Squirrelly bandits

Discounting, reinforcement learning, and evolution

Eric Rozon
Department of Mathematics
University of British Columbia
Math 564

Abstract

We put forward a simple, individual based model in which agents face an exploration/exploitation tradeoff. Optimal behavioural policies are analysed both from reinforcement learning (RL) and fitness optimization perspectives. Having fixed a minimal set of model parameters, we find the RL and evolutionary approaches can be made equivalent by appropriate choice of discounting parameter γ in the RL setup. Our results lend support to the use of exponential discounting in reinforcement learning algorithms.

Keywords— Multi-armed bandit, reinforcement learning, evolution, exploitation, exploration, foraging.

Contents

1 Background and motivation

Agents (people, algorithms, squirrels) tend to prefer larger rewards to smaller ones. At the same time, they tend to prefer immediate (as opposed to delayed) gratification. Intuitively, we see two rationales justifying impatience.

1. Risk that a future reward not be received (due to death of recipient, for instance).
2. Opportunity cost – an immediate reward can be translated into further compounding benefits, whereas while waiting for a future reward, we cannot compound.

For a broad review of discounting, see [FLO02]. The quantitative discounting problem with which we concern ourselves presently is: how ought an agent to decide between flows of rewards $r = (r_0, r_1, r_2, \ldots)$ and $r' = (r'_0, r'_1, r'_2, \ldots)$? Of particular interest is the case where r' is initially lower than r, but eventually gets larger. For instance, compare

- $r = (1, 1, 1, 1, \ldots)$ and
- $r' = (0, 0, 0, \ldots, 0, 2, 2, 2, 2, \ldots)$.

An agent's preference for larger rewards is in opposition to its impatience. This case is of particular interest to us, and finds applications in all of biology, zoology, economics, psychology, and computer science, to name but a few.

1.1 Classical discounting approach

Determining the present value of future rewards is common in economic problems. The approach almost universally used is to discount future rewards to present value, and pick the alternative resulting in maximal present value. A one time reward F received at a delay $t \geq 0$ is discounted to present value P via $P = \Delta(t) \cdot F$. We call $\Delta(t)$ a *discounting function*.

An agent using discounting function $\Delta(t)$ evaluates the present value of flow $(r_0, r_1, r_2, \ldots)$ as

$$P = \sum_{t=0}^{\infty} \Delta(t) \cdot r_t.$$

1.1.1 Hyperbolic discounting

Experiments with both human and non-hum animals find that *hyperbolic discounting* is prevalent; see [Soz98] for a review. Under hyperbolic discounting, agents use

$$\Delta(t) = \frac{1}{1 + ht}$$

where $h > 0$. Intuitive rationale for hyperbolic discounting is scarce, and seemingly irrational behaviour abound. The reader will with enthusiasm verify that *preference reversals* are predicted by hyperbolic discounting: if $F_1 < F_2$ and $t < s$ one finds that

$$\frac{F_1}{1 + ht} > \frac{F_2}{1 + hs}, \qquad \text{whereas} \qquad \frac{F_1}{1 + h(t + t')} < \frac{F_2}{1 + h(s + t')}$$

for appropriate choice of t'. The imposition of an additional delay to receipt of future rewards can propmpt agents to change from a smaller, more immediate reward to a larger, more delayed reward.

1.1.2 Exponential discounting

In contrast to hyperbolic discounting, *exponential discounting* finds plenty of motivation in simple and intuitive models, whereas it is less commonly observed empirically. At a high level, the models which motivate exponential discounting posit that a risk neutral agent

should be indifferent between rewards which are equal in expectation, so that we have

$$\Delta(t) = e^{-rt}$$

for some $r > 0$. Exponential discounting also finds natural motivation in fincance, but "that is another story and shall be told another time." Exponential discounting's abundance of intuitive rationale, as well as well as its simplification of analytic formulae, have made it the most popular choice in economic models involving intertemporal choice.

1.2 Reinforcement learning

It is beyond the scope of this project to provide anything resembling an appropriate treatment of reinforcement learning (RL) to the uninitiated (such as the author). Nevertheless, a high level description of some processes is possible, and provides thought provoking connection to the theory of discounting. Reinforcement learning is so called because it involves an agent who interacts with its environment, receives feedback (in the form of a reward), and then repeats the process. Repeated interactions with an environment result in an RL

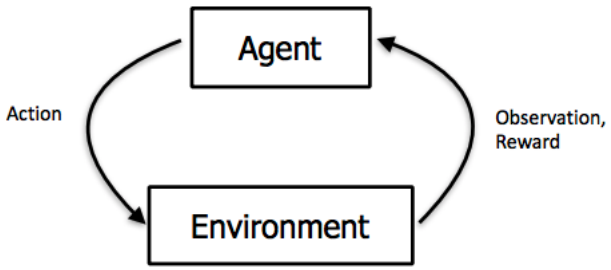

Figure 1.1: *Reinforcement learning, at a high level.*

agent receiving a (potentially stochastic) stream of rewards $(r_0, r_1, r_2, \ldots)$. The process via which an RL agent makes its decisions is called its *policy*, denoted π. For an agent with policy

π, denote its *expected reward* (at time t) by $\mathbb{E}_\pi[r_t]$. RL agents wish to choose an optimal policy π^*; standard practice in applied and theoretical RL is to fix a *discount parameter* $\gamma \in (0,1)$, and choose a policy π^* which maximizes its total discounted reward:

$$\pi^* \in \operatorname{argmax}_\pi \sum_{t=1}^{\infty} \gamma^t \mathbb{E}_\pi[r_t].$$

RL algorithms almost universally discount exponentially, with little in the way of rigorous theoretical justification for this choice. Furthermore, while the discounting parameter γ almost certainly impacts optimal policy choice, how one is to choose $\gamma \in (0,1)$ is most often arbitrary. See [FGB+19] for some background on the problem of intertemporal discounting in RL processes.

1.3 Evolutionary methods and perspectives

Fitness is the currency of evolutionary theory. For some background, see [SK13]. Throughout, we assume all populations considered grow exponentially, that is, according to $\text{Pop}(t) = \text{Pop}_0 e^{rt}$. Fitness is then defined by the *Lyapunov exponent*, r, which is the population's unit time growth rate. A *demographic matrix* is a square matrix $\mathcal{B} = (b_{ij})_{i,j=1}^n$, where b_{ij} is the probability that an individual move to category i from j. It is a standard result in the theory of structured population growth that the largest positive eigenvalue of $\mathcal{B}$ is its stable growth rate, that is, its Lyapunov exponent, or our chosen measure of fitness.

1.4 Motivating philosophy

Conjecture 1.1. *Humans and non-human animals discount hyperbolically because it is evolutionarily optimal to do so.*

Beyond simplisti models coming from risk analysis or financial math, there is little reason to assume agents discount exponentially. Nevertheless, exponential discounting is omnipresent, and in particular is used as shown by RL algorithms. In this paper, we proceed to introduce a simple model which can be analysed both as an evolutionary process and as an RL problem. We seek to understand if the exponential discounting employed in RL algorithms can be reconciled with a natural, evolutionary, optimzation process. While a priori the two approaches seem entirely disconnected and independent, we will see that they are intricately linked.

2 The model

2.1 The bandit

Let us assume a biological setting. Model the daily foraging for nuts by a squirrel via a bandit (slot machine), and assume that payoffs from the bandit are translated into reproductive rewards. That is, if the bandit pays β, then we interpret that a squirrel has β offspring.

- Bandit with infinitely many arms.
- Three types of arms:
    1. null giving arms (no reward received when pulled)
    2. low giving arms (reward of β_l)
    3. high giving arms (reward of β_h)
- Null, low, and high arms have frequency $(1-l-h), l$, and h respectively and each one in $[0,1]$.

2.2 Model squirrels

When playing with the bandit, squirrels face an exploration/exploitation tradeoff when at a low arm. Choosing to stay results in a guaranteed, albeit low payout. Choosing to explore results in a duration of time during which null and low arms are pulled, until after a delay (in expectation of length $1/h$) a high arm is pulled. To build our model, we need some assumptions on squirrels.

- All squirrels are biologically the same. Squirrels *only* vary in searching policy.
- Squirrels survive each night with probability $s \in [0, 1]$.
- Squirrels, at the beginning of each day, can choose to *stay* or *search*.
  - To stay means to pull the exact same arm as yesterday, receiving the same reward.
  - To search means to go to a different arm. Since we assume infinitely many arms, being at a specific arm type does not vary the probability of finding other arm types. Searching incurs no direct cost.
- Rewards translate to offspring directly. Assume offspring start without arm assignment, and so search immediately by default. Reproduction is asexual – a single squirrel pulls an arm and gets offspring. In fact, squirrels do not interact in any way.
- Terminology: a squirrel is defined by its *search policy*, which we will denote π.

2.3 Policy space and fitness

Consider an evolutionary setting, within which model squirrels inherit the search policy of their parent. We wonder: which strategy optimizes fitness? To address optimal fitness concerns, consider the following reduction.

- At a high arm, model squirrels should never go searching since they won't get anything higher.
- At a null arm, model squirrels should always search since you can't possibly do any worse and searching incurs no direct cost.
- Therefore, **the only time model squirrels ever wonder whether to search is when they are at a low arm**.

We consider the greatly reduced space of search policies $\{\pi_\varepsilon : \varepsilon \in [0,1]\}$, where having policy π_ε means that, when at a low arm, you search with probability ε and stay otherwise. From our setup naturally emerges a demographic matrix. For a model squirrel with search policy π_ε, set

$$\mathcal{B}_\varepsilon = \begin{pmatrix} 1-l-h & \beta_l + \varepsilon(1-l-h) & \beta_h \\ l & (1-\varepsilon)+\varepsilon l & 0 \\ h & \varepsilon h & 1 \end{pmatrix} \cdot s.$$

With the machinery of $\mathcal{B}_\varepsilon$ in place, we can address the question of optimal fitness. We optimize fitness by choosing ε^* so as to maximize the leading eigenvalue of $\mathcal{B}_\varepsilon$:

$$\varepsilon^* = \text{argmax}_{\varepsilon\in[0,1]} \text{ Leading Eigenvalue } (\mathcal{B}_\varepsilon) .$$

2.4 Example and questions

The model in abstract can be a bit tough to wrap one's head around, so here is a worked example of the life of a single model squirrel, who we call Sam. Assume that $l = 0.3$, $h = 0.05$, $\beta_l = 1$, $\beta_h = 10$, and $s = 0.99$. Suppose Sam's search policy involves $\varepsilon = 0.5$.

1. Sam is born without an assigned arm, so he searches. He finds an arm, pulls it, and finds no nuts. He is at a null arm!

2. The next day, Sam has no desire to stay at his null arm, so he searches for another one. Low and behold, he pulls another arm which gives him no nuts.

3. The following day, Sam again chooses to search. He finds a new arm, pulls it, and gets one nut. He has found a low arm, and is relieved. As a result of finding one nut, he produces an offspring, who goes on to live a life analogous to Sam (but entirely independently – offspring do not follow their parents).

4. Today, Sam faces a dilemma. If he leaves his current arm in favour of trying to find a high arm, he will likely find nothing for a while. He flips a coin, and chooses to stay, producing another offspring in the process.

5. Having chosen to stay yesterday, Sam again flips a coin today. It comes up opposite to yesterday, and so he chooses to search.

6. ...

7. Eventually, after this process goes on for quite some time, Sam pulls a high arm and therefore stays there until eventually his 1 in 100 luck runs out and he dies.

Given parameters $l, h, \beta_l, \beta_h,$ and s, we propose the following list of questions.

- What search probability ε^* optimizes population growth?
- For fixed discount parameter $\gamma \in (0, 1)$, what search probability $\varepsilon_{\text{RL}}(\gamma)$ maximizes the reinforcement learning objective function?
- How do each of ε^* and $\varepsilon_{\text{RL}}(\gamma)$ vary as functions of the model parameters l, h, β_l, β_h, and s?

- What is the relationship between ε^* and $\varepsilon_{\text{RL}}(\gamma)$? Are they at all similar, or fundamentally different? Can the evolutionary objective of fitness maximization be expressed in the language of reinforcement learning?

3 Results

Proposition 3.1. *The evolutionarily optimal search probability ε^* is independent of survival probability s.*

Proof. If A is any matrix, then $Av = \lambda v \iff sAv = s\lambda v$. So all eigenvalues are multiplied by s, and thus the maximal eigenvalue of our demographic matrix corresponds to the same value ε^* for all $s \in (0, 1]$. □

Proposition 3.2. *For any model parameters $l, h, \beta_l, \beta_h, s$, and γ, both the evolutionarily and the RL optimal search probabilities are corner solutions: $\varepsilon^*, \varepsilon_{RL}(\gamma) \in \{0, 1\}$.*

Remark. This result hold numerically across a wide variety of parameter values. An analytic proof has yet to be found.

Theorem 3.3. *Let the low arm payout $\beta_l > 0$ be fixed, and set $s = 1$, so that model squirrels live forever. Then there exists a unique $\gamma_1^* \in (0, 1)$ such that, for every l, h, and β_h, we have $\varepsilon^* = \varepsilon_{RL}(\gamma_1^*)$. More explicitly,*

$$\varepsilon^*(l, h, \beta_l, \beta_h, s) = \text{argmax}_\varepsilon \sum_{t=0}^{\infty} (\gamma_1^*)^t \, \mathbb{E}_\varepsilon [r_t] .$$

Remark. In words, the theorem says that for fixed low payout, the problem of optimizing fitness is equivalent to a uniquely determined RL optimization process. The theorem provides strong evidence for exponential discounting by squirrels in this model setup. As with the previous proposition, this result holds true numerically for all parameters tested, but no analytic argument has been found. Figure 3.1 displays a numerical estimate of $\gamma_1^* = \gamma_1^*(\beta_l)$.

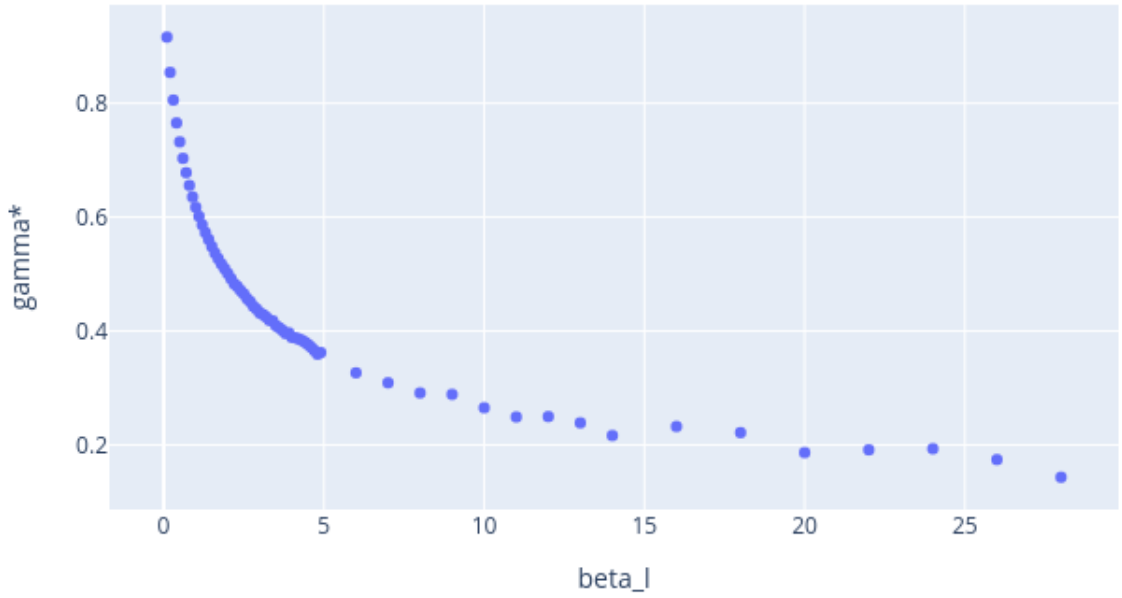

Figure 3.1: *The opportunity cost parameter γ_1^* as a function of β_l.*

Theorem 3.4. *Let β_l be fixed, and let γ_1^* be as in the previous theorem. The RL/evolutionary equivalence result from the previous theorem holds for for $s < 1$, and the corresponding discounting parameter is given by $\gamma_s^* = \gamma_1^* \cdot s$.*

Proof. Let r_t be the (stochastic) reward received at time t, and let $\hat{r}_t$ be the reward conditional on still being alive. Then $\mathbb{E}[r_t] = s^t \cdot \mathbb{E}[\hat{r}_t]$. By Proposition 3.1 and Theorem 3.3,

$$\sum_{t=0}^{\infty} \gamma_1^t \mathbb{E}[\hat{r}_t] = \sum_{t=0}^{\infty} (s\gamma_s)^t \mathbb{E}[\hat{r}_t],$$

from which the result follows. □

Remark. Theorem 3.4 decomposes the RL discounting parameter γ^* into two componenents: risk and opportunity cost. Risk accounts for the possibility that a future reward might not be realized due to death of the recipient, whereas opportunity cost is the lost opportunity for compounding benefits of an immediate reward. Risk is accounted for in s, while opportunity cost is in γ_1^*.

4 Conclusion

In this paper, we put forward a foraging model in which squirrels are routinely faced with an exploration/exploitation tradeoff. Squirrels are made to choose between a small but assured payout, β_l, and searching an unknown length of time for a larger reward, β_h. The model is analysed both from the perspective of maximizing evolutionary fitness (measured by the Lyapunov exponent) and as an RL problem. From both perspectives, we find that optimal search policies are deterministic: squirrels should either always search or always stay put. Perhaps more surprisingly, we find that having fixed a survival probability $s \in (0, 1]$ and low arm payout β_l , there exists a unique discounting parameter $\gamma_s^* \in (0, 1)$ for which the RL and evolutionary approaches lead to the same optimal search policy. We show that the discounting parameter γ_s^* can be decomposed into an opportunity cost component as well as a risk component: we can write $\gamma_s^* = \gamma_1^* \cdot s$, where γ_1 is the unique γ^* for fixed β_l when $s = 1$.

Our results provide a simple and intuitive rationale for exponential discounting in the context of foraging and exploitation/exploration tradeoffs. Further research is needed in order to determine the exact origin of the opportunity cost parameter γ_1^*. We posit that our present model is too hands on, in that its specificities obfuscate the likely simple and intuitive ab-

stract explanation for our findings. A future area of reserch, therefore, is to propose more abstract models which produce the same results. These models will allow us to prove key results which are known to be true in this paper, but for which we are unable to write a rigorous mathematical proof. For instance, it would be nice to prove that optimal search policies are deterministic, and that the right choice of γ^* leads to equivalence between RL and evolutionary optimizing. Nevertheless, this project is a reasonable start to understand under exactly what circumstances individual exponential discounting yields optimal evolutionary outcomes.

References

[FGB+19] William Fedus, Carles Gelada, Yoshua Bengio, Marc G. Bellemare, and Hugo Larochelle. Hyperbolic discounting and learning over multiple horizons. `https://arxiv.org/pdf/1902.06865.pdf`, 2019.

[FLO02] Shane Frederick, George Loewenstein, and Ted O'Donoghue. Time discounting and time preference: A critical review. *Journal of Economic Literature*, 40:351–401, June 2002.

[SK13] Janet Steven and James Kirkwood. *Mathematical Concepts and Methods in Modern Biology*, chapter Predicting Population Growth: Modeling with Projection Matrices. Elsevier Inc., 2013.

[Soz98] Peter Sozou. On hyperbolic discounting and uncertain hazard rates. *Proceedings of the Royal Society of London*, 265:2015–2020, 1998.

A Bandit results

Here are some bandit results.

A.1 Optimal policies are corner solutions

Our aim is to select policy π_ε among $\varepsilon \in [0,1]$ for which the leading eigenvalue of $\mathcal{B}_\varepsilon$ is maximal. Denote the optimal policy search probability by ε^*. Existence of such a value is clear, since the mapping $\mathcal{B}_\varepsilon \mapsto$(its leading eigenvalue) is a continuous function on the compact domain $[0,1]$. We claim that $\varepsilon^* \in \{0,1\}$. While a rigorous proof is beyond the scope of our (my) capabilities, the result holds experimentally across all values for l, h, β_l, and β_h tested. Furthermore, there is an intuitive argument to be made from the perspective of biology: all days are the same, and searches are independent. If a squirrel has any inclination to search today, then it should just as well search tomorrow. On the other hand, if it chooses not to search today, then it should not search tomorrow. The reader may not be entirely satisfied with such an argument, and I empathize with them – nevertheless, we shall take it as given throughout the remainder of this paper that $\varepsilon^* \in \{0,1\}$.

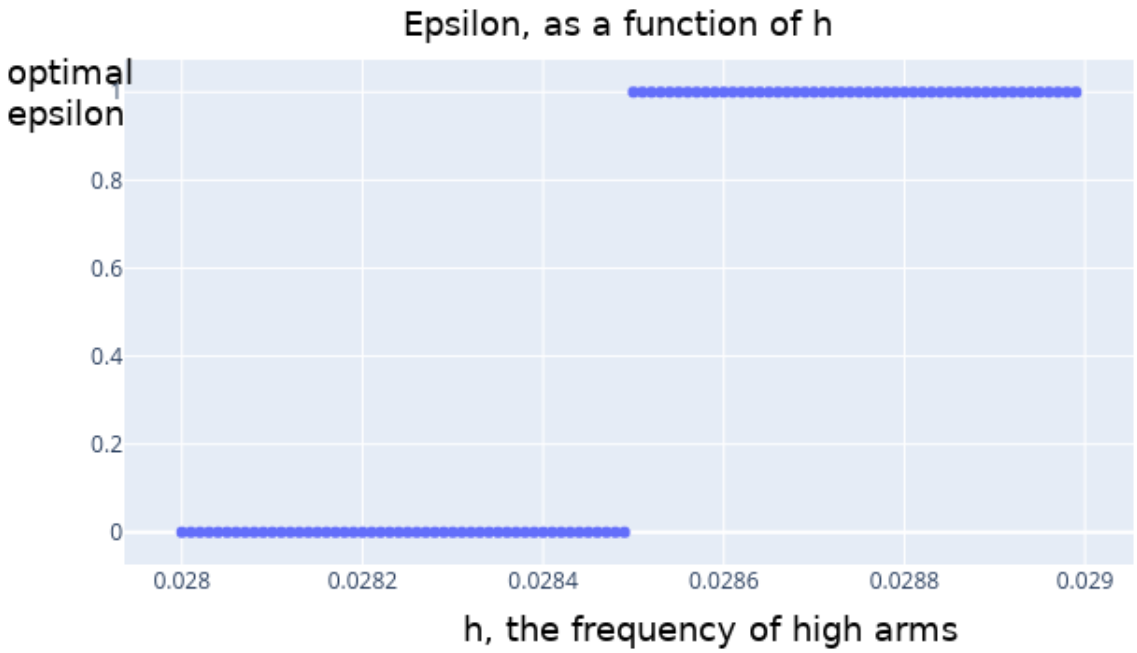

Figure A.1: $\varepsilon = \varepsilon(h)$, *with* $l = 0.3$, $\beta_l = 1$, *and* $\beta_h = 10$

A.2 Understanding $\varepsilon^* = \varepsilon^*(l, h, \beta_l, \beta_h)$

The optimal policy ε is determined by a large number of parameters. As such, it can be difficult to wrangle and properly understand. That $\varepsilon* \in \{0, 1\}$ helps us significantly in this task. The figure above illustrates an important point. To understand the plot of $\varepsilon^*(h)$ (with other terms fixed), it is sufficient to understand where it is discontinuous. Let h_{disc} the value of h such that $\varepsilon^*(h)$ is discontinuous at h_{disc}. We can then allow another parameter to vary, and determine its corresponding h_{disc}. For instance, if we allow β_h to float, we can consider $h_{\text{disc}} = h_{\text{disc}}(\beta_h)$.

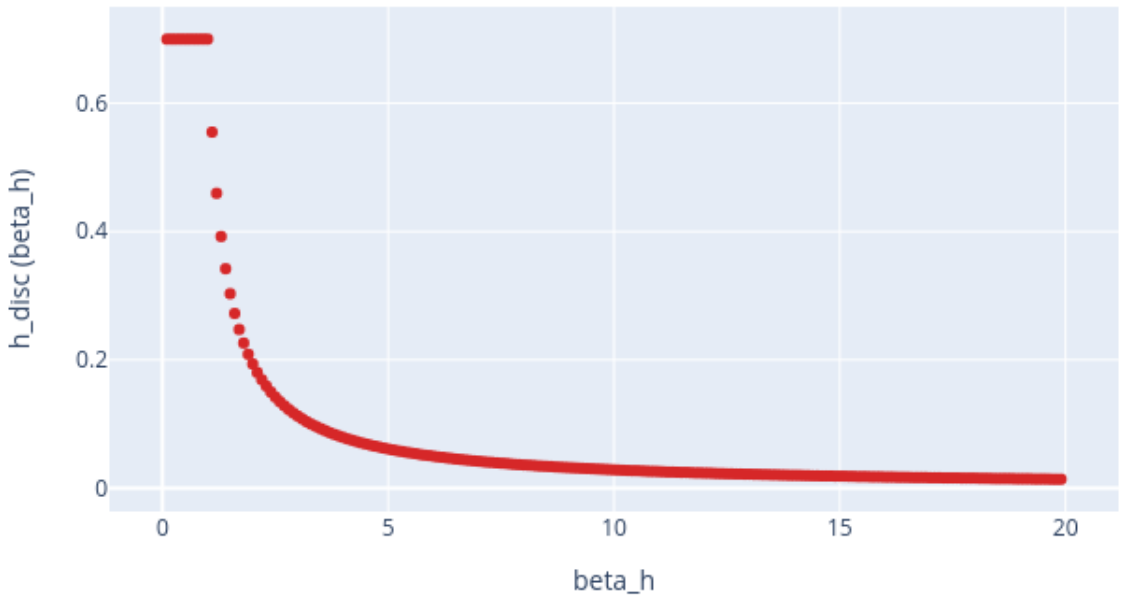

Figure A.2: $h_{\text{disc}} = h_{\text{disc}}(\beta_h)$, *with* $l = 0.3$ *and* $\beta_l = 1$

The above figure has a lot of information packed into it, so some unpacking might be helpful. For $0 \leq \beta_h \leq 1$, we interpret that $\varepsilon^*(h)$ is discontinuous at $h = 0.7$. Given our fixed value of $l = 0.3$, this means that $\varepsilon^* = 0$ no matter what h is. Is this reasonable? Well, if $\beta_h \leq 1 = \beta_l$, then our "high" payment is in fact lower than the "low" payment, so that it is indeed desireable that $\varepsilon^* = 0$. A relatively higher paying arm should never be left in search of a lower paying arm. As β_h increases past 1, the discontinuity point drops. Initially, when β_h is only marginally larger than β_l, the discontinuity point occurs when h is large, not far off from its maximal value of 70. This means that only marginally larger payments necessitate a high degree of confidence that a squirrel will find it quickly in order to make searching worthwhile. Later, for instance as $\beta_h = 20$, the relative abundance of high arms dips down

below 2%, which has a natural interpretation. Squirrels exhibit more patience for larger rewards.

B RL results

Here are some results from the RL optimization process.

B.1 Computational work

In an ideal world, this would be a section on probability and I would present a neat derivation for clean, closed form expression for $\mathbb{E}_\varepsilon[\hat{r}_t]$. Alas, it was not to be. All results used for computation come from the following clever trick: remove all birth and death terms from $\mathcal{B}_\varepsilon$. Then what remains is a migration matrix. Since $v_0 := \mathcal{B}_\varepsilon \cdot (1,0,0)^\intercal$ is a three dimensional vector with each component representing the probability of being at a null, low, or high arm respectively, we have that $\mathbb{E}_\varepsilon[\hat{r}_0] = v_0^2\beta_l + v_0^3\beta_h$. More generally, with $v_t = \mathcal{B}_\varepsilon^{t+1} \cdot (1,0,0)^\intercal$, we have $\mathbb{E}_\varepsilon[\hat{r}_t] = v_t^2\beta_l + v_t^3\beta_h$. Superscripts denote components, that is, $v_t = (v_t^1, v_t^2, v_t^3)$. The point here is that we efficiently compute the expected payout on a given day conditional on still being alive, and this is precisely how we do it. To compute the RL objective function, we simply take the first 1,000 terms in the summand (or so, depending on the context).

B.2 Corner solutions again

When trying to optimize the Lyapunov exponent we found ε^*. We employ the convention that ε^* corresponds to the optimal policy search probability in the evolutionary sense, whereas ε_{RL} is the search probability maximizing the RL objective function

$$\sum_{t=0}^{\infty} \gamma^t \mathbb{E}_\varepsilon[\hat{r}_t]\,.$$

Similarly to before, I claim that $\varepsilon_{\text{RL}} \in \{0, 1\}$, and as before, I am unable to prove this claim. The claim has been checked fairly exhaustively for different values of l, h, β_l, β_h, and γ. Nevertheless, a proof would be nice, but not in this paper.

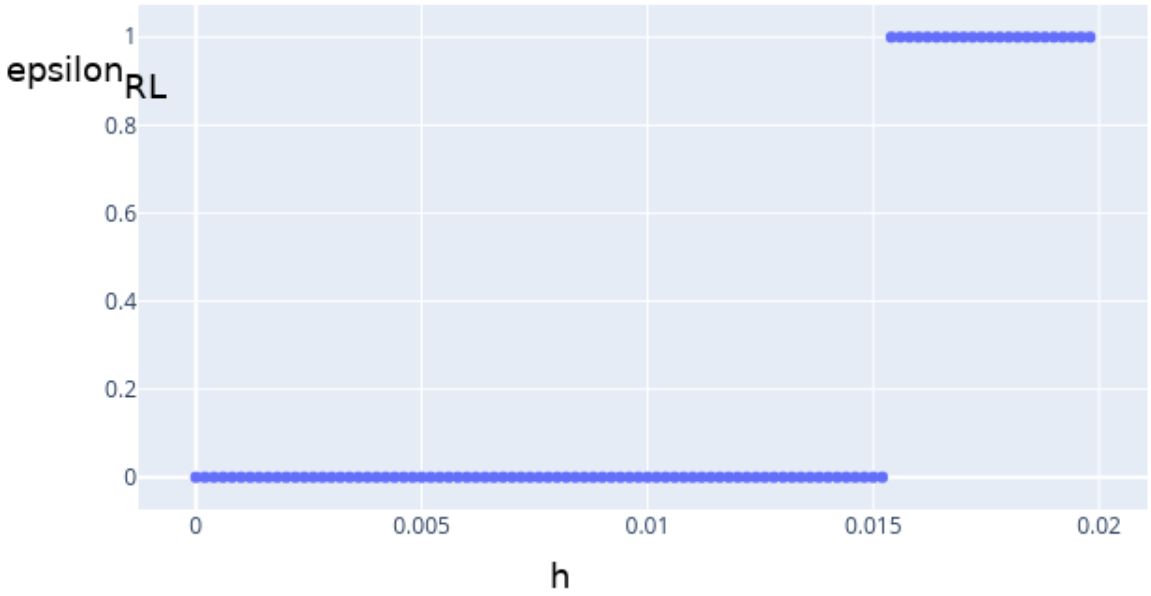

Figure B.1: *$\varepsilon_{RL} = \varepsilon_{RL}(h)$, with $l = 0.3$, $\beta_l = 1$, $\beta_h = 10$, and $\gamma = 0.8$.*

In this instance, a value of $\gamma = 0.8$ was arbitrarily chosen. The value of γ changes the discontinuity point; higher γ lead to more patience, that is, a higher value of γ will lead to lower value of h_{disc}.

