A Distributional Perspective on Reinforcement Learning

Ruslan Khaidurov

The 13th of March

Outline

RL Recap

- s for state, a for action, π for policy, r for reward.
- $\pi(s, a) = \pi(a|s)$ is a distribution over actions in a fixed state s.
- Discounted return:

$$G(s_k, a_k) = \sum_{i=0}^{\infty} \gamma^i r(s_{k+i}, a_{k+i}), \ \gamma \in [0, 1]$$

- Value function (expected discounted return):

$$Q_\pi(s, a) = \mathrm{E}_\pi[G(s_k, a_k)|s_k = s, a_k = a]$$

RL Recap

Optimal expected value: $Q^*(s, a) = \max_\pi Q_\pi(s, a)$.
Optimal policy: $\pi^*(s, a) = \arg\max_\pi Q_\pi(s, a)$
How about approximating $Q^*(s, a)$ with a neural network? (spoiler: naive approach doesn't work)

The setting

- $(\mathcal{S}, \mathcal{A}, R, \mathrm{P}, \gamma)$ — Markov decision process;
- $\mathcal{S}$ for state space, $\mathcal{A}$ for action space, R for reward function;
- P for transition kernel:
  $\mathrm{P}(s_{k+1}|s_k, a_k, \ldots, s_0, a_0) = \mathrm{P}(s_{k+1}|s_k, a_k)$, $\gamma \in [0, 1]$.

Bellman equations

- Fundamental result in reinforcement learning is to describe the value function like this:

$$Q_\pi(s, a) = \mathsf{E}R(s, a) + \gamma \mathsf{E}_{\mathsf{P},\pi} Q_\pi(s', a')$$

- Sometimes it is useful to rewrite it in the operator form:

$$\mathcal{T}_\pi Q(s, a) := \mathsf{E}R(s, a) + \gamma \mathsf{E}_{\mathsf{P},\pi} Q(s', a')$$

$$\mathcal{T} Q(s, a) := \mathsf{E}R(s, a) + \gamma \mathsf{E}_{\mathsf{P},\pi} \max_{a' \in \mathcal{A}} Q(s', a')$$

  and to find a fixed point of these operators.
- $\mathcal{T}_\pi$ and $\mathcal{T}$ are called Bellman's operator and Bellman's optimality operator respectively.

Recap: Wesserstein metric

Geven two distributions F and G in the probability space $(\Omega, \mathcal{F}, \mathrm{P})$. Let $U \sim F$.

- $||U||_p = \left(\mathrm{E}[||U(\omega)||_p^p]\right)^{1/p}$ — the norm of a random variable;
- Wasserstein metric:

$$d_p(F, G) = \inf_{U \sim F, V \sim G} ||U - V||_p$$

- For $p < \infty$ it can be explicitly written as:

$$d_p(F, G) = \left(\int_0^1 |F^{-1}(q) - G^{-1}(q)| dq \right)^{1/p}$$

Let's go beyond!

- The random return:

$$Z_\pi(s, a) = R(s, a) + \gamma Z_\pi(s', a')$$

  is a sum of random reward $R(s, a)$ and a random value of a random transition $s' \sim P(\cdot|s, a)$, $a' \sim \pi(s, a)$.
- How about model $Z_\pi(s, a)$ instead of $Q_\pi(s, a)$?

Recap: Wesserstein metric

Let $\mathcal{Z}$ be a space of all value distributions with bounded moments. For any $Z_1, Z_2, Z_3 \in \mathcal{Z}$ let

$$\overline{d_p}(Z_1, Z_2) = \sup_{s,a} d_p(Z_1(s,a), Z_2(s,a))$$

We can prove that $\overline{d_p}$ is a metric!

