Value Function in Frequency Domain and the Characteristic Value Iteration Algorithm

Amir-massoud Farahmand*
Vector Institute & University of Toronto
Toronto, Canada
farahmand@vectorinstitute.ai

Abstract

This paper considers the problem of estimating the distribution of returns in reinforcement learning, i.e., distributional RL problem. It presents a new representational framework to maintain the uncertainty of returns and provides mathematical tools to compute it. We show that instead of representing a probability distribution function of returns, one can represent their characteristic function, the Fourier transform of their distribution. We call the new representation Characteristic Value Function (CVF). The CVF satisfies a Bellman-like equation, and its corresponding Bellman operator is contraction with respect to certain metrics. The contraction property allows us to devise an iterative procedure to compute the CVF, which we call Characteristic Value Iteration (CVI). We analyze CVI and its approximate variant and show how approximation errors affect the quality of the computed CVF.

1 Introduction

The object of focus of the conventional RL is the expected return of following a policy, i.e., the value function [Sutton and Barto, 2019]. The goal is to find a policy that maximizes that expectation over all states, i.e., the optimal policy. This leads to agents that do not consider the distribution of returns in their decision making, but only its first moment. This might be of concern in scenarios where the risk is of paramount importance. Estimating the distribution of the return facilitates designing agents that consider objectives more general than maximizing the expected return, such as various notions of risk [Tamar et al., 2012, Prashanth and Ghavamzadeh, 2013, García and Fernández, 2015, Chow et al., 2018].

The Distributional RL (DistRL) literature [Engel et al., 2005, Morimura et al., 2010b, Bellemare et al., 2017, Barth-Maron et al., 2018, Lyle et al., 2019], on the other hand, moves away from the conventional goal of estimating the expectation of return and attempts to estimate a richer representation of the return, such as the distribution itself [Morimura et al., 2010b,a] or some statistical functional of it [Rowland et al., 2018, Dabney et al., 2018, Rowland et al., 2019]. It is notable that so far the focus of the DistRL literature has mostly been on designing better performing agents according to the expected return, and not any risk-related performance measure, but it is conceivable that those methods can be be used for designing risk-aware agents too.

This paper develops a new framework for maintaining the information available in the distribution of returns. Instead of estimating the distribution function itself, we maintain the Characteristic Function (CF) of the returns. The CF of a random variable (r.v.) is the Fourier transform of its probability distribution function (PDF). Similar to PDF, the CF of a r.v. contains all the information available about the distribution of that r.v., i.e., CF and PDF have a bijection relationship. They are nonetheless

*Homepage: http://academic.sologen.net.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

different representations of the uncertainty of a r.v., hence they allow different types of manipulations and processing. The benefit of a new representation is that it opens up the possibility of designing new algorithms. An example from the field of control theory is that we have both time and frequency domain representations of a dynamical system. Although they are equivalent in many cases, designing a controller in the frequency domain is sometimes easier and may provide better insights. This work brings the frequency-based representation of uncertainty to DistRL.

The estimation procedures based on CF are not novel. Methods based on the Empirical Characteristic Function (ECF) have a long history in the statistics and econometrics literature [Feuerverger and Mureika, 1977, Feuerverger and McDunnough, 1981, Feuerverger, 1990, Knight and Yu, 2002, Yu, 2004]. These methods are considered as alternatives to the maximum likelihood estimation (MLE), because as opposed to MLE, whose computation might be infeasible for some distributions, one can always define and compute the ECF. This paper is inspired from that literature and develops similar tools for RL and approximate dynamic programming.

The main idea of this work is that by transforming the return, which is a r.v., to the frequency domain through the Fourier transform, we can define Characteristic Value Function (CVF), which essentially captures all information about the distribution of the return. A contribution of this work is that we prove that CVF indeed satisfies a Bellman-like equation $\tilde{T}^\pi \tilde{V} = \tilde{V}$ (Section 3). The corresponding Bellman operator, however, is different from the conventional ones or those in the DistRL literature. Instead of having an additive form, it is multiplicative, i.e., $(\tilde{T}^\pi \tilde{V})(\omega; x) \triangleq \tilde{R}(\omega; x) \int \mathcal{P}^\pi(\mathrm{d}y|x)\tilde{V}(\gamma\omega; y)$ with ω being the frequency variable, x being the state variable, and $\tilde{R}$ being the Fourier transform of the immediate reward distribution (we will define these quantities later). We also prove that the new Bellman operator is contraction with respect to (w.r.t.) some specific metrics defined in the frequency domain (Section 3.1). The contraction property suggests that one might find the CVF through an iterative procedure similar to value iteration, which we call the Characteristic Value Iteration (CVI) algorithm (Section 4). This is the algorithmic contribution of this work.

Any procedure that implements CVI, however, may not perform it exactly, for example because we only have data as opposed to the actual transition probability distribution or because the state space is very large and we need to use function approximation. In case we can only approximately perform CVI, which we call Approximate CVI (ACVI), we inevitably have some errors. To understand the effect of using function approximation on these errors better, we consider a class of band-limited (in the frequency domain) functions, and study their function approximation and covering number properties . Another contribution of this work is the analysis of how the errors caused at each iteration of ACVI propagate throughout iterations and affect the quality of the outcome CVF (Section 5). We show that the errors in earlier iterations decay exponentially fast, i.e., the past errors are forgotten quickly. This is the same phenomenon observed in the conventional approximate value iteration. Finally, we show how to convert the error of CVF in the frequency domain to an error in distributions, measured according to the p-smooth Wasserstein distance (Section 6).

2 Distributional Bellman equation

We consider a discounted Markov Decision Process (MDP) $(\mathcal{X}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \gamma)$ [Szepesvári, 2010]. Here $\mathcal{X}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{P} : \mathcal{X} \times \mathcal{A} \to \mathcal{M}(\mathcal{X})$ is the transition probability kernel, $\mathcal{R} : \mathcal{X} \times \mathcal{A} \to \mathcal{M}(\mathbb{R})$ is the immediate reward distribution, and $0 \leq \gamma < 1$ is the discount factor.[2] The (Markov stationary) policy $\pi : \mathcal{X} \to \mathcal{M}(\mathcal{A})$ induces the transition probability kernel $\mathcal{P}^\pi : \mathcal{X} \to \mathcal{M}(\mathcal{X})$ and the immediate reward distribution for the policy $\mathcal{R}^\pi : \mathcal{X} \to \mathcal{M}(\mathbb{R})$.

An MDP together with an initial state distribution $\rho \in \mathcal{M}(\mathcal{X})$ encode the laws governing the temporal evolution of a discrete-time stochastic process controlled by an agent as follows: The controlled process starts at time $t = 0$ with random initial state X_0 drawn from ρ, i.e., $X_0 \sim \rho$. The agent following a policy π chooses action $A_t \in \mathcal{A}$ according to $A_t \sim \pi(\cdot|X_t)$ (stochastic policy) or $A_t = \pi(X_t)$ (deterministic policy). In response, the next state is $X_{t+1} \sim \mathcal{P}(\cdot|X_t, A_t)$ and the agent

[2]Here $\mathcal{M}(\Omega)$ refers to the space of all probability distributions on an appropriately defined σ-algebra of Ω, e.g., the Borel σ-algebra on $\mathbb{R}$. We do not deal with the measure theoretic considerations in this work. Refer to Appendix C of Bertsekas [2013] or Chapter 7 of Bertsekas and Shreve [1978]. We occasionally use $\bar{X}$ to denote the probability distribution μ of the r.v. X.

receives reward $R_t \sim \mathcal{R}(\cdot|X_t, A_t)$. This process repeats. We may occasionally use $R(x,a)$ or $R^\pi(x)$ to denote to the r.v. that is drawn from $\mathcal{R}(\cdot|x,a)$ or $\mathcal{R}^\pi(\cdot|x)$. Also we may use $z = (x,a)$ as a shorthand. When we refer to a r.v. $Z = (X, A)$, this should be interpreted as a r.v. defined with $A \sim \pi(\cdot|X)$, where the policy should be clear from the context.

The return of the agent starting from a state $x \in \mathcal{X}$ and following a policy π is the following random variable:

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

The (conventional) value function V^π is the first moment of this r.v., i.e.,

$$V^\pi(x) = \mathbb{E}\left[G^\pi(X_0)|X_0 = x\right].$$

Likewise, one may define the return $G^\pi(x,a)$ for starting from state x, choosing action a, and following policy π afterwards. The corresponding first moment of $G^\pi(x,a)$ would be the action-value function $Q^\pi(x,a)$.

From $G^\pi(x) = R_0 + \gamma \sum_{i \geq 0} \gamma^i R_{i+1}$, we see that $G^\pi(x)$ is the addition of two r.v. R_0 and $\gamma G^\pi(X_1)$ with $X_1 \sim \mathcal{P}^\pi(\cdot|X_0 = x)$. Therefore, the law (probability distribution) of $G^\pi(x)$ is the same as the law of $R_0 + \gamma G^\pi(X_1)$, i.e.,

$$G^\pi(x) \stackrel{(\mathrm{D})}{=} R_0 + \gamma G^\pi(X_1). \tag{1}$$

Here we use the symbol $\stackrel{(\mathrm{D})}{=}$ to emphasize that we are comparing two probability distributions. This is the Bellman-like distributional equation in the conventional DistRL.

We can also have a similar equation that relates $\bar{G}^\pi$ (the distribution of the r.v. G^π) and $\bar{R}(x) = \mathcal{R}^\pi(\cdot|x)$ (the distribution of the r.v. $R^\pi(x)$) [Rowland et al., 2018]. We use $\bar{G}^\pi(A; x)$ to denote the probability that the return r.v. G^π takes value in the Borel set $A \subset \mathbb{R}$, e.g., $G^\pi(\mathrm{d}g; x)$ is the density of the return at state x, if it exists. Also recall the definition of the pushforward measure: Given a probability distribution $\nu \in \mathcal{M}(\mathbb{R})$ and a measurable function $f : \mathbb{R} \to \mathbb{R}$, the pushforward measure $f_{\#}\nu \in \mathcal{M}(\mathbb{R})$ is defined as $(f_{\#}\nu)(A) = \nu(f^{-1}(A))$ for all Borel sets $A \subset \mathbb{R}$.

The Bellman operator $\bar{T}^\pi : \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathcal{X})$ between distributions is defined as

$$(\bar{T}^\pi \bar{G})(x) \triangleq \int \left[(r,g) \mapsto r + \gamma g\right]_{\#} \bar{G}(\mathrm{d}g; y) \mathcal{R}^\pi(\mathrm{d}r|x) \mathcal{P}^\pi(\mathrm{d}y|x), \qquad \forall x \in \mathcal{X}.$$

With this notation, the distributional Bellman equation is

$$\bar{G}^\pi(x) = (\bar{T}^\pi \bar{G}^\pi)(x), \qquad \forall x \in \mathcal{X}. \tag{2}$$

The distributional Bellman equation represents the intrinsic uncertainty of the return due to the randomness of the dynamics and policy. We may occasionally use $\bar{V}^\pi$ to refer to $\bar{G}^\pi$, to show its close relation to the conventional value function.

3 Characteristic value function

The conventional approach to representing the uncertainty of a r.v. is through its probability distribution function. This is not the only way to characterize a r.v. though. An alternative is to characterize the r.v. through the Fourier transform of its distribution function. This is known as the Characteristic Function (CF) of the random variable [Williams, 1991].

In this section we show that instead of representing the distribution function of the return G^π, we may represents its characteristic function. Interestingly, the CF of return satisfies a Bellman-like equation, which is quite different from the conventional ones (1) (2) that we have encountered so far.

Let us briefly recall the definition of a CF of a random variable. Given a real-valued r.v. X with the probability distribution $\mu \in \mathcal{M}(\mathbb{R})$, its corresponding CF $c_X : \mathbb{R} \to \mathbb{C}$ is the function defined as[3]

$$c_X(\omega) \triangleq \mathbb{E}\left[e^{jX\omega}\right] = \int \exp(jx\omega)\mu(\mathrm{d}x), \qquad \omega \in \mathbb{R} \tag{3}$$

[3]Here X is a generic r.v. and does not refer to the state. The particular r.v. will be clear from the context.

where $j = \sqrt{-1}$ is the imaginary unit. The CF of a probability distribution is closely related to the Fourier transform of its distribution function. If the probability density function is well-defined, CF is its Fourier transform, though CF exists even if the density does not. Several properties of CF are summarized in Appendix A. Thinking in the terms of the spatial-frequency duality common in the Fourier analysis, the probability distribution function is the spatial representation of a r.v. (with the magnitude of the r.v. corresponding to the space dimension), and the CF is its frequency representation.

Consider the recursive relation $G^\pi(x) = R^\pi(x) + \gamma G^\pi(X')$, with $X' \sim \mathcal{P}^\pi(\cdot|x)$, between the return $G^\pi(x)$ (a r.v.) and the random reward $R^\pi(x)$ and the return at the next step $G^\pi(X')$. By the distributional equality of both sides (cf. (1)), we have

$$c_{G^\pi(x)}(\omega) = \mathbb{E}\left[\exp\left(j\omega G^\pi(x)\right)\right] = \mathbb{E}\left[\exp\left(j\omega\left(R^\pi(x) + \gamma G^\pi(X')\right)\right)\right], \qquad \forall \omega \in \mathbb{R}. \tag{4}$$

The right-hand side (RHS) of (4) is

$$\begin{aligned}
\mathbb{E}\left[\exp\left(j\omega\left(R^\pi(x) + \gamma G^\pi(X')\right)\right)\right] &= \mathbb{E}\left[\mathbb{E}\left[\exp\left(j\omega\left(R^\pi(x) + \gamma G^\pi(X')\right)\right) \mid X = x, A\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\exp\left(j\omega R^\pi(x)\right) \mid X = x, A\right]\mathbb{E}\left[\exp\left(j\omega\gamma G^\pi(X')\right) \mid X = x, A\right]\right] \\
&= c_{R^\pi(x)}(\omega)\,\mathbb{E}\left[\mathbb{E}\left[\exp\left(j\omega\gamma G^\pi(X')\right) \mid X = x, A\right]\right] \\
&= c_{R^\pi(x)}(\omega)\,\mathbb{E}\left[\exp\left(j\omega\gamma G^\pi(X')\right) \mid X = x\right],
\end{aligned} \tag{5}$$

where A is a r.v. drawn from $\pi(\cdot|x)$. Here we benefitted from the fact that the r.v. $R^\pi(x)$ and $G^\pi(X')$ are conditionally independent given $X = x$ and A.

Let us consider the CF of $G^\pi(X')$ conditioned on $X = x$:

$$\begin{aligned}
\mathbb{E}\left[\exp\left(j\omega G^\pi(X')\right) \mid X = x\right] &= \mathbb{E}\left[\mathbb{E}\left[\exp\left(j\omega G^\pi(X')\right) \mid X'\right] \mid X = x\right] \\
&= \int \mathcal{P}^\pi(\mathrm{d}x'|x)\mathbb{E}\left[\exp\left(j\omega G^\pi(x')\right)\right] \\
&= \mathbb{E}\left[c_{G^\pi(X')}(\omega) \mid X = x\right],
\end{aligned} \tag{6}$$

where we conditioned the inner expectation on the next-state X' (so its randomness comes from the return from that point onward), and used the definition of CF.

Plugging (6) in (5) gives the RHS of (4). So we get

$$\begin{aligned}
c_{G^\pi(x)}(\omega) &= c_{R^\pi(x)}(\omega)\,\mathbb{E}\left[\exp\left(j\omega\gamma G^\pi(X')\right) \mid X = x\right] \\
&= c_{R^\pi(x)}(\omega)\mathbb{E}\left[c_{\gamma G^\pi(X')}(\omega) \mid X = x\right] \\
&= c_{R^\pi(x)}(\omega)\mathbb{E}\left[c_{G^\pi(X')}(\gamma\omega) \mid X = x\right] = c_{R^\pi(x)}(\omega)\int \mathcal{P}^\pi(\mathrm{d}y|x)c_{G^\pi(y)}(\gamma\omega),
\end{aligned} \tag{7}$$

where the penultimate equality is because of the scaling property of CF (Lemma 8 in Appendix A).

We 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 $\tilde{V}^\pi(\omega; x)$ for all $x \in \mathcal{X}$ and $\omega \in \mathbb{R}$. Here the symbol $\tilde{\cdot}$ is used to remind us that we are referring to a CF of a random variable. With these notations, we can write (7) in more compact form of

$$\tilde{V}^\pi(\omega; x) = \tilde{R}(\omega; x)\int \mathcal{P}^\pi(\mathrm{d}y|x)\tilde{V}^\pi(\gamma\omega; y). \tag{8}$$

This is the Bellman-like equation between the CF of return and the reward. The function $\tilde{V}^\pi : \mathbb{R} \times \mathcal{X} \to \mathbb{C}_1$ (where $\mathbb{C}_1$ is the area within the unit circle in the complex plane, i.e., $\mathbb{C}_1 = \{\, z \in \mathbb{C} : |z| \leq 1 \,\}$) is the CF of the $G^\pi(x)$ for all $x \in \mathcal{X}$. We call $\tilde{V}^\pi$ the Characteristic Value Function (CVF).

We also define the Bellman operator between the CF functions:

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

With this notation, the Bellman equation can be written more compactly as

$$\tilde{V}^\pi = \tilde{T}^\pi \tilde{V}^\pi.$$

It is worth mentioning that for any fixed $x \in \mathcal{X}$, $\omega \mapsto \tilde{V}^\pi(\omega; x)$ is a CF. A CF is continuous function of ω and its magnitude is bounded by 1 (Lemma 8 in Appendix A).

3.1 Bellman operator is contraction

We show that the Bellman operator $\tilde{T}^\pi$ is a contraction w.r.t. certain metrics, to be specified. This allows us to devise a value iteration-like procedure that converges to the CVF $\tilde{V}^\pi$ of a policy π.

We first define some distance metrics between CFs. Given two CF $c_1, c_2 : \mathbb{R} \to \mathbb{C}$, and $p \geq 1$, we define

$$d_{\infty,p}(c_1, c_2) \triangleq \sup_{\omega \in \mathbb{R}} \left| \frac{c_1(\omega) - c_2(\omega)}{\omega^p} \right|, \qquad d_{1,p}(c_1, c_2) \triangleq \int \left| \frac{c_1(\omega) - c_2(\omega)}{\omega^p} \right| \mathrm{d}\omega. \tag{9}$$

Here we use the convention that $\frac{0}{0} = 0$.[4]

We also define similar metrics for functions such as $\tilde{R}$ and $\tilde{V}^\pi$. Given $\tilde{V}_1, \tilde{V}_2 : \mathbb{R} \times \mathcal{X} \to \mathbb{R}$, we define

$$d_{\infty,p}(\tilde{V}_1, \tilde{V}_2) \triangleq \sup_{x \in \mathcal{X}} \sup_{\omega \in \mathbb{R}} \left| \frac{\tilde{V}_1(\omega; x) - \tilde{V}_2(\omega; x)}{\omega^p} \right|, \; d_{1,p}(\tilde{V}_1, \tilde{V}_2) \triangleq \sup_{x \in \mathcal{X}} \int \left| \frac{\tilde{V}_1(\omega; x) - \tilde{V}_2(\omega; x)}{\omega^p} \right| \mathrm{d}\omega. \tag{10}$$

There are similar to the distances for comparing two CFs, with the difference that we take the supremum over all states $x \in \mathcal{X}$. To be more precise about how the distances are calculated (e.g., sup over $\mathcal{X}$, etc.), we could use $d_{\mathcal{X}(\infty),\omega(\infty,p)}(\tilde{V}_1, \tilde{V}_2)$ instead of $d_{\infty,p}(\tilde{V}_1, \tilde{V}_2)$. To simplify the notations, however, we use the overloaded symbols $d_{\infty,p}$ and $d_{1,p}$ instead.

Based on these distances, we define the following norms for a function $\tilde{V} : \mathbb{R} \times \mathcal{X} \to \mathbb{R}$

$$\left\| \tilde{V} \right\|_{\infty,p} = d_{\infty,p}(\tilde{V}, 0), \qquad \left\| \tilde{V} \right\|_{1,p} = d_{1,p}(\tilde{V}, 0),$$

where 0 is a constant function $(\omega; x) \mapsto 0$. We sometimes refer to the supremum w.r.t. $x \in \mathcal{X}$ of $\tilde{V}$ by $\|\tilde{V}(\omega; \cdot)\|_\infty = \sup_{x \in \mathcal{X}} |\tilde{V}(\omega; x)|$. This should not be confused with $\|\tilde{V}\|_{\infty,p}$, whose supremum is over both ω and x, and the ω variable is weighted by w^{-p}.

