Understanding the Mulmuley-Sohoni Approach to P vs. NP

Kenneth W. Regan*
University at Buffalo

Abstract

We explain the essence of K. Mulmuley and M. Sohoni, "Geometric Complexity Theory I: An Approach to the P vs. NP and Related Problems" [MS02] for a general complexity-theory audience. We evaluate the power and prospects of the new approach. The emphasis is not on probing the deep mathematics that underlies this work, but rather on helping computational complexity theorists not versed in its background to understand the combinatorics involved.

1 Introduction

Consider a group G of $n \times n$ matrices A and a vector v in an n-dimensional vector space V over a field F. The *orbit* Gv is the set of images Av over all $A \in G$. We ask, how "nice" a subset of V does Gv form? Does it swing arbitrarily close to zero? If not, and/or if it remains nice when we replace V by the *projective space* V^*—then one says that the action of G on v is *stable*. We can ask similar questions for other kinds of *group actions* $\alpha : G \times V \to V$ besides $A, v \mapsto Av$, including cases where the dimension of V is much larger than n.

Stability is informally a notion of not being "chaotic," and has developed into a major branch of algebraic geometry under the guiding influence of D.A. Mumford among others. Ketan Mulmuley and Milind Sohoni [MS02] observe that many questions about complexity classes can be re-cast as questions about the nature of group actions on certain vectors in certain spaces that encode problems in these classes. This survey explains their framework from a lay point of view, and attempts to evaluate whether this approach truly adds new power to attacks on the P. vs. NP question.

2 Key geometric and algebraic concepts

First we need to define "nice." Given a field F, take V to be the finite-dimensional vector space F^n. Also write $F[x_1, \ldots, x_n]$ for the *ring* of polynomials in n variables with coefficients in F. The solution space S of a finite set of polynomial equations $p_1(x_1, \ldots, x_n) = 0, \ldots p_s(x_1, \ldots, x_n) = 0$ is then a subset of V, and we call S a *basic closed set*. Finite unions and arbitrary intersections of basic closed sets form the *closed sets* of the

*Supported in part by NSF grant CCR-9821040. regan@cse.buffalo.edu

Zariski topology on V. Their complements are Zariski-*open*. Proper Z(ariski)-closed subsets of V are "nice" *algebraic sets*. Halfway nice are subsets that are the intersection of a closed set and an open set—these are called *locally closed*. Finite unions of locally-closed sets are the same as all finite Boolean combinations of closed sets—these are called *constructible sets*.

Every Z-closed set is closed in the familiar "Euclidean" topology, but not conversely, because every Z-closed set other than V itself is at most $(n-1)$-dimensional. Thus every Z-closed set other than V has measure zero on V, and being Z-open gives a particularly strong notion of "almost everywhere." Just as in the Euclidean topology, every set $R \subseteq V$ has a *closure* $\bar{R}$ in the Zariski topology, defined to be the intersection of all Z-closed sets that contain R. It is also definable as $\bar{R} = \mathcal{V}(\mathcal{I}(R))$, where $\mathcal{I}(R)$ is the set of all n-ary polynomials that vanish on R, and $\mathcal{V}(\mathcal{I})$ means the set of common zeroes of those polynomials. Here $\mathcal{I} = \mathcal{I}(R)$ forms an *ideal*, meaning that for all $p, q \in \mathcal{I}$ and arbitrary polynomial α, $\alpha p + q \in \mathcal{I}$. Not all ideals have the form $\mathcal{I}(R)$ for some R—those that do are *radical*, meaning that whenever some power p^e of a polynomial p belongs to $\mathcal{I}$, p itself belongs to $\mathcal{I}$. *Hilbert's Nullstellensatz* says that if the field F is algebraically closed, then there is a 1-1 correspondence between Z-closed subsets of V and radical ideals in $F[x_1, \ldots, x_n]$. Every ideal of polynomials I is *finitely generated*, meaning that there exist $p_1, \ldots, p_s \in \mathcal{I}$ such that $\mathcal{I}$ comprises all the "algebraic consequences" $\alpha_1 p_1 + \ldots + \alpha_s p_s$ over all polynomials $\alpha_1, \ldots, \alpha_s$ of the equations p_i. The Nullstellensatz implies that the equations are unsolvable over the algebraically closed field iff 1 is a consequence. The Z-closure of R may balloon out a lot further than its Euclidean closure—for instance, any "open ball" (of full dimension) in the Euclidean topology closes out to all of V in the Zariski topology.

The Zariski topology may also be defined on *projective space*, and there it has an important special-case connection with the familiar topology. Define $P(V)$ to be the space of equivalence classes of V under scalar multiplication—i.e., of one-dimensional linear subspaces. This is the *projective space* associated to V. If each polynomial p_i in $\{p_1, \ldots, p_s\}$ is *homogeneous*, meaning that all its monomials in unfactored form have the same degree d_i (which can be different for different i), then their solution space S is invariant under scalar multiples. Thus S becomes a subset of projective space (ignoring the zero vector), and is a closed set in the Zariski topology on $P(V)$. The Z-closure of an arbitrary $R \subseteq P(V)$ is defined analogously. The key fact is that when V is a vector space over an algebraically closed field, and S is a *constructible* subset of $P(V)$, *the Z-closure of S coincides with its classical closure* (see [Mum76]).

2.1 Group actions

A *group action* of a group G *on* a set S is a mapping $\alpha : G \times S \to S$ that converts group multiplication into composition: for all $g, h \in G$ and $u \in S$, $\alpha(gh, u) = \alpha(g, \alpha(h, u))$. If e is the identity of G, it is customary—but not always necessary—to stipulate that $\alpha(e, u) = u$ for all u. The canonical example is when $S = V = \mathbf{C}^n$ and G is a group of complex $n \times n$ matrices A, with $\alpha(A, u) = Au$. Then $\alpha(AB, u) = ABu$ since multiplication is associative. Another action is $\alpha'(A, u) = AuA^{-1}$; note that $\alpha'(AB, u) = (AB)u(AB)^{-1} = ABuB^{-1}A = \alpha'(A, \alpha'(B, u))$. This is subsumed by the action of the product group $G' = G \times G$ given by $\alpha_2((A, B), u) = AuB^{-1}$.

