1 Theoretical Analysis of GME Exploration Bonuses

1.1 Background: Linear MDP and LSVI-UCB

In linear MDPs, transition kernels and the reward function are assumed to be linear. (1) formalizes the definition of linear MDPs as follows:

Definition 1.1 (Linear MDP)**.** *An MDP* $(\mathcal{S}, \mathcal{A}, H, \mathcal{P}, r)$ *is a* linear MDP *with a feature map* $\boldsymbol{\phi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ *if for any* $h \in [H]$*, there exist* d unknown *(signed) measures* $\boldsymbol{\mu}_h = (\mu_h^{(1)}, \ldots, \mu_h^{(d)})$ *over* $\mathcal{S}$ *and an* unknown *vector* $\boldsymbol{\theta}_h \in \mathbb{R}^d$*, such that for any* $(x, a) \in \mathcal{S} \times \mathcal{A}$*:*

$$\mathcal{P}_h(\cdot|x, a) = \langle \boldsymbol{\phi}(x, a), \boldsymbol{\mu}_h(\cdot) \rangle, \qquad r_h(x, a) = \langle \boldsymbol{\phi}(x, a), \boldsymbol{\theta}_h \rangle. \tag{1}$$

In linear MDPs, LSVI-UCB (1) achieves near-optimal worst-case regret. The key idea of LSVI-UCB is to use optimistic Q-values obtained by adding an UCB bonus r^{ucb} to the estimated Q-values. The UCB bonus is defined as:

$$r_t^{\text{ucb}} = \beta \cdot \left[\eta(s_t, a_t)^\top \Lambda_t^{-1} \eta(s_t, a_t)\right]^{1/2},$$

where β is a constant, $\Lambda_t = \sum_{i=0}^m \eta(x_t^i, a_t^i)\eta(x_t^i, a_t^i)^\top + \lambda \cdot \mathbf{I}$ is the Gram matrix, and m is the index of the current episode. The UCB bonus measures the epistemic uncertainty of state-action pairs and has been proven to be efficient. The LSVI-UCB algorithm is described in Algorithm 1. Each iteration of LSVI-UCB consists of two parts: first, in lines 3-6, the agent executes a policy based on Q_t; second, in lines 7-11, the Q-function parameters χ_t are updated via regularized least squares:

$$\chi_t \leftarrow \arg\min_{\chi \in \mathbb{R}^d} \sum_{i=0}^{m} \left[r_t(s_t^i, a_t^i) + \max_{a \in \mathcal{A}} Q_{t+1}(s_{t+1}^i, a) - \chi^\top \eta(s_t^i, a_t^i) \right]^2 + \lambda \|\chi\|^2, \tag{2}$$

where m is the number of episodes and i is the episode index. This least squares problem has a closed-form solution:

$$\chi_t = \Lambda_t^{-1} \sum_{\tau=0}^{m} \eta(x_t^i, a_t^i) \left[r_t(x_t^i, a_t^i) + \max_a Q_{t+1}(x_{t+1}^i, a) \right],$$

where Λ_t is the Gram matrix. The action-value function is estimated via $Q_t(s,a) \approx \chi_t^\top \eta(s,a)$.

LSVI-UCB constructs confidence intervals for the Q-function using the UCB bonus (line 10): $r^{\mathrm{ucb}} = \beta \left[\eta(s,a)^\top \Lambda_t^{-1} \eta(s,a)\right]^{1/2}$, which measures the epistemic uncertainty of state-action pairs. Theoretical analysis shows that with appropriate choices of β and λ, LSVI-UCB achieves near-optimal worst-case regret $\tilde{\mathcal{O}}(\sqrt{d^3T^3L^3})$, where L is the total number of steps. Next, we establish a theoretical connection between the exploration bonus in GME and the UCB bonus.

