COMS-E6998: Information Theory in Computer Science **Fall 2017**

Lecture 4: IC, Hellinger, and A $\Omega(n)$ LB for Disjointness

Lecturer: Omri Weinstein *Scribes: scribe-name1,2,3*

4.1 Information Cost

At least in the distributional world, it makes sense to look at a different complexity measure of CC protocols, namely, on the *information* that the protocol conveys to the parties about each other's inputs. Indeed, a natural extension of Shannon's entropy to the interactive setting is the *Information Complexity* of a function $\mathsf{IC}^{\varepsilon}_{\mu}(f)$, which informally measures the average amount of information the players need to disclose each other about their inputs in order to solve f with some prescribed error under the input distribution μ. From this perspective, communication complexity can be viewed as the extension of transmission problems to general tasks performed by two parties over a noiseless channel.

Definition 4.1.1 (Internal Information Cost [ChakrabartiSWY01, BBCR])**.** *The (internal) information cost of a protocol over inputs drawn from a distribution* μ *on* $\overline{X} \times \overline{Y}$, *is given by:*

$$\mathsf{IC}_{\mu}(\pi) := I(\Pi; X|Y) + I(\Pi; Y|X). \tag{4.1}$$

Recalling the KL-MI relationship, we can write the IC of π as:

$$\mathsf{IC}_{\mu}(\pi) = I(\Pi; X|Y) + I(\Pi; Y|X) = \mathbb{E}_{x,y}\left[\mathsf{D}\left(\frac{\Pi|x,y}{\Pi|y}\right)\right] + \mathbb{E}_{x,y}\left[\mathsf{D}\left(\frac{\Pi|x,y}{\Pi|x}\right)\right].$$

This view will be very useful later when we talk about the operational meaning of IC and interactive compression. Intuitively, the definition in (4.1) captures how much additional information the two parties learn about each other's inputs by observing the protocol's transcript. For example, the information cost of the trivial protocol in which Alice and Bob simply exchange their inputs, is simply the sum of their conditional marginal entropies $H(X|Y) + H(Y|X)$ (notice that, in contrast, the *communication* cost of this protocol is $|X| + |Y|$ which can be arbitrarily larger than the former quantity).

Remark 4.1.2 (External IC)**.** *Another information measure which makes sense at certain contexts is the* external *information cost of a protocol,* $\mathsf{IC^{ext}}_{\pi,\mu} := I(\Pi; XY)$, *which captures what an* external *observer learns on average about both player's inputs by observing the transcript of* π. *This quantity will be of minor interest in this survey (though it playes a central role in many applications). The external information cost of a protocol is always at least as large as its (internal) information cost, since intuitively an external observer is "more ignorant" to begin with. We remark that when* μ *is a* product *distribution, then* $\mathsf{IC^{ext}}_{\pi,\mu} = \mathsf{IC}_{\mu}(\pi)$ *(see, e.g., [Bra12]).*

One can now define the *information complexity* of a function f with respect to μ and error ε as the least amount of information the players need to reveal to each other in order to compute f with error at most ε:

Definition 4.1.3. *The Information Complexity* of f with respect to μ (and error ε) is

$$\mathsf{IC}^{\varepsilon}_{\mu}(f) := \inf_{\pi:\ \Pr_{\mu}[\pi(x,y) \neq f(x,y)] \leq \varepsilon} \mathsf{IC}_{\mu}(\pi).$$

What is the relationship between the information and communication complexity of f? This question and its (major) implications within TCS will occupy us for a full lecture in 2-3 weeks from now, but at least one direction of this question is easy – Since one bit of communication can never reveal more than one bit of information, the communication cost of any protocol is always an upper bound on its information cost over *any* distribution μ:

Lemma 4.1.4 (IC $\leq$ CC, [BravermanR11])**.** *For any distribution μ, $\mathsf{IC}_\mu(\pi) \leq \|\pi\|$.*

The proof is easy by induction on $\|\pi\|$. What about the other direction? Can we compares *interactive conversations*? In other words, is it true that

$$\mathsf{D}^\mu_{1/3}(f) \leq^? O(\mathsf{IC}^\mu_{1/3}(f))$$

We saw in the beginning of this course that the answer to this question was affirmative (in fact, an equality) for *one-way* protocols where only Alice speaks (Shannon's noiseless coding them and the Slepian-Wolf amortized them, Huffman one-shot). It turns out that, at least in the amortized sense, the answer is "YES" (we'll see this later on in the course). The answer to the *one-shot* problem, i.e., "is there an interactive analogue of the Huffman code?", is one of the most fascinating open problems in communication complexity, and we will devote an entire lecture to this problem and its major implications within complexity theory. But for the purpose of the next 2 lectures, we will study IC as an (elegant) tool for proving LBs on distributional (i.e., randomized) CC.

