of the form

$$h(x) = g(f_1(x), \ldots, f_k(x)) \quad f_i \in \mathcal{F}, \quad k = \text{poly}(1/\epsilon, 1/\delta) \quad \exp(k) \text{ complexity}$$

s.t.

$$|\mathbb{E}h(R) - \mathbb{E}h(U_X)| \geq \text{poly}(\epsilon, \delta)$$
The Results

**Theorem (Tao-Ziegler 2006)**

Suppose for all $M$ $\delta$-dense in $U_X$, some function in $\mathcal{F}$ $\epsilon$-distinguishes $M$ and $D$. Then, there is a function $h : X \rightarrow \{0, 1\}^n$ of the form

$$h(x) = g(f_1(x), \ldots, f_k(x)) \quad f_i \in \mathcal{F}, \quad k = \text{poly}(1/\epsilon, 1/\delta)$$

**exp(k) complexity**

s.t.

$$|\mathbb{E}h(R) - \mathbb{E}h(U_X)| \geq \text{poly}(\epsilon, \delta)$$

**Theorem (RTTV 2007)**

Suppose for all $M$ $\delta$-dense in $U_X$, some function in $\mathcal{F}$ $\epsilon$-distinguishes $M$ and $D$. Then, there is a function $h : X \rightarrow \{0, 1\}^n$ of the form

$$h(x) = g(f_1(x), \ldots, f_k(x)) \quad f_i \in \mathcal{F}, \quad k = \text{poly}(1/\epsilon, \log 1/\delta)$$

**O(k) complexity**

s.t.

$$|\mathbb{E}h(R) - \mathbb{E}h(U_X)| \geq \Omega(\epsilon \delta)$$
The Proof

- Switching the quantifiers

\[ \forall M \ \exists f \quad E_f(D) - E_f(M) \geq \epsilon \]
The Proof

Switching the quantifiers

$$\forall M \ \exists f \ \mathbb{E}f(D) - \mathbb{E}f(M) \geq \epsilon$$

$$\implies \exists \bar{f} \ \forall M \ \mathbb{E}\bar{f}(D) - \mathbb{E}\bar{f}(M) \geq \epsilon$$

where $\bar{f} : X \to [0, 1]$ is a convex combination of functions from $\mathcal{F}$. 
Switching the quantifiers

\[ \forall M \ \exists f \ \mathbb{E}f(D) - \mathbb{E}f(M) \geq \epsilon \]

\[ \implies \exists \bar{f} \ \forall M \ \mathbb{E}\bar{f}(D) - \mathbb{E}\bar{f}(M) \geq \epsilon \]

where \( \bar{f} : X \to [0, 1] \) is a convex combination of functions from \( \mathcal{F} \).

Proof: min-max.
The Proof

- **Switching the quantifiers**

\[
\forall M \exists f \quad \mathbb{E}f(D) - \mathbb{E}f(M) \geq \epsilon
\]

\[\implies \exists \bar{f} \quad \forall M \quad \mathbb{E}\bar{f}(D) - \mathbb{E}\bar{f}(M) \geq \epsilon\]

where \(\bar{f} : X \to [0, 1]\) is a convex combination of functions from \(\mathcal{F}\).

Proof: min-max.

- **Getting a threshold distinguisher**

\[\mathbb{E}\bar{f}(D) - \mathbb{E}\bar{f}(M) \geq \epsilon\]

\[\implies \exists t \in (0, 1) \quad \mathbb{P}(\bar{f}(D) \geq t) - \mathbb{P}(\bar{f}(M) \geq t) \geq \epsilon\]
The Proof

- **Switching the quantifiers**

  \[ \forall M \; \exists f \; \mathbb{E}f(D) - \mathbb{E}f(M) \geq \epsilon \]

  \[ \implies \exists \bar{f} \; \forall M \; \mathbb{E}\bar{f}(D) - \mathbb{E}\bar{f}(M) \geq \epsilon \]

  where \( \bar{f} : X \rightarrow [0, 1] \) is a convex combination of functions from \( \mathcal{F} \).

