2018 IEEE 59th Annual Symposium on Foundations of Computer Science

Near log-convexity of measured heat in (discrete) time and consequences

Mert Sağlam
University of Washington
saglam@uw.edu

***Abstract*—Let $u, v \in \mathbb{R}_+^{\Omega}$ be positive unit vectors and $S \in \mathbb{R}_+^{\Omega\times\Omega}$ be a symmetric substochastic matrix. For an integer $t \geq 0$, let $m_t = \langle v, S^t u\rangle$, which we view as the heat measured by v after an initial heat configuration u is let to diffuse for t time steps according to S. Since S is entropy improving, one may intuit that m_t should not change too rapidly over time. We give the following formalizations of this intuition.**

We prove that $m_{t+2} \geq m_t^{1+2/t}$, an inequality studied earlier by Blakley and Dixon (also Erdős and Simonovits) for $u = v$ and shown true under the restriction $m_t \geq e^{-4t}$. Moreover we prove that for any $\epsilon > 0$, a stronger inequality $m_{t+2} \geq t^{1-\epsilon} \cdot m_t^{1+2/t}$ holds unless $m_{t+2}m_{t-2} \geq \delta m_t^2$ for some δ that depends on ϵ only. Phrased differently, $\forall\epsilon > 0, \exists\delta > 0$ such that $\forall S, u, v$

$$\frac{m_{t+2}}{m_t^{1+2/t}} \geq \min\left\{t^{1-\epsilon}, \delta\frac{m_t^{1-2/t}}{m_{t-2}}\right\}, \quad \forall t \geq 2,$$

which can be viewed as a truncated log-convexity statement.

Using this inequality, we answer two related open questions in complexity theory: Any property tester for k-linearity requires $\Omega(k \log k)$ queries and the randomized communication complexity of the k-Hamming distance problem is $\Omega(k \log k)$. Further we show that any randomized parity decision tree computing k-Hamming weight has size $\exp(\Omega(k \log k))$.

***Keywords*-communication complexity, property testing**

I. Introduction

Suppose that some initial heat configuration $u\colon \Omega \to \mathbb{R}_+$ is given over a finite space Ω and the configuration evolves according to the map $w \mapsto Sw$ in each time step $t = 0, 1, \ldots$, for some symmetric stochastic matrix $S\colon \Omega \times \Omega \to \mathbb{R}_+$. Assume that we are interested in the amount of heat contained in a certain region $R \subseteq \Omega$ and how this quantity changes over time. In notation, assuming $\|u\|_2 = 1$ for normalization purposes and $v(x) := \mathbb{1}_{x\in R}/|R|^{1/2}$ for $x \in \Omega$, we would like to understand how

$$m_t := \langle v, S^t u\rangle$$

changes as a function of t. In this paper we derive local bounds that $\{m_t\}_{t=0}^{\infty}$ must obey for any S, u and v satisfying the symmetry, magnitude and positivity constraints above (in fact our bounds work for any countable Ω, arbitrary non-negative unit vector v and symmetric non-negative S). Further, we establish a tight connection between such bounds and the well-studied k-Hamming distance problem [1]–[10] and the k-Hamming weight problem [8], [11], [12] and obtain the first tight bounds for respectively the communication complexity and parity decision tree complexity of them.

Our tight $\Omega(k \log(k/\delta))$ lower bound for the δ-error communication complexity of the k-Hamming distance problem (that applies whenever $k^2 < \delta n$) answers affirmatively a conjecture stated in [9] (Conjecture 1.4). Prior to our work, the best impossibility results for this problem were an $\Omega(k \log^{(r)} k)$ bits lower bound ($\log^{(r)}$ being the iterated logarithm) that applies to any randomized r-round communication protocol [13], and an $\Omega(k \log(1/\delta))$ lower bound that applies to any δ-error randomized protocol for $k < \delta n$ [9].

Our parity decision tree lower bound shows that any δ-error parity decision tree solving the k-Hamming weight problem has size $\exp\Omega(k \log(k/\delta))$, which directly implies an $\Omega(k \log(k/\delta))$ bound on the depth of any such decision tree. Previously no nontrivial lower bound was known for the parity decision tree size of this problem and an $\Omega(k \log(1/\delta))$ bound on the parity decision tree depth followed from the communication complexity bound of [9]. Prior to [9], the best bound on the parity decision tree depth was $\Omega(k)$, derived in [7] and [12].

Either by combining our communication complexity lower bound with the reduction technique developed in [7] or by combining our parity decision tree lower bound with a reduction given in [14], one obtains an $\Omega(k \log(k/\delta))$ bound for any (potentially adaptive) property tester for the δ-error probability k-linearity testing problem. This establishes the correct bound for this problem which was studied extensively [7], [8], [12], [14]–[16] since [15] or earlier.

A. Motivating our bounds on m_t

We would like provide some intuition as to why one should expect

$$m_{t+2} \geq m_t^{1+2/t}, \text{ and} \tag{I.1}$$

$$m_{t+2} \geq m_t^{1+2/t} \cdot \min\left\{t^{1-\epsilon}, \delta\frac{m_t^{1-2/t}}{m_{t-2}}\right\} \tag{I.2}$$

to hold for appropriate ϵ, δ. Recalling that S is a symmetric matrix with maximum eigenvalue 1, we may write $S = QDQ^{\intercal}$ for an orthonormal matrix Q having columns q_x, $x \in \Omega$ and a diagonal matrix D with entries $\lambda_x \leq 1$, $x \in \Omega$. Plugging this into $m_t = \langle v, S^t u\rangle$, we get

$$m_t = \sum_{x\in\Omega} \lambda_x^t \langle u, q_x\rangle \langle v, q_x\rangle . \tag{I.3}$$

For sake of analogy let us drop our assumption that S, u, v are coordinate-wise nonnegative for a moment but instead assume that each summand in the right hand side of Eq. (I.3) is nonnegative by some coincidence. In this case we can consider $\{m_t\}_{t=0}^{\infty}$ as the moment sequence of a random variable supported on $[0, 1]$ that takes the value $|\lambda_x|$ with probability $|\langle u, q_x\rangle \langle v, q_x\rangle|$ and the value 0 with probability $1 - \sum_x |\langle u, q_x\rangle \langle v, q_x\rangle|$ (which is nonnegative by Cauchy-Schwarz inequality). This would imply that $\{m_t\}_{t=0}^{\infty}$ is *completely monotone* by Hausdorff's characterization [17] and therefore log-convex (e.g., [18], Section 2.1, Example 6).

One particular implication of the log-convexity of $\{m_t\}_{t=0}^{\infty}$, that $\frac{1}{t}\log m_t + \frac{t-1}{t}\log m_0 \geq \log m_1$, when combined with the fact $0 \leq m_0 \leq 1$ (that follows from our assumption on the terms of Eq. (I.3)), leads to $m_t \geq m_1^t$. In 1958, Mandel and Hughes showed

2575-8454/18/$31.00 ©2018 IEEE
DOI 10.1109/FOCS.2018.00095

that if $u = v$, rather surprisingly, one can trade the assumption that the summands of Eq. (I.3) are nonnegative with the assumption that S and $u = v$ are coordinate-wise nonnegative and still obtain the conclusion $m_t \geq m_1^t$:

Theorem I.1 (Mandel and Hughes [19])**.** *Let u be a nonnegative unit vector and S be a symmetric matrix with nonnegative entries. For an integer*[1] *$t \geq 1$ we have $\langle u, S^t u\rangle \geq \langle u, Su\rangle^t$.*

A more general implication of the log-convexity of $\{m_t\}_{t=0}^{\infty}$ and that $m_0 \leq 1$ is that for $k \geq t$, $\frac{t}{k}\log m_k + \frac{k-t}{k}\log m_0 \geq \log m_t$, therefore $m_k^t \geq m_t^k$. In 1966, Blakley and Dixon [20] investigated whether $m_k^t \geq m_t^k$ holds in the case $u = v$ when the nonnegativity assumption on the summands of Eq. (I.3) is replaced by the coordinate-wise nonnegativity of S, $u = v$. They note that the inequality $m_k^t \geq m_t^k$ fails when k and t have different parity and otherwise holds true under the restriction $m_t \geq e^{-4t}$. While the following is not explicitly stated as a conjecture in [20], they write

> if $t > 1$, [...] we cannot show that the inequality Eq. (I.1) holds for each nonnegative $|\Omega|$-vector u if S is nonnegative.

so with the earlier caveat we attribute the following to Blakley and Dixon [20]:

Conjecture I.2 (Blakley and Dixon [20])**.** *Let $S\colon \Omega \times \Omega \to \mathbb{R}_+$ be a symmetric matrix with nonnegative entries and let $u\colon \Omega \to \mathbb{R}_+$ be a nonnegative unit vector. For positive integers $k \geq t$ of the same parity, we have*

$$\left\langle u, S^k u\right\rangle^t \geq \langle u, S^t u\rangle^k .$$

In Section III we prove the following theorem which shows that a generalization of Conjecture I.2 holds true.

Theorem I.3. *Let $S\colon \Omega \times \Omega \to \mathbb{R}_+$ be a symmetric matrix with nonnegative entries and $u, v\colon \Omega \to \mathbb{R}_+$ be nonnegative unit vectors. For positive integers $k \geq t$ of the same parity, we have*

$$\left\langle v, S^k u\right\rangle^t \geq \langle v, S^t u\rangle^k .$$

It goes without saying that Eq. (I.1) is equivalent to Theorem I.3 as we can rearrange Eq. (I.1) to $m_{t+2}^{1/(t+2)} \geq m_t^{1/t}$ and apply it iteratively to obtain $m_k^{1/k} \geq \cdots \geq m_{t+2}^{1/(t+2)} \geq m_t^{1/t}$ whenever $k \geq t$ and k, t have the same parity. Moreover, while defining Eq. (I.1) we assumed S to be substochastic only to illustrate our interpretation of the inequality: indeed any nonnegative S can be scaled to be substochastic as both sides of Eq. (I.1) are $(t+2)$-homogeneous in S.

In Theorem I.1 and Theorem I.3 we observed that increasingly more general implications of the log-convexity of $\{m_t\}_{t=0}^{\infty}$ can be derived by only assuming the coordinate-wise nonnegativity of S, u and v. One may naturally wonder if the coordinate-wise nonnegativity of S, u and v implies the log-convexity of $\{m_t\}_{t=0}^{\infty}$ in its entirety. Unfortunately the following example shows that this is far from the truth.

[1]Since $u = v$ here, the summands inside Eq. (I.3) are nonnegative when t is even so this theorem is most interesting for t odd.

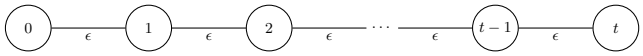

Figure 1. $\Omega = \{0, 1, \ldots, t\}$, $S(i, i+1) = S(i+1, i) = \epsilon$ for $i = 0, \ldots, t-1$.

Consider the transition matrix S on $\Omega = \{0, 1, \ldots, t\}$ such that $S(i, i+1) = S(i+1, i) = \epsilon$ for $i = 0, 1, \ldots, t-1$ and $S(i, j) = 0$ otherwise. Let u and v be the point masses respectively on states 0 and t; namely $u = [1, 0, \ldots, 0]^\intercal$ and $v = [0, 0, \ldots, 1]^\intercal$. We have $m_{t-2} = 0$, $m_t = \epsilon^t$ and $m_{t+2} = t\epsilon^{t+2}$. Therefore $m_{t-2}m_{t+2} = 0 \not\geq \epsilon^{2t} = m_t^2$. In this example the log-convexity breaks (in the strongest possible way) because the states 0 and t are separated by t hops according to S and the point mass at state 0 cannot reach state t before the tth time step.

Our next theorem shows that such reachability issues are essentially the only way the log-convexity property can fail to hold:

Theorem I.4. *For every $\epsilon > 0$ there is a $\delta > 0$ such that for any symmetric matrix $S\colon \Omega \times \Omega \to \mathbb{R}_+$ and unit vectors $u, v\colon \Omega \to \mathbb{R}_+$ with nonnegative entires, defining m_t as before, we have*

$$\frac{m_{t+2}}{m_t^{1+2/t}} \geq \min\left\{t^{1-\epsilon}, \delta\frac{m_t^{1-2/t}}{m_{t-2}}\right\}, \qquad \forall t \geq 2. \tag{I.4}$$

In other words, Theorem I.4 shows that one can recover a truncated version of the log-convexity of $\{m_t\}_{t=0}^{\infty}$ from just the coordinate-wise nonnegativity assumption of S, u and v. We stress that Theorem I.4 is tight up to the appearance of ϵ and the choice of $\delta = \delta(\epsilon)$. A direct calculation on Figure 1 for time steps $t, t+2, t+4$ shows that Eq. (I.4) cannot be improved to

$$\frac{m_{t+2}}{m_t^{1+2/t}} \geq \min\left\{t^{1-2/t}, \left(\frac{1+\eta}{2}\right)\frac{m_t^{1-2/t}}{m_{t-2}}\right\}$$

for $\eta > 0$.