The role of private randomness in information complexity. A subtle but vital issue when dealing with information complexity, is understanding the role of private vs. public randomness. In public-coin communication complexity, one often ignores the usage of private coins in a protocol, as they can always be simulated by public coins. When dealing with *information complexity*, the situation is somewhat the opposite: Public coins are essentially a redundant resource (as it can be easily shown via the chain rule that $\mathsf{IC}_\mu(\pi) = \mathbb{E}_R[\mathsf{IC}^{\pi_R}_\mu]$), while the usage of private coins is crucial for minimizing the information cost, and fixing these coins is prohibitive (once again, for communication purposes in the distributional model, one may always fix the entire randomness of the protocol, via the averaging principle). To illustrate this point, consider the simple example where in the protocol π, Alice sends Bob her 1-bit input $X \sim Ber(1/2)$, XORed with some random bit Z. If Z is private, Alice's message clearly reveals 0 bits of information to Bob about X. However, for any fixing of Z, this message would reveal an entire bit(!). The general intuition is that a protocol with low information cost would reveal information about the player's inputs in a "careful manner", and the usage of private coins serves to "conceal" parts of their inputs. Indeed, it was recently shown that the restriction to public coins may cause an exponential blowup in the information revealed compared to private-coin protocols ([GKR14,BMY14]).

4.1.1 Additivity of Information Complexity

Perhaps the single most remarkable property of information complexity is that it is a fully additive measure over composition of tasks. This property is what primarily makes information complexity a natural "relaxation" for addressing direct sum and product theorems. The main ingredient of the following lemma appeared first in the works of [Razborov08,Raz98] and more explicitly in [BBCR,BravermanR11,Bra12]. In the following, f^n denotes the function that maps the tuple $((x_1, \ldots, x_n), (y_1, \ldots, y_n))$ to $(f(x_1, y_1), \ldots, f(x_n, y_n)))$.

Lemma 4.1.5 (Additivity of Information Complexity)**.** $\mathsf{IC}^\varepsilon_{\mu^n}(f^n) = n \cdot \mathsf{IC}^\varepsilon_\mu(f)$.

Proof. The ($\leq$) direction of the lemma is easy, and follows from a simple argument that applies the single-copy optimal protocol independently to each copy of f^n, with independent randomness. We leave the simple analysis of this protocol as an exercise to the reader.

The ($\geq$) direction is the main challenge. Will will prove it in a contra-positive fashion: Let Π be an ε-error protocol for f^n, such that $\mathsf{IC}^{\Pi}_{\mu^n} = I$ (recall that here ε denotes the per-copy error of Π in computing $f(x_i, y_i)$). We shall use Π to produce a *single-copy* protocol for f whose information cost is $\leq I/n$, which would complete the proof. The guiding intuition for this is that Π should reveal I/n bits of information about an average coordinate.

To formalize this intuition, let $(x,y) \sim \mu$, and denote $\overline{X} := X_1 \ldots X_n$, $X_{\leq i} := X_1 \ldots X_i$ and $X_{-i} := X_1 \ldots X_{i-1}, X_{i+1}, \ldots, X_n$, and similarly for $\overline{Y}, Y_{\leq i}, Y_{-i}$. A natural idea is for Alice and Bob to "embed" their respective inputs (x,y) to a (publicly chosen) random coordinate $i \in [n]$ of Π, and execute Π. However, Π is defined over n input copies, so in order to execute it, the players need to somehow "fill in" the rest $(n-1)$ coordinates, each according to μ. How should this step be done? The first attempt is for Alice and Bob to try and complete X_{-i}, Y_{-i} privately. This approach fails if μ is a non-product distribution, since there's no way the players can sample X and Y privately, such that $(X,Y) \sim \mu$ if μ correlates the inputs. The other extreme – sampling X_{-i}, Y_{-i} using public randomness only – would resolve the aforementioned correctness issue, but might leak too much information: An instructive example to consider is where, in the first message of Π, Alice sends Bob the XOR of the n bits of her uniform input X: $M = X_1 \oplus X_2 \oplus \ldots \oplus X_n$. Conditioned on X_{-i}, Y_{-i}, M reveals 1 bit of information about X_i to Bob, while we want to argue that in this case, only $1/n$ bits are revealed about X_i. So this approach reveals too much information.

