Electronic Colloquium on Computational Complexity, Report No. 31 (2015)

Amplification of One-Way Information Complexity via Codes and Noise Sensitivity

Marco Molinaro
Delft University of Technology

David P. Woodruff
IBM Almaden

Grigory Yaroslavtsev
University of Pennsylvania

Abstract

We show a new connection between the information complexity of one-way communication problems under product distributions and a relaxed notion of list-decodable codes. As a consequence, we obtain a characterization of the information complexity of one-way problems under product distributions for *any error rate* based on covering numbers. This generalizes the characterization via VC dimension for constant error rates given by Kremer, Nisan, and Ron (CCC, 1999). It also provides an *exponential improvement in the error rate*, yielding tight bounds for a number of problems. In addition, our framework gives a new technique for analyzing the complexity of composition (e.g., XOR and OR) of one-way communication problems, connecting the difficulty of these problems to the *noise sensitivity* of the composing function. Using this connection, we strengthen the lower bounds obtained by Molinaro, Woodruff and Yaroslavtsev (SODA, 2013) for several problems in the distributed and streaming models, obtaining optimal lower bounds for finding the approximate closest pair of a set of points and the approximate largest entry in a matrix product. Finally, to illustrate the utility and simplicity of our framework, we show how it unifies proofs of existing 1-way lower bounds for sparse set disjointness, the indexing problem, the greater than function under product distributions, and the gap-Hamming problem under the uniform distribution.

1 Introduction

We consider the two-party one-way communication complexity model where Alice and Bob want to jointly compute a function $f : \mathcal{X} \times \mathcal{Y} \to \{0, 1\}$. More precisely, Alice holds an input $x \in \mathcal{X}$, Bob holds an input $y \in \mathcal{Y}$, and they have access to common random bits; Alice sends a (random) message to Bob, who then tries to output the value $f(x, y)$. The *cost* of a protocol is the maximum (over the inputs and the randomness) number of bits sent by Alice. The goal is to find a randomized protocol of minimum cost that for all inputs computes $f(x, y)$ with probability at least $1 - \alpha$; this minimum cost is denoted by $\mathsf{R}(f)^{\rightarrow}_{\alpha}$.

The one-way communication model has been studied in a number of works, including Yao [25], Papadimitriou and Sipser [20], Ablayev [1], Newman and Szegedy [17], and Kremer et al. [10]. It is particularly relevant to the *data stream* model in which an algorithm sees a stream of elements one at a time, and tries to compute a relation of these elements using as little space (in bits) as possible [16]. One way of lower-bounding the space complexity of data stream algorithms is to set up a one-way communication protocol in which Alice's message consists of the state of the streaming algorithm run on a stream created by Alice. Bob then continues the execution of the streaming algorithm on a stream he creates, and if from the output the players can solve a communication

ISSN 1433-8092

problem f, then the space complexity of the streaming algorithm must be at least the one-way communication complexity of f.

We will consider a distributional version of one-way communication complexity, in which Alice and Bob have inputs $(x, y) \sim \mu \times \nu$, where $\mu \times \nu$ is a *product distribution* on domains $\mathcal{X}$ and $\mathcal{Y}$. That is, Alice's input is drawn from μ, while Bob's input is drawn from ν, and the inputs are independent. We define $\mathsf{R}(f)^{\rightarrow,\square}_{\alpha}$ to be the maximum, over product distributions $\mu \times \nu$, of $\mathsf{D}(f)^{\rightarrow}_{\mu\times\nu,\alpha}$, where $\mathsf{D}(f)^{\rightarrow}_{\mu\times\nu,\alpha}$ is the minimum cost over deterministic protocols which compute f with error probability at most α when the input is drawn from $\mu \times \nu$. Kremer, Nisan, and Ron [10] show that for constant α and Boolean functions f, $\mathsf{R}(f)^{\rightarrow,\square}_{\alpha} = \Theta(VC)$, where VC is the VC-dimension of the class $\{f_x : \mathcal{Y} \to \{0, 1\} \mid x \in \mathcal{X}\}$ obtained by seeing the rows of the communication matrix of f as functions. Equivalently, VC is the dimension of the largest hypercube which is a submatrix of the communication matrix.

Unfortunately, a characterization for constant α does not suffice for streaming applications. This was the focus of work by Jayram and Woodruff [9], who showed that for a number of streaming problems, such as estimating the empirical entropy and Euclidean norm (and more generally the ℓ_p-norm for $p \leq 2$), the problem requires an extra multiplicative $\log(1/\delta)$ in the space complexity if the algorithm succeeds with probability at least $1 - \delta$. This was shown using one-way communication under a product distribution, and so obtaining the extra $\log(1/\delta)$ factor had to be shown by ways other than resorting to the VC-dimension, since we do not have a general characterization of problems showing how their communication cost scales with the error probability.

Besides single-shot problems, the gap in our understanding of the dependence on the error probability also manifests itself for solving a composition of many copies of a problem simultaneously with constant probability. The authors [14] previously showed that for several streaming problems, the communication cost of solving n copies of a problem simultaneously with probability $2/3$ scales as n times the cost of solving each copy with probability $1 - 1/n$. This composition theorem critically uses that a protocol must obtain a correct output for each of the n instances, and it is unknown if such a statement holds for other composition functions, such as the OR or XOR functions. This has led to $\log n$ factor gaps in the upper and lower bounds for streaming problems such as

- $\mathsf{ClosestPair}$: Alice has n points $p_1, \ldots, p_n$ in $\mathbb{R}^d$, Bob has n points $q_1, \ldots, q_n$ in $\mathbb{R}^d$, and they would like to find a pair p_i, q_j for which $\|p_i - q_j\|_2 \leq (1 + \epsilon) \min_{i',j'} \|p_{i'} - q_{j'}\|_2$, and
- $\mathsf{MatrixProduct}$: Alice has an $n \times d$ matrix A with rows of unit norm, Bob has a $d \times n$ matrix B with columns of unit norm. They want to approximate $\max_{i,j} |AB|_{i,j}$ up to an additive ϵ.

Our Contributions: We introduce the notion of an (α, β)-code and use it to capture the distance between rows of a communication matrix of a function $f : \mathcal{X} \times \mathcal{Y} \to \{0, 1\}$. Informally speaking, this notion says that under Alice's distribution μ, with probability at most β, two independently sampled rows have relative Hamming distance at most α when weighted with respect to Bob's input distribution ν. This notion thus captures the (pairwise) correlation of rows of a communication matrix, with respect to distributions μ and ν. We show that the one-way information cost of protocols under distribution $\mu \times \nu$ with error probability α is $\Omega(\log 1/\beta)$. This result is based on a Fano's inequality for list-decoding that may be of independent interest. This gives a surprisingly generic way of characterizing lower bounds in terms of the error probability.

Characterization Theorem: We use our characterization in terms of codes to obtain a characterization of 1-way communication complexity in terms of *packing numbers*. Here, given a pseudo-metric space $(\mathcal{X}, d)$, the α-packing number is the largest set of points in $\mathcal{X}$ with pairwise distance at least α. We show that $\max_\nu \Omega(\log p_{8\alpha,\nu}) \leq \mathsf{R}(f)_\alpha^{\rightarrow,[]} \leq \max_\nu O(\log p_{\alpha,\nu})$, where $p_{\alpha,\nu}$ is the packing number of the pseudo-metric space $(\{f(x)\}_{x\in\mathcal{X}}, \|\cdot\|_\nu)$, where $\{f(x)\}_{x\in\mathcal{X}}$ is the family of functions corresponding to rows of the communication matrix, and $\|\cdot\|_\nu$ is the weighted relative Hamming distance according to ν. This gives a strengthening of the result of Kremer, Nisan, and Ron [10] since it gives a tight characterization in terms of the error probability α (up to the distinction of α in the upper bound and 8α in the lower bound). We need to resort to packing numbers, since as observed by Jayram and Woodruff [9], there is no characterization possible in terms of the VC-dimension (as used by [10]). However, by relating packing numbers to VC-dimension, we considerably strengthen the result of [10] which states that $(1 - H(\alpha))VC \leq \mathsf{R}(f)_\alpha^{\rightarrow,[]} \leq O(VC\frac{1}{\alpha}\log\frac{1}{\alpha})$, where VC denotes the VC-dimension of f. We obtain the stronger result that $(1-H(\alpha))VC \leq \mathsf{R}(f)_\alpha^{\rightarrow,[]} \leq O(VC\log(\frac{1}{\alpha}))$. As an example, we use this to show that $\mathsf{R}_\alpha^{\rightarrow,[]}(GT) = \Theta(\log\frac{1}{\alpha})$ where GT is the greater-than function. This is an exponential improvement over the result based on VC-dimension.

Composition Theorem: Next we introduce the notion of noise sensitivity, which captures how a communication problem f whose rows form an (α, β)-code behaves under composition. There is a line of work on understanding how primitive problems behave under composition [12, 18, 3, 11, 22]; our work adds to this by characterizing the composition in terms of codes. The noise sensitivity of a composing function g on k inputs with respect to an input distribution μ^k intuitively captures how likely two independent samples of inputs to g from μ^k are likely to result in differing outputs of g. We show that if f is an (α, β)-code with respect to $\mu \times \nu$, then $g \circ f$ is an (α', β')-code with respect to $\mu^k \times \nu^k$ for certain α' and β' related to the noise sensitivity of g, as well as to α and β.

Streaming Applications: As the main application of our composition theorem, we consider the primitive problem f in which Alice holds a string $x \in [k]^m$, Bob has an $\ell \in [k]$ and an index $j \in [m]$, and Bob would like to know if $x_j = \ell$. We show that f is an (α, β)-code for sufficiently good α and β, and we lower bound the noise sensitivity of the OR function. These results imply that solving the OR of k copies of f, denoted $\mathrm{OR}^k \circ f$, with constant probability has one-way communication complexity $\Omega(km \log k)$. For our streaming applications, we further consider an augmented version of this problem, in which Alice has t independent instances of $\mathrm{OR}^k \circ f$, and Bob would like to solve one of these t instances i chosen uniformly at random. Bob is also given Alice's input for the first $i-1$ instances. For this we show an $\Omega(tkm \log k)$ one-way communication lower bound for constant probability protocols. These results greatly strengthen the results in [14], which could only show this if $\mathrm{OR}^k \circ f$ were replaced with $\mathsf{ALLCOPIES}\ ^k \circ f$, the latter requiring a correct output to all k instances of f rather than just an OR of the k instances. Note that the output of $\mathrm{OR}^k \circ f$ is only a single bit, whereas the output of $\mathsf{ALLCOPIES}\ ^k \circ f$ consists of k bits, making the latter a significantly easier problem. Our result directly improves the streaming application lower bounds in [14], leading to the first tight one-way lower bounds for $\mathsf{ClosestPair}$ and $\mathsf{MatrixProduct}$. The details are in Section 6.

Unified Lower Bounds: To illustrate the power of the framework developed, we recover in a unified way several 1-way lower bounds from the literature, including sparse set disjointness [6, 4, 21] and indexing [9] under product distributions, and the gap-Hamming problem under the uniform distribution [24].

2 Preliminaries

Information Theory. We use the following notions from information theory (see [5] for more details). Given random variables X, Y and Z on a common probability space, we use $H(X)$ to denote the binary entropy of X and $H(X \mid Y)$ its conditional entropy given Y. The mutual information between X and Y is then defined as $\mathrm{I}(X;Y) = H(X) - H(X \mid Y)$, and the conditional mutual information given Z is $I(X;Y \mid Z) = H(X \mid Z) - H(X \mid (Y,Z))$. We will need the *data processing inequality*: for any arbitrary functions g, h, $\mathrm{I}(X;Y) \geq \mathrm{I}(g(X);h(Y))$.

Distributional and Information Complexity. Consider a function $f : \mathcal{X} \times \mathcal{Y} \to \{0,1\}$ and a distribution μ over $\mathcal{X} \times \mathcal{Y}$. The one-way *distributional complexity* of f with respect to μ, denoted $\mathsf{D}(f)^{\to}_{\mu,\alpha}$, is the smallest communication cost of a one-way deterministic protocol that outputs $f(x,y)$ on all but an α fraction of inputs weighted according to μ. The one-way *distributional complexity* of f, denoted $\mathsf{D}(f)^{\to}_{\alpha}$, is the supremum of $\mathsf{D}(f)^{\to}_{\mu,\alpha}$ over all distributions μ. The classic Yao's Minimax Theorem [25] shows that randomized and distributional complexity are the same: $\mathsf{R}(f)^{\to}_{\alpha} = \mathsf{D}(f)^{\to}_{\alpha}$. Motivated by this observation, define the product distribution complexity $\mathsf{R}(f)^{\to,\square}_{\alpha}$ as the supremum of $\mathsf{D}(f)^{\to}_{\mu\times\nu,\alpha}$ over all distributions μ for $\mathcal{X}$ and ν for $\mathcal{Y}$.

