# A discrepancy lower bound for information complexity

Mark Braverman* and Omri Weinstein**

**Abstract.** This paper provides the first general technique for proving information lower bounds on two-party unbounded-rounds communication problems. We show that the discrepancy lower bound, which applies to randomized communication complexity, also applies to information complexity. More precisely, if the discrepancy of a two-party function $f$ with respect to a distribution $\mu$ is $Disc_\mu f$, then any two party randomized protocol computing $f$ must reveal at least $\Omega(\log(1/Disc_\mu f))$ bits of information to the participants. As a corollary, we obtain that any two-party protocol for computing a random function on $\{0,1\}^n \times \{0,1\}^n$ must reveal $\Omega(n)$ bits of information to the participants.
In addition, we prove that the discrepancy of the Greater-Than function is $\Omega(1/\sqrt{n})$, which provides an alternative proof to the recent proof of Viola [Vio11] of the $\Omega(\log n)$ lower bound on the communication complexity of this well-studied function and, combined with our main result, proves the tight $\Omega(\log n)$ lower bound on its information complexity.
The proof of our main result develops a new simulation procedure that may be of an independent interest. In a very recent breakthrough work of Kerenidis et al. [KLL$^+$12], this simulation procedure was a building block towards a proof that almost all known lower bound techniques for communication complexity (and not just discrepancy) apply to information complexity.

## 1 Introduction

The main objective of this paper is to expand the available techniques for proving information complexity lower bounds for communication problems. Let $f : \mathcal{X} \times \mathcal{Y} \to \{0,1\}$ be a function, and $\mu$ be a distribution on $\mathcal{X} \times \mathcal{Y}$. Informally, the information complexity of $f$ is the least amount of *information* that Alice and Bob need to exchange on average to compute $f(x,y)$ using a randomized communication protocol if initially $x$ is given to Alice, $y$ is given to Bob, and $(x,y) \sim \mu$. Note that information here is measured in the Shannon sense, and the amount of information may be much smaller than the number of bits exchanged. Thus the randomized communication complexity of $f$ is an upper bound on its information complexity, but may not be a lower bound.

Within the context of communication complexity, information complexity has first been introduced in the context of direct sum theorems for randomized

* Princeton University and the University of Toronto, mbraverm@cs.princeton.edu. Partially supported by an NSERC Discovery Grant, an Alfred P. Sloan Fellowship, and an NSF CAREER award.
** Princeton University, oweinste@cs.princeton.edu.

communication complexity [CSWY01,BYJKS04,BBCR10]. These techniques are also being used in the related direction of direct product theorems [KSDW04,LSS08,Jai10,Kla10]. A direct sum theorem in a computational model states that the amount of resources needed to perform $k$ independent tasks is roughly $k$ times the amount of resources $c$ needed for computing a single task. A direct product theorem, which is a stronger statement, asserts that any attempt to solve $k$ independent tasks using $o(kc)$ resources would result in an exponentially small success probability $2^{-\Omega(k)}$.

The direct sum line of work [HJMR07,JSR08,BBCR10,BR11] has eventually led to a tight connection (equality) between amortized communication complexity and information complexity. Thus proving lower bounds on the communication complexity of $k$ copies of $f$ for a growing $k$ is equivalent to proving lower bounds on the information complexity of $f$. In particular if $f$ satisfies $IC(f) = \Omega(CC(f))$, i.e. that its information cost is asymptotically equal to its communication complexity, then a strong direct sum theorem holds for $f$. In addition to the intrinsic interest of understanding the amount of information exchange that needs to be involved in computing $f$, direct sum theorems motivate the development of techniques for proving lower bounds on the information complexity of functions.

Another important motivation for seeking lower bounds on the information complexity of functions stems from understanding the limits of security in two-party computation. In a celebrated result Ben-Or et al. [BOGW88] (see also [AL11]) showed how a multi-party computation (with three or more parties) may be carried out in a way that reveals no information to the participants except for the computation's output. The protocol relies heavily on the use of random bits that are shared between some, but not all, parties. Such a resource can clearly not exist in the two-party setting. While it can be shown that perfect information security is unattainable by two-party protocols [CK89,BYCKO93], quantitatively it is not clear just how much information the parties must "leak" to each other to compute $f$. The quantitative answer depends on the model in which the leakage occurs, and whether quantum computation is allowed [Kla04]. Information complexity answers this question in the strongest possible sense for classical protocols: the parties are allowed to use private randomness to help them "hide" their information, and the information revealed is measured on average. Thus an information complexity lower bound of $I$ on a problem implies that the *average* (as opposed to worst-case) amount of information revealed to the parties is at least $I$.

As mentioned above, the information complexity is always upper bounded by the communication complexity of $f$. The converse is not known to be true. Moreover, lower bound techniques for communication complexity do not readily translate into lower bound techniques for information complexity. The key difference is that a low-information protocol is not limited in the amount of communication it uses, and thus rectangle-based communication bounds do not readily convert into information lower bounds. No general technique has been known to yield sharp information complexity lower bounds. A linear lower bound on the communication complexity of the disjointness function has been shown in

[Raz92]. An information-theoretic proof of this lower bound [BYJKS04] can be adapted to prove a linear *information* lower bound on disjointness [Bra11]. One general technique for obtaining (weak) information complexity lower bounds was introduced in [Bra11], where it has been shown that any function that has $I$ bits of information complexity, has communication complexity bounded by $2^{O(I)}$. This immediately implies that the information complexity of a function $f$ is at least the log of its communication complexity ($IC(f) \geq \Omega(\log(CC(f)))$). In fact, this result easily follows from the stronger result we prove below (Theorem 2).

### 1.1 Our results

In this paper we prove that the discrepancy method – a general communication complexity lower bound technique – generalizes to information complexity. The discrepancy of $f$ with respect to a distribution $\mu$ on inputs, denoted $Disc_\mu(f)$, measures how "unbalanced" $f$ can get on any rectangle, where the balancedness is measured with respect to $\mu$:

$$Disc_\mu(f) \stackrel{def}{=} \max_{\text{rectangles } R} \left| \Pr_\mu[f(x,y) = 0 \wedge (x,y) \in R] - \Pr_\mu[f(x,y) = 1 \wedge (x,y) \in R] \right|.$$

A well-known lower bound (see e.g [KN97]) asserts that the distributional communication complexity of $f$, denoted $D^\mu_{1/2-\epsilon}(f)$, when required to predict $f$ with advantage $\varepsilon$ over a random guess (with respect to $\mu$), is bounded from below by $\Omega(\log 1/Disc_\mu(f))$:

$$D^\mu_{1/2-\epsilon}(f) \geq \log(2\epsilon/Disc_\mu(f)).$$

Note that the lower bound holds even if we are merely trying to get an advantage of $\varepsilon = \sqrt{Disc_\mu(f)}$ over random guessing in computing $f$. We prove that the information complexity of computing $f$ with probability $9/10$ with respect to $\mu$ is also bounded from below by $\Omega(\log(1/Disc_\mu(f)))$.

**Theorem 1.** *Let $f : \mathcal{X} \times \mathcal{Y} \to \{0,1\}$ be a Boolean function and let $\mu$ be any probability distribution on $\mathcal{X} \times \mathcal{Y}$. Then*

$$\mathsf{IC}_\mu(f, 1/10) \geq \Omega(\log(1/Disc_\mu(f))).$$

*Remark 1.* The choice of $9/10$ is somewhat arbitrary. For randomized worst-case protocols, we may replace the success probability with $1/2 + \delta$ for a constant $\delta$, since repeating the protocol constantly many times ($1/\delta^2$) would yield the aforementioned success rate, while the information cost of the repeated protocol differs only by a constant factor from the original one. In particular, using prior-free information cost [Bra11] this implies $\mathsf{IC}(f, 1/2 - \delta) \geq \Omega(\delta^2 \log(1/Disc_\mu(f)))$.