B. Related work on m_t

Almost simultaneously with the work of Mandel and Hughes [19], Mulholland and Smith also prove Theorem I.1 in [21] and moreover they characterize the equality conditions of the inequality. Independently, in 1965, Blakley and Roy [22] prove the same inequality and characterize the equality conditions and [23] provides an alternative proof to that of [21] in 1966. We remark that Theorem I.1 is most commonly referred to as the Blakley-Roy bound or "Sidorenko's conjecture for paths". Note these results show that Conjecture I.2 is true whenever t divides k. Finally in 2012, Pate shows that $m_t \geq m_1^t$ without the restriction $u = v$:

Theorem I.5 (Pate [24])**.** *Let $S\colon \Omega \times \Omega \to \mathbb{R}_+$ be a symmetric matrix with nonnegative entries and let $u, v\colon \Omega \to \mathbb{R}_+$ be nonnegative unit vectors. It holds that*

$$\left\langle v, S^{2t+1}u\right\rangle \geq \langle v, Su\rangle^{2t+1},$$

with equality if and only if $\left\langle v, S^{2t+1}u\right\rangle = 0$ or $Su = \lambda v$ and $Sv = \lambda u$ for some $\lambda \in \mathbb{R}_+$

This result already shows that Theorem I.3 is true when t divides k but such a bound does not have any implications for our applications

in complexity theory. In [25], Erdős and Simonovits conjecture the following.

Conjecture I.6 (Erdős and Simonovits [25], Conjecture 6)**.** *For a graph $G = (V, E)$, let $w_k(G)$ be the number length k walks in G divided by $|V|$. For an undirected graph G, we have $w_k(G)^t \geq w_t(G)^k$ for $k > t$ of the same parity.*

Note that Conjecture I.6 is a specialization of Conjecture I.2 to S having 0-1 entries and $u = \mathbf{1}/\sqrt{|V|}$ therefore our Theorem I.3 verifies Conjecture I.6 as well.

C. Our results in complexity theory

Here we list our results in complexity theory; see Section V for the definition of the models and the problems. The following theorem (which was already known [9]) is a consequence of Theorem I.3 and uses the standard corruption technique in communication complexity.

Theorem I.7. *Any two party δ-error randomized protocol solving the k-Hamming distance problem over length-n strings communicates at least $\Omega(k \log(1/\delta))$ bits for $k^2 \leq \delta n$.*

The next is our main result for the communication complexity of the k-Hamming distance problem and is a consequence of Theorem I.4. This result cannot be obtained by the standard corruption technique and requires a suitable modification similar to [26].

Theorem I.8. *Any two party δ-error randomized protocol solving the k-Hamming distance problem over length-n strings communicates at least $\Omega(k \log(k/\delta))$ bits for $k^2 \leq \delta n$.*

Theorem I.9. *Any δ-error parity decision tree deciding the k-Hamming weight predicate over length-n strings has size $\exp \Omega\left(k \log(k/\delta)\right)$ for $k^2 < \delta n$.*

Corollary I.10. *Any δ-error probability property tester for k-linearity requires $\Omega(k \log(k/\delta))$ queries.*

Note the bound $m_k^t \geq m_t^k$ obtained in [20] under the condition $m_t \geq e^{-4t}$ does not have any implications for the communication complexity of the k-Hamming distance problem as our reduction crucially uses the fact that u and v are arbitrary, however it does lead to an $\exp \Omega(k)$ lower bound for the parity decision tree size of the k-Hamming weight problem when combined with our reduction.

Remark. Note that in Theorem I.3, when $u = v$ and either both k, t are even or S is positive semidefinite, the summands of Eq. (I.3) become nonnegative and the inequality holds trivially. For our application in communication complexity we crucially use the fact that u and v are arbitrary and for our application in parity decision trees, one can do away with $u = v$ but only at the expense of having to choose k, t odd. In both results the S we choose has eigenvalues 1 and -1 with equal multiplicities and therefore far away from being positive semidefinite. In either case, the implications of Theorem I.3 in complexity theory follow from the interesting cases of this theorem.

II. Preliminaries

We denote by $[n]$ the set $\{1, 2, \ldots, n\}$. We take exp and log functions to the base 2. Let Ω be a countable set. For a function $\mu \colon \Omega \to \mathbb{R}_+$ and a set $\Psi \subseteq \Omega$, we use the shorthand

$$\mu(\Psi) := \sum_{x \in \Psi} \mu(x).$$

A function $\mu \colon \Omega \to \mathbb{R}_+$ is said to be a distribution on Ω if $\mu(\Omega) = 1$ and a subdistribution if $\mu(\Omega) \leq 1$. For a function μ on Ω, we define

$$\operatorname{supp}(\mu) := \{x \in \Omega \mid \mu(x) > 0\}.$$

For two distributions $\mu \colon \Omega_1 \to \mathbb{R}_+$ and $\nu \colon \Omega_2 \to \mathbb{R}_+$, let us denote by $\mu\nu$ the distribution on $\Omega_1 \times \Omega_2$ given by $(\mu\nu)(x_1, x_2) = \mu(x_1)\nu(x_2)$.

For a discrete random variable X, we denote by $\operatorname{dist}(X)$ the distribution function of X and we define $\operatorname{supp}(X) := \operatorname{supp}(\operatorname{dist}(X))$. If X is so that $\operatorname{dist}(X) \colon \Omega \to \mathbb{R}_+$, then we say that X has sample space Ω. Two random variables X and Y are said to be independent if $\operatorname{dist}(XY) = \operatorname{dist}(X) \operatorname{dist}(Y)$.

Lemma II.1 (Jensen [27], Formula (5))**.** *Let X be a real-valued random variable and f be a convex function. We have $\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$. When f is strictly convex, the inequality holds with equality if and only if X is constant with probability 1.*

A. Information theory

In this section we review the definitions and facts we use from information theory. Let μ and ν be two nonnegative functions on Ω. The Kullback-Leibler divergence [28], [29] of μ from ν, denoted $\mathbf{D}(\mu \,\|\, \nu)$, is defined by

$$\mathbf{D}(\mu \,\|\, \nu) := \sum_{x \in \Omega} \mu(x) \log \frac{\mu(x)}{\nu(x)}. \tag{II.1}$$

Here, if $\mu(x) = 0$ for some x, then its contribution to the summation is taken as 0, even when $\nu(x) = 0$. The divergence is undefined if there is an $x \in \Omega$ such that $\mu(x) > 0$ and $\nu(x) = 0$. It can be shown that if the related series converges for the right hand side of Eq. (II.1), it converges absolutely, which justifies leaving the summation order unspecified. A fundamental property of $\mathbf{D}(\cdot \,\|\, \cdot)$ is that the divergence of a distribution from a subdistribution is always nonnegative.

Lemma II.2 (Kullback and Leibler [29], Lemma 3.2)**.** *Let $\mu, \nu \colon \Omega \to \mathbb{R}_+$ be so that μ is a distribution on Ω and $\operatorname{supp}(\mu) = \Psi \subseteq \Omega$. We have*

$$\mathbf{D}(\mu \,\|\, \nu) \geq -\log \nu(\Psi)$$

with equality if an only if $\mu(x) = \nu(x)/\nu(\Psi)$ for $x \in \Psi$ and $\mu(x) = 0$ for $x \notin \Psi$.

We extend the divergence notation $\mathbf{D}(\cdot \,\|\, \cdot)$ to apply to random variables as follows. Let X, Y be discrete random variables on the same sample space Ω. Define

$$\mathbf{D}(X \,\|\, Y) := \mathbf{D}(\operatorname{dist}(X) \,\|\, \operatorname{dist}(Y)). \tag{II.2}$$

With this notation in hand, we are ready to define the conditional divergence. Let X_1X_2 and Y_1Y_2 be random variables defined on the sample space $\Omega_1 \times \Omega_2$. The divergence of $X_1 \mid X_2$ from $Y_1 \mid Y_2$ is defined by

$$\mathbf{D}(X_1 \mid X_2 \,\|\, Y_1 \mid Y_2) := \mathop{\mathbb{E}}_{x_2 \sim X_2} \mathbf{D}(X_1 \mid X_2 = x_2 \,\|\, Y_1 \mid Y_2 = x_2). \tag{II.3}$$

Here, for each $x_2 \in \mathrm{supp}(X_2)$, $X_1 \,|\, X_2 = x_2$ and $Y_1 \,|\, Y_2 = x_2$ are random variables on the sample space Ω_1 obtained from, respectively X_1X_2 and Y_1Y_2, by conditioning on the second coordinate equaling x_2.

Lemma II.3 (e.g., [30])**.** *Let X_1X_2 and Y_1Y_2 be random variables, both on the sample space $\Omega_1 \times \Omega_2$. We have*

$$\mathbf{D}(X_1X_2 \,\|\, Y_1Y_2) = \mathbf{D}(X_1 \,\|\, Y_1) + \mathbf{D}(X_2 \,|\, X_1 \,\|\, Y_2 \,|\, Y_1)\,.$$

The mutual information: Let X and Y be jointly distributed random variables. The mutual information of X and Y, denoted $\mathbf{I}(X:Y)$, is defined as

$$\mathbf{I}(X:Y) := \mathbf{D}(\mathrm{dist}(X,Y) \,\|\, \mathrm{dist}(X)\,\mathrm{dist}(Y))\,. \tag{II.4}$$

The mutual information of a random variable with itself, i.e., the quantity $\mathbf{I}(X:X)$ is called the Shannon entropy of X.

III. Monotonicity of $t \mapsto m_{2t}^{1/(2t)}$ and $t \mapsto m_{2t+1}^{1/(2t+1)}$

Let $S\colon \Omega \times \Omega \to \mathbb{R}_+$ be a symmetric matrix with nonnegative entries, $u, v\colon \Omega \to \mathbb{R}_+$ nonnegative unit vectors and $m_t = \langle v, S^t u\rangle$. In this section we prove Theorem I.3 which we restate here (with additional equality conditions) for the convenience of the reader. Recall this theorem confirms Conjecture I.2 and Conjecture I.6.

Theorem I.3 (restated)**.** *Let $S\colon \Omega \times \Omega \to \mathbb{R}_+$ be a symmetric matrix with nonnegative entries and $u, v\colon \Omega \to \mathbb{R}_+$ be nonnegative unit vectors. For positive integers $k \geq t$ of the same parity, we have*

$$\left\langle v, S^k u\right\rangle^t \geq \langle v, S^t u\rangle^k\,, \tag{III.1}$$

with equality if and only if $\langle v, S^k u\rangle = 0$ or $Su = \lambda v$ and $Sv = \lambda u$ for some $\lambda \in \mathbb{R}_+$ when t is odd and $u = v$ is an eigenvector of S^2 when t is even.

We prove Theorem I.3 by an information theoretic argument. Define the distributions $\mu := u/\|u\|_1$ and $\nu := v/\|v\|_1$. Since either side of Eq. (III.1) is kt-homogeneous in S, we may assume that S is substochastic by scaling as needed. Having fixed this normalization, we view Eq. (III.1) as a statement about random walks on Ω that start from a state sampled according to μ or ν and evolve according to the transition matrix S.

A. Reference random walks

Let $\Omega_\circ = \Omega \cup \{r\}$ for some state $r \notin \Omega$ and t be a positive integer. Recall that $\mu = u/\|u\|_1$ and $\nu = v/\|v\|_1$. We start by defining random walks F^t, B^t on $\Omega_\circ$ that evolve in discrete time steps $-1, 0, 1, \ldots, t, t+1$.

The random walk F^t starts at r and transitions to a state $x \in \Omega$ with probability $\mu(x)$ at time step -1. In steps $0, 1, \ldots, t-1$, the random walk proceeds according to the transition matrix S. At the time step t, each state $x \in \Omega$ transitions to r with probability $\nu(x)$ and transitions to an arbitrary state in Ω with probability $1 - \nu(x)$ (say, all of them to the same arbitrary state). We view F^t as a joint random variable $F^t = (F^t_{-1}, F^t_0, \ldots, F^t_{t+1})$, where F^t_i is the location of the walk in time step i.

