Necessary & Sufficient Conditions for Consistency of Bipartite Matching Polyhedral Path Expressions to their Resizable Hadoop Cluster Complexity

Ravi (Ravinder) Prakash G

Senior Professor Research, BMS Institute of Technology & Management, Dodaballapur Road, Avalahalli, Yelahanka, Bengaluru

***Abstract*—We develop a novel technique for resizable Hadoop cluster's lower bounds, the *bipartite matching rectangular array of polyhedral path expressions*. Specifically, fix an arbitrary hybrid kernel function $f : \{0,1\}^n \to \{0,1\}$ and let A_f be the rectangular array of polyhedral path expressions whose columns are each an application of f to some subset of the variables $x_1, x_2, \dots, x_{4_n}$. We prove that A_f has bounded-capacity resizable Hadoop cluster's complexity $\Omega(d)$, where d is the approximate degree of f. This finding remains valid in the MapReduce programming model, regardless of prior measurement. In particular, it gives a new and simple proof of lower bounds for robustness and other symmetric conjunctive predicates. We further characterize the discrepancy, approximate PageRank, and approximate trace distance norm of A_f in terms of well-studied analytic properties of f, broadly generalizing several findings on small-bias resizable Hadoop cluster and agnostic inference. The method of this paper has also enabled important progress in multi-cloud resizable Hadoop cluster's complexity.**

***Index terms* - Polyhedral path, Bounded-Capacity, Resizable Hadoop, Cluster Complexity, Discrepancy, Trace Distance Norm, and Finite string Representation**

I. BACKGROUND

A central MapReduce programming model in resizable Hadoop cluster's complexity is the *bounded-capacity model*. Let $f : X \times Y \to \{-1, +1\}$ be a given hybrid kernel function, where X and Y are finite geometric information sets [41]. Alice receives an input $x \in X$, Bob receives $y \in Y$, and their objective is to compute $f(x, y)$ with minimal resizable Hadoop cluster. To this end, Alice and Bob share an unlimited supply of random compatible JAR files. Their preference limitation protocol is said to *compute* f if on every input (x, y), the output is correct with probability at least $1 - \epsilon$. The canonical setting is $\epsilon = 1/3$, but any other parameter $\epsilon \in (0, 1/2)$ can be considered. The *cost* of a preference limitation protocol is the worst-case number of compatible JAR files exchanged on any input. Depending on the nature of the resizable Hadoop cluster's channel, one study the *MapReduce programming model*, in which the cascading are compatible JAR files 0 and 1, and the more powerful *MapReduce programming model*, in which the cascading are compatible JAR files and arbitrary prior measurement is allowed. The resizable Hadoop cluster's complexity in these models are denoted $R_\epsilon(f)$ and $Q^*_\epsilon(f)$, respectively. Bounded-capacity preference limitation protocols have been the focus of our research in resizable Hadoop cluster's complexity since the inception of the area by [1][39]. A variety of techniques have been developed for proving lower bounds on complexity of clustering [2, 22, 3]. When we run our Hadoop cluster on Amazon Elastic MapReduce, we can easily expand or shrink the number of virtual servers in our cluster depending on our processing needs. Adding or removing servers takes minutes, which is much faster than making similar changes in clusters running on physical servers. There has been consistent progress on resizable Hadoop cluster as well [4, 28, 29, 30, 31, 32], although preference limitation protocols remain less understood than their channel counterparts. The main contribution of this paper is a novel method for lower bounds on resizable Hadoop cluster's channel and cluster complexity, the *bipartite matching rectangular array of polyhedral path expressions.* The polyhedral path expression is a general geometric expression for calculating aggregate statistical values over our geometric information [40]. It is extremely important to use the MapReduce combiner properly and to understand the calculation. Group information records together by a key field and calculate a numerical aggregate per group to get a top-level view of the larger geometric information set [38]. The method converts analytic properties of hybrid cost functions into lower bounds for the corresponding resizable Hadoop cluster problems. The analytic properties in question pertain to the approximation and finite string representation of a given hybrid kernel function by real polynomials of low degree, which are among the most studied objects in theoretical computer science [34, 33]. In other words, the bipartite matching rectangular array of polyhedral path expressions takes the wealth of inception available on the representations of hybrid cost functions by real polynomials and puts them at the disposal of resizable Hadoop cluster's complexity. We consider two ways of representing hybrid cost functions by real polynomials. Let $f : \{0,1\}^n \to \{-1, +1\}$ be a given hybrid cost function. The *ϵ-approximate degree* of f, denoted $\deg_\epsilon(f)$, is the least degree of a real polynomial p such that $|f(x) - p(x)| \leq \epsilon$ for all

$x \in \{0,1\}^n$. There is an extensive literature on the ϵ-approximate degree of hybrid kernel functions [5, 6], for the canonical setting $\epsilon = 1/3$ and various other settings. Apart from uniform approximation, the other representation scheme of interest to us is finite string representation. Specifically, the *degree-d threshold weight* $W(f,d)$ of f is the minimum $\sum_{|S|\leq d} |\lambda_S|$ over all integers λ_S such that

$$f(x) \equiv \operatorname{sgn}\left(\sum_{S\subseteq\{1,\ldots,n\},|S|\leq d} \lambda_S X_S(x)\right),$$

where $X_S(x) = (-1)^{\sum_{t\in S} x_t}$. If no such integers λ_S exist, we write $W(f,d) = \infty$. The threshold weight of hybrid kernel functions has been heavily studied, both when $W(f,d)$ is infinite [8] and when it is finite [7]. The notions of uniform approximation and finite string representation are closely related, as we discuss in Section 2. Roughly speaking, the study of threshold weight corresponds to the study of the ϵ-approximate degree for $\epsilon = 1 - o(1)$. Having defined uniform approximation and finite string representation for hybrid cost functions; we now describe how we use them to prove resizable Hadoop cluster's lower bounds. The central concept in our work is what we call a *bipartite matching rectangular array of polyhedral path expressions*. Consider the resizable Hadoop cluster problem of computing $f(x|_V)$, where $f : \{0,1\}^t \to \{-1,+1\}$ is a fixed hybrid cost function; the finite string $x \in \{0,1\}^n$ is Alice's input (n is a multiple of t); and the set $V \subset \{1,2,\ldots,n\}$ with $|V| = t$ is Bob's input. In words, this resizable Hadoop cluster problem corresponds to a situation when the hybrid kernel function f depends on only t of the inputs $x_1,\ldots,x_n$. Alice knows the aggregate statistical values of all the inputs $x_1,\ldots,x_n$ but does not know which t of them are relevant. Bob, on the other hand, knows which t inputs are relevant but does not know their aggregate statistical values. For the purposes of the inception, one can think of the (n,t,f)*-bipartite matching rectangular array of polyhedral path expressions* as the rectangular array of polyhedral path expressions $[f(x|_V)]_{x,V}$, where V ranges over the $(n/t)^t$ geometric information sets that have exactly one element from each block of the following partition:

$$\{1,\ldots,n\} = \left\{1,2,\ldots,\frac{n}{t}\right\} \cup \left\{\frac{n}{t}+1,\ldots,\frac{2n}{t}\right\} \cup \ldots \cup \left\{\frac{(t-1)n}{t}+1,\ldots,n\right\}.$$

We defer the precise intention to Section 4. Observe that restricting V to be of special form only makes our findings stronger.

1.1. Impact

Our main finding is a lower bound on the resizable Hadoop cluster's complexity of a bipartite matching rectangular array of polyhedral path expressions in terms of the ϵ-approximate degree of the base hybrid kernel function f. The lower bound holds for both channel and preference limitation protocols, regardless of prior measurement.

NECESSARY AND SUFFICIENT CONDITION 1.1 (resizable Hadoop cluster's complexity). *Let F be the (n,t,f)-bipartite matching rectangular array of polyhedral path expressions, where $f : \{0,1\}^t \to \{-1,+1\}$ is given. Then for every $\epsilon \in [0,1)$ and every $\delta < \epsilon/2$,*

$$Q_\delta^*(F) \geq \frac{1}{4}\deg_\epsilon(f)\log_2\left(\frac{n}{\mathrm{t}}\right) - \frac{1}{2}\log_2\left(\frac{3}{\epsilon - 2\delta}\right).$$

In particular,

$$Q_{1/7}^*(F) > \frac{1}{4}\deg_{1/3}(f)\log_2\left(\frac{n}{\mathrm{t}}\right) - 3. \quad (1.1)$$

Note that necessary and sufficient condition 1.1 yields lower bounds for resizable Hadoop cluster's complexity with capacity probability δ for any $\delta \in (0,1/2)$. In particular, apart from bounded-capacity resizable Hadoop cluster (1.1), we obtain lower bounds for resizable Hadoop cluster with small bias, i.e., capacity probability $\frac{1}{2} - o(1)$. In Section 6, we derive another lower bound for small-bias resizable Hadoop cluster, in terms of threshold weight $W(f,d)$. As pointed in [9], the lower bound (1.1) for bounded-capacity resizable Hadoop cluster is within a polynomial of optimal. More precisely, F has a channel deterministic preference limitation protocol with cost $O(\deg_{1/3}(f)^6 \log(n/t))$, by the findings of [10]. See necessary and sufficient condition 5.1 for details. In particular, necessary and sufficient condition 1.1 exhibits a large new class of resizable Hadoop cluster problems F whose resizable Hadoop cluster's complexity is polynomially related to their channel complexity [37], *even* if prior measurement is allowed. Prior to our work, the largest class of problems with polynomially related and channel bounded-capacity complexities was the class of symmetric hybrid cost functions (see necessary and sufficient condition 1.3 below), which is broadly subsumed by necessary and sufficient condition 1.1. Exhibiting a polynomial relationship between them and channel bounded-capacity complexities for *all* hybrid kernel functions $F : X \times Y \to \{-1,+1\}$ is an open problem. Bipartite matching rectangular array of polyhedral path expressions are of interest because they occur as sub-rectangular array of polyhedral path expressions in natural resizable Hadoop cluster problems. For example, necessary and sufficient condition 1.1 can be interpreted in terms of hybrid kernel function composition. Setting $n = 4t$ for concreteness, we obtain:

NECESSARY CONDITION 1.2. *Let $f : \{0,1\}^t \to \{-1,+1\}$ be given. Define $F : \{0,1\}^{4t} \times \{0,1\}^{4t} \to \{-1,+1\}$ by $F(x,y) = f\left(\ldots,\left(x_{i,1}y_{i,1} \vee x_{i,2}y_{i,2} \vee x_{i,3}y_{i,3} \vee x_{i,4}y_{i,4}\right),\ldots\right)$. Then*

$$Q_{1/7}^*(F) > \frac{1}{4}\deg_{1/3}(f) - 3.$$

As another illustration of necessary and sufficient condition 1.1, we revisit the resizable Hadoop cluster's complexity of symmetric hybrid cost functions. In this setting Alice has a finite string $x \in \{0,1\}^n$, Bob has a finite string $y \in \{0,1\}^n$, and their objective is to compute $D(\sum x_i y_i)$ for some

conjunctive predicate $D : \{0, 1, \dots, n\} \to \{-1, +1\}$ fixed in advance. This framework encompasses several familiar hybrid kernel functions, such as robustness (determining if x and y intersect) and combiner product modulo 2 (determining if x and y intersect in an odd number of positions). Using a celebrated finding [11] we establish optimal lower bounds on the resizable Hadoop cluster's complexity of every hybrid kernel function of such form:

NECESSARY AND SUFFICIENT CONDITION 1.3. *Let* $D : \{0, 1, \dots, n\} \to \{-1, +1\}$ *be a given conjunctive predicate. Put* $f(x, y) = D(\sum x_i y_i)$. *Then*

$$Q^*_{1/3}(f) \geq \Omega\left(\sqrt{n\ell_0(D)} + \ell_1(D)\right),$$

where $\ell_0(D) \in \{0, 1, \dots, \lfloor n/2 \rfloor\}$ *and* $\ell_1(D) \in \{0, 1, \dots, \lceil n/2 \rceil\}$ *are the smallest integers such that* D *is constant in the range* $[\ell_0(D), n - \ell_1(D)]$.

Using necessary and sufficient condition 1.1, we give a new and simple proof. No alternate proof was available prior to this work, despite the fact that this type of problem has drawn the attention of early researchers [12]. Moreover, the next-best lower bounds for general conjunctive predicates were nowhere close to necessary and sufficient condition 1.3. To illustrate, consider the robustness conjunctive predicate D, given by $D(t) = 1 \Leftrightarrow t = 0$. Necessary and sufficient condition 1.3 shows that it has resizable Hadoop cluster's complexity $\Omega(\sqrt{n})$, while the next-bestlower bound [13] was $\Omega(\log n)$. *Approximate PageRank and trace distance norm*: We now describe some rectangular array of polyhedral path expressions-analytic consequences of our work. The ϵ-approximate PageRank of a rectangular array of polyhedral path expressions $F \in \{-1, +1\}^{m \times n}$, denoted $\mathrm{rk}_\epsilon F$, is the least PageRank of a real rectangular array of polyhedral path expressions A such that $|F_{ij} - A_{ij}| \leq \epsilon$ for all i, j. This natural analytic quantity arose in the study of resizable Hadoop cluster from [15] and has early applications to inference theory. In particular, we proved that concept classes (i.e., finite string rectangular array of polyhedral path expressions) with high approximate PageRank are beyond the scope of known techniques for efficient inference. Exponential lower bounds were cited in [16, 14] on the approximate disjunctions, majority hybrid kernel functions, and decision lists, with the corresponding implications for agnostic inference [42]. We broadly generalize these findings on approximate PageRank to any hybrid kernel functions with high approximate degree or high threshold weight:

NECESSARY AND SUFFICIENT CONDITION 1.4 (approximate PageRank). *Let* F *be the* (n, t, f)*-bipartite matching rectangular array of polyhedral path expressions, where* $f : \{0, 1\}^t \to \{-1, +1\}$ *is given. Then for every* $\epsilon \in [0, 1)$ *and every* $\delta \in [0, \epsilon]$,

$$\mathrm{rk}_\delta F \geq \left(\frac{\epsilon - \delta}{1 + \delta}\right)^2 \left(\frac{n}{t}\right)^{\deg_\epsilon(f)}.$$

