Correlation in Hard Distributions in Communication Complexity

Ralph Christian Bottesch[1], Dmitry Gavinsky*[2], and Hartmut Klauck†[1,3]

1 **Nanyang Technological University**
50 Nanyang Avenue, Singapore 639798
2 **Institute of Mathematics, Czech Academy of Sciences**
Praha, Czech Republic
3 **Centre for Quantum Technologies, National University of Singapore**
Block S15, 3 Science Drive 2, Singapore 117543

Abstract

We study the effect that the amount of correlation in a bipartite distribution has on the communication complexity of a problem under that distribution. We introduce a new family of complexity measures that interpolates between the two previously studied extreme cases: the (standard) randomised communication complexity and the case of distributional complexity under product distributions.

- We give a tight characterisation of the randomised complexity of Disjointness under distributions with mutual information k, showing that it is $\Theta(\sqrt{n(k+1)})$ for all $0 \le k \le n$. This smoothly interpolates between the lower bounds of Babai, Frankl, Simon [4] for the product distribution case ($k = 0$), and the bound of Razborov [22] for the randomised case. The upper bounds improve and generalise what was known for product distributions, and imply that any tight bound for Disjointness needs $\Omega(n)$ bits of mutual information in the corresponding distribution.
- We study the same question in the distributional *quantum* setting, and show a lower bound of $\Omega((n(k+1))^{1/4})$, and an upper bound (via constructing communication protocols), matching up to a logarithmic factor.
- We show that there are total Boolean functions f_d that have distributional communication complexity $O(\log n)$ under all distributions of information up to $o(n)$, while the (interactive) distributional complexity maximised over all distributions is $\Theta(\log d)$ for $n \le d \le 2^{n/100}$. This shows, in particular, that the correlation needed to show that a problem is hard can be much larger than the communication complexity of the problem.
- We show that in the setting of one-way communication under product distributions, the dependence of communication cost on the allowed error ϵ is multiplicative in $\log(1/\epsilon)$ – the previous upper bounds had the dependence of more than $1/\epsilon$. This result, for the first time, explains how one-way communication complexity under product distributions is stronger than PAC-learning: both tasks are characterised by the VC-dimension, but have very different error dependence (learning from examples, it costs more to reduce the error).

1998 ACM Subject Classification F.1.3 Complexity Measures and Classes

Keywords and phrases communication complexity, information theory

Digital Object Identifier 10.4230/LIPIcs.APPROX-RANDOM.2015.544

* Partially funded by the grant P202/12/G061 of GA ČR and by RVO: 67985840. Most of this work was done while DG was visiting the Centre for Quantum Technologies at the National University of Singapore, and was partially funded by the Singapore Ministry of Education and the NRF.

† This work is funded by the Singapore Ministry of Education (partly through the Academic Research Fund Tier 3 MOE2012-T3-1-009) and by the Singapore National Research Foundation.

18th Int'l Workshop on Approximation Algorithms for Combinatorial Optimization Problems (APPROX'15) /
19th Int'l Workshop on Randomization and Computation (RANDOM'15).
Editors: Naveen Garg, Klaus Jansen, Anup Rao, and José D. P. Rolim; pp. 544–572
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

The standard way to attack the problem of showing a lower bound on the randomised communication complexity of a function f is to choose a probability distribution μ on the inputs, and then show that the deterministic distributional complexity is large for f w.r.t. μ – i.e., that any deterministic protocol that computes f with small error under μ must communicate much. This approach eliminates the need to argue about the randomness used by the protocol.[1]

It is well known that this approach can be used without loss of generality, due to von Neumann's minimax theorem (see [20]; the same principle applies to many nonuniform computational models):

$$\max_{\mu} D^{\mu}_{\epsilon}(f) = R_{\epsilon}(f),$$

where $D^{\mu}_{\epsilon}(f)$ denotes the deterministic complexity of protocols that compute f with error ϵ under the distribution μ of input to f, and $R_{\epsilon}(f)$ is the public coin randomised communication complexity of f with worst-case error ϵ.[2]

As a matter of convenience, one first tries to use a simple distribution μ, for instance the uniform distribution, or more generally, product distributions over the inputs to Alice and Bob. This works for some problems, like Inner Product modulo 2 [7]. However, Babai, Frankl, and Simon [4] observed that for the Disjointness problem DISJ one cannot obtain lower bounds larger than $\Omega(\sqrt{n}\log n)$ under *any* product distribution, i.e., they show that an upper bound of $O(\sqrt{n}\log n)$ holds for every product distribution. They also give a lower bound of $\Omega(\sqrt{n})$ under a product distribution. Later, Kalyanasundaram and Schnitger [16] obtained the tight $\Theta(n)$ bound, and Razborov [22] showed that indeed $D^{\mu}_{\epsilon}(DISJ) = \Theta(n)$ for an explicit simple distribution μ, for any sufficiently small constant $\epsilon > 0$ (that such a μ exists is immediate from the result in [16] and the minimax theorem, but their proof does not exhibit such a distribution explicitly). Distributional complexity under product distributions has been also frequently used to show structural properties like direct product theorems (e.g., [15, 12]). Furthermore, distributional communication complexity is the natural average case version of communication complexity, and it makes sense to study this for distributions that are 'easy', in order to get a different model than randomised complexity. It seems natural to measure "easiness" via mutual information.

For many years it was open how large the gap between $R^{I=0}_{\epsilon}(f) = \max_{\mu \text{ product}} D^{\mu}_{\epsilon}(f)$ and $R_{\epsilon}(f)$ (for constant $\epsilon > 0$) can be. Sherstov [25] finally gave a proof that there are total Boolean functions f, where the former is $O(1)$ and the latter is $\Omega(n)$. In his result f is not given explicitly. Very recently Alon et al. [2] give the following optimal explicit separation. Consider the problem where Alice gets a point and Bob a line from a projective plane containing $2^{\Theta(n)}$ points and lines. In this case the VC-dimension of the projective plane is at most 2, which implies that the distributional complexity under any product distribution is at most $O(1)$ (even for one-way protocols), whereas the sign-rank of the communication matrix is $2^{\Omega(n)}$, and hence the randomised (even unbounded error) communication complexity is $\Omega(n)$.

This leaves open a more precise investigation of the *amount* of correlation in μ needed to make $D^{\mu}(f)$ equal to $R(f)$. It is natural to quantify this via the mutual information

[1] We note that the popular information complexity method (see e.g.[5]) also uses distributional complexity, but does not seek to eliminate randomness from protocols.

[2] Throughout the paper we do not consider private coin randomised protocols.

$I(X:Y)$, when the input (X,Y) is drawn from μ. We define the following measure:

$$R_\epsilon^{I\leq k}(f) = \max_{\mu: I(X:Y)\leq k} D_\epsilon^\mu(f).$$

We note here that the quantity on the right hand side does not change if randomised or deterministic protocols are allowed, because in the distributional setting the randomness can be fixed without increasing the error (under any distribution). The investigation of this measure has been initiated by Jain and Zhang [14] in the setting of one-way communication complexity (we discuss their contribution at the end of Section 1.3). We note that $R_\epsilon^{I\leq n}(f) = R_\epsilon(f)$ for all functions $f : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$.

This family of complexity measures allows us to investigate how much correlation is needed in the input distribution to get good lower bounds. We have 3 main applications. First, we closely investigate the case of the Disjointness problem. Second, we show that a certain problem exhibits a threshold behaviour, i.e., only with almost maximal correlation can a tight lower bound be proved, and this correlation can also be larger than the actual communication complexity of the problem. Third, we investigate the dependence of one-way communication complexity under product distributions on the allowed error.

1.1 The Disjointness problem

In the *Disjointness problem (DISJ)*, Alice and Bob receive, respectively, subsets $x, y \subseteq \{1,\ldots,n\}$, and their task is to decide whether x and y are disjoint. This is one of the most-studied problem in communication complexity, which arguably has the biggest number of known applications to other models (see [20]). We give a complete characterisation of the information-bounded distributional complexity of Disjointness for all values of $k = I(X:Y)$, both in the randomised and in the quantum case.

▶ **Theorem 1.** *For all* $0 \leq k \leq n$ *and constant* ϵ *we have*

1. $R_\epsilon^{I\leq k}(DISJ) = \Theta(\sqrt{n(k+1)})$.
2. $Q_\epsilon^{I\leq k}(DISJ) = \tilde{O}((n(k+1))^{1/4})$.
3. $Q_\epsilon^{I\leq k}(DISJ) = \Omega((n(k+1))^{1/4})$.

Previously, a lower bound of $\Omega(\sqrt{n})$ was known for a product distribution [4], and the $\Omega(n)$ lower bound by Razborov [22] uses a distribution μ with $I^\mu(X:Y) = \Theta(n)$. Babai et al. [4] also gave an upper bound of $O(\sqrt{n}\log n)$ for the case of product distributions, which we improve by a log-factor. Our results interpolate between the previously-known extreme cases, and also show that one needs input correlation $\Omega(n)$ to prove tight lower bounds. Interestingly, the bounds depend inverse-polynomially on the error probability, except for the extreme cases of zero correlation and of maximal correlation. We also note that a nearly-optimal complexity for randomised protocols can be achieved in a protocol with two rounds of communication (though not in one round).

The tight bound in the randomised case is based on a two-phase protocol, in which the players first remove "uninteresting" elements from their sets, until they are (essentially) small enough to be communicated. For the quantum case this two-phase approach cannot be optimal, because the first phase reveals "too much" information about the input. Therefore we give a completely different protocol for the quantum case, in which the players identify uninteresting elements a priori. This approach is tight up to a log-factor.

1.2 Mutual information in hard distributions

Note that for DISJ the complexity increases with the information parameter, and the randomised communication complexity bound $\Theta(n)$ is reached only once the information in the hard distribution reaches $\Omega(n)$. For other problems like Inner Product mod 2 the tight bound of $\Omega(n)$ is reached already under product distributions [7]. But can the mutual information between the input sides that is required to show a tight lower bound ever be *larger* than the actual communication complexity? I.e., is it ever necessary to use distributions that are (much) more strongly correlated than the communication lower bound we want to show, or is it always possible to prove a tight lower bound for a (total) function f by using a hard distribution with $I(X:Y) \le poly(R(f))$? A weak example is the quantum complexity of Disjointness, where the tight $\Omega(\sqrt{n})$ bound is only reached when the information reaches $\Omega(n)$, but even here the complexity increases gradually with the information. We resolve this question, although our example is not explicit.

▶ **Theorem 2.** *For every $n \le d \le 2^{n/100}$ there is a function $f_d : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$ that has $R(f_d) = \Theta(\log d)$, but under all bipartite distributions with mutual information less than $n/1000$ the communication bound is $R^{I \le n/1000}_{1/10}(f_d) \le O(\log n)$.*

Hence for f_d the complexity stays low until the information is almost maximal, and then shoots up.

1.3 Dependence of $R^{A\to B, I=0}_{\epsilon}(f)$ on ϵ [3]

Finally, we investigate the error dependence of $R^{I\le k}_{\epsilon}(f)$ for arbitrary f. In the unrestricted case, by standard boosting techniques we have $R_\epsilon(f) \le O(R_{1/3}(f) \cdot \log(1/\epsilon))$. We call a function f and a class C of distributions on the inputs with $\max_{\mu\in C} D^{\mu}_{\epsilon}(f) \le O(\max_{\mu\in C} D^{\mu}_{1/3}(f) \cdot \log(1/\epsilon))$ *boost-able*. For this definition we require the above to be true for all ϵ. One can easily show that there are distributions μ and functions f, such that e.g. $D^{\mu}_{1/4}((f) = \Omega(n)$ and $D^{\mu}_{1/3}(f) = 0$, by placing a hard problem with weight 1/3 in an otherwise constant matrix, so for a fixed distribution μ one cannot in general expect the error dependence to behave nicely.

Boost-ability is a property of a class of distributions. The class of all distributions clearly has the property, but what about the class of distributions with information at most I? In particular, what about $I = 0$?

The issue is particularly interesting for product distributions, because boost-ability can be used to derive upper bounds on $R^{I\le k}(f)$ from upper bounds on $R^{I=0}(f)$: due to the substate theorem (Fact 4 below), a protocol that solves f under all product distributions with error $\epsilon 2^{-9k/\epsilon}$ can be used to solve f under distributions with $I(X:Y) = k$ with error ϵ, hence boost-ability implies $R^{I\le k}_{\epsilon}(f) \le O((k+1) \cdot R^{I=0}_{1/3}(f)/\epsilon)$.

We will use the super-script "$A \to B$" to denote one-way communication. In this model the class of product distributions is boost-able:

▶ **Theorem 3.** $R^{A\to B,I=0}_{\epsilon}(f) \le O(R^{A\to B,I=0}_{1/3}(f) \cdot \log(1/\epsilon))$.

We also show that when the information is between 1 and $n^{1-\Omega(1)}$, then neither randomised nor distributional quantum protocols are, in general, boost-able, see our Corollaries 20 and 27.

[3] The same result has been obtained recently by Molinaro et al. [21] independently. The methods being used in the two works are similar; [21] has been published prior to the current publication, while our results have been presented during a public talk prior to either publication.

It is well known that $R_{1/3}^{A\to B,I=0}(f)=\Theta(VC(f))$ [19], where $VC(f))$ is the VC-dimension of the set of rows of the communication matrix. This even extends to the quantum case [3, 18]. The VC-dimension is also known to characterise the hardness of PAC-learning (see the monograph by Kearns and Vazirani [17]) – in fact, the previous proofs of the upper bound on $R_{\epsilon}^{A\to B,I=0}(f)$ in terms of VC-dimension have been done by explicitly simulating learning algorithms in the one-way communication model: Random examples are generated using a public coin, and Alice classified the examples in order to teach Bob a row of the communication matrix of f in the PAC sense (examples were generated from the public coin, and Alice labelled those examples spending 1 bit per example).

The main limitation of this approach is that for PAC learning one needs $\Omega(1/\epsilon)$ examples to achieve error $1/\epsilon$. On the other hand, this approach ignores two strengths of the one-way model: First, Alice and Bob know the underlying distribution; second, Alice can do more than simply label examples. One can interpret the one-way communication model under product distributions as a learning model, in which Alice is an (old-fashioned) teacher, who teaches by monologue, but using shared randomness that does not count towards the communication. Does such a teacher offer any advantage over learning from random examples? At first glance no, since both models are characterised by the VC-dimension, and one could conclude that learning from experience is all it takes. Our Theorem 3, however, shows that the final error can be made much smaller when learning from a teacher, comparing to learning "just from experience". Note that in practice $1/\epsilon$ can also easily become the dominating factor in the complexity of a learning algorithm.

The main idea in our protocol is that Alice and Bob can beforehand agree on an ϵ-net among the rows of the communication matrix, and Alice simply sends the name of the nearest row in the net. During a PAC learning algorithm, on the other hand, the ϵ-net is generated from examples, which is more costly.

We can now discuss the previous result of Jain and Zhang [14]. They show that for all total Boolean functions f in the one-way model:

$$R_{\epsilon}^{A\to B,I\leq k}(f)\leq O((k+1)\cdot R_{1/3}^{A\to B,I=0}(f)\cdot 1/\epsilon^{2}\cdot\log(1/\epsilon)).$$

This extends the VC-dimension upper bound to distributions with nonzero information. Their protocol for information-k distributions is constructed by simulating the PAC learning algorithm for the row x, and by generating examples $y', f(x,y')$ using a rejection-sampling protocol. We can improve the error dependence to $1/\epsilon$ by the following idea. Due to the Substate Theorem (Fact 4 below) it is enough to find a protocol that has error $2^{-9k/\epsilon}$ under the product of the marginal distributions of a distribution μ (with information k). But this can be achieved with communication $O((k+1)/\epsilon\cdot R_{1/3}^{A\to B,I=0}(f))$ according to Theorem 3.

