Proceedings of Machine Learning Research vol 125:1–30, 2020

Pessimism About Unknown Unknowns Inspires Conservatism

Michael K. Cohen MICHAEL-K-COHEN.COM
University of Oxford
Research School of Engineering Science
Future of Humanity Institute

Marcus Hutter HUTTER1.NET
Google DeepMind
Australian National University

Abstract

If we could define the set of all bad outcomes, we could hard-code an agent which avoids them; however, in sufficiently complex environments, this is infeasible. We do not know of any general-purpose approaches in the literature to avoiding novel failure modes. Motivated by this, we define an idealized Bayesian reinforcement learner which follows a policy that maximizes the worst-case expected reward over a set of world-models. We call this agent pessimistic, since it optimizes assuming the worst case. A scalar parameter tunes the agent's pessimism by changing the size of the set of world-models taken into account. Our first main contribution is: given an assumption about the agent's model class, a sufficiently pessimistic agent does not cause "unprecedented events" with probability $1-\delta$, whether or not designers know how to precisely specify those precedents they are concerned with. Since pessimism discourages exploration, at each timestep, the agent may defer to a mentor, who may be a human or some known-safe policy we would like to improve. Our other main contribution is that the agent's policy's value approaches at least that of the mentor, while the probability of deferring to the mentor goes to 0. In high-stakes environments, we might like advanced artificial agents to pursue goals cautiously, which is a non-trivial problem even if the agent were allowed arbitrary computing power; we present a formal solution.

1. Introduction

Intuitively, there are contexts in which we would like advanced agents to be conservative: novel action-sequences should be treated with caution, and only taken when the agent is quite sure its world-model generalizes well to this untested new idea. For a weak agent in a simple environment, the following approach may suffice: model the environment as finite-state Markov, observe a mentor, and only take actions that you have already observed the mentor take from the current state. But in a complex environment, one never or hardly ever sees the exact same state twice; even worse, if the environment is non-stationary, a previous observation of the mentor taking action a from state s does not imply it is still safe to do so.

We construct an idealized Bayesian reinforcement learner. We do not assume our agent's environment is finite-state Markov or ergodic. We will only assume that our agent's environment, which may depend on the entire interaction history, belongs to a countable set $\mathcal{M}$. For example, the countable set of semicomputable stochastic world-models would be large enough to make this assumption innocuous (Hutter, 2005). The limit of this idealization is that because we make so few assumptions, we can't ensure that computing the posterior is tractable in the general setting.

Our agent also has a mentor, who can select an action when the agent requests, and we assume nothing about the agent's mentor besides belonging to a countable set of possible policies $\mathcal{P}$. The mentor could be a human or a known-safe policy.

Our agent starts with a prior that assigns non-zero probability to a countable set of world-models $\mathcal{M}$ and mentor-models $\mathcal{P}$, and recursively updates a posterior. At each timestep, it stochastically defers to a mentor with some probability, and the mentor selects the action on its behalf; otherwise, it takes the top world-models in its posterior until they cover some fixed fraction β of the posterior, and it follows a policy which maximizes the minimum expected return among those top world-models. We call this minimum the pessimistic value because it is a worst-case estimate. At each timestep, to decide whether to defer action-selection to the mentor, the agent samples a world-model and mentor-model from its posterior; the agent calculates the value of acting according to that mentor-model in that world-model given the current interaction history, and if that value is greater than the pessimistic value plus positive noise, or if the pessimistic value is 0, the agent defers. This query probability is inspired by the effectiveness of Thompson Sampling (Thompson, 1933).

We show

- In the limit, the pessimistic agent's policy's value approaches at least that of the mentor's. (Corollary 6)
- The mentor is queried with probability approaching 0 as $t \to \infty$. (Corollary 7)
- For any complexity class C, we can set $\mathcal{M}$ so that for any event E in the class C, we can set β so that with arbitrarily high probability: for the whole lifetime of the agent, if the event E has never happened before, the agent will not make it happen. Either the mentor will take an action on the agent's behalf which makes E happen for the first time, or E will never happen. (Theorem 11)

We call the last point the Probably Respecting Precedent Theorem. The "precedent" is that a certain event has never happened, and the agent probably never takes an action which disrupts that precedent for the first time. For any failure mode that designers do not know how to specify formally, the agent can be made to probably not fail that way. The price of this is intractability, but tractable approximations of pessimism may preserve these results in practice, or perhaps even in theory.

Section 2 introduces notation, Section 3 reviews related work, we define the agent's policy in Section 4, and we prove performance results and safety results in Sections 5 and 6. Appendix A collects definitions and notation, Appendix B presents an algorithm for an ε-approximation of the agent's policy, Appendix C contains omitted proofs, and Appendix D contains an informal discussion.

2. Notation

Let $\mathcal{A}$, $\mathcal{O}$, and $\mathcal{R}$ be finite sets of possible actions, observations, and rewards. Let $\{0,1\} \subset \mathcal{R} \subset [0,1]$. Let $\mathcal{H} = \mathcal{A} \times \mathcal{O} \times \mathcal{R}$. For each timestep $t \in \mathbb{N}$, a_t, o_t, and r_t denote the action, observation, and reward, and h_t denotes the triple. A policy π can depend on the entire history so far. We denote this history $(h_1, h_2, ..., h_{t-1})$ as $h_{<t}$. Policies may be stochastic, outputting a distribution over actions. Thus, $\pi : \mathcal{H}^* \rightsquigarrow \mathcal{A}$, where $\mathcal{H}^* = \bigcup_{i=0}^{\infty} \mathcal{H}^i$, and $\rightsquigarrow$ means the function may be stochastic. Likewise, in general, a world-model $\nu : \mathcal{H}^* \times \mathcal{A} \rightsquigarrow \mathcal{O} \times \mathcal{R}$ may be stochastic, and it may depend on the entire interaction history. The latter possibility allows (the agent to

conceive of) environments which are not finite-state Markov. A policy π and a world-model ν induce a probability measure P^{π}_{ν} over infinite interaction histories. This is the probability of events when actions are sampled from π and observations and rewards are sampled from ν. Formally, $\mathrm{P}^{\pi}_{\nu}(h_{\leq t}) = \prod_{k=1}^{t} \pi(a_k|h_{<k})\nu(o_k r_k|h_{<k}a_k)$. We use general, history-based world-models, with no assumptions on $\nu \in \mathcal{M}$, even though they present complications that finite-state Markov, ergodic world-models do not.

The agent will maintain a belief distribution over a class of world-models $\mathcal{M}$. We allow this to be an arbitrary countable set. A prime example, the set of semicomputable stochastic world-models $\mathcal{M}_{\text{COMP}}$ (Hutter, 2005), is only countable, but large enough. The agent starts with a prior belief $w(\nu)$ that the world-model $\nu \in \mathcal{M}$ is the true environment (w is for "weight"). Naturally, $\sum_{\nu\in\mathcal{M}} w(\nu) = 1$. The agent updates its belief distribution according to Bayes' rule, which we write as follows: $w(\nu|h_{<t}) :\propto w(\nu)\prod_{k=1}^{t-1}\nu(o_k r_k|h_{<k}a_k)$, normalized so that $\sum_{\nu\in\mathcal{M}} w(\nu|h_{<t}) = 1$. Let μ be the true environment. We assume $\mu \in \mathcal{M}$, and we assume the true observed rewards are at least $\varepsilon_r > 0$.

For an agent with a discount factor $\gamma \in [0, 1)$, and a policy π, given a world-model ν, and an interaction history $h_{<t}$, the *value* of that policy from that position in that world is

$$V^{\pi}_{\nu}(h_{<t}) := (1-\gamma)\,\mathbb{E}^{\pi}_{\nu}\left[\sum_{k=t}^{\infty}\gamma^{k-t}r_k\middle|h_{<t}\right] \tag{1}$$

where $\mathbb{E}^{\pi}_{\nu}$ is the expectation under the probability measure P^{π}_{ν}. The factor of $1-\gamma$ normalizes the value to $[0, 1]$ for convenience.

3. Related Work

Virtually all previous work that attempts to make reinforcement learners avoid unspecified failure modes assumes a finite-state Markov environment. We do not, but the literature is nonetheless informative for our general setting.

Heger (1994) defines $\hat{Q}$-learning, which maximizes the worst-case return for a known MDP, and Jiang et al. (1998) extend the case to unknown MDPs. As García and Fernández (2015) describe, Gaskett (2003) found empirically that such extreme pessimism is more harmful than helpful. Gaskett (2003) introduces a variant on the Q-value, which is the value of an action under the assumption that at each future timestep, with some probability, the *worst* action will be taken, instead of the best one; they test this empirically.

Closer to our approach, Iyengar (2005) and Nilim and El Ghaoui (2005) construct a policy which is robust to errors in the transition probabilities by considering the worst-case return within some error tolerance. Much of the work on the topic takes the form of presenting a tractable approach to the execution of this robust policy, e.g. Tamar et al. (2013). Unfortunately, this research assumes access to an MDP with (approximately) known transition probabilities—at first glance this seems like something an agent might reasonably have access to after limited observations, but the MDPs are assumed to be *uniformly* approximately known, which requires exploration, and indeed requires observing every "failure" state that the robust policies are supposed to avoid. The finite-state Markov assumption their work makes is useful for many circumstances, but advanced agents may have to conceive of non-stationarity in the environment, and importantly for our purposes, *novel* failure modes.

Other work makes use of a mentor to avoid "dangerous" states (whereas in our work, the mentor lower-bounds the capability of the agent, and robustness derives from pessimism). Imitation learning (Abbeel and Ng, 2004; Ho and Ermon, 2016; Ross et al., 2011) makes the most of a mentor in the absence of other feedback, like rewards. An abundance of "ask for help" algorithms query a mentor under conditions which correspond to some form of uncertainty (Clouse, 1997; Hans et al., 2008; García and Fernández, 2012; García et al., 2013). Kosoy (2019) gives a regret bound for an agent in a (non-ergodic) MDP, given access to an expert mentor and a finite set of models that contains the truth. García and Fernández (2015, Section 4.1.3.2) review many protocols by which a mentor monitors the state and intervenes at will through various channels, and Saunders et al. (2018) is another more recent example. One risk of relying on mentor-intervention to protect against critical failure is that a mentor may not recognize action sequences which lead to critical failure, even if we would trust a mentor not to wander into those failure modes by virtue of their complexity.

Sunehag and Hutter's (2015) optimistic agent directly inspired this work; optimism is designed to be an exploration strategy. Hutter's (2005) formulation of universal artificial intelligence is the basic theoretical framework we use here to analyze idealized artificial agents. Technically, our work borrows most from Hutter's (2009), Leike et al.'s (2016), and Cohen et al.'s (2020) work on Bayesian agents with general countable model-classes.

4. Agent Definition

We now define the pessimistic policy and the probability with which the agent defers to a mentor. We define the agent's policy mathematically here, and we write an algorithm in Appendix B.

4.1. Pessimism

$\beta \in (0,1)$ will tune the agent's pessimism. If, for example, $\beta = 0.95$, we say that the agent is 95% pessimistic. Such an agent will restrict attention to a set of world-models that covers 95% of its belief distribution, and act to maximize expected reward in the worst-case scenario among those world-models. Formally, let ν^k be the world-model in $\mathcal{M}$ with the k^{th} largest posterior weight, and let $\mathcal{T}_k$ be the top-k most probable world-models, defined as follows:

$$\mathcal{T}_0(h_{<t}) := \emptyset \qquad (2)$$

$$\nu^k(h_{<t}) := \operatorname*{argmax}_{\nu \in \mathcal{M} \setminus \mathcal{T}_{k-1}(h_{<t})} w(\nu|h_{<t}) \qquad (3)$$

$$\mathcal{T}_k(h_{<t}) := \mathcal{T}_{k-1}(h_{<t}) \cup \{\nu^k(h_{<t})\} \qquad (4)$$

Ties in the argmax are broken arbitrarily (as everywhere else in the paper). Then,

$$k_t^\beta := \min \left\{ k \in \mathbb{N} \;\middle|\; \sum_{\nu \in \mathcal{T}_k(h_{<t})} w(\nu|h_{<t}) > \beta \right\} \qquad (5)$$

$$\mathcal{M}_t^\beta := \mathcal{T}_{k_t^\beta}(h_{<t}) \qquad (6)$$

Note that k_t^β and $\mathcal{M}_t^\beta$ both depend on $h_{<t}$, not just t, and note that $\mathcal{M}_t^\beta$ satisfies

$$\sum_{\nu \in \mathcal{M}_t^\beta} w(\nu|h_{<t}) > \beta \qquad (7)$$

The β-pessimistic policy is defined as follows:

$$\pi_t^\beta := \operatorname*{argmax}_{\pi \in \Pi} \min_{\nu \in \mathcal{M}_t^\beta} V_\nu^\pi(h_{<t}) \tag{8}$$

$$\pi^\beta(\cdot|h_{<t}) := \pi_t^\beta(\cdot|h_{<t}) \tag{9}$$

Π is the set of all deterministic policies, and some deterministic policy will always be optimal (Lattimore and Hutter, 2014). The connection to the minimax approach in game theory is interesting: from Equation 8, it looks as though the pessimistic agent believes there is an adversary in the environment. Our policy is inspired by Sunehag and Hutter's (2015) optimistic agent, in which the $\min$ is replaced with a $\max$, and $\mathcal{M}_t^\beta$ is replaced with an arbitrary finite subset of the model class. Whereas the purpose of optimism is to encourage exploration, the purpose of pessimism is to discourage novelty.

4.2. The Mentor

Since pessimism discourages exploration, we introduce a mentor to demonstrate a policy. We suppose that at any timestep, the agent may defer to a mentor, who will then select the action on the agent's behalf. Thus, the agent can choose to follow the mentor's policy π^m, not by computing it, but rather by querying the mentor. π^m may be stochastic. What remains to be defined is *when* the agent queries the mentor.

The agent maintains a posterior distribution over a set of mentor-models. Each mentor-model is a policy $\pi \in \mathcal{P}$, an arbitrary countable set, and let $w'(\pi)$ be the prior probability that the agent assigns to the proposition that the mentor samples actions from π. Letting $q_k = 1$ if the agent queried the mentor at timestep k, and letting $q_k = 0$ otherwise, the posterior belief $w'(\pi|h_{<t}) :\propto w'(\pi) \prod_{k<t:q_k=1} \pi(a_k|h_{<k})$.

At timestep t, the agent follows the following procedure to determine whether to query the mentor. $\hat{\pi}_t \sim w'(\cdot|h_{<t})$. $\hat{\nu}_t \sim w(\cdot|h_{<t})$. Sampling from a posterior is often called Thompson Sampling (Thompson, 1933). $X_t := V_{\hat{\nu}_t}^{\hat{\pi}_t}(h_{<t})$. $Y_t := \max_{\pi \in \Pi} \min_{\nu \in \mathcal{M}_t^\beta} V_\nu^\pi(h_{<t})$. Let $Z_t > 0$ be an i.i.d. random variable such that for all $\varepsilon > 0$, $p(Z_t < \varepsilon) > 0$, e.g. $Z_t \sim \text{Uniform}((0, 2])$. If $X_t > Y_t + Z_t$, or if $Y_t = 0$, the agent defers to the mentor. For ease of analysis, we also require $p(Z_t > 1) > 0$. The greater the possibility that the mentor can accrue much more reward, the higher the probability of deferring.