Several properties of $d_{\infty,p}$ and $d_{1,p}$ are presented in Appendix B. Briefly, we show that $d_{1,p}$ and $d_{\infty,p}$ are metrics. We also show that the space of VCFs $\mathcal{V} = \{\tilde{V} : \mathbb{R} \times \mathcal{X} \to \mathbb{C}_1 : \tilde{V}(0; x) = 1\}$, which is a superset of the space of all feasible VCFs, endowed with $d_{\infty,p}$ is complete.

The following result shows that the Bellman operator for VCF is a contraction operator w.r.t. $d_{1,p}$ and $d_{\infty,p}$. This is the main result of this section.

Lemma 1. *Let $0 < \gamma < 1$. The operator $\tilde{T}^\pi$ is a γ^p-contraction in $d_{\infty,p}$ (for $p > 0$) and γ^{p-1}-contraction in $d_{1,p}$ (for $p > 1$). That is, for any $\tilde{V}_1, \tilde{V}_2 : \mathbb{R} \times \mathcal{X} \to \mathbb{C}$ with $d_{\infty,p}(\tilde{V}_1, \tilde{V}_2) < \infty$ or $d_{1,p}(\tilde{V}_1, \tilde{V}_2) < \infty$, we have*

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

Proof. Consider any $\tilde{V}_1, \tilde{V}_2 : \mathbb{R} \times \mathcal{X} \to \mathbb{C}$. From the definition of the Bellman operator $\tilde{T}^\pi$, we have that for any $x \in \mathcal{X}$ and $\omega \in \mathbb{R}$

$$(\tilde{T}^\pi \tilde{V}_1)(\omega; x) - (\tilde{T}^\pi \tilde{V}_2)(\omega; x) = \tilde{R}(\omega; x) \int \mathcal{P}^\pi(\mathrm{d}y|x) \left(\tilde{V}_1(\gamma\omega; y) - \tilde{V}_2(\gamma\omega; y) \right).$$

[4] The metric $d_{\infty,p}$ has been studied under the name of Fourier-based metric Carrillo and Toscani [2007], and is called Toscani distance by Villani [2008].

We first prove the result for $d_{\infty,p}$. For any $x \in \mathcal{X}$, we have

$$
\begin{aligned}
d_{\infty,p}\left((\tilde{T}^\pi \tilde{V}_1)(\cdot;x), (\tilde{T}^\pi \tilde{V}_2)(\cdot;x)\right) &= \sup_\omega \left| \frac{(\tilde{T}^\pi \tilde{V}_1)(\omega;x) - (\tilde{T}^\pi \tilde{V}_2)(\omega;x)}{\omega^p} \right| \\
&\leq \sup_\omega \left[\underbrace{|\tilde{R}(\omega;x)|}_{\leq 1} \int \mathcal{P}^\pi(\mathrm{d}y|x) \frac{\left|\tilde{V}_1(\gamma\omega;y) - \tilde{V}_2(\gamma\omega;y)\right|}{|\omega|^p} \right] \\
&\leq \int \mathcal{P}^\pi(\mathrm{d}y|x) \sup_\omega \left| \frac{\tilde{V}_1(\gamma\omega;y) - \tilde{V}_2(\gamma\omega;y)}{|\omega|^p} \right|. \qquad (11)
\end{aligned}
$$

We benefited from $|\tilde{R}(\omega;x)| \leq 1$ to get rid of the term depending of $\tilde{R}$, see Lemma 8 in Appendix A. Denote $\nu = \gamma\omega$, and write

$$
\begin{aligned}
\sup_{\omega\in\mathbb{R}} \left| \frac{\tilde{V}_1(\gamma\omega;y) - \tilde{V}_2(\gamma\omega;y)}{|\omega|^p} \right| = \sup_{\nu\in\mathbb{R}} \left| \frac{\tilde{V}_1(\nu;y) - \tilde{V}_2(\nu;y)}{|(\frac{\nu}{\gamma})|^p} \right| &= |\gamma|^p \sup_{\nu\in\mathbb{R}} \left| \frac{\tilde{V}_1(\nu;y) - \tilde{V}_2(\nu;y)}{|\nu|^p} \right| \\
&= |\gamma|^p d_{\infty,p}\left(\tilde{V}_1(\cdot;y), \tilde{V}_2(\cdot;y)\right). \qquad (12)
\end{aligned}
$$

Plugging this result in the RHS of (11), and taking the supremum over y, we get that

$$
\begin{aligned}
d_{\infty,p}\left((\tilde{T}^\pi \tilde{V}_1)(\cdot;x), (\tilde{T}^\pi \tilde{V}_2)(\cdot;x)\right) &\leq \gamma^p \int \mathcal{P}^\pi(\mathrm{d}y|x) \sup_y d_{\infty,p}\left(\tilde{V}_1(\cdot;y), \tilde{V}_2(\cdot;y)\right) \\
&= \gamma^p d_{\infty,p}\left(\tilde{V}_1, \tilde{V}_2\right) \underbrace{\int \mathcal{P}^\pi(\mathrm{d}y|x)}_{=1} = \gamma^p d_{\infty,p}\left(\tilde{V}_1, \tilde{V}_2\right).
\end{aligned}
$$

Since this holds for any $x \in \mathcal{X}$, we get the desired result by taking the supremum over $x \in \mathcal{X}$ in the left-hand side (LHS).

The proof of the result w.r.t. $d_{1,p}$ is similar. For any $x \in \mathcal{X}$, we have

$$
\begin{aligned}
d_{1,p}\left((\tilde{T}^\pi \tilde{V}_1)(\cdot;x), (\tilde{T}^\pi \tilde{V}_2)(\cdot;x)\right) &= \int \left| \frac{\tilde{R}(\omega;x) \int \mathcal{P}^\pi(\mathrm{d}y|x) \left(\tilde{V}_1(\gamma\omega;y) - \tilde{V}_2(\gamma\omega;y)\right)}{\omega^p} \right| \mathrm{d}\omega \\
&\leq \int \underbrace{\left|\tilde{R}(\omega;x)\right|}_{\leq 1} \left[\int \mathcal{P}^\pi(\mathrm{d}y|x) \left| \frac{\tilde{V}_1(\gamma\omega;y) - \tilde{V}_2(\gamma\omega;y)}{\omega^p} \right| \right] \mathrm{d}\omega \\
&\leq \int \mathcal{P}^\pi(\mathrm{d}y|x) \int \left| \frac{\tilde{V}_1(\gamma\omega;y) - \tilde{V}_2(\gamma\omega;y)}{\omega^p} \right| \mathrm{d}\omega \\
&= \int \mathcal{P}^\pi(\mathrm{d}y|x) \int \left| \frac{\tilde{V}_1(\nu;y) - \tilde{V}_2(\nu;y)}{\left(\frac{\nu}{\gamma}\right)^p} \right| \mathrm{d}\nu(\frac{1}{\gamma}) \\
&= \gamma^{p-1} \int \mathcal{P}^\pi(\mathrm{d}y|x) \int \left| \frac{\tilde{V}_1(\nu;y) - \tilde{V}_2(\nu;y)}{\nu^p} \right| \mathrm{d}\nu \\
&\leq \gamma^{p-1} \int \mathcal{P}^\pi(\mathrm{d}y|x) \sup_{y\in\mathcal{X}} \int \left| \frac{\tilde{V}_1(\nu;y) - \tilde{V}_2(\nu;y)}{\nu^p} \right| \mathrm{d}\nu \\
&= \gamma^{p-1} d_{1,p}\left(\tilde{V}_1, \tilde{V}_2\right) \underbrace{\int \mathcal{P}^\pi(\mathrm{d}y|x)}_{=1} = \gamma^{p-1} d_{1,p}\left(\tilde{V}_1, \tilde{V}_2\right).
\end{aligned}
\qquad (13)
$$

We used the change of variable $\nu = \gamma\omega$, which entails that $\mathrm{d}\omega = \frac{1}{\gamma}\mathrm{d}\nu$. $\square$

For the contraction to be non-trivial, and avoid having a trivial inequality such as $\infty \leq \gamma^p \infty$, we require the boundedness of $d_{\infty,p}(\tilde{V}_1, \tilde{V}_2)$ or $d_{1,p}(\tilde{V}_1, \tilde{V}_2)$. This is a condition that should be verified, and as we shall soon see holds under certain conditions.

Remark 1. The conventional Bellman operator is a γ-contraction w.r.t. the supremum norm. Following this commonly used norm, one could similarly define a supremum-based norm for a VCF $\tilde{V}$ as

$$\left\| \tilde{V} \right\|_\infty = \sup_{x \in \mathcal{X}} \sup_{\omega \in \mathbb{R}} \tilde{V}(\omega; x).$$

The Bellman operator $\tilde{T}^\pi$, however, is not a contraction w.r.t. this norm. To see this, consider a simple MDP with $\mathcal{P}^\pi = \mathbf{I}$ (each state returns to itself) and $\tilde{R}(\omega; x) = 1$, which corresponds to the choice of $R^\pi(\mathrm{d}y|x) = \delta(y - x)$, a Dirac's delta function. Let $\gamma > 0$. For any $\tilde{V}_1, \tilde{V}_2$, we have

$$\begin{aligned} \left\| \tilde{T}^\pi \tilde{V}_1 - \tilde{T}^\pi \tilde{V}_2 \right\|_\infty &= \sup_{x \in \mathcal{X}} \sup_{\omega \in \mathbb{R}} \left| \tilde{R}(\omega; x) \left(\tilde{V}_1(\gamma\omega; x) - \tilde{V}_2(\gamma\omega; x) \right) \right| \\ &= \sup_{x \in \mathcal{X}} \sup_{\omega \in \mathbb{R}} \left| \tilde{V}_1(\gamma\omega; x) - \tilde{V}_2(\gamma\omega; x) \right| \\ &= \sup_{x \in \mathcal{X}} \sup_{\nu \in \mathbb{R}} \left| \tilde{V}_1(\nu; x) - \tilde{V}_2(\nu; x) \right| = \left\| \tilde{V}_1 - \tilde{V}_2 \right\|_\infty . \end{aligned}$$

Therefore, the Bellman operator $\tilde{T}^\pi$ is a non-expansion, but not a contraction, w.r.t. $\|\cdot\|_\infty$. Having the ω term in the denominator of $d_{\infty,p}$ and $d_{1,p}$ is important to get a contraction.

The importance of showing that the Bellman operator for VCF is a contraction is that we can then apply the Banach fixed point theorem (e.g., Theorem 3.2 of Hunter and Nachtergaele [2001]) to show the uniqueness of the fixed point $\tilde{V}^\pi$ (we also require the completeness of the space, which is shown for $d_{\infty,p}$). Moreover, it suggests that we can find the fixed point by iterative application of the operator. This is the path we pursue in the next section.

4 Characteristic value iteration

The contraction property of the Bellman operator $\tilde{T}^\pi$ (Lemma 1) suggests that we can find $\tilde{V}^\pi$ by an iterative procedure, similar to the conventional value iteration. The procedure is

$$\begin{aligned} &\tilde{V}_1 \leftarrow \tilde{R}, \\ &\tilde{V}_{k+1} \leftarrow \tilde{T}^\pi \tilde{V}_k = \tilde{R} \mathcal{P}^\pi \tilde{V}_k. \qquad (k \geq 1) \end{aligned} \tag{14}$$

We call this procedure Characteristic Value Iteration (CVI).

CVI converges under certain conditions. To see this, notice that $\tilde{V}^\pi = \tilde{T}^\pi \tilde{V}^\pi$, so for $p \geq 1$ by Lemma 1 we have

$$d_{\infty,p}(\tilde{T}^\pi \tilde{V}_k, \tilde{V}^\pi) = d_{\infty,p}(\tilde{T}^\pi \tilde{V}_k, \tilde{T}^\pi \tilde{V}^\pi) \leq \gamma^p d_{\infty,p}(\tilde{V}_k, \tilde{V}^\pi),$$

under the condition that $d_{\infty,p}(\tilde{V}_k, \tilde{V}^\pi) < \infty$. Similarly, we have $d_{1,p}(\tilde{T}^\pi \tilde{V}_k, \tilde{V}^\pi) \leq \gamma^{p-1} d_{1,p}(\tilde{V}_k, \tilde{V}^\pi)$ (for $p > 1$). By the iterative application of this upper bound, assuming that $d_{\infty,p}(\tilde{R}, \tilde{V}^\pi) < \infty$, we get that

$$d_{\infty,p}(\tilde{V}_{k+1}, \tilde{V}^\pi) \leq \gamma^p d_{\infty,p}(\tilde{V}_k, \tilde{V}^\pi) \leq \cdots \leq (\gamma^p)^k d_{\infty,p}(\tilde{V}_1, \tilde{V}^\pi) = (\gamma^p)^k d_{\infty,p}(\tilde{R}, \tilde{V}^\pi). \tag{15}$$

Likewise, assuming that $d_{1,p}(\tilde{R}, \tilde{V}^\pi) < \infty$, we obtain

$$d_{1,p}(\tilde{V}_{k+1}, \tilde{V}^\pi) \leq (\gamma^{p-1})^k d_{1,p}(\tilde{R}, \tilde{V}^\pi). \tag{16}$$

As long as $d_{\infty,p}(\tilde{R}, \tilde{V}^\pi)$ (or $d_{1,p}(\tilde{R}, \tilde{V}^\pi)$) is finite for some $p \geq 1$ ($p > 1$), CVI converges geometrically fast. Lemma 11 in Appendix B specifies the condition when the $d_{\infty,p}$ distance of two CF would be finite. For $p = 1$, it is sufficient that the immediate reward $R^\pi(x) \sim \mathcal{R}(\cdot; x)$ and the return $G^\pi(\cdot; x)$ be integrable, i.e., $\mathbb{E}[|R^\pi(x)|], \mathbb{E}[|G^\pi(\cdot; x)|] < \infty$ for all states $x \in \mathcal{X}$. Since we deal with discounted MDP, the integrability of $R^\pi(x)$ (uniformly over $\mathcal{X}$) entails the integrability of $G^\pi(\cdot; x)$. Therefore under very mild conditions, CVI is convergent w.r.t. $d_{\infty,1}$.

For integer valued $p \geq 2$, the condition becomes more restrictive. The first requirement is that $\mathbb{E}\left[|R^\pi(x)|^p\right]$ and $\mathbb{E}\left[|G^\pi(\cdot;x)|^p\right]$ are finite. This is not restrictive, and holds for many problems. The restrictive condition is that the first $k = 1, \dots, p-1$ moments of the reward and the return should match, i.e., $\mathbb{E}\left[R^\pi(x)^k\right] = \mathbb{E}\left[G^\pi(x)^k\right]$ for all $x \in \mathcal{X}$. This does not seem realistic, perhaps except for $p = 2$ when problems with zero expected immediate reward for all states but with varying variance are imaginable.

One can show that the fixed point of $\tilde{T}^\pi$ is unique.

Proposition 2. *Consider an MDP with a discount factor $0 \leq \gamma < 1$. Consider $\mathcal{V}_B \triangleq \left\{ \tilde{V} : \tilde{V} \in \mathcal{V}, d_{\infty,1}(\tilde{V}, \tilde{V}^\pi) \leq B \right\}$ for a finite $B > 0$. The Bellman operator admits a unique fixed point in $\mathcal{V}_B$, which is $\tilde{V}^\pi$. Furthermore, the CVI procedure (14) starting from $\tilde{V}_1 \in \mathcal{V}_B$ generates a sequence $(\tilde{V}_k) \subset \mathcal{V}_B$ that converges to the fixed point. If we assume that the first absolute moment of the reward distribution is uniformly finite (i.e., $\bar{r}_{max} \triangleq \sup_{x \in \mathcal{X}} \mathbb{E}\left[|R^\pi(x)|\right] < \infty$), we may choose $\tilde{V}_1 = \tilde{R}$ and set $B = \frac{2-\gamma}{1-\gamma}\bar{r}_{max}$.*

Proof. Proposition 10 in Appendix B shows that the metric space $(\mathcal{V}, d_{\infty,1})$ is complete. The subset $\mathcal{V}_B$ is a closed ball in $\mathcal{V}$, hence $(\mathcal{V}_{\tilde{R}}, d_{\infty,1})$ is complete.

For any $\tilde{V}_1, \tilde{V}_2 \in \mathcal{V}_B$, we have that $d_{\infty,1}(\tilde{V}_1, \tilde{V}_2) \leq d_{\infty,1}(\tilde{V}_1, \tilde{V}^\pi) + d_{\infty,1}(\tilde{V}^\pi, \tilde{V}_2) \leq 2B < \infty$. Therefore, the finiteness condition of the upper bound in Proposition 1 is satisfied, and the Bellman operator $\tilde{T}^\pi$ is a γ-contraction within $\mathcal{V}_B$.

Moreover, the application of the Bellman operator on any $\tilde{V} \in \mathcal{V}_B$ leaves it within $\mathcal{V}_B$. To see this, notice that $d_{\infty,1}(\tilde{T}^\pi \tilde{V}, \tilde{V}^\pi) = d_{\infty,1}(\tilde{T}^\pi \tilde{V}, \tilde{T}^\pi \tilde{V}^\pi) \leq \gamma d_{\infty,1}(\tilde{V}, \tilde{V}^\pi) \leq \gamma B$, which entails that $\tilde{T}^\pi \tilde{V} \in \mathcal{V}_B$. By induction, the sequence generated by the repeated application of the Bellman operator remains within $\mathcal{V}_B$.

Given the contraction property of the Bellman operator and the completeness of the space, the Banach fixed-point theorem shows that $\tilde{T}^\pi$ admits a unique fixed point within $\mathcal{V}_B$, and the fixed point is the limit of a CVI procedure starting from any $\tilde{V}_1 \in \mathcal{V}_B$.

To show that the CVI procedure with $\tilde{V}_1 = \tilde{R}$ converges to $\tilde{V}^\pi$, we should verify that $\tilde{R}$ is within $\mathcal{V}_B$ with the choice of $B = \frac{2-\gamma}{1-\gamma}\bar{r}_{\max}$. Lemma 11 in Appendix B shows that $d_{\infty,1}(\tilde{R}, \tilde{V}^\pi) \leq \sup_{x \in \mathcal{X}} \{\mathbb{E}\left[|R^\pi(x)|\right] + \mathbb{E}\left[|G^\pi(x)|\right]\} \leq \bar{r}_{\max} + \frac{1}{1-\gamma}\bar{r}_{\max} = B$. $\square$

4.1 Approximate characteristic value iteration

Performing CVI (14) exactly may not be practical, for at least two reasons. First, for problems with large state space, we cannot represent $\tilde{V}^\pi$ exactly and we need to rely on function approximation. Second, for learning scenario where we do not have access to the model $\mathcal{P}^\pi$, but only observe data from interacting with the environment, we cannot apply the Bellman operator $\tilde{T}^\pi$ exactly either.

We can extend CVI to Approximate CVI (ACVI) similar to how exact VI can be extended to Approximate Value Iteration, also known as Fitted Value Iteration or Fitted Q-Iteration. Various variants of AVI have been empirically and theoretically studied in the literature [Ernst et al., 2005, Munos and Szepesvári, 2008, Farahmand et al., 2009, Silver et al., 2016, Tosatto et al., 2017, Chen and Jiang, 2019]. We would like to build the same general framework for CVF and CVI.

Suppose that for whatever reason we perform each iteration of CVI only approximately, that is, $\tilde{V}_{k+1} \approx \tilde{T}^\pi \tilde{V}_k$. The resulting procedure can be described as

$$\begin{aligned} &\tilde{V}_1 \leftarrow \tilde{R} + \tilde{\varepsilon}_1, \\ &\tilde{V}_{k+1} \leftarrow \tilde{T}^\pi \tilde{V}_k + \tilde{\varepsilon}_{k+1}. \qquad (k \geq 1) \end{aligned} \tag{17}$$

Here $\tilde{\varepsilon}_k : \mathbb{R} \times \mathcal{X} \to \mathbb{C}$ is the error in the frequency-state space. Recall that the value of a valid CF at frequency $\omega = 0$ is equal to one, i.e., $c(0) = 1$. To ensure that $\tilde{V}_k(\cdot; x)$ is a CF for all $x \in \mathcal{X}$, we must have $\tilde{V}_k(0; x) = 1$. This is satisfied if we require that $\tilde{\varepsilon}_k(0; x) = 0$ for all $k = 1, 2, \dots$ and $x \in \mathcal{X}$. We can interpret this requirement by noticing that the condition $c(0) = 1$ is simply a requirement that $c(0) = \mathbb{E}\left[e^{jX0}\right] = \mathbb{E}\left[1\right] = \int \mu(\mathrm{d}x)$ be equal to 1. So we are essentially requiring that we do not lose or add probability mass at each iteration of ACVI.

