tags: `2024 年 下學期讀書計畫` `Reinforcement Learning`

A Note on Conservative Offline Distributional Reinforcement Learning

- Remark: this note can be found in https://hackmd.io/@Origamyee/Sy7nkbsfR (https://hackmd.io/@Origamyee/Sy7nkbsfR)

Short opening

- Suppose you are a driver operating an autocar on a road.
- You want to minimize the time cost while still avoiding risky events.
- How do you train your autocar's direct model?
- More generally, how to avoid taking unsafe actions while still maximizing the expected reward ?
- This Problem we also call Conservative Reinforcement Learning
- Are you interested in how to combine current techniques to solve this problem, especially distributional techniques ?
- Let's collaborate to explore which techniques we can utilize

Introduction

Basic information

- Title: [Conservative Offline Distributional Reinforcement Learning](https://arxiv.org/pdf/2107.06106) (https://arxiv.org/pdf/2107.06106)
- Authors: Yecheng Jason Ma, Dinesh Jayaraman, Osbert Bastani
- Publication Date: 10/26, 2021
- Main Content: Conservative Offline Distributional Actor-Critic

Main challenges

- high uncertainty on out-of-distribution state-action pair
- value estimates for state-action pairs are high variance
- train a uncorrected policy (due to finite data)

High-level technical

Conservative Q-learning

- penalize Q values for out-of-distribution state-action pairs to ensure
  - the learned Q-function lower bounds the true Q-function
  - the quantiles of the learned return distribution lower bound those of the true return

distribution

Main contributions

- combing previous techniques (imitation learning and regularize the Q-function estimates), and they obtain conservative estimates of all quantile values of the return distribution

Personal perspective

- The estimator idea is simple: penalize the predicted quantiles of the return for out-of-distribution actions
- For example, if children go against their parents' expectations, then the children will be penalized in traditional Taiwanese families
- These papers demonstrate that the "penalize" approach is somehow feasible (with some theoretical guarantee) to meet their expectation
- Moreover, if I think something was wrong, then I will use this color to denote

Preliminaries

Offline RL

Goal

- learn the optimal policy π^*
- such that $Q^{\pi^*}(s, a) \geq Q^{\pi}(s, a)$ for all $s \in \mathcal{S}, a \in \mathcal{A}$ and all π

Markov Decision Process (MDP)

consist of five tuples $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$

- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $P\left(s' \mid s, a\right)$ transition distribution
- $R(r \mid s, a)$: reward distribution
- $\gamma \in (0, 1)$: discount factor

Notations

- $\pi(a \mid s)$: stochastic policy
- $\hat{\pi}_\beta(a \mid s)$: empirical behavior policy
- $Q^\pi(s, a) = \mathbb{E}_{D^\pi(\xi \mid s, a)}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right]$: Q-function
- $\xi = ((s_0, a_0, r_0), (s_1, a_1, r_1), \ldots)$: trajectory (rollout)
- $D^\pi(\xi \mid s, a)$: distribution over rollouts
- $(s, a, r, s') \sim \mathcal{D}$: a uniformly random sample from dataset
- actions not drawn from $\hat{\pi}_\beta(\cdot \mid s)$: we call out-of-distribution (OOD)

Distributional RL

Goal

- learn distribution of discounted cumulative rewards

notations

- $Z^\pi(s,a)=\sum_{t=0}^{\infty}\gamma^t r_t$: return distribution
- $F_{Z(s,a)}(x)$: cumulative density function (CDF) for return distribution $Z(s,a)$
- $F_{R(s,a)}$: CDF of $R(\cdot \mid s,a)$

- X, Y: random variables
- p-Wasserstein distance between X and Y: $W_p(X,Y)=\left(\int_0^1\left|F_Y^{-1}(\tau)-F_X^{-1}(\tau)\right|^p d\tau\right)^{1/p}$
- $\bar{d}_p\left(Z_1,Z_2\right)$: largest Wasserstein distance over (s,a)
- $\mathcal{Z}$: space of distributions over $\mathbb{R}$ with bounded p-th moment
- F_X^{-1}: quantile function (inverse CDF) of X``
- $F_{Z(s,a)}^{-1}(\tau)$: return distribution Z

- Given a distribution $g(\tau)$ over $[0,1]$
- distorted expectation of Z: $\Phi_g(Z(s,a))=\int_0^1 F_{Z(s,a)}^{-1}(\tau)g(\tau)d\tau$
- corresponding policy: $\pi_g(s):=\arg\max_a \Phi_g(Z(s,a))$

Optimization problem

$$\tilde{Z}^{k+1}=\underset{Z}{\arg\min}\alpha\cdot\mathbb{E}_{U(\tau),\mathcal{D}(s,a)}\left[c_0(s,a)\cdot F_{Z(s,a)}^{-1}(\tau)\right]+\mathcal{L}_p\left(Z,\hat{\mathcal{T}}^\pi\tilde{Z}^k\right) \tag{5}$$

- minimize $\tilde{Z}^{k+1}$

inequalities

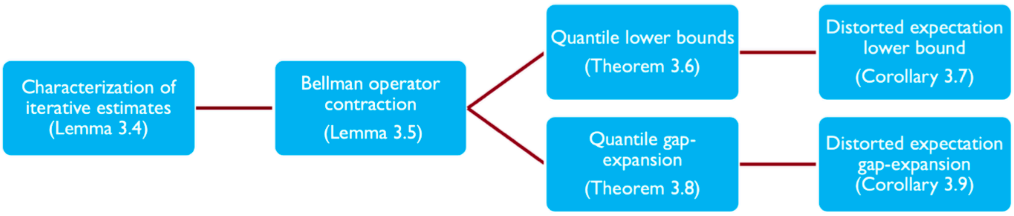
$$\hat{Z}^{k+1}=\underset{Z}{\arg\min}\mathcal{L}_p\left(Z,\hat{\mathcal{T}}^\pi\hat{Z}^k\right)\quad \text{where}\quad \mathcal{L}_p\left(Z,Z'\right)=\mathbb{E}_{\mathcal{D}(s,a)}\left[W_p\left(Z(s,a),Z'(s,a)\right)^p\right] \tag{4}$$

Technical assumptions

- learning algorithm only has access to a fixed dataset $\mathcal{D}:=\{(s,a,r,s')\}$ without any interaction with environment
- Assumption 3.1. $\hat{\pi}_\beta(a\mid s)>0$ for all $s\in\mathcal{D}$ and $a\in\mathcal{A}$
- Assumption 3.2. There exists $\zeta\in\mathbb{R}_{>0}$ such that for all $s\in\mathcal{S}$ and $a\in\mathcal{A}$, we have $F'_{Z^\pi(s,a)}(x)\geq\zeta$ (ζ-strongly monotone)
- Assumption 3.3. The search space of the minimum over Z in (5) is over all smooth functions $F_{Z(s,a)}$ (for all $s\in\mathcal{S}$ and $a\in\mathcal{A}$) with support on $[V_{\min},V_{\max}]$

Supporting Lemmas and Theoretical Analysis

Proof framework



- we will proof the above Lemmas and theorem
- And briefly tell their intuitions

Lemma 3.4.

