Symmetric Measures via Moments

Alexey Koloydenko

April 7, 2004

1 Introduction

This work is about objects that, when acted upon, do not change, or stay *invariant*. The notion of invariance is fundamental in many realms of human thought but we specialize it here to a collection of mathematical objects that can represent data observed in real experiments.

Our focus is probability distributions on $\Omega \subset \mathbb{R}^m$, where Ω is invariant under a finite group G of nonsingular linear transformations. Within this class of distributions we are most interested in ones that assign the same mass to all g-transforms ($g \in G$) of every (measurable) set $B \subset \Omega$. These are G-invariant distributions.

We set two goals for this work. The first one is to generalize the problem of unique determinacy of (multivariate) measures by their moments in the following way:

In the ordinary formulation, one studies whether or not a measure with finite (absolute) mixed moments, is uniquely determined by its mixed moments, or simply *determinate*, [1], [2], [8], [11], [20], [21], [31], [35].

Several sufficient conditions for determinacy ([1], [2], [8], [11], [31]) and indeterminacy ([31], [35]) are commonly known for measures on $\mathbb{R}$ or $\mathbb{R}^+$. For determinacy of measures on $\mathbb{R}^m$, [8] generalizes some of those conditions and gives several new ones. Now, we think of these (multivariate) measures in question as G-invariant where G is the trivial group of the identity transformation. Action of a non-trivial G narrows down the class of G-invariant measures under investigation. Hence, adapting the standard conditions for determinacy, we expect to need only a subset of all the moments in order to uniquely identify a G-invariant measure among all G-invariant ones.

Toward this goal, §2 reviews basic notions of group action and associated invariance, introduces G-invariant measures, and minimal sets of generators $\{f_1, \dots, f_N\}$ of the ring (algebra) of G-invariant polynomials in m indeterminates ([7], [9], [34], [36]). We also introduce Reynolds operators $\mathcal{R}$ ([7], [9], [34], [36]) that average real functions to make them G-invariant. Finally, we explain the sufficiency of $f = (f_1, \dots, f_N)$ to represent any G-invariant function on $\mathbb{R}^m$. Relevant proofs are given in Appendix A.

We continue in §3 by defining G-invariant moments and formulating the notion of determinacy of G-invariant measures by their G-invariant moments. Paralleling the main results of [8] obtained for the case of ordinary determinacy, we state several sufficient conditions for determinacy of G-invariant measures by their G-invariant moments. These include the *Extended Carleman Theorem for G-invariant moments*, and some integral conditions based on *quasi-analytic weights* (§3.1). All of these results rely on the one-to-one correspondence between the invariant measures on $\mathbb{R}^m$ and measures on $\mathbb{R}^N$ established via an extension of the multinomial map $f = (f_1, \ldots, f_N)$ (Lemma 20). Auxiliary proofs are deferred till Appendix B.

We acknowledge that to a certain extent, symmetry has already been studied in connection with the problem of moments. Thus, for instance, [21] studies the existence and uniqueness of symmetric measures on $\mathbb{R}$ with given moments. Also, [8] generalizes this case and studies determinacy of multivariate measures supported in the positive cone ("C-determinacy"). In one dimension, the correspondence between symmetric measures and measures on the nonnegative half-line is rather obvious and well-known [11]. Apparently, this correspondence generalizes easily to the multivariate setting (proof of Theorem 5.1 of [8] and Example 1), also illustrating the significance of our Lemma 20. The symmetry with respect to the continuous group of all the rotations on $\mathbb{R}^m$ is discussed, for example, in [1], [2]. In this case all of the invariant functions are "generated" by a single invariant polynomial $\sum_{i=1}^m x_i^2$, which is a *maximal invariant* in the language of *equivariance theory.* We, however, focus on finite subgroups of $GL(m, \mathbb{R})$. Finally, note the difference between our theme and the related notion of equivariance in statistics [27], [33]. In the latter case it is entire (parametric) families of distributions and not individual measures that are fixed under groups of transformations. Also, the relevant groups in the equivariance theory are continuous. However, there appear not so many interesting examples (besides the aforementioned one with the rotational symmetry) of finite measures individually fixed by an infinite subgroup of $GL(m, \mathbb{R})$.

Our second goal is to develop a framework for model selection in the presence of the above types of invariance. The main motivation comes from modeling distributions of very small square subimages of digitized natural images [15], [24], [26], [32]. §7 describes the particular state space Ω, the symmetry group G acting on Ω, a minimal set of generators of the corresponding G-invariant polynomials, and several other relevant details of the studies undertaken in [24]. Thus, §7 illuminates most of the concepts developed in this work, and proofs of the results from this section are given in Appendices D and E.

The framework that we propose is based on the constrained *Entropy Maximization Principle* ([4], [6], [20], [25], [28], [38], [41]). Recall that according with this principle, the knowledge of the distribution to be modeled is formulated by a finite set of consistent constraints of the form $\mathbb{E}_{P'}\phi(X) = \nu_\phi$. Among all distributions P' that satisfy the constraints, one chooses P' that maximizes the *entropy* $H(P)$ that represents mathematically our intuitive notion of distributional uncertainty. Equivalently, such

P' maximizes the likelihood under the exponential family of distributions for which ϕ's are a sufficient statistics.

We work with moment constraints, i.e. $\phi(X) = X^\alpha$, $\alpha \in A \subset \mathbb{N}^N$, considering sequences of maximum entropy problems with expanding A's. Unlike in the related works of [16], [20], [29], [37] on maximum entropy problems with moment constraints, our moment matching, or pursuit, is multidimensional, adaptive, and G-invariant. Adaptiveness (also see below) here refers to a certain optimality in the sequential expansion of A's, and is meant to accelerate the approximation of the modeled distribution. The connection with the notion of determinacy is that a determinate P can be approximated arbitrarily well by progressively matching all its moments. The one-dimensional version of the latter result was already successfully used for density estimation in, for example, solid state and quantum physics [29] and econometrics [16], [37].

We take advantage of a key observation that entropy maximization forces the resulting distributions to inherit G-invariance of the constraining functions (proof of Theorem 26).

Although for us the pivotal case is that of Ω finite, §5, in §4 we nonetheless lay a foundation for a more general sequential reconstruction of a G-invariant distribution by G-invariant moment pursuit (Theorem 24). We also touch on the continuous case (Theorem 26) for completeness of the presentation. We term our modeling approach "Adaptive minimax learning" in recognition of its origin in texture modeling [39], [40], [41]. Minimax learning of an unknown distribution P refers to an incremental model construction, in which at each step l the entropy maximization problem is solved with one new constraint added at a time. In the original formulation, the l-th constraint is chosen to minimize the Kullback-Leibler divergence of the candidate maximum entropy distribution (with l constraints) from the target distribution.

However, this formulation stops short of balancing model fit and model complexity, which is the main task of model selection. In response to this, we discuss in §5.1 a simple ad hoc way to prevent overfitting within the original formulation. In §5.2, we explain that our adaptive minimax learning allows one to replace the Kullback-Leibler divergence with a more suitable cost function of their own choice, such as, for example, *description length* [18]. In §8 we summarize the main features of our modeling framework as *minimax learning*, *incorporation of multivariate moments*, and *incorporation of G-invariance*.

In §6 we discuss several computational issues of our modeling, such as computation of G-invariant polynomial generators, Reynolds operators, and the partition $\mathcal{S}_\Omega$ (§6.2). §6.3 is dedicated to a computational result (Theorem 36 and Corollary 37) on dimensionality reduction in the entropy maximization problem with constraints that have nontrivial finite constancy classes. §6.4 discusses efficient computations of additional terms for the minimax learning algorithm. In §8 we also discuss directions for future work that include model selection experiments based on real data pertinent to our example in §7.

2 Group action, invariance, polynomial generators

In this section we review several notions from algebra and introduce relevant notation.

Definition 1 *A* group action *of a group G on a set A is a map from $G \times A$ to A (written as ga, for all $g \in G$ and $a \in \Omega$) satisfying the following properties ([10]):*

1.) $g_1(g_2 a) = (g_1 g_2)a$, for all $g_1, g_2 \in G$, $a \in A$, and

2.) $1a = a$, for all $a \in A$.

Definition 2 *Let G act on A and let $a \in A$. a is said to be fixed under G, or G-invariant, if $ga = a$ $\forall g \in G$. $B \subset A$ is said to be fixed under G, or G-invariant, if $\forall b \in B$ $\forall g \in G$ $gb \in B$.*

We will also use the following observations that show how the original G action on A induces G actions on objects from various categories involving A:

Proposition 3

1.) Let $B \subset A$ be fixed under G. Then the restriction of the original G action on A is a well-defined G action on B.

2.) The following defines a G action on $\mathbb{R}^A$, the set of all real valued functions on A:

$$(gf)(a) = f(g^{-1}a), \quad \text{where} \quad g \in G \quad \text{and } f \in \mathbb{R}^A \text{ and } a \in A. \tag{1}$$

3.) The following defines a G action on $\mathcal{P}^A$, the power set of A:

$$gB = \{gb : \ \omega \in B\} \text{ for } B \subset A. \tag{2}$$

Let a finite group G act on $W = \mathbb{R}^m$ in a way that admits a linear (matrix) representation $\rho : G \hookrightarrow GL(W)$ ($\cong GL(m, \mathbb{R})$). We will simply identify the original action of G on W with its matrix representation, ρ and will therefore think of $g \in G$ as an $m \times m$ matrix.

Instantiating Proposition 3, we introduce the following G actions:

Proposition 4 *The following actions are well-defined.*

1.) The (restricted) action of G on an invariant $\Omega \subset W$.

2.) The G action on $\mathcal{B}$, the Borel σ-algebra on Ω:

$$gB = \{g\omega : \ \omega \in B\}. \tag{3}$$

3.) The G action on $\mathcal{M}$, the set of (positive) measures on $\mathcal{B}$:

$$(gP)(B) = P(g^{-1}B), \quad B \in \mathcal{B}, \quad P \in \mathcal{M}. \tag{4}$$

4.) The G action on $\mathbb{R}[W]$, the set of real polynomials in m indeterminates:

$$(gf)(v) = f(g^{-1}v), \quad \textit{where} \;\; g \in G \quad \textit{and } f \in \mathbb{R}[W] \textit{ and } v \in W. \tag{5}$$

Proposition 5 *Any group action partitions the set on which it acts.*

Definition 6 *Let $\mathcal{S}_\Omega = \Omega/G$ be the set of equivalence classes (also called* orbits*) of the given G action on Ω.*

Proposition 7 *For any $\Omega_1 \subset \Omega_2$, two invariant subsets of W, $\mathcal{S}_{\Omega_1} \subset \mathcal{S}_{\Omega_2}$.*

We will also need the following sets of invariant measures on $\mathcal{B}$:

Definition 8

$$\mathcal{M}^G = \{P \in \mathcal{M} : \; gP = P \; \forall g \in G\} \quad \textit{and} \quad \mathcal{M}^G_* = \mathcal{M}^G \cap \mathcal{M}^*,$$

where

$$\mathcal{M}^* = \{P \in \mathcal{M} : \mathbb{E}_P|X^\alpha| < \infty \; \alpha \in \mathbb{N}^m\}, \; \textit{and } X = (X_1, \ldots, X_m).$$

The multiindex notation f^α for $f \in \mathbb{R}^N$ and $\alpha \in \mathbb{N}^N$ means $f_1^{\alpha_1} \cdots f_N^{\alpha_N}$, thus: $X^\alpha = X_1^{\alpha_1} \cdots X_m^{\alpha_m}$. We also extend the expectation notation $\mathbb{E}_P$ to all $P \in \mathcal{M}$.

Proposition 9

$$\mathcal{M}^* = \{P \in \mathcal{M} : \mathbb{E}_P\|X\|^d < \infty \; \forall d \geq 0\}$$

Other useful invariant objects include:

1. $\mathcal{P}^G$, the set of invariant probability measures on Ω.
2. $(\mathbb{R}^\Omega)^G$, the set of invariant real functions on Ω.
3. $\mathcal{B}^G$, the σ-algebra of invariant Borel sets.
4. $\mathbb{R}[W]^G$ (alternatively $\mathbb{R}[x]^G$), the ring of invariant polynomials on W ($\ni x$).

The following operator projects $\mathbb{R}^\Omega$, the linear space of real functions on Ω, onto $(\mathbb{R}^\Omega)^G$, the linear subspace of G-invariant real functions on Ω, and plays a key role in the ensuing development (see also §A):

$$\mathcal{R}(f) = \frac{1}{|G|}\sum_{g \in G} gf. \tag{6}$$

We will also be interested in the restricted operator $\mathcal{R} \; : \mathbb{R}[W] \to \mathbb{R}[W]^G$, and in the adjoint $\mathcal{R}^* \; : \mathcal{M} \to \mathcal{M}^G$:

$$\mathcal{R}^*(P) = \frac{1}{|G|}\sum_{g \in G} gP \tag{7}$$

Proposition 10 *Consider $\mathcal{R}$ mapping the space of measurable functions on W onto $(\mathbb{R}^W)^G$ and the linear functionals $f \mapsto \int_W f(x)dP(x)$ indexed by $P \in \mathcal{M}$. Then $\mathcal{R}$ and $\mathcal{R}^*$ are adjoint.*

Proposition 11

1.) Let $P \in \mathcal{M}$ have a density p relative to some reference measure μ. Then $\mathcal{R}(p)$ is a density of $\mathcal{R}^*(P)$ relative to μ.

2.) Let p be a density of a G-invariant measure P relative to μ, then p is μ-a.e. G-invariant.

Our main ingredients are invariant polynomials from $\mathbb{R}[W]^G$ and their special representatives that *generate* the entire ring:

Definition 12 *Polynomials $f_1, \ldots, f_N$ from $\mathbb{R}[W]^G$ are said to generate $\mathbb{R}[W]^G$ if any $f \in \mathbb{R}[W]^G$ can be expressed as a polynomial in terms of $f_1, \ldots, f_N$. We will also refer to such $f_1, \ldots, f_N$ as* generators.

Definition 13 *Let $f_1, \ldots, f_N$ generate $\mathbb{R}[W]^G$. We call $f_1, \ldots, f_N$ a* minimal system of generators *if none of the generators can be expressed as a polynomial in terms of the others. In this case, we will also refer to such $f_1, \ldots, f_N$ as* fundamental integral invariants.

The fact that there always exists a finite system of such generators was proved by Hilbert for polynomials with coefficients from fields of characteristic zero (e.g. $\mathbb{R}$), and later extended for certain fields of positive characteristic by Noether ([13], [34]).

Remark 14 *Let $\mathbb{C}[W]^G$ be the ring (also, a complex algebra) of G-invariant polynomials with complex coefficients. Then note that for any $r(x) \in \mathbb{C}[W]^G$,* $\mathrm{Re}(r(x))$, $\mathrm{Im}(r(x)) \in \mathbb{R}[W]^G$ *since the complex conjugation on $\mathbb{C}[W]$ commutes with the G action on $\mathbb{C}[W]$.*

The next well-known fact is also fundamental for our discussion and follows from more general results in *Invariant Theory* [7], [30], [34], [36]. In §A we give a short, basic proof of this result.

Proposition 15 *Let $f_1, \ldots, f_N$ generate $\mathbb{R}[W]^G$ and let $f = (f_1, \ldots, f_N) : W \to \mathbb{R}^N$. Then the map $\bar{f} : \mathcal{S}_W \to \mathbb{R}^N$ mapping $[w]$, the equivalence class of $w \in W$, to $f(w)$, is well-defined and injective. Thus $\mathcal{S}_W \cong f(W)$, the image of f in $\mathbb{R}^N$.*

Example 1 *Let $G \cong \mathbb{Z}_2^m$ be the group of order 2^m generated by the component-wise sign inversions. As a matrix group, G is generated by m matrices whose all off-diagonal entries equal 0, and all but one diagonal entries equal to 1. The k-th matrix has its k-th diagonal entry equal to -1. It can be shown that $\{f_i = x_i^2,\ i = 1, \ldots, m\}$ is a minimal set of generators of $\mathbb{R}[W]^G$. $[w]$, the equivalence class of $w \in W$,* *is the smallest set containing w and symmetric with respect to reflections about all hyperplanes $x_i = 0$ $i = 1, \ldots, m$. The size of $[w]$ is 2^l, where l is the number of nonzero components of w, which also stays invariant under the transformations in G.*

In particular, in one dimension this is simply the symmetry around 0. Also, if such an invariant measure has a density, then the density must be an even function, i.e. function of x^2.

3 Invariant Moments, Determinacy of Invariant Measures

The problem of moments is whether a measure exists with prescribed moments and if so, whether it is unique within the class of all measures with finite moments. We are going to generalize the latter question to include situations when measures are to be determined within special subclasses of the original class and by, one would then expect, "fewer" moments. In particular, we are introducing the notion of determinacy of G-invariant measures by "G-invariant moments". Our notation intentionally resembles that from [1] and [8].

Let $f_1, \ldots, f_N$ be a minimal set of generators. Let $P \in \mathcal{M}$, and let $\alpha \in \mathbb{N}^N$ be the degree multi index.

Definition 16 *Given the f generators, we call $\mathbb{E}_P f^\alpha = \int_W f^\alpha dP(x)$ the mixed G-invariant moment of order α, or,* invariant α-moment *and denote it by $s_\alpha(P)$.*

Let us also denote by $s(P)$ the set of all such moments $(s_\alpha(P))_{\alpha \in \mathbb{N}^N}$ for a given measure P. When the measure P is clear from the context, we will overload the notation $s_n(k) = \mathbb{E}_P f_n^k$ for $k \in \mathbb{N}$ and $1 \leq n \leq N$.