Where the action referred to is clear, it is written simply with product notation, namely $g \cdot u$ in place of $\alpha(g, u)$. The action is said to "give S the structure of a G-module."

The action used most critically by Mulmuley and Sohoni takes G to be a group of invertible $m \times m$ matrices B and h to be an m-variable polynomial that is homogeneous of some degree $d \leq m$. Such h belong to the vector space $\mathcal{V}_d$ over $\mathbf{C}$ of dimension $D = \binom{m+d-1}{d}$, with basis given by the set of monomials of degree d. The action is $\alpha_G(B, h) = h'$, where h' is defined for all $v \in \mathbf{C}^m$ by $h'(v) = h(vB^{-1})$. Again, if C then sends h' to $h'' = \lambda w.h'(wC^{-1})$, then for all v, $h''(v) = h(vC^{-1}B^{-1})$, which is where BC sends h.

Definition 2.1 (see discussions in [MS02]). Let a nonzero homogeneous polynomial $h \in \mathcal{V}_d$ be regarded also projectively as a member of $P(\mathcal{V}_d)$ Then h is *stable* under the G action if the orbit Gh in $P(\mathcal{V}_d)$ is Z-closed, and *semi-stable* if the Z-closure of the orbit Gh in $\mathcal{V}_d$ does not contain the zero polynomial. Otherwise it is *unstable* or *nilpotent*. The set of unstable polynomials is the *null cone* of $\mathcal{V}_d$.

Given a closed set S, h is *S-stable* under the action if the orbit Gh has empty intersection with S. Then 0-stable is the same as semi-stable.

Orbits are always locally closed (see section 8.3 in the text [Hum81]), and hence constructible, but need not be closed. The boundary of the orbit, namely its Z-closure minus the orbit, is also preserved by G, so it is a union of orbits. Since in projective space this is the same as the classical boundary, it is lower-dimensional than the original orbit, so all orbits in the union are lower-dimensional. Iterating this reason yields the existence of orbits of minimum dimension in these boundaries that themselves are closed, so stable points always exist.

Not to be confused with "stable" are the following key concepts. The *isotropy subgroup* of a vector $v \in V$ under a given action by a group G is defined by $G_v = \{ g \in G : gv = v \}$. The orbit is then the singleton $\{ v \}$, which is closed, so such v are stable. The *stabilizer* of a vector subspace W of V is given by $G_W = \{ g \in G : (\forall w \in W)\ gw \in W \}$. The isotropy subgroup G_w is also called the "stabilizer" of w, but we try to minimize the usage of "stable..." and prefer to say w is *fixed* by G_w. Note that stabilizing W is weaker than fixing every $w \in W$, when we say W is *pointwise fixed*.

Example This expands "5.3.1 Example 2" in [MS02], itself ascribed to [PV91]. Let $V = \mathbf{C}^4$ be thought of as the space of homogeneous polynomials of degree 3 via coefficients of the four basic monomials x^3, x^2y, xy^2, y^3. Take $G = SL_2(\mathbf{C})$, $f = x^2y$, and $g = x^3 + y^3$. Then f is identified with $(0, 1, 0, 0)$, and g with $(1, 0, 0, 1)$. Take $A = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)^{-1}$, so that the action $A \cdot f = f \circ A^{-1}$ gives $A \cdot f = f(ax + by, cx + dy)$ and $A \cdot g = (ax + by)^3 + (cx + dy)^3$. Then subject to $ad - bc = 1$, the orbits of f and g are:

$$
\begin{aligned}
Gf &= \{ (a^2c, 2abc + a^2d, 2abd + b^2c, b^2d) : ad - bc = 1 \} \\
Gg &= \{ (a^3 + c^3, 3a^2b + 3c^2d, 3ab^2 + 3cd^2, b^3 + d^3) : ad - bc = 1 \}
\end{aligned}
$$

To determine the isotropy subgroup G_f for the action on V, we solve $a^2c = 0$, $2abc + a^2d = 1$, $2abd + b^2c = 0$, and $b^2d = 0$, together with $ad - bc = 1$. We cannot have $a = 0$ by the second equation, so $c = 0$, and then by $ad - bc = 1$ we cannot have $d = 0$ either. So $b = 0$ and we are left to solve $a^2d = 1$ and $ad = 1$. This forces $a = d = 1$, so we get the identity matrix only, meaning that G_f on V is trivial. *However*, if we work in $P(V)$, then the second equation becomes $2abc + a^2d = m$ for a general nonzero multiplier m, and then we're left to solve $a^2d = m$ and $ad = 1$. This is solved by $a = m$, $d = 1/m$, giving us a one-dimensional isotropy subgroup of diagonal matrices. Since $P(V)$ is 3-dimensional, this implies that the orbit Gf is two-dimensional.

For G_g in V, we solve $a^3 + c^3 = 1$, $a^2b + c^2d = 0$, $ab^2 + cd^2 = 0$, and $b^3 + d^3 = 1$, together with $ad - bc = 1$. If $a = 0$, then we get $c^3 = 1$ and $c^2d = 0$, so $d = 0$ and we're left with $bc = -1$, $c^3 = 1$, and $b^3 = 1$. This is impossible because no two of the cube roots of unity multiply to -1. So $a \neq 0$, and symmetrically, $d \neq 0$. Thus we can divide by a to get $b = -c^2d/a^2$, and then $ab^2 = -cd^2$ simplifies to $c^4/a^3 = -c$. If $c \neq 0$, this gives $c^3/a^3 = -1$, but that contradicts $a^3 + c^3 = 1$. So $c = 0$, and symmetrically, $b = 0$. This leaves $a^3 = 1$, $d^3 = 1$, and $ad = 1$, so G_g is the finite set $\left(\begin{smallmatrix}\omega & 0\\ 0 & \bar{\omega}\end{smallmatrix}\right)$ over the three cube roots ω of unity. Thus G_g is finite.

