Bellman GAN:

Distributional Multivariate Policy Evaluation and Exploration

Dror Freirich, Tzahi Shimkin, Ron Meir, Aviv Tamar

Viterbi Faculty of Electrical Engineering
Technion

ICML 2019

Outline

- Distributional RL $\longleftrightarrow$ GANs
- Multivariate rewards
- Exploration

Distributional RL

Objective

Learning value distribution, rather than expectation

$$Z^\pi(s,a) \overset{D}{=} \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \quad ; \quad s_0 = s, a_0 = a$$

Z obeys distributional Bellman equation – Fixed Point!

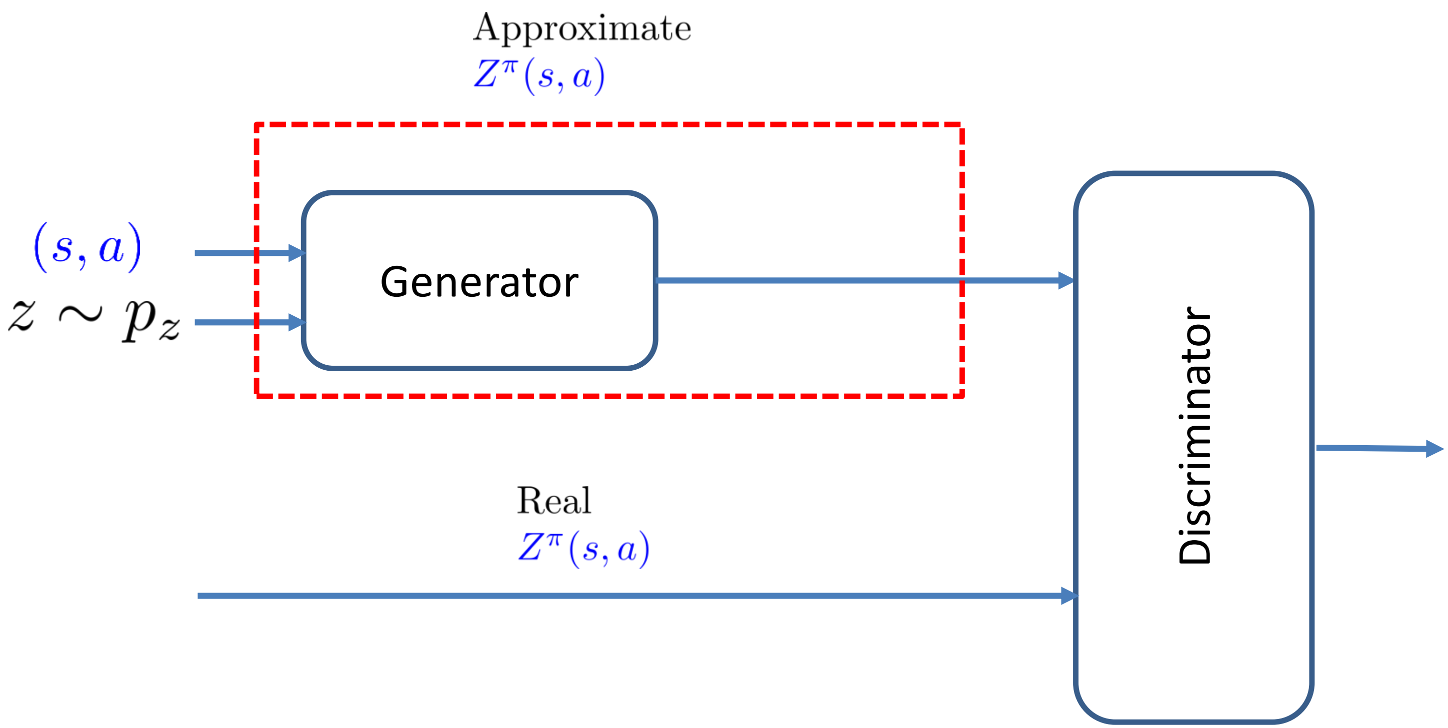
$$Z^\pi(s,a) \overset{D}{=} T^\pi Z^\pi(s,a)$$

