APPROXIMATION RESISTANT PREDICATES FROM PAIRWISE INDEPENDENCE

Per Austrin and Elchanan Mossel

Abstract. We study the approximability of predicates on k variables from a domain $[q]$, and give a new sufficient condition for such predicates to be approximation resistant under the Unique Games Conjecture. Specifically, we show that a predicate P is approximation resistant if there exists a balanced pairwise independent distribution over $[q]^k$ whose support is contained in the set of satisfying assignments to P.

Using constructions of pairwise independent distributions this result implies that

- For general $k \geq 3$ and $q \geq 2$, the Max k-CSP$_q$ problem is UG-hard to approximate within $\mathcal{O}(kq^2)/q^k + \epsilon$.
- For the special case of $q = 2$, i.e., boolean variables, we can sharpen this bound to $(k + \mathcal{O}(k^{0.525}))/2^k + \epsilon$, improving upon the best previous bound of $2k/2^k + \epsilon$ (Samorodnitsky and Trevisan, STOC'06) by essentially a factor 2.
- Finally, again for $q = 2$, assuming that the famous Hadamard Conjecture is true, this can be improved even further, and the $\mathcal{O}(k^{0.525})$ term can be replaced by the constant 4.

Keywords. Approximation Resistance, Constraint Satisfaction, Unique Games Conjecture

Subject classification. MSC Primary 68Q17; Secondary 41A52

1. Introduction

In the Max k-CSP problem, we are given a set of constraints over a set of boolean variables, each constraint being a boolean function acting on at most k of the variables. The objective is to find an assignment to the variables satisfying as many of the constraints as possible. This problem is NP-hard for

any $k \geq 2$, and as a consequence, a lot of research has been focused on studying how well the problem can be approximated. We say that a (randomized) algorithm has *approximation ratio* α if, for all instances, the algorithm is guaranteed to find an assignment which (in expectation) satisfies at least $\alpha \cdot \mathrm{Opt}$ of the constraints, where Opt is the maximum number of simultaneously satisfied constraints, in any assignment.

A particularly simple approximation algorithm is the algorithm which simply picks a random assignment to the variables. This algorithm has a ratio of $1/2^k$. It was first improved by Trevisan (1998) who gave an algorithm with ratio $2/2^k$ for Max k-CSP. Recently, Hast (2005a) gave an algorithm with ratio $\Omega(k/(2^k \log k))$, which was subsequently improved by Charikar *et al.* (2007) who gave an algorithm with approximation ratio $c \cdot k/2^k$, where $c > 0.44$ is an absolute constant.

The PCP Theorem implies that the Max k-CSP problem is NP-hard to approximate within $1/c^k$ for some constant $c > 1$. Samorodnitsky & Trevisan (2000) improved this hardness to $2^{2\sqrt{k}}/2^k$, and this was further improved to $2^{\sqrt{2k}}/2^k$ by Engebretsen & Holmerin (2005). Finally, Samorodnitsky & Trevisan (2006) proved that, if the Unique Games Conjecture (Khot 2002) is true, then the Max k-CSP problem is hard to approximate within $2k/2^k$. To be more precise, the hardness they obtained was $2^{\lceil \log_2 k+1 \rceil}/2^k$, which is $(k+1)/2^k$ for $k = 2^r - 1$, but can be as large as $2k/2^k$ for general k. Thus, the current gap between hardness and approximability is a small constant factor of $2/0.44$.

For a predicate $P : \{0,1\}^k \to \{0,1\}$, the Max CSP($P$) problem is the special case of Max k-CSP in which all constraints are of the form $P(l_1, \ldots, l_k)$, where each literal l_i is either a variable or a negated variable. For this problem, the random assignment algorithm achieves a ratio of $m/2^k$, where m is the number of satisfying assignments of P. Surprisingly, it turns out that for certain choices of P, this is the best possible algorithm. In a celebrated result, Håstad (2001) showed that for $P(x_1, x_2, x_3) = x_1 \oplus x_2 \oplus x_3$, the Max CSP(P) problem is hard to approximate within $1/2 + \epsilon$.

Predicates P for which it is hard to approximate the Max CSP(P) problem better than a random assignment, are called *approximation resistant*. A slightly stronger notion is that of *hereditary* approximation resistance – a predicate P is hereditary approximation resistant if all predicates implied by P are approximation resistant. A natural and important question is to understand the structure of approximation resistance. For $k = 2$ and $k = 3$, this question is resolved – predicates on 2 variables are never approximation resistant, and a predicate on 3 variables is approximation resistant if and only if it is implied by an XOR of the three variables (Håstad 2001; Zwick 1998). For

$k = 4$, Hast (2005b) managed to classify most of the predicates with respect to to approximation resistance, but for this case there does not appear to be as nice a characterization as in the case $k = 3$. It turns out that, assuming the Unique Games Conjecture, most predicates are in fact hereditary approximation resistant – as k grows, the fraction of such predicates tend to 1 (Håstad 2007). Thus, one possible approach to cope with the seemingly complicated structure of approximation resistance, is to instead characterize the possibly simpler structure of hereditarily resistant predicates, since these constitute the vast majority of all resistant predicates.

Additionally, approximation resistance is useful for proving inapproximability results for MAX k-CSP in general – a natural approach for obtaining such inapproximability is to search for approximation resistant predicates with very few accepting inputs. This is indeed how all mentioned hardness results for MAX k-CSP come about (except the one implied by the PCP Theorem).

It is natural to generalize the MAX k-CSP problem to variables over a domain of size q, rather than just boolean variables. Without loss of generality we may assume that the domain is $[q]$. We call this the MAX k-CSP$_q$ problem. For MAX k-CSP$_q$, the random assignment gives a $1/q^k$-approximation. By an observation of Yury Makarychev, the algorithm of Charikar et al. for $q = 2$ can be used to obtain a $0.44k\lfloor \log_2 q \rfloor/q^k$-approximation for general q (see Corollary B.2). The best previous inapproximability for the MAX k-CSP$_q$ problem is due to Engebretsen (2004), who showed that the problem is NP-hard to approximate within $q^{\mathcal{O}(\sqrt{k})}/q^k$.

Similarly to $q = 2$, we can define the MAX CSP(P) problem for $P : [q]^k \to \{0, 1\}$. Here, there are several natural ways of generalizing the notion of a literal. One possible definition is to say that a literal l is of the form $\pi(x_i)$, for some variable x_i and permutation $\pi : [q] \to [q]$. A stricter definition is to say that a literal is of the form $x_i + a$, where, again, x_i is a variable, and $a \in [q]$ is some constant. In this paper, we use the second, stricter, definition. As this is a special case of the first definition, our hardness results apply also to the first definition.

