Dichotomy for Holant* Problems with a Function on Domain Size 3

Jin-Yi Cai* Pinyan Lu† Mingji Xia‡

Abstract

Holant problems are a general framework to study the algorithmic complexity of counting problems. Both counting constraint satisfaction problems and graph homomorphisms are special cases. All previous results of Holant problems are over the Boolean domain. In this paper, we give the first dichotomy theorem for Holant problems for domain size greater than two. We discover unexpected tractable families of counting problems, by giving new polynomial time algorithms. This paper also initiates holographic reductions in domains of size greater than two. This is our main algorithmic technique, and is used for both tractable families and hardness reductions. The dichotomy theorem is the following: For any complex-valued symmetric function $\mathbf{F}$ with arity 3 on domain size 3, we give an explicit criterion on $\mathbf{F}$, such that if $\mathbf{F}$ satisfies the criterion then the problem Holant*($\mathbf{F}$) is computable in polynomial time, otherwise Holant*($\mathbf{F}$) is #P-hard.

1 Introduction

The study of computational complexity of counting problems has been a very active research area recently. Three related frameworks in which counting problems can be expressed as partition functions have received the most attention: Graph Homomorphisms (GH), Constraint Satisfaction Problems (CSP) and Holant Problems.

Graph Homomorphism was first defined by Lovász [38]. It captures a wide variety of graph properties. Given any fixed $k \times k$ symmetric matrix $\mathbf{A}$ over $\mathbb{C}$, the partition function $Z_{\mathbf{A}}$ maps any input graph $G = (V, E)$ to $Z_{\mathbf{A}}(G) = \sum_{\xi: V \to [k]} \prod_{(u,v)\in E} \mathbf{A}_{\xi(u),\xi(v)}$. When $\mathbf{A}$ is a 0-1 matrix, then the product $\prod_{(u,v)\in E}$ is essentially a Boolean AND function. The product value $\prod_{(u,v)\in E} \mathbf{A}_{\xi(u),\xi(v)} = 0$ or 1, and it is 1 iff every edge $(u, v) \in E$ is mapped to an edge in the graph H whose adjacency matrix is $\mathbf{A}$. Hence for a 0-1 matrix $\mathbf{A}$, $Z_{\mathbf{A}}(G)$ counts the number of homomorphisms from G to H. For example, if $\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$ then $Z_{\mathbf{A}}(G)$ counts the number of INDEPENDENT SETS in G. If $\mathbf{A} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$ then $Z_{\mathbf{A}}(G)$ is the number of valid 3-COLORINGS. When $\mathbf{A}$ is not 0-1, $Z_{\mathbf{A}}(G)$ is a weighted sum of homomorphisms. Each $\mathbf{A}$ defines a graph property on graphs G. Clearly if G and G' are isomorphic then $Z_{\mathbf{A}}(G) = Z_{\mathbf{A}}(G')$. While individual graph properties are fascinating to study, Lovász's intent is to study a wide class of graph properties representable as graph homomorphisms. The use of more general matrices $\mathbf{A}$ brings us into contact with another tradition, called *partition functions of spin systems* from statistical physics (see [3, 39]). The case of a 2×2 matrix $\mathbf{A} = \begin{bmatrix} \beta & 1 \\ 1 & \gamma \end{bmatrix}$ is called a 2-spin system, and the special case $\beta = \gamma$ is the Ising model [32, 33, 29]. The Potts model [28] with interaction strength γ is defined by a $k \times k$ matrix $\mathbf{A}$ where all off-diagonal entries equal to 1 and all diagonal entries equal to $1 + \gamma$. In classical physics, the matrix $\mathbf{A}$ is always real-valued. However, in a quantum system for which complex number is the right language, the partition function is in general complex-valued [25]. In particular, if the physics model is over a discrete graph and is non-orientable, then the edge weights are given by a symmetric complex matrix. We will see that the use of complex numbers is not just a modeling issue, it provides an inner unity in the algorithmic theory of partition functions.

A more general framework than GH is called counting CSP. Let $\mathcal{F}$ be any finite set of (complex-valued) constraint functions defined on some domain set D. It defines a counting CSP problem #CSP($\mathcal{F}$): An input consists of a bipartite graph $G = (X, Y, E)$, each $x \in X$ is a variable on D, each $y \in Y$ is labeled by a constraint function $f \in \mathcal{F}$, and the edges in E indicate how each constraint function is applied. The output is the sum of product of evaluations of the constraint functions over all assignments for the variables [18, 7, 20, 6, 14, 22, 11]. Again if all constraint functions in $\mathcal{F}$ are 0-1 valued then it counts the number of solutions. In general, this *sum of product* a.k.a. *partition function* is a weighted sum of solutions, and has occupied a central position. It reaches many areas ranging from AI, machine learning, tensor networks, statistical physics and coding theory. Note that GH is the special case of CSP where $\mathcal{F}$ consists of

*University of Wisconsin-Madison, and Peking University. jyc@cs.wisc.edu. Supported by NSF CCF-0914969.

†Microsoft Research Asia. pinyanl@microsoft.com

‡Max-Planck Institut für Informatik. xmjljx@gmail.com. Supported by NSFC 61003030.

a single binary symmetric function.

The strength of these frameworks derives from the fact that they can express many problems of interest and simultaneously it is possible to achieve a complete classification of its worst case complexity.

While GH (or spin systems) can express a great variety of natural counting problems, Freedman, Lovász and Schrijver [26] showed that GH cannot express the problem of counting Perfect Matchings. It is well known that the FKT algorithm [36, 42] can count the number of perfect matchings in a planar graph in polynomial time. This is one basic component of holographic algorithms recently introduced by Valiant [44, 43]. (The second basic component is holographic reduction.) To capture this extended class of problems typified by Perfect Matchings, the framework of Holant problems was introduced [13, 14, 15]. Briefly, an input instance of a Holant problem is a graph $G = (V, E)$ where each edge represents a variable and each vertex is labeled by a constraint function. The partition function is again the sum of product of the constraint function evaluations, over all edge assignments. E.g., if edges are Boolean variables (i.e., domain size 2), and the constraint function at every vertex is the Exact-One function which is 1 if exactly one incident edge is assigned true and 0 otherwise, then the partition function counts the number of perfect matchings. If each vertex has the At-Most-One function then it counts all (not necessarily perfect) matchings. It can be shown easily that the Holant framework can simulate spin systems but, as shown by [26], the converse is not true. The Holant framework turns out to be a very natural setting and captures many interesting problems. E.g., it was independently discovered in coding theory, where it is called Normal Factor Graphs or Forney Graphs [34, 35, 2, 1].

A complexity dichotomy theorem for counting problems classifies every problem within a class to be either in P or #P-hard. For GH, this is proved for $Z_{\mathbf{A}}$ for all symmetric complex matrices $\mathbf{A}$ [10]. This is a culmination of a long series of results [21, 8, 27]. The proof of [10] is difficult, but the tractability criterion is very explicit: $Z_{\mathbf{A}}$ is in polynomial time if $\mathbf{A}$ is a suitable rank-one modification of a tensor product of Fourier matrices, and is #P-hard otherwise. Explicit dichotomy theorems were also proved for counting CSP on the Boolean domain (i.e., $|D| = 2$): unweighted [18], non-negative weighted [20], real weighted [4], and finally complex weighted [14], where holographic reductions played an important role in the final result. Complex numbers make their appearance naturally as eigenvalues, and provide an internal logic to the theory, even if one is only interested in 0-1 valued constraint functions.

When we go from the Boolean domain to domain size greater than two, there is a huge increase in difficulty to prove dichotomy theorems. This is already seen in decision CSP, where the dichotomy (i.e., any decision CSP is either in P or NP-complete) for the Boolean domain is Schaefer's theorem [40], but the dichotomy for domain size 3 is a major achievement by Bulatov [5]. A long standing conjecture by Feder and Vardi [24] states that a dichotomy for decision CSP holds for all domain sizes, but this is open for domain size greater than three. The assertion that every decision CSP is either solvable in polynomial time or NP-complete is by no means obvious, since assuming P $\neq$ NP, Ladner showed that NP contains problems that are neither in P nor NP-complete [37]. This is also valid for P versus #P.

With respect to counting problems, for any finite set of 0-1 valued functions $\mathcal{F}$ over a general domain, Bulatov [6] proved a dichotomy theorem for #CSP($\mathcal{F}$), which uses deep results from Universal Algebra. Dyer and Richerby [22, 23] gave a more direct proof which has the advantage that their tractability criterion is decidable. Decidable dichotomy theorems are more desirable since they tell us not only every $\mathcal{F}$ belongs to either one or the other class, but also how to decide for a given $\mathcal{F}$ which class it belongs to. A decidable dichotomy theorem for #CSP($\mathcal{F}$), where all functions in $\mathcal{F}$ take non-negative values, is given in [11]. Finally a dichotomy theorem for all complex-valued #CSP($\mathcal{F}$) is proved in [9]. This last dichotomy is not known to be decidable.

More than giving a formal classification, the deeper meaning of a dichotomy theorem is to provide a comprehensive structural understanding as to what makes a problem easy and what makes it hard. This deeper understanding goes beyond the validity of a dichotomy, and even beyond the decidability of the dichotomy criterion. Decidabilitty is: Given $\mathcal{F}$, decide whether it satisfies the tractability criterion so that #CSP($\mathcal{F}$) is in P. Ideally we hope for dichotomy theorems that are *explicit* in the sense that the tractability criteria provide a mathematical characterization that can be applied symbolically to an arbitrary $\mathcal{F}$. An explicit dichotomy can also be readily used to prove broader dichotomy theorems, as we will see in this paper. The known dichotomy theorems for GH and for CSP on general domains have very different flavors. Dichotomy theorems for #CSP($\mathcal{F}$) for all domain sizes greater than two are not explicit. The tractability criterion is infinitary. This is in marked contrast with the dichotomy theorems for GH. For Holant problems all previous results are over the Boolean domain and are mostly explicit. In this paper, we give the first dichotomy theorem for Holant problems for domain size greater than two, and it is explicit.

Our main theorem can be stated as follows: For any complex-valued symmetric function $\mathbf{F}$ with arity 3 on domain size 3, we give an explicit criterion on $\mathbf{F}$, such that if $\mathbf{F}$ satisfies the criterion then the problem Holant$^*(\mathbf{F})$ is computable in polynomial time, otherwise Holant$^*(\mathbf{F})$ is #P-hard. (Formal definitions will be given in Section 2.) It is known that in the Holant framework any set of binary functions is tractable. A ternary function is the basic setting in the Holant framework where both tractable and intractable cases occur. A single ternary function in the Holant framework is the analog of GH as the basic setting in the CSP framework with a single binary function. Therefore this case is interesting in its own right. Furthermore, as demonstrated many times in the Boolean domain [14, 15, 12, 30, 31], a dichotomy for a single ternary function serves as the starting point for more general dichotomies in the Holant framework.

In order to prove this dichotomy theorem, we have to discover new tractable classes of Holant problems, and design new polynomial time algorithms. Many intricacies of the interplay between tractability and intractability do not occur in the Boolean domain. However these new algorithms actually provide fresh insight to our previous dichotomy theorems for the Boolean domain. They offer a deeper and more complete understanding of what makes a problem easy and what makes it hard.

Our main algorithmic innovation is to initiate the theory of holographic reductions in domain size three using 3×3 matrices. It is a recurring theme in our proof techniques here. This is a new development; all previous work on holographic reductions have been focused on the Boolean domain. Holographic transformation offers a perspective on internal connections and equivalences between different looking problems, that is unavailable by any other means. In particular since it naturally uses eigenvalues and eigenvectors, the field of complex numbers $\mathbb{C}$ is the natural setting to formulate the class of problems, even if one is only interested in 0-1 valued or non-negative valued constraint functions. Using complex-valued constraints in defining Holant problems we can see the internal logical connections between various problems. Completely different looking problems can be seen as one and the same problem under holographic transformations. The proof of our dichotomy theorem would be impossible without working over $\mathbb{C}$. Even the dichotomy criterion would be impossible to state without it. To quote Jacques Hadamard: *"The shortest path between two truths on the real line passes through the complex plane."*

Suppose our domain set is $\{B, G, R\}$, named for the three colors Blue, Green and Red. We isolate several classes of tractable cases of $\mathbf{F}$. One of them is a generalization of Fibonacci signatures from the Boolean domain, under an orthogonal transformation. Another involves a concept called isotropic vectors, which self-annihilates under dot product. The third type involves a more intricate interplay between an isotropic vector in some dimension and another function primarily "living" in the other dimensions. This last type was only discovered after we failed to push through certain hardness proofs.

For hardness proofs, the first main idea is to construct a binary function which acts as an EQUALITY function when restricted to $\{G, R\}$, and is zero elsewhere. This construction allows us to restrict a function on $\{B, G, R\}$ to a domain of size 2, and employ the known (and explicit) dichotomy theorems for the Boolean domain. The plan is to use it to restrict $\mathbf{F}$ to $\{G, R\}$ and, assuming it is non-degenerate, to *anchor* the entire hardness proof on that. Here it is crucial that the known Boolean domain dichotomy is explicit. This part of the proof is quite demanding and heavily depends on holographic reductions. A central motif is to show that after a holographic reduction, $\mathbf{F}$ must possess *fantastic* regularity to escape #P-hardness.

What perhaps took us by surprise is that when $\mathbf{F}$ restricted to $\{G, R\}$ is degenerate, there is still considerable technical difficulty remaining. These are eventually overcome by using unsymmetric functions (in the full paper [17], this part of the proof starts from Section 5.4.)

This work has been a marathon for us. During the process, repeatedly, we failed to clinch the hardness proof for some subclasses of functions and then new tractable cases were found. So we had to reformulate the final dichotomy several times. The discovery process is mutually reinforcing between new algorithms and hardness proofs. On many occasions we believed that we had overcome one last hurdle, only to be stymied by yet another. However the struggle has also paid handsome dividends. For example, our SODA paper two years ago [16] was obtained as part of the program to achieve this dichotomy. We realized we needed a dichotomy for unsymmetric functions over the Boolean domain, and indeed that is used to overcome a major difficulty in the proof here.

2 Preliminary

Holant problems are designed to capture a broad class of locally constrained counting type problems. Let D be a finite domain set, and $\mathcal{F}$ be a finite set of constraint functions called signatures. Each $f \in \mathcal{F}$ is a mapping from $D^k \to \mathbb{C}$ for some arity k. We assume signatures take complex algebraic numbers.

A *signature grid* $\Omega = (G, \mathcal{F}, \pi)$ consists of a graph $G = (V, E)$ where each vertex $v \in V$ is labeled by a function $f_v \in \mathcal{F}$, and π is the labeling. The Holant problem on instance Ω is to evaluate

$$\text{Holant}_\Omega = \sum_{\sigma} \prod_{v \in V} f_v(\sigma \mid_{E(v)}),$$

a sum over all edge assignments $\sigma : E \to D$, where $E(v)$ denotes the incident edges at v.

