Dense Subsets of Pseudorandom Sets

O. Reingold[1] L. Trevisan[2] M. Tulsiani[2] S. Vadhan[3]

[1]Weizmann Institute

[2]UC Berkeley

[3]Harvard University

Progressions in Subsets of Integers

Theorem (Szemeredi 1975)

Any set of A of δN integers 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 δN 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 δN 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.

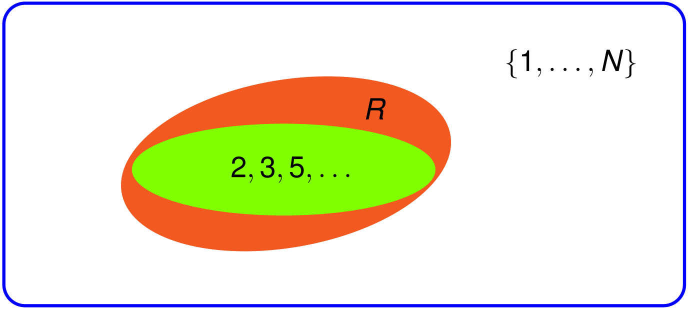
Green-Tao showed that a property of dense subsets of the integers (having progressions) also holds for the primes.

The Green-Tao Proof

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

The Green-Tao Proof

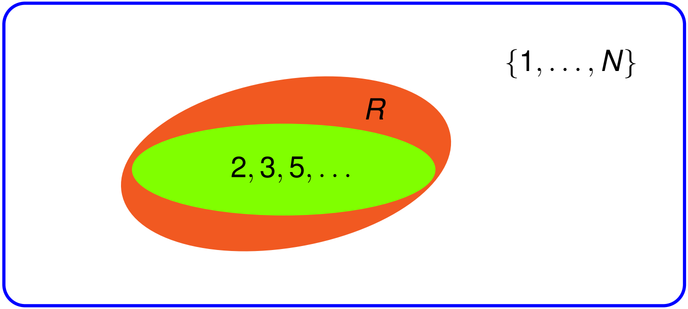
Thm 1 There is a pseudorandom set $R \subseteq \{1, \ldots, N\}$ such that primes have constant density in R.



The Green-Tao Proof

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

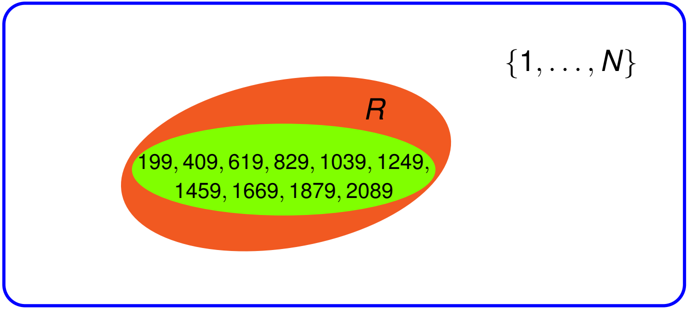
Thm 2 If R is a pseudorandom subset of $\{1, \ldots, N\}$ and if D is a dense subset i.e. $|D| \geq \delta R$, then D contains a length k-AP.



The Green-Tao Proof

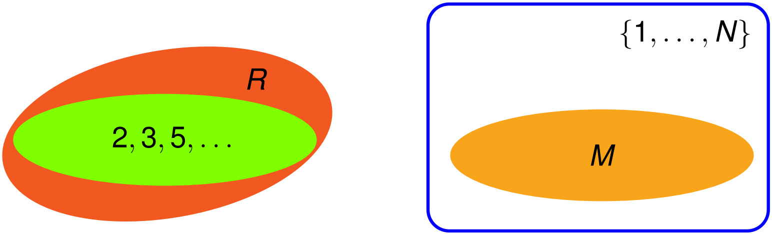
Thm 1 There is a pseudorandom set $R \subseteq \{1, \ldots, N\}$ such that primes have constant density in R.

