Online Model Selection for Reinforcement Learning with Function Approximation

Jonathan N. Lee Stanford University
Aldo Pacchiano UC Berkeley
Vidya Muthukumar Georgia Tech
Weihao Kong University of Washington
Emma Brunskill Stanford University

Abstract

Deep reinforcement learning has achieved impressive successes yet often requires a very large amount of interaction data. This result is perhaps unsurprising, as using complicated function approximation often requires more data to fit, and early theoretical results on linear Markov decision processes provide regret bounds that scale with the dimension of the linear approximation. Ideally, we would like to automatically identify the minimal dimension of the approximation that is sufficient to encode an optimal policy. Towards this end, we consider the problem of model selection in RL with function approximation, given a set of candidate RL algorithms with known regret guarantees. The learner's goal is to adapt to the complexity of the optimal algorithm without knowing it *a priori*. We present a meta-algorithm that successively rejects increasingly complex models using a simple statistical test. Given at least one candidate that satisfies realizability, we prove the meta-algorithm adapts to the optimal complexity with regret that is only marginally suboptimal in the number of episodes and number of candidate algorithms. The dimension and horizon dependencies remain optimal with respect to the best candidate, and our meta-algorithmic approach is flexible to incorporate multiple candidate algorithms and models. Finally, we show that the meta-algorithm automatically admits significantly improved instance-dependent regret bounds that depend on the gaps between the maximal values attainable by the candidates.

1 INTRODUCTION

Deep reinforcement learning has achieved impressive successes, yet often requires a very large amount of interaction data. This result is perhaps unsurprising, as more complicated function approximations often require more data to fit. Recent work on theoretical reinforcement learning for some structured function approximation settings has shown regret bounds that scale with a parameter characterizing the complexity of a particular function class. For example, for a type of function approximation by a d-dimensional linear model in Markov decision processes (MDPs), prior work has provided bounds that scale as $O(d^{3/2})$ regret (Jin et al., 2020b), which have been improved to $O(d)$ even given small inherent Bellman error (Zanette et al., 2020). When the dynamics can be expressed using a matrix, $O(d^{3/2})$ regret bounds have also been provided (Yang and Wang, 2020). The choice of dimension d is important: on one hand, if d is underspecified, such regret bounds typically either fail to hold or incur linear regret. On the other hand, if d is over-specified, the above regret bounds are unnecessarily large. Thus, a natural goal is to use the most compact representation suitable to encode the optimal policy for a domain (which we denote as d_*). This optimal representation is typically unknown a priori.

In this paper we frame this as a model selection question among a set of algorithms with model classes, parameterized by dimensions $\{d \geq 1\}$, that are nested in their regret bound guarantees. We assume that at least one class can realize the true underlying domain. We ask if is there an algorithm that can achieve regret bounds that scale with the minimal realizable model class, given by d_*. Doing so seems subtle: provably efficient reinforcement learning algorithms typically rely heavily on strategic exploration, and using the wrong model class during learning may alias states, resulting in performance that appears strong under the current (incorrect) model class but is actually suboptimal. Conversely, forced exploration under more complex classes mitigates this problem, but could introduce regret that scales with the more complex model

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

class dependence, even when a simpler model suffices.

Most prior work on model selection for online decision making has focused on contextual bandit settings. Here, minimax-optimal guarantees were recently shown under eigenvalue assumptions on the features by leveraging the special structure of the stochastic linear contextual bandit setting (Chatterji et al., 2020; Foster et al., 2019). These results also assume the *knowledge* of a good exploration policy, but such knowledge cannot be relied on in the reinforcement learning setting, where some "high-reward" states may only be observed under specific, initially unknown sequences of actions. Slightly weaker model selection guarantees can also be obtained under far more general assumptions by using a *corralling framework* that assumes access to a set of base algorithms, and provides a meta-algorithm that aims to realize the best regret of the (unknown) best algorithm (Agarwal et al., 2017; Arora et al., 2020; Pacchiano et al., 2020).

Our contributions We tackle the challenge of model selection in RL under minimal assumptions. Our main insight is to leverage the knowledge of expected regret that is achievable under a particular model *when it realizes the data.* Thus, we propose an algorithm in Section 4 that maintains a candidate set of model classes at every round, and statistically tests whether each of them is well-specified, or not, by comparing the observed returns under that model class to the regret we should expect from a well-specified model. Model classes detected as misspecified at any round are permanently eliminated there-after in a manner reminiscent of *active-arm elimination* in the multi-armed bandit problem (Even-Dar et al., 2006); this is a significant simplification over previous meta-algorithms for model selection that were based on adversarial bandit algorithms. Our choice of action at every round carefully interleaves executing the candidate model class of minimal complexity with executing algorithms using higher-order models. This procedure is shown to automatically satisfy the needed exploration-exploitation trade-off for model selection. In Section 5, we show the regret bounds exactly match the model complexity of the unknown best model in d_* (and the finite episode length H in RL), and achieve a $T^{2/3}$ rate when the underlying algorithms have a $T^{1/2}$ rate under minimal assumptions about the underlying dynamics process. This is similar to recent model selection algorithms under general assumptions (Pacchiano et al., 2020) which sacrifice either a tight dependence on T or d_*. We also demonstrate how our approach is compatible with multiple recently introduced RL results, and provide specific bounds for model selection in such settings. In addition to our algorithm being simpler than a recent model-selection approach (Pacchiano et al., 2020), we provide new, significantly improved bounds for instances in which there is a constant gap in performance between model classes in Section 6. These guarantees are in part *instance-dependent*, as they scale inversely with this performance gap. From a practical perspective, our wrapper algorithm can be used given any input algorithms with regret guarantees that are nested, which will allow it to directly inherit future advances in provably efficient reinforcement learning. Finally, the computational complexity of our meta-algorithm only adds an extra factor on the order of the total number of model classes over and above the computational complexity of a single base algorithm.

2 RELATED WORK

The problem of model selection in online decision-making environments with limited-information feedback (which includes both bandits and reinforcement learning), has been an active area of recent research (Agarwal et al., 2017; Chatterji et al., 2020; Foster et al., 2019; Pacchiano et al., 2020) and poses challenges that are both statistical and algorithmic.

Nearly Optimal Online Model Selection The best available guarantees for online model selection have been obtained for the linear contextual bandits setting (Abbasi-Yadkori et al., 2011; Chu et al., 2011). Here, the best worst-case bound when the optimal model class is given is of the form $\mathcal{O}(\sqrt{d_* T})$, where d_* is the dimension of the minimal feature space that realizes the data and T is the total number of rounds: in model selection, several models with different d are provided and the minimal d_* is unknown. When the contextual information is stochastic, Foster et al. (2019) obtain model selection guarantees of the form $\mathcal{O}(d_*^{1/3} T^{2/3})$ under an action-averaged eigenvalue condition, and Chatterji et al. (2020) match the optimal guarantee when choosing between multi-armed bandits and contextual bandits under a stronger universal eigenvalue condition that ensures that contexts corresponding to all arms are sufficiently diverse. The results of Foster et al. (2019) leverage the fact that it is possible to estimate the optimal value under the optimal model (what we will denote as V^* in this paper) at a faster rate of $\sqrt{d}/n$ as compared to finding the optimal policy under the complex model (which has estimation error rate d/n). Both critically leverage both stochasticity of contextual information and linearity of the model. These bandit approaches also rely on *a priori* access to a policy that explores the environment and allows for off-policy estimation. However, reward-free exploration in RL (Jin et al., 2020a; Wang et al., 2020; Zanette et al., 2020) can sometimes be as or more complex than estimating the optimal policy.