A function f_v is listed by its values lexicographically as a truth table, or as a tensor in $(\mathbb{C}^{|D|})^{\otimes \deg(v)}$. We can identify a unary function $f(x) : D \to \mathbb{C}$ with a vector $\mathbf{u} \in \mathbb{C}^{|D|}$. Given two vectors $\mathbf{u}$ and $\mathbf{v}$ of dimension $|D|$, the tensor product $\mathbf{u} \otimes \mathbf{v}$ is a vector in $\mathbb{C}^{|D|^2}$, with entries $u_i v_j$ $(1 \le i, j \le |D|)$. Similarly for matrices A and B, $A \otimes B$ has entries $a_{i,j} b_{k,l}$ indexed by $((i,k),(j,l))$. We write $\mathbf{u}^{\otimes k}$ for $\mathbf{u} \otimes \ldots \otimes \mathbf{u}$ with k copies of $\mathbf{u}$. $A^{\otimes k}$ is similarly defined. We have $A^{\otimes k} \mathbf{u}^{\otimes k} = (A\mathbf{u})^{\otimes k}$, and $A^{\otimes k} B^{\otimes k} = (AB)^{\otimes k}$.

A signature f of arity k is *degenerate* if $f = \mathbf{u}_1 \otimes \mathbf{u}_2 \otimes \ldots \otimes \mathbf{u}_k$ for some vectors $\mathbf{u}_i$. Equivalently there are unary functions f_i such that $f(x_1, \ldots, x_k) = f_1(x_1) \cdots f_k(x_k)$. Such a signature is very weak; there is no interaction between the variables. If every function in $\mathcal{F}$ is degenerate, then Holant_Ω for any $\Omega = (G, \mathcal{F}, \pi)$ is computable in polynomial time in a trivial way: Simply split every vertex v into $\deg(v)$ many vertices each assigned a unary f_i and connected to the incident edge. Then Holant_Ω becomes a product over each component of a single edge. Thus degenerate signatures are weak and should be properly understood as made up by unary signatures. To concentrate on the essential features that divide tractability from intractability, we introduced Holant^* problems. These are Holant problems where unary signatures are assumed to be present [14, 15].

DEFINITION 2.1. *Given a set of signatures $\mathcal{F}$,* Holant($\mathcal{F}$) *is the class of all Holant problems using (any finite subset of) $\mathcal{F}$, and* $\text{Holant}^*(\mathcal{F})$ *denotes* Holant($\mathcal{F} \cup \mathcal{U}$)*, where $\mathcal{U}$ is the set of all unary signatures.*

A signature f on k variables is symmetric if $f(x_1, \ldots, x_k) = f(x_{\sigma(1)}, \ldots, x_{\sigma(k)})$ for all $\sigma \in \mathfrak{S}_k$, the *symmetric group*. It can be shown easily that a symmetric signature f is degenerate iff $f = \mathbf{u}^{\otimes k}$ for some unary $\mathbf{u}$.

A symmetric signature has a clear combinatorial meaning. A symmetric signature f on k Boolean variables can be expressed as $[f_0, f_1, \ldots, f_k]$, where f_j is the value of f on inputs of Hamming weight j.

THEOREM 2.1. (THEOREM 3.1 IN [14]) *Let $\mathcal{F}$ be any set of non-degenerate, symmetric, complex-valued signatures in Boolean variables. If $\mathcal{F}$ is of one of the following types, then* $\text{Holant}^*(\mathcal{F})$ *is in P, otherwise it is #P-hard.*

1. *Any signature in $\mathcal{F}$ is of arity at most 2;*
2. *There exist two constants a and b (not both zero, depending only on $\mathcal{F}$), such that for all signatures $[f_0, f_1, \ldots, f_n]$ in $\mathcal{F}$ one of the two conditions is satisfied: (1) for every $k = 0, 1, \ldots, n-2$, we have $af_k + bf_{k+1} - af_{k+2} = 0$; (2) $n = 2$ and the signature $[f_0, f_1, f_2]$ is of the form $[2a\lambda, b\lambda, -2a\lambda]$.*
3. *For every signature $[f_0, f_1, \ldots, f_n] \in \mathcal{F}$ one of the two conditions is satisfied: (1) For every $k = 0, 1, \ldots, n-2$, we have $f_k + f_{k+2} = 0$; (2) $n = 2$ and the signature $[f_0, f_1, f_2]$ is of the form $[\lambda, 0, \lambda]$.*

There are alternative forms of this dichotomy theorem for Holant^* problems over the Boolean domain. When there is a single ternary symmetric function f, the Holant problem $\text{Holant}^*(\{f\})$ is tractable in the following cases, and is #P-hard otherwise (see [15, 16]).

1. $f = H^{\otimes 3}[a, 0, 0, b]^{\mathtt{T}}$, where H is a 2×2 orthogonal matrix, $ab \neq 0$.
2. $f = Z^{\otimes 3}[a, 0, 0, b]^{\mathtt{T}}$, where $Z = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ i & -i \end{bmatrix}$, $ab \neq 0$.
3. $f = Z^{\otimes 3}[a, b, 0, 0]^{\mathtt{T}}$, where $Z = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ i & -i \end{bmatrix}$ or $Z = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -i & i \end{bmatrix}$, $b \neq 0$.
4. f is degenerate.

To introduce the idea of holographic reductions, it is convenient to consider bipartite graphs. For a general graph, we can always transform it into a bipartite graph preserving the Holant value, as follows: For each edge in the graph, we replace it by a path of length 2, and assign to the new vertex the binary EQUALITY function $(=_2)$.

We use $\text{Holant}(\mathcal{R} \mid \mathcal{G})$ to denote the Holant problem on bipartite graphs $H = (U, V, E)$, where each signature for a vertex in U or V is from $\mathcal{R}$ or $\mathcal{G}$, respectively. An input instance for the bipartite Holant problem is a bipartite signature grid and is denoted as $\Omega = (H;\ \mathcal{R} \mid \mathcal{G};\ \pi)$. Signatures in $\mathcal{R}$ are considered as row vectors (or covariant tensors); signatures in $\mathcal{G}$ are considered as column vectors (or contravariant tensors) [19].

For a $|D| \times |D|$ matrix T and a signature set $\mathcal{F}$, define $T\mathcal{F} = \{g \mid \exists f \in \mathcal{F}$ of arity $n,\ g = T^{\otimes n} f\}$, similarly for $\mathcal{F}T$. Whenever we write $T^{\otimes n} f$ or $T\mathcal{F}$, we view the signatures as column vectors; similarly for $fT^{\otimes n}$ or $\mathcal{F}T$ as row vectors. A holographic transformation by T is the following operation: given a signature grid $\Omega = (H;\ \mathcal{R} \mid \mathcal{G};\ \pi)$, for the same graph H, we get a new grid $\Omega' = (H;\ \mathcal{R}T \mid T^{-1}\mathcal{G};\ \pi')$ by replacing each signature in $\mathcal{R}$ or $\mathcal{G}$ with the corresponding signature in $\mathcal{R}T$ or $T^{-1}\mathcal{G}$.

THEOREM 2.2. (VALIANT'S HOLANT THEOREM [44]) *If there is a holographic transformation mapping signature grid Ω to Ω', then* $\text{Holant}_\Omega = \text{Holant}_{\Omega'}$.

Therefore, an invertible holographic transformation does not change the complexity of the Holant problem in the bipartite setting. We illustrate the power of holographic transformation by an example. Let $f = [\frac{3}{2}, 0, \frac{1}{2}, 0, \frac{3}{2}]$. Consider $\text{Holant}(f)$ on the Boolean domain. For a 4-regular graph G, $\text{Holant}(f)$ is a sum over all 0-1 edge assignments of products of local evaluations. Each vertex contributes a factor $\frac{3}{2}$ if all incident edges are assigned the same truth value, a factor $\frac{1}{2}$ if exactly half are assigned 1 and the other half 0. Before anyone consigns this problem to be artificial and unnatural, consider a holographic transformation by $Z = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ i & -i \end{bmatrix}$. Then $\text{Holant}(f) = \text{Holant}(=_2 \mid f) = \text{Holant}((=_2)Z^{\otimes 2} \mid (Z^{-1})^{\otimes 4} f)$. Let $\hat{f} = [0, 0, 1, 0, 0]$, and writing it as a symmetrized sum of tensor products, then

$$
\begin{aligned}
& Z^{\otimes 4}\hat{f} \\
=& Z^{\otimes 4}\left\{\begin{bmatrix}1\\0\end{bmatrix}\otimes\begin{bmatrix}1\\0\end{bmatrix}\otimes\begin{bmatrix}0\\1\end{bmatrix}\otimes\begin{bmatrix}0\\1\end{bmatrix}+\begin{bmatrix}1\\0\end{bmatrix}\otimes\begin{bmatrix}0\\1\end{bmatrix}\otimes\begin{bmatrix}1\\0\end{bmatrix}\otimes\begin{bmatrix}0\\1\end{bmatrix}\right. \\
& \left. +\cdots+\begin{bmatrix}0\\1\end{bmatrix}\otimes\begin{bmatrix}0\\1\end{bmatrix}\otimes\begin{bmatrix}1\\0\end{bmatrix}\otimes\begin{bmatrix}1\\0\end{bmatrix}\right\} \\
=& \tfrac{1}{4}\left\{\begin{bmatrix}1\\i\end{bmatrix}\otimes\begin{bmatrix}1\\i\end{bmatrix}\otimes\begin{bmatrix}1\\-i\end{bmatrix}\otimes\begin{bmatrix}1\\-i\end{bmatrix}+\begin{bmatrix}1\\i\end{bmatrix}\otimes\begin{bmatrix}1\\-i\end{bmatrix}\otimes\begin{bmatrix}1\\i\end{bmatrix}\right. \\
& \left.\otimes\begin{bmatrix}1\\-i\end{bmatrix}+\cdots+\begin{bmatrix}1\\-i\end{bmatrix}\otimes\begin{bmatrix}1\\-i\end{bmatrix}\otimes\begin{bmatrix}1\\i\end{bmatrix}\otimes\begin{bmatrix}1\\i\end{bmatrix}\right\} \\
=& \tfrac{1}{2}[3, 0, 1, 0, 3] = f;
\end{aligned}
$$

Hence the contravariant transformation $(Z^{-1})^{\otimes 4} f = \hat{f}$. Meanwhile, a covariant transformation by Z transforms $(=_2)$ to the binary DISEQUALITY function $(\neq_2)$

$$
\begin{aligned}
(=_2)Z^{\otimes 2} &= \begin{pmatrix}1 & 0 & 0 & 1\end{pmatrix} Z^{\otimes 2} \\
&= \left\{\begin{pmatrix}1 & 0\end{pmatrix}^{\otimes 2} + \begin{pmatrix}0 & 1\end{pmatrix}^{\otimes 2}\right\} Z^{\otimes 2} \\
&= \tfrac{1}{2}\left\{\begin{pmatrix}1 & 1\end{pmatrix}^{\otimes 2} + \begin{pmatrix}i & -i\end{pmatrix}^{\otimes 2}\right\} \\
&= [0, 1, 0] = (\neq_2).
\end{aligned}
$$

So $\text{Holant}(f) = \text{Holant}((\neq_2) \mid [0, 0, 1, 0, 0])$; they are really one and the same problem. A moment's reflection shows that this latter formulation is counting the number of Eulerian orientations on 4-regular graphs, an eminently natural problem!

Furthermore, holographic transformation by an orthogonal matrix T preserves the binary equality and thus can be used freely in the standard setting.

THEOREM 2.3. *Suppose T is an orthogonal matrix ($TT^{\mathsf{T}} = I$) and let $\Omega = (G, \mathcal{F}, \pi)$ be a signature grid. Under a holographic transformation by T, we get a new grid $\Omega' = (G, T\mathcal{F}, \pi')$ and* $\text{Holant}_\Omega = \text{Holant}_{\Omega'}$.

Let $\mathbf{F}$ be a symmetric signatures of arity 3 over domain $\{B, G, R\}$. We use the following notation.

$$
\mathbf{F} = [F_{BBB}; F_{BBG}, F_{BBR}; F_{BGG}, F_{BGR}, F_{BRR};
$$

$$
F_{GGG}, F_{GGR}, F_{GRR}, F_{RRR}].
$$

Alternatively we also use the following notation:

(2.1)

$$
\begin{array}{ccccccc}
 & & & F_{BBB} & & & \\
 & & F_{BBG} & & F_{BBR} & & \\
 & F_{BGG} & & F_{BGR} & & F_{BRR} & \\
F_{GGG} & & F_{GGR} & & F_{GRR} & & F_{RRR}
\end{array}
$$

For a signature $\mathbf{F}$ of arity two, we also use a symmetric 3×3 matrix to represent it, $M = M_{\mathbf{F}} = \begin{bmatrix} F_{BB} & F_{BG} & F_{BR} \\ F_{BG} & F_{GG} & F_{GR} \\ F_{BR} & F_{GR} & F_{RR} \end{bmatrix}$. The rank of a binary signature is the rank of its 3×3 matrix. Also the matrix form of $T^{\otimes 2}\mathbf{F}$ is TMT^{T}. The matrix form of $(=_2)$ is the identity matrix I. Thus for an orthogonal T, and $\mathbf{F}$ of arity r, $\text{Holant}(\mathbf{F}) = \text{Holant}(=_2 \mid \mathbf{F}) = \text{Holant}((=_2)(T^{\mathsf{T}})^{\otimes 2} \mid T^{\otimes r}\mathbf{F}) = \text{Holant}(T\mathbf{F})$.

A unary function can be represented as $[F_B; F_G, F_R]$ in symmetric notation, or (F_B, F_G, F_R) as a vector. For a function of arity r, We use $\mathbf{F}^{i=A}$, where $i \in [r]$ and $A \in \{B, G, R\}$, to denote a signature of arity $r - 1$ by fixing the i-th input of $\mathbf{F}$ to A. For example for the ternary function $\mathbf{F}$ in (2.1),

$$
\mathbf{F}^{1=B} = \begin{bmatrix} F_{BBB} & F_{BBG} & F_{BBR} \\ F_{BBG} & F_{BGG} & F_{BGR} \\ F_{BBR} & F_{BGR} & F_{BRR} \end{bmatrix}. \tag{2.2}
$$

Sometimes, we also restrict the i-th input of $\mathbf{F}$ to a subset S of $\{B, G, R\}$, and we use $\mathbf{F}^{i \to S}$ (for example $\mathbf{F}^{2 \to \{B,R\}}$) to denote it. We use $\mathbf{F}^{* \to S}$ to denote the case when we restrict all inputs of $\mathbf{F}$ to S. For example $\mathbf{F}^{* \to \{G,R\}} = [F_{GGG}, F_{GGR}, F_{GRR}, F_{RRR}]$. The above notation can be combined, for example $\mathbf{F}^{1=B; 2,3 \to \{G,R\}} = [F_{BGG}, F_{BGR}, F_{BRR}]$. We also use $F_{a,b,c}$, $(a, b, c \in \mathbf{N}, a + b + c = r)$ to denote the value of $\mathbf{F}$ when the numbers of B's, G's and R's among the inputs are respectively a, b and c. For example, $F_{1,2,0} = F_{BGG}$.

