

PRACTICAL ϵ -EXPLORING THOMPSON SAMPLING FOR REINFORCEMENT LEARNING WITH CONTINUOUS CONTROLS

Anonymous authors

Paper under double-blind review

A ADDITIONAL RESULTS

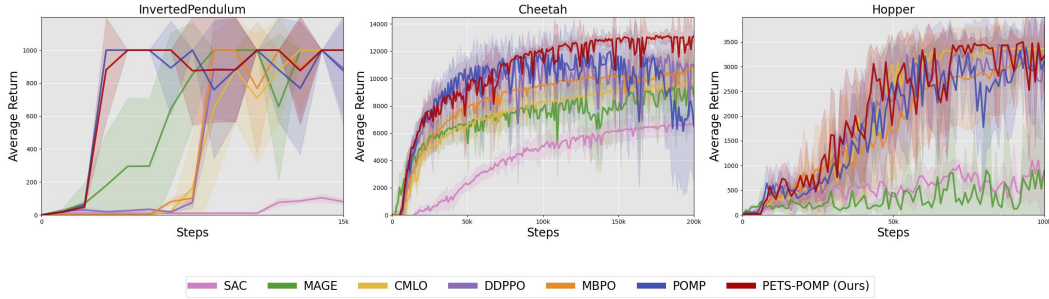


Figure 1: Learning curves of PETS-POMP (Ours) and baselines on three continuous control tasks from OpenAI Gym benchmark. The solid lines represent the mean and the shaded areas represent the standard deviation among trials of 8 different seeds. As shown in the figure, our method achieves better performance compared to baselines on InvertedPendulum and Cheetah while achieving comparable results on Hopper. Moreover, PETS-POMP outperforms POMP on all tasks while also improving its training stability.

B IMPLEMENTATION DETAILS

In this section, we provide some details for the implementation of our method.

B.1 DBAS DETAILS

Here we provide more implementation details for DBAS (?). The pseudocode for DBAS procedure can be found in Algorithm 1. In our method, we use a mixture of Gaussians as the generative model G . We use the Expectation-Maximization (E-M) algorithm (?) to train the mixture of Gaussians. Furthermore, we use an initial set of actions drawn from the policy for x_{init} . We repeat the DBAS procedure (Line 5 of Algorithm 1) for 10-20 iterations.

B.2 HYPERPARAMETERS

In Table 1 and Table 2 and 3 we present the set of hyperparameters used in PETS-POMP, PETS-MBPO and PETS-SAC respectively.

B.3 ACTION INITIALIZATION

As we leverage gradient-based optimization to find the approximate optimal action, action initialization can make a difference in the quality of the approximation. This is also supported in our regret analysis in Appendix C. We empirically have found that initializing the action from the action that

Algorithm 1 DBAS

Input: predictor oracle $O(x)$, $\text{GenTrain}(x_i)$, percentage of least-performing samples $[q = 0.9]$, number of samples $[M = 1000]$, initial data set $[x_{\text{init}} = \emptyset]$

```

1:  $set \leftarrow x_{\text{init}}$ 
2: if  $x_{\text{init}}$  is empty then
3:    $set \leftarrow$  randomly initialized data
4: end if
5: while not converged do
6:    $G \leftarrow \text{GenTrain}(set)$ 
7:    $set \leftarrow x_i \sim G$ 
8:    $scores_i \leftarrow O(x_i)$ 
9:    $set \leftarrow x_i$  if it is not among the  $q^{th}$  percentage least-performing samples based on  $scores$ 
10: end while
11: return  $set_0$ 

```

would have been taken by the underlying algorithm without the exploration helps with getting better approximations.

Table 1: Set of hyperparameters used in PETS-POMP.

		Inverted Pendulum	Walker2d	Cheetah	Ant	Humanoid	Hopper
ϵ	exploration probability	0.6	0.3	0.8	0.3	0.8	0.8
n_{samples}	number of posterior samples	5	5	50	10	5	10
$n_{\text{grad_steps}}$	number of gradient ascent steps	40	100	100	60	100	50
η'	gradient ascent learning rate	0.002	0.01	0.01	0.05	0.05	0.005

Table 2: Set of hyperparameters used in PETS-POMP.

