Communication Complexity of Statistical Distance[*][†]

Thomas Watson

University of Memphis, Memphis, TN, USA
Thomas.Watson@memphis.edu

Abstract

We prove nearly matching upper and lower bounds on the randomized communication complexity of the following problem: Alice and Bob are each given a probability distribution over n elements, and they wish to estimate within $\pm\epsilon$ the statistical (total variation) distance between their distributions. For some range of parameters, there is up to a $\log n$ factor gap between the upper and lower bounds, and we identify a barrier to using information complexity techniques to improve the lower bound in this case. We also prove a side result that we discovered along the way: the randomized communication complexity of n-bit Majority composed with n-bit Greater-Than is $\Theta(n \log n)$.

1998 ACM Subject Classification F.1.2 Modes of Computation

Keywords and phrases Communication, complexity, statistical, distance

Digital Object Identifier 10.4230/LIPIcs.APPROX/RANDOM.2017.49

1 Introduction

Statistical (a.k.a. total variation) distance is a standard measure of the distance between two probability distributions, and is ubiquitous in theoretical computer science. Expressing the distributions (over a universe of n elements) as vectors of probabilities $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$, the statistical distance is defined as

$$\begin{aligned}\Delta(x,y) \;\coloneqq\; \tfrac{1}{2}\textstyle\sum_{i\in[n]} |x_i - y_i| \;&=\; \max_{S\subseteq[n]} \big|\textstyle\sum_{i\in S} x_i - \sum_{i\in S} y_i\big| \\ &=\; \max_{S\subseteq[n]} \big(\textstyle\sum_{i\in S} x_i - \sum_{i\in S} y_i\big).\end{aligned}$$

This measure has various interpretations, such as the minimum over all couplings of the probability that the sample from x and the sample from y are unequal, or as twice the maximum advantage an observer can achieve in guessing whether a random sample came from x or from y (where x or y is used with probability $1/2$ each).

Given its pervasiveness, it is natural to inquire about the computational complexity of estimating the statistical distance between two distributions x and y that are given as input. This topic has been studied before in several contexts:

- [25] showed that when each of x and y is succinctly represented by an algorithm that takes uniform random bits and produces a sample from that distribution (so our actual input is the description of this pair of algorithms), then (a decision version of) the problem of estimating $\Delta(x,y)$ is complete for the complexity class SZK (statistical zero knowledge). (For results about the complexity of other problems where the inputs are succinctly represented distributions, see [12, 13, 3, 14, 30, 29].)

[*] A full version of the paper is available at https://eccc.weizmann.ac.il/report/2016/170/.

[†] Supported by NSF grant CCF-1657377.

Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2017).
Editors: Klaus Jansen, José D. P. Rolim, David Williamson, and Santosh S. Vempala; Article No. 49; pp. 49:1–49:10
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

- [2, 27, 9] studied the complexity of statistical distance estimation when an algorithm is only given black-box access to oracles that produce samples from the distributions specified by x and y. (For results about the complexity of other problems where the inputs are black-box samples from distributions, see the surveys [14, 24, 7].)
- [10, 11] studied the space complexity of (a generalization of) statistical distance estimation when the vectors x and y are provided as data streams.

1.1 Communication Upper and Lower Bounds

We study the statistical distance estimation problem in the context of communication complexity: Alice is given the vector x, Bob is given the vector y, and they wish to output a value in the range $\big[\Delta(x,y)-\epsilon, \Delta(x,y)+\epsilon\big]$. We let $\textsc{Stat-Dist}_{n,\epsilon}$ denote this two-party search problem. For any two-party search problem F, we let $\mathsf{R}(F)$ denote the minimum worst-case communication cost of any randomized protocol (allowing both public and private coins) such that for each input, the output is correct with probability at least 3/4. (For our problem $\textsc{Stat-Dist}_{n,\epsilon}$, the 3/4 can be replaced by any constant in the range $(1/2, 1)$ since we can amplify success probability by taking the median of multiple trials.) The following is a clean summary of our bounds.

▶ **Theorem 1.**

$$\mathsf{R}(\textsc{Stat-Dist}_{n,\epsilon}) \text{ is } \begin{cases} \Theta(1/\epsilon^2) & \text{if } 1 > \epsilon \geq 1/O(\sqrt{n}) \\ \Omega(n) \text{ and } O(n \log n) & \text{if } 1/\omega(\sqrt{n}) \geq \epsilon \geq 1/2^{o(n \log n)} \\ \Theta(\log(1/\epsilon)) & \text{if } 1/2^{\Omega(n \log n)} \geq \epsilon > 0 \end{cases} .$$

We also go ahead and ascertain the deterministic communication complexity (denoted with D instead of R) of this problem. We prove Theorem 1 and Theorem 2 in Section 2.

