Under Review at COLT vol TBD:1–22, 2021

Fully General Online Imitation Learning

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

Marcus Hutter HUTTER1.NET
Google DeepMind
Australian National University

Neel Nanda NEELNANDA27@GMAIL.COM
Google DeepMind

Abstract

In imitation learning, imitators and demonstrators are policies for picking actions given past interactions with the environment. If we run an imitator, we probably want events to unfold similarly to the way they would have if the demonstrator had been acting the whole time. No existing work provides formal guidance in how this might be accomplished, instead restricting focus to environments that restart, making learning unusually easy, and conveniently limiting the significance of any mistake. We address a fully general setting, in which the (stochastic) environment and demonstrator never reset, not even for training purposes. Our new conservative Bayesian imitation learner underestimates the probabilities of each available action, and queries for more data with the remaining probability. Our main result: if an event would have been unlikely had the demonstrator acted the whole time, that event's likelihood can be bounded above when running the (initially totally ignorant) imitator instead. Meanwhile, queries to the demonstrator rapidly diminish in frequency.

1. Introduction

Supervised learning of independent and identically distributed data is often practiced in two phases: training and deployment. This separation makes less sense if the learner's predictions affect the distribution of future tasks, since the deployment phase could lose all resemblance to the training phase. When a program's output changes its future percepts, we often call its output "actions". Supervised learning in that regime is commonly called "imitation learning", where labels are the actions of a "demonstrator" (Syed and Schapire, 2010). Our agent, acting in a general environment that responds to its actions, tries to pick actions according to the same distribution as a demonstrator.

Even in imitation learning, where it is understood that actions can change the distribution of contexts that the agent will face, it is common to separate a training phase from a deployment phase. This assumes away the possibility that the task distribution will shift significantly upon deployment and render the training data increasingly irrelevant. Here, we present an online imitation learner that is robust to this possibility.

The obvious downside is that the training never ends. The agent can always make queries for more data, but importantly, it does this with diminishing probability. It transitions smoothly from a mostly-training phase to a mostly-deployed phase. Our agent also handles totally general stochastic environments (environments serve new contexts for the agent to act in) and totally general stochastic

demonstrator-policies. No finite-state-Markov-style stationarity assumption is required for either. The lack of assumptions about the environment is a mundane point, because imitation learners don't have to learn the dynamics of the environment, but the lack of assumptions on the prediction target—the demonstrator's policy—makes these results highly non-trivial. The only assumption is that the demonstrator's policy belongs to some known countable class of possibilities. Moreover, stochasticity makes single-elimination-style learning (Gold, 1967) impossible.

For demonstrator policies this general, we present formal results that are unthinkable in the train-then-deploy paradigm. The ℓ_1 distance between the imitator and demonstrator policies converges to 0 in-mean-cube, when conditioned on a high-probability event (Theorem 7). Conditioned on the same high-probability event, we bound the KL divergence from imitator to demonstrator (Theorem 9), and we upper bound the probability of an arbitrary event under the imitator's policy, given a low probability of occurrence under the demonstrator's policy (Theorem 10). Instead of having a finite training phase, our agent's query probability converges to 0 in-mean-cube (Theorem 6).

Our imitator maintains a posterior over demonstrator-models. At each timestep, it takes the top few demonstrator-models in the posterior, in a way that depends on a scalar parameter α. Then, for each action, it considers the minimum over those models of the probability that the demonstrator picks that action. The imitator samples an action according to those probabilities, and if no action is sampled (since model-disagreement makes the probabilities to sum to less than 1), it defers to the demonstrator.

We review theoretical developments in imitation learning in Section 2, define our formal setting in Section 3, define our imitation learner in Section 4, and present formal results in Sections 5 and 6, some of which are proven in Appendix B. Appendix A collects notation and definitions.

2. Related Work

Recall that a key difficulty of imitation learning over supervised learning is the removal of a standard i.i.d. assumption. However, all existing formal work in imitation learning studies repeated finite episodes of length T; even though the dynamics are not i.i.d. from timestep to timestep within an episode, the agent learns from a sequence of episodes that are, as a whole, independent and identically distributed. Thus, the scope of existing formal work is limited to environments that "restart". A driving agent that gets housed in a new car every time it crashes (or gets hopelessly lost) enjoys a "restarting" environment, whereas a driving agent with only one car to burn does not.

In this context, Syed and Schapire (2010) reduce the problem of predicting a demonstrator's behavior to i.i.d. classification. The only assumption about the demonstrator is that the value of its policy as a function of state is arbitrarily well approximated by the value of a deterministic policy, which is only slightly weaker than assuming the demonstrator is deterministic itself. They make no assumptions about the environment, other than that we can access identical copies of it repeatedly. They show that if a classifier guessing the demonstrator's actions has an error rate of ε, then the value of the imitator's policy that uses the classifier is within $O(\sqrt{\varepsilon})$ of the demonstrator.

Judah et al. (2014) improve the label complexity of Syed and Schapire's (2010) reduction by actively deciding when to query the demonstrator, instead of simply observing N full episodes before acting. Making the same assumptions as that paper, and also assuming a realizable hypothesis class with a finite VC dimension, they attempt to reduce the number of queries before the agent can act for a whole episode on its own with an error rate less than ε. Compared to Syed and Schapire's (2010) $O(T^3/\varepsilon)$ labels, they achieve $O(T\log(T^3/\varepsilon))$.

Ross and Bagnell (2010) also reduce the problem to classification. In a trivial reduction, the imitator observes the demonstrator act from the distribution of states induced by the demonstrator policy. In this reduction, if the classifier has an error rate of ε per action on the demonstrator's state distribution, the error rate of the imitator on its own distribution is at most $T^2\varepsilon$, where T is the length of the episode. Their main contribution is to introduce a cleverer training regime for the classifier to reduce this bound to $T\varepsilon$ in environments with approximate recoverability.

Ross et al. (2011) reduce imitation learning to something else: a no-regret online learner, for which the average error rate over its lifetime approaches 0, even with a potentially changing loss function. With access to an online learner with average regret $O(1/N_{\text{predictions}})$, they construct an imitation learner with regret of the same order. However, this reduction is of dubious relevance. The no-regret online learners that they cite (e.g. Cesa-Bianchi et al. (2004)) expect i.i.d. tasks with a potentially changing bounded loss function. The no-regret online learner that their agent *needs* is one that accommodates a changing distribution of tasks; with a bounded loss function, this mismatch cannot simply be fixed with importance-weighted losses. Their agent needs this because as the imitation learner changes, the distribution over states that it visits does too. We are willing to believe that the changing data distribution that they feed to their putative no-regret online learner eventually converges, but there is no discussion of how long that might take.

There is a great deal of empirical study of imitation learning, given the practical applications, which Hussein et al. (2017) review. We call one experiment to the reader's attention, since it resembles our work in taking an active approach to querying, with an eye to risk aversion, not just label efficiency (Brown et al., 2018); they find it works.

3. Preliminaries

Let $a_t \in \mathcal{A}$ and $o_t \in \mathcal{O}$ be the action and observation at timestep $t \in \mathbb{N}$. Let $q_t \in \{0,1\}$ denote whether the imitator ($q_t = 0$) or demonstrator ($q_t = 1$) selects a_t. Let $\mathcal{H} = \{0,1\} \times \mathcal{A} \times \mathcal{O}$, and let $h_t = (q_t, a_t, o_t) \in \mathcal{H}$. Let $h_{<t} = (h_0, h_1, ..., h_{t-1})$. Let $\pi : \mathcal{H}^* \rightsquigarrow \{0,1\} \times \mathcal{A}$, where $\mathcal{H}^* = \bigcup_{i=0}^{\infty} \mathcal{H}^i$, and $\rightsquigarrow$ denotes that π gives a distribution over $\{0,1\} \times \mathcal{A}$. ϵ will denote the empty string; it is the element of $\mathcal{H}^0$. π is called a policy, and will typically be written $\pi(q_t a_t | h_{<t})$. $\pi(a_t | h_{<t})$ denotes the marginal distribution over the action. Let $\mu : \mathcal{H}^* \times \{0,1\} \times \mathcal{A} \rightsquigarrow \mathcal{O}$. μ is called the environment, and will typically be written $\mu(o_t | h_{<t} q_t a_t)$.

Let P_μ^π be the probability measure over $\mathcal{H}^\infty$ where query records and actions are sampled from π, and observations are sampled from μ. The event space is the standard sigma algebra over cylinder sets $\sigma(\{\{h_{<t} h_{t:\infty} : h_{t:\infty} \in \mathcal{H}^\infty\} : h_{<t} \in \mathcal{H}^*\})$.

Let Π be a finite or countable set of policies, and for $\pi \in \Pi$, let $w(\pi) > 0$ be a prior weight assigned to π, such that $\sum_{\pi \in \Pi} w(\pi) = 1$. This represents the imitator's initial belief distribution over the demonstrator's policy. For convenience, let Π only contain policies which assign zero probability to $q_t = 0$, since demonstrator-models may as well be convinced that the demonstrator is picking the action.

4. Imitation

Let $w(\pi|h_{<t})$ be the posterior weight after observing $h_{<t}$ that demonstrator-chosen actions were sampled from π. That is,

$$w(\pi|h_{<t}) :\propto w(\pi) \prod_{k<t:q_k=1} \pi(q_k a_k|h_{<k}) \tag{1}$$

normalized such that $\sum_{\pi\in\Pi} w(\pi|h_{<t}) = 1$. Ranking the policies by posterior weight, let $\pi_n^{h_{<t}}$ be the one with the n^{th} largest posterior weight $w(\pi|h_{<t})$, breaking ties arbitrarily. Now let $\Pi^\alpha_{h_{<t}}$ be the set of policies with posterior weights at least α times the sum of the posterior weights of policies that are at least as likely as it; that is,

$$\Pi^\alpha_{h_{<t}} := \{\pi_n^{h_{<t}} \in \Pi : w(\pi_n^{h_{<t}}|h_{<t}) \geq \alpha \sum_{m\leq n} w(\pi_m^{h_{<t}}|h_{<t})\} \tag{2}$$

Let π^d denote the demonstrator's policy, defined such that $\pi^d(q_t = 1|h_{<t}) = 1$ for all values of $h_{<t}$. The imitator's policy π^i_α is defined as follows,

$$\pi^i_\alpha(0, a|h_{<t}) := \min_{\pi'\in\Pi^\alpha_{h_{<t}}} \pi'(1, a|h_{<t}) \tag{3}$$

The 0 on the l.h.s. means the imitator is picking the action itself instead of deferring to the demonstrator, and the 1 on the r.h.s. means this is the probability of the demonstrator-model π' picking that same action.

Let $\theta_q(h_{<t}) := 1 - \sum_{a\in\mathcal{A}} \pi^i_\alpha(0, a|h_{<t})$. θ_q is the probability with which the imitator queries the demonstrator to have it pick the action. Thus,

$$\pi^i_\alpha(1, a|h_{<t}) := \theta_q(h_{<t})\pi^d(1, a|h_{<t}) \tag{4}$$

Taking the minimum over a set of models with high posterior weights is an approach to conservatism inspired by Cohen and Hutter's (2020) pessimistic agent. The pessimistic agent, unlike ours, is a reinforcement learner, but it is also designed to keep certain (risky) events unlikely.

