Impagliazzo hardcore predicate lemma

Satyadev Nandakumar

August 17, 2023

The following lemma applies to *all* Boolean functions such that at size S, every circuit of size is guaranteed to make at least $\delta 2^n$ errors.

Definition 0.1. A distribution D over Σ^n is said to have *density* $\delta` > 0$ if for every string $x \in \Sigma^n$, we have $D(x) \leq \frac{1}{\delta 2^n}$.

This is related to a concept studied in pseudorandomness, called a δ-flat distribution.

Definition 0.2. A distribution on Σ^n is K-flat if there is a set of $0 < K \leq 2^n$ strings on which it is uniformly distributed.

Note that every K-flat distribution is a $K2^n$-dense distribution. Conversely, the set of density δ distributions can be viewed as the set of distributions over $\delta 2^n$-flat distributions.

We use the observation about the converse in the proof of the following theorem.

Theorem 0.3 (Impagliazzo hardcore lemma)**.** *For every $\delta > 0$, $f \in \Sigma^n \to \Sigma$, and $\forall \epsilon > 0$, if $H_a^{1-\delta}(f) \geq S$, then there is a δ-density distribution D such that for every circuit of size $\leq \frac{\epsilon^2}{100n} S$, we have*

$$Pr_{x \sim D}[C(x) = f(x)] \leq \frac{1}{2} + \epsilon. \tag{1}$$

Point to ponder:

Why did we not write the conclusion of the above theorem simply by saying

$$H_a^{1/2+\epsilon} \geq \frac{\epsilon^2}{100n} S? \tag{2}$$

A: The H_a notation is only when the input $x \sim U_n$. In the above, we conclude for $x \sim D$. We *do not* have the result for $x \sim U_n$. We can think of the above as a result as x being almost uniformly distributed over a $\delta 2^n$-sized subset of Σ^n, ignoring the subtle difference between a flat distribution and a δ-dense distribution.

Proof overview: The simplified alternative proof, from Arora-Barak, is non-constructive: it does not explicitly construct the δ-dense distribution H, but it shows that it exists. Impagliazzo's original proof is constructive.

content of the theorem

The content of the Impagliazzo hardcore lemma is the following. If a function if $(1-\delta)$ hard on average, then either

1. $\forall\, C$ of size S, the fraction of inputs it gets wrong is approximately $\delta 2^n$. However, these "mistake sets" may be disjoint. As a result, if we take a few circuits, it may turn out that every instance is solvable by at least one of those circuits.
2. extremely hard on a few instances but may even be easy on others: $\exists$ a small set T approximately of size $(\delta 2^n)$ on which every circuit of size S is wrong on nearly half the strings in T. On T^c, the function may be even easy to compute.

The theorem says that it is always the second case for every Boolean function f which is $(1-\delta)$ hard on average!! This is somewhat hard to believe, except if one keeps in mind that the circuit sizes involved may be small for "easy" Boolean functions.

Proof. Let $f : \Sigma^n \to \Sigma$ be such that $H_a^{1-\delta}(f) \geq S$. Assume the lemma is false.

Consider the following two-player game: There are two players who play randomized strategies:

Complexity Theorist (**C**): plays first, and chooses a distribution over δ-density distributions. This is equivalent to selecting a δ-density distribution D. (see inset)

Algorithmist (**A**): Chooses a distribution $\mathcal{C}$ over circuits of size $\leq \frac{\epsilon^2 S}{100n}$.

Now, the game proceeds: we draw a string x at random according to D and a circuit C at random according to $\mathcal{C}$. If $C(x) = f(x)$, then the complexity theorist **C** pays 1 unit to player **A**. Otherwise, there is no reward for **A**.

This is a zero-sum game, since, if **C** starts with c dollars and **A** starts with a dollars, then at the end of the game, the combined capital of both players is still $c + a$.

The von Neuman minmax theorem for zero sum games states that if both players adopt randomized strategies, then the order of players does not matter: **A** can attain the same *expected* value even playing first.