▶ **Theorem 2.** $\mathsf{D}(\textsc{Stat-Dist}_{n,\epsilon}) = \Theta(n \log(1/\epsilon))$ *provided ϵ is at most a sufficiently small constant.*

Closing the gap in Theorem 1 is a principal open problem. We get slightly better bounds in certain narrow ranges of ϵ (see the proof), but e.g., it remains open to prove our conjecture that $\mathsf{R}(\textsc{Stat-Dist}_{n,1/2^n}) \geq \omega(n)$. A natural strategy is to use information complexity lower bound techniques; however, in the full version we exhibit a barrier to accomplishing this. Specifically, for a large class of inputs having a certain type of product structure (which arises naturally from attempts to use the direct sum property of information complexity), and for a wide range of ϵ, $\textsc{Stat-Dist}_{n,\epsilon}$ can be solved with $O(n)$ information cost and 0 error probability. This suggests that to improve the $\Omega(n)$ bound, we may need to look at inputs not having the aforementioned product structure, and we are at a loss for techniques in this case.

1.2 Composing with Majority

We take this opportunity to prove other results that we discovered in the process of trying to analyze $\textsc{Stat-Dist}_{n,\epsilon}$. Recall the famous direct sum conjecture stating that computing k independent copies of a two-party function should require $\Omega(k)$ times as much randomized communication as computing 1 copy. A somewhat stronger version of the conjecture states that even just computing the $\textsc{And}$ of k independent copies should still require $\Omega(k)$ times as much communication. [15] proved the query complexity analogue of this $\textsc{And}$-composition conjecture, as well as a communication complexity version that is weaker than the full conjecture in two senses: it is *qualitatively* weaker since instead of converting a protocol for $\textsc{And}_k$ composed with F into a plain randomized (BPP-type) protocol for F with factor $\Omega(k)$ savings, the conversion results in a protocol in a slightly stronger model (which has been variously called 2WAPP [16, 15], two-sided smooth rectangle bound [18], and relaxed partition bound [19]); it is *quantitatively* weaker since besides the $\Omega(k)$ savings, the conversion incurs a logarithmic additive loss due to the use of the "information odometer" of [5]. (We provide the precise statement in Section 3.)

We prove that when composing with the k-bit Majority function $\textsc{Maj}_k$ instead of $\textsc{And}_k$, the above quantitative deficiency can be avoided: we get a perfect $\Omega(k)$ factor savings by circumventing the need for the odometer (although we retain the qualitative deficiency). For the applications in [15, 1], the logarithmic additive loss in the $\textsc{And}$-composition result was immaterial albeit perhaps a slight nuisance. In some settings, however, that loss would be damaging; one such setting is the following corollary (which holds by combining our $\textsc{Maj}$-composition result with the lower bound of [6] for the Greater-Than function $\textsc{Gt}_n$ on n-bit inputs).

▶ **Theorem 3.** $\mathsf{R}(\textsc{Maj}_n \circ \textsc{Gt}_n^n) = \Theta(n \log n)$.

Evaluating the function $\textsc{Maj}_n \circ \textsc{Gt}_n^n$ can be described by a story: Alice and Bob have taken some exams and know their own scores, and they wish to determine the victor of their rivalry: who got a higher score on the most exams?

We prove the $\textsc{Maj}$-composition result and provide details about Theorem 3 in Section 3. We make the stronger conjecture that Theorem 3 should hold even with $\textsc{And}_n$ instead of $\textsc{Maj}_n$; this would follow from an $\Omega(\log n)$ information complexity lower bound for $\textsc{Gt}_n$ with respect to a distribution only over 1-inputs (which is open but may be doable).

1.3 Preliminaries

We define $\textsc{And}_n$, $\textsc{Or}_n$, $\textsc{Maj}_n$ as the And, Or, and Majority functions on n bits, and $\textsc{Eq}_n$, $\textsc{Gt}_n$, $\textsc{Disj}_n$, $\textsc{Gh}_n$ as the Equality, Greater-Than, Set-Disjointness, and Gap-Hamming two-party functions where Alice and Bob each get n bits. We use $\mathbb{P}$ for probability, $\mathbb{E}$ for expectation, $\mathbb{H}$ for Shannon entropy, and $\mathbb{I}$ for mutual information. We generally use upper-case letters for random variables and corresponding lower-case letters for particular outcomes.

Randomized protocols by default have both public and private coins. We let $CC(\Pi)$ denote the worst-case communication cost of protocol Π. We let $IC_D(\Pi) := \mathbb{I}(T\,;\,X \mid Y, R) + \mathbb{I}(T\,;\,Y \mid X, R)$ denote the (internal) information cost with respect to (X, Y) sampled from the input distribution D, where the random variables T and R represent the communication transcript and public coins of Π, respectively.

2 Communication Upper and Lower Bounds

We now prove Theorem 1 and Theorem 2. As a preliminary technicality, we note that for the upper bounds, we may assume each of the probabilities x_i and y_i can be written exactly in binary with $\log(n/\epsilon) + O(1)$ bits. This is because if we truncate the binary representations to that many bits and reassign the lost probability to an arbitrary element in both x and y, this ensures at most $\epsilon/4$ mass has been shifted within each distribution, so their statistical distance changes by at most $\epsilon/2$; then to obtain an ϵ-estimation for the original x and y, we can run a protocol to get an $(\epsilon/2)$-estimation for the new x and y.