In particular, Theorem 1 implies a linear lower bound on the information complexity of the inner product function $IP(x,y) = \sum_{i=1}^n x_i y_i \mod 2$, and on a random boolean function $f_r : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$, expanding the (limited) list of functions for which nontrivial information-complexity lower bounds are known:

**Corollary 1.** *The information complexity* $\mathsf{IC}_{uniform}(IP, 1/10)$ *of* $IP(x, y)$ *is* $\Omega(n)$*. The information complexity* $\mathsf{IC}_{uniform}(f_r, 1/10)$ *of a random function* $f_r$ *is* $\Omega(n)$*, except with probability* $2^{-\Omega(n)}$*.*

We study the communication and information complexity of the Greater-Than function ($GT_n$) on numbers of length $n$. This is a very well-studied problem [Smi88,MNSW95,KN97]. Only very recently the tight lower bound of $\Omega(\log n)$ in the public-coins probabilistic model was given by Viola [Vio11]. We show that the discrepancy of the $GT_n$ function is $\Omega(1/\sqrt{n})$:

**Lemma 1.** *There exist a distribution* $\mu_n$ *on* $\mathcal{X} \times \mathcal{Y}$ *such that the discrepancy of* $GT_n$ *with respect to* $\mu_n$ *satisfies*

$$Disc_{\mu_n}(GT_n) < \frac{20}{\sqrt{n}}.$$

We defer the proof to the appendix. Lemma 1 provides an alternative (arguably simpler) proof of Viola's [Vio11] lower bound on the *communication complexity* of $GT_n$. By Theorem 1, Lemma 1 immediately implies a lower bound on the *information complexity* of $GT_n$:

**Corollary 2.** $\mathsf{IC}_{\mu_n}(GT_n, 1/10) = \Omega(\log n)$

This settles the information complexity of the GT function, since this problem can be solved by a randomized protocol with $O(\log n)$ communication (see [KN97]). This lower bound is particularly interesting since it demonstrates the first tight information-complexity lower bound for a natural function that is not linear.

The key technical idea in the proof of Theorem 1 is a new simulation procedure that allows us to convert any protocol that has information cost $I$ into a (two-round) protocol that has communication complexity $O(I)$ and succeeds with probability $> 1/2 + 2^{-O(I)}$, yielding a $2^{-O(I)}$ advantage over random guessing. Combined with the discrepancy lower bound for communication complexity, this proves Theorem 1.

### 1.2 Comparison and connections to prior results

The most relevant prior work is an article by Lee, Shraibman, and Špalek [LSS08]. Improving on an earlier work of Shaltiel [Sha03], Lee et al. show a direct product theorem for discrepancy, proving that the discrepancy of $f^{\otimes k}$ — the $k$-wise XOR of a function $f$ with itself — behaves as $Disc(f)^{\Omega(k)}$. This implies in particular that the communication complexity of $f^{\otimes k}$ scales at least as $\Omega(k \cdot \log Disc(f))$. Using the fact that the limit as $k \to \infty$ of the amortized communication complexity of $f$ is equal to the information cost of $f$ [BR10], the result of Lee et al. "almost" implies the bound of Theorem 1. Unfortunately, the amortized communication complexity in the sense of [BR10] is the amortized cost of $k$ copies of $f$, where *each* copy is allowed to err with some probability (say 1/10). Generally speaking, this task is much easier than computing the

XOR (which requires *all* copies to be evaluated correctly with high probability). Thus the lower bound that follows from Lee et al. applies to a more difficult problem, and does not imply the information complexity lower bound.

Another generic approach one may try to take is to use compression results such as [BBCR10] to lower bound the information cost from communication complexity lower bounds. The logic of such a proof would go as follows: "Suppose there was a information-complexity-$I$ protocol $\pi$ for $f$, then if one can compress it into a low-communication protocol one may get a contradiction to the communication complexity lower bound $f$". Unfortunately, all known compression results compress $\pi$ into a protocol $\pi'$ whose communication complexity depends on $I$ but also on $CC(\pi)$. Even for external information complexity (which is always greater than the internal information complexity, the bound obtained in [BBCR10] is of the form $I_{ext}(\pi) \cdot polylog(CC(\pi))$. Thus compression results of this type cannot rule out protocols that have low information complexity but a very high (e.g. exponential) communication complexity.

Our result can be viewed as a weak compression result for protocols, where a protocol for computing $f$ that conveys $I$ bits of information is converted into a protocol that uses $O(I)$ bits of *communication* and giving an advantage of $2^{-O(I)}$ in computing $f$. This strengthens the result in [Bra11] where a compression to $2^{O(I)}$ bits of communication has been shown. We still do not know whether compression to a protocol that uses $O(I)$ bits of communication and succeeds with high probability (as opposed to getting a small advantage over random) is possible.

In a very recent breakthrough work of Kerenidis, Laplante, Lerays, Roland, and Xiao [KLL$^+$12], our main protocol played an important role in the proof that almost all known lower bound techniques for communication complexity (and not just discrepancy) apply to information complexity. The results of [KLL$^+$12] shed more light on the information complexity of many communication problems, and the question of whether interactive communication can be compressed.

## 2 Preliminaries

In an effort to make this paper as self-contained as possible, we provide some background on information theory and communication complexity, which is essential to proving our results. For further details and a more thorough treatment of these subjects see [BR10] and references therein.

*Notation.* We reserve capital letters for random variables and distributions, calligraphic letters for sets, and small letters for elements of sets. Throughout this paper, we often use the notation $|b$ to denote conditioning on the event $B = b$. Thus $A|b$ is shorthand for $A|B = b$.

We use the standard notion of *statistical/total variation* distance between two distributions.

**Definition 1.** *Let $D$ and $F$ be two random variables taking values in a set $\mathcal{S}$. Their* statistical distance *is* $|D - F| \stackrel{def}{=} \max_{\mathcal{T} \subseteq \mathcal{S}}(|\Pr[D \in \mathcal{T}] - \Pr[F \in \mathcal{T}]|) = \frac{1}{2}\sum_{s \in \mathcal{S}} |\Pr[D = s] - \Pr[F = s]|$

### 2.1 Information Theory

**Definition 2 (Entropy).** *The* entropy *of a random variable $X$ is $H(X) \stackrel{def}{=} \sum_x \Pr[X = x] \log(1/\Pr[X = x])$. The* conditional entropy $H(X|Y)$ *is defined as* $\mathbf{E}_{y \in_R Y}[H(X|Y = y)]$.

**Definition 3 (Mutual Information).** *The* mutual information *between two random variables $A, B$, denoted $I(A; B)$ is defined to be the quantity $H(A) - H(A|B) = H(B) - H(B|A)$. The* conditional mutual information $I(A; B|C)$ *is* $H(A|C) - H(A|BC)$.

We also use the notion of *divergence* (also known as Kullback-Leibler distance or relative entropy), which is a different way to measure the distance between two distributions:

**Definition 4 (Divergence).** *The informational divergence between two distributions is*

$$\mathbf{D}(A||B) \stackrel{def}{=} \sum_x A(x) \log(A(x)/B(x)).$$

**Proposition 1.** *Let $A, B, C$ be random variables in the same probability space. For every $a$ in the support of $A$ and $c$ in the support of $C$, let $B_a$ denote $B|A = a$ and $B_{ac}$ denote $B|A = a, C = c$. Then $I(A; B|C) = \mathbf{E}_{a,c \in_R A,C}[\mathbf{D}(B_{ac}||B_c)]$.*

### 2.2 Communication Complexity

We use the standard definitions of the computational model defined in [Yao79]. For complete details see section A of the appendix.

