Lecture 10: Condensers

Topics in Complexity Theory and Pseudorandomness (Spring 2013)
Rutgers University
Swastik Kopparty
Scribes: Ian Mertz, Yun Kuen Cheung

1 A Quick Recap

Last class we were studying randomness extractors, which convert impure sources of randomness into uniformly distributed random variables (which could be then be used to fuel our randomized algorithms, say). The impure sources from which we want to extract randomness must have some amount of randomness in them, and the measure of randomness that we used for this purpose was min-entropy.

Definition 1. *The min-entropy of a distribution X, written as $H_\infty(X)$ is defined by: $H_\infty(X) \leq k$ iff $\forall x, Pr[X = x] \leq 2^{-k}$. The units of min-entropy are* bits.

Note that if X is distrubuted over $\{0,1\}^n$, then $H_\infty(X) = n$ iff $X = U_n$.

Definition 2. *A function $E : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ is a (k, ϵ) extractor if for all distributions X on $\{0,1\}^n$ with $H_\infty(X) \geq k$, $E(X, U_d)$ is ϵ-close to U_m in statistical distance.*

Thus if we have a sample $w \in \{0,1\}^n$ which came from a k-bit min-entropy source, then by investing some d bits of pure randomness, we can get ourselves m bits of almost-pure randomness.

This would be interesting only if $m > d$, and indeed that is the case.

Theorem 3. *Extractors exist with $d = \log(n-k) + 2\log(\frac{1}{\epsilon}) + O(1)$, $m = k + d - 2\log(\frac{1}{\epsilon})$.*

Thus one can get back all d bits of randomness plus almost all of the k-bits of entropy that lie in the weak source!

This motivates the problem of constructing an extractor explicitly. Last lecture we constructed an extractor with $m = k + d - 2\log\frac{1}{\epsilon}$ (which is optimal), but which required $d = 2n$ seed length. In this lecture we will construct an extractor with both seed length and output length optimal upto constant factors (i.e., $d = O(\log n)$ and $m = \Omega(k + d)$).

2 Extractor for High Min-Entropy Sources

We first construct an extractor which works when $k = 0.999n$.

This will be based on the following Chernoff-like bound for random walks on expanders.

Theorem 4. *Let G be a C-regular, λ-absolute eigenvalue expander. Pick $x_0 \in V$ uniformly. Let $x_1 \ldots x_D$ be a random walk on G. Then $Pr[\frac{\#i \text{ for which } x_i \in S}{D} > \frac{|S|}{|V|} + \epsilon] \leq e^{-\epsilon^2 D(1-\frac{\lambda}{C})/4}$.*

Since the first vertex x_0 of this walk is picked uniformly from V, it is easy to see that each x_i is distributed uniformly over V. Thus the expected fraction of $i \in [D]$ for which $x_i \in S$ equals $\frac{|S|}{|V|}$. The above theorem says that the x_i sample with error bounds which are as nearly good as what one would get if they were sampled independently.

The way we get an extractor out of this is the following. We will first fix a large constant degree expander graph G. We identify $\{0,1\}^n$ with the space of random walks of length D on G. We will use the seed to index an integer $i \in [D]$. The output of the extractor $E(x,i)$, given an input $x \in \{0,1\}^n$ and $i \in [D]$, will be the ith vertex of the random walk corresponding to x.

Thus the bipartite graph associated to this extractor looks like the following:

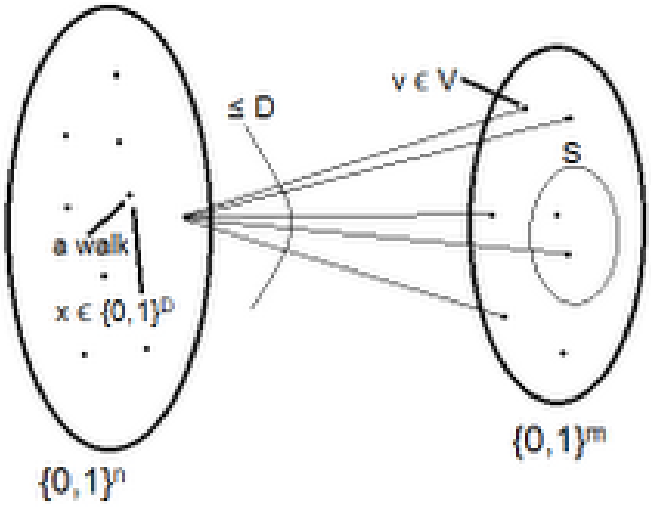

Figure 1: The expander walk extractor

The total number of walks $= 2^n = |V|C^D$. The total number of vertices $= 2^m = |V|$. Choose $D = \Theta(n)$ (so that the seed length $d = \log n + O(1)$).

We now prove that this is (k, ϵ) extractor with $k = (1 - \Omega(\epsilon^2)) \cdot n$. Let X be a distribution over $\{0,1\}^n$ with $H_\infty(X) \geq k$. Without loss of generality, we assume that X is a flat distribution.

Suppose $E(X, U_d)$ is not ϵ-close to U_n. Then by the "distinguisher" characterization of statistical distance, there exists a set $S \subseteq \{0,1\}^m$ such that:

$$\Pr[E(X, U_d) \in S] - \Pr[U_m \in S] \geq \epsilon.$$

Then we have that

$$\Pr_{x \in X, i \in U_d}[E(x,i) \in S] \geq \frac{|S|}{|V|} + \epsilon.$$

This implies that most $x \in X$ are bad:

$$\Pr_{x \in X}\left[\Pr_{i \in U_d}[E(x,i) \in S] \geq \frac{|S|}{|V|} + \frac{\epsilon}{2}\right] \geq \frac{\epsilon}{2}.$$

(this kind of manipulation shows up in many many places and is very useful).

This means that many walks in X have the property that they do not sample the set S well. But by the Expander Chernoff bound, there are not too many such walks:

$$\left|\{x \in \{0,1\}^n \text{ s.t. } \Pr_{i \in U_d}[E(x,i) \in S] \geq \frac{|S|}{|V|} + \frac{\epsilon}{2}\}\right| \leq e^{-\Omega(\epsilon^2 D)} 2^n$$

Putting these together:

$$\frac{\epsilon}{2}|X| \leq e^{-\Omega(\epsilon^2 D)},$$
$$\frac{\epsilon}{2}2^k \leq e^{-\Omega(\epsilon^2 D)}$$
$$k \leq n - \Omega(\epsilon^2 D)$$

But since $D = \Theta(n)$, this implies that

$$k \leq n - \Omega(\epsilon^2 D) \leq n(1 - \Omega(\epsilon^2)),$$

which contradicts our assumption on k. Thus, $E(X, U_d)$ is in fact ϵ-close to U_m.

3 Condensers

Having constructed an extractor that can extract from sources with high min-entropy, this naturally motivates us to define condensers: these will be devices that can convert sources distributed over $\{0,1\}^n$ with min-entropy k to a random variable distributed over $\{0,1\}^{(1.01)\cdot k}$ with min-entropy k.

Definition 5. *A function $C : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ is a (k, ϵ, s) condenser if for all distributions X on $\{0,1\}^n$ with $H_\infty(X) \geq k$, we have that $C(X, U_d)$ is ϵ-close to some distribution Z on $\{0,1\}^m$ with $H_\infty(Z) \geq s$ in the L_1 norm.*

The importance of this definition is the following theorem:

Theorem 6. *If $C : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^{n'}$ is a (k, ϵ, k') condenser, and $E : \{0,1\}^{n'} \times \{0,1\}^{d'} \to \{0,1\}^m$ is a (k', ϵ') extractor, then $\overline{E} : \{0,1\}^n \times \{0,1\}^{d+d'} \to \{0,1\}^m$ given by $\overline{E}(x, (y_1, y_2)) = E(C(x, y_1), y_2)$ is a $(k, \epsilon + \epsilon')$ extractor.*