		Cheetah	Ant	Hopper
ϵ	exploration probability	0.8	0.4	0.4
n_{samples}	number of posterior samples	50	5	5
$n_{\text{grad_steps}}$	number of gradient ascent steps	100	40	40
η'	gradient ascent learning rate	0.01	0.01	0.01

Table 3: Set of hyperparameters used in PETS-SAC.

		Cheetah	Ant	Hopper
ϵ	exploration probability	0.7	0.4	1.0
n_{samples}	number of posterior samples	5	5	5
$n_{\text{grad_steps}}$	number of gradient ascent steps	20	40	80
η'	gradient ascent learning rate	0.02	0.01	0.01

C REGRET ANALYSIS

We expand upon the regret analysis presented in ?, extending it to approximate greedy policies. We demonstrate that under certain assumptions on the action-value function, the regret bound of $\tilde{O}\left(d^{3/2}H^{3/2}\sqrt{T}\right)$ can be achieved, even when the optimal action a^* cannot be trivially identified and has to be approximated.

To this end, we first state the analysis setting. Consider an episodic MDP of the form $(S, \mathcal{A}, H, \mathbb{P}, r)$ where S is the state space, \mathcal{A} is the continuous action space, H is the episode length, $\mathbb{P} = \{\mathbb{P}_h\}_{h=1}^H$ are the state transition probability distributions, and $r = \{r_h\}_{h=1}^H$ are the reward functions where $r_h : S \times \mathcal{A} \rightarrow [0, 1]$.

Furthermore, $\pi_h(x)$ denotes the action that the agent takes in the state x at the h -th step in the episode, and π is the set of policies. The value and action-value functions are defined as:

$$V_h^\pi(x) = \mathbb{E}_\pi \left[\sum_{h'=h}^H r_{h'}(x_{h'}, a_{h'}) \mid x_h = x \right].$$

$$Q_h^\pi(x, a) = \mathbb{E}_\pi \left[\sum_{h'=h}^H r_{h'}(x_{h'}, a_{h'}) \mid x_h = x, a_h = a \right].$$

The Bellman equation and Bellman optimality equations are as follows:

$$Q_h^\pi(x, a) = (r_h + \mathbb{P}_h V_{h+1}^\pi)(x, a), \quad V_h^\pi(x) = Q_h^\pi(x, \pi_h(x)), \quad V_{H+1}^\pi(x) = 0.$$

$$Q_h^*(x, a) = (r_h + \mathbb{P}_h V_{h+1}^*)(x, a), \quad V_h^*(x) = Q_h^*(x, \pi_h^*(x)), \quad V_{H+1}^*(x) = 0.$$

where $V_h^*(x) = V_h^{\pi^*}(x)$, $Q_h^*(x, a) = Q_h^{\pi^*}(x, a)$, π^* is the optimal policy and $[\mathbb{P}_h V_{h+1}](x, a) = \mathbb{E}_{x' \sim \mathbb{P}_h(\cdot \mid x, a)} V_{h+1}(x')$.

We measure the suboptimality of an agent by the total regret defined as

$$\text{Regret}(K) = \sum_{k=1}^K \left[V_1^*(x_1^k) - V_1^{\pi^k}(x_1^k) \right] \quad (1)$$

where x_1^k is the initial state and π_k is the policy agent uses for episode k . and K is the total number of episodes which the agent interacts with the environment with the goal of learning the optimal policy.

Consider the following loss function for the action-value function from ?:

$$L_h^k(w_h) = \sum_{\tau=1}^{k-1} \left[r_h(x_h^\tau, a_h^\tau) + \max_{a \in \mathcal{A}} Q_{h+1}^k(x_{h+1}^\tau, a) - Q(w_h; \phi(x_h^\tau, a_h^\tau)) \right]^2 + \lambda \|w_h\|^2 \quad (2)$$

where $\phi(\cdot, \cdot)$ is a feature vector, w_h is the Q function parameters and λ is the regularization constant. We consider a linear function approximation for the Q function and:

$$Q_h^k(\cdot, \cdot) \leftarrow \min \left\{ \phi(\cdot, \cdot)^\top w_h^{k, J_k}, H - h + 1 \right\}^+ \quad (3)$$

where w_h^{k, J_k} is the parameter vector obtained after J_k iterations of the Langevin Monte Carlo (LMC) process, applied to the loss function defined in Eq. equation 4, as described in Algorithm 2. We further denote $V_h^k(x_h^k) = \max_{a \in \mathcal{A}} Q_h^k(x_h^k, a)$.