A further motivation for studying the approximability of the MAX k-CSP$_q$ problem is that it is closely related to the optimum soundness-query tradeoff of probabilistically checkable proof systems. To be specific, let naPCP$_{c,s}[r, q, d]$ denote the set of languages for which there is a probabilistically checkable proof system in which the verifier has completeness c and soundness s, uses r bits of randomness and makes q queries from a proof with symbols from an alphabet of size d. Then if the MAX q-CSP$_d$ problem is NP-hard to approximate within

a factor α, it holds that

$$\mathrm{NP} \subseteq \mathrm{naPCP}_{c,\alpha c}[\mathcal{O}(\log n), k, q],$$

for some $c \in [0,1]$. For instance, Theorem 1.3 below shows that if the UGC is true, one can take $c = 1 - \epsilon$ and $s = \alpha c = (1 + o(1))q^2 k/q^k$, where the $o(1)$ term refers to something which tends to 0 as $q \to \infty$ (the fact that one can take $c = 1 - \epsilon$ does not follow from Theorem 1.3 as stated, but follows from the fact that the theorem is derived from an approximation resistant predicate).

1.1. Our contributions. Our main result is the following:

THEOREM 1.1. *Let $P : [q]^k \to \{0,1\}$ be a k-ary predicate over $[q]$, and let μ be a distribution over $[q]^k$ such that*

$$\Pr_{x \in ([q]^k, \mu)}[P(x) = 1] = 1$$

and for all $1 \leq i \neq j \leq k$ and all $a, b \in [q]$, it holds that

$$\Pr_{x \in ([q]^k, \mu)}[x_i = a, x_j = b] = 1/q^2.$$

Then, for any $\epsilon > 0$, the UGC implies that the MAX CSP(P) *problem is NP-hard to approximate within*

$$\frac{|P^{-1}(1)|}{q^k} + \epsilon,$$

i.e., P is hereditary approximation resistant.

Using constructions of pairwise independent distributions, we obtain the following corollaries:

THEOREM 1.2. *For any $k \geq 3$, $q = p^e$ for some prime p, and $\epsilon > 0$, it is UG-hard to approximate the* MAX k-CSP$_q$ *problem within*

$$\frac{k(q-1)q}{q^k} + \epsilon.$$

In the special case that $k = (q^r - 1)/(q - 1)$ for some r, the hardness ratio improves to

$$\frac{k(q-1)+1}{q^k} + \epsilon < \frac{kq}{q^k} + \epsilon.$$

Using an observation due to Yury Makarychev (see Corollary B.3), this gives essentially the same hardness bounds for arbitrary q.

THEOREM 1.3. *For any $k \geq 3$, $q \geq 2$, and $\epsilon > 0$, it is UG-hard to approximate the* MAX k-CSP$_q$ *problem within*

$$kq^2(1 + o(1))/q^k + \epsilon$$

These results constitute a significant improvement upon the hardness by Engebretsen (2004) of $q^{\mathcal{O}(\sqrt{k})}/q^k$. However, they do not improve upon the results of Samorodnitsky & Trevisan (2006) for the case of $q = 2$. In this case, we obtain the following theorem.

THEOREM 1.4. *For any $k \geq 3$ and $\epsilon > 0$, it is UG-hard to approximate the* MAX k-CSP *problem within*

$$\frac{k + \mathcal{O}(k^{0.525})}{2^k} + \epsilon.$$

If the Hadamard Conjecture is true, it is UG-hard to approximate the MAX k-CSP *problem within*

$$\frac{4\lceil (k+1)/4 \rceil}{2^k} + \epsilon \leq \frac{k+4}{2^k} + \epsilon$$

Thus, we improve the hardness of Samorodnitsky & Trevisan (2006) by essentially a factor 2, decreasing the gap to the best algorithm from roughly 2/0.44 to 1/0.44.

1.2. Related work. It is interesting to compare our results to the results of Samorodnitsky & Trevisan (2006). Recall that using the Gowers norm, they prove that the MAX k-CSP problem has a hardness factor of $2^{\lceil \log_2 k+1 \rceil}/2^k$, which is $(k+1)/2^k$ for $k = 2^r - 1$, but can be as large as $2k/2^k$ for general k.

Our proof uses the same version of the UGC, but the analysis is more direct and general. The proof of Samorodnitsky & Trevisan (2006) requires us to work specifically with a hypergraph linearity test for the long codes. For this test, the success probability is shown to be closely related to the Gowers inner product of the long codes. In particular, in the soundness analysis it is shown that if the value of this test is too large, it follows that the Gowers norm is larger than for "random functions". From this it is shown that at least two of the functions have large influences which in turns allows us to obtain a good solution to the Unique Games instance.

The work of Samorodnitsky & Trevisan (2006) builds upon earlier work of Samorodnitsky & Trevisan (2000), and Håstad & Wigderson (2001), which analyse graph linearity tests (rather than hypergraph tests).

Our construction on the other hand allows any pairwise distribution to define a long code test. Using Mossel (2007) we show that if a collection of supposed long codes does better than random for this long code test, then at least two of them have coordinates with large influences.

Our approach has a number of advantages. From the quantitative point of view, we improve existing hardness results for Max k-CSP$_q$, even in the already thoroughly explored $q = 2$ case. These improvements may seem very small, being an improvement only by a constant multiplicative factor of essentially 2. However, as discussed in Section 5, the approach of getting inapproximability from approximation resistant predicates can at best give a hardness of $2\lceil (k+1)/2 \rceil / 2^k$, and thus, in this respect, our hardness of $4\lceil (k+1)/4 \rceil / 2^k$ assuming the Hadamard Conjecture is optimal to within an *additive* $2/2^k$ for $k \equiv 0, 1 \pmod 4$ and exactly optimal for $k \equiv 2, 3 \pmod 4$. Also, our results give approximation resistance of Max CSP(P) for a much larger variety of predicates (any P containing a balanced pairwise independent distribution).

From a qualitative point of view, our analysis is very direct, and general enough to accomodate any domain $[q]$ with virtually no extra effort. Also, our proof uses bounds on expectations of products under certain types of correlation, putting it in the same general framework as many other UGC-based hardness results, in particular those for 2-CSPs (Austrin 2007a,b; Khot *et al.* 2007; Khot & O'Donnell 2006; O'Donnell & Wu 2008; Raghavendra 2008).

Parallel to this work, Guruswami & Raghavendra (2008) extended the "Gowers-based" approach of Samorodnitsky & Trevisan (2006) to larger values of q. For a prime q, the inapproximability they obtain is kq^2/q^k, which is slightly weaker than our bound of $kq(q-1)/q^k$ (which holds for any prime power). As in the $q = 2$ case, our analysis is also simpler and more direct than that of Guruswami & Raghavendra (2008).

Also independently of this work, Raghavendra (2008) obtained very general hardness results for Max CSP(P). He proved that if a natural SDP relaxation of Max CSP(P) has a certain integrality gap, then it is UG-hard to approximate Max CSP(P) within a factor better than that gap. We remark that, since we do not have any good integrality gaps for Max CSP(P), his results can not (currently) be used to obtain the results of this paper.

In a subsequent work by Austrin & Håstad (2009), the results of this paper were used to derive strong lower bounds on the probability that a random predicate is resistant. To be specific, it was shown that a random predicate on k variables, accepting $c(q) \cdot k^2$ inputs, supports a pairwise independent distribution with high probability (and hence is approximation resistant assuming the UGC). The previous bounds on random approximation resistant predicates by

Håstad (2007) only showed that a (boolean) predicate with $\Theta(2^t/\sqrt{t})$ accepting inputs is resistant (also under the UGC). This previous result uses the hardness result of Samorodnitsky & Trevisan (2006), and it is shown that $\Theta(2^t/\sqrt{t})$ is the best that can be hoped for by that method. This further demonstrates the strength of this paper: predicates supporting pairwise independence are a lot more common than the predicates obtained by the result of Samorodnitsky & Trevisan (2006).

