Electronic Colloquium on Computational Complexity, Revision 4 of Report No. 11 (2021)

Approximability of all Boolean CSPs with linear sketches[*]

Chi-Ning Chou[†] Alexander Golovnev[‡] Madhu Sudan[§] Santhoshini Velusamy[¶]

Abstract

A Boolean constraint satisfaction problem (CSP), $\mathsf{Max\text{-}CSP}(f)$, is a maximization problem specified by a constraint $f : \{-1,1\}^k \to \{0,1\}$. An instance of the problem consists of m constraint applications on n Boolean variables, where each constraint application applies the constraint to k literals chosen from the n variables and their negations. The goal is to compute the maximum number of constraints that can be satisfied by a Boolean assignment to the n variables. In the (γ,β)-approximation version of the problem for parameters $\gamma \geq \beta \in [0,1]$, the goal is to distinguish instances where at least γ fraction of the constraints can be satisfied from instances where at most β fraction of the constraints can be satisfied.

In this work we consider the approximability of $\mathsf{Max\text{-}CSP}(f)$ in the context of sketching algorithms and completely characterize the approximability of all Boolean CSPs. Specifically, given f, γ and β we show that either (1) the (γ,β)-approximation version of $\mathsf{Max\text{-}CSP}(f)$ has a linear sketching algorithm using $O(\log n)$ space, or (2) for every $\varepsilon > 0$ the $(\gamma-\varepsilon,\beta+\varepsilon)$-approximation version of $\mathsf{Max\text{-}CSP}(f)$ requires $\Omega(\sqrt{n})$ space for any sketching algorithm. We also prove lower bounds against streaming algorithms for several CSPs. In particular, we recover the streaming dichotomy of [CGV20] for $k = 2$ and show streaming approximation resistance of all CSPs for which $f^{-1}(1)$ supports a distribution with uniform marginals.

Our positive results show wider applicability of bias-based algorithms used previously by [GVV17] and [CGV20] by giving a systematic way to discover biases. Our negative results combine the Fourier analytic methods of [KKS15], which we extend to a wider class of CSPs, with a rich collection of reductions among communication complexity problems that lie at the heart of the negative results.

[*] This paper replaces the paper [CGSV21] by the authors. The previous version had errors and is now withdrawn.

[†] School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts, USA. Supported by NSF awards CCF 1565264 and CNS 1618026. Email: `chiningchou@g.harvard.edu`.

[‡] Department of Computer Science, Georgetown University. Email: `alexgolovnev@gmail.com`.

[§] School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts, USA. Supported in part by a Simons Investigator Award and NSF Award CCF 1715187. Email: `madhu@cs.harvard.edu`.

[¶] School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts, USA. Supported in part by a Simons Investigator Award and NSF Award CCF 1715187. Email: `svelusamy@g.harvard.edu`.

ISSN 1433-8092

Contents

1 Introduction

In this paper we give a complete characterization of the approximability of Boolean constraint satisfaction problems (CSPs) by sketching algorithms. We describe the exact class of problems below, and give a brief history of previous work before giving our results.

1.1 Boolean CSPs

In this paper we use $\mathbb{N}$ to denote the set of natural numbers $\{1, 2, 3, \ldots\}$. For $n \in \mathbb{N}$ we use $[n]$ to denote the set $\{1, 2, \ldots, n\}$. We refer to a variable taking values in $\{-1, 1\}$ as a Boolean variable. Given a Boolean variable X, we refer to X and $-X$ as the literals associated with X. For vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ we use $\mathbf{a} \odot \mathbf{b}$ to denote their coordinate-wise product. I.e., if $\mathbf{a} = (a_1, \ldots, a_n)$ and $\mathbf{b} = (b_1, \ldots, b_n)$ then $\mathbf{a} \odot \mathbf{b} = (a_1b_1, \ldots, a_nb_n)$.

In this paper, a Boolean CSP is a maximization problem, $\mathsf{Max\text{-}CSP}(f)$, specified by a single constraint function $f : \{-1, 1\}^k \to \{0, 1\}$ for some positive integer k. Given n Boolean variables $x_1, \ldots, x_n$, an application of the constraint function f to these variables, which we term simply a *constraint*, is given by two k-tuples $\mathbf{j} = (j_1, \ldots, j_k) \in [n]^k$ and $\mathbf{b} = (b_1, \ldots, b_k) \in \{-1, 1\}^k$ where the j_i's are distinct, and represents the application of the constraint function f to the literals $b_1x_{j_1}, \ldots, b_kx_{j_k}$. Specifically an assignment $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_n) \in \{-1, 1\}^n$ satisfies a constraint given by $(\mathbf{j}, \mathbf{b})$ if $f(b_1\sigma_{j_1}, \ldots, b_k\sigma_{j_k}) = 1$. For a constraint $C = (\mathbf{j}, \mathbf{b})$ and assignment $\boldsymbol{\sigma}$ we use $\boldsymbol{\sigma}|_{\mathbf{j}}$ as shorthand for $(\sigma_{j_1}, \ldots, \sigma_{j_k})$ and $C(\boldsymbol{\sigma})$ as shorthand for $f(\mathbf{b} \odot \boldsymbol{\sigma}|_{\mathbf{j}}) = f(b_1\sigma_{j_1}, \ldots, b_k\sigma_{j_k})$. An instance Ψ of weighted $\mathsf{Max\text{-}CSP}(f)$ consists of m constraints $C_1, \ldots, C_m$ applied to n variables $x_1, \ldots, x_n$, along with m non-negative weights $w_1, \ldots, w_m$. The value of an assignment $\boldsymbol{\sigma} \in \{-1, 1\}^n$ on an instance $\Psi = ((C_1, w_1), \ldots, (C_m, w_m))$, denoted $\mathsf{val}_\Psi(\boldsymbol{\sigma})$, is the fraction of weight of constraints satisfied by $\boldsymbol{\sigma}$, i.e., $\mathsf{val}_\Psi(\boldsymbol{\sigma}) = \frac{\sum_{i \in [m]} w_i C_i(\boldsymbol{\sigma})}{\sum_{i \in [m]} w_i}$. The goal of the *exact* problem is to compute the maximum, over all assignments, of the value of the assignment on the input instance, i.e., to compute, given Ψ, the quantity $\mathsf{val}_\Psi = \max_{\boldsymbol{\sigma} \in \{-1,1\}^n} \{\mathsf{val}_\Psi(\boldsymbol{\sigma})\}$. [1]

In this work we consider the approximation version of $\mathsf{Max\text{-}CSP}(f)$, which we study in terms of the "gapped promise problems". Specifically given $0 \leq \beta < \gamma \leq 1$, the (γ, β)-approximation version of $\mathsf{Max\text{-}CSP}(f)$, abbreviated (γ, β)-$\mathsf{Max\text{-}CSP}(f)$, is the task of distinguishing between instances from $\Gamma = \{\Psi | \operatorname{opt}(\Psi) \geq \gamma\}$ and instances from $B = \{\Psi | \operatorname{opt}(\Psi) \leq \beta\}$. It is well-known that this distinguishability problem is a refinement of the usual study of approximation which usually studies the ratio of γ/β for tractable versions of (γ, β)-$\mathsf{Max\text{-}CSP}(f)$. See Proposition 2.10 for a formal statement in the context of streaming approximability of $\mathsf{Max\text{-}CSP}(f)$ problems.

1.2 Streaming algorithms

We study the complexity of (γ, β)-$\mathsf{Max\text{-}CSP}(f)$ in the setting of randomized streaming algorithms. Here, an instance $\Psi = (C_1, \ldots, C_m)$ is presented as a stream $\sigma_1, \sigma_2, \ldots, \sigma_m$ with $\sigma_i = (\mathbf{j}_i, \mathbf{b}_i)$ representing the ith constraint. We study the space required to solve the (γ, β)-approximation version of $\mathsf{Max\text{-}CSP}(f)$. Specifically we consider algorithms that are allowed to use internal randomness and s bits of space. The algorithms output a single bit at the end. They are said to solve the

[1] We note that the literature on CSPs has several generalizations: one may allow an entire set of constraint functions, not just a single one. One may restrict the constraint applications to be applicable only to variables and not literals. And finally one can of course consider non Boolean CSPs. We do not do any of those in this paper, though extending our techniques to classes of functions seems immediately feasible. See more discussion in Section 1.7.

(γ, β)-approximation problem correctly if they output the correct answer with probability at least $2/3$ (i.e., they err with probability at most $1/3$).

The main focus of this work is sketching algorithms, a special class of streaming algorithms, where the algorithm's output is determined by a small sketch it produces of the input stream, and the sketch itself has the property that the sketch of the concatenation of two streams can be computed from the sketches of the two component streams. (See Definition 3.3 for a formal definition.) We define the space of the sketching algorithm to be the length of the sketch.

Our main dividing line is between algorithms that work with space $O(\mathsf{poly}\log n)$, versus algorithms that require space at least n^ε for some $\varepsilon > 0$. In informal usage we refer to a streaming problem as "easy" if it can be solved with polylogarithmic space (the former setting) and "hard" if it requires polynomial space (the latter setting). We note that all the positive results (algorithms) given in this paper are linear sketching algorithms which are more restrictive than general sketching algorithms. We also note that many of our lower bounds work against general streaming algorithms and we elaborate on this in Section 1.4.

1.3 Past work

To the best of our knowledge, streaming algorithms for Boolean CSPs have not been investigated extensively. Here we cover the few results we are aware of. On the positive side, it may be surprising that there exists any non-trivial algorithm at all. Here, and later, we describe algorithms solving the $(1, \rho(f) - \varepsilon)$-approximation problem for $\varepsilon > 0$ as "trivial", where $\rho(f) = 2^{-k} \sum_{\mathbf{a} \in \{-1,1\}^k} f(\mathbf{a})$ is the fraction of clauses satisfied by a random assignment. Note that the algorithm that always outputs 1 correctly solves the $(1, \rho(f) - \varepsilon)$-approximation version of the $\mathsf{Max\text{-}CSP}(f)$ problem.

It turns out that there do exist some non-trivial approximation algorithms for Boolean CSPs. This was established by the work of Guruswami, Velingker, and Velusamy [GVV17] who, in our notation, gave an algorithm for the $(\gamma, 2\gamma/5 - \varepsilon)$-approximation version of $\mathsf{Max\text{-}2AND}$, for every $\gamma \in [0, 1]$ ($\mathsf{Max\text{-}2AND}$ is the $\mathsf{Max\text{-}CSP}(f)$ problem corresponding to $f(a, b) = 1$ if $a = b = 1$ and 0 otherwise). A central ingredient in their algorithm is the ability of streaming algorithms to approximate the ℓ_1 norm of a vector in the turnstile setting, which allows them to estimate the "bias" of n variables (how often they occur positively in constraints, as opposed to negatively). Subsequently, the work of Chou, Golovnev, and Velusamy [CGV20] further established the utility of such algorithms, which we refer to as bias-based algorithms, by giving optimal algorithms for all Boolean CSPs on 2 variables. In particular they give a better (optimal!) analysis of bias-based algorithms for $\mathsf{Max\text{-}2AND}$, and show that $\mathsf{Max\text{-}2SAT}$ also has an optimal algorithm based on bias. We note that $\mathsf{Max\text{-}2SAT}$ is again not covered by the results of the current paper since it involves two functions corresponding to clauses of length 1, and clauses of length 2.

On the negative side, the problem that has been explored the most is $\mathsf{Max\text{-}CUT}$, or in our language $\mathsf{Max\text{-}2XOR}$, which corresponds to $f(x, y) = x \oplus y = (1 - xy)/2$.[2] Kapralov, Khanna, and Sudan [KKS15] showed that $\mathsf{Max\text{-}2XOR}$ does not have a $(1, 1/2 + \varepsilon)$-approximation algorithm using $o(\sqrt{n})$-space, for any $\varepsilon > 0$. This was subsequently improved upon by Kapralov, Khanna, Sudan, and Velingker [KKSV17], and Kapralov and Krachun [KK19]. The final paper [KK19] completely resolves $\mathsf{Max\text{-}CUT}$ and $\mathsf{Max\text{-}2XOR}$ showing that $(1, 1/2 + \varepsilon)$-approximation for these problems requires $\Omega(n)$ space. Turning to other problems, the work by [GVV17] notices that the

[2] Strictly speaking this work does not include $\mathsf{Max\text{-}CUT}$, which does not allow constraints to be placed on arbitrary literals. $\mathsf{Max\text{-}2XOR}$ is however very closely related and in particular is harder than $\mathsf{Max\text{-}CUT}$.

$(1, 1/2 + \varepsilon)$-inapproximability of Max-2XOR immediately yields $(1, 1/2 + \varepsilon)$-inapproximability of Max-2AND as well. In [CGV20] more sophisticated reductions are used to improve the inapproximability result for Max-2AND to a $(\gamma, 4\gamma/9 + \varepsilon)$-inapproximability for some positive γ, which turns out to be the optimal ratio by their algorithm and analysis. As noted earlier their work gives optimal algorithms for all functions $f : \{-1, 1\}^2 \to \{0, 1\}$.

1.4 Our results

Our main theorem is a decidable dichotomy theorem for (γ, β)-Max-CSP(f) with sketching algorithms.

Theorem 1.1. *For every $k \in \mathbb{N}$, for every function $f : \{-1, 1\}^k \to \{0, 1\}$, and for every $0 \leq \beta < \gamma \leq 1$, at least one of the following always holds:*

1. *(γ, β)-Max-CSP(f) has a $O(\log n)$-space linear sketching algorithm.*
2. *For every $\varepsilon > 0$, any sketching algorithm that solves $(\gamma - \varepsilon, \beta + \varepsilon)$-Max-CSP$(f)$ requires $\Omega(\sqrt{n})$ space. If $\gamma = 1$, then any sketching algorithm that solves $(1, \beta + \varepsilon)$-Max-CSP(f) requires $\Omega(\sqrt{n})$ space.*

Furthermore, there is an algorithm using space $\mathsf{poly}(2^k, \ell)$ that decides which of the two conditions holds, given the truth-table of f, and γ and β as ℓ-bit rationals[3].

In analogy with the terminology used in the study of CSP approximation in polynomial time, we define a problem to be "approximation-resistant" if it is hard to beat the random assignment with $n^{o(1)}$-space. Recall $\rho(f)$ denotes the fraction of assignments that satisfy a function f. We say that Max-CSP(f) is *approximation-resistant* if, for every $\varepsilon > 0$ there exists $\delta > 0$ such that $(1, \rho + \varepsilon)$-Max-CSP(f) requires $\Omega(n^\delta)$ space.
We get the following dichotomy for approximation-resistance to sketching algorithms.

Corollary 1.2. *For every $k \in \mathbb{N}$, for every function $f : \{-1, 1\}^k \to \{0, 1\}$, if Max-CSP$(f)$ is approximation-resistant to sketching algorithms, then for every $\varepsilon > 0$, any sketching algorithm that solves $(1, \rho(f) + \varepsilon)$-approximation version of Max-CSP(f) requires $\Omega(\sqrt{n})$ space. If Max-CSP(f) is not approximation-resistant, then there exists $\varepsilon > 0$ such that $(1 - \varepsilon, \rho(f) + \varepsilon)$-Max-CSP$(f)$ can be solved by a linear sketching algorithm in logarithmic space . Furthermore, given the truth-table of the function f, there is an algorithm running in space $\mathsf{poly}(2^k)$ that decides whether or not Max-CSP(f) is approximation-resistant to sketching algorithms.*

The results above (and in particular the negative results) apply only to sketching algorithms for streaming CSPs. For general streaming algorithm, we get some partial results. To describe our next result, we define the notion of a function supporting a one-wise independent distribution.

We say that a function f *supports one-wise independence* if there exists a distribution D supported on the satisfying assignments to f, i.e., on $f^{-1}(1) \subseteq \{-1, 1\}^k$ such that its marginals are all uniform, i.e., for every $j \in [k]$, we have $\mathbb{E}_{\mathbf{a} \sim D}[a_j] = 0$.

Theorem 1.3. *If $f : \{-1, 1\}^k \to \{0, 1\}$ supports one-wise independence then Max-CSP(f) is approximation resistant in the streaming setting.*

[3] $\alpha \in \mathbb{R}$ is said to be an ℓ-bit rational if there exist integers $-2^\ell < p, q < 2^\ell$ such that $\alpha = p/q$.

We also give a (very) partial converse, showing that symmetric functions are approximation resistant if and only if they support one-wise independence (see Lemma 2.14).

While we do believe that there are other approximation-resistant problems in the streaming setting, we do not know of one even approximation-resistant to sketching algorithms (and in particular do not give one in this paper). We discuss this more in the next section.

We also give theorems capturing hardness in the streaming setting beyond the one-wise independent case. Stating the full theorem requires more notions (see Section 2.3), but as a consequence we get the following extension of the work of [CGV20] who study the setting of $k = 2$.

Theorem 1.4. *For every function $f : \{-1,1\}^2 \to \{0,1\}$, and for every $0 \leq \beta < \gamma \leq 1$, at least one of the following always holds:*

1. *(γ, β)-$\mathsf{Max\text{-}CSP}(f)$ has a $O(\log n)$-space linear sketching algorithm.*
2. *For every $\varepsilon > 0$, every streaming algorithm that solves $(\gamma - \varepsilon, \beta + \varepsilon)$-$\mathsf{Max\text{-}CSP}(f)$ requires $\Omega(\sqrt{n})$ space. If $\gamma = 1$, then $(1, \beta + \varepsilon)$-$\mathsf{Max\text{-}CSP}(f)$ requires $\Omega(\sqrt{n})$ space.*

Furthermore, there is an algorithm using space $\mathsf{poly}(\ell)$ that decides which of the two conditions holds given the truth-table of f, and γ and β as ℓ-bit rationals.

This reproduces the results of [CGV20] while giving a more refined picture of the approximability by considering all $\beta < \gamma$. In Section 2.4, we show how to apply our theorem above to get a full characterization of the approximation profile of the $\mathsf{Max\text{-}2AND}$ problem (i.e., the $\mathsf{Max\text{-}CSP}(f)$ problem for $f(x, y) = 1$ if $x = y = 1$ and 0 otherwise).

This version: This version of the paper replaces a previous version [CGSV21]. The previous version, now withdrawn, claimed Theorem 1.1 in the streaming setting, but that version had an error and the status of Theorem 1.1 in [CGSV21] is currently open.

1.5 Contrast with dichotomies in the polynomial time setting

The literature on dichotomies of $\mathsf{Max\text{-}CSP}(f)$ problems is vast. One broad family of results here [Sch78, Bul17, Zhu17] considers the exact satisfiability problems (corresponding to distinguishing between instances from $\{\Psi | \operatorname{opt}(\Psi) = 1\}$ and instances from $\{\Psi | \operatorname{opt}(\Psi) < 1\}$. Another family of results [Rag08, AM09, KTW14] considers the approximation versions of $\mathsf{Max\text{-}CSP}(f)$ and gets "near dichotomies" along the lines of this paper — i.e., they either show that the (γ, β)-approximation is easy (in polynomial time), or for every $\varepsilon > 0$ the $(\gamma - \varepsilon, \beta + \varepsilon)$-approximation version is hard (in some appropriate sense). Our work resembles the latter series of works both in terms of the nature of results obtained, the kinds of characterizations used to describe the "easy" and "hard" classes, and also in the proof approaches (though of course the streaming setting is much easier to analyze, allowing for much simpler proofs overall). We summarize their results giving comparisons to our theorem and then describe a principal contrast.

In a seminal work, Raghavendra [Rag08] gave a characterization of the polynomial time approximability of the $\mathsf{Max\text{-}CSP}(f)$ problems based on the unique games conjecture [Kho02]. Our Theorem 1.1 is analogous to his theorem, though restricted to a single function, with Boolean variables, with ability to complement variables. A characterization of approximation resistant functions is given by Khot, Tulsiani and Worah [KTW14]. Our Corollary 1.2 is analogous to this. Austrin and Mossel [AM09] show that all functions supporting a pairwise independent distribution are approximation-resistant. Our Theorem 1.3 is analogous to this theorem.

While our results run in parallel to the work on polynomial time approximability our characterizations are not immediately comparable. Indeed there are some significant differences which we highlight below. Of course there is the obvious difference that our negative results are unconditional (and not predicated on a complexity theoretic assumption like the unique games conjecture or P≠NP). But more significantly our characterization is a bit more "explicit" than those of [Rag08] and [KTW14]. In particular the former only shows decidability of the problem which take ε as an input (in addition to γ, β and f) and distinguishes (γ, β)-approximable problems from $(\gamma - \varepsilon, \beta + \varepsilon)$-inapproximable problems. The running time of their decision procedure grows with $1/\varepsilon$. In contrast our distinguishability separates (γ, β)-approximability from "$\forall \varepsilon > 0$, $(\gamma - \varepsilon, \beta + \varepsilon)$-inapproximability" — so our algorithm does not require ε as an input - it merely takes γ, β and f as input. Indeed this difference is key to the understanding of approximation resistance. Due to the stronger form of our main theorem (Theorem 1.1), our characterization of approximation-resistance to sketching algorithms is explicit (decidable in PSPACE), whereas a decidable characterization of approximation-resistance in the polynomial time setting seems to be still open.

Our characterizations also seem to differ from the previous versions in terms of the features being exploited to distinguish the two classes. This leads to some strange gaps in our knowledge. For instance, it would be natural to suspect that (conditional) inapproximability in the polynomial time setting should also lead to (unconditional) inapproximability in the streaming setting. But we don't have a formal theorem proving this.[4] One (unfulfilling) consequence of this gap in knowledge is that we do not yet have an approximation-resistant problem, even to sketching algorithms, that is not covered by Theorem 1.3. In the polynomial time setting, Potechin [Pot19] gives a balanced linear threshold function that is approximation-resistant. Balanced linear threshold functions do not support one-wise independence and so that function would be a good candidate for a streaming-approximation-resistant function that is not covered by Theorem 1.3.

1.6 Overview of our analysis

At the heart of our characterization is a family of linear sketching algorithms for Max-CSP(f). We will describe this family soon, but the main idea of our proof is that if no algorithm in this family solves (γ, β)-Max-CSP(f), then we can extract a single pair of instances, roughly a γ-satisfiable "yes" instance and an at most β-satisfiable "no" instance, that certify this inability. We then show how this pair of instances can be exploited as gadgets in a negative result. Up to this part our approach resembles that in [Rag08] (though of course all the steps are quite different). The main difference is that we are able to use the structure of the algorithm and the lower bound construction to show that we can afford to consider only instances on k variables. (This step involves a non-trivial choice of definitions that we elaborate on shortly.) This bound on the number of variables allows us to get a very "decidable" separation between approximable and inapproximable problems. Specifically we show that distinction between approximable setting and the inapproximable one can be expressed by a quantified formula over the reals with a constant number of quantifiers over 2^k variables and equations — a problem that is known to be solvable in PSPACE. We give more details below.

[4] Of course, if this were false, it would be a breakthrough result giving a polynomial time (even log space) algorithm for the unique games!

Bias-based algorithms. For every $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k) \in \mathbb{R}^k$ we define the $\boldsymbol{\lambda}$-bias measure of an instance Ψ of $\textsf{Max-CSP}(f)$ as follows. Let p_{ij} denote the number of occurrences of the literal x_i as the jth variable in a constraint, and let n_{ij} denote the same quantity for the literal $-x_i$. Let $\textsf{bias}_{i,j} = \frac{1}{m}(p_{ij} - n_{ij})$. We define the $\boldsymbol{\lambda}$-bias of the ith variable to be a weighted sum of $\textsf{bias}_{i,j}$ as follows: $\textsf{bias}_{\boldsymbol{\lambda}}(\Psi)_i = \sum_{j=1}^{k} \lambda_j \textsf{bias}_{i,j}$. Let the bias vector of the instance Ψ be $\textsf{bias}_{\boldsymbol{\lambda}}(\Psi) = (\textsf{bias}_{\boldsymbol{\lambda}}(\Psi)_1, \ldots, \textsf{bias}_{\boldsymbol{\lambda}}(\Psi)_n)$. It turns out that the ability to estimate the ℓ_1 norm of a vector in the "turnstile setting" implies that for any given $\boldsymbol{\lambda}$ vector, we can estimate the ℓ_1 norm of $\textsf{bias}_{\boldsymbol{\lambda}}(\Psi)$ (to within a multiplicative factor of $(1 \pm \varepsilon)$ for arbitrarily small $\varepsilon > 0$) dynamically. We refer to an algorithm that aims to solve the (γ, β)-$\textsf{Max-CSP}(f)$ using only an estimate of the ℓ_1 norm of $\textsf{bias}_{\lambda}(\Psi)$ (for some $\boldsymbol{\lambda}$ based on f, γ, β) as a "bias-based algorithm". A priori it is not clear how to choose a $\boldsymbol{\lambda}$ vector for a given problem. The crux of our analysis is to identify two (bounded, closed) convex sets $K_\gamma^Y, K_\beta^N \subseteq \mathbb{R}^k$ such that if the two sets are disjoint then the hyperplane separating them gives us the desired $\boldsymbol{\lambda}$.

We now give some insight into the sets K_γ^Y and K_β^N. Roughly these sets capture properties of instances of $\textsf{Max-CSP}(f)$ on k *variables*, say $x_1, \ldots, x_k$. The instances we consider are special in that x_i always appears as the ith variable in every constraint: the only variability being in whether it appears positively or negatively. The set K_γ^Y consists of the bias vectors $\textsf{bias}_{\boldsymbol{\lambda}}(\Psi)$ of all instances Ψ that have $\textsf{val}_\Psi(1^k) \geq \gamma$, i.e., the assignment of all 1's satisfied γ fraction of the constraints of Ψ. The set K_β^N is similarly supposed to capture the biases $\textsf{bias}_{\boldsymbol{\lambda}}(\Psi)$ of instances Ψ for which the value is at most β. Determining exactly which assignments achieve this bounded value turns out to be subtle and we defer describing it here. But given our choice, our analysis roughly works as follows: Given an instance Ψ on n variables, we create a distribution $\mathcal{D}(\Psi) \in \Delta(\{-1,1\}^k)$ and its projection $\boldsymbol{\mu}$ onto $\mathbb{R}^k$ such that if Ψ is a YES instance, then $\boldsymbol{\mu}$ ends up being in K_γ^Y, while if Ψ is a NO instance, $\boldsymbol{\mu} \in K_\beta^N$. Most crucially, the ℓ_1 norm of $\textsf{bias}_{\boldsymbol{\lambda}}(\Psi)$ exactly corresponds to the distance from $\boldsymbol{\mu}$ to the hyperplane separating K_γ^Y and K_β^N, which allows us to distinguish the YES and NO cases. Details of the definition of sets can be found in Section 2 and the analysis of the algorithm can be found in Section 4.

Lower bounds via a new set of communication problems. Hardness results in streaming are usually obtained by appealing to lower bounds for communication complexity problems. In our case, both our lower bounds for sketching algorithms and general streaming algorithms are derived from lower bounds on the one-way communication complexity of a class of 2-player problem we call the "Randomized Mask Detection" (RMD) problems. (See Definition 5.2.) We first describe this problem and our results about this problem before returning to the streaming lower bounds.

An RMD problem is specified by two distributions $\mathcal{D}_Y$ and $\mathcal{D}_N$ supported on $\{-1,1\}^k$. In this problem Alice gets a vector $\mathbf{x}^* \in \{-1,1\}^n$ chosen uniformly at random which we view as a 2-coloring of the vertex set $[n]$, and Bob gets a random k-uniform hypermatching M with αn hyperedges on $[n]$, along with a vector $\mathbf{z} \in \{-1,1\}^{k\alpha n}$ whose distribution depends on whether we are in the YES case or NO case (here α is some small but positive constant). Specifically, $\mathbf{z}$ specifies the values of $\mathbf{x}^*$ on the vertices touched by M, but this information is hidden partially by picking for each edge (independently) a masking vector $\mathbf{b} \in \{-1,1\}^k$ and letting $\mathbf{z}$ for this edge be the information for $\mathbf{x}^*$ masked by (xor'ing with) $\mathbf{b}$. See Section 5.2 for a mathematically precise statement. The key difference between the YES instance and the NO instance is the distribution of $\mathbf{b}$: In the YES case, for every edge, the masking vector $\mathbf{b}$ is chosen independently according to some distribution $\mathcal{D}_Y$ supported on $\{-1,1\}^k$ whose marginals are in K_γ^N; and in the NO case, they come independently from the distribution $\mathcal{D}_N$ whose marginals are in K_β^Y. In the settings of interest to us K_γ^Y and K_β^N intersect and we ignore K_γ^N and K_β^Y, and just consider two arbitrary distributions $\mathcal{D}_Y$ and $\mathcal{D}_N$ with matching marginals. The technical meat of our negative result is proving that for an arbitrary pair of distributions $\mathcal{D}_Y$ and $\mathcal{D}_N$ with matching marginals, any one-way communication protocol with $o(\sqrt{n})$ communication from Alice to Bob has $o(1)$-advantage in distinguishing the YES and NO cases. See Theorem 5.3.

The proof of Theorem 5.3 starts with the work of Kapralov, Khanna, and Sudan [KKS15] which roughly shows that $(\mathcal{D}_Y, \mathcal{D}_N)$-RMD is hard on the special case where $\mathcal{D}_Y$ is uniform on $\{(1,1),(-1,-1)\}$ and $\mathcal{D}_N$ is uniform on $\{-1,1\}^2$. Strictly speaking their formalism is slightly different[5] — and one in which we are not able to express all our problems, but their proof for this case certainly applies to our formalism. The proof of [KKS15] is Fourier analytic, based on prior work of Gavinsky, Kempe, Kerenidis, Raz, and de Wolf [GKK$^+$09]. The first step of our analysis extends this Fourier analytic approach to the case of distributions over $\{-1,1\}^k$ for all values of k, and to all distributions $\mathcal{D}_Y$ and $\mathcal{D}_N$ that have *uniform marginals*. This is reported in Section 6.

The Fourier analytic proof does not seem to extend to the case where $\mathcal{D}_Y$ and $\mathcal{D}_N$ have arbitrary but matching marginals (at least we were unable to do so). To get the full case, we turn to reductions. Specifically we show that while we cannot directly prove the indistinguishability of general $\mathcal{D}_Y$ and $\mathcal{D}_N$ with matching marginals, we can use the indistinguishability for uniform marginals as a tool (via reductions) to show indistinguishability of some restricted pairs of distributions $(\mathcal{D}, \mathcal{D}')$. The key to the final result is that for any pair of distributions $\mathcal{D}_Y$ and $\mathcal{D}_N$ with matching marginals, there is a path from one to the other of finite length (our upper bound is $\mathsf{poly}(k!)$) such that every adjacent pair of distributions on the path is indistinguishable by our aforementioned reductions for restricted pairs. We remark that while $\mathcal{D}_Y$ and $\mathcal{D}_N$ are typically chosen to have interesting properties with respect to their value on various assignments, the intermediate distributions may not have any interesting properties for the underlying optimization problem! But the generality of the framework turns out to be a strength in that we can refer to these problems anyway and use their indistinguishability features. The path from $\mathcal{D}_Y$ to $\mathcal{D}_N$ allows us to use triangle-inequality for indistinguishability to get the final result on indistinguishability of RMD on distributions with matching marginals. Details of this part can be found in Section 7.

The actual lower bounds. Returning to the streaming problems, the rough idea is to use the two player lower bounds to derive lower bounds for a streaming version of the RMD, and then to reduce this problem to our target $\mathsf{Max\text{-}CSP}(f)$ problem. An instance of the streaming RMD problem with distributions $\mathcal{D}_Y, \mathcal{D}_N$ generates an Alice input $\mathbf{x}^*$ as in the RMD problem, and T sets of Bob inputs $(M_1, \mathbf{z}_1), \ldots, (M_T, \mathbf{z}_T)$ independently conditioned on $\mathbf{x}^*$. It then creates a stream concatenating the T Bob inputs and the streaming challenge is to determine if the underlying mask distribution is $\mathcal{D}_Y$ or $\mathcal{D}_N$. Note that in the streaming setting, there is no player corresponding to Alice, making the streaming RMD problem potentially harder to solve than the 2-player problem. Indeed our initial hope (and claim) was that the streaming RMD problem reduces to the 2-player RMD problem, but this hope turns out to be false. We are however able to establish such a reduction when $\mathcal{D}_Y$ and $\mathcal{D}_N$ have uniform marginals as claimed in Theorem 1.3. In fact we get a slight generalization which allows the two distribution $\mathcal{D}_Y$ and $\mathcal{D}_N$ to be derived from two distributions $\mathcal{D}'_Y$ and $\mathcal{D}'_N$

[5] In order to handle the general $\mathsf{Max\text{-}CSP}$ problem, in RMD we extend the previous framework with a more detailed encoding of the hypermatching M, and also allow for a general masking vector $\mathbf{b}$. Due to these extensions, we cannot immediately conclude hardness of RMD from previous results, and we prove it from scratch.

with uniform marginals, by padding with a common distribution $\mathcal{D}_0$ (see Theorem 2.11).

While the hope of reducing the 2-player RMD problem to the streaming problem fails in generality, it turns out that we can get a reduction to a T-player simultaneous communication version of RMD. (In this simultaneous communication version, the tth player gets $(M_t, \mathbf{z}_t)$ as input and needs to send a message to a referee who collects the messages from the T players and attempts to guess if the mask distribution is $\mathcal{D}_Y$ or $\mathcal{D}_N$.) Since a sketching algorithm can be turned into a protocol for the simultaneous communication game, we are able to show that whenever K_γ^Y and K_β^N intersect, any sketching algorithm that solves the (γ, β)-approximation version of $\mathsf{Max\text{-}CSP}(f)$ requires $\Omega(\sqrt{n})$ space.

Section 5 describes the various RMD problems discussed above and how they can be used to get proofs of Theorem 5.1 and Theorem 2.11.

1.7 Future questions and work

Some of the main questions left open in this work are listed below:

1. Does the characterization given by Theorem 2.3 actually hold for general streaming algorithms? Resolving this either way would be quite interesting.

2. Can the methods be extended to handle the case where the constraints come from a family of functions, rather than a single function? We believe this should be straightforward to achieve.

3. Can we further extend the results to the setting where the constraints are not placed on literals, but rather only on variables? Such an extension seems to require new ideas beyond those in this paper.

4. Can we extend the results to the non-Boolean setting, i.e., when the variables take on values from an arbitrary finite set, as opposed to $\{-1, 1\}$. We stress that both the positive and negative results in this paper exploit restrictions of the Boolean setting! In this direction, Guruswami and Tao [GT19] proved that $(1/p + \varepsilon)$-approximation for the unique games with alphabet size p requires $\tilde{\Omega}(\sqrt{n})$ space in the streaming setting.

5. Can the lower bound for the hard problems be improved to linear space lower bounds? Such an improvement was given by Kapralov and Krachun [KK19] for the $\mathsf{Max\text{-}2LIN}$ problem ($\mathsf{Max\text{-}CSP}(f)$ where $f(x, y) = x \oplus y$) in a technical tour-de-force. Extending this work to other optimization problems seems non-trivially challenging.

6. Finally, our work and all the questions above only consider the setting of single-pass streaming algorithms. Once this is settled, it would make sense to extend the analyses to multi-pass algorithms. While there are several multi-pass streaming algorithms and lower bounds (see, e.g., [Cha20, McG14, GM08] and references therein), we note that Assadi, Kol, Saxena, and Yu [AKSY20] recently suggested a multi-round version of the Boolean Hidden Hypermatching problem that allows to extend some previous single-pass results (including a lower bound for approximate $\mathsf{Max\text{-}2LIN}$) to the multi-pass setting.

1.8 Structure of rest of the paper

In Section 2, we describe our result in detail. In particular we build our convex set framework and give an explicit criterion to distinguish the easy and hard $\mathsf{Max\text{-}CSP}(f)$ problems. We also describe sufficient conditions for the hardness of some streaming problems in the streaming setting. Section 3 contains some of the preliminary background used in the rest of the paper. In Section 4, we describe and analyze our algorithm that yields our easiness result. In Section 5, we define the central family of communication problems that lie at the heart of our lower bounds and show how the communication complexity of this problem leads to the streaming space lower bounds claimed in Section 2. In Section 6, we first establish the desired lower bounds for a subclass of the problems using Fourier analytic methods. In Section 7, we establish reductions between various communication problems that allow us to prove our most general lower bounds.

2 Our Results

We start with some notation needed to state our results. We use $\mathbb{R}^{\geq 0}$ to denote the set of non-negative real numbers. For a finite set Ω, let $\Delta(\Omega)$ denote the space of all probability distributions over Ω, i.e.,

$$\Delta(\Omega) = \{\mathcal{D} : \Omega \to \mathbb{R}^{\geq 0} | \sum_{\omega \in \Omega} \mathcal{D}(\omega) = 1\}.$$

We view $\Delta(\Omega)$ as being contained in $\mathbb{R}^{|\Omega|}$. We use $X \sim \mathcal{D}$ to denote a random variable drawn from the distribution $\mathcal{D}$.