The random walk B^t proceeds backwards in time. At time step $t+1$ the walk B^t starts at r and transitions to a state $x \in \Omega$ with probability $\nu(x)$. In time steps $t, t-1, \ldots, 1$, the random walk proceeds as prescribed by S. At time step 0, each state $x \in \Omega$ transitions to r with probability $\mu(x)$ and to an arbitrary state in Ω with probability $1 - \mu(x)$. Similarly, B^t denotes the joint random variable $B^t = (B^t_{-1}, B^t_0, \ldots, B^t_{t+1})$, where B^t_i is the location of the walk at time step i.

The following facts about F^t and B^t are immediate. The random variables F^t_{-1} and B^t_{t+1} are fixed to a single value r. The random variables F^t, B^t are Markovian, namely, $\mathrm{dist}(F^t_i \,|\, F^t_{i-1}, \ldots, F^t_{-1}) = \mathrm{dist}(F^t_i \,|\, F^t_{i-1})$ and $\mathrm{dist}(B^t_{i-1} \,|\, B^t_i, \ldots B^t_{t+1}) = \mathrm{dist}(B^t_{i-1} \,|\, B^t_i)$ for $i \in \{0, \ldots, t+1\}$.

B. Random walks returning to the origin

Assume that $\Pr[F^t_{t+1} = r] > 0$. Let X be the walk F^t conditioned on $F^t_{t+1} = r$. Note that X is a random variable on the sample space $\Omega_\circ^{t+3}$. The next two lemmas explicitly calculate the distribution of X.

For a matrix $M\colon \Omega \times \Omega \to \mathbb{R}_+$, functions $f, g\colon \Omega \to \mathbb{R}_+$, and $x, y \in \Omega$ we use the shorthands

$$\begin{aligned} M(f,y) &:= \sum_{x\in\Omega} f(x)M(x,y) = (M^\intercal f)(y)\\ M(x,g) &:= \sum_{y\in\Omega} M(x,y)g(y) = (Mg)(x)\\ M(f,g) &:= \sum_{x,y\in\Omega} f(x)M(x,y)g(y) = f^\intercal M g, \end{aligned}$$

where the last expression in each line is understood as a matrix vector multiplication.

Lemma III.1. *Under our assumption $S^t(\mu,\nu) > 0$,*

(i) we have $\Pr[X_i = x] = \frac{S^i(\mu,x)S^{t-i}(x,\nu)}{S^t(\mu,\nu)}$, and
(ii) if $S^{t-i}(x,\nu) > 0$, we have $\Pr[X_{i+1} = y \,|\, X_{\leq i} = x_{\leq i}] = \frac{S(x_i,y)S^{t-i-1}(y,\nu)}{S^{t-i}(x_i,\nu)}$.

With Lemma III.1 we confirm that the random variable $X = (X_{-1}, X_0, \ldots, X_{t+1})$ is Markovian; in particular a time inhomogeneous random walk on $\Omega_\circ$. Next we observe that the random variable B^t conditioned on $B^t_{-1} = r$ is precisely X also.

Lemma III.2. *Under our assumption $S^t(\mu,\nu) > 0$,*

(i) we have $\mathrm{dist}(X) = \mathrm{dist}(B^t \,|\, B^t_{-1} = r)$, and
(ii) if $S^i(\mu,x) > 0$, we have $\Pr[X_{i-1} = y \,|\, X_{\geq i} = x_{\geq i}] = \frac{S(x_i,y)S^{i-1}(y,\mu)}{S^i(x_i,\mu)}$.

Lemma III.3. *We have $\mathbf{D}\left(X \,\middle\|\, F^t\right) = -\log S^t(\mu,\nu) = \mathbf{D}\left(X \,\middle\|\, B^t\right)$.*

Proof: Recall that $\Pr[F^t_{t+1} = r] = S^t(\mu,\nu)$. Since X is obtained from F^t by conditioning on $F^t_{t+1} = r$, the equality criteria of Lemma II.2 are fulfilled and thus $\mathbf{D}\left(X \,\middle\|\, F^t\right) = -\log S^t(\mu,\nu)$. The derivation of $\mathbf{D}\left(X \,\middle\|\, B^t\right)$ is identical as per Lemma III.2(i). ∎

C. Longer random walks

Let J be an integer valued random variable taking the values $\{1, 2, \ldots, t\}$, each with equal probability. For each fixing j of J we perform a random walk $Z \,|\, J = j$ on $\Omega_\circ$ that evolves in time steps $-1, 0, 1, \ldots, t, t+1, t+2, t+3$ as follows.

The random walk starts at r and for each time step $-1 \leq i < j$, proceeds according to the transition kernel $\mathrm{dist}(X_{i+1} \,|\, X_i)$. At time

step j, the random walk proceeds according to $\mathrm{dist}(X_{j-1} \mid X_j)$ and in time steps $j < i \le t+3$ proceeds according to the transition kernel $\mathrm{dist}(X_{i-1} \mid X_{i-2})$. We view Z as a joint random variable $Z = (Z_{-1}, Z_0, \ldots, Z_{t+3})$, where Z_i denotes the location of the random walk at time step i.

Lemma III.4. *For $-1 \le i \le j$, we have $\mathrm{dist}(Z_i \mid J = j) = \mathrm{dist}(X_i)$ and for $j < i \le t+3$, $\mathrm{dist}(Z_i \mid J = j) = \mathrm{dist}(X_{i-2})$.*

Proof: This follows from the fact that

$$\mathrm{dist}(F^t \mid F^t_{t+1} = r) = \mathrm{dist}(X) = \mathrm{dist}(B^t \mid B^t_{-1} = r)$$

and that X is an actual random walk (i.e., Markovian) on $\Omega_\circ$. ∎

From this we can deduce that Z always ends up in r at time step $t+3$. We next argue that if X does not diverge too much from the reference random walk F^t, then Z does not diverge too much from F^{t+2}.

Lemma III.5. *We have*

$$\begin{aligned}\mathbf{D}\left(Z \mid J \,\middle\|\, F^{t+2}\right) &= \frac{t+2}{t}\,\mathbf{D}\left(X \,\middle\|\, F^t\right)\\ &\quad - \frac{1}{t}\left(\mathbf{D}\left(X_0 \,\middle\|\, F^t_0\right) + \mathbf{D}\left(X_{t+1} \mid X_t \,\middle\|\, F^t_{t+1} \mid F^t_t\right)\right)\\ &\quad - \frac{1}{t}\left(\mathbf{D}\left(X_t \,\middle\|\, B^t_t\right) + \mathbf{D}\left(X_{-1} \mid X_0 \,\middle\|\, B^t_{-1} \mid B^t_0\right)\right).\end{aligned}$$

Proof: For a fixing j of J, we have

$$\begin{aligned}\mathbf{D}\left(Z \mid J = j \,\middle\|\, F^{t+2}\right) &= \sum_{i=-1}^{j-1} \mathbf{D}\left(X_{i+1} \mid X_i \,\middle\|\, F^{t+2}_{i+1} \mid F^{t+2}_i\right)\\ &\quad + \mathbf{D}\left(X_{j-1} \mid X_j \,\middle\|\, F^{t+2}_{j+1} \mid F^{t+2}_j\right)\\ &\quad + \sum_{i=j+1}^{t+2} \mathbf{D}\left(X_{i-1} \mid X_{i-2} \,\middle\|\, F^{t+2}_{i+1} \mid F^{t+2}_i\right),\end{aligned}$$

where we have used the chain rule for divergence (cf. Lemma II.3), the fact that $Z \mid J = j$ and F^{t+2} are Markovian and Lemma III.4. Recalling that F^t and B^t evolve according to S in time steps $0, 1, \ldots, t-1$, and $\mathrm{dist}(F^t_{t+1} \mid F^t_t) = \mathrm{dist}(F^{t+2}_{t+3} \mid F^{t+2}_{t+2})$, we write

$$\begin{aligned}&\mathbf{D}\left(Z \mid J = j \,\middle\|\, F^{t+2}\right) = \sum_{i=-1}^{t} \mathbf{D}\left(X_{i+1} \mid X_i \,\middle\|\, F^t_{i+1} \mid F^t_i\right)\\ &\quad + \mathbf{D}\left(X_{j-1} \mid X_j \,\middle\|\, B^t_{j-1} \mid B^t_j\right) + \mathbf{D}\left(X_j \mid X_{j-1} \,\middle\|\, F^t_j \mid F^t_{j-1}\right)\\ &= \mathbf{D}\left(X \,\middle\|\, F^t\right) + \mathbf{D}\left(X_{j-1} \mid X_j \,\middle\|\, B^t_{j-1} \mid B^t_j\right)\\ &\qquad + \mathbf{D}\left(X_j \mid X_{j-1} \,\middle\|\, F^t_j \mid F^t_{j-1}\right)\end{aligned}$$

again by the chain rule for divergence (Lemma II.3) and the fact that X and F^t are Markovian. Now taking an expectation over all $j \in \mathrm{supp}(J)$, we have

$$\begin{aligned}&\mathbf{D}\left(Z \mid J \,\middle\|\, F^{t+2}\right) = \frac{1}{t}\sum_j \mathbf{D}\left(Z \mid J = j \,\middle\|\, F^{t+2}\right)\\ &= \mathbf{D}\left(X \,\middle\|\, F^t\right) + \frac{1}{t}\sum\left(\mathbf{D}\left(X_{j-1} \mid X_j \,\middle\|\, B^t_{j-1} \mid B^t_j\right)\right)\\ &\qquad + \frac{1}{t}\sum\left(\mathbf{D}\left(X_j \mid X_{j-1} \,\middle\|\, F^t_j \mid F^t_{j-1}\right)\right)\\ &= \mathbf{D}\left(X \,\middle\|\, F^t\right) + \frac{1}{t}\,\mathbf{D}\left(X \,\middle\|\, B^t\right) - \frac{1}{t}\,\mathbf{D}\left(X_t \,\middle\|\, B^t_t\right)\\ &- \frac{1}{t}\,\mathbf{D}\left(X_{-1} \mid X_0 \,\middle\|\, B^t_{-1} \mid B^t_0\right) + \frac{1}{t}\,\mathbf{D}\left(X \,\middle\|\, F^t\right) - \frac{1}{t}\,\mathbf{D}\left(X_0 \,\middle\|\, F^t_0\right)\\ &- \frac{1}{t}\,\mathbf{D}\left(X_{t+1} \mid X_t \,\middle\|\, F^t_{t+1} \mid F^t_t\right).\end{aligned}$$

Since $\mathbf{D}\left(X \,\middle\|\, B^t\right) = \mathbf{D}\left(X \,\middle\|\, F^t\right)$ by Lemma III.3, collecting the $\mathbf{D}\left(X \,\middle\|\, F^t\right)$ terms we finish the proof. ∎

Finally, we lower bound the negative terms in the statement of Lemma III.5.

Lemma III.6. *We have*

$$\begin{aligned}&\mathbf{D}\left(X_0 \,\middle\|\, F^t_0\right) + \mathbf{D}\left(X_{-1} \mid X_0 \,\middle\|\, B^t_{-1} \mid B^t_0\right) && \text{(III.2)}\\ &\qquad \ge \mathrm{H}_2(\mu) := -\log\|\mu\|_2^2, \text{ and} && \text{(III.3)}\\ &\mathbf{D}\left(X_t \,\middle\|\, B^t_t\right) + \mathbf{D}\left(X_{t+1} \mid X_t \,\middle\|\, F^t_{t+1} \mid F^t_t\right) && \text{(III.4)}\\ &\qquad \ge \mathrm{H}_2(\nu) := -\log\|\nu\|_2^2, && \text{(III.5)}\end{aligned}$$

where $\mathrm{H}_2(\cdot)$ denotes the second order Rényi entropy.

