Fitting Points on the Real Line and its Application to RH Mapping

Johan Håstad, Lars Ivansson, and Jens Lagergren

Department of Numerical Analysis and Computing Science
Royal Institute of Technology
SE–100 44 Stockholm
SWEDEN
E-mail: johanh@nada.kth.se, ivan@nada.kth.se, and jensl@nada.kth.se

A natural problem is that of, given an $n \times n$ symmetric matrix D, finding an arrangement of n points on the real line such that the so obtained distances agree as well as possible with the by D specified distances. We refer to the variation in which the difference in distance is measured in maximum norm as the MATRIX-TO-LINE problem. The MATRIX-TO-LINE problem has previously been shown to be NP-complete [12]. We show that it can be approximated within 2, but unless P=NP not within $7/5 - \delta$ for any $\delta > 0$. We also show a tight lower bound under a stronger assumption. We show that the MATRIX-TO-LINE problem cannot be approximated within $2 - \delta$ unless 3-colorable graphs can be colored with $\lceil 4/\delta \rceil$ colors in polynomial time. Currently, the best polynomial time algorithm colors a 3-colorable graph with $\tilde{O}(n^{3/14})$ colors [4].

We apply our MATRIX-TO-LINE algorithm to a problem in computational biology, namely, the Radiation Hybrid (RH) problem. That is, the algorithmic part of a physical mapping method called RH mapping. This gives us the first algorithm with a guaranteed convergence for the general RH problem.

Key Words: line fitting, RH mapping, approximation, lower bounds

1. INTRODUCTION

We study the MATRIX-TO-LINE problem, that is, the problem of, given a set of n points $\{p_i\}_{i=1}^n$ and an $n \times n$ distance matrix D, finding an arrangement $A : \{p_i\}_{i=1}^n \to \boldsymbol{R}^+$ which minimizes

$$\max_{i,j \in [n]} |D[i,j] - |A(p_i) - A(p_j)|| \tag{1}$$

over all such functions. The given distance matrix must be positive, symmetric, and with an all zero diagonal, but we do not require that the distances must satisfy the triangle inequality. The MATRIX-TO-LINE problem has previously been

shown to be NP-complete [12]. We give an algorithm that approximates it within a factor 2. In contrast to this, we show that the MATRIX-TO-LINE problem cannot be approximated within a factor $7/5 - \delta$, for any $\delta > 0$, unless P=NP. This proof is computer aided, since it involves solving a number of linear programs obtained from a gadget construction. We also show that the MATRIX-TO-LINE problem cannot be approximated within $2-\delta$ in polynomial time, unless 3-colorable graphs can be colored with $\lceil 4/\delta \rceil$ colors in polynomial time. It is NP-hard to find a 4-coloring of a 3-colorable graph [9]. The problem of k-coloring a 3-colorable graph is not known to be NP-hard for $k \geq 5$. However, it is a very well studied problem, and despite this there is currently no polynomial time algorithm that colors a 3-colorable graph with less than $\tilde{O}(n^{3/14})$ colors [4].

Sufficient conditions and non-polynomial time algorithms for a more general form of the MATRIX-TO-LINE problem have been given earlier [5]. The MATRIX-TO-LINE problem is an example of a general type of problems, where a distance matrix D for n points is given, and the points should be embedded in some metric space. The goal is to embed the points so that the obtained distances are as close as possible to the distances specified by D, with respect to some norm. This general problem has been studied in [11], and variations of it have been considered in [1, 7].

We apply our MATRIX-TO-LINE algorithm to a physical mapping problem. Physical mapping is an important problem used in large-scale sequencing of DNA as well as for locating genes. Using RH mapping (which is described in Section 4) one can construct a physical map of, for instance, a human chromosome with respect to n markers, which can be genes or arbitrary DNA sequences; that is, one can find the order between these markers and the distance between them on the chromosome by performing a series of experiments and then performing an algorithmic analysis of the outcomes. However, experiments are costly and for this reason one naturally strives to perform as few as possible.

By applying our MATRIX-TO-LINE algorithm, we obtain an algorithm with a guaranteed convergence rate for the RH-problem. Most previous algorithms, see for instance [3, 10, 14], are heuristics that do not guarantee convergence. In [2], Ben-Dor and Chor showed that after approximately $\delta_{\min}^{-2} \log n$ experiments, where $\delta_{\min}$ is a lower bound on the minimum marker distance, the laboratory data is with high probability, what they call, consistent. They also show that a number of rather straightforward algorithms always find the correct marker order when given consistent laboratory data; and so they obtain an algorithm that, given a prior lower bound on $\delta_{\min}$, with high probability finds the correct marker order. We show that the distances between the markers computed by our algorithm converges to the true distances as the number of experiments increases. We also show how this implies that after $O(\delta_{\min}^{-2} \log n)$ experiments, our algorithm finds the correct marker order as well. Furthermore, we show that the probability distribution on the output of an RH-experiment, induced by the arrangement of the markers produced by our algorithm, converges to the distribution induced by the true arrangement of the

markers. We also show a lower bound on this convergence rate for any algorithm for the RH-problem.

The remainder of this paper is organized as follows. In Section 2, the 2-approximation algorithm for the MATRIX-TO-LINE problem is presented. In Section 3 the lower bound $7/5$ on the approximability of the MATRIX-TO-LINE problem is proven. There, it is also shown that MATRIX-TO-LINE cannot be approximated within $2-\delta$ in polynomial time, unless 3-colorable graphs can be colored with $\lceil 4/\delta \rceil$ colors in polynomial time. In Section 4, a probabilistic model of an RH experiment is given. In Section 5, we show how our MATRIX-TO-LINE algorithm can be applied to yield an algorithm for the RH problem. Finally, in Section 6 we show lower bounds on the performance of any algorithm for the RH problem.

2. MATRIX-TO-LINE

In this section, we give an approximation algorithm for the MATRIX-TO-LINE problem.

DEFINITION 2.1. For two $n \times n$ matrices D and D', define

$$||D, D'||_\infty = \max_{i,j \in [n]} |D[i,j] - D'[i,j]|. \tag{2}$$

An arrangement A is a mapping from a set of points $\{p_i\}_{i=1}^n$ to $\boldsymbol{R}^+$. Each arrangement A has an associated distance matrix D_A defined by $D_A[i,j] = |A(p_i) - A(p_j)|$. To avoid multiple subscripts, we will abuse the notation above and write $||D, A||_\infty$ for $||D, D_A||_\infty$ and $||A, A'||_\infty$ for $||D_A, D_{A'}||_\infty$.

DEFINITION 2.2. Given an $n \times n$ distance matrix D, let A^* be an optimal solution to the MATRIX-TO-LINE instance given by D and let ϵ^* be the corresponding optimal value, i.e, $\epsilon^* = ||D, A^*||_\infty$.

Throughout the derivation of the algorithm, we will assume that the leftmost point in the optimal arrangement is known. If this is false, we can always try all possible choices, without increasing the running time with more than a factor n. For simplicity, we assume that p_1 is the leftmost point in A^* and that $A^*(p_1) = 0$.

DEFINITION 2.3. Let p_1 be the leftmost point in an optimal arrangement. Define the arrangement A^1 by $A^1(p_i) = 0$ if $i = 1$, and $A^1(p_i) = D[1,i]$ otherwise.

LEMMA 2.1. $|A^*(p_i) - A^1(p_i)| \leq \epsilon^*$ *for all points* p_i.

Proof. We know that $A^1(p_1) = A^*(p_1) = 0$, so for any point p_i

$$\begin{aligned} |A^*(p_i) - A^1(p_i)| &= |A^*(p_i) - D[1,i]| \\ &= ||A^*(p_i) - A^*(p_1)| - D[1,i]| \leq \epsilon^* \end{aligned} \tag{3}$$

∎

A corollary to Lemma 2.1 is that A^1 approximates the optimal arrangement within a factor 3.

COROLLARY 2.1. $||A^1, D||_\infty \leq 3\epsilon^*$.

