Distributional Reinforcement Learning

Meshal Alharbi
MIT Computational Science and Engineering
meshal@mit.edu

1 Introduction

The majority of reinforcement learning (RL) literature focuses on modeling and learning the expected random return incurred by RL agents. Operating solely based on expectations has a solid footing in the decision and expected utility theories [Von Neumann and Morgenstern, 1947, Barbera et al., 1999]; Nonetheless, recent works argue in favor of modeling and learning the *distribution* of the random returns, which gave rise to a sub-field known as distributional reinforcement learning [Bellemare et al., 2017, Barth-Maron et al., 2018, Mavrin et al., 2019, Urpí et al., 2021]. This project aims to explore the theories and methods behind this field.

1.1 Why learn distributions?

Although distributions provide richer information about RL agents' behavior, learning distributions can generally be considered a modeling choice. We mention several compelling examples where this choice is beneficial:

1. **Risk-sensitive policies:** Optimizing for the expectation of return is, by definition, a risk-neutral behavior. As RL becomes increasingly adapted for real-world problems, being able to learn policies that are sensitive to risks and adhere to constraints will be an intrinsic requirement. Distributional RL offers a natural way of modeling risk-sensitive behavior.
2. **Efficient exploration:** Manly RL algorithms solely employ $\epsilon-$greedy approaches to deal with the exploration-exploitation trade-off. With richer signals about the return distribution, one can easily adapt more elaborate concepts (like optimistic exploration), hopefully leading to better exploration strategies.
3. **Better approximations:** The task of learning distributions might seem harder than learning expectations; Surprisingly, empirical research has shown that distributional RL does not incur higher sample complexity but, on the contrary, lead to a more robust and efficient learning in the context of deep RL.

The early sections of this report provide a general treatment about the theory of learning distributions in a dynamic programming manner. Discussions about specific issues (e.g., distributions parameterization) and applications (e.g., risk-sensitive policies) will be offered later.

1.2 Notations and setting

We consider a time-homogeneous Markov decision process (MDP) described by the tuple $(\mathcal{S}, \mathcal{A}, R, P, \gamma)$. $\mathcal{S}$ and $\mathcal{A}$ are the finite state and the action spaces, $R(s, a)$ is the state-action dependent reward function which we assume to be random with finite first-moment, $P(\cdot|s, a)$ is the transition kernel, and $\gamma \in (0, 1)$ is a discount factor. We use $\pi(\cdot|s)$ to donate a stationary policy that maps states to a distribution over actions. For a policy π, we define its discounted return $Z^\pi(s, a)$ starting from state $s_0 = s$ and after committing to initial action $a_0 = a$ as:

$$Z^\pi(s,a) := \sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \qquad s_{t+1} \sim P(\cdot|s_t, a_t), \quad a_t \sim \pi(\cdot|s_t) \tag{1}$$

The goal of distributional RL is to learn the random quantity $Z^\pi(s, a)$. Note that $Z^\pi(s, a)$ is related to the well known Q-values through $Q^\pi(s, a) = \mathbb{E}[Z^\pi(s, a)]$.

Additionally, for the ease of notations, we define a policy transition operator $P^\pi : \mathcal{Z} \to \mathcal{Z}$ as:

$$P^\pi Z(s,a) := Z(S', A') \qquad S' \sim P(\cdot|s,a), \quad A' \sim \pi(\cdot|S') \tag{2}$$

2 Distributional Bellman Operators

We begin by discussing a dynamic programming approach for learning return distributions. We do that in two parts: first for policy evaluation (fixed π) and then for optimal control (optimizing π).

