Regret Bounds for Robust Online Decision Making

Alexander Appel   ALEXAPPEL8@GMAIL.COM
Computational Rational Agents Laboratory

Vanessa Kosoy   VANESSA@ALTER.ORG.IL
Faculty of Mathematics, Technion - Israel Institute of Technology

Editors: Nika Haghtalab and Ankur Moitra

Abstract

We propose a framework which generalizes "decision making with structured observations" from Foster et al. (2023) by allowing *robust* (i.e. multivalued) models. In this framework, each model associates each decision with a *convex set* of probability distributions over outcomes. Nature can choose distributions out of this set in an arbitrary (adversarial) manner, that can be non-oblivious and depend on past history. The resulting framework offers much greater generality than classical bandits and reinforcement learning, since the realizability assumption becomes much weaker and more realistic. We then derive a theory of regret bounds for this framework, which extends the "decision-estimation coefficients" of Foster et al. (2023). Although our lower and upper bounds are not tight, they are sufficient to fully characterize power-law learnability. We demonstrate this theory in two special cases: robust linear bandits (previously studied in Kosoy (2024)) and tabular robust online reinforcement learning (previously studied in Tian et al. (2021)). In both cases, we derive regret bounds that improve state-of-the-art (except that we do not address computational efficiency).

Keywords: adversarial learning and robustness, reinforcement learning, multi-armed bandits

1. Introduction

Traditional approaches to formal guarantees and statistical complexity analysis for interactive decision-making (e.g. multi-armed bandits, reinforcement learning theory) often need strong assumptions about the environment. Typical of assumptions include

- Assuming that the outcomes of any particular decision are IID samples from a fixed distribution (stochastic multi-armed bandits, see e.g. Lattimore and Szepesvári (2020)).

- Assuming that the state of environment changes according to a constant Markovian law (see e.g. Azar et al. (2017)).

- Assuming that the distribution of outcomes belongs to a known parametric family of distributions (see e.g. Russo and Van Roy (2013); Osband and Van Roy (2014); Jin et al. (2020)).

- Assuming that the MDP value function or some related quantity belongs to some known parametric family of functions (model-free reinforcement learning, see e.g. Wang et al. (2020); Du et al. (2021); Jin et al. (2021)).

Clearly, such realizability assumptions are often unrealistic, especially for agents operating in the physical world (as opposed to a game or a simulation). Any exact Markovian description of the real world would have an astronomically large state space, and any family of distributions that includes an exact description of physical law would be intractably large, at least because the computational complexity of the physical world is far beyond that of the agent itself.

There are several approaches for relaxing these assumptions, but each has major limitations.

The first approach is allowing a linear term in the regret bound, whose coefficient scales with some measure of distance between the real environment and the algorithm's class of models (see e.g. Zanette et al. (2020)). However, most work in this approach is limited to the model-free setting. Moreover, this typically requires the model to well-approximate the environment *uniformly* over the state space, whereas realistic approximations often break down outside of particular regions.

The second approach is allowing the environment to adversarially change its behavior between episodes/trials: adversarial multi-armed bandits (see e.g. Lattimore and Szepesvári (2020)) and adversarial reinforcement learning (see e.g. Liu et al. (2022); Foster et al. (2022)). However, the latter is plagued by statistical intractability results. Even in the limited special cases when adversarial learning is tractable, it still requires the environment to have a simple unambiguous description within each trial: e.g. in linear adversarial bandits the reward has to be a linear function of the arm.

The third approach, which is the basis for our own, is using *robust* Markov Decision Processes (RMDP) (see e.g. Suilen et al. (2025)). As opposed to an ordinary MDP, an RMDP has a *multivalued* transition kernel, and nature can adversarially choose the distribution upon each state transition. However, most work on reinforcement learning with RMDP's (see e.g. Wang and Zou (2021); Dong et al. (2022); Panaganti and Kalathil (2022)) has two assumptions which we dispense with:

- The set of distributions assigned by the transition kernel to a state-action pair has to be a ball relative to some metric or divergence.

- The training date for the algorithm consists of interaction with the MDP defined by taking the *centers* of these balls.

The typical motivation is training a reinforcement learning agent in a simulation (corresponding to the center MDP) and deploying it in a physical environment (corresponding to the RMDP). This requires that a sufficiently accurate simulation is available, and that the realizability assumption is satisfied for the simulation. On the other hand, we want to study robust *online* reinforcement learning, where the agent has to learn directly from the real world (or the simulation is so complex that the realizability assumption fails even there). One work which *does* consider learning from interacting with the actual RMDP is Lim et al. (2013). Their setting is a very narrow special case of our framework, where the multivalued kernel is always either a singleton or equal to a *known* set.

We propose to study an agent that makes a sequence of decisions. On each round, the agent chooses an action a from a set $\mathcal{A}$, receives an observation o from a set $\mathcal{O}$, and calculates its reward from a, o. A class $\mathcal{H}$ of models is given, where a model is a *multivalued*

function from $\mathcal{A}$ to distributions over $\mathcal{O}$. $\mathcal{H}$ acts as a set of possible constraints which nature may fulfill. In rich environments, models of this form may be much easier to specify and learn than a full model of the environment. Given a model M, the set $M(a)$ may be interpreted as a menu of choices available to an adversarial environment[1]. Each model M has an associated maximin value, and we define the *regret* relatively to that value[2] assuming that nature is constrained to be consistent with M[3].

Notice that even though this setting has the appearance of a multi-armed bandit, it allows episodic reinforcement learning as a special case: in this case, $\mathcal{A}$ consists of within-episode policies and $\mathcal{O}$ is the set of possible trajectories within an episode.

Here are some examples of possible applications.

Example 1 *A robot learns to complete tasks (e.g. moving objects, cleaning, cooking) in an environment that has inanimate objects and people. The robot's own movement and the interaction with objects can be described fairly well by fixed probability distributions. However, the behavior of people can only be described as a set of probability distributions. Moreover, their behavior might be affected by previous interactions with the robot in hard-to-predict ways.*

Example 2 *A self-driving car needs to bring passengers to their destination, while taking into account speed, convenience, fuel consumption and safety. The movement of cars can be partially predicted based on physical laws and common driver habits, but partially it depends on hard-to-predict idiosyncrasies of individual drivers. Moreover, the car's sensors only give it information about a certain neighborhood around the car. Previously unknown objects can enter this area, and the probability distribution of such new objects can change between days or geographical areas in complex ways.*

Example 3 *An AI system manages an investment portfolio on the stock market. Not only that some price movements are hard to predict (e.g. caused by a technological breakthrough or a political event), but some are the result of adversarial actors trying to profit on expense of the AI's portfolio that learning from its previous behavior.*

In Foster et al. (2023), regret bounds for online decision-making were studied for conventional (single-valued) models. They proved that a certain function called the *decision-estimation coefficient* enabled an almost tight characterization of the regret bound for any model class. Here, we define an analogue of the decision-estimation coefficient for the robust setting, and show a similar upper bound on regret. We also show a lower bound on regret, which is weaker than the non-robust analogue, but still sufficient to characterize model classes which admit a sublinear power-law regret bound.

---

1. This can be regarded as decision-making with uncertainty expressed as *imprecise probability*, see e.g. Augustin et al. (2014).

2. The typical choice in adversarial reinforcement learning is to measure the regret against the best action in hindsight. However, Theorem 1 of Tian et al. (2021) gives a family of robust MDP's with a lower bound on best-action regret which is exponential in the horizon. Maximin regret avoids this impossibility result.

3. Importantly, we don't require nature to choose a *fixed* mapping from $\mathcal{A}$ to distributions on $\mathcal{O}$. Nature's policy can be non-stationary and non-oblivious.

In Foster et al. (2023), the algorithm that implements that upper bound requires access to an online distribution learning oracle, and the bound scales with the statistical complexity of the latter. This statistical complexity has bounds in terms of model class cardinality or covering number, which are derived in Foster et al. (2021) using the techniques of Vovk (1995). In our setting, we require a similar oracle for *robust* online distribution learning. Again, we derive similar bounds on statistical complexity, but this requires different techniques: our algorithm imitates a *prediction market*, an idea inspired by Shafer et al. (2005); Garrabrant et al. (2017); Kosoy (2017).

We apply our methods to derive upper bounds on regret in two special cases. One case is tabular episodic robust online reinforcement learning, a generalization of which was studied in Tian et al. (2021). The motivation of Tian et al. (2021) was studying multi-agent learning, but since an RMDP can be rephrased as a zero-sum two-player stochastic game[4], their results apply to our framework. There, we get a regret bound of $\tilde{\mathcal{O}}(H\sqrt{S^3AT})$, where H is the episode length, S is the number of states, A is the number of actions, and T is the number of episodes. For $T \gg 0$, this is an improvement on Tian et al. (2021)'s bound of $\tilde{\mathcal{O}}(H^2\sqrt[3]{SAT^2})$. However, our algorithm is not computationally efficient. It remains an open question whether there exists a polynomial-time algorithm with regret of $\tilde{\mathcal{O}}(\text{poly}(H,S,A)\sqrt{T})$[5].

The other special case is robust linear bandits, previously studied in Kosoy (2024). There, we get a similar $\tilde{\mathcal{O}}(\sqrt{T})$ regret bound, except that the dependence on the dimension Z of the model class is improved from Z^2 to Z.

1.1. Notation and General Framework

Lambda notation for functions is used throughout. ΔX is the space of probability distributions on a space X. An *imprecise belief* on X is a nonempty closed convex subset of ΔX. $\square X$ is the space of imprecise beliefs on a space X. μ, ν are used for probability distributions, and Ψ, Φ are used for imprecise beliefs. Expectations of functions with respect to imprecise beliefs are defined to be the worst-case expectation.

$$\mathbb{E}_\Psi[f] := \min_{\mu \in \Psi} \mathbb{E}_\mu[f]$$

Our general framework for robust reinforcement learning is as follows. There are spaces $\mathcal{A}, \mathcal{O}$ of actions a and observations o, and a known reward function $r : \mathcal{A} \times \mathcal{O} \to [0,1]$. T is the number of timesteps. An algorithm $\pi : (\mathcal{A} \times \mathcal{O})^{<T} \to \Delta\mathcal{A}$ repeatedly interacts with an environment $\theta : (\mathcal{A} \times \mathcal{O})^{<T} \to (\mathcal{A} \to \Delta\mathcal{O})$ to produce a history. $\theta \bowtie \pi : \Delta\left((\mathcal{A} \times \mathcal{O})^T\right)$ is the distribution on histories produced by θ and π interacting.

A model M is of type $\mathcal{A} \to \square\mathcal{O}$ and $\mathcal{H} \subseteq \mathcal{A} \to \square\mathcal{O}$ is a hypothesis class[6]. To link environments θ and models M, we say that θ is *consistent* with M if, for all h, a,

---

4. One player is choosing the RMDP action, the other player is choosing the distribution out of the multi-valued transition kernel.

5. In Xie et al. (2020), a polynomial-time algorithm for learning a zero-sum game is presented, which has regret $\tilde{\mathcal{O}}(\sqrt{H^3S^3A_1^3A_2^3T})$, where A_1 and A_2 are the cardinalities of the action sets of the two players. However, they assume that the player can observe the opponent's action, which with our motivation is an unnatural assumption.

6. Notice that, unlike in the classical setting of Foster et al. (2023), environments and models have different type signatures.

$\theta(h)(a) \in M(a)$. This is written as $\theta \models M$. If θ is the true environment, and $\theta \models M$, we say that M is a true model.[7] The notation M^* denotes an arbitrary true model.

The (worst-case) expected reward of a model M for an action a is denoted as $f^M(a)$, which abbreviates $\min_{\mu \in M(a)} \mathbb{E}_{o \sim \mu}[r(a, o)]$. $\max(f^M)$ abbreviates $\max_{a \in \mathcal{A}} f^M(a)$, the maximin expected reward for M. We define the regret of an algorithm π against a model M for T timesteps as

$$\mathbf{REG}(\pi, M, T) := \max_{\theta : \theta \models M} \left(T \cdot \max\left(f^M\right) - \mathbb{E}_{\theta \bowtie \pi}\left[\sum_{t=1}^{T} r(a_t, o_t) \right] \right)$$

Something to note is that any algorithm π with low regret on all $M \in \mathcal{H}$ will exploit non-adversarial environments θ, in the sense that the average reward against θ will be comparable to or exceed the most optimistic maximin value $\max_{M : \theta \models M} \max(f^M)$.

2. Generalizing the Decision-Estimation Coefficient

The decision-estimation coefficient (DEC) was introduced in Foster et al. (2021) to quantify the difficulty of learning to behave optimally in various classical reinforcement learning problems, and provides nearly matching upper and lower bounds on minimax regret.

The definition of the DEC and its variants depends on the *Hellinger distance* between probability distributions, which is defined as

$$D_H(\mu, \nu) := \sqrt{\frac{1}{2} \int \left(\sqrt{\frac{d\mu}{d\xi}} - \sqrt{\frac{d\nu}{d\xi}} \right)^2 d\xi}$$

where μ and ν are absolutely continuous with respect to ξ. The choice of ξ doesn't affect the Hellinger distance. $D_H^2(\mu, \nu)$ is the square of the Hellinger distance.

There are many variants of the decision-estimation coefficient, and we introduce the constrained DEC from Foster et al. (2023) as an illustrative example. This quantity depends on the hypothesis space $\mathcal{H}$, a belief $\overline{M} : \mathcal{A} \to \Delta\mathcal{O}$, and a parameter $\varepsilon > 0$, and is defined in the classical case as

$$\text{dec}_\varepsilon^c\left(\mathcal{H}, \overline{M}\right) := \min_{p \in \Delta\mathcal{A}} \max_{M \in \mathcal{H}} \left\{ \max\left(f^M\right) - \mathbb{E}_{a \sim p}\left[f^M(a)\right] \;\middle|\; \mathbb{E}_{a \sim p}\left[D_H^2\left(\overline{M}(a), M(a)\right)\right] \leq \varepsilon^2 \right\}$$

This quantity is relevant to decision-making because a good algorithm should act to either attain low regret, or acquire information to distinguish models. If actions are selected from p, the M where $\mathbb{E}_{a \sim p}[D_H^2(\overline{M}(a), M(a))] \leq \varepsilon^2$ nearly mimic the belief $\overline{M}$ and the information gained doesn't effectively distinguish among this cluster of models. Therefore, the regret should be low among this cluster, which justifies minimizing the maximum regret among near-mimic models.

To generalize the DEC to our setting, we must first change our notion of what counts as a "near-mimic" of the beliefs $\overline{M}$. If Ψ is our belief state, and $\Psi \subseteq \Phi$, then any distribution

---

7. In particular, there may be *multiple* true models.

from Ψ could have been produced by Φ, so Φ can mimic Ψ. This motivates the following definition.

Definition 1 (Asymmetric Distance) *Given a distance metric D on ΔX, we define the asymmetric distance from the set Ψ to Φ as*

$$D(\Psi \to \Phi) := \max_{\mu \in \Psi} \min_{\nu \in \Phi} D(\mu, \nu)$$

$D(\Psi \to \Phi)$ is low when Ψ is almost a subset of Φ. For $D^2(\Psi \to \Phi)$, it doesn't matter whether the squaring happens inside or outside the maximin. Accordingly, the models M where $\mathbb{E}_{a \sim p}[D_H^2(\overline{M}(a) \to M(a))] \leq \varepsilon^2$ are considered to be the "near-mimics" of $\overline{M}$.

Our second change is to the regret term $\max(f^M) - \mathbb{E}_{a \sim p}[f^M(a)]$. Our chosen notion of regret is the gap between the maximin reward and the expected reward against the true environment θ, but $\mathbb{E}_{a \sim p}[f^M(a)]$ is the expected *worst-case* reward. We do not know θ, so we substitute with the expected reward against the *beliefs* $\overline{M}$ so the regret term becomes $\max(f^M) - \mathbb{E}_{a \sim p}[f^{\overline{M}}(a)]$. The expected reward (according to $\overline{M}$) still needs to be linked to the true reward, but this is a matter of prediction, not decision-making.

We now proceed to our third change to the definition of the DEC. The E2D+ algorithm from Foster et al. (2023), with a regret bound that scaled with the constrained DEC, did not generalize well to our setting. The primary obstacle was that no regret guarantee could be shown for models M which were too far from $\overline{M}$ to count as a near-mimic. To remedy this matter, we introduce the "fuzzy DEC", which is a relaxation of the constrained DEC. To present the definition, we use sub-probability distributions, which are measures μ with $\sum_x \mu(x) \in [0, 1]$. The space of these measures is notated as $\Delta^s X$. Given some $\mu : \Delta^s X$ and $\nu : \Delta Y$, the notation $\mathbb{E}_{x, y \sim \mu, \nu}[f(x, y)]$ abbreviates $\sum_{x,y} \mu(x)\nu(y)f(x, y)$.

Definition 2 (Fuzzy Decision-Estimation Coefficient)

$$dec_\varepsilon^f(\mathcal{H}, \overline{M}) := \min_{p \in \Delta \mathcal{A}} \max_{\mu \in \Delta^s \mathcal{H}} \left\{ \mathbb{E}_{M, a \sim \mu, p}[\max(f^M) - f^{\overline{M}}(a)] \ \middle| \ \mathbb{E}_{M, a \sim \mu, p}\left[D_H^2(\overline{M}(a) \to M(a))\right] \leq \varepsilon^2 \right\}$$

Comparing this to the constrained DEC, we see that the fuzzy decision-estimation coefficient implements a softer cutoff for which models count as near-mimics of $\overline{M}$. If $D_H^2(\overline{M}(a) \to M(a)) = 2\varepsilon^2$, then μ can put 0.5 measure on M and fulfill the strict distance requirement, causing the regret term in the fuzzy DEC to be half of the true regret against M. This imposes nontrivial regret bounds on models which aren't near-mimics of $\overline{M}$.

We now adapt another variant of the DEC from Foster et al. (2023).

Definition 3 (Offset Decision-Estimation Coefficient)

$$dec_\gamma^o(\mathcal{H}, \overline{M}) := \min_{p \in \Delta \mathcal{A}} \max_{M \in \mathcal{H}} \left(\max(f^M) - \mathbb{E}_{a \sim p}\left[f^{\overline{M}}(a)\right] - \gamma \mathbb{E}_{a \sim p}\left[D_H^2(\overline{M}(a) \to M(a))\right] \right)$$

These quantities are closely related, and the fuzzy DEC can be viewed as the offset DEC with the optimal choice of the information-to-regret ratio γ.

Proposition 1 *If $\mathcal{H}$ is a compact subset of $\mathcal{A} \to \Box \mathcal{O}$ then*

$$dec_\varepsilon^f(\mathcal{H}, \overline{M}) = \min_{\gamma \geq 0} \left(\max(dec_\gamma^o(\mathcal{H}, \overline{M}), 0) + \gamma \varepsilon^2 \right)$$

Given $\mathcal{H}, \varepsilon$, we can consider the worst-case value of the fuzzy DEC, and define

$$\text{dec}_\varepsilon^f(\mathcal{H}) := \max_{\overline{M} \in \mathcal{A} \to \Box\mathcal{O}} \text{dec}_\varepsilon^f(\mathcal{H}, \overline{M})$$

It is natural to ask whether restricting to probabilistic beliefs of type $\mathcal{A} \to \Delta\mathcal{O}$ can decrease the worst-case fuzzy DEC. We answer this question negatively.

Proposition 2 *For every $\overline{M} : \mathcal{A} \to \Box\mathcal{O}$, there is a probabilistic model $\overline{\theta} : \mathcal{A} \to \Delta\mathcal{O}$ consistent with $\overline{M}$ where $\text{dec}_\varepsilon^f(\mathcal{H}, \overline{M}) \leq \text{dec}_\varepsilon^f(\mathcal{H}, \overline{\theta})$*

When proving upper bounds, Proposition 2 lets us assume that $\overline{M}$ is probabilistic.

2.1. Upper Bounds with the E2D Algorithm

The Estimations to Decisions (E2D) algorithm was introduced in Foster et al. (2021), and reduces decision-making to online estimation. An estimation oracle produces beliefs $\widehat{M}_t$ on each timestep, and actions are selected by sampling from the distribution p which minimizes the DEC. For all reinforcement learning problems, this process gives an upper bound on regret which scales with the DEC and the performance of the estimation oracle. Accordingly, we must introduce some notation to quantify the quality of an estimation oracle.

Given a history and a timestep t, let $\widehat{M}_t : \mathcal{A} \to \Box\mathcal{O}$, $\pi_t : \Delta\mathcal{A}$, and $\theta_t : \mathcal{A} \to \Delta\mathcal{O}$ denote the estimated model, the distribution over actions, and the behavior of the environment on that timestep. For full generality, let $\mathcal{L} : (\mathcal{A} \to \Box\mathcal{O}) \times (\mathcal{A} \to \Box\mathcal{O}) \times \mathcal{A} \to \mathbb{R}^{\geq 0}$ be some function which measures how far apart two models are for a given action. By default, we have $\mathcal{L}(\overline{M}, M, a) = D_H^2(\overline{M}(a) \to M(a))$, but squared Hellinger distance is not always the most convenient loss to work with in later proofs.

Definition 4 (Inaccuracy Bound) *A function $\beta : \mathbb{N}^{>0} \times [0,1] \to \mathbb{R}^{\geq 0}$ is an inaccuracy bound on an online estimator $\widehat{M}$ for $\mathcal{L}$ if, for all T, δ, algorithms π, models $M \in \mathcal{H}$, and environments $\theta \models M$, with $1 - \delta$ probability according to $\theta \bowtie \pi$, we have*

$$\sum_{t=1}^T \mathop{\mathbb{E}}_{a \sim \pi_t} \left[\mathcal{L}(\widehat{M}_t, M, a) \right] \leq \beta(T, \delta)$$

Definition 5 (Optimism Bound) *A function $\alpha : \mathbb{N}^{>0} \times [0,1] \to \mathbb{R}^{\geq 0}$ is an optimism bound on an online estimator $\widehat{M}$ if, for all T, δ, algorithms π, models $M \in \mathcal{H}$, and environments $\theta \models M$, with $1 - \delta$ probability according to $\theta \bowtie \pi$, we have*

$$\sum_{t=1}^T \mathop{\mathbb{E}}_{a \sim \pi_t} \left[f^{\widehat{M}_t}(a) - f^{\theta_t}(a) \right] \leq \alpha(T, \delta)$$

$\beta_{\widehat{M}}$ and $\alpha_{\widehat{M}}$ denote inaccuracy and optimism bounds for an estimator. In the classical case, low inaccuracy implies low optimism, but this doesn't hold here[8]. Both quantities

---

8. A toy example is the multi-armed bandit problem, where a true model M^* says that a certain arm returns ≥ 0.7 reward. If the environment θ reliably returns 0.8 reward on that arm, while the models $\widehat{M}_t$ keep predicting 0.9 reward on that arm, the sequence $\widehat{M}_t$ is perfectly accurate relative to M^*, but the optimism is high.

appear in the regret bound for the Estimations to Decisions algorithm, so high-quality regret bounds require an estimation oracle with a low α and β. If an online estimator predicts by averaging past outcomes together, the optimism bound may be linear in T, so more sophisticated approaches are needed.

The variant of the E2D algorithm we present here depends on the time horizon T, for ease of analysis. However, converting it into an anytime algorithm with comparable performance may be done via the doubling trick (see e.g. Besson and Kaufmann (2018); Lattimore and Szepesvári (2020)), as well as more refined methods.

Algorithm 1: Estimations to Decisions (E2D)

Parameters: Rounds T, failure odds δ, oracle $\widehat{M}$, loss $\mathcal{L}$, hypothesis class $\mathcal{H}$, inaccuracy bound $\beta_{\widehat{M}}$ for $\widehat{M}, \mathcal{L}, \mathcal{H}$.

$\mathcal{H}_1 \leftarrow \mathcal{H}$

$h_{<1} \leftarrow \varnothing$

for $1 \leq t \leq T$ **do**

$\quad \widehat{M}_t \leftarrow \widehat{M}(h_{<t})$

$\quad p_t \leftarrow \underset{p \in \Delta\mathcal{A}}{\operatorname{argmin}} \underset{\mu \in \Delta^s \mathcal{H}_t}{\max} \left\{ \underset{\mu, p}{\mathbb{E}} \left[\max(f^M) - f^{\widehat{M}_t}(a) \right] \, \Big| \, \underset{\mu, p}{\mathbb{E}} \left[\mathcal{L}(\widehat{M}_t, M, a) \right] \leq \frac{\beta_{\widehat{M}}(T, \delta)}{T} \right\}$

$\quad$ **return** $a_t \sim p_t$

$\quad$ Receive o_t from the environment.

$\quad \mathcal{H}_{t+1} \leftarrow \mathcal{H} \cap \left\{ M \, \Big| \, \sum_{k=1}^t \underset{a \sim p_k}{\mathbb{E}} \left[\mathcal{L}(\widehat{M}_k, M, a) \right] \leq \beta_{\widehat{M}}(T, \delta) \right\}$

$\quad h_{<t+1} \leftarrow h_{<t}, a_t, o_t$

end

Given a function $\mathcal{L}$, $\operatorname{dec}_\varepsilon^{f,\mathcal{L}}(\mathcal{H})$ is the variant of the fuzzy DEC with $\mathbb{E}_{M, a \sim \mu, p} \left[\mathcal{L}(\overline{M}, M, a) \right]$ in place of $\mathbb{E}_{M, a \sim \mu, p} \left[D_H^2(\overline{M}(a) \to M(a)) \right]$.

Theorem 1 *For all $T, \delta, \mathcal{L}$, oracles $\widehat{M}$, and models $M \in \mathcal{H}$, if $\varepsilon = \sqrt{\frac{\beta_{\widehat{M}}(T, \delta)}{T}}$, we have*

$$\boldsymbol{REG}(E2D, M, T) \leq 2T \operatorname{dec}_\varepsilon^{f, \mathcal{L}}(\mathcal{H}) + \alpha_{\widehat{M}}(T, \delta) + 2T\delta$$

Theorem 1 lets us produce an upper bound on regret for any problem in our framework by picking a function $\mathcal{L}$, upper-bounding the fuzzy DEC for $\mathcal{L}$, constructing an estimator $\widehat{M}$, and upper-bounding $\beta_{\widehat{M}}$ and $\alpha_{\widehat{M}}$.

A robust reinforcement learning problem is said to be *learnable* if there is some algorithm π (which may depend on the timestep T), and $p < 1$, where

$$\max_{M \in \mathcal{H}} (\boldsymbol{REG}(\pi, M, T)) \in \mathcal{O}(T^p)$$

In one direction, if the fuzzy DEC shrinks as ε^p for some $p > 0$, and there is an estimator with sub-linear inaccuracy and optimism bounds, the E2D algorithm witnesses learnability.

Corollary 1 *If there is a $p > 0, q > 0, r < 1, s < 1$ and online estimator $\widehat{M}$ such that*

$$\limsup_{\varepsilon \to 0} \frac{\operatorname{dec}_\varepsilon^f(\mathcal{H})}{\varepsilon^p} < \infty, \limsup_{T \to \infty} \frac{\alpha_{\widehat{M}}(T, T^{-q})}{T^r} < \infty, \limsup_{T \to \infty} \frac{\beta_{\widehat{M}}(T, T^{-q})}{T^s} < \infty$$

then the E2D algorithm with $\widehat{M}$ as an oracle has $\mathcal{O}\left(T^{\max\left(1-\frac{p(1-s)}{2},r,1-q\right)}\right)$ regret[9] on all $M \in \mathcal{H}$.

2.2. Lower Bounds

Proposition 3 *For all $p \in (0,1)$ such that $\liminf_{\varepsilon\to 0} \frac{dec_\varepsilon^f(\mathcal{H})}{\varepsilon^p} = \infty$, if $\varepsilon_T = \sqrt{\frac{1}{T\ln(T)}}$, we have*

$$\min_\pi \max_{M\in\mathcal{H}}(\boldsymbol{REG}(\pi,M,T)) \in \Omega\left(T \cdot dec^f_{\frac{1}{\varepsilon_T^{1-p}}}(\mathcal{H})\right)$$

This lower bound relies on Hellinger distance, and cannot be adapted to general notions of estimation error. To compare the upper bound of Theorem 1 and the lower bound of Proposition 3, consider three cases where $\text{dec}_\varepsilon^f(\mathcal{H})$ scales as $\varepsilon, \varepsilon^{1/2}$, and $\varepsilon^{1/3}$. Neglecting estimation complexity, the upper bound of Theorem 1 would scale as $T^{1/2}, T^{3/4}$, and $T^{5/6}$. The lower bound above would be inapplicable in the first case, and scale as $T^{1/2}$ and $T^{3/4}$ in the other two cases. The lower bound on regret in the classical setting in Foster et al. (2023) is significantly tighter than this result. However, our lower bound still establishes that slow decay of the fuzzy DEC implies unlearnability.

Corollary 2 *If, for all $p > 0$, we have $\liminf_{\varepsilon\to 0} \frac{dec_\varepsilon^f(\mathcal{H})}{\varepsilon^p} = \infty$, then, for all $q < 1$ and π (which may depend on T), we have $\liminf_{T\to\infty} \frac{\max_{M\in\mathcal{H}}(\boldsymbol{REG}(\pi,M,T))}{T^q} = \infty$*

Putting Corollaries 1 and 2 together, if the fuzzy DEC declines more slowly than ε^p for every $p > 0$, then unlearnability holds. If the fuzzy DEC declines more quickly than ε^p for some $p > 0$, and sub-linear estimation error is attainable, the E2D algorithm certifies learnability.

3. Robust Online Estimation

3.1. The Robust Universal Estimator

We will now present an explicit online estimation algorithm which is suitable for use as an oracle, and computable if the sets $\mathcal{O}$ and $\mathcal{H}$ are finite. The RUE algorithm (Robust Universal Estimator) can be thought of as a prediction market. There is a set of bettors, $\mathcal{B}$, which have wealth and make bets against a market. The market odds $\widehat{M}_t : \mathcal{A} \to \Delta\mathcal{O}$ are generated, and the bettors bet against the market odds if they disagree. An outcome a_t, o_t is observed, and the bets resolve, which updates the wealth distribution ζ_t. The key part of RUE is that $\widehat{M}_t$ is set to ensure that the total bettor wealth is conserved no matter which outcome occurs. A bettor which reliably profits will acquire more than all of the wealth, which is impossible.

To enforce accuracy, we associate each model $M \in \mathcal{H}$ with a bettor which reliably profits if M is a true model and $\mathbb{E}_{a\sim\pi_t}\left[D_H^2\left(\widehat{M}_t(a) \to M(a)\right)\right]$ is high[10]. To enforce non-optimism,

---

9. Note that the quantity in the exponent is less than 1, witnessing learnability
10. Reliable profit is impossible, so if M is a true model, the sum of expected Hellinger-squared error must be low.

we introduce a pessimistic bettor $\bullet$, which reliably profits if $\mathbb{E}_{a\sim\pi_t}\left[f^{\widehat{M}_t}(a) - f^{\theta_t}(a)\right]$ is high[11]. There is also a uniform bettor u which exists to avert division-by-zero errors. Given a bettor B which isn't the pessimistic bettor, $M_B : \mathcal{A} \to \square\mathcal{O}$ denotes their corresponding model. For the uniform bettor, their model maps all actions to the uniform distribution on $\mathcal{O}$.

Mixtures of convex functions are convex, and the squared Hellinger distance to a convex set is convex by Lemma 4, so all minimization in RUE is over convex functions.

Algorithm 2: Robust Universal Estimator (RUE)

Parameters: Hypothesis class $\mathcal{H}$, rounds T, reward function r, prior ζ_1

$\varepsilon \leftarrow \min\left(\frac{1}{2}, \sqrt{\frac{\ln(2)}{T}}\right)$

Function `estimate`$(\zeta,\ a)$:

    return $\underset{\mu\in\Delta\mathcal{O}}{\text{argmin}}\ \underset{B\sim\zeta}{\mathbb{E}}\left[\text{if } B\neq\bullet, 2D_H^2(\mu \to M_B(a)),\ \text{else } \varepsilon\cdot\underset{o\sim\mu}{\mathbb{E}}\left[r(a,o)\right]\right]$

Function `update`$(\zeta,\ \overline{M},\ a,\ o)$:

    for $B \in \mathcal{H}\cup\{u\}$ **do**

        $\mu_B \leftarrow \text{argmin}_{\mu\in M_B(a)}\ D_H^2(\overline{M}(a),\mu)$

        $\xi(B) \leftarrow \zeta(B)\cdot\left(\sqrt{\frac{\mu_B(o)}{\overline{M}(a)(o)}} + D_H^2(\overline{M}(a) \to M_B(a))\right)$

    end

    $\xi(\bullet) \leftarrow \zeta(\bullet)\cdot\left(1 + \varepsilon(\mathbb{E}_{o'\sim\overline{M}(a)}[r(a,o')] - r(a,o))\right)$

    return ξ

for $1 \leq t \leq T$ **do**

    $\widehat{M}_t \leftarrow \lambda a.\texttt{estimate}(\zeta_t,\ a)$

    return $\widehat{M}_t$

    Receive a_t, o_t from the environment.

    $\zeta_{t+1} \leftarrow \texttt{update}(\zeta_t,\ \widehat{M}_t,\ a_t,\ o_t)$

end

Proposition 4 *If the uniform bettor has positive probability according to ζ_1, the RUE algorithm never divides by zero and all ζ_t are probability distributions.*

Theorem 2 *If our estimator $\widehat{M}$ is the RUE algorithm with a suitable choice of prior[12], then $\beta_{\widehat{M}}(T,\delta) \leq \ln\left(\frac{2|\mathcal{H}|}{\delta}\right)$, and $\alpha_{\widehat{M}}(T,\delta) \leq \sqrt{T}\left(2\sqrt{\ln(2)} + \sqrt{2\ln\left(\frac{1}{\delta}\right)}\right)$*

The "suitable choice of prior" is to assign an arbitrarily low probability ε' to the uniform bettor, $\frac{1}{2} - \varepsilon'$ probability to the pessimistic bettor, and $\frac{1}{2|\mathcal{H}|}$ probability to all other bettors.

---

11. Reliable profit is impossible, so the sum of expected reward overestimation must be low

12. Technically, the bound on $\alpha_{\widehat{M}}(T,\delta)$ in Theorem 2 only holds for $T > 2$ in the limit as the uniform bettor probability according to ζ_1 approaches zero. However, the uniform bettor is only present to avert division-by-zero errors, and any positive probability for it is sufficient to do so. If the starting probability of the uniform bettor is, say, 0.001, this only has a minor effect on constants and doesn't affect the asymptotics with respect to T and $\ln\left(\frac{1}{\delta}\right)$.

The inaccuracy bound is close to the inaccuracy bound of Bayesian updating in the classical case.

3.2. Covering Numbers

The robust universal estimator only applies to finite hypothesis spaces, and must be generalized to infinite hypothesis spaces via covering numbers. We use the symmetric notion of Hausdorff distance between sets of distributions here, instead of asymmetric distance.

Definition 6 (ε-Covering Number) *The ε-covering number of $\mathcal{H}$, $\mathcal{N}(\mathcal{H}, \varepsilon)$, is defined as*

$$\mathcal{N}(\mathcal{H}, \varepsilon) := \min \left\{ |X| \; \middle| \; X \subseteq \mathcal{H}, \forall M \in \mathcal{H} \; \exists N \in X : \max_{a \in \mathcal{A}} D_H(N(a), M(a)) \leq \varepsilon \right\}$$

Proposition 5 *If $\mathcal{O}$ is finite, then for all $\mathcal{H}$, there exists an online estimator $\widehat{M}$ where $\beta_{\widehat{M}}(T, \delta) \leq \min_{\varepsilon > 0} \left(2 \ln \left(\frac{2\mathcal{N}(\mathcal{H}, \varepsilon)}{\delta} \right) + 8T\varepsilon^2 \right)$, and $\alpha_{\widehat{M}}(T, \delta) \leq \sqrt{T} \left(2\sqrt{\ln(2)} + \sqrt{2 \ln \left(\frac{1}{\delta} \right)} \right)$*

The estimator constructed in the proof of Proposition 5 works as follows. For each point N in the minimal cover, a "fattened model" N' is constructed where $N'(a) = \{ \mu \mid D_H(\mu \to N(a)) \leq \varepsilon \}$. Then, RUE is run on the finite set of fattened models. This differs from the classical proof by using imprecision in an essential way. To simplify Proposition 5, we introduce the Minkowski-Bougliand dimension of $\mathcal{H}$, defined as

$$MB(\mathcal{H}) := \limsup_{\varepsilon \to 0} \frac{\ln(\mathcal{N}(\mathcal{H}, \varepsilon))}{\ln(\frac{1}{\varepsilon})}$$

Swapping $\ln \left(\frac{2\mathcal{N}(\mathcal{H}, \varepsilon)}{\delta} \right)$ for $\ln \left(\frac{2}{\delta} \right) + MB(\mathcal{H}) \cdot \ln \left(\frac{1}{\varepsilon} \right)$ in Proposition 5 and minimizing over ε yields an estimation complexity in $\mathcal{O} \left(\log \left(\frac{1}{\delta} \right) + MB(\mathcal{H}) \cdot \log(T) \right)$, just as in the classical case.

4. Novel Regret Bounds

4.1. Robust Linear Bandits

We begin with a setting from Kosoy (2024), robust linear bandits. In this setting, $\mathcal{A}$ and $\mathcal{O}$ are large finite sets of actions and observations, and there is a reward function $r : \mathcal{A} \times \mathcal{O} \to [0, 1]$. There is a vector space $\mathcal{Z}$, with a compact subset $\mathcal{H}$ of hypotheses. There is also an auxiliary vector space $\mathcal{W}$, and a function $F : \mathcal{A} \times \mathcal{Z} \times \mathbb{R}^{\mathcal{O}} \to \mathcal{W}$ that is bilinear in $\mathcal{Z}$ and $\mathbb{R}^{\mathcal{O}}$. F specifies how points in $\mathcal{Z}$ correspond to models. A point $z \in \mathcal{Z}$ corresponds to the model defined by

$$a \mapsto \{ \mu \in \Delta \mathcal{O} \mid F(a, z, \mu) = 0 \}$$

Kosoy's regret bound for this setting depended on several unusual quantities which are absent from the classical theory of linear bandits. R is a parameter which acts as a generalized condition number, and sine is another parameter which measures the angle between the walls of the probability simplex, and the set $\{ \mu \in \Delta \mathcal{O} \mid F(a, z, \mu) = 0 \}$.

If Z, W are the dimensions of the spaces $\mathcal{Z}$ and $\mathcal{W}$, then the previous regret bound for this setting was of the form $\tilde{\mathcal{O}}\left(Z^2(\frac{1}{\text{sine}} + 1)R\sqrt{WT}\right)$, which does not depend on $\mathcal{A}$ or $\mathcal{O}$.

To improve on this result, we upper-bound the fuzzy DEC and the ε-covering number of $\mathcal{H}$.

Theorem 3 *In the robust linear bandit setting, for all $\varepsilon < \frac{1}{e^2}$ (Euler's constant),*

$$dec_\varepsilon^f(\mathcal{H}) \leq 16 \left(\frac{1}{sine} + 1\right) R\sqrt{WZ}\varepsilon \ln\left(\frac{1}{\varepsilon}\right)$$

Proposition 6 *For the robust linear bandit setting,* $\mathcal{N}(\mathcal{H}, \varepsilon) \leq \left(\frac{4\left(\frac{1}{sine}+1\right)RZ}{\varepsilon^2} + 1\right)^Z$

The Minkowski-Bougliand dimension of $\mathcal{H}$ is then $2Z$ or less, so by Proposition 5 there is an estimator with $\beta_{\widehat{M}}(T, T^{-1/2}) \in \mathcal{O}(Z \log(T))$, and $\alpha_{\widehat{M}}(T, T^{-1/2}) \in \mathcal{O}(\sqrt{T \log(T)})$. Such an estimator, along with Theorems 1 and 3, show that a regret of $\tilde{\mathcal{O}}\left(Z\left(\frac{1}{\text{sine}} + 1\right)R\sqrt{WT}\right)$ is attainable in this setting, matching the classical linear bandit regret bound which scales linearly with the dimension of the hypothesis space.

4.2. Tabular Episodic RMDP Learning

A Robust Markov Decision Process (RMDP) may be thought of as an MDP where the transition kernel $\mathbb{M}$ produces an imprecise belief over the next state and reward, consisting of the distributions which the environment might select.

We work in the non-stationary tabular setting where the states $\mathcal{S}$ and actions $\mathcal{A}$ are finite sets, H is the time horizon, and the transition kernel $\mathbb{M}$ has type $\{0, ..., H\} \times \mathcal{S} \times \mathcal{A} \to \square([0,1] \times \mathcal{S})$. The initial state, initial action, and terminal state s_0, a_0, s_{H+1} are considered to be unique, so trajectories are of the form $r_0, s_1, a_1, r_1, ..., r_H$, which lets us specify an RMDP by the transition kernel alone.

To phrase episodic RMDP learning in our framework, we consider each episode to be a single interaction with the environment, so T is the number of episodes. The action for an episode is the choice of policy. We only consider randomized nonstationary policies, also called Markov policies. This is the space $\{1, ..., H\} \times \mathcal{S} \to \Delta\mathcal{A}$, also denoted as Π_{RNS}. The observation for an episode is the trajectory. To convert an RMDP to a model of type $\Pi_{\text{RNS}} \to \square\left([0,1] \times (\mathcal{S} \times \mathcal{A} \times [0,1])^{\{1,...,H\}}\right)$, we introduce the following definition.

Definition 7 (Selection of an RMDP) *A function $\sigma : (\mathcal{S} \times \mathcal{A} \times [0,1])^{\leq H} \times \mathcal{S} \times \mathcal{A} \to \Delta([0,1] \times \mathcal{S})$ is a selection of an RMDP $\mathbb{M}$ if, for all $h, tr_{<h}, s, a$, we have $\sigma(tr_{<h}, s, a) \in \mathbb{M}(h, s, a)$. This is denoted by $\sigma \models \mathbb{M}$.*

The selections of $\mathbb{M}$ are the possible behaviors of the environment. Given a selection σ, and a policy π, $\sigma \bowtie \pi$ denotes the distribution over trajectories produced by σ interacting with π. The model associated with an RMDP $\mathbb{M}$ is then $\pi \mapsto \{\sigma \bowtie \pi \mid \sigma \models \mathbb{M}\}$. We do not require that the sum of rewards in an episode be bounded in $[0, 1]$, although we impose an analogous requirement.

Definition 8 (1-Bounded RMDP) *An RMDP $\mathbb{M}$ is 1-bounded if there is some $\sigma \models \mathbb{M}$ where, for all π, h, s, a, we have $\mathbb{E}_{\sigma \bowtie \pi} \left[\sum_{k=h}^{H} r_k \mid s_h = s, a_h = a \right] \leq 1$*

Our hypothesis class $\mathcal{H}$ of interest then consists of all 1-bounded RMDP's, with $\mathcal{S}, \mathcal{A}, H$ fixed. The best existing regret bound in this setting is the bound of Tian et al. (2021), which was of the form $\tilde{\mathcal{O}}(H^2 S^{1/3} A^{1/3} T^{2/3})$. Tian et al. (2021) did not assume 1-boundedness, which accounts for an extra factor of H in their regret bound. Partially answering one of Tian et al. (2021)'s open questions, we show that $\tilde{\mathcal{O}} \left(\sqrt{\text{poly}(H, S, A) T} \right)$ regret is attainable in this specialization of their setting.

