PYCON KOREA 2018

RL Adventure

Distributional RL

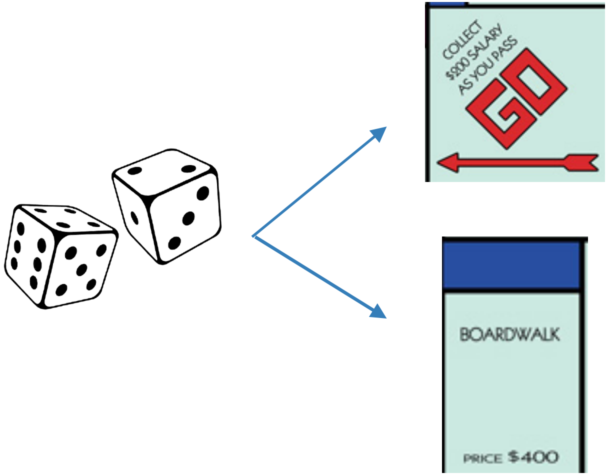
이의령

C51
Distributional RL

목차

1. Motivation

Motivation



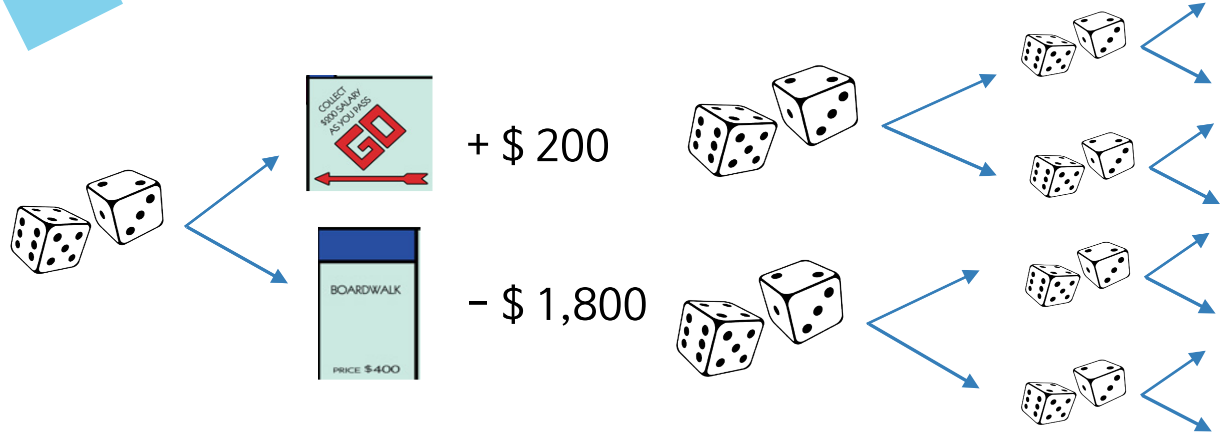
+ $ 200

- $ 1,800

$$E\,[R(x)] = \frac{35}{36} \times 200 - \frac{1}{36} \times 1{,}800$$

$$= 144$$

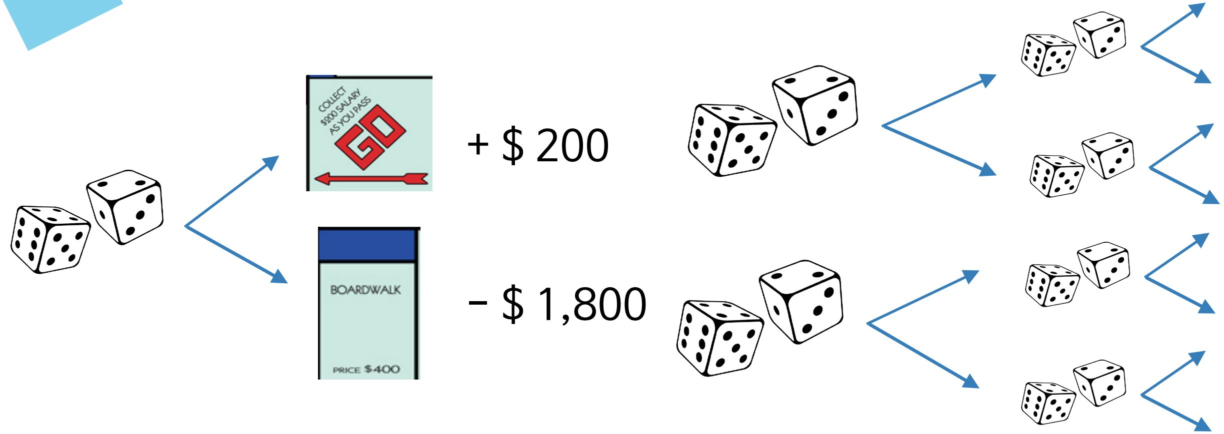
Motivation



보상의 합

$$R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{T-t-1} R_T$$

Expected RL



벨만 방정식

$$v(\mathcal{X}) = \boldsymbol{E}\,[R_{t+1} + \gamma R_{t+2} + \cdots \mid S_t = \mathcal{X}]$$