For all $s \in \mathcal{D}, a \in \mathcal{A}, k \in \mathbb{N}$, and $\tau \in [0,1]$, we have

$$F^{-1}_{\tilde{Z}^{k+1}(s,a)}(\tau) = F^{-1}_{\hat{\mathcal{T}}^\pi \tilde{Z}^k(s,a)}(\tau) - c(s,a),$$

where $c(s,a) = \left|\alpha p^{-1} c_0(s,a)\right|^{1/(p-1)} \cdot \operatorname{sign}\left(c_0(s,a)\right)$

- high level: help us to iteratively compute $\tilde{Z}^{k+1}(s,a)$

Lemma 3.4. Proof

$$\begin{aligned}
&= \alpha \cdot \mathbb{E}_{U(\tau),\mathcal{D}(s,a)}\left[c_0(s,a) \cdot F^{-1}_{Z(s,a)}(\tau)\right] + \mathcal{L}_p\left(Z, \hat{\mathcal{T}}^\pi \tilde{Z}^k\right) \\
&= \alpha \cdot \mathbb{E}_{U(\tau),\mathcal{D}(s,a)}\left[c_0(s,a) \cdot F^{-1}_{Z(s,a)}(\tau)\right] + \mathbb{E}_{\mathcal{D}(s,a)} \int_0^1 \left|F^{-1}_{Z(s,a)}(\tau) - F^{-1}_{\hat{\mathcal{T}}^\pi \tilde{Z}^k(s,a)}(\tau)\right|^p d\tau \quad \text{(by the inequality (4))} \\
&= \int_0^1 \mathbb{E}_{\mathcal{D}(s,a)}\left[\alpha \cdot c_0(s,a) \cdot F^{-1}_{Z(s,a)}(\tau) + \left|F^{-1}_{Z(s,a)}(\tau) - F^{-1}_{\hat{\mathcal{T}}^\pi \tilde{Z}^k(s,a)}(\tau)\right|^p\right] d\tau \quad \text{(by the definition of expectation}
\end{aligned}$$

objective

$$\begin{aligned}
&= \alpha \cdot \mathbb{E}_{U(\tau),\mathcal{D}(s,a)}\left[c_0(s,a) \cdot F^{-1}_{Z(s,a)}(\tau)\right] + \mathcal{L}_p\left(Z, \hat{\mathcal{T}}^\pi \tilde{Z}^k\right) \\
&= \alpha \cdot \mathbb{E}_{U(\tau),\mathcal{D}(s,a)}\left[c_0(s,a) \cdot F^{-1}_{Z(s,a)}(\tau)\right] + \mathbb{E}_{\mathcal{D}(s,a)} \int_0^1 \left|F^{-1}_{Z(s,a)}(\tau) - F^{-1}_{\hat{\mathcal{T}}^\pi \hat{Z}^k(s,a)}(\tau)\right|^p d\tau \quad \text{(by the inequality (4))} \\
&= \int_0^1 \mathbb{E}_{\mathcal{D}(s,a)}\left[\alpha \cdot c_0(s,a) \cdot F^{-1}_{Z(s,a)}(\tau) + \left|F^{-1}_{Z(s,a)}(\tau) - F^{-1}_{\hat{\mathcal{T}}^\pi \tilde{Z}^k(s,a)}(\tau)\right|^p\right] d\tau \quad \text{(by the definition of expectation}
\end{aligned}$$

- We consider a perturbation, replace $F^{-1}_{Z(s,a)}(\tau)$ to $G^\epsilon_{s,a}(\tau)$, where

$$G^\epsilon_{s,a}(\tau) = F^{-1}_{Z(s,a)}(\tau) + \epsilon \cdot \phi_{s,a}(\tau)$$

- for arbitrary smooth functions $\phi_{s,a}$ with compact support $[V_{\min}, V_{\max}]$, yielding new objective

$$\int_0^1 \mathbb{E}_{\mathcal{D}(s,a)}\left[\alpha c_0(s,a) \cdot G^\epsilon_{s,a}(\tau) + \left|G^\epsilon_{s,a}(\tau) - F^{-1}_{\hat{\mathcal{T}}^\pi \hat{Z}^k(s,a)}(\tau)\right|^p\right] d\tau$$

- Taking the derivative with respect to ϵ at $\epsilon = 0$, we have

$$
\begin{aligned}
&\frac{d}{d\epsilon}\int_0^1 \mathbb{E}_{\mathcal{D}(s,a)}\left[\alpha c_0(s,a)\cdot G^{\epsilon}_{s,a}(\tau)+\left|G^{\epsilon}_{s,a}(\tau)-F^{-1}_{\hat{\mathcal{T}}^{\pi}\hat{Z}^k(s,a)}(\tau)\right|^p\right]d\tau\Bigg|_{\epsilon=0}\\
&=\mathbb{E}_{\mathcal{D}(s,a)}\int_0^1\left[\alpha c_0(s,a)+p\left|F^{-1}_{Z(s,a)}(\tau)-F^{-1}_{\hat{\tau}^{\pi}\hat{Z}^k(s,a)}(\tau)\right|^{p-1}\operatorname{sign}\left(F^{-1}_{Z(s,a)}(\tau)-F^{-1}_{\hat{\mathcal{T}}^{\pi}\hat{Z}^k(s,a)}(\tau)\right)\right]\phi_{s,a}(\tau)d\tau\\
&=0
\end{aligned}
$$

- Then

$$
\int_0^1\left[\alpha c_0(s,a)+p\left|F^{-1}_{Z(s,a)}(\tau)-F^{-1}_{\hat{\tau}^{\pi}\hat{Z}^k(s,a)}(\tau)\right|^{p-1}\operatorname{sign}\left(F^{-1}_{Z(s,a)}(\tau)-F^{-1}_{\hat{\tau}^{\pi}\hat{Z}^k(s,a)}(\tau)\right)\right]\phi_{s,a}(\tau)d\tau=0,
$$

  for all (s,a)
- By the fundamental lemma of the calculus of variations, for each s,a, if this term is zero for all $\phi_{s,a}$, then the integrand must be zero

$$
\alpha c_0(s,a)+p\left|F^{-1}_{Z(s,a)}(\tau)-F^{-1}_{\hat{\mathcal{T}}^{\pi}\hat{Z}^k(s,a)}(\tau)\right|^{p-1}\operatorname{sign}\left(F^{-1}_{Z(s,a)}(\tau)-F^{-1}_{\hat{\mathcal{T}}^{\pi}\hat{Z}^k(s,a)}(\tau)\right)=0
$$

- if and only if

$$
F^{-1}_{Z(s,a)}(\tau)=F^{-1}_{\hat{\mathcal{T}}^{\pi}\hat{Z}^k(s,a)}(\tau)-c(s,a)\quad(\text{sort the previous ineqaulity}),
$$

  where $c(s,a)=\left|\alpha p^{-1}c_0(s,a)\right|^{1/(p-1)}\cdot\operatorname{sign}\left(c_0(s,a)\right)$
- Clearly, this choice of Z is valid, so the claim follows

Lemma 3.5.