Proof. Take a weak random source X with $H_\infty(X) \geq k$. By the condenser hypothesis, $C(X, U_d)$ is ϵ-close to some Z with $H_\infty(Z) \geq k'$. It follows that $E(Z, U_{d'})$ is ϵ'-close to U_m. Now we can say that $C(X, U_d)$ is ϵ-close to Z, which implies that $E(C(X, U_d), U_{d'})$ is ϵ-close to $E(Z, U_{d'})$, which is ϵ'-close to U_m. Therefore, $E(C(X, U_d), U_{d'})$ is $\epsilon + \epsilon'$-close to U_m, and so $\overline{E}(X, (U_d, U_{d'}))$ is $\epsilon + \epsilon'$-close to U_m. □

Our intuition here is to start with $d + d'$ bits of randomness, and instead of just using an extractor, to use the composition of a condenser and an extractor to (hopefully) recover more entropy.

Now the

Theorem 7. *There exist (k, ϵ, s) condensers with $d = \log(n) + \log(\frac{1}{\epsilon})$, $m = s + \log(\frac{1}{\epsilon})$, and $s = k + d$.*

In other words, there exist *lossless* condensers that recovers all the original entropy of the weakly random variable, in addition to all the bits of randomness used.

Given a condenser, one can form the associated $(2^n, 2^m)$ bipartite graph. For a lossless condenser, the associated bipartite graph is an expander where each left vertex has degree D, and where sets of size k on the left have $kD(1-\epsilon)$ neighbors (such expanders are called lossless expanders). Furthermore, if the bipartite graph associated with a condenser is a lossless expander, the condenser is itself lossless. (Prove this!)

4 Simple Univariate Polynomial Condenser

Here we construct two condensers using finite fields and polynomials, the first weaker condenser motivating the second optimal one. Their analysis involves some very nice ideas from linear algebra.

4.1 Version 1: condensers from Reed-Solomon codes

Fix a prime power q and positive integer c. Let $\mathcal{P}$ be the set of polynomials with coefficients in $\mathbb{F}_q$ and degree at most $c-1$, and identify it with $\{0,1\}^n$; identify $\mathbb{F}_q$ with $\{0,1\}^d$; identify $\mathbb{F}_q^2$ with $\{0,1\}^m$. For any $f \in \mathcal{P}$ and $\alpha \in \mathbb{F}_q$, define $C(f,\alpha) = (\alpha, f(\alpha))$.

We view the condenser C as a bipartite graph with left vertices from $\mathcal{P}$ and right vertices from $\mathbb{F}_q^2$. For each $f \in \mathcal{P}$, its neighborhood $\Gamma(f)$ is of size q. We want this graph be a good bipartite expander, i.e. for each $S \subset \mathcal{P}$ with $|S| = K$, $\Gamma(S) \geq AK$ for some constant A (the values of K and A determined later).

The analysis of the condenser will proceed as follows: we will show that for any $T \subset \mathbb{F}_q^2$ with $|T| < AK$, $|\{f \in \mathcal{P} \,|\, \Gamma(f) \subset T\}| < K$.

Lemma 8. *If $A + cK \leq q$, then for any $T \subset \mathbb{F}_q^2$ with $|T| < AK$, $|\{f \in \mathcal{P} : \Gamma(f) \subset T\}| < K$.*

Proof. Take any set $T \subseteq \mathbb{F}_q^2$ with $|T| < AK$.

The crucial step: Interpolate a nonzero bivariate polynomial $Q(X,Y) = \sum_{i=0}^{A-1} \sum_{j=0}^{K-1} \alpha_{ij} X^i Y^j$ such that $Q(x,y) = 0$ for all $(x,y) \in T$. Such a Q exists because (1) the space of all bivariate polynomials of those given degrees is a vector space of dimension AK (2) Each vanishing condition $Q(x,y) = 0$ imposes one homogeneous linear constraint on the coefficients of the bivariate polynomial, and there are $< AK$ such homogeneous constraints.

