Electronic Colloquium on Computational Complexity, Report No. 145 (2009)

The Complexity of Rationalizing Network Formation*

Shankar Kalyanaraman
Computer Science Department
California Institute of Technology
Pasadena, CA 91125
shankar@cs.caltech.edu

Christopher Umans
Computer Science Department
California Institute of Technology
Pasadena, CA 91125
umans@cs.caltech.edu

December 19, 2009

Abstract

We study the complexity of *rationalizing* network formation. In this problem we fix an underlying model describing how selfish parties (the vertices) produce a graph by making individual decisions to form or not form incident edges. The model is equipped with a notion of stability (or equilibrium), and we observe a set of "snapshots" of graphs that are assumed to be stable. From this we would like to infer some unobserved data about the system: edge prices, or how much each vertex values short paths to each other vertex.

We study two rationalization problems arising from the network formation model of Jackson and Wolinsky [JW96]. When the goal is to infer edge prices, we observe that the rationalization problem is easy. The problem remains easy even when rationalizing prices do not exist and we instead wish to find prices that maximize the stability of the system.

In contrast, when the edge prices are given and the goal is instead to infer valuations of each vertex by each other vertex, we prove that the rationalization problem becomes NP-hard. Our proof exposes a close connection between rationalization problems and the Inequality-SAT (I-SAT) problem.

Finally and most significantly, we prove that an approximation version of this NP-complete rationalization problem is NP-hard to approximate to within better than a 1/2 ratio. This shows that the trivial algorithm of setting everyone's valuations to infinity (which rationalizes all the edges present in the input graphs) or to zero (which rationalizes all the non-edges present in the input graphs) is the best possible assuming P $\neq$ NP. To do this we prove a tight $(1/2+\delta)$-approximation hardness for a variant of I-SAT in which all coefficients are non-negative. This in turn follows from a tight hardness result for MAX-LIN$_{\mathbb{R}_+}$ (linear equations over the reals, with non-negative coefficients), which we prove by a (non-trivial) modification of the recent result of Guruswami and Raghavendra [GR07] which achieved tight hardness for this problem without the non-negativity constraint.

Our technical contributions regarding the hardness of I-SAT and MAX-LIN$_{\mathbb{R}_+}$ may be of independent interest, given the generality of these problems.

1 Introduction

In any market setting where different goods are available at different prices, consumers demonstrate a *revealed preference* for a particular bundle by choosing to buy it. They do so in spite of the existence of other bundles of goods that may be more affordable. *Revealed preference theory* in economics is a well-studied

*Supported by NSF CCF-0346991, CCF-0830787, BSF 2004329, and a Sloan Research Fellowship.

ISSN 1433-8092

area [Sam48, Afr67, Die73, Var82] that uses these observed choices to infer preference profiles, and, more subtly, to argue about the limitations of such inferences by identifying settings in which simple classes of preference profiles (e.g. *linear* utility functions) can *always* rationalize data satisfying basic axioms — the conclusion being that experiments of this sort cannot be expected to provide evidence that a more complicated (e.g. non-linear) preference profile is actually in use.

From a computer science perspective, a natural question is: how hard is it, computationally, to infer unobserved quantities (e.g. preferences profiles, utility functions, prices) that explain or *rationalize* the observed market data? Such problems abound in systems where there is an underlying notion of stability or equilibrium (and one assumes that the observed data represent stable instances).

In certain classical settings [Afr67], rationalization amounts to solving a linear program, and so it is immediately seen to be easy. Other settings have a more combinatorial feel, such as rationalizing matchings, which we studied in previous work [KU08]. In this problem the input is a collection of bipartite matchings, each meant to represent the observed outcome of a two-sided market. We showed in [KU08] that determining the existence of preference orders under which the matchings are all *stable matchings* is NP-complete, and NP-hard to approximate to within a constant factor.

In this work, we consider the rationalization problem for network formation games. We study two variants of a prominent model due to Jackson and Wolinsky [JW96] that describes how selfish parties (the vertices) produce a graph by making individual decisions to form or not form incident edges. The model is equipped with a notion of stability (or equilibrium), and we observe a set of "snapshots" of graphs that are assumed to be stable. From this we would like to infer some unobserved data about the system: in one variant we are interested in edge prices; in the other, we are interested in how much each vertex values short paths to each other vertex. Both variants resemble the settings in which the rationalization problem can be solved using linear programming (in the sense that the equilibrium conditions can be expressed as linear inequalities), and yet they have a combinatorial component because the participants' total utility depends on the length of various shortest paths in the network.

In this paper we show an interesting contrast: inferring "per-edge" quantities (i.e., prices) is easy, while inferring "end-to-end" quantities (i.e., the value each vertex u assigns to having a short path to each other vertex v) is hard. In the latter case we show a tight $(1/2 + \delta)$ inapproximability result (and this is our most technically significant contribution). The $1/2$ ratio implies that the trivial approximation algorithm that sets everyone's valuations to infinity (which rationalizes all the edges present in the input graphs) or to zero (which rationalizes all the non-edges present in the input graphs) is the best possible assuming P $\neq$ NP.

1.1 The network formation model and two rationalization problems

The network formation model we study is one proposed by Jackson and Wolinsky [JW96]. In this model, there are n vertices, and each pair (u, v) ("potential edge") has an associated price and a distance. A network formation process produces a graph G, and given this outcome, the utility that accrues to each vertex v depends on two additional features of the model: (1) a non-increasing function f from distances to the non-negative reals (think of $f(d)$ as representing the value of having a connection of length d), and (2) "intrinsic values" of vertex u by v for each $u \neq v$. The utility realized by vertex v is then the *aggregate distance* minus the price of the edges in v's subset, where the *aggregate distance* is the sum over vertices u of v's intrinsic value of u times f applied to the shortest path length in G to each u.

The equilibrium concept here is *not* a Nash equilibrium[1], but rather a simpler notion of pairwise stability;

[1]This is because we only consider unilateral deviations of a player to an *adjacent* strategy – one in which a single edge has been added or removed – instead of to *any* alternative strategy.

the vertex strategies are *stable* if (1) for each edge (u, v) in G, both v's and u's marginal utility of forming edge (u, v) is non-negative, and (2) for each non-edge (u, v) in G either u's or v's marginal utility of forming edge (u, v) is non-positive.

We consider two rationalization problems arising under this model. In the first, which we call STABLE-PRICES, we are trying to infer edge prices, and we assume the other data (distances, the function f, and the pairwise "intrinsic values") are fixed or given. Specifically, we are given a collection of distance-weighted graphs $G_1, G_2, \ldots, G_m$ on the same underlying vertex set, that arise from equilibrium play. In addition, we are given the function f (as a circuit computing it), and the pairwise intrinsic values (which are the same across the different graphs). We do not observe the (potential) edge prices (which are the same across the different graphs). We are interested in determining edge prices that *rationalize* $G_1, G_2, \ldots, G_m$; i.e., for which each G_i is stable, or "in equilibrium," in the above sense.

In the second rationalization problem under consideration, which we call STABLE-VALUES, we are trying to infer the pairwise "intrinsic values," and we assume the other data (latencies, the function f, and the edge prices) are given. Specifically, as above, we are given a collection of distance-weighted graphs $G_1, G_2, \ldots, G_m$ on the same underlying vertex set, that arise from equilibrium play. In addition, we are given the function f (as a circuit[2]), and the edge prices (which are the same across the different graphs); we do not observe the pairwise "intrinsic values" (which are the same across the different graphs). We are interested in determining pairwise intrinsic values that *rationalize* $G_1, G_2, \ldots, G_m$; i.e., for which each G_i is stable, or "in equilibrium", in the above sense.

We also consider an optimization version of STABLE-VALUES. In it, we are seeking pairwise "intrinsic values" that maximize the number of stable edges/non-edges across all m input graphs *among active pairs*. We deem a pair (u, v) *active* unless (1) it is an edge in all of the input graphs, with price zero (which means effectively that edge (u, v) is present and fixed no matter how the other quantities are varied) or (2) it is a non-edge in all of the input graphs, with price infinity (which means effectively that edge (u, v) is permanently absent regardless of the other relevant quantities). Non-active pairs are "part of the landscape" and intuitively do not contribute to the stability of the system. After this consideration, our optimization problem is to infer intrinsic values with the maximum explanatory power (and note that edges/non-edges are counted separately for each graph in which they appear).

For concreteness, we briefly describe an example scenario in which this rationalization problem naturally arises. Social networks are formed among groups of people who ascribe a certain value ("friendship") to one another but establish connections with only those that they perceive to be most intrinsically valuable to them. If, for instance, everybody in the group was in close physical proximity to one another (they all went to the same high school or college) then the price of connecting to any one person is insignificant compared to the value derived in return, no matter how small that may be. This would result in a clique as a stable network. However, once this group becomes geographically spread out, the network formed in equilibrium can become sparser, such as a star network, where all connections are made to a single person since the cost of building mutual connections outweighs the utility gained. This illustrates that (when holding the intrinsic value people in such a group have for one another to be invariant) temporal and spatial dynamics affect the manner of how social networks coalesce and stabilize. While prices and distances might be readily observable, the intrinsic value each individual has for each other individual is generally private. The problem STABLE-VALUES in this paper asks to infer these values given (say) a series of snapshots taken over time of a single social network of individuals.