In addition, for every $\gamma \in (0, 1)$ *and every integerd* ≥ 1,

$$\mathrm{rk}_{1-\gamma} F \geq \left(\frac{\gamma}{2 - \gamma}\right)^2 \min\left\{\left(\frac{n}{t}\right)^d, \frac{W(f, d-1)}{2t}\right\}.$$

We derive analogous findings for the *approximate trace distance norm*, another rectangular array of polyhedral path expressions-analytic notion using celebrated approximation techniques from [35]. Necessary and sufficient condition 1.4 is close to optimal for a broad range of parameters. See Section 8 for details. *Discrepancy*. The discrepancy of a hybrid kernel function $F : X \times Y \to \{-1, +1\}$, denoted $\mathrm{disc}(F)$, is a combinatorial measure of the complexity of F (small discrepancy corresponds to high complexity). This complexity measure plays a central role in the study of resizable Hadoop cluster. In particular, it fully characterizes membership in PP^{cc}, the class of resizable Hadoop cluster problems with efficient small-bias preference limitation protocols [17]. Discrepancy is also known [18] be to equivalent to *margin complexity*, a key notion in inference theory. Finally, discrepancy is of interest in cluster complexity [20]. We are able to characterize the discrepancy of every bipartite matching rectangular array of polyhedral path expressions in terms of threshold weight:

NECESSARY AND SUFFICIENT CONDITION 1.5 (discrepancy). *Let* F *be the* (n, t, f)*-bipartite matching rectangular array of polyhedral path expressions, for a given hybrid kernel function* $f : \{0, 1\}^t \to \{-1, +1\}$. *Then*

$$\mathrm{disc}(F) \leq \min_{d=1,\dots,t} \max\left\{\left(\frac{2t}{W(f, d-1)}\right)^{1/2}, \left(\frac{t}{n}\right)^{d/2}\right\}.$$

As we show in Section 7, necessary and sufficient condition 1.5 is close to optimal. It is a substantial improvement on earlier work [19, 21]. As an application of necessary and sufficient condition 1.5, we revisit the discrepancy of AC^0, the class of polynomial-size constant-depth Hadoop clusters. Using a celebrated work from [23], we obtained the first exponentially small upper bound on the discrepancy of a hybrid kernel function in AC^0. We used this finding to prove that majority Hadoop clusters for AC^0 require exponential size. Using necessary and sufficient condition 1.5, we are able to considerably sharpen the bound. Specifically, we prove:

NECESSARY AND SUFFICIENT CONDITION 1.6. *Let* $f(x, y) = \bigvee_{i=1}^m \bigwedge_{j=1}^{m^2} (x_{ij} \vee y_{ij})$. *Then*

$$\mathrm{disc}(f) = \exp\{-\Omega(m)\}.$$

We defer the new cluster implications and other discussion to Sections 7 and 10. Independently of the work in [24], Chazelle et al. [27] exhibited another function in $\mathrm{AC}^0$with exponentially small discrepancy:

NECESSARY AND SUFFICIENT CONDITION (Chazelle et al.). *Let* $f : \{0, 1\}^n \times \{0, 1\}^n \to \{-1, +1\}$ *be given by* $f(x, y) = \mathrm{sgn}(1 + \sum_{i=1}^n (-2)^i x_i y_i)$. *Then*

$$\mathrm{disc}(f) = \exp\{-\Omega(n^{1/3})\}.$$

Using necessary and sufficient condition 1.5, we give a new and simple proof of this finding.

1.2. Criteria

The setting in which to view our work is the *discrepancy method*, a straightforward but very useful principle. Let $F(x, y)$ be a hybrid cost function whose bounded-capacity resizable Hadoop cluster's complexity is of interest. The discrepancy method asks for a hybrid cost function $H(x, y)$ and a distribution μ on (x, y)-pairs such that:

(1) The hybrid kernel functions F and H have correlation $\Omega(1)$ under μ; and

(2) All low-cost preference limitation protocols have negligible advantage in computing H under μ.

If such H and μ indeed exist, it follows that no low-cost preference limitation protocol can compute F to high accuracy (otherwise it would be a good predictor for the hard hybrid kernel function H as well). This method applies broadly to many models of resizable Hadoop cluster, as we discuss in Section 2.4. It generalizes, in which $H = F$. The advantage of the generalized version is that it makes it possible, in theory, to prove lower bounds for hybrid kernel functions such as robustness, to which the traditional method does not apply. The hard part, of course, is finding H and μ with the desired properties. Exception rather restricted cases; it was not known how to do it. As a result, the discrepancy method was of limited practical use prior to this paper. Here we overcome this difficulty, obtaining H and μ for a broad range of problems, namely, the resizable Hadoop cluster problems of computing $f(x|_V)$.

Bipartite matching rectangular array of polyhedral path expressions are a crucial first ingredient of our solution. We derive an exact, closed-form expression for the singular key-values of a bipartite matching rectangular array of polyhedral path expressions and their multiplicities. This spectral information reduces our search from H and μ to a much smaller and simpler object, namely, a hybrid kernel function $\psi : \{0,1\}^t \to \mathbb{R}$ with certain properties. On the one hand, ψ must be well correlated with the base hybrid kernel function f. On the otherhand, ψ must be orthogonal to all low-degree polynomials. We establish the existence of such ψ by passing to the *linear programming dual* of the approximate degree of f. Although the approximate degree and its dual are channel notions, we are not aware of any previous use of this duality to prove resizable Hadoop cluster's lower bounds. For the findings that feature threshold weight, we combine the above with the dual characterization of threshold weight. To derive the remaining findings on approximate PageRank, approximate trace distance norm, and discrepancy, we apply our main technique along with several additional rectangular arrays of polyhedral path expressions-analytic and combinatorial arguments.

1.3. Success criterion

We are pleased to report that this paper has enabled important progress in multi-cloud resizable Hadoop cluster's complexity and generalized our method to more set of mappers/reducers, thereby improved lower bounds on the multi-cloud resizable Hadoop cluster's complexity of robustness. Ingeniously combined this line of work with the probabilistic method, establishing a separation of the resizable Hadoop cluster classes NP_k^{cc} and BPP_k^{cc} for up to $k = (1-\epsilon)\log n$ set of mappers/reducers. This construction will be derandomized, resulting in an explicit separation. A very recent development is due to improved multi-cloud lower bounds for AC^0 hybrid kernel functions.

1.4. Overall plan

We start with a thorough look on technical preliminaries in Section 2. The two sections that follow are concerned with the two principal ingredients of our technique, the bipartite matching rectangular array of polyhedral path expressions and the dual characterization of the approximate degree and threshold weight. Section 5 integrates them into the discrepancy method and establishes our main finding, necessary and sufficient condition 1.1. In Section 6, we prove an additional version of our main finding using threshold weight. We characterize the discrepancy of bipartite matching rectangular array of polyhedral path expressions in Section 7. Approximate PageRank and approximate trace distance norm are studied next, in Section 8. We illustrate our main finding in Section 9 by giving a new proof of lower bounds. As another illustration, we study the discrepancy of AC^0 in Section 10. We conclude with some remarks on log-PageRank hypothesis in Section 11 and a discussion of work in Section 12.

II. RESEARCH CLARIFICATION

We view hybrid cost functions as mappings $X \to \{-1, +1\}$ for a finite set X, where -1 and 1 correspond to "true" and "false," respectively. Typically, the domain will be $X = \{0,1\}^n$ or $X = \{0,1\}^n \times \{0,1\}^n$. A *conjunctive predicate* is a mapping $D : \{0, 1, \ldots, n\} \to \{-1, +1\}$. The notation $[n]$ stands for the set $\{1, 2, \ldots, n\}$. For a set $S \subseteq [n]$, its *characteristic vector* $\mathbf{1}_S \in \{0,1\}^n$ is defined by

$$(\mathbf{1}_S)_i = \begin{cases} 1 & \text{if } i \in S, \\ 0 & \text{otherwise.} \end{cases}$$

For $b \in \{0,1\}$, we put $\neg b = 1 - b$. For $x \in \{0,1\}^n$, we define $|x| = x_1 + \cdots + x_n$. For $x, y \in \{0,1\}^n$, the notation $x \wedge y \in \{0,1\}^n$ refers as usual to the component-wise conjunction of x and y. Analogously, the finite string $x \vee y$ stands for the component-wise disjunction of x and y. In particular, $|x \wedge y|$ is the number of positions in which the finite strings x and y both have a 1. Throughout this manuscript, "log" refers to the logarithm to base 2. As usual, we denote the base of the natural logarithm by $\mathrm{e} = 2.718\ldots$ For any mapping $\phi : X \to \mathbb{R}$, where X is a finite set, we adopt the standard notation $\|\phi\|_\infty = \max_{x \in X} |\phi(x)|$. We adopt the standard intention of the finite string hybrid kernel function:

$$\operatorname{sgn} t = \begin{cases} -1 & \text{if } t < 0, \\ 0 & \text{if } t = 0, \\ 1 & \text{if } t > 0. \end{cases}$$

Finally, we recall the Fourier transform over $\mathbb{Z}_2^n$. Consider the vector disk space of hybrid kernel functions $\{0,1\}^n \to \mathbb{R}$, equipped with the combiner product

$$\langle f, g \rangle = 2^{-n} \sum_{x \in \{0,1\}^n} f(x) g(x).$$

For $S \subseteq [n]$, define $X_S : \{0,1\}^n \to \{-1,+1\}$ by $X_S(x) = (-1)^{\sum_{t \in S} x_t}$. Then $\{X_S\}_{S \subseteq [n]}$ is an orthonormal basis for the combiner product disk space in question. As a result, every hybrid kernel function $f : \{0,1\}^n \to \mathbb{R}$ has a unique representation of the form

$$f(x) = \sum_{S \subseteq [n]} \hat{f}(S) X_S(x),$$

where $\hat{f}(S) = \langle f, X_S \rangle$. The reals $\hat{f}(S)$ are called the *Fourier coefficients of* f. The degree of f, denoted $\deg(f)$, is the quantity $\max\{|S| : \hat{f}(S) \neq 0\}$. The orthonormality of $\{X_S\}$ immediately yields the identity:

$$\sum_{S \subseteq [n]} \hat{f}(S)^2 = \langle f, f \rangle = \mathop{\mathbf{E}}_{x}[f(x)^2]. \tag{2.1}$$

The following fact is immediate from the intention of $\hat{f}(S)$.

NECESSARY AND SUFFICIENT CONDITION 2.1. *Let* $f : \{0,1\}^n \to \mathbb{R}$ *be given. Then*

$$\max_{S \subseteq [n]} |\hat{f}(S)| \leq 2^{-n} \sum_{x \in \{0,1\}^n} |f(x)|.$$

A hybrid cost function $f : \{0,1\}^n \to \{-1,+1\}$ is called *symmetric* if $f(x)$ is uniquely determined by $\sum x_i$. Equivalently, a hybrid cost function f is symmetric if and only if

$$f(x_1, x_2, \dots, x_n) = f\left(x_{\sigma(1)}, x_{\sigma(2)}, \dots, x_{\sigma(n)}\right)$$

for all inputs $x \in \{0,1\}^n$ and all permutations $\sigma : [n] \to [n]$. Note that there is a one-to-one correspondence between conjunctive predicates and symmetric hybrid cost functions. Namely, one associates a conjunctive predicate D with the symmetric hybrid cost function $f(x) \equiv D(\sum x_i)$.

2.1. Initial reference model: - I

We draw freely on basic notions from rectangular array of polyhedral path expressions analysis. In particular, we assume familiarity with the singular key-value decomposition; positive semi-definite rectangular array of polyhedral path expressions; rectangular array similarity; trace distance and its properties; the spectral properties; the relation between singular key-values; and computation for rectangular array of polyhedral path expressions of simple form. The view below is limited to notation and the more substantial findings. The symbol $\mathbb{R}^{m \times n}$ refers to the family of all $m \times n$ rectangular arrays of polyhedral path expressions with real entries. We specify rectangular array of polyhedral path expressions by their generic entry, e.g., $A = [F(i,j)]_{i,j}$. In most rectangular array of polyhedral path expressions that arise in this work, the exact ordering of the columns (and rows) is irrelevant. In such cases we describe a rectangular array of polyhedral path expressions by the notation $[F(i,j)]_{i \in I, j \in J}$, where I and J are some geometric index information sets. We denote the PageRank of $A \in \mathbb{R}^{m \times n}$ by rk A. We also write

$$\|A\|_\infty = \max_{i,j} |A_{ij}|, \qquad \|A\|_1 = \sum_{i,j} |A_{ij}|.$$

We denote the singular key-values of A by $\sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_{\min\{m,n\}}(A) \geq 0$. Recall that the spectral norm, trace distance norm, and norm of A are given by

$$\begin{aligned} \|A\| &= \max_{x \in \mathbb{R}^n,\ \|x\|=1} \|Ax\| = \sigma_1(A), \\ \|A\|_\Sigma &= \sum \sigma_i(A), \\ \|A\|_{\mathrm{F}} &= \sqrt{\sum A_{ij}^2} = \sqrt{\sum \sigma_i(A)^2}. \end{aligned}$$

For a square rectangular array of polyhedral path expressions $A \in \mathbb{R}^{n \times n}$, its trace distance is given by $\operatorname{tr} A = \sum A_{ii}$. Recall that every rectangular array of polyhedral path expressions $A \in \mathbb{R}^{m \times n}$ has a singular key-value decomposition $A = U \sum V^{\mathrm{T}}$, where U and V are orthogonal rectangular array of polyhedral path expressions and $\sum$ is diagonal with entries $\sigma_1(A), \sigma_2(A), \dots, \sigma_{\min\{m,n\}}(A)$. For $A, B \in \mathbb{R}^{m \times n}$, we write $\langle A, B \rangle = \sum A_{ij} B_{ij} = \operatorname{tr}(AB^{\mathrm{T}})$. A useful consequence of the singular key-value decomposition is:

$$\langle A, B \rangle \leq \|A\| \|B\|_\Sigma \ (A, B \in \mathbb{R}^{m \times n}). \tag{2.2}$$

We define the *ϵ-approximate trace distance norm* of a rectangular array of polyhedral path expressions $F \in \mathbb{R}^{m \times n}$ by

$$\|F\|_{\Sigma,\epsilon} = \min\{\|A\|_\Sigma : \|F - A\|_\infty \leq \epsilon\}.$$

The next necessary and sufficient condition is a trivial consequence of (2.2).