Proof. For any pair of points p_i, p_j

$$\begin{aligned} &||A^1(p_i) - A^1(p_j)| - D[i,j]| \\ &\qquad \leq ||A^*(p_i) - A^*(p_j)| - D[i,j]| + 2\epsilon^* \leq 3\epsilon^*, \end{aligned} \tag{4}$$

and thus $||A^1, D||_\infty \leq 3\epsilon^*$. ∎

The key observation behind the 2-approximation algorithm is that if the arrangement A^1 can be modify in such a way that each point p_i, $i > 1$, is moved a distance $\epsilon^*/2$ to the side of $A^1(p_i)$ where $A^*(p_i)$ is located, then the new arrangement will have error $\leq 2\epsilon^*$. Unfortunately, both the optimal arrangement A^* and the optimal value ϵ^* are unknown. We will therefore associate a 0/1-variable x_i to each point p_i ($i > 1$). For any $\epsilon > 0$, each assignment to these variables will uniquely define an arrangement of the points $p_1, \ldots, p_n$ in the following way. If $x_i = 1$, the point p_i is located $\epsilon/2$ to the right of $A^1(p_i)$ and if $x_i = 0$, the point p_i will be located $\epsilon/2$ to the left of $A^1(p_i)$.

DEFINITION 2.4. Let $X = \{x_2, \ldots, x_n\}$ be a set of 0/1-variables. For any $\epsilon > 0$ and any truth assignment $\mathcal{S} : x_i \mapsto \{0, 1\}$, for the variables in X, the arrangement $A^{\mathcal{S}}_\epsilon$ is defined by

$$A^{\mathcal{S}}_\epsilon(p_i) = \begin{cases} 0 & \text{if } i = 1, \\ A^1(p_i) - \epsilon/2 + \mathcal{S}(x_i)\epsilon & \text{otherwise.} \end{cases} \tag{5}$$

With $\epsilon = \epsilon^*$ and the assignment corresponding to the optimal solution, this arrangement will be the arrangement mentioned above.

LEMMA 2.2. *Let $\mathcal{S}^*$ be the truth assignment defined by*

$$\mathcal{S}^*(x_i) = \begin{cases} 1 & \textit{if } A^*(p_i) > A^1(p_i), \\ 0 & \textit{otherwise.} \end{cases} \tag{6}$$

Then $||A_{\epsilon^*}^{\mathcal{S}^*}, D||_\infty \leq 2\epsilon^*$.

Proof. For any pair of points p_i, p_j,

$$\begin{aligned} |A_{\epsilon^*}^{\mathcal{S}^*}(p_i) - A_{\epsilon^*}^{\mathcal{S}^*}(p_j)| &\leq |A_{\epsilon^*}^{\mathcal{S}^*}(p_i) - A^*(p_i)| \\ &\quad + |A_{\epsilon^*}^{\mathcal{S}^*}(p_j) - A^*(p_j)| + |A^*(p_i) - A^*(p_j)| \\ &\leq |A^*(p_i) - A^*(p_j)| + \epsilon^*. \end{aligned} \tag{7}$$

This implies that

$$\begin{aligned} &||A_{\epsilon^*}^{\mathcal{S}^*}(p_i) - A_{\epsilon^*}^{\mathcal{S}^*}(p_j)| - D[i,j]| \\ &\qquad \leq ||A^*(p_i) - A^*(p_j)| - D[i,j]| + \epsilon^* \leq 2\epsilon^*, \end{aligned} \tag{8}$$

which means that $||A_{\epsilon^*}^{\mathcal{S}^*}, D||_\infty \leq 2\epsilon^*$. ∎

However, to obtain a 2-approximation algorithm, it is not necessary to find the truth assignment $\mathcal{S}^*$. It is sufficient to find any assignment $\mathcal{S}$ and parameter $\epsilon > 0$ for which $||A_\epsilon^{\mathcal{S}}, D||_\infty \leq 2\epsilon^*$. For each pair of points p_i and p_j, there are four possible ways to assign values to the variables x_i and x_j.

DEFINITION 2.5. An assignment $\mathcal{S}$ is *ϵ-allowed* for the pair of variables x_i and x_j if $||A_\epsilon^{\mathcal{S}}(p_i) - A_\epsilon^{\mathcal{S}}(p_j)| - D[i,j]| \leq 2\epsilon$. If an assignment is not ϵ-allowed for a pair of points, it is said to be ϵ-forbidden for this pair.

LEMMA 2.3. *Let $\mathcal{S}$ be a truth assignment for the variables $x_2, \ldots, x_n$, such that $\mathcal{S}$ is ϵ-allowed for all pairs x_i, x_j. Then $||A_\epsilon^{\mathcal{S}}, D||_\infty \leq 2\epsilon$.*

Proof. Follows immediately from Definition 2.5. ∎

It is easy to construct a 2-SAT-clause that forbids a certain assignment to a pair x_i, x_j (see Table 1). If we create a 2-SAT-formula forbidding all ϵ-forbidden assignments, any satisfying assignment to that formula will have the property we are looking for.

THEOREM 2.1. *For each pair x_i, x_j, let $\varphi_{i,j}^\epsilon$ be the conjunction of the at most four clauses forbidding all ϵ-forbidden assignments for the pair, and let $\Phi_\epsilon = \bigwedge_{i \neq j} \varphi_{i,j}^\epsilon$. Every satisfying assignment $\mathcal{S}$ to Φ_ϵ satisfies $||A_\epsilon^{\mathcal{S}}, D||_\infty \leq 2\epsilon$. Furthermore $\mathcal{S}^*$ is a satisfying assignment for Φ_{ϵ^*}.*

Proof. Let $\mathcal{S}$ be any satisfying assignment for Φ_ϵ. By construction, $\mathcal{S}$ is ϵ-allowed for each pair of variables x_i, x_j, so from Lemma 2.3 follows that $||A_\epsilon^{\mathcal{S}}, D||_\infty \leq 2\epsilon$.

TABLE 1.
Clauses forbidding ϵ-forbidden assignments

x_i	x_j	Clause
0	0	$(x_i \vee x_j)$
1	0	$(\overline{x_i} \vee x_j)$
0	1	$(x_i \vee \overline{x_j})$
1	1	$(\overline{x_i} \vee \overline{x_j})$

From Lemma 2.2 follows that $||A^{\mathcal{S}^*}_{\epsilon^*}, D||_\infty \leq 2\epsilon^*$ so $\mathcal{S}^*$ is ϵ^*-allowed for all pairs of points and thus a satisfying assignment for Φ_{ϵ^*}. ▮

To find an $\epsilon \leq \epsilon^*$ for which Φ_ϵ has a satisfying assignment, we will use some properties of these formulae that is due to their construction.

LEMMA 2.4. *If c is a clause in Φ_ϵ, then c is a clause in $\Phi_{\epsilon'}$ for all $0 \leq \epsilon' \leq \epsilon$.*

Proof. Let $\epsilon = \epsilon' + \delta$, where $0 \leq \delta \leq \epsilon$, and assume that c is a clause in Φ_ϵ forbidding an assignment to the pair x_i, x_j. Let $\mathcal{S}$ be an assignment that does not satisfy c, i.e.,

$$||A^{\mathcal{S}}_\epsilon(p_i) - A^{\mathcal{S}}_\epsilon(p_j)| - D[i,j]| > 2\epsilon. \tag{9}$$

From Definition 2.4, using the triangle inequality and Eq. (9), follows that

$$\begin{aligned}
&||A^{\mathcal{S}}_{\epsilon'}(p_i) - A^{\mathcal{S}}_{\epsilon'}(p_j)| - D[i,j]| \\
&\quad = ||A^1(p_i) - A^1(p_j) + \epsilon'(\mathcal{S}(x_i) - \mathcal{S}(x_j))| - D[i,j]| \\
&\quad = ||A^1(p_i) - A^1(p_j) + (\epsilon - \delta)(\mathcal{S}(x_i) - \mathcal{S}(x_j))| - D[i,j]| \\
&\quad \geq ||A^1(p_i) - A^1(p_j) + \epsilon(\mathcal{S}(x_i) - \mathcal{S}(x_j))| - D[i,j]| \\
&\qquad - \delta|\mathcal{S}(x_i) - \mathcal{S}(x_j)| \\
&\quad = ||A^{\mathcal{S}}_\epsilon(p_i) - A^{\mathcal{S}}_\epsilon(p_j)| - D[i,j]| - \delta|\mathcal{S}(x_i) - \mathcal{S}(x_j)| \\
&\quad > 2\epsilon - \delta|\mathcal{S}(x_i) - \mathcal{S}(x_j)|.
\end{aligned} \tag{10}$$

But $|\mathcal{S}(x_i) - \mathcal{S}(x_j)| \leq 1$, so

$$||A^{\mathcal{S}}_{\epsilon'}(p_i) - A^{\mathcal{S}}_{\epsilon'}(p_j)| - D[i,j]| > 2\epsilon', \tag{11}$$

which means that the assignment is ϵ'-forbidden as well; so c is a clause in $\Phi_{\epsilon'}$. ▮

DEFINITION 2.6. An $\epsilon' \in \boldsymbol{R}^+$ is a *breakpoint* if $\Phi_{\epsilon'} \neq \Phi_\epsilon$ for all $\epsilon < \epsilon'$.

