CSE 533: The PCP Theorem and Hardness of Approximation **(Autumn 2005)**

Lecture 13: "Confuse/Match" Games (II)

Nov. 14, 2005

Lecturer: Ryan O'Donnell *Scribe: Ryan O'Donnell and Sridhar Srinivasan*

1 Reminders from Lecture 12

Our goal is to show the following theorem:

Theorem 1.1. *Let $s < 1$ be a constant. Suppose G is a "confuse/match"-style game with $\omega(G) \leq s$. Then if $k = \mathrm{poly}(1/\epsilon)$, $\omega(G^k) < \epsilon$.*

If G is repeated in parallel $\mathrm{poly}(1/\epsilon)$ times, with overwhelming probability (as a function of ϵ), there will be $\mathrm{poly}(1/\epsilon)$ many "confuse rounds" and $\mathrm{poly}(1/\epsilon)$ many "match rounds". These rounds will be randomly ordered. Further, it only helps the provers if we fix the questions in some rounds and tell them everything chosen in those rounds. For all of these reasons, it suffices (and indeed is equivalent) to prove the following theorem:

Theorem 1.2. *Let $s < 1$ be constant. Let $C = (1/\epsilon)^{39}$, $m = (1/\epsilon)^{11}$, and write $C' = C + m$. Given any 2P1R game G (with the projection property), let G' be the game with C' parallel rounds, in which C confuse rounds of G and m match rounds of G are played, in a random order. Then $\omega(G') \leq \epsilon$.*

To prove this theorem we will refer to the following two theorems proved in the last class:

Fact 1.3. *Let X be a set and γ be a probability distribution on X. Let $f : X^{C'} \to \{0, 1\}$, $C \geq 1$, where we think of $X^{C'}$ as having product probability distribution $\gamma^{C'}$. Let*

$$\mu = \mathop{\mathbf{Pr}}_{\vec{q} \in \gamma^{C'}}[f(\vec{q}) = 1].$$

Suppose we pick $i \in [C']$ at random and then pick $q_i \leftarrow \gamma$ at random. Let

$$\widetilde{\mu} = \widetilde{\mu}_{i,q_i} = \mathop{\mathbf{Pr}}_{i,q_i}[f \mid i, q_i].$$

Then,

$$\mathop{\mathbf{Pr}}_{i,q_i}\left[|\widetilde{\mu} - \mu| \geq 1/\sqrt[3]{C'}\right] \leq 1/\sqrt[3]{C'}.$$

Fact 1.4. *Let $R \subseteq [C]$ be any nonempty set, and $P : Q^{C'} \to A^C$. Suppose i and q_i are chosen randomly as in Fact 1.3; then*

$$0 \leq \mathop{\mathbf{E}}_{i,q_i}\left[\mathrm{Predictability}_R(P \mid i, q_i)\right] - \mathrm{Predictability}_R(P)] \leq 1/C'.$$

Remark 1.5. *Note that the above two facts are lemmas about general functions; i.e., they have nothing to do with 2P1R games.*

2 Intuition for the proof

Let $P : Q^{C'} \to A^{C'}$ be the strategy of Prover 2. The key component of our overall proof is codifying the intuition that P_2's strategy is either "mostly serial" or "highly dependent on many coordinates". (Note that this key theorem still has nothing to do with games; it is just a theorem about the structure of prover strategies.) To prove a rigorous statement along these lines takes some work. *Roughly speaking*, our key theorem will say something like the following:

High-level version of the key theorem. Suppose we pick a block of coordinates $R \subseteq [C']$ and the questions for that block, $\vec{q}_R$, randomly. Then either:

- It's very likely (over the choice of $R, \vec{q}_R$ that P's answers on the coordinates of R are highly unpredictable (over the choice of the remaining questions). OR,
- [There's a decent chance that P's answers on R are decently predictable, BUT...] Conditioned on any plausible set of answers in the coordinates of R, P is forced into a highly serial strategy on the remaining coordinates $[C'] \setminus R$.