To establish this we use Theorem 1 and a non-standard notion of loss detailed in Appendix H. The term "modified fuzzy DEC" and notation $\text{dec}_\varepsilon^{f'}$ denotes the fuzzy DEC with this new notion of loss.

In Section 5.2 of Foster et al. (2021), the DEC was upper-bounded in an analogous classical setting by an algorithm which randomized between Markov policies. We prove a comparable result in a more general setting, by showing that a *single* Markov policy can certify that the modified fuzzy DEC is low.

Theorem 4 *In the episodic RMDP setting, if $\overline{M} : \Pi_{RNS} \to \Delta([0,1] \times (\mathcal{S} \times \mathcal{A} \times [0,1])^{\{1,\ldots,H\}})$ is continuous and policy-coherent[13], then $\text{dec}_\varepsilon^{f'}(\mathcal{H}, \overline{M}) \leq 2\sqrt{2(HSA+1)\varepsilon}$*

We now turn to constructing an estimation oracle for $\mathcal{H}$. Unfortunately, this hypothesis class is infinite-dimensional. To address this, we retreat to a smaller hypothesis class $\mathcal{H}_{\text{parhalf}}$, of Partial Halfspace RMDP's[14], where estimation is tractable. The general insights of the RUE algorithm let us construct a custom estimator for $\mathcal{H}_{\text{parhalf}}$ with the properties we need, in Appendix L.

Theorem 5 *There is an online estimator $\widehat{M}$ for hypotheses in $\mathcal{H}_{parhalf}$ where $\beta_{\widehat{M}}(T, \delta) \in \mathcal{O}\left(HS^2 \log(T) + HS \log\left(\frac{HSAT}{\delta}\right) \right)$ and $\alpha_{\widehat{M}}(T, \delta) \in \mathcal{O}\left(H\sqrt{T} \log\left(\frac{1}{\delta}\right) \right)$.*

Proposition 7 *For the estimator of Theorem 5, all estimates $\widehat{M}_t : \Pi_{RNS} \to \Delta([0,1] \times (\mathcal{S} \times \mathcal{A} \times [0,1])^{\{1,\ldots,H\}})$ are continuous and policy-coherent.*

Combining these results with Theorem 1, we have an algorithm which attains $\tilde{\mathcal{O}}\left(\sqrt{H^2 S^3 A T} \right)$ regret on all 1-bounded RMDP's in $\mathcal{H}_{\text{parhalf}}$. The final insight is that every $\mathbb{M} \in \mathcal{H}$ has a 1-bounded "surrogate" $\mathbb{M}'$ within $\mathcal{H}_{\text{parhalf}}$. $\mathbb{M}'$ will be a true model if $\mathbb{M}$ is, and low regret on $\mathbb{M}'$ implies low regret on $\mathbb{M}$, so an algorithm with low regret on $\mathcal{H}_{\text{parhalf}}$ has low regret on all of $\mathcal{H}$.

Corollary 3 *There is an algorithm which attains $\tilde{\mathcal{O}}\left(\sqrt{H^2 S^3 A T} \right)$ regret on all 1-bounded RMDP's in the episodic RMDP setting.*

---

13. A belief $\overline{M}$ is policy-coherent if, for all π, there exists some σ such that $\overline{M}(\pi) = \sigma \bowtie \pi$. Policies must be mapped to distribution on trajectories which could have been produced by that policy.

14. All Partial Halfspace RMDP's have the following form. For each h, s, there is a recommended action $a_{h,s} : \mathcal{A}$, function $f_{h,s} : \mathcal{S} \to [0,1]$, and constant $c_{h,s} : [0,1]$. $\mathbb{M}(h,s,a)$ for non-recommended a is $\Delta([0,1] \times \mathcal{S})$, so the RMDP is "partial" by only predicting one action. $\mathbb{M}(h,s,a_{h,s}) = \{\mu \in \Delta([0,1] \times \mathcal{S}) \mid \mathbb{E}_\mu[f_{h,s} + r] \geq c_{h,s}\}$, so the imprecise belief for a recommended action is a halfspace.

Acknowledgments

References

T. Augustin, F.P.A. Coolen, G. de Cooman, and M.C.M. Troffaes. *Introduction to Imprecise Probabilities*. Wiley Series in Probability and Statistics. Wiley, 2014. ISBN 9781118763148. URL https://books.google.co.il/books?id=sntcAwAAQBAJ.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 263–272. PMLR, 2017. URL http://proceedings.mlr.press/v70/azar17a.html.

Lilian Besson and Emilie Kaufmann. What doubling tricks can and can't do for multi-armed bandits, 2018. URL https://arxiv.org/abs/1803.06971.

Jing Dong, Jingwei Li, Baoxiang Wang, and Jingzhao Zhang. Online policy optimization for robust mdp, 2022. URL https://arxiv.org/abs/2209.13841.

Simon Du, Sham Kakade, Jason Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in rl. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2826–2836. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/du21a.html.

Dylan J. Foster, Sham M. Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *CoRR*, abs/2112.13487, 2021. URL https://arxiv.org/abs/2112.13487.

Dylan J Foster, Alexander Rakhlin, Ayush Sekhari, and Karthik Sridharan. On the complexity of adversarial decision making. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 35404–35417. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/e5daf532498ebba4fe6544d49c811678-Paper-Conference.pdf.

Dylan J. Foster, Noah Golowich, and Yanjun Han. Tight guarantees for interactive decision making with the decision-estimation coefficient. In Gergely Neu and Lorenzo Rosasco, editors, *The Thirty Sixth Annual Conference on Learning Theory, COLT 2023, 12-15 July 2023, Bangalore, India*, volume 195 of *Proceedings of Machine Learning Research*, pages 3969–4043. PMLR, 2023. URL https://proceedings.mlr.press/v195/foster23b.html.

Scott Garrabrant, Tsvi Benson-Tilsen, Andrew Critch, Nate Soares, and Jessica Taylor. A formal approach to the problem of logical non-omniscience. In Jérôme Lang, editor, Proceedings Sixteenth Conference on *Theoretical Aspects of Rationality and Knowledge*, Liverpool, UK, 24-26 July 2017, volume 251 of *Electronic Proceedings in Theoretical Computer Science*, pages 221–235. Open Publishing Association, 2017. doi: 10.4204/EPTCS.251.16.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2137–2143. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/jin20a.html.

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 13406–13418. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/6f5e4e86a87220e5d361ad82f1ebc335-Paper.pdf.

Vanessa Kosoy. Forecasting using incomplete models. *CoRR*, abs/1705.04630, 2017. URL http://arxiv.org/abs/1705.04630.

Vanessa Kosoy. Imprecise multi-armed bandits, 2024. URL https://arxiv.org/abs/2405.05673.

T. Lattimore and C. Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. ISBN 9781108486828. URL https://books.google.co.il/books?id=bbjpDwAAQBAJ.

Shiau Hong Lim, Huan Xu, and Shie Mannor. Reinforcement learning in robust markov decision processes. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/0deb1c54814305ca9ad266f53bc82511-Paper.pdf.

Qinghua Liu, Yuanhao Wang, and Chi Jin. Learning Markov games with adversarial opponents: Efficient algorithms and fundamental limits. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 14036–14053. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/liu22r.html.

Ian Osband and Benjamin Van Roy. Model-based reinforcement learning and the eluder dimension. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/1141938ba2c2b13f5505d7c424ebae5f-Paper.pdf.

Kishan Panaganti and Dileep Kalathil. Sample complexity of robust reinforcement learning with a generative model. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 9582–9602. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/panaganti22a.html.

Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/41bfd20a38bb1b0bec75acf0845530a7-Paper.pdf.

Glenn Shafer, V Vovk, and A Takemura. Defensive forecasting. In *AISTATS*, volume 2005, pages 365–372. Citeseer, 2005.

Marnix Suilen, Thom Badings, Eline M. Bovy, David Parker, and Nils Jansen. Robust markov decision processes: A place where ai and formal methods meet. In Nils Jansen, Sebastian Junges, Benjamin Lucien Kaminski, Christoph Matheja, Thomas Noll, Tim Quatmann, Mariëlle Stoelinga, and Matthias Volk, editors, *Principles of Verification: Cycling the Probabilistic Landscape : Essays Dedicated to Joost-Pieter Katoen on the Occasion of His 60th Birthday, Part III*, pages 126–154. Springer Nature Switzerland, Cham, 2025. ISBN 978-3-031-75778-5. doi: 10.1007/978-3-031-75778-5_7. URL https://doi.org/10.1007/978-3-031-75778-5_7.

Yi Tian, Yuanhao Wang, Tiancheng Yu, and Suvrit Sra. Online learning in unknown markov games. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 10279–10288. PMLR, 2021. URL http://proceedings.mlr.press/v139/tian21b.html.

V. G. Vovk. A game of prediction with expert advice. In *Proceedings of the Eighth Annual Conference on Computational Learning Theory*, COLT '95, page 51–60, New York, NY, USA, 1995. Association for Computing Machinery. ISBN 0897917235. doi: 10.1145/225298.225304. URL https://doi.org/10.1145/225298.225304.

Ruosong Wang, Russ R Salakhutdinov, and Lin Yang. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6123–6135. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/440924c5948e05070663f88e69e8242b-Paper.pdf.

Yue Wang and Shaofeng Zou. Online robust reinforcement learning with model uncertainty. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 7193–7206. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/3a4496776767aaa99f9804d0905fe584-Paper.pdf.

Qiaomin Xie, Yudong Chen, Zhaoran Wang, and Zhuoran Yang. Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 3674–3682. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/xie20a.html.

Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent Bellman Error. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10978–10989. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/zanette20a.html.

Appendix A. Basic Propositions

Proposition 1 *If $\mathcal{H}$ is a compact subset of $\mathcal{A} \to \Box\mathcal{O}$, then*

$$dec_{\varepsilon}^{f}(\mathcal{H}, \overline{M}) = \min_{\gamma \geq 0}\left(\max(dec_{\gamma}^{o}(\mathcal{H}, \overline{M}), 0) + \gamma\varepsilon^2\right)$$

Unpack definitions and reexpress the maximization.

$$dec_{\varepsilon}^{f}(\mathcal{H}, \overline{M})$$

$$= \min_{p \in \Delta\mathcal{A}} \max_{\mu \in \Delta^s\mathcal{H}} \left\{ \mathop{\mathbb{E}}_{M, a \sim \mu, p}\left[\max(f^M) - f^{\overline{M}}(a)\right] \ \middle| \ \mathop{\mathbb{E}}_{M, a \sim \mu, p}\left[D_H^2(\overline{M}(a) \to M(a))\right] \leq \varepsilon^2 \right\}$$

$$= \min_{p \in \Delta\mathcal{A}} \max_{\mu \in \Delta^s\mathcal{H}} \min_{\gamma \geq 0}\left(\mathop{\mathbb{E}}_{\mu, p}\left[\max(f^M) - f^{\overline{M}}\right] - \gamma\left(\mathop{\mathbb{E}}_{\mu, p}\left[D_H^2(\overline{M}(a) \to M(a))\right] - \varepsilon^2\right)\right)$$

If $\mathcal{H}$ is a compact subspace of $\mathcal{A} \to \Box\mathcal{O}$, we can use Sion's Minimax Theorem.

$$= \min_{p \in \Delta\mathcal{A}} \min_{\gamma \geq 0} \max_{\mu \in \Delta^s\mathcal{H}}\left(\mathop{\mathbb{E}}_{\mu, p}\left[\max(f^M) - f^{\overline{M}}\right] - \gamma\left(\mathop{\mathbb{E}}_{\mu, p}\left[D_H^2(\overline{M}(a) \to M(a))\right] - \varepsilon^2\right)\right)$$

Now, swap the two mins, distribute the γ, regroup parentheses, and reshuffle expectations.

$$= \min_{\gamma \geq 0}\left(\min_{p \in \Delta\mathcal{A}} \max_{\mu \in \Delta^s\mathcal{H}} \mathbb{E}_\mu\left[\max(f^M) - \mathbb{E}_p\left[f^{\overline{M}}\right] - \gamma\mathbb{E}_p\left[D_H^2(\overline{M}(a) \to M(a))\right]\right]\right) + \gamma\varepsilon^2$$

For the maximization over sub-probability distributions, the maximizer will be either 100 percent probability on a specific model, or the zero measure.

$$= \min_{\gamma \geq 0}\left(\min_{p \in \Delta\mathcal{A}} \max\left(0, \max_{M \in \mathcal{H}}\left(\max(f^M) - \mathbb{E}_p\left[f^{\overline{M}}\right] - \gamma\mathbb{E}_p\left[D_H^2(\overline{M}(a) \to M(a))\right]\right)\right)\right) + \gamma\varepsilon^2$$

$$= \min_{\gamma \geq 0} \max\left(0, \min_{p \in \Delta\mathcal{A}} \max_{M \in \mathcal{H}}\left(\max(f^M) - \mathbb{E}_p\left[f^{\overline{M}}\right] - \gamma\mathbb{E}_p\left[D_H^2(\overline{M}(a) \to M(a))\right]\right)\right) + \gamma\varepsilon^2$$

$$= \min_{\gamma \geq 0}\left(\max(dec_{\gamma}^{o}(\mathcal{H}, \overline{M}), 0) + \gamma\varepsilon^2\right)$$

∎

Proposition 2 *For every* $\overline{M} : \mathcal{A} \to \square\mathcal{O}$, *there is a probabilistic model* $\overline{\theta} : \mathcal{A} \to \Delta\mathcal{O}$ *consistent with* $\overline{M}$ *where* $dec_\varepsilon^f(\mathcal{H}, \overline{M}) \le dec_\varepsilon^f(\mathcal{H}, \overline{\theta})$

Let $\overline{\theta}(a) := \underset{\mu\in\overline{M}(a)}{\operatorname{argmin}} f^\mu(a)$. This is the probabilistic model which attains minimal expected reward while being consistent with $\overline{M}$. Unpacking the definition of the fuzzy DEC, we have

$$\mathrm{dec}_\varepsilon^f(\mathcal{H}, \overline{M})$$

$$= \min_{p\in\Delta\mathcal{A}} \max_{\mu\in\Delta^s\mathcal{H}} \left\{ \underset{M,a\sim\mu,p}{\mathbb{E}}\left[\max(f^M) - f^{\overline{M}}(a)\right] \;\middle|\; \underset{M,a\sim\mu,p}{\mathbb{E}}\left[D_H^2(\overline{M}(a) \to M(a))\right] \le \varepsilon^2 \right\}$$

Swapping out $\overline{M}$ for $\overline{\theta}$ does not affect the regret terms, because $f^{\overline{M}}(a) = \mathbb{E}_{\overline{M}(a)}[r] = \mathbb{E}_{\overline{\theta}(a)}[r] = f^{\overline{\theta}}(a)$. However, swapping out the set $\overline{M}(a)$ for the point $\overline{\theta}(a)$ can only decrease the greatest distance from the beliefs to M, for all M. Because the Hellinger distances all go down, more sub-probability distributions $\mu \in \Delta^s\mathcal{H}$ count as near-mimics of our beliefs, so the maximum expected regret over near-mimics can only go up. So, we have

$$\le \min_{p\in\Delta\mathcal{A}} \max_{\mu\in\Delta^s\mathcal{H}} \left\{ \underset{M,a\sim\mu,p}{\mathbb{E}}\left[\max(f^M) - f^{\overline{\theta}}(a)\right] \;\middle|\; \underset{M,a\sim\mu,p}{\mathbb{E}}\left[D_H^2(\overline{\theta}(a) \to M(a))\right] \le \varepsilon^2 \right\} = \mathrm{dec}_\varepsilon^f(\mathcal{H}, \overline{\theta})$$

∎

Appendix B. Upper Bounds

Theorem 1 *For all* $T, \delta, \mathcal{L}$, *oracles* $\widehat{M}$, *and models* $M \in \mathcal{H}$, *if* $\varepsilon = \sqrt{\frac{\beta_{\widehat{M}}(T,\delta)}{T}}$, *we have*

$$\textbf{REG}(E2D, M, T) \le 2T\,dec_\varepsilon^{f,\mathcal{L}}(\mathcal{H}) + \alpha_{\widehat{M}}(T,\delta) + 2T\delta$$

$$\textbf{REG}(\text{E2D}, M, T) = \max_{\theta\models M \theta\bowtie\text{E2D}} \mathbb{E} \left[\sum_{t=1}^{T} \underset{a\sim\text{E2D}_t}{\mathbb{E}}\left[\max(f^M) - f^{\theta_t}(a)\right] \right]$$

Fix a maximizing environment θ which is consistent with M, and reexpress the regret. Abbreviate E2D_t as p_t.

$$= \underset{\theta\bowtie\text{E2D}}{\mathbb{E}} \left[\sum_{t=1}^{T} \underset{a\sim p_t}{\mathbb{E}}\left[\max(f^M) - f^{\theta_t}(a)\right] \right]$$

$$= \underset{\theta\bowtie\text{E2D}}{\mathbb{E}} \left[\sum_{t=1}^{T} \max(f^M) - \mathbb{E}_{p_t}\left[f^{\widehat{M_t}}\right] + \mathbb{E}_{p_t}\left[f^{\widehat{M_t}} - f^{\theta_t}\right] \right]$$

M is falsely rejected from the sequence of hypotheses $\mathcal{H}_t$ iff $\sum_{t=1}^{T} \underset{a\sim p_t}{\mathbb{E}}\left[\mathcal{L}(\widehat{M_t}, M, a)\right] > \beta_{\widehat{M}}(T,\delta)$, and by the definition of $\beta_{\widehat{M}}$ with respect to $\mathcal{L}$, this only happens with $\le \delta$ probability. Similarly, there is a $\le \delta$ probability of the $\alpha_{\widehat{M}}$ bound failing to apply. So, we may upper-bound the expected regret by adding $2\delta T$ to account for

the regret if a failure event occurs, and passing to a history h where no failure events occur but the regret is otherwise as high as possible.

$$\leq \sum_{t=1}^{T} \left(\max(f^M) - \mathbb{E}_{p_t}\left[f^{\widehat{M}_t} \right] + \mathbb{E}_{p_t}\left[f^{\widehat{M}_t} - f^{\theta_t} \right] \right) + 2\delta T$$

The $\alpha_{\widehat{M}}$ bound holds for this history, so we may rewrite as

$$\leq \sum_{t=1}^{T} \left(\max(f^M) - \mathbb{E}_{p_t}\left[f^{\widehat{M}_t} \right] \right) + \alpha_{\widehat{M}}(T, \delta) + 2\delta T$$

Let ε be defined as $\sqrt{\frac{\beta_{\widehat{M}}(T,\delta)}{T}}$, and $\mu_t \in \Delta^s \mathcal{H}_t$ be defined as follows. If $\mathbb{E}_{p_t}\left[\mathcal{L}(\widehat{M}_t, M, a) \right] \leq \varepsilon^2$, then μ_t assigns M a probability of 1. If not, μ_t assigns M a probability of $\frac{\varepsilon^2}{\mathbb{E}_{p_t}\left[\mathcal{L}(\widehat{M}_t, M, a) \right]}$. We always have $\mu_t \in \Delta^s \mathcal{H}_t$ because M isn't falsely rejected from any $\mathcal{H}_t$. We can now rewrite as

$$= \sum_{t=1}^{T} \max \left(1, \frac{\mathbb{E}_{p_t}\left[\mathcal{L}(\widehat{M}_t, M, a) \right]}{\varepsilon^2} \right) \underset{M' \sim \mu_t}{\mathbb{E}} \left[\max(f^{M'}) - \mathbb{E}_{p_t}\left[f^{\widehat{M}_t} \right] \right] + \alpha_{\widehat{M}}(T, \delta) + 2\delta T$$

By the definition of the E2D algorithm, p_t is the fuzzy-DEC (with respect to $\mathcal{L}$)-minimizing distribution on actions. Also, μ_t was defined to ensure that $\underset{\mu_t, p_t}{\mathbb{E}} \left[\mathcal{L}(\widehat{M}_t, M, a) \right] \leq \varepsilon^2 = \frac{\beta_{\widehat{M}}(T,\delta)}{T}$, so we can upper-bound with the fuzzy DEC with respect to $\mathcal{L}$.

$$\leq \sum_{t=1}^{T} \max \left(1, \frac{\mathbb{E}_{p_t}\left[\mathcal{L}(\widehat{M}_t, M, a) \right]}{\varepsilon^2} \right) \mathrm{dec}_{\varepsilon}^{f, \mathcal{L}}(\mathcal{H}_t, \widehat{M}_t) + \alpha_{\widehat{M}}(T, \delta) + 2\delta T$$

Now, $\mathcal{H}_t \subseteq \mathcal{H}$. The DEC is larger for larger hypotheses classes, and worst-casing over estimates makes it larger still.

$$\leq \sum_{t=1}^{T} \max \left(1, \frac{\mathbb{E}_{p_t}\left[\mathcal{L}(\widehat{M}_t, M, a) \right]}{\varepsilon^2} \right) \mathrm{dec}_{\varepsilon}^{f, \mathcal{L}}(\mathcal{H}) + \alpha_{\widehat{M}}(T, \delta) + 2\delta T$$

Upper bound $\max(a, b)$ by $a + b$, and use that $\varepsilon = \sqrt{\frac{\beta_{\widehat{M}}(T,\delta)}{T}}$.

$$\leq \sum_{t=1}^{T} \left(1 + \frac{T}{\beta_{\widehat{M}}(T, \delta)} \right) \mathbb{E}_{p_t}\left[\mathcal{L}(\widehat{M}_t, M, a) \right] \mathrm{dec}_{\varepsilon}^{f, \mathcal{L}}(\mathcal{H}) + \alpha_{\widehat{M}}(T, \delta) + 2\delta T$$

On this history, no failure events occur, so the sum of the expected loss is upper-bounded by $\beta_{\widehat{M}}(T, \delta)$, which cancels, yielding

$$\leq 2T \mathrm{dec}_{\varepsilon}^{f, \mathcal{L}}(\mathcal{H}) + \alpha_{\widehat{M}}(T, \delta) + 2T\delta$$

∎

Corollary 1 *If there is a $p > 0, q > 0, r < 1, s < 1$ and online estimator $\widehat{M}$ such that:*

$$\limsup_{\varepsilon \to 0} \frac{dec_\varepsilon^f(\mathcal{H})}{\varepsilon^p} < \infty, \limsup_{T \to \infty} \frac{\alpha_{\widehat{M}}(T, T^{-q})}{T^r} < \infty, \limsup_{T \to \infty} \frac{\beta_{\widehat{M}}(T, T^{-q})}{T^s} < \infty$$

then the E2D algorithm with $\widehat{M}$ as an oracle has $\mathcal{O}\left(T^{\max\left(1 - \frac{p(1-s)}{2}, r, 1-q\right)}\right)$ regret on all $M \in \mathcal{H}$.

The regret bound produced by the E2D algorithm is $\mathcal{O}(T\mathrm{dec}_\varepsilon^f(\mathcal{H}) + \alpha_{\widehat{M}}(T, \delta) + T\delta)$ by Theorem 1. Now, let δ scale as T^{-q}, and use that $\alpha(T, T^{-q}) \in \mathcal{O}(T^r)$, and $\mathrm{dec}_\varepsilon^f(\mathcal{H}) \in \mathcal{O}(\varepsilon^p)$, to get a regret on the order of $\mathcal{O}(T\varepsilon^p + T^r + T^{1-q})$. Now, use the specific choice of ε from the E2D algorithm to show

$$\varepsilon = \sqrt{\frac{\beta_{\widehat{M}}(T, \delta)}{T}} = \sqrt{\frac{\beta_{\widehat{M}}(T, T^{-q})}{T}} \in \mathcal{O}\left(\sqrt{\frac{T^s}{T}}\right) = \mathcal{O}\left(T^{\frac{1}{2}(s-1)}\right)$$

Plugging this value of ε in, we get that the regret of the E2D algorithm is on the order of

$$\mathcal{O}\left(T^{1 + \frac{p}{2}(s-1)} + T^r + T^{1-q}\right) = \mathcal{O}\left(T^{\max\left(1 - \frac{p(1-s)}{2}, r, 1-q\right)}\right)$$

∎

Appendix C. Lower Bounds

Proposition 3 *For all $p \in (0, 1)$ such that $\liminf_{\varepsilon \to 0} \frac{dec_\varepsilon^f(\mathcal{H})}{\varepsilon^p} = \infty$, if $\varepsilon_T = \sqrt{\frac{1}{T \ln(T)}}$, we have*

$$\min_\pi \max_{M \in \mathcal{H}} (\boldsymbol{REG}(\pi, M, T)) \in \Omega(T \, dec_{\varepsilon_T^{\frac{1}{1-p}}}^f(\mathcal{H}))$$

For this proof, define the following quantities, which may depend on $\mathcal{H}, \pi, T, p$.

$u(T)$ is the uniform distribution on timesteps. $\overline{M} : \mathcal{A} \to \Delta\mathcal{O}$ is the belief which maximizes $\mathrm{dec}_{\varepsilon_T^{\frac{1}{1-p}}}^f(\mathcal{H}, \overline{M})$. The fuzzy DEC will never be zero, or else we would get a contradiction with the assumption that the DEC shrinks slower than ε^p for some $p \in (0, 1)$. Proposition 2 enables us to assume that $\overline{M}$ is probabilistic with no loss of generality.

Given a model $M \in \mathcal{H}$, we can define $\theta^M : \mathcal{A} \to \Delta\mathcal{O}$ as M's closest approximation to $\overline{M}$ in Hellinger distance.

$$\theta^M(a) := \underset{\mu \in M(a)}{\mathrm{argmin}} D_H^2(\mu, \overline{M}(a))$$

p^M and $\overline{p}$, of type $\Delta\mathcal{A}$, are the distributions over actions produced by the policy π interacting with θ^M or $\overline{M}$ respectively, where θ^M and $\overline{M}$ are treated as environments. In the definition, δ_{a_t} is the probability distribution which assigns all measure to action a_t. $\overline{p}$ is defined similarly.

$$p^M := \underset{\theta^M \bowtie \pi}{\mathbb{E}} [\underset{t \sim u(T)}{\mathbb{E}} [\delta_{a_t}]]$$

μ' is the sub-probability distribution on $\mathcal{H}$ which maximizes regret against $\overline{p}$, while being a near-mimic of $\overline{M}$. More specifically, it is

$$\mu' := \underset{\mu \in \Delta^s \mathcal{H}}{\operatorname{argmax}} \left\{ \underset{\mu, \overline{p}}{\mathbb{E}} \left[\max(f^M) - f^{\overline{M}} \right] \;\middle|\; \underset{\mu, \overline{p}}{\mathbb{E}} \left[D_H^2(\overline{M}(a) \to M(a)) \right] \leq \left(\varepsilon_T^{\frac{1}{1-p}} \right)^2 \right\}$$

μ' cannot be the all-zero sub-probability distribution, because if it was, that would witness that the fuzzy DEC for $\overline{M}$ is zero, which is impossible. So, μ' can be uniquely written as $\lambda\mu$, where $\lambda \in (0, 1]$, and $\mu \in \Delta\mathcal{H}$. We now proceed with the proof. Fix an arbitrary π, and we will lower-bound the maximum regret. Start by replacing the max over M with an expectation, and chose θ^M for each M.

$$\max_{M \in \mathcal{H}}(\mathbf{REG}(\pi, M, T)) = \max_{M \in \mathcal{H}, \theta \models M\theta \bowtie \pi} \mathbb{E} \left[\sum_{t=1}^{T} \max(f^M) - r_t \right]$$

$$\geq \underset{M \sim \mu}{\mathbb{E}} \left[\underset{\theta^M \bowtie \pi}{\mathbb{E}} \left[\sum_{t=1}^{T} \max(f^M) - r_t \right] \right]$$

Now, rewrite the sum as T times an expectation over the uniform distribution, and pull the T and $\max(f^M)$ out. Then swap the expectations, and expand the expectation over histories as an expectation over partial histories up to a_t, and an expectation of what r_t will be.

$$= T \cdot \mathbb{E}_\mu \left[\max(f^M) - \underset{\theta^M \bowtie \pi}{\mathbb{E}} \left[\underset{t \sim u(T)}{\mathbb{E}} [r_t] \right] \right] = T \cdot \mathbb{E}_\mu \left[\max(f^M) - \mathbb{E}_{u(T)} \left[\underset{\theta^M \bowtie \pi}{\mathbb{E}} \left[\underset{\theta^M(a_t)}{\mathbb{E}} [r] \right] \right] \right]$$

Rewrite the inner expectation as $f^{\theta^M}(a)$, and interchange expectations again.

$$= T \cdot \mathbb{E}_\mu \left[\max(f^M) - \mathbb{E}_{u(T)} \left[\underset{\theta^M \bowtie \pi}{\mathbb{E}} \left[f^{\theta^M}(a_t) \right] \right] \right] = T \cdot \mathbb{E}_\mu \left[\max(f^M) - \underset{\theta^M \bowtie \pi}{\mathbb{E}} \left[\mathbb{E}_{u(T)} \left[f^{\theta^M}(a_t) \right] \right] \right]$$

Note that θ^M doesn't depend on the history, just the action, and the action was generated by sampling a random history and random timestep, which can be viewed as sampling from p^M.

$$= T \cdot \mathbb{E}_\mu \left[\max(f^M) - \underset{a \sim p^M}{\mathbb{E}} \left[f^{\theta^M}(a) \right] \right]$$

$$= T \left(\underset{\mu, \overline{p}}{\mathbb{E}} \left[\max(f^M) - f^{\overline{M}} \right] + \underset{\mu, \overline{p}}{\mathbb{E}} \left[f^{\overline{M}} - f^{\theta^M} \right] + \mathbb{E}_\mu \left[\mathbb{E}_{\overline{p}} \left[f^{\theta^M} \right] - \mathbb{E}_{p^M} \left[f^{\theta^M} \right] \right] \right)$$

This can be lower-bounded by the first term, minus the absolute value of the second two terms. Then move the absolute value in.

$$\geq T \left(\underset{\mu, \overline{p}}{\mathbb{E}} \left[\max(f^M) - f^{\overline{M}} \right] - \left| \underset{\mu, \overline{p}}{\mathbb{E}} \left[f^{\overline{M}} - f^{\theta^M} \right] \right| - \left| \mathbb{E}_\mu \left[\mathbb{E}_{\overline{p}} \left[f^{\theta^M} \right] - \mathbb{E}_{p^M} \left[f^{\theta^M} \right] \right] \right| \right)$$

$$\geq T \left(\underset{\mu, \overline{p}}{\mathbb{E}} \left[\max(f^M) - f^{\overline{M}} \right] - \underset{\mu, \overline{p}}{\mathbb{E}} \left[\left| f^{\overline{M}} - f^{\theta^M} \right| \right] - \mathbb{E}_\mu \left[\left| \mathbb{E}_{\overline{p}} \left[f^{\theta^M} \right] - \mathbb{E}_{p^M} \left[f^{\theta^M} \right] \right| \right] \right)$$

21

These terms can be bounded by the total variation distance between $\overline{M}(a)$ and $\theta^M(a)$, and the total variation distance between $\bar{p}$ and p^M, because the expected rewards are in $[0, 1]$.

$$\geq T \left(\mathop{\mathbb{E}}_{\mu,\bar{p}} \left[\max(f^M) - f^{\overline{M}} \right] - \mathop{\mathbb{E}}_{\mu,\bar{p}} \left[D_{TV}(\overline{M}(a), \theta^M(a)) \right] - \mathbb{E}_\mu \left[D_{TV}(\bar{p}, p^M) \right] \right)$$

By the data processing inequality, $D_{TV}(\bar{p}, p^M)$ can be upper-bounded by the total variation distance between $\overline{M} \bowtie \pi$ and $\theta^M \bowtie \pi$. Then use that total variation distance is upper-bounded by $\sqrt{2}$ times the Hellinger distance.

$$\geq T \left(\mathop{\mathbb{E}}_{\mu,\bar{p}} \left[\max(f^M) - f^{\overline{M}} \right] - \mathop{\mathbb{E}}_{\mu,\bar{p}} \left[D_{TV}(\overline{M}(a), \theta^M(a)) \right] - \mathbb{E}_\mu \left[D_{TV}(\overline{M} \bowtie \pi, \theta^M \bowtie \pi) \right] \right)$$

$$\geq T \left(\mathop{\mathbb{E}}_{\mu,\bar{p}} \left[\max(f^M) - f^{\overline{M}} \right] - \mathop{\mathbb{E}}_{\mu,\bar{p}} \left[\sqrt{2D_H^2(\overline{M}(a), \theta^M(a))} \right] - \mathbb{E}_\mu \left[\sqrt{2D_H^2(\overline{M} \bowtie \pi, \theta^M \bowtie \pi)} \right] \right)$$

By Lemma A.13 of Foster et al. (2021), that latter Hellinger-squared term can be upper bounded by $100T \ln(T)$ times the expected (under $\bar{p}$) Hellinger distance between $\overline{M}(a)$ and $\theta^M(a)$. Then apply concavity of square root.

$$\geq T(\mathop{\mathbb{E}}_{\mu,\bar{p}} \left[\max(f^M) - f^{\overline{M}} \right] - \mathop{\mathbb{E}}_{\mu,\bar{p}} \left[\sqrt{2D_H^2(\overline{M}(a), \theta^M(a))} \right]$$

$$-\mathbb{E}_\mu \left[\sqrt{200T \ln(T) \mathbb{E}_{\bar{p}} \left[D_H^2(\overline{M}(a), \theta^M(a)) \right]} \right])$$

$$\geq T(\mathop{\mathbb{E}}_{\mu,\bar{p}} \left[\max(f^M) - f^{\overline{M}} \right] - \sqrt{2 \mathop{\mathbb{E}}_{\mu,\bar{p}} \left[D_H^2(\overline{M}(a), \theta^M(a)) \right]}$$

$$-\sqrt{200T \ln(T) \mathop{\mathbb{E}}_{\mu,\bar{p}} \left[D_H^2(\overline{M}(a), \theta^M(a)) \right]})$$

Rewrite the expectation over μ (a distribution) as an expectation over μ' times $\frac{1}{\lambda}$, then use that $\theta^M(a)$ was picked to minimize Hellinger distance to $\overline{M}(a)$, so $D_H^2(\overline{M}(a), \theta^M(a)) = D_H^2(\overline{M}(a) \to M(a))$.

$$\geq T(\frac{1}{\lambda} \mathop{\mathbb{E}}_{\mu',\bar{p}} \left[\max(f^M) - f^{\overline{M}} \right] - \sqrt{\frac{2}{\lambda} \mathop{\mathbb{E}}_{\mu',\bar{p}} \left[D_H^2(\overline{M}(a) \to M(a)) \right]}$$

$$-\sqrt{\frac{200T \ln(T)}{\lambda} \mathop{\mathbb{E}}_{\mu',\bar{p}} \left[D_H^2(\overline{M}(a) \to M(a)) \right]})$$

By how $\overline{M}$ and μ' were constructed, the expectation of the regret gap exceeds $\mathrm{dec}^f_{\frac{1}{\varepsilon_T^{1-p}}}(\mathcal{H})$ and the Hellinger expectation is less than $\varepsilon_T^{\frac{2}{1-p}}$. Then use that $\varepsilon_T = (T \ln(T))^{-1/2}$, so $T \ln(T) = \varepsilon_T^{-2}$.

$$\geq T \left(\frac{1}{\lambda} \mathrm{dec}^f_{\frac{1}{\varepsilon_T^{1-p}}}(\mathcal{H}) - \sqrt{\frac{2}{\lambda} \varepsilon_T^{\frac{2}{1-p}}} - \sqrt{\frac{200T \ln(T)}{\lambda} \varepsilon_T^{\frac{2}{1-p}}} \right)$$

$$= T \left(\frac{1}{\lambda} \mathrm{dec}^f_{\varepsilon_T^{\frac{1}{1-p}}} (\mathcal{H}) - \sqrt{\frac{2}{\lambda} \varepsilon_T^{\frac{2}{1-p}}} - \sqrt{\frac{200}{\lambda} \varepsilon_T^{\frac{2}{1-p} - 2}} \right)$$

At this point, the value of λ is the only quantity remaining which depends on the choice of policy π. This function is convex in λ, so we will compute the minimizing value of λ, which is

$$\left(\frac{2\mathrm{dec}^f_{\varepsilon_T^{\frac{1}{1-p}}} (\mathcal{H})}{\varepsilon_T^{\frac{1}{1-p}} \left(\sqrt{2} + \sqrt{200} \frac{1}{\varepsilon_T} \right)} \right)^2$$

If this minimizer exceeds 1, then because $\lambda \in (0, 1]$, the true minimizer will be 1. We will now show that, for all sufficiently large T, this occurs. The minimizing λ is

$$\geq \left(\frac{2\mathrm{dec}^f_{\varepsilon_T^{\frac{1}{1-p}}} (\mathcal{H})}{\varepsilon_T^{\frac{1}{1-p}} \left(2\sqrt{200} \frac{1}{\varepsilon_T} \right)} \right)^2 = \left(\frac{\mathrm{dec}^f_{\varepsilon_T^{\frac{1}{1-p}}} (\mathcal{H})}{\sqrt{200} \varepsilon_T^{\frac{1}{1-p} - 1}} \right)^2 = \left(\frac{\mathrm{dec}^f_{\varepsilon_T^{\frac{1}{1-p}}} (\mathcal{H})}{\sqrt{200} \left(\varepsilon_T^{\frac{1}{1-p}} \right)^p} \right)^2$$

Since we assumed that $\liminf_{\varepsilon \to 0} \frac{dec^f_\varepsilon(\mathcal{H})}{\varepsilon^p} = \infty$, the above quantity diverges to infinity. Therefore, for sufficiently large T (which doesn't depend on the choice of π), the minimizing λ is 1. So we continue to lower bound by

$$\geq T \left(\mathrm{dec}^f_{\varepsilon_T^{\frac{1}{1-p}}} (\mathcal{H}) - \sqrt{2\varepsilon_T^{\frac{2}{1-p}}} - \sqrt{200\varepsilon_T^{\frac{2}{1-p} - 2}} \right) = T \left(\mathrm{dec}^f_{\varepsilon_T^{\frac{1}{1-p}}} (\mathcal{H}) - \sqrt{2}\varepsilon_T^{\frac{1}{1-p}} - \sqrt{200}\varepsilon_T^{\frac{1}{1-p} - 1} \right)$$

Using that $\varepsilon_T < 1$, and $\frac{1}{1-p} - 1 = \frac{p}{1-p}$, we can proceed to

$$\geq T \left(\mathrm{dec}^f_{\varepsilon_T^{\frac{1}{1-p}}} (\mathcal{H}) - 2\sqrt{200}\varepsilon_T^{\frac{1}{1-p} - 1} \right) = T \left(\mathrm{dec}^f_{\varepsilon_T^{\frac{1}{1-p}}} (\mathcal{H}) - 2\sqrt{200} \left(\varepsilon_T^{\frac{1}{1-p}} \right)^p \right)$$

Putting our inequalities together, given a p where $\liminf_{\varepsilon \to 0} \frac{dec^f_\varepsilon(\mathcal{H})}{\varepsilon^p} = \infty$, for all T above some finite threshold, every policy π has the property that

$$\max_{M \in \mathcal{H}} (\mathbf{REG}(\pi, M, T)) \geq T \left(\mathrm{dec}^f_{\varepsilon_T^{\frac{1}{1-p}}} (\mathcal{H}) - 2\sqrt{200} \left(\varepsilon_T^{\frac{1}{1-p}} \right)^p \right)$$

As T rises, $\varepsilon_T^{\frac{1}{1-p}}$ shrinks. Applying our assumption on p, in the limit, the DEC term far exceeds the $2\sqrt{200}$ term, so we have $\min_\pi \max_{M \in \mathcal{H}} (\mathbf{REG}(\pi, M, T)) \in \Omega(T\mathrm{dec}^f_{\varepsilon_T^{\frac{1}{1-p}}} (\mathcal{H}))$ as desired. ∎

Corollary 2 *If, for all $p > 0$, we have $\liminf_{\varepsilon \to 0} \frac{dec^f_\varepsilon(\mathcal{H})}{\varepsilon^p} = \infty$, then, for all $q < 1$ and π (which may depend on T), we have $\liminf_{T \to \infty} \frac{\max_{M \in \mathcal{H}} (\mathbf{REG}(\pi, M, T))}{T^q} = \infty$*

23

Let $q < 1$. This implies $\frac{2-2q}{3-2q} \in (0,1)$. Accordingly, let p be an arbitrary positive number strictly less than $\frac{2-2q}{3-2q}$. By Proposition 3, we can show

$$\liminf_{T\to\infty} \frac{\max_{M\in\mathcal{H}}(\mathbf{REG}(\pi, M, T))}{T^q} \geq \liminf_{T\to\infty} T^{1-q}\cdot\mathrm{dec}^f_{\varepsilon_T^{\frac{1}{1-p}}}(\mathcal{H})) = \liminf_{T\to\infty} T^{1-q}\varepsilon_T^{\frac{p}{1-p}} \frac{\mathrm{dec}^f_{\varepsilon_T^{\frac{1}{1-p}}}(\mathcal{H})}{\left(\varepsilon_T^{\frac{1}{1-p}}\right)^p}$$

Apply our assumption that, for every p, the liminf of that fraction diverges to infinity, and use that ε_T was defined to be $(T\ln(T))^{-1/2}$, to get

$$\geq \liminf_{T\to\infty} T^{1-q}\varepsilon_T^{\frac{p}{1-p}} = \liminf_{T\to\infty} T^{1-q}(T\ln(T))^{\frac{-p}{2(1-p)}}$$

If the exponent on T exceeds 0, then it will outgrow the $\ln(T)$ raised to a negative power, establishing our desired result that this quantity diverges to infinity. We now switch to proving $1 - q - \frac{p}{2(1-p)} > 0$. By recalling that $p < \frac{2-2q}{3-2q}$, we can compute

$$1-q-\frac{p}{2(1-p)} > 1-q-\frac{\frac{2-2q}{3-2q}}{2\left(1-\frac{2-2q}{3-2q}\right)} = 1-q-\frac{2-2q}{2((3-2q)-(2-2q))} = 1-q-(1-q) = 0$$

And our result follows. ∎

Appendix D. Hellinger Distance Lemmas

Lemma 1 *Let X be a nonempty compact convex Polish space, and $f : X \to \mathbb{R}$ be continuous and strictly convex. Then there is a unique minimizer of f.*

f is continuous, and, because X is a compact space, a minimizer exists. To prove that the minimizer is unique, assume there are two distinct minimizers, x, x'. In such a case, by the strict convexity of f, we'd have

$$f(0.5x + 0.5x') < 0.5f(x) + 0.5f(x') = \min_x f(x)$$

By convexity of X, $0.5x + 0.5x'$ is also in X, but this choice of value makes f lower than its minimum value, and we have a contradiction. Therefore, the minimizer must be unique. ∎

Lemma 2 *Let X be a nonempty compact convex Polish space, and Y be a Polish space, and $f : X \times Y \to \mathbb{R}$ be a function which is continuous in $X \times Y$, and strictly convex in X for all y. Then the function $\lambda y.argmin_{x\in X} f(x, y)$ is continuous.*

Use x_y as an abbreviation for $argmin_{x\in X} f(x, y)$. This denotes a unique point, by Lemma 1. Fix an arbitrary sequence y_n converging to y_∞ in Y. By compactness of X, the sequence x_{y_n} has at least one limit point. Fix an arbitrary limit point x_∞, and a subsequence where

the x_{y_n} converge to x_∞. Then by continuity of f, the fact that x_{y_n} is the minimizer for y_n, continuity of f, and the fact that x_{y_∞} is the minimizer for y_∞, we have

$$f(x_{y_\infty}, y_\infty) = \lim_{n\to\infty} f(x_{y_\infty}, y_n) \geq \lim_{n\to\infty} f(x_{y_n}, y_n) = f(x_\infty, y_\infty) \geq f(x_{y_\infty}, y_\infty)$$

The left and right sides are equal, so all inequalities are equalities, and we have $f(x_\infty, y_\infty) = f(x_{y_\infty}, y_\infty)$. However, by Lemma 1, minimizers are unique, so $x_\infty = x_{y_\infty}$. x_∞ was an arbitrary limit point of the x_{y_n}, so all limit points are equal, so x_{y_n} converges to x_{y_∞}. Our convergent sequence y_n was arbitrary, so the function $\lambda y.x_y$ is continuous. Unpacking the definition of x_y, this shows continuity of $\lambda y.\text{argmin}_{x\in X} f(x, y)$, as desired. ∎

Lemma 3 *Let X be a nonempty compact convex Polish space, and Y be a convex open subset of a normed vector space A, and $f : X \times Y \to \mathbb{R}$ be a function which is continuous in $X \times Y$, convex in Y for all x, strictly convex in X for all y, Frechet-differentiable in Y for all x, and (letting $df^x : Y \to A^*$ be the Frechet derivative in Y at x), has $\lambda x.df_y^x : X \to A^*$ being continuous for all y. In such a case, letting $g(y) := \min_{x\in X} f(x, y)$, g is also Frechet-differentiable, and $dg_y = df_y^{argmin_{x\in X} f(x,y)}$.*

Use the notation x_y as an abbreviation for $\text{argmin}_{x\in X} f(x, y)$. By Lemma 1, this denotes a unique point. To show Frechet-differentiability of g, and that the Frechet differential is as desired, we must show that

$$\lim_{h\to 0} \frac{g(y+h) - g(y) - df_y^{x_y}(h)}{||h||} = 0$$

The h are vectors in the space A. We will prove this by showing that the limsup is below 0 and the liminf is above 0. For the upper bound, we compute

$$\limsup_{h\to 0} \frac{g(y+h) - g(y) - df_y^{x_y}(h)}{||h||} = \limsup_{h\to 0} \frac{f(x_{y+h}, y+h) - f(x_y, y) - df_y^{x_y}(h)}{||h||}$$

$$\leq \limsup_{h\to 0} \frac{f(x_y, y+h) - f(x_y, y) - df_y^{x_y}(h)}{||h||} = 0$$

In order, this was by expanding the definition of g, using that $f(x_{y+h}, y+h) \leq f(x_y, y+h)$, and using that $df_y^{x_y}$ being the Frechet differential of f at x_y and y implies the relevant limit is zero, by the definition of the Frechet derivative. For the lower bound, we have

$$\liminf_{h\to 0} \frac{g(y+h) - g(y) - df_y^{x_y}(h)}{||h||} = \liminf_{h\to 0} \frac{f(x_{y+h}, y+h) - f(x_y, y) - df_y^{x_y}(h)}{||h||}$$

$$\geq \liminf_{h\to 0} \frac{f(x_{y+h}, y+h) - f(x_{y+h}, y) - df_y^{x_y}(h)}{||h||}$$

$$\geq \liminf_{h\to 0} \frac{f(x_{y+h}, y) + df_y^{x_{y+h}}(h) - f(x_{y+h}, y) - df_y^{x_y}(h)}{||h||}$$

$$= \liminf_{h\to 0} \frac{(df_y^{x_{y+h}} - df_y^{x_y})(h)}{||h||} \geq \liminf_{h\to 0} \frac{-||df_y^{x_{y+h}} - df_y^{x_y}|| \cdot ||h||}{||h||}$$

25

$$= \liminf_{h \to 0} -||df_y^{x_{y+h}} - df_y^{x_y}|| = 0$$

In order, this was expanding definitions, and using that $f(x_y, y) \leq f(x_{y+h}, y)$. To derive the third line, we used the convexity of f for all x, specifically the part where the linear approximation to a convex function at a point always undershoots the original convex function. Then we cancel, use that the differentials are linear functions, lower-bound with the operator norm, and cancel again. The final step uses Lemma 2, that x_y is continuous in y, to show that x_{y+h} converges to x. Combining this with the starting assumption that $\lambda x.df_y^x : X \to A^*$ is continuous for all y, $df_y^{x_{y+h}}$ converges to $df_y^{x_y}$, so the operator norm of their difference shrinks to zero.

