1 Theoretical Analysis of GME Exploration Bonuses

1.1 Background: Linear MDP and LSVI-UCB[1]

In linear MDPs, transition kernels and the reward function are assumed to be linear. 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* $\boldsymbol{\eta} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ *if for any* $h \in [H]$*, there exist* d unknown *(signed) measures* $\boldsymbol{\mu}_h = (\mu_h^{(1)}, \dots, \mu_h^{(d)})$ *over* $\mathcal{S}$ *and an* unknown *vector* $\boldsymbol{\theta}_h \in \mathbb{R}^d$*, such that for any* $(x, a) \in \mathcal{S} \times \mathcal{A}$*:*

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

In linear MDPs, LSVI-UCB 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): $r^{\text{ucb}} = \beta \left[\eta(s,a)^\top \Lambda_t^{-1} \eta(s,a)\right]^{1/2}$, 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: Λ_t ← λ · I and w_h ← 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∈A} Q_t(s_t, a) and observe s_{t+1}
6:      end for
7:      for step t = T − 1 downto 0 do
8:          Λ_t ← Σ_{i=0}^m η(x_t^i, a_t^i) η(x_t^i, a_t^i)^⊤ + λ · I
9:          χ_t ← Λ_t^{-1} Σ_{i=0}^m η(x_t^i, a_t^i) [r_t(x_t^i, a_t^i) + max_a Q_{t+1}(x_{t+1}^i, a)]
10:         Q_t(·,·) = min { χ_t^⊤ η(·,·) + α [η(·,·)^⊤ Λ_t^{-1} η(·,·)]^{1/2}, T }
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., $s_{t+1} = W^\top \eta(s_t, a_t) + \epsilon_t$, where $\epsilon_t \sim \mathcal{N}(0, \sigma^2 I)$, and we assume the parameters follow a prior distribution $W \sim \mathcal{N}(0, \Lambda_0^{-1})$.

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:

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

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

where $\eta_i = \eta(s_i, a_i) \in \mathbb{R}^d$ 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 $z_t = W\eta(s_t, a_t)$ follows a Gaussian distribution $z_t|\mathcal{D}_t \sim \mathcal{N}(0, \eta_t^\top \Lambda_t^{-1} \eta_t I)$, there exist constants $\beta_1, \beta_2 > 0$ such that the GME exploration bonus satisfies:*

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

where $\eta_t = \eta(s_t, a_t)$ and $\Lambda_t = \lambda I + \sum_{i=1}^{t} \eta_i \eta_i^\top$.

Proof. The exploration bonus in GME is formulated as:

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

Let W follow the matrix normal distribution $\mathcal{MN}(0, \Lambda_t^{-1}, I)$ with vectorization $\text{vec}(W) \sim \mathcal{N}(0, \Lambda_t^{-1} \otimes I)$. The latent variable $z_t = W\eta_t$ has covariance:

$$\begin{aligned}
\Sigma_t &= \mathbb{E}[(z_t - \mathbb{E}z_t)(z_t - \mathbb{E}z_t)^\top] \\
&= \eta_t^\top \mathbb{E}[WW^\top] \eta_t \cdot I \\
&= \eta_t^\top \Lambda_t^{-1} \eta_t \cdot I \quad (\text{by } \mathbb{E}[WW^\top] = \Lambda_t^{-1})
\end{aligned} \tag{4}$$

For $p(z_t|\mathcal{D}_t) = \mathcal{N}(0, \eta_t^\top \Lambda_t^{-1} \eta_t \cdot I)$:

$$\begin{aligned}
\mathcal{H}[p(z_t|\mathcal{D}_t)] &= \frac{d}{2} \log(2\pi e) + \frac{1}{2} \log\det(\Sigma_t) \\
&= \frac{d}{2} \log(2\pi e) + \frac{d}{2} \log(\eta_t^\top \Lambda_t^{-1} \eta_t)
\end{aligned} \tag{5}$$

The covariance matrix is calculated as follows:

$$\Sigma_t = \mathbb{V}[W\eta_t] \tag{6}$$
$$= \mathbb{E}[(W\eta_t - \mathbb{E}W\eta_t)(W\eta_t - \mathbb{E}W\eta_t)^\top]$$
$$= \eta_t^\top \mathbb{E}[(W - \hat{W}_t)(W - \hat{W}_t)^\top]\eta_t$$
$$= \eta_t^\top \left(\mathbb{E}[WW^\top] - \hat{W}_t\hat{W}_t^\top\right)\eta_t$$
$$= \eta_t^\top \Lambda_t^{-1}\eta_t \cdot I \quad \text{(Based on posterior covariance } \mathbb{V}[W] = \Lambda_t^{-1}\text{)} \tag{7}$$