It turns out that the "right" way of breaking the dependence across the coordinates is to use a combination of public and private randomness. Let us define, for each $i \in [n]$, the public random variable

$$R_i := X_{<i}, Y_{>i}.$$

Note that given R_i, Alice can complete all her missing inputs $X_{>i}$ *privately* according to μ, and Bob can do the same for $Y_{<i}$. Let us denote by $\theta(x,y,i,R_i)$ the protocol transcript produced by running $\Pi(X_1, ..., X_{i-1}, x, X_{i+1}, ..., X_n \ , \ Y_1, ..., Y_{i-1}, y, Y_{i+1}, ..., Y_n)$ and outputting its answer on the i'th coordinate. Let $\Theta(x,y)$ be the protocol obtained by running $\theta(x,y,i,R_i)$ on a uniformly selected $i \in [n]$.

By definition, Π computes f^n with a *per-copy* error of ε, and thus in particular $\Theta(x,y) = f(x,y)$ with probability $\geq 1 - \varepsilon$. To analyze the information cost of Θ, we write:

$$\begin{aligned} I(\Theta; x|y) = \mathbb{E}_{i,R_i}[I(\theta; x|y, R_i)] &= \sum_{i=1}^{n} \frac{1}{n} \cdot I(\Pi; X_i \mid Y_i, R_i) \\ = \frac{1}{n} \sum_{i=1}^{n} I(\Pi; X_i \mid Y_i, X_{<i} Y_{>i}) &= \frac{1}{n} \sum_{i=1}^{n} I(\Pi; X_i \mid X_{<i} Y_{\geq i}) \\ \leq \frac{1}{n} \sum_{i=1}^{n} I(\Pi; X_i \mid X_{<i} \overline{Y}) &= \frac{1}{n} \cdot I(\Pi; \overline{X} \mid \overline{Y}), \end{aligned}$$

where the inequality follows from Lemma 4.1.6, since $I(Y_{<i}; X_i | X_{<i}) = 0$ by construction, and the last transition is by the chain rule for mutual information. By symmetry of construction, an analogous argument shows that $I(\Theta; y|x) \leq I(\Pi; \overline{Y} \mid \overline{X})/n$, and combining these facts gives

$$\mathsf{IC}^{\Theta}_{\mu} \leq \frac{1}{n} \left(I(\Pi; \overline{X} \mid \overline{Y}) + I(\Pi; \overline{Y} \mid \overline{X}) \right) = \frac{I}{n}. \tag{4.2}$$

□

Lemma 4.1.6 (Conditioning on independent variables does not decrease information)**.** *Let A, B, C, D be four random variables in the same probability space. If A and D are conditionally independent given C, then it holds that $I(A; B|C) \leq I(A; B|CD)$.*

Proof. We apply the chain rule for mutual information twice. On one hand, we have $I(A;BD|C) = I(A;B|C)+I(A;D|CB) \geq I(A;B|C)$ since mutual information is nonnegative. On the other hand, $I(A;BD|C) = I(A;D|C) + I(A;B|CD) = I(A;B|CD)$ since $I(A;D|C) = 0$ by the independence assumption on A and D. Combining both equations completes the proof. □

Exercise 4.1.7. *Prove that $IC_{\mu}^{1/3}(Index_n) = \Theta(\lg n)$, where μ is uniform over $x \in_R \{0,1\}^n$ and $i \in_R [n]$. (Hint: Use Fano's inequality).*

4.2 Hellinger Distance

So far we've seen 2 ways of measuring distance b/w distributions: TV and KL (and their relation via Pinsker's inequality). Each of them had its special properties and was the "right" measure to work with depending on the underlying application we sought. The last distance metric we'll encounter in this course is the *Hellinger distance*, which will play a central role in the next two lectures.