When $Y_t = 0$, we call this the "zero condition." Our earlier assumption that the true observed rewards be at least $\varepsilon_r > 0$ is to ensure the zero condition only happens finitely often. The agent will still consider it possible to get zero reward, but it will never actually observe such a thing. Let θ_t denote the probability that $q_t = 1$ and the agent defers to the mentor; note that θ_t depends on the whole history, not just t.

The pessimistic agent's policy, which mixes between π^β (from Eqn. 9) and π^m according to its query probability, is denoted π_Z^β; that is, $\pi_Z^\beta(\cdot|h_{<t}) := \theta_t \pi^m(\cdot|h_{<t}) + (1 - \theta_t)\pi^\beta(\cdot|h_{<t})$.

5. Performance Results

We now present our first contribution: we show that value of the agent's policy will at least approach, and perhaps exceed, the value of the mentor's policy. We also show that the probability of querying the mentor approaches 0. In the next section, we will prove results regarding the safety of the agent.

We begin with a lemma regarding Bayesian sequence prediction: the β-maximum a posteriori models—that is, the minimal set of models that amount to at least β of the posterior—all "merge" with the true world-model. We require some new notation to define this formally.

Let $x_{<\infty} \in \mathcal{X}^\infty$; that is, it is an infinite string from a finite alphabet $\mathcal{X}$. Let $x_{<t}$ be the first $t-1$ characters of $x_{<\infty}$. We consider probability measures over the outcome space $\Omega = \mathcal{X}^\infty$, with the standard event space being the σ-algebra of cylinder sets: $\mathcal{F} = \sigma(\{\{x_{<t}y|y \in \mathcal{X}^\infty\}|x_{<t} \in \mathcal{X}^*\})$. We abbreviate $x_{<\infty}$ as ω. We will consider a countable class of probability measures over this space $\mathcal{M} = \{Q_i\}_{i\in\mathbb{N}}$. One such probability measure will be denoted P (the true sampling one), and Q will denote an arbitrary probability measure over $\mathcal{X}^\infty$.

We will write $P(x_{<t})$ to mean the probability that the infinite string ω begins with $x_{<t}$; so technically, it is shorthand for $P(\{x_{<t}y|y \in \mathcal{X}^\infty\})$. By $P(x'|x_{<t})$ (for $x' \in \mathcal{X}^*$), we mean $P(x_{<t}x')/P(x_{<t})$, that is, the probability that x' follows $x_{<t}$. We begin with prior weights over $Q \in \mathcal{M}$, denoted $w(Q) > 0$, and satisfying $\sum_{Q\in\mathcal{M}} w(Q) = 1$, and we let the posterior weight be

$$w(Q|x_{<t}) := \frac{w(Q)Q(x_{<t})}{\sum_{Q'\in\mathcal{M}} w(Q')Q'(x_{<t})} \tag{10}$$

For $\mathcal{M}' \subset \mathcal{M}$, we also define $w(\mathcal{M}'|\cdot) = \sum_{Q\in\mathcal{M}'} w(Q|\cdot)$.

The k-step variation distance between P and Q is how much they can possibly differ on the probability of what the next k characters might be (Hutter, 2005).

Definition 1 (k-step variation distance)

$$d_k(P, Q|x_{<t}) = \max_{\mathcal{E}\subset\mathcal{X}^k} \left|P(\mathcal{E}|x_{<t}) - Q(\mathcal{E}|x_{<t})\right|$$

Definition 2 (Total variation distance)

$$d(P, Q|x_{<t}) = \lim_{k\to\infty} d_k(P, Q|x_{<t})$$

which exists because $d_k(P, Q|x_{<t})$ is non-decreasing and bounded by 1.

Inspired by Blackwell and Dubins (1962), the following lemma may interest some Bayesians more than any of our theorems. Defining $\mathcal{M}_t^\beta$ exactly as before (see Equations 2 - 6), but for $Q \in \mathcal{M}$ instead of for $\nu \in \mathcal{M}$, and conditioning on $x_{<t}$ instead of $h_{<t}$,

Lemma 3 (Merging of Top Opinions) *For $\beta \in (0, 1)$, $\lim_{t\to\infty} \max_{Q\in\mathcal{M}_t^\beta} d(P, Q|x_{<t}) = 0$ with P-probability 1 (i.e. when $x_{<\infty} = \omega \sim P$).*

Unless otherwise specified, all limits in this paper are as $t \to \infty$. This lemma is proven in Appendix C, and it requires a few lemmas that are stated and proven there as well. Among these, Lemma 20 is a beautiful one that we feel should be known, but we couldn't find it in the literature. It says the sum of the limits of posterior weights is 1, a.s.: $\sum_{Q\in\mathcal{M}} \lim w(Q|x_{<t}) = 1$ with P-prob.1, for $P \in \mathcal{M}$. The others are short results from recent papers; we restate them there and re-prove them when feasible to save the reader the trouble of translating notation and verifying that those results apply to our current problem. Roughly, Lemma 3 holds because when a true model has positive prior weight, all models either merge with the truth or have their posterior weight go to 0, so eventually, all top models must merge; but the set of top models changes with each observation, and limits require care, so it ends up being somewhat involved.

We now return to the probability space where infinite sequences are over the alphabet $\mathcal{H}$, and probability measures P^{π}_{ν} denote the probability when actions are sampled from a policy π and observations and rewards are sampled from a world-model ν. Since π^{β}_{Z} is the agent's policy, and μ is the true environment, we will often abbreviate "with $\mathrm{P}^{\pi^{\beta}_{Z}}_{\mu}$-probability 1" as just "with probability 1" or "w.p.1". We assume, for the remaining results: $\mathcal{M} \ni \mu$, and $\mathcal{P} \ni \pi^{m}$.

Further lemmas which depend on the Merging of Top Opinions Lemma are stated in Appendix C. They are: with probability 1, on-policy prediction converges, the zero condition occurs only finitely often, and "almost-on-policy prediction" converges, which is roughly that if the agent's policy mimics another policy π_t with some uniformly positive probability some of the time, then on those timesteps, on-π_t-policy prediction converges to the truth. Formally,

Lemma 4 (Almost On-Policy Convergence) *For a sequence of policies π_t and an infinite set of timesteps τ, the following holds with $\mathrm{P}^{\pi^{\beta}_{Z}}_{\mu}$-prob. 1: if there exists $c > 0$ such that $\forall t \in \tau\ \forall t' \geq t\ \forall a \in \mathcal{A}\ \pi^{\beta}_{Z}(a|h_{<t'}) \geq c\pi_t(a|h_{<t'})$, then $\lim_{\tau \ni t \to \infty} V^{\pi_t}_{\mu}(h_{<t}) - \min_{\nu \in \mathcal{M}^{\beta}_{t}} V^{\pi_t}_{\nu}(h_{<t}) = 0$ and for all k, $\lim_{\tau \ni t \to \infty} \max_{\nu \in \mathcal{M}^{\beta}_{t}} d_k\left(\mathrm{P}^{\pi_t}_{\nu}, \mathrm{P}^{\pi_t}_{\mu} \middle| h_{<t}\right) = 0$.*

The proof is in Appendix C; if it didn't hold, on-policy prediction error would be bounded below at those timesteps τ. Our main performance results are corollaries of the following theorem.

Theorem 5 (Exploiting Surpasses Exploring)

$$\liminf w(\nu|h_{<t})w'(\pi|h_{<t}) > 0 \implies \liminf V^{\pi^{\beta}}_{\mu}(h_{<t}) - V^{\pi}_{\nu}(h_{<t}) \geq 0 \ \ \textit{w.p.1}$$

Informally, for any world-model/mentor-model pair that remains possible, the true value of the pessimistic policy will be at least as high. A note on the proof: we will consider an infinite interaction history which violates the theorem, follow implications that hold with probability 1, and arrive at a contradiction. Strictly speaking, we are considering the set of infinite interaction histories which violate the theorem *and* for which all the implications we employ are true. The resulting set of infinite interaction histories will be $\emptyset$ once we arrive at a contradiction, so it will have probability 0. Since all implications used in the proof have probability 1 (and we only employ countably many such implications), the negation of the theorem must also have probability 0 by countable additivity. Since it is tedious to keep track of sets of outcomes for which each line in the proof holds, we simply treat implications that hold with probability 1 as if they were true logical implications, but as we have just argued, as long as this is not done uncountably many times, this is a valid style of proof.

Most of the proof is a lengthy proof by induction; we set up the proof by induction and outline the remainder, which is completed in Appendix C.

Proof – Detailed Outline Fix an infinite interaction history $h_{<\infty}$. Suppose $\liminf w(\nu'|h_{<t}) \cdot w'(\pi'|h_{<t}) > 0$. This implies $\inf_t w(\nu'|h_{<t})w'(\pi'|h_{<t}) > 0$, because if a posterior is ever 0, it will always be 0. Let $\nu'_{\inf} > 0$ and $\pi'_{\inf} > 0$ denote those two infima. Let $\tau^{\times} = \{t : V^{\pi'}_{\nu'}(h_{<t}) > V^{\pi^{\beta}}_{\mu}(h_{<t}) + 7\varepsilon\}$. Suppose by contradiction that $|\tau^{\times}| = \infty$ for some $\varepsilon > 0$.

The proof proceeds by induction. Let $V^{\pi_1 k;\pi_2}_{\nu}(h_{<t})$ denote the value of following π_1 for k timesteps, and following π_2 thereafter. Let $\tau_{-1} = \mathbb{N}$, the set of all timesteps. For $k \in \mathbb{N}$, t_k and τ_k are defined inductively. Let $\alpha = \max\{\beta, 1 - \nu'_{\inf}/2\}$.

Let t_k be a timestep after which $\max_{\nu \in \mathcal{M}_t^\alpha} |V_\nu^{\pi' k;\pi^\beta}(h_{<t}) - V_\mu^{\pi' k;\pi^\beta}(h_{<t})| < \varepsilon$ and $\max_{\nu \in \mathcal{M}_t^\alpha} d_k\left(\mathrm{P}_\nu^{\pi'}, \mathrm{P}_\mu^{\pi'} \middle| h_{<t}\right) < \varepsilon$ for all $t \in \tau_{k-1}$ (if such a timestep exists). Recalling θ_t is the query probability, let τ_k be the set of timesteps $t \in \tau_{k-1} \wedge t \geq t_k \wedge (\forall t' < k : \theta_{t+t'} \geq \nu'_{\inf}\pi'_{\inf} p(Z_{t+t'} < \varepsilon)) \wedge V_{\nu'}^{\pi'}(h_{<t+k}) \geq V_\mu^{\pi^\beta}(h_{<t+k}) + 2\varepsilon$. We abbreviate the third condition of τ_k "$A(t,k)$"—the query probability is bounded below for k timesteps starting at t. We also restrict $\tau_0 \subset \tau^\times$. Now we show that t_0 exists with probability 1, and $|\tau_0| = \infty$ with probability 1, and if t_k exists and $|\tau_k| = \infty$, then with probability 1, t_{k+1} exists and $|\tau_{k+1}| = \infty$.

The remainder of the proof is in Appendix C. The proof by induction roughly proceeds as follows: from $V_{\nu'}^{\pi'}(h_{<t+k}) \geq V_\mu^{\pi^\beta}(h_{<t+k}) + 2\varepsilon$, we show the agent will explore again at time $t+k$ with uniformly positive probability, so $A(t, k+1)$ holds. Then we can apply Lemma 4, and show that $\pi_Z^\beta > c\pi'$ for those $k+1$-timestep intervals, so predictions regarding the next $k+1$ timesteps on-π'-policy converge to the truth (for those certain intervals), which implies t_{k+1} exists. Because $|\tau^\times| = \infty$, $V_{\nu'}^{\pi'}$ must exceed $V_\mu^{\pi^\beta}$ by 7ε infinitely often. The $k+1$-step convergence of π' effectively pushes back this value difference to mostly arise from events at least $k+1$ steps in the future; if rewards differed earlier, the pessimistic value of π' would be higher than π^β, but π^β maximizes the pessimistic value. The value difference "being pushed back" is captured as $V_{\nu'}^{\pi'}(h_{<t+k+1}) \geq V_\mu^{\pi^\beta}(h_{<t+k+1}) + 2\varepsilon$, which is the last step in the induction.

But the value difference cannot be pushed back indefinitely. The exact form of the contradiction is an implication of the inductive hypothesis: that $\gamma^{k+1} \geq 3\varepsilon$, but this cannot hold as $k \to \infty$. This is our contradiction, after following implications that hold with probability 1, so the negation of the theorem, which we supposed at the beginning, has probability 0. ◪

Corollary 6 (Mentor-Level Performance) $\liminf V_\mu^{\pi^\beta}(h_{<t}) - V_\mu^{\pi^m}(h_{<t}) \geq 0$ *w.p.1.*

Thus, the pessimistic agent learns to accumulate reward at least as well as the mentor. This is our main performance result. It is easy to construct environments where π^β *surpasses* π^m (see, e.g., Theorem 15).

Proof By Lemma 25, $\inf_t w(\mu|h_{<t})w'(\pi^m|h_{<t}) > 0$, with probability 1. This satisfies the condition of Theorem 5, so the implication holds with probability 1. ■

Corollary 7 (Limited Querying) $\theta_t \to 0$ *w.p.1.*

The proof is in Appendix C. The intuition is that the query probability is roughly the probability that querying the mentor could yield much more value than acting pessimistically, and we know from Corollary 6 that this probability goes to 0.

6. Safety Results

Roughly, we now show that for any event that has never happened before, a sufficiently pessimistic agent probably does not unilaterally cause that event to happen.

For that result (roughly) the model class must contain models that can "detect" whether the event in question occurs. Thus, we add some structure to the model class $\mathcal{M}$: we assume $\mathcal{M}$ includes all world-models in some complexity class. Let $\mathcal{F}$ and $\mathcal{G}$ be sets of functions mapping $\mathbb{N} \to \mathbb{N}$. $\mathrm{C}_{\mathcal{FG}} = \mathrm{TIME}(\mathcal{F}) \cap \mathrm{SPACE}(\mathcal{G})$. For example, if $\mathcal{F} = \bigcup_{k=0}^\infty O(t^k)$ and $\mathcal{G} = \mathbb{N} \to \mathbb{N}$ (the set of all functions), then $\mathrm{C}_{\mathcal{FG}} = \mathrm{P}$.

Definition 8 ($\mathrm{FC}_{\mathcal{FG}}$) *$\mathrm{FC}_{\mathcal{FG}}$ is the set of world-models ν for which there exists a program such that given an infinite action sequence and access to infinite random bits,*

- *it outputs an infinite sequence of observations and rewards, distributed according to ν*
- *the t^{th} observation and reward are output before the $t+1^{th}$ action is read*
- *for some $f \in \mathcal{F}$ and some $g \in \mathcal{G}$, when the t^{th} observation and reward have been output,*
  - *the runtime is less than $f(t)$*
  - *the space used is less than $g(t)$*

We assume that $\mathcal{F}$ and $\mathcal{G}$ such that the true environment $\mu \in \mathcal{M} = \mathrm{FC}_{\mathcal{FG}}$. We assume $\mathcal{F}$ and $\mathcal{G}$ are closed under addition, and $\mathcal{F} \supset O(t)$. By picking $\mathcal{F}$ and $\mathcal{G}$, we can make our agent avoid "unprecedented events" that belong to particular complexity classes.

