Distributional Reinforcement Learning and Quantile Optimization

Internship Oral Defense

Alexandre Marthe

ENS de Lyon

June 7, 2023

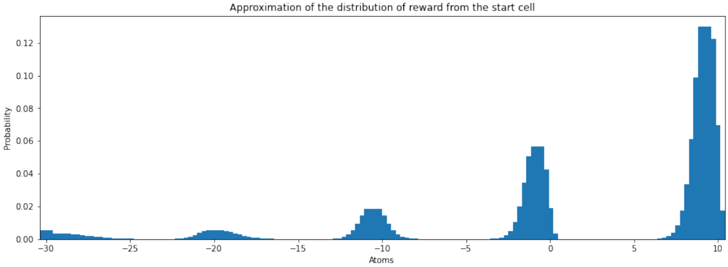


Figure: Example of distribution where the mean gives little information

Table of Contents

Table of Contents

Markov Decision Process[1]

Definition (Markov Decision Processes)

An MPD is a tuple $\mathcal{M}(\mathcal{X}, \mathcal{A}, P, \gamma)$, where:

- $\mathcal{X}$ is a finite state space
- $\mathcal{A}$ a finite action space
- P a transition probability kernel that assigns to each pair $(x, a) \in \mathcal{X} \times \mathcal{A}$ a probability measure on $\mathcal{X} \times \mathbb{R}$
- $\gamma \in [0, 1[$ the discount

The *return*, that we aim to optimize is:

$$R = \mathbb{E}\left[r(x_0, a_0) + \gamma r(x_1, a_1) + \gamma^2 r(x_2, a_2) + \dots\right] = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(x_t, a_t)\right]$$

Definition

A decisision rule d is a function that maps each state to a probability distribution on the action space :

$$d : \mathcal{X} \mapsto \mathscr{P}(\mathcal{A})$$

It is said *deterministic* if it of the form d: $\mathcal{X} \mapsto \mathcal{A}$

Definition

A policy is a sequence of decision rule:

$$\pi = (d_0, d_1, d_2, \dots)$$

It is said *stationnary* if it uses a unique decision rule.

Theorem (Bertsekas, 2007)

If an optimal policy exists, it can be chosen to be stationnary.

Proposition (Bellman optimality principle[2])

An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision.

Corollary

If an optimal policy exists, then it can be chosen to be deterministic.

Definition

The Value functions V and Q are defined by:

$$V^{\pi}(x) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(x_t, a_t) | x_0 = x\right]$$

$$Q^{\pi}(x, a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(x_t, a_t) | x_0 = x, a_0 = a\right]$$

with $x_t \sim p(\cdot|x_{t-1}, a_{t-1})$ and $a_t \sim \pi(\cdot|x_t)$

Definition

The Optimal Value functions $V^\star$ and $Q^\star$ are defined by:

$$V^\star(x) = \max_{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(x_t, a_t) | x_0 = x\right] = V^{\pi^\star}(x)$$

$$Q^\star(x, a) = \max_{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(x_t, a_t) | x_0 = x, a_0 = a\right] = Q^{\pi^\star}(x, a)$$

Definition (Bellman Operator)

Let $V : \mathcal{X} \mapsto \mathbb{R}$ or $Q : \mathcal{X} \times \mathcal{A} \mapsto \mathbb{R}$, π a policy. The Bellman operator $\mathcal{T}^\pi$ is defined by:

$$\forall x \in \mathcal{X}, \qquad \mathcal{T}^\pi V(x) = \sum_{a \in \mathcal{A}} \pi(a|x) \left(\mathbb{E}\left[r(x,a)\right] + \gamma \sum_{x' \in \mathcal{X}} p(x'|x,a) V(x') \right)$$

$$\forall x, a \in \mathcal{X} \times \mathcal{A}, \qquad \mathcal{T}^\pi Q(x,a) = \mathbb{E}\left[r(x,a)\right] + \gamma \sum_{x', a' \in \mathcal{X} \times \mathcal{A}} p(x'|x,a) \pi(a'|x') Q(x',a')$$

Definition (Optimal Bellman Operator)