This establishes that the Frechet derivative of g exists and equals $df_y^{x_y}$ at y. The lemma then follows because $x_y = \text{argmin}_{x \in X} f(x, y)$. ∎

Lemma 4 *Given a finite space of observations $\mathcal{O}$, a $\nu \in \Delta\mathcal{O}$ with full support, and an imprecise belief Ψ, $D_H^2(\nu \to \Psi)$ is convex and Frechet-differentiable in ν. Letting $\mu_\nu := argmin_{\mu \in \Psi} D_H^2(\mu, \nu)$, we have that $d(\lambda\nu'.D_H^2(\nu' \to \Psi))_\nu = d(\lambda\nu'.D_H^2(\mu_\nu, \nu'))_\nu$.*

We apply Lemma 3. To verify the conditions, the compact convex Polish space is Ψ, because imprecise beliefs are nonempty compact convex subsets of $\Delta\mathcal{O}$. Letting u denote the uniform measure over $\mathcal{O}$, then Y would be $\{\nu - u \mid \nu \in \Delta\mathcal{O}, \nu \text{ has full support}\}$. This is the convex open set of probability distributions with full support, but moved so that it's in a subspace of $\mathbb{R}^{\mathcal{O}}$. For Lemma 3, f would be $f(\mu, \nu - u) = D_H^2(\mu, \nu)$, and $D_H^2(\mu, \nu) = 1 - \sum_o \sqrt{\mu(o) \cdot \nu(o)}$. This is clearly seen to be continuous in both arguments, convex in its second argument (by concavity of square root), and strictly convex in its first argument (by strict concavity of square root, and $\nu(o)$ being nonzero for all observations). Now we only need to verify the Frechet-differentiability conditions on Hellinger distance to invoke Lemma 3. Computing the Frechet derivative for some fixed μ, and letting h be in $\mathbb{R}^{\mathcal{O}}$, we get

$$d(\lambda\nu - u.f(\mu, \nu - u))(h) = d\left(\lambda\nu.1 - \sum_o \sqrt{\mu(o) \cdot \nu(o)}\right)(h)$$

$$= \sum_o \frac{-1}{2\sqrt{\nu(o)}} \sqrt{\mu(o)} d(\lambda\nu.\nu(o))(h) = \sum_o \frac{-1}{2} \sqrt{\frac{\mu(o)}{\nu(o)}} h(o) = \left\langle -\frac{1}{2}\sqrt{\frac{\mu}{\nu}}, h \right\rangle$$

Because this derivative can be written as an inner product with a vector which is well-defined in all entries (because all ν are assumed to have full support), f is Frechet-differentiable in Y for all μ. Further, for all ν with full support, continuity in μ holds. All conditions have been verified, so we can apply Lemma 3 to show

$$d(\lambda\nu'.D_H^2(\nu' \to \Psi))_\nu = d(\lambda\nu'.\min_{\mu \in \Psi} D_H^2(\mu, \nu'))_\nu = d(\lambda\mu, \nu'.D_H^2(\mu, \nu'))_\nu^{\mu_\nu} = d(\lambda\nu'.D_H^2(\mu_\nu, \nu'))_\nu$$

This establishes Frechet-differentiability, but not convexity. To show convexity, we compute

$$D_H^2(p\nu + (1-p)\nu' \to \Psi) = \min_{\mu \in \Psi} D_H^2(p\nu + (1-p)\nu', \mu)$$

Using $\mu_\nu, \mu_{\nu'}$ for the Hellinger minimizers, we then use that Ψ is convex by the definition of an imprecise belief, so $p\mu_\nu + (1-p)\mu_{\nu'} \in \Psi$, then apply convexity of Hellinger-squared error in both arguments.

$$\leq D_H^2(p\nu + (1-p)\nu', p\mu_\nu + (1-p)\mu_{\nu'}) \leq pD_H^2(\nu, \mu_\nu) + (1-p)D_H^2(\nu, \mu_{\nu'})$$

$$= pD_H^2(\nu \to \Psi) + (1-p)D_H^2(\nu' \to \Psi)$$

And convexity is shown. $\blacksquare$

Lemma 5 *For imprecise beliefs Ψ, Φ, Θ, $D_H^2(\Psi \to \Theta) \leq 2D_H^2(\Psi \to \Phi) + 2D_H^2(\Phi \to \Theta)$*

First, we establish the analogous fact for probability distributions, that $D_H^2(\mu, \xi) \leq 2D_H^2(\mu, \nu) + 2D_H^2(\nu, \xi)$. Applying the triangle inequality, squaring both sides, and using the AM/GM inequality yields the result.

$$D_H^2(\mu, \xi) \leq (D_H(\mu, \nu) + D_H(\nu, \xi))^2 = D_H^2(\mu, \nu) + D_H^2(\nu, \xi) + 2D_H(\mu, \nu)D_H(\nu, \xi)$$

$$= D_H^2(\mu, \nu) + D_H^2(\nu, \xi) + 2\sqrt{D_H^2(\mu, \nu)D_H^2(\nu, \xi)} \leq 2D_H^2(\mu, \nu) + 2D_H^2(\nu, \xi)$$

Letting ν_μ be the point in Φ which is closest to μ, we can use definitions and our probabilistic result above to yield

$$D_H^2(\Psi \to \Theta) = \max_{\mu \in \Psi}\min_{\zeta \in \Theta}D_H^2(\mu, \zeta) \leq \max_{\mu \in \Psi}\min_{\zeta \in \Theta}(2D_H^2(\mu, \nu_\mu) + 2D_H^2(\nu_\mu, \zeta))$$

$$= 2\max_{\mu \in \Psi}D_H^2(\mu, \nu_\mu) + 2\min_{\zeta \in \Theta}D_H^2(\nu_\mu, \zeta) \leq 2\max_{\mu \in \Psi}D_H^2(\mu, \nu_\mu) + 2\max_{\nu' \in \Phi}\min_{\zeta \in \Theta}D_H^2(\nu', \zeta)$$

$$= 2\max_{\mu \in \Psi}\min_{\nu \in \Phi}D_H^2(\mu, \nu) + 2\max_{\nu' \in \Phi}\min_{\zeta \in \Theta}D_H^2(\nu', \zeta)) = 2D_H^2(\Psi \to \Phi) + 2D_H^2(\Phi \to \Theta)$$

$\blacksquare$

Appendix E. Robust Online Estimation

Lemma 6 *Given a finite set $\mathcal{B}$, a $T : \mathbb{N}^{>0}$, a history $h : (\mathcal{A} \times \mathcal{O})^{\leq T}$, a sequence $\widehat{M}_t : \mathcal{A} \to \Delta\mathcal{O}$, a distribution $\zeta_1 : \Delta\mathcal{B}$ with full support, and a betting function $bet : \mathcal{B} \times (\mathcal{A} \to \Delta\mathcal{O}) \times \mathcal{A} \times \mathcal{O} \to \mathbb{R}^{\geq 0}$, we can recursively define*

$$\bigstar_t := \mathbb{E}_{B \sim \zeta_t}[bet(B, \widehat{M}_t, a_t, o_t)]$$

$$\zeta_{t+1}(B) := \frac{\zeta_t(B) \cdot bet(B, \widehat{M}_t, a_t, o_t)}{\bigstar_t}$$

If all $\bigstar_t > 0$, then every $B \in \mathcal{B}$ fulfills

$$\ln(\zeta_{T+1}(B)) - \ln(\zeta_1(B)) = \sum_{t=1}^{T} \ln(bet(B, \widehat{M}_t, a_t, o_t)) - \sum_{t=1}^{T} \ln(\bigstar_t)$$

The proof is as follows.

$$\ln(\zeta_{T+1}(B)) - \ln(\zeta_1(B)) = \sum_{t=1}^{T} \ln(\zeta_{t+1}(B)) - \ln(\zeta_t(B)) = \sum_{t=1}^{T} \ln\left(\frac{\zeta_{t+1}(B)}{\zeta_t(B)}\right)$$

$$= \sum_{t=1}^{T} \ln\left(\frac{\frac{\zeta_t(B) \cdot \mathrm{bet}(B, \widehat{M}_t, a_t, o_t)}{\bigstar_t}}{\zeta_t(B)}\right) = \sum_{t=1}^{T} \ln\left(\frac{\mathrm{bet}(B, \widehat{M}_t, a_t, o_t)}{\bigstar_t}\right)$$

$$= \sum_{t=1}^{T} \ln(\mathrm{bet}(B, \widehat{M}_t, a_t, o_t)) - \sum_{t=1}^{T} \ln(\bigstar_t)$$

This proof above only works if all ζ_t assign nonzero probability to B. If some ζ_t assigns zero probability to B, then the same holds for all greater t, and this can only have occurred by one of the bets returning 0. So, as long as we still have $\bigstar_t > 0$ for all t, we can verify our desired equation via

$$\ln(\zeta_{T+1}(B)) - \ln(\zeta_1(B)) = -\infty = \sum_{t=1}^{T} \ln(\mathrm{bet}(B, \widehat{M}_t, a_t, o_t)) - \sum_{t=1}^{T} \ln(\bigstar_t)$$

This is because $\zeta_{T+1}(B) = 0$, $\zeta_1(B) > 0$, some t has $\mathrm{bet}(B, \widehat{M}_t, a_t, o_t) = 0$, and all $\bigstar_t > 0$. ∎

Proposition 4 *If the uniform bettor has positive probability according to ζ_1, the RUE algorithm never divides by zero and all ζ_t are probability distributions.*

This proof proceeds in three phases. First, we show that if the uniform bettor B_{uniform} has nonzero probability according to some ζ_t which is a probability distribution, then for all a, $\widehat{M}_t(a)$ will have full support, eliminating division-by-zero errors if the precondition holds. Second, we show that if the uniform bettor has nonzero probability according to some ζ_t which is a probability distribution, then in the language of Lemma 6, for all a, o, $\mathbb{E}_{B \sim \zeta_t}[\mathrm{bet}(B, \widehat{M}_t, a, o)] = 1$ holds. Finally, with these two results, we carry through an inductive proof that the uniform bettor always has nonzero probability and all ζ_t are probability distributions. Therefore, $\widehat{M}_t(a)$ always has full support, eliminating all division-by-zero errors by the first phase of the proof.

Let ε' be $\zeta_1(B_{\mathrm{uniform}})$. To recap some notation, given an bettor B which isn't the pessimism bettor, M_B denotes the model in $\mathcal{H}$ which B corresponds to, or the function mapping all actions to the uniform distribution over $\mathcal{O}$ for the uniform bettor. We use $\mu_{\widehat{M}_t, B, a}$ for $\mathrm{argmin}_{\mu \in M_B(a)} D_H^2(\mu, \widehat{M}_t(a))$. If $\widehat{M}_t(a)$ has full support, this distribution is unique, but until then, it is an arbitrary distribution.

For the first phase, assume for purposes of contradiction that there is some a, o' such that $\mathbb{P}_{\widehat{M}_t(a)}(o') = 0$ and yet the uniform bettor has nonzero probability according to ζ_t which is a probability distribution. Fix an observation o_a which $\widehat{M}_t(a)$ assigns nonzero probability to. Restating definitions, we have

$$\widehat{M}_t(a) = \mathrm{argmin}_{\mu \in \Delta \mathcal{O}} \mathbb{E}_{B \sim \zeta_t}[\text{ if } B = \bullet, \varepsilon \mathbb{E}_{o \sim \mu}[r(a, o)] \text{ else } 2D_H^2(\mu \to M_B(a))]$$

28

Perturb the minimizing $\widehat{M}_t(a)$ to $\widehat{M}'_t(a)$ by moving an arbitrarily small ε'' amount of probability measure from o_a to the (zero probability) observation o'. Now we compute.

$$\mathop{\mathbb{E}}_{B\sim\zeta_t}[\text{ if } B = \bullet, \varepsilon \mathop{\mathbb{E}}_{o\sim\widehat{M}_t(a)}[r(a,o)] \text{ else } 2D_H^2(\widehat{M}_t(a) \to M_B(a))]$$

$$-\mathop{\mathbb{E}}_{B\sim\zeta_t}[\text{ if } B = \bullet, \varepsilon \mathop{\mathbb{E}}_{o\sim\widehat{M}'_t(a)}[r(a,o)] \text{ else } 2D_H^2(\widehat{M}'_t(a) \to M_B(a))]$$

$$=\mathop{\mathbb{E}}_{B\sim\zeta_t}[\text{ if } \bullet, \varepsilon(\mathop{\mathbb{E}}_{o\sim\widehat{M}_t(a)}[r(a,o)] - \mathop{\mathbb{E}}_{o\sim\widehat{M}'_t(a)}[r(a,o)])$$

$$\text{else } 2(D_H^2(\widehat{M}_t(a) \to M_B(a)) - D_H^2(\widehat{M}'_t(a) \to M_B(a)))]$$

$$\geq \mathop{\mathbb{E}}_{B\sim\zeta_t}\left[\text{if } \bullet, -\varepsilon \cdot D_{TV}(\widehat{M}_t(a), \widehat{M}'_t(a)) \text{ else } 2\left(D_H^2(\widehat{M}_t(a), \mu_{\widehat{M}_t,B,a}) - D_H^2(\widehat{M}'_t(a), \mu_{\widehat{M}_t,B,a})\right)\right]$$

The above line was because the reward was in $[0, 1]$, so we can lower-bound by the negative total variation distance. For the Hellinger-squared error term, we use that $\mu_{\widehat{M}_t,B,a}$ was defined as the error-minimizer, and swapping "distance from a point to a set" for "distance from a point to a point" can only increase the distance, which leads to a decrease because of the negative sign. Now, split out the uniform bettor from the rest $\mathcal{H}$, use that $\mu_{\widehat{M}_t,B_{\text{uniform}},a}$ is the uniform distribution (denoted as u), decrease further by minimizing over all μ, and write $\min_\mu(f(\mu))$ as $-\max_\mu(-f(\mu))$.

$$\geq \mathop{\mathbb{E}}_{B\sim\zeta_t}[\text{if } \bullet, -\varepsilon \cdot D_{TV}(\widehat{M}_t(a), \widehat{M}'_t(a)) \text{ else if } B_{\text{uniform}},$$

$$-2\left(D_H^2(\widehat{M}'_t(a), u) - D_H^2(\widehat{M}_t(a), u)\right) \text{ else } -2\max_\mu\left(D_H^2(\widehat{M}'_t(a), \mu) - D_H^2(\widehat{M}_t(a), \mu)\right)]$$

Unpacking the Hellinger-squared error as $1 - \sum_o \sqrt{\widehat{M}_t(a)(o) \cdot \mu(o)}$, and canceling out as much as we can because $\widehat{M}_t(a)$ and $\widehat{M}'_t(a)$ are the same except on two observations, we get

$$=\mathop{\mathbb{E}}_{B\sim\zeta_t}[\text{ if } \bullet, -\varepsilon \cdot D_{TV}(\widehat{M}_t(a), \widehat{M}'_t(a)) \text{ else if } B_{\text{uniform}},$$

$$-2\left(\sqrt{\widehat{M}_t(a)(o')u(o')} + \sqrt{\widehat{M}_t(a)(o_a)u(o_a)} - \sqrt{\widehat{M}'_t(a)(o')u(o')} - \sqrt{\widehat{M}'_t(a)(o_a)u(o_a)}\right) \text{ else}$$

$$-2\max_\mu\left(\sqrt{\widehat{M}_t(a)(o')\mu(o')} + \sqrt{\widehat{M}_t(a)(o_a)\mu(o_a)} - \sqrt{\widehat{M}'_t(a)(o')\mu(o')} - \sqrt{\widehat{M}'_t(a)(o_a)\mu(o_a)}\right)]$$

We now use that $\widehat{M}_t(a)(o')$ is zero, $\widehat{M}'_t(a)(o')$ is ε'', and $\widehat{M}'_t(a)(o_a) = \widehat{M}_t(a)(o_a) - \varepsilon''$.

$$=\mathop{\mathbb{E}}_{B\sim\zeta_t}[\text{if } \bullet, -\varepsilon \cdot D_{TV}(\widehat{M}_t(a), \widehat{M}'_t(a))$$

$$\text{else if } B_{\text{uniform}}, -2\left(\sqrt{u(o_a)}\left(\sqrt{\widehat{M}_t(a)(o_a)} - \sqrt{\widehat{M}_t(a)(o_a) - \varepsilon''}\right) - \sqrt{\varepsilon'' \cdot u(o')}\right)$$

$$\text{else } -2\max_\mu\left(\sqrt{\mu(o_a)}\left(\sqrt{\widehat{M}_t(a)(o_a)} - \sqrt{\widehat{M}_t(a)(o_a) - \varepsilon''}\right) - \sqrt{\varepsilon'' \cdot \mu(o')}\right)]$$

29

Remove some positive terms, and pull one of them out of the expectation.

$$\geq \underset{B \sim \zeta_t}{\mathbb{E}} [\text{if } \bullet, -\varepsilon \cdot D_{TV}(\widehat{M}_t(a), \widehat{M}'_t(a))$$

$$\text{else if } B_{\text{uniform}}, -2\sqrt{u(o_a)}\left(\sqrt{\widehat{M}_t(a)(o_a)} - \sqrt{\widehat{M}_t(a)(o_a) - \varepsilon''}\right)$$

$$\text{else } -2\max_\mu \sqrt{\mu(o_a)}\left(\sqrt{\widehat{M}_t(a)(o_a)} - \sqrt{\widehat{M}_t(a)(o_a) - \varepsilon''}\right)] + 2\zeta_t(B_{\text{uniform}})\sqrt{\varepsilon'' \cdot u(o')}$$

Then upper-bound $\mu(o_a)$ and $u(o_a)$ by 1, letting us fold the uniform case, and the $B \in \mathcal{H}$ case together.

$$\geq \underset{B \sim \zeta_t}{\mathbb{E}}\left[\text{ if } \bullet, -\varepsilon \cdot D_{TV}(\widehat{M}_t(a), \widehat{M}'_t(a)) \text{ else } -2\left(\sqrt{\widehat{M}_t(a)(o_a)} - \sqrt{\widehat{M}_t(a)(o_a) - \varepsilon''}\right)\right]$$

$$+2\zeta_t(B_{\text{uniform}})\sqrt{\varepsilon'' \cdot u(o')}$$

Upper-bound ε by 1, and use that the total variation distance between $\widehat{M}_t(a)$ and $\widehat{M}'_t(a)$ is ε''. Further, $\widehat{M}_t(a)(o_a)$ was stipulated to be > 0, so we can take a limit as ε'' approaches zero, to yield, to within $\mathcal{O}\left(\varepsilon''^2\right)$,

$$\geq \underset{B \sim \zeta_t}{\mathbb{E}}\left[\text{if } \bullet, -\varepsilon'', \text{ else } -\frac{1}{\sqrt{\widehat{M}_t(a)(o_a)}}\varepsilon''\right] + 2\zeta_t(B_{\text{uniform}})\sqrt{\varepsilon'' \cdot u(o')}$$

$$\geq -\frac{1}{\sqrt{\widehat{M}_t(a)(o_a)}}\varepsilon' + 2\zeta_t(B_{\text{uniform}})\sqrt{\varepsilon' \cdot u(o')}$$

However, $\widehat{M}_t(a)(o_a)$ was assumed to be positive, as was $\zeta_t(B_{\text{uniform}})$, and the outcome space $\mathcal{O}$ was assumed to be finite, so this is a term on the order of $\sqrt{\varepsilon''}$ minus a term on the order of ε'', which is positive when ε'' is sufficiently small. So our net inequality, for sufficiently small ε'', is

$$\underset{B \sim \zeta_t}{\mathbb{E}}[\text{ if } B = \bullet, \varepsilon \underset{o \sim \widehat{M}_t(a)}{\mathbb{E}}[r(a,o)] \text{ else } 2D_H^2(\widehat{M}_t(a) \to M_B(a))]$$

$$-\underset{B \sim \zeta_t}{\mathbb{E}}[\text{if } B = \bullet, \varepsilon \underset{o \sim \widehat{M}'_t(a)}{\mathbb{E}}[r(a,o)] \text{ else } 2D_H^2(\widehat{M}'_t(a) \to M_B(a))] > 0$$

However, this is impossible, because $\widehat{M}_t(a)$ was assumed to be the minimizer of the equation, so perturbing it cannot produce a strictly lower value. By contradiction, if the uniform bettor B_{uniform} has nonzero probability according to some ζ_t which is a probability distribution, then for all a, $\widehat{M}_t(a)$ has full support.

Now we move on to phase two, and show that if the uniform bettor has nonzero probability according to some ζ_t which is a probability distribution (implying that $\widehat{M}_t(a)$ always

has full support), then for *every* a, o, we have $\mathop{\mathbb{E}}\limits_{B \sim \zeta_t}[\text{bet}(B, \widehat{M}_t, a, o)] = 1$. Restating the estimate of RUE, it is

$$\widehat{M}_t(a) = \operatorname*{argmin}_{\nu \in \Delta \mathcal{O}} \mathop{\mathbb{E}}_{B \sim \zeta_t} [\text{if } B = \bullet, \varepsilon \mathop{\mathbb{E}}_{o \sim \nu}[r(a,o)] \text{ else } 2D_H^2(\nu \to M_B(a))]$$

To abbreviate this, let $g^{B,a} : \Delta \mathcal{O} \to \mathbb{R}$ be $\lambda \nu . \varepsilon \mathop{\mathbb{E}}_{o \sim \nu}[r(a,o)]$ for $B = \bullet$ (the pessimism bettor), and $\lambda \nu . 2 D_H^2(\nu \to M_B(a))$ otherwise. This lets us restate the estimate as $\widehat{M}_t(a) = \operatorname*{argmin}_{\nu \in \Delta \mathcal{O}} \mathop{\mathbb{E}}_{B \sim \zeta_t}[g^{B,a}]$. Note that, for all B, a, $g^{B,a}$ is Frechet-differentiable on the interior of $\Delta \mathcal{O}$, and convex. This is easy to show for the pessimism bettor, but for Hellinger-squared error, we appeal to Lemma 4, because $\widehat{M}_t(a)$ has full support. $\mathop{\mathbb{E}}_{B \sim \zeta_t}[g^{B,a}]$ is a finite number of Frechet-differentiable convex functions mixed together, so it is a Frechet-differentiable convex function.

Because $\widehat{M}_t(a)$ is the minimizer of a differentiable convex function, the derivative in all directions is 0. Let o be an arbitrary observation, and consider our direction of movement to be $\widehat{M}_t(a) - \mathbf{1}_o$, where $\mathbf{1}_o$ is the distribution which places all measure on o. This is a probability distribution minus a probability distribution, and represents moving from certainty in an observation towards our prediction. Use that the derivative in all directions is 0, the derivative of a mixture of functions is the mixture of the derivatives, and linearity. Then unpack the definition of $g^{B,a}$.

$$0 = d\left(\lambda \nu . \mathop{\mathbb{E}}_{B \sim \zeta_t}\left[g^{B,a}(\nu)\right]\right)_{\widehat{M}_t(a)}(\widehat{M}_t(a) - \mathbf{1}_o) = \mathbb{E}_{\zeta_t}\left[d(g^{B,a})_{\widehat{M}_t(a)}(\widehat{M}_t(a) - \mathbf{1}_o)\right]$$

$$= \mathbb{E}_{\zeta_t}[\text{if } B = \bullet, d(\lambda \nu . \varepsilon \mathop{\mathbb{E}}_{o \sim \nu}[r(a,o)])_{\widehat{M}_t(a)}(\widehat{M}_t(a) - \mathbf{1}_o)$$

$$\text{else } d(\lambda \nu . 2 D_H^2(\nu \to M_B(a)))_{\widehat{M}_t(a)}(\widehat{M}_t(a) - \mathbf{1}_o)]$$

Applying Lemma 4, we get

$$= \mathbb{E}_{\zeta_t}[\text{if } B = \bullet, d(\lambda \nu . \varepsilon \mathop{\mathbb{E}}_{o \sim \nu}[r(a,o)])_{\widehat{M}_t(a)}(\widehat{M}_t(a) - \mathbf{1}_o)$$

$$\text{else } d(\lambda \nu . 2 D_H^2(\mu_{\widehat{M}_t, B, a}, \nu))_{\widehat{M}_t(a)}(\widehat{M}_t(a) - \mathbf{1}_o)]$$

In Lemma 4 we computed the differential for Hellinger distance and expressed it as an inner product, yielding

$$d(\lambda \nu . 2 D_H^2(\mu_{\widehat{M}_t, B, a}, \nu))_{\widehat{M}_t(a)}(\widehat{M}_t(a) - \mathbf{1}_o) = \left\langle -\sqrt{\frac{\mu_{\widehat{M}_t, B, a}}{\widehat{M}_t(a)}}, \widehat{M}_t(a) - \mathbf{1}_o \right\rangle$$

For the other differential, it is easy to compute.

$$d(\lambda \nu . \varepsilon \mathop{\mathbb{E}}_{o \sim \nu}[r(a,o)])_{\widehat{M}_t(a)}(\widehat{M}_t(a) - \mathbf{1}_o) = \left\langle \lambda o' . \varepsilon \cdot r(a, o'), \widehat{M}_t(a) - \mathbf{1}_o \right\rangle$$

Substituting these in, we have an overall equality of

$$0 = \mathbb{E}_{\zeta_t}\left[\text{if } B = \bullet, \left\langle \lambda o' . \varepsilon \cdot r(a, o'), \widehat{M}_t(a) - \mathbf{1}_o \right\rangle \text{ else } \left\langle -\sqrt{\frac{\mu_{\widehat{M}_t, B, a}}{\widehat{M}_t(a)}}, \widehat{M}_t(a) - \mathbf{1}_o \right\rangle\right]$$

31

We can rewrite this as

$$0 = \mathbb{E}_{\zeta_t}[\text{if } \bullet, \varepsilon \left(\mathop{\mathbb{E}}_{o' \sim \widehat{M}_t(a)} [r(a, o')] - r(a, o) \right)$$

$$\text{else } \sqrt{\frac{\mu_{\widehat{M}_t, B, a}(o)}{\widehat{M}_t(a)(o)}} - \sum_{o'} \sqrt{\mu_{\widehat{M}_t, B, a}(o') \cdot \widehat{M}_t(a)(o')]}$$

Adding 1 to both sides, and using that $D_H^2(\mu, \nu) = 1 - \sum_o \sqrt{\mu(o)\nu(o)}$, and that $D_H^2(\mu_{\widehat{M}_t, B, a}, \widehat{M}_t(a)) = D_H^2(\widehat{M}_t(a) \to M_B(a))$, we get

$$1 = \mathbb{E}_{\zeta_t}[\text{if } B = \bullet, 1 + \varepsilon \left(\mathop{\mathbb{E}}_{o' \sim \widehat{M}_t(a)} [r(a, o')] - r(a, o) \right)$$

$$\text{else } \sqrt{\frac{\mu_{\widehat{M}_t, B, a}(o)}{\widehat{M}_t(a)(o)} + D_H^2(\widehat{M}_t(a) \to M_B(a))]}$$

$\zeta_t(B)$ times the associated bet of B is $\zeta_{t+1}(B)$ according to the RUE algorithm, so the above equation shows that ζ_{t+1} is a probability distribution because $\sum_{B \in \mathcal{B}} \zeta_{t+1}(B) = 1$. In the language of Lemma 6, the above is precisely the definition of the betting functions for RUE, so we get an overall equality of $1 = \mathop{\mathbb{E}}_{B \sim \zeta_t} \left[\text{bet}(B, \widehat{M}_t, a, o) \right]$. This argument works for any a, o, t if ζ_t of the uniform bettor is nonzero. So, also, in the language of Lemma 6, we have just proven that $\zeta_t(B_{\text{uniform}}) > 0$ and ζ_t being a probability distribution implies that $\bigstar_t = 1$ and ζ_{t+1} is a probability distribution.

For our final phase, we give an inductive proof that the uniform bettor always has nonzero probability and all ζ_t are probability distributions. In the base case, this holds by assumption. For the induction step, if $\zeta_t(B_{\text{uniform}}) > 0$ and ζ_t is a probability distribution, then ζ_{t+1} is also a probability distribution, as proved earlier. Then we have $\zeta_{t+1}(B_{\text{uniform}}) = \zeta_t(B_{\text{uniform}}) \text{bet}(B_{\text{uniform}}, \widehat{M}_t, a_t, o_t)$. The probability of the uniform bettor according to ζ_t is nonzero (by induction assumption), so we just need to show that the bet made is nonzero. We may compute the bet of the uniform bettor (because it always bets on the uniform distribution u) as

$$\sqrt{\frac{u(o_t)}{\widehat{M}_t(a_t)(o_t)} + D_H^2(\widehat{M}_t(a_t), u)}$$

Because the observation space is finite, nonzero probability is assigned to all observations, so the square root is always nonzero, establishing $\zeta_{t+1}(B_{\text{uniform}}) > 0$. By induction, our desired result then follows, that the uniform bettor always has nonzero probability and all ζ_t are probability distributions. All $\bigstar_t$ are 1 by the second phase of the proof, and by the first phase of the proof, we can conclude that $\widehat{M}_t(a)$ has full support for all a, t, and division-by-zero errors are impossible. ∎

Theorem 2 *If our estimator $\widehat{M}$ is the RUE algorithm with a suitable choice of prior,*
$\beta_{\widehat{M}}(T, \delta) \leq \ln \left(\frac{2|\mathcal{H}|}{\delta} \right)$, *and* $\alpha_{\widehat{M}}(T, \delta) \leq \sqrt{T} \left(2\sqrt{\ln(2)} + \sqrt{2 \ln \left(\frac{1}{\delta} \right)} \right)$

The choice of prior is $\frac{1}{2|\mathcal{H}|}$ probability assigned to all bettors corresponding to $M \in \mathcal{H}$, some arbitrarily low ε' probability assigned to the uniform bettor, and $\frac{1}{2} - \varepsilon'$ probability assigned to the pessimism bettor. Technically, the bound on β will only hold for $\varepsilon' \leq \frac{1}{2}$, and the bound on α only holds for $T > 2$ in the $\varepsilon' \to 0$ limit.

We use $\mu_{\widehat{M}_t, B, a}$ to denote $\underset{\mu \in M_B(a)}{\operatorname{argmin}} D_H^2(\mu, \widehat{M}_t(a))$. The RUE algorithm is well-defined by Proposition 4. Now, we can invoke Lemma 6 to analyze the estimation complexity.

$$\ln(\zeta_{T+1}(B)) - \ln(\zeta_1(B)) = \sum_{t=1}^{T} \ln(\operatorname{bet}(B, \widehat{M}_t, a_t, o_t)) - \sum_{t=1}^{T} \ln(\bigstar_t)$$

This can be simplified further. All $\bigstar_t = 1$, as proved in Proposition 4. Our choice of ζ_1 was half of the measure uniformly distributed on $\mathcal{H}$, and approximately half on the pessimism bettor, and an arbitrarily small quantity on the uniform bettor. Substituting these in, we have shown that for all B which aren't the pessimism or uniform bettor, we have $\ln(2|\mathcal{H}|) \geq \sum_{t=1}^{T} \ln(\operatorname{bet}(B, \widehat{M}_t, a_t, o_t))$, and for the pessimism bettor, the left-hand side is just $\ln(\frac{1}{\frac{1}{2} - \varepsilon'})$, which is approximately $\ln(2)$.

We will now bound $\beta_{\widehat{M}_t}$ and $\alpha_{\widehat{M}_t}$ using martingale arguments. Fix some true hypothesis $M^* \in \mathcal{H}$, environment $\theta \models M^*$, and algorithm π. Let B^* denote the bettor corresponding to M^*. By Lemma A.4 from Foster et al. (2021), with $1 - \delta$ probability according to $\theta \bowtie \pi$, we will have

$$\sum_{t=1}^{T} \ln(\operatorname{bet}(B^*, \widehat{M}_t, a_t, o_t)) \geq -\sum_{t=1}^{T} \ln \left(\underset{a \sim \pi_t, o \sim \theta_t(a)}{\mathbb{E}} \left[e^{-\ln(\operatorname{bet}(B^*, \widehat{M}_t, a, o))} \right] \right) - \ln \left(\frac{1}{\delta} \right)$$

The history dependence is implicit in the notation π_t, θ_t. Composing this with the previous inequality, and moving the logarithm over, we have

$$\ln(2|\mathcal{H}|) + \ln \left(\frac{1}{\delta} \right) \geq -\sum_{t=1}^{T} \ln \left(\underset{a \sim \pi_t, o \sim \theta_t(a)}{\mathbb{E}} \left[e^{-\ln(\operatorname{bet}(B^*, \widehat{M}_t, a, o))} \right] \right)$$

We can re-express this as

$$\ln \left(\frac{2|\mathcal{H}|}{\delta} \right) \geq -\sum_{t=1}^{T} \ln \left(\underset{a \sim \pi_t, o \sim \theta_t(a)}{\mathbb{E}} \left[\frac{1}{\operatorname{bet}(B^*, \widehat{M}_t, a, o)} \right] \right)$$

Unpacking the definition of the betting function in the robust universal estimator, we have

$$= -\sum_{t=1}^{T} \ln \left(\mathbb{E}_{\pi_t, \theta_t} \left[\frac{1}{\sqrt{\frac{\mu_{\widehat{M}_t, B^*, a}(o)}{\widehat{M}_t(a)(o)}} + D_H^2(\widehat{M}_t(a) \to M^*(a))} \right] \right)$$

Removing the Hellinger-squared error makes this term smaller, and permits us to flip the square root. Re-expressing the expectations, we have

$$\geq -\sum_{t=1}^{T} \ln \left(\mathbb{E}_{\pi_t} \left[\mathbb{E}_{\theta_t(a)} \left[\sqrt{\frac{\widehat{M}_t(a)(o)}{\mu_{\widehat{M}_t, B^*, a}(o)}} \right] \right] \right)$$

Now, if we start at $\mu_{\widehat{M}_t,B^*,a}$ and travel in the direction of $\theta_t(a)$, the Hellinger distance to $\widehat{M}_t(a)$ can only increase, because $\mu_{\widehat{M}_t,B^*,a}$ is the Hellinger distance minimizer to $\widehat{M}_t(a)$ within $M^*(a)$, and $\theta_t(a)$ is also a distribution within $M^*(a)$, by the definition of θ being consistent with M^*. This justifies the equation

$$0 \le d(\lambda\mu.2D_H^2(\mu,\widehat{M}_t(a)))_{\mu_{\widehat{M}_t,B^*,a}}(\theta_t(a) - \mu_{\widehat{M}_t,B^*,a})$$

By explicitly computing this derivative, we arrive at

$$0 \le \left\langle -\sqrt{\frac{\widehat{M}_t(a)}{\mu_{\widehat{M}_t,B^*,a}}}, \theta_t(a) - \mu_{\widehat{M}_t,B^*,a} \right\rangle$$

Writing the inner product as an expectation and reshuffling, we have derived

$$\mathbb{E}_{\theta_t(a)}\left[\sqrt{\frac{\widehat{M}_t(a)(o)}{\mu_{\widehat{M}_t,M^*,a}(o)}}\right] \le \mathbb{E}_{\mu_{\widehat{M}_t,M^*,a}}\left[\sqrt{\frac{\widehat{M}_t(a)(o)}{\mu_{\widehat{M}_t,M^*,a}(o)}}\right]$$

Because of this fact, along with the definition of Hellinger-squared error, we can derive

$$-\sum_{t=1}^T \ln\left(\mathbb{E}_{\pi_t}\left[\mathbb{E}_{\theta_t(a)}\left[\sqrt{\frac{\widehat{M}_t(a)(o)}{\mu_{\widehat{M}_t,B^*,a}(o)}}\right]\right]\right) \ge -\sum_{t=1}^T \ln\left(\mathbb{E}_{\pi_t}\left[\mathbb{E}_{\mu_{\widehat{M}_t,B^*,a}}\left[\sqrt{\frac{\widehat{M}_t(a)(o)}{\mu_{\widehat{M}_t,B^*,a}(o)}}\right]\right]\right)$$

$$= -\sum_{t=1}^T \ln\left(\mathbb{E}_{\pi_t}\left[\sum_o \sqrt{\widehat{M}_t(a)(o)\cdot\mu_{\widehat{M}_t,B^*,a}(o)}\right]\right) = -\sum_{t=1}^T \ln\left(\mathbb{E}_{\pi_t}\left[1 - D_H^2(\mu_{\widehat{M}_t,B^*,a},\widehat{M}_t(a))\right]\right)$$

$$= -\sum_{t=1}^T \ln\left(1 - \mathbb{E}_{\pi_t}\left[D_H^2(\widehat{M}_t(a) \to M^*(a))\right]\right)$$

Chaining all our inequalities together, and finishing off with $-\ln(1-x) \ge x$, we have derived that

$$\ln\left(\frac{2|\mathcal{H}|}{\delta}\right) \ge \sum_{t=1}^T \mathbb{E}_{a\sim\pi_t}\left[D_H^2(\widehat{M}_t(a) \to M^*(a))\right]$$

holds with $> 1 - \delta$ probability according to $\theta \bowtie \pi$. M^*, θ, δ, and π were arbitrary, so this provides a value for $\beta_{\widehat{M}}(T,\delta)$.

