CS 229r Information Theory in Computer Science Apr 9, 2019

Lecture 19

Instructor: Madhu Sudan *Scribe: Shyam Narayanan*

1 Outline

1. Brief review of 2-Prover games, Parallel Repetition Theorem, and key lemma
2. Outline of Parallel Repetition Theorem

We will not prove the Parallel Repetition Theorem, but will instead outline some of the main ideas and main challenges in proving this theorem. There are a lot of subtleties in the proof, so we won't have time to do the entire proof. To see full proofs of the Parallel Repetition Theorem, we point the reader to Raz [3] (the original proof), Holenstein [2] (a simplified proof), and notes by Barak [1].

2 Brief Review and commentary

Here, we review the previous lecture, as well as provide some commentary on the main lemma we will be focusing on for this lecture.

2.1 Two-Prover Games

Recall that a *two-prover game* $G = (X, Y, A, B, \mu, V)$ works as follows. We have a verifier, and two provers Alice and Bob. The verifier chooses $(x, y) \in X \times Y$ according to distribution μ on $X \times Y$, and sends Alice x and Bob y. Alice has some strategy $f : X \to A$ and Bob has some strategy $g : Y \to B$, so if Alice sees x she sends back $a = f(x)$ and if Bob sees y he sends back $b = g(y)$. The verifier has a verifying function $V : X \times Y \times A \times B \to \{0, 1\}$ with 1 being accept and 0 being reject. Importantly, Alice and Bob **do not** get to see the other person's input (so Alice doesn't see Y and Bob doesn't see X).

The value of the game G for some fixed strategies f, g is defined as $Val_{f,g}(G) = \mathbb{E}_{(X,Y)\sim\mu} V(X, Y, f(X), g(Y))$. Finally, the value of the game G, $\omega(G)$ is defined as $\max_{f,g}\{Val_{f,g}(G)\}$.

2.2 Parallel Repetition of Games

For a game $G = (X, Y, A, B, \mu, V)$, we define the k-fold repetition of the game as $G^{\otimes k} := (X^k, Y^k, A^k, B^k, \mu^k, V^k)$. Here, X^k, Y^k, A^k, B^k are just the k-fold product alphabets, μ^k is the product distribution on $(X \times Y)^k = X^k \times Y^k$, and

$$V^k(\overline{x}, \overline{y}, \overline{a}, \overline{b}) = \bigwedge_{i=1}^{k} V(x_i, y_i, a_i, b_i),$$

where $\overline{x} = x_1 x_2 \cdots x_k$ and similarly for $\overline{y}, \overline{a}, \overline{b}$. Note that our functions $f : X^k \to A^k$ and $g : Y^k \to B^k$ no longer have to be entrywise functions, so $f(\overline{x})_i$ can depend on x_j for $j \neq i$.

2.3 Raz's Parallel Repetition Theorem

We saw in the last class that $\omega(G^{\otimes k}) \geq \omega(G)^k$, but that equality didn't necessarily hold, by looking at Feige's counterexample, which gaves us a game G such that $\omega(G^{\otimes 2}) = \frac{1}{2} = \omega(G)$. However, we have the following theorem, proven by Raz.

Theorem 1. *[3] For all alphabets A, B and all $\varepsilon > 0$, there exists $\delta > 0$ such that for any game G with $\omega(G) \leq 1 - \varepsilon$, then $\omega(G^{\otimes k}) \leq (1-\delta)^k$.*

To prove this, we will need to introduce a key lemma that we introduced last time, and we focus the remainder of the lecture on this lemma. Fix strategies f, g, and for any set $S \subset [k]$, let w_S be the event that $\wedge_{i \in S} V(x_i, y_i, f(x)_i, g(y)_i) = 1$, i.e. $V(x_i, y_i, f(x)_i, f(y)_i) = 1$ for all $i \in S$. We have the following lemma:

Lemma 2. *For fixed strategies f and g, there exists some γ only depending on ε, A, B such that for all $|S| \leq \gamma \cdot k$, $\mathbb{P}[w_S] \leq 2^{-\gamma k}$ or $\exists i \notin S$ such that $\mathbb{P}[w_i | w_S] \leq 1 - \frac{\varepsilon}{100}$.*

Exercise 3. *Show how Lemma 2 implies Theorem 1.*

Proof. If suffices to show for any fixed strategies f, g, that there is some uniform δ such that $\mathbb{P}[w_{[1:k]}] \leq (1-\delta)^k$. We know that by definition of w_S, w_S is implied by w_T for all $S \subset T$. Therefore, if we ever have some $|S| \leq \gamma \cdot k$ such that $\mathbb{P}[w_S] \leq 2^{-\gamma k}$, then we can set $\delta = 1 - 2^{-\gamma}$, which is greater than 0 and does not depend on f, g. Therefore, we assume for all $|S| \geq \gamma \cdot k$, there exists some $i \notin S$ such that $\mathbb{P}[w_i | w_S] \leq 1 - \frac{\varepsilon}{100}$, so $\mathbb{P}[w_{S \cup \{i\}}] \leq \mathbb{P}[w_S] \cdot \left(1 - \frac{\varepsilon}{100}\right)$. Therefore, there exists some chain $i_1, i_2, \cdots, i_{\gamma \cdot k}$ of distinct elements such that if $S_j := \{i_1, \cdots, i_j\}$, $\mathbb{P}[w_{S_{j+1}}] \leq \mathbb{P}[w_{S_j}] \cdot \left(1 - \frac{\varepsilon}{100}\right)$, which means $\mathbb{P}[w_{[1:k]}] \leq \mathbb{P}[w_{S_{\gamma \cdot k}}] \leq \left(1 - \frac{\varepsilon}{100}\right)^{\gamma \cdot k}$. Therefore, we can set $\delta = 1 - \left(1 - \frac{\varepsilon}{100}\right)^{\gamma}$, which is again greater than 0 and does not depend on f, g. □

2.4 Commentary

We note some commentary for why we may need a lemma like Lemma 2.

First, why do we need to stop the lemma at $|S| \leq \gamma \cdot k$? Recall Feige's counterexample from last lecture. Define $X = \{1, 2\}$, $Y = \{3, 4\}$, $A = B = \{1, 2, 3, 4\}$, $\mu = Unif(X \times Y)$, and $V(x, y, a, b) = 1$ if $a = b$ and $b \in \{x, y\}$. Then, $\omega(G) = \omega(G^{\otimes 2}) = \frac{1}{2}$, and since the value of the product of two games is at least the product of the values, $\omega(G^{\otimes k}) \geq 2^{-\lfloor k/2 \rfloor}$. If we have a 20-fold product of the game, we can pair up the games i and $10 + i$ for $1 \leq i \leq 10$ so that if the verifier accepts on the first 10 coordinates, that can guarantee winning on the next 10 coordinates.

Second, why do we need to have it be for all $|S| \leq \gamma \cdot k$ and $\exists i \notin S$? Why can't we just induct on $S = \{1, 2, \ldots, r\}$, i.e. show that if $r \leq \gamma \cdot k$, either $\mathbb{P}[w_{[r]}] \leq 2^{-\gamma k}$ or $\mathbb{P}[w_{r+1} | w_{[r]}] \leq 1 - \frac{\varepsilon}{100}$? This is true by a very similar reason as before? We can choose any pairing of the coordinates (such as pairing 1 with 2, 3 with 4, and so on) so that if we know the verifier accepts coordinate 1, it must accept coordinate 2. Therefore, we can only guarantee that at each step, there is **some** additional index that we can add to get $\mathbb{P}[w_{S \cup \{i\}}] \leq \left(1 - \frac{\varepsilon}{100}\right) \cdot \mathbb{P}[w_S]$.