Once we rigorize this statement, it is not too hard to use it to show that the provers cannot succeed in G' with probability more than ϵ.

3 Finding a "good block size"

Fix once and for all the strategy of Prover 2, $P_2 : Q^{C'} \to A^{C'}$. (We have switched notation, Q instead of Y.) Our first task is to determine a "good block size" r^* for the block R discussed in the previous section on intuition. This will be a number between 1 and $m/2$. To find r^*, we consider the following "thought experiment" regarding P_2:

We imagine filling in up to $m/2$ questions in $Q^{C'}$ at random, one by one. I.e., we pick $i_1, i_2, \ldots, i_{m/2}$ at random from $[C']$ (all distinct) and also $q_{i_1}, q_{i_2}, \ldots, q_{m/2}$, independently from γ, P_2's natural distribution on questions (inputs). Given $1 \leq r \leq m/2$, let R_r denote the (random) set $\{i_1, \ldots, i_r\}$, and write $\vec{q}_{R_r}$ for the associated questions. We now look at the following Predictabilities:

(1) $\text{Predictability}_{R_1}(P_2 \mid R_1, \vec{q}_{R_1})$

(1′) $\text{Predictability}_{R_1}(P_2 \mid R_2, \vec{q}_{R_2})$

(2) $\text{Predictability}_{R_2}(P_2 \mid R_2, \vec{q}_{R_2})$

(2′) $\text{Predictability}_{R_2}(P_2 \mid R_3, \vec{q}_{R_3})$

(3) $\text{Predictability}_{R_3}(P_2 \mid R_3, \vec{q}_{R_3})$

...

(m/2) $\text{Predictability}_{R_{m/2}}(P_2 \mid R_m, \vec{q}_{R_{m/2}})$

I.e., we see how Predictability of P_2 varies as we do the following two things: a) Give a new random question in a random coordinate (this makes Predictability go up); b) Require prediction on this new coordinate (this makes Predictability go down).

Consider now the list of *expectations* of the above quantities, (1), (1′), etc., where the expectation is over the choice of $R_{m/2}$ and $\vec{q}_{R_{m/2}}$.

Lemma 3.1. *There exists some "special block size" $1 \leq r^* < m/2$ such that in going from the (r^*) expectation to the (r^*+1) expectation, the expected Predictability goes down by at most $O(1)/m$.*

Proof. The expected value of any of the quantities (r) or (r') is a number in the range $[0,1]$, since Predictabilities are always in this range. In the $r \to r'$ steps, Predictability goes up. However, by Fact 1.4, it goes up by at most $1/(C'-r) \leq 1/C$. This is very small, and so the total amount the expected Predictability goes up over all $m/2$ steps is at most $(m/2)(1/C) \ll 1$, using the fact that $C \gg m$. Since all numbers are in the range $[0,1]$, and the total increase from beginning to end is at most 1, the total decrease, from the $r' \to r+1$ steps, must be at most 2. Thus there must be at least one step $(r^*)' \to r^*+1$, where the decrease in the Predictabilities' expectation is *at most* $2/(m/2) = 4/m$. (And, going from r^* to r^*+1 only makes the decrease less.) □

Thus we have identified a "special block size" $1 \leq r^* < m/2$ with the following property:

Corollary 3.2. *Let $R \subseteq [C']$ be a random set of cardinality r^* and let $\vec{q}_R$ be random questions for P_2 on these coordinates. Further, let i be a random coordinate from $[C'] \setminus R$ and let q_i be a random question for this coordinate. Then*

$$\mathbf{E}[\text{Predictability}_R(P_2 \mid \vec{q}_R)] - \mathbf{E}[\text{Predictability}_{R\cup\{i\}}(P_2 \mid \vec{q}_R, q_i)] \leq O(1)/m.$$