Intuitively, Hellinger distance is a "geometric" interpretation of the probability distributions :

Definition 4.2.1 (Hellinger distance)**.** *The* Hellinger distance *between two (discrete) distributions μ, ν in the same probability space, is defined as*

$$h(\mu,\nu) := \frac{1}{\sqrt{2}}\|\sqrt{\mu} - \sqrt{\nu}\|_2.$$

Since $\|v\|_2^2 = \langle v, v\rangle$, it'll be more convenient to work with the squared Hellinger distance, which simplifies to

$$h^2(\mu,\nu) = 1 - \sum_x \sqrt{\mu(x)\nu(x)}$$

4.2.1 Properties and relationships to other metrics

Recall that the total variation distance is $\Delta(\mu\,\nu) := \max_A \mu(A) - \nu(A) = \frac{1}{2}\|\mu - \nu\|_1$

Claim 4.2.2.

$$h^2(\mu,\nu) \leq \Delta(\mu,\nu) \leq \sqrt{h^2(2 - h^2(\mu,\nu))} \leq \sqrt{2}h(\mu,\nu).$$

Proof. The inequalities essentially follow from the basic inequalities b/w ℓ_1 and ℓ_2 norms. The basic identity to use in these inequalities is simply $a^2 - b^2 = (a+b)(a-b)$.

First inequality:

$$h^2(\mu,\nu) := \frac{1}{2}\cdot\sum_i |\sqrt{\mu_i} - \sqrt{\nu_i}|^2 \leq \frac{1}{2}\cdot\sum_i |\sqrt{\mu_i} - \sqrt{\nu_i}|\cdot|\sqrt{\mu_i} + \sqrt{\nu_i}|$$
$$= \frac{1}{2}\sum_i |\mu_i - \nu_i| = \Delta(\mu,\nu).$$

Second and third inequalities:

$$\begin{aligned}
\Delta^2(\mu,\nu) &= \frac{1}{4}\left(\sum_i |\mu_i - \nu_i|\right)^2 = \frac{1}{4}\left(\sum_i (\sqrt{\mu_i} - \sqrt{\nu_i})(\sqrt{\mu_i} + \sqrt{\nu_i})\right)^2 \\
&\leq \frac{1}{4}\left(\sum_i (\sqrt{\mu_i} - \sqrt{\nu_i})^2\right) \cdot \left(\sum_i (\sqrt{\mu_i} + \sqrt{\nu_i})^2\right) \quad \text{by Cauchy-Schwartz} \\
&\leq \frac{1}{2} h^2(\mu,\nu) \cdot \left(2 + 2\sum_i \sqrt{\mu_i \nu_i}\right) = h^2(\mu,\nu) \cdot \left(1 + (1 - h^2(\mu,\nu))\right)) = \\
&\leq h^2(\mu,\nu) \cdot \left(2 - h^2(\mu,\nu)\right) \leq \sqrt{2} h^2(\mu,\nu).
\end{aligned}$$

□

Cut-and-paste property. In a sense, the reason to use Hellinger distance for analyzing CC protocols, is that it allows to formulate the *randomized analogue* of the *fooling-set* argument, i.e, of combinatorial rectangles. Recall that in the fooling set argument, we saw that if inputs $(x,y),(x',y')$ the same transcript in a *deterministic* communication protocol, then (x,y') and (x',y) must have the same transcript as well (since the transcript forms a combinatorial rectangle). This rectangle property can be extended to private-coin *randomized* protocols using *Hellinger distance* in the follows sense: If the transcript distributions for inputs (x,y) and (x',y') are close in Hellinger distance, then so are the transcript distributions for (x,y') and (x',y) [TODO : draw rectangle with crossing segments for illustration].

Lemma 4.2.3 (Cut-and-Paste, BJSK'04)**.** *For any private-coin randomized communication protocol π, it holds that:*

$$h^2(\pi_{xy}, \pi_{x'y'}) = h^2(\pi_{xy'}, \pi_{x'y}),$$

where $\pi_{ab}(\tau) := \Pr_R[\pi = \tau | X = a, Y = b]$ is the distribution on transcripts.

Proof. (Done in class) □

4.3 A linear Lower Bound for Set Disjointness

4.4 [BJKS]'s Information-Complexity Based Proof

Theorem 4.4.1. $\mathsf{R}_{1/2-\varepsilon}(\mathsf{Disj_n}) := \Omega(\varepsilon^2 n)$.

