Tight Approximability Results for Test Set Problems in Bioinformatics

Piotr Berman[1], Bhaskar DasGupta[2], and Ming-Yang Kao[3]

[1] Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA 16802. Email: berman@cse.psu.edu.
[2] Department of Computer Science, University of Illinois at Chicago, Chicago, IL 60607-7053. Email: dasgupta@cs.uic.edu.
[3] Department of Computer Science, Northwestern University, Evanston, IL 60201. Email: kao@cs.northwestern.edu.

Abstract. In this paper, we investigate the test set problem and its variations that appear in a variety of applications. In general, we are given a universe of objects to be "distinguished" by a family of "tests", and we want to find the smallest sufficient collection of tests. In the simplest version, a test is a subset of the universe and two objects are distinguished by our collection if one test contains exactly one of them. Variations allow tests to be multi-valued functions or unions of "basic" tests, and different notions of the term distinguished. An important version of this problem that has applications in DNA sequence analysis has the universe consisting of strings over a small alphabet and tests that are detecting presence (or absence) of a substring. For most versions of the problem, including the latter, we establish matching lower and upper bounds on approximation ratio. When tests can be formed as unions of basic tests, we show that the problem is as hard as the graph coloring problem.

1 Introduction and Motivation

One of the test set problems was on the classic list of NP-complete problems given by Garey and Johnson [6]; these problems arise naturally in many other applications. Below we provide an *informal* description of the basic problem with its motivating applications in various settings; precise descriptions and definitions appear in Section 1.1. In every version of the test set problem, we are given a universe of objects, family of subsets (*tests*) of the universe and a notion of *distinguishability* of pairs of elements of the universe by a collection of these tests. Our goal is to select a subset of these tests of minimum size that distinguishes every pair of elements of the universe. This framework captures problems in several areas in bioinformatics and biological modeling.

Minimum Test Collection Problem: This problem has applications in diagnostic testing. Here a collection of tests distinguishes two objects if a test from the collection contains exactly one of them. Garey and Johnson [6, pp. 71] showed a proof of NP-hardness of this problem via a reduction from

the 3-dimensional matching problem. Moret and Shairo [12] discussed some heuristics and experimental results for this problem. Finally, very recently the authors in [2, 8] established a $(1-\varepsilon)\ln n$ lower bound for approximation for any polynomial-time algorithm under standard complexity-theoretic assumptions where n is the number of objects and $\varepsilon > 0$ is an arbitrary constant.

Condition Cover Problem: Karp *et al.* [10] considered a problem of verifying a multi-output feedforward Boolean circuit as a model of biological pathways. This problem can be phrased like the Minimum Test Collection Problem, except that two elements are distinguished by a collection of tests if one tests contains exactly one of them, and another contains both or none of them.

String Barcoding Problem: In this problem, discussed by Rash and Gusfield [13], the universe U consists of sequences (strings), and for every possible string v we can form a test T_v as a collection of strings from U in which v appears. The name "string barcoding" derives from the fact that the Boolean vector indicating the occurrence (as a substring) of the tests from an arbitrary collection of tests in a given input sequence is referred to as the "barcode" of the given sequence with respect to this collection of tests. Motivations for investigating these problems come from several sources such as: **(a)** database compression and fast database search for DNA sequences and **(b)** DNA microarray designs for efficient virus identification in which the immobilized DNA sequences at an array element are from a set of barcodes. In [13], Rash and Gusfield left open the exact complexity and approximability of String Barcoding. We also consider a version in which a test can be defined by a set T of strings, with some limit on the set size, and $u \in U$ passes test T if one of strings in T is a substring of u; such tests are as feasible in practice as the one-string tests.

Minimum Cost Probe Set Problem with a Threshold: This problem is very similar to String Barcoding and it was considered by Borneman *et al.* [3]. They used this in [3] for minimizing the number of oligonucleotide probes needed for analyzing populations of ribosomal RNA gene (rDNA) clones by hybridization experiments on DNA microarrays. Borneman *et al.* [3] noted that this problem was NP-complete assuming that the lengths of the sequences in the prespecified set were unrestricted, but no other nontrivial theoretical results are known.

1.1 Notation and Definitions

Each problem discussed in this paper is obtained by fixing parameters in our general *test set* problem $\mathrm{TS}^{\Gamma}(k)$. The following notation and terminology is used throughout this paper:

- $[i, j]$ denotes the set of integers $\{i, i+1, \ldots, j-1, j\}$.
- $\mathcal{P}(S) = \{A : A \subseteq S\}$ denotes the power set of S.
- $|X|$ denote the cardinality (resp. length) of X if X is a set (resp. sequence).

– For two sequences (strings) u and v over an alphabet Σ, v is a substring of x (denoted by $v \prec x$) if $x = uvw$ for some $u, w \in \Sigma^*$.
– For two sets of numbers A and B and a number a, $a \times A$ denotes the set $\{ai |\ i \in A\}$ and $A + B$ denotes the set $\{a + b |\ a \in A \,\&\, b \in B\}$.

Definition 1. (Problem $\mathrm{TS}^{\Gamma}(k)$ with parameters $\Gamma \subseteq \mathcal{P}([0, 2])$ and a positive integer k)

Instance: $(n, \mathcal{S})$ where $\mathcal{S} \subset \mathcal{P}([0, n-1])$.
Terminologies:
– A *k-test* is a union of at most k sets from $\mathcal{S}$.
– For a $\gamma \in \Gamma$ and two distinct elements $x, y \in [0, n-1]$, a k-test T *γ-distinguishes* x and y if $|\{x, y\} \cap T| \in \gamma$.
Valid solutions: A collection $\mathcal{T}$ of k-tests such that
$(\forall x, y \in [0, n-1]\ \ \forall \gamma \in \Gamma)\ \ x \neq y \implies \exists T \in \mathcal{T}$ such that T γ-distinguishes x and y.
Objective: *minimize* $|\mathcal{T}|$.