Proof: We only prove the first inequality as the second one is symmetric. By Lemma III.1, we have

$$\mathbf{D}\left(X_0 \,\middle\|\, F^t_0\right) = \sum_{x\in\Omega} \frac{\mu(x)S^t(x,\nu)}{S^t(\mu,\nu)} \log\frac{S^t(x,\nu)}{S^t(\mu,\nu)} \quad \text{and}$$

$$\mathbf{D}\left(X_{-1} \mid X_0 \,\middle\|\, B^t_{-1} \mid B^t_0\right) = \sum_{x\in\Omega} \frac{\mu(x)S^t(x,\nu)}{S^t(\mu,\nu)} \log\frac{1}{\mu(x)}.$$

Let $\Psi = \mathrm{supp}(X_0)$. By adding the two terms we get

$$\begin{aligned}&\mathbf{D}\left(X_0 \,\middle\|\, F^t_0\right) + \mathbf{D}\left(X_{-1} \mid X_0 \,\middle\|\, B^t_{-1} \mid B^t_0\right) && \text{(III.6)}\\ &\quad = -\sum_{x\in\Psi} \frac{\mu(x)S^t(x,\nu)}{S^t(\mu,\nu)} \log\frac{\mu(x)S^t(\mu,\nu)}{S^t(x,\nu)}\\ &\quad \ge -\log\sum_{x\in\Psi} \frac{\mu(x)^2 S^t(x,\nu)S^t(\mu,\nu)}{S^t(\mu,\nu)S^t(x,\nu)} && \text{(III.7)}\\ &\quad = -\log\sum_{x\in\Psi}\mu(x)^2\\ &\quad \ge -\log\sum_{x\in\Omega}\mu(x)^2, && \text{(III.8)}\end{aligned}$$

where the first inequality is by concavity of $z \mapsto \log z$ and the second inequality is true as the summands are nonnegative. ∎

D. Combining the inequalities

Proof of Theorem I.3: Note that Z_{-1} is fixed to r by definition (cf. Section III-C) and Z_{t+3} is fixed to r by Lemma III.4. Therefore by Lemma II.2 we have

$$\begin{aligned}-\log S^{t+2}(\mu,\nu) &\le \mathbf{D}\left(Z \,\middle\|\, F^{t+2}\right) && \text{(III.9)}\\ &= \mathbf{D}\left(Z \mid J \,\middle\|\, F^{t+2}\right) - \mathbf{I}\left(J : Z\right) && \text{(III.10)}\\ &\le \mathbf{D}\left(Z \mid J \,\middle\|\, F^{t+2}\right) && \text{(III.11)}\\ &\le \frac{t+2}{t}\,\mathbf{D}\left(X \,\middle\|\, F^t\right) + \frac{\log\|\mu\|_2^2 + \log\|\nu\|_2^2}{t}.\end{aligned}$$

Here Eq. (III.10) follows from the chain rule for the divergence (Lemma II.3) and the definition of mutual information (cf. Eq. (II.4)), Eq. (III.11) follows from the nonnegativity of mutual information and the last line follows from Lemma III.5 and Lemma III.6. Plugging in $\mathbf{D}\left(X \,\middle\|\, F^t\right) = -\log S^t(\mu,\nu)$, provided by Lemma III.3, we obtain

$$-\log S^{t+2}(\mu,\nu) \le -\frac{t+2}{t}\log S^t(\mu,\nu) + \frac{\log\|\mu\|_2^2 + \log\|\nu\|_2^2}{t}.$$

Arranging, we get

$$\|\mu\|_2^2\,\|\nu\|_2^2\left\langle\nu, S^{t+2}\mu\right\rangle^t \ge \left\langle\nu, S^t\mu\right\rangle^{t+2}$$

and substituting $\mu = u/\|u\|_1$, $\nu = v/\|v\|_1$, and recalling that u, v are unit vectors, we obtain

$$\langle v, S^{t+2}u\rangle^t \geq \langle v, S^t u\rangle^{t+2}, \text{ i.e.,}$$
$$m_{t+2} \geq m_t^{1+2/t}. \tag{III.12}$$

By applying this inequality iteratively, we get $\langle v, S^k u\rangle^t \geq \langle v, S^t u\rangle^k$ or written differently $m_k^{1/k} \geq m_t^{1/t}$ as long as $k > t$ and k, t have the same parity. ∎

IV. Near log-convexity of $t \mapsto m_{2t}$ and $t \mapsto m_{2t+1}$

In this section we would like to prove the following improvement to Eq. (III.12): for all $\epsilon > 0$ there exists a $\delta > 0$ such that

$$m_{t+2} \geq m_t^{1+2/t} \cdot \min\left\{t^{1-\epsilon}, \left\lceil \delta \frac{m_t^{1-2/t}}{m_{t-2}} \right\rceil\right\}, \quad \forall t \geq 2. \tag{IV.1}$$

Recall that in proving Eq. (III.12), in line (III.11), we used the relaxation $\mathbf{I}(J : Z) \geq 0$. Note that J is uniformly distributed on $[t]$ therefore has $\log t$ bits of entropy and provided that it is possible to infer J from Z (i.e., it is possibly to locate the time reversal we have inserted in Z) the $\mathbf{I}(J : K)$ term appears to be large enough to recover the factor

$$\min\left\{t^{1-\epsilon}, \left\lceil \delta \frac{m_t^{1-2/t}}{m_{t-2}} \right\rceil\right\}.$$

Note moreover that intuitively we are able to infer J from Z better when $\frac{m_t^{1-2/t}}{m_{t-2}}$ is high, as in such cases on average for a time step $i \in [t]$ and a typical $x \sim X_i$, the distributions $\mathrm{dist}(X_{i-1} \,|\, X_i = x)$ and $\mathrm{dist}(X_{i+1} \,|\, X_i = x)$ should be far from each other, as otherwise we can argue that there should be many $t-2$ walks as follows. If $\mathrm{dist}(X_{i-1} \,|\, X_i = x)$ and $\mathrm{dist}(X_{i+1} \,|\, X_i = x)$ are close to each other, there should be many $p \in \Omega$ which has high probability in both these distributions. Sample such a p, and attach to it a walk sampled from $X_{-1}X_1 \dots X_{i-1} \,|\, X_{i-1} = p$ and another walk sampled from $X_{i+1}X_{i+2} \dots X_{t+1} \,|\, X_{i+1} = p$, which leads to a length $t-2$ walk returning to the origin. However if m_{t-2} is low, this should not happen and therefore $\mathrm{dist}(X_{i-1} \,|\, X_i = x)$ and $\mathrm{dist}(X_{i+1} \,|\, X_i = x)$ on average should be far apart, which means that we can notice when we take a step backwards in time and therefore infer J. In particular, Figure 1 gives such an example where $m_{t-2} = 0$ and we can always recover J with certainty from a sample from Z: whenever we take a step to the left, it must be that we are at time step J.

Given this discussion, a direct approach to proving Eq. (IV.1) appears to bound

$$\mathbf{I}(Z : J) \geq \log \min\left\{t^{1-\epsilon}, \left\lceil \delta \frac{m_t^{1-2/t}}{m_{t-2}} \right\rceil\right\}. \tag{IV.2}$$

Unfortunately, this approach does not seem to work as we demonstrate with an example in the full version of this paper. The problem here appears to be that we fix a single distribution Z to explore the two cases of Eq. (IV.1). In our final approach, we pick different distributions depending on the case we would like to prove. Namely, if $\mathbf{I}(J : Z) \geq (1-\epsilon)\log t$, then carrying out the calculations in Eq. (III.9) through (III.12) with the assumption $\mathbf{I}(J : Z) \geq (1-\epsilon)\log t$, we prove the first case, namely $m_{t+2} \geq t^{1-\epsilon} m_t^{1+2/t}$ using the distribution Z. If $\mathbf{I}(J : Z) < (1-\epsilon)\log t$ on the other hand, we demonstrate two new random variables W, Y which are distributed respectively on length $t+2$ and length $t-2$ paths so that $\mathbf{D}(W \,\|\, F^{t+2}) + \mathbf{D}(Y \,\|\, F^{t-2}) \leq -2\log S^t(\mu,\nu) - \log\delta$, which implies that $m_{t-2}m_{t+2} \geq \delta m_t^2$. While W and Y are constructed by modifying X in suitable ways, which is how Z was constructed also, we do so with the hindsight of having inspected what causes $\mathbf{I}(J : Z)$ to be smaller than $(1-\epsilon)\log t$. It is precisely this adaptivity which enables this approach to overcome the difficulties encountered by the one suggested in Eq. (IV.2).

If $\mathbf{I}(J : Z) \geq (1-\epsilon)\log t$, by plugging this into Eq. (III.11) and carrying out the following calculations, we get $m_{t+2} \geq t^{1-\epsilon} \cdot m_t^{1+2/t}$. Therefore it remains to show there exists a $\delta > 0$ such that assuming $\mathbf{I}(J : Z) < (1-\epsilon)\log t$, we have $m_{t+2}m_{t-2} \geq \delta m_t^2$. To do so, we will demonstrate distributions W and Y on walks that start from $r \in \Omega_\circ$ and return to r after spending respectively $t+2$ and $t-2$ time steps in Ω such that $\mathbf{D}(W \,\|\, F^{t+2}) + \mathbf{D}(Y \,\|\, F^{t-2}) \leq -2\log S^t(\mu,\nu) - \log\delta$. Notice that by Lemma II.1 this indeed implies that $m_{t+2}m_{t-2} \geq \delta m_t^2$. The distributions W and Y will be mixture of $\Theta(t)$ random walks, in particular, they are not Markovian in general.

For brevity let us set $\mu_i^x := \mathrm{dist}(X_{i-1} \,|\, X_i = x)$ and $\nu_i^x := \mathrm{dist}(X_{i+1} \,|\, X_i = x)$. Let U be the unary encoding of J: a length t bit vector of which only the Jth coordinate is set. First we would like to understand the contribution of each bit of U to $\mathbf{I}(Z : J) = \mathbf{I}(Z : U)$. Using the chain rule, we write

$$\begin{aligned}
(1-\epsilon)\log t &> \mathbf{I}(U : Z) \\
&= \sum_{i=1}^{t} \mathbf{I}(U_i : Z \,|\, U_{<i}) && \text{(IV.3)} \\
&= \sum_{i=1}^{t} \frac{t-i+1}{t} \mathbf{I}(U_i : Z \,|\, U_{<i} = 0) && \text{(IV.4)} \\
&\geq \sum_{i=1}^{t} \frac{t-i+1}{t} \mathbf{I}(U_i : Z_i Z_{i+1} \,|\, U_{<i} = 0) && \text{(IV.5)} \\
&= \sum_{i=1}^{t} \frac{1}{t} \mathop{\mathbb{E}}_{x\sim X_i} \mathbf{D}(\mu_i^x \,\|\, \lambda_i \mu_i^x + (1-\lambda_i)\nu_i^x) \\
&\qquad + \sum_{i=1}^{t} \frac{t-i}{t} \mathop{\mathbb{E}}_{x\sim X_i} \mathbf{D}(\nu_i^x \,\|\, \lambda_i \mu_i^x + (1-\lambda_i)\nu_i^x) && \text{(IV.6)}
\end{aligned}$$

where we set $\lambda_i := 1/(t-i+1)$, which is the probability that $U_i = 1 \,|\, U_{<i} = 0$. Here, Eq. (IV.3) follows from the chain rule, Eq. (IV.4) is true because if $U_{<i} \neq 0$ then $U_i = 0$ (as U has a single coordinate that is one) and consequently the mutual information is zero, and Eq. (IV.5) is the data processing inequality. Next we lower bound Eq. (IV.6) by its first term (which is valid since μ_i^x, ν_i^x are distributions hence the second term of Eq. (IV.6) is nonnegative), obtaining

$$(1-\epsilon)\log t > \mathop{\mathbb{E}}_{i\sim J} \mathop{\mathbb{E}}_{x\sim X_i} \mathbf{D}(\mu_i^x \,\|\, \lambda_i \mu_i^x + (1-\lambda_i)\nu_i^x). \tag{IV.7}$$

To simplify the presentation, here we only provide the proof of Theorem I.4 for $\epsilon > 7/8$ which demonstrates the ideas in their simplest form. This bound already implies all our results in complexity theory, with a constant factor loss of no more than 8. The proof for any $\epsilon > 0$ can be found in the full version of this paper.