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

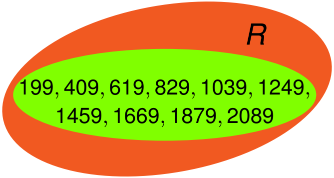
- If D is a dense in a pseudorandom set R ($|D| \geq \delta|R|$), then there is a dense model set M ($|M| \geq \delta N$) indistinguishable from D.

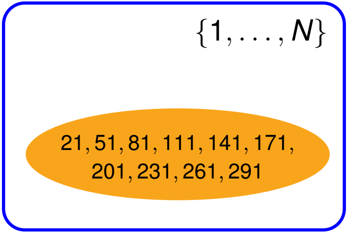


Proof of Theorem 2

- 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.

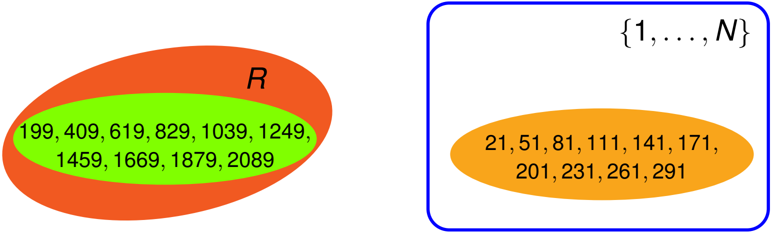
- M must contain length k-APs (Szemeredi). So does D.





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.



"A dense subset of a pseudorandom set has a dense model."
Can we prove this in general?

Abstracting out...

- A finite universe X (e.g. $\{1, \ldots, N\}, \{0,1\}^n$).
- A family of distinguishers $\mathcal{F} = \{f : X \to \{0,1\}\}$ (e.g. Circuits of size s).

Abstracting out...

- A finite universe X (e.g. $\{1, \ldots, N\}, \{0,1\}^n$).
- A family of distinguishers $\mathcal{F} = \{f : X \to \{0,1\}\}$ (e.g. Circuits of size s).
- Distributions A and B are ϵ-indistinguishable by $\mathcal{F}$ if

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

R is ϵ-pseudorandom if R is ϵ-indistinguishable from U_X (uniform on X).

Abstracting out...

- A finite universe X (e.g. $\{1, \dots, N\}, \{0,1\}^n$).
- A family of distinguishers $\mathcal{F} = \{f : X \to \{0,1\}\}$ (e.g. Circuits of size s).
- Distributions A and B are ϵ-indistinguishable by $\mathcal{F}$ if

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

  R is ϵ-pseudorandom if R is ϵ-indistinguishable from U_X (uniform on X).
- A is δ-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 $\implies \mathbb{P}(A = x) = \frac{1}{\delta|X|}$).

What should a "Dense Model Theorem" be?

D is δ-dense in R, R is ϵ-pseudorandom w.r.t $\mathcal{F}$.

$$\Downarrow$$

There is M δ-dense in U_X, ϵ-indistinguishable from D by $\mathcal{F}$.

What should a "Dense Model Theorem" be?

D is δ-dense in R, R is ϵ-pseudorandom w.r.t $\mathcal{F}$.

$$\Downarrow$$

There is M δ-dense in U_X, ϵ-indistinguishable from D by $\mathcal{F}$.

equivalently,

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

$$\Downarrow$$

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

What should a "Dense Model Theorem" be?

D is δ-dense in R, R is ϵ'-pseudorandom w.r.t $\mathcal{F}'$.

$$\Downarrow$$

There is M δ-dense in U_X, ϵ-indistinguishable from D by $\mathcal{F}$.

equivalently,

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

$$\Downarrow$$

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

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

The Results

Theorem (Tao-Ziegler 2006)

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

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

s.t.