An example to illustrate Definition 1: *Let* $n = 3$, $k = 1$, $\Gamma = \{\ \{1\}\ \}$ *and* $\mathcal{S} = \{\ \{0\}, \{1\}, \{0, 1\}\ \}$. *Then,* $\mathcal{T} = \{\ \{0\}, \{0, 1\}\ \}$ *is a valid solution since the* 1*-test* $\{0, 1\}$ $\{1\}$*-distinguishes* 0 *from* 2 *as well as* 1 *from* 2 *while the* 1*-test* $\{0\}$ $\{1\}$*-distinguishes* 0 *from* 1.

Now we precisely state the relationship of the $\mathrm{TS}^{\Gamma}(k)$ problem to several other problems in bioinformatics and biological modeling that we discussed before:

Minimum Test Collection Problem (Garey and Johnson [6]): This is precisely $\mathrm{TS}^{\{1\}}(1)$.

Condition Cover Problem (Karp *et al.* [10]): Assuming that the allowed perturbations are given as part of the input, this problem is identical to $\mathrm{TS}^{\{1\},\{0,2\}}(1)$.

String Barcoding Problem: Define a *k-sequence* as a collection of at most k distinct sequences. In this problem, considered by Rash and Gusfield [13] for the case when $k = 1$, we are given a set $\mathcal{S}$ of sequences over some alphabet Σ. For a *fixed* set of m k-sequences $\mathbf{t} = (\mathbf{t}_0, \ldots, \mathbf{t}_{m-1})$, the *barcode* $code(s, \mathbf{t})$ for each $s \in S$ is defined to be the Boolean vector (c_0, c_1, c_{m-1}) where c_i is 1 iff there exists a $t \in \mathbf{t}_i$ such that $t \prec s$. We say that $\mathbf{t}$ defines a *valid* barcode if for any two distinct strings $s, s' \in \mathcal{S}$, $code(s, \mathbf{t})$ is different from $code(s', \mathbf{t})$. The string barcoding problem over alphabet Σ, denoted by $\mathrm{SB}^{\Sigma}(k)$, has a parameter $k \in \mathbb{N}$ and is defined as follows:

Instance: $(n, \mathcal{S})$ where $\mathcal{S} \subset \Sigma^*$ and $1 \leq k \leq n = |\mathcal{S}|$.
Valid solutions: a set of k-sequences $\mathbf{t}$ defining a valid barcode.
Objective: *minimize* $|\mathbf{t}|$.

$\mathrm{SB}^{\Sigma}(k)$ is a special case of $\mathrm{TS}^{\{1\}}(k)$ in which $U = \mathcal{S}$ and for each substring p of each sequence in S there is a test $\{s \in \mathcal{S}\ :\ p \prec s\}$; valid barcodes can be identified with valid sets of k-tests.

Minimum Cost Probe Set with Threshold r (Borneman *et al.* [3]):
This problem, denoted by $\mathrm{MCP}^{\Sigma}(r)$, is a variation of $\mathrm{TS}^{\{1\}}(1)$. Denote by $oc(x, y)$ the number of occurrences of x in y as a substring, For a fixed set of m sequences $\mathbf{t} = (t_0, t_1, \ldots, t_{m-1})$, an r-barcode $code(s, \mathbf{t})$ for any sequence s is defined to be the vector $(c_0, c_1, \ldots, c_{m-1})$ where $c_i = \min\{r, oc(t_i, s)\}$. Given a set $\mathcal{S}$ of sequences over some alphabet Σ, $\mathbf{t}$ defines a *valid* r-barcode if for any two distinct strings $s, s' \in \mathcal{S}$, $code(s, \mathbf{t})$ is different from $code(s', \mathbf{t})$. $\mathrm{MCP}^{\Sigma}(r)$ is now defined as follows:

Instance: $(n, r, \mathcal{S}, \mathcal{P})$ where $\mathcal{S}, \mathcal{P} \subset \Sigma^*$ and $|\mathcal{S}| = n$.
Valid solutions: a set of sequences $\mathbf{t} \in \mathcal{P}^*$ defining a valid r-barcode.
Objective: *minimize* $|\mathbf{t}|$.

If $\mathcal{P}$ is the set of all substrings of sequences in $\mathcal{S}$, $\mathrm{MCP}^{\Sigma}(1)$ is precisely $\mathrm{SB}^{\Sigma}(1)$. All our results on $\mathrm{SB}^{\Sigma}(1)$ apply to $\mathrm{MCP}^{\Sigma}(r)$ with appropriate modifications.

2 Summary of Our Results

We provide matching upper and lower bounds on approximation ratios of polynomial time algorithms for $\mathrm{TS}^{\{1\}}(1)$, $\mathrm{TS}^{\{1\},\{0,2\}}(1)$, $\mathrm{SB}^{\Sigma}(1)$ and $\mathrm{MCP}^{\Sigma}(r)$ and strong lower bounds on approximation ratios of polynomial time algorithms for $\mathrm{TS}^{\{1\}}(k)$, $\mathrm{TS}^{\{1\},\{0,2\}}(k)$ and $\mathrm{SB}^{\Sigma}(k)$ for large k; these results are summarized in Table 1.