Performing ACVI can be quite similar to the conventional AVI. Suppose that we are given 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 $R_i \sim \mathcal{R}^\pi(\cdot|X_i)$. Given this dataset and a CVF $\tilde{V}$, we define the empirical Bellman operator as the following mapping:

$$(\hat{\tilde{T}}^\pi \tilde{V})(\omega; X_i) \triangleq e^{j\omega R_i} \tilde{V}(\gamma\omega; X'_i), \qquad \forall \omega \in \mathbb{R}, \forall i = 1, \dots, n.$$

For any fixed function $\tilde{V}$ and at any fixed state X_i, with a r.v. $A_i \sim \pi(\cdot|X_i)$, we have

$$\begin{aligned}
\mathbb{E}\left[(\hat{\tilde{T}}^\pi \tilde{V})(\omega; X) \mid X = X_i\right] &= \mathbb{E}\left[e^{j\omega R_i}\tilde{V}(\gamma\omega; X'_i) \mid X = X_i\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[e^{j\omega R_i}\tilde{V}(\gamma\omega; X'_i) \mid X = X_i, A_i\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[e^{j\omega R_i} \mid X = X_i, A_i\right]\mathbb{E}\left[\tilde{V}(\gamma\omega; X'_i) \mid X = X_i, A_i\right]\right] \\
&= \mathbb{E}\left[e^{j\omega R_i} \mid X = X_i\right]\mathbb{E}\left[\tilde{V}(\gamma\omega; X'_i) \mid X = X_i\right] \\
&= \tilde{R}(\omega; X_i) \int \mathcal{P}^\pi(\mathrm{d}y|X_i)\tilde{V}(\gamma\omega; y) = (\tilde{T}^\pi \tilde{V})(\omega; X_i).
\end{aligned}$$

This shows that the random process $(\hat{\tilde{T}}^\pi \tilde{V})(\omega; X_i)$ is an unbiased estimate of $(\tilde{T}^\pi \tilde{V})(\omega; X_i)$. In other words, $(\tilde{T}^\pi \tilde{V})(\omega; X_i)$ is the conditional mean of $(\hat{\tilde{T}}^\pi \tilde{V})(\omega; X_i)$. Finding the conditional mean of a r.v. is the regression problem (i.e., estimating $m(x) = \mathbb{E}\left[Y|X = x\right]$ by $\hat{m}(x)$ using a dataset of $\{(X_i, Y_i)\}_{i=1}^n$), which has been extensively studied in the statistics and machine learning literature [Györfi et al., 2002, Wasserman, 2007, Hastie et al., 2009, Goodfellow et al., 2016]. A powerful estimator that generalizes well across states and ω allows us to approximately perform one step of ACVI.

One approach to finding a regression estimator is to solve an empirical risk minimization problem:

$$\tilde{V}_{k+1} \leftarrow \operatorname*{argmin}_{\tilde{V} \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^n \int \left|\tilde{V}(\omega; X_i) - e^{j\omega R_i}\tilde{V}_k(\gamma\omega; X'_i)\right|^2 w(\omega)\mathrm{d}\omega, \tag{18}$$

where $\mathcal{F} \subset \mathcal{V}$ is a space of functions from $\mathbb{R} \times \mathcal{X}$ to $\mathbb{C}_1$, which can be represented by various types of function approximators (including decision trees, kernel-based ones, and neural networks), and $w : \mathbb{R} \mapsto \mathbb{R}$ is a weighting function that indicates the importance of different frequencies ω. This is similar to the usual Fitted Value Iteration procedure [Ernst et al., 2005, Munos and Szepesvári, 2008, Farahmand et al., 2009, Silver et al., 2016, Tosatto et al., 2017, Chen and Jiang, 2019], which solves

$$V_{k+1} \leftarrow \operatorname*{argmin}_{V \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^n \left|V(X_i) - (R_i + \gamma V_k(X'_i))\right|^2, \tag{19}$$

with appropriately chosen function space $\mathcal{F}$ (and similar for Fitted Q Iteration and the action-value function Q). One clear difference between (18) and (19) is that we have an integral over the frequency domain in the former. This one-dimensional integral can be numerically integrated, for example, by discretizing the low-frequency domain $[-b, +b]$ (with $b > 0$) with resolution ε_{int}. This incurs some controlled numerical error that is a function of ε_{int}. For some function approximators, such as a decision tree, one might be able to calculate the integral more efficiently by benefitting from the constancy of values within a leaf.

The quality of approximating $\tilde{T}^\pi \tilde{V}_k$ by $\tilde{V}_{k+1}$ determines the error $\tilde{\varepsilon}_k$. The error depends on the regression method being used, as well as the number of data points available, capacity and expressibility of the function space $\mathcal{F}$, etc. We do not analyze this regression problem in this paper. We are nevertheless interested in knowing whether one can hope to have a small error with a reasonably selected $\mathcal{F}$. Two relevant questions are whether one can approximate $\tilde{T}^\pi \tilde{V}_k$ within $\mathcal{F}$ well enough (function approximation error), and whether $\mathcal{F}$ has enough regularity to allow reasonable convergence rate for the estimation error. Appendices C and D study these questions in detail. We only briefly mention that if the reward distribution is smooth in a certain sense, a band-limited function class

$\mathcal{F}_b = \{\tilde{V} : \mathbb{R} \times \mathcal{X} \to \mathbb{C}_1 : \tilde{V}(0; x) = 1, \tilde{V}(\omega; x) = 0 \ \forall |\omega| > b\}$ provides an approximation error that goes to zero as the bandwidth b increases. More specifically, the $d_{\infty,1}$ distance-based norm of the approximation error behaves like $O(b^{-\frac{1}{1+\beta}})$ with β being the smoothness parameter. Furthermore, if the first s absolute moments of the reward distribution are finite, the CVF $\tilde{V}(\cdot; x)$ belongs to the smoothness class $C^s([-b, b]) \cap \mathcal{F}_b$ (Theorems 12 and 16). This leads to a well-behaving covering number, which can be used to obtain a convergence rate for the estimation error (Theorem 19). A side benefit of working with a band-limited function space is that the integral in (18) can be converted to a definite integral, which is easier to integrate numerically.

Next we analyze how these errors, however generated, affect the quality of the outcome $\tilde{V}_K$ after performing K steps of ACVI.

5 Error propagation analysis

We analyze how the errors in the ACVI procedure (17) propagate throughout the iterations and affect the quality of the outcome CVF $\tilde{V}_K$, where K is the number of times the iteration is performed.

Let us first define some new notations. For a frequency-state function $\tilde{V}$ and an integer $k \geq 0$, we use $\tilde{V}^{[k]}$ to denote the same function with the difference that its frequency domain variable ω is γ^k-scaled, that is,

$$\tilde{V}^{[k]}(\omega; x) \triangleq \tilde{V}(\gamma^k \omega; x), \qquad \forall \omega \in \mathbb{R}, \forall x \in \mathcal{X}. \tag{20}$$

Notice that $\tilde{V}^{[0]}$ is the same as $\tilde{V}$. The definition of $\tilde{R}^{[k]}$ is the same.

We define a new operator $\tilde{\mathcal{U}}^{[k]} = \tilde{R}^{[k]} \mathcal{P}^\pi$, which means that for $\tilde{V}$, we have

$$\left(\tilde{\mathcal{U}}^{[k]} \tilde{V}\right)(\omega; x) = \tilde{R}^{[k]}(\omega; x) \int \mathcal{P}^\pi(\mathrm{d}y|x) \tilde{V}(\omega; y).$$

With this notation, we have

$$\begin{aligned}
(\tilde{T}^\pi \tilde{V})(\omega; x) &= \tilde{R}(\omega; x) \int \mathcal{P}^\pi(\mathrm{d}y|x) \tilde{V}(\gamma\omega; y) \\
&= \tilde{R}(\omega; x) \int \mathcal{P}^\pi(\mathrm{d}y|x) \tilde{V}^{[1]}(\omega; y) = \left(\tilde{\mathcal{U}}^{[0]} \tilde{V}^{[1]}\right)(\omega; x).
\end{aligned}$$

The presence of the $\tilde{V}^{[1]}$ term in the RHS indicates that each application of the Bellman operator $\tilde{T}^\pi$ expands the frequency domain of the function to which it is applied by a factor of γ. Our choice of notation (20) makes this more transparent, and allows us to keep track of how much the frequency domain is expanded.

We define the operator $\mathcal{L}$, which is simply the multiplication of several $\tilde{\mathcal{U}}$, with different frequency domain scaling.

$$\mathcal{L}_i = \begin{cases} \mathbf{I} & i = 0 \\ \tilde{\mathcal{U}}^{[0]} \ldots \tilde{\mathcal{U}}^{[i-1]} & i \geq 1 \end{cases} \tag{21}$$

The repeated application of the Bellman operator to $\tilde{R}$ (or any other frequency-state function $\tilde{V}$) can be written as

$$(\tilde{T}^\pi)^k \tilde{R} \triangleq \underbrace{\tilde{T}^\pi \cdots \tilde{T}^\pi}_{k \text{ times}} \tilde{V} = \tilde{\mathcal{U}}^{[0]} \tilde{\mathcal{U}}[1] \ldots \tilde{\mathcal{U}}^{[k-1]} \tilde{R}^{[k]} = \mathcal{L}_k \tilde{R}^{[k]}. \tag{22}$$

The next lemma is a pointwise quantification on how the errors $\tilde{\varepsilon}_k$ propagate throughout iteration of ACVI.

Lemma 3. *Consider the ACVI procedure (17). For $k \geq 0$, we have*

$$\tilde{V}_{k+1} = \mathcal{L}_k \tilde{R}^{[k]} + \sum_{i=0}^{k} \mathcal{L}_i \tilde{\varepsilon}_{k+1-i}^{[i]}.$$

Proof. We expand ACVI, and simplify using the introduced notations.

$$
\begin{aligned}
\tilde{V}_1 &\leftarrow \tilde{R}^{[0]} + \tilde{\varepsilon}_1^{[0]} \\
\tilde{V}_2 &\leftarrow \tilde{T}^\pi \tilde{V}_1 + \tilde{\varepsilon}_2^{[0]} &&= \tilde{T}^\pi \left(\tilde{R}^{[0]} + \tilde{\varepsilon}_1^{[0]} \right) + \tilde{\varepsilon}_2^{[0]} \\
& &&= \tilde{R}\mathcal{P}^\pi \left(\tilde{R}^{[1]} + \tilde{\varepsilon}_1^{[1]} \right) + \tilde{\varepsilon}_2^{[0]} \\
& &&= \tilde{\mathcal{U}}^{[0]} \tilde{R}^{[1]} + \tilde{\mathcal{U}}^{[0]} \tilde{\varepsilon}_1^{[1]} + \tilde{\varepsilon}_2^{[0]} \\
\tilde{V}_3 &\leftarrow \tilde{T}^\pi \tilde{V}_2 + \tilde{\varepsilon}_3^{[0]} &&= \tilde{T}^\pi \left(\tilde{\mathcal{U}}^{[0]} \tilde{R}^{[1]} + \tilde{\mathcal{U}}^{[0]} \tilde{\varepsilon}_1^{[1]} + \tilde{\varepsilon}_2^{[0]} \right) + \tilde{\varepsilon}_3^{[0]} \\
& &&= \tilde{R}\mathcal{P}^\pi \left(\tilde{\mathcal{U}}^{[1]} \tilde{R}^{[2]} + \tilde{\mathcal{U}}^{[1]} \tilde{\varepsilon}_1^{[2]} + \tilde{\varepsilon}_2^{[1]} \right) + \tilde{\varepsilon}_3^{[0]} \\
& &&= \tilde{\mathcal{U}}^{[0]} \tilde{\mathcal{U}}^{[1]} \tilde{R}^{[2]} + \tilde{\mathcal{U}}^{[0]} \tilde{\mathcal{U}}^{[1]} \tilde{\varepsilon}_1^{[2]} + \tilde{\mathcal{U}}^{[0]} \tilde{\varepsilon}_2^{[1]} + \tilde{\varepsilon}_3^{[0]} \\
& &&\vdots \\
\tilde{V}_{k+1} &\leftarrow \tilde{T}^\pi \tilde{V}_k + \tilde{\varepsilon}_{k+1}^{[0]} &&= \tilde{\mathcal{U}}^{[0]} \cdots \tilde{\mathcal{U}}^{[k-1]} \tilde{R}^{[k]} + \\
& && \quad \tilde{\mathcal{U}}^{[0]} \cdots \tilde{\mathcal{U}}^{[k-1]} \tilde{\varepsilon}_1^{[k]} + \tilde{\mathcal{U}}^{[0]} \cdots \tilde{\mathcal{U}}^{[k-2]} \tilde{\varepsilon}_2^{[k-1]} + \ldots + \tilde{\mathcal{U}}^{[0]} \tilde{\varepsilon}_k^{[1]} + \tilde{\varepsilon}_{k+1}^{[0]}.
\end{aligned}
$$

Substituting the notation $\mathcal{L}_i = \tilde{\mathcal{U}}^{[0]} \cdots \tilde{\mathcal{U}}^{[i-1]}$ (21) gets us to the desired result. □

Example 1. Consider an MDP with self-returning state transition, i.e., $\mathcal{P}^\pi = \mathbf{I}$. Suppose that we run the exact CVI for k iterations, that is, $\tilde{\varepsilon}_k = 0$ for all $i = 1, \ldots, k+1$. By Lemma 3, we have

$$
\tilde{V}_{k+1} = \mathcal{L}_k \tilde{R}^{[k]} = \tilde{\mathcal{U}}^{[0]} \cdots \tilde{\mathcal{U}}^{[k-1]} \tilde{R}^{[k]}.
$$

As $\tilde{\mathcal{U}}^{[i]} = \tilde{R}^{[i]} \mathcal{P}^\pi = \tilde{R}^{[i]} \mathbf{I}$, we have $\tilde{V}_{k+1} = \tilde{R}^{[0]} \cdots \tilde{R}^{[k]}$. For each state x, the CF of the computed return is $\tilde{V}_{k+1}(\omega; x) = \tilde{R}(\omega; x) \tilde{R}(\gamma\omega; x) \ldots \tilde{R}(\gamma^k \omega; x)$. By Lemma 8, given two independent random variables R_1 and R_2 and constants a_1 and a_2, their CFs satisfy $c_{a_1 R_1 + a_2 R_2}(\omega) = c_{R_1}(a_1 \omega) c_{R_2}(a_2 \omega)$. So $\tilde{V}_{k+1}(\omega; x)$ is the CF of a r.v. $R_0 + \gamma R_1 + \ldots + \gamma^k R_k$ with each $R_i \sim R^\pi(\cdot|x)$ independently. This r.v. is the k-step Monte Carlo approximation of the return. And running CVI for k iterations computes its CF. This observation is more general and holds for arbitrary $\mathcal{P}^\pi$.

We move from pointwise quantification of the errors (Lemma 3) to computing their norm. We first state the following intermediate result.

Lemma 4. *For any bounded* $\tilde{V} : \mathbb{R} \times \mathcal{X} \to \mathbb{R}$*, and any integers* $k, l \geq 0$*, we have*

$$
\begin{aligned}
\left\| \left(\tilde{\mathcal{U}}^{[k]} \tilde{V} \right) (\omega; \cdot) \right\|_\infty &\leq \left\| \tilde{R}^{[k]} (\omega; \cdot) \right\|_\infty \left\| \tilde{V} (\omega; \cdot) \right\|_\infty, \\
\left\| \left(\tilde{\mathcal{U}}^{[k]} \tilde{\mathcal{U}}^{[l]} \tilde{V} \right) (\omega; \cdot) \right\|_\infty &\leq \left\| \tilde{R}^{[k]} (\omega; \cdot) \right\|_\infty \left\| \tilde{R}^{[l]} (\omega; \cdot) \right\|_\infty \left\| \tilde{V} (\omega; \cdot) \right\|_\infty.
\end{aligned}
$$

Proof. For any $x \in \mathcal{X}$, $\omega \in \mathbb{R}$ and $k \geq 0$, we have

$$
\left| \left(\tilde{\mathcal{U}}^{[k]} \tilde{V} \right) (\omega; x) \right| = \left| \tilde{R}^{[k]} (\omega; x) \int \mathcal{P}^\pi (\mathrm{d}y|x) \tilde{V}(\omega; y) \right| \leq \left| \tilde{R}^{[k]} (\omega; x) \right| \left\| \tilde{V} (\omega; \cdot) \right\|_\infty.
$$

Taking the supremum over $x \in \mathcal{X}$, we get $\|(\tilde{\mathcal{U}}^{[k]} \tilde{V})(\omega; \cdot)\|_\infty \leq \|\tilde{R}^{[k]}(\omega; \cdot)\|_\infty \|\tilde{V}(\omega; \cdot)\|_\infty$.

Similarly, we have

$$
\begin{aligned}
\left| \left(\tilde{\mathcal{U}}^{[k]} \tilde{\mathcal{U}}^{[l]} \tilde{V} \right) (\omega; x) \right| &= \left| \tilde{R}^{[k]} (\omega; x) \int \mathcal{P}^\pi (\mathrm{d}y|x) \left(\tilde{R}^{[l]} (\omega; y) \int \mathcal{P}^\pi (\mathrm{d}z|y) \tilde{V}(\omega; z) \right) \right| \\
&\leq \left\| \tilde{V} (\omega; \cdot) \right\|_\infty \left| \tilde{R}^{[k]} (\omega; x) \int \mathcal{P}^\pi (\mathrm{d}y|x) \left| \tilde{R}^{[l]} (\omega; y) \right| \right| \\
&\leq \left\| \tilde{V} (\omega; \cdot) \right\|_\infty \left| \tilde{R}^{[k]} (\omega; x) \right| \left\| \tilde{R}^{[l]} (\omega; \cdot) \right\|_\infty.
\end{aligned}
$$

Taking the supremum over the state space leads to the desired result. □

The following theorem is the main result of this section.

Theorem 5. *Consider the ACVI procedure (17) 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 have*

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

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

Proof. We decompose the error to two parts, one from stopping the exact CVI after K iterations instead of letting $K \to \infty$, and the other is from only approximately performing CVI, which is encoded by having $\tilde{\varepsilon}_k \neq 0$.

We denote the CVF of applying K iterations of the exact CVI by $\tilde{V}^\circ_{K+1}$. It is equal to K-times application $\tilde{T}^\pi$ to $\tilde{R}$, that is $\tilde{V}^\circ_{K+1} = (\tilde{T}^\pi)^{[K]}\tilde{R} = \mathcal{L}_K \tilde{R}^{[K]}$ (cf. (22)).

