

# A distributional perspective on RL

Marc G. Bellemare, Will Dabney, Remi Munos

Aleksey Grinchuk

# Classic perspective

- Transition operator

$$P^\pi : \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|} \rightarrow \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$$

$$P^\pi Q(s, a) := \mathbb{E}_{s' | s, a} \mathbb{E}_{a' | s'} Q(s', a')$$

- Bellman operator

$$\mathcal{T}^\pi : \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|} \rightarrow \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$$

$$\mathcal{T}^\pi Q(s, a) := R(s, a) + \gamma P^\pi Q(s, a)$$

- Bellman equation

$$Q(s, a) = \mathcal{T}^\pi Q(s, a)$$

# Distributional perspective

- Probability distribution

$$Z(s, a) \sim q(z|s, a)$$

- Value distribution

$$\mathbb{E}_{z|s,a} Z(s, a) = \sum_z z q(z|s, a) = Q(s, a)$$

- Each element  $Z$  of the **set of value distributions  $\mathcal{Z}$**  is defined by  $|\mathcal{S}| \times |\mathcal{A}|$  probability distributions

$$Z = \{Z(s_1, a_1), \dots, Z(s_n, a_m)\} \in \mathcal{Z}$$

# Distributional perspective

- Distributional transition operator

$$P_D^\pi : \mathcal{Z} \rightarrow \mathcal{Z}$$

$$P_D^\pi Z(s, a) := \mathbb{E}_{s' | s, a} \mathbb{E}_{a' | s'} Z(s', a')$$

- Distributional Bellman operator

$$\mathcal{T}_D^\pi : \mathcal{Z} \rightarrow \mathcal{Z}$$

$$\mathcal{T}_D^\pi Z(s, a) := R(s, a) + \gamma P_D^\pi Z(s, a)$$

- Distributional Bellman equation

$$Z(s, a) \stackrel{D}{=} \mathcal{T}_D^\pi Z(s, a)$$

# Distributional $\rightarrow$ Classic

- Distributional Bellman equation

$$Z(s, a) \stackrel{D}{=} \mathcal{T}_D^\pi Z(s, a)$$

$$\mathbb{E}_{z|s,a} Z(s, a) = \mathbb{E}_{z|s,a} \mathcal{T}_D^\pi Z(s, a)$$

$$Q(s, a) = \mathbb{E}_{z|s,a} R(s, a) + \gamma \mathbb{E}_{z|s,a} \mathbb{E}_{s'|s,a} \mathbb{E}_{a'|s'} Z(s', a')$$

$$Q(s, a) = R(s, a) + \gamma \mathbb{E}_{s'|s,a} \mathbb{E}_{a'|s'} \mathbb{E}_{z|s',a'} Z(s', a')$$

- Classic Bellman equation

$$Q(s, a) = R(s, a) + \gamma P^\pi Q(s, a)$$

# Probability distance

- Distributional Bellman equation

$$Z(s, a) \stackrel{D}{=} \mathcal{T}_D^\pi Z(s, a)$$

- How to define probability distance ?

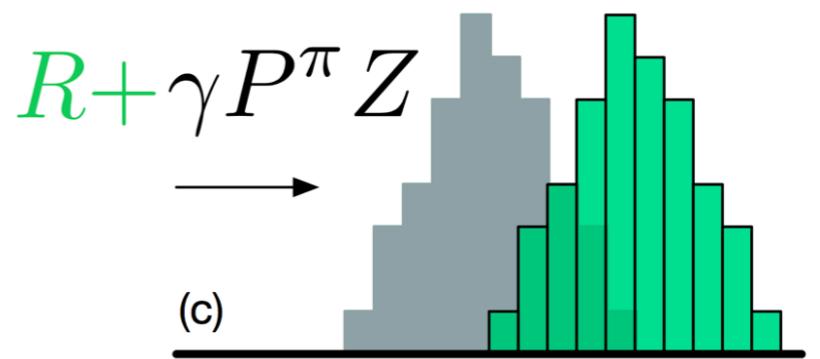
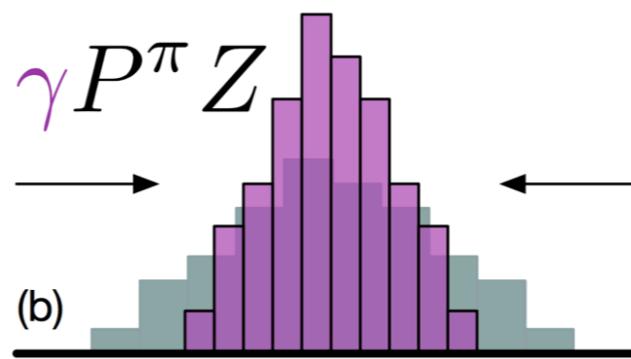
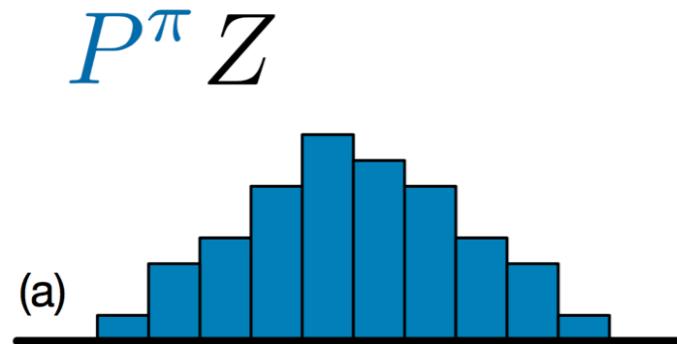
- Kullback-Leibler divergence
- Total variation
- Wasserstein metric

# KL divergence

- Kullback-Leibler divergence

$$D_{\text{KL}}(p\|q) = \mathbb{E}_p \left[ \log \frac{p(z)}{q(z)} \right]$$

- **Problem:** goes to infinity if distributions have disjoint supports  $D_{\text{KL}}(p\|q) = \infty$

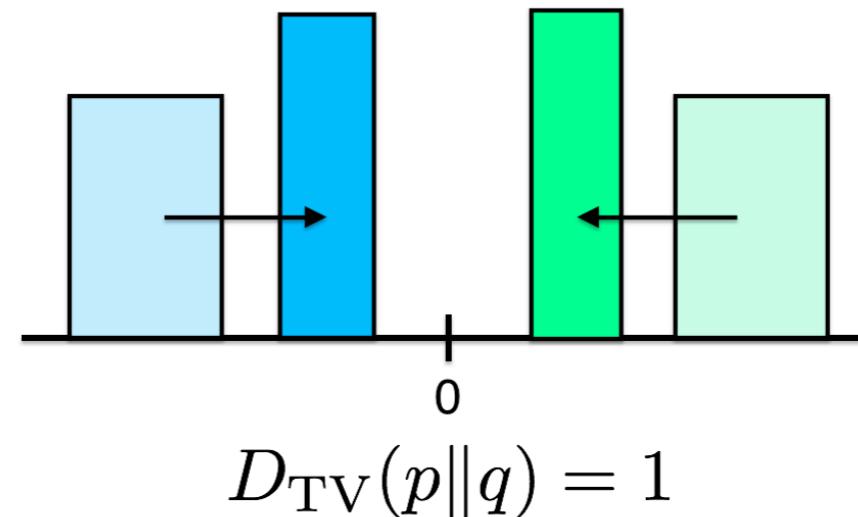
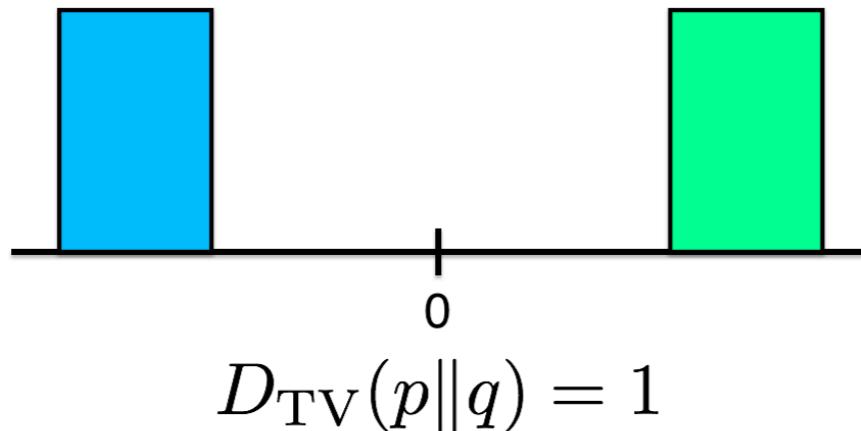