Problem	Approximation Ratio				Theorem(s)
	Upper Bound		Lower Bound		
	Time	the bound	the bound	Assumptions	
$\mathrm{TS}^{\{1\}}(1)$	$O(n^2\|\mathcal{S}\|)$	$1 + \ln n$	$(1-\varepsilon)\ln n$	NP$\not\subset$DTIME($n^{\log\log n}$)	1 and 5
$\mathrm{TS}^{\{1\},\{0,2\}}(1)$	$O(n^2\|\mathcal{S}\|)$	$1 + \ln 2 + \ln n$	$(1-\varepsilon)\ln n$	NP$\not\subset$DTIME($n^{\log\log n}$)	1 and 5
$\mathrm{SB}^{\Sigma}(1)$	$O(n^3\ell^2)$	$1 + \ln n$	$(1-\varepsilon)\ln n$	NP$\not\subset$DTIME($n^{\log\log n}$) $\|\Sigma\| > 1$	1 and 5
$\mathrm{MCP}^{\Sigma}(r)$	$O(n^2\|\mathcal{P}\| + L\|\mathcal{P}\|)$	$[1 + o(1)]\ln n$	$(1-\varepsilon)\ln n$	NP$\not\subset$DTIME($n^{\log\log n}$) $\|\Sigma\| > 1$	1 and 5
$\mathrm{TS}^{\{1\}}(n^{\delta})$			n^{ε}	NP$\neq$co-RP $0 < \varepsilon < \delta < 1$	9
$\mathrm{SB}^{\Sigma}(n^{\delta})$			n^{ε}	NP$\neq$co-RP $0 < \varepsilon < \delta < \frac{1}{2}$	9

Table 1. Summary of our approximability results: $(n, \mathcal{S})$ is an input instance of $\mathrm{TS}^{\Gamma}(k)$ and $\mathrm{SB}^{\Sigma}(k)$, $(n, \mathcal{S}, \mathcal{P})$ is an input instance of $\mathrm{MCP}^{\Sigma}(r)$, ℓ is the maximum length of any sequence in $\mathcal{S}$, L is the total length of all sequences in $\mathcal{S}$ and ε and δ are constants. The column "Assumptions" contains *sufficient* condition(s) for the respective lower bound.

Techniques Used
(a) Our algorithm to achieve the tight approximation bound in Theorem 1 for $\mathrm{TS}^{\{1\}}(1)$, $\mathrm{TS}^{\{1\},\{0,2\}}(1)$ and $\mathrm{MCP}^{\Sigma}(r)$ is a greedy algorithm that selects tests based on *information content* defined in terms of the change in the partition of the universe when the test is applied. This notion is directly related to the Shannon information complexity [1, 14]. A careful analysis yields an upper bound on the approximation ratio that matches the lower bound in Theorem 5 within a *small additive term*. We believe the analysis will be useful in the context of analyzing other problems involving recursive partitioning of a given universe as well.

(b) The inapproximability results of Theorem 5 are proved by approximation preserving reductions from the set cover problem. To handle the barcode problem for $\Sigma = \{0, 1\}$ we introduce an artificial intermediate problem (the "test set with order" problem) in which some tests are provided almost for free but they help very little in constructing a good set of tests. This roughly corresponds to the fact that we cannot avoid tests that do not correspond to sets in the original set cover instance, but we can make them cheap.

(c) The inapproximability results in Theorem 9 are obtained by approximation preserving reductions from the graph coloring problem.

Conparison of our results with those in [2, 8]: The authors in [2, 8] proved a $(1-\varepsilon)\ln n$ lower bound for approximation for $\mathrm{TS}^{1}(1)$. In this paper, we prove a lower bound of $(1-\varepsilon)\ln n$ for $\mathrm{SB}^{\{0,1\}}$, an *extremely restricted special case* of $\mathrm{TS}^{1}(1)$ that is of utmost importance to the bioinformatics community in detecting unknown virus sequences and designing probes for DNA microarrays. The proof in [2, 8] from set-cover to $\mathrm{TS}^{1}(1)$ does not seem to be easily transformable to provide a lower bound for $\mathrm{SB}^{\{0,1\}}$ with a similar quality of non-approximability because of the special nature of $\mathrm{SB}^{\{0,1\}}$. We therefore needed to introduce an artificial intermediate problem (the "test set with order" problem, denoted by TSO^{k}) which we could then translate to $\mathrm{SB}^{0,1}$ in a non-trivial manner. It should be noted that, for general k, TSO^{k} is neither equivalent to or nor a special case of $\mathrm{TS}^{1}(1)$.

Notational simplifications: We will skip (1) in $\mathrm{TS}^{\{1\}}(1)$, $\mathrm{TS}^{\{1\},\{0,2\}}(1)$ and $\mathrm{SB}^{\{0,1\}}(1)$, write "{1}-distinguishes" simply as "distinguishes" or "separates", and 1-tests simply as tests. Also, unless otherwise stated, all "computations", "transformations" or "reductions" take polynomial time.

The Map. Proofs of some of the claims in Theorems 1, 5 and 9 appear in Sections 3 , 4 ,and 5, respectively.