Note that while the action-value function Q_h^k is linear with respect to the parameter vector w , it is not necessarily linear in the action a . Furthermore, the loss function $L_h^k(w_h)$ includes the term $V_{h+1}^k(x_{h+1}^\tau)$, which, in a high-dimensional continuous action space, cannot be computed exactly due to infinite actions. Consequently, in Algorithm 2 and our regret analysis, we substitute this term with the approximate value function, \hat{V}_h^k , as defined in Equation equation ???. This approach leads to the formulation of the following modified loss function:

Algorithm 2 Langevin Monte Carlo Least-Squares Value Iteration (LMC-LSVI) with Approximate Greedy Policy

Input: step sizes $\{\eta_k > 0\}_{k \geq 1}$, inverse temperature $\{\beta_k\}_{k \geq 1}$, loss function $L_k(w)$

```

1: Initialize  $w_h^{1,0} = \mathbf{0}$  for  $h \in [H]$ ,  $J_0 = 0$ 
2: for episode  $k = 1, 2, \dots, K$  do
3:   Receive the initial state  $s_1^k$ 
4:   for step  $h = H, H-1, \dots, 1$  do
5:      $w_h^{k,0} = w_h^{k-1, J_{k-1}}$ 
6:     for  $j = 1, \dots, J_k$  do
7:        $\epsilon_h^{k,j} \sim \mathcal{N}(0, I)$ 
8:        $w_h^{k,j} = w_h^{k,j-1} - \eta_k \nabla L_h^k(w_h^{k,j-1}) + \sqrt{2\eta_k \beta_k^{-1}} \epsilon_h^{k,j}$ 
9:     end for
10:     $Q_h^k(\cdot, \cdot) \leftarrow \min \left\{ Q(w_h^{k, J_k}; \phi(\cdot, \cdot)), H - h + 1 \right\}^+$ 
11:    initialize set of actions  $a_{h,0}^k$ 
12:    for iteration  $t = 1, 2, \dots, t^*$  do
13:       $a_{h,t}^k = a_{h,t-1}^k + \nabla Q(\cdot, a_{h,t-1}^k)$  {as outlined in Eq ??}
14:    end for
15:     $\hat{V}_h^k(\cdot) \leftarrow Q(\cdot, a_{h,t^*}^k)$  {approximate optimal value}
16:  end for
17:  for step  $h = 1, 2, \dots, H$  do
18:    Take approximate optimal action  $a_h^k$  based on  $a_{h,t^*}^k$ , observe reward  $r_h^k(s_h^k, a_h^k)$  and next state  $s_{h+1}^k$ 
19:  end for
20: end for

```

$$L_h^k(w_h) = \sum_{\tau=1}^{k-1} \left[r_h(x_h^\tau, a_h^\tau) + \hat{V}_{h+1}^k(x_{h+1}^\tau) - Q(w_h; \phi(x_h^\tau, a_h^\tau)) \right]^2 + \lambda \|w_h\|^2 \quad (4)$$

We present a modified version of the Langevin Monte Carlo Least Squares Value Iteration (LMC_LSVI) algorithm (?) in Algorithm 2. Contrasting with Algorithm 1 in ?, Algorithm 2 incorporates the use of an approximate value function, denoted as \hat{V} , and employs approximate optimal actions. In the subsequent sections, we demonstrate that under certain assumptions about the action-value function, Algorithm 2 achieves the same target regret bound.

Proposition C.1. *As defined in ?, let w^{k, J_k} be the approximation of posterior parameters after J_k iterations of LMC as defined in Eq ?? for the k 'th episode where h is the horizon step. Under the loss defined in Eq 4, w^{k, J_k} follows a Gaussian distribution where the mean vector and covariance matrix are defined as:*

$$\mu_h^{k, J_k} = A_k^{J_k} \dots A_1^{J_1} w_h^{1,0} + \sum_{i=1}^k A_k^{J_k} \dots A_{i+1}^{J_{i+1}} (I - A_i^{J_i}) \hat{w}_h^i, \quad (5)$$

$$\Sigma_h^{k, J_k} = \sum_{i=1}^k \frac{1}{\beta_i} A_k^{J_k} \dots A_{i+1}^{J_{i+1}} (I - A_i^{2J_i}) (\Lambda_h^i)^{-1} (I + A_i)^{-1} A_{i+1}^{J_{i+1}} \dots A_k^{J_k}, \quad (6)$$

where $A_i = I - 2\eta_i \Lambda_h^i$ for $i \in [k]$.

