CS 710: Complexity Theory	Date: Mar. 30th

Lecture 17: Valiant-Vazirani Theorem and Efficient Amplification

Instructor: Jin-Yi Cai	Scribe: Jialu Bao

1 Valiant-Vazirani Theorem

Let $\#\phi$ denote the number of satisfying assignments of formula ϕ, then the unique SAT problem is to return 0 if $\#\phi = 0$, return 1 if $\#\phi = 1$. When $\#\phi$ is neither 0 nor 1, we do not care the result.

Theorem 1 (Valiant-Vazirani Theorem)**.** *If there exists a polynomial time algorithm that solves unique SAT problem, then $NP = RP$.*

Recall that if a language is L in RP, then there exists a polynomial time decision algorithm D and polynomial p such that for any $x \in L$, the success probability

$$\Pr_{y \in \{0,1\}^{p(|x|)}}[D(x, y) = 1] \geq \frac{1}{2}$$

and if $x \notin L$, then

$$\Pr_{y \in \{0,1\}^{p(|x|)}}[D(x, y) = 1] = 0$$

(For convenience of the notation, we use n to denote $p(|X|)$ in the following.)

Proof. We first want to show that if there exists a polynomial time algorithm A, then it can be used to show SAT problem is in RP. The main idea of the following proof is, given the formula ϕ of an SAT instance, to formulate another formula ϕ' that is likely to uniquely satisfied if ϕ is satisfied and not satisfied if ϕ is not satisfied. Then, we use the answer of A on ϕ' as the answer for whether ϕ is satisfiable.

To formulate ϕ', we first guess $\#\phi$. Say ϕ has n variables, then we pick an integer k in $[1, n]$ uniformly and guess $2^{k-1} \leq \#\phi \leq 2^k$.

When ϕ is satisfiable, there exists exactly one k that gives the correct range. We sample k uniformly, so with probability at least $\frac{1}{n}$, the guessed range is correct.

We then choose a target set T of size 2^{k+1}. Assuming that our guess of $\#\phi$ is correct, we have $2(\#\phi) \leq |T| \leq 4(\#\phi)$. We then set a 2-universal family of hash functions H from $0, 1^n$ to T. Pick $h \in H$ and an element α from T uniformly at random. Then let ϕ' satisfiable if and only if there exists α such that $\phi(\sigma) \wedge h(\sigma) = \alpha$. With Cook's reduction, we can transform ϕ' into conjunctive normal form that SAT problems takes.

Then we want to show that with non-trivial probability, ϕ' is uniquely satisfiable. As H is 2-universal, for any $a \neq b$, the collision probability $Pr_{h \in H}(h(a) = h(b)) = \frac{1}{|T|}$. Then the number of pairwise collisions could be expressed as

$$C^h = \sum_{a \neq b,\ \phi(a) = \phi(b) = 1} X^h_{a,b}$$

where $X^h_{a,b}$ are random variables that is 1 if $h(a) = h(b)$ and 0 otherwise. By linearity of expectation,

$$\begin{aligned} E_h[C^h] &= E_h\left[\sum_{a\neq b,\ \phi(a)=\phi(b)=1} X^h_{a,b}\right] \\ &= \sum_{a\neq b,\ \phi(a)=\phi(b)=1} E_h[X^h_{a,b}] \\ &= \sum_{a\neq b,\ \phi(a)=\phi(b)=1} Pr_{h\in H}(h(a)=h(b)) \\ &= \sum_{a\neq b,\ \phi(a)=\phi(b)=1} \frac{1}{|T|} \\ &= \binom{\#\phi}{2}\frac{1}{|T|} \leq \frac{\#\phi}{4} \end{aligned}$$

Then, using Markov's inequality, we have

$$\Pr[C \geq \frac{\#\phi}{3}] = \Pr\left[C \geq \frac{4}{3}\mathrm{E}[C]\right] \leq \frac{3}{4}$$

Therefore, with probability at least $\frac{1}{4}$, $C \leq \frac{\#\phi}{3}$. We wan to show that in this case ($C \leq \frac{\#\phi}{3}$), there are significant number of ϕ satisfying assignments whose images under h do not collide with others' images(we call them **the injective part**). With the number of collision pairs fixed, how they collide determines the size of injective part. In one extreme scenario, x collision pairs could be produced by y assignments that all mapped to one element in T and $x = \binom{y}{2}$. In another extreme scenario, x collision pairs could be produced by $2x$ assignments that each pair maps to a distinct element in T. There are many other scenario between these two, and among them, the second scenario minimizes the size of injective part to $\phi - 2C$. When $C \leq \frac{\#\phi}{3}$, $\phi - 2C$ is at least $\frac{1}{3}\#\phi$ and thus at least $\frac{1}{3}\frac{|T|}{4}$.