To compute $\alpha_{\widehat{M}}(T,\delta)$, we will use Azuma's inequality. For the pessimistic better, we were previously at $\ln(2) \ge \sum_{t=1}^T \ln(\text{bet}(B,\widehat{M}_t,a_t,o_t))$ from Lemma 6. This holds approximately, and the left-hand side is technically $\ln\left(\frac{1}{\frac{1}{2}-\varepsilon'}\right)$, but in the $\varepsilon' \to 0$ limit, this inequality holds. Substituting in the pessimism bet, we get

$$\ln(2) \ge \sum_{t=1}^T \ln(1 + \varepsilon(\mathbb{E}_{\widehat{M}_t(a_t)}[r(a_t,o)] - r(a_t,o_t)))$$

ε was defined to ensure that it was always $\frac{1}{2}$ or less, so we'll use the fact that $\ln(1+x) \geq x - x^2$ over the interval $[-0.5, 0.5]$, to get

$$\ln(2) \geq \sum_{t=1}^{T} \varepsilon(\mathbb{E}_{\widehat{M}_t(a_t)}[r(a_t, o)] - r(a_t, o_t)) - \varepsilon^2(\mathbb{E}_{\widehat{M}_t(a_t)}[r(a_t, o)] - r(a_t, o_t))^2$$

Upper-bounding the difference in rewards by 1, and adding to both sides, we get

$$\ln(2) + T\varepsilon^2 \geq \sum_{t=1}^{T} \varepsilon(\mathbb{E}_{\widehat{M}_t(a_t)}[r(a_t, o)] - r(a_t, o_t))$$

Azuma's inequality can be used to show that, no matter what θ, π are, with $1 - \delta$ probability under $\theta \bowtie \pi$, we have

$$\sum_{t=1}^{T} \varepsilon \mathop{\mathbb{E}}_{a \sim \pi_t, o \sim \theta_t(a)} \left[\mathop{\mathbb{E}}_{o' \sim \widehat{M}_t(a)}[r(a, o')] - r(a, o) \right]$$

$$\leq \sum_{t=1}^{T} \varepsilon(\mathop{\mathbb{E}}_{o \sim \widehat{M}_t(a_t)}[r(a_t, o)] - r(a_t, o_t)) + \sqrt{2\ln\left(\frac{1}{\delta}\right) \sum_{t=1}^{T} \varepsilon^2}$$

Rearranging and combining with the previous inequality, we get

$$\ln(2) + T\varepsilon^2 + \sqrt{2\ln\left(\frac{1}{\delta}\right) \sum_{t=1}^{T} \varepsilon^2} \geq \sum_{t=1}^{T} \varepsilon \mathop{\mathbb{E}}_{a \sim \pi_t, o \sim \theta_t(a)} \left[\mathop{\mathbb{E}}_{o' \sim \widehat{M}_t(a)}[r(a, o')] - r(a, o) \right]$$

Simplifying the contents of the square root, using the f notation to abbreviate expected rewards, and dividing both sides by epsilon, we get

$$\frac{\ln(2)}{\varepsilon} + T\varepsilon + \sqrt{2\ln\left(\frac{1}{\delta}\right) T} \geq \sum_{t=1}^{T} \mathop{\mathbb{E}}_{a \sim \pi_t} \left[f^{\widehat{M}_t}(a) - f^{\theta_t}(a) \right]$$

Finally, if $\sqrt{\frac{\ln(2)}{T}} \leq \frac{1}{2}$ (which holds for all $T \geq 3$), then $\varepsilon = \sqrt{\frac{\ln(2)}{T}}$ and plugging that in yields a bound of

$$\sqrt{T}\left(2\sqrt{\ln(2)} + \sqrt{2\ln\left(\frac{1}{\delta}\right)}\right) \geq \sum_{t=1}^{T} \mathop{\mathbb{E}}_{a \sim \pi_t} \left[f^{\widehat{M}_t}(a) - f^{\theta_t}(a) \right]$$

This bound covers all $T \geq 3$. This establishes a value for $\alpha_{\widehat{M}}(T, \delta)$. ∎

Proposition 5 *If $\mathcal{O}$ is finite, then for all $\mathcal{H}$, there exists an online estimator $\widehat{M}$ where* $\beta_{\widehat{M}}(T, \delta) \leq \min_{\varepsilon > 0}\left(2\ln\left(\frac{2\mathcal{N}(\mathcal{H}, \varepsilon)}{\delta}\right) + 8T\varepsilon^2\right)$, *and* $\alpha_{\widehat{M}}(T, \delta) \leq \sqrt{T}\left(2\sqrt{\ln(2)} + \sqrt{2\ln\left(\frac{1}{\delta}\right)}\right)$

Given a model N, $B_{N,\varepsilon}$ will denote the ball of all models which are ε-close to N in Hellinger distance.

$$B_{N,\varepsilon} = \{N \mid \forall a \in \mathcal{A} : D_H(N(a), M(a)) \leq \varepsilon\}$$

Here we are using Hausdorff distance between the sets $N(a)$ and $M(a)$, instead of an asymmetric notion of distance. Given a model N, let N_ε denote the ε-thickened version of N, defined as

$$N_\varepsilon(a) := \{\mu \mid D_H(\mu \to N(a)) \leq \varepsilon\}$$

By the definition of $\mathcal{N}(\mathcal{H}, \varepsilon)$ (the minimum number of models needed for an ε-approximate cover of $\mathcal{H}$), there is some finite set $\mathcal{C}_\varepsilon$ of models, where $\mathcal{H} \subseteq \bigcup\limits_{M' \in \mathcal{C}} B_{M',\varepsilon}$, and $|\mathcal{C}| = \mathcal{N}(\mathcal{H}, \varepsilon)$.

Our estimator will be RUE, with its finite hypothesis space being the N_ε models, for $N \in \mathcal{C}_\varepsilon$. This establishes the value of $\alpha_{\widehat{M}}(T, \delta)$, by Theorem 2. We will now compute $\beta_{\widehat{M}}(T, \delta)$ for this estimator. If M^* is a true hypothesis, let N^* denote some model $N \in \mathcal{C}_\varepsilon$ such that $M^* \in B_{N,\varepsilon}$. Such a model will always exist because the balls cover $\mathcal{H}$. By Lemma 5 applied twice, for any algorithm π, we have

$$\sum_{t=1}^{T} \mathop{\mathbb{E}}_{a \sim \pi_t} \left[D_H^2(\widehat{M}_t(a) \to M^*(a)) \right]$$

$$\leq 2 \sum_{t=1}^{T} \mathbb{E}_{\pi_t} \left[D_H^2(\widehat{M}_t(a) \to N_\varepsilon^*(a)) \right] + 2 \sum_{t=1}^{T} \mathbb{E}_{\pi_t} \left[D_H^2(N_\varepsilon^*(a) \to M^*(a)) \right]$$

$$\leq 2 \sum_{t=1}^{T} \mathbb{E}_{\pi_t} \left[D_H^2(\widehat{M}_t(a) \to N_\varepsilon^*(a)) \right] + 4 \sum_{t=1}^{T} \mathbb{E}_{\pi_t} \left[D_H^2(N_\varepsilon^*(a) \to N^*(a)) + D_H^2(N^*(a) \to M^*(a)) \right]$$

$N_\varepsilon^*(a)$ consists of all distributions with a Hellinger distance of ε or less from $N^*(a)$, so the middle term is at most ε^2. $N^*(a)$ has a symmetric Hellinger distance of ε or less from $M^*(a)$ because $M^* \in B_{N^*,\varepsilon}$, so the latter sum is at most ε^2 as well, yielding

$$\leq 2 \sum_{t=1}^{T} \mathbb{E}_{\pi_t} \left[D_H^2(\widehat{M}_t(a) \to N_\varepsilon^*(a)) \right] + 8T\varepsilon^2$$

For the first term, observe that, because all actions have $M^*(a) \subseteq N_\varepsilon^*(a)$, and the environment is consistent with M^*, N_ε^* is a true hypothesis as well. Therefore, by Theorem 2 along with $|\mathcal{C}_\varepsilon| = \mathcal{N}(\mathcal{H}, \varepsilon)$, with $1 - \delta$ probability according to $\theta \bowtie \pi$, we have

$$\leq 2 \ln \left(\frac{2\mathcal{N}(\mathcal{H}, \varepsilon)}{\delta} \right) + 8T\varepsilon^2$$

Because ε can be whatever we wish, we can take the minimum over all values of ε, yielding our conclusion. ∎

Appendix F. Definitions for Robust Linear Bandits:

$\mathcal{A}, \mathcal{O}$ are finite sets of actions and observations. We consider the space $\Delta\mathcal{O}$ to be embedded in the vector space $\mathbb{R}^{\mathcal{O}}$ in the typical way. $r : \mathcal{A} \times \mathcal{O} \to [0,1]$ is some known reward function.

$\mathcal{Z}, \mathcal{W}$ are finite-dimensional vector spaces. Z and W are their dimensions. $\mathcal{H}$ is a compact subset of $\mathcal{Z}$, consisting of the hypotheses.

$F : \mathcal{A} \times \mathcal{Z} \times \mathbb{R}^{\mathcal{O}} \to \mathcal{W}$ is a function which is bilinear in $\mathcal{Z}$ and $\mathbb{R}^{\mathcal{O}}$. Given some $M : \mathcal{Z}$ and $a : \mathcal{A}$, $F_{a,M} : \mathbb{R}^{\mathcal{O}} \to \mathcal{W}$ is the induced linear function, and $K_{a,M} \subseteq \mathbb{R}^{\mathcal{O}}$ is the nullspace of $F_{a,M}$.

Given an $M \in \mathcal{H}$, we will overload notation so that M also denotes a model of type $\mathcal{A} \to \square\mathcal{O}$, via the definition $M(a) := K_{a,M} \cap \Delta\mathcal{O}$. We will also overload notation so that $\mathcal{H}$ denotes the set of models which $\mathcal{H}$ induces.

a_M^* is $\underset{a \in \mathcal{A}}{\operatorname{argmax}} f^M(a)$, the action which maximizes the worst-case expected reward for M.

The spaces $\mathbb{R}^{\mathcal{O}}, \mathcal{W}, \mathcal{Z}$ are equipped with a variety of norms. Given a norm $||.||_{\text{norm}}$, B_{norm}^u denotes the unit ball of that norm, and D_{norm} denotes distances measured with respect to that norm.

The space $\mathbb{R}^{\mathcal{O}}$ is equipped with the L1 norm. $||y||_{\mathbb{R}^{\mathcal{O}}}$ denotes this norm.

The space $\mathcal{W}$ is equipped with the following three norms. $||w||_{\text{small},W}$ is the norm where $B_{\text{small},W}^u := \bigcap_{a \in \mathcal{A}, M \in \mathcal{H}} F_{a,M}(B_{\mathbb{R}^{\mathcal{O}}}^u)$. Equivalently, $||w||_{\text{small},W} := \underset{a \in \mathcal{A}, M \in \mathcal{H}}{\max} \underset{y : F_{a,M}(y)=w}{\min} ||y||_{\mathbb{R}^{\mathcal{O}}}$.

$||w||_{\text{big},W}$ is the norm where $||w||_{\text{big},W} := \dfrac{||w||_{\text{small},W}}{\underset{a \in \mathcal{A}, M \in \mathcal{H}, y \in B_{\mathbb{R}^{\mathcal{O}}}^u}{\max} ||F_{a,M}(y)||_{\text{small},W}}$. $B_{\text{big},W}^u$ can be thought of as $B_{\text{small},W}^u$, expanded so that it engulfs the set $\underset{a, M \in \mathcal{H}}{\bigcup} F_{a,M}(B_{\mathbb{R}^{\mathcal{O}}}^u)$.

Finally, $||w||_{2,W}$ is the norm where $B_{2,W}^u$ is the outer John ellipsoid of $B_{\text{big},W}^u$. It is an L2 norm and induces an inner product on $\mathcal{W}$.

The space $\mathcal{Z}$ is equipped with the following two norms.

$$||M||_Z := \max_{a \in \mathcal{A}, y \in B_{\mathbb{R}^{\mathcal{O}}}^u} ||F_{a,M}(y)||_{\text{big},W}$$

$||M||_{2,Z}$ is the norm where $B_{2,Z}^u$ is the outer John ellipsoid of B_Z^u. It is an L2 norm.

The quantity R is $\underset{a \in \mathcal{A}, M \in \mathcal{H}, y \in B_{\mathbb{R}^{\mathcal{O}}}^u}{\max} ||F_{a,M}(y)||_{\text{small},W}$. We have $R||w||_{\text{big},W} = ||w||_{\text{small},W}$.

sine is a parameter which is defined in Kosoy (2024).

Appendix G. Robust Linear Bandit Theorems

Lemma 7 *In the robust linear bandits setting, letting $\zeta : \Delta\mathcal{H}$, and $\overline{M} : \mathcal{A} \to \Delta\mathcal{O}$, and $\mathcal{G}$ be a set of functions of type $\mathcal{A} \to \mathbb{R}^{\geq 0}$ defined as $\mathcal{G} := \left\{ \lambda a.||F_{a,M}(\overline{M}(a))||_{2,W} \mid M \in \mathcal{H} \right\}$, we have that for any $\Delta', \varepsilon' > 0$, Foster's disagreement coefficient, $dis(\mathcal{G}, \Delta', \varepsilon', \zeta) \leq Z$.*

Foster's disagreement coefficient, from Foster et al. (2021), is defined as

$$\text{dis}(\mathcal{G}, \Delta', \varepsilon', \zeta) = \max_{\Delta \geq \Delta', \varepsilon \geq \varepsilon'} \frac{\Delta^2}{\varepsilon^2} \mathop{\mathbb{P}}_{N \sim \zeta} \left(\exists g \in \mathcal{G} : g(a_N^*) > \Delta, \mathop{\mathbb{E}}_{M \sim \zeta}[g^2(a_M^*)] \leq \varepsilon^2 \right)$$

Define the following subset of $\mathcal{Z}$.

$$G_{\varepsilon,\zeta} := \left\{ M \in \mathcal{Z} \;\middle|\; \mathop{\mathbb{E}}_{N \sim \zeta}[g_M^2(a_N^*)] \leq \varepsilon^2 \right\}$$

Note that the second condition in the existence statement of Foster's disagreement coefficient is precisely the requirement that the function g correspond to some $M \in G_{\varepsilon,\zeta} \cap \mathcal{H}$. Therefore, we can upper-bound the disagreement coefficient by

$$\leq \max_{\Delta \geq \Delta', \varepsilon \geq \varepsilon'} \frac{\Delta^2}{\varepsilon^2} \mathop{\mathbb{P}}_{N \sim \zeta} \left(\exists M \in G_{\varepsilon,\zeta} : g_M(a_N^*) > \Delta \right) = \max_{\Delta \geq \Delta', \varepsilon \geq \varepsilon'} \frac{\Delta^2}{\varepsilon^2} \mathop{\mathbb{P}}_{N \sim \zeta} \left(\max_{M \in G_{\varepsilon,\zeta}} g_M^2(a_N^*) > \Delta^2 \right)$$

By Markov's inequality, we can derive

$$\leq \max_{\Delta \geq \Delta', \varepsilon \geq \varepsilon'} \frac{\Delta^2}{\varepsilon^2} \frac{\mathop{\mathbb{E}}\limits_{N \sim \zeta} \left[\max\limits_{M \in G_{\varepsilon,\zeta}} g_M^2(a_N^*) \right]}{\Delta^2} = \max_{\varepsilon \geq \varepsilon'} \frac{\mathop{\mathbb{E}}\limits_{N \sim \zeta} \left[\max\limits_{M \in G_{\varepsilon,\zeta}} g_M^2(a_N^*) \right]}{\varepsilon^2}$$

We will now upper-bound the expectation. In the robust linear bandit setting, the space $\mathcal{W}$ is equipped with a notion of inner product. We'll use this to equip the hypothesis space $\mathcal{Z}$ with something which is almost an inner product (but which does not correspond to either of the two specified norms on $\mathcal{Z}$), namely,

$$\langle M, M' \rangle_\zeta := \mathop{\mathbb{E}}_{N \sim \zeta} \left[\left\langle F_{a_N^*, M}(\overline{M}(a_N^*)), F_{a_N^*, M'}(\overline{M}(a_N^*)) \right\rangle_W \right]$$

The reason for this definition is

$$g_M^2(a_N^*) = ||F_{a_N^*, M}(\overline{M}(a_N^*))||_{2,W}^2 = \langle F_{a_N^*, M}(\overline{M}(a_N^*)), F_{a_N^*, M}(\overline{M}(a_N^*)) \rangle_W$$

Which implies that $\mathop{\mathbb{E}}_{N \sim \zeta}[g_M^2(a_N^*)] = \langle M, M \rangle_\zeta$. Due to the inner product on $\mathcal{W}$, and the linearity of F in the second subscript argument, this "inner product" on $\mathcal{Z}$ obeys all properties of an inner product except one. There might be nonzero vectors M where $\langle M, M \rangle_\zeta = 0$. These are exactly the M where, with probability one according to $N \sim \zeta$, $F_{a_N^*, M}(\overline{M}(a_N^*)) = 0$.

By linear algebra, we have the following. There is a subspace $\mathcal{Z}'$ where this "inner product" is a true inner product, and a canonical surjection $pr : \mathcal{Z} \to \mathcal{Z}'$, and for all $M \in \mathcal{Z}$, we have $\langle M, M \rangle_\zeta = \langle pr(M), pr(M) \rangle_\zeta$.

We'll now prove that, with probability 1 according to $N \sim \zeta$, we have that for all $M \in \mathcal{Z}$, $g_M^2(a_N^*) = g_{pr(M)}^2(a_N^*)$. This occurs because, by the definition of g, the inner product on $\mathcal{W}$, and linearity of F in its second subscript argument, we have

$$g_M^2(a_N^*) = ||F_{a_N^*, M}(\overline{M}(a_N^*))||_{2,W}^2 = \langle F_{a_N^*, M}(\overline{M}(a_N^*)), F_{a_N^*, M}(\overline{M}(a_N^*)) \rangle_W =$$

$$\left\langle F_{a_N^*, pr(M)}(\overline{M}(a_N^*)) + F_{a_N^*, M-pr(M)}(\overline{M}(a_N^*)), F_{a_N^*, pr(M)}(\overline{M}(a_N^*)) + F_{a_N^*, M-pr(M)}(\overline{M}(a_N^*)) \right\rangle_W$$

There's a finite collection of vectors z_i which serve as a basis for the subspace of vectors of the form $M - pr(M)$, and for all of these vectors, with probability 1 when $N \sim \zeta$, $F_{a_N^*, z_i}(\overline{M}(a_N^*)) = 0$. So, with probability 1, for every M, we have

$$= \langle F_{a_N^*, pr(M)}(\overline{M}(a_N^*)), F_{a_N^*, pr(M)}(\overline{M}(a_N^*)) \rangle_W = g_{pr(M)}^2(a_N^*)$$

Now, we can return to the term we were trying to upper-bound, and go

$$\max_{\varepsilon \geq \varepsilon'} \frac{\underset{N \sim \zeta}{\mathbb{E}}\left[\underset{M \in G_{\varepsilon,\zeta}}{\max} g_M^2(a_N^*)\right]}{\varepsilon^2} = \max_{\varepsilon \geq \varepsilon'} \frac{\underset{N \sim \zeta}{\mathbb{E}}\left[\underset{M \in G_{\varepsilon,\zeta}}{\max} g_{pr(M)}^2(a_N^*)\right]}{\varepsilon^2}$$

Then, by the definition of $G_{\varepsilon,\zeta}$ as the M where $\underset{N \sim \zeta}{\mathbb{E}}[g_M^2(a_N^*)] \leq \varepsilon^2$, and our rephrasing of this as an "inner product", and the fact that this "inner product" doesn't change under projection, we can rewrite as

$$= \max_{\varepsilon \geq \varepsilon'} \frac{\underset{N \sim \zeta}{\mathbb{E}}\left[\underset{M \in \mathcal{Z}:\langle M,M\rangle_\zeta \leq \varepsilon^2}{\max} g_{pr(M)}^2(a_N^*)\right]}{\varepsilon^2} = \max_{\varepsilon \geq \varepsilon'} \frac{\underset{N \sim \zeta}{\mathbb{E}}\left[\underset{M \in \mathcal{Z}:\langle pr(M),pr(M)\rangle_\zeta \leq \varepsilon^2}{\max} g_{pr(M)}^2(a_N^*)\right]}{\varepsilon^2}$$

$$= \max_{\varepsilon \geq \varepsilon'} \frac{\underset{N \sim \zeta}{\mathbb{E}}\left[\underset{M' \in \mathcal{Z}':\langle M',M'\rangle_\zeta \leq \varepsilon^2}{\max} g_{M'}^2(a_N^*)\right]}{\varepsilon^2}$$

Now, since the "inner product" is a true inner product on $\mathcal{Z}'$, we're maximizing over an L2 ball, so every M' fulfilling those conditions can be written as a spherical combination of orthogonal basis vectors $z_i \in \mathcal{Z}'$, which all have an inner product of ε^2. Then unpack definitions and use linearity of F in its second subscript argument.

$$= \max_{\varepsilon \geq \varepsilon'} \frac{\underset{N \sim \zeta}{\mathbb{E}}\left[\underset{\alpha_i:\sum_i \alpha_i^2 \leq 1}{\max} g_{\sum_i \alpha_i z_i}^2(a_N^*)\right]}{\varepsilon^2} = \max_{\varepsilon \geq \varepsilon'} \frac{\underset{N \sim \zeta}{\mathbb{E}}\left[\underset{\alpha_i:\sum_i \alpha_i^2 \leq 1}{\max} ||F_{a_N^*,\sum_i \alpha_i z_i}(\overline{M}(a_N^*))||_{2,W}^2\right]}{\varepsilon^2}$$

$$= \max_{\varepsilon \geq \varepsilon'} \frac{\underset{N \sim \zeta}{\mathbb{E}}\left[\underset{\alpha_i:\sum_i \alpha_i^2 \leq 1}{\max} ||\sum_i \alpha_i F_{a_N^*,z_i}(\overline{M}(a_N^*))||_{2,W}^2\right]}{\varepsilon^2}$$

Upper-bound the norm of the sum by the sum of the norms, and use the Cauchy-Schwartz inequality.

$$\leq \max_{\varepsilon \geq \varepsilon'} \frac{\underset{N \sim \zeta}{\mathbb{E}}\left[\underset{\alpha_i:\sum_i \alpha_i^2 \leq 1}{\max} \left(\sum_i |\alpha_i| \cdot ||F_{a_N^*,z_i}(\overline{M}(a_N^*))||_{2,W}\right)^2\right]}{\varepsilon^2}$$

$$\leq \max_{\varepsilon \geq \varepsilon'} \frac{\underset{N \sim \zeta}{\mathbb{E}}\left[\underset{\alpha_i:\sum_i \alpha_i^2 \leq 1}{\max} \left(\sum_i \alpha_i^2\right)\left(\sum_i ||F_{a_N^*,z_i}(\overline{M}(a_N^*))||_{2,W}^2\right)\right]}{\varepsilon^2}$$

$$= \max_{\varepsilon \geq \varepsilon'} \frac{\underset{N \sim \zeta}{\mathbb{E}}\left[\sum_{i=1}^{\dim \mathcal{Z}'} ||F_{a_N^*,z_i}(\overline{M}(a_N^*))||_{2,W}^2\right]}{\varepsilon^2} = \max_{\varepsilon \geq \varepsilon'} \frac{\sum_{i=1}^{\dim \mathcal{Z}'} \underset{N \sim \zeta}{\mathbb{E}}\left[||F_{a_N^*,z_i}(\overline{M}(a_N^*))||_{2,W}^2\right]}{\varepsilon^2}$$

Pack it back up as an inner product, use that the z_i were selected to all have an inner product of ε^2, and use that Z is the dimension of $\mathcal{Z}$.

$$= \max_{\varepsilon \geq \varepsilon'} \frac{\sum_{i=1}^{\dim_{Z'}} \langle z_i, z_i \rangle_\zeta}{\varepsilon^2} \leq \max_{\varepsilon \geq \varepsilon'} \frac{Z\varepsilon^2}{\varepsilon^2} = Z$$

And so the disagreement coefficient has a maximum value of Z, the dimension of the hypothesis space. ∎

Lemma 8 *Let κ be $7\left(\frac{1}{sine} + 1\right) R\sqrt{WZ}$ and γ be $\geq 2e^2\kappa$ where e is Euler's constant. We then have that for all beliefs $\overline{M} : \mathcal{A} \to \Delta\mathcal{O}$,*

$$dec_\gamma^o(\mathcal{H}, \overline{M}) \leq \frac{\kappa^2 \ln^2\left(\frac{\gamma}{2\kappa}\right)}{\gamma}$$

Begin by unpacking the definition of the offset DEC.

$$\text{dec}_\gamma^o(\mathcal{H}, \overline{M}) = \min_{p \in \Delta\mathcal{A}} \max_{\mu \in \Delta\mathcal{H}} \mathbb{E}_{M \sim \mu}[\max(f^M)] - \mathbb{E}_{a \sim p}[f^{\overline{M}}(a)] - \gamma \mathbb{E}_{a,M \sim p,\mu}[D_H^2(\overline{M}(a) \to M(a))]$$

This is affine in both arguments. The hypothesis space and action space are compact, so $\Delta\mathcal{H}$ and $\Delta\mathcal{A}$ are compact. Therefore, we can use Sion's minimax theorem, to swap the max and the min.

$$= \max_{\mu \in \Delta\mathcal{H}} \min_{p \in \Delta\mathcal{A}} \mathbb{E}_\mu\left[\max(f^M)\right] - \mathbb{E}_p\left[f^{\overline{M}}\right] - \gamma\mathbb{E}_{p,\mu}[D_H^2(\overline{M}(a) \to M(a))]$$

Given μ, let p be the distribution over actions produced by sampling $M \sim \mu$, and selecting the optimal action, a_M^*.

$$\leq \max_{\mu \in \Delta\mathcal{H}} \mathbb{E}_{M \sim \mu}\left[f^M(a_M^*) - f^{\overline{M}}(a_M^*)\right] - \gamma \mathbb{E}_{M,N \sim \mu}[D_H^2(\overline{M}(a_N^*) \to M(a_N^*))]$$

Use that $\sqrt{2}$ times the Hellinger distance exceeds total variation distance, so 2 times the Hellinger-squared distance exceeds the total variation distance squared.

$$\leq \max_{\mu \in \Delta\mathcal{H}} \mathbb{E}_\mu\left[f^M(a_M^*) - f^{\overline{M}}(a_M^*)\right] - \frac{\gamma}{2} \mathbb{E}_{M,N \sim \mu}\left[D_{TV}^2(\overline{M}(a_N^*) \to M(a_N^*))\right]$$

Letting ν_M be the point in $M(a_M^*)$ which minimizes total variation distance to $\overline{M}(a_M^*)$, we have

$$f^M(a_M^*) - f^{\overline{M}}(a_M^*) = \min_{\nu \in M(a_M^*)} \mathbb{E}_{o \sim \nu}[r(a_M^*, o)] - \mathbb{E}_{o \sim \overline{M}(a_M^*)}[r(a_M^*, o)]$$

$$\leq \mathbb{E}_{\nu_M}[r(a_M^*, o)] - \mathbb{E}_{\overline{M}(a_M^*)}[r(a_M^*, o)] \leq D_{TV}(\nu_M, \overline{M}(a_M^*)) = D_{TV}(\overline{M}(a_M^*) \to M(a_M^*))$$

So, we can continue to upper-bound by

$$\leq \max_{\mu \in \Delta\mathcal{H}} \mathbb{E}_\mu[D_{TV}(\overline{M}(a_M^*) \to M(a_M^*))] - \frac{\gamma}{2} \mathbb{E}_{M,N \sim \mu}\left[D_{TV}^2(\overline{M}(a_N^*) \to M(a_N^*))\right]$$

At this point, we use that the norm on $\mathbb{R}^{\mathcal{O}}$ is the L1 norm, which coincides with total variation distance, to rewrite as

$$\leq \max_{\mu \in \Delta \mathcal{H}} \mathbb{E}_\mu [D_{\mathbb{R}^{\mathcal{O}}}(\overline{M}(a_M^*) \to M(a_M^*))] - \frac{\gamma}{2} \mathop{\mathbb{E}}_{M,N \sim \mu} \left[D_{\mathbb{R}^{\mathcal{O}}}^2(\overline{M}(a_N^*) \to M(a_N^*)) \right]$$

Now, note that the squared distance term (with γ) being 1 or more on all μ implies that the offset DEC is 0 or less and the lemma holds, so for the μ which is picked, we may assume that the squared distance times $\frac{\gamma}{2}$ is ≤ 1. $\gamma \geq 14e^2$ because we assumed that $\gamma \geq 2\kappa e^2$ and $\kappa \geq 7$, so the squared distance term (without γ) is $\leq \frac{1}{7e^2}$. This will be used later. By Lemmas A.2 and A.7 from Kosoy (2024), we have

$$\leq \max_{\mu \in \Delta \mathcal{H}} \mathbb{E}_\mu \left[4 \left(\frac{1}{\mathrm{sine}(K_{a_M^*,M}, \Delta \mathcal{O})} + 1 \right) D_{\mathbb{R}^{\mathcal{O}}}(\overline{M}(a_M^*) \to K_{a_M^*,M}) \right]$$

$$- \frac{\gamma}{2} \mathop{\mathbb{E}}_{M,N \sim \mu} \left[D_{\mathbb{R}^{\mathcal{O}}}^2(\overline{M}(a_N^*) \to M(a_N^*)) \right]$$

Letting sine be defined as $\min_{M \in \mathcal{H}, a \in \mathcal{A}} \mathrm{sine}(K_{a,M}^b(a), \Delta \mathcal{O})$, we can then reexpress as

$$\leq \max_{\mu \in \Delta \mathcal{H}} \mathbb{E}_\mu \left[4 \left(\frac{1}{\mathrm{sine}} + 1 \right) D_{\mathbb{R}^{\mathcal{O}}}(\overline{M}(a_M^*) \to K_{a_M^*,M}) \right] - \frac{\gamma}{2} \mathop{\mathbb{E}}_{M,N \sim \mu} \left[D_{\mathbb{R}^{\mathcal{O}}}^2(\overline{M}(a_N^*) \to M(a_N^*)) \right]$$

Now, we'll move to measuring distance in $\mathcal{W}$. It is immediate from the definitions of the relevant norms that, for any action a, model M, and points $y, y' \in \mathbb{R}^{\mathcal{O}}$, $D_{\mathbb{R}^{\mathcal{O}}}(y, y') \geq ||F_{a,M}(y) - F_{a,M}(y')||_{2,W}$, because all $F_{a,M}$ map the unit ball of the norm on $\mathbb{R}^{\mathcal{O}}$ inside the unit ball of the L2 norm on $\mathcal{W}$. Using this fact, along with the fact that $M(a_N^*)$ consists of all the points in $\Delta \mathcal{O}$ which $F_{a_N^*,M}$ maps to zero, we have that $D_{\mathbb{R}^{\mathcal{O}}}^2(\overline{M}(a_N^*) \to M(a_N^*)) \geq ||F_{a_N^*,M}(\overline{M}(a_N^*))||_{2,W}^2$ which yields

$$\leq \max_{\mu \in \Delta \mathcal{H}} \mathbb{E}_\mu \left[4 \left(\frac{1}{\mathrm{sine}} + 1 \right) D_{\mathbb{R}^{\mathcal{O}}}(\overline{M}(a_M^*) \to K_{a_M^*,M}) \right] - \frac{\gamma}{2} \mathop{\mathbb{E}}_{M,N \sim \mu} \left[||F_{a_N^*,M}(\overline{M}(a_N^*))||_{2,W}^2 \right]$$

In the other direction, we have $D_{\mathbb{R}^{\mathcal{O}}}(y, y') \leq ||F_{a,M}(y) - F_{a,M}(y')||_{\mathrm{small},W}$ by the definition of the small norm on $\mathcal{W}$. Again using that $K_{a_M^*,M}$ consists of all the points which $F_{a_M^*,M}$ maps to zero, we have $D_{\mathbb{R}^{\mathcal{O}}}(\overline{M}(a_M^*) \to K_{a_M^*,M}) \leq ||F_{a_M^*,M}(\overline{M}(a_M^*))||_{\mathrm{small},W}$ which yields

$$\leq \max_{\mu \in \Delta \mathcal{H}} \mathbb{E}_\mu \left[4 \left(\frac{1}{\mathrm{sine}} + 1 \right) ||F_{a_M^*,M}(\overline{M}(a_M^*))||_{\mathrm{small},W} \right] - \frac{\gamma}{2} \mathop{\mathbb{E}}_{M,N \sim \mu} \left[||F_{a_N^*,M}(\overline{M}(a_N^*))||_{2,W}^2 \right]$$

Now we use that $||w||_{\mathrm{small},W} = R \cdot ||w||_{\mathrm{big},W} \leq R\sqrt{W}||w||_{2,W}$, by the definition of the big $\mathcal{W}$ norm, the definition of the L2 norm on $\mathcal{W}$, and John's Ellipsoid Theorem.

$$\leq \max_{\mu \in \Delta \mathcal{H}} \mathbb{E}_\mu \left[4 \left(\frac{1}{\mathrm{sine}} + 1 \right) R\sqrt{W}||F_{a_M^*,M}(\overline{M}(a_M^*))||_{2,W} \right] - \frac{\gamma}{2} \mathop{\mathbb{E}}_{M,N \sim \mu} \left[||F_{a_N^*,M}(\overline{M}(a_N^*))||_{2,W}^2 \right]$$

At this point, we introduce the abbreviations

$$g_M(a) := ||F_{a,M}(\overline{M}(a_M^*))||_{2,W}$$

$$H := \mathbb{E}_{M,N\sim\mu}[g_M^2(a_N^*)]$$

To abbreviate our equation as

$$\leq \max_{\mu\in\Delta\mathcal{H}} 4\left(\frac{1}{\sin e}+1\right) R\sqrt{W} \mathop{\mathbb{E}}_{M\sim\mu}[g_M(a_M^*)] - \frac{\gamma}{2}H \tag{1}$$

Now, we take a detour to upper-bound that expectation. Via page 105, Lemma E.2 of Foster et al. (2021), we can upper-bound the expectation term with Foster's disagreement coefficient. We have that $g_M(a) \leq 1$, because regardless of action or hypothesis, F maps $\Delta\mathcal{O}$ into the unit ball of our L2 norm, so Lemma E.2 can be invoked. Expressing the result in our notation, define $\mathcal{G} := \{g_M \mid M \in \mathcal{H}\}$. Foster's result from Lemma E.2 was

$$\mathop{\mathbb{E}}_{M\sim\mu}[g_M(a_M^*)]$$

$$\leq \min_{\Delta,\varepsilon\in(0,1],\eta>0} \Delta + \frac{\eta\varepsilon^2}{2} + \frac{1}{2\eta}\left(4\,\mathrm{dis}\left(\mathcal{G},\Delta,\varepsilon,\mu\right)\ln\left(\frac{1}{\varepsilon}\right)\ln\left(\frac{1}{\Delta}\right)+1\right) + \frac{\eta}{2}\mathop{\mathbb{E}}_{M,N\sim\mu}[g_M^2(a_N^*)]$$

Where $\mathrm{dis}(\mathcal{G},\Delta,\varepsilon,\mu)$ is the disagreement coefficient. Applying Lemma 7 to upper-bound the disagreement coefficient by Z, and using our abbreviation H, we can proceed to

$$\mathop{\mathbb{E}}_{M\sim\mu}[g_M(a_M^*)] \leq \min_{\Delta,\varepsilon\in(0,1],\eta>0} \Delta + \frac{\eta\varepsilon^2}{2} + \frac{1}{2\eta}\left(4Z\ln\left(\frac{1}{\varepsilon}\right)\ln\left(\frac{1}{\Delta}\right)+1\right) + \frac{\eta}{2}H$$

Let Δ,ε be $\sqrt{H}$, and rewrite to

$$\leq \min_{\eta>0} \sqrt{H} + \eta H + \frac{1}{2\eta}\left(Z\ln^2\left(\frac{1}{H}\right)+1\right)$$

Now, choose η to be $\sqrt{\frac{Z}{2H}}\ln\left(\frac{1}{H}\right)$, and remember that $Z = \dim(\mathcal{Z}) \geq 1$ to get

$$\leq \sqrt{H} + \frac{1}{\sqrt{2}}\sqrt{ZH}\ln\left(\frac{1}{H}\right) + \frac{1}{\sqrt{2}}\sqrt{ZH}\ln\left(\frac{1}{H}\right) + \frac{1}{\sqrt{2}}\sqrt{\frac{H}{Z}}\ln^{-1}\left(\frac{1}{H}\right)$$

$$= \sqrt{\frac{ZH}{2}}\ln\left(\frac{1}{H}\right)\left(\sqrt{\frac{2}{Z}}\ln^{-1}\left(\frac{1}{H}\right)+2+\frac{1}{Z}\ln^{-2}\left(\frac{1}{H}\right)\right)$$

$$\leq \sqrt{\frac{ZH}{2}}\ln\left(\frac{1}{H}\right)\left(\sqrt{2}\ln^{-1}\left(\frac{1}{H}\right)+2+\ln^{-2}\left(\frac{1}{H}\right)\right)$$

Following the definition of that H term, it equals $\mathop{\mathbb{E}}_{M,N\sim\mu}\left[\|F_{a,M}(\overline{M}(a_M^*))\|_{2,W}^2\right]$. We had previously derived that $D_{\mathbb{R}\mathcal{O}}^2(\overline{M}(a_N^*) \to M(a_N^*)) \geq \|F_{a_N^*,M}(\overline{M}(a_N^*))\|_{2,W}^2$, so H is upper-bounded by $\mathop{\mathbb{E}}_{M,N\sim\mu}\left[D_{\mathbb{R}\mathcal{O}}^2(\overline{M}(a_N^*) \to M(a_N^*))\right]$. This quantity was previously demonstrated to be $\leq \frac{1}{7e^2}$. Putting these inequalities together, we can upper-bound H by $\frac{1}{7e^2}$, and proceed to an upper bound of

$$\leq 1.72\sqrt{ZH}\ln\left(\frac{1}{H}\right)$$

Plugging this upper bound back into 1 and using that $4 \cdot 1.72 \le 7$, we get

$$\le \max_{\mu \in \Delta \mathcal{H}} 7 \left(\frac{1}{\text{sine}} + 1 \right) R\sqrt{WZH} \ln \left(\frac{1}{H} \right) - \frac{\gamma}{2} H$$

The only quantity that depends on μ is H, though it is suppressed in the notation. We shift to maximizing over the value of H explicitly. $H \in [0, 1]$ by its definition and the fact that the g functions are in $[0, 1]$. We also abbreviate $7 \left(\frac{1}{\text{sine}} + 1 \right) R\sqrt{WZ}$ as κ, to get

$$\le \max_{H \in [0,1]} 7 \left(\frac{1}{\text{sine}} + 1 \right) R\sqrt{WZH} \ln \left(\frac{1}{H} \right) - \frac{\gamma}{2} H = \max_{H \in [0,1]} \kappa\sqrt{H} \ln \left(\frac{1}{H} \right) - \frac{\gamma}{2} H$$

$$\le \max \left(\max_{H \in \left[0, \frac{4\kappa^2}{\gamma^2} \right]} \left(\kappa\sqrt{H} \ln \left(\frac{1}{H} \right) - \frac{\gamma}{2} H \right), \max_{H \in \left[\frac{4\kappa^2}{\gamma^2}, 1 \right]} \left(\kappa\sqrt{H} \ln \left(\frac{\gamma^2}{4\kappa^2} \right) - \frac{\gamma}{2} H \right) \right)$$

We'll show that the first term has a positive derivative (so the maximizing H must be at the edge, and subsumed by the second term), and then exactly maximize the second term. For the first term, take the derivative to yield

$$-\frac{\kappa}{\sqrt{H}} + \frac{\kappa}{2\sqrt{H}} \ln \left(\frac{1}{H} \right) - \frac{\gamma}{2} \ge -\frac{\kappa}{\sqrt{H}} + \frac{\kappa}{2\sqrt{H}} \ln \left(\frac{\gamma^2}{4\kappa^2} \right) - \frac{\gamma}{2} = -\frac{\kappa}{\sqrt{H}} + \frac{\kappa}{\sqrt{H}} \ln \left(\frac{\gamma}{2\kappa} \right) - \frac{\gamma}{2}$$

$$= \frac{\kappa}{\sqrt{H}} \left(\ln \left(\frac{\gamma}{2\kappa} \right) - 1 \right) - \frac{\gamma}{2} \ge \frac{\gamma}{2} \left(\ln \left(\frac{\gamma}{2\kappa} \right) - 1 \right) - \frac{\gamma}{2} = \frac{\gamma}{2} \left(\ln \left(\frac{\gamma}{2\kappa} \right) - 2 \right) \ge 0$$

The first and second inequality follow because $\frac{1}{H} \ge \frac{\gamma^2}{4\kappa^2}$. The third inequality was our starting assumption that $\gamma \ge 2e^2\kappa$ and $\ln(e^2) = 2$. So, because the derivative is never negative, the maximizing value for H is $\frac{4\kappa^2}{\gamma^2}$, which is subsumed by the later max, so our maximization over H can be written as

$$= \max_{H \in \left[\frac{4\kappa^2}{\gamma^2}, 1 \right]} \left(\kappa\sqrt{H} \ln \left(\frac{\gamma^2}{4\kappa^2} \right) - \frac{\gamma}{2} H \right) = \max_{H \in \left[\frac{4\kappa^2}{\gamma^2}, 1 \right]} \left(2\kappa\sqrt{H} \ln \left(\frac{\gamma}{2\kappa} \right) - \frac{\gamma}{2} H \right)$$

The exact maximizing value of H is $\frac{4\kappa^2 \ln^2\left(\frac{\gamma}{2\kappa} \right)}{\gamma^2}$. Because we assumed that $\gamma \ge 2e^2\kappa$, this value is compatible with the range that H may lie within. The exact maximum is then

$$= \frac{2\kappa^2 \ln^2 \left(\frac{\gamma}{2\kappa} \right)}{\gamma}$$

And our result follows by chaining all inequalities together. ∎

Theorem 3 *In the robust linear bandit setting, for all $\varepsilon < \frac{1}{e^2}$ (Euler's constant),*

$$dec_\varepsilon^f(\mathcal{H}) \le 16 \left(\frac{1}{\text{sine}} + 1 \right) R\sqrt{WZ}\varepsilon \ln \left(\frac{1}{\varepsilon} \right)$$

By Proposition 1, to express the fuzzy DEC in terms of the offset DEC, and Proposition 2, to assume, without loss of generality, that probabilistic $\overline{M}$ suffice to upper-bound the fuzzy DEC with respect to squared Hellinger loss, we may consider only probabilistic $\overline{M}$, and compute

$$\text{dec}_\varepsilon^f(\mathcal{H}, \overline{M}) = \min_{\gamma \geq 0} \max(\text{dec}_\gamma^o(\mathcal{H}, \overline{M}), 0) + \gamma\varepsilon^2$$

Select $\gamma := \frac{\sqrt{2}\kappa}{\varepsilon} \ln\left(\frac{1}{\varepsilon}\right)$, where $\kappa := 7\left(\frac{1}{\text{sine}} + 1\right) R\sqrt{WZ}$, as defined in Lemma 8. Note that, in particular, since $\varepsilon \leq e^{-2}$ by assumption, we have $\gamma \geq 2e^2\kappa$, so we can safely invoke Lemma 8 to bound the offset DEC. So, we can rewrite as

$$\leq \frac{2\kappa^2 \ln^2\left(\frac{\gamma}{2\kappa}\right)}{\gamma} + \gamma\varepsilon^2 = \sqrt{2}\kappa\varepsilon \left(\frac{\ln^2\left(\frac{1}{\sqrt{2}\varepsilon} \ln\left(\frac{1}{\varepsilon}\right)\right)}{\ln\left(\frac{1}{\varepsilon}\right)} + \ln\left(\frac{1}{\varepsilon}\right)\right)$$

The ratio of the fraction on the left side, and $\ln(\frac{1}{\varepsilon})$, never exceeds 1.6 for ε in our range of interest, as can be verified by a graphing calculator, so we can upper-bound by

$$< \sqrt{2}\kappa\varepsilon \cdot 2.6 \ln\left(\frac{1}{\varepsilon}\right) < 16\left(\frac{1}{\text{sine}} + 1\right) R\sqrt{WZ}\varepsilon \ln\left(\frac{1}{\varepsilon}\right)$$

∎

Lemma 9 *For any two hypotheses $M, M' \in \mathcal{H}$,*
$\max_a D_H^2(M(a), M'(a)) \leq 4\left(\frac{1}{sine} + 1\right) R\sqrt{Z}||M' - M||_{Z,2}$, *where the Hellinger distance is symmetric Hellinger distance, and the norm is the L2 norm on $\mathcal{Z}$ from the definitions for robust linear bandits.*

We start with the fact that total variation distance upper-bounds Hellinger-squared distance, and that total variation distance coincides with the norm on $\mathbb{R}^{\mathcal{O}}$.

$$\max_a D_H^2(M(a) \to M'(a)) \leq \max_a D_{TV}(M(a) \to M'(a))$$

$$= \max_a D_{\mathbb{R}^{\mathcal{O}}}(M(a) \to M'(a)) = \max_a \max_{\mu \in M(a)} d_{\mathbb{R}^{\mathcal{O}}}(\mu \to M'(a))$$

Using Lemmas A.2 and A.7 from Kosoy (2024), and Kosoy's definition of the sine parameter, we have that $D_{\mathbb{R}^{\mathcal{O}}}(\mu \to M'(a)) \leq 4\left(\frac{1}{\text{sine}} + 1\right) D_{\mathbb{R}^{\mathcal{O}}}(\mu \to K_{a,M'})$, yielding

$$\leq 4\left(\frac{1}{\text{sine}} + 1\right) \max_a \max_{\mu \in M(a)} D_{\mathbb{R}^{\mathcal{O}}}(\mu \to K_{a,M'})$$

We have that, for all $y \in \mathbb{R}^{\mathcal{O}}$, $D_{\mathbb{R}^{\mathcal{O}}}(\mu, y) \leq ||F_{a,M'}(\mu) - F_{a,M'}(y)||_{\text{small},W}$ by the definition of the small norm on $\mathcal{W}$. Using that $K_{a,M'}$ consists of all the points which $F_{a,M'}$ maps to zero, we have that $D_{\mathbb{R}^{\mathcal{O}}}(\mu \to K_{a,M'}) \leq ||F_{a,M'}(\mu)||_{\text{small},W}$. Additionally, we have that $||w||_{\text{small},W} = R \cdot ||w||_{\text{big},W}$, so we may upper-bound by

$$\leq 4\left(\frac{1}{\text{sine}} + 1\right) R\max_a \max_{\mu \in M(a)} ||F_{a,M'}(\mu)||_{W,\text{big}}$$

Now, because $\mu \in M(a)$, $F_{a,M}(\mu) = 0$. Therefore, with linearity of F, the definitions of our norms on $\mathcal{Z}$, and John's ellipsoid theorem, we get

$$= 4\left(\frac{1}{\text{sine}} + 1\right) R \max_a \max_{y \in M(a)} ||F_{a,M'-M}(y)||_{W,\text{big}}$$

$$\leq 4\left(\frac{1}{\text{sine}} + 1\right) R \max_a \max_{y:||y||_{\mathbb{R}\mathcal{O}} \leq 1} ||F_{a,M'-M}(y)||_{W,\text{big}}$$

$$= 4\left(\frac{1}{\text{sine}} + 1\right) R ||M' - M||_Z \leq 4\left(\frac{1}{\text{sine}} + 1\right) R\sqrt{Z} ||M' - M||_{Z,2}$$

And so, our overall bound, for any M, M', is

$$\max_a D_H^2(M(a) \to M'(a)) \leq 4\left(\frac{1}{\text{sine}} + 1\right) R\sqrt{Z} ||M' - M||_{Z,2}$$

This same bound applies if we switch M and M', so our desired upper bound on $\max_a D_H^2(M(a), M'(a))$ follows. ∎

Lemma 10 *An upper bound on the number of L2 balls of radius ε needed to cover the cube $[-1, 1]^Z$ is $\left(\frac{\sqrt{Z}}{2\varepsilon} + 1\right)^Z$.*

Cover the $[-1, 1]^Z$ cube with a uniform grid that has one point evenly spaced every $\frac{2\varepsilon}{\sqrt{Z}}$ distance. There are, at most, $\left(\frac{\sqrt{Z}}{2\varepsilon} + 1\right)^Z$-many grid points. To show that this cover is suitable, and every point is within an L2 distance of ε of a point on this grid, we can take an arbitrary point x and round it off to the closest value on the grid, in every coordinate, to produce some x'. The difference in every coordinate is $\leq \frac{\varepsilon}{\sqrt{Z}}$, and then we can compute

$$||x - x'||_2 = \sqrt{\sum_{i=1}^{Z} (x_i - x'_i)^2} \leq \sqrt{\sum_{i=1}^{Z} \left(\frac{\varepsilon}{\sqrt{Z}}\right)^2} = \frac{\varepsilon}{\sqrt{Z}}\sqrt{Z} = \varepsilon$$

Thereby verifying that we have constructed an appropriately fine cover. ∎

Proposition 6 *For the robust linear bandit setting, $\mathcal{N}(\mathcal{H}, \varepsilon) \leq \left(\frac{4\left(\frac{1}{\text{sine}}+1\right)RZ}{\varepsilon^2} + 1\right)^Z$*

By Lemma 10, there is a cover of $[-1, 1]^Z$ (the unit ball in $\mathcal{Z}$ according to L_∞ distance, with the orthonormal basis given by the L_2 norm on $\mathcal{Z}$) with $\left(\frac{4\left(\frac{1}{\text{sine}}+1\right)RZ}{\varepsilon^2} + 1\right)^Z$-many balls of radius $\frac{\varepsilon^2}{8\left(\frac{1}{\text{sine}}+1\right)R\sqrt{Z}}$. This produces an ε-cover of $\mathcal{H}$ with respect to Hellinger distance, by taking every ball with a nonempty intersection with $\mathcal{H}$, and picking an arbitrary point in it. Here is why.

