Lecture 18: October 25, 2022

Lecturer: Eshan Chattopadhyay *Scribe: Jenny Chen, Tomas Alvarez*

1 Introduction

Recall the following claim from the previous lecture:

Definition 1.1 (Sampler)**.** Samp : $\{0,1\}^n \to [M]^D$ is a (k, ϵ, δ)-sampler if for all functions $f : [M] \to [0,1]$, and for all (n, k)-sources X,

$$\Pr\left[\left|\frac{1}{D}\sum_{i=1}^{D} f(y_i) - \mu(f)\right| > \epsilon\right] < \delta$$

where $(y_1, \cdots, y_D) = Samp(x)$ for $x \sim X$.

We propose a construction of a (k, ϵ, δ)-sampler based on extractors. We start with a (k', ϵ') extractor Ext: $[N] \times [D] \to [M]$ where the constants k' and ϵ' remain to be determined. We say that $G_{\text{Ext}} = ([N] \cup [M], E)$ is the bipartite graph on $[N] \cup [M]$ with edges $e = (x, z) \in E$ if $\exists y$ such that $\text{Ext}(x, y) = z$. We refer to the neighbors of x as $N(x)$ and the proposed construction is $\text{Samp}(x) = N(x)$.

We make use of the following two sets

$$Bad^+ = \{x \in \{0,1\}^n : \frac{1}{D}\sum_{y \in N(x)} f(y) - \mu(f) > \epsilon\}$$

$$Bad^- = \{x \in \{0,1\}^n : \frac{1}{D}\sum_{y \in N(x)} f(y) - \mu(f) < -\epsilon\}$$

Claim 1.2. $|Bad^+|, |Bad^-| < 2^{k'}$.

Proof. The proof is the same for both sets so we only prove it for Bad^+. Suppose for contradiction that $|Bad^+| \geq 2^{k'}$. Let X^+ be a flat distribution on Bad^+. There are at least $2^{k'}$ elements and X^+ is flat so $H_\infty \geq k'$. Since we have $H_\infty \geq k'$. and an (k', ϵ') extractor, then $\text{Ext}(X^+, U_d) \approx_{\epsilon'} U_m$. Let us now denote $\text{Ext}(X^+, U_d)$ by z^+. Because X^+ is the set of x's such that the sampled mean is larger than the true mean by ϵ we know $\mathbb{E}[f(z^+))] - \mathbb{E}[f(U_m)] = \mathbb{E}[f(z^+))] - \mu(f) > \epsilon$. Since $z^+ \approx_{\epsilon'} U_m$ we use the following fact from last lecture: $|\mathbb{E}[f(z^+))] - \mu(f)| < 2\epsilon'$. If we choose $\epsilon' = \epsilon/2$ then we have the inequality $\mathbb{E}[f(z^+))] - \mu(f) > \epsilon$ and $|\mathbb{E}[f(z^+))] - \mu(f)| < 2\epsilon' = \epsilon$ which is a contradiction. Thus, $|Bad^+|, |Bad^-| < 2^{k'}$. □

Notice that if $x \in Bad^+$ or $x \in Bad^-$ then $\left|\frac{1}{D}\sum_{i=1}^{D} f(y_i) - \mu(f)\right| > \epsilon$ by definition. Thus, $\Pr\left[\left|\frac{1}{D}\sum_{i=1}^{D} f(y_i) - \mu(f)\right| > \epsilon\right] = \Pr\left[x \in Bad^+ \cup Bad^-\right] \leq \frac{2 \cdot 2^{k'}}{2^k}$. Therefore, if we let $k' = k - \log(1/\delta) - 1$ we get $\frac{2 \cdot 2^{k'}}{2^k} < \delta$ which implies that this is a (k, ϵ, δ)-sampler.

2 Construction of Seeded Extractors

Recall the existential bound of a (strong) seeded extractor Ext: $[N] \times [D] \to [M]$, which is a (k, ϵ) extractor:

- $m = k - 2log(\frac{1}{\epsilon}) - O(1)$
- $d = log(n-k) + 2log(\frac{1}{\epsilon}) + O(1)$

This is the parameter we can reach with a random seeded extractor. We're going to show an explicit construction that uses $O(n)$ seed length but can extract a good amount of randomness from the weak source.

Construction 2.1. *Take a universal hash family $\mathcal{H} = \{h : [N] \to [M]\}$ of size D. Recall that the hash functions satisfy the following property: $Pr_{h\sim\mathcal{H}}[h(x) = h(y)] \leq \frac{1}{M}, \forall x \neq y$. Define the extractor as $Ext(x, h) = h(x)$.*

In other words, the extractor gets a seed and use it to pick a hash function. Then it gets a sample from the weak source and apply the hash function to the sample. Since the number of random bits we need to sample such functions is at least n, $d = O(n)$ here.

Before we prove the construction gives a valid (k, ϵ) extractor, we need to talk about the collision probability first.

