Action-Constrained Markov Decision Processes With Kullback–Leibler Cost

Ana Bušić ANA.BUSIC@INRIA.FR
Inria, Paris research centre
DI ENS, École normale supérieure, CNRS, PSL Research University, Paris, France

Sean Meyn MEYN@ECE.UFL.EDU
Department of Electrical and Computer Engineering at the University of Florida

Editors: Sébastien Bubeck, Vianney Perchet and Philippe Rigollet

Abstract

This paper concerns computation of optimal policies in which the one-step reward function contains a cost term that models Kullback-Leibler divergence with respect to nominal dynamics. This technique was introduced by Todorov in 2007, where it was shown under general conditions that the solution to the average-reward optimality equations reduce to a simple eigenvector problem. Since then many authors have sought to apply this technique to control problems and models of bounded rationality in economics.

A crucial assumption is that the input process is essentially unconstrained. For example, if the nominal dynamics include randomness from nature (e.g., the impact of wind on a moving vehicle), then the optimal control solution does not respect the exogenous nature of this disturbance.

This paper introduces a technique to solve a more general class of action-constrained MDPs. The main idea is to solve an entire parameterized family of MDPs, in which the parameter is a scalar weighting the one-step reward function. The approach is new and practical even in the original unconstrained formulation.

Keywords: Markov decision processes, Computational methods.

1. Introduction

Consider a Markov Decision Process (MDP) with finite state space X, general action space U, and one-step reward function $w\colon \mathsf{X}\times\mathsf{U}\to\mathbb{R}$. Two standard optimal control criteria are *finite-horizon*:

$$\mathcal{W}_T^*(x) = \max \sum_{t=0}^{T} \mathsf{E}[w(X(t),U(t)) \mid X(0)=x] \tag{1}$$

where $T \geqslant 0$ is fixed, and *average reward*:

$$\eta^*(x) = \max\Big\{ \liminf_{T\to\infty} \frac{1}{T}\mathsf{E}[w(X(t),U(t)) \mid X(0)=x]\Big\}\,. \tag{2}$$

where $\boldsymbol{X} = \{X(t) : t \geqslant 0\}$, $\boldsymbol{U} = \{U(t) : t \geqslant 0\}$ denote the state and input sequences.

In either case, the maximum is over all admissible input sequences; it is obtained as deterministic state feedback under general conditions. In the average-reward framework the optimal policy is typically stationary: $U(t) = \phi^*(X(t))$ for a mapping $\phi^*\colon \mathsf{X}\to\mathsf{U}$, and $\eta^*(x)$ does not depend upon the initial condition x (see Puterman (2014); Bertsekas and Shreve (1996)).

A special class of MDP models was introduced by Todorov (2007), for which either optimal control problem has an attractive solution. The reward function is assumed to be the sum of two terms: the first is a function $\mathcal{U}\colon \mathsf{X} \to \mathbb{R}$ that is completely unstructured. The second term is a "control cost", defined using Kullback–Leibler (K-L) divergence (also known as *relative entropy*). The control cost is based on deviation from nominal (control-free) behavior; modeled by a nominal transition matrix P_0.

It is shown that the solution with respect to the average reward criterion is obtained as the solution to the following eigenvector problem: let (λ, v) denote the Perron-Frobenius eigenvalue-eigenvector pair for the positive matrix with entries $\widehat{P}(x, x') = \exp(\mathcal{U}(x))P_0(x, x')$, $x, x' \in \mathsf{X}$. The eigenvector property $\widehat{P}v = \lambda v$ implies that the "twisted" matrix

$$\check{P}(x, x') = \frac{1}{\lambda}\frac{v(x')}{v(x)}\widehat{P}(x, x')\,, \quad x, x' \in \mathsf{X}\,. \tag{3}$$

is a transition matrix on X. This transition matrix defines the dynamics of the model under optimal control. A similar model was introduced in the earlier work of Kárný (1996), but without the complete solution reviewed here.

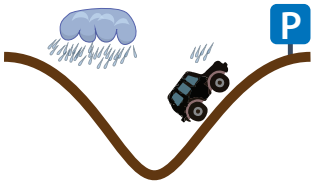


Figure 1: Optimal hill climb

Since the publication of Todorov (2007) there has been significant theoretical advancement, with proposed applications to economics Guan et al. (2014), distributed control Meyn et al. (2015), and neuroscience Doya (2009).

It is appealing to imagine that rational economic agents are solving an eigenvector problem to maximize their utility. However, a careful look at the controlled dynamics (3) suggests a limitation of this MDP formulation: *how can this transformation respect exogenous disturbances from nature?* An essential assumption in this prior work is that for each x, and any pmf μ, it is possible to choose the action so that $P(x, x') = \mu(x')$. This is equivalent to the assumption that the action space U consists of all probability mass functions on X, and the controlled transition matrix is entirely determined by the input as follows:

$$\mathsf{P}\{X(t+1) = x' \mid X(t) = x, U(t) = \mu\} = \mu(x')\,, \qquad x, x \in \mathsf{X},\ \mu \in \mathsf{U}\,. \tag{4}$$

This modeling assumption presents a significant limitation, as pointed out in Todorov (2009): "*It prevents us from modeling systems subject to disturbances outside the actuation space*".

Fig. 1 is based on an example of Todorov (2009). Reaching the parking spot at the top of the hill in minimum time (or minimal fuel) is formulated as a total *cost* problem, similar to (1). The figure has been modified to indicate that wind and rain influence the behavior of the car on the track. The optimal solution cannot take the form (3) when this additional randomness is included in the model, since this would mean our control action would modify the weather.

Contributions In this paper the K-L cost framework is broadened to include constraints on the pmf μ appearing in (4). The new approach to computation is based on the solution of an entire family of MDP problems, parameterized by a scalar ζ appearing as a weighting factor in the one-step reward function. Letting X_t denote the state, and R_t denote the randomized policy at time t, this one-step reward is of the form

$$w(X_t, R_t) = \zeta\mathcal{U}(X_t) - c_{\text{KL}}(X_t, R_t) \tag{5}$$

in which c_{KL} denotes relative entropy with respect to nominal dynamics (see (13)).

