Value Function in Frequency Domain and the Characteristic Value Iteration Algorithm

Amir-massoud Farahmand
Vector Institute & University of Toronto, Toronto, Canada

High-level Summary

Motivation: The goal of conventional RL is finding a policy that maximizes the **expected** return. This, however, ignores the distribution of returns. If we want to design a risk-aware RL agent, knowledge of the return distribution can be useful.

Distributional RL:

- Conventional: Learn the probability distribution function of return
- **This work: Learn the characteristic function of returns**

Q: Why should we care?

- A new representation opens up the possibility of designing new algorithms
- Fitting a PDF using MLE might be intractable

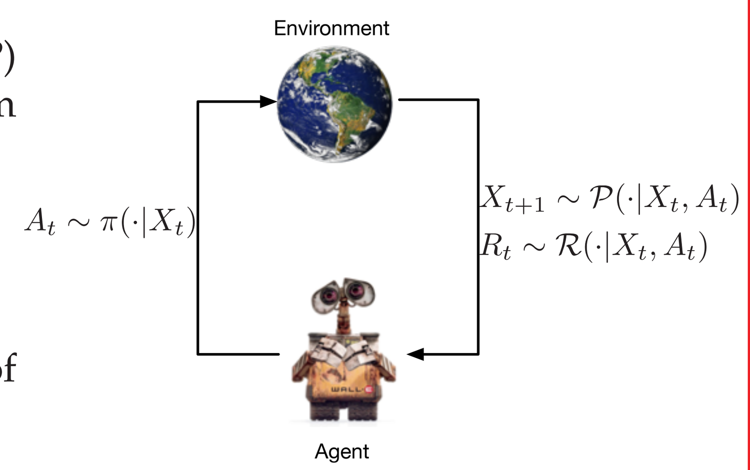
Contributions:

- Bring the frequency domain representation of uncertainty of returns to RL
- Algorithm: Characteristic Value Iteration
- Error Propagation theory
- Function approximation and covering number properties

From Distributional RL to Characteristic Value Function

Consider a discounted Markov Decision Process (MDP) $(\mathcal{X}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \gamma)$. Return of following a policy π starting from state x:

$$G^\pi(x) = \sum_{i \geq 0} \gamma^i R_i.$$



The (conventional) value function V^π is the first moment of this r.v., i.e., $V^\pi(x) = \mathbb{E}\left[G^\pi(X_0)|X_0 = x\right]$.

We have $G^\pi(x) = R_0 + \gamma \sum_{i \geq 0} \gamma^i R_{i+1} = R_0 + \gamma G^\pi(X')$, with $X' \sim \mathcal{P}^\pi(\cdot|X_0 = x)$. The probability distribution (law) of $G^\pi(x)$ is the same as the distribution of $R_0 + \gamma G^\pi(X')$, i.e.,

$$G^\pi(x) \overset{(D)}{=} R_0 + \gamma G^\pi(X'). \qquad \text{(Distributional Bellman Equation)}$$

Let us compute the CF of both sides:

$$\begin{aligned} c_{G^\pi(x)}(\omega) &= \mathbb{E}\left[\exp\left(j\omega G^\pi(x)\right)\right] = \mathbb{E}\left[\exp\left(j\omega\left(R^\pi(x) + \gamma G^\pi(X')\right)\right)\right], \qquad \forall \omega \in \mathbb{R} \\ &= c_{R^\pi(x)}(\omega)\, \mathbb{E}\left[\exp\left(j\omega\gamma G^\pi(X')\right) \mid X = x\right] = c_{R^\pi(x)}(\omega) \int \mathcal{P}^\pi(\mathrm{d}y|x) c_{G^\pi(y)}(\gamma\omega). \end{aligned}$$

