# 8. Disjointness over product distributions

*Lecturer: Prahladh Harsha* *Scribe: Gaurav Rattan*

In the last several lectures, we appealed to the following lower bound on disjointness for proving a variety of lower bounds in various contexts.

**Theorem 8.1** (Kalyanasundaram-Schnitger [KS92], Razborov [Raz92])**.**

$$\mathrm{R}(\mathsf{DISJ}_n) = \Theta(n).$$

In today's lecture, we build towards this result by showing a weaker lower bound for randomized communication complexity of the disjointness problem due to Babai, Frankl and Simon [BFS86]:

$$\mathrm{R}(\mathsf{DISJ}_n) = \Omega(\sqrt{n}).$$

The proof of this weaker bound relies on the equivalence between randomized communication complexity and distributional communication complexity as discussed in the previous lecture. We will use an appropriately chosen *product distribution* $\mu$ to show that the corresponding distributional complexity $\mathrm{D}^{\mu}(\mathsf{DISJ}_n)$ and hence the randomized complexity $\mathrm{R}(\mathsf{DISJ}_n)$ is at least $\Omega(\sqrt{n})$. In the second part of the lecture, we show that, if one works with just product distributions, one cannot improve this bound significantly and achieve the lower bound in Theorem 8.1. More precisely, we show that

$$\max_{\text{product } \mu} \mathrm{D}^{\mu}(\mathsf{DISJ}_n) = O(\sqrt{n}\log n).$$

The latter result shows that the hard distribution $\mu$ which witnesses $\mathrm{R}(\mathsf{DISJ}_n) = \Omega(n)$ must necessarily be a *non-product distribution*.

## 8.1 Preliminaries

Before we proceed to prove the Babai-Frankl-Simon lower bound for disjointness, we revisit the equivalence between distributional and randomized communication complexity. We also define and motivate the product distribution we work with.

In the previous lecture, we showed that:

**Lemma 8.2** (Yao's lemma)**.** $\mathrm{R}^{\mathrm{pub}}_{\varepsilon}(f) = \max_{\mu} \mathrm{D}^{\mu}_{\varepsilon}(f)$.

**Corollary 8.3.** *For $c > 0$, $\mathrm{R}^{pub}_{\varepsilon}(f) \geq c \iff \exists \mu : \mathrm{D}^{\mu}_{\varepsilon}(f) \geq c$.*

Therefore, to show a lower bound for $\mathrm{R}(f)$, it is sufficient to exhibit a suitable $\mu$ for which $\mathrm{D}^{\mu}(f)$ is large. Our next task is to choose such a good $\mu$ for $\mathsf{DISJ}_n$.

### 8.1.1 Choosing a good distribution for $\mathsf{DISJ}_n$

Let us first formalize what we mean by a *product* distribution.

**Definition 8.4** (product distribution)**.** *A distribution $\mu : 2^{[n]} \times 2^{[n]} \to [0,1]$ is called a product distribution if there exist distributions $\lambda, \rho : 2^{[n]} \to [0,1]$, such that $\mu(X,Y) = \lambda(X) \times \rho(Y)$.*

Recall that we used the *uniform* distribution $\mathsf{unif}$ for showing a lower bound for the inner product function $\mathsf{IP}$. Would $\mathsf{unif}$ work for $\mathsf{DISJ}_n$ too? Observe that $\Pr_{(X,Y)\sim\mathsf{unif}}[X \cap Y = \emptyset] \approx (3/4)^n$. This implies that under the measure $\mathsf{unif}$, the truth table $M_{\mathsf{DISJ}_n}$ is highly biased towards $0's$ (at least for large $n$). This implies there exists an $O(1)$ protocol (that does nothing but output 0) that works for $\mathsf{DISJ}_n$ under the uniform distribution. The reason for this bias towards "intersecting sets" is due to the large set size in $\mathsf{unif}$. We can salvage the uniform distribution by allowing its support to be suitably-sized sets only. Formally, our product distribution is $\mu = \lambda \times \rho$ where $\mu, \lambda$ are uniform over subsets of $[n]$ of size $\sqrt{n}$:

$$\lambda(S) = \rho(S) = \begin{cases} 1/\binom{n}{\sqrt{n}} & \text{if } |S| = \sqrt{n} \\ 0 & \text{otherwise} \end{cases}.$$

Under this measure, $M_{\mathsf{DISJ}_n}$ is not biased since it can be easily seen that:

$$\Pr_{(X,Y)\sim\mu}[X \cap Y = \emptyset] \approx \left(\frac{n-\sqrt{n}}{n}\right)^{\sqrt{n}} \approx \frac{1}{e}. \tag{8.1.1}$$

Next, we will show that $\mathrm{D}^{\mu}(\mathsf{DISJ})$ is large for the above-mentioned product distribution.

