A. Proof of Theorem 1

In this section, we provide the regret analysis of the UCRL-VTR Algorithm (Algorithm 1). We will explain the motivation for our construction of confidence sets for general nonlinear squared estimation, and establish the regret bound for a general class of transition models, $\mathcal{P}$.

A.1. Preliminaries

Recall that a finite horizon MDP is $M = (\mathcal{S}, \mathcal{A}, P, r, H, s_\circ)$ where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P = (P_a)_{a\in\mathcal{A}}$ is a collection of $P_a : \mathcal{S} \to M_1(\mathcal{S})$ Markov kernels, $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function, $H > 0$ is the horizon and $s_\circ \in \mathcal{S}$ is the initial state. For a state $s \in \mathcal{S}$ and an action $a \in \mathcal{A}$, $P_a(s)$ gives the distribution of the next state that is obtained when action a is executed in state s. For a bounded (measurable) function $V : \mathcal{S} \to \mathbb{R}$, we will use $\langle P_a(s), V\rangle$ as the shorthand for the expected value of V at a random next state s' whose distribution is $P_a(s)$.

Given any policy π (which may or may not use the history), its value function is

$$V^\pi(s) = \mathbb{E}_{\pi,\delta_s}\left[\sum_{i=1}^{H} r(s_i, a_i)\right],$$

where E_{π,δ_s} is the expectation operator underlying the probability measure P_{π,δ_s} induced over sequences of state-action pairs of length H by executing policy π starting at state s in the MDP M and s_h is the state visited in stage h and action a_h is the action taken in that stage after visiting s_h. For a nonstationary Markov policy $\pi = (\pi_1, \dots, \pi_H)$, we also let

$$V_h^\pi(s) = \mathbb{E}_{\pi_{h:H},\delta_s}\left[\sum_{i=1}^{H-h+1} r(s_i, a_i)\right]$$

be the value function of π from stage h to H. Here, $\pi_{h:H}$ denotes the policy $(\pi_h, \dots, \pi_H)$. The optimal value function $V^* = (V_1^*, \dots, V_H^*)$ is defined via $V_h^*(s) = \max_\pi V_h^\pi(s)$, $s \in \mathcal{S}$.

For simplicity assume that r is known. To indicate the dependence of V^* on the transition model P, we will write $V_P^* = (V_{P,1}^*, \dots, V_{P,H}^*)$. For convenience, we define $V_{P,H+1}^* = 0$.

Algorithm 1 is an instance of the following general model-based optimistic algorithm:

Algorithm 2 Generic Algorithm 1-Schema for finite horizon problems

Input: $\mathcal{P}$ – a set of transition models, K – number of episodes, s_0 – initial state
Set $\mathcal{B}_1 = \mathcal{P}$
for $k = 1, \dots, K$ **do**
  $P^k = \text{argmax}\{V_{\tilde{P}}^*(s_0) : \tilde{P} \in \mathcal{B}_k\}$
  $V_k = V_{P^k}^*$
  $s_1^k = s_0$
  for $h = 1, \dots, H$ **do**
    Choose $a_h^k = \text{argmax}_{a\in\mathcal{A}} r(s_h^k, a) + \langle P_a^k(s_h^k), V_{h+1,k}\rangle$
    Observe transition to s_{h+1}^k
  end for
  Construct $\mathcal{B}_{k+1}$ based on $(s_1^k, a_1^k, \dots, s_H^k, a_H^k)$
end for

Specific instances of Algorithm 2 differ in terms of how $\mathcal{B}_{k+1}$ is constructed. In particular, UCRL-VTR uses the construction described in Section 3.2.

Recall that $V_k = (V_{1,k}, \dots, V_{H,k}, V_{H+1,k})$ (with $V_{H+1,k} = 0$) in Algorithm 2. Let π_k be the nonstationary Markov policy chosen in episode k by Algorithm 2. Let

$$R_K = \sum_{k=1}^{K} V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k)$$

be the pseudo-regret of Algorithm 1 for K episodes. The following standard lemma bounds the kth term of the expression on the right-hand side.

Lemma 5. *Assuming that $P \in \mathcal{B}_k$, we have*

$$V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) \le \sup_{\tilde{P}\in\mathcal{B}_k} \sum_{h=1}^{H-1} \langle \tilde{P}_{a_h^k}(s_h^k) - P_{a_h^k}(s_h^k), V_{h,k}\rangle + \sum_{h=1}^{H-1} \xi_{h+1,k}\,,$$

where

$$\xi_{h+1,k} = \langle P_{a_h^k}(s_h^k), V_{h+1,k} - V_{h+1}^{\pi_k}\rangle - \left(V_{h+1,k}(s_{h+1}^k) - V_{h+1}^{\pi_k}(s_{h+1}^k)\right)\,.$$

Note that $(\xi_{2,1}, \xi_{3,1}, \dots, \xi_{H,1}, \xi_{2,2}, \xi_{3,2}, \dots, \xi_{H,2}, \xi_{2,3}, \dots)$ is a sequence of martingale differences.

Proof. Because $P \in \mathcal{B}_k$, $V_1^*(s_1^k) \le V_{1,k}(s_1^k)$ by the definition of the algorithm. Hence,

$$V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) \le V_{1,k}(s_1^k) - V_1^{\pi_k}(s_1^k)\,.$$

Fix $h \in [H]$. In what follows we bound $V_{h,k}(s_h^k) - V_h^{\pi_k}(s_h^k)$. By the definition of π_k, P^k and a_h^k, we have

$$\begin{aligned} V_{h,k}(s_h^k) &= r(s_h^k, a_h^k) + \langle P_{a_h^k}^k(s_h^k), V_{h+1,k}\rangle \text{ and} \\ V_h^{\pi_k}(s_h^k) &= r(s_h^k, a_h^k) + \langle P_{a_h^k}(s_h^k), V_{h+1}^{\pi_k}\rangle\,. \end{aligned}$$

Hence,

$$\begin{aligned} V_{h,k}(s_h^k) - V_h^{\pi_k}(s_h^k) &= \langle P_{a_h^k}^k(s_h^k), V_{h+1,k}\rangle - \langle P_{a_h^k}(s_h^k), V_{h+1}^{\pi_k}\rangle \\ &= \langle P_{a_h^k}^k(s_h^k) - P_{a_h^k}(s_h^k), V_{h+1,k}\rangle + \langle P_{a_h^k}(s_h^k), V_{h+1,k} - V_{h+1}^{\pi_k}\rangle\,. \end{aligned}$$

Therefore, by induction, noting that $V_{H+1,k} = 0$, we get that

$$\begin{aligned} V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) &\le \sum_{h=1}^{H-1} \langle P_{a_h^k}^k(s_h^k) - P_{a_h^k}(s_h^k), V_{h+1,k}\rangle + \sum_{h=1}^{H-1} \xi_{h+1,k} \\ &\le \sup_{\tilde{P}\in\mathcal{B}_k} \sum_{h=1}^{H-1} \langle \tilde{P}_{a_h^k}(s_h^k) - P_{a_h^k}(s_h^k), V_{h+1,k}\rangle + \sum_{h=1}^{H-1} \xi_{h+1,k}\,. \end{aligned}$$

□

A.2. The confidence sets for Algorithm 1

The previous lemma suggests that at the end of the kth episode, the model could be estimated using

$$\hat{P}_k = \operatorname{argmin}_{\tilde{P}\in\mathcal{P}} \sum_{k'=1}^{k} \sum_{h=1}^{H-1} \left(\langle \tilde{P}_{a_h^{k'}}(s_h^{k'}), V_{h+1,k'}\rangle - V_{h+1,k'}(s_{h+1}^{k'})\right)^2 \tag{9}$$

For a confidence set construction, we get inspiration from Proposition 5 in the paper of Osband & Van Roy (2014). The set is centered at $\hat{P}_k$:

$$\mathcal{B}_k = \{\tilde{P} \in \mathcal{P} \,:\, L_k(\hat{P}_k, \tilde{P}) \le \beta_k\}\,, \tag{10}$$

where

$$L_k(\hat{P}, \tilde{P}) = \sum_{k'=1}^{k} \sum_{h=1}^{H-1} \left(\langle \tilde{P}_{a_h^{k'}}(s_h^{k'}) - \hat{P}_{a_h^{k'}}(s_h^{k'}), V_{h+1,k'}\rangle\right)^2\,.$$

Note that this is the same confidence set as described in Section 3.2. To obtain the value of β_k, we now consider the nonlinear least-squares confidence set construction from Russo & Van Roy (2014). The next section is devoted to this construction.

A.3. Confidence sets for general nonlinear least-squares

Let $(X_p, Y_p)_{p=1,2,\dots}$ be a sequence of random elements, $X_p \in \mathcal{X}$ for some measurable set $\mathcal{X}$ and $Y_p \in \mathbb{R}$. Let $\mathcal{F}$ be a subset of the set of real-valued measurable functions with domain $\mathcal{X}$. Let $\mathbb{F} = (\mathbb{F}_p)_{p=0,1,\dots}$ be a filtration such that for all $p \geq 1$, $(X_1, Y_1, \dots, X_{p-1}, Y_{p-1}, X_p)$ is $\mathbb{F}_{p-1}$ measurable and such that there exists some function $f_* \in \mathcal{F}$ such that $\mathbb{E}[Y_p \mid \mathbb{F}_{p-1}] = f_*(X_p)$ holds for all $p \geq 1$. The (nonlinear) least-squares predictor given $(X_1, Y_1, \dots, X_t, Y_t)$ is $\hat{f}_t = \operatorname{argmin}_{f \in \mathcal{F}} \sum_{p=1}^{t} (f(X_p) - Y_p)^2$. We say that Z is conditionally ρ-subgaussian given the σ-algebra $\mathbb{F}$ if for all $\lambda \in \mathbb{R}$, $\log \mathbb{E}[\exp(\lambda Z) | \mathbb{F}] \leq \frac{1}{2}\lambda^2\rho^2$. For $\alpha > 0$, let N_α be the $\|\cdot\|_\infty$-covering number of $\mathcal{F}$ at scale α. That is, N_α is the smallest integer for which there exist $\mathcal{G} \subset \mathcal{F}$ with N_α elements such that for any $f \in \mathcal{F}$, $\min_{g \in \mathcal{G}} \|f - g\|_\infty \leq \alpha$. For $\beta > 0$, define

$$\mathcal{F}_t(\beta) = \{ f \in \mathcal{F} : \sum_{p=1}^{t} (f(X_p) - \hat{f}_t(X_p))^2 \leq \beta \} .$$

We have the following theorem, the proof of which is given in Section A.6.

Theorem 6. *Let $\mathbb{F}$ be the filtration defined above and assume that the functions in $\mathcal{F}$ are bounded by the positive constant $C > 0$. Assume that for each $s \geq 1$, $(Y_p - f_*(X_p))_p$ is conditionally σ-subgaussian given $\mathbb{F}_{p-1}$. Then, for any $\alpha > 0$, with probability $1 - \delta$, for all $t \geq 1$, $f_* \in \mathcal{F}_t(\beta_t(\delta, \alpha))$, where*

$$\beta_t(\delta, \alpha) = 8\sigma^2 \log(2N_\alpha/\delta) + 4t\alpha \left(C + \sqrt{\sigma^2 \log(4t(t+1)/\delta)} \right) .$$

The proof follows that of Proposition 6, Russo & Van Roy (2014), with minor improvements, which lead to a slightly better bound. In particular, with our notation, Russo & Van Roy stated their result with

$$\beta_t^{\mathrm{RvR}}(\delta, \alpha) = 8\sigma^2 \log(2N_\alpha/\delta) + 2t\alpha \left(8C + \sqrt{8\sigma^2 \log(8t^2/\delta)} \right) .$$

While $\beta_t(\delta, \alpha) \leq \beta_t^{\mathrm{RvR}}(\delta, \alpha)$, the improvement is only in terms of smaller constants.

A.4. The choice of β_k in Algorithm 1

To use this result in our RL problem recall that $\mathcal{P}$ is the set of transition probabilities parameterized by $\theta \in \Theta$. We index time $t = 1, 2, \dots$ in a continuous fashion. Episode $k = 1, 2, \dots$ and stage $h = 1, \dots, H-1$ corresponds to time $t = (k-1)(H-1) + h$:

episode (k)	1	1	...	1	2	2	...	2	3	...
stage (h)	1	2	...	$H-1$	1	2	...	$H-1$	1	...
time step (t)	1	2	...	$H-1$	H	$H+1$	...	$2H-2$	$2H-1$	...

Note that the transitions at stage $h = H$ are skipped and the time index at the end of episode $k \geq 1$ is $k(H-1)$.

Let $V_{(t)}$ be the value function used by Algorithm 1 at time t ($V_{(t)}$ is constant in periods of length $H-1$), while let $(s_{(t)}, a_{(t)})$ be the state-action pair visited at time t.

Let $\mathcal{V}$ be the set of optimal value functions under some model in $\mathcal{P}$: $\mathcal{V} = \{V^*_{P'} : P' \in \mathcal{P}\}$. Note that $\mathcal{V} \subset \mathcal{B}(\mathcal{S}, H)$, where $\mathcal{B}(\mathcal{S}, H)$ denotes the set of real-valued measurable functions with domain $\mathcal{S}$ that are bounded by H. Note also that for all t, $V_{(t)} \in \mathcal{V}$. Define $\mathcal{X} = \mathcal{S} \times \mathcal{A} \times \mathcal{V}$. We also let $X_t = (s_{(t)}, a_{(t)}, V_{(t)})$, $Y_t = V_{(t)}(s_{(t+1)})$ when $t+1 \notin \{H+1, 2H+1, \dots\}$ and $Y_t = V_{(t)}(s^k_{H+1})$, and choose

$$\mathcal{F} = \left\{ f : \mathcal{X} \to \mathbb{R} : \exists \tilde{P} \in \mathcal{P} \text{ s.t. } f(s, a, v) = \int \tilde{P}_a(ds'|s) v(s') \right\} . \tag{11}$$

Note that $\mathcal{F} \subset \mathcal{B}_\infty(\mathcal{X}, H)$.

Let $\phi : \mathcal{P} \to \mathcal{F}$ be the natural surjection to $\mathcal{F}$: $\phi(P) = f$ where $f(s, a, v) = \int P_a(ds'|s) v(s')$ for $(s, a, v) \in \mathcal{X}$. We know show that ϕ is in fact a bijection. If $P \neq P'$, this means that for some $(s, a) \in \mathcal{S} \times \mathcal{A}$ and $U \subset \mathcal{S}$ measurable, $P_a(U|s) \neq P'_a(U|s)$. Choosing v to be the indicator of U, note that $(s, a, v) \in \mathcal{X}$. Hence, $\phi(P)(s, a, v) = P_a(U|s) \neq$

$P'_a(U|s) = \phi(P')(s,a,v)$, and hence $\phi(P) \neq \phi(P')$: ϕ is indeed a bijection. For convenience and to reduce clutter, we will write $f_P = \phi(P)$.

Choose $\mathbb{F} = (\mathbb{F}_t)_{t\geq 0}$ so that $\mathbb{F}_{t-1}$ is generated by $(s_{(1)}, a_{(1)}, V_{(1)}, \ldots, s_{(t)}, a_{(t)}, V_{(t)})$. Then $\mathbb{E}[Y_t|\mathbb{F}_{t-1}] = \int P_{a_{(t)}}(ds'|s_{(t)})V_{(t)}(s') = f_P(X_t)$ and by definition $f_P \in \mathcal{F}$. Now, $Y_t \in [0,H]$, hence, $Z_t = Y_t - f_P(X_t)$ is conditionally $H/2$-subgaussian given $\mathbb{F}_{t-1}$.