Recall that the purpose of q_t is to record whether the demonstrator selected the action. We will also consider counterfactual imitator policies if the demonstrator policy were something else; for an arbitrary demonstrator policy π, let $\hat{\pi}_\alpha$ denote the corresponding imitator policy, so $\pi^i_\alpha = \hat{(\pi^d)}_\alpha$. This paper will investigate the probability distribution $\mathrm{P}_\mu^{\pi^i_\alpha}$ and compare it to $\mathrm{P}_\mu^{\pi^d}$.

5. General Sequence Prediction

We begin with results about a general sequence prediction setting. The main difficulty with then applying these results directly to the imitation learning setting is that the imitator doesn't observe how the demonstrator would act at every timestep. All omitted proofs appear in Appendix B.

Let $\mathcal{X}$ be an arbitrary finite alphabet. Let ν be a probability measure over $\mathcal{X}^\infty$ with the event space generated by the cylinder sets $\{\{x_{<t}\omega \mid \omega \in \mathcal{X}^\infty\}|x_{<t} \in \mathcal{X}^*\}$. Let $\mathcal{M}$ be a countable set of such probability measures, and let $w(\nu)$ be a prior weight over these measures such that

$\sum_{\nu\in\mathcal{M}} w(\nu) = 1$. Let $x_{<t} \in \mathcal{X}^t$, let $\nu(x_{<t})$ denote the probability that the infinite sequence begins with $x_{<t}$, and let $\nu(x|x_{<t}) = \nu(x_{<t}x)/\nu(x_{<t})$. We define the Bayes mixture measure

$$\xi := \sum_{\nu\in\mathcal{M}} w(\nu)\nu \tag{5}$$

and the truncated Bayes-mixture

$$\rho_i(x_{<t}) := \max_{\mathcal{M}'\subset\mathcal{M}:|\mathcal{M}'|=i} \sum_{\nu\in\mathcal{M}'} w(\nu)\nu(x_{<t}) \tag{6}$$

Let $\rho_i(x_t|x_{<t}) := \rho_i(x_{<t}x_t)/\rho_i(x_{<t})$, but note that ρ_i is not a probability measure. Instead, it is an anti-semimeasure:

$$\begin{aligned}\rho_i(x_{<t}) = \max_{\mathcal{M}'\subset\mathcal{M}:|\mathcal{M}'|=i} \sum_{\nu\in\mathcal{M}'} w(\nu)\nu(x_{<t}) &= \max_{\mathcal{M}'\subset\mathcal{M}:|\mathcal{M}'|=i} \sum_{\nu\in\mathcal{M}'} w(\nu) \sum_{x\in\mathcal{X}} \nu(x_{<t}x) \\ &\le \sum_{x\in\mathcal{X}} \max_{\mathcal{M}'\subset\mathcal{M}:|\mathcal{M}'|=i} \sum_{\nu\in\mathcal{M}'} w(\nu)\nu(x_{<t}x) = \sum_{x\in\mathcal{X}} \rho_i(x_{<t}x)\end{aligned} \tag{7}$$

since each term in the sum can be maximized separately on the r.h.s. We construct a probability measure by normalizing:

$$\rho_i^{\text{norm}}(x|x_{<t}) := \frac{\rho_i(x|x_{<t})}{\sum_{x'\in\mathcal{X}} \rho_i(x'|x_{<t})} = \frac{\rho_i(x_{<t}x)}{\sum_{x'\in\mathcal{X}} \rho_i(x_{<t}x')} \tag{8}$$

and naturally $\rho_i^{\text{norm}}(x_{<t}) := \prod_{k=0}^{t-1} \rho_i^{\text{norm}}(x_k|x_{<k})$. Let

$$\mathcal{M}_i^{x_{<t}} := \operatorname*{argmax}_{\mathcal{M}'\subset\mathcal{M}:|\mathcal{M}'|=i} \sum_{\nu\in\mathcal{M}} w(\nu)\nu(x_{<t}) \tag{9}$$

where $\mathcal{M}$ contains an arbitrary fixed order for breaking ties, and ties for $\mathcal{M}'$ are broken lexicographically. Then let

$$\rho_i^{\text{stat}}(x|x_{<t}) := \frac{\sum_{\nu\in\mathcal{M}_i^{x_{<t}}} w(\nu)\nu(x_{<t}x)}{\sum_{\nu\in\mathcal{M}_i^{x_{<t}}} w(\nu)\nu(x_{<t})} \le \frac{\sum_{\nu\in\mathcal{M}_i^{x_{<t}x}} w(\nu)\nu(x_{<t}x)}{\rho_i(x_{<t})} = \frac{\rho_i(x_{<t}x)}{\rho_i(x_{<t})} \tag{10}$$

Note that ρ_i^{stat} is a measure, since each ν is a measure. ρ_i^{stat} takes the current top-performing measures and uses a mixture over those to predict the next character. Let μ be a particular measure in $\mathcal{M}$, which we call the "true" measure. We identify some trivial relations between these functions.

$$\xi(x_{<t}) \ge \rho_i(x_{<t}) \tag{11}$$

$$\rho_i(x_{<t}) \ge w(\mu)\mu(x_{<t}) \tag{12}$$

$$\rho_i(x|x_{<t}) \ge \rho_i^{\text{norm}}(x|x_{<t}) \tag{13}$$

$$\rho_i(x|x_{<t}) \ge \rho_i^{\text{stat}}(x|x_{<t}) \tag{14}$$

where 13 follows from Equations 7 and 8. Our ρ_i, ρ_i^{norm}, and ρ_i^{stat} are closely inspired by Poland and Hutter (2005), who constructed (in our notation) ρ_1, ρ_1^{norm}, and ρ_1^{stat}.

Our first lemma bounds the deviation from ρ_i being a measure.

Lemma 1

$$0 \leq \mathbb{E}_\mu \sum_{t=0}^{\infty} \frac{\sum_{x \in \mathcal{X}} \rho_i(x_{<t}x)}{\rho_i(x_{<t})} - 1 \leq w(\mu)^{-1}$$

Proof idea ρ_i is bounded above and below by measures, save a multiplicative constant (Inequalities 11 and 12), so ρ_i converges to being a measure. ⬚

Lemma 2

$$\textit{(i)} \qquad \mathbb{E}_\mu \sum_{t=0}^{\infty} \sum_{x \in \mathcal{X}} \left| \rho_i(x|x_{<t}) - \rho_i^{\text{stat}}(x|x_{<t}) \right| \leq w(\mu)^{-1}$$

$$\textit{(ii)} \qquad \mathbb{E}_\mu \sum_{t=0}^{\infty} \sum_{x \in \mathcal{X}} \left| \rho_i(x|x_{<t}) - \rho_i^{\text{norm}}(x|x_{<t}) \right| \leq w(\mu)^{-1}$$

Proof

$$\mathbb{E}_\mu \sum_{t=0}^{\infty} \sum_{x \in \mathcal{X}} \left| \rho_i(x|x_{<t}) - \rho_i^{\text{stat}}(x|x_{<t}) \right| \overset{(a)}{=} \mathbb{E}_\mu \sum_{t=0}^{\infty} \sum_{x \in \mathcal{X}} \rho_i(x|x_{<t}) - \rho_i^{\text{stat}}(x|x_{<t}) =$$
$$\mathbb{E}_\mu \sum_{t=0}^{\infty} \frac{\sum_{x \in \mathcal{X}} \rho_i(x_{<t}x)}{\rho_i(x_{<t})} - 1 \overset{(b)}{\leq} w(\mu)^{-1} \tag{15}$$

where (a) follows from Inequality 14 and (b) follows from Lemma 1. The proof is identical for ρ_i^{norm}, except now (a) follows from Inequality 13. ■

Lemma 3 *Recalling $\nu(\cdot|x_{<t})$ is a measure over $\mathcal{X}$,*

$$\mathbb{E}_\mu \sum_{t=0}^{\infty} \text{KL}\left(\mu(\cdot|x_{<t}) \middle\| \rho_i^{\text{norm}}(\cdot|x_{<t})\right) \leq w(\mu)^{-1} + \log w(\mu)^{-1}$$

Proof idea The KL divergence telescopes over timesteps. The $\log w(\mu)^{-1}$ term comes from a gap between μ and ρ_i, and the $w(\mu)^{-1}$ term comes from a gap between ρ_i and ρ_i^{norm}. ⬚

Theorem 4

$$\mathbb{E}_\mu \sum_{t=0}^{\infty} \sum_{x \in \mathcal{X}} \left(\rho_i^{\text{stat}}(x|x_{<t}) - \mu(x|x_{<t})\right)^2 \leq 6w(\mu)^{-1} + 3$$

Proof idea ρ_i^{stat} is close to ρ_i in an ℓ_1 sense, and likewise for ρ_i and ρ_i^{norm}, and ρ_i^{norm} is close to μ in an ℓ_2 squared sense. ⬚

Proof We abbreviate $w(\mu)^{-1}$ as c. Let $[N] := (0, ..., N-1)$. We define an $N|\mathcal{X}|$-dimensional random vector depending on the infinite sequence $x_{<\infty}$:

$$\overrightarrow{\nu_1 : \nu_2}^N := (\nu_1(x|x_{<t}) - \nu_2(x|x_{<t}))_{t \in [N], x \in \mathcal{X}} \tag{16}$$

In this notation, we aim to show $\mathbb{E}_\mu ||\overrightarrow{\rho_i^{\text{stat}}:\mu^N}||_2^2 \leq 6c+3$. Lemma 2 (i) and (ii) become

$$\mathbb{E}_\mu ||\overrightarrow{\rho_i:\rho_i^{\text{stat}N}}||_1 \leq c \tag{17}$$

$$\mathbb{E}_\mu ||\overrightarrow{\rho_i:\rho_i^{\text{norm}N}}||_1 \leq c \tag{18}$$

Therefore,

$$\mathbb{E}_\mu ||\overrightarrow{\rho_i^{\text{stat}}:\rho_i^{\text{norm}N}}||_1 \leq 2c \tag{19}$$

Since each element in this vector is in $[-1,1]$, squaring them makes the magnitude no larger, so

$$\mathbb{E}_\mu ||\overrightarrow{\rho_i^{\text{stat}}:\rho_i^{\text{norm}N}}||_2^2 \leq 2c \tag{20}$$

The KL-divergence is larger than the sum of the squares of the probability differences (proven, for example, in (Hutter, 2005, §3.9.2)), so Lemma 3 implies

$$\mathbb{E}_\mu ||\overrightarrow{\rho_i^{\text{norm}}:\mu^N}||_2^2 \leq c + \log c \tag{21}$$

By the triangle inequality,

$$||\overrightarrow{\rho_i^{\text{stat}}:\mu^N}||_2 \leq ||\overrightarrow{\rho_i^{\text{stat}}:\rho_i^{\text{norm}N}}||_2 + ||\overrightarrow{\rho_i^{\text{norm}}:\mu^N}||_2 \tag{22}$$

so

$$||\overrightarrow{\rho_i^{\text{stat}}:\mu^N}||_2^2 \leq ||\overrightarrow{\rho_i^{\text{stat}}:\rho_i^{\text{norm}N}}||_2^2 + ||\overrightarrow{\rho_i^{\text{norm}}:\mu^N}||_2^2 + 2||\overrightarrow{\rho_i^{\text{stat}}:\rho_i^{\text{norm}N}}||_2||\overrightarrow{\rho_i^{\text{norm}}:\mu^N}||_2 \tag{23}$$

and because $\mathbb{E}[XY] \leq \sqrt{\mathbb{E}[X^2]\,\mathbb{E}[Y^2]}$ (the Cauchy-Schwarz Inequality),

$$\mathbb{E}_\mu ||\overrightarrow{\rho_i^{\text{stat}}:\mu^N}||_2^2 \leq 2c + (c+\log c) + 2\sqrt{2c(c+\log c)} < 6c+3 \tag{24}$$