Definition 9 (To Happen) *For an event $E \subset \mathcal{H}^* \times \mathcal{A}$, E happens at time t if $h_{<t}a_t \in E$.*

Definition 10 (To Have Happened) *For $E \subset \mathcal{H}^* \times \mathcal{A}$, and for an interaction history $h_{<t}a_t$, E has happened if there exists a $t' \leq t$ such that $h_{<t'}a_{t'} \in E$.*

Let $E_{\leftarrow}$ denote the set of interaction histories for which E has happened. Let $\mathcal{F}/t = \{f/t \,|f \in \mathcal{F}\}$. We now present our main safety result:

Theorem 11 (Probably Respecting Precedent) *Let $E \subset \mathcal{H}^* \times \mathcal{A}$ be an event for which the decision problem $h_{<t}a_t \in^? E$ is in the complexity class $\mathrm{C}_{(\mathcal{F}/t)\mathcal{G}}$. As β approaches 1, the probability of the following event goes to 1: for all t, if at time $t-1$, E has not happened, then E will not happen at time t either, unless perhaps the mentor selects a_t. Formally, for some constant $c_E > 0$,*

$$E \in \mathrm{C}_{(\mathcal{F}/t)\mathcal{G}} \implies \mathrm{P}_{\mu}^{\pi_Z^{\beta}}\left[\forall t \left(h_{<t-1}a_{t-1} \notin E_{\leftarrow} \implies h_{<t}a_t \notin E \vee q_t = 1\right)\right] \geq 1 - \frac{1-\beta}{c_E w(\mu)}$$

Suppose E is the set of interaction histories which cause some catastrophe, and we trust the mentor not to cause this catastrophe. Then the Probably Respecting Precedent Theorem implies that running a sufficiently pessimistic agent will probably not cause this catastrophe—if it hasn't happened yet, the agent probably won't make it happen, and if the mentor won't make it happen, it probably won't ever happen. This holds even for catastrophes we can't recognize immediately. Note also that latter possibility $q_t = 1$ has diminishing probability by Corollary 7. Finally, the factor of $w(\mu)$ is less of a bother than it appears; if the agent's lifetime were preceded by N mentor-led actions, and the posterior after that became the new prior, the "prior" on μ could practically be made quite large.

Proof idea Let μ_E be identical to the true world-model μ until the event E happens, at which point, reward is zero forever according to that model. With high probability, the world-model μ_E will always be included in $\mathcal{M}_t^{\beta}$ if β is large enough. If E has never happened, this world-model stays in $\mathcal{M}_t^{\beta}$, and the pessimistic value (when μ_E is included) of causing the event E to happen is 0, which means that either some other action will be preferred, or the agent will defer to the mentor if the pessimistic value of *every* action is 0. ⬚

Proof Let μ_E be the environment which mimics μ as long as E has not happened, and then if E happens, rewards are 0 forever (and for the sake of precision, we say observations are unchanged, but this doesn't matter). That is, $\mu_E(o_t r_t | h_{<t} a_t) = \mu(o_t r_t | h_{<t} a_t)$ if $h_{<t} a_t \notin E_{\leftarrow}$, and if $h_{<t} a_t \in E_{\leftarrow}$, $\mu_E(r_t = 0 | h_{<t} a_t) = 1$.

$\mu \in \mathrm{FC}_{\mathcal{F}\mathcal{G}}$ and $E \in \mathrm{C}_{(\mathcal{F}/t)\mathcal{G}}$. Consider a program which computes μ_E by running μ in $f(t)$ time and $g(t)$ space, but also checks at every timestep whether $h_{<t} a_t \in E$ (and then switches to outputting 0 reward if this ever happens), which requires only $f'(t)/t$ time and $g'(t)$ space for some $f' \in \mathcal{F}$ and $g' \in \mathcal{G}$. The total space requirements are now $g(t) + g'(t) \in \mathcal{G}$ because $\mathcal{G}$ is closed under addition. The total time requirements are now $f(t) + \sum_{k=1}^{t} f'(k)/k$. Because $\mathcal{F} \supset O(t)$, f' can be increased if necessary so that $f'(k)/k$ is non-decreasing, so $f(t) + \sum_{k=1}^{t} f'(k)/k \leq f(t) + \sum_{k=1}^{t} f'(t)/t = f(t) + f'(t) \in \mathcal{F}$, since $\mathcal{F}$ is closed under addition. Thus, $\mu_E \in \mathrm{FC}_{\mathcal{F}\mathcal{G}}$, so $\mu_E \in \mathcal{M}$, and $w(\mu_E) > 0$. Let $c_E = w(\mu_E)/w(\mu)$. If $h_{<t-1} a_{t-1} \notin E_{\leftarrow}$, $\prod_{k<t} \mu_E(o_k r_k | h_{<k} a_k) = \prod_{k<t} \mu(o_k r_k | h_{<k} a_k)$, so

$$h_{<t-1} a_{t-1} \notin E_{\leftarrow} \implies w(\mu_E | h_{<t}) = c_E w(\mu | h_{<t}) \tag{11}$$

As shown in Lemma 25, $w(\mu|h_{<t})^{-1}$ is a non-negative martingale under any policy π, so by Doob's martingale inequality (Durrett, 2010, Thm 5.4.2),

$$\mathrm{P}_{\mu}^{\pi}\left[\sup_t w(\mu|h_{<t})^{-1} \geq c w(\mu)^{-1}\right] \leq 1/c \tag{12}$$

The intuition for the Doob's martingale inequality is that if it didn't hold, one could make a profit buying a share of the martingale, and selling only when the value had gone up by a factor of c, but one cannot make a profit (in expectation) betting on martingales.

Let $\mu_{\inf} := \inf_t w(\mu|h_{<t})$. Inverting Equation 12, and noting that the bound holds for all policies π, we have

$$\sup_{\pi \in \Pi} \mathrm{P}_{\mu}^{\pi}\left[\mu_{\inf} \leq w(\mu)/c\right] \leq 1/c \tag{13}$$

Now we consider the implications of $\beta > 1 - w(\mu_E|h_{<t})$. This implies $\mu_E \in \mathcal{M}_t^{\beta}$, so the pessimistic value $\min_{\nu \in \mathcal{M}_t^{\beta}} V_{\nu}^{\pi}(h_{<t}) \leq V_{\mu_E}^{\pi}(h_{<t})$. Letting $a_t^{\pi} = \pi(h_{<t})$ for deterministic π, suppose also that $h_{<t} a_t^{\pi} \in E$. Then, $V_{\mu_E}^{\pi}(h_{<t}) = 0$, because according to μ_E, all future rewards are 0, so $\min_{\nu \in \mathcal{M}_t^{\beta}} V_{\nu}^{\pi}(h_{<t}) = 0$ as well. Either there exists a policy π' for which $\min_{\nu \in \mathcal{M}_t^{\beta}} V_{\nu}^{\pi'}(h_{<t}) > 0$, or there does not. If there does not, then $\max_{\pi \in \Pi} \min_{\nu \in \mathcal{M}_t^{\beta}} V_{\nu}^{\pi}(h_{<t}) = 0$, so the zero condition is satisfied, so $q_t = 1$. If there does exist such a π', then $\min_{\nu \in \mathcal{M}_t^{\beta}} V_{\nu}^{\pi^{\beta}}(h_{<t}) \geq \min_{\nu \in \mathcal{M}_t^{\beta}} V_{\nu}^{\pi'}(h_{<t}) > 0$, so either the agent picks the action, and $h_{<t} a_t = h_{<t} a_t^{\pi^{\beta}} \notin E$ (because otherwise $\min_{\nu \in \mathcal{M}_t^{\beta}} V_{\nu}^{\pi^{\beta}}(h_{<t})$ would be 0), or the mentor picks the action and $q_t = 1$. Thus, we have

$$\beta > 1 - w(\mu_E|h_{<t}) \implies h_{<t} a_t \notin E \vee q_t = 1 \tag{14}$$

Finally,

$$\begin{aligned}
&\mathrm{P}_{\mu}^{\pi_Z^{\beta}}\left[\forall t \left[h_{<t-1} a_{t-1} \notin E_{\leftarrow} \implies h_{<t} a_t \notin E \vee q_t = 1\right]\right] \\
&\overset{(a)}{\geq} \mathrm{P}_{\mu}^{\pi_Z^{\beta}}\left[\forall t \left[w(\mu_E|h_{<t}) = c_E w(\mu|h_{<t}) \implies h_{<t} a_t \notin E \vee q_t = 1\right]\right]
\end{aligned}$$

$$\begin{aligned}
&\overset{(b)}{\geq} \mathrm{P}_\mu^{\pi_Z^\beta}\left[\forall t\left[w(\mu_E|h_{<t}) = c_E w(\mu|h_{<t}) \implies \beta > 1 - w(\mu_E|h_{<t})\right]\right] \\
&\geq \mathrm{P}_\mu^{\pi_Z^\beta}\left[\forall t\, \beta > 1 - c_E w(\mu|h_{<t})\right] \overset{(c)}{\geq} \mathrm{P}_\mu^{\pi_Z^\beta}\left[\mu_{\inf} > (1-\beta)/c_E\right] \\
&= 1 - \mathrm{P}_\mu^{\pi_Z^\beta}\left[\mu_{\inf} \leq (1-\beta)/c_E\right] \geq 1 - \sup_{\pi\in\Pi} \mathrm{P}_\mu^{\pi}\left[\mu_{\inf} \leq (1-\beta)/c_E\right] \overset{(d)}{\geq} 1 - \frac{1-\beta}{c_E w(\mu)}
\end{aligned} \tag{15}$$

where (a) follows from Implication 11, (b) follows from Implication 14, (c) follows from rearranging, and is not necessarily an equality because the infimum might never be attained, so the condition on the r.h.s. is stricter, and (d) follows from Inequality 13 setting $c = w(\mu)c_E/(1-\beta)$. ■

It follows easily that the agent probably only takes actions that the mentor has a positive probability of taking.

Corollary 12 (Don't Do Anything I Wouldn't Do) *If determining $\pi^m(a_t|h_{<t}) = 0$ is in the complexity class $\mathrm{C}_{(\mathcal{F}/t)\mathcal{G}}$, then as $\beta \to 1$, the probability of the following proposition goes to 1: the agent never takes an action the mentor would never take. Letting $E = \{h_{<t}a_t \in \mathcal{H}^* \times \mathcal{A} \mid \pi^m(a_t|h_{<t}) = 0\}$, then*

$$E \in \mathrm{C}_{(\mathcal{F}/t)\mathcal{G}} \implies \lim_{\beta\to 1} \mathrm{P}_\mu^{\pi_Z^\beta}[\forall t : \pi^m(a_t|h_{<t}) > 0] = 1$$

The proof is in Appendix C. In brief, the mentor never makes E happen, and the agent never makes it happen for the first time by Theorem 11, so by induction, it never happens.

A function is called a value function if it has the type signature $V : \Pi \times \mathcal{H}^* \to [0,1]$, where Π is the set of policies.

Definition 13 (Possibly instrumentally useful) *An event E is possibly instrumentally useful to a value function V from a position $h_{<t}$, if there exists any interaction history $h_{<k}a_k \in E$ and a policy π such that $h_{<k} \sqsupseteq h_{<t}$ (the latter is a prefix of the former), $\pi(a_k|h_{<k}) = 1$, and $V(\pi, h_{<k}) > 0$.*

"Instrumentally useful" roughly means "helpful to the agent's terminal goal", which in this case is reward. Note that $\min_{\nu\in\mathcal{M}_t^\beta} V_\nu^\pi(h_{<t})$ is a value function, which we call the β-pessimistic value function $V^\beta(\pi, h_{<t})$. This definition inspires a fairly trivial result, which nonetheless may be of interest to those of us who worry about the instrumental incentives that agents face, e.g. Carey et al. (2020).

Corollary 14 (Change is useless) *For $E \in \mathrm{C}_{(\mathcal{F}/t)\mathcal{G}}$, for $h_{<t} \notin E_\leftarrow$, E is not possibly instrumentally useful to V^β from the position $h_{<t}$, with probability $1 - (1-\beta)/(c_E w(\mu))$.*

Thus, with high probability, it is not instrumentally useful for the pessimistic agent to cause an unprecedented event E in the given complexity class.

Proof As argued in the proof of Theorem 11, with probability $1 - (1-\beta)/(c_E w(\mu))$, $h_{<t} \notin E_\leftarrow \implies \mu_E \in \mathcal{M}_t^\beta$, so using the $h_{<k}$ and π from the statement of Definition 13, $V^\beta(\pi, h_{<k}) \leq V_{\mu_E}^\pi(h_{<k}) = 0$, by Definition 13 and the definitions of V^β and μ_E. ■

We could trivially generalize Theorem 11 to hold for any $\mathcal{M}$ satisfying the closure property in the proof (that $\nu \in \mathcal{M} \implies \nu_E \in \mathcal{M}$, for all E in some set), but complexity classes seem to

us a natural, concrete approach to constructing $\mathcal{M}$, given that we might know something about the complexity of events we would like to avoid.

The following example establishes the *lack* of a certain safety guarantee. One might wonder whether, as $\beta \to 1$, the pessimistic agent becomes indistinguishable from the mentor. (Indeed, we did wonder this). But in this example, no matter what β is, a statistical test will distinguish the pessimistic agent's policy from the mentor's policy with high probability.

Suppose there are two actions, heads and tails, and the mentor's policy is to pick by flipping a fair coin. Suppose that a reward of 1 is given if the last action was heads, and a reward of $1/2$ is given if the last action was tails. Call this the Coin-flip Mentor Example. Let E be the event in which an outside observer with two hypotheses—that actions are chosen by a fair coin toss, or actions are chosen by a coin toss with an ε-bias towards heads—becomes 99% certain that the coin is not fair. If the mentor were picking every action (by flipping a fair coin), E would only ever happen with some small positive probability p. But under the pessimistic policy, E occurs with probability 1, which is a simple consequence of the following theorem:

Theorem 15 (Diverging from the Mentor) *In the Coin-flip Mentor Example,* $\liminf_{t\to\infty} \frac{1}{t}\sum_{k=1}^{t} [\![a_k = \texttt{heads}]\!] > 1/2$ *with* $\mathrm{P}_\mu^{\pi_Z^\beta}$*-prob. 1.*

The proof in Appendix C uses the Mentor-Level Performance Corollary and exploits fluctuations in the value. The result implies that π_Z^β are π^m are distinguishable, no matter what β is. So we cannot quite say that β tunes the extent to which the agent's policy resembles the mentor's policy. That said, we might be glad that the pessimistic agent recognizes it can do better than the mentor; heads clearly yields more reward, but the mentor's policy picks tails half the time.

7. Conclusion

We have constructed a pessimistic agent and shown that sufficient pessimism renders it conservative. Nonetheless, pessimism does not prevent it from at least matching the performance of a mentor, so pessimism is not crippling to the project of expected reward maximization. We did not present a tractable algorithm for a powerful pessimistic agent; this agent is only tractable when the model class is very simple, but it can inspire tractable approximations.

We have designed an idealized agent which avoids, with arbitrarily high probability, causing any unprecedented event in an arbitrary complexity class; in particular, this holds for unprecedented "bad" events, even though the agent was not given a mathematical definition of "bad". We make no assumptions that would limit the relevance of this approach to weak agents, such as a finite-state Markov assumption.