## 8.2 Babai-Frankl-Simon Theorem

**Theorem 8.5** (Babai-Frankl-Simon [BFS86])**.**

$$\mathrm{R}(\mathsf{DISJ}_n) = \Omega(\sqrt{n}).$$

This theorem is proved using the corruption bound, which proceeds as follows. Any $(\varepsilon,\mu)$-deterministic protocol[1] for $\mathsf{DISJ}_n$ partitions $M_{\mathsf{DISJ}_n}$ into "almost" monochromatic rectangles. We show that any such "almost" monochromatic rectangle in $M_{\mathsf{DISJ}_n}$ must be small (Lemma 8.6). This would imply that the number of such rectangles is large, which would imply the large distributional complexity of $\mathsf{DISJ}_n$ and prove Theorem 8.5.

**Lemma 8.6.** *There exists a constant $c > 0$ such that for all sufficiently large $n$, the following holds. If $R = S \times T$ is an almost $\varepsilon$-unbalanced 1-rectangle i.e.,*

$$\Pr_{(X,Y)\sim\mu}[\mathsf{DISJ}_n(X,Y) = 0 \mid (X,Y) \in R] \leq \varepsilon$$

*then either $|S|$ or $|T|$ is at most $2^{-c\sqrt{n}}\binom{n}{\sqrt{n}}$.*

[1] An $(\varepsilon,\mu)$-deterministic protocol for a function $f$ is a deterministic protocol that makes error at most $\varepsilon$ under the distribution $\mu$.

*Proof.* The proof idea is as follows: we can assume wlog. that $|S|$ is large (else the lemma is trivial). Then, $S$ contains a lot of sets which are almost disjoint with each other (Claim 8.7). Hence, the sets in $S$ span a large fraction of the universe $[n]$. Then, it cannot be the case that there exist a large number of sets which are disjoint with most of the sets in $S$. Hence, $T$ is small. The formal proof is as follows.

Assume that $|S| > 2^{-c'\sqrt{n}}\binom{n}{\sqrt{n}}$ (where $c'$ is as in Claim 8.7 below). Define $S' = \{x \in S :$ *set $x$ intersects with at most $2\varepsilon$ fraction of sets $y$ in $T$* $\}$. We note that $|S'| \geq \frac{|S|}{2}$ (by an averaging argument). Now we claim that

**Claim 8.7.** $\exists c' > 0$ *: if* $|S'| > (\frac{1}{2})2^{-c'\sqrt{n}}\binom{n}{\sqrt{n}}$*, then for* $k = \frac{\sqrt{n}}{3}$*, there exist* $x_1, \ldots, x_k \in S'$ *such that* $\forall i = 1 \ldots k$,

$$\left| x_i \cap \bigcup_{j<i} x_j \right| < \frac{\sqrt{n}}{2}$$

Define $S'' = \{x_1, \ldots, x_k\}$. Claim 8.7 states that each set $x_i \in S''$ brings in at least $\sqrt{n}/2$ distinct elements. We postpone the proof of the claim. Clearly, $\left|\bigcup_{x \in S} x\right| \geq (\sqrt{n}/3) \times (\sqrt{n}/2) = n/6$. Hence, the sets in $S''$ span a good fraction of the universe $U = [n]$. We work with $S''$ henceforth.

Define $T' = \{y \in T :$ *set $y$ intersects with at most $4\varepsilon$-fraction of sets $x_i$ in $S''$*$\}$. Since $S'' \subseteq S'$, from the definition of $S'$ and an averaging argument, it follows that that $|T'| \geq \frac{|T|}{2}$. Let us now (upper) bound the size of set $T'$. Recall that every set $y \in T'$ intersects at most $4\varepsilon k$ sets in $S''$. Thus for every set $y \in T'$, there exists $(1 - 4\varepsilon)k$ sets $x_{i_1}, \ldots, x_{i_{(1-4\varepsilon)k}} \in S''$ such that $y \subseteq [n] \setminus \bigcup_j x_{i_j}$. Observe that $|[n] \setminus \bigcup_j x_{i_j}|$ is at most $n - (1 - 4\varepsilon)k\sqrt{n}/2 \leq n - n/9 = 8n/9$ (assuming $\varepsilon \leq 1/100$). We can thus upper bound the number of different possible $y$ in $T'$ by

