1

We have seen how to decrease the amount of randomness required by randomized algs
- PRGs (assuming hard fns)
- Random walks on an expander to get error reduction.

AGENDA
1. Extractors
2. Extractors from hash fns
3. Extractors from expanders
4. Extractors from PRGs

Extractors allow us to decrease the quality of randomness.
(In the def of BPP, we assume iid uniform bits. Is that reasonable?)

An "EXTRACTOR" extracts the randomness from a low-quality random source.

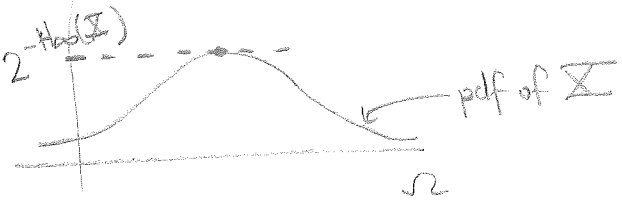
"weak" source $X \in \{0,1\}^n$ → [EXT] → something "close" to uniform

- "Close" means in total variation distance: for X, X' r.v.'s on Ω,

$$\Delta(X,X') := \frac{1}{2}\sum_{x\in\Omega} \left|\mathbb{P}(X=x) - \mathbb{P}(X'=x)\right| = \max_{A\subseteq\Omega} \left|\mathbb{P}(X\in A) - \mathbb{P}(X'\in A)\right|$$

- "Weak" means high min-entropy: For X r.v. on Ω,

$$H_\infty(X) := \max\{k : \mathbb{P}(X=x) \le 2^{-k} \ \forall x\in\Omega\}$$

picture: 

so high min-entropy means pretty flat.

- If $X \in \{0,1\}^n$ has $H_\infty(X) \ge k$, say X is a "(n,k)-source."

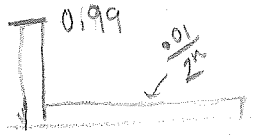
EXAMPLES

(1) k bits are iid uniform, remaining $n-k$ are fixed ("bit fixing")

(2) Conditioned on what's happened so far, next bit is pretty random:
$\forall i, \forall x_1,\dots,x_i$, $\mathbb{P}\{X_i=1 \mid X_j=x_j \ \forall j<i\} \in [\varepsilon, 1-\varepsilon]$
("Santha-Vazirani")

(3) Uniform on a set of size 2^k in $\{0,1\}^n$ ("k-flat")

Why min-entropy?
- Shannon entropy is a little silly, since



has $H(X) = \Omega(n)$.
- other notions do make sense, but let's go w/ this one.

2 It turns out that k-flat sources are as hard as any k-source:

FACT Any (n,k)-source is a convex combination of k-flat sources.

Aka, drawing from an (n,k)-src is the same as:
1) Choose k-flat src Y according to some dist
2) Draw from Y.

So if your extractor works on all k-flat Y, it will work on all k-src's.

Def For a class of sources $\mathcal{C}$, a (deterministic) $(\mathcal{C},\varepsilon)$-extractor

$$EXT: \{0,1\}^n \to \{0,1\}^m \quad \text{has}$$

$$\forall X \in \mathcal{C},\ \Delta(EXT(X), U_m) \le \varepsilon.$$

Observation $(\{k\text{-sources}\},\varepsilon)$-extractors don't exist. $\quad = 1$

pf. Say EXT was one, and consider X which is uniform on $\{x : EXT(x)_1 = 0\}$ $\quad = \frac{1}{2}$

SOLUTION: add a seed.

Def A (seeded) (k,ε)-extractor

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

$$\forall \text{ } k\text{-sources } X \in \{0,1\}^n,$$

$$\Delta(EXT(X, U_d), U_m) \le \varepsilon.$$

ASIDE: Actually, these do exist for bit-fixing sources

Now these DO exist, even w/ short seeds:

Thm $\forall\ k, n \in \mathbb{N}$, $\varepsilon > 0$, $\exists$ (k,ε)-extractor EXT w/

$$m = k + d - 2\log(1/\varepsilon) - O(1)$$

$$d = \log(n-k) + 2\log(1/\varepsilon) + O(1)$$

(and this is optimal).

← notice that something like $k+d$ is clearly the right answer, since that's how much randomness we put in

[3] Proof] Probabilistic method. Choose $EXT : \{0,1\}^{n+d} \to \{0,1\}^m$ randomly.

WTS $\forall$ k-flat X, $\forall A \subseteq \{0,1\}^m$, $\left| \mathbb{P}_{x,s}\{EXT(x,s) \in A\} - \frac{|A|}{2^m} \right| \le \varepsilon$

$$\frac{1}{2^{k+d}} \left| \sum_{x,s} \mathbb{1}_{\{EXT(x,s)\in A\}} - \mathop{\mathbb{E}}_{EXT} \mathbb{1}_{\{EXT(x,s)\in A\}} \right| \le \varepsilon \qquad (*)$$

These are 2^{k+d} indep, 0-1 r.v.'s, so use Chernoff.