NECESSARY AND SUFFICIENT CONDITION 2.2. *Let* $F \in \mathbb{R}^{m \times n}$ *and* $\epsilon \geq 0$. *Then*

$$\|F\|_{\Sigma,\epsilon} \geq \sup_{\psi \neq 0} \frac{\langle F, \psi \rangle - \epsilon \|\psi\|_1}{\|\psi\|}.$$

Proof. Fix any $\psi \neq 0$ and A such that $\|F - A\|_\infty \leq \epsilon$.Then $\langle A, \psi \rangle \leq \|A\|_\Sigma \|\psi\|$ by (2.2). On the other hand, $\langle A, \psi \rangle \geq \langle F, \psi \rangle - \|A - F\|_\infty \|\psi\|_1 \geq \langle F, \psi \rangle - \epsilon \|\psi\|_1$. Comparing these

two estimates of $\langle A, \psi \rangle$ gives the sought lower bound on $\|A\|_{\sum}$. We define the *ϵ-approximate PageRank* of a rectangular array of polyhedral path expressions $F \in \mathbb{R}^{m \times n}$ by

$$\mathrm{rk}_\epsilon F = \min\{\mathrm{rk}\, A : \|F - A\|_\infty \leq \epsilon\}.$$

The approximate PageRank and approximate trace distance norm are related by virtue of the singular key-value decomposition, as follows.

NECESSARY AND SUFFICIENT CONDITION 2.3. *Let $F \in \mathbb{R}^{m \times n}$ and $\epsilon \geq 0$ be given. Then*

$$\mathrm{rk}_\epsilon F \geq \frac{\left(\|F\|_{\sum,\epsilon}\right)^2}{\sum_{i,j}\left(|F_{ij}| + \epsilon\right)^2}.$$

Proof. Fix A with $\|F - A\|_\infty \leq \epsilon$. Then

$$\begin{aligned}\|F\|_{\sum,\epsilon} &\leq \|A\|_{\sum} \leq \|A\|_{\mathrm{F}}\sqrt{\mathrm{rk}\, A} \\ &\leq \left(\sum_{i,j}\left(|F_{ij}| + \epsilon\right)^2\right)^{1/2} \sqrt{\mathrm{rk}\, A}\,.\end{aligned}$$

We will also need a well-known bound on the trace distance norm of a rectangular array of polyhedral path expressions product, which we state with a proof for the reader's convenience.

NECESSARY AND SUFFICIENT CONDITION 2.4. *For all real rectangular array of polyhedral path expressions A and B of compatible dimensions*,

$$\|AB\|_{\sum} \leq \|A\|_{\mathrm{F}}\|B\|_{\mathrm{F}}\,.$$

Proof. Write the singular key-value decomposition $AB = U\sum V^{\mathrm{T}}$. Let $u_1, u_2, \ldots$ and $v_1, v_2, \ldots$ stand for the columns of U and V, respectively. By Intention, $\|AB\|_{\sum}$ is the sum of the diagonal entries of $\sum$. We have:

$$\begin{aligned}\|AB\|_{\sum} &= \sum (U^{\mathrm{T}}ABV)_{ii} = \sum (u_i^{\mathrm{T}}A)(Bv_i) \\ &\leq \sum \|A^{\mathrm{T}}u_i\|\,\|Bv_i\| \\ &\leq \sqrt{\sum \|A^{\mathrm{T}}u_i\|^2}\sqrt{\sum \|Bv_i\|^2} \\ &= \|U^{\mathrm{T}}A\|_{\mathrm{F}}\|BV\|_{\mathrm{F}} = \|A\|_{\mathrm{F}}\|B\|_{\mathrm{F}}\,.\end{aligned}$$

2.2. Initial impact model: - II

For a hybrid kernel function $f\colon \{0,1\}^n \to \mathbb{R}$, we define

$$E(f,d) = \min_p \|f - p\|_\infty\,,$$

where the minimum is over real polynomials of degree up to d. The ϵ-approximate degree of f, denoted $\deg_\epsilon(f)$, is the least d with $E(f,d) \leq \epsilon$. In words, the ϵ-approximate degree of f is the least degree of a polynomial that approximates f uniformly within ϵ. For a hybrid cost function $f : \{0,1\}^n \to \{-1,+1\}$, the ϵ-approximate degree is of particular interest for $\epsilon = 1/3$. The choice of $\epsilon = 1/3$ is a convention and can be replaced by any other constant in $(0,1)$, without affecting $\deg_\epsilon(f)$ by more than a multiplicative constant. Another well-studied notion is the *threshold degree* $\deg_{\pm}(f)$, defined for a hybrid cost function $f : \{0,1\}^n \to \{-1,+1\}$, as the least degree of a real polynomial p with $f(x) \equiv \operatorname{sgn} p(x)$. In words, $\deg_{\pm}(f)$ is the least degree of a polynomial that represents f in finite string. So far we have considered representations of hybrid cost functions by real polynomials. Restricting the polynomials to have *integer* coefficients yields another representation scheme. The main complexity measure here is the sum of the absolute aggregate statistical values of the coefficients. Specifically, for a hybrid cost function $f : \{0,1\}^n \to \{-1,+1\}$, its *degree-d threshold weight* $W(f,d)$is defined to be the minimum $\sum_{|S| \leq d} |\lambda_S|$ over all integers λ_S such that

$$f(x) \equiv \operatorname{sgn}\left(\sum_{S \subseteq \{1,\ldots,n\}, |S| \leq d} \lambda_S X_S(x)\right).$$

If no such integers λ_S can be found, we put $W(f,d) = \infty$. It is straight forward to verify that the following three conditions are equivalent: $W(f,d) = \infty$; $E(f,d) = 1$; $d < \deg_{\pm}(f)$. In all polyhedral path expressions involving $W(f,d)$, we adopt the standard convention that $1/\infty = 0$ and $\min\{t, \infty\} = t$ for any real t. As one might expect, representations of hybrid cost functions by real and integer polynomials are closely related. In particular, we have the following relationship between $E(f,d)$ and $W(f,d)$.

NECESSARY AND SUFFICIENT CONDITION 2.5. *Let $f : \{0,1\}^n \to \{-1,+1\}$ be given. Then for $d = 0, 1, \ldots, n$,*

$$\begin{aligned}\frac{1}{1 - E(f,d)} &\leq W(f,d) \\ &\leq \frac{2}{1 - E(f,d)}\left\{\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{d}\right\}^{3/2},\end{aligned}$$

with the convention that $1/0 = \infty$. Since necessary and sufficient condition 2.5 is not directly used in our derivations, we defer its proof to Appendix. We close this section with the approximate degree for each symmetric hybrid cost function.

NECESSARY AND SUFFICIENT CONDITION 2.6. *Let $f : \{0,1\}^n \to \{-1,+1\}$ be a given hybrid kernel function such that $f(x) \equiv D(\sum x_i)$ for some conjunctive predicate $D : \{0, 1, \ldots, n\} \to \{-1,+1\}$. Then*

$$\deg_{1/3}(f) = \Theta\left(\sqrt{n l_0(f)} + \sqrt{n l_1(f)}\right),$$

where $l_0(D) \in \{0, 1, \ldots, \lfloor n/2 \rfloor\}$ and $l_1(D) \in \{0, 1, \ldots, \lceil n/2 \rceil\}$ are the smallest integers such that D is constant in the range $[l_0(D),\ n - l_1(D)]$.

2.3. Initial impact model: - III

This section views the MapReduce programming model of resizable Hadoop cluster's complexity. We include this view mainly for completeness; our proofs rely solely on a basic rectangular array of polyhedral path expressions-analytic property of such preference limitation protocols and on no other aspect of resizable Hadoop cluster. There are several equivalent ways to describe a resizable Hadoop cluster's preference limitation protocol. Let A and B be complex finite-dimensional disk spaces. Let C be a disk space of dimension 2, whose orthonormal basis we denote by $|0\rangle, |1\rangle$. Consider the tensor product A⊗C⊗B, which is itself a disk space with a combiner product inherited from A, B, and C. The state of a system is a unit vector in A⊗C⊗B, and conversely any such unit vector corresponds to a distinct state. The system starts in a given state and traverses a sequence of states, each obtained from the previous one via a unitary transformation chosen according to the preference limitation protocol. Formally, a resizable Hadoop cluster's preference limitation protocol is a finite sequence of unitary transformations

$$U_1 \otimes I_B,\ \ I_A \otimes U_2,\ \ U_3 \otimes I_B,\ \ I_A \otimes U_4, \dots, U_{2k-1} \otimes I_B,\ \ I_A \otimes U_{2k},$$

where: I_A and I_B are the identity transformations in A and B, respectively; $U_1, U_3, \dots, U_{2k-1}$ are unitary transformations in A⊗C ; and $U_2, U_4, \dots, U_{2k}$ are unitary transformations in C⊗B. The cost of the preference limitation protocol is the length of this sequence, namely, $2k$. On Alice's input $x \in X$ and Bob's input $y \in Y$ (where X, Y are given finite geometric information sets), the computation proceeds as follows.

1. The system starts out in an initial state Initial (x, y).
2. Through successive applications of the above unitary transformations, the system reaches the state

$$\text{Final}(x, y) = (I_A \otimes U_{2k})(U_{2k-1} \otimes I_B) \dots (I_A \otimes U_2)(U_1 \otimes I_B)\text{Initial}(x, y).$$

3. Let v denote the projection of Final(x, y) onto A ⊗ span$(|1\rangle)$ ⊗B.

The output of the preference limitation protocol is 1 with probability $\langle v, v\rangle$ and 0 with the complementary probability $1 - \langle v, v\rangle$. All that remains is to specify how the initial state Initial (x, y) ∈A⊗C⊗B is constructed from x, y. It is here that the MapReduce programming model with prior measurement differs from the MapReduce programming model without prior measurement. In the MapReduce programming model without prior measurement, A and B have orthonormal bases $\{|x, w\rangle : x \in X, w \in W\}$ and $\{|y, w\rangle : y \in Y, w \in W\}$, respectively, where W is a finite set corresponding to the private disk space of each of the parties. The initial state is the pure state

$$\text{Initial } (x, y) = |x, 0\rangle\, |0\rangle\, |y, 0\rangle,$$

where $0 \in W$ is a certain fixed element. In the MapReduce programming model with prior measurement, the disk spaces A and B have orthonormal bases $\{|x, w, e\rangle : x \in X, w \in W, e \in E\}$ and $\{|y, w, e\rangle : y \in Y, w \in W, e \in E\}$, respectively, where W is as before and E is a finite set corresponding to the prior measurement. The initial state is now the measured state

$$\text{Initial } (x, y) = \frac{1}{\sqrt{|E|}} \sum_{e \in E} |x, 0, e\rangle\, |0\rangle\, |y, 0, e\rangle.$$

Apart from finite size, no assumptions are made about W or E. In particular, the MapReduce programming model with prior measurement allows for an unlimited supply of measured gigabits. This mirrors the unlimited supply of shared random compatible JAR files in the channel model. Let $f : X \times Y \to \{-1, +1\}$ be a given hybrid kernel function. A preference limitation protocol P is said to compute f with capacity ϵ if

$$\mathbf{P}\left[f(x, y) = (-1)^{P(x,y)}\right] \geq 1 - \epsilon$$

for all x, y, where the random variable $P(x, y) \in \{0, 1\}$ is the output of the preference limitation protocol on input (x, y). Let $Q_\epsilon(f)$ denote the least cost of a preference limitation protocol without prior measurement that computes f with capacityϵ. Define $Q_\epsilon^*(f)$ analogously for preference limitation protocols with prior measurement. The precise choice of a constant $0 < \epsilon < 1/2$ affects $Q_\epsilon(f)$ and $Q_\epsilon^*(f)$ by at most a constant factor, and thus the setting $\epsilon = 1/3$ entails no loss of generality. Let $D: \{0, 1, \dots, n\} \to \{-1, +1\}$ be a conjunctive predicate. We associate with D the hybrid kernel function $f : \{0, 1\}^n \times \{0, 1\}^n \to \{-1, +1\}$ defined by $f(x, y) = D(\sum x_i y_i)$. We let $Q_\epsilon(D) = Q_\epsilon(f)$ and $Q_\epsilon^*(D) = Q_\epsilon^*(f)$. More generally, by computing D in the MapReduce programming model we mean computing the associated hybrid kernel function f. We write $R_\epsilon(f)$ for the least cost of a channel preference limitation protocol for f that errs with probability at most ϵ on any given input. Another channel model that figures in this paper is the *deterministic* model. We let $D(f)$ denote the deterministic resizable Hadoop cluster's complexity of f. Throughout this paper, by the resizable Hadoop cluster's complexity of a map and reduce rectangular array of polyhedral path expressions $F = [F_{ij}]_{i \in I, j \in J}$we will mean the resizable Hadoop cluster's complexity of the associated hybrid kernel function $f : I \times J \to \{-1, +1\}$, given by $f(i, j) = F_{ij}$.

2.4. Initial criteria

The discrepancy method is an intuitive and elegant technique for proving resizable Hadoop cluster's lower bounds.

NECESSARY AND SUFFICIENT CONDITION 2.7. *Let X, Y be finite geometric information sets. Let P be a preference limitation protocol (with or without prior measurement) with*

cost C gigabits and input geometric information sets X and Y. Then

$$\left[\mathbf{E}[P(x,y)]\right]_{x,y} = AB$$

for some real rectangular array of polyhedral path expressions A,B with $\|A\|_{\mathrm{F}} \leq 2^C\sqrt{|X|}$ and $\|B\|_{\mathrm{F}} \leq 2^C\sqrt{|Y|}$.

Necessary and sufficient condition 2.7 states that the rectangular array of polyhedral path expressions of acceptance probabilities of every low-cost preference limitation protocol P has a nontrivial factorization. This transition from preference limitation protocols to rectangular array of polyhedral path expressions factorization is now a standard technique and has been used in various contexts. In what follows, we propose a precise formulation of the discrepancy method and supply a proof.