$$= \boldsymbol{E}\,[R_{t+1} + \gamma\, v(x) \mid S_t = \mathcal{X}]$$

$$= \boldsymbol{E}\ \mathrm{R}(\mathrm{x}) + \gamma\ \boldsymbol{E}\ v(x)$$

Expected RL

Reward를 Random Variable 관점에서 바라보면…

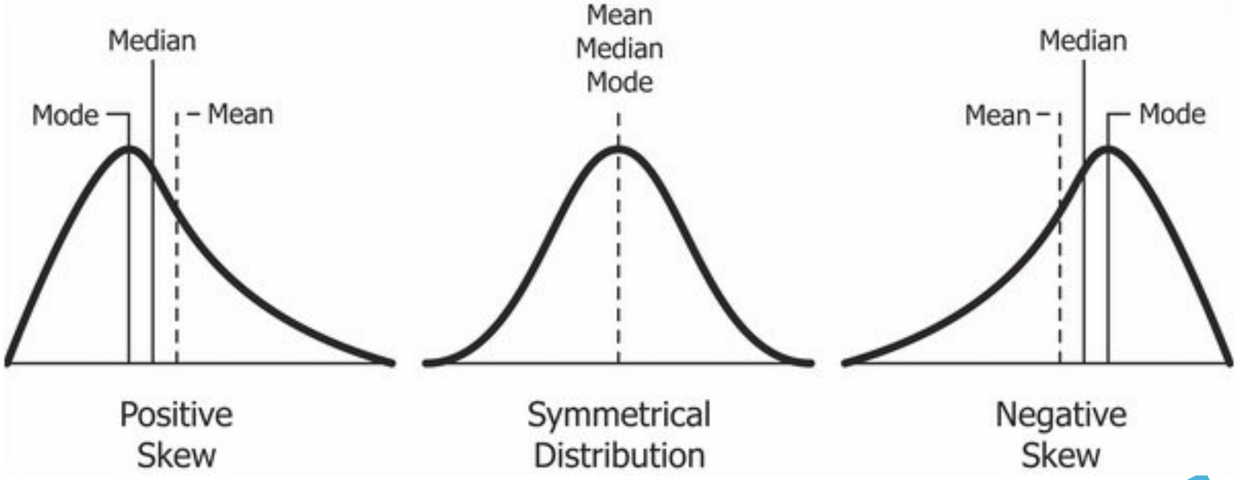
- 가치함수는 discount된 미래 보상에 대한 기댓값을 리턴한다.
- 기댓값 = Scalar(o) / Distribution(x)
- 미래 보상 값들은 complex, Multimodal의 특성을 가진다.
- 기댓값은 각 보상들이 가지는 intrinsic(본질적인)한 특성을 담아내지 못한다.

$$E\,[R(x)] = \frac{35}{36} \times 200 - \frac{1}{36} \times 1{,}800$$

$$= 144$$

Expected RL

Reward를 Random Variable 관점에서 바라보면...



이러한 Expected RL의 한계점을 보완책

-> A Distributional Perspective on RL (C51)

Return을 Distribution으로 만들어

Randomness한 특성과 정보를 최대한 반영해보자

$$V^{\pi} = E[\,Z^{\pi}(x)\,] = E[\,R(x)\,] + E[\,Z^{\pi}(X')\,]$$

Return을 Distribution으로 만들어

Randomness한 특성과 정보를 최대한 반영해보자

$$V^{\pi} = \cancel{E}[Z^{\pi}(x)] = \cancel{E}[R(x)] + \cancel{E}[Z^{\pi}(X')]$$

$$Z^{\pi}(x) = R(x) + Z^{\pi}(X')$$

2. Distributional RL

Distributional RL

A Distributional Perspective on Reinforcement Learning (C51)

https://arxiv.org/abs/1707.06887

- Expected RL → Distributional RL
- Return에 대한 Value Distribution을 만들자.
- C51 = Categorical / 이산형 분포
- 51개의 bin을 이용하여 분포를 만든다.

