The Leftover Hash Lemma and ϵ-HSG

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

1 A family of hash functions for Nisan's generator

Recall that for Nisan's generator we claimed the existence of a 2UFOHF $\mathcal{H}$ of functions $h : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$. We now present such a family.

We start by defining the collision probability of a probability distribution:

Definition 1. *Let P be a probability distribution over n elements and denote by $P(i) = \Pr_{x \sim P}[x = i]$. We will sometimes also consider P as a vector of length n where the ith entry $P_i = P(i)$. The* collision probability *of P is* $\mathrm{col}(P) = \sum_i P(i)^2 = ||P||_2^2$

The intuition behind the definition is that $\mathrm{col}(P)$ measures the likelihood of two independent samples from P colliding.

Observation 2. *Let P be any arbitrary distribution over a set X where $|X| = n$ and U be the uniform distribution over said set. We observe that:*

$$||P - U||_2^2 = \langle P - U, P - U \rangle = \langle P, P \rangle - 2 \langle P, U \rangle + \langle U, U \rangle \tag{1}$$

$$= \mathrm{col}(P) - 2 \sum_i P_i \cdot \frac{1}{n} + \mathrm{col}(U) = \mathrm{col}(P) - \frac{2}{n} \sum_i P_i + \frac{1}{n} \tag{2}$$

$$= \mathrm{col}(P) - 2\,\mathrm{col}(U) + \mathrm{col}(U) = \mathrm{col}(P) - \mathrm{col}(U) \tag{3}$$

Observation 3. *By Cauchy-Schwartz and Observation 2:*

$$|P - U|_1^2 \leqslant n|P - U|_2^2 = n \cdot (\mathrm{col}(P) - \mathrm{col}(U))$$

1.1 The construction

We assume wlog that n is a prime power. We will work over the field $\mathbb{F} = \mathbb{F}_n$ of n elements. We define $\mathcal{H} = \{h_{a,b} : a, b \in \mathbb{F}\}$ where $h_{a,b}(x) = ax + b$ restricted to its first m bits.

Claim 4. *$\mathcal{H}$ is a* 2UFOHF

Proof. Let $M = 2^m \leqslant N = 2^n$ and $x_1 \neq x_2$. A simple computation shows that:

$$\Pr_{h \in \mathcal{H}} [h(x_1) = \sigma_1 \wedge h(x_2) = \sigma_2] = \Pr_{a,b \in \mathbb{F}} [h(x_1) = \sigma_1 \wedge h(x_2) = \sigma_2] \tag{4}$$

$$= \Pr_{a,b \in \mathbb{F}} \left[\begin{pmatrix} x_1 & 1 \\ x_2 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sigma_1 \\ \sigma_2 \end{pmatrix} \right] \tag{5}$$

$$= \frac{1}{M^2} \tag{6}$$

As $\begin{pmatrix} x_1 & 1 \\ x_2 & 1 \end{pmatrix}$ is full rank and we are working over a field $\square$

Next, we define our extractor $E : \{0,1\}^n \times \mathcal{H} \to \{0,1\}^m$. We note that $d = \log|\mathcal{H}| = \log|\mathbb{F}|^2 = 2n$ and we claim:

Claim 5 (The Leftover Hash Lemma, [ILL89])**.** *The extractor $E : \{0,1\}^n \times \mathcal{H} \to \{0,1\}^m$ defined by $E(x,h) = h(x)$ is a (k,ϵ)-strong extractor for $d = 2n, k = m + 2\log\frac{1}{\epsilon}$*

Proof. We compute the collision probability of $U_d \circ E(X, U_d)$ where X is a flat k-source over $K = 2^k$ elements:

$$\operatorname{col}(U_d \circ E(X,U_d)) = \Pr_{h_1,h_2,x_1,x_2}[(h_1,h_1(x_1)) = (h_2,h_2(x_2))] \tag{7}$$

$$= \Pr_{h_1,h_2}[h_1 = h_2] \Pr_{x_1,x_2}[h_1(x_1) = h_2(x_2) \mid h_1 = h_2] \tag{8}$$

$$\leqslant \operatorname{col}(\mathcal{H}) \left[\Pr_{x_1,x_2}[x_1 = x_2] + \Pr_{x_1,x_2,h}[h(x_1) = h(x_2) \mid x_1 \neq x_2]\right] \tag{9}$$