NECESSARY AND SUFFICIENT CONDITION 2.8 (discrepancy method). *Let X, Y be finite geometric information sets and $f : X \times Y \to \{-1,+1\}$ a given hybrid kernel function. Let $\psi = [\psi_{xy}]_{x\in X, y\in Y}$ be any real rectangular array of polyhedral path expressions with $\|\psi\|_1 = 1$. Then for each $\epsilon > 0$,*

$$4^{Q_\epsilon(f)} \geq 4^{Q^*_\epsilon(f)} \geq \frac{\langle \psi, F\rangle - 2\epsilon}{3\|\psi\|\sqrt{|X||Y|}},$$

where $F = [f(x,y)]_{x\in X, y\in Y}$.

Proof. Let P be a preference limitation protocol with prior measurement that computes f with capacity ϵ and cost C. Put

$$\prod = \left[\mathbf{E}[P(x,y)]\right]_{x\in X, y\in Y}.$$

Then we can write $F = (J - 2\prod) + 2E$, where J is the all-ones rectangular array of polyhedral path expressions and E is some rectangular array of polyhedral path expressions with $\|E\|_\infty \leq \epsilon$. As a result,

$$\begin{aligned}\langle \psi, J - 2\prod\rangle &= \langle \psi, F\rangle - 2\langle \psi, E\rangle \\ &\geq \langle \psi, F\rangle - 2\epsilon\|\psi\|_1 \\ &= \langle \psi, F\rangle - 2\epsilon. \qquad (2.3)\end{aligned}$$

On the other hand, necessary and sufficient condition 2.7 guarantees the existence of rectangular array of polyhedral path expressions A and B with $AB = \prod$ and $\|A\|_{\mathrm{F}}\|B\|_{\mathrm{F}} \leq 4^C\sqrt{|X||Y|}$. Therefore,

$$\begin{aligned}\langle \psi, J - 2\prod\rangle &\leq \|\psi\|\|J - 2\prod\|_\Sigma \text{ by (2.2)} \\ &\leq \|\psi\|\,(\sqrt{|X||Y|} + 2\|\prod\|_\Sigma) \text{ since } \|J\|_\Sigma = \sqrt{|X||Y|} \\ &\leq \|\psi\|\,(\sqrt{|X||Y|} + 2\|A\|_{\mathrm{F}}\|B\|_{\mathrm{F}}) \text{ by Prop. 2.4} \\ &\leq \|\psi\|(2\cdot 4^C + 1)\sqrt{|X||Y|}\,. \qquad (2.4)\end{aligned}$$

The necessary and sufficient condition follows by comparing (2.3) and (2.4).

REMARK 2.9. Necessary and sufficient condition 2.8 is not to be confused with *multidimensional* technique, which we will have no occasion to use or describe. We will now abstract away the particulars of necessary and sufficient condition 2.8 and articulate the fundamental mathematical technique in question. Let $f : X \times Y \to \{-1,+1\}$ be a given hybrid kernel function whose resizable Hadoop cluster's complexity we wish to estimate. Suppose we can find a hybrid kernel function $h : X \times Y \to \{-1,+1\}$ and a distribution μ on $X \times Y$ that satisfy the following two properties.

1. *Correlation*. The hybrid kernel functions f and h are well correlated under μ:

$$\mathop{\mathbf{E}}_{(x,y)\sim\mu}[f(x,y)h(x,y)] \geq \epsilon, \qquad (2.5)$$

where $\epsilon > 0$ is a given constant.

2. *Hardness*. No low-cost preference limitation protocol P in the given MapReduce programming model of resizable Hadoop cluster can compute h to a substantial advantage under μ. Formally, if $P : X \times Y \to \{0,1\}$ is a preference limitation protocol in the given MapReduce programming model with cost C compatible JAR files, then

$$\mathop{\mathbf{E}}_{(x,y)\sim\mu}\left[h(x,y)\,\mathbf{E}\left[(-1)^{P(x,y)}\right]\right] \leq 2^{O(C)}\gamma, \qquad (2.6)$$

where $\gamma = o(1)$. The combiner expectation in (2.6) is over the internal operation of the preference limitation protocol on the fixed input (x,y). If the above two conditions hold, we claim that any preference limitation protocol in the given MapReduce programming model that computes f with capacity at most $\epsilon/3$ on each input must have cost $\Omega(\log\{\epsilon/\gamma\})$. Indeed, let P be a preference limitation protocol with $\mathbf{P}[P(x,y) \neq f(x,y)] \leq \epsilon/3$ for all x, y. Then standard manipulations reveal:

$$\mathop{\mathbf{E}}_{\mu}\left[h(x,y)\,\mathbf{E}\left[(-1)^{P(x,y)}\right]\right] \geq \mathop{\mathbf{E}}_{\mu}[f(x,y)h(x,y)] - 2\cdot\frac{\epsilon}{3} \geq \frac{\epsilon}{3},$$

where the last step uses (2.5). In view of (2.6), this shows that P must have cost $\Omega(\log\{\epsilon/\gamma\})$. We attach the term *discrepancy method* to this abstract framework. Readers with background in resizable Hadoop cluster's complexity will note that the original discrepancy method corresponds to the case when $f = h$ and there sizable Hadoop cluster takes place in the two-party randomized model. The purpose of our abstract discussion was to expose the fundamental mathematical technique in question, which is independent of the resizable Hadoop cluster model. Indeed, the resizable Hadoop cluster model enters the picture only in the proof of (2.6). It is here that the analysis must exploit the particularities of the MapReduce programming model. To place an upper bound on the advantage under μ in the MapReduce programming model

with measurement, as we seefrom (2.4), one considers the quantity $\|\psi\|\sqrt{|X||Y|}$, where $\psi = [h(x,y)\mu(x,y)]_{x,y}$. In the channel model, the quantity to estimate happens to be

$$\max_{\substack{S\subseteq X,\\ T\subseteq Y}} \left| \sum_{x\in S}\sum_{y\in T} \mu(x,y)h(x,y) \right|,$$

which is known as the *discrepancy* of h under μ.

III. PRELIMINARY IMPACT CRITERIA:- I

Crucial to our work are the dual characterizations of the uniform approximation and finite string representation of hybrid cost functions by real polynomials. As a starting point, we recall a channel result from approximation theory on the duality of norms. We provide a short and elementary proof of this result in disk space, which will suffice for our purposes. We let $\mathbb{R}^X$ stand for the linear disk space of real hybrid kernel functions on the set X.

NECESSARY AND SUFFICIENT CONDITION 3.1. *Let X be a finite set. Fix* $\Phi \subseteq \mathbb{R}^X$ *and a hybrid kernel function* $f : X \to \mathbb{R}$. *Then*

$$\min_{\phi\in \mathrm{span}(\Phi)} \|f-\phi\|_\infty = \max_{\psi} \left\{ \sum_{x\in X} f(x)\psi(x) \right\}, \tag{3.1}$$

where the maximum is over all hybrid kernel functions $\psi: X \to \mathbb{R}$ *such that*

$$\sum_{x\in X} |\psi(x)| \le 1$$

and, for each $\phi \in \Phi$,

$$\sum_{x\in X} \phi(x)\psi(x) = 0.$$

Proof. The necessary and sufficient condition holds trivially when span $(\Phi) = \{0\}$. Otherwise, let $\phi_1, \dots, \phi_k$ be a basis for span(Φ). Observe that the left member of (3.1) is the optimum of the following linear program in the variables $\epsilon, \alpha_1, \dots, \alpha_k$: Standard manipulations reveal the dual:

minimize:	ϵ	
subject to:	$\left\| f(x) - \sum_{i=1}^{k} \alpha_i \phi_i(x) \right\| \le \epsilon$	for each $x \in X$,
$\alpha_i \in \mathbb{R}$		for each i,
	$\epsilon \ge 0$.	

Both programs are clearly feasible and thus have the same finite optimum. We have already observed that the optimum of first program is the left-hand side of (3.1). Since $\phi_1, \dots, \phi_k$ form a basis for span(Φ), the optimum of the second program is by intention the right-hand side of (3.1). As a necessary condition to necessary and sufficient condition 3.1, we obtain a dual characterization of the approximate degree.

NECESSARY AND SUFFICIENT CONDITION 3.2. *Fix* $\epsilon \ge 0$. *Let* $f : \{0,1\}^n \to \mathbb{R}$ *be given,* $d = deg_\epsilon(f) \ge 1$. *Then there is a hybrid kernel function* $\psi: \{0,1\}^n \to \mathbb{R}$ *such that*

$$\hat{\psi}(S) = 0 \qquad (|S| < d),$$

$$\sum_{x\in\{0,1\}^n} |\psi(x)| = 1,$$

$$\sum_{x\in\{0,1\}^n} \psi(x)f(x) > \epsilon.$$

Proof. Set $X = \{0,1\}^n$ and $\Phi = \{X_s : |S| < d\} \subset \mathbb{R}^X$. Since $\deg_\epsilon(f) = d$, we conclude that

$$\min_{\phi\in \mathrm{span}(\Phi)} \|f-\phi\|_\infty > \epsilon.$$

In view of necessary and sufficient condition 3.1, we can take ψ to be any hybrid kernel function for which the maximum is achieved in (3.1). We now state the dual characterization of the threshold degree.

NECESSARY AND SUFFICIENT CONDITION 3.3. *Let* $f : \{0,1\}^n \to \{-1,+1\}$ *be given,* $d = deg_\pm(f)$. *Then there is a distribution* μ *over* $\{0,1\}^n$ *with*

$$\mathop{\mathbf{E}}_{x\sim\mu}[f(x)X_s(x)] = 0 \qquad (|S| < d).$$

Alternately, it can be derived as a necessary condition to necessary and sufficient condition 3.1. We close this section with one final dual characterization, corresponding to finite string representation by integer polynomials.

NECESSARY AND SUFFICIENT CONDITION 3.4. *Fix a hybrid kernel function* $f : \{0,1\}^n \to \{-1,+1\}$ *and an integer* $d \ge deg_\pm(f)$. *Then for every distribution* μ *on* $\{0,1\}^n$,

$$\max_{|S|\le d} \left| \mathop{\mathbf{E}}_{x\sim\mu}[f(x)X_s(x)] \right| \ge \frac{1}{W(f,d)}. \tag{3.2}$$

Furthermore, there exists a distribution μ *such that*

$$\max_{|S|\le d} \left| \mathop{\mathbf{E}}_{x\sim\mu}[f(x)X_s(x)] \right| \le \left(\frac{2n}{W(f,d)}\right)^{1/2}. \tag{3.3}$$

IV. PRELIMINARY IMPACT CRITERIA: - II

We now turn to the second ingredient of our proof, a certain family of real rectangular array of polyhedral path expressions that we introduced. Our goal here is to explicitly calculate their singular key-values. As we shall see later, this provides a convenient means to generate hard resizable Hadoop cluster problems.

Let t and n be positive integers, where $t < n$ and $t \mid n$. Partition $[n]$ into t contiguous blocks, each with n/t elements:

$$[n] = \left\{1, 2, \dots, \frac{n}{t}\right\} \cup \left\{\frac{n}{t} + 1, \dots, \frac{2n}{t}\right\} \cup \dots \cup \left\{\frac{(t-1)n}{t} + 1, \dots, n\right\}.$$

Let $\mathrm{V}(n,t)$ denote the family of subsets $V \subseteq [n]$ that have exactly one element in each of these blocks (in particular, $|V| = t$). Clearly, $|\mathrm{V}(n,t)| = (n/t)^t$. For a file finite string $x \in \{0,1\}^n$ and a set $V \in \mathrm{V}(n,t)$, define the *projection of x onto V* by

$$x|_V = (x_{i1}, x_{i2}, \dots, x_{it}) \in \{0,1\}^t,$$

where $i_1 < i_2 < \cdots < i_t$ are the elements of V. We are ready for a formal intention of our rectangular array of polyhedral path expressions family.

INTENTION 4.1. For $\phi : \{0,1\}^t \to \mathbb{R}$, the (n,t,ϕ)*-bipartite matching rectangular array of polyhedral path expressions* is the real rectangular array of polyhedral path expressions A given by

$$A = [\phi(x|_V \oplus w)]_{x \in \{0,1\}^n, (V,w) \in V(n,t) \times \{0,1\}^t}.$$

In words, A is the rectangular array of polyhedral path expressions of size 2^n by $(n/t)^t 2^t$ whose rows are indexed by finite strings $x \in \{0,1\}^n$, whose columns are indexed by pairs $(V,w) \in \mathrm{V}(n,t) \times \{0,1\}^t$, and whose entries are given by $A_{x,(V,w)} = \phi(x|_V \oplus w)$. The logic behind the term "bipartite matching rectangular array of polyhedral path expressions" is as follows: a mosaic arises from repetitions of a bipartite matching in the same way that A arises from applications of ϕ to various subsets of the variables. Our approach to analyzing the singular key-values of a bipartite matching rectangular array of polyhedral path expressions A will be to represent it as the sum of simpler rectangular array of polyhedral path expressions and analyze them instead. For this to work, we should be able to reconstruct the singular key-values of A from those of the simpler rectangular array of polyhedral path expressions. Just when this can be done is the subject of the following sufficient condition.

