Markov Decision Processes in Large State Spaces

Lawrence K. Saul and Satinder P. Singh
lksaul@psyche.mit.edu, singh@psyche.mit.edu

Center for Biological and Computational Learning
Massachusetts Institute of Technology
79 Amherst Street, E10-243
Cambridge, MA 02139

Abstract

In this paper we propose a new framework for studying Markov decision processes (MDPs), based on ideas from statistical mechanics. The goal of learning in MDPs is to find a policy that yields the maximum expected return over time. In choosing policies, agents must therefore weigh the prospects of short-term versus long-term gains. We study a simple MDP in which the agent must constantly decide between exploratory jumps and local reward mining in state space. The number of policies to choose from grows exponentially with the size of the state space, N. We view the expected returns as defining an energy landscape over policy space. Methods from statistical mechanics are used to analyze this landscape in the thermodynamic limit $N \rightarrow \infty$. We calculate the overall distribution of expected returns, as well as the distribution of returns for policies at a fixed Hamming distance from the optimal one. We briefly discuss the problem of learning optimal policies from empirical estimates of the expected return. As a first step, we relate our findings for the entropy to the limit of high-temperature learning. Numerical simulations support the theoretical results.

1 Introduction

Many real-world tasks in machine learning, particularly those in navigation and control, require agents with decision-making abilities. Markov decision processes (MDPs) [3] provide a theoretical framework for modeling tasks in which an agent must constantly monitor its environment and take appropriate courses of action. The fundamental problem in MDPs is one of temporal credit assignment—determining which actions have important long-term consequences. The goal of learning is to find a policy, or set of actions, that yields the maximum expected return over time. Successful strategies for learning must therefore look beyond immediate rewards and concentrate on long-term gains.

There is a large literature on methods for finding optimal policies in MDPs [3, 2, 6]. If a model of the Markov environment is assumed known, then there exist classical methods such as value iteration for finding optimal policies in polynomial time [11]. For many problems of interest, a model of the environment is not available. In this case, there are two basic strategies—direct and indirect—for finding optimal policies [2]. Direct methods attempt to learn good approximations to optimal policies without estimating a model of the environment. Reinforcement learning algorithms, such as TD(λ) [10] and Q-learning [13, 14], are examples of direct methods. The main results for these algorithms are that they converge with probability one in the limit of infinite experience; no rate of convergence results are available. Unlike direct methods, indirect methods attempt to estimate a model of the Markov environment and then derive control policies from the estimated model. Recently, Fiechter [4] has studied the problem of model estimation in MDPs and given a PAC-learning algorithm for finding near-optimal policies.

In this paper we propose an alternative framework for studying MDPs, based on ideas from statistical mechanics. Our approach draws on previous work in statistical mechanics and computational learning theory [5, 9, 12]. We have not made an effort to be rigorous, relying instead on numerical simulations to check the soundness of our methods. The main contributions of this paper are the following: to view the expected returns as defining an energy landscape over policy space, to analyze this landscape with tools from statistical mechanics, and to introduce a particularly tractable example that makes this analysis possible. The main shortcomings, on the other hand, are that our methods do not generalize to arbitrary MDPs and that we do not adequately address the problem of learning. Real-world problems in decision and control necessarily involve a large number of degrees of freedom. Our motivation was to build a statistical mechanical framework for these problems, similar to ones that exist for other problems in memory and learning [1, 12]. This paper does not achieve this goal, but should be viewed as a first step in this direction.

The organization of the paper is as follows. Section 2 reviews the basic elements of MDPs: rewards and transitions in state space, value functions, and the special consequences of the Markov property. It also introduces the problem that serves as a focus for the rest of the paper. This particular example was chosen because it epitomizes the dilemma of balancing short-term versus long-term payoffs. What makes it especially tractable is a simple, closed-form expression for the expected return.

Section 3 uses techniques from statistical mechanics to analyze the distribution of expected returns over policy space. Here, we introduce the important role of entropy and the thermodynamic limit. We also examine how the metric of Hamming distance relates to the landscape of expected returns. In particular, we calculate the typical loss in expected return as a function of Hamming distance from the optimal policy, as well as upper and lower bounds on this loss. In the last part of the section, we discuss the problem of learning optimal policies from empirical estimates of the expected return. We relate our findings for the entropy to the well-known limit of high temperature learning [9]. Numerical evidence is presented to support the theoretical results.

Finally, section 4 presents our conclusions and ideas for future work. The appendix contains technical details of the calculations that appear in section 3.

2 Markov Decision Processes

This section presents a brief review of MDPs, concentrating on those aspects most relevant to our work. A more thorough introduction may be found in [3].

