**0368-4283: Space-Bounded Computation** 08/05/2018 – Lecture 8

# Extractors

*Amnon Ta-Shma and Dean Doron* *Scribe: Noam Parzenchevski*

## 1 Motivation

Consider the following scenario: Alice and Bob hold a random string $x \in \{0,1\}^n$ and wish to use it to communicate securely. Meanwhile, Eve gained access to $n/3$ bits of information about $x$. Can Alice and Bob somehow modify $x$ to get an $x'$ of length roughly $2n/3$ which would appear (almost-) random to Eve?

In this lecture we will try to achieve this goal - given a "flawed" distribution $X \subseteq \{0,1\}^n$ along with a small auxiliary random seed $d$, we will construct a distribution $X'$ which is $\epsilon$-close to uniform over $\{0,1\}^m$ (where $m < n$), while trying to minimize $d$ and maximize $m$.

## 2 Preliminaries

A good measure of the amount of "randomness" in a distribution is its min entropy:

**Definition 1.** *(weak source) Let $X$ be a distribution over $\{0,1\}^n$. The* min-entropy *of $X$ is $H_\infty(X) = \log \frac{1}{\max_a X(a)}$. We say $X$ is a $k$-source if $H_\infty(x) \geq k$, or, equivalently,* $\Pr(X = x) \leq 2^{-k}$ *for every $x \in X$.*

*As an example, $U_d$, the uniform distribution over $\{0,1\}^d$, is a $d$-source.*

**Definition 2.** *(statistical distance) For two distributions $X, Y \subseteq \Omega$, we define the statistical distance:*

$$|X - Y| = \frac{1}{2} \cdot \sum_{x \in \Omega} |\Pr[X = x] - \Pr[Y = x]| = \max_{\Lambda \subseteq \Omega} |\Pr[X \in \Lambda] - \Pr[Y \in \Lambda]|$$

*If $|X - Y| \leqslant \epsilon$ we say that $X$ is $\epsilon$-close to $Y$. We will sometimes omit the $\frac{1}{2}$ factor.*

The statistical distance between two distributions $X, Y$ captures the best way to distinguish between the two. It is not hard to see that the test $T \subseteq \Omega$ which separates $X$ and $Y$ best is the test defined by $T = \{x \in \Omega \mid \Pr[X = x] > \Pr[Y = x]\}$ the distance given by this test is exactly $\Pr[X \in T] - \Pr[Y \in T]$.

We're now ready to define an extractor:

**Definition 3.** *(extractor) Let* $\mathsf{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *be a function*

- *Let $\mathcal{F}$ be a family of distributions over $\{0,1\}^n$. We say* $\mathsf{Ext}$ *is an $(\mathcal{F}, \epsilon)$-extractor, if for every $X \in \mathcal{F}$, $|E(X, U_d) - U_m| \leq \epsilon$.*
- *We say* $\mathsf{Ext}$ *is a $(k, \epsilon)$-extractor, if it is an $(\mathcal{F}, \epsilon)$ extractor for the family $\mathcal{F}$ of all $k$-sources.*

We think of $\mathsf{Ext}(X, U_d)$ as a random variable defined as follows: Pick an $x \sim X$, independently pick $y \sim U_d$, and output $\mathsf{Ext}(x, y)$.

Before we proceed, we want to show that wlog, when talking about $k$-source we can consider only flat sources (that is, uniform sources over $2^k$ elements). We first claim:

**Claim 4.** *Any $k$-source $X$ is a convex combination of flat sources over $2^k$ elements*

*Proof.* Any $k$-source $X$ over $\{0,1\}^n$ can be defined by the following system of linear equations where $\Pr[X = x_i] = p_i$ :

- $\sum_{i=1}^{2^n} p_i = 1$
- For any $i : 0 \leqslant p_i \leqslant 2^{-k}$

This set of equations defines a convex polytope whose vertices are given by sources where the maximal number of inequality constraints are satisfied tightly. I.e. - for any $i : p_i \in \{0, 2^{-k}\}$. By convexity, any point in the polytope can be expressed as a convex combination of the vertices of the polytope. The claim follows □

With that, we prove the following:

**Claim 5.** *If $\mathsf{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ is an $\epsilon$-extractor for the family of flat $k$-sources then it is a $(k, \epsilon)$-extractor*