When a satisfying assignment σ of ϕ is in the injective part and $h(\sigma) = \alpha$, ϕ' such that $\phi'(x) = h(x) = \alpha \wedge \phi(x) = 1$ is uniquely satisfiable by σ. Thus, assuming that we have correctly guessed the interval for $\#\phi$, and $C \leq \frac{\#\phi}{3}$, the probability of selecting an $\alpha \in T$ such that ϕ' is uniquely satisfiable is $\frac{1}{12}$.

Thus, when ϕ is satisfiable, with probability of ϕ' uniquely satisfiable is at least

$$\begin{aligned} &P(\textbf{Guessed correct } k) \cdot P(C \leq \frac{\#\phi}{3} \mid \textbf{Guessed correct } k) \\ &\cdot P(\phi\textbf{uniquely satisfiable} \mid C \leq \frac{\#\phi}{3}\textbf{and guessed correct } k) \\ =&\frac{1}{n}\cdot\frac{1}{4}\cdot\frac{1}{12} = \frac{n}{48} \end{aligned}$$

Thus, with probability at least $\frac{1}{48n}$, A would accept ϕ'.

When ϕ is unsatisfiable, ϕ' is also unsatisfiable, and for sure A would reject ϕ'.

Thus, when A accept ϕ', we are sure that ϕ is satisfiable, but when A rejects ϕ', ϕ might still be satisfiable. If we repeat this process and formulate a bunch of ϕ' out of randomly sampled k, h, α, we can amplify the probability of having A accepts at least one ϕ'. After $48n$ repetition, that probability would become $1 - (1 - \frac{1}{48n})^{48n} \sim 1 - \frac{1}{e}$. Thus, when ϕ is satisfiable, the probability of A accepting with $m = poly(n)$ repetition becomes at least $\frac{1}{2}$, and thus the satisfaction of ϕ would be in RP.

Therefore, $SAT \in RP$ and $NP \subseteq RP$. For the other direction $RP \subseteq NP$, the proof is simpler: for any $L \in RP$, if $x \in L$, then a fraction of $y \in \{0,1\}^n$ would witness x in the verifier D , so a non-deterministic Turing Machine would accept $\exists y.D(x,y)$ in polynomial time; if $x \notin L$, then no $y \in \{0,1\}^n$ would witness x and satisfy $D(x,y)$, so a non-deterministic Turing Machine would reject x in polynomial time. As a non-deterministic Turing Machine suffices to decide any $L \in RP$, $RP \subseteq NP$. □

2 Efficient Amplification

Now we are going to discuss a bunch of amplification techniques in the context of amplifying for RP setting. The naive way of amplification is to the success probability is to run k trials with independent $y_1, ..., y_k \in \{0,1\}^n$. This would amplify the success probability to $\frac{1}{2^k}$ with the use of $n \cdot k$ random bits. Now we will show several techniques that saves us random bits.

2.1 Chor-Goldreich Generator

Here we use a universal family of hash function $\{h_s\}$, and the strategy is to pick a random s and then take $y_i = h_s(i)$ as witness strings in trials. Here $h_s(i)$ are pairwise independent but not fully independent – not even 3-wise independent. Then we want to bound the probability that $x \in L$ but for all trial i $(1 \le i \le k)$, $D(x, y_i)$, $D(x, y_i) = 0$.

Let Z_i be the 0-1 random variable that takes value 1 if and only if $D(x, y_i) = 1$. Z_i are pairwise independent, with expectation $\mu = \frac{1}{2}$, and variance is at most $\frac{1}{4}$. Then by Chebychev Inequality,

$$\begin{aligned}
\Pr_s[\forall 1 \le i \le k, D(x, y_i)] &= \Pr_s[\sum_{i=1}^{k} Z_i = 0] \\
&\le \Pr_s\left[\left|\sum_{i=1}^{k} Z_i - \frac{k}{2}\right| \ge \frac{k}{2}\right] \\
&\le \Pr_s\left[\left|\sum_{i=1}^{k} Z_i - k\mu\right| \ge \sqrt{k} \cdot \sqrt{k}\sigma\right]
\end{aligned}$$

By linearity of expectation, $E_s[\sum_{i=1}^{k} Z_i] = k\mu$, and by pairwise independence of Z_i, the variance of $\sum_{i=1}^{k} Z_i$ is $k \cdot \sigma^2$. Thus, Chebyshev Inequality would bound the probability above to be at most $\frac{1}{k}$.

This technique, named Chor-Goldreich generator, amplify the success rate to $1 - \frac{1}{k}$ with $2n$ random bits, which are used to pick hash function h_s.

