A OMITTED PROOFS

In this section, we collect proofs for Theorem 1 that were omitted from the main paper.

A.1 Proof of Lemma 2

Here, we restate and prove Lemma 2.

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

Proof. We will show that each of the three events holds with high probability and the apply the union bound.

Corollary 1 of Section B shows event E_1 holds with probability at least $1 - 4LT^2\delta\log_2 T$.

For event E_2, i_* is the index of the algorithm that is $\mathcal{R}_{i_*}$-compatible and anytime. Let $\pi^{i_*}_{(k)}$ denote the policy played by $\mathcal{A}_{i_*}$ at the k^{th} call to i_*. For $K \in [T]$, these properties guarantee its regret bound holds, with probability at least $1-\delta$,

$$\sum_{k\in[K]} V^* - V^{\pi^{i_*}_{(k)}} \le \mathcal{R}_{i_*}(d_{i_*}, H, \log(T/\delta)) \cdot \sqrt{K}$$

Taking the union bound over all $K \in [T]$ shows that event E_2 holds with probability at least $1 - T\delta$.

As in the previous case, we can view the process $\epsilon^i_{(1)}, \ldots, \epsilon^i_{(T)}$ as the pre-drawn differences between the observed and expected returns for the 1 through (at most) T times of playing model $\mathcal{A}_i$. Applying the Azuma-Hoeffding inequality with $|\epsilon^i_{(k)}| \le H$ and taking the union bound over all $K \in [T]$,

$$|\sum_{k\in[K]} \epsilon^i_{(k)}| \le H\sqrt{2K\log(2/\delta)}$$

with probability at least $1 - T\delta$. Taking the union bound over all models, event E_3 occurs with probability at least $1 - LT\delta$.

Taking these events together and $\delta' = 10LT^2\delta\log_2 T$, event E holds with probability at least $1 - \delta'$. □

A.2 Full Proof of Theorem 1

Here, we restate and complete the proof of Theorem 1.

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 $\mathcal{R}$.

Proof. Let $\tau_* := \tau_{i_*}$ denote the time that $\mathcal{A}_{i_*}$ is chosen as the candidate. Recall that $\delta = \frac{\delta'}{10LT^2\log_2 T}$. The analysis can be divided into three phases when conditioned on the event E.

1. $t < \tau_{\min}(\delta)$: the test to determine switching to i_* is not valid yet.
2. $\tau_{\min}(\delta) < t \le \tau_*$: the test is eligible but ECE is still switching among incompatible algorithms.
3. $t > \tau_*$: ECE has switched to $\mathcal{A}_{i_*}$.

Note that it is possible that $\tau_* \ge T$. That is, the algorithm only uses incompatible algorithms; however, we will show that this case still guarantees regret that adapts to the optimal algorithm i_*.

Case 1: Invalid Test We require $t \geq \tau_{\min}(\delta)$ in order for the condition in Lemma 1 to hold under E when $\hat{\imath}_t = i_*$. Therefore, we can view this period $t < \tau_{\min}(\delta)$ as an unavoidable burn-in period. The regret during this interval can then be upper bounded in the worst case as

$$\text{Regret}_{1:\tau_{\min}(\delta)-1} = \sum_{t=1}^{\tau_{\min}-1} V^* - V^{\pi_t} \leq H\tau_{\min} = O\left(HL^{\frac{2}{1-\kappa}} \log^{\frac{1}{1-\kappa}}(1/\delta)\right)$$

Case 2: Misspecified Case In the second phase, the test is valid, but ECE is either utilizing algorithms below i_* or switching among them in the event the test fails. The regret can be decomposed across each set $\mathcal{T}^j_{\tau_*}$ of times playing $\mathcal{A}_j$ up to time τ_*:

$$\begin{aligned}
\text{Regret}_{\tau_{\min}(\delta):\tau_*} &= \sum_{j\in[L]} \sum_{t\in\mathcal{T}^j_{\tau_*}} V^* - V^{\pi_t} \\
&\leq 4H(L-i_*)\tau_*^{1-\kappa} + \sum_{j<i_*} \sum_{t\in\mathcal{T}^j_{\tau_{j+1}}} V^* - V^{\pi_t} \\
&\leq 4H(L-i_*)\tau_*^{1-\kappa} + Hi_* + \sum_{j<i_*} \sum_{t\in\mathcal{T}^j_{\tau_{j+1}-1}} V^* - V^{\pi_t}
\end{aligned}$$

The second line follows from the fact that for $j > i_*$, algorithm j is not selected yet (if ever), so maximal regret is paid for those algorithms during exploration. Event E_1 upper bounds the number of times that can be in $\mathcal{T}^j_{\tau_*}$ at time τ_*, since the regret due to j is only due to exploration. Furthermore, for $j < i_*$, once j is rejected, it is never used for exploration again, so we can replace $\mathcal{T}^j_{\tau_*}$ with $\mathcal{T}^j_{\tau_{j+1}}$ for $j < i_*$. The third line is necessary as no guarantee is given during episodes when a test fails and there can be at most i_* failing tests since the condition in Lemma 1 is always true under event E.

Then, we focus on bounding the right-hand term. Fix $j < i_*$. Observe that for $t \in \mathcal{T}^j_{\tau_{j+1}-1}$ the tests succeed for all comparisons including with i_*:

$$\mathcal{G}_{\tau_{j+1}-1}(j,i) \leq \mathcal{W}(|\mathcal{T}^j_{\tau_{j+1}-1}|, \mathcal{R}_j, d_j, \delta)$$

for all $i > j$. Therefore, since $i_* > j$, the definition of $\mathcal{G}$ can be used the bound the following:

$$\begin{aligned}
\sum_{t\in\mathcal{T}^j_{\tau_{j+1}-1}} V^* - V^{\pi_t} &= \sum_{t\in\mathcal{T}^j_{\tau_{j+1}-1}} (V^* - g_t) + \sum_{t\in\mathcal{T}^j_{\tau_{j+1}-1}} \epsilon_t \\
&\leq \frac{|\mathcal{T}^j_{\tau_{j+1}-1}|}{|\mathcal{T}^{i_*}_{\tau_{j+1}-1}|} \sum_{t\in\mathcal{T}^{i_*}_{\tau_{j+1}-1}} (V^* - g_t) + \mathcal{W}(|\mathcal{T}^j_{\tau_{j+1}-1}|, \mathcal{R}_j, d_j, \delta) + \sum_{t\in\mathcal{T}^j_{\tau_{j+1}-1}} \epsilon_t \\
&\leq \frac{|\mathcal{T}^j_{\tau_{j+1}-1}|}{|\mathcal{T}^{i_*}_{\tau_{j+1}-1}|} \sum_{t\in\mathcal{T}^{i_*}_{\tau_{j+1}-1}} (V^* - V^{\pi_t}) + \mathcal{W}(|\mathcal{T}^j_{\tau_{j+1}-1}|, \mathcal{R}_j, d_j, \delta) \\
&\quad + \sum_{t\in\mathcal{T}^j_{\tau_{j+1}-1}} \epsilon_t + \frac{|\mathcal{T}^j_{\tau_{j+1}-1}|}{|\mathcal{T}^{i_*}_{\tau_{j+1}-1}|} \sum_{t\in\mathcal{T}^{i_*}_{\tau_{j+1}-1}} \epsilon_t
\end{aligned}$$

Now we can use the fact that E_2 and E_3 hold to bound the regret and estimation errors:

$$\begin{aligned}
\sum_{t\in\mathcal{T}^j_{\tau_{j+1}-1}} V^* - V^{\pi_t} &\leq O\left(\mathcal{R}_{i_*}(d_{i_*}, H, \log(T/\delta)) \cdot \sqrt{\frac{|\mathcal{T}^j_{\tau_{j+1}-1}|^2}{|\mathcal{T}^{i_*}_{\tau_{j+1}-1}|}}\right) + \mathcal{W}(|\mathcal{T}^j_{\tau_{j+1}-1}|, \mathcal{R}_j, d_j, \delta) \\
&\quad + O\left(H\sqrt{|\mathcal{T}^j_{\tau_{j+1}-1}| \cdot \log(1/\delta)}\right) + O\left(H\sqrt{\frac{|\mathcal{T}^j_{\tau_{j+1}-1}|^2}{|\mathcal{T}^{i_*}_{\tau_{j+1}-1}|} \cdot \log(1/\delta)}\right)
\end{aligned} \tag{4}$$

Using E_1 and the fact that $\tau_{\min}(\delta) \leq \tau_{j+1} - 1 \leq \tau_*$, we have that

$$|\mathcal{T}^{i_*}_{\tau_{j+1}-1}| \geq \frac{(\tau_{j+1}-1)^{1-\kappa}}{8L} \geq \frac{|\mathcal{T}^{i_*}_{\tau_{j+1}-1}|^{1-\kappa}}{8L}.$$

Then the terms in (4) that contain $|\mathcal{T}^{i_*}_{\tau_{j+1}-1}|$ in the denominator can be upper bounded:

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

$$O\left(H\sqrt{\frac{|\mathcal{T}^{j}_{\tau_{j+1}-1}|^2}{|\mathcal{T}^{j}_{\tau_{j+1}-1}|} \cdot \log(1/\delta)}\right) \leq O\left(HL^{1/2}|\mathcal{T}^{j}_{\tau_{j+1}-1}|^{\frac{1+\kappa}{2}} \cdot \log^{1/2}(1/\delta)\right)$$