Let $t = k(H-1)$ for some $k \geq 1$. Thus, this time step corresponds to finishing episode k and thus $V_{(t)} = V_k$. Furthermore, letting $\hat{f}_t = \operatorname{argmin}_{f\in\mathcal{F}} \sum_{p=1}^t (f(X_p) - Y_p)^2$, since ϕ is an injection, we see that $\hat{f}_t = f_{\hat{P}_k}$ where $\hat{P}_k$ is defined using (9). For $P', P'' \in \mathcal{P}$, we have $L_k(P', P'') = \sum_{p=1}^t (f_{P'}(X_p) - f_{P''}(X_p))^2$ and thus

$$\mathcal{B}_k = \{\tilde{P} \in \mathcal{P} \,:\, L_k(\hat{P}_k, \tilde{P}) \leq \beta_k\} = \{\tilde{P} \in \mathcal{P} \,:\, \sum_{p=1}^t (\hat{f}_t(X_p) - f_{\tilde{P}}(X_p))^2 \leq \beta_k\}$$

$$= \{\phi^{-1}(f) \,:\, f \in \mathcal{F} \text{ and } \sum_{p=1}^t (\hat{f}_t(X_p) - f(X_p))^2 \leq \beta_k\} = \phi^{-1}(\mathcal{F}_t(\beta_k))\,.$$

Corollary 7. *For $\alpha > 0$ and $k \geq 1$ let*

$$\beta_k = 2H^2 \log\left(\frac{2\mathcal{N}(\mathcal{F}, \alpha, \|\cdot\|_\infty)}{\delta}\right) + 2H(kH-1)\alpha\left\{2 + \sqrt{\log\left(\frac{4kH(kH-1)}{\delta}\right)}\right\}\,.$$

Then, with probability $1-\delta$, for any $k \geq 1$, $P \in \mathcal{B}_k$ where $\mathcal{B}_k$ is defined by (10).

A.5. Regret of Algorithm 1

Recall that $\mathcal{X} = \mathcal{S} \times \mathcal{A} \times \mathcal{V}$ where $\mathcal{V} \subset \mathcal{B}_\infty(\mathcal{S}, H)$ is the set of value functions that are optimal under some model in $\mathcal{P}$. We will abbreviate $(x_1, \ldots, x_t) \in \mathcal{X}^t$ as $x_{1:t}$. Further, we let $\mathcal{F}|_{x_{1:t}} = \{(f(x_1), \ldots, f(x_t)) \,:\, f \in \mathcal{F}\} (\subset \mathbb{R}^t)$ and for $S \subset \mathbb{R}^t$, let $\operatorname{diam}(S) = \sup_{u,v\in S} \|u - v\|_2$ be the diameter of S. We will need the following lemma, extracted from Russo & Van Roy (2014):

Lemma 8 (Lemma 5 of Russo & Van Roy (2014))**.** *Let $\mathcal{F} \subset B_\infty(\mathcal{X}, C)$ be a set of functions bounded by $C > 0$, $(\mathcal{F}_t)_{t\geq 1}$ and $(x_t)_{t\geq 1}$ be sequences such that $\mathcal{F}_t \subset \mathcal{F}$ and $x_t \in \mathcal{X}$ hold for $t \geq 1$. Then, for any $T \geq 1$ and $\alpha > 0$ it holds that*

$$\sum_{t=1}^T \operatorname{diam}(\mathcal{F}_t|_{x_t}) \leq \alpha + C(d \wedge T) + 2\delta_T\sqrt{dT}\,,$$

where $\delta_T = \max_{1\leq t\leq T} \operatorname{diam}(\mathcal{F}_t|_{x_{1:t}})$ and $d = \dim_{\mathcal{E}}(\mathcal{F}, \alpha)$.

Let

$$W_k = \sup_{\tilde{P}\in\mathcal{B}_k} \sum_{h=1}^{H-1} \langle \tilde{P}_{a_h^k}(s_h^k) - P_{a_h^k}(s_h^k), V_{h,k}\rangle\,.$$

From Lemma 5, we get

$$R_K \leq \sum_{k=1}^K W_k + \sum_{k=1}^K \sum_{h=1}^{H-1} \xi_{h+1,k}\,. \tag{12}$$

Lemma 9. *Let $\alpha > 0$ and $d = \dim_{\mathcal{E}}(\mathcal{F}, \alpha)$ where $\mathcal{F}$ is given by (11). Then, for any nondecreasing sequence $(\beta_k^2)_{k=1}^K$, on the event when $P \in \cap_{k\in[K]}\mathcal{B}_k$,*

$$\sum_{k=1}^K W_k \leq \alpha + H(d \wedge K(H-1)) + 4\sqrt{d\beta_K K(H-1)}\,.$$

Proof. Let $P \in \cap_{k\in[K]}\mathcal{B}_k$ holds. Using the notation of the previous section, letting $\tilde{\mathcal{F}}_t = \mathcal{F}_t(\beta_k)$ for $(k-1)(H-1)+1 \le t \le k(H-1)$, we have

$$\begin{aligned}
\sum_{k=1}^{K} W_k &\le \sum_{k=1}^{K} \sup_{\tilde{P}\in\mathcal{B}_k} \sum_{h=1}^{H-1} \left(f_{\tilde{P}}(s_h^k, a_h^k, V_{h+1,k}) - f_P(s_h^k, a_h^k, V_{h+1,k})\right) \\
&\le \sum_{t=1}^{K(H-1)} \operatorname{diam}(\tilde{\mathcal{F}}_t|_{X_t}) \qquad \text{(because } P \in \cap_{k\in[K]}\mathcal{B}_k\text{)} \\
&\le \alpha + H(d \wedge K(H-1)) + 2\delta_{K(H-1)}\sqrt{dK(H-1)}\,,
\end{aligned}$$

where X_t is defined in Section A.4 and where the last inequality is by Lemma 8, which is applicable because $\mathcal{F} \subset \mathcal{B}_\infty(\mathcal{X}, H)$ holds by choice, and $\delta_{K(H-1)} = \max_{1\le t\le K(H-1)} \operatorname{diam}(\tilde{\mathcal{F}}_t|_{X_{1:t}})$. Thanks to the definition of $\tilde{\mathcal{F}}_t$, $\delta_{K(H-1)} \le 2\sqrt{\beta_K}$. Plugging this into the previous display finishes the proof. □

A.5.1. Proof of Theorem 1

Proof. Note that for any $k \in [K]$ and $h \in [H-1]$, $\xi_{h+1}, k \in [-H, H]$. As noted beforehand, $\xi_{2,1}, \xi_{3,1}, \ldots, \xi_{H,1}, \xi_{2,2}, \xi_{3,2}, \ldots, \xi_{H,2}, \xi_{2,3}, \ldots$ is a martingale difference sequence. Thus, with probability $1-\delta$, $\sum_{k=1}^{K}\sum_{h=1}^{H-1} \xi_{h+1,k} \le H\sqrt{2K(H-1)\log(1/\delta)}$. Consider the event when this inequality holds and when $P \in \cap_{k\in[K]}\mathcal{B}_k$. By using Corollary 7 and a union bound, this event holds with probability at least $1-2\delta$. On this event, by (12) and Lemma 9, we obtain

$$R_K \le \alpha + H(d \wedge K(H-1)) + 4\sqrt{d\beta_K K(H-1)} + H\sqrt{2K(H-1)\log(1/\delta)}\,.$$

Using $\alpha \le 1$, which holds by assumption, finishes the proof. □

A.5.2. Proof of Corollary 2

Proof. Note that

$$\begin{aligned}
\|f_{P'} - f_{P''}\|_\infty &= \sup_{s,a,v} \left|\int (P'_a(ds'|s) - P''_a(ds'|s))v(s')\right| \le H \sup_{s,a} \int |P'_a(ds'|s) - P''_a(ds'|s)| \\
&= H \sup_{s,a} \|P'_a(s) - P''_a(s)\|_1 =: H\|P' - P''\|_{\infty,1}\,.
\end{aligned}$$

For $\alpha > 0$ let $\mathcal{N}(\mathcal{P}, \alpha, \|\cdot\|_{\infty,1})$ denote the $(\alpha, \|\cdot\|_{\infty,1})$-covering number of $\mathcal{P}$. Then we have

$$\mathcal{N}(\mathcal{F}, \alpha, \|\cdot\|_\infty) \le \mathcal{N}(\mathcal{P}, \alpha/H, \|\cdot\|_{\infty,1}).$$

Then, by Corollary 7,

$$\beta_K = 2H^2\log(2\mathcal{N}(\mathcal{F}, \alpha, \|\cdot\|_\infty)/\delta) + C \le 2H^2\log(2\mathcal{N}(\mathcal{P}, \alpha/H, \|\cdot\|_{\infty,1})/\delta) + C$$

with some universal constant $C > 0$. Let $f : (\Theta, \|\cdot\|) \to (\mathcal{P}, \|\cdot\|_{\infty,1})$ be defined by $\theta \mapsto \sum_j \theta_j P_j$. Note that $\|f(\theta) - f(\theta')\|_{\infty,1} \le \sup_{s,a}\sum_j \|(\theta_j - \theta'_j)P_{j,a}(s)\|_1 = \sum_j |\theta_j - \theta'_j| = \|\theta - \theta'\|_1$. Hence, any $(\epsilon, \|\cdot\|_1)$ covering of Θ induces an $(\epsilon, \|\cdot\|_{\infty,1})$-covering of $\mathcal{P}$ and so $\mathcal{N}(\mathcal{P}, \alpha/H, \|\cdot\|_{\infty,1}) \le \mathcal{N}(\Theta, \alpha/H, \|\cdot\|_1) \le C'(RH/\alpha)^d$ with some universal constant $C' > 0$.

Now, choose $1/\alpha = K\sqrt{\log(KH/\delta)}$. Hence,

$$\beta_K \le 2H^2(\log(2C'/\delta) + d\log(RH/\alpha)) + C\,.$$

Suppressing $\log$ factors (e.g., $\log(RH)$), $\log\log$ terms and constants, we have $\beta_K = H^2(d + \log(1/\delta))$.

Let $\mathcal{F}$ be given by (11). We now bound $\dim_{\mathcal{E}}(\mathcal{F}, \alpha)$. Let $\mathcal{X} = \mathcal{S}\times\mathcal{A}\times B(\mathcal{S})$ as before. Define $z : \mathcal{S}\times\mathcal{A}\times B(\mathcal{S}) \to \mathbb{R}^d$ using $z(s,a,v)_j = \langle P_{j,a}(s), v\rangle$ and note that if $x \in \mathcal{X}$ is $(\epsilon, \mathcal{F})$-independent of $x_1, \ldots, x_k \in \mathcal{X}$ then $z(x) \in \mathbb{R}^d$ is (ϵ, Θ)-independent of $z(x_1), \ldots, z(x_k) \in \mathbb{R}^d$. This holds because if $P = \sum_j \theta_j P_j \in \mathcal{P}$ then $f_P(s,a,v) = \langle \theta, z(s,a,v)\rangle$

for any $(s,a,v) \in \mathcal{X}$. Hence, $\dim_{\mathcal{E}}(\mathcal{F},\alpha) \le \dim_{\mathcal{E}}(\mathrm{Lin}(\mathcal{Z},\Theta),\alpha)$, where $\mathrm{Lin}(\mathcal{Z},\Theta)$ is the set of linear maps with domain $\mathcal{Z} = \{z(x) : x \in \mathcal{X}\} \subset \mathbb{R}^d$ and parameter from Θ: $\mathrm{Lin}(\mathcal{Z},\Theta) = \{h : h : \mathcal{Z} \to \mathbb{R} \text{ s.t. } \exists \theta \in \Theta : h(z) = \langle \theta, z\rangle, z \in \mathcal{Z}\}$. Now, by Proposition 11 of Russo & Van Roy (2014), $\dim_{\mathcal{E}}(\mathrm{Lin}(\mathcal{Z},\Theta),\alpha) = O(d\log(1+(S\gamma/\alpha)^2)$ where S is the $\|\cdot\|_2$ diameter of Θ and $\gamma = \sup_{z\in\mathcal{Z}} \|z\|_2$. We have

$$\|z\|_2^2 = \sum_j (\langle P_{j,a}(s), v\rangle)^2 \le H^2 d\,,$$

hence $\gamma \le H\sqrt{d}$. By the relation between the 1 and 2 norms, the 2-norm diameter of Θ is at most $\sqrt{d}R$. Dropping $\log$ terms, $\dim_{\mathcal{E}}(\mathcal{F},\alpha) = \tilde{O}(d)$.

Plugging into Theorem 1 gives the desired result. □

A.6. Proof of Theorem 6

Recall the following:

Definition 3. *A random variable X is σ-subgaussian if for all $\lambda \in \mathbb{R}$, it holds that $\mathbb{E}[\exp(\lambda X)] \le \exp\left(\lambda^2\sigma^2/2\right)$.*

The proof of the next couple of statements is standard and is included only for completeness.

Theorem 10. *If X is σ-subgaussian, then for any $\lambda > 0$, with probability at least $1-\delta$,*

$$X < \frac{1}{\lambda}\log\left(\frac{1}{\delta}\right) + \lambda\frac{\sigma^2}{2}\,. \tag{13}$$

Proof. Let $\lambda > 0$. We have, $\{X \ge \epsilon\} = \{\exp(\lambda(X-\epsilon)) \ge 0\}$. Hence, Markov's inequality gives $\mathbb{P}(X \ge \epsilon) \le \exp(-\lambda\epsilon)\mathbb{E}[\exp(\lambda X)] \le \exp(-\lambda\epsilon + \frac{1}{2}\lambda^2\sigma^2)$. Equating the right-hand side with δ and solving for ϵ, we get that $\log(\delta) = -\lambda\epsilon + \frac{1}{2}\lambda^2\sigma^2$. Solving for ϵ gives $\epsilon = \log(1/\delta)/\lambda + \frac{\sigma^2}{2}\lambda$, finishing the proof. □

Choosing the λ that minimizes the right-hand side of the bound gives the usual form:

$$\mathbb{P}(X \ge \sqrt{2\sigma^2\log(1/\delta)}) \le \delta\,. \tag{14}$$

Lemma 11 (Lemma 5.4 of Lattimore & Szepesvári (2020))**.** *Suppose that X is σ-subgaussian and X_1 and X_2 are independent and σ_1 and σ_2-subgaussian, respectively, then:*

1. $\mathbb{E}[X] = 0$.
2. *cX is $|c|\sigma$-subgaussian for all $c \in \mathbb{R}$.*
3. *$X_1 + X_2$ is $\sqrt{\sigma_1^2+\sigma_2^2}$-subgaussian.*

Let $(Z_p)_p$ be an $\mathbb{F} = (\mathbb{F}_p)_p$-adapted process. Recall that $(Z_p)_p$ is conditionally σ-subgaussian given $\mathbb{F}$ if for all $p \ge 1$,

$$\log\mathbb{E}[\exp(\lambda Z_p)|\mathbb{F}_{p-1}] \le \frac{1}{2}\lambda^2\sigma^2\,, \quad \text{for all } \lambda \in \mathbb{R}\,.$$

A standard calculation gives that $S_t = \sum_{p=1}^t Z_p$ is $\sqrt{t}\sigma$-subgaussian (essentially, a refinement of the calculation that is need to show Part (3) of Lemma 11) and thus, in particular, for any $t \ge 1$ and $\lambda > 0$, with probability $1-\delta$,

$$S_t < \frac{1}{\lambda}\log\left(\frac{1}{\delta}\right) + \lambda\frac{t\sigma^2}{2}\,.$$

In fact, by slightly strengthening the argument, one can show that the above inequality holds simultaneously for all $t \ge 1$:

Theorem 12 (E.g., Lemma 7 of Russo & Van Roy (2014))**.** *Let $\mathbb{F}$ be a filtration and let $(Z_p)_p$ be an $\mathbb{F}$-adapted, conditionally σ-subgaussian process. Then for any $\lambda > 0$, with probability at least $1-\delta$, for all $t \ge 1$,*

$$S_t < \frac{1}{\lambda}\log\left(\frac{1}{\delta}\right) + \lambda\frac{t\sigma^2}{2}\,, \tag{15}$$

where $S_t = \sum_{p=1}^t Z_p$.

Proof of Theorem 6 Let us introduce the following helpful notation: For vectors $x, y \in \mathbb{R}^t$, let $\langle x, y\rangle_t = \sum_{p=1}^t x_p y_p$, $\|x\|_t^2 = \langle x, x\rangle_t$, and for $f : \mathcal{X} \to \mathbb{R}$, $\|f\|_t^2 = \sum_{p=1}^t f^2(X_p)$. More generally, we will overload addition and subtraction such that for $x \in \mathbb{R}^t$, $x + f \in \mathbb{R}^t$ is the vector whose pth coordinate is $x_p + f(X_p)$ (x_p and X_p both appear on purpose here). We also overload $\langle \cdot, \cdot\rangle_t$ such that $\langle x, f\rangle_t = \langle f, x\rangle_t = \sum_{p=1}^t x_p f(X_p)$.