Proof. We refer readers to Proposition B.1 of ? for the proof. \square

Definition C.2. (Model Prediction Error). For any $(k, h) \in [K] \times [H]$, we define the model prediction error associated with the reward r_h ,

$$l_h^k(x, a) = r_h(x, a) + \mathbb{P}_h \hat{V}_{h+1}^k(x, a) - Q_h^k(x, a). \quad (7)$$

Proposition C.3. For an action-value function $Q(x, \cdot)$ that has an L -Lipschitz continuous gradient and satisfies the Polyak-Łojasiewicz Inequality (PL) inequality for some $\mu > 0$ as stated below:

$$\frac{1}{2} \|\nabla Q_h^k(x, a)\|^2 \geq \mu (Q_h^k(x, a) - Q_h^k(x, a^*)), \quad \forall a. \quad (8)$$

where

$$a^* = \operatorname{argmax}_a Q_h^k(x_h, a)$$

is the optimal action. For $t^* \geq \left(\frac{L}{\mu} \log(KH \frac{Q_h^k(x_h, a^*) - Q_h^k(x_h, a_0)}{d^{3/2} H^{3/2} \sqrt{T}}) \right)$ iterations we have:

$$V_h^k(x_h) - \hat{V}_h^k(x_h) \leq \varepsilon_{t^*} \quad (9)$$

where $\varepsilon_{t^*} = \frac{d^{3/2} H^{3/2} \sqrt{T}}{KH}$.

Proof. As proved in Theorem 1. of ?, with the step size of $\frac{1}{L}$, for an action-value function $Q(x, \cdot)$ that has an L -Lipschitz continuous gradient and satisfies the PL inequality, we have

$$Q_h^k(x_h, a^*) - Q_h^k(x_h, a_t) \leq \left(1 - \frac{\mu}{L}\right)^t (Q_h^k(x_h, a^*) - Q_h^k(x_h, a_0)) \quad (10)$$

where a_t is the action after t iteration of gradient ascent:

$$a_{t+1} = a_t + \nabla Q_h^k(x_h, a_t) \quad (11)$$

using $1 - u \leq \exp(-u)$ on Eq 10 we have

$$Q_h^k(x_h, a^*) - Q_h^k(x_h, a_t) \leq \exp\left(-t \frac{\mu}{L}\right) (Q_h^k(x_h, a^*) - Q_h^k(x_h, a_0)).$$

So for $t^* \geq \frac{L}{\mu} \log(KH \frac{Q_h^k(x_h, a^*) - Q_h^k(x_h, a_0)}{d^{3/2} H^{3/2} \sqrt{T}})$ we have:

$$Q_h^k(x_h, a^*) - Q_h^k(x_h, a_{t^*}) \leq \frac{d^{3/2} H^{3/2} \sqrt{T}}{KH} = \varepsilon_{t^*} \quad (12)$$

by applying the definition of V_h^k we have:

$$V_h^k(x_h) - \hat{V}_h^k(x_h) \leq \varepsilon_{t^*} \quad (13)$$

□

Proposition C.4. Under the approximate value function \hat{V}_h^k we have:

$$\mathbb{P}_h V_{h+1}^k(x, a) - \mathbb{P}_h \hat{V}_{h+1}^k(x, a) \leq \varepsilon_{t^*} \quad (14)$$

where ε_{t^*} is defined in Proposition C.3

Proof. Applying the definition of \mathbb{P}_h we have:

$$[\mathbb{P}_h V_{h+1}^k](x, a) = \mathbb{E}_{x' \sim \mathbb{P}_h(\cdot | x, a)} V_{h+1}^k(x') \leq \mathbb{E}_{x' \sim \mathbb{P}_h(\cdot | x, a)} \hat{V}_{h+1}^k(x') + \varepsilon_{t^*} \quad (15)$$