$$\binom{k}{4\varepsilon k}\binom{8n/9}{\sqrt{n}} = \binom{\sqrt{n}/3}{4\varepsilon \cdot \sqrt{n}/3}\binom{8n/9}{\sqrt{n}} < 2^{-1-c''\sqrt{n}}\binom{n}{\sqrt{n}},$$

where $c''$ is a suitable constant (obtained from the Stirling's approximation). Hence,

$$|T| \leq 2|T'| < 2^{-c''\sqrt{n}}\binom{n}{\sqrt{n}}$$

Set $c = \min(c', c'')$. Consequently, if $|S| > 2^{-c\sqrt{n}}\binom{n}{\sqrt{n}}$, then $|S| > 2^{-c'\sqrt{n}}\binom{n}{\sqrt{n}}$ as well. We have already shown that this implies $|T| < 2^{-c''\sqrt{n}}\binom{n}{\sqrt{n}} < 2^{-c\sqrt{n}}\binom{n}{\sqrt{n}}$. Therefore, either $|S|$ or $|T|$ must be at most $2^{-c\sqrt{n}}\binom{n}{\sqrt{n}}$. Hence proved. □

We now prove Claim 8.7.

*Proof.* We will prove this claim by showing inductively that for every $l < \sqrt{n}/3$ and every choice of $l$ sets $x_1, \ldots, x_l$ in $S'$, there exist another set $x^*$ in $S'$ (since $S'$ is large) such that $\left|x^* \cap \bigcup_{i \leq l} x_i\right| < \sqrt{n}/2$. Suppose we have only $l < \sqrt{n}/3$ sets $x_1, \ldots, x_l$. Then, the size of

their union $Z = \bigcup_{i \le l} x_i$ is at most $\sqrt{n} \times \sqrt{n}/3 = n/3$. Now, we bound the number of $x \in S'$ such that $|x \cap Z| \ge \sqrt{n}/2$. Number of such $x's$ must be at most

$$\begin{aligned}
\sum_{i=\sqrt{n}/2}^{\sqrt{n}} \binom{n/3}{i}\binom{2n/3}{\sqrt{n}-i} &\le n\binom{n/3}{\sqrt{n}/2}\binom{2n/3}{\sqrt{n}/2} \\
&< 2^{-1-c'\sqrt{n}}\binom{n}{\sqrt{n}} \qquad [c' \text{ from Stirling's approximation}] \\
&< |S'|
\end{aligned}$$

Therefore, there exists a choice of $x^* \in S'$ such that $|x^* \cap Z| < \sqrt{n}/2$. Repeating this argument for each $l < \sqrt{n}/3$ proves the claim. □

Having proved Lemma 8.6, we are ready to prove Theorem 8.5 via the corruption bound argument.

*Proof.* From Lemma 8.6, we can conclude that for any $(\varepsilon, \mu)$-deterministic protocol for $\mathsf{DISJ}_n$, the $\mu$-measure of the largest "almost" monochromatic rectangle $R = S \times T$ is bounded as follows. Wlog. assume $|T| \le |S|$. Hence, if $R = S \times T$ is an almost $\varepsilon$-unbalanced 1-rectangle, we have $|T| \le 2^{-c\sqrt{n}}\binom{n}{\sqrt{n}}$.

$$\mu(R) = \sum_{s \in S, t \in T} \mu(s,t) \qquad = \left(\sum_{s \in S} \lambda(s)\right) \cdot \left(\sum_{t \in T} \rho(t)\right) \qquad \le 1 \cdot \frac{|T|}{\binom{n}{\sqrt{n}}} \qquad \le 2^{-c\sqrt{n}}.$$

Thus, every unbalanced 1-rectangle has $\mu$-measure at most $2^{-c\sqrt{n}}$. Recall from (8.1.1) that $\mu$ is a distribution that puts at least a constant mass on the 1-inputs of $\mathsf{DISJ}_n$, i.e., $\mu(\mathsf{DISJ}_n^{-1}(1)) \approx 1/e$. Applying the corruption bound (proved in Problem Set 2), we obtain

$$2^{\mathrm{D}^{\mu}_{\delta}(\mathsf{DISJ}_n)} \ge 2^{c\sqrt{n}} \cdot \left(\mu(\mathsf{DISJ}_n^{-1}(1)) - \delta \cdot \frac{\varepsilon}{1-\varepsilon}\right).$$