2.1 Background

A Markov decision process (MDP) models an agent's environment by a set of N states. In each of these states, the agent is required to choose from a set of possible actions. Here, we focus on MDPs in which the agent must decide on one of two possible actions. In this case, a policy π is an N-bit string that assigns an action to each state in the environment. We denote the prescribed action at state i by a_i, so that $\pi = \{a_1, a_2, \ldots, a_N\} \in \{0,1\}^N$.

At each time step, the agent executes an action and receives a positive or negative reward from the environment. The reward R_i depends on the current state and the selected action, so that

$$R_i = \bar{r}_i(1-a_i) + r_i a_i, \tag{1}$$

where $\bar{r}_i$ is the reward that results from taking action $a_i = 0$, and r_i the reward for $a_i = 1$. The agent's actions also lead to stochastic changes in the state of the environment. In particular, the probability of making a transition from state i to state j is given by

$$P_{ij} = \bar{p}_{ij}(1-a_i) + p_{ij} a_i, \tag{2}$$

where $\bar{p}_{ij}$ represents the transition probability that results from taking action $a_i = 0$, and p_{ij} the probability for $a_i = 1$. The actions thus determine both the rewards and the transition probabilities at each time step.

The usual goal of learning in MDPs is to find a policy that yields the maximum expected return over time. For a fixed policy π, the value function, or the expected return as a function of the start state, is given by the expected sum of discounted rewards

$$V_i^\pi = \mathrm{E}\left[\sum_{t=0}^{\infty} \gamma^t R_{i_t} \,\middle|\, i_0 = i\right], \tag{3}$$

when the agent starts in state i and executes policy π forever. The expectation is taken over all possible paths $\{i_t\}_{t=0}^{\infty}$ through state space that start at state i and result from actions dictated by π. The discount factor $0 \le \gamma < 1$ causes rewards later in time to be weighted less than rewards earlier in time. In particular, eq. (3) weights the reward at time t by γ^t, setting an effective horizon time

$$\tau = \sum_{t=0}^{\infty} \gamma^t = (1-\gamma)^{-1} \tag{4}$$

for the decision process. The Markov property leads to a recurrence relation for the value functions [3]:

$$V_i^\pi = R_i + \gamma \sum_j P_{ij} V_j^\pi. \tag{5}$$

Note that the rewards R_i and the transition probabilities P_{ij} implicitly depend on the policy π through eqs. (1) and (2). Solving for V_i^π gives

$$V_i^\pi = \sum_j (I-\gamma P)^{-1}_{ij} R_j \tag{6}$$

where I stands for the $N \times N$ identity matrix and the exponent denotes a full matrix inversion.

For concreteness, let us suppose that the agent starts the decision process with equal probability in each state i. Then the normalized return

$$v^\pi = \frac{1-\gamma}{N} \sum_i V_i^\pi \tag{7}$$

provides a reasonable measure-of-goodness for policy π: it is simply the total expected return divided by the effective horizon time of the decision process.

2.2 Example

In this section, we introduce the MDP that will serve as an example for the rest of the paper. The two actions in this MDP correspond to exploratory jumps and local reward-mining in state space. In particular, the action $a_i = 0$ causes the agent to jump with equal probability to any state in state space, while the action $a_i = 1$ causes the agent to remain in place. Eq. (2) gives the transition matrix

$$P_{ij} = \frac{1}{N}(1-a_i) + \delta_{ij} a_i, \tag{8}$$

where δ_{ij} is the Kronecker delta function. The agent receives zero reward $\bar{r}_i = 0$ for exploratory actions and a state-dependent reward r_i for remaining in state i. The task for the agent is to choose a high-reward state at which to stop exploring. The rewards r_i are assumed to be independently chosen from a distribution $\rho(r)$ and remain fixed for all time. Following eq. (1), we have

$$R_i = r_i a_i, \tag{9}$$

where r_i varies from state to state according to the distribution $\rho(r)$.

A basic strategy for maximizing the expected return, defined by eq. (3), is to jump out of states with low rewards and remain at states with high rewards. The dilemma is that the agent has an effective lifetime set by the discount factor γ. If $\gamma \ll 1$, then the agent must hope to quickly find a state with positive reward. On the other hand, if γ is close to unity, then the agent can afford to ignore modest rewards and explore the state space until it "hits the jackpot". The optimal expected return thus depends crucially on the effective horizon time and the distribution of rewards $\rho(r)$.