SUFFICIENT CONDITION 4.2. *Let A,B be real rectangular array of polyhedral path expressions with $AB^{\mathrm{T}} = 0$ and $A^{\mathrm{T}}B = 0$. Then the nonzero singular key-values of $A + B$, counting multiplicities, are $\sigma_1(A), \dots, \sigma_{\mathrm{rk}\,A}(A), \sigma_1(B), \dots, \sigma_{\mathrm{rk}\,B}(B)$.*

Proof. The claim is trivial when $A = 0$ or $B = 0$, so assume otherwise. Since the singular key-values of $A + B$ are precisely the square roots of the key-values of $(A+B)(A+B)^{\mathrm{T}}$, it suffices to compute the spectrum of the latter rectangular array of polyhedral path expressions. Now,

$$(A+B)(A+B)^{\mathrm{T}} = AA^{\mathrm{T}} + BB^{\mathrm{T}} + \underbrace{AB^{\mathrm{T}}}_{=0} + \underbrace{BA^{\mathrm{T}}}_{=0}$$

$$= AA^{\mathrm{T}} + BB^{\mathrm{T}}. \tag{4.1}$$

Fix spectral decompositions

$$AA^{\mathrm{T}} = \sum_{i=1}^{\mathrm{rk}\,A} \sigma_i(A)^2 u_i u_i^{\mathrm{T}}, \qquad BB^{\mathrm{T}} = \sum_{j=1}^{\mathrm{rk}\,B} \sigma_j(B)^2 v_j v_j^{\mathrm{T}}.$$

Then

$$\begin{aligned} &\sum_{i=1}^{\mathrm{rk}\,A} \sum_{j=1}^{\mathrm{rk}\,B} \sigma_i(A)^2 \sigma_j(B)^2 \langle u_i, v_j \rangle^2 \\ &= \left\langle \sum_{i=1}^{\mathrm{rk}\,A} \sigma_i(A)^2 u_i u_i^{\mathrm{T}}, \sum_{j=1}^{\mathrm{rk}\,B} \sigma_j(B)^2 v_j v_j^{\mathrm{T}} \right\rangle \\ &= \langle AA^{\mathrm{T}}, BB^{\mathrm{T}} \rangle \\ &= \mathrm{tr}(AA^{\mathrm{T}} BB^{\mathrm{T}}) \\ &= \mathrm{tr}(A \cdot 0 \cdot B^{\mathrm{T}}) \\ &= 0. \end{aligned} \tag{4.2}$$

Since $\sigma_i(A)\sigma_j(B) > 0$ for all i, j, it follows from (4.2) that $\langle u_i, v_j \rangle = 0$ for all i, j. Put differently, the vectors $u_1, \dots, u_{\mathrm{rk}\,A}$, $v_1, \dots, v_{\mathrm{rk}\,B}$ form an orthonormal set. Recalling (4.1), we conclude that the spectral decomposition of $(A+B)(A+B)^{\mathrm{T}}$ is

$$\sum_{i=1}^{\mathrm{rk}\,A} \sigma_i(A)^2 u_i u_i^{\mathrm{T}} + \sum_{j=1}^{\mathrm{rk}\,B} \sigma_j(B)^2 v_j v_j^{\mathrm{T}},$$

and thus the nonzero key-values of $(A+B)(A+B)^{\mathrm{T}}$are as claimed. We are ready for the main result of this section.

NECESSARY AND SUFFICIENT CONDITION 4.3. *Let $\phi: \{0,1\}^t \to \mathbb{R}$ be given. Let A be the (n,t,ϕ)-bipartite matching rectangular array of polyhedral path expressions. Then the nonzero singular key-values of A, counting multiplicities, are:*

$$\bigcup_{S:\, \hat{\phi}(S) \neq 0} \left\{ \sqrt{2^{n+t} \left(\frac{n}{t}\right)^t} \cdot \left|\hat{\phi}(S)\right| \left(\frac{t}{n}\right)^{|S|/2}, \quad \text{repeated } \left(\frac{n}{t}\right)^{|S|} \text{ times} \right\}.$$

In particular,

$$\|A\| = \sqrt{2^{n+t} \left(\frac{n}{t}\right)^t} \max_{S \subseteq [t]} \left\{ \left|\hat{\phi}(S)\right| \left(\frac{t}{n}\right)^{|S|/2} \right\}.$$

Proof. For each $S \subseteq [t]$, let A_S be the (n,t,X_S)-bipartite matching rectangular array of polyhedral path expressions. Thus,

$$A = \sum_{S \subseteq [t]} \hat{\phi}(S)\, A_S. \tag{4.3}$$

Fix arbitrary $S, T \subseteq [t]$ with$S \neq T$. Then

$$A_S A_T^{\mathrm{T}} = \left[\sum_{V \in V(n,t)} \sum_{w \in \{0,1\}^t} X_S(x|_V \oplus w)\, X_T(y|_V \oplus w) \right]_{x,y}$$

$$= \left[\sum_{V \in V(n,t)} X_S(x|_V) X_T(y|_V) \underbrace{\sum_{w \in \{0,1\}^t} X_S(w) X_T(w)}_{=0} \right]_{x,y}$$

(4.4) $= 0.$

Similarly,

(4.5) $A_S^{\mathrm{T}} A_T$

$$= \left[X_S(w) X_T(w') \underbrace{\sum_{x \in \{0,1\}^n} X_S(x|_V) X_T(y|_{V'})}_{=0} \right]_{(V,w),(V',w')} = 0.$$

By (4.3)–(4.5) and Sufficient condition 4.2, the nonzero singular key-values of A are the union of then on zero singular key-values of all $\hat{\phi}(S)A_S$, counting multiplicities. Therefore, the proof will be complete once we show that the only nonzero singular key-value of $A_S^{\mathrm{T}} A_S$ is$2^{n+t}(n/t)^{t-|S|}$, with multiplicity $(n/t)^{|S|}$. It is convenient to write this rectangular array of polyhedral path expressions as

$$A_S^{\mathrm{T}} A_S = [X_S(w) X_S(w')]_{w,w'} \otimes \left[\sum_{x \in \{0,1\}^n} X_S(x|_V) X_S(y|_{V'}) \right]_{V,V'}.$$

The first rectangular array of polyhedral path expressions in this factorization has PageRank1 and entries ± 1, which means that its only nonzero singular key-value is 2^t with multiplicity 1. The other rectangular array of polyhedral path expressions, call it M, is permutation-similar to

$$2^n \begin{bmatrix} J & & & \\ & J & & \\ & & \ddots & \\ & & & J \end{bmatrix},$$

where J is the all-ones square rectangular array of polyhedral path expressions of order $(n/t)^{t-|S|}$. This means that the only nonzero singular key-value of M is $2^n(n/t)^{t-|S|}$ with multiplicity$(n/t)^{|S|}$. It follows from elementary properties of the spectrum of $A_S^{\mathrm{T}} A_S$ is as claimed.

V. DESCRIPTIVE STUDY I: - I

The previous two sections examined relevant dual representations and the spectrum of bipartite matching rectangular array of polyhedral path expressions. Having studied these notions in their pure and basic form, we now apply our findings to resizable Hadoop cluster's complexity. Specifically, we establish the *bipartite matching rectangular array of polyhedral path expressions* for resizable Hadoop cluster's complexity, which gives strong lower bounds for every bipartite matching rectangular array of polyhedral path expressions generated by a hybrid cost function with high approximate degree.

NECESSARY AND SUFFICIENT CONDITION 1.1 (restated). *Let F be the (n,t,f)-bipartite matching rectangular array of polyhedral path expressions, where $f : \{0,1\}^t \to \{-1,+1\}$ is given. Then for every $\epsilon \in [0,1)$ and every $\delta < \epsilon/2$,*

$$(5.1) \qquad Q_\delta^*(F) \geq \frac{1}{4} \deg_\epsilon(f) \log\left(\frac{n}{t}\right) - \frac{1}{2} \log\left(\frac{3}{\epsilon - 2\delta}\right).$$

In particular,

$$(5.2) \qquad Q_{1/7}^*(F) > \frac{1}{4} \deg_{1/3}(f) \log\left(\frac{n}{t}\right) - 3.$$

Proof. Since (5.1) immediately implies (5.2), we will focus on the former in the remainder of the proof. Let $d = \deg_\epsilon(f) \geq 1$. By necessary and sufficient condition 3.2, there is a hybrid kernel function $\psi : \{0,1\}^t \to \mathbb{R}$ such that:

$$(5.3) \qquad \hat{\psi}(S) = 0 \qquad (|S| < d),$$

$$(5.4) \qquad \sum_{z \in \{0,1\}^t} |\psi(z)| = 1,$$

$$(5.5) \qquad \sum_{z \in \{0,1\}^t} \psi(z) f(z) > \epsilon.$$

Let ψ be the $(n,t,2^{-n}(n/t)^{-t}\psi)$-bipartite matching rectangular array of polyhedral path expressions. Then (5.4) and (5.5) show that

$$(5.6) \qquad \|\psi\|_1 = 1, \qquad \langle F, \psi \rangle > \epsilon.$$

Our last task is to calculate $\|\psi\|$. By (5.4) and necessary and sufficient condition 2.1,

$$(5.7) \qquad \max_{S \subseteq [t]} |\hat{\psi}(S)| \leq 2^{-t}.$$

Necessary and sufficient condition 4.3 yields, in view of (5.3) and (5.7):

$$(5.8) \qquad \|\psi\| \leq \left(\frac{t}{n}\right)^{d/2} \left(2^{n+t} \left(\frac{n}{t}\right)^t\right)^{-1/2}.$$

Now (5.1) follows from (5.6), (5.8), and necessary and sufficient condition 2.8.

Necessary and sufficient condition 1.1 gives lower bounds not only for bounded-capacity resizable Hadoop cluster but also for resizable Hadoop cluster's preference limitation protocols

with capacity probability $\frac{1}{2} - o(1)$. For example, if a hybrid kernel function $f : \{0,1\}^t \to \{-1,+1\}$ requires a polynomial of degree d for approximation within $1 - o(1)$, equation (5.1) gives a lower bound for small-bias resizable Hadoop cluster. We will complement and refine that estimate in the next section, which is dedicated to small-bias resizable Hadoop cluster.

We now prove the necessary condition to necessary and sufficient condition 1.1 on hybrid kernel function composition, stated in the inception.

Proof of Necessary condition 1.2. The $(2t, t, f)$-bipartite matching rectangular array of polyhedral path expressions occurs as a subset of rectangular array of polyhedral path expressions of $[F(x,y)]_{x,y\in\{0,1\}^{4t}}$.

Finally, we show that the lower bound (5.2) derived above for bounded-capacity resizable Hadoop cluster's complexity is tight up to a polynomial factor, even for deterministic preference limitation protocols.

NECESSARY AND SUFFICIENT CONDITION 5.1. *Let F be the (n, t, f)-bipartite matching rectangular array of polyhedral path expressions, where $f : \{0,1\}^t \to \{-1,+1\}$ is given. Then*

$$D(F) \le O(\mathrm{dt}(f)\log(n/t)) \le O(\deg_{1/3}(f)^6 \log(n/t)),$$

where $\mathrm{dt}(f)$ *is the least depth of a decision tree for* f. *In particular,* (5.2) *is tight up to a polynomial factor*.

Proof. That $\mathrm{dt}(f) \le O(\deg_{1/3}(f)^6)$ for all hybrid cost functions f. Therefore, it suffices to prove an upper bound of $O(d\log(n/t))$ on the deterministic resizable Hadoop cluster's complexity of F, where $d = \mathrm{dt}(f)$.

The needed deterministic preference limitation protocol is not well known. Fix a depth-d decision tree for f. Let $(x,(V,w))$ be a given input. Alice and Bob start at the root of the decision tree, labeled by some variable $i \in \{1,\dots,t\}$. By exchanging $\lceil\log(n/t)\rceil + 2$ compatible JAR files, Alice and Bob determine $(x|_V)_i \oplus w_i \in \{0,1\}$ and take the corresponding branch of thetree. The process repeats until a leaf is reached, at which point both parties learn $f(x|_V \oplus w)$

VI. DESCRIPTIVE STUDY I: - II

As we have already mentioned, necessary and sufficient condition 1.1 of the previous section can be used to obtain lower bounds not only for bounded-capacity resizable Hadoop cluster but also small-bias resizable Hadoop cluster. In the latter case, one first needs to show that the base hybrid kernel function $f : \{0,1\}^t \to \{-1,+1\}$ cannot be approximated point wise within $1 - o(1)$ by a real polynomial of a given degree d. In this section, we derive a different lower bound for small-bias resizable Hadoop cluster, this time using the assumption that the threshold weight $W(f,d)$ is high. We will see that this new lower bound is nearly optimal and closely related to the lower bound in necessary and sufficient condition 1.1.

NECESSARY AND SUFFICIENT CONDITION 6.1. *Let F be the (n, t, f)-bipartite matching rectangular array of polyhedral path expressions, where $f : \{0,1\}^t \to \{-1,+1\}$ is given. Then for every integer $d \ge 1$ and real $\gamma \in (0,1)$,*

$$(6.1)\qquad Q^*_{1/2-\gamma/2}(F) \ge \frac{1}{4}\min\left\{d\log\frac{n}{\mathrm{t}}, \log\frac{W(f,d-1)}{2t}\right\} - \frac{1}{2}\log\frac{3}{\gamma}.$$

In particular,

$$(6.2)\qquad Q^*_{1/2-\gamma/2}(F) \ge \frac{1}{4}\deg_{\pm}(f)\log\left(\frac{n}{\mathrm{t}}\right) - \frac{1}{2}\log\frac{3}{\gamma}.$$

Proof. Letting $d = \deg_{\pm}(f)$ in (6.1) yields (6.2), since $W(f, d-1) = \infty$ in that case. In the remainder of the proof, we focus on (6.1) alone. We claim that there exists a distribution μ on $\{0,1\}^t$ such that

$$(6.3)\qquad \max_{|S|<d}\left|\mathop{\mathbf{E}}_{z\sim\mu}[f(z)X_s(z)]\right| \le \left(\frac{2t}{W(f,d-1)}\right)^{1/2}.$$

For $d \le \deg_{\pm}(f)$, the claim holds by necessary and sufficient condition 3.3 since $W(f,d-1) = \infty$ in that case. For $d > \deg_{\pm}(f)$, the claim holds by necessary and sufficient condition 3.4. Now, define $\psi : \{0,1\}^t$ by $\psi(z) = f(z)\mu(z)$. It follows from (6.3) that

$$(6.4)\qquad |\widehat{\psi}(S)| \le 2^{-t}\left(\frac{2t}{W(f,d-1)}\right)^{1/2} (|S| < d),$$

$$(6.5)\qquad \sum_{z\in\{0,1\}^t} |\psi(z)| = 1,$$

$$(6.6)\qquad \sum_{z\in\{0,1\}^t} \psi(z)f(z) = 1.$$

Let ψ be the $(n, t, 2^{-n}(n/t)^{-t}\psi)$-bipartite matching rectangular array of polyhedral path expressions. Then (6.5) and (6.6) show that

$$(6.7)\qquad \|\psi\|_1 = 1, \qquad \langle F, \psi\rangle = 1.$$

It remains to calculate $\|\psi\|$. By (6.5) and necessary and sufficient condition 2.1,

$$(6.8)\qquad \max_{S\subseteq[t]} |\widehat{\psi}(S)| \le 2^{-t}.$$

Necessary and sufficient condition 4.3 yields, in view of (6.4) and (6.8):

$$(6.9)\quad \|\psi\| \le \max\left\{\left(\frac{t}{n}\right)^{d/2}, \left(\frac{2t}{W(f,d-1)}\right)^{1/2}\right\}\left(2^{n+t}\left(\frac{n}{t}\right)^t\right)^{-1/2}.$$

Now (6.1) follows from (6.7), (6.9), and necessary and sufficient condition 2.8. Recall from necessary and sufficient condition 2.5 that the quantities $E(f,d)$ and $W(f,d)$ are related for all f and d. In particular, the lower bounds for small-bias resizable Hadoop cluster in Propositions 1.1and 6.1 are quite close, and either one can be approximately deduced from the other. In deriving both findings from scratch, as we did, our motivation was to obtain the tightest bounds and to illustrate the bipartite matching rectangular array of polyhedral path expressions in different contexts. We will now see that the lower bound in necessary and sufficient condition 6.1 is close to optimal, even for channel preference limitation protocols.