Let $q(z_t|s_t) = \mathcal{N}(0, \eta_t^\top \Lambda_0^{-1}\eta_t \cdot I)$. The KL divergence is:

$$D_{\mathrm{KL}}[p\|q] = \frac{1}{2}\left[\log\frac{\det\Sigma_q}{\det\Sigma_p} + \mathrm{tr}(\Sigma_q^{-1}\Sigma_p) - d\right]$$
$$= \frac{1}{2}\left[d\log\frac{\eta_t^\top\Lambda_0^{-1}\eta_t}{\eta_t^\top\Lambda_t^{-1}\eta_t} + \frac{\eta_t^\top\Lambda_t^{-1}\eta_t}{\eta_t^\top\Lambda_0^{-1}\eta_t}\cdot d - d\right] \tag{8}$$

Combining the two terms, we have:

$$r_t^{\mathrm{GME}} = \frac{d}{2}\log(\eta_t^\top\Lambda_t^{-1}\eta_t) + \frac{d}{2}\left[\frac{\eta_t^\top\Lambda_t^{-1}\eta_t}{\eta_t^\top\Lambda_0^{-1}\eta_t} - \log\frac{\eta_t^\top\Lambda_t^{-1}\eta_t}{\eta_t^\top\Lambda_0^{-1}\eta_t} - 1\right] + C \tag{9}$$

where $C = \frac{d}{2}[\log(2\pi e) + \log(\eta_t^\top\Lambda_0^{-1}\eta_t)]$.

Now we let $v_t = \sqrt{\eta_t^\top\Lambda_t^{-1}\eta_t}$ and note that $\lambda_{\min}(\Lambda_t) \geq \lambda$, then:

$$\frac{1}{\lambda}\|\eta_t\|^2 \geq \eta_t^\top\Lambda_t^{-1}\eta_t \geq \frac{\|\eta_t\|^2}{\lambda + t} \geq 0 \quad \text{(by } \Lambda_t \preceq (\lambda + t)I\text{)} \tag{10}$$

Define $f(x) = \frac{d}{2}\log x + \frac{d}{2}\left[\frac{x}{c} - \log\frac{x}{c} - 1\right]$ where $c = \eta_t^\top\Lambda_0^{-1}\eta_t$. Using inequalities:

$$\frac{x}{2} \leq \log(1 + x) \leq x \quad \text{for } 0 \leq x \leq 1 \tag{11}$$
$$\log x \leq x - 1 \quad \forall x > 0 \tag{12}$$

We can show $\exists\beta_1, \beta_2 > 0$ such that:

$$\beta_1 v_t \leq \sqrt{f(v_t^2) + C} \leq \beta_2 v_t \quad \text{(via case analysis on } v_t \leq 1 \text{ and } v_t > 1\text{)} \tag{13}$$

Specifically, choosing:

$$\beta_1 = \sqrt{\frac{d}{2} \min\left\{1, \frac{1}{\lambda}\right\}} \tag{14}$$