1. $\overline{d_p}(Z_1, Z_2) \geqslant 0$
2. $\overline{d_p}(Z_1, Z_2) = 0 \Leftrightarrow Z_1 = Z_2$
3. $\overline{d_p}(Z_1, Z_2) = \overline{d_p}(Z_2, Z_1)$
4. $\overline{d_p}(Z_1, Z_3) \geqslant \overline{d_p}(Z_1, Z_2) + \overline{d_p}(Z_2, Z_3)$

Distributional Bellman Operators

Just like in ordinary Bellman operators:

- $P_\pi Z(s,a) :\overset{d}{=} Z(s',a'),\ s' \sim P(\cdot|s,a),\ a' \sim \pi;$
- $\mathcal{T}_\pi Z(s,a) :\overset{d}{=} R(s,a) + \gamma P_\pi Z(s,a).$

Control setting

A greedy policy maximizes the expectation of $Q(s, a)$:

$$\pi^* \text{ is greedy } \Leftrightarrow \mathsf{E}_{\mathsf{P},\pi^*} Z(s, a) = \mathsf{E}_{\mathsf{P}} \max_{a' \in \mathcal{A}} Z(s', a');$$

Distributional Bellman operator:

$$\mathcal{T} Z = \mathcal{T}_\pi Z \text{ for some greedy policy } \pi$$

Let $\{Z_k\}_{k=1}^\infty$ be a sequence of value distributions such that

$$Z_{k+1} = \mathcal{T} Z_k$$

Then $Q_k(s, a) = \mathsf{E} Z_k(s, a)$ converges uniformly to Q^* exponentially fast in L_∞ metric.

Control setting

We may expect that $\mathcal{T}$ has a unique fixed point Z^*.

Control setting

We may expect that $\mathcal{T}$ has a unique fixed point Z^*. (not really)

- $\mathcal{T}$ is not a contraction;
- Not all optimality operators $\mathcal{T}$ have a unique fixed point.

Control setting

We may expect that $\mathcal{T}$ has a unique fixed point Z^*. (not really)

- $\mathcal{T}$ is not a contraction;
- Not all optimality operators $\mathcal{T}$ have a unique fixed point.

All we can expect is convergence to a set of optimal value distributions in $\overline{d_p}$ metric.

Approximate distributional learning

How about model a discrete parametric distribution with parameters $N \in \mathbb{N}, V_{min}, V_{max} \in \mathbb{R}$, $\Delta z = (V_{max} - V_{min})/N$, $z_i = V_{min} + i\Delta z$:

$$Z_\theta(s, a) = z_i \text{ w.p. } p_i = \frac{\exp(\theta_i(s, a))}{\sum_j \exp(\theta_j(s, a))}$$

Where $\theta \colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}^N$.

Approximate distributional learning

How about model a discrete parametric distribution with parameters $N \in \mathbb{N}, V_{min}, V_{max} \in \mathbb{R}$, $\Delta z = (V_{max} - V_{min})/N$, $z_i = V_{min} + i\Delta z$:

$$Z_\theta(s, a) = z_i \text{ w.p. } p_i = \frac{\exp(\theta_i(s, a))}{\sum_j \exp(\theta_j(s, a))}$$

Where $\theta \colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}^N$.
Then hope that it will converge to Z^* such that $\mathrm{E}Z^* = Q^*$.

Approximate distributional learning

How about model a discrete parametric distribution with parameters $N \in \mathbb{N}, V_{min}, V_{max} \in \mathbb{R}$, $\Delta z = (V_{max} - V_{min})/N$, $z_i = V_{min} + i\Delta z$:

$$Z_\theta(s, a) = z_i \text{ w.p. } p_i = \frac{\exp(\theta_i(s, a))}{\sum_j \exp(\theta_j(s, a))}$$

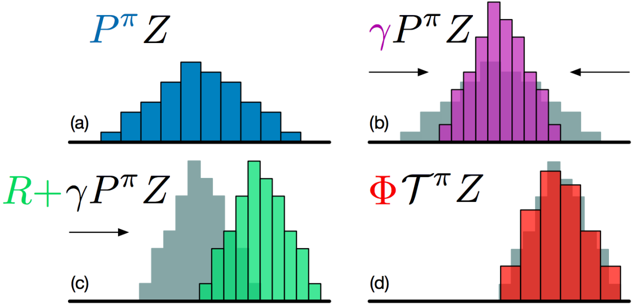
Where $\theta \colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}^N$.
Then hope that it will converge to Z^* such that $\mathrm{E}Z^* = Q^*$.