By the triangle inequality,

$$\begin{aligned} d_{\infty,p}(\tilde{V}_{K+1}, \tilde{V}^\pi) &\leq d_{\infty,p}(\tilde{V}_{K+1}, \tilde{V}^\circ_{K+1}) + d_{\infty,p}(\tilde{V}^\circ_{K+1}, \tilde{V}^\pi), \\ d_{1,p}(\tilde{V}_{K+1}, \tilde{V}^\pi) &\leq d_{1,p}(\tilde{V}_{K+1}, \tilde{V}^\circ_{K+1}) + d_{1,p}(\tilde{V}^\circ_{K+1}, \tilde{V}^\pi). \end{aligned} \tag{23}$$

By (15) and (16), we get

$$\begin{aligned} d_{\infty,p}(\tilde{V}^\circ_{K+1}, \tilde{V}^\pi) &\leq (\gamma^p)^K d_{\infty,p}(\tilde{R}, \tilde{V}^\pi), \\ d_{1,p}(\tilde{V}^\circ_{K+1}, \tilde{V}^\pi) &\leq (\gamma^{p-1})^K d_{1,p}(\tilde{R}, \tilde{V}^\pi). \end{aligned} \tag{24}$$

Let us attend to $d_{\infty,p}(\tilde{V}_{K+1}, \tilde{V}^\circ_{K+1})$ and $d_{1,p}(\tilde{V}_{K+1}, \tilde{V}^\circ_{K+1})$. Lemma 3 shows that

$$\tilde{V}_{K+1} = \mathcal{L}_K \tilde{R}^{[K]} + \sum_{i=0}^{K} \mathcal{L}_i \tilde{\varepsilon}^{[i]}_{K+1-i}.$$

As $d_{\infty,p}(\tilde{V}_{K+1}, \tilde{V}^\circ_{K+1}) = d_{\infty,p}(\tilde{V}_{K+1}, \mathcal{L}_K \tilde{R}^{[K]})$, and likewise for $d_{1,p}$, we need to provide upper bounds for

$$\begin{aligned} d_{1,p}(\tilde{V}_{K+1}, \mathcal{L}_K \tilde{R}^{[K]}) &\leq d_{1,p}\left(\sum_{i=0}^{K} \mathcal{L}_i \tilde{\varepsilon}^{[i]}_{K+1-i}, 0\right), \\ d_{\infty,p}(\tilde{V}_{K+1}, \mathcal{L}_K \tilde{R}^{[K]}) &\leq d_{\infty,p}\left(\sum_{i=0}^{K} \mathcal{L}_i \tilde{\varepsilon}^{[i]}_{K+1-i}, 0\right). \end{aligned}$$

By the repeated application of Lemma 4, we have that for any $\tilde{\varepsilon}$ and $i \geq 0$,

$$\left\|\left(\mathcal{L}_i \tilde{\varepsilon}^{[i]}\right)(\omega; \cdot)\right\|_\infty = \left\|\left(\tilde{\mathcal{U}}^{[0]} \cdots \tilde{\mathcal{U}}^{[i-1]} \tilde{\varepsilon}^{[i]}\right)(\omega; \cdot)\right\|_\infty \leq \left(\prod_{j=0}^{i-1} \left\|\tilde{R}^{[j]}(\omega; \cdot)\right\|_\infty\right) \left\|\tilde{\varepsilon}^{[i]}\right\|_\infty \tag{25}$$

We consider the case of $d_{\infty,p}$ first. Using (25) and the fact that the absolute value $|\tilde{R}(\omega;x)| \leq 1$ (for all $\omega \in \mathbb{R}$) because $\tilde{R}(\cdot;x)$ is a CF (Lemma 8), we get that

$$\begin{aligned}
d_{\infty,p}\left(\sum_{i=0}^{K}\mathcal{L}_i\tilde{\varepsilon}^{[i]}_{K+1-i},0\right) &= \sup_x\sup_\omega\left|\frac{\sum_{i=0}^{K}\left(\mathcal{L}_i\tilde{\varepsilon}^{[i]}_{K+1-i}\right)(\omega;x)}{\omega^p}\right| \\
&\leq \sum_{i=0}^{K}\sup_x\sup_\omega\left|\frac{\left(\mathcal{L}_i\tilde{\varepsilon}^{[i]}_{K+1-i}\right)(\omega;x)}{\omega^p}\right| \\
&\leq \sum_{i=0}^{K}\sup_\omega\frac{\left(\prod_{j=0}^{i-1}\left\|\tilde{R}^{[j]}(\omega;\cdot)\right\|_\infty\right)\left\|\tilde{\varepsilon}^{[i]}_{K+1-i}(\omega;\cdot)\right\|_\infty}{|\omega|^p} \\
&\leq \sum_{i=0}^{K}\sup_\omega\frac{\left\|\tilde{\varepsilon}^{[i]}_{K+1-i}(\omega;\cdot)\right\|_\infty}{|\omega|^p}
\end{aligned} \tag{26}$$

For any $c:\mathbb{R}\to\mathbb{R}$ and $\gamma > 0$ (cf. (12)),

$$\sup_{\omega\in\mathbb{R}}\left|\frac{c(\gamma^i\omega)}{\omega^p}\right| = (\gamma^p)^i\sup_{\omega\in\mathbb{R}}\left|\frac{c(\omega)}{\omega^p}\right|.$$

This allows us to simplify (26) to

$$d_{\infty,p}\left(\sum_{i=0}^{K}\mathcal{L}_i\tilde{\varepsilon}^{[i]}_{K+1-i},0\right) \leq \sum_{i=0}^{K}(\gamma^p)^i\sup_\omega\frac{\|\tilde{\varepsilon}_{K+1-i}(\omega;\cdot)\|_\infty}{|\omega|^p} = \sum_{i=0}^{K}(\gamma^p)^i\left\|\tilde{\varepsilon}_{K+1-i}\right\|_{\infty,p}. \tag{27}$$

We now consider the case of $d_{1,p}$, which is similar.

$$\begin{aligned}
d_{1,p}\left(\sum_{i=0}^{K}\mathcal{L}_i\tilde{\varepsilon}^{[i]}_{K+1-i},0\right) &= \sup_x\int\left|\frac{\sum_{i=0}^{K}\left(\mathcal{L}_i\tilde{\varepsilon}^{[i]}_{K+1-i}\right)(\omega;x)}{\omega^p}\right|\mathrm{d}\omega \\
&\leq \sum_{i=0}^{K}\sup_x\int\left|\frac{\left(\mathcal{L}_i\tilde{\varepsilon}^{[i]}_{K+1-i}\right)(\omega;x)}{\omega^p}\right|\mathrm{d}\omega \\
&= \sum_{i=0}^{K}\int\frac{\sup_x\left|\left(\mathcal{L}_i\tilde{\varepsilon}^{[i]}_{K+1-i}\right)(\omega;x)\right|}{|\omega|^p}\mathrm{d}\omega \\
&\leq \sum_{i=0}^{K}\int\left(\prod_{j=0}^{i-1}\left\|\tilde{R}^{[j]}(\omega;\cdot)\right\|_\infty\right)\frac{\left\|\tilde{\varepsilon}^{[i]}_{K+1-i}(\omega;\cdot)\right\|_\infty}{|\omega|^p}\mathrm{d}\omega \\
&\leq \sum_{i=0}^{K}\int\frac{\left\|\tilde{\varepsilon}^{[i]}_{K+1-i}(\omega;\cdot)\right\|_\infty}{|\omega|^p}\mathrm{d}\omega.
\end{aligned} \tag{28}$$

By using the same change of variable used in (13) in the proof of Lemma 1, we have that for $c:\mathbb{R}\to\mathbb{R}$, $\gamma > 0$, and $p > 1$,

$$\int\frac{|c(\gamma^i\omega)|}{|\omega|^p}\mathrm{d}\omega = \int\frac{|c(\nu)|}{|\frac{\nu}{\gamma}|^p}\frac{1}{\gamma}\mathrm{d}\nu = (\gamma^{p-1})^i\int\frac{|c(\omega)|}{|\omega|^p}\mathrm{d}\omega.$$

This allows us to simplify (28) to

$$d_{1,p}\left(\sum_{i=0}^{K}\mathcal{L}_i\tilde{\varepsilon}^{[i]}_{K+1-i},0\right) \leq \sum_{i=0}^{K}(\gamma^{p-1})^i\int\frac{\|\tilde{\varepsilon}_{K+1-i}(\omega;\cdot)\|_\infty}{|\omega|^p}\mathrm{d}\omega = \sum_{i=0}^{K}(\gamma^{p-1})^i\left\|\tilde{\varepsilon}_{K+1-i}\right\|_{1,p}. \tag{29}$$

Plugging (24), (27), and (29) in (23) leads to the final result. □

This result shows how the errors $\tilde{\varepsilon}_k$ in the ACVI procedure propagate throughout iterations and affect the quality of the approximation of $\tilde{V}^\pi$ by $\tilde{V}_{K+1}$. The error is measured according to the distances $d_{1,p}$ and $d_{\infty,p}$. The upper bounds show that errors in the earlier iterations are geometrically decayed. This entails that if the resources are limited, it is better to ensure the smallness of errors in later iterations. This phenomenon is similar to what we have observed in the conventional value iteration [Farahmand et al., 2010].

As discussed in Section 4, the condition that $d_{\infty,p}(\tilde{R}, \tilde{V}^\pi)$ is finite might be very restrictive for $p > 2$ and even for $p = 2$, it might hold only in special problems. But the finiteness of $d_{\infty,1}$ requires mild conditions. For the finiteness of $d_{\infty,1}(\tilde{R}, \tilde{V}^\pi)$ in the upper bound, the finiteness of the first absolute moment of the reward function is sufficient, as discussed after (16). For the finiteness of $\|\tilde{\varepsilon}_i\|_{\infty,1}$ terms, it is sufficient that $\tilde{\varepsilon}_i(0; x) = 0$ and that its first derivative w.r.t. ω is bounded for all states $x \in \mathcal{X}$, i.e., $|\tilde{\varepsilon}^{(1)}(\omega; x)| < \infty$ (this can be seen from the proof of Lemma 11 in Appendix B). Based on these, so from now on we focus on $p = 1$.

6 From error in frequency domain to error in probability distributions

Theorem 5 in the previous section relates the errors at each iteration of ACVI to the quality of the obtained approximation of $\tilde{V}^\pi$. The error is measured according to the metrics $d_{1,p}$ and $d_{\infty,p}$. These are metrics in the frequency domain. What does having a small error in the frequency domain imply about the quality of approximating the distribution of returns $\bar{V}^\pi$?

From Levy's continuity theorem we know that the pointwise convergence of CF implies the convergence in distribution of their corresponding distributions. This suggest that we could define the error in the frequency domain

$$d_{\text{unif}}(\tilde{V}, \tilde{V}^\pi) = \sup_{x \in \mathcal{X}} \sup_{\omega \in \mathbb{R}} \left| \tilde{V}(\omega; x) - \tilde{V}^\pi(\omega; x) \right| .$$

Nevertheless, we did not define the distance this way because the Bellman operator would not be a contraction w.r.t. to it. So a valid question is whether, or in what sense, the smallness of $d_{\infty,p}(\tilde{V}, \tilde{V}^\pi)$ implies anything about the closeness of their corresponding probability distribution functions $\bar{V}$ and $\bar{V}^\pi$? In this section we show that such a relation indeed exists. We relate $d_{\infty,p}$ and $d_{1,p}$ to the p-smooth Wasserstein distance of the probability distribution functions [Arras et al., 2017].

Definition 1. *Let $p \geq 1$, $\mathcal{C}^p(\Omega)$ be the space of p-times continuous differentiable functions on domain Ω, and $\mathcal{F}_p(\Omega) = \left\{ f \in \mathcal{C}^p(\Omega) : \|f^{(k)}\|_\infty \leq 1, 0 \leq k \leq p \right\}$. For two probability distributions $\mu_1, \mu_2 \in \mathcal{M}(\Omega)$, the p-smooth Wasserstein distance is defined as*

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

Remark 2. Note that the conventional 1-Wasserstein distance is defined as

$$\mathcal{W}_1(\mu_1, \mu_2) = \sup_{f \in \mathrm{Lip}_1(\Omega)} \left| \int f(x) \left(\mathrm{d}\mu_1(x) - \mathrm{d}\mu_2(x) \right) \right| ,$$

where Lip_1 is the space of 1-Lipschitz functions. As $\left\| f^{(1)} \right\|_\infty \leq 1$ implies 1-Lipschitz functions, but not necessarily vice versa, $\mathcal{W}_{\mathcal{C}_1}(\mu_1, \mu_2) \leq \mathcal{W}_1(\mu_1, \mu_2)$.

Let us also define the p-smooth Wasserstein between $\bar{V}_1$ and $\bar{V}_2$ as follows:

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

This is the maximum over states $x \in \mathcal{X}$ of the value of the p-smooth Wasserstein between the distribution of return according to the probability distributions $\bar{V}_1(\cdot; x)$ and $\bar{V}_2(\cdot; x)$.

The following result provides an upper bound for the p-smooth Wasserstein distances of two r.v. based on the distance of their CFs (9) in the frequency domain.

Lemma 6. *Consider the domain $\Omega = [-B, B]$ with $0 < B < \infty$. Let X_1 and X_2 be two random variables with the probability distribution functions $\mu_1, \mu_2 \in \mathcal{M}(\Omega)$, and their corresponding CF*

$c_1, c_2 : \mathbb{R} \to \mathbb{C}$. *Let* $p \geq 1$ *be an integer. We have*

$$\mathcal{W}_{\mathcal{C}_{p+1}}(\mu_1, \mu_2) \leq \frac{2\sqrt{2B}}{\sqrt{\pi}} d_{\infty,p}(c_1, c_2),$$
$$\mathcal{W}_{\mathcal{C}_p}(\mu_1, \mu_2) \leq \frac{\sqrt{2}B}{\sqrt{\pi}} d_{1,p}(c_1, c_2).$$

Proof. Let $f \in \mathcal{C}_c^\infty(\mathbb{R})$ with the support in $[-B, B]$. Denote its Fourier transform $\tilde{f}$, i.e., $\tilde{f} = \frac{1}{\sqrt{2\pi}} \int f(x) e^{-j\omega x} \mathrm{d}x$. So we have $f(x) = \frac{1}{\sqrt{2\pi}} \int \tilde{f}(\omega) e^{+j\omega x} \mathrm{d}\omega$. This is a unitary convention for the Fourier transform. The difference between the expectation of f w.r.t. $X_1 \sim \mu_1$ and $X_2 \sim \mu_2$ is

$$\begin{aligned}
\mathbb{E}\left[f(X_1) - f(X_2)\right] &= \mathbb{E}\left[\frac{1}{\sqrt{2\pi}} \int \tilde{f}(\omega) \left(e^{+j\omega X_1} - e^{+j\omega X_2}\right) \mathrm{d}\omega\right] \\
&= \frac{1}{\sqrt{2\pi}} \int \tilde{f}(\omega) \left(\mathbb{E}\left[e^{j\omega X_1}\right] - \mathbb{E}\left[e^{j\omega X_2}\right]\right) \mathrm{d}\omega \\
&= \frac{1}{\sqrt{2\pi}} \int \tilde{f}(\omega) \left(c_{X_1}(\omega) - c_{X_1}(\omega)\right) \mathrm{d}\omega \\
&= \frac{1}{\sqrt{2\pi}} \int \tilde{f}(\omega) \omega^p \frac{c_{X_1}(\omega) - c_{X_1}(\omega)}{\omega^p} \mathrm{d}\omega \\
&\leq \frac{1}{\sqrt{2\pi}} \int \left|\frac{c_{X_1}(\omega) - c_{X_1}(\omega)}{\omega^p}\right| |\omega|^p |\tilde{f}(\omega)| \mathrm{d}\omega, \qquad (30)
\end{aligned}$$

where we used the definition of a CF. We consider the two parts of the result separately.

Part I) $d_{\infty,p}(c_1, c_2)$**:** We can upper bound (30) by

$$\mathbb{E}\left[f(X_1) - f(X_2)\right] \leq \frac{1}{\sqrt{2\pi}} d_{\infty,p}(c_1, c_2) \int |\omega|^p |\tilde{f}(\omega)| \mathrm{d}\omega. \qquad (31)$$

The integral can be upper bound by using the Cauchy-Schwarz inequality:

$$\begin{aligned}
\int |\omega|^p |\tilde{f}(\omega)| \mathrm{d}\omega &= \int |\omega|^p |\frac{1 + |\omega|}{1 + |\omega|} |\tilde{f}(\omega)| \mathrm{d}\omega \\
&\leq \sqrt{\int |\omega|^{2p} (1 + |\omega|)^2 |\tilde{f}(\omega)|^2 \mathrm{d}\omega} \sqrt{\int \frac{1}{(1 + |\omega|)^2} \mathrm{d}\omega} \\
&\leq 2\sqrt{\int \left(|\omega|^{2p} + |\omega|^{2p+2}\right) |\tilde{f}(\omega)|^2 \mathrm{d}\omega}. \qquad (32)
\end{aligned}$$

Here we used $\int_{-\infty}^{\infty} \frac{1}{(1+|\omega|)^2} \mathrm{d}\omega = 2$ and $(1 + |\omega|)^2 \leq 2(1 + |\omega|^2)$.

The Fourier transform of the k-th derivative of a function satisfies $\mathcal{F}\{f^{(k)}\} = (j\omega)^k \tilde{f}(\omega)$. So by Parseval's theorem, we have

$$\int \left|\omega^p \tilde{f}(\omega)\right|^2 + \left|\omega^{p+1} \tilde{f}(\omega)\right|^2 \mathrm{d}\omega = \int \left|f^{(p)}(x)\right|^2 + \left|f^{(p+1)}(x)\right|^2 \mathrm{d}x.$$

As the support of f is $[-B, +B]$, we have that

$$\begin{aligned}
\int \left|f^{(p)}(x)\right|^2 + \left|f^{(p+1)}(x)\right|^2 \mathrm{d}x &= \int_{-B}^{+B} \left|f^{(p)}(x)\right|^2 + \left|f^{(p+1)}(x)\right|^2 \mathrm{d}x \\
&\leq (2B) \left[\left\|f^{(p)}\right\|_\infty^2 + \left\|f^{(p+1)}\right\|_\infty^2\right].
\end{aligned}$$

Now for any $f \in \mathcal{F}_{p+1}([-B, +B])$, the value of $\|f^{(p)}\|_\infty$ and $\|f^{(p+1)}\|_\infty$ are both less than or equal to 1, and therefore the integral (32) is upper bounded by $2\sqrt{4B}$. By combining this with (31), we get

$$\mathcal{W}_{\mathcal{C}_{p+1}}(\mu_1, \mu_2) = \sup_{f \in \mathcal{F}_{p+1}([-B,+B])} \mathbb{E}\left[f(X_1) - f(X_2)\right] \leq \frac{2\sqrt{2B}}{\sqrt{\pi}} d_{\infty,p}(c_1, c_2).$$

Part 2) $d_{1,p}(c_1, c_2)$**:** We can upper bound (30) by

$$\mathbb{E}\left[f(X_1) - f(X_2)\right] \leq \frac{1}{\sqrt{2\pi}} \sup_{\omega} \left|\omega^p \tilde{f}(\omega)\right| d_{1,p}(c_1, c_2). \tag{33}$$

Observe that for any integrable function $g : \mathbb{R} \to \mathbb{R}$ and its corresponding Fourier transform $\tilde{g}$, for any ω we have that

$$|\tilde{g}(\omega)| = \left|\int g(x) e^{-j\omega x} \mathrm{d}x\right| \leq \int \left|g(x) e^{-j\omega x}\right| \mathrm{d}x \leq \int |g(x)| \mathrm{d}x.$$

So as $(j\omega)^p \tilde{f}(\omega)$ is the Fourier transform of $f^{(p)}$, we get that

$$\sup_{\omega} \left|\omega^p \tilde{f}(\omega)\right| \leq \int \left|f^{(p)}(x)\right| \mathrm{d}x \leq 2B \left\|f^{(p)}\right\|_{\infty},$$

where we used the boundedness of the support of f.

By combining this with (33), we get

$$\mathcal{W}_{\mathcal{C}_p}(\mu_1, \mu_2) = \sup_{f \in \mathcal{F}_p([-B,+B])} \mathbb{E}\left[f(X_1) - f(X_2)\right] \leq \frac{\sqrt{2}B}{\sqrt{\pi}} d_{1,p}(c_1, c_2).$$

□

The proof of this theorem closely follows the same line of argument as the proof of Theorem 1 by Arras et al. [2017]. Our result is both a simplification and an extension. It is a simplification because it considers a bounded support of random variables, whereas Arras et al. [2017] allows an unbounded, but decaying, tail. The result on $d_{1,p}$ is an extension, as Arras et al. [2017] only consider $d_{\infty,p}$.

Given this result, we can use it alongside Theorem 5 to prove the main result of this section.

Theorem 7. *Consider the ACVI procedure (17) after $K \geq 1$ iterations. Assume that $\tilde{\varepsilon}_k(0; x) = 0$ for all $x \in \mathcal{X}$ and $k = 1, \ldots, K+1$. Furthermore, assume that the immediate reward distribution $\mathcal{R}^\pi(\cdot|x)$ is R_{max}-bounded. We then have*

$$\mathcal{W}_{\mathcal{C}_2}(\bar{V}_{K+1}, \bar{V}^\pi) \leq \frac{2\sqrt{2}}{\sqrt{\pi}} \sqrt{\frac{R_{max}}{1-\gamma}} \left[\sum_{i=0}^{K} \gamma^i \left\|\tilde{\varepsilon}_{K+1-i}\right\|_{\infty,1} + \frac{2\gamma^K}{1-\gamma} R_{max}\right].$$

Proof. We evoke Theorem 5 with the choice of $p = 1$ to upper bound $d_{\infty,1}(\tilde{V}_{K+1}, \tilde{V}^\pi)$. This in turn provides a pointwise (over states $x \in \mathcal{X}$) upper bound guarantee for $d_{\infty,1}(\tilde{V}_{K+1}(\cdot; x), \tilde{V}^\pi(\cdot; x))$. This allows us to focus on the CF of each state separately. So Lemma 6 shows that for each $x \in \mathcal{X}$,

$$\mathcal{W}_{\mathcal{C}_2}(\bar{V}_{K+1}(\cdot; x), \bar{V}^\pi(\cdot; x)) \leq \frac{2\sqrt{2}}{\sqrt{\pi}} \sqrt{\frac{R_{\max}}{1-\gamma}} \left[\sum_{i=0}^{K} \gamma^i \left\|\tilde{\varepsilon}_{K+1-i}\right\|_{\infty,1} + \gamma^K d_{\infty,1}(\tilde{R}(\cdot; x), \tilde{V}^\pi(\cdot; x))\right].$$

It remains to upper bound $d_{\infty,1}(\tilde{R}(\cdot; x), \tilde{V}^\pi(\cdot; x))$. Because the immediate reward distribution is $R_{\max}$-bounded, the distribution of random returns $\bar{V}^\pi$ would be $\frac{R_{\max}}{1-\gamma}$-bounded. Lemma 11 allows us to upper bound $d_{\infty,1}(\tilde{R}(\cdot; x), \tilde{V}^\pi(\cdot; x))$ by $\mathbb{E}\left[|R^\pi(x)|\right] + \mathbb{E}\left[|G^\pi(x)|\right] \leq R_{\max} + \frac{R_{\max}}{1-\gamma} = \frac{2-\gamma}{1-\gamma} R_{\max} \leq \frac{2}{1-\gamma} R_{\max}$. This finishes the proof. □

This upper bound can be simplified if we are willing to provide a uniform over iterations upper bound on $\|\tilde{\varepsilon}_{K+1-i}\|_{\infty,1}$. In that case, we have

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

We note that the 2-smooth Wasserstein distance $\mathcal{W}_{\mathcal{C}_2}$, which is an integral probability metric [Müller, 1997], is only one of the many distances between probability distributions [Gibbs and Su, 2002]. The choice of the right probability distance most likely depends on the performance measure we would like the policy to optimize. Studying this further is an interesting topic of future research.