Though there has been some work on offline feature selection and model selection for RL given a batch of data (see e.g. Farahmand and Szepesvári (2011); Hallak et al. (2013); Jiang et al. (2015); Parr et al. (2008)), there has been very little work specifically on online model selection in reinforcement learning. Prior work provided PAC results for online feature selection for factored tabular MDPs (Guo and Brunskill, 2018). More recent work provides regret bounds (Abbasi-Yadkori et al., 2020) and PAC bounds (Modi et al., 2020) for model selection in online RL when the optimal value V^* is given: however, unlike in contextual bandits (Foster et al., 2019; Kong et al., 2020), there are no known algorithms for estimating V^* faster than identifying the optimal policy in RL settings.

Corralling Methods Other researchers have provided general-purpose meta-algorithms designed for model selection for bandit settings that yield weaker, but still non-trivial and interesting statistical guarantees of the form $\mathcal{O}(\mathcal{R}_*^\alpha T^\beta)$ for arbitrary $\alpha \geq 1, \beta < 1$, where $\mathcal{R}_*$ depends generally on the complexity of the best model class or algorithm and other problem parameters. The early corralling algorithms for stochastic and adversarial bandits (Agarwal et al., 2017), have recently been simplified and improved under a mild stochastic assumption on the data (Pacchiano et al., 2020), using a novel smoothing technique broadly applicable to base algorithms with a regret guarantee. This *stochastic corralling* approach obtains model selection rates with $\alpha = 2, \beta = 1/2$ or $\alpha = 1, \beta = 2/3$ under very general assumptions including the RL setting; however, for technical reasons it still requires a complex two step smoothing procedure to modify the base algorithms to satisfy its regret guarantees. Our approach recovers rates of the form $\alpha = 1, \beta = 2/3$ (provided in Section 5) matching prior work without sacrificing generality and with a significantly simplified and interpretable algorithm design. This simplicity largely arises from using a stochastic master rather than an adversarial master. As a consequence, our same algorithm can be analyzed to provide new significantly stronger model selection guarantees for instances that have a constant gap in performance between model classes; these guarantees are provided in Section 6. Moreover, side information or faster estimators of the optimal value V^*, if available, can be naturally incorporated into our design to provide near-optimal rates; see Appendix D for precise statements of these guarantees.

3 SETTING

We consider the setting of an episodic Markov decision process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{U}, H, r, P, \rho)$, where $\mathcal{S}$ and $\mathcal{U}$ are state and action spaces, $H \in \mathbb{N}$ is the length of an episode, $r = \{r_h(s_h, u_h)\}$ is the per-step reward function with $r_h(s_h, u_h) \in [0, 1]$, $P = \{P_h(s_{h+1}|s_h, u_h)\}$ is the transition dynamics, and $\rho(s)$ is a fixed initial state distribution. A policy maps times and states to actions, $\pi : [H] \times \mathcal{S} \to \mathcal{U}$.

For a given $h \in [H]$ and $s \in \mathcal{S}$, the value function is the expected cumulative reward following policy π:

$$V_h^\pi(s) := \mathbb{E}_\pi \left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, u_{h'}) | s_h = s \right]$$

and similarly the action-value function is defined as the expected return from first taking action u and then following policy π: $Q_h^\pi(s, u) = r_h(s, u) + \mathbb{E}_{s' \sim P_h(\cdot|s,u)} V_{h+1}^\pi(s')$. The optimal value function is denoted $V_h^*(s) = \sup_\pi V_h^\pi(s)$. We write $V^\pi := \mathbb{E}_{s\sim\rho} V_1^\pi(s)$ and denote the optimal value under ρ as $V^* = \sup_\pi V^\pi$. In this work we primarily evaluate the quality of an algorithm $\mathcal{A}$ in an MDP $\mathcal{M}$ by its regret[1] with respect to the (unknown) optimal policy value V^* over T episodes:

$$\text{Regret}_T(\mathcal{A}; \mathcal{M}) := \sum_{t=1}^{T} V^* - V^{\pi_t}. \qquad (1)$$

We are interested in settings where the size of the state space $\mathcal{S}$ and/or action space $\mathcal{U}$ could be very large. Hence, we focus on function approximation methods for minimizing regret. A function approximation algorithm takes as input a model class $\mathcal{F}$ to generalize across states and actions (Agarwal et al., 2019). Several natural examples include value-based classes where $\mathcal{F} : \mathcal{S} \times \mathcal{U} \to \mathbb{R}$ is used to predict action-value functions Q^π and model-based classes where $\mathcal{F} : \mathcal{S} \times \mathcal{U} \times \mathcal{S} \to \mathbb{R}$ is used to predict the transition dynamics P and reward r. Concretely, linear MDPs (Jin et al., 2020b; Yang and Wang, 2020) model the transition dynamics as $\langle \phi(s, a), \mu(s') \rangle$, where $\phi \in \mathbb{R}^d$ and μ_h is a d-dimensional vector of measures.

We let $(\mathcal{A}, \mathcal{F})$ denote the pair of algorithm $\mathcal{A}$ equipped with model class $\mathcal{F}$. Recent high probability regret (upper) bounds in this setting are sublinear in T and typically depend polynomially on $d_\mathcal{F}$, H, and $\log(T/\delta)$, where $d_\mathcal{F}$ is a measure of statistical complexity of $\mathcal{F}$ and $\delta \in (0, 1)$ is a failure probability. For example, if $\mathcal{F}$ is finite, we often have $d_\mathcal{F} = \log |\mathcal{F}|$ and if $\mathcal{F}$ is a class of linear functions of dimension d, we have $d_\mathcal{F} = d$. However, provably sublinear regret bounds in T are generally only known for algorithms under problem-specific assumptions for $\mathcal{F}$—for

[1] Note that regret is here defined with respect to the optimal value. We will also consider algorithms satisfying "best-in-class" regret guarantees in Section 6.

example, there exists $f^* \in \mathcal{F}$ such that the function approximation error is 0. If this condition holds, we say that $\mathcal{F}$ *realizes* the MDP $\mathcal{M}$. Conversely, if $\mathcal{F}$ does not realize $\mathcal{M}$, then it is *misspecified.* Since we consider settings where $\mathcal{F}$ may or may not realize $\mathcal{M}$ and realizability is almost universally assumed among modern RL algorithms with function approximation, we define a general notion of the regret of $\mathcal{A}$ using $\mathcal{F}$ under realizability, following Pacchiano et al. (2020).

