On the Role of Discount Factor in Offline Reinforcement Learning

Hao Hu*, Yiqin Yang*, Qianchuan Zhao, Chongjie Zhang

June 14, 2022

Machine Intelligence Group

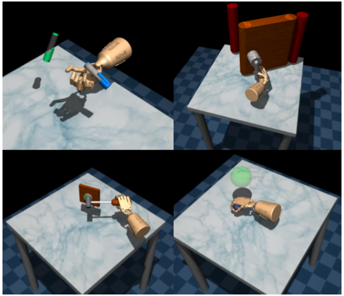
清華大學 交叉信息研究院
Tsinghua University Institute for Interdisciplinary Information Sciences

Offline Reinforcement Learning

Pessimism is the key for offline RL

- Constraining Policy (He and Hou, 2020; Fujimoto et al., 2019)
- Penalizing Uncertainty (Kumar et al., 2020; Wu et al., 2021; Yu et al., 2021)

Is there a simpler solution?



On the Role of Discount Factor in Offline Reinforcement Learning

Hao Hu, Yiqin Yang*, Qianchuan Zhao, Chongjie Zhang*

Observation:

- A lower discount factor can boost offline RL performance

Question:

- Is discount factor a proper way for pessimism? What affects the effectiveness of a lower guidance discount factor?

Analysis:

- Regularization Effect
- Pessimistic Effect

Linear MDPs

We say an episodic MDP $(\mathcal{S}, \mathcal{A}, H, \mathcal{P}, r)$ is a linear MDP with a known feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ if there exist d (unknown) measures $\mu_h = (\mu_h^{(1)}, \ldots, \mu_h^{(d)})$ over $\mathcal{S}$ and an unknown vector $\theta_h \in \mathbb{R}^d$ such that

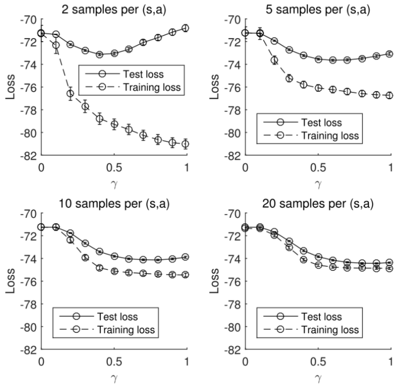
$$\mathcal{P}_h(x' \mid x, a) = \langle \phi(x, a), \mu_h(x') \rangle, \quad \mathbb{E}\big[r_h(s_h, a_h) \,\big|\, s_h = x, a_h = a\big] = \langle \phi(x, a), \theta_h \rangle \quad (1)$$

for all $(x, a, x') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ at each step $h \in [H]$.

- Tabular MDPs is a special case of linear MDPs.
- The condition above implies Q-function is linear.

Regularization Effect

- In online RL, use a smaller discount factor can be benificial (Jiang et al., 2015)
- Why? Does it applies to offline RL settings?



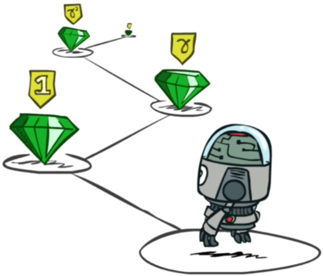
Regularization Effect

Lemma (Jiang et al. (2015))

For any MDP M with rewards in $[0, r_{\max}]$,
$\forall \pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and $\gamma \leq \gamma_e$,

$$\begin{aligned} V_{M,\gamma}(\pi) &\leq V_{M,\gamma_e}(\pi) \\ &\leq V_{M,\gamma}(\pi) + \frac{\gamma_e - \gamma}{(1-\gamma)(1-\gamma_e)} r_{\max}, \end{aligned}$$

where γ_e is the evaluation discount factor.