  **Proof:** min-max.

- **Getting a threshold distinguisher**

  \[ \mathbb{E}\bar{f}(D) - \mathbb{E}\bar{f}(M) \geq \epsilon \]

  \[ \implies \exists t \in (0, 1) \quad \mathbb{P}(\bar{f}(D) \geq t) - \mathbb{P}(\bar{f}(M) \geq t) \geq \epsilon \]

  **Proof:** \( \mathbb{E}Z \) is the average of \( \mathbb{P}(Z \geq t) \) over \( t \in (0, 1) \).
The Proof

- Switching the quantifiers

\[
\forall M \ \exists f \ \mathbb{E}f(D) - \mathbb{E}f(M) \geq \epsilon
\]

\[
\implies \ \exists \tilde{f} \ \forall M \ \mathbb{E}\tilde{f}(D) - \mathbb{E}\tilde{f}(M) \geq \epsilon
\]

where \( \tilde{f} : X \rightarrow [0, 1] \) is a convex combination of functions from \( \mathcal{F} \).

Proof: min-max.

- Getting a threshold distinguisher

\[
\mathbb{E}\tilde{f}(D) - \mathbb{E}\tilde{f}(M) \geq \epsilon
\]

\[
\implies \ \exists \ t \in (0, 1) \ \mathbb{P}(\tilde{f}(D) \geq t) - \mathbb{P}(\tilde{f}(M) \geq t) \geq \epsilon
\]

Proof: \( \mathbb{E}Z \) is the average of \( \mathbb{P}(Z \geq t) \) over \( t \in (0, 1) \).

In fact,

\[
\exists t \ \mathbb{P}(\tilde{f}(D) \geq t + \epsilon/2) - \mathbb{P}(\tilde{f}(M) \geq t) \geq \epsilon/2
\]
Using the distinguisher for $R$
Using the distinguisher for $R$

Let $S$ be the set of $\delta|X|$ elements where $\bar{f}$ is maximized.

$$\mathbb{P}(\bar{f}(D) \geq t + \epsilon/2) - \mathbb{P}(\bar{f}(U_S) \geq t) \geq \epsilon/2$$

$$\implies \mathbb{P}(\bar{f}(R) \geq t + \epsilon/2) - \mathbb{P}(\bar{f}(U_X) \geq t) \geq \epsilon\delta/2$$
The Proof (contd...)

- **Using the distinguisher for** $R$

Let $S$ be the set of $\delta |X|$ elements where $\bar{f}$ is maximized.

\[
\mathbb{P}(\bar{f}(D) \geq t + \epsilon/2) - \mathbb{P}(\bar{f}(U_S) \geq t) \geq \epsilon/2
\]

\[
\implies \mathbb{P}(\bar{f}(R) \geq t + \epsilon/2) - \mathbb{P}(\bar{f}(U_X) \geq t) \geq \epsilon\delta/2
\]
Using the distinguisher for $R$

Let $S$ be the set of $\delta|X|$ elements where $\bar{f}$ is maximized.

$$\mathbb{P}(\bar{f}(D) \geq t + \epsilon/2) - \mathbb{P}(\bar{f}(U_S) \geq t) \geq \epsilon/2$$

$$\Rightarrow \mathbb{P}(\bar{f}(R) \geq t + \epsilon/2) - \mathbb{P}(\bar{f}(U_X) \geq t) \geq \epsilon\delta/2$$
Using the distinguisher for $R$

Let $S$ be the set of $\delta|X|$ elements where $\bar{f}$ is maximized.

\[
\mathbb{P}(\bar{f}(D) \geq t + \epsilon/2) - \mathbb{P}(\bar{f}(U_S) \geq t) \geq \epsilon/2
\]

\[
\Rightarrow \mathbb{P}(\bar{f}(R) \geq t + \epsilon/2) - \mathbb{P}(\bar{f}(U_X) \geq t) \geq \epsilon\delta/2
\]
Using the distinguisher for $R$

Let $S$ be the set of $\delta|X|$ elements where $\bar{f}$ is maximized.

\[ \mathbb{P}(\bar{f}(D) \geq t + \epsilon/2) - \mathbb{P}(\bar{f}(U_S) \geq t) \geq \epsilon/2 \]

\[ \Rightarrow \mathbb{P}(\bar{f}(R) \geq t + \epsilon/2) - \mathbb{P}(\bar{f}(U_X) \geq t) \geq \epsilon\delta/2 \]
Using the distinguisher for $R$

Let $S$ be the set of $\delta|X|$ elements where $\bar{f}$ is maximized.

\[
\mathbb{P}(\bar{f}(D) \geq t + \epsilon/2) - \mathbb{P}(\bar{f}(U_S) \geq t) \geq \epsilon/2
\]

\[
\implies \mathbb{P}(\bar{f}(R) \geq t + \epsilon/2) - \mathbb{P}(\bar{f}(U_X) \geq t) \geq \epsilon \delta/2
\]
The Proof (almost done now...)

