18.408 Topics in Theoretical Computer Science Fall 2022
Lectures 14,15

Dor Minzer

In this lecture we present the parallel repetition theorem, an amplification result that allows us to prove a PCP theorem with small soundness.

1 Gap Amplification via Parallel Repetition

Recall the basic PCP theorem:

Theorem 1.1. *There are absolute constant $\varepsilon > 0$ and $k \in \mathbb{N}$ such that the problem gap-Label-Cover$[1, 1-\varepsilon]$ is NP-hard on instances with alphabet size at most k.*

Our goal in this lecture is to prove an improved form of Theorem 1.1, in which the soundness is small:

Theorem 1.2. *For all $\varepsilon > 0$, there is $k \in \mathbb{N}$ such that the problem gap-Label-Cover$[1, \varepsilon]$ is NP-hard on instances with alphabet size at most k.*

We intend to use Theorem 1.1 to prove Theorem 1.2; how can we do that? Given a label-cover instance, how do we construct a harder label-cover instance? To motivate this discussion, we take the 2-prover-1-round view on Theorem 1.1.

Suppose we have a computationally weak verifier V and two all powerful provers P_1 and P_2 that do not communicate. All 3 parties have a common label cover instance Ψ, and the verifier wishes to distinguish between the case that $\mathsf{val}(\Psi) = 1$ and the case that $\mathsf{val}(\Psi) \leqslant 1 - \varepsilon$. To do that, the verifier may ask each one of the provers a question, get an answer and decided whether to accept or reject.

Write $\Psi = (G = (L \cup R, E), \Sigma_L, \Sigma_R, \{\Phi_e\}_{e \in E})$. To execute the task above, the verifier can sample an edge $e = (u, v) \in E$ uniformly, send u to P_1 and v to P_2, and get labels $\sigma_u \in \Sigma_L$ and $\sigma_v \in \Sigma_R$ from them. The verifier then checks that these labels satisfy the constraint on e, that is that $(\sigma_u, \sigma_v) \in \Phi_e$, and if so accepts (and otherwise rejects).

It is easy to see that if $\mathsf{val}(\Psi) = 1$, then there are provers' strategies that make V accept with probability 1. It is also true, and not very difficult to show (try!), that if $\mathsf{val}(\Psi) \leqslant 1 - \varepsilon$, then no provers' strategy can make V accept with probability more than $1 - \varepsilon$. Hence, the prover has a good advantage in distinguishing between two cases. Still it is only an ε advantage, and a natural question is how can V increase it?

1.1 Sequential Repetition

The first idea that comes to mind is that V should just repeat this protocol several times. Namely, after sampling an edge, sending each one of the endpoints to one of the provers, receiving answers and checking the constraint, the verifier could repeat this process again by sampling another edge and so on. Thus, we get a 2-prover-multiple-round game, an it is easy to show that the advantage of V indeed increases. Unfortunately, 2-prover-multiple-round games do not have a simple PCP interpretation and we will not be

able to prove Theorem 1.2 using this idea. An essentially identical idea is for V to have access to t pairs of provers (P_i, P'_i) for $i = 1, \ldots, t$, and for each one of the pairs run the basic 2-prover-1-round game as before, independently. This operation, too, increases the advantage of the verifier, however the analogous PCP interpretation this would give us is a PCP with more than 2 queries ($2t$ queries to be precise), which is not good enough for proving Theorem 1.2.

1.2 Parallel Repetition

The next idea is that we should adapt the idea above while making sure to keep the number of provers to be exactly 2. A natural variant is to simply send each one of them multiple challenges. That is, the t-fold repeated game proceeds by the verifier V picking t edges $e_1, \ldots, e_t \in E$ uniformly at random, denoting $e_i = (u_i, v_i)$ for $i = 1, \ldots, t$ and send all of the challenges to each one of the provers in a single shot. That is, the verifier sends the first prover $(u_1, \ldots, u_t)$ and sends to the second prover $(v_1, \ldots, v_t)$, and excepts to get from each one of them a tuple of labels, say $(\sigma_{u_1}, \ldots, \sigma_{u_t})$ and $(\sigma_{v_1}, \ldots, \sigma_{v_t})$. The verifier then checks that for all $i = 1, \ldots, t$, the corresponding pair of labels $(\sigma_{u_i}, \sigma_{v_i})$ satisfies the constraint on e_i, that is that $(\sigma_{u_i}, \sigma_{v_i}) \in \Phi_{e_i}$, and if so accepts and otherwise rejects. We call this game the t-fold repeated game, and denote it by $\Psi^{\otimes t}$.