Definition 1. *For an MDP $\mathcal{M}$, let algorithm $\mathcal{A}$ be equipped with a model class $\mathcal{F}$. Let $\mathcal{R}$ be a known function that is* $\mathrm{poly}(d_{\mathcal{F}}, H, \log(T/\delta))$*. The pair $(\mathcal{A}, \mathcal{F})$ is said to be $\mathcal{R}$-compatible if $\mathcal{F}$ realizes $\mathcal{M}$ and we have*

$$\mathit{Regret}_t(\mathcal{A}; \mathcal{M}) \leq \mathcal{R}(d_{\mathcal{F}}, H, \log(T/\delta)) \cdot \sqrt{t}.$$

for all t with probability at least $1-\delta$. $\mathcal{R}$ is called a nominal regret coefficient[2] *for $(\mathcal{A}, \mathcal{F})$.*

The rationale behind $\mathcal{R}$-compatible algorithms is the following. For any $(\mathcal{A}, \mathcal{F})$, we may have a regret coefficient $\mathcal{R}$ in mind (from a provable guarantee) that holds if $\mathcal{F}$ realizes $\mathcal{M}$. The regret $\mathcal{R} \cdot \sqrt{t}$ reflects what we hope to achieve if $\mathcal{F}$ does actually realize $\mathcal{M}$, and $(\mathcal{A}, \mathcal{F})$ is only defined to be compatible if this happens. We remark that realizability is not necessary for a sublinear regret guarantee to hold, but most RL algorithms using function approximation assume it holds, so it is convenient to view both conditions together.

Note that Definition 1 requires that $\mathcal{A}$ be anytime, meaning the bound holds at any arbitrary round index $t \in [T]$ even though only the maximal number of rounds, T, may be specified. For algorithms without automatic anytime guarantees, this can be remedied up to constant factors via the doubling trick (Cesa-Bianchi and Lugosi, 2006). We will later give examples of how our model selection algorithm can be used with some recent single task RL algorithms with formal bounds in the function approximation setting.

Problem Statement Here, our goal in model selection is to obtain a regret guarantee that adapts on-the-fly to the model class of minimal complexity that remains competitive with the optimal value. That is, we wish to find the combination of algorithm $\mathcal{A}$ and model class $\mathcal{F}$, that is compatible in the sense of Definition 1, with the smallest possible leading coefficient $\mathcal{R}(d_{\mathcal{F}}, \cdot, \cdot)$. We consider a setting where we are choosing among a set of candidate algorithms $\mathcal{A}_1, \mathcal{A}_2, \ldots \mathcal{A}_L$ with model classes $\{\mathcal{F}_i\}_{i\in[L]}$, *known* nominal regret coefficients $\{\mathcal{R}_i\}_{i\in[L]}$, and complexities $\{d_i\}_{i\in[L]}$ where $d_i := d_{\mathcal{F}_i}$ and $\mathcal{F}_i$ is the model class of $\mathcal{A}_i$. Without loss of generality, we assume the algorithm-model class pairs can be ordered by their regret such that we have

$$\mathcal{R}_i(d_i, H, \log(T/\delta)) \leq \mathcal{R}_{i+1}(d_{i+1}, H, \log(T/\delta)) \quad (2)$$

for all $i \in [L-1]$, $T, H \in \mathbb{N}$, and $\delta \in (0,1)$. For example, if $\{\mathcal{A}_i\}$ are all instances of the same algorithm that use as input nested model classes $\{\mathcal{F}_i\}$, then (2) is satisfied by ordering $d_1 \leq \ldots \leq d_L$. This naturally captures, among other cases, linear models with nested features (Foster et al., 2019). We also assume[3] that at least one algorithm is $\mathcal{R}_i$-compatible for its respective regret coefficient $\mathcal{R}_i$. Define $i_* = \min\{i \in [L] : (\mathcal{A}_i, \mathcal{F}_i) \text{ is } \mathcal{R}_i\text{-compatible}\}$.

We aim to design a meta-algorithm $\mathcal{A}$ that selects among $\{\mathcal{A}_i\}_{i=1}^L$ without knowing i_* *a priori* and, for some $\alpha \geq 0$ and $\beta \in [1/2, 1)$, achieves a guarantee of

$$\mathrm{Regret}_T(\mathcal{A}) = O\left(\mathcal{R}_{i_*}(d_{i_*}, H, \log(T/\delta)) \cdot L^\alpha T^\beta\right).$$

4 MODEL SELECTION APPROACH

In this section, we present our model selection meta-algorithm, Explore-Commit-Eliminate (ECE) and detail the simple statistical test underlying our approach.

4.1 Algorithm

Our meta-algorithm for model selection is described in Algorithm 1. At a high level, the algorithm proceeds in the following way. It takes as input the base algorithms and model classes, their nominal regret coefficients, and their model complexities; mathematically, the input is given by $\{\mathcal{A}_i, \mathcal{F}_i, \mathcal{R}_i, d_i\}_{i\in[L]}$. The number of algorithms L, episodes $T \in \mathbb{N}$ and failure probability $\delta' \in (0, 1/e)$ are also specified. First, we set $\delta = \frac{\delta'}{10LT^2 \log_2 T}$. The meta-algorithm tracks a candidate algorithm index $\hat{\imath}_t$, corresponding to pair $(\mathcal{A}_{\hat{\imath}_t}, \mathcal{F}_{\hat{\imath}_t})$ that is believed to be $\mathcal{R}_{\hat{\imath}_t}$-compatible at any given time — as well as a set B_t of indices of algorithms with more complex models. At the start of each episode, the meta-algorithm determines whether to use the algorithm $\mathcal{A}_{\hat{\imath}_t}$ or explore using a randomly selected algorithm from the indices B_t, based on the outcome of a Bernoulli variable U_t with success probability $1/t^\kappa$ where $\kappa \in (0, 1/2]$. This random variable U_t represents an indicator that model exploration will occur. After executing the policy from the chosen algorithm, the data is fed back to the algorithm to update, and a test is run to determine whether the algorithm

[2] It is not necessary that $\mathcal{R}$ depend only on these arguments; but these arguments are typically of interest in RL regret bounds.

[3] Note that for all other misspecified algorithms, their nominal regret bounds will, in general, not hold. As regret is being measured with respect to V^*, it will include the misspecification error terms.

should reject $\mathcal{A}_{\hat{\imath}_t}$. The test checks the following condition for each $j \in B_t$:

$$\mathcal{G}_t(\hat{\imath}_t, j) > \mathcal{W}(|\mathcal{T}_t^{\hat{\imath}_t}|, \mathcal{R}_{\hat{\imath}_t}, d_{\hat{\imath}_t}, \delta)$$

where for all $i < j \in [L]$, $t \in [T]$, $\mathcal{T}_t^i$ is the set of times when $\mathcal{A}_i$ is chosen up to t, and $\mathcal{G}$ is a scaled estimate of the excess gap between models i and j, given by

$$\mathcal{G}_t(i, j) := \frac{|\mathcal{T}_t^i|}{|\mathcal{T}_t^j|} \sum_{t' \in \mathcal{T}_t^j} g_{t'} - \sum_{t' \in \mathcal{T}_t^i} g_{t'}$$

and $\mathcal{W}$ is defined as

$$\begin{aligned} \mathcal{W}(t, \mathcal{R}, d, \delta) := {} & C_{\mathcal{W}} \cdot \mathcal{R}(d, H, \log(T/\delta)) \cdot \sqrt{t} \\ & + C_{\mathcal{W}} \cdot H \sqrt{L t^{1+\kappa} \cdot \log(1/\delta)} \\ & + C_{\mathcal{W}} \cdot H \sqrt{t \cdot \log(1/\delta)} \end{aligned}$$

for a sufficiently large constant $C_{\mathcal{W}} > 0$. The test is only valid after a minimal burn-in period, $t \geq \tau_{\min}(\delta) = C_{\min} \cdot L^{\frac{2}{1-\kappa}} \log^{\frac{1}{1-\kappa}}(1/\delta)$ for a sufficiently large $C_{\min} > 0$, so this condition is also checked. If these conditions are true for some $j \in B_t$, meaning that the test fails, then ECE rejects $\mathcal{A}_{\hat{\imath}_t}$ and switches to $\mathcal{A}_{\hat{\imath}_t+1}$. This process repeats until episode T.