Proof of Theorem 1. In fact, we show that $\mathsf{R}(\textsc{Stat-Dist}_{n,\epsilon})$ is always

(i) $O(1/\epsilon^2)$,
(ii) $O(\max(n \log n, \log(1/\epsilon)))$,
(iii) $\Omega(\min(1/\epsilon^2, n))$,
(iv) $\Omega(\log(1/\epsilon))$,

which gives a slightly more detailed picture than the statement of Theorem 1.

The proof of (i) is inspired by the "correlated sampling lemma" that has been used in the context of parallel repetition [17, 22, 23] and earlier in the context of LP rounding [20]. As noted above, we may assume each probability x_i and y_i is a multiple of $1/m$ for some integer $m \coloneqq O(n/\epsilon)$. We make use of an $O(1)$-communication equality testing protocol that accepts with probability 1 when the inputs are equal and accepts with probability exactly 1/2 when the inputs are unequal (e.g., by using the inputs to index into a uniformly random public string and comparing the bits at those indices).

Here is the protocol witnessing (i). Alice and Bob repeat the following $O(1/\epsilon^2)$ times:

- Publicly sample a uniformly random ordering of $[n] \times [m]$.
- Alice finds the first (i_A, j_A) in the ordering such that $x_{i_\text{A}} \geq j_\text{A}/m$.
- Bob finds the first (i_B, j_B) in the ordering such that $y_{i_\text{B}} \geq j_\text{B}/m$.
- Run the equality test on (i_A, j_A) and (i_B, j_B).

Then they output $q/(1-q)$ where $q \coloneqq \min(1/2, \text{fraction of iterations where equality test rejected})$.

To analyze the correctness, let $\delta \coloneqq \Delta(x, y)$ and let p denote the probability the equality test rejects in a single iteration of the loop. We claim that $p = \delta/(1+\delta)$ (and hence $\delta = p/(1-p)$). To see this, define the following subsets of $[n] \times [m]$: $A \coloneqq \big\{(i,j) : x_i \geq j/m \text{ and } y_i < j/m\big\}$, $B \coloneqq \big\{(i,j) : x_i < j/m \text{ and } y_i \geq j/m\big\}$, and $C \coloneqq \big\{(i,j) : x_i \geq j/m \text{ and } y_i \geq j/m\big\}$. Then $|A| = |B| = \delta m$ and $|C| = (1-\delta)m$. The first (i^*, j^*) in the ordering to land in $A \cup B \cup C$ is uniformly distributed in that set. Thus with probability $\delta/(1+\delta)$ we have $(i^*, j^*) \in A$, in which case $(i_\text{A}, j_\text{A}) = (i^*, j^*) \neq (i_\text{B}, j_\text{B})$, and with probability $\delta/(1+\delta)$ we have $(i^*, j^*) \in B$, in which case $(i_\text{A}, j_\text{A}) \neq (i^*, j^*) = (i_\text{B}, j_\text{B})$, and with probability $(1-\delta)/(1+\delta)$ we have $(i^*, j^*) \in C$, in which case $(i_\text{A}, j_\text{A}) = (i^*, j^*) = (i_\text{B}, j_\text{B})$. It follows that the equality test rejects with probability $\frac{\delta}{1+\delta} \cdot \frac{1}{2} + \frac{\delta}{1+\delta} \cdot \frac{1}{2} + \frac{1-\delta}{1+\delta} \cdot 0 = \delta/(1+\delta)$.

By a Chernoff bound, the number of iterations guarantees that with probability at least 3/4, $|q-p| \leq \epsilon/8$. Since $\frac{d}{dp}\big[p/(1-p)\big] = 1/(1-p)^2 \in [1,4]$ for all $p \in [0, 1/2]$, it follows that $|\text{output} - \delta| = \big|q/(1-q) - p/(1-p)\big| \leq \epsilon/2$ whenever $|q-p| \leq \epsilon/8$ and $q \in [0, 1/2]$. This proves (i).

To prove (ii), we exploit the fact that the Greater-Than function $\textsc{Gt}_k$ with k-bit inputs can be computed with error probability $\gamma > 0$ and $O(\log(k/\gamma))$ bits of communication (by running the standard binary-search-based protocol [21, p. 170] for $O(\log(k/\gamma))$ many steps). As noted above, we may assume each probability x_i and y_i has $\log(n/\epsilon) + O(1)$ bits.

Here is the protocol witnessing (ii). For each $i \in [n]$, Alice and Bob compute $\textsc{Gt}(x_i, y_i)$ with error probability $1/(4n)$. Then Alice sends Bob the sum of x_i over all i for which the protocol for $\textsc{Gt}(x_i, y_i)$ accepted, and Bob sends Alice the sum of y_i over the same i's. They output Alice's sum minus Bob's sum. By a union bound, with probability at least 3/4 each of the $\textsc{Gt}$ tests returns the correct answer, in which case the final output is correct by definition. The communication cost is $O\big(n \log(n \log(n/\epsilon)) + \log(n/\epsilon)\big) \leq O(\max(n \log n, \log(1/\epsilon)))$.