2 Preliminaries and Definitions

2.1 Information Theory

We refer to [8] for standard definitions concerning information theory.

The relative entropy of two distributions on a discrete support is denoted by $D(\rho||\sigma)$. The relative max-entropy is $D_{\infty}(\rho||\sigma)=\max_{x}\log(\rho(x)/\sigma(x))$. Note that these quantities are infinite, if the support of σ does not contain the support of ρ. We mostly consider bipartite distributions on $\{0,1\}^{n}\times\{0,1\}^{n}$. The mutual information is $I(X:Y)=D(\mu||\mu_{X}\times\mu_{Y})$, where μ is the joint distribution of (X,Y) and μ_{X}, and μ_{Y} are the two marginal distributions of μ. We also use the quantity $I_{\infty}(X:Y)=D_{\infty}(\mu||\mu_{X}\times\mu_{Y})$. If we want to indicate the distribution used we write its name as a superscript, like $I^{\mu}(X:Y)$.

We first state the following well-known fact, see [13].

▶ **Fact 4** (Substate Theorem).

1. $I(X:Y) \leq I_\infty(X:Y)$.
2. *For a given μ there is a μ' with $||\mu - \mu'|| \leq \epsilon$, and $I_\infty^{\mu'}(X:Y) \leq I^\mu(X:Y) \cdot 4/\epsilon$, where $||\mu - \mu'||$ is the total variation distance between μ and μ'.*

We will use the following lemmas and facts. The first follows from the definition of relative entropy.

▶ **Lemma 5.** *Let μ be a bipartite distribution, $\rho = \mu_A \times \mu_B$, and $\sigma = \sigma_A \times \sigma_B$ any product distribution.*

Then $D(\mu||\sigma) = D(\mu||\rho) + D(\rho||\sigma) = I^\mu(X:Y) + D(\rho||\sigma)$.

The following is a consequence of the log-sum inequality.

▶ **Lemma 6.** *Let μ, σ be distributions (for concreteness on $\{0,1\}^n \times \{0,1\}^n$), and E an event. Then we have that $\sum_{x,y\in E} \mu(x,y)\log(\mu(x,y)/\sigma(x,y)) \geq \max\{-1, \mu(E)\log(\mu(E)/\sigma(E))\}$.*

▶ **Lemma 7.** *Let μ be a distribution on $\{0,1\}^n \times \{0,1\}^n$, E an event, and μ' the distribution μ restricted to E. Furthermore, assume that under μ we have that $Prob(E) = \alpha$. Then $D(\mu'||\sigma) \leq (D(\mu||\sigma) + 1)/\alpha - \log\alpha$.*

Proof. For all $x, y \in E$ we have $\mu'(x,y) = \mu(x,y)/\alpha$, otherwise $\mu'(x,y) = 0$.

$$\begin{aligned}
& D(\mu||\sigma) \\
= \;& \sum_{x,y} \mu(x,y)\log(\frac{\mu(x,y)}{\sigma(x,y)}) \\
\overset{(*)}{\geq} \;& \sum_{x,y\in E} \mu(x,y)\log(\frac{\mu(x,y)}{\sigma(x,y)}) - 1 \\
\geq \;& \sum_{x,y\in E} \mu'(x,y)\cdot\alpha\cdot\log(\frac{\mu'(x,y)\cdot\alpha}{\sigma(x,y)}) - 1 \\
= \;& D(\mu||\sigma)\cdot\alpha + \alpha\log\alpha - 1,
\end{aligned}$$

where for (*) we use Lemma 6 with the event $\{0,1\}^n \times \{0,1\}^n - E$. ◀

We will use the following *rejection sampling* protocol from [10].

▶ **Fact 8.** *Let μ and ν be distributions on $\{0,1\}^n$ with $D(\mu||\nu) = k$. Assume that Alice and Bob both know ν, and can create samples from ν using a public coin. Then Alice can send a message of expected length $k + 2\log k + O(1)$ to Bob, which allows Bob (and Alice) to obtain a shared sample from the distribution μ. The expectation is over the public coin tosses, and Bob's sample is distributed exactly with μ.*

The next lemma follows from a calculation and shows that a distribution can decrease a joint probability compared to the product of marginal distributions only in the presence of mutual information.

▶ **Lemma 9.** *Let X, Y be Boolean random variables with a joint distribution μ and marginal distributions μ_A, μ_B. If $\mu_A(X=1)\mu_B(Y=1) \geq 2\mu(X=Y=1)$, then $I^\mu(X:Y) \geq \mu_A(X=1)\mu_B(Y=1)/4$.*

Finally, we show that this is true for any product distribution, not just the product of marginals.

▶ **Lemma 10.** *Let X, Y be Boolean random variables with a joint distribution μ (and set $\rho = \mu_A \times \mu_B$), and σ any product distribution. If $\sigma_A(X = 1)\sigma_B(Y = 1) \geq 4\mu(X = Y = 1)$ then $D(\mu||\sigma) \geq \sigma(X = Y = 1)/16$.*

Proof. If $\rho(X = Y = 1) \geq \sigma(X = Y = 1)/2$, then by the above lemma $D(\mu||\sigma) \geq D(\mu||\rho) = I^\mu(X : Y) \geq \rho(X = Y = 1)/4 \geq \sigma(X = Y = 1)/8$, because σ is a product distribution and the relative entropy of μ and a product distribution is minimal for ρ. If $\rho(X = Y = 1) \leq \sigma(X = Y = 1)/2$, then we can bound $D(\mu||\sigma) \geq D(\rho||\sigma) = D(\mu_A||\sigma_A) + D(\mu_B||\sigma_B)$. Assume that $\alpha = \mu_A(X = 1) \leq \beta/\sqrt{2} = \sigma_A(X = 1)/\sqrt{2}$. Then $(1-\alpha)\log((1-\alpha)/(1-\beta)) + \alpha\log(\alpha/\beta) \geq \beta/16$. Hence in this case $D(\rho||\sigma) \geq D(\mu_A||\sigma_A) \geq \beta/16 = \sigma_A(X = 1)/16 \geq \sigma_A(X = 1)\sigma_B(Y = 1)/16$. Other cases follow by symmetry. ◀

2.2 Communication Complexity

We assume familiarity with classical and quantum communication complexity. For the former consult [20], the latter is surveyed in [9]. We concentrate on distributional complexity, which we define here.

▶ **Definition 11.** The distributional complexity $D^\mu_\epsilon(f)$ is the minimal worst case communication cost of any deterministic protocol that computes f with error ϵ under μ. Similarly we define $R^\mu_\epsilon(f)$ for randomised public coin protocols and $Q^\mu_\epsilon(f)$ for quantum protocols (we consider quantum protocols with shared entanglement, but do not use the entanglement in our protocols). When we drop the error ϵ from the notation, we set $\epsilon = 1/3$. When we drop the superscript we mean the ordinary, worst-case communication complexity.

We observe that $R^\mu_\epsilon(f) = D^\mu_\epsilon(f)$ for all f, μ, ϵ, because one can fix the public coin randomness without increasing the error. Hence, we adopt the R-notation, and use randomness in upper bounds and deterministic protocols in lower bounds. Note that $Q^\mu_\epsilon(f)$ can be smaller than $R^\mu_\epsilon(f)$, for instance for Disjointness under the hard distribution exhibited by Razborov [22], where $R^\mu(DISJ) = \Theta(n)$, since the quantum complexity of DISJ is at most $O(\sqrt{n})$ [1].

We consider functions $f : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$.

▶ **Definition 12.** Define by $D(k)$ the set of distributions on the inputs that have $I(X : Y) \leq k$.

We define $R^{I\leq k}_\epsilon(f) = \max_{\mu\in D(k)} R^\mu_\epsilon(f)$ and use an analogous definition for the quantum case.

Clearly $R(f) = R^{I\leq n}(f)$ and $R^{I=0}(f)$ is the complexity under the hardest product distribution.

▶ **Definition 13.** One-way protocols allow only a single message from Alice to Bob, who produces the output. We indicate this model by a superscript, like $R^{A\to B, I\leq k}(f)$.

Finally, we note the following fingerprinting technique [20].

▶ **Fact 14.** *There is a public coin protocol that can check equality of strings (of any length) with error $1/2^k$ and communication k.*

3 Randomised Complexity of Disjointness

3.1 Upper Bound

In this section we prove the upper bound for DISJ under bounded information distributions.

First we consider the case of 0 mutual information, for which we show an upper bound of $O(\sqrt{n}\log(1/\epsilon))$. Let μ be a product distribution on the inputs to DISJ. Babai et al. [4] already show a protocol of cost $O(\sqrt{n}\log n\log(1/\epsilon))$ [they do not state the dependence on ϵ, which is however easy to derive from their proof]. Note that one can combine their protocol for product distributions with the Substate Theorem (Fact 4) to get a bound of $O(\sqrt{n}(k+1)\log n/\epsilon)$ on the distributional complexity under distributions with information k: every distribution with information k approximately sits with probability $1/2^{4k/\epsilon}$ inside the product of its marginal distributions, hence it is enough to use a product distribution protocol with very small error. This bound is worse in the dependence on k than what is proved below.

▶ **Theorem 15.** $R_\epsilon^{I=0}(DISJ) \leq O(\sqrt{n}\cdot\log(1/\epsilon))$.

The proof is in the appendix. The main issue here is to achieve the small error dependence. The protocol has a 2-phase structure, where in phase 1, assuming that Bob holds a large set and that the probability that $x\cap y' = \emptyset$ is large, random y' are drawn using the public coin and, if disjoint from x, removed from the universe (initially $\{1,\dots,n\}$). After doing this sufficiently many times, the universe becomes small, and in phase 2 we use the small set disjointness protocol due to Hastad and Wigderson [11].

Now we turn to distributions with more information. The protocol has the same structure, but we need to sample from a distribution of y' that is not independent of x, which takes communication. The protocol also does not have the same error dependence, which we show is unavoidable later. Due to this we may just analyse expected communication, and show that the worst case communication cannot be more than $1/\epsilon$ the established bound by appealing to the Markov bound.

▶ **Theorem 16.** $R_\epsilon^{I\leq k}(DISJ) \leq O(\sqrt{n(k+1)}/\epsilon^2)$.

The proof is in the appendix. The main idea is to follow the 2-phase approach, and shrink the universe until is has size $S = \sqrt{n(k+1)}$. At this point the Hastad-Wigderson small set Disjointness protocol [11] takes over. To shrink the universe we need to sample inputs y' from the distribution conditioned on x, and on being disjoint from x. This is achieved using the rejection sampling protocol of Fact 8. We need to carefully bound the information increase, but on average we remove S elements from the universe with communication cost $O(k/\epsilon)$, and there are at most n/S iterations in phase 1, hence the expected communication is at most $n/S\cdot k/\epsilon$.

In the next section we will also show a lower bound of $\Omega(\sqrt{n/\epsilon})$, so the error dependence cannot be made logarithmic, in contrast to the the 0 information case.

One more issue we would like to consider is the number of rounds used. The above protocol can easily use a large number of rounds, and it is not immediately clear whether this is necessary. It is well known that the complexity of DISJ under product distributions for *one-way protocols* is $\Theta(n)$ [19]. We have the following modification that saves most of the interaction.

▶ **Theorem 17.**

1. *The complexity of DISJ under distributions with information at most k for protocols with 2 rounds is at most $O(\sqrt{n(k+1)}\log n/\epsilon^2)$.*
2. *The complexity of DISJ under distributions with information at most k for $O(\log^* n)$ rounds is at most $O(\sqrt{n(k+1)}/\epsilon^2)$.*
3. *In the case of 0 mutual information, the error dependence drops to a factor of $\log(1/\epsilon)$.*

Proof. For the first item we observe that in phase 1 Alice can simply act as if Bob's set was large, and continue to let him discover y_i''s that are disjoint with x until U_i is guaranteed to be small. This does not increase the bound on the communication. After this Bob can tell Alice, in which 'round' his set really became small, so that she can recover the proper universe U_j. He also sends her his set using $\sqrt{n(k+1)}\log n$ bits. Note that in this protocol only Alice learns the result.

For the second item we do as above, but when Bob's set is small also repeat the same in reverse until both sets are small. Saglam and Tardos [24] have a protocol that solves the small set disjointness problem in phase 2 in $O(\log^* n)$ rounds with communication $O(\sqrt{n(k+1)}\log(1/\epsilon))$.

Finally, note that for product distributions we can use the same modifications to the protocol described in Theorem 15. ◀

3.2 Lower Bound

In this section we prove that the protocol of the previous section is optimal (except regarding the exact dependence on ϵ).

For the lower bound we employ a distribution, depending on n and k, such that the mutual information of the two inputs according to the marginal distributions is at most k; we then prove an $\Omega(\sqrt{n(k+1)})$ lower-bound for the distributional complexity under this distribution. In what follows we consider $k = k(n)$ as being $\in o(n)$, since for $k = \Omega(n)$ the upper bound on the information is trivial and the lower bound on the communication is known.

Let $c = \frac{1}{\log e}$ and $m = c\sqrt{n(k+1)}$. Note that $m = o(n)$ as well. Now $\mu_{n,k}$ can be defined as the distribution obtained by mixing two distributions, one where a pair of disjoint subsets of $\{1, \ldots, n\}$ of size m is chosen uniformly among all such pairs, and one where a pair of subsets of size m with intersection of size 1 is chosen uniformly among all such pairs. This is essentially the distribution used in the proof by Razborov [22], but with smaller sets.

We show in the appendix that the information is bounded by k.

▶ **Theorem 18.** *For any sufficiently small $\epsilon > 0$ we have that $D_\epsilon^{\mu_{n,k}}(DISJ) = \Omega(\sqrt{n(k+1)})$, and hence that $R_\epsilon^{I\leq k}(DISJ) = \Omega(\sqrt{n(k+1)})$.*

While the proof is similar to that of the original proof of Razborov [22], two difficulties arise when working with smaller sets: The first is that by mixing the two distributions with equal probability, the weight of any pair of intersecting sets is much larger than that of a pair of disjoint sets. Since the proof relies heavily on the properties of the distribution when conditioned on certain events, and in particular on the proportion of the weight of 1-inputs and the weight of 0-inputs when conditioning, this imbalance complicates several computations.

The second difficulty comes from the fact that Razborov's entropy "counting" argument no longer works in our case, because in that argument a linear number of terms have their entropy upper-bounded as $H\left(\frac{1}{2}\right) = 1$. Since we still have to deal with a linear number of terms while having much less total entropy, we require a finer combinatorial counting argument instead.

Now we give a simple argument that shows that error dependence cannot be logarithmic in $1/\epsilon$.

▶ **Theorem 19.** $R_\epsilon^{I\leq 1}(DISJ) = \Omega(\sqrt{n/\epsilon})$ *for* $\epsilon \geq \Omega(1/n)$.

Proof. Above we have described a distribution $\mu_{n,k}$ with information at most k such that $\Omega(\sqrt{n(k+1)})$ communication is needed for some constant error δ. We define $\sigma_{n,k}$ to be $1/(2k)\cdot\mu_{n,k}+(1-1/(2k))\rho$, where ρ is some product distribution for DISJ that puts weight $1/2$ on 1-inputs. Clearly, for error $\delta/(4k)$ the communication must be at least $\Omega(\sqrt{n(k+1)})$. Set $k=4\delta/\epsilon$ (note that $k\le n$).