We use $\mathrm{Sym}(\mathbf{F})$ to denote the symmetrization of $\mathbf{F}$ as follows: For $i_1, i_2, \ldots, i_r \in \{B, G, R\}$, $(\mathrm{Sym}(\mathbf{F}))_{(i_1 i_2 \ldots i_r)} = \sum_{\sigma \in \mathfrak{S}_r} F_{i_{\sigma 1} i_{\sigma 2} \ldots i_{\sigma r}}$, where the summation is over the *symmetric group* $\mathfrak{S}_r$. [1]

Let $\mathbf{F}$ be a ternary symmetric signature, and let $\mathbf{u} = (\alpha, \beta, \gamma)$ be a unary signature, both on domain $\{B, G, R\}$, we can form a binary symmetric signature by connecting one input of $\mathbf{F}$ with $\mathbf{u}$. Since $\mathbf{F}$ is symmetric, connecting to any one of the input wires defines the same symmetric signature on the other input wires. We denote this signature by $\langle \mathbf{u}, \mathbf{F} \rangle$. This (contraction) operation can be performed on any signature $\mathbf{F}$ of arity at least 1. For $\mathbf{F}$ of arity at least 2, $\langle \mathbf{v}, \langle \mathbf{u}, \mathbf{F} \rangle \rangle = \langle \mathbf{u}, \langle \mathbf{v}, \mathbf{F} \rangle \rangle$. For two unary functions $\mathbf{u}$ and $\mathbf{v}$, $\langle \mathbf{u}, \mathbf{v} \rangle$ is simply the dot product value. A vector $\mathbf{v}$ is *isotropic* if $\langle \mathbf{v}, \mathbf{v} \rangle = 0$. For example $\mathbf{v} = (1, i)$. Any non-zero isotropic vector of length 3 can be transformed to $(1, i, 0)$ by an orthogonal transformation.

3 Statement of the Dichotomy Theorem

THEOREM 3.1. *Let $\mathbf{F}$ be a symmetric ternary function over domain $\{B, G, R\}$. Then Holant$^*(\mathbf{F})$ is #P-hard unless $\mathbf{F}$ is of one of the following forms, in which case the problem is in polynomial time.*

1. *There exist three vectors $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ of dimension 3 such that they are mutually orthogonal to each other, i.e. $\langle \boldsymbol{\alpha}, \boldsymbol{\beta} \rangle = 0$, $\langle \boldsymbol{\alpha}, \boldsymbol{\gamma} \rangle = 0$ and $\langle \boldsymbol{\beta}, \boldsymbol{\gamma} \rangle = 0$, and $\mathbf{F} = \boldsymbol{\alpha}^{\otimes 3} + \boldsymbol{\beta}^{\otimes 3} + \boldsymbol{\gamma}^{\otimes 3}$;*

2. *There exist three vectors $\boldsymbol{\alpha}$, $\boldsymbol{\beta_1}$, and $\boldsymbol{\beta_2}$ of dimension 3 such that $\langle \boldsymbol{\alpha}, \boldsymbol{\beta_1} \rangle = 0$, $\langle \boldsymbol{\alpha}, \boldsymbol{\beta_2} \rangle = 0$, $\langle \boldsymbol{\beta_1}, \boldsymbol{\beta_1} \rangle = 0$, $\langle \boldsymbol{\beta_2}, \boldsymbol{\beta_2} \rangle = 0$ and $\mathbf{F} = \boldsymbol{\alpha}^{\otimes 3} + \boldsymbol{\beta_1}^{\otimes 3} + \boldsymbol{\beta_2}^{\otimes 3}$;*

3. *There exist two vectors $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ of dimension 3 and a function $\mathbf{F}_\beta$ of arity three, such that $\boldsymbol{\beta} \neq \mathbf{0}$, $\langle \boldsymbol{\beta}, \boldsymbol{\beta} \rangle = 0$, $\langle \mathbf{F}_\beta, \boldsymbol{\beta} \rangle = \mathbf{0}$ and $\mathbf{F} = \mathbf{F}_\beta + \boldsymbol{\beta}^{\otimes 2} \otimes \boldsymbol{\gamma} + \boldsymbol{\beta} \otimes \boldsymbol{\gamma} \otimes \boldsymbol{\beta} + \boldsymbol{\gamma} \otimes \boldsymbol{\beta}^{\otimes 2}$.*

Remarks: (I) In the forms above, the vectors $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\beta_1}$, $\boldsymbol{\beta_2}$ can be the zero vector (except $\boldsymbol{\beta}$ in *3.*)
(II) In form *3*, $\mathbf{F}$ is the sum of $\mathbf{F}_\beta$ with (1/2 of) the symmetrization of $\boldsymbol{\beta}^{\otimes 2} \otimes \boldsymbol{\gamma}$. The constant factor 1/2 doesn't matter, and can be absorbed in $\boldsymbol{\gamma}$.
(III) Let T be an orthogonal 3×3 matrix, then $\mathbf{F}$ is of one of the three forms above iff $T^{\otimes 3}\mathbf{F}$ is.

Theorem 3.1 gives a complete list of tractable cases for Holant$^*(\mathbf{F})$. We now give various canonical forms of these tractable cases, under an orthogonal transformation T.

[1]Usually, there is a normalization factor $\frac{1}{r!}$ in front of the summation, however a global factor does not change the complexity and we ignore this factor for notational simplicity.

THEOREM 3.2. *Let $\mathbf{F}$ be a symmetric ternary function over domain $\{B, G, R\}$. Then Holant$^*(\mathbf{F})$ is #P-hard unless there is an orthogonal transformation T, such that the function $T^{\otimes 3}\mathbf{F}$ is of one of the following forms, in which case the problem is in P.*

1. *For some $a, b, c \in \mathbb{C}$, $T^{\otimes 3}\mathbf{F} = a\boldsymbol{e_1}^{\otimes 3} + b\boldsymbol{e_2}^{\otimes 3} + c\boldsymbol{e_3}^{\otimes 3}$.*

2. *For some $c \neq 0$ and $\lambda \in \mathbb{C}$, $cT^{\otimes 3}\mathbf{F} = \boldsymbol{\beta_0}^{\otimes 3} + \overline{\boldsymbol{\beta_0}}^{\otimes 3} + \lambda \boldsymbol{e_3}^{\otimes 3}$, where $\boldsymbol{\beta_0} = \frac{1}{\sqrt{2}}(1, i, 0)^{\mathtt{T}}$, and $\overline{\boldsymbol{\beta_0}}$ is its conjugate $\frac{1}{\sqrt{2}}(1, -i, 0)^{\mathtt{T}}$.*

3. *For $\epsilon \in \{0, 1\}$, $T^{\otimes 3}\mathbf{F} = \mathbf{F}_0 + \epsilon \mathrm{Sym}(\boldsymbol{\beta_0} \otimes \boldsymbol{\beta_0} \otimes \overline{\boldsymbol{\beta_0}})$, where $\mathbf{F}_0$ satisfies the annihilation condition $\langle \mathbf{F}_0, \boldsymbol{\beta_0} \rangle = \mathbf{0}$.*

4 Tractability

Suppose $\mathbf{F} = [3; 1, 1; 5, 1, 3; 7, 5, 1, 1]$. Is Holant$^*(\mathbf{F})$ computable in polynomial time? It turns out that there are three pairwise orthogonal vectors $(1, -1, 1)^{\mathtt{T}}, (1, 0, -1)^{\mathtt{T}}$ and $(1, 2, 1)^{\mathtt{T}}$ such that $\mathbf{F} = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}^{\otimes 3} + \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}^{\otimes 3} + \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}^{\otimes 3}$. By Theorem 3.1, Holant$^*(\mathbf{F})$ is tractable. If we take $T = \frac{1}{\sqrt{6}} \begin{bmatrix} \sqrt{2} & \sqrt{3} & 1 \\ -\sqrt{2} & 0 & 2 \\ \sqrt{2} & -\sqrt{3} & 1 \end{bmatrix}$, then T is orthogonal, and $\mathbf{F} = T^{\otimes 3}\mathbf{F}'$, where $\mathbf{F}' = \sqrt{27}\boldsymbol{e_1}^{\otimes 3} + \sqrt{8}\boldsymbol{e_2}^{\otimes 3} + \sqrt{216}\boldsymbol{e_3}^{\otimes 3}$. Hence we can perform an orthogonal transformation by T, then the problem Holant$^*(\mathbf{F})$ is transformed to Holant$^*(\mathbf{F}')$. For $\mathbf{F}'$ the polynomial time algorithm on any input G is simple: In each connected component of G, any color from $\{B, G, R\}$ at a vertex v uniquely determines the same color at all its neighbors, and the vertex contributes a factor $\sqrt{27}$ or $\sqrt{8}$ or $\sqrt{216}$ respectively. These values are multiplied over the connected component. Thus, if G has connected components $C_1, C_2, \ldots, C_k$, and C_j has n_j vertices, then the Holant values is $\prod_{1 \le j \le k} (\sqrt{27}^{n_j} + \sqrt{8}^{n_j} + \sqrt{216}^{n_j})$.

We believe for countless such questions, not only the problem is very natural, but also the answer is not obvious without the underlying theory. Note that even though the function $\mathbf{F}$ above takes only positive values, the vectors can have negative entries. Armed with the dichotomy theorem, any interested reader can find many more examples.

In this section we prove that Holant$^*(\mathbf{F})$ is computable in polynomial time, for any symmetric ternary function $\mathbf{F}$ given in the three forms of Theorem 3.1, or equivalently Theorem 3.2.

For any 3×3 orthogonal matrix T, it keeps the binary equality $(=_2)$ over $\{B, G, R\}$ unchanged, namely

$T^{\mathtt{T}} I_3 T = I_3$ in matrix notation. Hence Holant$^*(\mathbf{F})$ is tractable iff Holant$^*(T^{\otimes 3}\mathbf{F})$ is tractable.

The above argument proves that Holant$^*(\mathbf{F})$ is computable in polynomial time if $\mathbf{F}$ has form *1.*

$$a\boldsymbol{e_1}^{\otimes 3} + b\boldsymbol{e_2}^{\otimes 3} + c\boldsymbol{e_3}^{\otimes 3}.$$

In form *2.*, let $\mathbf{F}$ be

$$\boldsymbol{\beta_0}^{\otimes 3} + \overline{\boldsymbol{\beta_0}}^{\otimes 3} + \lambda\boldsymbol{e_3}^{\otimes 3}.$$

Under the matrix $M = \begin{bmatrix} Z^{-1} & 0 \\ 0 & 1 \end{bmatrix}$, where $Z = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ i & -i \end{bmatrix}$, $Z^{-1} = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & -i \\ 1 & i \end{bmatrix}$, the function $\mathbf{F}$ is transformed to

$$M^{\otimes 3}\mathbf{F} = \boldsymbol{e_1}^{\otimes 3} + \boldsymbol{e_2}^{\otimes 3} + \lambda\boldsymbol{e_3}^{\otimes 3}.$$

Meanwhile the covariant transformation on the binary equality is $(=_2)(M^{-1})^{\otimes 2}$, which has the matrix form $(M^{-1})^{\mathtt{T}} I M^{-1} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$. This can be viewed as a Disequality on $\{B, G\}$ and Equality on $\{R\}$, with a separated domain. Now it is clear that Holant$^*(\mathbf{F})$ is computable in polynomial time by a connectivity argument. Within each connected component, any assignment of R will be uniquely propagated as R; any assignment of B or G will be exchanged to G or B along every edge.

The proof of tractability for form *3.* is more involved. We refer to the more generic expression of form *3.* in Theorem 3.1. First, under an orthogonal transformation we may assume $\boldsymbol{\beta} = \begin{bmatrix} 1 & i & 0 \end{bmatrix}^{\mathtt{T}}$. The function $\mathbf{F}$ is expressed as a sum $S + \boldsymbol{\beta}^{\otimes 2} \otimes \boldsymbol{\gamma} + \boldsymbol{\beta} \otimes \boldsymbol{\gamma} \otimes \boldsymbol{\beta} + \boldsymbol{\gamma} \otimes \boldsymbol{\beta}^{\otimes 2}$, where $\langle S, \boldsymbol{\beta}\rangle = \mathbf{0}$. We denote by $T_0 = S$, and T_j for the remaining three terms respectively, $1 \le j \le 3$. The value Holant$^*(\mathbf{F})$ is the sum over all $\{B, G, R\}$ edge assignments, $\sum_\sigma \prod_v f_v(\sigma \mid_{E(v)})$, where $E(v)$ are the edges incident to v, and all f_v are the function $\mathbf{F}$, or some unary function.

Without loss of generality, we can assume the input graph is connected. In the first step, we handle all vertices of degree one. Such a vertex v is connected to another vertex p of degree d. We can calculate a function of arity $d-1$ by combining the unary function at v with the function at p. This is a symmetric function and we can replace the vertex p together with v by a vertex q of degree $d-1$ and given this function. If $d = 1$, since the graph is connected, there is no vertex left and we have computed the value of the problem. If $d = 2$, the new function at q is a unary function. If $d = 3$, then f_p is $\mathbf{F}$. We may repeat this process until all vertices are of degree 2 or 3 and given either $\mathbf{F}$ or $\langle \mathbf{u}, \mathbf{F}\rangle = \sum_{j=0}^{3} T'_j$ for some unary $\mathbf{u}$, where $T'_j = \langle \mathbf{u}, T_j\rangle$.

For every vertex v of degree 2 or 3, we can express the function f_v as $\sum_{j=0}^{3} T'_j$ or $\sum_{j=0}^{3} T_j$ with the incident edges assigned as (ordered) input variables to each T'_j or T_j. (Note that T'_j and T_j are in general not symmetric, for $1 \le j \le 3$.) Then Holant$^*(\mathbf{F}) = \sum_\sigma \prod_v f_v(\sigma \mid_{E(v)}) = \sum_\sigma \prod_v \sum_{j=0}^{3} f_{v,j}(\sigma \mid_{E(v)}) = \sum_\tau \sum_\sigma \prod_v f_{v,\tau(v)}(\sigma \mid_{E(v)})$, where the first summation is over all assignments τ from all vertices $v \in V$ to some $j = \tau(v) \in \{0, 1, 2, 3\}$ which also assigns a copy of T'_j or T_j as $f_{v,\tau(v)}$ at v.

We are given that $\langle \boldsymbol{\beta}, T_0\rangle = \mathbf{0}$, then $\langle \boldsymbol{\beta}, T'_0\rangle = \mathbf{0}$ as well. Meanwhile $T'_1 = c_1 \boldsymbol{\beta}^{\otimes 2}$, $T'_2 = c_2 \boldsymbol{\beta} \otimes \boldsymbol{\gamma}$, and $T'_3 = c_3 \boldsymbol{\gamma} \otimes \boldsymbol{\beta}$, where the constants $c_1 = \langle \mathbf{u}, \boldsymbol{\gamma}\rangle$, and $c_2 = c_3 = \langle \mathbf{u}, \boldsymbol{\beta}\rangle$. Note that T'_j and T_j, for $1 \le j \le 3$, are all degenerate functions, and can be decomposed as unary functions. We also note that they all have at least as many copies of $\boldsymbol{\beta}$ as $\boldsymbol{\gamma}$.