2. Definitions

2.1. Unique Games. Originally, we proved our results under a variant of the Unique Games Conjecture that we call the (t, k)-UGC, which we could show was implied by the UGC. It was our hope that the (t, k)-UGC could be slightly less difficult to prove than the UGC, or that it could be the case that the (t, k)-UGC was true even in the event that the UGC turns out to be false.

However, using the recent strong parallel repetition result of Rao (2008), it turns out that the (t, k)-UGC is equivalent to the UGC. Nevertheless, we still give the details of the (t, k)-UGC and its equivalence with the UGC, as this may be of independent interest.

Definition 2.1. *An instance of the* k-ary Unique Label Cover *problem is a* k*-uniform hypergraph where for each hyperedge* $(v_1, \ldots, v_k)$ *there are* k *permutations* $\pi_1, \ldots, \pi_k$ *on* $[L]$.

We say that a hyperedge $(v_1, \ldots, v_k)$ with permutations $\pi_1, \ldots, \pi_k$ is t-wise satisfied by a labelling $\ell : V \to [L]$ if there are $i_1 < i_2 < \ldots < i_t$ such that $\pi_{i_1}(\ell(v_{i_1})) = \pi_{i_2}(\ell(v_{i_2})) = \ldots = \pi_{i_t}(\ell(v_{i_t}))$. We say that a hyperedge is completely satisfied by a labelling if it is k-wise satisfied.

We denote by $\mathrm{Opt}_t(\Psi) \in [0, 1]$ the maximum fraction of t-wise satisfied hyperedges, over any labelling. Note that $\mathrm{Opt}_{t+1}(\Psi) \leq \mathrm{Opt}_t(\Psi)$.

Conjecture 2.2. *For any* $2 \leq t \leq k$, *and* $\delta > 0$, *there exists an* $L > 0$ *such that it is NP-hard to distinguish between* k*-ary Unique Label Cover instances* Ψ *with label set* $[L]$ *with* $\mathrm{Opt}_k(\Psi) \geq 1 - \delta$, *and* $\mathrm{Opt}_t(\Psi) \leq \delta$.

For particular values of t and k we will refer to the corresponding special case of the above conjecture as the (t, k)-*Unique Games Conjecture* (or the (t, k)-UGC). In Appendix A we prove that the (t, k)-UGC is equivalent to the $(2, 2)$-UGC for all $2 \leq t \leq k$.

Khot's original formulation of the Unique Games Conjecture (Khot 2002) is then exactly the $(2,2)$-UGC. Khot & Regev (2003) proved that this conjecture is equivalent to the $(2,k)$-UGC for all k, which is what Samorodnitsky & Trevisan (2006) used to obtain hardness for Max k-CSP. The methods of this paper can be used to show that any predicate supporting a t-wise independent distribution is approximation resistant under the $(t+1,k)$-UGC. As these conjectures are all equivalent, the presentation only deals with the case of pairwise independence, and uses the $(2,k)$-UGC to simplify the presentation.

2.2. Influences. It is well known (see e.g. Khot *et al.* (2007)) that each function $f : [q]^n \to \mathbb{R}$ admits a unique *Efron-Stein decomposition:*

$$f = \sum_{S \subseteq [n]} f_S$$

where

- The function f_S depends on $x_S = (x_i : i \in S)$ only.
- For every $S \not\subseteq S'$, and every $y_{S'} \in [q]^{S'}$ it holds that

$$\mathbb{E}[f_S(x_S) | x_{S'} = y_{S'}] = 0.$$

For $d \leq n$ we write $f^{\leq d} = \sum_{S:|S|\leq d} f_S$ for the degree d part of f. We now define the *influence of the ith coordinate on f*, denoted by $\mathrm{Inf}_i(f)$ by

$$\mathrm{Inf}_i(f) = \mathop{\mathbb{E}}_{x}[\mathop{\mathrm{Var}}_{x_i}[f(x)]].$$

We define the *degree m influence of the ith coordinate on f*, denoted by $\mathrm{Inf}_i^{\leq d}(f)$ by $\mathrm{Inf}_i(f^{\leq d})$.

Recall that the influence $\mathrm{Inf}_i(f)$ measures how much the function f depends on the i'th variable, while the low degree influence $\mathrm{Inf}_i^{\leq d}(f)$ measures this for the low degree part of the part of f. The later quantity is closely related to the influence of a "smoothed" version of f.

An important property of low-degree influences is that

$$\sum_{i=1}^{n} \mathrm{Inf}_i^{\leq d}(f) \leq d \,\mathrm{Var}[f],$$

implying that the number of coordinates with large low-degree influence must be small. In particular, we have the following fact.

FACT 2.3. *For $f : [q]^n \to [0,1]$, the the number of coordinates with degree d influence at least τ is at most d/τ.*

2.3. Correlated Probability Spaces. We will be interested in probability distributions supported on $P^{-1}(1) \subseteq [q]^k$, where P is some predicate defining a MAX CSP(P) problem.

It would be useful to follow Mossel (2007) and view $[q]^k$ with such a probability measure as a collection of k *correlated spaces* corresponding to the k coordinates. We proceed with formal definitions of two and k correlated spaces.

We use $\mathrm{Cov}[X,Y] = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$ to denote the covariance between two random variables X and Y, and $\mathrm{Var}[X] = \mathrm{Cov}[X,X]$ to denote the variance of a random variable.

DEFINITION 2.4. *Let (Ω,μ) be a probability space over a finite product space $\Omega = \Omega_1 \times \Omega_2$. The* correlation *between Ω_1 and Ω_2 (with respect to μ) is*

$$\rho(\Omega_1,\Omega_2;\mu) = \sup\{\,\mathrm{Cov}[f_1(x_1)f_2(x_2)] \;:\; f_i : \Omega_i \to \mathbb{R}, \mathrm{Var}[f_i(x_i)] = 1\,\},$$

where (x_1,x_2) is drawn from (Ω,μ).

DEFINITION 2.5. *Let (Ω,μ) be a probability space over a finite product space $\Omega = \prod_{i=1}^k \Omega_i$. The correlation of $\Omega_1,\ldots,\Omega_k$ (with respect to μ) is*

$$\rho(\Omega_1,\ldots,\Omega_k;\mu) = \max_{1\le i\le k} \rho\left(\Omega_i, \prod_{j\neq i}\Omega_j;\mu\right)$$

Of particular interest to us is the case where correlated spaces are defined by a measure that is t-wise independent.

