# Reining Generalization in Offline Reinforcement Learning via Representation Distinction

**Yi Ma[1], Hongyao Tang[2,*], Dong Li[3], Zhaopeng Meng[1]**

1 College of Intelligence and Computing, Tianjin University

2 Université de Montréal, Mila

3 Noah's Ark Lab, Huawei Technology

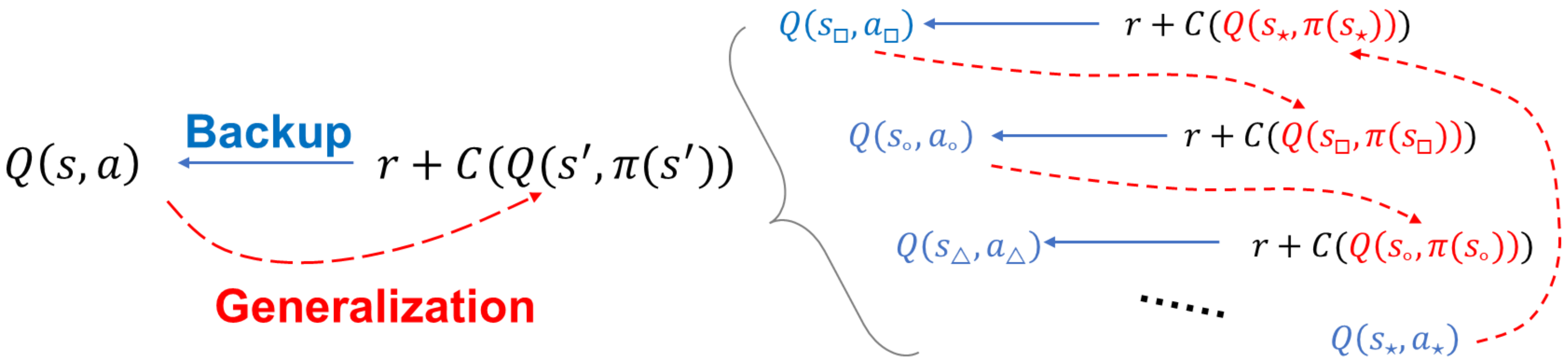
* Corresponding authors

# Outline

- **Problem Formulation**
- **Method**
- **Experimental Results**

# Problem Formulation: Backup-Generalization Cycle in Offline RL

- We introduce a view called Backup-Generalization Cycle. This view, as depicted in the Fig below, fosters an understanding of typical offline value function learning via two key components: **Backup** and **Generalization**:




This dynamic interplay forms a cycle:

(1) the backups on $(s,a) \in D$ consistently influence the values of $(s,a) \notin D$ through generalization;

(2) the consistently changing $Q(s', \pi(s'))$ participates in the backups on $(s,a) \notin D$;

The two kinds of dynamics iterate and twine during the learning process.

# Problem Formulation: Overgeneralization in Offline RL

Further, we consider how Q function update caused by typical Temporal-Difference (TD) learning on a single state-action pair $(s, a) \in D$ (denoted as $\phi \to \phi'$), affects the Q-value of an arbitrary state-action pair $(\bar{s}, \bar{a})$.

The post-update parameter $\phi'$ can be formalized as follows:

$$\phi' = \phi + \big(\mathcal{T} Q_\phi(s, a) - Q_\phi(s, a)\big)\nabla_\phi Q_\phi(s, a)$$

By Taylor expansion at the pre-update parameter $\phi$:

$$Q_{\phi'}(\bar{s}, \bar{a}) = Q_\phi(\bar{s}, \bar{a}) + \nabla_\phi Q_\phi(\bar{s}, \bar{a})^T(\phi' - \phi) + \mathcal{O}(\| \phi' - \phi \|^2)$$

By plugging the first Eq to the second Eq:

$$Q_{\phi'}(\bar{s}, \bar{a}) = Q_\phi(\bar{s}, \bar{a}) + k_\phi(\bar{s}, \bar{a}, s, a)\big(\mathcal{T} Q_\phi(s, a) - Q_\phi(s, a)\big) + \mathcal{O}(\| \phi' - \phi \|^2)$$

where $k_\phi(\bar{s}, \bar{a}, s, a) = \nabla_\phi Q_\phi(\bar{s}, \bar{a})^T \nabla_\phi Q_\phi(s, a)$, which us called Neural Tangent Kernel. We can control the generalization by mainly adjusting this kernel.

