Investigating Action-Space Generalization in RL for Recommendation Systems

Abhishek Naik Bo Chang Alexandros Karatzoglou Martin Mladenov Ed H. Chi Minmin Chen

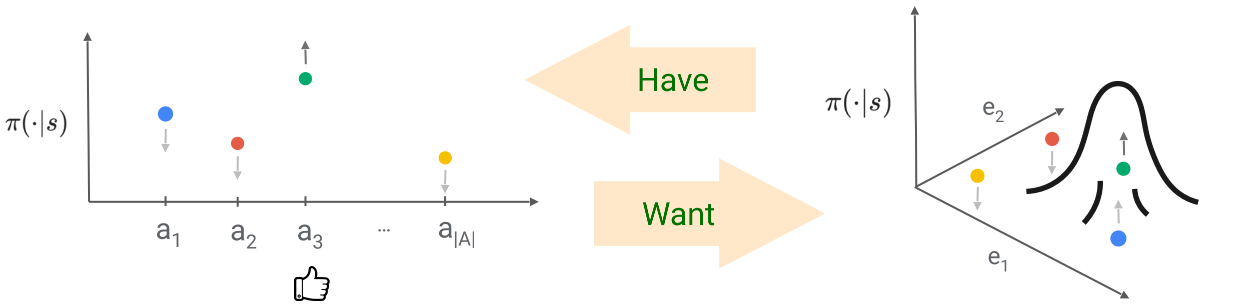
abhishek.naik@ualberta.ca bochang@google.com

UNIVERSITY OF ALBERTA, EDMONTON·ALBERTA·CANADA — Google

- Action-space generalization methods learn quickly compared to the baseline when there are very few interactions per item.
- The Gaussian parameterization for action-space generalization appears more robust to item/action features than the softmax parameterization.

	Softmax	Gaussian
Policy	$\pi_\theta(a) = \dfrac{\exp h_\theta(e(a))}{\sum_b \exp h_\theta(e(b))}$ $\quad h_\theta(e(a)) \doteq \theta^\top e(a)$	$\pi_\theta(a) \doteq \mathcal{N}'(\mu_\theta, \sigma_\theta^2)$ Biased! $\quad \mu_\theta = \theta_\mu \quad \sigma_\theta = \exp[\theta_\sigma]$
Gradient	$\nabla_\theta \ln \pi_\theta(a) = e(a) - \sum_b \pi_\theta(b) e(b)$ Costly!	$\nabla_{\theta_\mu} \ln \pi_\theta(a) = \dfrac{1}{\sigma_\theta^2}(e(a) - \mu_\theta)$ $\nabla_{\theta_\sigma} \ln \pi_\theta(a) = \left(\dfrac{(e(a) - \mu_\theta)^2}{\sigma_\theta^2} - 1\right)$
Query	1. Compute the softmax 2. Sample a discrete action O(|A|) logit computations	1. Sample action features 2. Find a nearby discrete action O(|A|) distance computations

What is Action-Space Generalization?



Check out the paper

Methods

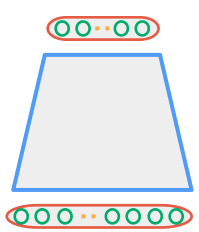
Base RL method: Policy Gradient

$$\theta_{t+1} \doteq \theta_t + \alpha \delta_t \nabla_{\theta_t} \ln \pi_{\theta_t}(A_t | S_t)$$

Softmax parameterization

$$\pi_\theta(a|s) = \frac{\exp h_\theta(s,a)}{\sum_b \exp h_\theta(s,b)}$$

$$h_\theta(s,a) = f(x(s))^\top g(e(a))$$

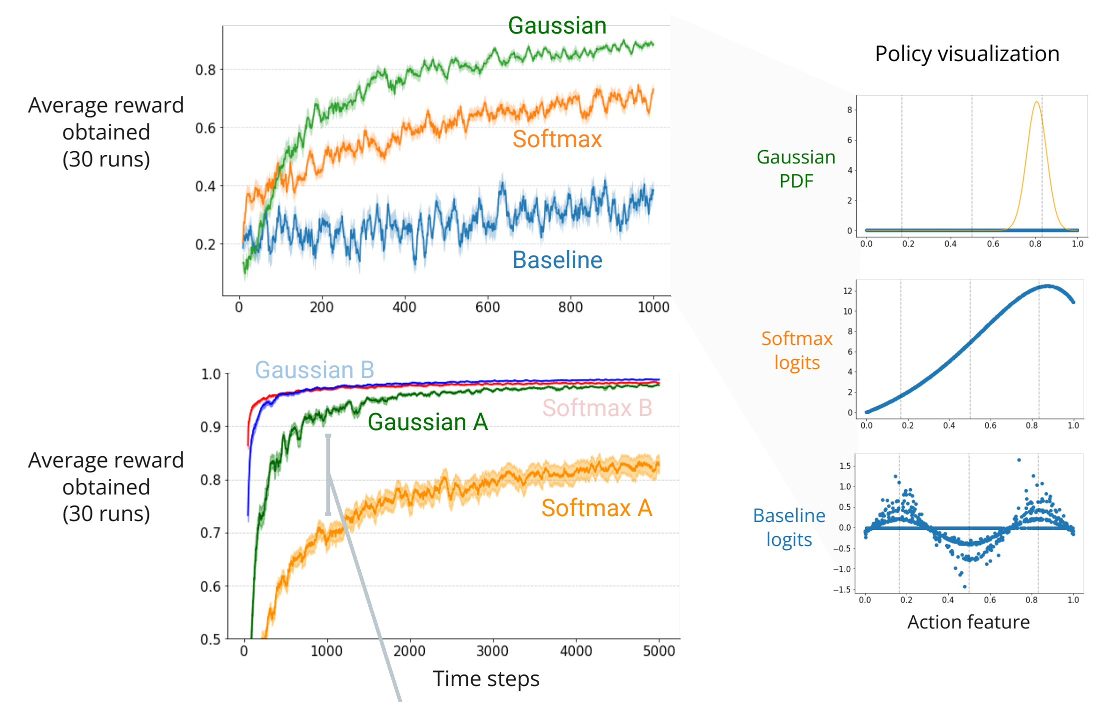


Gaussian parameterization

$$\pi_\theta(a|s) = \mathcal{N}'(\mu_\theta(s), \sigma_\theta^2(s))$$

$$\doteq \int_{\mathcal{M}(a)} \frac{1}{\sqrt{(2\pi)^k |\sigma_\theta|}} \exp\left[-\frac{1}{2}(e-\mu_\theta)^\top \sigma_\theta^{-1}(e-\mu_\theta)\right] de$$

Preliminary Empirical Results



The Gaussian parameterization is promising in real-world settings when features are learned and not perfect.

This work was done during AN's internship at Google Research, which all the other co-authors are affiliated to.