# Distributional Hamilton-Jacobi-Bellman Equations for Continuous-Time Reinforcement Learning

**Harley Wiltzer** [1 2] **David Meger** [1] **Marc G. Bellemare** [2 3 4]

## Abstract

Continuous-time reinforcement learning offers an appealing formalism for describing control problems in which the passage of time is not naturally divided into discrete increments. Here we consider the problem of predicting the distribution of returns obtained by an agent interacting in a continuous-time, stochastic environment. Accurate return predictions have proven useful for determining optimal policies for risk-sensitive control, learning state representations, multiagent coordination, and more. We begin by establishing the distributional analogue of the Hamilton-Jacobi-Bellman (HJB) equation for Itô diffusions and the broader class of Feller-Dynkin processes. We then specialize this equation to the setting in which the return distribution is approximated by $N$ uniformly-weighted particles, a common design choice in distributional algorithms. Our derivation highlights additional terms due to *statistical diffusivity* which arise from the proper handling of distributions in the continuous-time setting. Based on this, we propose a tractable algorithm for approximately solving the distributional HJB based on a JKO scheme, which can be implemented in an online control algorithm. We demonstrate the effectiveness of such an algorithm in a synthetic control problem.

## 1. Introduction

In continuous-time reinforcement learning (Munos, 1997; Munos & Bourgine, 1997), the expected return or *value function* is characterized by a partial differential equation (PDE) known as the Hamilton-Jacobi-Bellman (HJB) equation (Krylov, 1980; Fleming & Soner, 2006). This equation can be solved using numerical methods (Munos, 2004), producing a policy that is optimal in the sense it maximises the expected return and avoids the error and computational costs associated with discretizing time.

This paper presents an analysis of the behavior of the *distribution* over returns in the continuous-time limit, as opposed to solely its expectation. Existing literature in *distributional* reinforcement learning (DRL) has demonstrated that modeling return distributions aids the policy learning process, even when decisions are based only on the expectations of the return distributions (Bellemare et al., 2017; Hessel et al., 2018; Rowland et al., 2019). Beyond that, statistics of the return distributions may provide useful signals for exploration (Mavrin et al., 2019) and risk-sensitive behavior (Prashanth & Ghavamzadeh, 2013; Chow & Ghavamzadeh, 2014; Tamar et al., 2015; Dabney et al., 2018a; Yang et al., 2019; Halperin, 2021; Prashanth & Fu, 2021).

**A distributional HJB equation.** We first establish the distributional analogue to the HJB equation for a broad class of continuous-time environments, when the policy is fixed (the *policy evaluation* setting). Because return distribution functions are infinite-dimensional objects (both in state and return), they are in general quite complex. However, we obtain a concise form of the distributional HJB by appealing to the notion of an infinitesimal generator (Rogers & Williams, 1994), specifically applied to the cumulative distribution function of the return distribution. This basic result extends to the expected-return control setting by obtaining an optimal policy from the usual HJB equation and subsequently solving the distributional HJB equation with this policy.

**Specialization to finitely-supported distributions.** In distributional RL, it is common to represent return distributions parametrically, for example with a finite collection of Dirac deltas. With care, this makes it possible to derive practical algorithms that find finite-memory approximations to the return distribution function. Our second contribution is to specialize the distributional HJB equation to finite collections of statistical functionals and subsequently to what Bellemare et al. (2022) call the *quantile probability representation*. The result is effectively a set of HJB equations and associated distributional constraints, one per parameter.

[1] McGill University, Montreal, Canada [2] Mila – Quebec AI Institute [3] Google Brain, Montreal, Canada [4] CIFAR Fellow. Correspondence to: Harley Wiltzer <harley.wiltzer@mail.mcgill.ca>.

*Proceedings of the 39<sup>th</sup> International Conference on Machine Learning*, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

**Finite-difference algorithm for continuous-time distributional RL.** Finally, we extend the algorithm of Munos & Bourgine (1997) for optimal control of continuous-time environments to the distributional setting. In particular, the inner loop of our algorithm involves finding distributional approximations by means of a JKO scheme previously employed by Martin et al. (2020). Effectively, this method solves the quantile HJB equation at the desired level of accuracy, without explicitly discretizing time. In a synthetic experiment, we find that our technique produces far fewer artefacts than the equivalent discrete-time method.

## 2. Setting

In this section we establish the mathematical framework that enables us to characterize the random return. To describe a continuous-time environment, we use the formalisms of Feller-Dynkin processes and Itô diffusions. This is sufficient to establish the general distributional HJB equation; to derive more practical equations, however, we must also introduce notions from statistical functional theory.