Pessimistic Value Iteration

Algorithm 1 Pessimistic Value Iteration

```
Require: Dataset $\mathcal{D} = \{(s_\tau, a_\tau, r_\tau)\}_{\tau=1}^{T}$.
Initialization: Set $\widehat{V}(\cdot) \leftarrow 0$ and construct $\Gamma(\cdot,\cdot)$.
while not converged do
    Construct $(\widehat{\mathbb{B}}_\gamma \widehat{V})(\cdot,\cdot)$
    Set $\widehat{Q}(\cdot,\cdot) \leftarrow (\widehat{\mathbb{B}}_\gamma \widehat{V})(\cdot,\cdot) - \Gamma(\cdot,\cdot)$.
    Set $\widehat{\pi}(\cdot \mid \cdot) \leftarrow \arg\max_\pi \mathbb{E}_\pi\left[\widehat{Q}(\cdot,\cdot)\right]$.
    Set $\widehat{V}(\cdot) \leftarrow \mathbb{E}_{\widehat{\pi}}\left[\widehat{Q}(\cdot,\cdot)\right]$.
end while
Return $\widehat{\pi}$
```

Regularization Effect

Lemma (PAC Guarantee in Discount Setting)

Suppose there exists an absolute constant $c^\dagger > 0$ *such that with probability* $1 - \xi/2$,

$$c^\dagger \cdot \sum_{\tau=1}^{N} \phi(s_\tau, a_\tau)\phi(s_\tau, a_\tau)^\top \succeq N \cdot \mathbb{E}_{\pi^*}\left[\phi(s_t, a_t)\phi(s_t, a_t)^\top \,\middle|\, s_0 = s\right],$$

for all $s \in \mathcal{S}$. *We set*

$$\lambda = 1, \quad \beta = c \cdot dV_{\max}\sqrt{\zeta}, \quad \zeta = \log\left(4dN/(1-\gamma)\xi\right),$$

where $V_{\max} = r_{\max}/(1-\gamma)$. *Then with probability* $1-\xi$, *the policy* $\widehat{\pi}$ *generated by pessimistic value iteration satisfies*

$$\mathrm{SubOpt}\left(\widehat{\pi}, s; \gamma\right) \leq 2c\frac{r_{\max}}{(1-\gamma)^2}\sqrt{c^\dagger d^3\zeta/N}, \ \forall s \in \mathcal{S}$$

Putting together

Theorem

We set

$$\lambda = 1, \quad \beta = c \cdot dV_{\max}\sqrt{\zeta}, \quad \zeta = \log\left(4dN/(1-\gamma)\xi\right), \tag{2}$$

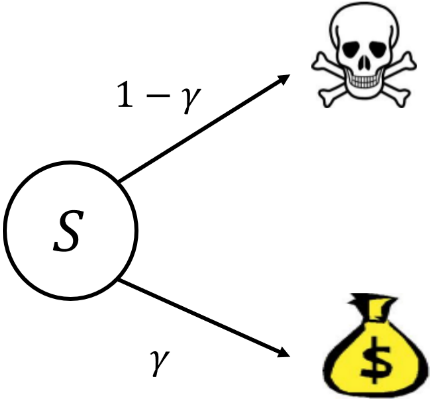
Then with probability $1-\xi$*, the suboptimality bound of the policy* $\widehat{\pi}$ *generated by pessimistic value iteration satisfies*

$$\begin{aligned}\mathrm{SubOpt}\left(\widehat{\pi};\gamma_e\right) \leq & \frac{2c}{(1-\gamma)^2}\sqrt{c^{\dagger}d^3\zeta/N} \cdot r_{\max} \\ & + \frac{\gamma_e - \gamma}{(1-\gamma)(1-\gamma_e)} r_{\max}.\end{aligned}$$

Pessimism Effect

An interesting equivalence

- Recall that discount factor can be intepreted as the probability of dying.
- A lower γ means that the probability of "dying" is higher.
- Does it sound like a kind of pessimism?



