北京航空航天大學
BEIHANG UNIVERSITY

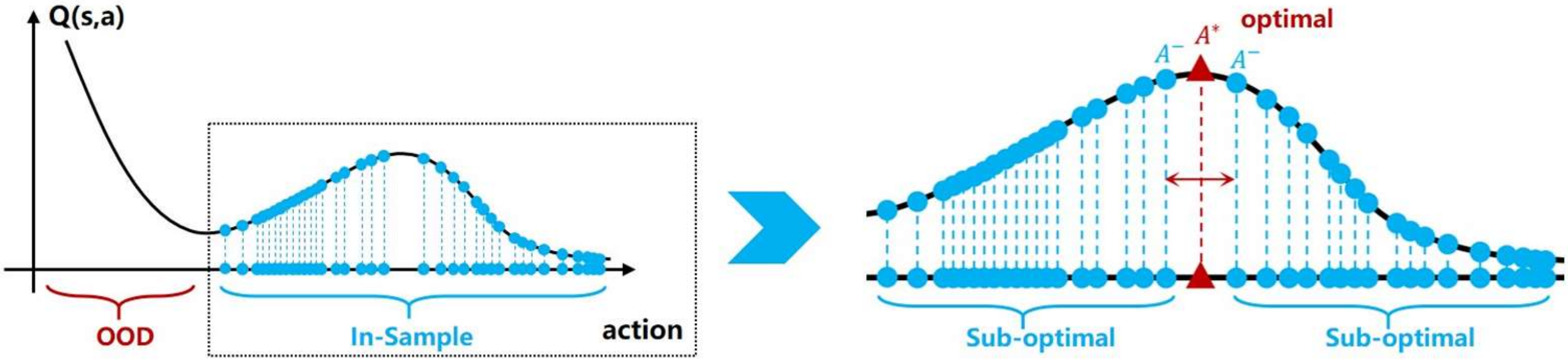
Value-aligned Behavior Cloning for Offline Reinforcement Learning via Bi-level Optimization

Xingyu Jiang, Ning Gao, Xiuhui Zhang, Hongkun Dou, Yue Deng

Beihang University

ICLR

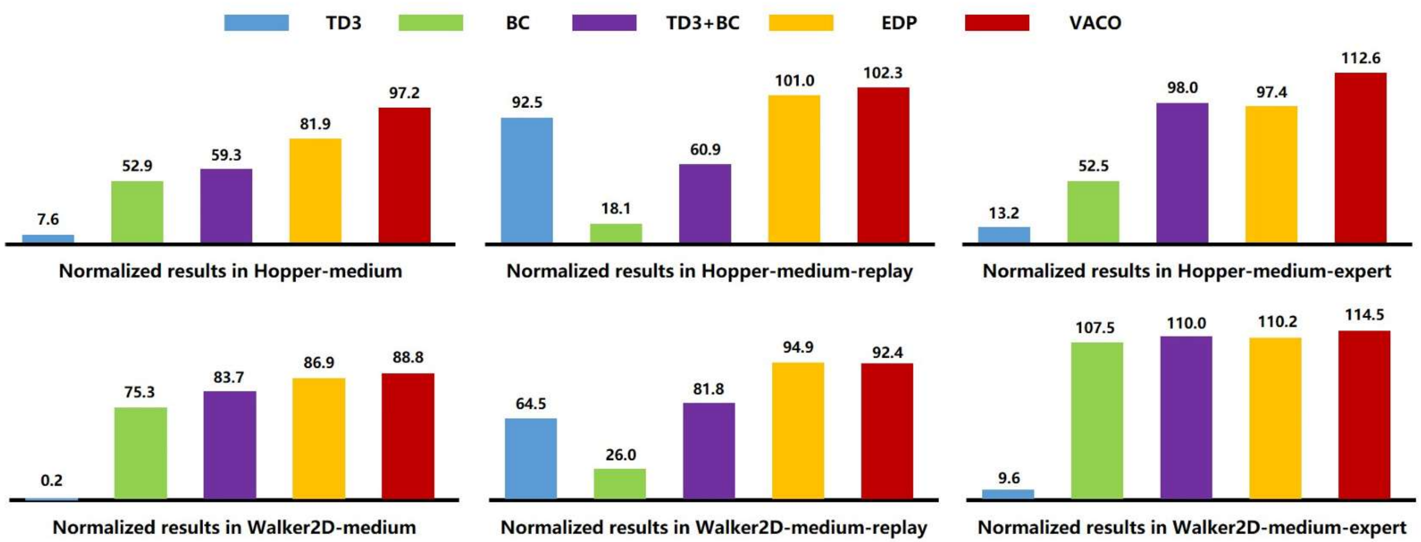
Two main challenges in Offline Reinforcement Learning