The main results of the paper are contained in Theorems 1 and 4, with parallel results for the total- and average-reward control problems. In each case, it is shown that *the solution to an entire family of MDPs can be obtained through the solution of a single ordinary differential equation* (ODE).

The ODE solution is most elegant in the average-reward setting. For each ζ, the solution to the average-reward optimization problem is based on a relative value function $h^*_\zeta \colon \mathsf{X} \to \mathbb{R}$. For the MDP with d states, each function is viewed as a vector in $\mathbb{R}^d$ with entries $\{h^*_\zeta(x^i) : 1 \leqslant i \leqslant d\}$. A vector field $\mathcal{V} \colon \mathbb{R}^d \to \mathbb{R}^d$ is constructed so that these functions solve the ODE

$$\tfrac{d}{d\zeta} h^*_\zeta = \mathcal{V}(h^*_\zeta)\,, \qquad \text{with boundary condition } h^*_0 \equiv 0.$$

One step in the construction of $\mathcal{V}$ is differentiating each side of the dynamic programming equations; a starting point of the 50 year old sensitivity theory of Schweitzer (1968), and more recent Sutton et al. (1999). More closely related is the sensitivity theory surrounding Perron-Frobenius eigenvectors that appears in the theory of large deviations (Kontoyiannis and Meyn, 2003, Prop. 4.9). The goals of this prior work are different, and we are not aware of comparable algorithms that simultaneously solve the family of control problems.

The optimal control formulation is far more general than in the aforementioned work Todorov (2007); Guan et al. (2014); Meyn et al. (2015), as it allows for inclusion of exogenous randomness in the MDP model. The dynamic programming equations become significantly more complex in this generality, so that in particular, the Perron-Frobenious computational approach used in prior work is no longer applicable.

In addition to its value as a computational tool, there is a significant benefit to solve the entire collection of optimal control problems for a range of the parameter ζ. For example, this provides a means to understand the tradeoff between state cost and control effort. Simultaneous computation of the optimal policies is also an essential ingredient of the distributed control architecture introduced in Meyn et al. (2015).

The ODE algorithm is easily implemented for problems of moderate size. In this paper an example is provided in which the the size of the state space d is greater than 1,000; the action space is an open subset $\mathbb{R}^{d-1}$ since actions correspond to randomized decision rules. The optimal solutions for the desired range of ζ were obtained in less than one hour using a standard laptop running Matlab.

The remainder of the paper is organized as follows. Section 2 describes the new Kullback–Leibler cost criterion and numerical techniques for the MDP solutions. This is applied to a path-finding problem in Section 3. Conclusions and topics for future research are contained in Section 4.

2. MDPs with Kullback–Leibler Cost

2.1. MDP model

The dynamics of the MDP are assumed of the form (4), where the action space consists of a convex subset of probability mass functions (pmf) on X. An explanation of the one-step reward (5) will be provided after a few preliminaries.

A transition matrix P_0 is given that describes nominal (control-free) behavior. It is assumed to be *irreducible and aperiodic*. It follows that P_0 admits a unique invariant pmf, denoted π_0. For any

other transition matrix, with unique invariant pmf π, the *Donsker-Varadhan rate function* is denoted,

$$K(P\|P_0) = \sum_{x,x'} \pi(x)P(x,x') \log\Big(\frac{P(x,x')}{P_0(x,x')}\Big) \tag{6}$$

under the usual convention that "$0\log(0) = 0$". It is called a "rate function" because it defines the relative entropy rate between two stationary Markov chains, see Dembo and Zeitouni (1998).

As in Todorov (2007); Guan et al. (2014); Meyn et al. (2015), the rate function is used here to model the cost of deviation from the nominal transition matrix P_0. The two control objectives surveyed in the introduction will be specialized as follows, based on the utility function $\mathcal{U}\colon \mathsf{X} \to \mathbb{R}$ and a scaling parameter $\zeta \geqslant 0$. For the finite-horizon optimal control problem,

$$\mathcal{W}_T^*(x,\zeta) = \max \sum_{t=0}^{T} \big\{\zeta \mathsf{E}_x[\mathcal{U}(X(t))] - K(P_t\|P_0)\big\}\,, \tag{7}$$

where the expectation is conditional on $X(0) = x$. The average reward optimization problem is analogous:

$$\eta^*(\zeta) = \max\Big(\liminf_{T\to\infty} \frac{1}{T}\big\{\zeta \mathsf{E}_x[\mathcal{U}(X(t))] - K(P_t\|P_0)\big\}\Big)\,. \tag{8}$$

In each case, the maximum is over all transition matrices $\{P_t\}$. In this context, the one-step reward appearing in (1, 2) is a function of pairs (x, P):

$$w(x,P) := \zeta\mathcal{U}(x) - \sum_{x'} P(x,x') \log\Big(\frac{P(x,x')}{P_0(x,x')}\Big) \tag{9}$$

for any $x \in \mathsf{X}$ and transition matrix P.

There is practical value to considering a parameterized family of reward functions. For one, it is useful to understand the sensitivity of the control solution to the relative weight given to utility and the penalty on control action. This is well understood in classical linear control theory – consider for example the celebrated symmetric root locus in linear optimal control Franklin et al. (1997).

Nature & nurture Exogenous randomness from nature imposes additional constraints in the optimal control problem (7) or (8).