The bound then becomes

$$\begin{aligned}\sum_{t \in \mathcal{T}^{j}_{\tau_{j+1}-1}} V^* - V^{\pi_t} \leq{}& O\left(L^{1/2}\mathcal{R}_{i_*}(d_{i_*}, H, \log(T/\delta)) \cdot |\mathcal{T}^{j}_{\tau_{j+1}-1}|^{\frac{1+\kappa}{2}}\right) + \mathcal{W}(|\mathcal{T}^{j}_{\tau_{j+1}-1}|, \mathcal{R}_j, d_j, \delta) \\ &+ O\left(H|\mathcal{T}^{j}_{\tau_{j+1}-1}|^{1/2} \cdot \log^{1/2}(1/\delta)\right) + O\left(HL^{1/2}|\mathcal{T}^{j}_{\tau_{j+1}-1}|^{\frac{1+\kappa}{2}} \cdot \log^{1/2}(1/\delta)\right)\end{aligned}$$

Since $\mathcal{R}_j \leq \mathcal{R}_{i_*}$, the regret for j in this case is

$$\sum_{t \in \mathcal{T}^{j}_{\tau_{j+1}-1}} V^* - V^{\pi_t} \leq O\left(L^{1/2}\mathcal{R}_{i_*}(d_{i_*}, H, \log(T/\delta)) \cdot |\mathcal{T}^{j}_{\tau_{j+1}-1}|^{\frac{1+\kappa}{2}} + HL^{1/2}|\mathcal{T}^{j}_{\tau_{j+1}-1}|^{\frac{1+\kappa}{2}} \cdot \log^{1/2}(1/\delta)\right)$$

Observe that $\sum_{j<i_*} |\mathcal{T}^{j}_{\tau_{j+1}-1}| \leq T$ and the right-hand side is a sum of concave functions of each $|\mathcal{T}^{j}_{\tau_{j+1}-1}|$. Using Jensen's inequality with the uniform distribution over $|\mathcal{T}^{j}_{\tau_{j+1}-1}|$ for $j < i_*$ and then upper bounding by T yields the bound:

$$\text{Regret}_{\tau_{\min}(\delta):\tau_*} \leq O\left(HLT^{1-\kappa} + Hi_* + \left(\mathcal{R}_{i_*}(d_{i_*}, H, \log(T/\delta)) + H\log^{1/2}(1/\delta)\right) \cdot i_*^{\frac{1-\kappa}{2}} L^{1/2} \cdot T^{\frac{1+\kappa}{2}}\right)$$

Case 3: Selecting $\mathcal{A}_{i_*}$ Starting at $\tau_* + 1$, $\mathcal{A}_{i_*}$ is selected. Note that the condition in Lemma 1 holds under event E, so ECE will never reject i_*. Then

$$\begin{aligned}\text{Regret}_{\tau_*+1:T} &\leq \sum_{j \in [i_*+1, L]} H|\mathcal{T}^{j}_{T}| + \sum_{t \in \mathcal{T}^{i_*}_{T}} V^* - V^{\pi_t} \\ &\leq \sum_{j \in [i_*+1, L]} H|\mathcal{T}^{j}_{T}| + O\left(\mathcal{R}_{i_*}(d_{i_*}, H, \log(T/\delta) \cdot \sqrt{T}\right) \\ &\leq O\left(HLT^{1-\kappa} + \mathcal{R}_{i_*}(d_{i_*}, H, \log(T/\delta) \cdot \sqrt{T}\right)\end{aligned}$$

Adding the terms from these three phases gives the final bound:

$$\text{Regret}_T = O\left(HL^{\frac{2}{1-\kappa}} \log^{\frac{1}{1-\kappa}}(1/\delta) + HLT^{1-\kappa} + Hi_* + \left(\mathcal{R}_{i_*}(d_{i_*}, H, \log(T/\delta)) + H\log^{1/2}(1/\delta)\right) \cdot i_*^{\frac{1-\kappa}{2}} L^{1/2} \cdot T^{\frac{1+\kappa}{2}}\right)$$

Then we choose $\kappa = 1/3$ to recover the statement in the theorem. □

A.3 Proof of Theorem 2

Here, we restate an prove Theorem 2.

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

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

Proof. First we will show that the sufficient events to prove this result occur with high probability. While the other events remain the same. we must modify event E_2 from Lemma 2 slightly because we are interested in the case when all algorithms are compatible with respect to their own policy classes. Let E_2' denote the following event: for all $t \in [T]$ and $i \in [L]$,

$$\textstyle\sum_{t' \in \mathcal{T}_t^i} V_i^* - V^{\pi_{t'}} \leq \mathcal{R}_i^{\Pi_i}(d_i, H, \log(T/\delta))\sqrt{|\mathcal{T}_t^i|}$$

As in Lemma 2, this almost follows from Definition 2; however, we also union bound over all algorithms. Thus E_2' occurs with probability at least $1 - LT\delta$. Let $E_1' = E_1$ and $E_3' = E_3$. Then $E' = \bigcap_{i \in 1,2,3} E_i'$ occurs with probability at least $1 - 10LT^2\delta \log_2 T$, as before.

Recall that $i_* = \min B_*$ where B_* is the set of indices that achieve maximal value, $\arg\max_i V_i^*$. For shorthand, we will let $\mathcal{R}_j := \mathcal{R}_j^{\Pi_j}(d_j, H, \log(T/\delta))$. We now verify that the statistical test will not fail once ECE reaches some $i_* \in B_*$. This is nearly identical to Lemma 1, but we must verify it with respect to values that are not the optimal value.

Lemma 3. *Let $(\mathcal{A}_i, \mathcal{F}_i)$ be an $\mathcal{R}_i^{\Pi_i}$-compatible algorithm with respect to Π_i for all $i \in [L]$ and let $i_* = \min B_*$. Given that event E' holds and $t \geq \tau_{\min}(\delta)$, then, for all $j \in [i_*+1, L]$, it holds that $\mathcal{G}_t(i_*, j) \leq \mathcal{W}(|\mathcal{T}_t^{i_*}|, \mathcal{R}_{i_*}, d_{i_*}, \delta)$.*

Proof. From the definition of $\mathcal{G}$,

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

where the last step uses the fact that $V_{i_*}^* = \max_i V_i^*$. Since $(\mathcal{A}_{i_*}, \mathcal{F}_{i_*})$ is $\mathcal{R}_{i_*}^{\Pi_{i_*}}$-compatible, the remainder of the proof is identical to that of Lemma 1 by applying the conditions in E'. □

As before, in the full proof we handle three cases: (1) before the test is valid, (2) while $i < i_*$ is chosen, (3) after i_* is chosen. In the first case, we again pay the burn-in period regret of $\text{Regret}_{1:\tau_{\min}(\delta)-1} = O(H\tau_{\min}(\delta))$. In the third, we showed that the test will never fail once $\hat{\imath}_t = i_*$. Therefore, $\text{Regret}_{\tau_*:T} = O\left(HLT^{1-\kappa} + \mathcal{R}_{i_*} \cdot \sqrt{T}\right)$.

To bound the regret during the misspecified phase, we construct an upper bound on the number of times $\mathcal{A}_j$ can be played for $j < i_*$. Let t be a time such that $\hat{\imath}_t = j < i_*$ and the test succeeds. First, we bound the size of the gaps.

Note that by definition $V_j^* \geq \frac{1}{|\mathcal{T}_t^j|} \sum_{t' \in \mathcal{T}_t^j} V^{\pi_{t'}}$ and event E' ensures that $V_{i_*}^* \leq \frac{\mathcal{R}_{i_*}}{|\mathcal{T}_t^{i_*}|^{1/2}} + \frac{1}{|\mathcal{T}_t^{i_*}|} \sum_{t' \in \mathcal{T}_t^{i_*}} V^{\pi_{t'}}$.

Then,

$$\begin{aligned}
\Delta_{j,i_*} &= V^*_{i_*} - V^*_j \\
&\le \frac{1}{|\mathcal{T}_t^{i_*}|}\sum_{t'\in\mathcal{T}_t^{i_*}} V^{\pi_{t'}} + \frac{\mathcal{R}_{i_*}}{|\mathcal{T}_t^{i_*}|^{1/2}} - \frac{1}{|\mathcal{T}_t^{j}|}\sum_{t'\in\mathcal{T}_t^{j}} V^{\pi_{t'}} \\
&= \frac{\mathcal{R}_{i_*}}{|\mathcal{T}_t^{i_*}|^{1/2}} + \frac{1}{|\mathcal{T}_t^{i_*}|}\sum_{t'\in\mathcal{T}_t^{i_*}} (g_{t'} - \epsilon_{t'}) - \frac{1}{|\mathcal{T}_t^{j}|}\sum_{t'\in\mathcal{T}_t^{j}} (g_{t'} - \epsilon_{t'}) \\
&\le \frac{\mathcal{W}(|\mathcal{T}_t^{j}|, \mathcal{R}_j, d_j, \delta)}{|\mathcal{T}_t^{j}|} + \frac{\mathcal{R}_{i_*}}{|\mathcal{T}_t^{i_*}|^{1/2}} - \frac{1}{|\mathcal{T}_t^{i_*}|}\sum_{t'\in\mathcal{T}_t^{i_*}} \epsilon_{t'} + \frac{1}{|\mathcal{T}_t^{j}|}\sum_{t'\in\mathcal{T}_t^{j}} \epsilon_{t'} \\
&\le C_{\mathcal{W}} \cdot \left(\frac{\mathcal{R}_j}{|\mathcal{T}_t^{j}|^{1/2}} + H\sqrt{\frac{16L\log(2/\delta)}{|\mathcal{T}_t^{j}|^{1-\kappa}}} + H\sqrt{\frac{2\log(2/\delta)}{|\mathcal{T}_t^{j}|}} \right) \\
&\quad + \frac{\mathcal{R}_{i_*}}{|\mathcal{T}_t^{i_*}|^{1/2}} + H\sqrt{\frac{2\log(2/\delta)}{|\mathcal{T}_t^{i_*}|}} + H\sqrt{\frac{2\log(2/\delta)}{|\mathcal{T}_t^{j}|}}
\end{aligned}$$

where we have applied the definition of $\mathcal{W}$ and event E_3 to bound the noise of the returns. Let $C'_{\mathcal{W}} = \max\{1, C_{\mathcal{W}}\}$. Since i_* has not been selected yet $|\mathcal{T}_t^{i_*}| \ge \frac{t^{1-\kappa}}{8L} \ge \frac{|\mathcal{T}_t^{j}|^{1-\kappa}}{8L}$. Then, since $\mathcal{R}_j \le \mathcal{R}_{i_*}$,

$$\Delta_{j,i_*} \le C'_{\mathcal{W}} \cdot \left(\frac{2\sqrt{8L}\mathcal{R}_{i_*}}{|\mathcal{T}_t^{j}|^{\frac{1-\kappa}{2}}} + H\frac{2\sqrt{16L\log(2/\delta)}}{|\mathcal{T}_t^{j}|^{\frac{1-\kappa}{2}}} \right)$$