# Total variation

- Total variation

$$D_{\text{TV}}(p\|q) = \sup_{A \in \mathcal{F}} |p(A) - q(A)|$$

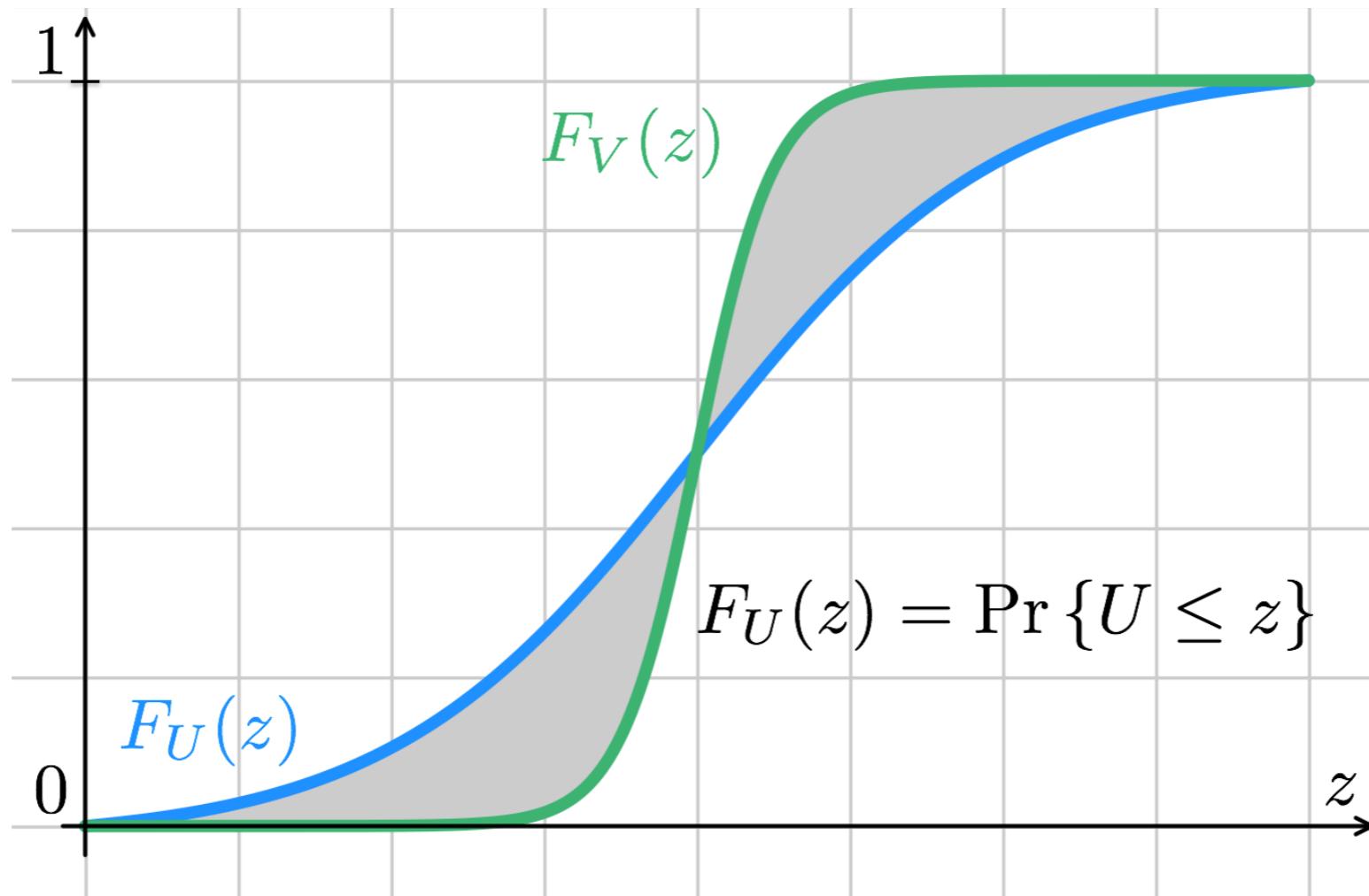
- **Problem:** distance may remain the same after shift



# Wasserstein metric

- Wasserstein metric

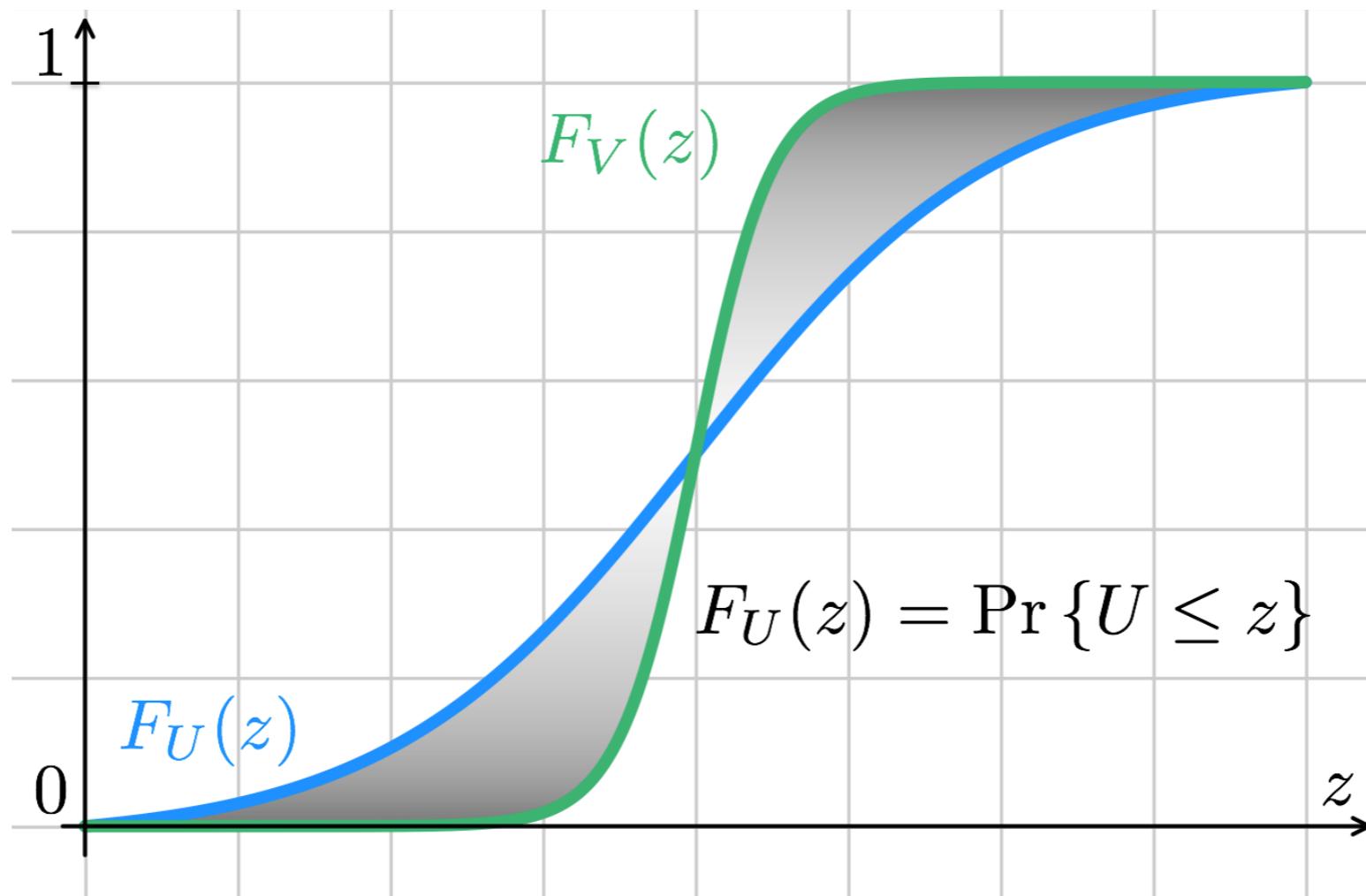
$$w_1(U, V) = \int_0^1 |F_U^{-1}(y) - F_V^{-1}(y)| dy$$