Now we define information complexity. Again we are given a distribution μ over $\mathcal{X} \times \mathcal{Y}$. Given a *randomized* one-way protocol for computing f, with $\mathsf{A}(x,r)$ denoting the message sent by Alice on input x and private randomness r, the *information cost* of this protocol is defined as $\mathrm{I}(\mathsf{A}(X,R);X \mid Y)$, where the pair (X,Y) is sampled from μ (and R is Alice's randomness, which is independent from X,Y). The *information complexity* with respect to μ, denoted $\mathsf{IC}(f)^{\to}_{\mu,\alpha}$, is the smallest information cost of a randomized one-way protocol computing $f(X,Y)$ with probability at least $1-\alpha$ (with respect to $(X,Y) \sim \mu$ and the private randomness of Alce and Bob). Finally the *information complexity* $\mathsf{IC}(f)^{\to}_{\alpha}$ is the supremum of $\mathsf{IC}(f)^{\to}_{\mu,\alpha}$ over all distributions μ. Similarly, the *information complexity over product distributions* $\mathsf{IC}(f)^{\to,\square}_{\alpha}$ is the supremum of $\mathsf{IC}(f)^{\to}_{\mu\times\nu,\alpha}$ over all distributions μ on $\mathcal{X}$ and ν on $\mathcal{Y}$. Notice that under a product distribution $(X,Y) \sim \mu \times \nu$ the information cost of a protocol becomes $\mathrm{I}(\mathsf{A}(X,R);X)$.

We have the following known relationship between information and distributional complexity (which follows from the entropy span bound and non-negativity of entropy): $\mathsf{R}(f)^{\to,\square}_{\alpha} \geq \mathsf{IC}(f)^{\to,\square}_{\alpha}$.

Notation. Given a function $f : \mathcal{X} \times \mathcal{Y} \to \{0,1\}$ and $x \in \mathcal{X}$, we use $f(x) : \mathcal{Y} \to \{0,1\}$ to denote the function $f(x)(y) = f(x,y)$. We say that $f(x)$ is a *row* of f (i.e., when f is seen as a matrix with rows indexed by $\mathcal{X}$ and columns indexed by $\mathcal{Y}$). Given a distribution ν over a set $\mathcal{Y}$, we define the semi-norm $\|.\|_{\nu}$ as $\|v\|_{\nu} = \mathbb{E}_{Y\sim\nu}[v(Y)]$ for all $v \in \mathbb{R}^{\mathcal{Y}}$. We also use $\|v\|_0$ to denote the number of non-zero entries of v. Finally, given a pseudo-metric space $(\mathcal{X}, d)$ and $x \in \mathcal{X}$, we use $B(x,\alpha)$ to denote the set of points in $\mathcal{X}$ at distance at most α from x.

3 Information Complexity and Relaxed Codes

Definition 3.1. *Consider a pseudo-metric space $(\mathcal{X}, d)$. A subset $\mathcal{C}$ is an (α, β)-*code *w.r.t. a distribution μ supported on $\mathcal{C}$ if for C, C' chosen independently from μ*

$$\Pr_{C,C'}(d(C, C') \leq \alpha) \leq \beta.$$

The following is the main result of this section which gives a lower bound on the information complexity of communication problems based on (α, β)-codes.

Theorem 3.2. *Consider a communication problem $f : \mathcal{S} \times \mathcal{L} \to \{0, 1\}$. Consider distributions μ (over $\mathcal{S}$) and ν (over $\mathcal{L}$) and suppose that the rows $\{f(s)\}_{s \in \mathcal{S}}$ form an (α, β)-code with respect to μ and the distance $\|.\|_\nu$. Then*

$$\mathsf{IC}(f)^{\rightarrow}_{(\mu\times\nu), \frac{\alpha}{8}} \geq \frac{1}{4} \log \frac{1}{4\beta} - 1.$$

The intuition is that if the rows of the communication problem are quite distinct from each other, a low error protocol allows Bob to recover the identity of the row that Alice's input is indexing, leading to a high information cost.

To make this intuition formal, we start by developing a list-decoding variant of Fano's inequality where a predictor outputs a prediction set, which might be of independent interest; the proof is deferred to Appendix B.

Lemma 3.3. *Consider a finite set $\mathcal{X}$ and an arbitrary set $\mathcal{R}$, and let μ and λ be distributions over $\mathcal{X}$ and $\mathcal{R}$ respectively. Also consider a (predictor) function $g : \mathcal{X} \times \mathcal{R} \to 2^{\mathcal{X}}$ such that for some $\beta \in (0, 1)$ we have $\Pr_{X\sim\mu, R\sim\lambda}(X \in g(X, R)$ and $\mu(g(X, R)) \leq \beta) \geq p$. Then $\mathrm{I}(X; g(X, R)) \geq p \log \frac{1}{\beta} - 1$.*

The next theorem connects this list-decoding version of Fano's inequality with (α, β)-codes; the mapping M next can be thought as an approximate decoder.

Theorem 3.4. *Consider a finite pseudo-metric space $(\mathcal{X}, d)$. Let $\mathcal{C} \subseteq \mathcal{X}$ be an (α, β)-code with respect to a distribution μ over $\mathcal{C}$. Consider an arbitrary space $\mathcal{R}$ with distribution λ. Consider the random variables $C \sim \mu$, $R \sim \lambda$ and a mapping $M : \mathcal{C} \times \mathcal{R} \to \mathcal{X}$satisfying $\Pr_{C,R}(d(M(C, R), C) \geq \frac{\alpha}{2}) \leq \frac{1}{4}$. Then*

$$\mathrm{I}(C; M(C, R)) \geq \frac{1}{4} \log \frac{1}{4\beta} - 1.$$

Proof. We employ Lemma 3.3 to the space $\mathcal{C} \times \mathcal{R}$. Construct the predictor $g : \mathcal{C} \times \mathcal{R} \to 2^{\mathcal{C}}$ given by $g(c, r) = B\big(M(c, r), \frac{\alpha}{2}\big)$; notice that $g(c, r)$ only depends on $M(c, r)$. We claim that

$$\Pr_{C\sim\mu, R\sim\lambda}(C \in g(C, R) \text{ and } \mu(g(C, R)) \leq 4\beta) \geq \frac{1}{2}. \tag{1}$$

Let $\mathcal{E}$ denote the event $\{C \in g(C, R)$ and $\mu(g(C, R)) \leq 4\beta\}$, and change the second term to define the event $\mathcal{E}' = \{d(M(C, R), C) \leq \frac{\alpha}{2}$ and $\mu(B(C, \alpha)) \leq 4\beta\}$ (notice that $C \in g(C, R)$ is equivalent to $d(M(C, R), C) \leq \frac{\alpha}{2}$). We claim that $\mathcal{E}'$ implies $\mathcal{E}$: if $\mathcal{E}'$ holds then using its first part and the triangle inequality we get $B(M(C, R), \frac{\alpha}{2}) \subseteq B(C, \alpha)$, so its second part gives $\mu(g(C, R)) = \mu(B(M(C, R), \frac{\alpha}{2}) \leq \mu(B(C, R)) \leq 4\beta$, proving the claim. So to prove inequality (1) it suffices to show $\Pr(\mathcal{E}') \geq \frac{1}{2}$.

Directly from the guarantees of M we have $\Pr(d(M(C,R),C) \leq \frac{\alpha}{2}) \geq \frac{3}{4}$. For $\mu(B(C,\alpha) \leq 4\beta$, notice that for a random variable $C' \sim \mu$ independent of C we have $\Pr_{C'}(d(c,C') \leq \alpha) = \mu(B(c,\alpha))$ for all $c \in \mathcal{C}$, and since $\mathcal{C}$ is an (α,β)-code, $\beta \geq \Pr_{C,C'}(d(C,C') \leq \alpha) = \mathbb{E}_C\left[\mu(B(C,\alpha))\right]$. Then from Markov's inequality we get that $\Pr_C(\mu(B(C,\alpha)) \geq 4\beta) \leq \frac{1}{4}$. Taking a union bound, $\mathcal{E}'$ holds with probability at least $\frac{1}{2}$, thus proving inequality (1).

Then we can apply Lemma B.1 with $p = \frac{1}{2}$ and 4β to get that $I(C; g(C,R)) \geq \frac{1}{2}\log\frac{1}{4\beta} - 1$. Since $M(C,R)$ determines $g(C,R)$, the data processing inequality implies that $\mathrm{I}(C; M(C,R)) \geq \mathrm{I}(C; g(C,R))$, thus completing the proof. □

Proof of Theorem 3.2: Consider random variables $(S,L) \sim \mu\times\nu$ and a randomized one-way protocol for $f(S,L)$ with error probability (with respect to S,L and private randomness) at most $\frac{\alpha}{8}$. Let $\mathsf{A}(s,r_A)$ be the message that Alice sends on this protocol over input s and her private randomness r_A, and let $\mathsf{B}(m,\ell,r_B)$ be the output of Bob when he has input ℓ, private randomness r_B and receives message m from Alice. We want to show $\mathrm{I}(S; \mathsf{A}(S,R_A)) \geq \frac{1}{4}\log\frac{1}{4\beta} - 1$.

For that, define $M(f(s),r_A,r_B) : \mathcal{L} \to \{0,1\}$ by setting $M(f(s),r_A,r_B)(\ell) = \mathsf{B}(\mathsf{A}(s,r_A),\ell,r_B)$ for all $s \in \mathcal{S}$ and $\ell \in \mathcal{L}$. Given the guarantees of the protocol, we have

$$\begin{aligned}&\mathbb{E}_{S\sim\mu,R_A,R_B}[\|M(f(S),R_A,R_B) - f(S)\|_\nu]\\ &= \Pr_{S\sim\mu,L\sim\nu,R_A,R_B}(M(f(S),R_A,R_B)(L) \neq f(S,L)) \leq \frac{\alpha}{8}.\end{aligned}$$

By Markov's inequality, $\Pr_{S\sim\mu,R_A,R_B}\left(\|M(f(S),R_A,R_B) - f(S)\|_\nu \geq \frac{\alpha}{2}\right) \leq \frac{1}{4}$.

Then we can employ Theorem 3.4 with $\mathcal{C}$ set to $\{f(s)\}_{s\in\mathcal{S}}$ to obtain that $\mathrm{I}(f(S); M(f(S),R_A,R_B)) \geq \frac{1}{4}\log\frac{1}{4\beta} - 1$. But the random variable S determines the row $f(S)$ and $(\mathsf{A}(S,R_A),R_B)$ determines the vector $M(f(S),R_A,R_B)$, so by the data processing inequality we get $\mathrm{I}(S; \mathsf{A}(S,R_A),R_B) \geq \frac{1}{4}\log\frac{1}{4\beta} - 1$. Finally, since R_B is independent from S and R_A, we have $\mathrm{I}(S; \mathsf{A}(S,R_A),R_B) = \mathrm{I}(S; \mathsf{A}(S,R_A))$. This concludes the proof of the theorem. □

We show how we can use relaxed codes to recover the lower bounds for k-sparse set disjointness of Dasgupta et al. [6] in Appendix D, and for the indexing problem of Jayram and Woodruff [9] in Appendix E.

4 Characterization via Packing Numbers

We now show how the lower bounds from the previous section lead to our main characterization theorem of the one-way information complexity under product distributions in terms of packing numbers. Given a pseudo-metric space $(\mathcal{X},d)$, its *α-packing number* is the size of the largest set of points in $\mathcal{X}$ with pairwise distances at least α; we denote this by $\mathcal{P}(\mathcal{X},d,\alpha)$. The base of the characterization is a new connection between relaxed codes and packing numbers.

Lemma 4.1. *Consider a pseudo-metric space $(\mathcal{C},d)$ and an $\alpha \in (0,1]$. Then $\mathcal{C}$ is an $\left(\alpha, \frac{1}{\mathcal{P}(\mathcal{C},d,\alpha)}\right)$-code with respect to some distribution μ over $\mathcal{C}$.*

Proof. Let $\mathcal{C}' \subseteq \mathcal{C}$ be a set of size $\mathcal{P}(\mathcal{C},d,\alpha)$ such that distinct points in $\mathcal{C}'$ have distance at least α. Let μ be the uniform distribution on $\mathcal{C}'$. Then $\Pr_{C,C'\sim\mu}(d(C,C') \leq \alpha) = \Pr_{C,C'\sim\mu}(C = C') = \frac{1}{|\mathcal{C}'|} = \frac{1}{\mathcal{P}(\mathcal{C},d,\alpha)}$, and hence $\mathcal{C}$ is an $\left(\alpha, \frac{1}{\mathcal{P}(\mathcal{C},d,\alpha)}\right)$-code with respect to μ. □

Theorem 4.2. *Consider a communication problem $f : \mathcal{S} \times \mathcal{L} \to \{0,1\}$ and let ν be a distribution over $\mathcal{L}$. Let $p_{\alpha,\nu}$ denote the α-packing number of the pseudo-metric space $(\{f(s)\}_{s\in\mathcal{S}}, \|.\|_\nu)$. Then for every $\alpha \in (0,1]$,*

$$\max_\mu \mathsf{IC}(f)^{\rightarrow}_{(\mu\times\nu),\frac{\alpha}{8}} \geq \frac{1}{4}\log\frac{p_{\alpha,\nu}}{4} - 1 \tag{2}$$

$$\max_\mu \mathsf{D}(f)^{\rightarrow}_{(\mu\times\nu),\alpha} \leq \log p_{\alpha,\nu} + 1, \tag{3}$$

where the $\max_\mu$ range over all distributions over $\mathcal{S}$. In particular, letting p^*_α denote the maximum $p_{\alpha,\nu}$ over all ν, we have for $\alpha \in (0, \frac{1}{8}]$

$$\Omega(\log p^*_{8\alpha}) \leq \mathsf{R}(f)^{\rightarrow,\square}_\alpha \leq \log p^*_\alpha + 1. \tag{4}$$

Proof. Inequality (2) follows directly from Theorem 3.2 and Lemma 4.1.