3 Approximation Algorithms for Test Set and Minimum Cost Probe Problems

The Set Cover (SC) Problem is defined on an input instance $(U, \mathcal{S})$ such that $\mathcal{S} \subset \mathcal{P}(U)$ with the goal of finding a $\mathcal{C} \subseteq \mathcal{S}$ such that $\bigcup_{A\in\mathcal{C}} A = U$ and $|\mathcal{C}|$ is

minimized. We can translate the $\mathrm{TS}^{\{1\}}$ problem to SC as follows. Given instance $(n, \mathcal{S})$ of $\mathrm{TS}^{\{1\}}$, we define instance $(U, \tau(\mathcal{S}))$ where $U = \{e \subset [0, n-1] : |e| = 2\}$, $\tau(T) = \{e \in U : |e \cap T| = 1\}$, and $\tau(\mathcal{S}') = \{\tau(T) : T \in \mathcal{S}\}$. The best proven approximation ratio for SC is achieved by a greedy heuristic [9] that, starting from the empty partial set cover, keeps adding new sets to the solution that maximize the number of elements that are not covered as yet. This heuristic for set cover runs in $O(\sum_{T\in\mathcal{S}} |\tau(T)|)$ time and has an approximation ratio of $1 + \ln(\max_{T\in\mathcal{S}} |\tau(T)|)$. Since $\max_{T\in\mathcal{S}} |\tau(T)| = |T|\ (n - |T|) \leq \frac{n^2}{4}$, the above translation offers a $O(n^2|\mathcal{S}|)$ time greedy heuristic for $\mathrm{TS}^{\{1\}}$ with an approximation ratio of $(2\ln n) - \ln 4$. A similar reduction for the $\mathrm{TS}^{\{1\},\{0,2\}}$ (resp. $\mathrm{MCP}^{\Sigma}(r)$) to the SC problem can also be given providing a greedy heuristic with an approximation ratio of $(2\ln n) - \ln\frac{4}{3}$ (resp. $2\ln n$). The main result of this section improves upon that simple heuristic as follows.

Theorem 1 *There is an $O(n^2|\mathcal{S}|)$ time approximation algorithm for TS^{Γ} with approximation ratio $1+\ln n$ for $\Gamma = \{\{1\}\}$ and $1+\ln 2+\ln n$ for $\Gamma = \{\{1\},]\{0,2\}\}$. There is an $O(n^2|P| + L|P|)$ time approximation algorithm for $MCP^{\Sigma}(r)$ with approximation ratio $1 + \ln n + \ln\log_2(r'+1)$, where $r' = \min\{r, n\}$ and L is the total length of the sequences in $\mathcal{S}$.*

3.1 Proof of Theorem 1 for $\mathrm{TS}^{\{1\}}$

In this section we provide a greedy heuristic for $\mathrm{TS}^{\{1\}}$ running in time $O(n^2|\mathcal{S}|)$ time with an improved approximation ratio of $1 + \ln n$. Notice that the upper bound almost matches the lower bound in Theorem 5 for $\mathrm{SB}^{\{0,1\}}$, a special case of $\mathrm{TS}^{\{1\}}$.

First, we consider the problem $\mathrm{TS}^{\{1\}}$. In the definition below and throughout the rest of this section we use $\mathcal{T} + T$ to denote $\mathcal{T} \cup \{T\}$.

Definition 2. *A set of tests $\mathcal{T} \subset \mathcal{S}$ defines the following:*

- *an equivalence relation $\overset{\mathcal{T}}{\equiv}$ on $[0, n-1]$ given by $i \overset{\mathcal{T}}{\equiv} j$ if and only if $\forall T \in \mathcal{T}(i \in T \equiv j \in T)$,*
- *a set of permutations $\Pi_{\mathcal{T}} = \{\pi \in (\text{permutations of } [0, n-1]) : \forall i \in [0, n-1]\ i \overset{\mathcal{T}}{\equiv} \pi(i)\}$,*
- *entropy $H_{\mathcal{T}} = \log_2 |\Pi_{\mathcal{T}}|$.*
- *information content of a $T \in \mathcal{S}$ with respect to $\mathcal{T}$, $IC(T, \mathcal{T}) = H_{\mathcal{T}} - H_{\mathcal{T}+T} = \log_2 \frac{|\Pi_{\mathcal{T}}|}{|\Pi_{\mathcal{T}+T}|}$.*

Our definition of entropy is very similar to the one suggested in [12]. Suppose that the equivalence relation $\overset{\mathcal{T}}{\equiv}$ on $[0, n-1]$ produces q equivalence classes of size $s_1, s_2, \ldots, s_q$. Then, the entropy suggested in [12] is $\frac{1}{n}\log_2(\Pi_{i=1}^{q} s_i^{s_i})$ whereas our entropy $H_{\mathcal{T}}$ is $\log_2(\Pi_{i=1}^{q} s_i!)$.

The *information content heuristic* (ICH for short) is the following simple greedy heuristic:

$\mathcal{T} = \emptyset$
while $H_{\mathcal{T}} \neq 0$ **do**
    select a $T \in \mathcal{S} - \mathcal{T}$ that *maximizes* $IC(T, \mathcal{T})$
    $\mathcal{T} = \mathcal{T} + T$
endwhile

The correctness of ICH follows from the fact that $H_{\mathcal{T}} = 0$ implies the equivalence classes of $\overset{\mathcal{T}}{\equiv}$ are n singleton sets $\{0\}, \{1\}, \ldots, \{n-1\}$ and the fact that if $H_{\mathcal{T}} \neq 0$ then there exists a $T \in \mathcal{S} - \mathcal{T}$ with $IC(T, \mathcal{T}) > 0$ (otherwise our problem instance has no feasible solution). It is also not very difficult to implement this algorithm efficiently within our claimed time bounds.

