Mori-Zwanzig Representation of Partially-Observable Markov Decision Processes

Debnil Sur*
Department of Computer Science
Stanford University
Stanford, CA 95039
debnil@stanford.edu

Mykel J. Kochenderfer
Department of Aeronautics and Astronautics
Stanford University
Stanford, CA 95039
mykel@stanford.edu

Abstract

Model reduction significantly improves the computational tractability of high-dimensional problem spaces. Mori-Zwanzig theory provides one such method that has been theoretically outlined for Hidden Markov Models. In this paper, we extend this formalism to Partially Observable Markov Decision Processes, a setting with control inputs. We establish the theoretical setting and define projections for state and belief updates using the smaller observation space. Finally, we outline experiments to measure the theoretical accuracy and computational efficiency of this procedure.

1 Markov Chain Model Reduction

Model reduction methods have been discussed across myriad computationally-oriented fields; examples include economics [11], speech and signal processing [7], and statistical mechanics [2, 3], to name a few. These efforts seek to find a simple mathematical model that adequately represents the behavior of a given complex system. An approach is then judged upon the computational complexity of implementation, amount of reduction in dimension, ease of calculation, and ability to capture given system behavior to some desired level. This paper focuses on model reduction from a controls and dynamical systems perspective. In particular, we attempt to reduce partially observable Markov decision processes (POMDPs) at a large scale in complex systems. We first situate this work in research in this area.

Most model complexity in a POMDP comes from the underlying Markov chain. We will thus begin by discussing approaches to reduce the size of these chains, primarily from control theory.

One of the most common approaches uses the system's decomposability to aggregate states into meta-states [11]. In this approach, the individual states have dynamics which evolve along a similar short-run time scale. Each meta-state is an aggregation of individual states, such that every state is represented in some meta-state. The interactions between these aggregated states evolve on a long-run time scale. This approach provides a basis for singular perturbation methods [10]. Another state aggregation approach is directly related to lumpability [5]. As in the first approach, the chain's states are partitioned into aggregated sets. But this approach focuses on ensuring that these larger sets exhibit similar dynamics and observation statistics. A third approach uses spectral graph partitioning methods to analyze multi-scale behavior [4]. This views the Markov model as a graph with edge weights defined by the entries of the accompanying transition matrix. The graph can then be partitioned into invariant subgraphs to determine the level of interaction between the aggregated states.

*Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

Researchers in statistical mechanics have also developed their own approaches to model reduction—namely, the Mori-Zwanzig theory [12, 9]. Here, a system is described in terms of microstates on a fine scale. Macrostates are then observable quantities characterized by a probability distribution over some group of microstates. Given dynamics on the microstates, this theory then studies the generated dynamics on the macrostates.

Much work remains on model reduction in both fields. Error bounds on many of these reduction processes have yet to be determined, though some such methods have been suggested in the context of hidden Markov models (HMMs) [6, 8].

This paper reviews existing work by Beck et al. (2009) that has cast reduction results from existing fields in a form useful for applications to hidden Markov models [1]. We take this work and apply it to partially observable Markov decision processes—a version of this model with multiple actions and rewards. We then discuss this simplified form in the context of some well-studied sample problems in POMDPs.

Such reduction procedures can save significant memory and computational time and power, with minimal loss of accuracy. POMDPs are used in many different intelligent infrastructures that must plan under uncertainty, such as robot navigation and aircraft collision avoidance. In particular, these decision processes typically have very large, sparse transition matrices representing the state space and significantly smaller observation spaces. Transforming the model from the state space to observation space can thus significantly reduce model size and complexity. The theories above represent such an approach. Thus, even marginal reductions in model complexity can yield significant performance improvements in the algorithms underlying these and other systems.

2 Partially Observable Markov Decision Processes

Markov decision processes (MDPs) provide a mathematical framework for modeling decision-making in situations where outcomes are partly random and partly controlled by a decision maker. POMDPs extend this agent decision process and assume that system dynamics are determined by an MDP, but the agent cannot directly observe the underlying state. Instead, it must maintain a probability distribution over all possible states using a set of observations, their probabilities, and the underlying MDP.