[2]This permits unnatural functions f, but note that all of our reductions produce instances with very simple piecewise linear and non-increasing f that one can easily envision occurring in the real world.

1.2 Rationalization problems and I-SAT

As mentioned above, we show that STABLE-PRICES is easy, while STABLE-VALUES is hard. Our hardness result is based on a reduction from a variant of the Inequality Satisfiability problem (abbreviated as I-SAT) introduced recently by Hochbaum and Moreno-Centeno [HMC08]. An instance of I-SAT is a conjunction of *inequality-clauses*, where each inequality-clause is a disjunction of linear inequalities over n real variables $x_1, x_2, \ldots, x_n$. The instance is a "yes" instance iff there exists an assignment of real values to the variables simultaneously satisfying all of the inequality-clauses. Hochbaum and Moreno-Centeno showed by a simple reduction from 3-SAT that this class of problems is NP-complete even in the case when each inequality-clause is a disjunction of only two inequalities.

The variant of I-SAT that we need for our reduction satisfies two additional constraints: (1) all of the coefficients are non-negative (and we are seeking a solution only in the non-negative reals), and (2) there is a partition of the variables into two sets S, T such that every inequality-clause is either the disjunction of two $\leq$ inequalities, one supported in S and one supported in T, or a conjunction of two $\geq$ inequalities, one supported in S and one in T. We call this variant I-SAT*.

To achieve our main hardness results, we show that I-SAT* is NP-complete, and that the optimization version (maximize the number of inequality-clauses simultaneously satisfied) is NP-hard to approximate to within $(1/2 + \delta)$. Note that, just as it is trivial to achieve approximation ratio 1/2 in the rationalization problem to which we reduce, it is trivial to achieve approximation ratio 1/2 here by either setting all variables to zero (satisfying all the inequality-clauses of the first type) or setting all variables sufficiently large (satisfying all the inequality-clauses of the second type).

The ease of translating between these problems brings us to an important observation. Not only is I-SAT useful as a starting point for reductions involving the hard rationalization problem in this paper, but we contend it is the abstract computational problem that captures rationalization problems more generally. It is common for the "stability conditions" arising in a rationalization problem to be expressible by a finite Boolean formula whose inputs are inequalities in the (real) quantities being inferred. This is true, e.g., for the bipartite matchings problem studied in [KU08] (the quantities being inferred are the values each left node has for each right node, and the familiar stability condition for stable matchings is expressible as the disjunction of two inequalities involving these quantities), for the rationalization problems studied here, and those mentioned in the introduction. Even the positivity constraint we add arises naturally in many such settings, as utilities, prices, etc. are often assumed to be non-negative.

Thus we expect that a more complete understanding of the approximability of I-SAT (which to our knowledge has not been studied prior to this paper) can serve as a useful starting point for understanding the approximability of rationalization problems more generally, and we view this as an important contribution of this paper.

1.3 Hardness of approximating I-SAT* via MAX-LIN$_{\mathbb{R}_+}$

For the general I-SAT problem, there is an easy reduction from MAX-LIN$_{\mathbb{R}}$ (linear equations over the reals). Namely, for each equation $\sum_i a_i x_i = b$, we produce the pair of I-SAT clauses $\sum_i a_i x_i \leq b$ and $\sum_i a_i x_i \geq b$.

MAX-LIN$_{\mathbb{R}}$ was (only recently) shown to have a PCP system with $(1-\epsilon)$ completeness and γ soundness [GR07] (with ϵ, γ close to 0), which gives rise to $(1/2+\delta)$-inapproximability for the general I-SAT problem via this reduction (although, the non-perfect completeness means this gap is between classes of unsatisfiable instances, which is a minor drawback).

We need a similar hardness result for our variant, I-SAT*, which crucially entails a positivity constraint. In the [GR07] inapproximability result (and similar inapproximability results using the basic framework of

Hastad [Hås01]), the equations all have the form $x_i + x_j - x_k = 0$ since they arise from linearity tests performed by the verifier in the PCP system. Thus, they are not suitable for proving inapproximability for I-SAT*. Simple transformations like translating the origin do not work, and the natural idea of introducing new variables x'_i and the constraints $x_i + x'_i = 0$ (and using x'_i in place of $-x_i$ to remove the negative coefficients) does not preserve the inapproximability.

It is also important to note that while Hastad's inapproximability results for MAX-LIN$_{\mathbb{F}_p}$ can be easily transformed into similar inapproximability results for MAX-LIN$_{\mathbb{Z}}$, this transformation introduces large coefficients (of magnitude p), which prohibit the clever trick in [GR07] that is used to argue that the inapproximability carries over to the reals.

So our hands are somewhat tied: to obtain the $(1/2 + \delta)$ inapproximability for I-SAT*, we really need an exact analog of [GR07], but one that produces equations with positive coefficients. In Section 5, we give such a result for MAX-LIN$_{\mathbb{R}_+}$, showing that it is NP-hard to distinguish between an instance with a $(1 - \epsilon)$ fraction satisfiable assignment and one with at most δ fraction satisfiable, and in turn a $(1/2 + \delta)$ inapproximability result for I-SAT*. Doing so requires more than a superficial modification of the proof in [GR07]. In stating our results, we abstract properties of the distribution used for the verifier's queries that are sufficient for the general proof strategy of [GR07] to work, and then utilize a different distribution (and some minor changes in the Fourier analysis) to eventually produce equations with all coefficients $+1$. This result is our most significant technical contribution.

1.4 Related work

Rationalizability has been well-studied under the domain of revealed preference theory and social choice theory by economists [Sam48, Var82, Spr00, FST04, BV06, Var06, Ech08]. Traditionally, the questions have been connected with characterizing the implications of various solution concepts to games and market settings, and whether these implications can be tested based on data obtained from consumer choices.

In connection with studying network formation games, while the question of understanding the properties and limitations of equilibria is not new [JW96, DM97, JW01, DJ03, FLM$^+$03], to the best of our knowledge there is no previous work done with respect to either the rationalizability question for these games in general, or the Jackson-Wolinsky model of network formation in particular.

1.5 Outline

In Section 3 we formally define STABLE-PRICES, and observe that it is easy (and even the optimization variant is easy to solve exactly). In Section 4 we define STABLE-VALUES and give a reduction from I-SAT* to it. We then show that I-SAT* is NP-complete (this is not subsumed by our eventual inapproximability result, since this reduction has perfect completeness). In Section 5 we state an approximation preserving reduction from MAX-LIN$_{\mathbb{R}_+}$ to I-SAT*, and we then describe the PCP system (based on a non-trivial modification of [GR07]) that implies ϵ inapproximability for it. This yields the tight $(1/2 + \delta)$ inapproximability for STABLE-VALUES.

2 Jackson-Wolinsky model for network formation games

We describe formally the seminal model for stability of network formation games as formulated by Jackson and Wolinsky [JW96]. The model comprises:

- n agents V

- pairwise distance function $d : V \times V \to \mathbb{R}_+$
- pairwise intrinsic value function $w : V \times V \to \mathbb{R}_+$
- a function $f : \mathbb{R}_+ \to \mathbb{R}$ that defines the contribution of a path of length d
- price profile $c : V \times V \to \mathbb{R}_+$

For a graph G and any two $i, j \in V$, we define $d_G(i, j)$ to be the distance of the shortest path $P(i, j)$ from i to j given by $\sum_{(u,v)\in P(i,j)} d(u, v)$. Let $\Gamma(v)$ be the set of neighbors of v. The utility agent i derives in graph G, denoted $u_i(G)$, is given by:

$$u_i(G) = \sum_{j\in V} f(d_G(i, j))w(i, j) - \sum_{k\in\Gamma(i)} c(i, k)$$

Definition 2.1. *A graph $G = (V, E)$ is said to be pairwise stable with respect to d, w, f, c if:*

1. *for all $i, j \in V$ such that $(i, j) \in G$,*

$$u_i(G) \geq u_i(G - (i, j)) \text{ and } u_j(G) \geq u_j(G - (i, j))$$

2. *for all $i, j \in V$ such that $(i, j) \notin G$,*

$$u_i(G) \geq u_i(G + (i, j)) \text{ or } u_j(G) \geq u_j(G + (i, j))$$

In this definition *both* endpoints must prefer that an edge in G exists (or at least be indifferent to its existence) for the edge to be stable; correspondingly a non-edge in G is stable if even *one* endpoint prefers that is not be present (or at least is indifferent to its presence).