2.1 Policy evaluation

Using the Markov property and time-homogeneity of the MDP, Bellemare et al. [2017] showed that the return of a fixed policy $Z^\pi(s,a)$ adheres to the following distributional Bellman equation:

$$Z^\pi(s,a) \overset{D}{=} R(s,a) + \gamma P^\pi Z^\pi(s,a) \tag{3}$$

Although this equation seems similar to the standard Bellman equation (in fact, taking expectation reveals the usual Bellman equation), it is fundamentally different; The equation relates the laws that govern the distributions of the two sides and does not imply dependency between random variables. Formally, we say that a two real-valued random variables Z_1 and Z_2 are equal in distribution, and write $Z_1 \overset{D}{=} Z_2$, if for any subset $S \subseteq \mathbb{R}$:

$$\mathbb{P}(Z_1 \in S) = \mathbb{P}(Z_2 \in S)$$

For completeness, we note that it is possible to develop the distributional Bellman equation in a more precise mathematical language that uses probability measures and pushforward maps instead of random variables [Morimura et al., 2010, Rowland et al., 2018].

As per usual, we can then define a fixed policy Bellman operator $\mathcal{T}_\pi$ as:

$$\mathcal{T}_\pi Z(s,a) \overset{D}{:=} R(s,a) + \gamma P^\pi Z(s,a) \tag{4}$$

We now ask questions about the existence and uniqueness of solutions to the equation $Z = \mathcal{T}_\pi Z$. Moreover, if a solution Z^π exist, does the sequence $\{\mathcal{T}_\pi{}^k Z_0\}_{k\geq 0} \to Z^\pi$ for all Z_0? Discussing convergence will require a notion of distance between distributions, and we achieve that by defining a generic metric d on probability space:

Definition 2.1. Let d be an extended metric on the space of probability distributions $d : \mathscr{P}(\mathbb{R}) \times \mathscr{P}(\mathbb{R}) \to [0,\infty]$. Further, for all $\gamma, \alpha \in (0,1)$ and independent random variable W, d is:

(i) *c-homogeneous*: $d(\gamma Z, \gamma Z') = \gamma^c d(Z, Z')$

(ii) *regular*: $d(Z + W, Z' + W) \leq d(Z, Z')$

(iii) *p-convex*: $d^p(\alpha Z_1 + (1-\alpha)Z_2, \alpha Z_1' + (1-\alpha)Z_2') \leq \alpha d^p(Z_1, Z_1') + (1-\alpha) d^p(Z_2, Z_2')$

Bellemare et al. [2022] showed that the p-Wasserstein metric, ℓ_p probability metric, and maximum mean discrepancy satisfy the conditions of d. Refer to [Rachev et al., 2013] for a formal treatment of probability metrics. We also define the supremum extension $\bar{d}$ as:

$$\bar{d}(Z, Z') = \sup_{s,a} d(Z(s,a), Z'(s,a)) \tag{5}$$

One can shows that $\bar{d}$ possess the sufficient conditions for the contractivity of $\mathcal{T}_\pi$.

Lemma 2.1. *The operator $\mathcal{T}_\pi$ is γ^c contraction in $\bar{d}$.*

Proof. For any two return distributions Z and Z':

$$\begin{aligned}
d^p(\mathcal{T}_\pi Z, \mathcal{T}_\pi Z') &= d^p(R(s,a) + \gamma P^\pi Z(s,a), R(s,a) + \gamma P^\pi Z'(s,a)) \\
&\overset{(a)}{\leq} d^p(\gamma P^\pi Z(s,a), \gamma P^\pi Z'(s,a)) \\
&\overset{(b)}{=} \gamma^{cp} d^p(P^\pi Z(s,a), P^\pi Z'(s,a)) \\
&\overset{(c)}{\leq} \gamma^{cp} P^\pi d^p(Z(s,a), Z'(s,a)) \\
&\leq \gamma^{cp} \sup_{s,a} d^p(Z(s,a), Z'(s,a))
\end{aligned}$$

Where (a) follows from the regularity of d, (b) from the c-homogeneity, and (c) from an extension of p-convexity to the mixtures of finitely-many distributions. Then:

$$\begin{aligned}\bar{d}^p(\mathcal{T}_\pi Z, \mathcal{T}_\pi Z') &= \sup_{s,a} d^p(Z(s,a), Z'(s,a)) \\ &\leq \gamma^{cp} \sup_{s,a} d^p(Z(s,a), Z'(s,a)) \\ &= \gamma^{cp} \bar{d}^p(Z, Z')\end{aligned}$$