A. The bound for $\epsilon > 7/8$

If we condition on the event $i \in \{1, \dots, \lceil t/2 \rceil\}$, this expectation increases by a factor of at most 2; namely

$$\mathop{\mathbb{E}}_{i\sim[t/2]} \mathop{\mathbb{E}}_{x\sim X_i} \mathbf{D}(\mu_i^x \,\|\, \lambda_i \mu_i^x + (1-\lambda_i)\nu_i^x) < 2(1-\epsilon)\log t.$$

By Markov's inequality

$$\Pr_{i\sim[t/2], x\sim X_i} [\mathbf{D}(\mu_i^x \,\|\, \lambda_i \mu_i^x + (1-\lambda_i)\nu_i^x) \geq 8(1-\epsilon)\log t] < 1/4,$$

so it follows that there is a set $T \subseteq [\lceil t/2 \rceil]$ of size at least $\lfloor t/4 \rfloor$ such that if $i \in T$ we have

$$\Pr_{x\sim X_i} [\mathbf{D}(\mu_i^x \,\|\, \lambda_i \mu_i^x + (1-\lambda_i)\nu_i^x) \geq 8(1-\epsilon)\log t] < 1/2. \tag{IV.8}$$

For each $i \in T$ let X_i' be the random variable obtained from X_i by conditioning on those $x \in \mathrm{supp}(X_i)$ satisfying $\mathbf{D}(\mu_i^x \,\|\, \lambda_i \mu_i^x + (1-\lambda_i)\nu_i^x) < 8(1-\epsilon)\log t$. Furthermore, for each $i \in T$ and $x \in \mathrm{supp}(X_i')$, we construct distributions $\pi_i^x \colon \Omega \to \mathbb{R}_+$ to be specified later. Let P_i be sampled by $x \sim X_i'$ first and then picking $p \sim \pi_i^x$.

B. The distributions W and Y

Let K be an integer sampled uniformly at random from the set T (constructed in the previous section). For each fixing k of K, the random variables $W \mid K = k$ and $Y \mid K = k$ are random walks (i.e., they are Markovian) constructed as follows. We first pick $x, p \sim X_k' P_k$. The walk $Y \mid K = k$ is generated by concatenating a sample from $X_{-1}X_0 \dots X_{k-1} \mid X_{k-1} = p$ and an independent sample from $X_{k+1} \dots X_{t+1} \mid X_{k+1} = p$. The walk W is generated by concatenating a sample from $X_{-1}X_0 \dots X_k \mid X_k = x$, the path (x, p) and (p, x') for an independent sample $x' \sim (X_k' \mid P_k = p)$ and an independent sample from $X_k \dots X_{t+1} \mid X_k = x'$.

For $k \in T$ we define another random walk $\check{X}^k = (\check{X}^k_{-1}, \dots, \check{X}^k_{t+1})$, only to be used in the analysis of W and Y. We sample $x \sim X_k'$ and set $\check{X}^k_k = x$. We pick the rest of the coordinates of $\check{X}^k$ according to the distribution $X \mid X_k = x$. Note that for any $k \in T$, we have

$$\mathbf{D}(\check{X}^k \,\|\, X) = \mathbf{D}\left(X_k' \,\middle\|\, X_k\right) \leq 1$$

by Eq. (IV.8) and Lemma II.2 and the fact that both X and $\check{X}^k$ are Markovian.

Lemma IV.1. *We have*

$$\begin{aligned}&\mathbf{D}\left(W \mid K = k \,\middle\|\, F^{t+2}\right) + \mathbf{D}\left(Y \mid K = k \,\middle\|\, F^{t-2}\right)\\ &\leq -2\log S^t(\mu,\nu) + 2 + \mathop{\mathbb{E}}_{x\sim X_k'} \mathbf{D}(\pi_k^x \,\|\, \mu_k^x) + \mathop{\mathbb{E}}_{x\sim X_k'} \mathbf{D}(\pi_k^x \,\|\, \nu_k^x).\end{aligned}$$

Proof: We have

$$\begin{aligned}\mathbf{D}\left(W \mid K = k \,\middle\|\, F^{t+2}\right) &= \mathbf{D}(\check{X}^k \,\|\, F^t)\\ &\quad + \mathbf{D}\left(P_k \mid X_k' \,\middle\|\, F_{k+1} \mid F_k\right) + \mathbf{D}\left(X_k' \mid P_k \,\middle\|\, F_{k+1} \mid F_k\right)\end{aligned}$$

and further

$$\begin{aligned}&\mathbf{D}\left(Y \mid K = k \,\middle\|\, F^{t-2}\right) + \mathbf{D}\left(P_k \mid X_k' \,\middle\|\, F_{k+1} \mid F_k\right)\\ &\quad + \mathbf{D}\left(X_k' \mid P_k \,\middle\|\, F_{k+1} \mid F_k\right)\\ &= \mathbf{D}(\check{X}^k \,\|\, F^t) + \mathop{\mathbb{E}}_{x\sim X_k'} \mathbf{D}(\pi_k^x \,\|\, \mu_k^x) + \mathop{\mathbb{E}}_{x\sim X_k'} \mathbf{D}(\pi_k^x \,\|\, \nu_k^x).\end{aligned}$$

Summing up the two inequalities and substituting $\mathbf{D}(\check{X}^k \,\|\, X) \leq 1$ we get the result. ∎

At this point, in light of Lemma IV.1, we could pick each π_k^x so that it minimizes $\mathbf{D}(\pi_k^x \,\|\, \mu_k^x) + \mathbf{D}(\pi_k^x \,\|\, \nu_k^x)$: the unique minimizer is given by $\pi_k^x = \sqrt{\mu_k^x \nu_k^x} / \langle \sqrt{\mu_k^x}, \sqrt{\nu_k^x} \rangle$. However doing so leads to W, Y which diverge from the F walk by more than a constant, and therefore is not good enough for our needs. To obtain better random variables W and Y, we crucially use the fact that W is a mixture of $\Theta(t)$ random walks. Namely, if we consider the entropy coming from the $\mathbf{I}(W : K)$ term also, a better strategy for picking the distributions π_k^x becomes available. By contrast, we do not use the fact that Y is a mixture and, in fact, it can be replaced by $Y \mid K = k_0$ where $k_0 = \arg\min_k \mathbf{D}\left(Y \mid K = k \,\middle\|\, F^{t-2}\right)$, however the averaged quantity $\mathbf{D}\left(Y \mid K \,\middle\|\, F^{t-2}\right)$ is far more convenient to work with.

C. The contribution of $\mathbf{I}(K : W)$

Similar to Eq. (IV.7), we would like to understand the contribution of each time step $t \in T$ to $\mathbf{I}(K : W)$. Let V be the unary encoding of K: a length t bit vector of which only the Vth coordinate is set. Using the chain rule for mutual information

$$\begin{aligned}\mathbf{I}(W : V) &= \sum_{i\in T} \mathbf{I}(V_i : W \mid V_{<i})\\ &\geq \mathop{\mathbb{E}}_{k\sim K} \mathop{\mathbb{E}}_{x\sim X_k'} \mathbf{D}(\pi_k^x \,\|\, \eta_i \pi_k^x + (1-\eta_i)\widetilde{v}_k^x),\end{aligned}$$

where $\eta_k = 1/\operatorname{rank}_T(k)$ and $\widetilde{v}_k^x := \mathbb{E}_{j>k : j\in T} \operatorname{dist}(\check{X}^j_{k+1} \mid \check{X}^j_k = x)$. Here $\operatorname{rank}_T(i)$ denotes the position of $i \in T$ when the elements of T are sorted in decreasing order. By Eq. (IV.8), and the definition of X_k', we have $\widetilde{v}_k^x(y) \leq 2\nu_k^x(y)$ for all $y \in \Omega$. Therefore we conclude that

$$\mathbf{I}(W : K) \geq \mathop{\mathbb{E}}_{k\sim K} \mathop{\mathbb{E}}_{x\sim X_k'} \mathbf{D}(\pi_k^x \,\|\, \eta_k \pi_k^x + 2(1-\eta_k)\nu_k^x). \tag{IV.9}$$

Note in the above divergence expression the reference measure is not a probability distribution, which our definition permits (cf. Eq. (II.1)).

Recall our goal in this section is to upper bound $\mathbf{D}\left(W \,\middle\|\, F^{t+2}\right) + \mathbf{D}\left(Y \,\middle\|\, F^{t-2}\right) + 2\log S^t(\mu,\nu)$ by $\log 1/\delta$. Let us write

$$\begin{aligned}&\mathbf{D}\left(W \,\middle\|\, F^{t+2}\right) + \mathbf{D}\left(Y \,\middle\|\, F^{t-2}\right) + 2\log S^t(\mu,\nu)\\ &\quad\leq \mathbf{D}\left(W \mid K \,\middle\|\, F^{t+2}\right) + \mathbf{D}\left(Y \mid K \,\middle\|\, F^{t-2}\right) \qquad \text{(IV.10)}\\ &\qquad - \mathbf{I}(K : W) + 2\log S^t(\mu,\nu)\\ &\leq 2 + \mathop{\mathbb{E}}_{k\sim K, x\sim X_k'} \mathbf{D}\left(\pi_x^k \,\middle\|\, \mu_x^k\right) + \mathbf{D}\left(\pi_x^k \,\middle\|\, \nu_x^k\right) - \mathbf{I}(K : W)\\ &\leq 2 + \mathop{\mathbb{E}}_{k\sim K, x\sim X_k'} \mathop{\mathbb{E}}_{y\sim\pi_k^x} \log \frac{\eta_k \pi_k^x(y)^2 + 2(1-\eta_k)\nu_k^x(y)\pi_k^x(y)}{\mu_k^x(y)\nu_k^x(y)}, \qquad \text{(IV.11)}\end{aligned}$$

where the second inequality follows from Lemma IV.1 and the last inequality is obtained by plugging in Eq. (IV.9). Note that the function $z \mapsto z\log(az^2 + bz)$ is strictly convex in $\mathbb{R}_+$ whenever $ab > 0$, therefore for each k, x there is a unique minimizer $(\pi_k^x)^*$ of Eq. (IV.11), which can be calculated, say, using Lagrange multipliers. However, instead of the minimizer, we work with a simple approximation of it. For each $k \in T$ and $x \in \mathrm{supp}(X_k')$, we let

$$\Psi_k^x := \left\{ y \in \Omega \;\middle|\; \nu_k^x(y) \geq \frac{\lambda_k}{1-\lambda_k} \mu_k^x(y) \right\}.$$

By definition of X'_k, we have $\mathbf{D}\left(\mu_i^x \,\|\, \lambda_i\mu_i^x + (1-\lambda_i)\nu_i^x\right) < 8(1-\epsilon)\log t$. Let $\gamma = 1 - 8(1-\epsilon)$, which is positive by our assumption $\epsilon > 7/8$. By Markov's inequality, and the fact that $\lambda_k \le 2/t$, we get

$$\mu_k^x(\Psi_k^x) \ge \gamma$$

for large enough t. Let π_k^x be $\mu_k^x \,|\, \Psi_k^x$, namely we have $\pi_k^x(y) = \mu_k^x(y)/\mu_k^x(\Psi_k^x)$ if $y \in \Psi_k^x$, and $\pi_k^x(y) = 0$ otherwise. Continuing from Eq. (IV.11), we have

$$\le 2 + \mathop{\mathbb{E}}_{k\sim K, x\sim X'_k} \mathop{\mathbb{E}}_{y\sim\pi_k^x} \log \frac{\eta_k \pi_k^x(y)^2 + 2(1-\eta_k)\nu_k^x(y)\pi_k^x(y)}{\mu_k^x(y)\nu_k^x(y)}$$

$$\le 2 + \mathop{\mathbb{E}}_{k\sim K} \log\left(\frac{\eta_k(1-\lambda_k)}{\lambda_k\gamma^2} + \frac{2}{\gamma}\right), \tag{IV.12}$$

where the second inequality is true by definition of Ψ_k^x and π_k^x. Now we argue that the expectation term in Eq. (IV.12) is maximized when T is the set containing the smallest $|T|$ elements of $[\lceil t/2 \rceil]$. To see this suppose there is an $i \notin T$ which is smaller than the maximum element of T. Let j be the smallest item in T which is greater than i. We see that the expectation term increases if we replace T by $T \setminus \{j\} \cup \{i\}$ as $\log\left(\frac{C(1-\lambda_k)}{\lambda_k\gamma^2} + \frac{2}{\gamma}\right)$ is decreasing in k and the ranks do not change after swapping j with i. Therefore,

$$\begin{aligned}
&\mathbf{D}\left(W \,\middle\|\, F^{t+2}\right) + \mathbf{D}\left(Y \,\middle\|\, F^{t-2}\right) + 2\log S^t(\mu,\nu) \\
&\qquad \le 2 + \log\left(\prod_{i=1}^{|T|} \frac{t/2+3i}{i\gamma^2}\right)^{1/|T|} \\
&\qquad = \log\frac{12}{\gamma^2} + \log\left(\prod_{i=1}^{|T|} \frac{t/6+i}{i}\right)^{1/|T|} \\
&\qquad \le \log\frac{12}{\gamma^2} + \log\binom{2|T|}{|T|}^{1/|T|} \qquad \text{(IV.13)} \\
&\qquad \le \log\frac{48}{\gamma^2}, \qquad \text{(IV.14)}
\end{aligned}$$

where the $\binom{2|T|}{|T|}$ is the middle binomial coefficient, in the second inequality we use the fact $|T| > t/6$, and the last inequality is true as $\binom{2n}{n} < 2^{2n}$. Therefore it is enough to choose $\epsilon > 7/8$ and $\delta \le \frac{(1-8(1-\epsilon))^2}{48} = \frac{4}{3}(\epsilon - 7/8)^2$. We have established the following.

