Metrics for Finite Markov Decision Processes

Norm Ferns
School of Computer Science
McGill University
Montréal, Canada, H3A 2A7
(514)398-7071 ext 09183
nferns@cs.mcgill.ca

Prakash Panangaden
Doina Precup
School of Computer Science
McGill University
Montréal, Canada, H3A 2A7
prakash@cs.mcgill.ca
dprecup@cs.mcgill.ca

Markov decision processes (MDPs) offer a popular mathematical tool for planning and learning in the presence of uncertainty (Boutilier, Dean, & Hanks 1999). MDPs are a standard formalism for describing multi-stage decision making in probabilistic environments. The objective of the decision making is to maximize a cumulative measure of long-term performance, called the *return*. Dynamic programming algorithms, e.g., value iteration or policy iteration (Puterman 1994), allow us to compute the optimal expected return for any state, as well as the way of behaving (policy) that generates this return. However, in many practical applications the state space of an MDP is simply too large, possibly even continuous, for such standard algorithms to be applied. A typical means of overcoming such circumstances is to partition the state space in the hope of obtaining an "essentially equivalent" reduced system. One defines a new MDP over the partition blocks, and if it is small enough, it can be solved by classical methods. The hope is that optimal values and policies for the reduced MDP can be extended to optimal values and policies for the original MDP.

The notion of equivalence for stochastic processes is problematic because it requires that the transition probabilities agree *exactly*. This is not a robust concept, especially considering that usually, the numbers used in probabilistic models come from experimentation or are approximate estimates; what is needed is a quantitative notion of equivalence. In our work we provide such a notion via semimetrics, distance functions on the state space that assign distances quantifying "how equivalent" states are. These semimetrics could potentially be used as a new theoretical tool to analyze current state compression algorithms for MDPs, or in practice to guide state aggregation directly. The ultimate goal of this research is to efficiently compress and analyze continuous state space MDPs. Here we focus on finite MDPs, but note that most of our results should hold, with slight modifications, in the context of continuous state spaces.

Recent MDP research on defining equivalence relations on MDPs (Givan, Dean, & Greig 2003) has built on the notion of strong probabilistic bisimulation from concurrency theory. Bisimulation was introduced by Larsen and Skou (1991) based on ideas of Park (1981) and Milner (1980). Roughly speaking, two states of a process are deemed equivalent if all the transitions of one state can be matched by transitions of the other state, and the results are themselves bisimilar. The extension of bisimulation to transition systems with rewards was carried out in the context of MDPs by Givan, Dean and Greig (2003). Suppose $M = (S, A, \{r^a_s | s \in S, a \in A\}, \{P^a_{ss'}\})$ is a given finite MDP consisting of a finite state space, a finite action space, numerical rewards (for convenience assumed constrained to the unit interval), and transition probabilities, respectively. *Stochastic bisimulation* is the largest relation on S that satisfies the following property: it is an equivalence relation and states are equivalent precisely when for each action they share the same expected immediate rewards and 1-step transition probabilities to equivalence classes, i.e. $s \sim s' \iff \forall a \in A.(r^a_s = r^a_{s'}$ and $\forall C \in S/\sim .(P^a_s(C) = P^a_{s'}(C)))$. Here $\sim$ denotes the bisimulation equivalence relation and $S/\sim$ denotes the resulting quotient space.

We quantify the notion of bisimulation in terms of bisimulation semimetrics (henceforth "metrics"), which assign distance zero to states precisely when those states are bisimilar. Our goal is to construct a class of bisimulation metrics useful for MDP state compression. Specifically, we require such metrics to be easily computable and to provide information concerning the optimal values of states. However, it is not hard to show that the bisimulation metric that assigns distance 1 to states that are not bisimilar satisfies both requirements, while possessing no more distinguishing power than that of bisimulation itself. So we additionally require that metric distances should vary smoothly and proportionally with differences in rewards and differences in probabilities. Formally, we will construct bisimulation metrics via a metric on rewards and a metric on probability functions (called a probability metric). The choice of metric on rewards is an obvious one: we simply use the usual Euclidean distance. However, there are many ways of defining useful probability metrics (Gibbs & Su 2002). Two of the most important for our purposes are the Kantorovich metric and the total variation metric. They lead to two classes of bisimulation metrics.

Our original bisimulation metrics use the Kantorovich metric, whose origins lie in the theory of mass transportation. Suppose d is a metric on S. Consider two copies of the state space, one in which states are labeled as supply nodes,

Copyright © 2004, American Association for Artificial Intelligence (www.aaai.org). All rights reserved.

and the other in which states are labeled as demand nodes. Each supply node has a supply whose value is equal to the probability mass of the corresponding state under probability function P. Each demand has a value equal to the probability mass of the corresponding state under probability function Q. Furthermore, imagine there is a transportation arc from each supply node to each demand node, labeled with a cost equal to the distance of the corresponding states under d. This constitutes a transportation network. A flow with respect to this network is an assignment of quantities to be shipped along each arc subject to the conditions that the total flow leaving a supply node is equal to its supply, and the total flow entering a demand node is equal to its demand. The cost of a flow along an arc is the value of the flow along that arc multiplied by the cost assigned to that arc. The goal of the Kantorovich optimal mass transportation problem is to find the best total flow for the given network, i.e. the flow of minimal cost. If we denote by $T_K(d)$ the Kantorovich metric, then the distance assigned to P and Q, $T_K(d)(P, Q)$, is the cost of the optimal flow. It is known to be computable in strongly polynomial time.