$$= \frac{1}{|\mathcal{H}|}\left[\operatorname{col}(X) + \frac{1}{M}\right] \tag{10}$$

$$= \frac{1}{|\mathcal{H}|} \cdot \frac{1}{M}\left[1 + \frac{M}{K}\right] \tag{11}$$

$$= \operatorname{col}(U_{\mathcal{H}} \times U_m) \cdot (1 + \epsilon^2) \tag{12}$$

where $\Pr_{x_1,x_2,h}[h(x_1) = h(x_2) \mid x_1 \neq x_2] = \frac{1}{M}$ since $\mathcal{H}$ is a 2UFOHF.

Therefore, by Observation 2 we have:

$$||(U_d \circ E(X,U_d)) - U_d \times U_m||_2^2 = \operatorname{col}(U_d \circ E(X,U_d)) - \operatorname{col}(U_d \times U_m) \tag{13}$$

$$\leqslant \operatorname{col}(U_d \times U_m)(1+\epsilon^2) - \operatorname{col}(U_d \times U_m) \tag{14}$$

$$= \epsilon^2 \frac{1}{|\mathcal{H}| \cdot M} \tag{15}$$

And thus by Observation 3 $|U_d \circ E(X,U_d) - U_d \times U_m|_1 \leqslant \sqrt{|\mathcal{H}|M\frac{\epsilon^2}{|\mathcal{H}|M}} = \epsilon$ □

We note that while the seed length is long $d = 2n$, the entropy loss $k - m = 2\log\frac{1}{\epsilon}$ is optimal. For Nisan's generator, setting the input length as $\ell = O(\log n)$ yields a family of hash functions $h : \{0,1\}^\ell \to \{0,1\}^\ell$ where h is described using $2\ell = O(\log n)$ bits.

2 ϵ-Hitting Set Generators

Nisan's generator ϵ-fools $[W,T]$ branching programs using a seed of length $O(\log T \cdot \log\frac{TW}{\epsilon})$. We now present a result which "liberates" the $\log\frac{1}{\epsilon}$ factor. The result is a modification of the construction in [HZ18]. Our construction will use Nisan's generator whereas the original work gives a more general construction using arbitrary space bounded PRGs, combined with dispersers.

Before we start, we need to define hitting set generators:

Definition 6. *A function $G : \{0,1\}^\ell \to \{0,1\}^r$ is an ϵ-Hitting Set Generator (ϵ-HSG) for $[W, r]$ branching programs if for any such branching program with machine M we have*

$$\Pr_{y \in U_r}[M(y) = 1] > \epsilon \implies \exists z \in U_\ell : M(G(z)) = 1$$

An ϵ-HSG is a useful tool for derandomizing probabilistic classes of one-sided errors.

Theorem 7 ([HZ18])**.** *For any W, T, ϵ there exists an ϵ-HSG for $[W, T]_{\{0,1\}}$ branching programs with seed length $\ell = O(\log T \cdot \log WT) + O(\log \frac{1}{\epsilon})$*

Let $n = \max\{W, T\}$ and think of $\epsilon \ll \frac{1}{n}$, e.g. $\epsilon \approx n^{-\log n}$. We begin with a useful claim:

Claim 8. *Let v be a vertex in M's branching program such that $\Pr[v \rightsquigarrow v_{acc}] = \alpha$. Denote*

$$\Gamma_v = \left\{ w \mid \alpha n \geqslant \Pr[w \rightsquigarrow v_{acc}] \geqslant \frac{\alpha n}{2} \right\}$$

and let M_j denote the jth layer of M's branching program. Then for $\Gamma_{v,j} = \Gamma_v \cap M_j$ there exists a j such that $\Pr[v \rightsquigarrow \Gamma_{v,j}] \geqslant \frac{1}{n^2}$

Proof. We first note that on any path $v \rightsquigarrow v_{acc}$ there exists a $w \in \Gamma_v$. To see this, fix a path, let $u_1 \to u_2$ be adjacent vertices on the path and denote by u_3 the other outneighbor of u_1. As $\Pr[u_1 \rightsquigarrow v_{acc}] = \frac{1}{2}\Pr[u_2 \rightsquigarrow v_{acc}] + \frac{1}{2}\Pr[u_3 \rightsquigarrow v_{acc}]$ clearly $\Pr[u_2 \rightsquigarrow v_{acc}] \leqslant 2\Pr[u_1 \rightsquigarrow v_{acc}]$. Additionally, the path finishes at v_{acc} where obviously $\Pr[v_{acc} \rightsquigarrow v_{acc}] = 1$. As the probability of acceptance grows by at most 2 at each stage and eventually reaches 1, clearly we have a w whose acceptance probability is in the given interval.

