Proceedings of the
47th IEEE Conference on Decision and Control
Cancun, Mexico, Dec. 9-11, 2008

Optimal Control of Observable Continuous Time Markov Chains

Roger Brockett

Abstract— **This paper considers the optimal control of time varying, finite horizon, continuous time Markov chains under the assumption that their behavior can be influenced by the adjustment of selected transition rates. We assume a quadratic penalty on the amount of the rate adjustment and that the system is completely observable. We derive an ordinary differential equation whose solution gives the minimum return function and describe how the optimal feedback control law is obtained from it. The results bear some resemblance to the solution of the quadratic regulator problem for linear systems, but because of the bilinear structure of these problems, the details are significantly different.**

I. Introduction

From its beginnings as a identifiable subject in the late 1950's [1], [2] the control of Markov processes has enjoyed growing success and is, by now, supported by a large literature in areas as diverse as operations research, economics, financial engineering, artificial intelligence and, more recently, learning theory. The abundant interest notwithstanding, few explicitly solvable problems have been identified , most papers recommend some variation of dynamic programming and/or possible heuristics as the solution technique; few explicitly solvable problems have been identified. The purpose of this paper is to describe a broad class of such problems for which there is an effective optimization procedure involving only the solution of a differential equation with a specified initial condition, analogous to, but distinct from, the way the Riccati equation is used in optimal control.

The problems considered here can be characterized as

1) Finite or infinite horizon; time varying parameters
2) Arbitrary running cost on the state, quadratic on the controls
3) Finite or denumerably many states
4) Perfect observation of the state
5) Affine dependence of the transition rates on the controls

Although the basic results require neither time invariance nor infinite horizon, we do describe some simplifications which occur in these special cases. The development is carried out in a continuous time setting; there is an analogous development for the discrete time problems.

Essential elements of the development here include

1) A representation of sample paths for Markov processes using Poisson counters
2) A "basis vector" representation of the states of the Markov process
3) A linear functional representation of the minimum return function.

The reader is directed to the paper [4] where some of these ideas play a role and also to the very recent paper [5] where aspects of this formalism are used as part of a new design methodology for hybrid systems.

II. Preliminaries

The Model: Because we seek optimal feedback control laws for observed Markov processes, it is necessary to have a representation of the sample paths. Our choice is to consider the states to be the standard basis vectors in $\mathbb{R}^n$. This makes it possible to give a convenient description in terms of Itô equations involving Poisson counters. The sample path $x(t)$ takes on values in the finite set $e_1, e_2, ..., e_n$ and the model for the evolution is

$$dx = \sum_{i=1}^{m} G_i x dN_i \;\; ; \;\; G_i \in \hat{G}$$

where $N_1, N_2, ..., N_m$ are Poisson counters with rates λ_i and $\hat{G}$ is the set of matrices of the form $E_{k,l} - E_{l,l}$ where E_{ij} is the matrix of all zeros except for a one in the i^{th} row and j^{th} column. The resulting Itô equation generates a Markov process whose transition probabilities are related to the rates of the Poisson counters in accordance with

$$P(t) = \sum_{i=1}^{m} G_i \lambda_i(t)$$

The rates of the counters are allowed to depend on controls in accordance with

$$P(t, u(t)) = \sum_{i=1}^{m} G_i \left(\nu_i + \sum_{j=1}^{m} \mu_{ij} u_j(t) \right)$$

This will be abbreviated as

$$P(t, u(t)) = A(t) + \sum u_i B_i(t)$$

Let $\mathcal{G}$ denote the set of square matrices whose columns sum to zero and whose off diagonal entries are nonnegative. These are the so called infinitesimal generators of continuous time Markov processes. Because the counting rates must be nonnegative, A is necessarily in $\mathcal{G}$, and the B_i are matrices whose columns must sum to zero; the B_i but they are not necessarily infinitesimal generators. This is consistent with the idea that nonzero values of u may correspond to either increasing or decreasing the resources available when the system is in a specific state.