*Proof.* Let $X$ be a $k$-source. By Fact 4 we can write $X = \sum \lambda_i F_i$ where $0 \leqslant \lambda_i \leqslant 1$, $\sum \lambda_i = 1$ and $F_i$ are flat sources. Let $F_{\max} = \max_{F_i} |\mathsf{Ext}(F_i, U_d) - U_m|$. If we think of $X$ as picking a flat source $F_i$ w.p. $\lambda_i$ and then sampling $F_i$, it is easy to see that:

$$\begin{aligned}
|\mathsf{Ext}(X, U_d) - U_m| &= \left|\sum \lambda_i \left(\mathsf{Ext}(F_i, U_d - U_m\right)\right| \\
&\leqslant \sum \lambda_i \left|\mathsf{Ext}(F_i, U_d) - U_m\right| \\
&\leqslant \left(\sum \lambda_i\right) \left|\mathsf{Ext}(F_{\max}, U_d) - U_m\right| \\
&\leqslant \left|\mathsf{Ext}(F_{\max}, U_d) - U_m\right| \\
&\leqslant \epsilon
\end{aligned}$$

□

As a warm-up, we show that there are no deterministic extractors for general $k$-sources. Indeed, even if we get $n-1$ bits of entropy we cannot output a single uniform bit:

**Claim 6.** *For any $\mathsf{Ext} : \{0,1\}^n \to \{0,1\}$ there exists an $(n-1)$-source $X$ s.t. $|\mathsf{Ext}(X) - U_1| = 1$*

*Proof.* Assume wlog that $|\mathsf{Ext}^{-1}(0)| \geqslant 2^{n-1}$ (otherwise take $\mathsf{Ext}^{-1}(1)$) and let $X$ be the uniform distribution over $\mathsf{Ext}^{-1}(0)$. Clearly, $X$ is an $(n-1)$-source, however, $\mathsf{Ext}(X) = 0$, thus $|\mathsf{Ext}(X) - U_1| = 1$ □

## 3 Affine extractors

As an example, we first turn our attention to a specific family of distributions - uniform distributions over affine spaces:

**Definition 7.** *(Set of affine spaces) For a vector space $V$ of dimension $n$, let*

$$\mathrm{Affn}_{n,k} = \{A \subseteq V : \exists z \in V \text{ and a subspace } U \subseteq V \text{ s.t. } \dim U = k \text{ and } A = U + z\}$$

*Each $X \in \mathrm{Affn}_{n,k}$ induces a distribution which is simply the uniform distribution over $X$. In this lecture, we will consider only the case $V = \mathbb{F}_2^n$.*

Clearly, $\mathrm{Affn}_{n,k}$ is a family of $k$-sources. We will use the fact that it is a fairly small family to show the existence of a deterministic extractor against this family.

**Theorem 8.** *For all $k \geqslant 2\log n + 2\log(\frac{1}{\varepsilon}) + O(1)$ there exists an $(\mathrm{Affn}_{n,k}, \epsilon)$-extractor $\mathsf{Ext} : \{0,1\}^n \to \{0,1\}^m$ where $m = k - 2\log(\frac{1}{\varepsilon}) - O(1)$*

*Proof.* We us the probabilistic method. For a distribution $X \in \mathrm{Affn}_{n,k}$ and $S \subseteq \mathbb{F}_2^m$ we'll say that $\mathsf{Ext}$ fails on $(X, S)$ if:

$$\left| \Pr_{x \in X}[\mathsf{Ext}(x) \in S] - \rho(S) \right| > \epsilon$$

By the definition of statistical distance, it is easy to see that if $\mathsf{Ext}$ passes over all $(X, S) \in \mathrm{Affn}_{n,k} \times \mathbb{F}_2^m$ then it is an $(\mathrm{Affn}_{n,k}, \epsilon)$-extractor.

Fix then a pair $(X, S)$. For each element $x \in X$ define the indicator r.v. $Y_x = 1$ iff $\mathsf{Ext}(x) \in S$. Note that $\Pr_x[\mathsf{Ext}(x) \in S] = \frac{1}{2^k}\sum_{x'} Y_{x'}$. Clearly, for a random $\mathsf{Ext}$ we have $\frac{1}{2^k}\mathbb{E}(\sum_x Y_x) = \rho(S)$, thus by Chernoff:

$$\Pr_{\mathsf{Ext}}\left[ \left| \Pr_{x \in X}[\mathsf{Ext}(x) \in S] - \rho(S) \right| > \epsilon \right] \leqslant 2^{-2\epsilon^2 2^k}$$

