Distributional Value Gradients for Stochastic Environments

Baptiste Debes, Tinne Tuytelaars

KU Leuven

ICLR

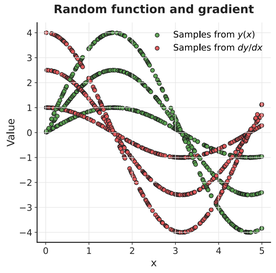
KU LEUVEN

Funded by the European Union

European Research Council
Established by the European Commission

TL;DR

- Observations
  - Stochastic returns
    - ↪ Value gradients are stochastic
    - ↪ Deterministic value gradients may struggle
- Goals
  - Robustness & sample efficiency
- Contributions
  - Distributional Value Gradients
    - ↪ Joint law over return and gradient
  - Sobolev TD framework
    - ↪ Contraction theory
    - ↪ Trade-off intepretation



1

Value Gradients

- Policy optimization
  - Actor update from $\nabla_a Q^\pi(s,a)$

$$\nabla_\theta J(\theta) \propto \nabla_\theta \pi_\theta(s)\, \nabla_a Q^\pi(s,a)\big|_{a=\pi_\theta(s)}$$

- Learned dynamics
  - Gradient through one-step model

$$\hat{s}' = f(s,a), \qquad \hat{r} = r(s,a)$$
$$\hat{\delta}(s,a) = \hat{r} + \gamma\, Q_{\phi'}\big(\hat{s}', \pi(\hat{s}')\big)$$

- TD on value and gradient

$$\mathcal{L}(\phi) = \|Q_\phi(s,a) - \hat{\delta}(s,a)\|^2 + \lambda \|\nabla_a Q_\phi(s,a) - \nabla_a \hat{\delta}(s,a)\|^2$$

2

Distributional RL

- Distribution over returns

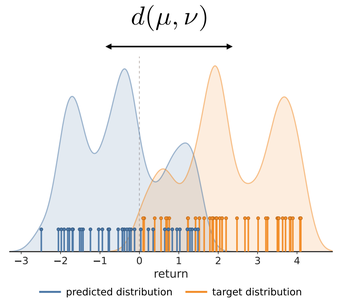
$$\eta^\pi(s,a) = \mathrm{Law}[Z^\pi(s,a)]$$
$$\eta_{\mathrm{tgt}}(s,a) = \mathrm{Law}[r + \gamma\, Z_{\phi'}(s', \pi(s'))]$$

- Distance between distributions

$$d(\mu,\nu) \quad \text{i.e. } W_p,\ \mathrm{MMD}$$

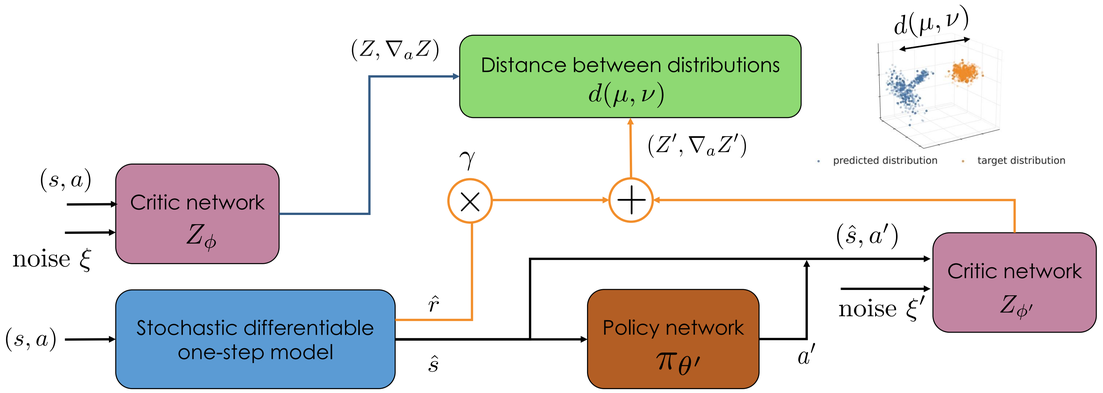
- TD on distributions

$$\mathcal{L}(\phi) = d(\eta_\phi(s,a),\ \eta_{\mathrm{tgt}}(s,a))$$



3

Distributional Value Gradients



- TD on joint distribution

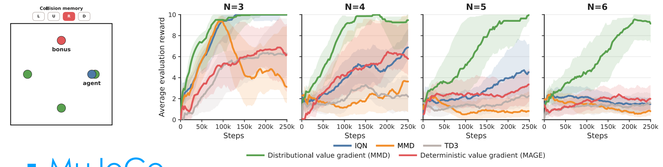
$$\big(Z_\phi(s,a,\xi), \nabla_a Z_\phi(s,a,\xi)\big), \quad \xi \sim \mathcal{N}(0,I)$$

- Stochastic world model
  - Environment aleatoric uncertainty
  - Cheap sampling and gradients
- Theory
  - Contraction in Wasserstein $\gamma\kappa < 1$
  - Smoothness κ
  - Trade-off
    - ↪ Large physical gradients
    - ↪ Lower horizon γ

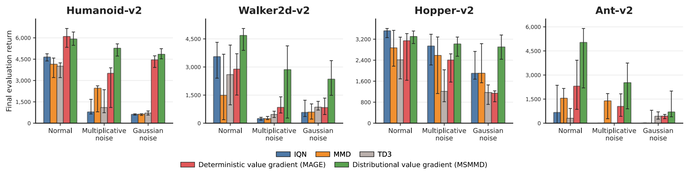
4

Results

- Toy environment



- MuJoCo



5