Learning Curve Bounds for Markov Decision Processes with Undiscounted Rewards

Lawrence K. Saul and Satinder P. Singh*
Center for Biological and Computational Learning
Massachusetts Institute of Technology
79 Amherst Street, E10-243
Cambridge, MA 02139
{lksaul,singh}@psyche.mit.edu

Abstract

Markov decision processes (MDPs) with undiscounted rewards represent an important class of problems in decision and control. The goal of learning in these MDPs is to find a policy that yields the maximum expected return per unit time. In large state spaces, computing these averages directly is not feasible; instead, the agent must estimate them by stochastic exploration of the state space. In this case, longer exploration times enable more accurate estimates and more informed decision-making. The learning curve for an MDP measures how the agent's performance depends on the allowed exploration time, $\mathcal{T}$. In this paper we analyze these learning curves for a simple control problem with undiscounted rewards. In particular, methods from statistical mechanics are used to calculate lower bounds on the agent's performance in the thermodynamic limit $\mathcal{T} \to \infty$, $N \to \infty$, $\alpha = \mathcal{T}/N$ (finite), where $\mathcal{T}$ is the number of time steps allotted per policy evaluation and N is the size of the state space. In this limit, we provide a lower bound on the return of policies that appear optimal based on imperfect statistics.

1 Introduction

Many problems in decision and control can be modeled as Markov decision processes (MDPs)[2]; applications include game-playing, network routing, robot navigation, elevator scheduling, and shortest-path problems. The goal of learning in MDPs is to discover the course of actions, or policy, that yields the maximum expected return over time. There are many classical algorithms for finding optimal policies given a complete description of the Markov environment. Based on this description, these algorithms compute the temporal statistics that reveal which actions lead to consistent, long-term gains.

The classical algorithms for solving MDPs do not apply in two important situations: (i) when the parameters of the Markov environment are not a priori known, and (ii) when it is not feasible (due to the size of the state space) to perform the required matrix operations. In these cases, one may devise stochastic approximations[1, 11, 13] to these algorithms that estimate the required statistics rather than computing them exactly. These approximations rely on stochastic exploration of the state space to measure the returns attached to particular courses of action. In MDPs, these returns are the sums of rewards that the agent accumulates over time.

In general, longer exploration times lead to more accurate estimates and more informed decision-making. The learning curve for an MDP measures how the agent's performance depends on the allowed exploration time, $\mathcal{T}$, per policy evaluation. An important question for solving MDPs is how these stochastic approximations converge to the correct answer as the exploration time tends to infinity.

Most work on this issue has focused on MDPs whose rewards are attenuated with time. This is done by introducing a discount factor, $0 < \gamma < 1$, and weighting the rewards at time t by γ^t. The discount factor is convenient for bounding estimation errors, as it sets an effective horizon time $\tau_\gamma = (1-\gamma)^{-1}$ for the decision process. In particular, actions beyond this horizon have relatively little effect on the agent's performance. If we demand a fixed level of performance and allow γ to vary, the required exploration times scale as $\mathcal{T} \sim \tau_\gamma$. Likewise, for fixed γ, one can relate the agent's performance to the allowed exploration time. This was done in a PAC framework by Fiechter[4].

Less is known about MDPs with undiscounted rewards. In this case, the goal of the agent is to find the policy that yields the maximum expected return per unit time. Here the prescription $\mathcal{T} \sim \tau_\gamma$ is vacuous, as undiscounted rewards correspond to the limit $\tau_\gamma \to \infty$. The

*Current address: Harlequin Inc., One Cambridge Center, Cambridge, MA 02142.

size of the state space also figures differently in MDPs with discounted and undiscounted rewards. Roughly speaking, in the former the only important states are those that can be reached within τ_γ actions of likely initial states. Hence, the effective number of accessible states may be much smaller than the size of the state space. These considerations do not apply to MDPs with undiscounted rewards.

Our analysis employs a particular limiting method—the so-called thermodynamic limit—developed in the statistical physics literature[9, 12]. For MDPs, this is the combined limit that the allowed exploration time, $\mathcal{T}$, and the size of the state space, N, grow to infinity at a fixed rate: $\mathcal{T} \to \infty, N \to \infty, \mathcal{T}/N = \alpha$ (finite). Ref. [5] gives a rigorous treatment of this method from the viewpoint of computational learning theory. Though formulated originally for problems in supervised learning, it can also be used to study problems in decision and control. Of course, important differences between these two types of problems must be kept in mind. In MDPs, individual rewards are temporally correlated by the agent's path through state space—a path over which the agent, by virtue of its actions at each time step, exerts a direct influence. This fundamentally distinguishes the learning problem in MDPs, one of decision and control, from problems in supervised learning where the goal is function approximation based on an *i.i.d.* set of training examples.

In this paper we analyze the agent's performance on a simple control problem with undiscounted rewards. In general, this performance depends on both the allowed exploration time, $\mathcal{T}$, and the size of the state space, N. The thermodynamic limit has two main virtues: it simplifies our task by collapsing these parameters into a single one, $\alpha = \mathcal{T}/N$, and it focuses our attention on the limit of large state spaces. This limit seems appropriate, since it is precisely for large state spaces that stochastic approximations are necessary to solve MDPs. This paper extends earlier work[8] in which we introduced a thermodynamic limit for MDPs with discounted rewards. Focusing on undiscounted rewards has enabled us to derive much stronger results. In particular, here we account explicitly for the temporal correlations introduced by the agent's path through state space, and we also derive learning curve bounds without taking any additional limits. Notwithstanding these improvements, the present work is essentially self-contained.

We focus on a simple example that makes our analysis tractable. Our example has three features that make it desirable for study. First, it captures the basic problem of temporal credit assignment; actions that yield short-term gains can have negative long-term consequences. Second, the policy space, though it consists of exponentially many distinct policies, has a simple topology that can be visualized in two dimensions. Third, even from this simple example, we have discovered asymptotic rates of convergence that would be difficult to explain otherwise.

The rest of the paper is organized as follows. In section 2, we review the basic elements of MDPs and introduce the example that serves as our case study. In section 3, we examine a particular representation of the policy space for this example; this leads to the notion of entropy and the limit of large state spaces, $N \to \infty$. In section 4, we use tools from large deviation theory to bound the probability of error in estimating value functions. In particular, for estimates based on random walks of length $\mathcal{T}$, we calculate the asymptotic ($\mathcal{T} \to \infty$) error rates when policies are ranked on the basis of sampled returns. Section 5 combines the results of the two previous sections to compute lower bounds on the agent's performance as a function of the allowed exploration time. This is done for the thermodynamic limit, $\mathcal{T} \to \infty$, $N \to \infty$, $\alpha = \mathcal{T}/N$ (finite). Finally, section 6 contains our conclusions, as well as issues for future research.