Given a communication protocol $\pi$, $\pi(x, y)$ denotes the concatenation of the public randomness with all the messages that are sent during the execution of $\pi$. We call this the *transcript* of the protocol. When referring to the random variable denoting the transcript, rather than a specific transcript, we will use the notation $\Pi(x, y)$ — or simply $\Pi$ when $x$ and $y$ are clear from the context, thus $\pi(x, y) \in_R \Pi(x, y)$. When $x$ and $y$ are random variables themselves, we will denote the transcript by $\Pi(X, Y)$, or just $\Pi$.

**Definition 5 (Communication Complexity notation).** *For a function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{Z}_K$, a distribution $\mu$ supported on $\mathcal{X} \times \mathcal{Y}$, and a parameter $\epsilon > 0$, $D^{\mu}_{\epsilon}(f)$ denotes the communication complexity of the cheapest deterministic protocol computing $f$ on inputs sampled according to $\mu$ with error $\epsilon$.*

**Definition 6 (Combinatorial Rectangle).** *A Rectangle* in $\mathcal{X} \times \mathcal{Y}$ is a subset $R \subseteq \mathcal{X} \times \mathcal{Y}$ which satisfies

$$(x_1, y_1) \in R \text{ and } (x_2, y_2) \in R \Longrightarrow (x_1, y_2) \in R$$

### 2.3 Information + Communication: The information cost of a protocol

The following quantity, which is implicit in [BYJKS04] and was explicitly defined in [BBCR10], is the central notion of this paper.

**Definition 7.** *The (internal) information cost* of a protocol $\pi$ over inputs drawn from a distribution $\mu$ on $\mathcal{X} \times \mathcal{Y}$, is given by:

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

Intuitively, Definition 7 captures how much the two parties learn about each other's inputs from the execution transcript of the protocol $\pi$. The first term captures what the second player learns about $X$ from $\Pi$ – the mutual information between the input $X$ and the transcript $\Pi$ given the input $Y$. Similarly, the second term captures what the first player learns about $Y$ from $\Pi$.

Note that the information of a protocol $\pi$ depends on the prior distribution $\mu$, as the mutual information between the transcript $\Pi$ and the inputs depends on the prior distribution on the inputs. To give an extreme example, if $\mu$ is a singleton distribution, i.e. one with $\mu(\{(x, y)\}) = 1$ for some $(x, y) \in \mathcal{X} \times \mathcal{Y}$, then $\mathsf{IC}_\mu(\pi) = 0$ for all possible $\pi$, as no protocol can reveal anything to the players about the inputs that they do not already know *a-priori*. Similarly, $\mathsf{IC}_\mu(\pi) = 0$ if $\mathcal{X} = \mathcal{Y}$ and $\mu$ is supported on the diagonal $\{(x, x) : x \in \mathcal{X}\}$. As expected, one can show that the communication cost $\mathsf{CC}(\pi)$ of $\pi$ is an upper bound on its information cost over *any* distribution $\mu$:

**Lemma 2.** *[BR10] For any distribution $\mu$, $\mathsf{IC}_\mu(\pi) \leq \mathsf{CC}(\pi)$.*

On the other hand, as noted in the introduction, the converse need not hold. In fact, by [BR10], getting a strict inequality in Lemma 2 is equivalent to violating the direct sum conjecture for randomized communication complexity.

As one might expect, the information cost of a function $f$ with respect to $\mu$ and error $\rho$ is the least amount of information that needs to be revealed by a protocol computing $f$ with error $\leq \rho$:

$$\mathsf{IC}_\mu(f, \rho) := \inf_{\pi:\ \mathbf{P}_\mu[\pi(x,y) \neq f(x,y)] \leq \rho} \mathsf{IC}_\mu(\pi).$$

The (prior-free) information cost was defined in [Bra11] as the minimum amount of information that a worst-case error-$\rho$ randomized protocol can reveal against *all* possible prior distributions.

$$\mathsf{IC}\,(f, \rho) := \inf_{\pi \text{ is a protocol with } \mathbf{P}[\pi(x,y) \neq f(x,y)] \leq \rho \text{ for all } (x,y)} \max_{\mu} \mathsf{IC}_\mu(\pi).$$

This information cost matches the amortized *randomized* communication complexity of $f$ [Bra11]. It is clear that lower bounds on $\mathsf{IC}_\mu(f, \rho)$ *for any distribution* $\mu$ also apply to $\mathsf{IC}\,(f, \rho)$.

## 3 Proof of Theorem 1

To establish the correctness of Theorem 1, we prove the following theorem, which is the central result of this paper:

**Theorem 2.** *Suppose that* $\mathsf{IC}_\mu(f, 1/10) = I_\mu$. *Then there exist a protocol* $\pi'$ *such that*

- $\mathsf{CC}(\pi') = O(I_\mu)$.
- $\mathbf{P}_{(x,y)\sim\mu}[\pi'(x,y) = f(x,y)] \geq 1/2 + 2^{-O(I_\mu)}$

We first show how Theorem 1 follows from Theorem 2:

**Proof of Theorem 1.** Let $f, \mu$ be as in theorem 1, and let $\mathsf{IC}_\mu(f, 1/10) = I_\mu$. By theorem 2, there exists a protocol $\pi'$ computing $f$ with error probability $1/2 - 2^{-O(I_\mu)}$ using $O(I_\mu)$ bits of communication. Applying the discrepancy lower bound for communication complexity we obtain

$$O(I_\mu) \geq D^\mu_{1/2-2^{-O(I_\mu)}}(f) \geq \log(2 \cdot 2^{-O(I_\mu)}/Disc_\mu(f)) \tag{1}$$

which after rearranging gives $I_\mu \geq \Omega(\log(1/Disc_\mu(f)))$, as desired.

We now turn to prove Theorem 2. The main step is the following sampling lemma.

**Lemma 3.** *Let* $\mu$ *be any distribution over a universe* $\mathcal{U}$ *and let* $I \geq 0$ *be a parameter that is known to both* $A$ *and* $B$. *Further, let* $\nu_A$ *and* $\nu_B$ *be two distributions over* $\mathcal{U}$ *such that* $\mathbf{D}(\mu||\nu_A) \leq I$ *and* $\mathbf{D}(\mu||\nu_B) \leq I$. *The players are each given a pair of real functions* $(p_A, q_A)$, $(p_B, q_B)$, $p_A, q_A, p_B, q_B : \mathcal{U} \to [0,1]$ *such that for all* $x \in \mathcal{U}$, $\mu(x) = p_A(x) \cdot p_B(x)$, $\nu_A(x) = p_A(x) \cdot q_A(x)$, *and* $\nu_B(x) = p_B(x) \cdot q_B(x)$. *Then there is a (two round) sampling protocol* $\Pi_1 = \Pi_1(p_A, p_B, q_A, q_B, I)$ *which has the following properties:*

1. *at the end of the protocol, the players either declare that the protocol "fails", or output* $x_A \in \mathcal{U}$ *and* $x_B \in \mathcal{U}$, *respectively ("success").*
2. *let* $\mathcal{S}$ *be the event that the players output "success". Then* $\mathcal{S} \Rightarrow x_A = x_B$, *and* $0.9 \cdot 2^{-50(I+1)} \leq \Pr[\mathcal{S}] \leq 2^{-50(I+1)}$.
3. *if* $\mu_1$ *is the distribution of* $x_A$ *conditioned on* $\mathcal{S}$, *then* $|\mu - \mu_1| < 2/9$.