To prove (iii), we use a reduction from the Gap-Hamming partial function $\textsc{Gh}_{n,\epsilon}$, in which the goal is to determine whether the relative Hamming distance between Alice's and Bob's length-n bit strings is $> 1/2 + \epsilon$ or $< 1/2 - \epsilon$. It is known that $\mathsf{R}(\textsc{Gh}_{n,\epsilon}) \geq \Omega(\min(1/\epsilon^2, n))$

[8, 28, 26]. Here is the reduction: Alice transforms $a \in \{0,1\}^n$ into a distribution x over $[2n]$ by letting $x_{2i-a_i} = 1/n$ for each $i \in [n]$ (and letting all other entries of x be 0). Bob transforms b into y in the same way. Then $\Delta(x,y)$ equals the relative Hamming distance between a and b, so a protocol for $\textsc{Stat-Dist}_{2n,\epsilon}$ can distinguish the two cases (by whether the output is above or below $1/2$).

To prove (iv), consider any correct randomized protocol for $\textsc{Stat-Dist}_{n,\epsilon}$, and fix any set of $1/(3\epsilon)$ many pairs of distributions having statistical distances 0, 3ϵ, 6ϵ, 9ϵ, $\ldots$. There must exist some outcome of the randomness of the protocol such that the induced deterministic protocol is correct on at least three fourths of those inputs. But then the same transcript cannot occur for any two of these $1/(4\epsilon)$ inputs since the statistical distances are more than 2ϵ apart. Thus at least $1/(4\epsilon)$ transcripts are necessary, so the communication cost must be at least $\log(1/\epsilon) - 2$. ◀

Proof of Theorem 2. For the upper bound, assuming each probability x_i and y_i is a multiple of $1/m$ for some integer $m := O(n/\epsilon)$, we employ the trivial protocol where Alice sends a specification of her distribution to Bob (who then responds with the $(\log(n/\epsilon)+O(1))$-bit answer). We just need to count the number of such distributions: $\binom{m+n-1}{n-1} \leq \left(\frac{e\cdot(m+n-1)}{n-1}\right)^{n-1} \leq \big(O(1/\epsilon)\big)^n$. Hence only $O(n\log(1/\epsilon))$ bits are needed to specify a distribution.

The proof of the lower bound is basically a Gilbert–Varshamov argument for codes in the Manhattan metric. Specifically, we claim that there is a set of $2^{\Omega(n\log(1/\epsilon))}$ many distributions over $[n]$ that pairwise have statistical distance $> 2\epsilon$. Then for any distinct distributions x and x' from this set, the inputs (x,x) and (x',x') cannot share the same transcript in any correct protocol for $\textsc{Stat-Dist}_{n,\epsilon}$, because if they did then (x,x') would also share that transcript, but (x,x) requires output $\leq \epsilon$ while (x,x') requires output $> \epsilon$. Hence any correct protocol has at least $2^{\Omega(n\log(1/\epsilon))}$ transcripts and so has communication cost $\Omega(n\log(1/\epsilon))$.

To see the claim, first note that the number of distributions whose probabilities are multiples of $1/m$ is $\big(\Omega(1/\epsilon)\big)^n$, while the number of such distributions within statistical distance $\leq 2\epsilon$ of any fixed such distribution can be simply upper bounded by $2^n \cdot \binom{4\epsilon m+n}{n} \leq \big(O(1)\big)^n$. Hence if we keep greedily adding to a set any distribution that has statistical distance $> 2\epsilon$ from every distribution we picked so far, then the number of iterations this process can continue is at least $\big(\Omega(1/\epsilon)\big)^n / \big(O(1)\big)^n \geq \big(\Omega(1/\epsilon)\big)^n$, which is $2^{\Omega(n\log(1/\epsilon))}$ provided ϵ is at most a sufficiently small constant. ◀

3 Composing with Majority

In this section, we follow a convention that has become common in recent literature: For a two-party (possibly partial) function $F\colon \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$ and a complexity class name $\mathcal{C}$, we let $\mathcal{C}(F)$ denote the minimum worst-case cost of any protocol for F in the model corresponding to $\mathcal{C}$, and we also use $\mathcal{C}$ to denote the class of (families of) F's such that $\mathcal{C}(F) \leq \mathrm{polylog}(n)$. In particular, $\mathsf{BPP}(F)$ is an alias for the plain randomized communication complexity $\mathsf{R}(F)$ in the case of $\{0,1\}$-valued F, but we use the complexity class notation now for aesthetic consistency. We also need the following "2-sided WAPP" model.[1]

[1] There are two ways to define this model, which are equivalent up to a factor of 2 in ϵ. Our way was also used in [16] and is the same as the relaxed partition bound [19]. In [15], a "starred" notation was used for this, while the notation 2WAPP was reserved for the other definition, which is the same as the two-sided smooth rectangle bound [18].

▶ **Definition 4.** $\mathsf{2WAPP}_\epsilon(F) := \min\big(CC(\Pi) + \log(1/\alpha)\big)$ over all $\alpha > 0$ and protocols Π with output values $\{0, 1, \perp\}$ such that for all (x, y), $\mathbb{P}[\Pi(x, y) \neq \perp] \leq \alpha$ and $\mathbb{P}[\Pi(x, y) = F(x, y)] \geq (1 - \epsilon)\alpha$.

