UNITARY REPRESENTATIONS OF REDUCTIVE LIE GROUPS

DAVID A. VOGAN, JR.

ABSTRACT. One of the fundamental problems of abstract harmonic analysis is the determination of the irreducible unitary representations of simple Lie groups. After recalling why this problem is of interest, we discuss the present state of knowledge about it. In the language of Kirillov and Kostant, the problem finally is to "quantize" nilpotent coadjoint orbits.

1. INTRODUCTION

In the 1930s I. M. Gelfand outlined a program of abstract harmonic analysis, which offered a paradigm for the use of symmetry to study a very wide class of mathematical problems. In this paper I want to explain Gelfand's program, and to look in some detail at one of the unsolved problems standing in the way of further applications of it.

This paper is based on a lecture delivered at the conference "Mathematics towards the third millenium" in May, 1999. I am very grateful to the organizers for the opportunity to participate in this conference.

The setting for Gelfand's program is a group G acting on a space X. We have in mind some class of "interesting questions" about X. Almost anything is allowed. If X is finite, we can ask for its cardinality; if it is a topological space, about its homology; if it is a Riemannian manifold, about eigenspaces of the Laplace operator.

In this setting, Gelfand's program proceeds in four steps. The first step is to attach to X a vector space V, so that questions about X can be translated into questions about V. Roughly speaking V should be thought of as a space of functions on X, although often something a little different is needed. For example, to study the cohomology of a manifold, we might look not at functions on X but at the whole complex of differential forms. A fundamental requirement is that the action of G on X should lift to a linear action

$$G \times V \to V, \qquad (g, v) \mapsto \pi(g)v \tag{1.1)(a}$$

of G on V.

The second step in Gelfand's program is to find the finest possible G-invariant decomposition

$$V = \sum_i V_i. \tag{1.1)(b}$$

1991 *Mathematics Subject Classification.* Primary 22E46.

Supported in part by NSF grant 9721441

Typeset by $\mathcal{A}_{\mathcal{M}}\mathcal{S}$-TEX

We have already translated our questions about X into questions about V; now we want to use this decomposition to translate them into questions about each V_i. The decomposition (1.1)(b) should be thought of as a G-equivariant analogue of a basis of V. When V is infinite-dimensional, the algebraic notion of basis must often be replaced by a topological one. A simple example is an orthonormal basis of a Hilbert space; a more subtle one is thinking of $L^2(X)$ as a "direct sum" of delta functions at the points of X. In the G-equivariant setting the possible complications multiply, and precise versions of (1.1)(b) are available only under fairly strong hypotheses (for example, when V is a Hilbert space and G acts by unitary operators). We will not address these questions here. Finally the hope is that each V_i is an irreducible representation of G. (Careful definitions of this and other technical terms are collected in section 3. For the moment all that matters is that the irreducibility of each V_i corresponds to the requirement that the decomposition in (1.1)(b) be as fine as possible.)

The third step in Gelfand's program (and the main topic for this paper) is to understand the set

$$\widehat{G} = \{\text{equivalence classes of irreducible representations of } G\}. \tag{1.1)(c}$$

Our questions about X have been translated into certain questions about the irreducible representations V_i; so "understand" here should mean "be able to answer these questions."

The fourth and final step in Gelfand's program is to assemble our information about the irreducible representations V_i into answers to our questions about V, and so finally into information about X.

The rest of this introduction is devoted to two examples of Gelfand's program. The first is a toy example, designed to show what the words mean. The second is more serious; it shows the kind of information about a real problem that one can hope to find just by studying irreducible representations.

For the first example, let $G = S_n$, the symmetric group on n letters; and let X^p be the collection of p-element subsets of $\{1, \dots, n\}$. (Often we will tacitly assume n is at least one.) The question we ask is this: what is the cardinality of X^p? The first step of Gelfand's program asks us to translate this into a linear algebra question about a vector space of something like functions on X^p. In this case it is natural to define

$$V^p = \text{functions on } X^p; \tag{1.2)(a}$$

then the cardinality of X^p is equal to the dimension of V^p. Of course G acts on V^p by

$$(\pi(g)f)(x) = f(g^{-1}x) \qquad (f \in V^p,\ x \in X^p). \tag{1.2)(b}$$

There is an obvious G-equivariant identification $X^p \simeq X^{n-p}$ (by taking complements of subsets); so we may as well assume henceforth that $p \le n - p$.

The second step in Gelfand's program is to find the finest possible G-equivariant decomposition of V^p. There is one obvious G-invariant subspace, consisting of constant functions:

$$V_{(n)} = \{f \in V^p \mid f(x) = f(y) \quad (x, y \in X^p)\}. \tag{1.2)(c}$$

(The subscript (n) refers to the classification of irreducible representations of S_n by partititions of n. The trivial representation corresponds to the trivial partition $n = n$.) With a little more thought one can invent a complementary G-invariant subspace

$$W = \{f \in V^p \mid \sum_{x \in X^p} f(x) = 0\}; \tag{1.2)(d}$$

then there is a G-invariant decomposition

$$V^p = V_{(n)} \oplus W.$$

If $p = 0$ or 1, then this decomposition cannot be further refined. (In fact if $p = 0$, then $W = 0$; the one-dimensional space $V^p = V_{(n)}$ admits no further decomposition.) For p between 2 and $n - 2$, however, we can decompose W. One way to do this is using the "Radon transforms," defined for $q \leq p$ by

$$T^{p,q}: V^p \to V^q \qquad (T^{q,p} f)(y) = \sum_{x \in X^p,\ y \subset x} f(x),$$

$$S^{q,p}: V^q \to V^p \qquad (S^{q,p} g)(x) = \sum_{y \in X^q,\ y \subset x} g(y).$$

These linear transformations are "intertwining operators" for the representations of G (see section 2). In particular, their images and kernels are G-invariant subspaces. Notice that $V_{(n)}$ is the image of $S^{0,p}$, and W is the kernel of $T^{p,0}$. I will not continue the analysis of these Radon transforms, but here is the conclusion. There is a G-invariant direct sum decomposition

$$V^p = \sum_{0 \leq r \leq p} V_{(n-r,r)}. \tag{1.2)(e}$$

The image of $S^{q,p}$ is $\sum_{0 \leq r \leq q} V_{(n-r,r)}$, and the kernel of $T^{p,q}$ is $\sum_{q < r \leq p} V_{(n-r,r)}$. This decomposition has no G-invariant refinement. Again the subscript $(n - r, r)$ refers to the fact that the representation of S_n on $V_{(n-r,r)}$ is the one parametrized by the partition $n = (n - r) + r$. Recall that we want to translate our question about V (what is the dimension of V?) into questions about the subspaces $V_{(n-r,r)}$. This is very easy:

$$\dim V = \sum_{0 \leq r \leq p} \dim V_{(n-r,r)}. \tag{1.2)(f}$$

The third step in Gelfand's program is to understand all the irreducible representations of S_n. "Understand" in this problem means "be able to calculate the dimension." In our case consideration of arbitrary representations may seem like unnecessary generality; the representations we need are explicitly given by the Radon transforms. For example,

$$V_{(n-r,r)} = \operatorname{im} S^{r,p} \cap \ker T^{p,r-1}.$$

The difficulty is that this description does not immediately reveal the dimension of $V_{(n-r,r)}$. This difficulty can be overcome, but I will instead follow Gelfand's program more literally. What one discovers is that

$$\widehat{S_n} = \{\text{partitions of } n\}. \tag{1.2)(g}$$

Writing V_τ for the irreducible representation corresponding to a partition τ, then it turns out that

$$\dim V_\tau = \text{number of standard Young tableaux of shape } \tau \tag{1.2)(h}$$

There are many ways to count the standard tableaux of shape $(n-r,r)$; one finds

$$\dim V_{(n-r,r)} = \binom{n}{r}\left(\frac{n-2r+1}{n-r+1}\right). \tag{1.2)(i}$$

The fourth step in Gelfand's program is to assemble all of this information to answer our original question about X^p. Combining (1.2)(e), (1.2)(f), and (1.2)(i), we find (for $p \leq n-p$)

$$\text{cardinality of } X^p = \sum_{r=0}^{p} \binom{n}{r}\left(\frac{n-2r+1}{n-r+1}\right) = \binom{n}{p}; \tag{1.2)(j}$$

the last equality is a fairly easy exercise (by induction on p).

For the second example, we take G to be a connected linear real reductive Lie group (see Definition 5.1 below), $\Gamma \subset G$ a discrete cocompact subgroup, and $X = G/\Gamma$. Let $K \subset G$ be a maximal compact subgroup; then

$$Z = K\backslash G/\Gamma = K\backslash X \tag{1.3)(a}$$

is a Riemannian locally symmetric space. The problem we consider is to understand the de Rham cohomology $H^p(Z)$. The group G does not act on the double coset space Z; so even the first step in Gelfand's program requires some ingenuity. (For details, background, and motivation, one can consult [BW].) Since we are interested in cohomology of Z, it would be natural to consider the space of p-forms on Z; but G does not act on this space. We use instead

$$V = C^\infty(X). \tag{1.3)(b}$$

The action of G on this space is differentiable, and so gives rise to a representation of the complexified Lie algebra $\mathfrak{g}$. Now using any representation W of $\mathfrak{g}$ equipped with a compatible representation of K, one can construct "relative Lie algebra cohomology groups" $H^p(\mathfrak{g}, K; W)$. These are the cohomology groups of a certain complex

$$\mathrm{Hom}_K(\wedge^p(\mathfrak{g}/\mathfrak{k}), W);$$

a definition of the differential (and more motivation) may be found in [BW]. Because $\mathfrak{g}/\mathfrak{k}$ is the (complexified) tangent space to Z at the identity coset, it is not hard to see that this complex for our choice of V may be identified with differential forms on Z:

$$\mathrm{Hom}_K(\wedge^p(\mathfrak{g}/\mathfrak{k}), C^\infty(X)) \simeq \text{smooth } p\text{-forms on } K\backslash X. \tag{1.3(c)}$$

The identification respects the differential. (This is no accident: relative Lie algebra cohomology was introduced in order to study the cohomology of homogeneous spaces for compact groups. The isomorphism in (1.3)(c) holds whenever G is a Lie group acting on X with open orbits, and K is a subgroup acting freely on X.) At any rate, we have expressed the cohomology of Z as a certain invariant (the relative Lie algebra cohomology) of the representation V.

The second step in Gelfand's program is to decompose V in a G-invariant way. This is a version of the basic problem of automorphic representation theory; our assumption that Γ is cocompact simplifies the analytic aspects of the problem enormously, but does not help with the far more difficult arithmetic ones. We will be content with a qualitative statement: there is a decomposition

$$V = \sum_{\pi \in \widehat{G}_u} m_\pi(\Gamma) V_\pi^\infty. \tag{1.3(d)}$$

The sum is over irreducible unitary representations π of G; that is the meaning of the subscript u on $\widehat{G}$. The multiplicities m_π are all finite, and they are positive only for a countable set of π. The space V_π^∞ is the space of smooth vectors in the Hilbert space V_π. The direct sum requires a topological interpretation, which we omit. (If we had worked instead with $V = L^2(X)$, then we would get a Hilbert space direct sum of the Hilbert spaces V_π. Then the difficulty would appear at (1.3)(c), where we would have instead something like L^2 p-forms.) What (1.3)(c) and (1.3)(d) suggest is that

$$\dim H^p(Z) = \sum_\pi m_\pi(\Gamma) \dim H^p(\mathfrak{g}, K; V_\pi^\infty). \tag{1.3(e)}$$

This result, due to Matsushima, is true. Of course the Betti numbers on the left are finite; what happens is that there are only finitely many π for which $H^p(\mathfrak{g}, K; V_\pi^\infty) \neq 0$, and for those π the cohomology is finite-dimensional.

The third step in Gelfand's program is the study of irreducible representations. For the present problem we can narrow it to this: classify the irreducible unitary representations π with $H^p(\mathfrak{g}, K; V_\pi^\infty) \neq 0$. For reasons that are beautifully illuminated in [Gu], the main example is the trivial representation $V_\pi = \mathbb{C}$. This always appears in $C^\infty(X)$ with multiplicity $m_\pi = 1$, realized on the space of constant functions. Its relative Lie algebra cohomology can be interpreted as follows. As a connected linear reductive group, G may be realized as a subgroup of $GL(n, \mathbb{C})$ stable under inverse conjugate transpose, with $K = G \cap U(n)$. Let $\mathfrak{p}_0$ be the space of Hermitian matrices in the Lie algebra $\mathfrak{g}_0$ (regarded as an algebra of $n \times n$ matrices), so that $\mathfrak{g}_0 = \mathfrak{k}_0 + \mathfrak{p}_0$. The "compact dual" G^c of G is by definition the connected subgroup of $U(n)$ with Lie algebra

$$\mathfrak{g}_0^c = \mathfrak{k}_0 + i\mathfrak{p}_0.$$

It is a compact group containing K. The "compact dual of G/K" is the compact symmetric space

$$U = G^c/K.$$

(Actually G^c and the space U may change by a finite covering if the embedding of G in $GL(n,\mathbb{C})$ changes, but the de Rham cohomology of U is well-defined.) An argument beginning with (1.3)(c) for $X = G^c$ shows that

$$H^p(\mathfrak{g}, K; \mathbb{C}) \simeq H^p(U). \tag{1.3)(f}$$

A complete classification of the irreducible unitary representations π for which $H^p(\mathfrak{g}, K; V_\pi^\infty) \neq 0$ may be found in [VZ]. Each of the non-trivial unitary representations π in this classication is attached to a smaller compact symmetric space U_π, and its relative Lie algebra cohomology is given by

$$H^p(\mathfrak{g}, K; V_\pi) \simeq H^{p-d_\pi}(U_\pi), \qquad d_\pi = \frac{1}{2}(\dim U - \dim U_\pi). \tag{1.3)(g}$$

This formula was actually first proved by Kumaresan in [Ku]. One very interesting feature of it is that the cohomology vanishes for $p < d_\pi$.

The last step in Gelfand's program is to assemble this information about irreducible representations into information about the cohomology of the locally symmetric space Z. One explicit form of the result is

$$\dim H^p(Z) = \dim H^p(U) + \sum_{\pi \neq 1} m_\pi(\Gamma) \dim H^{p-d_\pi}(U_\pi). \tag{1.3)(h}$$

Recall that here U is the compact symmetric space dual to Z, and the various U_π are smaller compact symmetric spaces. Because the numbers $m_\pi(\Gamma)$ (which are essentially dimensions of certain spaces of automorphic forms) are so difficult to compute, this formula does not at first appear to be very informative. But the remark at the end of the last paragraph suggests writing

$$H^p(K\backslash G/\Gamma) \simeq H^p(U), \qquad (p \leq \min_{\pi \neq 1} d_\pi). \tag{1.3)(i}$$

This is the Kumaresan vanishing theorem. (The idea is that the contribution of $H^p(U)$ is the trivial part of the cohomology of Z. What is "vanishing" is the non-trivial part.) The possible spaces U_π are easy to enumerate, so the minimum (which is some positive integer depending only on G) is easy to compute; it is tabulated in [VZ], Table 8.2.

