1 Lecture 3: More on Hard-core Bits

1.1 Goldreich-Levin Proof, $\frac{3}{4}$ version

1.1.1 Proof

We begin by assuming that we have some (probabilistic) adversary R whose probability of predicting a Hard-core bit $b = \langle x, p \rangle$ is $> \frac{3}{4} + \varepsilon$ over all strings p and x, for some non-negligible ε. That is, over all possible x, p, and random sets of coin flips w during its execution, R correctly computes b at least $\frac{3}{4} + \varepsilon$ of the time.

We show that this assumption allows R to be used to compute x from $f(x)$.

Just as with the probability 1 version from the previous section, we want to figure out the bits of x one at a time. However, we don't have the same level of certainty as before, so we have to use a slightly more creative strategy.

Consider some random string p, and define p^i to be the same as p, but with the i^{th} bit flipped:

$$p = p_1 p_2 \cdots p_i \cdots p_{|x|} \iff p^i = p_1 p_2 \cdots \overline{p_i} \cdots p_{|x|}.$$

Then, we are given $f(x)$, and we may have R attempt to find both $b_1 = \langle x, p \rangle$ and $b_2 = \langle x, p^i \rangle$. If both of the values b_1, b_2 that R gives are correct, then we may XOR them to get $b_1 \oplus b_2 = \langle x, p \rangle \oplus \langle x, p^i \rangle = x_i$.

And, over all p and x we have correctness happening at least $\frac{3}{4} + \varepsilon$ of the time. So if we do this over a large sample set of p strings, ignoring the effect of having a fixed x for now, we can say that on average, each trial will have the following probability of giving us the correct x_i value:

$$\begin{aligned} Pr[x_i \text{ correct}] &\geq 1 - Pr[b_1 \text{ wrong}] - Pr[b_2 \text{ wrong}] \\ &= 1 - \left(\frac{1}{4} - \varepsilon\right) - \left(\frac{1}{4} - \varepsilon\right) = \frac{1}{2} + 2\varepsilon. \end{aligned}$$

So this probability is bounded away from $\frac{1}{2}$ by a non-negligible amount (namely, 2ε, where ε is non-negligible). So, we would expect that over trials for a large sample set of random strings p, the value of x_i that appears more often (i.e., more than $\frac{1}{2}$ of the time) would be the correct value. We formalize this with an idea from statistics called the **"Chernoff Bound"**:

1.1.2 Chernoff Bound and Amplification

The statement of the Chernoff Bound is the following:

Given random variables $X_1, X_2, \cdots, X_n$ with identical distributions (thus, identical expected values), we define $X = \sum X_i$. Then, $E(X) = n \cdot E(X_1)$, and

$$Pr[X \geq (1+\beta)E(X)] \ < \ e^{\frac{-\beta^2 E(X)}{2}}.$$

Note that since the expected values are the same, we can identically write the inequality above as:

$$Pr[X \geq (1+\beta)nE(X_1)] \ < \ e^{\frac{-\beta^2 nE(X_1)}{2}}.$$

Which makes it clearer that as we **increase the number of trials linearly** (or according to any polynomial), the probability of exceeding the expected value by some constant factor **decreases exponentially**.

In particular, for this proof, we would like to show that we can make the probability of getting the wrong result for x_i in more than half of some n trials. Since, on average, we get the wrong value of x with probability at most $1/2 - 2\varepsilon$, the expected number of occurrences of the incorrect value is $n \cdot (\frac{1}{2} - 2\varepsilon)$. Of course, the trials, even with different values of p, are all from the same distribution of p and w over a fixed x, so we may apply the Chernoff Bound:

$$Pr\left[X \geq \frac{1}{2}n\right] = Pr\left[X \geq (1+\beta)\left(\frac{1}{2} - 2\varepsilon\right)n\right] \implies \beta = \frac{4\varepsilon}{1-4\varepsilon}$$

$$\implies Pr\left[X \geq \frac{1}{2}n\right] < e^{\frac{-\beta^2 E(X)}{2}} = e^{\frac{-4\varepsilon^2}{(1-4\varepsilon)}n}.$$

Since $\frac{3}{4} < \frac{3}{4} + \varepsilon < 1$, we certainly have $0 < \varepsilon < \frac{1}{4}$, so the coefficient of n in the exponent must be positive. Thus, as we increase n linearly, the probability that the wrong answer shows up more than half of the time decreases exponentially. So, we can increase n to be some value for which we are very confident in our result and get all of the x_i with near-certain probability.