An attractive feature of this MDP is that the expected return for a fixed policy π has a simple closed form. In particular, let

$$\mu^\pi = \frac{1}{N}\sum_i a_i \tag{10}$$

be the fraction of states where the agent chooses *not* to explore under policy π. Then the normalized return, defined by eq. (7), is given by

$$v^\pi = \frac{\frac{1}{N}\sum_i r_i a_i}{1-\gamma+\gamma\mu^\pi}. \tag{11}$$

Note that all dependence on π enters through the fraction μ^π and the weighted sum $\sum_i r_i a_i$. To prove this result, let

$$\Lambda_j = \sum_i (I-\gamma P)^{-1}_{ij}. \tag{12}$$

Taking the product $(I-\gamma P)^T \Lambda$ and substituting eq. (8) for the transition matrix gives a set of linear equations for Λ_j:

$$\sum_j \left[\delta_{jk}(1-\gamma a_j) - \frac{\gamma}{N}(1-a_j)\right]\Lambda_j = 1. \tag{13}$$

The solution to these equations,

$$\Lambda_j = \frac{(1-\gamma)^{-1}a_j + (1-a_j)}{1-\gamma+\gamma\mu^\pi}, \tag{14}$$

may be verified by substitution. Noting from eq. (7) that $v^\pi = N^{-1}(1-\gamma)\sum_j \Lambda_j R_j$, and substituting eq. (9) for the rewards, one arrives at the desired expression for v^π.

3 Statistical Mechanics

The simple result, eq. (11), makes it possible to study the structure of this MDP in detail. This structure is encoded in the distribution of expected returns and may be analyzed independent of any particular learning algorithm. For the example of the previous section, the analysis simplifies considerably in the limit $N \to \infty$. This is the thermodynamic limit of an infinitely large state space, and it leads naturally to the formalism of statistical mechanics.

3.1 Entropy

A fundamental point of interest is how the normalized returns v^π are distributed over policy space. Let

$$\Omega(v) = \sum_\pi \delta(v - v^\pi), \tag{15}$$

where $\delta(v)$ is the Dirac delta function, and the sum over π traces over all 2^N policies in $\{0,1\}^N$. Properly smoothed for finite N, the function $\Omega(v)$ becomes a histogram of v over policy space. The entropy

$$s(v) = \lim_{N\to\infty} \frac{1}{N}\ln\Omega(v) \tag{16}$$

corresponds to this histogram on a log scale. As $N \to \infty$, there emerges a continuum of expected returns, and we expect $s(v)$ to be a smooth function of v. The function $s(v)$ also characterizes the optimal return, $v^* = \max_\pi v^\pi$: assuming there exists a unique optimal policy, then v^* may be found by solving for the largest root of $s(v) = 0$.

The entropy can be calculated from eq. (15) by rewriting the sum as an integral and using the method of saddlepoint integration [7]. The details of this calculation are presented in the appendix. There we show that

$$s(v) = \min_\phi \left\{-\phi(1-\gamma)v + \int dr\ \rho(r)\ln\left[1+e^{\phi(r-\gamma v)}\right]\right\}. \tag{17}$$

Given a discount factor γ and a reward distribution $\rho(r)$, eq. (17) can be solved numerically for $s(v)$. Figure 1 shows a plot of $s(v)$ for discount factors $\gamma = 0.75$ and $\gamma = 0.95$ and the box distribution

$$\rho(r) = \begin{cases} \frac{1}{2} & \text{for } |r| < 1. \\ 0 & \text{otherwise.} \end{cases} \tag{18}$$

The asymptotic behavior of $s(v)$ for $(v^* - v) \ll 1$ may be calculated by performing a Sommerfeld expansion [7] in the parameter ϕ of eq. (17). The results

$$s(v) \sim (v^*-v)^{1/2}, \tag{19}$$

$$v^* = \gamma^{-2}\left[2-\gamma-2\sqrt{1-\gamma}\right] \tag{20}$$

are derived in the appendix for the box distribution of rewards, eq. (18). The exponent of eq. (19) determines the asymptotic behavior for learning curves in the limit of high temperature learning [9]. We will return to this point in section 3.3, where we also provide numerical evidence for the predicted form of $s(v)$.

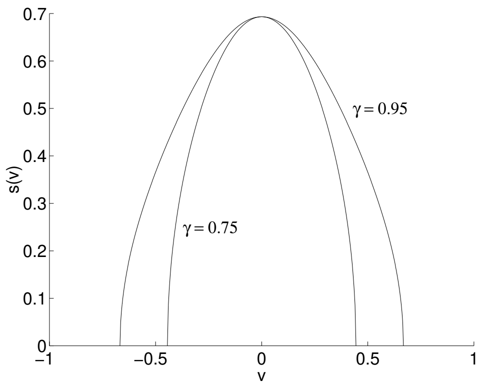

Figure 1: Plots of entropy, $s(v)$, versus normalized return, v, for discount factors $\gamma = 0.75$, and $\gamma = 0.95$. The right and left roots of $s(v) = 0$ indicate the best and worst possible returns.