3 Proof Outline of the Main Lemma

3.1 Outline of Strategy

The idea for proving the main lemma is a reduction. Namely, suppose we have a contradiction to the lemma, i.e. $|S| \leq \gamma k$ but $\mathbb{P}[w_S] \leq 2^{-\gamma k}$ and for all $i \notin S, \mathbb{P}[w_i | w_S] \leq 1 - \frac{\varepsilon}{100}$. Then, it suffices to reduce this to finding a strategy for the original game G with value more than $1 - \varepsilon$. Let's say for simplicity that $S = \{1, 2, \ldots, r-1\}$ and $i = r$ for now.

The general goal is as follows. If Alice is just given x_r and Bob is just given y_r with $(x_r, y_r) \sim \mu$, then we want Alice to inject x_r into $X_1, \ldots, X_r = x_r, \ldots, X_k$ and Bob to inject y_r into $Y_1, \ldots, Y_r = y_r, \ldots, Y_k$. We want Alice to sample the other coordinates X_i and Bob to sample the other coordinates Y_i so that the distribution of $(X_1, Y_1), \ldots, (X_n, Y_n)$ looks like the total distribution conditioned on the verifier accepting the first $r - 1$ coordinates. Then, we get a way to get a probability of success of a single iteration of the game to be close to $\mathbb{P}[w_r | w_{[r-1]}]$ since Alice can just output the rth bit of $f(X_1, \ldots, x_r, \ldots, X_k)$ and Bob can just output the rth bit of $g(Y_1, \ldots, y_r, \ldots, Y_k)$. We will end up being forced not to condition on the first $r - 1$ questions but rather the first $r - 1$ answers.

The general strategy is as follows. We will use shared randomness to completely determine $X_{<r}, Y_{<r}$. (Eventually, we will need to remove the common randomness, but we can just choose an instance of the common randomness that maximizes probability of success.) Remember that r is quite small compared to k, since $r \leq \gamma \cdot k$. Letting $X^k := X_1 \cdots X_k$ and $Y^k := Y_1 \cdots Y_k$, we ideally want to sample X^k, Y^k such that $V(X_j, Y_j, f(X^k)_j, g(Y^k)_j) = 1$ for $j = 1, \dots, r-1$. Assuming we can do this perfectly, we will have $\mathbb{P}(V(X_r, Y_r, f(X^k)_r, g(Y^k)_r) = 1) \geq 1 - \frac{\varepsilon}{100}$. Therefore, we can even have $V(X_j, Y_j, f(X^k)_j, g(Y^k)_j) = 1$ with probability more than $1 - \frac{99\varepsilon}{100}$ and we will get a strategy for the original game with success probability above $1 - \varepsilon$.

But we note that $f(X^k)_{<r}, g(X^k)_{<r}$ also depends on $X_{>r}, Y_{>r}$. So we need to sample them as well! Unfortunately, there are too many coordinates: we have about $k - r \approx k$ coordinates since $r \leq \gamma k$, which is much smaller than k. However, we don't need to have the verifier accept in the "$> r$" coordinates (i.e. $V(X_j, Y_j, f(X)_j, f(Y)_j)$ can be anything for $j > r$), so all we will need is for $X_{>r}, Y_{>r}$ to roughly be from the correct distribution. We will use some stuff we saw from amortized communication results. Let $T_{>r}$ be the common randomness: T_j will either equal $(0, x_j)$ with probability $1/2$ or $(1, y_j)$ with probability $1/2$. Thus, for each coordinate, Alice and Bob either get just a sample of x_j or a sample of y_j for each $j > r$, so $X_j = x_j$ if $T_j = (0, x_j)$ and $Y_j = y_j$ if $T_j = (1, y_j)$. If this strategy "wins" in the first $r - 1$ coordinates (i.e. the verifier accepts), then we are happy: otherwise, repeat the algorithm.