Define Z_p using $Y_p = f_*(X_p) + Z_p$ and collect $(Y_p)_{p=1}^t$ and $(Z_p)_{p=1}^t$ into the vectors Y and Z. As in the statement of the theorem, let $\mathbb{F} = (\mathbb{F}_p)_{p=0,1,\dots}$ be such that for any $s \geq 1$, $(X_1, Y_1, \dots, X_{p-1}, Y_{p-1}, X_p)$ is $\mathbb{F}_{p-1}$-measurable. Note that for any $p \geq 1$, $Z_p = Y_p - f_*(X_p)$ is $\mathbb{F}_p$-measurable, hence $(Z_p)_{p\geq 1}$ is $\mathbb{F}$-adapted.

With this, elementary calculation gives

$$\|Y - f\|_t^2 - \|Y - f_*\|_t^2 = \|f_* - f\|_t^2 + 2\langle Z, f_* - f\rangle_t \,.$$

Splitting $\|f_* - f\|_t^2$ and rearranging gives

$$\frac{1}{2}\|f_* - f\|_t^2 = \|Y - f\|_t^2 - \|Y - f_*\|_t^2 + E(f) \tag{16}$$

where

$$E(f) = -\frac{1}{2}\|f_* - f\|_t^2 + 2\langle Z, f - f_*\rangle_t \,.$$

Recall that $\hat{f}_t = \operatorname{argmin}_{f\in\mathcal{F}} \|Y - f\|_t^2$. Plugging $\hat{f}_t$ into 16 in place of f and using that thanks to $f_* \in \mathcal{F}$, $\|Y - \hat{f}_t\|_t^2 \leq \|Y - f_*\|_t^2$, we get

$$\frac{1}{2}\|f_* - \hat{f}_t\|_t^2 \leq E(\hat{f}_t) \,. \tag{17}$$

Thus, it remains to bound $E(\hat{f}_t)$. For this fix some $\alpha > 0$ to be chosen later and let $\mathcal{G}(\alpha) \subset \mathcal{F}$ be an α-cover of $\mathcal{F}$ in $\|\cdot\|_\infty$. Let $g \in \mathcal{G}(\alpha)$ be a random function, also to be chosen later. We have

$$E(\hat{f}_t) = E(\hat{f}_t) - E(g) + E(g) \leq E(\hat{f}_t) - E(g) + \max_{\tilde{g}\in\mathcal{G}(\alpha)} E(\tilde{g}) \tag{18}$$

We start by bounding the last term above. A simple calculation gives that for any fixed $f \in \mathcal{F}$, w.p. $1-\delta$, $2\langle Z, f - f_*\rangle_t$ is $2\sigma\|f - f_*\|_t$-subgaussian. Hence, with probability $1 - \delta$, simultaneously for all $t \geq 1$,

$$E(f) \leq -\frac{1}{2}\|f_* - f\|_t^2 + \frac{1}{\lambda}\log\left(\frac{1}{\delta}\right) + \lambda\,\frac{4\sigma^2\|f - f_*\|_t^2}{2} = 4\sigma^2\log\left(\frac{1}{\delta}\right) \,,$$

where the equality follows by choosing $\lambda = 1/(4\sigma^2)$ (which makes the first and last terms cancel). (Note how splitting $\|f - f_*\|_t^2$ into two halves allowed us to bound the "error term" $E(f)$ independently of t.) Now, by a union bound, it follows that with probability at least $1 - \delta$, the second term is bounded by $4\sigma^2\log(|\mathcal{G}(\alpha)|/\delta)$.

Let us now turn to bounding the first term. We calculate

$$\begin{aligned}
E(\hat{f}_t) - E(g) &= \frac{1}{2}\|g - f_*\|_t^2 - \frac{1}{2}\|\hat{f}_t - f_*\|_t^2 + 2\langle Z, \hat{f}_t - g\rangle_t \\
&\leq \frac{1}{2}\left(\langle g - \hat{f}_t, g + \hat{f}_t + 2f_*\rangle_t\right) + 2\|Z\|_t\|\hat{f}_t - g\|_t \\
&\leq \frac{1}{2}4C\alpha\, t + 2\|Z\|_t\alpha\sqrt{t} \,,
\end{aligned}$$

where for the last inequality we chose $g = \operatorname{argmin}_{\tilde{g}\in\mathcal{G}(\alpha)} \|\hat{f}_t - \tilde{g}\|_\infty$ so that $\|\hat{f}_t - g\|_t \leq \alpha\sqrt{t}$ and used Cauchy-Schwartz, together with that $\|g\|_t, \|\hat{f}_t\|_t, \|f_*\|_t \leq C\sqrt{t}$, which follows from $g, \hat{f}_t, f_* \in \mathcal{F}$ and that by assumption all functions in $\mathcal{F}$ are bounded by C.

It remains to bound $\|Z\|_t$. For this, we observe that with probability $1-\delta$, simultaneously for all $t \geq 1$,

$$\|Z\|_t \leq \sigma\sqrt{2t\log(2t(t+1)/\delta)}\,.$$

Indeed, this follows because with probability $1-\delta$, simultaneously for any $s \geq 1$, $|Z_p|^2 \leq 2\sigma^2\log(2s(s+1)/\delta)$ holds because of a union bound and Eq. (14). Therefore, for the above choice g, with probability $1-\delta$, simultaneously for all $t \geq 1$, it holds that

$$E(\hat{f}_t) - E(g) \leq 2C\alpha\, t + 2t\alpha\sqrt{\sigma^2\log(2t(t+1)/\delta)}\,.$$

Merging this with Eqs. (17) and (18) and with another union bound, we get that with probability $1-\delta$, for any $t \geq 1$,

$$\|f_* - \hat{f}_t\|_t^2 \leq 8\sigma^2\log(2N_\alpha/\delta) + 4t\alpha\left(C + \sqrt{\sigma^2\log(4t(t+1)/\delta)}\right)\,,$$

where N_α is the $(\alpha, \|\cdot\|_\infty)$-covering number of $\mathcal{F}$. □

B. Proof of Theorem 3

In this section we establish a regret lower bound by reduction to a known result for tabular MDP.

Proof. We assume without loss of generality that d is a multiple of 4 and $d \geq 8$. We set $S = 2$ and $A = d/4 \geq 2$. According to (Azar et al., 2017), (Osband & Van Roy, 2016), there exists an MDP $\mathcal{M}(\mathcal{S}, \mathcal{A}, P, r, H)$ with S states, A actions and horizon H such that any algorithm has regret at least $\Omega(\sqrt{HSAT})$. In this case, we have $|\mathcal{S}\times\mathcal{A}\times\mathcal{S}| = d$. We use $\sigma(s,a,s')$ to denote the index of (s,a,s') in $\mathcal{S}\times\mathcal{A}\times\mathcal{S}$. Letting

$$P_i(s'|s,a) = \begin{cases} 1 & \text{if } \sigma(s,a,s') = i, \\ 0 & \text{otherwise,} \end{cases}$$

and $\theta^i = P(s'|s,a)$ if $\sigma(s,a,s') = i$, we will have $P(s'|s,a) = \sum_{i=1}^d \theta^i P_i(s'|s,a)$. Therefore P can be parametrized using (19). Therefore, the known lower bound $\Omega(\sqrt{HSAT})$ implies a worst-case lower bound of $\Omega(\sqrt{H\cdot d/2\cdot T}) = \Omega(\sqrt{HdT})$ for our model. □

C. The Special Case of Linear Transition Models

We derive a modification of UCRL-VTR when P_θ is a linear model of the form $P_\theta = \sum_{j=1}^d \theta_j P_j$, which is captured in the following assumption:

Assumption 3 (Linear Parameterized Transition Model)**.** *There exists a vector $\theta_* \in \mathbb{R}^d$ such that $\|\theta_*\|_2 \leq C_\theta$ ($C_\theta \geq 1$) and*

$$P(s'|s,a) = \sum_{j=1}^d (\theta_*)_j P_j(s'|s,a) = P_\cdot(s'|s,a)^\top\theta_*, \tag{19}$$

where P_j's are known basis models such that $\sup_{j\in[d],(s,a)\in\mathcal{S}\times\mathcal{A}}\|P_j(\cdot|s,a)\|_1 \leq 1$, and $P_\cdot(s'|s,a)$ denotes the d-dimensional vector $P_\cdot(s'|s,a) = [P_1(s'|s,a),\ldots,P_d(s'|s,a)]^\top$[3]. Note that we do not require each basis model P_j to be a probability transition model.

By modifying the algorithm and using optimistic Q-update, we obtain an algorithm that can be implemented using efficient recursive update. See Algorithm 3 for full details of implementation.

Estimating θ_* by recursive regression. We let $X_{h,k}^\top\theta := \mathbb{E}_\cdot[V_{h+1,k}(s)|s_h^k, a_h^k]^\top\theta = \langle P_\theta(\cdot|s,a), V_{h+1,k}\rangle$ be the predicted expected value of next state. In this case, each new observation adds the following loss to regression:

$$\begin{aligned}&\left(X_{h,k}^\top\theta - y_{h,k}\right)^2 : \\ &= \left(\mathbb{E}_\cdot[V_{h+1,k}(s)|s_h^k, a_h^k]^\top\theta - V_{h+1,k}(s_{h+1}^k)\right)^2\end{aligned}$$

[3]We also use $P_\cdot(\cdot|s,a)$ to denote a $d\times S$ matrix.

Algorithm 3 UCRL-VTR with linear transition model

Input: MDP, $d, H, T = KH$;
Initialize: $M_{1,1} \leftarrow H^2 dI, \quad w_{1,1} \leftarrow 0 \in \mathbb{R}^{d\times 1}, \quad \theta_1 \leftarrow M_{1,1}^{-1} w_{1,1} \qquad$ for $1 \le h \le H$;
Initialize: $\delta \leftarrow 1/K$, and for $1 \le k \le K$,

$$\beta_k \leftarrow 16 C_\theta^2 H^2 d \log(1+Hk) \log^2((k+1)^2 H/\delta);$$

Compute Q-function $Q_{h,1}$ using $\theta_{1,1}$ according to (3);
for $k = 1 : K$ **do**
  Obtain initial state s_1^k for episode k;
  for $h = 1 : H$ **do**
    Choose action greedily by

$$a_h^k = \arg\max_{a\in\mathcal{A}} Q_{h,k}(s_h^k, a)$$

    and observe the next state s_{h+1}^k.
    Compute the predicted value vector: ▷ Evaluate the expected value of next state

$$\begin{aligned} X_{h,k} &\leftarrow \mathbb{E}_{\cdot}[V_{h+1,k}(s)|s_h^k, a_h^k] \\ &= \sum_{s\in\mathcal{S}} V_{h+1,k}(s) \cdot P_{\cdot}(s|s_h^k, a_h^k). \end{aligned}$$

    $y_{h,k} \leftarrow V_{h+1,k}(s_{h+1}^k)$ ▷ Update regression parameters
    $M_{h+1,k} \leftarrow M_{h,k} + X_{h,k} X_{h,k}^\top$
    $w_{h+1,k} \leftarrow w_{h,k} + y_{h,k} \cdot X_{h,k}$
  end for
  Update at the end of episode: ▷ Update Model Parameters

$$\begin{aligned} M_{1,k+1} &\leftarrow M_{H+1,k}, \\ w_{1,k+1} &\leftarrow w_{H+1,k}, \\ \theta_{k+1} &\leftarrow M_{1,k+1}^{-1} w_{1,k+1}; \end{aligned}$$

  Compute $Q_{h,k+1}, h = H, \ldots, 1$, using θ_{k+1} according to (20) ▷ Computing Q functions
end for

By aggregating the value prediction losses constructed from all past experiences, we formulate a ridge regression problem to estimate θ_* by

$$\begin{aligned}&\theta_{k+1}\\&= \arg\min_{\theta\in\mathbb{R}^d}\left[\theta^\top M_{1,1}\theta + \sum_{(h',k')\le(H,k)}\left(X_{h',k'}^\top\theta - y_{k',h'}\right)^2\right],\end{aligned}$$

where $M_{1,1} = H^2 dI$ acts as a regularization term.

To solve the above regression problem, we can first calculate $X_{h',k'}$ and recursively compute estimates of θ_* by letting

$$\begin{aligned}M_{1,k+1} &= M_{1,1} + \sum_{(h',k')\le(H,k)} X_{h',k'}X_{h',k'}^\top\\ w_{1,k+1} &= w_{1,1} + \sum_{(h',k')\le(H,k)} y_{h',k'}\cdot X_{h',k'},\end{aligned}$$

with $M_{1,1} = H^2 d\cdot I$ and $w_{1,1} = 0$. Then we obtain the estimated θ_{k+1} easily by

$$\theta_{k+1} = M_{1,k+1}^{-1} w_{k+1}.$$

Confidence ball. We construct B_k as follows:

$$B_k = \{\theta | (\theta-\theta_k)^\top M_k(\theta-\theta_k)\le\beta_k\}.$$

where β_k is preselected (see the algorithm).

Our model parameter update, θ_k and M_k, can be via a recursive update in an incremental fashion. In this way, one does not need to re-train the model parameter from scratch every episode. A similarly simple recursion was used in (Jin et al., 2019) for model-free Q learning. Our method differs in that our Q functions cannot be parameterized by d parameters and our updates are made on the transition model rather than Q functions.

Optimistic Q-update. Instead of solving the optimistic planning problem $\theta_k = \text{argmax}_\theta\{V_\theta^*(s_1)|\theta\in B_k\}$ as in Algorithm 1, we incorporate optimism into iterative Q-update:

$$\begin{aligned}&Q_{H+1,k}(s,a) = 0,\\&V_{h,k}(s) = \max_{a\in\mathcal{A}} Q_{h,k}(s,a),\\&Q_{h,k}(s,a) = r(s,a) + \max_{\theta\in B_k}\sum_{j=1}^d(\theta)_j P_j(\cdot|s,a)V_{h+1,k}.\end{aligned}$$

Since the confidence sets are ellipsoids, the preceding Q update has a closed-forms solution

$$\begin{aligned}&Q_{h,k}(s,a)\\&= r(s,a) + \max_{\theta\in B_k}\langle P_\theta(\cdot|s,a), V_{h+1,k}\rangle\\&= r(s,a) + X_{h,k}^\top\theta_k + \sqrt{\beta_k}\sqrt{X_{h,k}^\top M_k^{-1}X_{h,k}}.\end{aligned}\tag{20}$$

The last term in the above is the "bonus" term that quantifies uncertainty and encourages exploration. This optimistic Q value allows us to greedily pick actions while sufficiently exploring the state space.

Algorithm 3 is a modification of UCRL-VTR and uses a different construction of confidence set. we provide an independent regret analysis using techniques from linear bandit theory. The next theorem gives a egret upper bound for Algorithm 3.

Theorem 13. *Let Assumption 3 hold. If we choose*

$$\beta_k = \left(H\sqrt{d\log\left(\frac{1+Hk\cdot H^2 d}{\delta}\right)} + C_\theta H\sqrt{d}\right)^2,$$

then T-time-step regret of Algorithm 1 satisfies

$$\mathbb{E}\left[R(T)\right] = \tilde{\mathcal{O}}\left(C_\theta \cdot d\sqrt{H^3 T}\right),$$

where $T = HK$ is the total number of steps in K episodes, C_θ ($C_\theta \geq 1$) is a known constant such that $\|\theta_*\| \leq C_\theta$ and $\tilde{\mathcal{O}}$ hides polylog factors of H, T.

Let us outline the proof ideas. In the first part of the proof, we show that if $\theta_* \in B_{h,k}$, then the estimated Q-functions are optimistic estimates of the true Q-value functions. That is, $Q_{h,k}(s)$ is greater than the true Q-value $Q_h(s)$ for every $s \in \mathcal{S}$. Using this fact, we can bound the regret by the sum of $Q_{1,k}(s_1^k) - Q_1^{\pi_k}(s_1^{\pi_k})$, which can be decomposed into the sum of state-action confidence bounds on the sample path. In the second part, we construct martingale difference sequences and apply a concentration argument to show that $\theta_* \in B_{h,k}$ for all (h, k) with high probability. The full proof is deferred to the Appendix E.

D. Proof of Theorem 4

In this section, we will present the full proof of Theorem 4. To handle the mispecification error, we will modify the bonus term by replacing it with

$$\beta_k = 8H^2 \log\left(\frac{2\mathcal{N}(\mathcal{F}, \alpha, \|\cdot\|_\infty)}{\delta}\right) + 4H(kH-1)\alpha\left\{2 + \sqrt{\log\left(\frac{4kH(kH-1)}{\delta}\right)}\right\} + 8H^3 k\varepsilon^2.$$

The last term in the above choice of β_k can be viewed as an "error tolerance."