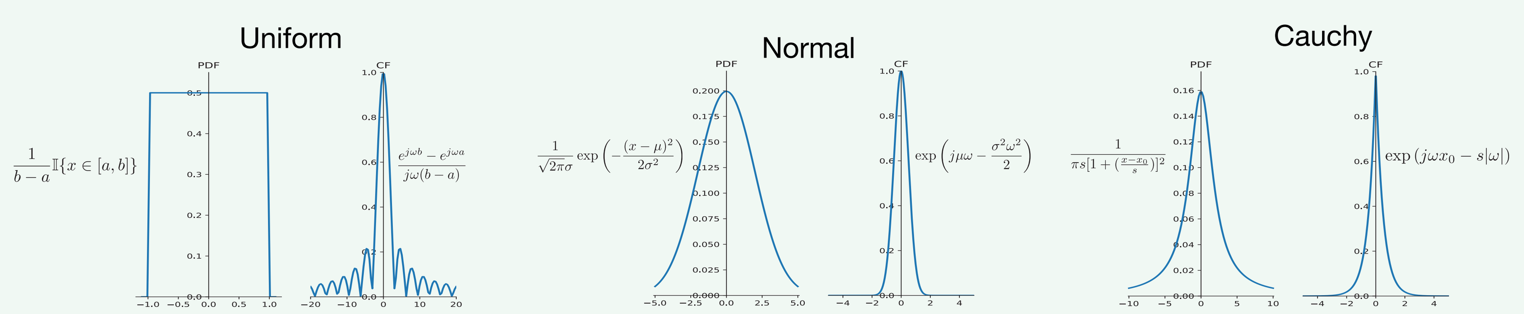
Denote the CF of the reward $c_{R^\pi(x)}(\omega)$ by $\tilde{R}(\omega; x)$, and the CF of the return $c_{G^\pi(x)}(\omega)$ by $\tilde{V}^\pi(\omega; x)$. We call the function $\tilde{V}^\pi : \mathbb{R} \times \mathcal{X} \to \mathbb{C}_1$ the Characteristic Value Function (CVF).

Overview of Characteristic Functions

Given a real-valued r.v. X with the probability distribution μ, its corresponding CF $c_X : \mathbb{R} \to \mathbb{C}$ is

$$c_X(\omega) \triangleq \mathbb{E}\left[e^{jX\omega}\right] = \int \exp(jx\omega)\mu(\mathrm{d}x), \qquad \omega \in \mathbb{R}.$$

- Closely related to Fourier transform of μ.
- Bijection relationship with μ, i.e., if we know one, we can know the other one.
- If X and Y are two (conditionally) independent random variables, $c_{X+Y}(\omega) = c_X(\omega)c_Y(\omega)$.
- $c_{aX+b}(\omega) = e^{jb\omega}c_X(a\omega)$.



Characteristic Value Function

Bellman equation in the frequency domain:

$$\boxed{\tilde{V}^\pi(\omega; x) = \tilde{R}(\omega; x) \int \mathcal{P}^\pi(\mathrm{d}y|x)\tilde{V}^\pi(\gamma\omega; y)}$$

Bellman operator between the CF functions:

$$(\tilde{T}^\pi \tilde{V})(\omega; x) \triangleq \tilde{R}(\omega; x) \int \mathcal{P}^\pi(\mathrm{d}y|x)\tilde{V}(\gamma\omega; y).$$