Note, however, that Alice doesn't know what $Y_{>r}$ is and Bob doesn't know what $X_{>r}$ is. What we will have to do is fix the answers for the first $r-1$ coordinates: Alice will assume her answers are $a_1, \cdots, a_{r-1}$ and Bob will assume his answers are $b_1, \cdots, b_{r-1}$. We give an outline of the methods Alice and Bob will follow:

1. We first pick "typical" $x_1, \dots, x_{r-1}, y_1, \dots, y_{r-1}, a_1, \dots, a_{r-1}, b_1, \dots, b_{r-1}$ and fix these (we can do so with common randomness). By typical, we mean that most coordinates $j \geq r$ will satisfy W_j with probability at least $1 - \frac{\varepsilon}{50}$, conditioned on $X_i = x_i, Y_i = y_i, f(X)_i = a_i$, and $f(Y)_i = b_i$ for all $i < r$, and that $\mathbb{P}(X_i = x_i, Y_i = y_i, f(X)_i = a_i, f(Y)_i = b_i \ \forall i < r)$ is about $2^{-\Theta(r)}$. Indeed, we can pick such a typical set because we are assuming $\mathbb{P}[w_i | w_S] \geq 1 - \frac{\varepsilon}{100}$ for all $i \notin S = \{1, \cdots, r-1\}$ and $\mathbb{P}[w_S] \geq 2^{-r}$. Let W be the event that the first $r-1$ questions equal $X_{<r}$ for Alice, the first $r-1$ questions equal Y_{r-1} for Bob, the first $r-1$ answers are $a_{<r}$ from Alice, and $b_{<r}$ from Bob, over the probability distribution μ^k of questions asked by the verifier. Then, we will have for most values $i \notin S$,

   $$(X_i, Y_i)|_W \approx (U, V),$$

   i.e. the distribution of (X_i, Y_i) will be close in Total Variation Distance from the correct distribution (U, V). We will state this claim more rigorously and prove this in the next subsection.

2. Alice will set $X_1, \dots, X_{r-1}$ as above, and Bob will set $Y_1, \dots Y_{r-1}$ as above. Also, Alice and Bob will set $X_i = U$ and $Y_i = V$ if the verifier sends U to X and V to Y, for some randomly chosen i between r and k (which can be done with public randomness). We claim without proof that for most values of $i \geq r$,

   $$T_{j \geq r, j \neq i}|_{W, X_i, Y_i} \approx T_{j \geq r, j \neq i}|_{W, X_i} \approx T_{j \geq r, j \neq i}|_{W, Y_i}.$$

   By $T_{j \geq r, j \neq i}$, we are looking at the T_j's for $j \geq r$ and $j \neq i$, since we want to exclude the T_j's for $j \in S$, as well as $j = i$ as we are conditioning on X_i, Y_i.

3. Alice will sample $T_{j \geq r, j \neq i}|_{W, X_r}$ and Bob will sample $T_{j \geq r, j \neq i}|_{W, Y_r}$, though Alice and Bob will use a correlated sampling method.

4. Alice and Bob will complete this to get $X_{j \geq r, j \neq i} | W$ and $Y_{j \geq r, j \neq i} | W$. For $T_j = (0, x_j)$, Alice will set $X_j = x_j$ and Bob will sample Y_j conditioned on $X_j = x_j$ and the first $r-1$ coordinates of $g(Y_1, \cdots, Y_n)$ being $b_1, \cdots, b_{r-1}$. We do an analogous procedure for $T_j = (1, y_j)$. Finally, we will resample if necessary.

If you do this carefully, winning at the r^{th} coordinate will occur with probability about

$$\mathbb{P}(\text{Winning at } r^{\text{th}} \text{ coordinate} | \text{Winning at coordinates } 1, \dots, r-1),$$

which gives us a reduction from $\mathbb{P}[w_i|w_S]$ being large for all $i \neq S$ to a high-success strategy for the original game $G = G^{\otimes 1}$.

3.2 Proof of a Proximity Result

Finally, we ask how to prove proximity lemmas like $(X_r, Y_r)|_W \approx (U, V) \sim \mu$. To do this, we will need to show that for a random i chosen between r and k, (X_i, Y_i) conditioned on the verifier accepting the first $r-1$ coordinates does not differ too much from a randomly chosen (U, V). The event W should occur with at least $2^{-\Theta(\gamma k)}$ probability, so a result like Lemma 6 applied to $(X_r, Y_r), \ldots, (X_k, Y_n)$ will be useful.