It is assumed that the state space is the cartesian product of two finite sets: $\mathsf{X} = \mathsf{X}_\mathsf{u} \times \mathsf{X}_\mathsf{n}$, and the state is similarly expressed $X(t) = (X_u(t), X_n(t))$. At a given time t it is assumed that $X_n(t+1)$ is conditionally independent of the input at time t, given the value of $X(t)$. This is formalized by the following conditional-independence assumption:

$$P(x,x') = R(x,x'_u)Q_0(x,x'_n), \quad x = (x_u, x_n) \in \mathsf{X},\ x'_u \in \mathsf{X}_\mathsf{u},\ x'_n \in \mathsf{X}_\mathsf{n} \tag{10}$$

The matrix R defines the randomized decision rule for $X_u(t+1)$ given $X(t)$. The matrix Q_0 is fixed and models the distribution of $X_n(t+1)$ given $X(t) = x$, and each are subject to the pmf constraint: $\sum_{x'_u} R(x,x'_u) = \sum_{x'_n} Q_0(x,x'_n) = 1$ for each x.

Subject to the constraint (10), the two optimal control problems (8, 9) are transformed to the final forms considered in this paper:

$$\mathcal{W}_T^*(x,\zeta) = \max \sum_{t=0}^{T} \mathsf{E}_x[w(X(t),R(t))] \tag{11}$$

$$\eta^*(\zeta) = \max\Big\{\liminf_{T\to\infty} \mathsf{E}_x[w(X(t),R(t))]\Big\} \tag{12}$$

where in each case the maximum is over sequences of randomized decision rules $\{R(0), \dots, R(T)\}$,

$$\begin{aligned} w(x, R) &:= \zeta \mathcal{U}(x) - c_{\text{KL}}(x, R) \\ \textit{and} \quad c_{\text{KL}}(x, R) &:= \sum_{x'} P(x, x') \log\Big(\frac{P(x, x')}{P_0(x, x')}\Big) = \sum_{x'_u} R(x, x'_u) \log\Big(\frac{R(x, x'_u)}{R_0(x, x'_u)}\Big) \end{aligned} \tag{13}$$

2.2. Notation

For any transition matrix P, an invariant pmf is interpreted as a row vector, so that invariance can be expressed $\pi P = \pi$. Any function $f\colon \mathsf{X} \to \mathbb{R}$ is interpreted as a d-dimensional column vector, and we use the standard notation $Pf\,(x) = \sum_{x'} P(x, x') f(x')$, $x \in \mathsf{X}$. The *fundamental matrix* is the inverse,

$$Z = [I - P + 1 \otimes \pi]^{-1} \tag{14}$$

where $1 \otimes \pi$ is a matrix in which each row is identical, and equal to π. If P is irreducible and aperiodic, then it can be expressed as the power series $Z = \sum_{n=0}^{\infty} [P - 1 \otimes \pi]^n$, with $[P - 1 \otimes \pi]^0 := I$ (the $d \times d$ identity matrix), and $[P - 1 \otimes \pi]^n = P^n - 1 \otimes \pi$ for $n \geqslant 1$.

Any function $g\colon \mathsf{X} \times \mathsf{X} \to \mathbb{R}$ is regarded as an unnormalized log-likelihood ratio: Denote for $x, x' \in \mathsf{X}$,

$$P_g(x, x') := P_0(x, x') \exp\big(g(x' \mid x) - \Lambda_g(x)\big), \tag{15}$$

in which $g(x' \mid x)$ is the value of g at $(x, x') \in \mathsf{X} \times \mathsf{X}$, and $\Lambda_g(x)$ is the normalization constant,

$$\Lambda_g(x) := \log\Big(\sum_{x'} P_0(x, x') \exp\big(g(x' \mid x)\big)\Big) \tag{16}$$

The rate function can be expressed in terms of its invariant pmf π_g, the bivariate pmf $\Pi_g(x, x') = \pi_g(x) P_g(x, x')$, and the log moment generating function (16):

$$\begin{aligned} K(P_g \| P_0) &= \sum_{x, x'} \Pi_g(x, x') \big[g(x' \mid x) - \Lambda_g(x)\big] \\ &= \sum_{x, x'} \Pi_g(x, x') g(x' \mid x) - \sum_{x} \pi_g(x) \Lambda_g(x) \end{aligned} \tag{17}$$

The unusual notation is introduced because $g(x' \mid x)$ will take the form of a conditional expectation in all of the results that follow: given any function $h\colon \mathsf{X} \to \mathbb{R}$ we denote

$$h(x'_u \mid x) = \sum_{x'_n} Q_0(x, x'_n) h(x'_u, x'_n)\,. \tag{18}$$

In this case the transformation only transforms the dynamics of $\boldsymbol{X}_u$:

$$P_h(x, x') = R_h(x, x'_u) Q_0(x, x'_n)\,, \quad R_h(x, x'_u) := R_0(x, x'_u) \exp\big(h(x'_u \mid x) - \Lambda_g(x)\big)\,.$$

2.3. ODE for finite time horizon

Here an ODE is constructed to compute the value functions $\{\mathcal{W}^*_\tau(x,\zeta) : 1 \leqslant \tau \leqslant T\,, \zeta \geqslant 0\}$. To aide exposition it is helpful to first look at the general problem: Assume that the state space X is finite, the action space U is *general*, and let $\{P_u(x,x')\}$ denote the controlled transition matrix. The one-step reward on state-action pairs is of the form $w(x,u) = \zeta\mathcal{U}(x) - c(x,u)$, where $c\colon \mathsf{X}\times\mathsf{U} \to \mathbb{R}_+$. Assume that $c(x,u) \equiv 0$ for a unique value $u = u_0$.

For each $1 \leqslant \tau \leqslant T$ denote, as in (1),

$$\mathcal{W}^*_\tau(x,\zeta) = \max \sum_{t=0}^{\tau} \mathsf{E}_x[w(X(t),U(t))] \tag{19}$$

where the maximum is over all admissible inputs $\{U(t) = \phi_t(X(0),\dots,X(t))\}$. Each value function can be regarded as the maximum over functions $\{\phi_t\}$ (subject to measurability conditions and hard constraints on the input). It is assumed that the maximum (19) is finite for each (x,ζ).