2.2 Hash Mixing Lemma

Now we consider a more sophisticated technique, which gives a better bound. We began with construct a rather strange G, hoping it can generate many pseudo-random bits when given a small number of random bits. Let $\{h_s\}$ be a family of universal hash function $\{0,1\}^n \to \{0,1\}^n$. Then inductively define

- $G_0(y) = y$
- $G_{i+1}(y; s_1, ..., s_{s+1}) = G_i(y; s1, ..., s_i) \circ G_i(h_{s_{i+1}}(y); s_1, ..., s_i)$ for $i > 0$.

where $\circ$ denotes the concatenation of strings. For instance, $G_1(y; s_1) = y \circ h_{s_1}(y)$. Here G_{i+1} generates a string based on input random bits $y, s_1, .., s_{i+1}$ through concatenating $G_i(y; s1, ..., s_i)$ with $G_i(y'; s_1, ..., s_i)$, where $y' = h_{s_{i+1}}$.

When i is small, it does not seem efficient. For instance, when $i = 1$, it takes $2n$ bits to sample s_1 and n bits to sample y, but $G_1(y; s_1)$ only outputs $2n$ bits. But the number of bits generated by G_{k+1} always double from the number of bits generated by G_k, so for general k, G_k can generate $2^k n$ bits, with $(2k+1)n$ input random bits.

Now we want to show that the bits output by G_k are good approximation of fully random to some degree. Formally, we show the following lemma,

Lemma 1 (Hash Mixing Lemma)**.** *Let $\epsilon = 2^{-\frac{n}{3}}$. Then for all E subset of possible domain cross range of h_s, i.e. $E \subseteq \{0,1\}^n \times \{0,1\}^n$, for $1 - \frac{\epsilon}{4}$ fraction of s,*

$$\left| \Pr_{y \in \{0,1\}^n}[y \circ h_s(y) \in E] - \mu[E] \right| < \epsilon$$

where $\mu[E]$ is the probability measure of the set E, i.e., $\mu[E] = \Pr_{y,z \in \{0,1\}^n}[y \circ z \in E]$

Intuitively, if this lemma hold, then given fully random y, $y \circ h_s(y)$ is close to fully random.

Proof. Again, we use the trick of "decomposing" the event $\Pr_y[y \circ h_s(y)]$ to be an event on the sum of a set of indicator variables. Here, we define $Z_y^{h_s}$ that takes value 1 if $y \circ h_s(y) \in E$ and takes value 0 otherwise (the random variable is taken over fixed y and random s).

Then,

$$\Pr_y[y \circ h_s(y)] = \frac{1}{2^n} \sum_{y \in \{0,1\}^n} Z_y^{h_s}$$

Thus, by linearity of expectation, the expected probability (with respect to s) is the expected value of the sum,

$$\begin{aligned}
E_s[\Pr_y[y \circ h_s(y)]] &= E_s[\frac{1}{2^n} \sum_{y \in \{0,1\}^n} Z_y^{h_s}] \\
&= \frac{1}{2^n} \sum_{y \in \{0,1\}^n} E_s[Z_y^{h_s}] \\
&= \frac{1}{2^n} \sum_{y \in \{0,1\}^n} \Pr_s[y \circ h_s(y) \in E] \\
&= \frac{1}{2^n} \sum_{y \in \{0,1\}^n} \frac{|\{z \mid y \circ z \in E\}|}{2^n} \\
&= \frac{1}{2^n} |\{z \mid y \circ z \in E\}| = \mu[E]
\end{aligned}$$

We want to apply the Chebyshev Inequality, so we calculate the variance of $\Pr_y[y \circ h_s(y)]$:

$$\begin{aligned}\mathbf{Var}\left[\Pr_y[y \circ h_s(y)]\right] &= \mathbf{Var}\left[\frac{1}{2^n}\sum_y Z_y^{h_s}\right] \\ &= \frac{1}{2^{2n}}\sum_y \mathbf{Var}[Z_y^{h_s}] \;\; Z_y^{h_s} \text{ are pair-wise independent} \\ &\leq \frac{1}{2^{2n}}\sum_y \frac{1}{4} = \frac{1}{4\cdot 2^n}\end{aligned}$$

Thus, by Chebyshev inequality,

$$\Pr_s\left[\left|\Pr_{y\in\{0,1\}^n}[y \circ h_s(y) \in E] - \mu[E]\right| \geq \epsilon\right] \leq \frac{\epsilon}{4}$$

Thus, for $1 - \frac{\epsilon}{4}$ fraction of s,

$$\left|\Pr_{y\in\{0,1\}^n}[y \circ h_s(y) \in E] - \mu[E]\right| < \epsilon$$

□

With the lemma, we show the following theorem,

Theorem 2. *If $x \in L$ and the set of witness for x is $W_x \subseteq \{0,1\}^n$, then*

$$\Pr[G_k(y; s_1, ..., s_k) \subseteq \overline{W_x}] \leq (\mu(\overline{W_x}))^{2^k} + (\frac{k}{4} + 2)\epsilon$$

where $\epsilon = 2^{\frac{n}{3}}$, $\overline{W_x}$ is the complement of W_x.