7 Conclusion

This paper laid the groundwork for a new class of distributional RL algorithms. We have shown that one might represent the uncertainty about the return in the frequency domain, and such a representation (called Characteristic Value Function) enjoys properties such as satisfying a Bellman equation and having a contractive Bellman operator. This in turn allows us to compute the CVF by an iterative method called Characteristic Value Iteration. We also showed the effect of errors in the iterative procedure, and provided error propagation results, in both the frequency domain and the probability distribution space.

This paper is only the first step towards understanding CVFs and their properties. Among remaining questions is how to perform the regression step (18) of ACVI properly and efficiently. Specifically, how should we set the weighting function $w(\omega)$ in order to achieve accurate CVF in frequencies that are relevant for the tasks we want to solve. Studying other distances between CFs and their properties is another interesting research directions. This work only focused on the policy evaluation problem, so another obvious direction is designing risk-aware policy optimization algorithms based on CVF. Finally, empirically evaluating this approach for return uncertainty representation may lead to better understanding of its strengths and weaknesses.

A Characteristic function of a random variable

Given a real-valued random variable X with the probability distribution $\mu \in \mathcal{M}(\mathbb{R})$, the space of probability distributions over $\mathbb{R}$, its corresponding CF $c_X : \mathbb{R} \to \mathbb{C}$ is the function defined as

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

where $j = \sqrt{-1}$ is the imaginary unit. If the distribution has a density $p(x) = \frac{\mathrm{d}\mu}{\mathrm{d}\lambda}$ w.r.t. the Lebesgue measure λ, we have $c_X(\omega) = \int \exp(jx\omega)p(x)\mathrm{d}x$ too.

The CF of a probability distribution is closely related to the Fourier transform of its probability distribution function. The Fourier transform of a function $f : \mathbb{R} \to \mathbb{C}$ is defined as[5]

$$\tilde{f}(\omega) \triangleq \mathcal{F}\left\{f\right\}(\omega) = \int f(x)e^{-j\omega x}\mathrm{d}x.$$

Hence, c_X is $\overline{\mathcal{F}\left\{p\right\}}$, the complex conjugate of the Fourier transform of the density p.

If X has a probability distribution μ_θ parameterized by θ, we may refer to its CF by c_θ.

Given independent samples $X_1, \dots, X_n$ from μ, the Empirical Characteristic Function (ECF) is defined as

$$c_n(w) \triangleq \frac{1}{n}\sum_{i=1}^{n} e^{jX_i\omega}. \qquad \forall \omega \in \mathbb{R}$$

ECF can be seen as the CF of the empirical measure, which assigns the probability

$$\mu_n(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{X_i \in A\},$$

to any measurable set A of $\mathbb{R}$ (of an appropriate σ-algebra, e.g., Borel σ-algebra). It is easy to see that because of the law of large numbers, $c_n(\omega) \to c_X(\omega)$ (a.s.) for any fixed ω.

We collect some useful properties of the CF in the following lemma, see e.g., Chapters 16 and 18 of Williams [1991] or Chapter 11 of Rosenthal [2006].

Lemma 8. *The characteristic function of a random variable X has the following properties:*

- $c_X(0) = 1$.
- $|c_X(\omega)| \leq 1$ *for all* $\omega \in \mathbb{R}$.

[5]There are several conventions regarding notations and normalization factors.

- *The function $\omega \mapsto c_X(\omega)$ is uniformly continuous in $\mathbb{R}$.*
- $c_{(-X)}(\omega) = \overline{c_X(\omega)}$.
- $c_{aX+b}(\omega) = e^{jb\omega} c_X(a\omega)$.
- *If X and Y are two (conditionally) independent random variables, $c_{X+Y}(\omega) = c_X(\omega) c_Y(\omega)$.*
- *If $k \in \mathbb{N}$ and $\mathbb{E}\left[|X|^k\right] < \infty$, the function $c_X(\omega)$ is k times differentiable and we have $c_X^{(k)}(\omega) = \mathbb{E}\left[(jX)^k e^{jX\omega}\right]$. In particular, the k-th moment of X satisfies $c_X^{(k)}(0) = j^k \mathbb{E}\left[X^k\right]$.*
- *(Levy Inversion Formula) If $\int |c_X(\omega)| \mathrm{d}\omega < \infty$, then X has a continuous probability density function $p(x)$ and $p(x) = \frac{1}{2\pi} \int \exp(-j\omega x) c_X(\omega) \mathrm{d}\omega$.*
- *(Levy's Convergence Theorem) Let (μ_n) be a sequence of probability distributions, and let (c_n) denote their corresponding CF. Suppose that $c(\omega) = \lim c_n(\omega)$ exists for all $\omega \in \mathbb{R}$ and c is continuous at 0. Then c is a CF of some distribution μ and $\mu_n \to \mu$ in distribution.*

Note that we stated a simplified Levy Inversion formula; a more general inversion formula exists even when the r.v. X does not have a density.

B Properties of the distance metrics $d_{1,p}$ and $d_{\infty,p}$

We provide some properties of the distances $d_{\infty,p}$ and $d_{1,p}$.

Proposition 9. *The distance functions $d_{1,p}$ and $d_{\infty,p}$ are metrics.*

Proof. We first consider $d_{\infty,p}$ defined for the CF and verify the properties of being a metric. The verification for $d_{\infty,p}$ for $\tilde{V}$ is similar.

It is clear that $d_{\infty,p}(c_1, c_2) \geq 0$. Whenever $d_{\infty,p}(c_1, c_2) = \sup_\omega |\frac{c_1(\omega) - c_2(\omega)}{\omega^p}|$ is equal to zero, it entails that $c_1(\omega) = c_2(\omega)$ for all $\omega \in \mathbb{R}$.

We also have $d_{\infty,p}(c_1, c_2) = d_{\infty,p}(c_2, c_1)$.

Finally, notice that for $c_1, c_2, c_3 : \mathbb{R} \to \mathbb{C}$, we have

$$
\begin{aligned}
d_{\infty,p}(c_1, c_2) = \sup_\omega \left| \frac{c_1(\omega) - c_2(\omega)}{\omega^p} \right| &= \sup_\omega \left| \frac{c_1(\omega) - c_3(\omega) + c_3(\omega) - c_2(\omega)}{\omega^p} \right| \\
&\leq \sup_\omega \left[\frac{|c_1(\omega) - c_3(\omega)| + |c_3(\omega) - c_2(\omega)|}{|\omega|^p} \right] \\
&\leq \sup_\omega \left| \frac{c_1(\omega) - c_3(\omega)}{\omega^p} \right| + \sup_\omega \left| \frac{c_3(\omega) - c_2(\omega)}{\omega^p} \right| \\
&= d_{\infty,p}(c_1, c_3) + d_{\infty,p}(c_3, c_2)
\end{aligned}
$$

The proof for $d_{1,p}$ is the same except that $\int \left| \frac{c_1(\omega) - c_2(\omega)}{\omega^p} \right| \mathrm{d}\omega = 0$ only entails that $c_1(\omega) = c_2(\omega)$ almost surely. □

We define the space of VCF $\mathcal{V} = \left\{ \tilde{V} : \mathbb{R} \times \mathcal{X} \to \mathbb{C}_1 \, : \, \tilde{V}(0; x) = 1 \right\}$.[6] We want to show that $\mathcal{V}$ with metric $d_{\infty,p}$ is a complete space. Showing that a space is complete allows us to use Banach fixed point theorem to show that the fixed point of a contraction operator is within the space; if the space is not complete, the fixed point might be outside the space.

Proposition 10. *The metric space $(\mathcal{V}, d_{\infty,p})$ is complete.*

[6] This space is larger than the space of feasible VCFs because a CF is uniformly continuous, but $\mathcal{V}$ does not have any continuity restriction.

Proof. Let $(\tilde{V}_n)$ be a Cauchy sequence in $\mathcal{V}$ w.r.t. $d_{\infty,p}$. To show that $\mathcal{V}$ is complete, we have to prove that there exists a $\tilde{V} \in \mathcal{V}$ such that $d_{\infty,p}(\tilde{V}_n, \tilde{V}) \to 0$ as $n \to \infty$.[7]

The fact that $(\tilde{V}_n)$ is a Cauchy sequence in $d_{\infty,p}$ means that for any $\varepsilon > 0$, there exists an integer N such that for any $n, m \geq N$, we have $d_{\infty,p}(\tilde{V}_n, \tilde{V}_m) < \varepsilon$. So

$$\sup_{x\in\mathcal{X}} \sup_{\omega} \left| \frac{\tilde{V}_n(\omega; x) - \tilde{V}_m(\omega; x)}{\omega^p} \right| < \varepsilon \Rightarrow \left| \tilde{V}_n(\omega; x) - \tilde{V}_m(\omega; x) \right| < \varepsilon|\omega|^p, \quad \forall \omega \in \mathbb{R}\backslash\{0\}, \forall x \in \mathcal{X}.$$

We would like to show that for any fix $x \in \mathcal{X}$ and $\omega \neq 0$, the sequence $(\tilde{V}_n(\omega; x))$ is Cauchy too. For any $\varepsilon' > 0$, let us pick $\varepsilon = \frac{\varepsilon'}{|\omega|^p}$. As $(\tilde{V}_n)$ is Cauchy w.r.t. $d_{\infty,p}$, there exists an integer number N such that for any $n, m \geq N$, we have $\left| \frac{\tilde{V}_n(\omega;x) - \tilde{V}_m(\omega;x)}{\omega^p} \right| < \varepsilon$. This is equivalent to $\left| \tilde{V}_n(\omega; x) - \tilde{V}_m(\omega; x) \right| < \varepsilon|\omega|^p = \varepsilon'$, which shows that $(\tilde{V}_n(\omega; x))$ is indeed a Cauchy sequence.

As the real-valued sequence $(\tilde{V}_n(\omega; x))$ is a Cauchy sequence and $\mathbb{R}$ is complete, the pointwise sequence $\tilde{V}_n(\omega; x)$ converges to a limit. We define the following function for all $x \in \mathcal{X}$ and $\omega \in \mathbb{R} \setminus \{0\}$.

$$\tilde{V}(\omega; x) = \lim_{n\to\infty} \tilde{V}_n(\omega; x)$$

For $\omega = 0$, we pick $\tilde{V}(0; x) = 1$.

As $\tilde{V}(\omega; x) \to \tilde{V}_m(\omega; x)$ when $m \to \infty$, we have

$$\begin{aligned} d_{\infty,p}(\tilde{V}_n, \tilde{V}) &= \sup_{x,\omega} \left| \frac{\tilde{V}_n(\omega; x) - \tilde{V}(\omega; x)}{\omega^p} \right| \\ &= \sup_{x,\omega} \lim_{m\to\infty} \left| \frac{\tilde{V}_n(\omega; x) - \tilde{V}_m(\omega; x)}{\omega^p} \right| \\ &\leq \liminf_{m\to\infty} \sup_{x,\omega} \left| \frac{\tilde{V}_n(\omega; x) - \tilde{V}_m(\omega; x)}{\omega^p} \right| = \liminf_{m\to\infty} d_{\infty,p}(\tilde{V}_n, \tilde{V}_m). \end{aligned}$$

As $(\tilde{V}_n)$ is a Cauchy sequence, for all $\varepsilon > 0$, there exists N such that for all $n, m \geq N$, $d_{\infty,p}(\tilde{V}_n, \tilde{V}_m) < \varepsilon$, which along the inequality above show that $d_{\infty,p}(\tilde{V}_n, \tilde{V}) < \varepsilon$ for all $n \geq N$. This proves that $\lim_{n\to\infty} d_{\infty,p}(\tilde{V}_n, \tilde{V}) = 0$, as desired. ☐

Showing whether $d_{1,p}$ also defines a complete metric space is an interesting question postponed to a future work.

We provide a condition under which the distance $d_{\infty,p}$ between two CF c_1, c_2 would be finite.

Lemma 11. *Consider two random variables X_1 and X_2 with their corresponding CF $c_1, c_2 : \mathbb{R} \to \mathbb{C}$. Let $p \geq 1$. If*

1. *(Matched first $p-2$-th moments) $\mathbb{E}\left[X_1^k\right] = \mathbb{E}\left[X_2^k\right]$ for $k = 1, \ldots, p-1$ (for $p \geq 2$),*
2. *(Finite p-th moments) $\mathbb{E}\left[|X_1|^p\right], \mathbb{E}\left[|X_2|^p\right] < \infty$,*

the distance $d_{\infty,p}(c_1, c_2)$ would be finite and can be upper bounded by $\frac{\mathbb{E}[|X_1|^p] + \mathbb{E}[|X_2|^p]}{p!}$.

Proof. We use Taylor series expansion of c_1 and c_2 to provide an upper bound on $\frac{c_1(\omega) - c_2(\omega)}{\omega^p}$. As $\mathbb{E}\left[|X|^p\right]$ is finite for both X_1 and X_2, they are finite for $k = 0, \ldots, p-1$ too. By Lemma 8, the functions $c_1(\omega)$ and $c_2(\omega)$ are differentiable for $k = 0, \ldots, p$, and for both of them we have

$$c(\omega) = c(0) + c^{(1)}(0)\frac{\omega^1}{1!} + \ldots + c^{(p-1)}(0)\frac{\omega^{p-1}}{(p-1)!} + c^{(p)}(\nu)\Big|_{0\leq\nu\leq\omega} \frac{\omega^p}{p!}.$$

[7] The proof of this result closely follows the proof of Theorem 2.4 of Hunter and Nachtergaele [2001].

Therefore for any ω, we can write

$$\frac{c_1(\omega) - c_2(\omega)}{\omega^p} = \sum_{k=0}^{p-1} \left(c_1^{(k)}(0) - c_1^{(k)}(0) \right) \frac{1}{k!\omega^{p-k}} + \frac{1}{p!} \left(c_1^{(p)}(\nu_1)\Big|_{0<\nu_1<\omega} - c_2^{(p)}(\nu_1)\Big|_{0<\nu_1<\omega} \right).$$

Note that if $c_1^{(k)}(0) \neq c_2^{(k)}(0)$ (for $k = 0, \ldots, p-1$), the corresponding term in the summation would be singular at $\omega = 0$. Because $c_X^{(k)}(0) = j^k \mathbb{E}\left[X^k\right]$, the condition of moments of the random variables being matched implies that the summation is zero. Under that condition, we have

$$\frac{c_1(\omega) - c_2(\omega)}{\omega^p} = \frac{c_1^{(p)}(\nu_1) - c_2^{(p)}(\nu_2)}{p!},$$

for some $0 \leq \nu_1, \nu_2 \leq \omega$. We use the definition of CF to write

$$\begin{aligned} \left| c_1^{(p)}(\nu_1) - c_2^{(p)}(\nu_2) \right| &= \left| \mathbb{E}\left[(jX_1)^p e^{j\nu_1 X_1} \right] - \mathbb{E}\left[(jX_2)^p e^{j\nu_2 X_2} \right] \right| \\ &\leq \mathbb{E}\left[\left| (jX_1)^p e^{j\nu_1 X_1} \right| \right] + \mathbb{E}\left[\left| (jX_2)^p e^{j\nu_2 X_2} \right| \right] \\ &\leq \mathbb{E}\left[|X_1|^p \right] + \mathbb{E}\left[|X_2|^p \right]. \end{aligned}$$

Therefore, if in addition to the first condition, the p-th moments $\mathbb{E}\left[|X_1|^p\right]$ and $\mathbb{E}\left[|X_2|^p\right]$ are finite too,

$$d_{\infty,p}(c_1, c_2) = \sup_\omega \left| \frac{c_1(\omega) - c_2(\omega)}{\omega^p} \right| \leq \frac{\mathbb{E}\left[|X_1|^p\right] + \mathbb{E}\left[|X_2|^p\right]}{p!}$$

is finite too, which is the desired result. □

Observe that $d_{\infty,1}(c_1, c_2)$ is bounded as long as $\mathbb{E}\left[|X_1|\right]$ and $\mathbb{E}\left[|X_2|\right]$ are finite, which is quite mild. A similar result holds if we replace CF with CVF, i.e., $d_{\infty,p}(\tilde{V}_1, \tilde{V}_2)$. The required condition would then be on the moments of the set of random variables, indexed by $x \in \mathcal{X}$, that have CF of $\tilde{V}_1(\cdot; x)$ and $\tilde{V}_2(\cdot; x)$.

We leave the result specifying the conditions for $d_{1,p}(c_1, c_2)$ to be bounded as a future work.

C A Study on function approximation error

We argued in Section 4.1 that performing CVI exactly may not be practical and we have to perform it approximately. This might be done by solving a sequence of regression problems, which find $\tilde{V}_{k+1} \approx \tilde{T}^\pi \tilde{V}_k$. A regression estimator such as (18) finds the estimate within a function space $\mathcal{F}$, which is often smaller than the space of all possible CVF, which is a subset of $\mathcal{V} = \{\tilde{V} : \mathbb{R} \times \mathcal{X} \to \mathbb{C}_1 : \tilde{V}(0; x) = 1\}$. As a result, we might have some function approximation error. We would like to know whether it is possible to have small function approximation error under some reasonable assumptions on the reward distribution and the choice of $\mathcal{F}$. This section attends to this question.

The function approximation error, however, is only one source of error in the analysis of a regression algorithm. Another source of error is the estimation error, which reflects the effect of having a finite number of samples. The estimation error depends on the complexity of the function space, which can be quantified in terms of its covering number. We provide a covering number result for a particular choice of function space in Appendix D. We should note that even though studying the function approximation error and covering number of a function space are crucial steps in the error analysis of a regression method, we do not provide a complete analysis of the regression problem that should be solved at each step of CVI in this work.

Before going into the detail, we briefly describe the result: If the reward distribution is smooth in a certain sense, a band-limited function class $\mathcal{F}_b$, to be defined shortly, provides an approximation error that goes to zero as the bandwidth b increases. Furthermore, if the first s absolute moments of the reward distribution are finite (uniformly for all $x \in \mathcal{X}$), the CVF $\tilde{V}(\cdot; x)$ belongs to the smoothness class $C^s([-b, b]) \cap \mathcal{F}_b$. This leads to a well-behaving covering number, which can be used to obtain a convergence rate for estimation error.

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

$$\mathcal{F}_b = \left\{ \tilde{V} : \mathbb{R} \times \mathcal{X} \to \mathbb{C}_1 \, : \, \tilde{V}(0; x) = 1, \tilde{V}(\omega; x) = 0 \; \forall |\omega| > b \right\}. \tag{34}$$

These are functions whose frequency component can be non-zero only in $|\omega| \leq b$. Soon we show that this function space can approximate the CF of a large class of distributions.

We define the class of β-smooth and super-smooth reward distributions, following the definition by Fan [1991].

Definition 2. *The reward distribution $\mathcal{R}^\pi$ is β-smooth if for all $x \in \mathcal{X}$,*

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

for some $c_0, c_1, \beta > 0$ and for $|\omega| \geq \omega_0$ with some finite $\omega_0 \geq 0$.

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

$$c_0 |\omega|^{\beta_0} \exp\left(- \frac{|\omega|^\beta}{\tau} \right) \leq |\tilde{R}(\omega; x)| \leq c_1 |\omega|^{\beta_1} \exp\left(- \frac{|\omega|^\beta}{\tau} \right),$$

for some positive constants c_0, c_1, τ, β and constants β_0 and β_1, and for $|\omega| \geq \omega_0$ with some finite $\omega_0 \geq 0$.