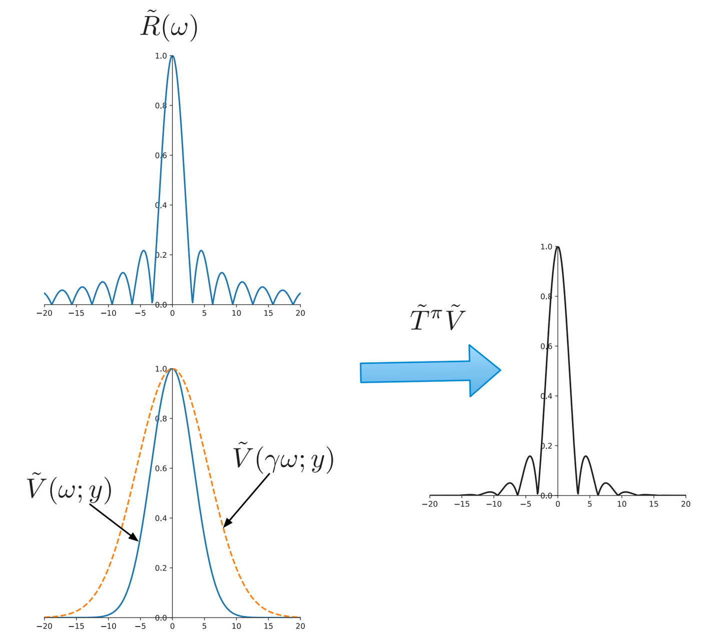
Bellman equation (compact): $\tilde{V}^\pi = \tilde{T}^\pi \tilde{V}^\pi$

This new Bellman operator is a contraction, but not with respect to the supremum norm.

Given $\tilde{V}_1, \tilde{V}_2 : \mathbb{R} \times \mathcal{X} \to \mathbb{R}$, we define

$$d_{\infty,p}(\tilde{V}_1, \tilde{V}_2) \triangleq \sup_{x \in \mathcal{X}} \sup_{\omega \in \mathbb{R}} \left| \frac{\tilde{V}_1(\omega; x) - \tilde{V}_2(\omega; x)}{\omega^p} \right|,$$

$$d_{1,p}(\tilde{V}_1, \tilde{V}_2) \triangleq \sup_{x \in \mathcal{X}} \int \left| \frac{\tilde{V}_1(\omega; x) - \tilde{V}_2(\omega; x)}{\omega^p} \right| \mathrm{d}\omega.$$



Norms: $\|\tilde{V}\|_{\infty,p} = d_{\infty,p}(\tilde{V}, 0)$ and $\|\tilde{V}\|_{1,p} = d_{1,p}(\tilde{V}, 0)$.

Lemma 1. *Bellman operator $\tilde{T}^\pi$ is contraction:*

$$d_{\infty,p}(\tilde{T}^\pi\tilde{V}_1, \tilde{T}^\pi\tilde{V}_2) \leq \gamma^p d_{\infty,p}(\tilde{V}_1, \tilde{V}_2), \qquad d_{1,p}(\tilde{T}^\pi\tilde{V}_1, \tilde{T}^\pi\tilde{V}_2) \leq \gamma^{p-1} d_{1,p}(\tilde{V}_1, \tilde{V}_2).$$

Being a contraction leads to nice properties, such as having a unique fixed point. It also suggests a way to find a CVF.

Characteristic Value Iteration

Q: How can we compute CVF?

A: Since the Bellman operator is a contraction, we can find $\tilde{V}^\pi$ using an iterative procedure similar to Value Iteration.

$$\begin{aligned} &\tilde{V}_1 \leftarrow \tilde{R}, \\ &\tilde{V}_{k+1} \leftarrow \tilde{T}^\pi \tilde{V}_k = \tilde{R}\mathcal{P}^\pi \tilde{V}_k. \qquad (k \geq 1) \end{aligned} \tag{1}$$

CVF converges: $d_{\infty,1}(\tilde{V}_{k+1}, \tilde{V}^\pi) \leq \gamma d_{\infty,1}(\tilde{V}_k, \tilde{V}^\pi) \leq \cdots \leq \gamma^k d_{\infty,1}(\tilde{V}_1, \tilde{V}^\pi) = \gamma^k d_{\infty,1}(\tilde{R}, \tilde{V}^\pi)$.

Performing CVI (1) exactly may not be practical:

- Large state space: $\tilde{V}^\pi$ cannot be represented exactly; we have to use function approximator.
- Learning: We do not have access to $\mathcal{P}^\pi$, but only observed data.