For inequality (3), let $\mathcal{S}' \subseteq \mathcal{S}$ be a set of size $p_{\alpha,\nu}$ such that $\|f(s) - f(s')\|_\nu \geq \alpha$ for all distinct $s, s' \in \mathcal{S}'$. The maximality of $\mathcal{S}'$ implies that the balls $\{B(f(s),\alpha)\}_{s\in\mathcal{S}'}$ cover all of $\{f(s)\}_{s\in\mathcal{S}}$. Then Alice and Bob, on inputs s and ℓ respectively, can do the following: Alice uses $\lceil \log p_{\alpha,\nu} \rceil$ bits to send Bob the index of a point $\psi(s)$ in $\mathcal{S}'$ such that $\|f(s) - f(\psi(s))\|_\nu \leq \alpha$; Bob then outputs $f(\psi(s),\ell)$. For any distribution μ, the distributional error of this protocol with respect to $\mu \times \nu$ is at most α: for any $s \in \mathcal{S}$, $\Pr_{L\sim\nu}(f(\psi(s),L) \neq f(s,L)) = \|f(\psi(s)) - f(s)\|_\nu \leq \alpha$. This concludes the proof of inequality (3).

Inequality (4) follows directly by taking a maximum over ν on inequalities (2) and (3) and using the bound $\mathsf{R}(f)^{\rightarrow,\square}_\alpha \geq \mathsf{IC}(f)^{\rightarrow,\square}_\alpha$. □

Notice that this characterization implies that Theorem 3.2 is tight up to constants (and up to constants in the error rate) given the right distributions μ and ν.

4.1 Relationship with VC Dimension

We recall the characterization of distributional complexity for *constant error* rate α in terms of VC-Dimension given by [10] and [2]. The *VC-dimension* of a subset $\mathcal{C} \subseteq \{0,1\}^n$ is the largest set of indices $I \subseteq [n]$ such that the projection onto I given by $\{(x_i)_{i\in I} : x \in \mathcal{C}\}$ equals the whole of $\{0,1\}^{|I|}$.

Theorem 4.3 ([10, 2])**.** *Consider a communication problem $f : \mathcal{S} \times \mathcal{L} \to \{0,1\}$ and $\alpha \in (0, \frac{1}{4}]$. Then, if VC denotes the VC-dimension of the rows $\{f(s)\}_{s\in\mathcal{S}}$,*

$$(1 - H(\alpha))VC \leq \mathsf{R}(f)^{\rightarrow,\square}_\alpha \leq O\left(VC \cdot \frac{1}{\alpha}\log\frac{1}{\alpha}\right). \tag{5}$$

Notice that, *for constant error* α, this characterizes the distributional complexity up to constant factors. Known bounds on the relationship between VC-dimension and packing numbers allow us to directly recover this characterization from Theorem 4.2. First, we need the dual of packing numbers: Given a pseudo-metric space $(\mathcal{X}, d)$, its *α-covering number* is the smallest number of balls $B(x,\alpha)$ of radius α needed to cover $\mathcal{X}$; we denote this by $\mathcal{N}(\mathcal{X}, d, \alpha)$. It is well-known that packing and covering numbers are closely related: for all $\alpha > 0$,

$$\mathcal{N}(\mathcal{X}, d, \alpha) \leq \mathcal{P}(\mathcal{X}, d, \alpha) \leq \mathcal{N}(\mathcal{X}, d, \alpha/2). \tag{6}$$

We have the following relationships between VC-dimension and packing/covering numbers (for completeness we provide a proof of the first one in the appendix).

Lemma 4.4. *Let $\mathcal{C}$ be a subset of $\{0,1\}^n$ and let VC denote its VC-dimension. Then for every $\alpha \in (0, \frac{1}{2}]$,*

$$\max_{\nu} \log \mathcal{N}(\mathcal{C}, \|.\|_\nu, \alpha) \geq (1 - H(\alpha))VC,$$

where the maximum is taken over all distributions on $[n]$ and $H(\alpha) = \alpha \log \frac{1}{\alpha} + (1-\alpha)\log \frac{1}{1-\alpha}$ denotes the binary entropy.

Lemma 4.5 ([7, 8])**.** *Let $\mathcal{C}$ be a subset of $\{0,1\}^n$ and let VC be its VC-dimension. Then for every distribution ν over $[n]$ and $\alpha \in (0,1]$, we have*

$$\log \mathcal{P}(\mathcal{C}, \|.\|_\nu, \alpha) \leq VC \cdot \log\left(\frac{5}{\alpha} \log \frac{10}{\alpha}\right).$$

Using these two lemmas and inequality (6), we get that for $\alpha \in (0, \frac{1}{4}]$

$$(1 - H(\alpha)) \cdot VC \leq \max_{\nu} \log \mathcal{P}(\mathcal{C}, \|.\|_\nu, \alpha) \leq VC \cdot \log\left(\frac{5}{\alpha} \log \frac{10}{\alpha}\right).$$

Using these bounds on Theorem 4.2 recovers the VC-dimension characterization from Theorem 4.3; in fact, it gives the improved dependence $O(\log \frac{1}{\epsilon})$ on ϵ.

Corollary 4.6. *Consider a communication problem $f : \mathcal{S} \times \mathcal{L} \to \{0,1\}$ and $\alpha \in (0, \frac{1}{16}]$. Then, letting VC denote the VC-dimension of the rows $\{f(s)\}_{s \in \mathcal{S}}$,*

$$(1 - H(8\alpha)) \cdot \Omega(VC) \leq \mathsf{R}(f)_{\alpha}^{\to,[]} \leq O\left(VC \cdot \log \frac{1}{\alpha}\right). \tag{7}$$

In Appendix F we consider the greater-than function to show the difference between the characterizations in terms of VC-dimension and packing numbers.

5 Composition of Communication Problems and Noise Sensitivity

In this section we are interested in compositions of communication problems. More precisely, given a communication problem $f : \mathcal{X} \times \mathcal{Y} \to \{0,1\}$ and a composition function $g : \{0,1\}^k \to \{0,1\}$, we use $g \odot f$ to denote the composition $g(f(x_1, y_1), \dots, f(x_k, y_k))$ (so it is a function mapping $(\mathcal{X} \times \mathcal{Y})^k \to \{0,1\}$). We will used relaxed codes to understand how the composed communication problem $g \odot f$ amplifies the hardness of the base problem f. We will see that the hardness amplification is governed by a generalization of the *noise sensitivity* [19] of g.

Definition 5.1 ((t,γ)-correlation)**.** *Given $\gamma \in [0,1]$, we say that two random variables Z, Z' are* γ-correlated *if $\Pr(Z = Z') \leq \gamma$. Given $t \in [k]$, we say that two random vectors $(Z_1, \dots, Z_k)$ and $(Z'_1, \dots, Z'_k)$ are* (t,γ)-correlated *if there is a subset $I \subseteq [k]$ of size t such that for all $i \in I$, Z_i and Z'_i are γ-correlated.*

Definition 5.2 ((t,γ)-Noise sensitivity)**.** *Consider a function $g : \{0,1\}^k \to \{0,1\}$ and fix $t \in [k]$ and $\gamma \in [0,1]$. Let $\mathfrak{D}$ be a family of distributions over $\{0,1\}^k$ such that there are (t,γ)-correlated random vectors $\boldsymbol{Z}, \boldsymbol{Z}'$ with distributions in $\mathfrak{D}$. Then the* (t,γ)-noise sensitivity *of g with respect to $\mathfrak{D}$ is given by*

$$\mathrm{NS}^t_{\gamma,\mathfrak{D}}(g) \triangleq \min_{\boldsymbol{Z},\boldsymbol{Z}'} \Pr(g(\boldsymbol{Z}) \neq g(\boldsymbol{Z}')),$$

where the minimum is taken over all (t,γ)-correlated random vectors $\boldsymbol{Z}, \boldsymbol{Z}'$ with distributions in $\mathfrak{D}$.

Now we try to give some intuition why noise sensitivity captures how a composition function amplifies the relaxed code of a base function. Consider a communication problem $f : \mathcal{X} \times \mathcal{Y} \to \{0, 1\}$, with a "hard" distribution $\mu \times \nu$, and a composition function $g : \{0, 1\}^k \to \{0, 1\}$. To understand the information complexity of $g \odot f$ under $(\mu \times \nu)^k$, we want to check if it forms an (α, β)-code, which informally means that for "typical" $\boldsymbol{x}, \boldsymbol{x'} \in \mathcal{X}^k$, $\Pr_{\boldsymbol{Y} \sim \nu^k}(g \odot f(\boldsymbol{x}, \boldsymbol{Y}) \neq g \odot f(\boldsymbol{x'}, \boldsymbol{Y})) \geq \alpha$. Expanding the left-hand side shows that it is related to the (t, γ)-sensitivity of g, where the noise level γ is given by $\Pr_{Y \sim \nu}(f(x, Y) = f(x', Y))$, again for "typical" $x, x' \in \mathcal{X}$; this noise level is in turn related to how good a relaxed code the rows $\{f(x)\}_{x \in \mathcal{X}}$ are with respect to μ and $\|.\|_\nu$. Formally:

Theorem 5.3. *Consider a communication problem $f : \mathcal{X} \times \mathcal{Y} \to \{0, 1\}$. Let μ and ν be distributions over $\mathcal{X}$ and $\mathcal{Y}$, respectively, such that $\{f(x)\}_{x \in \mathcal{X}}$ forms an (α, β)-code with respect to μ and the distance $\|.\|_\nu$. Let $\mathfrak{D}$ be the set of distributions of the random vectors $(f(x_1, Y_1), \ldots, f(x_k, Y_k))$ with $x_1, \ldots, x_k \in \mathcal{X}$, where $Y_1, \ldots, Y_k$ are independently sampled from ν. Consider a function $g : \{0, 1\}^k \to \{0, 1\}$. Then for $w \in (0, 1 - \beta]$, the rows $\{g \odot f(\boldsymbol{x})\}_{\boldsymbol{x} \in \mathcal{X}^k}$ form an (α_w, β_w)-code with respect to μ^k and the distance $\|.\|_{\nu^k}$, where*

$$\alpha_w = \mathrm{NS}^{k(1-\beta-w)}_{1-\alpha, \mathfrak{D}}(g)$$
$$\beta_w = \left(\frac{e^w}{(1 + w/\beta)^{\beta + w}}\right)^k \leq \left(\frac{e\beta}{w}\right)^{wk}.$$

Proof. It suffices to show that for a $1 - \beta_w$ fraction of the independent random vectors $\boldsymbol{X}, \boldsymbol{X'} \sim \mu^k$, we have $\Pr_{\boldsymbol{Y} \sim \nu^k}(g \odot f(\boldsymbol{X}, \boldsymbol{Y}) \neq g \odot f(\boldsymbol{X'}, \boldsymbol{Y})) \geq \mathrm{NS}^{k(1-\beta-w)}_{1-\alpha, \mathfrak{D}}(g)$.

Let $\Omega \subseteq \mathcal{X}^2$ be the set of pairs (x, x') such that $\|f(x) - f(x')\|_\nu > \alpha$, namely $\Pr_{Y \sim \nu}\big(f(x, Y) \neq f(x', Y)\big) > \alpha$. For two vectors $\boldsymbol{x}, \boldsymbol{x'}$ in $\mathcal{X}^k$, let $\#(\boldsymbol{x}, \boldsymbol{x'})$ denote the number of coordinates i such that (x_i, x'_i) belongs to Ω.