*Furthermore,* $\Pi_1$ *can be "compressed" to a protocol* $\Pi_2$ *such that* $\mathsf{CC}(\Pi_2) = 211I + 1$, *whereas* $|\Pi_1 - \Pi_2| \leq 2^{-59I}$ *(by an abuse of notation, here we identify* $\Pi_i$ *with the random variable representing its output).*

We will use the following technical fact about the information divergence of distributions.

*Claim (3).* [Claim 5.1 in [Bra11]] Suppose that $\mathbf{D}(\mu||\nu) \leq I$. Let $\varepsilon$ be any parameter. Then

$$\mu\left\{x:\ 2^{(I+1)/\varepsilon}\cdot\nu(x) < \mu(x)\right\} < \varepsilon.$$

For completeness, we repeat the proof in the appendix..

*Proof (***Proof of Lemma 3** ). Throughout the execution of $\Pi_1$, Alice and Bob interpret their shared random tape as a source of points $(x_i, \alpha_i, \beta_i)$ uniformly distributed in $\mathcal{U} \times [0, 2^{50(I+1)}] \times [0, 2^{50(I+1)}]$. Alice and Bob consider the first $T = |\mathcal{U}| \cdot 2^{100(I+1)} \cdot 60I$ such points. Their goal will be to discover the first index $\tau$ such that $\alpha_\tau \leq p_A(x_\tau)$ and $\beta_\tau \leq p_B(x_\tau)$ (where they wish to find it using a minimal amount of communication, even if they are most likely to fail). First, we note that the probability that an index $t$ satisfies $\alpha_t \leq p_A(x_t)$ and $\beta_t \leq p_B(x_t)$ is exactly $1/|\mathcal{U}|2^{50(I+1)}2^{50(I+1)} = 1/|\mathcal{U}|2^{100(I+1)}$. Hence the probability that $\tau > T$ (i.e. that $x_\tau$ is not among the $T$ points considered) is bounded by

$$\left(1 - 1/|\mathcal{U}|2^{100(I+1)}\right)^T < e^{-T/|\mathcal{U}|2^{100(I+1)}} = e^{-60I} < 2^{-60I} \tag{2}$$

Denote by $\mathcal{A}$ the following set of indices $\mathcal{A} := \{i \leq T:\ \alpha_i \leq p_A(x_i)$ and $\beta_i \leq 2^{50(I+1)} \cdot q_A(x_i)\}$, the set of potential candidates for $\tau$ from A's viewpoint. Similarly, denote $\mathcal{B} := \{i \leq T:\ \alpha_i \leq 2^{50(I+1)} \cdot q_B(x_i)$ and $\beta_i \leq p_B(x_i)\}$.

The protocol $\Pi_1$ is very simple. Alice takes her bet on the first element $a \in \mathcal{A}$ and sends it to Bob. Bob outputs $a$ only if (it just so happens that) $\beta_\tau \leq p_B(a)$. The details are given in Figure 2 in the appendix.

We turn to analyze $\Pi_1$. Denote the set of "Good" elements by

$$\mathcal{G} \stackrel{def}{=} \{x:\ 2^{50(I+1)} \cdot \nu_A(x) \geq \mu(x)\ \text{ and }\ 2^{50(I+1)} \cdot \nu_B(x) \geq \mu(x)\}\}.$$

Then by Claim 3, $\mu(\mathcal{G}) \geq 48/50 = 24/25$. The following claim asserts that if it succeeds, the output of $\Pi_1$ has the "correct" distribution on elements in $\mathcal{G}$.

**Proposition 2.** *Assume $\mathcal{A}$ is nonempty. Then for any $x_i \in \mathcal{U}$, the probability that $\Pi_1$ outputs $x_i$ is at most $\mu(x_i) \cdot 2^{-50(I+1)}$. If $x_i \in \mathcal{G}$, then this probability is* exactly $\mu(x_i) \cdot 2^{-50(I+1)}$.

*Proof.* Note that if $\mathcal{A}$ is nonempty, then for any $x_i \in \mathcal{U}$, the probability that $x_i$ is the first element in $\mathcal{A}$ (i.e, $a = x_i$) is $p_A(x_i)q_A(x_i) = \nu_A(x_i)$. By construction, the probability that $\beta_i \leq p_B(a)$ is $\min\{p_B(x_i)/(2^{50(I+1)}q_A(x_i)), 1\}$, and thus

$$\Pr[\Pi_1 \text{ outputs } x_i] \leq p_A(x_i)q_A(x_i) \cdot \frac{p_B(x_i)}{2^{50(I+1)}q_A(x_i)} = \mu(x_i) \cdot 2^{-50(I+1)}.$$

On the other hand, if $x_i \in \mathcal{G}$, then we know that $p_B(x_i)/q_A(x_i) = \mu(x_i)/\nu_A(x_i) \leq 2^{50(I+1)}$, and so the probability that $\beta_i \leq p_B(a)$ is *exactly* $p_B(x_i)/(2^{50(I+1)}q_A(x_i))$. Since $\Pr[\Pi_1 \text{ outputs } x_i] = \Pr[a = x_i]\Pr[\beta_i \leq p_B(a)]$ (assuming $\mathcal{A}$ is nonempty), we conclude that:

$$x_i \in \mathcal{G} \Longrightarrow \Pr[\Pi_1 \text{ outputs } x_i] = p_A(x_i)q_A(x_i) \cdot \frac{p_B(x_i)}{2^{50(I+1)}q_A(x_i)} = \mu(x_i) \cdot 2^{-50(I+1)}.$$

We are now ready to estimate the success probability of the protocol.

**Proposition 3.** *Let $\mathcal{S}$ denote the event that $\mathcal{A} \neq 0$ and $a \in \mathcal{B}$ (i.e, that the protocol succeeds). Then*

$$0.9 \cdot 2^{-50(I+1)} \leq \Pr[\mathcal{S}] \leq 2^{-50(I+1)}.$$

*Proof.* Using Proposition 2, we have that

$$\begin{aligned} \Pr[\mathcal{S}] \leq \mathbf{P}[a \in \mathcal{B} \mid \mathcal{A} \neq \emptyset] = \sum_{i \in \mathcal{U}} \Pr[a = x_i] \Pr[\beta_i \leq p_B(a)] \leq \\ \leq \sum_{i \in \mathcal{U}} \mu(x_i) \cdot 2^{-50(I+1)} = 2^{-50(I+1)} \end{aligned} \tag{3}$$

For the lower bound, we have

$$\begin{aligned} \Pr[\mathcal{S}] &\geq \Pr[\beta_i \leq p_B(a) \mid \mathcal{A} \neq \emptyset] \cdot \Pr[\mathcal{A} \neq \emptyset] \geq \\ &\geq (1 - 2^{-60I}) \left( \sum_{i \in \mathcal{U}} \Pr[a = x_i] \Pr[\beta_i \leq p_B(a)] \right) \geq \\ &\geq (1 - 2^{-60I}) \left( \sum_{i \in \mathcal{G}} \Pr[a = x_i] \Pr[\beta_i \leq p_B(a)] \right) = \\ &= (1 - 2^{-60I}) \left( 2^{-50(I+1)} \sum_{i \in \mathcal{G}} \mu(x_i) \right) = (1 - 2^{-60I}) \left( 2^{-50(I+1)} \mu(\mathcal{G}) \right) \geq \\ &\geq \frac{24}{25} (1 - 2^{-60I}) 2^{-50(I+1)} \geq 0.9 \cdot 2^{-50(I+1)} \end{aligned} \tag{4}$$

where the equality follows again from claim 2. This proves the second claim of Lemma 3.