Note that although the algorithm uniformly explores among the algorithms in B_t, it does not require any explicit uniform or directed exploration within episodes that may be a tougher problem in RL settings than regret-minimization—one can simply run the algorithms as they were prescribed. In fact, we can interpret our meta-algorithm as automatically leveraging the exploration already in-built in the regret-minimizing base algorithms.

4.2 Statistical test on excess gap

The ability of ECE to judiciously accept or reject base algorithms lies in the simple statistical test at the end of each episode. The test can be viewed as a comparison between the scaled expected return obtained by a "higher-order" algorithm, $\mathcal{A}_j$, corresponding to index $j \in B_t$ during exploration rounds; and that of the active candidate algorithm $\mathcal{A}_{\hat{\imath}_t}$ during all rounds of its usage. If we find that the return of $\mathcal{A}_j$ is significantly higher than that of $\mathcal{A}_{\hat{\imath}_t}$, it suggests that switching to the more complex algorithm $\mathcal{A}_j$ would yield significantly higher return, despite the fact that $\mathcal{A}_j$ has a larger nominal regret bound and might have received much less data than $\mathcal{A}_{\hat{\imath}_t}$ (as it is also competing for data with the other algorithms in B_t). The requirement that $t \geq \tau_{\min}(\delta)$ and our special choice of exploration schedule ensures that the algorithms in B_t will have sufficient data to be useful in the test with high probability, while still exploiting the candidate model $\mathcal{A}_{\hat{\imath}_t}$ whenever possible.

Algorithm 1 Explore-Commit-Eliminate (ECE)

```
**Input**: ${\mathcal{A}_i, \mathcal{F}_i, \mathcal{R}_i, d_i}_{i \in [L]}, L, T, \delta', \tau_{\min}(\cdot)$
$\delta \leftarrow \frac{\delta'}{10LT^2 \log_2 T}$, $\hat{\imath}_t \leftarrow 1$, $\mathcal{T}_1^i = \emptyset$ for all $i \in [L]$, $B_1 = [2, L]$.
$U_t = \begin{cases} 0 & \text{w.p. } 1 - \frac{1}{t^\kappa} \\ 1 & \text{w.p. } \frac{1}{t^\kappa} \end{cases}$ for all $t \in [T]$.
**for** $t = 1, \ldots, T$ **do**
    Set $j = \begin{cases} \hat{\imath}_t & U_t = 0 \\ J_t \sim \text{Unif}\{B_t\} & U_t = 1 \end{cases}$
    $\mathcal{T}_t^j \leftarrow \mathcal{T}_t^j \cup \{t\}$ and $\mathcal{T}_t^k \leftarrow \mathcal{T}_t^k$ for all $k \neq j$.
    Rollout policy $\pi_t$ from $\mathcal{A}_j$
    Observe $z_t := (s_{t,1}, u_{t,1}, \ldots, u_{t,H}, s_{t,H+1})$ and $g_t := \sum_{h \in [H]} r_{t,h}$
    Update $\mathcal{A}_j$ with $t, z_t, g_t$
    **if** $t \geq \tau_{\min}(\delta)$ and there exists $j \in B_t$ such that $\mathcal{G}_t(\hat{\imath}_t, j) > \mathcal{W}(|\mathcal{T}_t^{\hat{\imath}_t}|, \mathcal{R}_{\hat{\imath}_t}, d_{\hat{\imath}_t}, \delta)$ **then**
        $\hat{\imath}_{t+1} \leftarrow \hat{\imath}_t + 1$
        $B_{t+1} \leftarrow B_t \setminus \{\hat{\imath}_{t+1}\}$
        If $\hat{\imath}_{t+1} = L$, break and run $\mathcal{A}_L$ to end of time
    **else**
        $B_{t+1} \leftarrow B_t$
    **end if**
**end for**
```

While we want ECE to reject lower-order models when they perform poorly, the test cannot be too sensitive. Otherwise, it could reject the optimal i_* and choose some unnecessarily large $j > i_*$, leading to highly suboptimal model complexity dependence in the regret bound. Our statistical test is designed to avoid this situation, as we prove in Section 5.

To give some additional intuition behind the test, it is useful to view the expected returns $\frac{1}{|\mathcal{T}_t^j|} \sum_{s \in \mathcal{T}_t^j} g_s$ as a noisy *lower bound* of the optimal value V^*; meanwhile the expected returns of $\frac{1}{|\mathcal{T}_t^{i_*}|} \sum_{s \in \mathcal{T}_t^{i_*}} g_s$ plus the regret incurred, Regret($\mathcal{A}_{i_*}$), should be an *upper bound* of the optimal value V^* up to some noise as well, if $(\mathcal{A}_{i_*}, \mathcal{F}_{i_*})$ is $\mathcal{R}_{i_*}$-compatible. Thus, as long as these intervals intersect, the test should succeed and i_* continues to be accepted. If the intervals separate, the current candidate is rejected. This intuition is reflected in the three terms comprising the definition of $\mathcal{W}$. The first is the nominal regret one expects to see from $\mathcal{A}_{\hat{\imath}_t}$ if it is compatible. The last two follow from concentration of the averaging over returns of the algorithms.

5 MAIN RESULT

Our main result shows that the meta-algorithm automatically adapts to the regret of the optimal pair $(\mathcal{A}_{i_*}, \mathcal{F}_{i_*})$ that is $\mathcal{R}_{i_*}$-compatible. One of the main mechanisms behind this result is ensuring the validity of the test. The following lemma shows that ECE will never reject $(\mathcal{A}_{i_*}, \mathcal{F}_{i_*})$ with high probability.

Lemma 1. *We have $\mathcal{G}_t(i_*, j) \leq \mathcal{W}(|\mathcal{T}_t^{i_*}|, \mathcal{R}_{i_*}, d_{i_*}, \delta)$ with probability at least $1-\delta'$ for all $j \in [i_*+1, L]$ and $t \geq \tau_{\min}(\delta'/10LT^2 \log_2 T)$.*

Thus, since the meta-algorithm steps through the base-algorithms incrementally, Lemma 1 shows that once it reaches $(\mathcal{A}_{i_*}, \mathcal{F}_{i_*})$, the first $\mathcal{R}_{i_*}$-compatible pair, an algorithm with a more complex model class will never be selected. Our main theorem combines this result with the fact that, if the ECE has not rejected a misspecified algorithm $(\mathcal{A}_j, \mathcal{F}_j)$ with $j < i_*$, then the suboptimality of $\mathcal{A}_j$ must not be significant.