DEFINITION 2.6. *Let (Ω,μ) be a probability space over a product space $\Omega = \prod_{i=1}^k \Omega_i$. We say that μ is t-wise independent if, for any choice of $i_1 < i_2 < \ldots < i_t$ and $b_1,\ldots,b_t$ with $b_j \in \Omega_{i_j}$, we have that*

$$\Pr_{w\in(\Omega,\mu)}[w_{i_1} = b_1,\ldots,w_{i_t} = b_t] = \prod_{j=1}^{t} \Pr_{w\in(\Omega,\mu)}[w_{i_j} = b_j]$$

We say that (Ω,μ) is balanced *if for every $i \in [k], b \in \Omega_i$, we have that $\Pr_{w\in(\Omega,\mu)}[w_i = b] = 1/|\Omega_i|$.*

The following theorem considers low influence functions that act on correlated spaces where the correlation is given by a t-wise independent probability measure for $t \geq 2$. It shows that in this case, the outputs of the functions have almost the same distribution as if the inputs were completely independent. Moreover, the result holds even if some of the functions have large influences as long as in each coordinate not more than t functions have large influences.

THEOREM 2.7 (Mossel 2007, Theorem 6.6 and Lemma 6.9). *Let (Ω, μ) be a finite probability space over $\Omega = \prod_{i=1}^{k} \Omega_i$ with the following properties:*

(a) μ is t-wise independent.

(b) For all $i \in [k]$ and $b_i \in \Omega_i$, $\mu_i(b_i) > 0$.

(c) $\rho(\Omega_1, \ldots, \Omega_k; \mu) < 1$.

Then for all $\epsilon > 0$ there exists a $\tau > 0$ and $d > 0$ such that the following holds for all n. Let $f_1, \ldots, f_k$ be functions $f_i : \Omega_i^n \to [0, 1]$ satisfying that, for all $1 \leq j \leq n$,

$$|\{\, i \,:\, \mathrm{Inf}_j^{\leq d}(f_i) \geq \tau \,\}| \leq t.$$

Then

$$\left| \mathop{\mathbb{E}}_{w_1,\ldots,w_n} \left[\prod_{i=1}^{k} f_i(w_{1,i}, \ldots, w_{n,i}) \right] - \prod_{i=1}^{k} \mathop{\mathbb{E}}_{w_1,\ldots,w_n} [f_i(w_{1,i}, \ldots, w_{n,i})] \right| \leq \epsilon,$$

where $w_1, \ldots, w_n$ are drawn independently from (Ω, μ), and $w_{i,j} \in \Omega_j$ denotes the jth coordinate of w_i.

Roughly speaking, the basic idea behind the theorem and its proof is that low influence functions cannot detect dependencies of high order – in particular if the underlying measure is pairwise independent, then low influence functions of different coordinates are essentially independent.

Since condition (c) of the theorem is somewhat inconvenient to work with, we will instead use the following much simpler condition, which is a special case of Lemma 2.9 in Mossel (2007).

FACT 2.8. *A sufficient condition for (c) to hold in Theorem 2.7 is that for all $w \in \Omega$, $\mu(w) > 0$.*

3. Main theorem

In this section, we prove our main theorem. Note that it is a generalization of Theorem 1.1.

THEOREM 3.1. *Let $P : [q]^k \to \{0,1\}$ be a k-ary predicate over a (finite) domain of size q, and let μ be a balanced pairwise independent distribution over $[q]^k$ such that $\Pr_{x\in([q]^k,\mu)}[P(x) = 1] > 0$. Then, for any $\epsilon > 0$, the UGC implies that the* MAX CSP(P) *problem is NP-hard to approximate within*

$$\frac{|P^{-1}(1)|}{q^k \cdot \Pr_{x\in([q]^k,\mu)}[P(x) = 1]} + \epsilon$$

In particular, note that if $\Pr_{x\in([q]^k,\mu)}[P(x) = 1] = 1$, i.e., if the support of μ is entirely contained in the set of satisfying assignments to P, then P is approximation resistant. It is also hereditary approximation resistant, since the support of μ will still be contained in $P^{-1}(1)$ when we add more satisfying assignments to P.

Reduction. We will construct a probabilistically checkable proof system for Unique Label Cover. Given a k-ary Unique Label Cover instance Ψ, the prover writes down the table of a function $f_v : [q]^L \to [q]$ for each v, which is supposed to be the dictator function $f_v(x) = x_{\ell(v)}$ corresponding to the label $\ell(v)$ of vertex v. Furthermore, we will assume that f_v is folded, i.e., that for every $x \in [q]^k$ and $a \in [q]$, we have

$$f_v(x + (a, \ldots, a)) = f_v(x) + a$$

(where the definition of "+" in $[q]$ is arbitrary as long as $([q], +)$ is an Abelian group). When reading the value of $f_v(x_1, \ldots, x_L)$, the verifier can enforce this condition by instead querying $f_v(x_1 - x_1, x_2 - x_1, \ldots, x_L - x_1)$ and adding x_1 to the result. Let $\eta > 0$ be a parameter, the value of which will be determined later. Define a probability distribution μ' on $[q]^k$ by

$$\mu'(w) = (1 - \eta) \cdot \mu(w) + \eta \cdot \mu_U(w),$$

where μ_U is the uniform distribution on $[q]^k$, i.e., $\mu_U(w) = 1/q^k$. Finally, for $f : [q]^L \to \mathbb{R}$ and a permutation $\pi : [L] \to [L]$, we define $f\pi : [q]^L \to \mathbb{R}$ by $f\pi(x) = f(x_{\pi(1)}, x_{\pi(2)}, \ldots, x_{\pi(L)})$. Given a proof $\Sigma = \{f_v\}_{v\in V}$ of supposed long codes for a good labelling of Ψ, the verifier checks Σ as follows.

Algorithm 1: The verifier $\mathcal{V}$
$\mathcal{V}(\Psi, \Sigma = \{f_v\}_{v \in V})$
(1) Pick a random hyperedge $e = (v_1, \ldots, v_k)$ with permutations $\pi_1, \ldots, \pi_k$.
(2) For each $i \in [L]$, draw w_i randomly from $([q]^k, \mu')$.
(3) For each $j \in [k]$, let $x_j = w_{1,j} \ldots w_{L,j}$, and let $b_j = f_{v_j}\pi_j(x_j)$.
(4) Accept if $P(b_1, \ldots, b_k) = 1$.

LEMMA 3.2 (Completeness). *For any δ, if $\mathrm{Opt}_k(\Psi) \geq 1 - \delta$, then there is a proof Σ such that*

$$\Pr[\mathcal{V}(\Psi, \Sigma) \text{ accepts}] \geq (1 - \delta)(1 - \eta) \Pr_{w \in ([q]^k, \mu)}[P(w) = 1]$$

PROOF. Take a labelling ℓ for Ψ such that a fraction $\geq 1 - \delta$ of the hyperedges are k-wise satisfied, and let $f_v : [q]^L \to [q]$ be the long code of the label $\ell(v)$ of vertex v.

Let $(v_1, \ldots, v_k)$ be an hyperedge that is k-wise satisfied by ℓ. Then for each $j \in [k]$, $f_{v_j}\pi_j$ is the dictator function return the $\pi_j(\ell(v_j))$'th coordinate. Hence since the hyperedge is k-wise satisfied it holds that $f_{v_1}\pi_1 = f_{v_2}\pi_2 = \ldots = f_{v_k}\pi_k$, each being the dictator function returning the i'th input for some $i \in [L]$. The probability that $\mathcal{V}$ accepts is then exactly the probability that $P(w_i)$ is true. Since w_i is drawn from $([q]^k, \mu)$ with probability $1 - \eta$, $P(w_i)$ is true with probability at least $(1 - \eta) \Pr_{w \in ([q]^k, \mu)}[P(w) = 1]$.