□

Rösler [1992] has established a fixed-point theorem for distributions analogous to Banach fixed-point theorem, but contraction alone is not sufficient to invoke it; Notice that $\bar{d}$ can take infinite values, rendering the contraction bound meaningless. Thus, the return distributions and the Bellman operator must be restricted to a subset of probability space where all distances are finite:

Definition 2.2 (finite domain of d)**.** Let d be a probability metric. Its finite domain $\mathscr{P}_d(\mathbb{R}) \subseteq \mathscr{P}(\mathbb{R})$ is the set of probability distributions with finite first moment and finite d-distance to the distribution that puts all of its mass on zero:

$$\mathscr{P}_d(\mathbb{R}) = \{Z \in \mathscr{P}(\mathbb{R}) \mid d(Z, \delta_0) < \infty,\ \mathbb{E}[|Z|] < \infty\}$$

Finally, we state the main theorem in the case of policy evaluation:

Theorem 2.1. *Let d be a metric satisfying Definition 2.1 with a finite domain $\mathscr{P}_d(\mathbb{R})$. Suppose $\mathscr{P}_d(\mathbb{R})$ is closed under the operator $\mathcal{T}_\pi$ and $Z^\pi, Z_0 \in \mathscr{P}_d(\mathbb{R})$. Then:*

(a) Z^π is the unique solution to $Z = \mathcal{T}_\pi Z$ in $\mathscr{P}(\mathbb{R})$.

(b) $\{{\mathcal{T}_\pi}^k Z_0\}_{k\geq 0} \to Z^\pi$ with respect to $\bar{d}$ with rate γ^c.

In terms of p-Wasserstein metric, the conditions of Theorem 2.1 translate to the rewards having bounded p^{th} moments. The generic proof presented in this section is due to Bellemare et al. [2022].

2.2 Optimal control

Now we discuss the distributional setting when we optimize policies with respect to the criterion:

$$J(\pi^*) = \max_\pi \mathbb{E}[Z^\pi(s,a)] \tag{6}$$

Bellemare et al. [2017] showed that the distributional optimality Bellman operator is:

$$\begin{aligned}\mathcal{T}_* Z(s,a) &\overset{D}{:=} R(s,a) + \gamma P Z(s, \mathcal{G}_Z(s)) \\ \mathcal{G}_Z(s) &:= \arg\max_{a'} \mathbb{E}[P Z(s,a')]\end{aligned} \tag{7}$$

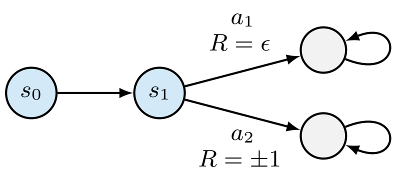
Where $\mathcal{G}_Z(s)$ is a greedy action selection rule. Making $\mathcal{T}_*$ explicitly depends on $\mathcal{G}_Z(s)$ is essential, as different rules will results in different optimal returns. For example, compare the rule (for all actions with equal expectation, choose the one with smallest variance) to (for all actions with equal expectation, choose one randomly). Thus, in general, there are many optimal return distributions.

Unfortunately, the guarantees that $\mathcal{T}_*$ has is weaker than $\mathcal{T}_\pi$. We list some of them in this section.

Lemma 2.2. *The operator $\mathcal{T}_*$ is is not a contraction in any metric.*

Proof. One can construct a simple MDP where $\mathcal{T}_*$ being a contraction in any metric results in a violation of triangular inequality. Figure 1 shows this simple MDP with distances measured using the 1-Wasserstein metric (± 1 means $+1$ and -1 with equal probabilities). □

Note that not being a contraction does not, however, rule out the convergence of $\mathcal{T}_*$ to a fixed point; In the example shown in Figure 1, $\{{\mathcal{T}_*}^k Z_0\}_{k\geq 0}$ eventually convergences to Z^*. On the other hand, this make arguing about the convergence of $\{{\mathcal{T}_*}^k Z_0\}_{k\geq 0}$ more involved than simply appealing to fixed-point theorems.