Let $V : \mathcal{X} \mapsto \mathbb{R}$ or $Q : \mathcal{X} \times \mathcal{A} \mapsto \mathbb{R}$, π a policy. The Bellman operator $\mathcal{T}^\star$ is defined by:

$$\forall x \in \mathcal{X}, \qquad \mathcal{T}^\star V(x) = \max_{a \in \mathcal{A}} \mathbb{E}\left[r(x,a)\right] + \gamma \sum_{x' \in \mathcal{X}} p(x'|x,a) V^\star(x')$$

$$\forall x, a \in \mathcal{X} \times \mathcal{A}, \qquad \mathcal{T}^\star Q(x,a) = \mathbb{E}\left[r(x,a)\right] + \gamma \sum_{x' \in \mathcal{X}} p(x'|x,a) \max_{a' \in \mathcal{A}} Q^\star(x',a')$$

Proposition

The Bellman Operators are γ-contractions.

Theorem (Banach fixed point[3])

Let (X, d) be a non-empty complete metric space with a contraction mapping $T : X \mapsto X$. Then T has admits a unique fixed-point $x^\star \in X$ and

$$\forall x \in X, \quad T^n(x) \longrightarrow x^\star \text{ exponentially}$$

Corollary (Algorithms)

Iterating the Bellman operators is an algorithm to compute the value functions.

Table of Contents

Metrics

Definition (Wasserstein Metric[4])

Let $p \geq 1$ and $\mathscr{P}_p(\mathbb{R})$ the space of distributions with finite p^{th} moment. Let $\nu_1, \nu_2 \in \mathscr{P}_p(\mathbb{R})$ with respective cumulative distribution function F and G. The p-Wasserstein distance d_p is then defined as :

$$d_p(\nu_1, \nu_2) = \left(\int_0^1 \left| F^{-1}(u) - G^{-1}(u) \right|^p du \right)^{\frac{1}{p}}$$

Definition ([4])

Let $\nu_1, \nu_2 \in \mathscr{P}(\mathbb{R})$. We define the family of metrics ℓ_p by :

$$\ell_p(\nu_1, \nu_2) = \left(\int_{\mathbb{R}} (F_{\nu_1}(x) - F_{\nu_2}(x))^p \mathrm{d}x \right)^{\frac{1}{p}}$$

ℓ_2 is called the Cramer distance.

Framework

The random return is the sum of the discounted random reward:

$$Z(x,a) = \sum_{t=0}^{\infty} \gamma R_t \mid X_0 = x, A_0 = a \tag{1}$$

The idea is that the distribution of the reward would follow similar Bellman equations:

$$Z(x,a) \overset{D}{=} R(x,a) + \gamma Z(X', A') \tag{2}$$

with X', A' the random next state-action.

Policy Evaluation

Let's consider a policy π. The distribution of the random return under π will be written as follows:

$$\eta_\pi^{(x,a)} = \text{Law}_\pi\left(\sum_{t=0}^{\infty} \gamma R_t \mid X_0 = x, A_0 = a\right)$$

The random return associated to policy π verifies the *distributional Bellman equation*:

$$\eta_\pi = \mathcal{T}^\pi \eta_\pi$$

where $\mathcal{T}^\pi$ is the Bellman operator defined by:

$$\mathcal{T}^\pi \eta^{(x,a)} = \int_{\mathbb{R}} \sum_{(x',a') \in \mathcal{X} \times \mathcal{A}} (f_{r,\gamma})_\# \eta^{(x',a')} \pi(a'|x') p(r, x'|x, a) dr$$

with $(f_{r,\gamma})_\# \eta$ is the pushforward measure define by $f_\# \eta(A) = \eta(f^{-1}(A))$ for all Borel sets $A \subseteq R$ and $f_{r,\gamma}(x) = r + \gamma x$ for all $x \in R$.

Proposition

$\mathcal{T}^\pi$ is a γ-contraction under the maximal p-Wasserstein metric $\overline{d}_p$ (for $p \geq 1$).

Corollary

$$\forall \eta \in \mathscr{P}(\mathbb{R})^{\mathcal{X} \times \mathcal{A}}, \quad (\mathcal{T}^\pi)^n \eta \underset{n \to \infty}{\longrightarrow} \eta_\pi$$

with an exponential convergence for the norm $\overline{d}_p$.