It remains to show that the information in σ is at most 1. Let E be an indicator variable that indicates that x,y have been chosen according to μ_k. Then $I(X:Y)\le I(XE:Y)=I(E:Y)+I(X:Y|E)\le H(E)+(1/2k)\cdot k\le H(1/(2k))+1/2\le 1$. ◀

▶ **Corollary 20.** *The class of distributions with information k with $1\le k\le n^{1-\Omega(1)}$ is not boost-able for randomised protocols.*

Proof. Consider $k=1$. We have that $R^{I\le 1}_{1/3}(DISJ)\le O(\sqrt{n})$. If distributions with at most 1 bit information were boost-able, then we would have $R^{I\le 1}_{\epsilon}(DISJ)\le O(\sqrt{n}\log(1/\epsilon))$. But the left hand side is at least $\Omega(\sqrt{n/\epsilon})$, which puts a lower bound on ϵ, whereas boost-ability should work for all ϵ.

In the case of larger k we use the same proof, to get that $\sqrt{\epsilon}\cdot\log(1/\epsilon)\ge\Omega(1/\sqrt{k+1})$, which remains a restriction on ϵ until k exceeds $n^{1-\Omega(1)}$, and the assumption of Theorem 19 is violated. ◀

4 Quantum Complexity of Disjointness

4.1 Upper Bound: First Attempt

Consider the two-phase approach from the previous section. The second phase ‘quantises’ readily, if we do not care about log-factors: Simply use distributed quantum search by amplitude amplification to obtain a quadratic speedup in this part [6]. We mention here that the tight protocol for DISJ due to Aaronson and Ambainis [1] does not seem to work well for the small set case and so we do not know if the logarithmic factor is needed or not.

The problem is the first phase of the classical protocol, which seems impossible to quantise. Since phase 2 is now cheaper one can re-balance the costs of the two phases (details are left to the reader) and find a protocol with cost $\tilde{O}((n(k+1))^{1/3}$.

In the next section we will show that this bound is not optimal. We do note here, however, that the error dependence for the case $I(X:Y)=0$ is a factor of $O(-\log\epsilon)$ for the above, which will not be the case in the protocol we present next.

4.2 Upper Bound: Almost Optimal Protocol

We now describe a different approach that also works in the classical case, but loses a logarithmic factor and has a worse error dependence for product distributions. The approach we use identifies two blocks of “interesting” positions (i.e., the blocks are subsets of $[n]$), such that Alice can conduct a search efficiently on one block, and Bob on the other one, because their sets are expected to be small on their respective blocks, and on the other hand, the situation when the input sets intersect but not on any interesting position is unlikely. This conforms to the rough intuition that if “large” x and y come from a product distribution that puts constant weight on 1-inputs, then there must be many “semi-interesting” and “uninteresting” positions – i.e., such $i\in[n]$ that not both $i\in x$ and $i\in y$ is likely.

Let μ be a distribution on $\{0,1\}^n\times\{0,1\}^n$ with $I^{\mu}(X:Y)\le k$. Denote by E_i the event that x,y drawn from μ satisfy $\sum_{1\le j\le i-1}x_iy_i=0$, i.e., x and y are disjoint on $\{1,\ldots,i-1\}$.

We set $s_i = Prob_\mu(E_i)$. We assume that $s_i \geq \alpha$ for all i, and some $\alpha \geq \epsilon$. If this is not the case, then the probability that x, y are disjoint is less than ϵ, and the distribution is trivial.

Define $q_i'^x = Prob(Y_i = 1 | X = x, X_i = 1, E_i)$ and $p_i'^y = Prob(X_i = 1 | Y = y, Y_i = 1, E_i)$. Our protocol follows the simple idea that Alice should search among those positions i, such that $q_i'^x$ is large, similarly for Bob and $p_i'^y$.

The Protocol. A position is *chosen* by Alice, if $q_i'^x \geq \epsilon^3/\sqrt{80000(k+1)n}$ and $i \in x$ and *chosen* by Bob, if $p_i'^y \geq \epsilon^3/\sqrt{(80000(k+1)n}$ and $i \in y$. Denote the former set by C_A and the latter by C_B. Alice is responsible for finding intersecting positions in C_A, Bob for finding intersecting positions in C_B. In the protocol Alice organises a search for an intersecting position based on amplitude amplification on her positions C_A (using a distributed Grover search as in [6]). More precisely, Alice creates a superposition over all positions in C_A, and the two players can mark intersecting positions like in Grover search by communicating $\log n$ qubits back and forth, and conduct amplitude amplification to find an intersection there. In phase 2 the same is done with C_B and the roles of the players reversed. If the players find an intersecting position, they reject, otherwise they accept.

Communication. For all x we have $\sum_{i \in x} q_i'^x s_i \leq Prob(DISJ(x,y) = 0) \leq 1$. Hence $|C_A| \leq O(\sqrt{(k+1)n}/\epsilon^4)$, since all $s_i \geq \alpha \geq \epsilon$. Amplitude amplification needs $O(((k+1)n)^{1/4}/\epsilon^2 \cdot \log(1/\epsilon))$ iterations, each taking $\log n$ communication.

Error Analysis. Let us define some probabilities. By $\vec{x}, \vec{y}$ we denote prefixes of strings x, y of length $i-1$, where i is usually clear from the context. The random variable $\vec{X}$ is the prefix of length $i-1$ of the random variable X (Alice's inputs), and similarly for Y.

Denote $p_i^{\vec{x}} = Prob(X_i = 1 | E_i, \vec{X} = \vec{x})$, $q_i^{\vec{x}} = Prob(Y_i = 1 | E_i, \vec{X} = \vec{x})$, and similarly for $p_i^{\vec{y}}$ and $q_i^{\vec{y}}$. Denote also $p_i'^{\vec{y}} = Prob(X_i = 1 | E_i, Y_i = 1, \vec{Y} = \vec{y})$, and similarly for $p_i'^{\vec{x}}$, $q_i'^{\vec{x}}$ and $q_i'^{\vec{y}}$. Denote also $q_i'^{\vec{x},\vec{y}} = Prob(Y_i = 1 | E_i, X_i = 1, \vec{X} = \vec{x}, \vec{Y} = \vec{y})$, and similarly for $p_i'^{\vec{x},\vec{y}}$. Denote by $r_i^{\vec{x}} = p_i^{\vec{x}} q_i'^{\vec{x}} = p_i'^{\vec{x}} q_i^{\vec{x}}$ the probability that $X_i = Y_1 = 1$ under the conditions $\vec{X} = \vec{x}$ and E_i, and similarly for other conditions (i.e., the super-script specifies the condition): say, r_i is the probability that $X_i = Y_i = 1$ conditioned on E_i, and so on.

As a first step we "get rid" of the input positions that are very unlikely to contribute, compared to the average for a position. We say that x with $x_i = 1$ is *A-bad* for i, if $q_i'^{\,x} \leq \epsilon q_i'^{\,\vec{x}}/10$. These are the positions where x depresses the probability of intersection compared to $\vec{x}$. Similarly, y with $y_i = 1$ is *B-bad* for i, if $p_i'^{\,y} \leq \epsilon p_i'^{\,\vec{y}}/10$. Denote by V_i the event that x is A-bad for i, and by W_i the event that y is B-bad for i. Finally, set $\tilde{q}_i'^{\,\vec{x}} = Prob(Y_i = 1 \wedge V_i | X_i = 1, \vec{X} = \vec{x}, E_i)$ and $\tilde{p}_i'^{\,\vec{y}} = Prob(X_i = 1 \wedge W_i | Y_i = 1, \vec{Y} = \vec{y}, E_i)$.

Note that $r_i^{\vec{x}} s_i = p_i^{\vec{x}} q_i'^{\,\vec{x}} s_i$ is the probability that the first intersection between X and Y is on position i when $\vec{X} = \vec{x}$. The probability that x is A-bad for i and the first intersection is on i is $p_i^{\vec{x}} \tilde{q}_i'^{\vec{x}} s_i$. Similarly, the probability that y is B-bad for i and the first intersection is on i is $\tilde{p}_i'^{\vec{y}} q_i^{\vec{y}} s_i$. The following lemma shows that possible intersections at such positions i that either x is A-bad for i or y is B-bad for i can be safely ignored.

▶ **Lemma 21.** $\tilde{q}_i'^{\vec{x}} \leq \epsilon q_i'^{\vec{x}}/10$ *and* $\tilde{p}_i'^{\vec{y}} \leq \epsilon p_i'^{\vec{y}}/10$.

Proof. Denote by $Bad(x,i)$ the property that x is A-bad for i and by $E_i(x,y)$ the property that x, y are disjoint on $\{1, \dots, i-1\}$.

$$\begin{aligned}
\tilde{q}_i^{\prime\vec{x}} &= \frac{Prob(V_i \wedge Y_i = 1 \wedge X_i = 1 \wedge \vec{X} = \vec{x} \wedge E_i)}{Prob(X_i = 1 \wedge \vec{X} = \vec{x} \wedge E_i)} \\
&= \sum_{x: x_i = 1, x_1, \dots, x_{i-1} = \vec{x}, Bad(x,i)} \ \sum_{y: y_i = 1, E_i(x,y)} \mu(x,y) / Prob(X_i = 1 \wedge \vec{X} = \vec{x} \wedge E_i) \\
&= \sum_{x: x_i = 1, x_1, \dots, x_{i-1} = \vec{x}, Bad(x,i)} q_i^{\prime x} \cdot Prob(X = x \wedge E_i) / Prob(X_i = 1 \wedge \vec{X} = \vec{x} \wedge E_i) \\
&\leq (\epsilon/10) \cdot q_i^{\prime\vec{x}} \cdot \sum_{x: x_i = 1, x_1, \dots x_{i-1} = \vec{x}} Prob(X = x \wedge E_i) / Prob(X_1 = 1 \wedge \vec{X} = \vec{x} \wedge E_i) \\
&\leq (\epsilon/10) \cdot q_i^{\prime\vec{x}}.
\end{aligned}$$

◀

Therefore, ignoring possible intersections where x or y are bad for i, one can introduce error at most $\epsilon/5$, because the probability of an A-bad (first) intersections is at most $p_i^{\vec{x}} \tilde{q}_i^{\prime\vec{x}} s_i \leq (\epsilon/10) p_i^{\vec{x}} q_i^{\prime\vec{x}} s_i$ for any $\vec{x}$, with a similar bound for B-bad. Hence in the following we assume that all x, y are not bad for i.

We call a position i and inputs x, y *lucky*, if $p_i^{\vec{x}} \leq 400(k+1) p_i^{\prime\vec{y}} / \epsilon^3$. The remaining positions are unlucky for x, y. There are four possible sources of error in our protocol: There are “bad” intersections (considered above). Among the positions for which the input is not bad, there may be *unchosen lucky* positions and *unchosen unlucky* positions. Finally, some error comes from the amplitude amplification quantum searches.

“Bad” intersections contribute error at most $\epsilon/5$, as shown above. The amplitude amplification error can be pushed below $\epsilon/20$ by increasing communication by a factor of $O(-\log \epsilon)$, which is already absorbed in the stated communication bound above. It remains to deal with the unchosen lucky and unlucky positions (for which the input is not bad).

We first consider the error contributed by lucky positions i that are not chosen by either Alice or Bob – denote these by L. Fix the input prefixes $\vec{x}, \vec{y}$ and assume that the inputs are not bad for i. Positions that are not chosen satisfy $p_i^{\prime\vec{y}} q_i^{\prime\vec{x}} \leq (10/\epsilon)^2 \cdot p_i^{\prime y} q_i^{\prime x} \leq \epsilon^4 / (800(k+1)n)$. We have that the probability that the first intersection is at position $i \in L$ but i is not chosen, is (conditioned on $\vec{x}$)

$$r_i^{\vec{x}} s_i = p_i^{\vec{x}} q_i^{\prime\vec{x}} s_i \leq 400(k+1) p_i^{\prime\vec{y}} q_i^{\prime\vec{x}} s_i / \epsilon^3 \leq \epsilon/(2n) \cdot s_i \leq \epsilon/(2n),$$

where the first inequality is because of ‘lucky’, and the second because of ‘unchosen’. Summing up, and taking expectations (over $\vec{x}$ under μ_i), this gives $\sum_i Prob(X_i = Y_i = 1 \wedge E_i \wedge i$ lucky, not chosen$) \leq \epsilon/2$, hence error at most $\epsilon/2$.

Now we turn to the error contributed by unlucky positions. For these we have that $p_i^{\vec{x}} > 400(k+1)/\epsilon^3 \cdot p_i^{\prime\vec{y}}$.

We use the following lemma.

▶ **Lemma 22.** *Assume that for no x or y the conditional probability of non-intersection is less than α, and that for no x and i the probability that $X_i = Y_i = 1$ conditioned on $X = x$ and E_i is larger than $1/2$, and the same for all y, i. Then*

$$\sum_i \mathbf{E}_{\vec{x},\vec{y}}^{\mu_i} \ p_i^{\vec{x}} q_i^{\vec{y}} \leq 16k/\alpha + 68/\alpha^2,$$

where $\vec{x}, \vec{y}$ are prefixes of length $i-1$.

Note that the first assumption of the lemma can be made since otherwise we can just reject x (resp., y) with error α. The second assumption can be made since otherwise i is chosen by Alice (resp., Bob), and no error happens there.

Now we can bound the probability that an unlucky position i has the first intersection (conditioned on $\vec{y}$) by

$$r_i^{\vec{y}} s_i \leq r_i^{\vec{y}} = p_i'^{\vec{y}} q_i^{\vec{y}} \leq \epsilon^3 p_i^{\vec{x}} q_i^{\vec{y}}/(400(k+1)).$$

Summing up over all i (not just the unlucky ones) and taking expectation over μ_i we get by our lemma that

$$\begin{aligned} & \sum_i \mathbf{E}_{\vec{x},\vec{y}}\ \epsilon^3/(400(k+1)) \cdot p_i^{\vec{x}} q_i^{\vec{y}} \\ \leq\ & \epsilon^3/(400(k+1)) \cdot ((16k/\alpha) + 68/\alpha^2) \\ \leq\ & \epsilon/4. \end{aligned}$$

Hence the total error is not more than $\epsilon/20 + \epsilon/4 + \epsilon/5 + \epsilon/2 \leq \epsilon$ and we get the following.

▶ **Theorem 23.** $Q_\epsilon^{I \leq k}(f) \leq O((n(k+1))^{1/4}/\epsilon^2 \cdot \log n \cdot \log(1/\epsilon))$.

It remains to prove the lemma.

Proof of Lemma 22. Denote by μ_i the probability distribution μ, restricted to the event E_i. We know that $k \geq I^\mu(X:Y) = D(\mu||\sigma)$, where σ is the product of marginals of μ. Denote by σ_i the product of marginals of μ_i, and by $\mu_i^{\vec{x},\vec{y},j}$ the distribution μ_i, conditioned on the event $X_1 = x_1, \ldots, X_j = x_j, Y_1 = y_1, \ldots, Y_j = y_j$, which we abbreviate by $F^{\vec{x},\vec{y},j}$. Similarly, $\sigma_i^{\vec{x},\vec{y},j}$ is σ_i conditioned on $F^{\vec{x},\vec{y},j}$. Note that for the latter probability distribution we first take the product of marginals of μ_i, and then condition. This is different from considering conditional mutual information, in which one would first condition and then take the product of marginals. We also stress that here j denotes the length of $\vec{x}, \vec{y}$, unlike before. In the following, when we do not mention j explicitly, it is $i-1$: e.g., $\mu_i^{\vec{x},\vec{y}} = \mu_i^{\vec{x},\vec{y},i-1}$.