From Lemma 2.4 follows that if $\mathcal{S}$ is a satisfying assignment for Φ_ϵ, then $\mathcal{S}$ also satisfies $\Phi_{\epsilon'}$ for all $\epsilon' > \epsilon$. This means that, if there is only a small number of breakpoints, we could use binary search over the breakpoints to find an $\epsilon \leq \epsilon^*$ for which Φ_ϵ has a satisfying assignment. (Note that $\mathcal{S}^*$ is a satisfying assignment for the 2-SAT-formula corresponding to the greatest breakpoint $\leq \epsilon^*$.)

THEOREM 2.2. *There can be at most* $3\binom{n}{2}$ *breakpoints.*

Proof. From Lemma 2.4 and Definition 2.6 it follows that for each breakpoint ϵ' there exists a clause c such that: 1) c is a clause in Φ_ϵ for all $\epsilon < \epsilon'$, and 2) c is not a clause in Φ_ϵ for all $\epsilon \geq \epsilon'$. Each clause thus corresponds to at most one breakpoint. Furthermore there are $\binom{n}{2}$ pairs of variables x_i, x_j and 4 possible clauses for each such pair. However, if $\mathcal{S}$ and $\mathcal{S}'$ are two assignments such that $\mathcal{S}(x_i) = \mathcal{S}(x_j) \neq \mathcal{S}'(x_i) = \mathcal{S}'(x_j)$, then

$$\begin{aligned} ||A_\epsilon^{\mathcal{S}}(p_i) - A_\epsilon^{\mathcal{S}}(p_j)| - D[i,j]| &= ||A^1(p_i) - A^1(p_j)| - D[i,j]| \\ &= ||A_\epsilon^{\mathcal{S}'}(p_i) - A_\epsilon^{\mathcal{S}'}(p_j)| - D[i,j]|; \end{aligned} \tag{12}$$

so the two clauses $(x_i \vee x_j)$ and $(\overline{x_i} \vee \overline{x_j})$, forbidding assignments where $x_i = x_j$, will correspond to the same breakpoint. This makes the total number of breakpoints at most $3\binom{n}{2}$. ∎

We are now ready to formulate the approximation algorithm for MATRIX-TO-LINE

ALGORITHM 2.1.
1. Construct the set of breakpoints.
2. Use binary search over the breakpoints to find the smallest breakpoint ϵ for which Φ_ϵ has a satisfying assignment $\mathcal{S}$.
3. Return $A_\epsilon^{\mathcal{S}}$.

THEOREM 2.3. *Algorithm 2.1 approximates* MATRIX-TO-LINE *within 2 in time* $O(n^2 \log(n))$*, if the leftmost point in an optimal arrangement is known.*

Proof. The correctness of the algorithm follows from the derivations above. What we need to show is the time bound. Given the leftmost point in the optimal arrangement, we can compute the $O(n^2)$ breakpoints in time $O(n^2)$ and sort them in time $O(n^2 \log(n))$. In each step of the binary search we construct and solve a 2-SAT-formula with $O(n^2)$ clauses. 2-SAT can be solved in linear time, so the total time for that part of the algorithm is $O(n^2 \log(n))$. Hence the total running time of Algorithm 2.1 is $O(n^2 \log(n))$. ∎

If the leftmost point in an optimal arrangement is unknown, we can try all possible choices to find the correct one.

THEOREM 2.4. *The* MATRIX-TO-LINE *problem can be approximated within 2 in time* $O(n^3 \log(n))$.

Proof. From Theorem 2.3 follows that, given that the leftmost point is known, Algorithm 2.1 approximates MATRIX-TO-LINE within 2 in time $O(n^2 \log(n))$. There are at most n choices for the leftmost point, so the time required for running Algorithm 2.1 for each such choice is $O(n^3 \log(n))$. ∎

Although the number of choices for the leftmost point is n in the worst case, simple heuristics should limit the number of choices considerably in most cases.

3. LOWER BOUNDS

In this section, we first prove a lower bound of $7/5$ on the approximability of MATRIX-TO-LINE under the assumption P $\neq$ NP; thereafter, we show that if MATRIX-TO-LINE is approximable within $2 - \delta$ in polynomial time, then every 3-colorable graph can be $\lceil 4/\delta \rceil$-colored in polynomial time. The problem of k-coloring a 3-colorable graph is a well studied problem. The problem is not known to be NP-hard. The best result so far is from [9] where they show that it is NP-hard to find a 4-coloring of a 3-colorable graph. However, the best approximation algorithm known for the corresponding optimization problem MINIMUM GRAPH COLORING for 3-colorable graphs, has performance ratio $\tilde{O}(n^{3/14})$ [4] (i.e., $O(n^{3/14} \log^k(n))$ for some constant k).

Deciding NOT-ALL-EQUAL-3-SAT was shown to be NP-complete by Schaefer [13], and it is defined as follows.

DEFINITION 3.1. Let X be a set of variables and let C be a collection of clauses over X, such that each clause $c \in C$ has three literals. Then (X, C) belongs to NOT-ALL-EQUAL-3-SAT if there is a truth assignment that for each clause $c \in C$ assigns at least one literal of c the value true and at least one literal of c the value false.

The lower bound of $7/5$ is obtained by the following reduction.

REDUCTION 3.1. *Given an* NOT-ALL-EQUAL-3-SAT *instance* (X, C)*, we define a corresponding* MATRIX-TO-LINE *instance* (P, D) *in the following way. For each variable* x *and its complement* $\bar{x}$*, there are two points,* p_x *and* $p_{\bar{x}}$*, in* P*. For each clause* c*, there are three points* c_1*,* c_2*, and* c_3*, in* P*. In addition to these points,* P *contains the point* b*. The distance matrix* D *has the following entries. For all points* $p \in P$*,* $D[b, p] = 0$*; for all variables* $x \in X$*,* $D[p_x, p_{\bar{x}}] = 9$*; furthermore, for all clauses* $(u \vee v \vee w) \in C$*,*

$$
\begin{array}{llll}
D[c_1, p_u] = 6 & D[c_1, p_v] = 0 & D[c_1, p_{\bar{u}}] = 0 & D[c_1, p_{\bar{v}}] = 6 \\
D[c_2, p_v] = 6 & D[c_2, p_w] = 0 & D[c_2, p_{\bar{v}}] = 0 & D[c_2, p_{\bar{w}}] = 6 \\
D[c_3, p_w] = 6 & D[c_3, p_u] = 0 & D[c_3, p_{\bar{w}}] = 0 & D[c_3, p_{\bar{u}}] = 6 \\
D[c_1, c_2] = 6 & D[c_1, c_3] = 6 & D[c_2, c_3] = 6. &
\end{array}
$$

All other distances are equal to 3.

The intuition behind the construction is that a literal x assigned with the value true corresponds to a point p_x being to the right of b, and a literal assigned with the value false corresponds to a point being to the left of b.

LEMMA 3.1. *If an instance (X, C) belongs to* NOT-ALL-EQUAL-3-SAT, *then the corresponding* MATRIX-TO-LINE *instance (P, D) has optimal value 3.*

Proof. Since $D[p_x, p_{\bar{x}}] = 9$ and $D[b, p_x] = D[b, p_{\bar{x}}] = 0$ it is clear that any arrangement of the points in P will have an error ≥ 3. Hence, to prove the lemma it suffices to construct an arrangement with error 3. If all points are located within an interval of length 6, then the error for all pairs where the specified distance is 3 will be at most 3. Therefore we will arrange all points within an interval of length 6 centered around the point b.