We will prove weaker (but somewhat simpler) theorem, namely, that $\mathsf{R}_{1/2-\varepsilon}(\mathsf{Disj_n}) := \Omega(\varepsilon^4 n)$ We'll use the minimax principle – in fact, we are actually going to prove that $\mathsf{IC}^{\varepsilon}_{\eta}(\mathsf{Disj_n}) \geq \Omega(\varepsilon^4 n)$ for some hard distribution η, which immediately implies the theorem. To this end, denote

$$\mathsf{R}_{\varepsilon}(\mathsf{Disj_n}) := \delta n$$

(for some $\delta > 0$) and let Π be a (public-coin) randomized protocol achieving the above communication. The first important observation of the proof is that $\mathsf{Disj_n}$ is in some sense "equivalent" to solving n (independent) copies of the 1-bit AND function, i.e.,

$$\mathsf{Disj_n}(x,y) = \neg \bigvee_{i=1}^{n} (x_i \wedge y_i).$$

This motivates a direct-sum approach for the proof – the idea is that a too-good-to-be-true protocol Π for $\mathsf{Disj_n}$ that reveals $o(n)$ bits of information (with respect to some carefully chosen hard distribution), can be used to solve the primitive 1-bit AND function with $o_\varepsilon(1)$ information against a non-trivial distribution, which would be a contradiction.

The first complication to the above program is that Disj is not exactly equivalent to solving n independent copies AND^n, but only the conjunction – hence in order for the hard distribution μ^n to be nontrivial, and for our "embedding" to AND to go through, μ better be (essentially) supported only on *non-intersecting inputs*, i.e., on $\{(0,1),(1,0),(0,0)\}$, but on the other hand the hard inputs must have *sufficiently large entropy* (to avoid sending the compressed inputs by one party). The following distribution achieves this:

$$\mu \sim \begin{cases} (0,0), & \text{w.p}1/3 \\ (1,0), & \text{w.p}1/3 \\ (0,1), & \text{w.p}1/3 \end{cases} \tag{4.3}$$

Of course, μ^n is a trivial distribution for $\mathsf{Disj_n}$ as the protocol can just output "1" and the players would be done with 0 communication . The key idea is to require the protocol to be correct on *every* input $(x,y) \in \{0,1\}^n \times \{0,1\}^n$ but measure its information only w.r.t μ^n (to facilitate the direct-sum argument above): **Key point:** Design protocol π for $\mathsf{AND}(X,Y)$, that is correct on *all* inputs $\{0,1\}$, but has low information *w.r.t* μ.

The players will use the "embedding" into a random coordinate of Π we saw last lecture. Formally, the run the following protocol π:

The Communication Protocol π for AND

1. **Inputs :** $(a,b) \in \{0,1\}$.
2. The players sample a publicly random coordinate $i \in_R [n]$ and set $X_i \leftarrow a, Y_i \leftarrow b$, respectively.
3. The players sample $R_i := X_{<i}Y_{>i} \sim \mu_X^{<i}\mu_Y^{>i}$ using public randomness, and "complete" the rest of their inputs ($X_{>i}$ and $Y_{<i}$) privately, according to $\mu_{X|Y}, \mu_{Y_X}$ respectively.
4. The players run $\Pi(X_1,\ldots,X_{i-1},a,X_i\ldots,X_n\ ,\ Y_1,\ldots,Y_{i-1},b,Y_i,\ldots,Y_n\ )$ and output its answer.

Figure 4.1: A communication protocol.

Since Π it an ε-error randomized protocol, it by definition succeeds w.p $\geq 1-\varepsilon$ on any input, and since μ is supported on *non-overlapping inputs*, it is clear that $\Pi(X,Y) = \mathsf{Disj_n}(X,Y) \Leftrightarrow \pi(x,y) = \mathsf{AND}(x,y)$. Let us henceforth denote by π_{ab} the conditional distribution of (the transcript of) π conditioned on $A=a, B=b$, that is:

$$\pi_{ab}(\tau) := \Pr_{i,R_i}[\Pi = \tau \mid A=a, B=b].$$

Note that even conditioned on r_i, a, b, π is still randomized as the players crucially use *private randomness* (R_A, R_B). By the above, we have

$$\Delta(\pi_{00}, \pi_{11}) \geq \varepsilon. \tag{4.4}$$

Now, by the additivity of IC that we saw last lecture ($\mathsf{IC}^{\varepsilon}_{\mu^n}(f^n) = n \cdot \mathsf{IC}^{\varepsilon}_{\mu}(f)$), we have that when $(A, B) \sim \mu$ (hence $(X, Y) \sim \mu^n$), the information cost of π is upper bounded as follows

$$\mathsf{IC}_{\mu}(\pi) \leq \frac{1}{n} \cdot \mathsf{IC}_{\mu^n}(\Pi) \leq \frac{\|\Pi\|}{n} = \frac{\delta n}{n} = \delta. \tag{4.5}$$

