FLAMBE: Structural complexity and representation learning of low rank MDPs

Alekh Agarwal, Sham M. Kakade,
Akshay Krishnamurthy, Wen Sun

Microsoft Research

NEURAL INFORMATION PROCESSING SYSTEMS

Goal: Sample efficient exploration in RL

How can we get reinforcement learning algorithms to explore efficiently when operating in complex environments? Such algorithms would be immensely useful in scaling RL into high stakes scenarios where sample-efficiency is a primary concern.

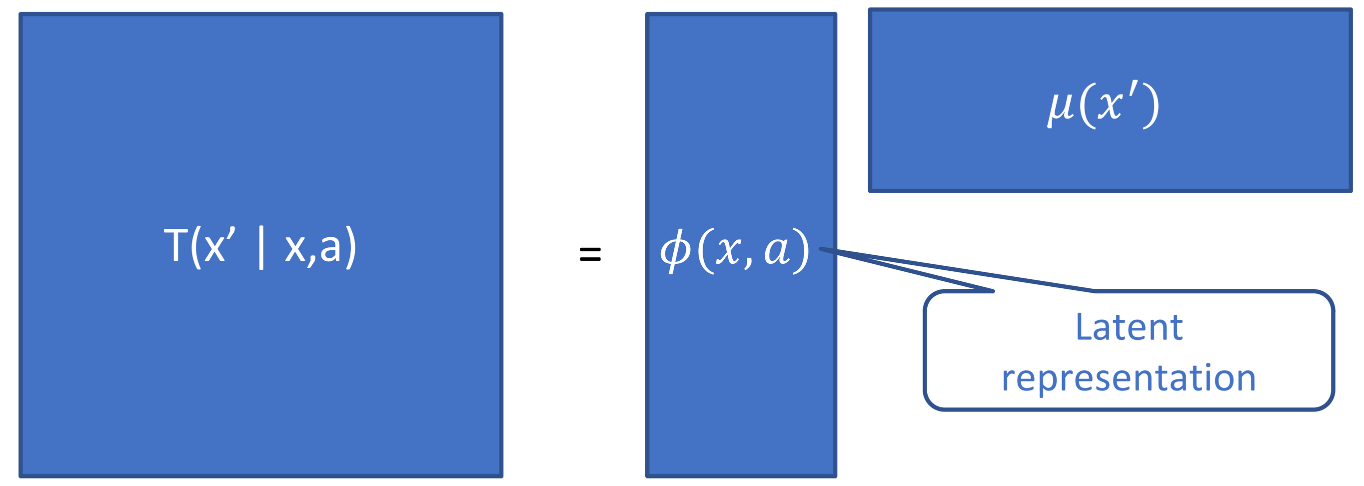
A possible solution is through representation learning, where we discover some simple underlying structure that enables us to efficiently explore and approximate the key quantities, such as value functions and policies.

Key question:

1. **What does it mean to have a good representation?**
2. **How do we learn one in a sample-efficient manner, while exploring?**

We answer these questions in the context of low rank MDPs

The low rank MDP



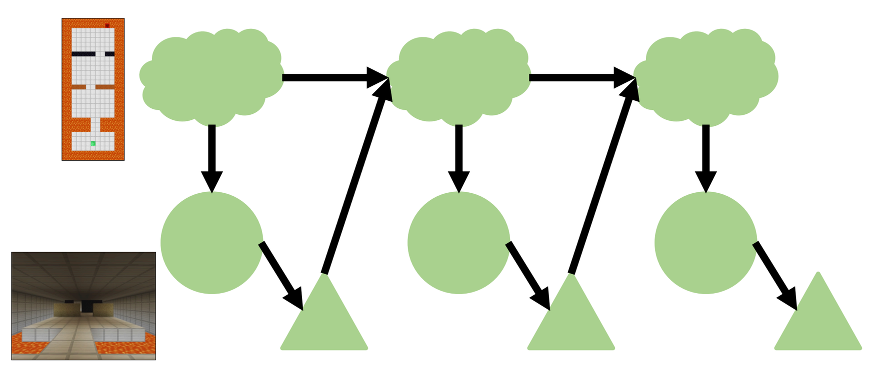
Transition operator T admits a low rank factorization into feature maps ϕ, μ.