Rearranging gives

$$|\mathcal{T}_t^{j}| = O\left(\frac{L^{\frac{1}{1-\kappa}} \left(\mathcal{R}_{i_*} + H\log^{1/2}(1/\delta) \right)^{\frac{2}{1-\kappa}}}{\Delta_{j,i_*}^{\frac{2}{1-\kappa}}} \right)$$

Now this bound can be used to bounding the regret with dependence on the gap. The regret during this phase is again

$$\begin{aligned}
\mathrm{Regret}_{\tau_{\min}(\delta):\tau_*} &\le H(L - i_*)\tau_*^{1-\kappa} + \sum_{j<i_*}\sum_{t\in\mathcal{T}^j_{\tau_{j+1}}} V^*_{i_*} - V^{\pi_t} \\
&\le H(L - i_*)\tau_*^{1-\kappa} + Hi_* + \sum_{j<i_*}\sum_{t\in\mathcal{T}^j_{\tau_{j+1}-1}} V^*_{i_*} - V^{\pi_t}
\end{aligned}$$

As in the proof of Theorem 1, we focus on bounding the right-hand term. For a fixed $j < i_*$, at time $\tau_{j+1} - 1$ we have that the test succeeds so $\mathcal{G}_{\tau_{j+1}-1}(j, i_*) \le \mathcal{W}(|\mathcal{T}^j_{\tau_{j+1}-1}|, \mathcal{R}_{i_*}, d_{i_*}, \delta)$. Then, applying the bound on the number of times j can be played,

$$\begin{aligned}
\sum_{t\in\mathcal{T}^j_{\tau_{j+1}-1}} V^*_{i_*} - V^{\pi_t} &\le \Delta_{j,i_*}|\mathcal{T}^j_{\tau_{j+1}-1}| + \mathcal{R}_j \cdot \sqrt{|\mathcal{T}^j_{\tau_{j+1}-1}|} \\
&\le O\left(\frac{L^{\frac{1}{1-\kappa}} \left(\mathcal{R}_{i_*} + H\log^{1/2}(1/\delta) \right)^{\frac{2}{1-\kappa}}}{\Delta_{j,i_*}^{\frac{1+\kappa}{1-\kappa}}} + \frac{\mathcal{R}_{i_*} L^{\frac{1}{2(1-\kappa)}} \left(\mathcal{R}_{i_*} + H\log^{1/2}(1/\delta) \right)^{\frac{1}{1-\kappa}}}{\Delta_{j,i_*}^{\frac{1}{1-\kappa}}} \right) \\
&= O\left(\frac{L^{\frac{1}{1-\kappa}} \left(\mathcal{R}_{i_*} + H\log^{1/2}(1/\delta) \right)^{\frac{2}{1-\kappa}}}{\Delta_{j,i_*}^{\frac{1+\kappa}{1-\kappa}}} \right)
\end{aligned}$$

Therefore, the regret in this phase can be upper bounded by

$$\text{Regret}_{\tau_{\min}(\delta):\tau_*} \leq O\left(H(L-i_*)T^{1-\kappa} + Hi_* + L^{\frac{1}{1-\kappa}}\left(\mathcal{R}_{i_*} + H\log^{1/2}(1/\delta)\right)^{\frac{2}{1-\kappa}} \sum_{j<i_*} \frac{1}{\Delta_{j,i_*}^{\frac{1+\kappa}{1-\kappa}}}\right)$$

Combining these three phases, the total regret is

$$O\left(HL^{\frac{2}{1-\kappa}}\log^{\frac{1}{1-\kappa}}(1/\delta) + HLT^{1-\kappa} + Hi_* + L^{\frac{1}{1-\kappa}}\left(\mathcal{R}_{i_*} + H\log^{1/2}(1/\delta)\right)^{\frac{2}{1-\kappa}} \sum_{j<i_*} \frac{1}{\Delta_{j,i_*}^{\frac{1+\kappa}{1-\kappa}}} + \mathcal{R}_{i_*}\sqrt{T}\right)$$

Choosing either $\kappa = 1/3$ or $\kappa = 1/2$ gives us the statements of Theorem 2. This completes the proof. □

B FREEDMAN INEQUALITY

In this section, we use a Freedman inequality to lower and upper bound with high probability the number of times a particular algorithm is played both during exploration and while it is chosen by the meta-algorithm (Lemma 2). First, we state a variant of the Freedman inequality from Bartlett et al. (2008).

Lemma 4 (Lemma 2, Bartlett et al. (2008))**.** *Suppose $X_1, \cdots, X_T$ is a martingale difference sequence with $|X_s| \leq b$. We define*

$$\text{Var}_s X_s = \mathbf{Var}(X_s | X_1, \cdots, X_{s-1})$$

Further, let $V_T = \sum_{s=1}^T \text{Var}_s X_s$ be the sum of conditional variances of X_s's, and $\sigma_T = \sqrt{V_T}$. Then we have, for any choice of $\delta < 1/e$ and $T \geq 4$:

$$\mathbb{P}\left(\sum_{s=1}^T X_s > 2\max(2\sigma_T, b\sqrt{\ln(1/\delta)})\sqrt{\ln(1/\delta)}\right) \leq \log_2(T)\delta \tag{5}$$

Recall that B_s denotes the indices of algorithms that have not been selected by time s. Note that $|B_s| \leq L$. For all $i \in [L]$ and $t \in [T]$, define the event

$$\mathcal{E}_{i,t} := \begin{cases} ||\mathcal{T}_t^i| - \sum_{s\in[t]} \frac{1}{|B_s|s^\kappa}| \leq 4\sqrt{\sum_{s\in[t]} \frac{1}{s^\kappa}\log(1/\delta)} & \tau_i \geq t \\ ||\mathcal{T}_t^i| - \sum_{s\in[\tau_i]} \frac{1}{|B_s|s^\kappa} - \sum_{s\in[\tau_i+1,t]} \left(1 - \frac{1}{s^\kappa}\right)| \leq 4\sqrt{\sum_{s\in[t]} \frac{1}{s^\kappa}\log(1/\delta)} & \tau_i < t \end{cases}$$

Lemma 5. *The event $\mathcal{E} = \cap_{i\in[L],t\in[T]}\mathcal{E}_{i,t}$ holds with probability at least $1 - 4LT^2\delta\log_2 T$*

Proof. Define

$$S_i(t, t') = \sum_{s\in[t']} Y_{s,i} + \sum_{s\in[t'+1,t]} \overline{Y}_{s,i}$$

where $Y_{s,i} \sim \text{Ber}\left(\frac{1}{s^\kappa|B_s|}\right)$ and $\overline{Y}_{s,i} \sim \text{Ber}\left(1 - \frac{1}{s^\kappa}\right)$. Then define

$$Z_i(t,t') := \sum_{s\in[t]} \mathbf{1}_{s\leq t'} \cdot \left(Y_{s,i} - \frac{1}{|B_s|s^\kappa}\right) + \mathbf{1}_{s>t'}\left(\overline{Y}_{s,i} - \left(1 - \frac{1}{s^\kappa}\right)\right)$$

$$V_i(t,t') := \sum_{s\in[t]} \mathbf{Var}_s\left(\mathbf{1}_{t\leq t'} \cdot \left(Y_{s,i} - \frac{1}{|B_s|s^\kappa}\right) + \mathbf{1}_{t>t'} \cdot \left(\overline{Y}_{s,i} - \left(1 - \frac{1}{s^\kappa}\right)\right)\right)$$

where $\mathbf{Var}_s$ denotes the conditional variance up to time s. By definition, $\{Z_i(t,t')\}_{t\geq 1}$ is a martingale sequence and $V_i(t,t') \leq \sum_{s\in[t]} \frac{1}{s^\kappa}$. By the Freedman inequality from Lemma 4,

$$\mathbb{P}\left(|Z_i(t,t')| \geq 4\sqrt{\sum_{s\in[t]} \frac{1}{s^\kappa} \cdot \log(1/\delta) + 4\log(1/\delta)}\right) \leq 2\delta\log_2 T$$