# Wasserstein metric

- Wasserstein metric

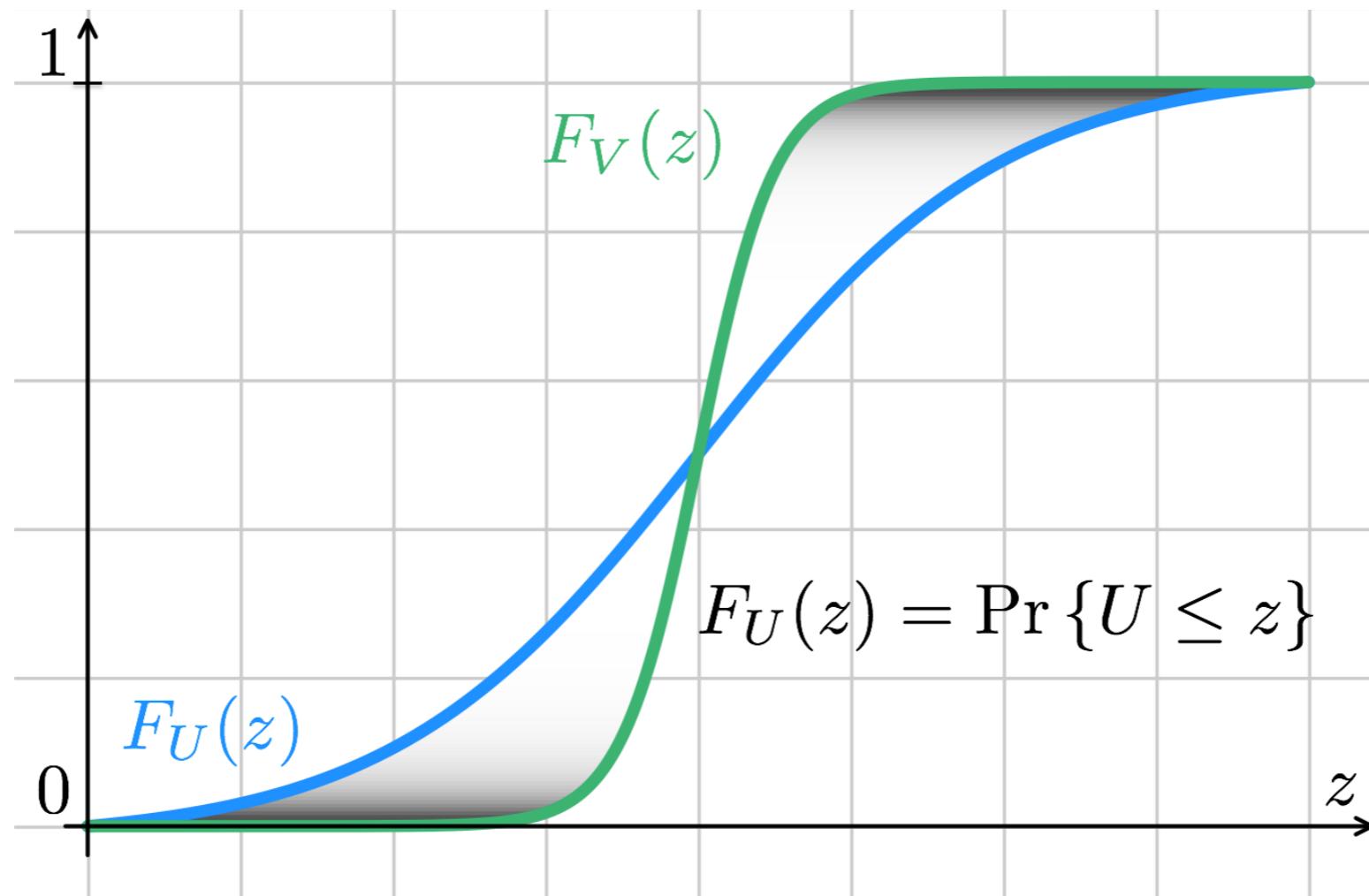
$$w_2(U, V) = \left( \int_0^1 |F_U^{-1}(y) - F_V^{-1}(y)|^2 dy \right)^{\frac{1}{2}}$$



# Wasserstein metric

- Wasserstein metric

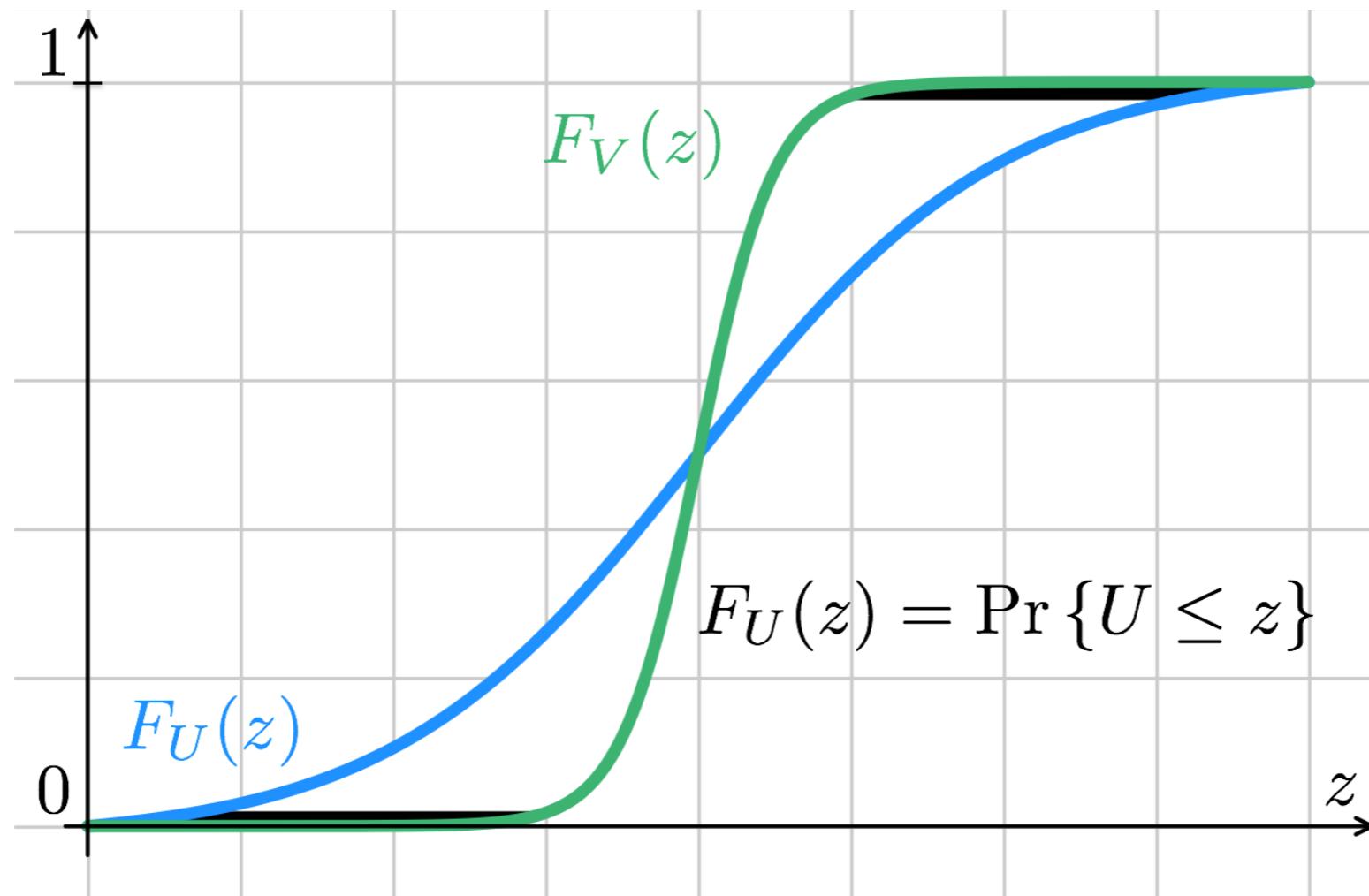
$$w_p(U, V) = \left( \int_0^1 |F_U^{-1}(y) - F_V^{-1}(y)|^p dy \right)^{\frac{1}{p}}$$



# Wasserstein metric

- Wasserstein metric

$$w_\infty(U, V) = \sup_{0 \leq y \leq 1} |F_U^{-1}(y) - F_V^{-1}(y)|$$



# Metric properties

- Wasserstein metric properties

$$w_p(A + U, A + V) \leq w_p(U, V)$$

$$w_p(aU, aV) \leq |a|w_p(U, V)$$

$$w_p(AU, AV) \leq \|A\|_p w_p(U, V)$$

- Maximal form of the Wasserstein metric

$$\bar{w}_p(Z_1, Z_2) := \sup_{s,a} w_p(Z_1(s, a), Z_2(s, a))$$

# Metric properties

- Wasserstein metric properties

$$w_p(A + U, A + V) \leq w_p(U, V)$$

$$w_p(aU, aV) \leq |a|w_p(U, V)$$

$$w_p(AU, AV) \leq \|A\|_p w_p(U, V)$$

- Maximal form of the Wasserstein metric

$$\bar{w}_p(\mathcal{T}^\pi Z_1, \mathcal{T}^\pi Z_2) = \bar{w}_p(R + \gamma P^\pi Z_1, R + \gamma P^\pi Z_2)$$

# Metric properties

- Wasserstein metric properties

$$w_p(A + U, A + V) \leq w_p(U, V)$$

$$w_p(aU, aV) \leq |a|w_p(U, V)$$

$$w_p(AU, AV) \leq \|A\|_p w_p(U, V)$$

- Maximal form of the Wasserstein metric

$$\begin{aligned} \bar{w}_p(\mathcal{T}^\pi Z_1, \mathcal{T}^\pi Z_2) &= \bar{w}_p(R + \gamma P^\pi Z_1, R + \gamma P^\pi Z_2) \\ &\leq \bar{w}_p(\gamma P^\pi Z_1, \gamma P^\pi Z_2) \end{aligned}$$