This work was supported in part by the Army Research Office,under the MURI award N0. N00014-01-1-0859. The views expressed in this paper are those of the author alone and not of the sponsor

Roger Brockett is with School of Engineering and Applied Sciences, Harvard University , Cambridge MA USA, 02138 brockett@seas.harvard.edu

978-1-4244-3124-3/08/$25.00 ©2008 IEEE

The Performance Measure: We assume that nonzero values of u have an associated cost and that each state of the process can be assigned a cost (or reward) that is expressible in terms of an integral over time plus an end point penality. More specifically, we assume that there is a cost function associated with each of the states such that

$$\eta = \mathcal{E}\int_0^T \sum_{i=1}^n c_i(t)x_i(t) + u^T(t)u(t)\;dt + \langle \phi_f, x(T)\rangle$$

measures the performance. Observe that if c is bounded then

$$\int_0^T c^T(t)x(t) + u^T u d\sigma \geq \inf_{i,t} c_i(t)T$$

and thus there are upper and lower bounds on the optimal value of η provided that T is finite.

The Admissible Controls: Defining $\mathcal{U}$: Because x only takes on values in the set $\{e_1, e_2, ..., e_n\}$ any real valued function of x, say $\phi(x)$, can be expressed as a linear functional, $\phi(x) = \langle\tilde{\phi}, x\rangle$. In fact, $\tilde{\phi}_i = \phi(e_i)$. We will use this representation repeatedly. Our optimization problem consists of finding a feedback control $u(t,x)$ that minimizes η. The controls are constrained by the fact that P must be an infinitesimal generator. To make this constraint explicit, let $\mathcal{U}$ denote the set of functions mapping the state space, $\{e_i\}_{i=1}^n$, to the space of controls, $\mathbb{R}^m$, such that the matrix with j^{th} column

$$f_j = Ae_j + \sum_{i=1}^m u_i(e_j)B_i e_j \;\; ; \;\; j = 1, 2, ..., n$$

is an infinitesimal generator. Such controls can be identified with a convex subset of the set of m by n matrices of the form

$$U = \begin{bmatrix} u_1(e_1) & u_1(e_2) & ... & u_1(e_n) \\ u_2(e_1) & u_2(e_2) & ... & u_2(e_n \\ ... & ... & ... & ... \\ u_m(e_1) & u_m(e_2) & ... & u_m(e_n) \end{bmatrix}$$

with the entries possibly time dependent. The set of such m by n matrices which result in P being an infinitesimal generator is a convex because the requirements on P are expressible as a a set of linear inequalities constraining its entries and P depends linearly on u. The set $\mathcal{U}$ can be characterized as the intersection of mn or fewer half spaces. It is not necessarily compact.

III. The Minimum Return Function

In Theorem 1 below the function

$$\phi(k,x) = \langle\phi(k), x\rangle = \min_{u(x)\in\mathcal{A}} \left(\sum_{i=1}^m u_i k^T Bx + u_i^2\right)$$

plays a role. If there were no constraints on u the minimum would be achieved at

$$u_i(x) = -\frac{1}{2}k^T B_i x$$

and the minimizing value would be

$$\phi^*(x) = -\frac{1}{4}\sum_{i=1}^m k^T B_i^T x x^T B_i k$$

In this situation it is clear that both the value of the minimum and the minimizing choice of u are continuous functions of k. When $-\frac{1}{2}k^T Bx$ lies outside the constraint set the minimizing feedback control $u(x)$ lies on the boundary of $\mathcal{U}$. Because of the $u^T u$ term , $\phi(k,x)$ is a strictly convex function of u. The minimizing value of u for fixed k need not be unique, but all local minima are global minima and because ϕ depends continuously on k the minimum value is continuous as a function of k. For values of k for which the minimizing u is not unique the strict convexity implies that there are at most a finite number of local minima and these are isolated. At any such value of u ϕ is a Lipschitz continuous function of k in a neighborhood and thus the choice that minimizes ϕ results in ϕ^* being a Lipschitz continuous function of k.

