Distributional Bellman Operators over Mean Embeddings

Li Kevin Wenliang, Grégoire Delétang, Matthew Aitchison, Marcus Hutter, Anian Ruoss, Arthur Gretton, Mark Rowland

Google DeepMind · UCL · GATSBY

- A new family of distributional Bellman operators for distributional reinforcement learning
- Return distributions encoded as mean embeddings, Bellman updates performed in ME space
- The algorithm is simple, fast and elegant
- Convergence by error analysis, nice Atari suite result

Distributional Bellman Consistency

Consider policy evaluation over a Markov reward process

$$x \in \mathcal{X}, \quad a \in \mathcal{A}, \quad P_X(X|x,a), \quad P_R(R|x,a), \quad \gamma$$

states *actions* *transition* *reward* *discount*

Trajectory under policy π: $X_0 = x,\ A_0,\ R_0,\ X_1,\ A_1,\ R_1,\ \dots$

In *value-based* RL, we estimate the **value** for each state

$$V^\pi(x) := \mathbb{E}_x^\pi[G^\pi(x)], \quad G^\pi(x) := \sum_{t=0}^{\infty} \gamma^t R_t \ \mid X_0 = x$$

In *distributional* RL we estimate the return **distribution**

$$\eta^\pi(x) := \mathrm{Law}(G^\pi(x)) \quad \Leftrightarrow \quad G^\pi(x) \sim \eta^\pi(x)$$

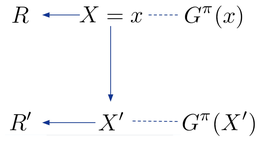
Bellman consistencies: for a transition x, A, R, X'

Value consistency

$$V^\pi(x) = \mathbb{E}_x^\pi[R + \gamma V^\pi(X')]$$

Distributional consistency

$$G^\pi(x) \overset{d}{=} R + \gamma G^\pi(X')$$



Distributional Bellman Update

Dynamical programming (DP)

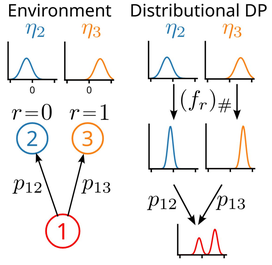
$$V(x) \leftarrow \mathbb{E}_x^\pi[R + \gamma V(X')]$$

Easy to implement with vectors

Distributional DP (Bellemare+ 23)

$$G(x) \overset{?}{\leftarrow} R + \gamma G(X')$$

How to implement?



Main Idea: Bellman + Mean Embeddings

The mean embedding (ME) encodes distributions by feature functions (statistical functional or sketch):

$$\text{for } \begin{cases} \eta \in \mathscr{P}(\mathbb{R}), \\ \phi : \mathbb{R} \to \mathbb{R}^m, \end{cases} \quad \text{define } u := \mathbb{E}_{Z\sim\eta}[\phi(Z)]\,.$$

Let's write the Bellman consistency using ME

$$U^\pi(x) := \mathbb{E}_x^\pi[\phi(G^\pi(x))] = \mathbb{E}_x^\pi[\phi(R + \gamma G^\pi(X'))]\,.$$

If the feature function satisfies:

$$\phi(r + \gamma z) = B_r \phi(z) \quad \text{for some } B_r \in \mathbb{R}^{m\times m}, \qquad (2)$$

then we have a "closed" expression:

$$\begin{aligned} U^\pi(x) &= \mathbb{E}_x^\pi[\phi(R + \gamma G^\pi(X'))] = \mathbb{E}_x^\pi[B_R \phi(G^\pi(X'))] \\ &= \mathbb{E}_x^\pi[B_R U^\pi(X')]\,. \end{aligned}$$

Leading to Sketch-DP $U(x) \leftarrow \mathbb{E}_x^\pi[B_R U(X')]$

On the Big If...

Does the assumption in (2) hold?

Rowland+ (2019) showed that only *polynomials* satisfy this assumption.

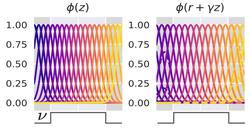
$$\begin{bmatrix} 1 \\ r+\gamma g \\ (r+\gamma g)^2 \\ \vdots \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \cdots \\ r & \gamma & 0 & \cdots \\ r & 2r\gamma & \gamma^2 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} 1 \\ g \\ g^2 \\ \vdots \end{bmatrix}$$

BUT let's relax (2) to an approximation: $\phi(r + \gamma z) \approx B_r \phi(z)$

To ensure small error, solve linear regression for each r

$$\min_{B\in\mathbb{R}^{m\times m}} \mathbb{E}_{Z\sim\nu} \|B\phi(Z) - \phi(r + \gamma Z)\|_2^2$$