So what does this operation do? Well clearly, if $\mathsf{val}(\Psi) = 1$, the provers can simply assign their vertices according to some pair of satisfying assignments $A_L \colon L \to \Sigma_L$ and $A_R \colon R \to \Sigma_R$ and make the verifier accept with probability 1. Also, if $\mathsf{val}(\Psi) \leqslant 1 - \varepsilon$, then on each one of the challenges e_i, the provers manage to win with probability at most $1 - \varepsilon$, and since the challenges are chosen independently the probability that they win on all t of them, and thus make the verifier accept, is at most $(1 - \varepsilon)^t$. Or is it?

1.2.1 An Instructive Example for Pitfalls in Parallel Repetition

Let us consider an example of a 2-prover-1-round game that exhibits an interesting possibility that may occur in parallel repetition. In the basic game Ψ, the verifier picks as challenges (x, y) uniformly from $\{0, 1\}^2$, sends x to P_1 and y to P_2, and expects to get as answer from P_1 a vector $a \in \{1, 2\} \times \{0, 1\}$ and from P_2 a vector $b \in \{1, 2\} \times \{0, 1\}$. The verifier accepts if and only if $a = b = (i, \sigma)$ and prover i received σ as a challenge.

In Ψ, the provers may use the following strategy: P_1 can give an answer which is $(2, 0)$ and P_2 can give the answer $(2, 0)$, and the probability that they win is the probability that $y = 0$ which is $1/2$. In general, it can be shown that $\mathsf{val}(\Psi) \leqslant 1/2$, since to win the provers must choose the same i, and conditioned on i – say $i = 1$ – the second prover must send σ which is equal to x, but his answer is independent of x hence equal to x with probability at most $1/2$.

What about the value of the 2-fold repeated game, that is $\Psi^{\otimes 2}$? The argument above says that we should have that $\mathsf{val}(\Psi^{\otimes 2}) \leqslant 1/4$, alas this is false. Indeed, consider the setting of the 2-fold repeated game in which P_1 receives challenges (x_1, x_2) and P_2 receives challenges (y_1, y_2), and they need to generate $a(1), a(2) \in \{1, 2\} \times \{0, 1\}$ and $b(1), b(2) \in \{1, 2\} \times \{0, 1\}$ so that $a(1) = b(1) = (i_1, \sigma_1)$ and P_{i_1} received σ_1 in their first coordinate, and $a(2) = b(2) = (i_2, \sigma_2)$ and P_{i_2} received σ_2 in their second coordinate. To do that, P_1 simply outputs $(1, x_1)$ and $(2, x_1)$ and P_2 simply outputs $(1, y_2)$ and $(2, y_2)$. Note that if $x_1 = y_2$, the provers win using this strategy, hence they make the verifier accept with probability at least $1/2$!

1.2.2 Where Did We Go Wrong?

For the game above Ψ, it turns out that $\mathsf{val}(\Psi^{\otimes 2}) = \mathsf{val}(\Psi)$, so 2-fold repetition does not change the value of the game at all, let alone square it (as we claimed). The way the provers managed to do that is by correlating their answers to the challenges on both coordinates, so that with some probability they fail miserably (on both coordinates), but as a by-product they manage to win all coordinates with probability which is higher than expected. Inspecting our earlier "proof" of $\mathsf{val}(\Psi^t) \leqslant \mathsf{val}(\Psi)^t$, we see that we implicitly assumed that the answer that each prover gives to the challenge on the i^{th} coordinate only depends on that coordinate. This need not be the case, and as the above example shows that there are cases that the provers can use this to their advantage.