Distributional Bellman operator

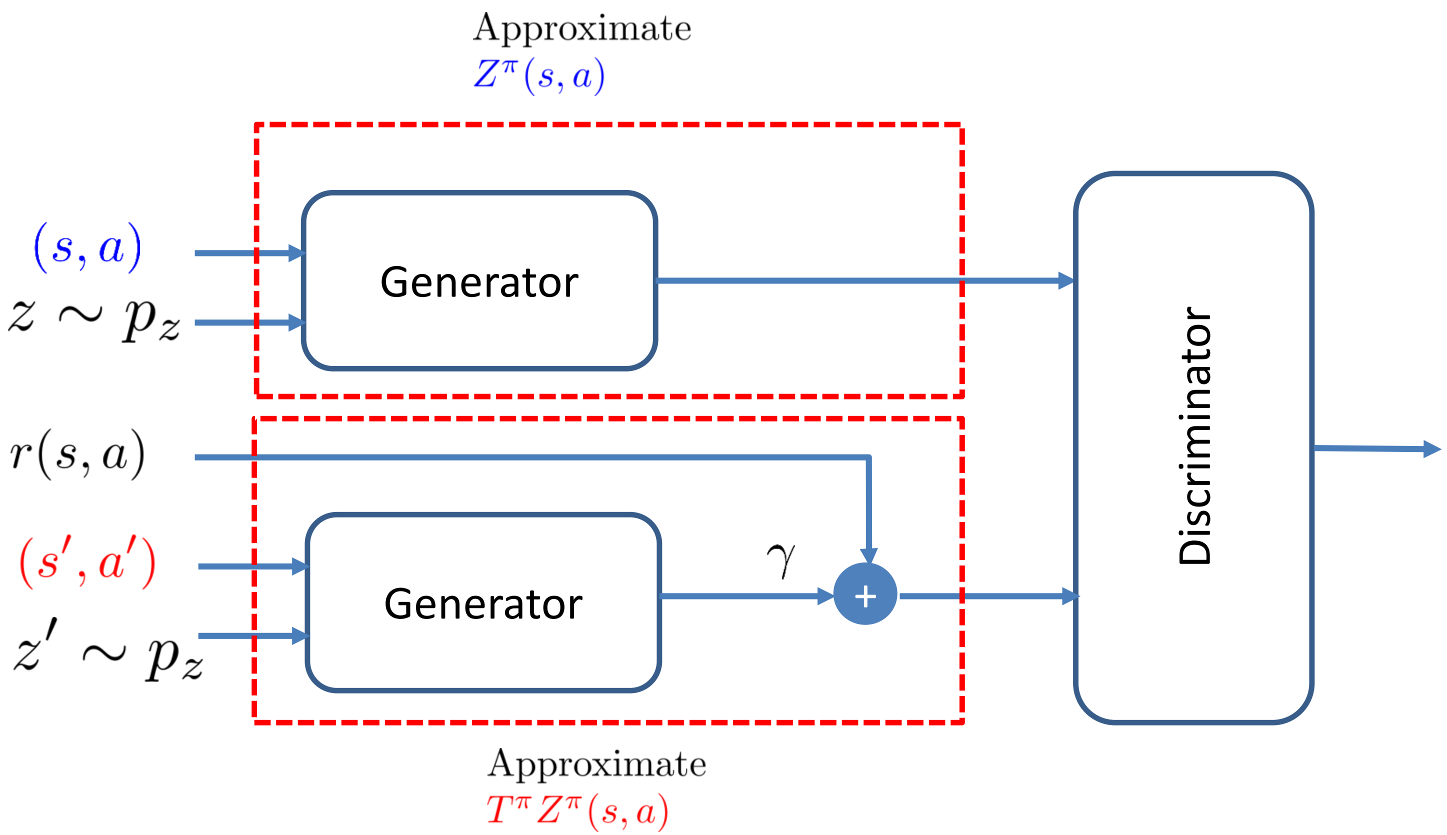
$$T^\pi Z^\pi(s,a) \triangleq R(s,a) + \gamma Z^\pi(s', a')$$

Bellemare et al, ICML 2017

Bellman GAN


Approximate
$Z^{\pi}(s,a)$
(s,a)
$z \sim p_z$
Generator
Discriminator
Real
$Z^{\pi}(s,a)$