Let $\mathcal{S}$ be an assignment of the variables in X such that at least one literal in each clause is true and at least one literal is false. Since (X, C) belongs to NOT-ALL-EQUAL-3-SAT we know that such an assignment exists. For each literal x that is assigned with the value true we arrange the corresponding point p_x a distance 3 to the right of b. For each literal x that is assigned with the value false we arrange the corresponding point p_x a distance 3 to the left of b. In this way, the error for any pair p_x, $p_{\bar{x}}$ and any pair p_x, b will be 3.

To complete the construction of the arrangement we show how to arrange the points corresponding to the clauses, so that the error is at most 3. Let $c = (u \vee v \vee w)$ be an arbitrary clause in C. By construction, at least one of the points corresponding to literals in c will be arranged to the right of b, and at least one to the left. Fig. 1 shows how to arrange the points c_1, c_2 and c_3 in all the six possible cases. It is easy to check that the error for any pair is at most 3. This completes the proof. ∎

LEMMA 3.2. *If an instance (X, C) does not belong to* NOT-ALL-EQUAL-3-SAT, *then the corresponding* MATRIX-TO-LINE *instance (P, D) has optimal value $\geq 21/5$.*

Proof. If (X, C) does not belong to NOT-ALL-EQUAL-3-SAT, then for any assignment $\mathcal{S}$ there is some clause $c \in C$ such that all literals are true or all literals are false. This implies that in any arrangement of the points in P, three points

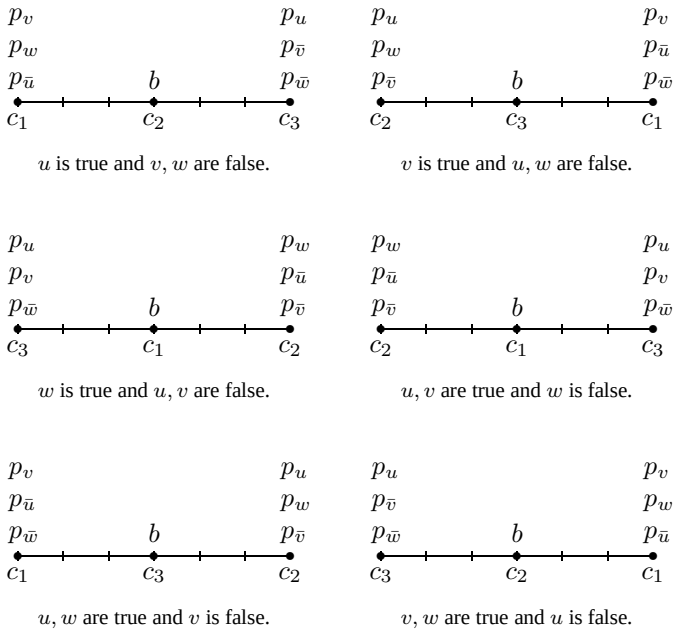

FIG. 1. The arrangement of a clause $c = (u \vee v \vee w)$.

corresponding to literals in some clause $c = (u \vee v \vee w)$ will be on the same side of b. Hence, to prove the lemma, it suffices to show that the error for any arrangement of the ten points: p_u, p_v, p_w, $p_{\bar{u}}$, $p_{\bar{v}}$, $p_{\bar{w}}$, c_1, c_2, c_3 and b, such that p_u, p_v, p_w are to the left of b, has error $\geq 21/5$.

For a fixed permutation of the points, the problem of finding the optimal arrangement can be formulated as a linear program. Let x_i be the location of the i:th point in the permutation and let $D[i, j]$ be the specified distance between the i:th and the j:th point in the permutation. Then the linear program can be formulated in the following way.

$$\begin{aligned} &\text{minimize} \quad z \\ &\text{s.t.} \begin{cases} x_j \geq x_i & \forall i < j \\ z \geq D[i,j] - (x_j - x_i) & \forall i < j \\ z \geq (x_j - x_i) - D[i,j] & \forall i < j \\ z \geq 0,\ x_i \geq 0 & \forall i \end{cases} \end{aligned} \tag{13}$$

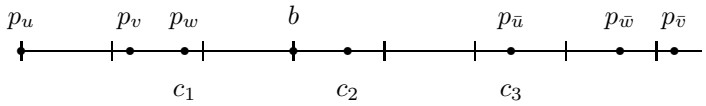

FIG. 2. The arrangement of a clause $c = (u \vee v \vee w)$, where the corresponding points are on the same side of b. In the arrangement $p_u = 0$, $p_v = 1.2$, $p_w = 1.8$, $p_{\bar{u}} = 5.4$, $p_{\bar{v}} = 7.2$, $p_{\bar{w}} = 6.6$, $b = 3$, $c_1 = 1.8$, $c_2 = 3.6$, $c_3 = 5.4$.

By symmetry, we can fix the order of the three points p_u, p_v and p_w. Furthermore, it is easy to see that any arrangement of the points such that p_x and $p_{\bar{x}}$ are on the same side of b will have an error $\geq 9/2$. This means that there are $3! \times 8 \times 9 \times 10 = 4320$ permutations that have to be considered. We solved the 4320 linear programs, using the publicly available package LP_SOLVE by Michael Berkelaar. The result shows that the smallest optimal value for these linear programs is $21/5$. Fig. 2 shows one of the optimal solutions corresponding to the optimal value $21/5$.

We have thus shown that if an arrangement of the points in P, for some clause c, has all points corresponding to literals in c on the same side of b, then the error is $\geq 21/5$. ∎

THEOREM 3.1. *It is NP-hard to approximate* MATRIX-TO-LINE *within* $7/5 - \delta$*, for any* $\delta > 0$.

Proof. From Lemma 3.1 follows that if an instance (X, C) belongs to NOT-ALL-EQUAL-3-SAT, then the corresponding MATRIX-TO-LINE instance (P, D) has optimal value 3. Furthermore, from Lemma 3.2 follows that if (X, C) does not belong to NOT-ALL-EQUAL-3-SAT, then the optimal value is $\geq 21/5$. This means that if there exists a polynomial time approximation algorithm for MATRIX-TO-LINE with performance ratio less than $7/5$, then this algorithm will decide NOT-ALL-EQUAL-3-SAT in polynomial time. ∎

Now, we show that it is at least as hard to approximate MATRIX-TO-LINE within $2 - \delta$ as it is to $\lceil 4/\delta \rceil$-color a 3-colorable graph. Consider the following reduction.

REDUCTION 3.2. *Given a 3-colorable graph* $G = (V, E)$ *with* n *vertices, we define the corresponding* MATRIX-TO-LINE *instance* (P, D)*, where each point*

$p_i \in P$ corresponds to a vertex $v_i \in V$ and

$$D[i,j] = \begin{cases} 2 & \textit{if } (v_i, v_j) \in E, \\ 1 & \textit{otherwise.} \end{cases} \tag{14}$$

The idea is to find an arrangement of the points in P, using a $(2-\delta)$-approximation algorithm for MATRIX-TO-LINE; and from this arrangement construct a coloring of G with $\lceil 4/\delta \rceil$ colors. To do this, we will use an upper bound on the error for an optimal arrangement of the points in P. Such a bound is possible to obtain using the 3-coloring of G, that exists by assumption.

DEFINITION 3.2. Given a 3-coloring, $c : V \to \{0,1,2\}$, of a graph G, let A_c be the arrangement defined by $A_c(p_i) = c(v_i)$, for each point p_i in the corresponding MATRIX-TO-LINE instance.