Theorem 1: Let G_i, and N_i be as described with the rates of the N_i being $\lambda_{i0} + \sum \mu_{ij}u_j$. The λ_{i0} and μ_{ij} may be time varying but are assumed to be bounded. Assume that x satisfies the Itô equation

$$dx = \sum_{i=1}^m G_i x dN_i \;\; ; \;\; , x(t) \in \{e_1, e_2, ..., e_n\}$$

Define A and B_i as

$$A = \sum_{i=1}^m G_i\lambda_{i0} \;\; ; \; B_i = \sum_{j=1}^m \mu_{ij}G_j$$

and let $\mathcal{U}$ be the constraint set defined above. Then for any $T > 0$ and any $\phi_f \in \mathbb{R}^n$ there exists a unique solution of the equation

$$\dot{k} = -A^T k - c + \min_{u(x)\in\mathcal{U}} \left(\sum_{i=1}^m u_i k^T B_i x + u_i^2\right) \;\; ; \; k(T) = \phi_f$$

on the interval $[0,T]$. Moreover, the control law

$$u(x) = \arg\min_{u(x)\in\mathcal{A}} \left(\sum_{i=1}^m u_i k^T B_i x + u_i^2\right)$$

minimizes

$$\eta = \mathcal{E}\int_0^T c^T(t)x(t) + u^T u d\sigma + \mathcal{E}\langle\phi, x(T)\rangle$$

relative to all other past measureable control laws and the minimizing value of η is

$$\eta^*(x(0)) = k^T(0)x(0)$$

Proof: Consider a candidate for the the minimum return function written as $\langle k(t), x(t)\rangle = k^T(t)x(t)$ and observe that the Itô differentiation rule gives

$$d\langle k, x\rangle = \langle\dot{k}, x\rangle dt + \sum_{i=1}^m \langle k, G_i x dN_i\rangle$$

Note that $dN_i - (\lambda_{i0} + \sum_{j=1}^m u_j\mu_{ji})dt$ is a martingale, and from the definitions of A and B_i, the expectation of

$$\sum_{j=1}^m G_i x dN_i - \left(Ax + \sum_{i=1}^m u_i B_i x\right) dt$$

vanishes. Thus we see that

$$\eta - \mathcal{E}k^T(t)x(t)\big|_{t=0}^{T} = \phi_f^T x(T) +$$
$$\mathcal{E}\int_0^T \dot{k}^T x + c^T x + k^T\left(Ax + \sum B_i x u_i\right) + u^T u dt$$

Introducing a minimization with respect to u we get the inequality

$$\eta - \mathcal{E}k^T(t)x(t)\big|_{t=0}^{T} - \phi^T x(T) \geq$$
$$\mathcal{E}\int_0^T \dot{k}^T x + k^T Ax + \min_{u(x)\in\mathcal{U}}\left(\sum_{i=1}^m u_i k^T Bx + u_i^2\right) dt$$

Now suppose that there exists a k satisfying the differential equation of the theorem statement with the boundary condition $k(T) = \phi_f$. In that case we see that

$$\eta \geq \mathcal{E}k^T(0)x(0)$$

and that equality can be obtained by letting $u(x)$ be the minimizing function of x identified in the theorem statement.

It remains only to show that there actually exists a solution to the differential equation for k. We have already argued for the Lipschitz continuity of the right-hand side of this equation. Thus there is local existence and uniqueness theorem for small values of $|t - T|$, $t < T$. Moreover, we have from the argument above

$$k^T(t)x = \min_{u\in\mathcal{U}}\mathcal{E}\int_t^T c^T x + u^2 dt$$

As noted, the right-hand side has at most linear growth in $|t - T|$ so we see that for $t \leq T$, $k(t)$ must be bounded and hence we have existence and uniqueness for all $t \leq T$.

This completes the proof.

Notation: If x is a vector in $\mathbb{R}^n$ then the vector in $\mathbb{R}^n$ whose entries are the squares of the corresponding entries in x will be written as $x^{\cdot 2}$.

Corollary: Under the hypothesis of theorem one, if for

$$\dot{k} = -A^T k - c + \frac{1}{4}\sum\left(B_i^T k\right)^{\cdot 2} \quad ; \quad k(T) = \phi_f$$

the feedback control

$$u_i(x) = -\frac{1}{2}\sum k^T B_i x$$

lies in $\mathcal{U}$ then it is the optimal control.