The condition of being β-smooth is satisfied by many distributions such as exponential ($\beta = 1$), uniform ($\beta = 1$), gamma ($\beta = k$, the shape parameter), etc. The condition of being super-smooth is satisfied for distributions such as normal ($\beta = 2$), Cauchy ($\beta = 1$), etc.

We would like to know that if the reward distribution is smooth (or super-smooth), what is the smallest $\|\tilde{\varepsilon}\|_{\infty,p}$ for $\tilde{\varepsilon} = \tilde{V} - \tilde{T}^\pi \tilde{V}'$ with $\tilde{V}$ being restricted to be in $\mathcal{F}_b$. The following proposition answers this question.

Theorem 12. *Consider function space $\mathcal{F}_b$ with $b \geq \omega_0$ (cf. Definition 2). If $\mathcal{R}$ is a β-smooth distribution, we have*

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

$$\inf_{\tilde{V} \in \mathcal{F}_b} \left\| \tilde{V} - \tilde{R} \right\|_{\infty,p} \leq \frac{c_1}{b^{p+\beta}}.$$

If $\mathcal{R}$ is β-super smooth distribution, under the condition that either (1) $\beta_1 \leq p$ or (2) $\beta_1 > p$ and $b > \sqrt[\beta]{\frac{\tau(\beta_1 - p)}{\beta}}$, we have

$$\sup_{\tilde{V}' \in \mathcal{V}} \inf_{\tilde{V} \in \mathcal{F}_b} \left\| \tilde{V} - \tilde{T}^\pi \tilde{V}' \right\|_{\infty,p} \leq c_1 |b|^{\beta_1 - p} \exp\left(- \frac{b^\beta}{\tau} \right),$$

$$\inf_{\tilde{V} \in \mathcal{F}_b} \left\| \tilde{V} - \tilde{R} \right\|_{\infty,p} \leq c_1 |b|^{\beta_1 - p} \exp\left(- \frac{b^\beta}{\tau} \right).$$

Proof. For any $\tilde{V}' \in \mathcal{V}$, consider $(\tilde{T}^\pi \tilde{V}')(\omega; x) = \tilde{R}(\omega; x) \int \mathcal{P}^\pi(\mathrm{d}y|x) \tilde{V}'(\gamma\omega; y)$ and define a function $\tilde{V}(\omega; x) = (\tilde{T}^\pi \tilde{V}')(\omega; x) \mathbb{I}\{\omega \in [-b, +b]\}$. This function is a CF and is zero outside $[-b, b]$, so it belongs to $\mathcal{F}_b$. So $\tilde{\varepsilon} = \tilde{V} - \tilde{T}^\pi \tilde{V}'$ is $\tilde{\varepsilon}(\omega; x) = (\tilde{T}^\pi \tilde{V}')(\omega; x) \mathbb{I}\{\omega \notin [-b, +b]\}$. As $|\tilde{V}'(\omega; x)| \leq 1$ for all x and ω, we have

$$\begin{aligned}
|\tilde{\varepsilon}(\omega; x)| &= \left| \mathbb{I}\{\omega \notin [-b, +b]\} \tilde{R}(\omega; x) \int \mathcal{P}^\pi(\mathrm{d}y|x) \tilde{V}^\pi(\gamma\omega; y) \right| \\
&\leq \mathbb{I}\{\omega \notin [-b, +b]\} \left| \tilde{R}(\omega; x) \right| \int \mathcal{P}^\pi(\mathrm{d}y|x) \left| \tilde{V}^\pi(\gamma\omega; y) \right| \\
&\leq \mathbb{I}\{\omega \notin [-b, +b]\} \left| \tilde{R}(\omega; x) \right|.
\end{aligned}$$

Under the β-smooth condition, the norm of $\tilde{\varepsilon}$ can be upper bounded by

$$\|\tilde{\varepsilon}\|_{\infty,p} = \sup_x \sup_{|\omega| > b} \left| \frac{\tilde{R}(\omega; x)}{\omega^p} \right| \leq \sup_{|\omega| > b} \frac{c_1}{|\omega|^{p+\beta}} = \frac{c_1}{b^{p+\beta}}. \tag{35}$$

As this holds for any $\tilde{V}' \in \mathcal{V}$, by taking the supremum over $\mathcal{V}$ we obtain the first statement for the β-smooth case. The proof of the second statement is essentially the same with the difference that we choose $\tilde{V}(\omega; x) = \tilde{R}(\omega; x)\mathbb{I}\{\omega \in [-b, +b]\} \in \mathcal{F}_b$ and compute the norm of $\tilde{\varepsilon} = \tilde{R}(\omega; x)\mathbb{I}\{\omega \notin [-b, +b]\}$.

For the β-super smooth case, the argument is similar except that instead of (35), we have

$$\|\tilde{\varepsilon}\|_{\infty,p} \leq \sup_{|\omega|>b} \frac{c_1 \exp\left(-\frac{|\omega|^\beta}{\tau}\right)}{|\omega|^{p-\beta_1}}. \tag{36}$$

The function in the upper bound is not monotonically non-increasing as a function of ω (it might increase for small ω before the exponential term dominates), but under the conditions specified in the statement of the theorem, it is. In that case, we simply replace ω with b. $\square$

This result shows that for the class of β-smooth or super smooth reward distributions, the error in approximating $\tilde{T}^\pi \tilde{V}'$ (for any $\tilde{V}' \in \mathcal{V}$) decreases as the bandwidth b increases. The rate depends on β and p for the smooth distributions as well as τ and β_1 for the super smooth ones. As discussed in Section 5, the choice of $p > 2$ is restrictive. So we summarize the result by stating that for $p = 1$, which we often care about, we have

$$\sup_{\tilde{V}' \in \mathcal{V}} \inf_{\tilde{V} \in \mathcal{F}_b} \left\| \tilde{V} - \tilde{T}^\pi \tilde{V}' \right\|_{\infty,1} \leq \begin{cases} \frac{c_1}{b^{1+\beta}}, & \text{(smooth)} \\ c_1 |b|^{\beta_1 - 1} \exp\left(-\frac{b^\beta}{\tau}\right). & \text{(super smooth)} \end{cases}$$

Analyzing the approximation error is only one part of the error analysis of a regression estimator. Another part is the analysis of the estimation error. One may use any universally consistent regression estimator, such as a K-NN estimator or many other partitioning-based estimator, to show that the estimation error goes to zero as the number of samples increases. This along with the above approximation error lead to a controlled asymptotic upper bound.

Providing a convergence rate for the estimation error, however, requires some more (mild) assumptions on the complexity of the function class. The function class $\mathcal{F}_b$ (34) is quite large. Even if we fix a single state x, the function $\tilde{V}(\cdot; x)$ with $\tilde{V} \in \mathcal{F}_b$ belongs to the space of 1-bounded functions on the domain $[-b, b]$. Learning such a function can be arbitrary slow, cf. Theorem 3.1 of Györfi et al. [2002]. This might appear hopeless, but it is not.

First of all, the function space $\mathcal{F}_b$ was chosen needlessly large. A CF is uniformly continuous, so we could choose to work with the smaller space of 1-bounded continuous functions. Moreover, with some extra mild assumptions, we can define a function space that is reasonably small and can potentially lead to a convergence rate for the estimation error. In particular, if the reward distribution has s finite absolute moments, its CF $\tilde{R}(\cdot; x)$ is s-times differentiable (Lemma 8). The space of s-times differentiable function is regular enough for a relatively fast convergence rate (depending on s). We provide the covering number result for such a space in Appendix D. In the rest of this section, we show that choosing this smaller function space still leads to reasonable function approximation properties. Let us introduce the necessary definitions.

Given an open set $\Omega \subset \mathbb{R}^d$, denote $\mathcal{C}^s(\Omega)$ as the class of s-times differentiable functions with the norm defined as

$$\|f\|_{\mathcal{C}^s} \triangleq \sum_{i=0}^{s} \left\| f^{(i)} \right\|_\infty. \tag{37}$$

If we want to emphasize that the domain is Ω, we use $\|\cdot\|_{\mathcal{C}^s(\Omega)}$ and $\|\cdot\|_{\infty(\Omega)}$, for the supremum norm. We use $c\Omega$ (with $c > 0$) to denote the set $\{\, c\omega \,:\, \omega \in \Omega \,\}$. In our applications, Ω would be an interval in $\mathbb{R}$, e.g., $(-b, b)$ for some $b > 0$.

Denote the class of CVF functions that are s-smooth function over ω with bandwidth of b and an r-bounded norm by

$$\mathcal{F}_{b,r}^s = \left\{ \tilde{V} \in \mathcal{F}_b \,:\, \|\tilde{V}(\cdot; x)\|_{\mathcal{C}^s((-b,b))} \leq r, \, \forall x \in \mathcal{X} \right\}. \tag{38}$$

This is a subset of $\mathcal{F}_b$ that contains VCFs that are s-smooth in the frequency domain with a certain norm of smoothness.

We denote the smoothness norm of $\tilde{V} \in \mathcal{F}^s_{b,r}$ by

$$\left\|\tilde{V}\right\|_{\mathcal{C}^s} = \left\|\tilde{V}\right\|_{\mathcal{C}^s(\Omega)} = \sup_{x \in \mathcal{X}} \left\|\tilde{V}(\cdot; x)\right\|_{\mathcal{C}^s(\Omega)}.$$

We would like to study the function approximation properties of $\mathcal{F}^s_{b,r}$, similar to Theorem 12. For simplicity, we focus only on β-smooth reward distributions, and not the super smooth ones.

We first provide some intermediate results.

Proposition 13. *Suppose that the reward random variable $R(x) \sim \mathcal{R}^\pi(\cdot|x)$ has finite absolute moments up to order s, with*

$$\mathbb{E}\left[|R(x)|^i\right] \leq m_i^i. \qquad i = 1, \dots, s$$

Then, for any $\omega \in \mathbb{R}$, we have

$$|\tilde{R}^{(i)}(\omega; x)| \leq m_i^i. \qquad i = 1, \dots, s$$

Proof. This is the direct consequence of Lemma 8 applied to the random variable $R(x) \sim \mathcal{R}^\pi(\cdot|x)$ i.e., $|\tilde{R}^{(i)}(\omega; x)| = |\mathbb{E}\left[(jR)^i e^{jR\omega}\right]| \leq \mathbb{E}\left[|(jR)^i||e^{jR\omega}|\right] = \mathbb{E}\left[|R|^i\right]$. $\square$

This result guarantees that as long as the reward distribution has s finite absolute moments, the CF $\tilde{R}^\pi$ is s-times continuously differentiable too, i.e., $\tilde{R}^\pi(\cdot; x) \in \mathcal{C}^s(\mathbb{R})$. This result along with an argument similar to the proof of Theorem 12 can be used to show that at the first iteration of ACVI, where we would like to approximate $\tilde{R}$ with $\tilde{V}_1$, we can choose $\tilde{V}_1(\cdot; x)$ to be from the smoothness class $\mathcal{C}^s((-b, b))$ and incur a small error. The error depends on the bandwidth b. The moment condition shows that $\tilde{R}^\pi \in \mathcal{F}^s_{b,r}$ with $b = \sum_{i=0}^s m_i^i$.

The challenge, however, is to show that after applying the Bellman operator $\tilde{T}^\pi$ to $\tilde{V}_1$ (and to $\tilde{V}_k$ with $k > 1$ in later iterations), it still remains in the same (or similar) smoothness class. If not, our function approximation result would only be applicable for the first iteration of the ACVI.

Let us denote two notations for the supremum norm of $\tilde{V}$. Given a domain Ω and a fixed $x \in \mathcal{X}$, we use

$$\left\|\tilde{V}(\cdot; x)\right\|_{\infty(\Omega)} = \sup_{\omega \in \Omega} |\tilde{V}(\omega; x)|,$$

if we want to emphasize the domain where the supremum is taken over. We denote the supremum norm of a VCF $\tilde{V}$ over both state and frequency space by

$$\left\|\tilde{V}\right\|_\infty = \left\|\tilde{V}\right\|_{\infty(\Omega), \infty} = \sup_{x \in \mathcal{X}} \sup_{\omega \in \Omega} \left|\tilde{V}(\omega; x)\right|.$$

We use $\|\tilde{V}\|_{\infty(\Omega),\infty}$ whenever we want to emphasize the domain Ω, and use $\|\tilde{V}\|_\infty$ otherwise. This notation should not be confused with $\|\tilde{V}\|_{\infty,p}$ defined in Section 3.1.

To show that applying the Bellman operator $\tilde{T}^\pi$ on a function $\tilde{V}$ in a smoothness class does not take the function outside the class, we study the operator's effect on the smoothness norm, i.e., $\|\tilde{T}^\pi \tilde{V}\|_{\mathcal{C}^s}$. The next intermediate result studies the the effect of taking the i-th derivative of $\mathcal{P}^\pi(\cdot|x)\tilde{V}(\gamma\omega; \cdot)$ w.r.t. the frequency parameter ω, and upper bounds it by $\gamma^i \|\tilde{V}^{(i)}\|_{\infty(\gamma\Omega),\infty}$

Proposition 14. *Let Ω be an open interval in $\mathbb{R}$ and $i \in \mathbb{N}_0$ an integer number. Assume that $\tilde{V}(\cdot; x)$ is i-times differentiable for all $x \in \mathcal{X}$. For any $x \in \mathcal{X}$, we have*

$$\sup_{\omega \in \Omega} \left|\frac{\mathrm{d}^i}{\mathrm{d}\omega^i} \int \mathcal{P}^\pi(\mathrm{d}y|x)\tilde{V}(\gamma\omega; y)\right| \leq \gamma^i \left\|\tilde{V}^{(i)}\right\|_{\infty(\gamma\Omega),\infty}.$$

Proof. Note that by the chain rule, we have $\frac{\mathrm{d}\tilde{V}(\gamma\omega;x)}{\mathrm{d}\omega} = \gamma\frac{\mathrm{d}\tilde{V}(u;x)}{\mathrm{d}u}|_{u=\gamma\omega}$. By the repeated application of the chain rule, we get that $\frac{\mathrm{d}^i\tilde{V}(\gamma\omega;x)}{\mathrm{d}\omega^i} = \gamma^i\frac{\mathrm{d}^i\tilde{V}(u;x)}{\mathrm{d}u^i}|_{u=\gamma\omega}$. We have

$$\begin{aligned}
\sup_{\omega\in\Omega}\left|\frac{\mathrm{d}^i}{\mathrm{d}\omega^i}\int\mathcal{P}^\pi(\mathrm{d}y|x)\tilde{V}(\gamma\omega;y)\right| &= \sup_{\omega\in\Omega}\left|\int\mathcal{P}^\pi(\mathrm{d}y|x)\frac{\mathrm{d}^i\tilde{V}(\gamma\omega;y)}{\mathrm{d}\omega^i}\right| \\
&= \sup_{\omega\in\Omega}\left|\gamma^i\int\mathcal{P}^\pi(\mathrm{d}y|x)\frac{\mathrm{d}^i\tilde{V}(u;y)}{\mathrm{d}u^i}\Big|_{u=\gamma\omega}\right| \\
&\le \gamma^i\int\mathcal{P}^\pi(\mathrm{d}y|x)\sup_{u\in\gamma\Omega}\left|\frac{\mathrm{d}^i\tilde{V}(u;y)}{\mathrm{d}u^i}\right| \\
&\le \gamma^i\left\|\tilde{V}^{(i)}\right\|_{\infty(\gamma\Omega),\infty}.
\end{aligned}$$

□

The next result shows the effect of applying the Bellman operator on the smoothness norm of a CVF.

Proposition 15. *Consider Ω to be an open interval in $\mathbb{R}$. Suppose that $\tilde{V}(\cdot;x)\in\mathcal{C}^s(\Omega)$ for all $x\in\mathcal{X}$. Assume that there exist finite constants $m_0,\dots,m_s$ such that the absolute moments of the rewards satisfy $\mathbb{E}\left[|R(x)|^i\right]\le m_i^i$ for all $x\in\mathcal{X}$. Let us denote $m=\max_{i=0,\dots,s}m_i$. We then have*

$$\left\|\tilde{T}^\pi\tilde{V}\right\|_{\mathcal{C}^s(\Omega)}\le s(m+\gamma)^s\left\|\tilde{V}\right\|_{\mathcal{C}^s(\gamma\Omega)}.$$

Proof. Consider the Bellman operator applied to $\tilde{V}$, which is $(\tilde{T}^\pi\tilde{V})(\omega;x) = \tilde{R}(\omega;x)\mathcal{P}^\pi(\cdot|x)\tilde{V}(\gamma\omega;\cdot)$. To simplify the notation, we denote $\tilde{h}(\omega;x) = (\tilde{T}^\pi\tilde{V})(\omega;x)$ and $\tilde{g}(\omega;x)=\mathcal{P}^\pi(\cdot|x)\tilde{V}(\gamma\omega;\cdot)$, so $\tilde{h}(\omega;x)=\tilde{R}(\omega;x)\tilde{g}(\omega;x)$. For any $k=0,\dots,s$, by the Leibniz product rule we have

$$\tilde{h}^{(k)}(\omega;x)=\sum_{i=0}^{k}\binom{k}{i}\tilde{R}^{(i)}(\omega;x)\tilde{g}^{(k-i)}(\omega;x).$$

We take the supremum over $\omega\in\Omega$ of both sides, and use Propositions 13 and 14 to get

$$\begin{aligned}
\left\|\tilde{h}^{(k)}(\cdot;x)\right\|_{\infty(\Omega)} &\le \sum_{i=0}^{k}\binom{k}{i}\left\|\tilde{R}^{(i)}(\cdot;x)\right\|_{\infty(\Omega)}\left\|\tilde{g}^{(k-i)}(\cdot;x)\right\|_{\infty(\Omega)} \\
&\le \sum_{i=0}^{k}\binom{k}{i}\left\|\tilde{R}^{(i)}(\cdot;x)\right\|_{\infty(\mathbb{R})}\left\|\tilde{g}^{(k-i)}(\cdot;x)\right\|_{\infty(\Omega)} \\
&\le \sum_{i=0}^{k}\binom{k}{i}m_i^i\gamma^{k-i}\left\|\tilde{V}^{(k-i)}(\cdot;x)\right\|_{\infty(\gamma\Omega)}.
\end{aligned}$$

We use $m_i\le m$ and $\|\tilde{V}^{(i)}(\cdot;x)\|_{\infty(\gamma\Omega)}\le\|\tilde{V}(\cdot;x)\|_{\mathcal{C}^k(\gamma\Omega)}$ for any $i\le k$ to get that

$$\left\|\tilde{h}^{(k)}(\cdot;x)\right\|_{\infty(\Omega)}\le\left\|\tilde{V}(\cdot;x)\right\|_{\mathcal{C}^k(\gamma\Omega)}\sum_{i=0}^{k}\binom{k}{i}m^i\gamma^{k-i}=\left\|\tilde{V}(\cdot;x)\right\|_{\mathcal{C}^k(\gamma\Omega)}(m+\gamma)^k,$$

where the last equality is because of the binomial theorem. As $m\ge1$ and $\|\tilde{h}(\cdot;x)\|_{\mathcal{C}^k}\le\|\tilde{h}(\cdot;x)\|_{\mathcal{C}^s}$ (for any $k\le s$), we have

$$\begin{aligned}
\left\|\tilde{h}(\cdot;x)\right\|_{\mathcal{C}^s(\Omega)}=\sum_{k=0}^{s}\left\|\tilde{h}^{(k)}(\cdot;x)\right\|_{\infty(\Omega)} &\le \left\|\tilde{V}(\cdot;x)\right\|_{\mathcal{C}^s(\gamma\Omega)}\sum_{k=0}^{s}(m+\gamma)^k \\
&\le s(m+\gamma)^s\left\|\tilde{V}(\cdot;x)\right\|_{\mathcal{C}^s(\gamma\Omega)}.
\end{aligned}$$

Taking the supremum over $x\in\mathcal{X}$ from both sides leads to the stated result. □

This result shows that after applying the Bellman operator to a CVF $\tilde{V}$ that has a finite smoothness norm $\|\tilde{V}\|_{\mathcal{C}^s(\Omega)}$, its smoothness norm $\|\tilde{T}^\pi \tilde{V}\|_{\mathcal{C}^s(\Omega)}$ remains finite. The upper bound shows that the smoothness norm might expand by a factor that depends on the absolute moments of the reward distribution and the smoothness degrees s.

We can now show a result similar to Theorem 12, but for when we choose the current and the next iteration's VCF from $\mathcal{F}^s_{b,r}$.