By the chain rule for relative entropy we get that

$$\sum_{j=1,\ldots,n} \mathbf{E}^{\mu_i}_{\vec{x},\vec{y},j-1} D(\mu_i^{\vec{x},\vec{y},j-1}(X_j, Y_j)||\sigma_i^{\vec{x},\vec{y},j-1}(X_j, Y_j)) = D(\mu_i||\sigma_i) = I^{\mu_i}(X:Y),$$

where the expectation is over the prefixes $\vec{x}, \vec{y}$ of length $j-1$ under μ_i. We are interested in

$$k_i = \mathbf{E}^{\mu_i}_{\vec{x},\vec{y}} k_i^{\vec{x},\vec{y}} = \mathbf{E}^{\mu_i}_{\vec{x},\vec{y}} D(\mu_i^{\vec{x},\vec{y}}(X_i, Y_i)||\sigma_i^{\vec{x},\vec{y}}(X_i, Y_i)).$$

For this, i determines both the condition on previous positions, and the choice of distribution. The chain rule can be used if we fix μ_i, but here we want to vary μ_i as well. For the moment suppose we can bound $\sum k_i$ by a k' not much larger than k.

Observe that $p_i^{\vec{x}} q_i^{\vec{y}}$ is the probability that $X_i = Y_i = 1$ under the distribution $\sigma_i^{\vec{x},\vec{y}}(X_i, Y_i)$. σ_i is a product distribution, and hence conditioning on $Y's$ does note change the probability of $X_i = 1$ etc., and so we get that $p_i^{\vec{x}} = Prob_{\sigma_i}(X_i = 1|\vec{X} = \vec{x}, \vec{Y} = \vec{y})$.

We can now apply Lemma 10 to learn that either $p_i^{\vec{x}} q_i^{\vec{y}} \leq 4r_i^{\vec{x},\vec{y}}$ or $D(\mu_i^{\vec{x},\vec{y}}(X_i, Y_i)||\sigma_i^{\vec{x},\vec{y}}(X_i, Y_i)) \geq p_i^{\vec{x}} q_i^{\vec{y}}/16$. Hence

$$p_i^{\vec{x}} q_i^{\vec{y}} \leq 4r_i^{\vec{x},\vec{y}} + 16D(\mu_i^{\vec{x},\vec{y}}(X_i, Y_i)||\sigma_i^{\vec{x},\vec{y}}(X_i, Y_i)).$$

Then

$$\sum_i p_i^{\vec{x}} q_i^{\vec{y}} \leq \sum_i 4r_i^{\vec{x},\vec{y}} + 16k_i^{\vec{x}\vec{y}}.$$

Noting that $\sum_i \mathbf{E}^{\mu_i}_{\vec{x},\vec{y}} r_i^{\vec{x},\vec{y}} s_i \leq \sum r_i s_i \leq 1$ and hence $\sum_i \mathbf{E}^{\mu_i}_{\vec{x},\vec{y}} r_i^{\vec{x},\vec{y}} \leq 1/\alpha$ it remains to bound $k' = \sum k_i$ by $k/\alpha + 4/\alpha^2$. For this we need to first compare $\mu_{i+1}(x,y)$ and $\mu_i(x,y)$. If $(x,y) \in E_{i+1}$, then we get $\mu_{i+1}(x,y) = \mu_i(x,y)/(1-r_i)$. Also, we have $\sigma_{i+1}(x,y) = \sum_{y':(x,y')\in E_{i+1}} \mu_i(x,y')/(1-r_i) \cdot \sum_{x':(x',y)\in E_{i+1}} \mu_i(x',y)/(1-r_i)$, and, denoting $r_i^y = Prob_{\mu_i}(X_i = Y_i = 1|Y = y)$ that is equal to $\sum_{y':(x,y')\in E_i} \mu_i(x,y') \cdot (1-r_i^x)/(1-r_i) \cdot \sum_{x':(x',y)\in E_i} \mu_i(x',y) \cdot (1-r_i^y)/(1-r_i)$. Which is $\mu_i(x) \cdot \mu_i(y) \cdot (1-r_i^x)(1-r_i^y)/(1-r_i)^2$.

Let us compute an upper bound on $D(\mu_{i+1}||\sigma_{i+1})$

$$\begin{aligned}
&= \sum_{(x,y)\in E_{i+1}} \mu_{i+1}(x,y) \log \frac{\mu_{i+1}(x,y)}{\sigma_{i+1}(x,y)} \\
&= \sum_{(x,y)\in E_{i+1}} \mu_i(x,y)/(1-r_i) \cdot \log \frac{\mu_i(x,y) \cdot (1-r_i)}{\sigma_i(x,y)(1-r_i^x)(1-r_i^y)} \\
&\overset{(*)}{\leq} \sum_{(x,y)\in E_i} \mu_i(x,y)/(1-r_i) \cdot \log \frac{\mu_i(x,y) \cdot (1-r_i)}{\sigma_i(x,y)(1-r_i^x)(1-r_i^y)} \\
&- \mu_i(E_i - E_{i+1})/(1-r_i) \cdot \log(4\mu_i(E_i - E_{i+1})/\sigma_i(E_i - E_{i+1})) \\
&\leq \sum_{(x,y)\in E_i} \mu_i(x,y) \log \left(\frac{\mu_i(x,y)}{\sigma_i(x,y)}\right) /(1-r_i) \\
&+ \sum_{(x,y)\in E_i} \mu_i(x,y)/(1-r_i) \cdot \log \frac{1-r_i}{(1-r_i^x)(1-r_i^y)} - r_i/(1-r_i) \cdot \log(4r_i) \\
&\overset{(**)}{\leq} D(\mu_i||\sigma_i)/(1-r_i) + 2 \sum_{(x,y)\in E_i} \mu_i(x,y) \cdot 2 \cdot (r_i^x + r_i^y) - 2r_i \log(4r_i) \\
&\leq D(\mu_i||\sigma_i)/(1-r_i) + 12r_i - 2r_i \log(r_i),
\end{aligned}$$

where in (*) we use Lemma 6, in (**) we use that $-\log(1-\lambda) = -\ln(1-\lambda)/\ln(2) \leq 2\lambda$, for all $0 \leq \lambda \leq 1/2$, and in general use that $r_i^x, r_i^y, r_i \leq 1/2$ by the assumption in the lemma. The conclusion is that the relative entropy increases only slightly.

Now we turn to the terms k_i in the chain rule expansion. Fix $X_1 = x_1, \ldots, X_i = x_i$ and $Y_1 = y_1, \ldots, Y_i = y_i$. We are interested in $D(\mu_{i+1}^{\vec{x},\vec{y},i}||\sigma_{i+1}^{\vec{x},\vec{y},i})$ and its relation to to $D(\mu_i^{\vec{x},\vec{y},i}||\sigma_i^{\vec{x},\vec{y},i})$. Note that the distributions involved are on $X_{i+1}, \ldots, X_n, Y_{i+1}, \ldots, Y_n$. We assume $x_j y_j \neq 1$ for all $j < i+1$, otherwise the inputs are not in E_{i+1} and have no weight under μ_{i+1}. We have

$$\begin{aligned}
& D(\mu_{i+1}^{\vec{x},\vec{y}}(X_{i+1}, \ldots, X_n, Y_{i+1}, \ldots, Y_n)||\sigma_{i+1}^{\vec{x},\vec{y}}(X_{i+1}, \ldots, X_n, Y_{i+1}, \ldots, Y_n)) \\
&= \sum_{x,y \in E_{i+1}: x_1,\ldots,x_i=\vec{x}, y_1,\ldots,y_i=\vec{y}} \mu_{i+1}(x,y|\vec{x},\vec{y}) \log \left(\frac{\mu_{i+1}(x,y|\vec{x},\vec{y})}{\sigma_{i+1}(x,y|\vec{x},\vec{y})}\right) \\
&\leq \sum_{x,y \in E_i: x_1,\ldots,x_i=\vec{x}, y_1,\ldots,y_i=\vec{y}} \mu_i(x,y|\vec{x},\vec{y}) \log \left(\frac{\mu_i(x,y|\vec{x},\vec{y})}{\sigma_i(x,y|\vec{x},\vec{y}) \cdot (1-r_i^x)(1-r_i^y)}\right) \\
&\leq D(\mu_i^{\vec{x},\vec{y},i}(X_{i+1}, \ldots, X_n, Y_{i+1}, \ldots, Y_n)||\sigma_i^{\vec{x},\vec{y},i}(X_{i+1}, \ldots, X_n, Y_{i+1}, \ldots, Y_n)) \\
&+ \sum_{x,y: x_1,\ldots,x_i=\vec{x}, y_1,\ldots,y_i=\vec{y}} \mu_i(x,y|\vec{x},\vec{y}) \cdot \log \left(\frac{1}{(1-r_i^x)(1-r_i^y)}\right) \\
&\leq D(\mu_i^{\vec{x},\vec{y},i}(X_{i+1}, \ldots, X_n, Y_{i+1}, \ldots, Y_n)||\sigma_i^{\vec{x},\vec{y},i}(X_{i+1}, \ldots, X_n, Y_{i+1}, \ldots, Y_n)) \\
&+ \sum_{x,y: x_1,\ldots,x_i=\vec{x}, y_1,\ldots,y_i=\vec{y}} \mu_i(x,y|\vec{x},\vec{y})(2r_i^x + 2r_i^y) \\
&\leq D(\mu_i^{\vec{x},\vec{y},i}(X_{i+1} \ldots)||\sigma_i^{\vec{x},\vec{y},i}(X_{i+1}, \ldots)) + 2r_i^{\vec{y}} + 2r_i^{\vec{x}}.
\end{aligned}$$

Note here that conditioned on $\vec{x}, \vec{y}$, the condition E_{i+1} is satisfied for all inputs, and no re-scaling happens going from μ_{i+1} to μ_i conditioned on $\vec{x}, \vec{y}$.

We can now bound $\sum_i k_i = \sum_i \mathbf{E}^{\mu_i}_{\vec{x},\vec{y}}\, k_i^{\vec{x},\vec{y}}$. Note that $\mu_1 = \mu$.

$$\begin{array}{rl}
k \geq & D(\mu||\sigma) \\
= & D(\mu_1(X_1,Y_1)||\sigma_1(X_1,Y_1)) \\
+ & \mathbf{E}^{\mu_1}_{x_1,y_1} D(\mu_1^{x_1,y_1}(X_2,\ldots,X_n,Y_2,\ldots,Y_n)||\sigma_1^{x_1,y_1}(X_2,\ldots,X_n,Y_2,\ldots,Y_n)) \\
\geq & D(\mu_1(X_1,Y_1)||\sigma_1(X_1,Y_1)) \\
+ & \mathbf{E}^{\mu_1}_{x_1,y_1} D(\mu_2^{x_1,y_1}(X_2,\ldots,X_n,Y_2,\ldots,Y_n)||\sigma_2^{x_1,y_1}(X_2,\ldots X_n,Y_2,\ldots,Y_n)) - 4r_1 \\
\geq & D(\mu_1(X_1,Y_1)||\sigma_1(X_1,Y_1)) \\
+ & \mathbf{E}^{\mu_2}_{x_1,y_1} D(\mu_2^{x_1,y_1}(X_2,\ldots)||\sigma_2^{x_1,y_1}(X_2,\ldots)) \cdot (1-r_1) - 4r_1 \\
= & D(\mu_1(X_1,Y_1)||\sigma(X_1,Y_1)) \\
+ & \mathbf{E}^{\mu_2}_{x_1,y_1} D(\mu_2^{x_1,y_1}(X_2,Y_2)||\sigma^{x_1,y_1}(X_2,Y_2)) \cdot (1-r_1) \\
+ & \mathbf{E}^{\mu_2}_{x_1,x_2,y_1,y_2} D(\mu_2^{x_1,x_2,y_1,y_2}(X_3,\ldots)||\mu_2^{x_1,x_2,y_1,y_2}(X_3,\ldots)) \cdot (1-r_1) - 4r^1 \\
\geq & D(\mu_1(X_1,Y_1)||\sigma_1(X_1,Y_1)) + \mathbf{E}^{\mu_2}_{x_1,y_1} D(\mu_2^{x_1,y_1}(X_2,Y_2)||\sigma_2^{x_1,y_1}(X_2,Y_2)) \cdot (1-r_1) \\
+ & \mathbf{E}^{\mu_2}_{x_1,x_2,y_1,y_2} D(\mu_3^{x_1,x_2,y_1,y_2}(X_3,\ldots)||\mu_3^{x_1,x_2,y_1,y_2}(X_3,\ldots)) \cdot (1-r_1) \\
- & 4r_1 - 4r_2 \cdot (1-r_1) \\
\geq & D(\mu_1(X_1,Y_1)||\sigma_1(X_1,Y_1)) + \mathbf{E}^{\mu_2}_{x_1,y_1} D(\mu_2^{x_1,y_1}(X_2,Y_2)||\sigma_2^{x_1,y_1}(X_2,Y_2)) \cdot (1-r_1) \\
+ & \mathbf{E}^{\mu_3}_{x_1,x_2,y_1,y_2} D(\mu_3^{x_1,x_2,y_1,y_2}(X_3,\ldots)||\mu_3^{x_1,x_2,y_1,y_2}(X_3,\ldots)) \cdot (1-r_1)(1-r_2) \\
- & 4r_1 - 4r_2 \cdot (1-r_1) \\
\vdots & \\
\geq & \displaystyle\sum_i \mathbf{E}^{\mu_i}_{\vec{x},\vec{y}} D(\mu_i^{\vec{x},\vec{y}}(X_i,Y_i)||\sigma_i^{\vec{x},\vec{y}}(X_i,Y_i)) \cdot \alpha - 4\sum_i r_i \\
= & \displaystyle\sum_i k_i \cdot \alpha - 4/\alpha,
\end{array}$$

where in the last step we use that $\prod_{i=1,\ldots,n}(1-r_i) \geq \alpha$ and $\sum_i r_i \leq 1/\alpha$.

This means that $\sum k_i \leq k/\alpha + 4/\alpha^2$. ◀

4.3 Lower Bound

We use exactly the same hard distribution for the quantum case as for the classical case, see Section 3.2, where also the mutual information of this distribution is shown to be at most k. Conveniently, Razborov [23] has done most of the hard work for us by analysing the quantum complexity of Disjointness for all set sizes. We get the following main result:

▶ **Theorem 24.** *The distributional quantum communication complexity of Disjointness under $\mu_{n,k}$ is at least $\Omega((n(k+1))^{1/4})$.*

Proof. Recall the distributions $\nu_{n,k}, \sigma_{n,k}$ as defined in Section 3.2. These are the distributions of sets of size $s = O(\sqrt{n(k+1)}$ from a size n universe (not intersecting resp. intersecting). We employ the following result by Razborov [23]:

▶ **Fact 25.** *Any quantum protocol that solves DISJ with error ϵ under $\nu_{n,k}$ and error ϵ under $\sigma_{n,k}$ needs communication $\Omega(\sqrt{s}) = \Omega((n(k+1))^{1/4})$.*

This follows from Razborov's proof, in which given a quantum protocol with communication c for DISJ (on inputs of size s from a size n universe), a uni-variate polynomial of degree $O(c)$ on $\{0, 1, \ldots, s\}$ is constructed such that $p(i)$ is close to 0 for all $\{0, 1, \ldots, s-1\}$ and $p(s) = 1$. Such a polynomial must have degree $\Omega(\sqrt{s})$. The construction is done by averaging of the acceptance probabilities on all inputs x, y where x, y have size s, and hence it is enough if the given protocol for DISJ is correct on average inputs under $\nu_{n,k}$ and under $\sigma_{n,k}$. But any protocol with small error under $\mu_{n,k}$ must also have small error under both of these distributions, and we get the same lower bound under this distribution as in the worst case, as stated by Razborov. ◀

We also note that again, the error dependence cannot be polylogarithmic. The proof is the same as in the classical case.