# Metric properties

- Wasserstein metric properties

$$w_p(A + U, A + V) \leq w_p(U, V)$$

$$w_p(aU, aV) \leq |a|w_p(U, V)$$

$$w_p(AU, AV) \leq \|A\|_p w_p(U, V)$$

- Maximal form of the Wasserstein metric

$$\begin{aligned} \bar{w}_p(\mathcal{T}^\pi Z_1, \mathcal{T}^\pi Z_2) &= \bar{w}_p(R + \gamma P^\pi Z_1, R + \gamma P^\pi Z_2) \\ &\leq \bar{w}_p(\gamma P^\pi Z_1, \gamma P^\pi Z_2) \\ &\leq \gamma \bar{w}_p(P^\pi Z_1, P^\pi Z_2) \end{aligned}$$

# Metric properties

- Wasserstein metric properties

$$w_p(A + U, A + V) \leq w_p(U, V)$$

$$w_p(aU, aV) \leq |a|w_p(U, V)$$

$$w_p(AU, AV) \leq \|A\|_p w_p(U, V)$$

- Maximal form of the Wasserstein metric

$$\begin{aligned} \bar{w}_p(\mathcal{T}^\pi Z_1, \mathcal{T}^\pi Z_2) &= \bar{w}_p(R + \gamma P^\pi Z_1, R + \gamma P^\pi Z_2) \\ &\leq \bar{w}_p(\gamma P^\pi Z_1, \gamma P^\pi Z_2) \\ &\leq \gamma \bar{w}_p(P^\pi Z_1, P^\pi Z_2) \\ &\leq \gamma \bar{w}_p(Z_1, Z_2) \end{aligned}$$

# Policy evaluation

- Classic

The Bellman operator  $\mathcal{T}^\pi : \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|} \rightarrow \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$

$$\mathcal{T}^\pi Q(s, a) := R(s, a) + \gamma P^\pi Q(s, a)$$

is a  $\gamma$ -contraction in  $l_\infty$ .

- Distributional

The Bellman operator  $\mathcal{T}_D^\pi : \mathcal{Z} \rightarrow \mathcal{Z}$

$$\mathcal{T}_D^\pi Z(s, a) := R(s, a) + \gamma P_D^\pi Z(s, a)$$

is a  $\gamma$ -contraction in  $\bar{w}_p$ ,  $1 \leq p \leq \infty$ .

# Control

- Definition. An **optimal value distribution** is the value distribution of an optimal policy. The set of optimal value distributions is  $\mathcal{Z}^* := \left\{ Z^{\pi^*} : \pi^* \in \Pi^* \right\}$

- Definition. The **set of greedy policies** for  $Z \in \mathcal{Z}$  is

$$\mathcal{G}_Z := \left\{ \pi : \sum_{a \in \mathcal{A}} \pi(a|s) \mathbb{E}Z(s, a) = \max_{a' \in \mathcal{A}} \mathbb{E}Z(s, a') \right\}.$$

- Definition. A **distributional Bellman opt. operator**

$$\mathcal{T}_D Z(s, a) := \mathcal{T}_D^\pi Z(s, a) \text{ for some } \pi \in \mathcal{G}_\pi.$$

# Control

- Classic

The Bellman opt. operator  $\mathcal{T} : \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|} \rightarrow \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$

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

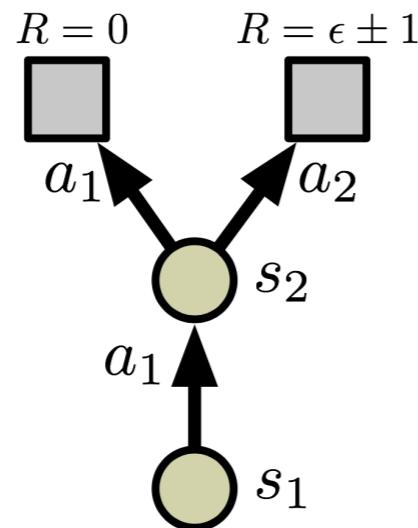
is a  $\gamma$ -contraction in  $l_\infty$ .

- Distributional

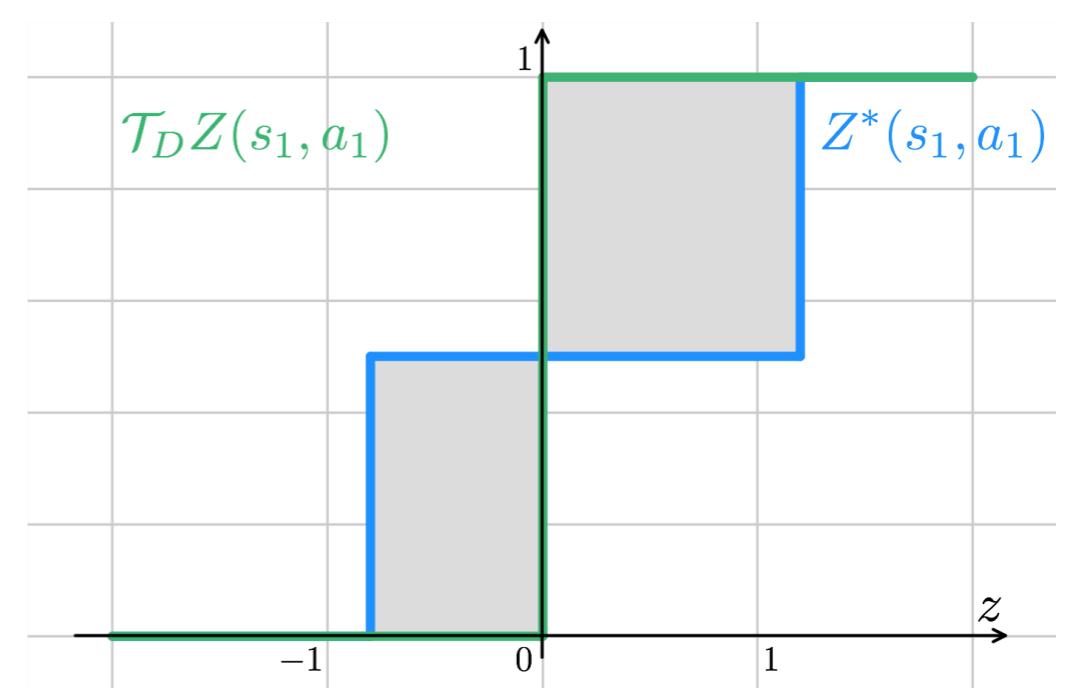
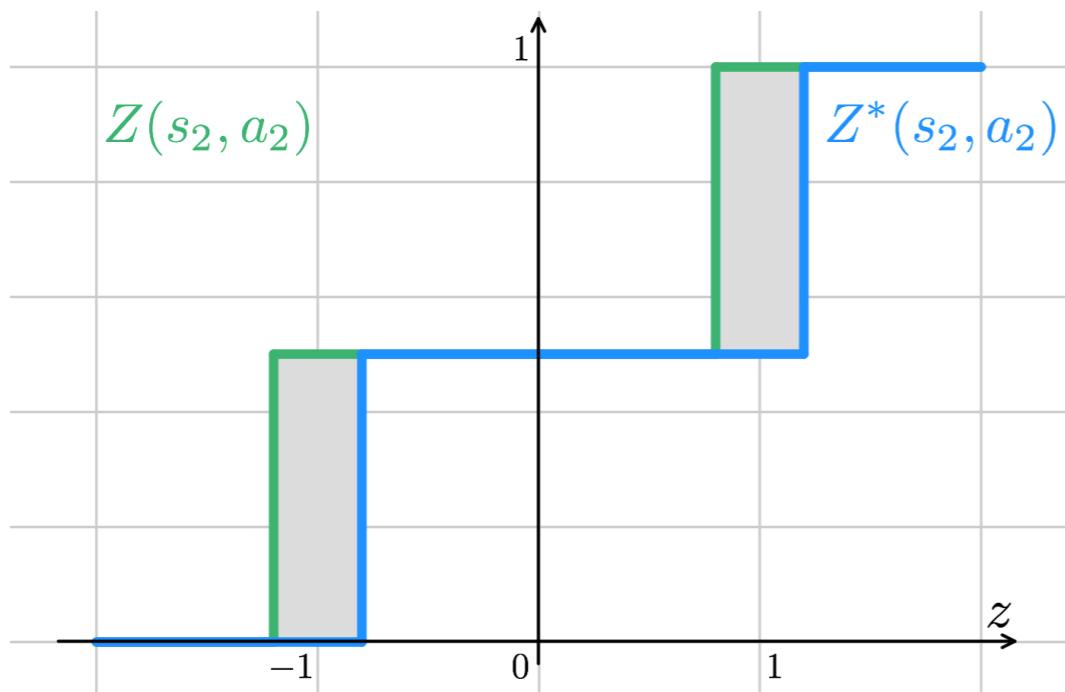
The Bellman opt. operator  $\mathcal{T}_D : \mathcal{Z} \rightarrow \mathcal{Z}$

is **not** a  $\gamma$ -contraction in  $\bar{w}_p$ ,  $1 \leq p \leq \infty$ .