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 ref. [2].

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^π depends on the current state and the selected action, so that

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

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

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

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 goal of learning in MDPs is to find a policy that yields the maximum expected return over time. This return is just the sum of rewards accumulated at each time step. In MDPs with discounted rewards, a discount factor $0 \le \gamma < 1$ is introduced to attenuate the rewards with time. In this case, the value function, or the expected return as a function of the start state, is defined as the expected sum of discounted rewards:

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

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 γ 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

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

for the decision process.

In this paper we shall focus on MDPs with undiscounted rewards. In this case, the above definitions need to be slightly modified. For MDPs with undiscounted rewards, the goal of the agent is to maximize the expected return per unit time. The value function

$$v_i^\pi = \lim_{\mathcal{T}\to\infty} \mathrm{E}\left[\frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}} R_{i_t}^\pi \,\middle|\, i_0 = i\right], \tag{5}$$

measures this time-averaged return for fixed policy π and initial state i. An equivalent definition for the value function is given by:

$$v_i^\pi = \lim_{\mathcal{T}\to\infty}\lim_{\gamma\to 1} \frac{\mathrm{E}\left[\sum_{t=1}^{\mathcal{T}} \gamma^t R_{i_t}^\pi \,\middle|\, i_0 = i\right]}{\sum_{t=1}^{\mathcal{T}} \gamma^t}, \tag{6}$$

which makes plain the limiting ($\gamma \to 1$) relationship between the case of discounted and undiscounted rewards. Note that the expectation in eq. (5) is independent of the initial state i provided that the policy π has an ergodic transition matrix P_{ij}^π. This will be true for all the examples we discuss. For simplicity, then, we shall drop the subscript labeling the initial state and denote the value functions by v^π.

2.2 Gibbs learning

There are many algorithms for learning optimal policies based on stochastic estimates of the value functions. These algorithms have two goals: first, to efficiently search through the space of 2^N policies, and second, to discriminate (based on imperfect statistics) which policies are best. Since it is this second goal we wish to focus on, we will study an algorithm that has unlimited resources for search, but limited resources for policy evaluation.

The so-called Gibbs algorithm[5] works as follows. For each policy π, it selects a random initial state, then estimates the value function v^π by the time-averaged return from a random walk of $\mathcal{T}$ steps:

$$\hat{v}^\pi = \frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}} R_{i_t}^\pi. \tag{7}$$

This is done in parallel for each of the 2^N policies, $\pi \in \{0,1\}^N$. Having collected empirical estimates $\hat{v}^\pi$ for the true value functions v^π, the Gibbs algorithm then outputs the policy π^{gibbs} with the highest empirical return:

$$\pi^{\mathrm{gibbs}} = \arg\max_\pi \hat{v}^\pi. \tag{8}$$

Thus the Gibbs algorithm selects the policy that appears best based on random walks of length $\mathcal{T}$.

The Gibbs algorithm is highly idealized in that it performs an exhaustive search over all policies $\pi \in \{0,1\}^N$. Direct methods based on policy iteration[2] are more practical for MDPs with large state spaces; roughly speaking, they search through policy space in a stepwise manner, favoring moves that lead to policies with higher returns. These methods approximate the exhaustive search of the Gibbs algorithm with a narrower (but more deliberate) hunt for the optimal policy. Focusing on the Gibbs algorithm allows us to avoid the complicated issue of how any particular algorithm searches through the policy space. Instead, we can concentrate on the more universal issue of decision and control based on imperfect statistics.

We measure the performance of the Gibbs algorithm by comparing the expected return of the Gibbs policy, π^{gibbs}, with that of the optimal one, π^*. As $\mathcal{T} \to \infty$, the estimated value functions approach the true ones, and we expect the return of the Gibbs policy to converge to the optimal expected return. The learning curve plots the difference between these two returns, $(v^* - v^{\mathrm{gibbs}})$, as a function of the exploration time allotted per policy evaluation, $\mathcal{T}$. Our goal is to provide upper bounds on this difference and thus guarantee a minimal level of performance from the Gibbs algorithm.

2.3 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 EXPLORE action $a_i = 0$ causes the agent to jump with equal probability to any state in state space, while the MINING action $a_i = 1$ causes the agent to remain in place with probability $\Delta < 1$. Eq. (2) gives the transition matrix

$$P_{ij}^\pi = \frac{1}{N}(1-a_i) + \left[\Delta\delta_{ij} + \frac{1-\Delta}{N}\right] a_i, \tag{9}$$

where δ_{ij} is the Kronecker delta function. Note that P_{ij}^π has no zero elements, and hence it describes an ergodic transition matrix for any policy π.

The rewards in our example are designed to set up a classic dilemma in decision and control: exploration versus exploitation. In particular, the agent receives zero reward $\bar{r}_i = 0$ for exploratory actions and a state-dependent reward r_i for choosing to mine the ith state. The task for the agent is to choose which states to mine and which to ignore. 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

$$R_i^\pi = r_i a_i, \tag{10}$$

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

A basic strategy for maximizing the expected return per unit time, defined by eq. (5), is to MINE the states with high rewards and jump out of the states with low ones. The dilemma is knowing how to classify the states as rich or poor based on the distribution of rewards, $\rho(r)$, and the mining probability Δ. The mining probability defines an effective lifetime, $\tau_\Delta = (1-\Delta)^{-1}$, during which the agent reaps its rewards. If τ_Δ is of order unity, then the agent does not lose much return by mining a state with low rewards. On the other hand, if $\tau_\Delta \gg 1$, then the agent must learn to pass on states with modest rewards; the same time is better spent looking for more profitable states. The optimal strategy thus depends crucially on the mining probability Δ and the distribution of rewards $\rho(r)$.