(a) OOD issues in offline RL

(b) alignment issues in offline RL

Method	Out-Of-Distribution (OOD) Issues	Alignment Issues
Behavior Cloning	√	×
TD3	×	√



Question:

How to combine TD3 and BC efficiently?

First step: introduce meta-scoring network

Behavior Cloning

$$J_{BC}(\phi) = \mathbb{E}_{(s,a)\sim D}[\pi_\phi(s) - a]^2$$

Meta-scoring Network

$w(s, a, Q_\theta(s, a))$

Weighted Behavior Cloning

$$J_{BC}^{w}(\phi) = \mathbb{E}_{(s,a)\sim D}\{w_\alpha(s, a, Q_\theta(s, a)) \cdot [\pi_\phi(s) - a]^2\}$$

Second step: introduce bi-level framework

High-level: **TD3**

Low-level: **Weighted Behavior Cloning**

$$\min_{\alpha} \quad J_\pi(\phi) := \mathbb{E}_{s\sim D}[-Q_\theta(s, \pi_\phi(s + N(0, \sigma)))]$$

$$\text{s.t. } \phi^*(\alpha) = \arg\min_{\phi} J_{BC}^{w}(\phi) := \mathbb{E}_{(s,a)\sim D}\{w_\alpha(s, a, Q_\theta(s, a)) \cdot [\pi_\phi(s) - a]^2\}$$

Optimization Loop:

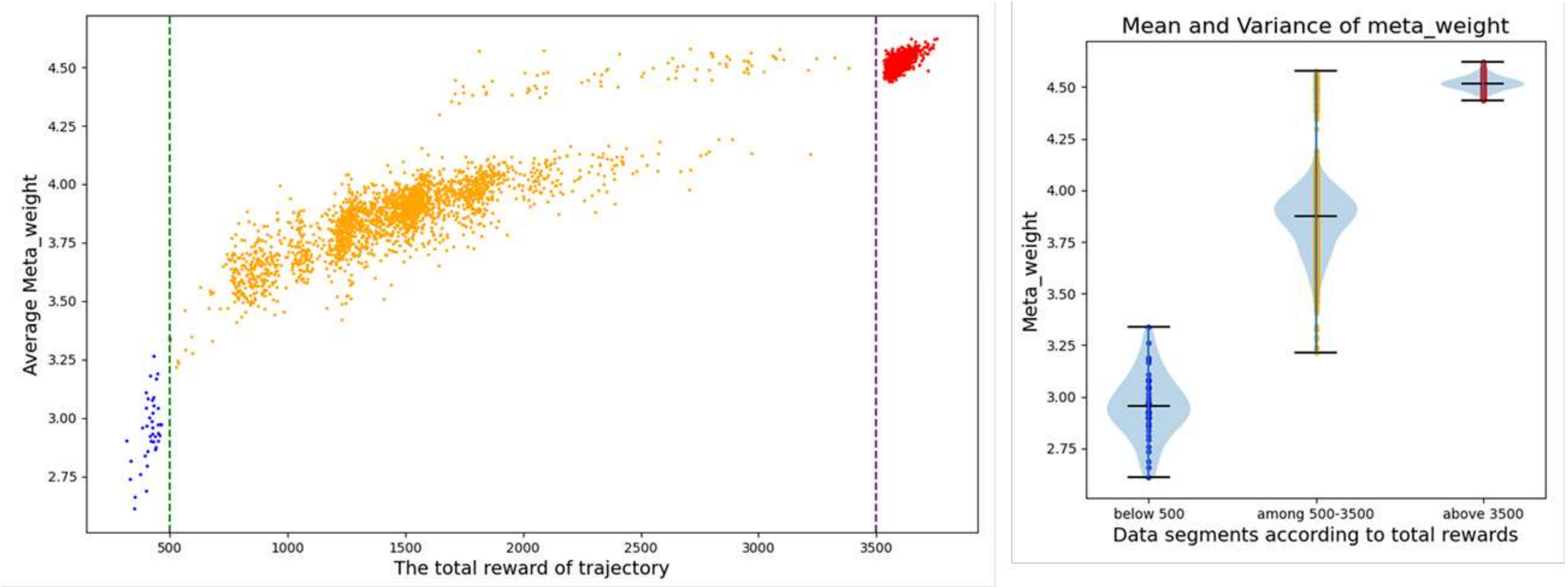
$$\phi_t \leftarrow \phi_{t-1} - \eta_1 \nabla_\phi J_{BC}^w(\phi) \qquad\qquad \alpha_t \leftarrow \alpha_{t-1} + \eta_2 \frac{\partial J_\pi(\phi)}{\partial \phi_t} \cdot \frac{\partial^2 J_{BC}^w(\phi_{t-1}))}{\partial \phi_{t-1} \partial \alpha}$$

Algorithm 1: Value-aligned Behavior Cloning via Bi-level Optimization (VACO)

Input: Fixed offline dataset $\mathcal{D}$, value network Q_θ, policy network π_ϕ, meta-scoring network w_α, update steps for value phase K_1, update steps for bi-level phase K_2

// Value Training Phase
for *update step* $k = 1...K_1$ **do**
    Sample a minibatch sample pairs (s, a, r, a') from $\mathcal{D}$
    Update value θ according to IQL's([27]) TD learning
end
// Bi-level Optimization Phase
for *update step* $k = 1...K_2$ **do**
    Sample a minibatch sample pairs (s, a, r, a') from $\mathcal{D}$
    Fix meta-scoring α and update policy ϕ according to Eq.7
    Fix policy ϕ and update meta-scoring α according to Eq.9
end

The effectiveness of the learned meta-scoring weights



(a) **(b)**

The Pearson correlation coefficient of meta weight and trajectory total rewards: 0.954

Above findings indicate that the learned meta-weights effectively reflect the quality of the data: higher meta-weights are associated with state-action pairs more likely generated by "expert"(good) policies, while lower meta-weights correspond to those generated by "random" (bad) policies. This aligns indeed with the original design intent of the meta-scoring network—to distinguish among data of varying quality.

Summary

We present VACO, a novel bi-level framework to balance OOD problem and value alignment issue concurrently for offline reinforcement learning.

$$\min_{\alpha} \quad J_{\pi}(\phi) := \mathbb{E}_{s\sim D}[-Q_{\theta}(s, \pi_{\phi}(s + N(0,\sigma)))]$$

$$\text{s.t. } \phi^*(\alpha) = \arg\min_{\phi} J^{w}_{BC}(\phi) := \mathbb{E}_{(s,a)\sim D}\{w_{\alpha}(s, a, Q_{\theta}(s,a)) \cdot [\pi_{\phi}(s) - a]^2\}$$

The framework comprises of:

- the internal loop for weighted behavior cloning.
- the external loop for policy-value alignment.
- meta-scoring network for assigning different importance weights to in-sample data.