Stochastic Ensemble Policy Transfer Framework

Tian Chang

2021/1/28

Outline

- Background
- Motivation
- Algorithm
- Experiment

Background

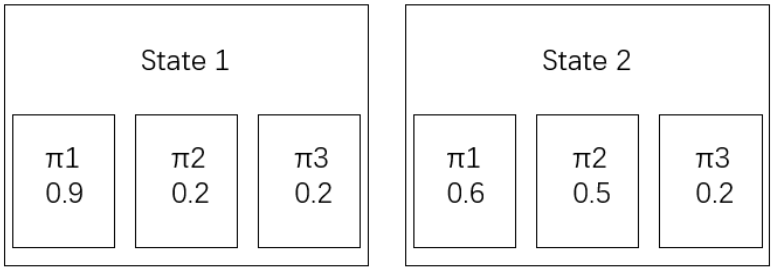
- **Problem Setting.** (Policy Transfer) A set of source tasks $M_1, M_2, \ldots, M_k$ are provided along with their expert (teacher) policies: $\pi_1, \pi_2, \ldots, \pi_k$. A student policy π_s for a target domain is learned by transferring knowledge from each π_i, with $1 \leq i \leq k$.

- **Policy Transfer Framework.** Modeling multi-policy transfer as an option learning problem，using ε-greedy to select option， Another option will be selected if the option is terminated according to the termination probability.

Motivation

- 1.The estimation of option value may be imperfect. in sufficiently complex environments, the value estimation will almost always be imperfect. As a result, agent will choose an unsuitable policy, which can lead to negative transfer.

- 2. Using ε-greedy to select option does not take advantage of value information, for example



SEPTF

- To solve these problem, we propose interpolating all of the source policies to produce a teacher policy which is better than any individual.

- The teacher policy is ensembled by weighting the candidate policies in a way that balances errors in the learned Q-function and value of policies.

Algorithm

- The idea comes from *Sample-Efficient Reinforcement Learning with Stochastic Ensemble Value Expansion*.(NeurIPS2018)
- 作者指出MVE算法特别依赖于调整展开的步数H来获得较好的效果，由于复杂的环境中展开过长的步数H的话，会引入很多误差，而在简单的环境中，展开过短的步数H的话，则会降低预估的精度。所以，作者提出了在不同环境中直接展开特定的步数H，并通过计算每一步的uncertainty来对 (H+1)个MVE的权重进行动态插值，以从模型中得到一个更好的Q值估计。

Algorithm

The key point is to balances errors in the learned Q-function and value of policies!

- 1. Estimate uncertainty of source policy

$$Q = \{Q_1, \ldots, Q_N\}$$

Each parameterization is trained on different subsets of the data in replay buffer.

- 2. Compute the variance for each policy $Q_{\pi_i}^{\sigma^2}$.
- 3.Ensemble teacher policy $\pi_{teacher} = \sum_{i=0}^{j} \frac{\widetilde{w}_i}{\sum_j \widetilde{w}_j} \pi_i$

Algorithm

4.Two parts to minimize: MSE of estimated value; difference between $Q^{Teacher}$ and Q^{Max}

$$\mathbb{E}\left[\left(\sum_{i=0}^{j} w_i Q(s,\pi_i) - Q(s,\pi_t)\right)^2\right] = Bias\left(\sum_i w_i Q(s,\pi_i)\right)^2 + Var\left(\sum_i w_i Q(s,\pi_i)\right)$$

$$\approx Bias\left(\sum_i w_i Q(s,\pi_i)\right)^2 + \sum_i w_i^2 Var\left(Q(s,\pi_i)\right)$$

$$\mathbb{E}\left[\left(Q(s,\pi_t) - Q^{Max}(s,\pi)\right)^2\right] \approx \left(\sum_{i=0}^{j} w_i Q(s,\pi_i) - \max_i\left(Q(s,\pi_i)\right)\right)^2$$

Minimize part1 + balancing factor * part2:

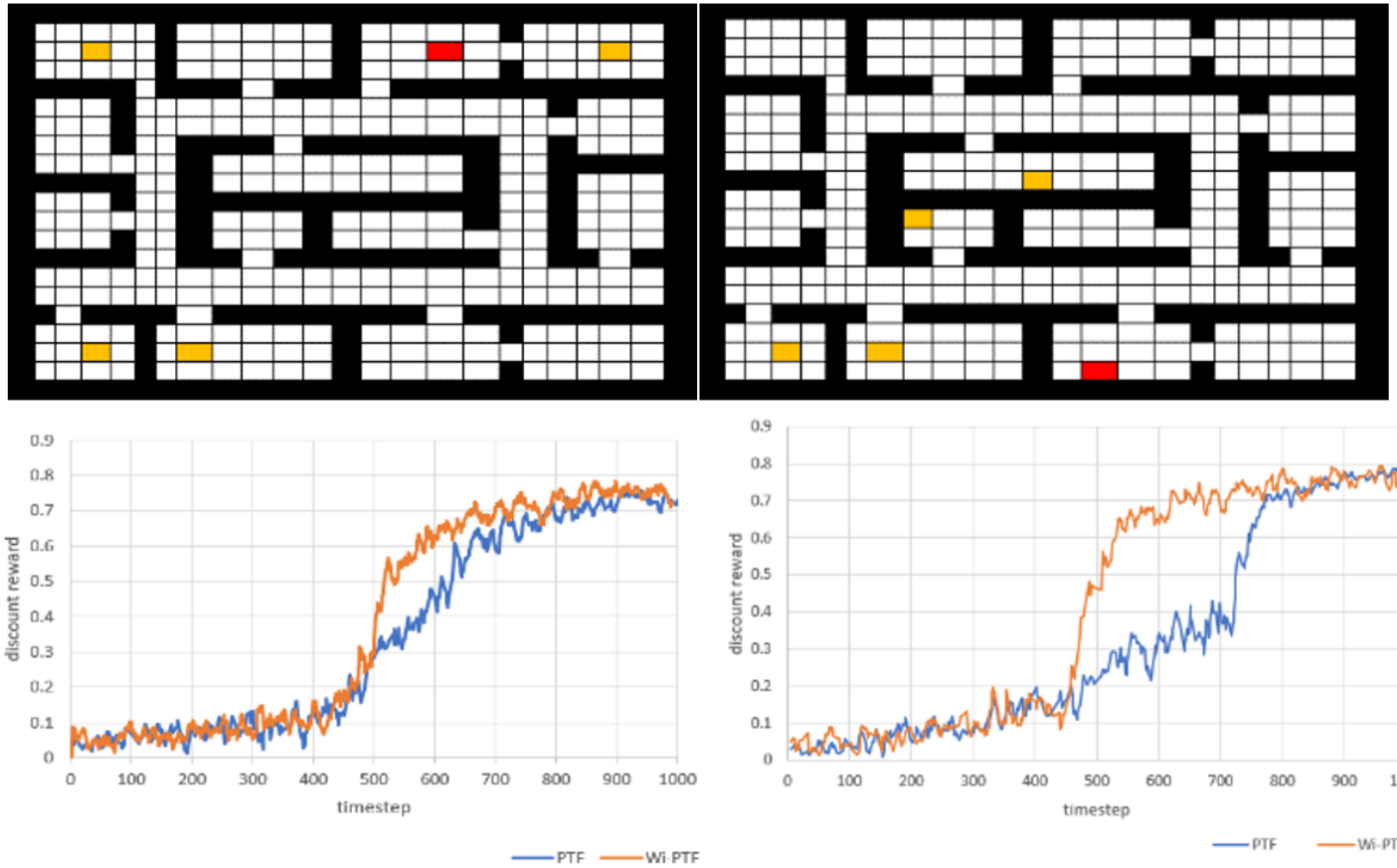
$$\begin{cases} min \sum_i w_i^2 Var\left(Q(s,\pi_i)\right) + b\left(\sum_{i=0}^{j} w_i Q(s,\pi_i) - \max_i\left(Q(s,\pi_i)\right)\right)^2 \\ s.t. \sum_{i=0}^{j} w_i = 1 \end{cases}$$

Experiment

- 下面为Weight-PTF的实验结果
- Weight-PTF:将源策略线性组合成教师策略，权重为源策略的终止概率做softmax
- Weight-PTF主要意图是利用次优策略的知识，没有衡量策略价值估计的不确定性

Experiment

- 实验证明，集成策略确实可以比基于选项的单策略更有效果



Conclusion

- 在策略迁移中，对源策略的价值估计不准会导致agent选择错误的策略进行迁移。
- SEPTF的目的：通过策略集成，在最小化选择策略的不准确性和最大化选择策略的价值期望中取得平衡
- 之前实验证明集成策略可以比选择单策略的PTF表现更好

http://lamda.nju.edu.cn

Thanks !