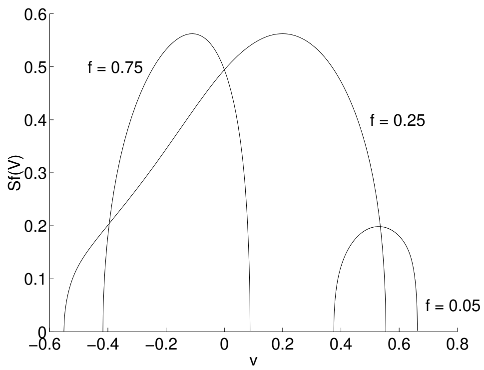

Figure 2: Plots of entropy, $s_f(v)$, versus normalized return, v, for Hamming radii $f = 0.05, 0.25$, and 0.75. The discount factor is $\gamma = 0.95$; the optimal return is $v^* = 0.66791$.

3.2 Hamming Shells

The entropy $s(v)$ characterizes the overall distribution of returns in policy space. The structure of this distribution is based on eq. (11), which identifies the two trademarks that determine a policy's expected return: the fraction of states devoted to reward-mining, and the policy-weighted sum over rewards. How well do other metrics over policy space succeed in characterizing this structure?

In this section, we analyze the Hamming metric in the neighborhood of the optimal policy, π^*. Let

$$d_H(\pi, \pi^*) = \sum_i [a_i(1 - a_i^*) + (1 - a_i)a_i^*] \qquad (21)$$

denote the Hamming distance between π and π^*. We refer to $f^\pi = N^{-1} d_H(\pi, \pi^*)$ as the Hamming radius of π: it is the fraction of bits that differ between π and π^*. Analogous to eq. (15), the function

$$\Omega_f(v) = \sum_\pi \delta(v - v^\pi)\delta(f - f^\pi) \qquad (22)$$

counts the number of policies with normalized return v and Hamming radius f. As $N \to \infty$, we expect a continuum of Hamming *shells*, i.e. sets of policies equidistant from π^*; we define the corresponding entropy by

$$s_f(v) = \lim_{N\to\infty} \frac{1}{N} \ln \Omega_f(v). \qquad (23)$$

The entropy curves $s_f(v)$ characterize the distributions of expected returns on concentric Hamming shells about the optimal policy. For large N, the histogram $\Omega_f(v)$ is sharply peaked about the value of v that maximizes $s_f(v)$. In the limit $N \to \infty$, the typical return for policies with Hamming radius f thus occurs at

$$v_f = \max_v s_f(v). \qquad (24)$$

Equivalently, $(v^* - v_f)$ represents the typical loss for these policies. In the example from section 2.2, there are unique policies in each Hamming shell that yield the best and worst expected returns. Lower and upper bounds on the loss are therefore given by the right-most and left-most roots of $s_f(v) = 0$.

The entropy curves for fixed Hamming radius are derived in the appendix. The final result is

$$\begin{aligned} s_f(v) \;&=\; \min_{\phi,\omega} \{-\phi(1-\gamma)v + \omega(f - \mu^*) \\ &+ \int dr\, \rho(r) \ln\left[1 + e^{\phi(r-\gamma v)+\omega \mathrm{sgn}(r-\gamma v^*)}\right]\Big\} \end{aligned} \qquad (25)$$

where μ^* is the fraction of states in π^* devoted to reward-mining, and $\mathrm{sgn}(x)$ denotes the sign (± 1) of x. Plots of $s_f(v)$ may be computed by numerically minimizing the right hand side of eq. (25) over the parameters ϕ and ω. Figure 2 shows three overlaid plots of $s_f(v)$ for Hamming radii $f = 0.05$, 0.25, and 0.75, discount factor $\gamma = 0.95$, and the box distribution of rewards, eq. (18). Each value of f has a range of possible returns, with v_f bounded on either side by the minimum and maximum loss. Of course, as f goes from 0 to 1, the peak of the entropy curve (and hence v_f) shifts away from v^*. The range of possible returns increases with f for small f, but eventually contracts, until at $f = 1$ only one policy in $\{0, 1\}^N$ remains to be counted, namely the complement of π^*.