# Non-contraction example



	$s_1, a_1$	$s_2, a_1$	$s_2, a_2$
$Z^*$	$\epsilon \pm 1$	0	$\epsilon \pm 1$
$Z$	$\epsilon \pm 1$	0	$-\epsilon \pm 1$
$\mathcal{T}_D Z$	0	0	$\epsilon \pm 1$

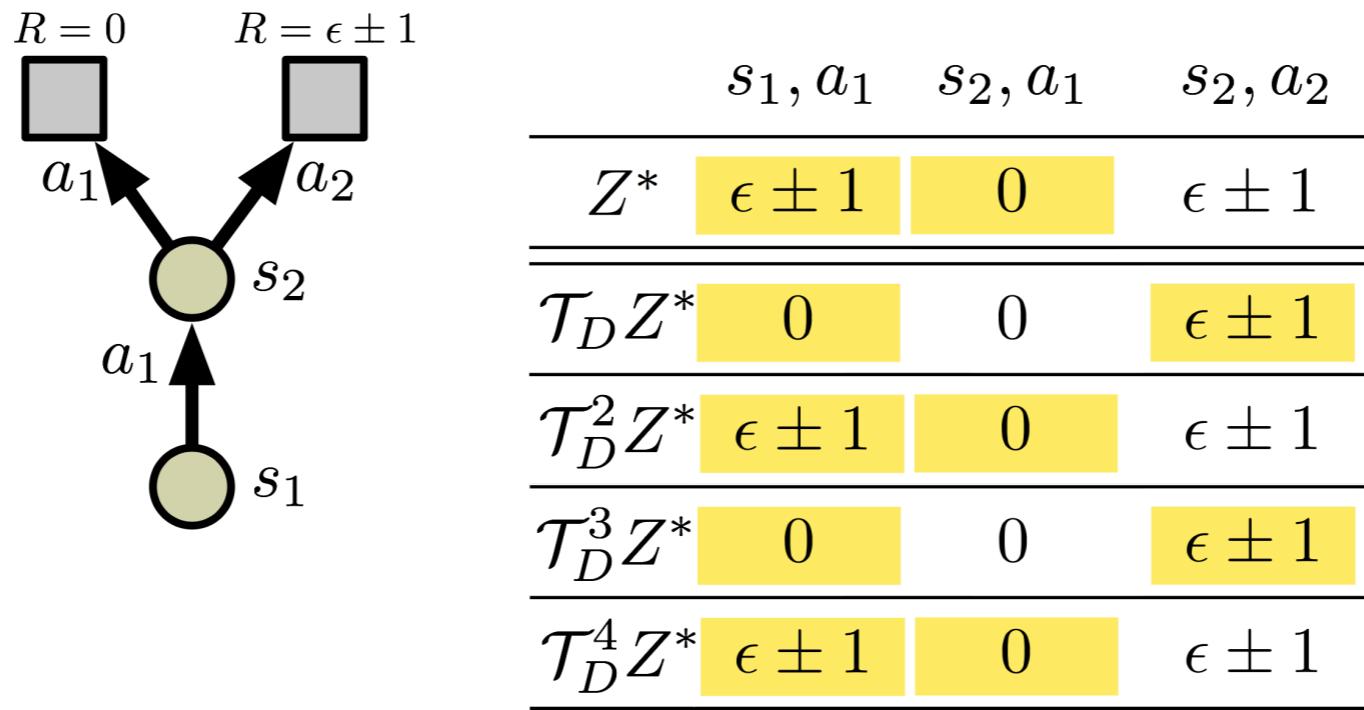


$$\begin{aligned} \bar{w}_1(Z, Z^*) &= w_1(Z(s_2, a_2), Z^*(s_2, a_2)) \\ &= \frac{1}{2}2\epsilon + \frac{1}{2}2\epsilon = 2\epsilon \end{aligned}$$

$$\begin{aligned} \bar{w}_1(\mathcal{T}_D Z, \mathcal{T}_D Z^*) &= w_1(\mathcal{T}_D Z(s_1, a_1), Z^*(s_1, a_1)) \\ &= \frac{1}{2}(1 - \epsilon) + \frac{1}{2}(1 + \epsilon) = 1 \end{aligned}$$

# No fixed point

- Proposition. Not all optimality operators have a fixed point  $Z^* = \mathcal{T}_D Z^*$ .



$$\mathcal{T}_D Z(s_1, a_1) = \begin{cases} Z(s_2, a_2), & Z(s_1, a_1) = 0, \\ Z(s_2, a_1), & Z(s_1, a_1) \neq 0. \end{cases}$$

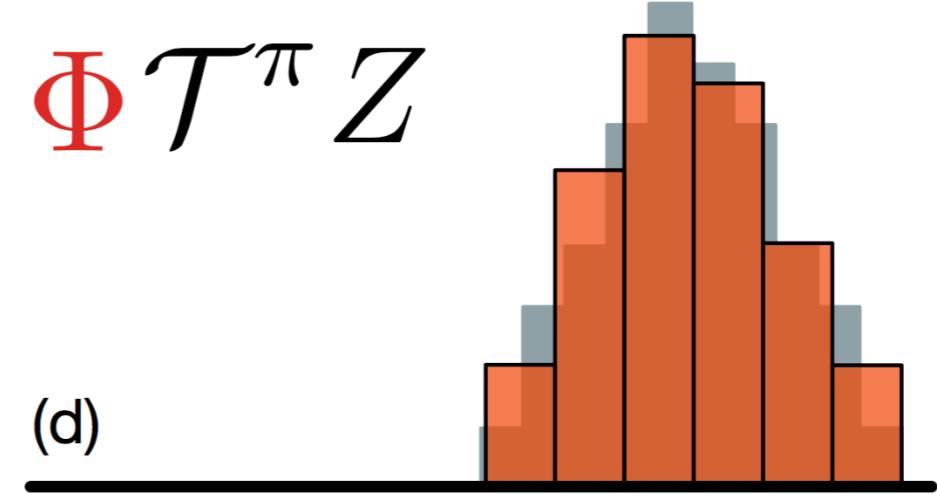
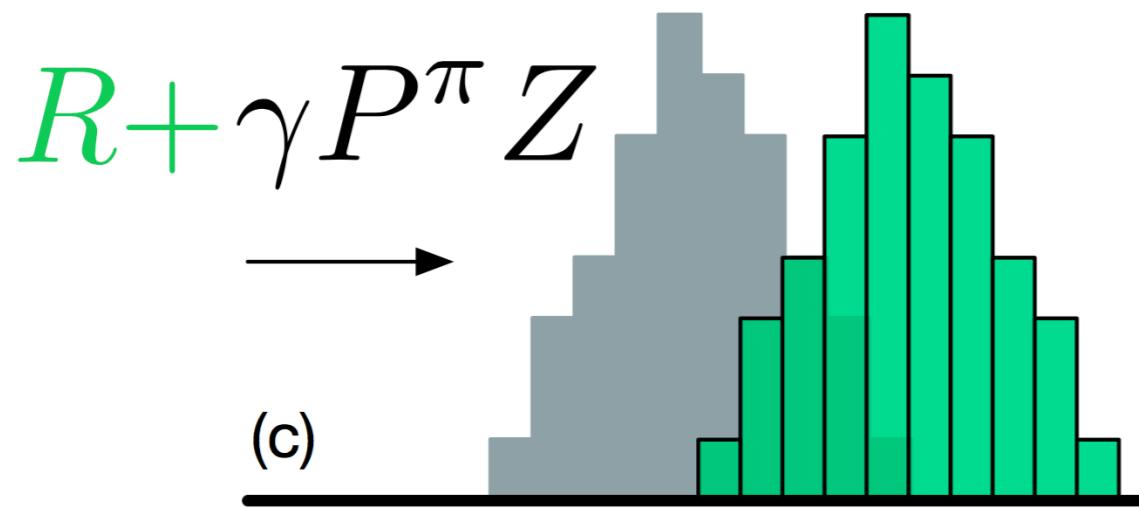
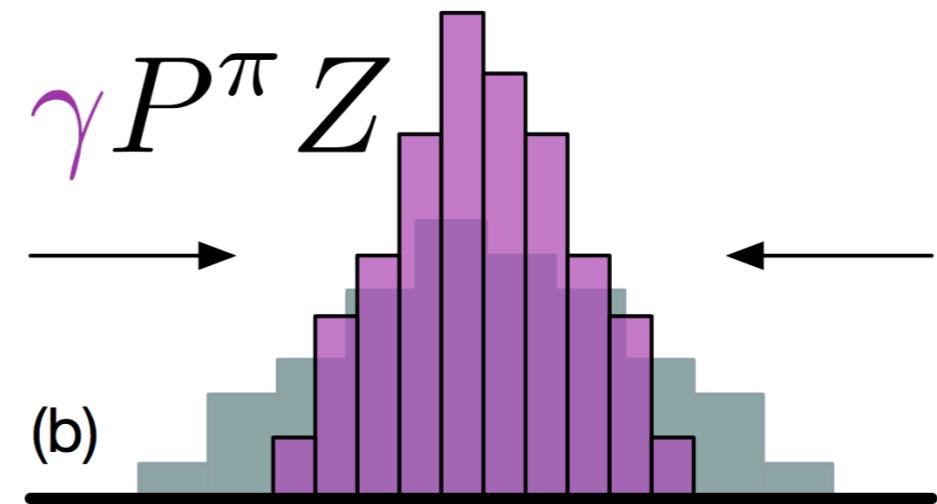
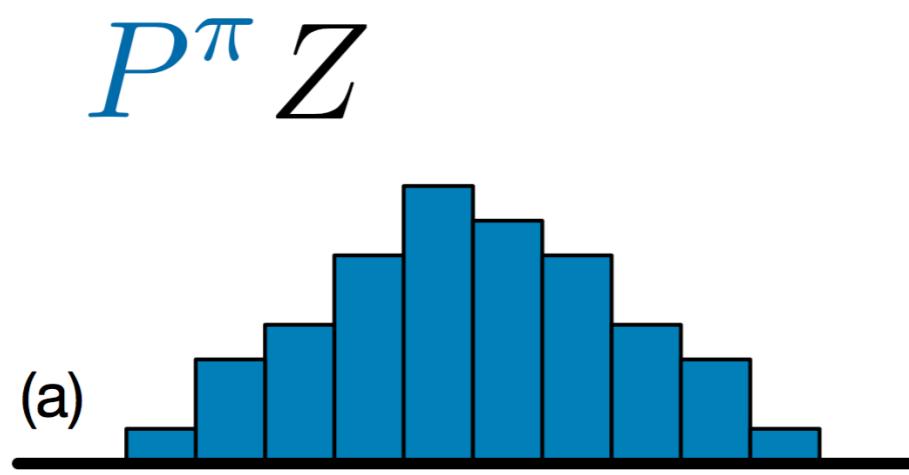
# Convergence in control