This framework originated in operations research and then spread to artificial intelligence and automated planning. Current popular applications include robot navigation, machine maintenance, and general planning under uncertainty. An exact solution to this problem yields the optimal action for each possible belief over the states. Such an action would maximize the reward or minimize the cost of the agent over some (possibly infinite) horizon of action. The sequence of optimal actions is called the optimal policy for the agent acting in the environment.

2.1 Formalism

Let us formalize this problem as follows. Consider a finite state space $S \in \mathbb{Z}^n = \{s_0, s_1, ..., s_n\}$, a finite observation space $O \in \mathbb{Z}^m = \{o_0, o_1, ..., o_m\} (m \leq n)$, and a finite action space $A \in \mathbb{Z}^{\text{l}} = \{a_0, a_1, ..., a_p\}$. Here, there are n states the underlying MDP can take; m possible observed states, which constitute the distribution of beliefs; and p actions that can be taken. Functions on these spaces are then defined on $C_S = \mathbb{R}^n$, $C_O = \mathbb{R}^m$, and $C_A = \mathbb{R}^p$ respectively. Distributions are defined on subsets of the respective dual spaces $C_S^* = \mathbb{R}^{n*}$, $C_O^* = \mathbb{R}^{m*}$, and $C_A^* = \mathbb{R}^{p*}$.

POMDPs provide a rigorous formulation for a stochastic optimal control process. At each time step, the environment is in some state $s \in S$. The agent takes an action $a \in A$. The environment then transitions to a new state $s' \in S$ with some probability $Pr[s'|s, a]$. The agent then makes an observation $o \in O$ with probability $Pr[o|s', a]$. This observation represents the agent's belief of its current state; it is used to determine the next action.

States and actions both have the Markov property. The next state depends only on the state and action of the current time step. Mathematically, $Pr[s(t+1) = s'|s(0) = s_0, s(1) = s_1, ..., s(t) = s_t, a(t) = a_t] = Pr[s(t+1) = s'|s(t) = s_t, a(t) = a_t]$.

Similarly, the next observation depends only on the next state and the action taken. $Pr[o(t+1) = o'|s(0) = s_0, a(0) = a_0, o(0) = o_0, ..., s(t) = s_t, a(t) = a_t, o(t) = o_t, s(t+1) = s_{t+1}] = Pr[o(t+1) = t'|s(t+1) = s_{t+1}, a(t) = a_t]$.

At each step, the agent receives a reward $R(s, a, s')$. It also acts on a horizon H, such that we consider actions on the time steps $0 \leq t \leq H$.

2.2 Transitions and Beliefs

To account for the stochastic nature of this system, we define two sets of matrices: $\{T\}$ and $\{B\}$. Each action has a corresponding transition matrix and observation matrix.
For each action $a' \in A$, let $P = T^{a'} \in \mathbb{R}^{n \times n}$. Then, $P_{ij} = Pr[s(t+1) = s_j|s(t) = s_i, a(t) = a']$. We fix the action a' that we are considering at this time step. The i^{th} row then becomes the distribution over the next state for the current state s_i.
Similarly, for each action $a' \in A$, let $Q = B^{a'} \in \mathbb{R}^{m \times n}$. Then, $Q_{ij} = Pr[o(t+1) = o_j|s(t+1) = s_i, a(t) = a']$. Again, we fix the action a' that we are considering at this time step. Row i is the distribution over the next observed state for the possible next state s_i.

2.3 Functions and Distributions

Finally, we establish some properties of the evolution of functions and distributions on these spaces. Again, we fix an action a' taken at some time t and consider the consequent evolution of the system under the fixed transition matrix $P = T^{a'}$. Distributions evolve according to $d(t+1) = P^T d(t)$, and functions evolve on S by $g(t) = Pg(t+1)$. This thus satisfies a natural duality pairing: $g(t+1)^T d(t+1) = g(t)^T d(t)$. Furthermore, we assume each such P is irreducible and positive recurrent. This means there exists a unique invariant distribution π such that $P^T \pi = \pi$, $\pi > 0$. Finally, we assume each $Q = B^{a'}$ has no zero columns, so $(Q^T \pi)_i > 0$ for all i.