NECESSARY AND SUFFICIENT CONDITION 6.2. *Let F be the (n,t,f)-bipartite matching rectangular array of polyhedral path expressions, where $f:\{0,1\}^t \to \{-1,+1\}$ is given. Then for every integer $d \geq deg_{\pm}(f)$,*

$$Q^*_{1/2-\gamma/2}(F) \leq R_{1/2-\gamma/2}(F) \leq d \log\left(\frac{n}{\mathrm{t}}\right) + 3,$$
where $\gamma - 1/W(f,d)$.

Proof. The resizable Hadoop cluster's preference limitation protocol that we will describe is standard. Put $W = W(f,d)$ and fix a representation

$$f(z) \equiv \operatorname{sgn}\left(\sum_{S\subseteq[t],|S|\leq d} \lambda_S X_S(z)\right),$$

where the integers λ_S satisfy $\sum|\lambda_S| = W$. On input $\big(x,(V,w)\big)$, the preference limitation protocol proceeds as follows. Let $i_1 < i_2 < \cdots < i_t$ be the elements of V. Alice and Bob use their shared randomness to pick a set $S \subseteq [t]$ with $|S| \leq d$, according to the probability distribution $|\lambda_S|/W$. Next, Bob sends Alice the indices $\{i_j : j \in S\}$ as well as the file $X_S(w)$. With this information, Alice computes the product $\operatorname{sgn}(\lambda_S)X_S(x|_V)X_S(w) = \operatorname{sgn}(\lambda_S)X_S(x|_V \oplus w)$ and announces the result as the output of the preference limitation protocol. Assuming an optimal encoding of the compatible JAR files, the resizable Hadoop cluster's cost of this preference limitation protocol is bounded by

$$\left\lceil \log\binom{n}{t}^d \right\rceil + 2 \leq d\log\left(\frac{n}{t}\right) + 3,$$

as desired. On each input x, V, w, the output of the preference limitation protocol is a random variable that $P(x,V,w) \in \{-1,+1\}$ obeys

$$\begin{aligned} &f(x|_V \oplus w)\ \mathbf{E}[P(x,V,w)] \\ &= f(x|_V \oplus w) \sum_{|S|\leq d} \frac{|\lambda_S|}{W} \operatorname{sgn}(\lambda_S) X_S(x|_V \\ &\oplus w) \\ &= \frac{1}{W}\left|\sum_{|S|\leq d} \lambda_S X_S(x|_V \oplus w)\right| \\ &\geq \frac{1}{W}, \end{aligned}$$

which means that the preference limitation protocol produces the correct answer with probability $\frac{1}{2} + \frac{1}{2W}$ or greater.

VII. PRESCRIPTIVE STUDY: - I

We now restate some of the findings of the previous section in terms of *discrepancy*, a key notion already mentioned in Section 2.4. This quantity figures prominently in the study of small-bias resizable Hadoop cluster as well as various applications, such as inference theory and cluster complexity [36]. For a hybrid cost function $f : X \times Y \to \{-1,+1\}$ and a probability distribution λ on $X \times Y$, the discrepancy of f under λ is defined by

$$\operatorname{disc}_\lambda(f) = \max_{\substack{S\subseteq X,\\ T\subseteq Y}} \left|\sum_{x\in S}\sum_{y\in T} \lambda(x,y) f(x,y)\right|.$$

We put

$$\operatorname{disc}(f) = \min_\lambda \operatorname{disc}_\lambda(f).$$

As usual, we will identify a hybrid kernel function $f : X \times Y \to \{-1,+1\}$ with its resizable Hadoop cluster rectangular array of polyhedral path expressions $F = [f(x,y)]_{x,y}$ and use the conventions $\operatorname{disc}_\lambda(F) = \operatorname{disc}_\lambda(f)$ and $\operatorname{disc}(F) = \operatorname{disc}(f)$. The above intention of discrepancy is not convenient to work with, and we will use a well-known rectangular array of polyhedral path expressions-analytic reformulation. For rectangular array of polyhedral path expressions $A = [A_{xy}]$ and $B = [B_{xy}]$, recall that their product is given by $A \circ B = [A_{xy}B_{xy}]$.

NECESSARY AND SUFFICIENT CONDITION 7.1. *Let X, Y be finite geometric information sets, $f : X \times Y \to \{-1,+1\}$ a given hybrid kernel function. Then*

$$\operatorname{disc}_P(f) \leq \sqrt{|X||Y|}\|P \circ F\|,$$

where $F = [f(x,y)]_{x\in X, y\in Y}$ and P is any rectangular array of polyhedral path expressions whose entries are nonnegative and sum to 1 *(viewed as a probability distribution). In particular,*

$$\operatorname{disc}(f) \leq \sqrt{|X||Y|}\min_P \|P \circ F\|,$$

where the minimum is over rectangular array of polyhedral path expressions P whose entries are nonnegative and sum to 1.

Proof. We have

$$\begin{aligned} \operatorname{disc}_P(f) &= \max_{S,T} \left|\mathbf{1}_S^{\mathrm{T}}(P \circ F)\mathbf{1}_T\right| \\ &\leq \max_{S,T}\{\|\mathbf{1}_S\| \cdot \|P \circ F\| \cdot \|\mathbf{1}_T\|\} \\ &= \|P \circ F\|\sqrt{|X||Y|}, \end{aligned}$$

as claimed. We will need one last ingredient, a well-known lower bound on resizable Hadoop cluster's complexity in terms of discrepancy.

NECESSARY AND SUFFICIENT CONDITION 7.2. *For every hybrid kernel function* $f : X \times Y \to \{-1, +1\}$ *and every* $\gamma \in (0, 1)$,

$$R_{1/2-\gamma/2}(f) \geq \log \frac{\gamma}{\operatorname{disc}(f)}.$$

Using Propositions 6.1 and 6.2, we will now characterize the discrepancy of bipartite matching rectangular array of polyhedral path expressions in terms of threshold weight.

NECESSARY AND SUFFICIENT CONDITION 7.3. *Let F be the* (n, t, f)*-bipartite matching rectangular array of polyhedral path expressions, where* $f : \{0,1\}^t \to \{-1, +1\}$ *is given. Then for every integer* $d \geq 0$,

$$\operatorname{disc}(F) \geq \frac{1}{8W(f, d)} \left(\frac{t}{n}\right)^d \tag{7.1}$$

and

$$\operatorname{disc}(F)^2 \leq \max\left\{\frac{2t}{W(f, d-1)}, \left(\frac{t}{n}\right)^d\right\}. \tag{7.2}$$

In particular,

$$\operatorname{disc}(F) \leq \left(\frac{t}{n}\right)^{\deg_{\pm}(f)/2}. \tag{7.3}$$

Proof. The lower bound (7.1) is immediate from necessary and sufficient condition 6.2 and necessary and sufficient condition 7.2. For the upper bound (7.2), construct the rectangular array of polyhedral path expressions ψ as in the proof of necessary and sufficient condition 6.1. Then (6.7) shows that $\psi = F \circ P$ for a nonnegative rectangular array of polyhedral path expressions P whose entries sum to 1. As a result, (7.2) follows from (6.9) and necessary and sufficient condition 7.1. Finally, (7.3) follows by taking $d = \deg_{\pm}(f)$ in (7.2), since $W(f, d-1) = \infty$ in that case. This settles necessary and sufficient condition 1.5 from the inception. Necessary and sufficient condition 7.3 follows up and considerably improves on the *Degree/Discrepancy necessary and sufficient condition*.

NECESSARY AND SUFFICIENT CONDITION 7.4. *Let* $f : \{0,1\}^t \to \{-1, +1\}$ *be given. Fix an integer* $n \geq t$. *Let* $M = [f(x|_S)]_{x,S}$, *where the row index* x *ranges over* $\{0,1\}^n$ *and the column index* S *ranges over all* t*-element subsets of* $\{1, 2, \ldots, n\}$. *Then*

$$\operatorname{disc}(M) \leq \left(\frac{4et^2}{n \deg_{\pm}(f)}\right)^{\deg_{\pm}(f)/2}.$$

Note that (7.3) is already stronger than necessary and sufficient condition 7.4. In Section 10, we will see an example when necessary and sufficient condition 7.3 gives an exponential improvement on necessary and sufficient condition 7.4. Threshold weight is typically easier to analyze than the approximate degree. For completeness, however, we will now supplement necessary and sufficient condition 7.3 with an alternate bound on the discrepancy of a bipartite matching rectangular array of polyhedral path expressions in terms of the approximate degree.

NECESSARY AND SUFFICIENT CONDITION 7.5. *Let F be the* (n, t, f)*-bipartite matching rectangular array of polyhedral path expressions, for a given hybrid kernel function* $f : \{0,1\}^t \to \{-1, +1\}$. *Then for every* $\gamma > 0$,

$$\operatorname{disc}(F) \leq \gamma + \left(\frac{t}{n}\right)^{\deg_{\pm}(f)/2}.$$

Proof. Let $d = \deg_{1-\gamma}(f) \geq 1$. Define $\epsilon = 1 - \gamma$ and construct the rectangular array of polyhedral path expressions ψ as in the proof of necessary and sufficient condition 1.1. Then (5.6) shows that $\psi = H \circ P$, where H is a finite string rectangular array of polyhedral path expressions and P is a nonnegative rectangular array of polyhedral path expressions whose entries sum to 1. Viewing P as a probability distribution, we infer from (5.8) and necessary and sufficient condition 7.1 that

$$\operatorname{disc}_P(H) \leq \left(\frac{t}{n}\right)^{d/2}. \tag{7.4}$$

Moreover,

$$\begin{aligned} \operatorname{disc}_P(F) &\leq \operatorname{disc}_P(H) + \|(F - H) \circ P\|_1 \\ &= \operatorname{disc}_P(H) + 1 - \langle F, H \circ P \rangle \\ &\leq \operatorname{disc}_P(H) + \gamma, \end{aligned} \tag{7.5}$$

where the last step follows because $\langle F, \psi \rangle > \epsilon = 1 - \gamma$ by (5.6). The proof is complete in view of (7.4) and (7.5).

VIII. PRESCRIPTIVE STUDY: - II

We will now use the findings of the previous sections to analyze the approximate PageRank and approximate trace distance norm of bipartite matching rectangular array of polyhedral path expressions. These notions were originally motivated by lower bounds on resizable Hadoop cluster. However, they also arise in inference theory and are natural analytic quantities in their own right.

NECESSARY AND SUFFICIENT CONDITION 8.1. *Let F be the* (n, t, f)*-bipartite matching rectangular array of polyhedral path expressions, where* $f : \{0,1\}^t \to \{-1, +1\}$ *is given. Let* $s = 2^{n+t}(n/t)^t$ *be the number of entries in* F. *Then for every* $\epsilon \in [0, 1)$ *and every* $\delta \in [0, \epsilon]$,

(8.1) $$\|F\|_{\sum,\delta} \geq (\epsilon - \delta)\left(\frac{n}{t}\right)^{\deg_\epsilon(f)/2}\sqrt{s}$$

and

(8.2) $$\mathrm{rk}_\delta F \geq \left(\frac{\epsilon-\delta}{1+\delta}\right)^2\left(\frac{n}{t}\right)^{\deg_\epsilon(f)}.$$

Proof. We may assume that $\deg_\epsilon(f) \geq 1$, since otherwise f is a constant hybrid kernel function and the claims hold trivially by taking $\psi = F$ in necessary and sufficient condition 2.2. Construct ψ asin the proof of necessary and sufficient condition 1.1. Then the claimed lower bound on $\|F\|_{\sum,\delta}$ follows from (5.6), (5.8), and necessary and sufficient condition 2.2. Finally, (8.2) follows immediately from (8.1) and necessary and sufficient condition 2.3. We prove an additional lower bound in the case of small-bias approximation.

NECESSARY AND SUFFICIENT CONDITION 8.2. *Let F be the (n,t,f)-bipartite matching rectangular array of polyhedral path expressions, where $f:\{0,1\}^t \to \{-1,+1\}$ is given. Let $s = 2^{n+t}(n/t)^t$ be the number of entries in F. Then for every $\gamma \in (0,1)$ and every integer $d \geq 1$,*

(8.3) $$\|F\|_{\sum,1-\gamma} \geq \gamma \min\left\{\left(\frac{n}{t}\right)^{d/2}, \left(\frac{W(f,d-1)}{2t}\right)^{1/2}\right\}\sqrt{s}$$

and

(8.4) $$\mathrm{rk}_{1-\gamma} F \geq \left(\frac{\gamma}{2-\gamma}\right)^2 \min\left\{\left(\frac{n}{t}\right)^d, \frac{W(f,d-1)}{2t}\right\}.$$

In particular,

(8.5) $$\|F\|_{\sum,1-\gamma} \geq \gamma\left(\frac{n}{t}\right)^{\deg_\pm(f)/2}\sqrt{s}$$

and

(8.6) $$\mathrm{rk}_{1-\gamma} F \geq \left(\frac{\gamma}{2-\gamma}\right)^2\left(\frac{n}{t}\right)^{\deg_\pm(f)}.$$

Proof. Construct ψ as in the proof of necessary and sufficient condition 6.1. Then the claimed lower bound on $\|F\|_{\sum,\delta}$ follows from (6.7), (6.9), and necessary and sufficient condition 2.2. Now (8.4) follows from (8.3) and necessary and sufficient condition 2.3. Finally, (8.5) and (8.6) follow by taking $d = \deg_\pm(f)$ in (8.3) and (8.4), respectively, since $W(f,d-1) = \infty$ in that case. Propositions 8.1 and 8.2 settle necessary and sufficient condition 1.4 from the inception. Recall that necessary and sufficient condition 4.3 gives an easy way to calculate the trace distance norm and PageRank of a bipartite matching rectangular array of polyhedral path expressions. In particular, it is straight forward to verify that the lower bounds in (8.2) and (8.4) are close to optimal for various choices of ϵ, δ, γ. For example, one has $\|F - A\|_\infty \leq 1/3$ by taking F and A to be the (n,t,f)-and (n,t,ϕ)-bipartite matching rectangular array of polyhedral path expressions, where $\phi:\{0,1\}^t \to \mathbb{R}$ is any polynomial of degree $\deg_{1/3}(f)$ with $\|F-\phi\|_\infty \leq 1/3$.