Next we show that $P^* \in B_k$ with high probability.

We first present a theorem which is nearly identical to Theorem 6 but tolerates misspecification. We use the same notations as in the proof of Theorem 6.

Theorem 14. *Let $\mathbb{F}$ be the filtration defined above and assume that the functions in $\mathcal{F}$ and also f_* are all bounded by the positive constant $C > 0$ at values X_t for all t. Assume that there exists $\tilde{f} \in \mathcal{F}$ such that $|\tilde{f}(X) - f_*(X)| \leq \zeta$ for all $X = (s, a, v)$ with $\|v\|_\infty \leq H$, and also for each $s \geq 1$, $(Y_p - f_*(X_p))_p$ is conditionally σ-subgaussian given $\mathbb{F}_{p-1}$. We define*

$$\hat{f}_t = \arg\min_{f \in \mathcal{F}} \sum_{p=1}^{t} (f(X_p) - Y_p)^2$$

and

$$\mathcal{F}_t(\beta) = \left\{ f : \mathcal{X} \to \mathbb{R}, s.t. \sum_{p=1}^{t} (f(X_p) - \hat{f}(X_p))^2 \leq \beta \right\}.$$

Then, for any $\alpha > 0$, with probability $1 - \delta$, for all $t \geq 1$, $f_* \in \mathcal{F}_t(\beta_t(\delta, \alpha))$, where

$$\beta_t(\delta, \alpha) = 16\sigma^2 \log(4N_\alpha/\delta) + 4t\alpha\left(C + \sqrt{\sigma^2 \log(8t(t+1)/\delta)}\right) + 3t\zeta^2.$$

Note that here the last term is due to the misspecification error.

Proof. The proof of this theorem is also nearly identical to Theorem 6, except for the modifications below.

Due to model misspecification, we no longer have $f_* \in \mathcal{F}$, and hence we may not have $\|Y - \hat{f}_t\|_t \leq \|Y - f_*\|_t$ (Here notation $\|\cdot\|_t$ is defined to be the same as the notations in Theorem 6). To handle the misspecification error, we will use the function $\tilde{f}$ as a bridge to bound the error between $\hat{f}_t$ and f_*. Hence since $\hat{f}_t = \arg\min_{f \in \mathcal{F}} \|Y - f\|_t^2$, we have $\|\hat{f}_t - Y\|_t^2 \leq \|\tilde{f} - Y\|_t^2$, which indicates that $\|\hat{f}_t - f_* - Z\|_t^2 \leq \|\tilde{f} - f_* - Z\|_t^2$. (Recall the notations $Z_p = Y_p - f_*(X_p)$ and $Z = (Z_1, \cdots, Z_p)$.) Therefore, we have

$$\|\hat{f}_t - f_*\|_t^2 - 2\langle \hat{f}_t - f_*, Z\rangle_t \leq \|\tilde{f} - f_*\|_t^2 - 2\langle \tilde{f} - f_*, Z\rangle_t.$$

We then obtain

$$\begin{aligned}\frac{1}{2}\|\hat{f}_t - f_*\|_t^2 &\le -\frac{1}{2}\|\hat{f}_t - f_*\|_t^2 + 2\langle \hat{f}_t - f_*, Z\rangle_t + \|\tilde{f} - f_*\|_t^2 - 2\langle \tilde{f} - f_*, Z\rangle_t \\ &= E(\hat{f}_t) + \tilde{E}(\tilde{f}) + \frac{3}{2}\|\tilde{f} - f_*\|_t^2,\end{aligned} \tag{21}$$

where we define

$$E(f) = -\frac{1}{2}\|f - f_*\|_t^2 + 2\langle Z, f - f_*\rangle_t \tag{22}$$

$$\tilde{E}(f) = -\frac{1}{2}\|f - f_*\|_t^2 - 2\langle Z, f - f_*\rangle_t \tag{23}$$

In the next, we will bound $E(\hat{f}_t)$ and also $\tilde{E}(\tilde{f})$. Similar to the proof of Theorem 6, we can show that

$$E(\hat{f}_t) \le 4\sigma^2 \log(|N_\alpha|/\delta) + 2C\alpha\, t + 2t\alpha\sqrt{\sigma^2 \log(2t(t+1)/\delta)},$$

holds with probability at least $1-\delta$, where N_α is the α-covering number of $\mathcal{F}$.

Now we analyze $\tilde{E}(\tilde{f})$ where $\tilde{f} \in \mathcal{F}$. Similarly, a simple calculation gives that for any fixed $f \in \mathcal{F}$, $2\langle -Z, f - f_*\rangle_t$ is $2\sigma\|f - f_*\|_t$-subgaussian. Hence, with probability $1-\delta$, simultaneously for all $t \ge 1$,

$$\tilde{E}(f) \le -\frac{1}{2}\|f_* - f\|_t^2 + 4\sigma^2 \log\left(\frac{1}{\delta}\right) + \frac{1}{4\sigma^2}\cdot\frac{4\sigma^2\|f - f_*\|_t^2}{2} = 4\sigma^2 \log\left(\frac{1}{\delta}\right),$$

which indicates that with probability at least $1-\delta$, we have

$$\tilde{E}(\tilde{f}) \le 4\sigma^2 \log\left(\frac{1}{\delta}\right).$$

Finally, as for the last term $\|\tilde{f} - f_*\|_t^2$ in (21), we have the following estimation due to the bound of the misspecification error:

$$\|\tilde{f} - f_*\|_t^2 = \sum_{p=1}^{t}(\tilde{f}(X_p) - \tilde{f}(X_p))^2 \le t\cdot\zeta^2,$$

where we use the fact that $X_p = (s_p, a_p, v_p)$ satisfies that $\|v_p\|_\infty \le H$.

We combine those bounds on the three terms in (21) above, and obtain that with probability at least $1-2\delta$, the following inequality holds:

$$\frac{1}{2}\|\hat{f} - f_*\|_t^2 \le 2tC\alpha + 2t\alpha\sqrt{\sigma^2\log(2t(t+1)/\delta)} + 8\sigma^2\log(2N_\alpha/\delta) + \frac{3}{2}t\zeta^2.$$

Finally, we switch δ into $\delta/2$ and multiply the above inequality by 2 on both sides. And the proof of Theorem 14 is completed. □

Next we apply this theorem to prove the following lemma:

Lemma 15. *For any transition model P, we define its corresponding function $f_p : \mathcal{X} \to \mathbb{R}$:*

$$f_P(s,a,v) = \int P(ds'|s,a)v.$$

Then with probability at least $1-\delta$, we have $P_* \in \mathcal{B}_k$, where

$$\mathcal{B}_k = \left\{\tilde{P} \in \mathcal{P} : \sum_{p=1}^{t}(f_{\tilde{P}}(X_t) - f_{\hat{P}_t}(X_t))^2 \le \beta_k\right\}, \quad t = k(H-1).$$

Here $\hat{P}_t$ is defined in (9) and we choose

$$\beta_k = 8H^2\log\left(\frac{4\mathcal{N}(\mathcal{F},\alpha,\|\cdot\|_\infty)}{\delta}\right) + 4H(kH-1)\alpha\left\{2 + \sqrt{\log\left(\frac{8kH(kH-1)}{\delta}\right)}\right\} + 8H^3k\varepsilon^2,$$

Proof. In the following proof, the notation of $X_t, Y_t, \mathcal{F}$ are the same as the proof of Corollary 7. We notice that $Y_t - f_P(X_t) \in [-H, H]$ for every $X_t = (s_t, a_t, V_t)$, and

$$\mathbb{E}[Y_t|\mathbb{F}_t] = \mathbb{E}[V_t(s_{t+1})|\mathbb{F}_t] = \int P(ds'|s_t, a_t)V_t(s') = f_P(s_t, a_t, v_t) = f_P(X_t).$$

Hence $Z_t = Y_t - f_P(X_t)$ is $\frac{H}{2}$-subgaussian given $\mathbb{F}_t$.

For every $f \in \mathcal{F}$, there exists some $\tilde{P} \in \mathcal{P}$ such that $f(s, a, v) = \int \tilde{P}(ds'|s, a)v(s')$, which indicates that $|f(X_t)| \le H$. Moreover, we also have $|f_P(X_t)| \le H$.

We next apply Theorem 14 with $C = H$ and $\sigma = \frac{H}{2}$ and $f_* = f_P$ and $\zeta = H\epsilon$ and $\tilde{f} = f_{P^*}$. According to Assumption 2, we notice that, for all $X = (s, a, v)$ with $\|v\|_\infty \le H$, we have

$$|f_*(X) - \tilde{f}(X)| = \left|\int (P(s'|s, a) - P^*(s'|s, a))v(s')ds'\right| \le \|P(s'|s, a) - P^*(s'|s, a)\|_1\|v\|_\infty \le H\varepsilon = \zeta.$$

Hence we have verified all the assumptions in Theorem 14. Hence we obtain that: for any $\alpha > 0$, with probability at least $1 - \delta$, for all $t \ge 1$, we have

$$\sum_{p=1}^{t}(f_P(X_p) - f_{\hat{P}_t}(X_p))^2 \le 4H^2 \log\left(\frac{4\mathcal{N}(\mathcal{F}, \alpha, \|\cdot\|_\infty)}{\delta}\right) + 2H(kH-1)\alpha\left\{2 + \sqrt{\log\left(\frac{8kH(kH-1)}{\delta}\right)}\right\} + 3H^3k\varepsilon^2.$$

Moreover, noticing that

$$(f_P(X_t) - f_{P_*}(X_t))^2 = \left(\int (P(ds'|s_t, a_t) - P^*(ds'|s_t, a_t))V_t\right)^2 \le (H\varepsilon)^2,$$

we have

$$\begin{aligned}
&\sum_{p=1}^{t}(f_{P^*}(X_p) - f_{\hat{P}_t}(X_p))^2 \le \sum_{p=1}^{t} 2(f_P(X_p) - f_{P^*}(X_p))^2 + 2(f_P(X_p) - f_{\hat{P}_t}(X_p))^2 \\
&\le 8H^2 \log\left(\frac{4\mathcal{N}(\mathcal{F}, \alpha, \|\cdot\|_\infty)}{\delta}\right) + 4H(kH-1)\alpha\left\{2 + \sqrt{\log\left(\frac{8kH(kH-1)}{\delta}\right)}\right\} + 6H^3k\varepsilon^2 + 2H^2\varepsilon^2 t \\
&\le 8H^2 \log\left(\frac{4\mathcal{N}(\mathcal{F}, \alpha, \|\cdot\|_\infty)}{\delta}\right) + 4H(kH-1)\alpha\left\{2 + \sqrt{\log\left(\frac{8kH(kH-1)}{\delta}\right)}\right\} + 8H^3k\varepsilon^2.
\end{aligned}$$

which indicates that $P^* \in \mathcal{B}_k$. This finishes the proof of this corollary. □

We next provide a lemma similar to Lemma 5, only adding the misspecification analysis.

Lemma 16. *Assuming that $P^* \in \mathcal{B}_k$, we have*

$$V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) \le \sup_{\tilde{P} \in \mathcal{B}_k} \sum_{h=1}^{H-1} \langle \tilde{P}(\cdot|s_h^k, a_h^k) - P^*(\cdot|s_h^k, a_h^k), V_{h,k}\rangle + \sum_{h=1}^{H-1} \xi_{h+1,k} + H^2\varepsilon\,,$$

where

$$\xi_{h+1,k} = \langle P(\cdot|s_h^k, a_h^k), V_{h+1,k} - V_{h+1}^{\pi_k}\rangle - \left(V_{h+1,k}(s_{h+1}^k) - V_{h+1}^{\pi_k}(s_{h+1}^k)\right)\,.$$

Note that $(\xi_{2,1}, \xi_{3,1}, \dots, \xi_{H,1}, \xi_{2,2}, \xi_{3,2}, \dots, \xi_{H,2}, \xi_{2,3}, \dots)$ is a sequence of martingale differences.

Proof. We first prove by induction that

$$V_{h,k}(s_h^k) \geq V_h^*(s_h^k) - (H+1-h)\varepsilon, \quad \forall 1 \leq h \leq H+1$$

by induction on h according to the fact that $P^* \in \mathcal{B}_k$ (but not $P \in \mathcal{B}_k$). When $h = H+1$, this inequality holds since both sides equal to 0. We assume it holds for $h+1$ and we consider the case of h. Actually we have

$$\begin{aligned}
Q_{h,k}(s_h^k) &= r(s_h^k, a_h^k) + \langle P^k(\cdot|s_h^k, a_h^k), V_{h+1,k}\rangle \geq r(s_h^k, a_h^k) + \langle P^*(\cdot|s_h^k, a_h^k), V_{h+1,k}\rangle \\
&= r(s_h^k, a_h^k) + \langle P(\cdot|s_h^k, a_h^k), V_{h+1,k}\rangle - \langle P(\cdot|s_h^k, a_h^k) - P^*(\cdot|s_h^k, a_h^k), V_{h+1,k})\rangle \\
&\geq r(s_h^k, a_h^k) + \langle P(\cdot|s_h^k, a_h^k), V_{h+1}^* - (H-h)\varepsilon \mathbf{1}\rangle - \|P(\cdot|s_h^k, a_h^k) - P^*(\cdot|s_h^k, a_h^k)\|_1 \|V_{h+1,k}\|_\infty \\
&\geq r(s_h^k, a_h^k) + \langle P(\cdot|s_h^k, a_h^k), V_{h+1}^*\rangle - (H+1-h)\xi = Q_h^*(s_h^k, a_h^k) - (H+1-h)\xi,
\end{aligned}$$

where in the third line we use the induction and in the last line we use the fact that $\|V_{h+1,k}\|_\infty \leq H$. This indicates that $V_{h,k}(s_h^k) \geq V_h^*(s_h^k) - (H+1-h)\varepsilon$, which completes the induction at h. Hence we know that $V_{h,k}(s_h^k) \geq V_h^*(s_h^k) - (H+1-h)\varepsilon$ holds for all $1 \leq h \leq H+1$.

Therefore,

$$V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) \leq V_{1,k}(s_1^k) - V_1^{\pi_k}(s_1^k) + H\varepsilon\,.$$

Fix $h \in [H]$. In what follows we bound $V_{h,k}(s_h^k) - V_h^{\pi_k}(s_h^k)$. By the definition of π_k, P^k and a_h^k, we have

$$\begin{aligned}
V_{h,k}(s_h^k) &= r(s_h^k, a_h^k) + \langle P^k(\cdot|s_h^k, a_h^k), V_{h+1,k}\rangle \text{ and} \\
V_h^{\pi_k}(s_h^k) &= r(s_h^k, a_h^k) + \langle P(\cdot|s_h^k, a_h^k), V_{h+1}^{\pi_k}\rangle\,.
\end{aligned}$$

Hence,

$$\begin{aligned}
V_{h,k}(s_h^k) - V_h^{\pi_k}(s_h^k) &= \langle P^k(\cdot|s_h^k, a_h^k), V_{h+1,k}\rangle - \langle P_{a_h^k}(s_h^k), V_{h+1}^{\pi_k}\rangle \\
&= \langle P^k(\cdot|s_h^k, a_h^k) - P(\cdot|s_h^k, a_h^k), V_{h+1,k}\rangle + \langle P(\cdot|s_h^k, a_h^k), V_{h+1,k} - V_{h+1}^{\pi_k}\rangle\,.
\end{aligned}$$

Therefore, by induction, noting that $V_{H+1,k} = 0$, we get that

$$\begin{aligned}
V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) &\leq \sum_{h=1}^{H-1} \langle P^k(\cdot|s_h^k, a_h^k) - P(\cdot|s_h^k, a_h^k), V_{h+1,k}\rangle + \sum_{h=1}^{H-1} \xi_{h+1,k} + H\varepsilon \\
&\leq \sup_{\tilde{P} \in \mathcal{B}_k} \sum_{h=1}^{H-1} \langle \tilde{P}(\cdot|s_h^k, a_h^k) - P(\cdot|s_h^k, a_h^k), V_{h+1,k}\rangle + \sum_{h=1}^{H-1} \xi_{h+1,k} + H\varepsilon\,.
\end{aligned}$$

Finally noticing that

$$\begin{aligned}
\langle \tilde{P}(\cdot|s_h^k, a_h^k) - P(\cdot|s_h^k, a_h^k), V_{h+1,k}\rangle &= \langle \tilde{P}(\cdot|s_h^k, a_h^k) - P^*(\cdot|s_h^k, a_h^k), V_{h+1,k}\rangle + \langle \tilde{P}^*(\cdot|s_h^k, a_h^k) - P(\cdot|s_h^k, a_h^k), V_{h+1,k}\rangle \\
&\leq \langle \tilde{P}(\cdot|s_h^k, a_h^k) - P^*(\cdot|s_h^k, a_h^k), V_{h+1,k}\rangle + H\varepsilon,
\end{aligned}$$

we have

$$\begin{aligned}
V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) &\leq \sum_{h=1}^{H-1} \langle \tilde{P}(\cdot|s_h^k, a_h^k) - P^*(\cdot|s_h^k, a_h^k), V_{h+1,k}\rangle + \sum_{h=1}^{H-1} \xi_{h+1,k} + \frac{H(H-1)}{2}\xi + H\varepsilon \\
&\leq \sum_{h=1}^{H-1} \langle \tilde{P}(\cdot|s_h^k, a_h^k) - P^*(\cdot|s_h^k, a_h^k), V_{h+1,k}\rangle + \sum_{h=1}^{H-1} \xi_{h+1,k} + H^2\varepsilon,
\end{aligned}$$

which completes the proof of this lemma. □

Equipped with these two lemmas, we are ready to prove Theorem 4.