To implement ICH, we iteratively maintain the equivalence classes of $\overset{\mathcal{T}}{\equiv}$ as sorted lists. We also precompute and store $\log_2(i!)$ for each $i \in [1, n]$. Given a specific $T \in \mathcal{S} - \mathcal{T}$, it is easy to compute in $O(n)$ time the equivalence classes of $\overset{\mathcal{T}+T}{\equiv}$ from the equivalence classes of $\overset{\mathcal{T}}{\equiv}$ since an equivalence class E of $\overset{\mathcal{T}}{\equiv}$ is either an equivalence class of $\overset{\mathcal{T}+T}{\equiv}$ or it is partitioned into two equivalence classes $E_1 = E \cap T$ and $E_2 = E - E_1$ of $\overset{\mathcal{T}+T}{\equiv}$; the first case contributes nothing to $IC(T, \mathcal{T})$ while the second case adds $\log_2 \binom{|E|}{|E_1|}$ to $IC(T, \mathcal{T})$. Finally, notice that the while loop is executed at most n times.

Now we analyze the approximation ratio of ICH. We will use the convention $x = |X|$ for a set X.

Lemma 2 *If $\mathcal{T}_0 \subset \mathcal{T}_1$ then $IC(T, \mathcal{T}_0) \geq IC(T, \mathcal{T}_1)$.*

Lemma 3 *$IC(T, \emptyset) < n$ for every test T.*

Lemma 4 *If $IC(T, \mathcal{T}) > 0$ then $IC(T, \mathcal{T}) \geq 1$.*

Now we are ready for an amortized analysis of ICH. Suppose that an optimum solution of $(n, \mathcal{S})$ is $\mathcal{T}^* = \{T_1^*, \ldots, T_k^*\}$. During the execution of ICH, for a current partial test set $\mathcal{T}$, let $\mathcal{T}_i = \mathcal{T} + T_1^* + \cdots + T_i^*$ (accordingly, $\mathcal{T}_0 = \mathcal{T}$) and $h_i = IC(\mathcal{T}_{i-1}, T_i^*)$. Notice that $\sum_{i=1}^{k} h_i = \sum_{i=1}^{k} (H_{\mathcal{T}_{i-1}} - H_{\mathcal{T}_{i-1}+T_i^*}) = H_{\mathcal{T}} - H_{\mathcal{T}+\mathcal{T}^*} = H_{\mathcal{T}}$, since $H_{\mathcal{T}+\mathcal{T}^*} = 0$. Let $h_i^* < n$ denote the initial value of h_i *i.e.* the value of h_i with $\mathcal{T} = \emptyset$.

During the j^{th} iteration of the while loop, ICH selects a test T (with, say, $IC(T, \mathcal{T}) = \Delta_j$) and changes $\mathcal{T}$ into $\mathcal{T} + T$. As a result, $H_{\mathcal{T}}$ drops by Δ_j and h_i drops by some $\delta_{i,j}$ with $\sum_{i=1}^{k} \delta_{i,j} = \Delta_j$. This iteration adds 1 to the solution cost. We distribute this cost among the elements of $\mathcal{T}^*$ by charging T_i^* with $\delta_{i,j}/\Delta_j$. Because $h_i = IC(\mathcal{T}_{i-1}, T_i^*) \leq IC(\mathcal{T}, T_i^*)$, we know that $\Delta_j \geq h_i$ since otherwise ICH would select T_i^* rather then T. Therefore reducing the current h_i by $\delta_{i,j}$ is associated with a charge that is at most $\delta_{i,j}/h_i$. Let $m(h)$ be the supremum of possible sums of charges that some T_i^* may receive starting from the time when $h_i = h$. By induction on the number of such positive charges we will show that $m(h) \leq 1 + \ln h$. If this number is 1, then $h > 0$ and hence $\ln h \geq 0$ (by Lemma 4), while the charge is at most 1. In the inductive step, we

consider a situation when T_i^* starts with $h_i = h$, receives a single charge δ/h, h_i is reduced to $h - \delta$ and afterwards, by inductive assumption, T_i^* receives at most $m(h-\delta)$ charges. Because $h - \delta > 0$ we know by Lemma 4 that $h - \delta \geq 1$. Therefore

$$m(h) \leq m(h-\delta) + \frac{\delta}{h} \leq 1 + \ln(h-\delta) + \frac{\delta}{h} < 1 + \int_1^{h-\delta} \frac{dx}{x} + \int_{h-\delta}^{h} \frac{dx}{x} = 1 + \ln h.$$

By Lemma 3, $h < n$. This proves our claim on the approximation ratio for $\mathrm{TS}^{\{1\}}$.

4 Inapproximability Results for Test Set, String Barcoding and Minimum Cost Probe Set Problems

The NP-hardness of $\mathrm{TS}^{\{1\}}$ follows from the NP-hardness of the minimum test collection problem in [6] from a reduction from the 3-dimensional matching problem and minor modifications of this reduction can be used to prove the NP-hardness of $\mathrm{TS}^{\{1\},\{0,2\}}$ as well. NP-hardness of $\mathrm{MCP}^{\Sigma}(r)$ from the vertex cover problem was mentioned without a proof in [3]. Our goal is to show that it is impossible (under reasonable complexity theoretic assumptions) to approximate these problems any better than mentioned in Theorem 1.

Theorem 5 *For any given constant $0 < \rho < 1$, it is impossible to approximate $SB^{\{0,1\}}$ (a restricted case of $TS^{\{1\}}$), $TS^{\{1\},\{0,2\}}$ or $MCP^{\{0,1\}}(r)$ within a factor of $(1-\rho)\ln n$ in polynomial time unless NP$\subset$DTIME$(n^{\log\log n})$.*

Our proof of Theorem 5 proceed in two stages:

- In Section 4.1 we introduce the Test Set with Order (TSO) problem and provide a reduction from the set cover problem to the TSO problem preserving apprpximation.
- Our complete reduction from the set cover problem to $\mathrm{SB}^{\{0,1\}}$, described in Section 4.2, uses a composition of the abovementioned reduction and another approximation-preserving reduction from the TSO problem to $\mathrm{SB}^{\{0,1\}}$.