Given any $M \in \mathcal{H}$, it is in some occupied L2 ball, which has had an arbitrary $M' \in \mathcal{H}$ selected from it. By the triangle inequality, $||M' - M||_{Z,2} \leq \frac{\varepsilon^2}{4\left(\frac{1}{\text{sine}}+1\right)R\sqrt{Z}}$. Then, by Lemma 9, this yields

$$\max_a D_H^2(M(a), M'(a)) \leq \varepsilon^2$$

So, every $M \in \mathcal{H}$ has some M' in our finite covering set, where, for all actions, $\varepsilon^2 \geq \max_a D_H^2(M(a), M'(a))$, which occurs iff $\varepsilon \geq \max_a D_H(M(a), M'(a))$. Therefore, we have constructed an ε-cover of our hypothesis space. $\blacksquare$

Appendix H. Definitions for Robust Markov Decision Processes

H is the number of timesteps in an episode. T is the number of episodes. $\mathcal{S}$ is a finite state space, and S is its cardinality. $\mathcal{A}$ is a finite space of actions, and A is its cardinality. Timesteps, states, actions, and rewards are denoted by h, s, a, r. s_h, a_h, r_h denote the state, action, and reward on timestep h.

To denote a set of timesteps, we use the notation $[H]$ to denote the set $\{1, 2...H\}$, and $[H]_0$ to denote the set $\{0, 1, ...H\}$. $\mathbb{P}$ is used for probabilities of events, instead of expectations.

The space of policies Π_{RNS} is defined as $[H] \times \mathcal{S} \to \Delta\mathcal{A}$. Elements are denoted π. $\pi(h, s)$ denotes the probability distribution over actions that π plays when it is in state s on timestep h, so $\pi(h, s)(a)$ is the probability of action a. Given a selection σ of an RMDP, defined in Definition 5, $\sigma \bowtie \pi$ is the corresponding distribution over trajectories produced by σ interacting with π.

We identify RMDP's with their transition kernels $\mathbb{M}$, of type $[H]_0 \times \mathcal{S} \times \mathcal{A} \to \square([0, 1] \times \mathcal{S})$. The initial state, initial action, and terminal state s_0, a_0, s_{H+1} are considered to be unique, so trajectories are of the form $r_0, s_1, a_1, r_1, ..., r_H$. We also identify RMDP's with their imprecise models $\pi \mapsto \{\sigma \bowtie \pi \mid \sigma \models \mathbb{M}\}$. $\sigma \models \mathbb{M}$ is defined in Definition 5.

Given a distribution over trajectories μ, $\mu|h, s, a$ is an abbreviation for $\mu|s_h = s, a_h = a$. Similarly, taking the probability of h, s, a is an abbreviation for the probability of $s_h = s, a_h = a$. The same pattern generalizes to conditioning on, or asking for the probability of, h, s. $\mu_{\downarrow h}$ denotes this distribution projected to the r_h, s_{h+1} coordinates.

$\mathbb{M}$ is 1-bounded if there is a selection $\sigma \models \mathbb{M}$ where, for all π, h, s, a, $\underset{\sigma \bowtie \pi | h, s, a}{\mathbb{E}} \left[\sum_{k=h}^H r_k\right] \leq 1$. For a fixed $H, \mathcal{S}, \mathcal{A}$, $\mathcal{H}$ denotes the set of all 1-bounded RMDP's.

A function $\overline{M}$ of type $\Pi_{RNS} \to \Delta\left([0, 1] \times (\mathcal{S} \times \mathcal{A} \times [0, 1])^{[H]}\right)$ is policy-coherent if it fulfills the following property. For all policies, partial trajectories that end in a state s, and actions, the probability of a according to $\overline{M}(\pi)$ conditioned on the trajectory equals $\pi(h, s)(a)$. Intuitively, $\overline{M}$ is policy-coherent if $\overline{M}(\pi)$ is always a distribution over trajectories that looks like the actions were generated by the policy π.

There is a natural function of type $[0, 1] \to \Delta\{0, 1\}$, which induces an affine function convert : $\Delta([0, 1] \times \mathcal{S}) \to \Delta(\{0, 1\} \times \mathcal{S})$, which induces a function convert : $\square([0, 1] \times \mathcal{S}) \to$

$\Box(\{0,1\} \times \mathcal{S})$. $D^2_{H_{\{0,1\}}}(\Psi \to \Phi)$ is an abbreviation for $D^2_H(\text{convert}(\Psi) \to \text{convert}(\Phi))$.

For the RMDP proofs, our notion of loss $\mathcal{L}(\overline{M}, \mathbb{M}, \pi)$ is

$$\mathbb{E}_{\overline{M}(\pi)} \left[\sum_{h=0}^H D^2_{H_{\{0,1\}}} \left((\overline{M}(\pi)|h, s_h, a_h)_{\downarrow h} \to \mathbb{M}(h, s_h, a_h) \right) \right]$$

$D^2_{H_{\{0,1\}}}$ measures squared Hellinger error, after converting distributions over $[0,1]$ to distributions over $\{0,1\}$. $\overline{M}(\pi)$ is a distribution over trajectories, so we have to condition it on arriving at a specific state and action on timestep h, and project out the next reward r_h and next state s_{h+1}, to find its prediction for what happens next. $\mathbb{M}(h, s, a)$ is just the transition kernel for our RMDP of choice. The modified offset DEC and modified fuzzy DEC are defined by taking the definitions of the offset DEC and fuzzy DEC, and swapping out Hellinger-squared loss for the above notion of loss.

The function $\text{odec}(\gamma, p, M, N) : \mathbb{R}^{\geq 0} \times \Delta\mathcal{A} \times (\mathcal{A} \to [0,2]) \times (\mathcal{A} \to [0,2]) \to (-\infty, 1]$ is defined as

$$\text{odec}(\gamma, p, M, N) := \max_{a'} N(a') - \mathbb{E}_{a \sim p}[M(a)] - \gamma \mathbb{E}_{a \sim p} \left[(N(a) - M(a))^2 \right]$$

Given a policy-coherent belief $\overline{M}$, a policy π, and a $\gamma \geq 0$, $b_{h,s}^{\overline{M}(\pi)}$ (the reward bonus at h, s), $\overline{V}_{h,s}^{\overline{M}(\pi)}$ (the expected future reward plus bonuses at h, s, but clipped to be 1 at most) and $\overline{Q}_{h,s,a}^{\overline{M}(\pi)}$ (the Q-value of $\overline{V}^{\overline{M}(\pi)}$ at h, s, a) are defined by downwards induction as follows. $\overline{V}_{H+1,s_{H+1}}^{\overline{M}(\pi)}$ is taken to be zero. Then, we have

$$\overline{Q}_{h,s,a}^{\overline{M}(\pi)} = \mathbb{E}_{\overline{M}(\pi)|h,s,a} \left[r_h + \overline{V}_{h+1,s_{h+1}}^{\overline{M}(\pi)} \right]$$

$$b_{h,s}^{\overline{M}(\pi)} := \min_p \max_N \text{odec} \left(\frac{\gamma}{8} \mathbb{P}_{\overline{M}(\pi)}(h,s), p, \lambda a.\overline{Q}_{h,s,a}^{\overline{M}(\pi)}, N \right)$$

$$\overline{V}_{h,s}^{\overline{M}(\pi)} := \min \left(1, b_{h,s}^{\overline{M}(\pi)} + \mathbb{E}_{a \sim \pi(h,s)} \left[\overline{Q}_{h,s,a}^{\overline{M}(\pi)} \right] \right)$$

This is well-defined for all h, s and h, s, a which $\overline{M}(\pi)$ does not assign zero probability to. The definitions may be extended to zero with our convention that s_0 is an initial state and a_0 is an initial action. $\overline{Q}_{0,s_0,a_0}^{\overline{M}(\pi)}$ reduces to $\mathbb{E}_{\overline{M}(\pi)} \left[r_0 + \overline{V}_{1,s_1}^{\overline{M}(\pi)} \right]$. $b_{0,s_0}^{\overline{M}(\pi)}$ is rendered well-defined and nonzero by looking at the value of the odec function for the one-armed bandit, instead of the A-armed bandit. Finally, we have $\overline{V}_{0,s_0}^{\overline{M}(\pi)} = \min \left(1, b_{h,s}^{\overline{M}(\pi)} + \overline{Q}_{0,s_0,a_0}^{\overline{M}(\pi)} \right)$.

$\Delta_{\geq \varepsilon}\mathcal{A}$ denotes the space of all probability distributions on $\mathcal{A}$ such that, for every action, the probability of that action is $\geq \frac{\varepsilon}{A}$. $\mathbf{1}$ is used to denote the indicator function which is 1 if an event happens and 0 if it doesn't.

47

Appendix I. RMDP Lemmas

Lemma 11 *Given two compact Polish spaces A, B, and a continuous function $f : A \times B \to \mathbb{R}$, the function $\lambda b.\min_a f(a, b)$ is continuous, as is $\lambda b.\max_a f(a, b)$.*

By the definition of "Polish space", we can equip A, B with metrics D_A, D_B, to make them into compact metric spaces. The product of compact metric spaces can be made into a compact metric space via a metric of $D((a, a'), (b, b')) = \max(D_A(a, a'), D_B(b, b'))$. We then establish continuity of $\lambda b.\max_a f(a, b)$ as follows. If b_n limits to b, then the functions $\lambda a.f(a, b_n)$ limit (in the sup-norm) to $\lambda a.f(a, b)$, by

$$\limsup_{n \to \infty} \max_a |f(a, b_n) - f(a, b)| \leq \limsup_{n \to \infty} \max_{a, b': D((a, b'), (a, b)) \leq D_B(b_n, b)} |f(a, b') - f(a, b)| = 0$$

The inequality is a simple consequence of our definitions of distance. The equality is because, by the Heine-Cantor theorem, f is uniformly continuous. For any $\varepsilon > 0$, uniform continuity of f yields a δ where points being δ apart implies that f can only vary by ε between those two points, and as n increases, b_n will eventually stay only δ apart from b, witnessing that the limsup is ε or less. This argument works for any ε, so the limsup is zero.

Therefore, $\lambda a.f(a, b_n)$ limits to $\lambda a.f(a, b)$ in the sup-norm, so the difference between their maximum values limit to zero, and we have that $\max_a f(a, b_n)$ converges to $\max_a f(a, b)$. This argument works for any convergent sequence b_n, so $\lambda b.\max_a f(a, b)$ is a continuous function. The same line of argument works for min as well. ∎

Lemma 12 *Given three compact Polish spaces X, Y, Z, and a continuous function $f : X \times Y \times Z \to \mathbb{R}$, the function $\lambda z.\min_x \max_y f(x, y, z)$ is continuous.*

By assumption, $\lambda x, y, z.f(x, y, z)$ is continuous and everything is a compact Polish space. A product of two compact Polish spaces is compact Polish. Invoking Lemma 11 with $A = Y$ and $B = X \times Z$ we get that $\lambda x, z.\max_y f(x, y, z)$ is continuous. Invoking Lemma 11 a second time with $A = X$ and $B = Z$, we get that $\lambda z.\min_x \max_y f(x, y, z)$ is continuous and Lemma 12 is proven. ∎

For Lemma 13, define the following construction. Given a compact Polish space X, $X^\dagger$ denotes the quotient of the space $X \times [0, 1]$ formed by identifying all $(x, 0)$ pairs.

Lemma 13 *If a policy-coherent belief $\overline{M} : \Pi_{RNS} \to \Delta([0, 1] \times (\mathcal{S} \times \mathcal{A} \times [0, 1])^{[H]})$ is continuous, then for all h, s, a and $\varepsilon > 0$, the following functions of type signature $([H] \times \mathcal{S} \to \Delta_{\geq \varepsilon} \mathcal{A}) \to [0, 2]^\dagger$ are continuous. The Q-value function $\lambda \pi. \left(\overline{Q}_{h,s,a}^{\overline{M}(\pi)}, \mathbb{P}_{\overline{M}(\pi)}(h, s) \right)$, the bonus function $\lambda \pi. \left(b_{h,s}^{\overline{M}(\pi)}, \mathbb{P}_{\overline{M}(\pi)}(h, s) \right)$ and the expected value function $\lambda \pi. \left(\overline{V}_{h,s}^{\overline{M}(\pi)}, \mathbb{P}_{\overline{M}(\pi)}(h, s) \right)$*

To establish this, we use a downwards induction proof, where we assume continuity of these functions at level $h+1$, and derive their continuity at level h. To show the base case that continuity holds at $h = H+1$, we use that the expected value $\overline{V}_{H+1,s}^{\overline{M}(\pi)}$ is always zero by definition and does not depend on the choice of π. The probability of $H+1, s$ is zero if s isn't the terminal state, and one if it is, and again, there is no dependence on π. Constant functions are continuous, so our base case is proven.

For the induction step, we first notice that, for the various functions, by the definition of $[0,2]^{\dagger}$, continuity is proven by showing that, if π_n converges to π, the probability term converges, and if the probability converges to > 0, the values converge as well.

Convergence of probabilities holds because, for any h, s, by the starting assumption of the continuity of $\overline{M}$, it is immediate that $\mathbb{P}_{\overline{M}(\pi_n)}(h, s)$ converges to $\mathbb{P}_{\overline{M}(\pi)}(h, s)$.

Now we must show convergence of values if the limiting probability for h, s is nonzero so we may freely assume that $\mathbb{P}_{\overline{M}(\pi)}(h, s) > 0$.

For Q-values, we can expand the definition.

$$\lim_{n\to\infty} \overline{Q}_{h,s,a}^{\overline{M}(\pi_n)} = \lim_{n\to\infty} \mathbb{E}_{\overline{M}(\pi_n)|h,s,a} \left[r_h + \overline{V}_{h+1,s_{h+1}}^{\overline{M}(\pi_n)} \right]$$

$$= \lim_{n\to\infty} \frac{\mathbb{E}_{\overline{M}(\pi_n)} \left[\mathbf{1}_{s_h=s} \cdot \mathbf{1}_{a_h=a} \cdot \left(r_h + \overline{V}_{h+1,s_{h+1}}^{\overline{M}(\pi_n)} \right) \right]}{\mathbb{E}_{\overline{M}(\pi_n)} \left[\mathbf{1}_{s_h=s} \cdot \mathbf{1}_{a_h=a} \right]}$$

$$= \lim_{n\to\infty} \frac{\mathbb{E}_{\overline{M}(\pi_n)} \left[\mathbf{1}_{s_h=s} \cdot \mathbf{1}_{a_h=a} \cdot \left(r_h + \overline{V}_{h+1,s_{h+1}}^{\overline{M}(\pi_n)} \right) \right]}{\mathbb{P}_{\overline{M}(\pi_n)}(h, s) \cdot \pi_n(h, s)(a)}$$

The quantity in the denominator clearly converges to $\mathbb{P}_{\overline{M}(\pi)}(h, s) \cdot \pi(h, s)(a)$. This limiting value is nonzero, because $\mathbb{P}_{\overline{M}(\pi)}(h, s) > 0$ by assumption, and $\pi(h, s)(a) \geq \frac{\varepsilon}{A}$ because we are restricting our attention to policies of type $[H] \times \mathcal{S} \to \Delta_{\geq\varepsilon}\mathcal{A}$.

$$= \lim_{n\to\infty} \frac{\mathbb{E}_{\overline{M}(\pi_n)} \left[\mathbf{1}_{s_h=s} \cdot \mathbf{1}_{a_h=a} \cdot \left(r_h + \overline{V}_{h+1,s_{h+1}}^{\overline{M}(\pi_n)} \right) \right]}{\mathbb{P}_{\overline{M}(\pi)}(h, s) \cdot \pi(h, s)(a)}$$

Now, as π_n limits to π, for any given s', by our induction assumption of continuity at $h+1$, either $\overline{V}_{h+1,s'}^{\overline{M}(\pi_n)}$ limits to $\overline{V}_{h+1,s'}^{\overline{M}(\pi)}$, or the probability of s' at time $h+1$ limits to zero. If we have convergence of values only holding for states with nonzero limiting probability, the expectation values will converge anyways, and we have

$$= \lim_{n\to\infty} \frac{\mathbb{E}_{\overline{M}(\pi_n)} \left[\mathbf{1}_{s_h=s} \cdot \mathbf{1}_{a_h=a} \cdot \left(r_h + \overline{V}_{h+1,s_{h+1}}^{\overline{M}(\pi)} \right) \right]}{\mathbb{P}_{\overline{M}(\pi)}(h, s) \cdot \pi(h, s)(a)}$$

Now, by continuity of $\overline{M}$, $\overline{M}(\pi_n)$ converges to $\overline{M}(\pi)$. The contents of the expectation in the numerator are continuous as a function from trajectories to $\mathbb{R}$, so their expectation

49

values converge, and we have

$$= \frac{\mathbb{E}_{\overline{M}(\pi)}\left[\mathbf{1}_{s_h=s} \cdot \mathbf{1}_{a_h=a} \cdot \left(r_h + \overline{V}_{h+1,s_{h+1}}^{\overline{M}(\pi)}\right)\right]}{\mathbb{P}_{\overline{M}(\pi)}(h,s) \cdot \pi(h,s)(a)} = \underset{\overline{M}(\pi)|h,s,a}{\mathbb{E}}\left[r_h + \overline{V}_{h+1,s_{h+1}}^{\overline{M}(\pi)}\right] = \overline{Q}_{h,s,a}^{\overline{M}(\pi)}$$

And we have shown continuity for the Q-values for h and all s, a.

To show continuity of the bonus values for h and all s, we use continuity of the Q-values for h and all s, a, and can freely assume that $\mathbb{P}_{\overline{M}(\pi)}(h,s) > 0$. By definition, we have

$$\lim_{n\to\infty} b_{h,s}^{\overline{M}(\pi_n)} = \lim_{n\to\infty} \min_p \max_N \text{ odec}\left(\frac{\gamma}{8}\mathbb{P}_{\overline{M}(\pi_n)}(h,s), p, \lambda a.\overline{Q}_{h,s,a}^{\overline{M}(\pi_n)}, N\right)$$

The odec function can be verified, by casual inspection of its definition, to be continuous in all arguments. All parameters come from compact Polish spaces, namely $[0, \frac{\gamma}{8}]$, $\Delta\mathcal{A}$, $\mathcal{A} \to [0,2]$, and $\mathcal{A} \to [0,2]$. Therefore, we can invoke Lemma 12 to get

$$= \min_p \max_N \text{ odec}\left(\lim_{n\to\infty}\frac{\gamma}{8}\mathbb{P}_{\overline{M}(\pi_n)}(h,s), p, \lim_{n\to\infty}\left(\lambda a.\overline{Q}_{h,s,a}^{\overline{M}(\pi_n)}\right), N\right)$$

The probabilities converge, as previously argued, and we have already proven that the Q-values converge, as long as the limiting probability of h, s is above zero, which it is assumed to be, so we have

$$= \min_p \max_N \text{ odec}\left(\frac{\gamma}{8}\mathbb{P}_{\overline{M}(\pi)}(h,s), p, \lambda a.\overline{Q}_{h,s,a}^{\overline{M}(\pi)}, N\right) = b_{h,s}^{\overline{M}(\pi)}$$

And continuity is shown for the bonus values.

Now, to show continuity for the expected values we have

$$\lim_{n\to\infty} \overline{V}_{h,s}^{\overline{M}(\pi_n)} = \lim_{n\to\infty} \min\left(1, b_{h,s}^{\overline{M}(\pi_n)} + \underset{a\sim\pi_n(h,s)}{\mathbb{E}}\left[\overline{Q}_{h,s,a}^{\overline{M}(\pi_n)}\right]\right)$$

By continuity of the bonus values and Q-values (which we have just proven in our induction step) and our assumption that h, s has a nonzero limiting probability, the expected values converge.

$$= \min\left(1, b_{h,s}^{\overline{M}(\pi)} + \underset{a\sim\pi(h,s)}{\mathbb{E}}\left[\overline{Q}_{h,s,a}^{\overline{M}(\pi)}\right]\right) = \overline{V}_{h,s}^{\overline{M}(\pi)}$$

And so, continuity is established for the expected values. The downwards induction step has been proven, so Lemma 13 follows, that for all h, s, a, the Q-values, bonus values, and expected values are continuous as a function of π. ∎

Lemma 14 *If a policy-coherent belief $\overline{M} : \Pi_{RNS} \to \Delta([0,1] \times (\mathcal{S} \times \mathcal{A} \times [0,1])^{[H]})$ is continuous, then for any $\varepsilon > 0, \gamma > 0$, there exists a policy $\tilde{\pi} \in \Pi_{RNS}$, where, for all $h, s \in [H]_0 \times \mathcal{S}$ that $\overline{M}(\tilde{\pi})$ assigns nonzero probability,*

$$b_{h,s}^{\overline{M}(\tilde{\pi})} + \left(2 + \frac{\gamma}{2}\right)\varepsilon$$

$$\geq \max_{N \in \mathcal{A} \to [0,2]} \left(\max_{a'} N(a') - \mathbb{E}_{a \sim \tilde{\pi}(h,s)} \left[\overline{Q}_{h,s,a}^{\overline{M}(\tilde{\pi})} \right] - \frac{\gamma}{8} \mathbb{P}_{\overline{M}(\tilde{\pi})}(h,s) \mathbb{E}_{a \sim \tilde{\pi}(h,s)} \left[\left(N(a) - \overline{Q}_{h,s,a}^{\overline{M}(\tilde{\pi})} \right)^2 \right] \right)$$

As in the definitions, $\Delta_{\geq \varepsilon} \mathcal{A}$ is the set of all probability distributions which assign $\geq \frac{\varepsilon}{A}$ probability to every action. Using the definitions, and letting $u_{\mathcal{A}}$ be the uniform distribution on actions, consider the following set-valued function $([H] \times \mathcal{S} \to \Delta_{\geq \varepsilon} \mathcal{A}) \to \mathcal{P}([H] \times \mathcal{S} \to \Delta_{\geq \varepsilon} \mathcal{A})$. $\Pi_{h,s \in [H] \times \mathcal{S}}$ denotes the Cartesian product of sets in this formula.

$$\lambda \pi . \Pi_{h,s \in [H] \times \mathcal{S}} \text{ if } \mathbb{P}_{\overline{M}(\pi)}(h,s) = 0, \{\varepsilon u_{\mathcal{A}} + (1-\varepsilon)\nu \mid \nu \in \Delta \mathcal{A}\} \tag{2}$$

$$\text{else} \left\{ \varepsilon u_{\mathcal{A}} + (1-\varepsilon)\nu \;\middle|\; \nu \in \operatorname*{argmin}_{p} \max_{N} \text{odec}\left(\frac{\gamma}{8} \mathbb{P}_{\overline{M}(\pi)}(h,s), p, \lambda a . \overline{Q}_{h,s,a}^{\overline{M}(\pi)}, N \right) \right\}$$

First, we will show that any fixpoint of this set-valued function has our desired property. Then we will show that, by Kakutani's fixpoint theorem, a suitable fixpoint exists.

Accordingly, let $\tilde{\pi}$ be a fixpoint of the above equation. Then, for any h, s with nonzero probability according to $\overline{M}(\tilde{\pi})$, $\tilde{\pi}(h,s)$ is of the form $\varepsilon u_{\mathcal{A}} + (1-\varepsilon)\nu$ where ν is the distribution over actions which minimizes the odec function.

$$\max_{N \in \mathcal{A} \to [0,2]} \max_{a'} N(a') - \mathbb{E}_{a \sim \tilde{\pi}(h,s)} \left[\overline{Q}_{h,s,a}^{\overline{M}(\tilde{\pi})} \right] - \frac{\gamma}{8} \mathbb{P}_{\overline{M}(\tilde{\pi})}(h,s) \mathbb{E}_{a \sim \tilde{\pi}(h,s)} \left[\left(N(a) - \overline{Q}_{h,s,a}^{\overline{M}(\tilde{\pi})} \right)^2 \right]$$

$$= \max_{N \in \mathcal{A} \to [0,2]} \max_{a'} N(a') - \mathbb{E}_{a \sim \varepsilon u_{\mathcal{A}} + (1-\varepsilon)\nu} \left[\overline{Q}_{h,s,a}^{\overline{M}(\tilde{\pi})} \right]$$

$$- \frac{\gamma}{8} \mathbb{P}_{\overline{M}(\tilde{\pi})}(h,s) \mathbb{E}_{a \sim \varepsilon u_{\mathcal{A}} + (1-\varepsilon)\nu} \left[\left(N(a) - \overline{Q}_{h,s,a}^{\overline{M}(\tilde{\pi})} \right)^2 \right]$$

Because these functions are bounded in $[0,2]$ (for the Q values, it is because reward is bounded in $[0,1]$ and so is $\overline{V}$), we can upper-bound by

$$\leq \max_{N \in \mathcal{A} \to [0,2]} \max_{a'} N(a') - \mathbb{E}_{a \sim \nu} \left[\overline{Q}_{h,s,a}^{\overline{M}(\tilde{\pi})} \right] + 2 D_{TV}(\varepsilon u_A + (1-\varepsilon)\nu, \nu)$$

$$- \frac{\gamma}{8} \mathbb{P}_{\overline{M}(\tilde{\pi})}(h,s) \left(\mathbb{E}_{a \sim \nu} \left[\left(N(a) - \overline{Q}_{h,s,a}^{\overline{M}(\tilde{\pi})} \right)^2 \right] - 4 D_{TV}(\varepsilon u_A + (1-\varepsilon)\nu, \nu) \right)$$

The total variation terms are both upper-bounded by ε, which can be pulled out of the maximum. Upper-bounding the probability term by 1, we get

$$\leq \left(2 + \frac{\gamma}{2} \right) \varepsilon + \max_{N \in \mathcal{A} \to [0,2]} \max_{a'} N(a') - \mathbb{E}_{a \sim \nu} \left[\overline{Q}_{h,s,a}^{\overline{M}(\tilde{\pi})} \right] - \frac{\gamma}{8} \mathbb{P}_{\overline{M}(\tilde{\pi})}(h,s) \left(\mathbb{E}_{a \sim \nu} \left[\left(N(a) - \overline{Q}_{h,s,a}^{\overline{M}(\tilde{\pi})} \right)^2 \right] \right)$$

The above equation can be rephrased as

$$= \left(2 + \frac{\gamma}{2} \right) \varepsilon + \max_{N \in \mathcal{A} \to [0,2]} \text{odec}\left(\frac{\gamma}{8} \mathbb{P}_{\overline{M}(\tilde{\pi})}(h,s), \nu, \lambda a . \overline{Q}_{h,s,a}^{\overline{M}(\tilde{\pi})}, N \right)$$

51

Using that ν is the minimizer of the above equation, we can rephrase, and pack up the definition of $b_{h,s}^{\overline{M}(\tilde{\pi})}$.

$$= \left(2 + \frac{\gamma}{2}\right)\varepsilon + \min_p \max_{N \in \mathcal{A} \to [0,2]} \text{odec}\left(\frac{\gamma}{8}\mathbb{P}_{\overline{M}(\tilde{\pi})}(h,s), p, \lambda a.\overline{Q}_{h,s,a}^{\overline{M}(\tilde{\pi})}, N\right) = \left(2 + \frac{\gamma}{2}\right)\varepsilon + b_{h,s}^{\overline{M}(\tilde{\pi})}$$

Putting all inequalities together, we have proven our desired property, as long as there is a $\tilde{\pi}$ which is a fixpoint of 2.

We now shift to proving that a fixpoint exists. The topological conditions for Kakutani's fixpoint theorem are easy to verify. $[H] \times \mathcal{S} \to \Delta_{\geq\varepsilon}\mathcal{A}$ is clearly a compact, convex, nonempty subset of a Euclidean space. That just leaves showing nonemptiness, convexity, and closed graph of the set-valued function.

To show nonemptiness, the product of nonempty sets is nonempty, so it suffices to prove nonemptiness for all component sets of the product. If $\mathbb{P}_{\overline{M}(\pi)}(h,s) = 0$, nonemptiness is trivial. If it exceeds zero, then Lemma 13 shows the Q-values are well-defined, and then the arguments in Lemma 12, along with the continuity of the odec function, prove that the function
$\lambda p.\max_N \text{odec}\left(\frac{\gamma}{8}\mathbb{P}_{\overline{M}(\pi)}(h,s), p, \lambda a.\overline{Q}_{h,s,a}^{\overline{M}(\pi)}, N\right)$ is continuous. The distribution p is from $\Delta\mathcal{A}$
which is compact, and continuous functions from compact spaces to $\mathbb{R}$ have minimizers, so the set of minimizers is nonempty. And so, nonemptiness has been shown.

To show convexity, the product of convex sets is convex, so it suffices to prove convexity for all the component sets of the product. Given a convex set of probability distributions C mixing ε of the uniform distribution with all distributions in C preserves convexity, so we just need to show convexity holds before mixing with the uniform distribution. If $\mathbb{P}_{\overline{M}(\pi)}(h,s) = 0$, this is trivial, because $\Delta\mathcal{A}$ is convex. If it exceeds zero, then Lemma 13 shows the Q-values are well-defined, and then, letting ν_1 and ν_2 be minimizers, and $q \in [0,1]$, we have, by the odec function being affine in the distribution over actions, convexity of maximization, and the defining property of ν_1 and ν_2,

$$\max_N \text{odec}\left(\frac{\gamma}{8}\mathbb{P}_{\overline{M}(\pi)}(h,s), q\nu_1 + (1-q)\nu_2, \lambda a.\overline{Q}_{h,s,a}^{\overline{M}(\pi)}, N\right)$$

$$= \max_N(q \cdot \text{odec}\left(\frac{\gamma}{8}\mathbb{P}_{\overline{M}(\pi)}(h,s), \nu_1, \lambda a.\overline{Q}_{h,s,a}^{\overline{M}(\pi)}, N\right)$$

$$+ (1-q)\text{odec}\left(\frac{\gamma}{8}\mathbb{P}_{\overline{M}(\pi)}(h,s), \nu_2, \lambda a.\overline{Q}_{h,s,a}^{\overline{M}(\pi)}, N\right))$$

$$\leq q \max_N \text{odec}\left(\frac{\gamma}{8}\mathbb{P}_{\overline{M}(\pi)}(h,s), \nu_1, \lambda a.\overline{Q}_{h,s,a}^{\overline{M}(\pi)}, N\right)$$

$$+ (1-q)\max_N \text{odec}\left(\frac{\gamma}{8}\mathbb{P}_{\overline{M}(\pi)}(h,s), \nu_2, \lambda a.\overline{Q}_{h,s,a}^{\overline{M}(\pi)}, N\right)$$

$$= \min_{p \in \Delta \mathcal{A}} \max_N \text{odec}\left(\frac{\gamma}{8}\mathbb{P}_{\overline{M}(\pi)}(h,s), p, \lambda a.\overline{Q}_{h,s,a}^{\overline{M}(\pi)}, N\right)$$

However, this inequality cannot be strict because otherwise $q\nu_1 + (1-q)\nu_2$ would yield a lower value than the minimum possible, so it must be an equality, witnessing that $q\nu_1 + (1-q)\nu_2$ is a minimizer as well. q and the choices of ν_1 and ν_2 were arbitrary, so any mixture of minimizers is a minimizer, and the argmin set is convex. Convexity is thereby established.

To show closed graph, let π_n limit to π and assume that π_n' are in the sets associated with π_n, and that they limit to π'. Showing that the limiting π' lies in the set associated with π verifies the closed graph property. This occurs iff, for all h, s, we have

$$\text{if } \mathbb{P}_{\overline{M}(\pi)}(h,s) = 0, \pi'(h,s) \in \{\varepsilon u_{\mathcal{A}} + (1-\varepsilon)\nu \mid \nu \in \Delta\mathcal{A}\}$$

$$\text{else } \pi'(h,s) \in \left\{\varepsilon u_{\mathcal{A}} + (1-\varepsilon)\nu \ \middle| \ \nu \in \underset{p}{\text{argmin}} \max_N \text{odec}\left(\frac{\gamma}{8}\mathbb{P}_{\overline{M}(\pi)}(h,s), p, \lambda a.\overline{Q}_{h,s,a}^{\overline{M}(\pi)}, N\right)\right\}$$

Now, given $\pi'(h,s)$, it is possible to uniquely recover the corresponding distribution ν via $\frac{\pi'(h,s) - \varepsilon u_{\mathcal{A}}}{(1-\varepsilon)}$. This is a continuous function of the policy, so from the $\pi_n'(h,s)$ sequence, we get a sequence ν_n limiting to some ν, where, for all n, $\pi_n'(h,s) = \varepsilon u_{\mathcal{A}} + (1-\varepsilon)\nu_n$, and the same holds for the limiting π' and ν. Therefore, we can show that $\pi'(h,s)$ lies in the needed set by showing that the limiting ν fulfills the appropriate condition. Proving this for arbitrary h, s verifies the closed graph property, so let h, s be arbitrary.

If $\mathbb{P}_{\overline{M}(\pi)}(h,s) = 0$, we only need to verify that the limiting ν lies in $\Delta\mathcal{A}$, which is trivial. If this quantity exceeds zero, we must verify that

$$\max_N \text{odec}\left(\frac{\gamma}{8}\mathbb{P}_{\overline{M}(\pi)}(h,s), \nu, \lambda a.\overline{Q}_{h,s,a}^{\overline{M}(\pi)}, N\right) = \min_p \max_N \text{odec}\left(\frac{\gamma}{8}\mathbb{P}_{\overline{M}(\pi)}(h,s), p, \lambda a.\overline{Q}_{h,s,a}^{\overline{M}(\pi)}, N\right)$$

This equality can be verified as follows. By Lemma 13, and ν being the limit of the ν_n, we have

$$\max_N \text{odec}\left(\frac{\gamma}{8}\mathbb{P}_{\overline{M}(\pi)}(h,s), \nu, \lambda a.\overline{Q}_{h,s,a}^{\overline{M}(\pi)}, N\right)$$

$$= \max_N \text{odec}\left(\lim_{n\to\infty}\frac{\gamma}{8}\mathbb{P}_{\overline{M}(\pi_n)}(h,s), \lim_{n\to\infty}\nu_n, \lim_{n\to\infty}\lambda a.\overline{Q}_{h,s,a}^{\overline{M}(\pi_n)}, N\right)$$

By Lemma 11, and continuity of the odec function, we can show that the function $\lambda\gamma, \nu, M.\max_N \text{dec}(\gamma, \nu, M, N)$ is continuous, yielding

$$= \lim_{n\to\infty} \max_N \text{odec}\left(\frac{\gamma}{8}\mathbb{P}_{\overline{M}(\pi_n)}(h,s), \nu_n, \lambda a.\overline{Q}_{h,s,a}^{\overline{M}(\pi_n)}, N\right)$$

Then since $\mathbb{P}_{\overline{M}(\pi)}(h,s) > 0$, all but finitely many of the π_n fulfill $\mathbb{P}_{\overline{M}(\pi_n)}(h,s) > 0$, so their corresponding ν_n fulfill

$$\max_N \text{odec}\left(\frac{\gamma}{8}\mathbb{P}_{\overline{M}(\pi_n)}(h,s), \nu_n, \lambda a.\overline{Q}_{h,s,a}^{\overline{M}(\pi_n)}, N\right)$$

53

$$= \min_p \max_N \text{odec} \left(\frac{\gamma}{8} \mathbb{P}_{\overline{M}(\pi_n)}(h, s), p, \lambda a. \overline{Q}_{h,s,a}^{\overline{M}(\pi_n)}, N \right)$$

This yields

$$= \lim_{n \to \infty} \min_p \max_N \text{odec} \left(\frac{\gamma}{8} \mathbb{P}_{\overline{M}(\pi_n)}(h, s), p, \lambda a. \overline{Q}_{h,s,a}^{\overline{M}(\pi_n)}, N \right)$$

Then, by Lemma 12 and the continuity of the odec function, along with Lemma 13 witnessing that the Q-values and probabilities are continuous, we have

$$= \min_p \max_N \text{odec} \left(\lim_{n \to \infty} \frac{\gamma}{8} \mathbb{P}_{\overline{M}(\pi_n)}(h, s), p, \lim_{n \to \infty} \lambda a. \overline{Q}_{h,s,a}^{\overline{M}(\pi_n)}, N \right)$$

$$= \min_p \max_N \text{odec} \left(\lim_{n \to \infty} \frac{\gamma}{8} \mathbb{P}_{\overline{M}(\pi)}(h, s), p, \lim_{n \to \infty} \lambda a. \overline{Q}_{h,s,a}^{\overline{M}(\pi)}, N \right)$$

Our necessary equality, to show that the limiting ν is a minimizer, has been verified. This establishes that $\pi'(h, s)$ lies in the appropriate set. h, s were arbitrary, and so we have established closed graph for our set-valued function.

As nonemptiness, convexity, and closed graph have been verified, a fixpoint exists by Kakutani's Fixpoint Theorem, and our result follows. ∎

Appendix J. RMDP Theorems

Lemma 15 *In the episodic RMDP setting, if $\overline{M} : \Pi_{RNS} \to \Delta([0, 1] \times (\mathcal{S} \times \mathcal{A} \times [0, 1])^H)$ is continuous and policy-coherent, the modified offset DEC fulfills $\text{dec}_\gamma^{o'}(\mathcal{H}, \overline{M}) \leq \frac{2(HSA+1)}{\gamma}$*

First, we unpack the definition of the modified offset DEC, and distribute the expectations over actions inwards, using policy-coherence of $\overline{M}$.

$$\min_{p \in \Delta\Pi_{RNM}} \max_{\mathbb{M} \in \mathcal{H}} \max(f^{\mathbb{M}}) - \mathop{\mathbb{E}}_{\pi \sim p} \left[f^{\overline{M}}(\pi) \right]$$

$$-\gamma \mathop{\mathbb{E}}_{\pi \sim p} \left[\mathbb{E}_{\overline{M}(\pi)} \left[\sum_{h=0}^H \mathop{\mathbb{E}}_{a \sim \pi(h, s_h)} \left[D_{H_{\{0,1\}}}^2 \left((\overline{M}(\pi)|h, s_h, a)_{\downarrow h} \to \mathbb{M}(h, s_h, a) \right) \right] \right] \right]$$

Because $\overline{M}$ is assumed to be continuous, we can invoke Lemma 14 for an arbitrarily low ε and the γ of choice to construct a suitable policy $\tilde{\pi}$. Let p be the deterministic choice of $\tilde{\pi}$.

$$\leq \max_{\mathbb{M} \in \mathcal{H}} \max(f^{\mathbb{M}}) - f^{\overline{M}}(\tilde{\pi}) - \gamma \mathbb{E}_{\overline{M}(\tilde{\pi})} \left[\sum_{h=0}^H \mathop{\mathbb{E}}_{a \sim \tilde{\pi}(h, s_h)} \left[D_{H_{\{0,1\}}}^2 \left((\overline{M}(\tilde{\pi})|h, s_h, a)_{\downarrow h} \to \mathbb{M}(h, s_h, a) \right) \right] \right]$$

Fix the optimal choice of $\mathbb{M}$. Let $\pi_{\mathbb{M}}$ be the optimal choice of policy to maximize worst-case expected reward. We can now reexpress the reward terms.

$$= \min_{\sigma \models \mathbb{M}} \mathbb{E}_{\sigma \bowtie \pi_{\mathbb{M}}} \left[\sum_{h=0}^H r_h \right] - \mathbb{E}_{\overline{M}(\tilde{\pi})} \left[\sum_{h=0}^H r_h \right] \tag{3}$$

$$-\gamma \mathbb{E}_{\overline{M}(\tilde{\pi})} \left[\sum_{h=0}^{H} \mathop{\mathbb{E}}_{a \sim \tilde{\pi}(h,s_h)} \left[D^2_{H_{\{0,1\}}} \left((\overline{M}(\tilde{\pi})|h, s_h, a)_{\downarrow h} \to \mathbb{M}(h, s_h, a) \right) \right] \right]$$

We now define a new term. Fix the convention that $\overline{V}_{h,s}^{\overline{M}(\tilde{\pi})} = 1$ if $\mathbb{P}_{\overline{M}(\tilde{\pi})}(h, s) = 0$, rendering $\overline{V}_{h,s}^{\overline{M}(\tilde{\pi})}$ well-defined everywhere. Call h, s a "bad state" if $\overline{V}_{h,s}^{\overline{M}(\tilde{\pi})} = 1$. Given a trajectory $s_1, a_1, r_1 ... s_H, a_H, r_H$, define t^* (the dependence on trajectory is suppressed in the notation) as

$$t^* := \max \left\{ h \;\middle|\; 0 \le h \le H + 1 \wedge \forall 0 \le k < h : \overline{V}_{k,s_k}^{\overline{M}(\tilde{\pi})} < 1 \right\}$$

Intuitively, t^* is the first timestep where we encounter a bad state, and is $H+1$ otherwise.