Proof of Theorem 4. According to Lemma 15, we learn that $P^* \in \mathcal{B}_k$ holds with probability at least $1-\delta$. We next assume $P^* \in \mathcal{B}_k$ and bound the error $V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k)$.

According to Lemma 16, we have

$$V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) \le \sup_{\tilde{P}\in\mathcal{B}_k} \sum_{h=1}^{H-1} \langle \tilde{P}(\cdot|s_h^k, a_h^k) - P^*(\cdot|s_h^k, a_h^k), V_{h,k}\rangle + \sum_{h=1}^{H-1} \xi_{h+1,k} + H^2\varepsilon\,, \tag{24}$$

where

$$\xi_{h+1,k} = \langle P(\cdot|s_h^k, a_h^k), V_{h+1,k} - V_{h+1}^{\pi_k}\rangle - \left(V_{h+1,k}(s_{h+1}^k) - V_{h+1}^{\pi_k}(s_{h+1}^k)\right)\,.$$

We let

$$W_k = \sup_{\tilde{P}\in\mathcal{B}_k} \sum_{h=1}^{H-1} \langle \tilde{P}(\cdot|s_h^k, a_h^k) - P^*(\cdot|s_h^k, a_h^k), V_{h,k}\rangle,$$

and summing h from 1 to H in (24) we obtain the following bound on the regret up to horizon K:

$$R_K = \sum_{k=1}^{K} V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) \le \sum_{k=1}^{K} W_k + \sum_{k=1}^{K}\sum_{h=1}^{H-1} \xi_{h+1,k} + H^2K\varepsilon$$

We next bound $\sum_{k=1}^K W_k$. Actually we have

$$\begin{aligned}\sum_{k=1}^{K} W_k &= \sum_{k=1}^{K} \sup_{\tilde{P}\in\mathcal{B}_k} \sum_{h=1}^{H-1} \langle \tilde{P}(\cdot|s_h^k, a_h^k) - P^*(\cdot|s_h^k, a_h^k), V_{h,k}\rangle \\ &\le \sum_{k=1}^{K}\sum_{h=1}^{H-1} \sup_{\tilde{P}\in\mathcal{B}_k} \langle \tilde{P}(\cdot|s_h^k, a_h^k) - P^*(\cdot|s_h^k, a_h^k), V_{h,k}\rangle\end{aligned},$$

and each term inside satisfies

$$\sup_{\tilde{P}\in\mathcal{B}_k} \langle \tilde{P}(\cdot|s_h^k, a_h^k) - P^*(\cdot|s_h^k, a_h^k), V_{h,k}\rangle \le \mathrm{diam}(\tilde{\mathcal{F}}_t|_{X_t})$$

where

$$\tilde{\mathcal{F}}_t = \left\{ f = f_P : P \in \mathcal{P}, \sum_{p=1}^{t} (f(X_p) - f_{\hat{P}_t}(X_p))^2 \le \beta_k \right\}$$

We notice that $\mathcal{F}_t \subset \mathcal{F} \subset \mathcal{B}_\infty(\mathcal{X}, H)$. Hence we apply Lemma 9 and obtain that

$$\sum_{k=1}^{K} \mathrm{diam}(\tilde{\mathcal{F}}_t|_{X_t}) \le \alpha + H(d \wedge K(H-1)) + 2\delta_{K(H-1)}\sqrt{dK(H-1)},$$

where $\delta_{K(H-1)} = \max_{1\le t\le K(H-1)} \mathrm{diam}(\tilde{\mathcal{F}}_t|_{X_t})$. Thanks to the definition of $\tilde{\mathcal{F}}_t$, $\delta_{K(H-1)} \le 2\sqrt{\beta_K}$. Plugging this into the previous display finishes the proof.

Moreover, we also have $\sum_{k=1}^K \sum_{h=1}^{H-1} \xi_{h+1,k} \le H\sqrt{2K(H-1)\log(1/\delta)}$ holds with probability at least $1-\delta$. Hence combine these two inequality together, we obtain that with probability at least $1-2\delta$, the following bound holds

$$\begin{aligned} R_K &= \sum_{k=1}^{K} V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) \\ &\le \sum_{k=1}^{K} \sup_{\tilde{P}\in\mathcal{B}_k} \sum_{h=1}^{H-1} \langle \tilde{P}_{a_h^k}(s_h^k) - P_{a_h^k}(s_h^k), V_{h,k}\rangle + \sum_{k=1}^{K}\sum_{h=1}^{H-1} \xi_{h+1,k} + H^2K\varepsilon \\ &\le \alpha + H(d\wedge K(H-1)) + 4\sqrt{d\beta_K K(H-1)} + H\sqrt{2K(H-1)\log(1/\delta)} + H^2K\varepsilon, \end{aligned}$$

where we use $\alpha \le 1$. □

E. Proof of Theorem 13

Here we will provide the formal regret analysis for Algorithm 3, which differs from Algorithm 1. By leveraging the linear structures, we provide an independent proof of Theorem 13 using an analysis adapted from linear bandits.

The full proof is divided into five parts in the following five subsections respectively. In the first subsection, we decompose the regret into the sum of bonuses assuming the Q-functions indeed are optimistic estimates. In the second subsection, we discover some important properties of our algorithm. We provide an upper bound to the sum of bonuses in the third subsection. In the fourth subsection, we will prove that the optimism holds with high probability by constructing a particular martingale and showing that it concentrates, and in the final subsection, we will put together all the analysis to finish the proof of upper bound of expected regret.

We say $(h,k) \leq (h',k')$ if $k < k'$ or $k = k', h \leq h'$. Thus, $\leq$ stands for the lexicographic order with k being the variable that takes priority. We say $(h,k) < (h',k')$ if $k < k'$ or $k = k', h < h'$.Let $\mathbb{F}_{h,k}$ be the filtration generated by the random sample path $\{(s_{h'}^{k'}, a_{h'}^{k'}, r_{h'}^{k'})\}_{(h',k') \leq (h,k)}$.

E.1. Regret Analysis

The proof in this section is similar to Lemma 5. Throughout E.1 to E.3, we assume that $\theta_* \in B_k$ for all $1 \leq k \leq K$. And in subsection E.4 we will prove that this event holds with high probability.

E.1.1. Optimism

We will show by induction that $Q_h^*(s,a) \leq Q_{h,k}(s,a)$ for all (s,a), h and k. For $h = H+1$, this inequality obviously holds, since both sides equal to 0. Next suppose that this inequality holds for some $h+1 \leq H$. As a result, we have

$$V_{h+1}^*(s) = \prod_{[0,H]}\left[\max_{a\in\mathcal{A}} Q_{h+1}^*(s,a)\right] \leq \prod_{[0,H]}\left[\max_{a\in\mathcal{A}} Q_{h+1,k}(s,a)\right] = V_{h+1,k}(s),$$

which indicates that

$$\begin{aligned} Q_h^*(s,a) &= r(s,a) + P(\cdot|s,a)^\top V_{h+1}^* \leq r(s,a) + P(\cdot|s,a)^\top V_{h+1,k} \\ &= r(s,a) + \sum_{j=1}^d (\theta_*)_j P_j(\cdot|s,a)^\top V_{h+1,k} \leq r(s,a) + \max_{\theta\in B_k}\left[\sum_{j=1}^d (\theta)_j P_j(\cdot|s,a)^\top V_{h+1,k}\right] \\ &= Q_{h,k}(s,a). \end{aligned}$$

This completes the induction.

E.1.2. Regret Decomposition

Let us denote π_k to be the stationary policy used in the k episode, and let

$$\bar{\theta}_{h,k}(s,a) = \arg\max_{\theta\in B_k} \sum_{j=1}^d (\theta)_j P_j(\cdot|s,a)^\top V_{h+1,k}.$$

Using the fact that $\pi_k(s_h^k) = a_h^k$ and $\theta_* \in B_k$ and letting ξ_{h+1}^k be

$$\xi_{h+1}^k := P(\cdot|s_h^k, a_h^k)^\top (V_{h+1,k} - V_{h+1}^*) - \left[V_{h+1,k}(s_{h+1}^k) - V_{h+1}^*(s_{h+1}^k)\right],$$

we have

$$\begin{aligned}
V_{h,k}(s_h^k) - V_h^{\pi_k}(s_h^k) &= Q_{h,k}(s_h^k, a_h^k) - Q_h^{\pi_k}(s_h^k, a_h^k) \\
&= r(s_h^k, a_h^k) + \bar{\theta}_{h,k}(s_h^k, a_h^k)^\top P_\cdot(\cdot|s_h^k, a_h^k) V_{h+1,k} - r(s_h^k, a_h^k) - \theta_*^\top P_\cdot(\cdot|s_h^k, a_h^k) V_{h+1}^{\pi_k} \\
&= \left[\theta_* + \bar{\theta}_{h,k}(s_h^k, a_h^k) - \theta_k + \theta_k - \theta_*\right]^\top P_\cdot(\cdot|s_h^k, a_h^k) V_{h+1,k} - \theta_*^\top P_\cdot(\cdot|s_h^k, a_h^k) V_{h+1}^{\pi_k} \\
&\le \theta_*^\top P_\cdot(\cdot|s_h^k, a_h^k)(V_{h+1,k} - V_{h+1}^{\pi_k}) + 2 \max_{\theta \in B_k} \left|(\theta - \theta_k)^\top P_\cdot(\cdot|s_h^k, a_h^k) V_{h+1,k}\right| \\
&\le P(\cdot|s_h^k, a_h^k)^\top (V_{h+1,k} - V_{h+1}^{\pi_k}) + 2 \max_{\theta \in B_k} \sqrt{(\theta - \theta_k)^\top M_k (\theta - \theta_k)} \cdot \\
&\qquad \sqrt{\left[P_\cdot(\cdot|s_h^k, a_h^k) V_{h+1,k}\right]^\top M_k^{-1} \left[P_\cdot(\cdot|s_h^k, a_h^k) V_{h+1,k}\right]} \\
&\le V_{h+1,k}(s_{h+1}^k) - V_{h+1}^{\pi_k}(s_{h+1}^k) + \xi_{h+1}^k \\
&\qquad + 2\sqrt{\beta_k} \cdot \sqrt{\left[P_\cdot(\cdot|s_h^k, a_h^k) V_{h+1,k}\right]^\top M_k^{-1} \left[P_\cdot(\cdot|s_h^k, a_h^k) V_{h+1,k}\right]},
\end{aligned}$$

where the first inequality uses the fact that $\theta_*, \bar{\theta}_{h,k} \in B_k$, the second inequality uses the Cauchy-Schwarz inequality and the third inequality uses the definition of B_k.

Recall that $V_{h+1,k}(s) = V_{H+1}^*(s) = 0$ for any $s \in \mathcal{S}$. We apply the preceding inequality recursively and obtain

$$\begin{aligned}
V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k) &\le V_{1,k}(s_1^k) - V_1^{\pi_k}(s_1^k) \ \text{(by optimism of value estimates)} \\
&\le \sum_{h=1}^H \xi_{h+1}^k + 2 \sum_{h=1}^H \sqrt{\beta_k} \cdot \sqrt{\left[P_\cdot(\cdot|s_h^k, a_h^k) V_{h+1,k}\right]^\top M_k^{-1} \left[P_\cdot(\cdot|s_h^k, a_h^k) V_{h+1,k}\right]},
\end{aligned}$$

therefore the expected regret can be bounded by if we bound the expectation of

$$\begin{aligned}
\hat{R}(K) &= \sum_{k=1}^K \left[V_1^*(s_1^k) - V_1^{\pi_k}(s_1^k)\right] \\
&\le \sum_{k=1}^K \sum_{h=1}^H \xi_{h+1}^k + 2 \sum_{k=1}^K \sum_{h=1}^H \sqrt{\beta_k} \cdot \sqrt{\left[P_\cdot(\cdot|s_h^k, a_h^k) V_{h+1,k}\right]^\top M_k^{-1} \left[P_\cdot(\cdot|s_h^k, a_h^k) V_{h+1,k}\right]}.
\end{aligned} \tag{25}$$

Moreover, it is easy to observe that

$$\mathbb{E}\left[\xi_{h+1}^k \middle| \mathbb{F}_{h,k}\right] = 0,$$

therefore ξ_{h+1}^k is a martingale difference sequence w.r.t. $\mathcal{F}_{h,k}$. Since

$$0 \le V_h^*(s_h^k), V_{h,k}(s_h^k) \le H \quad \text{and} \quad P(\cdot|s_h^k, a_h^k) \text{ is a probability distribution over the state space,}$$

we have $|\xi_h^k| \le H$ with probability 1. By the Azuma-Hoeffding inequality, with probability at least $1 - \delta$, the following inequality holds

$$\sum_{k=1}^K \sum_{h=1}^H \xi_{h+1}^k \le \sqrt{2H^3 K \log(1/\delta)}. \tag{26}$$

It remains to analyze the second term of (25), ie., the sum of bonus given by

$$2 \sum_{k=1}^K \sum_{h=1}^H \sqrt{\beta_k} \cdot \sqrt{\left[P_\cdot(\cdot|s_h^k, a_h^k) V_{h+1,k}\right]^\top M_k^{-1} \left[P_\cdot(\cdot|s_h^k, a_h^k) V_{h+1,k}\right]}.$$

E.2. Some Properties of Algorithm 3

In this subsection we establish several useful properties of our algorithm, assuming that optimism holds throughout.

E.2.1.

Note that

$$M_{h,k} = M_{1,1} + \sum_{(h',k')<(h,k)} \left[P_\cdot(\cdot|s_{h'}^{k'}, a_{h'}^{k'})V_{h'+1,k'}\right] \left[P_\cdot(\cdot|s_{h'}^{k'}, a_{h'}^{k'})V_{h'+1,k'}\right]^\top.$$

Denote

$$l_{h,k} = \sqrt{\left[P_\cdot(\cdot|s_h^k, a_h^k)V_{h+1,k}\right]^\top M_{h,k}^{-1} \left[P_\cdot(\cdot|s_h^k, a_h^k)V_{h+1,k}\right]}.$$

Denote by $(h,k)+1$ the double index of the next time step after (h,k), that is $(h+1,k)$ if $h < H$ and $(h,k+1)$ otherwise. We can see $\{M_{h,k}\}$ satisfies $M_{1,k} = M_{H+1,k-1}$ and also a recursive formula

$$\begin{aligned} M_{(h,k)+1}^{-1} &= \left(M_{h,k} + \left[P_\cdot(\cdot|s_h^k, a_h^k)V_{h+1,k}\right] \left[P_\cdot(\cdot|s_h^k, a_h^k)V_{h+1,k}\right]^\top\right)^{-1} \\ &= M_{h,k}^{-1} - \frac{M_{h,k}^{-1} \left[P_\cdot(\cdot|s_h^k, a_h^k)V_{h+1,k}\right] \left[P_\cdot(\cdot|s_h^k, a_h^k)V_{h+1,k}\right]^\top M_{h,k}^{-1}}{1 + \left[P_\cdot(\cdot|s_h^k, a_h^k)V_{h+1,k}\right]^\top M_{h,k}^{-1} \left[P_\cdot(\cdot|s_h^k, a_h^k)V_{h+1,k}\right]}. \end{aligned}$$

It implies that

$$\left[P_\cdot(\cdot|s_h^k, a_h^k)V_{h+1,k}\right]^\top M_{(h,k)+1}^{-1} \left[P_\cdot(\cdot|s_h^k, a_h^k)V_{h+1,k}\right] = l_{h,k}^2 - \frac{l_{h,k}^2 \cdot l_{h,k}^2}{1 + l_{h,k}^2} = \frac{l_{h,k}^2}{1 + l_{h,k}^2}.$$

E.2.2.