Fix any two $\boldsymbol{x}, \boldsymbol{x'}$ in $\mathcal{X}^k$. For $\boldsymbol{Y} = (Y_1, \ldots, Y_k)$ sampled from ν^k, define $Z_i = f(x_i, Y_i)$ and $Z'_i = f(x'_i, Y_i)$. Then by definition of Ω, $\boldsymbol{x}$ and $\boldsymbol{x'}$, we have that the vectors $\boldsymbol{Z} = (Z_1, \ldots, Z_k)$ and $\boldsymbol{Z'} = (Z'_1, \ldots, Z'_k)$ are $(\#(\boldsymbol{x}, \boldsymbol{x'}), 1 - \alpha)$-correlated with distributions in $\mathfrak{D}$. Then by the definition of (t, γ)-noise stability,

$$\Pr_{\boldsymbol{Y}}\big(g \odot f(\boldsymbol{x}, \boldsymbol{Y}) \neq g \odot f(\boldsymbol{x'}, \boldsymbol{Y})\big) = \Pr_{\boldsymbol{Z}}\big(g(\boldsymbol{Z}) \neq g(\boldsymbol{Z'})\big) \geq \mathrm{NS}^{\#(\boldsymbol{x}, \boldsymbol{x'})}_{1-\alpha, \mathfrak{D}}(g).$$

To show that $\Pr\big(\#(\boldsymbol{X}, \boldsymbol{X'}) \geq k(1 - \beta - w)\big)$ is at least $1 - \beta_w$, we observe the following. Since f forms an (α, β)-code, we know that $\Pr\big((\boldsymbol{X}, \boldsymbol{X'}) \in \Omega\big) > 1 - \beta$, and thus $\mathbb{E}[\#(\boldsymbol{X}, \boldsymbol{X'})] \geq k(1 - \beta)$. By a multiplicative Chernoff bound (Appendix A), we have that the event $k - \#(\boldsymbol{X}, \boldsymbol{X'}) > (1 + w/\beta)k\beta$ happens with probability at most $(\frac{e^w}{(1+w/\beta)^{\beta+w}})^k = \beta_w$, and hence with probability at least $1 - \beta_w$ we have $\#(\boldsymbol{X}, \boldsymbol{X'}) \geq k(1 - \beta - w)$.

To conclude the proof, we show that $\beta_w \leq (e\beta/w)^{wk}$. First, by reducing the denominator we have $\beta_w \leq \left(\frac{e}{1+w/\beta}\right)^{wk}$. But this quantity is at most $\left(\frac{e\beta}{w}\right)^{wk}$, which can be shown using concavity of the map $\beta \mapsto \frac{e}{1+w/\beta}$, and the fact that its derivative at 0 is $\frac{e}{w}$. This concludes the proof. □

Together with the lower bound of Theorem 3.2 based on relaxed codes, this amplification theorem gives a powerful tool for constructing lower bounds.

5.1 Example: Stronger Direct Sum for XOR

Let $\mathsf{XOR}^k : \{0,1\}^k \to \{0,1\}$ denote the k-ary XOR function, namely it maps $(z_1, \ldots, z_k) \mapsto \sum_i z_i$ mod 2. The following lower bound on the (t, γ)-Noise sensitivity of XOR^k is proved in Appendix H.

Lemma 5.4. *Let* $\mathfrak{prod}$ *denote the set of all product distributions over* $\{0,1\}^k$*. Then* $\mathrm{NS}^t_{1-\alpha,\mathfrak{prod}}(\mathsf{XOR}^k) \geq \frac{1}{2} - \frac{1}{2}(1-2\alpha)^t$.

Theorem 5.3 and the above lemma (together with Lemma 4.1 and Theorem 3.2) give a stronger direct sum theorem for XOR (notice the error probability $\frac{1}{k}$ on the right-hand side); details are presented in Appendix H.

Corollary 5.5. *For any communication problem* $f : \mathcal{X} \times \mathcal{Y} \to \{0,1\}$,

$$\mathsf{IC}(\mathsf{XOR}^k \odot f)^{\to,\square}_{\frac{1}{4}} \geq \frac{k}{8} \cdot \left(\mathsf{R}(f)^{\to,\square}_{\frac{1}{k}} - 8 \right).$$

6 Streaming Applications

We have the following tight bounds for streaming.

6.0.1 Approximate Closest Pair

This problem is described as follows: Alice has n vectors $\mathbf{v}^1, \mathbf{v}^2, \ldots, \mathbf{v}^n \in [\pm M]^d$, Bob has n vectors $\mathbf{u}^1, \mathbf{u}^2, \ldots, \mathbf{u}^n \in [\pm M]^d$ and a threshold value θ, and his goal is to distinguish (with prob. $1-\delta$) the cases:

1. For all $i \in [n]$ it holds that $\|\mathbf{u}^i - \mathbf{v}^i\|_p^p \geq (1+\epsilon)\theta$.
2. There exists i such that $\|\mathbf{u}^i - \mathbf{v}^i\|_p^p \leq (1-\epsilon)\theta$.

Let $\ell_p(n, d, M, \epsilon, \theta)$ denote this problem.

Theorem 6.1. *Assume n is at least a sufficiently large constant and ϵ is at most a sufficiently small constant. Assume there is a constant $\gamma > 0$ such that $d^{1-\gamma} \geq \frac{1}{\epsilon^2} \log \frac{n}{\delta}$. Then* $R^{\to}_{\delta}(\ell_p(n, d, M, \epsilon, \theta)) \geq \Omega\left(\frac{n}{\epsilon^2} \log \frac{n}{\delta} (\log d + \log M))\right)$ *for* $p \in \{1, 2\}$.

6.0.2 Approximating Largest Entry in Matrix Product by Sketching.

Given a matrix A, let A_i denote its i-th row and use A^j to denote its j-th column.

Theorem 6.2. *Assume n is a sufficiently large constant and ϵ is at most a sufficiently small constant. Assume there is a constant $\gamma > 0$ such that $n^{1-\gamma} \geq \frac{1}{\epsilon^2} \log \frac{n}{\delta}$. Let S be an $n \times d$ matrix that has an estimation procedure f_θ satisfying: for every pair of matrices $A, B \in [\pm M]^{n \times n}$, with probability at least $1 - \delta$*

1. *$f_\theta(AS, B) = 1$ if $(AB)_{i,j} \geq (1+\epsilon)\theta$ for some $i, j \in [n]$.*
2. *$f_\theta(AS, B) = 1$ if $(AB)_{i,j} \leq \theta$ for all $i, j \in [n]$.*

Then the number of bits to specify AS is at least $\Omega(n \frac{1}{\epsilon^2} \log \frac{n}{\delta} (\log n + \log M))$.

References

[1] Ablayev, F.M.: Lower bounds for one-way probabilistic communication complexity and their application to space complexity. Theor. Comput. Sci. 157(2), 139–159 (1996)

[2] Bar-Yossef, Z., Jayram, T.S., Kumar, R., Sivakumar, D.: An information statistics approach to data stream and communication complexity. J. Comput. Syst. Sci. 68(4), 702–732 (2004)

[3] Beals, R., Buhrman, H., Cleve, R., Mosca, M., de Wolf, R.: Quantum lower bounds by polynomials. J. ACM 48(4), 778–797 (2001)

[4] Buhrman, H., García-Soriano, D., Matsliah, A., de Wolf, R.: The non-adaptive query complexity of testing k-parities. Chicago J. Theor. Comput. Sci. (2013)

[5] Cover, T.M., Thomas, J.A.: Elements of information theory (2. ed.). Wiley (2006)

[6] Dasgupta, A., Kumar, R., Sivakumar, D.: Sparse and lopsided set disjointness via information theory. In: RANDOM (2012)

[7] Dudley, R.M.: Central limit theorems for empirical measures. The Annals of Probability 6(6), 899–929 (12 1978)

[8] Haussler, D.: Decision theoretic generalizations of the PAC model for neural net and other learning applications. Inform. Comput. 100(1), 78 – 150 (1992)

[9] Jayram, T.S., Woodruff, D.P.: Optimal bounds for johnson-lindenstrauss transforms and streaming problems with sub-constant error. In: SODA (2011)

[10] Kremer, I., Nisan, N., Ron, D.: On randomized one-round communication complexity. Computational Complexity pp. 21–49 (1999)

[11] Lee, T., Shraibman, A.: Lower bounds in communication complexity. Foundations and Trends in Theoretical Computer Science 3(4), 263–399 (2009)

[12] Lee, T., Zhang, S.: Composition theorem in communication complexity. In: ICALP (2010)

[13] Matousek, J., Vondrak, J.: The Probabilistic Method (2008), manuscript

[14] Molinaro, M., Woodruff, D.P., Yaroslavtsev, G.: Beating the direct sum theorem in communication complexity with implications for sketching. In: SODA (2013)

[15] Motwani, R., Raghavan, P.: Randomized Algorithms. Cambridge University Press, New York, NY, USA (1995)

[16] Muthukrishnan, S.: Data streams: algorithms and applications. Found. Trends Theor. Comput. Sci. 1(2), 117–236 (Aug 2005)

[17] Newman, I., Szegedy, M.: Public vs. private coin flips in one round communication games (extended abstract). In: STOC (1996)

[18] Nisan, N., Szegedy, M.: On the degree of boolean functions as real polynomials. Computational Complexity 4, 301–313 (1994)

[19] O'Donnell, R.: Analysis of Boolean Functions. Cambridge University Press (2014)

[20] Papadimitriou, C.H., Sipser, M.: Communication complexity. J. Comput. Syst. Sci. 28(2), 260–269 (1984)

[21] Saglam, M., Tardos, G.: On the communication complexity of sparse set disjointness and exists-equal problems. In: FOCS (2013)

[22] Sherstov, A.: The pattern matrix method. SIAM J. Comput. 40(6), 1969–2000 (2011)

[23] Woodruff, D.P.: Efficient and Private Distance Approximation in the Communication and Streaming Models. Phd Thesis, MIT (2007)

[24] Woodruff, D.P.: The average-case complexity of counting distinct elements. In: ICDT (2009)

[25] Yao, A.C.: Lower bounds by probabilistic arguments (extended abstract). In: FOCS (1983)

Appendix

A Probabilistic Inequalities

Theorem A.1 (Theorem 4.1 of [15])**.** *Let $X_1, X_2, \ldots, X_n$ be independent random variables over $\{0,1\}$ and let $\mu = \mathbb{E}[\sum_i X_i]$. Then for any $\delta > 0$,*

$$\Pr\left(\sum_i X_i > (1+\delta)\mu\right) < \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu} \leq \left(\frac{e}{1+\delta}\right)^{(1+\delta)\mu}.$$

Theorem A.2 (Proposition 7.3.2 of [13])**.** *Let $X_1, X_2, \ldots, X_n$ be independent random variables uniformly distributed in $\{0,1\}$ and let $X = \sum_{i=1}^{n} X_i$. Then for any integer $t \in [0, \frac{n}{8}]$,*

$$\Pr\left(X \geq \left\lfloor\frac{n}{2}\right\rfloor + t\right) \geq \frac{1}{15} e^{-16t^2/n}.$$

B List-Decoding Fano's Inequality

We start with the following more general but weaker list-decoding Fano's inequality.

Lemma B.1. *Consider finite sets $\mathcal{X}, \mathcal{Y}$ and a (predictor) function $g : \mathcal{X} \to 2^{\mathcal{Y}}$. Let (X, Y) be a random variable over $\mathcal{X} \times \mathcal{Y}$ with arbitrary distribution. If for some $k \leq |\mathcal{X}|$ we have $\Pr(Y \in g(X) \text{ and } |g(X)| \leq k) \geq p$, then $H(Y \mid g(X)) \leq p \log k + (1-p)\log|\mathcal{Y}| + 1$.*

Proof. Let $\mathcal{E}$ be the event $\{Y \in g(X) \text{ and } |g(X)| \leq k\}$. Consider a set $U \subseteq \mathcal{Y}$ such that $\Pr(g(X) = U, \mathcal{E}) > 0$; this implies that U has size at most k and that the random variable Y conditioned on $\{g(X) = U, \mathcal{E}\}$ is supported over U, and hence $H(Y \mid g(X) = U, \mathcal{E}) \leq \log k$.

Then letting $\mathbf{1}_{\mathcal{E}}$ denote the indicator random variable of $\mathcal{E}$, we have

$$\begin{aligned}
&H(Y \mid g(X), \mathbf{1}_{\mathcal{E}}) \\
&= \sum_{U \subseteq \mathcal{X}} \Big(H(Y \mid g(X) = U, \mathcal{E}) \Pr(g(X) = U, \mathcal{E}) + H(Y \mid g(X) = U, \bar{\mathcal{E}}) \Pr(g(X) = U, \bar{\mathcal{E}}) \Big) \\
&\leq \Pr(\mathcal{E}) \log k + (1 - \Pr(\mathcal{E})) \log|\mathcal{Y}| \leq p \log k + (1-p) \log|\mathcal{Y}|.
\end{aligned}$$

The result then follows by employing the chain rule and non-negativity of entropy: $H(Y \mid g(X)) \leq H(Y, \mathbf{1}_{\mathcal{E}} \mid g(X)) \leq H(Y \mid g(X), \mathbf{1}_{\mathcal{E}}) + H(\mathbf{1}_{\mathcal{E}} \mid g(X)) \leq H(Y \mid g(X), \mathbf{1}_{\mathcal{E}}) + 1$. □

Proof of Lemma 3.3: Since Fano's inequality is tighter under the uniform distribution, we modify the space $(\mathcal{X}, \mu)$ into a space $(\tilde{\mathcal{X}}, \tilde{\mu})$ where $\tilde{\mu}$ is a uniform distribution, and then apply the above lemma to the latter.