Pessimism Effect

An interesting equivalence

The optimal value function with a lower discount factor is equivalent to the pessimistic value function over a set of models. Formally, let

$$\pi^*_{M_\varepsilon} \in \arg\max_{\pi \in \Pi} \arg\min_{M \in \mathcal{M}_\varepsilon} V_{M,\gamma}(\pi), \tag{3}$$

where

$$\mathcal{M}_\varepsilon = \left\{ M | \mathcal{P}_M(\cdot|s,a) = (1-\varepsilon)\mathcal{P}_{M_0}(\cdot|s,a) + \varepsilon P(\cdot) \right\},$$

and $P(\cdot)$ is an arbitrary distribution over $\mathcal{S}$, then we have

$$V^*_{M_0,(1-\varepsilon)\gamma} = V_{M_0,\gamma}(\pi^*_{M_\epsilon}) + \Delta, \tag{4}$$

where Δ is a constant.

Proof of Equivalence

Proof.
Consider the following iteration

$$
\begin{aligned}
V_{\min} &\leftarrow \min_{s'} V(s'), \\
Q(s,a) &\leftarrow r(s,a) + \gamma(1-\varepsilon)\mathbb{E}_{s' \sim P_0} V(s') + \gamma \varepsilon V_{\min}, \\
V(s) &\leftarrow \max_{a} Q(s,a).
\end{aligned} \tag{5}
$$

It is easy to see that if the iteration in (5) converges, it is the value function for the policies specified in Equation (3). Then it suffices to show that the solution to the value iteration with discount factor $(1-\varepsilon)\gamma$ is the same as the above stationary solution up to a constant.

Proof of Equivalence

Let $Q(s,a)$ and $V(s,a)$ be the value learned with discount factor $(1-\varepsilon)\gamma$, then we have

$$Q(s,a) = r(s,a) + (1-\varepsilon)\gamma\mathbb{E}_{s'}V(s'),$$

Let $\Delta = \gamma\varepsilon \min_s[\max_a Q(s,a)]/(1-\gamma)$ and $\widetilde{Q}(\cdot,\cdot) = Q(\cdot,\cdot) + \Delta, \widetilde{V}(\cdot) = V(\cdot) + \Delta$, then we have

$$\min_s[\max_a \widetilde{Q}(s,a)] = \frac{(1-\gamma+\gamma\varepsilon)\Delta}{\gamma\varepsilon}.$$

This leads to

$$\begin{aligned}
&\widetilde{Q}(s,a) \\
=&r(s,a) + \gamma(1-\varepsilon)\mathbb{E}_{s'}V(s') + \Delta \\
=&r(s,a) + \gamma(1-\varepsilon)\mathbb{E}_{s'}\widetilde{V}(s') + (1-\gamma+\gamma\varepsilon)\Delta \\
=&r(s,a) + \gamma(1-\varepsilon)\mathbb{E}_{s'}\widetilde{V}(s') + \gamma\varepsilon \min_s[\max_a \widetilde{Q}(s,a)].
\end{aligned}$$

Pessimistic Effect

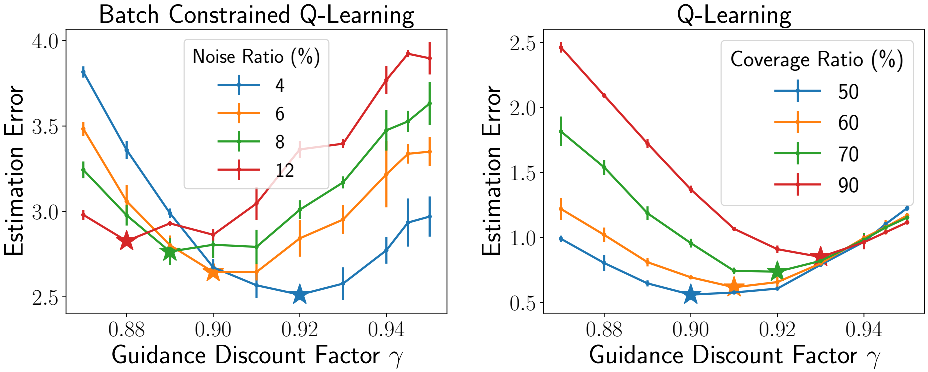
Theorem (Pessimistic Guarantees for a Lower γ)

Set $\gamma = (1-\varepsilon)\gamma_e$, *where* $\varepsilon \geq c_1 \log(c_2 N d/\xi)\sqrt{d/N}$. *Then with probability* $1-\xi$, *Learning with a guidance discount factor* γ *yields a policy* $\widehat{\pi}$ *such that*

$$\mathrm{SubOpt}(\widehat{\pi}; \gamma_e) \leq \frac{c_3}{(1-\gamma_e)^2}\sqrt{c^{\ddagger} d^2 \zeta / N} \cdot r_{max}, \tag{6}$$

where $c^{\ddagger} = \sup_{x \in \mathbb{R}^d} \frac{x^\top \Sigma_{\pi^*} x}{x^\top \Sigma_\rho x}$, $\Sigma_\rho = \mathbb{E}_\rho[\phi(s,a)\phi(s,a)^\top]$, $\Sigma_{\pi^*} = \mathbb{E}_{d^{\pi^*}}[\phi(s,a)\phi(s,a)^\top]$, *and* $c_1 \sim c_4$ *are universal constants.*

Tabular Experiments



The estimation error of BCQ and Q-Learning in the random MDP task. The star shapes mark the minimum of the curve.

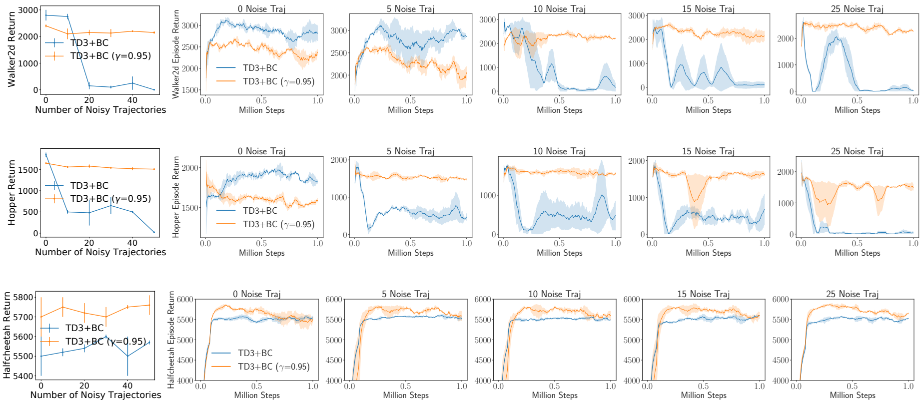
Results on D4RL Tasks

Experimental results on noised D4RL tasks with various offline RL methods

Tasks	BCQ	BCQ (γ)	TD3+BC	TD3+BC (γ)	COMBO	COMBO (γ)
walker2d (0 noised traj)	59.6±2.7	51.5±3.6	**62.0±3.2**	52.2±1.1	26.1±3.2	**65.5±1.7**
walker2d (10 noised traj)	53.7±2.5	51.8±1.3	**60.9.±1.2**	45.7±4.2	27.9±2.3	**63.1±1.6**
walker2d (50 noised traj)	20.3±3.3	**52.4±3.9**	4.3±1.2	**46.8±1.9**	27.2±1.6	**69.6±1.9**
walker2d (100 noised traj)	18.6±1.9	**52.1±2.2**	2.1±0.2	**46.6±1.3**	13.3±1.1	**70.7±2.3**
hopper (0 noised traj)	52.8±2.1	40.3±2.5	**52.5±1.8**	51.0±0.9	1.5±0.1	**53.5±3.2**
hopper (10 noised traj)	47.9±2.1	41.0±2.7	15.4±0.5	**47.9±0.3**	1.2±0.1	**56.5±2.5**
hopper (50 noised traj)	12.7±3.5	**44.1±1.9**	3.0±0.2	**47.0±0.5**	1.0±0.1	**48.6±4.2**
hopper (100 noised traj)	1.0±0.1	**41.6±0.6**	1.5±0.4	**46.3±0.7**	1.3±0.1	**52.3±1.7**
halfcheetah (0 noised traj)	40.2±1.3	**42.1±1.1**	45.3±1.5	**46.9±1.6**	32.6±1.6	27.6±1.5
halfcheetah (10 noised traj)	39.5±0.3	**40.2±3.3**	45.7±0.4	**47.3±1.6**	32.3±2.8	29.7±2.7
halfcheetah (50 noised traj)	36.5±0.9	**37.8±0.8**	45.9±0.3	**47.3±1.3**	31.1±4.7	28.0±1.6
halfcheetah (100 noised traj)	35.4±1.1	**36.4±1.7**	47.3±1.0	**46.1±1.8**	30.0±1.9	29.3±0.6