3 Finding stable prices when intrinsic values are known

In the first rationalization problem that we will call STABLE-PRICES, we consider a scenario where the intrinsic values are known but the edge-prices are not. We are given a collection of undirected graphs $G_1, \ldots, G_m$ all of which are formed over a common set of vertices V. In addition, we are given the pairwise distance functions for each G_i, $d_i : V \times V \to \mathbb{R}_+$. The rationalizability question entails inferring the prices that players in each of the graphs would have to pay given that the graphs are in pairwise equilibrium.

In the above problem (and in the subsequent version, STABLE-VALUES) we assume that f is given to us in the form of a circuit that takes as input numbers represented with some number of bits of precision that is polynomial in the size of the rest of the input. This precision is sufficient to exactly express the shortest path distance to which f is applied.

Theorem 3.1. *There is a polynomial-time algorithm for* STABLE-PRICES.

Proof. The algorithm is based on the simple intuition that for an edge to exist between two vertices v, w it must be the case that the *marginal utility* that v derives by building that edge must be at least 0 and the same for u, while for a non-edge (v, w), either v or w must have a marginal utility at most 0. These marginal utilities are easy to compute, and they yield a system of linear inequalities exactly describing stable edge prices. For details, please refer to a full version of this paper. □

Problem 1. STABLE-PRICES

Given: Collection of graphs $G_1, \ldots, G_m$ over common set of vertices V
Pairwise distance functions $d_i : V \times V \to \mathbb{R}_+$ for $i = 1, \ldots, m$
Intrinsic value function $w : V \times V \to \mathbb{R}_+$
Path distance contribution function $f : \mathbb{R}_+ \to \mathbb{R}$, where f is non-increasing.

Find: Price profile c that rationalizes $G_1, \ldots, G_m$, i.e. for which G_i is stable w.r.t. d_i, w, f, c for all i, if it exists.

We think of a price profile c as *rationalizing* (u, v) *in* G_i if the conditions in Definition 2.1 hold for (u, v). An optimization version of the STABLE-PRICES problem is, given a collection of graphs $G_1, \ldots, G_m$, to find a price profile that rationalizes the maximum number of "(u, v) in G_i" pairs. It is easy to see that the same algorithm described above will also work to find a price profile that would solve the optimization problem for STABLE-PRICES exactly.

Corollary 3.2. *Given an instance of* STABLE-PRICES *there exists a polynomial-time algorithm to construct a price profile c that rationalizes the maximum number of "(u, v) in G_i" pairs.*

4 Finding stable intrinsic values when prices are known

In the problem of STABLE-VALUES, we consider the scenario where the edge-prices are known but the intrinsic values function w is unknown. We define STABLE-VALUES below and show that STABLE-VALUES is NP-hard by a reduction from a special variant of I-SAT (defined below).

Problem 2. STABLE-VALUES

Given: Collection of graphs $G_1, \ldots, G_m$ over common set of vertices V
Pairwise distance functions $d_i : V \times V \to \mathbb{R}_+$ for $i = 1, \ldots, m$
Path distance contribution function $f : \mathbb{R}_+ \to \mathbb{R}$ where f is non-increasing
Price profile $c : V \times V \to \mathbb{R}_+$

Find: Intrinsic values function $w : V \times V \to \mathbb{R}_+$ that rationalizes $G_1, \ldots, G_m$, i.e. for which G_i is pairwise stable w.r.t. d_i, w, f, c for all i, if it exists.

Theorem 4.1. I-SAT* *is NP-complete.*

Proof. We give a reduction from 3-SAT. Consider an instance of 3-SAT given by n variables $x_1, \ldots, x_n$ and m clauses $C_1, \ldots, C_m$. For each variable x_i our instance of I-SAT* has two variables z_i^t, z_i^f and for each

Problem 3. I-SAT*

Given: n variables $x_1, \ldots, x_n$, and a partition $S \cup T$ of $[n]$.
m clauses that are either of type 1:
$\left(\sum_{i \in S} a_i x_i \leq 1\right)$ OR $\left(\sum_{j \in T} c_j x_j \leq 1\right)$
or of type 2:
$\left(\sum_{i \in S} a_i x_i \geq 1\right)$ AND $\left(\sum_{j \in T} c_j x_j \geq 1\right)$,
where all $a_i, c_j \geq 0$.

Find: An assignment for $x_1, \ldots, x_n \in \mathbb{R}_+$ satisfying all m clauses, if it exists.

clause C_j, we have three auxiliary variables a_j, b_j, c_j. In total, we will have $2n + 3m$ unknowns in our I-SAT* instance and we partition this into $S = \{z_i^t, z_i^f | i = 1, \ldots, n\}$ and $T = \{a_j, b_j, c_j | j = 1, \ldots, m\}$.

For each clause C_j, we produce three I-SAT* clauses of type 1 and one of type 2. Suppose $C_j = (x_p + \bar{x}_q + x_r)$ by way of example. Our reduction produces the type 2 clause:

$$\left(\frac{z_i^t}{3} + \frac{z_i^f}{3} \geq 1\right) \text{ AND } \left(\frac{a_j}{4} + \frac{b_j}{4} + \frac{c_j}{4} \geq 1\right) \tag{1}$$

and the following three type 1 clauses

$$\begin{aligned} \left(z_p^t \leq 1\right) \quad &\text{OR} \quad (a_j \leq 1) \\ (z_q^f \leq 1) \quad &\text{OR} \quad (b_j \leq 1) \\ \left(z_r^t \leq 1\right) \quad &\text{OR} \quad (c_j \leq 1) \end{aligned}$$

Suppose that there exists a satisfiable assignment for the 3-SAT instance. Then, for each x_i that is true in this assignment we let $z_i^t = 1, z_i^f = 2$ and for each x_i that is false, we let $z_i^t = 2, z_i^f = 1$. For each clause C_j, set one of the a_j, b_j, c_j variables to 2 that corresponds to a true literal, and the others to 1. These assignments satisfy all four I-SAT* clauses corresponding to clause C_j.

In the other direction, we claim that setting x_i to be true for exactly those i such that $z_i^t \leq 1$ is a satisfying assignment. For each clause C_j, it cannot be the case that a_j, b_j, c_j are all ≤ 1, as this would violate the associated type 2 I-SAT* clause. Thus at least one of the variable appearing positively in C_j must have $z_i^t \leq 1$ or one of the variables appearing negatively must have $z_i^f \leq 1$. Moreover, the type 2 I-SAT* clause ensures that it can't be the case that z_i^t and z_i^f are both ≤ 1, so as claimed, the derived assignment is consistent, and it is a satisfying assignment since it makes at least one literal in each clause true. □

The following reduction is actually approximation preserving, a fact that we will use in Section 5.

Theorem 4.2. STABLE-VALUES *is NP-complete.*

Proof. Our proof is by reduction from I-SAT*.

Suppose we are given an instance of I-SAT* with m clauses $C_1, \ldots, C_m$ over n unknowns $x_1, \ldots, x_n$ where each clause C_i is of type 1:

$$\left(\sum_{i \in S} a_i x_i \leq 1\right) \text{ OR } \left(\sum_{j \in T} c_j x_j \leq 1\right)$$

or of type 2

$$\left(\sum_{i\in S} a_i x_i \geq 1\right) \text{ AND } \left(\sum_{j\in T} c_j x_j \geq 1\right)$$

and all $a_i, c_j \geq 0$ and $S \cup T$ is a partition of $[n]$. Fix L to be the least integer that is greater than all of the a_i, c_j coefficients appearing in these clauses. Our reduction produces m edge-weighted graphs $G_1, \ldots, G_m$ on $(n+3)$ vertices labeled $v_1, \ldots, v_n, u, s, t$ as follows.

For each clause C (of either type), we construct a base graph G. For each $i \in S$, G contains edge (s, v_i) with weight $d(s, v_i) = L + a_i$ and edge (t, v_i) with weight $d(t, v_i) = L$. Similarly for each $j \in T$, G contains edge (s, v_j) with weight $d(s, v_j) = L$, and edge (t, v_j) with weight $d(t, v_j) = L + c_j$. We also include edges (s, u), (u, t) with weights $d(s, u) = d(t, u) = L/2$.

If C is of type 1, the base graph *is* the graph associated with C; refer to Figure 1. If C is of type 2, we add the edge (s, t) with weight $d(s, t) = 0$ to the base graph; refer to Figure 2.

We set the cost of an edge from s to t to be $c(s, t) = 1$. Note that with the exception of the edge (s, t), all of our graphs contain exactly the same set of edges. For any pair of vertices (p, q) other than (s, t), we set $c(p, q) = 0$ if (p, q) is an edge and $c(p, q) = \infty$ if (p, q) is a non-edge.

Finally, we define the piecewise-linear function f as follows:

$$f(x) = \begin{cases} -L & 0 \leq x \leq L \\ -x & x > L \end{cases}$$

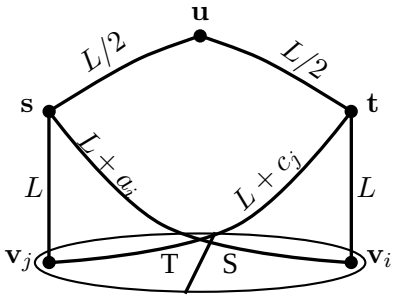

Figure 1: G from clause of type 1, with edge weights

This completes the description of the reduction. We now show that the stability conditions for each graph are exactly the conjunction or disjunction of linear inequalities of the associated I-SAT* clause.