Control

We define by optimal distribution a distribution associated to an optimal policy:

$$\eta^\star \in \{\eta_{\pi^\star} \mid \pi^\star \in \arg\max_\pi \mathbb{E}_{R\sim\eta_\pi}[R]\}$$

As expected, the optimal distributions verify the optimal distributional Bellman equation: $\eta^\star = \mathcal{T}\eta^\star$ with

$$\mathcal{T}\eta^{(x,a)} = \int_{\mathbb{R}} \sum_{(x',a')\in\mathcal{X}\times\mathcal{A}} (f_{r,\gamma})_\# \eta^{(x',a^\star(x'))} p(r,x'|x,a)dr$$

where $a^\star(x') = \arg\max_{a'\in\mathcal{A}} \mathbb{E}_{R\sim\eta^{(x',a')}}[R]$

Lemma

Let $\eta_1, \eta_2 \in \mathscr{P}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$*, we write* $\mathbb{E}[\eta] := \mathbb{E}_{Z\sim\eta}[Z]$*. Then:*

$$\|\mathbb{E}[\mathcal{T}\eta_1] - \mathbb{E}[\mathcal{T}\eta_2]\|_\infty \leq \gamma \|\mathbb{E}[\eta_1] - \mathbb{E}[\eta_2]\|_\infty$$

Which means that $\mathbb{E}[\mathcal{T}^n\eta] \underset{n\to\infty}{\longrightarrow} Q^\star$ *exponentially quickly.*

But also:

Theorem

Let $\mathcal{X}$ *and* $\mathcal{A}$ *be finite. Let* $\eta \in \mathscr{P}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$*. There exist an optimal policy* $\pi^\star$ *(potentially nonstationary), such that:*

$$\mathcal{T}^n\eta \underset{n\to\infty}{\longrightarrow} \eta_{\pi^\star} \text{ uniformly in } \overline{d}_p,\ p \geq 1$$

Proposition

The optimality operators are not always contractions.

Proposition

The optimality operators do not always have fixed points.

Distribution approximations

Main issue: we need a way to approximate the distributions. We have parametrizations:

- Categorical Parametrization
- Quantile Parametrization

Categorical Approach

The idea is to use the hypothesis of bounded reward to use evenly spread diracs on that reward support, and use the diracs weight as the parameters.

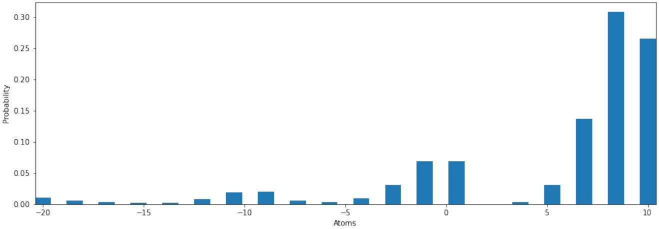


Figure: Example of a distribution projected by with the categorical approach

The projection operator is defined by:

$$\Pi_C(\delta_y) = \begin{cases} \delta_{z_0} & y \leq z_0 \\ \frac{z_{i+1}-y}{z_{i+1}-z_i}\delta_{z_i} + \frac{y-z_i}{z_{i+1}-z_i}\delta_{i+1} & z_i < y < z_{i+1} \\ \delta_{z_{N-1}} & y \geq z_{N-1} \end{cases} \quad (3)$$

Proposition

$\Pi_C \mathcal{T}^\pi$ is not a contraction for $\overline{d}_p$ with $p > 1$.

Proposition

$\Pi_C \mathcal{T}^\pi$ is a $\sqrt[p]{\gamma}$-contraction in $\overline{\ell}_p$.

$$\exists! \eta_C \in \mathcal{P}_C^{\mathcal{X} \times \mathcal{A}}, \forall \eta_0 \in \mathscr{P}(\mathbb{R})^{\mathcal{X} \times \mathcal{A}}, \quad (\Pi_C \mathcal{T}^\pi)^m \eta_0 \underset{m \to \infty}{\longrightarrow} \eta_C \quad \text{exponentially quickly in } \tag{4}$$

Lemma

Let η_C defined as in (4). Assume that η_π is supported on $[z_0, z_{N-1}]$. Then:

$$\overline{\ell}_2(\eta_C, \eta_\pi) \leq \frac{1}{1-\gamma} \Delta z$$

Quantile Approach

We define the quantile projection operator by $\Pi_{d_1}\nu = \frac{1}{N}\sum_{i=0}^{N-1}\delta_{z_i}$ with $z_i = F^{-1}\left(\frac{2i+1}{2N}\right)$. This leads to a minization of the Wasserstein metrics between the true distribution and the parametrized space.

