Experiments with Natural Policy Gradient for Low Rank MDP with log-linear Parametrization

Sihui Wang

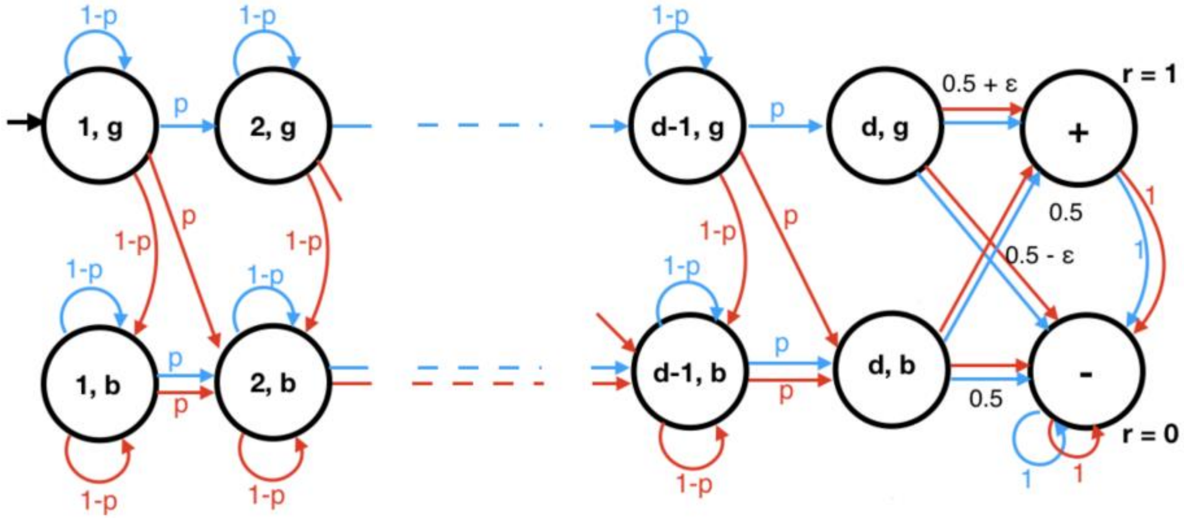
Dec 6th, 2022

Motivation

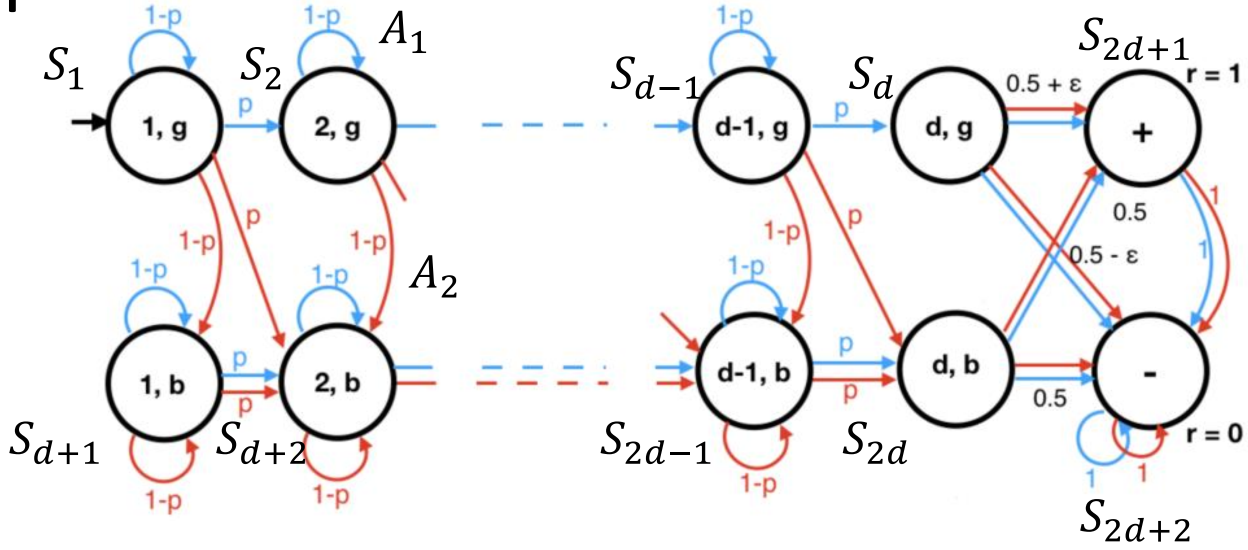
- For tabular softmax parametrization, NPG can obtain linear convergence rate with geometrically increasing step-size
- New findings suggest that, for log-linear policy parametrization, NPG can also obtain linear rate with geometrically increasing step-size, given that the environment has a low-rank structure:
- $\phi: \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d, \mathbb{E}_{s,a\sim v}[w^T\phi(s,a) - Q^{\pi_\theta}(s,a)] \leq \epsilon_{bias}$
- Our Goals:
  - Implement MDP with Low Rank Structure
  - Implement NPG with different step-size
  - Analyze Convergence Rate