Suppose $f(X) \in \mathcal{P}$ is such that $\Gamma(f) \subseteq T$. We then get that $Q(\alpha, f(\alpha)) = 0$ for all $\alpha \in \mathbb{F}_q$. Define the univariate polynomial$H(X) = Q(X, f(X))$. Note that $H(X)$ is a polynomial in X of degree at most $(A-1)+(c-1)(K-1)$, but $H(\alpha) = 0$ for all $\alpha \in \mathbb{F}_q$. Since $(A-1)+(c-1)(K-1) < A+cK \leq q$, $H(X)$ is the zero polynomial.

Thus $Q(X, f(X)) = H(X) \equiv 0$, and so $Y - f(X)$ divides $Q(X,Y)$ (this is the remainder theorem for polynomials). Since $Y - f(X)$ are pairwise coprime for distinct f's, and the Y-degree of $Q(X,Y) = K-1$, there must be $< K$ distinct f's with $\Gamma(f) \subseteq T$. □

Now we analyze the condenser. Without loss of generality, we just analyze the case where the weak

random source is a flat distribution. Let S be any subset of $\{0,1\}^n$ with size 2^k, and let U_S be the uniform distribution over S. It suffices to show that $C(U_S, U_d)$ is ϵ-close to having min-entropy $\geq k + d$.

Let $A = q - 2^k c$. By Lemma 8, $|\Gamma(S)| \geq 2^k A$. Thus $C(U_S, U_d)$ is $1 - \frac{A}{q} = \frac{2^k c}{q}$-close to the uniform distribution over a set T' with $|T'| = 2^k q$, which has min-entropy $\lg |T'| = k + \lg q = k + d$.

Thus $C(U_S, U_d)$ is $\frac{2^k c}{q}$-close to a distribution with min-entropy $k+d$, and so C is a $\left(k, \frac{2^k c}{q}, k+d\right)$-condenser.

This is interesting only when k is very small compared to n. Here is one example of the kind of parameters we can get in this setting: Choose $k = \frac{99}{100} \log q$, and $c = q^{\frac{1}{1000}}$. Then $n = q^{\frac{1}{1000}} \log q$, $d = \log q = O(\log n)$, $k = \frac{99}{100} \log q = O(\log n)$, $m = 2 \log q \leq (1.01) \cdot (k+d)$ and $\epsilon < q^{-\frac{1}{1000}} = \frac{1}{\mathtt{poly}(n)}$.

4.2 Version 2: condensers from Parvaresh-Vardy codes

The previous construction was naturally limited by the fact that the output was very small ($\mathbb{F}_q^2$). Our plan now is to increase the output size, by evaluating not one polynomial, but many polynomials.

Let $\ell > 0$ be an integer. We will first choose a special subset $\mathcal{P}' \subset \mathcal{P}^\ell$, and identify $\mathcal{P}'$ with $\{0,1\}^n$. For any $(f_1, f_2, \cdots, f_\ell) \in \mathcal{P}'$ and $\alpha \in \mathbb{F}_q$, let $C((f_1, f_2, \cdots, f_\ell), \alpha) = (\alpha, f_1(\alpha), f_2(\alpha), \cdots, f_j(\alpha))$. Thus the output space, $\mathbb{F}_q^{\ell+1}$, is identified with $\{0,1\}^m$. Pictorially, $\mathcal{P}'$ is a collection of curves, and the condenser uses the weak random source to pick a curve from this collection, and then outputs a random point on this curve.

The key property we will need of $\mathcal{P}'$ is a bound on the number of curves from $\mathcal{P}'$ on which any given multivariate polynomial formally vanishes. We will later see how to choose a large $\mathcal{P}'$ with this property.

Property Z: If $Q(X, Y_1, Y_2, \cdots, Y_\ell)$ is a multivariate polynomial in $\mathbb{F}_q$ with individual degrees $\leq (A-1, h-1, h-1, \cdots, h-1)$, then

$$\left|\{(f_1, f_2, \cdots, f_\ell) \in \mathcal{P}' \mid Q(X, f_1(X), f_2(X), \cdots, f_\ell(X)) \equiv 0\}\right| \leq h^\ell - 1.$$

Lemma 9. *Suppose $\mathcal{P}'$ has Property Z. Suppose $h = K^{1/\ell}$ is an integer. If $A + c\ell h \leq q$, then for any $T \subseteq \mathbb{F}_q^{j+1}$ with $|T| < AK$:*

$$\left|\{\vec{f} \in \mathcal{P}' \mid \Gamma(\vec{f}) \subset T\}\right| < K.$$

.

