Decoupled Q-Chunking

Qiyang Li, Seohong Park, Sergey Levine
UC Berkeley

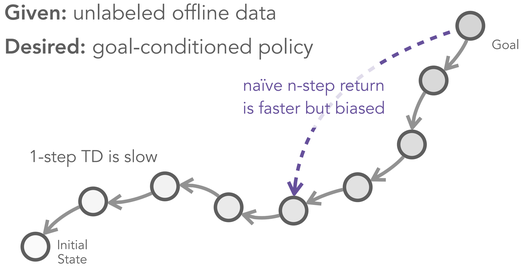
code

website

Problem and Challenges

Given: unlabeled offline data

Desired: goal-conditioned policy



Q-chunking (Li, Zhou, Levine; 2025)

- Learn Q-function for **action chunk** instead of **single action**.
- Use n-step return for fast value learning while not suffering from the bias. (There is a caveat!)

Open-Loop Consistency

What we should be doing...

$$Q(s_t, a_{t:t+n}) \leftarrow \mathbb{E}_{s_{t+1}\sim T(\cdot|s_t,a_t)}\left[\sum_{t'=t}^{t+h-1}\gamma^{t'-t}r(s_{t'},a_{t'}) + \gamma^h \hat{V}(s_{t+n})\right]$$

If we take $a_{t:t+n}$ open-loop in the environment, where would we end up be at s_{t+n}.

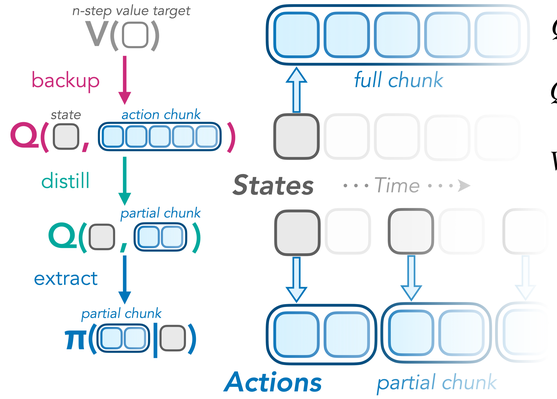
What we have been doing...

$$Q(s_t, a_{t:t+n}) \leftarrow \mathbb{E}_{s_{t+1:t+h+1}\sim P_D(\cdot|s_t,a_{t:t+h})}\left[\sum_{t'=t}^{t+h-1}\gamma^{t'-t}r(s_{t'},a_{t'}) + \gamma^h \hat{V}(s_{t+n})\right]$$

If we filter the prior data using $a_{t:t+n}$, what is the **conditional** distribution of s_{t+n} in the data.

How do we mitigate open-loop bias in Q-chunking?

Key Idea: Learn a policy that uses ***short*** action chunks ***(less open-loop bias)*** and a critic that uses ***long*** action chunks ***(faster value learning)***



Value Learning

(Implicit Value Backup)

$$Q_\phi(s_t, a_{t:t+h}) \leftarrow R_{t:t+h} + \gamma^h V_{\bar{\xi}}(s_{t+h})$$ (long)

$$Q^P_\psi(s_t, a_{t:t+h_a}) \leftarrow \text{Expectile}[Q_\phi(s_t, a_{t:t+h})]$$ (short)

$$V_\xi(s_t) \leftarrow \text{Expectile}[Q_\psi(s_t, a_{t:t+h_a})]$$

Policy Extraction

(Rejection Sampling)

$$a^1_{t:t+h_a}, \cdots, a^N_{t:t+h_a} \sim_{\text{i.i.d}} \pi_\beta(\cdot \mid s_t)$$

$$a^\star_{t:t+h_a} = \arg\max_{a_{t:t+h_a}} \left[Q_\psi(s_t, a_{t:t+h_a})\right]$$

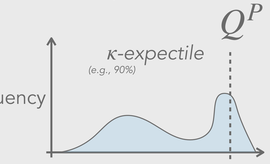
Why does it work?