Fix a graph G associated with I-SAT* clause C. We first note that all stability conditions not involving edge (s, t) are trivially satisfied. This is because for any other pair (p, q), the cost is 0 if (p, q) is an edge in G (so both p and q are indifferent to keeping it, which makes it stable), and the cost is ∞ if (p, q) is a non-edge in G (so neither p nor q benefit from adding it, making it stable).

Now we focus on the stability conditions for (s, t), which are

$$u_s(G + (s, t)) \leq u_s(G) \text{ OR } u_t(G + (s, t)) \leq u_t(G)$$

if (s, t) is a non-edge in G and

$$u_s(G - (s, t)) \geq u_s(G) \text{ AND } u_t(G - (s, t)) \geq u_t(G)$$

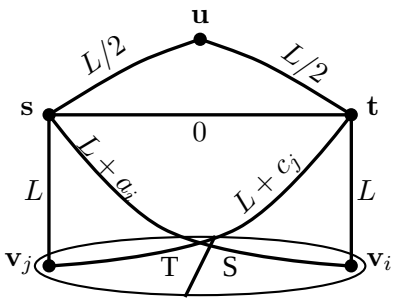

Figure 2: G from clause of type 2, with edge weights

if (s,t) is an edge in G. The following notation will be helpful. For a graph G, edge e and pair of vertices u, v, we define

$$\Delta f_{+e}(u,v) = f(d_{G+e}(u,v)) - f(d_G(u,v))$$

which can be thought of as the "value added" (for vertices u and v) by adding edge e to the graph G. Similarly we define

$$\Delta f_{-e}(u,v) = f(d_G(u,v)) - f(d_{G-e}(u,v))$$

which can be thought of as the "value added" (for vertices u and v) by deleting edge e from the graph G. After expanding and rearranging terms, the pairwise stability conditions for (s,t) then become:

$$\sum_v \Delta f_{+(s,t)}(s,v) \cdot w(s,v) \leq c(s,t)$$
OR
$$\sum_v \Delta f_{+(s,t)}(t,v) \cdot w(t,v) \leq c(s,t)$$

if (s,t) is a non-edge in G and

$$\sum_v \Delta f_{-(s,t)}(s,v) \cdot w(s,v) \geq c(s,t)$$
AND
$$\sum_v \Delta f_{-(s,t)}(t,v) \cdot w(t,v) \geq c(s,t)$$

if (s,t) is an edge in G.

Table 1: Shortest path distance for G constructed from a type 1 clause

(p,q)	$d_G(p,q)$	$d_{G+(s,t)}(p,q)$	$\Delta f_{+(s,t)}(p,q)$
$(s,v_i)\ i \in S$	$L + a_i$	L	a_i
$(t,v_i)\ i \in S$	L	L	0
$(s,v_j)\ j \in T$	L	L	0
$(t,v_j)\ j \in T$	$L + c_j$	L	c_j
(s,t)	L	0	0
$(s,u),(t,u)$	$L/2$	$L/2$	0

Table 2: Shortest path distance for G constructed from a type 2 clause

(p,q)	$d_{G-(s,t)}(p,q)$	$d_G(p,q)$	$\Delta f_{-(s,t)}(p,q)$
(s,v_i) $i \in S$	$L + a_i$	L	a_i
(t,v_i) $i \in S$	L	L	0
(s,v_j) $j \in T$	L	L	0
(t,v_j) $j \in T$	$L + c_j$	L	c_j
(s,t)	L	0	0
$(s,u),(t,u)$	$L/2$	$L/2$	0

If (s,t) is a non-edge in G (so G was constructed from a type 1 clause C), then Table 1 calculates the coefficients of $w(s,\cdot)$, and $w(t,\cdot)$ in the above inequalities; if (s,t) is an edge in G (so G was constructed from a type 2 clause C) then Table 2 calculates the coefficients. Plugging these in, and using the fact that $c(s,t) = 1$, we get exactly

$$\sum_{i \in S} a_i \cdot w(s,v_i) \leq 1$$
$$\text{OR}$$
$$\sum_{j \in T} c_i \cdot w(t,v_j) \leq 1$$

in the case that G was constructed from type 1 clause C_i, and

$$\sum_{i \in S} a_i \cdot w(s,v_i) \geq 1$$
$$\text{AND}$$
$$\sum_{j \in T} c_i \cdot w(t,v_j) \geq 1$$

in the case that G was constructed from type 2 clause C_i.

Thus in both cases the stability conditions arising in graph G constructed from I-SAT* clause C are exactly the constraints given by the clause, with the intrinsic values $w(s,v_i)$ playing the role of the variables x_i for $i \in S$, and the intrinsic values $w(s,v_j)$ playing the role of the variables x_j for $j \in T$.

This completes the proof of Theorem 4.2. □

We now get to the meat of the paper, where we give a tight inapproximability result for an optimization version of STABLE-VALUES.

5 A tight inapproximability result for STABLE-VALUES

Before defining the optimization version of STABLE-VALUES, we need to define the notion of *active pairs*:

Definition 5.1. *Given an instance of* STABLE-VALUES *containing a collection of graphs $G_1, \ldots, G_m$ all over a set of vertices V and a price profile $c : V \times V \to \mathbb{R}_+$, a pair $(u,v) \in V \times V$ is said to be an* active pair *if it is not the case that i) $(u,v) \in G_i$ for all i AND $c(u,v) = 0$, or ii) $(u,v) \notin G_i$ for any i AND $c(u,v) = \infty$.*

As explained in the introduction, a pair (u,v) that is not an active pair effectively "comes for free" since regardless of what the intrinsic values for u, v are, the pairwise stability conditions are trivially satisfied because $c(u,v) = 0$ if $(u,v) \in G_i$ for all i and $c(u,v) = \infty$ if $(u,v) \notin G_i$ for any i.

As before, we think of an intrinsic values function w as rationalizing an active pair (u, v) in G_i if the stability conditions in Definition 2.1 hold for (u, v). Then, the MAX-STABLE-VALUES problem is to seek intrinsic values that rationalize the *maximum* number of *active pairs* (u, v), counted separately for each G_i. We observed in the introduction that there is a trivial $1/2$-factor approximation algorithm. More precisely, for each non-edge, either adding it does not change the aggregate distance contributions for either of its endpoints (in which case it is stable regardless of the intrinsic valuations), or else it is rationalized when the intrinsic valuations are all set to 0; similarly for each edge, either removing it does not change the aggregate distance contributions for its endpoints (in which case it is stable iff its cost is 0, regardless of the intrinsic valuations), or else it is rationalized when the intrinsic valuations are all set to ∞. Thus one of the two extremes (setting all intrinsic valuations to 0, or all intrinsic valuations to ∞) rationalizes at least 1/2 of the "(u, v) in G_i" pairs that are rationalized by an optimum solution. We prove in this section that this $1/2$ factor is tight assuming $P \neq NP$.

To our end of showing a hardness result for MAX-STABLE-VALUES we show an inapproximability result for MAX-LIN$_{\mathbb{Z}_+}$, which is the main technical contribution of this paper. Although MAX-STABLE-VALUES is actually defined over the reals and would admittedly require us to show a hardness result for MAX-LIN$_{\mathbb{R}_+}$, we are able to employ a clever trick shown in [GR07] that makes it sufficient for us to work with MAX-LIN$_{\mathbb{Z}_+}$ and then carry the result over to the reals as long as we can ensure that the co-efficients in the MAX-LIN$_{\mathbb{Z}_+}$ instance we obtain are bounded and the equations have sparse support.

Problem 4. MAX-LIN$_{\mathbb{Z}_+}$

Given: n variables $x_1, \ldots, x_n$
m equations, each of which is of the type
$\sum_i a_i x_i = b$
where $a_i, b \in \mathbb{Z}_+$ for all $i = 1, \ldots, n$

Find: An assignment for $x_1, \ldots, x_n \in \mathbb{Z}_+$ that satisfies the maximum number of equations.

Theorem 5.2. *Given an instance of* MAX-LIN$_{\mathbb{Z}_+}$*, for all positive constants ϵ, δ it is NP-hard to distinguish between the following two cases:*

- *There exists a solution satisfying at least a $(1 - \epsilon)$ fraction of the linear equations in the instance.*
- *Every solution satisfies at most a δ fraction of the linear equations.*

The above promise problem will be referred to as MAX-LIN$_{\mathbb{Z}_+}(1 - \epsilon, \delta)$. Our proof follows the outline of the proof for Theorem 3.4 in [GR07]. In the remainder of this section, we go through the proof of Theorem 5.2, pointing out the crucial points where our proof needs to differ from [GR07]. We first define the LABEL-COVER problem below.