Notation. In the remainder of the proof, we will often use the following notation:

- $(R, \vec{q}_R)$ will be a "random block", meaning R is a randomly chosen subset of $[C']$ of size r^*, and $\vec{q}_R$ is a random set of questions for those coordinates.
- $\vec{a}$ will denote a string of answers in A^R.
- Notation like "$\mathbf{Pr}[\vec{a} \mid (R, \vec{q}_R)]$" will mean "the probability, over the choice of questions to P_2 outside of R, that P_2 will output the string of answers $\vec{a}$ in the positions R, *given* that it gets the questions $\vec{q}_R$ in the coordinates R".
- In particular, we will also use the notation $\mathbf{Pr}[\vec{a} \mid (R, \vec{q}_R), (i, q_i)]$ and $\mathbf{Pr}[\vec{a}, a \mid (R, \vec{q}_R), (i, q_i)]$, where (i, q_i) is a coordinate and a question outside R, and a is a single answer associated with the coordinate i.

4 The key theorem

In this section we use Corollary 3.2 to prove the key theorem, giving a dichotomy of the possibilities for P_2's strategy.

Definition 4.1. *Given $(R, \vec{q}_R)$, we say answer string $\vec{a} \in A^R$ is* dead *if $\mathbf{Pr}[\vec{a} \mid (R, \vec{q}_R)] \leq \epsilon$. Otherwise it is* alive. *We further say that the whole block $(R, \vec{q}_R)$ is* dead *if all answer strings $\vec{a} \in A^R$ are dead for it.*

Recall that ϵ is our target value for the repeated game. Define also $\eta = \epsilon^3$. We now make the following slightly tricky definition:

Definition 4.2. *We say that the block $(R, \vec{q}_R)$ "$(1-\eta)$ induces serial strategies" if*

- *$(R, \vec{q}_R)$ is alive.*
- *For every associated live answer $\vec{a} \in A^R$, there is a* serial strategy

$$S_{\vec{a}} : ([C'] \setminus R) \times Q \to A$$

such that with probability $\geq 1 - \eta$ over the choice of an additional random question (i, q_i), it holds that

$$\mathbf{Pr}[\vec{a}, S_{\vec{a}}(i, q_i) \mid (R, \vec{q}_R), (i, q_i)] \geq (1-\eta)\mathbf{Pr}[\vec{a} \mid (R, \vec{q}_R), (i, q_i)].$$

In other words, for almost all additional questions, the answer P_2 gives in the associated coordinate is essentially forced.

The idea for the key theorem is this. From Corollary 3.2 we know that for a typical r^*-block $(R, \vec{q}_R)$, there is almost no loss in Predictability between predicting on R and predicting on R plus one more random coordinate. By the definition of Predictability, this can only happen in one of two ways: First, either Predictability was negligible to begin with (i.e., almost all answer strings are dead); or, Predictability was non-negligible, *but*, every answer string on R forces a mostly serial way of answering most other questions.

Theorem 4.3 (Key Theorem)**.** *Given the strategy P_2, one of the following two cases holds:*

- *(Case 1) When $(R, \vec{q}_R)$ is a random r^*-block, with probability at least $1 - \epsilon$, $(R, \vec{q}_R)$ is dead.*
- *(Case 2) [At least an ϵ fraction of $(R, \vec{q}_R)$ are alive, but...] If $(R, \vec{q}_R)$ is a random* live *block, then with probability at least $1 - \epsilon$, $(R, \vec{q}_r)$ $(1-\eta)$-induces serial strategies.*

Proof. The proof essentially just involves unpacking the definitions involved in Corollary 3.2 and the phrase "$(1-\eta)$"-induces serial strategies. The proof is by contradiction. Assuming that neither Case 1 or Case 2 holds, we get that if you pick the block $(R, \vec{q}_R)$ at random, with probability at least ϵ^2 it is alive and *fails* to $(1-\eta)$-induce serial strategies. This means that there is some particular live answer string $\vec{a}$ such that, conditioned on P_2 answering $\vec{a}$, at least an η fraction of future

(i, q_i) pairs are not "$(1-\eta)$-determined". This is enough to show that in fact there is a slightly substantial loss in expected Predictability in predicting on r^*-sized blocks versus (r^*+1)-sized blocks, contradicting Corollary 3.2.