2.1 Key definitions

The main objects that allow us to derive our characterization are the space of distributions on constraints that either allow a large number of constraints to be satisfied, or only a few constraints to be satisfied. To see where the distributions come from, note that distributions of constraints over n variables can naturally be identified with instances of weighted constraint satisfaction problem (where the weight associated with a constraint is simply its probability). In what follows we will consider instances on exactly k variables $x_1, \ldots, x_k$. Furthermore all constraints will use x_i as the ith variable. Hence, a constraint on k variables is specified by $\mathbf{b} \in \{-1, 1\}^k$, specifying the constraint $f(b_1 x_1, \ldots, b_k x_k)$. Thus in what follows we will equate "instances on k variables" with distributions on $\{-1, 1\}^k$.

Given $0 \leq \beta \leq \gamma \leq 1$ we will consider two sets of instances/distributions. The first set $S_\gamma^Y = S_\gamma^Y(f)$ will be instances where γ fraction of the constraints are satisfied by the assignment 1^k. The second set $S_\beta^N = S_\beta^N(f)$ is a bit more subtle: it consists of instances where no "independent identical distribution" on the variables satisfies more that β-fraction of the clauses. To elaborate, recall that the only distributions on a single variable taking values in $\{-1, 1\}$ are the Bernoulli distributions. Let $\mathsf{Bern}(p)$ denote the distribution that takes the value 1 with probability p and -1 with probability $1 - p$. Then an instance belongs to S_β^N if for every p, when $(x_1, \ldots, x_k)$ gets a random assignment chosen according to $\mathsf{Bern}(p)^k$, the expected fraction of satisfied clauses is at most β. The following is our formal definition.

Definition 2.1 (Space of Yes/No Distributions)**.** *For $\gamma, \beta \in \mathbb{R}$, we define*

$$S_\gamma^Y = S_\gamma^Y(f) = \{\mathcal{D}_Y \in \Delta(\{-1,1\}^k) \mid \mathop{\mathbb{E}}_{\mathbf{b}\sim\mathcal{D}_Y}[f(\mathbf{b})] \geq \gamma\}$$

$$\text{and } S_\beta^N = S_\beta^N(f) = \{\mathcal{D}_N \in \Delta(\{-1,1\}^k) \mid \mathop{\mathbb{E}}_{\mathbf{b}\sim\mathcal{D}_N}\mathop{\mathbb{E}}_{\mathbf{a}\sim\mathsf{Bern}(p)^k}[f(\mathbf{b}\odot\mathbf{a})] \leq \beta, \forall p\}\,.$$

For $\gamma > \beta$ the sets S_γ^Y and S_β^N are clearly disjoint. But their marginals, when projected to single coordinates need not be, and this is the crux of our characterization. In what follows, we define sets K_γ^Y and K_β^N to be the marginals of the distributions in S_γ^Y and S_β^N respectively. For a distribution $\mathcal{D} \in \Delta(\{-1,1\}^k)$, let $\boldsymbol{\mu}(\mathcal{D})$ denote its marginals, i.e., $\boldsymbol{\mu}(\mathcal{D}) = (\mu_1,\ldots,\mu_k)$ where $\mu_i = \mathbb{E}_{\mathbf{b}\sim\mathcal{D}}[b_i]$.

Definition 2.2 (Marginals of Yes/No Distributions)**.** *For $\gamma, \beta \in \mathbb{R}$, we define*

$$K_\gamma^Y = K_\gamma^Y(f) = \{\ \boldsymbol{\mu}(\mathcal{D}_Y) \mid \mathcal{D}_Y \in S_\gamma^Y\}$$

$$\text{and } K_\beta^N = K_\beta^N(f) = \{\ \boldsymbol{\mu}(\mathcal{D}_N) \mid \mathcal{D}_N \in S_\beta^N\}\,.$$

With the two definitions above in hand we are ready to describe our characterizations of easy vs. hard approximation versions of $\mathsf{Max\text{-}CSP}(f)$.

2.2 Results on sketching algorithms

The following theorem now formalizes the informal statement that low space sketching algorithms (see Definition 3.3) can only capture the marginals of distributions.

Theorem 2.3 (Dichotomy for Sketching Algorithms)**.** *For every function $f : \{-1,1\}^k \to \{0,1\}$ and for every $0 \leq \beta < \gamma \leq 1$, the following hold:*

1. *If $K_\gamma^Y(f) \cap K_\beta^N(f) = \emptyset$, then (γ,β)-$\mathsf{Max\text{-}CSP}(f)$ admits a a probabilistic linear sketching algorithm (see Definition 3.3) that uses $O(\log n)$ space on instances on n variables.*
2. *If $K_\gamma^Y(f) \cap K_\beta^N(f) \neq \emptyset$, then for every $\varepsilon > 0$, every sketching algorithm for $(\gamma-\varepsilon,\beta+\varepsilon)$-$\mathsf{Max\text{-}CSP}(f)$ requires $\Omega(\sqrt{n})$ space[6] on instances on n variables. Furthermore, if $\gamma = 1$, then every sketching algorithm for $(1,\beta+\varepsilon)$-$\mathsf{Max\text{-}CSP}(f)$ requires $\Omega(\sqrt{n})$ space.*

Proof of Theorem 2.3. Part (1) of the theorem is restated and proved as Theorem 4.1 in Section 4. Part (2) is proved as Theorem 5.1 in Section 5. □

We now turn to the implications of this theorem. First, to get Theorem 1.1 from Theorem 2.3, we need to show that the question "Is $K_\gamma^Y \cap K_\beta^N = \emptyset$?" can be decided in polynomial space. To this end, we first make the following observation.

Lemma 2.4. *For every $\beta, \gamma \in [0,1]$ the sets $S_\gamma^Y, S_\beta^N, K_\gamma^N$ and K_β^Y are bounded, closed and convex. Furthermore, $K_\gamma^Y \cap K_\beta^N = \emptyset$ can be expressed in the quantified theory of the reals with 2 quantifier alternations, $O(2^k)$ variables, and polynomials of degree at most $k+1$.*

[6] The constant hidden in the Ω notation may depend on k and ε.

Proof. We start by considering the sets S_γ^Y and S_β^N. It is straightforward to see that S_γ^Y is a bounded and convex polytope in $\mathbb{R}^{2^k}$. S_β^N is a bit more subtle due to the universal quantification over $p \in [0,1]$. It is now specified by infinitely many linear inequalities in $\mathbb{R}^{2^k}$ and so is still a bounded and convex set (though not necessarily a polytope). K_γ^Y (resp. K_β^N) is obtained by a linear projection from $\mathbb{R}^{2^k}$ to $\mathbb{R}^k$. So K_γ^Y is a bounded, closed, and convex polytope in $\mathbb{R}^k$, while K_β^N is still a bounded, closed, and convex set.

To get an intersection detection algorithm we use one more property. Note that for variable p, the condition $\mathbb{E}_{\mathbf{a}\sim\mathcal{D}_N}\mathbb{E}_{\mathbf{b}\sim\mathsf{Bern}(p)^k}[f(\mathbf{b}\odot\mathbf{a})] \leq \beta$ is a polynomial inequality in p of degree at most k, with coefficients that are linear forms in $\mathcal{D}_N(\mathbf{b})$, $\mathbf{b} \in \{-1,1\}^k$. This allows us to express the condition $K_\gamma^Y \cap K_\beta^N \neq \emptyset$ using the following system of quantified polynomial inequalities:

$$\exists \quad \mathcal{D}_Y, \mathcal{D}_N \in \mathbb{R}^{2^k},\ \forall p \in [0,1] \text{ s.t.}$$

$$\mathcal{D}_Y, \mathcal{D}_N \text{ are distributions,} \tag{2.5}$$

$$\forall i \in [k],\ \mathop{\mathbb{E}}_{\mathbf{b}\sim\mathcal{D}_Y}[b_i] = \mathop{\mathbb{E}}_{\mathbf{b}\sim\mathcal{D}_N}[b_i], \tag{2.6}$$

$$\mathop{\mathbb{E}}_{\mathbf{b}\sim\mathcal{D}_Y}[f(\mathbf{b})] \geq \gamma, \tag{2.7}$$

$$\mathop{\mathbb{E}}_{\mathbf{b}\sim\mathcal{D}_N}\ \mathop{\mathbb{E}}_{\mathbf{a}\sim\mathsf{Bern}(p)^k}[f(\mathbf{a}\odot\mathbf{b})] \leq \beta. \tag{2.8}$$

Note that Equations (2.5), (2.6) and (2.7) are just linear inequalities in the variables $\mathcal{D}_Y, \mathcal{D}_N$ and do not depend on p. As noticed above Equation (2.8) is an inequality in p, and $\mathcal{D}_N$, of degree k in p, and 1 in $\mathcal{D}_N$. We thus get that the intersection problem can be expressed in the quantified theory of the reals by an expression with two quantifier alternations, 2^k variables and $O(2^k)$ polynomial inequalities, with polynomials of degree at most $k+1$. (Most of the inequalities are of the form $\mathcal{D}_Y(\mathbf{b}) \geq 0$ or $\mathcal{D}_N(\mathbf{b}) \geq 0$. Only $O(k)$ inequalities are not of that form; and of these, only one is non-linear.) □

The quantified theory of the reals is known to be solvable in PSPACE. In particular we use the following theorem.

Theorem 2.9 ([BPR06, Theorem 14.11, see also Remark 13.10])**.** *The truth of a quantified formula with w quantifier alternations over K variables and polynomial (potentially strict) inequalities can be decided in space $K^{O(w)}$ and time $2^{K^{O(w)}}$.*

(Specifically, Theorem 14.11 in [BPR06] asserts the time complexity above, and Remark 13.10 yields the space complexity.)

Theorem 1.1 now follows immediately.

Proof of Theorem 1.1. Theorem 2.3 asserts that the (γ,β)-approximation version of $\mathsf{Max\text{-}CSP}(f)$ is easy if and only if $K_\gamma^Y \cap K_\beta^N = \emptyset$. Lemma 2.4 asserts that this condition is in turn expressible in the quantified theory of the reals with 2 quantifier alternations. Finally Theorem 2.9 asserts that this can be decided in polynomial space. The theorem follows. □

We note that the literature on approximation algorithms usually considers a single parameter version of the problem. In our context we would say that an algorithm A is a α-approximation algorithm for $\mathsf{Max\text{-}CSP}(f)$ if for every instance Ψ, we have

$$\alpha \cdot \mathsf{val}_\Psi \leq A(\Psi) \leq \mathsf{val}_\Psi\,.$$

The following proposition converts our main theorem in terms of this standard notion.

Proposition 2.10. *Fix $f : \{-1,1\}^k$ and let K_γ^Y and K_β^N denote the space of marginals for this function f. Let*

$$\alpha = \inf_{\beta \in [0,1]} \left\{ \sup_{\gamma \in (\beta,1] \text{ s.t } \mathrm{K}_\gamma^\mathrm{Y} \cap \mathrm{K}_\beta^\mathrm{N} = \emptyset} \{\beta/\gamma\} \right\}.$$

Then for every $\varepsilon > 0$, there is an $(\alpha - \varepsilon)$-approximation algorithm for $\mathsf{Max\text{-}CSP}(f)$ that uses $O(\log n)$ space. Conversely every $(\alpha + \varepsilon)$-approximation algorithm for $\mathsf{Max\text{-}CSP}(f)$ requires $\Omega(\sqrt{n})$ space.

Proof. For the positive result, let $\tau \triangleq \varepsilon \cdot \rho(f)/2$, where $\rho(f) = 2^{-k} \sum_{\mathbf{a} \in \{-1,1\}^k} f(\mathbf{a})$ is the fraction of clauses satisfied by a random assignment. Let

$$A_\tau = \{(i\tau, j\tau) \in [0,1]^2 \mid i,j \in \mathbb{Z}^{\geq 0}, i > j, K_{i\tau}^Y \cap K_{j\tau}^N = \emptyset\}.$$

By Theorem 2.3, for every $(\gamma, \delta) \in A_\tau$ there is a $O(\log n \log(1/\tau))$-space algorithm for (γ,β)-$\mathsf{Max\text{-}CSP}(f)$ with error probability $1/(10\tau^2)$, which we refer to as the (γ,β)-distinguisher below. In the following we consider the case where all $O(\tau^{-2})$ distinguishers output correct answers, which happens with probability at least $2/3$.

Our $O(\tau^{-2}\log(1/\tau)\log n) = O(\log n)$ space $(\alpha - \varepsilon)$-approximation algorithm for $\mathsf{Max\text{-}CSP}(f)$ is the following: On input Ψ, run in parallel all the (γ,β)-distinguishers on Ψ, for every $(\gamma,\beta) \in A_\tau$. Let

$$\beta_0 = \arg\max_\beta [\exists \gamma \text{ such that the } (\gamma,\beta)\text{-distinguisher outputs YES on } \Psi]\,.$$

Output $\beta' = \max\{\rho(f), \beta_0\}$.

We now prove that this is an $(\alpha - \varepsilon)$-approximation algorithm. First note that by the correctness of the distinguisher we have $\beta' \leq \mathsf{val}_\Psi$. Let γ_0 be the smallest multiple of τ satisfying $\gamma_0 \geq (\beta_0 + \tau)/\alpha$. By the definition of α, we have that $K_{\gamma_0}^Y \cap K_{\beta_0+\tau}^N = \emptyset$. So $(\gamma_0, \beta_0 + \tau) \in A_\tau$ and so the $(\gamma_0, \beta_0 + \tau)$-distinguisher must have output NO on Ψ (by the maximality of β_0). By the correctness of this distinguisher we conclude $\mathsf{val}_\Psi \leq \gamma_0 \leq (\beta_0 + \tau)/\alpha + \tau \leq (\beta' + \tau)/\alpha + \tau$. We now verify that $(\beta' + \tau)/\alpha + \tau \leq \beta'/(\alpha - \varepsilon)$ and this gives us the desired approximation guarantee. We have

$$(\beta' + \tau)/\alpha + \tau \leq (\beta' + 2\tau)/\alpha \leq (\beta'/\alpha) \cdot (1 + 2\tau/\rho(f)) = (\beta'/\alpha)(1 + \varepsilon) \leq (\beta'/(\alpha(1 - \varepsilon))),$$

where the first inequality uses $\alpha \leq 1$, the second uses $\beta' \geq \rho(f)$, the equality comes from the definition of τ and the final inequality uses $(1+\varepsilon)(1-\varepsilon) \leq 1$. This concludes the positive result.

The negative result is simpler. Given γ, β with $\beta/\gamma \geq \alpha + \varepsilon$, we can use an $(\alpha+\varepsilon)$-approximation algorithm A to solve the (γ,β)-$\mathsf{Max\text{-}CSP}(f)$, by outputting YES if $A(\Psi) \geq \beta$ and NO otherwise. □

2.2.1 Approximation resistance

We now turn to Corollary 1.2. Recall that for a function $f : \{-1,1\}^k \to \{0,1\}$, we define $\rho(f) = 2^{-k} \cdot |\{\mathbf{a} \in \{-1,1\}^k : f(\mathbf{a}) = 1\}|$ to be the probability that a uniformly random assignment satisfies f. Recall further that f is approximation-resistant if for every $\varepsilon > 0$, the $(1, \rho(f) + \varepsilon)$-approximation version of $\mathsf{Max\text{-}CSP}(f)$ requires polynomial space.

Proof of Corollary 1.2. By Theorem 2.3 we have that $\mathsf{Max\text{-}CSP}(f)$ is approximation-resistant if and only if $K_1^Y \cap K_{\rho(f)+\varepsilon}^N \neq \emptyset$ for every $\varepsilon > 0$. In turn, this is equivalent to saying $\mathsf{Max\text{-}CSP}(f)$ is approximation resistant if and only if $K_1^Y \cap K_{\rho(f)}^N \neq \emptyset$. If $K_1^Y \cap K_{\rho(f)}^N = \emptyset$, then by the property that these sets are closed, we have that there must exist $\varepsilon > 0$ such that $K_{1-\varepsilon}^Y \cap K_{\rho(f)+\varepsilon}^N = \emptyset$. In turn this implies, again by Theorem 2.3, that the $(1-\varepsilon, \rho(f)+\varepsilon)$-approximation version of $\mathsf{Max\text{-}CSP}(f)$ can be solved by a linear sketching algorithm with $O(\log n)$ space. Finally, from Lemma 2.4 and Theorem 2.9 the condition "Is $K_1^Y \cap K_{\rho(f)}^N = \emptyset$?" can be checked in polynomial space. □

2.3 Lower bounds in the streaming setting

For two broad sets of special cases we are able to get lower bounds =in the streaming setting with general streaming algorithms where the lower bounds match the upper bounds derived using linear sketches. To define these classes we need some definitions.

We say that a distribution $\mathcal{D} \in \Delta(\{-1,1\}^k)$ is *one-wise-independent* if $\boldsymbol{\mu}(\mathcal{D}) = 0^k$. We say that a pair of distributions $(\mathcal{D}_Y, \mathcal{D}_N)$ form a *padded one-wise pair* if there exists $\tau \in [0,1]$ and distributions $\mathcal{D}_0, \mathcal{D}'_Y$ and $\mathcal{D}'_N$ such that (1) $\mathcal{D}'_Y$ and $\mathcal{D}'_N$ are one-wise independent and (2) $\mathcal{D}_Y = \tau\mathcal{D}_0 + (1-\tau)\mathcal{D}'_Y$ and $\mathcal{D}_N = \tau\mathcal{D}_0 + (1-\tau)\mathcal{D}'_N$.

Our main lower bound in the streaming setting asserts that if $S_\gamma^Y \times S_\beta^N$ contains a padded one-wise pair $(\mathcal{D}_Y, \mathcal{D}_N)$ then (γ, β)-$\mathsf{Max\text{-}CSP}(f)$ requires $\Omega(\sqrt{n})$-space.

Theorem 2.11 (Streaming lower bound)**.** *For every function $f : \{-1,1\}^k \to \{0,1\}$ and for every $0 \le \beta < \gamma \le 1$, if there exists a padded one-wise pair of distributions $\mathcal{D}_Y \in S_\gamma^Y$ and $\mathcal{D}_N \in S_\beta^N$ then, for every $\varepsilon > 0$, then every streaming algorithm that solves $(\gamma-\varepsilon, \beta+\varepsilon)$-$\mathsf{Max\text{-}CSP}(f)$ requires $\Omega(\sqrt{n})$ space. Furthermore, if $\gamma = 1$ then $(1, \beta+\varepsilon)$-$\mathsf{Max\text{-}CSP}(f)$ requires $\Omega(\sqrt{n})$ space.*

Theorem 2.11 is proved in Section 5.2.4. As stated above the theorem is more complex to apply than, say, Theorem 2.3, owing to the fact that the condition for hardness depends on the entire distribution (and the sets S_γ^Y and S_β^N) rather than just marginals (or the sets K_γ^Y and K_β^N). However it can be used to derive some clean results, specifically Theorem 1.3 and Theorem 1.4, that do depend only on the marginals. We prove these (assuming Theorem 2.11) below.

Recall that we say that a function f *supports one-wise independence* if there exists a one-wise-independent distribution D supported on the satisfying assignments to f. Note that this is equivalent to saying $0^k \in K_1^Y$. Theorem 1.3 asserts that every function that supports a one-wise independent distribution is approximation resistant in the streaming setting.

Proof of Theorem 1.3. Let $\rho = \rho(f)$. We first show that the vector 0^k belongs to both K_1^Y and K_ρ^N. We then note that this implies the existence of a padded one-wise pair of distributions $\mathcal{D}_Y \in S_1^Y$ and $\mathcal{D}_N \in S_\rho^N$ allowing us to apply Theorem 2.11 to get the theorem.

Let $\mathcal{D}_Y$ be the distribution proving that f supports a one-wise independent distribution, i.e., $\mathcal{D}_Y$ is supported on $f^{-1}(1)$ and satisfies $\mathbb{E}_{\mathbf{b} \in \mathcal{D}_Y}[b_i] = 0$ for every $i \in [k]$. It follows that $\mathcal{D}_Y \in S_1^Y$ and $0^k \in K_1^Y$. Let $\mathcal{D}_N$ be the uniform distribution on $\{-1,1\}^k$. Note that for every $\mathbf{a} \in \{-1,1\}^k$ we have $\mathbf{a} \odot \mathbf{b}$ is uniformly distributed over $\{-1,1\}^k$ if $\mathbf{b} \sim \mathcal{D}_N$. Consequently, for every $\mathbf{a}$ we get $\mathbb{E}_{\mathbf{b} \sim \mathcal{D}_N}[f(\mathbf{b} \odot \mathbf{a})] = \rho$, and so for every $p \in [0,1]$, we have

$$\mathop{\mathbb{E}}_{\mathbf{a} \sim \mathsf{Bern}(p)^k} \mathop{\mathbb{E}}_{\mathbf{b} \sim \mathcal{D}_N} [f(\mathbf{b} \circ \mathbf{a})] = \rho.$$

We conclude the $\mathcal{D}_N \in S_\rho^N$ and so $0^k \in K_\rho^N$.

Now by definition we have that $\mathcal{D}_Y$ and $\mathcal{D}_N$ form a padded one-wise pair (using $\mathcal{D}'_Y = \mathcal{D}_Y$, $\mathcal{D}'_N = \mathcal{D}_N$ and $\tau = 0$) and so Theorem 2.11 is applicable to show that $\mathsf{Max\text{-}CSP}(f)$ is not $(1, \rho - \varepsilon)$-approximable and so is approximation-resistant. □

We now turn to the proof of Theorem 1.4. Indeed we prove a more detailed statement along the lines of Theorem 2.3 in this case. For this part we use the fact, proved below, that any pair of distributions $\mathcal{D}_Y, \mathcal{D}_N \in \Delta(\{-1,1\}^2)$ with matching marginals form a padded one-wise pair.

Proposition 2.12. *If $\mathcal{D}_Y, \mathcal{D}_N \in \Delta(\{-1,1\}^2)$ satisfy $\mu(\mathcal{D}_Y) = \mu(\mathcal{D}_N)$ then $(\mathcal{D}_Y, \mathcal{D}_N)$ form a padded one-wise pair.*

Proof. Let $\mathcal{D}_Y = (p_{1,1}, p_{1,-1}, p_{-1,1}, p_{-1,-1})$ where $p_{i,j}$ denotes the probability $\Pr_{(a,b)\sim\mathcal{D}_Y}[a = i, b = j]$. If $\mathcal{D}_N$ has matching marginals with $\mathcal{D}_Y$ then there exists a $\delta \in [-1,1]$ such that $\mathcal{D}_N = (p_{1,1} - \delta, p_{1,-1} + \delta, p_{-1,1} + \delta, p_{-1,-1} - \delta)$. Assume without loss of generality that $\delta \geq 0$. Let $\tau = 1 - 2\delta$, $\mathcal{D}_0 = \frac{1}{1-2\delta}(p_{1,1} - \delta, p_{1,-1}, p_{-1,1}, p_{-1,-1} - \delta)$, $\mathcal{D}'_Y = (1/2, 0, 0, 1/2)$ and $\mathcal{D}'_N = (0, 1/2, 1/2, 0)$. It can be verified that $\mathcal{D}'_Y$ and $\mathcal{D}'_N$ are one-wise independent, $\mathcal{D}_Y = \tau\mathcal{D}_0 + (1-\tau)\mathcal{D}'_Y$ and $\mathcal{D}_N = \tau\mathcal{D}_0 + (1-\tau)\mathcal{D}'_N$, thus proving the proposition. □

Combining Proposition 2.12 and Theorem 2.11 we immediately get the following theorem, which in turn implies Theorem 1.4.

Theorem 2.13. *For every function $f : \{-1,1\}^2 \to \{0,1\}$, and for every $0 \leq \beta < \gamma \leq 1$, the following hold:*

1. *If $K_\gamma^Y(f) \cap K_\beta^N(f) = \emptyset$, then (γ,β)-$\mathsf{Max\text{-}CSP}(f)$ admits a linear sketching algorithm that uses $O(\log n)$ space.*
2. *If $K_\gamma^Y(f) \cap K_\beta^N(f) \neq \emptyset$, then for every $\varepsilon > 0$, every streaming algorithm that solves $(\gamma - \varepsilon, \beta + \varepsilon)$-$\mathsf{Max\text{-}CSP}(f)$ requires $\Omega(\sqrt{n})$ space[7]. Furthermore, if $\gamma = 1$, then $(1, \beta + \varepsilon)$-$\mathsf{Max\text{-}CSP}(f)$ requires $\Omega(\sqrt{n})$ space.*

Proof. Part (1) is simply the specialization of Part (1) of Theorem 2.13 to the case $k = 2$. For Part (2), suppose $\mu \in K_\gamma^Y \cap K_\beta^N$. Let $\mathcal{D}_Y \in S_\gamma^Y$ and $\mathcal{D}_N \in S_\beta^N$ be distributions such that $\mu(\mathcal{D}_Y) = \mu(\mathcal{D}_N) = \mu$. Then by Proposition 2.12 we have that $\mathcal{D}_Y$ and $\mathcal{D}_N$ form a padded one-wise pair, and so Theorem 2.11 can be applied to get Part (2). □

2.4 Examples

We illustrate the applicability of our results with two examples. The first is of the specific function $\mathsf{Max\text{-}2AND}$, i.e., $\mathsf{Max\text{-}CSP}(f)$ for $f(a,b) = a \wedge b$, i.e., $f(a,b) = 1$ if and only if $a = b = 1$. Here, since we are working with $k = 2$ we get to use the stronger separation from Theorem 2.13 (with general streaming lower bounds).

[7] The constant hidden in the Ω notation may depend on k and ε.

Example 1 (Max-2AND).

For the function $f : \{-1,1\} \to \{0,1\}$ given by $f(1,1) = 1$ and $f(a,b) = 0$ otherwise, we would like to calculate the quantity $\inf_\beta \sup_{\gamma | K^Y_\gamma \cap K^N_\beta = \emptyset} \beta/\gamma$. We first note that due to the symmetry of f, we have K^Y_γ is symmetric, i.e., $(\mu_1, \mu_2) \in K^Y_\gamma \Leftrightarrow (\mu_2, \mu_1) \in K^Y_\gamma$. Similarly with K^β_N. Further by convexity of K^Y_γ and K^N_β we get there exists a pair $(\mu_1, \mu_2) \in K^Y_\gamma \cap K^N_\beta$ if and only if there exists a μ such that $(\mu, \mu) \in K^Y_\gamma \cap K^N_\beta$. We now define two functions that will help us answer the question if such a μ exists. Let

$$\gamma(\mu) := \max_{\gamma \,|\, (\mu,\mu) \in K^Y_\gamma} \{\gamma\} \quad \& \quad \beta(\mu) := \min_{\beta \,|\, (\mu,\mu) \in K^N_\beta} \{\beta\}.$$

Note that $K^Y_\gamma \cap K^N_\beta \neq \emptyset$ if and only if there exists a μ such that $\gamma \leq \gamma(\mu)$ and $\beta \geq \beta(\mu)$. With some minimal calculations for $\gamma(\mu)$ and some slightly more involved ones for $\beta(\mu)$ we can show

$$\gamma(\mu) = \frac{1+\mu}{2} \quad \text{and} \quad \beta(\mu) = \begin{cases} |\mu| & , |\mu| \geq \frac{1}{3} \\ \frac{(1-|\mu|)^2}{4(1-2|\mu|)} & , \text{else.} \end{cases}$$

With the above in hand we can analyze when $K^Y_\gamma \cap K^N_\beta = \emptyset$.
First, when $\beta < 1/4$ then $K^N_\beta = \emptyset$. Next, when $\gamma \leq 1/2$, note that $(0,0) \in K^Y_\gamma$ and hence $K^Y_\gamma \cap K^N_\beta \neq \emptyset$ for all $\beta \geq 1/4$. When $\gamma > 1/2$, we set $\mu = 2\gamma - 1$ which leads to $|\mu| = \mu = 2\gamma - 1$ and so we get

$$\beta(\mu)\big|_{\mu = 2\gamma - 1} = \begin{cases} \frac{(1-\gamma)^2}{3-4\gamma} & , 1/2 \leq \gamma < 2/3 \\ 2\gamma - 1 & , 2/3 \leq \gamma\,. \end{cases}$$

We thus get that the set $H^\cap \triangleq \{(\gamma, \beta) \in [0,1]^2 | K^Y_\gamma \cap K^N_\beta \neq \emptyset\}$ (of *hard* problems) is given by:

$$\begin{aligned} H^\cap = \quad & \left[0, \frac{1}{2}\right] \times \left[\frac{1}{4}, 1\right] \\ \cup \quad & \left\{(\gamma, \beta) | \gamma \in \left[\frac{1}{2}, \frac{2}{3}\right], \beta \in \left[\frac{(1-\gamma)^2}{3-4\gamma}, 1\right]\right\} \\ \cup \quad & \left\{(\gamma, \beta) | \gamma \in \left[\frac{2}{3}, 1\right], \beta \in [2\gamma - 1, 1]\right\}. \end{aligned}$$

The quantity $\alpha(\beta) = \sup_{\gamma \in [\beta,1] \,|\, K^Y_\gamma \cap K^N_\beta = \emptyset} \beta/\gamma$ is minimized at $\beta = 4/15$.
At this point $\alpha = 4/9$, which is consistent with the findings in [CGV20] for the Max-2AND problem. Our more refined analysis also shows that $\alpha(\beta)$ approaches 1 as $\beta \to 1$ (suggesting that "almost-satisfiable" instances are better approximated).

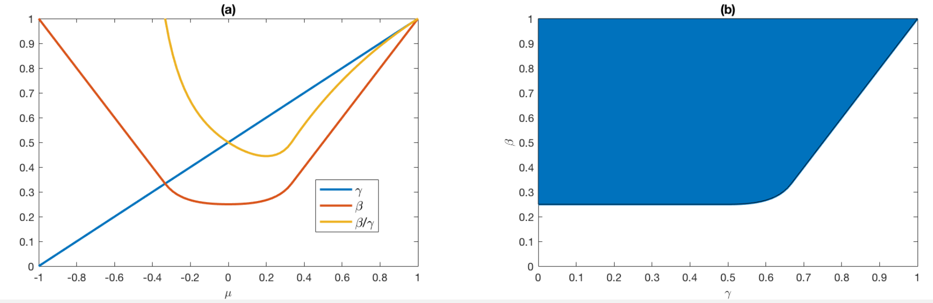

Figure 1: (a) A plot of γ, β, and β/γ with respect to μ. (b) A plot of $H^{\cap}$.

The second example we consider includes an entire family of functions.

Lemma 2.14 (one-wise independence implies approximation resistance)**.** *For a symmetric function $f : \{-1,1\}^k \to \{0,1\}$, Max-CSP$(f)$ is approximation resistant if and only if it supports a one-wise independent distribution.*

Proof. One direction of the implication directly follows from Corollary 1.2. For the other direction, we use Fourier analysis. The necessary definitions are included in Section 3.4. A symmetric function f is given by a set of "levels" $L = \{\ell_1, \ldots, \ell_t\} \subseteq \{-k, \ldots, k\}$ such that $f(a_1, \ldots, a_k) = 1$ if and only if $\|\mathbf{a}\|_1 = \sum_{i=1}^k a_i \in L$. If L contains 0, or if L contains both positive and negative elements, then f supports a one-wise independent distribution.[8] So we conclude L contains only positive elements or only negative elements. Without loss of generality we consider the case where L contains only positive elements.

Let $\rho = \rho(f)$, first note that both K_1^Y and K_ρ^N are symmetric since f is symmetric. Thus, by the convexity of the sets, it suffices to consider vectors of the form $\mu^k = (\mu, \mu, \ldots, \mu)$ in K_1^Y and K_ρ^N. Since L contains only positive elements, it follows that for $\mu^k \in K_1^Y$, we must have $\mu > 0$. To prove that Max-CSP(f) is not approximation resistant, it suffices to show that for $\mu > 0$, μ^k is not contained in K_ρ^N. Consider a distribution $\mathcal{D} \in S_\rho^N$ with $\mu(\mathcal{D}) = \mu^k$. It can be shown by elementary Fourier analysis that if $\mathbf{a} \sim \mathsf{Bern}(1/2 + \varepsilon)^k$ and $\mathbf{b} \sim \mathcal{D}$ then

$$\mathop{\mathbb{E}}_{\mathbf{b}\sim\mathcal{D}} \mathop{\mathbb{E}}_{\mathbf{a}\sim\mathsf{Bern}(p)^k} [f(\mathbf{b} \odot \mathbf{a})] = \rho + \Omega(\mu\tau\varepsilon) - O(\varepsilon^2),$$

where τ is the sum of the first level Fourier coefficients of f (i.e., $\tau = \sum_{||w||_1=1} \hat{f}(w)$), and the $\Omega(\cdot)$ and $O(\cdot)$ notations hide constants depending on f and $\mathcal{D}$, but not on $\varepsilon > 0$. Due to the symmetry of f, all the first level Fourier coefficients are equal, and due to the positivity of L, all these coefficients are positive. It follows that for some sufficiently small $\varepsilon > 0$, the expected probability of satisfying a constraint is strictly larger than ρ thus proving $\mu^k \notin K_\rho^N$. We conclude $K_1^Y \cap K_\rho^N = \emptyset$, and so Max-CSP$(f)$ is not approximation-resistant. □

[8] Indeed, if $\ell_1, \ell_2 \in L$, where $\ell_1 < 0$ and $\ell_2 > 0$, then a distribution $\mathcal{D}$ that with probability $p = \ell_2/(\ell_2 - \ell_1)$ samples a random $\mathbf{a}$ of Hamming weight $\|\mathbf{a}\|_1 = \ell_1$ and with probability $1 - p$ samples a random $\mathbf{a}$ of weight $\|\mathbf{a}\|_1 = \ell_2$ is one-wise independent and is supported on $f^{-1}(1)$.

3 Preliminaries

We will follow the convention that n denotes the number of variables in the CSP as well as the communication game, m denotes the number of constraints in the CSP, and k denotes the arity of the CSP. We use $\mathbb{N}$ to denote the set of natural numbers $\{1, 2, 3, \ldots\}$ and use $[n]$ to denote the set $\{1, 2, \ldots, n\}$. By default, the Boolean variable in this paper takes value in $\{-1, 1\}$.

For variables of a vector form, we write them in boldface, *e.g.,* $\mathbf{x} \in \{-1, 1\}^n$, and its i-th entry is written without boldface, *e.g.,* x_i. For variable being a vector of vectors, we write it, for example, as $\mathbf{b} = (\mathbf{b}(1), \mathbf{b}(2), \ldots, \mathbf{b}(m))$ where $\mathbf{b}(i) \in \{-1, 1\}^k$. The j-th entry of the i-th vector of $\mathbf{b}$ is then written as $\mathbf{b}(i)_j$. Let $\mathbf{x}$ and $\mathbf{y}$ be two vectors of the same length, $\mathbf{x} \odot \mathbf{y}$ denotes the entry-wise product of them.

For every $p \in [0, 1]$, $\mathsf{Bern}(p)$ denotes the Bernoulli distribution that takes value 1 with probability p and takes value -1 with probability $1 - p$.

3.1 Approximate Constraint Satisfaction

Let $f : \{-1, 1\}^k \to \{0, 1\}$ be a Boolean constraint function of arity k and $x_1, \ldots, x_n$ be variables. A constraint C consists of $\mathbf{j} = (j_1, \ldots, j_k) \in [n]^k$ and $\mathbf{b} = (b_1, \ldots, b_k) \in \{-1, 1\}^k$ where the j_i's are distinct. The constraint C reads as requiring $f(\mathbf{b} \odot \mathbf{x}|_{\mathbf{j}}) = f(b_1 x_{j_1}, \ldots, b_k x_{j_k}) = 1$. A $\mathsf{Max\text{-}CSP}(f)$ instance Ψ contains m constraints $C_1, \ldots, C_m$ with non-negative weights $w_1, \ldots, w_m$ where $C_i = (\mathbf{j}(i), \mathbf{b}(i))$ and $w_i \in \mathbb{R}$ for each $i \in [m]$. For an assignment $\boldsymbol{\sigma} \in \{-1, 1\}^n$, the value $\mathsf{val}_\Psi(\boldsymbol{\sigma})$ of $\boldsymbol{\sigma}$ on Ψ is the fraction of weight of constraints satisfied by $\boldsymbol{\sigma}$, i.e., $\mathsf{val}_\Psi(\boldsymbol{\sigma}) = \frac{1}{W} \sum_{i \in [m]} w_i \cdot f(\mathbf{b}(i) \odot \boldsymbol{\sigma}|_{\mathbf{j}(i)})$, where $W = \sum_{i=1}^m w_i$. The optimal value of Ψ is defined as $\mathsf{val}_\Psi = \max_{\boldsymbol{\sigma} \in \{-1,1\}^n} \mathsf{val}_\Psi(\boldsymbol{\sigma})$. The approximation version of $\mathsf{Max\text{-}CSP}(f)$ is defined as follows.