▶ **Theorem 26.** $Q_{\epsilon}^{I \leq 1}(DISJ) \geq \Omega((n/\epsilon)^{1/4})$.

We again obtain this following.

▶ **Corollary 27.** *The class of distributions with information k with $1 \leq k \leq n^{1-\Omega(1)}$ is not boost-able for quantum protocols.*

5 Large Correlation is Needed for Tight Bounds

In this section we show that there is a function, for which the distributional communication complexity is far from the randomised communication complexity if the information in the distribution is less than $\Omega(n)$. The main idea is that random sparse problems make it hard for low information distributions to 'focus' on the 1-inputs.

Define $f_{n,d}$ as a random variable that takes as its values functions $f : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$. The functions are generated randomly as follows. Each input x, y is chosen to be a 1-input independently with probability $d/2^n$.

Note that the communication matrix of $f_{n,d}$ has expected d 1-inputs for each row and column. In the following d should be thought of as some value like $2^{\sqrt{n}}$. We need $2^{n/100} \geq d \geq 6n$.

We first show that the complexity of $f_{n,d}$ is $\Theta(\log d)$ with high probability. Then, we show that with high probability $f_{n,d}$ has a property that allows an $O(\log n)$ protocol under all low information distributions.

First we note that by the Chernoff bound the probability that a row or column has more than $2d$ or less than $d/2$ 1-inputs is at most $2e^{-d/3} \leq 2^{-2n}$. By the union bound it is true for all rows and columns (with high probability) that they contains between $d/2$ and $2d$ 1-inputs. Throughout this section we assume that $f_{n,d}$ has this property.

▶ **Lemma 28.** *$R(f_{n,d}) \leq O(\log d)$ with high probability.*

Proof. With high probability there are at most $2d$ 1-inputs $(x_1, y), \ldots, (x_{2d}, y)$ in Bob's column. If Alice sends a fingerprint of x as in Fact 14, using $2 \log d$ bits, then Bob can check whether $x = x_j$ for some $1 \leq j \leq 2d$ with error $2d \cdot 2^{-2 \log d} \leq 2/d$. If so, then he accepts, otherwise he rejects. ◀

▶ **Lemma 29.** *$R(f_{n,d}) \geq \Omega(\log d)$ with high probability.*

Proof. The proof is by the probabilistic method. We use the minimax theorem and the following hard distribution: Put 1/2 weight on 1-inputs and 1/2 weight on 0-inputs to $f_{n,d}$. Note that the mutual information of this distribution is $\Omega(n)$: for 1-inputs, given x there are at most d inputs y out of 2^n such that x, y is a 1-input. Hence the information is at least $(n - \log d)/2$.

We employ the 1-sided discrepancy method. The 1-sided discrepancy under a distribution μ is $disc'(f, \mu) = \max_R \mu(f^{-1}(1) \cap R) - \mu(f^{-1}(0) \cap R)$, where the maximum is over all rectangles. Then $R^{\mu}(f) \geq -\log disc'(f, \mu)$ for all μ that put weight 1/2 on the 1-inputs. Our goal is to show that the 1-sided discrepancy is small with high probability over the choice of $f_{n,d}$.

Fix a rectangle R and consider a random $f_{n,d}$. We would like to compute the probability that $disc'(R) = \mu(f_{n,d}^{-1}(1) \cap R) - \mu(f_{n,d}^{-1}(0) \cap R)$ is large. Note that this is a random variable and that μ depends on $f_{n,d}$

If $\mu(R \cap f^{-1}(1)) \leq 4/d^{1/4}$, then $disc'(R) \leq 4/d^{1/4}$ and we are done. Hence we assume the opposite. For R to contain at least a $4/d^{1/4}$ fraction of all 1-inputs it must be the case that R contains at least $(4/d^{1/4}) \cdot 2^n d/2$ 1-inputs, and no row or column contains more than $2d$ of them, which implies that R must have at least $2^n/d^{1/4}$ rows and columns.

Write $R = A \times B$, where $|A|, |B| \geq 2^n/d^{1/4}$. The expected number of 1-inputs in R is at most $|A| \cdot |B| \cdot d/2^n$. The 1-inputs are chosen independently, and the Chernoff bound yields that $Prob(R$ contains more than $(1 + d^{-1/2})|A||B|d/2^n$ 1-inputs$) \leq e^{-|A||B|d/(3 \cdot 2^n d^{1/4})} \leq e^{-2^n d^{1/4}/3}$. Similarly, we can bound $Prob(R$ contains less than $(1 - d^{-1/2})|A||B|d/2^n$ 1-inputs$)$.

Furthermore, since there are at most $2^{2^{n+1}}$ rectangles, and by the union bound with high probability these estimates are correct for *all* rectangles with enough rows and columns (in particular the rectangle consisting of all inputs).

Note that R contains at least $|A| \cdot |B| - |A| 2d$ 0-inputs, each of which have weight at least $1/(2^{2n+1})$, for a total 0-weight of at least $|A||B|/2^{2n+1} - d/2^n$. The weight of a single 1-input is at most $1/(1 - d^{-1/2}) \cdot 1/(d2^{n+1})$ and the total 1-weight of R is at most $(1 + d^{-1/2})/(1 - d^{-1/2}) \cdot |A||B|/2^{2n+1}$ by the above. Hence the one-sided discrepancy is at most $O(d^{-1/2}|A||B|/2^{2n+1}) \leq O(d^{-1/2})$. ◀

We will now show that most functions $f_{n,d}$ are easy under all low information distributions, but hard for information n distributions, by showing that $f_{n,d}$ has a certain property with high probability. We assume in the following that $d \leq 2^{\epsilon^2 n}$ and set $\epsilon = 1/10$.

▶ **Definition 30.** We say a Boolean $2^n \times 2^n$ matrix is *good*, if it is true that every rectangle $A \times B$ with $\min\{|A|, |B|\} \leq 2^{2n/3}$ has no more than $100 \max\{|A|, |B|\}$ 1-entries. We also call any rectangle $A \times B$ with $\min\{|A|, |B|\} \leq 2^{2n/3}$ in a good matrix *good*.

▶ **Lemma 31.** *With high probability the communication matrix of $f_{n,d}$ is good.*

Proof. Fix A, B. Assume that $|B| \geq |A|$ and that $|A| \leq 2^{2n/3}$. The probability that a fixed x, y is a 1-input is $d/2^n$. The probability that there are at least $100|B|$ 1-inputs in R is at most $\binom{|A||B|}{100|B|} \cdot (d/2^n)^{100|B|} \leq (\frac{|A|d}{2^n})^{100|B|} \leq \frac{d}{2^{n/3}}^{100|B|}$.

There are $\binom{2^n}{|A|}\binom{2^n}{|B|} \leq (e2^n/|B|)^{2|B|}$ rectangles of this size. By the union bound the probability that there is a rectangle that is not good is small. ◀

Now assume that f (or rather its matrix) is good. Consider any ν such that $I(X : Y) \leq \epsilon^3 n$. We have to give a protocol for f under ν. By Fact 4 there is another distribution μ, that is $\epsilon/2$-close to ν and has $I_\infty(X : Y) \leq 8\epsilon^2 n$. We describe a protocol for f under μ with error $\epsilon/2$. The same protocol has error at most ϵ under ν. We assume $d \leq 2^{\epsilon^2 n}$.

Alice and Bob consider the marginal distributions μ_A and μ_B. Alice sends 0, if $\mu_A(x) \leq 2^{-n/2 - \epsilon n}$, and 1, otherwise, and Bob does the same for $\mu_B(y)$. We first consider the rectangle

R_{00} the messages were 00. Then $\mu_A(x) \cdot \mu_B(y) \leq 2^{-n-2\epsilon n}$ for all x, y in R_{00}. Hence on this rectangle $\sum_{x,y \in R: f(x,y)=1} \mu_A(x)\mu_B(y) \leq 2d2^{-2\epsilon n}$. That means that under $\mu_A \times \mu_B$ the probability of 1-inputs in R_{00} is at most $2d2^{-2\epsilon n}$. But since $I_\infty(X : Y) \leq 8\epsilon^2 n$, the probability of 1-inputs there under μ is at most $2^{-2\epsilon n + O(\epsilon^2 n)}$. We can reject on R_{00} without introducing much error.

Now consider one of the remaining rectangles, say $R_{10} = A \times B$. Clearly, this rectangle has $|A| \leq 2^{n/2+\epsilon n}$. Assume $|A| \leq |B|$. By the above lemma this means that $A \times B$ is good, i.e., contains relatively few 1-inputs, on average only 100 per column.

On R_{10} Alice and Bob send public coin fingerprints of x, y each, with error guarantee $\epsilon/1000$. This takes communication $O(-\log \epsilon)$. If a row or column contains few 1-inputs Alice resp. Bob can test with the fingerprint whether x, y is one of these. But R_{10} only contains few 1-inputs on average, and it is quite possible that both the row and the column of x, y have many 1-inputs.

Let $A = A_0$ and $B = B_0$. Assume that $|A| \leq |B|$. Define A_i as the set of $x \in A_{i-1}$ such that there are at least 1000 1-inputs x, y' with $y' \in B_{i-1}$ and B_i the set of $y \in B_{i-1}$ such that there are at least 1000 1-inputs x', y with $x' \in A_{i-1}$.

Clearly, all $A_i \times B_i$ are good. Assume that $|A_i| \leq |B_i|$. $A_i \times B_i$ has at most $100|B_i|$ 1-inputs. $A_i \times B_{i+1}$ has at least $1000|B_{i+1}|$ 1-inputs, hence $|B_{i+1}| \leq |A_i|/10$, because $A_i \times B_{i+1}$ is good: $1000|B_{i+1}| \leq 100 \max\{|A_i|, |B_{i+1}|\}$. That means that for odd i we have $|B_i| \leq |A_{i-1}|/10$ and for even i we have $|A_i| \leq |B_{i-1}|/10$.

All sets A_i, B_i are known to Alice and Bob without communication. Also, due to the shrinking sizes, all $i \leq O(n)$.

The protocol works as follows: Alice determines the first i such that on $A_i \times B_{i-1}$ her row contains at most 1000 1-inputs and sends this information. Bob also sends the index j, such that on $A_{j-1} \times B_j$ his column contains at most 1000 1-inputs. If $i < j$, then Bob also sends a fingerprint of y with error guarantee 1/10000 (see Fact 14). If there is a $y' \in B_{i-1}$ with the same fingerprint and $f(x, y') = 1$ then Alice accepts, otherwise she rejects. If $i > j$, then Alice sends the fingerprint, and Bob accepts if and only if there is an $x' \in A_{j-1}$ with $f(x', y) = 1$. Clearly the communication is $2 \log n + O(1)$, and is done in 2 rounds.

Correctness: Assume $i < j$. The players can be sure that $x, y \in A_i \times B_{i-1}$. There are at most 1000 1-inputs in row x in B_{i-1}. If $f(x, y) = 1$, then certainly the fingerprints will coincide, and Alice accepts. Otherwise the probability that the fingerprints equal is at most $100/10000 = 1/10$.

▶ **Lemma 32.** *Under ν with information at most $\epsilon^3 n$ and for $6n \leq d \leq 2^{\epsilon^2 n}$ we have that $R^\nu_\epsilon(f) \leq O(\log n)$, if f is good.*

▶ **Theorem 33.** *For every $6n \leq d \leq 2^{n/100}$ there is a function f_d such that*

- $R(f_d) = \Theta(\log d)$,
- $R^{I \leq n/1000}_{1/10}(f_d) \leq O(\log n)$.

6 One-Round Error Dependence

We now consider the general question of error dependence under distributions with limited information. In the case, where the information is bounded only by n, we get the standard randomised (resp. quantum) communication complexity, for which the usual boosting techniques (i.e., the Chernoff bound) show that the error dependence is at most factor of $O(\log(1/\epsilon))$. Furthermore, Corollary 20 shows that for all information parameters $1 < k < n^{1-\Omega(1)}$ the error dependence is polynomial. This leaves the case of product distributions, where in the randomised two-way communication case DISJ has logarithmic error dependence. In this section we show that for *all* total functions, in the case of one-way communication complexity the error dependence is small under product distributions. The corresponding statement about two-way protocols remains open.

In [19] Kremer et al. show that the complexity of one-way protocols for total functions under product distributions is determined by the VC-dimension (see also [17]).

▶ **Definition 34.** The VC-dimension of a Boolean matrix M is the largest k such that there is a $2^k \times k$ rectangle R in M such that R contains all Boolean strings of length k as rows.

The VC-dimension in turn characterises the number of examples needed to PAC-learn the concept class given by the rows of the communication matrix of f, under any distribution on the columns. Usually in learning theory a concept class is a set of Boolean functions ('concepts'), and here we view rows of the communication matrix of f as functions $f_x(y) = f(x,y)$. The task of PAC learning is for the learner to be able to compute $f_x(y)$ for most y under a distribution μ, after having seen labelled examples from the same distribution. It is well known, that $O(VC(f) \cdot 1/\epsilon \cdot \log(1/\epsilon))$ examples suffice [17].

Kremer et al. [19] proved the following upper bound on one-way communication complexity: $R_\epsilon^{A\to B,I=0}(f) \leq O(VC(f)\cdot 1/\epsilon \cdot \log(1/\epsilon))$. The idea is that Alice and Bob can choose examples y' from the public coin, which Alice can label by sending $f(x,y')$. Bob simulates the PAC learning algorithm for the rows of the communication matrix, and hence he can successfully predict $f(x,y)$ for most y, including (likely) his own input. Note that there is also a lower bound of $Q^{A\to B,I=0}(f) \geq (1-H(\epsilon))VC(f)$ (which is even true in the entanglement assisted case with an additional factor of 1/2)[19, 3, 18].

While it is known, that the number of examples needed to PAC-learn is at least $\Omega(VC(f)/\epsilon)$ [17], we get an exponentially better dependence on the error here for the one-way communication model under product distributions.

Our result has an appealing interpretation. Both the one-way model under product distributions and the PAC model can be viewed as learning models (for this it is crucial that the distributional one-way model is considered under product distributions). In the PAC model Alice (or nature) labels random examples drawn from a distribution, and Bob has to end up being able to label new examples mostly correct (under the same, unknown distribution). In the one-way model, there is a known distribution on examples (columns), and a known distribution on concepts (rows). The one-way model under product distributions can clearly simulate any PAC algorithm. But Alice can send any information she deems useful, not just label examples. Nevertheless, in both models the complexity is determined by the VC-dimension. Is a teacher like Alice not more useful than random labelled examples? We show that the one-way model (i.e., a teacher) is better in the sense that making the error small is exponentially cheaper there, compared to the PAC model.

▶ **Theorem 35.** *For all total f: $R_\epsilon^{A\to B,I=0}(f) \leq O(Q_{1/3}^{A\to B,I=0}(f)\cdot \log(1/\epsilon))$*

Proof. First, $Q_{1/3}^{A\to B,I=0}(f) = \Theta(VC(f))$. Hence we need to show only that $R_\epsilon^{A\to B,I=0}(f) \leq O(VC(f)\cdot \log(1/\epsilon))$.

For a given distribution μ on the columns, an ϵ-net among the rows of the communication matrix is a subset N of the set of rows, such that for every row x there is a row $x' \in N$ which coincides with x with probability $1-\epsilon$ under μ. We have the following simple observation, due to the fact that Alice can simply send the name of the closest $x' \in N$ to Bob.

