Pre-trained Q-function

Online Fine-tuned Q-function

No-retention Online Update

Dataset

OOD

Pessimism on OOD (S,A) offline makes TD target pessimistic online

Dataset

Replay Buffer

Online Dist.

(S, A)

- - - - Ground Truth Q-function
——— Learned Offline Q-function
——— Learned Online Q-function

Pessimistic TD target makes Q-values more underestimated

Conservative regularizer (if present) acts on (S, A) outside replay buffer