Definition 3.1 ((γ, β)-$\mathsf{Max\text{-}CSP}(f)$)**.** *Let* $f \colon \{-1, 1\}^k \to \{0, 1\}$ *be a constraint function and* $0 \leq \beta < \gamma \leq 1$. *For each* $m \in \mathbb{N}$, *let* $\Gamma_m = \{\Psi = (C_1, \ldots, C_m; w_1, \ldots, w_m) \mid \mathsf{val}_\Psi \geq \gamma\}$ *and* $B_m = \{\Psi = (C_1, \ldots, C_m; w_1, \ldots, w_m) \mid \mathsf{val}_\Psi \leq \beta\}$.

The task of (γ, β)-*Max-CSP*(f) *is to distinguish between instances from* $\Gamma = \cup_{m \leq \mathsf{poly}(n)} \Gamma_m$ *and instances from* $B = \cup_{m \leq \mathsf{poly}(n)} B_m$. *Specifically we desire algorithms that output* 1 *w.p. at least* 2/3 *on inputs from* Γ *and output* 1 *w.p. at most* 1/3 *on inputs from* B.

Let $\rho(f) = 2^{-k} \cdot |\{\mathbf{a} \in \{-1, 1\}^k \mid f(\mathbf{a}) = 1\}|$ denote the probability that a uniformly random assignment satisfies f. We say f is *streaming-approximation-resistant* if for every $\varepsilon > 0$, the $(1, \rho(f) + \varepsilon)$-$\mathsf{Max\text{-}CSP}(f)$ requires $\Omega(n^\delta)$ space for some constant $\delta > 0$.

We now define streaming and sketching algorithms in the context of $\mathsf{Max\text{-}CSP}(f)$. Note that the input to both algorithms are sequences of constraints. We use $\mathsf{C}_{\mathcal{F},n}$ to denote the set of all constraints of $\mathsf{Max\text{-}CSP}(f)$ on n variables. A stream is thus an element of $(\mathsf{C}_{\mathcal{F},n})^*$ and we use λ to denote the empty stream.

Definition 3.2 (Streaming algorithm)**.** *A space s general streaming algorithm* **ALG** *for Max-CSP*(f) *on n variables is given by a (state-evolution) function* $S : \{0, 1\}^s \times \mathsf{C}_{\mathcal{F},n} \to \{0, 1\}^s$ *and a (output) function* $v : \{0, 1\}^s \to [0, 1]$. *Let* $\widetilde{S} : (\mathsf{C}_{\mathcal{F},n})^* \to \{0, 1\}^s$ *given by* $\widetilde{S}(\lambda) = 0^s$ *and* $\widetilde{S}(\sigma_1, \ldots, \sigma_m)) = S(\widetilde{S}(\sigma_1, \ldots, \sigma_{m-1}), \sigma_m)$ *denote the iterated state-evolution map. Then the output of* **ALG** *on input* $\sigma = (\sigma_1, \ldots, \sigma_m)$ *is* $v(\widetilde{S}(\sigma))$. *For the purposes of this paper, a randomized streaming algorithm is simply a distribution on the pairs* (S, v).

Sketching algorithms are a special class of streaming algorithms that have been widely used in both upper bounds and lower bounds.

Definition 3.3 (Sketching algorithms)**.** *A (deterministic) space s streaming algorithm* $\mathbf{ALG} = (S, v)$ *is a sketching algorithm if there exists a compression function* $\mathsf{COMP} : (\mathsf{C}_{\mathcal{F},n})^* \to \{0,1\}^s$ *and a combination function* $\mathsf{COMB} : \{0,1\}^s \times \{0,1\}^s \to \{0,1\}^s$ *such that the following hold:*

- $S(z, C) = \mathsf{COMB}(z, \mathsf{COMP}(C))$ *for every* $z \in \{0,1\}^s$ *and* $C \in \mathsf{C}_{\mathcal{F},n}$.
- *For every pair of streams* $\sigma, \tau \in (\mathsf{C}_{\mathcal{F},n})^*$, *we have*

$$\mathsf{COMB}(\mathsf{COMP}(\sigma), \mathsf{COMP}(\tau)) = \mathsf{COMP}(\sigma \circ \tau)$$

  where $\sigma \circ \tau$ *represents the concatenation of the streams* σ *and* τ. *A randomized algorithm* $\mathbf{ALG}$ *is a randomized sketching algorithm if it is a distribution over deterministic sketching algorithms.*

We remark that a linear sketching algorithm roughly associates with elements of a vector space V (over some field) and COMB is simply vector addition in V.

3.2 Total variation distance

The total variation distance between probability distributions plays an important role in our analysis.

Definition 3.4 (Total variation distance of discrete random variables)**.** *Let* Ω *be a finite probability space and* X, Y *be random variables with support* Ω. *The total variation distance between* X *and* Y *is defined as follows.*

$$\|X - Y\|_{tvd} := \frac{1}{2} \sum_{\omega \in \Omega} |\Pr[X = \omega] - \Pr[Y = \omega]| \ .$$

We will use the triangle and data processing inequalities for the total variation distance.

Proposition 3.5 (E.g.,[KKS15, Claim 6.5])**.** *For random variables* X, Y *and* W*:*

- *(Triangle inequality)* $\|X - Y\|_{tvd} \geq \|X - W\|_{tvd} - \|Y - W\|_{tvd}$.
- *(Data processing inequality) If* W *is independent of both* X *and* Y, *and* f *is a function, then* $\|f(X, W) - f(Y, W)\|_{tvd} \leq \|X - Y\|_{tvd}$.

3.3 Concentration inequality

We will use the following concentration inequality which is essentially an Azuma-Hoeffding style inequality for submartingales. The form we use is based on [KK19, Lemma 2.5], and allows for variables with different expectations. The analysis is a very slight modification of theirs.

Lemma 3.6. *Let* $X = \sum_{i \in [N]} X_i$ *where* X_i *are Bernoulli random variables such that for every* $k \in [N]$, $\mathbb{E}[X_k \,|\, X_1, \ldots, X_{k-1}] \leq p_k$ *for some* $p_k \in (0, 1)$. *Let* $\mu = \sum_{k=1}^{N} p_k$. *For every* $\Delta > 0$, *we have:*

$$\Pr[X \geq \mu + \Delta] \leq \exp\left(-\frac{\Delta^2}{2\mu + 2\Delta}\right) .$$

Proof. Let $v = \Delta/(\mu+\Delta)$ and $u = \ln(1+v)$. We have

$$\mathbb{E}[e^{uX}] = \mathbb{E}[\prod_{k=1}^{N} e^{uX_k}] \le (1+p_N(e^u-1)) \cdot \mathbb{E}[\prod_{k=1}^{N-1} e^{uX_k}] \le \prod_{i=1}^{N}(1+p_k(e^u-1)) = \prod_{i=1}^{N}(1+p_k v) \le e^{v\mu},$$

where the final inequality uses $1+x \le e^x$ for every x (and the definition of μ). Applying Markov to the above, we have:

$$\Pr\left[X \ge \mu+\Delta\right] = \Pr\left[e^{uX} \ge e^{u(\mu+\Delta)}\right] \le \mathbb{E}[e^{uX}]/e^{u(\mu+\Delta)} \le e^{v\mu-u\mu-u\Delta}.$$

From the inequality $e^{v-v^2/2} \le 1+v$ we infer $u \ge v - v^2/2$ and so the final expression above can be bounded as:

$$\Pr\left[X \ge \mu+\Delta\right] \le e^{v\mu-u\mu-u\Delta} \le e^{\frac{v^2}{2}(\mu+\Delta)-v\Delta} = e^{-\frac{\Delta^2}{2(\mu+\Delta)}},$$

where the final equality comes from our choice of v. □

3.4 Fourier analysis

We will need the following basic notions from Fourier analysis over the Boolean hypercube (see, for instance, [O'D14]). For a Boolean function $f : \{-1,1\}^k \to \mathbb{R}$ its Fourier coefficients are defined by $\widehat{f}(\mathbf{v}) = \mathbb{E}_{\mathbf{a}\in\{-1,1\}^k}[f(\mathbf{a}) \cdot (-1)^{\mathbf{v}^\top \mathbf{a}}]$, where $\mathbf{v} \in \{0,1\}^k$. We need the following two important tools.

Lemma 3.7 (Parseval's identity)**.** *For every function $f : \{-1,1\}^k \to \mathbb{R}$,*

$$\|f\|_2^2 = \frac{1}{2^k} \sum_{\mathbf{a}\in\{-1,1\}^k} f(\mathbf{a})^2 = \sum_{\mathbf{v}\in\{0,1\}^k} \widehat{f}(\mathbf{v})^2 .$$

Note that for every distribution f on $\{-1,1\}^k$, $\widehat{f}(0^k) = 2^{-k}$. For the uniform distribution U on $\{-1,1\}^k$, $\widehat{U}(\mathbf{v}) = 0$ for every $\mathbf{v} \ne 0^k$. Thus, by Lemma 3.7, for any distribution f on $\{-1,1\}^k$:

$$\|f-U\|_2^2 = \sum_{\mathbf{v}\in\{0,1\}^k} \left(\widehat{f}(\mathbf{v}) - \widehat{U}(\mathbf{v})\right)^2 = \sum_{\mathbf{v}\in\{0,1\}^k\setminus\{0^k\}} \widehat{f}(\mathbf{v})^2 . \tag{3.8}$$

Next, we will use the following consequence of hypercontractivity for Boolean functions as given in [GKK$^+$09, Lemma 6] which in turns relies on a lemma from [KKL88].

Lemma 3.9. *Let $f : \{-1,1\}^n \to \{-1,0,1\}$ and $A = \{\mathbf{a} \in \{-1,1\}^n \mid f(\mathbf{a}) \ne 0\}$. If $|A| \ge 2^{n-c}$ for some $c \in \mathbb{N}$, then for every $\ell \in \{1,\ldots,4c\}$, we have*

$$\frac{2^{2n}}{|A|^2} \sum_{\substack{\mathbf{v}\in\{0,1\}^n \\ \|\mathbf{v}\|_1=\ell}} \widehat{f}(\mathbf{v})^2 \le \left(\frac{4\sqrt{2}c}{\ell}\right)^{\ell} .$$

4 A Streaming Approximation Algorithm for $\textsf{Max-CSP}(f)$

In this section we give our main algorithmic result — a $O(\log n)$-space linear sketching algorithm for (γ,β)-$\textsf{Max-CSP}(f)$ if $K_\gamma^Y = K_\gamma^Y(f)$ and $K_\beta^N = K_\beta^N(f)$ are disjoint. (See Definition 2.2.)

The algorithm in fact works in the (general) dynamic setting where the input $\Psi = (C_1,\ldots,C_m; w_1,\ldots,w_m)$ is obtained by inserting and deleting (unweighted) constraints, possibly with repetitions and thus leading to a (integer) weighted instance. Formally $\Psi = (C_1,\ldots,C_m; w_1,\ldots,w_m)$ is presented as a stream $\sigma_1,\ldots,\sigma_\ell$ where $\sigma_t = (C'_t, w'_t)$ and $w'_t \in \{-1,1\}$ such that $w_i = \sum_{t\in[\ell]:C_i=C'_t} w'_t$. For the algorithmic result to hold, we require that w_i's are non-negative at the end of the stream but the intermediate values can be arbitrary. Furthermore the algorithm requires that the length of the stream be polynomial in n (or else there will be a logarithmic multiplicative factor in the length of the stream in the space usage).

We state our main theorem of this section which simply restates Part (1) of Theorem 2.3.

Theorem 4.1. *For every function $f : \{-1,1\}^k \to \{0,1\}$ and for every $0 \le \beta < \gamma \le 1$, if $K_\gamma^Y(f) \cap K_\beta^N(f) = \emptyset$, then (γ,β)-$\textsf{Max-CSP}(f)$ admits a probabilistic streaming algorithm in the dynamic setting that uses $O(\log n)$ space and succeeds with probability at least $2/3$.*

The overview of the algorithm is as follows: We use the separability of K_γ^Y and K_β^N to obtain a hyperplane with normal vector $\boldsymbol{\lambda}$ that separates the two sets. We then estimate a $\boldsymbol{\lambda}$-weighted bias of a given instance Ψ and accept Ψ if this bias falls on the K_γ^Y side of the hyperplane. We note that the bias can be approximated arbitrarily well using well-known ℓ_1-norm approximators in the turnstile setting. The bulk of the work is in analyzing the correctness of our algorithm.

We will use the following streaming algorithm for approximating the ℓ_1 norm of a vector.

Proposition 4.2 ([Ind00],[KNW10, Theorem 2.1])**.** *Given a stream S of $\textsf{poly}(n)$ updates $(i,v) \in [n] \times \{-M,-(M-1),\ldots,M-1,M\}$ where $M = \textsf{poly}(n)$, let $x_i = \sum_{(i,v)\in S} v$ for $i \in [n]$. For every $\varepsilon > 0$, there exists a linear sketch that uses $O(\log n)$ bits of memory and outputs a $(1\pm\varepsilon)$-approximation to the value $\|x\|_1 = \sum_i |x_i|$ with probability at least $2/3$.*

4.1 Algorithm

Let us start with the definition of $\boldsymbol{\lambda}$-bias.

Definition 4.3 (Bias (vector))**.** *For $\boldsymbol{\lambda} = (\lambda_1,\ldots,\lambda_k) \in \mathbb{R}^k$, and instance $\Psi = (C_1,\ldots,C_m; w_1,\ldots,w_m)$ of $\textsf{Max-CSP}(f)$ where $C_i = (\mathbf{j}(i),\mathbf{b}(i))$ and $w_i \ge 0$, we let the $\boldsymbol{\lambda}$-*bias vector *of Ψ, denoted $\textsf{bias}_{\boldsymbol{\lambda}}(\Psi)$, be the vector in $\mathbb{R}^n$ given by*

$$\textsf{bias}_{\boldsymbol{\lambda}}(\Psi)_\ell = \frac{1}{W} \cdot \sum_{i\in[m],t\in[k]:j(i)_t=\ell} \lambda_t w_i b(i)_t\,,$$

for $\ell \in [n]$, where $W = \sum_{i\in[m]} w_i$. The $\boldsymbol{\lambda}$-bias of Ψ, denoted $B_{\boldsymbol{\lambda}}(\Psi)$, is the ℓ_1 norm of $\textsf{bias}_{\boldsymbol{\lambda}}(\Psi)$, i.e., $B_{\boldsymbol{\lambda}}(\Psi) = \sum_{\ell=1}^n |\textsf{bias}_{\boldsymbol{\lambda}}(\Psi)_\ell|$.

By directly applying the known ℓ_1-sketching algorithm (i.e., Proposition 4.2), the following lemma shows that $\boldsymbol{\lambda}$-bias can be estimated in $O(\log n)$ space.

Lemma 4.4. *For every vector $\boldsymbol{\lambda} \in \mathbb{R}^k$ and $\varepsilon > 0$, there exists a $O(\log n)$ space algorithm $\mathcal{A}$ that on input a stream $\sigma_1,\ldots,\sigma_\ell$, representing an instance $\Psi = (C_1,\ldots,C_m; w_1,\ldots,w_m)$, outputs a $(1\pm\varepsilon)$-approximation to $B_{\boldsymbol{\lambda}}(\Psi)$, i.e., for every Ψ, $(1-\varepsilon)B_{\boldsymbol{\lambda}}(\Psi) \le \mathcal{A}(\Psi) \le (1+\varepsilon)B_{\boldsymbol{\lambda}}(\Psi)$, with probability at least $2/3$.*

Proof. Note that since k and ε are constants with respect to n, we can without loss of generality assume that each entry of λ is an integer and ε has constant bit complexity. [9]

On input a stream $\sigma_1, \ldots, \sigma_\ell$ representing an instance $\Psi = (C_1, \ldots, C_m; w_1, \ldots, w_m)$ (see **??**) with $\sigma_i = (C'_i = (\mathbf{j}(i), \mathbf{b}(i)), w'_i)$, the algorithm $\mathcal{A}$ proceeds as follows. It implicitly maintains a vector $\mathbf{v} \in \mathbb{R}^n$ which is initially zero. Each stream element σ_i is converted into k updates to $\mathbf{v}$ given by $(\mathbf{j}(i)_1, w'_i \cdot \lambda_1), \ldots, (\mathbf{j}(i)_k, w'_i \cdot \lambda_k)$ (where the notation of "updating by (i, x)" indicates that x is added to v_i). It then applies the algorithm from Proposition 4.2 to compute a $(1 \pm \varepsilon)$ approximation B' to $\|v\|_1 = \sum_{i \in [m], t \in [k] : j(i)_t = \ell} \lambda_t w_i b(i)_t$. (Note that since $\ell = \mathsf{poly}(n)$ and k is a constant, we know that there are only $\mathsf{poly}(n)$ updates and each update is a constant integer and so the conditions of Proposition 4.2 are satisfied, and so B' is a $(1 \pm \varepsilon)$ approximation to $\|v\|_1$ with probability at least $2/3$.) Finally $\mathcal{A}$ outputs B'/W which is a $(1 \pm \varepsilon)$-approximation to $B_{\boldsymbol{\lambda}}(\Psi)$ if and only if B' is a $(1 \pm \varepsilon)$ approximation to $\|v\|_1$. □

We will use the following form of the hyperplane separation theorem for convex bodies (see, e.g., [BV04, Exercise 2.22]).

Proposition 4.5. *Let K^Y and K^N be two disjoint nonempty closed convex sets in $\mathbb{R}^k$ at least one of which is compact. Then there exists a nonzero vector $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k)$ and real numbers $\tau_Y > \tau_N$ such that*

$$\forall \mathbf{x} \in K^Y, \quad \langle \boldsymbol{\lambda}, \mathbf{x} \rangle \geq \tau_Y \quad \textit{and} \quad \forall \mathbf{x} \in K^N, \quad \langle \boldsymbol{\lambda}, \mathbf{x} \rangle \leq \tau_N \,.$$

We are now ready to describe our algorithm for (γ, β)-$\mathsf{Max\text{-}CSP}(f)$.

Algorithm 1 A streaming algorithm for (γ, β)-$\mathsf{Max\text{-}CSP}(f)$

```
Input: a stream $\sigma_1, \ldots, \sigma_\ell$ representing an instance $\Psi$ of $\mathsf{Max\text{-}CSP}(f)$.
1: Let $\boldsymbol{\lambda} \in \mathbb{R}^k$ and $\tau_N < \tau_Y$ be as given by Proposition 4.5 separating $K^Y_\gamma(f)$ and $K^N_\beta(f)$.
2: Let $\varepsilon = \frac{\tau_Y - \tau_N}{2(\tau_Y + \tau_N)}$ (so that $(1-\varepsilon)\tau_Y > (1+\varepsilon)\tau_N$).
3: Use the algorithm $\mathcal{A}$ from Lemma 4.4 to compute $\tilde{B}$ to be a $(1 \pm \varepsilon)$ approximation to $B_{\boldsymbol{\lambda}}(\Psi)$,
   i.e., $(1-\varepsilon)B_{\boldsymbol{\lambda}}(\Psi) \leq \tilde{B} \leq (1+\varepsilon)B_{\boldsymbol{\lambda}}(\Psi)$ with probability at least $2/3$.
4: if $\tilde{B} \leq \tau_N(1+\varepsilon)$ then
   Output: NO.
5: else
   Output: YES.
```

It is clear that the algorithm above runs in $O(\log n)$ space (in particular by using Proposition 4.2 via Lemma 4.4 for Step 3). We now turn to analyzing the correctness of the algorithm.

4.2 Analysis of the correctness of Algorithm 1

Lemma 4.6. *Algorithm 1 correctly solves (γ, β)-$\mathsf{Max\text{-}CSP}(f)$, if $K^Y_\gamma(f)$ and $K^N_\beta(f)$ are disjoint. Specifically, for every Ψ, let $\tau_Y, \tau_N, \varepsilon, \boldsymbol{\lambda}, \tilde{B}$ be as given in Algorithm 1, we have:*

$$\mathsf{val}_\Psi \geq \gamma \quad \Rightarrow \quad B_{\boldsymbol{\lambda}}(\Psi) \geq \tau_Y \textit{ and } \tilde{B} > \tau_N(1+\varepsilon)\,,$$
$$\textit{and } \mathsf{val}_\Psi \leq \beta \quad \Rightarrow \quad B_{\boldsymbol{\lambda}}(\Psi) \leq \tau_N \textit{ and } \tilde{B} \leq \tau_N(1+\varepsilon)\,,$$

[9] Concretely, round ε to 2^{-t} where t is the smallest integer such that $\varepsilon \geq 2^{-t}$. As for $\boldsymbol{\lambda}$, let $\lambda_{\min} = \min_{j \in [k]} |\lambda_j|$ and round it the same way as we did for ε. Next, for each $j \in [k]$, scale and round λ_j to $\lceil \frac{4\lambda_j}{\lambda_{\min}} \rceil$. It is not difficult to verify that scaling down the new $\boldsymbol{\lambda}$-bias by a factor of $\lambda_{\min}/4$, it is a $(1 \pm \varepsilon/2)$-approximation to the original $\boldsymbol{\lambda}$-bias.

provided $(1-\varepsilon)B_\lambda(\Psi) \leq \tilde{B} \leq (1+\varepsilon)B_\lambda(\Psi)$.

In the rest of this section, we will prove Lemma 4.6. The key to our analysis is a distribution $\mathcal{D}(\Psi^{\mathbf{a}}) \in \Delta(\{-1,1\}^k)$ that we associate with every instance Ψ and assignment $\mathbf{a} \in \{-1,1\}^n$ to the variables of Ψ. Recall that in Definition 2.2, we define $\boldsymbol{\mu}(\mathcal{D}) = (\mu_1,\ldots,\mu_k)$ where $\mu_i = \mathbb{E}_{\mathbf{b}\sim\mathcal{D}}[b_i]$. If Ψ is γ-satisfied by assignment $\mathbf{a}$, we prove that $\boldsymbol{\mu}(\mathcal{D}(\Psi^{\mathbf{a}})) \in K_\gamma^Y$. On the other hand, if Ψ is not β-satisfiable by any assignment, we prove that for every $\mathbf{a}$, $\boldsymbol{\mu}(\mathcal{D}(\Psi^{\mathbf{a}})) \in K_\beta^N$. Finally we also show that the bias $B_{\boldsymbol{\lambda}}(\Psi)$ relates to $\boldsymbol{\lambda}(\mathcal{D}(\Psi^{\mathbf{a}})) \triangleq \langle \boldsymbol{\mu}(\mathcal{D}(\Psi^{\mathbf{a}})), \boldsymbol{\lambda}\rangle$, where the latter quantity is exactly what needs to be computed (by Proposition 4.5) to distinguish the membership of $\boldsymbol{\mu}(\mathcal{D}(\Psi^{\mathbf{a}}))$ in K_γ^Y versus the membership in K_β^N.

We start with recalling some notations. For an instance $\Psi = (C_1,\ldots,C_m;w_1,\ldots,w_m)$ on n variables with $C_i = (\mathbf{j}(i),\mathbf{b}(i))$, and an assignment $\mathbf{a} \in \{-1,1\}^n$, let $\Psi^{\mathbf{a}}$ denote the new instance obtained by flipping the variables according to $\mathbf{a}$. Specifically $\Psi^{\mathbf{a}} = (C_1^{\mathbf{a}},\ldots,C_m^{\mathbf{a}};w_1,\ldots,w_m)$ where $C_i^{\mathbf{a}} = (\mathbf{j}(i), \mathbf{a}|_{\mathbf{j}(i)} \odot \mathbf{b}(i))$.

Given instance Ψ, let $\mathcal{D}(\Psi) \in \Delta(\{-1,1\}^k)$ be the distribution obtained by sampling a constraint at random (according to its weight) from Ψ and outputting the "negation pattern". Formally, to sample a random vector $\mathbf{b} \sim \mathcal{D}(\Psi)$, we sample $i \in [m]$ with probability w_i/W where $W = \sum_{i\in[m]} w_i$, and output $\mathbf{b}(i)$ where $C_i = (\mathbf{j}(i),\mathbf{b}(i))$.

The next lemma relates the $\boldsymbol{\lambda}$-bias vector of Ψ to $\boldsymbol{\lambda}(\mathcal{D}(\Psi^{\mathbf{a}}))$ and uses this to relate the bias of Ψ to the maximum over $\mathbf{a}$ of $\boldsymbol{\lambda}(\mathcal{D}(\Psi^{\mathbf{a}}))$.

Lemma 4.7. *For every vector* $\mathbf{a} \in \{-1,1\}^n$, *we have* $\boldsymbol{\lambda}(\mathcal{D}(\Psi^{\mathbf{a}})) = \langle \mathbf{a}, \mathsf{bias}_{\boldsymbol{\lambda}}(\Psi)\rangle$. *Consequently we have* $B_{\boldsymbol{\lambda}}(\Psi) = \max_{\mathbf{a}\in\{-1,1\}^n}\{\boldsymbol{\lambda}(\mathcal{D}(\Psi^{\mathbf{a}}))\}$.

Proof. We start with the first equality. Fix $\mathbf{a} \in \{-1,1\}^n$. We have

$$\begin{aligned}
\boldsymbol{\lambda}(\mathcal{D}(\Psi^{\mathbf{a}})) &= \langle \boldsymbol{\mu}(\mathcal{D}(\Psi^{\mathbf{a}})), \boldsymbol{\lambda}\rangle \quad \text{(By definition of } \lambda(\cdot))\\
&= \mathop{\mathbb{E}}_{\mathbf{y}\sim\mathcal{D}(\Psi^{\mathbf{a}})}[\langle \mathbf{y}, \boldsymbol{\lambda}\rangle] \quad \text{(By definition of } \boldsymbol{\mu}(\mathcal{D}) \text{ and linearity of inner product)}\\
&= \mathop{\mathbb{E}}_{i}[\langle \mathbf{b}^{\mathbf{a}}(i), \boldsymbol{\lambda}\rangle] \quad \text{(By definition of } \mathcal{D}(\Psi^{\mathbf{a}}))\\
&= \mathop{\mathbb{E}}_{i}\left[\sum_{t\in[k]} b^{\mathbf{a}}(i)_t \cdot \lambda_t\right] \quad \text{(Expanding the inner product)}\\
&= \frac{1}{W}\sum_{i\in[m]} w_i \sum_{\ell\in[n]}\sum_{t\in[k]} \mathbb{1}_{\mathbf{j}(i)_t=\ell}\cdot\lambda_t\cdot a_\ell\cdot b(i)_t \quad \text{(Using definition of } \Psi^{\mathbf{a}})\\
&= \frac{1}{W}\sum_{\ell\in[n]} a_\ell \sum_{t\in[k]}\lambda_t \sum_{i\in[m]} \mathbb{1}_{\mathbf{j}(i)_t=\ell}\cdot w_i\cdot b(i)_t \quad \text{(Exchanging summations)}\\
&= \sum_{\ell\in[n]} a_\ell \cdot \mathsf{bias}_{\boldsymbol{\lambda}}(\Psi)_\ell \quad \text{(By definition of } \mathsf{bias}_{\boldsymbol{\lambda}}(\cdot))\\
&= \langle \mathbf{a}, \mathsf{bias}_{\boldsymbol{\lambda}}(\Psi)\rangle\,,
\end{aligned}$$

yielding the first equality.

The second part is immediate from the observation that for every vector $\mathbf{v} \in \mathbb{R}^n$, we have $||\mathbf{v}||_1 = \max_{\mathbf{a}\in\{-1,1\}^n}\langle \mathbf{a}, \mathbf{v}\rangle$ and so

$$B_{\boldsymbol{\lambda}}(\Psi) = ||\mathsf{bias}_{\boldsymbol{\lambda}}(\Psi)||_1 = \max_{\mathbf{a}\in\{-1,1\}^n}\{\langle \mathbf{a}, \mathsf{bias}_{\boldsymbol{\lambda}}(\Psi)\rangle\} = \max_{\mathbf{a}\in\{-1,1\}^n}\{\boldsymbol{\lambda}(\mathcal{D}(\Psi^{\mathbf{a}}))\}\,.$$

□

We now turn to connecting val_Ψ to properties of $\mathcal{D}(\Psi^{\mathbf{a}})$.

Lemma 4.8. *For every Ψ and $\mathbf{a}$, if $\mathsf{val}_\Psi(\mathbf{a}) \geq \gamma$ then $\mathcal{D}(\Psi^{\mathbf{a}}) \in S^Y_\gamma$.*

Proof. Follows from the fact that

$$\mathop{\mathbb{E}}_{\mathbf{b}\sim\mathcal{D}(\Psi^{\mathbf{a}})}[f(\mathbf{b})] = \frac{1}{W}\sum_{i\in[m]} w_i \cdot f(\mathbf{b}(i)\odot \mathbf{a}|_{\mathbf{j}(i)}) = \frac{1}{W}\sum_{i\in[m]} w_i \cdot C_i(\mathbf{a}) = \mathsf{val}_\Psi(\mathbf{a}) \geq \gamma\,,$$

implying $\mathcal{D}(\Psi^{\mathbf{a}}) \in S^Y_\gamma$. □

Lemma 4.9. *For every Ψ, if $\mathsf{val}_\Psi \leq \beta$, then for all $\mathbf{a}$, we have $\mathcal{D}(\Psi^{\mathbf{a}}) \in S^N_\beta$.*

Proof. We claim if $\mathsf{val}_\Psi \leq \beta$, then $\mathcal{D}(\Psi) \in S^N_\beta$. This suffices to prove the lemma, since for every $\mathbf{a} \in \{-1,1\}^n$ we have $\mathsf{val}_{\Psi^{\mathbf{a}}} = \mathsf{val}_\Psi$. So if $\mathsf{val}_\Psi \leq \beta$ then $\mathsf{val}_{\Psi^{\mathbf{a}}} \leq \beta$ and so by the claim above applied to $\Psi^{\mathbf{a}}$, we have $\mathcal{D}(\Psi^{\mathbf{a}}) \in S^N_\beta$.

We prove the contrapositive, i.e., we assume $\mathcal{D}(\Psi) \notin S^N_\beta$ and show this implies $\mathsf{val}_\Psi > \beta$. If $\mathcal{D}(\Psi) \notin S^N_\beta$, then there exists $p \in [0,1]$ such that $\mathbb{E}_{\mathbf{b}\sim\mathcal{D}(\Psi)}\mathbb{E}_{\mathbf{c}\sim\mathsf{Bern}(p)^k}[f(\mathbf{b}\odot\mathbf{c})] > \beta$. But this implies, as we show below, that if $\boldsymbol{\sigma} \sim \mathsf{Bern}(p)^n$, then $\mathbb{E}_{\boldsymbol{\sigma}\sim\mathsf{Bern}(p)^n}[\mathsf{val}_\Psi(\boldsymbol{\sigma})] > \beta$. We have:

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{\sigma}\sim\mathsf{Bern}(p)^n}[\mathsf{val}_\Psi(\boldsymbol{\sigma})] &= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}\sim\mathsf{Bern}(p)^n}\mathop{\mathbb{E}}_{i}[C_i(\boldsymbol{\sigma})] \quad \text{(By definition of } \Psi\text{)}\\
&= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}\sim\mathsf{Bern}(p)^n}\mathop{\mathbb{E}}_{i}[f(\mathbf{b}(i)\odot\boldsymbol{\sigma}|_{\mathbf{j}(i)})] \quad \text{(By definition of } C_i\text{)}\\
&= \mathop{\mathbb{E}}_{i}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}|_{\mathbf{j}(i)}\sim\mathsf{Bern}(p)^k}[f(\mathbf{b}(i)\odot\boldsymbol{\sigma}|_{\mathbf{j}(i)})] \quad \text{(Exchanging summations)}\\
&= \mathop{\mathbb{E}}_{i}\mathop{\mathbb{E}}_{\mathbf{c}\sim\mathsf{Bern}(p)^k}[f(\mathbf{b}(i)\odot\mathbf{c})] \quad \text{(Renaming variables)}\\
&= \mathop{\mathbb{E}}_{\mathbf{b}\sim\mathcal{D}(\Psi)}\mathop{\mathbb{E}}_{\mathbf{c}\sim\mathsf{Bern}(p)^k}[f(\mathbf{b}\odot\mathbf{c})] \quad \text{(By definition of } \mathcal{D}(\Psi)\text{)}\\
&> \beta \quad \text{(By the contrapositive assumption)}
\end{aligned}$$

Since $\mathsf{val}_\Psi \triangleq \max_{\boldsymbol{\sigma}}\{\mathsf{val}_\Psi(\boldsymbol{\sigma})\} \geq \mathbb{E}_{\boldsymbol{\sigma}\sim\mathsf{Bern}(p)^n}[\mathsf{val}_\Psi(\boldsymbol{\sigma})]$, we get a contradiction to $\mathsf{val}_\Psi \leq \beta$. This concludes the proof of the claim and hence the lemma. □

Before turning to the proof of Lemma 4.6, we first do a quick post-analysis of the proof above. The proof above is the key reason why the definition of S^N_β is chosen as it is: In particular, from the fact that there was an i.i.d. distribution, namely $\mathsf{Bern}(p)^k$, according to which a random assignment satisfied the "instance" underlying $\mathcal{D}(\Psi)$ with value more than β allowed us to extend this to a (again i.i.d., but this was not necessary) distribution over assignments to Ψ that also achieved value of at least β. Note that the mere existence of an assignment of value greater than β on $\mathcal{D}(\Psi)$ would have been insufficient for this step to go through, explaining our choice of definition of S^N_β.

We are now ready to prove Lemma 4.6.

Proof of Lemma 4.6. Let $\mathsf{val}_\Psi \geq \gamma$. Then there exists $\mathbf{a} \in \{-1,1\}^n$ such that $\mathsf{val}_\Psi(\mathbf{a}) \geq \gamma$. By Lemma 4.8, we have that $\mathcal{D}(\Psi^{\mathbf{a}}) \in S^Y_\gamma$. By our choice of $\boldsymbol{\lambda}$, we have $\lambda(\mathcal{D}) \geq \tau_Y$ for every $\mathcal{D} \in S_\gamma^Y$ and so in particular we have $\boldsymbol{\lambda}(\mathcal{D}(\Psi^{\mathbf{a}})) \geq \tau_Y$. By Lemma 4.7, we have $B_{\boldsymbol{\lambda}}(\Psi) = \max_{\mathbf{c}\in\{-1,1\}^n}\{\boldsymbol{\lambda}(\mathcal{D}(\Psi^{\mathbf{c}}))\}$. Putting these together we have

$$B_{\boldsymbol{\lambda}}(\Psi) = \max_{\mathbf{c}\in\{-1,1\}^n}\{\boldsymbol{\lambda}(\mathcal{D}(\Psi^{\mathbf{c}}))\} \geq \boldsymbol{\lambda}(\mathcal{D}(\Psi^{\mathbf{a}})) \geq \tau_Y\,.$$

Finally, since $\tilde{B} \geq (1-\varepsilon)B_\lambda(\Psi)$, we get $\tilde{B} \geq (1-\varepsilon)\tau_Y > (1+\varepsilon)\tau_N$, where the final inequality holds by our choice of ε.

The case $\mathsf{val}_\Psi \leq \beta$ is similar. In this case, by Lemma 4.9 we have $\mathcal{D}(\Psi^{\mathbf{a}}) \in S_\beta^N$ for every $\mathbf{a}$. Now applying Lemma 4.7 we get that for every $\mathbf{a}$, $\langle \mathbf{a}, \mathsf{bias}_{\boldsymbol{\lambda}}\rangle = \boldsymbol{\lambda}(\mathcal{D}(\Psi^{\mathbf{a}})) \leq \tau_N$. We conclude that $B_{\boldsymbol{\lambda}}(\Psi) = \max_{\mathbf{a}\in\{-1,1\}^n}\{\langle \mathbf{a}, \mathsf{bias}_{\boldsymbol{\lambda}}\rangle\} \leq \tau_N$. since $\tilde{B} \leq (1+\varepsilon)B_{\boldsymbol{\lambda}}(\Psi)$, we get $\tilde{B} \leq (1+\varepsilon)\tau_N$. □

We now conclude the section with a formal proof of Theorem 4.1.