If we can only perform $\tilde{V}_{k+1} \approx \tilde{T}^\pi \tilde{V}_k$, we have **Approximate Characteristic Value Iteration (ACVI)**. Suppose that we have a dataset $\mathcal{D}_n = \{(X_i, R_i, X'_i)\}_{i=1}^n$, with $X_i \sim \mu$, $X'_i \sim \mathcal{P}^\pi(\cdot|X_i)$ and $R_i \sim \mathcal{R}^\pi(\cdot|X_i)$. For any fixed $\tilde{V}$, we can see that

$$\mathbb{E}\left[e^{j\omega R_i}\tilde{V}(\gamma\omega; X'_i)|X = X_i\right] = (\tilde{T}^\pi\tilde{V})(\omega; X_i),$$

Finding a good approximation of $\tilde{T}^\pi\tilde{V}$ given noisy samples is the regression problem.

Empirical Risk Minimization-based solution:

$$\tilde{V}_{k+1} \leftarrow \operatorname*{argmin}_{\tilde{V} \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \int \left| \tilde{V}(\omega; X_i) - e^{j\omega R_i}\tilde{V}_k(\gamma\omega; X'_i) \right|^2 w(\omega)\mathrm{d}\omega.$$

Similar to the usual Fitted Value Iteration procedure:

$$V_{k+1} \leftarrow \operatorname*{argmin}_{V \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \left| V(X_i) - (R_i + \gamma V_k(X'_i)) \right|^2.$$

Error Propagation for Approximate Characteristic Value Iteration

Q: How does the errors in ACVI affect the quality of the outcome estimate $\tilde{V}_K$?

$$\begin{aligned} &\tilde{V}_1 \leftarrow \tilde{R} + \tilde{\varepsilon}_1, \\ &\tilde{V}_{k+1} \leftarrow \tilde{T}^\pi \tilde{V}_k + \tilde{\varepsilon}_{k+1}. \qquad (k \geq 1) \end{aligned} \tag{2}$$

Theorem 2 (Error Propagation for ACVI — Simplified)**.** *Consider the ACVI procedure* (2) *after $K \geq 1$ iterations. Assume that $\tilde{\varepsilon}_k(0; x) = 0$ for all $x \in \mathcal{X}$ and $k = 1, \ldots, K+1$. We then have*

$$d_{\infty,1}(\tilde{V}_{K+1}, \tilde{V}^\pi) \leq \sum_{i=0}^K \gamma^i \|\tilde{\varepsilon}_{K+1-i}\|_{\infty,1} + \gamma^K d_{\infty,1}(\tilde{R}, \tilde{V}^\pi).$$

Q: How is the error in the frequency domain related to the error in the probability distribution functions?

A: Error according to $\|\cdot\|_{\infty,p}$ translates to an error w.r.t. $p+1$-smooth Wasserstein distance.

Let $\mathcal{F}_p(\Omega) = \left\{ f \in \mathcal{C}^p(\Omega) : \|f^{(k)}\|_\infty \leq 1, 0 \leq k \leq p \right\}$. For two probability distributions μ_1, μ_2, the p-smooth Wasserstein distance is defined as

$$\mathcal{W}_{\mathcal{C}_p}(\mu_1, \mu_2) = \sup_{f \in \mathcal{F}_p(\Omega)} \left| \int f(x)\left(\mathrm{d}\mu_1(x) - \mathrm{d}\mu_2(x)\right) \right|.$$

Given $\tilde{V}$, we denote $\bar{V}$ as its corresponding probability distribution function. Let us also define the p-smooth Wasserstein between $\bar{V}_1$ and $\bar{V}_2$:

$$\mathcal{W}_{\mathcal{C}_p}(\bar{V}_1, \bar{V}_2) \triangleq \sup_{x \in \mathcal{X}} \mathcal{W}_{\mathcal{C}_p}(\bar{V}_1(\cdot; x), \bar{V}_2(\cdot; x)).$$