Let this event be denoted by $\overline{\mathcal{E}}_i(t,t')$ for each $i \in [L]$ and $t,t' \in [T]$. Then, by the union bound, the event $\bigcup_{i,t,t'} \overline{\mathcal{E}}_i(t,t')$ holds with probability at most $4LT^2\delta \log_2 T$. Therefore, $\bigcap_{t,t' \geq 1} \mathcal{E}_i(t,t')$ holds with probability at least $1 - 4LT^2\delta \log_2 T$, and this event implies for all $i \in [L]$ and $t \in [T]$, if $t > \tau_i$, then

$$||\mathcal{T}_t^i| - \sum_{s \in [\tau_i]} \frac{1}{|B_s| s^\kappa} - \sum_{s \in [\tau_i + 1, t]} \left(1 - \frac{1}{s^\kappa}\right)| \leq 4\sqrt{\sum_{s \in [t]} \frac{1}{s^\kappa} \log(1/\delta)} + 4\log(1/\delta)$$

and if $\tau \leq \tau_i$, then

$$||\mathcal{T}_t^i| - \sum_{s \in [t]} \frac{1}{|B_s| s^\kappa}| \leq 4\sqrt{\sum_{s \in [t]} \frac{1}{s^\kappa} \log(1/\delta)} + 4\log(1/\delta)$$

□

Corollary 1. *With probability at least $1 - 4LT^2\delta \log_2 T$, for all $i \in [L]$ and $t \in [T]$ such that $t \geq \tau_{\min}(\delta)$, the following is true:*

1. *If $t \leq \tau_i$, then $\frac{t^{1-\kappa}}{8L} \leq |\mathcal{T}_t^i| \leq 4t^{1-\kappa}$.*
2. *If $t > \tau_i$, then $|\mathcal{T}_t^i| \leq t - \tau_i + 4t^{1-\kappa}$.*

Proof. Note that when $t \leq \tau_i$, it is also the case that $|B_s| \geq 1$ for all $s \leq t$. We condition on the event $\mathcal{E}$ from above, which occurs with probability at least $1 - 4LT^2\delta \log_2 T$. Given this event, it follows that if $t \leq \tau_i$, then

$$\begin{aligned}
|\mathcal{T}_t^i| &\geq \sum_{s \in [t]} \frac{1}{s^\kappa |B_s|} - 4\sqrt{\sum_{s \in [t]} \frac{1}{s^\kappa} \log(1/\delta)} - 4\log(1/\delta) \\
&\geq \frac{1}{2L} \sum_{s \in [t]} \frac{1}{s^\kappa} - 32L\log(1/\delta) \\
&\geq \frac{1}{2L}\left(t^{1-\kappa} - 2\right) - 32L\log(1/\delta) \\
&\geq \frac{t^{1-\kappa}}{4L} - 32L\log(1/\delta) \\
&\geq \frac{t^{1-\kappa}}{8L}
\end{aligned}$$

The second inequality uses the AM-GM inequality along with the fact that $|B_s| \leq L$, which implies

$$\sqrt{\sum_{s \in [t]} \frac{1}{Ls^\kappa} \cdot 16L\log(1/\delta)} \leq \frac{1}{2L} \sum_{s \in [t]} \frac{1}{s^\kappa} + 8L\log(1/\delta)$$

The third applies the integral approximation of the sum. The last two follow from the condition that $t \geq \tau_{\min}(\delta) = C_{\min} \cdot L^{\frac{2}{1-\kappa}} \log^{\frac{1}{1-\kappa}}(1/\delta)$ for a large enough constant $C_{\min} > 0$.

The other side follows similarly with

$$\begin{aligned}
|\mathcal{T}_t^i| &\leq 3t^{1-\kappa} + 32\log(1/\delta) \\
&\leq 4t^{1-\kappa}
\end{aligned}$$

when $t \geq (32\log(1/\delta))^{\frac{1}{1-\kappa}}$.

Similarly, for $t > \tau_i$, event $\mathcal{E}$ guarantees

$$\begin{aligned}|\mathcal{T}_t^i| &\leq \sum_{s \in [\tau_i]} \frac{1}{s^\kappa |B_s|} + \sum_{s \in [\tau_i + 1, t]} \left(1 - \frac{1}{s^\kappa}\right) + 4\sqrt{\sum_{s \in [t]} \frac{1}{s^\kappa} \log(1/\delta)} + 4\log(1/\delta) \\ &\leq t - \tau_i + 32\log(1/\delta) + \frac{3}{2}\sum_{s \in [\tau_i]} \frac{1}{s^\kappa} \\ &\leq t - \tau_i + 32\log(1/\delta) + 3t^{1-\kappa} \\ &\leq t - \tau_i + 4t^{1-\kappa}\end{aligned}$$

when $t \geq \tau_{\min}(\delta)$.

□

C APPLICATIONS

In this section, we expand on the applications of Theorem 1 to paradigms of function approximation in RL.

Linear MDPs Consider the setting of Jin et al. (2020b) which we mentioned as an example in Section 3. In this setting, we assume access to a set of 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 features generate linear model classes of the form

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

Nested-ness of the features ensures that $\mathcal{F}_i \subseteq \mathcal{F}_{i+1}$ for all i. In accordance with the setting of Jin et al. (2020b), we assume that there exists some minimal i_* such that for any $\mathcal{F}_i$ with $i \geq i_*$ there exist $\mu(\cdot)$ and $\omega_{i,h} \in \mathbb{R}^{d_i}$ that predict exactly the transition probabilities P and reward r:

$$\begin{aligned}P(s'|s,u) &= \langle \phi_i(s,u), \mu_i(s') \rangle \\ r_h(s,u) &= \langle \phi_i(s,u), \omega_{i,h} \rangle\end{aligned} \tag{6}$$

Here, $\mu_i(\cdot)$ is a d_i-dimensional vector of measures on $\mathcal{S}$. Let $\{\mathcal{A}_i\}$ be instances of LSVI-UCB equipped with the doubling trick and model classes $\{\mathcal{F}_i\}$. We further assume that the features and parameters for each of the models with $i \geq i_*$ satisfies the regularity conditions of Assumption A of Jin et al. (2020b), i.e. bounded ℓ_2 norms, $r \in [0,1]$.

Jin et al. (2020b) guarantees that for $i \geq i_*$ and $t \in [T]$ with probability at least $1 - \delta_0$, $\text{Regret}_t(\mathcal{A}_i) = O(\sqrt{d_i^3 H^4 t \cdot \log^2(d_i TH/\delta_0)})$. Adapting this to the framework of ECE, we let $\mathcal{R}_i = O\left(\sqrt{d_i^3 H^4 \cdot \log^2(d_i TH/\delta)}\right)$, which ensures $\mathcal{R}_i \leq \mathcal{R}_{i+1}$. A model selection corollary immediately follows from Theorem 1.

Corollary 2. *In the linear MDP setting of (6) with LSVI-UCB, ECE guarantees with probability at least* $1 - \delta'$

$$\mathit{Regret}_T = \widetilde{O}\left(\sqrt{d_{i_*}^3 H^4 \log^2(d_{i_*} LTH/\delta')} \cdot L^{5/6} T^{2/3}\right)$$

Yang and Wang (2020) consider a similar setting of linear MDPs where the transition dynamics P are linear. We again assume access to nested linear models but of the form

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

where $\{\phi_i\}_{i \in [L]}$ and $\{\psi_i\}_{i \in [L]}$ are nested features of dimension d_i and d_i' respectively. Yang and Wang (2020) assume that there is some minimal i_* such that for any $i \geq i_*$, there is $M \in \mathbb{R}^{d_i \times d_i'}$ such that

$$P(s'|s,u) = \phi_i(s,u)^\top M \psi_i(s') \tag{7}$$

for all $s, s' \in \mathcal{S}$, $u \in \mathcal{U}$. We further adhere to the regularity conditions of Assumption 2 of Yang and Wang (2020), who guarantee the MatrixRL $\mathcal{A}_i$ with model $\mathcal{F}_i$ has regret $\text{Regret}_t(\mathcal{A}_i) = \widetilde{O}\left(\sqrt{d_i^3 H^5 t} \cdot \log(d_i TH/\delta_0)\right)$

with probability at least $1-\delta_0$. Letting $\mathcal{R}_i = \widetilde{O}\left(\sqrt{d_i^3 H^5} \cdot \log(d_i TH/\delta)\right)$, we have the following model selection guarantee.

Corollary 3. *In the linear MDP setting of (7) with MatrixRL, ECE guarantees with probability at least* $1-\delta'$

$$\mathit{Regret}_T = \widetilde{O}\left(\sqrt{d_{i_*}^3 H^5 \log^2(d_{i_*} LTH/\delta')} \cdot L^{5/6} T^{2/3}\right)$$

The final linear setting we consider is that of low inherent Bellman error studied by Zanette et al. (2020). We let $\mathcal{F}_i$ be defined as it is in (3) and let $\mathcal{B} = \{\theta \in \mathbb{R}^{d_i} \ : \ \|\theta\| \leq D\}$ for some $D > 0$. Then assume there is a minimal i_* such that for any $i \geq i_*$ and $\theta_{h+1} \in \mathcal{B}$, there is θ_h such that

$$\langle \phi_i(s,u), \theta_h \rangle - \mathbf{B}_h Q_{h+1}(\theta_{h+1})(s,u) = 0$$

for all $s \in \mathcal{S}$ and $u \in \mathcal{U}$, where $Q_h(\theta)$ is the linear action-value function parameterized by θ (with features ϕ_i) and $\mathbf{B}_h$ is the Bellman operator with reward r_h. In other words, this condition asserts that $\mathcal{F}_{i_*}$ has zero inherent Bellman error. Under the same regularity conditions, for $i \geq i_*$, Zanette et al. (2020) guarantees ELEANOR achieves $\text{Regret}_t(\mathcal{A}_i) = \widetilde{O}\left(d_i \sqrt{H^4 t}\right)$ with probability at least $1-\delta_0$. Letting $\mathcal{R}_i = \widetilde{O}\left(d_i \sqrt{H^4}\right)$, we have the following model selection guarantee.