Proposition 17 *Let $f_1, \ldots, f_N$ be a minimal generating set. Then $\mathcal{M}_*^G = \{P \in \mathcal{M}^G : \ \mathbb{E}_P |f^\alpha| < \infty \ \alpha \in \mathbb{N}^N\}$.*

Definition 18 *Let $P \in \mathcal{M}_*^G$ have $s(P)$, its G-invariant moments, relative to some minimal generating set. Then P is said to be G-determinate by $s(P)$, or simply G-determinate, if no other measure in $\mathcal{M}_*^G$ has the same set of moments $s(P)$ relative to the chosen generating set.*

In §B we prove that this notion is well-defined, i.e. independent of the choice of the generators.

We next give a generalized version of the extended Carleman theorem (§C, [8]):

Theorem 19 *(Extended Carleman theorem for G-invariant measures). Let $f_1, \ldots, f_N$ be some minimal set of generators. Let $P \in \mathcal{M}_*^G$ and assume that for each $n = 1, \ldots, N$, $\{s_n(k)\}_{k=1}^\infty$ satisfies Carleman's condition*

$$\sum_{k=1}^{\infty} \frac{1}{s_n(2k)^{1/2k}} = \infty, \tag{8}$$

then P is determinate by G-invariant moments. Also, $\mathbb{C}[W]^G$ and $\mathrm{Span}_{\mathbb{C}}\{e^{i(\lambda,f)}|\lambda \in S\}$ are dense in $L_p^G(W,P)$, the G-invariant subspace of complex $L_p(W,P)$, for $1 \leq p < \infty$ and for every $S \in \mathbb{R}^N$ which is somewhere dense (i.e. $\bar{S}$, the closure of S, has a nonempty interior).

Proof. The proof of the first statement takes two steps. First, notice that the map $f = (f_1, \ldots, f_N) : W \to \mathbb{R}^N$ as in Proposition 15 induces an injection $\tilde{f}$ of $\mathcal{M}_*^G$ to $\tilde{\mathcal{M}}_*$, the set of probability measures on $\mathbb{R}^N$ with finite mixed absolute moments ($\mathbb{E}|X^\alpha| < \infty\ \forall \alpha \in \mathbb{N}^N$) via $\tilde{f}(P)(B) = P(f^{-1}(B))$ for any $B \in \mathcal{B}(\mathbb{R}^N)$.

Lemma 20 *The map $\tilde{f} : \mathcal{M}^G \to \tilde{\mathcal{M}}$ is one-to-one.*

Second, suppose $P, Q \in \mathcal{M}_*^G$, $P \neq Q$, and $s(P) = s(Q)$ that satisfy (8), the conditions of the Theorem. By Lemma 20, $\tilde{f}(P) \neq \tilde{f}(Q)$, and by definition the latter measures have all their mixed (ordinary N-dimensional) moments identical and satisfying the conditions of the extended Carleman theorem (§C). (Note that the definition of $\mathcal{M}^*$ in [8] and Definition 8 are equivalent by Proposition 9.) Thus, according to that theorem, $\tilde{f}(P)$ is determinate, i.e. $\tilde{f}(P) = \tilde{f}(Q)$, which contradicts our previous observation.

The proof of the denseness results closely parallels that of Theorem 2.3 of [8] (§C): Let $1 \leq p < \infty$ be fixed and let $h \in L_q^G(W,P)$, where $1/q + 1/p = 1$, and such that

$$\int_W r(x)h(x)dP(x) = 0 \tag{9}$$

$\forall r \in \mathbb{C}[W]^G$. In order to prove that $h = 0$ P-a.s., we first note that due to G-invariance of h combined with Proposition 15, there exists $\tilde{h} : \mathbb{R}^N \to \mathbb{C}$ such that $h = \tilde{h}(f)$. Next, following [8], we perform the following Fourier-like transform:

$$\hat{\xi}_h(\lambda) = \int_W e^{i(\lambda, f(x))}h(x)dP(x) = \int_{\mathbb{R}^N} e^{i(\lambda,y)}\tilde{h}(y)d[\tilde{f}(P)](y), \tag{10}$$

resulting in a smooth function on $\mathbb{R}^N$. All derivatives of this function vanish at $0 \in \mathbb{R}^N$ since (9) implies

$$\int_{\mathbb{R}^N} y^\alpha \tilde{h}(y)d[\tilde{f}(P)](y) = 0,\ \forall \alpha \in \mathbb{N}^N.$$

From this point, the corresponding part of the proof in [8] applies to conclude that under the hypotheses of the present Theorem, and based on Theorem 2.1 of [8] (see §C), $\hat{\xi}_h(\lambda)$ is identically 0. This in turn implies that $\tilde{h} = 0$ $\tilde{f}(P)$-a.s., which finally implies that $h = 0$ P-a.s.

The denseness of $\mathrm{Span}_{\mathbb{C}}\{e^{i(\lambda,f)}|\lambda \in S\}$ can be proved by a similar chain of arguments, replacing λ in the right-hand side of (10) by $\lambda + a$, where $a \in \mathrm{Interior}(\bar{S})$. $\diamond$

Example 1 continued.
Let $\mathcal{M}^C$ be the set of positive Borel measures with supports in $C = \{(w_1, \ldots, w_m) \in \mathbb{R}^m : w_i \geq 0, i = 1, \ldots, m\}$, the positive cone relative to the standard basis, and let $\mathcal{M}^C_* = \mathcal{M}_* \cap \mathcal{M}^C$. Then Lemma 20 applies to show $\mathcal{M}^G \cong \mathcal{M}^C$, and $\mathcal{M}^G_* \cong \mathcal{M}^C_*$ $\tilde{f}(\mathcal{M}^G) = \mathcal{M}^C$, and $\tilde{f}(\mathcal{M}^G_*) = \mathcal{M}^C_*$.

3.1 Integral criteria for G-invariant determinacy

In [8], it is argued that integral criteria for determinacy are more convenient in practice than series conditions such as Carleman's conditions, and the notion of *quasi-analytic weights* is introduced in order to formulate suitable integral conditions. Thus, following [8]:

Definition 21 *A* quasi-analytic weight on W *is a bounded nonnegative function* $w : W \to \mathbb{R}$ *such that*

$$\sum_{k=1}^{\infty} \frac{1}{||(v_j, x)^k w(x)||_\infty^{1/k}} = \infty$$

for $j = 1, \ldots, m$ *and* $v_1, \ldots, v_m$*, some basis for* W*.*

We next provide simple generalizations of Theorems 4.1 and 4.2 of [8] (§C) that provide sufficient integral conditions for determinacy by invariant moments. We omit proofs of these results since they are straightforward analogs of their prototypes in [8] and are based on the same "change of variable" argument that we used to prove Theorem 19.

Theorem 22 *Let* $P \in \mathcal{M}^G_*$ *be such that*

$$\int_W w(f(x))^{-1} dP < \infty$$

for some measurable quasi-analytic weight on $\mathbb{R}^N$*. Then* P *is determinate by its* G*-invariant moments. Furthermore,* $\mathbb{C}[W]^G$ *and* $\mathrm{Span}_{\mathbb{C}}\{e^{i(\lambda,f)} | \lambda \in S\}$ *are dense in (complex)* $L^G_p(W, P)$*, for* $1 \leq p < \infty$ *and for every* $S \subset \mathbb{R}^N$ *which is somewhere dense.*

Following [8], we point out that due to the rapidly-decreasing behavior of w, the premise of the Theorem implies that P is necessarily in $\mathcal{M}^G_*$.

Theorem 23 *For* $j = 1, \ldots, N$*, let* $R_j > 0$ *and let a non-decreasing function* $\rho_j : (R_j, \infty) \to \mathbb{R}^+$ *of class* C^1 *be such that*

$$\int_{R_j}^{\infty} \frac{\rho_j(s)}{s^2} ds = \infty.$$

Define $h_j : \mathbb{R} \to \mathbb{R}^+$ *by*

$$h_j(x) = \begin{cases} \exp\left(\int_{R_j}^{|x|} \frac{\rho_j(s)}{s} ds\right) & \textit{for} \quad |x| > R_j \\ 1 & \textit{for} \quad |x| \leq R_j. \end{cases}$$

Let A *be an affine automorphism of* $\mathbb{R}^N$*. If* P *is a positive Borel measure on* W *such that*

$$\int_W \prod_{j=1}^{N} h_j((Af(x))_j) dP(x) < \infty,$$

then P *is determinate by its* G*-invariant moments. Also,* $\mathbb{C}[W]^G$ *and* $\mathrm{Span}_{\mathbb{C}}\{e^{i(\lambda,f)}|\lambda \in S\}$ *are dense in (complex)* $L_p^G(W, P)$*, for* $1 \leq p < \infty$ *and for every* $S \in \mathbb{R}^N$ *which is somewhere dense.*

We conclude this part by pointing out that other integral criteria discussed in [8] also have their G-invariant formulations similar to the above ones. Thus, for example, Theorem 4.3 of [8] provides a significantly weakened version of the following classical condition for determinacy:

$$\int_W \exp(||x||) dP(x) < \infty$$

Both, the classical condition and its weakened versions due to [8], easily incorporate the G-invariant case by the appropriate adjustment of the radial integrands via: $||x|| \mapsto ||f(x)||$.

4 Sequential G-invariant modeling

From now on we specialize our discussion to probability measures $\mathcal{P}$. The following result lays a foundation for modeling invariant distributions via (invariant) moment constraints.

Theorem 24 *Let a sequence of* G*-invariant probability measures* $\{P_l\}_{l=1}^{\infty} \subset \mathcal{P}^G$ *be such that*

$$\forall \alpha \in \mathbb{N}^N \lim_{l \to \infty} \mathbb{E}_{P_l} f^{\alpha} = s_{\alpha}. \tag{11}$$

Assume that there can exist at most one G*-invariant* P *with such* s_{α}*. Then, such* P *indeed exists and* $P_l \Rightarrow P$*.*

Note that such P would necessarily be in $\mathcal{M}_*^G$.
Proof. Clearly ([12]), (11) implies that the m families of marginals of P_l's are individually *tight*, which immediately implies that the family $\{P_l\}_{l=1}^{\infty}$ is itself *tight*, and therefore ([3]) contains a weakly convergent subsequence. Since every subsequential

limit must also be G-invariant and have the same moments s_α, all such limits must be equal to each other by the uniqueness hypothesis of the Theorem. We take P to be the common value of those limits and finish the proof by invoking the well-known fact [3] that a tight sequence whose all (weak) subsequential limits are equal, converges weakly to that common measure. $\diamond$

We next introduce notation to describe G-invariant models based on the Entropy Maximization Principle (§1). Let a probability measure P be absolutely continuous with respect to some positive σ-finite reference measure μ, $P \ll \mu$, and let p be a density $dP/d\mu$. Let $H_\mu(P) = -\int_W p(x)\log p(x) d\mu(x)$ be the entropy of P relative to μ (for P discrete, a natural choice for μ is the counting measure on Ω, the support of P: $H(P) = -\sum_\Omega p(x)\log p(x)$ (the Shannon's entropy), and for P continuous - the Lebesgue measure on Ω: $H(P) = -\int_\Omega p(x)\log p(x)dx$). In the absence of ambiguity, we will suppress the reference measure in the subscript. Thus, let $D(P\|Q) = \int_W p(x)\log(p(x)/q(x))d\mu(x)$ stand for the Kullback-Leibler divergence between two probability measures P and Q with densities p and q relative to μ.

Proposition 25 *Let P have a density p relative to μ. Then*

$$0 \le H(P) \le H(\mathcal{R}^*(P)) \le H(P) + \log|G|.$$

The equality in place of the second inequality occurs if and only if P is G-invariant.

Let $\mathcal{F}$ be a finite set of (measurable) real-valued functions on (G-invariant) Ω, and $\{\nu_\phi \in \mathbb{R}\}_{\phi\in\mathcal{F}}$. Let

$$P_{\mathcal{F},\nu} = \arg\max_{\substack{P':\mathbb{E}_{P'}\phi=\nu_\phi \\ \forall\phi\in\mathcal{F}}} H(P'), \tag{12}$$

a *maximum entropy distribution* relative to the above constraints. Since we are going to work with (invariant) moment constraints (on P') of the form $\mathbb{E}_{P'}f^\alpha = \mathbb{E}_P f^\alpha$; $\alpha \in A \subset \mathbb{N}^N$, for some fixed measure P, we will write P_A for the maximum entropy distribution in such cases.

Theorem 26 *Let P be a probability measure on W supported on G-invariant Ω and having a density relative to some μ. Assume that $H_\mu(P) < \infty$ and that $\mathcal{R}^*(P)$ is G-determinate. (Note that G-invariance of Ω implies that $\mathcal{R}^*(P)$ is also a probability measure on Ω.) Let $f_1, \ldots, f_N$ be a minimal generating set for $\mathbb{R}[W]^G$. Let $A_1 \subset A_2 \subset \ldots$ be such that $\cup_{l=1}^\infty A_l = \mathbb{N}^N$ and that the corresponding maximum entropy problems* (12) *with $\nu_{f^\alpha} = \mathbb{E}_P f^\alpha$ $\alpha \in A_l$ have solutions $P_l = P_{A_l}$. Then $P_l \Rightarrow \mathcal{R}^*(P)$.*

Proof. First, note that for any (measurable) G-invariant function ϕ, $\mathbb{E}_P\phi = \mathbb{E}_P\mathcal{R}(\phi) = \mathbb{E}_{\mathcal{R}^*(P)}\phi$ (Proposition 10). Second, note that if P_l exists, then it is necessarily G-invariant (Proposition 25). This can also be seen from the exponential form of $p_l(x)$, the density of the maximum entropy distribution:

$$p_l(x) \;=\; \exp\left(\sum_{\alpha\in A_l}\lambda_\alpha f^\alpha(x) - \psi(\lambda)\right) \tag{13}$$

$$\psi(\lambda) = \log \int_\Omega \exp\left(\sum_{\alpha\in A_l} \lambda_\alpha f^\alpha(x)\right) d\mu(x) \quad (14)$$

$$\lambda = (\lambda_{\alpha_1}, \ldots, \lambda_{\alpha_{|A_l|}}) : \mathbb{E}_{P_l} f^\alpha = \mathbb{E}_P f^\alpha;\ \alpha \in A_l \quad (15)$$

Finally, Theorem 24 applies to finish the proof. $\diamond$

The above Theorem in its present form is too abstract to be immediately applied in practice. In general, the existence of a solution to the maximum entropy problem cannot be taken for granted as can be seen from the following well-known example [4], [6], [20]: There is no solution to the maximum entropy problem on $\mathbb{R}$ constraining only the mean. However, constraining additionally the second moment gives a unique maximum entropy distribution that is the normal distribution with the given first two moments. Thus, in order to produce feasible sets A_l as above, one may need to make more assumptions. For example, one sufficient condition for the well-posedness of the maximum entropy problems with moment constraints is given in [20] for Ω open but otherwise arbitrary. Using our notation, let $\Lambda(A_l) = \{\lambda \in \mathbb{R}^{|A_l|} :\ \psi(\lambda) < \infty\}$, where $\psi(\lambda)$ is as in (14) and the reference measure is the Lebesgue one. The condition then is that $\Lambda(A_l)$ be open, i.e. $\Lambda(A_l) \cap \partial\Lambda(A_l) = \emptyset$. Also, it is often a mild restriction in practice to assume compactness of Ω. In this case, first of all, the conclusion of Theorem 24 always holds (provided that $\{P_l\}_{l=1}^\infty$ are all supported on the same Ω) due to the uniform approximation of compactly-supported continuous functions by polynomials. Secondly, it can be seen that if one additionally required that p^G, the density of $\mathcal{R}^*(P)$ with respect to the Lebesgue measure on Ω, be non-zero almost everywhere on Ω and have finite entropy, then all subsets $A \in \mathbb{N}^N$ would give rise to well-posed maximum entropy problems with exponential solutions (13).

Alternatively, it is noted and used in [37] that all empirical distributions $\hat{P}$ on $[0, 1]$ give rise to well-posed maximum entropy problems with constraints on any set of first J moments (in order to keep all such constraints *active*, the sample data may not be identically equal to 1). Based on the multidimensional version of the Hausdorff's moment problem (see, for example, [23]) it appears that these latter one-dimensional results (Theorem 1 of [29] and Lemma 1 of [37]) also generalize to higher dimensions, in which case Theorem 28 below generalizes appropriately to include the case of empirical moment constraints. However, since in practice the use of the computer requires discretization of Ω, we leave aside the discussion of the well-posedness of the maximum entropy problem in the continuous case. Also, in our motivating example (§7) Ω is finite, and we therefore focus on this case in §5.

We next present a modification of Theorem 26 on accelerated convergence toward the target distribution. For completeness, we present the continuous version of this result before an appropriate algorithm for the finite case. We need the following notation: Let $\prec$ be a *total well-ordering* of $\mathbb{N}^N$ such that α, β, $\gamma \in \mathbb{N}^N$ and $\alpha \prec \beta$ imply $\alpha + \gamma \prec \beta + \gamma$ ([7]).

Definition 27 *A monomial ordering on* $\{f^\alpha\}_{\alpha\in\mathbb{N}^N}$ *is any relation* $\prec$ *on* $\mathbb{N}^N$ *as above.*

For $\alpha \in \mathbb{N}^N$ and for $A \subset \mathbb{N}^N$ define also

$$\begin{aligned} d_\prec(\alpha,\beta) &= |\{\gamma \in \mathbb{N}^N : \ \min_\prec(\alpha,\beta) < \gamma \le \max_\prec(\alpha,\beta)\}|, \\ d_\prec(\alpha,A) &= \min_{\beta\in A} d_\prec(\alpha,\beta), \end{aligned}$$

a discrete distance relative to $\prec$. Let r be a positive integer parameter.