More precisely, assume that $\mu(x)$ is rational for all $x \in \mathcal{X}$ (otherwise one can approximate μ by a rational probability distribution within total variation distance $\epsilon > 0$ and then take the limit as $\epsilon \to 0$ and the following proof will go through unchanged). We have $\mu(x) = \frac{p(x)}{q}$ for all $x \in \mathcal{X}$ with $p(x)$ and q integers (notice the common q). Then construct $(\tilde{X}, \tilde{\mu})$ as follows: for each $x \in \mathcal{X}$, add to $\tilde{X}$ $p(x)$ distinct copies $x_1, x_2, \ldots, x_{p(x)}$ of x, and set $\tilde{\mu}(x_i) = \frac{1}{q}$. It is convenient to have the map $\phi : \tilde{\mathcal{X}} \to \mathcal{X}$ that maps each element in $\tilde{\mathcal{X}}$ to its source in $\mathcal{X}$ defined by $\phi(x_i) = x$. Notice that for every subset $E \subseteq \mathcal{X}$, $\tilde{\mu}(\phi^{-1}(E)) = \mu(E)$ and hence $\phi(\tilde{X}) \sim \mu$ for $\tilde{X} \sim \tilde{\mu}$. In addition, define the predictor $\tilde{g} : \tilde{\mathcal{X}} \times \mathcal{R} \to 2^{\tilde{\mathcal{X}}}$ by extending g in the natural way: $\tilde{g}(\tilde{x}, r) = \phi^{-1}(g(\phi(\tilde{x}), r))$ for all $\tilde{x} \in \tilde{\mathcal{X}}$.

Now we want to apply Lemma B.1 to $\tilde{X} \times \mathcal{R}$, $\tilde{\mu}$, and $\tilde{g}$. For that, we claim that

$$\Pr_{\substack{\tilde{X} \sim \tilde{\mu} \\ R \sim \lambda}} \left(\tilde{X} \in \tilde{g}(\tilde{X}, R) \text{ and } |\tilde{g}(\tilde{X}, R)| \leq \beta |\tilde{\mathcal{X}}| \right) \geq p. \tag{8}$$

To see this, using the above observation about our construction we have that the event $\{\tilde{X} \in \tilde{g}(\tilde{X}, R) \text{ and } \tilde{\mu}(\tilde{g}(\tilde{X}), R) \leq \beta\}$ is the same as the event $\{\phi(\tilde{X}) \in g(\phi(\tilde{X}), R) \text{ and } \mu(g(\phi(\tilde{X}), R)) \leq \beta\}$. Since $\phi(\tilde{X}) \sim \mu$, we have

$$\Pr_{\substack{\tilde{X} \sim \tilde{\mu} \\ R \sim \lambda}} \left(\tilde{X} \in \tilde{g}(\tilde{X}, R) \text{ and } \tilde{\mu}(\tilde{g}(\tilde{X}), R) \leq \beta \right) = \Pr_{\substack{X \sim \mu \\ R \sim \lambda}} \left(X \in g(X, R) \text{ and } \mu(g(X, R)) \leq \beta \right) \geq p,$$

where the last inequality follows by the assumption on g. To recover (8) from this inequality simply notice that, since $\tilde{\mu}$ is the uniform distribution over $\tilde{\mathcal{X}}$, $\tilde{\mu}(\tilde{g}(\tilde{X}, R)) \leq \beta$ is equivalent to $|\tilde{g}(\tilde{X}, R)| \leq \beta |\tilde{\mathcal{X}}|$.

Then from Lemma B.1 we get that $H(\tilde{X} \mid \tilde{g}(\tilde{X}, R)) \leq p \log \beta |\tilde{\mathcal{X}}| + (1-p) \log |\tilde{\mathcal{X}}| + 1 = \log |\tilde{\mathcal{X}}| - p \log \frac{1}{\beta} + 1$. Since $H(\tilde{X}) = \log |\tilde{\mathcal{X}}|$, we get that $\mathrm{I}(\tilde{X}; \tilde{g}(\tilde{X}, R)) = H(\tilde{X}) - H(\tilde{X} \mid \tilde{g}(\tilde{X}, R)) \geq p \log \frac{1}{\beta} - 1$.

To conclude the proof, we claim that $\mathrm{I}(X; g(X, R)) \geq \mathrm{I}(\tilde{X}; \tilde{g}(\tilde{X}, R))$. For that, define the random variable (X, I) as follows: X is distributed according to μ, and I is uniform in $\{1, \ldots, p(X)\}$. So (X, I) can be thought of as a random element in $\tilde{\mathcal{X}}$; more precisely, the function $\psi : \mathcal{X} \times \mathbb{N} \to \tilde{\mathcal{X}}$ which maps $\psi(x, i)$ into the i-th copy of x in $\tilde{\mathcal{X}}$ satisfies $\psi(X, I) \sim \tilde{\mu}$. Then

$$\begin{aligned} \mathrm{I}(\tilde{X}; \tilde{g}(\tilde{X}, R)) = \mathrm{I}\left(\psi(X, I); \phi^{-1} g(\phi(\psi(X, I)), R)\right) &= \mathrm{I}\left(\psi(X, I); \phi^{-1}(g(X, R))\right) \\ &\leq \mathrm{I}(X, I; g(X, R)), \end{aligned}$$

where the last inequality follows from the data processing inequality. But by the chain rule for mutual information and by the independence of I and $g(X, R)$ conditioned on X, we get $\mathrm{I}(X, I; g(X, R)) = \mathrm{I}(X; g(X, R)) + \mathrm{I}(I; g(X, R) \mid X) = \mathrm{I}(X; g(X, R))$. This concludes the proof of the lemma. □

C Proof of Lemma 4.4

Let $I \subseteq [n]$ be a subset of size VC such that the projection $\{(x_i)_{i \in I} : x \in \mathcal{C}\}$ equals $\{0,1\}^{VC}$. Let ν be the uniform distribution over I. Then (after we identify points with distance 0) the space $(\mathcal{C}, \|.\|_\nu)$ is isometric to $(\{0,1\}^{VC}, \|.\|_{uni})$, where $\|x\|_{uni} = \frac{1}{VC} \sum_{i \in [VC]} x_i$ is the normalized Hamming distance; thus, their α-covering numbers are the same. One then just needs to lower bound the α-covering number of $(\{0,1\}^{VC}, \|.\|_{uni})$ by $2^{(1-H(\alpha))VC}$; this bounds follows from the fact that every ball in this space with radius α has at most $2^{VC \cdot H(\alpha)}$ points and the whole space has 2^{VC} points, hence at least $2^{(1-H(\alpha))VC}$ balls are needed to cover the whole space.

D Example: Sparse Set Disjointness

In this problem, the inputs for Alice and Bob are k-subsets of $[n]$ and we have the disjointness function $DISJ : \binom{[n]}{k} \times \binom{[n]}{k} \to \{0,1\}$ given by $DISJ(x,y) = 1$ iff x and y are disjoint sets. Dasgupta et al. exhibited the tight lower bound $\mathsf{D}^{\rightarrow}_{cst}(DISJ) = \Omega(k \log k)$, for $k \leq \sqrt{n}$ and a small enough constant cst (they also provide a matching upper bound).

To recover this bound, we start with the same construction used by Dasgupta et al. (see Section 3.2 of [6] for their existence). Let $\mathcal{X}$ and $\mathcal{Y}$ be subsets of $\binom{[n]}{k}$ with the following properties: 1) given two different $x \neq x' \in \mathcal{X}$, the rows $DISJ(x)$ and $DISJ(x')$ are distinct; 2) $|\mathcal{X}| \geq 2^{a \cdot k \log k}$ for a constant a independent of k; 3) $|\mathcal{Y}| \leq b \cdot k \log k$ for a constant b independent of k. It will be convenient to define $DISJ'$ as the restriction of $DISJ$ to the inputs $\mathcal{X} \times \mathcal{Y}$; it suffices to show $\mathsf{D}^{\rightarrow}_{cst}(DISJ') = \Omega(k \log k)$ for some constant cst.

For that, consider the uniform distribution $\mu \times \nu$ over $\mathcal{X} \times \mathcal{Y}$, so we have the distance $\|DISJ'(x) - DISJ'(x')\|_\nu = \|DISJ'(x) - DISJ'(x')\|_0 / |\mathcal{Y}|$. Now consider a row $DISJ'(x)$; since the rows of $DISJ'$ belong to $\{0,1\}^{|\mathcal{Y}|}$, standard bounds give that there are at most $2^{|\mathcal{Y}| \cdot H(cst)}$ rows $DISJ'(x')$ with $\|DISJ'(x) - DISJ'(x')\|_\nu \leq cst$, where $H(\alpha) = \alpha \log \frac{1}{\alpha} + (1-\alpha) \log \frac{1}{1-\alpha}$ denotes the binary entropy. Thus, for every $x \in \mathcal{X}$ we have $\Pr_{X' \sim \mu}(\|DISJ'(x) - DISJ'(X')\|_\nu \leq cst) \leq 2^{|\mathcal{Y}| \cdot H(cst)} / |\mathcal{X}| \leq 2^{(b \cdot H(cst) - a) k \log k}$; this implies $\Pr_{X,X' \sim \mu}(\|DISJ'(X) - DISJ'(X')\|_\nu \leq cst) \leq 2^{(b \cdot H(cst) - a) k \log k}$.

Setting cst a small enough constant we can make $b \cdot H(cst) \leq a/2$ so that $\Pr_{X,X' \sim \mu}(\|DISJ'(X) - DISJ'(X')\|_\nu \leq cst) \leq 2^{-(a/2) k \log k}$, and thus the rows of $DISJ'$ form a $(cst, 2^{-(a/2) k \log k})$-code. Theorem 3.2 then gives the desired lower bound $\mathsf{D}(DISJ')^{\rightarrow}_{cst/8} \geq \mathsf{IC}(DISJ')^{\rightarrow}_{(\mu \times \nu), cst/8} \geq \Omega(k \log k)$.

E Example: Indexing

We consider the *indexing problem* $\mathsf{ind}_{k,m}$, whose instance is given as follows: Alice has numbers $s_1, \ldots, s_m \in [k]$ and Bob has numbers $\ell_1, \ldots \ell_m$ and an index $j \in [m]$; the function $\mathsf{ind}_{k,m} : [k]^m \times ([k]^m \times [m]) \to \{0,1\}$ takes value 1 iff the input satisfies $s_j = \ell_j$; the goal is to compute $\mathsf{ind}_{k,m}$ over Alice's and Bob's inputs. To simplify the notation, we use $\boldsymbol{s} = (s_1, \ldots, s_m)$ and $\boldsymbol{\ell} = (\ell_1, \ldots, \ell_m)$.

We recover the tight lower bound on the indexing problem obtained in [9] using relaxed codes.

Theorem E.1. *Consider the indexing problem $\mathsf{ind}_{k,m}$ and let μ be the uniform distribution over Alice's inputs and let ν be the uniform distribution over Bob's inputs. Then the rows $\{\mathsf{ind}_{k,m}(\boldsymbol{s})\}_{\boldsymbol{s} \in [k]^m}$ of the communication matrix form a $\left(\frac{1}{k}, (\frac{k}{2e})^{-m/2}\right)$-code with respect to μ and $\|.\|_\nu$. In particular,*

$$\mathsf{IC}(\mathsf{ind}_{k,m})^{\rightarrow,[]}_{\frac{1}{k}} \geq \Omega(m \log k).$$

Proof. To simplify the notation, we use ind instead of $\mathsf{ind}_{k,m}$. To prove that the rows $\{\mathsf{ind}(\boldsymbol{s})\}_{\boldsymbol{s} \in [k]^m}$ form a $\left(\frac{1}{k}, (\frac{k}{e})^{-m/2}\right)$-code, start by taking any inputs $\boldsymbol{s}, \boldsymbol{s}' \in [k]^m$ with at least $m/2$ differing indices $s_i \neq s'_i$. We claim the lower bound $\|\mathsf{ind}(\boldsymbol{s}) - \mathsf{ind}(\boldsymbol{s}')\|_\nu \geq \frac{1}{k}$ on the distance between these rows. To see that, let $\Delta \subseteq [m]$ be the set of indices i where $s_i \neq s'_i$. Then

$$\begin{aligned} \|\mathsf{ind}(\boldsymbol{s}) - \mathsf{ind}(\boldsymbol{s}')\|_\nu &= \Pr_{(\boldsymbol{L},J) \sim \nu} \big(\mathsf{ind}(\boldsymbol{s}, (\boldsymbol{L}, J)) \neq \mathsf{ind}(\boldsymbol{s}', (\boldsymbol{L}, J))\big) \\ &\geq \mathbb{E}_J \big[\Pr(\mathsf{ind}(\boldsymbol{s}, (\boldsymbol{L}, J)) \neq \mathsf{ind}(\boldsymbol{s}', (\boldsymbol{L}, J)) \mid J \in \Delta)\big] \Pr(J \in \Delta). \end{aligned} \quad (9)$$

But for every $j \in \Delta$, it is easy to check that

$$\Pr(\mathsf{ind}(\boldsymbol{s}, (\mathbf{L}, j)) \neq \mathsf{ind}(\boldsymbol{s}', (\mathbf{L}, j)) = \frac{2}{k}.$$

Taking the average over all $j \in \Delta$ and using the fact that $\Pr(J \in \Delta) = 1/2$, equation (9) then gives that $\|\mathsf{ind}(\boldsymbol{s}) - \mathsf{ind}(\boldsymbol{s}')\|_\nu \geq \frac{1}{k}$.