The following claim asserts that if $\mathcal{S}$ occurs, then the distribution of $a$ is indeed close to $\mu$.

**Claim 4.** Let $\mu_1$ be the distribution of $a|\mathcal{S}$. Then $|\mu_1 - \mu| \leq 2/9$.
*Proof.* The claim follows directly from proposition 3. We defer the proof to the appendix.

We turn to the "Furthermore" part of of Lemma 3. The protocol $\Pi_1$ satisfies the premises of the lemma, except it has a high communication cost. This is due to the fact that Alice explicitly sends $a$ to Bob. To reduce the communication, Alice will instead send $O(I)$ random hash values of $a$, and Bob will add corresponding consistency constraints to his set of candidates. The final protocol $\Pi_2$ is given in Figure 1.

Let $\mathcal{E}$ denote the event that in step 3 of the protocol, Bob finds an element $x_i \neq a$ (that is, the probability that the protocol outputs "success" but

**Information-cost sampling protocol $\Pi_2$**

1. Alice computes the set $\mathcal{A}$. Bob computes the set $\mathcal{B}$.
2. If $\mathcal{A} = \emptyset$, the protocol fails. Otherwise, Alice finds the first element $a \in \mathcal{A}$ and sets $x_A = a$. She then computes $d = \lceil 211I \rceil$ random hash values $h_1(a), \ldots, h_d(a)$, where the hash functions are evaluated using public randomness.
3. Alice sends the values $\{h_j(a)\}_{1 \le j \le d}$ to Bob.
4. Bob finds the first index $\tau$ such that there is a $b \in \mathcal{B}$ for which $h_j(b) = h_j(a)$ for $j = 1..d$ (if such an $\tau$ exists). Bob outputs $x_B = x_\tau$. If there is no such index, the protocol fails.
5. Bob outputs $x_B$ ("success").
6. Alice outputs $x_A$.

**Fig. 1.** The sampling protocol $\Pi_2$ from Lemma 3

$x_A \neq x_B$). We upper bound the probability of $\mathcal{E}$. Given $a \in \mathcal{A}$ and $x_i \in \mathcal{B}$ such that $a \neq x_i$, the probability (over possible choices of the hash functions) that $h_j(a) = h_j(x_i)$ for $j = 1..d$ is $2^{-d} \le 2^{-211I}$. For any $t$, $\mathbf{P}[t \in \mathcal{B}] \le \frac{1}{|\mathcal{U}|} \sum_{x_i \in \mathcal{U}} p_B(x_i) q_B(x_i) \cdot 2^{50(I+1)} = \frac{1}{|\mathcal{U}|} \sum_{x_i \in \mathcal{U}} \nu_B(x_i) \cdot 2^{50(I+1)} = 2^{50(I+1)}/|\mathcal{U}|$. Thus, by a union bound we have

$$\begin{aligned} \mathbf{P}[\mathcal{E}] &\le \mathbf{P}[\exists x_i \in \mathcal{B} \; s.t \; x_i \neq a \; \wedge \; h_j(a) = h_j(x_i) \; \forall \; j = 1, \ldots, d] \le \\ &\le T \cdot 2^{50(I+1)} \cdot 2^{-d}/|\mathcal{U}| = 2^{150(I+1)} \cdot 60I \cdot 2^{-211I} \ll 2^{-60I}. \end{aligned} \tag{5}$$

By a slight abuse of notation, let $\Pi_2$ be the distribution of $\Pi_2$'s output. Similarly, denote by $\Pi_1$ the distribution of the output of protocol $\Pi_1$. Note that if $\mathcal{E}$ does not occur, then the outcome of the execution of $\Pi_2$ is identical to the outcome of $\Pi_1$. Since $\mathbf{P}[\mathcal{E}] \le 2^{-60I}$, we have

$$|\Pi_2 - \Pi_1| = \Pr[\mathcal{E}] \cdot |[\Pi_2 | \mathcal{E}] - [\Pi_1 | \mathcal{E}]| \le 2 \cdot 2^{-60I} \ll 2^{-59I}$$

which finishes the proof of the lemma.

*Remark 2.* The communication cost of the sampling protocol $\Pi_2$ can be reduced from $O(I_\mu)$ to $O(1)$ (more precisely, to only two bits) in the following way: Instead of having Alice compute the hash values privately and send them to Bob in step 2 and 3 of the protocol, the players can use their shared randomness to sample $d = O(I_\mu)$ random hash values $h_1(b_1), \ldots, h_d(b_d)$ (where the $b_i$'s are random independent strings in $\mathcal{U}$), and Alice will only send Bob a single bit indicating whether those hash values match the hashing of her string $a$ (i.e, $h_i(b_i) = h_i(a)$ for all $i \in [d]$). In step 4 Bob will act as before, albeit comparing the hashes of his candidate $b$ to the random hashes $h_i(b_i)$, and output success ("1") if the hashes match. Note that this modification incurs an additional loss of $2^{-d} = 2^{-O(I_\mu)}$ in the success probability of the protocol (as this is the probability that $h_i(b_i) = h_i(a)$ for all $i \in [d]$), but since the success probability we are shooting for is already of the order $2^{-O(I_\mu)}$, we can afford this loss. This modification was observed in [KLL+12].

With Lemma 3 in hand, we are now ready to prove Theorem 2.

**Proof of Theorem 2.** Let $\pi$ be a protocol that realizes the value $I_\mu := \mathsf{IC}_\mu(f, 1/10)$. In other words, $\pi$ has an error rate of at most $1/10$ and information cost of at most $I_\mu$ with respect to $\mu$. Denote by $\pi_{xy}$ the random variable that represents that transcript $\pi$ given the inputs $(x, y)$, and by $\pi_x$ (resp. $\pi_y$) the protocol conditioned on only the input $x$ (resp. $y$). We denote by $\pi_{XY}$ the transcripts where $(X, Y)$ are also a pair of random variables. By Claim 3, we know that

$$I_\mu = I(X; \pi_{XY}|Y) + I(Y; \pi_{XY}|X) = \mathbf{E}_{(x,y)\sim\mu}[\mathbf{D}\left(\pi_{xy}||\pi_x\right) + \mathbf{D}\left(\pi_{xy}||\pi_y\right)]. \quad (6)$$

Let us now run the sampling algorithm $\Pi_1$ from Lemma 3, with the distribution $\mu$ taken to be $\pi_{xy}$, the distributions $\nu_A$ and $\nu_B$ taken to be $\pi_x$ and $\pi_y$ respectively, and $I$ taken to be $20I_\mu$.

At each node $v$ of the protocol tree that is owned by player $X$ let $p_0(v)$ and $p_1(v) = 1 - p_0(v)$ denote the probabilities that the next bit sent by $X$ is 0 and 1, respectively. For nodes owned by player $Y$, let $q_0(v)$ and $q_1(v) = 1 - q_0(v)$ denote the probabilities that the next bit sent by $Y$ is 0 and 1, respectively, *as estimated by player $X$ given the input $x$*. For each leaf $\ell$ let $p_X(\ell)$ be the product of all the values of $p_b(v)$ from the nodes that are owned by $X$ along the path from the root to $\ell$; let $q_X(\ell)$ be the product of all the values of $q_b(v)$ from the nodes that are owned by $Y$ along the path from the root to $\ell$. The values $p_Y(\ell)$ and $q_Y(\ell)$ are defined similarly. For each $\ell$ we have $\mathbf{P}[\pi_{xy} = \ell] = p_X(\ell) \cdot p_Y(\ell)$, $\mathbf{P}[\pi_x = \ell] = p_X(\ell) \cdot q_X(\ell)$, and $\mathbf{P}[\pi_y = \ell] = p_Y(\ell) \cdot q_Y(\ell)$. Thus we can apply Lemma 3 so as to obtain the following protocol $\pi'$ for computing $f$:

- If $\Pi_1$ fails, we return a random unbiased coin flip.
- If $\Pi_1$ succeeds, we return the final bit of the transcript sample $T$. Denote this bit by $T_{out}$.