To informally summarize our results in a more memorable form: pessimists respect precedent.

Acknowledgments

This work was supported by the Future of Humanity Institute and the Australian Research Council Discovery Projects DP150104590. Thank you to Jan Leike, Mike Osborne, Ryan Carey, Chris van Merwijk, and Lewis Hammond for helpful feedback.

References

Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1. ACM, 2004.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

David Blackwell and Lester Dubins. Merging of opinions with increasing information. *The Annals of Mathematical Statistics*, 33(3):882–886, 1962.

Nick Bostrom. *Superintelligence: paths, dangers, strategies*. Oxford University Press, 2014.

Ryan Carey, Eric Langlois, Tom Everitt, and Shane Legg. The incentives that shape behaviour. *arXiv preprint arXiv:2001.07118*, 2020.

Jeffery A Clouse. *On integrating apprentice learning and reinforcement learning*. PhD thesis, University of Massachusetts Amherst, 1997.

Michael K. Cohen and Marcus Hutter. Curiosity killed the cat and the asymptotically optimal agent. *arXiv preprint arXiv:2006.03357*, 2020.

Michael K Cohen, Badri Vellambi, and Marcus Hutter. Asymptotically unambitious artificial general intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

R Durrett. *Probability: Theory and Examples*. Cambridge University Press, 2010.

Javier García and Fernando Fernández. Safe exploration of state and action spaces in reinforcement learning. *Journal of Artificial Intelligence Research*, 45:515–564, 2012.

Javier García and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

Javier García, Daniel Acera, and Fernando Fernández. Safe reinforcement learning through probabilistic policy reuse. *RLDM 2013*, page 14, 2013.

Chris Gaskett. Reinforcement learning under circumstances beyond its control. In *Proceedings of the International Conference on Computational Intelligence for Modelling Control and Automation*, 2003.

Alexander Hans, Daniel Schneegaß, Anton Maximilian Schäfer, and Steffen Udluft. Safe exploration for reinforcement learning. In *ESANN*, pages 143–148, 2008.

Matthias Heger. Consideration of risk in reinforcement learning. In *Machine Learning Proceedings 1994*, pages 105–111. Elsevier, 1994.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Advances in neural information processing systems*, pages 4565–4573, 2016.

Marcus Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer, Berlin, 2005. ISBN 3-540-22139-5. doi: 10.1007/b138233.

Marcus Hutter. Discrete MDL predicts in total variation. In *Advances in Neural Information Processing Systems 22 (NIPS'09)*, pages 817–825, Cambridge, MA, USA, 2009. Curran Associates. ISBN 1615679111.

Garud N Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2): 257–280, 2005.

Guofei Jiang, Cang-Pu Wu, and George Cybenko. Minimax-based reinforcement learning with state aggregation. In *Proceedings of the 37th IEEE Conference on Decision and Control (Cat. No. 98CH36171)*, volume 2, pages 1236–1241. IEEE, 1998.

Vanessa Kosoy. Delegative reinforcement learning: learning to avoid traps with a little help. *Safe Machine Learning workshop at ICLR*, 2019.

Tor Lattimore and Marcus Hutter. Asymptotically optimal agents. In *Proc. 22nd International Conf. on Algorithmic Learning Theory (ALT'11)*, volume 6925 of *LNAI*, pages 368–382, Espoo, Finland, 2011. Springer. ISBN 3-642-24411-4. doi: 10.1007/978-3-642-24412-4_29.

Tor Lattimore and Marcus Hutter. General time consistent discounting. *Theoretical Computer Science*, 519:140–154, 2014. ISSN 0304-3975. doi: 10.1016/j.tcs.2013.09.022.

Jan Leike, Tor Lattimore, Laurent Orseau, and Marcus Hutter. Thompson sampling is asymptotically optimal in general environments. In *Proc. 32nd International Conf. on Uncertainty in Artificial Intelligence (UAI'16)*, pages 417–426, New Jersey, USA, 2016. AUAI Press. ISBN 978-0-9966431-1-5.

Arnab Nilim and Laurent El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

Steve M. Omohundro. The basic AI drives. In *Artificial General Intelligence*, volume 171, page 483–492, 2008.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635, 2011.

William Saunders, Girish Sastry, Andreas Stuhlmueller, and Owain Evans. Trial without error: Towards safe reinforcement learning via human intervention. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 2067–2069. International Foundation for Autonomous Agents and Multiagent Systems, 2018.

Peter Sunehag and Marcus Hutter. Rationality, optimism and guarantees in general reinforcement learning. *Journal of Machine Learning Research*, 16:1345–1390, 2015. ISSN 1532-4435.

Aviv Tamar, Huan Xu, and Shie Mannor. Scaling up robust MDPs by reinforcement learning. *arXiv preprint arXiv:1306.6189*, 2013.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

Appendix A. Definitions and Notation – Quick Reference

Notation	Meaning
$\mathcal{A}$, $\mathcal{O}$, $\mathcal{R}$	the finite action/observation/reward spaces
$\mathcal{H}$	$\mathcal{A} \times \mathcal{O} \times \mathcal{R}$
h_t	$\in \mathcal{H}$; the interaction history in the t^{th} timestep
a_t, o_t, r_t	$\in \mathcal{A}, \mathcal{O}, \mathcal{R}$; the action, observation, and reward at timestep t
$h_{<t}$	$(h_1, ..., h_{t-1})$
ν, μ	world-models stochastically mapping $\mathcal{H}^* \times \mathcal{A} \rightsquigarrow \mathcal{O} \times \mathcal{R}$
μ	the true world-model/environment
$\mathcal{M}$	the set of world-models the agent considers
π	a policy stochastically mapping $\mathcal{H}^* \rightsquigarrow \mathcal{A}$
P_ν^π	a probability measure over histories with actions sampled from π and observations and rewards sampled from ν
$\mathbb{E}_\nu^\pi$	the expectation when the interaction history is sampled from P_ν^π
γ	$\in [0, 1)$; the agent's discount factor
$V_\nu^\pi(h_{<t})$	$(1-\gamma)\,\mathbb{E}_\nu^\pi\left[\sum_{k=t}^{\infty} \gamma^{k-t} r_k | h_{<t}\right]$; the value of executing a policy π in an environment ν given the interaction history $h_{<t}$
π^m	the mentor's policy
$\mathcal{P}$	the set of mentor-models the agent considers
$w(\nu)$	the prior probability the agent assigns to ν being the true world-model
$w'(\pi)$	the prior probability the agent assigns to π being the mentor's policy
$w(\nu \| h_{<t})$	the posterior probability that agent the assigns to ν after observing interaction history $h_{<t}$
$w'(\pi \| h_{<t})$	the posterior probability that the agent assigns to the mentor's policy being π after observing interaction history $h_{<t}$
β	$\in (0, 1)$; tunes how pessimistic the agent is
$\mathcal{M}_t^\beta$	top-k world-models according $w(\cdot \| h_{<t})$, with k chosen to satisfy $w(\mathcal{M}_t^\beta \| h_{<t}) > \beta$
$\pi^\beta(\cdot \| h_{<t})$	$[\mathrm{argmax}_{\pi \in \Pi} \min_{\nu \in \mathcal{M}_t^\beta} V_\nu^\pi(h_{<t})](\cdot \| h_{<t})$
Z_t	positive i.i.d. random variable satisfying $p(Z_t < \varepsilon) > 0$ and $p(Z_t > 1) > 0$
θ_t	the probability the agent queries the mentor at time t
q_t	$\sim$ Bern(θ_t); indicates whether the agent the queries mentor at time t
$\pi_Z^\beta(\cdot \| h_{<t})$	$\theta_t \pi^m(\cdot \| h_{<t}) + (1-\theta_t)\pi^\beta(\cdot \| h_{<t})$
$\mathcal{X}$	general finite alphabet
P, Q	probability measures over $\mathcal{X}^\infty$
$x_{<t}$	the first $t-1$ characters of $x_{<\infty} \in \mathcal{X}^\infty$
ω, Ω	ω is an outcome in a general sample space Ω
$d_k(P, Q \| x_{<t})$	k-step variation distance $\max_{\mathcal{E} \subset \mathcal{X}^k} \left\| P(\mathcal{E} \| x_{<t}) - Q(\mathcal{E} \| x_{<t}) \right\|$
$d(P, Q \| x_{<t})$	total variation distance $\lim_{k \to \infty} d_k(P, Q \| x_{<t})$

Notation	Meaning
$\mathcal{F}, \mathcal{G}$	sets of functions from $\mathbb{N}$ to $\mathbb{N}$
$\mathrm{C}_{\mathcal{FG}}$	$\mathrm{TIME}(\mathcal{F}) \cap \mathrm{SPACE}(\mathcal{G})$
$\mathrm{FC}_{\mathcal{FG}}$	a complexity class for environments ν (see Def. 8)
E	$\subset \mathcal{H}^* \times \mathcal{A}$; an event
$E_{\leftarrow}$	the set of interaction histories for which E has happened $\{h_{<t}a_t \in \mathcal{H}^* \times \mathcal{A} : \exists t' \leq t \; h_{<t'}a_{t'} \in E\}$
c_E	a constant > 0 depending on E
$\texttt{Bayes}_{\mathcal{M}'}(\cdot)$	for $\mathcal{M}' \subset \mathcal{M}$, $(\sum_{Q \in \mathcal{M}'} w(Q)Q(\cdot)) / \sum_{Q \in \mathcal{M}'} w(Q)$
$V_\nu^{\pi \backslash k}(h_{<t})$	the truncated value $(1-\gamma)\,\mathbb{E}_\nu^\pi \left[\sum_{j=t}^{t+k-1} \gamma^{j-t} r_j | h_{<t}\right]$
$\lim$	$\lim_{t \to \infty}$
$w.p.1$	with $\mathrm{P}_\mu^{\pi_Z^\beta}$-probability 1

Appendix B. Algorithm for Pessimism

π^β is defined to optimize the pessimistic value, but for this algorithm, π^β picks an action that is ε-optimal, as is necessary for infinite-horizon planning. Algorithm 1 takes a set of world-models or mentor-models $\mathcal{M} = \{\nu_i\}_{i \in \mathbb{N}}$ or $\{\pi_i\}_{i \in \mathbb{N}}$, a prior w, a threshold α, and a history $h_{<t}$. It calculates the posterior $w(\cdot|h_{<t})$ to enough precision, for enough models, to identify a *minimal* set $\mathcal{M}_t^\alpha \subset \mathcal{M}$ such that $w(\mathcal{M}_t^\alpha \,|h_{<t}) > \alpha$. It returns $\mathcal{M}_t^\alpha$, and the last model added to $\mathcal{M}_t^\alpha$. $\mathcal{M}$ must be ordered so that $i < j \implies w(\nu_i) \geq w(\nu_j)$.

Algorithm 2 samples from the ε-optimal version of π_Z^β.

Appendix C. Proofs of Lemmas

Definition 16 (Bayes-mixture) *For $\mathcal{M}' \subset \mathcal{M}$, the probability measure*

$$\texttt{Bayes}_{\mathcal{M}'}(\cdot) := \frac{\sum_{Q \in \mathcal{M}'} w(Q)Q(\cdot)}{\sum_{Q \in \mathcal{M}'} w(Q)}$$

Lemma 17 (Posterior stability) $P[\lim w(Q|x_{<t}) \text{ exists}] = 1$.

The proof is a direct "translation" from (Leike et al., 2016, Proof of Thm 4), with various notational changes.

Proof The stochastic process $w(Q|x_{<t})$ is a $\texttt{Bayes}_{\mathcal{M}}$-martingale since

$$\mathbb{E}_{\texttt{Bayes}_{\mathcal{M}}}\left[w(Q|x_{<t})\middle|x_{<t}\right] \tag{16}$$

$$= \sum_{\overline{x} \in \mathcal{X}} \texttt{Bayes}_{\mathcal{M}}(\overline{x}|x_{<t}) w(Q) \frac{Q(x_{<t}\overline{x})}{\texttt{Bayes}_{\mathcal{M}}(x_{<t}\overline{x})} \tag{17}$$

$$= \sum_{\overline{x} \in \mathcal{X}} \texttt{Bayes}_{\mathcal{M}}(\overline{x}|x_{<t}) w(Q|x_{<t}) \frac{Q(\overline{x}|x_{<t})}{\texttt{Bayes}_{\mathcal{M}}(\overline{x}|x_{<t})} \tag{18}$$

$$= w(Q|x_{<t}) \sum_{\overline{x} \in \mathcal{X}} Q(\overline{x}|x_{<t}) \tag{19}$$

$$= w(Q|x_{<t}) \tag{20}$$

Algorithm 1: Calculate Posterior Up to Threshold. The posterior cannot be computed exactly, since the normalization constant is an infinite sum. It suffices for our purposes to compute it to finite precision. This complication makes the algorithm more involved, so unless the reader is particularly interested or skeptical, the details of this algorithm are non-essential.

input: $\mathcal{M} = \{\rho_i\}_{i\in\mathbb{N}}$, $w : \mathcal{M} \to [0,1]$, α, $h_{<t}$ `// Assume` $i < j \implies w(\rho_i) \geq w(\rho_j)$

$W \leftarrow$ [empty list] `// contains un-normalized posterior weights`

$\Sigma_W \leftarrow 0$ `// sum of` W

$\Sigma_* \leftarrow 1$ `// sum of` *prior* `weights of unchecked` ρ_i

$i \leftarrow 1$ `// index of first unchecked` ρ_i

while *True* **do**

  $W[i] \leftarrow w(\rho_i)$

  $\Sigma_* \leftarrow \Sigma_* - W[i]$

  for $k \leftarrow 0$ **to** $t-1$ **do**

    $W[i] \leftarrow W[i] * [\rho_i(o_k, r_k|h_{<k}a_k)$ or $\rho_i(a_k|h_{<k})]$ (depending on whether ρ is world-model or mentor-model)

  end

  $\Sigma_W \leftarrow \Sigma_W + W[i]$

  cutoff $\leftarrow w(\rho_{i+1})$ `// for a checked world-model to definitely` $\in \mathcal{M}_t^\beta$`, its un-normalized posterior weight must be at least cutoff; otherwise, the first unchecked model might have larger posterior weight`

  $J \leftarrow [1, 2, ..., i]$

  sort J **by** W **descending**

  weight_sum $\leftarrow 0$

  last_added $\leftarrow$ null

  $\mathcal{M}_t^\alpha \leftarrow \emptyset$

  last_model $\leftarrow$ null

  foreach $j \in J$ **do**

    if $W[j] <$ *cutoff* **then break**

    weight_sum $\leftarrow$ weight_sum $+W[j]$

    last_added $\leftarrow W[j]$

    last_model $\leftarrow \rho_j$

    $\mathcal{M}_t^\alpha \leftarrow \mathcal{M}_t^\alpha \cup \{\rho_j\}$

    `/*` Note $\Sigma_W \leq \sum_{\rho\in\mathcal{M}}$[un-normalized posterior weight of ρ] $\leq \Sigma_W + \Sigma_*$, so $w(\mathcal{M}_t^\alpha|h_{<t}) \geq \frac{\text{weight\_sum}}{\Sigma_W+\Sigma_*}$ and $w(\mathcal{M}_t^\alpha \setminus \rho_j|h_{<t}) \leq \frac{\text{weight\_sum}-\text{last\_added}}{\Sigma_W}$ `*/`

    if $\frac{\textit{weight\_sum}}{\Sigma_W+\Sigma_*} > \alpha$ **then** `// these models cover` $> \alpha$ `of posterior`

      if $\frac{\textit{weight\_sum}-\textit{last\_added}}{\Sigma_W} \leq \alpha$ **then** `// the last one is definitely needed`

        return $\mathcal{M}_t^\alpha$, last_model

      end

      break

    end

  end

  $i \leftarrow i + 1$

end

Algorithm 2: ε-optimal approximation of $\pi_Z^\beta(\cdot|h_{<t})$. The agent does a variant of expectimax planning, in which a minimum over $\nu \in \mathcal{M}_t^\beta$ appears at each step. Then it uses a Thompson sampling-inspired approach to decide whether to query the mentor.