> $\tilde{\mathcal{T}}^{\pi}$ is a γ-contraction in $\bar{d}_p$, so $\tilde{Z}^k$ converges to a unique fixed point $\tilde{Z}^{\pi}$
>
> - shift operator $\mathcal{O}_c$ by $F^{-1}_{\mathcal{O}_c Z(s,a)}(\tau)=F^{-1}_{Z(s,a)}(\tau)-c(s,a)$
> - CDE operator $\tilde{\mathcal{T}}^{\pi}=\mathcal{O}_c\hat{\mathcal{T}}^{\pi}$

- high level: why two operator $\tilde{\mathcal{T}}^{\pi},\tilde{Z}^k$ are nice ?

Lemma 3.5. Proof

- first part: since $\hat{\mathcal{T}}^{\pi}$ is a γ-contraction in $\bar{d}_p$ (shown in [4, 7])
- and $\mathcal{O}_c$ is a non-expansion in $\bar{d}_p$, so by composition $\tilde{\mathcal{T}}^{\pi}$ is a γ-contraction in $\bar{d}_p$
- second: by the Banach fixed point theorem

Theorem 3.6.

For any $\delta \in \mathbb{R}_{>0}, c_0(s,a) > 0$, with probability at least $1-\delta$,

$$F^{-1}_{Z^\pi(s,a)}(\tau) \geq F^{-1}_{\tilde{Z}^\pi(s,a)}(\tau) + (1-\gamma)^{-1} \min_{s',a'} \left\{ c\left(s',a'\right) - \Delta\left(s',a'\right) \right\}$$
$$F^{-1}_{Z^\pi(s,a)}(\tau) \leq F^{-1}_{\tilde{Z}^\pi(s,a)}(\tau) + (1-\gamma)^{-1} \max_{s',a'} \left\{ c\left(s',a'\right) - \Delta\left(s',a'\right) \right\}$$

for all $s \in \mathcal{D}$, $a \in \mathcal{A}$, and $\tau \in [0,1]$, where $\Delta(s,a) = \frac{1}{\zeta}\sqrt{\frac{5|\mathcal{S}|}{n(s,a)} \log \frac{4|\mathcal{S}||\mathcal{A}|}{\delta}}$. Furthermore, for α sufficiently large (i.e., $\alpha \geq \max_{s,a}\left\{\frac{p \cdot \Delta(s,a)^{p-1}}{c_0(s,a)}\right\}$), we have $F^{-1}_{Z^\pi(s,a)}(\tau) \geq F^{-1}_{\tilde{Z}^\pi(s,a)}(\tau)$.

- high level: first inequality: the quantile estimates (computed by CDE) form a lower bound on the true quantiles as long as α satisfies the given condition
- second inequality: this lower bound is tight

Theorem 3.6. Proof

- we use Lemma A.1., Lemma A.2., Lemma A.6. to help us prove Theorem 3.6.
- and their proof are in Appendix

Lemma A.1.

$n(s,a) = |\{(s,a) \mid (s,a,r,s') \in \mathcal{D}\}|$: number of times (s,a) occurs in D.
For any return distribution Z with ζ-strongly monotone CDF $F_{Z(s,a)}$ and any $\delta \in \mathbb{R}_{>0}$, with probability at least $1-\delta$, for all $s \in \mathcal{D}$ and $a \in \mathcal{A}$, we have

$$\left\| F^{-1}_{\hat{\mathcal{T}}^\pi Z(s,a)} - F^{-1}_{\mathcal{T}^\pi Z(s,a)} \right\|_\infty \leq \Delta(s,a) \quad \text{where} \quad \Delta(s,a) = \frac{1}{\zeta}\sqrt{\frac{5|\mathcal{S}|}{n(s,a)} \log \frac{4|\mathcal{S}||\mathcal{A}|}{\delta}}$$

- high level: bound the estimation error of $\hat{\mathcal{T}}^\pi$ compared to $\mathcal{T}^\pi$

Lemma A.2.

If Z satisfies $\left\| F^{-1}_{Z(s,a)} - F^{-1}_{\mathcal{T}Z(s,a)} \right\|_\infty \leq \beta$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$, then

$$\left\| F^{-1}_{Z(s,a)} - F^{-1}_{Z^\pi(s,a)} \right\|_\infty \leq (1-\gamma)^{-1}\beta \quad (\forall s \in \mathcal{S}, a \in \mathcal{A})$$

- high level: relates one-step distributional Bellman contraction to an ∞-norm bound at the fixed point

Lemma A.6.

For any $\beta \in \mathbb{R}$, if Z satisfies

$$F^{-1}_{Z(s,a)}(\tau) \geq F^{-1}_{\mathcal{T}^\pi} Z(s,a)(\tau) + \beta \quad (\forall \tau \in [0,1]) \tag{12}$$

for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$, then we have

$$F^{-1}_{Z(s,a)}(\tau) \geq F^{-1}_{Z^\pi(s,a)}(\tau) + (1-\gamma)^{-1}\beta \quad (\forall \tau \in [0,1])$$

The result holds with $\geq$ replaced by $\leq$, or with $\mathcal{T}^\pi$ and Z^π replaced by $\hat{\mathcal{T}}^\pi$ and $\hat{Z}^\pi$ or $\tilde{\mathcal{T}}^\pi$ and $\tilde{Z}^\pi$

- back to proof Theorem 3.6.
- First, with probability at least $1-\delta$, we have

$$\begin{aligned} F^{-1}_{\tilde{\mathcal{T}}^\pi Z^\pi(s,a)}(\tau) &= F^{-1}_{\hat{\mathcal{T}}^\pi Z^\pi(s,a)}(\tau) - c(s,a) \\ &\text{(apply Lemma 3.4.(holds for any } \tilde{Z}^k\text{), substituting } \tilde{Z}^k = Z^\pi) \\ &\leq F^{-1}_{\mathcal{T}^\pi Z^\pi(s,a)}(\tau) - c(s,a) + \Delta(s,a) \\ &(\because Z^\pi \text{ is } \zeta\text{-strongly monotone, applying Lemma A.1. with } Z = Z^\pi) \quad (8) \\ &= F^{-1}_{Z^\pi(s,a)}(\tau) - c(s,a) + \Delta(s,a) \\ &(\because Z^\pi = \mathcal{T}^\pi Z^\pi) \end{aligned}$$