We thus, obtain that for small enough $\delta$, we have $\mathrm{R}_{\delta}(\mathsf{DISJ}_n) \ge \mathrm{D}^{\mu}_{\delta}(\mathsf{DISJ}_n) = \Omega(\sqrt{n})$. □

## 8.3 Limitations of Product Distributions

The next question is: can we improve the $\Omega(\sqrt{n})$ lower bound for $\mathrm{D}^{\mu}(\mathsf{DISJ}_n)$ for the distribution $\mu$ used above? The following result tells us that it is unlikely. (recall that $\mathsf{DISJ}_{k,n}$ is the disjointness problem restricted to sets of size $k$.)

**Theorem 8.8** (Håstad-Widgerson [HW07])**.**

$$\mathrm{R}^{pub}_{1/3}(\mathsf{DISJ}_{k,n}) = O(k)$$

Using this result and Corollary 8.3, we infer that for all distributions $\sigma$, $\mathrm{D}^{\sigma}_{\varepsilon}(\mathsf{DISJ}_{\sqrt{n},n}) = O(\sqrt{n})$. Since the distribution $\mu$ considered in the previous section is supported only on sets of size $\sqrt{n}$, we have that

$$\mathrm{D}^{\mu}(\mathsf{DISJ}_{\sqrt{n},n}) = O(\sqrt{n}).$$

In fact, now we will show that we cannot use any product distribution (even ones supported on sets of larger size) to obtain (significantly) better lower bounds for $\mathsf{DISJ}$.

**Theorem 8.9** ([BFS86])**.** *For all product distributions $\mu$,*

$$\mathrm{D}^{\mu}(\mathsf{DISJ}_n) = O(\sqrt{n}\log n).$$

*Proof.* Suppose that the inputs are distributed according to the product distribution $\mu = \lambda \times \rho$. We will a public coins randomized protocol $P$ for $\mathsf{DISJ}_n$ under the product distribution $\mu$ with expected error at most $\varepsilon$ and expected communication cost $O(\sqrt{n}\log n)$. The expectation (for both error and communication cost) is over both the public random coins as well as the distribution of inputs. Given such a public coins protocol, by fixing random coins, one obtains a $(\mu, \varepsilon)$-deterministic protocol for $\mathsf{DISJ}_n$ of cost $O(\sqrt{n}\log n)$, thus proving Theorem 8.9.

The public coin randomized protocol $P$ is described below.

Alice's Input: $x \subseteq [n]$ and Bob's Input: $y \subseteq [n]$.

1. Alice and Bob set $U \leftarrow [n]$.
2. Alice and Bob both have access to a public random stream of the form $R = X_1, X_2, \ldots$, where each $X_i$ is interpreted as follows. Each $X_i \subseteq [n]$ is distributed independently according to the distribution $\lambda|_{|X|>\sqrt{n}}$. (i.e, the distribution $\lambda$ conditioned that $X$ is of size at least $\sqrt{n}$.)
3. Repeat the following phase till a result is declared:
   (a) Alice: If $|x| \leq \sqrt{n}$, then Alice sends $x$ to Bob using $\sqrt{n}\log n$ bits, else she informs Bob (using constant bits) that $|x| > \sqrt{n}$.
   (b) Bob:
      i. If Alice has sent the set $x$, then Bob compares it with $y$ and declares 'Disjoint' or 'Intersecting' accordingly.
      ii. If Alice instead informs Bob that "$|x| > \sqrt{n}$", then Bob checks if:
      $$\varepsilon_y = \Pr_{X\sim\lambda}[X \cap y = \emptyset \mid |X| > \sqrt{n}] \leq \varepsilon/\sqrt{n} \tag{8.3.1}$$
         A. If (8.3.1) is true, then Bob declares 'Intersecting'.
         B. If (8.3.1) is not true, then Bob identifies the minimum $j$ such that the set $X_j$ in the random stream $R$ is disjoint with $y$. Bob sends the index $j$ to Alice.
   (c) Alice receives $j$ and updates $x' = x \setminus X_j$. *(Remark: $x \cap y = \emptyset$ iff $x' \cap y = \emptyset$)*
   (d) Alice updates her set $x \leftarrow x'$. Both Alice and Bob update $U \leftarrow U \setminus X_j$, the distribution $\lambda \leftarrow \lambda|_{X\subseteq U}$ and start the next phase with a fresh random-coins-stream $R = X_1, X_2, \ldots$, of sets distributed identically according to $\lambda|_{|X|>\sqrt{n}}$.