Theorem I.4 (restated)**.** *For any $\epsilon > 7/8$ there is a $\delta > 0$ such that $m_{t+2} \ge t^{1-\epsilon} m_t^{1+2/t}$ unless $m_{t+2}m_{t-2} \ge \delta m_t^2$.*

V. Randomized computational models

In this section we show the connection between Theorem I.3, Theorem I.4 and the randomized communication and query complexities of the Hamming distance problem.

A. Communication complexity

In a two player communication problem the players, named Alice and Bob, receive separate inputs, respectively $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, and they communicate in order to compute the value $f(x,y)$ of a function $f\colon \mathcal{X}\times\mathcal{Y} \to \{0,1\}$ (known to both players). In an r-round protocol, the players can take at most r turns alternately sending each other a message (that is, a bit string) and the last player to receive a message declares the output of the protocol. A protocol can be *deterministic* or *randomized*; in the latter case the players can base their actions on a common random source and we measure the *error probability*: the maximum over inputs $(x,y) \in \mathcal{X}\times\mathcal{Y}$, of the probability that the output of the protocol differs from $f(x,y)$. The *communication cost* of a protocol is the maximum, over the inputs and the random string, of the total number of bits sent between the players. For a function $f\colon \mathcal{X}\times\mathcal{Y} \to \{0,1\}$, an integer r and $\delta \in [0,1]$, we denote by $R_\delta^r(f)$ the minimum over all protocols for f having r-rounds and error probability at most δ, of the communication cost incurred. We define $R_\delta(f)$ similarly, but we take the maximum over δ-error protocols with no restriction on the number of rounds it uses.

In the k-Hamming distance problem, denoted Ham_k^n, the players receive length-n bit strings, respectively $x, y \in \{0,1\}^n$, and are required determine if $\|x-y\|_1 \le k$ or not. There is a well known one-round communication protocol which accomplishes this with error probability δ by communicating $O(k\log(k/\delta))$ bits.

Theorem V.1 (e.g., Huang, Shi, Zhang and Zhu [6])**.** *It holds that*

$$R_\delta^1(\mathrm{Ham}_k^n) = O(\min\{k\log(k/\delta), k\log(n/k)\}).$$

Highly related to the Ham_k^n is the k-disjointness problem Disj_k^n, wherein the players each receive a k-subset of $[n]$ and their goal is to determine if their sets intersect. Notice that Disj_k^n can be seen as a promise version of Ham_{2k-2}^n where each player is guaranteed to have a string with Hamming weight k: the sets are disjoint if and only if the Hamming distance between the characteristic vectors of the sets is more than $2k-2$. Therefore any upper bound for the Ham_k^n carries over to Disj_k^n and any lower bound for Disj_k^n carries over to Ham_k^n. Around 1993, Håstad and Wigderson [31] showed that there is a more efficient protocol for Disj_k^n than that implied by Theorem V.1, which communicates only $O(k)$ bits, but over $O(\log k)$ rounds.

On the lower bounds side, the result of [32] implies that $\Omega(k)$ bits is needed for these problems even if one uses arbitrarily large number of round protocols. In [8] it was shown that any 1-round protocol for Disj_k^n needs to communicate at least $\Omega(k\log k)$ bits when $k^2 < n$ (this result was proven later in [33] also). In Theorem 3.2 of [34], an $\Omega(k\log(1/\delta))$ bound for 1-round complexity of Disj_k^n was shown even when Bob receives just one element (i.e., the indexing problem) for $k < \delta n$ and a slightly more general result was shown in [35]. Finally in [13] the communication complexity of Disj_k^n was settled:

$$R_{1/3}^r(\mathrm{Disj}_k^n) = \Theta(k\log^{(r)} k)$$

for $1 \le r < \log^* k$ and $k < n^2$. Their upper bound solves the disjointness problem with error probability at most $1/\exp k + 1/\exp^{(r)}(c\log^{(r)} k)$ for any $c > 1$ by communicating $O(k\log^{(r)} k)$ bits over r rounds. In fact bulk of the bits is sent in the first round and the rest of the rounds amount to an $O(k)$ bits of communication. Taking $r = \log^* k$, this leads to an $O(k)$ bits protocol with error probability that is exponentially small in k. Their lower bound shows that at least one message of size $\Omega(k\log^{(r)} k)$ bits needs to be sent by any r-round protocol, even if it has error probability $1/3$. Prior to this work, this lower bound provided the strongest lower bound for Ham_k^n also, along with the incomparable bound of $\Omega(k\log(1/\delta))$ due to [9] which holds for any number of rounds, which we discuss shortly.

Problem	Upper bound	Rounds	Error	Lower bound	Reference
Ham_k^n	$O(k\log(k/\delta))$	1	δ		Folklore, [6]
	appl. ↓	any	δ	$\Omega(k\log(1/\delta))$	[9]
		any	δ	$\Omega(k\log(k/\delta))$	This work
Disj_k^n	$O(k\log(k/\delta))$	1	δ	applies ↑	Folklore
	$O(k)$	$O(\log k)$	$1/3$		[31]
	$O(k\log^{(r)} k)$	r	$1/\exp^{(r)}\left(c\log^{(r)} k\right)$		[13]
	$O(k)$	$\log^* k$	$1/\exp k$		[13]
		r	$1/3$	$\Omega(k\log^{(r)} k)$	[13]
		1	$1/3$	$\Omega(k\log k)$	[8], [33]
		1	δ	$\Omega(k\log(1/\delta))$	[34], [35]
		any	$1/3$	$\Omega(k)$	[32]

Table I
KNOWN BOUNDS FOR Disj_k^n AND Ham_k^n.

To summarize the above results, the 1-round communication complexity of both Disj_k^n and Ham_k^n is $\Theta(k\log(k/\delta))$ by [8], [34], [35] and [6]. We know that Disj_k^n can be solved much more efficiently if one is allowed many rounds: firstly the $\log k$ factor can be removed [31] and secondly the error probability can be brought down to $\exp(-k)$ [13], by using no more than $\log^* k$ rounds. It is an interesting question whether similar efficiency improvements can be obtained for Ham_k^n also, by using multiple rounds. The first separation of Disj_k^n and Ham_k^n was proven in [9], which shows that $\Omega(k\log(1/\delta))$ lower bound holds for any protocol solving Ham_k^n. Therefore in Ham_k^n, we get no improvements in error probability by interactive communication. It remained an open question whether *any* improvement can be made at all to the 1-round protocol by communicating interactively. In this work we answer this question negatively:

Theorem I.8 (restated)**.** *For $k^2 < \delta n$ we have $R_\delta(\mathrm{Ham}_k^n) = \Omega(k\log(k/\delta))$. The bound applies even to protocols that may output an arbitrary answer when $\|x-y\|_1 \notin \{k-2, k, k+2\}$.*

Before we proceed with proving Theorem I.8, let us first warm up by showing that Theorem I.3 implies an $\Omega(k\log(1/\delta))$ lower bound on $R_\delta(\mathrm{Ham}_k^n)$. To do so, let us review the so called *corruption bound* method. Let $f\colon \mathcal{X}\times\mathcal{Y}\to\{0,1\}$ be the function the players would like to compute with Alice having received $x\in\mathcal{X}$ and Bob $y\in\mathcal{Y}$. For a protocol P for f, define the matrix $A_P\colon \mathcal{X}\times\mathcal{Y}\to[0,1]$ such that $A_P(x,y)$ is the probability that the protocol outputs 1 on input (x,y). It is well known and not difficult to see that if P has communication cost c, then A_P is the average of matrices each of which is the sum of at most 2^c rank 1 matrices $uv^\intercal$ with $u\in\{0,1\}^{\mathcal{X}}$ and $v\in\{0,1\}^{\mathcal{Y}}$. Therefore to show the communication cost of a protocol P is more than c, it suffices to argue A_P lies outside 2^c times the polytope

$$\mathcal{T} := \mathrm{conv}\left\{uv^\intercal \mid u\in\{0,1\}^{\mathcal{X}}, v\in\{0,1\}^{\mathcal{Y}}\right\},$$

where conv denotes the convex hull. By convexity, A_p lies outside of $\mathcal{T}$ if and only if there is a hyperplane (with normal H) separating the two; namely that $\langle A_P, H\rangle > 2^c\langle R, H\rangle$ for all vertices R of the polytope $\mathcal{T}$.

Let $\mu_k\colon \{0,1\}^n\times\{0,1\}^n\to\mathbb{R}_+$ be the distribution on pairs (x,y) obtained as follows. Sample x uniformly at random and obtain y by flipping k coordinates of x chosen uniformly at random and with replacement (here if a coordinate gets flipped twice it reverts back to its initial value).

Theorem I.7 (restated)**.** *For $k^2 < \delta n$ we have $R_\delta(\mathrm{Ham}_k^n) = \Omega(k\log(1/\delta))$. The bound applies even to protocols that may output an arbitrary answer when $\|x-y\|_1 \notin \{k, k+2\}$.*

Proof: Suppose we have a randomized protocol for Ham_k^n with error probability δ. Form the matrix A, where $A(x,y)$ is the probability that the protocol reports $\|x-y\|_1 \le k$ on input (x,y).

Set $H = \mu_k - \mu_{k+2}/(3\delta)$. Let us first argue that $\langle A, H\rangle \ge 1/3$. We have $\langle A, \mu_k\rangle > (1-\delta)\left(1-\binom{k}{2}/n\right) > 1-3\delta/2$, and $\langle P, \mu_{k+2}\rangle \le \delta + \binom{k+2}{2}/n \le 3\delta/2$. Hence $\langle A, H\rangle \ge 1/3$ for $\delta \le 1/9$.

Next we argue that $\langle R, H\rangle < (3\delta)^{k/2}$ for any $R = uv^\intercal$ with $u, v\in\{0,1\}^n$. If $\langle R, \mu_k\rangle < (3\delta)^{k/2}$, we are done as $\langle R, \mu_{k+2}\rangle \ge 0$ and is a negative term in $\langle R, H\rangle$. If $\langle R, \mu_k\rangle \ge (3\delta)^{k/2}$ on the other hand, observing $\langle R, \mu_k\rangle = \left\langle v, W^k u\right\rangle/2^n$, where W is the normalized adjacency matrix of the Hamming cube, we have by Theorem I.3

$$\left(\frac{\|u\|_2\|v\|_2}{2^n}\right)^{2/k}\left\langle v, W^{k+2}u\right\rangle \ge \left\langle v, W^k u\right\rangle^{1+2/k}.$$

Note $\|u\|_2\|v\|_2 \le 2^n$ since u, v are 0-1 vectors, therefore $\langle R, \mu_{k+2}\rangle \ge 3\delta\langle R, \mu_k\rangle$ and hence $\langle R, H\rangle \le 0$. In either case we have shown that $\langle R, H\rangle < (3\delta)^{k/2}$. This implies an $\log((3\delta)^{-k/2}/3) = \Omega(k\log(1/\delta))$ bits lower bound on $R_\delta(\mathrm{Ham}_k^n)$. ∎

Interestingly, Theorem I.8 cannot be proved by a direct application of the corruption method described above. If we assume that the protocol is supposed to output 1 on inputs $\|x-y\|_1 \le k$, then there are vertices of the polytope $\mathcal{T}$ for which the $\Omega(k\log(1/\delta))$ bound of Theorem I.7 is tight. If we assume that the protocol is supposed to output 1 on inputs $\|x-y\|_1 > k$ on the other hand, no bound above $\Omega(k)$ can be obtained, as there are vertices for which this is tight. If we insist however that the protocol outputs 1 for $\|x-y\|_1 = k$ and 0 for $\|x-y\|_1 \in \{k-2, k+2\}$ then a protocol with cost smaller than $O(k\log(k/\delta))$ would be in violation of the near log-convexity principle we established in Theorem I.4 as we argue next. Of course, if we had a δ-error randomized protocol

P for Ham_k^{n+2} outputting 1 when $\|x-y\|_1 \in \{k-2,k\}$ and 0 if $\|x-y\|_1 = k+2$ (but without any guarantees for other types of inputs), then given inputs $a,b \in \{0,1\}^n$ Alice and Bob can run P (say, in parallel) on instances $(00a, 00b)$ and $(00a, 11b)$ and declare $\|a-b\|_1 = k$ if P returns 1 on $(00a, 00b)$ and 0 on $(00a, 11b)$. This would lead to a protocol with twice the error probability and communication cost of P, deciding between $\|a-b\|_1 = k$, $\|a-b\|_1 = k-2$ and $\|a-b\|_1 = k+2$. The table below shows that P outputting 1 on $(00a, 00b)$ and 0 on $(00a, 11b)$ implies $\|a-b\|_1 = k$ or at least one invocation of P erred.