To prove the correctness of the protocol, let $\mathcal{Z}$ denote the event that both $\mathbf{D}\left(\pi_{xy}||\pi_x\right) \leq 20I_\mu$ and $\mathbf{D}\left(\pi_{xy}||\pi_y\right) \leq 20I_\mu$. By (6) and Markov inequality, $\Pr[\mathcal{Z}] \geq 19/20$ (where the probability is taken with respect to $\mu$). Denote by $\delta$ the probability that $\Pi_1$ succeeds. By the assertions of Lemma 3, $\delta \geq 0.9 \cdot 2^{-50(I+1)}$. Furthermore, if $\Pi_1$ succeeds, then we have $|T - \pi_{xy}| \leq 2/9$, which in particular implies that $\mathbf{P}[T_{out} = \pi_{out}] \geq 7/9$. Finally, $\mathbf{P}[\pi_{out} = f(x, y)] \geq 9/10$, since $\pi$ has error at most $1/10$ with respect to $\mu$. Now, let $\mathcal{W}$ denote the indicator variable whose value is 1 iff $\pi'(x, y) = f(x, y)$. Putting together the above,

$$\mathbf{E}[\mathcal{W} \mid \mathcal{Z}] = (1 - \delta) \cdot \frac{1}{2} + \delta \cdot \left(\frac{7}{9} - \frac{1}{10}\right) > \frac{1}{2} + \delta \cdot \frac{1}{6} > \frac{1}{2} + \frac{1}{8} \cdot 2^{-50(I+1)}. \quad (7)$$

On the other hand, note that by lemma 3 the probability that $\Pi_1$ succeeds is at most $2^{-50(I+1)}$ (no matter how large $\mathbf{D}\left(\pi_{xy}||\pi_x\right)$ and $\mathbf{D}\left(\pi_{xy}||\pi_y\right)$ are!), and so $\mathbf{E}[\mathcal{W} \mid \neg\mathcal{Z}] \geq (1 - 2^{-50(I+1)})/2$.

Hence we conclude that

$$\mathbf{E}[\mathcal{W}] = \mathbf{E}[\mathcal{W} \mid \mathcal{Z}] \cdot \mathbf{P}[\mathcal{Z}] + \mathbf{E}[\mathcal{W} \mid \neg\mathcal{Z}] \cdot \mathbf{P}[\neg\mathcal{Z}] \geq \left(\frac{1}{2} + \frac{1}{8} \cdot 2^{-50(I+1)}\right) \cdot \frac{19}{20} +$$
$$+ \left(1 - 2^{-50(I+1)}\right) \cdot \frac{1}{2} \cdot \frac{1}{20} \geq \frac{1}{2} + \frac{1}{12} \cdot 2^{-50(I+1)} > \frac{1}{2} + \frac{1}{12} \cdot 2^{-1000(I_\mu+1)}.$$

Finally, Lemma 3 asserts that $|\mathit{\Pi}_1 - \mathit{\Pi}_2| < 2^{-59I}$. Thus if we replace $\mathit{\Pi}_1$ by $\mathit{\Pi}_2$ in the execution of protocol $\pi'$, the success probability decreases by at most $2^{-59I} \ll \frac{1}{12} \cdot 2^{-50(I+1)}$. Furthermore, the amount of communication used by $\pi'$ is now

$$211I = 4220I_\mu = O(I_\mu).$$

Hence we conclude that with this modification, $\pi'$ has the following properties:

- $\mathsf{CC}(\pi') = 4220 \cdot I_\mu$;
- $\mathbf{P}_{(x,y)\sim\mu}[\pi'(x,y) = f(x,y)] \geq 1/2 + 2^{-1000(I_\mu+1)-4}$;

which completes the proof.

*Remark 3.* Using similar techniques, it was recently shown in [Bra11] that any function $f$ whose information complexity is $I$ has communication cost at most $2^{O(I)}$ [1], thus implying that $IC(f) \geq \Omega(\log(CC(f)))$. We note that this result can be easily derived (up to constant factors) from Theorem 2. Indeed, applying the "compressed" protocol $2^{O(I)}\log(1/\epsilon)$ independent times and taking a majority vote guarantees an error of at most $\epsilon$ (by a standard Chernoff bound[2]), with communication $O(I) \cdot 2^{O(I)} = 2^{O(I)}$. Thus, our result is strictly stronger than the former one.


### Acknowledgments

We thank Ankit Garg and several anonymous reviewers from RANDOM 12' for their useful comments and helpful discussions.


## References


AL11. G. Asharov and Y. Lindell. A full proof of the bgw protocol for perfectly-secure multiparty computation. *Advances in CryptologyCRYPTO 2011*, 2011.

BBCR10. Boaz Barak, Mark Braverman, Xi Chen, and Anup Rao. How to compress interactive communication. In *Proceedings of the 42nd Annual ACM Symposium on Theory of Computing*, 2010.

BOGW88. M. Ben-Or, S. Goldwasser, and A. Wigderson. Completeness theorems for non-cryptographic fault-tolerant distributed computation. In *Proceedings of the twentieth annual ACM symposium on Theory of computing*, pages 1–10. ACM, 1988.


[1] More precisely, it shows that for any distribution $\mu$, $D^{\mu}_{\varepsilon+\delta}(f) = 2^{O(1+\mathsf{IC}_\mu(f,\varepsilon)/\delta^2)}$.

[2] See N.Alon and J. Spencer, "The Probabilistic Method" (Third Edition) ,Corollary A.1.14, p.312.

BR10. Mark Braverman and Anup Rao. Information equals amortized communication. *CoRR*, abs/1106.3595, 2010.

BR11. M. Braverman and A. Rao. Information equals amortized communication. *Arxiv preprint arXiv:1106.3595*, 2011.

Bra11. Mark Braverman. Interactive information complexity. *Electronic Colloquium on Computational Complexity (ECCC)*, 18:123, 2011.

BYCKO93. R. Bar-Yehuda, B. Chor, E. Kushilevitz, and A. Orlitsky. Privacy, additional information and communication. *Information Theory, IEEE Transactions on*, 39(6):1930–1943, 1993.

BYJKS04. Ziv Bar-Yossef, T. S. Jayram, Ravi Kumar, and D. Sivakumar. An information statistics approach to data stream and communication complexity. *Journal of Computer and System Sciences*, 68(4):702–732, 2004.

CK89. B. Chor and E. Kushilevitz. A zero-one law for boolean privacy. In *Proceedings of the twenty-first annual ACM symposium on Theory of computing*, pages 62–72. ACM, 1989.

CSWY01. Amit Chakrabarti, Yaoyun Shi, Anthony Wirth, and Andrew Yao. Informational complexity and the direct sum problem for simultaneous message complexity. In Bob Werner, editor, *Proceedings of the 42nd Annual IEEE Symposium on Foundations of Computer Science*, pages 270–278, Los Alamitos, CA, October 14–17 2001. IEEE Computer Society.