Let $\mathscr{P} = (\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathsf{P})$ be a filtered probability space. The notation $\mathcal{P}_p(A)$ refers to the space of all probability measures with bounded $p$-moments, and $\mathcal{P}(A) \equiv \mathcal{P}_1(A)$. Moreover, we denote by $\mathsf{H}_x, \mathsf{J}_x$ the Hessian and Jacobian operators, taken with respect to the $x$ variable.

### 2.1. Continuous-time reinforcement learning

We consider a continuous-time Markov decision process with a compact state space $\mathcal{X} \subseteq \mathbf{R}^d$ and a discrete action space $\mathcal{A}$. The state and action processes are respectively $(X_t)_{t\geq 0} : \mathbf{R}_+ \to \mathcal{X}$ and $(A_t)_{t\geq 0} : \mathbf{R}_+ \to \mathcal{A}$. The actions are determined by a stochastic policy $\pi : \mathcal{X} \to \mathcal{P}(\mathcal{A})$ such that $A_t \sim \pi(\cdot \mid X_t)$. For a fixed policy, the state process is assumed to be a Feller-Dynkin process (Rogers & Williams, 1994) with transition semigroup $(P_t^\pi)_{t\geq 0}$; a primer on Feller-Dynkin processes and other continuous-time objects is given in Appendix D. Finally, $r : \mathcal{X} \to \mathbf{R}$ is a bounded reward function.

When the agent exits the interior $\mathcal{O}$ of $\mathcal{X}$, we say that the process has stopped or *terminated*, and no further rewards are earned.[1] We will assume that $\mathcal{O}$ is Borel-measurable. The agent's (random) exit time $T$ from $\mathcal{O}$, expressed as

$$T = \inf\{t \in \mathbf{R}_+ : X_t \notin \mathcal{O}\}$$

is a stopping time with respect to the canonical filtration (Le Gall, 2016).

The (discounted) return $G^\pi(x)$ from state $x \in \mathcal{X}$ is the reward accumulated by following policy $\pi$ starting at state $x$, with rewards discounted exponentially in time by $\gamma \in (0,1)$:

$$G^\pi(x) \triangleq \int_0^T \gamma^t r(X_t)dt \quad X_0 = x \tag{1}$$

Since $r$ is bounded, it follows that the discounted return is also bounded, and we express the space of returns by the interval $\mathcal{R} = [V_{\min}, V_{\max}]$. The *value function* (Puterman, 2014) is the mapping $V^\pi : \mathcal{X} \to \mathcal{R}$ defined pointwise by

$$V^\pi(x) \triangleq \mathbf{E}[G^\pi(x)].$$

The distribution of $G^\pi(x)$ is denoted by the probability measure $\eta^\pi(x)$ for each $x \in \mathcal{X}$:

$$\eta^\pi(x) \triangleq \text{Law}\,(G^\pi(x)) \tag{2}$$

The mapping $\eta^\pi : \mathcal{X} \to \mathcal{P}(\mathcal{R})$ is referred to as the *return distribution function* (Bellemare et al., 2022). We equip $\mathcal{R}$ with the Borel $\sigma$-algebra $\mathscr{B}(\mathcal{R})$ using the usual topology of the reals. We overload notation and write $\eta^\pi(x, A) = (\eta^\pi(x))(A)$.

The optimal control problem seeks a policy that maximizes the expected return. An *optimal policy* $\pi^\star$ is one for which

$$V^{\pi^\star}(x) \geq V^\pi(x) \qquad \forall \pi : \mathcal{X} \to \mathcal{P}(\mathcal{A}), \ x \in \mathcal{X}$$

Because $(X_t)_{t\geq 0}$ is a Feller-Dynkin process, the value function is characterized by a partial differential equation (PDE) via the *infinitesimal generator* $\mathscr{L}$ of the process.[2] This is established via the probabilistic solutions to Kolmogorov backward PDEs (Kolmogorov, 1931; Le Gall, 2016):