NEURAL INFORMATION PROCESSING SYSTEMS

天津大學 Tianjin University

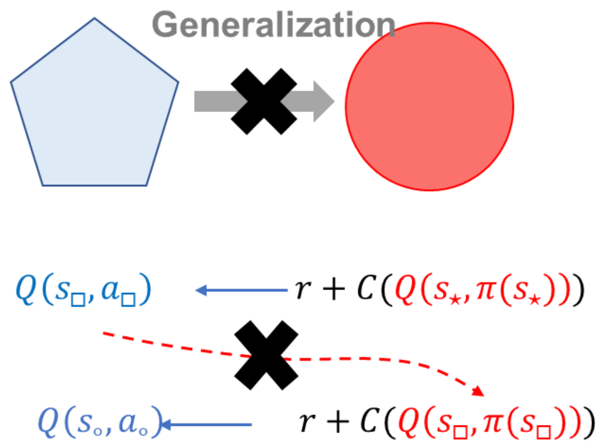
# Method: Reining Generalization via Two-Stage Kernel Control

Suppress $Q(s,a) \xrightarrow{\text{generalization}} Q(s,\pi(s))$ for $(s,a) \in D$ by $\boldsymbol{min_\phi \left|\nabla_\phi Q_\phi(s,a)^T \nabla_\phi Q_\phi(s,\pi(s))\right|}$

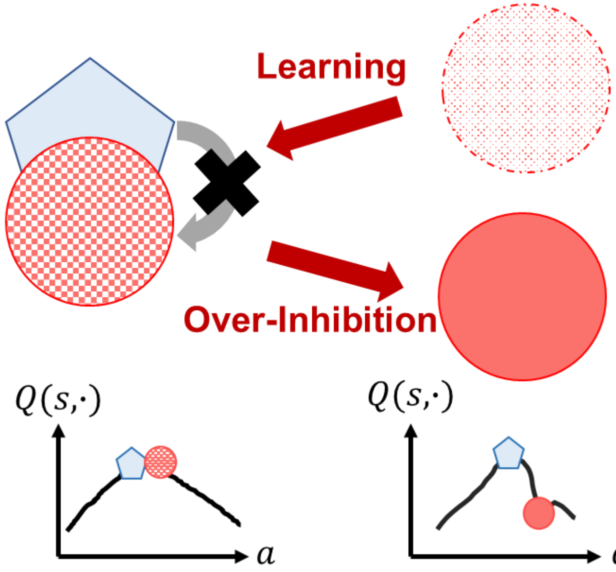
When the learning policy evolves and resembles the behavioral policy $min_\phi \left|\nabla_\phi Q_\phi(s,a)^T \nabla_\phi Q_\phi(s,\pi(s))\right|$ for $a \approx \pi(s)$ could lead to over-inhibition

Suppress the generalization between the learning policy distribution and the designed OOD policy distribution by $\boldsymbol{min_\phi \left|\nabla_\phi Q_\phi(s,\pi(s))^T \nabla_\phi Q_\phi(s,\pi_{ood}(s))\right|}$

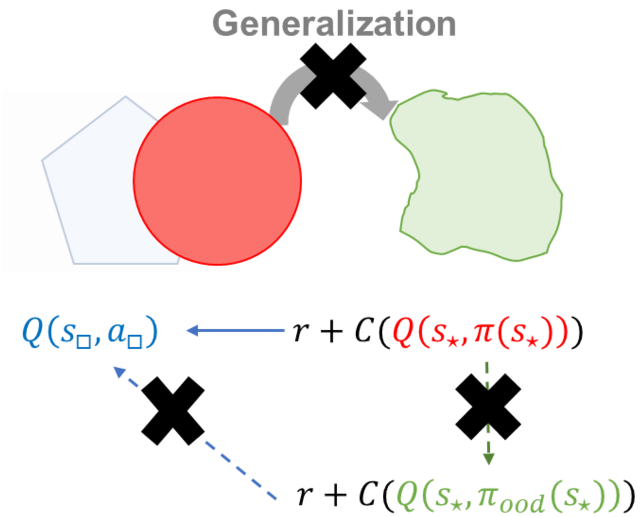
**Policy-Dataset Generalization Inhibition**




**Over-Inhibition when Distributions Overlap**




**Policy-OOD Generalization Inhibition**




Behavioral Policy Distribution  Learning Policy Distribution  Learned OOD Policy Distribution