Proof of Theorem 4.1. The desired algorithm is Algorithm 1. Its space complexity is bounded by the space required for Step 3, which by Lemma 4.4 is $O(\log n)$. Assuming Step 3 works correctly, which happens with probability at least 2/3, Lemma 4.6 shows that it correctly solves (γ,β)-$\mathsf{Max}\text{-}\mathsf{CSP}(f)$ whenever $K_\gamma^Y(f) \cap K_\beta^N(f) = \emptyset$. □

5 Sketching and Streaming Space Lower Bounds for Max-CSP(f)

In this section, we prove our main lower bound results, modulo a communication complexity lower bound which is proved in Section 6 and Section 7. We start by recalling the results to be proved. First we restate the lower bound in the general streaming setting. Recall that $(\mathcal{D}_Y, \mathcal{D}_N)$ form a padded one-wise pair if there exist $\tau \in [0,1]$, and $\mathcal{D}_0, \mathcal{D}'_Y, \mathcal{D}'_N$ such that for $i \in \{Y, N\}$ we have $\mathcal{D}_i = \tau\mathcal{D}_0 + (1-\tau)\mathcal{D}'_i$ and $\mathcal{D}'_i$ has uniform marginals.

Theorem 2.11 (Streaming lower bound)**.** *For every function $f : \{-1,1\}^k \to \{0,1\}$ and for every $0 \leq \beta < \gamma \leq 1$, if there exists a padded one-wise pair of distributions $\mathcal{D}_Y \in S_\gamma^Y$ and $\mathcal{D}_N \in S_\beta^N$ then, for every $\varepsilon > 0$, then every streaming algorithm that solves $(\gamma-\varepsilon, \beta+\varepsilon)$-Max-CSP$(f)$ requires $\Omega(\sqrt{n})$ space. Furthermore, if $\gamma = 1$ then $(1, \beta+\varepsilon)$-Max-CSP(f) requires $\Omega(\sqrt{n})$ space.*

We also restate the lower bound against sketching algorithms from Theorem 2.3 as a separate theorem below.

Theorem 5.1 (Lower bounds against sketching algorithms)**.** *For every function $f : \{-1,1\}^k \to \{0,1\}$ and for every $0 \leq \beta < \gamma \leq 1$, if $K_\gamma^Y(f) \cap K_\beta^N(f) \neq \emptyset$, then for every $\varepsilon > 0$, every sketching algorithm for $(\gamma-\varepsilon, \beta+\varepsilon)$-Max-CSP$(f)$ requires $\Omega(\sqrt{n})$ space*[10]*. Furthermore, if $\gamma = 1$. then $(1, \beta+\varepsilon)$-Max-CSP(f) requires $\Omega(\sqrt{n})$ space.*

To prove both theorems, we introduce the *Randomized Mask Detection (RMD)* communication game in Section 5.1. We then state a lower bound for the communication complexity of this game (Theorem 5.3), and use the lower bound to prove Theorem 2.11 in Section 5.2.4 and Theorem 5.1 in Section 5.3.2. The proof of Theorem 5.3 appears in Section 7.

[10] The constant hidden in the Ω notation may depend on k and ε.

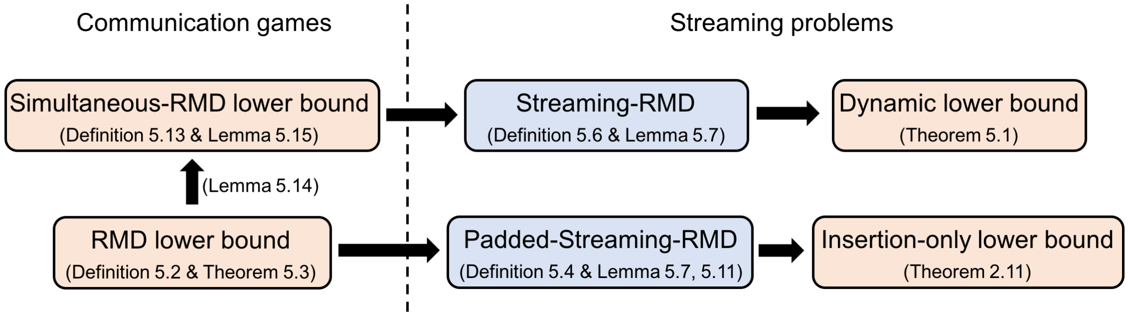
Communication games
Streaming problems
Simultaneous-RMD lower bound
(Definition 5.13 & Lemma 5.15)
Streaming-RMD
(Definition 5.6 & Lemma 5.7)
Dynamic lower bound
(Theorem 5.1)
(Lemma 5.14)
RMD lower bound
(Definition 5.2 & Theorem 5.3)
Padded-Streaming-RMD
(Definition 5.4 & Lemma 5.7, 5.11)
Insertion-only lower bound
(Theorem 2.11)

Figure 2: Roadmap of this section.

5.1 2-Player Communication Games and the Randomized Mask Detection Problem

In most of this section and the rest of this paper, we will be considering the complexity of 2-player 1-way communication games. Broadly such games are described by two (parameterized set of) distributions $\mathcal{Y}$ and $\mathcal{N}$. An instance of the game is a pair (X, Y) either drawn from $\mathcal{Y}$ or from $\mathcal{N}$ and X is given as input to Alice and Y to Bob. A (one-way communication) protocol $\Pi = (\Pi_A, \Pi_B)$ is a pair of functions with $\Pi_A(X) \in \{0,1\}^c$ denoting Alice's message to Bob, and $\Pi_B(\Pi_A(X), Y) \in \{\mathbf{YES}, \mathbf{NO}\}$ denoting the protocol's output. We denote this output by $\Pi(X, Y)$. The complexity of this protocol is the parameter c specifying the length of $\Pi_A(X)$ (maximized over all X). The advantage of the protocol Π is the quantity

$$\left| \Pr_{(X,Y)\sim\mathcal{Y}}[\Pi(X,Y) = \mathbf{YES}] - \Pr_{(X,Y)\sim\mathcal{N}}[\Pi(X,Y) = \mathbf{YES}] \right| .$$

The Randomized Mask Detection (RMD) communication game is an instance of such a communication game. Let $n, k \in \mathbb{N}$ and $\alpha \in (0,1)$ with $k \le n$ and $\alpha k \le 1$. Alice receives a private input $\mathbf{x}^*$ drawn uniformly at random from $\{-1,1\}^n$ while Bob receives private inputs of a k-uniform hypermatching of size αn and a vector $\mathbf{z} \in \{-1,1\}^{\alpha k n}$ of the form $\mathbf{z} = (\mathbf{z}(1), \ldots, \mathbf{z}(\alpha n))$ where $\mathbf{z}(i) \in \{-1,1\}^k$ for each $i \in [\alpha n]$. Alice's input $\mathbf{x}^*$ encodes a random bipartition of the vertex set according to the ±1 pattern. Bob's k-uniform hypermatching is encoded by a matrix $M \in \{0,1\}^{\alpha k n \times n}$ where the $(k(i-1)+1)$-th to the (ki)-th rows encode the i-th hyperedge by putting exactly one 1 in each row to the corresponding vertices. During the game, Alice sends a message to Bob and Bob has to discover the hidden structure of the vector $\mathbf{z}$. The following definition formally describes the problem.

Definition 5.2 (Randomized Mask Detection (RMD) Problem)**.** *For $k \in \mathbb{N}$, $\alpha \in (0, 1/k]$ and a pair of distributions $\mathcal{D}_Y, \mathcal{D}_N \in \Delta(\{-1,1\}^k)$, the $(\mathcal{D}_Y, \mathcal{D}_N; \alpha, k)$-RMD problem is the 2-player communication game given by a family of instances $(\mathcal{Y}_n, \mathcal{N}_n)_{n\in\mathbb{N}, n\ge 1/\alpha}$ where for a given n, $\mathcal{Y} = \mathcal{Y}_n$ and $\mathcal{N} = \mathcal{N}_n$ are as follows: Both $\mathcal{Y}$ and $\mathcal{N}$ are supported on triples $(\mathbf{x}^*, M, \mathbf{z})$ where $\mathbf{x}^* \in \{-1,1\}^n$, $M \in \{0,1\}^{k\alpha n \times n}$ and $\mathbf{z} \in \{-1,1\}^{k\alpha n}$, where $\mathbf{x}^*$ is Alice's input and the pair $(M, \mathbf{z})$ are Bob's inputs. We now specify the distributions of $\mathbf{x}^*, M$ and $\mathbf{z}$ in $\mathcal{Y}$ and $\mathcal{N}$:*

- *In both $\mathcal{Y}$ and $\mathcal{N}$, $\mathbf{x}^*$ is distributed uniformly over $\{-1,1\}^n$.*
- *In both $\mathcal{Y}$ and $\mathcal{N}$ the matrix $M \in \{0,1\}^{\alpha k n \times n}$ is chosen uniformly (and independently of*

$\mathbf{x}^*$) among matrices with exactly one 1 per row and at most one 1 per column. (Thus M represents a k-hypermatching where each block of k rows describes a hyperedge.)

- *The vector $\mathbf{z}$ is obtained by "masking" (i.e., xor-ing) $M\mathbf{x}^*$ by a random vector $\mathbf{b} \in \{-1,1\}^{\alpha kn}$ whose distribution differs in $\mathcal{Y}$ and $\mathcal{N}$. Specifically let $\mathbf{b} = (\mathbf{b}(1), \ldots, \mathbf{b}(\alpha n))$ be sampled from one of the following distributions (independent of $\mathbf{x}^*$ and M):*
  - *$\mathcal{Y}$: Each $\mathbf{b}(i) \in \{-1,1\}^k$ is sampled independently according to $\mathcal{D}_Y$.*
  - *$\mathcal{N}$: Each $\mathbf{b}(i) \in \{-1,1\}^k$ is sampled independently according to $\mathcal{D}_N$.*

  We now set $\mathbf{z} = (M\mathbf{x}^*) \odot \mathbf{b}$ (recall that that $\odot$ denotes coordinatewise product).

We will typically suppress k and α from the notation when they are clear from context and simply refer to the $(\mathcal{D}_Y, \mathcal{D}_N)$-RMD. We will refer to n as the length parameter or refer to "instances of length n" when the instances are drawn from $\mathcal{Y}_n$ vs. $\mathcal{N}_n$. The goal of a protocol solving RMD is to distinguish between case where the masks are sampled from $\mathcal{D}_Y$ from the case where the masks are sampled from $\mathcal{D}_N$ and advantage measures this probability of distinguishing.

We note that our communication game is slightly different from those in previous works: Specifically the problem studied in [GKK+09, KKS15] is called the *Boolean Hidden Matching (BHM)* problem from [GKK+09] and the works [KKSV17, KK19] study a variant called the *Implicit Hidden Partition* problem. While these problems are similar, they are less expressive than our formulation, and specifically do not seem to capture all Max-CSP(f) problems.

There are two main differences between the previous settings and our setting. The first difference is the way to encode the matching matrix M. In all the previous works, each edge (or hyperedge) is encoded by a single row in M where the corresponding columns are assigned to 1, so that $m = \alpha n$. However, it turns out that this encoding hides too much information and hence we do not know how to reduce the problem to general Max-CSP. We unfold the encoding by using k rows to encode a single k-hyperedge (leading to the setting of $m = k\alpha n$ in our case). The second difference is that we allow the masking vector $\mathbf{b}$ to be sampled from a more general distribution. This is also for the purpose of establishing a reduction to general Max-CSP. That being said, it is possible to describe some of the previous results in our language: all the papers consider the complexity of distinguishing the distribution $\mathcal{D}_Y = \mathsf{Unif}(\{(1,1),(-1,-1)\})$ from the distribution $\mathcal{D}_N = \mathsf{Unif}(\{-1,1\}^2)$. This problem is shown to have a communication lower bound of $\Omega(\sqrt{n})$ in [GKK+09]. And a variant of this problem (not captured by our formulation above) is shown to have an $\Omega(n)$ lower bound in [KK19].

Due to the above two differences, it is not clear how to derive communication lower bounds for general $\mathcal{D}_Y$ and $\mathcal{D}_N$ by reduction from the previous works. The main technical contribution of this part of the paper is a communication lower bound for RMD for general $\mathcal{D}_Y$ and $\mathcal{D}_N$. We summarize the result in the following theorem.

Theorem 5.3 (RMD Lower bound for distributions with matching marginals)**.** *For every $k \in \mathbb{N}$, there exists $\alpha_0(k) > 0$ such that for every $\alpha \in (0, \alpha_0(k))$ and $\delta > 0$ the following holds: For every pair of distributions $\mathcal{D}_Y, \mathcal{D}_N \in \Delta(\{-1,1\}^k)$ with $\boldsymbol{\mu}(\mathcal{D}_Y) = \boldsymbol{\mu}(\mathcal{D}_N)$ there exists $\tau > 0$ and n_0 such that for every $n \geq n_0$, every protocol for $(\mathcal{D}_Y, \mathcal{D}_n)$-RMD achieving advantage δ on instances of length n requires $\tau\sqrt{n}$ bits of communication.*

We prove Theorem 5.3 in two parts. First, in Section 6, we prove a communication lower bound for the special case where the marginals of $\mathcal{D}_Y$ and $\mathcal{D}_N$ are all zero. While this captures many new cases, it fails to capture the more interesting scenarios (involving non-approximation resistant problems). To get lower bounds for the general case, we reduce the 0-marginal case to the general case in Section 7.

In the rest of this section, we use Theorem 5.3 to prove Theorem 2.11 and Theorem 5.1.

5.2 The streaming lower bound

The hardness of RMD suggests a natural path for hardness of $\mathsf{Max\text{-}CSP}(f)$ problems in the streaming setting. Such a reduction would take two distributions $\mathcal{D}_Y \in S_\gamma^Y$ and $\mathcal{D}_N \in S_\beta^N$ with matching marginals, construct distributions $\mathcal{Y}$ and $\mathcal{N}$ of RMD, and then interpret these distributions (in a natural way) as distributions over instances of $\mathsf{Max\text{-}CSP}(f)$ that are indistinguishable to small space algorithms. While the exact details of this "interpretation" need to be spelled out, every step in this path can be achieved. Unfortunately this does not mean any hardness for $\mathsf{Max\text{-}CSP}(f)$ since the CSPs generated by this reduction would consist of instances that have at most one constraint per variable, and such instances are easy to solve!

To go from the instance suggested by the RMD problem to hard CSP instances, we instead pick T samples (somewhat) independently from the distributions $\mathcal{Y}$ and $\mathcal{N}$ suggested by the RMD problem and concatenate these. With an appropriate implementation of this notion (see Definition 5.4) it turns out it is possible to use the membership of the underlying distributions in S_γ^Y and S_β^N to argue that the resulting instances Ψ do (almost always) have $\mathsf{val}_\Psi \geq \gamma$ or $\mathsf{val}_\Psi \leq \beta$. (We prove this after appropriate definitions in Lemma 5.7.) But now to one needs to connect the streaming problem generated from the T-fold sampled version to the RMD problem.

To this end we formalize the T-fold streaming problem, which we call $(\mathcal{D}_Y, \mathcal{D}_N, T)$-$\mathsf{streaming\text{-}RMD}$ problem, in Definition 5.4. Unfortunately, we are not able to reduce the $(\mathcal{D}_Y, \mathcal{D}_N)$-$\mathsf{RMD}$ problem to $(\mathcal{D}_Y, \mathcal{D}_N, T)$-$\mathsf{streaming\text{-}RMD}$ problem for all $\mathcal{D}_Y$ and $\mathcal{D}_N$. (Roughly this problem arises from the fact that the T samples $(\mathbf{x}^*(t), M(t), \mathbf{z}(t))$ are not sampled independently from $\mathcal{Y}$ (or $\mathcal{N}$) for $t \in [T]$. Instead they are sampled independently conditioned on $\mathbf{x}^*(1) = \cdots = \mathbf{x}^*(T)$. This hidden correlation in *both* the **YES** and the **NO** cases turns out to be a serious problem.) But in the setting where $\mathcal{D}_Y$ and $\mathcal{D}_N$ have uniform marginals, we are able to effect the reduction and thus show that the streaming problem requires large space. This is a special case of Lemma 5.9 and Lemma 5.11 which we discuss next.

We are able to extend our reduction from RMD to $\mathsf{streaming\text{-}RMD}$ slightly beyond the uniform marginal case, to the case where $\mathcal{D}_Y$ and $\mathcal{D}_N$ form a padded one-wise pair, but both the streaming problem and the analysis of the resulting CSP value need to be altered to deal with this case, as elaborated next. Let $\tau \in [0,1]$ and $\mathcal{D}_0, \mathcal{D}'_Y, \mathcal{D}'_N$ be such that for $i \in \{Y, N\}$ we have $\mathcal{D}_i = \tau\mathcal{D}_0 + (1-\tau)\mathcal{D}'_i$ and $\mathcal{D}'_i$ has uniform marginals. Our padded streaming problem, denoted $(\mathcal{D}'_Y, \mathcal{D}'_N, T, \mathcal{D}_0, \tau)$-$\mathsf{padded\text{-}streaming\text{-}RMD}$ problem, includes an appropriately large number of constraints generated according to $\mathcal{D}_0$, followed by T samples chosen according to the $(\mathcal{D}'_Y, \mathcal{D}'_N, T)$-$\mathsf{streaming\text{-}RMD}$ problem. See Definition 5.4 for a formal definition. In Lemma 5.7 we show that the CSP value of the resulting streaming problem inherits the properties of $\mathcal{D}_Y$ and $\mathcal{D}_N$ (which is not as immediate for $\mathsf{padded\text{-}streaming\text{-}RMD}$ as for $\mathsf{streaming\text{-}RMD}$). We then show effectively that $(\mathcal{D}'_Y, \mathcal{D}'_N)$-$\mathsf{RMD}$ reduces to $(\mathcal{D}'_Y, \mathcal{D}'_N, T, \mathcal{D}_0, \tau)$-$\mathsf{padded\text{-}streaming\text{-}RMD}$. See Lemma 5.9 and Lemma 5.11. Putting these together leads to a proof of Theorem 2.11.

5.2.1 The (Padded) Streaming RMD Problem

Definition 5.4 ($(\mathcal{D}_Y, \mathcal{D}_N, T)$-streaming-RMD)**.** *For $k, T \in \mathbb{N}$, $\alpha \in (0, 1/k]$, distributions $\mathcal{D}_Y, \mathcal{D}_N$ over $\{-1,1\}^k$, the streaming problem $(\mathcal{D}_Y, \mathcal{D}_N, T; \alpha, k)$-streaming-RMD is the task of distinguishing, for every n, $\boldsymbol{\sigma} \sim \mathcal{Y}_{\textit{stream},n}$ from $\boldsymbol{\sigma} \sim \mathcal{N}_{\textit{stream},n}$ where for a given length parameter n, the distributions $\mathcal{Y}_{\textit{stream}} = \mathcal{Y}_{\textit{stream},n}$ and $\mathcal{N}_{\textit{stream}} = \mathcal{N}_{\textit{stream},n}$ are defined as follows:*

- *Let $\mathcal{Y}$ be the distribution over instances of length n, i.e., triples $(\mathbf{x}^*, M, \mathbf{z})$, from the definition of $(\mathcal{D}_Y, \mathcal{D}_N)$-RMD. For $\mathbf{x} \in \{-1,1\}^n$, let $\mathcal{Y}|_{\mathbf{x}}$ denote the distribution $\mathcal{Y}$ conditioned on $\mathbf{x}^* = \mathbf{x}$. The stream $\boldsymbol{\sigma} \sim \mathcal{Y}_{\textit{stream}}$ is sampled as follows: Sample $\mathbf{x}^*$ uniformly from $\{-1,1\}^n$. Let $(M^{(1)}, \mathbf{z}^{(1)}), \ldots, (M^{(T)}, \mathbf{z}^{(T)})$ be sampled independently according to $\mathcal{Y}|_{\mathbf{x}^*}$. Let $\boldsymbol{\sigma}^{(t)}$ be the pair $(M^{(t)}, \mathbf{z}^{(t)})$ presented as a stream of edges with labels in $\{-1,1\}^k$. Specifically for $t \in [T]$ and $i \in [\alpha n]$, let $\boldsymbol{\sigma}^{(t)}(i) = (e^{(t)}(i), \mathbf{z}^{(t)}(i))$ where $e^{(t)}(i)$ is the i-th hyperedge of $M^{(t)}$, i.e., $e^{(t)}(i) = (j^{(t)}(k(i-1)+1), \ldots, j^{(t)}(k(i-1)+k)$ and $j^{(t)}(\ell)$ is the unique index j such that $M^{(t)}_{j,\ell} = 1$. Finally we let $\boldsymbol{\sigma} = \boldsymbol{\sigma}^{(1)} \circ \cdots \circ \boldsymbol{\sigma}^{(T)}$ be the concatenation of the $\boldsymbol{\sigma}^{(t)}$s.*

- *$\boldsymbol{\sigma} \sim \mathcal{N}_{\textit{stream}}$ is sampled similarly except we now sample $(M^{(1)}, \mathbf{z}^{(1)}), \ldots, (M^{(T)}, \mathbf{z}^{(T)})$ independently according to $\mathcal{N}|_{\mathbf{x}^*}$ where $\mathcal{N}|_{\mathbf{x}}$ is the distribution $\mathcal{N}$ condition on $\mathbf{x}^* = \mathbf{x}$.*

Again when α and k are clear from context we suppress them and simply refer to the $(\mathcal{D}_Y, \mathcal{D}_N, T)$-streaming-RMD problem.

Remark 5.5. *We note that when $\mathcal{D}_N =$ Unif$(\{-1,1\}^k)$, then the distributions $\mathcal{N}|_{\mathbf{x}^*}$ are identical for all $\mathbf{x}^*$ (and the variables $\mathbf{z}^{(t)}(i)$ is distributed uniformly over $\{-1,1\}^k$ independently for every t, i).*

For technical reasons, we need the following *padded* version of streaming-RMD to extend our lower bound techniques in the streaming setting beyond uniform marginals.

Definition 5.6 ($(\mathcal{D}_Y, \mathcal{D}_N, T, \mathcal{D}_0, \tau)$-padded-streaming-RMD)**.** *For $k, T \in \mathbb{N}$, $\alpha \in (0, 1/k]$, $\tau \in [0, 1)$, distributions $\mathcal{D}_Y, \mathcal{D}_N, \mathcal{D}_0$ over $\{-1,1\}^k$, the streaming problem $(\mathcal{D}_Y, \mathcal{D}_N, T, \mathcal{D}_0, \tau; \alpha, k)$-padded-streaming-RMD is the task of distinguishing, for every n, $\boldsymbol{\sigma} \sim \mathcal{Y}_{\textit{pad-stream},n}$ from $\boldsymbol{\sigma} \sim \mathcal{N}_{\textit{pad-stream},n}$ where for a given length parameter n, the distributions $\mathcal{Y}_{\textit{pad-stream}} = \mathcal{Y}_{\textit{pad-stream},n}$ and $\mathcal{N}_{\textit{pad-stream}} = \mathcal{N}_{\textit{pad-stream},n}$ are defined as follows: Sample $\mathbf{x}^*$ from $\{-1,1\}^n$ uniformly. For each $i \in [\frac{\tau}{1-\tau}\alpha n T]$, uniformly sample a tuple $e^{(0)}(i) = (i_1, \ldots, i_k) \in \binom{[n]}{k}$ and $\mathbf{b}^{(0)}(i) \sim \mathcal{D}_0$, let $\boldsymbol{\sigma}^{(0)}(i) = (e^{(0)}(i), \mathbf{x}^*|_{e^{(0)}(i)} \odot \mathbf{b}^{(0)}(i))$. Next, sample $\boldsymbol{\sigma}^{(1)}, \ldots, \boldsymbol{\sigma}^{(T)}$ according to the Yes and No distribution of $(\mathcal{D}_Y, \mathcal{D}_N, T)$-streaming-RMD respectively. Finally, let $\boldsymbol{\sigma} = \boldsymbol{\sigma}^{(0)} \circ \cdots \circ \boldsymbol{\sigma}^{(T)}$ be the concatenation of the $\boldsymbol{\sigma}^{(t)}$s.*

Note that when $\tau = 0$, $(\mathcal{D}_Y, \mathcal{D}_N, T, \mathcal{D}_0, \tau)$-padded-streaming-RMD is the same as $(\mathcal{D}_Y, \mathcal{D}_N, T)$-streaming-RMD.

5.2.2 CSP value of padded-streaming-RMD

There is a natural way to convert instances of padded-streaming-RMD to a Max-CSP(f) problem. In this section we make this conversion explicit and show how to use properties of the underlying distributions $\mathcal{D}_0, \mathcal{D}_Y, \mathcal{D}_N$ to get bounds on the value of the instances produced.

Note that an instance $\boldsymbol{\sigma}$ of padded-streaming-RMD is simply a sequence $(\sigma(1), \ldots, \sigma(m))$ where each $\sigma(i) = (\mathbf{j}(i), \mathbf{z}(i))$ with $\mathbf{j}(i) \in [n]^k$ and $\mathbf{z}(i) \in \{-1,1\}^k$. This sequence is already syntactically very close to the description of a $\mathsf{Max\text{-}CSP}(f)$ instance. The only missing ingredient is any reference to the function f itself! Indeed the reduction from $\mathsf{padded\text{-}streaming\text{-}RMD}$ to $\mathsf{Max\text{-}CSP}(f)$ involves just applying this function f to the literals indicated by $\sigma(i)$.

Formally, given an instance $\boldsymbol{\sigma} = (\sigma(1), \ldots, \sigma(m))$ of $\mathsf{padded\text{-}streaming\text{-}RMD}$, let $\Psi(\boldsymbol{\sigma})$ denote the instance of $\mathsf{Max\text{-}CSP}(f)$ on variables $\mathbf{x} = (x_1, \ldots, x_n)$ with the constraints $C_1, \ldots, C_m$ with $C_i = \sigma(i) = (\mathbf{j}(i), \mathbf{z}(i))$ is the constraint satisfied if $f(\mathbf{z}(i) \odot \mathbf{x}|_{\mathbf{j}(i)}) = 1$.

In what follows we show that if $\tau\mathcal{D}_0 + (1-\tau)\mathcal{D}_Y \in S_\gamma^Y$ then for all sufficiently large constant T, and sufficiently large n, if we draw $\boldsymbol{\sigma} \sim \mathcal{Y}_{\text{pad-stream},n}$, then with high probability, $\Psi(\boldsymbol{\sigma})$ has value at least $\gamma - o(1)$. Conversely if $\tau\mathcal{D}_0 + (1-\tau)\mathcal{D}_N \in S_\beta^N$, then for all sufficiently large n, if we draw $\boldsymbol{\sigma} \sim \mathcal{N}_{\text{pad-stream},n}$, then with high probability $\Psi(\boldsymbol{\sigma})$ has value at most $\beta + o(1)$.

Lemma 5.7 (CSP value of $\mathsf{padded\text{-}streaming\text{-}RMD}$). *For every $k \in \mathbb{N}$, $f : \{-1,1\}^k \to \{0,1\}$, $0 \le \beta < \gamma \le 1$, $\varepsilon > 0$, $\tau = [0,1)$, distributions $\mathcal{D}_Y, \mathcal{D}_N, \mathcal{D}_0 \in \Delta(\{-1,1\}^k)$ there exists α_0 such that for every $\alpha \in (0, \alpha_0]$, there exists an integer T_0 such that for every $T \ge T_0$ the following holds:*

1. *If $\tau\mathcal{D}_0 + (1-\tau)\mathcal{D}_Y \in S_\gamma^Y$, then for every sufficiently large n, the $(\mathcal{D}_Y, \mathcal{D}_N, T, \mathcal{D}_0, \tau)$-padded-streaming-RMD **YES** instance $\boldsymbol{\sigma} \sim \mathcal{Y}_{\text{pad-stream},n}$ satisfies $\Pr[\mathsf{val}_{\Psi(\boldsymbol{\sigma})} < (\gamma - \varepsilon)] \le \exp(-n)$.*
2. *If $\tau\mathcal{D}_0 + (1-\tau)\mathcal{D}_N \in S_\beta^N$, then for every sufficiently large n, the $(\mathcal{D}_Y, \mathcal{D}_N, T, \mathcal{D}_0, \tau)$-padded-streaming-RMD **NO** instance $\boldsymbol{\sigma} \sim \mathcal{N}_{\text{pad-stream},n}$ satisfies $\Pr[\mathsf{val}_{\Psi(\boldsymbol{\sigma})} > (\beta + \varepsilon)] \le \exp(-n)$.*

Furthermore, if $\gamma = 1$ then $\Pr_{\boldsymbol{\sigma} \sim \mathcal{Y}_{\text{pad-stream},n}}\left[\mathsf{val}_{\Psi(\boldsymbol{\sigma})} = 1\right] = 1$.

Proof. We prove the lemma for $\alpha_0 = \varepsilon/(100k^2)$ and $T_0 = 1000/(\varepsilon^2\alpha)$. Roughly our proof uses the fact that the definition of S_γ^Y is setup so that $\Psi(\boldsymbol{\sigma})$ achieves value γ under the "planted" assignment $\mathbf{x}^*$. Similarly S_β^N is setup so that for every assignment, the expected value is not more than β.

We recall that the condition $\tau\mathcal{D}_0 + (1-\tau)\mathcal{D}_Y \in S_\gamma^Y$ implies that $\mathbb{E}_{\mathbf{a} \sim \tau\mathcal{D}_0 + (1-\tau)\mathcal{D}_Y}[f(\mathbf{a})] \ge \gamma$. Now consider a random **YES** instance $\boldsymbol{\sigma} \sim \mathcal{Y}_{\text{pad-stream},n}$ of $(\mathcal{D}_Y, \mathcal{D}_N, T, \mathcal{D}_0, \tau)$-$\mathsf{padded\text{-}streaming\text{-}RMD}$ and let $\mathbf{x}^*$ denote the underlying vector corresponding to this draw. We show that for $\Psi = \Psi(\boldsymbol{\sigma})$ we have $\mathsf{val}_\Psi(\mathbf{x}^*) \ge \gamma - \varepsilon$ with high probability. We consider the constraints given by $\sigma(i)$ one at a time. Let $m = \frac{\tau}{1-\tau}\alpha nT + \alpha nT = \frac{\alpha nT}{1-\tau}$ denote the total number of constraints of Ψ. Let $Z_i = C_i(\mathbf{x}^*) = f(\mathbf{z}(i) \odot \mathbf{x}^*|_{\mathbf{j}(i)})$ denote the indicator of the event that the ith constraint is satisfied by $\mathbf{x}^*$. By construction of $\mathbf{z}(i)$ (from Definition 5.2 and passed through Definition 5.4), we have $\mathbf{z}(i) = \mathbf{b}(i) \odot \mathbf{x}^*|_{\mathbf{j}(i)}$ where $\mathbf{b}(i) \sim \mathcal{D}_Y$ independently of all other choices. We thus have $Z_i = f(\mathbf{b}(i) \odot \mathbf{x}^*|_{\mathbf{j}(i)} \odot \mathbf{x}^*|_{\mathbf{j}(i)}) = f(\mathbf{b}(i))$. Thus Z_i is a random variable, chosen independently of $Z_1, \ldots, Z_{i-1}$, with expectation $\mathbb{E}[Z_i | Z_1, \ldots, Z_{i-1}] = \mathbb{E}_{\mathbf{b} \sim \mathcal{D}_0}[f(\mathbf{b})]$ when $i \le \frac{\tau}{1-\tau}\alpha nT$ and $\mathbb{E}[Z_i | Z_1, \ldots, Z_{i-1}] = \mathbb{E}_{\mathbf{b} \sim \mathcal{D}_Y}[f(\mathbf{b})]$ otherwise. In particular,

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^m Z_i\right] &= \frac{\tau}{1-\tau}\alpha nT \cdot \mathop{\mathbb{E}}_{\mathbf{b} \sim \mathcal{D}_0}[f(\mathbf{b})] + \alpha nT \cdot \mathop{\mathbb{E}}_{\mathbf{b} \sim \mathcal{D}_Y}[f(\mathbf{b})] \\ &= \frac{\alpha nT}{1-\tau} \cdot \mathop{\mathbb{E}}_{\mathbf{b} \sim \tau\mathcal{D}_0 + (1-\tau)\mathcal{D}_Y}[f(\mathbf{b})] \\ &\ge \frac{\alpha nT}{1-\tau} \cdot \gamma \;=\; \gamma \cdot m\,.\end{aligned}$$

By applying a concentration bound (Lemma 3.6 suffices, though even simpler Chernoff bounds would suffice) we get that $\Pr_{\boldsymbol{\sigma} \sim \mathcal{Y}_{\text{stream},n}}[\mathsf{val}_{\Psi(\boldsymbol{\sigma})} = \frac{1}{m}\sum_{i=1}^m Z_i < (\gamma - \varepsilon)] \le \exp(-\varepsilon^2 m) = \exp(-\varepsilon^2\alpha Tn)$. This yields Part (1) of the lemma.

Note that, if $\gamma = 1$, then $Z_i = 1$ deterministically for every i, and so we get $\mathsf{val}_\Psi = 1$ with probability 1, yielding the furthermore part of the lemma.

We now turn to the analysis of the **NO** case. By the condition $\tau\mathcal{D}_0 + (1-\tau)\mathcal{D}_N \in S_\beta^N$ we have that for every $p \in [0,1]$, we have

$$\mathop{\mathbb{E}}_{\mathbf{b}\sim\tau\mathcal{D}_0+(1-\tau)\mathcal{D}_N}\ \mathop{\mathbb{E}}_{\mathbf{a}\sim\mathsf{Bern}(p)^k}[f(\mathbf{b}\odot\mathbf{a})] \le \beta. \tag{5.8}$$

Now consider any fixed assignment $\boldsymbol{\nu} \in \{-1,1\}^n$. In what follows we show that for a random **NO** instance $\boldsymbol{\sigma} \sim \mathcal{N}_{\text{pad-stream},n}$ of $(\mathcal{D}_Y, \mathcal{D}_N, T, \mathcal{D}_0, \tau)$-padded-streaming-RMD if we let $\Psi = \Psi(\boldsymbol{\sigma})$, then $\Pr[\mathsf{val}_\Psi(\boldsymbol{\nu}) > (\beta+\varepsilon)] \le c^{-n}$ for $c > 2$. This allows us to take a union bound over the 2^n possible $\boldsymbol{\nu}$'s to claim $\Pr[\mathsf{val}_\Psi > (\beta+\varepsilon)] \le 2^n \cdot c^{-n} = o(1)$.

We thus turn to analyzing $\mathsf{val}_\Psi(\boldsymbol{\nu})$. Recall that $\boldsymbol{\sigma}$ is chosen by picking $\mathbf{x}^* \in \{-1,1\}^n$ uniformly and then picking $\sigma(i)$'s based on this choice — but our analysis will work for every choice of $\mathbf{x}^* \in \{-1,1\}^n$. Fix such a choice and let $\boldsymbol{\nu}^* = \boldsymbol{\nu}\odot\mathbf{x}^*$. Now for $i \in [m]$ (where $m = \frac{\alpha n T}{1-\tau}$) let Z_i denote the indicator of the event that $\boldsymbol{\nu}$ satisfies C_i. We have $Z_i = f(\mathbf{b}(i)\odot\mathbf{x}^*_{\mathbf{j}(i)}\odot\boldsymbol{\nu}_{\mathbf{j}(i)}) = f(\mathbf{b}(i)\odot\boldsymbol{\nu}^*|_{\mathbf{j}(i)})$. Our goal is to prove that $\Pr[\sum_{i=1}^m Z_i > (\beta+\varepsilon)\cdot m] \le c^{-n}$. To this end, let $\eta_i = \mathbb{E}[Z_i]$. Below we prove the following: (1) $\sum_{i=1}^m \eta_i \le (\beta + o(1))\cdot m$, and (2) For every i, and $Z_1,\ldots,Z_{i-1}$, $\mathbb{E}[Z_i|Z_1,\ldots,Z_{i-1}] \le \eta_i + \varepsilon/2$. With (1) and (2) in hand, a straightforward application of Azuma's inequality yields that $\Pr[\sum_i Z_i > (\beta+\varepsilon)\cdot m] \le c_1^{-m}$ for some $c_1 > 1$. Picking T large enough now ensures this is at most c^{-n} for some $c > 2$.