**Theorem 1** (Kolmogorov Backward Equation)**.** *Let $(Y_t)_{t\geq 0} : \mathbf{R}_+ \to \mathcal{Y}$ be a Feller-Dynkin process in some space $\mathcal{Y}$, driven by an infinitesimal generator $\mathscr{L}$. Let $\mathcal{Y}^\circ \subset \mathcal{Y}$ be a measurable set with respect to the Borel $\sigma$-algebra on $\mathcal{Y}$, and let $S \in \mathbf{R}_+$ be the (random) exit time of $(\mathcal{Y}_t)_{t\geq 0}$ from $\mathcal{Y}^\circ$. It is assumed that $Y_0 \in \mathcal{Y}^\circ$ and* $\mathsf{P}(S < \infty) = 1$. *For any measurable function $\phi$ that is absolutely continuous and differentiable almost everywhere, $u(t, y) = \mathbf{E}[\phi(Y_S) \mid Y_{t\wedge S} = y]$ solves*

$$\frac{\partial u(t,y)}{\partial t} = -\mathscr{L}u(t,y) \tag{3}$$

*with the terminal condition $u(t, y) = \phi(y)$ for all $y \notin \mathcal{Y}^\circ$.*

The process $(X_t)_{t\geq 0}$ is called an *Itô diffusion* when

$$dX_t = \mu_\pi(X_t)dt + \boldsymbol{\sigma}_\pi(X_t)dB_t \tag{4}$$

[1] A state $x \in \mathcal{X} \setminus \mathcal{O}$ is said to be *terminal*.

[2] Roughly, the infinitesimal generator of a Feller-Dynkin process with transition semigroup $(P_t)_{t\geq 0}$ satisfying $X_t \sim P_t X_0$ is the operator $\mathscr{L}$ satisfying $\mathscr{L}f = \lim_{t\downarrow 0} \mathbf{E}\frac{P_t f - f}{t}$ for each sufficiently smooth function $f$ on the state space. A formal definition is given in Appendix D.

where $\mu_\pi : \mathcal{X} \to \mathbf{R}^d, \boldsymbol{\sigma}_\pi : \mathcal{X} \to \mathbf{R}^{d\times d}$ are the mean and diffusion of the stochastic dynamics of the agent controlled by the policy $\pi$, and $(B_t)_{t\geq 0}$ is a $\mathsf{P}$-Brownian motion. [3] Such processes have infinitesimal generators given by

$$\begin{aligned}\mathscr{L}\psi(x) &= \langle \nabla_x \psi(x), \mu_\pi(x)\rangle \\ &\quad + \frac{1}{2}\mathsf{Tr}\left(\boldsymbol{\sigma}_\pi(x)^\top \mathsf{H}_x \psi(x) \boldsymbol{\sigma}_\pi(x)\right)\end{aligned} \tag{5}$$

where $\mathsf{H}_x$ is the Hessian operator with respect to $x$ and $\mathsf{Tr}$ is the trace operator. Additionally, we assume that $\mu_\pi, \boldsymbol{\sigma}_\pi$ are continuous and differentiable almost everywhere, and that $\boldsymbol{\sigma}_\pi(x) \succeq 0$ for each $x \in \mathcal{X}$.

Writing $u(t, x) = \mathbf{E}\left[G^\pi(x) \mid X_t = x\right] = V^\pi(x)$, we have

$$\frac{\partial}{\partial t} V^\pi(x) = -\mathscr{L}V^\pi(x)$$

Moreover, the Bellman equation (Bellman, 1957) gives

$$V(X_t) = \sup_\pi \mathop{\mathbf{E}}_{\substack{X_{t+\Delta}\\ \sim P_{\Delta,\pi}}} \left[\int_0^\Delta \gamma^s r(X_{t+s}) ds + \gamma^\Delta V^\pi(X_{t+\Delta})\right]$$

$$\frac{\partial}{\partial t} V(X_t) = r(X_t) + \log\gamma V(X_t)$$

Substituting into the Kolmogorov backward equation yields

$$r(x) + \log\gamma V(x) + \sup_\pi \{\mathscr{L}V^\pi(x)\} = 0$$

Expanding the expression for the generator in (5) we have

$$\begin{aligned}\sup_\pi \Big\{ & r(x) + \langle \nabla V(x), \mu_\pi(x)\rangle \\ & + \frac{1}{2}\mathsf{Tr}\left(\boldsymbol{\sigma}_\pi(x)^\top \mathsf{H} V(x) \boldsymbol{\sigma}_\pi(x)\right)\Big\} \\ & + \log\gamma V(x) = 0\end{aligned} \tag{HJB}$$

which is the stochastic Hamilton-Jacobi-Bellman (HJB) equation (Fleming & Soner, 2006).