The probability that the hyperedge e chosen by the verifier in step 1 is satisfied by ℓ is at least $1 - \delta$, and so we end up with the desired inequality. □

LEMMA 3.3 (Soundness). *For any $\epsilon > 0$, $\eta > 0$, there is a constant $\delta := \delta(\epsilon, \eta, k, q) > 0$, such that if $\mathrm{Opt}_2(\Psi) < \delta$, then for any proof Σ, we have*

$$\Pr[\mathcal{V}(\Psi, \Sigma) \text{ accepts}] \leq \frac{|P^{-1}(1)|}{q^k} + \epsilon$$

PROOF. Assume that

$$\Pr[\mathcal{V}(\Psi, \Sigma) \text{ accepts}] > \frac{|P^{-1}(1)|}{q^k} + \epsilon. \tag{3.4}$$

We need to prove that this implies that there is a $\delta := \delta(\epsilon, \eta, k, q) > 0$ such that $\mathrm{Opt}_2(\Psi) \geq \delta$.

Equation 3.4 implies that for a fraction of at least $\epsilon/2$ of the hyperedges e, the probability that $\mathcal{V}(\Psi, \Sigma)$ accepts when choosing e is at least $\frac{|P^{-1}(1)|}{q^k} + \epsilon/2$.

Let $e = (v_1, \ldots, v_k)$ with permutations $\pi_1, \ldots, \pi_k$ be such a "good" hyperedge. For $v \in V$ and $a \in [q]$, define $g_{v,a} : [q]^L \to \{0, 1\}$ by

$$g_{v,a}(x) = \begin{cases} 1 & \text{if } f_v(x) = a \\ 0 & \text{otherwise} \end{cases}.$$

The probability that $\mathcal{V}$ accepts when choosing e is then exactly

$$\sum_{a \in P^{-1}(1)} \mathop{\mathbb{E}}_{w_1,\ldots,w_L} \left[\prod_{i=1}^{k} g_{v_i,a_i} \pi_i(w_{1,i}, \ldots, w_{L,i}) \right],$$

which, by the choice of e, is greater than $|P^{-1}(1)|/q^k + \epsilon/2$. This implies that there is some $a \in P^{-1}(1)$ such that

$$\begin{aligned} \mathop{\mathbb{E}}_{w_1,\ldots,w_L} \left[\prod_{i=1}^{k} g_{v_i,a_i} \pi_i(w_{1,i}, \ldots, w_{L,i}) \right] &> 1/q^k + \epsilon' \\ = \prod_{i=1}^{k} \mathop{\mathbb{E}}_{w_1,\ldots,w_L} [g_{v_i,a_i} \pi_i(w_{1,i}, \ldots, w_{L,i})] + \epsilon', \end{aligned}$$

where $\epsilon' = \epsilon/(2|P^{-1}(1)|)$ and the last equality uses that, because f_{v_i} is folded and μ is balanced, we have $\mathbb{E}_{w_1,\ldots,w_L}[g_{v_i,a_i}(w_{1,i}, \ldots, w_{L,i})] = 1/q$.

Note that because both μ and μ_U are pairwise independent, μ' is also pairwise independent. Also, we have that for each $w \in [q]^k$, $\mu'(w) \geq \eta/q^k > 0$, which by Fact 2.8 implies both conditions (b) and (c) of Theorem 2.7. Then, the contrapositive formulation of Theorem 2.7 implies that there is an $i \in [L]$ and at least two indices $J \subseteq [k]$ such that $\mathrm{Inf}^{\leq d}_{\pi_j^{-1}(i)}(g_{v_j,a_j}) = \mathrm{Inf}^{\leq d}_{i}(g_{v_j,a_j}\pi_j) \geq \tau$ for all $j \in J$, where τ and d are functions of ϵ, η, t, k, and q.

The process of constructing a good labelling of Ψ from this point is standard. For each $v \in V$, let

$$C(v) = \{\, i \mid \mathrm{Inf}^{\leq d}_{i}(g_{v,a}) \geq \tau \text{ for some } a \in [q] \,\}.$$

Note that by Fact 2.3, $|C(v)| \leq q \cdot d/\tau$.

Define a labelling $\ell : V \to [L]$ by picking, for each $v \in V$, a label $\ell(v)$ uniformly at random from $C(v)$ (or an arbitrary label in case $C(v)$ is empty). Let $e = (v_1, \ldots, v_k)$ be one of the "good" hyperedges, so that there are at least

two indices $j \in [k]$ with values $a_j \in [q]$ such that $\mathrm{Inf}^{\leq d}_{\pi_j^{-1}(i)}(g_{v_j,a_j}) \geq \tau$. Then for all such j, $\pi_j^{-1}(i) \in C(v_j)$, and thus, the probability that $\pi_j(\ell(v_j)) = i$ is $1/|C(v_j)|$. This implies that the probability that this hyperedge is 2-wise satisfied is at least $\left(\frac{\tau}{d \cdot q}\right)^2$. Overall, the total expected fraction of hyperedges that are 2-wise satisfied by ℓ is at least $\delta = \epsilon \left(\frac{\tau}{d \cdot q}\right)^2$, and thus $\mathrm{Opt}_2(\Psi) \geq \delta$. □

It is now straightforward to prove Theorem 3.1.

Proof (of Theorem 3.1). Let $c = \Pr_{x \in ([q]^k, \mu)}[P(x) = 1]$, $s = |P^{-1}(1)|/q^k$, and $\eta = \min(1/4, \frac{\epsilon c}{8s})$. Note that since the statement of the Theorem requires $c > 0$ we also have $s > 0$ and $\eta > 0$. Assume that the $(2, k)$-UGC is true, and pick L large enough so that it is NP-hard to distinguish between k-ary Unique Label Cover instances Ψ with $\mathrm{Opt}_2(\Psi) \leq \delta$ and $\mathrm{Opt}_k(\Psi) \geq 1 - \delta$, where $\delta = \min(\eta, \delta(\epsilon c/4, \eta, k, q))$ and $\delta(\ldots)$ is the function from Lemma 3.3. By Lemmas 3.2 and 3.3, we then get that it is NP-hard to distinguish between Max CSP(P) instances with $\mathrm{Opt} \geq (1 - \delta)(1 - \eta)c \geq (1 - 2\eta)c$ and $\mathrm{Opt} \leq s + \epsilon c/4$. In other words, it is NP-hard to approximate the Max CSP(P) problem within a factor

$$\frac{s + \epsilon c/4}{(1 - 2\eta)c} \leq \frac{s(1 + 4\eta)}{c} + (1 + 4\eta)\epsilon/4 \leq s/c + \epsilon.$$

□

4. Inapproximability for Max k-CSP$_q$

As a simple corollary to Theorem 3.1, we have:

Corollary 4.1. *Let μ be a balanced pairwise independent distribution over $[q]^k$. Then the UGC implies that that* Max k-CSP$_q$ *problem is NP-hard to approximate within*

$$\frac{|\operatorname{Supp}(\mu)|}{q^k}$$

Thus, we have reduced the problem of obtaining strong inapproximability for Max k-CSP$_q$ to the problem of finding small pairwise independent distributions.