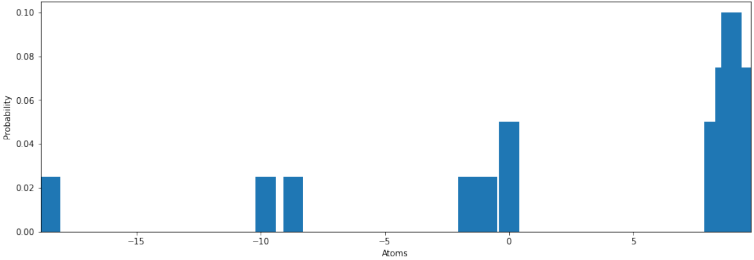


Figure: Example of a distribution projected with the quantile approach

Proposition

$\Pi_{d_1}\mathcal{T}^\pi$ is γ-contraction in $\overline{d}_\infty$:

$$\overline{d}_\infty(\Pi_{d_1}\mathcal{T}^\pi\eta_1, \Pi_{d_1}\mathcal{T}^\pi\eta_2) \leq \gamma\overline{d}_\infty(\eta_1, \eta_2)$$

Table of Contents

Framework

We are still considering MDPs of the form $\mathcal{M}(\mathcal{X}, \mathcal{A}, P, R, \gamma)$, but with another value to optimize. We consider $x \in \mathcal{X}$ a specific state, and $\tau \in [0, 1]$ the quantile of interest. Our objective is:

$$\max_{\pi} V_\tau(x) = q_\tau \left(\sum_{t=0}^{\infty} \gamma R_t \mid X_0 = x \right)$$

Cliff environment

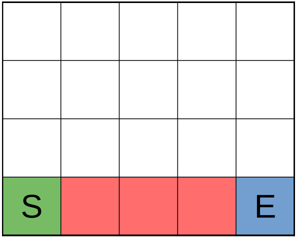


Figure: State space of the Cliff environment

The reward received when reaching E is set to 10. The reward received when falling is set to -10.
The agent can move in the 4 directions, but has only 0.7% chances to go in the chosen direction, and has 0.1% chances to go any other direction.

Policy Evaluation

In Practice:

- Iterating the Bellman algorithm and compute the quantile of the output distribution works well

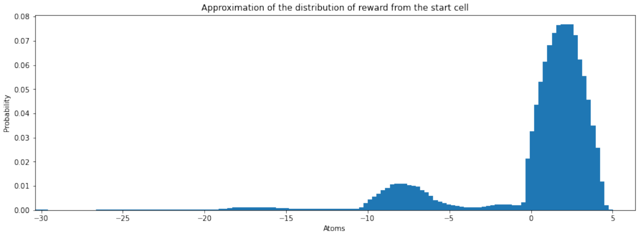
In Theory:

- No guaranteed bound on the difference between the computed quantile and the real one.

$$(\mathcal{T}^{\pi})^n \eta \underset{n\to\infty}{\longrightarrow} \eta_\pi \quad \not\Rightarrow \quad q_\tau\left((\mathcal{T}^{\pi})^n \eta\right) \underset{n\to\infty}{\longrightarrow} q_\tau(\eta_\pi)$$