Definition 5.3. *An instance of the* LABEL-COVER(c, s) *problem comprises a bipartite graph $\mathcal{H} = (\mathcal{A}, \mathcal{B}, \mathcal{E})$, a set of labels Σ and a set of projection mappings $\pi_e : \Sigma \to \Sigma$ for each edge $e \in \mathcal{E}$. An assignment $A : (\mathcal{A} \cup \mathcal{B}) \to \Sigma$ is a mapping from the set of vertices onto the set of labels and is* legal *for an edge $e = (u, v)$ if $\pi_e(A(u)) = A(v)$. We wish to ascertain for this instance of the problem if*

- *there exists an assignment A that is legal for at least a c fraction of edges, or*

- *every assignment is legal for at most an s fraction of the edges.*

The following theorem gives a hardness of approximation result for LABEL-COVER and is due to [ALM+98] based on a result of Raz's. [Raz98]

Theorem 5.4. *([ALM+98, Raz98]) There exists a constant $\gamma > 0$ such that for all sufficiently large constant-sized alphabets Σ, it is NP-hard to distinguish between an instance of* LABEL-COVER *that has an assignment legal for all edges and one for which every assignment is legal for at most $1/|\Sigma|^\gamma$ fraction of edges.*

As we noted in Section 1.3, there doesn't seem to be an easy reduction from MAX-LIN$_{\mathbb{Z}}$ (shown to be hard to approximate in [GR07]) or from MAX-LIN$_{\mathbb{F}_p}$ (shown to be hard to approximate in [Hås01]). Both those results are obtained by reductions from LABEL-COVER and involve constructing equations of the form $x + y - z = c$. Most of our effort in our proof is spent on giving an alternative reduction that gives rise to equations with coefficients in $\mathbb{Z}_+$. Specifically, our proof abstracts properties of the verifier query distribution that are sufficient for main steps of the [GR07] proof. We then specify a different distribution than the one in [GR07] that satisfies these properties, as well as an additional symmetry property that is key to our final PCP system for MAX-LIN$_{\mathbb{Z}_+}$.

5.1 Proof of Theorem 5.2

This subsection is devoted to the proof of Theorem 5.2. Consider an instance of the LABEL-COVER$(1, \delta)$ problem comprising the bipartite graph $G(U, V, E)$ over n vertices and m edges, a set of labels $\Sigma = \{1, \ldots, h\}$ and constraint relations $\pi_e : \Sigma \to \Sigma$.

In our PCP system, the proof comprises the labels for all vertices encoded using the Long Code we define below.

Definition 5.5. *[GR07] For a label $r \in [h]$, the codeword $\mathcal{C}(r)$ is an evaluation of the projection function $f_r : \mathbb{Z}_+^h \to \mathbb{Z}_+$ given by $f_r((z_1, \ldots, z_h)) = z_r$ over $\mathbb{Z}_+^h$. In other words, $\mathcal{C}(r)[x] = x_r$.*

In other words, the proof is given to the verifier as a sequence $(\mathcal{C}(A(v_1)), \ldots, \mathcal{C}(A(v_n)))$ where A is the purported legal assignment. The verifier makes queries to the proof at three locations. These locations are chosen based on probability distributions P_1, P_2, and Q over $\mathbb{Z}_+^h$. In the end, as in [GR07], we will discretize and truncate these distributions so that the verifier uses $O(\log n)$ randomness, so even though the proof is formally of infinite length, the verifier only ever sees a (fixed) polynomially large fragment.

Our proof will use the following properties of probability distributions over $\mathbb{Z}_+^h$:

Definition 5.6. *P is said to be* (M, δ)-heavy *if*

$$\sum_{x \in [M]^h} P(x) \geq (1 - \delta).$$

Definition 5.7. *P is said to be* (δ, L)-decay-resilient *if for all $x \in [L]^h$ and any $y \in \mathbb{Z}_+^h$*

$$\frac{P(y + x)}{P(y)} \geq \delta.$$

Definition 5.7 encapsulates and highlights a crucial property used in a technical step in the soundness analysis. We are now ready to define P_1, P_2, which have the following form (they are parameterized by $p, c_1, c_2, \Gamma_1, \Gamma_2$):

Definition 5.8. *For $j = 1, 2$ we define the functions P_j over $\mathbb{Z}_+^h$ to be*

$$P_j((x_1, \ldots, x_h)) \quad = \quad \Gamma_j \prod_{i=1}^{h} e^{-c_j |x_i - \frac{(p-1)}{2}|}. \tag{2}$$

where $\Gamma_j = (1 - e^{-c_j})^h$; $j = 1, 2$ are some constants.

The following proposition describes exactly how we must choose M, t, p, c_j:

Proposition 5.9. *There exist positive integers $M = M(h, \delta), t = t(h, \delta)$, a prime $p = \Omega(M)$, and positive reals $c_1, c_2, \Gamma_1, \Gamma_2$ for which P_1 and P_2 with these parameter settings are probability distributions, both (M, δ)-heavy, and P_2 is $(1/4, M + t)$-decay-resilient.*

Proof. We first prove that P_1 and P_2 as defined above in Definition 5.8 are probability distributions. This is easy to see since for any choice of c_1, c_2, we can appropriately choose normalization constants Γ_1, Γ_2 so as to ensure that P_1, P_2 are probability distributions. Next, we show that P_1, P_2 are (M, δ)-heavy. Note first that for any $(x_1, \ldots, x_h) \in \mathbb{Z}_+^h - [M]^h$, there exists some r such that $x_r > M$. Therefore:

$$\begin{aligned}
\sum_{(x_1,\ldots,x_h) \in \mathbb{Z}_+^h - [M]^h} P_j((x_1, \ldots, x_h)) \quad &= \quad \sum_{(x_1,\ldots,x_h) \in \mathbb{Z}_+^h - [M]^h} \Gamma_j e^{-c \sum_{i=1}^h |x_i - \frac{p-1}{2}|} \\
&\leq \quad \sum_{r=1}^{h} \sum_{(x_1,\ldots,x_h): x_r > M} \Gamma_j e^{-c_j \sum_{i=1}^h |x_i - \frac{p-1}{2}|} \\
&\leq \quad h(1 - e^{-c_j})^h e^{-c_j \left(M - \frac{p-1}{2}\right)}
\end{aligned}$$

The right-hand side is less than δ for

$$M - \frac{p-1}{2} \geq \frac{1}{c_j} \cdot \ln \left[\frac{h(1 - e^{-c_j})^h}{\delta} \right]$$

To show that P_2 is $(1/4, M + t)$-decay-resilient, we invoke Lemma 6.5 from [GR07], which we restate using our terminology below:

Lemma 5.10. *(Lemma 6.5, [GR07]) For all $N > 0$, and $c_2 < \frac{\ln 4}{hN}$, the following is true for any $x \in [N]^h, y \in \mathbb{Z}_+^h$:*

$$\frac{P_2(y + x)}{P_2(x)} \geq 1/4.$$

We compile all the constraints on our ability to choose M, c_j, t, p:

$$M - \frac{p-1}{2} \quad \geq \quad \frac{1}{c_j} \cdot \ln \left[\frac{h(1 - e^{-c_j})^h}{\delta} \right] \tag{3}$$

$$c_2 \quad \leq \quad \frac{\ln 4}{h(M + t)} \tag{4}$$

First, we set t to be the least integer greater than h^2/δ. Next, we proceed to set c_1, c_2. Our choice of c_1 is determined directly from [GR07] (pg 6), in order to ensure that the key lemma we use from their result in the soundness analysis (Lemma 5.15) goes through, and is given by:

$$c_1 = \min \left\{ \frac{1}{h} \ln \left(1 + 4\delta^2\right), \ln \left(1 + 4 \left(\frac{\delta}{4} \right)^5 \cdot \frac{\delta}{2h} \right) \right\}.$$

We set c_2, based on (3) and (4), to be:

$$c_2 = \ln\left[\frac{1}{1-(\delta 2^{1/h}/h)^{1/h}}\right],$$

Choosing M to be the smallest integer greater than $5\left(\frac{h}{\delta}\right)^{\frac{h}{\ln(1+4\delta^2)}}$, and p to be $p = \lceil 3M/2 \rceil$, we will satisfy all the required properties of P_1, P_2. □

From now on we will fix M, p and the parameters defining probability distributions P_1, P_2 as in the above proposition. We let μ be a random variable over $\mathbb{Z}_+^h$ generated by picking each coordinate to be 0 with probability $(1-\epsilon)$ and an integer chosen randomly from $[t]$ with probability ϵ. We denote Q to be the probability distribution with which μ is chosen.

