A NOTE ON RECENT ADVANCES IN RL WITH GENERAL FUNCTION APPROXIMATION

ABSTRACT

-

1 TL;DR

- GEC reduces out-of-sample V_{1,f^t} to in-sample error estimation:
  1 A low DEC: model-based + model-free;
  2 An effective in-sample error estimator;
  3 Handle the difference between $V_{1,f}$ and V_1^*;

$$\mathrm{Reg}(T) \lesssim \left[d_{\mathrm{GEC}} \cdot \sum_{t=1}^{T}\sum_{s=1}^{t-1} \ell^s(f^t)\right]^{1/2} \leq \underbrace{\gamma \sum_{t=1}^{T}\sum_{s=1}^{t-1} \ell^s(f^t)}_{\text{New target: in-sample estimation}} + \frac{1}{\gamma} \cdot d_{\mathrm{GEC}}.$$

- DEC reduces out-of-sample V_1^* to another out-of-sample target:
  1 A low DEC: model-based;
  2 An effective online learning oracle;

$$\mathbb{E}\mathrm{Reg}(T) \leq \underbrace{\sum_{t=1}^{T} \mathrm{dec}_\gamma^H(\mathcal{M}, \mu^t)}_{\text{Cost of transformation}} + \underbrace{\gamma \cdot \sum_{t=1}^{T} \mathbb{E}_{\pi_t \sim p^t} \mathbb{E}_{\widehat{M}_t \sim \mu^t} \left[D^{\pi_t}\left(\widehat{M}_t || M^*\right)\right]}_{\text{New target: online learning}}.$$

- DC/O-DEC reduces out-of-sample V_{1,f^t} to another optimistic out-of-sample target:
  1 A low complexity measure: model-based + model-free;
  2 An effective online learning oracle;
  3 Handle the difference between $V_{1,f}$ and V_1^*.

$$\mathbb{E}\mathrm{Reg}(T) \leq \underbrace{\sum_{t=1}^{T} \mathrm{odec}_\gamma^D(\mathcal{M}, \mu^t)}_{\text{Cost of transformation}} + \underbrace{\gamma \cdot \sum_{t=1}^{T} \mathbb{E}_{\pi_t \sim p^t} \mathbb{E}_{\widehat{M}_t \sim \mu^t} \left[D^{\pi_t}\left(\widehat{M}_t || M^*\right) - \gamma^{-1} \Delta V_{1,\widehat{M}_t}(x_1)\right]}_{\text{New target: online learning with feel-good term}}.$$

2 INTRODUCTION

One of the core problem in reinforcement learning is to identify the minimal structural assumption that permits sample-efficient learning. While the tabular MDP has been well studied, the minimax regret bound of tabular setting depends on the number of state S. This suggests that MDPs with large state space cannot be handled without further structural assumptions. Therefore, a line of works is devoted to the function approximation setting where we approximate the value function, policy, or model dynamic by an abstract hypothesis set $\mathcal{H}$.

Motivated by the results from supervised learning, one may expect that (1) realizability (the true model $f^* \in \mathcal{H}$) and (2) bounded statistical complexity (e.g., the covering number of $\mathcal{H}$) are sufficient for RL. Unfortunately, there exists a negative result, saying that learning a good policy is statistically hard even though the following two conditions are satisfied:

- for all $h \in [H]$, there exists $\theta_h^* \in \mathbb{R}^d$ such that $Q_h^*(\cdot, \cdot) = \phi(\cdot, \cdot)^\top \theta_h^*$;

- Constant gap: for all $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$, $\Delta_h(s,a) := V_h^*(s) - Q_h^*(s,a) \geq \Delta_{\min} > 0$ (Intuitively speaking, there exists a gap between the best action and the second-best one so learning is easier).

Therefore, we want to identify some natural additional structural assumption that permits sample-efficient learning for RL with general function approximation. In particular, we are interested in the following measures proposed recently:

- Decision-Estimation Coefficient (DEC) (Foster et al., 2021; 2022);
- Decoupling Coefficient (DC) (Zhang, 2022; Agarwal and Zhang, 2022);
- Generalized Eluder Coefficient (GEC) (Zhong et al., 2022).

For all the following discussions, we make the following standard realizability assumption.

Assumption 1 (Realizability)**.** *We assume that $f^* \in \mathcal{H}$.*

Learning objective. We consider the regret minimization problem:

$$\mathbb{E}\mathrm{Reg}(T) = \sum_{t=1}^{T} \mathbb{E}_{f^t \sim p^t} V_1^*(x_1) - V_1^{\pi_t}(x_1).$$

Notation. In general, the hypothesis class can be either value-based one where we approximate the optimal Q-value function or model-based one where we approximate the model dynamics. We will use $M \in \mathcal{M}$ to replace $\mathcal{H}$ when the hypothesis is a model-based set. To describe the discrepancy of a model $M \in \mathcal{M}$, we use the notation $V_{1,M}^{\pi}(x_1)$ to denote the value of policy π if M is the true model. If we omit the superscript π, it means that we take $\pi := \pi_M$. We also define the Bellman operator and Bellman residual as follows:

$$\begin{aligned} Q_h^*(x,a) &= (\mathcal{T}_h V_{h+1}^*)(x,a) := r_h(x,a) + \mathbb{E}_{x' \sim \mathbb{P}_h(\cdot|x,a)}; \\ \mathcal{E}_h(f,x,a) &= Q_{h,f}(x,a) - (\mathcal{T}_h V_{h+1,f})(x_h,a_h). \end{aligned} \tag{1}$$

Sometimes we will use $\zeta = \{(x_h,a_h,r_h)\}_{h=1}^H$ to denote a trajectory and we also write $\mathcal{E}_h(f,\zeta)$, meaning that we take x_h, a_h as the input. Given two distributions $\mathbb{P}$ and $\mathbb{Q}$, we also define the Hellinger distance as follows:

$$D_H^2(\mathbb{P},\mathbb{Q}) := \int \left(\sqrt{d\mathbb{P}} - \sqrt{d\mathbb{Q}}\right)^2. \tag{2}$$

3 Generalized Eluder Coefficient (GEC)

In this section, we consider the function approximation by either a valued-based $\mathcal{H}$ or a model-based $\mathcal{H}$. To motivate the GEC, we start with the following value decomposition lemma from Jiang et al. (2017).

Lemma 1 (Value Decomposition Lemma (Jiang et al., 2017))**.** *For any algorithm that achieves optimism such that $V_{1,f^t}(x_1) > V_1^*(x_1)$, by the value decomposition lemma, we have*

$$\begin{aligned} \sum_{h=1}^{T} V_1^*(x_1) - V_1^{\pi_{f^t}}(x_1) &= \sum_{h=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi_{f^t}}\left[\mathcal{E}_h\left(f^t, x_h^t, a_h^t\right)\right] - \underbrace{\left(V_{1,f^t}(x_1) - V_1^*(x_1)\right)}_{\Delta V_{1,f^t}(x_1)} \\ &\leq \sum_{h=1}^{T}\sum_{h=1}^{H} \mathbb{E}_{\pi_{f^t}}\left[\mathcal{E}_h\left(f^t, x_h^t, a_h^t\right)\right]. \end{aligned} \tag{3}$$

The last inequality is the main technical reason why we consider global optimism-based algorithms. However, if we carefully look at the RHS of (3), we can see a "mismatch" between Goal and Guarantee:

- Guarantee: f^t is good on the historical dataset: $\mathcal{D}^{t-1} = \{\zeta^1, \zeta^2, \cdots, \zeta^{t-1}\}$;
- Goal: f^t performs well on the unseen ζ^t.

Clearly, this requires certain "generalization" in an online manner so that a hypothesis consistent with the historical data should perform well for the future. The GEC captures the hardness of such a generalization, defined as follows.

Definition 1 (Generalized Eluder Coefficient (Zhong et al., 2022)). *Given a hypothesis class $\mathcal{H}$, a discrepancy function $\ell = \{\ell_f\}_{f\in\mathcal{H}}$, an exploration policy class $\Pi_{\exp}$, and $\epsilon > 0$, the generalized eluder coefficient $\mathrm{GEC}(\mathcal{H}, \ell, \Pi_{\exp}, \epsilon)$ is the smallest d ($d \geq 0$) such that for any sequence of hypotheses $\{f^t\}_{t=1}^T$:*

$$\sum_{h=1}^{T}\sum_{h=1}^{H}\underbrace{\mathbb{E}_{\pi_{f^t}}\left[\mathcal{E}_h\left(f^t, x_h^t, a_h^t\right)\right]}_{\textit{test error}} \leq \left[d\sum_{h=1}^{H}\sum_{t=1}^{T}\underbrace{\left(\sum_{s=1}^{t-1}\mathbb{E}_{\pi_{\exp}(f^s,h)}\ell_{f^s}(f^t,\zeta_h)\right)}_{\textit{training error}}\right]^{1/2} + \underbrace{2\sqrt{dHT}+\epsilon HT}_{\textit{burn-in cost}}. \tag{4}$$

In general, we can take $\epsilon > 0$ sufficiently small such that

$$\sum_{h=1}^{T}\sum_{h=1}^{H}\mathbb{E}_{\pi_{f^t}}\left[\mathcal{E}_h\left(f^t, x_h^t, a_h^t\right)\right] \lesssim \left[d_{\mathrm{GEC}}\sum_{h=1}^{H}\sum_{t=1}^{T}\left(\sum_{s=1}^{t-1}\mathbb{E}_{\pi_{\exp}(f^s,h)}\ell_{f^s}(f^t,\zeta_h)\right)\right]^{1/2}.$$

Motivation. On average, if $f^t \in \mathcal{H}$ is consistent with the historical data, then the test error on unseen t-th trajectory will also be small (but is amplified by GEC). In particular, based on GEC, the regret minimization problem is reduced to a sequential (optimistic) estimation problem.