For all F and constants $0 < \epsilon < 1/2$, we have $O(\mathsf{BPP}(F)) \geq \mathsf{2WAPP}_\epsilon(F) \geq \Omega(\mathsf{PP}(F))$, and thus $\mathsf{BPP} \subseteq \mathsf{2WAPP}_\epsilon \subseteq \mathsf{PP}$. It is not necessary to recall the communication complexity definition of PP, but we remark that $\mathsf{2WAPP}_\epsilon$ feels intuitively much closer to BPP, since there are many interesting classes sandwiched between $\mathsf{2WAPP}_\epsilon$ and PP [16]. The following is due to [16].

▶ **Theorem 5** (AND-composition). *For all F, k, and constants $0 < \epsilon < 1/2$, we have*

$$\mathsf{2WAPP}_\epsilon(F) \;\leq\; O\big(\mathsf{BPP}(\textsc{And}_k \circ F^k)/k + \log \mathsf{BPP}(\textsc{And}_k \circ F^k)\big).$$

We prove that by using $\textsc{Maj}_k$ instead of $\textsc{And}_k$, the logarithmic term can be avoided.

▶ **Theorem 6** (MAJ-composition). *For all F, k, and constants $0 < \epsilon < 1/2$, we have*

$$\mathsf{2WAPP}_\epsilon(F) \;\leq\; O\big(\mathsf{BPP}(\textsc{Maj}_k \circ F^k)/k + 1\big).$$

Proof of Theorem 3. As noted in the proof of Theorem 1, $\textsc{Gt}_n$ has a protocol with error probability $1/(4n)$ and communication cost $O(\log n)$. By running this on each of n coordinates, with probability at least $3/4$ all the outputs will be correct, so a protocol witnessing $\mathsf{BPP}(\textsc{Maj}_n \circ \textsc{Gt}_n^n) \leq O(n \log n)$ can be obtained by applying $\textsc{Maj}_n$ to all these outputs. The matching lower bound follows by combining Theorem 6 with the result that $\mathsf{PP}(\textsc{Gt}_n) \geq \Omega(\log n)$ [6]. ◀

Theorem 6 follows by stringing together the following three lemmas. For any input distribution D (over the domain of F), we define the distributions $D^b := (D \mid F^{-1}(b))$ for $b \in \{0, 1\}$. We say a protocol Π is δ-correct for F iff $\mathbb{P}[\Pi(x, y) = F(x, y)] \geq 1 - \delta$ for all (x, y).

▶ **Lemma 7.** *Fix any F, k, $0 < \delta < 1/2$, and input distribution D. For every δ-correct protocol Π for $\textsc{Maj}_k \circ F^k$ there exists a δ-correct protocol Π' for F such that $IC_{D^b}(\Pi') \leq O(CC(\Pi)/k)$ holds for both $b \in \{0, 1\}$.*

▶ **Lemma 8.** *Fix any F, input distribution D, and protocol Π (not necessarily correct). Then*

$$IC_D(\Pi) - 4 \;\leq\; \textstyle\sum_b \mathbb{P}_D[F^{-1}(b)] \cdot IC_{D^b}(\Pi) \;\leq\; IC_D(\Pi).$$

▶ **Lemma 9.** *Fix any F, constants $0 < \delta < \epsilon < 1/2$, and value c. If for every input distribution D there exists a δ-correct protocol Π for F such that $IC_D(\Pi) \leq c$, then $\mathsf{2WAPP}_\epsilon(F) \leq O(c + 1)$.*

Only the first inequality in Lemma 8 is needed for Theorem 6. Lemma 9 is due to [19]. Before we commence with the proofs of Lemma 7 and Lemma 8, we recall the following standard fact; see [4, §2.1] for a proof. (We apologize for overloading the D notation between this fact and the above lemmas, but there should be no confusion.)

▶ **Fact 10.** *Let A, B, C, D be four random variables. Then*
(i) $\mathbb{I}(A\,;\,B \mid C) \leq \mathbb{I}(A\,;\,B \mid C, D)$ *if* $\mathbb{I}(B\,;\,D \mid C) = 0$;
(ii) $\mathbb{I}(A\,;\,B \mid C) \geq \mathbb{I}(A\,;\,B \mid C, D)$ *if* $\mathbb{I}(B\,;\,D \mid A, C) = 0$.

Proof of Lemma 7. Assume k is odd for convenience. Consider a probability space with the following random variables: $Z \in \{0,1\}^k$ is a uniformly random string of Hamming weight $\lceil k/2 \rceil$, $S \coloneqq \{i : Z_i = 1\}$, (X, Y) is such that $(X_i, Y_i) \sim D^{Z_i}$ for each $i \in [k]$ independently, and T and R are the communication transcript and public coins (respectively) of Π on input (X, Y). We use the subscript notation $X_{<i}$ and $X_{>i}$ for restrictions to coordinates in $\{1, \ldots, i-1\}$ and $\{i+1, \ldots, k\}$, and we use the superscript notation X^S and X^{-S} for restrictions to coordinates in S and $[k] \smallsetminus S$, and we may combine these so e.g., $X^{-S}_{>i}$ is the restriction to coordinates in $\{i+1, \ldots, k\} \smallsetminus S$. We use corresponding notation for restrictions of Y. We have