By a union bound:

$$\begin{aligned}
\Pr_{\mathsf{Ext}}[\exists (X,S) : \mathsf{Ext} \text{ fails on } (X,S)] &\leqslant |\mathrm{Affn}_{n,k}| \cdot \mathcal{P}(\mathbb{F}_2^m) \cdot 2^{-2\epsilon^2 2^k} \\
&\leqslant \underbrace{2^n}_{\text{choose } z} \underbrace{\binom{2^n}{k}}_{\text{choose } U} \cdot 2^{2^m} \cdot 2^{-2\epsilon^2 2^k} \\
&\leqslant 2^{n(k+1) - \epsilon^2 2^k} \cdot 2^{2^m - \epsilon^2 2^k}
\end{aligned}$$

So it suffices to require both:

1. $2^{n(k+1) \cdot 2^{-\epsilon^2 2^k}} < 1$ for which $k > 2\log n + 2\log(\frac{1}{\varepsilon}) + O(1)$ suffices
2. $2^{2^m \cdot 2^{-\epsilon^2 2^k}} < 1$ which implies $k > m + 2\log(\frac{1}{\varepsilon}) + O(1)$

The claim now follows □

It is worth noting that the proof above did not use the structure of the source (the fact that it is a shift of a linear subspace) but only the fact that the size of the family of subspaces is small. Our next (and main) task will be to construct extractors against general $k$-sources.

We're now ready to construct our extractors. Before we start we mention that via a probabilistic argument similar to the one presented above one can show the existence of $(k, \epsilon)$-extractors where $d = \log(n-k) + 2\log(\frac{1}{\varepsilon}) + O(1)$ and $m \geqslant k + d - 2\log(\frac{1}{\varepsilon}) - O(1)$. Additionally, known lower bounds state that any $(k, \epsilon)$-extractor must have both $d \geqslant \log(n-k) + 2\log(\frac{1}{\varepsilon}) - O(1)$ and $m \leqslant k + d - 2\log(\frac{1}{\varepsilon}) - O(1)$, see for example [RTS00], Theorem 1.9.

# 4 Constructing extractors from expanders

We present two constructions based on expander graphs. Throughout this section we denote by captital letters exponential cardinality, e.g. - $A = 2^a$.

## 4.1 First attempt - single step on an expander

**Theorem 9.** *Let $k = n - O(1)$ then there exists a $(k, \epsilon)$-extractor:*

$$\mathsf{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^n$$

*Where $d = 2\log(\frac{1}{\varepsilon}) + O(1)$*

*Proof.* Given an $(N, D, \lambda)$-expander $G = (V, E)$, we have a natural function $\mathsf{Ext} : [N] \times [D] \to [N]$ induced by the structure of the graph: $\forall x \in V : \mathsf{Ext}(x, i) = x[i]$. A flat $k$-source over $V$ is simply a subset $X \subseteq V$ s.t. $|X| = K$ and we can think of $\mathsf{Ext}(X, U_d)$ as picking a vertex in $x \in X$ u.a.r and then stepping along a random edge of $x$ u.a.r to a neighbor $y$. As before, we want to require that for any $k$-source $X$ and $S \subseteq V$ we have

$$\left| \Pr_{x \in X, i \in [D]}[\mathsf{Ext}(x,i) \in S] - \rho(S) \right| \leqslant \epsilon$$

As $X, [D]$ are flat, it is easy to see that

$$\Pr_{x,i}[\mathsf{Ext}(x,i) \in S] = \frac{|E(X,S)|}{|X| \cdot D}$$

And from the Expander Mixing Lemma we know that for any $X, S$ as above:

$$\left| \frac{|E(X,S)|}{|X| \cdot D} - \rho(S) \right| \leqslant \lambda \sqrt{\frac{\rho(S)}{\rho(X)}} \leqslant \lambda \sqrt{\rho^{-1}(X)} = \lambda \cdot 2^{-\frac{k-n}{2}}$$

Thus it suffices to require $\lambda \cdot 2^{-\frac{k-n}{2}} \leqslant \epsilon$. Assuming $G$ is Ramanujan we have $\lambda \sim \frac{2}{\sqrt{D}}$ thus we need $\frac{2}{\sqrt{D}} \cdot 2^{-\frac{k-n}{2}} \leqslant \epsilon$ which rearranges to $D \geqslant \Omega\left(\frac{2^{n-k}}{\epsilon^2}\right)$ or equivalently $d = (n-k) + 2\log(\frac{1}{\varepsilon}) + O(1) = 2\log(\frac{1}{\varepsilon}) + O(1)$. We note that by the lower bound of [RTS00], this is tight up to the constant factor $\square$