■

We name the measure with the i^{th} largest posterior weight

$$\nu_i^{x_{<t}} :\in \mathcal{M}_i^{x_{<t}} \setminus \mathcal{M}_{i-1}^{x_{<t}} \tag{25}$$

with the posterior weight formally defined $w(\nu|x_{<t}) := \frac{w(\nu)\nu(x_{<t})}{\xi(x_{<t})}$, and $w(\mathcal{M}'|x_{<t}) := \sum_{\nu\in\mathcal{M}'} w(\nu|x_{<t})$. Now, we let

$$\phi_i^{x_{<t}} := \frac{w(\nu_i^{x_{<t}}|x_{<t})}{w(\mathcal{M}_i^{x_{<t}}|x_{<t})} \tag{26}$$

Theorem 5

(i)
$$\mathbb{E}_\mu \sum_{t=0}^{\infty} \sum_{x\in\mathcal{X}} \left[\mu(x|x_{<t}) - \min_{i:\phi_i^{x_{<t}}>\alpha} \nu_i^{x_{<t}}(x|x_{<t})\right]^2 \leq \alpha^{-3}(24w(\mu)^{-1}+12)$$

(ii)
$$\mathbb{E}_\mu \sum_{t=0}^{\infty} \left[1 - \sum_{x\in\mathcal{X}} \min_{i:\phi_i^{x_{<t}}>\alpha} \nu_i^{x_{<t}}(x|x_{<t})\right]^2 \leq |\mathcal{X}|\alpha^{-3}(24w(\mu)^{-1}+12)$$

Proof idea ρ_i^{stat} is a weighted average of $\nu_j^{x_{<t}}$ for $j \leq i$, so convergence results for ρ_i^{stat} and ρ_{i-1}^{stat} are leveraged for $\nu_i^{x_{<t}}$'s convergence. $\phi_i^{x_{<t}} > \alpha$ ensures the weights in the weighted average aren't too small, and that we only need to consider the top $\lfloor 1/\alpha \rfloor$ models. ⬚

6. Faithful Imitation

Now, we use the sequence prediction results to prove our main theorems about the online imitation learner. First, we bound the query probability.

Theorem 6 (Limited Querying)

$$\mathbb{E}_\mu^{\pi^i_\alpha}\left[\sum_{t=0}^{\infty}\theta_q(h_{<t})^3\right] \leq |\mathcal{A}|\alpha^{-3}(24w(\pi^d)^{-1}+12)$$

Proof idea The sort of model mismatch bounded by Theorem 5 (ii) is the basis for the definition of θ_q. Theorem 5 (ii) bounds model mismatch on *observed* data, and data is only observed with probability θ_q, so we go from an in-mean-square bound to a weaker in-mean-cube bound. ⬚

Proof Recall the agent considers a set of possible policies Π that includes the true demonstrator policy π^d, and assigns a strictly positive prior $w(\pi)$ to each policy in Π. Recall P_μ^π is a probability measure over $(\{0,1\}\times\mathcal{A}\times\mathcal{O})^\infty = \mathcal{H}^\infty$. Now we construct a class of measures over $\mathcal{H}^\infty$: let $\mathcal{M} := \{\mathrm{P}_\mu^{\hat{\pi}_\alpha} : \pi\in\Pi\}$ (see the last paragraph of Section 4 for the definition of $\hat{\pi}_\alpha$), and let $w(\mathrm{P}_\mu^{\hat{\pi}_\alpha}) := w(\pi)$. Let $w(\mathrm{P}_\mu^{\hat{\pi}_\alpha}|h_{<t}) :\propto w(\mathrm{P}_\mu^{\hat{\pi}_\alpha})\,\mathrm{P}_\mu^{\hat{\pi}_\alpha}(h_{<t})$. It follows straightforwardly from the definitions of the posterior that $w(\mathrm{P}_\mu^{\hat{\pi}_\alpha}|h_{<t}) = w(\pi|h_{<t})$, $w(\mathrm{P}_\mu^{\hat{\pi}_\alpha}|h_{<t}q_t) = w(\pi|h_{<t}q_t)$, and $w(\mathrm{P}_\mu^{\hat{\pi}_\alpha}|h_{<t}q_ta_t) = w(\pi|h_{<t}q_ta_t)$, since all measures in $\mathcal{M}$ assign the probabilities identically to actions after $q_t=0$, and to observations.

Instead of saying $\mathcal{M}$ contains measures over $\mathcal{X}^\infty$, we generalize slightly, and say that $\mathcal{M}$ contains measures over $\times_{k=0}^\infty\mathcal{X}_k$. For $k\equiv 0 \mod 3$, $\mathcal{X}_k=\{0,1\}$, for $k\equiv 1 \mod 3$, $\mathcal{X}_k=\mathcal{A}$, and for $k\equiv 2 \mod 3$, $\mathcal{X}_k=\mathcal{O}$. With $\nu_i^{x_{<k}}$ and $\phi_i^{x_{<k}}$ as defined before, we can apply Theorem 5 (i) to the class $\mathcal{M}$, after a trivial extension from fixed $\mathcal{X}$ to variable $\mathcal{X}_k$. Checking the definitions is enough to verify that $\{\nu_i^{x_{<k}} : \phi_i^{x_{<k}} > \alpha\}$ is exactly the set $\{\mathrm{P}_\mu^{\hat{\pi}_\alpha} : \pi\in\Pi^\alpha_{h_{<t}}\}$, where $h_j = (q_j,a_j,o_j) = (x_{3j},x_{3j+1},x_{3j+2})$, and $t=\lfloor(k+1)/3\rfloor$. Justifications of the upcoming lettered equations follow below the block. In short, for this $\mathcal{M}$, sequence prediction errors can only come from errors predicting actions after querying, since that's when models differ, so we can use Theorem 5 to bound the latter. Recalling that $\mathrm{P}_\mu^{\pi^i_\alpha}$ is the true probability measure,

$$\begin{aligned}
\alpha^{-3}(24w(\pi^d)^{-1}+12) &= \alpha^{-3}(24w(\mathrm{P}_\mu^{\pi^i_\alpha})^{-1}+12)\\
&\overset{(a)}{\geq} \mathbb{E}_\mu^{\pi^i_\alpha}\sum_{k=0}^{\infty}\sum_{x\in\mathcal{X}_k}\left[\mathrm{P}_\mu^{\pi^i_\alpha}(x|x_{<k}) - \min_{i:\phi_i^{x_{<k}}>\alpha}\nu_i^{x_{<k}}(x|x_{<k})\right]^2\\
&\overset{(b)}{=} \mathbb{E}_\mu^{\pi^i_\alpha}\sum_{t=0}^{\infty}\sum_{q\in\{0,1\}}\left[\mathrm{P}_\mu^{\pi^i_\alpha}(q|h_{<t}) - \min_{\pi\in\Pi^\alpha_{h_{<t}}}\mathrm{P}_\mu^{\hat{\pi}_\alpha}(q|h_{<t})\right]^2 +\\
&\qquad \sum_{a\in\mathcal{A}}\left[\mathrm{P}_\mu^{\pi^i_\alpha}(a|h_{<t}q_t) - \min_{\pi\in\Pi^\alpha_{h_{<t}}}\mathrm{P}_\mu^{\hat{\pi}_\alpha}(a|h_{<t}q_t)\right]^2 +\\
&\qquad \sum_{o\in\mathcal{O}}\left[\mathrm{P}_\mu^{\pi^i_\alpha}(o|h_{<t}q_ta_t) - \min_{\pi\in\Pi^\alpha_{h_{<t}}}\mathrm{P}_\mu^{\hat{\pi}_\alpha}(o|h_{<t}q_ta_t)\right]^2
\end{aligned}$$

$$
\begin{aligned}
&\overset{(c)}{=} \mathbb{E}^{\pi^i_\alpha}_\mu \sum_{t=0}^{\infty} \sum_{a\in\mathcal{A}} \left[\mathrm{P}^{\pi^i_\alpha}_\mu(a|h_{<t}q_t) - \min_{\pi\in\Pi^\alpha_{h_{<t}}} \mathrm{P}^{\hat{\pi}_\alpha}_\mu(a|h_{<t}q_t) \right]^2 \\
&= \mathbb{E}^{\pi^i_\alpha}_\mu \sum_{t=0}^{\infty} \sum_{q\in\{0,1\}} \mathrm{P}^{\pi^i_\alpha}_\mu(q|h_{<t}) \sum_{a\in\mathcal{A}} \left[\mathrm{P}^{\pi^i_\alpha}_\mu(a|h_{<t}q) - \min_{\pi\in\Pi^\alpha_{h_{<t}}} \mathrm{P}^{\hat{\pi}_\alpha}_\mu(a|h_{<t}q) \right]^2 \\
&\overset{(d)}{=} \mathbb{E}^{\pi^i_\alpha}_\mu \sum_{t=0}^{\infty} \mathrm{P}^{\pi^i_\alpha}_\mu(1|h_{<t}) \sum_{a\in\mathcal{A}} \left[\mathrm{P}^{\pi^i_\alpha}_\mu(a|h_{<t}1) - \min_{\pi\in\Pi^\alpha_{h_{<t}}} \mathrm{P}^{\hat{\pi}_\alpha}_\mu(a|h_{<t}1) \right]^2 \\
&= \mathbb{E}^{\pi^i_\alpha}_\mu \sum_{t=0}^{\infty} \theta_q(h_{<t}) |\,\mathcal{A}\,| \,\mathbb{E}_{a\sim\mathbb{U}(\mathcal{A})} \left[\mathrm{P}^{\pi^i_\alpha}_\mu(a|h_{<t}1) - \min_{\pi\in\Pi^\alpha_{h_{<t}}} \mathrm{P}^{\hat{\pi}_\alpha}_\mu(a|h_{<t}1) \right]^2 \\
&\overset{(e)}{\geq} \mathbb{E}^{\pi^i_\alpha}_\mu \sum_{t=0}^{\infty} \theta_q(h_{<t}) |\,\mathcal{A}\,| \left[\mathbb{E}_{a\sim\mathbb{U}(\mathcal{A})} \mathrm{P}^{\pi^i_\alpha}_\mu(a|h_{<t}1) - \min_{\pi\in\Pi^\alpha_{h_{<t}}} \mathrm{P}^{\hat{\pi}_\alpha}_\mu(a|h_{<t}1) \right]^2 \\
&= \mathbb{E}^{\pi^i_\alpha}_\mu \sum_{t=0}^{\infty} \theta_q(h_{<t}) |\,\mathcal{A}\,| \left[|\,\mathcal{A}\,|^{-1} \sum_{a\in\mathcal{A}} \mathrm{P}^{\pi^i_\alpha}_\mu(a|h_{<t}1) - \min_{\pi\in\Pi^\alpha_{h_{<t}}} \mathrm{P}^{\hat{\pi}_\alpha}_\mu(a|h_{<t}1) \right]^2 \\
&= |\,\mathcal{A}\,|^{-1} \mathbb{E}^{\pi^i_\alpha}_\mu \sum_{t=0}^{\infty} \theta_q(h_{<t}) \left[1 - \sum_{a\in\mathcal{A}} \min_{\pi\in\Pi^\alpha_{h_{<t}}} \frac{\hat{\pi}_\alpha(1,a|h_{<t})}{\hat{\pi}_\alpha(1|h_{<t})} \right]^2 \\
&\overset{(f)}{=} |\,\mathcal{A}\,|^{-1} \mathbb{E}^{\pi^i_\alpha}_\mu \sum_{t=0}^{\infty} \theta_q(h_{<t}) \left[1 - \sum_{a\in\mathcal{A}} \min_{\pi\in\Pi^\alpha_{h_{<t}}} \frac{\theta_q(h_{<t})\pi(1,a|h_{<t})}{\theta_q(h_{<t})} \right]^2 \\
&\overset{(g)}{=} |\,\mathcal{A}\,|^{-1} \mathbb{E}^{\pi^i_\alpha}_\mu \sum_{t=0}^{\infty} \theta_q(h_{<t}) \left[\theta_q(h_{<t})\right]^2 \qquad (27)
\end{aligned}
$$