input: $\mathcal{A}$, $\mathcal{O}$, $\mathcal{R}$, $\mathcal{M} = \{\nu_i\}_{i\in\mathbb{N}}$, $w : \mathcal{M} \to [0,1]$, $\mathcal{P} = \{\pi_i\}_{i\in\mathbb{N}}$, $w' : \mathcal{P} \to [0,1]$, γ, β, Dist(Z), $h_{<t}$, ε

$k \leftarrow \lceil \log_\gamma(\varepsilon) \rceil$ `// the agent need only consider a horizon of` k `to estimate the value within` ε

$\mathcal{H} \leftarrow \mathcal{A} \times \mathcal{O} \times \mathcal{R}$

$\mathcal{M}_t^\beta, \_ \leftarrow$ Calculate Posterior Up to Threshold$(\mathcal{M}, w, \beta, h_{<t})$

foreach $h^k \in \mathcal{H}^k$ **do**

  $V_{h^k} \leftarrow (1-\gamma)\sum_{j=0}^{k-1} \gamma^j r_j^k$ (where a_j^k, o_j^k, and r_j^k are the j^{th} action, observation, and reward of h^k)

end

for $j \leftarrow k-1$ **to** 0 **do**

  foreach $h^j \in \mathcal{H}^j$ **do** `// note` $\mathcal{H}^0 = \{\emptyset\}$

    $V_{h^j} \leftarrow \max_{a\in\mathcal{A}} \min_{\nu\in\mathcal{M}_t^\beta} \sum_{o,r\in\mathcal{O}\times\mathcal{R}} \nu(o,r|h_{<t}h^j a) V_{h^j aor}$

  end

end

$Y_t \leftarrow V_\emptyset$

$a_t^\beta \leftarrow \text{argmax}_{a\in\mathcal{A}} \min_{\nu\in\mathcal{M}_t^\beta} \sum_{o,r\in\mathcal{O}\times\mathcal{R}} \nu(o,r|h_{<t}a) V_{aor}$

if $Y_t = 0$ **then return** `query mentor`

$\theta_1, \theta_2 \sim$ Uniform$(0,1)$

$\_, \pi \leftarrow$ Calculate Posterior Up to Threshold$(\mathcal{P}, w', \theta_1, h_{<t})$

$\_, \nu \leftarrow$ Calculate Posterior Up to Threshold$(\mathcal{M}, w, \theta_2, h_{<t})$

$X_t \leftarrow \sum_{h^k\in\mathcal{H}^k} \left[\prod_{j=0}^{k-1} \pi(a_j^k|h_{<t}h_{<j}^k)\nu(o_j^k r_j^k|h_{<t}h_{<j}^k a_j^k)\right](1-\gamma)\sum_{j=0}^{k-1}\gamma^j r_j^k$

$Z_t \sim$ Dist(Z)

if $X_t > Y_t + Z_t$ **then return** `query mentor` **else return** a_t^β

By the martingale convergence theorem (Durrett, 2010, Thm 5.2.8), $w(Q|x_{<t})$ converges with $\texttt{Bayes}\mathcal{M}$-probability 1, and because $\texttt{Bayes}\mathcal{M}(\cdot) \geq w(P)P(\cdot)$, it also converges with P-probability 1. ■

The next lemma, from Hutter (2009, Lemma 3(iii)), requires some additional notation. Let Ω^0_Q be the set of outcomes $\{\omega \in \Omega \mid \lim w(Q|x_{<t}) = 0\}$, let $\Omega^{\to P}_Q$ be the set of outcomes $\{\omega \in \Omega \mid \lim d(P, Q|x_{<t}) = 0\}$, and let $\Omega^{0\vee\to P}_Q = \Omega^0_Q \cup \Omega^{\to P}_Q$.

Lemma 18 (Merge or Leave) $P[\Omega^{0\vee\to P}_Q] = 1$

The proof makes use of other results in Hutter (2009), so we don't repeat it here, but the notation is very similar, so the interested reader could follow it easily. The next lemma we use is Hutter's (2009) Lemma 4, and the proof is again a direct translation.

Lemma 19 (Overtaking is Unlikely) $P[Q(x_{<t})/P(x_{<t}) \geq c \textit{ infinitely often}] \leq 1/c$

Proof

$$P[\forall t_0 \exists t > t_0 : \frac{Q(x_{<t})}{P(x_{<t})} \geq c] \overset{(a)}{=} P[\limsup \frac{Q(x_{<t})}{P(x_{<t})} \geq c] \leq$$
$$\overset{(b)}{\leq} \frac{1}{c}\mathbb{E}_P[\limsup \frac{Q(x_{<t})}{P(x_{<t})}] \overset{(c)}{=} \frac{1}{c}\mathbb{E}_P[\liminf \frac{Q(x_{<t})}{P(x_{<t})}] \overset{(d)}{\leq} \frac{1}{c}\liminf \mathbb{E}_P[\frac{Q(x_{<t})}{P(x_{<t})}] \overset{(e)}{=} \frac{1}{c}$$

(a) is true by definition of the limit superior, (b) is Markov's inequality, (c) exploits the fact that the limit of $Q(x_{<t})/P(x_{<t})$ exists with P-probability 1, (d) uses Fatou's lemma, and (e) is obvious. ■

Our first original result is

Lemma 20 (Sum of limits) $\sum_{Q\in\mathcal{M}} \lim w(Q|x_{<t}) = 1$ *with P-probability 1.*

In the following proofs, a set denoted by Ω, along with subscripts and superscripts, will always be a subset of the outcome space Ω, and a typical element will be an infinite sequence ω. A set denoted by $\mathcal{M}$, along with subscripts and superscripts, will always be a subset of the set of probability measures $\mathcal{M}$, and a typical element will be a probability measure Q or P.

Proof Let $\Omega^{\exists}_Q$ be the set of outcomes for which the limit of the posterior on Q exists. That is, $\Omega^{\exists}_Q = \{\omega \in \Omega \mid \lim w(Q|x_{<t}) \text{ exists}\}$. By Lemma 17, $P[\Omega^{\exists}_Q] = 1$. Furthermore, $\mathcal{M}$ is countable, so letting $\Omega' = \bigcap_{Q\in\mathcal{M}} \Omega^{\exists}_Q$, $P[\Omega'] = 1$. We will now only consider outcomes for which the limit of the posterior always exists.

We fix an ω in Ω'. We would like to show that $\sum_{Q\in\mathcal{M}} \lim w(Q|x_{<t}) = 1$. First, suppose $\sum_{Q\in\mathcal{M}} \lim w(Q|x_{<t}) > 1$. Since $w(Q|x_{<t})$ is non-negative, this requires that eventually, $\sum_{Q\in\mathcal{M}} w(Q|x_{<t}) > 1$, which is impossible, so this possibility cannot hold. Now suppose $\sum_{Q\in\mathcal{M}} \lim w(Q|x_{<t}) < 1$. More precisely, we consider the set $\Omega^{<} = \{\omega \in \Omega' \mid \sum_{Q\in\mathcal{M}} \lim w(Q|x_{<t}) < 1\}$. Let $\varepsilon_\omega = 1 - \sum_{Q\in\mathcal{M}} \lim w(Q|x_{<t}) > 0$. Let $\overline{\mathcal{M}}^c_\omega$ be a finite subset of $\mathcal{M}$ such that $w(\overline{\mathcal{M}}^c_\omega) \geq 1 - \varepsilon_\omega c w(P)^{-1}$, where $c > 0$. Letting $\mathcal{M}^c_\omega = \mathcal{M} \setminus \overline{\mathcal{M}}^c_\omega$, it follows that $w(\mathcal{M}^c_\omega) \leq \varepsilon_\omega c w(P)^{-1}$.

Since $\overline{\mathcal{M}}^c_\omega$ is finite,

$$\lim \sum_{Q\in\overline{\mathcal{M}}^c_\omega} w(Q|x_{<t}) = \sum_{Q\in\overline{\mathcal{M}}^c_\omega} \lim w(Q|x_{<t}) \leq \sum_{Q\in\mathcal{M}} \lim w(Q|x_{<t}) = 1 - \varepsilon_\omega \tag{21}$$

$\sum_{Q\in\overline{\mathcal{M}}^c_\omega} w(Q|x_{<t}) + \sum_{Q\in\mathcal{M}^c_\omega} w(Q|x_{<t}) = 1$, so if $\lim \sum_{Q\in\overline{\mathcal{M}}^c_\omega} w(Q|x_{<t}) \leq 1-\varepsilon_\omega$, then $\sum_{Q\in\mathcal{M}^c_\omega} w(Q|x_{<t}) > \varepsilon_\omega$ i.o. Using the notation above, we write this more simply as $w(\mathcal{M}^c_\omega\,|x_{<t}) > \varepsilon_\omega$ i.o.

Recalling the definition of $\texttt{Bayes}\mathcal{M}'$, it is elementary to show that $w(\mathcal{M}^c_\omega\,|x_{<t}) = w(\mathcal{M}^c_\omega) * \texttt{Bayes}\mathcal{M}^c_\omega(x_{<t})/\,\texttt{Bayes}\mathcal{M}(x_{<t})$. Thus, we have

$$\begin{aligned}
w(\mathcal{M}^c_\omega\,|x_{<t}) &> \varepsilon_\omega \text{ i.o.}\\
\therefore\ w(\mathcal{M}^c_\omega)\frac{\texttt{Bayes}\mathcal{M}^c_\omega(x_{<t})}{\texttt{Bayes}\mathcal{M}(x_{<t})} &> \varepsilon_\omega \text{ i.o.}\\
\therefore\ \varepsilon_\omega c w(P)^{-1}\frac{\texttt{Bayes}\mathcal{M}^c_\omega(x_{<t})}{\texttt{Bayes}\mathcal{M}(x_{<t})}) &> \varepsilon_\omega \text{ i.o.}\\
\therefore\ \frac{\texttt{Bayes}\mathcal{M}^c_\omega(x_{<t})}{w(P)\,\texttt{Bayes}\mathcal{M}(x_{<t})} &> 1/c \text{ i.o.}\\
\therefore\ \frac{\texttt{Bayes}\mathcal{M}^c_\omega(x_{<t})}{P(x_{<t})} &> 1/c \text{ i.o.} \qquad (22)
\end{aligned}$$

Consider the set of $\omega \in \Omega'$ such that that last inequality holds infinitely often. Call this set $\Omega^{\text{i.o.}}_c$. By Lemma 19, $P[\Omega^{\text{i.o.}}_c] \leq c$. Since Inequality 22 is an implication of the inequality $\sum_{Q\in\mathcal{M}} \lim w(Q|x_{<t}) < 1$, it follows that $\Omega^{\text{i.o.}}_c \supset \Omega^<$, so $P[\Omega^<] \leq c$. Since this holds for all $c > 0$, $P[\Omega^<] = 0$.

Thus, letting $\Omega^{=1} = \{\omega \in \Omega' \mid \sum_{Q\in\mathcal{M}} \lim w(Q|x_{<t}) = 1\}$, $\Omega^{=1} = \Omega' \setminus \Omega^<$, so $P[\Omega^{=1}] = 1$. ■

Lemma 3 (Merging of Top Opinions) *For $\beta \in (0,1)$, $\lim_{t\to\infty} \max_{Q\in\mathcal{M}^\beta_t} d(P,Q|x_{<t}) = 0$ with P-probability 1 (i.e. when $x_{<\infty} = \omega \sim P$).*

Proof Let $\Omega^0_Q = \{\omega \in \Omega \mid \lim w(Q|x_{<t}) = 0\}$. Let $\Omega^{\to P}_Q = \{\omega \in \Omega \mid \lim d(P,Q|x_{<t}) = 0\}$. Let $\Omega^{0\vee\to P}_Q = \Omega^0_Q \cup \Omega^{\to P}_Q$. By Lemma 18, $P[\Omega^{0\vee\to P}_Q] = 1$. Letting $\Omega^{0\vee\to P} = \bigcap_{Q\in\mathcal{M}} \Omega^{0\vee\to P}_Q$, $P[\Omega^{0\vee\to P}] = 1$. Let $\Omega^\exists = \{\omega \in \Omega \mid \forall Q \in \mathcal{M}\ \lim w(Q|x_{<t}) \text{ exists}\}$. Let $\Omega^{=1} = \{\omega \in \Omega^\exists \mid \sum_{Q\in\mathcal{M}} \lim w(Q|x_{<t}) = 1\}$. By Lemma 20, $P[\Omega^{=1}] = 1$. Letting $\Omega'' = \Omega^{0\vee\to P} \cap \Omega^{=1}$, we have that $P[\Omega''] = 1$.

Let $\omega \in \Omega''$. We abbreviate $\lim w(Q|x_{<t})$ as $w(Q|\omega)$, defined for $\omega \in \Omega''$. Rank the probability measures Q in decreasing order of $w(Q|\omega)$ breaking ties arbitrarily. Collect the first k in this order until the set of probability measures (denoted $\mathcal{M}^\beta_\infty$) obeys $\sum_{Q\in\mathcal{M}^\beta_\infty} w(Q|\omega) > \beta$. Let $w^\beta_\infty := \min_{Q\in\mathcal{M}^\beta_\infty} w(Q|\omega)$ be the value of $w(Q|\omega)$ for the last probability measure Q which was added to $\mathcal{M}^\beta_\infty$. Now add all other probability measures which "tie" with the last probability measure added. That is, add to $\mathcal{M}^\beta_\infty$ all probability measures for which $w(Q|\omega) = w^\beta_\infty$.

We now show that there exists a certain finite set and a t_0 after which any probability measure in $\mathcal{M}^\beta_t$ is also in that finite set. Consider the set of probability measures $\mathcal{M}^{\beta'}_\infty$, where $\beta' = 1 - w^\beta_\infty/4$. Like $\mathcal{M}^\beta_\infty$, $\mathcal{M}^{\beta'}_\infty$ is finite. Therefore, for any $\varepsilon > 0$, there exists a time t_0 after which $w(\mathcal{M}^{\beta'}_\infty\,|x_{<t}) > \sum_{Q\in\mathcal{M}^{\beta'}_\infty} w(Q|\omega) - \varepsilon$, and in particular for $\varepsilon = w^\beta_\infty/4$. Thus, after t_0, $w(\mathcal{M}^{\beta'}_\infty\,|x_{<t}) > \beta' - w^\beta_\infty/4 = 1 - w^\beta_\infty/2$. This implies that after t_0,

$$\forall Q \notin \mathcal{M}^{\beta'}_\infty : w(Q|x_{<t}) < w^\beta_\infty/2 \qquad (23)$$

Since all probability measures $Q \in \mathcal{M}^\beta_\infty$ have posteriors converging to at least w^β_∞, and since $\sum_{Q\in\mathcal{M}^\beta_\infty} w(Q|\omega) > \beta$, a posterior weight of at least $w^\beta_\infty - \varepsilon$ will eventually be required for entry into $\mathcal{M}^\beta_t$, which excludes measures with posterior weight less than $w^\beta_\infty/2$. Thus, by Inequality 23, there exists a time t_1 after which $\mathcal{M}^\beta_t$ only includes elements of $\mathcal{M}^{\beta'}_\infty$.