Algorithm 1 Eluder TS

for $t = 1, \cdots, T$ **do**
  1 Optimistic planning: $f^t \sim p^t(f)$
    $p^t(f) \propto p^0(f)\exp(\gamma V_{1,f}(x_1))\exp(\sum_{h=1}^{H} -L_h^{1:t-1}(f))$;
  2 Data collection: for each $h \in [H]$,
    execute $\pi_{\exp}(f^t, h)$ for $\mathrm{N}_{\mathrm{batch}}$ times;
  3 Update the posterior p^t.
end for

Algorithm 2 Eluder UCB

for $t = 1, \cdots, T$ **do**
  1 Optimistic planning: $f^t := \operatorname{argmax}_{f\in\mathcal{B}^t} V_{1,f}(x_1)$
    $\mathcal{B}^t := \{\sum_{h=1}^{H} L_h^{1:t-1}(f) \leq \beta_t^2\}$;
  2 Data collection: for each $h \in [H]$,
    execute $\pi_{\exp}(f^t, h)$ for $\mathrm{N}_{\mathrm{batch}}$ times;
  3 Update the confidence set $\mathcal{B}^t$.
end for

Algorithmic design. The sequential (optimistic) estimation problem, which can be achieved by either Thompson sampling (Algorithm 1) or UCB (Algorithm 2). It remains to choose appropriate loss estimator $L_h^{1:t-1}(f)$ so that it converges to the training error $\sum_{s=1}^{t-1}\mathbb{E}_{\pi_{\exp}(f^s,h)}\ell_{f^s}(f,\zeta_h)$. We shall see that a majority of algorithmic choices scattered in the literature share this goal. We take the Q-type value-based problem and Q-type model-based problem as examples.

Remark 1. *We highlight that there are indeed two notions in Algorithm 1 and 2. We use the Eluder TS algorithm as an example. After the first step, we have constructed a distribution of the hypothesis. Then, such a distribution induces a distribution for the policy:*

$$q^t(\pi) := p^t\left(\{f \in \mathcal{H} | \pi_{\exp(f,h)} = \pi_h, \forall h \in [H]\}\right),$$

We refer such a inducement to as the simple strategy. Although such a mechanism is natural and is general enough to cover a majority of interesting iterative decision making problems, we do remark that it can be suboptimal in some cases due to Foster et al. (2022). We defer a detailed discussion to Section 6.

Example 1 (Q-type value-based problem). *We have $\ell_{f^s}(f^t, \zeta_h) = \mathcal{E}_h(f^t, \zeta_h)^2$ and $\pi_{\exp}(f, h) = \pi_f$. We suffer from the famous double-sampling issue (Antos et al., 2008):*

$$\mathbb{E}_{\pi_{f^s}}[Q_{h,f}(x_h^s, a_h^s) - r_h^s - V_{h+1,f}(x_{h+1}^s)]^2 = \underbrace{\mathbb{E}_{\pi_{f^s}}\mathcal{E}_h(f, x_h^s, a_h^s)^2}_{\textit{Goal}} + \underbrace{\sigma_{h,f}^2}_{\textit{Sampling variance}}.$$

This issue can be addressed by using two independent samples (hence the name) because $\mathbb{E}XY = \mathbb{E}X\mathbb{E}Y = (\mathbb{E}X)^2$ if X and Y are i.i.d.. Two typical strategies exist in the literature.

- *Minimax formulation (GOLF (Jin et al., 2021), Conditional PS (Dann et al., 2021))*[1]

$$L_h^{1:t-1}(f) = -\eta \sum_{s=1}^{t-1} [Q_{h,f}(x_h^s, a_h^s) - r_h^s - V_{h+1,f}(x_{h+1}^s)]^2 - \log \mathbb{E}_{\tilde{f}_h \sim p_h^0(\cdot)} \left[\exp\left(-\eta \sum_{s=1}^{t-1} [Q_{h,\tilde{f}}(x_h^s, a_h^s) - r_h^s - V_{h+1,f}(x_{h+1}^s)]^2 \right) \right],$$

  - *The introduced log term cancels the variance;*
  - *The log term requires Completeness to deal with;*

2 *Trajectory average with m i.i.d. samples (OLIVE (Jiang et al., 2017), BiLin-UCB (Du et al., 2021))*

$$L_h^{1:t}(f) = -\eta \sum_{s=1}^{t} \left(\frac{1}{m} \sum_{i=1}^{m} \left(Q_{h,f}(x_{i,h}^s, a_{i,h}^s) - r_{i,h}^s - V_{h+1,f}(x_{i,h+1}^s) \right) \right)^2;$$

  - *Sample mean admits a smaller variance:* $\mathrm{Var}[\bar{X}_m] = \frac{1}{m}\mathrm{Var}[X]$.

Example 2 (Q-type model-based problem)**.** *We have $\pi_{\exp}(f,h) = \pi_f$ and for all $f^s \in \mathcal{H}$, $\ell_{f^s}(f^t, \zeta_h) = D_H^2\big(\mathbb{P}_{h,f^t}(\cdot|x_h,a_h), \mathbb{P}_{h,f^*}(\cdot|x_h,a_h)\big)$. For model-based problem, we choose*

$$L_h^{1:t}(f) := \sum_{s=1}^{t} L_h^s(f) := \eta \sum_{s=1}^{t} \log \mathbb{P}_{h,f}(x_{h+1}^s \mid x_h^s, a_h^s).$$

Then, we can equivalently use $\Delta L_h^t(f) = L_h^t(f) - L_h^t(f^*)$ in the theoretical analysis (for instance, this will not influence the posterior distribution of Algorithm 1). Then, the MLE analysis directly leads to the desired Hellinger distance (see e.g. Lemma E.5 of Zhong et al. (2022)). Therefore, one only requires realizability for model-based approach.

Remark 2. *In the literature, we have known that model-based approach with realizability can achieve a $\sqrt{T}$-regret without further assumptions (e.g. Agarwal and Zhang (2022) v.s. Dann et al. (2021)). However, we remark that model-based realizability is stronger than the model-free one. Suppose that we are given a model class such that $M^* \in \mathcal{M}$. We can take $\mathcal{H} = \mathcal{H}_1 \times \cdots \mathcal{H}_H$ and $\mathcal{H}_h = \{Q_{h,M} : M \in \mathcal{M}\} \cup \mathcal{T}_h^M \mathcal{H}_{h+1}$ for all $M \in \mathcal{M}$. By doing so, we construct a value-based hypothesis $\mathcal{H}$ satisfying realizability and Bellman completeness and $|\mathcal{H}| = |\mathcal{M}|^2$.*

The point here is that all these approaches are designed for an effective (and sharper) estimation of in-sample training error. In particular, we have the following (informal) results.

Lemma 2. *For model-free approach, we have the following results.*

- *Minimax approach with Bellman completeness* [2]*:*

$$\sum_{s=1}^{t-1} \mathbb{E}_{\pi_{f^s}} \mathcal{E}_h(f^t, \zeta_h)^2 \lesssim \log(|\mathcal{H}|/\delta);$$

- *Trajectory average:*

$$\sum_{s=1}^{t-1} \mathbb{E}_{\pi_{f^s}} \mathcal{E}_h(f^t, \zeta_h)^2 \lesssim \frac{(t-1)}{m} \cdot \log(|\mathcal{H}|/\delta);$$

- *Model-based approach:*

$$\sum_{s=1}^{t-1} \mathbb{E}_{\pi_{f^s}} D_H^2\big(\mathbb{P}_{h,f^t}(\cdot|x_h,a_h), \mathbb{P}_{h,f^*}(\cdot|x_h,a_h)\big) \lesssim \log(|\mathcal{M}|/\delta).$$

[1] Also used in (Antos et al., 2008) and (Chen and Jiang, 2019).

[2] The $\mathcal{T}_h$-completeness can be generalized. See Bellman complete model in Du et al. (2021).

Combining Lemma 2 with Lemma 1 and Definition 1, we know that

Theorem 1 (Zhong et al. (2022)). *Suppose $|\mathcal{H}|$ is finite. With appropriate choices of hyper-parameters, the Eluder TS/UCB admit the following regret bounds:*

1 *Minimax formulation with Realizability + Completeness:* $\tilde{\mathcal{O}}\big(\sqrt{d_{\mathrm{GEC}} \cdot HT \cdot \log|\mathcal{H}|}\big)$;

2 *Trajectory average with Realizability:* $\tilde{\mathcal{O}}\big(d_{\mathrm{GEC}}^{2/3} \cdot H^{1/3}T^{2/3} \cdot (\log|\mathcal{H}|)^{1/3}\big)$;

3 *Model-based approach with Realizability:* $\tilde{\mathcal{O}}\big(\sqrt{d_{\mathrm{GEC}} \cdot HT \cdot \log|\mathcal{M}|}\big)$.

4 Decision-Estimation Coefficient (DEC)

We consider the function approximation with a **model class** $\mathcal{M}$. The original DEC (Foster et al., 2021) is defined as follows.

Definition 2 (Decision-Estimation Coefficient). *Given a class of model, and a reference model $\widehat{M}$ (typically is estimated by the historical dataset), we define*

$$\mathrm{dec}_\gamma(\mathcal{M}, \overline{M}) = \inf_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \mathbb{E}_{\pi\sim p}[\underbrace{V_{1,M}(x_1) - V_{1,M}^{\pi}(x_1)}_{\text{regret of decision}} - \gamma \cdot \underbrace{D_{\mathrm{H}}^2(M(\pi), \widehat{M}(\pi))}_{\text{Easy to control}}], \tag{5}$$

where we further define $\mathrm{dec}_\gamma(\mathcal{M}) := \sup_{\widehat{M}\in\mathcal{M}} \mathrm{dec}_\gamma(\mathcal{M}, \widehat{M})$.