2. Formalism of representation theory

In this section we collect some of the basic definitions associated with Gelfand's program, and recall very briefly some of the general results that justify our emphasis on unitary representations of reductive groups.

Representation theory is linear algebra with a group action. We will try to emphasize the correspondence between elementary notions from linear algebra and the representation-theoretic ideas. We begin with a topological group G. A *representation of* G is a complex topological vector space V endowed with a continuous action

$$G \times V \to V, \qquad (g, v) \mapsto \pi(g)v \tag{2.1)(a}$$

so that all the operators $\pi(g)$ are linear operators on V. We will often refer to the map π, or to the pair (π, V), as the representation. The *dimension of* π is by definition the dimension of V. An *invariant subspace of* V is a closed subspace $W \subset V$ preserved by all the operators $\pi(g)$:

$$\pi(g)w \in W \qquad (g \in G, w \in W). \tag{2.1)(b}$$

Two examples are $W = \{0\}$ and $W = V$. We say that V is *irreducible* if there are precisely two invariant subspaces. Because the zero vector space has only one subspace, an irreducible representation is non-zero. Any one-dimensional representation is irreducible.

The notion of irreducible representation seems at first not to have an analogue in linear algebra. In fact the analogue is a vector space with exactly two subspaces: that is, a one-dimensional vector space. The theory of bases is concerned with decomposing an arbitrary vector space as a sum of one-dimensional subspaces. In representation theory, one seeks in a parallel way to decompose an arbitrary representation as a sum of irreducible representations. In contrast with the linear algebra situation, such a decomposition is not always possible.

Suppose (π, V) and (ρ, W) are two representations of the same group G. An *intertwining operator from* V *to* W is a continuous linear map

$$T: V \to W, \qquad T \circ \pi = \rho \circ T. \tag{2.2}$$

The requirement just means that T should respect the actions of G on V and on W. Clearly the kernel of T is an invariant subspace of V, and the closure of the image of T is an invariant subspace of W. The vector space of all intertwining operators from V to W is written $\mathrm{Hom}_G(V, W)$. The composition of intertwining operators is an intertwining operator. In particular, $\mathrm{Hom}_G(V, V)$ is a complex associative algebra, with the identity operator as unit.

Intertwining operators play the role of linear operators in linear algebra.

Two representations (π, V) and (ρ, W) of G are called *equivalent* if there is an invertible intertwining operator $T \in \mathrm{Hom}_G(V, W)$ with T^{-1} continuous. (Once T^{-1} is known to be continuous, the fact that it is an intertwining operator is automatic.) Write (tentatively!)

$$\widehat{G} = \{\text{equivalence classes of irreducible representations of } G\}.$$

This definition turns out to be badly behaved for infinite-dimensional representations. The reason is that there are too many non-isomorphic topological vector spaces. Here is a family of examples. The group $G = GL(2, \mathbb{R})$ acts on the real projective space $X = \mathbb{RP}^1$ of lines in $\mathbb{R}^2$; this is just a circle. Consequently G acts on function spaces on the circle: $L^p(X)$ (for $1 \leq p < \infty$), $C^\infty(X)$, the analytic functions $C^\omega(X)$, various Sobolev spaces, distribution spaces, and so on. All of these are continuous representations of G. They are not quite irreducible, since each contains a one-dimensional invariant subspace W of constant functions; but in every case the quotient V/W is an irreducible representation. All of these representations of G are inequivalent. The only intertwining operators from $C^\infty(X)/W$ to $L^p(X)/W$, for example, come from the scalar multiples of the natural embedding $C^\infty(X) \hookrightarrow L^p(X)$. (This fact is not obvious, but it is not terribly difficult to prove.) The embedding is not invertible because it is not surjective: there are L^p functions that are not smooth.

On the other hand, all these representations of G are clearly closely related; for many purposes, it is convenient to identify them. There is no machinery available to do that in general, but at least two approaches are sometimes useful.

The first is to restrict enormously the class of representations considered, in such a way that only one of the examples for $GL(2,\mathbb{R})$ appears. A representation (π, V) of G is called *unitary* if V is a Hilbert space, and the action of G preserves the inner product; that is, if all the operators $\pi(g)$ are unitary. If W is an invariant subspace (closed by definition!) of the unitary representation (π, V), then the orthogonal complement $W^\perp$ is also invariant, and

$$V = W \oplus W^\perp. \tag{2.3)(a}$$

If W is neither 0 nor V, then this G-invariant decomposition is non-trivial. Consequently a non-zero unitary representation is irreducible if and only if it cannot be written as a non-trivial direct sum. (The corresponding statement in linear algebra is that a non-zero vector space has dimension one if and only if it cannot be written as a non-trivial direct sum.) We write

$$\widehat{G}_u = \{\text{equivalence classes of irreducible unitary representations of } G\}. \tag{2.3)(b}$$

This definition is fundamental, so we restate it: an irreducible unitary representation of G consists of a non-zero complex Hilbert space V and a continuous homomorphism π from G to the group of unitary operators on V, with the property that no proper closed subspace of V is invariant under all the operators $\pi(g)$. (Here the group of unitary operators is given the strong topology to define the continuity of the map π.)

Theorem 2.4 [Dix], Théorème 8.5.2. *Any unitary representation of a locally compact group G on a separable Hilbert space V may be written as a direct integral of irreducible unitary representations of G.*

The notion of direct integral generalizes that of Hilbert space direct sum. Details, and a discussion of the uniqueness of the decomposition, may be found in [Dix]. In the setting of unitary representations, this result corresponds to the second step of Gelfand's program of abstract harmonic analysis. More precisely, it is a fairly general guarantee that the second step is possible; as was evident in the examples in section 1, it is often important to know something about which irreducible unitary representations actually appear in the direct integral decomposition.

In light of Theorem 2.4, a reasonable special case of the third step in Gelfand's program is

Problem 2.5. For every locally compact group G, describe the irreducible unitary representations of G.

That is, we are asking for a description of $\widehat{G}_u$. This is the "unitary dual problem." In the case of an abelian locally compact group A, there is a beautiful solution. The set $\widehat{A}_u$ has in a natural way the structure of a locally compact abelian group, the "dual group" of A. The double dual of A is canonically isomorphic to A. The duality relation on locally compact abelian groups has been worked out explicitly in a wide range of examples, and the results are at the bottom of many deep and beautiful parts of mathematics: Fourier series, the Fourier transform, and class field theory, for example.

Understanding of the non-abelian case has come more slowly. Building on ideas of Eugene Wigner about the Lorentz group, George Mackey in the 1950s made a deep study of $\widehat{G}_u$ when G has a closed normal subgroup N. (A good place to read about this work is [Ma].) Roughly speaking, he showed how to build irreducible unitary representations of G from those of N and of G/N. This statement must be carefully qualified to be correct; but nevertheless it suggests that one should focus attention on groups G having no non-trivial closed normal subgroups; that is, on simple groups. Duflo in [Du] has made this idea precise for an algebraic Lie group G, using Mackey's work to give an explicit description of $\widehat{G}_u$ in terms of unitary duals of smaller reductive Lie groups. (A reductive Lie group is one that is locally isomorphic to a direct product of simple Lie groups.) For algebraic Lie groups, Problem 2.5 has therefore been reduced to

Problem 2.6. For every reductive Lie group G, describe the irreducible unitary representations of G.