Distributional RL

A Distributional Perspective on Reinforcement Learning (C51)

- Distributional Bellman Equation

$$Z(x,a) \overset{D}{=} R(x,a) + \gamma Z(X', A')$$

- Cf) Bellman Equation

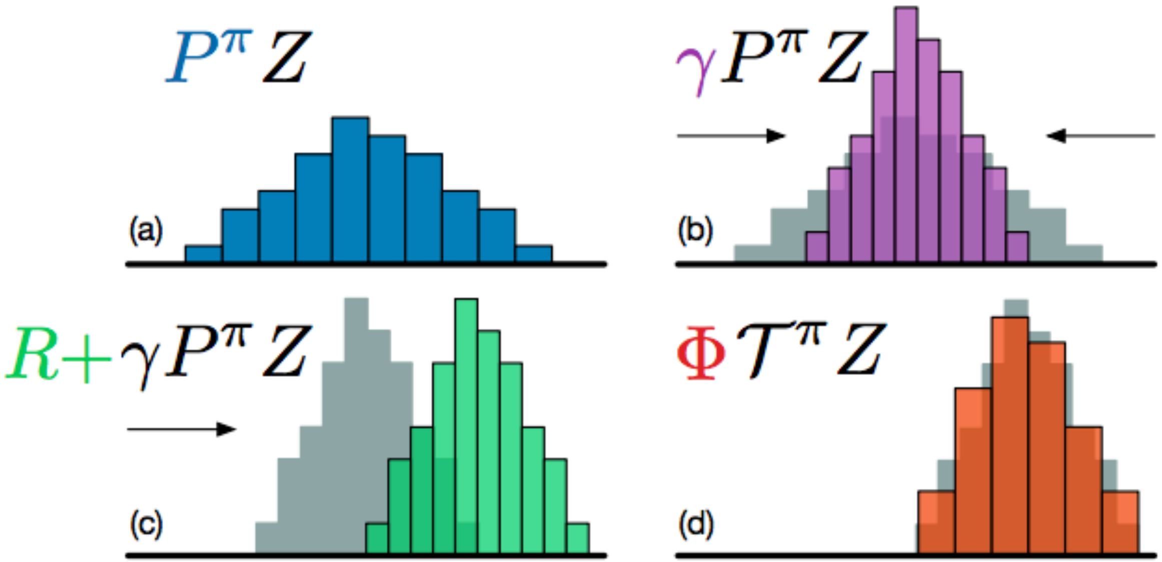
$$Q(x,a) = R(x,a) + \gamma Q_{\pi}(x', a')$$

- $Z(s,a)$ 는 Distribution을 의미, 이를 이용하여 Distribution을 생성

$$Q(s,a) = E[Z(s,a)] = \sum_{i=1}^{N} p_i x_i$$

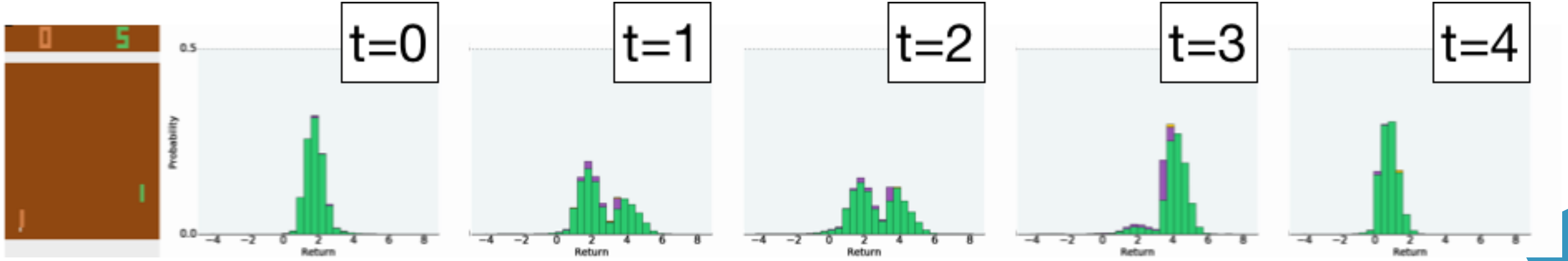
Distributional RL

A Distributional Perspective on Reinforcement Learning (C51)