Motivation. Intuitively speaking, for a fixed $M \in \mathcal{M}$ as the true model, DEC converts our target, immediate regret, (not easy to control) to the estimation error, Hellinger distance, (something we know how to control).

- The estimation error is the Hellinger distance between the true model M and our estimated reference model $\overline{M}$;
- DEC is in a worst-case manner, as it can be regarded as the worst-case cost for such a target transformation;
- DEC does not (at least not explicitly) incorporate optimism in the definition. Foster et al. (2022) extends the vanilla DEC to the optimistic version and discuss when such a optimistic modification can help (see Section 6 for details).

To illustrate how DEC works, we now give an upper bound for the Algorithm 3 (original algorithm of Foster et al. (2021) with with Option I).

Theorem 2 (Reduction for DEC). *Suppose that we have access to an online estimation oracle* $\mathbf{Alg}_{\mathrm{Est}} : \mathcal{S}^{t-1} \to \widehat{M}_t \in \mathcal{M}$ *for every* $t \in [T]$. *Then, if we adopt the posterior distribution as the solution of the minimax problem in the definition of DEC, then, we have*

$$\mathbb{E}\mathrm{Reg}(T) \le \underbrace{\sum_{t=1}^{T} \mathrm{dec}_\gamma(\mathcal{M}, \widehat{M}_t)}_{\text{Cost of transformation}} + \underbrace{\gamma \cdot \sum_{t=1}^{T} \mathbb{E}_{\pi_t\sim p^t}[D_H^2\big(M^*(\pi_t), \widehat{M}_t(\pi_t)\big)]}_{\text{New target: online learning}}. \tag{6}$$

Proof. We have

$$\begin{aligned}
\mathbb{E}\mathrm{Reg}(T) &= \sum_{t=1}^{T} \mathbb{E}_{f^t\sim p^t} V_1^*(x_1) - V_1^{\pi_t}(x_1) \\
&= \sum_{t=1}^{T} \mathbb{E}_{\pi_t\sim p^t}[V_1^*(x_1) - V_1^{\pi_t}(x_1)] - \gamma \mathbb{E}_{\pi_t\sim p^t}[D_H^2\big(M^*(\pi_t), \widehat{M}_t(\pi_t)\big)] \\
&\quad + \gamma \cdot \underbrace{\sum_{t=1}^{T} \mathbb{E}_{\pi_t\sim p^t}[D_H^2\big(M^*(\pi_t), \widehat{M}_t(\pi_t)\big)]}_{\mathbf{Est}_H} \\
&\le \sup_{\sup_{M\in\mathcal{M}}} \mathbb{E}_{\pi_t\sim p^t}\Big[V_{1,M}(x_1) - V_{1,M}^{\pi_t}(x_1)] - \gamma D_H^2\big(M(\pi_t), \widehat{M}_t(\pi_t)\big)\Big] + \gamma\cdot \mathbf{Est}_H \\
&= \inf_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \mathbb{E}_{\pi\sim p}\Big[V_{1,M}(x_1) - V_{1,M}^{\pi}(x_1)] - \gamma D_H^2\big(M(\pi), \widehat{M}_t(\pi)\big)\Big] + \gamma\cdot \mathbf{Est}_H \\
&= \sum_{t=1}^{T} \mathrm{dec}_\gamma(\mathcal{M}, \widehat{M}_t) + \gamma\cdot \mathbf{Est}_H,
\end{aligned}$$

where in the first inequality, we use the realizability $M^* \in \mathcal{M}$ and also the worst-case consideration of DEC. □

Algorithm 3 DEC-META

1: **Initialize:** $\gamma > 0$.
2: **for** $t = 1, 2, \cdots, T$ **do**
3: Compute estimate $\widehat{M}_t := \mathbf{Alg}_{\mathrm{Est}}\big(\mathcal{S}^{t-1}\big)$;
4: Solve the minimax problem in the definition of DEC:

$$p^t := \operatorname*{argmin}_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \mathbb{E}_{\pi\sim p}[V_{1,M}(x_1) - V_{1,M}^{\pi}(x_1) - \gamma\cdot D_{\mathrm{H}}^2(M(\pi), \widehat{M}_t(\pi))];$$

5: Execute $\pi_t \sim p^t$ and collect trajectory into $\mathcal{S}^t = \mathcal{S}^{t-1} \cup \{x_h^t, a_h^t, r_h^t\}$.
6: **end for**

Discussions According to the definition of DEC in Definition 2 and the proof of Theorem 2, we can see that two natural extensions are straightforward.

- We can replace the Hellinger distance $D_H^2(\cdot,\cdot)$ with other divergence $D\big(M(\pi)||\widehat{M}(\pi)\big)$:

$$\mathrm{dec}_\gamma^D(\mathcal{M}, \overline{M}) := \inf_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \mathbb{E}_{\pi\sim p}[V_{1,M}(x_1) - V_{1,M}^{\pi}(x_1) - \gamma\cdot D(M(\pi), \widehat{M}(\pi))].$$

  Accordingly, we have the following transformation:

$$\mathbb{E}\mathrm{Reg}(T) \le \sum_{t=1}^{T} \mathrm{dec}_\gamma^D(\mathcal{M}, \widehat{M}_t) + \gamma\cdot \sum_{t=1}^{T} \mathbb{E}_{\pi_t\sim p^t} D(M(\pi), \widehat{M}_t(\pi));$$

- We can further replace the point estimate $\widehat{M}_t$ with a randomized estimator with distribution ν and we use $\underline{\mathrm{dec}}$ to highlight such a randomization feature:

$$\underline{\mathrm{dec}}_\gamma^D(\mathcal{M}, \nu) := \inf_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \mathbb{E}_{\pi\sim p}[V_{1,M}(x_1) - V_{1,M}^{\pi}(x_1) - \gamma\cdot \mathbb{E}_{\widehat{M}\sim\nu} D(M(\pi), \widehat{M}(\pi))].$$

  Accordingly, we have the following transformation:

$$\mathbb{E}\mathrm{Reg}(T) \le \sum_{t=1}^{T} \mathrm{dec}_\gamma^D(\mathcal{M}, \nu_t) + \gamma\cdot \sum_{t=1}^{T} \mathbb{E}_{\pi_t\sim p^t}\mathbb{E}_{\widehat{M}_t\sim\nu_t} D(M(\pi), \widehat{M}_t(\pi)).$$

Also, we would like to highlight that a direct motivation for us to consider the extension is: similar to the discussions for GEC in Section 3 and the bilinear class (Du et al., 2021), the flexible choices of loss function can cover more approaches and problems, and also can lead to better estimation rate in some situations, which has been shown by Chen et al. (2022); Foster et al. (2022). For instance, it seems that the vanilla DEC (Foster et al., 2021) CANNOT handle model-free approach even though we replace the Hellinger distance with some suitable divergence (e.g. squared Bellman error) [3].

We defer a detailed discussion to Section 6 and first introduce a closely related notion of decoupling coefficient (Zhang, 2022), whose techniques are later leveraged to the extensions of DEC.

5 Decoupling Coefficient

We consider the contextual bandit problem with a **value-based** hypothesis space $\mathcal{H} := \{f : \mathcal{X} \times \mathcal{A} \to [-b, b]\}$ [4]. We also use the convention that $f(x) := \max_{a\in\mathcal{A}} f(x, a)$ and $a(f, x) \in \operatorname{argmax}_{\tilde{a}\in\mathcal{A}} f(x, \tilde{a})$. The DC shares similar spirits with the DEC to convert the immediate regret to something that is easier to deal with, defined as follows.

Definition 3. *Given $x \in \mathcal{X}$ and $\mathcal{H}' \subset \mathcal{H}$, we define $\mathrm{dc}(x, \mathcal{H}')$ as the smallest $K > 0$ such that for all $p \in \Delta(\mathcal{H}')$ and the induced random policy*

$$\pi_p(\tilde{a}|x) = \mathbb{E}_{f\sim p(\cdot)} I\big(\tilde{a} \in \operatorname*{argmax}_{a\in\mathcal{A}} f(x, a)\big),$$

we have the following immediate regret decoupling:

$$\begin{aligned}\mathbb{E}_{f\sim p(\cdot)}\mathrm{reg}_t &:= \mathbb{E}_{f\sim p(\cdot)}[f^*(x) - f^*\big(x, a(f, x)\big)] \\ &\le \mathbb{E}_{f\sim p(\cdot)} \underbrace{f(x) - f^*(x)}_{\Delta f(x)} + \inf_{\mu>0}\Big[\mu\mathbb{E}_{\tilde{a}\sim\pi_p(\tilde{a}|x)}\mathbb{E}_{f\sim p(\cdot)}\big(f(x, \tilde{a}) - f^*(x, \tilde{a})\big)^2 + \frac{K}{4\mu}\Big].\end{aligned}$$

We proceed to write down the decomposition for T rounds.

$$\mathbb{E}\mathrm{Reg}(T) \le \underbrace{\sum_{t=1}^{T} \frac{\mathrm{dc}(x^t, \Omega)}{4\mu}}_{\text{Transformation cost}} + \underbrace{\sum_{t=1}^{T}\Big[\mu\mathbb{E}_{\tilde{a}\sim\pi_{p^t}(\tilde{a}|x^t)}\mathbb{E}_{f^t\sim p^t(\cdot)}\big(f^t(x^t, \tilde{a}) - f^*(x^t, \tilde{a})\big)^2 + \mathbb{E}_{f^t\sim p^t(\cdot)}\Delta f^t(x^t)\Big]}_{\text{Optimistic Target}},$$

where $\Delta f^t(x^t)$ is referred to as the feel-good term used to compensate the difference between $f^*(x^t)$ and $f(x^t)$.