Value functions

Let us now compute the value functions v^π for this example. For MDPs with undiscounted rewards, this is done by finding the stationary distribution of the transition matrix P_{ij}^π. The stationary distribution ϕ_i^π obeys the left eigenvalue equation

$$\sum_i \phi_i^\pi P_{ij}^\pi = \phi_j^\pi \tag{11}$$

and the normalization condition $\sum_i \phi_i^\pi = 1$. It is straightforward to show that

$$\phi_i^\pi = \frac{a_i + (1-\Delta)(1-a_i)}{N(1-\Delta+\Delta\mu^\pi)} \tag{12}$$

satisfies these conditions, where

$$\mu^\pi = \frac{1}{N}\sum_i a_i \tag{13}$$

denotes the fraction of states in which the agent chooses to MINE under policy π.

The value function v^π measures the time-averaged return from policy π. As the transition matrix P_{ij}^π is ergodic, the expected return is found by averaging the agent's rewards over the stationary distribution, $v^\pi = \sum_i \phi_i^\pi R_i^\pi$. Combining eqs. (10) and (12), we find:

$$v^\pi = \frac{\omega^\pi}{1-\Delta+\Delta\mu^\pi} \tag{14}$$

where

$$\omega^\pi = \frac{1}{N}\sum_i a_i r_i \tag{15}$$

is a simple reward-weighted sum over state space. Note that the value function v^π depends on π only through the two parameters μ^π and ω^π.

Optimal policy

The optimal policy for this problem has a simple form: it is to EXPLORE at states for which r_i is less than some critical value, r_c, and to MINE at all the rest. Of course, the precise value of r_c depends on the mining probability, Δ as well as the rewards, r_i. In particular, to find r_c, we must maximize eq. (14) subject to the constraint that $a_i = \Theta(r_i - r_c)$, where $\Theta(\cdot)$ is the Heaviside step function.

Let us denote the optimal policy by π^* and the optimal value function by v^*. In the limit of large state spaces, $N \to \infty$, we have:

$$\mu^* = \frac{1}{N}\sum_i \Theta(r_i - r_c) \quad\rightarrow\quad \int_{r_c}^\infty dr\ \rho(r). \tag{16}$$

$$\omega^* = \frac{1}{N}\sum_i r_i\Theta(r_i - r_c) \quad\rightarrow\quad \int_{r_c}^\infty dr\ r\rho(r). \tag{17}$$

Note how sums over r_i converge to integrals over $\rho(r)$: this is because the rewards r_i are sampled independently from the distribution $\rho(r)$. It follows that the optimal expected return

$$v^* = \frac{\omega^*}{1-\Delta+\Delta\mu^*} \tag{18}$$

becomes an intrinsic property of the reward distribution $\rho(r)$ (as opposed to sampled values of r_i) in the limit $N \to \infty$.

To illustrate this, let us calculate v^* for the MDP whose mining rewards r_i are uniformly distributed over $[0,1]$:

$$\rho_u(r) = \begin{cases} 1 & \text{for } 0 \le r \le 1. \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

For this reward distribution, it is straightforward to evaluate the integrals in eqs. (16–17); substituting the results into eq. (18) gives:

$$v^* = \frac{1}{2}\left(\frac{1-r_c^2}{1-\Delta r_c}\right). \tag{20}$$

Maximizing this with respect to r_c, one obtains the final result:

$$v^* = \frac{1-\sqrt{1-\Delta^2}}{\Delta^2}, \tag{21}$$

with $r_c = \Delta v^*$. This is the optimal value function for the reward distribution $\rho_u(r)$ and arbitrary $\Delta < 1$. The calculation is equally straightforward for many other reward distributions.

3 Policy space

The form of the value function, eq. (14), gives rise to a simple two dimensional representation of the policy space. The coordinates in this representation are the parameters μ^π and ω^π, and each of the 2^N policies represents a point in the $\mu\omega$–plane. Note that different policies may map into the same point. Possible values

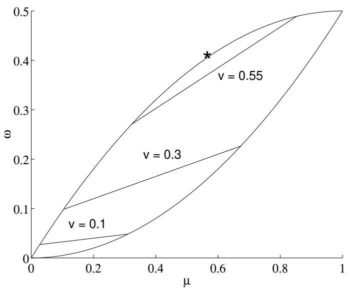

Figure 1: $\mu\omega$–representation of policy space. The lines represent policies with the same expected return. The asterisk denotes the location of the optimal policy.

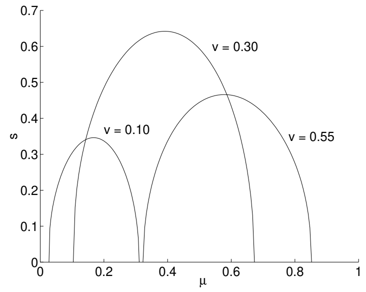

Figure 2: Cross-sectional plots of the entropy, $s(\mu,\omega)$, versus μ along lines of constant expected return $v = \frac{\omega}{1-\Delta+\Delta\mu}$ in the $\mu\omega$–plane.

of μ^π range from zero to one, with the extreme values corresponding to policies that uniformly EXPLORE or MINE. The range of ω^π in general depends on the values of the mining rewards, r_i. As mentioned earlier, we are interested in the limit of large state spaces, $N \to \infty$. In this limit, the parameters μ^π and ω^π assume a continuum of values, with the range of ω^π (for fixed μ^π) determined completely by the reward distribution, $\rho(r)$. The effect of this is that the policy space "fills in" a two dimensional region in the $\mu\omega$–plane.

Figure 1 shows this region for the MDP with mining probability $\Delta = 0.75$ and reward distribution $\rho_u(r)$. In this case the region is bounded by parabolas that delimit the minimum and maximum values of $\omega^\pi = \frac{1}{N}\sum a_i r_i$ for fixed $\mu^\pi = \frac{1}{N}\sum a_i$. These values correspond to policies whose MINE actions occur in the μth fraction of states with the lowest or highest rewards:

$$\omega^{\min}(\mu) = \int_0^\mu dr\ r = \frac{\mu^2}{2}. \qquad (22)$$

$$\omega^{\max}(\mu) = \int_{1-\mu}^1 dr\ r = \frac{\mu(2-\mu)}{2}. \qquad (23)$$

As before, we have used the limit $N \to \infty$ to replace sums over r_i by integrals over the reward distribution—in this case, $\rho_u(r)$ from eq. (19).

In this two-dimensional representation of policy space, there is an interesting geometric relationship among policies with the same expected return. In particular, from eq. (14) we see that all policies with $v^\pi = v$ lie on the line with slope Δv and intercept $(1-\Delta)v$ in the $\mu\omega$–plane. Some of these lines are shown in figure 1.