In the remainder of this section, we will focus on the details of standard constructions of pairwise independence, giving hardness for Max k-CSP$_q$ under the UGC.

4.1. Theorem 1.2. The pairwise independent distributions used to give Theorem 1.2 is based on the following simple lemma, which is well-known but stated here in a slightly more general form than usual:

LEMMA 4.2. *Let R be a finite commutative ring, and let $u, v \in R^n$ be two vectors over R such that $u_i v_j - u_j v_i \in R^*$ for some i, j.*[1] *Let $X \in R^n$ be a uniformly random vector over R^n and let μ be the probability distribution over R^2 of $(\langle u, X\rangle, \langle v, X\rangle) \in R^2$. Then μ is a balanced pairwise independent distribution.*

PROOF. Without loss of generality, assume that $i = 1$ and $j = 2$. It suffices to prove that, for all $(a, b) \in R^2$ and any choice of values of $X_3, \ldots, X_n$, we have

$$\Pr[(\langle u, X\rangle, \langle v, X\rangle) = (a, b) \mid X_3, \ldots, X_n] = 1/|R|^2.$$

For this to be true, we need that the system

$$\begin{cases} u_1 X_1 & + & u_2 X_2 & = & a' \\ v_1 X_1 & + & v_2 X_2 & = & b' \end{cases}$$

has exactly one solution, where $a' = a - \sum_{i=3}^n u_i X_i$ and $b' = b - \sum_{i=3}^n v_i X_i$. This in turn follows directly from the condition on u and v. □

Consequently, given a set of m vectors in R^n such that any pair of them satisfy the condition of Lemma 4.2, we can construct a pairwise independent distribution over R^m with support size $|R|^n$.

From this, it is easy to prove Theorem 1.2. The construction we use is essentially the same as that of O'Brien (1980).

PROOF (of Theorem 1.2). Define

$$r = \lceil \log_q(k(q-1)+1) \rceil \qquad\qquad n = (q^r - 1)/(q-1) \geq k.$$

Let $\mathbb{P}(\mathbb{F}_q^r)$ denote the projective space over $\mathbb{F}_q^r$, i.e.,

$$\mathbb{P}(\mathbb{F}_q^r) = (\mathbb{F}_q^r \setminus 0)/\!\sim.$$

Here $\sim$ is the equivalence relation defined by $(x_1, \ldots, x_r) \sim (y_1, \ldots, y_r)$ if there exists a $c \in \mathbb{F}_q^*$ such that $x_i = c y_i$ for all i, i.e., if $(x_1, \ldots, x_r)$ and $(y_1, \ldots, y_r)$ are linearly dependent. We then have

$$|\mathbb{P}(\mathbb{F}_q^r)| = (q^r - 1)/(q-1) = n.$$

[1] R^* denotes the set of units of R. In the case that R is a field, the condition is equivalent to saying that u and v are linearly independent.

Choose n vectors $u_1, \ldots, u_n \in \mathbb{F}_q^r$ as representatives from each of the equivalence classes of $\mathbb{P}(\mathbb{F}_q^r)$. Then any pair u_i, u_j satisfy the condition of Lemma 4.2. Thus, we have that $(\langle u_i, X\rangle)_{1\leq i\leq n}$ for a uniformly random $X \in \mathbb{F}_q^r$ induces a balanced pairwise independent distribution over $\mathbb{F}_q^n$ (and hence over $[q]^k$) with support size q^r.

When $k = (q^r - 1)/(q - 1)$, this gives a hardness of $\frac{k(q-1)+1}{q^k}$, and for general k, in particular

$$k = (q^{r-1} - 1)/(q - 1) + 1,$$

we lose a factor q in the hardness ratio. □

We remark that for $q = 2$ this construction gives exactly the predicate used by Samorodnitsky & Trevisan (2006), giving an inapproximability of $2k/2^k$ for all k, and $(k + 1)/2^k$ for all k of the form $2^l - 1$.

4.2. Theorem 1.4. Let us now look closer at the special case of boolean variables, i.e., $q = 2$. So far, we have only given a different proof of Samorodnitsky and Trevisan's result, but we will now show how to improve this.

An Hadamard matrix is an $n \times n$ matrix over ± 1 such that $HH^T = nI$, i.e., each pair of rows, and each pair of columns, are orthogonal. Let $h(n)$ denote the smallest $n' \geq n$ such that there exists an $n' \times n'$ Hadamard matrix. It is a well-known fact that Hadamard matrices give small pairwise independent distributions and thus give hardness of approximating Max k-CSP. To be specific, we have the following proposition:

Proposition 4.3. *For every $k \geq 3$, the UGC implies that the* Max k-CSP *problem is UG-hard to approximate within $h(k + 1)/2^k + \epsilon$.*

Proof. Let $n = h(k+1)$ and let A be an $n \times n$ Hadamard matrix, normalized so that one column contains only ones. Remove $n - k$ of the columns, including the all-ones column, and let A' be the resulting $n \times k$ matrix. Let $\mu : \{-1, 1\}^k \to [0, 1]$ be the probability distribution which assigns probability $1/n$ to each row of A'. Then μ is a balanced pairwise independent distribution with $|\operatorname{Supp}(\mu)| = h(k + 1)$. □

It is well known that Hadamard matrices can only exist for $n = 1$, $n = 2$, and $n \equiv 0 \pmod 4$. The famous *Hadamard Conjecture* asserts that Hadamard matrices exist for all n which are divisible by 4, in other words, that $h(n) = 4\lceil n/4\rceil \leq n + 3$. It is also possible to get useful unconditional bounds on $h(n)$. We now give one such easy bound.

THEOREM 4.4 (Paley 1933). *For every odd prime p and integers $e, f \geq 0$, there exists an $n \times n$ Hadamard matrix H_n where $n = 2^e(p^f + 1)$, whenever this number is divisible by 4.*

THEOREM 4.5 (Baker *et al.* 2001). *There exists an integer n_0 such that for every $n \geq n_0$, there is a prime p between n and $n + n^{0.525}$.*

COROLLARY 4.6. $h(n) \leq n + \mathcal{O}(n^{0.525})$.

PROOF. Let p be the smallest prime larger than $n/2$, and let $n' = 2(p + 1) \geq n$. Then, Theorem 4.4 asserts that there exists an $n' \times n'$ Hadamard matrix, so $h(n) \leq n'$. If n is sufficiently large ($n \geq 2n_0$), then by Theorem 4.5, $p \leq n/2 + (n/2)^{0.525}$ and $n' \leq n + 2n^{0.525}$, as desired. □

Theorem 1.4 follows from Proposition 4.3 and Corollary 4.6.

It is probably possible to get a stronger unconditional bound on $h(n)$ than the one given by Corollary 4.6, by using stronger construction techniques than the one of Theorem 4.4.

5. Discussion

We have given a strong sufficient condition for predicates to be hereditary approximation resistant under the Unique Games Conjecture: it suffices for the set of satisfying assignments to contain a balanced pairwise independent distribution. Using constructions of such distributions with small support, we were then able to construct approximation resistant predicates with few accepting inputs, which in turn gave improved hardness for the MAX k-CSP$_q$ problem.