To do the details, we need to overcome a slight technicality: Answer strings that are alive (i.e., have probability at least ϵ) for $(R, \vec{q}_R)$ might become significantly less probable once the extra question (i, q_i) is picked. However, by Fact 1.3, this is extremely unlikely. Leaving this detail for later, what we have is the following:

Proposition 4.4. *Assuming the statement of the theorem does not hold, let $(R, \vec{q}_R)$ be a random r^*-block and let (i, q_i) be an additional random question. Then with probability at least $\eta\epsilon^2/2$, there exists some answer string $\vec{a} \in A^R$ such that:*

1. *For all $a \in A$, $\mathbf{Pr}[\vec{a}, a \mid (R, \vec{q}_R), (i, q_i)] \leq (1-\eta)\mathbf{Pr}[\vec{a}, a \mid (R, \vec{q}_R)]$.*
2. *$\mathbf{Pr}[\vec{a} \mid (R, \vec{q}_R), (i, q_i)] \geq \epsilon/2$.*

We show that this proposition leads to a loss of $\eta^2\epsilon^4/8$ in expected Predictability. From Point 1 above it is easy to conclude that

$$\sum_{a \in A} \mathbf{Pr}[\vec{a}, a \mid (R, \vec{q}_R), (i, q_i)]^2 \leq (1 - 2\eta + 2\eta^2)\mathbf{Pr}[\vec{a} \mid (R, \vec{q}_R), (i, q_i)]$$

because, subject to Point 1, the left-hand side is maximized if there is one answer contributing a $(1-\eta)$-fraction of the probability, another contributing an η-fraction, and the rest contributing 0. Using $2\eta - 2\eta^2 \geq \eta$ and $\mathbf{Pr}[\vec{a} \mid (R, \vec{q}_R), (i, q_i)] \geq \epsilon/2$, we get

$$\mathbf{Pr}[\vec{a} \mid (R, \vec{q}_R), (i, q_i)]^2 - \sum_{a \in A} \mathbf{Pr}[\vec{a}, a \mid (R, \vec{q}_R), (i, q_i)]^2 \geq \eta(\epsilon/2)^2.$$

From the definition of Predictability, we conclude that whenever the events of the Proposition occur, at least $\eta(\epsilon/2)^2$ is contributed to

$$\mathbf{E}[\mathrm{Predictability}_R(P_2 \mid \vec{q}_R)] - \mathbf{E}[\mathrm{Predictability}_{R \cup \{i\}}(P_2 \mid \vec{q}_R, q_i)].$$

Thus the total loss is at least $(\eta\epsilon^2/2) \cdot \eta(\epsilon/2)^2 = \epsilon^1 0/8 \ll O(1)/m$, contradicting Corollary 3.2.

To complete the proof we now only need to use Fact 1.3 to overcome the technicality mentioned earlier. Fix any $(R, \vec{q}_R)$ and any live answer string $\vec{a} \in A^R$, so $\mathbf{Pr}[\vec{a} \mid (R, \vec{q}_R)] \geq \epsilon$. Since $1/\sqrt[3]{C} \ll \epsilon/2$, Fact 1.3 tells us that $\mathbf{Pr}[\vec{a} \mid (R, \vec{q}_R), (i, q_i)] \geq \epsilon/2$ except with probability at most $1/\sqrt[3]{C}$ over the choice of (i, q_i). (We are using $C \leq C' - r^*$ here.) Union-bounding over all live $\vec{a}$ (of which there at most $1/\epsilon$) we get that for a random choice of $(R, \vec{q}_R)$ and (i, q_i), except with probability $1/\epsilon\sqrt[3]{(C)} = (1/\epsilon)^{12}$ every answer string $\vec{a}$ that is alive for $(R, \vec{q}_R)$ still satisfies $\mathbf{Pr}[\vec{a} \mid (R, \vec{q}_R), (i, q_i)] \geq \epsilon/2$. Since $1/\epsilon^{12} \ll \eta\epsilon^2/2$, we are done. □