1. Low rank MDP is algorithmically tractable if ϕ is known [JYWJ19]
2. Low rank MDP is statistically tractable with unknown ϕ [JKALS17]

Representation learning in low rank MDPs: Discover ϕ.
Efficient algorithms?

Structural results

Proposition: A block MDP is low rank with $d = |\mathcal{S}|$. However, there exist low rank MDPs of embedding dimension 2 that admit no non-trivial block MDP representation



A stochastic factorization, where $\phi(x,a)$ is on the simplex, has a natural interpretation as a (fully observable) latent variable model. However:
Proposition: There exists low rank MDPs of rank d for which the stochastic factorization has dimension $2^{\Omega(\sqrt{d})}$.

Main result

Assume access to function class Φ, Υ such that $\phi \in \Phi, \mu \in \Upsilon$
Assume computational oracle for optimizing and sampling from Φ, Υ

Theorem [AKKS20]: FLAMBE learns a low rank MDP model such that

$$\forall\ \pi, h: \quad \mathbb{E}_\pi \left\| \langle \hat{\phi}_h(x_h, a_h), \hat{\mu}_h(\cdot) \rangle - T_h(\cdot \mid x_h, a_h) \right\|_{TV} \le \varepsilon$$

With sample complexity:

$$poly(d, |A|, H, \tfrac{1}{\varepsilon}, \log(|\Phi||\Upsilon|/\delta))$$

FLAMBE runs in polynomial time in oracle model.

FLAMBE

Algorithm 1 FLAMBE: Feature Learning And Model-Based Exploration

Input: Environment $\mathcal{M}$, function classes Φ, Υ, subroutines MLE and SAMP, parameters β, n.
Set ρ_0 to be the random policy, which takes all actions uniformly at random.
Set $D_h = \emptyset$ for each $h \in \{0, \ldots, H-1\}$.
for $j = 1, \ldots, J_{\max}$ **do**
  for $h = 0, \ldots, H-1$ **do**
    Collect n samples (x_h, a_h, x_{h+1}) by rolling into x_h with ρ_{j-1} and taking $a_h \sim \text{unif}(\mathcal{A})$.
    Add these samples to D_h.
    Solve maximum likelihood problem: $(\hat{\phi}_h, \hat{\mu}_h) \leftarrow \text{MLE}(D_h)$.
    Set $\hat{T}_h(x_{h+1} \mid x_h, a_h) = \left\langle \hat{\phi}_h(x_h, a_h), \hat{\mu}_h(x_{h+1}) \right\rangle$.
  end for
  For each h, call planner (Algorithm 2) with h step model $\hat{T}_{0:h-1}$ and β to obtain ρ_h^{pre}.
  Set $\rho_j = \text{unif}(\{\rho_h^{\text{pre}} \circ \text{random}\}_{h=0}^{H-1})$, to be uniform over the discovered h-step policies, augmented with random actions.
end for