We start by analyzing the η_i's. Let $m_0 = \tau\cdot m$ and let $m_1 = \alpha n$ so that $m = m_0 + m_1\cdot T$. Let p be the fraction of 1's in $\boldsymbol{\nu}^*$. When $i \le m_0$, we have $\eta_i \le \mathbb{E}_{\mathbf{b}(i)\sim\mathcal{D}_0}\mathbb{E}_{\mathbf{a}\sim\mathsf{Bern}(p)^k}[f(\mathbf{b}(i)\odot\mathbf{a})] + o(1)$, where the $o(1)$ term accounts for the difference between sampling k elements from n distinct elements with repetition and without. When $i > m_0$, we have $\eta_i \le \mathbb{E}_{\mathbf{b}(i)\sim\mathcal{D}_N}\mathbb{E}_{\mathbf{a}\sim\mathsf{Bern}(p)^k}[f(\mathbf{b}(i)\odot\mathbf{a})] + o(1)$. By linearity of expectations, we now get

$$\begin{aligned}\sum_{i=1}^m \eta_i &\le \tau m\cdot \mathop{\mathbb{E}}_{\mathbf{b}(i)\sim\mathcal{D}_0}\ \mathop{\mathbb{E}}_{\mathbf{a}\sim\mathsf{Bern}(p)^k}[f(\mathbf{b}(i)\odot\mathbf{a})] + (1-\tau)m\cdot \mathop{\mathbb{E}}_{\mathbf{b}(i)\sim\mathcal{D}_N}\ \mathop{\mathbb{E}}_{\mathbf{a}\sim\mathsf{Bern}(p)^k}[f(\mathbf{b}(i)\odot\mathbf{a})] + o(1)\cdot m\\ &\le m\cdot \mathop{\mathbb{E}}_{\mathbf{b}(i)\sim\tau\mathcal{D}_0+(1-\tau)\mathcal{D}_N}\ \mathop{\mathbb{E}}_{\mathbf{a}\sim\mathsf{Bern}(p)^k}[f(\mathbf{b}(i)\odot\mathbf{a})] + o(1)\cdot m\\ &\le (\beta+o(1))\cdot m\,. \quad \text{(By Equation 5.8)}\end{aligned}$$

This yields (1).

Turning to part (2) we need to understand how the Z_i's depend on each other. For the case $i \le m_0$, note that $Z_1,\ldots,Z_{m_0}$ are independent by construction (Definition 5.6). So we have $\mathbb{E}[Z_i\,|\,Z_1,\ldots,Z_{i-1}] = \eta_i$ in this case. We now consider $i > m_0$. For $t \in [T]$, let us partition the variables $Z_{m_0+1},\ldots,Z_m$ into T block $B_1,\ldots,B_T$ with $B_t = (Z_{m_0+(t-1)m_1+1},\ldots,Z_{m_0+tm_1})$ for $t \in [T]$. By construction (Definition 5.4 via Definition 5.6) we have that the blocks $B_1,\ldots,B_T$ are identically distributed and independent conditioned on $Z_1,\ldots,Z_{m_0}$. Thus the only dependence between the Z_i's is within the Z_i's in the same block. Within a block two Z_i's may depend on each other due to the restriction that the underlying set of variables are disjoint. Thus, in particular when choosing the variables of $\boldsymbol{\sigma}^{(t)}(i')$ (corresponding to $Z_{m_0+(t-1)m_1+i'}$, some subset $S \subseteq [n]$ of the variables may already be involved in constraints of the t-th block. Let S be the set of variables not assigned to constraints in the t-th block at this time, and let p_S denote the fraction of 1's in $\boldsymbol{\nu}^*|_S$. (Note both S and p_S are random variables.) Since $|S| \ge n - k\alpha n$ and

$\alpha \leq \varepsilon/(100k^2)$ we have $|S| \geq (1-\varepsilon/(100k))n$ and so $|p - p_S| \leq \varepsilon/(100k)$. In turn this implies that $\|\mathsf{Bern}(p)^k - \mathsf{Bern}(p_S)^k\|_{tvd} \leq \varepsilon/100$. Using these bounds we now have:

$$\begin{aligned}
\mathbb{E}[Z_i|Z_1,\ldots,Z_{i-1}] &= \mathbb{E}[Z_i|S] \\
&= \mathop{\mathbb{E}}_{\mathbf{j}(i),\mathbf{b}(i)}[f(\mathbf{b}(i)\odot \boldsymbol{\nu}^*|_{\mathbf{j}(i)})] \\
&\leq \mathop{\mathbb{E}}_{\mathbf{b}(i)\sim\mathcal{D}_N}\ \mathop{\mathbb{E}}_{\mathbf{a}\sim\mathsf{Bern}(p_S)^k}[f(\mathbf{b}(i)\odot\mathbf{a})] + \frac{k^2}{|S|} \\
&\leq \mathop{\mathbb{E}}_{\mathbf{b}(i)\sim\mathcal{D}_N}\ \mathop{\mathbb{E}}_{\mathbf{a}\sim\mathsf{Bern}(p)^k}[f(\mathbf{b}(i)\odot\mathbf{a})] + \frac{\varepsilon}{100} + \frac{k^2}{|S|} \\
&\leq \mathop{\mathbb{E}}_{\mathbf{b}(i)\sim\mathcal{D}_N}\ \mathop{\mathbb{E}}_{\mathbf{a}\sim\mathsf{Bern}(p)^k}[f(\mathbf{b}(i)\odot\mathbf{a})] + \frac{\varepsilon}{2} \\
&= \eta_i + \frac{\varepsilon}{2}\,.
\end{aligned}$$

(In the second equality above $\mathbf{j}(i)$ denotes a random sequence of distinct variables from S. The next inequality comes from the difference between sampling k elements from S with repetition and without. The following inequality is the key one, using $\|\mathsf{Bern}(p)^k - \mathsf{Bern}(p_S)^k\|_{tvd} \leq \varepsilon/100$. The final inequality uses the fact that n and hence $|S|$ are large enough, and the final equality uses the value of η_i from Part (1).) This concludes Part (2).

Finally we use a version of a concentration bound for submartingales to combine (1) and (2) to get the desired bound on $\Pr[\mathsf{val}_\Psi(\boldsymbol{\nu}) > (\beta+\varepsilon)] \leq c^{-n}$. Specifically, we apply Lemma 3.6 with $N = m$, $p_i = \eta_i$ for $i \in [N]$ and $\Delta = (\varepsilon/2)N$ to conclude that $\Pr[\mathsf{val}_\Psi(\boldsymbol{\nu}) > (\beta+\varepsilon)] = \Pr[\sum_i Z_i > (\beta+\varepsilon)\cdot m] \leq \exp(-O(\varepsilon^2\alpha n T))$. Given c we can choose T to be large enough so that this is at most c^{-n}. This concludes the analysis of the **NO** case, and thus the lemma. □

5.2.3 Reduction from one-way $(\mathcal{D}_Y, \mathcal{D}_N)$-RMD to padded-streaming-RMD

We start by reducing RMD to padded-streaming-RMD in the special case where $\mathcal{D}_N = \mathsf{Unif}(\{-1,1\}^k)$. Note that since the former is hard in this case for all $\mathcal{D}_Y$ with uniform marginals, applying this argument twice shows hardness of padded-streaming-RMD for all $\mathcal{D}_Y$ and $\mathcal{D}_N$ with uniform marginals.

Lemma 5.9. *Let $T, k \in \mathbb{N}$, $\alpha \in (0, \alpha_0(k)]$, $\tau \in [0,1)$, and $\mathcal{D}_Y, \mathcal{D}_N, \mathcal{D}_0 \in \Delta(\{-1,1\}^k)$ with $\boldsymbol{\mu}(\mathcal{D}_Y) = 0^k$ and $\mathcal{D}_N = \mathsf{Unif}(\{-1,1\}^k)$. Suppose there is a streaming algorithm* **ALG** *solves $(\mathcal{D}_Y, \mathcal{D}_N, T, \mathcal{D}_0, \tau)$-padded-streaming-RMD on instances of length n with advantage Δ and space s, then there is a one-way protocol for $(\mathcal{D}_Y, \mathcal{D}_N)$-RMD on instances of length n using at most sT bits of communication achieving advantage at least Δ/T.*

The proof of Lemma 5.9 is based on a hybrid argument (*e.g.,* [KKS15, Lemma 6.3]). We provide a proof here based on the proof of [CGV20, Lemma 4.11].

Proof of Lemma 5.9. Note that since we are interested in distributional advantage, we can fix the randomness in **ALG** so that it becomes a deterministic algorithm. By an averaging argument the randomness can be chosen to ensure the advantage does not decrease. Let Γ denote the evolution of function of **ALG** as it processes a block of αn edges. That is, if the algorithm is in state s and receives a stream $\boldsymbol{\sigma}$ of length αn then it ends in state $\Gamma(s, \boldsymbol{\sigma})$. Let s_0 denote its initial state.

We consider the following collection of (jointly distributed) random variables: Let $\mathbf{x}^* \sim \mathsf{Unif}(\{-1,1\}^n)$. Denote $\mathcal{Y} = \mathcal{Y}_{\text{pad-stream},n}$ and $\mathcal{N} = \mathcal{N}_{\text{pad-stream},n}$. Let $(\boldsymbol{\sigma}_Y^{(0)}, \boldsymbol{\sigma}_Y^{(1)}, \ldots, \boldsymbol{\sigma}_Y^{(T)}) \sim \mathcal{Y}|_{\mathbf{x}^*}$. Similarly, let $(\boldsymbol{\sigma}_N^{(0)}, \boldsymbol{\sigma}_N^{(1)}, \ldots, \boldsymbol{\sigma}_N^{(T)}) \sim \mathcal{N}|_{\mathbf{x}^*}$. Recall by Remark 5.5 that since $\mathcal{D}_N$ is the uniform distribution, we have $\mathcal{N}|_{\mathbf{x}^*}$ is independent of $\mathbf{x}^*$, a feature that will be crucial to this proof.

Let S_t^Y denote the state of **ALG** after processing $\boldsymbol{\sigma}_Y^{(0)}, \ldots, \boldsymbol{\sigma}_Y^{(t)}$, i.e., $S_0^Y = \Gamma(s_0, \boldsymbol{\sigma}_Y^{(0)})$ and $S_t^Y = \Gamma(S_{t-1}^Y, \boldsymbol{\sigma}_Y^{(t)})$ where s_0 is the fixed initial state (recall that **ALG** is deterministic). Similarly let S_t^N denote the state of **ALG** after processing $\boldsymbol{\sigma}_N^{(0)}, \ldots, \boldsymbol{\sigma}_N^{(t)}$. Note that since $\boldsymbol{\sigma}_Y^{(0)}$ has the same distribution (conditioned on the same $\mathbf{x}^*$) as $\boldsymbol{\sigma}_N^{(0)}$ by definition, we have $\|S_0^Y - S_0^N\|_{tvd} = 0$.

Let $S_{a:b}^Y$ denote the sequence of states $(S_a^Y, \ldots, S_b^Y)$ and similarly for $S_{a:b}^N$. Now let $\Delta_t = \|S_{0:t}^Y - S_{0:t}^N\|_{tvd}$. Observe that $\Delta_0 = 0$ while $\Delta_T \geq \Delta$. (The latter is based on the fact that **ALG** distinguishes the two distributions with advantage Δ.) Thus $\Delta \leq \Delta_T - \Delta_0 = \sum_{t=0}^{T-1}(\Delta_{t+1} - \Delta_t)$ and so there exists $t^* \in \{0, 1, \ldots, T-1\}$ such that

$$\Delta_{t^*+1} - \Delta_{t^*} = \|S_{0:t^*+1}^Y - S_{0:t^*+1}^N\|_{tvd} - \|S_{0:t^*}^Y - S_{0:t^*}^N\|_{tvd} \geq \frac{\Delta}{T}\,.$$

Now consider the random variable $\tilde{S} = \Gamma(S_{t^*}^Y, \boldsymbol{\sigma}_N^{(t^*+1)})$ (so the previous state is from the **YES** distribution and the input is from the **NO** distribution). We claim below that $\|S_{t^*+1}^Y - \tilde{S}\|_{tvd} = \mathbb{E}_{A \sim_d S_{0:t^*}^Y}[\|S_{t^*+1}^Y|_{S_{0:t^*}^Y = A} - \tilde{S}|_{S_{0:t^*}^Y = A}\|_{tvd}] \geq \Delta_{t^*+1} - \Delta_{t^*}$. Once we have the claim, we show how to get a space $T \cdot s$ protocol for $(\mathcal{D}_Y, \mathcal{D}_n)$-RMD with advantage $\Delta_{t^*+1} - \Delta_{t^*}$ concluding the proof of the lemma.

Claim 5.10. $\|S_{t^*+1}^Y - \tilde{S}\|_{tvd} \geq \Delta_{t^*+1} - \Delta_{t^*}$.

Proof. First, by triangle inequality for the total variation distance, we have

$$\|S_{t^*+1}^Y - \tilde{S}\|_{tvd} \geq \|S_{t^*+1}^Y - S_{t^*+1}^N\|_{tvd} - \|\tilde{S} - S_{t^*+1}^N\|_{tvd}\,.$$

Recall that $\tilde{S} = \Gamma(S_{t^*}^Y, \boldsymbol{\sigma}_N^{(t^*+1)})$ and $S_{t^*+1}^N = \Gamma(S_{t^*}^N, \boldsymbol{\sigma}_N^{(t^*+1)})$. Also, note that $\boldsymbol{\sigma}_N^{(t^*+1)}$ is uniformly distributed over $(\{-1,1\}^k)^{\alpha n}$ and in particular is independent of $S_{t^*}^Y$ and $S_{t^*}^N$. (This is where we rely crucially on the property $\mathcal{D}_N = \mathsf{Unif}(\{-1,1\}^k)$.) Furthermore Γ is a deterministic function, and so we can apply the data processing inequality (Item (2) of Proposition 3.5 with $X = S_{t^*}^Y$, $Y = S_{t^*}^N$, $W = \boldsymbol{\sigma}_N^{(t^*+1)}$, and $f = \Gamma$) to conclude

$$\|\tilde{S} - S_{t^*+1}^N\|_{tvd} = \|\Gamma(S_{t^*}^Y, \boldsymbol{\sigma}_N^{(t^*+1)}) - \Gamma(S_{t^*}^N, \boldsymbol{\sigma}_N^{(t^*+1)})\|_{tvd} \leq \|S_{t^*}^Y - S_{t^*}^N\|_{tvd}.$$

Combining the two inequalities above we get

$$\|S_{t^*+1}^Y - \tilde{S}\|_{tvd} \geq \|S_{t^*+1}^Y - S_{t^*+1}^N\|_{tvd} - \|S_{t^*}^Y - S_{t^*}^N\|_{tvd} = \Delta_{t^*+1} - \Delta_{t^*}$$

as desired.

□

We now show how a protocol can be designed for $(\mathcal{D}_Y, \mathcal{D}_N)$-RMD that achieves advantage at least $\theta = \mathbb{E}_{A \sim_d S_{0:t^*}^Y}[\|S_{t^*+1}^Y|_{S_{0:t^*} = A} - \tilde{S}|_{S_{0:t^*} = A}\|_{tvd}] \geq \Delta_{t^*+1} - \Delta_{t^*}$ concluding the proof of the lemma. The protocol uses the distinguisher $T_A : \{0,1\}^s \to \{0,1\}$ such that $\mathbb{E}_{A, S_{t^*+1}^Y, \tilde{S}}[T_A(S_{t^*+1}^Y)] - \mathbb{E}[T_A(\tilde{S})] \geq \theta$ which is guaranteed to exist by the definition of total variation distance.

Our protocol works as follows: Let Alice receive input $\mathbf{x}^*$ and Bob receive inputs $(M, \mathbf{z})$ sampled from either $\mathcal{Y}_{\mathsf{RMD}}|_{\mathbf{x}^*}$ or $\mathcal{N}_{\mathsf{RMD}}|_{\mathbf{x}^*}$ where $\mathcal{Y}_{\mathsf{RMD}}$ and $\mathcal{N}_{\mathsf{RMD}}$ are the Yes and No distribution of $(\mathcal{D}_Y, \mathcal{D}_N)$-RMD respectively.

1. Alice samples $(\boldsymbol{\sigma}^{(0)}, \boldsymbol{\sigma}^{(1)}, \ldots, \boldsymbol{\sigma}^{(T)}) \sim \mathcal{Y}|_{\mathbf{x}^*}$ and computes $A = S^Y_{0:t^*} \in \{0,1\}^{(t^*+1)s}$ and sends A to Bob.
2. Bob extracts $S^Y_{t^*}$ from A, computes $\widehat{S} = \Gamma(S^Y_{t^*}, \boldsymbol{\sigma})$, where $\boldsymbol{\sigma}$ is the encoding of $(M, \mathbf{z})$ as a stream, and outputs **YES** if $T_A(\widehat{S}) = 1$ and **NO** otherwise.

Note that if $(M, \mathbf{z}) \sim \mathcal{Y}_{\mathsf{RMD}}|_{\mathbf{x}^*}$ then $\widehat{S} \sim_d S^Y_{t^*+1}|_{S^Y_{0:t^*}=A}$ while if $(M, \mathbf{z}) \sim \mathcal{N}_{\mathsf{RMD}}|_{\mathbf{x}^*}$ then $\widehat{S} \sim \tilde{S}_{S^Y_{0:t^*}=A}$. It follows that the advantage of the protocol above exactly equals $\mathbb{E}_A[T_A(S^Y_{t+1})] - \mathbb{E}_A[T_A(\tilde{S})] \geq \theta \geq \Delta_{t^*+1} - \Delta_{t^*} \geq \Delta/T$. This concludes the proof of the lemma. □

By combining Lemma 5.9 with Theorem 5.3, we immediately have the following consequence.

Lemma 5.11. *For $k \in \mathbb{N}$ let $\alpha_0(k)$ be as given by Theorem 5.3. Let $T \in \mathbb{N}$, $\alpha \in (0, \alpha_0(k)]$, $\tau \in [0,1)$, and $\mathcal{D}_0, \mathcal{D}_Y, \mathcal{D}_N$ be three distributions over $\{-1,1\}^k$ with $\boldsymbol{\mu}(\mathcal{D}_Y) = \boldsymbol{\mu}(\mathcal{D}_N) = 0^k$. Then every streaming algorithm* **ALG** *solving $(\mathcal{D}_Y, \mathcal{D}_N, T, \mathcal{D}_0, \tau)$-padded-streaming-RMD with advantage $1/8$ for all lengths uses space $\Omega(\sqrt{n})$.*

Proof. Let **ALG** be an algorithm using space s solving $(\mathcal{D}_Y, \mathcal{D}_N, T)$-padded-streaming-RMD with advantage $1/8$. Let $\mathcal{D}_M = \mathsf{Unif}(\{-1,1\}^k)$. Then by the triangle inequality **ALG** solves either the $(\mathcal{D}_Y, \mathcal{D}_M, T, \mathcal{D}_0, \tau)$-padded-streaming-RMD with advantage $1/16$ or it solves the $(\mathcal{D}_N, \mathcal{D}_M, T, \mathcal{D}_0, \tau)$-padded-streaming-RMD with advantage $1/16$. Assume without loss of generality it is the former. Then by Lemma 5.9, there exists a one-way protocol for $(\mathcal{D}_Y, \mathcal{D}_M)$-RMD using at most sT bits of communication with advantage at least $1/(16T)$. Applying Theorem 5.3 with $\delta = 1/(16T) > 0$, we now get that $s = \Omega(\sqrt{n})$.

□

5.2.4 Proof of the streaming lower bound

We are now ready to prove Theorem 2.11.

Proof of Theorem 2.11. We combine Theorem 5.3, Lemma 5.11 and Lemma 5.7. So in particular we set our parameters α and T so that the conditions of these statements are satisfied. Specifically k and $\varepsilon > 0$, let $\alpha_0^{(1)}$ be the constant from Theorem 5.3 and let $\alpha_0^{(2)}$ be the constant from Lemma 5.7. Let $\alpha_0 = \min\{\alpha_0^{(1)}, \alpha_0^{(2)}\}$, Given $\alpha \in (0, \alpha_0)$ let T_0 be the constant from Lemma 5.7 and let $T = T_0$. (Note that these choices allow for both Theorem 5.3 and Lemma 5.7 to hold.) Suppose there exists a streaming algorithm **ALG** that solves $(\gamma - \varepsilon, \beta + \varepsilon)$-Max-CSP$(f)$. Let $\tau \in [0,1)$ and $\mathcal{D}_Y, \mathcal{D}_N, \mathcal{D}_0$ be distributions such that (i) $\boldsymbol{\mu}(\mathcal{D}_Y) = \boldsymbol{\mu}(\mathcal{D}_N) = 0^k$, (ii) $\tau\mathcal{D}_0 + (1-\tau)\mathcal{D}_Y \in S^Y_\gamma(f)$, (iii) $\tau\mathcal{D}_0 + (1-\tau)\mathcal{D}_N \in S^N_\beta(f)$, and (iv) $\boldsymbol{\mu} = \tau\boldsymbol{\mu}(\mathcal{D}_0)$. Let n be sufficiently large and let $\mathcal{Y}_{\text{stream},n}$ and $\mathcal{N}_{\text{stream},n}$ denote the distributions of **YES** and **NO** instances of $(\mathcal{D}_Y, \mathcal{D}_N, T, \mathcal{D}_0, \tau)$-padded-streaming-RMD of length n. Since α and T satisfy the conditions of Lemma 5.7, we have for every sufficiently large n

$$\Pr_{\boldsymbol{\sigma} \sim \mathcal{Y}_{\text{stream},n}} \left[\mathsf{val}_{\Psi(\boldsymbol{\sigma})} < (\gamma - \varepsilon)\right] = o(1) \quad \text{and} \quad \Pr_{\boldsymbol{\sigma} \sim \mathcal{N}_{\text{stream},n}} \left[\mathsf{val}_{\Psi(\boldsymbol{\sigma})} > (\beta + \varepsilon)\right] = o(1)\,.$$

We conclude that **ALG** can distinguish **YES** instances of $\mathsf{Max\text{-}CSP}(f)$ from **NO** instances with advantage at least $1/4 - o(1) \geq 1/8$. However, since $\mathcal{D}_Y, \mathcal{D}_N$ and α satisfy the conditions of Lemma 5.11 (in particular $\boldsymbol{\mu}(\mathcal{D}_Y) = \boldsymbol{\mu}(\mathcal{D}_N) = 0^k$ and $\alpha \in (0, \alpha_0(k))$) such an algorithm requires space at least $\Omega(\sqrt{n})$. Thus, we conclude that any streaming algorithm that solves $(\gamma - \varepsilon, \beta + \varepsilon)$-$\mathsf{Max\text{-}CSP}(f)$ requires $\Omega(\sqrt{n})$ space.

Finally, note that if $\gamma = 1$ then in Lemma 5.7, we have $\mathsf{val}_\Psi = 1$ with probability one. Repeating the above reasoning with this information, shows that $(1, \beta+\varepsilon)-\mathsf{Max\text{-}CSP}(f)$ requires $\Omega(\sqrt{n})$-space. □

5.3 The lower bound against sketching algorithms

In the absence of a reduction from RMD to streaming-RMD for general $\mathcal{D}_Y$ and $\mathcal{D}_N$, we turn to other means of using the hardness of RMD. In particular, we use lower bounds on the communication complexity of a T-player communication game in the *simultaneous communication* setting — one which is significantly easier to obtain lower bounds for than the one-way setting. Below we describe a family of T-player simultaneous communication games, which we call $(\mathcal{D}_Y, \mathcal{D}_N, T)$-simultaneous-RMD (See Definition 5.12.) We then show a simple reduction from $(\mathcal{D}_Y, \mathcal{D}_N)$-RMD to $(\mathcal{D}_Y, \mathcal{D}_N, T)$-simultaneous-RMD. Combining this reduction with our lower bounds on RMD and the reduction from $(\mathcal{D}_Y, \mathcal{D}_N, T)$-simultaneous-RMD to streaming complexity leads to the proof of Theorem 5.1.

5.3.1 T-Player Simultaneous Version of RMD

In this section, we consider the complexity of *T-player number-in-hand simultaneous message passing communication games* (abbrev. T-player simultaneous communication games). Such games are described by two distributions $\mathcal{Y}$ and $\mathcal{N}$. An instance of the game is a T-tuple $(X^{(1)}, \ldots, X^{(T)})$ either drawn from $\mathcal{Y}$ or from $\mathcal{N}$ and $X^{(t)}$ is given as input to the t-th player. A (simultaneous communication) protocol $\Pi = (\Pi^{(1)}, \ldots, \Pi^{(T)}, \Pi_{\text{ref}})$ is a $(T + 1)$-tuple of functions with $\Pi^{(t)}(X^{(t)}) \in \{0,1\}^c$ denoting the t-th player's message to the *referee*, and $\Pi_{\text{ref}}(\Pi^{(1)}(X^{(1)}), \ldots, \Pi^{(T)}(X^{(T)})) \in \{\mathbf{YES}, \mathbf{NO}\}$ denoting the protocol's output. We denote this output by $\Pi(X^{(1)}, \ldots, X^{(T)})$. The complexity of this protocol is the parameter c specifying the maximum length of $\Pi^{(1)}(X^{(1)}), \ldots, \Pi^{(T)}(X^{(T)})$ (maximized over all X). The advantage of the protocol Π is the quantity

$$\left| \Pr_{(X^{(1)},\ldots,X^{(T)})\sim\mathcal{Y}}[\Pi(X^{(1)}, \ldots, X^{(T)}) = \mathbf{YES}] - \Pr_{(X^{(1)},\ldots,X^{(T)})\sim\mathcal{N}}[\Pi(X^{(1)}, \ldots, X^{(T)}) = \mathbf{YES}] \right|.$$

Definition 5.12 ($(\mathcal{D}_Y, \mathcal{D}_N, T)$-simultaneous-RMD)**.** *For $k, T \in \mathbb{N}$, $\alpha \in (0, 1/k]$, distributions $\mathcal{D}_Y, \mathcal{D}_N$ over $\{-1,1\}^k$, the $(\mathcal{D}_Y, \mathcal{D}_N, T)$-simultaneous-RMD is a T-player communication game given by a family of instances $(\mathcal{Y}_{simul,n}, \mathcal{N}_{simul,n})_{n\in\mathbb{N}, n\geq 1/\alpha}$ where for a given n, $\mathcal{Y} = \mathcal{Y}_{simul,n}$ and $\mathcal{N} = \mathcal{N}_{simul,n}$ are as follows: Both $\mathcal{Y}$ and $\mathcal{N}$ are supported on tuples $(\mathbf{x}^*, M^{(1)}, \ldots, M^{(T)}, \mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(T)})$ where $\mathbf{x}^* \in \{-1,1\}^n$, $M^{(t)} \in \{0,1\}^{k\alpha n\times n}$, and $\mathbf{z}^{(t)} \in \{-1,1\}^{k\alpha n}$, where the pair $(M^{(t)}, \mathbf{z}^{(t)})$ are the t-th player's inputs for all $t \in [T]$. We now specify the distributions of $\mathbf{x}^*$, $M^{(t)}$, and $\mathbf{z}^{(t)}$ in $\mathcal{Y}$ and $\mathcal{N}$:*

- *In both $\mathcal{Y}$ and $\mathcal{N}$, $\mathbf{x}^*$ is distributed uniformly over $\{-1,1\}^n$.*
- *In both $\mathcal{Y}$ and $\mathcal{N}$, the matrix $M^{(t)} \in \{0,1\}^{\alpha kn\times n}$ is chosen uniformly (and independently of $\mathbf{x}^*$) among matrices with exactly one 1 per row and at most one 1 per column.*

- *The vector $\mathbf{z}^{(t)}$ is obtained by "masking" (i.e., xor-ing) $M^{(t)}\mathbf{x}^*$ by a random vector $\mathbf{b}^{(t)} \in \{-1,1\}^{\alpha kn}$ whose distribution differs in $\mathcal{Y}$ and $\mathcal{N}$. Specifically, let $\mathbf{b}^{(t)} = (\mathbf{b}^{(t)}(1), \ldots, \mathbf{b}^{(t)}(\alpha n))$ be sampled from one of the following distributions (independent of $\mathbf{x}^*$ and M):*
  - *$\mathcal{Y}$: Each $\mathbf{b}^{(t)}(i) \in \{-1,1\}^k$ is sampled independently according to $\mathcal{D}_Y$.*
  - *$\mathcal{N}$: Each $\mathbf{b}^{(t)}(i) \in \{-1,1\}^k$ is sampled independently according to $\mathcal{D}_N$.*

  We now set $\mathbf{z}^{(t)} = (M^{(t)}\mathbf{x}^*) \odot \mathbf{b}^{(t)}$ (recall that that $\odot$ denotes coordinatewise product).

Given an instance $\boldsymbol{\sigma} = (\mathbf{x}^*, M^{(1)}, \ldots, M^{(T)}, \mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(T)})$ and a function $f : \{-1,1\}^k \to \{0,1\}$, we will let $\Psi(\boldsymbol{\sigma})$ represent the instance of Max-CSP(f) it corresponds to, presented as a stream of $T\alpha n$ constraints.

Note that the instance $\Psi(\boldsymbol{\sigma})$ obtained in the **YES** and **NO** cases of $(\mathcal{D}_Y, \mathcal{D}_N, T)$-simultaneous-RMD are distributed exactly according to instances derived in the **YES** and **NO** cases of $(\mathcal{D}_Y, \mathcal{D}_N, T, \mathcal{D}_0, \tau = 0)$-padded-streaming-RMD and thus Lemma 5.7 can still be applied to conclude that **YES** instances usually have $\mathsf{val}_{\Psi(\boldsymbol{\sigma})} \geq \gamma - o(1)$ and **NO** instances usually have $\mathsf{val}_{\Psi(\boldsymbol{\sigma})} \leq \beta - o(1)$. We will use this property when proving Theorem 5.1.

We start by showing the simultaneous-RMD problems above do not have low-communication protocols when the marginals of $\mathcal{D}_Y$ and $\mathcal{D}_N$ match.

Lemma 5.13. *Let $k, T \in \mathbb{N}$, $\mathcal{D}_Y, \mathcal{D}_N \in \Delta(\{-1,1\}^k)$, and let $\alpha \in (0, 1/k]$. Suppose there is a protocol Π that solves $(\mathcal{D}_Y, \mathcal{D}_N, T)$-simultaneous-RMD on instances of length n with advantage Δ and space s, then there is a one-way protocol for $(\mathcal{D}_Y, \mathcal{D}_N)$-RMD on instances of length n using at most $s(T-1)$ bits of communication achieving advantage at least Δ/T.*

Proof. Let us first fix the randomness in Π so that it becomes a deterministic protocol. Note that by an averaging argument the advantage of Π does not decrease. Recall that $\mathcal{Y}$ and $\mathcal{N}$ are Yes and No input distribution of $(\mathcal{D}_Y, \mathcal{D}_N, T)$-simultaneous-RMD and we have

$$\Pr_{X\sim\mathcal{Y}}[\Pi(X) = \mathbf{YES}] - \Pr_{X\sim\mathcal{N}}[\Pi(X) = \mathbf{YES}] \geq \Delta\,.$$

Now, we define the following distributions $\mathcal{D}_0, \ldots, \mathcal{D}_T$. Let $\mathcal{D}_0 = \mathcal{Y}$ and $\mathcal{D}_T = \mathcal{N}$. For each $t \in [T-1]$, we define $\mathcal{D}_t$ to be the distribution of input instances of $(\mathcal{D}_Y, \mathcal{D}_N, T)$-simultaneous-RMD by sampling $\mathbf{b}^{(t')}(i)$ independently according to $\mathcal{D}_Y$ (resp. $\mathcal{D}_N$) for all $t' \leq t$ (resp. $t' > t$) and i (see Definition 5.12 to recall the definition). Next, for each $t \in [T]$, let

$$\Delta_t = \Pr_{X\sim\mathcal{D}_t}[\Pi(X) = \mathbf{YES}] - \Pr_{X\sim\mathcal{D}_{t-1}}[\Pi(X) = \mathbf{YES}]\,.$$

Observe that $\sum_{t\in[T]} \Delta_t = \Delta$ and hence there exists $t^* \in [T]$ such that $\Delta_{t^*} \geq \Delta/T$.

Now, we describe a protocol Π' for $(\mathcal{D}_Y, \mathcal{D}_N)$-RMD as follows. On input $(\mathbf{x}^*, M, \mathbf{z})$, Alice receives $\mathbf{x}^*$ and Bob receives $(M, \mathbf{z})$. Alice first samples matrices $M^{(1)}, \ldots, M^{(t^*-1)}, M^{(t^*+1)}, \ldots, M^{(T)}$ as the second item in Definition 5.12. Next, Alice samples $\mathbf{b}^{(t')}$ according to $\mathcal{D}_Y$ (resp. $\mathcal{D}_N$) for all $t' < t^*$ (resp. $t' > t^*$) and sets $\mathbf{z}^{(t')} = (M^{(t')}\mathbf{x}^*) \odot \mathbf{b}^{(t')}$ as the third item in Definition 5.12. Note that this is doable for Alice because she possesses $\mathbf{x}^*$. Finally, Alice sends $\{\Pi^{(t')}(M^{(t')}, \mathbf{z}^{(t')})\}_{t'\in[T]\setminus\{t^*\}}$ to Bob. After receiving Alice's message $(X^{(1)}, \ldots, X^{(t^*-1)}, X^{(t^*+1)}, \ldots, X^{(T)})$, Bob computes $\Pi^{(t^*)}(M, \mathbf{z})$ and outputs $\Pi'(M, \mathbf{z}) = \Pi_{\mathrm{ref}}(X^{(1)}, \ldots, X^{(t^*-1)}, \Pi^{(t^*)}(M, \mathbf{z}), X^{(t^*+1)}, \ldots, X^{(T)})$.

It is clear from the construction that the protocol Π' uses at most $s(T-1)$ bits of communication. To see Π' has advantage at least Δ/T, note that if $(\mathbf{x}^*, M, \mathbf{z})$ is sampled from the Yes distribution $\mathcal{Y}_{\mathsf{RMD}}$ of $(\mathcal{D}_Y, \mathcal{D}_N)$-$\mathsf{RMD}$, then $((M^{(1)}, \mathbf{z}^{(1)}), \ldots, (M^{(t^*-1)}, \mathbf{z}^{(t^*-1)}), (M, \mathbf{z}), (M^{(t^*+1)}, \mathbf{z}^{(t^*+1)}), \ldots, (M^{(T)}, \mathbf{z}^{(T)}))$ follows the distribution $\mathcal{D}_{t^*}$. Similarly, if $(\mathbf{x}^*, M, \mathbf{z})$ is sampled from the No distribution $\mathcal{N}_{\mathsf{RMD}}$ of $(\mathcal{D}_Y, \mathcal{D}_N)$-$\mathsf{RMD}$, then $((M^{(1)}, \mathbf{z}^{(1)}), \ldots, (M^{(t^*-1)}, \mathbf{z}^{(t^*-1)}), (M, \mathbf{z}), (M^{(t^*+1)}, \mathbf{z}^{(t^*+1)}), \ldots, (M^{(T)}, \mathbf{z}^{(T)}))$ follows the distribution $\mathcal{D}_{t^*-1}$. Thus, the advantage of Π' is at least

$$\begin{aligned}&\Pr_{(M,\mathbf{z})\sim\mathcal{Y}_{\mathsf{RMD}},\Pi'}[\Pi'(M,\mathbf{z}) = \mathbf{YES}] - \Pr_{(M,\mathbf{z})\sim\mathcal{N}_{\mathsf{RMD}},\Pi'}[\Pi'(M,\mathbf{z}) = \mathbf{YES}] \\ &= \Pr_{X\sim\mathcal{D}_{t^*}}[\Pi(X) = \mathbf{YES}] - \Pr_{X\sim\mathcal{D}_{t^*-1}}[\Pi(X) = \mathbf{YES}] = \Delta_{t^*} \geq \Delta/T\,.\end{aligned}$$

We conclude that there is a one-way protocol for $(\mathcal{D}_Y, \mathcal{D}_N)$-$\mathsf{RMD}$ using at most $s(T-1)$ bits of communication achieving advantage at least Δ/T. □

As an immediate consequence of Theorem 5.3 and Lemma 5.13 we get that $(\mathcal{D}_Y, \mathcal{D}_N, T)$-$\mathsf{simultaneous}$-$\mathsf{RMD}$ requires $\Omega(\sqrt{n})$ bits of communication when the marginals of $\mathcal{D}_Y$ and $\mathcal{D}_N$ match.

Lemma 5.14. *For every $k \in \mathbb{N}$, there exists $\alpha_0 > 0$ such that for every $\alpha \in (0, \alpha_0)$ and $\delta > 0$ the following holds: For every $T \in \mathbb{N}$ and every pair of distributions $\mathcal{D}_Y, \mathcal{D}_N \in \Delta(\{-1,1\}^k)$ with $\boldsymbol{\mu}(\mathcal{D}_Y) = \boldsymbol{\mu}(\mathcal{D}_N)$, there exists $\tau > 0$ and n_0 such that for every $n \geq n_0$, every protocol for $(\mathcal{D}_Y, \mathcal{D}_N, T)$-$\mathsf{simultaneous}$-$\mathsf{RMD}$ achieving advantage δ on instances of length n requires $\tau\sqrt{n}$ bits of communication.*

We are now ready to prove Theorem 5.1.