where (a) follows from Theorem 5, (b) groups triples $(x_{3t}, x_{3t+1}, x_{3t+2})$ into h_t, (c) follows because all $\mathrm{P}^{\hat{\pi}_\alpha}_\mu \in \mathcal{M}$ give identical conditional probabilities as $\mathrm{P}^{\pi^i_\alpha}_\mu$ on queries and observations, (d) follows because all $\mathrm{P}^{\hat{\pi}_\alpha}_\mu \in \mathcal{M}$ give identical conditional probabilities as $\mathrm{P}^{\pi^i_\alpha}_\mu$ for actions that follow $q_t = 0$, (e) follows from Jensen's Inequality, (f) follows from the definition of $\hat{\pi}_\alpha$, and (g) follows from the definition of $\theta_q(h_{<t})$. Rearranging Inequality 27 gives the theorem. ■

The following theorem is conditioned on a high-probability event: that the set of top demonstrator-models always contains the truth; at least, the probability is high for small α.

Theorem 7 (Predictive Convergence) $\mathrm{P}^{\pi^i_\alpha}_\mu(\forall t : \pi^d \in \Pi^\alpha_{h_{<t}}) \geq 1 - \alpha w(\pi^d)^{-1}$, *and for* $\alpha < w(\pi^d)$,

$$
\mathbb{E}^{\pi^i_\alpha}_\mu \left[\sum_{t=0}^{\infty} \left(\sum_{a\in\mathcal{A}} \left| \pi^i_\alpha(0,a|h_{<t}) - \pi^d(1,a|h_{<t}) \right| \right)^3 \middle| \forall t : \pi^d \in \Pi^\alpha_{h_{<t}} \right] \leq \frac{|\mathcal{A}|\alpha^{-3}(24w(\pi^d)^{-1} + 12)}{1 - \alpha w(\pi^d)^{-1}}
$$

Note that the prior on the truth $w(\pi^d)$ can practically be made quite large by pre-training with N consecutive demonstrator queries and calling the posterior at that point the new prior.

Proof idea The martingale dynamics of $w(\pi^d|h_{<t})^{-1}$ ensure that $\forall t : \pi^d \in \Pi^\alpha_{h_{<t}}$ with high probability, and conditioned on that event, $\pi^i_\alpha \leq \pi^d$. Thus, $\theta_q \geq$ the ℓ_1 norm between π^d and π^i_α, and we apply Theorem 6. ⬚

We now consider the restriction of probability measures over $\mathcal{H}^\infty$ to $(\mathcal{A} \times \mathcal{O})^\infty$; that is, we marginalize over the query record. For a history $h_{<t} = q_0 a_0 o_0 ... q_{t-1} a_{t-1} o_{t-1}$, let $h^{\setminus}_{<t}$ denote $a_0 o_0 ... a_{t-1} o_{t-1}$. We define the t-step KL-divergence as follows:

$$\underset{t}{\mathrm{KL}}(P||Q) := \sum_{h^{\setminus}_{<t} \in (\mathcal{A} \times \mathcal{O})^t} P(h^{\setminus}_{<t}) \log \frac{P(h^{\setminus}_{<t})}{Q(h^{\setminus}_{<t})} \tag{28}$$

We now define what it is for an environment and demonstrator policy to be "fair".

Definition 8 (Fair) *An environment $\mu : \mathcal{H}^* \times \{0,1\} \times \mathcal{A} \rightsquigarrow \mathcal{O}$ is fair if it does not depend on the query record; that is, $h^{\setminus}_{<t} = \overline{h}^{\setminus}_{<t} \implies \mu(o|h_{<t} q_t a_t) = \mu(o|\overline{h}_{<t} \overline{q}_t a_t)$.*

A demonstrator policy $\pi^d : \mathcal{H}^* \rightsquigarrow \{0,1\} \times \mathcal{A}$ is fair if it does not depend on the query record; that is, $h^{\setminus}_{<t} = \overline{h}^{\setminus}_{<t} \implies \pi^d(qa|h_{<t}) = \pi^d(qa|\overline{h}_{<t})$.

In an unfair environment, it is clearly impossible to expect events to unfold similarly whether the demonstrator or imitator is acting, since the environment treats their actions differently.

Theorem 9 (KL Bound) *Let E be the event that the truth is always in the set of top demonstrator-models: $\forall t : \pi^d \in \Pi^\alpha_{h_{<t}}$. Suppose that μ and π^d are fair. Letting the two probability measures below be restricted to $(\mathcal{A} \times \mathcal{O})^t$ (that is, marginalizing over the query record, and considering only the first t timesteps),*

$$\underset{t}{\mathrm{KL}}\left(\mathrm{P}^{\pi^i_\alpha}_\mu(\cdot|E) \middle\| \mathrm{P}^{\pi^d}_\mu(\cdot|E)\right) \leq \frac{\alpha^{-1}|\mathcal{A}|^{1/3}(24w(\pi^d)^{-1} + 12)^{1/3}}{(1 - \alpha/w(\pi^d))^2} t^{2/3} - \log(1 - \alpha/w(\pi^d))$$

Proof idea The KL divergence telescopes into KL-divergences per timestep, and the latter can be bounded above by a function of θ_q, at which point we use Theorem 6, but a factor of t^2 appears because the finite bound from the theorem applies to θ_q^3 not θ_q. ⬚

We will now construct an upper bound for the probability of an event given the probability of the event if the demonstrator were acting the whole time. This bound is mainly of interest for "bad" events.

Theorem 10 (Preserving Unlikeliness) *Fix t. Let E be the event $\forall k \leq t : \pi^d \in \Pi^\alpha_{h_{<k}}$. Let $B \subset (\mathcal{A} \times \mathcal{O})^t$ be a (bad) event, and extending B to the outcome space $(\{0,1\} \times \mathcal{A} \times \mathcal{O})^t = \mathcal{H}^t$, let $D = B \cap E$. Then, for fair μ and π^d,*

$$\mathrm{P}^{\pi^i_\alpha}_\mu(D) \leq \frac{t^2 s_\alpha}{\left(\log \frac{t^2 s_\alpha}{27\,\mathrm{P}^{\pi^d}_\mu(B)} - 3\log\log\left(1 + \frac{t^{2/3} s_\alpha^{1/3}}{3\,\mathrm{P}^{\pi^d}_\mu(B)^{1/3}}\right)\right)^3}$$

where $s_\alpha = |\,\mathcal{A}\,|\alpha^{-3}(24w(\pi^d)^{-1} + 12)$.

That is, as $\mathrm{P}_\mu^{\pi^d}(B)^{-1} \to \infty$, $\mathrm{P}_\mu^{\pi^i_\alpha}(D)^{-1} \to \infty$ at least polylogarithmicly. If an event would have been very unlikely under the demonstrator's policy, a similar event is unlikely when running the imitator. As an informal note, for most bad worldly events B, it would be quite strange for B to be correlated with E under any policy, which would imply $\mathrm{P}_\mu^{\pi^i_\alpha}(B) = \mathrm{P}_\mu^{\pi^i_\alpha}(D)/(1 - \alpha w(\pi^d)^{-1})$.

Proof idea $\mathrm{P}_\mu^{\pi^i_\alpha}(B \cap E)/\mathrm{P}_\mu^{\pi^d}(B)$ increases by a factor of at most $1+\theta_q$ per timestep. The factor of t^2 appears for the same reason as in Theorem 9. Finally, some complication arises from conditioning Theorem 6 on the event D.

Proof If $\pi^d \in \Pi^\alpha_{h_{<t}}$, then $\pi^i_\alpha(0, a|h_{<t}) \leq \pi^d(a|h_{<t})$, and of course $\pi^i_\alpha(1, a|h_{<t}) = \theta_q(h_{<t})\pi^d(a|h_{<t})$, so $\pi^i_\alpha(a|h_{<t}) \leq (1 + \theta_q(h_{<t}))\pi^d(a|h_{<t})$, that is

$$\frac{\pi^i_\alpha(a|h_{<t})}{\pi^d(a|h_{<t})} \leq \left[1 + \theta_q(h_{<t}) \text{ if } \pi^d \in \Pi^\alpha_{h_{<t}} \text{ else } \infty\right] \tag{29}$$

Thus, for fair μ and π^d, for $h_{<t} \in E$,

$$\frac{\mathrm{P}_\mu^{\pi^i_\alpha}(h^\backslash_{<t})}{\mathrm{P}_\mu^{\pi^d}(h^\backslash_{<t})} \leq \prod_{k=0}^{t-1} 1 + \theta_q(h_{<k}) \tag{30}$$

It follows from Theorem 6 that

$$\mathbb{E}_\mu^{\pi^i_\alpha}\left[\sum_{k=0}^{t-1} \theta_q(h_{<k})^3 \middle| D\right] \leq \frac{s_\alpha}{\mathrm{P}_\mu^{\pi^i_\alpha}(D)} \tag{31}$$

By the same derivation as in Inequality 57, we can thus bound the sum

$$\mathbb{E}_\mu^{\pi^i_\alpha}\left[\sum_{k=0}^{t-1} \theta_q(h_{<k}) \middle| D\right] \leq t^{2/3}\left(\frac{s_\alpha}{\mathrm{P}_\mu^{\pi^i_\alpha}(D)}\right)^{1/3} \tag{32}$$