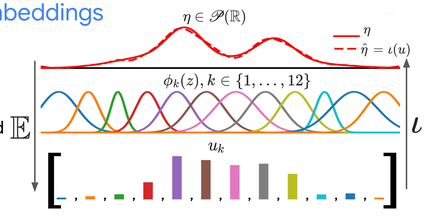
$$B_r = \mathbb{E}_\nu[\phi(r + \gamma Z)\phi(Z)^\top]\mathbb{E}_\nu[\phi(Z)\phi(Z)^\top]^{-1}$$



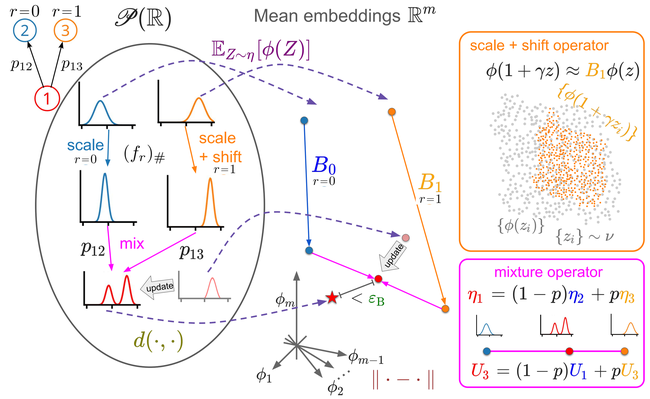
More on Mean Embeddings

$$u_k = \mathbb{E}_{Z\sim\eta}[\phi_k(Z)]$$

ι requires a constrained optimisation,



Algorithms: Sketch-DP and Sketch-TD



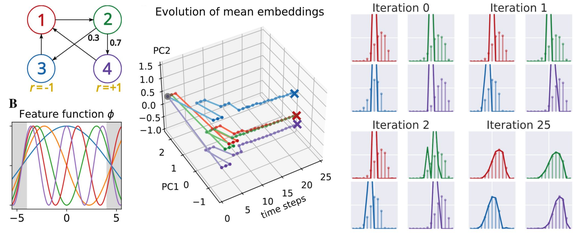
Simple DRL updates:

Initialise $U(x) \in \mathbb{R}^m, \ \forall x$

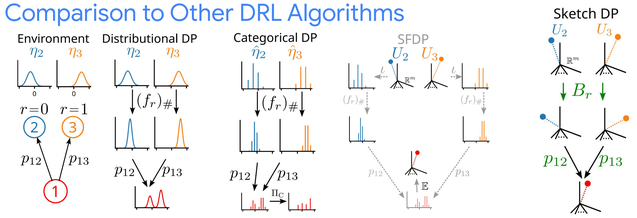
– DP update : $U(x) \leftarrow \mathbb{E}_x^\pi[B_R U(X')]$

– TD update on $x \to r \to x'$: $U(x) \leftarrow (1-\alpha)U(x) + \alpha B_r U(x')$

Example Sketch-DP Run



Comparison to Other DRL Algorithms



Convergence: Error Analysis

Intuitively, to ensure convergence to U^π, we want ϕ to be "rich":

1. approximate shift+scale versions well;
2. retain sufficient distributional information.

Informal theorem: *Sketch-DP iterates → neighbourhood of U^π if:*

1: Feature approx. error bounds distributional update error

$$\begin{aligned} &\|\mathbb{E}_x^\pi[\phi(R + \gamma G(X'))] - \mathbb{E}_x^\pi[B_R U(X')]\| \\ &\quad < \sup_{g\in[G_{\min}, G_{\max}]} \max_r \|\phi(r + \gamma g) - B_r \phi(g)\| \\ &\quad =: \varepsilon_{\mathrm{B}} \end{aligned}$$

2: U and U' of two return distributions η and η' must satisfy:

a. there exists a distributional metric d on which the distributional Bellman operator contracts;

b. $\exists\, \varepsilon_1, \varepsilon_2 > 0$ s.t. $d(\eta, \eta') - \varepsilon_1 \le \|U - U'\| \le d(\eta, \eta') + \varepsilon_2\,.$

See paper for a concrete example using a particular sketch.

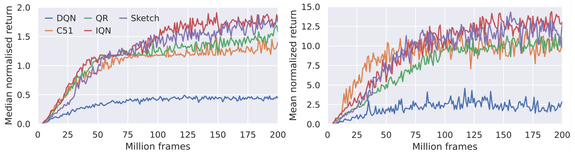
Experiments

Main Atari Suite

ε-greedy policy, value readout $Q(x,a) = \beta \cdot U(x,a)$

$$\arg\min_\beta \mathbb{E}_{G\sim\mu}[(G - \langle\beta, \phi(G)\rangle)^2]$$

sigmoid features ϕ, based on QR-DQN architecture



Synthetic Experiments

A few simple environments (column), feature types (hue)

Cramer distance $\max_x \ell_2[\iota(U(x)), \eta^\pi(x)]$