3 Optimal Prediction

The theory of optimal prediction studies stochastic dynamics on the observables of a system. An example is the numerical simulation of fluid dynamics: the coarse observables are the values of the state on various grid points, but subgrid information has been lost. This field focuses on how to update coarse variables given their time history and statistical information about the unobserved fine-scale model [2, 3]. In this section, we explicitly the derivation of a reduced form for HMMs found in Beck et al (2009) to derive an analogous formulation for POMDPs.

3.1 Function Evolution

We wish to define the evolution of a function g on the Markov chain. Though generalized, we can interpret this as the reward function that we seek to maximize. In this case, the study of optimal prediction and control will give us a reduced model for a greedy approach—one in which the agent acts to maximize reward at each time-step, which may not necessarily be globally optimal.

Suppose we have taken action a' at time t. Let $P = T^{a'}$ be the corresponding transition matrix and $Q = B^{a'}$ the observation matrix. Then, $(Pg)_i = E[g(s(t+1))|s(t) = s_i]$.

By definition, Pg is the optimal predictor of $g(s(t+1))$ that depends only on $s(t)$.

Equivalently, $Pg = \text{argmin}_h E[(h(s(t)) - g(s(t+1)))^2]$. It is the function that minimizes the squared loss between its value at the current state and the reward of the next state.

3.2 Reduced Chain Representation

We wish to find a reduced chain $\bar{P}$ defined on the observations o. Recall that this is defined for a specific action and therefore has a specific belief matrix $Q = B^{a'}$.

Now, consider a function f that evolves on the observation space. We take an optimal prediction given by $\bar{P}f = \lim_{t \to \infty} \text{argmin}_h E[(h(o(t)) - f(o(t+1)))^2]$. As before, we minimize the squared loss between a function of the current observation and some fixed function of the next observation.

Equivalently, $(\bar{P}f)a = \lim_{t\to\infty} E[f(o(t+1))|o(t) = o_a]$.

Then, $\bar{P}_a b = \lim_{t\to\infty} Pr[o(t+1) = o_b|o(t) = o_a]$, from the definition of expectation. We can thus compute the explicit formula $\bar{P}_a b = \frac{\sum_{i,j} \pi_i P_{ij} Q_{ia} Q_{jb}}{\sum_i \pi_i Q_{ia}}$.

We can use this formulation to compute the reduced form $\bar{P}$ for each $P = T^{a'}$ using its corresponding $Q = B^{a'}$. Note that this is simply one of many reduced chains.

3.3 Operator Representation

We will also utilize the operator representation from statistical mechanics to help us derive the desired Mori-Zwanzig representation. Let $\Psi_S g = \lim_{t\to\infty} \text{argmin}_f E[(f(o(t) - g(s(t)))^2]$ be the best predictor of $g(s)$ that depends only on o. In row-wise form, this becomes $(\Psi_S g)_i = \lim_{t\to\infty} E[g(s(t)|o(t) = o_i]$.
Similarly, let $\Psi_O f = \lim_{t\to\infty} \text{argmin}_g E[(g(s(t)) - f(o(t)))^2$ be the best predictor of $f(o)$ that depends only on s. In row-wise form, this becomes $(\Psi_Y f)_i = \lim_{t\to\infty} E[f(o(t))|s(t) = s_i]$. These limits exist if the chain is irreducible and aperiodic.