Figure 3 shows plots of the best, worst, and typical return versus Hamming radius. For the box distribution, eq. (18), the typical return is given by

$$v_f = \frac{(1-2f)(1-\gamma^2 v^{*2})}{2[2 - \gamma - \gamma^2 v^*(1-2f)]}. \qquad (26)$$

Empirical results for the typical return are also plotted in figure 3. The empirical results were obtained in finite-size MDPs by sampling policies on constant Hamming shells and averaging the returns computed from eq. (11). We approximated the box distribution for rewards by setting r_i at N equally spaced intervals between -1 and 1. Note that even for finite-size MDPs

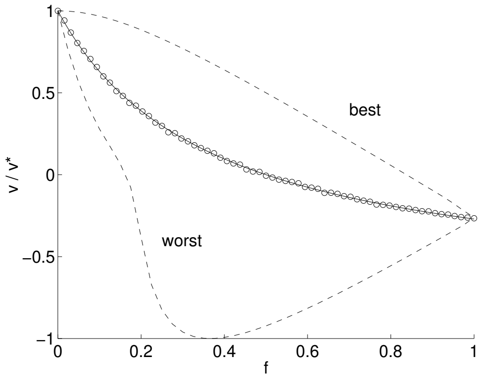

Figure 3: Plots of the best, worst, and typical return (normalized by v^*) versus Hamming radius from the optimal policy. The circles are empirical results from finite-size MDPs ($N = 128$).

($N = 128$), there is good agreement with the results for the thermodynamic limit.

3.3 High Temperature Learning

Several algorithms have been proposed for learning optimal policies from empirical estimates of the expected return. In this section, we examine a stochastic search algorithm in which the agent learns from repeated trials of the decision process. Our goal is not to introduce yet another algorithm for solving MDPs, but to study how inferences about the optimal policy improve with sample size, i.e. the number of trials available to the agent. For this purpose, we will borrow a framework from statistical mechanics [9], originally developed to analyze learning curves in simple perceptrons.

Our algorithm takes the following form. We suppose that, for each policy π, the agent is allowed to determine an estimate of the expected return by executing m trials of the decision process. In each of these trials, the agent begins from a random initial state of the MDP and accumulates the sum of discounted rewards under the chosen policy. Let $\hat{v}_\ell^\pi$ label the return sampled from the ℓ^{th} trial[1] under policy π. The sum of sampled returns

$$E_\pi = -\sum_{\ell=1}^{m} \hat{v}_\ell^\pi \tag{27}$$

defines an energy landscape over policy space; the negative sign is introduced so that high returns correspond to low energies. We imagine learning as a process in which the agent traverses this landscape, looking for policies with low energies.

[1] For simplicity, we assume that each trial of the decision process steps through an infinite sequence of state-action pairs. Taking only a finite number of steps introduces a truncation error into the discounted sum of rewards. For $\gamma < 1$, however, the magnitude of this error is trivially bounded by $\gamma^K(1-\gamma)^{-1}\mathcal{R}$, where K is the number of steps and $\mathcal{R} = \max_i |r_i|$.

Stochastic exploration of this landscape may be done by Monte Carlo simulation [8], in which policies are updated by the following rule. Let E denote the energy of the agent's current policy π, and E' the energy of a nearby policy π' with $d_H(\pi, \pi') = 1$. Then the agent changes from policy π to π' with probability

$$W(\pi \to \pi') = \begin{cases} 1 & \text{if } \Delta E < 0. \\ e^{-\beta \Delta E} & \text{otherwise.} \end{cases} \tag{28}$$

where $\Delta E = E' - E$. Here, β is a noise parameter that determines how often the agent opts for policies with higher energies. We assume that, during the course of exploration, the agent is able to keep a record of energies from previous trials; as a result, it does not need to resample the decision process in the event that it returns to a previously visited policy.

In what follows, we focus on the long-time, or equilibrium, properties of this learning procedure. At long times, the dynamics of eq. (28) generates a Gibbs probability distribution

$$P_\pi = Z^{-1} e^{-\beta E_\pi} \tag{29}$$

over policy space, with the noise parameter $\beta = 1/T$ playing the role of inverse temperature. The prefactor Z is the partition function

$$Z = \sum_\pi e^{-\beta E_\pi} \tag{30}$$

that normalizes the Gibbs distribution. We may now apply the formalism of statistical mechanics to calculate the equilibrium properties of this learning procedure. To obtain a proper thermodynamic limit requires the energy function, eq. (27), to scale linearly with N. We therefore take the combined limit

$$m \to \infty, \quad N \to \infty, \quad \tfrac{m}{N} = \alpha \text{ (finite)}, \tag{31}$$

such that the number of trials grows in direct proportion to the size of the state space.

The goal of this framework is to extract typical learning behaviors as a function of the parameters α and T. As T decreases, the agent tends to concentrate on policies with low energies. As α increases, the agent receives more accurate estimates of the expected returns, so that policies with low energies tend to be near-optimal. Exactly how close are the agent's choices to the optimal policy, and to what extent do they yield a near optimal expected return?