Now we can describe the verifier test. Let X_1, X_2 be random variables over $\mathbb{Z}^h$ distributed according to P_1, P_2, respectively. We use $x \circ \pi_e$ to denote the permutation of π_e applied to the co-ordinates of $x \in \mathbb{Z}^h$. In other words $(x \circ \pi_e)_i = x_{\pi_e(i)}$. Using this notation, define random variable $X_3 = \mathbf{p} - (X_1 \circ \pi_e + X_2 + \mu)$, where $\mathbf{p}$ denotes $(p, p, \ldots, p)$. The equation that the verifier checks is:

$$\mathcal{C}(A(u))[X_1] + \mathcal{C}(A(v))[X_2] + \mathcal{C}(A(v))[X_3] \quad = \quad p. \tag{5}$$

Lemma 5.11. *The PCP system for* MAX-LIN$_{\mathbb{Z}_+}$ *described above has* $(1-\epsilon)$ *completeness.*

Proof. Suppose A is indeed a legal assignment for all edges $e \in E(G)$. This means that for any edge $e = (u, v)$, $\pi_e(A(v)) = A(u)$. Therefore, $\mathcal{C}(A(u))[X_1] + \mathcal{C}(A(v))[X_2] + \mathcal{C}(A(v))[X_3]$

$$\begin{aligned} &= X_{1A(u)} + X_{2A(v)} + \mathbf{p}_{A(v)} - (X_1 \circ \pi_e + X_2 + \mu)_{A(v)} \\ &= X_{1A(u)} + X_{2A(v)} + p - X_{1\pi_e(A(v))} - X_{2A(v)} - \mu_{A(v)} \\ &= p - \mu_{A(v)} \end{aligned}$$

Recalling how we picked μ, we know that $\mu_{A(v)}$ is 0 with probability exactly $(1-\epsilon)$ and hence, (5) is satisfied with probability $(1-\epsilon)$. □

Lemma 5.12. *The PCP system for* MAX-LIN$_{\mathbb{Z}_+}$ *described above has* 19δ *soundness error.*

Proof. To argue for soundness, suppose A is an assignment that causes the verifier to accept with probability at least $\delta' = 19\delta$. This means that over all $e = (u, v)$ chosen uniformly at random from $E(G)$, and $\mathbf{x}_1, \mathbf{x}_2, \mu$ chosen according to their respective probability distributions P_1, P_2, Q from $\mathbb{Z}_+^h$:

$$\Pr_{e, X_1, X_2, \mu}[\mathcal{C}(A(u))[X_1] + \mathcal{C}(A(v))[X_2] + \mathcal{C}(A(v))[\mathbf{p} - (X_1 \circ \pi_e + X_2 + \mu)] = p] \quad \geq \quad 19\delta$$

The following fact is handy:

Fact 5.13. *Let P be a $(1/4, (M+t))$-decay-resilient probability distribution over $\mathbb{Z}_+^h$. Then, for any $y \in [M+t]^h$ and all $x \in \mathbb{Z}_+^h$:*

$$P(x) \leq 2\sqrt{P(x+y) \cdot P(x)}$$

Proof. Since P is $(1/4, (M+t))$-decay-resilient it satisfies the following inequality:

$$\begin{aligned} P(x) &\leq 4P(x+y) \\ &\leq 2\sqrt{P(x+y) \cdot P(x)} \end{aligned}$$

□

The following lemma is based on the first step of the proof technique used in [GR07] applied to our setting:

Lemma 5.14. *Let P_1, P_2 be probability distributions over $\mathbb{Z}_+^h$ and Q be a probability distribution over $[t]^h$ such that P_1, P_2, Q satisfy the following properties:*

1. *P_1, P_2, Q are $(p/3, \delta)$-heavy.*
2. *P_2 is $(1/4, M+t)$-decay-resilient.*
3. *P_2 is symmetric around $(p/2, \ldots, p/2)$, i.e. $P_2(x) = P_2(\mathbf{p} - x)$*

Suppose that X_1, X_2, μ are chosen respectively from distributions P_1, P_2, Q and $e = (u, v)$ is chosen uniformly at random. Denote X_3 to be $X_3 = \mathbf{p} - (X_1 \circ \pi_e + X_2 + \mu)$. Then, if:

$$\Pr_{e, X_1, X_2, \mu}[\mathcal{C}(A(u))[X_1] + \mathcal{C}(A(v))[X_2] + \mathcal{C}(A(v))[X_3] = p] \quad \geq \quad 23\delta,$$

letting $\Upsilon_p^{(u,v)}(X_1, X_2, \mu)$ be the indicator variable for the event

$$\mathcal{C}(A(u))[X_1] + \mathcal{C}(A(v))[X_2] + \mathcal{C}(A(v))[\mathbf{p} - (X_1 \circ \pi_e + X_2 + \mu)] = 0 \mod p$$

the following must hold:

$$E_{(u,v)}\left[\sum_{x_1, x_2, x_3 \in [p/3]^h} P_1(x_1)\sqrt{P_2(x_2) \cdot P_2(x_3)}Q(\mu)\Upsilon_p^{(u,v)}(x_1, x_2, \mu)\right] \quad \geq \quad 8\delta$$

Proof. Note first that since P_1 is $(p/3, \delta)$-heavy, and $p/3 < M$, it is also (M, δ)-heavy. Furthermore, since μ is by default chosen from $[t]^h$, with probability at most δ our choice of X_1 will lie outside $[M]^h$ and so:

$$\Pr_{e, X_1, X_2, \mu}\left[\mathcal{C}(A(u))[X_1] + \mathcal{C}(A(v))[X_2] + \mathcal{C}(A(v))[X_3] = p | X_1 \in [M]^h\right] \quad \geq \quad 22\delta$$

Denoting $\Upsilon^e(x_1, x_2, \mu)$ to be the indicator variable for the event:

$$\mathcal{C}(A(u))[x_1] + \mathcal{C}(A(v))[x_2] + \mathcal{C}(A(v))[\mathbf{p} - (x_1 \circ \pi_e + x_2 + \mu)] = p$$

we can rewrite the left-hand side above in terms of an expectation over all edges $e(u, v)$:

$$E_e\left[\sum_{x_1 \in [M]^h, x_2 \in \mathbb{Z}_+^h, \mu \in [t]^h} P_1(x_1)P_2(x_2)Q(\mu)\Upsilon^{(u,v)}(x_1, x_2, \mu)\right] \quad \geq \quad 22\delta \tag{6}$$

Combining (6) with Property 2 and Fact 5.13, we get:

$$E_e\left[\sum_{x_1 \in [M]^h, x_2 \in \mathbb{Z}_+^h, \mu \in [t]^h} P_1(x_1)\sqrt{P_2(x_2) \cdot P_2(x_2 + x_1 \circ \pi_e + \mu)}Q(\mu)\Upsilon^{(u,v)}(x_1, x_2, \mu)\right] \quad \geq \quad 11\delta$$

Since P_2 is symmetric around $(p/2, \ldots, p/2)$, $P_2(x_2 + x_1 \circ \pi + \mu) = P_2(\mathbf{p} - (x_2 + x_1 \circ \pi + \mu))$ and hence the above inequality becomes:

$$E_e \left[\sum_{x_1 \in [M]^h, x_2 \in \mathbb{Z}_+^h, \mu \in [t]^h} P_1(x_1) \sqrt{P_2(x_2) \cdot P_2(\mathbf{p} - (x_1 \circ \pi_e + x_2 + \mu))} Q(\mu) \Upsilon^{(u,v)}(x_1, x_2, \mu) \right] \geq 11\delta$$

Observing that $\Upsilon_p^{(u,v)}(x_1, x_2, \mu) \geq \Upsilon^{(u,v)}(x_1, x_2, \mu)$ and denoting $x_3 = \mathbf{p} - (x_1 \circ \pi_e + x_2 + \mu)$, we have:

$$E_e \left[\sum_{x_1 \in [M]^h, x_2 \in \mathbb{Z}_+^h, \mu \in [t]^h} P_1(x_1) \sqrt{P_2(x_2) \cdot P_2(x_3)} Q(\mu) \Upsilon_p^{(u,v)}(x_1, x_2, \mu) \right] \geq 11\delta$$

Since P_1, P_2, Q are $(p/3, \delta)$-heavy:

$$E_{(u,v)} \left[\sum_{x_1, x_2, \mu \in [p/3]^h} P_1(x_1) \sqrt{P_2(x_2) \cdot P_2(x_3)} Q(\mu) \Upsilon_p^{(u,v)}(x_1, x_2, \mu) \right] \geq 8\delta \tag{7}$$

□

Note that the function $\Upsilon_p^{(u,v)} : [p]^{3h} \to \{0, 1\}$ is given by

$$\Upsilon_p^{(u,v)}(x_1, x_2, \mu) = \begin{cases} 1 & \mathcal{C}(A(u))[x_1] + \mathcal{C}(A(v))[x_2] + \mathcal{C}(A(v))[x_3] = 0 \mod p, \\ 0 & \text{otherwise}. \end{cases}$$

$\Upsilon_p^{(u,v)}$ can equivalently be written as below:

$$\Upsilon_p^{(u,v)}(x_1, x_2, \mu) = \frac{1}{p} \sum_{k=0}^{p-1} e^{\frac{2\pi i k}{p}(\mathcal{C}(A(u))[x_1] + \mathcal{C}(A(v))[x_2] + \mathcal{C}(A(v))[x_3])}$$

