Technical Argumentation for "Balancing Exploration and Exploitation in Hierarchical Reinforcement Learning via Latent Landmark Graphs"

Anonymous submission

We provide theoretical support for the proposed graph-based approach (HILL) as follows:

Universal Markov Decision Process

An Universal Markov Decision Process (UMDP) (Schaul et al. 2015) extends an MDP with a set of subgoals G. UMDP has a pseudo-reward function $R_g : S \times A \times G \to \mathbb{R}$ and a pseudo-discount factor γ_g, where S is the state space and A is the action space. γ_g takes the double role of state-dependent discounting and soft termination. γ_g turns to 0 if and only if s is a terminal state according to the subgoal g. For any policy $\pi : S \to A$, a general value function is then defined to represent the expected cumulative pseudo-discounted future pseudo-return:

$$V_{g,\pi}(s) = E\Big[\sum_{t=0}^{\infty} R_g(s_t, a_t, g_t) \prod_{k=0}^{t} \gamma_g(s_k)\Big| s_0 = s\Big], \quad (1)$$

where $a_t = \pi(s_t)$. $V_{g,\pi}(s)$ is called goal-conditioned value, or universal value.

View UMDP as a Graph

HILL refers to a family of UMDP tasks. The state space S is a low-dimension manifold in the latent space. Assume that the low-level policy $\pi_{\theta_l}(a|s, g)$ takes at most T_g steps to realize g and the reward at each step r_g^t's absolute value is bounded by R_g^{max}. Let $q_{\pi_{\theta_l}}(s, g)$ be the expected total reward along an episode, $d_{\pi_{\theta_l}}(s, g) = -q_{\pi_{\theta_l}}(s, g)$ for all s, g and $\epsilon \in [0, 1]$. We can prove:

$$|V_{g,\pi_{\theta_l}}(s) - q_{\pi_{\theta_l}}(s, g)| \quad (2)$$

$$= |E[\sum_{t=1}^{T_g} r_g^t (1-\epsilon)^{t-1}] - E[\sum_{t=1}^{T_g} r_g^t]| \quad (3)$$

$$\approx |E[\sum_{t=1}^{T_g} r_g^t - (t-1)\epsilon r_g^t] - E[\sum_{t=1}^{T_g} r_g^t]| \quad (4)$$

$$= |\sum_{t=1}^{T_g} (t-1)\epsilon E[r_g^t]| \quad (5)$$

$$\leq T_g^2 R_g^{max} \epsilon, \quad (6)$$

where we use the Taylor expansion in above derivation: When $\epsilon \xrightarrow{+} 0$, we can approximate $(1-\epsilon)^t$ by its first-order Taylor expansion $1 - \epsilon t$. Thus when $\lambda_g \xrightarrow{-} 1$, which is often the case in RL, $T_g^2 R_g^{max} \epsilon \xrightarrow{+} 0$, and $V_{g,\pi_{\theta_l}}(s)$ can be approximated as:

$$V_{g,\pi_{\theta_l}}(s) \approx E[q_{\pi_{\theta_l}}(s, g)] = E[-d_{\pi_{\theta_l}}(s, g)]. \quad (7)$$

It shows that the value iteration based on Bellman Equation $V_{g,\pi^*_{\theta_l}}(s_t) = R_g(s_t, a_t, g_t) + \gamma_g \mathbb{E}[V_{g,\pi^*_{\theta_l}}(s_{t+1})]|_{s_{t+1} \sim \mathcal{P}_{\pi^*_{\theta_l}}(\cdot|s_t, a_t)}$ implies $q_{\pi^*_{\theta_l}}(s_t, g_t) \approx R_g(s_t, a_t, g_t) + q_{\pi^*_{\theta_l}}(s_{t+1}, g_t)|_{(s_{t+1} \sim \mathcal{P}_{\pi^*_{\theta_l}}(\cdot|s_t, a_t)}$, where $\mathcal{P}_{\pi^*_{\theta_l}}$ is the transition probability of optimal policy $\pi^*_{\theta_l}$.