To digress, letting σ^* be

$$\sigma^* := \operatorname*{argmin}_{\sigma \models \mathbb{M}} \mathbb{E}_{\sigma \bowtie \pi_{\mathbb{M}}} \left[\sum_{h=0}^{t^*-1} r_h + \mathbf{1}_{t^*=h} \right]$$

we can modify σ^* to construct a σ' which is also a selection of $\mathbb{M}$ as follows. σ' copies σ^*, except that if it ever hits a bad state, the transition probabilities from that point onward will behave in the way (consistent with $\mathbb{M}$) which minimizes the expected sum of future rewards. We have

$$\min_{\sigma \models \mathbb{M}} \mathbb{E}_{\sigma \bowtie \pi_{\mathbb{M}}} \left[\sum_{h=0}^{H} r_h \right] \le \mathbb{E}_{\sigma' \bowtie \pi_{\mathbb{M}}} \left[\sum_{h=0}^{H} r_h \right] \le \mathbb{E}_{\sigma' \bowtie \pi_{\mathbb{M}}} \left[\sum_{h=0}^{t^*-1} r_h + \mathbf{1}_{t^*=h} \right]$$

$$= \mathbb{E}_{\sigma^* \bowtie \pi_{\mathbb{M}}} \left[\sum_{h=0}^{t^*-1} r_h + \mathbf{1}_{t^*=h} \right] = \min_{\sigma \models \mathbb{M}} \mathbb{E}_{\sigma \bowtie \pi_{\mathbb{M}}} \left[\sum_{h=0}^{t^*-1} r_h + \mathbf{1}_{t^*=h} \right]$$

The first inequality and the last equality are obvious. The second inequality holds because one of our assumptions was that $\mathcal{H}$ consisted of the 1-bounded RMDP's, so for all h, s, π, there was a way that the RMDP could behave which would guarantee that the expected sum of future rewards was in $[0, 1]$. Switching from "receive 1 reward when a bad state is reached" to "receive the worst-case sum of future rewards when a bad state is reached" can only decrease the expected sum of rewards. The first equality holds because ν' only behaves differently than ν after a bad state is reached, so if the reward only depends on the trajectory up to the first bad state, the expected rewards must be the same. Combining this inequality with 3, we get

$$\le \min_{\sigma \models \mathbb{M}} \mathbb{E}_{\sigma \bowtie \pi_{\mathbb{M}}} \left[\sum_{h=0}^{t^*-1} r_h + \mathbf{1}_{t^*=h} \right] - \mathbb{E}_{\overline{M}(\tilde{\pi})} \left[\sum_{h=0}^{H} r_h \right]$$

$$-\gamma \mathbb{E}_{\overline{M}(\tilde{\pi})} \left[\sum_{h=0}^{H} \mathop{\mathbb{E}}_{a \sim \tilde{\pi}(h,s_h)} \left[D^2_{H_{\{0,1\}}} \left((\overline{M}(\tilde{\pi})|h, s_h, a)_{\downarrow h} \to \mathbb{M}(h, s_h, a) \right) \right] \right]$$

Now, let M be the Markov decision process which, in situation h, s, a, selects the distribution $\mu : \Delta([0,1] \times \mathcal{S})$ from $\mathbb{M}(h, s, a)$ which minimizes $D^2_{H_{\{0,1\}}} \left(\overline{M}(\tilde{\pi}) | h, s, a \right)_{\downarrow h}, \mu)$. If the conditional is undefined, M can select any distribution from $\mathbb{M}(h, s, a)$. This is a selection of $\mathbb{M}$, so it can only increase the expected reward. It also lets us rephrase the Hellinger error term. We then rephrase the $\overline{M}(\tilde{\pi})$-expected sum of rewards, to yield

$$\leq \mathbb{E}_{M \bowtie \pi_{\mathbb{M}}} \left[\sum_{h=0}^{t^*-1} r_h + \mathbf{1}_{t^*=h} \right] - \mathbb{E}_{\overline{M}(\tilde{\pi})} \left[\sum_{h=0}^{H} r_h \right]$$

$$-\gamma \mathbb{E}_{\overline{M}(\tilde{\pi})} \left[\sum_{h=0}^{H} \mathbb{E}_{a \sim \tilde{\pi}(h, s_h)} \left[D^2_{H_{\{0,1\}}} \left(\left(\overline{M}(\tilde{\pi}) | h, s_h, a \right)_{\downarrow h}, M(h, s_h, a) \right) \right] \right]$$

Then, creating some more terms, we have

$$= \mathbb{E}_{M \bowtie \pi_{\mathbb{M}}} \left[\sum_{h=0}^{t^*-1} r_h + \mathbf{1}_{t^*=h} \right] - \overline{V}^{\overline{M}(\tilde{\pi})}_{0, s_0} \tag{4}$$

$$+\overline{V}^{\overline{M}(\tilde{\pi})}_{0, s_0} - \mathbb{E}_{\overline{M}(\tilde{\pi})} \left[\sum_{h=0}^{H} r_h + b^{\overline{M}(\tilde{\pi})}_{h, s_h} \right] + \mathbb{E}_{\overline{M}(\tilde{\pi})} \left[\sum_{h=0}^{H} b^{\overline{M}(\tilde{\pi})}_{h, s_h} \right]$$

$$-\gamma \mathbb{E}_{\overline{M}(\tilde{\pi})} \left[\sum_{h=0}^{H} \mathbb{E}_{a \sim \tilde{\pi}(h, s_h)} \left[D^2_{H_{\{0,1\}}} \left(\left(\overline{M}(\tilde{\pi}) | h, s_h, a \right)_{\downarrow h}, M(h, s_h, a) \right) \right] \right]$$

Note that s_0 is the unique initial state. The terms in the second and third lines will be ignored until later, and we'll focus on the term in the first line. We can rewrite $\sum_{h=0}^{t^*-1} r_h$ as $\sum_{h=0}^{H} \mathbf{1}_{t^* \geq h+1} \cdot r_h$, and pull the sum out of the expectation. We also have $\overline{V}^{\overline{M}(\tilde{\pi})}_{0, s_0} = \mathbb{E}_{M \bowtie \pi_{\mathbb{M}}} \left[\mathbf{1}_{t^* \geq 0} \cdot \overline{V}^{\overline{M}(\tilde{\pi})}_{0, s_0} \right]$ holding because $t^* \geq 0$ always holds and expectations of constants are that same constant. $t^* \geq 0$ holds because 0 is always in the set that t^* is the maximum of, due to the relevant for-all statement being vacuously true. Finally, we have $0 = \mathbb{E}_{M \bowtie \pi_{\mathbb{M}}} \left[\mathbf{1}_{t^* \geq H+1} \cdot \overline{V}^{\overline{M}(\tilde{\pi})}_{H+1, s_{H+1}} \right]$ holding because $\overline{V}^{\overline{M}(\tilde{\pi})}_{H+1, s_{H+1}}$ is defined to be zero. Therefore, the first term we are focusing on can be rewritten as

$$\sum_{h=0}^{H} \left(\mathbb{E}_{M \bowtie \pi_{\mathbb{M}}} \left[\mathbf{1}_{t^* \geq h+1} \cdot r_h + \mathbf{1}_{t^*=h} \right] \right) + \mathbb{E}_{M \bowtie \pi_{\mathbb{M}}} \left[\mathbf{1}_{t^* \geq H+1} \cdot \overline{V}^{\overline{M}(\tilde{\pi})}_{H+1, s_{H+1}} \right] - \mathbb{E}_{M \bowtie \pi_{\mathbb{M}}} \left[\mathbf{1}_{t^* \geq 0} \cdot \overline{V}^{\overline{M}(\tilde{\pi})}_{0, s_0} \right]$$

It can be further rewritten as a telescoping sum now.

$$= \sum_{h=0}^{H} \mathbb{E}_{M \bowtie \pi_{\mathbb{M}}} \left[\mathbf{1}_{t^* \geq h+1} \cdot r_h + \mathbf{1}_{t^*=h} \right]$$

$$+ \sum_{h=0}^{H} \left(\mathbb{E}_{M \bowtie \pi_{\mathbb{M}}} \left[\mathbf{1}_{t^* \geq h+1} \cdot \overline{V}^{\overline{M}(\tilde{\pi})}_{h+1, s_{h+1}} \right] - \mathbb{E}_{M \bowtie \pi_{\mathbb{M}}} \left[\mathbf{1}_{t^* \geq h} \cdot \overline{V}^{\overline{M}(\tilde{\pi})}_{h, s_h} \right] \right)$$

Combining the sums and expectations, this can be rewritten as

$$= \sum_{h=0}^{H} \underset{M \bowtie \pi_M}{\mathbb{E}} \left[\mathbf{1}_{t^* \geq h+1} \left(r_h + \overline{V}_{h+1,s_{h+1}}^{\overline{M}(\tilde{\pi})} - \overline{V}_{h,s_h}^{\overline{M}(\tilde{\pi})} \right) + \mathbf{1}_{t^*=h} \left(1 - \overline{V}_{h,s_h}^{\overline{M}(\tilde{\pi})} \right) \right]$$

By the definition of t^* and $h \leq H$, if $t^* = h$, $\overline{V}_{h,s_h}^{\overline{M}(\tilde{\pi})} = 1$, so we can rewrite as

$$= \sum_{h=0}^{H} \underset{M \bowtie \pi_M}{\mathbb{E}} \left[\mathbf{1}_{t^* \geq h+1} \left(r_h + \overline{V}_{h+1,s_{h+1}}^{\overline{M}(\tilde{\pi})} - \overline{V}_{h,s_h}^{\overline{M}(\tilde{\pi})} \right) \right]$$

The expectation can be viewed as an expectation over the history up to s_h, and then an expectation over what r_h and s_{h+1} are. Once s_h has happened, that determines what $\overline{V}_{h,s_h}^{\overline{M}(\tilde{\pi})}$ is, and whether $t^* \geq h + 1$, so those terms can be treated as constants, and the second expectation can be moved inside.

$$= \sum_{h=0}^{H} \underset{M \bowtie \pi_M}{\mathbb{E}} \left[\mathbf{1}_{t^* \geq h+1} \left(\underset{r,s \sim M(h,s_h,\pi_M(h,s_h))}{\mathbb{E}} \left[r + \overline{V}_{h+1,s}^{\overline{M}(\tilde{\pi})} \right] - \overline{V}_{h,s_h}^{\overline{M}(\tilde{\pi})} \right) \right]$$

By $t^* \geq h + 1$ and the definition of t^*, we know that $\overline{V}_{h,s_h}^{\overline{M}(\tilde{\pi})} < 1$. We can now unpack the definition of $\overline{V}_{h,s_h}^{\overline{M}(\tilde{\pi})}$ to yield

$$= \sum_{h=0}^{H} \underset{M \bowtie \pi_M}{\mathbb{E}} \left[\mathbf{1}_{t^* \geq h+1} \left(\underset{r,s \sim M(h,s_h,\pi_M(h,s_h))}{\mathbb{E}} \left[r + \overline{V}_{h+1,s}^{\overline{M}(\tilde{\pi})} \right] - b_{h,s_h}^{\overline{M}(\tilde{\pi})} - \underset{a \sim \tilde{\pi}(h,s_h)}{\mathbb{E}} \left[\overline{Q}_{h,s_h,a}^{\overline{M}(\tilde{\pi})} \right] \right) \right]$$

$$\leq \sum_{h=0}^{H} \underset{M \bowtie \pi_M}{\mathbb{E}} \left[\mathbf{1}_{t^* \geq h+1} \left(-b_{h,s_h}^{\overline{M}(\tilde{\pi})} + \max_{a'} \underset{r,s \sim M(h,s_h,a')}{\mathbb{E}} \left[r + \overline{V}_{h+1,s}^{\overline{M}(\tilde{\pi})} \right] - \underset{a \sim \tilde{\pi}(h,s_h)}{\mathbb{E}} \left[\overline{Q}_{h,s_h,a}^{\overline{M}(\tilde{\pi})} \right] \right) \right]$$

$$(5)$$

Now, because $t^* \geq h + 1$, we have that $\overline{V}_{h,s_h}^{\overline{M}(\tilde{\pi})} < 1$, and this was defined to be one if $\mathbb{P}_{\overline{M}(\tilde{\pi})}(h, s_h) = 0$, so this state has nonzero probability. By our choice of $\tilde{\pi}$, we may now invoke Lemma 14, and choose N to be $\lambda a.\mathbb{E}_{r,s \sim M(h,s_h,a)} \left[r + \overline{V}_{h+1,s}^{\overline{M}(\tilde{\pi})} \right]$, to derive the following inequality for all h, s_h. Note that the ε term is disregarded, because ε can be set as close to zero as we desire.

$$b_{h,s_h}^{\overline{M}(\tilde{\pi})} \geq \max_{a'} \underset{r,s \sim M(h,s_h,a')}{\mathbb{E}} \left[r + \overline{V}_{h+1,s}^{\overline{M}(\tilde{\pi})} \right] - \underset{a \sim \tilde{\pi}(h,s_h)}{\mathbb{E}} \left[\overline{Q}_{h,s_h,a}^{\overline{M}(\tilde{\pi})} \right]$$

$$- \frac{\gamma}{8} \mathbb{P}_{\overline{M}(\tilde{\pi})}(h, s_h) \underset{a \sim \tilde{\pi}(h,s_h)}{\mathbb{E}} \left[\left(\underset{r,s \sim M(h,s_h,a)}{\mathbb{E}} \left[r + \overline{V}_{h+1,s}^{\overline{M}(\tilde{\pi})} \right] - \overline{Q}_{h,s_h,a}^{\overline{M}(\tilde{\pi})} \right)^2 \right]$$

This can be reshuffled into

$$b_{h,s_h}^{\overline{M}(\tilde{\pi})} + \frac{\gamma}{8} \mathbb{P}_{\overline{M}(\tilde{\pi})}(h, s_h) \underset{a \sim \tilde{\pi}(h,s_h)}{\mathbb{E}} \left[\left(\underset{r,s \sim M(h,s_h,a)}{\mathbb{E}} \left[r + \overline{V}_{h+1,s}^{\overline{M}(\tilde{\pi})} \right] - \overline{Q}_{h,s_h,a}^{\overline{M}(\tilde{\pi})} \right)^2 \right]$$

$$\geq \max_{a'} \mathop{\mathbb{E}}_{r,s \sim M(h,s_h,a')} \left[r + \overline{V}_{h+1,s}^{\overline{M}(\tilde{\pi})} \right] - \mathop{\mathbb{E}}_{a \sim \pi(h,s_h)} \left[\overline{Q}_{h,s_h,a}^{\overline{M}(\tilde{\pi})} \right]$$

Applying this upper bound to 5, the bonus terms cancel, and we have

$$\leq \sum_{h=0}^{H} \mathop{\mathbb{E}}_{M \bowtie \pi_{\mathbb{M}}} \left[\mathbf{1}_{t^* \geq h+1} \cdot \frac{\gamma}{8} \mathbb{P}_{\overline{M}(\tilde{\pi})}(h, s_h) \cdot \mathop{\mathbb{E}}_{a \sim \tilde{\pi}(h,s_h)} \left[\left(\mathop{\mathbb{E}}_{r,s \sim M(h,s_h,a)} \left[r + \overline{V}_{h+1,s}^{\overline{M}(\tilde{\pi})} \right] - \overline{Q}_{h,s_h,a}^{\overline{M}(\tilde{\pi})} \right)^2 \right] \right]$$

$$\leq \sum_{h=0}^{H} \mathop{\mathbb{E}}_{M \bowtie \pi_{\mathbb{M}}} \left[\frac{\gamma}{8} \mathbb{P}_{\overline{M}(\tilde{\pi})}(h, s_h) \cdot \mathop{\mathbb{E}}_{a \sim \tilde{\pi}(h,s_h)} \left[\left(\mathop{\mathbb{E}}_{r,s \sim M(h,s_h,a)} \left[r + \overline{V}_{h+1,s}^{\overline{M}(\tilde{\pi})} \right] - \overline{Q}_{h,s_h,a}^{\overline{M}(\tilde{\pi})} \right)^2 \right] \right]$$

$$\leq \frac{\gamma}{8} \sum_{h=0}^{H} \sum_{s \in \mathcal{S}} \mathbb{P}_{\overline{M}(\tilde{\pi})}(h, s_h) \cdot \mathop{\mathbb{E}}_{a \sim \tilde{\pi}(h,s_h)} \left[\left(\mathop{\mathbb{E}}_{r,s \sim M(h,s_h,a)} \left[r + \overline{V}_{h+1,s}^{\overline{M}(\tilde{\pi})} \right] - \overline{Q}_{h,s_h,a}^{\overline{M}(\tilde{\pi})} \right)^2 \right]$$

$$= \frac{\gamma}{8} \sum_{h=0}^{H} \mathbb{E}_{\overline{M}(\tilde{\pi})} \left[\mathop{\mathbb{E}}_{a \sim \tilde{\pi}(h,s_h)} \left[\left(\mathop{\mathbb{E}}_{r,s \sim M(h,s_h,a)} \left[r + \overline{V}_{h+1,s}^{\overline{M}(\tilde{\pi})} \right] - \overline{Q}_{h,s_h,a}^{\overline{M}(\tilde{\pi})} \right)^2 \right] \right]$$

$$= \frac{\gamma}{8} \mathbb{E}_{\overline{M}(\tilde{\pi})} \left[\sum_{h=0}^{H} \mathop{\mathbb{E}}_{a \sim \tilde{\pi}(h,s_h)} \left[\left(\mathop{\mathbb{E}}_{r,s \sim M(h,s_h,a)} \left[r + \overline{V}_{h+1,s}^{\overline{M}(\tilde{\pi})} \right] - \overline{Q}_{h,s_h,a}^{\overline{M}(\tilde{\pi})} \right)^2 \right] \right]$$

$$= \frac{\gamma}{8} \mathbb{E}_{\overline{M}(\tilde{\pi})} \left[\sum_{h=0}^{H} \mathop{\mathbb{E}}_{a \sim \tilde{\pi}(h,s_h)} \left[\left(\mathop{\mathbb{E}}_{r,s \sim M(h,s_h,a)} \left[r + \overline{V}_{h+1,s}^{\overline{M}(\tilde{\pi})} \right] - \mathop{\mathbb{E}}_{(\overline{M}(\tilde{\pi})|h,s_h,a)_{\downarrow h}} \left[r + \overline{V}_{h+1,s}^{\overline{M}(\tilde{\pi})} \right] \right)^2 \right] \right]$$

Passing $M(h, s_h, a)$ and its counterpart through the convert function (to convert $\Delta([0,1] \times \mathcal{S})$ into $\Delta(\{0,1\} \times \mathcal{S})$), does not affect the expected rewards. Because $\overline{V}_{h+1,s}^{\overline{M}(\tilde{\pi})} \leq 1$, and the rewards are in $[0,1]$, these functions are in $[0,2]$, so the difference in expectations can be upper bounded by 2 times the total variation distance between the converted distributions.

$$\leq \frac{\gamma}{8} \mathbb{E}_{\overline{M}(\tilde{\pi})} \left[\sum_{h=0}^{H} \mathop{\mathbb{E}}_{a \sim \tilde{\pi}(h,s_h)} \left[4 D_{TV_{\{0,1\}}}^2 \left((\overline{M}(\tilde{\pi})|h, s_h, a)_{\downarrow h}, M(h, s_h, a) \right) \right] \right]$$

The total variation distance between the converted distributions is less than $\sqrt{2}$ times the Hellinger distance between them. Then pull the 8 out and cancel.

$$\leq \gamma \mathbb{E}_{\overline{M}(\tilde{\pi})} \left[\sum_{h=0}^{H} \mathop{\mathbb{E}}_{a \sim \tilde{\pi}(h,s_h)} \left[D_{H_{\{0,1\}}}^2 \left((\overline{M}(\tilde{\pi})|h, s_h, a)_{\downarrow h}, M(h, s_h, a) \right) \right] \right]$$

Chaining these inequalities together, we have derived

$$\mathbb{E}_{M \bowtie \pi_{\mathbb{M}}} \left[\sum_{h=0}^{t^*-1} r_h + \mathbf{1}_{t^*=h} \right] - \overline{V}_{0,s_0}^{\overline{M}(\tilde{\pi})}$$

$$\leq \gamma \cdot \mathbb{E}_{\overline{M}(\tilde{\pi})} \left[\sum_{h=0}^{H} \mathop{\mathbb{E}}_{a \sim \tilde{\pi}(h,s_h)} \left[D_{H_{\{0,1\}}}^2 \left((\overline{M}(\tilde{\pi})|h, s_h, a)_{\downarrow h}, M(h, s_h, a) \right) \right] \right]$$

Applying this to 4, the Hellinger error terms cancel, and our overall upper bound on the offset DEC is now

$$\leq \overline{V}_{0,s_0}^{\overline{M}(\tilde{\pi})} - \mathbb{E}_{\overline{M}(\tilde{\pi})}\left[\sum_{h=0}^{H} r_h + b_{h,s_h}^{\overline{M}(\tilde{\pi})}\right] + \mathbb{E}_{\overline{M}(\tilde{\pi})}\left[\sum_{h=0}^{H} b_{h,s_h}^{\overline{M}(\tilde{\pi})}\right] \qquad (6)$$

Now, we'll prove that for all $h \in [H]_0$, we have

$$\mathbb{E}_{\overline{M}(\tilde{\pi})}\left[\sum_{k=0}^{h-1}\left(r_k + b_{k,s_k}^{\overline{M}(\tilde{\pi})}\right) + \overline{V}_{h,s_h}^{\overline{M}(\tilde{\pi})}\right] \leq \mathbb{E}_{\overline{M}(\tilde{\pi})}\left[\sum_{k=0}^{h}\left(r_k + b_{k,s_k}^{\overline{M}(\tilde{\pi})}\right) + \overline{V}_{h+1,s_{h+1}}^{\overline{M}(\tilde{\pi})}\right] \qquad (7)$$

Subtracting from both sides, it suffices to prove

$$\mathbb{E}_{\overline{M}(\tilde{\pi})}\left[\overline{V}_{h,s_h}^{\overline{M}(\tilde{\pi})}\right] \leq \mathbb{E}_{\overline{M}(\tilde{\pi})}\left[r_h + b_{h,s_h}^{\overline{M}(\tilde{\pi})} + \overline{V}_{h+1,s_{h+1}}^{\overline{M}(\tilde{\pi})}\right]$$

This is proven as follows. By the definition of $\overline{V}_{h,s_h}^{\overline{M}(\tilde{\pi})}$ and $\overline{Q}_{h,s_h,a}^{\overline{M}(\tilde{\pi})}$, we have

$$\mathbb{E}_{\overline{M}(\tilde{\pi})}\left[\overline{V}_{h,s_h}^{\overline{M}(\tilde{\pi})}\right] \leq \mathbb{E}_{\overline{M}(\tilde{\pi})}\left[b_{h,s_h}^{\overline{M}(\tilde{\pi})} + \mathbb{E}_{a\sim\tilde{\pi}(h,s_h)}\left[\mathbb{E}_{r'_h,s'_{h+1}\sim\overline{M}(\tilde{\pi})|h,s_h,a}\left[r'_h + \overline{V}_{h+1,s'_{h+1}}^{\overline{M}(\tilde{\pi})}\right]\right]\right]$$

$$= \mathbb{E}_{\overline{M}(\tilde{\pi})}\left[b_{h,s_h}^{\overline{M}(\tilde{\pi})} + \mathbb{E}_{r'_h,s'_{h+1}\sim\overline{M}(\tilde{\pi})|h,s_h}\left[r'_h + \overline{V}_{h+1,s'_{h+1}}^{\overline{M}(\tilde{\pi})}\right]\right] = \mathbb{E}_{\overline{M}(\tilde{\pi})}\left[b_{h,s_h}^{\overline{M}(\tilde{\pi})} + r_h + \overline{V}_{h+1,s_{h+1}}^{\overline{M}(\tilde{\pi})}\right]$$

The result in 7 holding for all h. We chain these inequalities together, and use $\overline{V}_{H+1,s_{H+1}}^{\overline{M}(\tilde{\pi})} = 0$, to yield $\overline{V}_{0,s_0}^{\overline{M}(\tilde{\pi})} \leq \mathbb{E}_{\overline{M}(\tilde{\pi})}\left[\sum_{h=0}^{H} r_h + b_{h,s_h}^{\overline{M}(\tilde{\pi})}\right]$. Plugging this into 6, canceling, and pulling out the sum, we get

$$\leq \sum_{h=0}^{H} \mathbb{E}_{\overline{M}(\tilde{\pi})}\left[b_{h,s_h}^{\overline{M}(\tilde{\pi})}\right] = \sum_{h=0}^{H}\sum_{s\in\mathcal{S}} \mathbb{P}_{\overline{M}(\tilde{\pi})}(h,s) \cdot b_{h,s}^{\overline{M}(\tilde{\pi})}$$

The bonus value was defined as the value of the offset DEC, with square loss, for a specific instance of the multi-armed bandit problem. The proof of Proposition 5.5 in Foster et al. (2021) may be used to upper-bound this quantity, where the γ in Foster's proof of Proposition 5.5 is chosen to be twice our value of γ. The net result is an upper bound on the offset DEC of $\frac{A}{4\gamma}$. However, the chosen value of γ fed into the odec function varies depending on the choice of h, s, and there is also one instance of the one-armed bandit problem at the start. Plugging in our values, we get

$$\leq \mathbb{P}_{\overline{M}(\tilde{\pi})}(0,s_0)\frac{1}{4\frac{\gamma}{8}\mathbb{P}_{\overline{M}(\tilde{\pi})}(0,s_0)} + \sum_{h=1}^{H}\sum_{s\in\mathcal{S}} \mathbb{P}_{\overline{M}(\tilde{\pi})}(h,s) \cdot \frac{A}{4\frac{\gamma}{8}\mathbb{P}_{\overline{M}(\tilde{\pi})}(h,s)} = \frac{2(HSA+1)}{\gamma}$$

And so, we have upper-bounded the modified offset DEC by $\frac{2(HSA+1)}{\gamma}$ and the lemma follows. ∎

Theorem 4 *In the episodic RMDP setting, if $\overline{M} : \Pi_{RNS} \to \Delta([0,1] \times (\mathcal{S} \times \mathcal{A} \times [0,1])^{[H]})$ is continuous and policy-coherent, then $dec_{\varepsilon}^{f'}(\mathcal{H}, \overline{M}) \leq 2\sqrt{2(HSA+1)\varepsilon}$*

The proof of Proposition 1 does not specifically rely on Hellinger-squared loss, and we may avoid the use of Sion's Minimax Theorem in the proof at the cost of introducing a $\leq$ instead of a strict equality, so we have

$$\text{dec}_{\varepsilon}^{f'}(\mathcal{H}, \overline{M}) \leq \min_{\gamma \geq 0} \left(\max \left(0, \text{dec}_{\gamma}^{o'}(\mathcal{H}, \overline{M}) \right) + \gamma \varepsilon^2 \right)$$

From Lemma 15 we have $\text{dec}_{\gamma}^{o'}(\mathcal{H}, \overline{M}) \leq \frac{2(HSA+1)}{\gamma}$. Plugging this in and computing the minimizing value of γ to be $\frac{\sqrt{2(HSA+1)}}{\varepsilon}$, and plugging this in, we get

$$\text{dec}_{\varepsilon}^{f'}(\mathcal{H}, \overline{M}) \leq 2\sqrt{2(HSA+1)}\varepsilon$$

And the result follows. ∎

Appendix K. Auxiliary Definitions for RMDP Estimator:

The estimator which achieves low estimation complexity for RMDP's is not the RUE algorithm, although it is conceptually similar. We define many functions and sets for the following lemmas, and the definition of our estimator of choice.

With the state space, action space, and time horizon H fixed, the space $\mathcal{RMDP}$ of compatible RMDP's (not necessarily 1-bounded), with our modifications (that the starting state and action are unique, and the ending state is unique), may be described by the transition kernel alone, so we define the type of RMDP's via

$$\mathcal{RMDP} := \square([0,1] \times \mathcal{S}) \times ([H-1] \times \mathcal{S} \times \mathcal{A} \to \square([0,1] \times \mathcal{S})) \times (\mathcal{S} \times \mathcal{A} \to \square[0,1])$$

The notation $[H-1]$ denotes the set of integers from 1 to $H-1$. The type $\mathcal{PRMDPR}$ of Partial RMDP's with Recommendations, is then defined as

$$\mathcal{PRMDPR} := \square([0,1] \times \mathcal{S}) \times ([H-1] \times \mathcal{S} \to \mathcal{A} \times \square([0,1] \times \mathcal{S})) \times (\mathcal{S} \to \mathcal{A} \times \square[0,1])$$

Intuitively, for each h, s, these recommend an action and predict the consequences of that action.

Given spaces X, Y, Z, there is a function shift of type $(X \to Y \times \square Z) \to (X \times Y \to \square Z)$, defined as

$$\text{shift}(f)(x,y) := \text{ if } f(x)_Y = y, f(x)_{\square Z} \text{ else } \Delta Z$$

This corresponds to total uncertainty about the consequences of non-recommended y, while making a prediction about the consequences of the recommended y.

PRMDPR's can be converted to RMDP's via (id, shift, shift), which adds complete uncertainty about non-recommended actions, and RMDP's can be converted to models of type $\Pi_{RNS} \to \square \left([0,1] \times (\mathcal{S} \times \mathcal{A} \times [0,1])^{[H]}\right)$ via $\mathbb{M} \mapsto \{\sigma \bowtie \pi \mid \sigma \models \mathbb{M}\}$. This produces an imprecise belief about trajectories by letting an RMDP interact with a policy.

Given an $f : \mathcal{S} \to [0,1]$ and a $c : [0,1]$, the imprecise belief $\Psi_{f,c}^{\{0,1\}} : \square(\{0,1\} \times \mathcal{S})$ is defined as

$$\Psi_{f,c}^{\{0,1\}} := \left\{ \mu \in \Delta(\{0,1\} \times \mathcal{S}) \ \middle| \ \underset{r,s \sim \mu}{\mathbb{E}}[f(s) + r] \geq c \right\}$$

The imprecise beliefs $\Psi_c^{\{0,1\}} : \square\{0,1\}$, $\Psi_{f,c}^{[0,1]} : \square([0,1] \times \mathcal{S})$, and $\Psi_c^{[0,1]} : \square[0,1]$ are defined similarly by adjusting the space of outcomes or probability distributions accordingly.

$\mathcal{H}_{\mathrm{mid}}^{\{0,1\}} \subseteq \square(\{0,1\} \times \mathcal{S})$, the set of "halfspace hypotheses" for the middle of an RMDP, is

$$\mathcal{H}_{\mathrm{mid}}^{\{0,1\}} := \left\{ \Psi_{f,c}^{\{0,1\}} \ \middle| \ f : \mathcal{S} \to [0,1], c : [0,1] \right\}$$

The spaces $\mathcal{H}_{\mathrm{end}}^{\{0,1\}} \subseteq \square\{0,1\}$, $\mathcal{H}_{\mathrm{mid}}^{[0,1]}$, and $\mathcal{H}_{\mathrm{end}}^{[0,1]}$, are defined similarly, using their corresponding Ψ sets.

Finally, the subset $\mathcal{H}_{\mathrm{parhalf}}$ of PRMDPR's, is defined as

$$\mathcal{H}_{\mathrm{parhalf}} := \mathcal{H}_{\mathrm{mid}}^{[0,1]} \times \left([H-1] \times \mathcal{S} \to \mathcal{A} \times \mathcal{H}_{\mathrm{mid}}^{[0,1]} \right) \times \left(\mathcal{S} \to \mathcal{A} \times \mathcal{H}_{\mathrm{end}}^{[0,1]} \right)$$

Because $\mathcal{H}_{\mathrm{parhalf}}$ is a set of PRMDPR's, it can be converted to a set of RMDP's, and a set of models. By abuse of notation, these induced sets are also be denoted $\mathcal{H}_{\mathrm{parhalf}}$. We now present some lemmas about these definitions.

Lemma 16 *Given a distribution ν supported on $\{0,1\} \times \mathcal{S}$, for any $f : \mathcal{S} \to [0,1]$, and $c : [0,1]$, $D_{\mathcal{H}_{\{0,1\}}}^2 \left(\nu \to \Psi_{f,c}^{[0,1]} \right) = D_{\mathcal{H}}^2 \left(\nu \to \Psi_{f,c}^{\{0,1\}} \right)$*

We use r' for rewards that are 0 or 1, and r for rewards in $[0,1]$. $\underset{r' \sim r}{\mathbb{E}}$ means expected value, when r is rounded up to 1 with r probability, and down to 0 with $1 - r$ probability.

To begin we have convert $\left(\Psi_{f,c}^{[0,1]} \right) = \Psi_{f,c}^{\{0,1\}}$, by the following argument. If $\mu \in$ convert $\left(\Psi_{f,c}^{[0,1]} \right)$, there exists a $\nu \in \Psi_{f,c}^{[0,1]}$ s.t. $\mu =$ convert(ν). We then have

$$\underset{r',s \sim \mu}{\mathbb{E}}[f(s) + r] = \underset{r',s \sim \mathrm{convert}(\nu)}{\mathbb{E}}[f(s) + r] = \underset{r,s \sim \nu}{\mathbb{E}}\left[\underset{r' \sim r}{\mathbb{E}}[f(s) + r'] \right] = \underset{r,s \sim \nu}{\mathbb{E}}[f(s) + r] \geq c$$

This was by the definition of conversion (a distribution ν can be converted by rounding the rewards up or down appropriately), pulling the constant function out of the inner expectation, using that $\underset{r' \sim r}{\mathbb{E}}[r'] = r$, and using that $\nu \in \Psi_{f,c}^{[0,1]}$. The fact that $\underset{r',s \sim \mu}{\mathbb{E}}[f(s) + r] \geq c$ then certifies that $\mu \in \Psi_{f,c}^{\{0,1\}}$. This establishes convert $\left(\Psi_{f,c}^{[0,1]} \right) \subseteq \Psi_{f,c}^{\{0,1\}}$.

In the other direction, for any $\mu \in \Psi_{f,c}^{\{0,1\}}$, the type signature of μ can be straightforwardly changed to be $\Delta([0,1] \times \mathcal{S})$. The expectation of f plus reward exceeded c because $\mu \in \Psi_{f,c}^{\{0,1\}}$, and altering its type signature doesn't change its expectation, so $\mu \in \Psi_{f,c}^{[0,1]}$. Converting μ back then involves rounding 1's up to 1 with 1 probability, and rounding 0's down to 0 with 1 probability, which has no effect, so $\mu \in$ convert$(\Psi_{f,c}^{[0,1]})$ as well, establishing convert$(\Psi_{f,c}^{[0,1]}) \supseteq \Psi_{f,c}^{\{0,1\}}$. Both subset inclusion directions have been shown, so the sets are equal.

Now, by applying the definition of the converted Hellinger distance, and our result proved above, we compute

$$D^2_{H_{\{0,1\}}}\left(\nu \to \Psi^{[0,1]}_{f,c}\right) = D^2_H\left(\text{convert}(\nu) \to \text{convert}(\Psi^{[0,1]}_{f,c})\right) = D^2_H\left(\nu \to \Psi^{\{0,1\}}_{f,c}\right)$$

We could remove the conversion from ν because it was already supported on $\{0,1\}$ by assumption. And the result follows. ∎

Lemma 17 *Given some $\varepsilon_1, \varepsilon_2$, $f : \mathcal{S} \to [0,1]$, and $c : [0,1]$, use $\lceil f \rceil$ to denote the function $\lambda s.1 - \left\lfloor \frac{1-f(s)}{\varepsilon_1} \right\rfloor \cdot \varepsilon_1$, and $\lfloor c \rfloor$ to denote $\left\lfloor \frac{c}{\varepsilon_2} \right\rfloor \cdot \varepsilon_2$. We have that $D^2_H\left(\Psi^{\{0,1\}}_{\lceil f \rceil, \lfloor c \rfloor} \to \Psi^{\{0,1\}}_{f,c}\right) \leq \varepsilon_1 + \varepsilon_2$.*

By definition,

$$D^2_H\left(\Psi^{\{0,1\}}_{\lceil f \rceil, \lfloor c \rfloor} \to \Psi^{\{0,1\}}_{f,c}\right) = \max_{\mu \in \Psi^{\{0,1\}}_{\lceil f \rceil, \lfloor c \rfloor}} \min_{\mu' \in \Psi^{\{0,1\}}_{f,c}} D^2_H(\mu, \mu')$$

Fixing the maximizing μ, we will construct a μ' which witnesses that the Hellinger distance is low. μ' is made by taking μ, and moving $\varepsilon_1 + \varepsilon_2$ probability mass from points $s, 0$ to their corresponding point $s, 1$. If there is less than $\varepsilon_1 + \varepsilon_2$ probability mass on points $s, 0$, move all of it up to their corresponding points $s, 1$.

By Hellinger distance squared being less than the total variation distance, and us moving $\varepsilon_1 + \varepsilon_2$ measure to construct μ' from μ, we may compute

$$D^2_H(\mu, \mu') \leq D_{TV}(\mu, \mu') \leq \varepsilon_1 + \varepsilon_2$$

yielding our result. However, we still have to show that $\mu' \in \Psi^{\{0,1\}}_{f,c}$. To show this, we must show that $\underset{r,s \sim \mu'}{\mathbb{E}}[f(s) + r] \geq c$. If less than $\varepsilon_1 + \varepsilon_2$ measure was moved, this implies that μ' has all of its measure on 1 reward, which trivially implies our desired inequality. Therefore, assume that $\varepsilon_1 + \varepsilon_2$ measure was moved to construct μ'. We may then compute

$$\mathbb{E}_{\mu'}[f + r] = \mathbb{E}_{\mu'}[f] + \mathbb{E}_{\mu'}[r] = \mathbb{E}_\mu[f] + (\mathbb{E}_\mu[r] + \varepsilon_1 + \varepsilon_2) = \mathbb{E}_\mu[\lceil f \rceil + r] + \mathbb{E}_\mu[f - \lceil f \rceil] + \varepsilon_1 + \varepsilon_2$$

$$\geq \lfloor c \rfloor + \mathbb{E}_\mu[f - \lceil f \rceil] + \varepsilon_1 + \varepsilon_2 \geq (c - \varepsilon_2) - \varepsilon_1 + \varepsilon_1 + \varepsilon_2 = c$$

In order, this is because μ' was made from μ by moving $\varepsilon_1 + \varepsilon_2$ measure from 0 to 1 reward while not affecting the distribution over states. Then, we used that $\mu \in \Psi_{\lceil f \rceil, \lfloor c \rfloor}$. Finally, we used that the rounding procedure only increases the value of f by ε_1 at most, and decreases c by ε_2 at most. Because $\mu' \in \Psi^{\{0,1\}}_{f,c}$, and μ' was constructed from the distance-maximizing μ, our result follows. ∎

Appendix L. Definition of the RMDP Estimator

T (the number of episodes), and three nonzero parameters $\varepsilon_{[0,1]}, \varepsilon_\mathcal{S}, \varepsilon'$, are taken as input. $\varepsilon_{\text{pess}}$ is taken to be $\sqrt{\frac{1}{T(H+1)^2}}$. We now define our estimator, and the associated functions, as follows.

Given an MDP M and an $h \geq 0$, $M^{>h}$ denotes the partial transition kernel consisting of $M(h', s, a)$ for all s, a and $h' > h$. Given a $\mu : \Delta(\{0, 1\} \times \mathcal{S})$, $(\mu, M^{>h})$ may be considered as an MDP which starts with the reward on timestep h, and policies can interact with this MDP. Accordingly, given an MDP M and timestep h, we define
$\widetilde{M}^{>h} : \Delta(\{0, 1\} \times \mathcal{S}) \times \Pi_{RNS} \rightarrow \Delta\left(\{0, 1\} \times (\mathcal{S} \times \mathcal{A} \times \{0, 1\})^{\{h+1, \ldots, H\}}\right)$ as

$$\widetilde{M}^{>h}(\mu, \pi) := (\mu, M^{>h}) \bowtie \pi$$

And we define $\widetilde{M}^{\geq h} : \mathcal{S} \times \mathcal{A} \times \Pi_{RNS} \rightarrow \Delta\left((\mathcal{S} \times \mathcal{A} \times \{0, 1\})^{\{h, \ldots, H\}}\right)$ as

$$\widetilde{M}^{\geq h}(s, a, \pi) := \delta_{s,a} \times ((M(h, s, a), M^{>h}) \bowtie \pi)$$

Where $\delta_{s,a}$ is the distribution which puts all measure on s, a. $\widetilde{M}^{>h}(\mu, \pi)$ takes the policy π and MDP M and counterfacts on r_h, s_{h+1} being distributed according to μ to derive a distribution over trajectories. $\widetilde{M}^{\geq h}(s, a, \pi)$ does the same, but counterfacts on a specific s, a appearing on timestep h instead.

For the case where $h = H$, $\widetilde{M}^{>H} : \Delta\{0, 1\} \times \Pi_{RNS} \rightarrow \Delta\{0, 1\}$, and $\widetilde{M}^{>H}(\mu, \pi) = \mu$. For the case where $h = 0$, $\widetilde{M}^{\geq 0}$ is only ever invoked on the unique initial state and action at timestep 0, and $M(0, s_0, a_0)$ is the initial distribution of the MDP M.