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

The Results

Theorem (Tao-Ziegler 2006)

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

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

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

The Results

Theorem (Tao-Ziegler 2006)

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

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

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

Theorem (RTTV 2007)

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

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

s.t. $$|\mathbb{E}h(R) - \mathbb{E}h(U_X)| \geq \Omega(\epsilon\delta)$$

The Proof

- Switching the quantifiers

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

The Proof

- Switching the quantifiers

$$\begin{aligned} & \forall M \;\; \exists f \quad \mathbb{E}f(D) - \mathbb{E}f(M) \geq \epsilon \\ \Longrightarrow \quad & \exists \bar{f} \;\; \forall M \quad \mathbb{E}\bar{f}(D) - \mathbb{E}\bar{f}(M) \geq \epsilon \end{aligned}$$

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

The Proof

- Switching the quantifiers

$$\begin{aligned} & \forall M \;\; \exists f \quad \mathbb{E}f(D) - \mathbb{E}f(M) \geq \epsilon \\ \Longrightarrow \quad & \exists \bar{f} \;\; \forall M \quad \mathbb{E}\bar{f}(D) - \mathbb{E}\bar{f}(M) \geq \epsilon \end{aligned}$$

where $\bar{f} : X \to [0, 1]$ is a convex combination of functions from $\mathcal{F}$.
Proof: min-max.

The Proof

- Switching the quantifiers

$$
\begin{aligned}
& \forall M \;\; \exists f \quad \mathbb{E} f(D) - \mathbb{E} f(M) \geq \epsilon \\
\implies \quad & \exists \bar{f} \;\; \forall M \quad \mathbb{E} \bar{f}(D) - \mathbb{E} \bar{f}(M) \geq \epsilon
\end{aligned}
$$

  where $\bar{f} : X \to [0,1]$ is a convex combination of functions from $\mathcal{F}$.
  Proof: min-max.

- Getting a threshold distinguisher

$$
\begin{aligned}
& \mathbb{E} \bar{f}(D) - \mathbb{E} \bar{f}(M) \geq \; \epsilon \\
\implies \quad \exists \, t \in (0,1) \quad & \mathbb{P}(\bar{f}(D) \geq t) - \mathbb{P}(\bar{f}(M) \geq t) \geq \; \epsilon
\end{aligned}
$$

The Proof

- Switching the quantifiers

$$\begin{aligned} & \forall M \;\; \exists f \quad \mathbb{E}f(D) - \mathbb{E}f(M) \geq \epsilon \\ \Longrightarrow \quad & \exists \bar{f} \;\; \forall M \quad \mathbb{E}\bar{f}(D) - \mathbb{E}\bar{f}(M) \geq \epsilon \end{aligned}$$

where $\bar{f} : X \to [0,1]$ is a convex combination of functions from $\mathcal{F}$.
Proof: min-max.

- Getting a threshold distinguisher

$$\begin{aligned} & \mathbb{E}\bar{f}(D) - \mathbb{E}\bar{f}(M) \geq \; \epsilon \\ \Longrightarrow \quad & \exists \; t \in (0,1) \quad \mathbb{P}(\bar{f}(D) \geq t) - \mathbb{P}(\bar{f}(M) \geq t) \geq \; \epsilon \end{aligned}$$

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

The Proof

- Switching the quantifiers

$$\begin{aligned} & \forall M \;\; \exists f \quad \mathbb{E}f(D) - \mathbb{E}f(M) \geq \epsilon \\ \Longrightarrow \quad & \exists \bar{f} \;\; \forall M \quad \mathbb{E}\bar{f}(D) - \mathbb{E}\bar{f}(M) \geq \epsilon \end{aligned}$$

where $\bar{f} : X \to [0,1]$ is a convex combination of functions from $\mathcal{F}$.
Proof: min-max.

- Getting a threshold distinguisher

$$\begin{aligned} & \mathbb{E}\bar{f}(D) - \mathbb{E}\bar{f}(M) \geq \epsilon \\ \Longrightarrow \quad \exists\, t \in (0,1) \quad & \mathbb{P}(\bar{f}(D) \geq t) - \mathbb{P}(\bar{f}(M) \geq t) \geq \epsilon \end{aligned}$$

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

In fact, $\exists t \quad \mathbb{P}(\bar{f}(D) \geq t + \epsilon/2) - \mathbb{P}(\bar{f}(M) \geq t) \geq \epsilon/2$

The Proof (contd...)

- Using the distinguisher for R

The Proof (contd...)