This problem will be the topic of the rest of the paper.

3. Quantum mechanics and classical mechanics

Problem 2.6 asks us to find Hilbert spaces equipped with nice families of unitary operators. The idea that we will describe for doing that comes from mathematical physics, by way of Kirillov and Kostant. Here is an outline. The mathematical setting for quantum mechanics is a Hilbert space and a nice family of unitary operators. Quantum mechanical systems often correspond formally to classical mechanical ones. One might therefore hope that there is a group-theoretic object that is a "classical analogue" of a unitary representation. What Kirillov and Kostant accomplished was to define just such a classical analogue, and then to classify completely the corresponding "irreducible" objects. Their hope (realized in many settings) was that there should be a method of "quantization" for passing from these classical analogues to actual unitary representations, and that one should get in this way something close to a solution to Problem 2.6.

In order to justify their definitions, we recall first some of the barest rudiments of the corresponding ideas in mathematical physics.

For quantum mechanics we refer to [MQ], in part because this book is written with connections to unitary representation theory in mind. A quantum-mechanical physical system corresponds to a complex Hilbert space $\mathcal{H}$. The possible states of the system are parametrized by lines in $\mathcal{H}$. Physical observables correspond to operators $\{A_j\}$ on $\mathcal{H}$. If the system is in a state parametrized by a unit vector $v \in \mathcal{H}$, then the result of the observation corresponding to an operator A has a certain probability distribution. The expectation of this distribution is $\langle Av, v\rangle$; notice that this depends only on the line in which the unit vector v lies. These expectations need not be finite for every v; correspondingly, the operators A_j may be unbounded.

There is a distinguished observable called the energy, corresponding to a skew-adjoint operator A_0. Attached to A_0 is a one-parameter group of unitary operators

$$U(t) = \exp(tA_0). \tag{3.1}$$

These operators govern the time evolution of the quantum-mechanical system, in the sense that if the state of the system at time t_0 is v_0, then the state at time $t+t_0$ is $U(t)v_0$. One consequence is that evaluating the observable A at time $t+t_0$ is like evaluating $\exp(-tA_0)A\exp(tA_0)$ at time t_0. Said briefly, the observable A evolves as $\exp(-tA_0)A\exp(tA_0)$. In particular, the observable A is conserved (constant in time) if and only if $[A_0, A] = 0$. To summarize: a quantum-mechanical system is a complex Hilbert space equipped with a family $\{A_j\}$ of operators. The commutation relations among these operators control some of the basic physics.

We turn next to the mathematical formalism of classical mechanics. A convenient reference is [Ar]. Recall first of all that a *symplectic manifold* is a manifold M endowed with a Lie algebra structure $\{,\}$ on $C^\infty(M)$ called the *Poisson bracket.* This bracket must satisfy

$$\{a, bc\} = \{a, b\}c + b\{a, c\} \tag{3.2}$$

and a certain nondegeneracy condition. (One excuse for omitting a statement of the nondegeneracy condition is this. Without it one gets not a symplectic manifold but a *Poisson manifold.* All of the formalism below still makes sense; and indeed it is sometimes convenient to work in this more general setting.) Each smooth function f on M defines a *Hamiltonian vector field*

$$\xi_f = \{f, \cdot\}. \tag{3.2)(b}$$

A classical mechanical system corresponds to a symplectic manifold M. A state of the system (usually corresponding to something like knowledge of the positions and velocities of all the particles) corresponds to a points in M. Physical observables correspond to smooth functions $\{a_j\} \subset C^\infty(M)$; the value of the observable a on the state m is the number $a(m)$. Again there is a distinguished observable called the energy, corresponding to a real-valued function a_0. The time evolution of the system is the flow of the corresponding Hamiltonian vector field ξ_{a_0}. That is, the time history of a particle is a smooth function m from $\mathbb{R}$ to M, satisfying the differential equation

$$\frac{dm}{dt} = \xi_{a_0}(m(t)). \tag{3.3)(a}$$

A consequence is that the evolution of any other observable a is governed by the differential equation

$$\frac{d(a \circ m)}{dt} = \{a_0, a\}. \tag{3.3)(b}$$

In particular, the observable a is conserved if and only if $\{a_0, a\} = 0$. To summarize, a classical mechanical system is a symplectic manifold equipped with a family $\{a_j\}$ of smooth functions. The Poisson bracket relations among these functions control some of the basic physics.

4. Unitary representations and "classical" representations.

We begin by recasting our description of a unitary representation $(\pi, \mathcal{H})$ of a Lie group G so as to emphasize the analogy with quantum mechanics. Write

$$\mathfrak{g}_0 = \operatorname{Lie} G \tag{4.1)(a}$$

for the Lie algebra of G. Each element $X \in \mathfrak{g}_0$ defines a one-parameter subgroup $\exp(tX)$ of G. Applying the unitary representation π gives a one-parameter group of unitary operators $\pi(\exp(tX))$ on $\mathcal{H}$. According to Stone's theorem, such a one-parameter group is attached to a (possibly unbounded) skew-adjoint operator $d\pi(X)$ on $\mathcal{H}$, by the requirement

$$\pi(\exp(tX)) = \exp(td\pi(X)) \qquad (X \in \mathfrak{g}_0, t \in \mathbb{R}). \tag{4.1)(b}$$

The Lie group structure on G is reflected in commutation relations

$$[d\pi(X), d\pi(Y)] = d\pi([X,Y]) \qquad (X, Y \in \mathfrak{g}_0). \tag{4.1)(c}$$

(There are serious problems about domains in forming these commutators of unbounded operators, but these will not affect our search for inspiration.)

In the analogy between unitary representations and quantum mechanical systems, the family of skew-adjoint ops $\{d\pi(X)\}$ corresponds to the physical observables. The irreducibility condition is that these operators should have no common closed invariant subspace but 0 and $\mathcal{H}$. It is worth noticing that the quantum mechanical systems arising in this way are extremely special ones, in that the collection of observables is both finite-dimensional and closed under commutation.

We can now define "classical" representations, following Kirillov and Kostant. A *Hamiltonian G-space* is first of all a symplectic manifold M with smooth action

$$G \times M \to M, \qquad (g, m) \mapsto g \cdot m \tag{4.2)(a}$$

respecting the symplectic structure. Each $X \in \mathfrak{g}_0$ defines by this smooth action a vector field $\xi(X)$ on M. The second requirement is a G-equivariant Lie algebra homomorphism

$$\mathfrak{g}_0 \to C^\infty(M), \qquad X \mapsto f(X) \tag{4.2)(b}$$

satisfying

$$\xi(X) = \xi_{f(X)} = \{f(X), \cdot\}. \tag{4.2)(c}$$

Notice that the quantum family of operators $\{d\pi(X) \mid X \in \mathfrak{g}_0\}$ (sending the Lie bracket to commutator) has been replaced by a classical family of functions $\{f(X) \mid X \in \mathfrak{g}_0\}$ (sending the Lie bracket to Poisson bracket). The classical analogue of "irreducibility" is the requirement that M be a homogeneous space for G.

Our goal is the classification of irreducible unitary representations. The physical analogy suggests that we warm up by trying to classify homogeneous Hamiltonian G-spaces. The linear map f from the vector space $\mathfrak{g}_0$ to the space of smooth functions on M is equivalent to a smooth map μ from M to the dual vector space $\mathfrak{g}_0^*$: explicitly,

$$\mu(m)(X) = f(X)(m) \qquad (m \in M, X \in \mathfrak{g}_0). \tag{4.2)(d}$$

This formulation suggests that we might look around $\mathfrak{g}_0^*$ for Hamiltonian G-spaces.