Now we can show that no multiple mf of f lies in the orbit Gg. Here we try to solve $a^3 + c^3 = 0$, $a^2b + c^2d = m$, $ab^2 + cd^2 = 0$, $b^3 + d^3 = 0$, and $ad - bc = 1$. If $d = 0$ then $b = 0$ and $ad - bc = 1$ is impossible; thus we can divide by d. Substituting $a = (1 + bc)/d$ in the third equation and multiplying through by d gives $b^2 + b^3c + cd^3 = 0$. But since $b^3 = d^3 = 0$ this gives $b = 0$ and $d = 0$, sending us back to the impasse. Thus without even considering the equation with m we cannot solve this. *However*, we can come within any desired $\epsilon > 0$ of f in each coefficient. Take $d = 0$, $b = \epsilon$, and $a = c = -1/\epsilon$. Then

$$Ag = (0, 1/\epsilon, -\epsilon, \epsilon^3).$$

Thus the multiple $(1/\epsilon)f$ of f lies within ϵ of the orbit, so in projective space, f lies within the Z-closure of the orbit. The fact that progressively higher multiples of f are needed in affine space is typical.

A simple example of isotropy is to note that symmetric polynomials are fixed under permutations of the variables, i.e. by permutation matrices applied to the argument variables. For another example, note that when A and B are invertible $n \times n$ matrices of equal determinant, and Y is an $n \times n$ matrix of variables, the action $Y \mapsto AYB^{-1}$ preserves the determinant, since $\det(AYB^{-1}) = \det(A)\det(Y)/\det(B) = \det(Y)$. This linear transformation of Y can be expressed as a matrix $S_{A,B}$ of size $n^2 \times n^2$ applied to Y unrolled as a vector. This matrix has $\det(S_{A,B}) = 1$, so $S_{A,B} \in SL_{n^2}(F)$. The set of all matrices arising as $S_{A,B}$ with A, B as above forms a subgroup R of $SL_{n^2}(F)$ that fixes $\det_n$.

If we only care about preserving det up to scalar multiples—i.e. if we consider det as existing in projective space—then we can define R' without the restriction $\det(B) = \det(A)$. Then R' is no longer a subgroup of $SL_{n^2}(F)$, but its intersection with $SL_{n^2}(F)$ is a subgroup R''. As asserted in [MS02], R'' comprises all matrices in $SL_{n^2}(F)$ that fix det, so the isotropy subgroup of det under the standard action by $SL_{n^2}(F)$ is R''. Moreover, the only polynomials fixed by R'' are multiples of det, so in projective space, $\det_n$ is characterized by R''. The isotropy subgroup of the permanent for $n \geq 3$ is shown ([MS02] citing Minc) to be generated by (the linear transformations in $SL_{n^2}(F)$ arising from) the subcases where A and B are either diagonal or permutation matrices, provided F is not of characteristic 2, and again the permanent is the unique polynomial up to multiples that it fixes.

Finally, an action by a group G on V can fix a polynomial function p on V in the sense that for all $g \in G$, $\lambda x.p(\alpha(g, x))$ equals p itself. Then p is likewise "G-invariant." When V is a space of polynomials, thinking of "polynomials with polynomial arguments" can seem hairy, but remembering correspondences such as $V =$ the span of $\{ x^3, x^2y, xy^2, y^3 \} = \mathbf{C}^4$ can help. As ascribed to Hilbert in [MS02], the null cone of V under a given group action is characterized as the set of points on which every non-constant homogeneous polynomial that is constant on orbits of G vanishes. This is an example of how the notion of isotropy ("stabilizing") interacts with stability (or being unstable).

2.2 Linear group representations

An action α_G on a vector space V is *linear* if for all $g \in G$, $u, v \in V$, and scalars $c \in F$, $\alpha(g, cu + v) = c\alpha(g, u) + \alpha(g, v)$. Then $\alpha(g, \cdot)$ induces a linear transformation on V, so there is a matrix A_g of the same dimension as V such that $\alpha(g, u) = A_g u$ for all $u \in V$. Thus the action of left multiplication by a matrix is canonical, and one can regard α itself as a homomorphism from G into a group of such matrices. In the action $\alpha(B, h) = \lambda x.h(B^{-1}x)$ used by Mulmuley and Sohoni above, "A_B" becomes an exponentially large matrix in terms of m (if $d = \Theta(m)$).

An action α_H by a homomorphic image $H = \phi(G)$ of G can be regarded as an action of G itself via $\alpha_G(g, u) = \alpha_H(\phi(g), u)$, since

$$\begin{aligned} \alpha_G(gh, u) &= \alpha_H(\phi(gh), u) \\ &= \alpha_H(\phi(g)\phi(h), u) \\ &= \alpha_H(\phi(g), \alpha_H(\phi(h), u)) \\ &= \alpha_G(g, \alpha_G(h, u)) \end{aligned}$$

. Thus if S is an H-module, it is a G-module for any G of which H is a subgroup.

Evidently because of these facts, it has become "cultural" to call either the matrix group or V itself a *representation* of G, even if the mapping "really" represents only a small image H of G. Another transmutation is that a group G of $n \times n$ matrices over a field F is-a subset of the vector space F^{n^2}, and one can attribute to G properties such as being Z-closed, locally closed, connected, and/or compact (the latter two with reference to the Euclidean topology).

For instance, the *special linear group* $SL_n(F)$ of $n \times n$ matrices A with $\det(A) = 1$ is Z-closed—because it is defined by the single equation $\det(x_{11}, \ldots, x_{nn}) = 1$. It is connected for $F = \mathbf{Q}, \mathbf{R}, \mathbf{C}$ because any two matrices of determinant 1 can be continuously varied one to the other through matrices of determinant 1. The *general linear group* $GL_n(F)$ of invertible $n \times n$ matrices is Z-open, because it is the complement of the Z-closed set defined by $\det(x_{11}, \ldots, x_{nn}) = 0$. It is, however, locally Z-closed.