$$\begin{aligned}
&2 \cdot CC(\Pi)\\
&\geq \mathbb{I}\big(T \,;\, X^S \,\big|\, X^{-S}, Y, R, S\big) + \mathbb{I}\big(T \,;\, Y^S \,\big|\, Y^{-S}, X, R, S\big)\\
&= \mathbb{E}_{s \sim S}\Big[\textstyle\sum_{i \in s} \mathbb{I}\big(T \,;\, X_i \,\big|\, X^s_{<i}, X^{-s}, Y, R, s\big) + \textstyle\sum_{i \in s} \mathbb{I}\big(T \,;\, Y_i \,\big|\, Y^s_{>i}, Y^{-s}, X, R, s\big)\Big]\\
&\geq \mathbb{E}_{s \sim S}\Big[\textstyle\sum_{i \in s} \mathbb{I}\big(T \,;\, X_i \,\big|\, Y_i, X_{<i}, Y_{>i}, R, s\big) + \textstyle\sum_{i \in s} \mathbb{I}\big(T \,;\, Y_i \,\big|\, X_i, Y_{>i}, X_{<i}, R, s\big)\Big]\\
&= \lceil k/2 \rceil \cdot \mathop{\mathbb{E}}_{\substack{s \sim S,\quad i \sim s,\quad r \sim R\\ x_{<i} \sim X_{<i},\quad y_{>i} \sim Y_{>i}}} \Big[\mathbb{I}\big(T \,;\, X_i \,\big|\, Y_i, x_{<i}, y_{>i}, r, s\big) + \mathbb{I}\big(T \,;\, Y_i \,\big|\, X_i, x_{<i}, y_{>i}, r, s\big)\Big]
\end{aligned}$$

where the second line is by the chain rule, the third line is by Fact 10.(i) since $X^{-s}_{>i}, Y_{<i}$ is independent of X_i given $Y_i, X_{<i}, Y_{>i}, R, s$ and since $Y^{-s}_{<i}, X_{>i}$ is independent of Y_i given $X_i, Y_{>i}, X_{<i}, R, s$, and where $i \sim s$ on the fourth line means i is sampled uniformly at random from the set s.

Note that sampling $s \sim S$ and $i \sim s$ is equivalent to sampling $i \sim [k]$ and a uniformly random balanced bit string $z_{-i} \sim Z_{-i}$ indexed by $[k] \smallsetminus \{i\}$ (and setting $z_i = 1$). We let $q \sim Q$ denote a sample of all the data $(i,\ z_{-i},\ r,\ x_{<i},\ y_{>i})$. In summary, we have

$$\mathbb{E}_{q \sim Q}\big[\mathbb{I}(T \,;\, X_i \,|\, Y_i, q) + \mathbb{I}(T \,;\, Y_i \,|\, X_i, q)\big] \;\leq\; (2/\lceil k/2 \rceil) \cdot CC(\Pi)$$

so by Markov's inequality, with probability $> 1/2$ over $q \sim Q$ we have

$$\mathbb{I}(T \,;\, X_i \,|\, Y_i, q) + \mathbb{I}(T \,;\, Y_i \,|\, X_i, q) \;\leq\; (4/\lceil k/2 \rceil) \cdot CC(\Pi) \tag{1}$$

where $(X_i, Y_i) \sim D^1$. By symmetric reasoning (interchanging the roles of 0 and 1), with probability $> 1/2$ over $q \sim Q$, (1) also holds if we instead have $(X_i, Y_i) \sim D^0$. Thus there exists a q (which we fix henceforth) such that (1) holds both when $(X_i, Y_i) \sim D^1$ and when $(X_i, Y_i) \sim D^0$ (and in either case, $(X_j, Y_j) \sim D^{z_j}$ for $j \neq i$).

Now consider the protocol Π' where the input is interpreted as (x_i, y_i), Alice privately samples $x_{>i} \sim (X_{>i} \,|\, y_{>i}, z_{>i})$, Bob privately samples $y_{<i} \sim (Y_{<i} \,|\, x_{<i}, z_{<i})$, and they run Π on the combined input (x, y) with public coins r. The conclusion of the previous paragraph is exactly that $IC_{D^b}(\Pi') \leq (4/\lceil k/2 \rceil) \cdot CC(\Pi) \leq O(CC(\Pi)/k)$ holds for both $b \in \{0,1\}$. Furthermore, Π' is δ-correct since Π is δ-correct and $F(x_i, y_i) = (\textsc{Maj}_k \circ F^k)(x, y)$ with probability 1, for every (x_i, y_i) in F's domain. ◀

Proof of Lemma 8. Consider a probability space with the following random variables: $(X, Y) \sim D$, $F \coloneqq F(X, Y)$, and T and R are the communication transcript and public coins (respectively) of Π on input (X, Y). Then we have

$$\begin{aligned}
IC_D(\Pi) \;&=\; \mathbb{I}(T \,;\, X \,|\, Y, R) \quad\;\; + \mathbb{I}(T \,;\, Y \,|\, X, R)\\
\textstyle\sum_b \mathbb{P}_D[F^{-1}(b)] \cdot IC_{D^b}(\Pi) \;&=\; \mathbb{I}(T \,;\, X \,|\, Y, R, F) + \mathbb{I}(T \,;\, Y \,|\, X, R, F)
\end{aligned}$$

and so the second inequality of Lemma 8 holds by Fact 10.(ii) since conditioned on X, Y, R, there is no remaining entropy in F and hence it is independent of T.