# Results

Table 1: Results of different algorithms and the ones equipped with RD

| Dataset | TD3-N-Unc | TD3-N-Unc +RD | SAC-N-Unc | SAC-N-Unc +RD | TD3BC | TD3BC +RD | CQL | CQL +RD |
|---|---|---|---|---|---|---|---|---|
| halfcheetah-m | 66.8 ± 0.5 | **66.8 ± 1.2** | 65.9 ± 1.0 | **65.9 ± 1.9** | 48.0 ± 0.3 | **48.3 ± 0.5** | 47.1 ± 0.2 | **53.0 ± 0.5** |
| halfcheetah-mr | 53.4 ± 3.9 | **57.7 ± 0.9** | 53.2 ± 5.4 | **61.5 ± 1.4** | 44.6 ± 0.3 | 44.6 ± 0.5 | 45.2 ± 0.6 | **51.6 ± 0.9** |
| halfcheetah-me | 97.7 ± 2.2 | **101.1 ± 0.4** | 99.4 ± 2.5 | **102.5 ± 1.8** | 90.5 ± 6.6 | **93.9 ± 2.9** | 81.1 ± 6.0 | **90.2 ± 5.8** |
| hopper-m | 41.9 ± 50.5 | **103.0 ± 0.8** | 45.7 ± 41.0 | **102.8 ± 0.2** | 60.4 ± 4.0 | **61.0 ± 2.6** | 65.0 ± 6.1 | **74.9 ± 7.1** |
| hopper-mr | 92.5 ± 18.1 | **104.1 ± 0.8** | 104.7 ± 0.9 | 104.6 ± 0.4 | 61.2 ± 20.5 | **72.1 ± 8.4** | 87.7 ± 14.4 | **100.3 ± 3.2** |
| hopper-me | 100.3 ± 22.6 | **110.7 ± 0.6** | 110.9 ± 0.2 | 110.6 ± 0.3 | 105.4 ± 6.1 | 104.8 ± 2.8 | 93.9 ± 14.3 | **98.2 ± 9.7** |
| walker2d-m | 69.9 ± 35.2 | **97.6 ± 3.4** | 24.2 ± 28.2 | **92.3 ± 1.3** | 82.7 ± 5.5 | **83.7 ± 2.7** | 80.4 ± 3.5 | **84.5 ± 1.0** |
| walker2d-mr | 91.6 ± 2.7 | **92.1 ± 2.7** | 85.2 ± 2.7 | **86.9 ± 3.1** | 82.1 ± 2.5 | **84.8 ± 1.4** | 79.2 ± 5.0 | **94.4 ± 2.5** |
| walker2d-me | 90.6 ± 45.0 | **118.8 ± 1.2** | 113.1 ± 9.6 | **116.4 ± 1.5** | 110.2 ± 0.5 | 110.1 ± 0.5 | 109.7 ± 0.5 | **113.0 ± 0.5** |

Table 2: Average normalized scores of our methods and previous methods on the D4RL benchmark

| Dataset | BC | DT | TD3BC | CQL | IQL | EDAC | Diffusion-QL | SAC-N-Unc +RD | TD3-N-Unc +RD |
|---|---|---|---|---|---|---|---|---|---|
| halfcheetah-r | 2.2 | 2.2 | 11.0 | 31.3 | 13.7 | 28.4 | 22.0 | 25.4 | 31.0 |
| hopper-r | 3.7 | 5.4 | 8.4 | 5.3 | 8.4 | 25.3 | 18.3 | **31.6** | **31.7** |
| walker2d-r | 1.3 | 2.2 | 1.7 | 5.4 | 5.9 | 16.6 | 5.5 | **21.2** | **21.7** |
| halfcheetah-m | 42.6 | 42.6 | 48.0 | 47.1 | 47.4 | 65.9 | 51.5 | **65.9** | **66.8** |
| hopper-m | 52.9 | 67.6 | 60.4 | 65.0 | 66.3 | 101.6 | 96.6 | **102.8** | **103.0** |
| walker2d-m | 75.3 | 74.0 | 82.7 | 80.4 | 78.3 | 92.5 | 87.3 | 92.3 | **97.6** |
| halfcheetah-mr | 36.6 | 36.6 | 44.6 | 45.2 | 44.2 | 61.3 | 48.3 | **61.5** | 57.7 |
| hopper-mr | 18.1 | 82.7 | 61.2 | 87.7 | 94.7 | 101.0 | 102.0 | **104.6** | **104.1** |
| walker2d-mr | 26.0 | 66.6 | 82.1 | 79.2 | 73.9 | 87.1 | 98.0 | 86.9 | 92.1 |
| halfcheetah-me | 55.2 | 86.8 | 90.5 | 81.1 | 86.7 | 106.3 | 97.2 | 102.5 | 101.1 |
| hopper-me | 52.5 | 107.6 | 105.4 | 93.9 | 91.5 | 110.7 | 112.3 | 110.6 | 110.7 |
| walker2d-me | 107.5 | 108.1 | 110.2 | 109.7 | 109.6 | 114.7 | 111.2 | **116.4** | **118.8** |
| halfcheetah-e | 91.8 | 87.7 | 96.7 | 97.3 | 94.9 | 106.8 | 96.3 | **108.8** | 103.1 |
| hopper-e | 107.7 | 94.2 | 107.8 | 106.5 | 108.8 | 110.1 | 102.6 | 109.8 | 108.8 |
| walker2d-e | 108.7 | 108.3 | 110.2 | 109.3 | 109.7 | 115.1 | 109.5 | 112.3 | 111.2 |
| MuJoCo Total | 782.1 | 972.6 | 1020.9 | 1044.4 | 1034.0 | 1243.4 | 1158.6 | **1252.6** | **1259.4** |
| pen-human | 25.8 | 73.9 | -1.9 | 35.2 | 71.5 | 52.1 | 75.7 | 61.1 | **77.9** |
| pen-cloned | 38.3 | 67.3 | 9.6 | 27.2 | 37.3 | 68.2 | 60.8 | 53.0 | 65.5 |
| Adroit Total | 64.1 | 141.2 | 7.7 | 62.4 | 108.8 | 120.3 | 136.5 | 114.1 | **143.4** |
| Total | 846.2 | 1113.8 | 1028.6 | 1106.8 | 1142.8 | 1363.7 | 1295.1 | **1366.7** | **1402.8** |

# Thanks

**Welcome to communicate and cooperate with Tianjin University Deep Reinforcement Learning Lab**

- **Homepage: http://rl.beiyang.ren/**