Bellman GAN



Mapping Distributional Bellman Eqn. to WGAN

High Dimensional Distributions

- GANs learn distributions of high-dim data

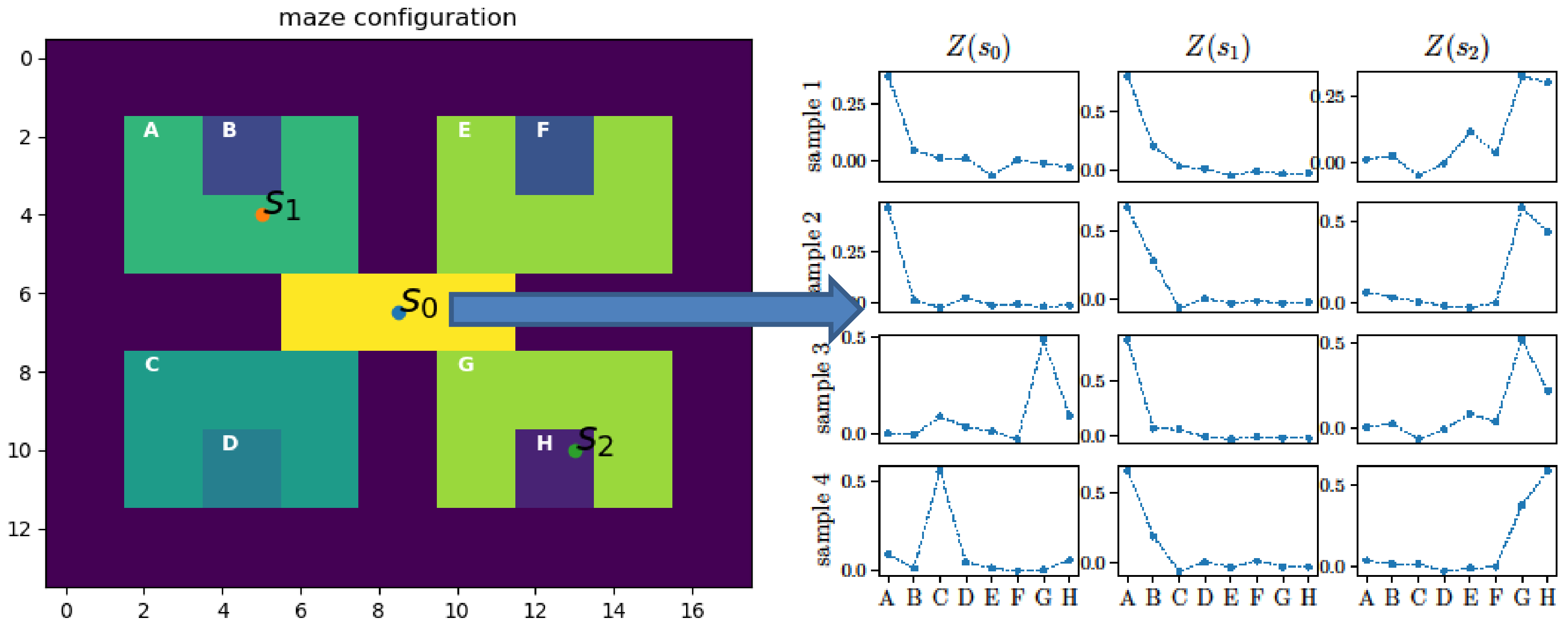
Brock et al, 2018

Main insight Framework applicable to vector rewards $r(s,a) \in \mathbb{R}^d$

Scalable DiRL algorithm for Multi-Objective RL

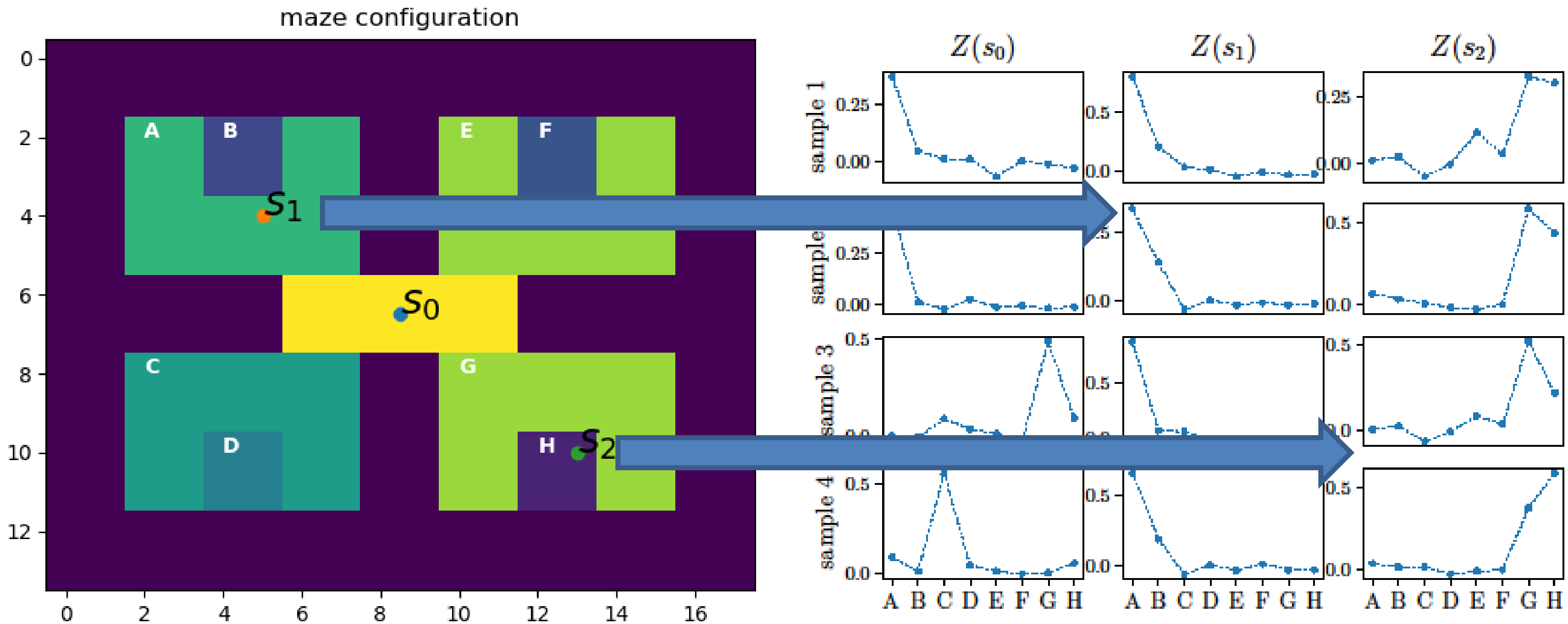
Multi-Reward Policy Evaluation

- Tabular state-space, 4 actions, Random policy.
- 8 reward types, 2 in each room.
- Trained BellGAN, sampled Generator at different locations.



Multi-Reward Policy Evaluation