4.1 Test Set with Order

To make the approximation preserving reduction from set cover to $\mathrm{SB}^{\{0,1\}}$ easier to follow, we introduce an intermediate problem called *Test Set with Order* with parameter $k \in \mathbb{N}$ (denoted by TSO^k):

Instance: $(n, k, \mathcal{S})$ where k is a positive integer, $(n, \mathcal{S})$ is an instance of $\mathrm{TS}^{\{1\}}$ and $\mathcal{S}$ includes the family of "cheap" sets $\mathcal{S}_0 = \{\{i\} |\ i \in [0, n-1]\} \cup \{[0, i] \mid i \in [0, n-1]\}$.
Valid solutions: a solution for the instance $(n, \mathcal{S})$ of $\mathrm{TS}^{\{1\}}$.
Objective: minimize $cost(\mathcal{T}) = |\mathcal{T} - \mathcal{S}_0| + \frac{1}{k}|\mathcal{T} \cap \mathcal{S}_0|$.

Note that TSO^1 is in fact a *special case* of $\mathrm{TS}^{\{1\}}$; hence any hardness results proved for TSO^1 would apply to $\mathrm{TS}^{\{1\}}$ as well. Our claim follows once the following theorem is proved.

Theorem 1. *For any integer constant $k > 0$ and any constant $0 < \rho < 1$, it is impossible to approximate TSO^k within a factor of $(1-\rho)\ln n$ in polynomial time unless $NP \subset DTIME(n^{\log\log n})$.*

In the rest of this section, we prove the above theorem. We need the following straightforward extension of the hardness result in [4] for a slightly restricted version of SC.

Fact 6 *Assuming $NP \not\subset DTIME(n^{\log\log n})$, instances of the SC problem for which the optimal cover requires at least $(\log_2 n)^2$ sets cannot be approximated to within a factor of $(1-\varepsilon')\ln n$ for any constant $\varepsilon' > 0$ in polynomial time.*

For notational simplicity, assume that kn is an exact power of 2 and $\ell = \log_2(kn)$. The following lemma gives a reduction from SC to TSO^k problem.

Lemma 7 *There exists a polynomial-time computable function τ that maps an instance $(n, \mathcal{S})$ of SC into instance $(2kn, k, \tau(\mathcal{S}))$ of TSO^k such that optimal solutions of $(n, \mathcal{S})$ and $(2kn, k, \tau(\mathcal{S}))$, C^* and $\mathcal{T}^*$ respectively, satisfy the following:*

$$|\mathcal{C}^*| \le cost(\mathcal{T}^*) \le |\mathcal{C}^*| + \ell + 1.$$

Moreover, given any solution X of $(2kn, k, \tau(\mathcal{S}))$, we can in polynomial time construct a solution Y of $(n, \mathcal{S})$ such that $|Y| \le cost(X)$.

Proof. $\tau(\mathcal{S})$ contains the following sets:

$$\begin{array}{l}\text{cover sets: } D(S) = 2 \times (k \times S + [0, k-1]) \text{ for } S \in \mathcal{S};\\ \text{cheap sets: } \{i\} \text{ and } [0, i] \text{ for each } i \in [0, 2kn-1];\\ \text{other sets: } A_i = \{j \in [0, 2kn-1] |\ j \bmod 2^{i+1} \ge 2^i\} \text{ for } i \in [1, \ell].\end{array}$$

First, we show that $cost(\mathcal{T}^*) \le |\mathcal{C}^*| + \ell$. Given a set cover $\mathcal{C}$ of $(n, \mathcal{S})$ we define the following test set that is a solution of $(2kn, \tau(\mathcal{S}))$: $\mathcal{T} = \{D(A) |\ A \in \mathcal{C}\} \cup \{A_i |\ i \in [1, \ell]\}$. To see that $\mathcal{T}$ is indeed a valid solution, consider $i, j \in [0, 2kn-1]$. Suppose that i is even and j is not. Then for some $A \in \mathcal{C}$ and $a \in 2 \times [0, k-1]$ we have $(i-2a)/2k \in A$, and thus $i \in D(A)$ while $j \notin D(A)$. On the other hand, if that i and j have the same parity then they differ on k^{th} bit for some $k \in [1, \ell]$, in which case i and j are distinguished by test A_k. Hence, $cost(\mathcal{T}^*) = |\mathcal{T}^*| \le |\mathcal{C}^*| + \ell$.

Next, we show that $|\mathcal{C}^*| \le cost(\mathcal{T}^*)$. Given a set of tests $\mathcal{T}$, consider the *partial cover* $\mathcal{C}' = \{A |\ D(A) \in \mathcal{T}\}$, and let $C = \bigcup_{S \in \mathcal{C}'} S$. Consider $i \in [0, n-1] - C$. For $a \in [0, k-1]$ we know that some set of $\mathcal{T}$ distinguished $2ki - 2a$ from $2ki - 2a + 1$. This distinguishing set can only be one of the three sets: $\{2ki - 2a\}$, $\{2ki - 2a + 1\}$ or $[0, 2ki - 2a]$. Note that for each $i \in [0, n-1] - C$ and each $a \in [0, k-1]$ we have a choice of different three sets, so in each such case we use a different element of $\mathcal{T}$. We can conclude that $\mathcal{T}$ contains $k(n - |C|)$ such sets, and thus $cost(\mathcal{T}) \ge |\mathcal{C}'| + n - |C|$. Since for each $i \in [0, n-1]$ $\mathcal{T}$ must distinguish

$2i-1$ from $2i$, $\mathcal{T}$ must contain one of these three sets: $\{2i-1\}, \{2i\}, [0, 2i-1]$. Note that each $i \in [0, n-1] - C$ has different possibilities, thus for each of them $\mathcal{T}$ contains a different set of choices. We can therefore extend $\mathcal{C}'$ to a cover $\mathcal{C}$ of $(n, \mathcal{S})$ by adding at most $n - |C|$ sets. Hence $|\mathcal{C}| \leq cost(\mathcal{T})$.