Fix any τ, let $\mathcal{S}$ (resp. $\mathcal{T}$) denote the set of vertices which are assigned the function T_0 or T'_0 (resp. T_j or T'_j, with $1 \le j \le 3$) by τ. Suppose neither $\mathcal{S}$ nor $\mathcal{T}$ is empty. Then by connectedness, there are edges between $\mathcal{S}$ and $\mathcal{T}$. All functions in $\mathcal{T}$ are decomposed into unary functions. There are at least as many copies of $\boldsymbol{\beta}$ as $\boldsymbol{\gamma}$. Some of these functions may be paired up by edges inside $\mathcal{T}$. If any two copies of $\boldsymbol{\beta}$ are paired up, the product is zero. If every copy of $\boldsymbol{\beta}$ is paired up with some $\boldsymbol{\gamma}$ within $\mathcal{T}$, then at least one copy of $\boldsymbol{\beta}$ is connected to some vertex in $\mathcal{S}$. But every function in $\mathcal{S}$ is annihilated by $\boldsymbol{\beta}$. Hence the total contribution for such τ to Holant$^*(\mathbf{F})$ is zero when $\mathcal{S}$ are $\mathcal{T}$ both non-empty.

Now consider $\sum_\sigma \prod_v f_{v,\tau(v)}(\sigma \mid_{E(v)})$ for those τ such that either $\mathcal{S}$ or $\mathcal{T}$ is empty. Suppose $\mathcal{S} = \emptyset$. Again we decompose every function in $\mathcal{T}$ into unary functions. Then in order to be non-zero, the number of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ must be exactly equal. Hence if there is any vertex of degree 3, the contribution is 0. We only need to consider a connected graph such that all vertices have degree 2, which is a cycle. Because each $\boldsymbol{\beta}$ must be paired up exactly with $\boldsymbol{\gamma}$, we only need to calculate the sum $\sum_\sigma \prod_v f_{v,\tau(v)}(\sigma \mid_{E(v)})$ for two τ, which is tractable, since the graph is just a cycle.

Finally suppose $\mathcal{T} = \emptyset$. Then there is only one assignment τ which assigns T_0 and T'_0 to every vertex of degree 3 and 2 respectively. Consider all edge assignments σ. Suppose $E = \{e_1, e_2, \ldots, e_m\}$ is the edge set, and $e_1 = (p, q)$. All assignments σ are divided into 3 sets Σ_B, Σ_G or Σ_R, according to the value $\sigma(e_1) = B$, G or R, respectively. There is a natural one-to-one mapping ϕ from Σ_B to Σ_G, such that $(\phi(\sigma))(e_j) = \sigma(e_j)$ for $j = 2, \ldots, m$. Let $\theta(\sigma)$ denote $\prod_v f_{v,\tau(v)}(\sigma \mid_{E(v)})$, where $E(v)$ are the edges incident to v. Notice that at all $v \ne p, q$, the value of

$f_{v,\tau(v)}$ is the same for σ and $\phi(\sigma)$, but at $v = p, q$, $f_{v,\tau(v)}(\phi(\sigma) \mid_{E(v)}) = i f_{v,\tau(v)}(\sigma \mid_{E(v)})$, which can be directly verified using the condition $\langle S, \boldsymbol{\beta} \rangle = \mathbf{0}$. Hence $\theta(\phi(\sigma)) = -\theta(\sigma)$. Therefore we only need to calculate $\theta(\sigma)$ for σ in Σ_R. We can use $\sigma(e_2)$ to divide Σ_R into 3 sets, to repeat this process. At last, we only need to calculate $\theta(\sigma)$ for the single σ mapping every edge to R. This concludes the proof of tractability.

5 Outline for Hardness Proof

The starting point of our hardness proof is the dichotomy for Holant$^*(\mathbf{F})$ problems on the Boolean domain. A natural hope is that Holant$^*(\mathbf{F})$ is #P-hard if the Boolean domain Holant* problem for the function $\mathbf{F}^{*\to\{G,R\}}$, which is the restriction of the function $\mathbf{F}$ to the two-element subdomain $\{G, R\}$, is already #P-hard. But this statement is false when stated in such full generality, as we can easily construct an $\mathbf{F}$ such that Holant$^*(\mathbf{F})$ is tractable while Holant$^*(\mathbf{F}^{*\to\{G,R\}})$ is #P-hard (e.g., the first example in Section 4). However, this would be true if we have another special binary function $(=_{G,R}) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$. The reduction is straightforward: Given an instance G of Holant$^*(\mathbf{F}^{*\to\{G,R\}})$, we construct an instance of Holant$^*(\mathbf{F})$ by inserting a vertex into each edge of G and assigning the binary function $=_{G,R}$ to these vertices. The binary function $=_{G,R}$ in each edge acts as an equality function in the Boolean subdomain $\{G, R\}$ while any assignment of B anywhere produces a zero.

Therefore, our first main step is to construct the function $=_{G,R}$. If we can construct a non-degenerate binary function with the form $\begin{bmatrix} 0 & 0 & 0 \\ 0 & * & * \\ 0 & * & * \end{bmatrix}$, we can use interpolation to interpolate $=_{G,R}$ by a chain of copies of the above binary function. The remaining task is to realize such a binary function.

However we find that it is difficult or impossible to realize it directly by gadget construction in most cases. Here we use the idea of holographic reduction. As shown in the tractability part, holographic reduction plays an essential role there in developing polynomial algorithms. It also plays an important role in the hardness proof part as a method to normalize functions. We can always apply an orthogonal holographic transformation to a signature function without changing its complexity as shown in Theorem 2.3. If we can realize a binary function with rank 2, which can be constructed directly with the help of unary functions, then we can hope to use a holographic reduction to transform the binary function to the above form. This fits well with the idea of holographic reduction. A binary function with rank 2 shows that there is a hidden structure with a domain of size 2. The holographic reduction mixes the domain elements in a suitable way so that this hidden Boolean subdomain becomes explicit.

There are certain rank 2 matrices such as $\begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & i \\ 1 & i & 0 \end{bmatrix}$, for which an orthogonal holographic transformation does not exist. The reason is that the eigenvector of this matrix corresponding to the eigenvalue 0 is isotropic. This is the first place where isotropic vectors present some obstacle to our proof. There are several places throughout the entire proof, where we have to deal with isotropic vectors separately. There are two reasons: (1) For an isotropic vector, we cannot normalize it to a unit vector by an orthogonal transformation; (2) There are indeed additional tractable functions which are related to isotropic vectors. Consequently we have to circumvent this obstacle presented by the isotropic eigenvectors.

Additionally, there are some exceptional cases where the above process cannot go through. For these cases, we either prove the hardness result directly or show that it belongs to one of the three forms in Theorem 3.1. In the second main step, we assume that we are already given $=_{G,R}$ and we further prove that Holant$^*(\mathbf{F})$ is #P-hard if $\mathbf{F}$ is not of one of the three forms in Theorem 3.1.

Given $=_{G,R}$, Holant$^*(\mathbf{F})$ is #P-hard if Holant$^*(\mathbf{F}^{*\to\{G,R\}})$ is #P-hard, which we use our previous dichotomy for Boolean Holant* to determine. Hence we may assume that $\mathbf{F}^{*\to\{G,R\}}$ takes a tractable form. At this point, we employ holographic reduction to normalize our function further. But we should be careful here since we do not want the transformation to destroy $=_{G,R}$. We introduce the idea of a domain separated holographic reduction. A basis for a domain separated holographic transformation is of the form $\begin{bmatrix} * & 0 & 0 \\ 0 & * & * \\ 0 & * & * \end{bmatrix}$, which mixes up the subdomain $\{G, R\}$ while keeping B separate. In particular, such orthogonal holographic transformations preserve $=_{G,R}$.

For example, when $\mathbf{F}^{*\to\{G,R\}}$ is $a(x, y)^{\otimes 2} + b(z, w)^{\otimes 2} = [ax^2 + by^2, axy + bzw, ay^2 + bw^2]$, where $\begin{pmatrix} x & y \\ z & w \end{pmatrix}$ is an orthogonal matrix (this corresponds to a tractable case in form 2 of Theorem 2.1), we can apply an orthogonal holographic transformation of the basis $\begin{pmatrix} 1 & 0 & 0 \\ 0 & x & y \\ 0 & z & w \end{pmatrix}$ so that $\mathbf{F}$ is transformed to $\mathbf{H} =$

$$\begin{array}{ccccccc} & & & H_{BBB} & & & \\ & & H_{BBG} & & H_{BBR} & & \\ & H_{BGG} & & H_{BGR} & & H_{BRR} & \\ a & & 0 & & 0 & & b \end{array}.$$

According to the Holant* dichotomy on domain size 2, when putting this $\mathbf{H}^{*\to\{G,R\}} = [a, 0, 0, b]$ and a binary function together, the problem is #P-hard unless the binary function is of the form $[*, 0, *]$, $[0, *, 0]$, or degenerate. We shall prove that we can always construct a binary function which is not of these forms unless the function $\mathbf{F}$ has an *uncanny* regularity such that it is one of the forms in Theorem 3.1.

One idea greatly simplifies our argument in this part. By gadget construction, we can realize some binary functions with some parameters, which we can set freely to any complex number. Then we want to prove that we can set these parameters suitably so that the signature escapes from all the known tractable forms. This is quite difficult since different values may make the signature belong to different tractable forms. A nice observation here is that the condition that a binary signature belongs to a particular form say $[*, 0, *]$ can be described by the zero set of a polynomial. Thus these values form an algebraic set. To escape from a finite union of such sets, it is sufficient to prove that for every form, we can set these parameters to escape from this particular form. We call this the *polynomial argument*.

The spirit of the proof for all the other tractable non-degenerate ternary forms for $\mathbf{F}^{*\to\{G,R\}}$ is similar although the details are very different. In particular, we need to employ a non-orthogonal holographic transformation $\begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & Z \end{bmatrix}$ where $Z = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ i & -i \end{bmatrix}$. This transformation does not preserve $=_{G,R}$, rather it transforms $=_{G,R}$ to $(\neq_{G,R}) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$.

When the ternary signature $\mathbf{F}^{*\to\{G,R\}}$ is degenerate, the proof structure is quite different. The reason is that any set of binary functions are tractable in the Holant framework. So we have to construct a non-degenerate signature with arity at least three. It is quite difficult to construct a totally symmetric function with high arity except with some simple gadgets such as a star or a triangle. These gadgets work for some signatures but fail for others. Due to this difficulty, we employ unsymmetric gadgets too. Fortunately, we also have a dichotomy for unsymmetric Holant* problems in the Boolean domain [16]. Since the dichotomy for this more general Boolean Holant* is more complicated, we use a different proof strategy here. We only show the existence of a non-degenerate signature with arity at least three, but do not analyze all possible forms case-by-case. We instead prove that we can always construct some binary signature in addition to the higher arity one, which makes the problem hard no matter what the high arity signature is, provided that $\mathbf{F}$ is not one of the tractable cases.

Finally, for a particular family of signatures which can be normalized to the following form:

$$\begin{array}{ccccccc} & & & 0 & & & \\ & & ix & & x & & \\ & 0 & & 0 & & 0 & \\ 1 & & i & & -1 & & -i. \end{array}$$

where two isotropic vectors $(1, i)$ and $(1, -i)$ interact in an unfavorable way, we have to use a different argument. Due to its special structure, we have to use a different hard problem to reduce from, namely the problem of counting perfect matchings on 3-regular graphs. The problem is also used when $\mathbf{F}^{*\to\{G,R\}} = [0, 0, 0, 0]$ is identically 0. This problem is #P-hard (although tractable over planar graphs. This also indicates that the holographic reduction theory developed here is distinct from the theory of matchgate based holographic algorithms [44, 43].)

6 Realize the Binary Function $=_{G,R}$

Recall that $(=_{G,R}) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, and the first step of our hardness proof is to realize it, which is stated formally by the following theorem.

THEOREM 6.1. *Let $\mathbf{F}$ be a symmetric ternary function over domain $\{B, G, R\}$. Then one of the following is true:*

1. *$\mathbf{F}$ is of one of the forms in Theorem 3.1, and Holant$^*(\mathbf{F})$ is in P;*
2. *Holant$^*(\mathbf{F})$ is #P-hard;*
3. *There exists an orthogonal 3×3 matrix T such that Holant$^*(\mathbf{F})$ is polynomial time equivalent to Holant$^*(\{T^{\otimes 3}\mathbf{F}, =_{G,R}\})$.*

This theorem is proved in 3 steps by the following 3 lemmas.

LEMMA 6.1. *If $\mathbf{F}$ does not take one of the three forms in Theorem 3.1, then we can either prove that Holant$^*(\mathbf{F})$ is #P-hard or construct a binary symmetric function f from $\mathbf{F}$ by connecting a unary function to it, such that (the matrix form of) f has rank 2.*

Proof. By connecting $\mathbf{F}$ to a unary $\mathbf{u} = (x, y, z)$, we can realize $x\mathbf{F}^{1=B} + y\mathbf{F}^{1=G} + z\mathbf{F}^{1=R}$. For notational simplicity, we denote the 3×3 matrices $X = \mathbf{F}^{1=B}$, $Y = \mathbf{F}^{1=G}$ and $Z = \mathbf{F}^{1=R}$. First suppose there exists a non-zero unary $\mathbf{u}$ such that $xX + yY + zZ = 0$. If $\mathbf{u}$ is isotropic, then $\mathbf{F}$ is in the third form of Theorem 3.1. Suppose $\mathbf{u}$ is not isotropic, we may assume $\mathbf{u}^{\mathtt{T}}\mathbf{u} = 1$. Then we can apply an orthogonal transformation by a matrix whose first vector is $\mathbf{u}$, to reduce the problem to an equivalent problem in domain size 2. The dichotomy theorem for Holant* problems over domain size 2 completes the proof. The conclusion is that if $\mathbf{F}$ is not of the three forms, then Holant*($\mathbf{F}$) is #P-hard. In the following, we assume that X, Y and Z are linearly independent as complex matrices.

Now we prove the lemma by analyzing the ranks of X, Y, Z. By linear independence, X, Y, Z all have rank at least one.