Motivation. Similar to the DEC, the main motivation for DC is to convert the regret minimization reg_t to another target that can be handled rather easily. The key feature is that in the new optimistic target, the action and the hypothesis are *independently* sampled from their posterior distributions, which can be handled with standard techniques in the analysis of online aggregation algorithms. As compared to DEC,

- DC allows an arbitrary posterior over the hypothesis class and takes the induced random policy $\pi_p(\cdot|x)$, which allows a more flexible choice of algorithms to minimize the new optimistic target;
- DC has an additional feel-good term.

In particular, the presence of the feel-good term is for the optimism at first but is later shown to be beneficial for a large class of problems as noted by Foster et al. (2022) and we will discuss this in Section 6.

Algorithmic design. Algorithm 4 presents the Thompson sampling for contextual bandit, which shares similar structure with Algorithm 1 because they are all based on *exponential weights update*. The main difference is that DC-TS involves the feel-good term as part of the estimation error, which tends out to be critical in some cases, while Eluder-TS adopts an optimistic prior. This difference stems from the different philosophies of the considered complexity measures.

[3] See discussion in Section 3.1 of Foster et al. (2022)

[4] We can handle unbounded reward with refined analysis as in Zhang (2022).

Algorithm 4 DC-TS for Contextual Bandit

1: **for** $t = 1, 2, \cdots, T$ **do**
2: Observe $x^t \in \mathcal{X}$;
3: Sample $f^t \sim p^t(\cdot|\mathcal{S}^{t-1})$ by

$$p^t(f|\mathcal{S}^{t-1}) \propto p^0(f) \cdot \exp\Big(-\sum_{s=1}^{t-1} L(f, x^s, a^s, r^s)\Big), \tag{7}$$

where $L(f, x, a, r) = \eta \cdot \big(f(x, a) - r\big)^2 - \gamma \cdot f(x)$.
4: **end for**

We only present the result when the action space is finite for simplicity.

Theorem 3 (DC-TS)**.** *Suppose that the contextual has a finite action space. Then, suppose further that the boundedness condition is satisfied by* 1*. Then, with appropriate choices of hyper-parameters, we have*

$$\mathbb{E}\mathrm{Reg}(T) \lesssim \sqrt{|\mathcal{A}|T \log |\mathcal{H}|}.$$

Discussion. The DC can be further bounded for linearly embeddable contextual bandit and certain parametric function class. It was later extended to model-based scenario with Hellinger distance (Agarwal and Zhang, 2022). In particular, we would like to highlight that from a technical viewpoint, the requirement of decoupling for arbitrary distribution in DC is unnecessary. Actually, we only require a special class of distribution (corresponding to a special class of exploration policy) to satisfy the decoupling.

6 Optimistic Extension of DEC

The three complexity measures share similar ideas of reducing the hard regret minimization problem to some new target that is easier to handle. Specifically, GEC reduces the out-of-sample test error to the in-sample training error (on average); DC and DEC convert the immediate regret incurred in one iteration to the estimation error of the hypothesis. There are two necessary conditions must be met if we would like to handle problems with such a reduction-based framework:

- The reduction holds with a mild complexity measure (GEC/DC/DEC);
- The new target can be handled (with a sharp rate).

Sometimes, we are in face of a trade-off between these two conditions. The GEC (Zhong et al., 2022) allows a flexible choice of the notion of training error, thus covering both the model-free and model-based approaches. DC has been applied to model-free approach for contextual bandit (Zhang, 2022) and model-based approach for RL (Agarwal and Zhang, 2022), with different choices of "optimistic target". However, the vanilla DEC (Foster et al., 2021) cannot handle the model-free approach[5].

In comparison, one of the key feature of GEC and DC is that they all adopt some optimistic modification to handle the difference between $V_{1,f}$ and V_1^* to encourage exploration, while GEC solves the minimax problem in the definition. The recent conclusions (Foster et al., 2022; Zhong et al., 2022) are as follows:

- In terms of the complexity measure, the optimistic modification, or more general, a flexible choice of target (beyond Hellinger distance) is beneficial for a large class of problems;
- In terms of the algorithmic design, after obtaining the (randomized) hypothesis estimation, direct usage of the induced randomized policy (as in Definition 3) can be suboptimal.

[5] Foster et al. (2021) deal with model-free case via a Bayesian approach. "This allows us to deduce existence of algorithms for the frequentist setting by exhibiting algorithms for the Bayesian setting". As shown by Zhang (2022), this can be done without the optimistic modification. See Section 2.1 of Foster et al. (2021) and page 13 of Foster et al. (2022) for a remark.

In what following, we first extend the DEC to handle the model-free approach. Then, we discuss whether the minimax mechanism to pick the policy is necessary.

Optimistic DEC We now combine DEC with the optimistic modification (Zhang, 2022; Dann et al., 2021; Agarwal and Zhang, 2022; Zhong et al., 2022), as done by Foster et al. (2022).

Definition 4 (Optimistic Decision-Estimation Coefficient (ODEC))**.**

$$\mathrm{odec}_\gamma^D(\mathcal{M},\mu) := \inf_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \mathbb{E}_{\pi\sim p}\mathbb{E}_{\widehat{M}_t\sim\mu^t}[V_{1,\widehat{M}_t}(x_1) - V_{1,M}^\pi(x_1) - \gamma\cdot D^\pi(\widehat{M}_t||M)]. \quad (8)$$

To improve readability and to facilitate discussion, we also write the definition of the vanilla DEC (with randomized estimator and general divergence) here:

$$\mathrm{dec}_\gamma^D(\mathcal{M},\mu) = \inf_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \mathbb{E}_{\pi\sim p}\mathbb{E}_{\widehat{M}_t\sim\mu^t}[V_{1,M}(x_1) - V_{1,M}^\pi(x_1) - \gamma\cdot D^\pi(\widehat{M}_t||M)].$$

In comparison, odec replace the original $V_{1,M}(x_1)$ with $V_{1,\widehat{M}_t}(x_1)$. We shall remark that we regard M as the true model and take a sup to consider the worst-case scenario. Therefore, this is exactly the difference between $V_1^*(x_1)$ and $V_{1,f}(x_1)$ we have mentioned for GEC and DC. Such a feel-good consideration play a critical role for DEC to cover more problems.

Algorithmic design. We again suppose that we have an online oracle such that take the historical dataset $\mathcal{S}^{t-1} = \{\zeta^1,\cdots,\zeta^{t-1}\}$ as an input, and produce a distribution $\mu^t\in\Delta(\mathcal{M})$ as output, where we write:

$$\mu^t = \mathbf{Alg}_{\mathrm{Est}}(\mathcal{S}^{t-1}).$$

For instance, (7) is such an example. For this randomized estimator, we define the *optimistic estimation error* as

$$\mathbf{OptEst}_\gamma^D := \sum_{t=1}^T \mathbb{E}_{\pi^t\sim p^t}\mathbb{E}_{\widehat{M}_t\sim\mu^t}\left[D^\pi\left(\widehat{M}_t||M^*\right) + \gamma^{-1}\underbrace{\left(V_1^*(x_1) - V_{1,\widehat{M}_t}(x_1)\right)}_{\text{Feel-good term}}\right].$$

Algorithm 5 shares similar structure with Algorithm 3, except that we now replace DEC, **Est** with Optimistic DEC and **OptEst**, respectively.

Theorem 4 (Reduction for O-DEC)**.** *Suppose that we have access to an online estimation oracle* $\mathbf{Alg}_{\mathrm{Est}}:\mathcal{S}^{t-1}\to\widehat{M}_t\in\Delta(\mathcal{M})$ *for every* $t\in[T]$*. Then, if we adopt the posterior distribution as the solution of the minimax problem in* (8)*, then, we have*

$$\mathbb{E}\mathrm{Reg}(T) \le \underbrace{\sum_{t=1}^T \mathrm{odec}_\gamma^D(\mathcal{M},\mu^t)}_{\textit{Cost of transformation}} + \underbrace{\gamma\cdot\sum_{t=1}^T \mathbb{E}_{\pi^t\sim p^t}\mathbb{E}_{\widehat{M}_t\sim\mu^t}\left[D^\pi\left(\widehat{M}_t||M^*\right) + \gamma^{-1}\left(V_1^*(x_1) - V_{1,\widehat{M}_t}(x_1)\right)\right]}_{\textit{New target: online learning with feel-good term}}. \quad (9)$$

Algorithm 5 Optimistic-DEC-META

1: **Initialize:** $\gamma>0$.
2: **for** $t=1,2,\cdots,T$ **do**
3: Compute randomized estimator $\mu^t = \mathbf{Alg}_{\mathrm{Est}}(\mathcal{S}^{t-1})$.
4: Solve the minimax problem in the definition of DEC:

$$p^t := \operatorname*{argmin}_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \mathbb{E}_{\pi\sim p}\mathbb{E}_{\widehat{M}_t\sim\mu^t}[V_{1,\widehat{M}_t}(x_1) - V_{1,M}^\pi(x_1) - \gamma\cdot D^\pi(\widehat{M}_t||M)];$$

5: Execute $\pi_t\sim p^t$ and collect trajectory into $\mathcal{S}^t = \mathcal{S}^{t-1}\cup\{x_h^t,a_h^t,r_h^t\}$[6].
6: **end for**

Motivation. As we mentioned before, to handle a RL problem under the reduction framework, we need (1) reduction holds: the DEC is mild and (2) the reduced estimation can be done. The following result shows that there exists some instance where the vanilla DEC does not admit a favorable bound for model-free approach.

