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 $\mathcal{B}$
for i in $[N]$ do
  $p_1 \sim$ Dirchlet distribution($\mathbb{1}$)
  $p_2 \sim$ point-mass distribution
  $\omega \sim$ Unif(0.1[10])
  action distribution $p = (1 - \omega)p_1 + \omega p_2$
  action means $\mu \sim \text{Unif}[0, 1]^5$

  for h in $[H = 500]$ do
    action $a_h \sim p$ (OHE)
    reward $r_h \sim N(\mu_a, \sigma^2)$ where $\sigma = 0.3$
    goal g_h = $(a_h, r_h) + g_{h+1}$
    append a_h, r_h to $\mathcal{B}$

2 Value Function Approximation

for i in [#iterations] do
  sample offline data $\{s_t^i, a_t^i, s_{t+1}^i, g_t^i\}_{i=1}^N \sim \mathcal{B}$, $\{s_0^i\}_{i=1}^M \sim \mu_0$
  obtain reward $\{R(s_t^i; g_t^i)\}_{i=1}^N \sim \mathcal{B}$
  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

3 DT Training

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

`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` π_ϕ`:` $\pi_\phi \leftarrow \pi_\phi - \alpha_\pi$

For bandit data, condition V on a and s instead

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`
`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 +=` $\mu_{a^*} - \mu_{\hat{a}}$
  `add` (a, r) `to D`