- Tabular state-space, 4 actions, Random policy.
- 8 reward types, 2 in each room.
- Trained BellGAN, sampled Generator at different locations.



Model Learning

Multivariate Bellman equation

$$Z^\pi(s,a) \stackrel{D}{=} T^\pi Z^\pi(s,a) \triangleq \tilde{r}(s,a,s') + \tilde{\Gamma} Z^\pi(s',a')$$

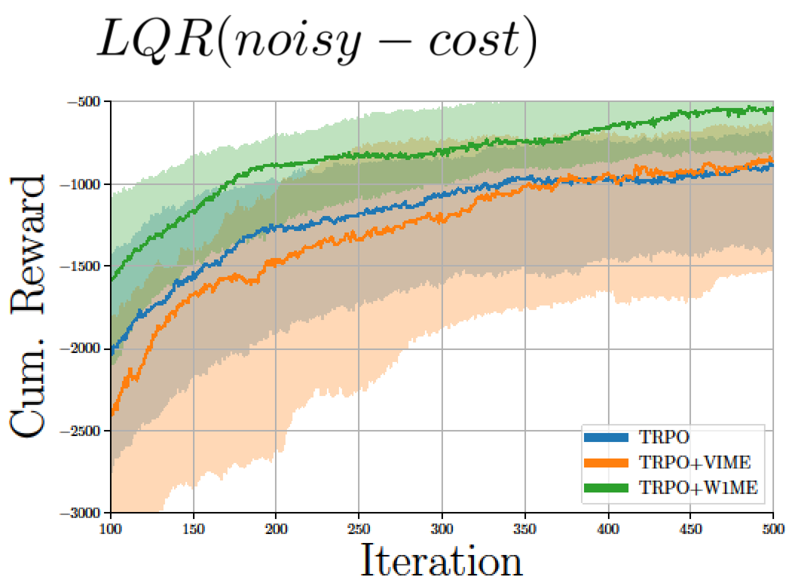
Special case: **Model Learning**

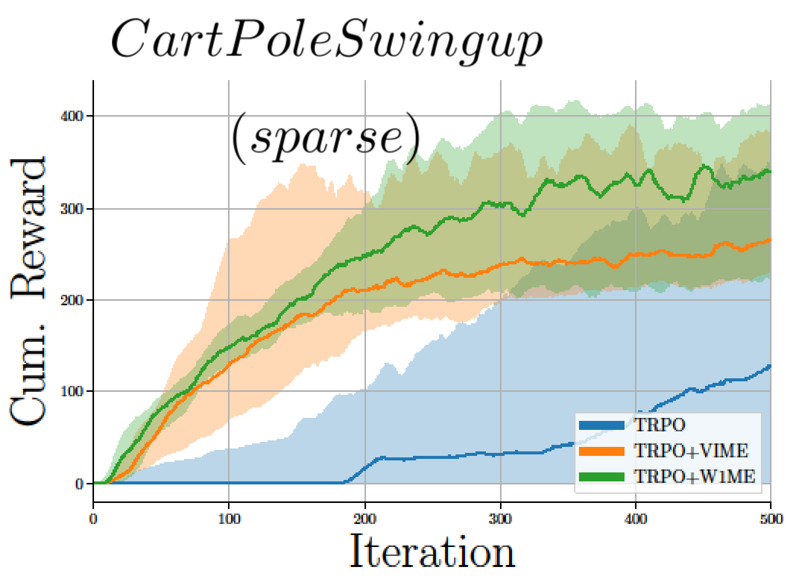
$$\tilde{r}(s,a,s') = \begin{pmatrix} r(s,a,s') \\ s' \end{pmatrix} \quad \tilde{\Gamma} = \begin{pmatrix} \gamma I & 0 \\ 0 & 0 \end{pmatrix}$$