Theorem 5 (Separation in Model-free Setting). *Let $\mathcal{M}$ be the class of all horizon-H tabular MDPs with $|\mathcal{X}| = 2$ and $|\mathcal{A}| = 2$. If we adopt $\ell_h^{\mathrm{est}}(Q; z_h) := \big(Q_h(x_h, a_h) - r_h - \max_{a' \in \mathcal{A}} Q_{h+1}(x_{h+1}, a')\big)$, and define the divergence as*

$$D^\pi(Q||M) := \sum_{h=1}^{H} \left(\mathbb{E}_{M,\pi} \ell_h^{\mathrm{est}}(Q; z_h)\right)^2,$$

where we use the induced Q-value of the model as the first input. Then, we have $\sup_\mu \mathrm{odec}_\gamma(\mathcal{M}, \mu) \lesssim \frac{H}{\gamma}$ but there exists $\overline{M} \in \mathcal{M}$ for which $\mathrm{dec}_\gamma(\mathcal{M}, \overline{M}) \gtrsim \frac{2^H}{\gamma} \wedge 1$.

On the other hand, with the feel-good term, DEC can handle the bilinear class with model-free approach, which we defer to the Appendix for details.

7 Discussion

A very interesting observation so far is that the vanilla DEC is sufficient for *model-based* approach, but only the optimistic DEC can handle the *model-free* approach [7]. This question has been addressed in Foster et al. (2022) as follows.

Further notations. In this section, we use the underline to denote the vanilla DEC with randomized estimator:

$$\underline{\mathrm{dec}}_\gamma^D(\mathcal{M}, \mu) = \inf_{p \in \Delta(\Pi)} \sup_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p}[V_{1,M}(x_1) - V_{1,M}^\pi(x_1) - \gamma \cdot \mathbb{E}_{\widehat{M} \sim \nu} D^\pi(M, \widehat{M})].$$

and define $\underline{\mathrm{dec}}_\gamma^D(\mathcal{M}) := \sup_\mu \underline{\mathrm{dec}}_\gamma^D(\mathcal{M})$.

Feel-good term helps for asymmetric divergence We use $\mathrm{co}(\mathcal{X})$ to denote the set of all finitely supported convex combinartions of elements in $\mathcal{X}$. We require the following assumption.

Assumption 2. *For all pairs of models $M, M' \in \mathrm{co}(\mathcal{M})$, there exists some $L > 0$ such that*

$$\left(V_{1,M'}^\pi - V_{1,M}^\pi\right)^2 \le L^2 D^\pi(M'||M).$$

Given a divergence D, we define the *flipped divergence* by swapping the first and second arguments:

$$\check{D}^\pi(M' \| M) := D^\pi(M \| M'). \tag{10}$$

We have the following result.

Lemma 3. *Whenever Assumption 2 holds, we have that for all $\gamma > 0$,*

$$\underline{\mathrm{dec}}_{1.5\gamma}^{\check{D}}(\mathcal{M}) - \frac{L^2}{2\gamma} \le \mathrm{odec}_\gamma^D(\mathcal{M}) \le \underline{\mathrm{dec}}_{0.5\gamma}^{\check{D}}(\mathcal{M}) + \frac{L^2}{2\gamma}.$$

Therefore, ignoring the possible loss of rate, as long as we can control the estimation error with respect to the flipped divergence, this lemma shows that optimism is not necessary. This implies the following result:

Ignoring possible lose of rate, for symmetric convergence, optimism offers no statistical advantage.

The following lemma further shows that under mild assumptions on the divergence D, randomization offers no improvement.

Lemma 4. *Suppose that D is a bounded divergence satisfying the following "triangle inequality": there exists some $C > 0$, such that for all $M, M', \widehat{M}$ and for all $\pi \in \Pi$, we have*

$$D^\pi(M \| M') \le C \cdot \left(D^\pi(\bar{M} \| M) + D^\pi(\bar{M} \| M')\right).$$

Then, for all $\gamma > 0$, we have

$$\underbrace{\underline{\mathrm{dec}}_\gamma^D(\mathcal{M})}_{\text{Vanilla DEC sup over deterministic reference}} \le \underbrace{\underline{\mathrm{dec}}_{\gamma/2C}^D(\mathcal{M})}_{\text{Randomized DEC sup over all distributions}}.$$

[7] It has been shown by Chen et al. (2022) that optimistic DEC can handle model-based approach.

On the other hand, whenever D is convex in the first input, we have

$$\underline{\mathrm{dec}}_\gamma^D(\mathcal{M}) \le \sup_{\widehat{M}\in\mathrm{co}(\mathcal{M})} \mathrm{dec}_\gamma^D(\mathcal{M},\widehat{M}) = \mathrm{dec}_\gamma^D(\mathcal{M}(.$$

We now consider the concrete examples.

- Model-based approach with Hellinger distance.
  - D_H^2 satisfies Assumption 2 with $L=1$ when the value functions are bounded by 1;
  - D_H^2 satisfies the condition of Lemma 4 with $C=2$;
  - Therefore, for Hellinger distance, we have

$$\mathrm{odec}_{2\gamma}^D(\mathcal{M}) - \frac{1}{\gamma} \le \sup_{\widehat{M}} \mathrm{dec}_\gamma(\mathcal{M},\widehat{M}) \le \mathrm{odec}_{\gamma/6}(\mathcal{M}) + \frac{3}{\gamma}.$$

> Optimism (or randomized estimator) offers no statistical advantage for model-based approach with Hellinger divergence.

- Model-free approach with bilinear class.
  - The bilinear divergence (detailed in the Appendix) is asymmetric:

$$D_{bi}^\pi(Q||M) = \sum_{h=1}^H \left(\mathbb{E}_{M,\pi}[\ell_h(Q;z_h)]\right)^2;$$

  - From the viewpoint of estimation, it is not a good idea to swap the input because we can only collect trajectory with the underlying M^*. Specifically, the following flipped estimation loss is hard to control:

$$\check{\mathbf{Est}} := \sum_{t=1}^T \mathbb{E}_{\pi_t\sim p^t}\mathbb{E}_{\widehat{M}_t\sim\mu^t} D_{bi}^\pi(Q^*||\widehat{M}_t) = \sum_{t=1}^T \mathbb{E}_{\pi_t\sim p^t}\mathbb{E}_{\widehat{M}_t\sim\mu^t}\left[\sum_{h=1}^H \left(\mathbb{E}_{\widehat{M}_t,\pi_t}\ell_h(Q^*;z_h)\right)^2\right];$$

  - On the other hand, as we mentioned before, although the estimation problem of the bilinear class can be handled by trajectory average, the vanilla DEC without feel-good term does not admit a favorable bound (Theorem 5).

On the minimax strategy for policy selection For DEC-based algorithms including Algorithm 3 and Algorithm 5, once we obtain a distribution over $\mathcal{M}$: μ^t (deterministic estimator is a special case), they solve a minimax problem for policy selection:

$$p^t := \operatorname*{argmin}_{p\in\Delta(\Pi)} \sup_{M\in\mathcal{M}} \mathbb{E}_{\pi\sim p}\mathbb{E}_{\widehat{M}_t\sim\mu^t}[V_{1,\widehat{M}_t}(x_1) - V_{1,M}^\pi(x_1) - \gamma\cdot D^\pi(\widehat{M}_t||M)].$$

Then, they pick a policy by $\pi_t\sim p^t$. On the other hand, algorithms in Zhang (2022); Dann et al. (2021); Agarwal and Zhang (2022); Zhong et al. (2022) sample the policy by sampling a hypothesis $f^t\sim\mu^t$. Equivalently, the induced distribution over the policy space is

$$q^t(\pi) := \int_{\{f^t\in\mathcal{H}:\tilde{\pi}(f^t)=\pi\}} d\mu^t(f^t),$$

where $\tilde{\pi}(f)$ is the policy induced by hypothesis f. We can show that such a simple mechanism is sufficient for large classes of tractable problems. However, the following result shows that it could be suboptimal in some corner case.

Theorem 6 (Insufficiency of simple strategy)**.** *Consider the Hellinger distance. For any $S\in\mathbb{N}$ and $H\ge\log_2 S$, there exists a class of horizon-H MDPs $\mathcal{M}$ with $|\mathcal{X}|=S$ and $|\mathcal{A}|=3$ that satisfies the following properties:*

- *There exists an estimation oracle such that $\mathbf{OptEst}_\gamma^H \lesssim \log(S/\delta)$ for all $\gamma>0$;*
- *The simple strategy has $\mathbb{E}\mathrm{Reg}(T) \gtrsim S\wedge 2^{\Omega(H)}$;*
- *The minimax mechanism has $\mathbb{E}\mathrm{Reg}(T)\le\tilde{O}(\sqrt{T\log S})$ where $\mathrm{odec}_\gamma\le\frac{1}{\gamma}$.*

Discussion: Average Error v.s. Squared Error By Zhong et al. (2022) and Foster et al. (2022), we have seen that a flexible choice of the notion of estimation error is beneficial to cover more problems and methods. For most of frameworks we have adopt the squared Bellman residual, while for OLIVE (Jiang et al., 2017), it uses an average one. Since the average loss does not suffer from the double-sampling issue, one may wonder whether the average loss is superior.

REFERENCES

Agarwal, A. and Zhang, T. (2022). Model-based rl with optimistic posterior sampling: Structural conditions and sample complexity. *arXiv preprint arXiv:2206.07659*.