Since the random stream is supported on $\sqrt{n}$-sized subsets $U$, on every completion of the phase the set $U$ drops in size by at least $\sqrt{n}$. Hence, the protocol has at most $\sqrt{n}$ phases. Though the above protocol was described for a specific input pair $(x, y)$, we will carry out the analysis averaged over input pairs $(X, Y)$ distributed according to $\mu$.

**Expected Error:** In any phase, error occurs only in Step 3(b)iiA of Bob's response i.e. when $X$ and $Y$ are disjoint, but Bob declares them to be intersecting based on the probability calculation in (8.3.1). Otherwise, correctness of the protocol stems from maintaining the invariant corresponding to disjointness/intersection of $x$ and $y$ in Step 3c. Averaged over Alice's input $X$, the error in Step 3(b)iiA is at most $\varepsilon/\sqrt{n}$ for a particular phase due to (8.3.1). Taking union bound over the $\sqrt{n}$ phases, the total error is at most $\sqrt{n} \cdot \varepsilon/\sqrt{n} = \varepsilon$.

**Expected Communication:** We will bound the expected cost of the protocol as follows. If in any phase, Alice sends her set of size at most $\sqrt{n}$ to Bob, then the protocol terminates and the communication cost for that phase is $O(\sqrt{n}\log n)$. Next, we bound the expected cost $c_i$ incurred in $i^{th}$ phase for communicating $J = j(X, Y; R)$ to Alice in Step 3(b)iiB of the protocol. Note that it costs $O(\log J)$ bits to communicate the integer $J$.

$$c_i = \mathop{\mathbb{E}}_{R,X\sim\lambda,Y\sim\rho}[O(\log J)] \leq O\left(\mathop{\mathbb{E}}_{Y\sim\rho}\left[\log\left(\mathop{\mathbb{E}}_{X\sim\lambda,,R}[j]\right)\right]\right),$$

where the last inequality follows from Jensen's inequality for the log function. We can bound the expected value fo $j$ over random $X$ and $R$ (for a fixed value of $y$) using (8.3.1) as follows.

$$\mathop{\mathbb{E}}_{X\sim\lambda,R}[j] = \frac{1}{\Pr_{X\sim\lambda}[X \textit{ and } y \textit{ are disjoint} \mid |X| > \sqrt{n}]} = 1/\varepsilon_y \leq \frac{\sqrt{n}}{\varepsilon}.$$

Hence,

$$c_i = O\left(\mathop{\mathbb{E}}_{Y\sim\rho}\left[\log\left(\frac{\sqrt{n}}{\varepsilon}\right)\right]\right) = O\left(\log n + \log\left(\frac{1}{\varepsilon}\right)\right).$$

Taken over all $\sqrt{n}$ phases, this cost is $O(\sqrt{n}\log n)$. Therefore, the overall communication cost for the protocol is $O(\sqrt{n}(\log n + \log(1/\varepsilon))$. □

## References


[BFS86] László Babai, Peter Frankl, and Janos Simon. *Complexity classes in communication complexity theory (preliminary version).* In *Proc. 27th IEEE Symp. on Foundations of Comp. Science (FOCS)*, pages 337–347. 1986. doi:10.1109/SFCS.1986.15.

[HW07] Johan Håstad and Avi Wigderson. *The randomized communication complexity of set disjointness.* Theory of Computing, 3(1):211–219, 2007. doi:10.4086/toc.2007.v003a011.

[KS92] Bala Kalyanasundaram and Georg Schnitger. *The probabilistic communication complexity of set intersection.* SIAM J. Discrete Math., 5(4):545–557, 1992. (Preliminary Version in *2nd Structure in Complexity Theory Conference*, 1987). doi:10.1137/0405044.

[Raz92] Alexander A. Razborov. *On the distributional complexity of disjointness.* Theoretical Comp. Science, 106(2):385–390, 1992. doi:10.1016/0304-3975(92)90260-M.