By Chernoff, for fixed X, A,

$$\mathbb{P}(*) \le \exp\left(-\varepsilon^2 \cdot 2^{k+d}\right)$$

By the union bound over all $\binom{2^n}{2^k}$ X and 2^{2^m} A's, ($\binom{2^n}{2^k} \le 2^{n2^k}$)

$$\mathbb{P}_{EXT}\{ EXT \text{ is } (k,\varepsilon)\text{-extractor}\} \ge 1 - \binom{2^n}{2^k} 2^{(2^m)} \exp\left(-\varepsilon^2 \cdot 2^{k+d}\right)$$

$$\ge 1 - \exp\left(n \cdot 2^k + 2^m - \varepsilon^2 \cdot 2^{k+d}\right)$$

$$> 0 \quad \text{if } m, d \text{ as above.} \qquad \blacksquare$$

Note: this logarithmic seed len is pretty cool since we can exhaust over it (if, say, $k = \delta n$ so $m = k+d = \Theta(n)$). What's not cool is that this isn't explicit.

Some ways to get explicit extractors (for $\varepsilon = \Omega(1)$):

	seed len	output len	
Pairwise indep hash fns	$O(n)$	$k+d-O(1)$	we'll see these
Expanders	$O(n-k)$	$k+d$	
PRGs	$c\log(n)$	$k^{1/c}$	
List-decodable codes w/ etc.	$O(\log(n))$	$(1-\alpha)k$ for any α ← [GUV]	
optimal + existential	$\log(n-k)+O(1)$	$k+d$	
OPEN	$O(\log(n))$	$k+d-O(1)$	
	$\log(n)$	$\Omega(k)$	

4

Thm "Leftover hash lemma"

Let $\mathcal{H}$ be a pairwise indep hash family of fns $h: \{0,1\}^n \to \{0,1\}^m$, for $m = k - 2\lg(1/\varepsilon)$.

Then $EXT: (x, h) \longmapsto (h(x), h)$ is a $(k, \frac{\varepsilon}{2})$-extractor.

Recall, this means $\forall x \neq x'$, $\forall y, y'$ $\mathbb{P}\{h(x)=y, h(x')=y'\} = \frac{1}{2^{2m}}$ hint

NOTE This is actually stronger, since the seed itself is part of the output. This is called a "strong extractor."

(We can make $|\mathcal{H}| = 2^{n+m}$, so $d = n+m$.)

Proof For any r.v. Z, define

$$CP(Z) = \mathbb{P}_{z,z' \sim Z}\{z=z'\} = \sum_y \mathbb{P}(Z=y)^2 =: \|Z\|_2^2$$

"collision probability"; abusing notation

First,

$$CP((H(X),H)) = \underbrace{CP(H)\cdot CP(X)}_{h=h' \text{ and } x=x'} + \underbrace{CP(H)\,\mathbb{P}\{H(X)=H(X') \mid X \neq X'\}}_{h=h' \text{ but } x \neq x'}$$

$$\leq \frac{1}{2^d}\left(\frac{1}{2^k} + \frac{1}{2^m}\right) \leq (1+\varepsilon^2)\left(\frac{1}{2^{d+m}}\right) \text{ by choice of } m$$

$$= (1+\varepsilon^2)\, CP(\mathcal{U}_{d+m}).$$

Next, for any RV Z, if $CP(Z) \leq (1+\varepsilon^2) CP(\mathcal{U})$, then

$$(1+\varepsilon^2) CP(\mathcal{U}) \geq CP(Z) = \|Z\|_2^2 = \|\mathcal{U} + (Z-\mathcal{U})\|_2^2 = \|\mathcal{U}\|_2^2 + \|Z-\mathcal{U}\|_2^2$$

$$= CP(\mathcal{U}) + \|Z-\mathcal{U}\|_2^2$$

(these are $\perp$, since $\sum_z \mathbb{P}(Z=z)\cdot\mathbb{P}(\mathcal{U}=z) = 1\cdot 1 = 0$.)

$$\Rightarrow \|Z-\mathcal{U}\|_2^2 \leq \varepsilon^2 CP(\mathcal{U}).$$

Thus,

$$\|(H(X),H) - (\mathcal{U}_{d+m})\|_2^2 \leq \varepsilon^2/2^{d+m}$$

$$\Rightarrow \|(H(X),H) - \mathcal{U}_{d+m}\|_1 \leq \varepsilon \quad \text{by Cauchy-Schwarz}$$

$$\Rightarrow \Delta((H(X),H), \mathcal{U}_{d+m}) \leq \varepsilon/2 \qquad \blacksquare$$

Note Extracting w/r/t ℓ_2 norm, like we just did, is called a "Renyi Extractor." For these we need $d = \Omega(\min(n-k, m))$, so the seed len $m+n$ isn't so bad, comparatively.

Extractors from expanders.

Let G be an expander on 2^m vertices of degree 2^d, w/ param. $\lambda = \lambda(G)$.