Corollary 4. *In the inherent Bellman error setting with ELEANOR, ECE guarantees with probability at least* $1-\delta'$

$$\mathit{Regret}_T = \widetilde{O}\left(d_{i_*} \sqrt{H^4} \cdot L^{5/6} T^{2/3}\right)$$

where $\widetilde{O}$ hides polylog dependencies.

Low Bellman Rank Another class of algorithms using more general function approximation considers the setting of MDPs with low Bellman rank (Jiang et al., 2017). In this setting, a finite model class $\mathcal{F} : \mathcal{S} \times \mathcal{U} \to \mathbb{R}$ realizes $\mathcal{M}$ if there exists $f^* \in \mathcal{F}$ such that $Q_h^*(s,a) = f^*(s,a)$, where Q^* is the optimal action-value function for all $h \in [H]$. For any $f \in \mathcal{F}$, define π_f as the greedy policy with respect to f, and the Bellman error at $h \in [H]$ as

$$\mathcal{E}(f, \pi, h) := \mathbb{E}\left[f(s, \pi_f(s)) - r(s, \pi_f(s)) - f(s', \pi_f(s'))\right],$$

where the expectation is over s from the state distribution of π at h and $s' \sim P(\cdot | s, \pi_f(s))$. In this setting, it is assumed that there is a Bellman rank $M \ll |\mathcal{F}|$ such that for any $f, g \in \mathcal{F}$, we have $\mathcal{E}(f, \pi_g, h) = \langle \nu_h(g), \xi_h(f) \rangle$ for $\nu_h(g), \xi_h(f) \in \mathbb{R}^M$ and $\|\nu\|\|\xi\| \leq \zeta$. We assume access to a set of finite model classes $\{\mathcal{F}_i\}_{i \in [L]}$ such that there is at least one that realizes $\mathcal{M}$, and the complexity of $\mathcal{F}_i$ is a function of its cardinality $|\mathcal{F}_i|$ and induced Bellman rank M_i. We consider instances of the AVE algorithm $\{\mathcal{A}_i\}$ of Dong et al. (2020) with the doubling trick, which has nominal regret $\widetilde{O}\left(\sqrt{M_i^2 |\mathcal{U}| H^4 t \log^3 |\mathcal{F}_i|}\right)$. Choose $\mathcal{R}_{\mathcal{F}_i} = \widetilde{O}\left(\sqrt{M_i^2 |\mathcal{U}| H^4 \log^3(|\mathcal{F}_i|)}\right)$ and let i_* be the smallest index that realizes $\mathcal{M}$. This yields the following corollary.

Corollary 5. *In the low Bellman rank setting with AVE, the model selection algorithm guarantees with probability at least* $1-\delta'$

$$\mathit{Regret}_T(\mathcal{A}) = \widetilde{O}\left(\sqrt{M_{i_*}^2 |\mathcal{U}| H^4 \log^3(|\mathcal{F}_{i_*}|)} \cdot L^{5/6} T^{2/3}\right).$$

D Implications of fast rates of estimating V^* and/or gap between policy classes

We previously discussed the recent results that prove PAC (Modi et al., 2020) and regret (Pacchiano et al., 2020) results for model selection in RL given knowledge of V^*. We now show an analogous result for our setting. We use the framework of Algorithm 1 but set the probability of forced exploration to zero, i.e. set $\kappa = \infty$. Then, the test is modified to check the following condition for eliminating model $\hat{\imath}_t$:

$$\sum_{t' \in \mathcal{T}_t^{\hat{\imath}_t}} V^* - g_{t'} > \mathcal{W}_{V^*}(|\mathcal{T}_t^{\hat{\imath}_t}|, \mathcal{R}_{\hat{\imath}}, d_{\hat{\imath}_t}, \delta)$$

where

$$\mathcal{W}_{V^*}(\Delta, \mathcal{R}, d, \delta) = C_{\mathcal{W}} \cdot \mathcal{R}(d, H, \log(1/\delta)) \cdot \sqrt{\Delta} \\ + C_{\mathcal{W}} \cdot H\sqrt{\Delta \cdot \log(1/\delta)}$$

for a sufficiently large constant $C_{\mathcal{W}_{V^*}} > 0$. The test effectively measures the regret of $\mathcal{A}_{\hat{\imath}_t}$ up to noise in g_t and rejects when we are confident that the regret does not match the nominal.

Proposition 1. *Given side information of the optimal value V^* for MDP $\mathcal{M}$, the above model selection algorithm $\mathcal{A}$ guarantees regret*

$$\mathit{Regret}_T(\mathcal{A}) = \widetilde{O}\left(\mathcal{R}_{i_*}(d_{i_*}, H, \log(LT/\delta')) \cdot \sqrt{LT}\right)$$

with probability at least $1 - \delta'$.

Proof. The proof is identical to that of Theorem 1 except for the handling of the misspecified case. For any model $j < i_*$ for which there is a time when the test succeeds,

$$\begin{aligned}\sum_{t \in \mathcal{T}^j_{\tau_{j+1}-1}} V^* - V^{\pi_t} &= \sum_{t \in \mathcal{T}^j_{\tau_{j+1}-1}} (V^* - g_t) + \sum_{t \in \mathcal{T}^j_{\tau_{j+1}-1}} \epsilon_t \\ &\leq \mathcal{W}_{V^*}(|\mathcal{T}^j_t|, \mathcal{R}_j, d_j, \delta) + \sum_{t \in \mathcal{T}^j_{\tau_{j+1}-1}} \epsilon_t \\ &= O\left(\left(\mathcal{R}_{i_*} + H\log^{1/2}(1/\delta)\right) \cdot \sqrt{|\mathcal{T}^j_t|}\right)\end{aligned}$$

Summing over all $j < i_*$ and using Jensen's inequality again shows that the dominant term remains $O(\mathcal{R}_{i_*}\sqrt{T})$ instead of $O(\mathcal{R}_{i_*}T^{2/3})$. □

This regret optimally matches the regret of the base algorithms in both $\mathcal{R}_{i_*}$ and T, but a dependence on L is still included.

Unfortunately, it is unclear whether such an assumption of knowing V^* is realistic in practice. An immediate alternative solution is to try to estimate V^* without first finding the optimal policy. The original test in Section 4.2 attempts this: the average returns of the algorithms in B_t act as a noisy lower bound of V^*. The test, however, is sensitive to the amount of exploration allocated to the base algorithms, and, since we are comparing to the nominal regret, the flat dependence on $\mathcal{R}$ is unlikely to improve. We hypothesize that better estimates of V^* can significantly improve the model selection guarantee.

In the following subsections, we consider the implications of having access to fast estimators, either of the optimal value $V^* := V^*_{i_*}$ or *gaps* between optimal values of different model orders, i.e. $\Delta_{i,j} := V^*_i - V^*_j$. We employ our instance-dependent analysis to show that improved regret rates can be obtained in both cases when the gap between the value of the optimal policy class and others is relatively large (i.e. constant). These consequences are demonstrated for the special case of linear contextual bandits, where such fast estimators are known to be available (Dicker, 2014; Kong and Valiant, 2018; Kong et al., 2020; Verzelen et al., 2018).

D.1 Implications for access to a fast rate of estimating *gaps* in policy class optimal values

We first consider the possibility of fast rates in estimating the *gap* in optimal policy values, i.e. $\Delta_{i,j} := V^*_j - V^*_i$ for all $i < j$. In this section, we show that a modification of our ECE algorithm with a direct estimator of the gap in maximal values would yield improved model selection rates if there is a constant gap between all lower-order models and the true model, i.e. $\Delta_{i,i_*} > 0$ for all i. Along with the replaced estimator, the radius of the statistical test is also modified according to the faster estimation error rate in the policy gap. For the special case of linear contextual bandits, these modifications will correspond *exactly* to the ModCB algorithm proposed by Foster et al. (2019).

Since our focus is on instance-dependent analysis, we carry over the assumptions from Section 6, and further assume model nested-ness in the sense that $V^*_j = V^*$ for $j \geq i_*$. Thus, we get $\Delta_{i_*,i} = 0$ for all $i \geq i_*$, and

Algorithm 2 Explore-Commit-Eliminate With Fast Gap Estimator And Forced Exploration Routines(ECE-Gap)

```
**Input**: $\{\mathcal{A}_i, \widetilde{\mathcal{A}}_i, \mathcal{F}_i, \mathcal{V}_i, d_i\}_{i\in[L]}, T, \delta', \tau_{\min}(\cdot)$
$\delta \leftarrow \frac{\delta'}{10LT^2\log_2 T}$, $\hat{\imath}_t \leftarrow 1$, $\mathcal{T}_1^i = \emptyset$ for $i \in [L]$, $B_1 = [2, L]$
$U_t = \begin{cases} 0 & \text{w.p. } 1 - \frac{1}{t^\kappa} \\ 1 & \text{w.p. } \frac{1}{t^\kappa} \end{cases}$ for all $t \in [T]$.
**for** $t = 1, \ldots, T$ **do**
  **if** $U_t = 0$ **then**
    Set $j \leftarrow \hat{\imath}$.
  **else**
    Sample $J_t \sim \mathrm{Unif}\{B_t\}$
    Set $j \leftarrow J_t$
  **end if**
  $\mathcal{T}_t^j \leftarrow \mathcal{T}_t^j \cup \{t\}$ and $\mathcal{T}_t^k \leftarrow \mathcal{T}_t^k$ for all $k \neq j$.
  IF $U_t = 0$: Rollout policy $\pi_t$ from $\mathcal{A}_j$.
  ELSE: Rollout policy $\pi_t$ from $\widetilde{\mathcal{A}}_j$.
  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 $\mathcal{A}_j$ if $U_t = 0$, else update $\widetilde{\mathcal{A}}_j$ with $t, z_t, g_t$
  **if** $t \geq \tau_{\min}(\delta)$ and there exists $j \in B_t$ such that $\widehat{\Delta}_{\hat{\imath}_t, j}(\mathcal{T}_t^j) > \mathcal{Z}(|\mathcal{T}_t^j|, \mathcal{V}_j)$ **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**
    $B_{t+1} = B_t$
  **end if**
**end for**
```

$\Delta_{i,i_*} > 0$ for all $i < i_*$. To estimate the gap during exploration episodes, rather than running $\mathcal{A}_i$ directly, we allow an exploration algorithm $\widetilde{\mathcal{A}}_i$ to be run. In the case of Foster et al. (2019) for contextual bandits, this would be an exploration policy that picks an arm uniformly at random from the set of K arms. Finally, we make the following assumption on the estimation error rate of the gaps.