Now, assume towards contradiction that for any j we have $\Pr[v \rightsquigarrow \Gamma_{v,j}] < \frac{1}{n^2}$ and note that

$$\Pr[v \rightsquigarrow v_{acc} \text{ via } \Gamma_{v,j}] = \Pr[v \rightsquigarrow \Gamma_{v,j}] \cdot \Pr[v \rightsquigarrow v_{acc} \mid v \rightsquigarrow \Gamma_{v,j}] < \frac{1}{n^2} \cdot \alpha n = \frac{\alpha}{n}$$

by the definition of $\Gamma_{v,j}$. On the other hand, as v must pass thru some $\Gamma_{v,j}$:

$$\begin{aligned} \alpha &= \Pr[v \rightsquigarrow v_{acc}] && (16)\\ &= \Pr[v \rightsquigarrow v_{acc} \text{ thru some } \Gamma_{v,j}] && (17)\\ &\leqslant \sum_j \Pr[v \rightsquigarrow v_{acc} \text{ via } \Gamma_{v,j}] && (18)\\ &< n \cdot \frac{\alpha}{n} = \alpha && (19) \end{aligned}$$

in contradiction ∎

Now, for a computation with acceptance probability ϵ (i.e. $\Pr[v_{init} \rightsquigarrow v_{acc}] = \epsilon$) fix a set of vertices $v_{init} = v_0, v_1, \ldots, v_k$ and a set of layers $\overline{\ell} = \ell_0 = 0, \ell_1, \ldots, \ell_k$ where $k = \log_n \frac{1}{\epsilon}$ such that $v_i \in \Gamma_{v_{i-1},\ell_i}$ where $\Pr[v_i \rightsquigarrow \Gamma_{v_{i-1},\ell_i}] \geqslant \frac{1}{n^2}$ (such vertices and layers exist by Claim 8), and note that by definition $\Pr[v_k \rightsquigarrow v_{acc}] \geqslant n^k \epsilon = 1$. We now show that we can construct a HSG for this path.

For any choice of a vertex v at layer i in the branching program and any layer $j > i$ we can define a new branching program $B_{v,j}$ such that $\Pr[B_{v,j} = 1] = \Pr[v \rightsquigarrow \Gamma_{v,j} \text{ in } M]$, this gives us a total of $WT^2 \leqslant n^3$ branching programs. Let $\mathcal{B} = \{B_{v,j} : v \in M, j \in [T]\}$ be the set of these BPs.

To construct our HSG, we first record a theorem which encapsulates what we will require from Nisan's generator:

Theorem 9. *Let M be a $[W,T]$ branching program, $\alpha > 0$ and let $h_1, \dots, h_{\log T} \in \mathcal{H}$ where $\mathcal{H}$ is a* 2UFOHF *and $h_i : \Sigma \to \Sigma$, then:*

1. *(A union bound on Claim 14 in Lecture 9) For a random $\overline{h} = h_1, \dots, h_{\log T}$:*

$$\Pr_{\overline{h}}[\overline{h} \textit{ is not } \alpha\textit{-good for } M] \leqslant \log T \cdot |W|^3 \frac{1}{\alpha^2 |\Sigma|}$$

2. *(Claim 17 in Lecture 9) If $\overline{h}$ is α-good for M then:*

$$||M_{\overline{h}} - M^T|| \leqslant TW^2 \alpha$$

With this, we claim:

Claim 10. *There exists an $\overline{h} = h_1, \dots, h_{\log T}$ which $\frac{1}{2n^2}$-fools $\mathcal{B}$*