Theorem 28 *Let* P *be a probability measure supported on compact and* G*-invariant* Ω*. Assume* p *is a density of* P *relative to some* μ *and that* $H_\mu(P) < \infty$ *and* $p^G > 0$ *(*μ*-) almost everywhere on* Ω*. Fix a monomial ordering* $\prec$ *(Definition 27), and let* $\mathbf{0} = (0,\dots,0) \in \mathbb{N}^N$*. Define* $P_l = P_{A_l}$ *in accordance with* (12) *and the scheme below:*

$$\begin{aligned} A_1 &= \{\alpha_1^*\} \text{ where } \alpha_1^* = \arg\min_{\alpha: d_\prec(\alpha,\mathbf{0})\le r} D(P\|P_{\{\alpha\}}) \\ A_l &= A_{l-1}\cup\{\alpha_l^*\} \text{ for } l = 2,\dots, \text{ where } \alpha_l^* = \arg\min_{\alpha: d_\prec(\alpha,A_{l-1})\le r} D(P\|P_{A_{l-1}\cup\{\alpha\}}). \end{aligned}$$

Then $P_l \Rightarrow \mathcal{R}^*(P)$*.*

Proof. Based on the above discussion of well-posedness of the maximum entropy problem, the conditions of the Theorem guarantee the existence and uniqueness of maximum entropy distributions for all finite subsets A and in particular for A_l, $l = 1,2,\dots$ as above. (The optimization of D is over a finite set, and hence α_l^* is always well-defined.) Compactness of Ω results in G-determinacy of $\mathcal{R}^*(P)$, and application of Theorem 26 completes the proof. ⋄

Remark 29 *If* $P \ne \mathcal{R}^*(P)$*,* $D(P\|Q)$ *need not in general equal* $D(\mathcal{R}^*(P)\|Q)$ *even if* $Q = \mathcal{R}^*(Q)$*. However, one should not worry about replacing the target distribution* P *by its symmetrized version thanks to the additivity of* D *on nested exponential models* $M_0 \subset M_1 \subset M_2$*:* $D(P_2|P_0) = D(P_2|P_1) + D(P_1|P_0)$*, which in our case gives:*

$$D(P\|P_A) = D(P\|\mathcal{R}^*(P)) + D(\mathcal{R}^*(P)\|P_A). \tag{16}$$

Hence, minimizing $D(P\|P_{A_{l-1}\cup\{\alpha\}})$ *is equivalent to minimizing* $D(\mathcal{R}^*(P)\|P_{A_{l-1}\cup\{\alpha\}})$*.*

5 Adaptive minimax learning of symmetric distributions

We now specialize this modeling scheme to Ω finite, which is often the case in practice.

Fix an enumeration $k(\cdot) : \Omega = \{\omega_1,\dots,\omega_K\} \to \mathbb{Z}_K$. Relative to this enumeration, identify f^α with $(f^\alpha(\omega_1),\dots,f^\alpha(\omega_K)) \in (\mathbb{R}^\Omega)^G$.

Proposition 30 *Let* $M = |\mathcal{S}_\Omega|$*. There exist* $\alpha_1,\dots,\alpha_M \in \mathbb{N}^N$ *such that* $\{f^{\alpha_k}\}_{k=1}^M$ *is a basis for* $(\mathbb{R}^\Omega)^G$*.*

Proof. Clearly, $(\mathbb{R}^\Omega)^G$ has a basis in terms of G-invariant polynomials. One such basis, for example, is given by $\{\mathbb{I}_\mathcal{O}\}_{\mathcal{O}\in\mathcal{S}_\Omega}$, the set of all the orbit indicators computed as follows:

$$\begin{aligned} \mathbb{I}_\mathcal{O}(x) &= \frac{\tilde{h}(x)}{\bar{f}(\mathcal{O})}, \text{ where} \\ \tilde{h}_\mathcal{O}(x) &= \prod_{\substack{\mathcal{O}'\in\mathcal{S}_\Omega \\ \mathcal{O}'\neq\mathcal{O}}} \sum_{i=1}^m \left[f_i(x) - \bar{f}_i(\mathcal{O}')\right]^2, \end{aligned} \tag{17}$$

and $\bar{f}([w]) = f(w)$ $[w] \in \mathcal{S}_\Omega$ (Proposition 15). Since $\mathbb{I}_\mathcal{O}(x) \in \mathbb{R}[W]^G$ and $M < \infty$, then the set of all $f^\alpha(x)$'s participating in polynomial expansions of $\tilde{h}_\mathcal{O}$ is finite. Evidently the corresponding set of K-dimensional vectors f^α spans $(\mathbb{R}^\Omega)^G$ and therefore contains a sought basis with M elements. $\diamond$

We introduce more notation:

Definition 31 *Let $A \subset \mathbb{N}^N$ and $d \in \mathbb{N}$ and $\prec$ be a monomial order.*

$$B_\prec(A,d) = \{\alpha \in \mathbb{N}^N : \ d_\prec(A,\alpha) \leq d\}, \qquad B^\perp_\prec(A,d) = \{\alpha \in B_\prec(A,d) : \ \alpha \perp A\},$$
$$C^\perp_\prec(A,r) = \cap_{d\in\mathbb{N}:\ |B^\perp_\prec(A,d)|\geq r} B^\perp_\prec(A,d), \qquad C^*_\prec(A,r) = \cup_{0<r'\leq r} C^\perp_\prec(A,r'),$$

where for $A \subset \mathbb{N}^N$ and $\beta \in \mathbb{N}^N$ we write $\beta \perp A$ if $\{f^\alpha\}_{A\cup\{\beta\}}$ is a linearly independent system.

Adaptive minimax learning of symmetric distributions

$$\begin{aligned} A_0 = \{\mathbf{0}\},\ A_l &= A_{l-1} \cup \{\alpha^*_l\} \text{ for } l = 1, 2, \ldots, \\ \text{where } \alpha^*_l &= \arg\min_{\alpha: C^*_\prec(A_{l-1},r)} D(P \| P_{A_{l-1}\cup\{\alpha\}}). \end{aligned} \tag{18}$$

Then $P_{M-1} = \mathcal{R}^*(P)$.

Remark 32

1.) P_0 can be included above as the uniform distribution on Ω: it maximizes the entropy without constraints.

2.) Suppose that P is an empirical distribution based on an i.i.d. sample. It can then be easily verified ([24]) that P_l gives the maximum likelihood estimate (of the data generating distribution) relative to the parametric family (13) *(parametrized by λ). In particular, $\mathcal{R}^*(P)$ gives the maximum likelihood estimate relative to $\mathcal{P}^G$.*

3.) At each step $l = 1, 2, \ldots, M-1$ the procedure "explores" upto r new dimensions each of which is linearly independent of $\mathrm{Span}\{f^\alpha : \alpha \in A_l\}$*, the span of the current model "factors". A dimension that promises a fastest approach toward $\mathcal{R}^*(P)$ (or, equivalently, toward P), is chosen and the current model is augmented accordingly (ties being broken arbitrarily).*

4.) Let $D_l = D(P\|P_l)$, *and* $H_l = H(P_l)$, *for* $l = 0, \ldots, M-1$. *It can be easily seen that* $\{D_l\}$ *and* $\{H_l\}$ *are strictly decreasing and* $D_{M-1} = D(P\|\mathcal{R}^*(P))$ *and* $H_{M-1} = H(\mathcal{R}^*(P))$. *Clearly, if* $\alpha \not\perp A_l$, *then* $D_l = D(P\|P_{A_l \cup \{\alpha\}})$, *i.e. adding a linearly dependent factor does not change the model and is therefore avoided by the minimization phase of the procedure.*

Even if $\mathcal{R}^*(P)$ is accepted as a working model of P, the utility of the above procedure would still be limited to simply finding $p^G(f(x))$, an analytic form for $\mathcal{R}(p)$. In fact, computing and working with $\mathcal{R}(p)$ (see §6.2) as the K-dimensional vector may also be acceptable depending on the application. Next, we explain how the ideas of adaptive minimax learning can combine with a variety of automated model selection schemes, which we view as the *main application* of our work.

Model selection is about balancing between fitting the data well and keeping the complexity of the model low. There are several criteria addressing this problem, and, for example, the *Minimal Description Length Principle* [18] appears to suite well our context. In short, many model selection principles including the MDL one, can be viewed as a minimization of a cost function C that balances the two penalties, namely for deficiency and for excess of fit. We now reexamine and generalize our "Adaptive minimax learning" with a view toward model selection.

5.1 Present approach based on D

In its present form, our "Adaptive minimax learning" is essentially a variation of the *minimax learning* [39], [40], [41] originally introduced for texture modeling. This latter principle considers image filter banks (in our notation, sets $\mathcal{F}$ of constraints ϕ), each corresponding to its maximum entropy model (maximization step). One then measures the Kullback-Leibler distance D from the empirical, or target, distribution to each of such maximum entropy models, and the model with the minimum distance is selected. In practice one fixes a very large but finite pool of filters to consider, and the cardinality of $\mathcal{F}$. Since $|\mathcal{F}|$ equals the number of model parameters, it can be thought of as a measure of model complexity that must be set in advance.

Based on our "Adaptive minimax learning", we propose a model selection that selects efficiently $p^G_{A_l}$, a suboptimal model within the class of the G-invariant ones, declaring it our best G-invariant approximation to the target P. Specifically, we propose to halt the model construction algorithm at step l as soon as $D(\mathcal{R}^*(P)\|P^G_{A_l}) \leq D(P\|\mathcal{R}^*(P))$, or, equivalently, $D(P\|P^G_{A_l}) \leq 2 * D(P\|\mathcal{R}^*(P))$.

We then propose to repeat the same minimax learning procedure using ordinary moments instead of G-invariant ones, and stopping at l. We then choose between $P_{A'_l}$ and $P^G_{A_l}$, the resulting generic and G-invariant models, respectively,based on their fit only: $D(P\|P_{A'_l}) \gtrless D(P\|P^G_{A_l})$.

5.2 General approach based on cost C

The choice of $D(P\|\cdot)$ in the minimization step of the above procedure is not the only one possible. In fact, it is precisely for that reason that D always drives the model selection toward the extreme fit, that we had to introduce an ad hoc stopping rule in §5.1 to prevent the overfitting. Suppose one employs a cost function C that favors neither extreme. For example, C could be a description length as in MDL [18]. One then modifies the adaptive minimax learning by using C instead of D in the minimization step, and terminating the model construction once C cannot be minimized further. Again, if one wants "to test" appropriateness of the G-invariance, one can repeat the construction with the ordinary moments in order to see if C can be further reduced outside the G-invariant class.

Clearly, this framework as well as the one of §5.1 applies to other situations, where f need no longer be generators of invariant polynomials.

6 Computational issues

6.1 Computing minimal generating sets

In Appendix E we compute f "by hand" for our example in §7. However, algorithms exist to compute such generating sets in a systematic fashion (see, for example, [9], [34] and [36]) and there are also computer algebra tools implementing those algorithms: *Gap* [14], *INVAR* [22], *Macaulay2* [17], *Magma* [5], to name a few.

6.2 Computing $\mathcal{R}$ and $\mathcal{S}_\Omega$

The operator defined in (6) and used throughout this work admits a natural decomposition

$$\mathcal{R} \;=\; \pi_2 \circ \pi_1, \tag{19}$$

where $\pi_1 : \mathbb{R}^\Omega \to \mathbb{R}^{\mathcal{S}_\Omega}$ surjectively and $\pi_2 : \mathbb{R}^{\mathcal{S}_\Omega} \to \mathbb{R}^\Omega$ injectively as follows:

$$(\pi_1(h))(\mathcal{O}) \;=\; \frac{1}{\sqrt{|\mathcal{O}|}} \sum_{\omega \in \mathcal{O}} h(\omega) \tag{20}$$

$$(\pi_2(\tilde{h}))(\omega) \;=\; \frac{1}{\sqrt{|[\omega]|}} \tilde{h}([\omega]). \tag{21}$$

Simply speaking, this operator averages a function h over the G-invariant orbits, in particular it computes the maximum likelihood estimate relative to $\mathcal{P}^G$ based on an i.i.d. sample (Remark 32). Thus, to implement this averaging with the computer, one needs to index the orbits of $\mathcal{S}_\Omega$. We briefly comment on two types of such indexings. The first type is based on a naive generation-elimination via $\rho : G \hookrightarrow GL(W)$, the matrix representation of G (for a concrete example, see (31)). Below is a sketch of a naive algorithm that computes $\chi : \mathbb{Z}_K \to \mathbb{Z}_M$, $(M = |\mathcal{S}_\Omega|)$, an orbit indexing map, assuming some ordering $k(\cdot)$ of Ω (§5):

$\chi(m) \Leftarrow 1$, $m = 1, \ldots, K$
$l = 0$, $m = 0$
$R = \{m' : m < m' < K, \chi(m') = 0\}$
while $R \neq \emptyset$ **do**
  $m \Leftarrow \min R$, $l \Leftarrow l + 1$
  $\chi\left(k\left(\rho(g) \cdot \omega_m\right)\right) = l$
end while

The second approach to calculating $\mathcal{S}$ is more algebraic. Recall that $\mathbb{I}_{\mathcal{O}}$, $\mathcal{O} \in \mathcal{S}$ can be computed using minimal generators f as in (17). Next note that writing $\mathbb{I}$ and h as K-dimensional column vectors, we have $(\pi_1(h))(\mathcal{O}) = \mathbb{I}_{\mathcal{O}}^{tr} \times h/\sqrt{|\mathcal{O}|}$. Thus, $\pi_1(h)$ can be computed as $\pi_1 \times h$, where, abusing the notation, π_1 become the matrix whose rows are $\mathbb{I}_{\mathcal{O}}^{tr}$, the transposed orbit indicator vectors renormalized by the square root of the orbit size. It can easily be seen that in this matrix formulation, $\pi_2 = \pi_1^{tr}$, which means that the corresponding linear operators are adjoint. Thus, we obtain the matrix representation of $\mathcal{R} = \pi_1^{tr} \times \pi_1$.

6.3 Entropy maximization. Sequential approach and dimensionality reduction.

To solve for λ, one uses numerical methods that require an initial guess. A certain computational saving has been noticed in experiments of [24] and [37] involving nested maximum entropy models with moment constraints. Namely, suppose $\lambda^{(l)} = (\lambda_1^{(l)}, \ldots, \lambda_l^{(l)})$ have been found at step l, i.e. the distribution P_l is computed, and suppose an $l+1$-st constraint f^α is added. One then seeks $\lambda^{(l+1)} = (\lambda_1^{(l+1)}, \ldots, \lambda_{l+1}^{(l+1)})$. It then often turns out in practice that $(\lambda_1^{(l)}, \ldots, \lambda_l^{(l)}, 0)$ is a good initial guess for $\lambda^{(l+1)}$. It is also noticed in [24] that the minimization step contributes significantly to the observed continuity in λ, i.e. when the "most informative" moments are added first, then the subsequent steps affect the corresponding parameters progressively less. Thus, the overall computations stay comparable to those of the baseline procedure without the minimization feature: Specifically, on one hand, the minimization requires at each step computing upto r models instead of just one, but on the other hand, such computations require progressively less time as the number of constraints grows.

We now show that the G-invariance allows us to translate the entropy maximization problem on the original space $\Omega \subset \mathbb{R}^m$ to the quotient space $\mathcal{S}_\Omega$, which for nontrivial G is "smaller" than Ω. We also show that in the most important in practice case of Ω finite, the dimension of the optimization problem indeed reduces from

$|\Omega|$ to $|\mathcal{S}_\Omega|$.