For the first inequality, we use the following result proven in [15].

▶ **Lemma 11.** *There exist numbers $c_{x,y}, c'_{x,y} \geq 0$ for each input (x, y) in the domain of F, such that*

- $IC_D(\Pi) = \mathbb{E}[c_{X,Y}]$,
- $IC_{D^b}(\Pi) = \mathbb{E}[c'_{X,Y} \mid F = b]$ *for both* $b \in \{0, 1\}$,
- *for each (x, y) in the domain of F, letting $b \coloneqq F(x, y)$ we have*

$$c_{x,y} \;\leq\; c'_{x,y} + \log\bigl(1/\mathbb{P}[F = b \mid y]\bigr) + \log\bigl(1/\mathbb{P}[F = b \mid x]\bigr).$$

Hence, letting $p_{x,y} \coloneqq \mathbb{P}[(X, Y) = (x, y)]$, we have

$$\begin{aligned} IC_D(\Pi) &= \textstyle\sum_{(x,y)} p_{x,y} \cdot c_{x,y} \\ &\leq \textstyle\sum_b \sum_{(x,y)\in F^{-1}(b)} p_{x,y} \cdot \bigl(c'_{x,y} + \log\bigl(1/\mathbb{P}[F = b \mid y]\bigr) + \log\bigl(1/\mathbb{P}[F = b \mid x]\bigr)\bigr) \\ &= \textstyle\sum_b \mathbb{P}[F = b] \cdot IC_{D^b}(\Pi) + \\ &\quad \textstyle\sum_b \sum_{(x,y)\in F^{-1}(b)} p_{x,y} \cdot \bigl(\log\bigl(1/\mathbb{P}[F = b \mid y]\bigr) + \log\bigl(1/\mathbb{P}[F = b \mid x]\bigr)\bigr). \end{aligned}$$

We claim that for both $b \in \{0, 1\}$ we have $\sum_{(x,y)\in F^{-1}(b)} p_{x,y} \cdot \log\bigl(1/\mathbb{P}[F = b \mid y]\bigr) \leq 1$ and $\sum_{(x,y)\in F^{-1}(b)} p_{x,y} \cdot \log\bigl(1/\mathbb{P}[F = b \mid x]\bigr) \leq 1$; it then follows that $IC_D(\Pi) \leq \sum_b \mathbb{P}[F = b] \cdot IC_{D^b}(\Pi) + 4$.

We just argue the claim for $b = 1$ and conditioning on y; the other three cases are completely analogous. For $a \in \{0, 1\}$ define $p^a_y \coloneqq \mathbb{P}[F = a \text{ and } Y = y] = \sum_{x\,:\,(x,y)\in F^{-1}(a)} p_{x,y}$. Then we have

$$\begin{aligned} \textstyle\sum_{(x,y)\in F^{-1}(1)} p_{x,y} \cdot \log\bigl(1/\mathbb{P}[F = 1 \mid y]\bigr) &= \textstyle\sum_y p^1_y \cdot \log\bigl((p^0_y + p^1_y)/p^1_y\bigr) \\ &\leq \textstyle\sum_y p^1_y \cdot \bigl((p^0_y + p^1_y)/p^1_y\bigr) \\ &= 1. \end{aligned}$$

This finishes the proof. ◀

Acknowledgements. I thank Mika Göös for discussions and anonymous reviewers for comments.

References

1 Anurag Anshu, Aleksandrs Belovs, Shalev Ben-David, Mika Göös, Rahul Jain, Robin Kothari, Troy Lee, and Miklos Santha. Separations in communication complexity using cheat sheets and information complexity. In *Proceedings of the 57th Symposium on Foundations of Computer Science (FOCS)*, pages 555–564. IEEE, 2016. doi:10.1109/FOCS.2016.66.

2 Tugkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren Smith, and Patrick White. Testing closeness of discrete distributions. *Journal of the ACM*, 60(1):4, 2013. doi:10.1145/2432622.2432626.

3 Andrej Bogdanov, Elchanan Mossel, and Salil Vadhan. The complexity of distinguishing Markov random fields. In *Proceedings of the 12th International Workshop on Randomization and Computation (RANDOM)*, pages 331–342. Springer, 2008. doi:10.1007/978-3-540-85363-3_27.

4 Mark Braverman. Interactive information complexity. *SIAM Journal on Computing*, 44(6):1698–1739, 2015. doi:10.1137/130938517.

5 Mark Braverman and Omri Weinstein. An interactive information odometer and applications. In *Proceedings of the 47th Symposium on Theory of Computing (STOC)*, pages 341–350. ACM, 2015. doi:10.1145/2746539.2746548.

6 Mark Braverman and Omri Weinstein. A discrepancy lower bound for information complexity. *Algorithmica*, 76(3):846–864, 2016. doi:10.1007/s00453-015-0093-8.

7 Clément Canonne. A survey on distribution testing: Your data is big. But is it blue? Technical Report TR15-063, Electronic Colloquium on Computational Complexity (ECCC), 2015. URL: http://eccc.hpi-web.de/report/2015/063.