Theorem 3 (Error in PDF — Simplified)**.** *Consider the same setting and assumption as in Theorem 2. Furthermore, assume that the immediate reward distribution $\mathcal{R}^\pi(\cdot|x)$ is R_{max}-bounded. We then have*

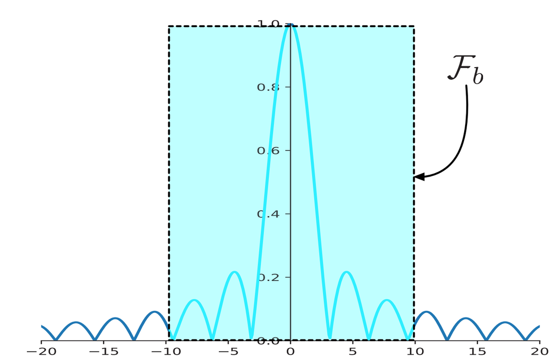
$$\mathcal{W}_{\mathcal{C}_2}(\bar{V}_{K+1}, \bar{V}^\pi) \leq \frac{2\sqrt{2R_{max}}}{\sqrt{\pi}(1-\gamma)^{3/2}} \left[\max_{i=1,\ldots,K+1} \|\tilde{\varepsilon}_i\|_{\infty,1} + 2\gamma^K R_{max} \right].$$

A Study on Function Approximation Error and Covering Numbers

Q: How well can we approximate a CVF given a restrictive function space?

Let $\mathcal{F}_b$ be defined as the b-band-limited CVF, i.e.,

$$\mathcal{F}_b = \left\{ \tilde{V} : \mathbb{R} \times \mathcal{X} \to \mathbb{C}_1 \,:\, \tilde{V}(0; x) = 1, \tilde{V}(\omega; x) = 0 \;\forall |\omega| > b \right\}.$$



The reward distribution $\mathcal{R}^\pi$ is β-smooth if for all $x \in \mathcal{X}$,

$$c_0|\omega|^{-\beta} \leq |\tilde{R}(\omega; x)| \leq c_1|\omega|^{-\beta}.$$

Examples: exponential, uniform, gamma, etc.

Theorem 4. *Consider function space $\mathcal{F}_b$, and assume that $\mathcal{R}$ is a β-smooth distribution. We have*

$$\sup_{\tilde{V}' \in \mathcal{V}} \inf_{\tilde{V} \in \mathcal{F}_b} \left\| \tilde{V} - \tilde{T}^\pi\tilde{V}' \right\|_{\infty,p} \leq \frac{c_1}{b^{p+\beta}}, \qquad \inf_{\tilde{V} \in \mathcal{F}_b} \left\| \tilde{V} - \tilde{R} \right\|_{\infty,p} \leq \frac{c_1}{b^{p+\beta}}.$$

Some Remarks:

- The β-smooth reward distributions can be well-approximated within $\mathcal{F}_b$. Moreover, if we apply $\tilde{T}^\pi$ to a member of $\mathcal{F}_b$, the result can still be well-approximated within $\mathcal{F}_b$.
- The function space $\mathcal{F}_b$ is very large.
- Similar results for much smaller space of band-limited smooth (in $\mathcal{C}^s(\Omega)$ sense) functions.
- Covering number result for the smooth band-limited function space:
  - $\log \mathcal{N}(\varepsilon, \mathcal{F}^s_{b,r}, L_{\infty,p}) \leq cb^{1+\frac{1}{sp}} |\mathcal{X}| \left(\frac{r}{\varepsilon}\right)^{\frac{1}{s}}$
  - $\log \mathcal{N}(\varepsilon, \mathcal{F}^s_{b,r}, d_{\infty,1}) \leq |\mathcal{X}|\; s \log\left(\frac{2erb^{\frac{s-1}{2}}}{\varepsilon}\right)$