Definition 2.2. (Collision Probability) Let Y be a distribution on a set T such that $|T| = A$. $CP(Y) = Pr[Y = Y']$, such that Y' is an independent copy of Y. $CP(Y) = Pr[Y = Y'] = \sum_{y\in T} Pr[Y = y]^2 = ||Y||_2^2$.

$CP(Y)$ equals to the L2 norm of Y, and if we have a uniform distribution on T, then $CP(U_T) = \frac{1}{A}$ (since each $Pr[Y = y] = \frac{1}{A^2}$ and there are a total of A such y's).

Claim 2.3. *If $CP(Y) \leq \frac{1}{A}(1 + \epsilon)$, then $|Y - U_m| \leq \frac{1}{2}\sqrt{\epsilon}$ (the statistical distance).*

Proof. From Cauchy-Schwarz inequality, we know $\forall u, v \in \mathbb{R}^n, < u, v > \leq ||u||_2||v||_2$. If we pick the $u = \vec{\mathbb{1}}$ and v be the difference between Y and U_m. Then $||u||_2||v||_2 = \sqrt{A} \cdot ||Y - U_m||_2$. Plug into the inequality we obtain $||Y - U_m||_1 \leq \sqrt{A} \cdot ||Y - U_m||_2$ (1).

We also know that $Y = U_m + (Y - U_m)$. And we claim that $< U_m, Y - U_m >= 0$. This inner product equals to the sum of all entries of vector $Y - U_m$, which is equivalent to $\sum_{y\in T} Pr[Y = y] - \sum_{y\in T}[U_m = y]$. Since the sum of the probability of all points in the distribution is simply 1, $\sum_{y\in T} Pr[Y = y] - \sum_{y\in T}[U_m = y] = 1 - 1 = 0$. So we know U_m and $Y - U_m$ are orthogonal to each other. Using Pythagorean theorem, $||Y||_2^2 = ||U_m||_2^2 + ||Y - U_m||_2^2$ (2).

Square both sides of (1) and plug in (2), we get $||Y - U_m|| \leq A(||Y||_2^2 - ||U_m||_2^2) = A(CP(Y) - CP(U_m)) = A(\frac{1}{A}(1 + \epsilon) - \frac{1}{A}) = \epsilon \Rightarrow ||Y - U_m|| \leq \sqrt{\epsilon}$. By definition, the statistical distance is half of the L1 norm, thus $|Y - U_m| \leq \frac{1}{2}\sqrt{\epsilon}$. □

Now we can prove the Leftover Hash Lemma.

Theorem 2.4. *(Leftover Hash Lemma). If $\mathcal{H} = \{h : \{0, 1\}^n \to \{0, 1\}^m\}$ is a pairwise independent family of hash functions, then $Ext(x, h) = h(x)$ is a strong (k, ϵ)-extractor for any (n, k)-source x.*

Proof. Let X be an arbitrary k-source. Essentially, we want to show that $\mathcal{H}(X), \mathcal{H} \simeq_\epsilon U_m, \mathcal{H}$.

$$\begin{aligned}
CP(\mathcal{H}(X), \mathcal{H}) &= Pr[(H(X), H) = (H'(X'), H')] && (1)\\
&= CP(H)((Pr_{h\sim H}[h(X) = h(X')])) = \frac{1}{D}(Pr_{h\sim H}[h(X) = h(X')]) && (2)\\
&= \frac{1}{D}(Pr[X = X'] + Pr_{h\sim H}[h(X) = h(Y)|X \neq Y]) && (3)\\
&= \frac{1}{D}(\frac{1}{k} + Pr_{h\sim H}[h(X) = h(Y)|X \neq Y]) && (4)\\
&= \frac{1}{D}(\frac{1}{k} + \frac{1}{M}) && (5)\\
&= \frac{1}{MD}(1 + \frac{M}{K}) \Rightarrow \epsilon' = \frac{M}{K} && (6)\\
&\Rightarrow \epsilon = 2^{\frac{m-k}{2}-1} && (7)
\end{aligned}$$

Line (1) comes from the definition of the collision probability.

In order for $(H(x), H) = (H'(x), H')$ to happen, we need $H = H'$. Since there are D hash functions, we get line (2).

For line (3) and (4), if we fix the h, then there are two cases for $h(X) = h(X')$: either $X = X'$ or $X \neq X'$ but $h(X) = h(X')$. The probability of $X = X'$ is just the collision probability of X. We know that X is an (n, k)-source, so $CP(X) \leq \frac{1}{K}$ since $H_\infty(X) \geq k$.

Line (5) comes from the definition of the hash function: $Pr_{h\sim H}[h(X) = h(Y)|X \neq Y] \leq \frac{1}{M}$. □

From those, it follows that $m = k - 2log(1/\epsilon) + 1$. Not that we used a very large seed to achieve that. Since we need to enumerate over all seeds which has a total of 2^d such seeds, we really want a seed length within $O(logn)$.