The optimal control problem defined and solved by this theorem requires the choice of a time dependent m by n matrix $U(t) \in \mathcal{U}$. We now describe a deterministic problem that is equivalent to the problem described in Theorem one in the sense that there is a simple mapping between its solution and the solution given by Theorem 1. This formulation provides an alternate way to conceptualize the situation and, because it is deterministic, can be treated as a open loop optimization problem.

Notation: Let e be the vector whose components are all ones,

$$e = \sum_{i=1}^n e_i$$

Theorem 2: Let A, B_i, c be as in Theorem 1 and let p satisfy the equation

$$\dot{p} = \left(A + \sum_{i=1}^m B_i D_i\right) p \quad ; \quad p(0) = \text{given}$$

with $p(0)$ a probability vector and the D_i arbitrary, time dependent, diagonal matrices, considered as controls. The minimization of

$$\eta = \int_0^T c^T p + \sum_{i=1}^m e^T D_i^2 p dt + \phi_f^T p(T)$$

subject to the constraint that

$$A + \sum_{i=1}^m B_i D_i \in \mathcal{G}$$

results in a choice for the j^{th} element of D_i which equals the optimal choice of $u_i(e_j)$ in Theorem 1.

Proof: Because there are m diagonal matrices, each with n potentially nonzero entries, there are mn controls to be determined. We use the maximum principle to find first order necessary conditions. Introduce the Hamiltonian with costate variable q (we assume normality)

$$h(p,q,D) = q^T\left(A + \sum_{i=1}^m B_i D_i)\right) p + c^T p + \sum e^T D_i^2 p$$

so that

$$D_i = \arg\min_{A+\sum B_i D_i\in\mathcal{G}} \sum q^T B_i D_i p + \sum e^T D_i^2 p$$

The costate equation is then

$$\dot{q} = -A^T q - c + \min_{A+\sum B_i D_i\in\mathcal{G}} \sum_{i=1}^m q^T(B_i D_i p + D_i^2 e)$$

In comparing this with the equation for k in theorem 1,

$$\dot{k} = -A^T k - c + \min_{u(x)\in\mathcal{U}}\left(\sum_{i=1}^m u_i k^T B_i x + u_i^2\right)$$

the important point to note is that the minimization of

$$\min_{A+\sum B_i D_i\in\mathcal{G}} \sum_{i=1}^m qB_i D_i p + e^T\sum D_i^2 p = \sum_{i=1}^n \gamma_i p_i$$

decouples in the sense that it can be done by considering one column of $A + \sum B_i D_i$ at a time. Moreover, the p_i are necessarily nonnegative and so the minimization is achieved by minimizing $q^T BD_i e_i$. However, this is what is required for the minimization of

$$\min_{u(x)\in\mathcal{U}}\left(\sum_{i=1}^m u_i k^T B_i x + u_i^2\right)$$

By observing that the constraints and the value of the minimum are preserved if we let

$$d_{ij} = u_i(e_j)$$

with d_{ij} being the j^{th} diagonal of D_i. Thus q and k are the same and we see that the costate is not just the local gradient of the minimum return function but represents it over the entire interval $[0,T]$.

This completes the proof.

Remark 1: Because the $\mathcal{E}x(t) = p$ and because $e^Tp = 1$, we see that the effect of changing c in the performance measure to $c+\beta e$ and simultaneously changing ϕ_f to $\phi_f + \beta e$ is simply to add βT to the minimum return function. Such a change has no effect the optimal control law because $B^Te = 0$. Consequently, it is only the projection of c onto the orthogonal complement of e, i.e., the vector $c-(c^Te/n)e$ that plays a role in determining the optimal control.

Remark 2: If $A - \frac{1}{2}\sum u_iB_i$ is an infinitesimal generator for

$$u_i(t,x) = -\frac{1}{2}k^(t)TB(t)x(t)$$

and

$$\dot{k} = -A^Tk - c - \frac{1}{4}\sum(B_i^Tk)^{\cdot 2} \ \ ; \ \ k(t_f) = \phi_f$$

then u is the optimal solution. For fixed values of A and B_i, this will be the case if the entries in A are strictly positive off the diagonal and if $c-(c^Te/n)e$ is sufficiently close to zero. It will also be the case if B is an infinitesimal generator and $-B^Tk$ is element-by-element nonnegative.