What matters to us most in a representation, however, is the correspondence between subgroups H of G and subspaces W of V that are stabilized (one also says *preserved* or *invariant*) under the action of G. If H stabilizes W, then W is a representation of H. A representation α_G is *irreducible* if there is no proper subspace W of V that is stabilized by G. If G is a direct sum $H + J$ of two groups H and J (i.e., G is isomorphic to $H \times J$ with the group product $(h_1, j_1) \cdot (h_2, j_2) = (h_1 \cdot h_2, j_1 \cdot j_2)$), and $V = W + X$ for subspaces W and X preserved under the induced actions α_H and α_J, respectively, then α_G is *reducible* and *factors* as $\alpha_G = \alpha_H + \alpha_J$. Writing $v \in V$ uniquely as $w + x$ for $w \in W$ and $x \in X$, and $g \in G$ corresponds to (h, j) with $h \in H$ and $j \in J$, we get

$$\alpha_G(g, u) = (h, j) \cdot (w + x) = hw + jx == \alpha_H(w) + \alpha_J(x).$$

Note that if G itself stabilizes W, then we can take $H = G$ and "J" = the identity subgroup $\mathbf{1}$ of G and factor

$$\alpha_G = \alpha_G \wr W + \tau \wr X,$$

where τ is the action of $\mathbf{1}$ and is called the *trivial representation*. (Here $\wr$ means "restricted to.") Note also that $\tau \wr X$ itself is irreducible if and only if the subspace X is 1-dimensional. The main theorem this leads to is that *every representation onto a finite-dimensional vector space can be factored into irreducible representations*, and this factorization is unique up to isomorphism. The number of times τ occurs in this representation gives information about the dimensionality of invariant subspaces.

Representations have immediate relevance via *Kempf's stability criterion*. A *one-parameter subgroup* G' of G is the image of a homomorphism from the multiplicative group F^* of non-zero elements of F to G. It is *central* if every element of G' commutes with every element of G. For example, $GL_n(F)$ has the nontrivial central one-parameter subgroup consisting of scalar multiples of the identity matrix, but $SL_n(F)$ does not—and has no such subgroup. Following Alperin and Bell [AB95], define a subgroup G' of G to be *parabolic* if it is the simultaneous stabilizer of a sequence of linear subspaces

$$0 \subset W_1 \subset W_2 \subset \ldots \subset W_{r-1} \subset W_r = V.$$

Such a sequence is called a *flag*, and by the proper containments, must have $r \leq n = \dim(V)$ terms. For groups G such as $SL_n(F)$ with no nontrivial central one-parameter subgroups, Kempf's criterion states that a projective point $y \in P(V)$ is stable if the isotropy subgroup G_y is not contained in any proper parabolic subgroup of G. In particular this follows if the representation of G_y on V is irreducible. The above descriptions of the isotropy subgroups of $\det_n$ and $perm_n$ are irreducible representations, so these polynomials are stable under the action by $SL_{n^2}(F)$.

3 Application to Complexity Theory

The jumping-off point for the Mulmuley-Sohoni method is Valiant's method [Val79] (see also [vzG87]) of reducing any polynomial-size family of arithmetical circuits to a polynomial-size family of determinant computations. This extends to saying that functions believed to be intractable, such as the permanent polynomials, have polynomial-size (arithmetical) circuits iff they belong to the (Z-closure of the) orbits of the determinant polynomials under certain group actions. It is important to note that these polynomials are represented as vectors of length exponential in "n," hence exponential in the size of the matrices involved at the outset. We henceforth alter the notation in [MS02] to make explicit a distinction between *large* and *small* objects.

1. n is always the reference parameter for the length of the input to a computational problem.

2. $N = n^2$ is the size of an $n \times n$ matrix. Since we are concerned only with polynomial complexity, we can measure in terms of n even if the input is a matrix of size N.

3. $m = n^{O(1)}$ is the number of rows/columns in paddings of $n \times n$ matrices that arise, and becomes the degree of the permanent, determinant, and other homogeneous polynomials associated to these matrices. Sometimes d stands for the degree of these polynomials.

4. $M = m^2$ is the size of this matrix, and also the number of variables of these polynomials.

5. $A, B, C, \ldots$ stand for $n \times n$ or $m \times m$ matrices, while $\mathcal{A}, \mathcal{B}, \mathcal{C}$ stand for matrices of exponential size.

6. $p(n) = n^{O(1)}$ stands for a polynomial running time.

7. $h, k, q, \ldots$ stand for polynomials in N or M variables—note that these are vectors of length exponential in m, i.e. *large* objects.

8. $U, V, W, \ldots$ stand for vector spaces of *small* objects, while $\mathcal{U}, \mathcal{V}, \mathcal{W}$ stand for vector spaces of *large* objects.

9. $G, H, K, \ldots$ stand for groups of small matrices, while $\mathcal{G}, \mathcal{H}, \mathcal{K}$ stand for groups of large matrices.

3.1 Permanent and Determinant

Fix F to be the complex numbers for the time being. Suppose the permanent polynomial $perm_n$ of an $n \times n$ matrix x_{ij} of indeterminates has arithmetical formulas over F of size s. Then as remarked by von zur Gathen [vzG87], there is an $(s+2) \times (s+2)$ matrix M whose entries are either constants or variables x_{ij}, such that $\det(M)$ re-creates the permanent polynomial. Thus $perm_n$ becomes a "Valiant projection" of the determinant polynomial of an $m \times m$ matrix of indeterminates y_{ij}, with $m = s+2$. In order to apply the Mulmuley-Sohoni setup, we need to pad $perm_n$ into a degree-m homogeneous polynomial over the y_{ij} variables of equivalent circuit complexity. Mulmuley and Sohoni do this via a map $\phi(f) = y_{mm}^{m-n} f'$, where f' is f with variables x_{ij} renamed to y_{ij} for $1 \leq i, j \leq n$ and the choice of y_{mm} outside this square being arbitrary. Write $\phi(f)$ as f^ϕ for short. Also let $s_{\det}(n)$ stand for an upper bound on the formula size of $\det_n$; currently $s_{\det(n)} = n^{O(\log n)}$ is best known (see [BCS97]). The connection to complexity theory is the following proposition and its near-converse, which hold for any f in place of $perm_n$.