### 2.2. Distributional reinforcement learning

In *distributional* RL, we aim to learn the probability distribution $\eta^\pi(x)$ over $G^\pi(x)$ as opposed to only its expectation. A good approximation to the return distribution can be obtained by modelling particular statistics of the distribution, based on the notion of statistical functionals (Rowland et al., 2019; Bellemare et al., 2022).

**Definition 1** (Statistical functional, sketch)**.** *A statistical functional maps each probability distribution to a real number. A* sketch *is a collection of statistical functionals, equivalently a mapping* $\mathbf{s} : \mathcal{P}(\mathbf{R}) \to \mathbf{R}^N$ *that maps probability measures to ordered sets of real numbers (statistics).*

In this paper we will be interested in *quantile functionals*, which effectively invert the CDF $F_\nu$ of a measure $\nu$. For a return distribution function $\eta$, let us write $F_\eta(x, z) = \eta(x, [V_{\min}, z])$. The quantile functionals $q_\tau$ are

$$q_\tau(\eta(x)) = \inf\{z \in \mathcal{R} : F_\eta(x, z) = \tau\} \qquad \tau \in (0, 1)$$

**Definition 2** (Imputation strategy)**.** *Define a set* $\mathcal{S}_\Phi \subset \mathbf{R}^N$ *corresponding to a space of statistics. An* imputation strategy *is a mapping* $\Phi : \mathcal{S}_\Phi \to \mathcal{P}(\mathbf{R})$*. The set* $\mathcal{S}_\Phi$ *is referred to as the set of* admissible statistics *for* $\Phi$*.*

As with return distribution functions, for a vector $\vec{\mathsf{s}} \in \mathcal{S}_\Phi$ and $A \subseteq \mathbf{R}$ we write

$$\Phi(\vec{\mathsf{s}}, A) = \Phi(\vec{\mathsf{s}})(A).$$

We use imputation strategies to map statistical functional values back to distributions. In the sequel we consider the imputation strategy that maps a set of quantiles to what Bellemare et al. (2022) call a *quantile distribution*.

**Definition 3** (Quantile Distribution)**.** *Let* $\{y_k\}_{k=1}^N$ *be elements of a set* $\mathcal{Y}$*. The quantile distribution over* $\mathcal{Y}$ *with quantiles* $\{y_k\}_{k=1}^N$ *is a probability measure* $\nu$ *given by*

$$\nu(A) = \frac{1}{N}\sum_{k=1}^N \delta_{y_k}(A), \ A \in \mathscr{B}(\mathbf{R}).$$

Our aim will be to incorporate these two elements – sketch and imputation strategy – into a distributional HJB equation in order to produce a system of equations that can be approximated with standard numerical methods.

In our continuous-time formulations, we will analyze differential quantities of $\eta^\pi$ with respect to both the state space and the return space. As such, we will often find it more convenient to express return distributions $\eta^\pi(x)$ by their CDFs, which have a substantially simpler domain to differentiate over. We express these CDFs by $F_\eta : \mathcal{X} \times \mathcal{R} \to [0, 1]$, where $F_\eta(x, z) = \eta(x, [V_{\min}, z])$.

## 3. Distributional HJB Equations

We will now shift our focus to formally representing the return distribution function for an RL agent evolving continuously in time with a fixed policy. In order to do so, it will be necessary to impose some structural and regularity properties on the dynamics of the environment and on the return distributions.

**Assumption 1.** *At every state* $x \in \mathcal{X}$*, the return distribution* $\eta^\pi(x)$ *is absolutely continuous with respect to the Lebesgue measure.*

Although Assumption 1 can be violated in various MDPs, particularly when dynamics are deterministic and the reward

[3] An overview of Brownian motion is given in Appendix D.1.

function is not continuous, we note that such issues can easily be remedied in practice by adding low-variance white noise to the rewards, for example.

**Assumption 2.** *The mapping $(x, z) \mapsto F_{\eta^\pi}(x, z)$ is twice differentiable over $\mathcal{X} \times \mathcal{R}$ almost everywhere, and its second partial derivatives are continuous almost everywhere.*

All omitted proofs in the sequel will be provided in Appendix A.

### 3.1. Stochastic Return Processes

We would like to understand how estimates of the random return should evolve over time, using the machinery of stochastic calculus (Le Gall, 2016). However, a function mapping states to (random) returns cannot be progressively measurable (see Appendix C), as it requires knowledge of an entire trajectory. Our solution is to introduce an intermediate stochastic process as a "gateway" to the random return.