Let

$$\tilde{\mathcal{B}} = \{\tilde{B} \subset \mathcal{S}_W | \cup_{\mathcal{O}\in\tilde{B}} \mathcal{O} \in \mathcal{B}\}, \tag{22}$$

which can be seen to be a σ-algebra on $\mathcal{S}_W$. Let $\tilde{\mathcal{M}}$ be the image of the following operator:

$$\pi_1^* : \mathcal{M} \to \tilde{\mathcal{M}} \quad \text{via} \quad \pi_1^*(P)(\tilde{B}) = P(\cup_{\mathcal{O}\in\tilde{B}}). \tag{23}$$

Note that π_1^* maps $\mathcal{P}$, the probability measures on $\mathcal{B}$, to $\tilde{\mathcal{P}}$, the probability measures on $\tilde{\mathcal{B}}$. π_1^* is also surjective since $\pi_1^* \circ \pi_2^* = id$, where

$$\pi_2^* : \tilde{\mathcal{M}} \to \mathcal{M} \quad \text{via} \quad \pi_2^*(\tilde{P})(B) = \int_{\mathcal{S}} \frac{|B \cap \mathcal{O}|}{|\mathcal{O}|} d\tilde{P}(\mathcal{O}). \tag{24}$$

The right hand side of (24) is well-defined as can be seen from the following:

Proposition 33 *Let* $h_B(\mathcal{O}) = \frac{|B\cap\mathcal{O}|}{|\mathcal{O}|}$. *Then* $h_B : \mathcal{S}_W \to \mathbb{R}$ *is* $\tilde{\mathcal{B}}$*-measurable, and* $h_B \circ [w] : W \to \mathbb{R}$ *is* $\mathcal{B}$*-measurable.*

We now observe the following:

Proposition 34

$$\mathcal{R}^* = \pi_2^* \circ \pi_1^*, \text{ and } \pi_1^* : \mathcal{M}^G \to \tilde{\mathcal{M}} \text{ and } \pi_2^* : \tilde{\mathcal{M}} \to \mathcal{M}^G \text{ are bijective.}$$

Next, we define the adjoints of π_1^* and π_2^*:

$$\pi_2 f(\mathcal{O}) = \frac{1}{|\mathcal{O}|} \sum_{w\in\mathcal{O}} f(w) \qquad \pi_1 \tilde{f}(w) = \tilde{f}([w]), \tag{25}$$

and notice:

Proposition 35 π_1 *and* π_2 *are indeed adjoints of* π_1^* *and* π_2^**, respectively, and*

$$\mathcal{R} = \pi_1 \circ \pi_2.$$

The last two ingredients needed to state the main result of this section are as follows:

$$\tau^*\mu(\tilde{B}) = \int_W \frac{\mathbb{I}_{\tilde{B}}([w])}{|[w]|} d\mu(w) \qquad \tau f(\mathcal{O}) = \sum_{w\in\mathcal{O}} f(w), \tag{26}$$

Theorem 36 *Let* $V : \mathbb{R}^m \to \mathbb{R}^J$ *be measurable and* G*-invariant. Then*

$$\underset{\substack{Q\in\mathcal{P} Q\ll\mu \\ \mathbb{E}_Q V = \mathbb{E}_P V}}{\text{argmax}} H_\mu(Q) \;=\; \pi_2^* \left\{ \underset{\substack{\tilde{Q}\in\tilde{\mathcal{P}} \;\; \tilde{Q}\ll\tau^*\mu \\ \mathbb{E}_{\tilde{Q}}\pi_2 V = \mathbb{E}_{\pi_1^* P}\pi_2 V}}{\arg\max} \left[H_{\tau^*\mu}(\tilde{Q}) + \mathbb{E}_{\tilde{Q}}(\log(|\mathcal{O}|)\right] \right\}.$$

Proof.

$$
\begin{aligned}
\underset{\substack{Q\in\mathcal{P}\ \ Q\ll\mu \\ \mathbb{E}_Q V=\mathbb{E}_P V}}{\operatorname{argmax}} H_\mu(Q) &\overset{\text{by Propositions 10, 25}}{=} \arg\max_{\substack{Q\in\mathcal{P}^G\ \ Q\ll\mu \\ \mathbb{E}_Q V=\mathbb{E}_P V}} H_\mu(\mathcal{R}^*Q) \\
&\overset{\text{by Proposition 34}}{=} \arg\max_{\substack{Q\in\mathcal{P}^G\ \ Q\ll\mu \\ \mathbb{E}_{\pi_2^*\circ\pi_1^*Q}V=\mathbb{E}_{\pi_2^*\circ\pi_1^*P}V}} H_\mu(\pi_2^*\circ\pi_1^*Q) \\
&\overset{\text{by Propositions 34, 35}}{=} \pi_2^*\left\{\arg\max_{\substack{\tilde{Q}\in\tilde{\mathcal{P}}\ \ \tilde{Q}\ll\tau^*\mu \\ \mathbb{E}_{\tilde{Q}}\pi_2 V=\mathbb{E}_{\pi_1^*P}\pi_2 V}} H_\mu(\pi_2^*\tilde{Q})\right\} \\
&= \pi_2^*\left\{\arg\max_{\substack{\tilde{Q}\in\tilde{\mathcal{P}}\ \ \tilde{Q}\ll\tau^*\mu\ \ \gamma=\frac{d\pi_2^*\tilde{Q}}{d\mu} \\ \mathbb{E}_{\tilde{Q}}\pi_2 V=\mathbb{E}_{\pi_1^*P}\pi_2 V}} -\int_W \gamma(w)\log(\gamma(w))d\mu\right\} \\
&\overset{\text{by Proposition 11}}{=} \pi_2^*\left\{\arg\max_{\substack{\tilde{Q}\in\tilde{\mathcal{P}}\ \ \tilde{Q}\ll\tau^*\mu\ \ \gamma=\frac{d\pi_2^*\tilde{Q}}{d\mu} \\ \mathbb{E}_{\tilde{Q}}\pi_2 V=\mathbb{E}_{\pi_1^*P}\pi_2 V}} -\int_W \frac{\tau\gamma([w])}{|[w]|}\log\left(\frac{\tau\gamma([w])}{|[w]|}\right)d\mu\right\} \\
&= \pi_2^*\left\{\arg\max_{\substack{\tilde{Q}\in\tilde{\mathcal{P}}\ \ \tilde{Q}\ll\tau^*\mu\ \ \gamma=\frac{d\pi_2^*\tilde{Q}}{d\mu} \\ \mathbb{E}_{\tilde{Q}}\pi_2 V=\mathbb{E}_{\pi_1^*P}\pi_2 V}} -\int_{\mathcal{S}_W} \tau\gamma(\mathcal{O})\log(\tau\gamma(\mathcal{O}))d\tau^*\mu \right. \qquad (27) \\
&\qquad \left. +\int_{\mathcal{S}_W} \tau\gamma(\mathcal{O})\log(|\mathcal{O}|)d\tau^*\mu\right\} \\
&= \pi_2^*\left\{\arg\max_{\substack{\tilde{Q}\in\tilde{\mathcal{P}}\ \ \tilde{Q}\ll\tau^*\mu \\ \mathbb{E}_{\tilde{Q}}\pi_2 V=\mathbb{E}_{\pi_1^*P}\pi_2 V}} \left[H_{\tau^*\mu}(\tilde{Q})+\mathbb{E}_{\tilde{Q}}(\log(|\mathcal{O}|)\right]\right\}. \qquad (28)
\end{aligned}
$$

It follows from (26) that

$$
\int_{\mathcal{S}} \tilde{f}(\mathcal{O})d(\tau^*\mu) = \int_W \frac{\tilde{f}([w])}{|[w]|}d(\tau^*\mu),
$$

hence (27). Also, τ maps probability densities on W relative to μ to probability densities on $\mathcal{S}_W$ relative to $\tau^*\mu$, and $\tau\gamma = d\tilde{Q}/d\tau^*\mu$, hence (28). Note, that $\pi_2\gamma = d\tilde{Q}/d\pi_1^*\mu$ is not a probability density. This fact and also the fact that τ^* preserves uniformity of the reference measure (e.g. counting measures on discrete $\Omega \subset W$ are transformed into counting measures on $\mathcal{S}_\Omega$) are the reasons to use the τ transforms despite the extra term in (28). $\diamond$

Corollary 37 *Let $|\Omega| = K$ and $|\mathcal{S}_\Omega| = M$. Let ρ be the distribution on $\mathcal{S}_\Omega$ defined via $\rho(\{\mathcal{O}\}) = |\mathcal{O}|/K$. Let μ be the counting measure on Ω, and let P be some fixed probability distribution on Ω. Let $V : \Omega \to \mathbb{R}^J$ be G-invariant. Then*

$$\underset{\substack{Q\in\mathcal{P}\\ \mathbb{E}_Q V=\mathbb{E}_P V}}{\operatorname{argmax}} H(Q) = \pi_2^* \left\{ \arg \min_{\substack{\tilde{Q}\in\tilde{\mathcal{P}}\\ \mathbb{E}_{\tilde{Q}}\pi_2 V=\mathbb{E}_{\pi_1^* P}\pi_2 V}} D(\tilde{Q}\|\rho) \right\}.$$

Proof. Rewrite (27) in the proof of the Theorem as follows:

$$\pi_2^* \left\{ \arg \min_{\substack{\tilde{Q}\in\tilde{\mathcal{P}}\ \ \gamma=\frac{d\pi_2^*\tilde{Q}}{d\mu}\\ \mathbb{E}_{\tilde{Q}}\pi_2 V=\mathbb{E}_{\pi_1^* P}\pi_2 V}} \sum_{\mathcal{O}\in\mathcal{S}_W} \tau\gamma(\mathcal{O}) \log\left(\frac{\tau\gamma(\mathcal{O})K}{|\mathcal{O}|K}\right) \right\}$$

$$= \pi_2^* \left\{ \arg \min_{\substack{\tilde{Q}\in\tilde{\mathcal{P}}\\ \mathbb{E}_{\tilde{Q}}\pi_2 V=\mathbb{E}_{\pi_1^* P}\pi_2 V}} D(\tilde{Q}\|\rho) - \log(K) \right\} = \pi_2^* \left\{ \arg \min_{\substack{\tilde{Q}\in\tilde{\mathcal{P}}\\ \mathbb{E}_{\tilde{Q}}\pi_2 V=\mathbb{E}_{\pi_1^* P}\pi_2 V}} D(\tilde{Q}\|\rho) \right\}.$$

$\diamond$

Unlike Theorem 36 that is very general, Corollary 37 emphasizes the practical significance of the main result, i.e. reduction of dimensionality of the original optimization problem. Note that the orbit sizes (or the distribution ρ) become available once the partition $\mathcal{S}_\Omega$ has been computed. Thus, if the original problem is solvable with all $|\lambda_j| < \infty$, one can manipulate the solution to the original problem given by (29) in order to obtain (30), the corresponding solution on $\mathcal{S}_\Omega$.

$$\begin{aligned} \gamma(w) &= \exp\left(\sum_{j=1}^J \lambda_j V_j(w) - \psi(\lambda)\right) \\ \psi(\lambda) &= \log \sum_{w\in\Omega} \exp\left(\sum_{j=1}^J \lambda_j V_j(w)\right) \\ \lambda &= (\lambda_1, \ldots, \lambda_J) : \mathbb{E}_{Q(\lambda)} V_j = \mathbb{E}_P V_j;\ j = 1, \ldots, J, \end{aligned} \tag{29}$$

where we assumed linear independence of $\vec{\mathbf{1}}, V_1, \ldots, V_J$ as K-dimensional real vectors. Thus, except for computing the orbits, the computations required to solve the problem on $\mathcal{S}_\Omega$ are essentially identical to those of entropy maximization: Solving (numerically or by simulation) a system of exponential equations to find the Lagrange multipliers λ. The only difference is therefore the reweighting of the summands of the equations according to the orbit sizes:

$$\begin{aligned}
\tau\gamma(\mathcal{O}) &= |\mathcal{O}| \exp\left(\sum_{j=1}^{J} \lambda_j \tilde{V}_j(\mathcal{O}) - \psi(\lambda)\right) \\
\psi(\lambda) &= \log \sum_{\mathcal{O} \in \mathcal{S}_\Omega} |\mathcal{O}| \exp\left(\sum_{j=1}^{J} \lambda_j \tilde{V}_j(\mathcal{O})\right) \\
\lambda &= (\lambda_1, \ldots, \lambda_J) : \mathbb{E}_{\tilde{Q}(\lambda)} \tilde{V}_j = \mathbb{E}_{\tilde{P}} \tilde{V}_j;\ j = 1, \ldots, J,
\end{aligned} \tag{30}$$

where we used $\tilde{V} = \pi_2 V$, $\tilde{P} = \pi_1^* P$.

Note finally that in the case of Ω finite, the assumption $\Omega \subset \mathbb{R}^m$ and $G \leq GL(m, \mathbb{R})$ is not necessary for the above reduction of dimensionality. Thus, in general Ω can be any finite set with an arbitrary partition $\mathcal{S}$, in which case G can always be recovered from $\mathcal{S}$ as a subgroup of the permutation group $S_{|\Omega|}$. $\mathcal{S}$, on the other hand, may emerge as the set of constancy classes of $V : \Omega \to \mathbb{R}^J$ as one usually defines models in terms of V and not $\mathcal{S}$.

6.4 Construction of $C^*_\prec(A, r)$ from Definition 31

Note that the algorithm (18) refers to the sets $C^*_\prec(A_{l-1}, r)$ that contain as many as possible upto r candidate terms f^α for model refinement. It would therefore help analyze the algorithm if we could, at least for some orders $\prec$, bound (from above) $S(A, r)$, the number of steps required to generate $C^*_\prec(A_{l-1}, r)$. Consider, for example, the *Graded Lex Order*: $\alpha >_{grlex} \beta$ if $\deg(\alpha) = \sum_{n=1}^N \alpha_n > \deg(\beta)$, or $\deg(\alpha) = \deg(\beta)$ and $\alpha >_{lex} \beta$. Let $\deg(A) = \sup\{\deg(\alpha), \alpha \in A\}$. Suppose at step $l < M - 1$ we seek $\alpha \perp A_{l-1}$. We would then like to predict

1. $S(A_{l-1}, 1)$ such that there is at least one $\alpha \perp A_{l-1}$ $\alpha >_{grlex} A_{l-1}$ with $\deg(\alpha) \leq \deg(A_{l-1}, 1) + S(A_{l-1})$.
2. $S(A_{l-1}, r) = \max\limits_{\substack{\alpha >_{grlex} A \\ \deg(\alpha) \leq A + S(A_{l-1}, r-1)}} S(A \cup \{\alpha\})$, for $r < M - l$.
3. S^* the total execution time of the algorithm.

7 Microimage Distributions

We consider an example from the area of *natural image statistics* which, in its broad formulation, studies various statistics defined on digitized images of sufficiently *complex* scenes. For example, we qualify photographs of a landscape or an urban scene as complex, or natural, as opposed to a photograph of an artificially arranged scene of an isolated chair in an otherwise empty room. Statistics of interest are usually *local*, i.e. defined on very small, relative to the image size, regular (e.g. square) subimages, or, *microimages*. Suppose that images and microimages are identified with $I \times I$ and $n \times n$ matrices ($n < I$), respectively, with entries from $\mathcal{C}_L = \{0, \ldots, L-1\}$ (e.g. $L = 256$). We denote the set of microimages by $\tilde{\Omega}_n^L$. Typical studies are based on large collections of digital grey scale images of a particular origin (e.g. optical or range imaging) and a particular domain (e.g. landscapes, terrains) followed by a comparative analysis of findings (e.g. topological and geometrical properties of percentiles). Distributional properties of such statistics are functions of P, the underlying *microimage distributions* on $\tilde{\Omega}_n^L$. Defining P is, however, application dependent and can be quite non obvious as one usually starts with fixing a microimage sampling scheme without worrying about a corresponding *microimage population*. The microimage sampling mechanism then also depends on a number of application-specific factors, and varies from low-density random sampling within the entire image [24] to high-density sampling within certain globally defined regions of interests, or from sampling at regular grid nodes [24] to conditional sampling at high contrast regions [15], [26], and [32]. In principle, every distinct sampling scheme leads to its own definition of the microimage population or, equivalently, P. Remarkably ([24]), certain properties of microimage samples appear stable regardless of the particular sampling scheme and the imaging domain. This, to a certain extent, allows one to think of *the microimage distribution* P. It is this "universal" P whose properties we discuss next.

7.1 The group G of Microimage Symmetries

There has been found ample evidence of P respecting the geometric symmetries of $\tilde{\Omega}_n^L$ (n is typically 3 or 2 and $I = 100, \ldots, 1500$. $\tilde{\Omega}_n^L$ is identified with the square-based *parallelepiped* whose bases correspond to the "all-dark" (0) and "all-bright" ($L-1$) configurations. This evidence includes visual inspection of graphs of various multidimensional local statistics [19], point estimates of probabilities of high contrast patches [15], [26], and P-values of statistical tests [24]. Some symmetries, such as "left-right" and "up-down", are more pronounced than the others, such as, for example, the intensity inversion one. Nonetheless, here we will consider the entire group G of the corresponding transformations, and one can easily specialize the discussion to the subgroups of G.

Thus, we define G via its three generators, r, s, and i: Let r represent the counterclockwise rotation of the square by $\pi/2$, and let s stand for the reflection of the square through its secondary diagonal. The resulting subgroup of G is isomorphic to D_8[1], the *dihedral group* of order 8, with the following presentation $\langle r, s | r^4 = s^2 = 1, rs = sr^3 \rangle$.

[1] We follow the notation of [10] in which D_{2n} stands for the group of all symmetries of a regular n-gon. Another popular notation for this group is D_n.

Recall that composite actions propagate right to left; for example, $rs\omega$ acts on ω by the diagonal reflection s followed by the rotation r.

The last symmetry required to generate G is that with respect to the *photometric inversion*, denoted here by i: $i(\omega) = L - \omega$, $\omega \in \tilde{\Omega}_n^L$. Finally, the group G generated by all the above symmetries has presentation $\langle r, s, i | r^4 = s^2 = i^2 = 1, si = is, ri = ir, rs = sr^3 \rangle$. Therefore, $G \cong D_8 \times C_2$, where $C_2 \cong \mathbb{Z}_2 \cong \langle i \rangle$ is the *cyclic* group of order two.

In order to simplify computations (including establishing a group isomorphism between G and the corresponding subgroup of $GL(n^2, \mathbb{R})$), we standardize intensity ranges $\mathcal{C}_L$: $\{\frac{1-L}{2L}, \frac{3-L}{2L}, \ldots, \frac{L-1}{2L}\}$, embedding them in $[-0.5, 0.5]$ via $c \mapsto \frac{2c-(L-1)}{2L}$, $c \in \mathcal{C}_L$. The corresponding state spaces are consequently embedded in $\Omega_n \overset{\text{def}}{=} [-0.5, 0.5]^{n^2}$ in the same manner ($\omega \mapsto \frac{2\omega-(L-1)}{2L}$), and will be written as Ω_n^L. Thus, by partitioning (quantizing) Ω_n uniformly as below

$$\left(\left(\frac{-L}{2L} + \frac{1+2\cdot 0}{2L}, \frac{-L}{2L} + \frac{1+2\cdot 1}{2L}\right] \cup \cdots \cup \left(\frac{-L}{2L} + \frac{1+2\cdot(L-1)}{2L}, \frac{-L}{2L} + \frac{1+2\cdot L}{2L}\right] \right)^{n^2}$$

one can think of $\omega = (\omega_{1,1}, \ldots, \omega_{n,n}) \in \tilde{\Omega}_n^L$ as the central point of $(\omega_{1,1} - \frac{1}{2L}, \omega_{1,1} + \frac{1}{2L}] \times \cdots \times (\omega_{n,n} - \frac{1}{2L}, \omega_{n,n} + \frac{1}{2L}]$, the corresponding n^2-dimensional partition cell.