Remark 3: The quadratic programming problem consisting of minimizing $\eta = a+b^Tu+u^Tu$ subject to $u \in U$ with U defined by the intersection of half-planes is quite standard. The minimum always exists and occurs either at $u = -\frac{1}{2}b$, if this point is in the interior of U, or else on the boundary of U. If u is a scalar constrained by $g_l \le u \le g_u$, the minimum is

$$\eta_{\min}(u) = \begin{cases} a + g_lb + g_l^2 & -\frac{1}{2}b \le g_l \\ a - \frac{1}{4}b^2 & g_l \le -\frac{1}{2}b \le g_u \\ a + g_ub + g_ug_u & g_u \le -\frac{1}{2}b \end{cases}$$

In the vector case the minimum may occur in the interior of U, on a hyperplane defined by $h^Tu = r$, at the intersection of two such hyperplanes, etc. If u is in the interior then $\eta = a - b^Tb/4$. In case u lies on a hyperplane $h^Tu = r$ we may always suppose that $h^Th = 1$. Using a Lagrange multiplier we are led to

$$\eta_\lambda = a + b^Tu + u^Tu + \lambda(h^Tu - r)$$

whose minimum occurs at

$$u = -\frac{1}{2}(b - \lambda h)$$

Enforcing $h^Tu = r$ gives an equation for λ

$$h^T(b - \lambda h) = -2r$$

Thus $\lambda = 2r + h^Tb$ and the minimizing value of u is given by

$$u^* = -\frac{1}{2}\left(b - (2r + h^Tb)h\right)$$

The minimum value of η is then

$$\eta^* = a - \frac{1}{4}b^Tb + r^2 + rh^Tb/2$$

Further details are left to the reader.

IV. The Constant Coefficient Cases

If we assume that the vector c and the matrices A and B_i entering in Lemma 1 are time invariant, some simplification occurs. In this case the problem becomes invariant with respect to a translation in time, and we may expect that the gains defining the optimal control will approach a constant as T goes to infinity. We will, in fact, show that under appropriate assumptions there exists a time invariant optimal control policy even though k is not bounded on $[0,\infty)$.

Lemma 1: Under the assumptions of Theorem 3, and with A, B, and c constant, suppose that for

$$\dot{p} = (A + \sum_{i=1}^m B_iD_i)p$$

there exists a finite interval $[0,T]$ and a time dependent choice of D_i on that interval such that $A+\sum B_iD_i$ is an infinitesimal generator that transfers p_1 to p_2. Also suppose there exists a control with these properties that transfers p_2 to p_1. Then there is a constant m such that

$$||k^T(t)(p_1 - p_2)|| \le m$$

for all probability vectors p_1, p_2 and all $t > 0$.

Proof: Without loss of generality we can consider an infinite horizon problem. For $t > T$ the minimum cost from p_2 can be expressed as

$$\int_0^T c^Tx(t) + \sum e^TD_i^2pdt + \int_T^\infty c^Tx(t) + \sum e^TD_i^2pdt$$

The minimum cost from p_1 can not exceed the minimum cost from p_2 by more than the cost of getting from p_1 to p_2 on the interval $[0,T]$ and this is finite. Reversing the roles of p_1 and p_2 completes the proof.

In the next section we will consider the controllability question in slightly more detail, but notice that if the set of points $(a-b)$ with the property that a can be steered to b and b can be steered to a has in its linear span all vectors whose entries sum to zero then $k - e(e^Tk/n$ is necessarily bounded.

Lemma 2: If A is an irreducible infinitesimal generator then the bordered matrix

$$M = \begin{bmatrix} A^T & -e \\ e^T & 0 \end{bmatrix}$$

is invertible and for $||c - e(e^Tc/n)||$ sufficiently small there is a real solution (α,β) to the pair of equations

$$\alpha e = A^T\beta + c - \frac{1}{4}\left(\sum_{i=1}^m B_i^T\beta)^{\cdot 2}\right)$$

$$e^T\beta = 0$$

Proof: To prove the statement on M we show that there is no nonzero vector in the null space of M^T. Suppose that $(g^T,h)^T$ is in the null space of M^T so that $Ag+eh = 0$. In that case $e^TAg + nh = 0$. But because A is an infinitesimal generator, $e^TA = 0$. and so $h = 0$. Because A is irreducible its null space is one dimensional and consists of multiples of some probability vector. However, $e^T p \neq 0$ for any vector whose components sum to something nonzero. Thus, 0 is the only element of the null space of M^T and M it is invertible. Rewrite the given equations as

$$f(\alpha,\beta) = M \begin{bmatrix} \beta \\ \alpha \end{bmatrix} - \frac{1}{4} \begin{bmatrix} \sum_{i=1}^{m} B_i^T \beta)^{\cdot 2} \\ 0 \end{bmatrix} = \begin{bmatrix} -c \\ 0 \end{bmatrix}$$

We can express the Jacobian of the left-hand side with the help of the following notation. Introduce the diagonal matrix D defined as

$$D_i(\beta) = \begin{bmatrix} e_1^T B_i^T \beta & 0 & ... & 0 \\ 0 & e_2^T B_i^T \beta & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & e_n^T B_i^T \beta \end{bmatrix}$$

The Jacobian of f evaluated at (α,β) is

$$J = \begin{bmatrix} A^T & -e \\ e^T & 0 \end{bmatrix} - \frac{1}{2} \begin{bmatrix} \sum_{i=1}^{m} D_i(k) B_i & 0 \\ 0 & 0 \end{bmatrix}$$

If $c = \beta e$ then the equation for $[\beta,\alpha]$ has $[\beta,0]^T$ as a solution and the Jacobian of the left-hand side, when evaluated at this solution, is just the invertible matrix M. Thus the inverse function theorem implies that there will be a solution for all $c - e(e^T c/n)$ in a neighborhood of 0.

Remark 4: This lemma shows that in the time invariant situation, and with mild additional assumptions, the time reversed differential equation for k admits a solution of the form

$$k(t) = \alpha e t + \beta$$

This solution can be used to define the optimal solution to an infinite horizon version of the problem

$$\eta = \mathcal{E} \int_0^T c^T x(t) + u^T u dt$$

The following theorem summarizes the situation.

Theorem 3: With A, B_i, c, k_f as in Theorem 1, but now constant, assume that there is a solution (α,β) of

$$\alpha e = A^T \beta + c - \sum_{i=1}^{m} \frac{1}{4} \left(B_i^T \beta\right)^{\cdot 2}$$

such that

$$A - \frac{1}{2} \sum_{i=1}^{m} B_i \text{ diag } (\mathrm{B_i^T} \beta) \in \mathcal{G}$$

Then the control law $u_i(x) = -\frac{1}{2}\beta^T B_i x$ gives

$$\lim_{T\to\infty} \frac{1}{T} \mathcal{E} \int_0^\infty c^T x + u^T u dt = \alpha/\mathrm{dim\ x}$$

and no other feedback control improves on this asymptotic rate. Moreover,

$$\eta_1 = \mathcal{E} \int_0^\infty (c^T x - \alpha e^T x + u^T u dt = \mathcal{E} \beta^T x(0)$$

and this is the smallest possible value of η_1

Notice that this implies that there is at most one solution of the α,β equations such that $A - \frac{1}{2}\sum_{i=1}^{m} B_i \text{ diag } (\mathrm{B_i^T}\beta) \in \mathcal{G}$

V. Controllability

Consider a system of the form

$$\dot{p} = Ap + \sum u_i B_i p$$

with A and B_i such that $e^T A = e^T B_i = 0$. For such systems we see that $e^T p$ is constant and that the system evolves on a hyperplane $e^T p(t) = e^T p(0)$. From the transition matrix point of view, solutions of

$$\dot{P} = AP + \sum u_i B_i P \;\; ; \;\; P(0) = I$$

evolve on the group of matrices $G = \{X \mid e^T X = e^T$. If the system is n-dimensional then this is a $n^2 - n$-dimensional group. A similarity transformation can be used to transform the eigenvector e^T to the vector e_n showing that as a Lie group this group is isomorphic to the group of affine transformations of the standard form,

$$G = \begin{bmatrix} A & b \\ 0 & 1 \end{bmatrix}$$

The reachable set from I will have a nonempty interior in this group if the Lie algebra generated by A and B_i is of dimension $n^2 - n$. In Lemma 1 we have assumed an explicit controllability condition but it could be replaced by this Lie algebraic condition.