Elements of $\mathfrak{g}_0$ may be regarded as smooth (linear) functions on $\mathfrak{g}_0^*$. The Lie bracket on $\mathfrak{g}_0$ extends uniquely to a Poisson bracket $\{,\}$ on $C^\infty(\mathfrak{g}_0^*)$. Now fix

$$M = G \cdot \lambda \simeq G/H \subset \mathfrak{g}_0^* \tag{4.3}$$

an orbit of G. (This is a *coadjoint orbit*, because the action of G on $\mathfrak{g}_0^*$ is called the *coadjoint action.*) The Poisson bracket on $\mathfrak{g}_0^*$ restricts to (G-invariant) symplectic structure on M. Elements $X \in \mathfrak{g}_0$ (regarded as functions on $\mathfrak{g}_0^*$) restrict to functions $f(X) \in C^\infty(M)$. In this way $M = G/H$ becomes a homogeneous Hamiltonian G-space. This structure is inherited by any G-equivariant covering space $\widetilde{M} = G/H_1$, with $H \supset H_1 \supset H_0$.

Theorem 4.4 (Kirillov-Kostant; see [K], Theorem 5.4.1). *Every homogeneous Hamiltonian G-space is (by means of the moment map $\mu: M \to \mathfrak{g}_0^*$) a G-equivariant cover of a coadjoint orbit.*

According to this theorem, classifying "classical" representations amounts to classifying the orbits of G on the dual of its Lie algebra. We will see that this is an entirely tractable problem. The analogy with mathematical physics outlined at the beginning of section 3 now suggests

Problem 4.5. For every reductive Lie group G, and all (appropriately "integral") coadjoint orbits $M = G \cdot \lambda \subset \mathfrak{g}_0^*$, find a "quantization procedure" to produce an associated unitary representation $\pi(M)$ of G.

This is the "philosophy of coadjoint orbits" of Kirillov and Kostant, and discussing it will occupy the rest of this paper. We can summarize where we are with a diagram of analogies and wishful thinking:

$$\begin{array}{ccc} \text{unitary representations} & \longleftrightarrow & \text{quantum mechanical systems} \\ \uparrow & & \uparrow \\ \text{coadjoint orbit covers} & \longleftrightarrow & \text{classical mechanical systems} \end{array}$$

The right vertical arrow is "quantization"; it should exist since the world exists, and is quantum-mechanical. The left vertical arrow is the wishful thinking part. It should exist by analogy with the right.

I will conclude this section with some remarks about the relationship between coadjoint orbits and the unitary dual (Problem 2.6). First, one cannot expect that quantization of coadjoint orbits will produce *all* the irreducible unitary representations of G. Already for $SL(2,\mathbb{R})$ it does not produce the "complementary series" of representations discovered by Bargmann. Nevertheless, it does better—that is, it predicts more unitary representations—than any other general approach that I know. (Next most effective are Arthur's conjectures, based on Langlands' philosophy and automorphic representation theory.)

Second, quantization will be possible only under an appropriate (and subtle) integrality constraint on the coadjoint orbit. This point is not yet fully understood; one can find information in [Du] and [AV]. When the integrality constraint can be satisfied at all, it can often be satisfied in several ways. What is called $\pi(M)$ in Problem 4.5 should therefore be understood as a small (usually finite) family of unitary representations.

5. Three kinds of coadjoint orbit and two kinds of quantization

In order to proceed further we need to be a little more precise about the notion of reductive. Here is a convenient definition, taken essentially from [Kn].

Definition 5.1. Write $GL(n)$ for the group of real or complex $n \times n$ matrices. The *Cartan involution* of $GL(n)$ is the automorphism conjugate transpose inverse:

$$\theta(g) = {}^t\overline{g}^{-1}.$$

A *linear reductive group* is a closed subgroup G_1 of some $GL(n)$, preserved by θ and having finitely many connected components. (Because of the natural inclusions

$$GL(n, \mathbb{R}) \subset GL(n, \mathbb{C}) \subset GL(2n, \mathbb{R}),$$

we may always arrange if necessary that the matrices appearing be real or that they be complex. For the construction of the compact dual symmetric space in the introduction, complex matrices were needed. For definition 5.2 below, real matrices are convenient.) A *reductive group* is a Lie group $\widetilde{G}$ endowed with a homomorphism $\pi: G \to G_1$ onto a linear reductive group, so that the kernel of π is finite.

Henceforth G will always be a reductive Lie group in this sense. The definition allows connected semisimple Lie groups with finite center, compact Lie groups, and real points of reductive algebraic groups; it also has convenient heriditary properties (in that many nice subgroups of a reductive group are automatically reductive).

Definition 5.2. Suppose G is a real reductive Lie group (Definition 5.1). The map

$$G \to GL(n, \mathbb{R})$$

gives rise to an inclusion

$$\mathfrak{g}_0 \subset \mathfrak{gl}(n, \mathbb{R}) = n \times n \text{ real matrices.}$$

The vector space $\mathfrak{gl}(n, \mathbb{R})$ is naturally isomorphic to its dual by means of the bilinear form

$$\langle X, Y \rangle = \operatorname{tr} XY.$$

This form is non-degenerate on the subspace $\mathfrak{g}_0$, defining an isomorphism $\mathfrak{g}_0^* \simeq \mathfrak{g}_0$ and therefore an inclusion

$$\mathfrak{g}_0^* \subset n \times n \text{ real matrices.}$$

An element $\lambda \in \mathfrak{g}_0^*$ is called *hyperbolic* if the corresponding matrix is diagonalizable; *elliptic* if it is diagonalizable over $\mathbb{C}$, with purely imaginary eigenvalues; and *nilpotent* if the corresponding matrix is nilpotent.

A basic problem about matrices is "normal form": to classify them up to the conjugation action of $GL(n, \mathbb{R})$. It can be solved very explicitly using the Jordan decomposition and these three special types. In the same way, general coadjoint orbits for G (and therefore homogeneous Hamiltonian G-spaces) can be classified. The answer (known by the 1960s, and described in some detail in [VP]) is not much more complicated than the theory of Jordan normal form for real matrices (which is just the special case $G = GL(n, \mathbb{R})$.)

We turn therefore to Problem 4.5, attaching a unitary representation $\pi(M)$ to a coadjoint orbit $M = G \cdot \lambda \simeq G/G^\lambda$. This can also be reduced to the three special cases described in Definition 5.2. Here is a summary of what is known. Recall that, as a homogeneous Hamiltonian G-space, M is in particular a symplectic manifold; so the dimension of M is an even integer $2m$. A submanifold N is called *coisotropic* if the ideal $J(N)$ of functions on M vanishing on N is closed under Poisson bracket. This condition forces $J(N)$ to be fairly small, and therefore forces N to be fairly large; in particular, it forces $\dim N \geq m$. We say that N is *Lagrangian* if it is coisotropic, and $\dim N = m$.

Theorem 5.3. *Suppose G is a real reductive Lie group, and $\lambda \in \mathfrak{g}_0^*$ is hyperbolic. Write $M = G \cdot \lambda$. Then there is a G-equivariant fibration $M \to Z$ with Lagrangian fibers and Z compact. Attached to M is a natural finite set of G-equivariant Hermitian vector bundles on Z; we can attach to M the unitary representations of G on spaces of L^2 sections of these Hermitian vector bundles on Z.*

The isotropy group for the action of G on Z is a real parabolic subgroup of G. These unitary representations (which are a slightly special case of *degenerate principal series representations*) were introduced by Gelfand-Naimark and others by the early 1950s.