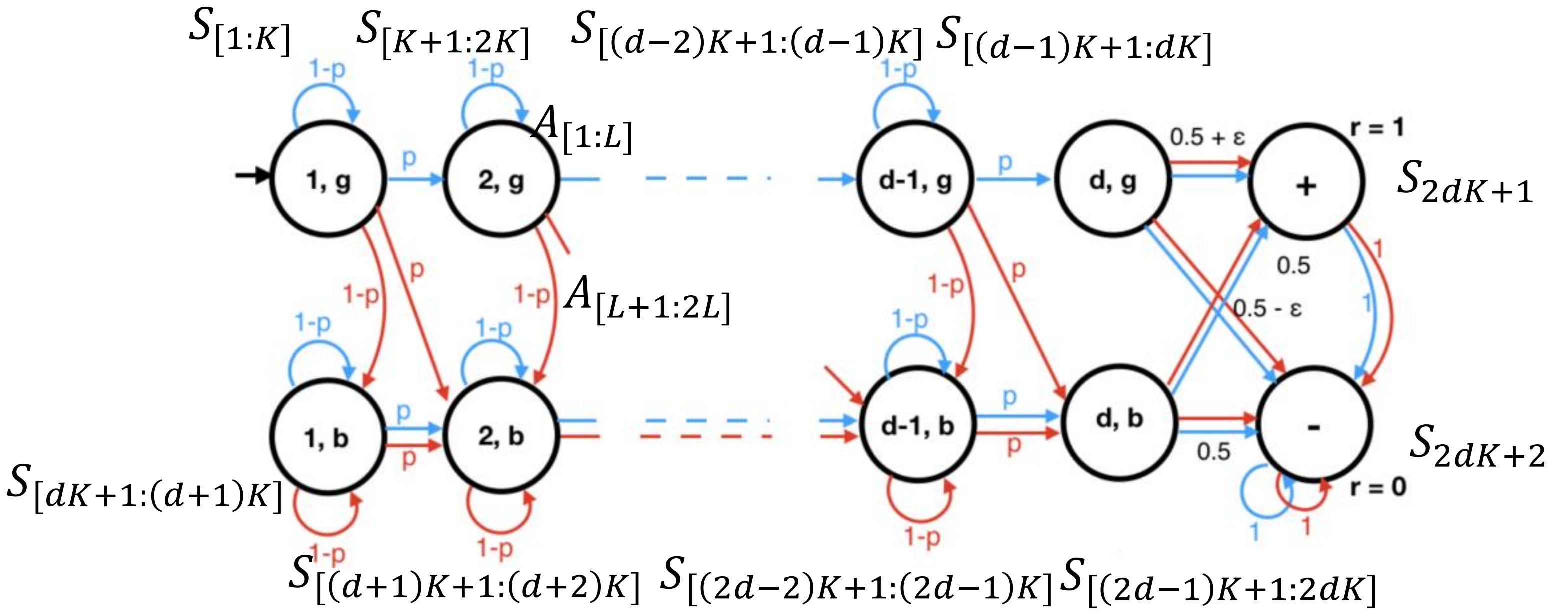
MDP: Latent State Construction

- Follow optimal actions (blue arrows), the agent can remain in good states (i,g) and eventually get reward 1 with $\frac{1}{2} + \epsilon$ probability.
- Once the agent chooses sub-optimal actions and transits to bad states (i,b), the following action choices make no difference, and the agent will get reward 1 with $\frac{1}{2}$ probability.



