Combinatorial characterizations of extractors and Kolmogorov extractors

Marius Zimand

Towson University

CCA/CCR 2011, Cape Town
Randomness extraction: algorithmical process that improves the quality of sources of randomness

Sources of randomness can be modeled in two ways:

1. A distribution on \( \{0, 1\}^n \), quality given by min-entropy
2. A string in \( \{0, 1\}^n \), quality is given by its Kolmogorov complexity

**THEOREM:** A procedure \( E \) extracts randomness in model(1) IFF it extracts randomness in model(2) (with some small loss in parameters).
This talk in two slides

- Fortnow, Hitchcock, Pavan, Vinodchandran, Wang, ’06: Model(1) $\implies$ Model(2).
- Hitchcock, Pavan, Vinodchandran, ’09: Model(2) $\implies$ Model(1)

We show that extractors in Model(1) and Model(2) admit similar combinatorial characterizations.

Equivalence Theorem follows.

Part of the survey paper: Z, Possibilities and impossibilities in Kolmogorov extraction, SIGACT News, Dec. 2010
Randomness Extractor = Algorithmical procedure that “repairs” one source or several sources of randomness

\[ x \text{ (low quality randomness)} \mapsto E(x) \text{ (high quality randomness)} \]

Extraction from a single source is impossible (unless some structural information about the source is known).

Therefore we consider 2 sources:

\[ x_1, x_2 \text{ (low quality randomness)} \mapsto E(x_1, x_2) \text{ (high quality randomness)} \]
Two models for **source** and **randomness quality**

1. Source $X$ is distribution over $\{0, 1\}^n$.
   - Quality: min-entropy $H_\infty(X)$.
     
     $H_\infty(X) = k$ means $\forall a \in \{0, 1\}^n, \Pr[X = a] \leq 2^{-k}$.

2. Source $x$ is a string in $\{0, 1\}^n$.
   - Quality: Kolmogorov complexity $C(x)$.
     
     $n - C(x) \geq C(x)$.
Two models for source and randomness quality

(1) Source $X$ is distribution over $\{0, 1\}^n$.
   Quality: min-entropy $H_\infty(X)$.
   $H_\infty(X) = k$ means $\forall a \in \{0, 1\}^n$, $\text{Prob}[X = a] \leq 2^{-k}$.
Two models for source and randomness quality

(2) Source $x$ is a string in $\{0, 1\}^n$.

Quality: Kolmogorov complexity $C(x)$.

$n - C(x)$

$C(x)$
Two models for source and randomness quality

(1) Source $X$ is distribution over $\{0, 1\}^n$.
   Quality: min-entropy $H_\infty(X)$.
   $H_\infty(X) = k$ means $\forall a \in \{0, 1\}^n$, $\text{Prob}[X = a] \leq 2^{-k}$.

(2) Source $x$ is a string in $\{0, 1\}^n$.
   Quality: Kolmogorov complexity $C(x)$.
Extractor (in Model (1))

Definition:

\((n, k)\) source: a r.v. \(X\) over \(\{0, 1\}^n\) with \(H_\infty(X) \geq k\).

\[\begin{array}{c}
\text{n-k} \\
\text{k}
\end{array}\]

Definition:

\((k, \varepsilon)\) Extractor = an ensemble of functions \(\{E_n\}_{n \in \mathbb{N}}\), each \(E_n : \{0, 1\}^n \times \{0, 1\}^n \rightarrow \{0, 1\}^m\), such that for every \(X_1, X_2\), independent \((n, k)\) sources, it holds that \(E_n(X_1, X_2)\) is \(\varepsilon\) close to \(U_m\) (the uniform distribution on \(\{0, 1\}^m\)).

\[\begin{array}{c}
\text{n-k} \\
\text{k} \\
\text{m}
\end{array}\]
## Large literature on extractors:

### Computable extractors

- [Chor, Goldreich,'88] $m = \frac{1}{3}(k - \log n - O(1))$.
- [Chor, Goldreich,'88] $m = 2k - n$.
- [Dodis, Oliveira,'03] any $k \geq \log(n) + 2\log(1/\epsilon)$, $m = 2k - 2\log(1/\epsilon)$.
- [Dodis, Oliveira,'03] any $k \geq \log(n) + 2\log(1/\epsilon)$, $m = k - 2\log(1/\epsilon)$, and $(X_1, f(X_1, X_2)) \approx_\epsilon (X_1, U_m)$.