- Second, rearranging (8), we have

$$\begin{aligned} F^{-1}_{Z^\pi(s,a)}(\tau) &\geq F^{-1}_{\tilde{\mathcal{T}}^\pi Z^\pi(s,a)}(\tau) + c(s,a) - \Delta(s,a) \\ &\geq F^{-1}_{\tilde{\mathcal{T}}^\pi Z^\pi(s,a)}(\tau) + \min_{s,a}\{c(s,a) - \Delta(s,a)\} \\ &\geq F^{-1}_{\tilde{Z}^\pi(s,a)}(\tau) + (1-\gamma)^{-1} \min_{s,a}\{c(s,a) - \Delta(s,a)\} \\ &\text{(applied Lemma A. 6 for the case } \geq \text{ and } \tilde{\mathcal{T}}^\pi\text{, with } \beta = \min_{s,a}\{c(s,a) - \Delta(s,a)\}) \end{aligned} \tag{9}$$

- Finally, note that for the last term in (9) to be positive, we need

$$\alpha p^{-1} c_0(s,a) \geq \Delta(s,a)^{p-1} \quad (\forall s,a)$$

- Since we have assumed that $c_0(s,a) > 0$, this expression is in turn equivalent to

$$\alpha \geq \max_{s,a}\left\{\frac{p \cdot \Delta(s,a)^{p-1}}{c_0(s,a)}\right\}$$

- so the claim holds

Corollary 3.7.

For any $\delta \in \mathbb{R}_{>0}, c_0(s,a) > 0, \alpha$ sufficiently large, and $g(\tau)$, with probability at least $1-\delta$, for all $s \in \mathcal{D}, a \in \mathcal{A}$, we have $\Phi_g\left(Z^\pi(s,a)\right) \geq \Phi_g\left(\tilde{Z}^\pi(s,a)\right)$.

- high level: integrals of the return quantiles version
- It extends Theorem 3.6

Theorem 3.8.

Under the choice

$$c_0(s,a) = \frac{\mu(a \mid s) - \hat{\pi}_\beta(a \mid s)}{\hat{\pi}_\beta(a \mid s)} \tag{6}$$

, $p = 2$, and α sufficiently large (satisfy the α condition in Theorem 3.6.), for all $s \in \mathcal{S}$ and $\tau \in [0,1]$, we have

$$\mathbb{E}_{\hat{\pi}_\beta(a|s)} F^{-1}_{\tilde{Z}^\pi(s,a)}(\tau) - \mathbb{E}_{\mu(a|s)} F^{-1}_{\tilde{Z}^\pi(s,a)}(\tau) \geq \mathbb{E}_{\hat{\pi}_\beta(a|s)} F^{-1}_{Z^\pi(s,a)}(\tau) - \mathbb{E}_{\mu(a|s)} F^{-1}_{Z^\pi(s,a)}(\tau)$$

- high level: the difference in quantile values between in-distribution and OOD actions is larger under $\tilde{\mathcal{T}}^\pi$ than under $\mathcal{T}^\pi$ ($\tilde{\mathcal{T}}^\pi$ is gap-expanding)
- $c_0(s,a)$ is large for actions a with higher probability under μ than under $\hat{\pi}_\beta$ (i.e., an OOD action)

Theorem 3.8. Proof

- we use Lemma A.3. to help us prove Theorem 3.8.

Lemma A.3.

For any Z and any $\bar{\Delta}$, for sufficiently large α, with probability at least $1-\delta$, we have
$\mathbb{E}_{\hat{\pi}_\beta(a|s)} F^{-1}_{\tilde{\mathcal{T}}^\pi Z(s,a)}(\tau) - \mathbb{E}_{\mu(a|s)} F^{-1}_{\tilde{\mathcal{T}}^\pi Z(s,a)}(\tau) \geq \mathbb{E}_{\hat{\pi}_\beta(a|s)} F^{-1}_{\mathcal{T}^\pi Z(s,a)}(\tau) - \mathbb{E}_{\mu(a|s)} F^{-1}_{\mathcal{T}^\pi Z(s,a)}(\tau) + \bar{\Delta}$

- high level: the difference in quantile values between in-distribution and OOD actions is larger under $\tilde{\mathcal{T}}^\pi Z$ than under $\mathcal{T}^\pi Z$

Lemma A.3. Proof

- First, by Lemma A.1., with probability at least $1-\delta$, we have

$$F^{-1}_{\mathcal{T}^\pi Z(s,a)}(\tau) - \Delta(s,a) \leq F^{-1}_{\hat{\mathcal{T}}^\pi Z(s,a)}(\tau) \leq F^{-1}_{\mathcal{T}^\pi Z(s,a)}(\tau) + \Delta(s,a)$$

- By Lemma 3.4 , we have

$$F^{-1}_{\tilde{\mathcal{T}}^\pi Z(s,a)}(\tau) = F^{-1}_{\hat{\tau}^\pi Z(s,a)}(\tau) - c(s,a)$$

- Then, when we combine them together, we have

$$F^{-1}_{\mathcal{T}^\pi Z(s,a)}(\tau) - \Delta(s,a) \leq F^{-1}_{\tilde{\mathcal{T}}^\pi Z(s,a)}(\tau) + c(s,a) \leq F^{-1}_{\mathcal{T}^\pi Z(s,a)}(\tau) + \Delta(s,a)$$

- substract $c(s,a)$ all sides, we get

$$F^{-1}_{\mathcal{T}^\pi Z(s,a)}(\tau) - c(s,a) - \Delta(s,a) \leq F^{-1}_{\tilde{\tau}^\pi Z(s,a)}(\tau) \leq F^{-1}_{\mathcal{T}^\pi Z(s,a)}(\tau) - c(s,a) + \Delta(s,a)$$

- Taking the expectation over $\hat{\pi}_\beta$ (resp., μ) of the lower (resp., upper) bound gives

$$\mathbb{E}_{\hat{\pi}_\beta(a|s)} F^{-1}_{\tilde{\mathcal{T}}^\pi Z(s,a)}(\tau) \geq \mathbb{E}_{\hat{\pi}_\beta(a|s)} F^{-1}_{\mathcal{T}^\pi Z(s,a)}(\tau) - \mathbb{E}_{\hat{\pi}_\beta(a|s)} c(s,a) - \mathbb{E}_{\hat{\pi}_\beta(a|s)} \Delta(s,a)$$
$$\mathbb{E}_{\mu(a|s)} F^{-1}_{\tilde{\mathcal{T}}^\pi Z(s,a)}(\tau) \leq \mathbb{E}_{\mu(a|s)} F^{-1}_{\mathcal{T}^\pi Z(s,a)}(\tau) - \mathbb{E}_{\mu(a|s)} c(s,a) + \mathbb{E}_{\mu(a|s)} \Delta(s,a),$$