5.3.2 Proof of Theorem 5.1

Proof of Theorem 5.1. The proof is a straightforward combination of Lemma 5.7 and Lemma 5.14 and so we pick parameters so that all these are applicable. Given ε and k, let $\alpha_0^{(1)}$ be as given by Lemma 5.7 and let $\alpha_0^{(2)}$ be as given by Lemma 5.14. Let $\alpha = \min\{\alpha_0^{(1)}, \alpha_0^{(2)}\}$. Given this choice of α, let T_0 be as given by Lemma 5.7. We set $T = T_0$ below. Let n be sufficiently large.

Throughout this proof we will be considering integer weighted instances of Max-$\mathsf{CSP}(f)$ on n variables with constraints. Note that such an instance Ψ can be viewed as a vector in $\mathbb{Z}^N$ where $N = O(n^k)$ represents the number of possibly distinct constraints applications on n variables. Let $\Gamma = \{\Psi | \mathsf{val}_\Psi \geq \gamma - \varepsilon\}$. Let $B = \{\Psi | \mathsf{val}_\Psi \leq \beta + \varepsilon\}$. Suppose there exists a sketching algorithm $\mathbf{ALG}_1$ that solves $(\gamma - \varepsilon, \beta + \varepsilon)$-$\mathsf{Max}$-$\mathsf{CSP}(f)$ using at most $s(n)$ bits of space. Note that $\mathbf{ALG}_1$ must achieve advantage at least $1/3$ on the problem (Γ, B). By running several independent copies of $\mathbf{ALG}_1$ and thresholding appropriately, we can get an algorithm $\mathbf{ALG}_2$ with space $O(s)$ and advantage $1 - \frac{1}{100}$ solving (Γ, B).

Now, let COMP and COMB be the compression and combination functions as given by this sketching algorithm (see Definition 3.3). We use these to design a protocol for $(\mathcal{F}, \mathcal{D}_Y, \mathcal{D}_N, T)$-$\mathsf{simultaneous}$-$\mathsf{RMD}$ as follows.

Let $(M^{(t)}, \mathbf{z}^{(t)})$ denote the input to the t-th player in $(\mathcal{D}_Y, \mathcal{D}_N, T)$-$\mathsf{simultaneous}$-$\mathsf{RMD}$. Each player turn his/her inputs into $\boldsymbol{\sigma}^{(t)} = (\sigma_1^{(t)}, \ldots, \sigma_{\alpha n}^{(t)})$ where $\sigma_i^{(t)}$ corresponds to the constraint $(\mathbf{j}^{(t)}(i), \mathbf{z}_i^{(t)})$ with $\mathbf{j}_i^{(t)} \in [n]^k$ the indicator vector for the i-th hyperedge of $M^{(t)}$. Next, the players use shared randomness to compute the sketch of his/her input $\mathsf{COMP}(\boldsymbol{\sigma}^{(t)})$

and send it to the referee. Finally, the referee computes the sketch for all streams $\mathsf{COMB}(\mathsf{COMP}(\boldsymbol{\sigma}^{(1)}), \mathsf{COMP}(\boldsymbol{\sigma}^{(2)}), \ldots, \mathsf{COMP}(\boldsymbol{\sigma}^{(T)})$ and outputs the corresponding answer.

To analyze the above, note that the communication is $O(s)$. Next, by the advantage of the sketching algorithm, we have that

$$\min_{\Psi \in \Gamma}[\mathbf{ALG}_2(\Psi) = 1] - \max_{\Psi \in B}[\mathbf{ALG}_2(\Psi) = 1] \geq 1 - 12/100. \tag{5.15}$$

Now we consider what happens when $\Psi \sim \mathcal{Y}_{\text{simul},n}$ and $\Psi \sim \mathcal{N}_{\text{simul},n}$. By Lemma 5.7 we have that $\Pr_{\Psi \sim \mathcal{Y}_{\text{simul},n}}[\Psi \in \Gamma] \geq 1 - o(1)$ and $\Pr_{\Psi \sim \mathcal{N}_{\text{simul},n}}[\Psi \in B] \geq 1 - o(1)$. Combining with Equation 5.15 we thus get

$$\Pr_{\Psi \sim \mathcal{Y}_{\text{simul},n}}[\mathbf{ALG}_2(\Psi) = 1] - \Pr_{\Psi \sim \mathcal{N}_{\text{simul},n}}[\mathbf{ALG}_2(\Psi) = 1] \geq 1 - 12/100 - o(1) \geq 1/2,$$

We thus get that there is a $O(s)$ simultaneous communication protocol for $(\mathcal{D}_Y, \mathcal{D}_N, T)$-$\mathsf{simultaneous\text{-}RMD}$ with advantage at least $1/2$.

Now we conclude by applying Lemma 5.14 with $\delta = 1/2$ to get that $s = \Omega(\sqrt{n})/T = \Omega(\sqrt{n})$, thus yielding the theorem. □

6 Communication Lower Bound: A Special Case of 1-wise Independence

The goal of this section is to prove a special case of Theorem 5.3 when the distributions are 1-wise independent, i.e., their marginals are all 0. The main theorem of this section is summarized below.

Theorem 6.1 (Lower bound for 1-wise distributions)**.** *For every $k \geq 2$, there exists an $\alpha_0 > 0$ such that for every $\alpha \in (0, 1/\alpha_0)$, $\delta \in (0, 1/2)$, and every $\mathcal{D}_Y, \mathcal{D}_N \in \Delta(\{-1,1\}^k)$ with $\mu(\mathcal{D}_Y) = \mu(\mathcal{D}_N) = 0^k$, there exists $\tau > 0$, and n_0 such that for every $n \geq n_0$, we have that every protocol for $(\mathcal{D}_Y, \mathcal{D}_N)$-RMD with parameter α that achieves advantage δ requires at least $\tau\sqrt{n}$ bits of communication on instances of length n.*

Our proof of Theorem 6.1 follows the methodology of [GKK$^+$09] with minor modifications as required by the RMD formulation. Their proof uses Fourier analysis to reduce the task of proving a communication lower bound to that of proving some combinatorial identities about randomly chosen matchings. We follow the same approach and this leads us to slightly different conditions about randomly chosen hypermatchings which requires a fresh analysis (though at the end our bounds are qualitatively similar to those in [GKK$^+$09]).

The proof is by contradiction. We show that if the number of bits communicated is $o(\sqrt{n})$, then the *posterior distribution* of Bob's input $\mathbf{z}$ is close to the uniform distribution in total variation distance, and hence contradicts the assumed advantage of the protocol. In Theorem 6.2 we show that this total variation distance is small when Alice's message is a "typical" one, in that the number of Alice inputs leading to this message is not too small. We show immediately after stating Theorem 6.2 how to go from the case of typical messages to all messages, and this gives a proof of Theorem 6.1.

For each k-uniform hypermatching M, distribution $\mathcal{D}$ over $\{-1,1\}^k$, and a fixed Alice's message, the posterior distribution function $p_{M,\mathcal{D}} : \{-1,1\}^{\alpha k n} \to [0,1]$ is defined as follows. For each

$\mathbf{z} \in \{-1,1\}^{\alpha k n}$, let

$$p_{M,\mathcal{D}}(\mathbf{z}) := \Pr_{\substack{\mathbf{x}^* \in \{-1,1\}^n \\ \mathbf{b} \sim \mathcal{D}^{\alpha n}}}[\mathbf{z} = (M\mathbf{x}^*) \odot \mathbf{b} \mid M, \text{ Alice's message}] = \mathop{\mathbb{E}}_{\mathbf{x}^* \in A} \mathop{\mathbb{E}}_{\mathbf{b} \sim \mathcal{D}^{\alpha n}} [\mathbf{1}_{\mathbf{z} = (M\mathbf{x}^*) \odot \mathbf{b}}] \,,$$

where $A \subset \{-1,1\}^n$ is the set of Alice's inputs that correspond to the message. If the number of bits communicated is at most c, then there exists a message such that the corresponding A satisfies $|A| \geq 2^{n-c}$.

Theorem 6.2. *For every $k \in \mathbb{N}$, there exists $\alpha_0 > 0$ such that for every $\alpha \in (0, \alpha_0)$, $\delta \in (0, 1/2)$, there exists a $\tau_0 > 0$ such that the following holds for every sufficiently large n. Let $A \subseteq \{-1,1\}^n$ be a set satisfying $|A| \geq 2^{n-\tau_0\sqrt{n}}$, and let $\mathcal{D}$ be a distribution over $\{-1,1\}^k$ satisfying $\mathbb{E}_{\mathbf{a}\sim\mathcal{D}}[a_j] = 0$ for all $j \in [k]$. Then*

$$\mathop{\mathbb{E}}_{M}\left[\|p_{M,\mathcal{D}} - U\|_{tvd}^2\right] \leq \delta^2 \,, \tag{6.3}$$

where U denotes the uniform distribution over $\{-1,1\}^{k\alpha n}$.

Assuming Theorem 6.2, we prove Theorem 6.1 below.

Proof of Theorem 6.1. Let δ be as in the theorem statement and let $\delta' = \delta/8$. Let τ_0 be the constant given by Theorem 6.2 when invoked with parameter α and δ'. Let $\tau = \tau_0/2$, $c' = \tau_0\sqrt{n}$, and $c = c' - \log(1/\delta')$. Note that for large enough n, we have $c \geq \tau\sqrt{n}$.

We will prove the theorem for this choice of τ. The proof is by contradiction. Suppose there exists a protocol for $(\mathcal{D}_Y, \mathcal{D}_N)$-$\mathsf{RMD}$ on instances of length n with advantage at least δ using at most c bits of communication. Let $\mathcal{D}_{unif}$ be the uniform distribution over $\{-1,1\}^k$. By triangle inequality, there is a protocol for either $(\mathcal{D}_Y, \mathcal{D}_{unif})$-$\mathsf{RMD}$ or $(\mathcal{D}_N, \mathcal{D}_{unif})$-$\mathsf{RMD}$ with advantage at least $\frac{\delta}{2}$ using at most c bits of communication. Without loss of generality, suppose there is protocol for $(\mathcal{D}_Y, \mathcal{D}_{unif})$-$\mathsf{RMD}$ with advantage at least $\frac{\delta}{2}$. We have

$$\|p_{M,\mathcal{D}_Y} - p_{M,\mathcal{D}_{unif}}\|_{tvd} \geq \frac{\delta}{2} \,.$$

Next, by Yao's principle [Yao77] we may assume that the message sent by Alice is deterministic. Namely, the message partitions the set $\{-1,1\}^n$ of $\mathbf{x}^*$ into 2^c sets $A_1, A_2, \ldots, A_{2^c}$. Using a simple counting argument, we can show that with probability at least $1 - \delta'$, the message sent by Alice corresponds to a set $A_i \subset \{-1,1\}^n$ of size at least $2^{n-c-\log 1/\delta'} \geq 2^{n-c'}$. We call such an event GOOD. That is,

$$\mathsf{GOOD} = \bigcup_{i \in [2^c] : |A_i| \geq 2^{n-c'}} A_i \,.$$

Now for each A_i with $|A_i| \geq 2^{n-c'}$, we apply Theorem 6.2 with parameters α and δ' to get

$$\|p_{M,\mathcal{D}_Y} - p_{M,\mathcal{D}_{unif}}\|_{tvd}|_{\mathbf{x}^* \in A_i} = \mathop{\mathbb{E}}_{M}[\|p_{M,\mathcal{D}_Y} - U\|_{tvd}|_{\mathbf{x}^* \in A_i}] \leq \delta' \,.$$

Now, for $\mathbf{x}^* \sim \mathsf{Unif}(\{-1,1\}^n)$, we have

$$\begin{aligned} \|p_{M,\mathcal{D}_Y} - U\|_{tvd} &= \Pr[\mathbf{x}^* \in \mathsf{GOOD}] \cdot \|p_{M,\mathcal{D}_Y} - U\|_{tvd}|_{\mathbf{x}^* \in \mathsf{GOOD}} \\ &\quad + \Pr[\mathbf{x}^* \notin \mathsf{GOOD}] \cdot \|p_{M,\mathcal{D}_Y} - U\|_{tvd}|_{\mathbf{x}^* \notin \mathsf{GOOD}} \\ &\leq 1 \cdot \delta' + \delta' \cdot 1 = \frac{\delta}{4} < \frac{\delta}{2} \,. \end{aligned}$$

But this contradicts our assumption that

$$\|p_{M,\mathcal{D}_Y} - U\|_{tvd} = \|p_{M,\mathcal{D}_Y} - p_{M,\mathcal{D}_{unif}}\|_{tvd} \geq \frac{\delta}{2}\,.$$

This completes the proof of Theorem 6.1. □

The rest of this section is devoted to the proof of Theorem 6.2. In Section 6.1, we reduce the upper bound for Equation 6.3 to a combinatorial problem. Next, we analyze the combinatorial problem in Section 6.2, and finally complete the proof of Theorem 6.2 in Section 6.3.

6.1 Reduction to a combinatorial problem

Let $A \subseteq \{-1,1\}^n$ be the set of Alice's inputs that correspond to the message. We define $f : \{-1,1\}^n \to \{0,1\}$ to be the indicator function of A, i.e., $f(\mathbf{x}^*) = 1$ iff $\mathbf{x}^* \in A$. In this subsection, we apply Fourier analysis on the left hand side of Equation 6.3 and get an upper bound in terms of a combinatorial quantity related to the random matching and the Fourier coefficients of f. The reduction is summarized in the following lemma.

In what follows, we will write a vector $\mathbf{s} \in \{0,1\}^{\alpha kn}$ as a concatenation of αn vectors, i.e., $\mathbf{s} = (\mathbf{s}(1), \ldots, \mathbf{s}(\alpha n))$ where $\mathbf{s}(i) \in \{0,1\}^k$. We use $|\mathbf{s}(i)|$ to denote the Hamming weight of $\mathbf{s}(i)$.

Lemma 6.4. *Let $A \subseteq \{-1,1\}^n$ and $f : \{-1,1\}^n \to \{0,1\}$ be its indicator function. Let $k \in \mathbb{N}$ and $\alpha \in (0, 1/100k)$. Let $\mathcal{D}$ be a distribution over $\{-1,1\}^k$ such that $\mathbb{E}_{\mathbf{a}\sim\mathcal{D}}[a_j] = 0$ for all $j \in [k]$. For each $\ell \in [n]$, let us denote by $\mathbf{v}_\ell \in \{0,1\}^n$, the vector where the first ℓ entries are 1, and the remaining entries are 0. We have*

$$\mathop{\mathbb{E}}_M[\|p_{M,\mathcal{D}} - U\|_{tvd}^2] \leq \frac{2^{2n}}{|A|^2} \sum_{\ell\geq 2}^{\alpha kn} g(\ell) \cdot \sum_{\substack{\mathbf{v}\in\{0,1\}^n \\ |\mathbf{v}|=\ell}} \widehat{f}(\mathbf{v})^2\,,$$

where

$$g(\ell) = \Pr_M\left[\exists \mathbf{s} \in \{0,1\}^{\alpha kn}\backslash\{0^{\alpha kn}\},\ |\mathbf{s}(i)| \neq 1\,\forall i,\ M^\top \mathbf{s} = \mathbf{v}_\ell\right]\,.$$

Proof. By Cauchy–Schwarz inequality and Equation 3.8,

$$\begin{aligned}\mathop{\mathbb{E}}_M\left[\|p_{M,\mathcal{D}} - U\|_{tvd}^2\right] &\leq 2^{2\alpha kn} \mathop{\mathbb{E}}_M\left[\|p_{M,\mathcal{D}} - U\|_2^2\right] \\ &= 2^{2\alpha kn} \mathop{\mathbb{E}}_M\left[\sum_{\mathbf{s}\in\{0,1\}^{\alpha kn}\backslash\{0^{\alpha kn}\}} \widehat{p_{M,\mathcal{D}}}(\mathbf{s})^2\right]\,. \qquad (6.5)\end{aligned}$$

The following claim shows that the expected sum of the Fourier coefficients (corresponding to non-empty subsets of $[\alpha kn]$) of the posterior distribution $p_{M,\mathcal{D}}$ can be upper bounded by an expected sum of certain Fourier coefficients of the indicator function f.

Claim 6.6.

$$\mathop{\mathbb{E}}_M[\|p_{M,\mathcal{D}} - U\|_{tvd}^2] \leq \frac{2^{2n}}{|A|^2} \sum_{\mathbf{s}\in \mathsf{GOOD}\backslash\{0^{\alpha kn}\}} \mathop{\mathbb{E}}_M\left[\widehat{f}(M^\top \mathbf{s})^2\right]\,.$$

Proof. For every $\mathbf{s} \in \{0,1\}^{\alpha kn} \backslash \{0^{\alpha kn}\}$, consider $\mathbf{s} \in \{0,1\}^{\alpha kn}$ to be αn blocks $\mathbf{s}(1), \ldots, \mathbf{s}(\alpha n) \in \{0,1\}^k$ of length k. Observe that

$$\widehat{p_{M,\mathcal{D}}}(\mathbf{s}) = \frac{1}{2^{\alpha kn}} \sum_{\mathbf{z} \in \{-1,1\}^{\alpha kn}} p_{M,\mathcal{D}}(\mathbf{z}) \prod_{\substack{i \in [\alpha n], j \in [k] \\ s(i)_j = 1}} z(i)_j \, .$$

By substituting $p_{M,\mathcal{D}}(\mathbf{z}) = \mathbb{E}_{\mathbf{x}^* \in A} \mathbb{E}_{\mathbf{b} \sim \mathcal{D}^{\alpha n}} [\mathbf{1}_{\mathbf{z} = M\mathbf{x}^* \odot \mathbf{b}}]$, the equation becomes

$$\begin{aligned}
\widehat{p_{M,\mathcal{D}}}(\mathbf{s}) &= \frac{1}{2^{\alpha kn}} \cdot \mathop{\mathbb{E}}_{\mathbf{x}^* \in A} \left[\prod_{\substack{i \in [\alpha n], j \in [k] \\ s(i)_j = 1}} (M\mathbf{x}^*)_{i,j} \right] \mathop{\mathbb{E}}_{\mathbf{b} \sim \mathcal{D}^{\alpha n}} \left[\prod_{\substack{i \in [\alpha n], j \in [k] \\ s(i)_j = 1}} b(i)_j \right] \\
&= \frac{1}{2^{\alpha kn}} \cdot \mathop{\mathbb{E}}_{\mathbf{x}^* \in A} \left[\prod_{\substack{i \in [\alpha n], j \in [k] \\ s(i)_j = 1}} (M\mathbf{x}^*)_{i,j} \right] \prod_{i \in [\alpha n]} \mathop{\mathbb{E}}_{\mathbf{b}(i) \sim \mathcal{D}} \left[\prod_{\substack{j \in [k] \\ s(i)_j = 1}} b(i)_j \right] .
\end{aligned}$$

Since $\mathbb{E}_{\mathbf{a} \sim \mathcal{D}}[a_j] = 0$ for all $j \in [k]$, the right hand side expression becomes zero if there exists $i \in [\alpha n]$ such that $|\mathbf{s}(i)| = 1$. Define $\mathsf{GOOD} := \{\mathbf{s} \in \{0,1\}^{\alpha kn} \,|\, |\mathbf{s}(i)| \neq 1 \ \forall i\}$. We have

$$\widehat{p_{M,\mathcal{D}}}(\mathbf{s}) \leq \frac{1}{2^{\alpha kn}} \cdot \left| \mathop{\mathbb{E}}_{\mathbf{x}^* \in A} \left[\prod_{\substack{i \in [\alpha n], j \in [k] \\ s(i)_j = 1}} (M\mathbf{x}^*)_{i,j} \right] \right| \cdot \mathbf{1}_{\mathbf{s} \in \mathsf{GOOD}} \, .$$

Since each row and column in M has at most one non-zero entry, we can rewrite the right hand side as

$$= \frac{1}{2^{\alpha kn}} \cdot \left| \mathop{\mathbb{E}}_{\mathbf{x}^* \in A} \left[\prod_{\substack{i \in [n] \\ (M^\top \mathbf{s})_i = 1}} \mathbf{x}_i^* \right] \right| \cdot \mathbf{1}_{\mathbf{s} \in \mathsf{GOOD}}$$

Now we relate the above quantity to the Fourier coefficients of f. Recall that f is the indicator function of the set A and hence for each $\mathbf{v} \in \{0,1\}^n$, we have

$$\widehat{f}(\mathbf{v}) = \frac{1}{2^n} \sum_{\mathbf{x}^*} f(\mathbf{x}^*) \prod_{i \in [n] : v_i = 1} \mathbf{x}_i^* = \frac{1}{2^n} \sum_{\mathbf{x}^* \in A} \prod_{i \in [n] : v_i = 1} \mathbf{x}_i^* \, .$$

Thus, the Fourier coefficient of p_M corresponding to a set $\mathbf{s} \in \{0,1\}^{\alpha kn}$ can be bounded as follows:

$$\widehat{p_{M,\mathcal{D}}}(\mathbf{s}) \leq \frac{1}{2^{\alpha kn}} \cdot \frac{2^n}{|A|} \left| \widehat{f}(M^\top \mathbf{s}) \right| \cdot \mathbf{1}_{\mathbf{s} \in \mathsf{GOOD}} \, . \tag{6.7}$$

By plugging Equation 6.7 into Equation 6.5, we have the desired bound, and this completes the proof of Claim 6.6. □

It follows from Claim 6.6 that

$$\mathop{\mathbb{E}}_{M}[\|p_{M,\mathcal{D}}-U\|_{tvd}^2] \le \frac{2^{2n}}{|A|^2} \sum_{\mathbf{s}\in\mathsf{GOOD}\backslash\{0^{\alpha k n}\}} \mathop{\mathbb{E}}_{M}\left[\widehat{f}(M^\top \mathbf{s})^2\right] .$$

Since for a fixed M, the map $M^\top$ is injective, the right hand side of the above inequality has the following combinatorial form.

$$\frac{2^{2n}}{|A|^2} \sum_{\mathbf{v}\in\{0,1\}^n\backslash\{0^n\}} \Pr_{M}\left[\exists \mathbf{s}\in\mathsf{GOOD}\backslash\{0^{\alpha k n}\},\ M^\top \mathbf{s}=\mathbf{v}\right] \widehat{f}(\mathbf{v})^2 .$$

By symmetry, the above probability term will be the same for $\mathbf{v}$ and $\mathbf{v}'$ which have the same Hamming weight. For each $\ell\in[n]$, denote $g(\ell)=\Pr_M\left[\exists \mathbf{s}\in\mathsf{GOOD}\backslash\{0^{\alpha k n}\},\ M^\top \mathbf{s}=\mathbf{v}_\ell\right]$. Therefore, the expression simplifies to

$$\frac{2^{2n}}{|A|^2} \sum_{\ell\ge 1}^{n} g(\ell)\cdot \sum_{\substack{\mathbf{v}\in\{0,1\}^n\\ |\mathbf{v}|=\ell}} \widehat{f}(\mathbf{v})^2 .$$

Note that for $\ell=1$ or $\ell>\alpha k n$, $g(\ell)=0$ by definition. Thus, the above expression further simplifies to the following:

$$\frac{2^{2n}}{|A|^2} \sum_{\ell\ge 2}^{\alpha k n} g(\ell)\cdot \sum_{\substack{\mathbf{v}\in\{0,1\}^n\\ |\mathbf{v}|=\ell}} \widehat{f}(\mathbf{v})^2 .$$

We conclude that

$$\mathop{\mathbb{E}}_{M}[\|p_{M,\mathcal{D}}-U\|_{tvd}^2] \le \frac{2^{2n}}{|A|^2} \sum_{\ell\ge 2}^{\alpha k n} g(\ell)\cdot \sum_{\substack{\mathbf{v}\in\{0,1\}^n\\ |\mathbf{v}|=\ell}} \widehat{f}(\mathbf{v})^2 .$$

This completes the proof of Lemma 6.4. □

6.2 An upper bound for the combinatorial problem

In this subsection, we upper bound the combinatorial term $g(\ell)$ in Lemma 6.4. The result is summarized in the following lemma.

Lemma 6.8. *For every k, there exists an $\alpha_0>0$ such that for every $\alpha\in(0,\alpha_0)$, and for every n and $\ell\le n/2$, we have*

$$g(\ell)=\Pr_{M}\left[\exists \mathbf{s}\neq 0,\ |\mathbf{s}(i)|\neq 1\ \forall i,\ M^\top \mathbf{s}=\mathbf{v}_\ell\right] \le \left(\frac{\ell}{n}\right)^{\ell/2} .$$

Proof. We set $\alpha_0=(1/(2e^2k))^k$ so that $2\alpha_0^{1/k}e^{3/2}k\le 1$. We reformulate our events. Instead of fixing $\mathbf{v}=\mathbf{v}_\ell$ and picking the matching M at random, we note that it is equivalent to fixing the matching M and letting $\mathbf{v}$ be a uniformly random vector of weight ℓ. We thus let M be the

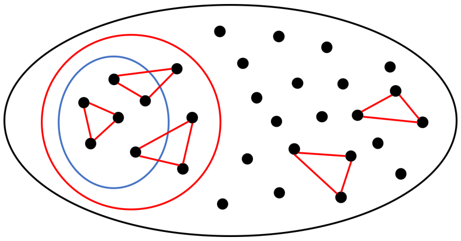

Figure 3: An example with $n = 30$, $k = 3$, $\alpha = 0.5$, and $\ell = 6$. The red triangles denote the edges $e_1, \dots, e_5$. The blue circle denotes the set V and the red circle denotes the set T with $t = 3$. Note that the figure illustrates the over-counting we do in the proof of the lemma - the set V actually intersects one of the edges just once, and so should not be counted. Our counting will nevertheless include the set since it is contained in at most $\ell/2 = 3$ edges.

matching $e_1, \dots, e_{\alpha n}$, where $e_i = \{(i-1)k+1, \dots, (i-1)k+k\}$. Letting V denote the support of the vector $\mathbf{v}$, the event we wish to consider is: "$V \subseteq [k\alpha n]$ and $|V \cap e_i| \neq 1$ for every $i \in [\alpha n]$."

We bound the probability as follows. Let $T = \{i \in [\alpha n] \mid V \cap e_i \neq \emptyset\}$ denote the set of edges that touch V, and let $|T| = t$. Note that $\ell/k \leq t \leq \ell/2$, where the latter inequality follows from the fact that every intersection is of size at least 2. We pick V by first picking T (there are at most $\binom{\alpha n}{t}$ ways of doing this), and then picking V as a subset of the vertices incident to the edges of T (there are $\binom{kt}{\ell}$ ways of doing this). (See Figure 3.) Summing over t and dividing by the total number of choices of V gives the final bound. We give the calculations below (which use the inequalities $(a/b)^b \leq \binom{a}{b} \leq (ea/b)^b$).

$$\begin{aligned}
&\Pr_V[V \subseteq [k\alpha n], |V \cap e_i| \neq 1, \forall i \in [\alpha n]] \\
&\leq \frac{\sum_{t=\ell/k}^{\ell/2} \binom{\alpha n}{t}\binom{kt}{\ell}}{\binom{n}{\ell}} \\
&\leq \sum_{t=\ell/k}^{\ell/2} \left(\frac{e\alpha n}{t}\right)^t \cdot \left(\frac{ekt}{\ell}\right)^\ell \cdot \left(\frac{n}{\ell}\right)^{-\ell} \\
&= \sum_{t=\ell/k}^{\ell/2} e^{t+\ell}\alpha^t k^\ell (t/n)^{\ell-t} \\
&\leq \alpha^{\ell/k} e^{3\ell/2} k^\ell (\ell/n)^{\ell/2} \sum_{t'=0}^{\infty} (\ell/n)^{t'} \\
&\leq 2(\alpha^{1/k} e^{3/2} k)^\ell (\ell/n)^{\ell/2} \\
&\leq (2\alpha^{1/k} e^{3/2} k)^\ell (\ell/n)^{\ell/2} \\
&\leq (\ell/n)^{\ell/2} .
\end{aligned}$$

□

6.3 Proof of Theorem 6.2

Proof of Theorem 6.2. By Lemma 6.4 and Lemma 6.8, we have

$$\mathop{\mathbb{E}}_{M}[\|p_{M,\mathcal{D}} - U\|_{tvd}^2] \le \frac{2^{2n}}{|A|^2} \cdot \sum_{\ell=2}^{\alpha kn} \frac{\ell^{\ell/2}}{n^{\ell/2}} \sum_{\substack{\mathbf{v}\in\{0,1\}^n \\ |\mathbf{v}|=\ell}} \widehat{f}(\mathbf{v})^2 \,.$$

We use Lemma 3.9 to upper bound the sum of level-ℓ Fourier coefficients for small ℓ as follows. Let $c = \tau_0\sqrt{n}$ so that $|A| \ge 2^{n-c}$. For $\ell \in [4c]$, we have

$$\frac{2^{2n}}{|A|^2} \sum_{\substack{\mathbf{v}\in\{0,1\}^n \\ |\mathbf{v}|=\ell}} \widehat{f}(\mathbf{v})^2 \le \left(\frac{4\sqrt{2}c}{\ell}\right)^{\ell} .$$

Next, we apply Parseval's inequality (Lemma 3.7) and have $\sum_{\mathbf{v}} \widehat{f}(\mathbf{v})^2 \le 1$. Thus,

$$\mathop{\mathbb{E}}_{M}[\|p_{M,\mathcal{D}} - U\|_{tvd}^2] \le \sum_{\ell=2}^{4c} \frac{\ell^{\ell/2}}{n^{\ell/2}} \cdot \left(\frac{4\sqrt{2}c}{\ell}\right)^{\ell} + \frac{2^{2n}}{|A|^2} \cdot \max_{4c<\ell\le\alpha kn} \left\{\frac{\ell^{\ell/2}}{n^{\ell/2}}\right\}$$

The second term on the right hand side is maximized at $\ell = 4c+1$, and hence

$$\begin{aligned}\mathop{\mathbb{E}}_{M}[\|p_{M,\mathcal{D}} - U\|_{tvd}^2] &\le \sum_{\ell=2}^{4c} \left(\frac{32c^2}{\ell\cdot n}\right)^{\ell/2} + \left(\frac{8c}{n}\right)^{2c} \\ &\le \sum_{\ell=2}^{4c} (16\tau_0^2)^{\ell/2} + (8\tau_0)^{2c} \\ &\le \delta^2 ,\end{aligned}$$

where the final expression determines our choice of τ_0. Specifically, we set $\tau_0 = \delta/2^6$ so that each term is at most $\delta^2/2$. This completes the proof of Theorem 6.2. □

7 Communication Lower Bound: General Case

In this section we finally prove Theorem 5.3. In other words we show that for every $\mathcal{D}_Y, \mathcal{D}_N \in \Delta(\{-1,1\}^k)$ with matching marginals, any protocol for $(\mathcal{D}_Y,\mathcal{D}_N)$-RMD with positive advantage requires $\Omega(\sqrt{n})$ bits of communication. We start with an overview.

The first step is to observe that we can prove indistinguishability of *some* distributions with matching non-zero marginals. For example, given that $\mathcal{D}_1 = \mathsf{Unif}(\{(-1,-1),(1,1)\}$ is indistinguishable from $\mathcal{D}_2 = \mathsf{Unif}(\{-1,1\}^2)$, it can also be shown that $\mathcal{D}_1' = \frac{1}{2}\{(1,1)\} + \frac{1}{2}\mathcal{D}_1$ is indistinguishable from $\mathcal{D}_2' = \frac{1}{2}\{(1,1)\} + \frac{1}{2}\mathcal{D}_2$ (see Lemma 7.7 for a related statement). Note that $\mathcal{D}_1'$ and $\mathcal{D}_2'$ are distributions with non-zero but matching marginals.

The bulk of this section is devoted to proving that for every pair of distributions $\mathcal{D}_Y$ and $\mathcal{D}_N$, we can find a path (a sequence) of intermediate distributions $\mathcal{D}_Y = \mathcal{D}_0, \mathcal{D}_1, \ldots, \mathcal{D}_L = \mathcal{D}_N$ such that adjacent pairs in this sequence are indistinguishable by a "basic" argument, where a basic argument is a combination of an indistinguishability result from Theorem 6.1 and a shifting argument formalized in Lemma 7.7. Our proof comes in the following steps:

1. For every marginal vector $\boldsymbol{\mu}$, we identify a *canonical* distribution $\mathcal{D}_{\boldsymbol{\mu}}$ that we use as the endpoint of the path. So it suffices to prove that for all $\mathcal{D}$, $\mathcal{D}$ is indistinguishable from $\mathcal{D}_{\boldsymbol{\mu}(\mathcal{D})}$, i.e., there is a path of finite length from $\mathcal{D}$ to $\mathcal{D}_{\boldsymbol{\mu}(\mathcal{D})}$.

2. We identify a measure $\Phi(\mathcal{D})$ associated with distributions that helps measure progress on a path. Among distributions with marginal $\boldsymbol{\mu}(\mathcal{D})$, this measure is uniquely maximized by $\mathcal{D}_{\boldsymbol{\mu}(\mathcal{D})}$. We show that for every distribution $\mathcal{D}$ that is not canonical one can take a basic step that increases $\boldsymbol{\mu}(\mathcal{D})$. Unfortunately the measure Φ is real-valued and the increases per step can be by arbitrarily small amounts, so we are not done.

3. We give a combinatorial proof that there is a path of finite length (some function of k) that takes us from an arbitrary distribution to the canonical one.

Putting the three ingredients together, along with a proof that a "basic step" is indistinguishable gives us the final theorem.

We start with the definition of the chain and the canonical distribution. For a distribution $\mathcal{D} \in \Delta(\{-1,1\}^k)$, its support is the set $\mathsf{supp}(\mathcal{D}) = \{\mathbf{a} \in \{-1,1\}^k \mid \mathcal{D}(\mathbf{a}) > 0\}$. Next, we consider the following partial order on $\{-1,1\}^k$. For vectors $\mathbf{a},\mathbf{b} \in \{-1,1\}^k$ we use the notation $\mathbf{a} \le \mathbf{b}$ if $a_i \le b_i$ for every $i \in [k]$. Further we use $\mathbf{a} < \mathbf{b}$ if $\mathbf{a} \le \mathbf{b}$ and $\mathbf{a} \neq \mathbf{b}$.

Definition 7.1 (Chain)**.** *We refer to a sequence* $\mathbf{a}(0) < \mathbf{a}(1) < \cdots < \mathbf{a}(\ell)$*,* $\mathbf{a}(i) \in \{-1,1\}^k$ *for every* $i \in \{0,\ldots,\ell\}$*, as a* chain *of length* ℓ*. Note that chains in* $\{-1,1\}^k$ *have length at most* k*.*

Definition 7.2 (Canonical distribution)**.** *Given a vector of marginals* $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_k) \in [-1,1]^k$*, the* canonical distribution *associated with* $\boldsymbol{\mu}$*, denoted* $\mathcal{D}_{\boldsymbol{\mu}}$*, is defined as follows: Let* $\rho : [k] \to [k]$ *be a permutation such that* $-1 \le \mu_{\rho(1)} \le \cdots \le \mu_{\rho(k)} \le 1$*. For* $i \in \{0,\ldots,k\}$*, let* $\mathbf{a}(i) \in \{-1,1\}^k$ *be given by* $\mathbf{a}(i)_j = -1$ *if* $j \in \{\rho(1),\ldots,\rho(k-i)\}$ *and* $\mathbf{a}(i)_j = 1$ *otherwise. (Note that* $\mathbf{a}(0) < \cdots < \mathbf{a}(k)$*.) Then* $\mathcal{D}_{\boldsymbol{\mu}}(\mathbf{a}(i)) = \frac{1}{2}(\mu_{\rho(k-i+1)} - \mu_{\rho(k-i)})$*, where we define* $\mu_{\rho(0)} = -1$ *and* $\mu_{\rho(k+1)} = 1$*. Finally,* $\mathcal{D}_{\boldsymbol{\mu}}(\mathbf{a}) = 0$ *for all* $\mathbf{a} \notin \{\mathbf{a}(0),\ldots,\mathbf{a}(k)\}$*.*

It is easy to verify that $\mathcal{D}_{\boldsymbol{\mu}}$ is indeed a distribution, and that it has the desired marginals, i.e., $\boldsymbol{\mu}(\mathcal{D}_{\boldsymbol{\mu}}) = \boldsymbol{\mu}$. Note that a distribution is a canonical distribution if and only if its support is a chain. Furthermore, the canonical distribution is uniquely determined even though ρ, and hence the chain $\mathbf{a}(0),\ldots,\mathbf{a}(k)$, may not be uniquely determined. This is so since ρ is non-unique only if $\mu_{\rho(i)} = \mu_{\rho(i+1)}$ for some i, and in this case $\mathcal{D}_{\boldsymbol{\mu}}(\mathbf{a}(i)) = 0$ so the "non-uniqueness of $\mathbf{a}(i)$ does not affect $\mathcal{D}_{\boldsymbol{\mu}}$.

