# Multivariate Distributional Reinforcement Learning Using Sliced Divergences

Baptiste Debes, Tinne Tuytelaars

KU Leuven

Funded by the European Union

European Research Council

ICML International Conference On Machine Learning

## Multivariate distributional RL

- Reward function can be multivariate $R \in \mathbb{R}^d$
- Multivariate Temporal Difference
  - Sobolev TD
  - Successor features

$$S' \sim P(\cdot \mid s,a), \qquad A' \sim \pi(\cdot \mid S'), \qquad Z(S',A') \sim \mu(S',A')$$
$$(\mathcal{T}^\pi \mu)(s,a) = \mathrm{Law}\left[R(s,a) + \Gamma(s,a) Z(S',A')\right]$$

$$\Gamma(s,a) = \gamma I_d = \begin{bmatrix} \gamma & 0 & \cdots & 0 \\ 0 & \gamma & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \gamma \end{bmatrix}$$

**Regular multivariate RL**

$$\Gamma(s,a) = \begin{bmatrix} \gamma_{11}(s,a) & \gamma_{12}(s,a) & \cdots & \gamma_{1d}(s,a) \\ \gamma_{21}(s,a) & \gamma_{22}(s,a) & \cdots & \gamma_{2d}(s,a) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{d1}(s,a) & \gamma_{d2}(s,a) & \cdots & \gamma_{dd}(s,a) \end{bmatrix}$$

**General case**

## Contraction property

- Divergence between return laws

$$D : \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}_+$$

- Contraction criterion

$$\overline{D}(\eta_1, \eta_2) := \sup_{(s,a)} D(\eta_1(s,a), \eta_2(s,a))$$
$$\overline{D}(\mathcal{T}^\pi \eta_1, \mathcal{T}^\pi \eta_2) \le \kappa\, \overline{D}(\eta_1, \eta_2), \qquad \kappa < 1$$

- Baseline is Wasserstein

$$\overline{W}_p(\mathcal{T}^\pi \eta_1, \mathcal{T}^\pi \eta_2) \le \overline{L}\, \overline{W}_p(\eta_1, \eta_2), \qquad \overline{L} := \sup_{(s,a)} \|\Gamma(s,a)\|_{\mathrm{op}} < 1$$

Exact empirical estimator is **intractable** $\mathcal{O}(n^3 \log n)$

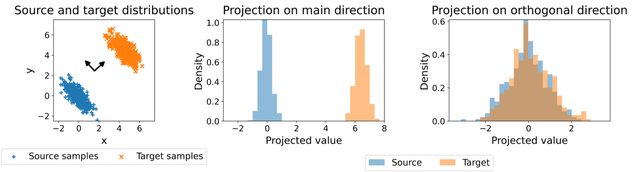
## Slicing

- Setup
  - Base 1D divergence $\Delta : \mathcal{P}(\mathbb{R}) \times \mathcal{P}(\mathbb{R}) \to \mathbb{R}_+$
  - Two distributions $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$
  - Random direction $\theta \in \mathbb{S}^{d-1}$
- Core idea : compare projections

$$\Delta((P_\theta)_\# \mu, (P_\theta)_\# \nu) \qquad P_\theta(x) = \langle \theta, x \rangle$$

- Aggregate

$$\theta_1, \ldots, \theta_K \overset{\text{i.i.d.}}{\sim} \sigma, \qquad \sigma = \mathrm{Unif}(\mathbb{S}^{d-1})$$
$$\widehat{\mathbf{S}}\Delta_p^p(\mu,\nu) = \frac{1}{K} \sum_{k=1}^{K} \Delta^p((P_{\theta_k})_\# \mu, (P_{\theta_k})_\# \nu)$$




## Base 1D divergences

- Cramér distance
  - Distance between CDFs

$$\mathbf{C}_2(\mu,\nu) = \int_{\mathbb{R}} |F_\mu(u) - F_\nu(u)|^2 \, du$$

- Maximum Mean Discrepancy
  - Kernel-based $k : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$

$$\mathbf{MMD}_k^2(\mu,\nu) = \mathbb{E}_{x,x' \sim \mu}[k(x,x')] + \mathbb{E}_{y,y' \sim \nu}[k(y,y')] - 2\mathbb{E}_{x \sim \mu, y \sim \nu}[k(x,y)]$$

| Divergence | Time complexity |
|---|---|
| Wasserstein $W_p$ | $\mathcal{O}(n \log n)$ |
| Cramér $C_2$ | $\mathcal{O}(n \log n)$ |
| $\mathrm{MMD}_k$ | $\mathcal{O}(n^2)$ |

## Contraction

- Result #1: standard case $\Gamma = \gamma I$
  - $\mathrm{S}\Delta_p$ inherits contraction from $\Delta$
- Result #2: dense case $\Gamma(s,a)$, $\overline{L} < 1$
  - $\mathrm{S}\Delta_p$ fails norm-based contraction
- Result #3: max-slicing
  - Search for most discriminative direction

$$\mathrm{MS}\Delta(\mu,\nu) = \sup_{\theta \in S^{d-1}} \Delta((P_\theta)_\# \mu, (P_\theta)_\# \nu)$$

  - $\mathrm{MS}\Delta$ recovers norm-based contraction

## Stochastic training

- True Bellman target

$$\mu = \mathbb{E}_{S',A'}\left[\mathrm{Law}\left(R + \Gamma Z^-(S',A')\right)\right]$$

- Sampled Bellman target

$$\hat{\mu} = \mathrm{Law}\left(R + \Gamma Z^-(s',a')\right)$$

- Unbiased gradient

$$\textbf{(U)} \quad \mathbb{E}_{S',A'}\left[\nabla_\phi D(\hat{\mu}, \nu_\phi)\right] = \nabla_\phi D(\mu, \nu_\phi)$$

| Method | Contraction $\overline{L} < 1$ | (U) |
|---|---|---|
| $\mathbf{S}\Delta_p$ | × | if base $\Delta$ has (U) |
| $\mathbf{MS}\Delta$ | ✓ | × |
| $\mathbf{W}_p$ | ✓ | × |

**Open question: can we get both ?**

## Experiments