$$= [\mathbb{P}_h \hat{V}_{h+1}^k](x, a) + \varepsilon_{t^*} \quad (16)$$

where the first to the second line is by using Proposition C.3. □

Lemma C.5. Let $\lambda = 1$ in Eq 4, Define the following event

$$\begin{aligned} \mathcal{E}(K, H, \delta) &= \left\{ \left| \phi(x, a)^\top \hat{w}_h^k - r_h(x, a) - \mathbb{P}_h V_{h+1}^k(x, a) \right| \right. \\ &\quad \left. \leq 5H\sqrt{d}C_\delta \|\phi(x, a)\|_{(\Lambda_h^k)^{-1}}, \forall (h, k) \in [H] \times [K] \text{ and } \forall (x, a) \in \mathcal{S} \times \mathcal{A} \right\}. \end{aligned} \quad (17)$$

where we denote

$$C_\delta = \left[\frac{1}{2} \log(K+1) + \log\left(\frac{2\sqrt{2}KB_{\delta/2}}{H}\right) + \log\frac{2}{\delta} \right]^{1/2}$$

and $B_\delta = \left(\frac{16}{3} H d \sqrt{K} + \sqrt{\frac{2K}{3\beta_K \delta}} d^{3/2} \right)$. Then we have $\mathbb{P}(\mathcal{E}(K, H, \delta)) \geq 1 - \delta$.

Proof. We refer readers to Lemma. B5. of (?) for the proof. \square

Lemma C.6. *Let $\lambda = 1$ in Eq 4. For any $\delta \in (0, 1)$ conditioned on the event $\mathcal{E}(K, H, \delta)$, for all $(h, k) \in [H] \times [K]$ and $(x, a) \in \mathcal{S} \times \mathcal{A}$, with probability at least $1 - \delta^2$, we have*

$$- (r_h(x, a) + \mathbb{P}_h V_{h+1}^k(x, a) - Q_h^k(x, a) l_h^k(x, a)) \quad (18)$$

$$\leq \left(5H\sqrt{d}C_\delta + 5\sqrt{\frac{2d\log(1/\delta)}{3\beta_K}} + 4/3 \right) \|\phi(x, a)\|_{(\Lambda_h^k)^{-1}}, \quad (19)$$

where C_δ is defined in Lemma C.5.

Proof. We refer readers to Lemma. B.6 of (?) for the proof. \square

Lemma C.7. (Error bound). *Let $\lambda = 1$ in Eq 4. For any $\delta \in (0, 1)$ conditioned on the event $\mathcal{E}(K, H, \delta)$, for all $(h, k) \in [H] \times [K]$ and $(x, a) \in \mathcal{S} \times \mathcal{A}$, with probability at least $1 - \delta^2$, we have*

$$-l_h^k(x, a) \leq \left(5H\sqrt{d}C_\delta + 5\sqrt{\frac{2d\log(1/\delta)}{3\beta_K}} + 4/3 \right) \|\phi(x, a)\|_{(\Lambda_h^k)^{-1}} + \varepsilon_{t^*}, \quad (20)$$

Proof. using Lemma C.6 and Proposition C.4 we have:

$$\begin{aligned} & - \left(r_h(x, a) + \mathbb{P}_h \hat{V}_{h+1}^k(x, a) - Q_h^k(x, a) l_h^k(x, a) + \varepsilon_{t^*} \right) \\ & = l_h^k(x, a) - \varepsilon_{t^*} \leq \left(5H\sqrt{d}C_\delta + 5\sqrt{\frac{2d\log(1/\delta)}{3\beta_K}} + 4/3 \right) \|\phi(x, a)\|_{(\Lambda_h^k)^{-1}} \end{aligned}$$