Hence, $cost(\mathcal{T}^*) \leq |\mathcal{C}^*| + \ell + \frac{1}{k}$.

We can now complete the proof of Theorem 1. Consider an instance of SC as mentioned in Fact 6, transform it to an instance of TSO^k as described in Lemma 7 and let $\mathcal{C}^*$ and $\mathcal{T}^*$ be optimal solutions to the instances of SC and TSO^k, respectively. Suppose that we can approximate TSO^k within a factor of $(1-\rho)\ln n$ and let T' be such an approximate solution. Then, by using Lemma 7 we can find a solution C' to the instance of SC such that

$$\begin{aligned} |C'| &\leq cost(T') \\ &\leq (1-\rho)\ln n\, cost(\mathcal{T}^*) \\ &\leq (1-\rho)\ln n\,(|\mathcal{C}^*| + \ell + 1) \\ &\leq (1-\rho+o(1))\ln n\,|\mathcal{C}^*| \quad \text{since } |\mathcal{C}^*| = \Omega(\ell^2) \text{ and } \ell = \Omega(\log n) \end{aligned}$$

which violates Fact 6 by choosing $\varepsilon' = 1 - \rho + o(1)$.

4.2 Proof of Theorem 5 for $\text{SB}^{\{0,1\}}$

As before, for notational simplicity, assume that kn is an exact power of 2 and $\ell = \log_2(kn)$. First, using the reduction described in the proof of Lemma 7, we provide a reduction of SC to $\text{SB}^{\{0,1\}}$.

Lemma 8 *For any given constant integer $k > 0$, there exists a polynomial-time computable function σ that maps an instance $(n, \mathcal{S})$ of SC into an instance $(2kn, \sigma(\mathcal{S}))$ of $SB^{\{0,1\}}$, so that if $\mathcal{C}^*$ and $\mathbf{t}^*$ are the optimal solutions for $(n, \mathcal{S})$ and $(2kn, \sigma(\mathcal{S}))$, respectively, then*

$$\frac{|\mathcal{C}^*|}{1+\frac{1}{k}} \leq |\mathbf{t}^*| \leq |\mathcal{C}^*| + \ell.$$

Moreover, given any solution $\mathbf{x}$ of $(2kn, \sigma(\mathcal{S}))$, we can in polynomial time construct a solution Y of $(n, \mathcal{S})$ such that $\frac{|Y|}{1+\frac{1}{k}} \leq |\mathbf{x}|$.

Proof. First, we define a family $\tau(\mathcal{S})$ of subsets of $[0, 2kn-1]$ using the function τ from Lemma 7. Let $\mathcal{S}_0$ be the family of "special" or "cheap" test sets, and $\mathcal{S}_1 = \tau(\mathcal{S}) - \mathcal{S}_0$. We number the elements of $\mathcal{S}_1$, so $\mathcal{S}_1 = \{B_0, \ldots, B_{m-1}\}$ and let $B_m = [0, 2kn-1] \in \mathcal{S}_0$. For each $i \in [0, 2kn-1]$ we define sequence s_i as a concatenation of alternating groups of 0^{i+1} and a distinct member from the set $\{1^{k+1} \mid i \in B_k\}$, begining and ending with 0^{i+1}. This completes the description of the function σ.

Consider any set cover $\mathcal{C}$ of $(n, \mathcal{S})$. As noted in the proof of Lemma 7, we can map it into a solution for TSO^k without using any cheap tests and with at most

$|\mathcal{C}^*| + \ell$ test sets. Then, we replace test B_j with a test sequence $01^{j+1}0$. Thus $|\mathbf{t}^*| \leq |\mathcal{C}^*| + \ell$.

Now consider a solution vector of sequences $\mathbf{t}$ for $\sigma(\mathcal{S})$. We show how to replace each sequence t of $\mathbf{t}$ with at most two sets such that the following two statements hold:

(a) if $(t \prec s_p) \neq (t \prec s_q)$ for two sequences s_p and s_q, then the replaced sets $\{1\}$-distinguish p from q;
(b) when we use two sets, one of them is cheap.

By **(a)**, the replacement sets form a solution for the instance $(2kn, k, \tau(\mathcal{S}))$ of TSO^k. By **(b)**, the cost of the this solution for $(2kn, k, \tau(\mathcal{S}))$ is at most $\left(1 + \frac{1}{k}\right)|\mathbf{t}|$. Finally, by Lemma 7, it is possible to construct from this solution for $(2kn, k, \tau(\mathcal{S}))$ a solution for the set cover instance $(n, \mathcal{S})$ with no more than $\left(1 + \frac{1}{k}\right)|\mathbf{t}|$ sets. Hence, it only remains to show the replacement. We have the following cases:

Case 1: t contains a substring 10^a1 for some $a > 0$. Then t can be a substring of only s_{a-1}, so we can replace t with a cheap test $\{a - 1\}$.
Case 2: Otherwise, t is of the form $0^*1^*0^*$.
- **Case 2.1:** $t = 0^a$ for some $a > 0$. Then t is a substring of all s_i's with $i \geq a - 1$, and therefore we can replace it with a cheap test $[0, i - 2]$.
- **Case 2.2:** $t = 0^a1^b$ for some $a, b > 0$. If $b > m + 1$, t is not a substring of any s_i, so we can discard it. If $b \leq m + 1$, then this test is equivalent to 0^a because every s_i contains 1^{m+1}.
- **Case 2.3:** $t = 1^a0^b$ for some $a, b > 0$. Similar to Case 2.2.
- **Case 2.4:** $t = 0^a1^b0^c$ where $a, b, c > 0$. Let $d = \max\{a, c\}$; one can see that we can replace t with B_{b-1} and $[0, d - 2]$.