We complete our proof with the following note. In the statement we are trying to prove, we assume that R can predict b from $f(x)$ and p with some probability $P + \varepsilon$, where P is some threshold probability, and ε is some non-negligible value. In our proof, we use $P = 3/4$, and in the actual theorem, Goldreich and Levin used $P = 1/2$. We will do the rest of this work in terms of P to show that the choice does not matter, though P shouldn't be less than $1/2$, since we are talking about an R which predicts one of two possible outcomes for a bit b.

However, the important thing to note is that this probability is taken **over *all* x, p, and w.**

So, must consider randomness and the probabilities as being from some arbitrary distribution throughout our proof. We manage the distribution in terms of the predicate string p and the adversary's coin flips w through the amplification. By running R many times, with many different p's, we end up sampling the distribution so that we can essentially ignore effects of choosing a specific p or w.

But, throughout this whole process, $f(x)$ and x are completely fixed and do not change. This poses a problem. Although we get good samples for p and w, there may be some x for which sampling over all p, w gives a probability of correctness that is almost 1, and some where sampling over all p, w gives a probability of correctness close to 0.

So, somehow, we have to deal with the fact that x and $f(x)$ are not going to change throughout our proof, and that there is no way to remove the effects of choosing a specific x from the sample. And yet, we somehow still must show that R can be used to invert f with some non-negligible probability.

What we have essentially assumed up to this point is that when we use R to invert $f(x)$, we have an x for which the probability of correctly predicting b from $f(x)$ and p, over all p and w, is greater than P, by some non-negligible amount like ε or $\varepsilon/2$. And, what we have shown is that for any such x, we can invert $f(x)$ to get x with as much accuracy as we want, due to amplification techniques.

We call any x of this form "good", and any x that is not of this form is "not good". For any x that is not good, all bets are off. We have no guarantee, at least by what we have shown, that any of the "not good" x allow R to be used to invert f.

So, in order to avoid having to go back and re-work this whole proof, we really need a non-negligible proportion of the x's to be "good". Then, for this non-negligible proportion, we can use R to invert f to find x with as much certainty as we want. What this ends up doing for us is guaranteeing that we properly invert f for some non-negligible proportion of all the x's, which means that R breaks the one-way function, according to its definition.

We show that the proportion of x's which are "good" is non-negligible by using **"Bayesian Conditioning"**.

1.1.3 Bayesian Conditioning

The main statement of Bayesian probability that we will be using is:

$$Pr[A] = Pr[A|B]Pr[B] + Pr[A|\neg B]Pr[\neg B].$$

The symbols are:

1. $Pr[X]$: the probability that some statement X is true.
2. $Pr[\neg X]$: the probability that some statement X is false.
3. $Pr[X|Y]$: the probability that some statement X is true, given that some other statement Y is true.

With this in mind, the statement is quite intuitive. We know that B is always either true or false. Then, we break down the probability of A being true in general into cases based on these two possible outcomes of B, and their probabilities.

For the purposes of our proof, we show that if the probability of predicting b from a given $f(x)$ over all x, p, w is $P + \varepsilon$, then at least $\frac{\varepsilon}{2}$ of the x must be "good", with a probability of success of at least $P + \varepsilon/2$.

Suppose, for the sake of contradiction, that less than $\varepsilon/2$ of all possible x are "good". We use bayesian conditioning, and let A be the statement "R correctly predicts b", and we let B be the statement "x is good". Then, we have the following:

1. $Pr[A] = P + \varepsilon$
2. $Pr[B] < \varepsilon/2$
3. $Pr[\neg B] \leq 1$
4. $Pr[A|B] \leq 1$
5. $Pr[A|\neg B] < P + \varepsilon/2$

$$\implies P + \varepsilon = Pr[A|B]Pr[B] + Pr[A|\neg B]Pr[\neg B] < \varepsilon/2 + P + \varepsilon/2 = P + \varepsilon.$$

But this says that $P + \varepsilon < P + \varepsilon$, which is clearly a contradiction. So, our original assumption must have been false, and at least $\varepsilon/2$ of all possible x must be "good".

This completes the proof. If an adversary R predicts the hard-core bit b from $f(x)$ and p with probability at least $\frac{3}{4} + \varepsilon$ over all x, p, w for some non-negligible ε, then R can be used as a subroutine to invert the one-way function f with non-negligible probability.

(In particular, there some non-negligible proportion of the values of x for which R can be used as a subroutine to invert $f(x)$ with probability arbitrarily close to 1.)

Thus, if any adversary R is able to predict a hard-core bit of the form $b = \langle x, p \rangle$ from $f(x), p$ with probability $3/4 + \varepsilon$ for some non-negligible ε, then R can be used to invert the 1WF f with non-negligible probability. So, predicting hardcore bits with this level of advantage is as difficult as inverting a 1WF.