- Then, subtracting the latter from the former and rearranging terms, we get

$$\begin{aligned}
&\mathbb{E}_{\hat{\pi}_\beta(a|s)} F^{-1}_{\tilde{\mathcal{T}}^\pi Z(s,a)}(\tau) - \mathbb{E}_{\mu(a|s)} F^{-1}_{\tilde{\mathcal{T}}^\pi Z(s,a)}(\tau) \\
&\geq \mathbb{E}_{\hat{\pi}_\beta(a|s)} F^{-1}_{\mathcal{T}^\pi Z(s,a)}(\tau) - \mathbb{E}_{\mu(a|s)} F^{-1}_{\mathcal{T}^\pi Z(s,a)}(\tau) + (\mathbb{E}_{\mu(a|s)} c(s,a) - \mathbb{E}_{\hat{\pi}_\beta(a|s)} c(s,a)) - \bar{\Delta}(s) \\
&\geq \mathbb{E}_{\hat{\pi}_\beta(a|s)} F^{-1}_{\mathcal{T}^\pi Z(s,a)}(\tau) - \mathbb{E}_{\mu(a|s)} F^{-1}_{\mathcal{T}^\pi Z(s,a)}(\tau) + (\alpha/2)\bar{c}(s) - \bar{\Delta}(s) \\
&(\text{notice we have } c_0(s,a) = \alpha p^{-1} c_0(s,a) = \frac{\alpha}{2} c_0(s,a))
\end{aligned}$$

- where

$$\begin{aligned}
\bar{c}(s) &= \mathbb{E}_{\mu(a|s)} c_0(s,a) - \mathbb{E}_{\hat{\pi}_\beta(a|s)} c_0(s,a) \\
\bar{\Delta}(s) &= \mathbb{E}_{\mu(a|s)} \Delta(s,a) + \mathbb{E}_{\hat{\pi}_\beta(a|s)} \Delta(s,a)
\end{aligned}$$

- Notice that if we have

$$(\alpha/2)\bar{c}(s) \geq \bar{\Delta}(s) + \bar{\Delta} \quad (\forall s) \tag{10}$$

- Then we can use (10) to obtain Lemma A.3.
- So we claim (10) holds for sufficient large α
- Note that

$$\mathbb{E}_{\hat{\pi}_\beta(a|s)} c_0(s,a) = \sum_a \left(\frac{\mu(a \mid s) - \hat{\pi}_\beta(a \mid s)}{\mu_\beta(a \mid s)} \right) \mu_\beta(a \mid s) = \sum_a \left(\mu(a \mid s) - \hat{\pi}_\beta(a \mid s) \right) = 0$$

- and

$$\begin{aligned}
&\mathbb{E}_{\mu(a|s)} c_0(s,a) \\
&= \sum_a \left(\frac{\mu(a \mid s) - \hat{\pi}_\beta(a \mid s)}{\hat{\pi}_\beta(a \mid s)} \right) \mu(a \mid s) \\
&= \sum_a \left(\frac{\mu(a \mid s) - \hat{\pi}_\beta(a \mid s)}{\hat{\pi}_\beta(a \mid s)} \right) \left(\mu(a \mid s) - \hat{\pi}_\beta(a \mid s) \right) + \sum_a \left(\frac{\mu(a \mid s) - \hat{\pi}_\beta(a \mid s)}{\hat{\pi}_\beta(a \mid s)} \right) \hat{\pi}_\beta(a \mid s) \\
&= \sum_a \left(\frac{\mu(a \mid s) - \hat{\pi}_\beta(a \mid s)}{\hat{\pi}_\beta(a \mid s)} \right) \left(\mu(a \mid s) - \hat{\pi}_\beta(a \mid s) \right) \\
&= \sum_a \frac{\left(\mu(a \mid s) - \hat{\pi}_\beta(a \mid s) \right)^2}{\hat{\pi}_\beta(a \mid s)}
\end{aligned}$$

- so we have

$$\bar{c}(s) = \mathbb{E}_{\mu(a|s)} c_0(s,a) - \mathbb{E}_{\hat{\pi}_\beta(a|s)} c_0(s,a) \sum_a \frac{\left(\mu(a \mid s) - \hat{\pi}_\beta(a \mid s) \right)^2}{\hat{\pi}_\beta(a \mid s)} = \mathrm{Var}_{\hat{\pi}_\beta(a|s)} \left[\frac{\mu(a \mid s) - \hat{\pi}_\beta(a \mid s)}{\hat{\pi}_\beta(a \mid s)} \right] > 0$$

- the last inequality holds since $\mu(a \mid s) \neq \hat{\pi}_\beta(a \mid s)$
- Thus, for 10 to hold, it suffices to have

$$\alpha \geq 2 \cdot \max_s \left\{ \mathrm{Var}_{\hat{\pi}_\beta(a|s)} \left[\frac{\mu(a \mid s) - \hat{\pi}_\beta(a \mid s)}{\hat{\pi}_\beta(a \mid s)} \right]^{-1} \cdot (\bar{\Delta}(s) + \bar{\Delta}) \right\}$$

- The Lemma A.3. follows.

- Now, let $Z_0 = \tilde{Z}_0$, and let $Z_k = (\mathcal{T}^\pi)^k Z_0$ and $\tilde{Z}_k = \left(\tilde{\mathcal{T}}^\pi\right)^k \tilde{Z}_0$
- Applying Lemma A.3 with $Z = \tilde{Z}^k$ and $\bar{\Delta} = 4V_{\max}$, we have