We now derive an alternate representation for the reduced chain found in the prior section. Fix $P = T^{a'}$, the transition matrix for a certain action. Then, $(\Psi_S P \Psi_Y f)_i = \lim_{t\to\infty} E[P\Psi_O f(s(t+1))|o(t+1) = o_i]$, by definition of Ψ_S. Applying P^{-1} gives the equivalent form $\lim_{t\to\infty} E[\Psi_O f(s(t))|o(t+1) = o_i]$. By definition of Ψ_O, this becomes $\lim_{t\to\infty} E[f(o(t))|o(t+1) = o_i]$. Hence, $\bar{P} = \Psi_S P \Psi_O$. This provides an alternate derivation for the reduced form that is more helpful in deriving the Mori-Zwanzig representation.

To compute an explicit form for this implicit expression, we must first compute explicit forms for Ψ_S and Ψ_O. Beck et al (2009) provide the explicit computation of $\Psi_S = M_O^{-1} Q^T M_S$ and $\Psi_O = Q$, where $M_S = \text{diag}(\pi)$ and $M_O = \text{diag}(Q^T\pi)$ are the induced inner products on C_S and C_O, respectively. This gives us the reduced transition matrix $\bar{P} = M_Y^{-1} Q^T M_S P Q$, which is the matrix form of the conditional probability computed before. We define $\bar{T}$ as the set of all reduced transition matrices $\bar{P}$ for all actions.

4 Mori-Zwanzig Representation

4.1 State Evolution

Above we found the best Markov model $\bar{T}$ on the observation space O for the partially observable Markov decision process T on the state space S. We decompose this set into $\bar{P}_t$, We can regard this reduced model $\bar{P}$ as the leading-order term in a representation of the evolution on O as a non-Markovian system. This is known as the Mori-Zwanzig representation and constructed as follows for a linear system with fixed operators defining transitions between states.

Theorem 4.1: Define linear operators $A \in R^{n\times n}$, $C \in R^{m\times n}$, and $R \in R^{n\times m}$. Define a linear system by $x(t+1) = Ax(t)$ with measurements $y(t) = Cx(t)$ and reconstruction $\hat{x}(t) = Ry(t)$. Then, $y(t+1) = CARy(t) + \sum_{k=0}^{t-1} C(A(I - RC))^{t-k} ARy(k) + C(A(I - RC))^{t+1} x(0)$.

Proof: Write $x(t) = RCx(t) + (I - RC)x(t) = Ry(t) + (I - RC)Ax(t-1)$. Then $y(t+1) = CAx(t) = CARy(t) + CA(I - RC)Ax(t-1)$. We can recursively substitute $x(k)$ for $k = (t-1), ..., 0$.

Now, we consider the stochastic case of a POMDP, in which each different time step can have a different action.

Theorem 4.2: Consider a single time step t. As before, we define linear operators $A_t \in R^{n\times n}$, $C_t \in R^{m\times n}$, and $R_t \in R^{n\times m}$. At this time step, the linear system evolves by $x(t+1) = A_t x(t)$ with measurements $y(t) = C_t x(t)$ and reconstruction $\hat{x}(t) = R_t y(t)$. Then, $y(t+1) = C_{t+1} A_t R_t y(t) + \sum_{k=0}^{t-1} C_{t+1} (\prod_{i=0}^{k} A_{t-i}(I - R_{t-i}C_{t-i})) A_k R_k y(k) + C_t (\prod_{k=0}^{t} A_{t-k}(I - R_{t-k}C_{t-k}))x(0)$.

Proof: We begin as in the previous proof. $x(t) = R_t C_t x(t) + (I - R_t C_t)x(t) = R_t y(t) + (I - R_t C_t)A_{t-1}x(t-1)$. Then, $y(t+1) = C_{t+1}x(t+1) = C_{t+1}A_t x(t) = C_{t+1}A_t(R_t y(t) + (I -$

$R_tC_t)A_{t-1}x(t-1))$. Now, $x(t-1) = R_{t-1}y(t-1) - (I - R_{t-1}C_{t-1})A_{t-2}x(t-2)$. We can continue the recursive substitution process of $x(k)$ for $k = (t-1), ..., 0$. This derives the closed form above.