## 4.2 Second attempt - walking on an expander

Our previous construction worked only when the entropy deficiency $n - k$ was constant, we now construct extractors with logarithmic seed length which work for $k$-sources where $k$ is some constant fraction of $n$ and $\epsilon$ is a constant. Let $G$ be an $(M, D, \lambda)$-expander as before. As in the previous section we will use our expander to construct a bipartite graph which induces an extractor, but this time we will consider the left side of the graph to be all length $t$-walks on $G$ and the right side as the vertices of $G$. We will connect each path with all vertices that lie on said path. It is easy to see that this construction yields a $([N = MD^t], [M], t)$ bipartite graph $G_{path}$, which induces a $\mathsf{Ext} : [MD^t] \times [t] \to [M]$ extractor. In what follows we will analyze the extraction properties of this function.

Recall the definition of a sampler. Informally, a sampler is a bipartite graph where for each subset $T$ of the right side, most vertices on the left side fall into roughly the density of $T$. In this sense, sampling vertices from the left side approximates the density of subsets on the right side. Formally:

**Definition 10.** *(Sampler) A $D$-left regular bipartite graph $S = (A, B)$ is a $(\delta, \epsilon)$-sampler if for any $T \subseteq B$ we have:*

$$|\mathrm{Bad}_T| = |\{\epsilon\text{-}bad\ elements\ in\ A\ w.r.t\ T\}| \leqslant \delta |A|$$

*Where we say that $v \in A$ is bad w.r.t $T$ if:*

$$\left| \Pr_{i \in [D]} [v[i] \in T] - \rho(T) \right| \geqslant \epsilon$$

*We think of $S$ as a function $S : [A] \times [D] \to [B]$ as before.*

**Claim 11.** *If $S = (A, B)$ is a $(\delta, \epsilon)$-sampler then it is also a $(k, 2\epsilon)$-extractor for $k = \log\left(\frac{\delta|A|}{\epsilon}\right)$*

*Proof.* As before, we know that $S$ is a $(k, 2\epsilon)$-extractor iff for all $X \subseteq A$ of size $|X| = K = \frac{\delta|A|}{\epsilon}$ we have:

$$\forall T \subseteq B : \left| \Pr_{x,i} [S(x, i) \in T] - \rho(T) \right| \leqslant 2\epsilon$$

Now, clearly:

$$\Pr_{x,i} [S(x, i) \in T] \leqslant \Pr_{x} [x \in \mathrm{Bad}_T] + \Pr_{x,i} [S(x, i) \in T \mid x \notin \mathrm{Bad}_T]$$

By our choice of $K$, $\Pr_x [x \in \mathrm{Bad}_T] \leqslant \epsilon$, and by defintion $\Pr_{x,i} [S(x, i) \in T \mid x \notin \mathrm{Bad}_T] = \rho(T) \pm \epsilon$, thus together $\Pr_{x,i} [S(x, i) \in T] \leqslant \rho(T) + 2\epsilon$ and therefore

$$\forall T \subseteq B : \left| \Pr_{x,i} [S(x, i) \in T] - \rho(T) \right| \leqslant |\rho(T) + 2\epsilon - \rho(T)| = 2\epsilon$$

as needed ☐

By the expander Chernoff bound [Hea08], we know that for any subset $S \subseteq M$ and a $t$-long walk $v_1, \ldots, v_t$ if we let $\mathbb{I}_{v_i \in S}$ be an indicator for $v_i \in S$ we have

$$\Pr\left[ \left| \frac{1}{t} \sum_i \mathbb{I}_{v_i \in S} - \rho(S) \right| > \epsilon \right] \leqslant \delta = 2e^{-\frac{\epsilon^2 \cdot \gamma \cdot t}{4}}$$

Where $\gamma = 1 - \lambda$. It is easy to see that this is equivalent to saying that $G_{path}$ is a $(\delta, \epsilon)$-sampler, thus by Claim 11 if $k = \frac{\delta N}{\epsilon}$ then $\mathsf{Ext}$ is $(k, 2\epsilon)$-extractor.