Next, we derive an upper bound to the quantity

$$\sum_{k=1}^{K} \sum_{h=1}^{H} \frac{l_{h,k}^2}{1 + l_{h,k}^2}.$$

Since

$$M_{(h,k)+1} = M_{h,k} + \left[P_\cdot(\cdot|s_h^k, a_h^k)V_{h+1,k}\right]^\top \left[P_\cdot(\cdot|s_h^k, a_h^k)V_{h+1,k}\right],$$

we have

$$\begin{aligned} \det M_{(h,k)+1} &= \det M_{h,k} \det\left(I + M_{h,k}^{-1/2} \left[P_\cdot(\cdot|s_h^k, a_h^k)V_{h+1,k}\right]^\top \left[P_\cdot(\cdot|s_h^k, a_h^k)V_{h+1,k}\right] M_{h,k}^{-1/2}\right) \\ &= \det M_{h,k} \left(1 + l_{h,k}^2\right). \end{aligned}$$

This indicates that

$$\sum_{(h',k')\le(h,k)} \log\left(1 + l_{h',k'}^2\right) = \log\det M_{(h,k)+1} - \log\det M_{1,1}.$$

Furthermore, since

$$\frac{l_{h,k}^2}{1 + l_{h,k}^2} \le \min\{1, l_{h,k}^2\} \le 2\log\left(1 + l_{h,k}^2\right),$$

we have

$$\begin{aligned} \sum_{(h',k')\le(h,k)} \frac{l_{h',k'}^2}{1 + l_{h',k'}^2} &\le \sum_{(h',k')\le(h,k)} \min\left\{1, l_{h',k'}^2\right\} \\ &\le \sum_{(h',k')\le(h,k)} 2\log(1 + l_{h',k'}^2) = 2\log\det M_{(h,k)+1} - 2\log\det M_{1,1}. \end{aligned}$$

E.2.3.

Given the initial value $M_{1,1} = H^2 dI$, we have

$$\begin{aligned}\mathbf{tr}(M_{(h,k)+1}) &= \mathbf{tr}(M_{1,1}) + \sum_{(h',k')\le(h,k)} \|P_{\cdot}(\cdot|s^{k'}_{h'}, a^{k'}_{h'})V_{h'+1,k'}\|^2 \\ &= H^2 d^2 + \sum_{(h',k')\le(h,k)} \sum_{j=1}^{d} \left(P_j(\cdot|s^{k'}_{h'}, a^{k'}_{h'})V_{h'+1,k'}\right)^2 \\ &\le H^2 d^2 + KdH^3,\end{aligned}$$

where the last inequality uses Assumption 3 and the fact that

$$P_j(\cdot|s^{k'}_{h'}, a^{k'}_{h'})V_{h'+1,k'} \le \|P_j(\cdot|s^{k'}_{h'}, a^{k'}_{h'})\|_1 \|V_{h'+1,k'}\|_\infty \le H.$$

Using the inequalities of arithmetic and geometric means, we get the following upper bound for the determinant of $M_{(h,k)+1}$:

$$\det M_{(h,k)+1} \le \left(\frac{\mathrm{tr}(M_{(h,k)+1})}{d}\right)^d \le (H^2 d + KH^3)^d,$$

which indicates that

$$\log\det M_{(H,k)+1} - \log\det M_{1,1} \le \log\left((H^2 d + KH^3)^d\right) - \log\left((H^2 d)^d\right) \le d\log(1+HK). \tag{27}$$

Hence we have

$$\sum_{(h',k')\le(h,k)} \frac{l^2_{h',k'}}{1+l^2_{h',k'}} \le \sum_{(h',k')\le(h,k)} \min\left\{1, l^2_{h',k'}\right\} \le 2d\log(1+HK). \tag{28}$$

E.3. Sum-of-Bonus Analysis

In this section, under the assumption that $\theta_* \in B_k$ for every k, we establish an upper bound for the following sum-of-bonus term

$$2\sum_{k=1}^{K}\sum_{h=1}^{H} \sqrt{\beta_k} \cdot \sqrt{\left[P_{\cdot}(\cdot|s^k_h, a^k_h)V_{h+1,k}\right]^\top M_k^{-1} \left[P_{\cdot}(\cdot|s^k_h, a^k_h)V_{h+1,k}\right]},$$

where we denote $M_k = M_{1,k}$ for simplicity. We let

$$u_{h,k} = \sqrt{\left[P_{\cdot}(\cdot|s^k_h, a^k_h)V_{h+1,k}\right]^\top M_k^{-1} \left[P_{\cdot}(\cdot|s^k_h, a^k_h)V_{h+1,k}\right]}.$$

Since $\beta_k \le \beta_K$ for any $1 \le k \le K$ and by letting

$$\begin{aligned}u^2_{h,k} &= \left[P_{\cdot}(\cdot|s^k_h, a^k_h)V_{h+1,k}\right]^\top M_k^{-1} \left[P_{\cdot}(\cdot|s^k_h, a^k_h)V_{h+1,k}\right] \\ &\le \left[P_{\cdot}(\cdot|s^k_h, a^k_h)V_{h+1,k}\right]^\top M_1^{-1} \left[P_{\cdot}(\cdot|s^k_h, a^k_h)V_{h+1,k}\right] \\ &= \frac{1}{H^2 d} \cdot \sum_{j=1}^{d} \left[P_j(\cdot|s^k_h, a^k_h)V_{h+1,k}\right]^2 \le \frac{1}{H^2 d} \cdot H^2 d = 1,\end{aligned}$$

we have

$$\begin{aligned}&2\sum_{k=1}^{K}\sum_{h=1}^{H} \sqrt{\beta_k} \cdot \sqrt{\left[P_{\cdot}(\cdot|s^k_h, a^k_h)V_{h+1,k}\right]^\top M_k^{-1} \left[P_{\cdot}(\cdot|s^k_h, a^k_h)V_{h+1,k}\right]} \\ &\le 2\sqrt{\beta_K} \cdot \sum_{k=1}^{K}\sum_{h=1}^{H} u_{h,k} \le 2\sqrt{\beta_K} \cdot \sum_{k=1}^{K}\sum_{h=1}^{H} \min\{1, u_{h,k}\} \\ &\le 2\sqrt{HK\beta_K} \cdot \sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H} \min\{1, u^2_{h,k}\}} \le 4\sqrt{HK\beta_K} \cdot \sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H} \log(1+u^2_{h,k})}\end{aligned} \tag{29}$$

where the third inequality uses the Cauchy-Schwarz inequality. Next we notice that

$$M_{k+1} = M_k + \sum_{h=1}^{H} \left[P_\bullet(\cdot|s_h^k, a_h^k)V_{h+1,k}\right]^\top \left[P_\bullet(\cdot|s_h^k, a_h^k)V_{h+1,k}\right].$$

Hence we have

$$\det(M_{k+1}) = \det(M_k) \cdot \det\left(I + \sum_{h=1}^{H} M_k^{-1/2} \left[P_\bullet(\cdot|s_h^k, a_h^k)V_{h+1,k}\right]^\top \left[P_\bullet(\cdot|s_h^k, a_h^k)V_{h+1,k}\right] M_k^{-1/2}\right).$$

We further notice that every eigenvalue of the matrix

$$I + \sum_{h=1}^{H} M_k^{-1/2} \left[P_\bullet(\cdot|s_h^k, a_h^k)V_{h+1,k}\right]^\top \left[P_\bullet(\cdot|s_h^k, a_h^k)V_{h+1,k}\right] M_k^{-1/2}$$

is at least 1, and we have the following bound of its trace:

$$\begin{aligned}
&\text{tr}\left(\sum_{h=1}^{H} M_k^{-1/2} \left[P_\bullet(\cdot|s_h^k, a_h^k)V_{h+1,k}\right]^\top \left[P_\bullet(\cdot|s_h^k, a_h^k)V_{h+1,k}\right] M_k^{-1/2}\right) \\
&= \sum_{h=1}^{H} \left[P_\bullet(\cdot|s_h^k, a_h^k)V_{h+1,k}\right]^\top M_k^{-1} \left[P_\bullet(\cdot|s_h^k, a_h^k)V_{h+1,k}\right] = \sum_{h=1}^{H} u_{h,k}^2.
\end{aligned}$$

This indicates that

$$\begin{aligned}
&\det\left(I + \sum_{h=1}^{H} M_k^{-1/2} \left[P_\bullet(\cdot|s_h^k, a_h^k)V_{h+1,k}\right]^\top \left[P_\bullet(\cdot|s_h^k, a_h^k)V_{h+1,k}\right] M_k^{-1/2}\right) \\
&\geq 1 + \text{tr}\left(I + \sum_{h=1}^{H} M_k^{-1/2} \left[P_\bullet(\cdot|s_h^k, a_h^k)V_{h+1,k}\right]^\top \left[P_\bullet(\cdot|s_h^k, a_h^k)V_{h+1,k}\right] M_k^{-1/2}\right) \\
&= 1 + \sum_{h=1}^{H} u_{h,k}^2,
\end{aligned}$$

where the first inequality follows from the following fact: $\prod_i (1 + w_i) \geq 1 + \sum_i w_i$ provided $w_i \geq 0$. Combining the above inequality with the following inequality

$$1 + \sum_{h=1}^{H} u_{h,k}^2 = \frac{\sum_{h=1}^{H}(1 + Hu_{h,k}^2)}{H} \geq \prod_{h=1}^{H}(1 + Hu_{h,k}^2)^{1/H} \geq \prod_{h=1}^{H}(1 + u_{h,k}^2)^{1/H},$$

we obtain that

$$\sum_{h=1}^{H} \log(1 + u_{h,k}^2) \leq H \log\left(1 + \sum_{h=1}^{H} u_{h,k}^2\right) \leq H \det(M_{k+1}) - H \det(M_k).$$

Therefore, we have

$$\begin{aligned}
&2\sum_{k=1}^{K}\sum_{h=1}^{H} \sqrt{\beta_k} \cdot \sqrt{\left[P_\bullet(\cdot|s_h^k, a_h^k)V_{h+1,k}\right]^\top M_k^{-1} \left[P_\bullet(\cdot|s_h^k, a_h^k)V_{h+1,k}\right]} \\
&\leq 4\sqrt{HK\beta_K} \cdot \sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H} \log(1 + u_{h,h}^2)} \\
&\leq 4\sqrt{HK\beta_K} \cdot \sqrt{\sum_{k=1}^{K} H \det(M_{k+1}) - H \det(M_k)} \\
&\leq 4\sqrt{HK\beta_K} \cdot \sqrt{H \det(M_{(H,k)+1}) - H \det(M_{1,1})} \\
&\leq 4\sqrt{H^2 dK\beta_K \log(1 + HK)},
\end{aligned}$$

where the last inequality uses (27).

E.4. Confidence Set of Value Target Regression

We adopt the result in (Abbasi-Yadkori et al., 2011). For $t = H(k-1)+h$, we choose

$$\begin{aligned}\lambda &= H^2 d,\\ \overline{V}_t &= M_{h,k},\\ S &= C_\theta,\\ R &= H,\\ L &= \sqrt{H^2 d}.\end{aligned}$$

Then we have

$$\theta_{h,k} = (\mathbf{X}_{1:t}^T \mathbf{X}_{1:t} + \lambda I)^{-1} \mathbf{X}_{1:t} \mathbf{Y}_{1:t} = \hat{\theta}_t, \quad \text{and} \quad \|\theta_*\|_2 \le C_\theta = S.$$

Moreover, if we since $|\eta_t| = |Y_t - \langle X_t, \theta_* \rangle| = |Y_{h,k} - P(\cdot|s_h^k, a_h^k)^T V_{h+1,k}| \le H$, η_t is H-subgaussian. We can also verify that

$$\|X_t\|_2^2 = \sum_{i=1}^{d} \left(P_i(\cdot|s_h^k, a_h^k) V_{h+1,k}\right)^2 \le H^2 d = L^2.$$

Hence according to Theorem 2 in (Abbasi-Yadkori et al., 2011), we obtain that with probability at least $1-\delta$, for any $(h,k) \le (H,K)$, the following inequality holds:

$$\|\theta_* - \theta_{h,k}\|_{M_{h,k}} \le H\sqrt{d \log\left(\frac{1 + Hk \cdot H^2 d}{\delta}\right)} + C_\theta H \sqrt{d}$$

Therefore, if we choose

$$\beta_k = \left(H\sqrt{d \log\left(\frac{1 + Hk \cdot H^2 d}{\delta}\right)} + C_\theta H \sqrt{d}\right)^2,$$

then we will have

$$\theta_* \in B_{h,k}$$

for all $(h,k) \le (H,K)$ with probability at least $1-\delta$.

E.5. Expected Regret Analysis

According to Section E.4, we have with probability at least $1-\delta$ that $\theta_* \in B_k$ for all $1 \le k \le K$. When this event happens, we enable the analysis of Sections E.1-E.3. We combine the error bounds (26) and (29) and apply them into the regret bound (25). It follows that, if $T = KN$,

$$\begin{aligned}R(T) &\le 2\sqrt{H^3 K}\log(1/\delta) + 4\sqrt{H^2 dK\beta_K \log(1+HK)}\\ &= 2\sqrt{H^3 K}\log\left(\frac{1}{\delta}\right) + 4H^2 d\sqrt{K\log(1+HK)} \cdot \left(\sqrt{\log\left(\frac{1+H^3Kd}{\delta}\right)} + C_\theta\right)\\ &\le 6H^2 d\sqrt{K}\left(C_\theta\sqrt{\log(1+HK)} + \log\left(\frac{1+H^3Kd}{\delta}\right)\right)\end{aligned}$$

with probability at least $1-2\delta$. Note the trivial upper bound $R(K) \le HK$. Therefore, by letting $\delta = 1/K$ and noticing $T = HK$, we get

$$\begin{aligned}\mathbb{E}\left[R(T)\right] &\le (1-2\delta)\cdot 6H^2 d\sqrt{K}\left(C_\theta\sqrt{\log(1+HK)} + \log\left(\frac{1+H^3Kd}{\delta}\right)\right) + 2\delta \cdot HK\\ &\le 6H^2 d\sqrt{K}\cdot\left(C_\theta\sqrt{\log(1+HK)} + \log\left(\frac{1+H^3Kd}{\delta}\right)\right)\\ &= \tilde{\mathcal{O}}(C_\theta \cdot H^2 d\sqrt{K}) = \tilde{\mathcal{O}}(C_\theta \cdot d\sqrt{H^3 T}).\end{aligned}$$

Thus we have completed the proof of Theorem 13.

F. Implementation

F.1. Analysis of Implemented Confidence Bounds

In the implementation of UCRL-VTR used in Section 6, we used different confidence intervals then the ones stated in the paper. The confidence intervals used in our implementation are the ones introduced in (Abbasi-Yadkori et al., 2011). These confidence intervals are much tighter in the linear setting than the ones introduced in Section 3 and thus have better practical performance. The purpose of this section is to formally introduce the confidence intervals used in our implementation of UCRL-VTR as well as show how these confidence intervals were adapted from the linear bandit setting to the linear MDP setting.

F.1.1. Linear MDP Assumptions

For our implementation of UCRL-VTR we used different confidence then was introduced in the paper. These are the tighter confidence bounds from the seminal work done by (Abbasi-Yadkori et al., 2011) and further expanded upon in Chapter 20 of (Lattimore & Szepesvári, 2020). Now we will state some assumptions in the MDP setting, then we will state the equivalent assumptions from the linear bandit setting, and lastly we will make the connections between the two that allow us to use the confidence bounds from the linear bandit setting in the RL setting.