- If at least one of X, Y, Z has rank 2, then we are done by choosing the corresponding coefficient to be 1 and the other two to be 0.
- If there are at least two of them (we assume they are X and Y) have rank 1, we shall prove that $X + Y$ has rank exactly 2. Firstly, the rank of $X + Y$ is at most 2 since both X and Y have rank 1. For symmetric matrices of rank 1, we can write $X = uu^{\mathtt{T}}$ and $Y = vv^{\mathtt{T}}$. We know that u and v are linearly independent, since X and Y are linearly independent. If $X + Y$ has rank at most 1, then there exists some w such that $uu^{\mathtt{T}} + vv^{\mathtt{T}} = ww^{\mathtt{T}}$. There exists a vector u' which is orthogonal to u but not to v. This can be seen by considering the dimensions of the null spaces of u and v. Then $\langle u', v\rangle v = \langle u', w\rangle w$. This implies that v is a linear multiple of w since $\langle u', v\rangle \neq 0$. Similarly, u is also a linear multiple of w. This contradicts the linear independence of u and v.
- In the remaining case, there are at least two of them (we assume they are X and Y) have rank 3. Then $\det(xX + Z) = 0$ is not a trivial equation since the coefficient of x^3 is $\det(X) \neq 0$. Let x_0 be a root for the equation. Then the rank of x_0X+Z is less than 3. If the rank is 2, then we are done. Otherwise, the rank is exactly 1; it cannot be zero since Z is not a linear multiple of X. Similarly, there exists a y_0 such that the rank of the non-zero matrix $y_0Y + Z$ is less than 3. Again, if the rank is 2, then we are done. Now we assume that both $x_0X + Z$ and $y_0Y + Z$ have rank 1. If $x_0X + Z$ and $y_0Y + Z$ are linearly independent, then x_0X+y_0Y+2Z has rank exactly 2, by the proof above, and we are done. If $x_0X + Z$ and $y_0Y + Z$ are linearly dependent, then a non-trivial combination is the zero matrix $\lambda(x_0X+Z)+\mu(y_0Y+Z) = 0$. Since they are both nonzero matrices, both $\lambda, \mu \neq 0$. Since X, Y, Z are linearly independent, we must have $x_0 = y_0 = 0$, and Z has rank 1. In this case, we consider $zX+Y$. Again we have some z_0 such that z_0X+Y has rank at most 2. If it is 2, we are done. It can't be 0, as X, Y are linearly independent. So $z_0X + Y$ has rank exactly 1. Then z_0X+Y+Z has rank exactly 2.

LEMMA 6.2. *If we can realize a rank 2 binary symmetric function A in Holant*($\mathbf{F}$), then we can either prove that $\mathbf{F}$ takes one of the forms in Theorem 3.1 and Holant*($\mathbf{F}$) is in P, or realize a rank 2 binary symmetric function whose eigenvector corresponding to the eigenvalue* 0 *is not isotropic.*

Proof. We only need to handle the case that the constructed rank 2 matrix A (binary function) has an isotropic eigenvector corresponding to 0.

Suppose A is the 3×3 matrix representing the binary function $\langle \mathbf{u}, \mathbf{F}\rangle$ for some unary function $\mathbf{u}$. By the canonical form in [41], there exists an orthogonal matrix T, such that

$$TAT^{\mathtt{T}} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & i \\ 1 & i & 0 \end{bmatrix}.$$

We may consider $T^{\otimes 3}\mathbf{F}$ instead of $\mathbf{F}$. Because $TAT^{\mathtt{T}}$ is the matrix form for $\langle T\mathbf{u}, T^{\otimes 3}\mathbf{F}\rangle$, to reuse the notation, we can assume there exists a $\mathbf{u}$, such that $\langle \mathbf{u}, \mathbf{F}\rangle$ has the matrix form $\begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & i \\ 1 & i & 0 \end{bmatrix}$. We will rename this matrix A.

Given any unary function $\mathbf{v}$ and a complex number x, we can realize the binary function $\langle x\mathbf{u} + \mathbf{v}, \mathbf{F}\rangle$ which has the matrix form $C = xA + \widetilde{A}$, where $\widetilde{A}$ is the matrix form of $\langle \mathbf{v}, \mathbf{F}\rangle$. If there exist some unary function $\mathbf{v}$ and a complex number x, such that C is nonsingular, and $\gamma = C^{-1}\begin{bmatrix} 1 \\ i \\ 0 \end{bmatrix}$ is not isotropic, then we can realize the binary symmetric function CAC of rank 2 as a chain of three binary symmetric functions, whose eigenvector corresponding to 0 is γ, and the conclusion holds.

Now, we prove that if there does not exist such $\mathbf{v}$ and x, then either Holant*($\mathbf{F}$) is in P, or we can realize a required binary function directly. We calculate the two conditions, C is singular and $\gamma = C^{-1}\begin{bmatrix} 1 \\ i \\ 0 \end{bmatrix}$ is isotropic, individually.

Suppose $\widetilde{A} = \begin{bmatrix} a & b & c \\ b & d & e \\ c & e & f \end{bmatrix}$. Then $C = xA + \widetilde{A} = \begin{bmatrix} a & b & c+x \\ b & d & e+xi \\ c+x & e+xi & f \end{bmatrix}$. Let $P(x) = \det(C)$. As a polynomial in x, $P(x)$ has degree at most 2, and the coefficient of x^2 is $a + 2bi - d$. If $a + 2bi - d \neq 0$, then for all complex x except at most two values, C is nonsingular.

Because $C\gamma = (1, i, 0)^{\mathtt{T}}$, γ is orthogonal to $\mu = (c + x, e + xi, f)$ and $\nu = (b - ai, d - bi, e - ci)$. Consider the cross-product vector $\theta = \left(\begin{vmatrix} e+xi & f \\ d-bi & e-ci \end{vmatrix}, \begin{vmatrix} f & c+x \\ e-ci & b-ai \end{vmatrix}, \begin{vmatrix} c+x & e+xi \\ b-ai & d-bi \end{vmatrix} \right)^{\mathtt{T}}$, which is orthogonal to μ and ν. Calculation shows that the inner product $\theta^{\mathtt{T}}\theta$ is a polynomial $Q(x)$ of degree at most 2, and the coefficient of x^2 is $(a + 2bi - d)^2$.

Assume $a + 2bi - d \neq 0$. Then, neither $P(x)$ nor $Q(x)$ is the zero polynomial. There exists an x such that C is nonsingular, which implies $\gamma \neq \mathbf{0}$ in particular, and $\theta^{\mathtt{T}}\theta \neq 0$. If μ and ν were linearly dependent, then $\theta = \mathbf{0}$ by the definition of θ, and $\theta^{\mathtt{T}}\theta = 0$, a contradiction. Hence, μ and ν are linearly independent. So γ is a nonzero linear multiple of θ, since they both belong to the 1-dimensional subspace orthogonal to μ and ν. Then $\gamma^{\mathtt{T}}\gamma$ is a nonzero multiple of $\theta^{\mathtt{T}}\theta \neq 0$, i.e., γ is not isotropic. Then CAC is the required function.

Now we assume that for any $\mathbf{v}$, $\widetilde{A} = \langle \mathbf{v}, \mathbf{F} \rangle$ satisfies $a + 2bi - d = 0$.

Substitute d by $a+2bi$, we get $P(x) = 2(b-ai)(e-ci)x - a(e-ci)^2 - f(b-ai)^2 + 2c(b-ai)(e-ci)$, and the coefficient of x in $Q(x)$ is $2i(e - ci)^3$.

For any fixed $\widetilde{A}$, either $e - ci = 0$, or $e - ci \neq 0$. If $e-ci \neq 0$, $Q(x)$ is not the zero polynomial. If $P(x)$ is not the zero polynomial as well, then by the same argument as above, we get a required function. Hence we assume $P(x)$ is the zero polynomial. Then by the expression for $P(x)$, it follows that $b - ai = 0$, and $a = 0$. Because we also have $a + 2bi - d = 0$, we get $a = b = d = 0$.

In this case $\widetilde{A}$ has the form $\widetilde{A} = \begin{bmatrix} 0 & 0 & c \\ 0 & 0 & e \\ c & e & f \end{bmatrix}$. It has rank ≤ 2. If it has rank ≤ 1, then $c = e = 0$. This is a contradiction to $e - ci \neq 0$. Hence it has rank 2. It is easy to check that the eigenvector corresponding to the eigenvalue 0 is a multiple of $(-e, c, 0)^{\mathtt{T}}$. If $c^2 + e^2 \neq 0$, then this eigenvector is non-isotropic and we are done. Since $e - ci \neq 0$, the only possibility of $c^2 + e^2 = 0$ is $e = -ci \neq 0$. In this case it is easy to check that $cA + \widetilde{A}$ has the form $\begin{bmatrix} 0 & 0 & 2c \\ 0 & 0 & 0 \\ 2c & 0 & f \end{bmatrix}$. It has rank 2, and a non-isotropic eigenvector $(0, 1, 0)^{\mathtt{T}}$ corresponding to the eigenvalue 0.

Finally we have for any $\widetilde{A}$, $e - ci = 0$, in addition to $d = a + 2bi$.

Consider the possible choices of $\mathbf{v}$ in $\widetilde{A} = \langle \mathbf{v}, \mathbf{F} \rangle$. We can set it to be $\mathbf{F}^{1=B}$, $\mathbf{F}^{1=G}$ or $\mathbf{F}^{1=R}$. Considering what entries a, b, c, d, e correspond to in the table (2.1) for these three cases of $\widetilde{A}$, we get the following: If $w \neq 0$, then $\mathbf{F}_{u,v,w} = i\mathbf{F}_{u+1,v-1,w}$ for $v \geq 1$ and $u + v + w = 3$. If $w = 0$, then $\mathbf{F}_{u,v,w} = \mathbf{F}_{u,v,0} = si^v + tvi^{v-1}$ for some coefficients s and t, where $u, v \geq 0$ and $u + v = 3$. This follows from $e = ci$ and $d = a + 2bi$ for $\widetilde{A}$. E.g., $e = ci$ in (2.2) gives a linear recurrence $F_{BGR} = iF_{BBR}$, and $d = a + 2bi$ in (2.2) gives a linear recurrence $F_{BGG} = 2iF_{BBG} + F_{BBB}$. Hence, $\mathbf{F} = S + T$ is the summation of two functions S and T, where $S_{u,v,w} = iS_{u+1,v-1,w}$, and $T(u, v, w) = 0$, if $w \neq 0$, and $T(u, v, 0) = tvi^{v-1}$, where $u + v + w = 3$. This T can be expressed as the symmetrization of simple tensor products,

$$\begin{aligned} T &= T_1 + T_2 + T_3 \\ &= t \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \otimes \begin{bmatrix} 1 \\ i \\ 0 \end{bmatrix} \otimes \begin{bmatrix} 1 \\ i \\ 0 \end{bmatrix} + t \begin{bmatrix} 1 \\ i \\ 0 \end{bmatrix} \otimes \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \otimes \begin{bmatrix} 1 \\ i \\ 0 \end{bmatrix} \\ &\quad + t \begin{bmatrix} 1 \\ i \\ 0 \end{bmatrix} \otimes \begin{bmatrix} 1 \\ i \\ 0 \end{bmatrix} \otimes \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \\ &= \frac{t}{2} \mathrm{Sym}(\begin{bmatrix} 1 \\ i \\ 0 \end{bmatrix} \otimes \begin{bmatrix} 1 \\ i \\ 0 \end{bmatrix} \otimes \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}). \end{aligned}$$

This is in form 3 given in Theorem 3.1 and we have shown that in this case $\text{Holant}^*(\mathbf{F})$ is tractable in Section 4.

Finally we use a holographic transformation and interpolation to get $=_{G,R}$ from the binary function obtained in Lemma 6.2. This will complete the proof of Theorem 6.1.

Let $\mathbf{v}$ be a non-isotropic eigenvector corresponding to the eigenvalue 0 of the binary function A constructed from $\mathbf{F}$. We may assume $\langle \mathbf{v}, \mathbf{v} \rangle = 1$. We can extend $\mathbf{v}$ to an orthogonal matrix T, such that $\mathbf{v}$ is the first column vector of T. Then the matrix form of the binary function after the holographic transformation by $T^{-1} = T^{\mathtt{T}}$ takes the form

$$T^{\mathtt{T}}AT = \begin{bmatrix} 0 & 0 & 0 \\ 0 & a & b \\ 0 & b & c \end{bmatrix} \tag{6.3}$$

with rank 2.

The next lemma shows that given this, we can interpolate $=_{G,R}$.

LEMMA 6.3. *Let $H : \{B, G, R\}^2 \to \mathbb{C}$ be a rank 2 binary function of the form (6.3). Then for any $\mathcal{F}$ containing H, we have*

$$\text{Holant}(\mathcal{F} \cup \{=_{G,R}\}) \leq_T \text{Holant}(\mathcal{F}).$$

Proof. Consider the Jordan normal form of H. There are two cases: there exist a non-singular $M = \mathrm{diag}(1, M_2)$, and either $\Lambda = \left[\begin{smallmatrix} 0 & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \mu \end{smallmatrix}\right]$, or $\Lambda' =$

$\begin{bmatrix} 0 & 0 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{bmatrix}$, such that $H = M\Lambda M^{-1}$, or $H = M\Lambda' M^{-1}$.

For the first case $H = M\Lambda M^{-1}$, consider an instance I of Holant($\mathcal{F} \cup \{=_{G,R}\}$). Suppose the function $=_{G,R}$ appears m times. Replace each occurrence of $=_{G,R}$ by a chain of M, $=_{G,R}$, M^{-1}. More precisely, we replace any occurrence of $=_{G,R}$ (x, y) by

$$M(x, z) \cdot (=_{G,R})(z, w) \cdot M^{-1}(w, y),$$

where z, w are new variables. This defines a new instance I'. Since $M\text{diag}(0, I_2)M^{-1} = \text{diag}(0, I_2)$, where I_2 denotes the 2×2 identity matrix, the Holant value of the instance I and I' are the same. To have a non-zero contribution to the Holant sum, the assignments given to any occurrence of the new EQUALITY constraints of the form $(=_{G,R})(z, w)$ must be (G, G) or (R, R). We can stratify the Holant sum defining the value on I' according to how many (G, G) and (R, R) assignments are given to these occurrences of $(=_{G,R})(z, w)$. Let ρ_j denote the sum, over all assignments with j many times (G, G) and $m - j$ many times (R, R), of the evaluation on I', including those of $M(x, z)$ and $M^{-1}(w, y)$. Then the Holant value on the instance I' can be written as $\sum_{j=0}^{m} \rho_j$.

Now we construct from I a sequence of instances I'_k indexed by k: Replace each occurrence of $(=_{G,R})(x, y)$ by a chain of k copies of the function H to get an instance I'_k of Holant($\mathcal{F}$). More precisely, each occurrence of $(=_{G,R})(x, y)$ is replaced by $H(x, x_1)H(x_1, x_2) \ldots H(x_{k-1}, y)$, where $x_1, x_2, \ldots, x_{k-1}$ are new variables specific for this occurrence of $(=_{G,R})(x, y)$. The function of this chain is $H^k = M\Lambda^k M^{-1}$. A moment of reflection shows that the value of the instance I'_k is

$$\sum_{j=0}^{m} \rho_j \lambda^{kj} \mu^{k(m-j)} = \mu^{mk} \sum_{j=0}^{m} \rho_j (\lambda/\mu)^{kj}.$$

If λ/μ is a root of unity, then take a k such that $(\lambda/\mu)^k = 1$. (Input size is measured by the number of variables and constraints. The functions in $\mathcal{F}$ are considered constants. Thus this k is a constant.) We have the value $\sum_{j=0}^{m} \rho_j \lambda^{kj} \mu^{k(m-j)} = \mu^{mk} \sum_{j=0}^{m} \rho_j$. As H has rank 2, $\mu \neq 0$, we can compute the value of I from the value of I'_k.

