

Making Linear MDPs Practical via Contrastive Representation Learning

Tianjun Zhang¹ Tongzheng Ren^{2, 3} Mengjiao Yang^{1, 3}
Joseph E. Gonzalez¹ Dale Schuurmans^{3, 4} Bo Dai³

¹UC Berkeley

²UT Austin

³Google Brain

⁴University of Alberta

ICML 2022



Motivation

- Linear MDPs provide an elegant framework for function approximation in reinforcement learning, which assumes that

$$P(s'|s, a) = \langle \phi(s, a), \mu(s') \rangle, \quad r(s, a) = \langle \phi(s, a), \theta_r \rangle.$$

- Easy to perform planning, as

$$\begin{aligned} Q_{P,r}^{\pi}(s, a) &= r(s, a) + \gamma \int V_{P,r}^{\pi}(s) P(s'|s, a) ds' \\ &= \left\langle \phi(s, a), \theta_r + \gamma \int V_{P,r}^{\pi}(s') \mu(s') ds' \right\rangle. \end{aligned}$$

- How to obtain $\phi(s, a)$, from a practical perspective?

Difficulties

For the normalization conditions:

- We requires

$$\left\| \int_{\mathcal{S}} \mu(s') g(s') ds' \right\|_2 \leq \sqrt{d}, \quad \forall \|g\|_{\infty} \leq 1$$

to ensure the linear weight of Q -function has bounded norm.

- However, it's an integral constraint with indefinite g , that cannot be embedded to the learning procedure.

Difficulties

For representation learning:

- A natural idea is to use Maximum Likelihood Estimation (MLE), as discussed in [UZS21].
- However, we have the constraints on marginal regularity, i.e.

$$\left\langle \phi(s, a), \int \mu(s') ds' \right\rangle = 1, \quad \forall (s, a)$$

- As we can have infinite state-action pairs, such constraints cannot be embedded to the learning procedure as well.

Proposed Solutions

For the normalization conditions:

- We modify the original linear MDP to

$$P(s'|s, a) = \langle \phi(s, a), p(s')\mu(s') \rangle,$$

where p is a probability measure supported on the whole state space.

- For $\|g\|_\infty \leq 1$, as

$$\begin{aligned} \left\| \int_{\mathcal{S}} p(s')\mu(s')'g(s')ds' \right\|_2^2 &\leq \int_{\mathcal{S}} p(s')\|\mu(s')g(s')\|_2^2 ds' \\ &\leq \int_{\mathcal{S}} p(s')\|\mu(s')\|_2^2 ds', \end{aligned}$$

we can introduce an additional regularizer $\mathbb{E}_p\|\mu(s')\|^2$ to the objective, which can be approximated by Monte Carlo estimation.

Proposed Solutions

For representation learning:

- We use noise contrastive estimation (NCE) [MC18] to estimate ϕ , whose objective can be written as:

$$\frac{1}{n} \sum_{i=1}^n \log \frac{\frac{\langle \phi(s_i, a_i), p(s'_i) \mu(s'_i) \rangle}{q(s'_i)}}{\frac{\langle \phi(s_i, a_i), p(s'_i) \mu(s'_i) \rangle}{q(s'_i)} + \sum_{j=1}^K \frac{\langle \phi(s_i, a_i), p(s'_{i,j}) \mu(s'_{i,j}) \rangle}{q(s'_{i,j})}},$$

where $s_{i,j} \sim q$ are the negative samples and q is a probability measure supported on the whole state space.

- We can choose q as p .
- To enforce the normalizing constant to be 1, we can also add the additional regularizer $\mathbb{E}_{(s,a)} \left[\left(\log \int_{\mathcal{S}} \phi(s, a)^\top \mu(s') p(s') ds' \right)^2 \right]$, that can be approximated with Monte Carlo estimation.

Full Algorithm

Algorithm 1 CTRL-UCB: Online Exploration with Representation Learning

Input: Regularizer λ_n , parameter α_n , Model class $\mathcal{M} = \{(\mu, \phi) : \mu \in \Psi, \phi \in \Phi\}$, Iteration N , Number of Negative Samples K .