This relationship allows us to view the MDP as a directed graph. The nodes are S, and the edges are sampled according to the transition probability in the MDP. And we can view the UMDP as a temporal-abstracted graph built on top of an MDP, with nodes forming a low-dimensional manifold where distant nodes can only be reached by long paths.

Balance Exploration and Exploitation via UMDP Graphs

The above derivation shows that the general value iteration for RL problems is exactly the shortest path algorithm in terms of $d_{\pi_{\theta_l}}(s, g)$ on the graph. This structure allows us to estimate the shortest paths efficiently and accurately by designing measures upon nodes and edges of the graph.

Our novelty measure $N(\cdot)$ consists of two sums: the outside is to sample historical episodes from the replay buffer, and the inside calculates the visit counts of z_i to z_T by using historical episodes containing the cell in which z_i is currently located, where T is the episodic length. $N(z)$ is an increasing function of the visited times of z. z with a smaller $N(z)$ is more worth exploring.

$$N(\phi(z_i)) = \sum_{\mathcal{T} \in \mathcal{B}_l} \sum_{j=0}^{\lfloor (T-i)/c \rfloor} \gamma^j n(\phi(z_{i+jc})), \quad (8)$$

The utility measure $U(\cdot)$ uses methods like shortest-path planning to calculate values of candidate subgoals and then choose one with the max value, thus leading to better exploitation. $U(w_{i,j})$ is an increasing function of the expected

reward gain of $w_{i,j}$. z_j with a greater $U(w_{i,j})$ is more worth exploiting.

$$U(w_{i,j}) = \mathbb{E}_{s_k \in \mathcal{B}_l}[I(\varphi(s_k) = z_i)V(s_k, z_j)], \tag{9}$$

For remote nodes, we obtain their utilities by applying the Bellman-Ford (McQuillan and Walden 1977) algorithm:

$$\begin{aligned} U(w_{i,f}) &= \max_j [U(w_{i,j}) + U(w_{j,f})] \\ &= \max_{j,\dots,n} [U(w_{i,j}) + \sum_{k=j}^{n-1} U(w_{k,k+1}) + U(w_{n,f})], \end{aligned} \tag{10}$$

Universal Value Function Approximators

Universal Value Function Approximators (UVFAs) (Schaul et al. 2015) use neural networks to model $V(s, g) \approx V_{g,\pi^*}(s)$, where π^* is the optimal policy, and apply the Bellman equation to train $V(s, g)$ in a bootstrapping way. However, the agent can receive non-trivial rewards only when it can reach the subgoal, which usually challenges the learning process of UVFAs.

Local Value Function Approximators

We adopt Hindsight experience replay (HER) (Andrychowicz et al. 2017) to overcome the difficulties that UVFAs met. HER has the idea of "turning failure into success," replacing the original failed subgoal with an achieved one, giving the agent more feedback. Thus, the agent will gradually learn to reach distant goals even if the reward is sparse. The value approximators trained with HER give more reliable estimates between nearby subgoals and master the skill to achieve subgoals of increasing difficulty in a curriculum way. We define these approximators enabling locally accurate value estimation as local value function approximators.

References

Andrychowicz, M.; Wolski, F.; Ray, A.; Schneider, J.; Fong, R.; Welinder, P.; McGrew, B.; Tobin, J.; Pieter Abbeel, O.; and Zaremba, W. 2017. Hindsight experience replay. *Advances in neural information processing systems*, 30.

McQuillan, J. M.; and Walden, D. C. 1977. The ARPA network design decisions. *Computer Networks*, 1(5): 243–289.

Schaul, T.; Horgan, D.; Gregor, K.; and Silver, D. 2015. Universal value function approximators. In *International Conference on Machine Learning*, 1312–1320.