The dynamic programming equation (principle of optimality) holds: for $\tau \geqslant 1$,

$$\mathcal{W}^*_\tau(x,\zeta) = \max_u \Big\{\zeta\mathcal{U}(x) - c(x,u) + \sum_{x'} P_u(x,x')\mathcal{W}^*_{\tau-1}(x')\Big\} \tag{20}$$

Assume that a maximizer $\phi^*_{\tau-1,\zeta}(x)$ exits for each τ,ζ, and x.

A crucial observation is that for each x, the value function appearing in (19) is the maximum of functions that are affine in ζ. It follows that $\mathcal{W}^*_\tau(x,\zeta)$ is convex as a function of ζ, and hence absolutely continuous. Consequently, the right derivative $H^*_\tau(x,\zeta) := \frac{d^+}{d\zeta}\mathcal{W}^*_\tau(x,\zeta)$ exists everywhere. A recursive equation follows from (20):

$$H^*_\tau(x,\zeta) = \mathcal{U}(x) + \sum_{x'} \check{P}_{\tau-1,\zeta}(x,x')H^*_{\tau-1}(x',\zeta) \tag{21}$$

where $\check{P}_{\tau-1,\zeta}(x,x') = P_{u^*}(x,x')$ with $u^* = \phi^*_{\tau-1,\zeta}(x)$.

In matrix notation this becomes $H^*_\tau = \check{Z}_{\tau-1,\zeta}\mathcal{U}$, where $\check{Z}_{0,\zeta} = I$, and for any $1 \leqslant \tau \leqslant T$,

$$\check{Z}_{\tau-1,\zeta} = I + \check{P}_{\tau-1,\zeta} + \check{P}_{\tau-1,\zeta}\check{P}_{\tau-2,\zeta} + \check{P}_{\tau-1,\zeta}\check{P}_{\tau-2,\zeta}\cdots\check{P}_{0,\zeta} \tag{22}$$

This is similar to a truncation of the power series representation of the fundamental matrix (14).

Denote $\mathcal{W}^*_\zeta(x) = \{\mathcal{W}^*_k(x,\zeta) : 0 \leqslant k \leqslant T\}$, regarded as a vector in $\mathbb{R}^{|\mathsf{X}|\times(T+1)}$, parameterized by the non-negative constant ζ. The following result follows from the preceding arguments:

Theorem 1 *The family of functions $\{\mathcal{W}^*_\zeta\}$ solves the ODE $\frac{d^+}{d\zeta}\mathcal{W}^*_\zeta = \mathcal{V}(\mathcal{W}^*_\zeta)$, $\zeta \geqslant 0$, with boundary condition $\mathcal{W}^*_0 = 0$. The vector field can be described in block-form as follows, with $T+1$ blocks:*

$$\tfrac{d^+}{d\zeta}\mathcal{W}^*_k(\,\cdot\,,\zeta) = \mathcal{V}_k(\mathcal{W}^*_\zeta)\,, \quad 0 \leqslant k \leqslant T\,.$$

The identity $\mathcal{V}_0(\mathcal{W}) = \mathcal{U}$ holds for any $\mathcal{W}$. For $k \geqslant 1$, the right hand side depends on its argument only through the associated policy: for any sequence of functions $\mathcal{W}^* = (\mathcal{W}^*_0,\dots,\mathcal{W}^*_T)$,

$$\begin{aligned}
\mathcal{V}_k(\mathcal{W}) &= Z_{k-1}\mathcal{U} \\
\textit{where} \qquad Z_{k-1} &= I + P_{k-1} + P_{k-1}P_{k-2} + P_{k-1}P_{k-2}\cdots P_0 \\
P_i(x,x') &= P_{\phi_i(x)}(x,x')\,, \quad \textit{all } i,\ x,\ x', \\
\phi_i(x) &= \arg\max_u \Big\{-c(x,u) + \sum_{x'} P_u(x,x')\mathcal{W}^*_i(x')\Big\}\,, \qquad 1 \leqslant i,k \leqslant T.
\end{aligned}$$

□

The theorem provides valuable computational tools for models of moderate cardinality and moderate time-horizon. Two questions remain:

(i) What is ϕ_i for the problem under study in this paper?

(ii) Can a tractable ODE be constructed in infinite-horizon optimal control problems?

The answer to the second question is the focus of Section 2.4. The answer to (i) is contained in the following. For any function $\mathcal{W}\colon \mathsf{X} \to \mathbb{R}$, denote

$$R_{\mathcal{W}}(x, \cdot) = \arg\max_{R} \Big\{ w(x, R) + \sum_{x'} P(x, x') \mathcal{W}(x') \Big\}, \quad x \in \mathsf{X},$$

subject to the constraint that P depends on R via (10), and with w defined in (13).

Proposition 2 *For any function $\mathcal{W}$ the maximizer $R_{\mathcal{W}}$ is unique and can be expressed*

$$R_{\mathcal{W}}(x, x'_n) = R_0(x, x'_n) \exp\big(\mathcal{W}(x'_u \mid x) - \Lambda(x)\big)$$

where $\mathcal{W}(x'_u \mid x) = \sum_{x'_n} Q_0(x, x'_n) \mathcal{W}(x'_u, x'_n)$ for each $x \in \mathsf{X}$, $x'_u \in \mathsf{X}_{\mathsf{u}}$, and $\Lambda(x)$ is a normalizing constant, defined so that $R_{\mathcal{W}}(x, \cdot)$ is a pmf for each x.

Proof Given the form of the reward w and the constraint on P, the optimization problem of interest here can be written, for each x, as

$$R_{\mathcal{W}}(x, \cdot) = \arg\max_{\mu} \big\{ \mu(\widehat{\mathcal{W}}) - D(\mu \| \mu_0) \big\}$$

where the variable $\mu(\cdot)$ represents $R(x, \cdot)$, $\mu_0 = R_0(x, \cdot)$, and

$$\mu(\widehat{\mathcal{W}}) = \sum_{x' = (x'_u, x'_n)} R(x, x'_u) Q_0(x, x'_n) \mathcal{W}(x'_u, x'_n) = \sum_{x'_u} \mu(x'_u) \mathcal{W}(x'_u \mid x)$$

The proposition is a consequence of this combined with Theorem 3.1.2 of Dembo and Zeitouni (1998) (i.e., convex duality between relative entropy and the log moment generating function). ■

It follows from the proposition that the vector field is smooth in a neighborhood of the optimal solution $\{\mathcal{W}^*_\zeta : \zeta \geqslant 0\}$. These results are central to the average-reward case considered next.