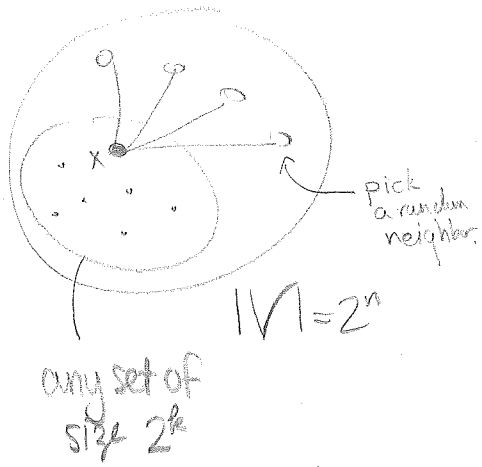
Consider $EXT_G : (x, s) \longmapsto s^{th}$ neighbor of x in G

$x \in V$, $s \in [2^d]$

What does it mean for this to be a good extractor?

$\forall$ k-flat X, $\forall A \subseteq V$,

$$\left| \Pr_{s}[EXT_G(X, U_d) \in A] - |A|/2^n \right| \leq \varepsilon$$



aka

$\forall$ $X \subseteq V$, $|X| = 2^k$, $\forall A$,

$$\left| E(X, A)\left(\frac{1}{|X| \cdot 2^d}\right) - \frac{|A|}{2^n} \right| \leq \varepsilon$$

or

$$\left| E(X, A) - |X||A|\left(\frac{2^d}{2^n}\right) \right| \leq \varepsilon |X| \cdot 2^d = \varepsilon \cdot 2^{k+d}$$

Recall EML says, $\forall X, A \subseteq V$,

$$\left| E(X, A) - |X||A|\left(\frac{2^d}{2^n}\right) \right| \leq \lambda 2^d \sqrt{|A||X|} \leq \lambda \cdot 2^{d + n/2 + k/2}$$

so, Thm EXT_G is a $\left(k, \frac{\lambda(G)}{2^{k/2 + n/2}}\right)$-extractor.

If $\lambda(G) = 2^{-d/2}$ (aka, G is Ramanujan), this gives the parameters in the table.

CATCH. We don't know how to make Ramanujan graphs for all (k, d), but you can get the same parameters by starting w/ constant $\lambda(G) = 1/2$ and taking a random walk from x instead of just picking a random nbr.

6 Extractors from PRGs ("Trevisan's extractor") [SKETCH]

Recall from hardness v. randomness, we saw

hard fn f ⟶ {ECC's, XoR lemma, N-W PRG} ⟶ PRG,

and the proof actually gave us:

f ⟶ [proof 1] ⟶ G^f,

and

small circuit distinguishing $G^f(U)$ from U ⟶ [proof 2] ⟶ small circuit that computes f

On inspection, what we ACTUALLY got was, for any "stretch" $k^{1/c}$

$f: \{0,1\}^{\log(n)} \to \{0,1\}$ ⟶ [proof 1] ⟶ $G^f: \{0,1\}^{c\log(n)} \to \{0,1\}^{k^{1/c}}$

(think of $f \in \{0,1\}^n$)

and

ANY fn D distinguishing $G^f(U)$ from U ⟶ [proof 2] ⟶ "small circuit w/ oracle access to D" that computes f (← time poly(n,k))

advice $a \in \{0,1\}^{k^{1/4}}$

x ⟶ [R] ⟶ $f(x)$

R ⟷ D

CLAIM

EXT: $(f,z) \longmapsto G^f(z)$

is a $(k, 1/5)$-extractor

output len is $m = k^{1/c}$

seed len is $d = c\log(n)$

7 "Proof." (sketch)

Suppose not. Then $\exists$ k-flat dist F on fns $f: \{0,1\}^{\log(n)} \to \{0,1\}$ st

$$\Delta\left(EXT(F, U_d), U_m\right) > 1/5$$

$\Leftrightarrow$

$$\exists \text{ some fn } D, \quad \left| \mathbb{P}\{D(G^F(U_d)) = 1\} - \mathbb{P}\{D(U_m) = 1\} \right| > 1/5$$

(not nec. a small circuit!)

So then for $\geq 1/10$ of the f's $\in supp(F)$,

$$\left| \mathbb{P}\{D(G^f(U_d)) = 1\} - \mathbb{P}\{D(U_m) = 1\} \right| > 1/10.$$

Call those f's "bad", so

$$\mathbb{P}_{f \sim F}\{f \text{ is bad}\} \geq 1/10.$$

But $(\#\text{bad } f\text{'s}) \leq \left(\begin{array}{c}\#\text{advice strings } a \\ \text{that our alg} \\ R \text{ takes}\end{array}\right) \leq 2^{k^{1/4}}$, so

$$\mathbb{P}_{f \sim F}\{f \text{ is bad}\} \leq \frac{1}{2^k} \cdot 2^{k^{1/4}} \ll 1/10 \quad \lightning$$

so then EXT is indeed an extractor. ∎