1.3 The Parallel Repetition Theorem

Still, it turns out that parallel repetition does work, in the sense that for large enough t it does decrease the probability V accepts in the case that $\mathsf{val}(\Psi) \leqslant 1 - \varepsilon$. More precisely, one has:

Theorem 1.3 (The Parallel Repetition Theorem)**.** *For all $\varepsilon > 0$, there exists $\delta > 0$ such that the following holds. Let Ψ be a projection 2-prover-1-round game, and suppose that $\mathsf{val}(\Psi) \leqslant 1 - \varepsilon$. Then*

$$\mathsf{val}(\Psi^{\otimes t}) \leqslant (1 - \delta)^t.$$

In words, Theorem 1.3 states that the value of the t-fold repeated game does decrease exponentially with t. There are other interesting aspects of Theorem 1.3, such as for example the precise dependency of δ on ε, and we may discuss that later on in the course.

We next show how to deduce Theorem 1.2 from Theorem 1.1 by appealing to the Parallel Repetition theorem.

Proof of Theorem 1.2. Let ε_0 be from Theorem 1.1, take δ_0 from Theorem 1.3 for ε_0 and choose $t = \frac{\log(1/\varepsilon)}{\delta_0}$. Given a label-cover instance $\Psi = (G = (L \cup R, E), \Sigma_L, \Sigma_R, \{\Phi_e\}_{e \in E})$ from Theorem 1.1, we construct a label cover instance $\Psi' = (G' = (L' \cup R', E'), \Sigma_{L'}, \Sigma_{R'}, \{\Phi'_e\}_{e \in E'})$ as follows. The sides of the bi-partite graph G' are

$$L' = L^t = \{\, (u_1, \ldots, u_t) \mid u_i \in L \; \forall i = 1, \ldots, t \,\}, \qquad R' = R^t = \{\, (v_1, \ldots, v_t) \mid v_i \in R \; \forall i = 1, \ldots, t \}$$

(corresponding to challenges in the setting of parallel repetition), and label sets

$$\Sigma_{L'} = \Sigma_L^t, \qquad \Sigma_{R'} = \Sigma_R^t.$$

The edge set E' has an edge between $(u_1, \ldots, u_t)$ and $(v_1, \ldots, v_t)$ if $(u_i, v_i) \in E$ for all $i = 1, \ldots, t$. The constraint on this edge then allows for tuples $(\sigma_1, \ldots, \sigma_t) \in \Sigma_{L'}$ and $(\tau_1, \ldots, \tau_t) \in \Sigma_{R'}$ such that $(\sigma_i, \tau_i) \in \Phi_{u_i, v_i}$ for all $i = 1, \ldots, t$. This completes the description of the reduction.

Note that the reduction runs in time $n^{O(t)}$, hence polynomial. Next, we analyze the completeness and the soundness of the reduction.

Completeness. If Ψ is fully satisfiable, then we can use a satisfying assignment pair A_L and A_R of it to assign all tuples in Ψ' accordingly, and notice that it satisfies all of the constraints of Ψ.

Soundness. If Ψ is at most $1-\varepsilon$ satisfiable, then as observed earlier the corresponding 2-prover-1-round game has value which is at most $1-\varepsilon$. Note that if we have an assignment to Φ' satisfying at least η fraction of the constraints, then the provers may use it to win the t-fold repeated game $\Psi^{\otimes t}$ with probability at least η; indeed the edges of Ψ' exactly correspond to challenges that they may face in the t-fold repeated game. Thus, in Ψ' at most $\mathsf{val}(\Psi^{\otimes t})$ of the constraints can be satisfied, and by Theorem 1.3 we have that $\eta \leqslant \mathsf{val}(\Psi^{\otimes t}) \leqslant (1-\delta_0)^t \leqslant \varepsilon$. □

2 On the Proof of the Parallel Repetition Theorem

There are several known approach to prove Theorem 1.3 but none of them is very easy. Roughly speaking, known proofs (including the original proof by Ran Raz) go via the route of information theory, or via spectral graph theory. Our goal here will be to give some flavor of the proof and thus we will omit many (very crucial) details. Our presentation will follow the information theoretic approach to parallel repetition.