HJMR07. P. Harsha, R. Jain, D. McAllester, and J. Radhakrishnan. The communication complexity of correlation. In *Computational Complexity, 2007. CCC'07. Twenty-Second Annual IEEE Conference on*, pages 10–23. IEEE, 2007.

Jai10. R. Jain. A strong direct product theorem for two-way public coin communication complexity. *Arxiv preprint arXiv:1010.0846*, 2010.

JSR08. R. Jain, P. Sen, and J. Radhakrishnan. Optimal direct sum and privacy trade-off results for quantum and classical communication complexity. *Arxiv preprint arXiv:0807.1267*, 2008.

Kla04. Hartmut Klauck. Quantum and approximate privacy. *Theory Comput. Syst.*, 37(1):221–246, 2004.

Kla10. H. Klauck. A strong direct product theorem for disjointness. In *Proceedings of the 42nd ACM symposium on Theory of computing*, pages 77–86. ACM, 2010.

KLL$^+$12. I. Kerenidis, S. Laplante, V. Lerays, J. Roland, and D. Xiao. Lower bounds on information complexity via zero-communication protocols and applications. *Arxiv preprint arXiv:1204.1505*, 2012.

KN97. Eyal Kushilevitz and Noam Nisan. *Communication complexity*. Cambridge University Press, New York, 1997. 96012840 96012840 Eyal Kushilevitz, Noam Nisan.

KSDW04. H. Klauck, R. Spalek, and R. De Wolf. Quantum and classical strong direct product theorems and optimal time-space tradeoffs. In *Foundations of Computer Science, 2004. Proceedings. 45th Annual IEEE Symposium on*, pages 12–21. IEEE, 2004.

LSS08. T. Lee, A. Shraibman, and R. Spalek. A direct product theorem for discrepancy. In *Computational Complexity, 2008. CCC'08. 23rd Annual IEEE Conference on*, pages 71–80. IEEE, 2008.

MNSW95. P.B. Miltersen, N. Nisan, S. Safra, and A. Wigderson. On data structures and asymmetric communication complexity. In *Proceedings of the twenty-seventh annual ACM symposium on Theory of computing*, pages 103–111. ACM, 1995.

Raz92. Alexander A. Razborov. On the distributional complexity of disjointness. *Theor. Comput. Sci.*, 106(2):385–390, 1992.
Sha03. R. Shaltiel. Towards proving strong direct product theorems. *Computational Complexity*, 12(1):1–22, 2003.
Smi88. D V Smirnov. Shannon's information methods for lower bounds for probabilistic communication complexity. Master's thesis, Moscow State University, 1988.
Vio11. Emanuele Viola. The communication complexity of addition. *Electronic Colloquium on Computational Complexity (ECCC)*, 18:152, 2011.
Yao79. Andrew Chi-Chih Yao. Some complexity questions related to distributive computing (preliminary report). In *STOC*, pages 209–213, 1979.

## A Communication Complexity

Let $\mathcal{X}, \mathcal{Y}$ denote the set of possible inputs to the two players, who we name A and B. We view a *private coins protocol* for computing a function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{Z}_K$ as a rooted tree with the following structure:

- Each non-leaf node is *owned* by A or by B.
- Each non-leaf node owned by a particular player has a set of children that are owned by the other player. Each of these children is labeled by a binary string, in such a way that this coding is prefix free: no child has a label that is a prefix of another child.
- Every node is associated with a function mapping $\mathcal{X}$ to distributions on children of the node and a function mapping $\mathcal{Y}$ to distributions on children of the node.
- The leaves of the protocol are labeled by output values.

A public coin protocol is a distribution on private coins protocols, run by first using shared randomness to sample an index $r$ and then running the corresponding private coin protocol $\pi_r$. Every private coin protocol is thus a public coin protocol. The protocol is called deterministic if all distributions labeling the nodes have support size 1.

**Definition 8.** *The* communication cost *(or communication complexity) of a public coin protocol $\pi$, denoted $\mathsf{CC}(\pi)$, is the maximum number of bits that can be transmitted in any run of the protocol.*

**Definition 9.** *The* number of rounds *of a public coin protocol is the maximum depth of the protocol tree $\pi_r$ over all choices of the public randomness.*

## B Proof of Claim 3 (from [Bra11])

*Proof.* Recall that $\mathbf{D}(\mu||\nu) = \sum_{x \in \mathcal{U}} \mu(x) \log \frac{\mu(x)}{\nu(x)}$. Denote by $\mathcal{N} = \{x : \ \mu(x) < \nu(x)\}$ – the terms that contribute a negative amount to $\mathbf{D}(\mu||\nu)$. First we observe that for all $0 < x < 1$, $x \log x > -1$, and thus

$$\sum_{x \in \mathcal{N}} \mu(x) \log \frac{\mu(x)}{\nu(x)} = \sum_{x \in \mathcal{N}} \nu(x) \cdot \frac{\mu(x)}{\nu(x)} \log \frac{\mu(x)}{\nu(x)} \geq \sum_{x \in \mathcal{N}} \nu(x) \cdot (-1) > -1.$$

Denote by $\mathcal{L} = \left\{x : \ 2^{(I+1)/\varepsilon} \cdot \nu(x) < \mu(x)\right\}$; we need to show that $\mu(\mathcal{L}) < \varepsilon$. For each $x \in \mathcal{L}$ we have $\log \frac{\mu(x)}{\nu(x)} > (I+1)/\varepsilon$. Thus

$$I \geq \mathbf{D}\left(\mu||\nu\right) \geq \sum_{x\in\mathcal{L}} \mu(x) \log \frac{\mu(x)}{\nu(x)} + \sum_{x\in\mathcal{N}} \mu(x) \log \frac{\mu(x)}{\nu(x)} > \mu(\mathcal{L}) \cdot (I+1)/\varepsilon - 1,$$

implying $\mu(\mathcal{L}) < \varepsilon$.

## C Proof of Claim 4

*Proof.* For any $x_i \in \mathcal{U}$,

$$\mu_1(x_i) = \Pr(a = x_i \mid \mathcal{S}) \leq \frac{\mu(x_i)2^{-50(I+1)}}{\Pr[\mathcal{S}]} \leq \frac{\mu(x_i)}{0.9} = (1 + 1/9)\mu(x_i) \qquad (8)$$

where the last inequality follows from Proposition 3. Hence, $|\mu_1 - \mu| =$

$$2\Bigg(\sum_{x_i:\mu_1(x_i)\geq\mu(x_i)} \mu_1(x_i) - \mu(x_i)\Bigg) \leq 2\Bigg(\sum_{x_i:\mu_1(x_i)\geq\mu(x_i)} (1+1/9)\mu(x_i) - \mu(x_i)\Bigg) \leq \frac{2}{9}$$

This proves claim (3) of the lemma.

## D Proof of Lemma 1: The discrepancy of the Greater-Than function

We consider the Greater-Than function on $n$-bit strings. We start by defining the "hard" distribution $\mu$. A pair $(x, y)$ is sampled as follows:

1. Sample an index $k \in \{1, \ldots, n\}$ uniformly at random.
2. Sample $z_1, \ldots, z_{k-1}, w, x_{k+1}, \ldots, x_n, y_{k+1}, \ldots, y_n$ — uniformly random bits.
3. Let $x = z_1, \ldots, z_{k-1}, w, x_{k+1}, \ldots, x_n$, $y = z_1, \ldots, z_{k-1}, \overline{w}, y_{k+1}, \ldots, y_n$.