Remark: if K is not a perfect ℓ-th power, this lemma is still true, but the argument involves one extra trick which we do not discuss here.

Proof. As argued in Lemma 8, there exists a multivariate polynomial $Q(X, Y_1, Y_2, \cdots, Y_\ell)$ vanishing on T with individual degrees at most $(A-1, h-1, h-1, \cdots, h-1)$. Let $\vec{f} = (f_1, f_2, \cdots, f_\ell)$ be such that $\Gamma(\vec{f}) \subseteq T$. Define $H(X) = Q(X, f_1(X), f_2(X), \cdots, f_\ell(X))$. By definition of Q, $H(\alpha) = 0$ for all $\alpha \in \mathbb{F}_q$. The degree of the polynomial H is at most $(A-1) + (c-1)j(h-1) < A + cjh \leq q$. Thus $H(X) = Q(X, f_1(X), \ldots, f_\ell(X))$ is the zero polynomial.

By Property Z, the number of such $\vec{f} \in \mathcal{P}'$ is at most $h^j - 1 = K - 1$. $\square$

As before, this lemma implies that C is a (lossless) $(\log K, \frac{c\ell h}{q}, \log K + \log q)$ condenser.

Now we move to the problem of constructing a large family $\mathcal{P}'$ with Property Z. (The reason we want $\mathcal{P}'$ to be large is so that the seed-length $d = \log q$ of the condenser, is logarithmic in $n = \log |\mathcal{P}'|$, the size of the weak random source).

Let us first see a similar problem in $\mathbb{Z}^\ell$.

Example 10. *The problem is to find a big set $S \subset \mathbb{Z}^\ell$ such that for any multivariate polynomial $Q(Y_1, Y_2, \cdots, Y_\ell) \in \mathbb{Z}[Y_1, Y_2, \cdots, Y_\ell]$ with degree of each variable at most $h-1$, $Q(s_1, s_2, \cdots, s_\ell) = 0$ for $< h^\ell$ vectors $(s_1, s_2, \cdots, s_\ell) \in S$.*

Here is a well known solution to this problem: take

$$S = \{(\alpha, \alpha^h, \alpha^{h^2}, \cdots, \alpha^{h^{\ell-1}}) \,:\, \alpha \in \mathbb{Z}\}.$$

Then for every element of S on which Q vanishes, we find a root of the polynomial $R(Z) = Q(Z, Z^h, Z^{h^2}, \ldots, Z^{h^{\ell-1}})$, which is a nonzero polynomial of degree at most $(h-1)(1+h+h^2+\cdots+h^{\ell-1}) = h^\ell - 1$. Thus there can be at most $h^\ell - 1$ elements of S on which Q vanishes.

One drawback of this idea is that the vectors in S are very large, and we pay for it when we look at an analogous construction of $\mathcal{P}'$. (Try to see why!)

Another choice is

$$S = \{(\alpha, \alpha^h, \alpha^{h^2}, \cdots, \alpha^{h^{\ell-1}}) \mod p \mid 0 \le \alpha < p\}$$

for some large prime p (here we use $a \mod p$ to denote the integer in $\{0, \ldots, p-1\}$ congruent to a mod p). Why does this work? Take any Q. First, we may assume that the coefficients of Q have GCD = 1 (because otherwise we can divide out by the GCD, which does not affect zeroness of evaluations). Thus if we reduce the coefficients of Q mod p, we get a nonzero polynomial $\tilde{Q}(Y_1, \ldots, Y_\ell) \in \mathbb{F}_p[Y_1, \ldots, Y_\ell]$.

Let $\tilde{R}(Z) = \tilde{Q}(Z, Z^h, \ldots, Z^{h^{\ell-1}}) \in \mathbb{F}_p[Z]$, which is again a nonzero polynomial. Notice that any element of S on which Q is zero gives us a root α of $\tilde{R}(Z)$. Thus the number of such zeroes in S of Q is at most the degree of $\tilde{R}$, which is at most $h^\ell - 1$, as desired.