2.1 A Little Bit of Information Theory

There are many basic and important notions of information theory, such as entropy, mutual information and KL-divergence and all of their condition counterparts. To simplify presentation we will define as little of them as possible, at the expense of appealing to intuition (instead of rigorous proofs).

2.1.1 Shannon Entropy

Still, we will need the most basic notion in information theory, namely the notion of Shannon Entropy defined as follows:

Definition 2.1. *Let* $\mathbf{X}$ *be a discrete random variable getting values in* X*. The Shannon Entropy of* $\mathbf{X}$ *is*

$$H(\mathbf{X}) = \sum_{x \in X} \Pr[\mathbf{X} = x] \log\left(\frac{1}{\Pr[\mathbf{X} = x]}\right).$$

Intuitively, $H(\mathbf{X})$ measures the amount of randomness the random variable $\mathbf{X}$. To verify this intuition, it makes sense to ask what is the maximal entropy a random variable $\mathbf{X}$ over X may have, and what sort of random variables achieve this maximum or values near it.

1. **Entropy of a random variable is at most logarithm of the size of the support.** Note that by Jensen's inequality,

$$\begin{aligned} H(\mathbf{X}) = \sum_{x \in X} \Pr[\mathbf{X} = x] \log\left(\frac{1}{\Pr[\mathbf{X} = x]}\right) &= \mathop{\mathbb{E}}_{x \sim \mathbf{X}}\left[\log\left(\frac{1}{\Pr[\mathbf{X} = x]}\right)\right] \\ &\leqslant \log\left(\mathop{\mathbb{E}}_{x \sim \mathbf{X}}\left[\frac{1}{\Pr[\mathbf{X} = x]}\right]\right) \end{aligned}$$

   since $\log(z)$ is concave. As $\mathbb{E}_{x \sim \mathbf{X}}\left[\frac{1}{\Pr[\mathbf{X}=x]}\right] = |X|$, it follows that $H(\mathbf{X}) \leqslant \log(|X|)$.

2. **Almost full entropy implies close to being uniform.** For a random variable $\mathbf{X}$ whose distribution is uniform over X, the previous bound it tight as then

$$H(\mathbf{X}) = \sum_{x \in X} \frac{1}{|X|} \log(|X|) = \log(|X|).$$

Moreover, one may observe that the uniform distribution over X is the unique distribution for which equality holds (by inspecting the equality case of Jensen's inequality). In fact, one can show that a random variable $\mathbf{X}$ that achieves near equality, that is a random variable $\mathbf{X}$ that has entropy at least $\log(|X|) - \varepsilon$, is close to being uniformly distributed over X. Here, closeness is with respect to the statistical distance between random variables, which is defined as: for random variables $\mathbf{X}$ and $\mathbf{Y}$ distributed over Ω, define

$$\mathsf{SD}(\mathbf{X}, \mathbf{Y}) = \frac{1}{2} \sum_{\omega \in \Omega} \left| \Pr[\mathbf{X} = \omega] - \Pr[\mathbf{Y} = \omega] \right|.$$

Then, we have the following result, which can be proved using a result known as Pinsker's inequality:

Claim 2.2. *If $\mathbf{X}$ is a distribution over X satisfying $H(\mathbf{X}) \geqslant \log(|X|) - \varepsilon$, then $\mathsf{SD}(\mathbf{X}, \mathbf{U}) \leqslant \sqrt{\varepsilon}$ where $\mathbf{U}$ is the uniform distribution over X.*

2.1.2 Conditional Shannon Entropy

We will also need the notion of conditional Shannon Entropies.

Definition 2.3. *Let $\mathbf{X}$ be a discrete random variable getting values in X, and let E be an event. Then the Shannon entropy of $\mathbf{X}|E$ is*

$$H(\mathbf{X}|E) = \sum_{x \in X} \Pr[\mathbf{X} = x \mid E] \log\left(\frac{1}{\Pr[\mathbf{X} = x \mid E]}\right).$$

Conditioning on an event can either increase or decrease the entropy of a random variable. Next, we define the Shannon entropy of a random variable conditioned on another random variable.