IX. DESCRIPTIVE STUDY II:- I

As an illustrative application of the bipartite matching rectangular array of polyhedral path expressions, we now give a short and elementary proof of optimal lower bounds for every conjunctive predicate $D:\{0,1,\dots,n\} \to \{-1,+1\}$. We first solve the problem for all conjunctive predicates D that change value close to 0. Extension to the general case will require an additional step.

NECESSARY AND SUFFICIENT CONDITION 9.1. *Let $D:\{0,1,\dots,n\} \to \{-1,+1\}$ be a given conjunctive predicate. Suppose that $D(\ell) \neq D(\ell-1)$ for some $\ell \leq \frac{1}{8}n$. Then*

$$Q^*_{1/3}(D) \geq \Omega(\sqrt{n\ell}).$$

Proof. It suffices to show that $Q^*_{1/7}(D) \geq \Omega(\sqrt{n\ell})$. Define $f:\{0,1\}^{\lfloor n/4\rfloor} \to \{-1,+1\}$ by $f(z) = D(|z|)$. Then $\deg_{1/3}(f) \geq \Omega(\sqrt{n\ell})$ by necessary and sufficient condition 2.6. necessary and sufficient condition 1.1 implies that

$$Q^*_{1/7}(F) \geq \Omega(\sqrt{n\ell}),$$

where F is the $(2\lfloor n/4\rfloor, \lfloor n/4\rfloor, f)$-bipartite matching rectangular array. Since F occurs as a subset of rectangular array of polyhedral path expressions of $[D(|x \wedge y|)]_{x,y}$, the proof is complete. The remainder of this section is a simple if tedious exercise in shifting and padding. We note that proof concludes in a similar way.

NECESSARY AND SUFFICIENT CONDITION 9.2. *Let $D:\{0,1,\dots,n\} \to \{-1,+1\}$ be a given conjunctive predicate. Suppose that $D(\ell) \neq D(\ell-1)$ for some $\ell \leq \frac{1}{8}n$. Then*

(9.1) $$Q^*_{1/3}(D) \geq c(n-\ell)$$

for some absolute constant $c > 0$.

Proof. Consider the resizable Hadoop cluster problem of computing $D(|x \wedge y|)$ when the last k compatible JAR files in x and y are fixed to 1. In other words, the new problem is to compute $D_k(|x' \wedge y'|)$ where $x', y' \in \{0,1\}^{n-k}$ and the conjunctive predicate $D_k:\{0,1,\dots,n-k\} \to \{-1,+1\}$, is given by $D_k(i) \equiv D(k+i)$. Since the new problem is a restricted version of the original, we have

(9.2) $$Q^*_{1/3}(D) \geq Q^*_{1/3}(D_k).$$

We complete the proof by placing a lower bound on $Q^*_{1/3}(D_k)$ for

$$k = \ell - \left\lfloor \frac{\alpha}{1-\alpha} \cdot (n-\ell) \right\rfloor,$$

where $\alpha = \frac{1}{8}$. Note that k is an integer between 1 and ℓ (because $\ell > \alpha n$). The equality $k = \ell$ occurs if and only if $\left\lfloor \frac{\alpha}{1-\alpha}(n-\ell) \right\rfloor = 0$, in which case (9.1) holds trivially for c suitably small. Thus, we can assume that $1 \leq k \leq \ell - 1$, in which case $D_k(\ell - k) \neq D_k(\ell - k - 1)$ and $\ell - k \leq \alpha(n - k)$. Therefore, necessary and sufficient condition 9.1 is applicable to D_k and yields:

(9.3) $$Q^*_{1/3}(D_k) \geq C\sqrt{(n-k)(\ell-k)},$$

where $C > 0$ is an absolute constant. Calculations reveal:

(9.4) $$n - k = \left\lfloor \frac{1}{1-\alpha} \cdot (n-\ell) \right\rfloor,$$

$$\ell - k = \left\lfloor \frac{\alpha}{1-\alpha} \cdot (n-\ell) \right\rfloor.$$

The necessary and sufficient condition is now immediate from (9.2)–(9.4). Together, Propositions 9.1 and 9.2 give the main result of this section:

NECESSARY AND SUFFICIENT CONDITION 1.3 (restated). *Let* $D : \{0, 1, \ldots, n\} \to \{-1, +1\}$. *Then*

$$Q^*_{1/3}(D) \geq \Omega\left(\sqrt{n\ell_0(D)} + \ell_1(D)\right),$$

where $\ell_0(D) \in \{0, 1, \ldots, \lfloor n/2 \rfloor\}$ *and* $\ell_1(D) \in \{0, 1, \ldots, \lceil n/2 \rceil\}$ *are the smallest integers such that* D *is constant in the range* $[\ell_0(D), n - \ell_1(D)]$.

Proof. If $\ell_0(D) \neq 0$, set $\ell = \ell_0(D)$ and note that $D(\ell) \neq D(\ell - 1)$ by intention. One of Propositions 9.1 and 9.2 must be applicable, and therefore $Q^*_{1/3}(D) \geq \min\{\Omega(\sqrt{n\ell}), \Omega(n - \ell)\}$. Since $\ell \leq n/2$, this simplifies to

(9.5) $$Q^*_{1/3}(D) \geq \Omega\left(\sqrt{n\ell_0(D)}\right).$$

If $\ell_1(D) \neq 0$, set $\ell = n - \ell_1(D) + 1 \geq n/2$ and note that $D(\ell) \neq D(\ell - 1)$ as before. By necessary and sufficient condition 9.2,

(9.5) $$Q^*_{1/3}(D) \geq \Omega\left(\ell_1(D)\right).$$

The necessary and sufficient condition follows from (9.5) and (9.6).

X. DESCRIPTIVE STUDY II: - II

As another application of the bipartite matching rectangular array of polyhedral path expressions, we revisit the discrepancy of AC^0, the class of polynomial-size constant-depth Hadoop clusters. Independently, [26] exhibited another hybrid kernel function in AC^0 with exponentially small discrepancy. We revisit this discrepancy below, considerably sharpening the bound in [25] and giving a new and simple proof of the bound. Consider the hybrid kernel function $\mathrm{MP}_m : \{0,1\}^{4m^3} \to \{-1, +1\}$ given by

$$\mathrm{MP}_m(x) = \bigvee_{i=1}^{m} \bigwedge_{j=1}^{4m^2} x_{ij}.$$

Using this hybrid kernel function and the Degree/Discrepancy necessary and sufficient condition (necessary and sufficient condition 7.4), an upper bound of $\exp\{-\Omega(n^{1/5})\}$ was derived on the discrepancy of an explicit AC^0cluster $f : \{0,1\}^n \times \{0,1\}^n \to \{-1, +1\}$ of depth 3. We will now sharpen that bound to $\exp\{-\Omega(n^{1/3})\}$.

NECESSARY AND SUFFICIENT CONDITION 1.6 (restated). *Let* $f(x, y) = \mathrm{MP}_m(x \vee y)$. *Then*

$$\mathrm{disc}(f) = \exp\{-\Omega(m)\}.$$

Proof. Put $d = \lfloor n/2 \rfloor$. We state that $\deg_{\pm}(\mathrm{MP}_d) \geq d$. Since the $(8d^3, 4d^3, \mathrm{MP}_d)$-bipartite matching rectangular array of polyhedral path expressions is a subset of rectangular array of polyhedral path expressions of $[f(x, y)]_{x,y}$, the proof is complete in view of equation (7.3) of necessary and sufficient condition 7.3. The ODD-MAX-BOUND hybrid kernel function $\mathrm{OMB}_n : \{0,1\}^n \to \{-1, +1\}$, is given by

(10.1) $$\mathrm{OMB}_n(x) = \mathrm{sgn}\left(1 + \sum_{i=1}^{n} (-2)^i x_i\right).$$

It is straight-forward to compute OMB_n by a linear-size DNF formula and even a decision list. In particular, OMB_n belongs to the class AC^0.

NECESSARY AND SUFFICIENT CONDITION 10.1 (Chazelle et al.). *Let* $f(x, y) = \mathrm{OMB}_n(x \wedge y)$. *Then*

$$\mathrm{disc}(f) = \exp\{-\Omega(n^{1/3})\}.$$

Using the celebrated findings of Chazelle's papers, we can give a short alternate proof of this necessary and sufficient condition.

Proof. Put $m = \lfloor n/4 \rfloor$. Shows that $W(\mathrm{OMB}_m, cm^{1/3}) \geq \exp(cm^{1/3})$ for some absolute constant $c > 0$. Since the $(2m, m, \mathrm{OMB}_m)$-bipartite matching rectangular array of polyhedral path expressions is a subset of rectangular array of polyhedral path expressions of $[f(x, y)]_{x,y}$, the proof is complete by necessary and sufficient condition 7.3.

REMARK 10.2. The above proofs illustrate that the characterization of the discrepancy of bipartite matching rectangular array of polyhedral path expressions. In particular, the representation (10.1) makes it clear that $\deg_{\pm}(\mathrm{OMB}_n) = 1$ and therefore necessary and sufficient condition 7.4 cannot yield an upper bound better than $n^{-\Omega(1)}$ on the discrepancy of

$\mathrm{OMB}_n(x \wedge y)$. Necessary and sufficient condition 7.3, on the other hand, gives an exponentially better upper bound. It is well known that the discrepancy of a hybrid kernel function f implies a lower bound on the size of majority Hadoop clusters that compute f. Following, we record the consequences of Propositions 1.6 and 10.1 in this regard.

NECESSARY AND SUFFICIENT CONDITION 10.3. *Any majority vote of threshold that computes the hybrid kernel function*

$$f(x, y) = \mathrm{MP}_m(x \vee y)$$

has size $\exp\{\Omega(m)\}$. *Analogously, any majority vote of threshold that computes the hybrid kernel function*

$$f(x, y) = \mathrm{OMB}_n(x \wedge y)$$

has size $\exp\{\Omega(n^{1/3})\}$.

XI. Conclusions

In previous sections, we characterized various rectangular array of polyhedral path expressions-analytic and combinatorial properties of bipartite matching rectangular array of polyhedral path expressions, including their channel and resizable Hadoop cluster's complexity, discrepancy, approximate PageRank, and approximate trace distance norm. We conclude this study with another fact about bipartite matching rectangular array of polyhedral path expressions. We observed that the deterministic resizable Hadoop cluster's complexity of a finite string rectangular array of polyhedral path expressions F satisfies $D(F) \geq \log \mathrm{rk}\, F$. The log-PageRank hypothesis is that this lower bound is always tight up to a polynomial factor, i.e.,$D(F) \leq (\log \mathrm{rk}\, F)^{O(1)} + O(1)$. Using the findings of the previous sections, we can givea short proof of this hypothesis in the case of bipartite matching rectangular array of polyhedral path expressions.

NECESSARY AND SUFFICIENT CONDITION 11.1. *Let* $f : \{0,1\}^t \to \{-1, +1\}$ *be a given hybrid kernel function,* $d = \deg(f)$. *Let* F *be the* (n, t, f)*-bipartite matching rectangular array of polyhedral path expressions. Then*

$$\mathrm{rk}\, F \geq \left(\frac{n}{t}\right)^d \geq \exp\{\Omega(D(F)^{1/4})\}. \tag{11.1}$$

In particular, F satisfies the log-PageRank hypothesis.

Proof. Since $\hat{f}(S) \neq 0$ for some set S with $|S| = d$, necessary and sufficient condition 4.3 implies that F has at least $(n/t)^d$ nonzero singular key-values. This settles the first inequality in (11.1). Necessary and sufficient condition 5.1 implies that $D(F) \leq O(\mathrm{dt}(f) \log(n/t))$, where $\mathrm{dt}(f)$ denotes the least depth of a decision tree for f that $\mathrm{dt}(f) \leq 2 \deg(f)^4$ for all f. Combining these two observations establishes the second inequality in (11.1).