### Poly-time computable extractors

- [Chor, Goldreich,'88] $k \geq (7/8)n$, $m = k - n/2 - 2\log(1/\epsilon)$.
- [Raz,05] $(n, k_1)$-source and one $(n, k_2)$-source with $k_1 > 0.5n$, $k_2 = \text{polylog}(n)$,
- [Bourgain,'05] $k = 0.4999n$, $m = \Omega(n)$,
- [Kalai, Li, Rao,'09] $k = \delta n$ (any $\delta > 0$), $m = n^{\Omega(1)}$ (under a hardness assumption: existence of one-way permutations with certain parameters).
Kolmogorov extractors (Model (2))

- $x \in \{0,1\}^n$, $y \in \{0,1\}^n$.
- Definition: $\text{dep}(x, y) = \max\{C(x \mid n) - C(x \mid y), C(y \mid n) - C(y \mid x)\}$.

Sources with complexity $k$ and dependency $\alpha$

We look at pairs of strings in $S_{k,\alpha} = \{(x, y) \in \{0,1\}^n \times \{0,1\}^n \mid C(x \mid n) \geq k, C(y \mid n) \geq k, \text{dep}(x, y) \leq \alpha\}$.

Definition:

$(k, \alpha, d)$ Kolmogorov extractor = an ensemble of functions $\{E_n\}_{n \in \mathbb{N}}$, each $E_n : \{0,1\}^n \times \{0,1\}^n \rightarrow \{0,1\}^m$, such that for every sources $(x, y) \in S_{k,\alpha}$, it holds that $C(E_n(x, y) \mid n) \geq m - d$.

Why $d$? Because from sources with dependency $\alpha$, it’s impossible to extract with randomness deficiency $< \alpha$ (i.e., $d$ has to be $\geq \alpha$).
Almost Extractor (in Model (1))

Definition:

\((k, \epsilon, d)\) almost extractor = an ensemble of functions \(\{E_n\}_{n \in \mathbb{N}}\), each \(E_n : \{0, 1\}^n \times \{0, 1\}^n \rightarrow \{0, 1\}^m\), such that for every \(X_1, X_2\), independent \((n, k)\) sources, it holds that \(E_n(X_1, X_2)\) is \(\epsilon\) close to having min-entropy \(m - d\).
Equivalence Theorem

Let $E : \{0,1\}^n \times \{0,1\}^n \rightarrow \{0,1\}^m$ be an ensemble of functions.

1. If $E$ is a $(k, \epsilon, d)$ almost extractor, then $E$ is a $(k', \alpha, \alpha + 2d + 1)$ Kolmogorov extractor, where $k' = k + \log n + O(1)$, $\alpha = \log(1/\epsilon) + d + 1$.

2. If $E$ is a $(k, \alpha, d)$ Kolmogorov extractor, then $E$ is a $(k', \epsilon, d')$ almost extractor, where $k' = k + \alpha$, $\epsilon = 2 \cdot 2^{-\alpha}$, $d' = d + O(1)$. 
Recall that an extractor has the form $E : \{0, 1\}^n \times \{0, 1\}^n \rightarrow \{0, 1\}^m$.

We view it as an $N \times N$ table colored with $M$ colors ($N = 2^n, M = 2^m$).

A table $T : [N] \times [N] \rightarrow [M]$ is balanced if for all “large” rectangles $B_1 \times B_2$, all colors appear “approximately” the same number of times.

Depending on “large” and “approximately”, we obtain different types of balanced tables.
Why are balanced tables related to randomness extraction?

- If \( E : \{0, 1\}^n \times \{0, 1\}^n \to \{0, 1\}^m \) is Kolmogorov extractor, then its table is balanced.

  Sketch of the argument:

- Suppose the table \( E \) is not balanced.
- Take the color \( a \) with the largest number of occurrences.
- \( a \) has low Kolmogorov complexity (the above line is a short description)
- \( a \) has many occurrences because the table is not balanced.
- \( a \) has many preimages.
- There are so many preimages, that one of them, say \( (x, y) \), must be in \( S_{k,\alpha} \).
- So \( E(x, y) \) has low Kolm. complexity, but \( (x, y) \) is in \( S_{k,\alpha} \).
- Contradiction.
Why are balanced tables related to randomness extraction? (cont.)