Assumption 1. *For any $i < j$, we define $\widehat{\Delta}_{i,j}^{(n)}$ as an estimate of $\Delta_{i,j}$ that is a functional of the (context and reward) feedback obtained after running n exploration episodes for $\widetilde{\mathcal{A}}_j$. Then, we say that our estimate is $\mathcal{V}_j := \mathcal{V}(d_j, H, \log(1/\delta))$-consistent if, for some positive constant $C > 1$, we have*

$$|\widehat{\Delta}_{i,j}^{(n)} - \Delta_{i,j}| \leq \frac{\Delta_{i,j}}{C} + \frac{\mathcal{V}_j}{\sqrt{n}} \text{ for all } n \in [T] \text{ and } i < j \tag{8}$$

with probability at least $1 - \delta$. As with the earlier definition[5]*, $\mathcal{V}_|$ is poly and non-decreasing in d_j, H, $|\mathcal{U}|$, and $\log(LT/\delta)$).*

The original estimator used in the ECE algorithm satisfies the above assumption with $\mathcal{V} := \mathcal{R}$. In what follows, we want to exploit situations in which we have available an estimator $\widehat{\Delta}_{i,j}$ with guarantee $\mathcal{V} \ll \mathcal{R}$; in particular, the dependence of the function $\mathcal{V}$ on dimension d could be significantly improved over any regret bound. While constructing such estimators is in general a open problem in RL, we do have one example for the linear contextual bandit problem where this is known to be possible.

Example 1. *[Linear contextual bandits.] Consider the stochastic d^{th}-order linear contextual bandits model as in Chu et al. (2011), parameterized by K context distributions $\{\Sigma_i\}_{i=1}^K$, reward parameter $\theta^* \in \mathbb{R}^d$, and σ-sub-Gaussian noise in the rewards. Further, we carry over the assumptions from Foster et al. (2019) of τ-sub-Gaussianity of the contexts and $\lambda_{min}(\overline{\Sigma}) \geq \nu > 0$ where $\overline{\Sigma} := \frac{1}{K}\sum_{i=1}^K \Sigma_i$ is the action-averaged covariance matrix. We assume that τ, ν are universal positive constants. Then, Assumption 1 holds with the choice of forced*

[5] Similar to $\mathcal{R}$, the definition of $\mathcal{V}_j$ can be general and include other problem dependent parameters as well.

exploration $\widetilde{\mathcal{A}}_i$ that chooses arms uniformly at random from the set $[K]$ (regardless of round index t and model index i), with the choices $C = 2$ and $\mathcal{V}_i(d_i, \log(1/\delta))$ scaling as $\widetilde{O}(d_i^{1/4})$ for the estimator based on the square loss gap, used in Foster et al. (2019). Meanwhile, the regret bound for the base algorithms (e.g. instances of Exp4-IX) would give $\mathcal{R}_i$ scaling as $\widetilde{O}(d_i^{1/2})$. Further, note that Algorithm 2 exactly becomes the ModCB *algorithm for this case.*

We now described the modified ECE algorithm, ECE-Gap, to work with a plugged-in estimate of $\Delta_{i,j}$ with the above guarantees. Note that the input now has extra "exploration algorithms" $\widetilde{\mathcal{A}}_i$, and what was earlier defined as regret bound leading factors, i.e. $\mathcal{R}_i$, is replaced by $\mathcal{V}_i$, the leading factors in the gap estimation error. Importantly, we are now using the fast estimator $\widehat{\Delta}_{i,j}(t)$ in place of the earlier estimator $\mathcal{G}_t(j,i)/|\mathcal{T}_t^j|$.

Moreover, the threshold is now defined as:

$$\mathcal{Z}(n, \mathcal{V}) := \frac{\mathcal{V}}{\sqrt{n}}$$

Note that the threshold *is always applied to the more complex model* $d := d_i$ for $i > j$. The algorithm is stated formally in Algorithm 2. We derive the following instance-dependent result for this algorithm.

Proposition 2. *For a given $\mathcal{M}$, let Assumption 1 hold and let $\{\Delta_{i,i_*}\}_{i<i_*}$ be the gaps. Then, with probability at least $1-\delta'$,* ECE-Gap *in Algorithm 2 satisfies the regret bound*

$$\widetilde{O}\left(HLT^{1-\kappa} + \mathcal{R}_{i_*}^{\Pi_{i_*}}\sqrt{LT} + \sum_{i=1}^{i_*-1} \min\{L^{\frac{1}{1-\kappa}}\mathcal{V}_{i_*}^{\frac{2}{1-\kappa}}\Delta_{i,i_*}^{-\frac{1+\kappa}{1-\kappa}}, \Delta_{i,i_*}T\}\right),$$

where regret is measured with respect to the optimal value V^*.

Before proving Proposition 2, let us consider its implication for the linear contextual bandits setting, ignoring dependence on $K = |\mathcal{U}|$ for now. Here, the modified ECE algorithm will essentially correspond to ModCB.

By choosing $\kappa = 1/3$ and using the gap estimator from Foster et al. (2019), we can achieve an instance-dependent result with lower d_{i_*} dependence than that of Theorem 2 for the same setting of κ under the assumption of constant gaps. Furthermore, in the case the case of variable gaps, this result can immediately imply a minimax guarantee that matches that of Foster et al. (2019).

Corollary 6. *For the linear contextual bandit problem, under the same setting as Corollary 7, with probability at least $1-\delta'$, Algorithm 2 with $\kappa = 1/3$ and constant gaps satisfies the instance-dependent regret bound*

$$\widetilde{O}\left(LT^{2/3} + \sqrt{d_{i_*}LT} + L^{3/2}d_{i_*}^{3/4}\sum_{i<i_*}\Delta_{i,i_*}^{-2}\right) = \widetilde{O}\left(LT^{2/3} + \sqrt{d_{i_*}LT}\right). \tag{9}$$

Furthermore, for variable gaps, let $Regret_T(\mathcal{A}; \mathcal{M}, \{\Delta_{i,i_*}\}_i)$ denote the regret as a function of the gaps. Since $\min\{L^{3/2}\mathcal{V}_{i_*}^3\Delta_{i_*,i}^{-2}, \Delta_{i_*,i}T\} \leq L^{1/2}\mathcal{V}_{i_*}T^{2/3}$, ECE-Gap *also satisfies the minimax regret bound*

$$\sup_{\Delta_{i,i_*}>0\ :\ i<i_*} Regret_T\left(\textsf{ECE-Gap}; \mathcal{M}, \{\Delta_{i,i_*}\}_i\right) = \widetilde{O}\left(Ld_{i_*}^{1/4}T^{2/3} + \sqrt{d_{i_*}LT}\right).$$

The equality in the (9) uses $d_i \ll T$ for all $i \in [L]$ and the constant gap assumption. If we knew *a priori* that the gaps are constant, the instance-dependent bound in (9) can be improved by a more aggressive choice of $\kappa = 1/2$, as in Theorem 2. We can then achieve the desired regret rate of $\widetilde{O}(\sqrt{d_{i_*}T})$ regret *if and only if* the gaps are constant. Again there is only sub-optimal d_{i_*}-dependence on the term independent of T.

Corollary 7. *For the linear contextual bandit problem under Assumption 1 with constant gaps $\{\Delta_{j,i_*}\}_{j<i_*}$, let $\mathcal{V}_{i_*} := \widetilde{O}(d_{i_*}^{1/4})$ and $\mathcal{R}_{i_*}^{\Pi_{i_*}} := \widetilde{O}(d_{i_*}^{1/2})$. Then, with probability at least $1-\delta'$, Algorithm 2 with $\kappa = 1/2$ satisfies the regret bound*

$$\widetilde{O}\left(L\sqrt{T} + \sqrt{d_{i_*}LT} + L^2 d_{i_*}\sum_{i<i_*}\Delta_{i,i_*}^{-3}\right) = \widetilde{O}\left(L\sqrt{T} + \sqrt{d_{i_*}LT}\right).$$

In summary, Proposition 2 not only recovers the minimax rate, but shows an improved instance-dependent guarantee for more favorable cases when the gap between optimal policy values is larger.

Let us now prove the proposition.

Proof. Let $\widehat{\Delta}^t_{i,j} := \widehat{\Delta}^{(|\mathcal{T}^i_t|)}_{i,j}$. First, we show that under the intersection of the event of Equation (8) and event E' of Theorem 2, we will never reach $\hat{\imath}_t > i_*$. For every $i > i_*$, and all $t \geq 1$, Equation (8) gives us

$$\widehat{\Delta}^t_{i_*,i} \leq \frac{\mathcal{V}_i}{\sqrt{|\mathcal{T}^i_t|}}$$

Thus, model order i_* is never rejected under this event, and higher order models have no contribution to the overall regret.