LEMMA 3.3. *Let c be a 3-coloring of a graph G. Then $||A_c, D||_\infty \leq 1$*

Proof. For any pair of points p_i, p_j, the distance $D[i,j]$ is either 1 or 2. From Definition 3.2 it follows immediately that $0 \leq |A_c(p_i) - A_c(p_j)| \leq 2$ for all i, j. So, if $D[i,j] = 1$, then $||A_c(p_i) - A_c(p_j)| - D[i,j]| \leq 1$. By construction $D[i,j] = 2$ if and only if there is an edge between v_i and v_j in G. But if $(v_i, v_j) \in E$, then $c(v_i) \neq c(v_j)$, implying that $1 \leq |A_c(p_i) - A_c(p_j)| \leq 2$; and hence $||A_c(p_i) - A_c(p_j)| - D[i,j]| \leq 1$. ∎

Let A be the arrangement produced by a $(2-\delta)$-approximation algorithm for MATRIX-TO-LINE. W.l.o.g. we assume that p_1 is the leftmost point in the arrangement A and that $A(p_1) = 0$. From Lemma 3.3 follows that

$$||A, D||_\infty \leq 2 - \delta, \tag{15}$$

which means that

$$A(p_i) \leq 4 - \delta \quad \forall i, \tag{16}$$

since $D[1,i] \leq 2$ for all i. Eq. (16) implies that we can cover the interval containing all points in A with $\lceil 4/\delta \rceil$ disjoint sub-intervals of equal length $d < \delta$. These sub-intervals will induce a coloring of the vertices in G.

DEFINITION 3.3. Let $G = (V, E)$ be a 3-colorable graph, let (P, D) be the corresponding MATRIX-TO-LINE instance, and let A be the arrangement of P produced by a $(2-\delta)$-approximation algorithm for MATRIX-TO-LINE. Define $c_A : V \to \{1, \ldots, \lceil 4/\delta \rceil\}$ in the following way. For each point $p_i \in P$, let $c_A(v_i) = j$, if and only if $A(p_i) \in [(j-1)d, jd)$, where $d = (4 - \delta/2)/\lceil 4/\delta \rceil$.

LEMMA 3.4. *c_A is a $\lceil 4/\delta \rceil$-coloring of G.*

Proof. We need to show that $c_A(v_i) \neq c_A(v_j)$ whenever $(v_i, v_j) \in E$. Assume that $(v_i, v_j) \in E$. This implies that $D[i, j] = 2$. Now, $||A, D||_\infty \leq 2 - \delta$ so in this case $|A(p_i) - A(p_j)| \geq \delta > d$, which means that $c_A(v_i) \neq c_A(v_j)$. ▮

We have thus proven the following theorem.

THEOREM 3.2. *If* MATRIX-TO-LINE *is approximable within $2 - \delta$ in polynomial time, then any 3-colorable graph can be colored with $\lceil 4/\delta \rceil$-colors in polynomial time.*

Proof. Let G be a three colorable graph. Given a $(2 - \delta)$ approximation of MATRIX-TO-LINE, we can construct c_A in Definition 3.3 in polynomial time. By Lemma 3.4, c_A is a $\lceil 4/\delta \rceil$-coloring of G. ▮

4. THE RH MODEL

A marker is a gene or an arbitrary DNA sequence for which there is an "easy" laboratory test for its presence in any fragment of DNA. Suppose that we want to construct a physical map of a human chromosome with respect to n markers; that is, we want to find the order in which the markers appear on the chromosome and the distance between them.

Since there is no direct procedure giving the orientation of a pair of markers on a fragment of DNA, an RH-experiment is performed. The chromosome is exposed to gamma radiation which shatters it into fragments. A subset of the fragments are incorporated into a hamster cell, which is grown to yield a hybrid cell line. Each marker is then tested for presence in cells from this cell line.

The outcome of one experiment is represented by a vector in $\{0, 1\}^n$ where 1 corresponds to presence of the marker in the hybrid cell line; a number of experiments are in the natural way represented by a 0/1-matrix. Such a 0/1-matrix is the laboratory data which is the input to the algorithmic problem. That is, the RH problem is that of, given a 0/1-matrix, finding the order of the markers and the distance between them.

We use the following model of an RH experiment, which is basically the same model as in [2, 10], but without the assumption that the markers are uniformly distributed. A genome is modeled by the unit interval $[0, 1]$. A set of n markers is modeled by a function $A : [n] \rightarrow [0, 1]$; that is, each marker is a point in $[0, 1]$. (The former is just a question of scaling. The latter is motivated by the fact that compared to the genome the markers are very short.) An *experiment*

for $A : [n] \to [0,1]$ is the following probabilistic procedure in which a vector $v \in \{0,1\}^n$ is produced.

1. Breaks are distributed in the unit interval $[0,1]$ according to a Poisson process with rate λ. This induces a division of $[0,1]$ into maximal subintervals without breaks, denoted $I_1, \ldots, I_l$. (This models how radiation breaks the genome.)

2. A set S is constructed by, for each subinterval I_i, letting I_i belong to the set S with probability p. (This models how some fragments are incorporated into the hamster genome.) Let $I = \cup_{I_i \in S} I_i$.

3. For each $i \in [n]$, if $A(i) \in I$ let $v(i) = 1$ with probability $1 - \beta$ and otherwise let $v(i) = 1$ with probability α. (This models the negative and positive errors that can occur when a hamster genome is tested for presence of a marker.)

In this way each $A : [n] \to [0,1]$ induces a probability distribution P_A on $\{0,1\}^n$; that is, for each $x \in \{0,1\}^n$, $P_A(x)$ is the probability that an experiment for A produces x.

Since everything to the left of the leftmost marker in a genome will be unknown to us we will assume that this marker is located in $x = 0$.

DEFINITION 4.1. A *marker function* is a function $A : [n] \to [0,1]$ such that $A(m) = 0$ for some $m \in [n]$, and the leftmost marker has lower index than the rightmost marker.

The last condition is present to assure that, for any probability distribution on $\{0,1\}^n$, there is a unique corresponding marker function.

5. FINDING THE GENOME MAP

In this section, we give an algorithm for the RH problem using the algorithm for MATRIX-TO-LINE from Section 2.

Let A be the unknown marker function representing the genome we want to study, and let D be the distance matrix defined by $D[i,j] = |A(i) - A(j)|$. We show how to, given m experiments for A, construct a marker function $\hat{A}$ such that $||\hat{A}, D||_\infty \leq O(\sqrt{\log(n)/m})$, using the 2-approximation algorithm from Section 2. In fact, any approximation algorithm with constant performance ratio will give this bound, but the constant hidden in the O notation will be proportional to $1 + \tau$, where τ is the performance ratio of the MATRIX-TO-LINE algorithm.

DEFINITION 5.1. Two markers i and j are *separated* by an experiment if $v(i) \neq v(j)$, where $v \in \{0,1\}^n$ is the outcome of the experiment.

In [2], an expression for the probability of two markers being separated was derived for the case when $\alpha = \beta$. This derivation, which involves case analysis, can easily be generalized to the case where $\alpha \neq \beta$, and yields the following

expression.

$$\varphi_{i,j} = 2pq(1 - e^{-\lambda D[i,j]})(1 - (\alpha + \beta))^2 + g(\alpha, \beta, p), \tag{17}$$

where $q = 1 - p$ and

$$g(\alpha, \beta, p) = 2p(\alpha - \beta)(\alpha + \beta - 1) + 2\alpha(1 - \alpha). \tag{18}$$

Since $0 \leq D[i,j] \leq 1$ for all i, j, the separation probability satisfies $\varphi_{\min} \leq \varphi_{i,j} \leq \varphi_{\max}$ for all i, j, where

$$\varphi_{\min} = g(\alpha, \beta, p) \tag{19}$$

and

$$\varphi_{\max} = 2pq(1 - e^{-\lambda})(1 - (\alpha + \beta))^2 + g(\alpha, \beta, p). \tag{20}$$

Solving for $D[i,j]$ in (17) shows that $D[i,j] = \delta(\varphi_{i,j})$, where

$$\delta(\varphi) = -\frac{1}{\lambda} \ln\big(1 - \frac{\varphi - g(\alpha, \beta, p)}{2pq(1 - (\alpha + \beta))^2}\big). \tag{21}$$

Let $X_{i,j} \in \{0, 1\}$ denote the output for the i:th marker in experiment j. By calculating the frequency of separation between the markers i and j,

$$\nu_{i,j} = \frac{1}{m} \sum_{k=1}^{m} |X_{ki} - X_{kj}|, \tag{22}$$

we get an estimate of $\varphi_{i,j}$ by

$$\hat{\varphi}_{i,j} = \begin{cases} \varphi_{\min} & \text{if } \nu_{i,j} < \varphi_{\min} \\ \nu_{i,j} & \text{if } \varphi_{\min} \leq \nu_{i,j} \leq \varphi_{\max} \\ \varphi_{\max} & \text{if } \nu_{i,j} > \varphi_{\max} \end{cases} \tag{23}$$

which we use in Eq. (21) to obtain an estimate $\hat{D}[i,j] = \delta(\hat{\varphi}_{i,j})$ of $D[i,j]$. The idea is to apply the 2-approximation algorithm for MATRIX-TO-LINE to the estimated distances $\hat{D}[i,j]$ in order to find a marker function $\hat{A}$ close to the true marker function A.

