Lecture Notes 6

18 Goldreich's "Toy Example"

18.1 Motivation

The goal of the next two lectures is to prove:

Theorem 1 *Weakly one-way functions exist iff strongly one-way functions exist.*

Proof: [Sketch] The backwards direction is trivial since every strongly one-way function is also weakly one-way.

For the forwards direction, we construct a strongly one-way function g from a weakly one-way function f. The construction is straightforward — g just evaluates f on a large number $t(n)$ of arguments and concatenates together the results. That is, $g(x) = (f(x_1), \ldots, f(x_{t(n)}))$, where $x = x_1, \ldots, x_{t(n)}$ is viewed as consisting of $t(n)$ length-n strings, and $t(n)$ is a suitably-chosen polynomial. Intuitively, $g(x)$ is hard to invert if $f(x_i)$ is hard to invert for at least one i. ■

Proving that this intuition correct requires careful attention to detail and a substantial amount of probabilistic reasoning. The full proof will be presented in the next lecture.

18.2 Some key ideas

Because of the complexity of the careful proof of theorem 1, it is useful to look first at a careful proof of a simpler theorem that illustrates many, but not all, of the ideas that go into the full proof.

We introduce some terminology to simplify discussions of the difficulty of inverting functions. We write A' *inverts* $f(x)$ to denote the event $A'(x, I^{|x|}) \in f^{-1}(f(x))$, and we write A' *fails on* $f(x)$ to denote its complement . With this shorthand, it follows that

$$\Pr[A' \text{ inverts } f(U_n)] = \Pr[A'(U_n, 1^n) \in f^{-1}(f(U_n))].$$

Definition: Let $\rho \in [0, 1]$. A function f is *deterministic ρ-one-way* if for all deterministic polynomial time algorithms A' and all sufficiently large n,

$$\Pr[A' \text{ fails on } f(U_n)] \geq \rho(n).$$

Theorem 2 (Proposition 2.3.3 in text) *Suppose f is deterministic 1/3-one-way. Define*

$$g(x_1, x_2) = (f(x_1), f(x_2)).$$

Then g is deterministic 0.55-one-way.

(Note that $0.55 < 1 - (2/3)^2$.)

Proof: Suppose g is not 0.55-one-way. Then there exists a polynomial time algorithm B' that inverts $g(U_{2n})$ with success probability $> 1 - 0.55 = 0.45$ for ∞-many n. Consider such an n and let $N = 2^n$.

Define the $N \times N$ matrix μ by $\mu(x_1, x_2) = 1$ if B' succeeds on input $g(x_1, x_2) = (f(x_1), f(x_2))$ and $\mu(x_1, x_2) = 0$ otherwise, where $|x_1| = |x_2| = n$. Then

$$\Pr[B'(U_{2n}) \text{ inverts } g(U_{2n})] = \frac{\text{\# 1's in } \mu}{N^2} > 0.45. \tag{1}$$

If B' operated independently on each component of the pair $(f(U_n), f(U_n))$, then μ would have a simple structure. Namely, one could define

$$\begin{aligned} R &= \{x_1 \mid B' \text{ inverts } f(x_1)\} \text{ (the "good" rows)} \\ C &= \{x_2 \mid B' \text{ inverts } f(x_2)\} \text{ (the "good" columns).} \end{aligned}$$

It would then follow that B' succeeds in inverting $g(x_1, x_2)$ iff it succeeds on both $f(x_1)$ and $f(x_2)$, so $\mu(x_1, x_2) = 1$ iff $x_1 \in R$ and $x_2 \in C$. By assumption f is deterministic 1/3-one-way, so $|R|, |C| \leq (2/3)N$. Hence,

$$\frac{\#1's \text{ in } \mu}{N^2} = \frac{|R|}{N} \times \frac{|C|}{N} \leq \left(\frac{2}{3}\right)^2 = \frac{4}{9} < 0.45.$$

which would contradict inequality 1.

However, B' does not necessarily have to operate in such a way. Perhaps B' can somehow "solve" the simultaneous equations

$$\begin{aligned} f(x_1) &= y_1 \\ f(x_2) &= y_2 \end{aligned}$$

for x_1 and x_2 given y_1 and y_2 with higher success probability that it can achieve by solving the two equations independently. We don't know if this can be done, but we can't rule it out, either. Hence, we need a more sophisticated argument to derive our contradiction.

Call a row (column) of μ *good* if at least 0.1% of the entries in it are 1 and *bad* otherwise.

Claim 1 *The fraction of good rows (columns) in μ is $\leq 66.8\%$.*

Proof: [of claim] Suppose the fraction of good rows $> 66.8\%$. Consider the algorithm I' for inverting f on input y:

1. Choose x_2 uniformly at random in $\{0,1\}^n$.
2. Run $B'(y, f(x_2)) = (x', x'')$.
3. If $f('x) = y$, halt with output x', and announce "success".

Now, algorithm $A'(y)$ runs $I'(y)$ repeatedly for 10,000 times. If any of the runs of $I'(y)$ succeeds, then A' succeeds and gives the same output as I'; otherwise, A' fails.

For every good x_1,

$$\Pr[I' \text{ fails on } f(x_1)] \leq 1 - 0.001 = 0.999 \ .$$

Hence,

$$\Pr[A' \text{ fails on } f(x_1)] \leq 0.999^{10,000} < 0.001$$

so

$$\Pr[A' \text{ inverts } f(x_1)] > 0.999 \ .$$

Thus, the overall success probability of $A'(f(U_n))$ is at least

$$\begin{aligned} &\Pr[U_n \text{ is good}] \cdot \Pr[A'(U_n) \text{ inverts } f(U_n) \mid U_n \text{ is good}] \\ &\qquad > \quad 0.668 \times 0.999 > \frac{2}{3} \ . \end{aligned}$$

This contradicts the assumption that f is deterministic $1/3$-one-way. ■

To finish the proof of the theorem, we conclude from Claim 1 that there are at most $(0.668N)^2$ 1's in the intersection of the good rows and good columns. The number of 1's in each bad row and bad column is less than $0.001N$. Hence,

$$\#1\text{'s in } \mu \leq (0.668N)^2 + 2N(0.001N) < 0.449N^2.$$

This contradicts inequality 1 and completes the proof. ■