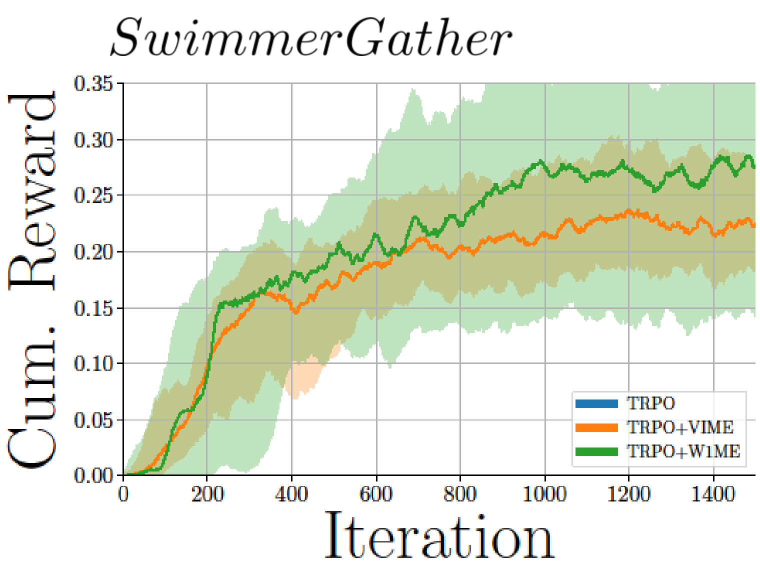
Advantages Framework for learning both **value and transition model** , and the dependencies between them.

Application **Exploration** – change in Wasserstein distance as reward bonus for curiosity.

Continuous Control Experiments







Epilogue

- Equivalence - Distributional Bellman Eqn and GANs
- GAN-based algorithm for DiRL
  - high-dimensional, multivariate rewards
  - Unify learning of return and next state distributions
- Novel exploration method based on DiRL

- Paves the way for a distributional approach to:
  - Multi-objective RL
  - Policy optimization

Thank You !

References

- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Advances in neural information processing systems, pp. 2672–2680, 2014.

- Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. arXiv preprint arXiv:1707.06887, 2017.

- Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. arXiv preprint arXiv:1701.07875, 2017.

- Jürgen Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). IEEE Transactions on Autonomous Mental Development, 2(3):230–247, 2010.

References

- Pierre-Yves Oudeyer, Frdric Kaplan, and Verena V Hafner. Intrinsic motivation systems for autonomous mental development. IEEE transactions on evolutionary computation, 11(2):265–286, 2007.

- Cederic Villani, Optimal transport old and new, 2008

- Brock et al, Large scale GAN training for high fidelity natural image synthesis, September 2018

- Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. Vime: Variational information maximizing exploration. In Advances in Neural Information Processing Systems, pp. 1109–1117, 2016.

Freirich, Shimkin, Meir, T. , Distributional multivariate policy evaluation and exploration with the Bellman GAN, ICML 2019

DiRL Driven Exploration

Bellman GAN objective

$$\mathcal{L}_\pi(G,D) \triangleq E_{z\sim p_z, a_{t+1}\sim\pi(\cdot|s_{t+1})}\Lambda(G_\theta, D_\omega)$$

Intrinsic reward function

$$r^i(s_t, a_t, r_t, s_{t+1}) \triangleq \|E_{z\sim P_z, a_{t+1}\sim\pi(\cdot|s_{t+1})}\nabla_\theta \Lambda(G_\theta, D_\omega)\|$$

Approx. contribution to learning

Combined reward function

$$\hat{r}(s_t, a_t, s_{t+1}) = \underbrace{r(s_t, a_t, s_{t+1})}_{\text{Exploitation}} + \underbrace{\eta r^i(s_t, a_t, r_t, s_{t+1})}_{\text{Exploration}}$$

Apply any RL algorithm