5 Bounding the success probability of the provers

We now turn to dealing with 2P1R games; specifically, Theorem 1.2. Fix now also Prover 1's strategy, $P_1 : X^{C'} \to A^{C'}$.

In the theorem the way questions are picked is that a random m our C' coordinates are "match rounds" and the rest are "confuse rounds". We will equivalently imagine the rounds types and questions to be picked as follows:

Step 1: Pick $R \subseteq [C']$ of size r^* at random. These will be match rounds. Pick also both provers' questions for these rounds.

Step 2: Pick $m - r^*$ more random coordinates to be match rounds, and also pick the provers' questions for these rounds. (Note there are at least $m/2$ such rounds, as $r^* \leq m/2$.)

Step 3: All other coordinates are "confuse rounds"; pick the provers' questions for these rounds randomly. Note that the two provers get *independent* questions in these rounds, since they are confuse rounds.

The proof that the provers succeed with probability at most ϵ (actually, we will just prove $O(\epsilon)$) now divides into two cases, the case from the Key Theorem 4.3.

Case 1: In this case, after Step 1 is complete, the question block $(R, \vec{q}_R)$ for P_2 is dead with probability at least $1 - \epsilon$. We give up the ϵ here to the provers' success probability. So assuming it's dead, no answer string $\vec{a} \in A^R$ has more than ϵ probability of ultimately being given by P_2, over the choice of its remaining questions.

We would like to argue that this is still true after Step 2, when P_2 has gotten its remaining match round questions. This follows pretty easily from Fact 1.3. We can't quite union bound over all answer strings (there may be too many), but we can do the following: Group all possible strings in A^R into "clusters", where the clusters have total probability between $\epsilon/2$ and ϵ. There are at most $2/\epsilon$ such clusters. By Fact 1.3, for each additional random match round question, the probability a cluster gets more than $1/\sqrt[3]{C}$ more probable is at most $1/\sqrt[3]{C}$. Union-bounding over all remaining match round questions (at most m) and all clusters, we get that even after Step 2, except with probability at most $m \cdot (2/\epsilon) \cdot (1/\sqrt[3]{C} = 2\epsilon$, all clusters — and therefore all answer strings in A^R — have probability at most $\epsilon + m/\sqrt[3]{C} \leq 2\epsilon$.

Again, we give up the 2ϵ chance of answer strings becoming significantly more likely to the provers' success probability. So assuming this doesn't happen, we are through Steps 1 and 2 and we know that for every answer string $\vec{a} \in A^R$ in P_2's R coordinates, the probability P_2 will answer with that string — over the choice of its remaining confuse round questions — is at most 2ϵ. But now in Step 3, the provers' questions are chosen independently. So first imagine filling Prover 1's questions. Then this prover has all of its questions, and so its complete answer string is decided. In particular, its answer string on the coordinates R — call it $\vec{b}$ — is decided. Since these coordinates are match coordinates, and since G has the projection property, this means that there is a unique

string $\vec{a}$ that Prover 2 must answer on these R coordinates ($\pi(\vec{b})$) in order for the provers to win. But we have already argued the probability P_2 will answer this string, over the choice of its remaining confuse round questions, is at most 2ϵ.

We have therefore shown that the overall success probability of the provers in Case 1 is at most $\epsilon + 2\epsilon + 2\epsilon = O(\epsilon)$.