Substituting this in (7), the left-hand side becomes:

$$E_{(u,v)} \left[\frac{1}{p} \sum_{x_1, x_2, \mu \in [\frac{p}{3}]^h} P_1(x_1) \sqrt{P_2(x_2) \cdot P_2(x_3)} Q(\mu) \left(\sum_{k=0}^{p-1} e^{\frac{2\pi i k}{p}(\mathcal{C}(A(u))[x_1] + \mathcal{C}(A(v))[x_2] + \mathcal{C}(A(v))[x_3])} \right) \right]$$

We further simplify the term within the expectation:

$$\begin{aligned} & \frac{1}{p} \sum_{x_1, x_2, \mu \in [\frac{p}{3}]^h} P_1(x_1) \sqrt{P_2(x_2) \cdot P_2(x_3)} Q(\mu) \left(\sum_{k=0}^{p-1} e^{\frac{2\pi i k}{p}(\mathcal{C}(A(u))[x_1] + \mathcal{C}(A(v))[x_2] + \mathcal{C}(A(v))[x_3])} \right) \\ = \; & \frac{1}{p} \sum_{k=0}^{p-1} \sum_{x_1, x_2, \mu \in [\frac{p}{3}]^h} Q(\mu) \left(P_1(x_1) e^{\frac{2\pi i k}{p} \mathcal{C}(A(u))[x_1]} \right) \left(\sqrt{P_2(x_2)} e^{\frac{2\pi i k}{p} \mathcal{C}(A(v))[x_2]} \right) \left(\sqrt{P_2(x_3)} e^{\frac{2\pi i k}{p} \mathcal{C}(A(v))[x_3]} \right) \end{aligned}$$

Setting $\mathcal{U}_u(x) = P_1(x)e^{\frac{2\pi ik}{p}\mathcal{C}(A(u))[x]}$ and $\mathcal{V}_v(x) = \sqrt{P_2(x)}e^{\frac{2\pi ik}{p}\mathcal{C}(A(v))[x]}$, (7) now simplifies to:

$$E_{(u,v)}\left[\frac{1}{p}\sum_{k=0}^{p-1}\sum_{x_1,x_2,\mu\in[\frac{p}{3}]^h} Q(\mu)\mathcal{U}_u(x_1)\mathcal{V}_v(x_2)\mathcal{V}_v(x_3)\right] \geq 8\delta \tag{8}$$

This is where our proof technique has a crucial point of departure from that used in [GR07]. Since our test has only positive co-efficients, we do not have the luxury to make the substitution

$$\sqrt{P_2(x)}e^{-\frac{2\pi ik}{p}\mathcal{C}(A(v))[x]} = \overline{\mathcal{V}_v(x)}$$

that is made in [GR07] which simplifies their analysis.

Consider the Fourier expansion for $\mathcal{U}_u$ described below:

$$\mathcal{U}_u(x) = \sum_{w\in[p]^h}\widehat{\mathcal{U}}_u(w)e^{\frac{2\pi i}{p}\langle w,x\rangle}$$

where

$$\widehat{\mathcal{U}}_u(w) = \frac{1}{p^h}\sum_{x\in[p]^h}\mathcal{U}_u(x)e^{-\frac{2\pi i}{p}\langle w,x\rangle}$$

We substitute this and a similar Fourier expansion for $\mathcal{V}_v$ back in (8):

$$E_{(u,v)}\left[\frac{1}{p}\sum_{k=0}^{p-1}\sum_{x_1,x_2,\mu} Q(\mu)\sum_{w_1}\widehat{\mathcal{U}}_u(w_1)e^{\frac{2\pi i}{p}\langle w_1,x_1\rangle}\sum_{w_2}\widehat{\mathcal{V}}_v(w_2)e^{\frac{2\pi i}{p}\langle w_2,x_2\rangle}\sum_{w_3}\widehat{\mathcal{V}}_v(w_3)e^{\frac{2\pi i}{p}\langle w_3,x_3\rangle}\right] \geq 8\delta$$

Substituting $x_3 = (\mathbf{p} - (x_1\circ\pi_e + x_2 + \mu))$ the left-hand side becomes:

$$\begin{aligned}
& E_{(u,v)}\left[\frac{1}{p}\sum_{k=0}^{p-1}\sum_{x_1,x_2,\mu} Q(\mu)\sum_{w_1}\widehat{\mathcal{U}}_u(w_1)e^{\frac{2\pi i}{p}\langle w_1,x_1\rangle}\sum_{w_2}\widehat{\mathcal{V}}_v(w_2)e^{\frac{2\pi i}{p}\langle w_2,x_2\rangle}\sum_{w_3}\widehat{\mathcal{V}}_v(w_3)e^{\frac{2\pi i}{p}\langle w_3,(\mathbf{p}-(x_1\circ\pi_e+x_2+\mu))\rangle}\right] \\
= \; & E_{(u,v)}\left[\frac{1}{p}\sum_{k=0}^{p-1}\sum_{x_1,x_2,\mu} Q(\mu)\sum_{w_1}\widehat{\mathcal{U}}_u(w_1)e^{\frac{2\pi i}{p}\langle w_1,x_1\rangle}\sum_{w_2}\widehat{\mathcal{V}}_v(w_2)e^{\frac{2\pi i}{p}\langle w_2,x_2\rangle}\sum_{w_3}\widehat{\mathcal{V}}_v(w_3)e^{-\frac{2\pi i}{p}\langle w_3,(x_1\circ\pi_e+x_2+\mu)\rangle}\right] \\
= \; & E_{(u,v)}\left[\frac{1}{p}\sum_{k=0}^{p-1}\sum_{w_1,w_2,w_3}\widehat{\mathcal{U}}_u(w_1)\widehat{\mathcal{V}}_v(w_2)\widehat{\mathcal{V}}_v(w_3)\sum_{x_1}e^{\frac{2\pi i}{p}\langle(w_1-w_3\circ\pi_e^{-1}),x_1\rangle}\sum_{x_2}e^{\frac{2\pi i}{p}\langle(w_2-w_3),x_2\rangle}\sum_{\mu}Q(\mu)e^{-\frac{2\pi i}{p}\langle w_3,\mu\rangle}\right]
\end{aligned}$$

where $w_3\circ\pi_e^{-1}$ denotes the vector obtained by setting $(w_3\circ\pi_e^{-1})_i = \sum_{j\in\pi_e^{-1}(i)} w_{3j}$ for $i = 1,\ldots,h$. Note that for $w_1 \neq w_3\circ\pi_e^{-1}$, $\sum_{x_1} e^{\frac{2\pi i}{p}\langle(w_1-w_3\circ\pi_e^{-1}),x_1\rangle} = 0$ and similarly, for $w_2 \neq w_3$, $\sum_{x_2} e^{\frac{2\pi i}{p}\langle(w_2-w_3),x_2\rangle} = 0$. Setting $w = w_3$, the overall inequality simplifies to:

$$E_{(u,v)}\left[\frac{1}{p}\sum_{k=0}^{p-1}\sum_{w}\left(p^h\widehat{\mathcal{U}}_u(w\circ\pi_e^{-1})\right)\left(p^h\widehat{\mathcal{V}}_v(w)^2\right)\sum_{\mu}Q(\mu)e^{-\frac{2\pi i}{p}\langle w,\mu\rangle}\right] \geq 8\delta \tag{9}$$

Also, note that $|\widehat{\mathcal{V}}_v(w)^2| = \sqrt{\left(\widehat{\mathcal{V}}_v(w)^2\cdot\overline{\widehat{\mathcal{V}}_v(w)^2}\right)} = \sqrt{\left(\widehat{\mathcal{V}}_v(w)\cdot\overline{\widehat{\mathcal{V}}_v(w)}\right)^2} = |\widehat{\mathcal{V}_v}(w)|^2$ using the simple identity that $\mathbf{z}^2\cdot\overline{\mathbf{z}^2} = (\mathbf{z}\cdot\overline{\mathbf{z}})^2$ for any complex number $\mathbf{z}$.

Substituting back in (9), we obtain:

$$E_{(u,v)}\left[\frac{1}{p}\sum_{k=0}^{p-1}\sum_{\omega}\left(p^h\widehat{\mathcal{U}}_u(w\circ\pi_e^{-1})\right)\left(|p^h\widehat{\mathcal{V}}_v(w)|^2\right)\left(\sum_{\mu}Q(\mu)e^{-\frac{2\pi i}{p}\langle w,\mu\rangle}\right)\right] \geq 8\delta \tag{10}$$

We are now ready to use the following lemma, again from [GR07] concerning probability distributions P_1, P_2, Q and some assignment A of labels to vertices in G satisfying (10).