In general, answering these questions requires the additional step of averaging over the agent's training data, i.e. all possible outcomes of the trials $\hat{v}_l^\pi$. The reason for this is that we do not expect the typical learning behavior to depend on the particular estimates that the agent receives for v^π. Performing the average over these estimates requires knowledge of the distribution from which they are generated. Characterizing this distribution remains a current area of research. In this paper, we therefore focus on a highly simplified limit in which this average is unnecessary: this is the combined limit of high temperatures and large sample size. In particular, let

$$\alpha \to \infty, \quad T \to \infty, \quad \frac{\alpha}{T} = \tilde{\alpha} \text{ (finite)}. \tag{32}$$

This is the limit of high temperature learning [9]. In this limit, the quantity $-\beta E_\pi$ in the exponent of the Gibbs distribution may be replaced by $N\tilde{\alpha}v^\pi$, so that the typical learning behavior becomes a function of the single parameter $\tilde{\alpha}$. Note that high temperature learning does not correspond to the situation in which all policies are equally likely; rather, the combined limit, eq. (32), ensures a non-trivial competition between energy and entropy for finite values of $\tilde{\alpha}$.

High temperature learning is clearly an artificial paradigm for learning in MDPs. The results are of interest, though, for two reasons: first, because they have analogs in the perceptron literature [9] that can guide our thinking, and second, because they provide a numerical check on the entropy curves of section 3.1. In the high temperature limit, the partition function reduces to

$$Z_0 = \sum_\pi e^{N\tilde{\alpha}v^\pi} = \int dv \; e^{N[s(v)+\tilde{\alpha}v]}, \tag{33}$$

where $s(v)$ is the entropy from eq. (16). Fluctuations about the most probable value of v vanish in the thermodynamic limit, so that

$$\frac{ds}{dv} + \tilde{\alpha} = 0 \tag{34}$$

gives the learning curve for the normalized return as a function of $\tilde{\alpha}$. The asymptotic behavior for large $\tilde{\alpha}$ follows from eq. (19):

$$(v^* - v) \sim \tilde{\alpha}^{-2}. \tag{35}$$

Figure 4 compares the theoretical prediction for the entire learning curve, based on eqs. (17) and (34), versus the results of Monte Carlo simulations on finite-size MDPs. These simulations were performed in the high temperature limit, with true expected returns replacing energies and the parameter $\tilde{\alpha}$ serving as an effective temperature. As before, the rewards were chosen uniformly between -1 and 1. There is good agreement between the theoretical and empirical results, thus validating the predicted form of the entropy curve $s(v)$.

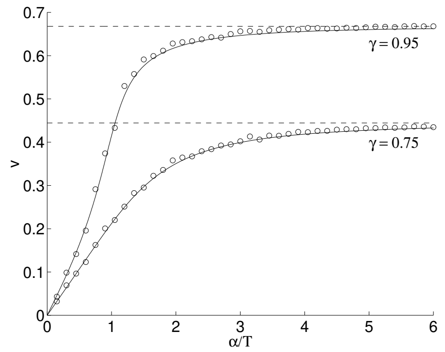

Figure 4: Learning curves, v vs. $\tilde{\alpha}$, in the high temperature limit. The dashed lines indicate optimal returns. The circles are empirical results for finite-size MDPs ($N = 128$).

4 Discussion

Our goal in this paper has been to explore how statistical mechanics can be applied to problems in decision and control, problems which typically involve a large number of degrees of freedom. The framework we have presented has a number of advantages for studying MDPs. In certain cases, it can provide a detailed understanding of the distribution of expected returns. It also focuses on the limit of large state spaces, a limit which is relevant for real-world problems. On the other hand, our approach has clear limitations. It is unlikely that the calculations can be done for arbitrary MDPs, since they rely on simple forms for the rewards and transition matrices. Moreover, there are important differences between the algorithms in this paper and the ones used by practitioners. Why pursue a framework based on contrived examples and naive algorithms? The main reason is to obtain an unambiguous benchmark against which to evaluate the relative merits of more sophisticated algorithms, such as TD(λ) and Q-learning, that exploit the structure of the decision process. Clear theoretical results, even on relatively simple MDPs, would be a step in this direction.

There is more work to be done for results that serve this purpose. In particular, we hope to extend the analysis of section 3.3 beyond the limit of high temperature learning. It would also be interesting to consider the limit $\gamma \to 1$, or possibly the combined limit $\gamma \to 1, N \to \infty, \gamma^N = \tilde{\gamma}$, where $0 < \tilde{\gamma} < 1$: this is the limit of infinite effective horizon times. Finally, we would like to find other MDPs, besides the example of section 2.2, that can be analyzed with the tools of this paper. These issues and others are left for future work.