Input	$k-2$	k	$k+2$
$(00a, 00b)$	1	1	0
$(00a, 11b)$	1	0	?

Proof of Theorem I.8: Suppose we have a δ-error randomized protocol that outputs 1 when $\|x-y\|_1 = k$ and 0 when $\|x-y\|_1 = k-2$ or $\|x-y\|_1 = k+2$.

Form the matrix A, where $A(x,y)$ is the probability that the protocol reports that $\|x-y\|_1 = k$ on input (x,y). Let $\alpha_1, \alpha_2 > 0$ be some reals so that Theorem I.4 implies $m_{t+2} \geq t^{\alpha_1} m^{1+2/t}$ or $m_{t+2} m_{t-2} \geq \alpha_2 {m_t}^2$ for m_t defined in statement of this theorem.

Set $H = \mu_k - (\mu_{k-2} + \mu_{k+2})/(6\delta)$. Let us argue first that $\langle A, H\rangle \geq 1/3$. One can verify that $\langle A, \mu_k\rangle \geq (1-\delta)(1-\binom{k}{2}/n) > 1 - 3\delta/2$ and $\langle A, \mu_{k-2}\rangle + \langle A, \mu_{k-2}\rangle < 3\delta$. Hence $\langle A, H\rangle \geq 1/3$ for $\delta \leq 1/9$.

We upper bound $\langle R, H\rangle$ for some rank-1 matrix $R = uv^\intercal$ with 0-1 values. Let W be the normalized adjacency matrix of the Hamming cube graph. Observe that $\langle R, \mu_k\rangle = \langle v, W^k u\rangle / 2^n$. By Theorem I.4, either $\langle R, \mu_{k+2}\rangle \langle R, \mu_{k-2}\rangle \geq \alpha_2 \langle R, \mu_k\rangle^2$ or

$$\left(\frac{\|u\|_2 \|v\|_2}{2^n}\right)^{2/k} \langle R, \mu_{k+2}\rangle \geq k^{\alpha_1} \langle R, \mu_k\rangle^{1+2/k}.$$

In the former case,

$$\frac{\langle R, \mu_{k+2}\rangle + \langle R, \mu_{k-2}\rangle}{2} \geq \sqrt{\langle R, \mu_{k+2}\rangle \langle R, \mu_{k-2}\rangle} \geq \sqrt{\alpha_2} \langle R, \mu_k\rangle,$$

which implies that $\langle R, H\rangle < 0$ whenever $\delta < 2\sqrt{\alpha_2}/6$ (recall α_2 is a constant). In the latter case, recalling $\|v\|_2 \|u\|_2 \leq 2^n$, we get $\langle R, H\rangle < 0$ unless $6\delta \langle R, \mu_{k+2}\rangle \leq \langle R, \mu_k\rangle$, which implies that $k^{\alpha_1} \langle R, \mu_k\rangle^{2/k} < 6\delta$. From this we get

$$\langle R, H\rangle \leq \langle R, \mu_k\rangle \leq \left(\frac{6\delta}{k^{\alpha_1}}\right)^{k/2},$$

and hence $\langle R, H\rangle < \left(\frac{6\delta}{k^{\alpha_1}}\right)^{k/2}$ in every case and $R_\delta(\mathrm{Ham}_k^n) = \Omega(k \log(k/\delta))$ whenever $k^2 < \delta n$. ∎

For a protocol P, denote by $\Pi = \Pi(x,y)$ the random variable entailing all the messages communicated between the players on input (x,y). So far we have considered the communication cost of a protocol which is the maximum length of Π over all inputs and the configurations of the random source (these together determine the value of Π). When a distribution μ on the inputs is available, we may speak of a more refined notion of cost, *the internal information cost*, for a protocol P which is defined as

$$\mathsf{IC}_\mu(P) := \mathbf{I}(\Pi : Y \mid X) + \mathbf{I}(\Pi : X \mid Y),$$

where $(X,Y) \sim \mu$. Combining our Theorem I.8 with a result of [36] which relates information and communication costs of a protocol under suitable circumstances, one can conclude that any randomized protocol for Ham_k^n has information cost $\Omega(k \log k)$ as well, under the distribution $\mu = (\mu_k + \mu_{k-2} + \mu_{k+2})/3$. However we note that instead of using Theorem I.4 black-box, taking a closer look at the proof of Theorem I.3 and not performing the relaxation provided in Lemma III.6, we get the following more directly.

Theorem V.2. *Let P be a protocol outputting 1 on pairs (x,y) having $\|x-y\|_1 = k$ with probability $1-\delta$ and outputting 0 on pairs (x,y) having $\|x-y\|_1 \in \{k-2, k+2\}$ with probability $1-\delta$. We have $\mathsf{IC}_{\mu_k}(P) = \Omega(k \log(k/\delta))$.*

Let us finally mention another highly related problem, the so called the gap Hamming distance problem. In GHD_k^n, each of the players receive a bit string, respectively $x, y \in \{0,1\}^n$, with the promise that either $\|x-y\|_1 \leq k$ or $\|x-y\|_1 \geq k + \sqrt{k}$. Their goal is to determine which is the case for any given input. In [37], an $\Omega(k)$ lower bound for this problem was shown, which applies to protocols with any number of rounds. Here we conjecture an improvement to this bound and argue that it would follow from a natural analogue of of Theorem I.4 for continuous time Markov chains, which we discuss in Section VI.

Conjecture V.3. *For $k < \delta n$, we have $R_\delta(\mathrm{GHD}_k^n) = \Omega(k \log(1/\delta))$.*

B. Parity decision trees

In the parity decision tree model, we are given a string $x \in \mathbb{F}_2^n$ and our goal is to determine whether x satisfies a fixed predicate $P\colon \mathbb{F}_2^n \to \{0,1\}$ by only making linear measurements of the form $\langle x, y\rangle$ for some $y \in \mathbb{F}_2^n$ we get to choose. Here, the inner product is over $\mathbb{F}_2^n$, and therefore we get a single bit answer for every measurement we make.

Such measurements can be identified by binary decision trees wherein each internal node is labeled by a $y \in \mathbb{F}_2^n$ denoting the linear measurement $\langle x, y\rangle$ we would make at that node and each leaf is labeled by a YES or a NO denoting the final decision we arrive. Given such a tree and an x, the output of the decision tree is obtained by a root to leaf walk, where at each internal node v with label y_v, we perform the measurement $\langle x, y_v\rangle$ and walk to the left child of v if $\langle x, y_v\rangle = 0$ and to the right child if $\langle x, y_v\rangle = 1$. If a leaf node is reached, the label of the node is taken as the answer of the decision tree. Two quantities we are concerned with are the depth and the size (i.e., the total number of nodes) of the tree.

A δ-error randomized decision tree is a distribution ν over deterministic trees such that for any fixed x, the sampled decision tree outputs the correct answer with probability at least $1-\delta$, where the randomness is over the choice of the decision tree from ν. The depth and the size of a randomized decision tree can be taken as the maximum over the decision trees in the support of ν (here, one can also take the average depth or size, however this choice leads to negligible changes in our bounds).

For a predicate $P\colon \mathbb{F}_2^n \to \{0,1\}$, let $\mathrm{PD}_\delta(P)$ be the minimum, over all randomized decision trees T computing P with probability $1-\delta$, of the depth of T. Let $\mathrm{PS}_\delta(P)$ be the minimum, over all randomized decision trees T computing P with probability $1-\delta$, of the size of T. The following inequalities are immediate

$$\begin{aligned} R_\delta(P \circ \oplus) &\leq 2\mathrm{PD}_\delta(P), \\ \log \mathrm{PS}_\delta(P) &\leq \mathrm{PD}_\delta(P), \end{aligned} \tag{V.1}$$

where $P \circ \oplus$ is the two player communication game in which the two players are given strings $x, y \in \mathbb{F}_2^n$ and are required to calculate $P(x+y)$. We remark that $\log \mathrm{PS}_\delta$ is incomparable to R_δ in general.

Here we study the predicate H_k^n which equals 1 if and only if the Hamming weight of its input is precisely k. By Eq. (V.1) and a padding argument similar to the one we gave before the proof of Theorem I.8, each lower bound for Ham_k^n listed in Table I applies to $\mathrm{PD}_\delta(H_k^n)$ as well. In [12] another direct $\Omega(k)$ bound for $\mathrm{PD}_\delta(H_k^n)$ was shown. In [14], showing an $\Omega(k \log k)$ lower bound to a variant of $\mathrm{PD}_\delta(H_k^n)$ to obtain tight bounds for k-linearity problem (see Section V-C) was suggested. Finally, our Theorem I.8 shows that $\mathrm{PD}_\delta(H_k^n) = \Omega(k \log(k/\delta))$, which is tight. Next we show the same bound holds even for $\log \mathrm{PS}_\delta(H_k^n)$.

Theorem I.9 (restated)**.** *For $k^2 < \delta n$, $\log \mathrm{PS}_\delta(H_k^n) = \Omega(k \log(k/\delta))$.*

Proof: The proof is very similar to that of Theorem I.8, so we only describe the differences.

Let T be a δ-error randomized parity decision tree computing H_k^n. Form $A\colon \mathbb{F}_2^n \to [0,1]$ so that $A(x)$ is the probability T outputs 1 on input $x \in \mathbb{F}_2^n$. Define the polytope

$$\mathcal{P} := \mathrm{conv}\left\{x \mapsto \mathbb{1}[Bx = c] \mid B \in \mathbb{F}_2^{n\times n}, c \in \mathbb{F}_2^n\right\}$$

whose vertices are indicator functions for affine subspaces of $\mathbb{F}_2^n$. Given a parity decision tree, the set of inputs that end up in a particular leaf of it is an affine subspace in $\mathbb{F}_2^n$. Therefore if T has at most s leaves, then A is inside $s\mathcal{P}$. It remains to demonstrate a hyperplane with normal H so that $\langle A, H\rangle > s\langle V, H\rangle$ for any vertex V of the polytope $\mathcal{P}$ for $s = \exp\Omega(k\log(k/\delta))$.

Let μ_k be a distribution on $\mathbb{F}_2^n$ obtained as follows. Start with the 0 vector, and flip a coordinate chosen uniformly at random with replacement k times. Here, flipping a coordinate an even number of times leaves it seemingly intact. Set $H = \mu_k - (\mu_{k-2} + \mu_{k+2})/(6\delta)$.

First observe that $\langle A, \mu_k\rangle > (1-\delta)(1-\binom{k}{2}/n) > 1 - 3\delta/2$ and $\langle A, \mu_{k+2}\rangle + \langle A, \mu_{k-2}\rangle < 3\delta$ so $\langle A, H\rangle \geq 1/3$ for $\delta \leq 1/9$. Next we would like to upper bound $\langle V, H\rangle$ for an indicator function V of an affine subspace $\{x \in \mathbb{F}_2^n \mid Bx = c\}$. The key observation is

$$\langle V, \mu_k\rangle = \left\langle \mathbb{1}_c, S^k \mathbb{1}_0 \right\rangle \tag{V.2}$$

where S is a stochastic matrix describing the following transition: For any $x \in \mathbb{F}_2^n$, sample a column y of $B \in \mathbb{F}_2^{n\times n}$ uniformly at random and transition to $x + y$. Namely, the right hand side of Eq. (V.2) describes the following probability. We start with the 0 vector in $\mathbb{F}_2^n$ and in each time step sample a uniform random column y of B and add y to the current state. We measure the probability of reaching $c \in \mathbb{F}_2^n$ at time step k. Having observed Eq. (V.2), and that $\|\mathbb{1}_0\|_2 = \|\mathbb{1}_c\|_2 = 1$, the rest of the proof is identical to that of Theorem I.8: by Theorem I.4, we either have

$$\left\langle \mathbb{1}_c, S^{k+2}\mathbb{1}_0\right\rangle \left\langle \mathbb{1}_c, S^{k-2}\mathbb{1}_0\right\rangle \geq \alpha_2 \left\langle \mathbb{1}_c, S^k \mathbb{1}_0\right\rangle^2$$

or

$$\left\langle \mathbb{1}_c, S^{k+2}\mathbb{1}_0\right\rangle \geq k^{\alpha_1} \left\langle \mathbb{1}_c, S^k \mathbb{1}_0\right\rangle^{1+2/k}.$$

In either event, we conclude that $\langle V, H\rangle \leq \left(\frac{6\delta}{k^{\alpha_1}}\right)^{k/2}$. This completes the proof. ∎

Note in Theorem I.8, we use Theorem I.4 with a simple and fixed S (i.e., the standard random walk on the Hamming cube), but with complicated vectors u, v that come from the particular communication protocol whose communication cost we would like to lower bound. By contrast, in Theorem I.9 the vectors u, v are simple point masses on states 0 and c but the matrix S is a convolution random walk on the Hamming cube that comes from the particular decision tree whose size we lower bound.