Entropy

Let us now examine the distribution of policies in the $\mu\omega$–plane. In particular, consider the indicator function

$$\Omega(\mu,\omega) = \sum_\pi \delta(\mu-\mu^\pi)\delta(\omega-\omega^\pi), \qquad (24)$$

where $\delta(\cdot)$ 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 $\Omega(\mu,\omega)$ defines a two-dimensional histogram that counts the number of policies with $\mu^\pi = \mu$ and $\omega^\pi = \omega$. The entropy

$$s(\mu,\omega) = \lim_{N\to\infty} \frac{1}{N} \ln \Omega(\mu,\omega) \qquad (25)$$

corresponds to this histogram on a log scale. In the limit $N \to \infty$, there emerges a continuum of values for μ and ω, and we expect $s(\mu,\omega)$ to be a smooth function of its arguments.

The entropy can be calculated from eq. (24) by rewriting the sum as an integral and using the method of saddle-point integration [6]. A calculation similar to the one in ref. [8] gives:

$$s(\mu,\omega) = \min_{\tilde\mu\tilde\omega} \left\{ -\tilde\mu\mu - \tilde\omega\omega + \int dr\ \rho(r) \ln\left[1 + e^{\tilde\mu + \tilde\omega r}\right] \right\}. \qquad (26)$$

Given a reward distribution $\rho(r)$, eq. (26) can be solved numerically for $s(\mu,\omega)$. Figure 2 shows some cross-sectional plots of $s(\mu,\omega)$ for the uniform distribution of rewards in eq. (19). These entropy curves are plotted along lines of constant expected return in the $\mu\omega$–plane: thus, the horizontal axis labels the μ coordinate while ω is determined implicitly by eq. (14).

Let us summarize the main points of this section. The policy space of the EXPLORE/MINE MDP can be represented as a two dimensional region in the $\mu\omega$–plane. The shape of this region, in the limit $N \to \infty$, is a property of the reward distribution $\rho(r)$ and the mining probability Δ. Policies with the same expected return v lie on straight lines of slope Δv and intercept $(1-\Delta)v$. Finally, the entropy $s(\mu,\omega)$, provides a smooth logarithmic measure of the number of policies with $\mu^\pi = \mu$ and $\omega^\pi = \omega$.

We have seen from eq. (14) that policies with the same parameters (μ,ω) also have the same expected return. The expected return, of course, only defines one statistic of the distribution from which these returns are sampled. One may also ask whether more general properties of this distribution are preserved by the $\mu\omega$–representation of policy space. Clearly, an important property (from the viewpoint of decision and control) is the probability that estimates of the expected return, based on finite exploration times $\mathcal{T}$, are in large error. How are these probabilities related for policies with the same parameters (μ, ω)? This is the subject of the next section.

4 Estimation errors

Consider how imperfect statistics undermine the agent's performance in the EXPLORE/MINE MDP with undiscounted rewards. In particular, let π^* denote the optimal policy, v^* its value function, and $\hat{v}^*$ an estimate of this value function based on eq. (7) and a random walk of length $\mathcal{T}$. Similarly, let v^π and $\hat{v}^\pi$ denote the corresponding statistics for another policy π. Assuming there is a unique optimal policy, it follows that $v^* > v^\pi$. For finite exploration times $\mathcal{T}$, however, it may happen that the empirical estimates are reversed: $\hat{v}^\pi > \hat{v}^*$. If so, the agent may be confused into adopting a suboptimal policy.

What is the probability of such an error, and how does it depend on the exploration time allowed for each policy evaluation? The probability of error clearly vanishes as $\mathcal{T} \to \infty$, since in this limit the empirical estimates converge to the true value functions. In fact, one can show this probability obeys a large deviation principle, decaying exponentially fast with $\mathcal{T}$:

$$\Pr[\hat{v}^\pi > \hat{v}^*] \sim e^{-\mathcal{T}\varepsilon^\pi}. \tag{27}$$

Here, the asymptotic notation "$\sim$" is used to hide multiplicative corrections that depend less strongly on $\mathcal{T}$; equivalently, we may write:

$$\lim_{\mathcal{T}\to\infty} \frac{1}{\mathcal{T}} \ln \Pr[\hat{v}^\pi > \hat{v}^*] = -\varepsilon^\pi. \tag{28}$$

The rate of decay, ε^π, is a function of the distributions (and thus ultimately, the policies) that generate the estimates $\hat{v}^\pi$ and $\hat{v}^*$; we have suppressed the dependence on π^* to avoid excessive notation. The goal of this section is to provide a lower bound on ε^π and thus an upper bound on the "confusion" probability that appears in eq. (27).

4.1 Large deviations

Our starting point is the following basic theorem from the large deviation theory of Markov processes[3]. Let P_{ij} be an $N \times N$ ergodic transition matrix with stationary distribution ϕ_i, and let X_i be a real-valued function over its state space. Without loss of generality, take[1] $\sum_i \phi_i X_i = 0$. Then for the Markov chain $\{i_t\}_{t=1}^{\mathcal{T}}$, the probability of overestimating $\mathrm{E}[X]$ decays as:

$$\Pr\left[\frac{1}{\mathcal{T}} \sum_{t=1}^{\mathcal{T}} X_{i_t} > \delta\right] \sim e^{-\mathcal{T}\varepsilon(\delta)}, \tag{29}$$

where

$$\varepsilon(\delta) = \max_{\psi \geq 0} \left\{\psi\delta - \ln \lambda(\psi)\right\}, \tag{30}$$

[1] This can always be achieved by adding a uniform constant to X_i (i.e. subtracting out the mean).

and $\lambda(\psi)$ is the largest eigenvalue of the "twisted" transition matrix, $e^{\psi X_i} P_{ij}$. Borrowing terminology from statistical mechanics, we will refer to matrices of the form $e^{\psi X_i} P_{ij}$ as transfer matrices[6]. The rate of decay $\varepsilon(\delta)$ is also known as the large deviation rate function[3].