Antos, A., Szepesvári, C., and Munos, R. (2008). Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129.

Chen, F., Mei, S., and Bai, Y. (2022). Unified algorithms for rl with decision-estimation coefficients: No-regret, pac, and reward-free learning. *arXiv preprint arXiv:2209.11745*.

Chen, J. and Jiang, N. (2019). Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pages 1042–1051. PMLR.

Dann, C., Mohri, M., Zhang, T., and Zimmert, J. (2021). A provably efficient model-free posterior sampling method for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 34:12040–12051.

Du, S., Kakade, S., Lee, J., Lovett, S., Mahajan, G., Sun, W., and Wang, R. (2021). Bilinear classes: A structural framework for provable generalization in rl. In *International Conference on Machine Learning*, pages 2826–2836. PMLR.

Foster, D. J., Golowich, N., Qian, J., Rakhlin, A., and Sekhari, A. (2022). A note on model-free reinforcement learning with the decision-estimation coefficient. *arXiv preprint arXiv:2211.14250*.

Foster, D. J., Kakade, S. M., Qian, J., and Rakhlin, A. (2021). The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. (2017). Contextual decision processes with low Bellman rank are PAC-learnable. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1704–1713. PMLR.

Jin, C., Liu, Q., and Miryoosefi, S. (2021). Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in Neural Information Processing Systems*, 34.

Zhang, T. (2022). Feel-good thompson sampling for contextual bandits and reinforcement learning. *SIAM Journal on Mathematics of Data Science*, 4(2):834–857.

Zhong, H., Xiong, W., Zheng, S., Wang, L., Wang, Z., Yang, Z., and Zhang, T. (2022). A posterior sampling framework for interactive decision making. *arXiv preprint arXiv:2211.01962*.

A BI-ODEC-E2D

In this section, we show how to handle the bilinear class by model-free approach and optimistic DEC. For simplicity, we consider the finite-dimensional case, which can be generalized by the notion of information gain (Du et al., 2021). To be consistent with the worst-case consideration ($\sup_{M\in\mathcal{M}}$) of DEC, we slightly modify the original definition in Du et al. (2021).

Definition 5 (Bilinear class). *An MDP M is said to a bilinear class with $\mathcal{H}$, discrepancy function $\ell : \mathcal{H}\times(\mathcal{X}\times\mathcal{A}\times\mathbb{R}\times\mathcal{X})\times\mathcal{H}\to\mathbb{R}$, if there exists functions $W_h : \mathcal{H}\to\mathbb{R}^d$ and $X_h : \mathcal{H}\to\mathbb{R}^d$ such that for all $f\in\mathcal{H}$ and $h\in[H]$, we have*

$$\left|\mathbb{E}_{\pi_f,M}\left[Q_{h,f}(x_h,a_H)-r_h-V_{h+1,f}(x_{h+1})\right]\right|\le|\,\langle X_h(f;M),W_h(f;M)\rangle\,|.$$

Moreover, we denote $z_h=(x_h,a_h,r_h,x_{h+1})$. Then, it holds that for all $f,g\in\mathcal{H}$:

$$\left|\mathbb{E}_{M,x_h\sim\pi_f,a_h\sim\tilde{\pi}}\ell_h(g;z_h)\right|=|\,\langle X_h(f;M),W_h(g,M)\rangle\,|,$$

where $\tilde{\pi}$ is either π_f (Q-type) or π_g (V-type). Furthermore, we assume that $|\mathbb{E}_{M,\pi}\ell_h(f^*;z_h)|=0$ for all π. We further assume that $|\ell_h|$ is upper bounded by $L\ge 1$.

Note that we assume that the discrepancy loss does not depend on the roll-in policy for simplicity and this can be generalized readily. We also list the original definition here.

Definition 6 (Bilinear Class). *An MDP M is said to a bilinear class with $\mathcal{H}$, discrepancy function $\ell : \mathcal{H}\times(\mathcal{X}\times\mathcal{A}\times\mathbb{R}\times\mathcal{X})\times\mathcal{H}\to\mathbb{R}$, if there exists functions $W_h : \mathcal{H}\to\mathbb{R}^d$ and $X_h : \mathcal{H}\to\mathbb{R}^d$ such that for all $f\in\mathcal{H}$ and $h\in[H]$, we have*

$$\begin{aligned}\left|\mathbb{E}_{\pi_f}\left[Q_{h,f}\left(x_h,a_h\right)-r\left(x_h,a_h\right)-V_{h+1,f}\left(x_{h+1}\right)\right]\right|&\le\left|\langle W_h(f)-W_h\left(f^*\right),X_h(f)\rangle\right|,\\ \left|\mathbb{E}_{x_h\sim\pi_f,a_h\sim\tilde{\pi}}\left[\ell_f\left(g,\zeta_h\right)\right]\right|&=\left|\langle W_h(g)-W_h\left(f^*\right),X_h(f)\rangle\right|,\end{aligned}\tag{11}$$

where $\tilde{\pi}$ is either π_f (Q-type) or π_g (V-type).

Optimistic DEC framework We still consider the model class $\mathcal{M}$ but in general we may take the induced value function $Q_{h,M}$ as input (referred to as the sufficient statistic in Foster et al. (2022)). With slight abuse of notations, we adopt the following $D^{\pi}_{bi}(f||M)$:

$$D^{\pi}_{bi}(f||M):=\sum_{h=1}^{H}\left(\mathbb{E}_{M,\pi}\ell_h(f;z_h)\right)^2.$$

In this case, the optimistic DEC is given by

$$\mathrm{odec}^{D}_{\gamma}(\mathcal{M},\mu^t):=\inf_{p\in\Delta(\Pi)}\sup_{M\in\mathcal{M}}\mathbb{E}_{\pi\sim p}\mathbb{E}_{\widehat{M}_t\sim\mu^t}[V_{1,\widehat{M}_t}(x_1)-V^{\pi}_{1,M}(x_1)-\gamma\cdot\sum_{h=1}^{H}\left(\mathbb{E}_{M,\pi}\ell_h(\widehat{M}_t;z_h)\right)^2].$$

and the estimation target is

$$\mathbf{OptEst}^{D}_{\gamma}:=\sum_{t=1}^{T}\mathbb{E}_{\pi_t\sim p^t}\mathbb{E}_{\widehat{M}_t\sim\mu^t}\left[\sum_{h=1}^{H}\left(\mathbb{E}_{M^*,\pi_t}\ell_h(\widehat{M}_t;z_h)\right)^2+\gamma^{-1}\underbrace{\left(V^*_1(x_1)-V_{1,\widehat{M}_t}(x_1)\right)}_{\text{Feel-good term}}\right].$$

Error estimation For each epoch k, we adopt the trajectory average technique to compute:

$$L^k(f):=\sum_{h=1}^{H}\left(\frac{1}{m}\sum_{i=1}^{m}\ell_h(f;z_h^{k,i})\right)^2-\frac{1}{8\gamma}V_{1,f}(x_1).$$

The randomized estimator is given by the following exponential weight update:

$$\mu^k(f)\propto\exp\left(-\eta\sum_{s=1}^{k-1}L^s(f)\right),$$

which is similar to MOPS (Agarwal and Zhang, 2022). The following lemma controls the cumulative estimation error.

Lemma 5 (Estimation error for bilinear class)**.** *If we take batch size as m, $\gamma \geq 1$, and $K := T/m$, with an appropriate choice of learning rate η, the exponential weight update gives*

$$\mathbf{OptEst}_{\gamma} \leq \frac{\sqrt{K \log |\mathcal{Q}|}}{\gamma} + HL^2 \log(|\mathcal{Q}|KH/\delta)\big(1 + \frac{K}{m}\big).$$

Note that the estimation error only depends the cardinality of $\mathcal{Q} := \{Q_{h,M} : M \in \mathcal{M}\}$, which can be much smaller than $|\mathcal{M}|$. Also note that we cannot take $m \geq K$ because it will not bring any help.

Proof. We start with the following standard online aggregation analysis.

Lemma 6 (online learning)**.** *For $t = k, \cdots, K$:*

- *Learner predicts a random hypothesis $g^k \in \mathcal{H}$;*
- *Nature reveals $L^k : \mathcal{H} \to \mathbb{R}$ and learner suffers loss $L^k(g^k)$.*

If we define the regret as:

$$\mathrm{Reg}_{OL} := \sum_{k=1}^{K} \mathbb{E}_{g^k \sim \mu^k} L^k(g^k) - \inf_{g \in \mathcal{H}} \sum_{k=1}^{K} L^k(g),$$

where $\mu^k(g) \propto \exp\big(-\eta \sum_{s=1}^{k-1} L^s(g)\big)$, then for any sequence of loss function $L^s, \cdots, L^T$ with $L^s(g) \in [-L, L]$, if we set $\eta \leq 1/(2L)$, then

$$\mathrm{Reg}_{OL} \leq 4\eta \sum_{s=1}^{K} \mathbb{E}_{g^t \sim \mu^t}\big[(L^k(g^k))^2\big] + \frac{\log |\mathcal{H}|}{\eta}.$$

For our algorithm, we take

$$L^k(f) := \sum_{h=1}^{H} \big(\frac{1}{m} \sum_{i=1}^{m} \ell_h(f; z_h^{k,i})\big)^2 - \alpha V_{1,f}(x_1).$$

Therefore, it holds that

$$\sum_{k=1}^{K} \mathbb{E}_{f^k \sim \mu^k} L^k(f^k) - \sum_{k=1}^{K} L^k(f^*) \leq 4\eta \sum_{k=1}^{K} \mathbb{E}_{f^k \sim \mu^k} \big(L^k(f^k)\big)^2 + \frac{\log |\mathcal{Q}|}{\eta}.$$