There are several aspects here where there is room for interesting further work:

A very natural and interesting question is whether our condition is also necessary for a predicate to be hereditary approximation resistant, i.e., if pairwise independence gives a complete characterization of hereditary approximation resistance.

Finally, it is natural to ask whether our results for MAX k-CSP$_q$ can be pushed a bit further, or whether they are tight. For the case of boolean variables, Hast (2005b) proved that any predicate accepting at most $2\lfloor k/2 \rfloor + 1$ inputs is *not* approximation resistant. For $k \equiv 2, 3 \pmod 4$ this exactly matches the result we get under the UGC and the Hadamard Conjecture (which for

$k = 2^r - 1$ and $k = 2^r - 2$ is the same hardness as in Samorodnitsky & Trevisan (2006)). For $k \equiv 0, 1 \pmod 4$, there is an additive constant 2 between how few satisfying assignments an approximation resistant predicate can and cannot have.

Thus, the hitherto very succesful approach of obtaining hardness for Max k-CSP by finding "small" approximation resistant predicates, can not be taken further, but there is still a small constant gap of roughly 1/0.44 to the best current algorithm. It would be interesting to know whether the algorithm can be improved, or whether the hardest instances of Max k-CSP are not Max CSP(P) instances for some approximation resistant P.

For large q, there is still a significant gap of $\Theta(q^2/\log q)$ between the best algorithm and the best inapproximability. However, on the positive side, we note that this gap is independent of k, and depends only on q.

Acknowledgements

We would like to thank Irit Dinur, Johan Håstad, and Oded Regev for interesting discussions. We are very grateful to Yury Makarychev, for the nice observation in Appendix B allowing us to get essentially the same inapproximability for general q as for prime powers, rather than the weaker bound we originally had.

Per Austrin was funded by Swedish Research Council Project Number 50394001. Elchanan Mossel was supported by BSF grant 2004105, NSF CAREER award DMS 0548249 and DOD ONR grant N0014-07-1-05-06.

References

Per Austrin (2007a). Balanced Max 2-Sat Might Not be the Hardest. In *ACM Symposium on Theory of Computing (STOC)*, 189–197.

Per Austrin (2007b). Towards Sharp Inapproximability For Any 2-CSP. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, 307–317. ISSN 0272-5428.

Per Austrin & Johan Håstad (2009). Randomly Supported Independence and Resistance. To appear in STOC.

R. C. Baker, G. Harman & J. Pintz (2001). The Difference Between Consecutive Primes, II. *Proceedings of the London Mathematical Society* **83**(3), 532–562.

Moses Charikar, Konstantin Makarychev & Yury Makarychev (2007). Near-Optimal Algorithms for Maximum Constraint Satisfaction Problems. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 62–68.

Lars Engebretsen (2004). The Nonapproximability of Non-Boolean Predicates. *SIAM Journal on Discrete Mathematics* **18**(1), 114–129.

Lars Engebretsen & Jonas Holmerin (2005). More Efficient Queries in PCPs for NP and Improved Approximation Hardness of Maximum CSP. In *Symposium on Theoretical Aspects of Computer Science (STACS)*, 194–205.

Venkatesan Guruswami & Prasad Raghavendra (2008). Constraint Satisfaction over a Non-Boolean Domain: Approximation Algorithms and Unique-Games Hardness. In *APPROX-RANDOM*, 77–90.

Gustav Hast (2005a). Approximating Max kCSP – Outperforming a Random Assignment with Almost a Linear Factor. In *International Colloquium on Automata, Languages and Programming (ICALP)*, 956–968.

Gustav Hast (2005b). *Beating a Random Assignment – Approximating Constraint Satisfaction Problems*. Ph.D. thesis, KTH – Royal Institute of Technology.

Johan Håstad (2001). Some optimal inapproximability results. *Journal of the ACM* **48**(4), 798–859.

Johan Håstad (2007). On the approximation resistance of a random predicate. In *APPROX-RANDOM*, 149–163.

Johan Håstad & Avi Wigderson (2001). Simple analysis of graph tests. In *2001 Conference on Computational Complexity*, 244–255.

Subhash Khot (2002). On the power of unique 2-prover 1-round games. In *ACM Symposium on Theory of Computing (STOC)*, 767–775. ISBN 1-58113-495-9.

Subhash Khot, Guy Kindler, Elchanan Mossel & Ryan O'Donnell (2007). Optimal Inapproximability Results for MAX-CUT and Other 2-variable CSPs? *Siam Journal on Computing* **37**, 319–357.

Subhash Khot & Ryan O'Donnell (2006). SDP gaps and UGC-hardness for MAXCUTGAIN. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, 217–226.

Subhash Khot & Oded Regev (2003). Vertex Cover Might be Hard to Approximate to within $2-\epsilon$. In *IEEE Conference on Computational Complexity*, 379–.

Elchanan Mossel (2007). Gaussian bounds for noise correlation of functions. arXiv Report math/0703683v3.

G. L. O'Brien (1980). Pairwise Independent Random Variables. *Annals of Probability* **8**(1), 170–175.

Ryan O'Donnell & Yi Wu (2008). An optimal SDP algorithm for Max-Cut, and equally optimal Long Code tests. In *ACM Symposium on Theory of Computing (STOC)*, 335–344.

Raymond E. A. C. Paley (1933). On orthogonal matrices. *Journal of Mathematics and Physics* **12**, 311–320.

Prasad Raghavendra (2008). Optimal Algorithms and Inapproximability Results For Every CSP? In *ACM Symposium on Theory of Computing (STOC)*.

Anup Rao (2008). Parallel Repetition in Projection Games and a Concentration Bound. In *ACM Symposium on Theory of Computing (STOC)*.

Alex Samorodnitsky & Luca Trevisan (2000). A PCP characterization of NP with optimal amortized query complexity. In *ACM Symposium on Theory of Computing (STOC)*, 191–199.

Alex Samorodnitsky & Luca Trevisan (2006). Gowers uniformity, influence of variables, and PCPs. In *ACM Symposium on Theory of Computing (STOC)*, 11–20. ISBN 1-59593-134-1.

Luca Trevisan (1998). Parallel Approximation Algorithms by Positive Linear Programming. *Algorithmica* **21**, 72–88.

Uri Zwick (1998). Approximation Algorithms for Constraint Satisfaction Problems Involving at Most Three Variables Per Constraint. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)*.

A. Equivalence of the (t, k)-conjectures

In this section, we prove that for every $2 \leq t \leq k$, the (t, k)-UGC (as defined in Section 2.1) is equivalent to Khot's Unique Games Conjecture.

As mentioned in Section 2.1, Khot & Regev (2003) proved that Khot's original Unique Games Conjecture is equivalent to the $(2, k)$-UGC for all $k \geq 2$. Clearly, since $\mathrm{Opt}_{t+1}(\Psi) \leq \mathrm{Opt}_t(\Psi)$, the (t, k)-UGC implies the $(t+1, k)$-UGC, so it suffices to prove the following proposition.