An immediate corollary therefore is:

**Theorem 12.** *For any* $\epsilon, \alpha$ *there exists a* $\zeta = \Omega(\epsilon^2 \cdot \alpha)$ *for which there exists a* $(k = (1-\zeta)n + \log(\frac{1}{\varepsilon}), 2 \cdot \epsilon)$*-extractor*

$$\mathsf{Ext} : \{0,1\}^n \times \{0,1\}^r \to \{0,1\}^m$$

*with* $r \leqslant \log \alpha n$ *and* $m = (1-\alpha)n$

*Proof.* Given $n, \alpha$, we set $m = (1-\alpha)n$ and take our original graph $G$ to be an $(M = 2^m, D = 2^d, \lambda)$-expander where $D$ and $\lambda < 1$ are absolute constants. We now build $G_{path}$ on $N = 2^n$ vertices. We consider each vertex as a register specifiying an initial vertex in $G$ (given by $m$ bits) and $t$ instructions for the next step, each given by $d$ bits. Thus, we require $n = m + td = (1-\alpha)n + td$ or equivalently, $t = \frac{\alpha n}{d}$ (note: $r \stackrel{\text{def}}{=} \log t = \log \frac{\alpha n}{d} \leqslant \log \alpha n$).

Next, we know that if $K = \frac{\delta N}{\epsilon}$ then $G_{path}$ induces a $(k, 2\epsilon)$-extractor. Thus, we set:

$$K = \frac{\delta N}{\epsilon} = \frac{1}{\epsilon} \cdot 2e^{-\frac{\epsilon^2 \gamma t}{4}} \cdot 2^n = \frac{1}{\epsilon} \cdot 2e^{-\frac{\epsilon^2 \gamma (n-m)}{4 \cdot d}} \cdot 2^n = \frac{1}{\epsilon} \cdot 2e^{-\frac{\epsilon^2 \gamma \alpha n}{4 \cdot d}} \cdot 2^n = 2^{n(1-\frac{\gamma}{4 \cdot d} \cdot \epsilon^2 \alpha) + \log(\frac{1}{\varepsilon})}$$

as $d, \gamma$ are absolute constants, the claim follows by setting $\zeta = \frac{\gamma}{4 \cdot d} \cdot \epsilon^2 \alpha$ □

We note that for any *constants* $\epsilon, \alpha$, $\mathsf{Ext}$ is a $(k, 2 \cdot \epsilon)$-extractor with logarithmic seed length requiring min-entropy of a constant fraction of $n$ as promised in the beginning of the section.

## 5 Condensing randomness

We want to conclude by addressing the case where we are given a $k$-source such that $k \ll n$. Let us assume that we have some extractor $\mathsf{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ which requires a seed of length $d = \Theta(\log^2 \frac{n}{\epsilon})$. Given a $k$-source $X$, if we were first able to convert $X \to X'$ where $X'$ is a $k'$-source over $\{0,1\}^{n'}$ with $n', k' \sim k$ then we could apply the above extractor using only $d' = \Theta(\log^2 \frac{k}{\epsilon})$ truly random bits. For the case where $k \ll n$, the difference between $d, d'$ could be prohibitive. It turns out that such a goal is possible, and for that we introduce the notion of condensers:

**Definition 13.** *(Condenser) A function:*

$$C : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$$

*is a* $k \to_\epsilon k'$ *condenser if for any* $k$*-source* $X$ *it holds that* $C(X, U_d)$ *is* $\epsilon$*-close to some* $k'$*-source. Furthermore, if* $k' = k + d$ *we say that* $C$ *is lossless.*

In a (fairly) recent line of work, lossless condensers were built with optimal parameters. We cite without proof the following condenser due to [GUV09], which is based on the Parvaresh-Vardy error correcting code:

**Theorem 14** ([GUV09])**.** *There exists an explicit, lossless,* $k \to_\epsilon k + d$ *condenser:*

$$C : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$$

*where* $d = O(\log n + \log(\frac{1}{\varepsilon}))$ *and* $m = 1.0001(k+d)$

With this condenser, a general scheme for building an extractor would be to work in two steps. Given a source over $\{0,1\}^n$ we first apply a condensing step and then extract the randomness using e.g. our expander walk extractor:

$$\{0,1\}^n \to_C \{0,1\}^{1.0001k} \to_{\mathsf{Ext}} U_m$$

giving us a $(k, \epsilon)$-extractor with $d = O(\log \frac{n}{\epsilon})$ and $m = \Omega(k)$

## References


[GUV09] Venkatesan Guruswami, Christopher Umans, and Salil Vadhan. Unbalanced expanders and randomness extractors from parvaresh–vardy codes. *Journal of the ACM (JACM)*, 56(4):20, 2009.

[Hea08] Alexander D Healy. Randomness-efficient sampling within nc. *Computational Complexity*, 17(1):3–37, 2008.

[RTS00] Jaikumar Radhakrishnan and Amnon Ta-Shma. Bounds for dispersers, extractors, and depth-two superconcentrators. *SIAM Journal on Discrete Mathematics*, 13(1):2–24, 2000.