- Using the distinguisher for R

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

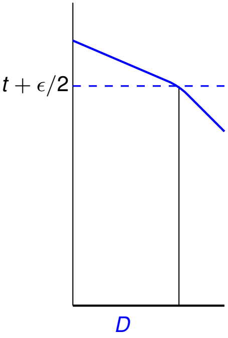
$$
\begin{aligned}
& \mathbb{P}(\bar{f}(D) \geq t + \epsilon/2) - \mathbb{P}(\bar{f}(U_S) \geq t) \geq \epsilon/2 \\
\implies \quad & \mathbb{P}(\bar{f}(R) \geq t + \epsilon/2) - \mathbb{P}(\bar{f}(U_X) \geq t) \geq \epsilon\delta/2
\end{aligned}
$$

The Proof (contd...)

- Using the distinguisher for R

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

$$
\begin{aligned}
& \mathbb{P}(\bar{f}(D) \geq t + \epsilon/2) - \mathbb{P}(\bar{f}(U_S) \geq t) \geq \ \epsilon/2 \\
\implies \quad & \mathbb{P}(\bar{f}(R) \geq t + \epsilon/2) - \mathbb{P}(\bar{f}(U_X) \geq t) \geq \ \epsilon\delta/2
\end{aligned}
$$

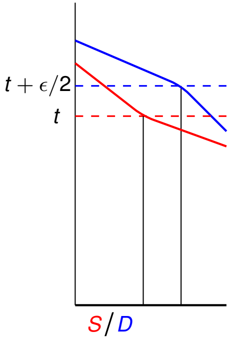


The Proof (contd...)

- Using the distinguisher for R

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

$$\begin{aligned} & \mathbb{P}(\bar{f}(D) \geq t + \epsilon/2) - \mathbb{P}(\bar{f}(U_S) \geq t) \geq \ \epsilon/2 \\ \implies \quad & \mathbb{P}(\bar{f}(R) \geq t + \epsilon/2) - \mathbb{P}(\bar{f}(U_X) \geq t) \geq \ \epsilon\delta/2 \end{aligned}$$

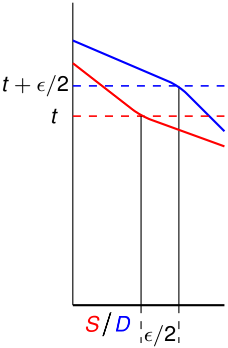


The Proof (contd...)

- Using the distinguisher for R

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

$$
\begin{aligned}
& \mathbb{P}(\bar{f}(D) \geq t + \epsilon/2) - \mathbb{P}(\bar{f}(U_S) \geq t) \geq \epsilon/2 \\
\Longrightarrow\ & \mathbb{P}(\bar{f}(R) \geq t + \epsilon/2) - \mathbb{P}(\bar{f}(U_X) \geq t) \geq \epsilon\delta/2
\end{aligned}
$$

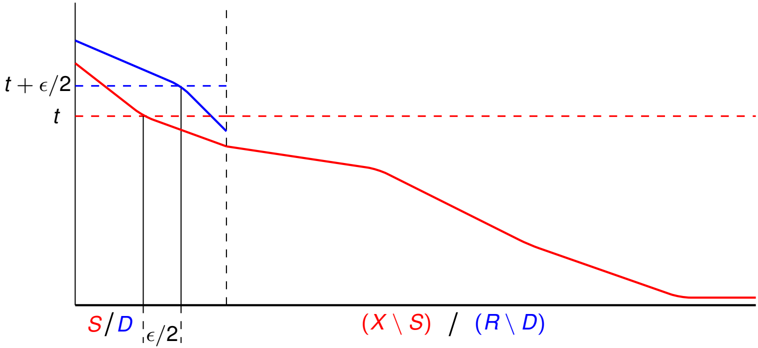


The Proof (contd...)