The finite set $\mathcal{H}'_{\text{mid}} \subseteq \square(\{0, 1\} \times \mathcal{S})$, is defined by

$$\left\{ \Psi^{\{0,1\}}_{f,c} \;\middle|\; \forall s \left(\exists i \in \left\{0, 1 \ldots \left\lfloor \frac{1}{\varepsilon_{\mathcal{S}}} \right\rfloor \right\} : f(s) = 1 - i \cdot \varepsilon_{\mathcal{S}} \wedge \exists j \in \left\{0, 1 \ldots \left\lfloor \frac{1}{\varepsilon_{[0,1]}} \right\rfloor \right\} : c = j \cdot \varepsilon_{[0,1]} \right) \right\}$$

The finite set $\mathcal{H}'_{\text{end}} \subseteq \square\{0, 1\}$, is defined by

$$\mathcal{H}'_{\text{end}} := \left\{ \Psi^{\{0,1\}}_c \;\middle|\; \exists j \in \left\{0, 1 \ldots \left\lfloor \frac{1}{\varepsilon_{[0,1]}} \right\rfloor \right\} : c = j \cdot \varepsilon_{[0,1]} \right\}$$

The finite sets $\mathcal{B}_{\text{frag}}, \mathcal{B}_{\text{cal}}, \mathcal{B}_{\text{unif}}$, of fragment, calibration, and uniform bettors, are indexed as follows. $+$ is taken to be disjoint union, in the type theory sense. $[H - 1]$ denotes the set of integers from 1 to $H - 1$. $\mathbf{1}$ denotes the unit type.

$$\mathcal{B}_{\text{frag}} := \mathcal{H}'_{\text{mid}} + ([H - 1] \times \mathcal{S} \times \mathcal{A} \times \mathcal{H}'_{\text{mid}}) + (\mathcal{S} \times \mathcal{A} \times \mathcal{H}'_{\text{end}})$$

$$\mathcal{B}_{\text{cal}} := ([H - 1] \times \mathcal{S} \times \mathcal{A} \times \mathcal{H}'_{\text{mid}}) + (\mathcal{S} \times \mathcal{A} \times \mathcal{H}'_{\text{end}})$$

$$\mathcal{B}_{\text{unif}} := \mathbf{1} + ([H - 1] \times \mathcal{S} \times \mathcal{A}) + (\mathcal{S} \times \mathcal{A})$$

The set $\mathcal{B}$ of all bettors consists of the fragment and calibration and uniform bettors, and one pessimism bettor.

$$\mathcal{B} := \mathcal{B}_{\text{frag}} + \mathcal{B}_{\text{cal}} + \mathcal{B}_{\text{unif}} + \mathbf{1}$$

Given a B which is a fragment, calibration, or uniform bettor, h_B, s_B, a_B, Ψ_B are their corresponding timestep, state, action, and hypothesis. For fragment bettors and the uniform bettor in the first component of their disjoint union, $h_B = 0$, and s_B, a_B are the unique initial state and initial action at time 0. For fragment, calibration, and uniform bettors in the last component of their disjoint union, $h_B = H$. For uniform bettors, their associated

hypothesis Ψ_B is the set containing only the uniform distribution on $\mathcal{S} \times \{0, 1\}$, or on $\{0, 1\}$, respectively.

The notation $(h, s, a)_B$ is shorthand for the tuple h_B, s_B, a_B, and similar for $(h, s)_B$. The probability of the event $(h, s)_B$ is shorthand for the probability of $s_{h_B} = s_B$, that the state on the h_B'th timestep matches s_B.

Given a calibration bettor B, policy $\pi : \Pi_{RNS}$, and an MDP M, define

$$X_{M,\pi}^B := \pi((h, s)_B)(a_B) \cdot D_H^2(M((h, s, a)_B) \to \Psi_B)$$

Given either a fragment or uniform bettor B, and an MDP M, define

$$\mu_M^B := \underset{\mu \in \Psi_B}{\operatorname{argmin}} \, D_H^2(M((h, s, a)_B), \mu)$$

If $h_B = 0$, then s_B, a_B will be the unique initial state and action, and $M((h, s, a)_B)$ is the initial distribution on rewards and states of the MDP.

Now, given a bettor B, an MDP M, a policy π, and a situation h, s, a with $h \geq 0$, define the g functions $g_{M,\pi}^{B,h,s,a} : \Delta(\{0, 1\} \times \mathcal{S}) \to \mathbb{R}$ and local betting functions $\operatorname{lbet}_{M,\pi}^{B,h,s,a} : [0, 1] \times \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ as follows, by case analysis on bettors.

If B is a fragment or uniform bettor and $h, s, a \neq (h, s, a)_B$, $g_{M,\pi}^{B,h,s,a}$ is the constant-zero function and $\operatorname{lbet}_{M,\pi}^{B,h,s,a}$ is the constant-one function. If $h, s, a = (h, s, a)_B$, we then have

$$g_{M,\pi}^{B,h,s,a}(\mu) := 2D_H^2(\mu \to \Psi_B)$$

$$\operatorname{lbet}_{M,\pi}^{B,h,s,a}(r, s', a') := \underset{r' \sim r}{\mathbb{E}} \left[\sqrt{\frac{\mu_M^B(r', s')}{M(h, s, a)(r', s')}} \right] + D_H^2(M(h, s, a) \to \Psi_B)$$

The notation $\underset{r' \sim r}{\mathbb{E}}$ denotes sampling a 1 with r probability, and a 0 otherwise, which is needed to convert between the continuous reward $r : [0, 1]$, and the distributions being over the reward space $\{0, 1\}$. As usual, if $h = 0$, s_B, a_B will be the unique initial state and action, and $M(h, s, a)$ is the initial distribution of the MDP M. If $h = H$, only the probability of reward is assessed and the equation doesn't depend on s'.

If B is a calibration bettor, and $h \geq h_B$, $g_{M,\pi}^{B,h,s,a}$ is the constant-zero function and $\operatorname{lbet}_{B,\pi}^{B,h,s,a}$ is the constant-one function. If $h < h_B$, we then have

$$g_{M,\pi}^{B,h,s,a}(\mu) := \frac{X_{M,\pi}^B}{4} \mathbb{P}_{\widetilde{M}^{>h}(\mu,\pi)}((h, s)_B)$$

$$\operatorname{lbet}_{M,\pi}^{B,h,s,a}(r, s', a') := 1 + \frac{X_{M,\pi}^B}{4} \left(\mathbb{P}_{\widetilde{M}^{\geq h}(s,a,\pi)}((h, s)_B) - \mathbb{P}_{\widetilde{M}^{\geq h+1}(s',a',\pi)}((h, s)_B) \right)$$

Rounding out our case analysis, if B is the unique pessimism bettor, we have

$$g_{M,\pi}^{B,h,s,a}(\mu) := \varepsilon_{\text{pess}} \cdot \underset{\widetilde{M}^{>h}(\mu,\pi)}{\mathbb{E}} \left[\sum_{k=h}^{H} r_k \right]$$

$$\text{lbet}_{M,\pi}^{B,h,s,a}(r,s',a') := 1 + \varepsilon_{\text{pess}} \cdot \left(\mathbb{E}_{\widetilde{M}^{\geq h}(s,a,\pi)} \left[\sum_{k=h}^{H} r_k \right] - \left(r - \mathbb{E}_{\widetilde{M}^{\geq h+1}(s',a',\pi)} \left[\sum_{k=h+1}^{H} r_k \right] \right) \right)$$

This defines the g functions and local betting functions lbet for all bettors B and situations h, s, a, relative to an MDP M and policy π.

Now we can finally present the core of the RMDP estimator. Given a $\zeta_t : \Delta\mathcal{B}$, and policy π, we define the MDP $M_{t,\pi}$ by downwards induction as

$$M_{t,\pi}(h,s,a) := \underset{\mu \in \Delta(\{0,1\} \times \mathcal{S})}{\text{argmin}} \ \underset{B \sim \zeta_t}{\mathbb{E}} \left[g_{M_{t,\pi},\pi}^{B,h,s,a}(\mu) \right]$$

Given a $\zeta_t : \Delta\mathcal{B}$, the final estimated model, $\widehat{M}_t : \Pi_{RNS} \to \Delta\left(\{0,1\} \times (\mathcal{S} \times \mathcal{A} \times \{0,1\})^{[H]}\right)$, is then given by $\lambda\pi.M_{t,\pi} \bowtie \pi$.

All entries of the $M_{t,\pi}$ transition kernel involve solving a convex optimization problem, and so are feasible to compute. Also, the definition of $M_{t,\pi}$ is well-founded. The base case holds because the g functions used at level H reference $\widetilde{M}_{t,\pi}^{>H}(\mu,\pi)$, but this function is just the identity on μ. The downwards induction step holds because all of the g functions used in computing the transition kernel at level h only depend on the transition kernel for $M_{t,\pi}$ at levels $h' > h$.

All that remains to fully define the estimator is to provide a prior distribution ζ_1, and show how to update the distribution ζ_t on incoming data. For fragment or calibration bettors, their prior probability is $\frac{1}{2(|\mathcal{B}_{\text{frag}}|+|\mathcal{B}_{\text{cal}}|)}$. For the pessimism bettor, its prior probability is $\frac{1}{2} - \varepsilon'$. For the uniform bettors, their prior probabilities are $\frac{\varepsilon'}{|\mathcal{B}_{\text{unif}}|}$. Updating is given as follows. Given a bettor B, MDP M and policy π, we define $\text{bet}_{M,\pi}^{B} : \left([0,1] \times (\mathcal{S} \times \mathcal{A} \times [0,1])^{[H]}\right) \to \mathbb{R}$, the aggregate betting function, as

$$\text{bet}_{M,\pi}^{B}(r_0,s_1,a_1,...r_H) := 1 + \sum_{h=0}^{H} \left(\text{lbet}_{M,\pi}^{B,h,s_h,a_h}(r_h,s_{h+1},a_{h+1}) - 1 \right)$$

For $h = 0$, as usual, s_0, a_0 are the unique initial state and action. For $h = H$, none of the betting functions depend on the final state and action, so it doesn't matter that s_{H+1}, a_{H+1} are ill-defined. We may then succinctly define updating as

$$\bigstar_t := \underset{B \sim \zeta_t}{\mathbb{E}} \left[\text{bet}_{M_{t,\pi_t},\pi_t}^{B}(tr_t) \right]$$

$$\zeta_{t+1}(B) := \frac{\zeta_t(B) \cdot \text{bet}_{M_{t,\pi_t},\pi_t}^{B}}{\bigstar_t}$$

Where tr_t and π_t are the trajectory in episode t and policy in episode t, respectively. This concludes our specification of the estimator for RMDP's. We now turn to analyzing the estimation complexity. When $M_{t,\pi}, \pi$ shows up in a subscript, we use a subscript of t, π instead, to abbreviate it.

Appendix M. RMDP Estimator Lemmas

Lemma 18 *The estimator $\widehat{M}$ is well-defined, and all $M_{t,\pi}(h,s,a)$ have full support.*

The only potential issues in the definition of the estimator are showing that all $M_{t,\pi}(h,s,a)$ have full support so no division-by-zero errors arise when computing the bets, and that all ζ_t are indeed probability distributions. To show that all $M_{t,\pi}(h,s,a)$ have full support, we may reuse the proof from Proposition 4, that the uniform bettors ensure that $M_{t,\pi}(h,s,a)$ has full support if they have nonzero probability, and these bettors never run out of probability mass completely. Minor adaptations must be made to deal with the presence of the calibration bettors, but their tedium greatly outweighs their impact on the overall proof, and the result holds anyways by the same line of argument. Showing that all ζ_t are probability distributions is slightly more difficult. To show this, we show by induction, that *if* all bets are ≥ 0 and there is a bettor B^* which never has a bet value of zero, all ζ_t are probability distributions. Then we show that these two properties indeed hold.

For the induction base case, ζ_1 is a probability distribution, and $\zeta_1(B^*) > 0$ holds. For the induction step, we assume that these two properties hold of ζ_t and prove them for ζ_{t+1} if all bets are ≥ 0 and the bet of B^* is > 0, as follows.

First, because $\zeta_t(B^*) > 0$ by the induction assumption, and $\mathrm{bet}_{t,\pi_t}^{B^*}(tr_t) > 0$ by assumption, and ζ_t is a probability distribution by the induction assumption, and all bets are ≥ 0 by assumption, $\bigstar_t > 0$. This ensures that no division by zero issues occur and ζ_{t+1} is well-defined. $\zeta_t(B^*) > 0$ and $\mathrm{bet}_{t,\pi_t}^{B^*}(tr_t) > 0$ also implies that $\zeta_{t+1}(B^*) > 0$, proving half of our induction step. To show that ζ_{t+1} is a probability distribution, we use that ζ_t is a probability distribution by induction assumption and all B fulfill $\mathrm{bet}_{t,\pi_t}^{B}(tr_t) \geq 0$ by assumption, which implies that all B fulfill $\zeta_{t+1}(B) \geq 0$. To show that the sum of measure is 1, we compute

$$\sum_{B \in \mathcal{B}} \frac{\zeta_t(B) \cdot \mathrm{bet}_{t,\pi_t}^{B}(tr_t)}{\bigstar_t} = \sum_{B \in \mathcal{B}} \frac{\zeta_t(B) \cdot \mathrm{bet}_{t,\pi_t}^{B}(tr_t)}{\sum_{B' \in \mathcal{B}} \zeta_t(B') \cdot \mathrm{bet}_{t,\pi_t}^{B'}(tr_t)} = \frac{\sum_{B \in \mathcal{B}} \zeta_t(B) \cdot \mathrm{bet}_{t,\pi_t}^{B}(tr_t)}{\sum_{B' \in \mathcal{B}} \zeta_t(B') \cdot \mathrm{bet}_{t,\pi_t}^{B'}(tr_t)} = 1$$

Therefore, the second half of the induction step goes through, proving our result, if all bets are non-negative and there's a bettor which is guaranteed to always make positive bets.

We now shift to proving those two properties. We use the definition of the global betting function bet and the local betting functions lbet to compute the overall bets. If $B \in \mathcal{B}_{\mathrm{frag}} \cup \mathcal{B}_{\mathrm{unif}}$, and $s_{h_B}, a_{h_B} \neq s_B, a_B$, all local bets are 1, and cancel out with the -1's in the sum of the global bet, leaving the global bet as 1, which is nonnegative. However, if $s_{h_B}, a_{h_B} = s_B, a_B$, the 1 in the global bet cancels out with the -1 at the nontrivial bet, yielding

$$\mathrm{bet}_{t,\pi}^{B}(r_0, (s,a,r)_{1:H}) = \mathop{\mathbb{E}}_{r' \sim r_{h_B}} \left[\sqrt{\frac{\mu_{t,\pi}^{B}(r', s_{h_B+1})}{M_{t,\pi}((h,s,a)_B)(r', s_{h_B+1})}} \right] + D_H^2(M_{t,\pi}((h,s,a)_B) \to \Psi_B)$$

This is clearly nonnegative and well-defined if $M_{t,\pi}((h,s,a)_B)$ has full support. If $B \in \mathcal{B}_{\mathrm{cal}}$, local bets at and after h_B are 1 and cancel out with the -1's in the sum of the global

bet. We can then compute

$$\text{bet}_{t,\pi}^{B}(r_0, (s,a,r)_{1:H}) = 1 + \sum_{h=0}^{h_B-1}\left(\text{lbet}_{t,\pi}^{B,h,s_h,a_h}(r_h, s_{h+1}, a_{h+1}) - 1\right)$$

$$= 1 + \sum_{h=0}^{h_B-1}\left(\left(1 + \frac{X_{t,\pi}^{B}}{4}\left(\mathbb{P}_{\widetilde{M}_{t,\pi}^{\geq h}(s_h,a_h,\pi)}((h,s)_B) - \mathbb{P}_{\widetilde{M}_{t,\pi}^{\geq h+1}(s_{h+1},a_{h+1},\pi)}((h,s)_B)\right)\right) - 1\right)$$

The 1's cancel and the sum is a telescoping sum, yielding

$$= 1 + \frac{X_{t,\pi}^{B}}{4}\left(\mathbb{P}_{\widetilde{M}_{t,\pi}^{\geq 0}(s_0,a_0,\pi)}((h,s)_B) - \mathbb{P}_{\widetilde{M}_{t,\pi}^{\geq h_B}(s_{h_B},a_{h_B},\pi)}((h,s)_B)\right)$$

The MDP's are always guaranteed to start with s_0 and a_0 at time 0, so $\widetilde{M}_{t,\pi}^{\geq 0}(s_0,a_0,\pi)$ is just $M_{t,\pi} \bowtie \pi$, ie, $\widehat{M}_t(\pi)$. Further, the event $(h,s)_B$ (reaching s_B at time h_B) is an abbreviation for $s_{h_B} = s_B$, so this probability is always either 1 or 0 and can be replaced with an indicator function, yielding

$$= 1 + \frac{X_{t,\pi}^{B}}{4}\left(\mathbb{P}_{\widehat{M}_t(\pi)}((h,s)_B) - \mathbf{1}_{s_{h_B}=s_B}\right)$$

Because $X_{t,\pi}^{B} \in [0,1]$, this quantity is always in $[3/4, 5/4]$, and we have found a family of bettors whose bets are not just nonnegative, but also bounded away from zero, providing our desired B^*.

Finally, for the pessimism bettor, we can compute

$$\text{bet}_{t,\pi}^{B}(r_0, (s,a,r)_{1:H}) = 1 + \sum_{h=0}^{H}\left(\text{lbet}_{t,\pi}^{B,h,s_h,a_h}(r_h, s_{h+1}, a_{h+1}) - 1\right)$$

$$= 1 + \sum_{h=0}^{H}\left(\left(1 + \varepsilon_{\text{pess}}\left(\mathbb{E}_{\widetilde{M}_{t,\pi}^{\geq h}(s_h,a_h,\pi)}\left[\sum_{k=h}^{H}r_k\right] - \left(r_h + \mathbb{E}_{\widetilde{M}_{t,\pi}^{\geq h+1}(s_{h+1},a_{h+1},\pi)}\left[\sum_{k=h+1}^{H}r_k\right]\right)\right)\right) - 1\right)$$

The 1's cancel and the sum is a telescoping sum, yielding

$$= 1 + \varepsilon_{\text{pess}}\left(\mathbb{E}_{\widetilde{M}_{t,\pi}^{\geq 0}(s_0,a_0,\pi)}\left[\sum_{k=0}^{H}r_k\right] - \sum_{h=0}^{H}r_h - \mathbb{E}_{\widetilde{M}_{t,\pi}^{\geq H+1}(s_{H+1},a_{H+1},\pi)}\left[\sum_{k=H+1}^{H}r_k\right]\right)$$

The latter sum is an empty sum, so it is zero, and this supersedes that $\widetilde{M}_{t,\pi}^{\geq H+1}(s_{H+1},a_{H+1},\pi)$ is ill-defined. As before, $\widetilde{M}_{t,\pi}^{\geq 0}(s_0,a_0,\pi)$ can be reexpressed as $\widehat{M}_t(\pi)$, so we are left with

$$= 1 + \varepsilon_{\text{pess}}\left(\mathbb{E}_{\widehat{M}_t(\pi)}\left[\sum_{k=0}^{H}r_k\right] - \sum_{h=0}^{H}r_h\right)$$

$\varepsilon_{\text{pess}} = \sqrt{\frac{1}{T(H+1)^2}}$. Therefore, $\varepsilon_{\text{pess}} \leq \frac{1}{H+1}$. Even in the worst possible case where the expected sum of rewards is 0, and the actual sum of rewards is $H+1$, this bet is non-negative. Because all bets are non-negative, and there are bettors whose bets are always bounded away from zero, the inductive proof goes through, and the estimator is well-defined. ∎

Lemma 19 *For every calibration bettor B, with $1 - \frac{\delta}{4(HS+1)}$ probability (over the true algorithm interacting with the true environment), we have*

$$6 \left(\ln\left(\frac{4(HS+1)}{\delta} \right) + \ln(2(|\mathcal{B}_{cal}| + |\mathcal{B}_{frag}|)) + \sum_{t=1}^{T} \ln\left(\bigstar_t\right) \right) + 2 \sum_{t=1}^{T} \mathbb{E}_{\pi, s_{h_B} \sim \xi_t} \left[X_{t,\pi}^{B} \cdot \mathbf{1}_{s_{h_B} = s_B} \right]$$

$$\geq \sum_{t=1}^{T} \mathbb{E}_{\pi \sim p_t} \left[X_{t,\pi}^{B} \cdot \mathbb{P}_{\widehat{M}_t(\pi)}((h,s)_B) \right]$$

Where p_t is the distribution over policies on episode t and ξ_t is the joint distribution over policies and trajectories on episode t according to the true environment.

We use ξ_t as an abbreviation for the distribution over policies π_t and trajectories tr_t, which is sampled from on the t'th episode. It depends on the true algorithm of the agent and the true environment it interacts with.

Let B be a calibration bettor. Via Lemma A.4 of Foster et al. (2021), we have that, with $1 - \frac{\delta}{4(HS+1)}$ probability according to the true algorithm interacting with the true environment, we have

$$\ln\left(\frac{4(HS+1)}{\delta} \right) + \sum_{t=1}^{T} \ln(\text{bet}_{t,\pi_t}^{B}(tr_t)) \geq -\sum_{t=1}^{T} \ln\left(\mathbb{E}_{\pi, tr \sim \xi_t} \left[e^{-\ln(\text{bet}_{t,\pi}^{B}(tr))} \right] \right)$$

Unpacking the overall bet for calibration bettors, which was analyzed in Lemma 18, yields

$$= -\sum_{t=1}^{T} \ln\left(\mathbb{E}_{\xi_t} \left[\frac{1}{\text{bet}_{t,\pi}^{B}(tr)} \right] \right) = -\sum_{t=1}^{T} \ln\left(\mathbb{E}_{\xi_t} \left[\frac{1}{1 + \frac{1}{4} X_{t,\pi}^{B} \left(\mathbb{P}_{\widehat{M}_t(\pi)}((h,s)_B) - \mathbf{1}_{s_{h_B} = s_B} \right)} \right] \right)$$

When $x \in [-1, 1]$, we have that $\frac{1}{1 + \frac{1}{4}x} \leq 1 - \frac{1}{4}x + \frac{1}{12}|x|$. Taking x to be $X_{t,\pi}^{B} \left(\mathbb{P}_{\widehat{M}_t(\pi)}((h,s)_B) - \mathbf{1}_{s_{h_B} = s_B} \right)$, and upper-bounding $|x|$ by $X_{t,\pi}^{B} \left(\mathbb{P}_{\widehat{M}_t(\pi)}((h,s)_B) + \mathbf{1}_{s_{h_B} = s_B} \right)$, yields

$$\geq -\sum_{t=1}^{T} \ln(\mathbb{E}_{\xi_t}[1 - \frac{1}{4} X_{t,\pi}^{B} \left(\mathbb{P}_{\widehat{M}_t(\pi)}((h,s)_B) - \mathbf{1}_{s_{h_B} = s_B} \right)$$

$$+ \frac{1}{12} X_{t,\pi}^{B} (\mathbb{P}_{\widehat{M}_t(\pi)}((h,s)_B) + \mathbf{1}_{s_{h_B} = s_B})])$$

$$= -\sum_{t=1}^{T} \ln\left(1 - \mathbb{E}_{\xi_t} \left[\frac{1}{6} X_{t,\pi}^{B} \cdot \mathbb{P}_{\widehat{M}_t(\pi)}((h,s)_B) - \frac{1}{3} X_{t,\pi}^{B} \cdot \mathbf{1}_{s_{h_B} = s_B} \right] \right)$$

By $-\ln(1 - x) \geq x$ we have

$$\geq \sum_{t=1}^{T} \mathbb{E}_{\xi_t} \left[\frac{1}{6} X_{t,\pi}^{B} \cdot \mathbb{P}_{\widehat{M}_t(\pi)}((h,s)_B) - \frac{1}{3} X_{t,\pi}^{B} \cdot \mathbf{1}_{s_{h_B} = s_B} \right]$$

The expectation was over the policy and trajectory. However, $X_{t,\pi}^B$ and $\mathbb{P}_{\widehat{M}_t(\pi)}((h,s)_B)$ don't depend on the trajectory, so we can rewrite as

$$= \sum_{t \leq T} \frac{1}{6} \mathbb{E}_{\pi \sim p_t} \left[X_{t,\pi}^B \cdot \mathbb{P}_{\widehat{M}_t(\pi)}((h,s)_B) \right] - \frac{1}{3} \mathbb{E}_{\zeta_t} \left[X_{t,\pi}^B \cdot \mathbf{1}_{s_{h_B}=s_B} \right]$$

Putting all the inequalities together, we have, for any calibration bettor B, with $1 - \frac{\delta}{4(HS+1)}$ probability,

$$\ln\left(\frac{4(HS+1)}{\delta}\right) + \sum_{t=1}^{T} \ln\left(\text{bet}_{t,\pi_t}^B(tr_t)\right)$$

$$\geq \sum_{t=1}^{T} \frac{1}{6} \mathbb{E}_{\pi \sim p_t} \left[X_{t,\pi}^B \cdot \mathbb{P}_{\widehat{M}_t(\pi)}((h,s)_B) \right] - \sum_{t \leq T} \frac{1}{3} \mathbb{E}_{\zeta_t} \left[X_{t,\pi}^B \cdot \mathbf{1}_{s_{h_B}=s_B} \right]$$

Multiply both sides by 6 and reshuffle.

$$6\left(\ln\left(\frac{4(HS+1)}{\delta}\right) + \sum_{t=1}^{T} \ln\left(\text{bet}_{t,\pi_t}^B(tr_t)\right) \right) + 2\sum_{t=1}^{T} \mathbb{E}_{\zeta_t} \left[X_{t,\pi}^B \cdot \mathbf{1}_{s_{h_B}=s_B} \right] \tag{8}$$

$$\geq \sum_{t=1}^{T} \mathbb{E}_{\pi \sim p_t} \left[X_{t,\pi}^B \cdot \mathbb{P}_{\widehat{M}_t(\pi)}((h,s)_B) \right]$$

Finally, by reshuffling Lemma 6, we have

$$\ln(\zeta_{T+1}(B)) - \ln(\zeta_1(B)) + \sum_{t=1}^{T} \ln(\bigstar_t) = \sum_{t=1}^{T} \ln\left(\text{bet}_{t,\pi_t}^B(tr_t)\right)$$

$\zeta_{T+1}(B)$ can be at most 1, and $\zeta_1(B)$, for calibration bettors, is known to be $\frac{1}{2(|\mathcal{B}_{\text{cal}}|+|\mathcal{B}_{\text{frag}}|)}$, yielding the inequality

$$\ln(2(|\mathcal{B}_{\text{cal}}| + |\mathcal{B}_{\text{frag}}|)) + \sum_{t=1}^{T} \ln(\bigstar_t) \geq \sum_{t=1}^{T} \ln\left(\text{bet}_{t,\pi_t}^B(tr_t)\right)$$

Which, when substituted into 8, yields the desired result. ■

Lemma 20 *For all fragment bettors B where a true RMDP $\mathbb{M}$ fulfills convert$(\mathbb{M}((h,s,a)_B)) \subseteq \Psi_B$, with $1 - \frac{\delta}{4(HS+1)}$ probability (over the true algorithm interacting with the true environment), we have $\ln\left(\frac{4(HS+1)}{\delta}\right) + \ln(2(|\mathcal{B}_{cal}| + |\mathcal{B}_{frag}|)) + \sum_{t=1}^{T} \ln(\bigstar_t)$*
$\geq \sum_{t=1}^{T} \mathbb{E}_{\pi,s_{h_{B'}} \sim \xi_t} \left[X_{t,\pi}^{B'} \cdot \mathbf{1}_{s_{h_{B'}}=s_{B'}} \right]$ for $h_B \geq 1$ and $\geq \sum_{t=1}^{T} \mathbb{E}_{\pi \sim p_t} \left[D_H^2(M_{t,\pi}((h,s,a)_B) \to \Psi_B) \right]$
for $h_B = 0$, where B' is the calibration bettor corresponding to B, ξ_t is the joint distribution over the policy and trajectory on episode t, and p_t is the distribution over the policy on episode t.

69

The proof of this lemma substantially follows the analysis of the estimation complexity in the proof of Theorem 2, so we will gloss over arguments spelled out in more detail there, and only highlight meaningful differences. The implicit uses of Lemma 4 in the proof of Theorem 2 are permissible because the "every transition kernel has full support" argument from Theorem 2 still holds in our setting by Lemma 18

Fix a fragment bettor B with the property that, for a true RMDP $\mathbb{M}$, $\text{convert}(\mathbb{M}((h, s, a)_B)) \subseteq \Psi_B$. Via Lemma A.4 of Foster et al. (2021), we have that, with $1 - \frac{\delta}{4(HS+1)}$ probability,

$$\ln\left(\frac{4(HS+1)}{\delta}\right) + \sum_{t=1}^{T} \ln(\text{bet}_{t,\pi_t}^{B}(tr_t)) \geq -\sum_{t=1}^{T} \ln\left(\underset{\pi, tr \sim \xi_t}{\mathbb{E}}\left[e^{-\ln(\text{bet}_{t,\pi}^{B}(tr))}\right]\right)$$

ξ_t is, as before, the joint distribution over policies and trajectories on episode t. Unpacking the definition of the bet for fragment bettors, as in Lemma 18, yields

$$= -\sum_{t=1}^{T} \ln\left(\mathbb{E}_{\xi_t}\left[\frac{1}{\text{bet}_{t,\pi}^{B}(tr)}\right]\right) = -\sum_{t=1}^{T} \ln(\mathbb{E}_{\xi_t}$$

$$\left[\frac{1}{\mathbf{1}_{(s,a)_{h_B}=(s,a)_B}\left(\underset{r'\sim r_{h_B}}{\mathbb{E}}\left[\sqrt{\frac{\mu_{t,\pi}^{B}(r',s_{h_B+1})}{M_{t,\pi}((h,s,a)_B)(r',s_{h_B+1})}}\right] + D_H^2(M_{t,\pi}((h,s,a)_B) \to \Psi_B)\right) + \mathbf{1}_{(s,a)_{h_B}\neq(s,a)_B}}\right])$$

$$\geq -\sum_{t=1}^{T} \ln\left(\mathbb{E}_{\xi_t}\left[\frac{1}{\mathbf{1}_{(s,a)_{h_B}=(s,a)_B}\left(\underset{r'\sim r_{h_B}}{\mathbb{E}}\left[\sqrt{\frac{\mu_{t,\pi}^{B}(r',s_{h_B+1})}{M_{t,\pi}((h,s,a)_B)(r',s_{h_B+1})}}\right]\right) + \mathbf{1}_{(s,a)_{h_B}\neq(s,a)_B}}\right]\right)$$

$$= -\sum_{t=1}^{T} \ln\left(\mathbb{E}_{\xi_t}\left[\mathbf{1}_{(s,a)_{h_B}=(s,a)_B} \cdot \frac{1}{\underset{r'\sim r_{h_B}}{\mathbb{E}}\left[\sqrt{\frac{\mu_{t,\pi}^{B}(r',s_{h_B+1})}{M_{t,\pi}((h,s,a)_B)(r',s_{h_B+1})}}\right]} + \mathbf{1}_{(s,a)_{h_B}\neq(s,a)_B}\right]\right)$$

By convexity of $\frac{1}{x}$, we have

$$\geq -\sum_{t=1}^{T} \ln\left(\mathbb{E}_{\xi_t}\left[\mathbf{1}_{(s,a)_{h_B}=(s,a)_B} \cdot \underset{r'\sim r_h}{\mathbb{E}}\left[\frac{1}{\sqrt{\frac{\mu_{t,\pi}^{B}(r',s_{h_B+1})}{M_{t,\pi}((h,s,a)_B)(r',s_{h_B+1})}}}\right] + \mathbf{1}_{(s,a)_{h_B}\neq(s,a)_B}\right]\right)$$

$$= -\sum_{t=1}^{T} \ln\left(\mathbb{E}_{\xi_t}\left[\mathbf{1}_{(s,a)_{h_B}=(s,a)_B} \cdot \underset{r'\sim r_h}{\mathbb{E}}\left[\sqrt{\frac{M_{t,\pi}((h,s,a)_B)(r',s_{h_B+1})}{\mu_{t,\pi}^{B}(r',s_{h_B+1})}}\right] + \mathbf{1}_{(s,a)_{h_B}\neq(s,a)_B}\right]\right)$$

The outer expectation can be thought of as two expectations. One is over the π and partial trajectory tr' extending up to the action on turn h_B. The other is over r_{h_B}, s_{h_B+1}.

We use $\sigma_{t,\pi}^{tr'}$ to denote this latter distribution, if we play π and have partial trajectory tr' on episode t. This inner expectation can be moved past the indicator functions to yield

$$= -\sum_{t=1}^{T} \ln \left(\mathop{\mathbb{E}}_{\pi,tr' \sim \xi_t} \left[\mathbf{1}_{(s,a)_{h_B}=(s,a)_B} \cdot \mathop{\mathbb{E}}_{r,s' \sim \sigma_{t,\pi}^{tr'}} \left[\mathop{\mathbb{E}}_{r' \sim r} \left[\sqrt{\frac{M_{t,\pi}((h,s,a)_B)(r',s')}{\mu_{t,\pi}^{B}(r',s')}} \right] \right] + \mathbf{1}_{(s,a)_{h_B}\neq(s,a)_B} \right] \right)$$

Now, sampling an r, s', and then sampling either 0 or 1 with r probability, is the same as sampling an r', s' from $\text{convert}(\sigma_{t,\pi}^{tr'})$.

$$= -\sum_{t=1}^{T} \ln \left(\mathop{\mathbb{E}}_{\pi,tr' \sim \xi_t} \left[\mathbf{1}_{(s,a)_{h_B}=(s,a)_B} \cdot \mathop{\mathbb{E}}_{r',s' \sim \text{convert}(\sigma_{t,\pi}^{tr'})} \left[\sqrt{\frac{M_{t,\pi}((h,s,a)_B)(r',s')}{\mu_{t,\pi}^{B}(r',s')}} \right] + \mathbf{1}_{(s,a)_{h_B}\neq(s,a)_B} \right] \right)$$

Because $\sigma_{t,\pi}^{tr'}$, for all t, π, tr', was chosen by the true environment, and $\mathbb{M}$ was a true RMDP (the environment respects its constraints), and this expectation is multiplied by an indicator function for being at $(h,s,a)_B$, when the indicator function is 1, $\sigma_{t,\pi}^{tr'}$ is guaranteed to be a selection from $\mathbb{M}((h,s,a)_B)$. Therefore, all of the $\text{convert}(\sigma_{t,\pi}^{tr'})$ distributions lie in $\text{convert}(\mathbb{M}(h,s,a)_B)$. By assumption, this set was a subset of Ψ_B. Because the distributions are all within our set of interest Ψ_B, we can apply the argument from the analysis of estimation complexity in Theorem 2 to yield

$$\geq -\sum_{t=1}^{T} \ln \left(\mathbb{E}_{\xi_t} \left[\mathbf{1}_{(s,a)_{h_B}=(s,a)_B} \cdot (1 - D_H^2(M_{t,\pi}((h,s,a)_B) \to \Psi_B)) + \mathbf{1}_{(s,a)_{h_B}\neq(s,a)_B} \right] \right)$$

$$= -\sum_{t=1}^{T} \ln \left(\mathbb{E}_{\xi_t} \left[1 - \mathbf{1}_{(s,a)_{h_B}=(s,a)_B} \cdot D_H^2(M_{t,\pi}((h,s,a)_B) \to \Psi_B) \right] \right)$$

$$= -\sum_{t=1}^{T} \ln \left(1 - \mathop{\mathbb{E}}_{\pi,s_{h_B},a_{h_B} \sim \xi_t} \left[\mathbf{1}_{s_{h_B}=s_B} \cdot \mathbf{1}_{a_{h_B}=a_B} \cdot D_H^2(M_{t,\pi}((h,s,a)_B) \to \Psi_B) \right] \right)$$

Now, the trajectory extends up to the action on turn h_B. It can be viewed as a trajectory which extends up to the state s_{h_B}, and the last action being selected via $\pi(h_B, s_{h_B})$. The expectation over this last action can be pushed past the Hellinger term and the indicator function over states, to yield the probability of taking the appropriate action.

$$= -\sum_{t=1}^{T} \ln \left(1 - \mathop{\mathbb{E}}_{\pi,s_{h_B} \sim \xi_t} \left[\mathbf{1}_{s_{h_B}=s_B} \cdot \pi(h_B, s_B)(a_B) \cdot D_H^2(M_{t,\pi}((h,s,a)_B) \to \Psi_B) \right] \right)$$

For $h_B = 0$, we can use that the indicator function for the correct state is always 1, as is the probability of the correct action, because only one state and action are possible. Also, there is no dependence on the trajectory, so we have

$$= -\sum_{t=1}^{T} \ln \left(1 - \mathop{\mathbb{E}}_{\pi \sim p_t} \left[D_H^2(M_{t,\pi}((h,s,a)_B) \to \Psi_B) \right] \right)$$

71

For $h \geq 1$, we can use that B', the calibration bettor corresponding to the fragment bettor B, has $(h, s, a)_{B'} = (h, s, a)_B$ and $\Psi_B = \Psi_{B'}$, so we can apply the definition of $X_{t,\pi}^{B'}$ to yield

$$= -\sum_{t=1}^{T} \ln \left(1 - \underset{\pi, s_{h_{B'}} \sim \xi_t}{\mathbb{E}} \left[X_{t,\pi}^{B'} \cdot \mathbf{1}_{s_{h_{B'}} = s_{B'}} \right] \right)$$

By $-\ln(1 - x) \geq x$ in both cases, for $h_B = 0$ we have

$$\geq \sum_{t=1}^{T} \underset{\pi \sim p_t}{\mathbb{E}} \left[D_H^2(M_{t,\pi}((h, s, a)_B) \to \Psi_B) \right]$$

And for $h_B \geq 1$ we have

$$\geq \sum_{t=1}^{T} \underset{\pi, s_{h_{B'}} \sim \xi_t}{\mathbb{E}} \left[X_{t,\pi}^{B'} \cdot \mathbf{1}_{s_{h_{B'}} = s_{B'}} \right]$$

Our net inequality is, for $h_B \geq 1$,

$$\ln \left(\frac{4(HS + 1)}{\delta} \right) + \sum_{t=1}^{T} \ln \left(\mathrm{bet}_{t,\pi_t}^B(tr_t) \right) \geq \sum_{t=1}^{T} \underset{\pi, s_{h_{B'}} \sim \xi_t}{\mathbb{E}} \left[X_{t,\pi}^{B'} \cdot \mathbf{1}_{s_{h_{B'}} = s_{B'}} \right]$$

And for $h_B = 0$,

$$\ln \left(\frac{4(HS + 1)}{\delta} \right) + \sum_{t=1}^{T} \ln \left(\mathrm{bet}_{t,\pi_t}^B(tr_t) \right) \geq \sum_{t=1}^{T} \underset{\pi \sim p_t}{\mathbb{E}} \left[D_H^2(M_{t,\pi}((h, s, a)_B) \to \Psi_B) \right]$$

To conclude in both cases, by reshuffling Lemma 6, we have

$$\ln(\zeta_{T+1}(B)) - \ln(\zeta_1(B)) + \sum_{t=1}^{T} \ln(\bigstar_t) = \sum_{t=1}^{T} \ln \left(\mathrm{bet}_{t,\pi_t}^B(tr_t) \right)$$

$\zeta_{T+1}(B)$ can be at most 1, and $\zeta_1(B)$, for fragment bettors, is $\frac{1}{2(|\mathcal{B}_{\mathrm{cal}}| + |\mathcal{B}_{\mathrm{frag}}|)}$ by construction yielding the inequality

$$\ln(2(|\mathcal{B}_{\mathrm{cal}}| + |\mathcal{B}_{\mathrm{frag}}|)) + \sum_{t=1}^{T} \ln(\bigstar_t) \geq \sum_{t=1}^{T} \ln \left(\mathrm{bet}_{t,\pi_t}^B(tr_t) \right)$$

Which, substituted into the previous equations, yields the desired result. ∎

Lemma 21 *With $1 - \frac{\delta}{2}$ probability, $\ln \left(\frac{2}{\delta} \right) \geq \sum_{t=1}^{T} \ln(\bigstar_t)$.*

By Lemma A.4 in Foster et al. (2021), with $1 - \frac{\delta}{2}$ probability, we have

$$\sum_{t=1}^{T} \ln(\bigstar_t) \leq \ln \left(\frac{2}{\delta} \right) + \sum_{t=1}^{T} \ln \left(\mathbb{E}_{\xi_t} \left[e^{\ln(\bigstar_t)} \right] \right) = \ln \left(\frac{2}{\delta} \right) + \sum_{t=1}^{T} \ln(\mathbb{E}_{\xi_t} [\bigstar_t])$$

Again, ξ_t is the distribution over the policy and the trajectory on episode t. Now, by applying definitions and expectation-shuffling, we have

$$\mathbb{E}_{\xi_t}[\bigstar_t] = \mathbb{E}_{\xi_t}\left[\mathop{\mathbb{E}}_{B\sim\zeta_t}\left[\mathrm{bet}_{t,\pi}^B(r_0,(s,a,r)_{1:H})\right]\right]$$

$$= \mathbb{E}_{\xi_t}\left[\mathop{\mathbb{E}}_{B\sim\zeta_t}\left[1 + \sum_{h=0}^{H}\left(\mathrm{lbet}_{t,\pi}^{B,h,s_h,a_h}(r_h,s_{h+1},a_{h+1}) - 1\right)\right]\right]$$

$$= 1 + \sum_{h=0}^{H}\mathbb{E}_{\xi_t}\left[\mathop{\mathbb{E}}_{B\sim\zeta_t}\left[\mathrm{lbet}_{t,\pi}^{B,h,s_h,a_h}(r_h,s_{h+1},a_{h+1})\right] - 1\right] \tag{9}$$

Letting $\nu_{r,s'} : \Delta(\{0,1\}\times\mathcal{S})$ be the distribution which assigns r measure to $1, s'$ and the rest to $0, s'$, define $\mathrm{res}_{t,\pi}^{B,h,s,a}(r,s',a')$ (a "residual" term) as

$$\mathrm{res}_{t,\pi}^{B,h,s,a}(r,s',a') := \mathrm{lbet}_{t,\pi}^{B,h,s,a}(r,s',a') - 1 + d(g_{t,\pi}^{B,h,s,a})_{M_{t,\pi}(h,s,a)}(M_{t,\pi}(h,s,a) - \nu_{r,s'})$$

That term is the Frechet derivative of the g function at $M_{t,\pi}(h,s,a)$, in a certain direction. With this definition, we can re-express 9 as

$$= 1 + \sum_{h=0}^{H}\mathbb{E}_{\xi_t}[\mathop{\mathbb{E}}_{B\sim\zeta_t}[\mathrm{res}_{t,\pi}^{B,h,s_h,a_h}(r_h,s_{h+1},a_{h+1})$$

$$-d(g_{t,\pi}^{B,h,s_h,a_h})_{M_{t,\pi}(h,s_h,a_h)}(M_{t,\pi}(h,s_h,a_h) - \nu_{r_h,s_{h+1}})]]$$

$$= 1 + \sum_{h=0}^{H}\mathbb{E}_{\xi_t}\left[\mathop{\mathbb{E}}_{B\sim\zeta_t}[\mathrm{res}_{t,\pi}^{B,h,s_h,a_h}(r_h,s_{h+1},a_{h+1})]\right]$$

$$- \mathop{\mathbb{E}}_{B\sim\zeta_t}\left[d(g_{t,\pi}^{B,h,s_h,a_h})_{M_{t,\pi}(h,s_h,a_h)}(M_{t,\pi}(h,s_h,a_h) - \nu_{r_h,s_{h+1}})\right]]$$

By linearity of differentiation, we get

$$= 1 + \sum_{h=0}^{H}\mathbb{E}_{\xi_t}[\mathop{\mathbb{E}}_{B\sim\zeta_t}\left[\mathrm{res}_{t,\pi}^{B,h,s_h,a_h}(r_h,s_{h+1},a_{h+1})\right]$$

$$-d\left(\mathop{\mathbb{E}}_{B\sim\zeta_t}\left[g_{t,\pi}^{B,h,s_h,a_h}\right]\right)_{M_{t,\pi}(h,s_h,a_h)}(M_{t,\pi}(h,s_h,a_h) - \nu_{r_h,s_{h+1}})]$$

Now, $M_{t,\pi}(h,s_h,a_h)$ was picked to be the minimizer of $\mathop{\mathbb{E}}_{B\sim\zeta_t}\left[g_{t,\pi}^{B,h,s_h,a_h}\right]$, a Frechet-differentiable convex function, so the derivative in every direction is zero. Interchanging expectations, we have

$$= 1 + \sum_{h=0}^{H}\mathop{\mathbb{E}}_{B\sim\zeta_t}\left[\mathbb{E}_{\xi_t}\left[\mathrm{res}_{t,\pi}^{B,h,s_h,a_h}(r_h,s_{h+1},a_{h+1})\right]\right] \tag{10}$$

73

Now we compute the residuals. By routine but tedious calculations of derivatives similar to those we did for the robust universal estimator in Theorem 2, the residual for the fragment bettors is zero. For the calibration bettors, the residuals are

$$\mathrm{res}_{t,\pi}^{B,h,s,a}(r,s',a') = \frac{X_{t,\pi}^B}{4}\left(\mathbb{P}_{\widehat{M}_t(\pi)|h+1,s'}((h,s)_B) - \mathbb{P}_{\widehat{M}_t(\pi)|h+1,s',a'}((h,s)_B)\right)$$

And for the pessimism bettor, the residual is

$$\mathrm{res}_{t,\pi}^{B,h,s,a}(r,s',a') = \varepsilon_{\mathrm{pess}}\left(\underset{\widehat{M}_t(\pi)|h+1,s'}{\mathbb{E}}\left[\sum_{k=h+1}^{H} r_k\right] - \underset{\widehat{M}_t(\pi)|h+1,s',a'}{\mathbb{E}}\left[\sum_{k=h+1}^{H} r_k\right]\right)$$

Importantly, in both cases, the residuals are zero in expectation, and we have

$$\underset{a'\sim\pi(h+1,s')}{\mathbb{E}}\left[\mathrm{res}_{t,\pi}^{B,h,s,a}(r,s',a')\right] = 0$$

We can expand 10 and break it up to yield

$$= 1 + \sum_{h=0}^{H}\underset{B\sim\zeta_t}{\mathbb{E}}\left[\underset{\pi,(s,a,r)_h,s_{h+1},a_{h+1}\sim\xi_t}{\mathbb{E}}\left[\mathrm{res}_{t,\pi}^{B,h,s_h,a_h}(r_h,s_{h+1},a_{h+1})\right]\right]$$

$$= 1 + \sum_{h=0}^{H}\underset{B\sim\zeta_t}{\mathbb{E}}\left[\underset{\pi,(s,a,r)_h,s_{h+1}\sim\xi_t}{\mathbb{E}}\left[\underset{a_{h+1}\sim\pi(h+1,s_{h+1})}{\mathbb{E}}\left[\mathrm{res}_{t,\pi}^{B,h,s_h,a_h}(r_h,s_{h+1},a_{h+1})\right]\right]\right] = 1$$

Chaining our equalities together, we have $\mathbb{E}_{\xi_t}[\bigstar_t] = 1$, which, substituting into the starting inequality, yields

$$\ln\left(\frac{2}{\delta}\right) \geq \sum_{t=1}^{T}\ln(\bigstar_t)$$

∎

Lemma 22 *The function* $\mathbb{E}_{B\sim\zeta_t}\left[g_{t,\pi}^{B,h,s,a}(\mu)\right]$ *has a unique minimum.*

First, any strictly convex function has a unique minimum, because if the set of minimizers contained two distinct points, a 50/50 mix of the minimizers would attain a strictly lower value, producing a contradiction. So, we just need to show that the mixture of g functions is strictly convex.