8 Amit Chakrabarti and Oded Regev. An optimal lower bound on the communication complexity of Gap-Hamming-Distance. *SIAM Journal on Computing*, 41(5):1299–1317, 2012. doi:10.1137/120861072.

9 Siu On Chan, Ilias Diakonikolas, Paul Valiant, and Gregory Valiant. Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the 25th Symposium on Discrete Algorithms (SODA)*, pages 1193–1203. ACM-SIAM, 2014. doi:10.1137/1.9781611973402.88.

10 Joan Feigenbaum, Sampath Kannan, Martin Strauss, and Mahesh Viswanathan. An approximate L^1-difference algorithm for massive data streams. *SIAM Journal on Computing*, 32(1):131–151, 2002. doi:10.1137/S0097539799361701.

11 Jessica Fong and Martin Strauss. An approximate L^p-difference algorithm for massive data streams. *Discrete Mathematics & Theoretical Computer Science*, 4(2):301–322, 2001.

12 Oded Goldreich, Amit Sahai, and Salil Vadhan. Can statistical zero knowledge be made non-interactive? or On the relationship of SZK and NISZK. In *Proceedings of the 19th International Cryptology Conference (CRYPTO)*, pages 467–484. Springer, 1999. doi:10.1007/3-540-48405-1_30.

13 Oded Goldreich and Salil Vadhan. Comparing entropies in statistical zero-knowledge with applications to the structure of SZK. In *Proceedings of the 14th Conference on Computational Complexity (CCC)*, pages 54–73. IEEE, 1999. doi:10.1109/CCC.1999.766262.

14 Oded Goldreich and Salil Vadhan. On the complexity of computational problems regarding distributions. *Studies in Complexity and Cryptography*, pages 390–405, 2011. doi:10.1007/978-3-642-22670-0_27.

15 Mika Göös, T. S. Jayram, Toniann Pitassi, and Thomas Watson. Randomized communication vs. partition number. In *Proceedings of the 44th International Colloquium on Automata, Languages, and Programming (ICALP)*. Schloss Dagstuhl, 2017. To appear.

16 Mika Göös, Shachar Lovett, Raghu Meka, Thomas Watson, and David Zuckerman. Rectangles are nonnegative juntas. *SIAM Journal on Computing*, 45(5):1835–1869, 2016. doi:10.1137/15M103145X.

17 Thomas Holenstein. Parallel repetition: Simplification and the no-signaling case. *Theory of Computing*, 5(1):141–172, 2009. doi:10.4086/toc.2009.v005a008.

18 Rahul Jain and Hartmut Klauck. The partition bound for classical communication complexity and query complexity. In *Proceedings of the 25th Conference on Computational Complexity (CCC)*, pages 247–258. IEEE, 2010. doi:10.1109/CCC.2010.31.

19 Iordanis Kerenidis, Sophie Laplante, Virginie Lerays, Jérémie Roland, and David Xiao. Lower bounds on information complexity via zero-communication protocols and applications. *SIAM Journal on Computing*, 44(5):1550–1572, 2015. doi:10.1137/130928273.

20 Jon Kleinberg and Éva Tardos. Approximation algorithms for classification problems with pairwise relationships: metric labeling and Markov random fields. *Journal of the ACM*, 49(5):616–639, 2002. doi:10.1145/585265.585268.

21 Eyal Kushilevitz and Noam Nisan. *Communication Complexity.* Cambridge University Press, 1997.

22 Anup Rao. Parallel repetition in projection games and a concentration bound. *SIAM Journal on Computing*, 40(6):1871–1891, 2011. doi:10.1137/080734042.

23 Ran Raz. A counterexample to strong parallel repetition. *SIAM Journal on Computing*, 40(3):771–777, 2011. doi:10.1137/090747270.

24 Ronitt Rubinfeld. Taming big probability distributions. *ACM Crossroads*, 19(1):24–28, 2012. doi:10.1145/2331042.2331052.

25 Amit Sahai and Salil Vadhan. A complete problem for statistical zero knowledge. *Journal of the ACM*, 50(2):196–249, 2003. doi:10.1145/636865.636868.

26 Alexander Sherstov. The communication complexity of Gap Hamming Distance. *Theory of Computing*, 8(1):197–208, 2012. doi:10.4086/toc.2012.v008a008.

27 Paul Valiant. Testing symmetric properties of distributions. *SIAM Journal on Computing*, 40(6):1927–1968, 2011. doi:10.1137/080734066.

28 Thomas Vidick. A concentration inequality for the overlap of a vector on a large set, with application to the communication complexity of the Gap-Hamming-Distance problem. *Chicago Journal of Theoretical Computer Science*, 2012(1):1–12, 2012. doi:10.4086/cjtcs.2012.001.

29 Thomas Watson. The complexity of deciding statistical properties of samplable distributions. *Theory of Computing*, 11:1–34, 2015. doi:10.4086/toc.2015.v011a001.

30 Thomas Watson. The complexity of estimating min-entropy. *Computational Complexity*, 25(1):153–175, 2016. doi:10.1007/s00037-014-0091-2.