Proposition 3.1 ([MS02], Props. 4.1 and 4.4) *(a) If the permanent has formulas of size $m-2$, then $\det_m$ is not $perm_n^\phi$-stable.*

(b) If $\det_m$ is not $perm_n^\phi$-stable, then for any $\epsilon > 0$, there are n^2-variable formulas F_ϵ of size $m^2 s_{\det}(m)$ that approximate $perm_n$, in the strong sense that the coefficients of the polynomial computed by F_ϵ are within ϵ of those of corresponding terms of $perm_n$. Only the constants in F_ϵ depend on ϵ.

Proof. (a) Form an $m \times m$ matrix M' by taking M above, multiplying every nonzero constant entry of M by y_{mm}, and renaming entries x_{ij} with $1 \leq i, j \leq n$ back to y_{ij}. Then $\det(M') = perm_n^\phi$. Since every entry of M' is a trivial linear combination of entries of the $m \times m$ matrix $Y = (y_{ij})$, there is a linear transformation $A : F^{m^2} \to F^{m^2}$ that produces M' from Y. Thus $perm_n^\phi = \det(AY)$. Although A may be a non-singular $m^2 \times m^2$ matrix, there are elements A' of GL_{m^2} arbitrarily close to A. In fact, $\det(A'Y)$ can be made to approximate $perm_n^\phi$ coefficient-wise in the sense of (b). (Note that Y is "unrolled" as a vector of length m^2 in the product $A'Y$, not kept as a matrix.) Thus $perm_n^\phi$ is in the classical closure of the

GL_{m^2}-orbit of $\det_m = \det(Y)$ in affine space $\mathcal{V}_m$. Now take $d = \det(A')$ and $A'' = A'/(d^{1/m^2})$. Then $\det(A'') = 1$, so $A'' \in SL_{m^2}$, and $\det(A''Y)$ is coefficient-wise a scalar multiple of the previous approximation to $perm_n^\phi$. Thus passing to projective space, $perm_n^\phi$ belongs to the classical closure of the SL_{m^2}-orbit of $\det_m$. Since the classical closure is always contained in the Z-closure, (a) is proved.

(b) Suppose $perm_n^\phi$, regarded as a point in $P(\mathcal{V}_m)$, lies in the Z-closure of the SL_{m^2}-orbit of $\det_m$. Here we use the "key facts" that this orbit is a locally Z-closed subset of a Z-closed set in $P(\mathcal{V}_m)$ (see [Hum81]), and thus its Z-closure coincides with its classical closure in $P(\mathcal{V}_m)$. Thus for any $\delta > 0$, we can find an $m^2 \times m^2$ matrix A_δ with $\det(A_\delta) = 1$ such that $\det(A_\delta Y)$ has coefficients within δ of some multiple c of $perm_n^\phi$. Now let $A'_\delta = A_delta/c^{1/m}$. Then $\det(A'_\delta Y)$ has coefficients within δ/c of $perm_n^\phi$. Provided $\delta < \epsilon/c$, $\det(A'_\delta Y)$ has coefficients within ϵ of $perm_n^\phi$.

Now in forming $\det(A'_\delta Y)$, we can zero out variables that do not occur in $perm_n^\phi$, as the terms involving these variables must have coefficients of magnitude less than ϵ anyway. Also set $y_{mm} = 1$ and rename the remaining variables y_{ij} to x_{ij}. Then every entry of the length-m^2 vector $A'_\delta Y$ is a linear combination of variables x_{ij} plus possibly a constant term, where the coefficients depend on ϵ. When we roll $A'_\delta Y$ back into a matrix and compose this with formulas ψ of size $s_{\det}(m)$ for the $m \times m$ determinant, we plug linear formulas of size at most m^2 at the input gates of ψ. The resulting size is at most $m^2 s_{\det}(m)$. Moreover the entries of A'_δ affect only the coefficients, not the structure, of the resulting formula. $\square$

When $m = n^{O(1)}$ or even $m = n^{(\log n)^{O(1)}}$, $m^2 s_{\det}(m) = n^{(\log n)^{O(1)}}$ = size quasipolynomial in n. Thus there is nearly an equivalence here between f^ϕ-stability of $\det_m$ and coefficient-wise approximability of f by formulae of quasi-polynomial size, for any f not just the (padded) permanent.

What is untreated in (b) is how the magnitudes of constants in A_δ depend on δ, hence on ϵ. High magnitudes would prevent a solid link being drawn to *computing* approximations to the permanent via Turing machines or RAM models with "fair-cost arithmetic." However, algebraic complexity theory has mainly developed to ignore the magnitudes of constants—see [Lok95] as a stem paper for bounded-magnitude lower bounds and Mulmuley's own prior work [Mul99] for the related issue of constraining access to individual bits of constants. The intuition leading Valiant to conjecture exponential formula size lower bounds for $perm_n$ over any fields not of characteristic 2, not caring about the constants, certainly seems to extend to intuition against such approximability by quasipolynomial-size formulas. Hence this backs up Mulmuley-Sohoni's

Conjecture 3.2 *The determinant is not padded-permanent-stable under any polynomial (or sub-exponential) amount of padding.*

The $64,000 question is whether this reformulation of the intuition adds something tangibly new to the basic problem. Historically it must be remarked that in the 1890s, early proponents of group-representation theory faced detractors such as Burnside, until a long development curve finally did start producing results that seemed unobtainable by other means.

The meat of the matter is how well the algebraic tools capture complexity properties of both the function f being lower-bounded and the "universal" function being orbited—here, the determinant. The determinant is attractively characterizable as the unique degree-m homogeneous function that is fixed by a certain natural group action, as we note below. (As remarked in [MS02], other families of functions that are similarly universal for small formulas can be used, if they have even nicer stabilizer characterizations.) So is the permanent. The question is how well we can get a direct understanding of the group actions on—and in the neighborhoods of—these functions.