Acknowledgements

It is a pleasure to acknowledge useful discussions with P. Dayan, T. Jaakkola, M. Jordan, M. Kearns, and H. Seung. Both authors are supported by grants to M. Jordan (MIT) from ATR Human Information Processing Research and Siemens Corporation. LS also acknowledges support from NSF grant CDA-9404932.

References

[1] D. J. Amit. *Modeling brain function*. Cambridge University Press, Cambridge, 1989.

[2] A. G. Barto, S. J. Bradtke, and S. P. Singh. Learning to act using real-time dynamic programming. *Artificial Intelligence* **72**: 81–138, 1995.

[3] D. P. Bertsekas. *Dynamic programming: deterministic and stochastic models*. Prentice Hall, Englewood Cliffs, NJ, 1987.

[4] C. N. Fiechter. Efficient reinforcement learning. In *Proc. 7th Annual Workshop on Comput. Learning Theory*, pages 88–97. Morgan Kaufmann, San Mateo, CA 1994.

[5] D. Haussler, M. Kearns, H. S. Seung, and N. Tishby. Rigorous learning curve bounds from statistical mechanics. In *Proc. 7th Annual Workshop on Comput. Learning Theory*, pages 76–87. Morgan Kaufmann, San Mateo, CA, 1994.

[6] R. Howard. *Dynamic programming and Markov processes.* MIT Press, Cambridge, MA, 1960.

[7] K. Huang. *Statistical Mechanics.* John Wiley & Sons, New York, NY, 1987.

[8] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipes.* Cambridge University Press, Cambridge, 1986.

[9] H. S. Seung, H. Sompolinsky, and N. Tishby. Statistical mechanics of learning from examples. *Physical Review A* **45**: 6056–6091, 1992.

[10] R. S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning* **3**:9–44, 1988.

[11] P. Tseng. Solving H-horizon, stationary Markov decision problems in time proportional to log(H), *Operations Research Letters*, **9**:287–297, 1990.

[12] T. Watkin, A. Rau, and M. Biehl. The statistical mechanics of learning a rule. *Reviews of Modern Physics* **65**:499-556, 1993.

[13] C. Watkins. Learning from delayed rewards. PhD thesis, Cambridge University, 1989.

[14] C. Watkins and P. Dayan. Q-learning. *Machine Learning* **8**: 279–292, 1992.

Appendix

In this appendix we demonstrate the methods used to derive the entropy curves from section 3. Consider first the distribution of expected returns, eq. (15). Substituting eq. (11) for v^π, we may rewrite this as

$$\Omega(v) = \sum_\pi \int d\mu \, \delta\left(\mu - \frac{1}{N}\sum_i a_i\right) \delta\left(v - \frac{\frac{1}{N}\sum_i r_i a_i}{1-\gamma+\gamma\mu}\right). \tag{36}$$

The constraints on μ and v may be recast as integrals using the representation $2\pi i \delta(x) = \int_{-i\infty}^{+i\infty} d\hat{x}\, e^{\hat{x}x}$ for the Dirac delta function. Introducing auxiliary variables $\hat{\mu}$ and $\hat{v}$, we find

$$\begin{aligned} \Omega(v) &= \sum_\pi \int \exp\left\{ N \left[\hat{\mu}\left(\mu - \frac{1}{N}\sum_i a_i\right) \right.\right. \qquad (37) \\ &\quad \left.\left. + \, \hat{v}\left(v - \frac{\frac{1}{N}\sum_i r_i a_i}{1-\gamma+\gamma\mu}\right)\right]\right\} \frac{N d\hat{\mu}}{2\pi i}\frac{N d\hat{v}}{2\pi i}\, d\mu. \end{aligned}$$

Note that the action variables a_i in the exponent are now factorized, making it possible to perform the sum over policies. Tracing over the action-state pairs $\{a_i\} \in \{0,1\}^N$ gives

$$\Omega(v) = \left(\frac{N}{2\pi i}\right)^2 \int d\mu\, d\hat{\mu}\, d\hat{v}\, \exp\left[N h(\hat{\mu},\hat{v},\mu)\right], \tag{38}$$

where

$$\begin{aligned} h(\hat{\mu},\hat{v},\mu) &= \frac{1}{N}\sum_i \ln\left\{1 + \exp\left[-\hat{\mu} - \frac{r_i \hat{v}}{1-\gamma+\gamma\mu}\right]\right\} \\ &\quad + \hat{\mu}\mu + \hat{v}v. \end{aligned} \tag{39}$$