Approximate distributional learning

In fact it does not work as $\mathcal{T} Z_\theta$ and Z_θ almost always have disjoint supports.

Approximate distributional learning

In fact it does not work as $\mathcal{T} Z_\theta$ and Z_θ almost always have disjoint supports.
But let's project with an operator Φ support of $\mathcal{T} Z_\theta \to$ support of Z_θ!



The algorithm

Algorithm 1 Categorical Algorithm

---

input A transition $x_t, a_t, r_t, x_{t+1}, \gamma_t \in [0, 1]$

$Q(x_{t+1}, a) := \sum_i z_i p_i(x_{t+1}, a)$

$a^* \leftarrow \arg\max_a Q(x_{t+1}, a)$

$m_i = 0, \quad i \in 0, \ldots, N-1$

for $j \in 0, \ldots, N-1$ **do**

Compute the projection of $\hat{\mathcal{T}} z_j$ onto the support $\{z_i\}$

$\hat{\mathcal{T}} z_j \leftarrow [r_t + \gamma_t z_j]_{V_{\text{MIN}}}^{V_{\text{MAX}}}$

$b_j \leftarrow (\hat{\mathcal{T}} z_j - V_{\text{MIN}})/\Delta z$ # $b_j \in [0, N-1]$

$l \leftarrow \lfloor b_j \rfloor, u \leftarrow \lceil b_j \rceil$

Distribute probability of $\hat{\mathcal{T}} z_j$

$m_l \leftarrow m_l + p_j(x_{t+1}, a^*)(u - b_j)$

$m_u \leftarrow m_u + p_j(x_{t+1}, a^*)(b_j - l)$

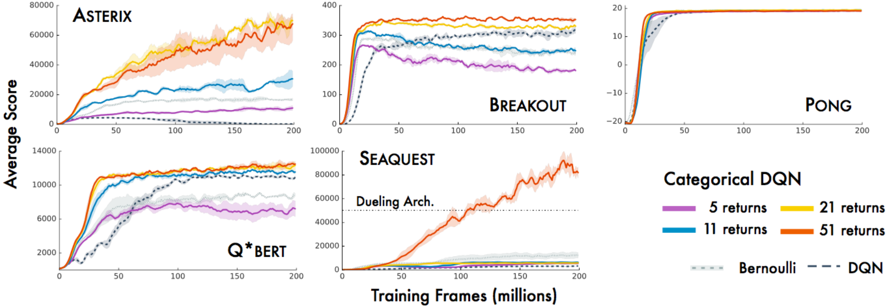
end for

output $-\sum_i m_i \log p_i(x_t, a_t)$ # Cross-entropy loss

Experimental setting

- DQN which predicts $p_i(s, a)$;
- ε-greedy policy over the expected action-values;
- $V_{min} = -10$, $V_{max} = 10$.

Experiments



Varying N.

State-of-the-art results

	Mean	Median	> H.B.	> DQN
DQN	228%	79%	24	0
DDQN	307%	118%	33	43
DUEL.	373%	151%	37	50
PRIOR.	434%	124%	39	48
PR. DUEL.	592%	172%	39	44
C51	**701%**	**178%**	**40**	**50**
UNREAL†	880%	250%	-	-

Average performance on Atari 57 games compared to human baseline (C51 is an agent with $N = 51$).

Problems of this approach

- Instability in Bellman optimality operator;

Problems of this approach

- Instability in Bellman optimality operator;
- In fact we don't minimize Wasserstein metric (but KL-divergence);

Summary

- We are trying approximate value distribution instead of it's expectation (which is exactly a value function);
- Any sequence of Bellman-operator value distributions converges to a set of optimal distributions (but in fact not uniformly);
- Outputs of a DQN are parameters of a modelled discrete distribution;

Example

http://youtu.be/yFBwyPuO2Vg

Thank you for your attention

Read more here: `https://arxiv.org/pdf/1707.06887.pdf`