▶ **Lemma 36.** *$R_\epsilon^{\mu_A \times \mu_B}(f)$ is upper bounded by the logarithm of the size of the smallest ϵ-net for f and μ_B.*

Hence instead of the simulation Alice and Bob can agree on an ϵ-net beforehand, and the size of the ϵ-net determines the complexity of the protocol. Note that PAC-learners also try to find an ϵ-net, but they are restricted to finding one from random examples. The size of the constructed ϵ-net is much smaller than the number of examples (this is not surprising, since otherwise the concept is not learned yet). Indeed, Sauer's lemma tells us enough about the size of the ϵ-net, when the specified number of examples have been chosen.

▶ **Fact 37** (Sauer). *Let M be a Boolean matrix with r rows and c columns and VC-dimension d. Then $r \leq \Phi(c,d)$, where $\Phi(c,d) = \sum_{i=0,\ldots,d} \binom{c}{i} \leq d \cdot \binom{c}{d}$.*

We now state the fundamental result from PAC learning (see Theorem 3.3 in [17]).

▶ **Fact 38.** *Consider any function $f : \{0,1\} \times \{0,1\}^n \to \{0,1\}$. Assume we fix any x, and there is a distribution μ on y's that does not depend on x. We are given $c = O(VC(f) \cdot 1/\epsilon \cdot \log(1/\epsilon))$ random examples $y_1, \ldots, y_c$ from the distribution and labels $\ell_1 = f(x,y_1), \ldots, \ell_c = f(x,y_c)$. If we use any x' that is consistent with these values, i.e., $f(x',y_i) = \ell_i$ for all $i = 1,\ldots,c$, then the probability that $f(x',y) \neq f(x,y)$ is at most ϵ under μ, i.e., if we choose a string x' consistent with any vector $\ell_1,\ldots,\ell_c$, then we get an ϵ-net for f,μ.*

The size of this ϵ-net is clearly at most 2^c. Sauer's lemma can be used to show that the constructed ϵ-net is actually much smaller. The size of the ϵ-net constructed in Fact 38 is at most the size of the set of *distinct* rows in the matrix for f, when we restrict the matrix to the c chosen columns (we may choose one x' for every distinct value of the c labels appearing and add it into the ϵ-net).

The size of the number of distinct rows is bounded now by Sauer's lemma as follows: $VC(f) \cdot \binom{c}{VC(f)} = VC(f) \cdot \binom{const \cdot VC(f) \cdot 1/\epsilon \cdot \log(1/\epsilon)}{VC(f)} \leq (1/\epsilon)^{O(VC(f))}$. Hence the communication is at most the logarithm of this size, which yields the theorem. ◀

7 Open Problems

- Can the error dependence of a tight upper bound on $Q_\epsilon^{I=0}(DISJ)$ be improved to $\log(1/\epsilon)$?
- Can the error dependence of $R_\epsilon^{I=0}(f)$ be improved to $\log(1/\epsilon)$ for *every* total function f?
- What is the trade-off between the number of rounds and the randomised complexity of DISJ under product distributions?
- What is the quantum communication complexity of DISJ where the inputs are sets of size $\sqrt{n}$ from a size n universe? The best known lower bound is $\Omega(n^{1/4})$, the best known upper bound is $O((n^{1/4})\log n)$.
- What is the largest gap between $Q^{I=0}(f)$ and $R^{I=0}(f)$? In the one-way model there is at most a constant gap for any total function. We have shown a quadratic gap for DISJ.

References

1 S. Aaronson and A. Ambainis. Quantum search of spatial regions. *Theory of Computing*, 1(1):47–79, 2005. Earlier version in FOCS'03. quant-ph/0303041.

2 Noga Alon, Shay Moran, and Amir Yehudayoff. Sign rank, VC dimension and spectral gaps. *Electronic Colloquium on Computational Complexity (ECCC)*, 21:135, 2014.

3 A. Ambainis, A. Nayak, A. Ta-Shma, and U. V. Vazirani. Dense quantum coding and quantum finite automata. *Journal of the ACM*, 49(4):496–511, 2002. Earlier version in STOC'99.

4 L. Babai, P. Frankl, and J. Simon. Complexity classes in communication complexity theory. In *Proceedings of 27th IEEE FOCS*, pages 337–347, 1986.

5 Z. Bar-Yossef, T. S. Jayram, R. Kumar, and D. Sivakumar. An information statistics approach to data stream and communication complexity. In *Proceedings of 43rd IEEE FOCS*, pages 209–218, 2002.

6 H. Buhrman, R. Cleve, and A. Wigderson. Quantum vs. classical communication and computation. In *Proceedings of 30th ACM STOC*, pages 63–68, 1998. quant-ph/9802040.

7 B. Chor and O. Goldreich. Unbiased bits from sources of weak randomness and probabilistic communication complexity. *SIAM Journal on Computing*, 17(2):230–261, 1988. Earlier version in FOCS'85.

8 T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley, 1991.

9 Ronald de Wolf. Quantum communication and complexity. *Theoretical Computer Science*, 287(1):337–353, 2002.

10 Prahladh Harsha, Rahul Jain, David A. McAllester, and Jaikumar Radhakrishnan. The communication complexity of correlation. *IEEE Transactions on Information Theory*, 56(1):438–449, 2010.

11 Johan Håstad and Avi Wigderson. The randomized communication complexity of set disjointness. *Theory of Computing*, 3(1):211–219, 2007.

12 R. Jain, H. Klauck, and A. Nayak. Direct product theorems for classical communication complexity via subdistribution bounds. In *Proc. of 40th ACM STOC*, pages 599–608, 2008.

13 R. Jain, J. Radhakrishnan, and P. Sen. Privacy and interaction in quantum communication complexity and a theorem about the relative entropy of quantum states. In *Proceedings of 43rd IEEE FOCS*, pages 429–438, 2002.

14 R. Jain and S. Zhang. New bounds on classical and quantum one-way communication complexity. *Theoretical Computer Science*, 410(26):2463–2477, 2009.

15 Rahul Jain, Jaikumar Radhakrishnan, and Pranab Sen. A direct sum theorem in communication complexity via message compression. In *ICALP*, page 187, 2003.

16 B. Kalyanasundaram and G. Schnitger. The probabilistic communication complexity of set intersection. *SIAM Journal on Discrete Mathematics*, 5(4):545–557, 1992. Earlier version in Structures'87.

17 Michael J. Kearns and Umesh V. Vazirani. *An Introduction to Computational Learning Theory.* MIT Press, 1994.

18 H. Klauck. On quantum and probabilistic communication: Las Vegas and one-way protocols. In *Proceedings of 32nd ACM STOC*, pages 644–651, 2000.

19 I. Kremer, N. Nisan, and D. Ron. On randomized one-round communication complexity. *Computational Complexity*, 8(1):21–49, 1999. Earlier version in STOC'95. Correction at `http://www.eng.tau.ac.il/~danar/Public/KNR-fix.ps`.

20 E. Kushilevitz and N. Nisan. *Communication Complexity.* Cambridge Univ. Press, 1997.

21 Marco Molinaro, David P. Woodruff, and Grigory Yaroslavtsev. Amplification of one-way information complexity via codes and noise sensitivity. In *ICALP*, pages 960–972, 2015.

22 A. Razborov. On the distributional complexity of disjointness. *Theoretical Computer Science*, 106(2):385–390, 1992.

23 A. Razborov. Quantum communication complexity of symmetric predicates. *Izvestiya of the Russian Academy of Sciences, mathematics*, 67(1):159–176, 2003. quant-ph/0204025.

24 Mert Saglam and Gábor Tardos. On the communication complexity of sparse set disjointness and exists-equal problems. In *54th Annual IEEE Symposium on Foundations of Computer Science, FOCS*, pages 678–687, 2013.

25 Alexander A. Sherstov. Communication complexity under product and nonproduct distributions. In *Proceedings of the 23rd Annual IEEE Conference on Computational Complexity, CCC*, pages 64–70, 2008.

A Randomised Protocol for DISJ under Product Distributions

Proof of Theorem 15. Fix any product distribution μ on $\{0,1\}^n \times \{0,1\}^n$. The main idea is (just like in [4]) to have a first phase in which large sets are reduced in size until both sets have size $O(\sqrt{n})$. In phase 2 we employ the randomised protocol for DISJ on small sets given by Hastad and Wigderson [11] (instead of communicating the sets). To simplify our presentation we describe a randomised protocol.

Set $S = \sqrt{n}$. In phase 1 Alice and Bob try to shrink the universe U (without removing positions in $x \cap y$) until the size of U is at most S. At that point also $|x \cap U|$ and $|y \cap U|$ are at most of size S and the players move to phase 2. The protocol starts with the universe $U_0 = \{1, \dots, n\}$. The players maintain a current universe U_i until U_i is small at some point.

The protocol proceeds in rounds during phase 1 (we later explain how to get rid of all but two rounds). In each round Alice and Bob exchange a bit each, indicating whether $|x|, |y| \geq S$ or not. If both are smaller, they move to phase 2. The players also maintain a current rectangle of inputs $R_i = A_i \times B_i$ (this would be immediate in a deterministic protocol, but needs to be maintained in the randomised case).

After this exchange, Alice and Bob each compute $Prob(x, y$ are disjoint) on the current distribution restricted to R_i and their row/column. If this probability is less than ϵ for someone, they reject and quit the protocol. Otherwise, one player who has a large set still, say Alice, uses the public coin to generate samples $y' \in B_i$. These are disjoint from x with probability at least ϵ. Hence, Alice can name a disjoint y' with expected communication $O(\log(1/\epsilon))$. Since $x \cap y$ is disjoint with y' they set $U_{i+1} = U_i - y'$. The size of the universe decreases by at least $\sqrt{n}$ in each round in phase 1, the communication is expected $O(\log(1/\epsilon))$ per round, and there are at most $\sqrt{n}$ rounds.

Phase 2, as mentioned, is the protocol from [11], which solves DISJ with communication $O(\sqrt{n}\log(1/\epsilon))$ and worst case error ϵ on sets of size at most $\sqrt{n}$.

Hence the total expected communication is at most $O(\sqrt{n}\log(1/\epsilon))$. We need a protocol with a *worst case* communication bound, though, but note that during each round in phase 1, using the public coin to pick a new y' corresponds to a Bernoulli trial with success probability at least ϵ. The communication cost is the logarithm of the number of the first successful trial. The probability that this is larger than $t\log(1/\epsilon)$ is at most $e^{-1/\epsilon^{t-1}}$. Assume there are T rounds in phase 1. The probability that the message length in any round is more than $(T+1)\log(1/\epsilon)$ is at most $T \cdot e^{-1/\epsilon^T} \leq \epsilon$. Hence we can assume that the message length is at most $(T+1)\log(1/\epsilon)$ in all rounds (the probability that this is not the case is bounded by ϵ).

We now bound the probability that the total message length is more than $10T\log(1/\epsilon)$, by appealing to the Hoeffding bound. Note that the message lengths of all rounds are (still) independent, and that we just established an upper bound on the message length. The Hoeffding bound now implies that the probability of the total message length being larger than the stated bound is at most ϵ. Furthermore, we have that $T \leq \sqrt{n}$ with certainty. This shows that the communication of phase 1 is at most $O(\sqrt{n}\log(1/\epsilon))$. Note that the protocol needs to be modified such that it aborts if the communication in phase 1 exceeds this bound. This introduces error at most ϵ. ◀

B Randomised Protocol and Distributions with Bounded Mutual Information

Proof of Theorem 16. Fix any distribution μ' that has information at most k. The protocol we describe again has 2 phases. Informally, the first phase shrinks the sets of Alice and

Bob (which could be arbitrarily large) until their sizes are both small enough. The second phase is small set disjointness, as considered before by Hastad and Wigderson [11], and more recently by Saglam and Tardos [24]. We will establish an upper bound of $O(\sqrt{n(k+1)}/\epsilon)$ on the *expected* communication complexity with error ϵ. Then the theorem (which claims a worst-case bound) follows via the Markov inequality: if the stated communication bound is violated, stop the protocol and output a random bit.

Set $S=\sqrt{(k+1)n}$. The goal of the first phase is to make both sets smaller than S. Suppose Alice holds x and Bob y. They communicate to determine one of them has a set larger than S. This needs communication $O(1)$. If both sets are small we move to phase 2 described below.

In phase 1 Alice and Bob try to shrink the universe U until the size of U is at most S. At that point also $|x\cap U|$ and $|y\cap U|$ are at most S and the players move to phase 2. The protocol starts with the universe $U_0=\{1,\ldots,n\}$. The players maintain a current universe U_i until U_i is small at some point.

Note that while the information under $\mu_0=\mu$ is at most k, in some branches of the protocol the information on the current sub-rectangle can grow, and we need that on average it is bounded by k. We keep a transcript $T_i=U_i,V_i,R_i$, which contains the messages exchanged in phase 1 up to round i (in every round either Alice or Bob sends a message, which goes into U_i resp. V_i), as well as the random variable R_i containing the public coins used so far. Note that conditioned on a fixed value r of R_i the message transcript $U_i(r)\times V_i(r)$ is a rectangle in the communication matrix.

Then $I(X:Y|T_i)=H(XU_iV_i|R_i)+H(YU_iV_i|R_i)-H(U_iV_i|R_i)-H(XYU_iV_i|R_i)\leq H(XU_i|R_i)+H(YV_i|R_i)-H(XYU_iV_i|R_i)+I(U_i:V_i|R_i)=I(XU_i:YV_i|R_i)=I(X:Y|R_i)=I(X:Y)$, hence the information does not increase on average.

Denote by μ_{t_i} the distribution on inputs conditional on the transcript being $T_i=t_i$. $\mu^x_{t_i}$ is μ_{t_i} restricted to the row $X=x$. $\tilde{\mu},\tilde{\mu}_{t_i},\tilde{\mu}^x_{t_i}$ denote the distributions restricted to 1-inputs of DISJ. $\mu_{t_i,Y}$ is the marginal of μ_{t_i} on Bob's inputs. $\tilde{\mu}^x_{t_i,Y}$ is the distribution on y's under $T_i=t_i$, for fixed x and conditioned on $x\cap y=\emptyset$. $\mu^x_{t_i,Y}$ is the distribution on y's under $T_i=t_i$, for fixed x.

Here is the protocol for phase 1. Explanations follow.

1. Alice and Bob check whether $|x|\leq S$ and $|y|\leq S$ on U_i. If both are, they move to phase 2. W.l.o.g. assume that $|y|\geq S$, otherwise the following steps are done by Bob in an analogous fashion.
2. Alice computes the probability that $DISJ(x,y')=1$ if y' is chosen from $\mu^x_{t_i}$. If this probability is less than $\epsilon/2$, she ends the protocol with output 0.
3. Alice computes $\tilde{\mu}^x_{t_i,Y}$. Another distribution, this one known to both players, is $\mu_{t_i,Y}$.
4. Alice and Bob use rejection sampling as in Fact 8 (using the distributions $\tilde{\mu}^x_{t_i,Y}$ and $\mu_{t_i,Y}$) to discover a y'_i distributed according to $\tilde{\mu}^x_{t_i,Y}$.
5. Alice and Bob set $U_{i+1}=U_i-y'_i$.
6. t_{i+1} is t_i together with the message and randomness from 1. $\mu_{t_{i+1}}$ is μ conditioned on $T_{i+1}=t_{i+1}$.
7. Move to step 1.

We note the following on the different steps.

1. Communication is $O(1)$.
2. Clearly the total error introduced by these steps under μ can never be more than $\epsilon/2$. If the protocol moves ahead the probability of $DISJ(x,y)=1$ is at least $\epsilon/2$ under $\mu^x_{t_i}$.
3. Since $I(X:Y|T_i)\leq k$ we have that $E_{t_i,x}D(\mu^x_{t_i,Y}||\mu_{t_i,Y})\leq k$.