Proving things like this is where the information theory we have developed will come in!

Proposition 4. *If X is some random variable and E is some event that happens with probability at least 2^{-d}, then the KL Divergence between $X|E$ and X satisfies $D((X|E)||X) \leq d$.*

Exercise 5. *Prove Proposition 4.*

Proof. Suppose that X has image $\Omega = [n]$ and $\mathbb{P}(X = i) = p_i, \mathbb{P}(X = i|E) = q_i$. Since $\mathbb{P}(E) \leq 2^{-d}$, we have

$$p_i = \mathbb{P}(X = i) \geq \mathbb{P}(X = i, E) = \mathbb{P}(E)\mathbb{P}(X = i|E) \geq 2^{-d} \cdot q_i.$$

Therefore, $q_i \leq 2^d \cdot p_i$, so $\frac{q_i}{p_i} \leq 2^d$. So by the definition of KL divergence, we have

$$D((X|E)||X) = \sum_{i=1}^{n} q_i \log_2 \frac{q_i}{p_i} \leq \sum_{i=1}^{n} q_i \cdot d \leq d,$$

since $\sum q_i = 1$. □

From this, we have the following Lemma:

Lemma 6. *If $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent events with distributions on $X \times Y$, and if E is an event that occurs with at least 2^{-d} probability, then we have $\mathbb{E}_i[\delta((X_i, Y_i)|E, (X_i, Y_i))] = O\left(\sqrt{d/n}\right)$, where i is uniformly chosen from $[n]$ and δ represents total variation distance.*

Proof. Note that by replacing X with $(X^n, Y^n) = (X_1, \ldots, X_n, Y_1, \ldots, Y_n)$ in Proposition 4, we have

$$D((X_1, \ldots, X_n, Y_1, \ldots, Y_n|E)||(X_1, \ldots, X_n, Y_1, \ldots, Y_n)) \leq d.$$

This equals

$$-H(X_1, \ldots, X_n, Y_1, .., Y_n|E) + \mathbb{E}_{X_1,\ldots,X_n,Y_1,\ldots,Y_n|E} \log \mathbb{P}(X_1 = x_1, \ldots, Y_n = y_n)$$
$$\geq -\sum_{i=1}^{n} H(X_i, Y_i|E) + \sum_{i=1}^{n} \mathbb{E}_{X_1,\ldots,X_n,Y_1,\ldots,Y_n|E} \log \mathbb{P}(X_i = x_i, Y_i = y_i)$$
$$= \sum_{i=1}^{n} \left(-H(X_i, Y_i|E) + \mathbb{E}_{X_i,Y_i|E} \log \mathbb{P}(X_i, Y_i)\right) = \sum_{i=1}^{n} D((X_i, Y_i|E)||(X_i, Y_i)),$$

using the fact that (X_i, Y_i)'s are independent events. Thus, we have by Pinsker's inequality,

$$\sum_{i=1}^{n} \delta((X_i, Y_i|E), (X_i, Y_i))^2 \leq \sum_{i=1}^{n} D((X_i, Y_i|E), (X_i, Y_i)) \leq d,$$

so we have that

$$\sum_{i=1}^{n} \delta((X_i, Y_i|E), (X_i, Y_i)) \leq \sqrt{d \cdot n},$$

which can be seen by either the Cauchy-Schwarz inequality or Jensen's inequality. The lemma follows by dividing both sides by n. □

References

[1] Barak, B. (2007). COS 522: Complexity Theory : Boaz Barak. Handout 10: Parallel Repetition Lemma. http://www.cs.princeton.edu/courses/archive/spr07/cos522/ho11.pdf.

[2] Holenstein, T. (2007). Parallel repetition: Simplifications and the no-signaling case. Proceedings of the Thirty-ninth Annual ACM Symposium on Theory of Computing, 411-419.

[3] Raz, R. (1998). A Parallel Repetition Theorem. SIAM Journal on Computing, 27(3), 763-803.