One thing that can help in a direct determination of whether the closure of an orbit of $\det_m$ touches a function f is whether the orbit of f itself is stable (under the same group action)! If f is stable, then known algebraic techniques come into play to analyze the neighborhoods of the orbits, and whether the orbits approach arbitrarily closely in the (here-classical-is-equivalent-to-) Zariski sense. Mulmuley and Sohoni observe that $perm_m$ is stable, but to the chagrin of all, the padded permanent $perm_n^\phi$ is not stable. However—and this is the jumping-point for the main technical content of their work—they develop a notion of "partial stability" that still allows much of the desired analytical tools to be recovered. In other words, the padding ϕ introduces a little bit of "nastiness" that can be carefully peeled away by refining the algebraic analysis. For this we need more definitions.

4 Partial stability

Recall that a parabolic subgroup $P \subseteq G$ stabilizes a flag of the form $0 \subset W_1 \subset \dots W_r = V$. P is *maximal* if it is not contained in any other parabolic subgroup. Again following [AB95] (for representation-tailored rather than "innate" definitions), the *unipotent radical* of P is the subgroup U_P of P that pointwise fixes every quotient subspace W_i/W_{i-1} in the flag. U_P is always a normal subgroup. A *reductive* group is one whose representation forms a connected set of points (unrolling the representing matrices as vectors) and whose unipotent radical is the identity. For subgroups of P, intuitively this means that the reductive subgroup poses no blockage for working down the flag. The identity group itself is not counted as reductive. If we write $W_i = W_{i-1} + Y_i$ for each i, then a *Levi subgroup* L of P stabilizes all the Y_i. The largest such L is denoted by L_P and called the *Levi complement of* U_P. L_P is always isomorphic to the direct product $GL_{y_1}(F) \times \dots \times GL_{y_r}(F)$, where each $y_i = \dim(Y_i) = \dim(W_i) - \dim(W_{i-1})$. The point is that $P = U_P L_P$ and $U_P \cap L_P$ is the identity, a situation summarized by saying that P is the *semi-direct product* of the normal subgroup U_P and the subgroup L_P.

The *rank* of an algebraic group G is the maximum dimension of a subgroup of G that is isomorphic to a group of diagonal matrices. For example, the rank of $GL_n(F)$ is n, but the rank of $SL_n(F)$ is only $n-1$ since one of the diagonal elements is constrained by the product of the others to make the determinant equal to 1. Finally, a subgroup of an algebraic group G is *regular* if its *root system* is a subsystem of that pf G. (We do not go into root systems here, but see the Appendix of [Hum81].) Although intuition in a short treatment (or even in [MS02]) has to lag behind, we now have all the definitions required to define partial stability.

Definition 4.1 ([MS02]). A projective point $y \in P(V)$ is *partially stable with defect* δ, Δ with respect to the action of an algebraic group G if there are a maximal parabolic subgroup P, a Levi subgroup L of P (not necessarily L_P itself), and a regular reductive subgroup K

of L such that for all affine points $x \in V$ on the line y:

(a) The isotropy subgroup G_x contains the unipotent radical U_P;

(b) $L \cap G_x$ is reductive;

(c) y is stable under the action by K;

(d) $rank(K) = rank(L) - \delta$; and

(e) $\dim(Lx)/\dim(Gx) \geq \Delta$, i.e., the dimension of the orbit Lx is sufficiently large compared to the dimension of the orbit Gx.

If $P = G$, $\delta = 0$, and $\Delta = 1$, then one gets $K = L = G$ also, and by (c) this co-incides with the definition of y being stable under G. Thus δ and Δ quantify the deviation from stability. In the cases of partial stability sought and used by Mulmuley and Sohoni, $\delta = 1$ and Δ is inverse-polynomial in n, so the deviation is "not too large." The intent, to be brought out in the companion paper [MS02], is to focus in on the action of K.

Theorem 4.1 (Theorem 4.3 in [MS02]) *The padded permanent $perm_n^\phi$ as a member of $P(\mathcal{V}_m)$ with m^2 variables is partially stable under the action by $SL_{m^2}(F)$, with defect $\delta = 1$ and Δ polynomial in m/n. This applies to any padded homogeneous form h of degree $n < m$ in n^2 variables, such that h is stable as a member of $P(\mathcal{V}_n)$ under the action by SL_{n^2}.*

5 Obstructions

Given polynomials f and g such as $perm^\phi$ and $\det_m$, respectively, and a relevant action by an algebraic group G, an *obstruction* is a witness that f does not lie in the closure of the orbit Gg in $P(\mathcal{V})$. This witness can be a "meta-polynomial" q. Namely, if f and g are homogeneous of degree d in r variables (above, $d = m$ and $r = m^2$), then f and g are points in a vector space of $R = \binom{m+d-1}{d}$ dimensions—as are all polynomials in the orbit Gg and its projective Z-closure. A polynomial q in R variables can thus be said to take these smaller polynomials as arguments. If q vanishes on all points in Gg, then it also vanishes on the closure. Further, if such a q gives $q(f) \neq 0$, then q witnesses that f does not belong to the orbit closure. However, such a witness could be a double-exponential-size object, and it is not clear that we have gained any information about the (f, g)-problem by doing this.

Rather, Mulmuley and Sohoni advance the goal of constructing representations of the isotropy subgroups G_f and G_g that serve as a witness. (In [MS02], G_f is called "H" and G_g is called "Q.")

Theorem 5.1 (Theorem 5.1 in [MS02]) *Let f be stable under the action of a group G on $P(\mathcal{V})$. A nonzero representation $\mathcal{W}$ of G is an obstruction for (f, g) if its unique factorization into irreducibles contains an occurrence of the trivial G_f-module τ_f but not a trivial G_g-module τ_g—and more generally, if τ_f occurs more often than τ_g in this factorization.*

This reads like a definition but is actually a theorem, replacing "is an obstruction" with the conclusion, "then $f \notin \overline{Gg}$." We can word the more-general condition a little more helpfully.

Define

$$\begin{aligned} E_f &= \{\, v \in \mathcal{W} : (\forall A \in G_f)\ A{\cdot}v = v \,\}, \\ E_g &= \{\, v \in \mathcal{W} : (\forall A \in G_g)\ A{\cdot}v = v \,\}. \end{aligned}$$

Then E_f and E_g are closed under scalar multiplication and under addition, so they are linear subspaces—respectively, the subspace pointwise fixed by G_f and the space pointwise fixed by G_g. The more-general condition is then simply $\dim(E_f) > \dim(E_g)$—i.e., construction of a representation W giving this implies that $f \notin \overline{Gg}$.