Theorem 1. *Let the model exploration parameter $\kappa = 1/3$. Then, the model selection algorithm ECE satisfies the regret bound*

$$\widetilde{O}\left(HLT^{2/3} + \mathcal{R}_{i_*}(d_{i_*}, H, \log(LT/\delta')) \cdot i_*^{1/3} L^{1/2} T^{2/3}\right).$$

with probability at least $1-\delta'$, where $\widetilde{O}$ hides logs and terms independent of T and $\mathcal{R}$.

The regret bound of the meta-algorithm matches that of the optimal algorithm in dependence on the complexity of its model class d_{i_*} and horizon H, i.e., the best dependence if the optimal algorithm were provided *a priori*. We do incur a worse dependence on T, which is now $T^{2/3}$, compared to the nominal $\sqrt{T}$ rate, and a dependence of $L^{1/2}$, the total number of algorithms, and i_*, the index of the optimal algorithm. Note that this type of trade-off in the parameter optimality for model selection is typical in recent results focused on contextual bandits, where methods making less strong assumptions typically incur sub-optimality in either the dependence on d_{i_*} or T. In particular, Theorem 1 matches the rate of Exp3.P (Pacchiano et al., 2020) and does so without non-trivially modifying the base algorithms. In addition to the minimax guarantee of Theorem 1, we show in Section 6 that this can be improved to instance-dependent bounds, in contrast to Exp3.P and Corral.

5.1 Proofs

All proofs of Theorem 1, when not provided here, are available in Appendix A. Due to space limitations, in this section, we prove Lemma 1 and provide a proof sketch for Theorem 1 to illustrate the main idea behind handling pairs $(\mathcal{A}_j, \mathcal{F}_j)$ that are not $\mathcal{R}_j$-compatible. In both cases, we require that three events hold and will show that they do with high probability. Define $\epsilon_t = g_t - V^{\pi_t}$ and let τ_i denote the first episode in which $\mathcal{A}_i$ is chosen as the candidate $\hat{\imath}_t$. If $\mathcal{A}_i$ is never chosen then default to $\tau_i = T$. Recall that $\delta = \frac{\delta'}{10LT^2 \log_2 T}$.

Alg.	Env.	Regret
ModCB	CB	$\widetilde{O}\left(L^{2/3} d_{i_*}^{1/3} T^{2/3}\right)$
OSOM	CB	$\widetilde{O}\left(d_{i_*}^{1/2} T^{1/2}\right)$
Corral	RL	$\widetilde{O}\left(L^{1/2} \mathcal{R}_{i_*}^2 T^{1/2}\right)$
Exp3.P	RL	$\widetilde{O}\left(L^{1/3} \mathcal{R}_{i_*} T^{2/3}\right)$
Ours	RL	MM: $\widetilde{O}\left(L^{5/6} \mathcal{R}_{i_*} T^{2/3}\right)$ ID: $\widetilde{O}\left(\frac{L^{5/2}\mathcal{R}_{i_*}^3}{\Delta_{\min}^2} + \mathcal{R}_{i_*} T^{1/2} + LT^{2/3}\right)$

Table 1: We compare the theoretical guarantees of our algorithm to recent model selection work: ModCB (Foster et al., 2019), OSOM (Chatterji et al., 2020), Corral (Agarwal et al., 2017; Pacchiano et al., 2020), and Exp3.P (Pacchiano et al., 2020). The first two apply to the contextual bandit (CB) setting and leverage distribution assumptions on the contexts to get nearly optimal regret. Corral and Exp3.P apply generally, but are suboptimal and require modifying the base algorithms in non-trivial ways. Our rate matches that of Exp3.P in the minimax (MM) setting without significant assumptions or modifications to the algorithms. We also achieve an improved instance-dependent (ID) rate when the gaps in performance between base algorithms are constant with minimal gap $\Delta_{\min}$.

1. Event E_1: For all $j \in [L]$ and all $t \in [T]$ such that $t \geq \tau_{\min}(\delta)$, if $t \leq \tau_i$, then $\frac{t^{1-\kappa}}{8L} \leq |\mathcal{T}_t^i| \leq 4t^{1-\kappa}$. If $t > \tau_i$, then $|\mathcal{T}_t^i| \leq t - \tau_i + 4t^{1-\kappa}$

2. Event E_2: For all $t \in [T]$,

   $$\textstyle\sum_{t' \in \mathcal{T}_t^{i_*}} V^* - V^{\pi_{t'}} \leq \mathcal{R}_{i_*}(d_{i_*}, H, \log(T/\delta))\sqrt{|\mathcal{T}_t^{i_*}|}$$

3. Event E_3: For all $j \in [L]$ and all $t \in [T]$, $|\sum_{t' \in \mathcal{T}_t^j} \epsilon_{t'}| \leq H\sqrt{2|\mathcal{T}_t^j| \log(2/\delta)}$

The first event ensures that the exploration schedule yields sufficient data to all the algorithms before they are chosen. The second states that the nominal anytime regret guarantee holds for $(\mathcal{A}_{i_*}, \mathcal{F}_{i_*})$. The third handles concentration of the noisy returns that the algorithm observes from deploying policies. The following lemma shows that all three events happen with high probability.

Lemma 2. *The event $E = \bigcap_{i \in \{1,2,3\}} E_i$ holds with probability at least $1 - 10LT^2\delta \log_2 T$.*

Lemma 2 is proved in Appendix A.1. The proof for the first event uses a Freedman inequality (details in Appendix B) to bound the sizes of all sets given that enough time has passed. The second event holds with high probability under the assumption that $(\mathcal{A}_{i_*}, \mathcal{F}_{i_*})$ is $\mathcal{R}_{i_*}$-compatible. The third event can be shown to hold with high probability using the Azuma-Hoeffding inequality with appropriate union bounds.

5.1.1 Proof of Lemma 1

We now prove the statement of Lemma 1 under the event E. Adding and subtracting the sum of appropriately scaled value functions $\sum_{t' \in \mathcal{T}_t^j} V^{\pi_{t'}}$ and $\sum_{t' \in \mathcal{T}_t^{i_*}} V^{\pi_{t'}}$, we can write $\mathcal{G}_t(i_*, j)$ in terms of value functions and conditionally zero-mean errors:

$$\begin{aligned}\mathcal{G}_t(i_*, j) &= \frac{|\mathcal{T}_t^{i_*}|}{|\mathcal{T}_t^j|} \sum_{t' \in \mathcal{T}_t^j} g_{t'} - \sum_{t' \in \mathcal{T}_t^{i_*}} g_{t'} \\ &= \frac{|\mathcal{T}_t^{i_*}|}{|\mathcal{T}_t^j|} \sum_{t' \in \mathcal{T}_t^j} (V^{\pi_{t'}} + \epsilon_{t'}) - \sum_{t' \in \mathcal{T}_t^{i_*}} (V^{\pi_{t'}} + \epsilon_{t'}) \\ &\leq \sum_{t' \in \mathcal{T}_t^{i_*}} (V^* - V^{\pi_{t'}}) + \frac{|\mathcal{T}_t^{i_*}|}{|\mathcal{T}_t^j|} \sum_{t' \in \mathcal{T}_t^j} \epsilon_{t'} - \sum_{t' \in \mathcal{T}_t^{i_*}} \epsilon_{t'}\end{aligned}$$