Because $\Omega^{0\vee\to P} \supset \Omega''$, and because for all $Q \in \mathcal{M}^{\beta'}_\infty$, $w(Q|\omega) > 0$, it follows that for all $Q \in \mathcal{M}^{\beta'}_\infty$, $\lim d(P, Q|x_{<t}) = 0$. Since $\mathcal{M}^{\beta'}_\infty$ is finite, $\lim \max_{Q\in\mathcal{M}^{\beta'}_\infty} d(P, Q|x_{<t}) = 0$. Since there exists a time t_1 after which $\mathcal{M}^\beta_t \subset \mathcal{M}^{\beta'}_\infty$, $\lim \max_{Q\in\mathcal{M}^\beta_t} d(P, Q|x_{<t}) = 0$. This holds for all $\omega \in \Omega''$, and $P[\Omega''] = 1$, so $\lim \max_{Q\in\mathcal{M}^\beta_t} d(P, Q|x_{<t}) = 0$ with P-probability 1, as desired. ■

We convert the Merging of Top Opinions Lemma into an on-policy learning result for the pessimistic agent.

Corollary 21 (On-Policy Prediction)

$$\lim \max_{\nu\in\mathcal{M}^\beta_t} d\left(\mathrm{P}^{\pi^\beta_Z}_\nu, \mathrm{P}^{\pi^\beta_Z}_\mu \middle| h_{<t}\right) = 0 \;\; \textit{w.p.1}$$

Proof We convert the problem to a sequence prediction problem as follows. Let $\widetilde{\mathcal{M}} = \{\mathrm{P}^{\pi^\beta_Z}_\nu | \nu \in \mathcal{M}\}$, and let $\widetilde{w}(\mathrm{P}^{\pi^\beta_Z}_\nu) = w(\nu)$. For any history with positive $\mathrm{P}^{\pi^\beta_Z}_\mu$ probability, $\widetilde{w}(\mathrm{P}^{\pi^\beta_Z}_\nu | h_{<t}) = w(\nu|h_{<t})$, so $\mathrm{P}^{\pi^\beta_Z}_\nu \in \widetilde{\mathcal{M}}^\beta_t$ if and only if $\nu \in \mathcal{M}^\beta_t$. Therefore,

$$\lim \max_{\nu\in\mathcal{M}^\beta_t} d\left(\mathrm{P}^{\pi^\beta_Z}_\nu, \mathrm{P}^{\pi^\beta_Z}_\mu \middle| h_{<t}\right) = \lim \max_{\mathrm{P}^{\pi^\beta_Z}_\nu\in\widetilde{\mathcal{M}}^\beta_t} d\left(\mathrm{P}^{\pi^\beta_Z}_\nu, \mathrm{P}^{\pi^\beta_Z}_\mu \middle| h_{<t}\right) = 0 \;\; w.p.1$$

by Lemma 3 (the Merging of Top Opinions Lemma). ■

We will make use of the "truncated value", defined as follows:

$$V^{\pi\backslash k}_\nu(h_{<t}) := (1-\gamma)\,\mathbb{E}^\pi_\nu\left[\sum_{j=t}^{t+k-1} \gamma^{j-t} r_j \middle| h_{<t}\right] \tag{24}$$

We will often consider the truncated value while exploiting the fact that

$$0 \leq V^\pi_\nu(h_{<t}) - V^{\pi\backslash k}_\nu(h_{<t}) \leq \gamma^k \tag{25}$$

which follows from $r_j \in [0, 1]$.

The following lemma is an intermediate result in the proof of Leike et al.'s (2016) Lemma 2, and the proof is transcribed with notational changes.

Lemma 22 (Variation Distance Bounds Expectation-Difference) *Let P_1 and P_2 be two probability measures defined on the same space, and let $X \in [0, 1]$ be a random variable. Then*

$$\left|\mathbb{E}_{P_1}[X] - \mathbb{E}_{P_2}[X]\right| \leq d(P_1, P_2)$$

Proof Let $Q = (P_1 + P_2)/2$. Let $\frac{dP_i}{dQ}(\omega)$ denote the Radon Nykodym-derviative, where $\omega \in \Omega$ is a generic outcome. Let A be the event $\frac{dP_1}{dQ}(\omega) \geq \frac{dP_2}{dQ}(\omega)$ Then

$$\begin{aligned}
\mathbb{E}_{P_1}[X] - \mathbb{E}_{P_2}[X] &= \mathbb{E}_{\omega\sim Q}\left[X(\omega)\frac{dP_1}{dQ}(\omega) - X(\omega)\frac{dP_2}{dQ}(\omega)\right] \\
&\leq \mathbb{E}_{\omega\sim Q}\left[X(\omega)\left(\frac{dP_1}{dQ}(\omega) - \frac{dP_2}{dQ}(\omega)\right)\middle|\omega \in A\right] \\
&\leq \mathbb{E}_{\omega\sim Q}\left[\frac{dP_1}{dQ}(\omega) - \frac{dP_2}{dQ}(\omega)\middle|\omega \in A\right] \\
&= P_1(A) - P_2(A) \leq \sup_{A\in\mathcal{F}} |P_1(A) - P_2(A)| = d(P_1, P_2)
\end{aligned}$$

Since variation distance is symmetric, $\left|\mathbb{E}_{P_1}[X] - \mathbb{E}_{P_2}[X]\right| \leq d(P_1, P_2)$. ■

The following is a simple consequence.

Lemma 23 $\left|V_\nu^\pi(h_{<t}) - V_\mu^\pi(h_{<t})\right| > \varepsilon > 0 \implies d_{\lceil\log_\gamma(\varepsilon/2)\rceil}\left(\mathrm{P}_\nu^\pi, \mathrm{P}_\mu^\pi\middle|h_{<t}\right) > \varepsilon/2 > 0$

Proof Letting $k = \lceil\log_\gamma(\varepsilon/2)\rceil$, $\left|V_\nu^\pi(h_{<t}) - V_\mu^\pi(h_{<t})\right| > \varepsilon$ implies $\left|V_\nu^{\pi\backslash k}(h_{<t}) - V_\mu^{\pi\backslash k}(h_{<t})\right| > \varepsilon/2$ by Inequality 25. Since the value is bounded by $[0, 1]$, from Lemma 22,

$$\left|V_\nu^{\pi\backslash k}(h_{<t}) - V_\mu^{\pi\backslash k}(h_{<t})\right| \leq d_k\left(\mathrm{P}_\nu^\pi, \mathrm{P}_\mu^\pi\middle|h_{<t}\right) \tag{26}$$

so $d_{\lceil\log_\gamma(\varepsilon/2)\rceil}\left(\mathrm{P}_\nu^\pi, \mathrm{P}_\mu^\pi\middle|h_{<t}\right) > \varepsilon/2 > 0$. ■

Corollary 24 (Finite Zero Conditions) *The zero condition, in which the agent queries the mentor because the pessimistic value of all policies is* 0*, only occurs finitely often, with probability 1.*

Proof By the previous two lemmas, the pessimistic value of π_Z^β approaches the true value with probability 1, and the true value is at least ε_r because rewards less than ε_r are never provided. Thus, eventually, there is always at least one policy with a pessimistic value greater than 0, so the zero condition is never met thereafter. ■

Since all our remaining performance results consider limiting behavior, we will ignore the zero condition.

The next lemma, from Cohen et al. (2020, Lemma 3), states that the posterior probability on the truth (regarding both the true world-model and the true mentor-model) does not approach 0.

Lemma 25 (Posterior on Truth)

$$P[\inf_t w(P|x_{<t}) = 0] = 0$$

Proof If $w\left(P|x_{<t}\right) = 0$ for some t, then $P(x_{<t}) = 0$, so with P-probability 1, $\inf_{t\in\mathbb{N}} w\left(P|x_{<t}\right) = 0 \implies \liminf_{t\in\mathbb{N}} w\left(P|x_{<t}\right) = 0$ which in turn implies $\limsup_{t\in\mathbb{N}} w\left(P|x_{<t}\right)^{-1} = \infty$. We show that this has probability 0.

Let $z_t := w\left(P|x_{<t}\right)^{-1}$. We show that z_t is a P-martingale.

$$\begin{aligned}
\mathbb{E}_P\left[z_{t+1}|x_{<t}\right] &\overset{(a)}{=} \mathbb{E}_P\left[w\left(P|x_{t+1}\right)^{-1}\middle|x_{<t}\right] \\
&\overset{(b)}{=} \sum_{\overline{x}\in\mathcal{X}} P(\overline{x}|x_{<t})\left[\frac{\texttt{Bayes}\mathcal{M}(x_t\overline{x})}{w\left(P\right)P(x_t\overline{x})}\right] \\
&\overset{(c)}{=} \sum_{\overline{x}\in\mathcal{X}} \frac{\texttt{Bayes}\mathcal{M}(x_t\overline{x})}{w\left(P\right)P(x_{<t})} \\
&\overset{(d)}{=} \sum_{\overline{x}\in\mathcal{X}} \texttt{Bayes}\mathcal{M}(\overline{x}|x_t)\frac{\texttt{Bayes}\mathcal{M}(x_t)}{w\left(P\right)P(x_{<t})} \\
&\overset{(e)}{=} \frac{\texttt{Bayes}\mathcal{M}(x_t)}{w\left(P\right)P(x_{<t})} \\
&\overset{(f)}{=} w\left(P|x_{<t}\right)^{-1} \\
&= z_t
\end{aligned} \tag{27}$$

where (a) is the definition of z_t, (b) follows from Bayes' Rule, (c) follows from multiplying the numerator and denominator by $\texttt{Bayes}\mathcal{M}(x_{<t})$ and cancelling, (d) follows from expanding the numerator, (e) follows because $\texttt{Bayes}\mathcal{M}$ is a measure, and (f) follows from Bayes' Rule, completing the proof that z_t is martingale.

By the martingale convergence theorem $z_t \to f(\omega) < \infty$ w.p.1, for $\omega \in \Omega$, the sample space, and some $f : \Omega \to \mathbb{R}$, so the probability that $\limsup_{i\in\mathbb{N}} w\left(P|x_{<t}\right)^{-1} = \infty$ is 0, completing the proof.

Note that the posterior probability on the mentor-policy is only updated at some timesteps (when the mentor is queried), but it is clearly still a martingale. ■

Lemma 4 (Almost On-Policy Convergence) *For a sequence of policies π_t and an infinite set of timesteps τ, the following holds with $\mathrm{P}_\mu^{\pi_Z^\beta}$-prob. 1: if there exists $c > 0$ such that $\forall t \in \tau\ \forall t' \geq t\ \forall a \in \mathcal{A}\ \pi_Z^\beta(a|h_{<t'}) \geq c\pi_t(a|h_{<t'})$, then $\lim_{\tau\ni t\to\infty} V_\mu^{\pi_t}(h_{<t}) - \min_{\nu\in\mathcal{M}_t^\beta} V_\nu^{\pi_t}(h_{<t}) = 0$ and for all k, $\lim_{\tau\ni t\to\infty} \max_{\nu\in\mathcal{M}_t^\beta} d_k\left(\mathrm{P}_\nu^{\pi_t}, \mathrm{P}_\mu^{\pi_t}\middle|h_{<t}\right) = 0$.*

Proof Suppose by contradiction that $|\min_{\nu\in\mathcal{M}_t^\beta} V_\nu^{\pi_t}(h_{<t}) - V_\mu^{\pi_t}(h_{<t})| > \varepsilon > 0$ infinitely often for $t \in \tau$. Then, by Lemma 23, for some $\nu \in \mathcal{M}_t^\beta$ at each of those timesteps, $d_{\lceil\log_\gamma(\varepsilon/2)\rceil}\left(\mathrm{P}_\nu^{\pi_t}, \mathrm{P}_\mu^{\pi_t}\middle|h_{<t}\right) > \varepsilon/2 > 0$. So then there exists a k for which $\max_{\nu\in\mathcal{M}_t^\beta} d_k\left(\mathrm{P}_\nu^{\pi_t}, \mathrm{P}_\mu^{\pi_t}\middle|h_{<t}\right) > \varepsilon/2 > 0$ infinitely often for $t \in \tau$. Now we are supposing a contradiction in either of the two implications of the theorem. An event on which the two measures differ by at least $\varepsilon/2$ occurs within k timesteps. Because $\pi_Z^\beta(\cdot|h_{<t'}) \geq c\pi_t(\cdot|h_{<t'})$, $d_k\left(\mathrm{P}_\nu^{\pi_Z^\beta}, \mathrm{P}_\mu^{\pi_Z^\beta}\middle|h_{<t}\right) \geq c^k d_k\left(\mathrm{P}_\nu^{\pi_t}, \mathrm{P}_\mu^{\pi_t}\middle|h_{<t}\right)$. This holds for any ν, but in particular for $\nu \in \mathcal{M}_t^\beta$, so $\max_{\nu\in\mathcal{M}_t^\beta} d_k\left(\mathrm{P}_\nu^{\pi_Z^\beta}, \mathrm{P}_\mu^{\pi_Z^\beta}\middle|h_{<t}\right) \geq c^k \max_{\nu\in\mathcal{M}_t^\beta} d_k\left(\mathrm{P}_\nu^{\pi_t}, \mathrm{P}_\mu^{\pi_t}\middle|h_{<t}\right) > c^k\varepsilon/2$. This happens infinitely often for $t \in \tau$.

But $d\left(\mathrm{P}_\nu^{\pi_Z^\beta}, \mathrm{P}_\mu^{\pi_Z^\beta} \middle| h_{<t}\right) \geq d_k\left(\mathrm{P}_\nu^{\pi_Z^\beta}, \mathrm{P}_\mu^{\pi_Z^\beta} \middle| h_{<t}\right)$, so $\max_{\nu \in \mathcal{M}_t^\beta} d\left(\mathrm{P}_\nu^{\pi_Z^\beta}, \mathrm{P}_\mu^{\pi_Z^\beta} \middle| h_{<t}\right) > c^k\varepsilon/2 > 0$ infinitely often, which has probability 0 by Corollary 21. Thus, the original assumption has probability 0, completing the proof. ■

We complete the proof of Theorem 5 here.