	$(s_0, \cdot)$	(s_1, a_1)	(s_1, a_2)	$\bar{w}_1(Z^*, \cdot)$
Z^*	$\gamma\epsilon$	ϵ	± 1	
Z_0	$\gamma\epsilon$	$-\epsilon$	± 1	2ϵ
$\mathcal{T}_* Z_0$	$\pm\gamma$	ϵ	± 1	γ
${\mathcal{T}_*}^2 Z_0$	$\gamma\epsilon$	ϵ	± 1	0

Figure 1: An example MDP where $\mathcal{T}_*$ is not a contraction for $\epsilon < 0.5\gamma$.

Lemma 2.3. *Not all optimality operators $\mathcal{T}_*$ have a fixed point $Z^* = \mathcal{T}_* Z^*$.*

Proof. Same example in Figure 1 but now with $\epsilon = 0$. Consider a greedy rule $\mathcal{G}_Z(s)$ that breaks ties by picking a_2 if $Z(s_0) = \gamma\epsilon$ and a_1 otherwise. This will cause $\{{\mathcal{T}_*}^k Z_0\}_{k \geq 0}$ to oscillate. □

It remains an open question whether such pathological examples affect the performance of distributional RL in practice, or if one can design $\mathcal{G}_Z(s)$ to achieve a theoretically more stable behaviour. Current empirical evidences have shown that distributional RL works well in many domains. Nonetheless, these examples affect what can be said about $\mathcal{T}_*$ theoretically [Bellemare et al., 2022]:

Theorem 2.2. *Let d be a metric satisfying Definition 2.1 with a finite domain $\mathscr{P}_d(\mathbb{R})$. let $\mathcal{Z}^*$ be the set of stationary optimal return distributions with mean Q^* and $\mathcal{Z}^{**}$ be the set of nonstationary optimal return distributions. Suppose $Z_0 \in \mathscr{P}_d(\mathbb{R})$. Then:*

(a) $\mathbb{E}[{\mathcal{T}_*}^k Z^0]$ converge geometrically to Q^* with rate γ.

(b) $\lim_{k\to\infty} \inf_{Z^{**} \in \mathcal{Z}^{**}} \bar{d}({\mathcal{T}_*}^k Z^0, Z^{**}) = 0$.

(c) If π^* is the unique optimal policy with return Z^*, then $\{{\mathcal{T}_*}^k Z_0\}_{k \geq 0} \to Z^*$.

3 Further Discussion

In this section, we discuss more practical topics of distributional RL in broader terms.

3.1 Distributions parameterization

Beside simplified settings, any distributional RL algorithm will need to approximate distributions to be computationally feasible. Consider the following classes of parameterized distributions:

$$\mathscr{F}_N = \left\{ \mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 \geq 0 \right\}$$
$$\mathscr{F}_C = \left\{ \sum_1^m p_i \delta_{\theta_i} : p_i \geq 0, \sum_1^m p_i = 1 \right\}$$

The first class uses one normal distribution and parameterize its mean and variance, while the second class discretize the sample space $\mathbb{R}$ at fixed locations $\theta_1, \theta_2, ..., \theta_m$ and parameterize the probabilities at these points. These parameterizations offer different trade-offs in terms of tractability, expressiveness, and uniqueness of the projection $\Pi : \mathscr{P}(\mathbb{R}) \to \mathscr{F}$.

Dabney et al. [2018a] have shown that introducing parameterization amount to solving a projected Bellman equation, where the projections are established on the subspace $\mathscr{P}_d(\mathbb{R})$. In general, one should expect that the solution of the projected Bellman equation will be different than the projection of the solution of the Bellman equation by a factor of $\frac{1}{1-\gamma^c}$ [Bellemare et al., 2022].

3.2 Linear and nonlinear function approximation

Lyle et al. [2019] compared distributional RL with the standard value-based RL in the presence of function approximation. They establish that, in the risk-neutral control, distributional algorithms cannot do better than value-based algorithms when using tabular representation or linear approximation. However, in the presence of nonlinear function approximation, distributional methods performed better and were generally more stable.