A straightforward extension of this theorem is to consider pairs of independent Markov processes. In particular, let P_{ij} and P'_{ij} be $N \times N$ ergodic transition matrices with stationary distributions ϕ_i and ϕ'_i, and let X_i and X'_i be real-valued functions over their respective state spaces. Suppose moreover that $\sum_i \phi_i X_i < \sum \phi'_i X'_i$. Then if $\{i_t\}_{t=1}^{\mathcal{T}}$ and $\{i'_t\}_{t=1}^{\mathcal{T}}$ are two independent Markov chains generated by P and P',

$$\Pr\left[\frac{1}{\mathcal{T}} \sum_{t=1}^{\mathcal{T}} X_{i_t} > \frac{1}{\mathcal{T}} \sum_{t=1}^{\mathcal{T}} X'_{i'_t}\right] \sim e^{-\mathcal{T}\varepsilon}, \tag{31}$$

where

$$\varepsilon = \max_{\psi \geq 0} \left\{-\ln\left[\lambda(\psi)\lambda'(-\psi)\right]\right\}, \tag{32}$$

and $\lambda(\psi)$ and $\lambda'(-\psi)$ are the largest eigenvalues of the transfer matrices, $e^{\psi X_i} P_{ij}$ and $e^{-\psi X'_i} P'_{ij}$. Eq. (32) may be derived by expressing the two independent Markov processes as a "meta-process" whose transition matrix is the cartesian product of P and P'. The result then follows from the previous theorem.

Let us now consider the decay rate ε^π that appears in eq. (27). The empirical estimates of the value functions, $\hat{v}^\pi$ and $\hat{v}^*$, are sums over time with the accumulated rewards playing the role of X_i and X'_i in eq. (31). Likewise, the paths through state space that dictate these rewards are generated by independent Markov processes (i.e. the random walks from π and π^*). Applying the previous theorem gives:

$$\varepsilon^\pi = \max_{\psi \geq 0} \left\{-\ln\left[\lambda^\pi(\psi)\lambda^*(-\psi)\right]\right\}, \tag{33}$$

where $\lambda^\pi(\psi)$ is the largest eigenvalue of the transfer matrix $e^{\psi R_i^\pi} P_{ij}^\pi$, and $\lambda^*(\psi)$ is the corresponding eigenvalue for the optimal policy, π^*.

4.2 Eigenvalues

Eq. (33) highlights the special role played by the largest eigenvalue of the transfer matrix $e^{\psi R_i^\pi} P_{ij}^\pi$. The form of the EXPLORE/MINE MDP enables one to derive a number of expressions satisfied by this eigenvalue. For example, we may obtain $\lambda^\pi(\psi)$ exactly by computing the largest solution to the equation:

$$\lambda = 1 - \mu^\pi + \left(\frac{1-\Delta}{N}\right) \sum_i a_i \left[\frac{\lambda}{\lambda e^{-\psi r_i} - \Delta}\right]. \tag{34}$$

The proof of this identity is straightforward and given in the appendix. In an effort to improve the readability of eq. (34), we have suppressed the dependence of λ on π and ψ. In general, we will follow this convention for the equations in this section while continuing to write out the full dependencies (e.g. $\lambda^\pi(\psi)$) in the surrounding text.

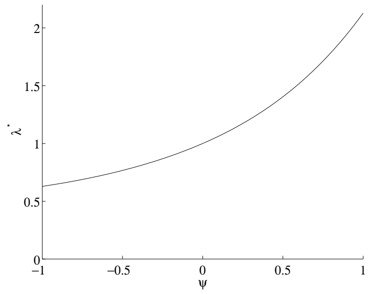

Figure 3: Plot of $\lambda^*(\psi)$, the largest eigenvalue of the optimal policy's transfer matrix, obtained from eq. (36).

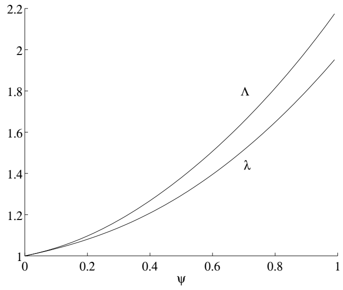

Figure 4: Plot of $\lambda^\pi(\psi)$ and its upper bound $\Lambda^\pi(\psi)$ for a randomly generated policy $\pi \in \{0,1\}^{100}$.

Optimal policy

Eq. (34) is particularly useful for computing $\lambda^*(\psi)$ in the limit of large state spaces, $N \to \infty$. In this case, $a_i = \Theta(r_i - r_c)$, where r_c is the critical reward value above which the agent decides to MINE. Substituting this into eq. (34) and taking the limit $N \to \infty$ gives:

$$\lambda^* = 1 - \mu^* + (1-\Delta)\int_{r_c}^{\infty} \frac{\lambda^* \rho(r)\; dr}{\lambda^* e^{-\psi r} - \Delta}. \tag{35}$$

For fixed Δ and reward distribution $\rho(r)$, eq. (35) can be solved numerically to obtain $\lambda^*(\psi)$ as a function of ψ.

As an example, consider again the uniform distribution of rewards from eq. (19). The value of r_c for this distribution was calculated at the end of section 2.3. Using $\rho_u(r)$ to evaluate the right hand side of eq. (35), we find:

$$\lambda^* = 1 - \mu^* + \frac{\lambda^*(1-\Delta)}{\Delta\psi} \ln\left[\frac{\lambda^* - \Delta e^{\psi \Delta v^*}}{\lambda^* - \Delta e^{\psi}}\right], \tag{36}$$

with $\mu^* = 1 - r_c$ and v^* given by eq. (21). Figure 3 shows a plot of $\lambda^*(\psi)$ obtained by numerically solving eq. (36) with $\Delta = 0.75$.

$\mu\omega$–policies

Though we have used eq. (34) to evaluate $\lambda^\pi(\psi)$ for the optimal policy π^*, it is clearly impractical to do this for each of the 2^N policies $\pi \in \{0,1\}^N$. Instead, we shall obtain a more useful characterization of $\lambda^\pi(\psi)$ by exploiting the $\mu\omega$–representation of policy space introduced in section 3. Our main result, proved in the appendix, is an upper bound on $\lambda^\pi(\psi)$ expressed in terms of the parameters μ^π and ω^π. There we show that if the rewards r_i are bounded between zero and one, then we obtain an upper bound $\Lambda^\pi(\psi) \geq \lambda^\pi(\psi)$ by computing the largest solution of the equation:

$$\Lambda = 1 + \left[\frac{\Delta(1-\Lambda)}{\Lambda - \Delta}\right]\mu^\pi + \left[\frac{\Lambda^2(1-\Delta)(1-e^{-\psi})}{(\Lambda e^{-\psi} - \Delta)(\Lambda - \Delta)}\right]\omega^\pi. \tag{37}$$

Because the bound $\Lambda^\pi(\psi)$ depends on π only through the parameters μ^π and ω^π, it extends naturally to the limit of large state spaces where we view these parameters as smoothly varying coordinates in the two-dimensional representation of policy space.