Now consider independent $\mathbf{S}, \mathbf{S}'$ uniformly distributed in $[k]^m$. We claim that

$$\Pr\left([\# \text{ indices } i \text{ such that } S_i \neq S_i'] \leq \frac{m}{2}\right) \leq \left(\frac{2e}{k}\right)^{m/2}.$$

Due to the product structure in $[k]^m$, notice that the number of indices i such that S_i is *equal* to S_i' is a binomially distributed random variable with m trials and success probability $\frac{1}{k}$; the claim then follows from applying the multiplicative Chernoff bound from Appendix A using $1 + \delta = \frac{k}{2}$.

Putting these claims together gives that $\{\mathsf{ind}(\boldsymbol{s})\}_{\boldsymbol{s} \in [k]^m}$ forms a $\left(\frac{1}{k}, (\frac{k}{2e})^{-m/2}\right)$-code. The lower bound on $\mathsf{IC}(\mathsf{ind})^{\rightarrow,[]}_{\frac{1}{k}}$ then follows from Theorem 3.2, thus concluding the proof. □

F Example: Greater-than Function

The *greater-than* function $GT : [n] \times [n] \to \{0,1\}$ is given by $GT(x,y) = 1$ iff $x > y$. It is easy to see that the VC-dimension of the rows of GT is equal to 1, so Theorem 4.3 gives the bounds $\Omega(1) \leq \mathsf{R}^{\rightarrow,[]}_{\alpha}(GT) \leq O(\frac{1}{\alpha} \log \frac{1}{\alpha})$. On the other hand, the characterization based on packing numbers gives the right bound $\mathsf{R}^{\rightarrow,[]}_{\alpha}(GT) = \Theta(\log \frac{1}{\alpha})$ for $\alpha \in [\frac{1}{n}, \frac{1}{3}]$.

To see this, consider a distribution ν over $[n]$. Notice that the rows $GT(x)$ and $GT(x')$ (for $x < x'$) have distance $\|GT(x) - GT(x')\|_\nu = \nu((x, x'])$. Then the α-packing number $p_{\alpha,\nu}$ of the rows of GT is at most $O(\frac{1}{\alpha})$: given rows $GT(x_1), \ldots, GT(x_k)$ (with $x_1 < \ldots < x_k$) with pairwise distances at least α, we have $\nu((x_i, x_{i+1}]) \geq \alpha$ for all i and $\sum_{i=1}^{k-1} \nu((x_i, x_{i+1}]) \leq 1$, thus giving $k \leq \frac{1}{\alpha} + 1$. Moreover, for ν being the uniform distribution we have α-packing number $p_{\alpha,\nu} = \Omega(\frac{1}{\alpha})$ (for $\alpha \geq \frac{1}{n}$): just notice that the rows $GT(1), GT(\lceil n\alpha \rceil + 1), GT(2\lceil n\alpha \rceil + 1), \ldots$ have pairwise distances at least α. Theorem 4.2 then gives the desired bound $\mathsf{R}^{\rightarrow,[]}_{\alpha}(GT) = \Theta(\log \frac{1}{\alpha})$.

G Example: Gap-Hamming Problem

We now show how the noise sensitivity approach recovers in a natural way the result from [23] that the Gap Hamming Problem is hard even with respect to the uniform distribution.

In this problem we have the (partial) function $GH' : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$ given by

$$GH'(x,y) = \begin{cases} 1, & \text{if } \|x - y\|_0 \geq \frac{n}{2} + \sqrt{n} \\ 0, & \text{if } \|x - y\|_0 \leq \frac{n}{2} - \sqrt{n} \end{cases}$$

Known lower bounds show that for a small constant error cst, $D(GH')^{\rightarrow}_{cst} = \Omega(n)$.

Consider the extension GH of the partial function GH' given by $GH(x,y) = 1$ if $\|x - y\|_0 > \frac{n}{2}$ and $GH(x,y) = 0$ if $\|x - y\|_0 \leq \frac{n}{2}$. Let $\bar{\mu}$ be the uniform distribution over $\{0,1\}^n$. It is easy to see that a lower bound $\mathsf{D}(GH)^{\rightarrow,[]}_{\bar{\mu}^2, \frac{1}{10}} = \Omega(n)$ under the uniform distributon implies the above lower bound $D(GH')^{\to}_{cst} = \Omega(n)$, since an input (X,Y) sampled uniformly from $\{0,1\}^n \times \{0,1\}^n$ has only constant probability of having $\frac{n}{2} - \sqrt{n} < \|X - Y\|_0 \leq \frac{n}{2} + \sqrt{n}$ (see Section 4.4 of [23] for more details).

Indeed, Woodruff showed this lower bound on GH under the uniform distribution, which we now recover via noise sensitivity. For that, notice that GH can be expressed as the composition $MAJ \odot NEQ$ of the not-equal function $NEQ : \{0,1\} \times \{0,1\} \to \{0,1\}$ given by $NEQ(x,y) = 1$ iff $x \neq y$, and the majority function $MAJ : \{0,1\}^n \to \{0,1\}$ given by $MAJ(z_1, \ldots, z_n) = 1$ iff $\sum_{i=1}^n z_i > \frac{n}{2}$.

To lower bound GH, first notice that the rows of the function NEQ form a $(\frac{1}{3}, \frac{1}{2})$-code with respect to the uniform distribution. Then employing Theorems 5.3 (with $w = 1/6$) and 3.2 we get $\mathsf{IC}(MAJ \circ NEQ)^{\to}_{\bar{\mu}^2, \alpha_w/8} \geq \Omega(n)$, where $\alpha_w = \mathrm{NS}^{n/3}_{\frac{2}{3},\bar{\mu}}(MAJ)$. The following lemma then gives the desired lower bound; the proof is similar to the lower bound for the regular noise sensitivity of the majority function (see [19]).

Lemma G.1. *For $n \geq 225^2$ we have $\mathrm{NS}^{n/3}_{\frac{2}{3},\bar{\mu}}(MAJ) \geq e^{-150}$.*

Proof. Let $(Z_1, \ldots, Z_n)$ and $(Z'_1, \ldots, Z'_n)$ be two $(\frac{n}{3}, \frac{2}{3})$-correlated vectors, each distributed uniformly in $\{0,1\}^n$; it suffices to show that $\Pr(MAJ(Z_1, \ldots, Z_n) = 0 \wedge MAJ(Z'_1, \ldots, Z'_n) = 1) = \Pr(\sum_{i=1}^n Z_i \leq \frac{n}{2} \wedge \sum_{i=1}^n Z'_i > \frac{n}{2})$ is at least $\frac{1}{1000}$.

Let I be an $(n/3)$-subset of $[n]$ such that for all $i \in I$, Z_i and Z'_i are $\frac{2}{3}$-correlated. We first control the indices outside I: let $\mathcal{E}_{out}$ be the event that both sums $\sum_{i \notin I} Z_i$ and $\sum_{i \notin I} Z'_i$ lie in $[\frac{n}{3} - \sqrt{n}, \frac{n}{3} + \sqrt{n}]$. Using Chebychev's inequality and a union bound gives that $\mathcal{E}_{out}$ holds with probability at least $1 - \frac{1}{3}$.

To control the indices in I, consider the random set $S = \{i \in I : Z_i \neq Z'_i\}$. Given a subset $s \subseteq I$, let $\mathcal{E}_s$ be the event that $\sum_{i \in I \setminus s} Z_i \in \frac{|I-s|}{2} \pm \sqrt{n}$ and $\sum_{i \in s}(1 - Z_i) \geq \frac{|s|}{2} + 3\sqrt{n}$. Notice that whenever the events $\mathcal{E}_{out}$ and $\mathcal{E}_S$ hold we have:

$$\sum_{i \in [n]} Z_i = \sum_{i \notin I} Z_i + \sum_{i \in I \setminus S} Z_i + \sum_{i \in S} Z_i \leq \frac{n}{3} + \frac{|I - S|}{2} + \frac{|S|}{2} - \sqrt{n} = \frac{n}{2} - \sqrt{n}, \qquad \text{and}$$
$$\sum_{i \in [n]} Z'_i = \sum_{i \notin I} Z'_i + \sum_{i \in I \setminus S} Z_i + \sum_{i \in S} (1 - Z_i) \geq \frac{n}{2} + \sqrt{n}.$$

So it suffices to show that $\Pr(\mathcal{E}_{out} \wedge \mathcal{E}_S) \geq \frac{1}{1000}$.

For that, let $s \subseteq I$ be a subset of size at least $\frac{n}{9} - \sqrt{n}$; we lower bound the probability $\Pr(\mathcal{E}_s \mid S = s)$. First, using the fact that Z_i and Z'_i are (uniform) 0/1 random variables, we can determine the joint probability of (Z_i, Z'_i); more precisely, $\Pr(Z_i = Z'_i = 0) = \Pr(Z_i = Z'_i = 1) = \Pr(Z_i = Z'_i)/2$, and $\Pr(Z_i = 0 \wedge Z'_i = 1) = \Pr(Z_i = 1 \wedge Z'_i = 0) = \Pr(Z_i \neq Z'_i)/2$. These allow us to see that Z_i is independent of the event $Z_i \neq Z'_i$: we compute $\Pr(Z_i = 0 \mid Z_i \neq Z'_i) = \Pr(Z_i = 0 \wedge Z'_i = 1)/\Pr(Z_i \neq Z'_i) = \frac{1}{2} = \Pr(Z_i = 0)$, and similarly $\Pr(Z_i = 1 \mid Z_i \neq Z'_i) = \Pr(Z_i = 1)$. Employing this independence over all $i \in I$, we see that $(Z_i)_{i \in I}$ is independent from the event $S = s$, and hence $\Pr(\mathcal{E}_s \mid S = s) = \Pr(\mathcal{E}_s)$. To bound the latter, from Chebychev's inequality we have $\Pr\left(\sum_{i \in I \setminus s} Z_i \in \frac{|I-s|}{2} \pm \sqrt{n}\right) \geq 1 - \frac{1}{12}$. Also, using the fact that the Z_i's are independent and uniform over $\{0,1\}$, standard anti-concentration bounds (see Theorem A.2 in the Appendix) give $\Pr\left(\sum_{i \in s}(1 - Z_i) \geq \frac{|s|}{2} + 3\sqrt{n}\right) \geq e^{-147}$ (this also uses the fact $3\sqrt{n} \leq |s|/8$, which follows from our assumptions on n and $|s|$). Finally, using independence of $(Z_i)_{i \in I \setminus s}$ and $(Z_i)_{i \in s}$, we have that $\Pr(\mathcal{E}_s \mid S = s) = \Pr(\mathcal{E}_s) \geq e^{-148}$.

Now we can proceed with lower bounding $\Pr(\mathcal{E}_{out} \wedge \mathcal{E}_S)$. By independence of $(Z_i)_{i \notin I}$ and $(Z_i)_{i \in I}$ we have $\Pr(\mathcal{E}_{out} \wedge \mathcal{E}_S) = \Pr(\mathcal{E}_{out}) \cdot \Pr(\mathcal{E}_S)$. Also

$$\begin{aligned} \Pr(\mathcal{E}_S) = \mathbb{E}_S[\Pr(\mathcal{E}_s \mid S = s)] &\geq \mathbb{E}_S \left[\Pr(\mathcal{E}_s \mid S = s) \middle| |S| \geq \frac{n}{9} - \sqrt{n}\right] \Pr\left(|S| \geq \frac{n}{9} - \sqrt{n}\right) \\ &\geq e^{-148} \cdot \left(1 - \frac{1}{12}\right) \geq e^{-149}, \end{aligned}$$

where for the second inequality we use the bound from the previous paragraph on the first term and Chebychev's inequality on the second term. Since $\Pr(\mathcal{E}_{out}) \geq \frac{2}{3}$, it follows that $\Pr(\mathcal{E}_{out} \wedge \mathcal{E}_S) \geq e^{-150}$. This concludes the proof. □

H Stronger Direct Sum for XOR

H.1 Proof of Lemma 5.4

To simplify the notation we omit the superscript in XOR^k. It is actually easier to work with $\{-1, 1\}$ domains, so define $\tilde{\mathsf{XOR}} : \{-1, 1\}^k \to \{-1, 1\}$ as the mapping $(z_1, \ldots, z_k) \mapsto \prod_i z_i$. It is easy to see that it suffices to prove the (t, γ)-noise sensitivity statement for $\tilde{\mathsf{XOR}}$.