We now assume $n = 2$. With the standard basis for $\mathbb{R}^4$, the matrix version of G is generated by

$$r \overset{\rho}{\mapsto} \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \quad s \overset{\rho}{\mapsto} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \quad i \overset{\rho}{\mapsto} \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \tag{31}$$

As explained in §6, knowing $\mathcal{S}_\Omega$ is important for understanding the complexity of $\mathcal{P}^G$, for obtaining the Reynolds operator $\mathcal{R}$ in its matrix form §6.2, and for efficient computation of the invariant models §6.3.

Proposition 38 *Let L be even. Then $|\mathcal{S}_{\Omega_2^L}| = \frac{L^4+2L^3+6L^2+4L}{16}$. There are L orbits of size two, $\frac{L^2}{4}$ orbits of size four, $\frac{2L^3+3L^2-10L}{8}$ orbits of size eight, and $\frac{L^4-2L^3-4L^2+8L}{16}$ orbits of size* 16.

This proposition and its proof (§D) suggest the following asymptotic result for any finite subgroup $G \leq GL(n^2, \mathbb{R})$ acting on Ω_n^L for any n and L: The leading term of $|\mathcal{S}_{\Omega_n^L}|$ is $\frac{|\Omega_n^L|}{|G|}$, i.e., $\frac{|\mathcal{S}_L||G|}{|\Omega_n^L|} \to 1$ as $L \to \infty$. In particular, not surprisingly the complexity of the corresponding models $\mathcal{P}^G$ grows as L^{n^2} $(= |\Omega_n^L|)$. However, one needs to recall the technical issues of computing invariant distributions (30) in order to appreciate this reduction of model dimensionality. Thus, for example, $L = 16$ and $n = 2$ give $|\Omega| = 65536$ and $|\mathcal{S}_\Omega| = 4708$, almost 14-fold reduction that is surely appreciated by any computational method of parameter estimation.

7.2 A minimal set of generators of $\mathbb{R}[\mathbb{R}^4]^G$.

Before we propose a particular set of invariant generators for $\mathbb{R}[x_1, x_2, x_3, x_4]^G$, let us recall that, according to (31) and (37), the G action on $\mathbb{R}[x_1, x_2, x_3, x_4]$ can be concisely expressed via the action of r, s, i, generators of G, on x_1, x_2, x_4, x_4, canonical generators of $\mathbb{R}[x]$:

$$\begin{aligned} rx_1 = x_2; \quad & rx_2 = x_3; \quad rx_3 = x_4; \quad rx_4 = x_1; \\ sx_1 = x_1; \quad & sx_2 = x_4; \quad sx_3 = x_3; \quad sx_4 = x_2; \\ ix_k = -x_k, \ & k = 1, 2, 3, 4 \end{aligned} \tag{32}$$

Theorem 39 *The following set of polynomials is a minimal set of generators of* $\mathbb{R}[x_1, x_2, x_3, x_4]^G$*:*

$$\begin{aligned} f_1(x) &= (x_1 + x_3)(x_2 + x_4), \\ f_2(x) &= x_1 x_3 + x_2 x_4, \\ f_3(x) &= x_1^2 + x_2^2 + x_3^2 + x_4^2, \\ f_4(x) &= x_1 x_2 x_3 x_4, \\ f_5(x) &= (x_1^2 + x_3^2)(x_2^2 + x_4^2). \end{aligned} \tag{33}$$

Also,

$$\mathbb{R}[x_1, x_2, x_3, x_4]^G \overset{(f_1,\ldots,f_5)}{\cong} \mathbb{R}[w_1, w_2, w_3, w_4, w_5]/J_F, \ \textit{where} \tag{34}$$

$$\begin{gathered} J_F = \{h \in \mathbb{R}[w_1, w_2, w_3, w_4, w_5] : h(f_1, f_2, f_3, f_4, f_5) = 0 \in \mathbb{R}[x_1, x_2, x_3, x_4]\} = \\ \langle q \rangle, \ \textit{and} \ q(w_1, w_2, w_3, w_4, w_5) = 4w_1^2 w_3 + 8w_1 w_2 w_5 + 2w_1 w_3 w_5 - 2w_1 w_4^2 w_5 + \\ 16w_2^2 - 8w_2 w_3 - 8w_2 w_4^2 + 4w_2 w_5^2 + w_3^2 - 2w_3 w_4^2 + w_4^4 \qquad . \end{gathered}$$

A proof of the theorem is given in §E. We base our proof on a very intuitive approach, which, in particular, does not require familiarity with algebraic geometry or invariant theory (§6.1). One classical upper bound due to Noether gives $m \leq N \leq \binom{m+|G|}{|G|}$. In our case the above upper bound is $\binom{4+16}{16} = 4845$. This is too large for a direct implementation of the corresponding algorithm to find such generators. Our case turns out to be special, however, in that we nearly achieve the lower bound determined by $\dim \mathbb{R}^4 = 4$. This small number of generators encourages one to use them in practice for orbit-indexing (§6.2).

8 Conclusion

Information theory, moments of measures on Euclidean spaces, polynomial invariants of finite groups, statistical model selection, and computational algebraic geometry, are the major faces of the polyhedral conglomerate of tools with which we articulate the concept of invariance, or symmetry. We have touched on each face just about as much as necessary to illuminate the central theme, that is *representing invariance under a finite group of symmetry transformations of the Euclidean space*. Indeed, we view establishment of connections among the relevant mathematical areas as the main contribution of this work. The types of symmetries that had motivated this work originate from our studies of statistics of small subimages of natural images. Hence, this work also contributes to the interdisciplinary efforts to analyze and to model vagaries of the natural image microworld.

To the probability and measure theory, this work offers a novel notion of determinacy within classes $\mathcal{M}_*^G$ of invariant measures indexed by the acting group G. This extends the ordinary notion of determinacy which formally corresponds to the action of the trivial group of the identity transformation. Specifically, we present a set of sufficient conditions, including a generalized Extended Carleman Theorem and some integral criteria, for determinacy of invariant measures by their invariant moments. The generalized notion is based on the algebra of invariant polynomials and a one-to-one correspondence between invariant measures on $\mathbb{R}^m$ and measures on $\mathbb{R}^N$. This correspondence is induced by a multinomial map $f = (f_1, \ldots, f_N)$, where $\{f_1, \ldots, f_N\}$ is any minimal set of generators of the ring of invariant polynomials. Thus, given this "change of variables", monomial terms in $f_1, \ldots, f_N$ replace ordinary moments. One important special example of this correspondence is that of measures supported in C, the positive cone of $\mathbb{R}^m$, where the acting group $G \cong \mathbb{Z}_2^m$ is generated by the sign inversions of all the coordinates, and a natural minimal set of generators has exactly $N = m$ elements, which makes it special. This case is well-known, at least for $m = 1$, and is unique in the following sense of "super-symmetry": $\mathbb{R}^m = \cup_{g \in G} gC$, $\dim(g_1 C \cap g_2 C) < m$ for all $g_1,\ g_2 \in G$.

In §7, we provide a less obvious example of this correspondence. This example is motivated by, and particularly suitable for models arising in natural image statistics, and is thus relevant for applied statistics. We present this example in great detail to show to applied statisticians that working with finite symmetries is possible within the basic algebraic theory, which also is becoming increasingly more accessible to nonspecialists through symbolic algebra software. Thus, our work also contributes to the field of algebraic statistics.

Determinate distributions can be approximated arbitrarily closely by matching in the limit all of their moments. We have shown that this combines perfectly with the notion of generalized determinacy via, for one example, the maximum entropy approach: First, given a sequence of invariant measures with all their mixed moments converging to corresponding mixed moments of P, an invariant measure determinate by its invariant moments, we obtain weak convergence of the sequence to P. Second, we construct special approximating sequences from the maximum entropy distributions that match subsets of the invariant moments of P. We therein make use of a key fact that, satisfying invariant constraints, the maximum entropy distribution inherits the underlying invariance. Requiring the above subsets of moments to cover in the limit all the moments, we again obtain convergence.

In the second part of this work, we specialize the above theory to modeling invariant distributions on finite state spaces. Instead of convergence in sequential approximation of P by increasingly refining invariant models, we address optimality of such approximations and efficiency of the involved computations. The former is closely related to statistical model selection where one balances model complexity and fit, or similarly, the amount of detail to be encoded in the model from one experiment. We propose a *framework for efficient modeling of invariant distributions* that combines well with many model selection principles and we give two examples. At the core of our framework is the fact that monomials $f_1^{\alpha_1}, \ldots, f_N^{\alpha_N}$ evaluated on the finite state space Ω span the linear space of the invariant functions on Ω. We give a family of algorithms to compute bases of nested subspaces of invariant functions on Ω.

A particular model selection principle, such as Minimal Description Length (MDL), may be applied naturally in this situation to flag the termination of the model construction at a minimum of an appropriate cost function C (description length in the case of MDL). Besides MDL, we give another example based on $D(P \| P^G_{A_l})$, the Kullback-Leibler divergence from P (or its symmetrized version $\mathcal{R}^*P$) to the invariant model $P^G_{A_l}$. That example employs the ad hoc penalized maximum likelihood criterion to terminate the model construction when the model $P^G_{A_l}$ approaches the best (in the maximum likelihood sense) invariant model $\mathcal{R}^*P$ at "distance" $D(\mathcal{R}^*(P) \| P^G_{A_l}$ comparable to that from P to $\mathcal{R}^*P$, e.g.: $D(\mathcal{R}^*(P) \| P^G_{A_l}) \leq D(P \| \mathcal{R}^*(P))$.

In fact, our main algorithms ("adaptive minimax learning") optimize this sequential model construction based on C or D by a "look ahead", or, "adaptive" model augmentation: Among a feasible set of directions outside the span of the current model, we choose one with the largest decrease in the cost function. Deriving performance bounds for these algorithms presents a direction for future work.

In summary, the proposed combination of our modeling framework and a model selection principle is essentially a technically special way to apply the model selection principle to the family of invariant distributions. One can then also "test the hypothesis" that P in fact possesses the given type of invariance: Carry out the same model selection including all the moments, and then, again using the same selection criterion, decide between the best invariant and "general" models. Thus, in selections based on cost function C, the invariance claim would be asserted if the minimum of C on the invariant family is lower than that obtained with general moments, and in the case of using D as above - if the dimension of the parameter space (i.e. the number of the monomials) of the best invariant model is smaller than that of the best model with general moments. Carrying out the outlined "testing", or "super" model selection experiments for our example of the natural microimage distribution (§7) shall be a natural continuation of this work.

From the statistical inference viewpoint, one would like to make an inference about P based on a sample distribution $\hat{P}$. Ideally, the population behind P is defined clearly and the sample is a simple random one. However, situations are common where, as in our microimage example, P is not obvious to define or its relation to the sample distribution $\hat{P}$ is difficult to establish. It is then also in response to such situations that we propose to use our "super" model selection principle to judge, however loosely, whether P is invariant. We finally note that the same methodology of "super" selection extends beyond invariant families by allowing arbitrary (as opposed to generating invariant polynomials) functions f.

This work at last discusses a number of computational issues related to invariant models. All these issues are rather basic, at least for specialists in the respective areas. However, the intuitively obvious result on dimensionality reduction in constrained entropy maximization with invariant, or "piecewise constant", constraints is, to our knowledge, presented here in full generality (Theorem 36) for the first time. Its finite version (Corollary 37), that is more important in practice, has already been presented in [24]. This observation may also be quite evident to statisticians preferring the equivalent "exponential family+likelihood maximization" viewpoint to the constrained entropy maximization one, chosen in this work.

We hope that this work provides a relatively self contained treatment of invariance under the action of a finite group of nonsingular transformations from the perspective of probability theory and applied statistics.

A Algebraic Supplements

This section presents proofs and remarks on the notions from §2.
Proposition 4 *The following actions are well-defined.*

1.) The (restricted) action of G on an invariant $\Omega \subset W$.

2.) The G action on $\mathcal{B}$, the Borel σ-algebra on Ω:

$$gB = \{g\omega : \ \omega \in B\}. \tag{35}$$

3.) The G action on $\mathcal{M}$, the set of (positive) measures on $\mathcal{B}$:

$$(gP)(B) = P(g^{-1}B), \quad B \in \mathcal{B}, \quad P \in \mathcal{M}. \tag{36}$$

4.) The G action on $\mathbb{R}[W]$, the set of real polynomials in m indeterminates:

$$(gf)(v) = f(g^{-1}v), \quad \text{where} \quad g \in G \quad \text{and } f \in \mathbb{R}[W] \text{ and } v \in W. \tag{37}$$

Proof.

1.) Straightforward verification.

2.) Clearly, $\forall B \in \mathcal{B}$ and $\forall g \in G\ gB \in \mathcal{B}$ (any g maps an open ball in Ω to an open set in Ω), and $(g_1 g_2)B = g_1 g_2 B$ immediately follows from its pointwise counterpart.

3.) Let g and P be arbitrary elements of G and $\mathcal{M}$, respectively. Clearly, $\forall B \in \mathcal{B}\ g^{-1}B \in \mathcal{B}$, hence gP is defined on the entire $\mathcal{B}$. It is also obvious that $gP(\emptyset) = P(g^{-1}\emptyset) = P(g\emptyset) = 0$. Note that this action is also preserved if $\mathcal{M}$ is restricted to the set of probability measures, since in that case $0 \leq gP(B) \leq 1$ and $gP(\Omega) = P(g^{-1}\Omega) = P(\Omega) = 1$ hold (all transformations $g \in G$ map Ω onto itself). Finally, for any collection $\{B_n\}_{n=1}^{\infty}$ of disjoint Borel sets, the Borel sets $\{g^{-1}B_n\}_{n=1}^{\infty}$ are clearly also disjoint (all transformations $g \in G$ are one-to-one), and thus: $gP(\cup_{n=1}^{\infty} B_n) = P(g^{-1} \cup_{n=1}^{\infty} B_n) = P(\cup_{n=1}^{\infty} g^{-1}B_n) = \sum_{n=1}^{\infty} P(g^{-1}B_n) = \sum_{n=1}^{\infty} gP(B_n)$.

4.) Straightforward verification.

$\diamond$

Proposition 9

$$\mathcal{M}^* = \{P \in \mathcal{M} : \mathbb{E}_P\|X\|^d < \infty\ \forall d \geq 0\}$$

Proof. Let $P \in \mathcal{M}^*$, and let $d \geq 0$ be arbitrary. Then, $\mathbb{E}_P\|X\|^d < P(B(0,1)) + \mathbb{E}_P\|X\|^D$, where $B(0,1)$ is the unit ball, D is even and $D > d$. The first term is finite as $\alpha = \mathbf{0}$ is included in the definition of $\mathcal{M}^*$ and the second term breaks down into a finite sum of "even" mixed moments, each of which is again finite by the definition of $\mathcal{M}^*$. To see the reverse inclusion, assume $\mathbb{E}_P\|X\|^d < \infty\ \forall d \geq 0$ and let $\alpha \in \mathbb{N}^N$ be arbitrary. Then, $\mathbb{E}_P|X|^\alpha \leq P(B(0,1)) + \mathbb{E}_P|X|^{2\alpha} \leq 1 + \mathbb{E}_P\|X\|^d < \infty$, where $d = 2\sum_{i=1}^m \alpha_i$. $\diamond$

More on Reynolds operator defined in (6)**.**

In polynomial algebra, this "averaging" map is called the *Reynolds Operator*. The orbit-averaging feature of this operator is apparent from its definition and the following property further underlines the correspondence with probabilistic averaging: $\forall f \in \mathbb{R}^\Omega$ and $\forall h \in (\mathbb{R}^\Omega)^G$, $\mathcal{R}(hf) = h\mathcal{R}(f)$. The probabilistic interpretation is that a random variable which is measurable relative to the σ-algebra on which conditioning is performed can almost surely be factorized through the conditional expectation.

Proposition 10

Consider $\mathcal{R}$ mapping the space of measurable functions on W onto $(\mathbb{R}^W)^G$ and the linear functionals $f \mapsto \int_W f(x)dP(x)$ defined by $P \in \mathcal{M}$. Then $\mathcal{R}$ and $\mathcal{R}^*$ are adjoint.

Proof. First show that for simple functions ϕ, $\int_W \mathcal{R}(\phi(x))dP(x)$ is indeed equal to $\int_W \phi(x)d(\mathcal{R}^*(P))(x)$ and then use the definition of the Lebesgue integral to extend this equality to all the measurable functions. $\diamond$

Proposition 11

1.) Let $P \in \mathcal{M}$ have a density p relative to some reference measure μ. Then $\mathcal{R}(p)$ is a density of $\mathcal{R}^*(P)$ relative to μ.

2.) Let p be a density of a G-invariant measure P relative to μ, then p is μ-a.e. G-invariant.