To understand the relationship between the maps for optimal prediction and the reduction in Mori-Zwanzig, we consider the evolution of the state function $s(t)$ on the observation space C_O. The distribution $b(t)$ represents the observer's belief of the current state, represented as a probability distribution over all states. We wish to decompose the evolution of this belief distribution. Let us fix an action a, and let P be the accompanying transition matrix, with unique invariant π, and Q the observation matrix. Then, we have the evolution of the linear system as $s(t+1) = Ps(t)$, the measurement $o(t) = Qs(t)$, and reconstruction $\hat{s}(t) = Ro(t)$.

Above, we defined the optimal maps $\Psi_S = \text{diag}(Q^T\pi)^{-1}Q^T\text{diag}(\pi)$ and $\Psi_O = Q$.

Since we are considering the evolution of a distribution on O, we have the maps $A = \bar{P}^T$, $C = \Psi_O^T$, and $R = \Psi_S^T$.

This gives us the evolution of $b(t)$ as $b(t+1) = \bar{P}b(t) + \sum_{k=0}^{t-1} \Psi_{S_{t+1}}(\prod_{i=0}^{k} P_{t-i}(I - \Psi_{S_{t-i}}\Psi_{O_{t-i}}))P_k\Psi_{S_k}o(k) + \Psi_{O_t}(\prod_{k=0}^{t} P_{t-k}(I - \Psi_{S_{t-k}}\Psi_{O_{t-k}}))s(0)$, where $s(0)$ is the initial state.

The leading-order term here is the optimal predictor chain $\bar{P}$, the reduced form of the transition matrix for the action a_t. The second and third terms are called the "history" and "noise" terms, representing the contributions from past observations and the unobservable components of initial state, respectively.

We can utilize this expression to test the computational efficiency and accuracy of the Mori-Zwanzig representation for a POMDP. To do this, we can numerically evaluate a policy, consisting of a sequence of actions over discrete time steps, and compare the process with the traditional method of state evolution.

4.2 State Prediction

We can also view this result from the perspective of state prediction, using the observer equations. In these equations, the state estimate $\hat{s}$ evolves by $\hat{s}(t+1) = A\hat{s}(t) + R(o(t+1) - CA\hat{s}(t)$. In this stochastic setting, that becomes $\hat{s}(t+1) = A_t\hat{s}(t) + R_{t+1}(o(t+1) - C_{t+1}A_ts(t))$.

To connect this with the Mori-Zwanzig representation, this update rule can be rewritten as $\hat{s}(t+1) = (I - R_{t+1}C_{t+1})A_t\hat{s}(t) + R_{t+1}o(t+1)$, which stochastically defines a convolution map from the sequence of observations $o(1), ..., o(t)$ to $\hat{s}(t)$. We also define the prediction $\hat{y}(t+1) = C_{t+1}A_tx(t)$. Recursively substituting the convolved update rule yields the stochastic Mori-Zwanzig equation defined above.

5 Experiments

The ultimate goal of a POMDP is to efficiently calculate the optimal policy for an agent. Algorithms for this computation fall into two groups: policy iteration and value iteration.

In the former, the algorithm starts with an arbitrary policy, computes its utility, and computes an improved policy. It terminates when there is no more improvement. Since there are finitely many policies, the algorithm terminates at an optimal solution.

In the latter, we first define b, the current belief distribution as above. Given a belief at a single time step, we attempt to solve the Bellman optimality equation: $V^*(b) = \max_{a \in A}[r(b,a) + \gamma \sum_{o \in O} Q(o|b,a)V^*(\tau(b,a,o))]$, where τ represents the update of belief b based on action a and observation o. Value iteration applies dynamic programming to update these values until it converges to an optimal value function, where the criteria for optimality can be determined and bounded accordingly.

We propose experimental design to test our formalism in each setting. For both, at least three different POMDP problems should be used, with small, medium, and very large state spaces. We expect that this formalism performs noticeably worse on smaller problems. However, as the size of the state space increases exponentially more than the accompanying belief space, the formalism's inherent

advantage—the transformation of a large, sparse transition matrix to a small, dense observation matrix—will yield computational benefits and comparable accuracy.