By $(a+b)^2 \leq 2a^2 + 2b^2$ and boundedness and also the Jensen's inequality, we have

$$\big(L^k(f)\big)^2 \leq 2HL^2 \sum_{h=1}^{H} \big(\frac{1}{m} \sum_{i=1}^{m} \ell_h(f; z_h^{k,i})\big)^2 + 2\alpha^2.$$

Therefore, we have

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}_{f^k \sim \mu^k} \big(\frac{1}{m} \sum_{i=1}^{m} \ell_h(f^k; z_h^{k,i})\big)^2 + \alpha \sum_{k=1}^{K} \mathbb{E}_{f^k \sim \mu^k}[V_1^*(x_1) - V_{1,f}(x_1)] \\
&\leq \sum_{k=1}^{K}\sum_{h=1}^{H} \big(\frac{1}{m} \sum_{i=1}^{m} \ell_h(f^*; z_h^{k,i})\big)^2 + 8\eta HL^2 \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}_{f^k \sim \mu^k} \big(\frac{1}{m} \sum_{i=1}^{m} \ell_h(f; z_h^{k,i})\big)^2 + 8\eta\alpha^2 K + \frac{\log |\mathcal{Q}|}{\eta}.
\end{aligned}$$

We take $\eta \leq \frac{1}{16HL^2}$ such that $8\eta HL^2 \leq 1/2$. Then, it follows

$$\begin{aligned}
&\frac{1}{2} \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}_{f^k \sim \mu^k} \big(\frac{1}{m} \sum_{i=1}^{m} \ell_h(f^k; z_h^{k,i})\big)^2 + \alpha \sum_{k=1}^{K} \mathbb{E}_{f^k \sim \mu^k}[V_1^*(x_1) - V_{1,f}(x_1)] \\
&\leq \sum_{k=1}^{K}\sum_{h=1}^{H} \big(\frac{1}{m} \sum_{i=1}^{m} \ell_h(f^*; z_h^{k,i})\big)^2 + 8\eta\alpha^2 K + \frac{\log |\mathcal{Q}|}{\eta}.
\end{aligned} \tag{12}$$

We now state the following uniform concentration inequality.

Lemma 7 (Uniform concentration)**.** *With probability at least* $1-\delta$*, for any* (k,h,f)*, we have*

$$0.5\big(\mathbb{E}_{M^*,\pi_k}\ell_h(f;z_h)\big)^2-\frac{2L^2\iota}{m}\le\big(\frac{1}{m}\sum_{i=1}^m\ell_h(f;z_h^{k,i})\big)^2\le 2\big(\mathbb{E}_{M^*,\pi_k}\ell_h(f;z_h)\big)^2+\frac{4L^2\iota}{m}.$$

Combine (12) with the lemma, we have

$$\begin{aligned}&\frac{1}{4}\sum_{k=1}^K\sum_{h=1}^H\mathbb{E}_{f^k\sim\mu^k}\big(\mathbb{E}_{M^*,\pi_k}\ell_h(f^k;z_h)\big)^2+\alpha\sum_{k=1}^K\mathbb{E}_{f^k\sim\mu^k}[V_1^*(x_1)-V_{1,f}(x_1)]\\&\le 8\eta\alpha^2K+\frac{\log|\mathcal{Q}|}{\eta}+\frac{6HKL^2\iota}{m},\end{aligned}$$

where we use $\mathbb{E}_{M^*,\pi}\ell_h(f^*,z)=0$. Now we apply the Freedman's inequality to obtain that

$$\sum_{k=1}^K\sum_{h=1}^H\mathbb{E}_{f^k\sim\mu^k}\big(\mathbb{E}_{M^*,\pi_k}\ell_h(f^k;z_h)\big)^2\ge\frac{1}{2}\sum_{k=1}^K\mathbb{E}_{\pi^k\sim p^k}\sum_{h=1}^H\mathbb{E}_{f^k\sim\mu^k}\big(\mathbb{E}_{M^*,\pi_k}\ell_h(f^k;z_h)\big)^2-O(HL^2\log(H/\delta)).$$

Putting everything together, we have

$$\frac{1}{8}\sum_{k=1}^K\mathbb{E}_{\pi_k\sim p^k}\mathbb{E}_{f^k\sim\mu^k}[D^{\pi_k}(f^k||M^*)+8\alpha(V_1^*(x_1)-V_{1,f}(x_1))]\lesssim\eta\alpha^2K+\frac{\log|\mathcal{Q}|}{\eta}+\frac{HKL^2\iota}{m}+HL^2\log(H/\delta).$$

We now choose $\alpha=\frac{1}{8\gamma}$ and $\eta=\sqrt{\frac{\log|\mathcal{Q}|}{\alpha^2K}}\wedge\frac{1}{16R}$ to obtain the desired result. □

The following lemma establishes the upper bound for the odec.

Lemma 8 (O-DEC for the bilinear class)**.** *For Q-type problem, we have that for all* $\gamma>0$*,*

$$\mathrm{odec}_\gamma^D(\mathcal{M})\lesssim\frac{H\cdot d}{\gamma}.$$

For V-type problem, we have that for all $\gamma\ge H^2\cdot d$*,*

$$\mathrm{odec}_\gamma^D(\mathcal{M})\lesssim\sqrt{\frac{H^2\cdot d}{\gamma}}.$$

We can obtain the regret as follows. We illustrate by the Q-type problems.

$$\mathrm{Reg}(T)\lesssim\mathrm{odec}_\gamma^D(\mathcal{M})\cdot T+\gamma m\cdot\mathbf{OptEst}_\gamma^D(K,m,\delta).$$

In our case, we have (we also take $L=1$ for simplicity):

$$\begin{aligned}\mathrm{Reg}(T)&\le\mathrm{odec}_\gamma^D(\mathcal{M})\cdot T+\sqrt{mT\log|\mathcal{Q}|}+\gamma HL^2\iota K+H\iota\gamma m\\&\le\frac{HdT}{\gamma}+\sqrt{mT\log|\mathcal{Q}|}+\gamma H\iota K+\gamma H\iota m\\&\lesssim T^{3/4}\end{aligned}$$

where $m=K=T^{1/2}$ and $\gamma=T^{1/4}$.

B BI-GEC-TS

In this section, we handle the bilinear class with GEC. We first study the UCB algorithm

Bi-LinUCB By the definition of GEC and optimism, we have

$$\begin{aligned}\mathrm{Reg}(T) &\le m\big(\sum_{k=1}^{K} V_{1,f^k}(x_1) - V_1^{\pi_{f^k}}(x_1)\big)\\ &\lesssim m\Big[d\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{s=1}^{k-1}\big(\mathbb{E}_{\pi_{f^s}}\ell_h(f^k,\zeta_h^s)\big)^2\Big]^{1/2} + md.\end{aligned}$$

It remains to determine the confidence level. We use the notation

$$\hat{\epsilon}_h^k(f) := \frac{1}{m}\sum_{i=1}^{m} l_h(f,\zeta_h^{k,i}).$$

By Hoeffding's inequality and a union bound, we have for all $k\in[K], h\in[H], f\in\mathcal{H}$, it holds that with probability at least $1-\delta$,

$$|\hat{\epsilon}_h^k(f) - \mathbb{E}_{\pi_{f^k}}\ell_h(f,\zeta_h^s)| \lesssim \sqrt{\frac{\iota}{m}},$$

where $\iota = c\cdot\log(|\mathcal{H}|KH/\delta)$. By $(a+b)^2\le 2a^2+2b^2$, we have

$$\big(\mathbb{E}_{\pi_{f^s}}\ell_h(f,\zeta_h)\big)^2 \le 2\big(\hat{\epsilon}_h^k(f)\big)^2 + \frac{2\iota}{m}$$

and

$$\big(\hat{\epsilon}_h^k(f)\big)^2 \le 2\big(\mathbb{E}_{\pi_{f^s}}\ell_h(f,\zeta_h)\big)^2 + \frac{2\iota}{m}.$$

Therefore, the confidence level is that

$$\sum_{s=1}^{k-1}\big(\mathbb{E}_{\pi_{f^s}}\ell_h(f^*,\zeta_h^s)\big)^2 \le \frac{2(k-1)\iota}{m}.$$

It holds that

$$\begin{aligned}\mathrm{Reg}(T) &\le m\big(\sum_{k=1}^{K} V_{1,f^k}(x_1) - V_1^{\pi_{f^k}}(x_1)\big)\\ &\lesssim m\Big[d\sum_{h=1}^{H}\sum_{k=1}^{K}\frac{(k-1)\iota}{m}\Big]^{1/2} + md\\ &\lesssim K\sqrt{mdH\iota} + md\\ &\lesssim T^{2/3}d^{2/3}(\iota H)^{1/3}\end{aligned}$$

where we set $K = T^{1/3}d^{1/3}\iota^{-1/3}H^{-1/3}$ and $mK = T$. This matches the result of Du et al. (2021) implied by online-to-batch (ignoring the H due to different boundedness assumption).