$$\begin{aligned}
&\mathbb{E}_{\hat{\pi}_\beta(a|s)} F^{-1}_{\tilde{\mathcal{T}}^\pi \tilde{Z}^k(s,a)}(\tau) - \mathbb{E}_{\mu(a|s)} F^{-1}_{\tilde{\mathcal{T}}^\pi \tilde{Z}^k(s,a)}(\tau) \\
&\geq \mathbb{E}_{\hat{\pi}_\beta(a|s)} F^{-1}_{\mathcal{T}^\pi \tilde{Z}^k(s,a)}(\tau) - \mathbb{E}_{\mu(a|s)} F^{-1}_{\mathcal{T}^\pi \tilde{Z}^k(s,a)}(\tau) + \bar{\Delta} \quad \text{(apply Lemma A.3)} \\
&= \mathbb{E}_{\hat{\pi}_\beta(a|s)} F^{-1}_{\mathcal{T}^\pi Z^k(s,a)}(\tau) - \mathbb{E}_{\mu(a|s)} F^{-1}_{\mathcal{T}^\pi Z^k(s,a)}(\tau) + \bar{\Delta} \\
&\quad + \left(\mathbb{E}_{\hat{\pi}_\beta(a|s)} F^{-1}_{\mathcal{T}^\pi \tilde{Z}^k(s,a)}(\tau) - \mathbb{E}_{\mu(a|s)} F^{-1}_{\mathcal{T}^\pi \tilde{Z}^k(s,a)}(\tau)\right) \\
&\quad - \left(\mathbb{E}_{\hat{\pi}_\beta(a|s)} F^{-1}_{\mathcal{T}^\pi Z^k(s,a)}(\tau) - \mathbb{E}_{\mu(a|s)} F^{-1}_{\mathcal{T}^\pi Z^k(s,a)}(\tau)\right) \\
&\geq \mathbb{E}_{\hat{\pi}_\beta(a|s)} F^{-1}_{\mathcal{T}^\pi Z^k(s,a)}(\tau) - \mathbb{E}_{\mu(a|s)} F^{-1}_{\mathcal{T}^\pi Z^k(s,a)}(\tau) \\
&\quad + \bar{\Delta} - 4V_{\max} \\
&= \mathbb{E}_{\hat{\pi}_\beta(a|s)} F^{-1}_{\mathcal{T}^\pi Z^k(s,a)}(\tau) - \mathbb{E}_{\mu(a|s)} F^{-1}_{\mathcal{T}^\pi Z^k(s,a)}(\tau) \quad (\because \bar{\Delta} = 4V_{\max})
\end{aligned}$$

- The Theorem 3.8. follows by taking the limit $k \to \infty$.

Corollary 3.9.

Under the choice

$$c_0(s,a) = \frac{\mu(a \mid s) - \hat{\pi}_\beta(a \mid s)}{\hat{\pi}_\beta(a \mid s)} \tag{6}$$

, $p = 2$, α sufficiently large (satisfy the α condition in Theorem 3.6.), and any $g(\tau)$, for all $s \in \mathcal{S}$,

$$\mathbb{E}_{\hat{\pi}_\beta(a|s)} \Phi_g\left(\tilde{Z}^\pi(s,a)\right) - \mathbb{E}_{\mu(a|s)} \Phi_g\left(\tilde{Z}^\pi(s,a)\right) \geq \mathbb{E}_{\hat{\pi}_\beta(a|s)} \Phi_g\left(Z^\pi(s,a)\right) - \mathbb{E}_{\mu(a|s)} \Phi_g\left(Z^\pi(s,a)\right)$$

- high level: gap-expansion of integrals of the quantiles
- Together, Corollaries 3.7 & 3.9: CDE provides conservative lower bounds on the return quantiles while being less conservative for in-distribution actions

Appendix Proof

Lemma A.1.

$n(s,a) = |\{(s,a) \mid (s,a,r,s') \in \mathcal{D}\}|$: number of times (s,a) occurs in D.
For any return distribution Z with ζ-strongly monotone CDF $F_{Z(s,a)}$ and any $\delta \in \mathbb{R}_{>0}$, with probability at least $1-\delta$, for all $s \in \mathcal{D}$ and $a \in \mathcal{A}$, we have

$$\left\|F^{-1}_{\hat{\mathcal{T}}^\pi Z(s,a)} - F^{-1}_{\mathcal{T}^\pi Z(s,a)}\right\|_\infty \leq \Delta(s,a) \quad \text{where} \quad \Delta(s,a) = \frac{1}{\zeta}\sqrt{\frac{5|\mathcal{S}|}{n(s,a)} \log \frac{4|\mathcal{S}||\mathcal{A}|}{\delta}}$$

- We first prove a bound on the concentration of the empirical CDF to the true CDF (Lemma A.4., Lemma A.5.)

Lemma A.4.

For all $\delta \in \mathbb{R}_{>0}$, with probability at least $1-\delta$, for any $Z \in \mathcal{Z}$, for all $(s,a) \in \mathcal{D}$,

$$\left\| F_{\hat{\mathcal{T}}^\pi Z(s,a)} - F_{\mathcal{T}^\pi Z(s,a)} \right\|_\infty \leq \sqrt{\frac{5|\mathcal{S}|}{n(s,a)} \log \frac{4|\mathcal{S}||\mathcal{A}|}{\delta}} \tag{11}$$

- high level: bound on the concentration of the empirical CDF to the true CDF

Lemma A.4. Proof

- By the definition of distributional Bellman operator applied to the CDF function, we have $F_{\hat{\mathcal{T}}^\pi Z(s,a)}(x) - F_{\mathcal{T}^\pi Z(s,a)}(x)$

$$= \sum_{s',a'} \hat{P}\left(s' \mid s,a\right) \pi\left(a' \mid s'\right) F_{\gamma Z(s',a') + \hat{R}(s,a)}(x) - \sum_{s',a'} P\left(s' \mid s,a\right) \pi\left(a' \mid s'\right) F_{\gamma Z(s',a') + R(s,a)}(x)$$

- Adding and subtracting $\sum_{s',a'} \hat{P}\left(s' \mid s,a\right) \pi\left(a' \mid s'\right) F_{\gamma Z(s',a') + R(s,a)}(x)$ from this expression gives

$$\begin{aligned} &\sum_{s',a'} \hat{P}\left(s' \mid s,a\right) \pi\left(a' \mid s'\right) \left(F_{\gamma Z(s',a') + \hat{R}(s,a)}(x) - F_{\gamma Z(s',a') + R(s,a)}(x)\right) \\ &\quad + \sum_{s',a'} \left(\hat{P}\left(s' \mid s,a\right) - P\left(s' \mid s,a\right)\right) \pi\left(a' \mid s'\right) F_{\gamma Z(s',a') + R(s,a)}(x) \end{aligned}$$

- We proceed by bounding the two terms in the summation.
- For the first term, observe that

$$\begin{aligned} &F_{\gamma Z(s',a') + \hat{R}(s,a)}(x) - F_{\gamma Z(s',a') + R(s,a)}(x) \\ &= \int \left[F_{\hat{R}(s,a)}(r) - F_{R(s,a)}(r)\right] dF_{\gamma Z(s',a')}(x-r) \quad \text{(convolution form)} \\ &\leq \int \left|F_{\hat{R}(s,a)}(r) - F_{R(s,a)}(r)\right| dF_{\gamma Z(s',a')}(x-r) \quad \text{(integrate a larger } f(x)) \\ &\leq \sup_r \left|F_{\hat{R}(s,a)}(r) - F_{R(s,a)}(r)\right| \int dF_{\gamma Z(s',a')}(x-r) \quad \text{(take maximum term)} \\ &= \left\|F_{\hat{R}(s,a)}(r) - F_{R(s,a)}(r)\right\|_\infty \quad \text{(latter term is 1, by probability axiom)} \end{aligned}$$