4. $D(\tilde{\mu}^x_{t_i,Y}||\mu_{t_i,Y}) \leq 2(D(\mu^x_{t_i,Y}||\mu_{t_i,Y})+1)/\epsilon - \log(\epsilon/2)$ due to Lemma 7 and hence the rejection sampling protocol from Fact 8 uses expected communication $O((k+1)/\epsilon)$. Drawn y'_i are always disjoint from x.
5. $|y'_i| \geq S$. Hence $|U_i - U_{i+1}| \geq S$. This step can be performed at most $n/\sqrt{n/(k+1)}$ times.

The protocol ends phase 1 with sets $x \cap U_j$ held by Alice and $y \cap U_j$ held by Bob, and $|x \cap U_j|, |y \cap U_j| \leq S$, and $DISJ(x,y) = 1 \Leftrightarrow DISJ(x \cap U_j, y \cap U_j) = 1$. The probability that the protocol ends during phase 1 and makes an error is at most $\epsilon/2$. The expected communication is at most $O(\sqrt{n(k+1)}/\epsilon$.

Phase 2 is simply the Hastad Wigderson protocol for small set disjointness [11], that finishes the protocol in communication $O(\sqrt{n(k+1)}\log(1/\epsilon))$ and with worst case error $\epsilon/2$. Hence we get a protocol with error ϵ, and expected communication $O(\sqrt{n(k+1)}/\epsilon)$. ◀

C Randomised Lower Bound for DISJ

We first bound the information. Letting X and Y follow the marginal distributions of $\mu_{n,k}$, respectively, we have:

$$\begin{aligned}
I(X:Y) = H(X) - H(X|Y) &= \log\binom{n}{m} - \mathbf{E}_{y\in Y}(Pr(y)H(X|Y=y)) \\
&= \log\binom{n}{m} - H(X|Y=y_0) \text{ (where } y_0 \text{ is any set with } P(y_0)>0) \\
&= \log\binom{n}{m} - \left(2\log\binom{n-m}{m} + 2\log\binom{n-m}{m-1} + 2\right) \\
&\leq \log\binom{n}{m} - \log\binom{n-m}{m} = \log\frac{n(n-1)\cdot\ldots\cdot(n-m+1)}{(n-m)\cdot\ldots\cdot(n-2m+1)} \\
&\leq \log\left(1+\frac{m}{n-2m+1}\right)^m \leq (\log e)\frac{m^2}{n-2m+1} \\
&= c^2(\log e)(1+o(1))(k+1) \leq k
\end{aligned}$$

for any sufficiently large n.

Proof of Theorem 18. We may assume that $k = o(n)$, since otherwise (if $k = \Omega(n)$), the original proof by Razborov [22] applies directly. Let $l \in \mathbb{N}$ be given and assume that $n = 4l-1$. Let $\gamma = \log_l(c\sqrt{n(k+1)})$, where $c = (\log e)^{-1}$. Thus $\gamma \in \left(\frac{1}{2},1\right)$ (for n sufficiently large) and our distribution will pick sets of size $l^\gamma = c\sqrt{n(k+1)}$. Throughout the proof we will treat numbers like l^γ as natural numbers, and avoid using the floor function for the sake of readability. We will also identify $\mathcal{P}(\{1,\ldots,n\})$ with $\{0,1\}^n$.

We now give an alternative definition for the distribution $\mu = \mu_{n,k}$, as the distribution induced by the following process: First, a triple $T = (T_1, T_2, i)$ is chosen uniformly among all such triples, where $|T_1| = |T_2| = 2l-1$ and $\{T_1, T_2, \{i\}\}$ form a partition of the set $\{1,\ldots,n\}$. Then, with probability $\frac{1}{2}$ the set x is chosen uniformly among all subsets of $T_1 \cup \{i\}$ with l^γ elements and such that they contain i, and with probability $\frac{1}{2}$ the set x is chosen as a subset of T_1 with l^γ elements, again uniformly among all such subsets of T_1. Similarly, and independently of the choice of x, y is chosen with probability $\frac{1}{2}$ uniformly as a subset of $T_2 \cup \{i\}$ with l^γ elements and such that it contains i, and with probability $\frac{1}{2}$ uniformly among the subsets of T_2 with l^γ elements (not containing i). Thus non-zero probabilities are assigned only on the set $\{(x,y) \mid x,y \subseteq \{1,\ldots,n\}, |x| = |y| = l^\gamma, |x \cap y| \in \{0,1\}\}$.

Now the statement that $D_{\epsilon}^{\mu_{n,k}}(DISJ) = \Omega(\sqrt{n(k+1)})$ for any sufficiently small constant $\epsilon > 0$, follows directly from Lemma 39 below. ◀

▶ **Lemma 39.** *Let γ and μ be defined as in the proof of Theorem 18. Let $A = \{(x,y) \mid \mu(x,y) > 0 \text{ and } x \cap y = \emptyset\}$ and $B = \{(x,y) \mid \mu(x,y) > 0 \text{ and } x \cap y \neq \emptyset\}$. For any sufficiently small $\epsilon > 0$ we have for any rectangle $R = C \times D \subseteq \mathcal{P}(\{1,\dots,n\})^2$ that*

$$\mu(B \cap R) \geq \Omega(\mu(A \cap R)) - 2^{-\Omega(n^{\gamma})}.$$

Proof. We consider $\epsilon > 0$ to be fixed (but will specify its value later). We begin by defining for any triple $T = (T_1, T_2, \{i\})$ as above, the numbers $Row(T) = Pr[x \in C \mid x \subseteq T_1 \cup \{i\}]$, $Row_0(T) = Pr[x \in C \mid x \subseteq T_1 \cup \{i\}, i \notin x]$ and $Row_1(T) = Pr[x \in C \mid x \subseteq T_1 \cup \{i\}, i \notin x]$, and similarly $Col(T) = Pr[y \in D \mid y \subseteq T_2 \cup \{i\}]$, $Col_0(T) = Pr[y \in D \mid y \subseteq T_2 \cup \{i\}, i \notin y]$ and $Col_1(T) = Pr[y \in D \mid y \subseteq T_2 \cup \{i\}, i \notin y]$. It is important to note that $Row(T) = \frac{1}{2}(Row_0(T) + Row_1(T))$ and $Col(T) = \frac{1}{2}(Col_0(T) + Col_1(T))$, just as in the case of Razborov's original distribution, and for the same reasons.

Next, for a triple $T = (T_1, T_2, \{i\})$ (and under the above distribution μ) we say that T is *x-bad* if $Row_1(T) < \frac{1}{6}Row_0(T) - 2^{-\epsilon n^{\gamma}}$, and that T is *y-bad* if $Col_0(T) < \frac{1}{6}Col_0(T) - 2^{-\epsilon n^{\gamma}}$. If T is x-bad or y-bad, we say that T is *bad*. Let $Bad_x(T)$, $Bad_y(T)$ and $Bad(T)$ be the respective event indicators.

▶ **Claim 40.** *For all $t_2 \subseteq \{1,\dots,n\}$, with $|t_2| = 2l-1$, we have that $Pr[Bad_x(T) = 1 \mid T_2 = t_2] \leq \frac{1}{5}$ and $Pr[Bad_y(T) = 1 \mid T_2 = t_2] \leq \frac{1}{5}$.*

Proof of the Claim. We prove the first statement, the second one having an almost identical proof.

Let $t_2 \subseteq \{1,\dots,n\}$, with $|t_2| = 2l-1$, be fixed. Under our distribution, $Row(T)$ can take different values even when T is restricted to partitions for which $T_2 = t_2$. Thus we first treat the case when $\max\{Row(T) \mid T_2 = t_2\} \leq 2^{-\epsilon n^{\gamma}}$. If this inequality holds, then for all T with $T_2 = t_2$ we have: $Row(T) \leq 2^{-\epsilon n^{\gamma}}$, and hence $Row_0(T) \leq 2Row(T) \leq 2 \cdot 2^{-\epsilon n^{\gamma}}$ so that $\frac{Row_0(T)}{6} - 2^{-\epsilon n^{\gamma}} < 0 \leq Row_1(T)$ holds trivially (and hence $Pr[Bad_x(T) = 1 \mid T_2] = 0$).

Next we treat the case where $\max\{Row(T) \mid T_2 = t_2\} > 2^{-\epsilon n^{\gamma}}$. Define $S = \{x \in C \mid |x| = l^{\gamma}, x \subset \{1,\dots,n\} \setminus t_2\}$. Note that for any T with $T_2 = t_2$, $Row(T)$ measures the conditional probability (conditioned on T) of the same set S, with each $x \in S$ having a different (conditional) probability depending on whether $i \in x$. Specifically, if $i \in x$ then the probability of x being chosen, conditioned on T, is $\frac{1}{2}\binom{2l-1}{l^{\gamma}-1}^{-1}$, otherwise the probability is $\frac{1}{2}\binom{2l-1}{l^{\gamma}}^{-1} = \frac{1}{2}\binom{2l-1}{l^{\gamma}-1}^{-1}\frac{l^{\gamma}}{2l-l^{\gamma}} = \frac{1}{2}\binom{2l-1}{l^{\gamma}-1}^{-1}\frac{1}{2l^{1-\gamma}-1}$. Thus, when T is fixed, the probability of each set x containing i is $2l^{1-\gamma} - 1$ times that of a set which does not contain i.

The proof of this case will proceed as follows: First, we show that under the assumption that a sufficiently large part of the partitions T with $T_2 = t_2$ are x-bad, three quarters of the elements of S (which are subsets of $\{1,\dots,n\} \setminus T_2$) must have at least $\frac{21}{25}$ of their elements in a subset of $\{1,\dots,n\} \setminus T_2$ of size $\frac{8l}{5}$. We will then upper-bound the number of subsets of $\{1,\dots,n\} \setminus T_2$ of size l^{γ} that have this property (regardless of whether they are in C or not). Next, we will lower-bound $\frac{3}{4}|S|$ in terms of ϵ, and show that for a suitable choice of ϵ, the lower bound for $\frac{3}{4}|S|$ is in fact larger than the upper bound we computed before, which is a contradiction showing that it is not possible for that T with $T_2 = t_2$ to be x-bad for that many choices of i.

Note first that whenever T_2 is fixed (in our case to t_2), the choice of $i \in \{1,\dots,n\} \setminus T_2$ also fixes T_1 and hence all of T, and that the choice of i determines the proportion of $x \in S$ whose weights are counted in $Row_1(T)$. If for a particular choice of i the resulting T is

x-bad, then by definition we have that $Row_1(T) < \frac{1}{6}Row_0(T) - 2^{-\epsilon n^\gamma}$, and in particular that $Row_1(T) < \frac{1}{6}Row_0(T)$. If we let S' be the set of $x \in S$ with $i \in x$, then we may rewrite this inequality as:

$$\frac{|S'|}{\binom{2l-1}{l^\gamma-1}} < \frac{|S|-|S'|}{6\binom{2l-1}{l^\gamma}} \iff |S'| < \frac{|S|-|S'|}{6(2l^{1-\gamma}-1)}$$

$$\iff |S'|\left(1+\frac{1}{6(2l^{1-\gamma}-1)}\right) < \frac{|S|}{6(2l^{1-\gamma}-1)},$$

and we may conclude that for l sufficiently large, $|S'| < \frac{|S|}{10l^{1-\gamma}}$ (under the assumption that T is x-bad). For the last inequality we have used the fact that $\lim_{n\to\infty} l^{1-\gamma} = \infty$, which holds because: $\lim_{n\to\infty} \log l^{1-\gamma} = \lim_{n\to\infty}(1-\gamma)\log l = \lim_{n\to\infty}(1-\log_l(c\sqrt{n(k+1)}))\log l \geq \lim_{n\to\infty}(\log l - \log\sqrt{n(k+1)}) \geq \lim_{n\to\infty}\log\sqrt{\frac{n+1}{16(k+1)}} = \infty$ (since $k = o(n)$).

Let $B = \{i \in \{1,\ldots,n\} \mid \text{the partition } (\{1,\ldots,n\} \setminus (t_2 \cup \{i\}), t_2, \{i\}) \text{ is } x\text{-bad}\}$, and assume that $|B| \geq \frac{2l}{5}$, that is, assume that for at least one fifth of the possible choices for i the corresponding partition is x-bad. By excluding some elements of B, we may assume that $|B| = \frac{2l}{5}$. Now, if we consider the number of pairs (x,i) with $x \in S$ and $i \in x$, we have by the inequality in the last paragraph that each of the $i \in B$ can be the second element of at most $\frac{|S|}{10l^{1-\gamma}}$ such pairs, and hence B can contribute the second element of at most $\frac{2l}{5}\frac{|S|}{10l^{1-\gamma}} = \frac{1}{25}l^\gamma|S|$ of the total of $l^\gamma|S|$ pairs. Applying the Colouring Lemma below with $X = S$, $Y = \{1,\ldots,l^\gamma\}$, $c(x,i) = 0$ if and only if the i-th smallest element of x is in B (so that $p \geq \frac{24}{25}$) and $r = \frac{21}{25}$, we have that at least three quarters of all $x \in S$ have the property that more than $\frac{21}{25}$ of their elements lie in $G = \{1,\ldots,n\} \setminus (t_2 \cup B)$. Let Q be the set of subsets $x \subseteq B \cup G = \{1,\ldots,n\} \setminus t_2$, with $|x| = l^\gamma$ and the property that $|x \cap G| \geq \frac{21}{25}l^\gamma$. Then we must have that $|Q| \geq \frac{3}{4}|S|$. We will now upper-bound the size of the set Q.

Since every $x \in Q$ can have a proportion of at most $4/25$ of its elements in B, we have that

$$\begin{aligned}
\log|Q| &\leq \log\left[\sum_{i=0}^{\frac{4}{25}l^\gamma}\binom{\frac{2l}{5}}{i}\binom{\frac{8l}{5}}{l^\gamma-i}\right] \leq \log\left[\sum_{i=0}^{\frac{4}{25}l^\gamma}\left(\frac{2le}{5i}\right)^i\left(\frac{8le}{5(l^\gamma-i)}\right)^{l^\gamma-i}\right] \\
&\leq \log\left[\frac{4}{25}l^\gamma\left(\frac{2le}{5}\frac{25}{4l^\gamma}\right)^{\frac{4}{25}l^\gamma}\left(\frac{8le}{5}\frac{25}{21l^\gamma}\right)^{\frac{21}{25}l^\gamma}\right] \\
&= \log\left[\frac{4}{25}l^\gamma\left(\frac{5e}{2}l^{1-\gamma}\right)^{\frac{4}{25}l^\gamma}\left(\frac{40e}{21}l^{1-\gamma}\right)^{\frac{21}{25}l^\gamma}\right] \\
&\leq \gamma\log l + \frac{4}{25}l^\gamma\log\left(\frac{5e}{2}l^{1-\gamma}\right) + \frac{21}{25}l^\gamma\log\left(\frac{40e}{21}l^{1-\gamma}\right) + O(1) \\
&= (1-\gamma)l^\gamma\log l + \left(\frac{4}{25}\log\frac{5e}{2} + \frac{21}{25}\log\frac{40e}{21}\right)l^\gamma + O(\log l) \\
&\leq (1-\gamma)l^\gamma\log l + 2.43508\cdot l^\gamma + O(\log l),
\end{aligned}$$

where in the first line we used the inequality $\binom{m}{k} \leq \left(\frac{em}{k}\right)^k$ for each term of the sum. The inequality sign between the first and second line can be justified as follows: For $x \in (0, \frac{1}{2})$, consider the expression

$$\log\left[\left(\frac{\frac{2}{5}el}{x\cdot l^\gamma}\right)^{x\cdot l^\gamma}\left(\frac{\frac{8}{5}el}{(1-x)l^\gamma}\right)^{(1-x)l^\gamma}\right] = (1-\gamma)l^\gamma\log l + l^\gamma\left(x\log\frac{2e}{5x} + (1-x)\log\frac{8e}{5(1-x)}\right)$$

and set $f(x) = x \log \frac{2e}{5x} + (1-x) \log \frac{8e}{5(1-x)} = x \log \frac{1}{x} + (1-x) \log \frac{1}{1-x} + (1 + \log e)x + (3 + \log e)(1-x) - \log 5 = 3 + \log e - \log 5 + H(x) - 2x$. Then $f'(x) = H'(x) - 2 = \log \frac{1-x}{x} - 2$. Note that the function $\frac{1-x}{x}$ is decreasing but positive on $(0,1)$, and we have that the smallest value $x_0 \in (0, \frac{1}{2})$ for which we can have $f'(x_0) = 0$ is $x_0 = \frac{1}{5}$, which implies that $f(x)$, and hence also the argument of the logarithm in the expression above, is strictly increasing on $(0, \frac{1}{5})$. Thus the terms of the sum

$$\sum_{i=0}^{\frac{4}{25} l^\gamma} \left(\frac{2le}{5i}\right)^i \left(\frac{8le}{5(l^\gamma - i)}\right)^{l^\gamma - i}$$

are increasing, so that each term is upper-bounded by the final term, which justifies the inequality between the the first and second line above.