Notice that a system with two noninteracting parts would typically have steady state minimum return functions with different rates of growth (their α's would not be equal) so that without some further assumption the asymptotic growth implied by Lemma 2 would not be described by a single constant α but rather by two or more different rates of growth. If the assumption that A is irreducible is dropped then the situation described by Lemma 2 does not follow without some mixing hypothesis. If A provides no interaction between subsystems the B_i may do so if the system is controllable. In this case the effect of an optimal control policy would be to move the probability from the higher rates of growth to the lowest rate of growth. Thus from this point of view it seems that the role of irreducibility should be replaced by a controllability condition which would assure that the state can be steered to the most advantageous invariant distribution.

VI. Examples

To illustrate the ideas involved here we give one quite specific example of a three state system worked out in detail. We also give an example which formulates a path planning problem as a Markov decision problem of the type considered here.

Example 1: Consider a three state markov chain having a one parameter family of possible rates. The differential equation for the probability law depends on a control u in accordance with

$$\frac{d}{dt} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} = \begin{bmatrix} -1 & 1 & 0 \\ 1 & -2 & 1 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} +$$

$$u\begin{bmatrix} -1 & 0 & 0 \\ 1 & 0 & 1 \\ 0 & 0 & -1 \end{bmatrix}\begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix}$$

In this case $A+uB \in \mathcal{G}$ if $u \geq -1$. The problem we consider is that of minimizing

$$\eta = \mathcal{E}\int_0^T 3x_1 + 3x_3 + u^2 dt$$

The idea is that it is advantageous to keep the system in state e_2 but there is a penalty associated with using u. More specifically, for $x = e_1$ or $x = e_3$ we need $u \geq -1$. There is no constraint on u if $x = e_2$. From Theorem 1 we see that if the system is in state e_1 and if $-k^T Be_1/2 \geq -1$ the optimal control is $u = (k_1 - k_2)/2$; otherwise the optimal control is $u = -1$. If $x = e_3$ and $-k^T Be_3/2 \geq -1$ the optimal control is $u = (k_3 - k_2)/2$; otherwise the optimal control is $u = -1$. If $x = e_2$ the optimal control is zero. According to Theorem 1 the equation for k is

$$\begin{bmatrix} \dot{k}_1 \\ \dot{k}_1 \\ \dot{k}_1 \end{bmatrix} = -\begin{bmatrix} -1 & 1 & 0 \\ 1 & -2 & 1 \\ 0 & 1 & -1 \end{bmatrix}\begin{bmatrix} k_1 \\ k_2 \\ k_3 \end{bmatrix} - \begin{bmatrix} 3 \\ 0 \\ 3 \end{bmatrix} +$$

$$\begin{bmatrix} \max\{(k_2-k_1)^2/4, k_2-k_1+1\} \\ 0 \\ \max\{(k_2-k_3)^2/4, k_3-k_2+1\}) \end{bmatrix}$$

There is no end point penalty term so the boundary condition is $x(T) = 0$.

We now proceed to show that in this case, the control $-k^T Bx/2$ does not exceed the limits on u so that for this problem the equation for k reduces to

$$\dot{k}_1 = k_1 - k_2 - 3 + (k_2 - k_1)^2/4$$

$$\dot{k}_2 = -k_1 + 2k_2 - k_3$$

$$\dot{k}_3 = -k_2 + k_3 - 3 + (k_2 - k_3)^2/4$$

Because the problem is time invariant, it is natural to look for the steady state solution. Following Theorem 3, we look for a constant α and a vector β such that $\alpha = \beta_1 - \beta_2 - 3 - (\beta_2 - \beta_1)^2/4$, $\alpha = \beta_1 + 2\beta_2 - \beta_3$ and $\alpha = -\beta_2 + \beta_3 - 3 + (\beta_2 - \beta_3)^2/4$. Clearly β_1 and β_3 enter this set of equations symmetrically and so for the relevant solution they will be equal. Let $a = \beta_1 - \beta_2 = \beta_3 - \beta_2$. From the second equation we have $\alpha = 2a$ and from the first