- Therefore, we have

$$\begin{aligned} &\sum_{s',a'} \hat{P}\left(s' \mid s,a\right) \pi\left(a' \mid s'\right) \left(F_{\gamma Z(s',a') + \hat{R}(s,a)}(x) - F_{\gamma Z(s',a') + R(s,a)}(x)\right) \\ &\leq \sum_{s',a'} \hat{P}\left(s' \mid s,a\right) \pi\left(a' \mid s'\right) \left\|F_{\hat{R}(s,a)}(r) - F_{R(s,a)}(r)\right\|_\infty \\ &= \left\|F_{\hat{R}(s,a)}(r) - F_{R(s,a)}(r)\right\|_\infty \quad \text{(former term is 1, by probability axiom)} \end{aligned}$$

- The second term can be bounded as follows:

$$
\begin{aligned}
&\sum_{s',a'}\left(\hat{P}\left(s' \mid s,a\right)-P\left(s' \mid s,a\right)\right)\pi\left(a' \mid s'\right)F_{\gamma Z(s',a')+R(s,a)}(x) \\
&=\sum_{s'}\left(\hat{P}\left(s' \mid s,a\right)-P\left(s' \mid s,a\right)\right)\sum_{a'}\pi\left(a' \mid s'\right)F_{\gamma Z(s',a')+R(s,a)}(x) \\
&\leq\|\hat{P}(\cdot \mid s,a)-P(\cdot \mid s,a)\|_1 \cdot\left\|\sum_{a'}\pi\left(a' \mid \cdot\right)F_{\gamma Z(\cdot,a')+R(s,a)}(x)\right\|_\infty \\
&\leq\|\hat{P}(\cdot \mid s,a)-P(\cdot \mid s,a)\|_1 \cdot\left\|\sum_{a'}\pi\left(a' \mid \cdot\right)\right\|_\infty \\
&=\|\hat{P}(\cdot \mid s,a)-P(\cdot \mid s,a)\|_1
\end{aligned}
$$

- Together, we have

$$
\left|F_{\hat{\mathcal{T}}^\pi Z(s,a)}(x)-F_{\mathcal{T}^\pi Z(s,a)}(x)\right| \leq\left\|F_{\hat{R}(s,a)}(r)-F_{R(s,a)}(r)\right\|_\infty+\left\|\hat{P}\left(s' \mid s,a\right)-P\left(s' \mid s,a\right)\right\|_1
$$

- By the DKW inequality, we have that with probability $1-\delta/2$, for all $(s,a)\in\mathcal{D}$,

$$
\left\|F_{\hat{R}(s,a)}(r)-F_{R(s,a)}(r)\right\|_\infty \leq\sqrt{\frac{1}{2n(s,a)}\ln\frac{4|\mathcal{S}||\mathcal{A}|}{\delta}}
$$

- Similarly, by Hoeffding's inequality and an ℓ_1 concentration bound for multinomial distribution, we have

$$
\max_{s,a}\|\hat{P}(\cdot \mid s,a)-P(\cdot \mid s,a)\|_1 \leq\sqrt{\frac{2|\mathcal{S}|}{n(s,a)}\ln\frac{4|\mathcal{S}||\mathcal{A}|}{\delta}}
$$

- The claim follows by combining the two inequalities.

Lemma A.5.

> Lemma A.5. Consider two CDFs F and G with support $\mathcal{X}$. Suppose that F is ζ-strongly monotone and that $\|F-G\|_\infty \leq \epsilon$. Then, $\left\|F^{-1}-G^{-1}\right\|_\infty \leq \epsilon/\zeta$.

- it says that if F and G is close, then F^{-1} and G^{-1} is not far away

Lemma A.5. Proof

- First, note that

$$
F^{-1}(y)-G^{-1}(y)=\int_{G^{-1}(y)}^{F^{-1}(y)} dx=\int_{F(G^{-1}(y))}^{y} dF^{-1}\left(y'\right)
$$

- first equality: by fundamental theorem of calculus
- the second: by a change of variable $y'=F(x)$
- Since $F\left(F^{-1}\left(y'\right)\right)=y'$, we have $F'\left(F^{-1}\left(y'\right)\right)dF^{-1}\left(y'\right)=dy'$, so

$$
dF^{-1}\left(y'\right)=\frac{dy'}{F'\left(F^{-1}\left(y'\right)\right)} \leq\frac{dy'}{\zeta}
$$

- inequality: by ζ-strong monotonicity

- As a consequence, we have

$$\int_{F(G^{-1}(y))}^{y} dF^{-1}\left(y'\right) \leq \int_{F(G^{-1}(y))}^{y} \frac{dy'}{\zeta} = \frac{\left(y - F\left(G^{-1}(y)\right)\right)}{\zeta} = \frac{G\left(G^{-1}(y)\right) - F\left(G^{-1}(y)\right)}{\zeta} \leq \frac{\epsilon}{\zeta}$$

- where the last inequality: $\|G - F\|_{\infty} \leq \epsilon$
- The Lemma A.5. follows

Lemma A.1. Proof

- substituting $F = F_{\hat{\mathcal{T}}^{\pi} Z(s,a)}(x), G = F_{\mathcal{T}^{\pi} Z(s,a)}(x)$, and $\epsilon = \sqrt{\frac{5|\mathcal{S}|}{n(s,a)}} \log \frac{4|\mathcal{S}||\mathcal{A}|}{\delta}$ into Lemma A.5. where the condition $\|F - G\|_{\infty} \leq \epsilon$ holds by Lemma A.4.
- Lemma A. 1 follows

Lemma A.2.

If Z satisfies $\left\|F_{Z(s,a)}^{-1} - F_{\mathcal{T} Z(s,a)}^{-1}\right\|_{\infty} \leq \beta$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$, then

$$\left\|F_{Z(s,a)}^{-1} - F_{Z^{\pi}(s,a)}^{-1}\right\|_{\infty} \leq (1 - \gamma)^{-1} \beta \quad (\forall s \in \mathcal{S}, a \in \mathcal{A})$$

- high level: relates one-step distributional Bellman contraction to an ∞-norm bound at the fixed point

Lemma A.2. Proof

- We prove the following slightly stronger result

Lemma A.6.

For any $\beta \in \mathbb{R}$, if Z satisfies

$$F_{Z(s,a)}^{-1}(\tau) \geq F_{\mathcal{T}^{\pi}}^{-1} Z(s,a)(\tau) + \beta \quad (\forall \tau \in [0,1]) \tag{12}$$

for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$, then we have

$$F_{Z(s,a)}^{-1}(\tau) \geq F_{Z^{\pi}(s,a)}^{-1}(\tau) + (1 - \gamma)^{-1} \beta \quad (\forall \tau \in [0,1])$$

The result holds with $\geq$ replaced by $\leq$, or with $\mathcal{T}^{\pi}$ and Z^{π} replaced by $\hat{\mathcal{T}}^{\pi}$ and $\hat{Z}^{\pi}$ or $\tilde{\mathcal{T}}^{\pi}$ and $\tilde{Z}^{\pi}$.