Proof. The second statement follows immediately from the first one. To prove the first, let $B \in \mathcal{B}$ be arbitrary and note

$$\begin{aligned} \mathcal{R}^*P(B) = & \quad \tfrac{1}{|G|}\sum_{g\in G} P(gB) & = \frac{1}{|G|}\sum_{g\in G}\int_{gB} p(x)\mu(dx) \\ = & \quad \tfrac{1}{|G|}\sum_{g\in G}\int_B p(gy)|\det(g)|\mu(dy) & = \int_B \mathcal{R}p(y)\mu(dy). \end{aligned}$$

$|\det(g)| = 1$ follows from the finiteness of $G \subset GL(m, \mathbb{R})$. $\diamond$

Remark 40 *Despite being finite, minimal generating sets need not in general have the same cardinality unless one explicitly requires the minimality of their cardinality.*

Proposition 15 Let $f_1, \ldots, f_N$ generate $\mathbb{R}[W]^G$ and let $f = (f_1, \ldots, f_N) : W \to \mathbb{R}^N$.Then the map $\bar{f} : \mathcal{S}_W \to \mathbb{R}^N$ mapping $[w]$, the equivalence class of $w \in W$, to $f(w)$, is well-defined and injective. Thus $\mathcal{S}_W \cong f(W)$, the image of f in $\mathbb{R}^N$.
Proof. The G-invariance of $f_1, \ldots, f_N$ means constancy of f on the orbits of $\mathcal{S}_W$. Thus $[w] \overset{\bar{f}}{\mapsto} f(w)$ is indeed well-defined as a map from $\mathcal{S}_W$ onto $f(W)$. Therefore, we need only prove that, given any two distinct orbits $\mathcal{O}_1, \mathcal{O}_2 \in \mathcal{S}_W$, $\bar{f}(\mathcal{O}_1) \neq \bar{f}(\mathcal{O}_2)$. We show this by exhibiting a G-invariant polynomial h that takes distinct values on $\mathcal{O}_1$ and $\mathcal{O}_2$, and then conclude that the values assumed by at least one of the N generators on these orbits must be distinct since h can be expressed (as a polynomial) in terms of the given generators.

The finite size of the orbits allows the following crude construction of h:

$$\tilde{h}_{\mathcal{O}_1}(x) = \prod_{g\in G}\sum_{l=1}^{m}\left[x_l - (g\omega)_l\right]^2, \qquad \omega \in \mathcal{O}_1 \tag{38}$$

$$h_{\mathcal{O}_1}(x) = \mathcal{R}(\tilde{h})(x). \tag{39}$$

The definition (38) ensures that $\tilde{h}_{\mathcal{O}_1}(v) = 0$ (and consequently $h(v) = 0$) if and only if $v \in \mathcal{O}_1$. In (39), we average $\tilde{h}_{\mathcal{O}_1}$ over all the G-orbits in order to guarantee G-invariance. Note that $h_{\mathcal{O}_1}$ separates $\mathcal{O}_1$ from the rest of the orbits, since for each $g \in G$ the only roots of $g\tilde{h}_{\mathcal{O}_1}$ are the points in $\mathcal{O}_1$. In particular, $h_{\mathcal{O}_1}$ assumes distinct values on $\mathcal{O}_1$ and $\mathcal{O}_2$. $\diamond$

B Invariant measures, moments, and determinacy

Proposition 17 Let $f_1, \ldots, f_N$ be a minimal generating set. Then $\mathcal{M}_*^G = \{P \in \mathcal{M}^G : \ \mathbb{E}_P|f^\alpha| < \infty \ \forall \alpha \in \mathbb{N}^N\}$.
Proof. The inclusion of $\mathcal{M}_*^G \subset$ into the right hand side is obvious. To show the other inclusion, we take $\alpha^* \in \mathbb{N}^N$ arbitrary and $P \in$ RHS and otherwise arbitrary. Let Σ_k be the set of all k-subsets of $\{1, \ldots, m\}$, and notice:

$$\begin{aligned}
\mathbb{E}_P|X^{\alpha^*}| &= \sum_{\substack{0 \le k \le m \\ \sigma \in \Sigma_k}} \int_{\substack{|x_j| \ge 1 \ \forall j \in \sigma \\ |x_j| < 1 \ \forall j \notin \sigma}} |x^{\alpha^*}| dP \\
&\le \sum_{\substack{0 \le k \le m \\ \sigma \in \Sigma_k}} \int_{\substack{|x_j| \ge 1 \ \forall j \in \sigma \\ |x_j| < 1 \ \forall j \notin \sigma}} \prod_{i \in \sigma} x_i^{2\alpha_i^*} dP \\
&\le \sum_{\substack{0 \le k \le m \\ \sigma \in \Sigma_k}} \int_{\mathbb{R}^m} \prod_{i \in \sigma} x_i^{2\alpha_i^*} dP \\
&= \sum_{\substack{0 \le k \le m \\ \sigma \in \Sigma_k}} \int_{\mathbb{R}^m} \prod_{i \in \sigma} x_i^{2\alpha_i^*} d\mathcal{R}^* P \\
&= \sum_{\substack{0 \le k \le m \\ \sigma \in \Sigma_k}} \int_{\mathbb{R}^m} \mathcal{R}(\prod_{i \in \sigma} x_i^{2\alpha_i^*}) dP < \infty.
\end{aligned}$$

In the above we used the fact $\mathcal{R}^*$ and $\mathcal{R}$ are adjoint (Proposition 10). The last inequality follows from that $\mathcal{R}(\prod_{i\in\sigma} x^{2\alpha^*})$ is G-invariant and hence is a polynomial in f-generators: $\sum_\alpha a_\alpha f^\alpha$, but $\mathbb{E}_P f^\alpha \le \mathbb{E}_P|f^\alpha| < \infty$ for all $\alpha \in \mathbb{N}^N$. $\diamond$

Definition 18 Let $P \in \mathcal{M}_*^G$ have $s(P)$, its G-invariant moments, relative to some minimal generating set. Then P is said to be G-determinate by $s(P)$, or simply G-determinate, if no other measure in $\mathcal{M}_*^G$ has the same set of moments $s(P)$ relative to the chosen generating set.
Let us prove that this notion is well-defined:
Proof. Let $f_1, \ldots, f_N$ and $h_1, \ldots, h_L$ be two distinct minimal sets of generators, and let $s_f(P)$ and $s_h(P)$ be the corresponding sets of G-invariant moments. Suppose that P is the only measure in $\mathcal{M}_*^G$ possessing $s_f(P)$, and suppose that there exists $Q \in \mathcal{M}_*^G$ such that $Q \ne P$ and $s_h(P) = s_h(Q)$. Then there must exist $\alpha \in \mathbb{N}^N$ such that $\mathbb{E}_P f^\alpha \ne \mathbb{E}_Q f^\alpha$. Since f^α is G-invariant, it can be written as a polynomial in h-generators: $\sum_\beta a_\beta h^\beta$, but then for each monomial we have $\mathbb{E}_P h^\beta = \mathbb{E}_Q h^\beta$. This clearly contradicts $\mathbb{E}_P f^\alpha \ne \mathbb{E}_Q f^\alpha$. $\diamond$

Lemma 20 The map $\tilde{f} : \mathcal{M}^G \to \tilde{\mathcal{M}}$ via $\tilde{f}(P)(B) = P(f^{-1}(B))$ for any $B \in \mathcal{B}(\mathbb{R}^N)$, is one-to-one.

Proof. Let $P, Q \in \mathcal{M}_*^G$ be distinct, and let $B \in \mathcal{B}(\Omega)$ be such that $P(B) > Q(B)$. Now, define $h(x) = \mathcal{R}(\mathbb{I}_B(x))$, the G-symmetrized indicator function of B. Next note that $P(B) = \mathbb{E}_P \mathbb{I}_B(X) = \mathbb{E}_P h(X)$, where the random vector X is distributed according to P, and the second equality is a consequence of G-invariance of P. Also note that similarly, $Q(B) = \mathbb{E}_Q h(X)$, and therefore $\mathbb{E}_P h(X) > \mathbb{E}_Q h(X)$.

Observe that the level sets $h^{-1}(x \geq c)$ for any $c \in \mathbb{R}$ are also G-invariant:

$$\begin{aligned} gh^{-1}(x \geq c) = \quad & \{gw : \; w \in W \; h(w) \geq c\} = \quad & \{w' : \; g^{-1}w' \in W h(g^{-1}w') \geq c\} = \\ = \quad & \{w' : \; g^{-1}w' \in W gh(w') \geq c\} = \quad & \{w' : \; g^{-1}w' \in W h(w') \geq c\} = \\ = \quad & \{w' : \; w' \in W h(w') \geq c\} = \quad & h^{-1}(x \geq c) \end{aligned}$$

Now, $\mathbb{E}_P h(X) = \sum_{c \in \{h(w): \; w \in W\}} P(h(X) \geq c)$, where the summation has a finite number of terms due to the special form of h. Hence, there must be at least one term such that $P(h(X) \geq c) > Q(h(X) \geq c)$, which gives us a G-invariant set $A = h^{-1}(x \geq c)$ (that is obviously also Borel) on which P and Q differ.

It now remains to prove that $\tilde{f}(P) \neq \tilde{f}(Q)$. To this end we show that

$$\begin{aligned} \tilde{f}(P)(fA) &= P(f^{-1}fA) \\ &= P(\bar{f}^{-1}\bar{f} \dot{\cup}_{\mathcal{O} \subset A} \mathcal{O}) && (40) \\ &= P(\dot{\cup}_{\mathcal{O} \subset A} \bar{f}^{-1}\bar{f}(\mathcal{O})) \\ &= P(\dot{\cup}_{\mathcal{O} \subset A} \mathcal{O}) && (41) \\ &= P(A) && (42) \end{aligned}$$

Proposition 5 gives $A = \dot{\cup}_{\mathcal{O} \subset A} \mathcal{O}$ used in (40) and (42), and Proposition 15 implies (41).

Summarizing the above, we get $\tilde{f}(P)(fA) > \tilde{f}(Q)(fA)$, finishing the proof of the Lemma. $\diamond$

C Some results from [8]

Theorem 2.1 of [8] on multidimensional quasi-analytic classes.
For $j = 1, \ldots, n$ let $\{M_j(m)\}_{m=0}^{\infty}$ be a sequence of non-negative real numbers such that

$$\sum_{m=1}^{\infty} \frac{1}{M_j(m)^{1/m}} = \infty.$$

Assume that $f : \mathbb{R}^n \mapsto \mathbb{C}$ is of class C^∞ and that there exists $C \geq 0$ such that

$$\left| \frac{\partial^\alpha f}{\partial \lambda^\alpha}(\lambda) \right| \leq C \prod_{j=1}^{n} M_j(\alpha_j)$$

for all $\alpha \in \mathbb{N}^n$ and all $\lambda \in \mathbb{R}^n$. Then, if $|\frac{\partial^\alpha f}{\partial \lambda^\alpha}(0)| = 0$ for all $\alpha \in \mathbb{N}^n$, f is actually identically zero on $\mathbb{R}^n$.

Theorem 2.3 of [8]: Extended Carleman Theorem.

Let $\mu \in \mathcal{M}^*$ and suppose $\{v_1, \ldots, v_n\}$ is a basis of $\mathbb{R}^n$. For $j = 1, \ldots, n$ and $m = 0, 1, 2, \ldots$ define

$$s_j(m) = \int_{\mathbb{R}^n} (v_j, x)^m d\mu(x).$$

If each of the sequences $\{s_j(m)\}_{m=1}^\infty$ $(j = 1, \ldots, n)$ satisfies Carleman's condition

$$\sum_{m=1}^\infty \frac{1}{s_j(2m)^{1/2m}} = \infty,$$

then μ is determinate. Furthermore, the polynomials and $\mathrm{Span}_\mathbb{C}\{\exp i(\lambda, x) | \lambda \in S\}$ are dense in $L_p^G(\mathbb{R}^n, \mu)$ for all $1 \leq p < \infty$ and for every $S \in \mathbb{R}^n$ which is somewhere dense.

Theorem 4.1 of [8].

Let μ be a positive Borel measure on $\mathbb{R}^n$ such that

$$\int_{\mathbb{R}^n} w(f(x))^{-1} d\mu < \infty$$

for some measurable quasi-analytic weight. Then μ is determinate. Furthermore, $\mathbb{R}[W]^G$ and $\mathrm{Span}_\mathbb{C}\{\exp i(\lambda, x) | \lambda \in S\}$ are dense in $L_p^G(\mathbb{R}^n, \mu)$, for $1 \leq p < \infty$ and for every $S \in \mathbb{R}^n$ which is somewhere dense.

Theorem 4.2. of [8].

For $j = 1, \ldots, n$, let $R_j > 0$ and let a non-decreasing function $\rho_j : (R_j, \infty) \to \mathbb{R}^+$ of class C^1 be such that

$$\int_{R_j}^\infty \frac{\rho_j(s)}{s^2} ds = \infty.$$

Define $h_j : \mathbb{R} \to \mathbb{R}^+$ by

$$h_j(x) = \begin{cases} \exp\left(\int_{R_j}^{|x|} \frac{\rho_j(s)}{s} ds\right) & \text{for} \quad |x| > R_j \\ 1 & \text{for} \quad |x| \leq R_j. \end{cases}$$

Let A be an affine automorphism of $\mathbb{R}^n$. If P is a positive Borel measure on W such that

$$\int_{\mathbb{R}^n} \prod_{j=1}^N h_j((Ax)_j) dP(x) < \infty,$$

then P is determinate by its G-invariant moments. Furthermore, the polynomials and $\mathrm{Span}_\mathbb{C}\{\exp i(\lambda, x) | \lambda \in S\}$ are dense in $L_p(\mathbb{R}^n, P)$, for $1 \leq p < \infty$ and for every S

of $\mathbb{R}^n$ which is somewhere dense.

Proposition 25 Let P have a density p relative to λ. Then

$$0 \le H(P) \le H(\mathcal{R}^* P) \le H(P) + \log |G|.$$

The equality in place of the second inequality occurs if and only if P is G-invariant.

Proof. Convexity of $x \log x$ and Jensen's inequality establish positiveness of H. To see the second inequality, first recall that $D(P|Q) \ge 0$ with the strict equality if and only if $P = Q$ (use $\log x \le x - 1$ with the strict equality only at $x = 1$). Then notice that

$$0 \le D(P|\mathcal{R}^*(P)) = -H(P) + \mathbb{E}_P \log(1/\mathcal{R}(p(X))),$$

and by Proposition 10:

$$\mathbb{E}_P \log(1/\mathcal{R}(p)(X)) = \mathbb{E}_{\mathcal{R}^*(P)} \log(1/\mathcal{R}(p)(X)) = H(\mathcal{R}^*(P)).$$

Finally, noticing that $|\mathcal{O}| \le |G|, \forall \mathcal{O} \in \mathcal{S}_W$, gives:

$$D(P|\mathcal{R}^*(P)) \le \int_W p(x) \log \frac{\max_{y \in [x]} p(y)}{\max_{y \in [x]} p(y)/|[x]|} d\mu(x) = \int_W p(x) \log |[x]| d\mu(x) \le \log |G|.$$

Summarizing the above: $H(\mathcal{R}^*(P)) = H(P) + D(P|\mathcal{R}^*(P)) \le H(P) + \log |G|$. $\diamond$

Remark 29 continued. In order to see more directly that minimizing $D(P|P_{A_{l-1} \cup \{\alpha\}})$ is equivalent to minimizing $D(\mathcal{R}^*(P)|P_{A_{l-1} \cup \{\alpha\}})$ note that the minimization takes place only within the term $-\mathbb{E}_P \log(p')$, where p' is a G-invariant density of $P_{A_{l-1} \cup \{\alpha\}}$ (Proposition 11). Recalling (Proposition 10) that the operators $\mathcal{R}$ and $\mathcal{R}^*$ are adjoint and Proposition 11, establishes $\mathbb{E}_P \log(p') = \mathbb{E}_P \mathcal{R}(\log(p')) = \mathbb{E}_{\mathcal{R}^*(P)} \log(p')$.

D The structure of $\mathcal{S}_{\Omega_L^2}$

Proposition 38 Let L be even. Then $|\mathcal{S}_{\Omega_L^2}| = \frac{L^4+2L^3+6L^2+4L}{16}$. There are L orbits of size two, $\frac{L^2}{4}$ orbits of size four, $\frac{2L^3+3L^2-10L}{8}$ orbits of size eight, and $\frac{L^4-2L^3-4L^2+8L}{16}$ orbits of size 16.

Proof. The $n = 1$ case is special but trivial. There are two orbits of size two:

$$\left\{ \begin{matrix} -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} \end{matrix}, \begin{matrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{matrix} \right\}, \left\{ \begin{matrix} -\frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{matrix}, \begin{matrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{matrix} \right\},$$

one orbit of size four:

$$\left\{ \begin{matrix} -\frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{matrix}, \begin{matrix} \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{matrix}, \begin{matrix} \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} \end{matrix}, \begin{matrix} -\frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{matrix} \right\},$$

and one orbit of size eight:

$$\left\{ \begin{matrix} \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{matrix}, \begin{matrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{matrix}, \begin{matrix} \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{matrix}, \begin{matrix} -\frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{matrix}, \begin{matrix} -\frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} \end{matrix}, \begin{matrix} -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{matrix}, \begin{matrix} -\frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{matrix}, \begin{matrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} \end{matrix} \right\}$$

To prove the general case, one first recalls that $\forall \mathcal{O}, \forall \omega \in \mathcal{O}$, $|\mathcal{O}| = |G : G_\omega|$, the size of the orbit $\mathcal{O}$ equals the index of the stabilizer G_ω.