The associated bisimulation metric, d_{fix}, is given by $d_{fix}(s, s') = \max_{a \in A}(c_R|r^a_s - r^a_{s'}| + c_T T_K(d_{fix})(P^a_s, P^a_{s'}))$ where c_R and c_T are fixed positive weights that sum to 1, i.e. it is a (unique) fixed-point metric. It is based on similar fixed-point metrics developed in the context of labeled Markov systems, roughly MDPs without rewards in (Desharnais *et al.* 1999) and (Desharnais *et al.* 2002). Like those, these metrics can be iteratively computed in time polynomial in the sizes of the state space and the action space, and can be equivalently formulated in terms of a real-valued modal logic characterizing bisimulation.

An alternative bisimulation metric, $d_\sim$, comes from replacing in the above equation the Kantorovich metric by one based on the total variation distance. We take the bisimulation probability metric T_B to be defined as $T_B(P, Q) = \frac{1}{2}\sum_{C \in S/\sim} |P(C) - Q(C)|$. Its calculation involves computing the stochastic bisimulation partition, which can be done iteratively in polynomial time (Givan, Dean, & Greig 2003). We then define $d_\sim(s, s') = \max_{a \in A}(c_R|r^a_s - r^a_{s'}| + c_T T_B(P^a_s, P^a_{s'}))$.

Both bisimulation metrics assign distances between 0 (bisimilar) and 1 (not bisimilar), giving a measure of "how bisimilar" two states are. They are related by $d_{fix} \leq d_\sim$ under a pointwise ordering and are in some sense the tightest and loosest smooth bisimulation metrics of the given form, respectively. Moreover, we can show that the optimal value function for M, V^*, is Lipschitz continuous for each. Specifically, $c_R|V^*(s) - V^*(s')| \leq d_{fix}(s, s') \leq d_\sim(s, s')$ provided the discount factor $\gamma \leq c_T$ (Ferns 2003). In particular, this shows that the more bisimilar two states are, the closer are their optimal values. Such results are extendable to a particular aggregate MDP, allowing us to similarly bound the difference between the optimal value of a state in M and the optimal value of its equivalence class in the aggregate in terms of the bisimulation distances.

Both metrics have advantages and disadvantages. Computation of d_{fix} does not require that the exact bisimulation partition be computed. Moreover, extending these metrics to continuous state space models in which rewards are uniformly bounded looks very promising. On the other hand, iteratively computing the Kantorovich metric (as is required for the overall computation) while requiring only polynomially many operations is still very time consuming. By contrast, $d_\sim$ is relatively fast to compute (as the bisimulation partition itself can be quickly computed). On the other hand since its computation relies on computing the exact bisimulation partition, which is subject to numerical inaccuracy, it may be numerically unstable. Nevertheless, for these reasons and more we argue that while the Kantorovich based bisimulation metrics are theoretically pleasing, the total variation based metrics are more pleasing in practice, and provide empirical results. The bulk of this work can be found in (Ferns 2003).

References

Boutilier, C.; Dean, T.; and Hanks, S. 1999. Decision-theoretic planning: Structural assumptions and computational leverage. *Journal of Artificial Intelligence Research* 11:1–94.

Desharnais, J.; Gupta, V.; Jagadeesan, R.; and Panangaden, P. 1999. Metrics for labeled markov systems. In *International Conference on Concurrency Theory*, 258–273.

Desharnais, J.; Gupta, V.; Jagadeesan, R.; and Panangaden, P. 2002. The metric analogue of weak bisimulation for probabilistic processes. In *Logic in Computer Science*, 413–422. Los Alamitos, CA, USA: IEEE Computer Society.

Ferns, N. 2003. Metrics for markov decision processes. Master's thesis, McGill University. URL: `http://www.cs.mcgill.ca/~nferns/mythesis.ps`.

Gibbs, A. L., and Su, F. E. 2002. On choosing and bounding probability metrics.

Givan, R.; Dean, T.; and Greig, M. 2003. Equivalence notions and model minimization in markov decision processes. *Artif. Intell.* 147(1-2):163–223.

Larsen, K., and Skou, A. 1991. Bisimulation through probabilistic testing. *In Information and Computation* 94:1–28.

Milner, R. 1980. *A Calculus of Communicating Systems*. Lecture Notes in Computer Science Vol. 92. Springer-Verlag. MIL r 80:1 1.Ex.

Park, D. 1981. Concurrency and automata on infinite sequences. In *Proceedings of the 5th GI-Conference on Theoretical Computer Science*, 167–183. Springer-Verlag.

Puterman, M. L. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc.