# Asymptotically Optimal Agents

Tor Lattimore[1] Marcus Hutter[1,2]

October 7, 2011

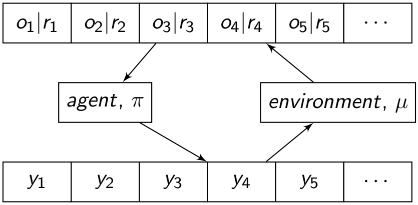
[1]Australian National University

[2]ETH Zürich

# The Questions

1. What is a reasonable definition of optimality for a general *learning* agent?
2. Do such optimal learning agents exist?

# Environments




Let $\mathcal{Y}$, $\mathcal{O}$, and $\mathcal{R} \subset \mathbb{R}^+$ be sets of actions, observations and rewards respectively. Let $\mathcal{X} = \mathcal{R} \times \mathcal{O}$. A deterministic environment $\mu$ is a function

$$\mu : (\mathcal{Y} \times \mathcal{X})^* \times \mathcal{Y} \to \mathcal{X}$$

# Environments

### Definition

A *policy* is a function $\pi : (\mathcal{Y} \times \mathcal{X})^* \to \mathcal{Y}$

### Definition

The value of policy $\pi$ after history $h_{<n}$ in environment $\mu$ is

$$V_\mu^\pi(h_{<n}) = \frac{1}{\Gamma_n} \sum_{k=n}^{\infty} \gamma_k r_k$$

where $r_k$ is the reward obtained at time $k$ when $\pi$ interacts with $\mu$ and

$$\sum_{t=1}^{\infty} \gamma_t < \infty \qquad\qquad \Gamma_n := \sum_{t=n}^{\infty} \gamma_t$$

### Definition

The optimal policy $\pi_\mu^*$ in environment $\mu$ is

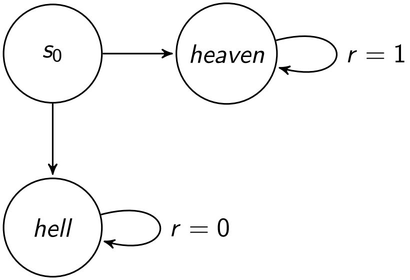
$$\pi_\mu^* := \arg\max_\pi V_\mu^\pi \qquad\qquad V_\mu^* := V_\mu^{\pi_\mu^*}$$

# Optimality

What does it mean to be optimal? In the planning problem, policy $\pi$ is optimal in environment $\mu$ if

$$V^*_\mu(h_{<t}) - V^\pi_\mu(h_{<t}) = 0, \qquad \forall h_{<t}$$

This is unreasonable for learning algorithms because they need time to explore.



# Optimality

## Definition (Strong/Weak Asymptotic Optimality)

Let $\mathcal{M}$ be a set of environments then $\pi$ is strong/weak asymptotically optimal in $\mathcal{M}$ if

$$\lim_{n\to\infty}\left[V^*_\mu(h_{<n}) - V^\pi_\mu(h_{<n})\right] = 0, \forall \mu \in \mathcal{M} \qquad \text{Strong}$$

$$\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\left[V^*_\mu(h_{<t}) - V^\pi_\mu(h_{<t})\right] = 0, \forall \mu \in \mathcal{M} \qquad \text{Weak}$$

1. A strong asymptotically optimal agent eventually makes no errors.
2. A weak asymptotically optimal agent makes a fraction of errors that decreases to zero in the limit.
3. The larger the class $\mathcal{M}$, the more powerful is policy $\pi$.

# Results

Let $\mathcal{M}$ be the class of all deterministic computable environments, which aside from the restriction to deterministic environments is an extremely large class.

Does there exist a single policy $\pi$ that is computable/incomputable and weak/strong asymptotically optimal in $\mathcal{M}$.

| | Computable | Incomputable |
|---|---|---|
| Weak | No | Depends on discounting |
| Strong | No | No |

# No Computable Asymptotically Optimal Agents

### Theorem

*If $\mathcal{M}$ is the class of all computable deterministic environments then no computable deterministic policy is weak/strong asymptotically optimal.*

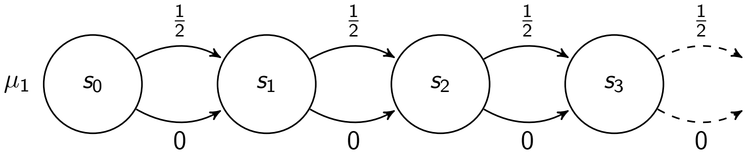
Computability of policy means the worst possible (very adversarial) environment is also computable. This implies any weak/strong asymptotically optimal policy is necessarily incomputable.

# No Strong Asymptotically Optimal Agents

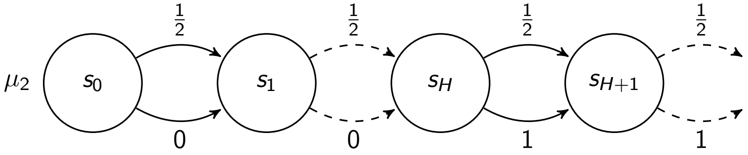
## Theorem

*There does not exist a strong asymptotically optimal agent for $\mathcal{M}$.*

## Proof sketch (geometric discounting).




To be strong asymptotically optimal in $\mu_1$, $\pi$ must eventually only go up. Let $H$ be the time-step at which it stops exploring.



# Weak Asymptotically Optimal Agents

**Theorem**

*Let $\mathcal{M} = \{\mu_1, \mu_2, \cdots\}$ be a countable class of deterministic environments. There exists a weak asymptotically optimal policy $\pi$ in $\mathcal{M}$ if discounting is geometric.*

**Proof sketch.**

- $\pi$ must explore infinitely often and arbitrarily deep.
- $\pi$ cannot explore too much.
- $\pi$ must be unpredictable.

$\pi$ is defined as follows.

- At time $t$, $\pi$ uses for its model $\mu_{t_i}$ which is the first environment consistent with the history seen so far.
- With probability $\frac{t-1}{t}$, $\pi(h_{<t}) = \pi^*_{\mu_{t_i}}(h_{<t})$.
- With probability $\frac{1}{t}$, $\pi$ enters an exploration phase, exploring randomly for $\log t$ time-steps.

# Large Horizons

**Theorem**

*Let $\mathcal{M}$ be the class of all computable deterministic environments. There does not exist a weak asymptotically optimal policy if discounting is* $\gamma_t := \frac{1}{t(t-1)}$.

1. This discount function has a growing effective horizon.
2. To find the true model an agent must explore for too long, which wrecks weak asymptotic optimality.

# Summary

- Strong asymptotic optimality is too strong.
- Existence of weak asymptotically optimal agents depends on discounting.
- Weak asymptotically optimal agents in the class of all deterministic computable environments must be stochastic (or incomputable). The one defined here is both stochastic *and* incomputable.
- Smaller, but still interesting, classes of environments admit computable weak asymptotically optimal agents.
- There should exist weak asymptotically optimal agents in the class of computable stochastic environments.
- Theorems apply only to computable discount functions.
- We also care about non asymptotic behaviour.