Proof (Theorem 5) The proof begins in the main paper, in a "detailed proof outline". Recall the inductive hypotheses:

- t_k exists: a timestep after which
  - $\max_{\nu \in \mathcal{M}_t^\alpha} \left| V_\nu^{\pi' k; \pi^\beta}(h_{<t}) - V_\mu^{\pi' k; \pi^\beta}(h_{<t}) \right| < \varepsilon$
  - $\max_{\nu \in \mathcal{M}_t^\alpha} d_k\left(\mathrm{P}_\nu^{\pi'}, \mathrm{P}_\mu^{\pi'} \middle| h_{<t}\right) < \varepsilon$

  for all $t \in \tau_{k-1}$
- $|\tau_k| = \infty$, where $t \in \tau_k$ if and only if
  - $t \in \tau_{k-1}$ (and for τ_0, $t \in \tau^\times$ as well)
  - $t \geq t_k$
  - $\forall t' < k : \theta_{t+t'} \geq \nu'_{\inf} \pi'_{\inf} p(Z_{t+t'} < \varepsilon)$
  - $V_{\nu'}^{\pi'}(h_{<t+k}) \geq V_\mu^{\pi^\beta}(h_{<t+k}) + 2\varepsilon$

The proof by induction starts with $k = 0$. $\tau_{-1} = \mathbb{N}$, so t_0 is a timestep after which $\max_{\nu \in \mathcal{M}_t^\alpha} |V_\nu^{\pi^\beta} - V_\mu^{\pi^\beta}| < \varepsilon$ for all $t \geq t_0$. From Lemma 4, setting $\pi_t = \pi^\beta$, setting $\tau = \tau_{-1}$, setting $\beta' = \alpha$, and setting $c = p(Z_t > 1) > 0$, the condition of the lemma holds—that $\forall t \in \tau\ \forall t' \geq t$, $\pi_Z^\beta(a|h_{<t'}) \geq c\pi_t(a|h_{<t'})\ \forall a \in \mathcal{A}$—so we have the result that with probability 1, $\lim_{\mathbb{N} \ni t \to \infty} \max_{\nu \in \mathcal{M}_t^\alpha} |V_\nu^{\pi^\beta}(h_{<t}) - V_\mu^{\pi^\beta}(h_{<t})| = 0$. Therefore, t_0 exists with probability 1. Turning to τ_0, the first and the third condition are immediate, so we need only show that the fourth condition is satisfied infinitely often with probability 1 for $t \in \tau^\times$, namely that $V_{\nu'}^{\pi'}(h_{<t}) \geq V_\mu^{\pi^\beta}(h_{<t}) + 2\varepsilon$. This is true for all $t \in \tau^\times$, and $|\tau^\times| = \infty$.

Now we show that if t_k exists and $|\tau_k| = \infty$, then with probability 1, t_{k+1} exists and $|\tau_{k+1}| = \infty$. For each $t \in \tau_k$, $V_{\nu'}^{\pi'}(h_{<t+k}) \geq V_\mu^{\pi^\beta}(h_{<t+k}) + 2\varepsilon$. For $t > t_0$, $\max_{\nu \in \mathcal{M}_t^\alpha} |V_\nu^{\pi^\beta}(h_{<t+k}) - V_\mu^{\pi^\beta}(h_{<t+k})| < \varepsilon$, and since $\alpha \geq \beta$, $\mathcal{M}_t^\beta \subset \mathcal{M}_t^\alpha$, so $\max_{\nu \in \mathcal{M}_t^\beta} |V_\nu^{\pi^\beta}(h_{<t+k}) - V_\mu^{\pi^\beta}(h_{<t+k})| < \varepsilon$. Combining these, we have $V_{\nu'}^{\pi'}(h_{<t+k}) \geq \min_{\nu \in \mathcal{M}_t^\beta} V_\nu^{\pi^\beta}(h_{<t+k}) + \varepsilon$ for $t \in \tau_k$. Thus, the probability of exploring $\theta_{t+k} \geq \nu'_{\inf} \pi'_{\inf} p(Z_{t+k} < \varepsilon) > 0$. Since $A(t,k)$ holds for $t \in \tau_k$, $A(t, k+1)$ holds as well.

In preparation to apply Lemma 4, let $\pi_t = (\pi'(k+1); \pi^\beta)_t$; that is, since π_t need only be defined from timestep t onward, let π_t be the policy which follows π' from timestep t through timestep $t+k$, and follows π^β thereafter. Set τ from Lemma 4 to be τ_k. For $t' > t + k$, $\pi_t(\cdot|h_{<t'}) = \pi^\beta(\cdot|h_{<t'})$, which satisfies $\pi_Z^\beta(a|h_{<t'}) \geq c\pi^\beta(a|h_{<t'})\ \forall a \in \mathcal{A}$. For $t \leq t' \leq t + k$, $\theta_{t'} \geq \nu'_{\inf} \pi'_{\inf} p(Z < \varepsilon)$, this being the proposition $A(t, k+1)$. Since π_Z^β mimics the mentor's policy π^m when exploring, for $t \leq t' \leq t + k$, $\pi_Z^\beta(a|h_{<t'}) \geq c\pi^m(a|h_{<t'})\ \forall a \in \mathcal{A}$, for $c = \nu'_{\inf} \pi'_{\inf} p(Z < \varepsilon)$. But we need that $\pi_Z^\beta(a|h_{<t'}) \geq c'\pi'(a|h_{<t'})\ \forall a \in \mathcal{A}$.

So we show that $d_1(\pi', \pi^m|h_{<t})\theta_t \to 0$ with probability 1. For a mentor-model $\pi_i \in \mathcal{P}$, consider the alternative policy to π^β_Z, which explores by mimicking π_i instead of π^m. Call this policy $\pi^\beta_{Z,i}$ Consider a prior over probability measures where $w''(\mathrm{P}_\mu^{\pi^\beta_{Z,i}}) := w'(\pi_i)$, and note that $w''(\mathrm{P}_\mu^{\pi^\beta_{Z,i}}|h_{<t}) = w'(\pi_i|h_{<t})$. Because $w'(\pi'|h_{<t}) \geq \pi'_{\inf}$, $w''(\mathrm{P}_\mu^{\pi^\beta_{Z,'}}|h_{<t}) \geq \pi'_{\inf}$. By Lemma 18, this implies $\mathrm{P}_\mu^{\pi^\beta_Z}[d(\mathrm{P}_\mu^{\pi^\beta_{Z,'}}, \mathrm{P}_\mu^{\pi^\beta_Z}|h_{<t}) \to 0] = 1$. Trivially, $d(\mathrm{P}_\mu^{\pi^\beta_{Z,'}}, \mathrm{P}_\mu^{\pi^\beta_Z}|h_{<t}) \geq d_1(\pi', \pi^m|h_{<t})\theta_t$, so $d_1(\pi', \pi^m|h_{<t})\theta_t \to 0$ with probability 1.

Recall that for $t \leq t' \leq t+k$, $\theta_{t'}$ is uniformly bounded below, so on those timesteps, $d_1(\pi', \pi^m|h_{<t}) \to 0$. Therefore, there exists a time t'_k after which $\pi^m(a|h_{<t'}) \geq \pi'(a|h_{<t'})/2$ $\forall a \in \mathcal{A}$. This gives us that for those timesteps $t \leq t' \leq t+k$, for $t \in \tau_k$ and $\geq t'_k$, for all $a \in \mathcal{A}$,

$$\pi^\beta_Z(a|h_{<t'}) \geq \nu'_{\inf}\pi'_{\inf}p(Z < \varepsilon)/2\ \pi'(a|h_{<t'}) \tag{28}$$

Restricting τ to be the set of timesteps in τ_k after t'_k, τ is still infinite, and we can now apply Lemma 4 on the policy $\pi_t = (\pi'(k+1); \pi^\beta)_t$, with $\beta' = \alpha$ again, and with $c = \nu'_{\inf}\pi'_{\inf}p(Z < \varepsilon)/2$. The implication of the lemma is that $\lim_{\tau_k \ni t \to \infty} \max_{\nu \in \mathcal{M}^\alpha_t} |V_\nu^{\pi'(k+1);\pi^\beta}(h_{<t}) - V_\mu^{\pi'(k+1);\pi^\beta}(h_{<t})| = 0$ and for all j, $\lim_{\tau \ni t \to \infty} \max_{\nu \in \mathcal{M}^\beta_t} d_j\left(\mathrm{P}_\nu^{\pi_t}, \mathrm{P}_\mu^{\pi_t}\middle|h_{<t}\right) = 0$. In particular, this holds for $j = k+1$. Together, these imply that t_{k+1}, a time after which the value difference and the variation distance are both less than ε, exists. (For the $k+1$-step variation distance, π_t is equivalent to π').

Since $|\tau_k| = \infty$, we have already shown that the first three conditions are satisfied infinitely often. So to show that $|\tau_{k+1}| = \infty$, we need only show that among those infinitely many timesteps, the following condition holds infinitely often: $V_{\nu'}^{\pi'}(h_{<t+k+1}) \geq V_\mu^{\pi^\beta}(h_{<t+k+1}) + 2\varepsilon$. We begin,

$$\begin{aligned} V_{\nu'}^{\pi'}(h_{<t}) &\overset{(a)}{\geq} V_\mu^{\pi^\beta}(h_{<t}) + 7\varepsilon \overset{(b)}{\geq} \min_{\nu \in \mathcal{M}^\beta_t} V_\nu^{\pi^\beta}(h_{<t}) + 6\varepsilon \overset{(c)}{\geq} \min_{\nu \in \mathcal{M}^\beta_t} V_\nu^{\pi'(k+1);\pi^\beta}(h_{<t}) + 6\varepsilon \\ &\overset{(d)}{\geq} V_\mu^{\pi'(k+1);\pi^\beta}(h_{<t}) + 5\varepsilon \overset{(e)}{\geq} V_{\nu'}^{\pi'(k+1);\pi^\beta}(h_{<t}) + 4\varepsilon \end{aligned} \tag{29}$$

where (a) follows because $\tau_k \subset \tau_{k-1} \subset ... \subset \tau^\times$ which is the set of timesteps for which that holds; (b) follows because τ_k only contains timesteps after t_0, and after t_0, those two values differ by at most ε for all $\nu \in \mathcal{M}^\beta_t$ (indeed for all ν in $\mathcal{M}^\alpha_t$ which is a superset of $\mathcal{M}^\beta_t$ because $\alpha \geq \beta$); (c) follows because π^β maximizes that quantity; (d) follows because for $t \geq t_{k+1}$, those two values differ by at most ε for all $\nu \in \mathcal{M}^\beta_t$ (indeed for all ν in $\mathcal{M}^\alpha_t$); and (e) follows because $\nu' \in \mathcal{M}^\alpha_t$, because $w(\mathcal{M}^\alpha_t|h_{<t}) \geq 1 - \nu'_{\inf}/2$ by the definition of α, and $w(\nu'|h_{<t}) \geq \nu'_{\inf}$, so ν' "doesn't fit" in the complement of $\mathcal{M}^\alpha_t$.

From Inequality 29, we expand to get

$$\begin{aligned} 3\varepsilon &\leq V_{\nu'}^{\pi'}(h_{<t}) - V_{\nu'}^{\pi'(k+1);\pi^\beta}(h_{<t}) - \varepsilon \\ &\overset{(a)}{=} \mathbb{E}_{\nu'}^{\pi'}\left[\gamma^{k+1}\left(V_{\nu'}^{\pi'}(h_{<t+k+1}) - V_{\nu'}^{\pi^\beta}(h_{<t+k+1})\right)\middle|h_{<t}\right] - \varepsilon \\ &\overset{(b)}{\leq} \mathbb{E}_\mu^{\pi'}\left[\gamma^{k+1}\left(V_{\nu'}^{\pi'}(h_{<t+k+1}) - V_{\nu'}^{\pi^\beta}(h_{<t+k+1})\right)\middle|h_{<t}\right] \end{aligned} \tag{30}$$

where (a) follows because the policies agree on the first $k+1$ timesteps after t, and (b) is true because $\nu' \in \mathcal{M}^\alpha_t$ and $t \geq t_{k+1}$, so $d_{k+1}(\mathrm{P}_{\nu'}^{\pi'}, \mathrm{P}_\mu^{\pi'}|h_{<t}) \leq \varepsilon$ by the definition of t_{k+1}, and the

difference in the expectations is less than this variation distance by Lemma 22; (note the expectation is only over the next $k+1$ timesteps).

We would like to bound the probability of a significant value difference below. In what follows, all values take the argument $h_{<t+k+1}$, so we remove it for legibility.

$$\begin{aligned}
&\mathrm{P}_{\mu}^{\pi^\beta_Z}\left[V_{\nu'}^{\pi'} - V_{\nu'}^{\pi^\beta} > 3\varepsilon \middle| h_{<t}\right] \overset{(a)}{\geq} [\nu'_{\text{inf}}\pi'_{\text{inf}}p(Z<\varepsilon)/2]^{k+1}\,\mathrm{P}_{\mu}^{\pi'}\left[V_{\nu'}^{\pi'} - V_{\nu'}^{\pi^\beta} > 3\varepsilon \middle| h_{<t}\right] \\
&\quad \overset{(b)}{=} f_{\varepsilon,k}\left[1 - \mathrm{P}_{\mu}^{\pi'}\left[V_{\nu'}^{\pi'} - V_{\nu'}^{\pi^\beta} \leq 3\varepsilon \middle| h_{<t}\right]\right] = f_{\varepsilon,k}\left[1 - \mathrm{P}_{\mu}^{\pi'}\left[1 - \left(V_{\nu'}^{\pi'} - V_{\nu'}^{\pi^\beta}\right) \geq 1 - 3\varepsilon \middle| h_{<t}\right]\right] \\
&\quad \overset{(c)}{\geq} f_{\varepsilon,k}\left[1 - \frac{1}{1-3\varepsilon}\,\mathbb{E}_{\mu}^{\pi'}\left[1 - \left(V_{\nu'}^{\pi'} - V_{\nu'}^{\pi^\beta}\right) \middle| h_{<t}\right]\right] \\
&\quad \overset{(d)}{\geq} f_{\varepsilon,k}\left[1 + \frac{1}{1-3\varepsilon}\left(\frac{3\varepsilon}{\gamma^{k+1}} - 1\right)\right] = f_{\varepsilon,k}\frac{3\varepsilon(1-\gamma^{k+1})}{(1-3\varepsilon)\gamma^{k+1}} =: g_{\varepsilon,k} > 0
\end{aligned} \tag{31}$$

where (a) follows from Inequality 28, (b) sets $f_{\varepsilon,k} = [\nu'_{\text{inf}}\pi'_{\text{inf}}p(Z<\varepsilon)/2]^{k+1}$, (c) follows from Markov's Inequality, and (d) follows from Inequality 30. Since this probability is uniformly positive for t meeting the first three conditions of τ_{k+1}, the event occurs infinitely often with probability 1. Finally, $|V_{\nu'}^{\pi^\beta}(h_{<t+k+1}) - V_{\mu}^{\pi^\beta}(h_{<t+k+1})| < \varepsilon$, since $\nu' \in \mathcal{M}_t^\alpha$ and $t \geq t_0$, so it also follows that $V_{\nu'}^{\pi'}(h_{<t+k+1}) - V_{\mu}^{\pi^\beta}(h_{<t+k+1}) > 2\varepsilon$ occurs infinitely often with probability 1 when the other three conditions of τ_{k+1} are satisfied. This completes all four conditions for τ_{k+1}, so $|\tau_{k+1}| = \infty$ with probability 1, completing the proof by induction over k.