Lemma A.6. Proof

- We prove the first case
- the cases with $\geq$, and the cases with $\hat{\mathcal{T}}^{\pi}$ and $\hat{Z}^{\pi}$ follow by the same argument

- First, we show that

$$F_{\mathcal{T}^\pi Z(s,a)}(x) \geq F_{Z(s,a)}(x+\beta) \quad (\forall x \in [V_{\min}, V_{\max}]) \tag{13}$$

- To this end, note that rearranging (12), we have

$$\bar{F}_{\mathcal{T}^\pi Z(s,a)}\left(F^{-1}_{Z(s,a)}(\tau) - \beta\right) \geq \tau$$

- Then, substituting $\tau = F_{Z^\pi \hat{(s,a)}}(x+\beta)$ yields (13)
- Next, we shot that

$$F^{-1}_{\mathcal{T}^\pi Z(s,a)}(\tau) \geq F^{-1}_{\mathcal{T}^\pi(\mathcal{T}^\pi Z(s,a))}(\tau) + \gamma\beta \quad (\forall \tau \in [0,1]) \tag{14}$$

- Intuitively, this claim says that $\mathcal{T}^\pi$ distributes additively to the constant β, and since $\mathcal{T}^\pi$ is a γ-contraction in $\bar{d}_p$, we have $\mathcal{T}^\pi\beta \leq \gamma\beta$
- To show (14), first note that

$$\begin{aligned}
F_{\mathcal{T}^\pi(\mathcal{T}^\pi Z(s,a))}(x) &= \sum_{s',a'} P^\pi\left(s', a' \mid s, a\right) \int F_{\mathcal{T}^\pi Z(s',a')}\left(\frac{x-r}{\gamma}\right) dF_{R(s,a)}(r) \\
&\geq \sum_{s',a'} P^\pi\left(s', a' \mid s, a\right) \int F_{Z(s',a')}\left(\frac{x-r}{\gamma} + \beta\right) dF_{R(s,a)}(r) \\
&= \sum_{s',a'} P^\pi\left(s', a' \mid s, a\right) \int F_{\gamma Z(s',a')}(x - r + \gamma\beta) dF_{R(s,a)}(r) \\
&= \sum_{s',a'} P^\pi\left(s', a' \mid s, a\right) F_{R(s,a)+\gamma Z(s',a')}(x+\gamma\beta) \\
&= F_{\mathcal{T}^\pi Z(s,a)}(x+\gamma\beta)
\end{aligned}$$

- where the first step follows by derivation of the Bellman operator for the CDF, the second step follows from (13), and the third step follows from the property of a CDF function. It follows that

$$F^{-1}_{\mathcal{T}^\pi Z(s,a)}\left(F_{\mathcal{T}^\pi(\mathcal{T}^\pi Z(s,a))}(x)\right) \geq x + \gamma\beta$$

- Setting $\tau = F_{\mathcal{T}^\pi(\mathcal{T}^\pi Z(s,a))}(x)$, we have

$$F^{-1}_{\mathcal{T}^\pi Z(s,a)}(\tau) \geq F^{-1}_{\mathcal{T}^\pi(\mathcal{T}^\pi Z(s,a))}(\tau) + \gamma\beta$$

for all $\tau \in [0,1]$

- Thus, we have shown (14). Now, by induction on $\mathcal{T}^\pi$, we have

$$F^{-1}_{(\mathcal{T}^\pi)^k Z(s,a)}(\tau) \geq F^{-1}_{(\mathcal{T}^\pi)^{k+1} Z(s,a)}(\tau) + \gamma^k\beta$$

for all $k \in \mathbb{N}$

- Summing these inequalities over $k \in \{0, 1, \ldots, n\}$ inequality gives

$$\sum_{k=0}^{n} F^{-1}_{(\mathcal{T}^\pi)^k Z(s,a)}(\tau) \geq \sum_{k=0}^{n} F^{-1}_{(\mathcal{T}^\pi)^k(\mathcal{T}^\pi Z(s,a))}(\tau) + \sum_{k=0}^{n} \gamma^k\beta$$

Subtracting common terms from both sides and evaluating the sum over γ^k, we have

$$F^{-1}_{Z(s,a)}(\tau) \geq F^{-1}_{(\mathcal{T}^\pi)^{n+1} Z(s,a)}(\tau) + \frac{1-\gamma^{n+1}}{1-\gamma}\beta$$

- Taking $n \to \infty$, we have

$$F^{-1}_{Z(s,a)}(\tau) \geq F^{-1}_{Z^\pi(s,a)}(\tau) - (1-\gamma)^{-1}\beta$$

- where we have used the fact that Z^π is the fixed point of $\mathcal{T}^\pi$
- The lemma A.6. follows.

Theorem A.7.

We have $\left\| F^{-1}_{\hat{Z}^\pi(s,a)} - F_{Z^\pi(s,a)} \right\|_\infty \leq (1-\gamma)^{-1}\Delta_{\max}$, where $\hat{Z}^\pi$ and Z^π are the fixed-points of $\hat{\mathcal{T}}^\pi$ and $\mathcal{T}^\pi$, respectively.

- high level: Bound on error of the fixed-point of the empirical distributional bellman operator

Theorem A.7. Proof

- Let $\Delta_{\max} = \max_{s,a} \Delta(s,a)$. We have $\left\| F^{-1}_{\hat{Z}^\pi(s,a)} - F_{\mathcal{T}^\pi \hat{Z}^\pi(s,a)} \right\|_\infty \leq \Delta_{\max}$ by Lemma A. 1 with $Z = \hat{Z}^\pi$
- Thus, we have $\left\| F^{-1}_{\hat{Z}^\pi(s,a)} - F_{Z^\pi(s,a)} \right\|_\infty \leq (1-\gamma)^{-1}\Delta_{\max}$ by Lemma A. 2

Discussions and Takeaways

- we train

$$\tilde{Z}^{k+1} = \arg\min_Z \alpha \cdot \mathbb{E}_{U(\tau),\mathcal{D}(s,a)}\left[c_0(s,a) \cdot F^{-1}_{Z(s,a)}(\tau)\right] + \mathcal{L}_p\left(Z, \hat{\mathcal{T}}^\pi \tilde{Z}^k\right)$$

- as our return distribution
- and we told a lot of reasons why the estimator is nice
- their proof is based on distributional Bellman operator
- Also, by Lemma 3.4., we can iterately compute $\tilde{Z}^{k+1}$
- This guarantees that the runtime should not be too bad.

References

- Conservative Offline Distributional Reinforcement Learning (https://arxiv.org/abs/2107.06106)