Intuitively, if the 2^k string generated by $G_k(y; s_1, ..., s_k)$ are fully random, then the probability that all of them are in $\overline{W_x}$/none of them is the witness is $(\mu(\overline{W_x}))^{2^k}$. $G_k(y; s_1, ..., s_k)$ are not fully random, so we have an error term, which is exponentially small by this careful construction.

Proof. The main idea is to analyze the probability step by step. For any $1 \leq i \leq k$, and fixed hash function seeds $s_i, ..., s_{i-1}$, define $A_i = \{y \mid G_{i-1}(y; s_1, ..., s_{i-1}) \subseteq \overline{W_x}\}$. By construction, $G_{i+1}(y; s_1, ..., s_{i+1}) = G_i(y; s_1, ..., s_i) \circ G_i(h(y); s_1, ..., s_i)$, so $G_{i+1}(y; s_1, ..., s_{i+1}) \subseteq \overline{W_x}$ if and only if $y \in A_i$ and $h_{s_i}(y) \in A_i$, which is equivalent to $y \circ h_{s_i}(y) \in A_i \times A_i$. Thus, applying the Hash Mixing Lemma above, we derive that with at most $\frac{\epsilon}{4}$ fraction of seeds s would make $\left|\Pr_{y\in\{0,1\}^n}[y \circ h_s(y) \in E] - \mu[E]\right| \geq \epsilon$. We say that s_i is bad if s_i is one of those seeds.

Then

$$\begin{aligned}Pr[G_k(y; s_1, ..., s_k) \subseteq \overline{W_x}] \leq{}& \Pr[s_1 \text{ bad}] + \Pr[s_1 \text{ good}] \cdot \Pr[s_2 \text{ bad} \mid s_1 \text{ good}] + ... \\ &+ \Pr[s_1, ..., s_k \text{ good}] \cdot \Pr_y[G_k(y; s_1, ..., s_k) \subseteq \overline{W_x}] \qquad (1)\end{aligned}$$

By the Hash mixing lemma,$\Pr[s_k \text{ good} \mid s_1, ..., s_{k-1} \text{ good}] = \frac{\epsilon}{4}$ for all k. So the first $k-1$ additive terms sum to $\frac{\epsilon}{4}(k-1)$. The last term is upper bounded by $\Pr_y[G_k(y; s_1, ..., s_k) \subseteq \overline{W_x}]$. Denoted it by $\int_y$, and write $\beta = \mu(\overline{W_x})$. We prove by induction that $\int_y \leq \beta^{2^k} + 2\epsilon$.

$k=0$: $\int_y = \Pr_y[y \subseteq \overline{W_x}] = \beta$

$k=1$:

$$\begin{aligned}\int_y &= \Pr_y[y \circ h_{s_i}(y) \subseteq \overline{W_x}] \\ &\leq \Pr_{y,z}[y \circ z \subseteq \overline{W_x}] + \epsilon \text{ By assumption that } s_1 \text{ is good} \\ &\leq \beta^2 + \epsilon\end{aligned}$$

$k=2:$

$$\begin{aligned}\int_y &\leq \Pr_{y,z}[G_1(y; s_1) \circ G_1(z; s_1) \subseteq \overline{W_x}] + \epsilon \text{ By assumption that } s_1, s_2 \text{ are good} \\ &= (\beta^2 + \epsilon)^2 + \epsilon \leq \beta^{2^2} + 2\epsilon\end{aligned}$$

$k>2:$

$$\begin{aligned}\int_y &\leq \Pr_{y,z}[G_{k-1}(y; s_1, ..., s_{k-1}) \circ G_{k-1}(z; s_1, ..., s_{k-1}) \subseteq \overline{W_x}] + \epsilon \text{ As } s_1, ..., s_k \text{ are good} \\ &= (\beta^{2^{k-1}} + \epsilon)^2 + \epsilon \leq \beta^{2^k} + 2\epsilon\end{aligned}$$

Thus, $\int_y \leq \beta^{2^k} + 2\epsilon$. Substituting it into equation 1, then we have

$$\Pr[G_k(y; s_1, ..., s_k) \subseteq \overline{W_x}] \leq \beta^{2^k} + 2\epsilon + \frac{k}{4}\epsilon = (\mu(\overline{W_x}))^{2^k} + (\frac{k}{4} + 2)\epsilon$$

□

Recall that with $\mu(\overline{W_x})$ probability, a set of 2^k fully-random strings would also be subset of $\overline{W_x}$. So in terms of the providing at witness of a given input, the set of pseudo-random strings generated by $G_k(y; s_1, ..., s_k)$ is not much worse than a set of fully random strings.

References

[1] J. Cai. Lectures in Computational Complexity, 2003