To show how close $\hat{A}$ is to A, we will use a bound on $||D, \hat{D}||_\infty$. The function $\delta(\varphi)$ is differentiable with respect to φ in $[\varphi_{\min}, \varphi_{\max}]$. Since

$$\delta'(\varphi) = \frac{1}{2\lambda pq(1 - (\alpha + \beta))^2 - \lambda(\varphi - g(\alpha, \beta, p))}, \tag{24}$$

it follows from the mean value theorem that for any pair i, j,

$$\begin{aligned} |D[i,j] - \hat{D}[i,j]| &\leq \frac{|\varphi_{i,j} - \hat{\varphi}_{i,j}|}{|2\lambda pq(1-(\alpha+\beta))^2 - \lambda(\varphi_{\max} - g(\alpha,\beta,p))|} \\ &= \frac{|\varphi_{i,j} - \hat{\varphi}_{i,j}|}{2\lambda pq(1-(\alpha+\beta))^2} e^{\lambda}. \end{aligned} \tag{25}$$

If we apply Hoeffding's inequality [8] for $\hat{\varphi}_{i,j}$, we see that for each pair i and j

$$\Pr[\,|\varphi_{i,j} - \hat{\varphi}_{i,j}| \geq \tau\,] \leq 2e^{-2m\tau^2}. \tag{26}$$

Hence, the probability σ that there is a pair i, j such that $|\varphi_{i,j} - \hat{\varphi}_{i,j}| \geq \tau$, is less than $n^2 e^{-2m\tau^2}$. We thus conclude that, with error probability σ,

$$||D, \hat{D}||_\infty < \frac{\sqrt{\ln(n^2/\sigma)}}{2pq\lambda\sqrt{2m}(1-(\alpha+\beta))^2} e^{\lambda} \tag{27}$$

LEMMA 5.1. *If $\hat{A}$ is the marker function obtained by applying the 2-approximation algorithm for* MATRIX-TO-LINE *on the matrix $\hat{D}$, then*

$$||\hat{A}, D||_\infty \leq \frac{3\sqrt{\ln(n^2/\sigma)}}{2pq\lambda\sqrt{2m}(1-(\alpha+\beta))^2} e^{\lambda} \tag{28}$$

with probability $1 - \sigma$.

Proof. Since the approximation algorithm has performance ratio 2 we know that $||\hat{A}, \hat{D}||_\infty \leq 2||D, \hat{D}||_\infty$. Together with the triangle inequality and Eq. (27) this show that

$$\begin{aligned} ||\hat{A}, D||_\infty &\leq ||\hat{A}, \hat{D}||_\infty + ||\hat{D}, D||_\infty \leq 3||\hat{D}, D||_\infty \\ &\leq \frac{3\sqrt{\ln(n^2/\sigma)}}{2pq\lambda\sqrt{2m}(1-(\alpha+\beta))^2} e^{\lambda}. \end{aligned} \tag{29}$$

∎

Lemma 5.1 states that the distances between the markers in the arrangement $\hat{A}$ are close to the distances in the true arrangement. This implies that the positions of the markers are close as well. However, to show how close the arrangements are, i.e., to give an upper bound on the distance between the arrangements, turns out to be surprisingly technical. The next lemma gives such a bound for functions satisfying certain technical conditions. We then show that this lemma is applicable for marker functions.

LEMMA 5.2. *Let $f, g : [n] \to [0, 1]$ be two functions such that $f(p) = g(q) = 0$, $f(q) \leq f(r)$, and $g(p) \leq g(r)$, for some $p, q, r \in [n]$. If*

$$\max_{i,j \in [n]} ||f(i) - f(j)| - |g(i) - g(j)|| \leq \epsilon, \tag{30}$$

then $|f(i) - g(i)| \leq 2\epsilon$, for all i.

Proof. If $\exists m \in [n]$ such that $f(m) = g(m) = 0$, then $|f(i) - g(i)| \leq \epsilon$ for all $i \in [n]$; and the lemma holds. By assumption, there exists integers $p, q, r \in [n]$ such that $f(p) = g(q) = 0$, $f(q) \leq f(r)$, and $g(p) \leq g(r)$; and, by symmetry, we can assume that $f(q) \leq g(p)$. If

$$||f(i) - f(j)| - |g(i) - g(j)|| \leq \epsilon, \tag{31}$$

for all $i, j \in [n]$, we know that

$$|f(r) - (g(r) - g(p))| \leq \epsilon, \tag{32}$$
$$|g(r) - (f(r) - f(q))| \leq \epsilon. \tag{33}$$

Using the triangle inequality on the sum of these equations we observe that

$$g(p) + f(q) \leq 2\epsilon. \tag{34}$$

Eq. (34) together with the inequality $f(q) \leq g(p)$ show that

$$f(q) \leq \epsilon. \tag{35}$$

Choose an arbitrary $j \in [n]$, and assume that $f(j) \leq f(q)$. If $g(j) \leq f(j)$, then it follows from Eq. (35) that $|f(j) - g(j)| \leq \epsilon$. If $g(j) > f(j)$, then

$$|f(j) - g(j)| \leq |g(j) - g(q)| \leq |f(j) - f(q)| + \epsilon \leq 2\epsilon. \tag{36}$$

Now, assume that $f(j) > f(q)$. If $f(j) \leq g(j)$, then

$$|f(j) - g(j)| \leq |g(j) - g(q)| - |f(j) - f(q)| \leq \epsilon. \tag{37}$$

If $f(j) > g(j)$, then

$$|f(j) - g(j)| \leq |f(j) - f(q)| + \epsilon - |g(j) - g(q)| \leq 2\epsilon. \tag{38}$$

We have thus shown that $|f(i) - g(i)| \leq 2\epsilon$ for all $i \in [n]$. ∎

Not every pair of functions will satisfy the conditions in Lemma 5.2. However, if the lemma is not applicable to f and g, then it is applicable to f and the reversal of g.

DEFINITION 5.2. Let $f : [n] \to [0, 1]$ be a function and let r be the integer satisfying $f(r) \geq f(i)$ for all i. Define the *reversal* $\bar{f}$ of f by $\bar{f}(i) = f(r) - f(i)$ for all i.

LEMMA 5.3. *For any pair of functions $f, g : [n] \to [0, 1]$ mapping some value to 0, there exists $p, q, r \in [n]$ such that either*

$$f(p) = g(q) = 0, \quad f(q) \leq f(r), \quad g(p) \leq g(r), \tag{39}$$

or

$$f(p) = \bar{g}(q) = 0, \quad f(q) \leq f(r), \quad \bar{g}(p) \leq \bar{g}(r). \tag{40}$$

Proof. Let p and q be defined by $f(p) = g(q) = 0$ and let r be defined by $f(r) \geq f(i)$ for all i. If $g(p) \leq g(r)$, then all conditions in (39) are satisfied. Assume therefore that $g(p) > g(r)$. Let s be defined by $g(s) \geq g(i)$ for all i. From Definition 5.2 follows that $\bar{g}(s) = 0$ and $\bar{g}(p) < \bar{g}(r)$. Hence, since $f(r) \geq f(i)$ for all i, this implies that $f(p) = \bar{g}(s) = 0$, $f(s) \leq f(r)$ and $\bar{g}(p) \leq \bar{g}(r)$. ∎

THEOREM 5.1. *If $\hat{A}$ is the marker function obtained by applying the 2-approximation algorithm for* MATRIX-TO-LINE *on the matrix $\hat{D}$, estimated using m experiments for the marker function A, then*

$$\max_i\{|\hat{A}(i) - \tilde{A}(i)|\} \leq \frac{6\sqrt{\ln(n^2/\sigma)}}{2pq\lambda\sqrt{2m}(1 - (\alpha + \beta))^2}e^{\lambda}, \tag{41}$$

with probability $1 - \sigma$, where $\tilde{A}$ is either A or $\bar{A}$.