The last inequality follows as $V^* \geq V^{\pi_{t'}}$ for all $t' \in [T]$. If events E_2 and E_3 hold then

$$\begin{aligned}\mathcal{G}_t(i_*, j) \leq{}& \mathcal{R}_{i_*}\left(d_{i_*}, H, \log(1/\delta)\right) \cdot \sqrt{|\mathcal{T}_t^{i_*}|} \\ &+ H\sqrt{2|\mathcal{T}_t^{i_*}| \log(2/\delta)} + H\sqrt{\frac{2|\mathcal{T}_t^{i_*}|^2}{|\mathcal{T}_t^j|} \log(2/\delta)}\end{aligned}$$

By event E_1 and the fact that $j > i_*$ and $t \geq \tau_{\min}(\delta)$, $|\mathcal{T}_t^j| \geq \frac{t^{1-\kappa}}{8L} \geq \frac{|\mathcal{T}_t^{i_*}|^{1-\kappa}}{8L}$. Therefore, for the third term,

$$H\sqrt{\frac{2|\mathcal{T}_t^{i_*}|^2}{|\mathcal{T}_t^j|} \log(2/\delta)} \leq H\sqrt{16L|\mathcal{T}_t^{i_*}|^{1+\kappa} \log(2/\delta)}$$

Applying this bound to the result in the previous display and given the definition of $\mathcal{W}$, it follows that $\mathcal{G}_t(i_*, j) \leq \mathcal{W}(|\mathcal{T}_t^{i_*}|, \mathcal{R}_{i_*}, d_{i_*}, \delta)$ for a sufficiently large constant $C_{\mathcal{W}} > 0$, independent of t, d_{i_*}, H, and δ.

5.1.2 Proof Sketch of Theorem 1

In bounding the regret of the meta-algorithm, there are three cases to handle: (1) before the test becomes valid, (2) once the test is valid but i_* has not been chosen yet, and finally (3) once i_* is chosen. We address the first and third cases before addressing the second, which is more involved. We define $\tau_* = \tau_{i_*}$ for shorthand.

Case (1): When $t < \tau_{\min}(\delta)$, the test to determine switching among any of the model classes is not yet valid. Here we simply pay the burn-in period giving $\text{Regret}_{1:\tau_{\min}(\delta)-1} \leq O(HL^{\frac{2}{1-\kappa}} \log^{\frac{1}{1-\kappa}}(1/\delta))$.

Case (3): If $t > \tau_*$, then the meta-algorithm has switched to $\mathcal{A}_{i_*}$. Under event E, the condition in Lemma 1 is met and so the test no longer fails. Therefore $(\mathcal{A}_{i_*}, \mathcal{F}_{i_*})$ which is $\mathcal{R}_{i_*}$-compatible is not rejected in the remaining episodes. The regret during this phase scales as $\mathcal{R}_{i_*}(d_{i_*}, H, \log(T/\delta)) \cdot \sqrt{T}$ plus additional $O(HLT^{1-\kappa})$ regret due to exploration of the remaining base algorithms in B_t.

Case (2) is when $\tau_{\min} < t \leq \tau_*$—the test is eligible but the meta-algorithm is either switching among misspecified models or unable to detect that they are misspecified. Since the misspecification is not detected for any of the algorithms in B_t, we know $\mathcal{G}_t(\hat{\imath}_t, i_*) \leq \mathcal{W}(|\mathcal{T}_t^{\hat{\imath}_t}|, \mathcal{R}_{\hat{\imath}_t}, d_{\hat{\imath}_t}, \delta)$. That is, the average reward for $\mathcal{A}_{\hat{\imath}_t}$ is not significantly different from that of $\mathcal{A}_{i_*}$. Since $\mathcal{A}_{i_*}$ is only played during exploration and $t \geq \tau_{\min}(\delta)$, its number of rounds played can be lower bounded by $t^{1-\kappa}/8L$ and thus its average regret is at most roughly

$$\widetilde{O}\left(\frac{L^{1/2}\mathcal{R}_{i_*}(d_{i_*}, H, \log(T/\delta))}{t^{\frac{1-\kappa}{2}}}\right).$$

The success of the test suggests that the average reward of $\mathcal{A}_{\hat{\imath}_t}$ should be close to this. Extrapolating over the rounds played by $\mathcal{A}_{\hat{\imath}_t}$, the regret for $\hat{\imath}_t$ will be

$$\widetilde{O}\left(\mathcal{R}_{i_*}(d_{i_*}, H, \log(T/\delta)) \cdot L^{1/2}|\mathcal{T}_t^{\hat{\imath}_t}|^{\frac{1+\kappa}{2}}\right)$$

up to a constant shift by $\mathcal{W}(|\mathcal{T}_t^{\hat{\imath}}|, \mathcal{R}_{\hat{\imath}_t}, d_{\hat{\imath}_t}, \delta)$. The shift is dominated by the above display because $\mathcal{R}_{\hat{\imath}_t} \leq \mathcal{R}_{i_*}$ and $\kappa \in (0, 1/2]$. Finally, since we must account for the cumulative effect for all $i < i_*$, Jensen's inequality shows the sum of these terms is bounded above by

$$\widetilde{O}\left(\mathcal{R}_{i_*}(d_{i_*}, H, \log(T/\delta)) \cdot i_*^{\frac{1-\kappa}{2}} L^{1/2} T^{\frac{1+\kappa}{2}}\right).$$

This becomes the dominant term in the regret. Additional regret of $O(HLT^{1-\kappa} + Hi_* + HT^{\frac{1+\kappa}{2}} \log^{1/2}(1/\delta))$ is also paid for exploration, switching costs, and estimation error of the averages. Summing these three cases and taking $\kappa = 1/3$ proves Theorem 1.

5.2 Applications

Though Theorem 1 is stated generally for any RL algorithms with nominal anytime regret bounds, we can easily specialize it to several important problem settings without knowing the optimal model class *a priori*. Formal details can be found in Appendix C.

Linear Models Recent work has considered linear MDPs where the transition dynamics and reward are linear in some feature vector (Jin et al., 2020b; Yang and Wang, 2020). We assume access to nested features $\phi_i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{d_i}$ for $i \in [L]$ such that $d_i \leq d_{i+1}$ and the first d_i components of ϕ_{i+1} are the same as ϕ_i. These feature generate linear model classes:

$$\mathcal{F}_i = \left\{(s, a) \mapsto \langle \phi_i(s, a), \theta \rangle \ : \ \theta \in \mathbb{R}^{d_i}\right\} \qquad (3)$$

$\mathcal{F}_i$ realizes $\mathcal{M}$ if it has zero approximation error for the transition dynamics $P(\cdot|s,a)$ and reward $r(s,a)$. Let i_* be the smallest index such that $\mathcal{F}_{i_*}$ realizes $\mathcal{M}$. The regret of LSVI-UCB (Jin et al., 2020b) under $\mathcal{F}_i$ for $i \geq i_*$ is $\widetilde{O}\left(\sqrt{d_i^3 H^4 T}\right)$. Using ECE with LSVI-UCB algorithms guarantees $\text{Regret}_T(\mathsf{ECE}) = \widetilde{O}\left(\sqrt{d_{i_*}^3 H^4} \cdot L^{5/6} T^{2/3}\right)$. MatrixRL (Yang and Wang, 2020) similarly assumes a linear function class:

$$\mathcal{F}_i = \left\{(s,u,s') \mapsto \phi_i(s,u)^\top M \psi_i(s') \ : \ M \in \mathbb{R}^{d_i \times d_i'}\right\}$$

for $\psi_i(s') \in \mathbb{R}^{d_i'}$. For $\mathcal{F}_i$ with $i \geq i_*$ that realizes the transition dynamics P, the regret is $\widetilde{O}\left(\sqrt{d_i^3 H^5 T}\right)$. Our model selection algorithm, ECE, achieves $\text{Regret}_T(\mathsf{ECE}) = \widetilde{O}\left(\sqrt{d_{i_*}^3 H^5} \cdot L^{5/6} T^{2/3}\right)$.