- Lemma. Let  $Z_1, Z_2 \in \mathcal{Z}$ . Then

$$\|\mathbb{E}\mathcal{T}_D Z_1 - \mathbb{E}\mathcal{T}_D Z_2\|_\infty \leq \gamma \|\mathbb{E}Z_1 - \mathbb{E}Z_2\|_\infty$$

and in particular  $\mathbb{E}Z_k \rightarrow Q^*$  exponentially quickly  
for the process  $Z_{k+1} := \mathcal{T}_D Z_k$ ,  $Z_0 \in \mathcal{Z}$ .

# From theory to practice



$$\mathcal{T}_D^\pi Z \stackrel{w_p}{=} Z$$

Theory

$$\Phi \mathcal{T}_D^\pi Z \stackrel{D_{\text{KL}}}{\approx} Z$$

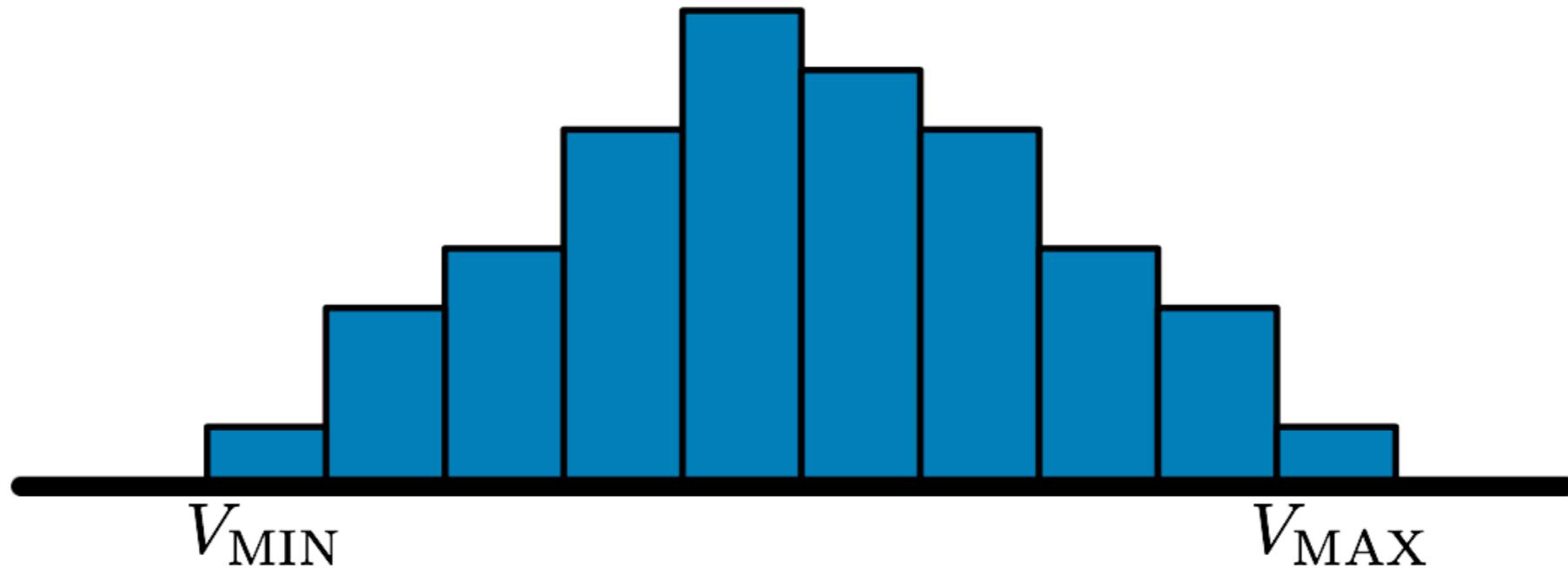
Practice

# Categorical DQN

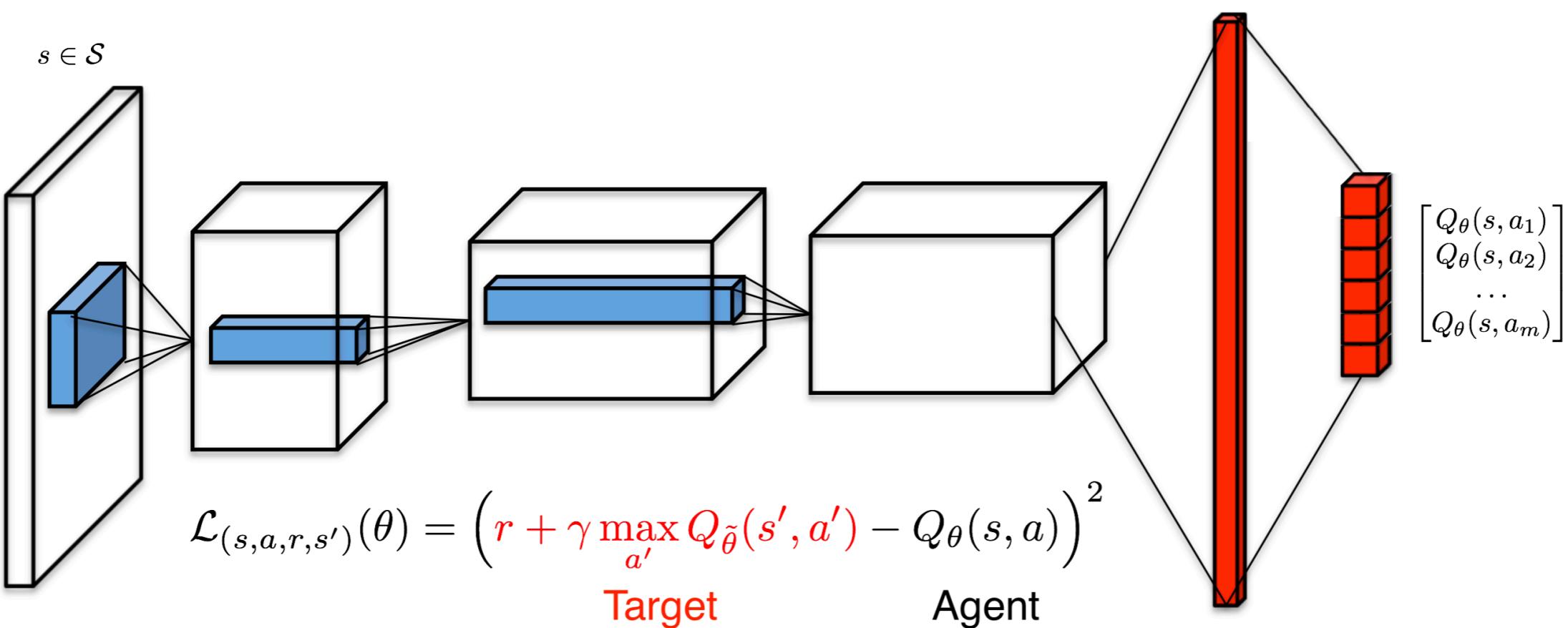
- Discrete probability distribution over  $Z_\theta(s, a)$

$$Z_\theta(s, a) = z_i \text{ w.p. } p_i(s, a; \theta) := \frac{e^{\theta_i(s, a)}}{\sum_{j=1}^N e^{\theta_j(s, a)}}$$