In the limit $N \to \infty$, the leading contribution to $\Omega(v)$ may be determined by the method of saddlepoint integration [7]. The saddlepoint of $h(\hat{\mu},\hat{v},\mu)$ is located by setting its derivatives equal to zero:

$$0 = \mu - \sum_i \frac{1}{1+e^{\hat{\mu}+\hat{v}r_i/(1-\gamma+\gamma\mu)}}. \tag{40}$$

$$0 = v - \sum_i \frac{r_i(1-\gamma+\gamma\mu)^{-1}}{1+e^{\hat{\mu}+\hat{v}r_i/(1-\gamma+\gamma\mu)}}. \tag{41}$$

$$0 = \hat{\mu} + \sum_i \frac{\gamma\hat{v}r_i(1-\gamma+\gamma\mu)^{-1}}{1+e^{\hat{\mu}+\hat{v}r_i/(1-\gamma+\gamma\mu)}}. \tag{42}$$

The entropy, defined by eq. (16), is given by the value of $h(\hat{\mu},\hat{v},\mu)$ at its saddlepoint. Eliminating the variable $\hat{\mu}$ through eqs. (41) and (42), and introducing $\phi = -\hat{v}/(1-\gamma+\gamma\mu)$, we obtain

$$s(v) = \min_\phi \left\{ -\phi(1-\gamma)v + \frac{1}{N}\sum_i \ln\left[1 + e^{\phi(r_i - \gamma v)}\right]\right\}.$$

In the limit $N \to \infty$, the sum over states may be replaced by the integral over the reward distribution, $\rho(r)$. This leads to the desired result for the entropy, eq. (17).

Let v^* denote the return of the optimal policy, π^*. We now consider the asymptotic behavior of the entropy curve for $(v^* - v) \ll 1$. The value of ϕ which minimizes the right hand size of eq. (17) satisfies

$$(1-\gamma)v = \int \frac{dr\, \rho(r)(r-\gamma v)}{1+e^{-\phi(r-\gamma v)}}. \tag{43}$$

As v approaches v^*, the solution for ϕ diverges to infinity. The asymptotic behavior of $s(v)$ may be calculated by performing a Sommerfeld expansion [7] of eq. (43) in powers of ϕ^{-2}:

$$(1-\gamma)v \approx \int dr\, \rho(r)(r-\gamma v)\Theta(r-\gamma v) + \frac{\pi^2\rho(\gamma v)}{6\phi^2}, \tag{44}$$

where $\Theta(x)$ is the unit step function. The integral in eq. (44) can be done exactly for the box distribution of rewards. In this case, one finds $\phi \sim (v^* - v)^{-1/2}$, with v^* given by eq. (20). From eq. (17) we have that

$$\frac{ds}{dv} = -\phi(1-\gamma+\gamma\mu), \tag{45}$$

where

$$\mu = \int \frac{dr\, \rho(r)}{1+e^{-\phi(r-\gamma v)}}. \tag{46}$$

As $v \to v^*$ and $\phi \to \infty$, the integral in eq. (46) reduces to

$$\mu^* = \int dr\ \rho(r)\Theta(r - \gamma v^*), \tag{47}$$

Hence, $ds/dv \sim -(1-\gamma+\gamma\mu^*)(v^* - v)^{-1/2}$, and by integration, we obtain the desired result for $s(v)$, eq. (19).

The profile of the optimal policy may also be deduced from the saddlepoint equations. From eq. (47), the fraction of states devoted to reward-mining in π^* is given by counting all the states with rewards greater than γv^*. Eq. (47) thus gives the following prescription for π^*: the agent stays at state i if $r_i > \gamma v^*$ and jumps out of it otherwise. Substituting $a_i^* = \Theta(r_i - \gamma v^*)$ into eq. (21) gives an expression for the Hamming distance to the optimal policy:

$$d_H(\pi, \pi^*) = N\mu^* - \sum_i \mathrm{sgn}(r_i - \gamma v^*) a_i. \tag{48}$$

The steps that lead from eq. (22) for $\Omega_f(v)$ to eq. (25) for $s_f(v)$ are essentially identical to those given for the calculation of $s(v)$. The only difference is that an additional auxiliary variable must be introduced to handle the delta function for f. The final result, eq. (25), therefore requires a minimization over two variables, as opposed to just one. For the box distribution of rewards, the entropy curve $s_f(v)$ has only one peak, so that solving $(ds_f/dv)|_{v_f} = 0$ yields the typical return. As before, the required integrals can be done exactly, yielding eq. (26) for v_f.