If λ/μ is not a root of unity, $(\lambda/\mu)^j$ are all distinct for $j \geq 1$. We can take $k = 1, \ldots, m+1$ and get a system of linear equations about ρ_j. Because the coefficient matrix is Vandermonde in $(\lambda/\mu)^j, j = 0, 1, \ldots m$, we can solve ρ_j and get the value of I.

For the second case $H = M\Lambda' M^{-1}$, the construction is the same, so we only show the difference with the proof in the first case. Again we can stratify the Holant sum for I' according to how many different types of assignments are given to the m occurrences of the new EQUALITY constraints of the form $(=_{G,R})(z, w)$. Any assignment other than assigning only (G, G) or (R, R) will produce a 0 contribution for I'. However, this time we cluster all assignments according to exactly j many times (G, G) *or* (R, R), and the rest $m - j$ are (G, R)'s, on all m occurrences of these $(=_{G,R})(z, w)$. Note that any assignment with a non-zero number of (R, G)'s in the corresponding m signatures in I'_k, after the substitution of each $(=_{G,R})(x, y)$ in I by $H(x, x_1)H(x_1, x_2) \ldots H(x_{k-1}, y)$, will produce a 0 contribution in the Holant value for I'_k. This is because, by this substitution, effectively each $(=_{G,R})(z, w)$ in I' is replaced by $\Lambda^k = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \lambda^k & k\lambda^{k-1} \\ 0 & 0 & \lambda^k \end{bmatrix}$. Let ρ_j be the sum over all assignments with j many (G, G) *or* (R, R), and $m - j$ many (G, R) of the evaluation (including those of $M(x, z)$ and $M^{-1}(w, y)$) on I'. Then the Holant value on the instance I' (and on I) is just ρ_m.

The value of I'_k is

$$\sum_{j=0}^{m} \rho_j \lambda^{kj} (k\lambda^{k-1})^{m-j} = \lambda^{(k-1)m} \sum_{j=0}^{m} (\lambda^j \rho_j) k^{m-j}.$$

We can take $k = 1, \ldots, m + 1$ and get a system of linear equations on $\lambda^j \rho_j$. Because the coefficient matrix is a Vandermonde matrix, we can solve $\lambda^j \rho_j$ and (since $\lambda \neq 0$ as H has rank 2) we can get the value of ρ_m, which is the value of I.

7 Reductions From Domain Size 2

THEOREM 7.1. *Let* $\mathbf{F}$ *be a symmetric ternary function over domain* $\{B, G, R\}$*, which is not of one of the forms in Theorem 3.1. Then Holant*$^*(\{\mathbf{F}, =_{G,R}\})$ *is #P-hard.*

Theorem 6.1 and 7.1 imply our main Theorem 3.1.

Using $=_{G,R}$ we can realize signatures over domain $\{G, R\}$ from $\mathbf{F}$ such as $\mathbf{F}^{*\to\{G,R\}}$. If Holant$^*(\mathbf{F}^{*\to\{G,R\}})$ is already #P-hard as a problem over the domain $\{G, R\}$ of size 2, then Holant$^*(\{\mathbf{F}, =_{G,R}\})$ is #P-hard and we are done. Therefore, we only need to deal with the cases when Holant$^*(\mathbf{F}^{*\to\{G,R\}})$ is tractable. They are listed as follows (see [15, 16]).

1. $\mathbf{F}^{*\to\{G,R\}} = H^{\otimes 3}[a, 0, 0, b]^{\mathtt{T}}$, where H is a 2×2 orthogonal matrix, $ab \neq 0$.

2. $\mathbf{F}^{*\to\{G,R\}} = Z^{\otimes 3}[a, 0, 0, b]^{\mathtt{T}}$, where $Z = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ i & -i \end{bmatrix}$, $ab \neq 0$.

3. $\mathbf{F}^{*\to\{G,R\}} = Z^{\otimes 3}[a,b,0,0]^{\mathtt{T}}$, where $Z = \frac{1}{\sqrt{2}}\begin{bmatrix}1 & 1\\ i & -i\end{bmatrix}$ or $Z = \frac{1}{\sqrt{2}}\begin{bmatrix}1 & 1\\ -i & i\end{bmatrix}$, $b \neq 0$.
4. $\mathbf{F}^{*\to\{G,R\}}$ is degenerate.

We will prove Theorem 7.1 by considering these four cases one by one. The overall proof approach for the first three cases is to construct a binary function over the domain $\{G,R\}$ such that, together with $\mathbf{F}^{*\to\{G,R\}}$ it is already #P-hard according to the dichotomy theorem for Holant* over domain size 2, Theorem 2.1. For some functions $\mathbf{F}$, we fail to do this; and whenever this happens, we show that $\mathbf{F}$ is indeed among the tractable cases in Theorem 3.1.

Definition 7.1. *A symmetric function F of arity $r \geq 2$ is said to have a domain separated form if its domain is the union of two disjoint non-empty subsets, such that the value F is 0 when not all input variables take values from the same subset.*

Because in all four listed cases $\mathbf{F}^{*\to\{G,R\}}$ has a good form, we can use domain separated holographic reductions to simplify $\mathbf{F}^{*\to\{G,R\}}$. The following is an observation about how a holographic transformation affects a function, when its basis is a block diagonal matrix.

Fact 7.1. *Suppose T is in the domain separated form separating $\{B\}$ and $\{G,R\}$, $T = \begin{pmatrix}e & 0 & 0\\ 0 & a & b\\ 0 & c & d\end{pmatrix}$. Let $M = \begin{pmatrix}a & b\\ c & d\end{pmatrix}$. We have,*

$$\begin{aligned}(T^{\otimes 3}\mathbf{F})^{*\to\{G,R\}} &= M^{\otimes 3}(\mathbf{F}^{*\to\{G,R\}}),\\ (T^{\otimes 3}\mathbf{F})^{1=B,2,3\to\{G,R\}} &= eM^{\otimes 2}(\mathbf{F}^{1=B,2,3\to\{G,R\}}),\\ (T^{\otimes 3}\mathbf{F})^{1=B,2=B,3\to\{G,R\}} &= e^2M(\mathbf{F}^{1=B,2=B,3\to\{G,R\}}).\end{aligned}$$

Proof. We prove the second formula as an example. Other formulae can be proved similarly.

$(T^{\otimes 3}\mathbf{F})^{1=B,2,3\to\{G,R\}}$ is the line $[(T^{\otimes 3}\mathbf{F})_{BGG}, (T^{\otimes 3}\mathbf{F})_{BGR}, (T^{\otimes 3}\mathbf{F})_{BRR}]$ in the triangular table form of $T^{\otimes 3}\mathbf{F}$, and $\mathbf{F}^{1=B,2,3\to\{G,R\}}$ is the corresponding line of $\mathbf{F}$.

Because one input of $T^{\otimes 3}\mathbf{F}$ is fixed to B, it is equivalent to connecting one unary function $(1,0,0)$ to $T^{\otimes 3}\mathbf{F}$. By associativity this unary can be combined with a copy of T in the gadget $T^{\otimes 3}\mathbf{F}$. This combination results in a unary function $\langle(1,0,0),T\rangle = (T_{B,B},0,0) = (e,0,0)$, which is then connected to $\mathbf{F}$. This creates a binary function $e\mathbf{F}^{1=B}$. Now, we get $(T^{\otimes 2}(e\mathbf{F}^{1=B}))^{*\to\{G,R\}}$. The two external edges of the gadget $T^{\otimes 2}(e\mathbf{F}^{1=B})$ are restricted to $\{G,R\}$. Because the domain of T is separated into $\{B\}$ and $\{G,R\}$, they force the two internal edges to take values in $\{G,R\}$. Since all 4 edges take values in $\{G,R\}$, this turns $T^{\otimes 2}(e\mathbf{F}^{1=B})$ into $eM^{\otimes 2}\mathbf{F}^{1=B,2,3\to\{G,R\}}$.

In some subcases, we reach a ternary function with a separated domain. The following lemma whose proof is omitted, takes care of these subcases.

Lemma 7.1. *If a ternary function $\mathbf{F}$ has a domain separated form then Holant$^*(\mathbf{F})$ is either #P-hard or is in one of the tractable forms of Theorem 3.1, and it is determined by the Holant* problem defined by the restriction of $\mathbf{F}$ to the separated domain of size two.*

Next we consider the first listed case for $\mathbf{F}^{*\to\{G,R\}}$.

Case 1: $\mathbf{F}^{*\to\{G,R\}} = H^{\otimes 3}[a,0,0,b]^{\mathtt{T}}$, $ab \neq 0$.

By Fact 7.1, after a domain separated holographic reduction under the orthogonal matrix $\begin{bmatrix}1 & \mathbf{0}\\ \mathbf{0} & H\end{bmatrix}$, we can assume that $\mathbf{F}^{*\to\{G,R\}} = [a,0,0,b]$, where we are given $ab \neq 0$. We note that this transformation does not change $=_{G,R}$. According to Theorem 2.1, when putting this $[a,0,0,b]$ and a binary function together, the problem is #P-hard unless the binary function is of the form $[*,0,*]$, $[0,*,0]$ or degenerate. Now $\mathbf{F}$ has the form

$$\begin{array}{ccccccc} & & & F_{BBB} & & & \\ & & F_{BBG} & & F_{BBR} & & \\ & F_{BGG} & & F_{BGR} & & F_{BRR} & \\ a & & 0 & & 0 & & b\end{array}$$

Suppose $F_{BGR} \neq 0$. We can realized a binary function $[F_{BGG}+at, F_{BGR}, F_{BRR}]$ over domain $\{G,R\}$ by connecting this ternary function to a unary function $(1,t,0)$, namely $\langle(1,t,0),\mathbf{F}\rangle$, and then putting $=_{G,R}$ on the other two dangling edges. Since $a \neq 0$ and we can choose any t, we can make the first entry of $[F_{BGG}+at, F_{BGR}, F_{BRR}]$ arbitrary and the function is out of all three tractable binary forms. Therefore the problem is #P-hard.

Now we can assume that $F_{BGR} = 0$. To simplify notations, we use variables to denote the function entries as follows

$$\begin{array}{ccccccc} & & & g & & & \\ & & y & & w & & \\ & x & & 0 & & z & \\ a & & 0 & & 0 & & b\end{array} \tag{7.4}$$

Then we use the gadget as depicted in Figure 1 to construct another binary function. The signature of this

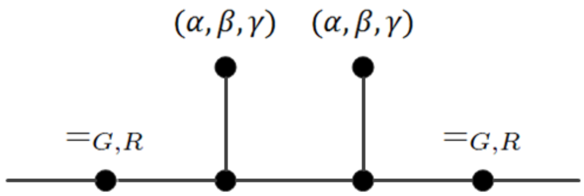

Figure 1: A binary gadget.

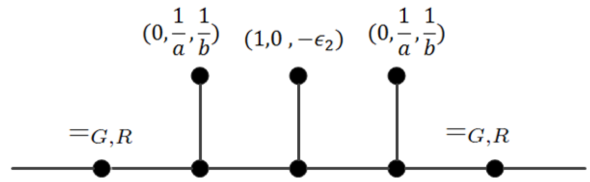

Figure 2: A binary gadget.

binary function is

$$[f_0, f_1, f_2] = [(\alpha x+\beta a)^2+(\alpha y+\beta x)^2, (\alpha y+\beta x)(\alpha w+\gamma z),$$
$$(\alpha z + \gamma b)^2 + (\alpha w + \gamma z)^2].$$

If there exists some (α, β, γ) such that this $[f_0, f_1, f_2]$ is not of the form $[*, 0, *]$, $[0, *, 0]$, or degenerate, then the problem is #P-hard and we are done.

All conditions are polynomials (1) $f_0 = f_2 = 0$, or (2) $f_1 = 0$, or (3) $f_1^2 = f_0 f_2$. By polynomial argument, we only need to solve that cases that one of them is zero polynomial.

If statement (1) $f_0 = f_2 = 0$ holds for all (α, β, γ), we have

$$(x^2 + y^2)\alpha^2 + 2(ax + xy)\alpha\beta + (a^2 + x^2)\beta^2 =$$
$$(z^2 + w^2)\alpha^2 + 2(bz + zw)\alpha\gamma + (b^2 + z^2)\gamma^2 = 0,$$

as identically zero polynomials in (α, β, γ). Therefore we have

$$x^2 + y^2 = ax + xy = a^2 + x^2 = z^2 + w^2 =$$
$$bz + zw = b^2 + z^2 = 0.$$

Since $a \neq 0$, we have $x \neq 0$ from $a^2 + x^2 = 0$. Similarly, we have $z \neq 0$. Then the conclusion is $x = \epsilon_1 a, \quad y = -a, \quad z = \epsilon_2 b, \quad w = -b$, where $\epsilon_1, \epsilon_2 \in \{i, -i\}$. Then we rewrite our function as follows

$$\begin{array}{ccccccc} & & & g & & & \\ & & -a & & -b & & \\ & \epsilon_1 a & & 0 & & \epsilon_2 b & \\ a & & 0 & & 0 & & b \end{array}$$

Next we use the gadget depicted in Figure 2 to construct another binary function over domain $\{G, R\}$, whose signature is

$$\begin{bmatrix} \epsilon_1 & 1 & 0 \\ \epsilon_2 & 0 & 1 \end{bmatrix} \begin{bmatrix} g + \epsilon_2 b & -a & 0 \\ -a & \epsilon_1 a & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \epsilon_1 & \epsilon_2 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} =$$
$$\begin{bmatrix} -g - \epsilon_1 a - \epsilon_2 b & \epsilon_1\epsilon_2(g + \epsilon_1 a + \epsilon_2 b) \\ \epsilon_1\epsilon_2(g + \epsilon_1 a + \epsilon_2 b) & -g - \epsilon_2 b \end{bmatrix}.$$

If $g + \epsilon_1 a + \epsilon_2 b \neq 0$, this symmetric binary signature can not be of the form $[*, 0, *]$ or $[0, *, 0]$, and it is not degenerate as its determinant is nonzero. Therefore the problem is #P-hard.

If $g + \epsilon_1 a + \epsilon_2 b = 0$, we show that this is indeed a tractable case in Theorem 3.1. It is of the second form in Theorem 3.1 where $\boldsymbol{\alpha} = (0, 0, 0)^{\mathtt{T}}, \boldsymbol{\beta_1} = \sqrt[3]{a}(\epsilon_1, 1, 0)^{\mathtt{T}}$ and $\boldsymbol{\beta_2} = \sqrt[3]{b}(\epsilon_2, 0, 1)^{\mathtt{T}}$.

If statement (2) $f_1 = 0$ holds for all (α, β, γ), we have $x = y = 0$ or $z = w = 0$. If $x = y = 0$, the ternary function (7.4) is as follows

$$\begin{array}{ccccccc} & & & g & & & \\ & & 0 & & w & & \\ & 0 & & 0 & & z & \\ a & & 0 & & 0 & & b \end{array}$$

Then G is separated from B-R, and by Lemma 7.1, we are done. The case $z = w = 0$ is similar.