$$Q^P(s_t, a_{t:t+h_a}) \approx Q(s_t, [a_{t:t+h_a}, a^\star_{t+h_a:t+h}]),$$

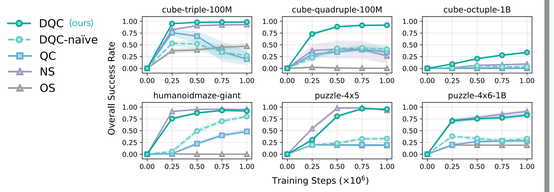
short critic — long critic — optimal action chunk completion

$$a^\star_{t+h_a:t+h} = \arg\max_{a_{t+h_a:t+h}} Q(s_t, [a_{t:t+h_a}, a_{t+h_a:t+h}])$$

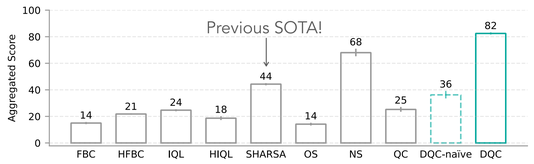
approximating argmax via expectile



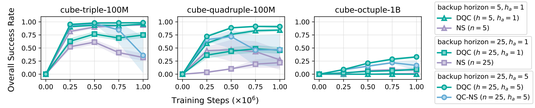
Empirical Results



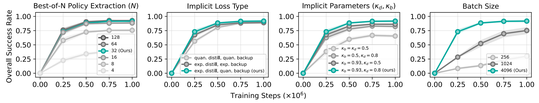
(A) How does it compare to prior methods?



(B) Ablation: n-step return



(C) Hyperparameter sensitivity



Theoretical Results

Definition: the prior data D is ε_h-OLC if for all $s_t, a_{t:t+h}$,

$$D_{\text{TV}}(\underbrace{T(s_{t+h'} \mid s_t, a_{t:t+h'})}_{\text{Take } a_{t:t+h'} \text{ open-loop}} \;\|\; P_D(s_{t+h'} \mid s_t, a_{t:t+h})) \le \varepsilon_h, \forall h' \in \{1,2,\cdots,h\}$$

OLC: Open-loop Consistency

BOV: Bounded Optimality Variability

$\pi^\star$: *Execute the first action in the chunk closed-loop*

$V^\star$: *The value of* $\pi^\star$

V_{QC} : *The value of the action chunking policy learned from* D

$V^\star$: *The value of the optimal closed-loop policy*

Result 1: Optimality of ACP under OLC

$$\|V^\star - V_{\text{QC}}\|_\infty \le \Theta\left(\frac{\varepsilon_h}{(1-\gamma)(1-\gamma^h)}\right)$$

Result 2: Optimality of Closed-loop ACP under OLC

$$\|V^\star - V^\bullet\|_\infty \le O\left(\frac{\varepsilon_h}{(1-\gamma)^2(1-\gamma^h)}\right)$$

Definition *(informal)*: Prior data D exhibits ϑ_h-BOV if for all $s_t, a_{t:t+h}$,

$$\max_{\text{supp}(P_D(\cdot|s_t,a_{t:t+h}))}\left[R_{t:t+h} + \gamma^h V^\star(s_{t+h})\right] - \min_{\text{supp}(P_D(\cdot|s_t,a_{t:t+h}))}\left[R_{t:t+h} + \gamma^h V^\star(s_{t+h})\right] \le \vartheta_h .$$

Result 3. Optimality of Closed-loop ACP under BOV

(Informal) If D contains optimal behavior, then

$$\|V^\star - V^\bullet\|_\infty \le \Theta\left(\frac{\vartheta_h}{(1-\gamma)(1-\gamma^h)}\right)$$

there exists an MDP and D where

$$V^\star - V_{\text{QC}} \ge O\left(\frac{1}{1-\gamma}\right)$$