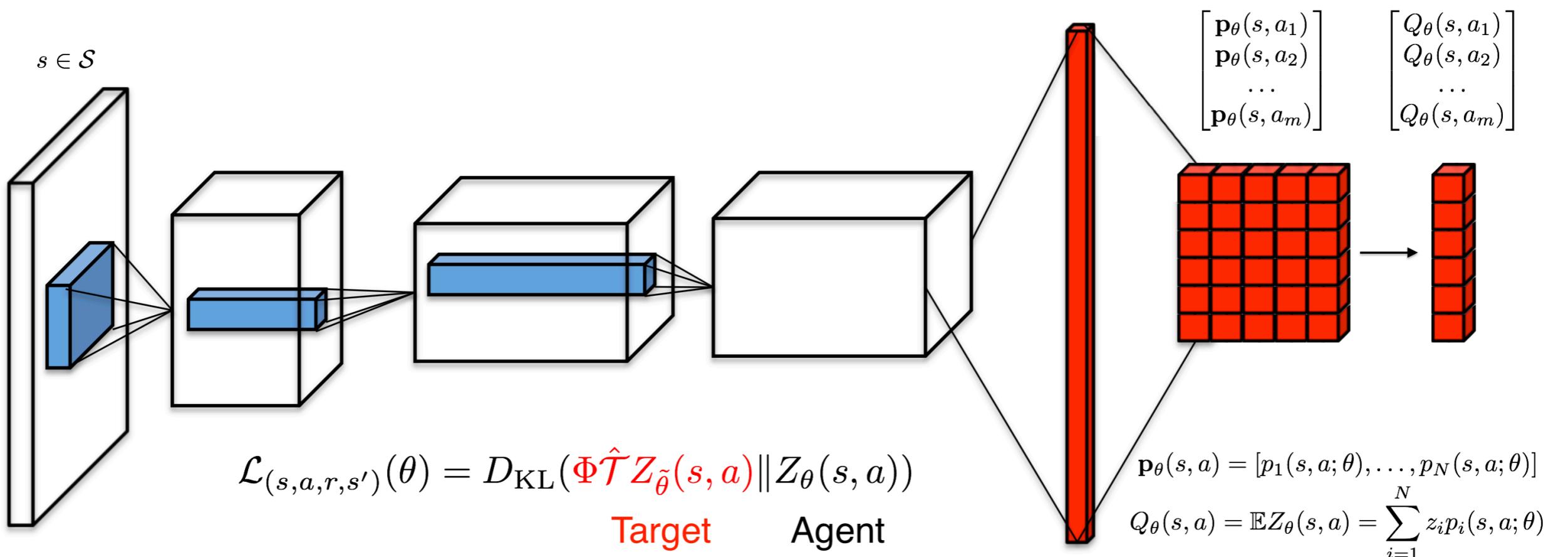
$$\{z_i = V_{\text{MIN}} + i\Delta z : 0 \leq i < N\}, \quad \Delta z := \frac{V_{\text{MAX}} - V_{\text{MIN}}}{N - 1}$$



# Classic DQN



# Categorical DQN



# Categorical algorithm

---

**Algorithm 1** Categorical Algorithm

---

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

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

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

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

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

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

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

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

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

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

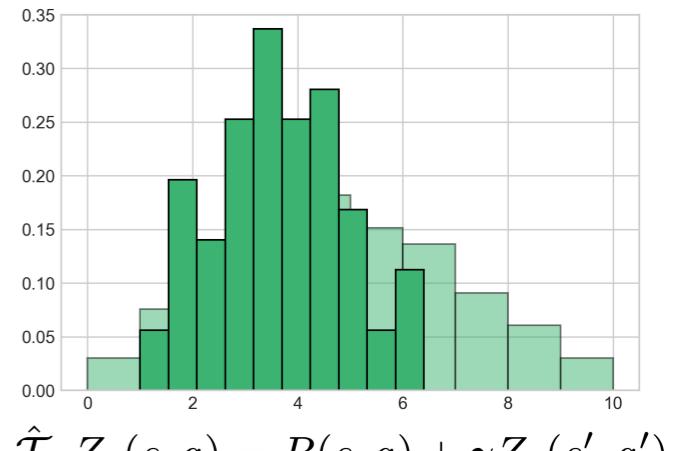
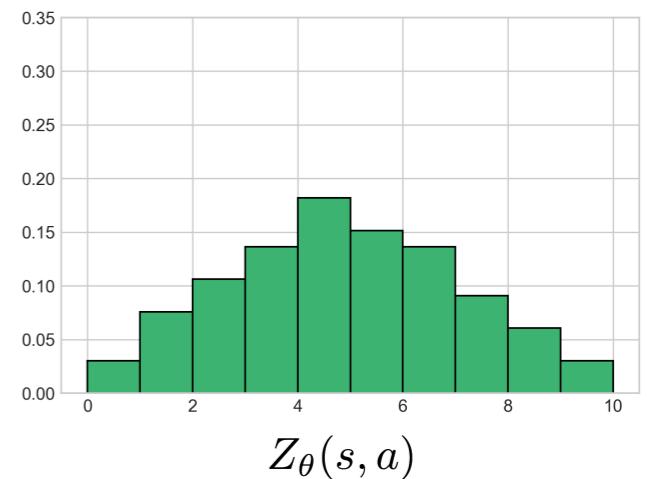
$$m_l \leftarrow m_l + p_j(x_{t+1}, a^*)(u - b_j)$$