Now, applying Inequality 29 repeatedly,

$$\begin{aligned}
&\mathbb{E}_\mu^{\pi^i_\alpha}\left[\prod_{k=0}^{t-1}(1 + \theta_q(h_{<k}))^{-1} \middle| D\right] \\
&= \frac{\sum_{h_{<t} \in D} \mathrm{P}_\mu^{\pi^i_\alpha}(h_{<t}) \prod_{k=0}^{t-1}(1 + \theta_q(h_{<k}))^{-1}}{\sum_{h_{<t} \in D} \mathrm{P}_\mu^{\pi^i_\alpha}(h_{<t})} \\
&\leq \frac{\sum_{h_{<t-1} \in E} \sum_{h_{t-1} \in \mathcal{H}: h^\backslash_{<t} \in B} \mathrm{P}_\mu^{\pi^i_\alpha}(h_{<t}) \prod_{k=0}^{t-1}(1 + \theta_q(h_{<k}))^{-1}}{\sum_{h_{<t} \in D} \mathrm{P}_\mu^{\pi^i_\alpha}(h_{<t})} \\
&= \frac{\sum_{h_{<t-1} \in E} \left[\mathrm{P}_\mu^{\pi^i_\alpha}(h_{<t-1}) \prod_{k=0}^{t-2}(1 + \theta_q(h_{<k}))^{-1}\right] \sum_{h^\backslash_{t-1} \in \mathcal{A} \times \mathcal{O}: h^\backslash_{<t} \in B} \mathrm{P}_\mu^{\pi^i_\alpha}(h^\backslash_{t-1}|h_{<t-1})(1 + \theta_q(h_{<t-1}))^{-1}}{\sum_{h_{<t} \in D} \mathrm{P}_\mu^{\pi^i_\alpha}(h_{<t})} \\
&\overset{(a)}{\leq} \frac{\sum_{h_{<t-1} \in E} \left[\mathrm{P}_\mu^{\pi^i_\alpha}(h_{<t-1}) \prod_{k=0}^{t-2}(1 + \theta_q(h_{<k}))^{-1}\right] \sum_{h^\backslash_{t-1} \in \mathcal{A} \times \mathcal{O}: h^\backslash_{<t} \in B} \mathrm{P}_\mu^{\pi^d}(h^\backslash_{t-1}|h_{<t-1})}{\sum_{h_{<t} \in D} \mathrm{P}_\mu^{\pi^i_\alpha}(h_{<t})}
\end{aligned}$$

$$\overset{(b)}{\leq} \frac{\sum_{h_{<t-2}\in E}\left[\mathrm{P}_{\mu}^{\pi_\alpha^i}(h_{<t-2})\prod_{k=0}^{t-3}(1+\theta_q(h_{<k}))^{-1}\right]\sum_{h^{\setminus}_{t-2}h^{\setminus}_{t-1}\in(\mathcal{A}\times\mathcal{O})^2 : h^{\setminus}_{<t}\in B}\mathrm{P}_{\mu}^{\pi^d}(h^{\setminus}_{t-2}h^{\setminus}_{t-1}|h_{<t-2})}{\sum_{h_{<t}\in D}\mathrm{P}_{\mu}^{\pi_\alpha^i}(h_{<t})}$$

$$\overset{(c)}{\leq} \frac{\sum_{h^{\setminus}_{<t}\in B}\mathrm{P}_{\mu}^{\pi^d}(h^{\setminus}_{<t})}{\sum_{h_{<t}\in D}\mathrm{P}_{\mu}^{\pi_\alpha^i}(h_{<t})} = \frac{\mathrm{P}_{\mu}^{\pi^d}(B)}{\mathrm{P}_{\mu}^{\pi_\alpha^i}(D)} \tag{33}$$

where (a) follows from Inequality 29 since $h_{<t-1} \in E$ (note the change from π_α^i to π^d), (b) iterates the previous three lines, and (c) iterates the logic down to 0.

Now we bound the expectation

$$\begin{aligned}
\mathbb{E}_{\mu}^{\pi_\alpha^i}\left[\prod_{k=0}^{t-1}(1+\theta_q(h_{<k}))^{-1}\middle| D\right] &\overset{(a)}{\geq} \prod_{k=0}^{t-1}(1+\mathbb{E}_{\mu}^{\pi_\alpha^i}\left[\theta_q(h_{<k})|D\right])^{-1} \\
&= \exp\left(-\sum_{k=0}^{t-1}\log\left(1+\mathbb{E}_{\mu}^{\pi_\alpha^i}\left[\theta_q(h_{<k})|D\right]\right)\right) \\
&\geq \exp\left(-\sum_{k=0}^{t-1}\mathbb{E}_{\mu}^{\pi_\alpha^i}\left[\theta_q(h_{<k})|D\right]\right) \\
&\overset{(b)}{\geq} e^{-t^{2/3}s_\alpha^{1/3}\mathrm{P}_{\mu}^{\pi_\alpha^i}(D)^{-1/3}}
\end{aligned} \tag{34}$$

where (a) follows from Jensen's Inequality (one can easily show the Hessian of $\prod_i 1/(1+x_i)$ is positive semidefinite for $x \succ 0$), and (b) follows from Inequality 32. Solving for $\mathrm{P}_{\mu}^{\pi_\alpha^i}(D)$ in terms of $\mathrm{P}_{\mu}^{\pi^d}(B)$, we get

$$\mathrm{P}_{\mu}^{\pi_\alpha^i}(D) \leq \frac{t^2 s_\alpha}{27W(\frac{t^{2/3}s_\alpha^{1/3}}{3\,\mathrm{P}_{\mu}^{\pi^d}(B)^{1/3}})^3} \tag{35}$$

where W is the Lambert-W function, defined by the property $W(z)e^{W(z)} = z$. A property of the Lambert-W function—that $W(z) \geq \log z - \log\log(1+z)$—yields the theorem:

$$\mathrm{P}_{\mu}^{\pi_\alpha^i}(D) \leq \frac{t^2 s_\alpha}{\left(\log\frac{t^2 s_\alpha}{27\,\mathrm{P}_{\mu}^{\pi^d}(B)} - 3\log\log\left(1+\frac{t^{2/3}s_\alpha^{1/3}}{3\,\mathrm{P}_{\mu}^{\pi^d}(B)^{1/3}}\right)\right)^3}$$

One can easily verify this inequality by supposing the opposite and showing that it violates Inequality 34, but we omit this.

■

Existing work on imitation learners attempts to be robust to a bounded loss function. In the real world, however, to quote Theon Greyjoy, "It can always be worse". But some bounds on badness are possible: we tolerate one-in-ten-chance events; they happen, and we get on with it. One-in-a-hundred-chance events can be meaningfully worse. But in a world largely governed by humans, we keep most truly devastating events below even a 1% chance. It's hard to apply similar bounds to the badness of one-in-a-billion-chance events, and in general, as the probability gets smaller, a loss

function should countenance steadily larger losses. When an event goes from a 1% to a 2% chance, we should be much less concerned than if it went from 10^{-9} to 1%. In the extreme, if an event has probability 0 under a demonstrator's policy, there might be an arbitrarily good reason for that. Whereas the bounded loss functions of all existing work ignore this effect, our Theorem 10 does not.

7. Conclusion

We present the first formal results for an imitation learner in a setting where the environment does not reset. We present the first formal results for an imitation learner that do not depend on a bounded loss assumption. We present the first finite error bounds for an agent acting in general environments; existing results only regard limiting behavior (although existing work considers reinforcement learning, a harder problem than imitation learning). If we would like to have an artificial agent imitate, with particular concern for keeping unlikely events unlikely, this is the first theory of how to do it.

References

Daniel S Brown, Yuchen Cui, and Scott Niekum. Risk-aware active inverse reinforcement learning. In *Conference on Robot Learning*, pages 362–372. PMLR, 2018.

Nicolò Cesa-Bianchi, Alex Conconi, and Claudio Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004.

M. Cohen, B. Vellambi, and M. Hutter. Asymptotically unambitious artificial general intelligence. In *Proc. 34rd AAAI Conference on Artificial Intelligence (AAAI'20)*, volume 34, New York, USA, 2020. AAAI Press.

Michael K Cohen and Marcus Hutter. Pessimism about unknown unknowns inspires conservatism. In *Conference on Learning Theory*, pages 1344–1373. PMLR, 2020.

R Durrett. Probability: Theory and examples. cambridge university press, 2010.

E Mark Gold. Language identification in the limit. *Information and control*, 10(5):447–474, 1967.

Ahmed Hussein, Mohamed Medhat Gaber, Eyad Elyan, and Chrisina Jayne. Imitation learning: A survey of learning methods. *ACM Computing Surveys (CSUR)*, 50(2):1–35, 2017.

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

Kshitij Judah, Alan P Fern, Thomas G Dietterich, and Prasad Tadepalli. Active imitation learning: Formal and practical reductions to iid learning. *Journal of Machine Learning Research*, 15(120): 4105–4143, 2014.

Jan Poland and Marcus Hutter. Asymptotics of discrete MDL for online prediction. *IEEE Transactions on Information Theory*, 51(11):3780–3795, 2005. ISSN 0018-9448. doi: 10.1109/TIT.2005.856956.

Stéphane Ross and Drew Bagnell. Efficient reductions for imitation learning. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 661–668, 2010.

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.

Umar Syed and Robert E Schapire. A reduction from apprenticeship learning to classification. *Advances in neural information processing systems*, 23:2253–2261, 2010.

Appendix A. Notation and Definitions

Notation	Meaning
Preliminary Notation	
$\mathcal{A}$, $\mathcal{O}$	the finite action/observation spaces
a_t, o_t	$\in \mathcal{A}, \mathcal{O}$; the action and observation at timestep t
q_t	$\in \{0,1\}$; indicates whether the demonstrator is queried at time t
$\mathcal{H}$	$\{0,1\} \times \mathcal{A} \times \mathcal{O}$
h_t	(q_t, a_t, o_t); the interaction history in the t^{th} timestep
$h_{<t}$	$(h_1, ..., h_{t-1})$
$h_t^{\setminus}$	(a_t, o_t)
ϵ	the empty history
π	policy stochastically mapping $\mathcal{H}^* \rightsquigarrow \{0,1\} \times \mathcal{A}$
μ	environment stochastically mapping $\mathcal{H}^* \times \{0,1\} \times \mathcal{A} \rightsquigarrow \mathcal{O}$
P_ν^π	a probability measure over histories with actions sampled from π and observations sampled from ν
$\mathbb{E}_\nu^\pi$	the expectation when the interaction history is sampled from P_ν^π
$w(\pi)$	(positive) prior weight that the policy π is the demonstrator's
$w(\pi \mid h_{<t})$	posterior weight on the policy π; $\propto w(\pi) \prod_{k<t:q_k=1} \pi(q_k a_k \mid h_{<k})$
Imitation Learner Definition	
α	$\in (0,1]$; lower values mean the imitator better resembles the demonstrator, but queries longer
$\Pi^\alpha_{h_{<t}}$	set of top models; $\{\pi_n^{h_{<t}} \in \Pi : w(\pi_n^{h_{<t}} \mid h_{<t}) \geq \alpha \sum_{m \leq n} w(\pi_m^{h_{<t}} \mid h_{<t})\}$
π^d	the demonstrator's policy
π^i_α	the imitator's policy; $\pi^i_\alpha(0, a \mid h_{<t}) = \min_{\pi' \in \Pi^\alpha_{h_{<t}}} \pi'(1, a \mid h_{<t})$, and $\pi^i_\alpha(1, a \mid h_{<t}) = \theta_q(h_{<t}) \pi^d(1, a \mid h_{<t})$
$\theta_q(h_{<t})$	the query probability; $1 - \sum_{a \in \mathcal{A}} \pi^i_\alpha(0, a \mid h_{<t})$
$\hat{\pi}_\alpha$	the imitator policy defined with respect to an arbitrary demonstrator π, not the real demonstrator π^d
General Sequence Prediction	
$\mathcal{X}$	finite alphabet
$x_{<t}$	an element of $\mathcal{X}^t$
$\mathcal{M}$	countable set of measures over $\mathcal{X}^\infty$
$w(\nu)$	prior weight on $\nu \in \mathcal{M}$
$w(\nu \mid x_{<t})$	posterior weight on $\nu \in \mathcal{M}$
ξ	$\xi(x_{<t}) = \sum_{\nu \in \mathcal{M}} w(\nu) \nu(x_{<t})$
ρ_i	$\rho_i(x_{<t}) = \max_{\mathcal{M}' \subset \mathcal{M} : \lvert\mathcal{M}'\rvert = i} \sum_{\nu \in \mathcal{M}'} w(\nu) \nu(x_{<t})$
ρ_i^{norm}	like ρ_i, but normalized to be a measure $\rho_i^{\text{norm}}(x \mid x_{<t}) = \rho_i(x \mid x_{<t}) / \sum_{x' \in \mathcal{X}} \rho_i(x' \mid x_{<t})$
$\mathcal{M}_i^{x_{<t}}$	$\operatorname{argmax}_{\mathcal{M}' \subset \mathcal{M} : \lvert\mathcal{M}'\rvert = i} \sum_{\nu \in \mathcal{M}} w(\nu) \nu(x_{<t})$
ρ_i^{stat}	a mixture over the top i models, sorted by posterior weight $\rho_i^{\text{stat}}(x \mid x_{<t}) = \sum_{\nu \in \mathcal{M}_i^{x_{<t}}} w(\nu) \nu(x_{<t} x) / \sum_{\nu \in \mathcal{M}_i^{x_{<t}}} w(\nu) \nu(x_{<t})$
$\phi_i^{x_{<t}}$	$w(\nu_i^{x_{<t}} \mid x_{<t}) / w(\mathcal{M}_i^{x_{<t}} \mid x_{<t})$