A more general linear setting considers learning under low Bellman error without directly assuming linearity of P. With $\mathcal{F}_i$ defined as in (3), we say $\mathcal{F}_i$ realizes $\mathcal{M}$ if it has zero inherent Bellman error (Definition 1, Zanette et al. (2020)). Then for $i \geq i_*$, ELEANOR (Zanette et al., 2020) guarantees improved regret $\widetilde{O}\left(d_i \sqrt{H^4 T}\right)$ and ECE achieves $\text{Regret}_T(\mathsf{ECE}) = \widetilde{O}\left(d_{i_*} \sqrt{H^4} \cdot L^{5/6} T^{2/3}\right)$.

As done by Foster et al. (2019), for nested model classes, the L dependence can be replaced by $\log T$ by only considering a subset of features such that $d_i = O(2^i)$ for $i \in [\lceil \log_2(T) \rceil]$.

Low Bellman Rank For more general function approximation, consider the setting of MDPs with low Bellman rank (Jiang et al., 2017) and finite (but not necessarily linear or nested) models $\{\mathcal{F}_i\}$ with $\mathcal{F}_i : \mathcal{S} \times \mathcal{U} \to \mathbb{R}$. $\mathcal{F}_i$ realizes $\mathcal{M}$ if there is $f^* \in \mathcal{F}_i$ such that $f^* = Q_h^*$ for all $h \in [H]$ and the induced Bellman rank is $M_i < |\mathcal{F}_i|$. For i such that $\mathcal{F}_i$ realizes $\mathcal{M}$, AVE (Dong et al., 2020) guarantees regret $\widetilde{O}\left(\sqrt{M_i^2 |\mathcal{U}| H^4 T \log^3 |\mathcal{F}_i|}\right)$. Let i_* be defined similarly as before. Then ECE achieves $\text{Regret}_T(\mathsf{ECE}) = \widetilde{O}\left(\sqrt{M_{i_*}^2 |\mathcal{U}| H^4 \log^3 |\mathcal{F}_{i_*}|} \cdot L^{5/6} T^{2/3}\right)$

6 INSTANCE-DEPENDENT BOUNDS

We now prove a stronger "instance-dependent" guarantee on online selection over more specialized base algorithms which have provable regret guarantees that are sublinear in T, but compared to the best policy within its respective policy class. For example, for an algorithm and model class $(\mathcal{A}, \mathcal{F})$ using value-based function approximation we might consider the greedy policy class:

$$\Pi_{\mathcal{F}} = \left\{(s,h) \mapsto \arg\max_{u \in \mathcal{U}} \ f(s,u,h) \ : \ f \in \mathcal{F}\right\}.$$

The regret with respect to the best-in-class is

$$\text{Regret}_T(\mathcal{A}, \Pi_{\mathcal{F}}; \mathcal{M}) = \max_{\pi \in \Pi_{\mathcal{F}}} \textstyle\sum_{t \in [T]} V^\pi - V^{\pi_t}$$

To consider algorithms that may obtain sublinear regret with respect to this weaker benchmark but not with respect to V^*, we give a refined definition of $\mathcal{R}$-compatible algorithms.

Definition 2. *The pair $(\mathcal{A}, \mathcal{F})$ is said to be $\mathcal{R}^{\Pi_{\mathcal{F}}}$-compatible with respect to $\Pi_{\mathcal{F}}$ on the MDP $\mathcal{M}$ if we have*

$$\textit{Regret}_T(\mathcal{A}, \Pi_{\mathcal{F}}; \mathcal{M}) \leq \mathcal{R}^{\Pi_{\mathcal{F}}}(d_{\mathcal{F}}, H, \log(T/\delta)) \cdot \sqrt{t}$$

for all t with probability at least $1 - \delta$.

The value of $\max_{\pi \in \Pi_{\mathcal{F}}} V^\pi$ is typically unknown because of the complex dependence between $\Pi_{\mathcal{F}}$ and $\mathcal{M}$, and because $\Pi_{\mathcal{F}}$ is often determined by $\mathcal{F}$. Given a set of algorithms with different policy classes, we would like to select the one with the smallest regret compared to the optimal best-in-class value. Formally, we assume there are given algorithms $\{(\mathcal{A}_i, \mathcal{F}_i)\}$ with policy classes $\{\Pi_i\}$ each having optimal values $V_i^* := \max_{\pi \in \Pi_i} V^\pi$ and regret coefficients $\{\mathcal{R}_i^{\Pi_i}\}$ such that *for all* i the pair $(\mathcal{A}_i, \mathcal{F}_i)$ is $\mathcal{R}_i^{\Pi_i}$-compatible and $\mathcal{R}_i(d_i, \cdot, \cdot) \leq \mathcal{R}_{i+1}(d_{i+1}, \cdot, \cdot)$. Our goal is to select $i_* \in B_* := \arg\max_{j \in [L]} V_j^*$ that has the smallest complexity dependence i.e. $i_* = \arg\min_{i \in B_*} \mathcal{R}_i^{\Pi_i}(d_i, \cdot, \cdot)$. We emphasize that even if no algorithm is compatible in the sense of Definition 1, we want the optimal best-in-class guarantee[4] in the sense of Definition 2.

The difference between this setting and the last is that all algorithms are assumed to be compatible with respect to their own policy classes now, but the differing Π_i mean that some can have lower V_i^*, which we want to eliminate. Note that although the regret coefficients are ordered as in (2), the values $\{V_i^*\}$ are unknown and not necessarily ordered. Observe that $i_* = \min B_*$, so that $V_{i_*}^* > V_i^*$ for all $i < i_*$ and $V_{i_*}^* \geq V_i^*$ for all $i > i_*$. Thus i_* has the lowest regret for the best policy class. We would like an algorithm $\mathcal{A}$ that bounds $\text{Regret}_T(\mathcal{A}, \Pi_{i_*}; \mathcal{M})$ with dependence on only the complexity of $\mathcal{F}_{i_*}$. The following result shows that Algorithm 1, without any modifications, can obtain an *instance-dependent* regret guarantee based on the size of the gaps $\Delta_{j,i_*} := V_{i_*}^* - V_j^*$ for $j < i_*$.

[4] In essence, the best-in-class guarantee needs to hold even under model misspecification. A good example of a base algorithm satisfying this condition would be Exp4 in the contextual bandits setting.