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

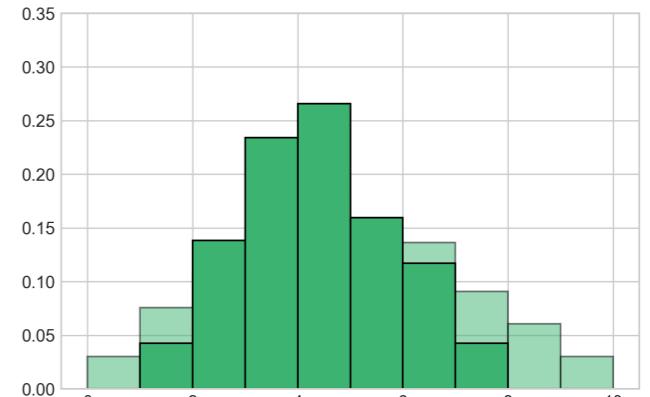
**end for**

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

---

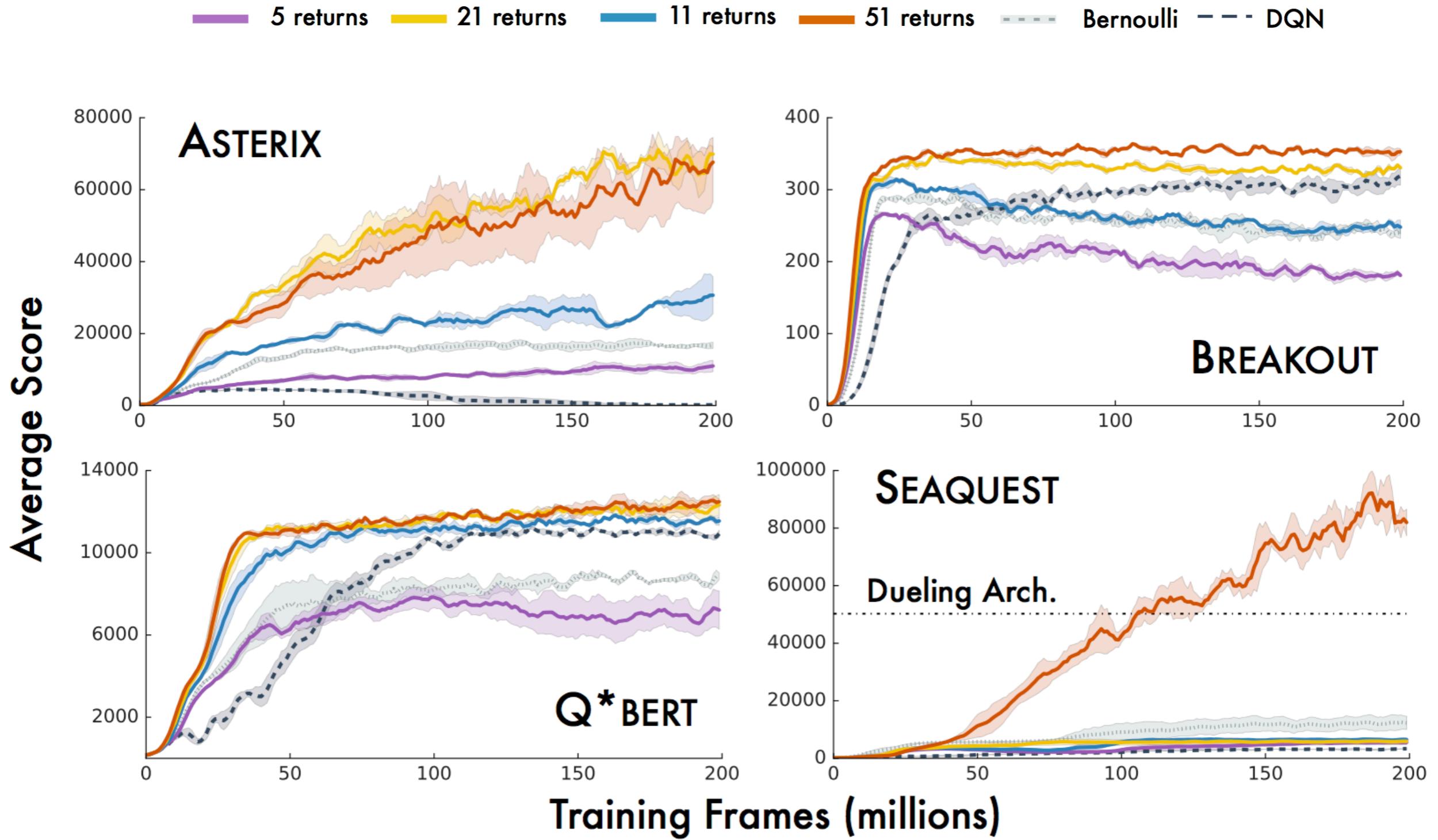


$$\hat{\mathcal{T}}_D Z_\theta(s, a) = R(s, a) + \gamma Z_\theta(s', a')$$



$$\Phi \hat{\mathcal{T}}_D Z_\theta(s, a)$$

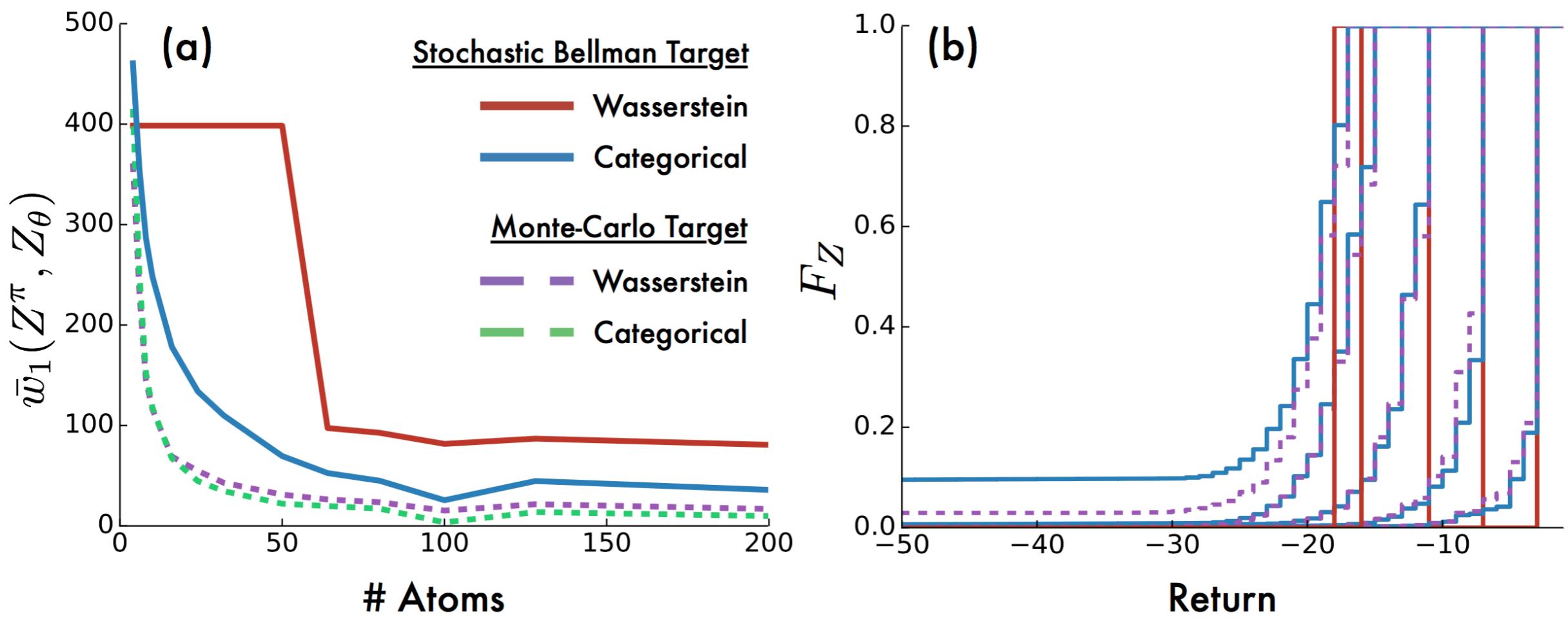
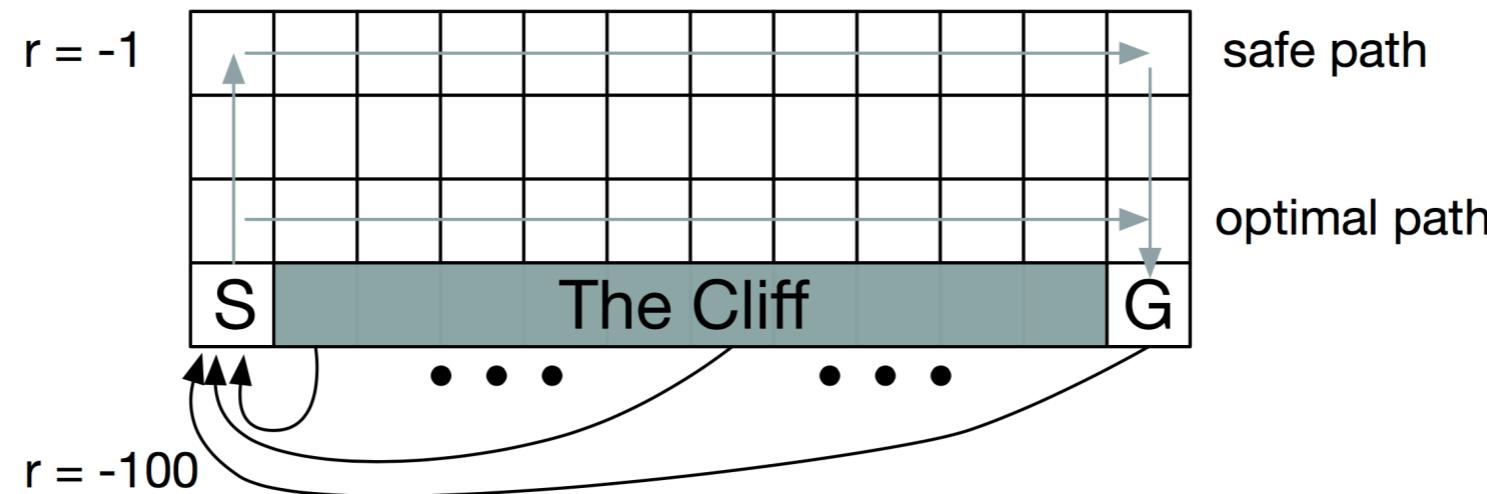
# Results



# C51 Hyperparameters

- Adam optimizer: learning rate  $\alpha = 2.5 \times 10^{-4}$   
batch size  $L = 32$   
epsilon  $\varepsilon_{\text{adam}} = 0.01/L$
- Distribution support  $V_{\text{MIN}} = -10$ ,  $V_{\text{MAX}} = 10$
- Number of atoms  $N = 51$
- Implementation  
[https://github.com/AlexGrinch/rl\\_algorithms](https://github.com/AlexGrinch/rl_algorithms)

# Sample Wasserstein loss



# Discussion

- Reduced chattering
- State aliasing
- A richer set of predictions
- Framework for inductive bias
- Well-behaved optimization