2.4. Average reward formulation

We consider now the case of average reward (12), subject to the structural constraint (10). The associated average reward optimization equation (AROE) is expressed as follows:

$$\max_{R} \Big\{ w(x, R) + \sum_{x'} P(x, x') h^*_\zeta(x') \Big\} = h^*_\zeta(x) + \eta^*(\zeta) \tag{23}$$

In which $\eta^*(\zeta)$ is the optimal average reward, and h^*_ζ is the *relative value function*. The maximizer defines a transition matrix:

$$\check{P}_\zeta = \arg\max_{P} \big\{ \zeta \pi(\mathcal{U}) - K(P \| P_0) : \pi P = \pi \big\} \tag{24}$$

Recall that the relative value function is not unique, since a new solution is obtained by adding a non-zero constant; the normalization $h^*_\zeta(x^\circ) = 0$ is imposed, where $x^\circ \in \mathsf{X}$ is a fixed state.

The proof of Theorem 3 (i) is a consequence of Prop. 2. The second result is obtained on combining Lemmas B.2–B.4 of Bušić and Meyn (2018).

Theorem 3 *There exist optimizers* $\{\check{\pi}_\zeta, \check{P}_\zeta : \zeta \in \mathbb{R}\}$*, and solutions to the AROE* $\{h^*_\zeta, \eta^*(\zeta) : \zeta \in \mathbb{R}\}$ *with the following properties:*

(i) *The optimizer* $\check{P}_\zeta$ *can be obtained from the relative value function* h^*_ζ *as follows:*

$$\check{P}_\zeta(x, x') := P_0(x, x') \exp\big(h_\zeta(x'_u \mid x) - \Lambda_{h_\zeta}(x)\big) \tag{25}$$

where for $x \in \mathsf{X}$*,* $x'_u \in \mathsf{X}_{\mathsf{u}}$*,*

$$h_\zeta(x'_u \mid x) = \sum_{x'_n} Q_0(x, x'_n) h^*_\zeta(x'_u, x'_n), \tag{26}$$

and $\Lambda_{h_\zeta}(x)$ *is the normalizing constant* (16) *with* $h = h_\zeta$*.*

(ii) $\{\check{\pi}_\zeta, \check{P}_\zeta, h^*_\zeta, \eta^*(\zeta) : \zeta \in \mathbb{R}\}$ *are continuously differentiable in the parameter* ζ*.* □

Representations for the derivatives in Theorem 3 (ii), in particular the derivative of $\Lambda_{h^*_\zeta}$ with respect to ζ, lead to a representation for the ODE used to compute the transition matrices $\{\check{P}_\zeta\}$.

It is convenient to generalize the problem slightly here: let $\{h^\circ_\zeta : \zeta \in \mathbb{R}\}$ denote a family of functions on X, continuously differentiable in the parameter ζ. They are not necessarily relative value functions, but we maintain the structure established in Theorem 3 for the family of transition matrices. Denote,

$$h_\zeta(x'_u \mid x) = \sum_{x'_n} Q_0(x, x'_n) h^\circ_\zeta(x'_u, x'_n), \quad x \in \mathsf{X},\ x'_u \in \mathsf{X}_{\mathsf{u}} \tag{27}$$

and then define as in (15),

$$P_\zeta(x, x') := P_0(x, x') \exp\big(h_\zeta(x'_u \mid x) - \Lambda_{h_\zeta}(x)\big) \tag{28}$$

The function $\Lambda_{h_\zeta} : \mathsf{X} \to \mathbb{R}$ is a normalizing constant, exactly as in (16):

$$\Lambda_{h^\circ_\zeta}(x) := \log\Big(\sum_{x'} P_0(x, x') \exp\big(h_\zeta(x'_u \mid x)\big)\Big)$$

We begin with a general method to construct a family of functions $\{h^\circ_\zeta : \zeta \in \mathbb{R}\}$ based on an ODE. The ODE is expressed,

$$\tfrac{d}{d\zeta} h^\circ_\zeta = \mathcal{V}(h^\circ_\zeta)\,, \qquad \zeta \in \mathbb{R}, \tag{29}$$

with boundary condition $h^\circ_0 \equiv 0$. A particular instance of the method will result in $h^\circ_\zeta = h^*_\zeta$ for each ζ. Assumed given is a mapping $\mathcal{H}^\circ$ from transition matrices to functions on X. Following this, the vector field $\mathcal{V}$ is obtained through the following two steps: For a function $h \colon \mathsf{X} \to \mathbb{R}$,

(i) Define a new transition matrix via (15),

$$P_h(x, x') := P_0(x, x') \exp\big(h(x'_u \mid x) - \Lambda_h(x)\big), \quad x, x' \in \mathsf{X}, \tag{30}$$

in which $h(x'_u \mid x) = \sum_{x'_n} Q_0(x, x'_n) h(x'_u, x'_n)$, and $\Lambda_h(x)$ is a normalizing constant.

(ii) Compute $H^\circ = \mathcal{H}^\circ(P_h)$, and define $\mathcal{V}(h) = H^\circ$. It is assumed that the functional $\mathcal{H}^\circ$ is constructed so that $H^\circ(x^\circ) = 0$ for any h.