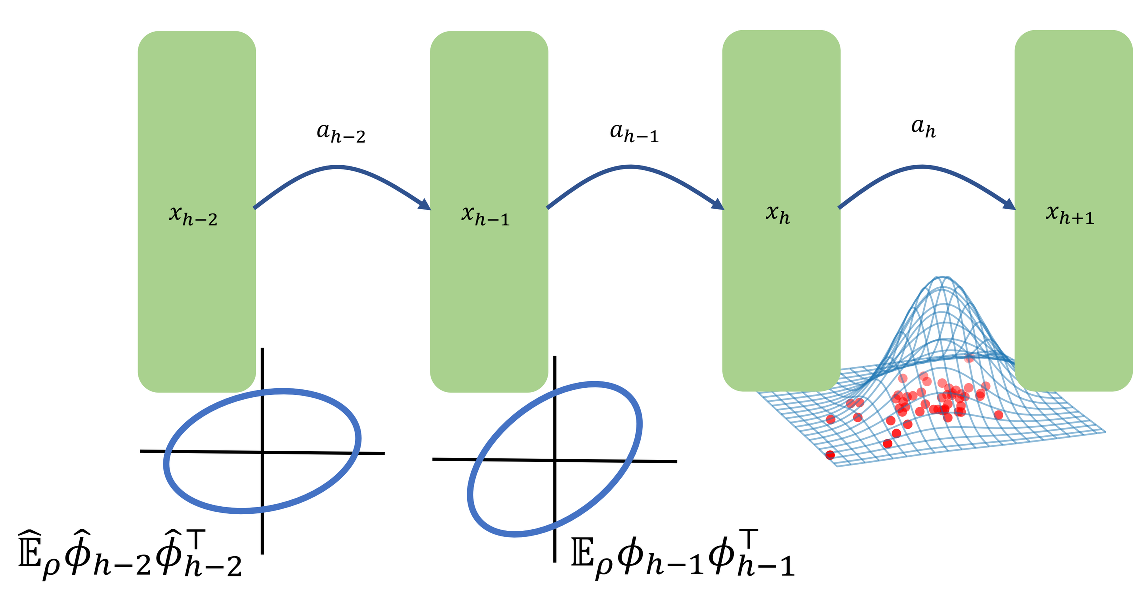
Algorithm 2 Elliptical planner

Input: MDP $\widetilde{\mathcal{M}} = (\phi_{0:\tilde{h}}, \mu_{0:\tilde{h}})$, subroutine SAMP, parameter $\beta > 0$. Initialize $\Sigma_0 = I_{d\times d}$.
for $t = 1, 2, \ldots,$ **do**
  Compute (see text for details)

$$\pi_t = \operatorname*{argmax}_\pi \mathbb{E}\left[\phi_{\tilde{h}}(x_{\tilde{h}}, a_{\tilde{h}})^\top \Sigma_{t-1}^{-1} \phi_{\tilde{h}}(x_{\tilde{h}}, a_{\tilde{h}}) \mid \pi, \widetilde{\mathcal{M}}\right]. \tag{2}$$

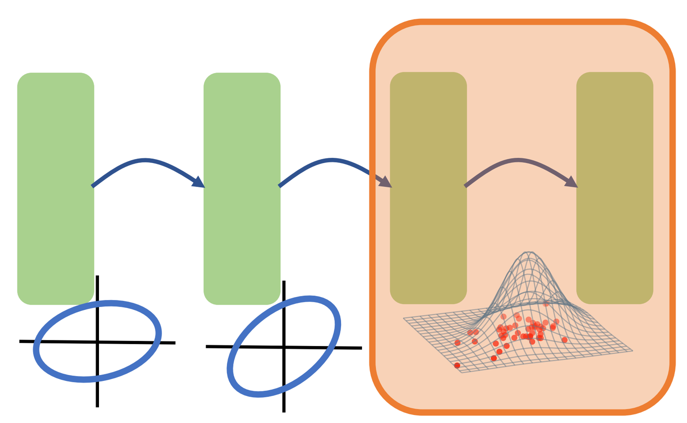
  If the objective is at most β, halt and output $\rho = \text{unif}(\{\pi_\tau\}_{\tau<t})$.
  Compute $\Sigma_{\pi_t} = \mathbb{E}\left[\phi_{\tilde{h}}(x_{\tilde{h}}, a_{\tilde{h}})\phi_{\tilde{h}}(x_{\tilde{h}}, a_{\tilde{h}})^\top \mid \pi, \widetilde{\mathcal{M}}\right]$. Update $\Sigma_t \leftarrow \Sigma_{t-1} + \Sigma_{\pi_t}$.
end for

Proof overview



Three key questions:

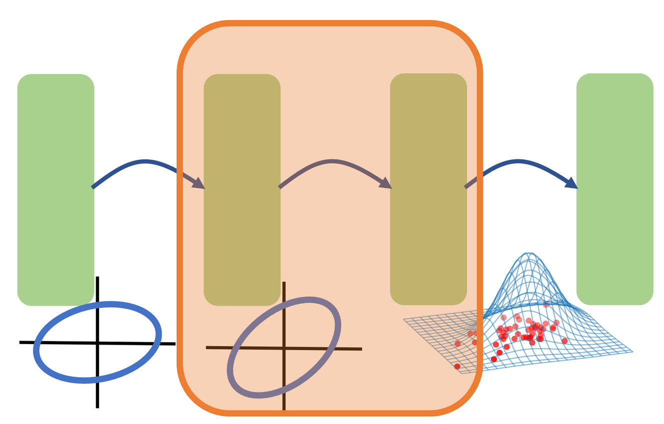
1. How to learn the dynamics?
2. How to measure coverage?
3. How to compute an exploratory policy to optimize coverage?



Learn the model with maximum likelihood:

$$\hat{\phi}_h, \hat{\mu}_h = \text{argmax}_{\phi\in\Phi, \mu\in\Upsilon} \sum_{x_h, a_h, x_{h+1}} \log \langle \phi(x_h, a_h), \mu(x_{h+1}) \rangle$$

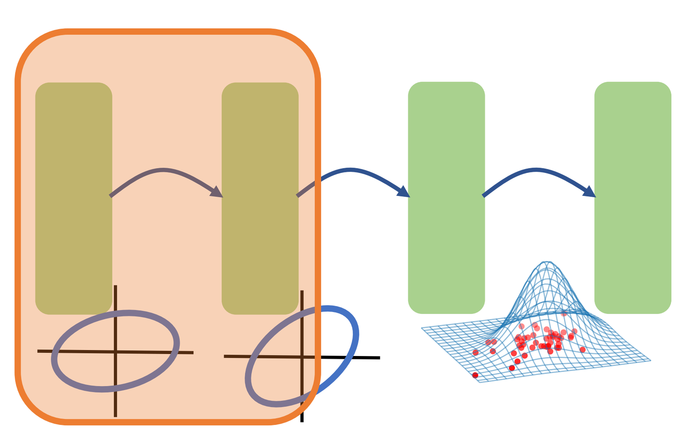
Can guarantee accuracy in TV-distance on training distribution [Z07]



Measure coverage using second moment of features at the previous time:

$$\mathbb{E}_\pi f(x_h) = \langle \mathbb{E}_\pi \phi(x_{h-1}, a_{h-1}), \textstyle\int \mu(x_h) f(x_h) \rangle$$

Addresses distribution shift with potential function



Optimize coverage by planning to visit all directions of *learned feature two steps behind!*

ρ guarantees: $\max_\pi \mathbb{E}_\pi[\hat{\phi}_{h-2}\, \hat{\Sigma}_\rho^{-1}\, \hat{\phi}_{h-2} \mid \hat{M}] \le O(d)$

By TV guarantee, ρ approximately visits all directions in the environment

Corollaries, discussion, references

Corollaries

1. For any reward, near-optimal policy and Q function are linear in $\hat{\phi}_{1:H}$
2. Can optimize any reward function with no additional experience
3. Simpler planner for stochastic factorization, with a much better sample complexity.

Discussion

1. Provable RL with general non-linear function approximation
2. Suggestions for practice: reward bonuses, model architecture, etc.
3. Future work: does it work in practice?

References

[JKALS17] Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, Robert E. Schapire. Contextual decision processes with low Bellman rank are PAC learnable. ICML, 2017.

[JYWJ20] Chi Jin, Zhuoran Yang, Zhaoran Wang, Michael I. Jordan. Provably efficient reinforcement learning with linear function approximation. COLT, 2020.

[Z07] Tong Zhang. From ϵ-entropy to KL-entropy: Analysis of minimum information complexity density estimation. Annals of Statistics, 2007.

https://arxiv.org/abs/2006.10814