Appendix B. Omitted Proofs

Lemma 1

$$0 \leq \mathbb{E}_\mu \sum_{t=0}^{\infty} \frac{\sum_{x\in\mathcal{X}} \rho_i(x_{<t}x)}{\rho_i(x_{<t})} - 1 \leq w(\mu)^{-1}$$

Proof All terms in the sum are non-negative, by Inequality 7. Recall ϵ denotes the empty string—the element of $\mathcal{X}^0$.

$$\begin{aligned}
&\mathbb{E}_\mu \sum_{t=0}^{N-1} \frac{\sum_{x\in\mathcal{X}} \rho_i(x_{<t}x)}{\rho_i(x_{<t})} - 1 \\
&= \sum_{t=0}^{N-1} \sum_{x_{<t}\in\mathcal{X}^t} \mu(x_{<t}) \frac{\sum_{x\in\mathcal{X}} \rho_i(x_{<t}x) - \rho_i(x_{<t})}{\rho_i(x_{<t})} \\
&\overset{(a)}{\leq} \sum_{t=0}^{N-1} \sum_{x_{<t}\in\mathcal{X}^t} w(\mu)^{-1} \left[\sum_{x\in\mathcal{X}} \rho_i(x_{<t}x) - \rho_i(x_{<t})\right] \\
&\overset{(b)}{=} w(\mu)^{-1} \left[\sum_{x_{<N}\in\mathcal{X}^N} \rho_i(x_{<N}) - \rho_i(\epsilon)\right] \\
&\overset{(c)}{\leq} w(\mu)^{-1} \sum_{x_{<N}\in\mathcal{X}^N} \xi(x_{<N}) = w(\mu)^{-1}
\end{aligned} \tag{36}$$

where (a) follows from Inequality 12, (b) cancels terms that are added then subtracted, and (c) follows from Inequality 11. ■

Lemma 3 *Recalling $\nu(\cdot|x_{<t})$ is a measure over $\mathcal{X}$,*

$$\mathbb{E}_\mu \sum_{t=0}^{\infty} \mathrm{KL}\left(\mu(\cdot|x_{<t}) \middle\| \rho_i^{\mathrm{norm}}(\cdot|x_{<t})\right) \leq w(\mu)^{-1} + \log w(\mu)^{-1}$$

Proof The KL-divergence is non-negative, so we bound an arbitrary finite sum.

$$\begin{aligned}
&\mathbb{E}_\mu \sum_{t=0}^{N-1} \mathrm{KL}\left(\mu(\cdot|x_{<t}) \middle\| \rho_i^{\mathrm{norm}}(\cdot|x_{<t})\right) \\
&= \sum_{t=0}^{N-1} \mathbb{E}_\mu \sum_{x_t\in\mathcal{X}} \mu(x_t|x_{<t}) \log \frac{\mu(x_t|x_{<t})}{\rho_i^{\mathrm{norm}}(x_t|x_{<t})} \\
&\overset{(a)}{=} \sum_{t=0}^{N-1} \sum_{x_{<t}\in\mathcal{X}^t} \mu(x_{<t}) \sum_{x_t\in\mathcal{X}} \mu(x_t|x_{<t}) \left[\log \frac{\mu(x_t|x_{<t})}{\rho_i(x_t|x_{<t})} + \log \frac{\sum_{x'\in\mathcal{X}} \rho_i(x_{<t}x')}{\rho_i(x_{<t})}\right] \\
&\overset{(b)}{\leq} w(\mu)^{-1} + \sum_{t=0}^{N-1} \sum_{x_{<t}\in\mathcal{X}^t} \mu(x_{<t}) \sum_{x_t\in\mathcal{X}} \mu(x_t|x_{<t}) \log \frac{\mu(x_t|x_{<t})}{\rho_i(x_t|x_{<t})}
\end{aligned}$$

$$
\begin{aligned}
&= w(\mu)^{-1} + \sum_{t=0}^{N-1} \sum_{x_{<t} \in \mathcal{X}^t} \mu(x_{<t}) \sum_{x_t \in \mathcal{X}} \mu(x_t|x_{<t}) \left[\log \frac{\mu(x_{<t}x_t)}{\rho_i(x_{<t}x_t)} - \log \frac{\mu(x_{<t})}{\rho_i(x_{<t})} \right] \\
&= w(\mu)^{-1} + \sum_{t=0}^{N-1} \sum_{x_{<t} \in \mathcal{X}^t} \mu(x_{<t}) \left[\sum_{x_t \in \mathcal{X}} \mu(x_t|x_{<t}) \log \frac{\mu(x_{<t}x_t)}{\rho_i(x_{<t}x_t)} - \log \frac{\mu(x_{<t})}{\rho_i(x_{<t})} \right] \\
&= w(\mu)^{-1} + \sum_{t=0}^{N-1} \left[\sum_{x_{\leq t} \in \mathcal{X}^{t+1}} \mu(x_{\leq t}) \log \frac{\mu(x_{\leq t})}{\rho_i(x_{\leq t})} - \sum_{x_{<t} \in \mathcal{X}^t} \mu(x_{<t}) \log \frac{\mu(x_{<t})}{\rho_i(x_{<t})} \right] \\
&\overset{(c)}{=} w(\mu)^{-1} + \sum_{x_{<N} \in \mathcal{X}^N} \mu(x_{<N}) \log \frac{\mu(x_{<N})}{\rho_i(x_{<N})} - \mu(\epsilon) \log \frac{\mu(\epsilon)}{\rho_i(\epsilon)} \\
&\overset{(d)}{\leq} w(\mu)^{-1} + \sum_{x_{<N} \in \mathcal{X}^N} \mu(x_{<N}) \log w(\mu)^{-1} = w(\mu)^{-1} + \log w(\mu)^{-1}
\end{aligned} \tag{37}
$$

where (a) follows from the definition of ρ_i^{norm} in Equation 8, (b) follows from Lemma 1 and the fact that $\log x \leq x - 1$, (c) cancels like terms, and (d) follows from Inequality 12. ■

Theorem 5

(i) $$\mathbb{E}_\mu \sum_{t=0}^{\infty} \sum_{x \in \mathcal{X}} \left[\mu(x|x_{<t}) - \min_{i : \phi_i^{x_{<t}} > \alpha} \nu_i^{x_{<t}}(x|x_{<t}) \right]^2 \leq \alpha^{-3}(24w(\mu)^{-1} + 12)$$

(ii) $$\mathbb{E}_\mu \sum_{t=0}^{\infty} \left[1 - \sum_{x \in \mathcal{X}} \min_{i : \phi_i^{x_{<t}} > \alpha} \nu_i^{x_{<t}}(x|x_{<t}) \right]^2 \leq |\mathcal{X}|\alpha^{-3}(24w(\mu)^{-1} + 12)$$

Proof $\rho_i^{\text{stat}}(x|x_{<t})$ is a weighted average of $\nu_j^{x_{<t}}(x|x_{<t})$ for $j \leq i$:

$$
\begin{aligned}
\rho_i^{\text{stat}}(x|x_{<t}) &= \frac{\sum_{\nu \in \mathcal{M}_i^{x_{<t}}} w(\nu)\nu(x_{<t}x)}{\sum_{\nu \in \mathcal{M}_i^{x_{<t}}} w(\nu)\nu(x_{<t})} \\
&= \frac{\sum_{\nu \in \mathcal{M}_i^{x_{<t}}} w(\nu)\nu(x_{<t})\nu(x|x_{<t})}{\sum_{\nu \in \mathcal{M}_i^{x_{<t}}} w(\nu)\nu(x_{<t})} \\
&= \frac{\sum_{\nu \in \mathcal{M}_i^{x_{<t}}} w(\nu|x_{<t})\xi(x_{<t})\nu(x|x_{<t})}{\sum_{\nu \in \mathcal{M}_i^{x_{<t}}} w(\nu|x_{<t})\xi(x_{<t})} \\
&= \sum_{\nu \in \mathcal{M}_i^{x_{<t}}} \frac{w(\nu|x_{<t})}{w(\mathcal{M}_i^{x_{<t}}|x_{<t})} \nu(x|x_{<t}) \\
&= \sum_{j=1}^{i} \frac{w(\nu_j^{x_{<t}}|x_{<t})}{w(\mathcal{M}_i^{x_{<t}}|x_{<t})} \nu_j^{x_{<t}}(x|x_{<t})
\end{aligned} \tag{38}
$$

Trivially,

$$\nu_1^{x_{<t}}(x|x_{<t}) = \rho_1^{\text{stat}}(x|x_{<t}) \tag{39}$$

but for $i > 1$, we would like to express $\nu_i^{x_{<t}}$ in terms of ρ_i^{stat} and ρ_{i-1}^{stat}:

$$\rho_i^{\text{stat}}(x|x_{<t}) = \frac{w(\mathcal{M}_{i-1}^{x_{<t}}|x_{<t})}{w(\mathcal{M}_i^{x_{<t}}|x_{<t})}\rho_{i-1}^{\text{stat}}(x|x_{<t}) + \frac{w(\nu_i^{x_{<t}}|x_{<t})}{w(\mathcal{M}_i^{x_{<t}}|x_{<t})}\nu_i^{x_{<t}}(x|x_{<t}) \tag{40}$$

Thus,

$$\nu_i^{x_{<t}}(x|x_{<t}) = \frac{w(\mathcal{M}_i^{x_{<t}}|x_{<t})}{w(\nu_i^{x_{<t}}|x_{<t})}\rho_i^{\text{stat}}(x|x_{<t}) - \frac{w(\mathcal{M}_{i-1}^{x_{<t}}|x_{<t})}{w(\nu_i^{x_{<t}}|x_{<t})}\rho_{i-1}^{\text{stat}}(x|x_{<t}) \tag{41}$$

Since $\frac{w(\mathcal{M}_i^{x_{<t}}|x_{<t})}{w(\nu_i^{x_{<t}}|x_{<t})} - \frac{w(\mathcal{M}_{i-1}^{x_{<t}}|x_{<t})}{w(\nu_i^{x_{<t}}|x_{<t})} = 1$, for $i > 1$,

$$\nu_i^{x_{<t}}(x|x_{<t}) - \mu(x|x_{<t}) = \frac{w(\mathcal{M}_i^{x_{<t}}|x_{<t})}{w(\nu_i^{x_{<t}}|x_{<t})}\left[\rho_i^{\text{stat}}(x|x_{<t}) - \mu(x|x_{<t})\right] - \frac{w(\mathcal{M}_{i-1}^{x_{<t}}|x_{<t})}{w(\nu_i^{x_{<t}}|x_{<t})}\left[\rho_{i-1}^{\text{stat}}(x|x_{<t}) - \mu(x|x_{<t})\right] \tag{42}$$