We now specify the functional $\mathcal{H}^\circ$, whose domain consists of transition matrices that are irreducible and aperiodic. For any transition matrix P in this domain, the fundamental matrix Z is obtained using (14), and then $H^\circ = \mathcal{H}^\circ(P)$ is defined as

$$H^\circ(x) = \sum_{x'} [Z(x, x') - Z(x^\circ, x')]\mathcal{U}(x'), \qquad x \in \mathsf{X} \tag{31}$$

The function H° is a solution to Poisson's equation,

$$PH^\circ = H^\circ - \mathcal{U} + \bar{u}\,, \qquad \text{where } \bar{u} := \pi(\mathcal{U}) := \sum_x \pi(x)\mathcal{U}(x). \tag{32}$$

Theorem 4 *Consider the ODE (29) with boundary condition $h_0^\circ \equiv 0$, and with $H^\circ = \mathcal{H}^\circ(P)$ defined using (31) for each transition matrix P that is irreducible and aperiodic. The solution to this ODE exists, and the resulting functions $\{h_\zeta^\circ : \zeta \in \mathbb{R}\}$ coincide with the relative value functions $\{h_\zeta^* : \zeta \in \mathbb{R}\}$. Consequently, $\check{P}_\zeta = P_{h_\zeta}$ for each ζ.*

Proof The proof requires validation of the representation $H_\zeta^* = \mathcal{H}^\circ(\check{P}_\zeta)$ for each ζ, where h_ζ^* is the relative value function, $\check{P}_\zeta$ is defined in (24), and

$$H_\zeta^* = \tfrac{d}{d\zeta} h_\zeta^* \tag{33}$$

Substituting the maximizer $\check{P}_\zeta$ in the form (25) into the AROE gives the fixed point equation $\zeta\mathcal{U} + \Lambda_{h_\zeta^*} = h_\zeta^* + \eta^*(\zeta)$. Differentiating each side then gives,

$$\mathcal{U} + \check{P}_\zeta H_\zeta^* = H_\zeta^* + \tfrac{d}{d\zeta}\eta^*(\zeta). \tag{34}$$

This is Poisson's equation, and it follows that $\check{\pi}_\zeta(\mathcal{U}) = \frac{d}{d\zeta}\eta^*(\zeta)$. Moreover, since $h_\zeta^*(x^\circ) = 0$ for every ζ, we must have $H_\zeta^*(x^\circ) = 0$ as well. Since the solution to Poisson's equation with this normalization is unique, we conclude that (33) holds, and hence $H_\zeta^* = \mathcal{H}^\circ(\check{P}_\zeta)$ as claimed. ■

3. Example

We consider a variant of the example of Al-Sabban et al. (2013) in which a UAV (unmanned aerial vehicle) needs to reach a target subject to energy costs, and subject to disturbances from wind. The location of the UAV at time t is denoted L_t, and evolves according to the controlled linear dynamics:

$$L_{t+1} = L_t + W_t + U_t \tag{35}$$

where $\boldsymbol{U} = \{U_t\}$ is the control sequence, and $\boldsymbol{W} = \{W_t\}$ models the impact of the wind. There are d_L locations across a two-dimensional grid.

Wind is location-dependent: It is assumed that the wind profile over the region is determined by a stochastic process $\boldsymbol{N} = \{N_t\}$ and a function ω such that for each t,

$$W_t = \omega(L_t, N_t).$$

The process $\boldsymbol{N}$ is assumed to be Markovian with finite state space $\{1, \dots, d_N\}$, and state transition matrix denoted Q_0. This is the nature component of the MDP model, with state process $X_t = (L_t, N_t)$, $t \geqslant 0$.

A nominal model is described by a randomized policy in which $U_t = 0$ with high probability. The specific form used in the experiments was constructed as follows. On denoting $L_t^+ = L_t + W_t$, a transition matrix R_0^L is constructed with the interpretation

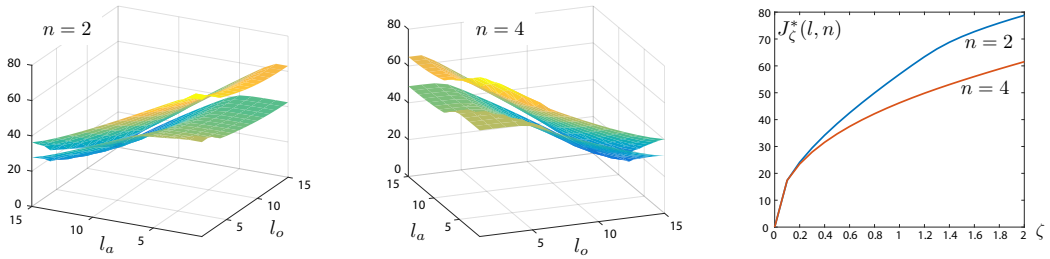
$$R_0^L(l^+, l') = \mathsf{P}\{L_{t+1} = l' \mid L_t^+ = l^+\}, \quad t \geqslant 0.$$

The nominal randomized strategy is the $d_L \times d_L$ matrix,

$$R_0(x, u) = \mathsf{P}\{U_t = u \mid X_t = x\} = R_0^L(l + \omega(l, n), l + \omega(l, n) + u), \qquad x = (l, n).$$

The overall transition matrix is the product:

$$P_0(x, x') = R_0^L(l + w(n, l), l')Q_0(n, n'), \qquad x = (n, l),\ x' = (n', l').$$



Surface Plots of $J^*_\zeta(\cdot, n)$, $\zeta = 1, 2$ Plots of $J^*_\zeta(l, n)$, $l = (1, 1)$

Figure 2: Cost to go for two values of the initial value $n = N_0$, $n = 2, 4$. Each surface plot indicates values of $J^*_\zeta(\cdot, n)$ for $\zeta = 1$ and $\zeta = 2$. The one of larger magnitude corresponds to $\zeta = 2$. The plot at the right shows $J^*_\zeta(l, n)$ as a function of ζ for these values of n, and $l = (1, 1)$.

The goal of the control problem is to reach a target location $l^\bullet$ and remain there. To ensure that the set $\{(l^\bullet, n) : 1 \leqslant n \leqslant d_N\}$ is absorbing, a separate rule is imposed on R_0 for these states: $W_t + U_t = 0$ if $L_t = l^\bullet$.