MDP: Duplicate States and Actions





Replace entries in the MDP matrix by "blocks"

Transition probabilities among blocks remain the same

The probabilities are set as:

$$p \cdot DirichletDist$$

or:

$$(1-p) \cdot DirichletDist$$

Hence, we create an $(2Kd+2) \times 2L$ dimensional MDP from $(2d+2) \times 2$ dimensional latent space

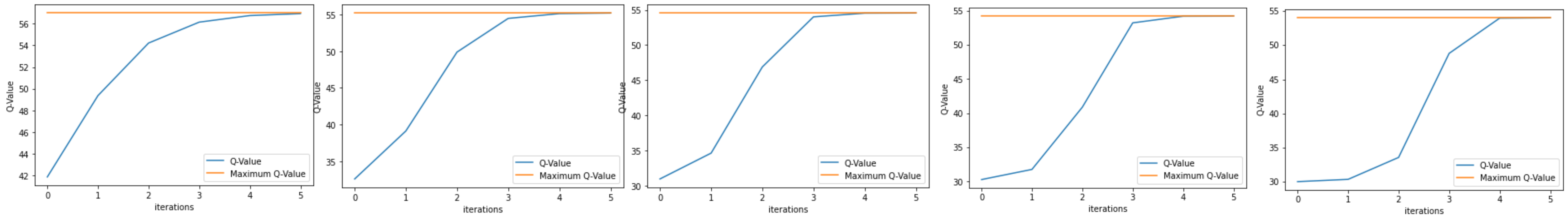
Choose Features

- Assumption: Q-function should be approximated by linear combinations of the features:
$$\phi : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}^d, \mathbb{E}_{s,a \sim v}[w^T \phi(s,a) - Q^{\pi_\theta}(s,a)] \leq \epsilon_{bias}$$
- $(2d + 2) \times 2$ dimensional Feature:
  - For S, A, find the corresponding s, a in latent space. Set $\phi_{s,a} = 1$ and set all other components as 0: $Q^t \phi = Q^t$
- 1-dimensional Feature:
  - Tentatively set $\phi(S, A) = Q^{\pi^*}(s, a)$

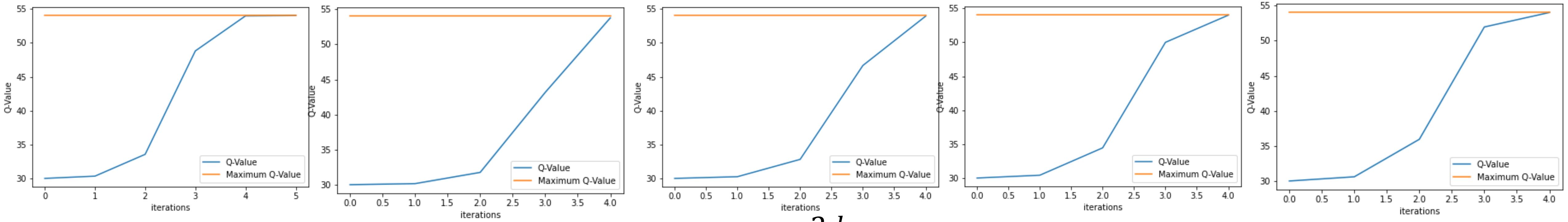
Calculation of Fisher Information Matrix

- $\pi_\theta(a|s) = \frac{e^{\theta \cdot \phi_{s,a}}}{\sum_{a' \in \mathcal{A}} e^{\theta \cdot \phi_{s,a'}}}$
- $\log(\pi_\theta(a|s)) = \theta \cdot \phi_{s,a} - \log\left(\sum_{a' \in \mathcal{A}} e^{\theta \cdot \phi_{s,a'}}\right)$
- $\nabla_\theta \left(log\left(\pi_\theta(a|s)\right)\right) = \phi_{s,a} - \sum_{a' \in \mathcal{A}} \frac{e^{\theta \cdot \phi_{s,a'}}}{\sum_{a' \in \mathcal{A}} e^{\theta \cdot \phi_{s,a'}}} \cdot \phi_{s,a'} \coloneqq \Phi(s, a, \theta)$
- $F = \mathbb{E}_{\substack{s \sim d_\rho^{\pi_\theta} \\ a \sim \pi_{\theta(\cdot|s)}}} \left[\nabla_\theta \left(log\left(\pi_\theta(a|s)\right)\right) \nabla_\theta^T \left(log\left(\pi_\theta(a|s)\right)\right)\right] =$
  $\sum_{s,a} d_\rho^{\pi_\theta}(s) \cdot \pi_\theta(a|s) \cdot [\Phi(s, a, \theta)\Phi^T(s, a, \theta)]$