Lemma 5.15. *([GR07]) Let $P_1, P_2, Q : \mathbb{Z}_+^h \to [0,1]$ be probability distributions and $A : V(G) \to [h]$ some assignment of labels to vertices in G satisfying (10). Then, there exists a constant C such that*

$$\Pr_{(u,v)}\left[A \text{ is legal for } (u,v)\right] \geq \delta^4/96C^2$$

By choosing our original instance of LABEL-COVER$(1,\delta)$ to be such that h is large enough, we can ensure that $\delta^4/96C^2 \geq 1/h^\gamma$. This gives us a soundness of 19δ as required. □

5.2 Tying it all together

Theorem 5.16. *For all constants $\epsilon, \delta > 0$, the problem* MAX-LIN$_{\mathbb{R}_+}(1-\epsilon,\delta)$ *is NP-hard.*

Proof. We just need the following argument from [GR07], which we apply to $\mathbb{Z}_+$ and $\mathbb{R}_+$ as opposed to $\mathbb{Z}$ and $\mathbb{R}$.

We give a reduction from MAX-LIN$_{\mathbb{Z}_+}(1-\epsilon,\delta/8)$. Note that our reduction in proving that this problem was hard produced an instance of MAX-LIN$_{\mathbb{Z}_+}$ where each equation consists of three variables:

$$x_i + x_j + x_k = c$$

where $x_i, x_j, x_k, c \in \mathbb{Z}_+$. The MAX-LIN$_{\mathbb{R}_+}$ instance we construct will have exactly the same set of these equations:

$$x'_i + x'_j + x'_k = c$$

with x'_i, x'_j, x'_k chosen from $\mathbb{R}_+$. A solution in integers to the original MAX-LIN$_{\mathbb{Z}_+}$ instance is automatically a solution to the MAX-LIN$_{\mathbb{R}_+}$ instance.

Suppose that with probability at least δ over the choice of equations in the instance, a solution is feasible. Then, for any such equation:

$$x'_i + x'_j + x'_k = c$$

by choosing each variable x_s to be either $\lfloor x'_s \rfloor$ or $\lceil x'_s \rceil$ uniformly at random for $s = i, j, k$ we will have satisfied the equation in the corresponding MAX-LIN$_{\mathbb{Z}_+}$ instance with probability at least $\delta/8$ thereby contradicting the hardness assumption we made for the instance. □

Theorem 5.17. *For all constants $\epsilon, \delta > 0$, the problem* I-SAT$^*(1-\epsilon, 1/2+76\delta)$ *is NP-hard.*

Proof. The proof is by reduction from MAX-LIN$_{\mathbb{R}_+}(1-\epsilon, 19\delta)$ over a set of unknowns $x_1, \ldots, x_n$. Our I-SAT* instance will have variables $y_1, \ldots, y_n; y'_1, \ldots, y'_n$. We define a partition $S \cup T$ where $S = \{y_i | i = 1, \ldots, n\}$ and $T = \{y'_j | j = 1, \ldots, n\}$. For each equation over the reals $\sum_i a_i x_i = b$ in an instance of MAX-LIN$_{\mathbb{R}_+}$, we construct the I-SAT* clauses:

$$\left(\begin{array}{c}\sum_{i\in S} a_i y_i \leq b \\ \text{OR} \\ \sum_{j\in T} a_j y'_j \leq b\end{array}\right), \left(\begin{array}{c}\sum_{i\in S} a_i y_i \geq b \\ \text{AND} \\ \sum_{j\in T} a_j y'_j \geq b\end{array}\right) \tag{11}$$

Let $x_1, \ldots, x_n$ be a solution for MAX-LIN$_{\mathbb{Z}_+}$. We set $y_i = y'_i = x_i$ for all $i = 1, \ldots, n$. A solution satisfying at least a $(1-\epsilon)$ fraction of the equations in the MAX-LIN$_{\mathbb{R}_+}$ instance will satisfy at least a $(1-\epsilon)$ fraction of the inequality-clauses in the I-SAT* instance. Conversely, suppose there exists a solution satisfying at least a $(1/2+76\delta)$ fraction of the inequality-clauses. Then there is at least a 38δ fraction of inequality-clause pairs of type (11) which must be satisfied. For each such pair, the type 2 clause implies that both $(\sum_{i\in S} a_i y_i \geq b)$ and $(\sum_{j\in T} a_j y'_j \geq b)$ and the type 1 clause implies that one of $\sum_{i\in S} a_i y_i \leq b$ and $\sum_{j\in T} a_j y'_j \geq b$ must hold. Therefore, for each such pair one of $\sum_{i\in S} a_i y_i = b$ or $\sum_{i\in T} a_i y'_i = b$ must hold. Setting x_i to be y_i for all i if there are more pairs for which $\sum_{i\in S} a_i y_i \leq b$ and y'_i otherwise, guarantees that at least a 19δ fraction of equations $\sum_i a_i x_i = b$ must be satisfied in the original MAX-LIN$_{\mathbb{Z}_+}$ instance thereby giving us the necessary gap reduction. This completes the proof of the theorem. □

Theorems 5.16 and 5.17 give us the corollary below that immediately yields the intended $(1/2 + \delta)$ hardness result:

Corollary 5.18. *For all $\epsilon, \delta > 0$ the problem* MAX-STABLE-VALUES$(1 - \epsilon, 1/2 + \delta)$ *is NP-hard.*

Proof. We argue that the reduction from I-SAT* shown in the proof of Theorem 4.2 is also a gap-preserving reduction and reduce from I-SAT$^*(1 - \epsilon, 1/2 + \delta)$. Arguing first for $(1 - \epsilon)$ completeness, we note that if a clause in the I-SAT* instance is satisfied then the corresponding pairwise stability condition is also satisfied.

Suppose now, that $(1/2 + \delta)$ fraction of the stability conditions for the active pairs are satisfied. But each such condition exactly corresponds to a clause being satisfied in the original I-SAT* instance thereby giving us the required $(1/2 + \delta)$ soundness for STABLE-VALUES. □

Acknowledgments

We thank Prasad Raghavendra for a helpful conversation, and the FOCS referees for their comments.

References

[Afr67] S.N. Afriat. The construction of utility functions from expenditure data. *International Economic Review*, 8(1):67–77, 1967.

[ALM+98] S. Arora, C. Lund, R. Motwani, M. Sudan, and M. Szegedy. Proof verification and the hardness of approximation problems. *Journal of the ACM*, 45(3):501–555, 1998.

[BV06] E. Beigman and R. Vohra. Learning from revealed preference. In *ACM Conference on Electronic Commerce*, pages 36–42, 2006.

[Die73] E. Diewert. Afriat and revealed preference theory. *Review of Economic Studies*, 40(3):419–425, 1973.

[DJ03] B. Dutta and M.O. Jackson. *Networks and groups: Models of strategic formation.* Studies in Economic Design. Springer-Verlag, 2003.

[DM97] B. Dutta and S. Mutuswami. Stable networks. *Journal of Economic Theory*, 76:322–344, 1997.

[Ech08] F. Echenique. What matchings can be stable? The testable implications of matching theory. *Mathematics of Operations Research*, 33(3):757–768, 2008.

[FLM+03] A. Fabrikant, A. Luthra, E. Maneva, C.H. Papadimitriou, and S. Shenker. On a network creation game. In *Proceedings of Principles of Distributed Computing, 22nd ACM Symposium (PODC)*, pages 347–351, 2003.

[FST04] A. Fostel, H.E. Scarf, and M.J. Todd. Two new proofs of Afriat's theorem. *Economic Theory*, 24(1):211–219, 2004.

[GR07] V. Guruswami and P. Raghavendra. A 3-query PCP over integers. In David S. Johnson and Uriel Feige, editors, *STOC*, pages 198–206. ACM, 2007.

[Hås01] J. Håstad. Some optimal inapproximability results. *J. ACM*, 48(4):798–859, 2001.

[HMC08] D. Hochbaum and E. Moreno-Centeno. The inequality-satisfiability problem. *Operations Research Letters*, 36(2):229–233, 2008.

[JW96] M.O. Jackson and A. Wolinsky. A strategic model of social and economic networks. *Journal of Economic Theory*, 71(1):44–74, October 1996.

[JW01] M.O. Jackson and A. Watts. The existence of pairwise stable networks. *Seoul Journal of Economics*, 14(3):299–321, 2001.

[KU08] S. Kalyanaraman and C. Umans. The complexity of rationalizing matchings. In *Proceedings of Algorithms and Computation, 19th International Symposium (ISAAC)*, pages 171–182, 2008.

[Raz98] R. Raz. A parallel repetition theorem. *SIAM J. Comput.*, 27(3):763–803, 1998.

[Sam48] P.A. Samuelson. Consumption theory in terms of revealed preference. *Economica*, 15(60):243–253, 1948.

[Spr00] Y. Sprumont. On the testable implications of collective choice theories. *Journal of Economic Theory*, 93:205–232, 2000.

[Var82] H.R. Varian. The nonparametric approach to demand analysis. *Econometrica*, 50(4):945–973, 1982.

[Var06] H.R. Varian. Revealed preference. In M. Szenberg, L. Ramrattan, and A.A. Gottesman, editors, *Samuelson Economics and the Twenty-First Century*, chapter 6, pages 99–115. Oxford University Press, 2006.

ECCC ISSN 1433-8092

http://eccc.hpi-web.de