The reward function $\mathcal{U}$ is taken to be a scaled negative cost: $\mathcal{U} = -c$, where $c\colon \mathsf{X}^L \to \mathbb{R}_+$, with $c(l^\bullet) = 0$ and $c(l) > 0$ for $l \neq l^\bullet$. The optimal steady-state mean is zero in this model, and the relative value function is the negative of the cost to go:

$$-h^*(x) = J^*(x) := \min \mathsf{E}_x\Big[\sum_{t=0}^{\tau_\bullet} \big\{\zeta c(L_t) + c_{\text{KL}}(X_t, R_t)\big\}\Big] \tag{36}$$

where $\tau_\bullet$ (unknown a-priori) is the first hitting time to $l^\bullet$. An example is illustrated in Fig. 2 — the details are provided in the following.

Details of the numerical experiment The set of locations X^L is taken to be a rectangular grid of the form $\mathsf{X}^L = \{(i, j) : 1 \leqslant i \leqslant d_a,\ 1 \leqslant j \leqslant d_o\}$, in which $d_a, d_o \geqslant 2$ and $d_L = d_a \times d_o$ (the subscripts are meant to suggest latitude and longitude). The function c appearing in (36) was taken to be the indicator function, $c(l) = \mathbb{I}\{l \neq l^\bullet\}$, with $l^\bullet = (d_a, d_o)$.

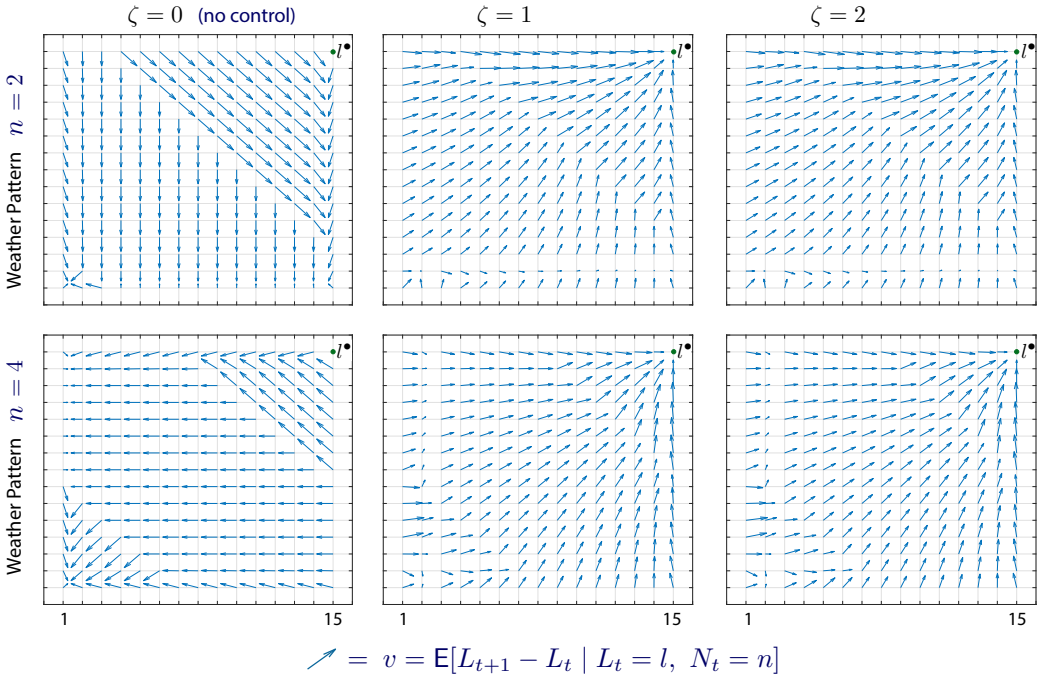


Figure 3: Vector field $v(l, n)$ for two values of n, and $\zeta = 0, 1, 2$: see eqn. (37)

The values $d_a = d_o = 15$, and $d_N = 5$ are fixed throughout. The size of the state space is thus $d_a \times d_o \times d_N = 1,125$, and the action space is a subset of the simplex in $\mathbb{R}^{1,125}$.

The transition matrix for nominal control was taken of the following form:

$$R_0^L(l, l') = \kappa(l) \exp\Big\{-\frac{1}{2\sigma_u^2}\|l' - l\|^2\Big\}, \quad l, l' \in \mathsf{X}^L,$$

where $\kappa(l) > 0$ is chosen so that $R_0^L(l, \cdot\,)$ is a pmf on X^L for each $l \in \mathsf{X}^L$. The value $\sigma_u^2 = 1/2$ was used in the numerical results that follow.

The Markov chain $\boldsymbol{N}$ was taken to be a skip-free symmetric random walk on the integers $\{1, \ldots, d_N\}$. For a fixed $\delta_n \in (0, 1)$ the probability of transition is $Q_0(n, n+1) = Q_0(n, n-1) = \frac{1}{2}\delta_n$, where addition is modulo d_N, and $Q_0(n, n) = 1 - \delta_n$ for any n. Recall that this means

$$\mathsf{P}\{N_{t+1} = n + 1 \mid N_t = n\} = \mathsf{P}\{N_{t+1} = n - 1 \mid N_t = n\} = \tfrac{1}{2}\delta_n.$$

The value $\delta_n = 0.05$ was chosen in these experiments.

Recall that $\omega \colon \mathsf{X}^L \to \mathbb{Z}^2$ is used to defined the wind process $\boldsymbol{W}$. For each value of n, the function $\omega(\,\cdot\,, n)$ can be interpreted as a vector field on X^L. For each n, a slowly varying continuous function was constructed on the two-dimensional rectangle $[1, d_a] \times [1, d_o]$. The function $\omega(\,\cdot\,, n)$ was taken to be its quantization to the lattice X^L. The values were restricted to the set of pairs $\{(i, j) : |i| \leqslant 1,\ |j| \leqslant 1\}$.