Experiment: $s \times a$ dimensional features

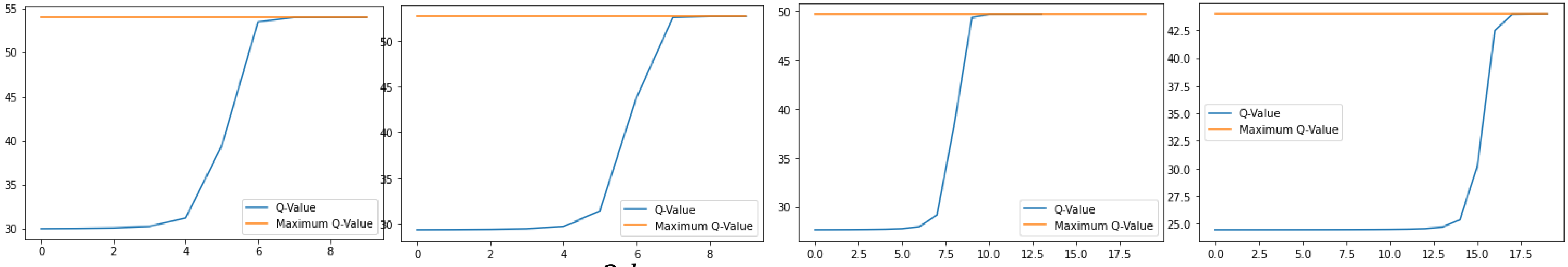


$$p = 0.95, K = 1, L = 1, R = 100.0, \epsilon = 0.4, \gamma = \frac{2d}{2d+1}, d = 2,4,6,8,10$$

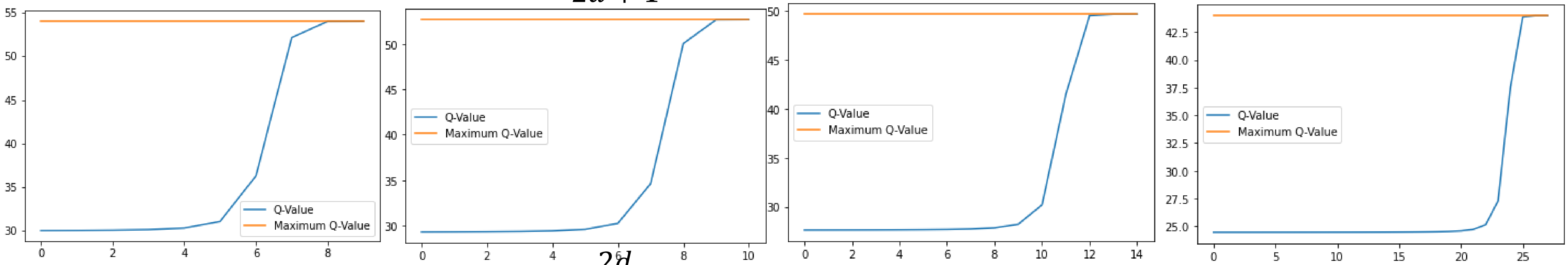


$$p = 0.95, R = 100.0, \epsilon = 0.4, \gamma = \frac{2d}{2d+1}, d = 10, K = L = 1,2,4,8,10$$

Experiment: 1 dimensional features



$$R = 100.0, \epsilon = 0.4, \gamma = \frac{2d}{2d+1}, d = 10, K = L = 10, p = 0.95, 0.90, 0.80, 0.65$$



$$R = 100.0, \epsilon = 0.4, \gamma = \frac{2d}{2d+1}, d = 10, K = L = 10, p = 0.95, 0.90, 0.80, 0.65$$

with suboptimal features

Summary

- Duplication of states and actions doesn't affect convergence rates
- For $s \times a$ dimensional features, as d increases, we don't observe that "flat gradient" should slow down convergence
- For inexact 1-dimensional features, as MDP becomes indeterministic, convergence tends to slow down
- For 1-dimensional features, sub-optimal features lead to slower convergence
- For 1-dimensional features, the convergence is slow at first, then it suddenly accelerates

Future Directions

- Experiment with different step-size
- Experiment with randomized features
- Experiment with good features with some stochasticity