$$\beta_2 = \sqrt{d\left(1 + \frac{1}{\lambda}\right)} \tag{15}$$

satisfies the inequality for all $t \geq 0$. $\square$

2 Atari100k and Atari1000k

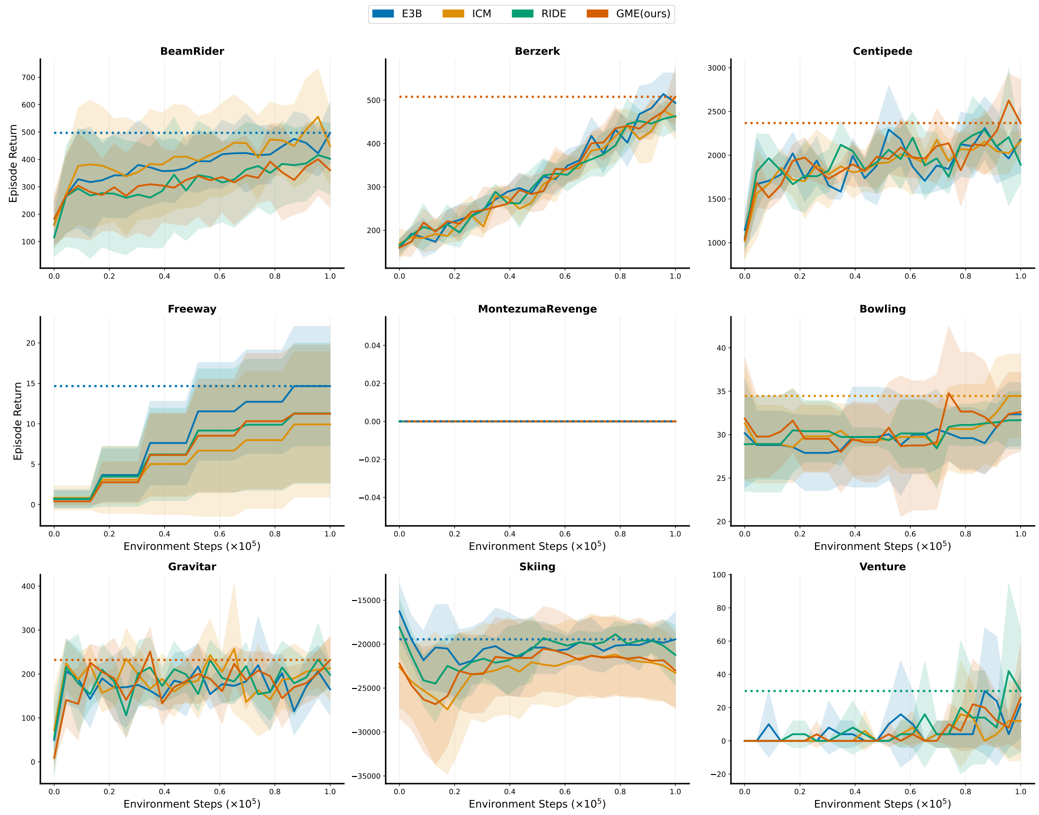

Figure 1: Results on 9 Atari games with 100k step

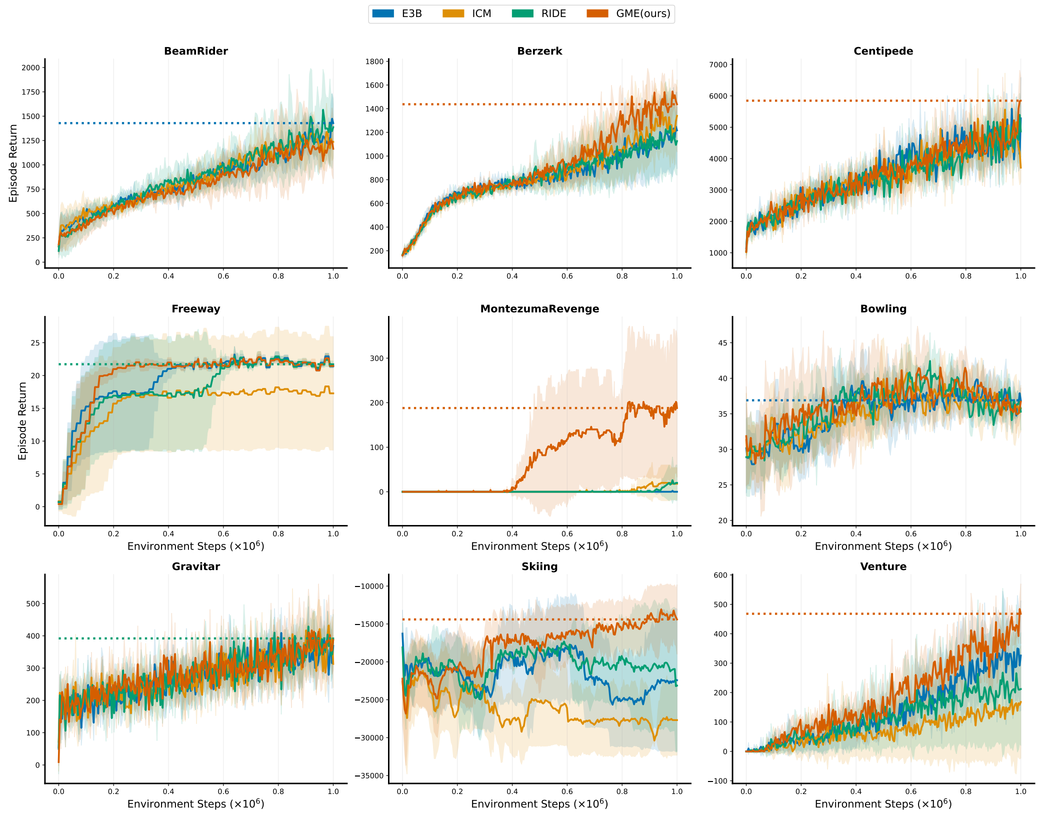

Figure 2: Results on 9 Atari games with 1000k step

References

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