Lecture 22: Basic Applications of Fourier Analysis (Extractors and Leftover Hash Lemma)
A probability distribution $X$ has min-entropy at least $k$ if $\Pr[X = x] \leq 2^{-k}$, for all $x$ in the sample space.

**Definition ((n,k)-Source)**

A source over sample space $\{0, 1\}^n$ with min-entropy at least $k$ is known as an $(n, k)$-source.

- There are other specialized imperfect randomness sources like, bit-fixing sources, Santha-Vazirani sources.
- Goal: Design an extractor to extract pure randomness from any min-entropy source from a class of sources.
- For example, design an extractor that extracts pure randomness from any $(n, k)$-source.
Definition \(((C_n, m, \varepsilon)\text{-Extractor})\)

Let \(C_n\) be a class of imperfect randomness sources over the sample space \(\{0, 1\}^n\). A \((C_n, m, \varepsilon)\)-extractor is a function 
\[
\text{Ext}: \{0, 1\}^n \rightarrow \{0, 1\}^m
\]
such that, for all \(X \in C_n\), we have 
\[
\text{SD}(\text{Ext}(X), U_m) \leq \varepsilon.
\]

Such a function is also known as a deterministic extractor.
Lemma (Negative Result)

Let $C_n$ be the set of all $(n, n-1)$-sources. For any $\varepsilon < 1/2$, there does not exist a $(C_n, 1, \varepsilon)$-extractor.

This result is extremely strong. Even if the sources have $(n-1)$ min-entropy, we cannot extract even one bit that is close to uniform!

If possible let there exists such an extractor $\text{Ext}$

Let $P_b = \text{Ext}^{-1}(b)$, for $b \in \{0, 1\}$

Note that at least one of $P_0$ or $P_1$ is of size $2^{n-1}$. Suppose $|P_{b^*}| \geq 2^{n-1}$

Let $X$ be the uniform distribution over the set $P_{b^*}$, represented by $U(P_{b^*})$, and $\Pr[X = x] \leq 2^{-(n-1)}$, for all $x \in \{0, 1\}^n$

Note that $\text{SD}(\text{Ext}(X), U_1) = 1/2$
Note that computing the distribution \( U(P_b^*) \) might be computationally inefficient. What if we restrict to distributions that are easy (or, efficient) to sample?

**Lemma (Efficient Negative Result)**

Let \( C_n \) be the sources that are samplable in time \( T \) (given uniform random bits as input) and have min-entropy at least \( k = (n - 1) - \log(3/2) \). Then, for all \( \epsilon < 1/4 \) there does not exist any \((C_n, 1, \epsilon)\)-extractor that has time complexity \( T' \), such that \( T' \leq T - 2n - \Theta(1) \).

Let \( P_b \) be the distribution that takes as input two uniform random strings \((r, r') \in \{0,1\}^{2n}\). If \( \text{Ext}(r) = b \), output \( r \); otherwise output \( r' \).
This technique is known as rejection sampling, i.e. “keep rejecting the samples till you get something you desire, or (after a threshold number of sample draws) give up and output the final sample.”

The time complexity $T$ to sample $P_b$ is $T' + 2n + \Theta(1)$, hence the bound $T \geq T' + 2n + \Theta(1)$ is satisfied.

Let $p_b$ be the probability of $\text{Ext}(U_n) = b$.

Then, we have

\[
\Pr[\text{Ext}(P_b) = b] = p_b \cdot 1 + (1 - p_b) \cdot p_b = p_b(2 - p_b)
\]

and, similarly, \[
\Pr[\text{Ext}(P_b) = \overline{b}] = p_b(2 - p_b)
\]

Maximum of these two probabilities is at least $3/4$.

So, the statistical distance from $U_1$ of one of these two distributions is at least $1/4$.

That distribution will have maximum probability $2^{-k} \leq 2^{-(n-1)} + 2^{-n}$, and this satisfies the min-entropy bound.
Take-away Message

- It is still possible to have the complexity of the extractor to be significantly larger than the sampling complexity of the sources.
- There are positive results where good deterministic extractors exist when the class of sources are simple, for example, bit-fixing sources, affine sources, sources samplable by small-depth circuits.
- In the computational setting, if hard to invert functions exist then we can construct an efficient extractor for sources samplable in time $p(n)$, where $p(\cdot)$ is a fixed polynomial.
- A more general version of the above statement is considered by Nisan-Wigderson.
Seeded Extractors

- These extractors take as inputs a uniform random string $s \sim U_d$ known as the seed.
- Goal: Given this initial investment of pure $d$ bits, we are interested in obtaining $m$ pure random bits as output from $k$ imperfect bits. We want $m \approx n + d$ and $d$ to be as small as possible.