$$2a = -a + 3 - a^2/4 \quad \Longrightarrow \quad a = -6(1 \pm \sqrt{1+1/3})$$

Because α defines the steady state performance, and because this is positive, we need to select the minus sign in the expression for a

$$a = 6\sqrt{1+1/3} - 6 \approx .92$$

Thus, the steady state feedback control law can be expressed as

$$u(e_1) = -.46 \;\; ; \;\; u(e_2) = 0 \;\; ; \;\; u(e_2) = -.46$$

This example admits an interpretation as a more classical control problem. Consider a scalar random process $z(t)$ that takes on the values +1, 0, -1. Suppose that it is desired to "reset" $z(t)$ to zero whenever it deviates from zero. Consider a performance measure penalizing the integral of the sum $z^2(t) + u^2(t)$. If the stochastic process description of the evolution of z matches the one given here the optimal solution can be identified from the solution above. In this interpretation the system might be compared with the more widely studied stochastic regulator problem

$$dy = -cydt + udt + dw \;;\; \eta = \lim_{T\to\infty}\frac{1}{T}\mathcal{E}\int_0^T qy^2 + u^2 dt$$

which has a steady state optimal control of similar form

$$u(t) = (c - \sqrt{c^2+q^2})y$$

Example 2: As an indication of the broader applicability of the model used here, consider a path planning problem in which an autonomous vehicle is to traverse a directed graph, (move along roads) moving from an initial vertex v_1 while seeking to end at a goal vertex v_g. Label the branches of the graph $x_1, x_2, ..., x_n$. Assume that the action of the robot cannot be described deterministically because of uncertainties in the environment. The path actually realized is modeled as evolving in continuous time with the robot traversing one branch after another. The time to traverse a path is determined by the transition times of the Markov process. This means that when a control is selected, the next branch traversed is determined probabilistically, subject to the allowed transitions coded by the G_i. There is a cost associated with being on a branch and a final end point cost as well. In terms of equations we have

$$dx = \sum G_j x dN_j \;\;;\;\; \eta = \int_0^T c^T x + F(u)dt + \phi(x(T)$$

and a performance measure

$$\eta = \mathcal{E}\int_0^T c^T(t)x(t) + u^T u d\sigma + k_f^T \mathcal{E}x(T)$$

VII. ACKNOWLEDGMENTS

The author thanks Steve Marcus of the University of Maryland and Xi-Ren Cao of the Hong Kong University of Science and Technology for their comments on an early draft.

REFERENCES

[1] Richard Bellman. *Dynamic Programming*, Princeton University Press, 1957.

[2] R. Howard. *Dynamic Programming and Markov Processes*, MIT Press and John Wiley & Sons, 1960.

[3] M.H.A. Davis, *Markov Models and Optimization*, Chapman and Hall, London, 1993.

[4] Nanayaa Twum-Danso and Roger Brockett, "Trajectory Estimation from Place Cell Data," *Neural Networks,* vol. 14 (2001), pp. 835-844.

[5] Roger W. Brockett, "Reduced Complexity Control Systems,", in *Plenary Papers, Milestone Reports, & Slected Survey Papers,* Myung Jin Chung and Pradeep Misra, Eds., 17th IFAC World Congress, Seoul Koera, 2008.

[6] D. Bertsekas and J. Tsitsiklis, *Neuro-Dynamic Programming*, Athena Scientific, Belmont, MA, 1996.

[7] Sutton, R. S. and Barto A. G. *Reinforcement Learning: An Introduction*, The MIT Press, Cambridge, MA, 1998.

[8] Xi-Ren Cao, *Stochastic Learning and Optimization*, Springer Verlag, New York, 2007