- If $E : \{0, 1\}^n \times \{0, 1\}^n \rightarrow \{0, 1\}^m$ is a balanced table, then $E$ is a Kolmogorov extractor.
- Sketch of the argument.
- Fix $(x, y) \in S_{k, \alpha}$.
- Let $B_x = \{ u \mid C(u \mid n) \leq C(x \mid n) \}$, $B_y = \{ v \mid C(v \mid n) \leq C(y \mid n) \}$
- $B_x \times B_y$ is a rectangle of size $\approx 2^{C(x\mid n)} \times 2^{C(y\mid n)} \geq 2^k \times 2^k$.
- Each color from $[M]$ appears in $B_x \times B_y$ at most a fraction of $O(1)/M$ times (because the table is balanced).
- $(x, y)$ is in $B_x \times B_y$, and the color $z = E(x, y)$ appears at most $(O(1)/M)2^{C(x\mid n)} \times 2^{C(y\mid n)} = 2^{C(x\mid n)+C(y\mid n)-m+O(1)}$ times.
Why are balanced tables related to randomness extraction? (cont.)

- \((x, y)\) is in \(B_x \times B_y\), and the color \(z = E(x, y)\) appears at most 
  \((O(1)/M)2^{C(x|n)} \times 2^{C(y|n)} = 2^{C(x|n)+C(y|n)-m+O(1)}\) times.

- Given \(C(x \mid n)\) and \(C(y \mid n)\), the elements of \(B_x \times B_y\) can be effectively enumerated.

- So the string \(xy\) can be described by \(z, C(x \mid n), C(y \mid n)\) and by the rank of cell \((x, y)\) in an enumeration of \(z\)-colored cells in \(B_x \times B_y\).

<table>
<thead>
<tr>
<th></th>
<th>(u_1)</th>
<th>(u_2)</th>
<th>(B_y)</th>
<th>(y)</th>
<th>(u_N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(u_1)</td>
<td></td>
<td>.</td>
<td>.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(u_2)</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>(B_x)</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>
<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>(u_N)</td>
<td></td>
<td>.</td>
<td>.</td>
<td>.</td>
<td></td>
</tr>
</tbody>
</table>
So,
\[ C(xy \mid n) \leq C(z \mid n) + C(C(x \mid n)) + C(C(y \mid n)) + C(\text{rank} \mid n) + O(\log n). \]

But \( C(\text{rank} \mid n) \leq C(x \mid n) + C(y \mid n) - m + O(1) \) (from first line).

So, \( C(xy \mid n) \leq C(z \mid n) + C(x \mid n) + C(y \mid n) - m + O(\log n) \)

Also, \( C(xy \mid n) \geq C(x \mid n) + C(y \mid n) - \alpha \) (because dependency \((x, y) \leq \alpha\)).

So, \( C(z \mid n) \geq m - \alpha - O(\log n) \).

Thus, \( E \) is a Kolmogorov extractor, qed.
In the previous argument we choose the right parameters.

The argument can be made tighter in a few places.

**THEOREM (informal):** \( E : \{0, 1\}^n \times \{0, 1\}^n \to \{0, 1\}^m \) is \((k, \alpha, d)\) Kolm. extractor

IFF in any rectangle of size \( \approx 2^k \times 2^k \), any set of colors \( U \subseteq [M] \) of appropriate size appears a fraction \( \approx (|U|/M) \cdot 2^d \).
Kolmogorov extractors and balanced tables

Theorem (Combinatorial characterization of Kolmogorov extractors)

Let $f : \{0, 1\}^n \times \{0, 1\}^n \rightarrow \{0, 1\}^m$ be an ensemble of functions.