Initialize $\pi_0(\cdot | s)$ to be uniform; set $\mathcal{D} = \emptyset$

for episode $n = 1, \dots, N$ **do**

Collect a transition (s, a, s') where $s \sim d_{P^*}^{\pi_{n-1}}$, $a \sim U(\mathcal{A})$, $s' \sim P^*(\cdot | s, a)$.

$\mathcal{D}_n = \mathcal{D}_{n-1} \cup \{s, a, s'\}$

Learn representation $\hat{\phi}(s, a)$ via NCE.

Update the empirical covariance matrix

$$\hat{\Sigma}_n = \sum_{s, a \in \mathcal{D}_n} \hat{\phi}_n(s, a) \hat{\phi}_n(s, a)^\top + \lambda_n I$$

Set the exploration bonus:

$$\hat{b}_n(s, a) = \min \left(\alpha_n \sqrt{\hat{\phi}_n(s, a)^\top \hat{\Sigma}_n^{-1} \hat{\phi}_n(s, a)}, 2 \right)$$

Update policy $\pi_n = \arg \max_{\pi} V_{\hat{P}_n, r + \hat{b}_n}^{\pi}$

end for

Return π_1, \dots, π_N

Theoretical Justification

- Our algorithm motivates from [UZS21].
- We show that as $K \rightarrow \infty$, the estimator obtained with NCE is identical to the estimator obtained by MLE, which shows the asymptotic sample efficiency as $K \rightarrow \infty$.

Experiment Results

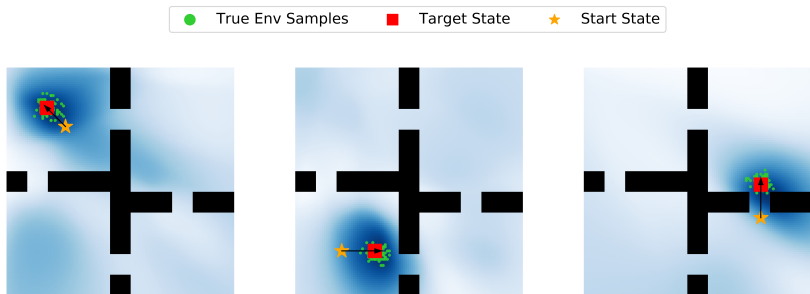


Figure: Heatmap of the Learned Transition: The agent transits from orange star to red square and the green dots are the samples drawn from environment simulation.

Experimental Results

Table: Performance of on various Deepmind Suite Control tasks. All the results are averaged across four random seeds and a window size of 10K

		cheetah_run	cheetah_run_sparse	walker_run	walker_run_sparse	humanoid_run	hopper_hop
Model-Free RL	PPO	227.7±57.9	5.4±10.8	51.6±1.5	0.0±0.0	1.1±0.0	0.7±0.8
	SAC (2-layer)	222.2±41.0	32.4±27.8	183.0±23.4	53.5±69.3	1.3±0.1	0.4±0.5
	SAC (3-layer)	595.2±96.0	419.5±73.3	700.9±36.6	311.5±361.4	1.2±0.1	28.6±19.5
Representation RL	DeepSF	295.3±43.5	0.0±0.0	27.9±2.2	0.1±0.1	0.9±0.1	0.3±0.1
	CTRL-UCBM	679.0±40.8	446.2±146.3	743.1±53.0	697.0±44.8	11.5±5.4	161.9±76.1

Conclusion

- We investigate the difficulties of practical linear MDP, and provide solutions for these difficulties.
- We show that NCE can be used for the representation learning in reinforcement learning, and prove the asymptotic efficiency of our algorithm.
- We show superior performance on several benchmarks, which demonstrates our effectiveness.

Reference I

- [MC18] Zhuang Ma and Michael Collins. Noise contrastive estimation and negative sampling for conditional models: Consistency and statistical efficiency. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3698–3707, 2018.
- [UZS21] Masatoshi Uehara, Xuezhou Zhang, and Wen Sun. Representation learning for online and offline rl in low-rank mdps. arXiv preprint arXiv:2110.04652, 2021.