Theorem 16. *Let $\mathcal{R}$ be a β-smooth distribution. Assume that the reward distribution has s finite absolute moments satisfying $\max_{i=0,\dots,s} \mathbb{E}\left[|R(x)|^i\right] \leq m^i$ for all $x \in \mathcal{X}$. Consider function space $\mathcal{F}^s_{b,r}$ with $b \geq \omega_0$ (cf. Definition 2). We have*

$$\sup_{\tilde{V}' \in \mathcal{F}^s_{b,r}} \inf_{\tilde{V} \in \mathcal{F}^s_{b,s(m+\gamma)^s r}} \left\| \tilde{V} - \tilde{T}^\pi \tilde{V}' \right\|_{\infty,p} \leq \frac{c_1}{b^{p+\beta}},$$

$$\inf_{\tilde{V} \in \mathcal{F}^s_{b,sm^s}} \left\| \tilde{V} - \tilde{R} \right\|_{\infty,p} \leq \frac{c_1}{b^{p+\beta}}.$$

Proof. Let $\Omega = (-b,b)$. For any $\tilde{V}' \in \mathcal{F}^s_{b,r}$, consider $(\tilde{T}^\pi \tilde{V}')(\omega;x) = \tilde{R}(\omega;x) \int \mathcal{P}^\pi(\mathrm{d}y|x) \tilde{V}'(\gamma\omega;y)$. Because of the frequency shrinkage $\gamma\omega$ term in $\tilde{V}'(\gamma\omega;y)$, the bandwidth of $\tilde{T}^\pi \tilde{V}'$ is at most $\frac{b}{\gamma}$, i.e., it is zero outside $\gamma^{-1}\Omega$. Proposition 15 shows that

$$\left\| \tilde{T}^\pi \tilde{V}' \right\|_{\mathcal{C}^s(\frac{\Omega}{\gamma})} \leq s(m+\gamma)^s \left\| \tilde{V}' \right\|_{\mathcal{C}^s(\Omega)} \leq s(m+\gamma)^s r. \tag{39}$$

We truncate the high frequency terms of $\tilde{T}^\pi \tilde{V}'$ in order to have a function $\tilde{V}$ that has a bandwidth of b. We define $\tilde{V}(\omega;x) = (\tilde{T}^\pi \tilde{V}')(\omega;x)\mathbb{I}\{\omega \in \Omega\}$. This function has a bandwidth of b by construction. Moreover, its smoothness norm satisfies

$$\left\| \tilde{V} \right\|_{\mathcal{C}^s(\Omega)} = \left\| \tilde{T}^\pi \tilde{V}' \right\|_{\mathcal{C}^s(\Omega)} \leq \left\| \tilde{T}^\pi \tilde{V}' \right\|_{\mathcal{C}^s(\frac{\Omega}{\gamma})} \leq s(m+\gamma)^s r,$$

where the last inequality is due to (39). Therefore, the function $\tilde{V}$ belongs to $\mathcal{F}^s_{b,r'}$ with $r' = s(m+\gamma)^s r$.

The error function $\tilde{\varepsilon} = \tilde{V} - \tilde{T}^\pi \tilde{V}'$ is $\tilde{\varepsilon}(\omega;x) = (\tilde{T}^\pi \tilde{V}')(\omega;x)\mathbb{I}\{\omega \notin \Omega\}$. As $|\tilde{V}'(\omega;x)| \leq 1$ for all x and ω, we have

$$|\tilde{\varepsilon}(\omega;x)| = \left| \mathbb{I}\{\omega \notin \Omega\} \tilde{R}(\omega;x) \int \mathcal{P}^\pi(\mathrm{d}y|x) \tilde{V}^\pi(\gamma\omega;y) \right| \leq \mathbb{I}\{\omega \notin \Omega\} \left| \tilde{R}(\omega;x) \right|.$$

The norm of $\tilde{\varepsilon}$ can be upper bounded by

$$\|\tilde{\varepsilon}\|_{\infty,p} = \sup_x \sup_{\omega \in \mathbb{R}-\Omega} \left| \frac{\tilde{R}(\omega;x)}{\omega^p} \right| \leq \sup_{|\omega| \geq b} \frac{c_1}{|\omega|^{p+\beta}} = \frac{c_1}{b^{p+\beta}}. \tag{40}$$

As this holds for any $\tilde{V}' \in \mathcal{F}^s_{b,r}$, by taking the supremum over $\mathcal{F}^s_{b,r}$, we obtain the first statement.

The proof of the second statement is similar. We choose $\tilde{V}(\omega;x) = \tilde{R}(\omega;x)\mathbb{I}\{\omega \in \Omega\}$. By construction, its bandwidth is b. By the assumption on the absolute moments, Proposition 13 indicates that $\tilde{R}^{(i)}(\omega;x) \leq m_i^i \leq m^i$ for all $i = 0,\dots,s$, all frequencies $\omega \in \mathbb{R}$, and all states $x \in \mathcal{X}$. Therefore,

$$\left\| \tilde{V} \right\|_{\mathcal{C}^s(\Omega)} = \left\| \tilde{R} \right\|_{\mathcal{C}^s(\Omega)} \leq \left\| \tilde{R} \right\|_{\mathcal{C}^s(\mathbb{R})} = \sum_{i=0}^{s} \left\| \tilde{R}^{(i)} \right\|_\infty \leq \sum_{i=0}^{s} m^i \leq sm^s,$$

where we used $m \geq 1$ in the last inequality. This shows that $\tilde{V}$ is in $\mathcal{F}^s_{b,sm^s}$. By computing the norm of $\tilde{\varepsilon} = \tilde{R}(\omega;x)\mathbb{I}\{\omega \notin \Omega\}$, as in (40), we obtain the second statement. $\square$

This theorem shows that for any function $\tilde{V}$ that belongs to $\mathcal{F}^s_{b,r}$, we can find an approximation in a slightly larger function space $\mathcal{F}^s_{b,s(m+\gamma)^s r}$. The approximation error depends on the β-smoothness of the reward distribution, the bandwidth b of functions represented by the function space, and the p parameter used in the definition of distance $\|\cdot\|_{\infty,p}$, and it behaves like $O(b^{-(p+\beta)})$. A similar result would hold for super smooth reward distributions, but we omit it here.

This result is comparable to Theorem 12. The main difference is that the approximation space is $\mathcal{F}^s_{b,r}$ (38) instead of $\mathcal{F}_b$ (34), and the target function is limited to $\mathcal{F}^s_{b,r}$ instead of any $\mathcal{V}$. The difference between $\mathcal{F}^s_{b,r}$ and $\mathcal{F}_b$ is in the smoothness regularity of the former function space. The function space $\mathcal{F}_b$ does not impose any smoothness in the frequency domain, and its restriction is only on the bandwidth of the functions. As already mentioned, $\mathcal{F}_b$ is a very large function space and may not allow us to provide a convergence rate for the estimation error. The addition of the smoothness regularity leads to a well-behaving complexity of $\mathcal{F}^s_{b,r}$, which is represented by its covering number. Studying its covering number is the topic of the next section.

D Covering number of $\mathcal{F}^s_{b,r}$

We provide a covering number result for the function space $\mathcal{F}^s_{b,r}$, defined in (38). The covering number (and its logarithm, the metric entropy) is a measure of the complexity of a function space, and appears in the analysis of the estimation error [Györfi et al., 2002, van de Geer, 2000]. We restrict our analysis to finite state spaces, i.e, $|\mathcal{X}| < \infty$. Our result, Theorem 19, shows that for a finite state space $\mathcal{X}$, the metric entropy w.r.t. the supremum norm behaves as $\log \mathcal{N}(\varepsilon, \mathcal{F}^s_{b,r}, L_\infty) \leq c|\mathcal{X}|b(\frac{r}{\varepsilon})^{-1/s}$ (and similar for other L_p-norms). Interestingly, the covering number w.r.t. $d_{\infty,1}$ behaves as $\log \mathcal{N}(\varepsilon, \mathcal{F}^s_{b,r}, d_{\infty,1}) \leq |\mathcal{X}|\ s \log(\frac{2reb^{\frac{s-1}{2}}}{\varepsilon})$ (and similar w.r.t. $d_{1,1}$), which shows a quite different behaviour, i.e., a logarithmic dependence on ε as opposed to a polynomial dependence.

To prepare for the main result of this section, we define some notations and state a few auxiliary results. Consider a function $f : \Omega \to \mathcal{C}$ with $\Omega \subset \mathbb{R}$. We denote its extension to $\mathbb{R}$ by $\bar{f}$, i.e.,

$$\bar{f}(\omega) = \begin{cases} f(\omega), & |\omega| < b \\ 0. & |\omega| \geq b \end{cases} \tag{41}$$

Let us consider the set of functions belonging to $\mathcal{C}^s(\Omega)$ with a domain Ω and r-bounded $\mathcal{C}^s(\Omega)$-norm and denote it by $B^s(r;\Omega)$, i.e.,

$$B^s(r;\Omega) \triangleq \left\{ f \in C^s(\Omega) \,:\, \|f\|_{\mathcal{C}^s(\Omega)} \leq r \right\}.$$

We sometimes use $B^s_b(r)$ instead of $B^s(r;(-b,+b))$ with $b > 0$. The value of a CVF at $\omega = 0$ is equal to 1, so in order to discuss functions whose restriction to domain Ω is smooth and takes the value of 1 at $\omega = 0$, we define

$$\bar{B}^s(r;\Omega) = \left\{ \bar{f} : \mathbb{R} \to \mathbb{C} \,:\, f \in B^s(r;\Omega), f(0) = 1 \right\}.$$

We sometimes use $\bar{B}^s_b(r)$ instead of $\bar{B}^s(r;(-b,+b))$.

The following result, which is based on an already known result on the metric entropy of $\mathcal{C}^s([0,1])$, provides an upper bound for the metric entropy of $B^s_b(r;(-b,b))$.

Proposition 17. *For any $1 \leq p \leq \infty$ and $b \geq 1$, the covering number of $B^s_b(r)$ satisfies*

$$\log \mathcal{N}\left(\varepsilon, B^s_b(r), L_p((-b,b))\right) \leq c b^{1+\frac{1}{sp}} \left(\frac{r}{\varepsilon}\right)^{\frac{1}{s}},$$

for a constant $c > 0$ that depends only on s and p.

Proof. Consider $f \in B^s(r;\Omega)$ with $\Omega = (-b,b)$. The domain Ω can be partitioned into $\lceil 2b \rceil$ intervals Ω_i with length of 1 (in the form of $(-b, b+1], (b+1, b+2], \dots$) such that $\Omega \subset \cup_i \Omega_i$. We denote the restriction of f on each interval by f_i, i.e., $f_i(\omega) = f(\omega)\mathbb{I}\{\omega \in \Omega_i\}$. The $\mathcal{C}^s$-norm of f_i over Ω_i is less than or equal to that of f over Ω, i.e., $\|f_i\|_{\mathcal{C}^s(\Omega_i)} \leq \|f\|_{\mathcal{C}^s(\Omega)}$. Therefore, each f_i, after a translation, belongs to $B^s(r;(0,1]) \subset B^s(r;[0,1])$.

The covering of $B^s(r;(0,1])$ for each $i = 1, 2, \dots, 2b$ induces a covering of $B^s(r;\Omega)$. To see this, suppose that N_ε is an ε-covering set of $B^s(r;(0,1])$ w.r.t. $L_p((0,1])$. For any function $f \in B^s(r;\Omega)$, we can write it as $f(\omega) = \sum_{i=1}^{2b} f_i(\omega)$. For each i, pick $f'_i \in N_\varepsilon$ (after a shift of domain so that Ω_i aligns with $(0, 1])$ so that $\int |f_i(\omega) - f'_i(\omega)|^p \mathrm{d}\omega \leq \varepsilon^p$ (for $1 \leq p < \infty$) or $\|f_i - f'_i\|_\infty \leq \varepsilon$ (for $p = \infty$). We construct $f'(\omega) = \sum_{i=1}^{2b} f'_i(\omega)$. The $L_p(\Omega)$-norm of the difference between f' and f is $\sqrt[p]{\sum_{i=1}^{2b} \int |f'_i(\omega) - f_i(\omega)|^p \mathrm{d}\omega} \leq \sqrt[p]{2b}\varepsilon$ (for $1 \leq p < \infty$) or $\| \sum_{i=1}^{2b}(f'_i - f_i)\|_\infty = \max_{i=1,\dots,2b} \|f'_i - f_i\|_\infty \leq \varepsilon$ (for $p = \infty$). This shows that we can construct an $\sqrt[p]{2b}\varepsilon$-covering (for $1 \leq p < \infty$) or an ε-covering (for $p = \infty$) of $B^s(r; \Omega)$ based on ε-covering $B^s(r; (0, 1])$. The number of choices for each f'_i is $|N_\varepsilon|$, so the number of functions to cover $B^s_b(r)$ is upper bounded by $|N_\varepsilon|^{\lceil 2b \rceil}$. As the covering of $B^s(r; [0, 1])$ implies a covering on $B^s(r; (0, 1])$, we get that

$$\log \mathcal{N}\left(\varepsilon, B^s(r; \Omega), L_p(\Omega)\right) \leq \lceil 2b \rceil \log \mathcal{N}\left(\varepsilon', B^s(r; [0, 1]), L_p([0, 1])\right), \tag{42}$$

with $\varepsilon' = \frac{\varepsilon}{\sqrt[p]{2b}}$ for $1 \leq p < \infty$ and $\varepsilon' = \varepsilon$ for $p = \infty$.

Corollary 4.3.38 of Giné and Nickl [2015] shows that

$$\log \mathcal{N}\left(\varepsilon, B^s(r; [0, 1]), L_p([0, 1])\right) \leq c \left(\frac{r}{\varepsilon}\right)^{1/s},$$

for some constant c that depends only on s and p.[8] This along with (42) finish the proof. □

This result can be generalized to other function spaces. There are two points in the proof where the properties of $\mathcal{C}^s$ are used. The first is that the norm of a function is greater or equal to the norm of that function over a restricted domain. This seems to be true for many other norms. The second is the covering number of $B^s(r; [0, 1])$. The same inequality holds for more general function spaces, including Sobolev space $\mathbb{W}^{s,2}([0, 1])$ and some Besov spaces with the same order of smoothness s, cf. Theorem 4.3.36 and Corollary 4.3.38 of Giné and Nickl [2015].

We can also provide a covering number result for $\bar{B}^s_b(r)$ w.r.t. $d_{\infty,1}$ and $d_{1,1}$, instead of the supremum norm for $B^s_b(r)$ in the previous proposition.

Proposition 18. *For any $s \geq 1$ and for any $0 \leq \varepsilon < r$, we have*

$$\log \mathcal{N}\left(\varepsilon, \bar{B}^s(r; [-b, +b]), d_{\infty,1}\right) \leq s \log\left(\frac{2erb^{\frac{s-1}{2}}}{\varepsilon}\right),$$

$$\log \mathcal{N}\left(\varepsilon, \bar{B}^s(r; [-b, +b]), d_{1,1}\right) \leq s \log\left(\frac{4ersb^{\frac{s+1}{2}}}{\varepsilon}\right).$$

Proof. Let $\Omega = (-b, +b)$. For any $f \in \bar{B}^s(r; \Omega)$, by the Taylor series expansion around $\omega = 0$, we have that for $\omega \in \Omega$,

$$f(\omega) = 1 + \sum_{i=1}^{s-1} f^{(i)}(0)\frac{w^i}{i!} + f^{(s)}(u)\Big|_{0<u<\omega} \frac{\omega^s}{s!},$$

for some $u \in (0, \omega)$. Here without loss of generality we supposed that ω is non-negative (otherwise, we could write $\omega < u < 0$). As $f \in \bar{B}^s(r; \Omega)$, its i-th derivative $f^{(i)}$ is uniformly bounded by r on Ω for any $i = 0, \dots, s$. So we can discretize the interval $[-r, +r]$ and approximate the value of the $f^{(i)}$ terms by the quantized value.

Let us discretize the interval $[-r, +r]$ with resolutions $\varepsilon_1, \dots, \varepsilon_s$, to be determined, and call the resulting sets $U_1, \dots U_s$, i.e., $U_i = \{-r, -r + \varepsilon_i, -r + 2\varepsilon_i, \dots, r - \varepsilon_i\}$. The set U_i has $N_i = |U_i| = \frac{2r}{\varepsilon_i}$ elements. We construct $f'(\omega)$ (for $\omega \in \Omega$) as

$$f'(\omega) = 1 + \sum_{i=1}^{s} a_i \frac{\omega^i}{i!},$$

[8] The definition of the $\mathcal{C}^s$-norm of a function f by Giné and Nickl [2015] is $\|f\|_\infty + \|f^{(s)}\|$, which is upper bounded by our definition (37) (see Section 4.3.3 of their book). Therefore, the function space $B^s_b(r; \Omega)$ is a subset of the function space defined with their norm, which is $\{f \in C^s(\Omega) : \|f\|_\infty + \|f^{(s)}\| \leq r\}$. So their covering number is an upper bound on the covering number of the function space we are interested in.

with $a_i \in U_i$ (for $i = 1, \dots, s$) being selected so that $|a_i - f^{(i)}(0)| \leq \varepsilon_i$ (for $i = 1, \dots, s-1$) and $|a_s - f^{(s)}(u)| \leq \varepsilon_s$, which exist by the construction of U_i. We use $\bar{f}'$ as the extension of f' from Ω to $\mathbb{R}$.

Both functions f' and $\bar{f}'$ can be identified by an element of the product set $U_1 \times \cdots \times U_s$, so the number of distinct f' and $\bar{f}'$ constructed this way is the number of elements in the product set $|U_1 \times \cdots \times U_s| = N_1 \times \cdots \times N_s$. We can provide an explicit number on the size of this set as soon as we decide on $\varepsilon_1, \dots, \varepsilon_s$.

Let us verify that this set provides a covering for $\bar{B}^s(r; \Omega)$ w.r.t. $d_{\infty,1}$ and $d_{1,1}$, with a resolution that depends on $\varepsilon_1, \dots, \varepsilon_s$. The difference between f and $\bar{f}'$ at any $\omega \in \Omega$ is

$$f(\omega) - \bar{f}'(\omega) = \sum_{i=1}^{s-1} (f^{(i)}(0) - a_i) \frac{w^i}{i!} + \left(f^{(s)}(u)\Big|_{0<u<\omega} - a_s \right) \frac{\omega^s}{s!}.$$

As the value of f and $\bar{f}'$ outside Ω is zero, we can decompose the $d_{\infty,1}(f, \bar{f}')$ distance as follows:

$$\begin{aligned} d_{\infty,1}(f, \bar{f}') &= \sup_{\omega \in \mathbb{R}} \left| \frac{f(\omega) - \bar{f}'(\omega)}{\omega} \right| = \sup_{\omega \in \Omega} \left| \frac{f(\omega) - f'(\omega)}{\omega} \right| + \sup_{\omega \in \mathbb{R} - \Omega} \left| \frac{0-0}{\omega} \right| \\ &= \sup_{\omega \in \Omega} \left| \sum_{i=1}^{s-1} (f^{(i)}(0) - a_i) \frac{w^{i-1}}{i!} + \left(f^{(s)}(u)\Big|_{0<u<\omega} - a_s \right) \frac{\omega^{s-1}}{s!} \right| \\ &\leq \sum_{i=1}^{s} \varepsilon_i \frac{b^{i-1}}{i!}. \end{aligned}$$

Likewise, we can decompose the $d_{1,1}(f, \bar{f}')$ distance as follows:

$$\begin{aligned} d_{\infty,1}(f, \bar{f}') &= \int_{-\infty}^{+\infty} \left| \frac{f(\omega) - \bar{f}'(\omega)}{\omega} \right| \mathrm{d}\omega = \int_{-b}^{+b} \left| \frac{f(\omega) - \bar{f}'(\omega)}{\omega} \right| \mathrm{d}\omega + \int_{\mathbb{R}-\Omega} \left| \frac{0-0}{\omega} \right| \mathrm{d}\omega \\ &= \int_{-b}^{+b} \left| \sum_{i=1}^{s-1} (f^{(i)}(0) - a_i) \frac{w^{i-1}}{i!} + \left(f^{(s)}(u)\Big|_{0<u<\omega} - a_s \right) \frac{\omega^{s-1}}{s!} \right| \mathrm{d}\omega \\ &\leq \sum_{i=1}^{s} \frac{\varepsilon_i}{i!} \int_{-b}^{+b} |\omega|^{i-1} \mathrm{d}\omega \\ &\leq \sum_{i=1}^{s} \varepsilon_i \frac{2b^i}{i \times i!}. \end{aligned}$$

By choosing $\varepsilon_i = \frac{\varepsilon}{eb^{i-1}}$ (for the $d_{\infty,1}$ case) and $\varepsilon_i = \frac{\varepsilon}{2eib^i}$ (for the $d_{1,1}$ case), we obtain that

$$d_{\infty,1}(f, \bar{f}'), d_{1,1}(f, \bar{f}') \leq \frac{\varepsilon}{e} \sum_{i=1}^{s} \frac{1}{i!} \leq \varepsilon.$$

This shows that $\bar{f}'$ provides an ε-covering for $\bar{B}^s(r; \Omega)$ w.r.t. $d_{\infty,1}$ and $d_{1,1}$. To count the number of elements in this covering set, note that for the $d_{\infty,1}$ case, we have $N_i = |U_i| = \frac{2r}{\varepsilon_i} = \frac{2erb^{i-1}}{\varepsilon}$, and as a result, the total number of elements of the product set $U_1 \times \cdots \times U_s$ is

$$N_1 \times \cdots \times N_s = \left(\frac{2er}{\varepsilon} \right)^s b^{\frac{s(s-1)}{2}}.$$

Similarly, for the $d_{1,1}$ case, we have $N_i = |U_i| = \frac{2r}{\varepsilon_i} = \frac{4erb^i}{\varepsilon}$, and

$$N_1 \times \cdots \times N_s = s! \left(\frac{4er}{\varepsilon} \right)^s b^{\frac{s(s+1)}{2}}.$$

Taking the logarithm of both sides provides the desired result. □

To provide a covering number for $\mathcal{F}^s_{b,r}$, we have to specify a distance between functions in $\mathcal{F}^s_{b,r}$. We provide results for two different types of distances. The first is for L_p-based norms, and the other is for $d_{\infty,1}$ and $d_{1,1}$ (10).