Since $|G| = 16$, $|\mathcal{O}|$ can only be $1, 2, 4, 8, 16$. Clearly, there is no ω with $G_\omega = G$ because $i(\omega) = \omega$ has no solution. For the same reason G_ω can not contain i, si, or r^2si among its generators. This leaves only two copies of D_8 (i.e. $\langle r, s | r^4 = s^2 = 1, rs = sr^3\rangle$ and $\langle ri, s | (ri)^4 = s^2 = 1, (ri)s = s(ri)^3\rangle$) as possible stabilizers of index two. The first group gives rise to the two equations $r(\omega) = \omega$ and $s(\omega) = \omega$ with L solutions of the form $\left(\begin{smallmatrix} \lambda & \lambda \\ \lambda & \lambda \end{smallmatrix}\right), \lambda \in \mathcal{C}_L$, thus yielding $L/2$ orbits of size two. The second choice implies that $(ri)(\omega) = \omega$ and $s(\omega) = \omega$, resulting in the 2^n patches of the form $\left(\begin{smallmatrix} -\lambda & \lambda \\ \lambda & -\lambda \end{smallmatrix}\right), \lambda \in \mathcal{C}_L$ that are partitioned into $L/2$ size-two orbits. Hence, the total number of size-two orbits becomes L.

We now count orbits of size four. The following subgroups are the only subgroups of G of index four not containing i, si, or r^2si: $\langle r\rangle$, $\langle ri\rangle$, $\langle r^3i\rangle$, $\langle r^2, s\rangle$, $\langle r^2, rs\rangle$, $\langle r^2, rsi\rangle$, $\langle r^2i, rs\rangle$, $\langle r^2i, rsi\rangle$. Since all the ω's fixed by the rotation group are necessarily fixed by the entire $\langle r, s | r^4 = s^2 = 1, rs = sr^3\rangle$ group, the rotation group can not be a proper stabilizer itself. Similarly, $(ri)(\omega) = \omega \Rightarrow s(\omega) = \omega$ implies that $\langle ri\rangle$ is a proper subgroup of a larger stabilizer, and for the same reason $(r^3i)(\omega) = \omega \Rightarrow s(\omega) = \omega$ makes it impossible for $\langle r^3i\rangle$ to be a stabilizer. Now notice, $\langle r^2, rs\rangle$ can not be a proper stabilizer since $[(r^2)(\omega) = \omega] \wedge [(rs)(\omega) = \omega] \Rightarrow r(\omega) = \omega^2$; $\langle r^2, rsi\rangle$ can not be a proper stabilizer because $[(r^2)(\omega) = \omega] \wedge [(rsi)(\omega) = \omega] \Rightarrow (ri)(\omega) = \omega$. Finally, $\langle rs, r^3s\rangle$ fails to be a stabilizer since $[(rs)(\omega) = \omega] \wedge [r^2(\omega) = \omega] \Rightarrow r(\omega) = \omega$.

Next, $\langle r^2, s\rangle$ is a stabilizer for all elements of the form: $\left(\begin{smallmatrix} \lambda & \gamma \\ \gamma & \lambda \end{smallmatrix}\right)$, where $\gamma, \lambda \in \mathcal{C}_L, \gamma \neq \lambda$, $\gamma \neq -\lambda$. Since there are $L(L-2)$ such matrices, and the orbit of each of them consists of matrices of the same form (up to renaming of λ and γ), they must form exactly $L(L-2)/4$ size-four orbits.

Matrices of the form $\left(\begin{smallmatrix} -\lambda & -\lambda \\ \lambda & \lambda \end{smallmatrix}\right)$, with $\lambda \in \mathcal{C}_L$ are stabilized by $\langle r^2i, rs\rangle$. In fact, these will represent only $L/2$ distinct matrices as λ runs effectively only through half of the range $\mathcal{C}_L$. Since no two distinct such matrices fall into the same orbit, we obtain $L^2/4$ as the total number of size-four orbits. We also notice that the subgroup $\langle r^2i, rsi\rangle$ is a stabilizer for the elements of the form $\left(\begin{smallmatrix} \lambda & -\lambda \\ \lambda & -\lambda \end{smallmatrix}\right)$, which are rotationally equivalent to the previous matrices, hence adding no new orbits.

The last task is to compute the number of orbits of size eight. First, we list all the subgroups of index eight (thus, order two) not containing i, si, or r^2si. These are: $\langle r^2\rangle$, $\langle r^2\rangle$, $\langle r^2i\rangle$, $\langle s\rangle$, $\langle r^2s\rangle$, $\langle rs\rangle$, $\langle r^3s\rangle$, $\langle rsi\rangle$, and $\langle r^3si\rangle$. $\langle r^2\rangle$ immediately leaves the list since it is a proper subgroup of a larger stabilizer ($r^2(\omega) = \omega \Rightarrow s(\omega) = \omega$). Matrices of the form $\left(\begin{smallmatrix} \lambda & \delta \\ \gamma & \lambda \end{smallmatrix}\right)$, where $\delta, \gamma, \lambda \in \mathcal{C}_L$, $\gamma \neq \delta$, are stabilized by $\langle s\rangle$, whereas rotationally equivalent to them matrices of the form $\left(\begin{smallmatrix} \delta & \lambda \\ \lambda & \gamma \end{smallmatrix}\right)$ are stabilized by $\langle r^2s\rangle$. Since size-eight orbits generated by these $2L^2(L-1)$ matrices are composed of these matrices only, we arrive at $L^2(L-1)/4$ distinct orbits of size eight. Next, observe that $\langle rs\rangle$ fixes $L(L-2)$ matrices of the form $\left(\begin{smallmatrix} \gamma & \gamma \\ \lambda & \lambda \end{smallmatrix}\right)$, with $\gamma, \lambda \in \mathcal{C}_L$, $\gamma \neq \lambda$, $\gamma \neq -\lambda$,

[2] We use "$\wedge$" to denote the logical *and*.

whereas their $L(L-2)$ rotational equivalents $\left(\begin{smallmatrix} \lambda & \gamma \\ \lambda & \gamma \end{smallmatrix}\right)$ are fixed by $\langle r^3 s\rangle$. Since all the matrices inside the corresponding orbits of size eight are of either of the two forms, we add $L(L-2)/4$ orbits of size eight. The same number of $L(L-2)/4$ size-eight orbits come from $L(L-2)$ matrices of the form $\left(\begin{smallmatrix} -\lambda & -\gamma \\ \lambda & \gamma \end{smallmatrix}\right)$ fixed by $\langle r^3 si\rangle$, with $\gamma, \lambda \in \mathcal{C}_L$, $\gamma \neq \lambda$, $\gamma \neq -\lambda$, and from their $L(L-2)$ rotational equivalents of the form $\left(\begin{smallmatrix} \gamma & -\gamma \\ \lambda & -\lambda \end{smallmatrix}\right)$ fixed by $\langle rsi\rangle$. The last source of size-eight orbits is matrices stabilized by $\langle r^2 i\rangle$. They are represented by $\left(\begin{smallmatrix} -\gamma & -\lambda \\ \lambda & \gamma \end{smallmatrix}\right)$, where $\gamma \neq \lambda$, $\gamma \neq -\lambda$. There are exactly $L(L-2)$ such matrices, producing the last $L(L-2)/8$ orbits of size eight.

Summing over orbits of sizes less than 16, we get $2 \times L + 4 \times L^2/4 + 8 \times (L^3/4 + 3L^2/8 - 5L/4)$ as the total number of elements in these orbits. Hence, the number of orbits of size 16 is $(L^4 - 2L^3 - 4L^2 + 8L)/16 = n^4 - n^3 - n^2 + n$. Finally, the total number of orbits is $\frac{L^4+2L^3+6L^2+4L}{16} = n^4 + n^3 + \frac{n(3n+1)}{2}$. $\diamond$

E Generators for $\mathbb{R}[x]^G$

Theorem 39. The following set of polynomials is a minimal set of generators of $\mathbb{R}[x_1, x_2, x_3, x_4]^G$:

$$\begin{aligned} f_1(x) &= (x_1 + x_3)(x_2 + x_4), \\ f_2(x) &= x_1 x_3 + x_2 x_4, \\ f_3(x) &= x_1^2 + x_2^2 + x_3^2 + x_4^2, (33) \\ f_4(x) &= x_1 x_2 x_3 x_4, \\ f_5(x) &= (x_1^2 + x_3^2)(x_2^2 + x_4^2). \end{aligned}$$

Also,

$$\mathbb{R}[x_1, x_2, x_3, x_4]^G \overset{(f_1,\ldots,f_5)}{\cong} \mathbb{R}[w_1, w_2, w_3, w_4, w_5]/J_F, \text{ where} \tag{34}$$
$$J_F = \{h \in \mathbb{R}[w_1, w_2, w_3, w_4, w_5] : h(f_1, f_2, f_3, f_4, f_5) = 0 \in \mathbb{R}[x_1, x_2, x_3, x_4]\} =$$
$$\langle q\rangle, \text{ and } q(w_1, w_2, w_3, w_4, w_5) = 4w_1^2 w_3 + 8w_1 w_2 w_5 + 2w_1 w_3 w_5 - 2w_1 w_4^2 w_5 +$$
$$16w_2^2 - 8w_2 w_3 - 8w_2 w_4^2 + 4w_2 w_5^2 + w_3^2 - 2w_3 w_4^2 + w_4^4 \quad .$$

Proof. It is immediate to see that $f_1, \ldots, f_5$ respect the action of r, s, i, generators of G. Therefore, they $f_1, \ldots, f_5 \in \mathbb{R}[x_1, x_2, x_3, x_4]^G$. We base our computations on a sequence of decompositions of the original G action, first step of which is given by:

$$\mathcal{S}_{\mathbb{R}^4} \cong (\mathbb{R}^4/G_1) \,/\, (G/G_1)\,,$$
$$\text{where } G_1 = \langle s, r^2 | s^2 = (r^2)^2 = 1, r^2 s = s r^2\rangle \trianglelefteq G \tag{43}$$

The equation above simply says that the original action of G on $\mathbb{R}^4$ decomposes into two actions as follows: First, G_1, a *normal subgroup* of G, acts on $\mathbb{R}^4$, producing the orbit set $\mathbb{R}^4/G_1$, and then the *quotient group* G/G_1 acts on $\mathbb{R}^4/G_1$, resulting in "the same" orbits $\mathcal{S}_{\mathbb{R}^4}$, just as if G acted on $\mathbb{R}^4$ directly. Thus, we first aim to find $y_1(x), \ldots, y_k(x)$ for some k, generators for $\mathbb{R}[x]^{G_1}$, and then will focus on the polynomials (in those generators) that are invariant under G/G_1.

Claim 41 $\mathbb{R}[x]^{G_1} = \mathbb{R}[x_1 + x_3, x_2 + x_4, x_1x_3, x_2x_4]$.

Proof. It suffices to prove that $\mathbb{R}[x]^{\langle r^2 s\rangle} = \mathbb{R}[x_1 + x_3, x_2, x_1x_3, x_4]$ and $\mathbb{R}[x]^{\langle s\rangle} = \mathbb{R}[x_1, x_2 + x_4, x_3, x_2x_4]$, since $\mathbb{R}[x_1 + x_3, x_2 + x_4, x_1x_3, x_2x_4] = \mathbb{R}[x_1 + x_3, x_2, x_1x_3, x_4] \cap \mathbb{R}[x_1, x_2 + x_4, x_3, x_2x_4]$. In fact, we only prove the first of these statements since the second one proves along the same lines interchanging x_1 with x_2 and x_3 with x_4. We argue by induction on the *degree* function, $\deg = \deg_1 + \deg_2 + \deg_3 + \deg_4$, where $\deg_k$ is the highest power of x_k ($k = 1, 2, 3, 4$) in a given polynomial. Let us begin by noticing that the result holds for all polynomials of $\deg = 0$ (i.e. constants.) Assume now that the result is true for $\deg \leq N$, $N \geq 0$ and show that it also holds for $\deg = N + 1$. A generic polynomial $r(x_1, x_2, x_3, x_4) \in \mathbb{R}^{\langle r^2 s\rangle}[x]$ such that $\deg(r) \leq N + 1$ has the form:

$$\sum_{\substack{i,j,k,l\geq 0\\ i+j+k+l\leq N+1}} a_{i,j,k,l}x_1^i x_2^j x_3^k x_4^l = \overbrace{\sum_{\substack{i,k\geq 0\\ i+k\leq N}} a_{i,0,k,0}x_1^i x_3^k}^{1} + \overbrace{a_{N+1,0,0,0}x_1^{N+1} + a_{0,0,N+1,0}x_3^{N+1}}^{2} + \tag{44}$$

$$\overbrace{x_1x_3 \sum_{\substack{i,k>0\\ i+k=N+1}} a_{i,0,k,0}x_1^{i-1}x_3^{k-1}}^{3} + \sum_{\substack{j,l\geq 0\\ 0<j+l\leq N+1}} \left(\overbrace{\sum_{\substack{i,k\geq 0\\ 0\leq i+k\leq N+1-j-l}} a_{i,j,k,l}x_1^i x_3^k}^{4} \right) x_2^j x_4^l \tag{45}$$

In order for the left hand side to be invariant under $x_1 \leftrightarrow x_3$, each of the terms $1 - 4$ in (44)-45 must be invariant under the same action. By the induction argument, terms of degree N and below are already in the desired form. Thus, the first sum and all the sums labeled 4 belong to $\mathbb{R}[x_1 + x_3, x_2, x_1x_3, x_4]$. This implies that the entire double sum of (45) is in $\mathbb{R}[x_1 + x_3, x_2, x_1x_3, x_4]$. The cofactor of x_1x_3 in the third term of (45) is also invariant and has degree N, hence lies in $\mathbb{R}[x_1 + x_3, x_2, x_1x_3, x_4]$ as well. The invariance of the second term of (44) forces $a_{N+1,0,0,0} = a_{0,0,N+1,0}$. We now notice that if $N = 0$, then

$$a_{N+1,0,0,0}x_1^{N+1} + a_{0,0,N+1,0}x_3^{N+1} = a_{1,0,0,0}(x_1 + x_3) \in R[x_1 + x_3, x_2, x_1x_3, x_4]$$

For $N \geq 1$, on the other hand,

$$x_1^{N+1} + x_3^{N+1} = (x_1 + x_3)(x_1^N + x_3^N) - x_1x_3(x_1^{N-1} + x_3^{N-1}) \in R[x_1 + x_3, x_2, x_1x_3, x_4]$$

by the induction argument. This shows that the left hand side of (44),(45) belongs to $\mathbb{R}[x_1 + x_3, x_2, x_1x_3, x_4]$. $\diamond$

Thus, we have obtained a set of generators for $\mathbb{R}[x]^{G_1}$:

$$y_1 = x_1 + x_3,\ y_2 = x_2 + x_4,\ y_3 = x_3 x_4,\ y_4 = x_2 x_4, \tag{46}$$

which are algebraically independent. We now want to find $\mathbb{R}^{G/G_1}[y_1, y_2, y_3, y_4]$. Recall that

$$G/G_1 = \{1, \overline{r}, \overline{\imath}, \overline{\imath r}\}$$

and that its action on the orbit set $\mathbb{R}^4/G_1$ translates into

$$\overline{r} : y_1 \leftrightarrow y_2, \qquad y_3 \leftrightarrow y_4$$
$$\overline{\imath} : y_1 \mapsto -y_1,\ y_2 \mapsto -y_2, \qquad y_3 \leftrightarrow y_3,\ y_4 \leftrightarrow y_4$$

Continuing (43) to decompose the original G action, we write:

$$(\mathbb{R}^4/G_1)/(G/G_1) \cong \big((\mathbb{R}^4/G_1)/G_2\big)\Big/\big(G/G_1/G_2\big), \text{ where } G_2 = \langle \overline{\imath} \rangle \trianglelefteq G/G_1 \tag{47}$$

Claim 42 $\mathbb{R}[y_1, y_2, y_3, y_4]^{G_2} = \mathbb{R}[y_1^2, y_2^2, y_1 y_2, y_3, y_4]$

Proof. Using induction just as in the proof of Claim 41, we can simply imagine replacing x_1 with y_1, x_3 with y_2, x_2 with y_3, and x_4 with y_4, which yields equations essentially identical to (44),(45):

$$\sum_{\substack{i,j,k,l \geq 0 \\ i+j+k+l \leq N+1}} a_{i,j,k,l} y_1^i y_2^j y_3^k y_4^l = \sum_{\substack{i,j \geq 0 \\ i+j \leq N}} a_{i,j,0,0} y_1^i y_2^j + \overbrace{a_{N+1,0,0,0} y_1^{N+1} + a_{0,N+1,0,0} y_2^{N+1}}^{2} + \tag{48}$$
$$y_1 y_2 \sum_{\substack{i,j > 0 \\ i+j = N+1}} a_{i,j,0,0} y_1^{i-1} y_2^{j-1} + \sum_{\substack{k,l \geq 0 \\ 0 < k+l \leq N+1}} \left(\sum_{\substack{i,j \geq 0 \\ 0 \leq i+j \leq N+1-k-l}} a_{i,j,k,l} y_1^i y_2^j \right) y_3^k y_4^l$$

The only other difference from the previous proof is as follows: The new second term 48 disappears if $N+1$ is odd, whereas even $N+1$ immediately yields the needed form, i.e. $y_{1,2}^{N+1} = (y_{1,2}^2)^{(N+1)/2}$. ⋄

Next, notice:

$$\mathbb{R}[y_1^2, y_2^2, y_1 y_2, y_3, y_4] \cong \mathbb{R}[z_1, z_2, z_3, z_4, z_5]\big/\langle z_1 z_2 - z_5^2 \rangle,$$

under:

$$y_1^2 \to z_1,\ y_2^2 \to z_2, \quad y_3 \to z_3, \quad y_4 \to z_4,\ y_1 y_2 \to z_5.$$