The accuracy of the bound is easily tested in large but finite ($N = 100$) state spaces. Figure 4 plots the upper bound $\Lambda^\pi(\psi)$ and the actual eigenvalue $\lambda^\pi(\psi)$ for a randomly chosen policy $\pi \in \{0,1\}^{100}$. Here, $\Lambda^\pi(\psi)$ was computed by solving eq. (37), while $\lambda^\pi(\psi)$ was computed by diagonalizing the 100×100 transfer matrix. As before, we used mining probability $\Delta = 0.75$ and rewards generated from the uniform distribution, $\rho_u(r)$.

4.3 Decay rates

We obtain a lower bound on the decay rate by performing the maximization

$$\varepsilon_l(\mu^\pi, \omega^\pi) = \max_{\psi \geq 0}\left\{-\ln\left[\Lambda^\pi(\psi)\lambda^*(-\psi)\right]\right\}. \tag{38}$$

The only difference between eq. (33) and eq. (38) is that we have substituted the upper bound $\Lambda^\pi(\psi)$ in place of the true eigenvalue $\lambda^\pi(\psi)$. Though not exact, this suffices to obtain a lower bound, $\varepsilon^\pi \geq \varepsilon_l(\mu^\pi, \omega^\pi)$.

Like $\Lambda^\pi(\psi)$, this bound depends on π only through the parameters μ^π and ω^π. Hence, eq. (38) provides the same upper bound on the decay rate for those policies represented by the same point in the $\mu\omega$–plane. At the end of section 3, we asked how the probabilities of estimation errors were related for policies with the same values of (μ, ω). The uniformity of our bound provides at least a partial answer to this question.

Figure 5 shows the results of one of these maximizations for the EXPLORE/MINE MDP with $\Delta = 0.75$ and reward distribution $\rho_u(r)$. The curve in the figure was calculated for a policy with $(\mu^\pi, \omega^\pi) = (0.55, 0.33)$, as compared to the optimal policy at $(\mu^*, \omega^*) = (0.55, 0.40)$. As these policies are rather close in the $\mu\omega$–plane, the bound on the decay rate is quite small: $\varepsilon_\pi \geq 0.0028$. This in turn suggests that the confusion probability $\Pr[\hat{v}^\pi > \hat{v}^*] \sim e^{-\mathcal{T}\varepsilon_\pi}$ remains substantial unless $\mathcal{T} \gg (0.0028)^{-1} \approx 350$.

5 Performance Bounds

In this section we evaluate the performance of the Gibbs algorithm in the so-called thermodynamic limit:

$$\mathcal{T} \to \infty, \quad N \to \infty, \quad \alpha = \tfrac{\mathcal{T}}{N} \quad \text{(finite)}. \tag{39}$$

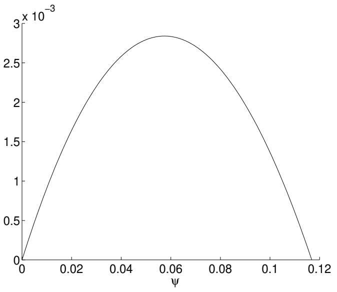

Figure 5: Maximization of $-\ln[\Lambda^\pi(\psi)\lambda^*(-\psi)]$ over ψ from eq. (38), whose result gives a lower bound on the decay rate, ε^π.

This is the combined limit that the exploration time per policy evaluation and the size of the state space grow to infinity at a fixed rate. The ratio α measures the exploration time $\mathcal{T}$ in what are natural units for an MDP with undiscounted rewards—namely the size of the state space, N. Why are these natural units? The reason is that in order to accurately estimate a policy's expected return, it may be necessary to explore the state space several times over. Large values of α ensure that each policy evaluation is based on enough steps to perform this exploration.

Let us sketch why the *combined* limit in eq. (39) is necessary for interesting learning behavior[5]. Recall that the Gibbs algorithm estimates each value function v^π by a random walk of length $\mathcal{T}$, then outputs the policy with the best empirical return, $\hat{v}^\pi$. As $\mathcal{T} \to \infty$, the "confusion" probability $\Pr[\hat{v}^\pi > \hat{v}^*]$ becomes exponentially small for any non-optimal policy π. As $N \to \infty$, however, there arise an exponentially large number of non-optimal policies. Roughly speaking, in the thermodynamic limit these two effects "balance out" to generate interesting learning behavior as a function of the ratio $\alpha = \mathcal{T}/N$. On the other hand, the Gibbs learner exhibits quite trivial behavior if only $\mathcal{T} \to \infty$ (always selecting the optimal policy) or $N \to \infty$ (never obtaining adequate statistics).

Developing these ideas further, let $\Pr[\pi^{\text{gibbs}} \in (\mu, \omega)]$ denote the probability that the Gibbs algorithm outputs a policy with $(\mu^\pi, \omega^\pi) = (\mu, \omega)$. Then we have the following chain of inequalities:

$$\begin{aligned} \Pr\left[\pi^{\text{gibbs}} \in (\mu,\omega)\right] &\leq \Pr\left[\exists \pi \in (\mu,\omega) \text{ with } \hat{v}^\pi > \hat{v}^*\right] \\ &\leq \sum_{\pi \in (\mu,\omega)} \Pr[\hat{v}^\pi > \hat{v}^*]. \end{aligned} \tag{40}$$

The first inequality notes that at least one policy $\pi \in (\mu, \omega)$ must register a higher empirical return than $\hat{v}^*$ for $\pi^{\text{gibbs}} \in (\mu, \omega)$. In general, however, this is not a sufficient condition since we also require that no other policy with different coordinates samples a higher return. The second inequality, eq. (40), follows from a standard union bound.