Theorem 5.4. *Suppose G is a real reductive Lie group, and $\lambda \in \mathfrak{g}_0^*$ is elliptic. Then there is a G-equivariant complex structure on $M = G \cdot \lambda$ making M an (indefinite) Kähler manifold. Under an integrality constraint on λ, we can construct a finite set of G-equivariant holomorphic vector bundles on M; and finally attach to M unitary representations of G realized in the Dolbeault cohomology of M with coefficients in these vector bundles.*

When G is compact, these homogeneous Kähler manifolds are projective algebraic varieties. The corresponding representations were understood in the 1950s (the Borel-Weil-Bott theorem). At the same time Harish-Chandra studied the case when M is Stein, which happens essentially only when G/K is a Hermitian symmetric space. In this way he was able to construct the "holomorphic discrete series" representations, realized in spaces of L^2 holomorphic sections of vector bundles on G/K. Whenever the isotropy group G^λ is compact, one expects to realize discrete series representations of G on $G \cdot \lambda$. Langlands made a precise conjecture along these lines in 1965, and this conjecture (which includes much of Theorem 5.4) was proved by Schmid in [Sc]. What remains are the singular elliptic coadjoint orbits. The necessary representations were constructed by Zuckerman in 1978, and their unitarity was established in [Vu]. Zuckerman's construction is sometimes called *cohomological parabolic induction*, and the representations are unfortunately often referred to as $A_{\mathfrak{q}}(\lambda)$ *modules.* (The term *elliptic representations* is more elegant and descriptive, but there seems to be little hope of popularizing it now.)

Theorem 5.5. *Suppose G is a real reductive Lie group, and $\lambda \in \mathfrak{g}_0^*$ is nilpotent. Then $M = G \cdot \lambda$ is a cone (closed under positive dilations in $\mathfrak{g}_0^*$).*

We have no general method to attach unitary representations (to be called *unipotent representations*) to a nilpotent coadjoint orbit. This is really all that is left of Problem 4.5: the constructions of representations underlying Theorems 5.3 and 5.4 are sufficiently flexible to solve Problem 4.5 in general once we know how to treat the nilpotent case. For the rest of the paper we concentrate on that.

6. Quantizing nilpotent orbits: motivation

We continue to assume G is a real reductive Lie group. Define

$$\mathcal{N}_{\mathbb{R}}^* = \text{cone of nilpotent elements in } \mathfrak{g}_0^*. \qquad (6.1)(a)$$

This is a finite union of orbits of G; we want to attach unitary representations to these orbits. There are two guiding principles.

The first principle is "compatibility with restriction to K." G has a maximal compact subgroup K, and we understand the unitary representations of K very well. In particular, we understand fairly well the relationship between the "classical" and "quantum" notions of representation theory. The main feature is that representations are constructed from classical objects using invariant complex structures. The basic example of this is the Borel-Weil construction of irreducible representations of K, mentioned after Theorem 5.4. Such representations extend naturally to the complexification $K_{\mathbb{C}}$ of K, which is a complex reductive algebraic group.

This principle suggests that we might try to replace the cone $\mathcal{N}_{\mathbb{R}}^*$ with a complex algebraic variety $\mathcal{N}_\theta^*$ carrying an algebraic $K_{\mathbb{C}}$ action, in such a way that the actions of K on $\mathcal{N}_{\mathbb{R}}^*$ and $\mathcal{N}_\theta^*$ are (more or less) equivalent. Here is a way to do that. Define

$$\mathfrak{g}^* = \mathrm{Hom}_{\mathbb{R}}(\mathfrak{g}_0, \mathbb{C}), \qquad (6.1)(b)$$

$$\mathcal{N}_\theta^* = \{\lambda \in \mathfrak{g}^* \mid \lambda\,|_{\mathfrak{k}} = 0 \text{ and } \lambda \text{ is nilpotent}\}. \qquad (6.1)(c)$$

Here are the basic facts about this new cone.

(1) (Kostant-Rallis) The cone $\mathcal{N}_\theta^*$ is a complex algebraic variety on which $K_{\mathbb{C}}$ acts with finitely many orbits.
(2) (Sekiguchi) These orbits are in one-to-one correspondence with the orbits of G on $\mathcal{N}_{\mathbb{R}}^*$.
(3) (Vergne) Corresponding orbits are K-equivariantly diffeomorphic.

The conclusion we want to draw from the first principle is that representations attached to nilpotent orbits should be realized (as representations of K) as (global sections of) $K_{\mathbb{C}}$-equivariant sheaves of modules on $\mathcal{N}_\theta^*$.

The second principle is "compatibility with classical limit." So far we have considered only the problem of constructing a quantum system from a classical one.

The mathematical models of quantum mechanics typically contain Planck's constant as a parameter, and the physical notion of "classical limit" is letting Planck's constant tend to zero. The usual effect of this process mathematically is to make operators more commutative. A fundamental example is the symbol calculus for differential operators on a manifold X, which relates the noncommutative algebra $\mathcal{D}(X)$ to the commutative algebra $C^\infty(T^*(X))$.

A corresponding idea in the setting of group representations is to make use of the Poincaré-Birkhoff-Witt isomorphism

$$\operatorname{gr} U(\mathfrak{g}) \simeq S(\mathfrak{g}). \tag{6.2)(a}$$

That is, we try to replace representations (modules for the noncommutative algebra $U(\mathfrak{g})$) by modules for the polynomial ring $S(\mathfrak{g})$.

Here is a construction. Suppose V is an irreducible Harish-Chandra module for G ([Kn], section 10.9). Choose a K-invariant good filtration

$$V_0 \subset V_1 \subset \cdots \qquad \bigcup_n V_n = V \qquad U_p(\mathfrak{g}) \cdot V_q \subset V_{p+q}. \tag{6.2)(b}$$

Then $\operatorname{gr} V$ is a finitely generated $K_{\mathbb{C}}$-equivariant $S(\mathfrak{g}/\mathfrak{k})$-module supported on $\mathcal{N}_\theta^*$. (A more detailed discussion of this construction may be found in [AV].) This module is a natural "classical limit" of V, and so its support (a closed union of $K_{\mathbb{C}}$ orbits on $\mathcal{N}_\theta^*$) is a natural candidate for a classical analogue of V.

The conclusion we want to draw from the second principle is that representations V attached to nilpotent orbits should have $\operatorname{gr} V$ an uncomplicated module with specified support in $\mathcal{N}_\theta^*$. (Some additional suggestions about what is meant by "uncomplicated" may be found in [AV].)

Problem 4.5 for a nilpotent orbit $G \cdot \lambda$ now looks like this: we seek a unitary representation of G so that the corresponding Harish-Chandra module V makes $\operatorname{gr} V$ an uncomplicated module with specified support (namely the closure of the $K_{\mathbb{C}}$ orbit on $\mathcal{N}_\theta^*$ corresponding to $G \cdot \lambda$). The Kazhdan-Lusztig conjectures (which are proved for linear groups) tell us how to construct many Harish-Chandra modules V' with specified support. There are two (related) difficulties: $\operatorname{gr} V'$ may be complicated, and V' may not be unitary.

One possible manifestation of the first difficulty is that the multiplicity of $\operatorname{gr} V'$ (along the generic part of its support) may be bigger than 1. (Any finitely generated module looks locally like several copies of the algebra of functions along an open set in its support. The "multiplicity" is equal to the number of copies. One way of saying that the module is uncomplicated is to require the generic multiplicity to be one.)