Note that the expected value that **A** gains in the above game is:

$$1 \times Pr_{x\sim D, C\sim\mathcal{C}}[C(x) = f(x)] \;+\; 0 \times Pr_{x\sim D, C\sim\mathcal{C}}[C(x) \neq f(x)] = Pr_{x\sim D, C\sim\mathcal{C}}[C(x) = f(x)]. \quad (4)$$

It is in the interest of **C** (the complexity theorist), to minimize this value as much as possible. By assumption, this value is at least $1/2 + \epsilon$.

By the von Neumann minimax theorem, we have

$$\min_D \max_{\mathcal{C}} \Pr_{x\sim D, C\sim\mathcal{C}}[C(x) = f(x)] = \max_{\mathcal{C}} \min_D \Pr_{C\sim\mathcal{C}, x\sim D}[C(x) = f(x)] \geq \frac{1}{2} + \epsilon. \quad (5)$$

Hence there is a distribution $\mathcal{C}_{\mathbf{max}}$ over circuits of this size such that

$$\Pr_{C\sim\mathcal{C}_{\mathbf{max}}, x\sim D}[C(x) = f(x)] \geq \frac{1}{2} + \epsilon. \quad (6)$$

About randomized choices

A can choose a *distribution* $\mathcal{C}$ over S'-sized circuits (i.e. circuits of size $\frac{\epsilon^2}{100n}$) since the set of such circuits is finite.

How can **C** choose a δ-dense distribution at random? This requires explanation.

First, we note that the set of δ-dense distributions is precisely the set of convex combinations of $\delta 2^n$ flat distributions, as we mentioned in the beginning of this chapter. Since there are 2^n strings in Σ^n and $\binom{2^n}{\delta 2^n}$ ways to select the subset of elements on which to place the probability, we have $\binom{2^n}{\delta 2^n}$ flat sources, a finite set - denote this by $F_1, \ldots, F_{\binom{2^n}{\delta 2^n}}$.

Now, a probability distribution over such δ-dense distributions is a convex combination of δ-dense distributions. Such convex combinations of convex combinations of $\delta 2^n$ flat distributions can be written merely as convex combinations of the finite number of $\delta 2^n$ flat distributions. This leads us to conclude that a distribution over δ-dense distributions is itself a δ-dense distribution.

So for the player **C**, all (s)he has to do is to pick a set of convex weights $a_1, \ldots, a_{\binom{2^n}{\delta 2^n}}$ - *i.e.* all $a_i \geq 0$, and $\sum a_i = 1$. The δ-dense distribution will be

$$\sum_{i=1}^{\binom{2n}{\delta 2^n}} a_i F_i. \tag{3}$$

Call a string x "tough" if $\text{Pr}_{C \sim \mathcal{C}_{\mathbf{max}}}[C(x) = f(x)] < 1/2 + \epsilon$, and easy otherwise. (Note that this is a probability over circuits for a given x.)

Then there are at most $\delta 2^n$ tough strings. Otherwise, we could let D be a uniform distribution over the tough strings, and this would violate our assumption.

Let us choose a circuit C as follows: Set $t = 50n/\epsilon^2$, pick $C_1, \ldots, C_t$ independently from $\mathcal{C}_{\mathbf{max}}$. Let

$$C(x) = \text{majority}\{C_1(x), \ldots, C_n(x)\}. \tag{7}$$

Using Chernoff bounds, the probability that for every easy string x,

$$Pr_{C_1, \ldots C_n \sim \mathcal{C}_{\mathbf{max}}}[C(x) \neq f(x)] < 2^{-n}. \tag{8}$$

Using the fact that the sizes of each C_i is less than S', we may verify that the circuit for computing C has size less than S. (there are n smaller circuits, and then a circuit on top to compute the majority of n bits.)

Since there are at most 2^n easy strings. By the union bound, there must be a circuit C such that $C(x) = f(x)$ for *every* easy x. But since there are less than $\delta 2^n$ tough strings, this means that

$$\text{Pr}_{x \sim U_n}[C(x) = f(x)] > 1 - \delta, \tag{9}$$

which contradicts our assumption that $H_a^{1-\delta}(f) \geq S$.

This completes the proof. □