Definition 2.4. *Let $\mathbf{X}, \mathbf{Y}$ be a discrete random variable. Then the Shannon Entropy of $\mathbf{X}|\mathbf{Y}$ is*

$$H(\mathbf{X}|\mathbf{Y}) = \mathop{\mathbb{E}}_{y \sim \mathbf{Y}} [H(\mathbf{X}|\mathbf{Y} = y)].$$

Conditioning on a random variable can never increase the entropy of a random variable:

Claim 2.5. *For jointedly distributed $(\mathbf{X}, \mathbf{Y})$ discrete random variables we have that $H(\mathbf{X}|\mathbf{Y}) \leqslant H(\mathbf{X})$.*

Proof. Write for convenience $p_{x,y} = \Pr[\mathbf{X} = x, \mathbf{Y} = y]$ and $p_{y|x} = \Pr[\mathbf{Y} = y \mid \mathbf{X} = x]$, by Jensen's inequality

$$H(\mathbf{X}|\mathbf{Y}) = \sum_x \mathop{\mathbb{E}}_{y \sim \mathbf{Y}} \left[p_{x|y} \log\left(\frac{1}{p_{x|y}}\right) \right] \leqslant \sum_x \mathop{\mathbb{E}}_{y \sim \mathbf{Y}} \left[p_{x|y}\right] \log\left(\frac{1}{\mathbb{E}_{y \sim \mathbf{Y}}\left[p_{x|y}\right]}\right),$$

and the proof is concluded by noting that $\mathbb{E}_{y \sim \mathbf{Y}}\left[p_{x|y}\right] = p_x$, so the last sum is exactly $H(\mathbf{X})$. □

2.1.3 Entropy Sub-additivity

The Shannon entropy has several important properties, which are all very plausible sounding. For example, if we think of $H(\mathbf{X})$ as the amount of randomness in $\mathbf{X}$, then one may expect the following connection. Suppose we have $(\mathbf{X}, \mathbf{Y})$ that is jointedly distribution, and we look at $H(\mathbf{X}, \mathbf{Y})$, which measures the amount of randomness in $(\mathbf{X}, \mathbf{Y})$. Then we expect it to be equal to the amount of randomness in $\mathbf{X}$, plus the amount of randomness in $\mathbf{Y}$ conditioned on knowing $\mathbf{X}$. In notations, we expect that it will be true that

$$H(\mathbf{X}, \mathbf{Y}) = H(\mathbf{X}) + H(\mathbf{Y} \mid \mathbf{X}).$$

Indeed, this is true and not difficult to prove; write for convenience $p_{x,y} = \Pr[\mathbf{X} = x, \mathbf{Y} = y]$ and $p_{y|x} = \Pr[\mathbf{Y} = y \mid \mathbf{X} = x]$, then

$$H(\mathbf{X}, \mathbf{Y}) = \sum_{x,y} p_{x,y} \log\left(\frac{1}{p_{x,y}}\right) = \sum_{x,y} p_x p_{y|x} \log\left(\frac{1}{p_x p_{y|x}}\right) = \sum_{x,y} p_x p_{y|x} \log\left(\frac{1}{p_{y|x}}\right) + \sum_{x,y} p_x p_{y|x} \log\left(\frac{1}{p_x}\right),$$

and the first term is equal to $H(\mathbf{Y} \mid \mathbf{X})$ while the second term is equal to $H(\mathbf{X})$ (pushing the sum over y inside and notion that the sum of $p_{y|x}$ over y is 1). Using Claim 2.5, we conclude that Shannon entropy sub-additivity:

Claim 2.6. *Let $(\mathbf{X}, \mathbf{Y})$ be jointedly distributed discrete random variables. Then $H(\mathbf{X}, \mathbf{Y}) \leqslant H(\mathbf{X}) + H(\mathbf{Y})$.*

2.1.4 Entropy Decrease by Conditioning on an Event

The last fact we need about Shannon entropies is that if we have a random variable $\mathbf{U}$ distributed uniformly over a set U and an event E which is not too unlikely, then the entropy of $\mathbf{U}|E$ is still somewhat large:

Claim 2.7. *Let $\mathbf{U}$ be a discrete uniform random variable over a universe U, and let E be some event. Then*

$$H(\mathbf{U} \mid E) \geqslant H(\mathbf{U}) - \log\left(\frac{1}{\Pr[E]}\right).$$