Case 2: In this case, let us first do Step 1, producing $(R, \vec{q}_R)$ (as well as Prover 1's questions on R). If $(R, \vec{q}_R)$ is dead then the analysis from Case 1 shows that the provers succeed with probability at most 4ϵ. Giving up this probability, we can condition on $(R, \vec{q}_R)$ being alive. Since we are in Case 2 of Theorem 4.3, we have that except with probability $1 - \epsilon$, $(R, \vec{q}_R)$ $(1 - \eta)$-induces serial strategies. We assume this is the case, giving up the remaining ϵ probability.

We now proceed to analyze the provers' success probability as follows:

$$\mathbf{Pr}[\text{success}] = \sum_{\vec{a} \in A^R} \mathbf{Pr}[\text{success AND } P_2 \text{ eventually answers } \vec{a} \text{ on } R].$$

We break up the sum above into the cases when $\vec{a}$ is alive and when it is dead. For the dead $\vec{a}$'s, we can again use the analysis from Case 1 to show that in total they contribute at most 4ϵ to the overall sum. Giving up this 4ϵ probability, we proceed to analyze the contribution from live $\vec{a}$'s, each of which we know $(1 - \eta)$-induces a serial strategy $S_{\vec{a}}$. We will show that for each live $\vec{a}$ the probability the provers succeed AND that P_2 eventually answers $\vec{a}$ is at most $O(\epsilon^2)$. Thus we conclude that the overall success probability is at most $4\epsilon + \epsilon + 4\epsilon + (1/\epsilon) \cdot O(\epsilon^2) \leq O(\epsilon)$, using the fact that there are at most $(1/\epsilon)$ live answers. This will complete the proof.

So let us fix a live answer $\vec{a}$ and its associated serial strategy $S := S_{\vec{a}}$. By definition of inducing serial strategies, we know that conditioning on P_2 answering with $\vec{a}$, the probability that it answers a random additional question (i, q_i) in accordance with S is at least $1 - 2\eta$. Hence the expected fraction of coordinates answered in accordance with S is at least $1 - 2\eta$. Indeed, this is true just of the coordinates in the remaining match rounds. Using Markov's inequality, we conclude:

Fact 5.1. *Conditioned on P_2 ultimately answering $\vec{a}$ in the coordinates R, except with probability ϵ^2, P_2 will answer at least a $1 - 2\eta/\epsilon^2 = 1 - 2\epsilon$ fraction of the coordinates in the Step 2 rounds according to the serial strategy S.*

Thus

$$\begin{aligned}\mathbf{Pr}[&\text{success AND } P_2 \text{ eventually answers } \vec{a}] \leq \epsilon^2 + \\ &\mathbf{Pr}[\text{success AND } P_2 \text{ eventually answers } \vec{a} \\ &\quad \text{AND } P_2 \text{ answers} \geq 1 - 2\epsilon \text{ fraction of remaining match coordinates according to } S].\end{aligned}$$

But as soon as we know P_2 plays a large fraction of coordinates according to a serial strategy, we can upper-bound its success probability. Since succeeding on all coordinates while using a serial

strategy on many coordinates is even harder than succeeding on many coordinates while using a serial strategy on all coordinates, we have the probability on the right, above, is at most

$$\mathbf{Pr}[P_2 \text{ matches on } \geq 1 - 2\epsilon \text{ fraction of the Step 2 match coordinates} \mid P_2 \text{ answers all Step 2 coordinates according to } S]. \quad (1)$$

But S is a serial strategy, and even with all questions in the Step 1 match rounds fixed, P_2 can match in any given Step 2 match round with probability at most $s < 1$. Thus it is expected to match in at most an s fraction of these coordinates — of which there are at least $m/2$. By a Chernoff bound, we easily get that (1) at most $\exp(-\Omega(\epsilon(m/2))) \ll O(\epsilon^2)$. This completes the proof.