We now show by induction that

$$\big(\mathbb{R}[z_1, z_2, z_3, z_4, z_5]\big/\langle z_1 z_2 - z_5^2 \rangle\big)^{(G/G_1)\big/G_2} \cong \mathbb{R}[z_1 + z_2, z_3 + z_4, z_3 z_4, z_1 z_3 + z_2 z_4, z_5], \tag{49}$$

where $(G/G_1)/G_2 = \langle\bar{\bar{r}}\rangle$, and its action results in exchanging z_1 with z_2 and z_3 with z_4. First, denote the right hand side of (49) by R and focus on the inductive transition from deg $\leq N$ to deg $= N+1$. A generic polynomial of interest splits into two sums, one with deg $\leq N$ and the other - with deg $= N+1$, each of which is separately invariant under the action of $\bar{\bar{r}}$. Since the first sum is in R by the induction assumption, we continue on to decompose the second one as follows:

$$\sum_{\substack{i,j,k,l\geq 0\\ i+j+k+l=N+1}} a_{i,j,k,l} z_1^i z_2^j z_3^k z_4^l = \overbrace{z_1 z_2 z_3 z_4 \sum_{\substack{i,j,k,l>0\\ i+j+k+l=N+1}} a_{i,j,k,l} z_1^{i-1} z_2^{j-1} z_3^{k-1} z_4^{l-1}}^{1} + \qquad (50)$$

$$\overbrace{z_1 z_2 \left(\sum_{\substack{i,j,k>0\\ i+j+k=N+1}} a_{i,j,k,0} z_1^{i-1} z_2^{j-1} z_3^k + \sum_{\substack{i,j,l>0\\ i+j+l=N+1}} a_{i,j,0,l} z_1^{i-1} z_2^{j-1} z_4^l \right)}^{2} +$$

$$\overbrace{z_3 z_4 \left(\sum_{\substack{i,k,l>0\\ i+k+l=N+1}} a_{i,0,k,l} z_1^i z_3^{k-1} z_4^{l-1} + \sum_{\substack{j,k,l>0\\ j+k+l=N+1}} a_{0,j,k,l} z_2^j z_3^{k-1} z_4^{l-1} \right)}^{3} +$$

$$\overbrace{z_1 z_2 \sum_{\substack{i,j>0\\ i+j=N+1}} a_{i,j,0,0} z_1^{i-1} z_2^{j-1}}^{4} + \overbrace{z_3 z_4 \sum_{\substack{k,l>0\\ k+l=N+1}} a_{0,0,k,l} z_3^{k-1} z_4^{l-1}}^{5} +$$

$$\overbrace{\sum_{\substack{i,k>0\\ i+k=N+1}} a_{i,0,k,0} z_1^i z_3^k + \sum_{\substack{j,l>0\\ j+l=N+1}} a_{0,j,0,l} z_2^j z_4^l}^{6} + \overbrace{\sum_{\substack{i,l>0\\ i+l=N+1}} a_{i,0,0,l} z_1^i z_4^l + \sum_{\substack{j,k>0\\ j+k=N+1}} a_{0,j,k,0} z_2^j z_3^k}^{7} +$$

$$\overbrace{a_{N+1,0,0,0} z_1^{N+1} + a_{0,N+1,0,0} z_2^{N+1}}^{8} + \overbrace{a_{0,0,N+1,0} z_3^{N+1} + a_{0,0,0,N+1} z_4^{N+1}}^{9}$$

An immediate inspection of (50) combined with the symmetry of the coefficients $a_{i,j,k,l} = a_{j,i,l,k}$ reveals that each of the terms numbered one through nine is individually invariant under the the given action. By the inductive argument, terms one through five are already in R, and following the pattern of the second term of (44) eventually shows that terms eight and nine are also in R. We now rewrite the sum of terms six and seven as follows:

$$\sum_{\substack{i,k>0\\ i+k=N+1}} a_{i,0,k,0} \left(z_1^i z_3^k + z_2^i z_4^k \right) + \sum_{\substack{i,k>0\\ i+k=N+1}} a_{i,0,0,k} \left(z_1^i z_4^k + z_2^i z_3^k \right)$$

Observe that for $i, k > 0$:

$$z_1^i z_3^k + z_2^i z_4^k = (z_1 z_3 + z_2 z_4)(z_1^{i-1} z_3^{k-1} + z_2^{i-1} z_4^{k-1}) - z_1^{i-1} z_2 z_3^{k-1} z_4 - z_1 z_2^{i-1} z_3 z_4^{k-1} \quad (51)$$
$$z_1^i z_4^k + z_2^i z_3^k = (z_1 z_4 + z_2 z_3)(z_1^{i-1} z_4^{k-1} + z_2^{i-1} z_3^{k-1}) - z_1 z_2^{i-1} z_3^{k-1} z_4 - z_1^{i-1} z_2 z_3 z_4^{k-1}$$

We conclude by considering the first of the two equations above and noticing that the second equation can be treated similarly due to that $z_1 z_4 + z_2 z_3$ equals $(z_1 + z_2)(z_3 + z_4) - (z_1 z_3 + z_2 z_4)$, and thus lies in R. The following expression in conjunction with the induction argument helps to see why the left hand side of (51) belongs to R:

$$z_1^{i-1} z_2 z_3^{k-1} z_4 + z_1 z_2^{i-1} z_3 z_4^{k-1} = \begin{cases} z_1 z_3 + z_2 z_4, \text{ if } \ i-1 = k-1 = 0 \\ z_3 z_4 (z_2 z_3^{k-2} + z_1 z_4^{k-2}), \text{ if } \ i-1 = 0, \ k-1 > 0 \\ z_1 z_2 (z_1^{i-2} z_4 + z_2^{i-2} z_3), \text{ if } \ i-1 > 0, \ k-1 = 0 \\ z_1 z_2 z_3 z_4 (z_1^{i-2} z_3^{k-2} + z_2^{i-2} z_4^{k-2}), \text{ if } \ i-1, k-1 > 0. \end{cases}$$

Summarizing the results proved to this point, we return to the initial x indeterminates:

$$\mathbb{R}[x_1, x_2, x_3, x_4]^G = \mathbb{R}[(x_1 + x_3)^2 + (x_2 + x_4)^2, x_1 x_3 + x_2 x_4, \quad (52)$$
$$x_1 x_2 x_3 x_4, (x_1 + x_3)^2 x_1 x_3 + (x_2 + x_4)^2 x_2 x_4, (x_1 + x_3)(x_2 + x_4)]$$

These generators are not unique, and recognizing that

$$(x_1 + x_3)^2 + (x_2 + x_4)^2 = f_3(x) + 2f_2(x),$$
$$(x_1 + x_3)^2 x_1 x_3 + (x_2 + x_4)^2 x_2 x_4 = \tfrac{1}{2}[f_5(x) - f_1^2(x)] +$$
$$f_2(x) f_3(x) + 2f_2^2(x) - 2f_4(x),$$

with f_1, f_2, f_3, f_4, f_5 as in (33), makes it clear that

$$\mathbb{R}[x_1, x_2, x_3, x_4]^G = \mathbb{R}[f_1(x), f_2(x), f_3(x), f_4(x), f_5(x)]^G.$$

A straightforward computation verifies that none of the above five generators can be expressed as a real polynomial in the remaining four. We conclude by instantiating a well-known fact (see, for example, [7]):

$$\mathbb{R}[x_1, x_2, x_3, x_4]^G \cong \mathbb{R}[w_1, w_2, w_3, w_4, w_5]/J_F, \text{ where} \quad (34)$$
$$J_F = \{h \in \mathbb{R}[w_1, w_2, w_3, w_4, w_5] : h(f_1, f_2, f_3, f_4, f_5) = 0 \in \mathbb{R}[x_1, x_2, x_3, x_4]\} =$$
$$\langle q \rangle, \text{ and } q(w_1, w_2, w_3, w_4, w_5) = 4w_1^2 w_3 + 8w_1 w_2 w_5 + 2w_1 w_3 w_5 - 2w_1 w_4^2 w_5 +$$
$$16w_2^2 - 8w_2 w_3 - 8w_2 w_4^2 + 4w_2 w_5^2 + w_3^2 - 2w_3 w_4^2 + w_4^4$$

In order to compute J_F, the *syzygy* ideal, one can use, for example, the *elimination method* based on computation of a *Gröbner basis* for the ideal $J_F = \langle f_2 - w_1, f_4 - w_2, f_5 - w_3, f_1 - w_4, f_3 - w_5 \rangle \subset \mathbb{R}[x_1, x_2, x_3, x_4, w_1, w_2, w_3, w_4, w_5]$ [7]. The above generator for J_F was computed analytically and also verified using *Macaulay2* [17]. $\diamond$

F Computational Issues

Proposition 33 Let $B \in \mathcal{B}$ and $h_B(\mathcal{O}) = \frac{|B \cap \mathcal{O}|}{|\mathcal{O}|}$. Then $h_B : \mathcal{S}_W \to \mathbb{R}$ is $\tilde{\mathcal{B}}$-measurable, and $h_B \circ [w] : W \to \mathbb{R}$ is $\mathcal{B}$-measurable.
Proof. Let $B \in \mathcal{B}$, then

$$h_B([w]) = \frac{1}{|G|} \sum_{g \in G} \mathbb{I}_B(gw). \tag{53}$$

To see this, notice

$$\begin{aligned} h_B([w]) &= \frac{1}{|G_w||[w]|} \sum_{hG_w \in G/G_w} |G_w| \mathbb{I}_B(hw) \\ &= \frac{1}{|G|} \sum_{hG_w \in G/G_w} |hG_w| \mathbb{I}_B(hw) && (54) \\ &= \frac{1}{|G|} \sum_{hG_w \in G/G_w} \sum_{g \in hG_w} \mathbb{I}_B(gw) && (55) \\ &= \frac{1}{|G|} \sum_{g \in G} \mathbb{I}_B(gw). && (56) \end{aligned}$$

Equalities (54)-(56) follow from the isomorphism between the orbit $[w]$ and G/G_w, the left cosets hG_w of G_w, the stabilizer of $[w]$. Evidently, $\mathbb{I}_B(gw)$ is measurable for all $g \in G$. ⋄

Proposition 34

$$\mathcal{R}^* = \pi_2^* \circ \pi_1^*$$

Proof. Let $B \in \mathcal{B}$, then

$$\begin{aligned} \mathcal{R}^*(P)(B) &= \frac{1}{|G|} \sum_{g \in G} P(gB) && (57) \\ &= \mathbb{E}_P h_B([\cdot]) \text{ by } (53) \\ &= \mathbb{E}_P \frac{|[w] \cap B|}{|[w]|} \\ &= \mathbb{E}_{\pi_1^*(P)} \frac{|\mathcal{O} \cap B|}{|\mathcal{O}|} && (58) \\ &= \pi_2^* \circ \pi_1^*(P)(B). && (59) \end{aligned}$$

Equality (57) is due to (36) and (7). Equalities (58) and (59) follow from the definitions (23) and (24). ⋄

Acknowledgments

I thank Professor D. Geman for repeatedly jump-starting my interest in statistics of natural microimages that has led to this work. Thanks to Professor A. van der Vaart for help with the proof of Lemma 20. Many thanks to Professor D. Rumynin for consulting me on the algebraic issues, and to Dr. J. Lember and Dr. K. Maulik for helpful discussions. Thanks to Professor R. Gill and Dr. P. Grünwald for their constructive criticism, useful comments, and approval. Thanks also to Professors A. Di Bucchianico, A. Cohen, and H. Wynn for discussing with me the algebraic and computational aspects of this work. Special thanks to my friend I. Snihir for inspiration.

References

[1] C. Berg. Recent results about moment problems. In *Probability measures on groups and related structures.*

[2] C. Berg. Moment problems and polynomial approximations. 100 ans après Th.-J. Stieltjes. *Ann. Fac. Sci. Toulouse Mathématiques*, pages 9–32, 1996.

[3] P. Billingsley. *Probability and Measure*, pages 380–381. John Wiley & Sons, 3d edition, 1995.

[4] J. Borwein. Maximum entropy-type methods & convex programming. Published on web at http://www.cecm.sfu.ca/personal/jborwein/lbl.pdf, May 2001. Presented at Lawrence Berkeley National Laboratory Workshop on "New Approaches to Phase Problem for Non-Periodic Objects".

[5] Computational Algebra Group School of Mathematics and Statistics University of Sydney. http://magma.maths.usyd.edu.au/magma. *The Magma Computational Algebra System. Release Notes V2.10.*

[6] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* John Wiley, New York, 1991.

[7] D. Cox, J. Little, and O'Shea D. *Ideals, Variety, and Algorithms.* Springer, 1996.

[8] M. de Jeu. Determinate multidimensional measures, the extended Carleman theorem and quasi-analytic weights. *The Annals of Probability*, 31(3), 2003.

[9] H. Derksen and G. Kemper. *Computational Invariant Theory*, volume 130 of *Encyclopaedia of Mathematical Sciences.* Springer, 2002.

[10] D. S. Dummit and R. M. Foote. *Abstract Algebra.* Prentice Hall, Inc., 1991.

[11] R. Durrett. *Probability: Theory and Examples*, pages 107–111. Duxbury Press, 2d edition, 1996.

[12] R. Durrett. *Probability: Theory and Examples*, page 90. Duxbury Press, 2d edition, 1996.

[13] J. Fogarty. On Noether's bound for the degrees of generating invariants. Technical report, Mathematics & Statistics Dept., Univ. of Massachusetts Amherst, 1999.

[14] The GAP Group. *GAP – Groups, Algorithms, and Programming, Version 4.3*. http://www.gap-system.org.

[15] D. Geman and A. Koloydenko. Invariant statistics and coding of natural microimages. In S.C. Zhu, editor, *IEEE Workshop on Statistical and Computational Theories of Vision*, Published on web at http://department.stat.ucla.edu/∼sczhu/Workshops/sctv99/Geman1.html, 1999.

[16] A. Golan, G. Judge, and D. Miller. *Maximum Entropy Econometrics*. Series in Financial Econometrics and Quantitative Analysis. Wiley, 1996).

[17] Daniel R. Grayson and Michael E. Stillman. Macaulay 2, a software system for research in algebraic geometry. Available at http://www.math.uiuc.edu/Macaulay2/.

[18] P. Grünwald, Myung I.J., and M. Pitt, editors. *Minimum Description Length: recent advances in theory and practice*, chapter 1. MIT Press, 2004.

[19] J. Huang and D. Mumford. Statistics of natural images and models. In *Computer Vision and Pattern Recognition*, 1999.

[20] M. Junk. Maximum entropy for reduced moment problems. *Mathematical Models and Methods in Applied Sciences*, 10:1001–10251, 2000.

[21] U. Keich. Krein's strings, the symmetric moment problem, and extending a real positive definite function. *Communications on Pure and Applied Mathematics*, (10):1315–1334, 1999.

[22] G. Kemper. INVAR. A Maple package for invariant theory of finite groups. Available at http://www.iwr.uni-heidelberg.de/∼Gregor.Kemper/invar.html.

[23] O. Knill. On Hausdorff's moment problem in higher dimensions. Published on web at http://abel.math.harvard.edu/∼knill/preprints/stability.ps, 1997.

[24] A. Koloydenko. *Modeling Natural Microimage Statistics*. PhD thesis, University of Massachusetts Amherst, 2000.

[25] S. Kullback. *Information Theory and Statistics*. Dover Publications, Inc., 1997.

[26] A. Lee, K. Pedersen, and D. Mumford. The complex statistics of high contrast patches in natural images. In S.C. Zhu, editor, *IEEE Workshop on Statistical and Computational Theories of Vision*, Published on web at http://department.stat.ucla.edu/∼sczhu/Workshops/sctv01/Lee.html, 2001.

[27] E. Lehmann. *Testing statistical hypotheses*. New York : Springer, 1997.

[28] A. Lippman. *A maximum entropy method for expert system construction*. PhD thesis, Brown University, 1986.

[29] L. Mead and N. Papnicolaou. Maximum entropy in the problem of moments. *Journal of Mathematical Physics*, 25(8):2404–2417, 1984.

[30] P. J. Olver. *Classical Invariant Theory*. Cambridge University Press, 1999.

[31] A. Pakes. Criteria for the unique determination of probability distributions by moments. *Aust NZ J Stat*, 43(1):101–101, 2001.

[32] K. Pedersen and A. Lee. Toward a full probability model of edges in natural images. In A. Heyden, G. Sparr, M. Nielsen, and P. Johansen, editors, *7th European Conference on Computer Vision*, Available at http://link.springer.de/link/service/series/0558/papers/2350/23500328.pdf, 2002.

[33] M. Schervish. *Theory of Statistics*. Springer-Verlag New York, Inc., 1997.

[34] L. Smith. *Polynomial Invariants of Finite Groups*. A K Peters, Ltd., 1995.

[35] J. Stoyanov. Krein condition in probabilistic moment problems. *Bernoulli*, 5(6):939–949, 2000.

[36] B. Sturmfels. *Algorithms in invariant theory*. Springer-Verlag/Wien, 1993.

[37] X. Wu. Calculation of maximum entropy densities with application to income distribution. *Journal of Econometrics*, (115):347–354, 2003.

[38] E. Zeidler. *Applied Functional Analysis: Applications to Mathematical Physics*, volume 108 of *Applied Mathematical Sciences*. Springer-Verlag New York, Inc., 1995.

[39] S. C. Zhu. Embedding Gestalt laws in Markov random fields. *IEEE Trans. PAMI*, 21, November 1999.

[40] S. C. Zhu, A. Lanterman, and M. Miller. Clutter modeling and performance and analysis in automatic target recognition. In *Workshop on Detection and Classification of Difficult Targets*. Redstone Arsenal, 1998.

[41] S. C. Zhu and D. Mumford. Prior learning and Gibbs reaction-diffusion. *IEEE Trans. PAMI*, 19:1236–1250, 1997.