To prove this, note first that this disparity of dimension is impossible if G_f contains AG_gA^{-1} for some $A \in G$. This is intuitively because $\dim(E_f) > \dim(E_g)$ says that G_f should be smaller than G_g, since larger groups fix fewer elements. Notice that AG_gA^{-1} forms a group, called a *conjugate* of G_g.

However, the stability of f implies by a result known as *Luna's slice theorem* that the orbit Gf in $\mathcal{V}$ has a neighborhood U preserved by G, such that the isotropy subgroup S_p of any point p in U is a conjugate of a subgroup of G_f. Let $[\cdot]$ denote the mapping from $\mathcal{V} \setminus \{\, 0 \,\}$ to $P(\mathcal{V})$. If $[f]$ lies in the (Z-equivalent-to-classical) closure of $G[g]$, then $[U]$ contains a point $[A{\cdot}g]$ for some $A \in G$. Thus U contains a point $p = \lambda A{\cdot}g$ with $\lambda \in F^*$ (i.e., $\lambda \neq 0$). Now the isotropy subgroup $S_p = G_{\lambda Ag}$ equals G_{Ag} for any nonzero scalar λ. Now observe that $G_{Ag} = AG_gA^{-1}$, because if $B{\cdot}g = g$, then

$$(ABA^{-1}){\cdot}(A{\cdot}g) = (ABA^{-1}A){\cdot}g = (AB){\cdot}g = A{\cdot}(B{\cdot}g) = A{\cdot}g.$$

Thus elements of AG_gA^{-1} are precisely those that fix $A{\cdot}g$, so S_p is a conjugate of G_g. Since a conjugate of a conjugate is a conjugate, Luna's slice theorem tells us that G_g is a conjugate of a subgroup of G_f. This is exactly what we argued couldn't happen from the dimensions of E_f and E_g in the representation. Hence $[f]$ cannot belong to the closure of the orbit $G[g]$ in $P(\mathcal{V})$.

This proof—and its exploitation of the stability of f itself—moves the focus onto constructing informative, extremal representations of G. This is a much-studied area of mathematics, where many kinds of demands have already been dealt with. The demands here are governed largely by the structure of the isotropy subgroups G_f and G_g, which in the case of the permanent and the determinant have been nicely characterized above.

Ah, but Conjecture 3.2 references the *padded* permanent. Partial stability does not yield the efficient conclusion of Luna's slice theorem. Mulmuley and Sohoni give one adaptation of Theorem 5.1 *toward* the partially-stable case, as "Theorem 5.3" in the part-I paper [MS02], but it still requires f to be stable. The full development of obstructions in the partially-stable case is the main subject of the part-II paper [MS02], to appear. According to the overview paper [MS01], this proceeds deeper into century-old unsolved problems in representation theory such as the "plethysm problem." At this point, as complexity theorists we should first take one step back to try to assess the combinatorial nature of the proof techniques and problems encountered.

6 Would this naturalize?

Razborov and Rudich [RR97] showed the existence of a new obstacle to circuit lower-bound proofs. They observed that basically all known proofs that certain Boolean functions h_n lie outside a circuit class $\mathcal{C}$ revolve around sequences Π_n of subsets of $\mathcal{F}_n$ (the set of 2^{2^n} Boolean functions of n variables) such that:

(a) no language L such that $L^{=n} \in \Pi_n$ for almost all (variantly, infinitely-many) n belongs to $\mathcal{C}$ ("usefulness");

(b) $|\Pi_n|/|\mathcal{F}_n|$ is bounded below by $1/p(2^n)$ for some polynomial p ("largeness");

(c) whether a given $f \in \mathcal{F}_n$ belongs to Π_n is decidable in time $2^{n^{O(1)}}$, which is quasi-polynomial in the length 2^n of the truth table of f given as input ("low complexity");

and $h_n \in \Pi_n$ for all n. Razborov and Rudich showed that if a sequence Π_n satisfies (a)–(c) with $\mathcal{C} = \mathrm{P}/poly$ (i.e., the class of languages having polynomial-sized circuits), then pseudorandom generators and one-way functions of exponential security do not exist. Since the factoring problem is widely believed to be hard enough to produce such generators, such "natural proofs" Π_n are conjectured not to exist. Granting this, a proof that NP-complete problems do not have polynomial-sized circuits (regarded on a par with proving Ptime $\neq$ NP) must surmount either the largeness or the low-complexity condition.

Although the Razborov-Rudich framework and results have not yet been carried over formally to arithmetic circuits, it is reasonable to speak as though they have—and to expect that recent techniques by Koiran [Koi96] and Bürgisser [Bür98, Bür00] using fields of finite characteristic as conduits from the algebraic to the Boolean case can extend to accomplish this. Thus assuming the Mulmuley-Sohoni method is useful against (the arithmetical analogue of) $\mathrm{P}/poly$, which of (b) and (c) does it escape from?

Mulmuley and Sohoni argue for (b). Note that if the hardness predicate $\Pi_n(h)$ were simply $D_n(h) =$ "h does not belong to the projective Z-closure of the orbit of the determinant," then $\Pi_n(h)$ would be very large—indeed *vast* in the Zariski sense, being the complement of the closure of the orbit. However, they emphasize that their hardness predicates will have the form $D_n(h) \wedge S_n(h)$, or more likely $D_n(h) \wedge P_n(h)$, where S_n [P_n] expresses the [partial] stability of h under the same group action—and/or the ability to compute explicit obstructions. The predicates S_n and P_n are expected to be small (under translation to the Boolean case). Moreover, for h such as the (padded) permanent polynomial(s), one can conjoin to S_n or P_n the clause, "and h has a nice, large isotropy subgroup such that h is the only thing it fixes."