The family of optimal policies was obtained using the ODE method, and the solution for three values of ζ is illustrated in Fig. 3. Each of the arrows shown is proportional to the conditional expectation:

$$v(l, n) := \mathsf{E}[L_{t+1} - L_t \mid L_t = l\ \ N_t = n] \tag{37}$$

in which $l \in \mathsf{X}^L$ is the position on the grid. The figure shows only the values $n = 2$ and $n = 4$ (the most interesting to view because of obvious spatial variability).

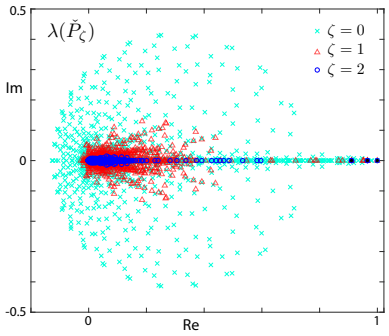


Figure 4: Eigenvalues of $\check{P}_\zeta$

If the position $l = (l_a, l_o)$ is far from the boundary of X^L, say, $\min(l_a, l_o) \geqslant 4$ and $\min(d_a - l_a, d_o - l_o) \geqslant 4$, then

$$\mathsf{E}[U_t \mid L_t = l \;\; N_t = n] \approx 0 \text{ and } v(l, n) \approx \omega(l, n), \quad \zeta = 0$$

For the case $\zeta = 1$ the vector field is transformed so that vectors near the target state point in this direction; for $\zeta = 2$ this behavior is more apparent. For states far from the target the control effort seems to be lower – most likely the optimal policy waits for more favorable weather that will push the UAV in the North-East direction.

The eigenvalues of $\check{P}_\zeta$ are shown in Fig. 4 for $\zeta = 0, 1, 2$. Most of the eigenvalues are driven near zero for $\zeta = 2$. Those three that are independent of ζ are the three eigenvalues of Q_0, $\{0.9095, 0.9655, 1\}$.

While the vector field and eigenvalues change significantly when ζ is doubled from 1 to 2, the cost to go J^* defined in (36) grows relatively slowly with ζ. Shown on the right hand side of Fig. 2 are comparisons for these two values of ζ. One plot with $n = 2$ and the other $n = 4$. The plot on the far right shows $J^*_\zeta(l, \zeta)$ for $0 \leqslant \zeta \leqslant 1$ and $l = (1, 1)$ (the location farthest from $l^\bullet$).

These plots are easily obtained because of the nature of the algorithm: the optimal policy and value function are generated for any range of ζ of interest.

4. Conclusions

The ODE approach for solving MDPs has simple structure for the class of models considered in this paper. We are currently looking at approaches to approximate dynamic programming as has been successful in the unconstrained model Todorov (2009).

It is likely that the ODE has special structure for other classes of MDPs, such as the "rational inattention" framework of Sims (2006); Shafieepoorfard et al. (2016). The computational efficiency of this approach will depend in part on numerical properties of the ODE, such as its sensitivity for complex models. Applications to distributed control were the original motivation for this work, with particular attention to "demand dispatch" Chen et al. (2017). It is believed that this paper will offer new computational tools in this ongoing research.

Acknowledgments

Funding from the ANR under grant ANR-16-CE05-0008, and NSF under awards EPCN 1609131, CPS 1646229 is gratefully acknowledged.

References

W. H. Al-Sabban, L. F. Gonzalez, and R. N. Smith. Wind-energy based path planning for unmanned aerial vehicles using Markov Decision Processes. In *Proc. IEEE Conf. Robotics and Automation (ICRA)*, pages 784–789. IEEE, 2013.

D. P. Bertsekas and S. E. Shreve. *Stochastic Optimal Control: The Discrete-Time Case*. Athena Scientific, 1996.

A. Bušić and S. Meyn. Ordinary differential equation methods for Markov decision processes and application to Kullback–Leibler control cost. *SIAM Journal on Control and Optimization*, 56(1): 343–366, 2018.

Y. Chen, U. Hashmi, J. Mathias, A. Bušić, and S. Meyn. Distributed control design for balancing the grid using flexible loads. In *IMA volume on the control of energy markets and grids*. Springer, 2017.

A. Dembo and O. Zeitouni. *Large Deviations Techniques And Applications*. Springer-Verlag, New York, second edition, 1998.

K. Doya. How can we learn efficiently to act optimally and flexibly? *Proceedings of the National Academy of Sciences*, 106(28):11429–11430, 2009.

G. F. Franklin, M. L. Workman, and D. Powell. *Digital Control of Dynamic Systems*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 3rd edition, 1997.

P. Guan, M. Raginsky, and R. M. Willett. Online Markov decision processes with Kullback-Leibler control cost. *IEEE Trans. Automat. Control*, 59(6):1423–1438, June 2014.

M. Kárný. Towards fully probabilistic control design. *Automatica*, 32(12):1719 –1722, 1996.

I. Kontoyiannis and S. P. Meyn. Spectral theory and limit theorems for geometrically ergodic Markov processes. *Ann. Appl. Probab.*, 13:304–362, 2003.

S. Meyn, P. Barooah, A. Bušić, Y. Chen, and J. Ehren. Ancillary service to the grid using intelligent deferrable loads. *IEEE Trans. Automat. Control*, 60(11):2847–2862, Nov 2015.

M. L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

P. J. Schweitzer. Perturbation theory and finite Markov chains. *J. Appl. Prob.*, 5:401–403, 1968.

E. Shafieepoorfard, M. Raginsky, and S. P. Meyn. Rationally inattentive control of Markov processes. *SIAM J. Control Optim.*, 54(2):987–1016, 2016.

C. A. Sims. Rational inattention: Beyond the linear-quadratic case. *The American economic review*, pages 158–163, 2006.

R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, NIPS'99, pages 1057–1063, Cambridge, MA, USA, 1999. MIT Press.

E. Todorov. Linearly-solvable Markov decision problems. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 1369–1376. MIT Press, Cambridge, MA, 2007.

E. Todorov. Efficient computation of optimal actions. *Proceedings of the National Academy of Sciences*, 106(28):11478–11483, 2009.