Proof. We have

$$H(\mathbf{U} \mid E) = \sum_{u \in U} p_{u|E} \log\left(\frac{1}{p_{u|E}}\right) = \sum_{u \in U} p_{u|E} \log\left(\frac{\Pr[E]}{\Pr[\mathbf{U} = u \wedge E]}\right).$$

Since $\Pr[\mathbf{U} = u \wedge E] \leqslant \Pr[\mathbf{U} = u] = \frac{1}{|U|}$, we get that

$$H(\mathbf{U} \mid E) \geqslant \sum_{u \in U} p_{u|E} \log\Big(|U| \cdot \Pr[E]\Big) = \log(|U|) - \log\left(\frac{1}{\Pr[E]}\right) = H(\mathbf{U}) - \log\left(\frac{1}{\Pr[E]}\right).$$

□

2.2 The Information Theoretic Approach to Parallel Repetition

2.2.1 The High Level Approach

Let $\Psi = (G = (L \cup R, E), \Sigma_L, \Sigma_R, \{\Phi_e\}_{e \in E})$ be a 2-player-1-round game as before, and consider the t-fold repeated game. In that game, the verifier samples challenges, which are uniformly chosen edges $(\mathbf{X}_1, \mathbf{Y}_1), \ldots, (\mathbf{X}_t, \mathbf{Y}_t) \in_R E$, sends the challenges $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_t)$ to the first prover, the challenges $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_t)$ to the second prover, and expects answers $\mathbf{A}(\mathbf{X}_1, \ldots, \mathbf{X}_t) = (\mathbf{A}_1, \ldots, \mathbf{A}_t)$ from the first prover and answers $\mathbf{B}(\mathbf{Y}_1, \ldots, \mathbf{Y}_t) = (\mathbf{B}_1, \ldots, \mathbf{B}_t)$ from the second prover. We note that each answer $\mathbf{A}_i$ and each answer $\mathbf{B}_i$ may depend on all of the challenges the respective player has received, but we omit this from the notation to make it less cumbersome.

We say the provers win on coordinate i if $(\mathbf{A}_i, \mathbf{B}_i) \in \Phi_{(\mathbf{X}_i, \mathbf{Y}_i)}$, and denote this event by W_i. We also denote by $W = W_1 \cap W_2 \cap \ldots \cap W_t$ the probability the players win all of the coordinates. In these notations, our goal is to show that $\Pr[W] \leqslant (1-\delta)^t$ for some $\delta > 0$ depending only on ε.

To show this, we assume that this is not the case, and show by induction on s that then we may find coordinates $i_1, i_2, \ldots, i_s$ such that $\Pr\left[W_{i_s} \mid W_{i_1} \cap \ldots W_{i_{s-1}}\right] \leqslant 1 - \varepsilon/2 + O(\sqrt{\delta}) < 1 - \varepsilon/4$. Once we show that we will be done, as then we get for $s = t/100$ that

$$\Pr[W] \leqslant \Pr\left[\bigwedge_{i=1}^{s} W_i\right] = \prod_{j=1}^{s} \Pr\left[W_j \mid \bigwedge_{i=1}^{j-1} W_i\right] \leqslant (1-\varepsilon/4)^s \leqslant (1-\varepsilon/4)^{t/100} < (1-\delta)^t.$$

2.2.2 Overview of the Argument

For $s = 1$, the claim is obvious. We can take the coordinate $i = 1$, and note that $\Pr[W_1]$ is at most the probability the provers win in a single repetition game, which is at most $1 - \varepsilon$. We now move on to the inductive part, which is where most of the action takes place. Suppose we proved the statement for $s \geqslant 0$, and let $i_1, \ldots, i_s$ be the coordinates we found so far. Then, our goal is to find a new coordinate i such that even conditioned on winning coordinates $i_1, \ldots, i_s$, the probability the provers win coordinate i is still bounded away from 1.