Algorithm 1 LSVI-UCB Algorithm for Linear MDPs

1: **Initialize:** $\Lambda_t \leftarrow \lambda \cdot \mathbf{I}$ and $w_h \leftarrow 0$
2: **for** episode $m = 0$ **to** $M-1$ **do**
3:     Receive initial state s_0
4:     **for** step $t = 0$ **to** $T-1$ **do**
5:         Execute action $a_t = \arg\max_{a \in \mathcal{A}} Q_t(s_t, a)$ and observe s_{t+1}
6:     **end for**
7:     **for** step $t = T-1$ **downto** 0 **do**
8:         $\Lambda_t \leftarrow \sum_{i=0}^{m} \eta(x_t^i, a_t^i)\eta(x_t^i, a_t^i)^\top + \lambda \cdot \mathbf{I}$
9:         $\chi_t \leftarrow \Lambda_t^{-1} \sum_{i=0}^{m} \eta(x_t^i, a_t^i) \left[r_t(x_t^i, a_t^i) + \max_a Q_{t+1}(x_{t+1}^i, a)\right]$
10:         $Q_t(\cdot,\cdot) = \min\left\{\chi_t^\top \eta(\cdot,\cdot) + \alpha\left[\eta(\cdot,\cdot)^\top \Lambda_t^{-1} \eta(\cdot,\cdot)\right]^{1/2}, T\right\}$
11:     **end for**
12: **end for**

1.2 Theoretical Connection Between GME and LSVI-UCB

In linear MDPs, we represent the prior model as a linear combination of state-action encodings, i.e., $s_{t+1} = W^\top \phi(s_t, a_t) + \epsilon_t$, where $\epsilon_t \sim \mathcal{N}(0, \sigma^2 I)$, and we assume the parameters follow a prior distribution $W \sim \mathcal{N}(0, \Lambda_0^{-1})$. We use standard Bayesian analysis to illustrate our conclusions.

The prior distribution of the parameter matrix $W \in \mathbb{R}^{d \times d}$ is:

$$p(W) = \mathcal{N}(W | \mathbf{0}, \Lambda_0^{-1})$$

where $\Lambda_0 = \lambda I$ is the prior variance matrix. Based on the conjugate prior property of Gaussian distributions, the posterior distribution after t observations is updated as follows:

$$\Lambda_t = \sum_{i=1}^{t} \phi_i \phi_i^\top + \Lambda_0 \quad \text{(Variance Matrix Update)}$$

$$\hat{W}_t = \Lambda_t^{-1} \left(\sum_{i=1}^{t} \phi_i s_{i+1}^\top \right) \quad \text{(Mean Matrix Update)}$$

where $\phi_i = \phi(s_i, a_i) \in \mathbb{R}^d$ is the state-action feature vector, $s_{i+1} \in \mathbb{R}^d$ is the next-state observation, $\Lambda_t \in \mathbb{R}^{d\times d}$ is the posterior variance matrix, and $\hat{W}_t \in \mathbb{R}^{d\times d}$ is the posterior mean matrix.

Theorem 1.2 (Equivalence Between GME Exploration Bonus and UCB)**.** *In linear MDPs, assuming the parameter matrix W follows a Gaussian prior $W \sim \mathcal{N}(0, \Lambda_0^{-1})$, and the latent state $z_t = W\phi(s_t, a_t)$ follows a Gaussian distribution $z_t|\mathcal{D}_t \sim \mathcal{N}(0, \phi_t^\top \Lambda_t^{-1} \phi_t I)$, there exist constants $\beta_1, \beta_2 > 0$ such that the GME exploration bonus satisfies:*

$$\beta_1 \cdot \sqrt{\phi_t^\top \Lambda_t^{-1} \phi_t} \leq r_t^{GME} \leq \beta_2 \cdot \sqrt{\phi_t^\top \Lambda_t^{-1} \phi_t} \tag{3}$$

where $\phi_t = \phi(s_t, a_t)$ and $\Lambda_t = \lambda I + \sum_{i=1}^{t} \phi_i \phi_i^\top$.

Proof. The exploration bonus in GME is formulated as:

$$r_t^{\text{GME}} = \mathcal{H}[p(z_t|\mathcal{D}_t)] + D_{\text{KL}}[p(z_t|\mathcal{D}_t) \| q(z_t|s_t)]$$

For a Gaussian distribution $p(z_t|\mathcal{D}_t) = \mathcal{N}(\mu_t, \Sigma_t)$, the entropy is derived as follows:

$$\begin{aligned}
\mathcal{H}[p(z_t|\mathcal{D}_t)] &= \frac{1}{2} \log\det(2\pi e \Sigma_t) \\
&= \frac{d}{2} \log(2\pi e) + \frac{1}{2} \log\det \Sigma_t \\
&= \frac{d}{2} \log(2\pi e) + \frac{1}{2} \log\det \left(\phi_t^\top \Lambda_t^{-1} \phi_t \cdot I \right) \\
&= \frac{d}{2} \log(2\pi e) + \frac{1}{2} \log \left[(\phi_t^\top \Lambda_t^{-1} \phi_t)^d \det I \right] \\
&= \frac{d}{2} \log(2\pi e) + \frac{d}{2} \log(\phi_t^\top \Lambda_t^{-1} \phi_t) \\
&\propto \frac{d}{2} \log(\phi_t^\top \Lambda_t^{-1} \phi_t)
\end{aligned} \tag{4}$$