1. $P^*(s' \mid s,a) = \sum_{i=1}^{d} (\theta_*^{MDP})_i P_i(s' \mid s,a)$
2. $s_{h+1}^k \sim P^*(\cdot \mid s_h^k, a_h^k)$
3. $\mathcal{C}_t^{MDP} = \{\theta^{MDP} \in \mathbb{R}^d : \|\theta^{MDP} - \hat{\theta}_t^{MDP}\|_{M_k} \leq \beta_t\}$

where t is defined in the table of A.4. Also note that in this section $(\cdot)_*$ denotes the true parameter or model, $(\cdot)^{MDP}$ denotes something derived or used in the linear MDP setting, and $(\cdot)^{LIN}$ denotes something derived or used in the linear bandit setting. Now, under 1-3 of F.1.1 we hope to construct a confidence set $\mathcal{C}_t^{MDP}$ such that

$$\theta^{MDP} \in \bigcap_{t=1}^{\infty} \mathcal{C}_t^{MDP}$$

with high probability. Now the choice of how to choose both $\mathcal{C}_t^{MDP}$ and β_t comes from the linear bandit literature. We will introduce the necessary theorems and assumptions to derive both $\mathcal{C}_t^{LIN}$ and β_t in the linear bandit setting and then adapt the results from the linear bandit setting to the linear MDP setting.

F.1.2. Tighter Confidence Bounds for Linear Bandits

The following results are introduced in the paper by (Abbasi-Yadkori et al., 2011) and are further explained in Chapter 20 of the book by (Lattimore & Szepesvári, 2020). In this section, we will introduce the theorems and lemmas that allows us to derive tighter confidence intervals for the linear bandit setting. Then we will carefully adapt the confidence intervals to the linear bandit setting. Now supposed a bandit algorithm has chosen actions $A_1, ..., A_t \in \mathbb{R}^d$ and received rewards $X_1^{LIN}, ..., X_t^{LIN}$ with $X_s^{LIN} = \langle A_t, \theta_*^{LIN} \rangle + \eta_s$ where η_s is some zero mean noise. The least squares estimator of θ_*^{LIN} is the minimizer of the following loss function

$$L_t(\theta^{LIN}) = \sum_{s=1}^{t} (X_s^{LIN} - \langle A_t, \theta^{LIN} \rangle)^2 + \lambda \|\theta^{LIN}\|_2^2$$

where $\lambda > 0$ is the regularizer. This loss function is minimized by

$$\hat{\theta}_t^{LIN} = W_t^{-1} \sum_{s=1}^{t} X_s^{LIN} A_s \text{ with } W_t = \lambda I + \sum_{s=1}^{t} A_s A_s^\top$$

notice how this linear bandit problem is very similar to the linear MDP problem introduced in section 3 of our paper. In our linear MDP setting, it is convenient to think of M and W as serving equivalent purposes (storing rank one updates) thus it is also convenient to think of A_t and X_t^{MDP} as serving equivalent purposes (the features by which we use to make our predictions), where X_t^{MDP} is defined in section 3 of our paper with some added notation to distinguish it from the X_t^{LIN} used here in the linear bandit setting. We will now build up some intuition by making some simplifying assumptions.

1. No regularization: $\lambda = 0$ and W_t is invertible.
2. Independent subgaussian noise: $(\eta_s)_s$ are independent and σ-subgaussian
3. Fixed Design: $A_1, ..., A_t$ are deterministically chosen without the knowledge of $X_1^{LIN}, ..., X_t^{LIN}$

finally it is also convenient to think of X_t^{LIN} and $V_{t+1}(s_{t+1})$ as serving equivalent purposes (the target of our predictions). Thus the statements we prove in the linear bandit setting can be easily adapted to the linear MDP setting. While none of the assumptions stated above is plausible in the bandit setting, the simplifications eases the analysis and provides insight.

Comparing θ_*^{LIN} and $\hat{\theta}_t^{LIN}$ in the direction $x \in \mathbb{R}^d$, we have

$$\begin{aligned}\langle \hat{\theta}_t^{LIN} - \theta_*^{LIN}, x\rangle &= \left\langle x, W_t^{-1} \sum_{s=1}^{t} A_s X_s^{LIN} - \theta_*^{LIN} \right\rangle = \left\langle x, W_t^{-1} \sum_{s=1}^{t} A_s (A_s^\top \theta_*^{LIN} + \eta_s) - \theta_*^{LIN} \right\rangle \\ &= \left\langle x, W_t^{-1} \sum_{s=1}^{t} A_s \eta_s \right\rangle = \sum_{s=1}^{t} \langle x, W_t^{-1} A_s\rangle \eta_s\end{aligned}$$

Since $(\eta_s)_s$ are independent and σ-subgaussian, by Lemma 5.4 and Theorem 5.3 (need to be stated),

$$\mathbb{P}\left(\langle \hat{\theta}_t^{LIN} - \theta_*^{LIN}, x\rangle \geq \sqrt{2\sigma^2 \sum_{s=1}^{t} \langle x, W_t^{-1} A_s\rangle^2 \log\left(\frac{1}{\delta}\right)} \right) \leq \delta$$

A little linear algebra shows that $\sum_{s=1}^{t} \langle x, W_t^{-1} A_s\rangle^2 = \|x\|_{W_t^{-1}}^2$ and so,

$$\mathbb{P}\left(\langle \hat{\theta}_t^{LIN} - \theta_*^{LIN}, x\rangle \geq \sqrt{2\sigma^2 \|x\|_{W_t^{-1}}^2 \log\left(\frac{1}{\delta}\right)} \right) \leq \delta \tag{30}$$

We now remove the limiting assumptions we stated above and use the newly stated assumptions for the rest of this section

1. There exists a $\theta_*^{LIN} \in \mathbb{R}^d$ such that $X_t^{LIN} = \langle \theta_*^{LIN}, A_t\rangle + \eta_t$ for all $t \geq 1$.
2. The noise is conditionally σ-subgaussian:

   $$\text{for all } \alpha \in \mathbb{R} \text{ and } t \geq 1, \mathbb{E}[\exp(\alpha \eta_t) \mid \mathcal{F}_{t-1}] \leq \exp\left(\frac{\alpha\sigma^2}{2}\right) a.s.$$

   where $\mathcal{F}_{t-1}$ is such that $A_1, X_1^{LIN}, ..., A_{t-1}, X_{t-1}^{LIN}$ are $\mathcal{F}_{t-1}$-measurable.
3. In addition, we now assume $\lambda > 0$.

The inclusion of A_t in the definition of $\mathcal{F}_{t-1}$ allows the noise to depend on past choices, including the most recent action. Since we want exponentially decaying tail probabilities, one is tempted to try the Cramer-Chernoff method:

$$\mathbb{P}(\|\hat{\theta}_t^{LIN} - \theta_*^{LIN}\|_{W_t}^2 \geq u^2) \leq \inf_{\alpha > 0} \mathbb{E}\left[\exp\left(\alpha \|\hat{\theta}_t^{LIN} - \theta_*^{LIN}\|_{W_t}^2 - \alpha u^2\right)\right].$$

Sadly, we do not know how to bound this expectation. Can we still somehow use the Cramer–Chernoff method? We take inspiration from looking at the special case of $\lambda = 0$ one last time, assuming that $W_t = \sum_{s=1}^{t} A_s A_s^\top$ is invertible. Let

$$S_t = \sum_{s=1}^{t} \eta_s A_s$$

Recall that $\hat{\theta}_t^{LIN} = W_t^{-1} \sum_{s=1}^{t} X_s^{LIN} A_s = \theta_*^{LIN} + W_t^{-1} S_t$. Hence,

$$\frac{1}{2}\|\hat{\theta}_t^{LIN} - \theta_*^{LIN}\|_{W_t}^2 = \frac{1}{2}\|S_t\|_{W_t^{-1}}^2 = \max_{x \in \mathbb{R}^d} \left(\langle x, S_t\rangle - \frac{1}{2}\|x\|_{W_t}^2 \right).$$

The next lemma shows that the exponential of the term inside the maximum is a supermartingale even when $\lambda \geq 0$.

Lemma 17. *For all $x \in \mathbb{R}^d$ the process $D_t(x) = \exp(\langle x, S_t\rangle - \frac{1}{2}\|x\|_{W_t^2})$ is an $\mathbb{F}$-adapted non-negative supermartingale with $D_0(x) \leq 1$.*

The proof for this Lemma can be found in Chapter 20 of the book by (Lattimore & Szepesvári, 2020). For simplicity, consider now again the case when $\lambda = 0$. Combining the lemma and the linearisation idea almost works. The Cramer–Chernoff method leads to

$$\mathbb{P}\left(\frac{1}{2}\|\hat{\theta}_t^{LIN} - \theta_*^{LIN}\|_{W_t}^2 \geq \log(1/\delta)\right) = \mathbb{P}\left(\exp\left(\max_{x\in\mathbb{R}^d}\left(\langle x, S_t\rangle - \frac{1}{2}\|x\|_{W_t}^2\right)\right) \geq \log(1/\delta)\right) \tag{31}$$

$$\leq \delta\mathbb{E}\left[\exp\left(\max_{x\in\mathbb{R}^d}\left(\langle x, S_t\rangle - \frac{1}{2}\|x\|_{W_t}^2\right)\right)\right] = \delta\mathbb{E}\left[\max_{x\in\mathbb{R}^d} D_t(x)\right] \tag{32}$$

Now Lemma 17 shows that $\mathbb{E}[D_t(x)] \leq 1$. Now using Laplace's approximation we write

$$\max_x D_t(x) \approx \int_{\mathbb{R}^d} D_t(x)dh(x),$$

where h is some measure on $\mathbb{R}^d$ chosen so that the integral can be calculated in closed form. This is not a requirement of the method, but it does make the argument shorter. The main benefit of replacing the maximum with an integral is that we obtain the following lemma

Lemma 18. *Let h be a probability measure on $\mathbb{R}^d$; then; $\bar{D}_t = \int_{\mathbb{R}^d} D_t(x)dh(x)$ is an $\mathbb{F}$-adapted non-negative supermartingale with $\bar{D}_0 = 1$.*

The proof of Lemma 18 can, again, be found in Chapter 20 of the book by (Lattimore & Szepesvári, 2020). Now the following theorem is the key result from which the confidence set will be derived.

Theorem 19. *For all $\lambda > 0$, and $\delta \in (0, 1)$*

$$\mathbb{P}\left(exists\ t \in \mathbb{N} : \|S_t\|_{W_t^{-1}}^2 \geq 2\sigma^2 \log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det W_t}{\lambda^d}\right)\right) \leq \delta$$

Furthermore, if $\|\theta_*^{LIN}\|_2 \leq m_2$, then $\mathbb{P}(exists\ t \in \mathbb{N}^+ : \theta_*^{LIN} \notin \mathcal{C}_t^{LIN}) \leq \delta$ with

$$\mathcal{C}_t^{LIN} = \left\{\theta \in \mathbb{R}^d : \|\hat{\theta}_{t-1}^{LIN} - \theta\|_{W_{t-1}} < m_2\sqrt{\lambda} + \sqrt{2\sigma^2 \log\left(\frac{1}{\delta}\right) + \log\left(\frac{W_{t-1}}{\lambda^d}\right)}\right\}.$$

The proof of Theorem 19 can be found in Chapter 20 of the book by (Lattimore & Szepesvári, 2020).

F.1.3. Adaptation of the Confidence Bounds to our Linear MDP Setting

Now with the Lemmas and Theorems introduced in the previous section we are ready to derive the confidence bounds used in our implementation of UCRL-VTR. Now using the notation from the linear bandit setting we set

1. The target $X_t^{MDP} = \int_j V_t(s')P_j(ds' \mid s_t, a_t)$
2. $Y_t = V_t(s_{t+1})$
3. $\mathcal{F}_{t-1} = \sigma(s_1, a_1, ..., s_{t-1}, a_{t-1})$, which just means the filtration is set to be the sigma-algebra generated by all past states and actions observed.
4. $\eta_t = Y_t - \langle X_t^{MDP}, \theta_*^{MDP}\rangle = V_t(s_{t+1}) - \int_j V_t(s')P_j^*(ds' \mid s_t, a_t)$, since θ_*^{MDP} is the true model of the MDP.
5. M_t in the linear MDP setting is defined equivalently to W_t in the linear bandit setting, i.e. they are both the sums of a regularizer term and a bunch of rank one updates.

it can be seen that our the noise in our system η_t has zero mean $\mathbb{E}[\eta_t \mid \mathcal{F}_{t-1}] = 0$ finally the noise in our system has variance $H/2$ thus our system in $H/2$-subgaussian.

Lemma 20. *(Hoeffding's lemma) Let $Z = Z - \mathbb{E}[Z]$ be a real centered random variable such that $Z \in [a, b]$ almost surely. Then $\mathbb{E}[\exp(\alpha Z)] \leq \exp(\alpha^2 \frac{(b-a)^2}{8})$ for any $\alpha \in \mathbb{R}$ or Z is subgaussian with variance $\sigma^2 = \frac{(b-a)^2}{4}$.*

Proof Define $\psi(\alpha) = \log \mathbb{E}[\exp(\alpha Z)]$ we can then compute

$$\psi'(\alpha) = \frac{\mathbb{E}[Z \exp(\alpha Z)]}{\mathbb{E}[\exp(\alpha Z)]}, \;\; \psi''(\alpha) = \frac{\mathbb{E}[Z^2 \exp(\alpha Z)]}{\mathbb{E}[\exp(\alpha Z)]} - \left(\frac{\mathbb{E}[Z \exp(\alpha Z)]}{\mathbb{E}[\exp(\alpha Z)]}\right)^2$$

Thus $\psi''(\alpha)$ can be interpreted as the variance of the random variable Z under the probability measure $d\mathbb{Q} = \frac{\exp(\alpha Z)}{\mathbb{E}[\exp(\alpha Z)]} d\mathbb{P}$, but since $Z \in [a, b]$ almost surely, we have, under any probability

$$\text{var}(Z) = \text{var}(Z - \frac{a+b}{2}) \leq \mathbb{E}\left[\left(Z - \frac{a+b}{2}\right)^2\right] \leq \left(\frac{b-a}{4}\right)^2$$

The fundamental theorem of calculus yields

$$\psi(\alpha) = \int_0^s \int_0^\mu \psi(\rho) d\rho d\mu = \frac{s^2(b-a)^2}{8}$$

using $\psi(0) = \log 1 = 0$ and $\psi'(0) = \mathbb{E}[Z] = 0$. □

Now using Lemma 20 and the fact that Y_t is bounded in the range of $[0, H]$, $\mathbb{E}[Y_t] = \langle X_t^{MDP}, \theta_*^{MDP} \rangle$, and $\eta_t = Y_t - \langle X_t^{MDP}, \theta_*^{MDP} \rangle = Y_t - \mathbb{E}[Y_t]$, the noise η_t in our linear MDP setting is $H/2$-subgaussian. This result is also stated in a proof from A.4.

Putting this all together we can derive the tighter confidence set for UCRL-VTR in the linear setting,

$$\mathcal{C}_t^{MDP} = \left\{\theta \in \mathbb{R}^d : \|\hat{\theta}_{t-1}^{MDP} - \theta\|_{M_{t-1}} < m_2\sqrt{\lambda} + \frac{H}{2}\sqrt{2\log\left(\frac{1}{\delta}\right) + \log\left(\frac{M_{t-1}}{\lambda^d}\right)}\right\}.$$

where here in the linear MDP setting M_t replaces W_t from the linear bandit setting and $\|\theta_*^{MDP}\|_2 \leq m_2$. The justification of using these bounds in the linear MDP setting follow exactly from the justification given above for using these bounds in the linear bandit setting.

F.2. UCRL-VTR

In the proceeding subsections we discuss the implementation of the algorithms studied in Section 6 of the paper. The first algorithm we present is the algorithm used to generate the results for UCRL-VTR.