To test the first, we propose utilizing a traditional solver to generate the optimal policy for a problem. Then, using the equations outlined above for state and belief update, we execute the policy and compare the performance of the Mori-Zwanzig form to traditional updating methods. Their difference can be measured in two ways. First, we can compute the accuracy of the belief state through, for example, the mean K-L divergence between the one-hot vector representing the true state and the belief state over all time steps. Second, we can use built-in timer functions to measure the computational speed of each approach. This setting can easily be extended to policy iteration by computing the utility of a policy, changing the action, and then continuing till convergence.

To test the second, we propose writing a value-iteration solver using the Mori-Zwanzig belief update rule for the update function τ. This solver can be tested against existing solvers, and its performance can be measured using the two criteria outlined for policy iteration. To test speed, we can time the performance of the algorithm. To test accuracy, we can measure whether or not the execution of the policy returns an optimal value.

6 Conclusions

In this paper, we extend the Mori-Zwanzig theory from Hidden Markov Models to Partially Observable Markov Decision Processes and provide formulas for state evaluation and prediction. To verify this theoretical extension, we outline experimental designs for both policy iteration and value iteration that can test the computational speed and accuracy of this representation against more traditional methods.

Future work can immediately build upon this through the implementation of the experimental procedures outlined above. This will demonstrate the efficiency and accuracy of this formalism. Additional efforts in this field can focus on other methods for state reduction, such as predictive state representations, and compare their performance with this formalism.

References

[1] Carolyn L Beck et al. "Model reduction, optimal prediction, and the Mori-Zwanzig representation of Markov chains". In: *Decision and Control, 2009 held jointly with the 2009 28th Chinese Control Conference. CDC/CCC 2009. Proceedings of the 48th IEEE Conference on*. IEEE. 2009, pp. 3282–3287.

[2] Alexandre J Chorin, Ole H Hald, and Raz Kupferman. "Optimal prediction and the Mori–Zwanzig representation of irreversible processes". In: *Proceedings of the National Academy of Sciences* 97.7 (2000), pp. 2968–2973.

[3] Alexandre Joel Chorin and Ole H Hald. *Stochastic tools in mathematics and science*. Vol. 3. Springer, 2009.

[4] Miroslav Fiedler. "A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory". In: *Czechoslovak Mathematical Journal* 25.4 (1975), pp. 619–633.

[5] John G Kemeny, James Laurie Snell, et al. *Finite markov chains*. Vol. 356. van Nostrand Princeton, NJ, 1960.

[6] Georgios Kotsalis, Alexandre Megretski, and Munther A Dahleh. "Balanced truncation for a class of stochastic jump linear systems and model reduction for hidden Markov models". In: *IEEE Transactions on Automatic Control* 53.11 (2008), pp. 2543–2557.

[7] Lahcène Mitiche, Amel BH Adamou-Mitiche, and Daoud Berkani. "Low-order model for speech signals". In: *Signal processing* 84.10 (2004), pp. 1805–1811.

[8] Bruce Moore. "Principal component analysis in linear systems: Controllability, observability, and model reduction". In: *IEEE transactions on automatic control* 26.1 (1981), pp. 17–32.

[9] Hazime Mori. "Transport, collective motion, and Brownian motion". In: *Progress of theoretical physics* 33.3 (1965), pp. 423–455.

[10] R Phillips and P Kokotovic. "A singular perturbation approach to modeling and control of Markov chains". In: *IEEE Transactions on Automatic Control* 26.5 (1981), pp. 1087–1094.

[11] Herbert A Simon and Albert Ando. “Aggregation of variables in dynamic systems”. In: *Econometrica: journal of the Econometric Society* (1961), pp. 111–138.

[12] Robert Zwanzig. “Problems in nonlinear transport theory”. In: *Systems far from equilibrium*. Springer, 1980, pp. 198–225.