Proof. A union bound over the first item in Theorem **??** gives:

$$\Pr[\exists B \in \mathcal{B} : \overline{h} \text{ is not } \alpha\text{-good for } B] \leqslant n^3 \cdot \log T \cdot |W|^3 \frac{1}{\alpha^2 |\Sigma|}$$

And by the second item if $\overline{h}$ is α-good for $\mathcal{B}$ then for any $M \in \mathcal{B}$:

$$||B_{\overline{h}} - B^T|| \leqslant TW^2 \cdot \alpha$$

Picking $\alpha = \frac{1}{2n^5}$ and $|\Sigma| = n^{17} = \text{poly}(W, T, \frac{1}{\alpha})$ we get that there exists an $\overline{\overline{h}}$ which is $\frac{1}{2n^2}$-good for $\mathcal{B}$. We note that $\log |\Sigma| = O(\log n)$ $\square$

Corollary 11. *For any vertex on the path we've defined earlier:*

$$\Pr[v_i \rightsquigarrow v_{i+1} \textit{ in } M_{\overline{\overline{h}}}] \geqslant \frac{1}{2n^2}$$

Proof. As $\overline{\overline{h}}$ $\frac{1}{2n^2}$-fools $\mathcal{B}$ we have:

$$|\Pr[v_i \rightsquigarrow v_{i+1} \text{ in } M^T] - \Pr[v_i \rightsquigarrow v_{i+1} \text{ in } M_{\overline{\overline{h}}}]| \leqslant \frac{1}{2n^2}$$

the corollary follows as $\Pr[v_i \rightsquigarrow v_{i+1} \text{ in } M^T] \geqslant \frac{1}{n^2}$ since $v_{i+1} \in \Gamma_{v_i, \ell_i}$ $\square$

We finally define our HSG. The input for the generator is composed of three parts

- $\overline{h} = h_1, \dots, h_{\log T}$ where $h_i \in \mathcal{H}$
- $\overline{i} = i_0 = 1 < i_1 < \cdots < i_k = T$ a segmentation of $[1, n]$
- $\overline{x} = x_1, \dots, x_k$ where $x_i \in \Sigma$

And the output is given by:

$$G(\overline{h}, \overline{i}, \overline{x}) = (\mathcal{N}_{\overline{h}}(x_1))_{i_1} \circ (\mathcal{N}_{\overline{h}}(x_2))_{i_2 - i_1} \circ \cdots \circ (\mathcal{N}_{\overline{h}}(x_k))_{i_k - i_{k-1}}$$

where $(\mathcal{N}_{\overline{h}}(x))_{j_1 - j_2}$ denotes the output of Nisan's generator restricted to its first $j_1 - j_2$ bits.

By Claim 9 we know that for $\overline{h} = \overline{\overline{h}}$ and $\overline{i} = \overline{\ell}$ we must have a set of inputs $\overline{x}$ such that for any j the generator's jth block $(\mathcal{N}_{\overline{h}}(x_j))_{i_j - i_{j-1}}$ takes $v_j \to v_{j+1}$. It follows that $G(\overline{\overline{h}}, \overline{\ell}, \overline{x})$ takes v_0 to a vertex v_k such that $\Pr[v_k \rightsquigarrow v_{acc}] = 1$, which is what we needed.

Claim 12. *The seed length of G is $O(\log^2 n) + O(\log \frac{1}{\epsilon})$*

Proof. A straightforward computation shows that:

- $|\overline{h}| = k \cdot 2 \log |\Sigma| = \log T \cdot O(\log n) = O(\log^2 n)$
- $|\overline{i}| = \log \binom{T}{k} \leqslant \log \binom{n}{k} \leqslant k \log n = \frac{\log \frac{1}{\epsilon}}{\log n} \cdot O(\log n) = O(\log \frac{1}{\epsilon})$
- $|\overline{x}| = k \cdot \log |\Sigma| = O(\log \frac{1}{\epsilon})$

The claim follows $\square$

References

[HZ18] William M Hoza and David Zuckerman. Simple optimal hitting sets for small-success rl. Technical Report TR18-063, ECCC, 2018.

[ILL89] Russell Impagliazzo, Leonid A Levin, and Michael Luby. Pseudo-random generation from one-way functions. In *Proceedings of the twenty-first annual ACM symposium on Theory of computing*, pages 12–24. ACM, 1989.