Next we define a potential associated with distributions. For a distribution $\mathcal{D} \in \Delta(\{-1,1\}^k)$ define its potential to be

$$\Phi(\mathcal{D}) = \mathop{\mathbb{E}}_{\mathbf{b}\sim\mathcal{D}}\left[\left(\sum_{j\in[k]} b_j\right)^2\right].$$

We will show shortly that $\mathcal{D}_{\boldsymbol{\mu}}$ is the distribution with maximum potential among all distributions with marginal $\boldsymbol{\mu}$. In the process of showing this we will introduce a "polarization operator" which maps a distribution $\mathcal{D}$ to a new one that increases the potential for typical distributions. Since this operator is useful also for further steps, we start with defining this operator and analyzing its effect on the potential.

7.1 Polarization

Briefly, suppose the support of a distribution contains both $(-1)^i(1)^{k-i}$ and $1^i(-1)^{k-i}$. Then the polarization operator moves some of this mass (as much as possible while maintaining the property that the result is a distribution) to the more "polarized" points $(-1)^k$ and 1^k. The operator is defined more generally to allow the two starting points to agree on some coordinates. To define this operator, the following notation will be useful.

For $\mathbf{u}, \mathbf{v} \in \{-1,1\}^k$, let $\mathbf{u} \wedge \mathbf{v} = (\min\{u_1, v_1\}, \ldots, \min\{u_k, v_k\})$ and let $\mathbf{u} \vee \mathbf{v} = (\max\{u_1, v_1\}, \ldots, \max\{u_k, v_k\})$. We say $\mathbf{u}$ and $\mathbf{v}$ are incomparable if $\mathbf{u} \not\leq \mathbf{v}$ and $\mathbf{v} \not\leq \mathbf{u}$. Note that if $\mathbf{u}$ and $\mathbf{v}$ are incomparable then $\{\mathbf{u}, \mathbf{v}\}$ and $\{\mathbf{u} \vee \mathbf{v}, \mathbf{u} \wedge \mathbf{v}\}$ are disjoint[11].

Definition 7.3 (Polarization (update) operator)**.** *Given a distribution $\mathcal{D} \in \Delta(\{-1,1\}^k)$ and incomparable elements $\mathbf{u}, \mathbf{v} \in \{-1,1\}^k$, we define the* $(\mathbf{u}, \mathbf{v})$-polarization *of $\mathcal{D}$, denoted $\mathcal{D}_{\mathbf{u},\mathbf{v}}$, to be the distribution as given below. Let $\varepsilon = \min\{\mathcal{D}(\mathbf{u}), \mathcal{D}(\mathbf{v})\}$.*

$$\mathcal{D}_{\mathbf{u},\mathbf{v}}(\mathbf{b}) = \begin{cases} \mathcal{D}(\mathbf{b}) - \varepsilon & , \mathbf{b} \in \{\mathbf{u}, \mathbf{v}\} \\ \mathcal{D}(\mathbf{b}) + \varepsilon & , \mathbf{b} \in \{\mathbf{u} \vee \mathbf{v}, \mathbf{u} \wedge \mathbf{v}\} \\ \mathcal{D}(\mathbf{b}) & , \textit{otherwise.} \end{cases}$$

We refer to $\varepsilon(\mathcal{D}, \mathbf{u}, \mathbf{v}) = \min\{\mathcal{D}(\mathbf{u}), \mathcal{D}(\mathbf{v})\}$ as the polarization amount.

It can be verified that the polarization operator preserves the marginals, i.e., $\boldsymbol{\mu}(\mathcal{D}) = \boldsymbol{\mu}(\mathcal{D}_{\mathbf{u},\mathbf{v}})$. Note also that this operator is non-trivial, i.e., $\mathcal{D}_{\mathbf{u},\mathbf{v}} = \mathcal{D}$, if $\{\mathbf{u}, \mathbf{v}\} \not\subseteq \mathsf{supp}(\mathcal{D})$. By correlating the "+1"s and "−1"s, the polarization operator makes the support of $\mathcal{D}$ more polarized in the sense quantified in the following lemma.

Lemma 7.4 (Polarization increases potential)**.** *Let $\mathcal{D} \in \Delta(\{-1,1\}^k)$ be a distribution with marginal vector $\boldsymbol{\mu} = \boldsymbol{\mu}(\mathcal{D})$ and let $\mathbf{u}, \mathbf{v} \in \mathsf{supp}(\mathcal{D})$ be incomparable. Then we have*

$$\Phi(\mathcal{D}_{\mathbf{u},\mathbf{v}}) = \Phi(\mathcal{D}) + 8 \cdot \varepsilon \cdot s \cdot t$$

where $\varepsilon = \varepsilon(\mathcal{D}, \mathbf{u}, \mathbf{v})$ is the polarization amount, and $s = |\{j \in [k] \,|\, u_j = -v_j = 1\}|$ and $t = |\{j \in [k] \,|\, u_j = -v_j = -1\}|$. In particular $\Phi(\mathcal{D}_{\mathbf{u},\mathbf{v}}) > \Phi(\mathcal{D})$.

Proof. We look at the difference $\Phi(\mathcal{D}_{\mathbf{u},\mathbf{v}}) - \Phi(\mathcal{D})$. Let $\ell = \sum_{j \in [k] : u_j = v_j} u_j$. We have:

$$\begin{aligned} \Phi(\mathcal{D}_{\mathbf{u},\mathbf{v}}) - \Phi(\mathcal{D}) &= \sum_{\mathbf{b} \in \{-1,1\}^k} (\mathcal{D}_{\mathbf{u},\mathbf{v}}(\mathbf{b}) - \mathcal{D}(\mathbf{b})) \cdot \Phi(\mathbf{b}) \\ &= \varepsilon \cdot (\Phi(\mathbf{u} \wedge \mathbf{v}) + \Phi(\mathbf{u} \vee \mathbf{v}) - \Phi(\mathbf{u}) - \Phi(\mathbf{v})) \\ &= \varepsilon \cdot ((\ell + s + t)^2 + (\ell - s - t)^2 - (\ell + s - t)^2 - (\ell - s + t)^2) \\ &= 8 \cdot \varepsilon \cdot s \cdot t\,. \end{aligned}$$

Finally note that $s, t > 0$ since $\mathbf{u}$ and $\mathbf{v}$ are incomparable, and $\varepsilon > 0$ since $\mathbf{u}, \mathbf{v} \in \mathsf{supp}(\mathcal{D})$, thus yielding $\Phi(\mathcal{D}_{\mathbf{u},\mathbf{v}}) > \Phi(\mathcal{D})$. □

Lemma 7.5 ($\mathcal{D}_{\boldsymbol{\mu}}$ maximizes potential)**.** *For every distribution $\mathcal{D} \in \Delta(\{-1,1\}^k)$ with $\boldsymbol{\mu} = \boldsymbol{\mu}(\mathcal{D})$ we have $\Phi(\mathcal{D}) \leq \Phi(\mathcal{D}_{\boldsymbol{\mu}})$. Furthermore the inequality is strict if $\mathcal{D} \neq \mathcal{D}_{\boldsymbol{\mu}}$.*

[11] To see this, suppose $\mathbf{u} = \mathbf{u} \wedge \mathbf{v}$, then we have $u_j = \min\{u_j, v_j\}$ for all $j \in [k]$ and hence $\mathbf{u} \leq \mathbf{v}$, which is a contradiction. The same analysis works for the other cases.

Proof. Let $\mathcal{D}^*$ be a distribution with marginal $\boldsymbol{\mu}$ that maximized $\Phi(\mathcal{D})$. Suppose there exist incomparable $\mathbf{u}, \mathbf{v} \in \mathsf{supp}(\mathcal{D}^*)$, then by Lemma 7.4 we have that $\Phi(\mathcal{D}^*) < \Phi(\mathcal{D}^*_{\mathbf{u},\mathbf{v}})$ contradicting the maximality of $\mathcal{D}^*$. It follows that there are no incomparable elements in $\mathsf{supp}(\mathcal{D}^*)$, or in other words, $\mathsf{supp}(\mathcal{D}^*)$ is a chain. We now show that this implies $\mathcal{D}^* = \mathcal{D}_{\boldsymbol{\mu}}$.

More specifically we show that any distribution $\mathcal{D}$ supported on a chain is uniquely determined by its marginal $\boldsymbol{\mu}$. To see this, let $\rho : [k] \to [k]$ be a bijection such that $\mu_{\rho(j)} \leq \mu_{\rho(j+1)}$ for all j. Let $\tau_0 < \tau_1 < \cdots < \tau_\ell$ be the attainable values of $\boldsymbol{\mu}$, i.e., $\{\tau \mid \exists j \in [k] \, s.t. \, \mu_j = \tau\} = \{\tau_0, \ldots, \tau_\ell\}$. For $0 \leq i \leq \ell$, let $\mathbf{a}(i)$ be given by $\mathbf{a}(i)_j = -1$ if $\mu_j \leq \tau_{\ell-i}$ and $\mathbf{a}(i)_j = 1$ otherwise. Note that $\mathbf{a}(0) < \cdots < \mathbf{a}(\ell)$. It can be verified that $\mathsf{supp}(\mathcal{D}^*) = \{\mathbf{a}(0), \ldots, \mathbf{a}(\ell)\}$, and $\mathcal{D}^*(\mathbf{a}(i))$ is uniquely defined for all i.

Claim 7.6. *$\mathsf{supp}(\mathcal{D}^*) = \{\mathbf{a}(0), \ldots, \mathbf{a}(\ell)\}$, and $\mathcal{D}^*(\mathbf{a}(i)) = (\tau_{\ell-i+1} - \tau_{\ell-i})/2$, where $\tau_{-1} = -1$ and $\tau_{\ell+1} = 1$.*

Proof. For the sake of contradiction, assume $\mathsf{supp}(\mathcal{D}^*) = \{\mathbf{a}'(0), \ldots, \mathbf{a}'(\ell')\} \neq \{\mathbf{a}(0), \ldots, \mathbf{a}(\ell)\}$ where $\mathbf{a}'(0) < \mathbf{a}'(1) < \cdots < \mathbf{a}'(\ell')$ is a chain. Let $0 \leq i \leq \min\{\ell, \ell'\}$ be the smallest i such that $\mathbf{a}(i) \neq \mathbf{a}'(i)$. Consider the following three situations: (i) $\mathbf{a}(i) < \mathbf{a}'(i)$, (ii) $\mathbf{a}(i) > \mathbf{a}'(i)$, and (iii) $\mathbf{a}(i)$ and $\mathbf{a}'(i)$ are incomparable.

For (i) and (iii), due to the construction of $\{\mathbf{a}(0), \ldots, \mathbf{a}(\ell)\}$ and the fact that $\{\mathbf{a}'(0), \ldots, \mathbf{a}'(\ell')\}$ is a chain, we have that for each $j, j' \in [k]$ with $\tau_{i-2} < \mu_j, \mu_{j'} \leq \tau_i$, $\mathbf{a}'(i')_j = \mathbf{a}'(i')_{j'}$ for all $0 \leq i' \leq \ell'$. This implies that $\mu_j = \mu_{j'}$ which is a contradiction because there are two attainable values τ_i and τ_{i-1} lie in the interval $(\tau_{i-2}, \tau_i]$. Similar argument also works for situation (ii).

We conclude that $\mathsf{supp}(\mathcal{D}^*) = \{\mathbf{a}(0), \ldots, \mathbf{a}(\ell)\}$. It is immediate to see that $\mathcal{D}^*(\mathbf{a}(i))$ is uniquely defined for all i by solving the following linear system.

$$\boldsymbol{\mu} = \begin{bmatrix} | & | & & | \\ \mathbf{a}(0) & \mathbf{a}(1) & \cdots & \mathbf{a}(\ell) \\ | & | & & | \end{bmatrix} \begin{bmatrix} \mathcal{D}^*(\mathbf{a}(0)) \\ \mathcal{D}^*(\mathbf{a}(1)) \\ \cdots \\ \mathcal{D}^*(\mathbf{a}(\ell)) \end{bmatrix} .$$

Note that by the construction of $\{\mathbf{a}(0), \ldots, \mathbf{a}(\ell)\}$, the matrix has full rank, and, hence, there is a unique solution. It can be verified that the solution is given by $\mathcal{D}^*(\mathbf{a}(i)) = (\tau_{\ell-i+1} - \tau_{\ell-i})/2$, where $\tau_{-1} = -1$ and $\tau_{\ell+1} = 1$. □

In summary, $\mathcal{D}^*$ is uniquely determined by $\boldsymbol{\mu}(\mathcal{D})$ and its support is a chain. This implies that $\mathcal{D}^* = \mathcal{D}_{\boldsymbol{\mu}}$, so $\mathcal{D}_{\boldsymbol{\mu}}$ is the unique distribution that maximizes the potential. □

7.2 Indistinguishability of a polarization update

Our next observation is that for every distribution $\mathcal{D}$ with incomparable elements $\mathbf{u}$, $\mathbf{v}$ in their support, $\mathcal{D}$ is indistinguishable, in the RMD problem, from its $(\mathbf{u}, \mathbf{v})$-polarization $\mathcal{D}_{\mathbf{u},\mathbf{v}}$.

Lemma 7.7 (Polarization update preserves indistinguishability)**.** *Let $\alpha_0(k)$ be as given in Theorem 6.1. Let $k \in \mathbb{N}$, $\alpha \in (0, \alpha_0)$, $\delta \in (0, 1/2)$. Then for every distribution $\mathcal{D} \in \Delta(\{-1,1\}^k)$ and incomparable $\mathbf{u}, \mathbf{v} \in \mathsf{supp}(\mathcal{D})$ there exists $\tau > 0$ and n_0 such that for every $n \geq n_0$ every protocol for $(\mathcal{D}, \mathcal{D}_{\mathbf{u},\mathbf{v}})$-RMD achieving advantage δ on instances of length n requires $\tau\sqrt{n}$ bits of communication.*

We prove Lemma 7.7 by a reduction. We show that there exists a pair of distributions $\mathcal{D}_Y$ and $\mathcal{D}_N$ with marginals being zero such that given a protocol Π for $(\mathcal{D}, \mathcal{D}_{\mathbf{u},\mathbf{v}})$-RMD, we can get a protocol Π' for $(\mathcal{D}_Y, \mathcal{D}_N)$-RMD. We then use Theorem 6.1 to get a lower bound on the communication of Π' and thus of Π. Specifically, we divide the proof into three steps. In step one, we define $\mathcal{D}_Y$ and $\mathcal{D}_N$ and provide intuition on the reduction. Next, we formally describe the reduction by designing a protocol for $(\mathcal{D}_Y, \mathcal{D}_N)$-RMD from a protocol for $(\mathcal{D}, \mathcal{D}_{\mathbf{u},\mathbf{v}})$-RMD. Finally, we prove the correctness of the reduction and wrap up the proof of Lemma 7.7.

Step 1: The auxiliary distributions $\mathcal{D}_Y$ and $\mathcal{D}_N$. We start by defining $\mathcal{D}_Y$ and $\mathcal{D}_N$. Let $S = \{i \in [k] \,|\, u_i \neq v_i\}$. Let $k' = |S|$. Without loss of generality, we re-index the coordinates and assume $S = \{1, 2, \ldots, k'\}$. Let $\mathbf{a} = \mathbf{u}|_S$ so that $\mathbf{v}|_S = -\mathbf{a}$. We also let $\tilde{\mathbf{u}} = \mathbf{u}|_{\bar{S}}$ denote the common parts of $\mathbf{u}$ and $\mathbf{v}$. Let $\mathcal{D}_Y$ be the uniform distribution over $\{\mathbf{a}, -\mathbf{a}\}$, and $\mathcal{D}_N$ be the uniform distribution over $\{1^{k'}, (-1)^{k'}\}$. Note that $\boldsymbol{\mu}(\mathcal{D}_Y) = \boldsymbol{\mu}(\mathcal{D}_N) = 0^{k'}$. Let $\mathcal{D}_1 = \mathsf{Unif}(\{\mathbf{u}, \mathbf{v}\})$ and $\mathcal{D}_2 = \mathsf{Unif}(\{\mathbf{u} \vee \mathbf{v}, \mathbf{u} \wedge \mathbf{v}\})$. Let $\varepsilon = \varepsilon(\mathcal{D}, \mathbf{u}, \mathbf{v})$ be the polarization amount. Let $\mathcal{D}_0 \in \Delta(\{-1, 1\}^k)$ be such that $\mathcal{D} = (1 - 2\varepsilon)\mathcal{D}_0 + 2\varepsilon\mathcal{D}_1$. Note that $\mathcal{D}_{\mathbf{u},\mathbf{v}} = (1 - 2\varepsilon)\mathcal{D}_0 + 2\varepsilon\mathcal{D}_2$.

We give an informal idea now, before giving the (potentially notationally complex) details. The rough idea is that Alice and Bob first pad their inputs with lots of dummy variables (whose values are known to both) and expand the masks from $\mathcal{D}_Y$ (or $\mathcal{D}_N$) into masks that are from $\mathcal{D}_1$ (respectively $\mathcal{D}_2$). They then augment the sequence of masks from $\alpha n'$ to $\alpha n = \Omega(\alpha n'/\varepsilon)$, injecting many random masks from $\mathcal{D}_0$. This gives them an instance of $(\mathcal{D}, \mathcal{D}_{\mathbf{u},\mathbf{v}})$-RMD to solve for which they use the protocol Π. It is not too hard to see all this can be done locally by Alice and Bob; and this is proved formally below.

Step 2: A reduction from $(\mathcal{D}_Y, \mathcal{D}_N)$-RMD to $(\mathcal{D}, \mathcal{D}_{\mathbf{u},\mathbf{v}})$-RMD. Consider a protocol $\Pi = (\Pi_A, \Pi_B)$ for $(\mathcal{D}, \mathcal{D}_{\mathbf{u},\mathbf{v}})$-RMD with parameter $\alpha \leq 1/(200k)$ using $C(n)$ bits of communication to achieve an advantage of δ on instances of length n. We let $n' = (k'\varepsilon/k)n$ where k' was chosen in the previous step. We also let $\alpha' = (2k/k')\alpha$ so that $\alpha' \leq 1/(100k')$. We use Π to design a protocol Π' for $(\mathcal{D}_Y, \mathcal{D}_N)$-RMD with parameter α' achieving advantage of at least $\delta/2$ on instances of length n' with communication $C'(n') = C(n)$. We conclude by Theorem 6.1 that there exists a constant τ' such that $C(n) \geq \tau'\sqrt{n'} = \tau\sqrt{n}$, where $\tau = \tau'\sqrt{\varepsilon k'/k} > 0$ as desired.

Our protocol Π' uses shared randomness between Alice and Bob (while we assume Π is deterministic). Let $n'' = kn'/k'$ so that $n = n''/(2\varepsilon)$. Let $\alpha'' = \alpha' n'/n'' = k\alpha/k'$. Recall that an instance of $(\mathcal{D}_Y, \mathcal{D}_N)$-RMD is determined by a four tuple $(\mathbf{x}', M', \mathbf{z}', \mathbf{b}')$ with $\mathbf{x}' \in \{-1, 1\}^{n'}$, $M' \in \{0, 1\}^{k'\alpha' n' \times n'}$ and $\mathbf{z}', \mathbf{b}' \in \{-1, 1\}^{k'\alpha' n'}$ with $\mathbf{z}' = M'\mathbf{x}' \odot \mathbf{b}'$. See Figure 4 for a pictorial description.

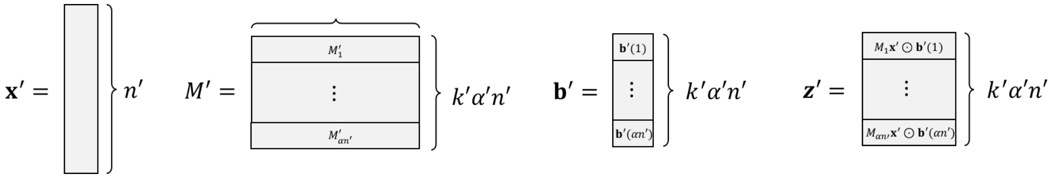

Figure 4: Pictorial description of $(\mathbf{x}', M', \mathbf{b}', \mathbf{z}')$.

We give two maps using shared randomness R' and R'':

(i) **From** $(\mathcal{D}_Y, \mathcal{D}_N)$**-RMD to** $(\mathcal{D}_1, \mathcal{D}_2)$**-RMD**: $(\mathbf{x}', M', \mathbf{b}', \mathbf{z}', R') \mapsto (\mathbf{x}'', M'', \mathbf{b}'', \mathbf{z}'')$ where $\mathbf{x}'' \in \{0,1\}^{n''}$, $M'' \in \{0,1\}^{k\alpha'' n'' \times n''}$ and $\mathbf{b}'', \mathbf{z}'' \in \{-1,1\}^{k\alpha'' n''}$.

(ii) **From** $(\mathcal{D}_1, \mathcal{D}_2)$**-RMD to** $(\mathcal{D}, \mathcal{D}_{\mathbf{u},\mathbf{v}})$**-RMD**: $(\mathbf{x}'', M'', \mathbf{b}'', \mathbf{z}'', R'') \mapsto (\mathbf{x}, M, \mathbf{b}, \mathbf{z})$, where $\mathbf{x} \in \{0,1\}^n$, $M \in \{0,1\}^{k\alpha n \times n}$ and $\mathbf{b}, \mathbf{z} \in \{-1,1\}^{k\alpha n}$.

Before describing the two maps, let us first state the desired conditions.

Success conditions for the reduction

(1) **The reduction is locally well-defined.** Namely, there exist random strings R' and R'' so that (i) Alice can get $\mathbf{x}$ through the maps $(\mathbf{x}', R') \mapsto \mathbf{x}''$ and $(\mathbf{x}'', R'') \mapsto \mathbf{x}$ while Bob can get $(M, \mathbf{z})$ through the maps $(M', \mathbf{z}', R') \mapsto (M'', \mathbf{z}'')$ and $(M'', \mathbf{z}'', R'') \mapsto (M, \mathbf{z})$.

(2) **The reduction is sound and complete.** Namely, (i) $\mathbf{z}'' = M''\mathbf{x}'' \odot \mathbf{b}''$ and $\mathbf{z} = M\mathbf{x} \odot \mathbf{b}$. (ii) If $\mathbf{b}' \sim \mathcal{D}_Y^{\alpha' n'}$ then $\mathbf{b}'' \sim \mathcal{D}_1^{\alpha'' n''}$ and $\mathbf{b} \sim \mathcal{D}^{\alpha n}$. Similarly if $\mathbf{b}' \sim \mathcal{D}_N^{\alpha' n'}$ then $\mathbf{b}'' \sim \mathcal{D}_2^{\alpha'' n''}$ and $\mathbf{b} \sim \mathcal{D}_{\mathbf{u},\mathbf{v}}^{\alpha n}$. (iii) $\mathbf{x}'' \sim \mathsf{Unif}(\{-1,1\}^{n''})$, $\mathbf{x} \sim \mathsf{Unif}(\{-1,1\}^n)$ and M is a uniformly random matrix conditioned on having exactly one "1" per row and at most one "1" per column.

In Claim 7.8 and Claim 7.9 we show that the above conditions hold except for an error event that occurs with tiny ($\exp(-n)$) probability. For now, let us show that these conditions imply the success of the reduction. Assuming conditions (1) and (2) the rest is simple. Alice computes $\mathbf{x}$ from $\mathbf{x}', R'$ and R'' and sends $m = \Pi_A(\mathbf{x})$ to Bob, who computes $(M, \mathbf{z})$ from $M', \mathbf{z}', R'$ and R'' and outputs $\Pi_B(m, M, \mathbf{z})$. Conditions (1)-(2) combined with the bound on the error event imply that if Π has advantage δ then Π' has advantage at least $\delta - \exp(-n) \geq \delta/2$ as desired.

In the rest of this subsection, we describe the two maps and show that they satisfy the described success conditions. We wrap up the reduction and the proof of Lemma 7.7 in the end.

Step 3: Specify and analyze the first map. We now turn to specifying the maps mentioned above and proving that they satisfy conditions (1)-(2). We start with $(\mathbf{x}', M', \mathbf{b}', \mathbf{z}', R') \mapsto (\mathbf{x}'', M'', \mathbf{b}'', \mathbf{z}'')$. For this part, we let $R' \sim \mathsf{Unif}(\{-1,1\}^{n''-n'})$. We set $\mathbf{x}'' = (\mathbf{x}', R')$. To get M'', $\mathbf{z}''$ and $\mathbf{b}''$ we need some more notations. First, note that $\alpha' n' = \alpha'' n''$ due to the choice of parameters. Next, note that M'' can be viewed as the stacking of matrices $M'_1, \ldots, M'_{\alpha' n'} \in \{0,1\}^{k' \times n'}$. We first extend M'_i by adding all-zero columns at the end to get $N''_i \in \{0,1\}^{k' \times n''}$. We then stack N''_i on top of $P''_i \in \{0,1\}^{(k-k') \times n''}$ to get M''_i, where $(P''_i)_{j\ell} = 1$ if and only if $\ell = n' + (i-1)k + j$. See Figure 6 for a pictorial description of N''_i and P''_i. We let M'' be the stacking of $M''_1, \ldots, M''_{\alpha'' n''}$. Next we turn to $\mathbf{b}''$. Let $\mathbf{b}' = (\mathbf{b}'(1), \cdots, \mathbf{b}'(\alpha'' n''))$. Let $\tilde{\mathbf{u}} = (u_{k'+1}, \ldots, u_k)$ denote the common parts of $\mathbf{u}$ and $\mathbf{v}$. We let $\mathbf{b}''(i) = (\mathbf{b}'(i), \tilde{\mathbf{u}})$ and $\mathbf{b}'' = (\mathbf{b}''(1), \cdots, \mathbf{b}''(\alpha'' n''))$. Finally we let $\mathbf{z}'' = M''\mathbf{x}'' \odot \mathbf{b}''$ as required. See Figure 5 for a pictorial description.

Now, we verify that the first map satisfies the success conditions mentioned above.

Claim 7.8. *The first map in the reduction is locally well-defined, sound, and complete.*

Proof. To see that the first map is locally well-defined, note that Alice can compute $\mathbf{x}'' = (\mathbf{x}', R')$ locally. Similarly, Bob can compute M'' locally by construction. As for $\mathbf{z}''$, note that $\mathbf{z}''$ interleaves

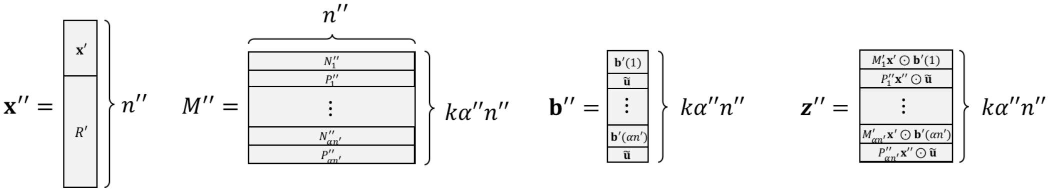

Figure 5: Pictorial description of $(\mathbf{x}'', M'', \mathbf{b}'', \mathbf{z}'')$.

(in a predetermined order) the bits of $\mathbf{z}'$ and those of $(P_i\mathbf{x}'' \odot \tilde{\mathbf{u}})_{i\in[\alpha n']}$. Furthermore $P_i\mathbf{x}''$ depends only on R' (since the first n' columns of all P_is are zero). Thus Bob can locally compute $P_i\mathbf{x}''$ for every i, and since $\tilde{\mathbf{u}}$ is also known Bob can compute $\mathbf{z}''$ locally.

To see the first map is sound and complete, (i) $\mathbf{z}'' = M''\mathbf{x}''\odot\mathbf{b}''$ follows from the construction. As for (ii), for each $i \in [\alpha' n'] = [\alpha'' n'']$, if $\mathbf{b}'_i \sim \mathcal{D}_Y = \mathsf{Unif}(\{\mathbf{a}, -\mathbf{a}\})$, then $\mathbf{b}''_i \sim \mathsf{Unif}(\{(\mathbf{a}, \tilde{\mathbf{u}}), (-\mathbf{a}, \tilde{\mathbf{u}})\})$. Note that $\mathbf{a}$ is chosen to be the uncommon part of $\mathbf{u}$ and $\mathbf{v}$ and hence $(\mathbf{a}, \tilde{\mathbf{u}}) = \mathbf{u}$ and $(-\mathbf{a}, \tilde{\mathbf{u}}) = \mathbf{v}$. Thus, $\mathbf{b}''_i \sim \mathsf{Unif}(\{\mathbf{u}, \mathbf{v}\}) = \mathcal{D}_1$ as desired. Similarly, one can show that if $\mathbf{b}'_i \sim \mathcal{D}_N$, then $\mathbf{b}''_i \sim \mathcal{D}_2$. Finally, we have $\mathbf{x}'' \sim \mathsf{Unif}(\{-1, 1\}^{n''})$ by construction and hence (iii) holds.

This completes the proof of conditions (1)-(2) for the first step of the reduction. □

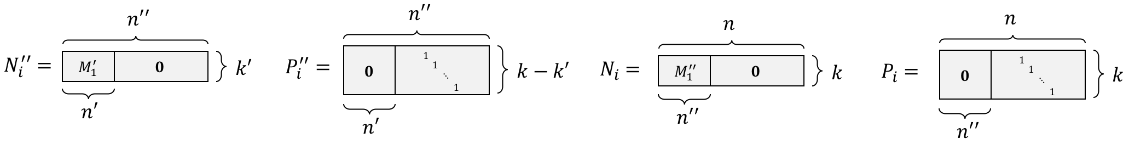

Figure 6: Pictorial description of N''_i, P''_i, N_i, P_i.

Step 4: Specify and analyze the second map. We now turn to the second map. Here R'' will be composed of many smaller parts which we introduce now. Let $\mathbf{y} \sim \mathsf{Unif}(\{-1, 1\}^{n-n''})$, $\mathbf{w} \sim \mathsf{Bern}(2\varepsilon)^{\alpha n}$. Let $\Gamma \in \{0,1\}^{n\times n}$ be a uniform permutation matrix. Let $\mathbf{c} = (\mathbf{c}(1), \ldots, \mathbf{c}((n - n'')/k))$ where $\mathbf{c}(i) \sim \mathcal{D}_0$ are chosen independently. We let $R'' = (\mathbf{y}, \mathbf{w}, \Gamma, \mathbf{c})$. Let $\#_w(i) = |\{j \in [i] \mid w_j = 1\}|$ denote the number 1's among the first i coordinates of $\mathbf{w}$. If $\#_w(\alpha n) \geq \alpha'' n''$ or if $\alpha n - \#_w(\alpha n) \geq (n - n'')/k$ we declare an error, Note $\mathbb{E}[\#_w(n)] = \alpha'' n''/2$ so the probability of error is negligible (specifically it is $\exp(-n)$).

We now define the elements of $(\mathbf{x}, M, \mathbf{b}, \mathbf{z})$. We set $\mathbf{x} = \Gamma(\mathbf{x}'', \mathbf{y})$ so $\mathbf{x}$ is a random permutation of the concatenation of $\mathbf{x}''$ and $\mathbf{y}$. Next, let $M'' = (M''_1, \ldots, M''_{\alpha'' n''})$ where $M''_i \in \{0,1\}^{k\times n''}$. We extend M''_i to $N_i \in \{0,1\}^{k\times n}$ by adding all-zero columns to the right. For $i \in \{1, \ldots, (n - n'')/k\}$, let $P_i \in \{0,1\}^{k\times n}$ be given by $(P_i)_{j\ell} = 1$ if and only if $\ell = n'' + (i-1)k + j$. See Figure 6 for a pictorial description of N_i and P_i. Next we define a matrix $\tilde{M} \in \{0,1\}^{k\alpha n\times n} = (\tilde{M}_1, \ldots, \tilde{M}_{\alpha n})$ where $\tilde{M}_i \in \{0,1\}^{k\times n}$ is defined as follows: If $w_i = 1$ then we let $\tilde{M}_i = N_{\#_w(i)}$ else we let $\tilde{M}_i = P_{i-\#_w(i)}$. Finally we let $M = \tilde{M} \cdot \Gamma^{-1}$. Next we turn to $\mathbf{b}$. Again let $\mathbf{b}'' = (\mathbf{b}''(1), \ldots, \mathbf{b}''(\alpha'' n''))$. We let $\mathbf{b} = (\mathbf{b}(1), \ldots, \mathbf{b}(\alpha n))$ where $\mathbf{b}(i)$ is defined as follows: If $w_i = 1$ then $\mathbf{b}(i) = \mathbf{b}''(\#_w(i))$, else $\mathbf{b}(i) = \mathbf{c}(i - \#_w(i))$. Finally, $\mathbf{z} = M\mathbf{x} \odot \mathbf{b}$. See Figure 7 for a pictorial description. This concludes the description of the map and we turn to analyzing its properties.

Now, we verify that the first map satisfies the success conditions mentioned above.

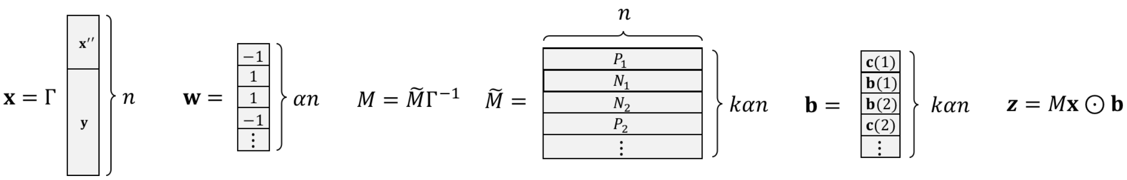

Figure 7: Pictorial description of $\mathbf{x}, \mathbf{w}, M, \mathbf{b}, \mathbf{z}$.

Claim 7.9. *If $\#_w(\alpha n) \leq \alpha'' n''$ and $\alpha n - \#_w(\alpha n) \leq (n - n'')/k$, then the second map in the reduction is locally well-defined, sound, and complete. In particular, the error event happens with probability at most $\exp(-\Omega(n))$ over the randomness of R''.*

Proof. To see that the second map is locally well-defined, first note that Alice can compute $\mathbf{x} = \Gamma(\mathbf{x}'', \mathbf{y})$ from $\mathbf{x}''$ and the shared randomness R'' locally. As for Bob, note that the maximum index needed for N and $\mathbf{b}''$ (resp. P and $\mathbf{c}$) is at most $\#_w(\alpha n)$ (resp. $\alpha n - \#_w(i)$). Namely, if $\#_w(\alpha n) \leq \alpha'' n''$ and $\alpha n - \#_w(\alpha n) \leq (n - n'')/k$, then M and $\mathbf{b}$ are well-defined. Also, using similar argument as in the proof of Claim 7.8, one can verify that M and $\mathbf{b}$ can be locally computed by M'', $\mathbf{b}''$, and the shared randomness R''.

To see the second map is sound and complete, (i) $\mathbf{z} = M\mathbf{x} \odot \mathbf{b}$ directly follows from the construction. As for (ii), if $\mathbf{b}' \sim \mathcal{D}_Y^{\alpha' n'}$, from Claim 7.8 we know that $\mathbf{b}'' \sim \mathcal{D}_1^{\alpha' n'} = \mathsf{Unif}(\{\mathbf{u}, \mathbf{v}\})^{\alpha' n'}$. Now, for each $i \in [\alpha n]$, $\mathbf{b}(i) = \mathbf{b}''(\#_w(i))$ with probability 2ε and $\mathbf{b}(i) = \mathbf{c}(i - \#_w(i))$ with probability $1 - 2\varepsilon$. As $\mathbf{b}''(i') \sim \mathcal{D}_1$ for every $i' \in [\alpha' n']$ and $\mathbf{c}(i'') \sim \mathcal{D}_0$ for every $i'' \in [(n - n'')/k]$, we have $\mathbf{b}(i) \sim (1 - 2\varepsilon)\mathcal{D}_0 + 2\varepsilon\mathcal{D}_1 = \mathcal{D}$ as desired. Similarly, one can show that for every $i \in [\alpha' n'] = [\alpha'' n'']$, if $\mathbf{b}'(i) \sim \mathcal{D}_N^{\alpha' n'}$, then $\mathbf{b}(i) \sim \mathcal{D}_{\mathbf{u},\mathbf{v}}$. Finally, we have $\mathbf{x} \sim \mathsf{Unif}(\{-1, 1\}^n)$ and M is a uniformly random matrix with exactly one "1" per row and at most one "1" per column (due to the application of a random permutation Γ) by construction.

This completes the proof of conditions (1)-(2) for the second step of the reduction. □

Step 5: Proof of Lemma 7.7.