Denote this distribution by $\mu_n$. Let $GT_n(x, y) = 1$ iff $x > y$. We will prove the following Lemma:

**Lemma 4.** *The discrepancy of $GT_n$ with respect to $\mu_n$ satisfies*

$$Disc_{\mu_n}(GT_n) < \frac{20}{\sqrt{n}}.$$

In fact, to facilitate an inductive proof, we will show a slightly stronger statement:

**Lemma 5.** *Let $R = S \times T$ be a rectangle in $\{0,1\}^n \times \{0,1\}^n$. Let $s := |S|/2^n$ and $t := |T|/2^n$ be the uniform size of $S$ and $T$ respectively. Then*

$$Disc_{\mu_n}(GT_n, R) < \frac{20\sqrt{st}}{\sqrt{n}}.$$

Note that Lemma 5 immediately implies Lemma 4.

*Proof.* We prove Lemma 5 by induction on $n$. The statement is trivially true for $n = 1$. Assume the statement is true for $n-1$. Our goal is to prove it for $n$. Let $R = S \times T$ be any rectangle in $\{0,1\}^n \times \{0,1\}^n$. By a slight abuse of notation we write:

$$Disc_{\mu_n}(GT_n, R) = \Pr_{\mu_n}[f(x,y) = 1 \wedge (x,y) \in R] - \Pr_{\mu_n}[f(x,y) = 0 \wedge (x,y) \in R],$$

and prove an upper bound on this quantity (without $|\cdot|$). The matching upper bound on $-Disc_{\mu_n}(GT_n, R)$ follows by an identical argument.

Let $s := |S|/2^n$ and $t := |T|/2^n$. Denote by $S_0$ the set of strings in $S$ that begin with a 0, and $S_1 := S \setminus S_0$. Similarly, define $T_0$ and $T_1$. Further, let $p := |S_0|/|S|$ and $q := |T_0|/|T|$.

Note that restricted to $S_0 \times T_0$, $\mu_n$ is the same distribution as $\mu_{n-1}$, scaled by a factor of $\frac{n-1}{2n}$. Moreover, $s_0 := |S_0|/2^{n-1} = ps2^n/2^{n-1} = 2ps$. Similarly, $s_1 := |S_1|/2^{n-1} = 2(1-p)s$, $t_0 := |T_0|/2^{n-1} = 2qt$, $t_1 := |T_1|/2^{n-1} = 2(1-q)t$. Putting these pieces together, and applying the inductive hypothesis we get:

$$\begin{gathered}
Disc_{\mu_n}(GT_n, S \times T) = Disc_{\mu_n}(GT_n, S_0 \times T_0) + Disc_{\mu_n}(GT_n, S_1 \times T_1) + \\
Disc_{\mu_n}(GT_n, S_1 \times T_0) + Disc_{\mu_n}(GT_n, S_0 \times T_1) = \\
\frac{n-1}{2n} \cdot Disc_{\mu_{n-1}}(GT_{n-1}, S_0 \times T_0) + \frac{n-1}{2n} \cdot Disc_{\mu_{n-1}}(GT_{n-1}, S_1 \times T_1) + \\
\frac{2}{n}(1-p)sqt - \frac{2}{n}ps(1-q)t < \\
\frac{n-1}{2n} \cdot \frac{20\sqrt{s_0 t_0}}{\sqrt{n-1}} + \frac{n-1}{2n} \cdot \frac{20\sqrt{s_1 t_1}}{\sqrt{n-1}} + \frac{2}{n}(q-p)st = \\
\frac{1}{\sqrt{n}} \left( \sqrt{\frac{n-1}{n}} \cdot \left( 20\sqrt{pq} \cdot \sqrt{st} + 20\sqrt{(1-p)(1-q)} \cdot \sqrt{st} \right) + \frac{2}{\sqrt{n}}(q-p)st \right).
\end{gathered} \tag{9}$$

If $q - p < 0$, we continue (9) as follows:

$$\begin{aligned}
RHS \leq \frac{1}{\sqrt{n}} \left( \sqrt{\frac{n-1}{n}} \cdot \left( 20\sqrt{pq} \cdot \sqrt{st} + 20\sqrt{(1-p)(1-q)} \cdot \sqrt{st} \right) \right) \leq \\
\frac{20\sqrt{st}}{\sqrt{n}} \cdot \left( \sqrt{pq} + \sqrt{(1-p)(1-q)} \right) \leq \frac{20\sqrt{st}}{\sqrt{n}},
\end{aligned}$$

where the last inequality follows from simple calculations.

On the other hand, in the more difficult case when $q-p \geq 0$, we use the fact the $st \leq 1$ to continue (9) as follows:

$$RHS \leq$$
$$\frac{1}{\sqrt{n}}\left(\sqrt{\frac{n-1}{n}} \cdot \left(20\sqrt{pq} \cdot \sqrt{st} + 20\sqrt{(1-p)(1-q)} \cdot \sqrt{st}\right) + \frac{2}{\sqrt{n}}(q-p)\sqrt{st}\right) =$$
$$\frac{20\sqrt{st}}{\sqrt{n}}\left(\sqrt{\frac{n-1}{n}} \cdot \left(\sqrt{pq} + \sqrt{(1-p)(1-q)}\right) + \frac{1}{10\sqrt{n}}(q-p)\right) \quad (10)$$

Next, we use the readily verifiable facts that $\sqrt{\frac{n-1}{n}} < 1 - \frac{1}{2n}$ and that $\sqrt{pq} + \sqrt{(1-p)(1-q)} \leq 1 - (q-p)^2/4$, to continue (10) as follows:

$$RHS \leq \frac{20\sqrt{st}}{\sqrt{n}}\left(\left(1-\frac{1}{2n}\right) \cdot \left(1-(q-p)^2/4\right) + \frac{1}{10\sqrt{n}}(q-p)\right) \leq$$
$$\frac{20\sqrt{st}}{\sqrt{n}}\left(\left(1-\frac{1}{4n}-(q-p)^2/8\right) + \frac{1}{10\sqrt{n}}(q-p)\right) =$$
$$\frac{20\sqrt{st}}{\sqrt{n}}\left(1-\left(\frac{1/(2n)+(q-p)^2/4}{2}\right) + \frac{1}{10\sqrt{n}}(q-p)\right) \leq$$
$$\frac{20\sqrt{st}}{\sqrt{n}}\left(1-\sqrt{\frac{1}{2n}\cdot\frac{(q-p)^2}{4}} + \frac{1}{10\sqrt{n}}(q-p)\right) =$$
$$\frac{20\sqrt{st}}{\sqrt{n}}\left(1-\frac{1}{\sqrt{8}\sqrt{n}}(q-p) + \frac{1}{10\sqrt{n}}(q-p)\right) \leq \frac{20\sqrt{st}}{\sqrt{n}}, \quad (11)$$

where the third-to-last inequality follows from the fact that for all $0 \leq a, b \leq 1$, $(1-a)(1-b) \leq 1 - a/2 - b/2$, and the second-to-last one in an application of the AM-GM inequality.

## E Sampling protocol from Lemma 3

| **Information-cost sampling protocol** $\Pi_1$ |
|---|
| 1. Alice computes the set $\mathcal{A}$. Bob computes the set $\mathcal{B}$.<br>2. If $\mathcal{A} = \emptyset$, the protocol fails, otherwise Alice finds the first element $a \in \mathcal{A}$, and sends $a$ to Bob.<br>3. Bob checks if $a \in \mathcal{B}$. If not, the protocol fails.<br>4. Alice and Bob output $a$ ("success"). |

**Fig. 2.** The sampling protocol $\Pi_1$ from Lemma 3