Given two VCF $\tilde{V}_1, \tilde{V}_2 \in \tilde{V}$, and $p, q \in [1, \infty]$, we define

$$\left\|\tilde{V}_1 - \tilde{V}_2\right\|_{L_{q,p}} = \begin{cases} \sqrt[q]{\sum_{x\in\mathcal{X}} \left\|\tilde{V}_1(\cdot;x) - \tilde{V}_2(\cdot;x)\right\|_p^q}, & 1 \le q < \infty \\ \max_{x\in\mathcal{X}} \left\|\tilde{V}_1(\cdot;x) - \tilde{V}_2(\cdot;x)\right\|_p. & q = \infty \end{cases}$$

With these definition, and equipped with Propositions 17 and 18, we are ready to state and prove our result.

Theorem 19. *Consider the function space* (38). *(Part I) For any* $1 \le p \le \infty$, $q \in \{p, \infty\}$ *and* $b \ge 1$, *the covering number of* $\mathcal{F}^s_{b,r}$ *satisfies*

$$\log \mathcal{N}\left(\varepsilon, \mathcal{F}^s_{b,r}, L_{q,p}\right) \le c b^{1+\frac{1}{sp}} \left(\frac{r}{\varepsilon}\right)^{\frac{1}{s}} \times \begin{cases} |\mathcal{X}|, & q = \infty \\ |\mathcal{X}|^{1+\frac{1}{sp}}. & q = p \end{cases}$$

for a constant $c > 0$ *that depends only on* s *and* p.

(Part II) It also holds that for any $0 < \varepsilon < r$,

$$\log \mathcal{N}\left(\varepsilon, \mathcal{F}^s_{b,r}, d_{\infty,1}\right) \le |\mathcal{X}|\; s \log\left(\frac{2erb^{\frac{s-1}{2}}}{\varepsilon}\right),$$

$$\log \mathcal{N}\left(\varepsilon, \mathcal{F}^s_{b,r}, d_{1,1}\right) \le |\mathcal{X}|\; s \log\left(\frac{4ersb^{\frac{s+1}{2}}}{\varepsilon}\right).$$

Proof. We decompose a function $\tilde{V} \in \mathcal{F}^s_{b,r}$ into $|\mathcal{X}|$ functions, each of which can be constructed based on a member of $B^s_b(r)$ (first part) or $\bar{B}^s_b(r)$ (second part). We then relate the covering number of $\mathcal{F}^s_{b,r}$ to the covering of $B^s_b(r)$ or $\bar{B}^s_b(r)$.

We let $\Omega = (-b, b)$. Recall that given a function $f : \Omega \to \mathcal{C}$, we denote the extension of its domain to $\mathbb{R}$ by $\bar{f}$ (41). We decompose a function $\tilde{V} \in \mathcal{F}^s_{b,r}$ into $|\mathcal{X}|$-functions

$$\tilde{V}(\omega; x) = \sum_{x_i\in\mathcal{X}} \mathbb{I}\{x = x_i\} \bar{f}_{x_i}(\omega).$$

For the first part of the result, $\bar{f}_{x_i}$ is an extension of a member of a subset of $B^s_b(r)$ that is 1-bounded and is equal to 1 at $\omega = 0$, i.e., $f_x \in B^s_b(r) \cap \{\, f : \omega \to \mathcal{C} \,:\, |f(\omega)| \le 1, f(0) = 0 \,\}$ for all $x \in \mathcal{X}$. For the second part, $\bar{f}_{x_i}$ is a member of $\bar{B}^s_b(r)$.

Let us focus on the first part. Consider an ε-covering set N_ε of $B^s_b(r)$ w.r.t. $L_p(\Omega)$, which entails that for any $f \in B^s_b(r)$, we can find $f' \in N_\varepsilon$ such that $\|f - f'\|_{L_p(\Omega)} \le \varepsilon$. Consider the product set $N^\times_\varepsilon = \prod_{x_i\in\mathcal{X}} N_\varepsilon$ constructed from each of the $|\mathcal{X}|$ covering sets. Any function $\tilde{V}(\omega; x) = \sum_{x_i\in\mathcal{X}} \mathbb{I}\{x = x_i\}\bar{f}_{x_i}(\omega) \in \mathcal{F}^s_{b,r}$ can be approximated by $\tilde{V}'(\omega; x) = \sum_{x_i\in\mathcal{X}} \mathbb{I}\{x = x_i\}\bar{f}'_{x_i}(\omega)$, with $\bar{f}'_{x_i}$ being an extension of $f'_{x_i} \in N_\varepsilon$ and f'_{x_i} itself is selected to satisfy $\left\|f_{x_i} - f'_{x_i}\right\|_p \le \varepsilon$ (which exists by the definition of the covering set). The function $\tilde{V}'$ is close to $\tilde{V}$, in the $\|\cdot\|_{L_{\infty,p}}$-norm, because

$$\begin{aligned} \left\|\tilde{V} - \tilde{V}'\right\|_{L_{\infty,p}} &= \left\|\sum_{x_i\in\mathcal{X}} \mathbb{I}\{x = x_i\}\left(\bar{f}_{x_i}(\omega) - \bar{f}'_{x_i}(\omega)\right)\right\|_{L_{\infty,p}} \\ &= \max_{x\in\mathcal{X}} \left\|\bar{f}_x - \bar{f}'_x\right\|_{L_p(\mathbb{R})} \\ &= \max_{x\in\mathcal{X}} \left\|f_x - f'_x\right\|_{L_p(\Omega)} \le \varepsilon. \end{aligned} \tag{43}$$

This shows that our choice of $\tilde{V}'$ is ε-close to $\tilde{V}$ w.r.t. $L_{\infty,p}$. Likewise for $1 \le p < \infty$, we have

$$
\begin{aligned}
\left\|\tilde{V} - \tilde{V}'\right\|_{p,p} &= \left\| \sum_{x_i \in \mathcal{X}} \mathbb{I}\{x = x_i\} \left(\bar{f}_{x_i}(\omega) - \bar{f}'_{x_i}(\omega)\right) \right\|_{p,p} \\
&= \sqrt[p]{\sum_{x \in \mathcal{X}} \left\| \sum_{x_i \in \mathcal{X}} \mathbb{I}\{x = x_i\} \left(\bar{f}_{x_i} - \bar{f}'_{x_i}\right) \right\|^p_{L_p(\mathbb{R})}} \\
&= \sqrt[p]{\sum_{x \in \mathcal{X}} \|f_x - f'_x\|^p_{L_p(\Omega)}} \le \sqrt[p]{|\mathcal{X}|}\,\varepsilon. \qquad (44)
\end{aligned}
$$

This shows that our choice of $\tilde{V}'$ is $\sqrt[p]{|\mathcal{X}|}\varepsilon$-close to $\tilde{V}$ w.r.t. $L_{p,p}$.

The selected functions $(f'_{x_1}, \ldots, f'_{x_{|\mathcal{X}|}})$ belong to the product space $N_\varepsilon^\times$, which has $|N_\varepsilon|^{|\mathcal{X}|}$ members. The upper bounds (43) and (44) show that this provides an ε-covering of $\mathcal{F}^s_{b,r}$ w.r.t. $L_{\infty,p}$ and a $\sqrt[p]{|\mathcal{X}|}\varepsilon$-covering of $\mathcal{F}^s_{b,r}$ w.r.t. $L_{p,p}$.

To complete the proof of this part, we use Proposition 17, which shows that one can find an ε-covering of $B^s_b(r)$ w.r.t. $L_p(\Omega)$ with $\log |N_\varepsilon| \le c b^{1+\frac{1}{sp}} \left(\frac{r}{\varepsilon}\right)^{\frac{1}{s}}$, with an appropriate choice of ε for each case.

The proof of the second part of the result is similar too. We construct an ε-cover N_ε of $\bar{B}^s_b(r)$ w.r.t. $d_{\infty,1}$ (or $d_{\infty,1}$), which entails that for any $\bar{f} \in \bar{B}^s_b(r)$, we can find $\bar{f}' \in N_\varepsilon$ such that $d_{\infty,1}(\bar{f}, \bar{f}') \le \varepsilon$ (or $d_{1,1}(\bar{f}, \bar{f}') \le \varepsilon$).

The product set $N_\varepsilon^\times = \prod_{x_i \in \mathcal{X}} N_\varepsilon$ is an ε-covering set for $\mathcal{F}^s_{b,r}$ w.r.t. $d_{\infty,1}$ (or $d_{\infty,1}$). To see this, we first notice that any function $\tilde{V} \in \mathcal{F}^s_{b,r}$ can be written as $\tilde{V} = \sum_{x_i \in \mathcal{X}} \mathbb{I}\{x = x_i\} \bar{f}_{x_i}(\omega)$ with $\bar{f}_{x_i} \in \bar{B}^r_b(s)$. The function $\tilde{V}$ can be approximated by $\tilde{V}'(\omega; x) = \sum_{x_i \in \mathcal{X}} \mathbb{I}\{x = x_i\} \bar{f}'_{x_i}(\omega)$ with $(\bar{f}'_{x_1}, \ldots, \bar{f}'_{x_{|\mathcal{X}|}}) \in N_\varepsilon^\times$. This is because for each x_i, by the definition of the covering set, one can find a $\bar{f}'_{x_i} \in N_\varepsilon$ such that $d_{\infty,1}(\bar{f}_{x_i}, \bar{f}'_{x_i}) \le \varepsilon$ (or $d_{1,1}(\bar{f}_{x_i}, \bar{f}'_{x_i}) \le \varepsilon$) , and hence

$$
\begin{aligned}
d_{\infty,1}(\tilde{V}, \tilde{V}') &= \max_{x \in \mathcal{X}} \sup_{\omega \in \mathbb{R}} \left| \frac{\sum_{x_i \in \mathcal{X}} \mathbb{I}\{x = x_i\} \left(\bar{f}_{x_i}(\omega) - \bar{f}'_{x_i}(\omega)\right)}{\omega} \right| \\
&= \max_{x \in \mathcal{X}} d_{\infty,1}\left(\bar{f}_x, \bar{f}'_x\right) \le \varepsilon, \\
d_{1,1}(\tilde{V}, \tilde{V}') &= \max_{x \in \mathcal{X}} \int \left| \frac{\sum_{x_i \in \mathcal{X}} \mathbb{I}\{x = x_i\} \left(\bar{f}_{x_i}(\omega) - \bar{f}'_{x_i}(\omega)\right)}{\omega} \right| \\
&= \max_{x \in \mathcal{X}} d_{1,1}\left(\bar{f}_x, \bar{f}'_x\right) \le \varepsilon.
\end{aligned}
$$

The result follows by noticing that the number of elements of $N_\varepsilon^\times$ is $|N_\varepsilon|^{|\mathcal{X}|}$ and then evoking Proposition 18 to provide an upper bound on $|N_\varepsilon|$. □

The behaviour of these two different covering numbers crucially depends on the choice of the distance. For the $L_{q,p}$ distance, the behaviour as a function of ε is $O(\varepsilon^{-\frac{1}{s}})$. This is the usual behaviour of the s-times differentiable functions over a bounded subset of $\mathbb{R}$. This is not surprising as $\mathcal{F}^s_{b,r}$ is a product space of smoothness class over a bounded frequency domain defined over $|\mathcal{X}|$ states.

The covering number behaviour w.r.t. $d_{\infty,1}$ and $d_{1,1}$, however, is a different story. Its behaviour as a function of ε is not polynomial, but logarithmic $O(\log(\frac{1}{\varepsilon}))$, hence has a much slower increase as ε decreases. In other words, measured according to these distances, the function space $\mathcal{F}^s_{b,r}$ is not very complex. One intuition is that the distances $d_{\infty,1}(\tilde{V}_1, \tilde{V}_2) = \sup_x \sup_\omega |\frac{\tilde{V}_1(\omega;x) - \tilde{V}_2(\omega;x)}{\omega}|$ and $d_{1,1}(\tilde{V}_1, \tilde{V}_2) = \sup_x \int |\frac{\tilde{V}_1(\omega;x) - \tilde{V}_2(\omega;x)}{\omega}| \mathrm{d}\omega$ between two VCFs $\tilde{V}_1$ and $\tilde{V}_2$ are less sensitive to high frequency differences between them as the difference is dampened by the ω in the denominator.

This result alone does not necessarily imply a faster convergence rate for the estimation error term of solving the regression problem (18). One has to study the estimation error more closely to see whether or not the covering number appearing in the analysis is in fact w.r.t. $d_{\infty,1}$ or $d_{1,1}$ (or $L_{q,p}$ or some other distance). It is in fact likely that the choice of the covering number depends on the choice of $w(\omega)$ in (18). This is a topic of future research.

Finally, we remark that the extension of this result to more general state spaces such as a subset of $\mathbb{R}^d$ requires making assumptions not only on the reward distribution $\mathcal{R}^\pi(\cdot|x)$ at each state x (which we have already done by assuming that it is β-smooth and has certain moment conditions), but also on how the reward distribution changes as a function of state, e.g., it belongs to a certain smoothness class. This is similar to assumptions required in the analysis of conventional RL methods for large state spaces, e.g., smoothness of the value function [Farahmand, 2011, Farahmand et al., 2016]. That type of analysis is possible to extend to our case too, but we postpone it to a future study.

Acknowledgments

I would like to thank the anonymous reviewers for their helpful feedback, particularly Reviewer #4. I acknowledge the funding from the Canada CIFAR AI Chairs program.

References

Benjamin Arras, Guillaume Mijoule, Guillaume Poly, and Yvik Swan. A new approach to the Stein-Tikhomirov method: with applications to the second Wiener chaos and Dickman convergence. *arXiv:1605.06819v2*, 2017. 14, 16

Gabriel Barth-Maron, Matthew W. Hoffman, David Budden, Will Dabney, Dan Horgan, Dhruva TB, Alistair Muldal, Nicolas Heess, and Timothy Lillicrap. Distributional policy gradients. In *International Conference on Learning Representations (ICLR)*, 2018. 1

Marc G. Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017. 1

Dimitri P. Bertsekas. *Abstract dynamic programming*. Athena Scientific Belmont, 2013. 2

Dimitri P. Bertsekas and Steven E. Shreve. *Stochastic Optimal Control: The Discrete-Time Case*. Academic Press, 1978. 2

José Antonio Carrillo and Giuseppe Toscani. Contractive probability metrics and asymptotic behavior of dissipative kinetic equations (notes of the Porto Ercole school, june 2006). *Riv. Mat. Univ. Parma*, 7(6):75–198, 2007. 5

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019. 8, 9

Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *Journal of Machine Learning Research (JMLR)*, 18(167):1–51, 2018. 1

Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018. 1

Yaakov Engel, Shie Mannor, and Ron Meir. Reinforcement learning with Gaussian processes. In *Proceedings of the 22nd International Conference on Machine learning (ICML)*, pages 201–208. ACM, 2005. 1

Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research (JMLR)*, 6:503–556, 2005. 8, 9

Jianqing Fan. On the optimal rates of convergence for nonparametric deconvolution problems. *Annals of Statistics*, 19(3):1257–1272, 09 1991. 21

Amir-massoud Farahmand. *Regularization in Reinforcement Learning*. PhD thesis, University of Alberta, 2011. 31

Amir-massoud Farahmand, Mohammad Ghavamzadeh, Csaba Szepesvári, and Shie Mannor. Regularized fitted Q-iteration for planning in continuous-space Markovian Decision Problems. In *Proceedings of American Control Conference (ACC)*, pages 725–730, June 2009. 8, 9

Amir-massoud Farahmand, Rémi Munos, and Csaba Szepesvári. Error propagation for approximate policy and value iteration. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems (NeurIPS - 23)*, pages 568–576. 2010. 14

Amir-massoud Farahmand, Mohammad Ghavamzadeh, Csaba Szepesvári, and Shie Mannor. Regularized policy iteration with nonparametric function spaces. *Journal of Machine Learning Research (JMLR)*, 17(139):1–66, 2016. 31

Andrey Feuerverger. An efficiency result for the empirical characteristic function in stationary time-series models. *Canadian Journal of Statistics*, 18(2):155–161, 1990. 2

Andrey Feuerverger and Philip McDunnough. On some Fourier methods for inference. *Journal of the American Statistical Association*, 76(374):379–387, 1981. 2

Andrey Feuerverger and Roman A. Mureika. The empirical characteristic function and its applications. *Annals of Statistics*, 5(1):88–97, 01 1977. 2

Javier García and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research (JMLR)*, 16:1437–1480, 2015. 1

Alison L. Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, 2002. 16

Evarist Giné and Richard Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press, 2015. 27

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. 9

László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer Verlag, New York, 2002. 9, 22, 26

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd edition)*. Springer, 2009. 9

John K. Hunter and Bruno Nachtergaele. *Applied analysis*. World Scientific Publishing Company, 2001. 7, 19

John L. Knight and Jun Yu. Empirical characteristic function in time series estimation. *Econometric Theory*, 18(3):691–721, 2002. 2

Clare Lyle, Pablo Samuel Castro, and Marc G. Bellemare. A comparative analysis of expected and distributional reinforcement learning. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, 2019. 1

Tetsuro Morimura, Masashi Sugiyama, Hisashi Kashima, Hirotaka Hachiya, and Toshiyuki Tanaka. Nonparametric return distribution approximation for reinforcement learning. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, 2010a. 1

Tetsuro Morimura, Masashi Sugiyama, Hisashi Kashima, Hirotaka Hachiya, and Toshiyuki Tanaka. Parametric return density estimation for reinforcement learning. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 368–375, 2010b. 1

Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997. 16

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research (JMLR)*, 9:815–857, 2008. 8, 9

L.A. Prashanth and Mohammad Ghavamzadeh. Actor-critic algorithms for risk-sensitive mdps. In *Advances in Neural Information Processing Systems (NIPS - 26)*. 2013. 1

Jeffrey S. Rosenthal. *A Fist Look at Rigorous Probability Theory*. World Scientific Publishing, 2nd edition, 2006. 17

Mark Rowland, Marc G. Bellemare, Will Dabney, Rémi Munos, and Yee Whye Teh. An analysis of categorical distributional reinforcement learning. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018. 1, 3

Mark Rowland, Robert Dadashi, Saurabh Kumar, Rémi Munos, Marc G. Bellemare, and Will Dabney. Statistics and samples in distributional reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019. 1

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 01 2016. 8, 9

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2019. 1

Csaba Szepesvári. *Algorithms for Reinforcement Learning*. Morgan Claypool Publishers, 2010. 2

Aviv Tamar, Dotan Di Castro, and Shie Mannor. Policy gradients with variance related risk criteria. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, 2012. 1

Samuele Tosatto, Matteo Pirotta, Carlo D'Eramo, and Marcello Restelli. Boosted fitted Q-iteration. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017. 8, 9

Sara A. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000. 26

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008. 5

Larry Wasserman. *All of Nonparametric Statistics*. Springer, 2007. 9

David Williams. *Probability with Martingales*. Cambridge University Press, 1991. 3, 17

Jun Yu. Empirical characteristic function estimation and its applications. *Econometric reviews*, 23(2): 93–123, 2004. 2