→	→	→	→	↓
↑	↑	↑	↑	↓
↑	↑	↑	↑	↓
S				E

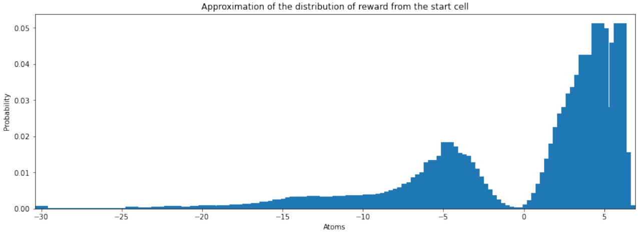
(a) Safe policy



(b) Distribution of return

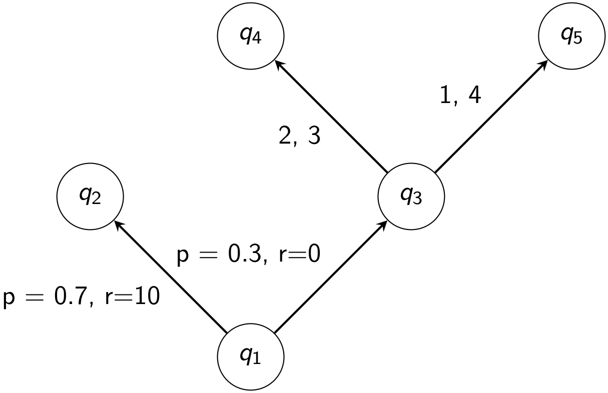
→	→	→	→	↓
→	→	→	→	↓
→	→	→	→	↓
S				E

(a) Risky policy



(b) Distribution of return

Counter example Bellman Optimality Principle



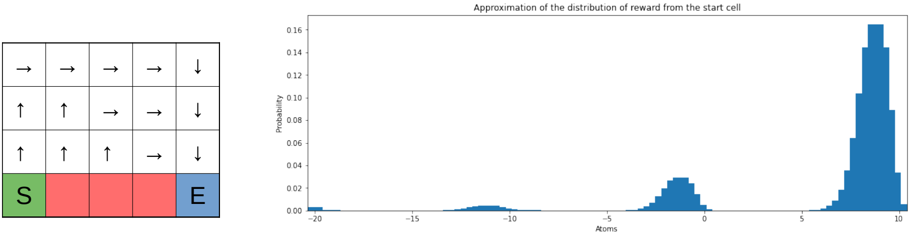


Figure: Behavior on mean optimization $\gamma = 0.99$

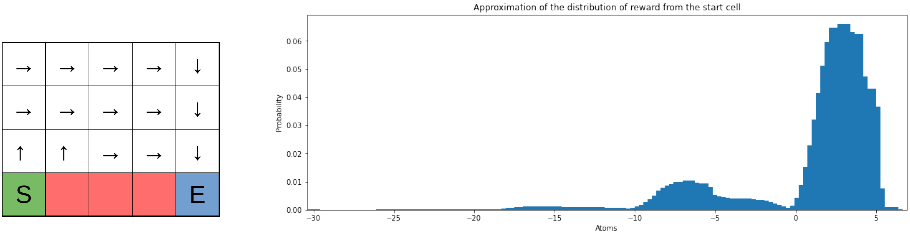


Figure: Behavior on mean optimization, $\gamma = 0.9$

Median case

No convergence:

→	↓	↓	↓	↓
→	→	→	↓	↓
→	→	→	→	↓
S				E

(a) 1st output policy

→	↓	↓	↓	↓
→	↓	→	↓	↓
→	→	→	→	↓
S				E

(b) 2nd output policy

↓	→	↓	↓	↓
←	↓	→	↓	↓
↑	→	→	→	↓
S				E

(c) 3rd output policy

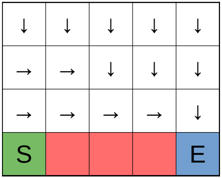
→	→	↓	↓	↓
↓	→	→	↓	↓
→	→	→	→	↓
S				E

(d) 4th output policy

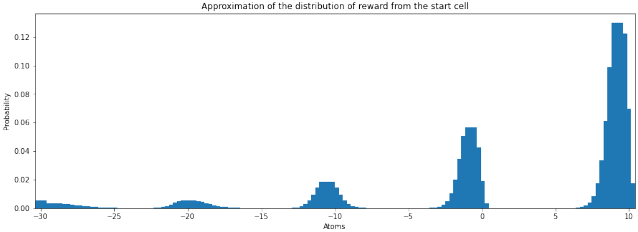
Figure: Policies output by median optimisation

Issues with equal medians due to the distribution approximation. The medians were still higher than the mean case.

Quantile Case



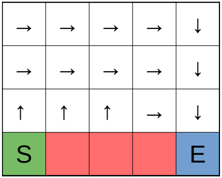
(a) output policy



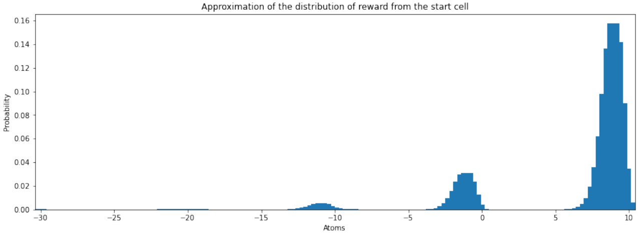
(b) Distribution of return

Figure: Behavior on 0.8 quantile optimazation

The policy is risky, as expected, and the quantile 0.8 is higher than the mean case.