Lemma C.8. *Let $\lambda = 1$ in Eq 4. Conditioned on the event $\mathcal{E}(K, H, \delta)$, for all $(h, k) \in [H] \times [K]$ and $(x, a) \in \mathcal{S} \times \mathcal{A}$, with probability at least $\frac{1}{2\sqrt{2e\pi}}$, we have*

$$- (r_h(x, a) + \mathbb{P}_h V_{h+1}^k(x, a) - Q_h^k(x, a)) \leq 0 \quad (21)$$

Proof. We refer readers to Lemma. B.7 of (?) for the proof. \square

Lemma C.9. (Optimism). *Let $\lambda = 1$ in Eq 4. Conditioned on the event $\mathcal{E}(K, H, \delta)$, for all $(h, k) \in [H] \times [K]$ and $(x, a) \in \mathcal{S} \times \mathcal{A}$, with probability at least $\frac{1}{2\sqrt{2e\pi}}$, we have*

$$l_h^k(x, a) \leq 0 \quad (22)$$

Proof. We immediately get the stated result by using Proposition C.4 on Lemma C.8. \square

We restate the main theorem:

Theorem C.10. *Let $\lambda = 1$ in Eq 4, $\frac{1}{\beta_k} = \tilde{O}(H\sqrt{d})$ in Algorithm 2 and $\delta \in \left(\frac{1}{2\sqrt{2e\pi}}, 1\right)$. For any episode $k \in [K]$, let the learning rate $\eta_k = 1 / (4\lambda_{\max}(\Lambda_h^k))$, the update number for LMC in Eq ?? be $J_k = 2\kappa_k \log(4HKd)$ where $\kappa_k = \lambda_{\max}(\Lambda_h^k) / \lambda_{\min}(\Lambda_h^k)$ is the condition number of Λ_h^k defined in Proposition C.1. Under the assumption that the action-value function Q_h^k in Eq 3 has an L -Lipschitz continuous gradient and satisfies the Polyak-Łojasiewicz Inequality (PL) inequality Eq 8, the regret of Algorithm 2 under the regret definition in Definition 1, satisfies*

$$\text{Regret}(K) = \tilde{O}\left(d^{3/2}H^{3/2}\sqrt{T}\right), \quad (23)$$

with probability at least $1 - \delta$.

Proof of Theorem C.10. By Lemma. 4.2 in (?), it holds that

$$\text{Regret}(T) = \sum_{k=1}^K \left(V_1^* (x_1^k) - V_1^{\pi^k} (x_1^k) \right) \quad (24)$$

$$= \underbrace{\sum_{k=1}^K \sum_{t=1}^H \mathbb{E}_{\pi^*} [\langle Q_h^k(x_h, \cdot), \pi_h^*(\cdot | x_h) - \pi_h^k(\cdot | x_h) \rangle | x_1 = x_1^k]}_{(i)} + \underbrace{\sum_{k=1}^K \sum_{t=1}^H \mathcal{D}_h^k}_{(ii)} \quad (25)$$

$$+ \underbrace{\sum_{k=1}^K \sum_{t=1}^H \mathcal{M}_h^k}_{(iii)} + \underbrace{\sum_{k=1}^K \sum_{h=1}^H (\mathbb{E}_{\pi^*} [l_h^k(x_h, a_h) | x_1 = x_1^k] - l_h^k(x_h^k, a_h^k))}_{(iv)}, \quad (26)$$

where $\langle \cdot, \cdot \rangle$ denotes inner product which in continuous spaces is defined as $\langle f, g \rangle = \int_D f(t)g(t) dt$. Furthermore, \mathcal{D}_h^k and \mathcal{M}_h^k are defined as

$$\begin{aligned} \mathcal{D}_h^k &:= \left\langle \left(Q_h^k - Q_h^{\pi^k} \right) (x_h^k, \cdot), \pi_h^k(\cdot, x_h^k) \right\rangle - \left(Q_h^k - Q_h^{\pi^k} \right) (x_h^k, a_h^k), \\ \mathcal{M}_h^k &:= \mathbb{P}_h \left(\left(V_{h+1}^k - V_{h+1}^{\pi^k} \right) (x_h^k, a_h^k) - \left(V_{h+1}^k - V_{h+1}^{\pi^k} \right) (x_h^k) \right). \end{aligned}$$

Bounding Term (i): Using Proposition C.3 we have $Q(x_h, a^*) - Q(x_h, a_{t^*}) \leq \varepsilon_{t^*}$.