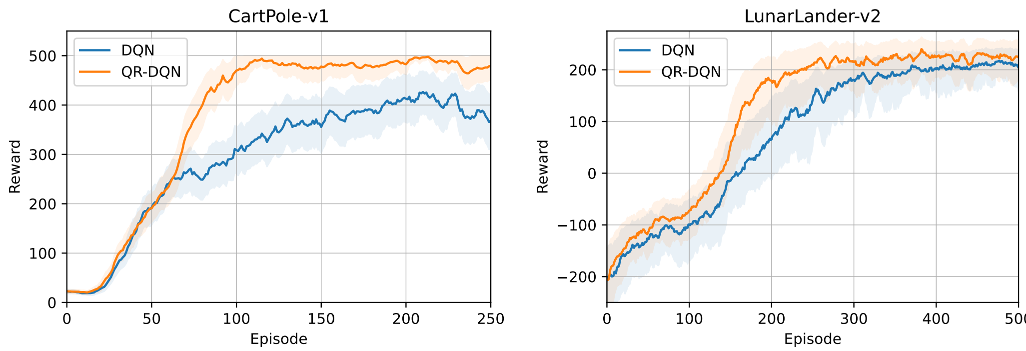


Figure 2: Comparing DQN and QR-DQN with environments from OpenAI Gym library.

3.3 Empirical performance in Deep RL

The literature has shown that the distributional counterpart of popular Deep RL algorithms like DQN and D3PG usually perform better and are more robust; See [Dabney et al., 2018a] for comparison on the Atari domain and [Mavrin et al., 2019] for continuous control tasks. We ran our own experiments comparing DQN with quantile regression DQN (QR-DQN) [Dabney et al., 2018b] using simple environments from OpenAI Gym library. The results are shown in Figure 2, and have been generated by taking the mean of 20 independent seeds.

3.4 Impacts on representation learning

One perspective that has been studied to explain the empirical performance of distributional RL is through its impact on representation learning [Bellemare et al., 2019, Lyle et al., 2021], where distributional RL is believed to generate richer and more general representations. The geometry of the value functions subspace visited by value and policy iteration algorithms have been studied by Dadashi et al. [2019], and they argued in favour of learning representations that are useful across this subspace. The work in [Dabney et al., 2020] considered a similar notion, but focused instead on the path visited by the series of policies generated while learning.

3.5 Efficient exploration

For the purpose of exploration, Mavrin et al. [2019] used the upper quantiles of the learned distributions from QR-DQN to devise optimism bonuses. They empirically showed that their approach outperforms the $\epsilon-$greedy strategy employed in standard QR-DQN. Tang and Agrawal [2018] adapted techniques from the Bayesian bandits literature, in which distributions were used to calculate an entropy bonuses that incentivize exploration.

3.6 Risk-averse behavior

Distributional RL methods have been successfully utilized to optimize risk-sensitive objectives. Singh et al. [2020] were able to optimize for the conditional value at risk (CVaR) in continuous state action spaces using a sample based distributional policy gradient method. Urpí et al. [2021] demonstrated that it is possible to learn risk-sensitive policies in the offline (batch) RL setting. While these works present promising results, the treatment of risk-averse learning seems to be primarily empirical.

Overall, distributional RL appears to follow well-established theories and provides several avenues for further research. In particular, studying the theoretical proprieties of risk-averse Bellman operators seems to be a viable direction. Moreover, offline or batch distributional RL is another underdeveloped and existing research direction.

References

John Von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1947. URL https://press.princeton.edu/books/paperback/9780691130613/theory-of-games-and-economic-behavior.

Salvador Barbera, Peter Hammond, and Christian Seidl. *Handbook of Utility Theory, Volume 1: Principles*. Springer New York, NY, 1999. URL https://link.springer.com/book/9780792381747.

Marc G. Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 449–458. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/bellemare17a.html.