XII. Discussions

Fix hybrid kernel functions $f : \{0,1\}^n \to \{-1, +1\}$ and $g : \{0,1\}^k \times \{0,1\}^k \to \{-1, +1\}$. Let $f \circ g^n$ denote the composition of f with n independent copies of g. More formally, the hybrid kernel function $f \circ g^n : \{0,1\}^{nk} \times \{0,1\}^{nk} \to \{-1, +1\}$ is given by

$$(f \circ g^n)(x, y) = f\left(g\left(x^{(1)}, y^{(1)}\right), \ldots, g\left(x^{(n)}, y^{(n)}\right)\right),$$

where $x = \left(x^{(1)}, \ldots, x^{(n)}\right) \in \{0,1\}^{nk}$ and $y = \left(y^{(1)}, \ldots, y^{(n)}\right) \in \{0,1\}^{nk}$. The resizable Hadoop cluster's complexity of $f \circ g^n$is that

$$Q^*_{1/3}(f \circ g^n) \geq \Omega\left(\deg_{1/3}(f)\right) \text{ provided that}$$
$$\rho(g) \leq \frac{\deg_{1/3}(f)}{2en},$$

where $\rho(g)$ is a new variant of discrepancy that the authors introduce. As an illustration, they re-prove a weaker version of lower bounds in necessary and sufficient condition 1.3. In our terminology (Section 2.4), their proof also fits in the framework of the discrepancy method. The quantity $\rho(g)$, which needs to be small. This poses two complications. First, the hybrid kernel function g will generally need to depend on many variables, from $k = \Theta(\log n)$ to $k = n^{\Theta(1)}$, which weakens the final lower bounds on resizable Hadoop cluster. A second complication, as the authors note, is that "estimating $\rho(g)$ is unfortunately difficult in general". For example, re-proving lower bounds reduces to estimating $\rho(g)$ for $g(x, y) = x_1 y_1 \vee \ldots \vee x_k y_k$. Our method avoids these complications altogether. For example, we prove (by taking $n = 2t$ in the bipartite matching rectangular array of polyhedral path expressions, necessary and sufficient condition 1.1) that

$$Q^*_{1/3}(f \circ g^n) \geq \Omega\left(\deg_{1/3}(f)\right)$$

for any hybrid kernel function $g : \{0,1\}^k \times \{0,1\}^k \to \{-1, +1\}$ such that the rectangular array of polyhedral path expressions $[g(x, y)]_{x,y}$ contains the following subset rectangular array of polyhedral path expressions, up to permutations of rows and columns:

$$\begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}.$$

To illustrate, one can take g to be

$$g(x, y) = x_1 y_1 \vee x_2 y_2 \vee x_3 y_3 \vee x_4 y_4$$

or

$$g(x, y) = x_1 y_1 y_2 \vee \overline{x_1} y_1 \overline{y_2} \vee x_2 \overline{y_1} y_2 \vee \overline{x_2 y_1 y_2}.$$

In summary, there is a simple hybrid kernel function g on $k = 2$ variables that works universally for all f. This means

no technical conditions to check, such as $\rho(g)$, and no blow-up in the number of variables. As a result, we are able to re-prove optimal lower bounds exactly. Moreover, the technical machinery of this paper is self-contained and disjoint from proof. A further advantage of the bipartite matching rectangular array of polyhedral path expressions is that it extends in a straightforward way to the multi-cloud model. This extension depends on the fact that the rows of a bipartite matching rectangular array of polyhedral path expressions are applications of the same hybrid kernel function to different subsets of the variables. In the general context of block composition, it is unclear how to carry out this extension.

REFERENCES

[1]. Ravi Prakash G, Kiran M and Saikat Mukherjee. 2014. On Randomized Preference Limitation Protocol for Quantifiable Shuffle and Sort Behavioral Implications in MapReduce Programming Model. *Parallel & Cloud Computing* **3**, Issue 1, 1-14.

[2]. Greenlaw, R. and Kantabutra. 2008. On the parallel complexity of hierarchical clustering and CC-complete problems. *Complexity* **14**, 18-28. (doi:10.1002/cplx.20238)

[3]. Ravi (Ravinder) Prakash G, Kiran M. 2014. On The Least Economical MapReduce Sets for Summarization Expressions. *International Journal of Computer Applications* **94**, 13-20. (doi: 10.5120/16354-5732)

[4]. Amazon Elastic MapReduce. http://aws.amazon.com/elasticmapreduce/

[5]. Ravi (Ravinder) Prakash G, Kiran M. "Problems on Inverted Index Summarization Expressions for Resizable Hadoop Cluster Channel and Cluster Complexity" International Journal of Latest Technology in Engineering, Management & Applied Science (IJLTEMAS), Volume V, Issue V, May 2016, Pages: 1-19, ISSN 2278 – 2540

[6]. N. Ailon, B. Chazelle, S. Comandur, D. Liu. 2007. Estimating the Distance to a Monotone Function. *Random Structures and Algorithms* **31**, 371-383. (doi:10.1002/rsa.20167)

[7]. A. Gavish, Abraham Lempel. 1996. Match-length functions for data compression. *IEEE Transactions on Information Theory* **42**, 1375-1380. (doi:10.1109/18.532879)

[8]. Ravi (Ravinder) Prakash G, Kiran M. "Is it Consistent with Counting that any Summarization Expressions with Resizable Hadoop Cluster Channel have a Cluster Complexity?" International Journal of Engineering Research and Management (IJERM), Volume-03, Issue-06, June 2016, Pages: 135-152, ISSN: 2349- 2058.

[9]. Ping Wah Wong. 1997. Rate distortion efficiency of subband coding with crossband prediction. *IEEE Transactions on Information Theory* **43**, 352-356. (doi:10.1109/18.567761)

[10]. A. Lafourcade, Alexander Vardy. 1996. Optimal sectionalization of a trellis. *IEEE Transactions on Information Theory* **42**, 689-703. (doi: 10.1109/18.490504)

[11]. T.M. Cover. 1998. Comments on Broadcast Channels. *IEEE Transactions on Information Theory* **44**, 2524-2530. (doi: 10.1109/18.720547)

[12]. A. Lapidoth and P. Narayan. 1998. Reliable Communication Under Channel Uncertainty. *IEEE Transactions on Information Theory* **44**, 2148-2177. (doi:10.1109/18.720535)

[13]. David K. Ruch, Patrick J. Van Fleet, (October 2009). Wavelet Theory: An Elementary Approach with Applications, 504 pages pages, SBN: 978-0-470-38840-2.

[14]. Alexander Schrijver, 2004, Combinatorial Optimization Polyhedra and Efficiency, Volume A-C, Algorithms and Combinatorics 24, Pages: CIV, 1879, Springer-Verlag, ISBN 978-3-540-44389-6.

[15]. Leo Breiman. 1993. Hinging hyperplanes for regression, classification, and function approximation. *IEEE Transactions on Information Theory* **39**, 999-1013. (doi:10.1109/18.256506)

[16]. S. R. Kulkarni, D. N.C. Tse. 1994. A paradigm for class identification problems. *IEEE Transactions on Information Theory* **40**, 696-705. (doi:10.1109/18.335881)

[17]. Donald Miner, Adam Shook, 2013, "MapReduce Design Patterns" O'Reilly Media, Inc.: 978-1-449-32717-0.

[18]. Rudolf F. Ahlswede, Zhen Zhang. 1994. On multiuser write-efficient memories. *IEEE Transactions on Information Theory* **40**, 674-686. (doi:10.1109/18.335880)

[19]. B. Chazelle. 2000. *The Discrepancy Method: Randomness and Complexity*. Cambridge University Press. 978-0-521-77093-9.

[20]. B. Chazelle, A. Lvov. 2001. A Trace Bound for the Hereditary Discrepancy. *Discrete Computational. Geom*. **26**, 221-231. (doi:10.1007/s00454-001-0030-2)

[21]. B. Chazelle, A. Lvov. 2001. The Discrepancy of Boxes in Higher Dimension. *Discrete Computational. Geom*. **25**, 519-524. (doi:10.1007/s00454-001-0014-2)

[22]. B. Chazelle, J. Matoušek, M. Sharir. 1995. An Elementary Approach to Lower Bounds in Geometric Discrepancy. *Discrete Comput. Geom*. **13**, 363-381. (doi:10.1007/BF02574050)

[23]. E. Arikan. 1994. An upper bound on the zero-error list-coding capacity. *IEEE Transactions on Information Theory* **40**, 1237-1240. (doi:10.1109/18.335947)

[24]. B. Chazelle, H. Edelsbrunner, L.J. Guibas, M. Sharir. 1991. A Singly Exponential Stratification Scheme for Real Semi-Algebraic Varieties and Its Applications. *Theoretical Computer Science* **84**, 77-105. (doi:10.1016/0304-3975(91)90261-Y)

[25]. Ravi (Ravinder) Prakash G, Kiran M. "How economical are Bounds on Inverted Index Summarization for Calculating Hadoop Channel?" International Journal of Applied Information Systems (IJAIS), Volume 11 – No. 1, June 2016, Pages: 19-35 ISSN ISSN : 2249-0868

[26]. B. Chazelle. 1999. Discrepancy Bounds for Geometric Set Systems with Square Incidence Matrices. *Advances in Discrete and Computational Geometry, Contemporary Mathematics AMS* **223**, 103-107.

[27]. B. Chazelle. 2004. The Discrepancy Method in Computational Geometry. *Handbook of Discrete and Computational Geometry, CRC Press* **44**, 983-996.

[28]. Fadika, Z.; Govindaraju, M. 2010. LEMO-MR: Low Overhead and Elastic MapReduce Implementation Optimized for Memory and CPU-Intensive Applications. *IEEE Second International Conference on Cloud Computing Technology and Science (CloudCom)*, 1-8. (doi:10.1109/CloudCom.2010.45)

[29]. Fadika, Z.; Govindaraju, M. 2011. DELMA: Dynamically Elastic MapReduce Framework for CPU-Intensive Applications. *11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid)*, 454-463. (doi: 10.1109/CCGrid.2011.71)

[30]. Iordache, A.; Morin, C.; Parlavantzas, N.; Feller, E.; Riteau, P. 2013. Resilin: Elastic MapReduce over Multiple Clouds. *13th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid)*, 261-268. (doi:10.1109/CCGrid.2013.48)

[31]. XiaoyongXu; Maolin Tang. 2013. A comparative study of the semi-elastic and fully-elastic mapreduce models. *IEEE International Conference on Granular Computing (GrC)*, 380-385. (doi:10.1109/GrC.2013.6740440)

[32]. Wei Xiang Goh; Kian-Lee Tan. 2014. Elastic MapReduce Execution. *14th IEEE/ACM, International Symposium on Cluster, Cloud and Grid Computing (CCGrid)*, 216-225. (doi:10.1109/CCGrid.2014.14)

[33]. B. Chazelle, W. Mulzer. 2011. Computing Hereditary Convex Structures. *Discrete Comput. Geom*. **45**, 796-823. (doi:10.1007/s00454-011-9346-8)

[34]. B. Chazelle, H. Edelsbrunner, M. Grigni, L.J. Guibas, M. Sharir, E. Welzl. 1995. Improved Bounds on Weak ε-Nets for Convex Sets. *Discrete Comput. Geom*. **13**, 1-15. (doi:10.1007/BF02574025)

[35]. David P. Williamson, David B. Shmoys. 2011. *The Design of Approximation Algorithms*.Cambridge University Press, 978-0-521-19527-0.

[36]. Oded Goldreich. 2008. *Computational Complexity: A Conceptual Perspective*.Cambridge University Press, 978-0-521-88473-0.

[37]. Sanjeev Arora, Boaz Barak. 2009. *Computational Complexity: A Modern Approach*.Cambridge University Press, 978-0-521-42426-4.

[38]. Dimitri P. Bertsekas, Convex Optimization Algorithms, Athena Scientific, Hardcover Edition ISBN: 1-886529-28-0, 978-1-886529-28-1, Publication: February, 2015, 576 pages.
[39]. Ravi (Ravinder) Prakash G, Kiran M. "Does there exist lower bounds on numerical summarization for calculating aggregate resizable Hadoop channel and complexity?" International Journal of Advanced Information Science and Technology, April 2016, Pages: 26-44, ISSN: 2319:2682.
[40]. Patrick Van Flee, (January 2008). Discrete Wavelet Transformations: An Elementary Approach with Applications, 572 pages, ISBN: 978-0-470-18311-3.
[41]. Kevin P. Murphy. 2012. Machine Learning: A Probabilistic Perspective. The MIT Press.
[42]. Koller and Nir Friedman. 2009. Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning. The MIT Press

.

APPENDIX

The purpose of this appendix is to prove necessary and sufficient condition 2.5 on the representation of a hybrid cost function by real versus integer polynomials.

NECESSARY AND SUFFICIENT CONDITION 2.5 (restated). *Let* $f: \{0,1\}^n \to \{-1,+1\}$ *be given. Then for* $d = 0, 1, \dots, n$,

$$\frac{1}{1-E(f,d)} \le W(f,d) \le \frac{2}{1-E(f,d)}\left\{\binom{n}{0} + \binom{n}{1} + \dots + \binom{n}{d}\right\}^{3/2},$$

with the convention that $1/0 = \infty$.

Proof. One readily verifies that $W(f,d) = \infty$ if and only if $E(f,d) = 1$. In what follows, we focus on the complementary case when $W(f,d) < \infty$ and $E(f,d) < 1$. For the lower bound on $W(f,d)$, fix integers λ_S with $\sum_{|S|\le d} |\lambda_S| = W(f,d)$ such that the polynomial $p(x) = \sum_{|S|\le d} \lambda_S X_S(x)$ satisfies $f(x) \equiv \operatorname{sgn} p(x)$.Then $1 \le f(x)p(x) \le W(f,d)$ and therefore

$$E(f,d) \le \left\| f - \frac{1}{W(f,d)} p \right\|_\infty \le 1 - \frac{1}{W(f,d)}.$$

To prove the upper bound on $W(f,d)$, fix any degree-d polynomial p such that $\|f-p\|_\infty = E(f,d)$. Define $\delta = 1 - E(f,d) > 0$ and $N = \sum_{i=0}^{d}\binom{n}{i}$. For a real t, let $\operatorname{rnd} t$ be the result of rounding t to the closest integer, so that $|t - \operatorname{rnd} t| \le 1/2$. We claim that the polynomial

$$q(x) = \sum_{|S|\le d} \operatorname{rnd}\big(M\hat{p}(S)\big) X_S(x),$$

where $M = 3N/(4\delta)$, satisfies $f(x) \equiv \operatorname{sgn} q(x)$. Indeed,

$$\begin{aligned}\left| f(x) - \frac{1}{M} q(x) \right| &\le |f(x) - p(x)| + \frac{1}{M}|Mp(x) - q(x)| \\ &\le 1 - \delta + \frac{1}{M}\sum_{|S|\le d} |M\hat{p}(S) - \operatorname{rnd}\big(M\hat{p}(S)\big)| \\ &\le 1 - \delta + \frac{N}{2M} \\ &< 1.\end{aligned}$$

It remains to examine the sum of the coefficients of q. We have:

$$\begin{aligned}\sum_{|S|\le d} |\operatorname{rnd}\big(M\hat{p}(S)\big)| &\le \frac{1}{2}N + M\sum_{|S|\le d} |\hat{p}(S)| \\ &\le \frac{1}{2}N + M\left(N \mathop{\mathbf{E}}_x [p(x)^2]\right)^{1/2} \\ &\le \frac{2N\sqrt{N}}{\delta},\end{aligned}$$

where the second step follows by an application of the inequality and identity (2.1).

AUTHORS PROFILE

Dr. Ravi (Ravinder) Prakash G. teaches Data Visualization, Networks, Kernel Methods for Pattern Analysis, Geometric Methods for Digital Image Analysis, Industrial Imaging, Statistical and Computational Inverse Problems, (Complementary Metal Oxide Semiconductor) CMOS Circuit Design, Layout, and Simulation, Convex Optimization.