However, in public talks subsequent to their paper, Razborov and Rudich have pointed out a philosophical obstacle to arguments that rely only on overcoming (b). Within the confines of the purported proof based on Π_n, the complement Π'_n becomes an "easyness" predicate. If Π'_n is vast, then one is willy-nilly arguing in the situation where "a random function f is easy." Now because $+$ is invertible (as with exclusive-or in the Boolean case), it follows that "a random function $g = h - f$" over random f is easy. One thus must admit cases where one's "hard" function h equals a sum $f+g$ of two "easy" functions. Roughly put, this prevents the lower-bound proof from working inductively on the arithmetical operators.

However, the argument over (b) may be effectively moot, as it seems transparent from the way the Mulmuley-Sohoni technique involves "large objects" that the complexity of the hardness predicate would be greater than the bound in (c). The large objects have size $2^{p(n)}$ where p is *not* a fixed polynomial but rather one that is universally quantified. For instance, the hardness predicate taken from their "Conjecture 4.3" for the permanent could be made to read: h is hard if for all polynomials $p(n)$, taking $m = p(n)$, h padded up to h^ϕ is not in the orbit closure of $\det_m$. The "Natural Proofs Obstacle," however, still seems to indicate here that their method will not be able to escape the necessity of quantifying over p and dealing both with the padding and with *multiple* source functions like $\det_m$ that one has to argue over.

Mulmuley and Sohoni do have a point in giving motivation to consider problems other than standard (NP-)complete ones as targets for lower bound proofs. In Sections 7 and 8 of [MS02], they introduce the following problem in NP, which we name "FULL RANK AVOIDANCE."

> INSTANCE: A matrix X with n rows and kn columns grouped as n blocks of k.
> QUESTION: Is it possible to choose one column from each block so that the resulting $n \times n$ matrix M has $\det(M) = 0$?

They conjecture that even for $k = 3$ and $F = \mathrm{GF}(2)$, this problem is not in Ptime. The closest problem that is known to us (personal communication from Mitsunori Ogihara, September 2002) to be NP-complete has $k = 2$ and $2r$ columns where the input can vary r subject to $r < n$, and asks if the resulting $n \times r$ matrix M has column rank less than r. The Mulmuley-Sohoni problem requires $r = n$, however, and Ogihara's reduction (from 3-not-all-equal-SAT) seems to require $r < n$ depending on the input formula to the reduction.

Mulmuley and Sohoni prove that provided the characteristic of the field F does not divide n, k, or $k - 1$, the following polynomial $E(X)$ associated to this problem is stable under the standard action by $SL_{kn^2}(F)$ (again "unrolling" the matrix into a vector):

$$E(X) = \prod_{\sigma} \det(X_\sigma),$$

where σ stands for functions choosing one column from each block and X_σ is the resulting $n \times n$ matrix. Clearly $E(X) = 0$ iff the answer to the problem instance is "yes." The stability of $E(X)$ then plays into partial stability of padded forms $E(X)_n^\phi$, and thus allows the meat of the analysis to come in [MS02] to be applied to it.

In conclusion, there is both much deep mathematical content *and* some new concrete combinatorics in their approach. It may not seem near to resolving Ptime vs. NP now, but it does talk about objects of the right kind of complexity for working on it.

Acknowledgments The author thanks Ketan Mulmuley, Mitsunori Ogihara, and Maurice Jansen for helpful answers and contributions.

References

[AB95] J.L. Alperin and R.B. Bell. *Groups and Representations*, volume 162 of *Graduate Texts in Mathematics*. Springer Verlag, 1995.

[BCS97] P. Bürgisser, M. Clausen, and M.A. Shokrollahi. *Algebraic Complexity Theory.* Springer Verlag, 1997.

[Bür98] Peter Bürgisser. On the structure of Valiant's complexity classes. In *15th Annual Symposium on Theoretical Aspects of Computer Science*, volume 1373 of *lncs*, pages 194–204, Paris France, 25–27 February 1998. Springer.

[Bür00] Peter Bürgisser. Cook's versus Valiant's hypothesis. *Theor. Comp. Sci.*, 235:71–88, 2000.

[Hum81] J.E. Humphreys. *Linear Algebraic Groups*, volume 21 of *Graduate Texts in Mathematics*. Springer Verlag, 1981. 2nd printing.

[Koi96] Pascal Koiran. Hilbert's Nullstellensatz is in the polynomial hierarchy. *Journal of Complexity*, 12(4):273–286, December 1996.

[Lok95] S. Lokam. Spectral methods for matrix rigidity with applications to size-depth tradeoffs and communication complexity. In *Proc. 36th Annual IEEE Symposium on Foundations of Computer Science*, pages 6–15, 1995.

[MS02] K. Mulmuley and M. Sohoni. Geometric complexity theory II: Explicit obstructions. Manuscript, in preparation, 2001–02.

[MS01] K. Mulmuley and M. Sohoni. Geometric complexity theory, P vs. NP, and explicit obstructions. In *Proceedings, International Conference on Algebra and Geometry, Hyderabad, 2001*, 2001.

[MS02] K. Mulmuley and M. Sohoni. Geometric complexity theory I: An approach to the P vs. NP and related problems. *SIAM J. Comput.*, 31(2):496–526, 2002.

[Mul99] K. Mulmuley. Lower bounds in a parallel model without bit operations. *SIAM J. Comput.*, 28:1460–1509, 1999.

[Mum76] D. Mumford. *Algebraic Geometry I: Complex Projective Varieties*, volume 221 of *Grundlehren der mathematischen Wissenschaften.* Springer Verlag, 1976.

[PV91] V. Popov and E. Vinberg. *Invariant Theory*, volume 55 of *Algebraic Geometry IV, Encyclopedia Math. Sci.* Springer Verlag, 1991.

[RR97] A. Razborov and S. Rudich. Natural proofs. *J. Comp. Sys. Sci.*, 55:24–35, 1997.

[Val79] L. Valiant. Completeness classes in algebra. Technical Report CSR-40-79, Dept. of Computer Science, University of Edinburgh, April 1979.

[vzG87] J. von zur Gathen. Feasible arithmetic computations: Valiant's hypothesis. *Journal of Symbolic Computation*, 4:137–172, 1987.