Recall

$$\phi_i^{x_{<t}} := \frac{w(\nu_i^{x_{<t}}|x_{<t})}{w(\mathcal{M}_i^{x_{<t}}|x_{<t})}$$

Since $w(\mathcal{M}_{i-1}^{x_{<t}}|x_{<t}) \leq w(\mathcal{M}_i^{x_{<t}}|x_{<t})$, we have

$$(\phi_i^{x_{<t}})^2\left[\nu_i^{x_{<t}}(x|x_{<t}) - \mu(x|x_{<t})\right]^2 \leq 2\left[\rho_i^{\text{stat}}(x|x_{<t}) - \mu(x|x_{<t})\right]^2 + 2\left[\rho_{i-1}^{\text{stat}}(x|x_{<t}) - \mu(x|x_{<t})\right]^2 \tag{43}$$

Now we consider all measures $\nu_i^{x_{<t}}$ for which $\phi_i^{x_{<t}} > \alpha$.

$$\mathbb{E}_\mu \sum_{t=0}^{N-1} \sum_{i:\phi_i^{x_{<t}}>\alpha} \sum_{x\in\mathcal{X}} \left[\nu_i^{x_{<t}}(x|x_{<t}) - \mu(x|x_{<t})\right]^2 \leq 2\alpha^{-2}\,\mathbb{E}_\mu \sum_{t=0}^{N-1} \sum_{i:\phi_i^{x_{<t}}>\alpha} \sum_{x\in\mathcal{X}} \left[\rho_i^{\text{stat}}(x|x_{<t}) - \mu(x|x_{<t})\right]^2 + \left[\rho_{i-1}^{\text{stat}}(x|x_{<t}) - \mu(x|x_{<t})\right]^2 \tag{44}$$

Now we note that $\{i : \phi_i^{x_{<t}} > \alpha\} \subset \{i : i < \alpha^{-1}\}$, since $w(\nu_i^{x_{<t}}|x_{<t}) \leq w(\nu_j^{x_{<t}}|x_{<t})$ for $i > j$. Thus,

$$\begin{aligned}
&\mathbb{E}_\mu \sum_{t=0}^{N-1} \sum_{i:\phi_i^{x_{<t}}>\alpha} \sum_{x\in\mathcal{X}} \left[\nu_i^{x_{<t}}(x|x_{<t}) - \mu(x|x_{<t})\right]^2 \\
\leq& 2\alpha^{-2}\,\mathbb{E}_\mu \sum_{t=0}^{N-1} \sum_{i:i<\alpha^{-1}} \sum_{x\in\mathcal{X}} \left[\rho_i^{\text{stat}}(x|x_{<t}) - \mu(x|x_{<t})\right]^2 + \left[\rho_{i-1}^{\text{stat}}(x|x_{<t}) - \mu(x|x_{<t})\right]^2 \\
=& 2\alpha^{-2} \sum_{i:i<\alpha^{-1}} \mathbb{E}_\mu \sum_{t=0}^{N-1} \sum_{x\in\mathcal{X}} \left[\rho_i^{\text{stat}}(x|x_{<t}) - \mu(x|x_{<t})\right]^2 + \left[\rho_{i-1}^{\text{stat}}(x|x_{<t}) - \mu(x|x_{<t})\right]^2 \\
\leq& 2\alpha^{-2} \sum_{i:i<\alpha^{-1}} 2(6w(\mu)^{-1} + 3) \leq \alpha^{-3}(24w(\mu)^{-1} + 12)
\end{aligned} \tag{45}$$

Considering only a subset of these conditional-probability-errors,

$$\mathbb{E}_\mu \sum_{t=0}^{N-1} \sum_{x\in\mathcal{X}} \left[\mu(x|x_{<t}) - \min_{i:\phi_i^{x<t}>\alpha} \nu_i^{x<t}(x|x_{<t})\right]^2 \leq$$
$$\mathbb{E}_\mu \sum_{t=0}^{N-1} \sum_{i:\phi_i^{x<t}>\alpha} \sum_{x\in\mathcal{X}} \left[\nu_i^{x<t}(x|x_{<t}) - \mu(x|x_{<t})\right]^2 \leq \alpha^{-3}(24w(\mu)^{-1} + 12) \quad (46)$$

This completes the proof of (i). Finally, with $\mathbb{U}$ being the uniform distribution,

$$\begin{aligned}
&\mathbb{E}_\mu \sum_{t=0}^{N-1} \left[1 - \sum_{x\in\mathcal{X}} \min_{i:\phi_i^{x<t}>\alpha} \nu_i^{x<t}(x|x_{<t})\right]^2 \\
=&\mathbb{E}_\mu \sum_{t=0}^{N-1} \left[\sum_{x\in\mathcal{X}} \mu(x|x_{<t}) - \min_{i:\phi_i^{x<t}>\alpha} \nu_i^{x<t}(x|x_{<t})\right]^2 \\
=&\mathbb{E}_\mu \sum_{t=0}^{N-1} \left[|\mathcal{X}|\,\mathbb{E}_{x\sim\mathbb{U}(\mathcal{X})}\,\mu(x|x_{<t}) - \min_{i:\phi_i^{x<t}>\alpha} \nu_i^{x<t}(x|x_{<t})\right]^2 \\
\overset{(a)}{\leq}&|\mathcal{X}|^2\,\mathbb{E}_\mu \sum_{t=0}^{N-1} \mathbb{E}_{x\sim\mathbb{U}(\mathcal{X})} \left[\mu(x|x_{<t}) - \min_{i:\phi_i^{x<t}>\alpha} \nu_i^{x<t}(x|x_{<t})\right]^2 \\
=&|\mathcal{X}|\,\mathbb{E}_\mu \sum_{t=0}^{N-1} \sum_{x\in\mathcal{X}} \left[\mu(x|x_{<t}) - \min_{i:\phi_i^{x<t}>\alpha} \nu_i^{x<t}(x|x_{<t})\right]^2 \\
\overset{(b)}{\leq}&|\mathcal{X}|\alpha^{-3}(24w(\mu)^{-1} + 12)
\end{aligned} \quad (47)$$

where (a) follows from Jensen's Inequality, and (b) follows from Theorem 5 (i), which completes the proof of (ii). ■

Theorem 7 (Predictive Convergence) $\mathrm{P}_\mu^{\pi_\alpha^i}(\forall t : \pi^d \in \Pi_{h_{<t}}^\alpha) \geq 1 - \alpha w(\pi^d)^{-1}$, *and for* $\alpha < w(\pi^d)$,

$$\mathbb{E}_\mu^{\pi_\alpha^i}\left[\sum_{t=0}^{\infty}\left(\sum_{a\in\mathcal{A}}\left|\pi_\alpha^i(0,a|h_{<t}) - \pi^d(1,a|h_{<t})\right|\right)^3 \middle| \forall t : \pi^d \in \Pi_{h_{<t}}^\alpha\right] \leq \frac{|\mathcal{A}|\alpha^{-3}(24w(\pi^d)^{-1} + 12)}{1 - \alpha w(\pi^d)^{-1}}$$

Proof First we show $\mathrm{P}_\mu^{\pi_\alpha^i}(\forall t : \pi^d \in \Pi_{h_{<t}}^\alpha) \geq 1 - \alpha w(\pi^d)^{-1}$. Since, $w(\pi^d|h_{<t}) > \alpha \implies \pi^d \in \Pi_{h_{<t}}^\alpha$, we show $\mathrm{P}_\mu^{\pi_\alpha^i}(\forall t : w(\pi^d|h_{<t}) > \alpha) \geq 1 - \alpha w(\pi^d)^{-1}$. $w(\pi^d|h_{<t})^{-1}$ is a non-negative $\mathrm{P}_\mu^{\pi_\alpha^i}$-martingale (see for example (Cohen et al., 2020, proof of Lem. 3) for a proof that the inverse of the posterior on the truth is a martingale). Thus, by Doob's martingale inequality (Durrett, 2010, Thm. 5.4.2),

$$\mathrm{P}_\mu^{\pi_\alpha^i}(\exists t : w(\pi^d|h_{<t})^{-1} \geq \alpha^{-1}) \leq \alpha w(\pi^d)^{-1} \quad (48)$$

so

$$\mathrm{P}_{\mu}^{\pi_\alpha^i}(\forall t : w(\pi^d|h_{<t}) > \alpha) \geq 1 - \alpha w(\pi^d)^{-1} \tag{49}$$

which implies

$$\mathrm{P}_{\mu}^{\pi_\alpha^i}(\forall t : \pi^d \in \Pi_{h_{<t}}^{\alpha}) \geq 1 - \alpha w(\pi^d)^{-1} \tag{50}$$

Recall $\pi_\alpha^i(0, a|h_{<t}) = \min_{\pi \in \Pi_{h_{<t}}^{\alpha}} \pi(1, a|h_{<t})$, so if $\pi^d \in \Pi_{h_{<t}}^{\alpha}$, then $\pi_\alpha^i(0, a|h_{<t}) \leq \pi^d(1, a|h_{<t})$. Thus, in that case,

$$\sum_{a \in \mathcal{A}} \left|\pi_\alpha^i(0, a|h_{<t}) - \pi^d(1, a|h_{<t})\right| = \sum_{a \in \mathcal{A}} \pi^d(1, a|h_{<t}) - \pi_\alpha^i(0, a|h_{<t}) \leq$$
$$1 - \sum_{a \in \mathcal{A}} \pi_\alpha^i(0, a|h_{<t}) = \theta_q(h_{<t}) \tag{51}$$

The rest follows trivially:

$$\begin{aligned}
&\mathbb{E}_{\mu}^{\pi_\alpha^i}\left[\sum_{t=0}^{\infty}\left(\sum_{a \in \mathcal{A}}\left|\pi_\alpha^i(0, a|h_{<t}) - \pi^d(1, a|h_{<t})\right|\right)^3 \middle| \forall t : \pi^d \in \Pi_{h_{<t}}^{\alpha}\right] \\
\leq &\mathbb{E}_{\mu}^{\pi_\alpha^i}\left[\sum_{t=0}^{\infty}\theta_q(h_{<t})^3 \middle| \forall t : \pi^d \in \Pi_{h_{<t}}^{\alpha}\right] \\
\overset{(a)}{\leq} &\mathbb{E}_{\mu}^{\pi_\alpha^i}\left[\sum_{t=0}^{\infty}\theta_q(h_{<t})^3\right] \Big/ \mathrm{P}_{\mu}^{\pi_\alpha^i}(\forall t : \pi^d \in \Pi_{h_{<t}}^{\alpha}) \\
\overset{(b)}{\leq} &\frac{|\mathcal{A}|\alpha^{-3}(24w(\pi^d)^{-1} + 12)}{1 - \alpha w(\pi^d)^{-1}}
\end{aligned} \tag{52}$$

where (a) follows because θ_q is non-negative, and (b) follows from Equation 50 and Theorem 6 (as long as $\alpha < w(\pi^d)$).