Proposition A.1. *The (k, k)-UGC implies the $(2, 2)$-UGC.*

Proof. Given a k-ary Unique Label Cover instance Ψ, create a 2-ary Unique Label Cover instance Ψ' as follows. The vertices and labels of Ψ' are the same as those of Ψ. For each hyperedge $(v_1, \ldots, v_k)$ with permutations $\pi_1, \ldots, \pi_k$ in

Ψ, and each $1 \leq i \leq k$, add the edge (v_i, v_{i+1}) with permutations π_i, π_{i+1} to Ψ' (where we define $v_{k+1} = v_1$ and $\pi_{k+1} = \pi_1$).

The optimums of Ψ and Ψ' can be related to each other as follows:

$$\mathrm{Opt}_k(\Psi) \leq \mathrm{Opt}_2(\Psi') \leq 1 - \frac{2}{k}(1 - \mathrm{Opt}_k(\Psi))$$

The first inequality is trivial. To see that the second inequality holds, pick an arbitrary labelling $\ell : V \to [L]$, and let $(v_1, \ldots, v_k)$ be an hyperedge in Ψ which is not k-wise satisfied by ℓ. Then, there is a non-trivial partition of $v_1, \ldots, v_k$ such that $\pi_i(v_i) \neq \pi_j(v_j)$ whenever v_i and v_j are not in the same part. Now, in Ψ' there are k edges corresponding to the hyperedge $v_1, \ldots, v_k$. The total number of these edges which are not satisfied by ℓ is at least the number of edges cut by the partition of $v_1, \ldots, v_k$, which in turn is at least 2, since this is the size of a minimum cut in the cycle. Thus, since a fraction of at least $1 - \mathrm{Opt}_k(\Psi)$ hyperedges of Ψ are not k-wise satisfied by ℓ, we conclude that the total fraction of not 2-wise satisfied edges in Ψ' must be at least $(1 - \mathrm{Opt}_k(\Psi)) \cdot 2/k$.

In particular, it follows that if $\mathrm{Opt}_k(\Psi) \geq 1 - \delta$, then $\mathrm{Opt}_2(\Psi') \geq 1 - \delta$, and if $\mathrm{Opt}_k(\Psi) \leq \delta$, then $\mathrm{Opt}_2(\Psi') \leq 1 - (1 - \delta)2/k \leq 1 - 1/k$ (assuming $\delta \leq 1/2$).

The Proposition follows by applying parallel repetition to Ψ'. In particular, it follows directly from Rao's recent strong characterization of the Unique Games Conjecture (Rao 2008):

THEOREM A.2 (Rao 2008). *Assume that there exists a $\delta^* > 0$ such that the following holds:*

For every $0 < \delta < \delta^*$ there exists an L such that it is NP-hard to distinguish between 2-ary Unique Label Cover instances Ψ with L labels in which $\mathrm{Opt}_2(\Psi) \leq 1 - \delta^{1/3}$ and $\mathrm{Opt}_2(\Psi) \geq 1 - \delta$.

Then the Unique Games Conjecture is true.

In particular, we have that if the (k, k)-UGC is true, then for all $\delta > 0$ it is NP-hard to distinguish between the case that $\mathrm{Opt}_2(\Psi) \leq 1 - 1/k$ and the case $\mathrm{Opt}_2(\Psi) \geq 1 - \delta$, which, by Theorem A.2 (with $\delta^* = 1/k^3$) implies that the Unique Games Conjecture, a.k.a. the $(2, 2)$-UGC, is true. □

B. Monotonicity of the approximability of Max k-CSP$_q$

In this section, we describe an observation due to Yury Makarychev, allowing us to relate inapproximability for general domain sizes q in terms of inapproximability for the case when $q = p^e$ is a prime power.

PROPOSITION B.1. *If the* MAX k-CSP$_q$ *problem can be approximated within a factor* $C(q,k)/q^k$ *then for any* $r \geq q$*, the* MAX k-CSP$_r$ *problem can be approximated within a factor* $C(q,k)/r^k$*.*

Note that $C(q,k)$ is the "multiplicative advantage" of the MAX k-CSP$_q$ algorithm over the random assignment algorithm. In other words, the proposition states that for fixed k this advantage *increases* as q increases.

PROOF. Given an instance Ψ of the MAX k-CSP$_r$ problem, pick, independently for each variable x_i, a set $S_i \subseteq [r]$ of size $|S_i| = q$ uniformly at random from $\binom{[r]}{q}$.

Construct a new MAX k-CSP$_r$ instance Ψ' by adding the constraints $x_i \in S_i$ for each variable x_i (i.e., for any constraint involving the variable x_i, we throw away all satisfying assignments to that constraint which are such that $x_i \notin S_i$). Clearly, Ψ' can be viewed as a MAX k-CSP$_q$ instance, and thus, we can find an assignment a to Ψ' with value at least $C(q,k)/q^k \operatorname{Opt}(\Psi')$, where $\operatorname{Opt}(\Psi')$ is the value of an optimal solution to Ψ'.

Finally, for any assignment a to Ψ with value V, the expected value of this assignment in Ψ' is exactly $(q/r)^k V$, since the sets S_i are all independent. It follows that $\mathbb{E}[\operatorname{Opt}(\Psi')] = (q/r)^k \operatorname{Opt}(\Psi)$, and in particular, the expected value of the assignment a to Ψ' is at least $C(q,k)/r^k \operatorname{Opt}(\Psi)$. Noting that the value of a as an assigment to Ψ is even larger, the proposition follows. □

In particular, Proposition B.1 has the following two easy corollaries:

COROLLARY B.2. *Assume that the* MAX k-CSP *problem can be approximated to within a factor* $C(k)/2^k$*. Then the* MAX k-CSP$_q$ *problem can be approximated to within a factor* $C(k\lfloor \log_2 q \rfloor)/q^k$*.*

PROOF. This follows by noting that the MAX k-CSP$_{2^l}$ problem can be approximated within a factor $C(kl)/(2^l)^k$ by encoding each value $i \in [2^l]$ as an l-bit string and solving the resulting MAX (kl)-CSP instance. We then apply Proposition B.1 with domain sizes $2^{\lfloor \log_2 q \rfloor}$ and q. □

COROLLARY B.3. *Assume that it is UG-hard to approximate* MAX k-CSP$_q$ *to within a factor* $C(q,k)/q^k$ *whenever* q *is a prime power. Then the* MAX k-CSP$_q$ *problem is UG-hard to approximate within* $C(q + o(q), k)/q^k$ *for all* q*.*

PROOF. Given q, let $p \geq q$ be the smallest prime larger than q. By Theorem 4.5, $p \leq q + \mathcal{O}(q^{0.525}) \leq q + o(q)$. The MAX k-CSP$_p$ problem is then

UG-hard to approximate within $C(q + o(q), k)/p^k$, and the corollary follows from the contrapositive formulation of Proposition B.1. □

Manuscript received September 2, 2008

Per Austrin
KTH – Royal Institute of Technology
Stockholm, Sweden
austrin@kth.se
http://www.csc.kth.se/~austrin

Elchanan Mossel
U.C. Berkeley, USA, &
Weizmann Institute, Rehovot, Israel
mossel@stat.berkeley.edu
http://www.stat.berkeley.edu/~mossel