Next, we bound the regret arriving from the misspecified models $i < i_*$. We do this by bounding the number of rounds during which model order $i < i_*$ is used, given by $|\mathcal{T}^i_T|$. From Equation (8), we get

$$\begin{aligned}
\Delta_{i,i_*} &\leq \widehat{\Delta}^t_{i,i_*} + \frac{\Delta_{i,i_*}}{C} + \frac{\mathcal{V}_{i_*}}{\sqrt{|\mathcal{T}^{i_*}_t|}} \\
\implies \Delta_{i,i_*} &\leq \frac{C}{C-1}\left(\widehat{\Delta}^t_{i_*,i} + \frac{\mathcal{V}_{i_*}}{\sqrt{|\mathcal{T}^{i_*}_t|}}\right) \\
&\leq \frac{C\mathcal{V}_{i_*}}{(C-1)\sqrt{|\mathcal{T}^{i_*}_t|}}
\end{aligned}$$

where the last inequality follows because the condition in the test has not yet been violated. More-over, since model i_* has not been selected yet, we have $|\mathcal{T}^{i_*}_t| \geq \frac{t^{1-\kappa}}{8L} \geq \frac{|\mathcal{T}^i_t|^{1-\kappa}}{8L}$. This gives us

$$\begin{aligned}
\Delta_{i,i_*} &\leq \frac{8(CL)^{1/2}\mathcal{V}_{i_*}}{\sqrt{C-1}|\mathcal{T}^i_t|^{\frac{1-\kappa}{2}}} \\
\implies |\mathcal{T}^i_t| &= \mathcal{O}\left(\frac{L^{\frac{1}{1-\kappa}}(\mathcal{V}_{i_*})^{\frac{2}{1-\kappa}}}{\Delta_{i,i_*}^{\frac{2}{1-\kappa}}}\right)
\end{aligned}$$

Thus, the total contribution to the regret from the misspecified model i is given by

$$\begin{aligned}
&T^{1-\kappa} + |\mathcal{T}^i_t|\Delta_{i,i_*} + \mathcal{R}^{\Pi_i}_i\sqrt{|\mathcal{T}^i_t|} \\
&\leq T^{1-\kappa} + |\mathcal{T}^i_t|\Delta_{i,i_*} + \mathcal{R}^{\Pi_{i_*}}_{i_*}\sqrt{|\mathcal{T}^i_t|}.
\end{aligned}$$

The first term comes from the forced exploration, and the last term is equivalent to the regret we would pay anyway if we knew $i_* = 2$ beforehand. Focusing on the second term, the contribution to regret is upper bounded by

$$\min\left\{\Delta_{i,i_*}T, \left(\frac{C_{\mathcal{Z}}L^{1/2}\mathcal{V}_{i_*}}{\Delta_{i,i_*}}\right)^{\frac{2}{1-\kappa}} \cdot \Delta_{i,i_*}\right\}$$

□

D.2 Implications for a fast rate of estimating V^*

An alternative setting is one where we have access to an estimator of V^* instead of an estimator of the gap. Corollary 1 of Kong et al. (2020) shows that an ϵ-close approximation of V^* is possible in $\widetilde{O}\left(\sqrt{d}/\epsilon^2\right)$ interactions in the disjoint linear bandit setting (where there is a different parameter vector for each arm) under Gaussian assumptions. Whether or not such fast estimators exist or are practical for other general settings is still open, but future work on this problem could be applied to the instance dependent results here.

We will retain the same problem assumptions as the previous subsection. We also assume there is $\widehat{V}_i$ for each $i \in [L]$. Each estimator offers a high-probability guarantee on the estimation error as a function of the number of exploration episodes using corresponding exploration algorithms $\{\widetilde{\mathcal{A}}_i\}$.

Assumption 2. *For all $i \in [L]$, we define the $\widehat{V}_i^{(n)}$ where $n \in [T]$ as the estimator of V_i^* given n exploration rounds with $\widetilde{\mathcal{A}}_i$. We assume with probability at least $1-\delta$, for all $i \geq i_*$, the estimator $\widehat{V}_i^{(n)}$ satisfies*

$$|V^* - \widehat{V}_i^{(n)}| \leq \frac{\mathcal{V}_i}{n^\alpha} + \frac{\mathcal{V}_i'}{n^\beta} \tag{10}$$

where $\mathcal{V}_i$ and $\mathcal{V}_i'$ are poly and increasing in d, H, $|\mathcal{U}|$, and $\log(LT/\delta))$ and $\alpha, \beta \in (0,1)$.

Let $\hat{V}_i^t := \widehat{V}_i^{(|\mathcal{T}_t^i|)}$. The algorithm will be of the same form as Algorithm 2, but instead we leverage the following alternative test:

$$\sum_{t \in \mathcal{T}_t^{\hat{\imath}_t}} \widehat{V}_j^t - g_{t'} \leq \mathcal{Z}_{\hat{\imath}}(|\mathcal{T}_t^{\hat{\imath}_t}|, \mathcal{V}_j, \mathcal{V}_j') \tag{11}$$

where

$$\mathcal{Z}_i(t, \mathcal{V}, \mathcal{V}') := C_{\mathcal{Z}} \left(\mathcal{V}_j L^\alpha t^{1-(1-\kappa)\alpha} + \mathcal{V}_j' L^\beta t^{1-(1-\kappa)\beta} + H\sqrt{t \log(1/\delta)} + \mathcal{R}_i^{\Pi_i}\sqrt{t} \right)$$

for a sufficiently large constant $C_{\mathcal{Z}} > 0$. That is, if the above inequality holds, then ECE continues to use $\hat{\imath}_t$; otherwise, ECE switches to $\hat{\imath}_t + 1$ for round $t+1$. First, we prove an analogous result to Lemma 1, showing that the test will not fail under the good event E''. Here, we let $E'' = E' \cap E_4$ where E' is the event from Theorem 2 and event E_4 is the following.

Event E_4: Let $\{\widehat{V}_i\}$ be the estimators from Assumption 2. For all $i \geq i_*$ and $n \in [T]$, equation (10) is satisfied.

Note that E_4 holds with probability at least $1-\delta$ by assumption. Therefore E'' still holds with probability at least $1 - 10LT^2\delta\log_2(T)$.

Lemma 6. *Given that event E' holds, then for all $t \geq \tau_{\min}$ and $j \in [i_*+1, L]$, it holds that $\sum_{t' \in \mathcal{T}_t^{i_*}} \hat{V}_t^j - g_{t'} \leq \mathcal{Z}_{i_*}(|\mathcal{T}_t^{i_*}|, \mathcal{V}_j, \mathcal{V}_j')$*

Proof. Since $j > i_*$, we use the assumption on the estimator $\widehat{V}_j$ to write the difference in terms of regret, estimation error and noise:

$$\begin{aligned}\sum_{t' \in \mathcal{T}_t^{i_*}} \widehat{V}_j^t - g_{t'} &\leq \sum_{t' \in \mathcal{T}_t^{i_*}} \widehat{V}_j^t - V^{\pi_{t'}} - \epsilon_{t'} \\ &\leq \frac{\mathcal{V}_j |\mathcal{T}_t^{i_*}|}{|\mathcal{T}_t^j|^\alpha} + \frac{\mathcal{V}_j' |\mathcal{T}_t^{i_*}|}{|\mathcal{T}_t^j|^\beta} + \sum_{t' \in \mathcal{T}_t^{i_*}} V^* - V^{\pi_{t'}} - \epsilon_{t'}\end{aligned}$$

Then note that $\sum_{t' \in |\mathcal{T}_t^{i_*}|} \epsilon_{t'} \leq H\sqrt{2|\mathcal{T}_t^{i_*}|\log(2/\delta)}$ and $\sum_{t' \in |\mathcal{T}_t^{i_*}|} V^* - V^{\pi_{t'}} \leq \mathcal{R}_{i_*}^{\Pi_{i_*}}\sqrt{|\mathcal{T}_t^{i_*}|}$ under event E'. Furthermore, under E', we have $|\mathcal{T}_t^j| \geq \frac{t^{1-\kappa}}{8L} \geq \frac{|\mathcal{T}_t^{i_*}|^{1-\kappa}}{8L}$, which implies

$$\sum_{t' \in \mathcal{T}_t^{i_*}} \widehat{V}_j^t - g_{t'} \leq C_{\mathcal{Z}} \left(\mathcal{V}_j L^\alpha |\mathcal{T}_t^{i_*}|^{1-(1-\kappa)\alpha} + \mathcal{V}_j' L^\beta |\mathcal{T}_t^{i_*}|^{1-(1-\kappa)\beta} + H\sqrt{|\mathcal{T}_t^{i_*}|\log(2/\delta)} + \mathcal{R}_{i_*}^{\Pi_{i_*}}\sqrt{|\mathcal{T}_t^{i_*}|} \right)$$

for $C_{\mathcal{Z}}$ large enough. Therefore, it holds that $\sum_{t' \in \mathcal{T}_t^{i_*}} \widehat{V}_j^t - g_{t'} \leq \mathcal{Z}_{i_*}(|\mathcal{T}_t^{i_*}|, \mathcal{V}_j, \mathcal{V}_j')$. □

The main proposition states that a better instance-dependent rate is available under less restrictive assumptions on "realizability" by utilizing the test based on the V^* estimators.