If statement (3) $f_1^2 = f_0 f_2$ holds for all (α, β, γ), we have

$$(7.5) \quad (\alpha x + \beta a)^2(\alpha z + \gamma b)^2 + (\alpha x + \beta a)^2(\alpha w + \gamma z)^2+$$
$$(\alpha y + \beta x)^2(\alpha z + \gamma b)^2 = 0.$$

Let $\alpha = a$ and $\beta = -x$, we have $(ay - x^2)^2(az + \gamma b)^2 = 0$ holds for all γ. Since $b \neq 0$, we can choose γ such that $az + \gamma b \neq 0$ and conclude that $ay - x^2 = 0$. Similarly, let $\alpha = b$ and $\gamma = -z$, we can get $bw - z^2 = 0$. Then let $\beta = \gamma = 1$ and $\alpha = 0$ in (7.5), we have

$$a^2 b^2 + a^2 z^2 + b^2 x^2 = 0.$$

Denote by $p = \frac{x}{a}$ and $q = \frac{z}{b}$, we have $p^2 + q^2 + 1 = 0$ and the ternary signature in (7.4) has the following form

$$\begin{array}{ccccccc} & & & g & & & \\ & & ap^2 & & bq^2 & & \\ & ap & & 0 & & bq & \\ a & & 0 & & 0 & & b \end{array}$$

If $p = 0$ or $q = 0$, then the function is separable and we are done by Lemma 7.1. In the following, we assume that $pq \neq 0$.

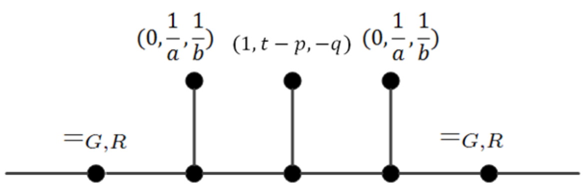

Figure 3: A binary gadget.

Then we use the gadget in Figure 3 to construct another binary function over domain $\{G, R\}$, whose signature is

$$\begin{bmatrix} p & 1 & 0 \\ q & 0 & 1 \end{bmatrix} \begin{bmatrix} g - bq^3 - ap^3 + ap^2t & apt & 0 \\ apt & at & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} p & q \\ 1 & 0 \\ 0 & 1 \end{bmatrix} =$$

$$\begin{bmatrix} p^2\delta + at(p^2+1)^2 & pq\delta + apqt(p^2+1) \\ pq\delta + apqt(p^2+1) & q^2\delta + ap^2q^2t \end{bmatrix},$$

where $\delta = g - ap^3 - bq^3$. We denote this symmetric binary function as $[g_0, g_1, g_2]$.

If $\delta = 0$, one can verify that this is indeed a tractable case of Theorem 3.1. This is of the third form of Theorem 3.1, where $\boldsymbol{\beta} = (1, -p, -q)^{\mathtt{T}}, \boldsymbol{\gamma} = (0, 0, 0)^{\mathtt{T}}$, and $\mathbf{F}_\beta$ is the given function $\mathbf{F}$.

Now we assume that $\delta \neq 0$. If there exists some t such that this binary function is not of the form $[*, 0, *]$, $[0, *, 0]$, or degenerate, then the problem is #P-hard and we are done. Otherwise, by the same argument as above, at least one of the three statements (i) $g_0 = g_2 = 0$, (ii) $g_1 = 0$, or (iii) $g_1^2 = g_0 g_2$ holds for all t. Choose $t = 0$, we have all three $g_0, g_1, g_2 \neq 0$. Therefore, the only possibility is that $g_1^2 = g_0 g_2$ holds for all t. However, this is also impossible which can be seen by choosing $t = \frac{1}{a}$. One can calculate the determinant $\det \begin{bmatrix} g_0 & g_1 \\ g_1 & g_2 \end{bmatrix} = \delta q^2 \neq 0$. This completes the proof for the case $\mathbf{F}^{*=\{G,R\}} = H[a, 0, 0, b]^{\mathtt{T}}$.

We omit the proof of the next two cases, (Case 2: $\mathbf{F}^{*\to\{G,R\}} = Z^{\otimes 3}[a, 0, 0, b]^{\mathtt{T}}$, and Case 3: $\mathbf{F}^{*\to\{G,R\}} = Z^{\otimes 3}[a, b, 0, 0]^{\mathtt{T}}$. See [17]) Case 4: $\mathbf{F}^{*\to\{G,R\}}$ is degenerate, turns out to be very difficult and requires a different proof technique. We will omit most of the proof of Case 4, except the first lemma of the second subcase in this Case 4.

For Case 4, where $\mathbf{F}^{*\to\{G,R\}}$ is degenerate, our approach is more complicated. If we can construct a gadget which realizes a symmetric function of arity 3 whose restriction to domain $\{G, R\}$ is non-degenerate, then we can go on with the proof to construct some proper binary gadget where the idea is similar to previous cases. We don't need such a construction in the first 3 cases, because $\mathbf{F}^{*\to\{G,R\}}$ is already good enough and F itself can take this task. Unfortunately, we did not find such a gadget that can be proved to work.

We find some unsymmetric gadget of arity 4 as a replacement. Then we have to utilize the complexity dichotomy theorem in [16], which is a generalization of Theorem 2.1 to function sets that are not necessary symmetric.

The first tractable class from [16] is denoted by $\langle \mathcal{T} \rangle$. $\mathcal{T}$ is the set of all functions of arity at most 2. We say a function set $\mathcal{F}$ is closed under tensor product, if for any $\mathbf{A}, \mathbf{B} \in \mathcal{F}$, $\mathbf{A} \otimes \mathbf{B} \in \mathcal{F}$. The tensor closure $\langle \mathcal{F} \rangle$ of a set $\mathcal{F}$ is the minimum set containing $\mathcal{F}$, closed under tensor product.

Our unsymmetric gadget will be good enough, so that its restriction to domain $\{G, R\}$ is not in $\langle \mathcal{T} \rangle$. Together with some proper binary gadget, we can get the hardness from the general complexity dichotomy theorem from [16].

Case 4: $\mathbf{F}^{*\to\{G,R\}}$ is degenerate.

Suppose $\mathbf{F}^{*\to\{G,R\}} = (a, b)^{\otimes 3}$. The first subcase is that (a, b) is not isotropic, which we omit here. Consider the second subcase that (a, b) is isotropic.

We only need to prove the case when $\mathbf{F}^{*\to\{G,R\}}$ is $[1, i, -1, -i]$, because we can perform an orthogonal domain separated holographic reduction. Let $\mathbf{F} = [u; t, r; s, p, q; 1, i, -1, -i]$, namely

$$\begin{array}{ccccccc} & & & u & & & \\ & & t & & r & & \\ & s & & p & & q & \\ 1 & & i & & -1 & & -i \end{array} \tag{7.6}$$

Suppose $\mathbf{T} = \langle (\alpha, \beta, \gamma), \mathbf{F} \rangle = \alpha \mathbf{A} + \beta \mathbf{B} + \gamma \mathbf{C}$, where $\mathbf{A} = \mathbf{F}^{1=B}$, $\mathbf{B} = \mathbf{F}^{1=G}$ and $\mathbf{C} = \mathbf{F}^{1=R}$. Construct

$$\mathbf{H}(x_1, x_2, x_3, x_4) =$$

$$\left(\sum_{\substack{y_1,\, y_2 \,\in \\ \{B,\, G,\, R\}}} \mathbf{F}(x_1, x_2, y_1) \mathbf{T}(y_1, y_2) \mathbf{F}(y_2, x_3, x_4) \right)^{*\to\{G,R\}}.$$

This H has some good properties, which can be utilized to prove that it is not in $\langle \mathcal{T} \rangle$.

FACT 7.2. *Suppose* $\mathbf{F}$ *satisfies* $\mathbf{H}(x_1, x_2, x_3, x_4) = \mathbf{H}(x_2, x_1, x_3, x_4) = \mathbf{H}(x_1, x_2, x_4, x_3) = \mathbf{H}(x_2, x_1, x_4, x_3)$. *If* $\mathbf{F} \in \langle \mathcal{T} \rangle$, *then there are binary functions* $\mathbf{A}, \mathbf{B}$, *either* $\mathbf{H} = \mathbf{A}(x_1, x_2)\mathbf{B}(x_3, x_4)$ *or* $\mathbf{H} = \mathbf{A}(x_1, x_3)\mathbf{B}(x_2, x_4)$.

We omit the proof of this fact.

Let $\mathbf{S} = \begin{bmatrix} s & 1 & i \\ p & i & -1 \\ p & i & -1 \\ q & -1 & -i \end{bmatrix}$, indexed by $\{G,R\}^2 \times \{B,G,R\}$ in lexicographic order. Then the arity 4 function $\mathbf{H}$ has a matrix form $\mathbf{H}_1 = \mathbf{S}\mathbf{T}\mathbf{S}^{\mathtt{T}}$, where the rows are indexed by $(x_1,x_2) \in \{G,R\}^2$ and the columns are indexed by $(x_3,x_4) \in \{G,R\}^2$. The other matrix form of $\mathbf{H}$ is $\mathbf{H}_2$ indexed by (x_1,x_3) and (x_2,x_4). $\mathbf{H}$ has a decomposed form $\mathbf{K}(x_1,x_2)\mathbf{L}(x_3,x_4)$ (respectively $\mathbf{K}(x_1,x_3)\mathbf{L}(x_2,x_4)$) iff $\mathbf{H}_1$ (resp. $\mathbf{H}_2$) has rank at most one.

Let $\mathbf{P} = \begin{bmatrix} s & 1 \\ p & i \\ p & i \\ q & -1 \end{bmatrix}$ and $\mathbf{Q} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & i \end{pmatrix}$. Then, $\mathbf{S} = \mathbf{P}\mathbf{Q}$. By associativity, we can multiply $\mathbf{Q}\mathbf{T}\mathbf{Q}^{\mathtt{T}}$ first in $\mathbf{H}_1 = \mathbf{P}\mathbf{Q}\mathbf{T}\mathbf{Q}^{\mathtt{T}}\mathbf{P}^{\mathtt{T}}$.

We have

$$\mathbf{Q}\mathbf{A}\mathbf{Q}^{\mathtt{T}} = \begin{pmatrix} u & t+ir \\ t+ir & s+2ip-q \end{pmatrix},$$

$$\mathbf{Q}\mathbf{B}\mathbf{Q}^{\mathtt{T}} = \begin{pmatrix} t & s+ip \\ s+ip & 0 \end{pmatrix},$$

$$\mathbf{Q}\mathbf{C}\mathbf{Q}^{\mathtt{T}} = \begin{pmatrix} r & p+iq \\ p+iq & 0 \end{pmatrix},$$

and in symmetric signature notation $\mathbf{Q}\mathbf{T}\mathbf{Q}^{\mathtt{T}}$ is

$$[u\alpha+t\beta+r\gamma, (t+ir)\alpha+(s+ip)\beta+(p+iq)\gamma, (s+2ip-q)\alpha].$$

LEMMA 7.2. *If $p \neq is$ or $q \neq ip$, then there exist some α, β, γ, such that $\mathbf{H} \notin \langle \mathcal{T} \rangle$.*

Proof. The proofs under both conditions are the same. W.l.o.g. we assume $p \neq is$. The proof is composed of three steps. We will use different matrix or vector representations of $\mathbf{H}$.

In the first step, we use the matrix form $\mathbf{H}_1 = \mathbf{P}(\mathbf{Q}\mathbf{T}\mathbf{Q}^{\mathtt{T}})\mathbf{P}^{\mathtt{T}}$ of $\mathbf{H}$, and show that for some α, β, γ, this matrix has rank at least 2.

The submatrix $\begin{pmatrix} s & 1 \\ p & i \end{pmatrix}$ of $\mathbf{P}$ has full rank. We only need to show that $\begin{pmatrix} s & 1 \\ p & i \end{pmatrix}(\mathbf{Q}\mathbf{T}\mathbf{Q}^{\mathtt{T}})\begin{pmatrix} s & 1 \\ p & i \end{pmatrix}^{\mathtt{T}}$, a 2×2 submatrix of $\mathbf{H}_1$, is of full rank. $\det(\mathbf{Q}\mathbf{T}\mathbf{Q}^{\mathtt{T}})$ is a polynomial whose coefficient of β^2 is the nonzero number $-(s+ip)^2$. For any fixed α and γ, there are 3 different values $c_1, c_2, (c_1+c_2)/2$ (these may depend on α, γ), such that when β takes any one of these values, $\det(\mathbf{Q}\mathbf{T}\mathbf{Q}^{\mathtt{T}}) \neq 0$ and consequently $\mathbf{H}_1$ has rank at least 2.

We attack the rank of $\mathbf{H}_2$ twice in following steps. Either we already get rank at least 2 in the second step, or we get some conditions utilized in the third step. In the second step, we consider the matrix form $\mathbf{H}_2 = \mathbf{H}_{x_1x_3,\ x_2x_4}$. If at least one of the three matrices $\mathbf{H}_2$ given by (α, c_1, γ), (α, c_2, γ) and $(\alpha, (c_1+c_2)/2, \gamma)$ has rank at least two, by Fact 7.2, $\mathbf{H} \notin \langle \mathcal{T} \rangle$.

Now suppose all three matrices $\mathbf{H}_2$ given by (α, c_1, γ), (α, c_2, γ) and $(\alpha, (c_1+c_2)/2, \gamma)$ have rank at most 1. By the conclusion $\det(\mathbf{Q}\mathbf{T}\mathbf{Q}^{\mathtt{T}}) \neq 0$ in the first step, the three ranks are exactly 1.

Let the matrices $\mathbf{H}_2$ given by (α, c_1, γ) and (α, c_2, γ) be $\mathbf{u}\mathbf{u}^{\mathtt{T}}$ and $\mathbf{v}\mathbf{v}^{\mathtt{T}}$ for some column vectors $\mathbf{u}$ and $\mathbf{v}$. Then the matrix $\mathbf{H}_2$ given by $(\alpha, (c_1+c_2)/2, \gamma)$ is $(\mathbf{u}\mathbf{u}^{\mathtt{T}} + \mathbf{v}\mathbf{v}^{\mathtt{T}})/2$. If $\mathbf{u}$ and $\mathbf{v}$ are linearly independent, then $(\mathbf{u}\mathbf{u}^{\mathtt{T}} + \mathbf{v}\mathbf{v}^{\mathtt{T}})/2$ has rank 2. (It certainly has rank at most two, as its image is contained in the span of $\{\mathbf{u}, \mathbf{v}\}$. By linear independence, there are $\mathbf{w}$ satisfying $\mathbf{u}^{\mathtt{T}}\mathbf{w} = 0$ but $\mathbf{v}^{\mathtt{T}}\mathbf{w} \neq 0$. Thus the image contains $\mathbf{v}$, and similarly it also contains $\mathbf{u}$.) Hence $\mathbf{u}$ and $\mathbf{v}$ are linearly dependent. It follows that the matrices $\mathbf{u}\mathbf{u}^{\mathtt{T}}$ and $\mathbf{v}\mathbf{v}^{\mathtt{T}}$ are also linearly dependent. This is so when we write these two matrices as vectors.