Let us now recall some of our previous results. From section 3, we know that in the limit of large state spaces, the entropy $s(\mu, \omega)$ provides a smooth logarithmic measure of the number of policies $\pi \in (\mu, \omega)$. From section 4, we have a lower bound $\varepsilon_l(\mu, \omega)$ on the asymptotic decay rate that characterizes $\Pr[\hat{v}^\pi > \hat{v}^*] \sim e^{-\mathcal{T}\varepsilon^\pi}$. Roughly speaking, then, for sufficiently large values of $\mathcal{T}$ and N, we have:

$$\sum_{\pi \in (\mu,\omega)} \Pr[\hat{v}^\pi > \hat{v}^*] \leq e^{N s(\mu,\omega) - \mathcal{T}\varepsilon_l(\mu,\omega)} \tag{41}$$

More formally, we can take the thermodynamic limit in eq. (39) to obtain:

$$\lim \left\{ \frac{1}{N} \ln \Pr\left[\pi^{\text{gibbs}} \in (\mu,\omega)\right] \right\} \;\leq\; s(\mu,\omega) - \alpha\varepsilon_l(\mu,\omega). \tag{42}$$

Note that if the right hand side of eq. (42) is less than zero, then the probability $\Pr[\pi^{\text{gibbs}} \in (\mu, \omega)]$ can be said to *vanish* in the thermodynamic limit. In general, this occurs at values of $\alpha = \mathcal{T}/N$ for which the Gibbs learner has sufficient statistics to rule out *all* the (suboptimal) policies $\pi \in (\mu, \omega)$.

The competition between $s(\mu, \omega)$ and $\varepsilon_l(\mu, \omega)$ to determine the sign of eq. (42) is the balancing act that gives rise to interesting behavior in the thermodynamic limit. Note how the parameter α modulates the relative contribution of these quantities to the overall value of the right hand side. The critical value

$$\alpha_c(\mu,\omega) = \frac{s(\mu,\omega)}{\varepsilon_l(\mu,\omega)} \tag{43}$$

is the value of α above which the Gibbs learner is able to eliminate from consideration (with probability one) all the policies $\pi \in (\mu, \omega)$. Thus the contours of constant $\alpha_c(\mu, \omega)$ in the $\mu\omega$–plane enclose regions whose policies have yet to be eliminated as candidates for the Gibbs policy. In particular, $\alpha_c(\mu, \omega) = 0$ encloses the entire policy space, while $\alpha_c(\mu, \omega) = \infty$ consists of a single point—the coordinates of the optimal policy, (μ^*, ω^*). Figure 6 shows several of these contours for the EXPLORE/MINE MDP with $\Delta = 0.75$ and reward distribution, $\rho_u(r)$. Note how the enclosed regions shrink in size as the value of α_c is increased.

In analogy to eq. (43), we may also define a critical value $\alpha_c(v)$ that suffices to rule out policies with expected return $v^\pi < v$. Recall that policies with $v^\pi = v$ lie on lines with slope Δv and intercept $(1 - \Delta)v$ in the $\mu\omega$–plane. Hence the value of $\alpha_c(v)$ may be found by maximizing eq. (43) over the range of μ and ω that lie on this line.

The learning curve for an MDP measures how the agent's performance improves with the allowed exploration time per policy evaluation, $\mathcal{T}$. In the thermodynamic limit, the exploration time is measured in units of the size of the state space, or $\alpha = \mathcal{T}/N$. Because the Gibbs algorithm is guaranteed to output a policy with $v^{\text{gibbs}} \geq v$ for exploration times $\alpha > \alpha_c(v)$, we obtain a lower bound on the agent's performance by plotting v versus $\alpha_c(v)$.

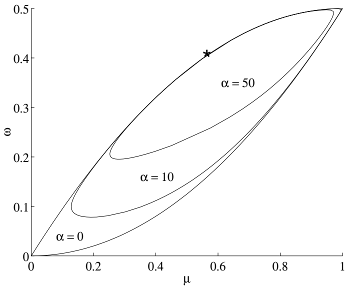

Figure 6: Contours of constant $\alpha_c(\mu, \omega)$ in the $\mu\omega$–plane. The contours enclose policies that remain candidates for the Gibbs policy. As $\alpha \to \infty$, all policies (except the optimal one) are eliminated from consideration. The asterisk denotes the location of the optimal policy.

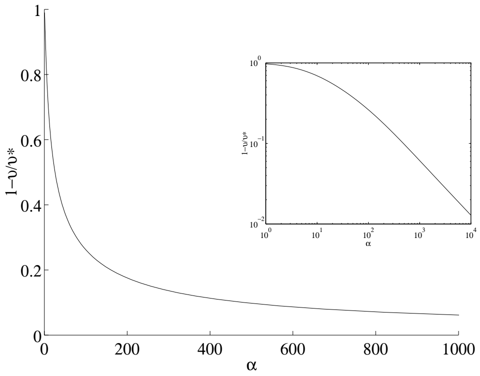

Figure 7: Upper bound on $(1 - v^{\text{gibbs}}/v^*)$ versus α for the EXPLORE/MINE MDP with $\Delta = 0.75$ and reward distribution, $\rho_u(r)$. The inset shows a log-log plot extended to larger values of α.

Figure 7 shows a plot of $(1 - v/v^*)$ versus $\alpha_c(v)$ for the EXPLORE/MINE MDP with $\Delta = 0.75$ and reward distribution, $\rho_u(r)$. The inset shows the plot on a logarithmic scale to highlight the asymptotic behavior of the bound. In particular, as $\alpha \to \infty$, the upper bound on $(v^* - v^{\text{gibbs}})$ behaves asymptotically as

$$(v^* - v^{\text{gibbs}}) \sim \frac{1}{\alpha^z} \text{ with } z = \frac{2}{3}, \tag{44}$$

where the exponent $z = 2/3$ is determined by the slope of the log-log plot at large α. The exponent can also be verified analytically by examining the behavior of $s(\mu, \omega)$ and $\varepsilon_l(\mu, \omega)$ in the neighborhood of the optimal policy. Nevertheless, we do not know of any obvious method by which this power law could have been guessed beforehand.

6 Conclusions

In this paper we have used methods from statistical mechanics to study the problem of decision and control based on imperfect statistics. This was done in the framework of Markov decision processes with undiscounted rewards.

One virtue of our approach is that for simple examples we can understand in great detail how the agent's performance improves with its capacity to acquire more accurate statistics. The shrinking regions in figure 6 and the performance curve in figure 7 reveal the learning behavior from start to finish—that is, from a state of impoverished statistics to one of perfect knowledge. They also reveal asymptotic rates of convergence that could not be predicted from simple statistical considerations. This detailed picture of learning is to be contrasted with the much weaker statement that optimal control emerges in the limit that the agent visits each state infinitely often. Yet even this weaker statement remains an open question for many simulation-based algorithms used to solve MDPs with undiscounted rewards[10].