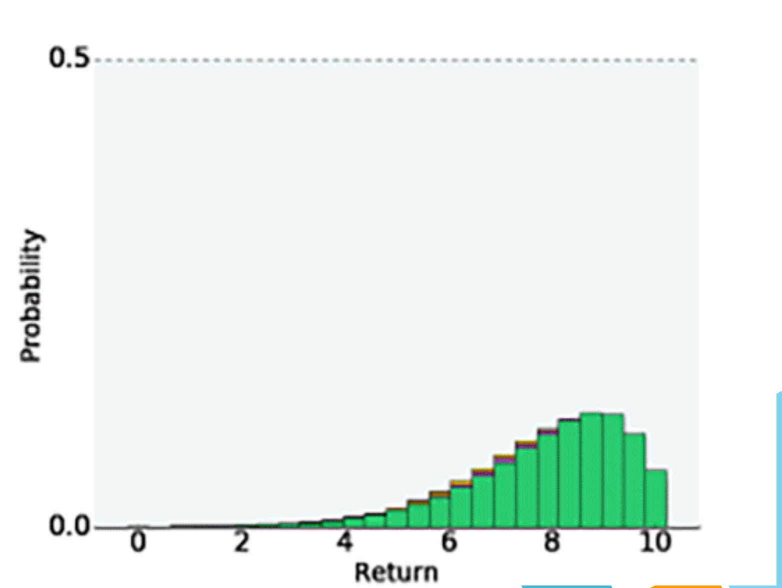
Distributional RL

A Distributional Perspective on Reinforcement Learning (C51)



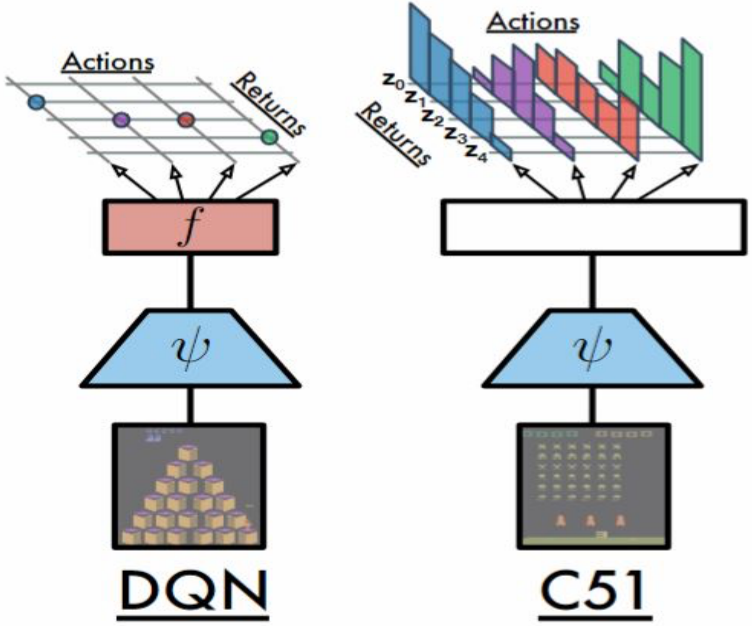
Distributional RL

A Distributional Perspective on Reinforcement Learning (C51)



Distributional RL

A Distributional Perspective on Reinforcement Learning (C51)



Distributional RL

A Distributional Perspective on Reinforcement Learning (C51)

C51 = DQN + Projection Distribution
(분포 만들기)

Distributional RL

A Distributional Perspective on Reinforcement Learning (C51)

Distributional DQN

1. Return에 대한 Value Distribution(51개 bin)을 만든다.
2. 각 스텝마다 만든 Value Distribution 들간의 거리를 구한다.
   → 논문에서 이론상 Wasserstein distance로 정의했지만
   실험에서 KL-divergence로 계산
3. Cross entropy로 분포간의 Loss 계산

Distributional RL

A Distributional Perspective on Reinforcement Learning (C51)

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

Distributional RL

A Distributional Perspective on Reinforcement Learning (C51)

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

Replay Buffer에서 Batch size만큼 추출

Distributional RL

A Distributional Perspective on Reinforcement Learning (C51)

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

Projection Distribution

(분포 만들기)

Distributional RL

A Distributional Perspective on Reinforcement Learning (C51)