To get some intuition, denote the event $W_{\leqslant s} = W_{i_1} \cap \ldots W_{i_s}$, and consider the distribution over the challenges conditioned on $W_{\leqslant s}$, that is the distribution of $(\mathbf{X}, \mathbf{Y}) \mid W_{\leqslant s}$. Intuitively, since the probability of $W_{\leqslant s}$ is not very small, the overall amount of information it provides about the challenges is small, so for a typical coordinate i we get very little information about the challenge there.

To formalize this intuition, we used the tools we developed in information theory. Let us view the joint distribution of $(\mathbf{X}, \mathbf{Y})$ as $(\mathbf{U}_1, \ldots, \mathbf{U}_t)$ where each $\mathbf{U}_i$ is a uniformly chosen edge from G. Then by Claim 2.7

$$H(\mathbf{X}, \mathbf{Y} \mid W_{\leqslant s}) = H(\mathbf{U}_1, \ldots, \mathbf{U}_t \mid W_{\leqslant s}) \geqslant H(\mathbf{U}_1, \ldots, \mathbf{U}_t) - \log\left(\frac{1}{\Pr[W_{\leqslant s}]}\right).$$

Note that $H(\mathbf{U}_1, \ldots, \mathbf{U}_t) = t\log(|E|)$, and that $\Pr[W_{\leqslant s}] \geqslant \Pr[W] \geqslant (1-\delta)^t$, so we get that

$$H(\mathbf{U}_1, \ldots, \mathbf{U}_t \mid W_{\leqslant s}) \geqslant t\log(|E|) - t\log\left(\frac{1}{1-\delta}\right),$$

and using $\log(1/(1-\delta)) \leqslant 2\delta$ which holds for sufficiently small δ, we get that

$$H(\mathbf{U}_1, \ldots, \mathbf{U}_t \mid W_{\leqslant s}) \geqslant t(\log(|E|) - 2\delta).$$

Thus, thinking of this intuitively, this says that on average, on each one of the $\mathbf{U}_i$'s we lost entropy of at most 2δ which is very little. We can formalize this by using the sub-additivity of entropy, namely Claim 2.6, to note that

$$H(\mathbf{U}_1, \ldots, \mathbf{U}_t \mid W_{\leqslant s}) \leqslant \sum_{i=1}^{t} H(\mathbf{U}_i \mid W_{\leqslant s}),$$

so combining we get that

$$\frac{1}{t}\sum_{i=1}^{t} H(\mathbf{U}_i \mid W_{\leqslant s}) \geqslant \log(|E|) - 2\delta.$$

In particular, there exists $i = 1, \ldots, t$ such that $H(\mathbf{U}_i \mid W_{\leqslant s}) \geqslant \log(|E|) - 2\delta$, and by Claim 2.2 we get that $\mathbf{U}_i \mid W_{\leqslant s}$ is close to uniform over E, namely that $\mathsf{SD}(\mathbf{U}_i \mid W_{\leqslant s}, \mathbf{U}) \leqslant \sqrt{2\delta}$.

Note that if coordinate i was sampled according to $\mathbf{U}$ without the conditioning, the provers would win on it with probability at most $1 - \varepsilon$ by the assumption on the single repetition game. Above, we proved that the actual distribution over challenges they have is close to $\mathbf{U}$, hence we expect them to win on it with probability which is at most $1 - \varepsilon + \sqrt{2\delta}$. This turns out to be true (up to little more error terms), but is non-trivial to prove and where much of the effort in the actual proof of the parallel repetition theorem. [1]

If we ignore all of these complications though, we have just proved that $\Pr[W_i \mid W_{\leqslant s}] \leqslant 1 - \varepsilon + \sqrt{2\delta}$, as we wished.

[1] The reason is that to make this argument formal, we have to show that if the probability the provers win on coordinate i with high probability $1 - \varepsilon/100$ and $\mathsf{SD}(\mathbf{U}_i \mid W_{\leqslant s}, \mathbf{U})$, then one can use that to construct too good of a strategy to the basic game (with no repetition). To do that, one has to use the provers strategies, and in particular to be able to sample the rest of the coordinates of the game conditioned $W_{\leqslant s}$.

18.408 Topics in Theoretical Computer Science: Probabilistically Checkable Proofs
Fall 2022