We use the vector form $\mathbf{H}_3$ of $\mathbf{H}$ to show the consequence of this observation. This form helps to explain getting rid of $\mathbf{P}$ and $\mathbf{P}^{\mathtt{T}}$. Let $\widetilde{\mathbf{A}}$ denote the column vector form of $\mathbf{Q}\mathbf{A}\mathbf{Q}^{\mathtt{T}}$, namely $\widetilde{\mathbf{A}} = (u, t+ir, t+ir, s+2ip-q)^{\mathtt{T}}$. Similarly, let $\widetilde{\mathbf{B}} = (t, s+ip, s+ip, 0)^{\mathtt{T}}$ and $\widetilde{\mathbf{C}} = (r, p+iq, p+iq, 0)^{\mathtt{T}}$ be the column vector forms of $\mathbf{Q}\mathbf{B}\mathbf{Q}^{\mathtt{T}}$ and $\mathbf{Q}\mathbf{C}\mathbf{Q}^{\mathtt{T}}$, respectively. Then $\mathbf{H}_3 = \mathbf{P}^{\otimes 2}(\alpha\widetilde{\mathbf{A}} + \beta\widetilde{\mathbf{B}} + \gamma\widetilde{\mathbf{C}})$, which lists all entries of $\mathbf{H}$, and therefore also all entries of $\mathbf{H}_2$, in some order. Notice that the submatrix $\begin{pmatrix} s & 1 \\ p & i \end{pmatrix}^{\otimes 2}$ of $\mathbf{P}^{\otimes 2}$ is of full rank. Let $\alpha = 1$ and $\gamma = 0$, we get $\widetilde{\mathbf{A}} + c_1\widetilde{\mathbf{B}}$ and $\widetilde{\mathbf{A}} + c_2\widetilde{\mathbf{B}}$ are linearly dependent, where $c_1 \neq c_2$. It follows that $\widetilde{\mathbf{A}}$ and $\widetilde{\mathbf{B}}$ are linearly dependent. Because the entry $s+ip$ in $\widetilde{\mathbf{B}}$ is nonzero, $\widetilde{\mathbf{A}}$ is a multiple of $\widetilde{\mathbf{B}}$, and $s+2ip-q = 0$ as the corresponding entry in $\widetilde{\mathbf{B}}$ is 0. This is just $(s+ip) + i(p+iq) = 0$. Hence we have $p+iq = i(s+ip) \neq 0$.

In the third step, we fix $\alpha = 0, \beta = 1, \gamma = 0$. Obviously, $\mathbf{H}_1$ has rank at least 2, since $\det(\mathbf{Q}\mathbf{B}\mathbf{Q}^{\mathtt{T}}) = -(s+ip)^2 \neq 0$. We consider $\mathbf{H}_2$. Since the matrix $\begin{pmatrix} s & 1 \\ p & i \end{pmatrix}$ has rank 2, and $\begin{pmatrix} t \\ s+ip \end{pmatrix}$ is a nonzero vector, we have either $\begin{pmatrix} s & 1 \end{pmatrix}\begin{pmatrix} t \\ s+ip \end{pmatrix} \neq 0$ or $\begin{pmatrix} p & i \end{pmatrix}\begin{pmatrix} t \\ s+ip \end{pmatrix} \neq 0$.

Suppose the first is not zero. Consider the $(GG, GR) \times (GG, GR)$ submatrix of $\mathbf{H}_2$, whose row index is by x_1x_3 and the column index is by x_2x_4. They are precisely the entries in the first row $(GG, GG), (GG, GR), (GG, RG)$ and (GG, RR) of $\mathbf{H}_1$. Recall that $\mathbf{H}_1 = \mathbf{P}(\mathbf{Q}\mathbf{T}\mathbf{Q}^{\mathtt{T}})\mathbf{P}^{\mathtt{T}} = \mathbf{P}(\mathbf{Q}\mathbf{B}\mathbf{Q}^{\mathtt{T}})\mathbf{P}^{\mathtt{T}}$, after our choice $\alpha = 0, \beta = 1, \gamma = 0$. The first row of $\mathbf{P}$ is $\begin{pmatrix} s & 1 \end{pmatrix}$. Let $\begin{pmatrix} a & b \end{pmatrix} = \begin{pmatrix} s & 1 \end{pmatrix} \begin{pmatrix} t & s+ip \\ s+ip & 0 \end{pmatrix}$. Then the first row of $\mathbf{H}_1$ is

$$\begin{pmatrix} s & 1 \end{pmatrix} \mathbf{Q}\mathbf{B}\mathbf{Q}^{\mathtt{T}}\mathbf{P}^{\mathtt{T}} =$$

$$\begin{pmatrix} s & 1 \end{pmatrix} \begin{pmatrix} t & s+ip \\ s+ip & 0 \end{pmatrix} \begin{pmatrix} s & p & p & q \\ 1 & i & i & -1 \end{pmatrix}$$

$$= \begin{pmatrix} a & b \end{pmatrix} \begin{pmatrix} s & p & p & q \\ 1 & i & i & -1 \end{pmatrix}.$$

Because $s + 2ip - q = 0$, which we proved in the second step, we have the linear dependence $\begin{pmatrix} s \\ 1 \end{pmatrix} + 2i \begin{pmatrix} p \\ i \end{pmatrix} - \begin{pmatrix} q \\ -1 \end{pmatrix} = 0$. Therefore the four entries in the first row of $\mathbf{H}_1$ are $(k, l, l, k+2il)$, where $k = as+b$ and $l = ap+bi$. If the submatrix of $\mathbf{H}_2$ indexed by $(GG, GR)\times(GG, GR)$ is not of full rank, then $l^2 = k(k + 2il)$, which is $(l - ik)^2 = 0$. Hence $l = ik$. It follows that $ap = ias$. Notice that $a = \begin{pmatrix} s & 1 \end{pmatrix} \begin{pmatrix} t \\ s+ip \end{pmatrix} \neq 0$. We get $p = is$, a contradiction.

If $\begin{pmatrix} p & i \end{pmatrix} \begin{pmatrix} t \\ s+ip \end{pmatrix} \neq 0$, the proof is similar. Consider the $(GG, GR) \times (RG, RR)$ submatrix of $\mathbf{H}_2$ indexed by x_1x_3 and x_2x_4. They are precisely the second row entries $(GR, GG), (GR, GR), (GR, RG), (GR, RR)$ of $\mathbf{H}_1$. The rest of the proof of Lemma 7.2 is the same as the previous case.

The full proof can be found in [17].

References

[1] Ali Al-Bashabsheh and Yongyi Mao. Normal factor graphs and holographic transformations. *IEEE Transactions on Information Theory*, 57(2):752–763, 2011.

[2] Ali Al-Bashabsheh, Yongyi Mao, and Pascal O. Vontobel. Normal factor graphs: A diagrammatic approach to linear algebra. In Alexander Kuleshov, Vladimir Blinovsky, and Anthony Ephremides, editors, *ISIT*, pages 2178–2182. IEEE, 2011.

[3] R.J. Baxter. *Exactly solved models in statistical mechanics*. Academic press London, 1982.

[4] A. Bulatov, M. Dyer, L.A. Goldberg, M. Jalsenius, and D. Richerby. The complexity of weighted boolean #CSP with mixed signs. *Theoretical Computer Science*, 410(38-40):3949–3961, 2009.

[5] Andrei A. Bulatov. A dichotomy theorem for constraint satisfaction problems on a 3-element set. *J. ACM*, 53(1):66–120, 2006.

[6] Andrei A. Bulatov. The complexity of the counting constraint satisfaction problem. In Luca Aceto, Ivan Damgård, Leslie Ann Goldberg, Magnús M. Halldórsson, Anna Ingólfsdóttir, and Igor Walukiewicz, editors, *ICALP (1)*, volume 5125 of *Lecture Notes in Computer Science*, pages 646–661. Springer, 2008.

[7] Andrei A. Bulatov and Víctor Dalmau. Towards a dichotomy theorem for the counting constraint satisfaction problem. In *FOCS*, pages 562–571. IEEE Computer Society, 2003.

[8] Andrei A. Bulatov and Martin Grohe. The complexity of partition functions. *Theor. Comput. Sci.*, 348(2-3):148–186, 2005.

[9] Jin-Yi Cai and Xi Chen. Complexity of counting CSP with complex weights. STOC 2012: 909-920. http://arxiv.org/abs/1111.2384

[10] Jin-Yi Cai, Xi Chen, and Pinyan Lu. Graph homomorphisms with complex values: A dichotomy theorem. In Samson Abramsky, Cyril Gavoille, Claude Kirchner, Friedhelm Meyer auf der Heide, and Paul G. Spirakis, editors, *ICALP (1)*, volume 6198 of *Lecture Notes in Computer Science*, pages 275–286. Springer, 2010.

[11] Jin-Yi Cai, Xi Chen, and Pinyan Lu. Non-negatively weighted #CSP: An effective complexity dichotomy. In *IEEE Conference on Computational Complexity*, pages 45–54, 2011.

[12] Jin-Yi Cai, Sangxia Huang, and Pinyan Lu. From holant to #CSP and back: Dichotomy for holant problems. In Otfried Cheong, Kyung-Yong Chwa, and Kunsoo Park, editors, *ISAAC (1)*, volume 6506 of *Lecture Notes in Computer Science*, pages 253–265. Springer, 2010.

[13] Jin-Yi Cai, Pinyan Lu, and Mingji Xia. Holographic algorithms by fibonacci gates and holographic reductions for hardness. In *FOCS '08: Proceedings of the 49th Annual IEEE Symposium on Foundations of Computer Science*, Washington, DC, USA, 2008. IEEE Computer Society.

[14] Jin-Yi Cai, Pinyan Lu, and Mingji Xia. Holant problems and counting CSP. In Michael Mitzenmacher, editor, *STOC*, pages 715–724. ACM, 2009.

[15] Jin-Yi Cai, Pinyan Lu, and Mingji Xia. Computational complexity of holant problems. *SIAM J. Comput.*, 40(4):1101–1132, 2011.

[16] Jin-Yi Cai, Pinyan Lu, and Mingji Xia. Dichotomy for holant* problems of boolean domain. In *SODA '11: Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms*, 2011.

[17] Jin-Yi Cai, Pinyan Lu, and Mingji Xia. Dichotomy for Holant* Problems with a Function on Domain Size 3. http://arxiv.org/pdf/1207.2354v1.pdf

[18] Nadia Creignou and Miki Hermann. Complexity of generalized satisfiability counting problems. *Inf. Comput.*, 125(1):1–12, 1996.

[19] C. T. J. Dodson and T. Poston. *Tensor Geometry.* Graduate Texts in Mathematics 130. Springer-Verlag, New York, 1991.

[20] Martin E. Dyer, Leslie Ann Goldberg, and Mark

Jerrum. The complexity of weighted boolean csp. *SIAM J. Comput.*, 38(5):1970–1986, 2009.

[21] M.E. Dyer and C. Greenhill. The complexity of counting graph homomorphisms. In *Proceedings of the 9th International Conference on Random Structures and Algorithms*, pages 260–289, 2000.

[22] M.E. Dyer and D.M. Richerby. On the complexity of #CSP. In *Proceedings of the 42nd ACM symposium on Theory of computing*, pages 725–734, 2010.

[23] M.E. Dyer and D.M. Richerby. The #CSP dichotomy is decidable. In *Proceedings of the 28th Symposium on Theoretical Aspects of Computer Science*, 2011.

[24] T. Feder and M.Y. Vardi. The computational structure of monotone monadic SNP and constraint satisfaction: A study through Datalog and group theory. *SIAM Journal on Computing*, 28(1):57–104, 1998.

[25] R.P. Feynman. *Feynman lectures on physics.* Addison Wesley Longman, 1970.

[26] M. Freedman, L. Lovász, and A. Schrijver. Reflection positivity, rank connectivity, and homomorphism of graphs. *J. AMS*, 20:37–51, 2007.

[27] Leslie Ann Goldberg, Martin Grohe, Mark Jerrum, and Marc Thurley. A complexity dichotomy for partition functions with mixed signs. *SIAM J. Comput.*, 39(7):3336–3402, 2010.

[28] Leslie Ann Goldberg and Mark Jerrum. Approximating the partition function of the ferromagnetic potts model. In *Proceedings of the 37th international colloquium conference on Automata, languages and programming*, ICALP '10, pages 396–407, Berlin, Heidelberg, 2010. Springer-Verlag.

[29] Leslie Ann Goldberg, Mark Jerrum, and Mike Paterson. The computational complexity of two-state spin systems. *Random Struct. Algorithms*, 23(2):133–154, 2003.

[30] Heng Guo, Pinyan Lu, and Leslie G. Valiant. The complexity of symmetric boolean parity holant problems - (extended abstract). In Luca Aceto, Monika Henzinger, and Jiri Sgall, editors, *ICALP (1)*, volume 6755 of *Lecture Notes in Computer Science*, pages 712–723. Springer, 2011.

[31] Sangxia Huang and Pinyan Lu. A Dichotomy for Real Weighted Holant Problems. IEEE Conference on Computational Complexity 2012: 96-106.

[32] E. Ising. Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik A Hadrons and Nuclei*, 31(1):253–258, 1925.

[33] Mark Jerrum and Alistair Sinclair. Polynomial-time approximation algorithms for the ising model. *SIAM Journal on Computing*, 22(5):1087–1116, 1993.

[34] G. David Forney Jr. Codes on graphs: Normal realizations. *IEEE Transactions on Information Theory*, 47(2):520–548, 2001.

[35] G. David Forney Jr. and Pascal O. Vontobel. Partition functions of normal factor graphs. *CoRR*, abs/1102.0316, 2011.

[36] P. W. Kasteleyn. The statistics of dimers on a lattice. *Physica*, 27:1209–1225, 1961.

[37] Richard E. Ladner. On the structure of polynomial time reducibility. *J. ACM*, 22(1):155–171, 1975.

[38] L. Lovász. Operations with structures. *Acta Math. Hung.*, 18:321–328, 1967.

[39] B.M. McCoy and T.T. Wu. *The two-dimensional Ising model*. Harvard University Press Cambridge, 1973.

[40] T.J. Schaefer. The complexity of satisfiability problems. In *Proceedings of the 10th annual ACM symposium on Theory of computing*, pages 216–226, 1978.

[41] N. H. Scott. A new canonical form for complex symmetric matrices. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 441(1913):625–640, 1993.

[42] H. N. V. Temperley and M. E. Fisher. Dimer problem in statistical mechanics c an exact result. *Philosophical Magazine*, 6:1061C 1063, 1961.

[43] Leslie G. Valiant. Accidental algorithms. In *FOCS '06: Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, pages 509–517, Washington, DC, USA, 2006. IEEE Computer Society.

[44] Leslie G. Valiant. Holographic algorithms. *SIAM J. Comput.*, 37(5):1565–1594, 2008.