Next we compute a lower bound for $\frac{3}{4}|S|$. Let T^* be a partition with $T_2^* = t_2$ and $\mathit{Row}(T^*) = \max\{\mathit{Row}(T) \mid T_2 = t_2\}$. Then we have that $\frac{3}{4}|S| = \frac{3}{4}[\mathit{Row}_0(T^*)\binom{2l-1}{l^\gamma} + \mathit{Row}_1(T^*)\binom{2l-1}{l^\gamma - 1}] \geq \frac{3}{4}(\mathit{Row}_0(T^*) + \mathit{Row}_1(T^*))\binom{2l-1}{l^\gamma - 1} = \frac{6}{4}\mathit{Row}(T^*)\binom{2l-1}{l^\gamma - 1} > 2^{-\epsilon n^\gamma}\binom{2l-1}{l^\gamma - 1}$. Finally we have:

$$\begin{aligned}
\log \frac{3}{4}|S| > \log \left[2^{-\epsilon n^\gamma}\binom{2l-1}{l^\gamma - 1}\right] &\geq l^\gamma \log\left((e - o(1))\frac{2l-1}{l^\gamma - 1}\right) - \epsilon n^\gamma - \Theta(\log l) \\
&\geq (1-\gamma) l^\gamma \log l + l^\gamma \log(2(e - o(1))) - \epsilon \cdot (4l - 1)^\gamma - \Theta(\log l) \\
(\text{for large } l) &\geq (1-\gamma) l^\gamma \log l + 2.4426 \cdot l^\gamma - \epsilon \cdot (4l)^\gamma - \Theta(\log l).
\end{aligned}$$

For $\epsilon \leq \frac{1}{1000 \cdot 4^\gamma}$ we get the desired contradiction, that $\frac{3}{4}|S| > |Q|$.

The lower-bound for $\binom{2l-1}{l^\gamma - 1}$ above can be obtained using the Stirling bounds for the factorial, $\sqrt{2\pi n}\left(\frac{n}{e}\right)^n \leq n! \leq e\sqrt{n}\left(\frac{n}{e}\right)^n$, as follows:

$$\begin{aligned}
\binom{2l-1}{l^\gamma - 1} &\geq \frac{\sqrt{2\pi(2l-1)} \cdot (2l-1)^{2l-1}}{e^2 \sqrt{(l^\gamma - 1)(2l - l^\gamma)} \cdot (l^\gamma - 1)^{l^\gamma - 1} \cdot (2l - l^\gamma)^{2l - l^\gamma}} \\
&= \frac{\sqrt{2\pi(2l-1)}}{e^2 \sqrt{(l^\gamma - 1)(2l - l^\gamma)}} \left(\frac{2l-1}{l^\gamma - 1}\right)^{l^\gamma - 1} \left(\frac{2l-1}{(2l-1) - (l^\gamma - 1)}\right)^{2l - l^\gamma} \\
&= \frac{\sqrt{2\pi(2l-1)}}{e^2 \sqrt{(l^\gamma - 1)(2l - l^\gamma)}} \left(\frac{2l-1}{l^\gamma - 1}\right)^{l^\gamma - 1} \left[\left(1 + \frac{l^\gamma - 1}{2l - l^\gamma}\right)^{\frac{2l - l^\gamma}{l^\gamma - 1}}\right]^{l^\gamma - 1} \\
&\geq \frac{\sqrt{2\pi}}{e^2} \frac{1}{\sqrt{(l^\gamma - 1)}} \left(\frac{2l-1}{l^\gamma - 1}\right)^{l^\gamma - 1} (e - o(1))^{l^\gamma - 1}.
\end{aligned}$$

◀

▶ **Claim 41.** $\mathbf{E}[\mathit{Row}_0(T)\mathit{Col}_0(T)(1 - \mathit{Bad}(T))] > \frac{1}{5}\mathbf{E}[\mathit{Row}_0(T)\mathit{Col}_0(T)]$.

Proof of the Claim. Since $\mathit{Bad}(T) \leq \mathit{Bad}_x(T) + \mathit{Bad}_y(T)$, it is enough to prove that $\mathbf{E}[\mathit{Row}_0(T)\mathit{Col}_0(T)\mathit{Bad}_x(T)] \leq \frac{2}{5}\mathbf{E}[\mathit{Row}_0(T)\mathit{Col}_0(T)]$, with a similar statement for $\mathit{Bad}_y(T)$ being proved in the same fashion. For each $t_2 \subseteq \{1, \ldots, n\}$, with $|t_2| = 2l - 1$, we will show that the desired inequality holds when conditioned on $T_2 = t_2$, which implies that the unconditioned inequality holds. All triples T with $T_2 = t_2$ have the same value for $\mathit{Col}_0(T)$, so let this value be called c'. Also let $r = \mathbf{E}[\mathit{Row}(T) \mid T_2 = t_2]$. Now we have:

$$\begin{aligned}
\mathbf{E}[\mathit{Row}_0(T)\mathit{Col}_0(T)\mathit{Bad}_x(T) \mid T_2 = t_2] &\leq c'\mathbf{E}[\mathit{Row}_0(T)\mathit{Bad}_x(T) \mid T_2 = t_2] \\
&\leq c'\mathbf{E}\left[2 \cdot \mathbf{E}[\mathit{Row}(T) \mid T_2 = t_2] \cdot \mathit{Bad}_x(T) \mid T_2 = t_2\right] \\
&\leq 2c'r\mathbf{E}[\mathit{Bad}_x(T) \mid T_2 = t_2]
\end{aligned}$$

$$\begin{aligned}
&\leq \frac{2}{5}c'r \text{ (by Claim 1)}\\
&= \frac{2}{5}c'\mathbf{E}[\mathit{Row}(T) \mid T_2 = t_2]\\
&= \frac{2}{5}c'\mathbf{E}[\mathit{Row}_0(T) \mid T_2 = t_2]\\
&= \frac{2}{5}\mathbf{E}[\mathit{Row}_0(T)\mathit{Col}_0(T) \mid T_2 = t_2]
\end{aligned}$$

The inequality between the second and the third line can be justified as follows: Recall that, as observed in the proof of Claim 1, even when considering only triples $T_2 = t_2$, the value of $\mathit{Row}(T)$ can differ by a factor of at most $2l^{1-\gamma} - 1$. This is due to the fact that $\mathit{Row}(T)$ measures the probability (conditioned on T) of the same set $S = \{x \in C \mid |x| = l^\gamma, x \subseteq \{1, \dots, n\} \setminus t_2\}$, but depending on whether $i \in x$ (for a particular choice of i and hence of T), an $x \in S$ will have (conditional) probability either $\frac{1}{2}\binom{2l-1}{l^\gamma-1}^{-1}$ or $\frac{1}{2}\binom{2l-1}{l^\gamma}^{-1}$. Thus if T^* is a triple with $T_2^* = t_2$ for which $\mathit{Row}_0(T^*) = \max\{\mathit{Row}_0(T) \mid T_2 = t_2\}$, then $\mathit{Row}(T^*)$ must be the minimum among all values of $\mathit{Row}(T)$ when $T_2 = t_2$, because when $T = T^*$ the largest portion of elements of S have probability $\frac{1}{2}\binom{2l-1}{l^\gamma}^{-1}$ instead of $\frac{1}{2}\binom{2l-1}{l^\gamma-1}^{-1}$. It follows that for all T with $T_2 = t_2$ we have $\mathit{Row}(T) \geq \mathit{Row}(T^*) \geq \frac{1}{2}\mathit{Row}_0(T^*)$, and hence that $\mathbf{E}[2\mathit{Row}(T) \mid T_2 = t_2] \geq \mathit{Row}_0(T^*)$. On the other hand we have that for all T with $T_2 = t_2$, $\mathit{Row}_0(T) \leq \mathit{Row}_0(T^*)$, so finally we get that $\mathit{Row}_0(T) \leq 2\mathbf{E}[\mathit{Row}(T) \mid T_2 = t_2]$ for all T with $T_2 = t_2$. ◀

▶ **Claim 42.** *For any rectangle R:* $\mu(B \cap R) = \frac{1}{4}\mathbf{E}[\mathit{Row}_1(T)\mathit{Col}_1(T)]$ *and* $\mu(A \cap R) = \frac{3}{4}\mathbf{E}[\mathit{Row}_0(T)\mathit{Col}_0(T)]$ *(with the expectation taken over all partitions T).*

The proof of this claim is identical to the case where μ is the distribution in Razborov's proof (see [20]), since the relevant observations also apply to our modified distribution: 1. $\mu(B) = \frac{1}{4}$ (and hence $\mu(A) = \frac{3}{4}$), because for every fixed partition T, $i \in x$ with probability $\frac{1}{2}$ and $i \in y$ with probability $\frac{1}{2}$, independently. 2. $i \in x$ and $i \in y$ are independent events (for the same reason). 3. For every (x, y) with $x \cap y = \emptyset$ we have that $Pr[(x, y) \mid (i \notin x) \wedge (i \notin y)] = Pr[(x, y) \mid ((i \notin x) \wedge (i \notin y)) \vee ((i \in x) \wedge (i \notin y)) \vee ((i \notin x) \wedge (i \in y))]$, because conditioning on either one of the two events induces the uniform distribution on the set $\{(x, y) \mid x, y \subset \{1, \dots, n\}, x \cap y = \emptyset, |x| = |y| = l^\gamma\}$.

We now use claims 2 and 3 to prove the statement of the lemma:

$$\begin{aligned}
&\mu(B \cap R) = \frac{1}{4}\mathbf{E}[\mathit{Row}_1(T)\mathit{Col}_1(T)]\\
\geq&\frac{1}{4}\mathbf{E}[\mathit{Row}_1(T)\mathit{Col}_1(T)(1 - \mathit{Bad}(T))]\\
\geq&\frac{1}{4}\mathbf{E}\left[\left(\frac{\mathit{Row}_0(T)}{6} - 2^{-\epsilon n^\gamma}\right)\left(\frac{\mathit{Col}_0(T)}{6} - 2^{-\epsilon n^\gamma}\right)(1 - \mathit{Bad}(T))\right] \text{ (by def. of } \mathit{Bad})\\
=&\frac{1}{4}\mathbf{E}\left[\left(\frac{\mathit{Row}_0(T)\mathit{Col}_0(T)}{36} - \frac{2^{-\epsilon n^\gamma}}{6}(\mathit{Row}_0(T) + \mathit{Col}_0(T)) + 2^{-2\epsilon n^\gamma}\right)(1 - \mathit{Bad}(T))\right]\\
\geq&\Omega\left(\mathbf{E}[\mathit{Row}_0(T)\mathit{Col}_0(T)(1 - \mathit{Bad}(T))]\right) - 2^{-\epsilon n^\gamma} \text{ (since } \mathit{Row}_0(T) + \mathit{Col}_0(T) \leq 2)\\
\geq&\Omega\left(\mathbf{E}[\mathit{Row}_0(T)\mathit{Col}_0(T)]\right) - 2^{-\epsilon n^\gamma} \text{ (by Claim 2)}\\
\geq&\Omega(\mu(A \cap R)) - 2^{-\epsilon n^\gamma} \text{ (by Claim 3)}
\end{aligned}$$

Choosing ϵ to be smaller than both the constant in front of $\mu(A \cap R)$ and $\frac{1}{1000 \cdot 4^\gamma}$ completes the proof. ◀

▶ **Lemma 43** (Colouring Lemma.). *Let X and Y be non-empty finite sets, and let $c : X \times Y \mapsto \{0,1\}$ be a colouring of $X \times Y$ such that a proportion $p \in (0,1)$ of the elements of $X \times Y$ are mapped to 1, that is, such that $|c^{-1}(1)|/|X \times Y| = p$. Then for any $r \in (0,p)$ such that $r|Y| \in \mathbb{N}$, we have that for at least $\frac{p-r}{1-r}|X|$ elements $x \in X$, $|(\{x\} \times Y) \cap c^{-1}(1)| > r|Y|$.*

Proof. We call sets of the form $\{x\} \times Y$ *rows*, and let the number $w(x) = \sum_{y \in Y} c(x,y) = |(\{x\} \times Y) \cap c^{-1}(1)|$ be the *weight* of the row $\{x\} \times Y$, for each $x \in X$. Let c be a colouring of $X \times Y$ as above, but such that the smallest possible proportion of rows have weight $> r|Y|$, and denote this proportion by q. Thus q is such that for any colouring c' satisfying the conditions of the lemma, at least $q|X|$ elements $x \in X$ satisfy $|(\{x\} \times Y) \cap c^{-1}(1)| > r|Y|$.

We may assume that all rows with weight $\leq r|Y|$ have weight exactly $r|Y|$: If this is not the case, we may repeatedly perform the operation of changing a 0 into 1 on a row with weight $< r|Y|$, and a 1 into 0 on a row with weight $> r|Y|$, until the above statement is true. (It is easy to see that the colouring c must have rows with weight $> r|Y|$, since otherwise the overall proportion of elements mapped to 1 would be $\leq r < p$.) This operation leaves the proportion of elements that are mapped to 1 unchanged, and the minimality of the chosen colouring c guarantees that the number of rows with weight $> r|Y|$ does not decrease (and therefore remains unchanged).

Next, we may assume that all but at most one of the rows with weight $> r|Y|$ have weight exactly $|Y|$: If this is not the case, we may fix one such row, replace all zeroes with ones on all other rows of weight $> r|Y|$ (thus making their weight exactly $|Y|$), and on the fixed row change the same number of ones into zeroes so as to match the changes made on all other rows. Again the overall proportion of elements being mapped to 1 does not change, and the minimality of the colouring c guarantees that the weight of the fixed row stays $> r|Y|$.

Based on the above we now have: $p|X||Y| = q|X||Y| - \alpha|Y| + (1-q)|X|r|Y|$, where $\alpha \in [0, 1-r)$ is the proportion of zeroes on the one row that has weight $> r|Y|$ but not necessarily $= |Y|$. Thus we have:

$$p \leq q + (1-q)r \iff p \leq (1-r)q + r \iff \frac{p-r}{1-r} \leq q. \qquad ◀$$