An important lesson from previous work in supervised learning is that the shapes of learning curves are not universal and vary from problem to problem. We expect the same to be true for MDPs; thus even within the EXPLORE/MINE MDP, it seems likely that changing the reward distribution $\rho(r)$ could affect the asymptotic rate of convergence for our bounds on $(v^* - v^{\text{gibbs}})$. This suggests two goals for future research: uncovering the variety of learning behaviors that can occur in MDPs, and understanding the features that make it easier or harder to learn optimal policies. These issues and others are left for future work.

Acknowledgements

The authors thank T. Jaakkola and P. Dayan for useful discussions. LS also acknowledges support from NSF grant CDA-9404932.

References

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

[2] D. P. Bertsekas. *Dynamic programming and optimal control*, vols. 1 & 2. Athena Scientific, Belmont, MA, 1995.

[3] J. A. Bucklew. *Large deviation techniques in decision, simulation, and estimation.* John Wiley & Sons, New York, 1990.

[4] C. N. Fiechter. Efficient reinforcement learning. In *Proceedings of the 7th Annual ACM Workshop on Computational Learning Theory*, pages 88–97. Morgan Kauffman, San Mateo, CA 1994.

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

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

[7] M. Marcus and H. Minc. *A survey of matrix theory and matrix inequalities.* Dover, New York, 1992.

[8] L. K. Saul and S. P. Singh. Markov decision processes in large state spaces. In *Proceedings of the 8th Annual Workshop on Computational Learning Theory,* pages 281–288. ACM Press, New York, 1995.

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

[10] S. P. Singh. Reinforcement learning algorithms for average-payoff markovian decision problems. In *Proceedings of the 12th National Conference on Artificial Intelligence,* pages 700–706. AAAI Press, Seattle, 1994.

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

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

A Eigenvalues

In this appendix we derive eq. (34) for the largest eigenvalue of the transfer matrix, $e^{\psi R_i^\pi} P_{ij}^\pi$, and eq. (37) for its upper bound. To this end, let ξ_i denote the elements of the largest eigenvector of this matrix and λ its corresponding eigenvalue:

$$\sum_j e^{\psi R_i^\pi} P_{ij}^\pi \xi_j = \lambda \xi_i. \tag{45}$$

The elements of the matrix $e^{\psi R_i^\pi} P_{ij}^\pi$ are all positive, and hence by the Perron-Frobenius theorem[7], we know that: (i) the largest eigenvalue is positive; (ii) the eigenvector corresponding to this eigenvalue has only positive elements; and (iii) no other eigenvector has this property. From (i), we can assume that the eigenvector satisfies the normalization condition $\sum_i \xi_i = 1$. Using this, it is straightforward to show that

$$\xi_i = \frac{1}{N\lambda}(1-a_i) + \frac{1}{N}\left[\frac{(1-\Delta)}{\lambda e^{-\psi r_i} - \Delta}\right] a_i \tag{46}$$

solves eq. (45), with λ chosen in a self-consistent fashion to satisfy $\sum_i \xi_i = 1$. Summing both sides of eq. (46) over i to enforce this constraint, we obtain:

$$1 = \frac{1-\mu^\pi}{\lambda} + \left(\frac{1-\Delta}{N}\right) \sum_i \frac{a_i}{\lambda e^{-\psi r_i} - \Delta}, \tag{47}$$

which is equivalent to eq. (34). Hence, by solving this equation (numerically), we obtain an eigenvalue of the transfer matrix whose eigenvector is given by eq. (46).

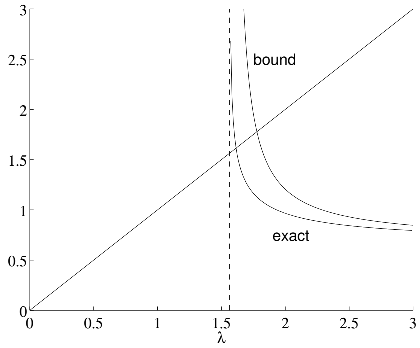

Figure 8: Graphical solution of eqs. (34) and (37). The dashed line is the pole that occurs at $\lambda_c = \max_i \left[a_i \Delta e^{\psi r_i}\right]$ in eq. (34). The lower point of intersection is the solution for the exact eigenvalue, λ; the upper point is the solution for its upper bound.

To ensure that our solution corresponds to the *largest* eigenvalue, it is enough to show (by the Perron-Frobenius theorem) that all the elements of ξ_i are positive. From eq. (46) this will be true if $\lambda > \lambda_c = \max_i \left[a_i \Delta e^{\psi r_i}\right]$. Note that the right hand side of eq. (34) has a pole at $\lambda = \lambda_c$ and that above this value, it monotonically decreases, approaching a finite value as $\lambda \to \infty$. Figure 8 illustrates this behavior, along with the graphical solution to eq. (34) for a random policy $\pi \in \{0,1\}^{100}$; the solution occurs where the plots of the left and right hand sides intersect. It is easy to see from the figure that the equation has a unique solution to the right of the pole at λ_c. Hence, this largest solution corresponds to the maximal eigenvalue of the transfer matrix, as claimed in section 4.2.

To obtain an upper bound on this eigenvalue, we note that the summand in eq. (47) is a convex function of r_i. In particular, suppose that r_i is bounded between 0 and 1. Then:

$$\frac{1}{\lambda e^{-\psi r_i} - \Delta} \leq \frac{1}{\lambda - \Delta} + \left[\frac{1}{\lambda e^{-\psi} - \Delta} - \frac{1}{\lambda - \Delta}\right] r_i, \tag{48}$$

where the upper bound in eq. (48) is the linear function in r_i that interpolates between the left hand side's value at $r_i = 0$ and $r_i = 1$. Substituting this linear function of r_i not only leads to an upper bound, but also enables one to perform the sum over states in terms of the variables $\mu^\pi = \frac{1}{N}\sum_i a_i$ and $\omega^\pi = \frac{1}{N}\sum_i a_i r_i$. Thus, substituting eq. (48) into eq. (47) and evaluating the sum over states gives (after some algebra):

$$\lambda \leq 1 + \left[\frac{\Delta(1-\lambda)}{\lambda - \Delta}\right] \mu^\pi + \left[\frac{\lambda^2(1-\Delta)(1-e^{-\psi})}{(\lambda e^{-\psi} - \Delta)(\lambda - \Delta)}\right] \omega^\pi. \tag{49}$$

Replacing this inequality by equality, we recover the prescription of eq. (37) for obtaining an upper bound on the eigenvalue λ. The graphical solution to this equation is also shown in Figure 8, where it is easily seen that it yields an upper bound on the solution to eq. (34).