C. Property testing

In the property testing model, given black box access to an otherwise unknown function $f\colon \mathbb{F}_2^n \to \mathbb{F}_2$, our goal is to tell apart whether $x \in P$ for some fixed set of functions P or $\|f - g\|_1 \geq \epsilon 2^n$ for any $g \in P$. Here, the black box queries are done by providing an input $x \in \mathbb{F}_2^n$ to the function and observing $f(x)$.

A function $f\colon \mathbb{F}_2^n \to \mathbb{F}_2$ is called k-linear if f is given by

$$f(x) = \sum_{i\in S} x_i$$

for some $S \subseteq [n]$ of size at most k. By combining our communication complexity lower bound Theorem I.8 with the reduction technique developed in [7] or by combining our parity decision tree lower bound Theorem I.9 with a reduction given in [14], one obtains the following.

Corollary I.10 (restated)**.** *Any δ-error property testing algorithm for k-linearity with $\epsilon = 1/2$ requires $\Omega(k\log(k/\delta))$ queries.*

In fact through this, one obtains similar lower bounds to property testing for k-juntas, k-term DNFs, size-k formulas, size-k decision trees, k-sparse $\mathbb{F}_2$-polynomials; see [38], [39].

VI. Discussion

We showed that for a symmetric matrix $S\colon \Omega\times\Omega \to \mathbb{R}_+$ and unit vectors $u, v\colon \Omega \to \mathbb{R}_+$, defining $m_t = \langle v, S^t u\rangle$ for $t = 0, 1, \ldots$, we have

$$m_{t+2} \geq m_t^{1+2/t}, \text{ and} \tag{VI.1}$$

$$m_{t+2} \geq m_t^{1+2/t} \cdot \min\left\{t^{1-\epsilon}, \left\lceil \delta \frac{m_t^{1-2/t}}{m_{t-2}}\right\rceil\right\} \tag{VI.2}$$

and argued that Eq. (VI.2) and (VI.1), in this order, are best viewed as gradual weakenings of the log-convexity of $\{m_t\}_{t=0}^\infty$. We conjecture that a similar principle holds true for continuous time Markov chains as well.

Call a function $f\colon \mathbb{R}_+ \to [0,1]$, whose logarithm is continuously twice differentiable (i.e., $\log f \in C^2(\mathbb{R}_+)$, *nearly-log-convex* if $x^2(\log f)''(x) \geq 2\log f(x)$ for $x \in \mathbb{R}_+$. Note that $\log f \leq 0$, therefore this is a weakening of the usual log-convexity definition, which requires $(\log f)'' \geq 0$.

Conjecture VI.1. *Let $S\colon \Omega\times\Omega \to \mathbb{R}_+$ be a symmetric substochastic matrix and $u, v\colon \Omega \to \mathbb{R}_+$ be unit vectors. The function*

$$t \mapsto \left\langle v, e^{t(S-I)}u\right\rangle$$

is nearly-log-convex.

By an argument similar to the proof of Theorem I.8, one can show the following.

Theorem VI.2. *Conjecture VI.1 implies Conjecture V.3.*

REFERENCES

[1] K. F. Pang and A. E. Gamal, "Communication complexity of computing the hamming distance," *SIAM J. Comput.*, vol. 15, no. 4, pp. 932–947, Nov. 1986.

[2] A. C.-C. Yao, "On the power of quantum fingerprinting," in *Proceedings of the Thirty-fifth Annual ACM Symposium on Theory of Computing*, ser. STOC '03, San Diego, CA, USA: ACM, 2003, pp. 77–81.

[3] G. Cormode, M. Paterson, S. C. Sahinalp, and U. Vishkin, "Communication complexity of document exchange," in *Proceedings of the Eleventh Annual ACM-SIAM Symposium on Discrete Algorithms*, ser. SODA '00, San Francisco, California, USA: Society for Industrial and Applied Mathematics, 2000, pp. 197–206.

[4] Z. Bar-Yossef, T. S. Jayram, R. Krauthgamer, and R. Kumar, "Approximating edit distance efficiently," in *45th Annual IEEE Symposium on Foundations of Computer Science*, Oct. 2004, pp. 550–559.

[5] D. Gavinsky, J. Kempe, and R. de Wolf, "Quantum communication cannot simulate a public coin," *CoRR*, vol. quant-ph/0411051, 2004.

[6] W. Huang, Y. Shi, S. Zhang, and Y. Zhu, "The communication complexity of the hamming distance problem," *Information Processing Letters*, vol. 99, no. 4, pp. 149–153, 2006.

[7] E. Blais, J. Brody, and K. Matulef, "Property testing lower bounds via communication complexity," *computational complexity*, vol. 21, no. 2, pp. 311–358, Jun. 2012.

[8] H. Buhrman, D. Garcia-Soriano, A. Matsliah, and R. de Wolf, *The non-adaptive query complexity of testing k-parities*, 2012. eprint: arXiv:1209.3849.

[9] E. Blais, J. Brody, and B. Ghazi, "The Information Complexity of Hamming Distance," in *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2014)*, K. Jansen, J. D. P. Rolim, N. R. Devanur, and C. Moore, Eds., ser. Leibniz International Proceedings in Informatics (LIPIcs), vol. 28, Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2014, pp. 465–489.

[10] A. Ambainis, W. Gasarch, A. Srinivasan, and A. Utis, "Lower bounds on the deterministic and quantum communication complexity of hamming-distance problems," *ACM Trans. Comput. Theory*, vol. 7, no. 3, 10:1–10:10, Jun. 2015.

[11] A. Ada, O. Fawzi, and H. Hatami, "Spectral norm of symmetric functions," in *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, A. Gupta, K. Jansen, J. Rolim, and R. Servedio, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 338–349.

[12] E. Blais and D. Kane, "Tight bounds for testing k-linearity," in *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, A. Gupta, K. Jansen, J. Rolim, and R. Servedio, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 435–446.

[13] M. Sağlam and G. Tardos, "On the communication complexity of sparse set disjointness and exists-equal problems," *CoRR*, vol. abs/1304.1217, 2013.

[14] A. Bhrushundi, S. Chakraborty, and R. Kulkarni, "Property testing bounds for linear and quadratic functions via parity decision trees," in *Computer Science - Theory and Applications*, E. A. Hirsch, S. O. Kuznetsov, J.-É. Pin, and N. K. Vereshchagin, Eds., Cham: Springer International Publishing, 2014, pp. 97–110.

[15] E. Fischer, E. Lehman, I. Newman, S. Raskhodnikova, R. Rubinfeld, and A. Samorodnitsky, "Monotonicity testing over general poset domains," in *Proceedings of the Thiry-fourth Annual ACM Symposium on Theory of Computing*, ser. STOC '02, Montreal, Quebec, Canada: ACM, 2002, pp. 474–483.

[16] O. Goldreich, "On testing computability by small width obdds," in *Proceedings of the 13th International Conference on Approximation, and 14 the International Conference on Randomization, and Combinatorial Optimization: Algorithms and Techniques*, ser. APPROX/RANDOM'10, Barcelona, Spain: Springer-Verlag, 2010, pp. 574–587.

[17] F. Hausdorff, "Summationsmethoden und momentfolgen i.," ger, *Mathematische Zeitschrift*, vol. 9, pp. 74–109, 1921.

[18] C. Niculescu and L. Persson, *Convex Functions and their Applications: A Contemporary Approach*, ser. CMS Books in Mathematics. Springer New York, 2005.

[19] S. Mandel and I. Hughes, "Change in mean viability at a multiallelic locus in a population under random mating," *Nature*, vol. 182, no. 4627, p. 63, 1958.

[20] G. Blakley and R. Dixon, "Hölder type inequalities in cones," *Journal of Mathematical Analysis and Applications*, vol. 14, pp. 1–4, 1966.

[21] H. P. Mulholland and C. A. B. Smith, "An inequality arising in genetical theory," *The American Mathematical Monthly*, vol. 66, no. 8, pp. 673–683, 1959.

[22] G. R. Blakley and P. Roy, "A hölder type inequality for symmetric matrices with nonnegative entries," *Proceedings of the American Mathematical Society*, vol. 16, no. 6, pp. 1244–1245, 1965.

[23] D. London, "Inequalities in quadratic forms," *Duke Math. J.*, vol. 33, no. 3, pp. 511–522, Sep. 1966.

[24] T. Pate, "Extending the hölder type inequality of blakley and roy to non-symmetric non-square matrices," *Transactions of the American Mathematical Society*, vol. 364, no. 8, pp. 4267–4281, 2012.

[25] P. Erdős and M. Simonovits, "Compactness results in extremal graph theory," *Combinatorica*, vol. 2, no. 3, pp. 275–288, Sep. 1982.

[26] A. A. Sherstov, "The communication complexity of gap hamming distance," *Theory of Computing*, vol. 8, no. 8, pp. 197–208, 2012.

[27] J. Jensen, "Sur les fonctions convexes et les inégalités entre les valeurs moyennes," *Acta Mathematica*, vol. 30, pp. 175–193, 1 Dec. 1906.

[28] A. Wald, "Sequential tests of statistical hypotheses," *Ann. Math. Statist.*, vol. 16, no. 2, pp. 117–186, Jun. 1945.

[29] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, Mar. 1951.

[30] T. M. Cover and J. A. Thomas, *Elements of information theory (2. ed.)* Wiley, 2006, pp. I–XXIII, 1–748.

[31] J. Håstad and A. Wigderson, "The randomized communication complexity of set disjointness," *Theory of Computing*, vol. 3, no. 1, pp. 211–219, 2007.

[32] B. Kalyanasundaram and G. Schnitger, "The probabilistic communication complexity of set intersection," *SIAM J. Discrete Math.*, vol. 5, no. 4, pp. 545–557, 1992.

[33] A. Dasgupta, R. Kumar, and D. Sivakumar, "Sparse and lopsided set disjointness via information theory," in *APPROX-RANDOM*, 2012, pp. 517–528.

[34] M. Sağlam, "Tight bounds for data stream algorithms and communication problems," Master's thesis, Simon Fraser University, School of Computing Science, 8888 University Drive, Burnaby, B.C. Canada V5A 1S6, Sep. 2011.

[35] T. S. Jayram and D. P. Woodruff, "Optimal bounds for johnson-lindenstrauss transforms and streaming problems with sub-constant error," in *SODA*, 2011, pp. 1–10.

[36] I. Kerenidis, S. Laplante, V. Lerays, J. Roland, and D. Xiao, "Lower bounds on information complexity via zero-communication protocols and applications," *SIAM Journal on Computing*, vol. 44, no. 5, pp. 1550–1572, 2015.

[37] A. Chakrabarti and O. Regev, "An optimal lower bound on the communication complexity of gap-hamming-distance," *SIAM Journal on Computing*, vol. 41, no. 5, pp. 1299–1317, 2012.

[38] E. Blais, "Testing juntas nearly optimally," in *Proceedings of the forty-first annual ACM symposium on Theory of computing*, ACM, 2009, pp. 151–158.

[39] S. Chakraborty, D. Garcia-Soriano, and A. Matsliah, "Efficient sample extractors for juntas with applications," in *Proceedings of the 38th International Colloquim Conference on Automata, Languages and Programming - Volume Part I*, ser. ICALP'11, Zurich, Switzerland: Springer-Verlag, 2011, pp. 545–556.