That is,

$$\mathbb{E}_{i,R_i}\left[I_{\mu}(A; \pi | B, R_i) + I_{\mu}(B; \pi | A, R_i)\right] \leq \delta.$$

This means that, on average, π reveals little information about the inputs $(a, b) \sim \mu$, which means that the distribution of $\pi_{a,b}$ should be close to π_a and to π_b, hence the latter distributions must be close to each other. Indeed, using exercise 4 from last homework (Pinsker's inequality), the above inequality implies

$$\mathbb{E}_{a,b,i,r_i}\left[\|\pi_{a,b,r_i} - \pi_{b,r_i}\|_1\right] \leq \sqrt{2 \ln 2\delta} \leq 2\sqrt{\delta},$$

and similarly

$$\mathbb{E}_{a,b,i,r_i}\left[\|\pi_{a,b,r_i} - \pi_{a,r_i}\|_1\right] \leq \sqrt{2 \ln 2\delta} \leq 2\sqrt{\delta}.$$

But in $\mu((0,0)) = 1/3$, hence $\|\pi_{0,0,R_i} - \pi_{b=0,R_i}\|_1 \leq 6\sqrt{\delta}$ and similarly $\mu((1,0)) = 1/3$, hence $\|\pi_{1,0,R_i} - \pi_{b=0,R_i}\|_1 \leq 6\sqrt{\delta}$, so by triangle inequality we have

$$\|\pi_{0,0,R_i} - \pi_{1,0,R_i}\|_1 \leq 12\sqrt{\delta}.$$

An analogues calculation (using the fact that $I_{\mu}(B; \pi | A, R_i) \leq \delta$) shows that

$$\|\pi_{0,0,R_i} - \pi_{0,1,R_i}\|_1 \leq 12\sqrt{\delta},$$

hence applying the triangle inequality one more time, we conclude that

$$\begin{aligned} & \|\pi_{0,1,R_i} - \pi_{1,0,R_i}\|_1 \leq 24\sqrt{\delta} \\ \iff & \Delta(\pi_{01}, \pi_{10}) \leq 48\sqrt{\delta}. \end{aligned}$$

In other words, the protocol π cannot distinguish well between the inputs $(1, 0)$ and $(0, 1)$. This is not quite what we need, as we would like to argue that $\Delta(\pi_{00}, \pi_{11})$ is small (in order to apply Equation (4.4) which would give a LB on $\delta \gtrsim \varepsilon^2$). So the question is whether a protocol that cannot distinguish well between $(1, 0)$ and $(0, 1)$, also cannot distinguish between $(0, 0)$ and $(1, 1)$? (Draw Picture). Since for *information cost* purposes we may always assume that π is a *private-coin* protocol (as Alice can send Bob R_i as the first message of π with no additional information), the Cut-and-Paste property of Hellinger distance precisely formalizes this "fooling set" argument:

Corollary 4.4.2. $h^2(\pi_{00}, \pi_{11}) = h^2(\pi_{10}, \pi_{01})$.

We can now finish the proof. By the relationship between Hellinger and statistical (total variation) distance (Claim 4.2.2), we have

$$48\sqrt{\delta} \geq \Delta(\pi_{10}, \pi_{01}) \geq h^2(\pi_{10}, \pi_{01}) = h^2(\pi_{00}, \pi_{11}) \geq \frac{1}{2} \cdot \Delta^2(\pi_{00}, \pi_{11}) \geq \frac{1}{2} \cdot \varepsilon^2.$$

where the last transition is by Equation (4.4). This implies

$$\delta \geq \Omega(\varepsilon^4)$$

, which finishes the proof.

We remark that it was recently shown [BM'13] that in fact $\mathsf{D}^{\mu}_{1/2-\varepsilon}(\mathsf{DISJ}_n) \geq \Omega(\varepsilon n)$, and this turns out to have implications for LP LBs (that we saw earlier in the course), but this is beyond our scope.

4.5 An $O(k)$ Protocol for Small-Set Disjointness

We conclude with a surprising upper bound for $DISJ_n^k$ due to Hasted and Wigderson. (Sketched in class).

References