- Using the distinguisher for R

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

$$\begin{aligned} & \mathbb{P}(\bar{f}(D) \geq t + \epsilon/2) - \mathbb{P}(\bar{f}(U_S) \geq t) \geq \ \epsilon/2 \\ \implies \quad & \mathbb{P}(\bar{f}(R) \geq t + \epsilon/2) - \mathbb{P}(\bar{f}(U_X) \geq t) \geq \ \epsilon\delta/2 \end{aligned}$$



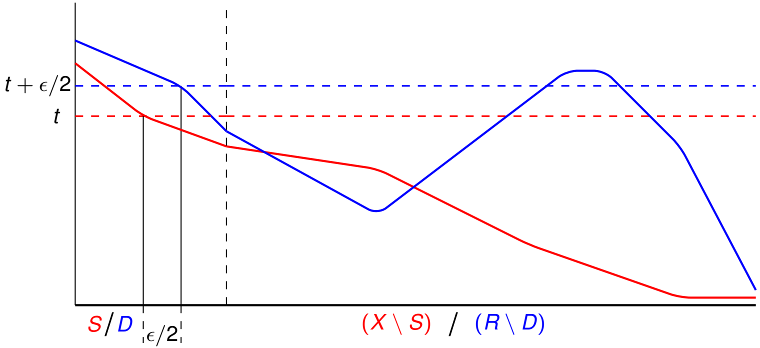
The Proof (contd...)

- Using the distinguisher for R

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

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

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



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 $k = poly(1/\epsilon, \log 1/\delta)$ 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

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

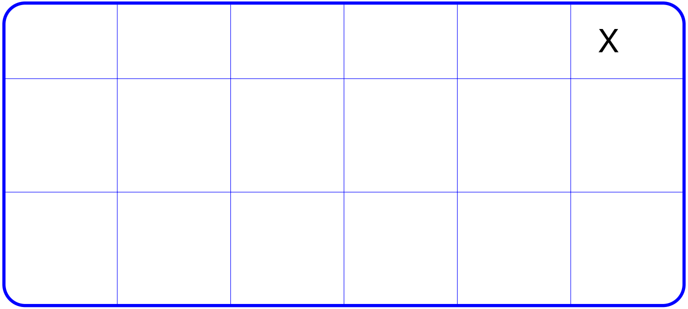
  Sample $k = poly(1/\epsilon, \log 1/\delta)$ functions $f_1, \ldots, f_k$

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

- Note that we combine $f_1, \ldots, f_k$ only as a linear threshold function. Complexity = $O(k)$.

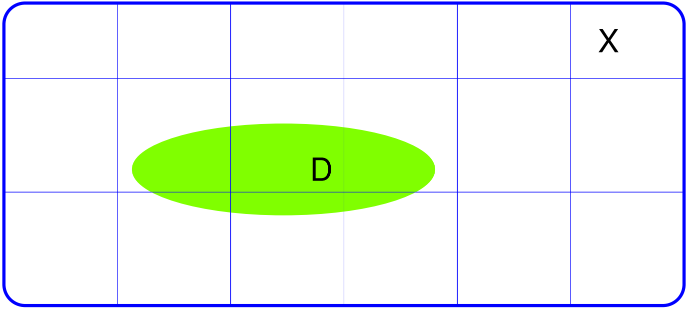
The Green-Tao proof (Iterative Partitioning)

- Partition X into pieces.



The Green-Tao proof (Iterative Partitioning)

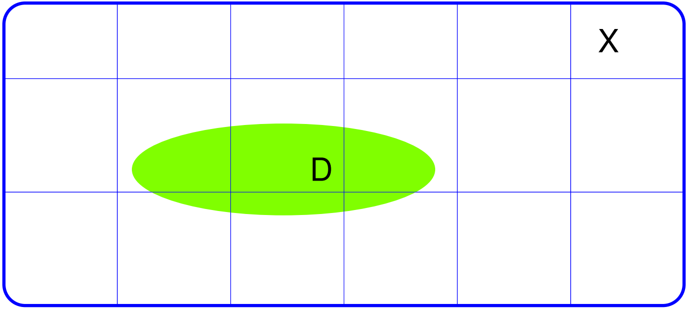
- Partition X into pieces.



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

The Green-Tao proof (Iterative Partitioning)

- 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 $f : X \to \{0, 1\}$ 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.

Smuggling techniques in the other direction

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

> If function $f : X \to \{0, 1\}$ 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 $\Leftrightarrow$ 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 $\Leftrightarrow$ 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.

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?