Second, a mixture of convex functions with one strictly convex function that has nonzero probability, is also strictly convex, because we get a non-strict inequality $\geq$ for the other convex functions, and a strict inequality for the strictly convex function. So, if we can verify that all of the g functions are convex, and that there is a g function corresponding to a bettor which retains nonzero probability, which is strictly convex, our desired result will follow.

The g functions associated with the calibration bettors and pessimism bettors are linear in μ, and the g functions associated with the fragment bettors and uniform bettor are convex in μ, by Lemma 4. To verify strict convexity for the uniform bettor, we may let ν,ν'

be unequal, and $p \in (0,1)$, and use that the set Ψ contains only a single point, the uniform distribution u. We then have

$$D_H^2(p\nu + (1-p)\nu', u) = 1 - \sum_{o,r} \sqrt{(p\nu(o,r) + (1-p)\nu'(o,r)) \cdot u(o,r)}$$

$$= 1 - \sum_{o,r} \sqrt{(p\nu(o,r)u(o,r) + (1-p)\nu'(o,r)u(o,r)}$$

The negative square root function is strictly convex. There is some o, r where $\nu(o,r) \neq \nu'(o,r)$ because $\nu \neq \nu'$, and $u(o,r)$ is nonzero because it is the uniform distribution, so we have a strict inequality for one o, r and a nonstrict inequality for the others, yielding

$$< 1 - p\sum_{o,r} \sqrt{\nu(o,r)u(o,r)} - (1-p)\sum_{o,r} \sqrt{\nu'(o,r)u(o,r)} = pD_H^2(\nu, u) + (1-p)D_H^2(\nu', u)$$

Because ν, ν', p were arbitrary, strict convexity of the bet made by the uniform bettor is verified. The uniform bettors always retain nonzero probability by the inductive argument from the end of Proposition 4, although we must also use that all $\bigstar_t > 0$ (proved in Lemma 18) to generalize the argument. Therefore, the function being minimized is a mixture of convex functions and a strictly convex function with nonzero probability, so it is strictly convex, so the minimum is unique. ∎

Appendix N. RMDP Estimator Theorems

Theorem 5 *There is an online estimator $\widehat{M}$ for hypotheses in $\mathcal{H}_{parhalf}$ where $\beta_{\widehat{M}}(T, \delta) \in \mathcal{O}\left(HS^2 \log(T) + HS \log\left(\frac{HSAT}{\delta}\right)\right)$ and $\alpha_{\widehat{M}}(T, \delta) \in \mathcal{O}\left(H\sqrt{T} \log\left(\frac{1}{\delta}\right)\right)$.*

We use the estimator defined in Appendix L, with ε' being arbitrarily small. By Lemma 18, it is well-defined. The parameters $\varepsilon_{\mathcal{S}}$ and $\varepsilon_{[0,1]}$ are taken as unspecified for now, and chosen later on in the proof to minimize the estimation error.

We begin by bounding $\beta_{\widehat{M}}$ for the hypothesis class $\mathcal{H}_{parhalf}$. Note that, to retain compatibility with our nonstandard notion of estimation complexity, we must use the associated definition of estimation error used in the previous theorems about the value of the modified DEC, so if $\mathbb{M} \in \mathcal{H}_{parhalf}$, we must upper-bound the following term with $1 - \delta$ probability.

$$\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\pi \sim p_t} \left[\mathop{\mathbb{E}}_{r_0,(s,a,r)_{1:H} \sim \widehat{M}_t(\pi)} \left[\sum_{h=0}^{H} \mathop{\mathbb{E}}_{a \sim \pi(h,s_h)} \left[D_{H_{\{0,1\}}}^2 \left((\widehat{M}_t(\pi)|h, s_h, a)_{\downarrow h} \to \mathbb{M}(h, s_h, a) \right) \right] \right] \right]$$

First, observe that $(\widehat{M}_t(\pi)|h, s_h, a)_{\downarrow h}$ is the distribution over the upcoming reward and state r_h, s_{h+1} if h, s_h, a occurs. Because $\widehat{M}_t(\pi)$ (our estimator) was defined from $M_{t,\pi}$ (our MDP), this is just $M_{t,\pi}(h, s_h, a)$.

$$= \sum_{t=1}^{T} \mathop{\mathbb{E}}_{\pi \sim p_t} \left[\mathop{\mathbb{E}}_{\widehat{M}_t(\pi)} \left[\sum_{h=0}^{H} \mathop{\mathbb{E}}_{a \sim \pi(h,s_h)} \left[D_{H_{\{0,1\}}}^2 \left(M_{t,\pi}(h, s_h, a) \to \mathbb{M}(h, s_h, a) \right) \right] \right] \right]$$

Pulling the sum over h out, we have

$$= \sum_{h=0}^{H} \sum_{t=1}^{T} \mathop{\mathbb{E}}_{\pi \sim p_t} \left[\mathop{\mathbb{E}}_{\widehat{M_t}(\pi)} \left[\mathop{\mathbb{E}}_{a \sim \pi(h,s_h)} \left[D^2_{H_{\{0,1\}}} \left(M_{t,\pi}(h, s_h, a) \to \mathbb{M}(h, s_h, a) \right) \right] \right] \right]$$

Because $\mathbb{E}_{\widehat{M_t}(\pi)}$ only determines s_h, we can write this as a sum over s of the probability of (h, s), and pull the sum out again, to yield

$$= \sum_{h=0}^{H} \sum_{s \in \mathcal{S}} \sum_{t=1}^{T} \mathop{\mathbb{E}}_{\pi \sim p_t} \left[\mathbb{P}_{\widehat{M_t}(\pi)}(h, s) \cdot \mathop{\mathbb{E}}_{a \sim \pi(h,s)} \left[D^2_{H_{\{0,1\}}} \left(M_{t,\pi}(h, s, a) \to \mathbb{M}(h, s, a) \right) \right] \right]$$

Now, because $\mathbb{M} \in \mathcal{H}_{\text{parhalf}}$ (the set of RMDP's), it was created from some PRMDPR in $\mathcal{H}_{\text{parhalf}}$ (the set of PRMDPR's). For this PRMDPR, every h, s is associated with some recommended action a, and an imprecise belief indexed by an $f : \mathcal{S} \to [0, 1]$ and a $c : [0, 1]$. We use $a_{h,s}, f_{h,s}, c_{h,s}$ to denote these. By how we go from PRMDPR's to RMDP's, this means that if $a \neq a_{h,s}$,
$\mathbb{M}(h, s, a) = \Delta([0, 1] \times \mathcal{S})$, but $\mathbb{M}(h, s, a_{h,s}) = \Psi^{[0,1]}_{f_{h,s}, c_{h,s}}$. This yields

$$= \sum_{h=0}^{H} \sum_{s \in \mathcal{S}} \sum_{t=1}^{T} \mathop{\mathbb{E}}_{\pi \sim p_t} \left[\mathbb{P}_{\widehat{M_t}(\pi)}(h, s) \cdot \pi(h, s)(a_{h,s}) \cdot D^2_{H_{\{0,1\}}} \left(M_{t,\pi}(h, s, a_{h,s}) \to \Psi^{[0,1]}_{f_{h,s}, c_{h,s}} \right) \right]$$

$M_{t,\pi}(h, s, a_{h,s})$ was already supported over 0 or 1 reward by its construction, so we may invoke Lemma 16 to go

$$= \sum_{h=0}^{H} \sum_{s \in \mathcal{S}} \sum_{t=1}^{T} \mathop{\mathbb{E}}_{\pi \sim p_t} \left[\mathbb{P}_{\widehat{M_t}(\pi)}(h, s) \cdot \pi(h, s)(a_{h,s}) \cdot D^2_{H} \left(M_{t,\pi}(h, s, a_{h,s}) \to \Psi^{\{0,1\}}_{f_{h,s}, c_{h,s}} \right) \right]$$

Now, given some $f : \mathcal{S} \to [0, 1]$, we use $\lceil f \rceil$ to denote the function $\lambda s. 1 - \left\lfloor \frac{1 - f(s)}{\varepsilon_{\mathcal{S}}} \right\rfloor \cdot \varepsilon_{\mathcal{S}}$, and given some $c : [0, 1]$, we use $\lfloor c \rfloor$ to denote $\left\lfloor \frac{c}{\varepsilon_{[0,1]}} \right\rfloor \cdot \varepsilon_{[0,1]}$. Intuitively, the hypothesis space $\mathcal{H}'_{\text{mid}}$ consists of the $\Psi_{f,c}$ where the values of f and c are discretized in this way, so we are rounding up and down as needed to find the closest discrete approximation to f, c. By Lemma 5, and linearity of expectation, we get

$$\leq 2 \sum_{h=0}^{H} \sum_{s \in \mathcal{S}} \sum_{t=1}^{T} \mathop{\mathbb{E}}_{\pi \sim p_t} \left[\mathbb{P}_{\widehat{M_t}(\pi)}(h, s) \cdot \pi(h, s)(a_{h,s}) \cdot D^2_{H} \left(M_{t,\pi}(h, s, a_{h,s}) \to \Psi^{\{0,1\}}_{\lceil f_{h,s} \rceil, \lfloor c_{h,s} \rfloor} \right) \right]$$

$$\tag{11}$$

$$+ 2 \sum_{h=0}^{H} \sum_{s \in \mathcal{S}} \sum_{t=1}^{T} \mathop{\mathbb{E}}_{\pi \sim p_t} \left[\mathbb{P}_{\widehat{M_t}(\pi)}(h, s) \cdot \pi(h, s)(a_{h,s}) \cdot D^2_{H} \left(\Psi^{\{0,1\}}_{\lceil f_{h,s} \rceil, \lfloor c_{h,s} \rfloor} \to \Psi^{\{0,1\}}_{f_{h,s}, c_{h,s}} \right) \right]$$

Focusing on that second term, we can rewrite it as

$$2 \sum_{t=1}^{T} \mathop{\mathbb{E}}_{\pi \sim p_t} \left[\sum_{h=0}^{H} \mathop{\mathbb{E}}_{s_h \sim \widehat{M_t}(\pi)} \left[\pi(h, s_h)(a_{h,s_h}) \cdot D^2_{H} \left(\Psi^{\{0,1\}}_{\lceil f_{h,s_h} \rceil, \lfloor c_{h,s_h} \rfloor} \to \Psi^{\{0,1\}}_{f_{h,s_h}, c_{h,s_h}} \right) \right] \right]$$

$$\leq 2\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\pi \sim p_t} \left[\sum_{h=0}^{H} \mathop{\mathbb{E}}_{s_h \sim \widehat{M}_t(\pi)} \left[D_H^2 \left(\Psi_{\lceil f_{h,s_h} \rceil, \lfloor c_{h,s_h} \rfloor}^{\{0,1\}} \to \Psi_{f_{h,s_h}, c_{h,s_h}}^{\{0,1\}} \right) \right] \right]$$

By Lemma 17, we may upper-bound this by

$$\leq 2T(H+1)(\varepsilon_{\mathcal{S}} + \varepsilon_{[0,1]})$$

Substituting this back in to 11 our upper bound on the estimation complexity is now

$$\leq 2\sum_{h=0}^{H}\sum_{s \in \mathcal{S}}\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\pi \sim p_t} \left[\mathbb{P}_{\widehat{M}_t(\pi)}(h,s) \cdot \pi(h,s)(a_{h,s}) \cdot D_H^2 \left(M_{t,\pi}(h,s,a_{h,s}) \to \Psi_{\lceil f_{h,s} \rceil, \lfloor c_{h,s} \rfloor}^{\{0,1\}} \right) \right]$$

$$+ 2T(H+1)(\varepsilon_{\mathcal{S}} + \varepsilon_{[0,1]})$$

We now focus on the first term. For every h, s, $\Psi_{\lceil f_{h,s} \rceil, \lfloor c_{h,s} \rfloor}^{\{0,1\}} \in \mathcal{H}'_{\text{mid}}$. Therefore, for every h, s with $h \geq 1$, the tuple $h, s, a_{h,s}, \Psi_{\lceil f_{h,s} \rceil, \lfloor c_{h,s} \rfloor}^{\{0,1\}}$ picks out a corresponding fragment bettor $B_{h,s}^{\text{frag}}$, and calibration bettor $B_{h,s}^{\text{cal}}$. If $h = 0$, only a fragment bettor is picked out, so we can rewrite our upper bound as

$$= 2\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\pi \sim p_t} \left[\mathbb{P}_{\widehat{M}_t(\pi)}(0,s_0) \cdot \pi(h_0,s_0)(a_0) \cdot D_H^2 \left(M_{t,\pi}(0,s_0,a_0) \to \Psi_{B_{h,s}^{\text{frag}}} \right) \right]$$

$$+ 2\sum_{h=1}^{H}\sum_{s \in \mathcal{S}}\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\pi \sim p_t} \left[\mathbb{P}_{\widehat{M}_t(\pi)}(h,s) \cdot \pi(h,s)(a_{h,s}) \cdot D_H^2 \left(M_{t,\pi}(h,s,a_{h,s}) \to \Psi_{B_{h,s}^{\text{cal}}} \right) \right]$$

$$+ 2T(H+1)(\varepsilon_{\mathcal{S}} + \varepsilon_{[0,1]})$$

For the first term, $\mathbb{P}_{\widehat{M}_t(\pi)}(0,s_0)$ and $\pi(h_0,s_0)(a_0)$ are both 1 because only one state and action is possible. For the second term, we pack up the definition of $X_{t,\pi}^{B_{h,s}^{\text{cal}}}$, and use that $(h,s) = (h,s)_{B_{h,s}^{\text{cal}}}$, to yield

$$= 2\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\pi \sim p_t} \left[D_H^2 \left(M_{t,\pi}(0,s_0,a_0) \to \Psi_{B_{h,s}^{\text{frag}}} \right) \right]$$

$$+ 2\sum_{h=1}^{H}\sum_{s \in \mathcal{S}}\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\pi \sim p_t} \left[X_{t,\pi}^{B_{h,s}^{\text{cal}}} \cdot \mathbb{P}_{\widehat{M}_t(\pi)}((h,s)_{B_{h,s}^{\text{cal}}}) \right] + 2T(H+1)(\varepsilon_{\mathcal{S}} + \varepsilon_{[0,1]})$$

By Lemma 19, and the union bound over h, s, we have that with $\geq 1 - \frac{\delta}{4}$ probability, we have an upper bound of

$$\leq 2\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\pi \sim p_t} \left[D_H^2 \left(M_{t,\pi}(0,s_0,a_0) \to \Psi_{B_{h,s}^{\text{frag}}} \right) \right]$$

$$+ 12HS \left(\ln \left(\frac{4(HS+1)}{\delta} \right) + \ln(2(|\mathcal{B}_{\text{cal}}| + |\mathcal{B}_{\text{frag}}|)) + \sum_{t=1}^{T} \ln(\bigstar_t) \right)$$

$$+4\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{t=1}^{T}\mathbb{E}_{\xi_t}\left[X_{t,\pi}^{B_{h,s}^{\text{cal}}}\cdot\mathbf{1}_{s_h=s}\right]+2T(H+1)(\varepsilon_{\mathcal{S}}+\varepsilon_{[0,1]})$$

From earlier in this theorem, we had that, for the true $\mathbb{M}$ in $\mathcal{H}_{\text{parhalf}}$, for all h, s, $\mathbb{M}(h, s, a_{h,s}) = \Psi_{f_{h,s}, c_{h,s}}^{[0,1]}$, by the definition of $a_{h,s}, f_{h,s}, c_{h,s}$. By the proof of Lemma 16, (specifically, the result that converting a $\Psi_{f,c}^{[0,1]}$ set just swaps the type signature in the superscript), this yields $\text{convert}(\mathbb{M}(h, s, a_{h,s})) = \Psi_{f_{h,s}, c_{h,s}}^{\{0,1\}}$. Also, by the we rounded the function up and the constant down, along with the definitions of the $\Psi_{f,c}^{\{0,1\}}$ beliefs, and the way we defined our fragment bettors of interest, we have $\Psi_{f_{h,s}, c_{h,s}}^{\{0,1\}} \subseteq \Psi_{\lceil f_{h,s}\rceil, \lfloor c_{h,s}\rfloor}^{\{0,1\}} = \Psi_{B_{h,s}^{\text{frag}}}$. Combining these, we have $\text{convert}(\mathbb{M}(h, s, a_{h,s})) \subseteq \Psi_{B_{h,s}^{\text{frag}}}$. This lets us invoke Lemma 20. By Lemma 20 and the union bound over the various h, s, we get that with $1 - \frac{\delta}{2}$ probability, our upper bound is now

$$\leq (16HS+2)\left(\ln\left(\frac{4(HS+1)}{\delta}\right) + \ln(2(|\mathcal{B}_{\text{cal}}| + |\mathcal{B}_{\text{frag}}|)) + \sum_{t=1}^{T}\ln(\bigstar_t)\right) + 2T(H+1)(\varepsilon_{\mathcal{S}}+\varepsilon_{[0,1]})$$

$$(12)$$

Then, by Lemma 21, and another application of the union bound, we get that with $1 - \delta$ probability, our upper bound on the estimation complexity is now

$$\leq (16HS+2)\left(\ln\left(\frac{4(HS+1)}{\delta}\right) + \ln(2(|\mathcal{B}_{\text{cal}}| + |\mathcal{B}_{\text{frag}}|)) + \ln\left(\frac{2}{\delta}\right)\right) + 2T(H+1)(\varepsilon_{\mathcal{S}}+\varepsilon_{[0,1]})$$

Now we compute the size of our hypothesis space. By the definitions of the sets of traders and our space of hypothesis fragments, we have

$$|\mathcal{B}|_{\text{cal}} < |\mathcal{B}_{\text{frag}}| \leq (HSA+1)|\mathcal{H}'_{\text{mid}}| \leq (HSA+1)\left(\frac{1}{\varepsilon_{\mathcal{S}}}+1\right)^{S}\left(\frac{1}{\varepsilon_{[0,1]}}+1\right)$$

Therefore, we have

$$\ln\left(2(|\mathcal{B}_{\text{frag}}| + |\mathcal{B}|_{\text{cal}})\right) < \ln\left(2(HSA+1)\left(\frac{1}{\varepsilon_{\mathcal{S}}}+1\right)^{S}\left(\frac{1}{\varepsilon_{[0,1]}}+1\right)\right)$$

$$= \ln(2(HSA+1)) + \ln\left(\frac{1}{\varepsilon_{[0,1]}}+1\right) + S\ln\left(\frac{1}{\varepsilon_{\mathcal{S}}}+1\right)$$

Combining this with 12, merging some logarithms together, and using that the $\mathbb{M}$, selection, algorithm choosing the policy, δ, and T were arbitrary, with $1 - \delta$ probability, the estimation complexity for $\mathcal{H}_{\text{parhalf}}$ is upper-bounded by

$$\leq (16HS+2)\left(\ln\left(\frac{1}{\varepsilon_{[0,1]}}+1\right) + S\ln\left(\frac{1}{\varepsilon_{\mathcal{S}}}+1\right) + \ln\left(\frac{16(HSA+1)(HS+1)}{\delta^2}\right)\right)$$

$$+2T(H+1)(\varepsilon_{\mathcal{S}}+\varepsilon_{[0,1]})$$

Minimizing over the two parameters, a near-optimal value for $\varepsilon_{\mathcal{S}}$ is $\frac{(8HS+1)S}{(H+1)T}$, and a near-optimal value for $\varepsilon_{[0,1]}$ is $\frac{(8HS+1)}{(H+1)T}$. Our estimator $\widehat{M}$ run with these two parameters would make the term $2T(H+1)(\varepsilon_{\mathcal{S}} + \varepsilon_{[0,1]})$ simplify to $(16HS+2)(S+1)$. The values $\ln\left(\frac{1}{\varepsilon_{[0,1]}}+1\right)$ and $\ln\left(\frac{1}{\varepsilon_{\mathcal{S}}}+1\right)$ may both be approximated as $\mathcal{O}(\log(T))$, yielding a final estimation complexity on the order of

$$\mathcal{O}\left(HS^2\log(T) + HS\log\left(\frac{HSAT}{\delta}\right)\right)$$

And so, this is our value of $\beta_{\widehat{M}}(T,\delta)$. To bound $\alpha_{\widehat{M}}(T,\delta)$, we follow the corresponding segment of the proof of Theorem 2, neglecting the arbitrarily low ε' in the same way. We have, from Lemma 6, rearranging, and using that around half of the probability mass was placed on the pessimism trader, in the $\varepsilon' \to 0$ limit,

$$\sum_{t=1}^{T}(\ln(\bigstar_t)) + \ln(2) \geq \sum_{t=1}^{T}\ln(\mathrm{bet}_{t,\pi_t}^{B}(tr_t))$$

Reusing our argument in Lemma 21 about the expected sum of $\ln(\bigstar_t)$ being less than $\ln\left(\frac{2}{\delta}\right)$ with $1 - \frac{\delta}{2}$ probability, and unpacking the bet for the pessimism trader, we get

$$\ln\left(\frac{4}{\delta}\right) \geq \sum_{t=1}^{T}\ln\left(1 + \varepsilon_{\mathrm{pess}}\left(\underset{\widehat{M}_t(\pi_t)}{\mathbb{E}}\left[\sum_{k=0}^{H}r_k\right] - \sum_{h=0}^{H}r_{h,t}\right)\right)$$

We may reuse the Azuma argument from Theorem 2 with $1 - \frac{\delta}{2}$ probability (to have a net $1 - \delta$ failure probability by the union bound). Reusing this argument requires making all necessary changes to account for the fact that the maximum value of the difference in reward is in $[0, H+1]$, not $[0,1]$. In order to apply our quadratic approximation to ln, we need that $\sqrt{\frac{1}{T(H+1)^2}}(H+1) \leq \frac{1}{2}$, which holds for all $T \geq 4$. Adding 4 to the final bound will account for this. Proceeding with Theorem 2's Azuma argument, making necessary changes, our net upper bound ends up being

$$\leq \frac{\ln\left(\frac{4}{\delta}\right)}{\varepsilon_{\mathrm{pess}}} + T \cdot \varepsilon_{\mathrm{pess}}(H+1)^2 + \sqrt{2\ln\left(\frac{2}{\delta}\right)T(H+1)^2}$$

Using that $\varepsilon_{\mathrm{pess}} = \sqrt{\frac{1}{T(H+1)^2}}$, factoring out appropriate terms, and adding the 4 onto the regret bound to account for the missing timesteps at the start, we are left with a net upper bound on $\alpha_{\widehat{M}}(T,\delta)$ of

$$(H+1)\sqrt{T}\left(\ln\left(\frac{4}{\delta}\right) + 1 + \sqrt{2\ln\left(\frac{2}{\delta}\right)}\right) + 4$$

This bound is $\mathcal{O}\left(H\sqrt{T}\ln\left(\frac{1}{\delta}\right)\right)$, as desired, and $\alpha_{\widehat{M}}(T,\delta)$ and $\beta_{\widehat{M}}(T,\delta)$ for this estimator have been computed. ∎

Proposition 7 *For the estimator of Theorem 5, all estimates*
$\widehat{M}_t : \Pi_{RNS} \to \Delta\left([0,1] \times (\mathcal{S} \times \mathcal{A} \times [0,1])^{[H]}\right)$ *are continuous and policy-coherent.*

Policy-coherence holds because the estimate $\widehat{M}(\pi)$ was constructed via having π interact with an MDP, which automatically produces policy-coherent estimates.

To show continuity of the estimator in π, we proceed as follows. The function $\lambda\pi.\widehat{M}(\pi)$ is an abbreviation for the function $\lambda\pi.\widehat{M}_t \bowtie \pi$, so we must show that this is a continuous function.

If $\lambda\pi.M_{t,\pi}(h,s,a)$ of type $\Pi_{RNS} \to \Delta(\{0,1\} \times \mathcal{S})$ were continuous in π for all h,s,a, then as π_n converges to a limiting π, every trajectory tr (and there are only finitely many) would have $\mathbb{P}_{M_{t,\pi_n} \bowtie \pi_n}(tr)$ limit to $\mathbb{P}_{M_{t,\pi} \bowtie \pi}(tr)$, because the probability of a trajectory can be written as a product of transition probabilities for the environment and the policy. All of the finitely many transition probabilities would converge as π_n converges to π, leading to convergence of the probability of every trajectory, so the distribution $M_{t,\pi_n} \bowtie \pi_n$ would converge to the distribution $M_{t,\pi} \bowtie \pi$, certifying continuity as desired.

Therefore, to prove continuity, it suffices to prove that, for all h,s,a, the function $\lambda\pi.M_{t,\pi}(h,s,a)$ is continuous. We will prove this fact by downwards induction, so we may select an arbitrary h,s,a, and freely assume that $\lambda\pi.M_{t,\pi}(h',s',a')$ is continuous for all s',a' and $h' > h$.

First, we verify continuity of the function $\lambda\pi,\mu.(\mu, M_{t,\pi}^{>h}) \bowtie \pi$. Let π_n, μ_n limit to π, μ. The MDP $(\mu_n, M_{t,\pi_n}^{>h})$ converges, in its starting distribution, and all of its transition probabilities, to $(\mu, M_{t,\pi}^{>h})$, by the induction assumption that $M_{t,\pi}(h',s',a')$ was continuous for all s',a' and $h' > h$, and convergence of μ_n. Then, by our earlier argument about there being finitely many trajectories, and convergence in the probability of each trajectory, the distributions $(\mu_n, M_{t,\pi_n}^{>h}) \bowtie \pi_n$ converge to $(\mu, M_{t,\pi}^{>h}) \bowtie \pi$. Our convergent sequence was arbitrary, so continuity is established.

This function that we just established the continuity of, is the same as the function $\lambda\pi,\mu.\widetilde{M}_{t,\pi}^{>h}(\mu,\pi)$. Taking a distribution over trajectories, and asking for the probability of some h',s' event, or the expected sum of future rewards, is a continuous function from probability distributions over trajectories to $\mathbb{R}$. Compositions of continuous functions are continuous, so, the functions

$\lambda\pi,\mu.\mathbb{E}_{\widetilde{M}_{t,\pi}^{>h}(\mu,\pi)}\left[\sum_{k=h}^{H} r_k\right]$ and $\lambda\pi,\mu.\mathbb{P}_{\widetilde{M}_{t,\pi}^{>h}(\mu,\pi)}(h',s')$ are continuous.

Further, the quantities $X_{t,\pi}^B$, are continuous in π, because they only depend on π through the probability of a particular action in a particular situation, and this probability converges as the policy does.

Now, for the various bettors B, we look at the continuity of the function $\lambda\pi,\mu.g_{t,\pi}^{B,h,s,a}(\mu)$. For fragment and uniform bettors, their g functions are either constants, or constants in π and continuous in μ. For calibration bettors, their g functions are either constants, or they depend on multiplying two continuous functions of π,μ together (the $X_{t,\pi}^B$ quantity, and the probability of some h',s'), so they are continuous. For the pessimism bettor, it is just a constant times a continuous function of π,μ (the expectation of future reward). Therefore, for all bettors B, the function $\lambda\pi,\mu.g_{t,\pi}^{B,h,s,a}(\mu)$ is continuous.

In the special case of $h = H$, the fragment and uniform g functions only depend continuously on μ, not π. The calibration g functions are all constants, and the pessimism

bettor only depends on μ, so continuity of all of the g functions in π, μ holds automatically (because it doesn't depend on π and depends continuously on μ), which provides a base case for the downwards induction that doesn't require the induction step to prove, rendering everything well-founded.

Because all of the functions $\lambda \pi, \mu.g_{t,\pi}^{B,h,s,a}(\mu)$ of type $\Pi_{RNS} \times \Delta(\{0,1\} \times \mathcal{S}) \to \mathbb{R}$ are continuous, their mixture, $\lambda \pi, \mu. \underset{B \sim \zeta_t}{\mathbb{E}}[g_{t,\pi}^{B,h,s,a}(\mu)]$ is as well. $\Pi_{RNS} \times \Delta(\{0,1\} \times \mathcal{S})$ is a compact space, so, by the Heine-Cantor theorem, the mixture function is uniformly continuous. Therefore, if π_n converges to π, $\underset{\mu}{\max}| \underset{B \sim \zeta_t}{\mathbb{E}}[g_{t,\pi_n}^{B,h,s,a}(\mu)] - \underset{B \sim \zeta_t}{\mathbb{E}}[g_{t,\pi}^{B,h,s,a}(\mu)]|$ converges to 0 (because for all μ, the low distance between π_n and π translates into a guarantee on the difference of the values, by uniform continuity), so the functions $\underset{B \sim \zeta_t}{\mathbb{E}}[g_{t,\pi_n}^{B,h,s,a}]$ uniformly converge to $\underset{B \sim \zeta_t}{\mathbb{E}}[g_{t,\pi}^{B,h,s,a}]$.

Now, we show that $M_{t,\pi_n}(h,s,a)$ limits to $M_{t,\pi}(h,s,a)$, as follows. Because $\Delta(\{0,1\} \times \mathcal{S})$ is a compact space, the sequence $M_{t,\pi_n}(h,s,a)$ of distributions has limit points, call them M. We will show that all limit points must equal $M_{t,\pi}(h,s,a)$, so the sequence converges to the desired point. To do this, we pass to a convergent subsequence, and compute

$$\underset{B \sim \zeta_t}{\mathbb{E}}[g_{t,\pi}^{B,h,s,a}(M)] = \lim_{m \to \infty} \underset{B \sim \zeta_t}{\mathbb{E}}[g_{t,\pi_{n_m}}^{B,h,s,a}(M_{t,\pi_{n_m}}(h,s,a))]$$

$$= \lim_{m \to \infty} \min(\underset{B \sim \zeta_t}{\mathbb{E}}[g_{t,\pi_{n_m}}^{B,h,s,a}]) = \min(\underset{B \sim \zeta_t}{\mathbb{E}}[g_{t,\pi}^{B,h,s,a}])$$

In order, this was by continuity of the mix of g functions in π and the distribution, because M was a limit of a sequence of distributions. The second equality was because $M_{t,\pi_{n_m}}(h,s,a)$ is the minimize of the associated function. Finally, we use our result that the functions $\underset{B \sim \zeta_t}{\mathbb{E}}[g_{t,\pi_{n_m}}^{B,h,s,a}]$ uniformly converge to $\underset{B \sim \zeta_t}{\mathbb{E}}[g_{t,\pi}^{B,h,s,a}]$, so their minimum values converge as well. This certifies that any limit point of the sequence $M_{t,\pi_n}(h,s,a)$ minimizes $\underset{B \sim \zeta_t}{\mathbb{E}}[g_{t,\pi}^{B,h,s,a}]$. However, by Lemma 22, the minimizer is unique. Therefore, the sequence $M_{t,\pi_n}(h,s,a)$ converges to $M_{t,\pi}(h,s,a)$. The convergent sequence π_n was arbitrary, as was s,a, so we have established that all the functions $\lambda \pi.M_{t,\pi}(h,s,a)$ are continuous, if continuity holds for all $h' > h$. Therefore, our downwards induction proof of continuity in π for the transition probabilities goes through (the base case at H does not rely on the induction assumption), which implies continuity of the overall estimator. ■

Corollary 3 *There is an algorithm which attains $\tilde{\mathcal{O}}(\sqrt{H^2 S^3 AT})$ regret on all 1-bounded RMDP's in the episodic RMDP setting.*

Take the failure probability, δ, to be $T^{-1/2}$. By Theorem 5 and Proposition 7, there is a continuous estimator for $\mathcal{H}_{\text{parhalf}}$ with $\beta_{\widehat{M}}(T,\delta) \in \widetilde{O}(HS^2)$, and $\alpha_{\widehat{M}}(T,\delta) \in \widetilde{O}(\sqrt{T})$. By Theorem 4, because the estimates are continuous, the modified DEC for the 1-bounded subset of $\mathcal{H}_{\text{parhalf}}$ is on the order of $\mathcal{O}(\sqrt{HSA}\varepsilon)$. By Theorem 1 applied to the expected sum of Hellinger losses, along with ε being $\sqrt{\frac{\beta_{\widehat{M}}(T,\delta)}{T}}$, the Theorem 5 estimator may be combined with the E2D algorithm to attain $\tilde{\mathcal{O}}(\sqrt{H^2 S^3 AT})$ regret on the 1-bounded subset of $\mathcal{H}_{\text{parhalf}}$.

To transfer this result to all 1-bounded RMDP's, we must show that, given any 1-bounded RMDP $\mathbb{M}$, there is a surrogate RMDP $\mathbb{M}' \in \mathcal{H}_{\text{parhalf}}$ where, if $\mathbb{M}$ is a 1-bounded true model of the environment, then the surrogate is also a 1-bounded true model with the same worst-case expected reward for optimal play. Then, if $\mathbb{M}$ was a true hypothesis (a hypothesis that the environment respects the constraints of), the same would apply to $\mathbb{M}'$. The E2D algorithm would guarantee low regret against $\mathbb{M}'$ because it is true and in $\mathcal{H}_{\text{parhalf}}$ and 1-bounded. Finally, because $\mathbb{M}, \mathbb{M}'$ ensure the same bound on expected reward, the low regret against $\mathbb{M}'$ implies low regret against $\mathbb{M}$. If this argument applies to all $\mathbb{M}$, the corollary follows.

To prove that every 1-bounded RMDP $\mathbb{M}$ has a surrogate with the desired properties, we introduce the following definition. Given an RMDP $\mathbb{M}$ and a policy $\pi \in \Pi_{RNS}$, define the value of h, s as follows. For $h = H + 1$, it is zero. For other h, s, it is defined as

$$V_{h,s}^{\mathbb{M},\pi} = \mathop{\mathbb{E}}_{a \sim \pi(h,s)} \left[\min_{\mu \in \mathbb{M}(h,s,a)} \mathop{\mathbb{E}}_{r,s' \sim \mu} \left[r + V_{h+1,s'}^{\mathbb{M},\pi} \right] \right]$$

This is the expected sum of future rewards at h, s if the environment will make the worst choices which are compatible with the RMDP $\mathbb{M}$.

For any RMDP, $f^{\mathbb{M}}(\pi)$ (the worst-case expected sum of rewards) equals $V_{0,s_0}^{\mathbb{M},\pi}$. 1-boundedness is equivalent to $\forall h, s, \pi : V_{h,s}^{\mathbb{M},\pi} \leq 1$. Further, given an RMDP, its optimal policy may be chosen to be deterministic, because, for any h, s, $V_{h,s}^{\mathbb{M},\pi}$ may be maximized by a deterministic choice of action.

Now, given an $\mathbb{M}$, define the surrogate $\mathbb{M}'$ as follows. Let $\pi_{\mathbb{M}}$ be a deterministic optimal policy for that RMDP. Then, $\mathbb{M}'$ is the RMDP generated by the PRMDPR which maps h, s to $\pi_{\mathbb{M}}(h,s), \Psi_{\lambda s'.V_{h+1,s'}^{\mathbb{M},\pi_{\mathbb{M}}}, V_{h,s}^{\mathbb{M},\pi_{\mathbb{M}}}}^{[0,1]}$. This is in $\mathcal{H}_{\text{parhalf}}$ by construction.

First, $\pi_{\mathbb{M}}$ is an optimal policy for the surrogate $\mathbb{M}'$, and any other policy attains the same score or worse in all circumstances. This holds because, when a PRMDPR is converted to an RMDP, the transition kernel for non-recommended actions is completely unconstrained, and it's possible that, for all actions which aren't recommended, the environment returns zero reward, and a transition to the worst state. Therefore, in any situation h, s, the recommended action produces the maximum guarantee on expected future reward. Because the optimal action in every situation coincides with the action of $\pi_{\mathbb{M}}$, this policy is optimal for the surrogate $\mathbb{M}'$.

Second, we will show that, for any h, s, we have $V_{h,s}^{\mathbb{M},\pi_{\mathbb{M}}} = V_{h,s}^{\mathbb{M}',\pi_{\mathbb{M}}}$. This is proved by downwards induction. It holds at the base case of $h = H + 1$ because the values are zero. For the induction step, we compute

$$V_{h,s}^{\mathbb{M}',\pi_{\mathbb{M}}} = \min_{\mu \in \mathbb{M}'(h,s,\pi_{\mathbb{M}}(h,s))} \mathop{\mathbb{E}}_{r,s' \sim \mu} \left[r + V_{h+1,s'}^{\mathbb{M},\pi} \right] = V_{h,s}^{\mathbb{M},\pi_{\mathbb{M}}}$$

The last equality was because $\mathbb{M}'(h, s, \pi_{\mathbb{M}}) = \Psi_{\lambda s'.V_{h+1,s'}^{\mathbb{M},\pi_{\mathbb{M}}}, V_{h,s}^{\mathbb{M},\pi_{\mathbb{M}}}}^{[0,1]}$, and this Ψ set is

$\left\{ \mu \,\middle|\, \mathop{\mathbb{E}}_{r,s' \sim \mu} [r + V_{h+1,s'}^{\mathbb{M},\pi_{\mathbb{M}}}] \geq V_{h,s}^{\mathbb{M},\pi_{\mathbb{M}}} \right\}$. By induction, we then have that $V_{h,s}^{\mathbb{M},\pi_{\mathbb{M}}} = V_{h,s}^{\mathbb{M}',\pi_{\mathbb{M}}}$ for

all h, s. Because $\mathbb{M}$ was 1-bounded (all the values are ≤ 1), $\mathbb{M}'$ is as well. Because $\pi_{\mathbb{M}}$ is optimal for $\mathbb{M}'$, we also have

$$\max(f^{\mathbb{M}}) = f^{\mathbb{M}}(\pi_{\mathbb{M}}) = V^{\mathbb{M},\pi_{\mathbb{M}}}_{0,s_0} = V^{\mathbb{M}',\pi_{\mathbb{M}}}_{0,s_0} = f^{\mathbb{M}'}(\pi_{\mathbb{M}}) = \max(f^{\mathbb{M}'})$$

All that remains to prove the corollary is to show that, if $\mathbb{M}$ is a true model of the environment (at every h, s, a, the distribution over the next reward and state is in $\mathbb{M}(h, s, a)$), the surrogate $\mathbb{M}'$ is as well. This may be proven by showing that, for all h, s, a we have $\mathbb{M}(h, s, a) \subseteq \mathbb{M}'(h, s, a)$. For an h, s, a where a is not the recommended action, $\mathbb{M}'(h, s, a)$ is the entire space of distributions, so this case is trivial. If the action is the recommended action $\pi_{\mathbb{M}}(h, s)$, we have, for any $\mu \in \mathbb{M}(h, s, \pi_{\mathbb{M}}(h, s))$,

$$\mathop{\mathbb{E}}_{r,s'\sim\mu}[r + V^{\mathbb{M},\pi_{\mathbb{M}}}_{h+1,s'}] \geq \min_{\mu'\in\mathbb{M}(h,s,\pi_{\mathbb{M}}(h,s))} \mathop{\mathbb{E}}_{r,s'\sim\mu'}[r + V^{\mathbb{M},\pi_{\mathbb{M}}}_{h+1,s'}] = V^{\mathbb{M},\pi_{\mathbb{M}}}_{h,s}$$

This inequality certifies $\mu \in \Psi^{[0,1]}_{\lambda s'.V^{\mathbb{M},\pi_{\mathbb{M}}}_{h+1,s'},V^{\mathbb{M},\pi_{\mathbb{M}}}_{h,s}} = \mathbb{M}'(h, s, \pi_{\mathbb{M}}(h, s))$. μ was arbitrary, so $\mathbb{M}(h, s, \pi_{\mathbb{M}}(h, s)) \subseteq \mathbb{M}'(h, s, \pi_{\mathbb{M}}(h, s))$. This concludes the proof. $\blacksquare$