■

Lemma 11 *For $a \in \mathcal{A}$, let $0 \leq i_a \leq d_a$, and let $\sum_{a \in \mathcal{A}} d_a = 1$. Let $\theta_q = 1 - \sum_{a \in \mathcal{A}} i_a$. Then,*

$$\Delta := \sum_{a \in \mathcal{A}} (i_a + \theta_q d_a) \log \frac{i_a + \theta_q d_a}{d_a} \leq \theta_q$$

Proof

$$\sum_{a \in \mathcal{A}} (i_a + \theta_q d_a) \log \frac{i_a + \theta_q d_a}{d_a} = \sum_{a \in \mathcal{A}} (i_a + \theta_q d_a) \log(\frac{i_a}{d_a} + \theta_q) \leq$$
$$\left(\sum_{a \in \mathcal{A}} i_a + \theta_q \sum_{a \in \mathcal{A}} d_a\right) \log(1 + \theta_q) = (1 - \theta_q + \theta_q) \log(1 + \theta_q) \leq \theta_q \tag{53}$$

■

Theorem 9 (KL Bound) *Let E be the event that the truth is always in the set of top demonstrator-models: $\forall t : \pi^d \in \Pi^\alpha_{h_{<t}}$. Suppose that μ and π^d are fair. Letting the two probability measures below be restricted to $(\mathcal{A} \times \mathcal{O})^t$ (that is, marginalizing over the query record, and considering only the first t timesteps),*

$$\operatorname*{KL}_t\left(\mathrm{P}^{\pi^i_\alpha}_\mu(\cdot|E)\middle\|\mathrm{P}^{\pi^d}_\mu(\cdot|E)\right) \leq \frac{\alpha^{-1}|\mathcal{A}|^{1/3}(24w(\pi^d)^{-1}+12)^{1/3}}{(1-\alpha/w(\pi^d))^2}t^{2/3} - \log(1-\alpha/w(\pi^d))$$

Proof We begin by restricting attention to a particular timestep t. Recall $\pi^i_\alpha(0,a|h_{<t}) = \min_{\pi' \in \Pi^\alpha_{h_{<t}}} \pi'(1,a|h_{<t})$. We abbreviate this quantity i_a. We also let d_a denote $\pi^d(1,a|h_{<t})$. Note that when $\pi^d \in \Pi^\alpha_{h_{<t}}$,

$$i_a \leq d_a \tag{54}$$

Recall that the query probability $\theta_q = 1 - \sum_{a\in\mathcal{A}} i_a$, and the marginalized probability $\pi^i_\alpha(a|h_{<t}) = i_a + \theta_q d_a$. Assuming $h_{<k}$ satisfies E, let

$$\Delta_k := \operatorname*{KL}_1\left(\pi^i_\alpha(\cdot|h_{<k})\middle\|\pi^d(\cdot|h_{<k})\right) = \sum_{a\in\mathcal{A}}(i_a+\theta_q d_a)\log\frac{i_a+\theta_q d_a}{d_a} \tag{55}$$

By Lemma 11, $\Delta_k \leq \theta_q$.

Now, we write the t-step KL-divergence KL_t as a sum of the expectation of 1-step KL-divergences. We'll abbreviate a measure $\mathrm{P}(\cdot|E)$ as ${}^E\mathrm{P}$.

$$\begin{aligned}
\operatorname*{KL}_t\left({}^E\mathrm{P}^{\pi^i_\alpha}_\mu\middle\|{}^E\mathrm{P}^{\pi^d}_\mu\right) &= \mathbb{E}_{h_{<t}\sim{}^E\mathrm{P}^{\pi^i_\alpha}_\mu}\log\frac{{}^E\mathrm{P}^{\pi^i_\alpha}_\mu(h^\backslash_{<t})}{{}^E\mathrm{P}^{\pi^d}_\mu(h^\backslash_{<t})} \\
&\overset{(a)}{\leq} \mathbb{E}_{h_{<t}\sim{}^E\mathrm{P}^{\pi^i_\alpha}_\mu}\log\frac{\mathrm{P}^{\pi^i_\alpha}_\mu(h^\backslash_{<t})/\,\mathrm{P}^{\pi^i_\alpha}_\mu(E)}{\mathrm{P}^{\pi^d}_\mu(h^\backslash_{<t})/\,\mathrm{P}^{\pi^d}_\mu(E)} \\
&= \mathbb{E}_{h_{<t}\sim{}^E\mathrm{P}^{\pi^i_\alpha}_\mu}\log\frac{\mathrm{P}^{\pi^i_\alpha}_\mu(h^\backslash_{<t})}{\mathrm{P}^{\pi^d}_\mu(h^\backslash_{<t})} + \log\frac{\mathrm{P}^{\pi^d}_\mu(E)}{\mathrm{P}^{\pi^i_\alpha}_\mu(E)} \\
&\leq \mathbb{E}_{h_{<t}\sim{}^E\mathrm{P}^{\pi^i_\alpha}_\mu}\log\frac{\mathrm{P}^{\pi^i_\alpha}_\mu(h^\backslash_{<t})}{\mathrm{P}^{\pi^d}_\mu(h^\backslash_{<t})} - \log\mathrm{P}^{\pi^i_\alpha}_\mu(E) \\
&=: \mathbb{E}_{h_{<t}\sim{}^E\mathrm{P}^{\pi^i_\alpha}_\mu}\log\frac{\mathrm{P}^{\pi^i_\alpha}_\mu(h^\backslash_{<t})}{\mathrm{P}^{\pi^d}_\mu(h^\backslash_{<t})} + C_\alpha \\
&\overset{(b)}{=} C_\alpha + \mathbb{E}_{h_{<t}\sim{}^E\mathrm{P}^{\pi^i_\alpha}_\mu}\sum_{k=0}^{t-1}\log\frac{\mathrm{P}^{\pi^i_\alpha}_\mu(h^\backslash_k|h_{<k})}{\mathrm{P}^{\pi^d}_\mu(h^\backslash_k|h_{<k})} \\
&= C_\alpha + \sum_{k=0}^{t-1}\mathbb{E}_{h_{<k}\sim{}^E\mathrm{P}^{\pi^i_\alpha}_\mu}\mathbb{E}_{h_k\sim{}^E\mathrm{P}^{\pi^i_\alpha}_\mu(\cdot|h_{<k})}\log\frac{\mathrm{P}^{\pi^i_\alpha}_\mu(h^\backslash_k|h_{<k})}{\mathrm{P}^{\pi^d}_\mu(h^\backslash_k|h_{<k})} \\
&= C_\alpha + \sum_{k=0}^{t-1}\mathbb{E}_{h_{<k}\sim{}^E\mathrm{P}^{\pi^i_\alpha}_\mu}\sum_{h^\backslash_k\in\mathcal{A}\times\mathcal{O}}{}^E\mathrm{P}^{\pi^i_\alpha}_\mu(h^\backslash_k|h_{<k})\log\frac{\mathrm{P}^{\pi^i_\alpha}_\mu(h^\backslash_k|h_{<k})}{\mathrm{P}^{\pi^d}_\mu(h^\backslash_k|h_{<k})}
\end{aligned}$$

$$
\begin{aligned}
&\leq C_\alpha + \sum_{k=0}^{t-1} \mathbb{E}_{h_{<k}\sim {}^E\mathrm{P}_\mu^{\pi_\alpha^i}} \sum_{h_k^{\setminus}\in\mathcal{A}\times\mathcal{O}} \frac{\mathrm{P}_\mu^{\pi_\alpha^i}(h_k^{\setminus}|h_{<k})}{\mathrm{P}_\mu^{\pi_\alpha^i}(E)} \log \frac{\mathrm{P}_\mu^{\pi_\alpha^i}(h_k^{\setminus}|h_{<k})}{\mathrm{P}_\mu^{\pi^d}(h_k^{\setminus}|h_{<k})} \\
&= C_\alpha + \sum_{k=0}^{t-1} \mathbb{E}_{h_{<k}\sim {}^E\mathrm{P}_\mu^{\pi_\alpha^i}} \frac{1}{\mathrm{P}_\mu^{\pi_\alpha^i}(E)} \underset{1}{\mathrm{KL}}\left(\mathrm{P}_\mu^{\pi_\alpha^i}(\cdot|h_{<k}) \middle\| \mathrm{P}_\mu^{\pi^d}(\cdot|h_{<k})\right) \\
&= C_\alpha + \frac{1}{\mathrm{P}_\mu^{\pi_\alpha^i}(E)} \sum_{k=0}^{t-1} \mathbb{E}_{h_{<k}\sim {}^E\mathrm{P}_\mu^{\pi_\alpha^i}} \underset{1}{\mathrm{KL}}\left(\pi_\alpha^i(\cdot|h_{<k}) \| \pi^d(\cdot|h_{<k})\right) \\
&\overset{(c)}{\leq} -\log \mathrm{P}_\mu^{\pi_\alpha^i}(E) + \frac{1}{\mathrm{P}_\mu^{\pi_\alpha^i}(E)} {}^E\mathbb{E}_\mu^{\pi_\alpha^i} \sum_{k=0}^{t-1} \theta_q(h_{<k}) \\
&\leq -\log \mathrm{P}_\mu^{\pi_\alpha^i}(E) + \frac{1}{\mathrm{P}_\mu^{\pi_\alpha^i}(E)^2} \mathbb{E}_\mu^{\pi_\alpha^i} \sum_{k=0}^{t-1} \theta_q(h_{<k}) \qquad (56)
\end{aligned}
$$

where (a) follows from $h_{<t}$ satisfying E with ${}^E\mathrm{P}_\mu^{\pi_\alpha^i}$-prob. 1, (b) follows because μ and π^d are fair, and (c) follows from Equation 55 and Lemma 11, and the others are trivial.

Finally,

$$
\begin{aligned}
\mathbb{E}_\mu^{\pi_\alpha^i} \sum_{k=0}^{t-1} \theta_q(h_{<k}) &= t\, \mathbb{E}_{k\sim\mathbb{U}([t])}\, \mathbb{E}_\mu^{\pi_\alpha^i}\, \theta_q(h_{<k}) \\
&= t \left(\left(\mathbb{E}_{k\sim\mathbb{U}([t])}\, \mathbb{E}_\mu^{\pi_\alpha^i}\, \theta_q(h_{<k}) \right)^3 \right)^{1/3} \\
&\overset{(a)}{\leq} t \left(\mathbb{E}_{k\sim\mathbb{U}([t])}\, \mathbb{E}_\mu^{\pi_\alpha^i}\, \theta_q(h_{<k})^3 \right)^{1/3} \\
&= t \left(\frac{1}{t} \sum_{k=0}^{t-1} \mathbb{E}_\mu^{\pi_\alpha^i}\, \theta_q(h_{<k})^3 \right)^{1/3} \\
&\overset{(b)}{\leq} t^{2/3} |\mathcal{A}|^{1/3} \alpha^{-1} (24 w(\pi^d)^{-1} + 12)^{1/3} \qquad (57)
\end{aligned}
$$

where (a) follows from Jensen's Inequality, and (b) follows from Theorem 6. Combining this with Inequality 56, and recalling $\mathrm{P}_\mu^{\pi_\alpha^i}(E) \geq 1 - \alpha/w(\pi^d)$, we have

$$
\underset{t}{\mathrm{KL}}\left({}^E\mathrm{P}_\mu^{\pi_\alpha^i} \middle\| {}^E\mathrm{P}_\mu^{\pi^d} \right) \leq \frac{\alpha^{-1}|\mathcal{A}|^{1/3}(24 w(\pi^d)^{-1} + 12)^{1/3}}{(1 - \alpha/w(\pi^d))^2} t^{2/3} - \log(1 - \alpha/w(\pi^d)) \qquad (58)
$$

■