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 $\{z_i\}$
$\boxed{\hat{\mathcal{T}} z_j \leftarrow [r_t + \gamma_t z_j]_{V_{\text{MIN}}}^{V_{\text{MAX}}}}$
$b_j \leftarrow (\hat{\mathcal{T}} z_j - V_{\text{MIN}})/\Delta z$ # $b_j \in [0, N-1]$
$l \leftarrow \lfloor b_j \rfloor, u \leftarrow \lceil b_j \rceil$
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

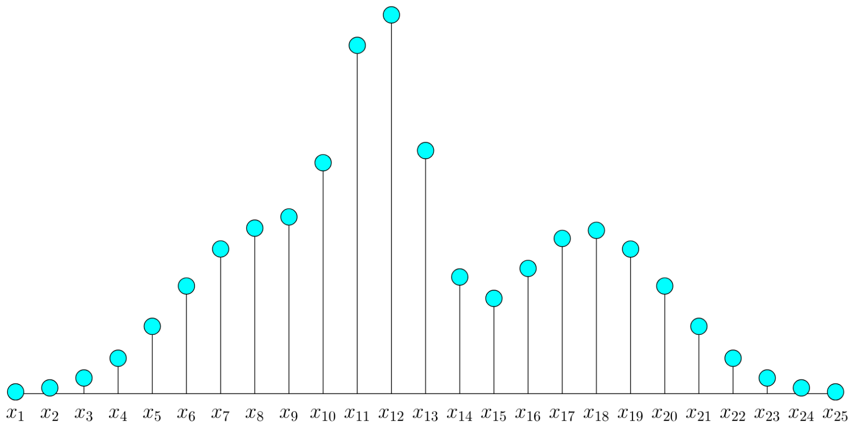
Bellman distributional operator

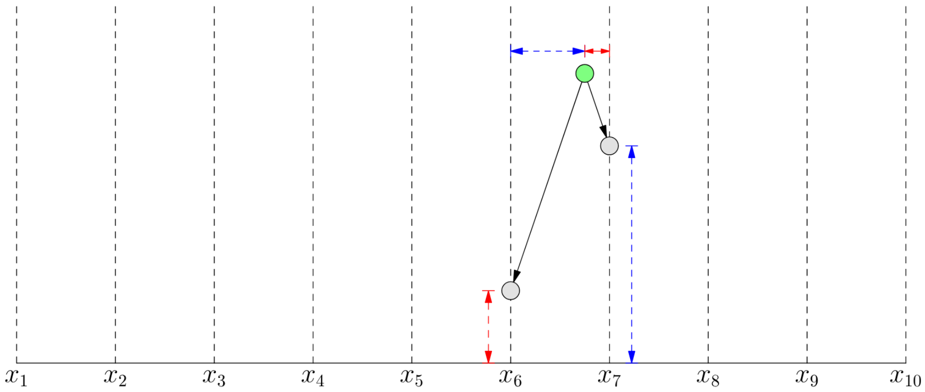
$V_{max} = 10$

$V_{mim} = -10$

Distributional RL

A Distributional Perspective on Reinforcement Learning (C51)





Distributional RL

A Distributional Perspective on Reinforcement Learning (C51)

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

KL-divergence(cross entropy)로

Loss 구하기

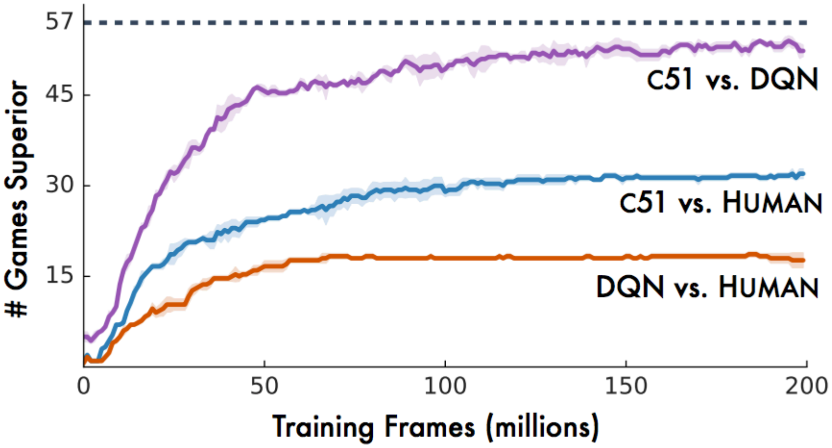
Performance

A Distributional Perspective on Reinforcement Learning (C51)

Comparison

	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%	-	-

Relative Performance



3. 코드 구현체 분석

감사합니다.

» urleee@naver.com