Proof. Lemma 5.1 states that

$$||\hat{A}, D||_\infty \leq \frac{3\sqrt{\ln(n^2/\sigma)}}{2pq\lambda\sqrt{2m}(1 - (\alpha + \beta))^2}e^{\lambda}. \tag{42}$$

From Lemma 5.2 and Lemma 5.3 then immediately follows that, for either $\tilde{A} = A$ or $\tilde{A} = \bar{A}$,

$$|\hat{A}(i) - \tilde{A}(i)| \leq \frac{6\sqrt{\ln(n^2/\sigma)}}{2pq\lambda\sqrt{2m}(1 - (\alpha + \beta))^2}e^{\lambda}, \tag{43}$$

for all $i \in [n]$. ∎

If there is a lower bound on the minimum distance between any pair of markers Theorem 5.1 gives an upper bound on the number of experiments required to find the true order of the markers.

COROLLARY 5.1. *Let $D_{\min}$ be the minimum distance between any pair of markers in the marker function A, i.e., $D_{\min} = \min_{i,j}\{D[i,j]\}$. If $\hat{A}$ is the marker function obtained by applying the 2-approximation algorithm for* MATRIX-TO-LINE *on the matrix $\hat{D}$, estimated using*

$$m \geq \frac{18\ln(n^2/\sigma)}{p^2q^2\lambda^2 D_{min}^2(1-(\alpha+\beta))^4}e^{2\lambda} \tag{44}$$

experiments for A, then the order of the markers in $\hat{A}$ will be the same as in A, with probability $1-\sigma$.

Proof. From Theorem 5.1 follows that, if the number of experiments m satisfy Eq. (44), then either $|\hat{A}(i) - A(i)| \leq D_{\min}/2$ or $|\hat{A}(i) - \bar{A}(i)| \leq D_{\min}/2$. This means that the order of the markers in $\hat{A}$ is the same as the order of the markers in either A or $\bar{A}$. However, since $\hat{A}$ is a marker function we know that the leftmost marker has lower index then the rightmost marker. Hence, the order of the markers in $\hat{A}$ must be the same as the order of the markers in A. ∎

Disregarding constants, the bound in Corollary 5.1 on the number of experiments is the same as the bound by Ben-Dor and Chor [2] for their algorithms to find the correct order of the markers.

Each marker function A induces a probability distribution P_A on the set $\{0,1\}^n$. The L^1-norm for these distributions can be used as a measure of the distance between marker functions.

DEFINITION 5.3. Let A and B be two marker functions. Define

$$L^1(P_A, P_B) = \sum_{x\in\{0,1\}^n} |P_A(x) - P_B(x)|. \tag{45}$$

This is the *variational distance* used for instance for Cavender-Farris trees in [6]. Following [6], it is possible to show that this is a metric for the marker functions. It is easy to check that it is symmetric, positive, and satisfies the triangle inequality. What we need to show is that $L^1(P_A, P_B) = 0$ implies $A = B$. However, this follows from the fact that the distances between two markers i and j can be expressed as a function of the separation probability for i and j.

Under this metric the constructed marker function $\hat{A}$ converges to the true marker function A, as the number of experiments m increases. We show that

$L^1(P_A, P_{\hat{A}}) \leq O(n\sqrt{\log(n)/m})$. Furthermore, we show that if an algorithm M given the output of m experiments for any marker function A, returns an approximation $\hat{A}$ of A such that $L^1(P_A, P_{\hat{A}}) \leq f(m)$, then $f(m) \geq \Omega(1/m)$. Finally, we show how this bound gives a bound on the sum of the difference in the positions of the markers in A and $\hat{A}$.

Lemma 5.2 enables us to get an upper bound on the L^1-norm of the difference in probability distribution for two marker functions, for which the differences in distance between markers are bounded.

LEMMA 5.4. *If A and B are two marker functions, then*

$$L^1(P_A, P_B) \leq 2\lambda||A - B||_1 \leq 4\lambda n||A, B||_\infty, \tag{46}$$

where

$$||A - B||_1 = \sum_{i=1}^{n} |A(i) - B(i)|. \tag{47}$$

Proof. From the definition of an experiment it is clear that $P_A = P_{\bar{A}}$ for every marker function A. Therefore, w.l.o.g. we assume that there exist $p, q, r \in [n]$ such that $A(p) = B(q) = 0$, $A(q) \leq A(r)$, and $B(p) \leq B(r)$.

Assume that one experiment is performed simultaneously for A and B. If the set of markers is partitioned differently for A and B, then the probability for a certain outcome of the experiment may differ, but otherwise it will not. We call such a break a *dangerous break*. Each marker i induces a subinterval $[A(i), B(i)]$ (or $[B(i), A(i)]$ if $A(i) > B(i)$) of $[0, 1]$, within which each break is dangerous. The length of the union of all these intervals will be at most $||A-B||_1 \leq 2n||A, B||_\infty$, according to Lemma 5.2.

Since the breaks are distributed in the interval $[0, 1]$ according to a Poisson process with rate λ, the probability that at least one dangerous break occurs is at most $1 - e^{-\lambda||A-B||_1}$. Let F be the event that at least one dangerous break occurs. Then $\Pr_A[\, x|\bar{F}\,] = \Pr_B[\, x|\bar{F}\,]$ for all x, since the probability of incorporation of fragments in hamster cells and false answers in the test of occurrences of markers are independent of the size of a fragment.

$$\begin{aligned} L^1(P_A, P_B) &= \sum_{x\in\{0,1\}^n} |P_A(x) - P_B(x)| \\ &\leq \Pr[\, F\,] \sum_{x\in\{0,1\}^n} |\Pr_A[\, x \mid F\,] - \Pr_B[\, x \mid F\,]| \\ &\quad + \Pr[\, \bar{F}\,] \sum_{x\in\{0,1\}^n} |\Pr_A[\, x \mid \bar{F}\,] - \Pr_B[\, x \mid \bar{F}\,]| \\ &\leq 2\Pr[\, F\,] \leq 2(1 - e^{-\lambda||A-B||_1}) \\ &\leq 2\lambda||A - B||_1 \leq 4n\lambda||A, B||_\infty \end{aligned} \tag{48}$$

where we have used the inequality $1 - x \leq e^{-x}$. ▌

Combining Lemma 5.1 and Lemma 5.4, we obtain the following bound on the difference in distribution between $\hat{A}$ and A.

THEOREM 5.2. *With probability* $1 - \sigma$,

$$L^1(P_A, P_{\hat{A}}) \leq \frac{3\sqrt{2}n\sqrt{\ln(n^2/\sigma)}}{pq\sqrt{m}(1-(\alpha+\beta))^2}e^{\lambda}. \tag{49}$$

Proof. Follows immediately from Lemma 5.1 and Lemma 5.4. ▌

6. LOWER BOUNDS FOR RH ALGORITHMS

In this section, we show a lower bound on the convergence rate for any algorithm for the RH problem.

LEMMA 6.1. *Let A_1 and A_2 be two marker functions and let M be any decision procedure that given the output from m experiments for either A_1 or A_2 decides whether the experiments were performed for A_1 or A_2. If $e_1(M)$ is the probability that M is incorrect when the experiments are performed for A_1 and $e_2(M)$ is the probability that M is incorrect when the experiments are performed for A_2, then*

$$e(M) \geq \frac{1 - m\epsilon}{2}, \tag{50}$$

where $e(M) = \max\{e_1(M), e_2(M)\}$ and $\epsilon = L^1(P_{A_1}, P_{A_2})$.