Results on D4RL Tasks

Experimental results on noised D4RL tasks with various noised trajectories



Pessimism Effects

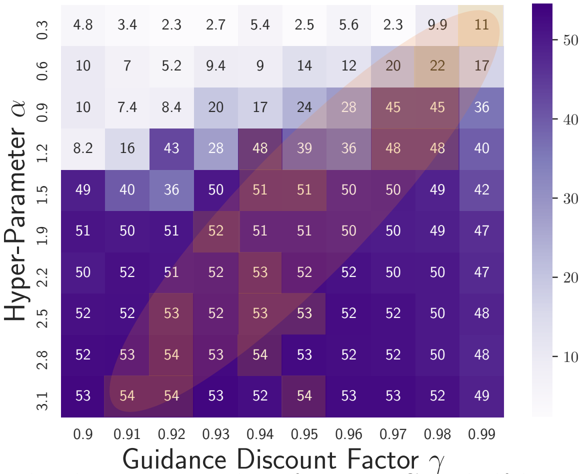
SAC-N	random-v2	medium-v2	medium-expert-v2	expert-v2
Halfcheetah (γ=0.95)	**30.0±1.6**	**65.1±0.9**	**51.4±2.2**	**82.7±0.8**
Halfcheetah (γ=0.99)	26.6±1.5	48.7±1.3	26.7±1.1	80.2±0.6
	random-v2	medium-v2	medium-expert-v2	expert-v2
Hopper (γ=0.95)	8.4±1.7	**22.4±2.1**	**23.1±1.9**	**14.5±2.6**
Hopper (γ=0.99)	14.5±3.5	7.1±2.0	15.4±1.4	2.3±0.3

Table 1: Results on Halfcheetah and Hopper tasks in D4RL. Q-ensemble size N is 2 in Halfcheetah and N is 50 in Hopper.

Adroit	pen-expert-v0	door-expert-v0	hammer-expert-v0
SAC-N (lower γ)	**97.1±3.2**	**106.4±1.9**	**100.6±2.3**
SAC-N (γ=0.99)	3.6±1.1	2.2±0.2	65.5±4.2

Table 2: Results on Adroit tasks in D4RL. Q-ensemble size N is 50 and $\gamma = 0.95$.

Discount Factor versus Other Trade-Offs



Relationship between γ and α of TD3+BC on halfcheetah task.

Summary

Discount factor plays an important role in offline RL

- Regularization Effect
  - Similar to online scenario, but affected by data size, coverage ratio etc.
  - More effective when data coverage is low and dataset is small
- Pessimistic Effect
  - A lower discount factor is equivalent to model-based pessimism
  - More effective when data coverage is sufficiently large

Thanks for Listening!

References I

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In International Conference on Machine Learning, pages 2052–2062. PMLR, 2019.

Qiang He and Xinwen Hou. Popo: Pessimistic offline policy optimization. arXiv preprint arXiv:2012.13682, 2020.

Nan Jiang, Alex Kulesza, Satinder Singh, and Richard Lewis. The dependence of effective planning horizon on model accuracy. In Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems, pages 1181–1189. Citeseer, 2015.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. Advances in Neural Information Processing Systems, 33:1179–1191, 2020.

Yue Wu, Shuangfei Zhai, Nitish Srivastava, Joshua Susskind, Jian Zhang, Ruslan Salakhutdinov, and Hanlin Goh. Uncertainty weighted actor-critic for offline reinforcement learning. arXiv preprint arXiv:2105.08140, 2021.

Tianhe Yu, Aviral Kumar, Rafael Rafailov, Aravind Rajeswaran, Sergey Levine, and Chelsea Finn. Combo: Conservative offline model-based policy optimization. Advances in Neural Information Processing Systems, 34, 2021.