Proposition 3. *For a given $\mathcal{M}$, let Assumption 2 hold some for α, β and $i \geq i_*$ and let $\kappa \in (0, 1/2]$. Then, with probability at least $1 - \delta'$, ECE in Algorithm 1 with the modified test (Equation 11) satisfies the regret bound*

$$\widetilde{O}\left(HLT^{1-\kappa} + \mathcal{R}_{i_*}^{\Pi_{i_*}}\sqrt{LT} + \sum_{j<i_*}\Delta_{j,i_*}\max\left\{\frac{L^{\frac{1}{1-\kappa}}\mathcal{V}_{i_*}^{\frac{1}{(1-\kappa)\alpha}}}{\Delta_{j,i_*}^{\frac{1}{(1-\kappa)\alpha}}},\ \frac{L^{\frac{1}{1-\kappa}}\mathcal{V}_{i_*}'^{\frac{1}{(1-\kappa)\beta}}}{\Delta_{j,i_*}^{\frac{1}{(1-\kappa)\beta}}},\ \frac{(\mathcal{R}_{i_*}^{\Pi_{i_*}} + H\log^{1/2}(LT/\delta'))^2}{\Delta_{j,i_*}^2}\right\}\right)$$

Proof. As discussed previously, the sufficient events occur with probability at least $1 - \delta'$. Similar to Theorem 2, we now show that the gaps Δ_{j,i_*} can be bounded by using the estimation error of $\hat{V}^{i_*}$ and the concentration bounds from E'. Let t be such that $\hat{\imath}_t = j$ and the test succeeds. Then,

$$\begin{aligned}
\Delta_{j,i_*} &= V^* - V_j^* \\
&\leq \widehat{V}_{i_*}^t + \frac{\mathcal{V}_{i_*}}{|\mathcal{T}_t^{i_*}|^\alpha} + \frac{\mathcal{V}_{i_*}'}{|\mathcal{T}_t^{i_*}|^\beta} - \frac{1}{|\mathcal{T}_t^j|}\sum_{t' \in \mathcal{T}_t^j} V^{\pi_{t'}} \\
&\leq \widehat{V}_{i_*}^t + \frac{\mathcal{V}_{i_*}}{|\mathcal{T}_t^{i_*}|^\alpha} + \frac{\mathcal{V}_{i_*}'}{|\mathcal{T}_t^{i_*}|^\beta} - \frac{1}{|\mathcal{T}_t^j|}\sum_{t' \in \mathcal{T}_t^j} g_{t'} + \frac{1}{|\mathcal{T}_t^j|}\sum_{t' \in \mathcal{T}_t^j} \epsilon_{t'} \\
&\leq C_{\mathcal{Z}}\left(\mathcal{V}_{i_*}L^\alpha|\mathcal{T}_t^j|^{-(1-\kappa)\alpha} + \mathcal{V}_{i_*}'L^\beta|\mathcal{T}_t^j|^{-(1-\kappa)\beta} + H\sqrt{\frac{\log(1/\delta)}{|\mathcal{T}_t^j|}} + \frac{\mathcal{R}_{i_*}^{\Pi_{i_*}}}{\sqrt{|\mathcal{T}_t^j|}}\right) + \frac{\mathcal{V}_{i_*}}{|\mathcal{T}_t^{i_*}|^\alpha} + \frac{\mathcal{V}_{i_*}'}{|\mathcal{T}_t^{i_*}|^\beta} + H\sqrt{\frac{\log(1/\delta)}{|\mathcal{T}_t^j|}}
\end{aligned}$$

Again noting that $|\mathcal{T}_t^{i_*}| \geq \frac{t^{1-\kappa}}{8L} \geq \frac{|\mathcal{T}_t^j|^{1-\kappa}}{8L}$, the above can be simplified to

$$\begin{aligned}
\Delta_{j,i_*} &\leq C_{\mathcal{Z}}' \cdot \left(2\mathcal{V}_{i_*}L^\alpha|\mathcal{T}_t^j|^{-(1-\kappa)\alpha} + 2\mathcal{V}_{i_*}'L^\beta|\mathcal{T}_t^j|^{-(1-\kappa)\beta} + \frac{2H\log^{1/2}(1/\delta) + \mathcal{R}_{i_*}^{\Pi_{i_*}}}{|\mathcal{T}_t^j|^{1/2}}\right) \\
&\leq 6C_{\mathcal{Z}}' \cdot \max\left\{\frac{\mathcal{V}_{i_*}L^\alpha}{|\mathcal{T}_t^j|^{(1-\kappa)\alpha}},\ \frac{\mathcal{V}_{i_*}'L^\beta}{|\mathcal{T}_t^j|^{(1-\kappa)\beta}},\ \frac{H\log^{1/2}(1/\delta) + \mathcal{R}_{i_*}^{\Pi_{i_*}}}{|\mathcal{T}_t^j|^{1/2}}\right\}
\end{aligned}$$

where $C_{\mathcal{Z}}' = \max\{1, C_{\mathcal{Z}}\}$. Then, we can consider the three potential cases to upper bound $|\mathcal{T}_t^j|$. Depending on the maximal term, one of the three possible cases occurs:

$$|\mathcal{T}_t^j| \leq \left(\frac{6C_{\mathcal{Z}}'\mathcal{V}_{i_*}L^\alpha}{\Delta_{j,i_*}}\right)^{\frac{1}{(1-\kappa)\alpha}}, \quad |\mathcal{T}_t^j| \leq \left(\frac{6C_{\mathcal{Z}}'\mathcal{V}_{i_*}'L^\beta}{\Delta_{j,i_*}}\right)^{\frac{1}{(1-\kappa)\beta}}, \quad |\mathcal{T}_t^j| \leq \left(\frac{6C_{\mathcal{Z}}'(H\log^{1/2}(1/\delta) + \mathcal{R}_{i_*}^{\Pi_{i_*}})}{\Delta_{j,i_*}}\right)^2$$

The regret during the misspecified phase becomes

$$\begin{aligned}
&\text{Regret}_{\tau_{\min}(\delta):\tau_*} \\
&= O\left(HLT^{1-\kappa} + Hi_* + \mathcal{R}_{i_*}^{\Pi_{i_*}}\sqrt{LT} + \sum_{j<i_*}\Delta_{j,i_*}\max\left\{\frac{L^{\frac{1}{1-\kappa}}\mathcal{V}_{i_*}^{\frac{1}{(1-\kappa)\alpha}}}{\Delta_{j,i_*}^{\frac{1}{(1-\kappa)\alpha}}},\ \frac{L^{\frac{1}{1-\kappa}}\mathcal{V}_{i_*}'^{\frac{1}{(1-\kappa)\beta}}}{\Delta_{j,i_*}^{\frac{1}{(1-\kappa)\beta}}},\ \frac{(\mathcal{R}_{i_*}^{\Pi_{i_*}} + H\log^{1/2}(LT/\delta'))^2}{\Delta_{j,i_*}^2}\right\}\right)
\end{aligned}$$

The total regret is

$$\begin{aligned}
&O\left(HL^{\frac{2}{1-\kappa}}\log^{\frac{1}{1-\kappa}}(1/\delta) + HLT^{1-\kappa} + Hi_*\right) \\
&\quad + O\left(\mathcal{R}_{i_*}^{\Pi_{i_*}}\sqrt{LT} + \sum_{j<i_*}\Delta_{j,i_*}\max\left\{\frac{L^{\frac{1}{1-\kappa}}\mathcal{V}_{i_*}^{\frac{1}{(1-\kappa)\alpha}}}{\Delta_{j,i_*}^{\frac{1}{(1-\kappa)\alpha}}},\ \frac{L^{\frac{1}{1-\kappa}}\mathcal{V}_{i_*}'^{\frac{1}{(1-\kappa)\beta}}}{\Delta_{j,i_*}^{\frac{1}{(1-\kappa)\beta}}},\ \frac{(\mathcal{R}_{i_*}^{\Pi_{i_*}} + H\log^{1/2}(LT/\delta'))^2}{\Delta_{j,i_*}^2}\right\}\right)
\end{aligned}$$

□

Consider again the implications of this bound in the contextual bandit setting. It is possible that to estimate an upper bound of V^* with rate $\widetilde{O}\left(\frac{d_j^{1/4}}{n^{1/2}} + \frac{1}{n^{1/4}}\right)$, where n is the number of samples and $j \geq i_*$ (Foster et al., 2019; Kong and Valiant, 2018). However, this would only give a one-sided estimation error bound. If a two-sided guarantee of the same form were possible, we would have $\alpha = 1/2$, $\beta = 1/4$, and $\mathcal{V}_{i_*} = \widetilde{O}\left(d^{1/4}\right), \mathcal{V}'_{i_*} = \widetilde{O}\left(1\right)$. We now state the following immediate corollary in this setting with constant gaps under the hypothesis that such an estimator for this problem exists and is given.

Corollary 8. *For the linear contextual bandit problem under Assumption 2 with constant gaps $\{\Delta_{j,i_*}\}_{j<i_*}$, let $\alpha = 1/2$, $\beta = 1/4$, $\mathcal{V}_{i_*} = \widetilde{O}(d_{i_*}^{1/4})$ and $\mathcal{V}'_{i_*} = \widetilde{O}(1)$. Let the exploration parameter $\kappa = 1/2$. Then with probability at least $1 - \delta'$, ECE in Algorithm 1 with the modified test (Equation 11) satisfies the regret bound*

$$\widetilde{O}\left(\sqrt{T} + \sqrt{d_{i_*}T} + \sum_{j<i_*} \max\left\{d_{i_*}\Delta_{j,i_*}^{-3},\ \Delta_{j,i_*}^{-7},\ d_{i_*}\Delta_{j,i_*}^{-1}\right\}\right) = \widetilde{O}\left(\sqrt{T} + \sqrt{d_{i_*}T} + d_{i_*}\right)$$

where $\widetilde{O}$ hides dependence on the number of models L, the number of actions $K = |\mathcal{U}|$, and log factors.

For constant gaps, the scalings in d and T are nearly same for this estimator and the gap estimator of the previous section. The main difference arises in the dependence on the gap, $O(\Delta_{\min}^{-5})$ in this case compared to $O\left(\Delta_{\min}^{-2}\right)$ in the previous case. In this case, it is clearly suboptimal.