Note that π_h^k is approximately greedy w.r.t Q_h^k and $\pi_h^k(a_{t^*}) = 1$ where a_{t^*} is the approximate optimal greedy action from Eq 12. The largest value that $\langle Q_h^k(x_h, \cdot), \pi_h^*(\cdot | x_h) - \pi_h^k(\cdot | x_h) \rangle$ in Eq 25 can take is $Q(x_h, a^*) - Q(x_h, a_{t^*})$ which happens if $\pi_h^*(a^*) = 1$ where $a^* = \arg\max_a Q_h^k(x_h, a)$. This completes the proof using Eq 12.

Bound for Term (ii): With probability $1 - \delta/3$ we have:

$$\sum_{k=1}^K \sum_{h=1}^H \mathcal{D}_h^k \leq \sqrt{2H^2T \log(3/\delta)} \quad (27)$$

We refer the readers to the Appendix. B.2 of (?) for the proof.

Bound for Term (iii): With probability $1 - \delta/3$ we have:

$$\sum_{k=1}^K \sum_{h=1}^H \mathcal{M}_h^k \leq \sqrt{2H^2T \log(3/\delta)}. \quad (28)$$

We refer the readers to the Appendix. B.2 of ? for the proof.

Bound for Term (iv): With probability at least $\left(1 - \frac{\delta}{3} - \frac{1}{2\sqrt{2e\pi}}\right)$ we have:

$$\sum_{k=1}^K \sum_{h=1}^H (\mathbb{E}_{\pi^*} [l_h^k(x_h, a_h) | x_1 = x_1^k] - l_h^k(x_h^k, a_h^k)) \leq \tilde{O} \left(d^{3/2} H^{3/2} \sqrt{T} \right) \quad (29)$$

Suppose the event $\mathcal{E}(K, H, \delta')$ holds. by union bound, with probability $1 - \left(\delta'^2 + \frac{1}{2\sqrt{2e\pi}}\right)$, we have,

$$\sum_{k=1}^K \sum_{h=1}^H (\mathbb{E}_{\pi^*} [l_h^k(x_h, a_h) \mid x_1 = x_1^k] - l_h^k(x_h^k, a_h^k)) \quad (30)$$

$$\leq \sum_{k=1}^K \sum_{h=1}^H -l_h^k(x_h^k, a_h^k) \quad (31)$$

$$\leq \sum_{k=1}^K \sum_{h=1}^H \left(5H\sqrt{d}C_{\delta'} + 5\sqrt{\frac{2d\log(1/\delta')}{3\beta_K}} + 4/3 \right) \|\phi(x_h^k, a_h^k)\|_{(\Lambda_h^k)^{-1}} + KH\varepsilon_{t^*} \quad (32)$$

$$= \left(5H\sqrt{d}C_{\delta'} + 5\sqrt{\frac{2d\log(1/\delta')}{3\beta_K}} + 4/3 \right) \sum_{k=1}^K \sum_{h=1}^H \|\phi(x_h^k, a_h^k)\|_{(\Lambda_h^k)^{-1}} + KH\varepsilon_{t^*} \quad (33)$$

$$\leq \left(5H\sqrt{d}C_{\delta'} + 5\sqrt{\frac{2d\log(1/\delta')}{3\beta_K}} + 4/3 \right) \sum_{h=1}^H \sqrt{K} \left(\sum_{k=1}^K \|\phi(x_h^k, a_h^k)\|_{(\Lambda_h^k)^{-1}}^2 \right)^{1/2} + KH\varepsilon_{t^*} \quad (34)$$

$$\leq \left(5H\sqrt{d}C_{\delta'} + 5\sqrt{\frac{2d\log(1/\delta')}{3\beta_K}} + 4/3 \right) H\sqrt{2dK\log(1+K)} + KH\varepsilon_{t^*} \quad (35)$$

$$= \left(5H\sqrt{d}C_{\delta'} + 5\sqrt{\frac{2d\log(1/\delta')}{3\beta_K}} + 4/3 \right) \sqrt{2dHT\log(1+K)} + KH\varepsilon_{t^*} \quad (36)$$

$$= \left(5H\sqrt{d}C_{\delta'} + 5\sqrt{\frac{2d\log(1/\delta')}{3\beta_K}} + 4/3 \right) \sqrt{2dHT\log(1+K)} + \left(d^{3/2}H^{3/2}\sqrt{T} \right) \quad (37)$$