A manifestation of the second difficulty is that V' may carry an invariant Hermitian form with indefinite signature.

In the next section I will explain how to prove (sometimes) that the second difficulty forces the first to occur: that is, that an indefinite invariant Hermitian form on V can exist only if the generic multiplicity of $\operatorname{gr} V'$ is greater than one. One special case is easy. If the nilpotent orbit we are considering is the point $\{0\}$, then the condition that the support of $\operatorname{gr} V'$ is $\{0\}$ is equivalent to requiring

$\dim V' < \infty$. In this case the generic multiplicity of V' is equal to the dimension of V'. It is certainly true that V' can carry an indefinite Hermitian form only if $\dim V' > 1$; that is, if and only if the generic multiplicity is greater than 1. The case of larger nilpotent orbits is more subtle, of course, but similar in spirit.

We may therefore hope to approach Problem 4.5 in the following way. Given a nilpotent orbit, we use ideas from the Kazhdan-Lusztig conjectures to find a Harish-Chandra module V whose support is the closure of the corresponding $K_{\mathbb{C}}$ orbit. The problem of calculating the generic multiplicity of $\operatorname{gr} V$ is sometimes tractable, although there is no general algorithm for solving it; so we can hope to show that the generic multiplicity of $\operatorname{gr} V$ is 1. The argument to be explained in the next section may then allow us to deduce that V comes from a unitary representation.

7. Quantizing nilpotent orbits: signatures and unitarity

The argument given at the end of the last section was that an indefinite Hermitian form can be defined only on a vector space of dimension at least two. The reason is that if the signature of the form is (p, q), then the dimension of the space is $p + q$, and the form is indefinite if and only if p and q are both non-zero. To extend this to infinite-dimensional Harish-Chandra modules, we need some control over their signatures. This is provided by the following theorem, in which the modules M^+ and M^- are like the integers p and q. Statement (1) corresponds to the dimension being $p + q$; statement (2) is a consequence; statement (3) is the indefiniteness criterion; and statement (4) is a technical condition that will allow us to control $V^{\pm}$.

Theorem 7.1 ([Vu]). *Suppose V is an irreducible Harish-Chandra module for a real reductive Lie group G admitting a non-degenerate invariant Hermitian form, and supported on a $K_{\mathbb{C}}$-orbit closure $X \subset \mathcal{N}^*_\theta$. Then there are $K_{\mathbb{C}}$-equivariant finitely generated graded $S(\mathfrak{g}/\mathfrak{k})$-modules $M^{\pm}$ supported on X, with the following properties.*

1. *We have $\operatorname{gr} V = M^+ + M^-$ in the Grothendieck group of $K_{\mathbb{C}}$-equivariant finitely generated graded $S(\mathfrak{g}/\mathfrak{k})$-modules supported on $\mathcal{N}^*_\theta$.*
2. *The generic multiplicity of $\operatorname{gr} V$ is equal to the sum of the generic multiplicities of M^+ and M^-.*
3. *The modules $M^{\pm}$ are both non-zero if and only if the form on V is indefinite.*
4. *There are virtual Harish-Chandra modules $V^{\pm}$ satisfying*
   - *(a) $\operatorname{gr} V^{\pm} = M^{\pm}$ in the Grothendieck group; and*
   - *(b) the sizes of infinitesimal characters satisfy*

$$|\textit{infinitesimal character}(V^{\pm})| \leq |\textit{infinitesimal character}(V)|.$$

We conclude from this theorem that if the generic multiplicity of $\operatorname{gr} V$ is 1, then one of the two modules $M^{\pm}$ must have generic multiplicity zero; that is, must be supported on the boundary of the orbit closure X. Possibly replacing the Hermitian form by its negative (which interchanges M^+ and M^-), we can arrange for $\operatorname{gr} V^-$

to have this smaller support. We also know (from (4)) an upper bound on the infinitesimal character of V^-. We are aiming (because of (3)) to show that V^- must be zero. In order to do that, we need results like

Desideratum 7.2. *Suppose W is a virtual Harish-Chandra module, and* $\operatorname{gr} W$ *is a non-zero virtual module with small support. Then the infinitesimal character of W is large. More precisely, suppose that the support of* $\operatorname{gr} W$ *is contained in the boundary of the nilpotent orbit closure $X \subset \mathcal{N}_\theta^*$. Then there is a constant c_X with*

$$|\textit{infinitesimal character}(W)| \geq c_X.$$

Techniques introduced by McGovern in [Mc] make it possible to prove results like this, with very explicit constants c_X. Some additional discussion may be found in [VP]. The calculations have so far been made only in quite special cases, but here is an example.

Theorem 7.3. *Suppose G is a reductive Lie group, and $X \subset \mathcal{N}_\theta^*$ is the closure of a minimal $K_{\mathbb{C}}$ orbit. (Therefore the boundary of X is the point $\{0\}$.) Suppose W is a virtual Harish-Chandra module with* $\operatorname{gr} W$ *finite-dimensional and non-zero. Then the infinitesimal character of W is at least as large as that of the trivial representation. That is, Desideratum 7.2 holds, with c_X the length of ρ (half the sum of a set of positive roots).*

Let us summarize the reasoning to this point. We begin with a $K_{\mathbb{C}}$-orbit closure $X \subset \mathcal{N}_\theta^*$. Suppose that Desideratum 7.2 has been established for X. Suppose finally that V is an irreducible Harish-Chandra module carrying a Hermitian form, and satisfying

(1) the support of $\operatorname{gr} V$ is equal to X;
(2) the generic multiplicity of $\operatorname{gr} V$ is 1; and
(3) the length of the infinitesimal character of V is strictly smaller than c_X.

Then the Hermitian form on V is definite; so V comes from a unitary representation of G. It is by this argument that we hope to solve Problem 4.5 for nilpotent orbits. Here is an example.

Corollary 7.4. *Suppose G is a reductive Lie group and V is a Hermitian ladder representation of infinitesimal character strictly shorter than half the sum of a set of positive roots. Then V is unitary.*

Sketch of proof. The "ladder" condition is that the highest weights of the K-types of V all have multiplicity one, and form a string with separation equal to a highest weight of K on $\mathfrak{g}/\mathfrak{k}$. It implies that the support of $\operatorname{gr} V$ is a minimal orbit closure, and that the generic multiplicity is one. We can therefore apply Theorems 7.1 and 7.3 to finish the argument. $\square$

Techniques for constructing Harish-Chandra modules V satisfying the conditions in Corollary 7.4 have been available for a long time (and the unitarity of most of these representations was known as well). If for example G is a complex simple group, then the quotient of the enveloping algebra by the Joseph ideal has the required properties. In this way one can attach unitary representations to minimal nilpotent orbits whenever they satisfy the "integrality" constraint of Problem 4.5.

(A completely different general approach to Problem 4.5 for minimal orbits may be found in [To]. Torasso's method requires even less case-by-case analysis than the one using Corollary 7.4, and applies also to reductive groups over other local fields. It is not clear how to extend it to general nilpotent orbits, however.)

For more general nilpotent orbits, making Desideratum 7.2 explicit is more difficult, but should not be impossible. I have already claimed that the Kazhdan-Lusztig conjectures can help exhibit representations satisfying condition (1) on the list above; explicit information about the infinitesimal characters is easy to get, so (3) should not present difficulties. (These two steps are treated for complex groups in [BV].) Condition (2) is the most difficult, but the same calculations involved in Desideratum 7.2 can help.