Let $\boldsymbol{Z}$ and $\boldsymbol{Z}'$ be $(t, 1 - \alpha)$-correlated random vectors over $\{-1, 1\}^k$, both with k-fold product distributions. Due to our choice of codomain $\{-1, 1\}$, notice that

$$\Pr\left(\tilde{\mathsf{XOR}}(\boldsymbol{Z}) \neq \tilde{\mathsf{XOR}}(\boldsymbol{Z}')\right) = \frac{1}{2} - \frac{1}{2}\mathbb{E}\big[\tilde{\mathsf{XOR}}(\boldsymbol{Z}) \cdot \tilde{\mathsf{XOR}}(\boldsymbol{Z}')\big], \tag{10}$$

so it suffices to lower bound the right-hand side. By the definition of $\tilde{\mathsf{XOR}}$, we have $\mathbb{E}\big[\tilde{\mathsf{XOR}}(\boldsymbol{Z}) \cdot \tilde{\mathsf{XOR}}(\boldsymbol{Z}')\big] = \mathbb{E}\big[\prod_i (Z_i \cdot Z_i')\big] = \prod_i \mathbb{E}[Z_i \cdot Z_i']$, where the last equality uses the k-fold product structure of the vectors $\boldsymbol{Z}$ and $\boldsymbol{Z}'$. But $\mathbb{E}[Z_i \cdot Z_i'] = -1 + 2\Pr(Z_i = Z_i')$. By $(t, 1 - \alpha)$-correlation, at least t indices i have $\mathbb{E}[Z_i \cdot Z_i'] \leq 1 - 2\alpha$, and the other indices have $\mathbb{E}[Z_i \cdot Z_i'] \leq 1$. Replacing these bounds on equation (10) concludes the proof.

H.2 Proof of Corollary 5.5

Let $\alpha = \frac{1}{k}$. From Theorem 4.2 we have $\mathsf{R}(f)_{\alpha}^{\to,[]} \leq \log p_{\alpha}^* + 1$; let $\beta = \frac{1}{p_{\alpha}^*}$. We assume that $\log p_{\alpha}^* = \log \frac{1}{\beta}$ is at least 7, otherwise there is nothing to prove (since $\mathsf{IC}(.)$ is non-negative).

Using Lemma 4.1, let μ and ν be distributions over $\mathcal{X}$ and $\mathcal{Y}$ respectively such that the rows of f form an (α, β)-code with respect to μ and $\|.\|_{\nu}$. Now apply Theorem 5.3 with $w = \frac{1}{2}$ (notice that our assumption on β implies $w \leq 1 - \beta$) to get that the rows of $\mathsf{XOR}^k \odot f$ form a (a_w, β_w)-code with $a_w = \mathrm{NS}_{1-\alpha,\mathfrak{prod}}^{k(\frac{1}{2}-\beta)}(\mathsf{XOR}^k)$ and $\beta_w \leq (2e\beta)^{k/2}$. Employing Lemma 5.4, we get the following lower bound on α_w:

$$\alpha_w \geq \frac{1}{2}\left(1 - (1 - 2\alpha)^{k(\frac{1}{2}-\beta)}\right) \geq \frac{1}{2}\left(1 - (1 - 2\alpha)^{2k/5}\right) \geq \frac{1}{2}\left(1 - e^{-4/5}\right) \geq \frac{1}{4},$$

where the second inequality uses the assumption $\log \frac{1}{\beta} \geq 7$, and the third uses $(1-p) \leq e^{-p}$ and $\alpha = \frac{1}{k}$. Then Theorem 3.2 gives that

$$\mathsf{IC}(\mathsf{XOR}^k \odot f)^{\to,[]}_{\frac{1}{4}} \geq \frac{1}{4}\log\frac{1}{4\beta_w} = \frac{1}{4}\left(\frac{k}{2}\log\frac{1}{2e\beta} - 2\right) \geq \frac{k}{8}\left(\log\frac{1}{\beta} - 7\right) \geq \frac{k}{8}\left(\mathsf{R}(f)^{\to,[]}_{\frac{1}{k}} - 8\right),$$

thus concluding the proof.

I Proofs for Streaming Applications

I.1 (t, γ)-Noise Sensitivity of OR

Let $\mathsf{OR}^k : \{0,1\}^k \to \{0,1\}$ denote the k-ary OR function, that is $\mathsf{OR}^k(z_1, \ldots, z_k) = 0$ iff $z_i = 0$ for all i. We also lower bound the (t, γ)-noise sensitivity of OR^k, but for that we need to restrict the distributions allowed.

Lemma I.1. *Let λ be the distribution over $\{0,1\}$ that puts mass $1/k$ on 1. Then for $\alpha \in [0, \frac{2}{k}]$,*

$$NS^t_{1-\alpha,\{\lambda^k\}}(\mathsf{OR}^k) \geq \left(1 - \frac{1}{k}\right)^k \left[1 - \left(1 - \frac{\alpha}{2}\right)^t\right].$$

Proof. To simplify the notation we drop the superscript on OR^k. Consider two $(t, 1-\alpha)$-correlated random vectors $\boldsymbol{Z}$ and $\boldsymbol{Z}'$ over $\{0,1\}^k$ with distribution λ^k. (Indeed, for $\alpha \leq \frac{2}{k}$ there are $1-\alpha$ correlated random variables Z, Z' with distributions equal to λ, for instance by defining Z' as follows: when $Z = 1$, set $Z' = 0$; when $Z = 0$, with probability $\frac{1}{k-1}$ set $Z' = 1$, and otherwise set $Z' = 0$.) Let $I \subseteq [k]$ be a set of size t such that $\Pr(Z_i \neq Z'_i) \geq \alpha$ for all $i \in I$. Conditioning on $\boldsymbol{Z} = \boldsymbol{0}$ we have

$$\begin{aligned}\Pr\big(\mathsf{OR}(\boldsymbol{Z}) \neq \mathsf{OR}(\boldsymbol{Z}') \mid \boldsymbol{Z} = \boldsymbol{0}\big) &= \Pr\left(\bigvee_{i=1}^k Z'_i \mid \boldsymbol{Z} = \boldsymbol{0}\right) = 1 - \prod_{i=1}^k \Pr(Z'_i = Z_i \mid Z_i = 0) \\ &\geq 1 - \prod_{i \in I} \frac{\Pr(Z_i = Z'_i = 0)}{\Pr(Z_i = 0)}, \end{aligned} \tag{11}$$

where the second equation uses the k-fold product structure of the vectors $\boldsymbol{Z}$ and $\boldsymbol{Z}'$.

To estimate the right-hand side, notice that

$$\Pr(Z_i = Z'_i = 0) = 1 - \frac{1}{2}\left[\Pr(Z_i = 1) + \Pr(Z'_i = 1) + \Pr(Z_i \neq Z'_i)\right].$$

So for $i \in I$ this implies $\Pr(Z_i = Z'_i = 0) \leq 1 - \frac{1}{k} - \frac{\alpha}{2}$. Replacing this bound on (11) and using $\Pr(Z_i = 0) = 1 - \frac{1}{k}$ we get

$$\Pr(\mathsf{OR}(\boldsymbol{Z}) \neq \mathsf{OR}(\boldsymbol{Z}') \mid \boldsymbol{Z} = \boldsymbol{0}) \geq 1 - \left(1 - \frac{\alpha}{2(1 - \frac{1}{k})}\right)^t \geq 1 - \left(1 - \frac{\alpha}{2}\right)^t.$$

Since $\Pr(\mathsf{OR}(\boldsymbol{Z}) \neq \mathsf{OR}(\boldsymbol{Z}')) \geq \Pr(\mathsf{OR}(\boldsymbol{Z}) \neq \mathsf{OR}(\boldsymbol{Z}') \mid \boldsymbol{Z} = \boldsymbol{0}) \Pr(\boldsymbol{Z} = \boldsymbol{0})$ and $\Pr(\boldsymbol{Z} = \boldsymbol{0}) = \big(1 - \frac{1}{k}\big)^k$, the result follows. □

I.2 Augmented Indexing of OR of Indexing

We consider the problem $\mathsf{IOI}^{t,n}_{k,m}$ which consists of taking an augmented indexing over $\mathsf{OR}^n \odot \mathsf{ind}_{k,m}$. To simplify the notation, let $\mathcal{X} = ([k]^m)^n$ and $\mathcal{Y} = ([k]^m \times [m])^n$ denote the set of Alice's and Bob's inputs in the problem $\mathsf{OR}^n \odot \mathsf{ind}_{k,m}$. The function $\mathsf{IOI}^{t,n}_{k,m}$ is defined as follows: Alice has as input $s_1, \dots, s_t \in \mathcal{X}$, and Bob has as input an index $i \in [t]$, part of Alice's input $s_1, \dots, s_{i-1}$, and also an $\ell \in \mathcal{Y}$; then $\mathsf{IOI}^{t,n}_{k,m}((s_1, \dots s_t), (i, s_1, \dots, s_{i-1}, \ell)) = \mathsf{OR}^n \odot \mathsf{ind}_{k,m}(s_i, \ell)$.

Lemma I.2. *For $n \geq 8e$, $m \geq 2$ and $0 < \delta \leq 1$,*

$$\mathsf{IC}(\mathsf{IOI}^{t,n}_{n/\delta,m})^{\to,[]}_{\frac{\delta}{400}} \geq \Omega(tnm \log(n/\delta)).$$

Proof. Let $k = n/\delta$. First we get $\mathsf{IC}(\mathsf{OR}^n \odot \mathsf{ind}_{k,m})^{\to,[]}_{\frac{\delta}{400}} \geq \Omega(nm \log k)$ by putting together our lower bound for indexing from Theorem E.1, the connection between codes and (t, γ)-noise sensitivity from Theorem 5.3, and our lower bound on the latter for OR^n from Lemma I.1.

To see this, let μ be the uniform distribution over $[k]^m$ and ν the uniform distribution over $[k]^m \times [m]$; from Theorem E.1 we have that the rows $\{\mathsf{ind}_{k,m}(u)\}_{u \in [k]^m}$ form a $(\frac{1}{k}, (\frac{k}{2e})^{-m/2})$-code with respect to μ and $\|.\|_\nu$; let $\beta = \left(\frac{k}{2e}\right)^{-m/2}$. Moreover, let λ be the distribution of $\mathsf{ind}_{k,m}(u, V)$ for $u \in [k]^m$ and $V \sim \nu$, and notice that this distribution is indeed independent of u. Since our assumption on k and m implies $\frac{1}{2} \leq 1 - \beta$, we can use Theorem 5.3 with $w = \frac{1}{2}$ to get that the rows $\{\mathsf{OR}^n \odot \mathsf{ind}_{k,m}(\boldsymbol{u})\}_{\boldsymbol{u} \in ([k]^m)^n}$ form a $(\alpha_{1/2}, \beta_{1/2})$-code w.r.t. μ^n and $\|.\|_{\nu^n}$, with $\alpha_{1/2} = \mathrm{NS}^{n(\frac{1}{2} - \beta)}_{1 - \frac{1}{k}, \{\lambda^n\}}(\mathsf{OR}^n)$ and $\beta_{1/2} = (2e\beta)^{n/2}$. Now notice that λ puts mass $\frac{1}{k}$ on the value 1; Lemma I.1 then gives

$$\begin{aligned}
\alpha_{1/2} \geq \mathrm{NS}^{\frac{n}{4}}_{1 - \frac{1}{k}, \{\lambda^n\}}(\mathsf{OR}^n) &\geq \left(1 - \frac{1}{n}\right)^n \left[1 - \left(1 - \frac{1}{2k}\right)^{n/4}\right] \\
&\geq \left(1 - \frac{1}{n}\right)^n \left[1 - e^{-\frac{\delta}{8}}\right] \geq \left(1 - \frac{1}{n}\right)^n \left[1 - \left(1 - \frac{\delta}{16}\right)\right] \geq \frac{\delta}{50},
\end{aligned}$$

where the first inequality uses the fact $\frac{1}{2} - \beta \geq \frac{1}{4}$ (from our assumption on n and m), the third uses the bound $1 - p \leq e^{-p}$, which holds for all p, the fourth uses $e^{-p} \leq 1 - p/2$ that holds for $0 \leq p \leq 1$, and the last uses $n \geq 4e$. Then using Theorem 3.2 and our bound on $\alpha_{1/2}$ and $\beta_{1/2}$, we get $\mathsf{IC}(\mathsf{OR}^n \odot \mathsf{ind}_{k,m})^{\to,[]}_{\frac{\delta}{400}} \geq \Omega\left(nm \log k\right)$, and the claim follows.