Algorithm 4 UCRL-VTR with Tighter Confidence Bounds

Input: MDP, $d, H, T = KH$;
Initialize: $M_{1,1} \leftarrow I, \quad w_{1,1} \leftarrow 0 \in \mathbb{R}^{d\times 1}, \quad \theta_1 \leftarrow M_{1,1}^{-1} w_{1,1} \qquad$ for $1 \le h \le H$, $d_1 = |\mathcal{S}| \times |\mathcal{A}|$;
Initialize: $\delta \leftarrow 1/K$, and for $1 \le k \le K$,
Compute Q-function $Q_{h,1}$ using $\theta_{1,1}$ according to (3);
for $k = 1 : K$ **do**
  Obtain initial state s_1^k for episode k;
  for $h = 1 : H$ **do**
    Choose action greedily by

$$a_h^k = \arg\max_{a\in\mathcal{A}} Q_{h,k}(s_h^k, a)$$

    and observe the next state s_{h+1}^k.
    Compute the predicted value vector: ▷ Evaluate the expected value of next state

$$X_{h,k} \leftarrow \mathbb{E}_{\cdot}[V_{h+1,k}(s)|s_h^k, a_h^k] \quad = \sum_{s\in\mathcal{S}} V_{h+1,k}(s) \cdot P_{\cdot}(s|s_h^k, a_h^k).$$

    $y_{h,k} \leftarrow V_{h+1,k}(s_{h+1}^k)$ ▷ Update regression parameters
    $M_{h+1,k} \leftarrow M_{h,k} + X_{h,k} X_{h,k}^\top$
    $w_{h+1,k} \leftarrow w_{h,k} + y_{h,k} \cdot X_{h,k}$
  end for
  Update at the end of episode: ▷ Update Model Parameters

$$M_{1,k+1} \leftarrow M_{H+1,k},$$
$$w_{1,k+1} \leftarrow w_{H+1,k},$$
$$\theta_{k+1} \leftarrow M_{1,k+1}^{-1} w_{1,k+1};$$

  Compute $Q_{h,k+1}$ for $h = H, \ldots, 1$, using θ_{k+1} according to (33) using

$$\sqrt{\beta_{h,k}} \leftarrow \sqrt{d_1} + \frac{H-h+1}{2}\sqrt{2\log\left(\frac{1}{\delta}\right) + \log\det(M_{1,k+1})};$$

  ▷ Computing Q functions
end for

The iterative Q-update for Algorithm 4 is

$$\begin{aligned} V_{h+1,k}(s) &= 0 \\ Q_{h,k}(s,a) &= r(s,a) + X_{h,k}^\top \theta_k + \sqrt{\beta_{h,k}}\sqrt{X_{h,k}^\top M_{1,k+1}^{-1} X_{h,k}} \\ V_{h,k}(s) &= \max_a Q_{h,k}(s,a) \end{aligned} \tag{33}$$

The choice of the confidence bounds used in Algorithm 4 comes from the tight bounds derived in (Abbasi-Yadkori et al., 2011) for linear bandits and further expanded upon in Chapter 20 of (Lattimore & Szepesvári, 2020). The details of which are shown and stated in F.1. We slightly tighten the values for the noise at each stage by using the fact that for each stage in the horizon, $h \in [H]$, the value $V_h^k(\cdot)$ is capped as to never be greater than $H - h + 1$. The appearance of the $\sqrt{d_1}$ comes from the fact that $\|\theta_*\|_2 \le \sqrt{d_1}$ for all $\theta_* \in \mathbb{R}^d$ in the tabular setting since θ_* in the tabular setting is equal to the true model of the environment.

F.3. EGRL-VTR

In this section we discuss the algorithm EGRL-VTR. This algorithm is very similar to UCRL-VTR expect it performs ε-greedy value iteration instead of optimistic value iteration and acts ε-greedy with respect to $Q_{h,k}$.

Algorithm 5 EGRL-VTR

Input: MDP, $d, H, T = KH, \varepsilon > 0$;
Initialize: $M_{1,1} \leftarrow I, \quad w_{1,1} \leftarrow 0 \in \mathbb{R}^{d\times 1}, \quad \theta_1 \leftarrow M_{1,1}^{-1} w_{1,1} \qquad$ for $1 \le h \le H$;
Compute Q-function $Q_{h,1}$ using $\theta_{1,1}$ according to (34);
for $k = 1 : K$ **do**
  Obtain initial state s_1^k for episode k;
  for $h = 1 : H$ **do**
    With probability $1 - \varepsilon$ do

$$a_h^k = \arg\max_{a\in\mathcal{A}} Q_{h,k}(s_h^k, a)$$

    else pick a uniform random action $a_h^k \in \mathcal{A}$. Observe the next state s_{h+1}^k.
    Compute the predicted value vector: ▷ Evaluate the expected value of next state

$$X_{h,k} \leftarrow \mathbb{E}_{\cdot}[V_{h+1,k}(s)|s_h^k, a_h^k] \quad = \sum_{s\in\mathcal{S}} V_{h+1,k}(s) \cdot P_{\cdot}(s|s_h^k, a_h^k).$$

    $y_{h,k} \leftarrow V_{h+1,k}(s_{h+1}^k)$ ▷ Update regression parameters
    $M_{h+1,k} \leftarrow M_{h,k} + X_{h,k} X_{h,k}^\top$
    $w_{h+1,k} \leftarrow w_{h,k} + y_{h,k} \cdot X_{h,k}$
  end for
  Update at the end of episode: ▷ Update Model Parameters

$$M_{1,k+1} \leftarrow M_{H+1,k},$$
$$w_{1,k+1} \leftarrow w_{H+1,k},$$
$$\theta_{k+1} \leftarrow M_{1,k+1}^{-1} w_{1,k+1};$$

  Compute $Q_{h,k+1}$ for $h = H, \ldots, 1$, using θ_{k+1} according to (34) ▷ Computing Q functions
end for

The iterative value update for EGRL-VTR is

$$\begin{aligned} V_{h+1,k}(s) &= 0 \\ Q_{h,k}(s,a) &= r(s,a) + X_{h,k}^\top \theta_k \\ V_{h,k}(s) &= (1-\varepsilon)\Pi_{[0,H]} \max_a Q_{h,k}(s,a) + \frac{\varepsilon}{|\mathcal{A}|} \sum_{a\in\mathcal{A}} Q_{h,k}(s,a) \end{aligned} \tag{34}$$

F.4. EG-Frequency

In this section we discuss the algorithm EG-Frequency. This algorithm is the ε-greedy version of UC-MatrixRL (Yang & Wang, 2019a).

Algorithm 6 EG-Frequency

Input: MDP, Features $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ and $\psi : \mathcal{S} \to \mathbb{R}^{|\mathcal{S}|}$, $\varepsilon > 0$, and the total number of episodes K;
Initialize: $A_1 \leftarrow I \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|}$, $M_1 \leftarrow 0 \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}|}$, and $K_\psi \leftarrow \sum_{s' \in \mathcal{S}} \psi(s')\psi(s')^\top$;
for $k = 1 : K$ **do**
  Let $Q_{h,k}$ be given in (35) using M_k;
  for $h = 1 : H$ **do**
    Let the current state be s_h^k;
    With probability $(1-\varepsilon)$ play action $a_h^k = \arg\max_{a \in \mathcal{A}} Q_{h,k}(s_h^k, a)$ else pick a uniform random action $a_h^k \in \mathcal{A}$.
    Record the next state s_{h+1}^k
  end for
  $A_{k+1} \leftarrow A_k + \sum_{h \le H} \phi(s_h^k, a_h^k)\phi(s_h^k, a_h^k)^\top$
  $M_{k+1} \leftarrow M_k + A_{k+1}^{-1} \sum_{h \le H} \phi(s_h^k, a_h^k)\psi(s_{h+1}^k)^\top K_\psi^{-1}$
end for

The iterative Q-update for EG-Frequency is

$$\begin{aligned} Q_{h+1,k}(s,a) &= 0 \text{ and} \\ Q_{h,k}(s,a) &= r(s,a) + \phi(s,a)^\top M_k \mathbf{\Psi}^\top V_{h+1,k} \\ V_{h,k} &= (1-\varepsilon)\Pi_{[0,H]} \max_a Q_{h,k}(s,a) + \frac{\varepsilon}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} Q_{h,k}(s,a) \end{aligned} \tag{35}$$

Note that $\mathbf{\Psi}$ is a $|\mathcal{S}| \times |\mathcal{S}|$ whose rows are the features $\psi(s')$ and $\mathbf{\Phi}$ is a $|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|$ whose rows are the features $\phi(s,a)$. In the tabular RL setting both $\mathbf{\Psi}$ and $\mathbf{\Phi}$ are the identity matrix which is what we used in our numerical experiments. In the tabular RL setting, EG-Frequency stores the counts of the number of times it transitioned to next state s' from the state-action pair (s,a) and fits the estimated model M_k accordingly.

F.5. Futher Implementation Notes

In this section, we include some further details on how we implemented Algorithms 4, 5, and 6. All code was written in Python 3 and used the Numpy and Scipy libraries. All plots were generated using MatPlotLib. In Algorithm 4, Numpy's logdet function was used to calculate the determinate in step 15 for numerical stability purposes. No matrix inversion was performed in our code, instead a Sherman-Morrison update was performed for each matrix in which a matrix inversion is performed at each (k,h) in order to save on computation. To read more about the Sherman Morrison update in the context of RL, we refer to the reader to Eqn (9.22) of (Sutton & Barto, 2018). When computing the weighted L1-norm, we added a small constant to each summation in the denominator to avoid dividing by zero. Finally, when computing UC-MatrixRL we also used the self-normalize bounds introduced in the beginning of this section. Some pseudocode for using self-normalized bounds with UC-MatrixRL can be found in step 5 of Alg 7.

G. Mixture Model

In this section, we introduce, analyze, and evaluate a Linear model-based RL algorithm that used both the canonical model and the VTR model for planning. We call this algorithm UCRL-MIX.

G.1. UCRL-MIX

Below a meta-algorithm for UCRL-MIX

Algorithm 7 UCRL-MIX

Compute Algorithm 4 and UC-MatrixRL (Yang & Wang, 2019a) simultaneously.
At end of episode k, perform value iteration and set $V_{H+1,k}(s) = 0$.
for $h = H : 1$ **do**
  for $s \in |\mathcal{S}|$ and $a \in |\mathcal{A}|$ **do**
    Compute the confidence set bonuses as follows

$$B_{h,k}^{VTR} \leftarrow \sqrt{d_1} + \frac{H-h+1}{2}\sqrt{2\log\left(\frac{2}{\delta}\right) + \log\det(M_{1,k+1})};$$

$$B_{h,k}^{MAT} \leftarrow \sqrt{|\mathcal{S}||\mathcal{A}|} + \frac{H-h+1}{2}\sqrt{2\log\left(\frac{2}{\delta}\right) + \log\det(A_{k+1})};$$

    if $B_{h,k}^{VTR}\sqrt{X_{h,k}^\top M_{1,k+1}^{-1} X_{h,k}} \leq B_{h,k}^{MAT}\sqrt{\phi^\top(s,a)A_n^{-1}\phi(s,a)}$ **then**
      Perform one step of value iteration using the VTR model as follows: $Q_{h,k}(s,a) = r(s,a) + X_{h,k}^\top \theta_k + \sqrt{\beta_{h,k}}\sqrt{X_{h,k}^\top M_{1,k+1}^{-1} X_{h,k}}$
    else
      Update $Q_{h,k}(s,a)$ according to Equation 8 (Yang & Wang, 2019a) using the UC-MatrixRL model A_k. Note that in (Yang & Wang, 2019a) they use n to denote the current episode, in our paper we use k to denote the current episode.
    end if
    $V_{h,k}(s) = \max_a Q_{h,k}(s,a)$
  end for
end for

We are now using multiple models instead of a single model, we must adjust our confidence sets accordingly. By using a union bound we replace δ with $\delta/2$ for our confidence parameter. This updated confidence parameter changes the term inside the logarithm. We now have $\log(2/\delta)$ where as before we had $\log(1/\delta)$.

G.2. Numerical Results

We will include the cumulative regret and the weighted L1 norm of UCRL-MIX on the RiverSwim environment as in Section 6. We also include a bar graph of the relative frequency with which the algorithm used the VTR-model for planning and the canonical model for planning.

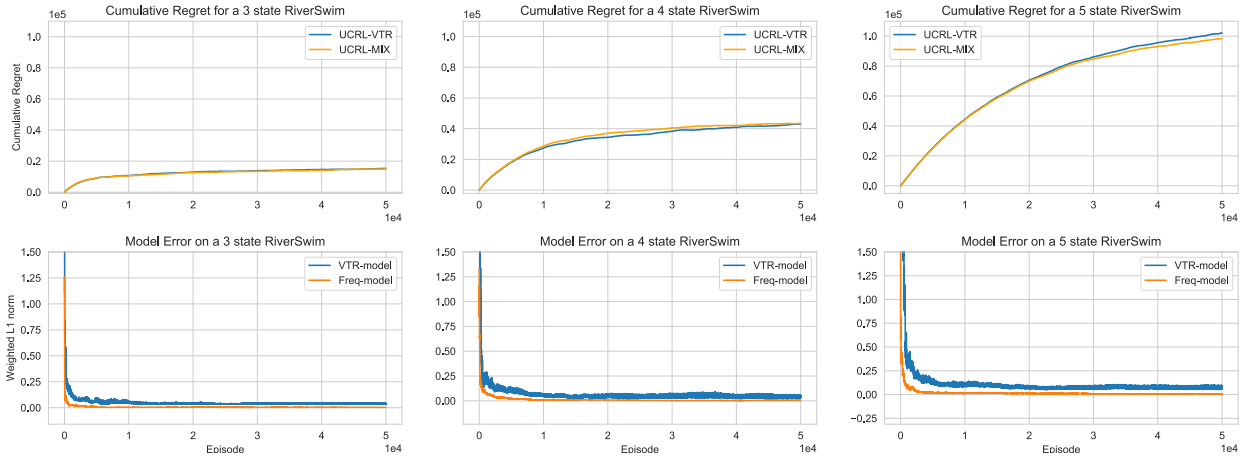

Figure 5. In the plots for the model error we include model error for both the VTR-model and the canonical model. Even though only one is used during planning both are updated at the end of each episode.

If we compare the results of Figure 5 with the results of Figure 2 from Section 6.2 we see that the cumulative regret of UCRL-MIX is almost identical to the cumulative regret of UCRL-VTR. The model errors of both the VTR and the canonical models are almost identical to the model errors of UCRL-VTR and UC-MatrixRL respectively.

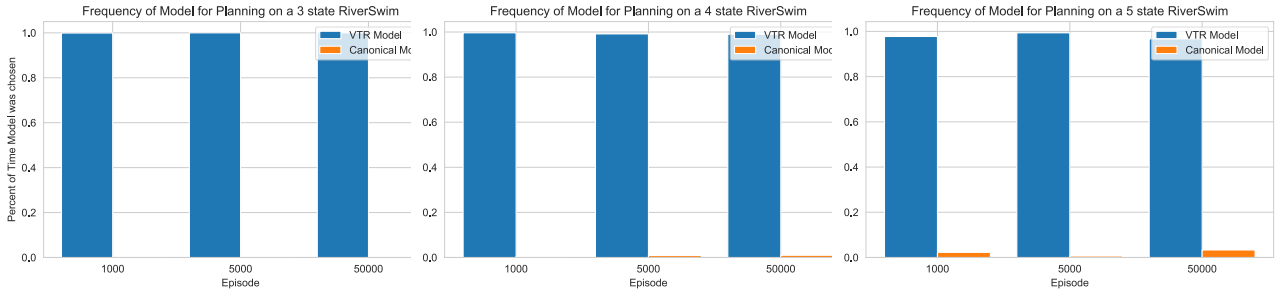

Figure 6. UCRL-MIX rarely, if ever, chooses the canonical model for planning on the RiverSwim environments.

From Figure 6, we see that on the RiverSwim environment, UCRL-MIX almost always uses the VTR-model for planning. We calculate this frequency by counting the number of times Step 7 of Alg 7 was observed up until episode k and by counting the number of times Step 9 of Alg 7 was observed up until episode k. We then divide these counts by the sum of the counts to get a percentage. We believe the reason the algorithm overwhelming chose the VTR-model was due to the fact that the confidence intervals for the VTR-model shrink much faster than the confidence intervals for the canonical model. The canonical model is forced to explore much longer than the VTR-model as its objective is to learn a globally optimal model rather than a model that yields high reward. Thus, the canonical model is forced to explore all state-action-next state tuples, even ones that do not yield high reward, in order to meet its objective of learning a globally optimal model while the VTR-model is only forced to explore state-action-next state tuples that fall in-line with its objective of accumulating high reward. The set of all state-action-next state tuples is much larger then the set of state-action-next state tuples that yield high reward which means the confidence intervals for the canonical model shrink slower than the confidence sets of the VTR-model on the RiverSwim environment.