(a) output policy



(b) Distribution of return

Figure: Behavior on 0.2 quantile optimazation

The policy isn't much safer, and the quantile 0.2 is lower than in the mean case.

About deterministic policies

Lemma

Let $n \in \mathbb{N}$, let $0 \leq \lambda_1, \lambda_2, \ldots, \lambda_n \leq 1$ such that $\sum_{i=0}^{n} \lambda_i = 1$, and $\mu_1, \ldots, \mu_n$ n distributions. Let q_τ the quantile function for $\tau \in [0,1]$. We have:

$$q_\tau\left(\sum_{i=0}^{n} \lambda_i \mu_i\right) \leq \max_{1 \leq i \leq n} q_\tau(\mu_i)$$

Corollary

Consider a finite MDP where no state can be visited twice (i.e, without any loops). Consider a state $x \in X$, and $\tau \in [0,1]$. There exist an deterministic policy π_x^* that optimizes the τ quantile for state x :

$$V_\tau^{\pi_x^*}(x) = \max_\pi V_\tau^\pi(x)$$

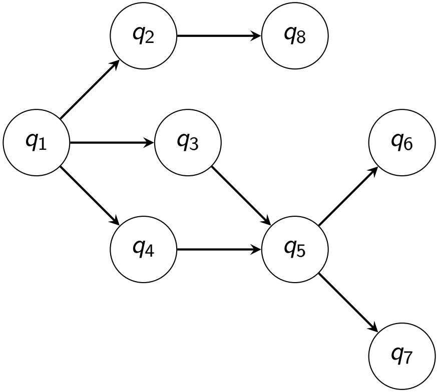


Figure: Example of an MDP on which the corollary applies

Table of Contents

Conclusion

Main work of the intership:

- Find the bibliography and understand the Distributionnal Framework.
- Develop a small librairy to experiment on this distributional framework.
- Experiment on it with quantile Optimization, understand behaviors.
- Find some counter examples and a little theoretical result.

Conclusion of the internship: Quantile Optimization is hard and the theoretical results are sparse. Some results are promising, but a quantity such as the expectile would be better to optimize on.

C. Szepesvári, "Algorithms for reinforcement learning," *Synthesis lectures on artificial intelligence and machine learning*, vol. 4, no. 1, pp. 1–103, 2010.

R. Bellman, "Dynamic programming," *Science*, vol. 153, no. 3731, pp. 34–37, 1966.

W. Rudin, *Functional Analysis*.
International series in pure and applied mathematics, McGraw-Hill, 1991.

M. Rowland, M. Bellemare, W. Dabney, R. Munos, and Y. W. Teh, "An Analysis of Categorical Distributional Reinforcement Learning," in *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, pp. 29–37, PMLR, Mar. 2018.
ISSN: 2640-3498.

M. G. Bellemare, W. Dabney, and R. Munos, "A Distributional Perspectiv e on Reinforcement Learning," *arXiv:1707.06887 [cs, stat]*, July 2017.
arXiv: 1707.06887.

W. Dabney, M. Rowland, M. G. Bellemare, and R. Munos, "Distributional Reinforcement Learning with Quantile Regression," *arXiv:1710.10044 [cs, stat]*, Oct. 2017.
arXiv: 1710.10044.

C. Lyle, M. G. Bellemare, and P. S. Castro, "A Comparative Analysis of Expected and Distributional Reinforcement Learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 4504–4511, July 2019.
Number: 01.

T. Morimura, M. Sugiyama, H. Kashima, H. Hachiya, and T. Tanaka, "Parametric Return Density Estimation for Reinforcement Learning," *arXiv:1203.3497 [cs, stat]*, Mar. 2012.
arXiv: 1203.3497.

R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*.
MIT press, 2018.

Wikipedia contributors, "Markov decision process — Wikipedia, the free encyclopedia," 2022.

D. Bertsekas, *Dynamic programming and optimal control: Volume I*, vol. 1.
Athena scientific, 2012.