Then standard direct sum arguments for information complexity, we get $\mathsf{IC}(\mathsf{IOI}^{t,n}_{k,m})^{\to,[]}_{\frac{\delta}{400}} \geq t \cdot \mathsf{IC}(\mathsf{OR}^n \odot \mathsf{ind}_{k,m})^{\to,[]}_{\frac{\delta}{400}} = \Omega(tnm \log k)$ (see Section B.1 of [14]). This concludes the proof. □

I.3 Proof of Theorem 6.1

We use the following reduction from [14]. As in Section I.2 , let $\mathcal{X} = ([k]^m)^n$ and $\mathcal{Y} = ([k]^m \times [m])^n$ denote the set of Alice's and Bob's inputs in the problem $\mathsf{OR}^n \odot \mathsf{ind}_{k,m}$. Let Alice have as input $s_1, \dots, s_t \in \mathcal{X}$, and Bob have as input an index $i \in [t]$, part of Alice's input $s_1, \dots, s_{i-1}$, and also an $\ell \in \mathcal{Y}$. For $s \in \mathcal{X}$ and $\ell \in \mathcal{Y}$ we denote their j-th components as s^j and ℓ^j respectively.

Theorem I.3 ([14])**.** *Let $k = n/\delta$ and $m = 1/4\epsilon^2$. There exist two encodings of $(s_1, \ldots, s_t)$ and $(i, s_1, \ldots, s_{i-1}, \ell)$ based on shared randomness into n vectors $\mathbf{v}^1, \mathbf{v}^2, \ldots, \mathbf{v}^n \in [\pm M]^d$ and n vectors $\mathbf{u}^1$, $\mathbf{u}^2$, $\ldots$, $\mathbf{u}^n \in [\pm M]^d$ respectively such that the first encoding has $t = c \log d$ and the second encoding has $t = c \log M$ for some constant $c > 0$, and both encodings satisfy for every $j \in [n]$:*

1. *If $\mathsf{ind}_{k,m}(s_i^j, \ell^j) = 0$, then with probability $1-\delta/n$ we have $||\mathbf{u}^j - \mathbf{v}^j||_p^p \geq \theta_i(1+\epsilon)$ for all $p > 0$.*
2. *If $\mathsf{ind}_{k,m}(s_i^j, \ell^j) = 1$ then with probability at least $1 - \delta/n$ we have $||\mathbf{u}^j - \mathbf{v}^j||_p^p \leq \theta_i(1-\epsilon)$ for all $p > 0$,*

where θ_i are functions of i, d, ϵ, δ and n defined by the encodings.

Using the encodings above the parties can solve $\mathsf{IOI}^{c \log d, n}_{n/\delta, 1/4\epsilon^2}$ and $\mathsf{IOI}^{c \log M, n}_{n/\delta, 1/4\epsilon^2}$ with success probability $1-\delta$ by constructing vectors $\mathbf{u}^1, \ldots, \mathbf{u}^n$ and $\mathbf{v}^1, \ldots, \mathbf{v}^n$ and outputting 1 if there exists a pair of such vectors $(\mathbf{u}^j, \mathbf{v}^j)$ such that $||\mathbf{u}^j - \mathbf{v}^j||_p^p \leq \theta_i(1-\epsilon)$ and 0 otherwise. This gives a reduction from the augmented OR-indexing problem to the closest pair problem, showing that:

$$\begin{aligned} R_\delta^{\rightarrow}(\ell_p(n,d,M,\epsilon)) &\geq \max\left(R_\delta^{\rightarrow}\left(\mathsf{IOI}^{c \log d, n}_{n/\delta, 1/4\epsilon^2}\right), R_\delta^{\rightarrow}\left(\mathsf{IOI}^{c \log M, n}_{n/\delta, 1/4\epsilon^2}\right)\right) \\ &\geq \Omega\left(n \frac{1}{\epsilon^2} \log \frac{n}{\delta} (\log d + \log M))\right), \end{aligned}$$

where the last inequality is by Lemma I.2.

I.4 Proof of Theorem 6.2

We use an encoding from [14], which has the following guarantee.

Theorem I.4 ([14])**.** *Let $k = n/\delta$ and $m = 1/4\epsilon^2$. There exist two encodings $(s_1, \ldots, s_t)$ and $(i, s_1, \ldots, s_{i-1}, \ell)$ based on shared randomness R into n vectors $\mathbf{u}^1, \ldots, \mathbf{u}^n$, $\underline{\mathbf{u}}^1, \ldots, \underline{\mathbf{u}}^n$ and $\mathbf{v}^1, \ldots, \mathbf{v}^n$, where $\mathbf{u}^j(s_1^j, \ldots, s_t^j, R)$, $\underline{\mathbf{u}}^j(s_1^j, \ldots, s_{i-1}^j, R)$, $\mathbf{v}^j(i, \ell^j, R) \in [\pm M]^d$ such that the first encoding has $t = c \log d$ and the second encoding has $t = c \log M$ for some constant $c > 0$ and both encodings satisfy for all $j \in [n]$ and $r = O(\frac{1}{\epsilon^2} \log \frac{n}{\delta})$:*

1. *If $\mathsf{ind}_{k,m}(s_i^j, \ell^j) = 0$, then with probability $1 - \delta/n$ we have $\langle \mathbf{u}^j - \underline{\mathbf{u}}^j, \mathbf{v}^j \rangle \leq 10^{t-i} r$.*
2. *If $\mathsf{ind}_{k,m}(s_i^j, \ell^j) = 1$ then with probability at least $1-\delta/n$ we have $\langle \mathbf{u}^j - \underline{\mathbf{u}}^j, \mathbf{v}^j \rangle \geq (1+\epsilon) 10^{t-i} r$.*

We augment the encodings above with extra coordinates by using vectors $\mathbf{c}_i^j \in \{0,1\}^{b 10^{t-i} r}$ for $i \in [t], j \in [n]$ and a constant b. For each $i \in [t]$ the set of vectors $\mathbf{c}_i^1, \ldots, \mathbf{c}_i^n$ is chosen to be a subset of different codewords of the following code $\mathcal{C}_w$ with $w = 6 \cdot 10^{t-i} r$. Note that such a choice implies that $\langle \mathbf{c}_i^j, \mathbf{c}_i^j \rangle = 6 \cdot 10^{t-i} r$ while for $j_1 \neq j_2$ we have $\langle \mathbf{c}_i^{j_1}, \mathbf{c}_i^{j_2} \rangle \leq 3 \cdot 10^{t-i} r$.

Fact I.5 (Combinatorial designs)**.** *For every sufficiently large w there exists a constant c and a family $\mathcal{C}_w$ of codewords over $\{0,1\}^{cw}$ of size $s = 2^w$, such that every codeword in $\mathcal{C}_w$ has Hamming weight w and the distance between every two codewords in $\mathcal{C}_w$ is at least w.*

Proof. The existence of the code above corresponds to existence of combinatorial $(2^w, cw, w, w/2)$-designs, which follows by a standard probabilistic argument. □

For two vectors $\mathbf{a}$ and $\mathbf{b}$ we denote their concatenation as $\mathbf{ab}$. Let $\mathbf{c^j} = \mathbf{c}^j_1\mathbf{c}^j_2\dots\mathbf{c}^j_t$ where $\mathbf{c}^j_i$ are defined as above. Let $\mathbf{c}^j_{-i} = \mathbf{c}^j_1\mathbf{c}^j_2\dots\mathbf{c}^j_{i-1}0^{b10^{t-i}r}\mathbf{c}^j_{i+1}\dots\mathbf{c}^j_t$ denote the same concatenation but with the entries corresponding to $\mathbf{c}^j_i$ zeroed out and let $\mathbf{c}^j_{+i} = 0^{b10^{t-1}r}0^{b10^{t-2}r}\dots0^{b10^{t-i+1}r}\mathbf{c}^j_i 0^{b10^{t-i-1}r}\dots0^{br}$ denote the concatenation of $\mathbf{c}^j_i$ with matching number of 0's on both sides.

In the reduction Alice constructs vectors $\mathbf{u}^{*j} = \mathbf{u}^j\mathbf{c}^j$. Bob constructs vectors $\underline{\mathbf{u}}^{*j} = \underline{\mathbf{u}}^j\mathbf{c}^j_{-i}$ and $\mathbf{v}^{*j} = \mathbf{v}^j\mathbf{c}^j_{+i}$, where $\mathbf{u}^j, \underline{\mathbf{u}}^j$ and $\mathbf{v}^j$ are constructed using one of the two constructions given by the Theorem I.4.

Proposition I.6. *The construction above satisfies that for all $j \in [n]$:*

1. *With probability at least $1-\delta/n$ it holds that $\langle \mathbf{u}^{*j} - \underline{\mathbf{u}}^{*j}, \mathbf{v}^{*j}\rangle \le 7 \cdot 10^{t-i}r$ if $\mathsf{ind}_{k,m}(s^j_i, \ell^j) = 0$.*

2. *With probability at least $1-\delta/n$ it holds that $\langle \mathbf{u}^{*j} - \underline{\mathbf{u}}^{*j}, \mathbf{v}^{*j}\rangle \ge (7+\epsilon) \cdot 10^{t-i}r$ if $\mathsf{ind}_{k,m}(s^j_i, \ell^j) = 1$.*

3. *$\langle \mathbf{u}^{*j} - \underline{\mathbf{u}}^{*j}, \mathbf{v}^{*j'}\rangle \le 6 \cdot 10^{t-i}r$ if $j \neq j'$.*

Proof. We have:

$$\begin{aligned}\langle \mathbf{u}^{*j} - \underline{\mathbf{u}}^{*j}, \mathbf{v}^{*j}\rangle = \langle \mathbf{u}^j - \underline{\mathbf{u}}^j, \mathbf{v}^j\rangle + \langle \mathbf{c}^j - \mathbf{c}^j_{-i}, \mathbf{c}^j_{+i}\rangle &= \langle \mathbf{u}^j - \underline{\mathbf{u}}^j, \mathbf{v}^j\rangle + \langle \mathbf{c}^j_{+i}, \mathbf{c}^j_{+i}\rangle \\ &= \langle \mathbf{u}^j - \underline{\mathbf{u}}^j, \mathbf{v}^j\rangle + 6 \cdot 10^{t-i}r,\end{aligned}$$

and the first two properties follow from Theorem I.4.

For the third property note that:

$$\begin{aligned}\langle \mathbf{u}^{*j} - \underline{\mathbf{u}}^{*j}, \mathbf{v}^{*j'}\rangle = \langle \mathbf{u}^j - \underline{\mathbf{u}}^j, \mathbf{v}^{j'}\rangle + \langle \mathbf{c}^j - \mathbf{c}^j_{-i}, \mathbf{c}^{j'}_{+i}\rangle &= \langle \mathbf{u}^j - \underline{\mathbf{u}}^j, \mathbf{v}^{j'}\rangle + \langle \mathbf{c}^j_{+i}, \mathbf{c}^{j'}_{+i}\rangle \\ &\le \langle \mathbf{u}^j - \underline{\mathbf{u}}^j, \mathbf{v}^{j'}\rangle + 3 \cdot 10^{t-i}r,\end{aligned}$$

The number of non-zero coordinates in the vector $\mathbf{u}^j - \underline{\mathbf{u}}^j$ is at most $\sum_{q=i}^t 2r10^{t-q} \le 2r \cdot 10^{t-i}\frac{10}{9} \le 3r \cdot 10^{t-i}$ and hence $\langle \mathbf{u}^j - \underline{\mathbf{u}}^j, \mathbf{v}^{j'}\rangle \le 3r \cdot 10^{t-i}$ completing the proof. □

Consider matrices $\mathbf{U}, \underline{\mathbf{U}}$ and $\mathbf{V}$, each with n rows formed by vectors $\mathbf{u}^j, \underline{\mathbf{u}}^j$ and $\mathbf{v}^j$. Note that by Proposition I.6 and a union bound with probability at least $1-\delta$ the largest entry in $(\mathbf{U} - \underline{\mathbf{U}})\,\mathbf{V}^T$ is at least $(7+\epsilon) \cdot 10^{t-i}r$ if there exists $j \in [n]$ such that $\mathsf{ind}_{k,m}(s^j_i, \ell^j) = 1$, otherwise it is at most $7 \cdot 10^{t-i}r$. Thus, by approximating the largest entry up to $(1+\epsilon/10)$ multiplicative error the parties can solve $\mathsf{IOI}^{c\log d, n}_{k,m}$ and $\mathsf{IOI}^{c\log M, n}_{k,m}$. Assuming the existence of the sketch matrix S and estimation procedure f_θ in the theorem statement this gives a protocol for these probelms with communication from Alice to Bob at most the bit size of $\mathbf{U}S$ since Bob can solve them by approximating the largest entry as $f_\theta((\mathbf{U} - \underline{\mathbf{U}})\,S, \mathbf{V}^T)$ as $f_\theta(\mathbf{U}S - \underline{\mathbf{U}}S, \mathbf{V}^T)$ with $\theta = 7 \cdot 10^{t-i}r$.