**Definition (Strong Extractor)**

A $(C_n, d, m, \varepsilon)$-strong-extractor $\text{Ext}: \{0, 1\}^n \times \{0, 1\}^d \rightarrow \{0, 1\}^m$ is a function such that, for any $X \in C_n$, we have:

$$\text{SD} \left( (U_d, \text{Ext}(X, U_d)), (U_d, U_m) \right) \leq \varepsilon$$

- For $C_n = (n, k)$-sources, our aim is to get $m \approx k$ and $d$ as small as possible.
2-Universal Hash Function Family

- Let $\mathcal{F}_{n,m}$ be the set of all function $f : \{0, 1\}^n \rightarrow \{0, 1\}^m$
- $H$ is a distribution over the sample space $\mathcal{F}_{n,m}$

**Definition (2-Universal Hash Function Family)**

For every distinct $x_1, x_2 \in \{0, 1\}^n$, we have:

$$\Pr_{h \sim H}[h(x_1) = h(x_2)] \leq \frac{1}{2^m}$$

- We want that the sampling $h \sim H$ can be efficiently performed by a randomized algorithm that takes a sample from $U_d$
- Intuitively, two separate inputs collide under $h$ at the same probability that they collide under a random function from $\mathcal{F}_{n,m}$
Theorem (LHL)

Let $H$ be a 2-universal Hash Function Family. For any $X$ that is an $(n, k)$-source, the following is true:

$$\text{SD} ((H, H(X)), (H, U_m)) \leq \varepsilon,$$

where $2\varepsilon = \sqrt{2^{-(k-m)} - 2^{-k}}$

- That is, $H$ is a good extractor for $(n, k)$-sources
- So, we need to construct the family $H$ that can be sampled using only $d$-bits of randomness, and we want $d$ to be as small as possible
- Note about the proof: We will see a more general Fourier-based proof, because there is another result, namely “Lopsided-LHL,” that (as far as I know) cannot be proven using elementary combinatorial techniques
Proof

- We will use $M = 2^m$ and $K = 2^k$
- We bound the SD as follows:

$$
2SD \left( (H, H(X)), (H, U_m) \right) = \mathbb{E}_{h \sim H} \left[ 2SD (h(X), U_m) \right] = \mathbb{E}_{h \sim H} \left[ \sum_{y \in \{0,1\}^m} \left| h(X)(y) - U_m(y) \right| \right] \\
\leq \mathbb{E}_{h \sim H} \left[ M^{1/2} \left( \sum_{y \in \{0,1\}^m} (h(X)(y) - U_m(y))^2 \right)^{1/2} \right], \quad \text{Cauchy-Schwarz}
$$

$$
= M \mathbb{E}_{h \sim H} \left[ \sqrt{\| h(X) - U_m \|_2^2} \right] \\
\leq M \sqrt{\mathbb{E}_{h \sim H} \left[ \| h(X) - U_m \|_2^2 \right]}, \quad \text{Jensen's}
$$
Proof

- Let us upper bound \( \| h(X) - U_m \|_2^2 \)

\[
\| h(X) - U_m \|_2^2 \\
= \sum_{S \subseteq [m]} (h(X) - U_m)(S)^2, \\
\text{Parseval's}
\]

\[
= \sum_{S \subseteq [m]: S \neq \emptyset} \hat{h}(X)(S)^2 \\
= \sum_{S \subseteq [m]} \hat{h}(X)(S)^2 - \hat{h}(X)(S = \emptyset)^2 \\
= \| h(X) \|_2^2 - 1/M^2
\]

- So, we have the bound:

\[
2SD \left( (H, H(X)), (H, U_m) \right) \leq M \sqrt{\mathbb{E}_{h \sim H} \left[ \| h(X) \|_2^2 - M^{-2} \right]}
\]
Proof

So, it suffices to upper bound $\mathbb{E}_{h \sim H} \left[ \| h(X) \|_2^2 \right]$

$$= \mathbb{E}_{h \sim H} \left[ \| h(X) \|_2^2 \right]$$

$$= \mathbb{E}_{h \sim H} \mathbb{E}_{y \sim U_m} \left[ h(X)(y)^2 \right]$$

$$= \mathbb{E}_{h \sim H} \mathbb{E}_{y \sim U_m} \left[ \Pr[h(X^{(1)}) = y \land h(X^{(2)}) = y] \right]$$

$$= \mathbb{E}_{h \sim H} \mathbb{E}_{y \sim U_m} \left[ \Pr[X^{(1)} = X^{(2)}] \Pr[h(X^{(1)}) = h(X^{(2)}) = y | X^{(1)} = X^{(2)}] \right]$$

$$+ \mathbb{E}_{h \sim H} \mathbb{E}_{y \sim U_m} \left[ \Pr[X^{(1)} \neq X^{(2)}] \Pr[h(X^{(1)}) = h(X^{(2)}) = y | X^{(1)} \neq X^{(2)}] \right]$$
Proof

The first term:

\[
\Pr[X^{(1)} = X^{(2)}] \sum_{y \in \{0, 1\}^m} \frac{1}{M} \Pr[h(X^{(1)}) = h(X^{(2)}) = y | X^{(1)} = X^{(2)}]
\]

\[
\leq \frac{1}{M} \cdot \Pr[X^{(1)} = X^{(2)}]
\]
Second Term:

\[
\frac{1}{M} \cdot \Pr[X^{(1)} \neq X^{(2)}] \sum_{h \sim H} \Pr[h(X^{(1)}) = h(X^{(2)}) | X^{(1)} \neq X^{(2)}] \\
\leq \frac{1}{M^2} \Pr[X^{(1)} \neq X^{(2)}] \\
= \frac{1}{M^2}(1 - \Pr[X^{(1)} = X^{(2)}])
\]
So, we have:

\[ E_{h \sim H} \left[ \| h(X) \|_2^2 \right] - \frac{1}{M^2} \]

\[ \leq \Pr[X^{(1)} = X^{(2)}] \left( \frac{1}{M} - \frac{1}{M^2} \right) \]

\[ \leq \frac{1}{K} \left( \frac{1}{M} - \frac{1}{M^2} \right) \]

So, overall we have:

\[ 2\text{SD}((H, H(X)), (H, U_m)) \leq \sqrt{\frac{M}{K} - \frac{1}{K}} \]

Hence the result.