In-Context RL

August 2023

1 Data Generation

1.1 5-Armed Bandit Problem

State space $S = \{0, 1, 2, 3, 4\}$; action space $A = \{0, 1, 2, 3, 4\}$

```
Initialize empty pretraining dataset 𝓑
for i in [N] do
   p_1 ~ Dirchlet distribution(𝟙)
   p_2 ~ point-mass distribution
   ω ~ Unif(0.1[10])
   action distribution p = (1 − ω)p_1 + ωp_2
   action means μ ~ Unif[0, 1]^5

   for h in [H = 500] do
      action a_h ~ p (OHE)
      reward r_h ~ N(μ_a, σ^2) where σ = 0.3
      append (a_h, r_h) to goal g
      append a_h, r_h to 𝓑
```

2 Value Function Approximation

for i in [#iterations] do
  sample offline data $\{s_t^i, a_t^i, s_{t+1}^i, g_t^i\}_{i=1}^N \sim \mathcal{B}$, $\{s_0^i\}_{i=1}^M \sim \mu_0$
  estimate the reward function $\hat{R}$ for each task in the current offline data
  Value objective: $L_V(\theta) = \frac{1-\gamma}{M} \sum_{i=1}^M [V_\theta(s_0^i; g_0^i)] + \frac{1}{N} \sum_{i=1}^N \left[f_\star(R_t^i + \gamma V(s_{t+1}^i; g_t^i) - V(s_t^i; g_t^i)) \right]$
  update V_θ: $V_\theta \leftarrow V_\theta - \alpha_V \nabla L_V(\theta)$

For bandit data, condition V on a and s instead, and there is no γV term

3 DT Training

for i in [#iterations] do
  sample offline data $\{s_t^i, a_t^i, s_{t+1}^i, g_t^i\}_{i=1}^N \sim \mathcal{B}$

```
   estimate the reward function $\hat{R}$ for each task in the current offline
data
   Policy objective:  $L_\pi(\phi) = \sum_{i=1}^{N} \left[\left(f'_\star\left(R_t^i + \gamma V_\theta(s_{t+1}^i; g_t^i) - V_\theta(s_t^i; g_t^i)\right)\right) \log \pi(a \mid s, g)\right]$
   Update $\pi_\phi$:  $\pi_\phi \leftarrow \pi_\phi - \alpha_\pi$
```

For bandit data, condition V on a and s instead, and there is no γV term

4 Test

4.1 Offline Test

```
# Bandit version
subopt = []
for $i$ in [500] do
   sample dataset $D$ with number of $i$ data $\sim \mathcal{B}_{\texttt{test}}$
   $s = s_0$
   $a^* = \arg\max_a \mu$
   $\hat{a} = \arg\max_{a \in \mathcal{A}} \pi_\phi(\cdot|s, D)$
   suboptimality = $\mu_{a^*} - \mu_{\hat{a}}$
   append suboptimality to subopt
```

4.2 Online Test

```
# Bandit version
suboptimality = 0
subopt = []
Initialize D = {}
for ep in [max_eps=500] do
   sample dataset $D \sim \mathcal{B}_{\texttt{test}}$
   $s = s_0 \sim$ Unif$[0, 1]$
   $\hat{a} = \pi_\phi(\cdot|s, D)$
   suboptimality += $\mu_{a^*} - \mu_{\hat{a}}$
   append suboptimality to subopt
   add $(a, r)$ to D
```