I will conclude with an example: the fourteen-dimensional simply connected split real group of type G_2. The calculations have been done hastily, and the details should not be trusted too far.

The real nilpotent orbits are of dimensions 0, 6, 8, 10, and 12. There is a unique orbit of each dimension, except that there are two orbits of dimension 10. Correspondingly the orbits of $K_{\mathbb{C}}$ on $\mathcal{N}_\theta^*$ are of complex dimensions 0, 3, 4, 5, and 6. We can realize the root system on the real Euclidean space

$$E = \{(x, y, z) \in \mathbb{R}^3 \mid x + y + z = 0\};$$

the roots are

$$\{e_i - e_j \mid i \neq j\} \cup \{\pm(2e_i - e_j - e_k) \mid \{i, j, k\} = \{1, 2, 3\}\}.$$

The Weyl group is generated by permutation of the coordinates and $-I$; it has order 12. An infinitesimal character is represented by an element of E, defined up to the action of the Weyl group; its length is just its length as a vector in E.

In Table 7.5 we attach several infinitesimal characters to each nilpotent orbit of $K_{\mathbb{C}}$ on $\mathcal{N}_\theta^*$. Since these are the same for the two five-dimensional orbits, we label the rows of the table only by the dimension of the orbit. The first infinitesimal character (labelled "lower bound") is a lower bound (in length) for the infinitesimal character of a virtual Harish-Chandra module V with $\operatorname{gr} V$ supported on this orbit closure. To compute the number c_X of Desideratum 7.2, we need to find the maximum length for these infinitesimal characters on the boundary of X. That is, c_X is the length of the "lower bound" infinitesimal character for the row above X. (In the case of the orbit $\{0\}$ the boundary is empty, so nothing can be supported there; so $c_X = \infty$.)

The next family of infinitesimal characters are those of unipotent representations attached to the corresponding real orbits in the sense of Problem 4.5. Most of these unitary representations are discussed in [V2]. Except for the orbits of dimensions 3 and 4, they are all Arthur's "special unipotent representations." One can show (in increasing order of difficulty) that they are Hermitian, that they have $\operatorname{gr} V$ supported on the indicated orbit closure, and that they have generic multiplicity one. In order to prove that they are unitary by the argument outlined before Corollary 7.4, we therefore just need to check that the infinitesimal characters are

$K_{\mathbb{C}}$-orbit dimension	infinitesimal character lower bound	infinitesimal characters of unipotent representations
0	$(3,-1,-2)$	$(3,-1,-2)$
3	$(\frac{3}{2},0,-\frac{3}{2})$	$(\frac{5}{3},-\frac{1}{3},-\frac{4}{3})$
4	$(1,\frac{1}{2},-\frac{3}{2})$	$(1,\frac{1}{2},-\frac{3}{2})$
5	$(\frac{1}{2},0,-\frac{1}{2})$	$(\frac{1}{2},0,-\frac{1}{2}),(1,-\frac{1}{2},-\frac{1}{2}),(1,0,-1)$
6	$(0,0,0)$	$(0,0,0)$

TABLE 7.5. Unipotent representations for split G_2

strictly smaller than the "lower bound" infinitesimal character on the previous line. This is obviously true, so the representations are all unitary (as was already known).

There is a final subtlety about Table 7.5 worth mentioning here. The "lower bound" column refers to infinitesimal characters of virtual Harish-Chandra modules supported on the orbit. The "unipotent representations" column refers to infinitesimal characters of actual Harish-Chandra modules supported on the orbit. It seems natural to expect that these two columns should be the same, or at least that the second column should be the smallest element of the third. This is usually the case, but not always. What happens here for the three-dimensional orbit is that one can can find a small collection of representations of infinitesimal characters $(\frac{3}{2},0,-\frac{3}{2})$ and smaller, mostly supported on the five-dimensional orbit. Taking a certain integer combination of these representations kills the restrictions to the four- and five-dimensional orbits, leaving just the three-dimensional one. This cancellation appears only on the level of associated graded modules; it can't be achieved for the group representations. (The Kazhdan-Lusztig conjectures control which nilpotent orbit closures can appear as supports of representations with a specified infinitesimal character. In this case they say that the three-dimensional orbit can appear only at a an infinitesimal character whose coordinates are distinct numbers all congruent to $\pm\frac{1}{3}$ modulo $\mathbb{Z}$.)

REFERENCES

[Ar] V. Arnold, *Mathematical Methods of Classical Mechanics*, Springer-Verlag, New York-Heidelberg-Berlin, 1978.

[BV] D. Barbasch and D. Vogan, *Unipotent representations of complex semisimple Lie groups*, Ann. of Math. **121** (1985), 41–110.

[BW] A. Borel and N. Wallach, *Continuous Cohomology, Discrete Subgroups, and Representations of Reductive Groups*, Princeton University Press, Princeton, New Jersey, 1980.

[Dix] J. Dixmier, *Les C^*-algèbres et leurs Représentations*, Gauthier-Villars, Paris, 1964.

[Du] M. Duflo, *Théorie de Mackey pour les groupes de Lie algébriques*, Acta Math. **149** (1982), 153–213.

[Gu] A. Guichardet, *Cohomologie des Groupes Topologiques et des Algèbres de Lie*, CEDIC, Paris, 1980.

[Kn] A. Knapp, *Representation Theory of Semisimple Groups: An Overview Based on Examples*, Princeton University Press, Princeton, New Jersey, 1986.

[K] B. Kostant, *Quantization and unitary representations*, Lectures in Modern Analysis and Applications (C. Taam, eds.), Lecture Notes in Mathematics, vol. 170, Springer-Verlag, Berlin-Heidelberg-New York, 1970.

[Ku] S. Kumaresan, *On the canonical $\mathfrak{k}$-types in the irreducible unitary $\mathfrak{g}$-modules with non-zero relative cohomology*, Invent. Math. **59** (1980), 1–11.

[MQ] G. Mackey, *Mathematical Foundations of Quantum Mechanics*, W. A. Benjamin, Inc., New York, 1963.

[Ma] G. Mackey, *Theory of Unitary Group Representations*, University of Chicago Press, Chicago, 1976.

[Mc] W. McGovern, *Rings of regular functions on nilpotent orbits and their covers*, Inv. Math. **97** (1989), 209–217.

[Sc] W. Schmid, *L_2 cohomology and the discrete series*, Ann. of Math. **103** (1976), 375–394.

[To] P. Torasso, *Méthode des orbites de Kirillov-Duflo et représentations minimales des groupes simples sur un corps local de caractéristique nulle*, Duke Math. J. **90** (1997), 261–377.

[Vu] D. Vogan, *Unitarizability of certain series of representations*, Ann. of Math. **120** (1984), 141–187.

[AV] D. Vogan, *Associated varieties and unipotent representations*, Harmonic Analysis on Reductive Groups (W. Barker and P. Sally, eds.), Birkhäuser, Boston-Basel-Berlin, 1991, pp. 315–388.

[V2] D. Vogan, *The unitary dual of G_2*, Invent. math. **116** (1994), 677–791.

[VP] D. Vogan, *The method of coadjoint orbits for real reductive groups*, Representation Theory of Lie Groups, IAS/Park City Mathematics Series, vol. 8, American Mathematical Society, Providence, RI, 1999.

[VZ] D. Vogan and G. Zuckerman, *Unitary representations with non-zero cohomology*, Compositio Math. **53** (1984), 51–90.

Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

E-mail address: dav@math.mit.edu