We also note that

$$\mathbb{E}_s(\mathbb{E}_h\hat{\epsilon}_h^s(f))^2 = \big(\mathbb{E}_s\hat{\epsilon}_h^2(f)\big)^2 + \frac{1}{m^2}\sum_{i=1}^{m}\sigma_i^2.$$

Therefore, we have

$$\hat{\epsilon}_h^k(f^*) \le 2\big(\mathbb{E}_s l_h(f^*,\zeta_h)\big)^2 + \frac{2\iota}{m} = \frac{2\iota}{m}.$$

and

$$\begin{aligned}\mathbb{E}_s(\mathbb{E}_h l_h(f,\zeta_h))^2 &\le 2\mathbb{E}_s\big(\mathbb{E}_h l_h(f,\zeta_h) - \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_h l_h(f,\zeta_h^{s,i})\big)^2 + 2\mathbb{E}_s\big(\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{i,h} l_h(f,\zeta_h^{s,i})\big)^2\\ &\lesssim 2\frac{\iota}{m} + 2\big(\mathbb{E}_s l_h(f,\zeta_h)\big)^2 + 2\frac{1}{m} \lesssim \frac{\iota}{m}.\end{aligned}$$

TS The first step is to equivalently write the posterior as:

$$p^k(f) \propto \exp\left(-\eta \sum_{s=1}^{k-1} L_h^s(f) + \ln p^0(f) + \gamma \cdot \underbrace{\big(V_{f,1}(x_1) - V_1^*(x_1)\big)}_{\Delta V_{1,f}(x_1)}\right), \tag{13}$$

where $L_h^s(f) := \big(\hat{\epsilon}_h^s(f)\big)^2$. We start with the definition of GEC.

$$\begin{aligned}
\sum_{k=1}^{K} V_1^*(x_1) - V_1^{\pi_{f^k}}(x_1) &\lesssim -\sum_{k=1}^{K} \Delta V_{1,f^k}(x_1) + \Big[d \sum_{h=1}^{H}\sum_{k=1}^{K}\Big(\sum_{s=1}^{k-1} \mathbb{E}_{\pi_{\exp}(f^s,h)} \ell_{f^s}(f^k, \xi_h)\Big)\Big]^{1/2} + \min\{d, HK\} \\
&\lesssim -\sum_{k=1}^{K} \Delta V_{1,f^k}(x_1) + \mu \sum_{h=1}^{H}\sum_{k=1}^{K}\Big(\sum_{s=1}^{k-1} \mathbb{E}_{\pi_{\exp}(f^s,h)} \ell_{f^s}(f^k, \xi_h)\Big) + (\frac{1}{\mu} + 1) \cdot d \\
&= -\sum_{k=1}^{K} \Delta V_{1,f^k}(x_1) + \mu \sum_{h=1}^{H}\sum_{k=1}^{K}\Big(\sum_{s=1}^{k-1} \mathbb{E}_{\pi_{\exp}(f^s,h)} (\mathbb{E}_h l_h(f^t, \xi_h))^2\Big) + (\frac{1}{\mu} + 1) \cdot d \\
&= -\sum_{k=1}^{K} \Delta V_{1,f^k}(x_1) + \frac{0.5\eta}{\gamma} \sum_{h=1}^{H}\sum_{k=1}^{K}\Big(\sum_{s=1}^{k-1} \mathbb{E}_{\pi_{\exp}(f^s,h)} (\mathbb{E}_h l_h(f^k, \xi_h))^2\Big) + (\frac{\gamma}{0.5\eta} + 1) \cdot d
\end{aligned}$$

where we use AM-GM inequality in the second inequality and take $\mu = \frac{0.5\eta}{\gamma}$ in the last equality. We now connect the training error to the potential function.

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}^{k-1}} \mathbb{E}_{f^k \sim p^k}\Big[\eta \sum_{s=1}^{k-1}\sum_{h=1}^{H} L_h^s(f^k) + \ln \frac{p^k(f^k)}{p^0(f^k)} - \gamma \cdot \Delta V_{1,f^k}(x_1)\Big] \\
&\geq \mathbb{E}_{\mathcal{S}^{k-1}} \mathbb{E}_{f^k \sim p^k}\Big[0.5\eta \sum_{s=1}^{k-1}\sum_{h=1}^{H} \mathbb{E}_s\big(\mathbb{E}_h l_h(f^k, \zeta_h^s)\big)^2 - \frac{\eta H(k-1)\iota}{m} - \gamma \cdot \Delta V_{1,f^k}(x_1)\Big]
\end{aligned}$$

We require the following lemma.

Lemma 9. *Let ν be a probability distribution over $x \in \mathcal{X}$. Then, $\mathbb{E}_{x\sim\nu}[f(x) + \log \nu(x)]$ is minimized at $\nu(x) \propto \exp(-f(x))$.*

This implies that

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}^{k-1}} \mathbb{E}_{f^k \sim p^k}\Big[\eta \sum_{s=1}^{k-1}\sum_{h=1}^{H} L_h^s(f^k) + \ln \frac{p^k(f^k)}{p^0(f^k)} - \gamma \cdot \Delta V_{1,f^k}(x_1)\Big] \\
&= \mathbb{E}_{\mathcal{S}^{k-1}} \inf_p \mathbb{E}_{f\sim p(\cdot)}\Big[\eta \sum_{s=1}^{k-1}\sum_{h=1}^{H} L_h^s(f) + \ln \frac{p(f)}{p^0(f)} - \gamma \cdot \Delta V_{1,f}(x_1)\Big] \\
&\leq \mathbb{E}_{\mathcal{S}^{k-1}} \inf_p \mathbb{E}_{f\sim p(\cdot)}\Big[2\eta \sum_{s=1}^{k-1}\sum_{h=1}^{H} \mathbb{E}_s\big(\mathbb{E}_h l_h(f, \zeta_h^s)\big)^2 + \frac{\eta H(k-1)\iota}{m} + \ln \frac{p(f)}{p^0(f)} - \gamma \cdot \Delta V_{1,f}(x_1)\Big] \\
&\lesssim \gamma\epsilon + 2\eta H(k-1)\epsilon^2 + \frac{\eta H(k-1)\iota}{m} + \ln|\mathcal{H}| \\
&= \frac{\eta H(k-1)\iota}{m} + \ln|\mathcal{H}|,
\end{aligned}$$

where we take $p(\cdot) := \delta(f^*)$. It follows that

$$\begin{aligned}
\mathrm{Reg}(T) &= m\sum_{k=1}^{K}\left[V_1^*(x_1) - V_1^{\pi_{f^k}}(x_1)\right] \\
&\lesssim m\cdot\Big[\frac{1}{\gamma}\sum_{k=1}^{K}\mathbb{E}_{\mathcal{S}^{k-1}}\mathbb{E}_{f^k\sim p^k}\Big[\eta\sum_{s=1}^{k-1}\sum_{h=1}^{H}L_h^s(f^k) + \ln\frac{p^k(f^k)}{p^0(f^k)} - \gamma\cdot\Delta V_{1,f^k}(x_1)\Big] + \frac{\eta HK^2\iota}{m\gamma}\Big] + (\frac{\gamma}{0.5\eta}+1)\cdot md \\
&\lesssim \frac{m}{\gamma}\left(\frac{\eta HK^2\iota}{m} + K\cdot\ln|\mathcal{H}|\right) + (\frac{\gamma}{0.5\eta}+1)\cdot md \\
&= \frac{\eta HK^2\iota}{\gamma} + \frac{T\cdot\ln|\mathcal{H}|}{\gamma} + (\frac{\gamma}{0.5\eta}+1)\cdot md \\
&\lesssim T^{2/3}
\end{aligned}$$

Implicit UCB An interesting observation is that if we set $\gamma \approx \eta$ sufficiently large, the realizability term $\ln|\mathcal{H}|$ can be ignored (as long as it is finite). To interpret such an observation, we consider the following modified algorithm, which does not explicitly construct a confidence set.

We consider the following algorithm:

$$f^k := \operatorname*{argmax}_{f\in\mathcal{H}}\Big[V_{1,f}(x_1) - \eta\sum_{s=1}^{k-1}\sum_{h=1}^{H}L_h^s(f)\Big].$$

It follows that

$$\begin{aligned}
&\sum_{k=1}^{K}V_1^*(x_1) - V_1^{\pi_k}(x_1) := \sum_{k=1}^{K}V_{1,f^k}(x_1) - V_1^{\pi_k}(x_1) - \Delta V_{1,f^k}(x_1) \\
&\lesssim -\sum_{k=1}^{K}\Delta V_{1,f^k}(x_1) + \mu\sum_{h=1}^{H}\sum_{k=1}^{K}\Big(\sum_{s=1}^{k-1}\mathbb{E}_{\pi_{\exp}(f^s,h)}(\mathbb{E}_h l_h(f^k,\xi_h))^2\Big) + (\frac{1}{\mu}+1)\cdot d \\
&\le \eta\sum_{k=1}^{K}\sum_{h=1}^{H}L_h^s(f^*) - \eta\sum_{k=1}^{K}\sum_{h=1}^{H}L_h^s(f^k) + \mu\sum_{h=1}^{H}\sum_{k=1}^{K}\Big(\sum_{s=1}^{k-1}\mathbb{E}_{\pi_{\exp}(f^s,h)}(\mathbb{E}_h l_h(f^k,\xi_h))^2\Big) + (\frac{1}{\mu}+1)\cdot d \\
&\lesssim \frac{\eta K^2H\iota}{m} + (\mu-\eta)\sum_{h=1}^{H}\sum_{k=1}^{K}\Big(\sum_{s=1}^{k-1}\mathbb{E}_{\pi_{\exp}(f^s,h)}(\mathbb{E}_h l_h(f^k,\xi_h))^2\Big) + (\frac{1}{\mu}+1)\cdot d \\
&\le \frac{\eta K^2H\iota}{m} + (\frac{1}{\mu}+1)\cdot d \\
&\lesssim \frac{K^2H\iota}{m} + d.
\end{aligned}$$

where the second inequality uses realizability and the definition of algorithm and we carefully tune η and μ to cancel the loss of f^k. It follows that

$$\mathrm{Reg}(T) \lesssim K^2H\iota + md \lesssim T^{2/3}.$$