Gabriel Barth-Maron, Matthew W Hoffman, David Budden, Will Dabney, Dan Horgan, Dhruva Tb, Alistair Muldal, Nicolas Heess, and Timothy Lillicrap. Distributed distributional deterministic policy gradients. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. URL https://doi.org/10.48550/arXiv.1804.08617.

Borislav Mavrin, Hengshuai Yao, Linglong Kong, Kaiwen Wu, and Yaoliang Yu. Distributional reinforcement learning for efficient exploration. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4424–4434. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/mavrin19a.html.

Núria Armengol Urpí, Sebastian Curi, and Andreas Krause. Risk-averse offline reinforcement learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2102.05371.

Tetsuro Morimura, Masashi Sugiyama, Hisashi Kashima, Hirotaka Hachiya, and Toshiyuki Tanaka. Parametric return density estimation for reinforcement learning. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, 2010. URL https://arxiv.org/abs/1203.3497.

Mark Rowland, Marc Bellemare, Will Dabney, Remi Munos, and Yee Whye Teh. An analysis of categorical distributional reinforcement learning. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 29–37. PMLR, 09–11 Apr 2018. URL https://proceedings.mlr.press/v84/rowland18a.html.

Marc G. Bellemare, Will Dabney, and Mark Rowland. *Distributional Reinforcement Learning*. MIT Press, 2022. http://www.distributional-rl.org.

Svetlozar T Rachev, Lev B Klebanov, Stoyan V Stoyanov, and Frank Fabozzi. *The methods of distances in the theory of probability and statistics*, volume 10. Springer, 2013. doi:10.1007/978-1-4614-4869-3.

Uwe Rösler. A fixed point theorem for distributions. *Stochastic Processes and their Applications*, 42(2):195–214, 1992. doi:10.1016/0304-4149(92)90035-O.

Will Dabney, Georg Ostrovski, David Silver, and Remi Munos. Implicit quantile networks for distributional reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1096–1105. PMLR, 10–15 Jul 2018a. URL https://proceedings.mlr.press/v80/dabney18a.html.

Clare Lyle, Marc G Bellemare, and Pablo Samuel Castro. A comparative analysis of expected and distributional reinforcement learning. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, volume 33, pages 4504–4511, 2019. URL https://ojs.aaai.org/index.php/AAAI/article/view/4365.

Will Dabney, Mark Rowland, Marc Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, volume 32, 2018b. URL https://ojs.aaai.org/index.php/AAAI/article/view/11791.

Marc Bellemare, Will Dabney, Robert Dadashi, Adrien Ali Taiga, Pablo Samuel Castro, Nicolas Le Roux, Dale Schuurmans, Tor Lattimore, and Clare Lyle. A geometric perspective on optimal representations for reinforcement learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/3cf2559725a9fdfa602ec8c887440f32-Paper.pdf.

Clare Lyle, Mark Rowland, Georg Ostrovski, and Will Dabney. On the effect of auxiliary tasks on representation dynamics. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1–9. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/lyle21a.html.

Robert Dadashi, Adrien Ali Taiga, Nicolas Le Roux, Dale Schuurmans, and Marc G Bellemare. The value function polytope in reinforcement learning. In *International Conference on Machine Learning*, pages 1486–1495. PMLR, 2019. URL https://arxiv.org/abs/1901.11524.

Will Dabney, André Barreto, Mark Rowland, Robert Dadashi, John Quan, Marc G Bellemare, and David Silver. The value-improvement path: Towards better representations for reinforcement learning. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence*, 2020. URL https://ojs.aaai.org/index.php/AAAI/article/view/16880.

Yunhao Tang and Shipra Agrawal. Exploration by distributional reinforcement learning. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 2710–2716. International Joint Conferences on Artificial Intelligence Organization, 7 2018. URL https://doi.org/10.24963/ijcai.2018/376.

Rahul Singh, Qinsheng Zhang, and Yongxin Chen. Improving robustness via risk averse distributional reinforcement learning. In *Learning for Dynamics and Control*, pages 958–968. PMLR, 2020. URL https://arxiv.org/abs/2005.00585.