The covariance matrix is calculated as follows:

$$
\begin{aligned}
\Sigma_t &= \mathbb{V}[W\phi_t] && (5)\\
&= \mathbb{E}[(W\phi_t - \mathbb{E}W\phi_t)(W\phi_t - \mathbb{E}W\phi_t)^\top] \\
&= \phi_t^\top \mathbb{E}[(W - \hat{W}_t)(W - \hat{W}_t)^\top]\phi_t \\
&= \phi_t^\top \left(\mathbb{E}[WW^\top] - \hat{W}_t\hat{W}_t^\top\right)\phi_t \\
&= \phi_t^\top \Lambda_t^{-1}\phi_t \cdot I \quad \text{(Based on posterior covariance } \mathbb{V}[W] = \Lambda_t^{-1}\text{)} && (6)
\end{aligned}
$$

The KL divergence term expands to:

$$
\begin{aligned}
D_{\mathrm{KL}}[p\|q] &= \frac{1}{2}\left[\mathrm{tr}(\Lambda_t^{-1}\phi_t\phi_t^\top) + (\mu_t - \hat{\mu}_t)^\top \Lambda_t(\mu_t - \hat{\mu}_t)\right] && (7)\\
&= \frac{1}{2}\phi_t^\top \Lambda_t^{-1}\phi_t \cdot \mathrm{tr}(I) + \mathcal{O}(\|\phi_t\|^3) && (8)
\end{aligned}
$$

Combining the two terms, we have:

$$
r_t^{GME} = \frac{d}{2}\log(\phi_t^\top \Lambda_t^{-1}\phi_t) + \frac{d}{2}\phi_t^\top \Lambda_t^{-1}\phi_t + C \tag{9}
$$

Define $v_t = \sqrt{\phi_t^\top \Lambda_t^{-1}\phi_t}$, then the GME exploration bonus can be rewritten as:

$$
\begin{aligned}
r_t^{\mathrm{GME}} &= \frac{d}{2}\log(v_t^2) + \frac{d}{2}v_t^2 + C \\
&= d\log v_t + \frac{d}{2}v_t^2 + C && (10)
\end{aligned}
$$

Applying the arithmetic-geometric mean (AM-GM) inequality:

$$
\begin{aligned}
\frac{\log v_t + v_t^2/2}{2} &\geq \sqrt{\log v_t \cdot v_t^2/2} \quad \text{(AM-GM)} \\
\Rightarrow \log v_t + \frac{v_t^2}{2} &\geq \sqrt{2\log v_t \cdot v_t^2} \\
&= v_t\sqrt{2\log v_t} \\
&\geq \beta_1 v_t \quad \text{(When } v_t \geq 1\text{)} && (11)
\end{aligned}
$$

Using the upper bound of the logarithmic function $\log x \leq x - 1$:

$$\begin{aligned}
r_t^{\mathrm{GME}} &= d \log v_t + \frac{d}{2} v_t^2 + C \\
&\leq d(v_t - 1) + \frac{d}{2} v_t^2 + C \quad \text{(Applying } \log v_t \leq v_t - 1\text{)} \\
&= \frac{d}{2} v_t^2 + d v_t + (C - d) \\
&\leq \frac{d}{2}(v_t^2 + 2v_t) \quad \text{(When } C \leq d\text{)} \\
&= \frac{d}{2}(v_t + 1)^2 - \frac{d}{2} \\
&\leq \beta_2 d v_t \quad \text{(When } v_t \geq 0 \text{ since } (v_t + 1)^2 \leq 2v_t^2 + 2\text{)}
\end{aligned} \tag{12}$$

Choosing $\beta_2 = \max\{1, \sqrt{(2C + 2)/d}\}$ yields:

$$r_t^{\mathrm{GME}} \leq \beta_2 \sqrt{\phi_t^\top \Lambda_t^{-1} \phi_t} \tag{13}$$

□

References

[1] Jin, Chi and Yang, Zhuoran and Wang, Zhaoran and Jordan, Michael I. *Provably efficient reinforcement learning with linear function approximation*. In: Conference on Learning Theory, 2020, pp. 2137–2143. PMLR.