- Getting few functions (Chernoff bound)

  \( \bar{f} \) is a distribution over functions such that

  \[
  \mathbb{P}(\bar{f}(R) \geq t + \epsilon/2) - \mathbb{P}(\bar{f}(U_X) \geq t) \geq \epsilon\delta/2
  \]

  Sample \( k = \text{poly}(1/\epsilon, \log 1/\delta) \) functions \( f_1, \ldots, f_k \)

  \[
  \mathbb{P} \left( \frac{\sum f_i(R)}{k} \geq t + \epsilon/4 \right) - \mathbb{P} \left( \frac{\sum f_i(U_X)}{k} \geq t + \epsilon/4 \right) \geq \epsilon\delta/4
  \]
The Proof (almost done now...)

- **Getting few functions (Chernoff bound)**
  
  $\bar{f}$ is a distribution over functions such that
  
  $$\mathbb{P}(\bar{f}(R) \geq t + \epsilon/2) - \mathbb{P}(\bar{f}(U_X) \geq t) \geq \epsilon\delta/2$$

  Sample $k = \text{poly}(1/\epsilon, \log 1/\delta)$ functions $f_1, \ldots, f_k$

  $$\mathbb{P} \left( \frac{\sum f_i(R)}{k} \geq t + \epsilon/4 \right) - \mathbb{P} \left( \frac{\sum f_i(U_X)}{k} \geq t + \epsilon/4 \right) \geq \epsilon\delta/4$$

- Note that we combine $f_1, \ldots, f_k$ only as a linear threshold function. **Complexity = $O(k)$**.
The Green-Tao proof (Iterative Partitioning)

- Partition $X$ into pieces.

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The Green-Tao proof (Iterative Partitioning)

1. Partition $X$ into pieces.

2. To get $M$, pick whole pieces according to density of $D$ in the piece.
Partition $X$ into pieces.

To get $M$, pick whole pieces according to density of $D$ in the piece.

If $D$ is distinguishable from $M$, then can refine partition.

Use pseudorandomness of $R$ to bound number of steps.
Smuggling techniques in the other direction

- We adapt the Green-Tao proof technique to prove Impagliazzo’s hardcore lemma:

  If function $f : X \rightarrow \{0, 1\}$ is hard to compute correctly on more than $1 - \delta$ fraction of inputs from $X$ then there is a set $H \subseteq X$, $|H| \geq \delta|X|$ such that $f$ is “very hard” to compute on $H$. 

Iterative partitioning gives a circuit for computing $H$. 

We adapt the Green-Tao proof technique to prove Impagliazzo's hardcore lemma:

If function $f : X \rightarrow \{0, 1\}$ is hard to compute correctly on more than $1 - \delta$ fraction of inputs from $X$ then there is a set $H \subseteq X$, $|H| \geq \delta|X|$ such that $f$ is “very hard” to compute on $H$.

Iterative partitioning gives a circuit for computing $H$. 
Further questions

- All this is good in theory... but how can it be applied?
Further questions

- All this is good in theory... but how can it be applied?
  - Pseudoentropy $\Leftrightarrow$ density in a pseudorandom distribution.
Further questions

- All this is good in theory... but how can it be applied?
  - Pseudoentropy $\iff$ density in a pseudorandom distribution.
  - New proof of regularity lemma for subgraphs of expanders.
Further questions

- All this is good in theory... but how can it be applied?

  - Pseudoentropy ⇔ density in a pseudorandom distribution.

  - New proof of regularity lemma for subgraphs of expanders.
    - Uniform distribution on edges of the complete graph.
    - Expanders are pseudorandom w.r.t. cuts.

- And?

- Other applications of “ergodic arguments” in complexity theory?
Further questions

- All this is good in theory... but how can it be applied?
  - Pseudoentropy $\Leftrightarrow$ density in a pseudorandom distribution.
  - New proof of regularity lemma for subgraphs of expanders.
    - Uniform distribution on edges of the complete graph.
    - Expanders are pseudorandom w.r.t. cuts.
  - And?
Further questions

All this is good in theory... but how can it be applied?

- Pseudoentropy $\Leftrightarrow$ density in a pseudorandom distribution.
- New proof of regularity lemma for subgraphs of expanders.
  - Uniform distribution on edges of the complete graph.
  - Expanders are pseudorandom w.r.t. cuts.

And?

Other applications of “ergodic arguments” in complexity theory?