Proof. This is a reformulation of Lemma 1 in [6]. ▌

Lemma 6.1 implies that it is interesting to study pairs of marker functions and their possible L_1 distances. We do this in Lemma 6.2 below.

LEMMA 6.2. *For each marker function A and constant K such that*

$$K \leq \frac{pq}{2}(1-\alpha-\beta)^2(1-e^{-\lambda}), \tag{51}$$

there is a marker function A' such that $L^1(P_A, P_{A'}) = K$.

Proof. Consider the two marker functions W_1 and W_2 defined in the following way.

$$W_1(k) = \begin{cases} 0 & \text{if } k \neq n, \\ 1 & \text{otherwise.} \end{cases} \qquad W_2(k) = \begin{cases} 0 & \text{if } k \neq n-1, \\ 1 & \text{otherwise.} \end{cases} \tag{52}$$

Let E be the event that $x_{n-2} = 1$ and $x_{n-1} = 1$ in the output x of an experiment. From the definition of W_1 and W_2 follows that

$$\Pr_{W_1}[\,E\,] = p(1-\beta)^2 + q\alpha^2, \tag{53}$$

$$\begin{aligned}\Pr_{W_2}[\,E\,] = (p^2(1-\beta)^2 + q^2\alpha^2 + 2pq\alpha(1-\beta))(1-e^{-\lambda}) \\ + (p(1-\beta)^2 + q\alpha^2)e^{-\lambda}.\end{aligned} \tag{54}$$

Together with the triangle inequality, this shows that

$$\begin{aligned}L^1(P_{W_1}, P_{W_2}) &= \sum_{x\in\{0,1\}^n} |P_{W_1}(x) - P_{W_2}(x)| \\ &\geq \sum_{\substack{x\in\{0,1\}^n \\ x_{n-2}=x_{n-1}=1}} |P_{W_1}(x) - P_{W_2}(x)| \\ &\geq |\Pr_{W_1}[\,E\,] - \Pr_{W_2}[\,E\,]| \\ &= (p(1-\beta)^2 + q\alpha^2)(1-e^{-\lambda}) \\ &\quad - (p^2(1-\beta)^2 + q^2\alpha^2 + 2pq\alpha(1-\beta))(1-e^{-\lambda}) \\ &= (pq(1-\beta)^2 + pq\alpha^2 - 2pq\alpha(1-\beta))(1-e^{-\lambda}) \\ &= pq(1-\alpha-\beta)^2(1-e^{-\lambda}).\end{aligned} \tag{55}$$

Finally, the triangle inequality for the L^1-norm together with Eq. (55) imply that

$$\max\{L^1(P_{W_1}, P_A), L^1(P_{W_2}, P_A)\} \geq \frac{pq}{2}(1-\alpha-\beta)^2(1-e^{-\lambda}). \tag{56}$$

By continuity, we thus conclude that there exists a marker function A' such that $L^1(P_A, P_{A'}) = K$, for any

$$K \leq \frac{pq}{2}(1-\alpha-\beta)^2(1-e^{-\lambda}). \tag{57}$$

∎

THEOREM 6.1. *Let $\mathcal{A}$ be any algorithm that for all marker functions A, given the output of m experiments for A, returns an approximation $\hat{A}$ such that $L^1(P_A, P_{\hat{A}}) \leq f(m)$ with probability $1-\sigma$, for some constant $\sigma < 1/2$. Then $f(m) \geq \Omega(1/m)$.*

Proof. Assume that $\mathcal{A}$ is an algorithm that, given the output of m experiments for any marker function A, returns an approximation $\hat{A}$ of A, such that $L^1(P_A, P_{\hat{A}}) \leq f(m)$, with probability $1-\sigma$. We want to show that

$f(m) \geq \Omega(1/m)$. Assume for contradiction that this is not the case, i.e., assume that $f(m) \leq o(1/m)$.

Let A be an arbitrary marker function. Since $f(m) \leq o(1/m)$, Lemma 6.2 implies that there exists a marker function A' such that $L^1(P_A, P_{A'}) = 3f(m)$, for large enough m.

Consider the following decision procedure M'. Given m experiments for the two marker functions A or A', it first runs $\mathcal{A}$ to obtain a marker function $\hat{A}$; Thereafter, it outputs A if $L^1(P_A, P_{\hat{A}}) \leq L^1(P_{A'}, P_{\hat{A}})$ and A' otherwise. Note that we do not have to consider the running time of M'. It is clear that with this definition, $e(M') \leq \sigma$. However, using Lemma 6.1, we see that

$$\sigma \geq \frac{1 - 3mf(m)}{2}, \tag{58}$$

which contradicts the assumption that $f(m) \leq o(1/m)$. ▮

The lower bound on $L^1(P_A, P_{\hat{A}})$ together with Lemma 5.4 immediately gives a lower bound on $||A - \hat{A}||_1$.

COROLLARY 6.1. *Let $\mathcal{A}$ be any algorithm that for all marker functions A, given the output of m experiments for A, returns an approximation $\hat{A}$ such that $||A - \hat{A}||_1 \leq f(m)$ with probability $1 - \sigma$, for some constant $\sigma < 1/2$. Then $f(m) \geq \Omega(1/m)$.*

Proof. Lemma 5.4 states that $L^1(P_A, P_B) \leq 2\lambda||A - B||_1$. ▮

ACKNOWLEDGMENT

We thank Timothy F. Havel for kindly answering questions concerning the complexity of the MATRIX-TO-LINE problem and pointing us to [12]. We also like to thank Gunnar Andersson for useful remarks.

REFERENCES

1. Richa Agarwala, Vineet Bafna, Martin Farach, Babu Narayanan, Mike Paterson, and Mikkel Thorup. On the approximability of numerical taxonomy (fitting distances by tree metrics). In *Proceedings of the Seventh Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 365–372, 1996.
2. Amir Ben-Dor and Benny Chor. On constructing radiation hybrid maps. In *Proceedings of the First International Conference on Computational Molecular Biology*, pages 17–26, 1997.
3. D. Timothy Bishop and Gillian P. Crockford. Comparison of radiation hybrid mapping and linkage mapping. *Cyt. Cell Genet.*, 59:93–95, 1992.
4. Avrim Blum and David Karger. An $\tilde{O}(n^{3/14})$-coloring algorithm for 3-colorable graphs. *Inform. Process. Lett.*, 61(1):49–53, 1997.

5. Andreas W. M. Dress and Timothy F. Havel. Bound smoothing under chirality constraints. *SIAM J. Disc. Math.*, 4:535–549, 1991.

6. Martin Farach and Sampath Kannan. Efficient algorithms for inverting evolution. In *Proceedings of the Twenty-Eighth Annual ACM Symposium on Theory of Computing*, pages 230–236, 1996.

7. Martin Farach, Sampath Kannan, and Tandy Warnow. A robust model for finding optimal evolutionary trees (extended abstract). In *Proceedings of the Twenty-Fifth Annual ACM Symposium on Theory of Computing*, pages 137–145, 1993.

8. Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58(301):13–30, March 1963.

9. Sanjeev Khanna, Nathan Linial, and Shmuel Safra. On the hardness of approximating the chromatic number. In *Proceedings of the 2nd Israel Symposium on Theory and Computing Systems, ISTCS*, pages 250–260. IEEE Computer Society Press, 1993.

10. Kenneth Lange, Michael Boehnke, David R. Cox, and Kathryn L. Lunetta. Statistical methods for polyploid radiation hybrid mapping. *Genome Res.*, 5:136–150, 1995.

11. Nathan Linial, Eran London, and Yuri Rabinovich. The geometry of graphs and some of its algorithmic applications. *Combinatorica*, 15:215–245, 1995.

12. James B. Saxe. Embeddability of graphs in k-space is strongly NP-hard. In *17th Allerton Conference in Communication, Control, and Computing*, pages 480–489, 1979.

13. Thomas J. Schaefer. The complexity of satisfiability problems. In *Conference Record of the Tenth Annual ACM Symposium on Theory of Computing*, pages 216–226, 1978.

14. Donna Slonim, Leonid Kruglyak, Lincoln Stein, and Eric Lander. Building human genome maps with radiation hybrids. In *Proceedings of the First International Conference on Computational Molecular Biology*, pages 277–286, 1997.