B.3 Discontinuity point in $\varepsilon_{\mathbf{RL}}(h)$ as a function of β_h

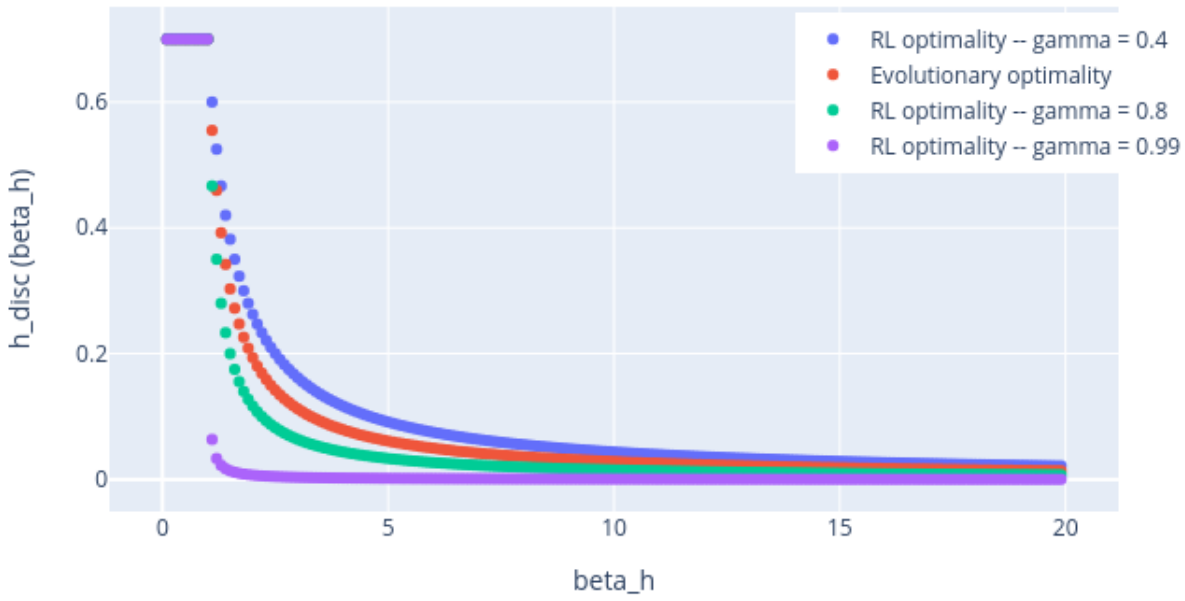

Figure B.2: *Discontinuity point in optimal ε for $l = 0.3$ and $\beta_l = 1$, various optimality conditions.*

We interpret the figure as meaning that more steeply descending curves indicate higher degrees of patience, meaning a higher tolerance for a low probability of pulling a high arm. Marginally increasing β_h beyond $1 = \beta_l$, a lower value of $h_{\text{disc}}(\beta_h)$ implies an earlier switch from never searching to always searching from a low arm. We see that for $\gamma = 0.99$ (weakly discounting in the RL objective function), the curve is most steeply downward sloping. This is to be expected, since highly valuing future rewards makes a squirrel more likely to search until finding a high arm.

Perhaps most interestingly displayed in the figure is that the function $h^{\gamma}_{\text{disc}}(\beta_h)$ may lie either below or above the evolutionary discontinuity curve, dependent on the value of γ. A natural question arises: for fixed l and β_l, does there exist a value, call it $\gamma_{\text{evolutionary}}$, such that $\varepsilon^*(l, h, \beta_l, \beta_h) = \varepsilon_{\text{RL}}(l, h, \beta_l, \beta_h, \gamma_{\text{evolutionary}})$ for all h and all β_h? Essentially, can we tune γ so as to have the RL curve be the same as the evolutionary curve? The answer, eyeballing it, is yes. Numerically, this result holds as well, and leads to our Theorem 3.3. We numerically compute the approximate value for γ^* and find that having fixed β_l and s, γ^* is uniquely determined.