1. If $f$ is a $(k, \alpha, d)$-Kolmogorov extractor, then for any rectangle $B_1 \times B_2 \subseteq [N] \times [N]$ of size $2^{k'} \times 2^{k'}$, where $k' = k + \alpha$, for any set of colors $U \subseteq [M]$, with size $|U| = 2^{-\alpha} \cdot M \cdot 2^{-(d+O(1))}$, it holds that

$$\frac{|\{U \text{-cells in } B_1 \times B_2\}|}{|B_1 \times B_2|} \leq \frac{|U|}{M} \cdot 2^{d+O(1)}.$$

2. Suppose that there exists a constant $d$ such that for all rectangles $B_1 \times B_2$ of size $2^k \times 2^k$, for any $U \subseteq [M]$ and for some $\epsilon$ computable from $n$, it holds that

$$\frac{|\{U \text{-cells in } B_1 \times B_2\}|}{|B_1 \times B_2|} \leq \frac{|U|}{M} \cdot 2^d + \epsilon.$$

Then $f$ is a $(k', \alpha, \alpha + 2d + 1)$ Kolmogorov extractor, where $k' = k + \log n + O(\log \log n)$, and $\alpha = \log(1/\epsilon) + d + 1$. 

Marius Zimand (Towson U.)
Combinatorial characterizations of extractors
2011 17 / 1
Almost extractors and balanced tables

Theorem (Combinatorial characterization of almost extractors)

Let \( f : \{0, 1\}^n \times \{0, 1\}^n \to \{0, 1\}^m \) be an ensemble of functions.

1. If \( f \) is a \((k, \epsilon, d)\) almost extractor, then for every rectangle \( B_1 \times B_2 \subseteq [N] \times [N] \) of size \( 2^k \times 2^k \), and for any set of colors \( U \subseteq [M] \),

\[
\frac{|\{ U \text{-cells in } B_1 \times B_2 \}|}{|B_1 \times B_2|} \leq \frac{|U|}{M} \cdot 2^d + \epsilon.
\]

2. Suppose that for every rectangle \( B_1 \times B_2 \subseteq [N] \times [N] \) of size \( 2^k \times 2^k \), for any set of colors \( U \subseteq [M] \) with \( |U| = \epsilon \cdot M \cdot 2^{-d} \),

\[
\frac{|\{ U \text{-cells in } B_1 \times B_2 \}|}{|B_1 \times B_2|} \leq \frac{|U|}{M} \cdot 2^d + \epsilon.
\]

Then \( f \) is a \((k, 2\epsilon, d)\) almost extractor.
**Proof idea**

- **Almost extractor** ⇒ **balanced table**
- Suppose there is a $2^k \times 2^k$ rectangle that is not balanced.
- Consider two r.v.’s, $X$ and $Y$ who put equal distribution on each row, resp. each column of the rectangle.
- $X$ and $Y$ have min-entropy $k$, but the most popular color has low entropy. **Contradiction.**

<table>
<thead>
<tr>
<th></th>
<th>$u_1$</th>
<th>$u_2$</th>
<th>$B_2$</th>
<th>$u_N$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u_1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$u_2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\cdot$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$B_1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\cdot$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\cdot$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\cdot$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\cdot$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\cdot$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$u_N$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Proof idea**

- **balanced table** $\Rightarrow$ **almost extractor**
- Flat distribution = a distribution that puts equal probability on the elements of a subset of the space, and nothing elsewhere.
- Since each $2^k \times 2^k$ rectangle is balanced, $E$ extracts well from flat distributions with min-entropy $k$.
- Each distribution with min-entropy $k$ is a convex combination of flat distributions with min-entropy $k$.
Combining the combinatorial characterizations of Kolmogorov extractors and almost extractors, we obtain the equivalence theorem.

**Theorem**

Let \( E : \{0, 1\}^n \times \{0, 1\}^n \rightarrow \{0, 1\}^m \) be an ensemble of functions.

1. If \( E \) is a \((k, \epsilon, d)\) almost extractor, then \( E \) is a \((k', \alpha, \alpha + 2d + 1)\) Kolmogorov extractor, where \( k' = k + \log n + O(1) \), \( \alpha = \log(1/\epsilon) + d + 1 \).

2. If \( E \) is a \((k, \alpha, d)\) Kolmogorov extractor, then \( E \) is a \((k', \epsilon, d')\) almost extractor, where \( k' = k + \alpha \), \( \epsilon = 2 \cdot 2^{-\alpha} \), \( d' = d + O(1) \).
Proofs and other goodies in the survey paper: Possibilities and impossibilities in Kolmogorov extraction, SIGACT News, Dec. 2010

Thank you.