Theorem 2. *For a given $\mathcal{M}$, let $(\mathcal{A}_i, \mathcal{F}_i)$ be $\mathcal{R}_i^{\Pi_i}$-compatible with respect to Π_i for all $i \in [L]$. Then, with probability at least $1-\delta'$, ECE with $\kappa = 1/3$ satisfies the regret bound with respect to policy class Π_{i_*}:*

$$\widetilde{O}\left(HLT^{2/3} + \mathcal{R}_{i_*}^{\Pi_{i_*}}\sqrt{T} + L^{3/2}(\mathcal{R}_{i_*}^{\Pi_{i_*}})^3 \textstyle\sum_{i<i_*} \Delta_{i,i_*}^{-2}\right)$$

If $\kappa = 1/2$, then it satisfies

$$\widetilde{O}\left(HL\sqrt{T} + \mathcal{R}_{i_*}^{\Pi_{i_*}}\sqrt{T} + L^2(\mathcal{R}_{i_*}^{\Pi_{i_*}})^4 \textstyle\sum_{i<i_*} \Delta_{i,i_*}^{-3}\right)$$

Comparing this result to Theorem 1, if ECE is run with the same $\kappa = 1/3$ and the gaps are constant, a significantly better rate is possible since the third term has no dependence on T. With a more aggressive exploration choice of $\kappa = 1/2$, an even stronger instance-dependent guarantee is possible, matching the optimal $\mathcal{R}_{i_*}^{\Pi_{i_*}}\sqrt{T}$ rate of the best algorithm. However, this comes at the price of worse dependence on the gaps and $\mathcal{R}_{i_*}^{\Pi_{i_*}}$ factors, in the term that does not increase polynomially with T. In either case, Theorem 2 shows that we can obtain optimal or near-optimal dependence in T and only suboptimal $\mathcal{R}_{i_*}^{\Pi_{i_*}}$-dependence on terms that do not grow with T, as long as the gaps are constant. In Appendix D, we show that these rates can be even further improved with only minimal modifications to ECE if given access to fast estimators of the gaps or V^*.

7 CONCLUSION

We present a new model selection meta-algorithm for RL with function approximation. Given a set of base algorithms in which one is well-specified, the meta-algorithm adapts to the regret of the optimal one using a simple and interpretable statistical test. The regret of the meta-algorithm retains optimal dependence on model complexity while increasing the dependence on the number of episodes, T, to $O(T^{2/3})$. Compared to past efforts, our meta-algorithm provides similarly strong worst-case regret bounds, is computationally efficient conditioned on efficiency of the base algorithms, works under minimal assumptions, and provides new instance-dependent results.

Of many interesting future directions, a particularly interesting one given the prior significance of access to V^* (Foster et al., 2019; Modi et al., 2020) and our even stronger instance-dependent regret rates (see Appendix D), is whether estimating V^* is easier than estimating the optimal policy.

Acknowledgements

This work is supported by an Office of Naval Research Young Investigator Program Award, the Stanford AI Lab, and a gift from Toyota. JNL is supported by the NSF GRFP. VM acknowledges support from a Simons-Berkeley Research Fellowship. Part of this work was done while some of the authors were visiting the Simons Institute for the Theory of Computing.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

Yasin Abbasi-Yadkori, Aldo Pacchiano, and My Phan. Regret balancing for bandit and rl model selection. *arXiv preprint arXiv:2006.05491*, 2020.

Alekh Agarwal, Haipeng Luo, Behnam Neyshabur, and Robert E Schapire. Corralling a band of bandit algorithms. In *Conference on Learning Theory*, pages 12–38. PMLR, 2017.

Alekh Agarwal, Nan Jiang, and Sham M Kakade. Reinforcement learning: Theory and algorithms. Technical report, Technical Report, Department of Computer Science, University of Washington, 2019.

Raman Arora, Teodor V Marinov, and Mehryar Mohri. Corralling stochastic bandit algorithms. *arXiv preprint arXiv:2006.09255*, 2020.

Peter L Bartlett, Varsha Dani, Thomas Hayes, Sham Kakade, Alexander Rakhlin, and Ambuj Tewari. High-probability regret bounds for bandit online linear optimization. 2008.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games.* Cambridge university press, 2006.

Niladri Chatterji, Vidya Muthukumar, and Peter Bartlett. Osom: A simultaneously optimal algorithm for multi-armed and linear contextual bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 1844–1854, 2020.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214, 2011.

Lee H Dicker. Variance estimation in high-dimensional linear models. *Biometrika*, 101(2):269–284, 2014.

Kefan Dong, Jian Peng, Yining Wang, and Yuan Zhou. n-regret for learning in markov decision processes with function approximation and low bellman rank. *Proceedings of Machine Learning Research vol*, 125:1–4, 2020.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(Jun): 1079–1105, 2006.

Amir-massoud Farahmand and Csaba Szepesvári. Model selection in reinforcement learning. *Machine learning*, 85(3):299–332, 2011.

Dylan Foster, Akshay Krishnamurthy, and Haipeng Luo. Model selection for contextual bandits. *Advances in neural information processing systems*, 2019.

Zhaohan Guo and Emma Brunskill. Sample efficient learning with feature selection for factored mdps. In *European Workshop on Reinforcement Learning*, 2018.

Assaf Hallak, Dotan Di-Castro, and Shie Mannor. Model selection in markovian processes. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 374–382, 2013.

Nan Jiang, Alex Kulesza, and Satinder Singh. Abstraction selection in model-based reinforcement learning. In *International Conference on Machine Learning*, pages 179–188, 2015.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pages 1704–1713. PMLR, 2017.

Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-free exploration for reinforcement learning. *arXiv preprint arXiv:2002.02794*, 2020a.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143, 2020b.

Weihao Kong and Gregory Valiant. Estimating learnability in the sublinear data regime. In *Advances in Neural Information Processing Systems*, pages 5455–5464, 2018.

Weihao Kong, Emma Brunskill, and Gregory Valiant. Sublinear optimal policy value estimation in contextual bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 4377–4387, 2020.

Aditya Modi, Nan Jiang, Ambuj Tewari, and Satinder Singh. Sample complexity of reinforcement learning using linearly combined model ensembles. In *International Conference on Artificial Intelligence and Statistics*, pages 2010–2020, 2020.

Aldo Pacchiano, My Phan, Yasin Abbasi Yadkori, Anup Rao, Julian Zimmert, Tor Lattimore, and Csaba Szepesvari. Model selection in contextual stochastic bandit problems. *Advances in Neural Information Processing Systems*, 33, 2020.

Ronald Parr, Lihong Li, Gavin Taylor, Christopher Painter-Wakefield, and Michael L Littman. An analysis of linear models, linear value-function approximation, and feature selection for reinforcement learning. In *Proceedings of the 25th international conference on Machine learning*, pages 752–759, 2008.

Nicolas Verzelen, Elisabeth Gassiat, et al. Adaptive estimation of high-dimensional signal-to-noise ratios. *Bernoulli*, 24(4B):3683–3710, 2018.

Ruosong Wang, Simon S Du, Lin F Yang, and Ruslan Salakhutdinov. On reward-free reinforcement learning with linear function approximation. *arXiv preprint arXiv:2006.11274*, 2020.

Lin Yang and Mengdi Wang. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pages 10746–10756. PMLR, 2020.

Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pages 10978–10989. PMLR, 2020.