We now construct $\mathcal{P}'$ with Property Z. The construction is completely analogous to the above example:

1. Fix an irreducible polynomial $E(X) \in \mathbb{F}_q[X]$ of degree c.
2. Take $\mathcal{P}' := \{(f(X), f(X)^h, f(X)^{h^2}, \cdots, f(X)^{h^{\ell-1}}) \mod E(X) \,:\, f(X) \in \mathcal{P}\}$. (here we use $a(X) \mod E(X)$ to denote the unique polynomial of degree $< c$ congruent to $a(X)$ mod $E(X)$).

We then have:

Lemma 11. *Suppose $Q(X, Y_1, Y_2, \cdots, Y_\ell)$ is a multivariate polynomial of degrees $(A, h-1, h-1, \cdots, h-1)$, then*

$$\left|\{(f_1, f_2, \cdots, f_\ell) \in \mathcal{P}' \,:\, Q(X, f_1(X), f_2(X), \cdots, f_\ell(X)) \equiv 0\}\right| \le h^\ell - 1.$$

Proof. View Q as a polynomial in $Y_1, \ldots, Y_\ell$ with coefficients in $\mathbb{F}_q[X]$. If the coefficients have any common factors, we divide Q by them (this does not affect the zeroness of evaluations), and thus we may assume that all the coefficients have GCD 1. Consider the reduction $Q(X, Y_1, \ldots, Y_\ell)$ mod $E(x)$, and call it $\tilde{Q}(Y_1, Y_2, \cdots, Y_\ell) \in (\mathbb{F}_q[x]/E(x))[Y_1, Y_2, \cdots, Y_\ell]$. Thus $\tilde{Q}(Y_1, \ldots, Y_\ell)$ is a non-zero polynomial with coefficients in the field $\mathbb{F}_{q^c}$ (here we used the fact that $E(X)$ is irreducible) and its degree is at most $(h-1, h-1, \cdots, h-1)$. Define $\tilde{R}(Z) = \tilde{Q}(Z, Z^h, Z^{h^2}, \cdots, Z^{h^{\ell-1}}) \in \mathbb{F}_{q^c}[Z]$. The degree of the polynomial $\tilde{R}(Z)$ is at most $(h-1)(1+h+h^2+\cdots+h^{j-1}) = h^j - 1$, and so $\tilde{R}$ has at most $h^j - 1$ roots in $\mathbb{F}_{q^c}$.

Now the lemma follows from the observation that if $f(X)$ is such that $(f(X), f(X)^h, \ldots, f(X)^{h^{\ell-1}})$ mod $E(X)$ is a zero of $Q(X, Y_1, \ldots, Y_\ell)$, then $f(X) \mod E(X)$ is a zero of $\tilde{R}(Z)$. □

Unwinding Parameters. We just saw that C is a $(\log K, \frac{c\ell h}{q}, \log K + \lg q)$-condenser, where $K = h^\ell$. We now describe a setting of parameters that makes this compose nicely with the previously constructed extractor, thus giving our final extractor construction.

Let us choose $h = q^{0.99}$, $c = q^{0.001}$ and $\ell = q^{0.0001}$. Note that the error $\epsilon < \frac{1}{q^{\Omega(1)}}$.

Since $2^n = q^c$, we have that $q \leq n^{O(1)}$, and so the seed length $d = \log q = O(\log n)$. Thus the error $\epsilon < \frac{1}{\mathtt{poly}(n)}$.

The output length, m, equals $\log(q^{\ell+1}) = (\ell+1)\log q$. The input min-entropy, k, equals $\log K = \log h^\ell = (0.99) \cdot \ell \cdot \log q \geq 0.98m$. Thus the output of the condenser is a random variable distributed over $\{0,1\}^m$ with entropy $0.98m$, and this can be plugged into the earlier extractor we constructed, to yield a nearly-optimal extractor which works for all min-entropies.