We can now complete the proof of our claim in a manner similar to that in the proof of Lemma 7. Consider an instance of SC as mentioned in Fact 6, transform it to an instance of $\text{SB}^{\{0,1\}}$ as described in Lemma 8 and let $\mathcal{C}^*$ and $\mathbf{t}^*$ be optimal solutions to the instances of SC and $\text{SB}^{\{0,1\}}$, respectively. Suppose that we can approximate $\text{SB}^{\{0,1\}}$ within a factor of $(1-\rho)\ln n$ and let $\mathbf{t}'$ be such an approximate solution. Then, by using Lemma 8 we can find a solution C' to the instance of SC such that

$$\begin{aligned}
|C'| &\leq \left(1 + \tfrac{1}{k}\right) cost(\mathbf{t}') \\
&\leq \left(1 + \tfrac{1}{k}\right)(1 - \rho)\ln n\, cost(\mathbf{t}^*) \\
&\leq \left(1 + \tfrac{1}{k}\right)(1 - \rho)\ln n\,(|\mathcal{C}^*| + \ell + 1) \\
&\leq (1 - \rho + o(1))\ln n\,|\mathcal{C}^*| \qquad \text{since } |\mathcal{C}^*| = \Omega(\ell^2) \text{ and } \ell = \Omega(\log n)
\end{aligned}$$

which violates Fact 6 by choosing $\varepsilon' = 1 - \rho + o(1)$.

5 Stronger Inapproximabilities for $\text{TS}^{\{1\}}(k)$, $\text{TS}^{\{1\},\{0,2\}}(k)$ and $\text{SB}^{\{0,1\}}(k)$

Theorem 9
(a) *For any two given constants* $0 < \rho < \delta < 1$, $TS^{\{1\}}(n^\delta)$ *and* $TS^{\{1\},\{0,2\}}(n^\delta)$

cannot be approximated to within a factor of n^{ρ} in polynomial time unless co-RP=NP.
(b) *The result in* **(a)** *also holds for $SB^{\Sigma}(n^{\delta})$ if $0 < \rho < \delta < \frac{1}{2}$.*

Acknowledgments. We would like to thank the anonymous reviewers for pointing out references [2, 8, 12] and for extremely useful comments. Berman's research was supported by NSF grant CCR-O208821, DasGupta's research was supported in part by NSF grants CCR-0296041, CCR-0206795 and CCR-0208749, and Kao's research was supported in part by NSF grant EIA-0112934.

References

1. Y. S. Abu-Mostafa (editor). *Complexity in Information Theory*, Springer Verlag, 1986.
2. K. M. J. De Bontridder, B. V. Halldórsson, M. M. Halldórsson, C. A. J. Hurkens, J. K. Lenstra, R. Ravi and L. Stougie. *Approximation algorithms for the test cover problem*, Mathematical Programming-B, Vol. 98, No. 1-3, 2003, pp. 477-491.
3. J. Borneman, M. Chrobak, G. D. Vedova, A. Figueroa and T. Jiang. *Probe Selection Algorithms with Applications in the Analysis of Microbial Communities*, Bioinformatics, Vol. 17 Suppl. 1, 2001, pp. S39-S48.
4. U. Feige. *A threshold for approximating set cover*, JACM, Vol. 45, 1998, pp. 634-652.
5. U. Feige and J. Kilian. *Zero knowledge and the chromatic number*, Journal of Computer and System Sciences, Vol. 57, No. 2, October 1998, pp. 187-199.
6. M. R. Garey and D. S. Johnson. *Computers and Intractability - A Guide to the Theory of NP-Completeness*, W. H. Freeman & Co., 1979.
7. D. Gusfield. *Algorithms on Strings, Trees and Sequences*, Cambridge University Press, 1997.
8. B. V. Halldórsson, M. M. Halldórsson and R. Ravi. *On the Approximability of the Minimum Test Collection Problem*, Proc. Ninth Annual European Symposium on Algorithms, Lecture Notes in Computer Science 2161, pp. 158-169, 2001.
9. D. S. Johnson. *Approximation Algorithms for Combinatorial Problems*, Journal of Computer and Systems Sciences, Vol. 9, 1974, pp. 256-278.
10. R. M. Karp, R. Stoughton and K. Y. Yeung, *Algorithms for Choosing Differential Gene Expression Experiments*, Proc. Third Annual International Conference on Computational Molecular Biology, 1999, pp. 208-217.
11. L. Lovasz. *On the Ratio of Optimal Integral and Fractional Covers*, Discrete Mathematics, Vol. 13, 1975, pp. 383-390.
12. B. M. E. Moret and H. D. Shapiro. *On minimizing a set of tests*, SIAM Journal on Scientific and Statistical Computing, Vol. 6, 1985, pp. 983-1003.
13. S. Rash and D. Gusfield. *String Barcoding: Uncovering Optimal Virus Signatures*, Proc. Sixth Annual International Conference on Computational Molecular Biology, 2002, pp. 254-261.
14. C. E. Shannon. *Mathematical Theory of Communication*, Bell Systems Technical Journal, Vol. 27, 1948, pp. 379-423, pp. 623-658.