$$= \tilde{O}\left(d^{3/2}H^{3/2}\sqrt{T}\right). \quad (38)$$

Here the first, the second, and the third inequalities follow from Lemma C.9, Lemma C.7 and the Cauchy-Schwarz inequality respectively. The last inequality follows from Lemma C.5 The last equality follows from $\frac{1}{\sqrt{\beta_K}} = 10H\sqrt{d}C_{\delta'} + \frac{8}{3}$ which we defined in Lemma C.9. By Lemma C.7, the event $\mathcal{E}(K, H, \delta')$ occurs with probability $1 - \delta'$. Thus, by union bound, the event $\mathcal{E}(K, H, \delta')$ occurs and it holds that

$$\sum_{k=1}^K \sum_{h=1}^H (\mathbb{E}_{\pi^*} [l_h^k(x_h, a_h) \mid x_1 = x_1^k] - l_h^k(x_h^k, a_h^k)) \leq \tilde{O}\left(d^{3/2}H^{3/2}\sqrt{T}\right)$$

By applying union bound for (i), (ii), (iii) and (iv), the final regret bound is $\tilde{O}\left(d^{3/2}H^{3/2}\sqrt{T}\right)$ with at least probability $1 - \delta$ where $\delta \in (\frac{1}{2\sqrt{2e\pi}}, 1)$.

Theorem C.11. Let $\lambda = 1$ in Eq 4, $\frac{1}{\beta_k} = \tilde{O}(H\sqrt{d})$ in Algorithm 2 and $\delta \in (\frac{1}{2\sqrt{2e\pi}}, 1)$. For any episode $k \in [K]$, let the learning rate $\eta_k = 1/(4\lambda_{\max}(\Lambda_h^k))$, the update number for LMC in Eq ?? be $J_k = 2\kappa_k \log(4HKd)$ where $\kappa_k = \lambda_{\max}(\Lambda_h^k)/\lambda_{\min}(\Lambda_h^k)$ is the condition number of Λ_h^k defined in Proposition C.1. Let $\vec{w} = [w_1, w_2, \dots, w_n]^T$ be the extended parameter space and $Q(x, a) = \max_{i \in [n]} Q_{w_i}(x, a)$ be the optimistic action-value function. Under the assumption that the action-value function Q_h^k in Eq 3 has an L -Lipschitz continuous gradient and satisfies the Polyak-Lojasiewicz Inequality (PL) inequality Eq 8, the regret of Algorithm 2 under the regret definition in Definition 1, satisfies

$$\text{Regret}(K) = \tilde{O}\left(d^{3/2}H^{3/2}\sqrt{T}\right), \quad (39)$$

with probability of $1 - \epsilon'$ for any $\epsilon' \in (0, 1)$.

Proof of Theorem C.11. In Lemma B.9 we prove that the estimation $Q_h^k(x, a)$ is optimistic with a constant probability of at least $\frac{1}{2\sqrt{2e\pi}}$. In other words, the failure probability is at most $1 - \frac{1}{2\sqrt{2e\pi}}$. By extending the parameter space $\vec{w} = [w_1, w_2, \dots, w_n]^T$ and modelling the optimistic action-value function using $Q(x, a) = \max_{i \in [n]} Q_{w_i}(x, a)$, the failure probability will be at most $(1 - \frac{1}{2\sqrt{2e\pi}})^n$. We want this probability to be arbitrarily small. To guarantee that the failure probability is less than ϵ' it suffices to find an n that is large enough such that $(1 - \frac{1}{2\sqrt{2e\pi}})^n < \epsilon'$. If we solve for n we have $n > \frac{\log \epsilon'}{\log(1 - \frac{1}{2\sqrt{2e\pi}})}$. We can express the latter quantity as $\frac{\log(1/\delta)}{\log(2\sqrt{2e\pi}) - \log(2\sqrt{2e\pi} - 1)} \in \Omega(\log(1/\epsilon'))$. So, we can extend the parameter space by a factor of $\Omega(\log(1/\epsilon'))$ to ensure that the failure probability is less than ϵ' . Finally, we can apply the union bound on (i), (ii), (iii), and (iv) to conclude that the regret bound in Theorem C.11 holds with a probability of $1 - \epsilon'$ for any $\epsilon' \in (0, 1)$.