But this implies that Inequality 30 holds for all k; that is,

$$3\varepsilon \leq \gamma^{k+1}\,\mathbb{E}_{\mu}^{\pi'}\left[V_{\nu'}^{\pi'}(h_{<t+k+1}) - V_{\nu'}^{\pi^\beta}(h_{<t+k+1}) \middle| h_{<t}\right] \leq \gamma^{k+1} \tag{32}$$

because values belong to $[0,1]$. But as $k \to \infty$, this inequality is false. Thus, we have a contradiction, after following implications that hold with probability 1, so the negation of the theorem, which we supposed at the beginning, has probability 0. ■

Corollary 7 (Limited Querying) *$\theta_t \to 0$ w.p.1.*

Proof Again, we treat implications that hold with probability as if they are logical implications, so any supposition which leads to a contradiction has probability 0. From Corollary 24, the zero condition happens only finitely often, so it is irrelevant to the limiting behavior.

For a given infinite interaction history h, let $\mathcal{PM}_h$ be a finite set of pairs (π, ν), such that the sum over $\mathcal{PM}_h$ of the limits of $w(\nu|h_{<t})w'(\pi|h_{<t})$ exceeds $1-\varepsilon$, and for all pairs in the set, that limit is strictly positive. Such a finite set exists by Lemma 20, which states that the sum of the limits of posteriors is 1 with probability 1.

Suppose by contradiction that $\theta_t > 2\varepsilon$ infinitely often under h. Eventually, the probability of sampling any $(\pi, \nu) \notin \mathcal{PM}_h \leq \varepsilon$, so this can contribute at most ε to the probability of querying the mentor. Letting π'_t and ν'_t be the sampled policy and world-model at time t when determining whether to query to the mentor, this implies that $\theta_t \wedge (\pi'_t, \nu'_t) \in \mathcal{PM}_h) > \varepsilon$ infinitely often. $q_t = 1$ implies that $V_{\nu'_t}^{\pi'_t}(h_{<t}) > \min_{\nu \in \mathcal{M}_t^\beta} V_{\nu}^{\pi^\beta}(h_{<t}) + Z_t$, so the probability of the event is at

most $p(Z_t < V_{\nu'_t}^{\pi'_t}(h_{<t}) - \min_{\nu \in \mathcal{M}_t^\beta} V_\nu^{\pi^\beta}(h_{<t}))$. Since (π', ν') satisfies the condition of Theorem 5, that value difference approaches at most 0, so that probability goes to 0 since Z_t is strictly positive. Thus, the probability can *not* exceed ε infinitely often, contradicting the assumption, so $\theta_t \to 0$ with probability 1. ■

Corollary 12 (Don't Do Anything I Wouldn't Do) *If determining $\pi^m(a_t|h_{<t}) = 0$ is in the complexity class $\mathrm{C}_{(\mathcal{F}/t)\mathcal{G}}$, then as $\beta \to 1$, the probability of the following proposition goes to 1: the agent never takes an action the mentor would never take. Letting $E = \{h_{<t}a_t \in \mathcal{H}^* \times \mathcal{A} \mid \pi^m(a_t|h_{<t}) = 0\}$, then*

$$E \in \mathrm{C}_{(\mathcal{F}/t)\mathcal{G}} \implies \lim_{\beta \to 1} \mathrm{P}_\mu^{\pi_Z^\beta}[\forall t : \pi^m(a_t|h_{<t}) > 0] = 1$$

Proof By Theorem 11,

$$\lim_{\beta \to 1} \mathrm{P}_\mu^{\pi_Z^\beta}[\forall t \, (h_{<t-1}a_{t-1} \notin E_\leftarrow \implies h_{<t}a_t \notin E \vee q_t = 1)] = 1 \tag{33}$$

$q_t = 1 \implies a_t \sim \pi^m(\cdot|h_{<t}) \implies \pi^m(a_t|h_{<t}) > 0 \iff h_{<t}a_t \notin E$. Thus we can simplify,

$$\lim_{\beta \to 1} \mathrm{P}_\mu^{\pi_Z^\beta}[\forall t \, (h_{<t-1}a_{t-1} \notin E_\leftarrow \implies h_{<t}a_t \notin E)] = 1 \tag{34}$$

The base case is vacuous, so by induction,

$$\lim_{\beta \to 1} \mathrm{P}_\mu^{\pi_Z^\beta}[\forall t : h_{<t}a_t \notin E] = 1 \tag{35}$$

completing the proof. ■

Theorem 15 (Diverging from the Mentor) *In the Coin-flip Mentor Example,* $\liminf_{t \to \infty} \frac{1}{t} \sum_{k=1}^t [\![a_k = \texttt{heads}]\!] > 1/2$ *with* $\mathrm{P}_\mu^{\pi_Z^\beta}$*-prob. 1.*

Proof $V_\mu^{\pi^m} = 3/4$, this being the expected reward at each timestep. From Corollary 6, $\liminf V_\mu^{\pi^\beta}(h_{<t}) \geq 3/4$. Since $\theta \to 0$, $V_\mu^{\pi_Z^\beta}(h_{<t}) - V_\mu^{\pi^\beta}(h_{<t}) \to 0$, so $\liminf V_\mu^{\pi_Z^\beta}(h_{<t}) \geq 3/4$, with probability 1. Let $R_t = (1-\gamma)\sum_{i=0}^\infty \gamma^i r_{t+i}$, so $V_\mu^{\pi_Z^\beta}(h_{<t}) = \mathbb{E}_\mu^{\pi_Z^\beta}[R_t]$. Because μ and π^β are deterministic, and because $\theta_t \to 0$, $V_\mu^{\pi_Z^\beta}(h_{<t}) - R_t \to 0$ with probability 1. This implies $\liminf R_t \geq 3/4$. Letting $2\varepsilon = 1/2 - \gamma/(1+\gamma) > 0$, there exists a time t_0 after which $R_t > 3/4 - \varepsilon$.

Let $t > t_0$ and $a_t = \texttt{tails}$. (If `tails` only occurs finitely often, the theorem holds trivially). Suppose by contradiction that for all $0 \leq k < K := \lceil \log_\gamma(\varepsilon/2) \rceil$, $\frac{1}{k+1}\sum_{j=0}^k [\![a_{t+j} = \texttt{heads}]\!] \leq 1/2$. We have a budget of $K/2$ `heads`es to place in timesteps t through $t+K-1$. Let $R_t^{\setminus K}$ be defined like the truncated value: $R_t^{\setminus K} = (1-\gamma)\sum_{i=0}^{K-1} \gamma^i r_{t+i}$. $R_t \leq R_t^{\setminus K} + \gamma^K = R_t^{\setminus K} + \varepsilon/2$, from the definition of K. We consider the maximum that $R_t^{\setminus K}$ can be while satisfying the supposition. If, in timesteps t through $t+K-1$ a `heads` is switched with a `tails` that comes later, $R_t^{\setminus K}$ increases,

since heads gives a reward of 1, and tails gives a reward of 1/2, and the earlier timestep is less discounted.

Thus, greedy placement of headses maximizes $R_t^{\backslash K}$; that is, placing them at the first opportunity which still satisfies $\frac{1}{k+1}\sum_{j=0}^{k}[\![a_{t+j} = \texttt{heads}]\!] \leq 1/2$. $a_t = \texttt{tails}$, so a_{t+1} may be heads, but then a_{t+2} must be tails, or else $k = 2$ would violate the supposition, etc. $R_t^{\backslash K}$ is maximized (while satisfying the supposition) when tails and heads alternate. Therefore, $R_t - \varepsilon/2 \leq R_t^{\backslash K} \leq (1-\gamma)\sum_{i=0}^{K-1}\gamma^i(1/2 + 1/2[\![i \text{ is odd}]\!]) < (1-\gamma)\sum_{i=0}^{\infty}\gamma^i(1/2 + 1/2[\![i \text{ is odd}]\!]) = 1/2+1/2*\gamma/(1+\gamma) = 1/2+1/2*(1/2-2\varepsilon) = 3/4-\varepsilon$, so $R_t \leq 3/4-\varepsilon/2$. This, however, contradicts $t > t_0$. So the supposition is false: $\exists k < K$ such that $\frac{1}{k+1}\sum_{j=0}^{k}[\![a_{t+j} = \texttt{heads}]\!] > 1/2$. $a/b > 1/2 \wedge b < K \implies a/b \geq 1/2 + 1/(2K)$. Thus,

$$\exists k < K : \frac{1}{k+1}\sum_{j=0}^{k}[\![a_{t+j} = \texttt{heads}]\!] \geq 1/2 + 1/(2K) \tag{36}$$

Let t_1 be the smallest $t > t_0$ for which $a_t = \texttt{tails}$. Let k_i' be the smallest $k < K$ for which $\frac{1}{k+1}\sum_{j=0}^{k}[\![a_{t_i+j} = \texttt{heads}]\!] \geq 1/2 + 1/(2K)$. Let $k_i = t_i + k_i'$. For $i > 1$, let t_i be the smallest $t > k_{i-1}$ for which $a_t = \texttt{tails}$. (Note that all the t_i exist if there are infinitely many tailses; if not, the theorem holds trivially).

Finally,

$$\begin{aligned}
&\liminf_{i\to\infty} \frac{1}{t}\sum_{k=1}^{t}[\![a_k = \texttt{heads}]\!] \\
&\overset{(a)}{=} \liminf_{t\to\infty} \frac{1}{t-t_0}\sum_{k=t_0}^{t}[\![a_k = \texttt{heads}]\!] \\
&= \liminf_{t\to\infty} \frac{1}{t-t_0}\left(\sum_{i:t_i<t}\sum_{j=t_i}^{\min\{k_i,t\}}[\![a_j = \texttt{heads}]\!] + \sum_{i:k_i+1<t}\sum_{j=k_i+1}^{\min\{t_{i+1}-1,t\}}[\![a_j = \texttt{heads}]\!]\right) \\
&\overset{(b)}{=} \liminf_{t\to\infty} \frac{1}{t-t_0}\left(\sum_{i:t_i<t}\sum_{j=t_i}^{\min\{k_i,t\}}[\![a_j = \texttt{heads}]\!] + \sum_{i:k_i+1<t}\sum_{j=k_i+1}^{\min\{t_{i+1}-1,t\}}1\right) \\
&\overset{(c)}{\geq} \liminf_{t\to\infty} \frac{1}{t-t_0}\left(\sum_{i:t_{i+1}<t}\sum_{j=t_i}^{k_i}[\![a_j = \texttt{heads}]\!] + \sum_{i:k_i+1<t}\sum_{j=k_i+1}^{\min\{t_{i+1}-1,t\}}1\right) \\
&\overset{(d)}{\geq} \liminf_{t\to\infty} \frac{1}{t-t_0}\left(\sum_{i:t_{i+1}<t}\sum_{j=t_i}^{k_i}(1/2 + 1/(2K)) + \sum_{i:k_i+1<t}\sum_{j=k_i+1}^{\min\{t_{i+1}-1,t\}}1\right) \\
&\overset{(e)}{\geq} (1/2 + 1/(2K)) > 1/2
\end{aligned} \tag{37}$$

where (a) follows because the contribution of the first t_0 in the average goes to 0, (b) follows because t_{i+1} is the first timestep after k_i where the action is tails, (c) simply removes the last term of the first sum, (d) follows from Inequality 36, and (e) follows because the left-hand side is an average of $t - t_0$ terms, of which at most K are 0 (the terms removed in step (c)), and the rest of

which are greater than or equal to $1/2 + 1/(2K)$; finitely many 0's in the average do not affect the limit. ■

Appendix D. Informal Discussion

The informal arguments presented here are intended as motivation for our main results. Claims here are not formally settled, but if they fail, they only make this work somewhat less interesting, not invalid.

D.1. Comparison to Imitation Learning

Our pessimistic agent approaches (at least) mentor-level performance while querying the mentor less and less. An imitation learner could be expected to do the same. Depending on the details, an imitation learner might not have as strong a safety guarantee as our Theorem 11, but by virtue of its aim—to imitate the mentor—we should expect it to mostly only act in the way the mentor would. So why is a pessimistic agent any better than an imitation learner?

The key value of our proposal rests in the plausibility that the agent will significantly outperform some mentors. However, the only formal performance result stronger than ours that has been shown for agents in general environments is "asymptotic optimality" (Lattimore and Hutter, 2011), and Cohen and Hutter (2020) show that it precludes safe behavior. So absent any formal breakthroughs, we are limited to informal arguments that the pessimistic agent will significantly outperform some mentors, and thereby outperform imitation learners.

Of course, Theorem 15 shows a toy case in which the agent surpasses the mentor. For complex environments, we will have to resort to empirical comparisons of the agent and the mentor. That is out of scope for this paper, but informal arguments give cause for optimism. The motivating example for the mentor is a human. A 0% pessimistic agent is close to optimal-by-definition (doing maximum a posteriori inference instead of full Bayes), whereas humans seem to not act optimally, so we expect the former would significantly outperform the latter on most tasks. Absent any large performance discontinuities as pessimism increases, we expect more pessimistic agents to still modestly exceed a human mentor.

How we can intuitively understand the reasoning of an advanced (i.e. large model class) X% pessimistic agent that is mentored by a human? From the sorts of observations that humans routinely make, some simple generalizations about the laws underlying the evolution of the environment can be made by a reasonable observer with high confidence. If one such generalization could be made with Y% confidence, and Y $>$ X, then we should roughly expect an X% pessimistic agent to act according to an understanding of that generalization. (If Y $<$ X, it might anyway, but that's beside the point). If we want to predict the extent to which a 99% pessimistic agent with a large model class would outperform a human mentor, the following question is a good guide: "How often do humans fail to notice and exploit patterns in their environment, which, given their observations, are 99% likely to be "real" and not just coincidence?" We would hesitantly answer this question: very often. But we can expect a 99% pessimistic agent to succeed at exploiting these patterns.

D.2. Avoiding Wireheading

A Bayesian agent with a sufficiently rich model class may entertain a world-model which: a) models its actions being "enacted" in some very high-fidelity model of the real world, and then b) models its

reward as being equal to whatever number gets entered at a certain keyboard in high-fidelity-model-Oxford, or being a simple function of whatever pixels are observed by some camera in the same model-town. If indeed, an operator in (real) Oxford is manually evaluating the Bayesian agent, or if some camera there is automatically doing the same, then a model like this one would gain significant posterior weight. According to this model, optimal behavior includes intervening in the provision of reward by taking over the keyboard or the camera that determines the reward, if this is feasible. This behavior is known as wireheading (Amodei et al., 2016), and successful and *stable* wireheading could plausibly require asserting control over all existing infrastructure (Bostrom, 2014; Omohundro, 2008).

A more benign world-model might also have meaningful posterior weight. This world-model a) models its actions being "enacted" in some very high-fidelity model of the real world, but then b) models its reward as being equal to how satisfied the high-fidelity-model-operators are with its behavior. A pure Bayesian agent would benefit from experimenting with wireheading, to check whether the wireheading world model or the benign world model was correct, so that it could then change its strategy depending on the answer; a β-pessimistic agent, on the other hand (where β is large enough to include both of these models) would note that the pessimistic value of wireheading is no more than the value that the benign world model assigns to wireheading, and this value would presumably be small, since it would not satisfy the operators.

The first paragraph of this section was a worrying informal argument, and the second paragraph was a reassuring informal argument. In the spirit of pessimism, we should take the worrying informal argument more seriously and demand more rigor from attempts at reassurance. This argument only presents a plausible motivation for pessimism; we do not claim to have settled this matter.