Proof of Lemma 7.7. Let us start with setting up the parameters. Given $k, \alpha \in (0, \alpha_0), n, \mathcal{D}$, and incomparable pair $(\mathbf{u}, \mathbf{v}) \in \mathsf{supp}(\mathcal{D})$ and polarization amount $\varepsilon = \varepsilon(\mathcal{D}, \mathbf{u}, \mathbf{v})$, let $k' = |\{i \in [k] \,|\, u_i \neq v_i\}|$, $n' = (k'\varepsilon/k)n$, $\alpha' = (2k/k')\alpha$, $n'' = kn'/k'$, $\alpha'' = \alpha' n'/n''$, and $\delta' = \delta/2$.

Now, for the sake of contradiction, we assume that there exists a protocol $\Pi = (\Pi_A, \Pi_B)$ for $(\mathcal{D}, \mathcal{D}_{\mathbf{u},\mathbf{v}})$-RMD with advantage δ and at most $\tau\sqrt{n}$ bits of communication.

First, observe that $n - n'' = (1 - \varepsilon)n$ and $\alpha'' n'' = 2\varepsilon\alpha n$. As $\mathbf{w} \sim \mathsf{Bern}(\varepsilon)^{\alpha n}$, we have $\#_w(\alpha n) \leq \alpha'' n''$ and $\alpha n - \#_w(\alpha n) \leq (n - n'')/k$ with probability at least $1 - \exp(-\Omega(n))$. Thus, combine with Claim 7.8 and Claim 7.9, if $(\mathbf{x}', M', \mathbf{z}')$ is a Yes (resp. No) instance of $(\mathcal{D}_Y, \mathcal{D}_N)$-RMD, then the output of the reduction, i.e., $(\mathbf{x}, M, \mathbf{z})$, is a Yes (resp. No) instance of $(\mathcal{D}, \mathcal{D}_{\mathbf{u},\mathbf{v}})$-RMD with probability at least $1 - \exp(-\Omega(n))$. Moreover, Claim 7.8 and Claim 7.9 also show that the reduction can be implemented locally and hence Alice and Bob can run the protocol Π on $(\mathbf{x}, M, \mathbf{z})$. In particular, Alice and Bob computes $\mathbf{x}$ and $(M, \mathbf{z})$ using their inputs and shared randomness respectively. Then, Alice sends $m = \Pi_A(\mathbf{x})$ to Bob and Bob outputs $\Pi_B(m, M, \mathbf{z})$. By the correctness of the reduction as well as that of the protocol, we know that Alice and Bob have advantage at least $\delta - \exp(-\Omega(n)) \geq \delta/2 = \delta'$ in solving $(\mathcal{D}_Y, \mathcal{D}_N)$-RMD with at most $\tau\sqrt{n} = \tau\sqrt{(k/(k'\varepsilon))n'}$ bits of communication.

Finally, by Theorem 6.1, we know that there exists a constant $\tau_0 > 0$ such that any protocol for $(\mathcal{D}_Y, \mathcal{D}_N)$-RMD with advantage δ' requires at least $\tau_0\sqrt{n'}$ bits of communication. This implies that $\tau \geq \tau_0\sqrt{k'\varepsilon/k}$. We conclude that any protocol for $(\mathcal{D}, \mathcal{D}_{\mathbf{u},\mathbf{v}})$-RMD with advantage δ requires at least $\tau\sqrt{n}$ bits of communication. □

7.3 Finite upper bound on the number of polarization steps

In this section we prove that there is a finite upper bound on the number of polarization steps needed to move from a distribution $\mathcal{D} \in \Delta(\{-1,1\}^k)$ to the canonical distribution with marginal $\boldsymbol{\mu}(\mathcal{D})$, i.e., $\mathcal{D}_{\boldsymbol{\mu}(\mathcal{D})}$. Together with the indistinguishability result from Lemma 7.7 this allows us to complete the proof of Theorem 5.3 by going from $\mathcal{D}_Y$ to $\mathcal{D}_{\boldsymbol{\mu}(\mathcal{D}_Y)} = \mathcal{D}_{\boldsymbol{\mu}(\mathcal{D}_N)}$ and then to $\mathcal{D}_N$ by using the triangle inequality for indistinguishability.

In this section we extend our considerations to functions $A : \{-1,1\}^k \to \mathbb{R}^{\geq 0}$. Let $\mathcal{F}(\{-1,1\}^k) = \{A : \{-1,1\}^k \to \mathbb{R}^{\geq 0}\}$. For $A \in \mathcal{F}(\{-1,1\}^k)$, let $\mu_0(A) = \sum_{\mathbf{a}\in\{-1,1\}^k} A(\mathbf{a})$. Note $\Delta(\{-1,1\}^k) \subseteq \mathcal{F}(\{-1,1\}^k)$ and $A \in \Delta(\{-1,1\}^k)$ if and only if $A \in \mathcal{F}(\{-1,1\}^k)$ and $\mu_0(A) = \sum_{\mathbf{a}\in\{-1,1\}^k} A(\mathbf{a}) = 1$. We extend the definition of marginals, support, canonical distribution, potential and polarization operators to $\mathcal{F}(\{-1,1\}^k)$. In particular we let $\boldsymbol{\mu}(A) = (\mu_0, \mu_1, \ldots, \mu_k)$ where $\mu_0 = \mu_0(A)$ and $\mu_j = \sum_{\mathbf{a}\in\{-1,1\}^k} a_j A(\mathbf{a})$ for $j \in [k]$. We also define canonical function and polarization operators so as to preserve $\boldsymbol{\mu}(A)$. So given arbitrary A, let $\mathcal{D} = \frac{1}{\mu_0(A)} \cdot A$. Note $\mathcal{D} \in \Delta(\{-1,1\}^k)$. For $\boldsymbol{\mu} = (\mu_0, \mu_1, \ldots, \mu_k) \in \mathbb{R}^{k+1}$, we define $A_{\boldsymbol{\mu}} = \mu_0 \cdot \mathcal{D}_{\boldsymbol{\mu}'}$ where $\boldsymbol{\mu}' = (\mu_1/\mu_0, \ldots, \mu_k/\mu_0)$ to be the canonical function associated with $\boldsymbol{\mu}$. We remark that by Lemma 7.4 and Lemma 7.5, $A_{\boldsymbol{\mu}(A)}$ is the unique function such that (i) it has the same marginals as A and (ii) it supports a chain.

Definition 7.10 (Polarization length)**.** *For distribution $A \in \mathcal{F}(\{-1,1\}^k)$, let $N(A)$ be the smallest t such that there exists a sequence $\mathbf{A} = A_0, A_1, \ldots, A_t$ such that $A_0 = A$, $A_t = A_{\boldsymbol{\mu}(A)}$ is canonical and for every $i \in [t]$ it holds that there exists incomparable $\mathbf{u}_i, \mathbf{v}_i \in \mathsf{supp}(A_{i-1})$ such that $A_i = (A_{i-1})_{\mathbf{u}_i,\mathbf{v}_i}$. If no such finite sequence exists then let $N(A)$ be infinite. Let $N(k) = \sup_{A\in\mathcal{F}(\{-1,1\}^k)}\{N(A)\}$. Again, if $N(A) = \infty$ for some A or if no finite upper bound exists, $N(k)$ is defined to be ∞.*

Note that if $\mathcal{D} \in \Delta(\{-1,1\}^k)$ so is every element in the sequence, so the polarization length bound below applies also to distributions. Our main lemma in this subsection is the following:

Lemma 7.11 (A finite upper bound on $N(k)$)**.** *$N(k)$ is finite for every finite k. Specifically $N(k) \leq (k^2 + 3)(1 + N(k-1))$.*

We prove Lemma 7.11 constructively in the following four steps.

Step 1: Description of the algorithm Polarize**.** Let us start with some notations. For $A \in \mathcal{F}(\{-1,1\}^k)$ we let $A|_{x_\ell=b}$ denote the function A restricted to the subcube $\{-1,1\}^{\ell-1} \times \{b\} \times \{-1,1\}^{k-\ell}$. Note that A restricted to subcubes is effectively a $(k-1)$-dimensional function and we will use this reduction in dimension in our recursive algorithm.

Algorithm 2 Polarize$(\cdot)$

```
Input: $A \in \mathcal{F}(\{-1,1\}^k)$.
if k=2 then
    Output: $A_{(-1,1),(1,-1)}$.
$(A_0)|_{x_k=-1} \leftarrow \textsc{Polarize}(A|_{x_k=-1})$ ; $(A_0)|_{x_k=1} \leftarrow \textsc{Polarize}(A|_{x_k=1})$ ; $t \leftarrow 0$.
Let $(-1)^k = \mathbf{a}_t(0) < \cdots < \mathbf{a}_t(k-1) = (1^{k-1}, -1)$ be a chain supporting $(A_t)|_{x_k=-1}$.
Let $((-1)^{k-1}, 1) = \mathbf{b}_t(0) < \cdots < \mathbf{b}_t(k-1) = 1^k$ be a chain supporting $(A_t)|_{x_k=1}$.
while $\exists (i,j)$ with $j < k-1$ such that $\mathbf{a}_t(i) \vee \mathbf{b}_t(j) = 1^k$ and $A_t(\mathbf{a}_t(i)), A_t(\mathbf{b}_t(j)) > 0$ do
    Let $(i_t, j_t)$ be the lexicographically smallest such pair $(i,j)$.
    $B_t \leftarrow (A_t)_{\mathbf{a}_t(i_t), \mathbf{b}_t(j_t)}$.
    $(A_{t+1})|_{x_k=-1} \leftarrow \textsc{Polarize}(B_t|_{x_k=-1})$ ; $(A_{t+1})|_{x_k=1} \leftarrow (B_t)|_{x_k=1}$.
    $t \leftarrow t+1$.
    Let $(-1)^k = \mathbf{a}_t(0) < \cdots < \mathbf{a}_t(k-1) = (1^{k-1}, -1)$ be a chain supporting $(A_t)|_{x_k=-1}$.
    Let $((-1)^{k-1}, 1) = \mathbf{b}_t(0) < \cdots < \mathbf{b}_t(k-1) = 1^k$ be a chain supporting $(A_t)|_{x_k=1}$.
Let $\ell \in [k]$ be such that for every $\mathbf{a} \in \{-1,1\}^k \setminus \{1^k\}$ we have $A_t(\mathbf{a}) > 0 \Rightarrow a_\ell = -1$.
$(A_{t+1})|_{x_\ell=-1} \leftarrow \textsc{Polarize}(A_t)|_{x_\ell=-1}$. $(A_{t+1})|_{x_\ell=1} \leftarrow (A_t)|_{x_\ell=1}$.
Output: $A_{t+1}$.
```

The goal of the rest of the proof is to show that Algorithm 2 terminates after a finite number of steps and outputs $A_{\boldsymbol{\mu}(A)}$.

Step 2: Correctness assuming Polarize terminates.

Claim 7.12 (Correctness condition of Polarize)**.** *For every $A \in \mathcal{F}(\{-1,1\}^k)$, if* Polarize *terminates, then* Polarize$(A) = A_{\boldsymbol{\mu}(A)}$*. In particular,* Polarize(A) *has the same marginals as A and is supported on a chain.*

Proof. First, by the definition of the polarization operator (Definition 7.3), the marginals of A_t are the same for every t. So in the rest of the proof, we focus on inductively showing that if Polarize terminates, then Polarize(A) is supported on a chain.

For the base case where $k = 2$, we always have Polarize$(A) = A_{(-1,1),(1,-1)}$ supported on a chain as desired.

When $k > 2$, note that when the algorithm enters the Clean-up stage, if we let m and n denote the largest indices such that $A_t(\mathbf{a}_t(m)), A_t(\mathbf{b}_t(n)) > 0$ and $A_t(\mathbf{b}_t(n)) \neq 1^k$, then the condition that $\mathbf{a}_t(m) \vee \mathbf{b}_t(n) \neq 1^k$ implies that there is a coordinate ℓ such that $\mathbf{a}_t(m)_\ell = \mathbf{b}_t(n)_\ell = -1$. Since every $\mathbf{c}$ such that $A_t(\mathbf{c}) > 0$ and $c_k = -1$ satisfies $\mathbf{c} \leq \mathbf{a}_t(m)$, we have $A_t(\mathbf{c}) > 0$ implies $c_\ell = -1$. Similarly for every $\mathbf{c} \neq 1^k$ such that $c_k = 1$, we have $A_t(\mathbf{c}) > 0$ implies $c_\ell = -1$. We conclude that A_t is supported on $\{1^k\} \cup \{\mathbf{c} \,|\, c_\ell = -1\}$. Thus, by the induction hypothesis, after polarizing the subcube $x_\ell = -1$ and leaving the subcube $x_\ell = 1$ unchanged, we get that the resulting function A_{t+1} is supported on a chain as desired and complete the induction. We conclude that if Polarize terminates, we have Polarize$(A) = A_{\boldsymbol{\mu}(A)}$. □

Step 3: Invariant in Polarize. Now, in the rest of the proof of Lemma 7.11, the goal is to show that for every input A, the number of iterations of the while loop in Algorithm 2 is finite. The key claim (Claim 7.16) here asserts that the sequence of pairs (i_t, j_t) is monotonically increasing in lexicographic order. Once we establish this claim, it follows that there are at most k^2 iterations of the while loop and so $N(k) \leq (k^2+3)\cdot(1+N(k-1))$, proving Lemma 7.11. Before proving Claim 7.16, we establish the following properties that remain invariant after every iteration of the while loop.

Claim 7.13. *For every $t \geq 0$, we have $\forall b \in \{-1,1\}$, $(A_t)|_{x_k=b}$ is supported on a chain.*

Proof. For $b=-1$, the claim follows from the correctness of the recursive call to Polarize. For $b=1$, we claim by induction on t that the supporting chain $\mathbf{b}_t(0) < \cdots < \mathbf{b}_t(k-1)$ never changes (with t). To see this, note that $\mathbf{b}_t(k-1) = 1^k$ is the only point in the subcube $\{x_k = 1\}$ that increases in value compared to A_t, and this is already in the supporting chain. Thus $\mathbf{b}_t(0) < \cdots < \mathbf{b}_t(k-1)$ continues to be a supporting chain for $(A_{t+1})|_{x_k=1}$. □

For $\mathbf{c} \in \{-1,1\}^k$, we say that a function $A : \{-1,1\}^k \to \mathbb{R}^{\geq 0}$ is $\mathbf{c}$-*subcube-respecting* ($\mathbf{c}$-respecting, for short) if for every $\mathbf{c}'$ such that $A(\mathbf{c}') > 0$, we have $\mathbf{c}' \geq \mathbf{c}$ or $\mathbf{c}' \leq \mathbf{c}$. We say that A is $\mathbf{c}$-*downward-respecting* if A is $\mathbf{c}$-respecting and the points in the support of A above $\mathbf{c}$ form a partial chain, specifically, if $\mathbf{u},\mathbf{v} > \mathbf{c}$ have $A(\mathbf{u}), A(\mathbf{v}) > 0$ then either $\mathbf{u} \geq \mathbf{v}$ or $\mathbf{v} \geq \mathbf{u}$.

Note that if A is supported on a chain then A is $\mathbf{c}$-respecting for every point $\mathbf{c}$ in the chain. Conversely, if A is supported on a chain and A is $\mathbf{c}$-respecting, then A is supported on a chain that includes $\mathbf{c}$.

Claim 7.14 (Polarization on subcubes)**.** *Let A be a $\mathbf{c}$-respecting function and let $\tilde{A}$ be obtained from A by a finite sequence of polarization updates, as in Definition 7.3. Then $\tilde{A}$ is also $\mathbf{c}$-respecting. Furthermore if A is $\mathbf{c}$-downward-respecting and $\mathbf{w} > \mathbf{c}$ then $\tilde{A}$ is also $\mathbf{c}$-downward-respecting and $A(\mathbf{w}) = \tilde{A}(\mathbf{w})$.*

Proof. Note that it suffices to prove the claim for a single update by a polarization operator since the rest follows by induction. So let $\tilde{A} = A_{\mathbf{u},\mathbf{v}}$ for incomparable $\mathbf{u},\mathbf{v} \in \mathsf{supp}(A)$.

Since A is $\mathbf{c}$-respecting, and $\mathbf{u},\mathbf{v}$ are incomparable, either $\mathbf{u} \leq \mathbf{c}, \mathbf{v} \leq \mathbf{c}$ or $\mathbf{u} \geq \mathbf{c}, \mathbf{v} \geq \mathbf{c}$. Suppose the former is true, then $\mathbf{u} \vee \mathbf{v} \leq \mathbf{c}$ and $\mathbf{u} \wedge \mathbf{v} \leq \mathbf{c}$, and hence, $\tilde{A}$ is $\mathbf{c}$-respecting. Similarly, in the case when $\mathbf{u} \geq \mathbf{c}, \mathbf{v} \geq \mathbf{c}$, we can show that $\tilde{A}$ is $\mathbf{c}$-respecting. The furthermore part follows by noticing that for $\mathbf{u}$ and $\mathbf{v}$ to be incomparable if A is $\mathbf{c}$-downward-respecting and $A(\mathbf{u}), A(\mathbf{v}) > 0$, then $\mathbf{u},\mathbf{v} \leq \mathbf{c}$, and so the update changes A only at points below $\mathbf{c}$. □

The following claim asserts that in every iteration of the while loop, by the lexicographically minimal choice of (i_t, j_t), there exists a coordinate $h \in [k-1]$ such that every vector $c < a_t(i_t)$ in the support of A_t, B_t, or A_{t+1} has $c_h = -1$, and every vector $c \neq 1^k$ in the support of $(A_t)|_{x_k=1}$ has $c_h = -1$.

Claim 7.15. *For every $t \geq 0$, $\exists h \in [k-1]$ such that $\forall \mathbf{c} \in \{-1,1\}^k$, if $\mathbf{c} \in \mathsf{supp}(A_t) \cup \mathsf{supp}(B_t) \cup \mathsf{supp}(A_{t+1})$, then the following hold:*

- *If $\mathbf{c} < \mathbf{a}_t(i_t)$, then $c_h = -1$.*
- *If $c_k = 1$ and $\mathbf{c} \neq 1^k$, then $c_h = -1$.*

Proof. Since (i_t, j_t) is lexicographically the smallest incomparable pair in the support of A_t, for $i < i_t$, $j < k-1$, and $A_t(\mathbf{a}(i)), A_t(\mathbf{b}(j)) > 0$, we have $\mathbf{a}(i) \vee \mathbf{b}(j) \neq 1^k$. Let m be the largest index smaller than i_t such that $A_t(\mathbf{a}_t(m)) > 0$. Similarly, let $n < k-1$ be the largest index such that

$A_t(\mathbf{b}_t(n)) > 0$. Then the fact that $\mathbf{a}_t(m) \vee \mathbf{b}_t(n) \neq 1^k$ implies that there exists $h \in [k-1]$ such that $\mathbf{a}_t(m)_h = \mathbf{b}_t(n)_h = -1$. Now, using the fact (from Claim 7.13) that $(A_t)|_{x_k=-1}$ is supported on a chain, we conclude that for every $\mathbf{c} < \mathbf{a}_t(i_t)$, $A_t(\mathbf{c}) > 0$ implies that $\mathbf{c} \leq \mathbf{a}_t(m)$ and hence, $c_h = -1$. Similarly, for every vector $\mathbf{c} \neq 1^k$ in the support of $(A_t)|_{x_k=1}$, by the maximality of n, we have $c_h = -1$.

We now assert that the same holds for B_t. First, recall that $\mathsf{supp}(B_t) \subset \mathsf{supp}(A_t) \cup \{1^k, \mathbf{a}_t(i_t) \wedge \mathbf{b}_t(j_t)\}$ since $B_t = (A_t)_{\mathbf{a}_t(i_t),\mathbf{b}_t(j_t)}$. Next, note that the only point (other than 1^k) where B_t is larger than A_t is $\mathbf{a}_t(i_t) \wedge \mathbf{b}_t(j_t)$. It suffices to show that $(\mathbf{a}_t(i_t) \wedge \mathbf{b}_t(j_t))_h = -1$. We have $\mathbf{a}_t(i_t) \wedge \mathbf{b}_t(j_t) \leq \mathbf{b}_t(j_t) \leq \mathbf{b}_t(n)$ and hence $(\mathbf{a}_t(i_t) \wedge \mathbf{b}_t(j_t))_h = -1$.

Finally, we assert that same holds also for A_{t+1}. Since $A_{t+1}|_{x_k=1} = B_t|_{x_k=1}$, the second item in the claim follows trivially. To prove the first item, let us consider $\mathbf{a}' \in \{-1,1\}^k$ defined as follows: $\mathbf{a}'_h = -1$ and $\mathbf{a}'_r = \mathbf{a}_t(i_t)_r$ for $r \neq h$. Note that $B_t|_{x_k=-1}$ is $\mathbf{a}_t(i_t)$-respecting since potentially the only new point in its support (compared to $A_t|_{x_k=-1}$) is $\mathbf{a}_t(i_t) \wedge \mathbf{b}_t(j_t) \leq \mathbf{a}_t(i_t)$. From the previous paragraph we also have that if $B_t(\mathbf{c}) > 0$ and $\mathbf{c} < \mathbf{a}_t(i_t)$, then $c_h = -1$ and hence, $\mathbf{c} \leq \mathbf{a}'$. On the other hand, if $B_t(\mathbf{c}) > 0$ and $\mathbf{c} \geq \mathbf{a}_t(i_t)$, then $\mathbf{c} \geq a'$. Therefore, $B_t|_{x_k=-1}$ is $\mathbf{a}'$-respecting. By applying Claim 7.14, we conclude that $(A_{t+1})|_{x_k=-1}$ is also $\mathbf{a}'$-respecting. It follows that if $\mathbf{c} < \mathbf{a}(i_t)$ and $A_{t+1}(\mathbf{c}) > 0$, then $\mathbf{c} \leq \mathbf{a}'$ and so $c_h = -1$. □

Step 4: Proof of Lemma 7.11. The following claim establishes that the while loop in the Polarize algorithm terminates after a finite number of iterations.

Claim 7.16. *For every $t \geq 0$, $(i_t, j_t) < (i_{t+1}, j_{t+1})$ in lexicographic ordering.*

Proof. Consider the chain $\mathbf{a}_{t+1}(0) < \cdots < \mathbf{a}_{t+1}(k-1)$ supporting $A_{t+1}|_{x_k=-1}$. Note that for $i \geq i_t$, $A_{t+1}|_{x_k=-1}$ is $\mathbf{a}_t(i)$-respecting (since $A_t|_{x_k=-1}$ and $B_t|_{x_k=-1}$ were also so). In particular, $A_t|_{x_k=-1}$ is $\mathbf{a}_t(i)$-respecting because it is supported on a chain containing $a_t(i)$. Next $B_t|_{x_k=-1}$ is $\mathbf{a}_t(i)$-respecting since potentially the only new point in its support is $\mathbf{a}_t(i_t) \wedge \mathbf{b}_t(j_t) \leq \mathbf{a}_t(i)$. Finally, $A_{t+1}|_{x_k=-1}$ is also $\mathbf{a}_t(i)$-respecting using Claim 7.14. Thus we can build a chain containing $\mathbf{a}_t(i)$ that supports $A_{t+1}|_{x_k=-1}$. It follows that we can use $\mathbf{a}_{t+1}(i) = \mathbf{a}_t(i)$ for $i \geq i_t$. Now consider $i < i_t$. We must have $\mathbf{a}_{t+1}(i) < \mathbf{a}_{t+1}(i_t) = \mathbf{a}_t(i_t)$. By Claim 7.15, there exists $h \in [k-1]$ such that for $i < i_t$, $\mathbf{a}_{t+1}(i)_h = -1$.

We now turn to analyzing (i_{t+1}, j_{t+1}). Note that by definition, $A_{t+1}(\mathbf{a}_{t+1}(i_{t+1})) > 0$ and $A_{t+1}(\mathbf{b}_{t+1}(b_{t+1})) > 0$. First, let us show that $i_t \leq i_{t+1}$. On the contrary, let us assume that $i_{t+1} < i_t$. It follows from the above paragraph that $\mathbf{a}_{t+1}(i_{t+1})_h = -1$. Also, for every $\mathbf{b}_{t+1}(j)$ with $j < k-1$ and $A_{t+1}(\mathbf{b}_{t+1}(j)) > 0$, we have $\mathbf{b}_{t+1}(j)_h = -1$. Therefore, $\mathbf{a}(i_{t+1}) \vee \mathbf{b}(j_{t+1}) \neq 1^k$ (in particular $(\mathbf{a}(i_{t+1}) \vee \mathbf{b}(j_{t+1}))_h = -1$), which is a contradiction.

Next, we show that if $i_{t+1} = i_t$, then $j_{t+1} \geq j_t$. By the minimality of (i_t, j_t) in the t-th round, for $j < j_t$ such that $A_t(b_t(j)) > 0$, we have $a_t(i_t) \vee b_t(j) \neq 1^k$. Since $i_{t+1} = i_t$, $a_{t+1}(i_{t+1}) = a_{t+1}(i_t) = a_t(i_t)$. We already noted in the proof of Claim 7.13 that $\mathbf{b}_t(0) < \cdots < \mathbf{b}_t(k-1)$ is also a supporting chain for $(A_{t+1})|_{x_k=1}$. The only point where the function $A_{t+1}|_{x_k=1}$ has greater value than $A_t|_{x_k=1}$ is 1^k. Therefore, for $j < j_t$ such that $A_{t+1}(b_{t+1}(j)) > 0$, we have $a_{t+1}(i_{t+1}) \vee b_{t+1}(j) \neq 1^k$ and hence, $j_{t+1} \geq j_t$.

So far, we have established that $(i_{t+1}, j_{t+1}) \geq (i_t, j_t)$ in lexicographic ordering. Finally, we will show that $(i_{t+1}, j_{t+1}) \neq (i_t, j_t)$ by proving that at least one of $A_{t+1}(\mathbf{a}_{t+1}(i_t))$ and $A_{t+1}(\mathbf{b}_{t+1}(j_t))$ is zero. The polarization update ensures that at least one of $B_t(\mathbf{a}_t(i_t))$ and $B_t(\mathbf{b}_t(j_t))$ is zero. If $B_t(\mathbf{b}_t(j_t)) = 0$, then by definition, we have $A_{t+1}(\mathbf{b}_{t+1}(j_t)) = A_{t+1}(\mathbf{b}_t(j_t)) = 0$. Finally to handle the case $B_t(\mathbf{a}_t(i_t)) = 0$, let us again define $\mathbf{a}'$ as: $\mathbf{a}'_h = -1$ and $\mathbf{a}'_r = \mathbf{a}_t(i_t)_r$ for $r \neq h$, where h is as given by Claim 7.15. We assert that $B_t|_{x_k=-1}$ is $\mathbf{a}'$-downward-respecting. As shown in the proof of Claim 7.15, we have $B_t|_{x_k=-1}$ is $\mathbf{a}'$-respecting. The support of $B_t|_{x_k=-1}$ is contained in $\{\mathbf{a}_t(0), \cdots, \mathbf{a}_t(k-1)\} \cup \{\mathbf{a}_t(i_t) \wedge \mathbf{b}_t(j_t)\}$ and $\mathbf{a}_t(i_t) \wedge \mathbf{b}_t(j_t) < \mathbf{a}_t(i_t)$, and by Claim 7.15, $\mathbf{a}_t(i_t) \wedge \mathbf{b}_t(j_t) \leq \mathbf{a}'$. It follows that $B_t|_{x_k=-1}$ is $\mathbf{a}'$-downward-respecting. Finally, by the furthermore part of Claim 7.14 applied to $B_t|_{x_k=-1}$ and $\mathbf{w} = \mathbf{a}_t(i_t)$, we get that $A_{t+1}(\mathbf{a}_{t+1}(i_t)) = A_{t+1}(\mathbf{a}_t(i_t)) = B_t(\mathbf{a}_t(i_t)) = 0$. It follows that $(i_{t+1}, j_{t+1}) \neq (i_t, j_t)$. □

Proof of Lemma 7.11. By Claim 7.12, we know that if Algorithm 2 terminates, we have Polarize$(A) = A_{\boldsymbol{\mu}(A)}$. Hence, the maximum number of polarization updates used in Polarize (on input from $\mathcal{F}(\{-1,1\}^k)$) serves as an upper bound for $N(k)$. By Claim 7.16, we know that there are at most k^2 iterations of the while loop and so $N(k) \leq (k^2+3) \cdot (1 + N(k-1))$ as desired. □

7.4 Putting it together

We now have the ingredients in place to prove Theorem 5.3.

Proof of Theorem 5.3. Given distribution $\mathcal{D}_Y, \mathcal{D}_N$ with $\boldsymbol{\mu} = \boldsymbol{\mu}(\mathcal{D}_Y) = \boldsymbol{\mu}(\mathcal{D}_N)$, first we apply Lemma 7.11 to $\mathcal{D}_Y$ to get $\mathcal{D}_0 = \mathcal{D}_Y, \mathcal{D}_1, \ldots, \mathcal{D}_t = \mathcal{D}_{\boldsymbol{\mu}}$ such that $\mathcal{D}_{i+1} = (\mathcal{D}_i)_{\mathbf{u}(i),\mathbf{v}(i)}$, i.e., $\mathcal{D}_i$ is an update of $\mathcal{D}_i$, with $t \leq N(k) < \infty$. Similarly, we apply Lemma 7.11 to $\mathcal{D}_N$ to get $\mathcal{D}'_0 = \mathcal{D}_N, \mathcal{D}'_1, \ldots, \mathcal{D}'_{t'} = \mathcal{D}_{\boldsymbol{\mu}}$ such that $\mathcal{D}'_{i+1} = (\mathcal{D}'_i)_{\mathbf{u}'(i),\mathbf{v}'(i)}$ with $t' \leq N(k) < \infty$.

Now, Lemma 7.7, applied to the pairs $\mathcal{D}_i$ and $\mathcal{D}_{i+1}$ with $\delta' = \delta/(2N(k))$, gives is τ_i such that every protocol for $(\mathcal{D}_i, \mathcal{D}_{i+1})$-RMD requires $\tau_i\sqrt{n}$ bits of communication to achieve advantage δ'. Similarly applying Lemma 7.7 again with $\delta' = \delta/(2N(k))$ to the pairs $\mathcal{D}'_i$ and $\mathcal{D}'_{i+1}$, we get τ'_i such that every protocol for $(\mathcal{D}'_i, \mathcal{D}'_{i+1})$-RMD requires $\tau'_i\sqrt{n}$ bits of communication to achieve advantage δ'.

Letting $\tau = \min\left\{\min_{i\in[t]}\{\tau_i\}, \min_{i\in[t']}\{\tau'_i\}\right\}$, we get, using the triangle inequality for indistinguishability, that every protocol Π for $(\mathcal{D}_Y, \mathcal{D}_N)$-RMD achieving advantage $\delta \geq (t+t')\delta'$ requires $\tau\sqrt{n}$ communication. □

Acknowledgments

We are grateful to Lijie Chen, Gillat Kol, Dmitry Paramonov, Raghuvansh Saxena, Zhao Song, and Huacheng Yu, for detecting a fatal error in an earlier version of this paper [CGSV21] and then for pinpointing the location of the error. As a result the main theorem of the current paper is significantly different than the theorem claimed in the previous version.

Thanks to Johan Håstad for many pointers to the work on approximation resistance and answers to many queries. Thanks to Dmitry Gavinsky, Julia Kempe and Ronald de Wolf for prompt and detailed answers to our queries on outdated versions of their work [GKK+09]. Thanks to Prasad Raghavendra for answering our questions about the approximation resistance dichotomy from his work [Rag08]. Thanks to Saugata Basu for the pointers to the algorithms for quantified theory of the reals. Thanks to Jelani Nelson for pointers to ℓ_1 norm estimation algorithms.

Thanks to Michael Hwang and Tarun Prasad for pointing out some errors in Example 1 in a previous version of this paper. Thanks to Noah Singer for pointing out some typos in the paper.

References

[AKSY20] Sepehr Assadi, Gillat Kol, Raghuvansh R Saxena, and Huacheng Yu. Multi-Pass Graph Streaming Lower Bounds for Cycle Counting, MAX-CUT, Matching Size, and Other Problems. In *FOCS 2020*, 2020.

[AM09] Per Austrin and Elchanan Mossel. Approximation resistant predicates from pairwise independence. *Comput. Complex.*, 18(2):249–271, 2009.

[BPR06] Saugata Basu, Richard Pollack, and Marie-Françoise Roy. *Algorithms in Real Algebraic Geometry*. Springer, 2006.

[Bul17] Andrei A. Bulatov. A dichotomy theorem for nonuniform CSPs. In Chris Umans, editor, *FOCS 2017*, pages 319–330. IEEE, 2017.

[BV04] Stephen P. Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

[CGSV21] Chi-Ning Chou, Alexander Golovnev, Madhu Sudan, and Santhoshini Velusamy. Classification of the streaming approximability of Boolean CSPs. *CoRR*, abs/2102.12351v1, 2021.

[CGV20] Chi-Ning Chou, Alexander Golovnev, and Santhoshini Velusamy. Optimal streaming approximations for all Boolean Max-2CSPs and Max-kSAT. In *FOCS 2020*. IEEE, 2020.

[Cha20] Amit Chakrabarti. Data stream algorithms. *Lecture notes*, page 94, 2020.

[GKK$^+$09] Dmitry Gavinsky, Julia Kempe, Iordanis Kerenidis, Ran Raz, and Ronald de Wolf. Exponential separation for one-way quantum communication complexity, with applications to cryptography. *SIAM J. Comput.*, 38(5):1695–1708, 2009.

[GM08] Sudipto Guha and Andrew McGregor. Tight lower bounds for multi-pass stream computation via pass elimination. In *ICALP 2008*, pages 760–772. Springer, 2008.

[GT19] Venkatesan Guruswami and Runzhou Tao. Streaming hardness of unique games. In *APPROX 2019*, pages 5:1–5:12. LIPIcs, 2019.

[GVV17] Venkatesan Guruswami, Ameya Velingker, and Santhoshini Velusamy. Streaming complexity of approximating Max 2CSP and Max Acyclic Subgraph. In *APPROX 2017*. LIPIcs, 2017.

[Ind00] Piotr Indyk. Stable distributions, pseudorandom generators, embeddings and data stream computation. In *FOCS 2000*, pages 189–197. IEEE, 2000.

[Kho02] Subhash Khot. On the power of unique 2-prover 1-round games. In *STOC 2002*, pages 767–775. ACM, 2002.

[KK19] Michael Kapralov and Dmitry Krachun. An optimal space lower bound for approximating MAX-CUT. In *STOC 2019*, pages 277–288. ACM, 2019.

[KKL88] Jeff Kahn, Gil Kalai, and Nathan Linial. The influence of variables on Boolean functions. In *FOCS 1988*, pages 68–80. IEEE, 1988.

[KKS15] Michael Kapralov, Sanjeev Khanna, and Madhu Sudan. Streaming lower bounds for approximating MAX-CUT. In *SODA 2015*, pages 1263–1282. SIAM, 2015.

[KKSV17] Michael Kapralov, Sanjeev Khanna, Madhu Sudan, and Ameya Velingker. $(1 + \omega(1))$-approximation to MAX-CUT requires linear space. In *SODA 2017*, pages 1703–1722. SIAM, 2017.

[KNW10] Daniel M. Kane, Jelani Nelson, and David P. Woodruff. On the exact space complexity of sketching and streaming small norms. In *SODA 2010*, pages 1161–1178. SIAM, 2010.

[KTW14] Subhash Khot, Madhur Tulsiani, and Pratik Worah. A characterization of strong approximation resistance. In *STOC 2014*, pages 634–643, 2014.

[McG14] Andrew McGregor. Graph stream algorithms: a survey. *SIGMOD Record*, 43(1):9–20, 2014.

[O'D14] Ryan O'Donnell. *Analysis of Boolean functions*. Cambridge University Press, 2014.

[Pot19] Aaron Potechin. On the approximation resistance of balanced linear threshold functions. In Moses Charikar and Edith Cohen, editors, *STOC 2019*, pages 430–441. ACM, 2019.

[Rag08] Prasad Raghavendra. Optimal algorithms and inapproximability results for every CSP? In *STOC 2008*, pages 245–254, 2008.

[Sch78] Thomas J. Schaefer. The complexity of satisfiability problems. In *STOC 1978*, pages 216–226. ACM, 1978.

[Yao77] Andrew Chi-Chin Yao. Probabilistic computations: Toward a unified measure of complexity. In *FOCS 1977*, pages 222–227. IEEE, 1977.

[Zhu17] Dmitriy Zhuk. A proof of CSP dichotomy conjecture. In *FOCS 2017*, pages 331–342. IEEE, 2017.