**Definition 4** (The Truncated Return Process)**.** *The* truncated return process *is a stochastic process $(J_t)_{t\geq 0} \in \mathbf{R}_+ \times \mathcal{X} \times \mathcal{R}$ given by*

$$J_t = (t, X_t, \overline{G}_t) \qquad \overline{G}_t = \int_0^t \gamma^s r(X_s) ds$$

The values $\overline{G}_t$ are simply the discounted rewards accumulated up to time $t$, and $\overline{G}_0 = 0$.

**Proposition 1.** *The truncated return process is a Markov process w.r.t. the canonical filtration.*

The (discounted) random return can be expressed in terms of the truncated return process. If the process $(X_t)_{t\geq 0}$ halts at the random exit time $T$, then $\overline{G}_T$ *is* the return:

$$\overline{G}_T \overset{\mathcal{L}}{=} G^\pi(x) \qquad X_0 = x, \tag{6}$$

where $\overset{\mathcal{L}}{=}$ denotes equality in distribution. It will be convenient to encapsulate this identity in a time-homogeneous manner, since we would like to evaluate return distributions at each state as opposed to only the initial state $X_0$. This is captured by the *conditional backward return process*.

**Definition 5** (Conditional Backward Return Process)**.** *Let $z \in \mathcal{R}$ be a desired target return. The* conditional backward return process $\left(\overleftarrow{G}(z)_t\right)_{t\geq 0} : \mathbf{R}_+ \to \mathcal{R}$ *is given by*

$$\overleftarrow{G}(z)_t = \gamma^{-t}(z - \overline{G}_t)$$

*Likewise, we define the joint process $\left(\overleftarrow{J}(z)_t\right)_{t\geq 0}$ where $\overleftarrow{J}(z)_t = (X_t, \overleftarrow{G}(z)_t)$.*

Unlike the truncated return process which accumulates rewards "forward in time", the conditional backward return process conditions on a given return $z$ and describes the residual discounted rewards needed to attain a return of $z$.

### 3.2. A Characterization of the Return Distributions

At this point, let us remark on the joint state-return process $(\overleftarrow{J}(z)_t)_{t\geq 0}$. Because $X_t$ is $d$-dimensional and the return is bounded in $[V_{\min}, V_{\max}]$, the joint process is effectively $(d+1)$-dimensional. Our goal is thus to derive, using the Kolmogorov backward equation, the PDE that characterizes the evolution of this joint process. The solution of this PDE is then the desired continuous-time return distribution function.

**Lemma 1.** *Let $z \in \mathcal{R}$ be a desired return, and suppose that $\left(\overleftarrow{J}(z)_t\right)_{t\geq 0}$ is a Feller-Dynkin process with infinitesimal generator $\mathscr{L}_J$. Then at each $x \in \mathcal{O}$ and $z' \in \mathcal{R}$, $\eta^\pi$ satisfies*

$$\mathscr{L}_J F_{\eta^\pi}(x, z') = 0 \tag{7}$$

We are now ready to introduce the characterization of the return distribution function in continuous time. In the remainder of the text, the notation $\iota_k$ will be used to denote the coordinate projection operators, where $\iota_k(a_1, a_2, \dots, a_k, \dots) = a_k$.

**Theorem 2** (Distributional HJB Equation for Policy Evaluation)**.** *Denote by $\mathscr{L}_X$ the infinitesimal generator of the process $(X_t)_{t\geq 0} = \left(\iota_1 \overleftarrow{J}(z)_t\right)_{t\geq 0}$. Moreover, suppose Assumptions 1 and 2 hold. Then $F_{\eta^\pi}$ satisfies*

$$(\mathscr{L}_X F_{\eta^\pi}(\cdot, z))(x) - (r(x) + z\log\gamma)\frac{\partial}{\partial z}F_{\eta^\pi}(x, z) = 0 \quad \mathsf{P}-\textit{almost surely} \tag{8}$$

Theorem 2 admits a useful corollary when the agent evolves according to an Itô diffusion.

**Corollary 1** (Policy Evaluation of Itô Diffusions)**.** *In the setting of Theorem 2, if the state process $(X_t)_{t\geq 0}$ is governed by the Itô diffusion of (4), the return distribution function $\eta^\pi$ satisfies for each $x \in \mathcal{X}$ and $z \in \mathcal{R}$,*

$$\begin{aligned} 0 = &\langle \nabla_x F_{\eta^\pi}(x, z), \mu_\pi(x)\rangle - (r(x) + z\log\gamma)\frac{\partial}{\partial z}F_{\eta^\pi}(x, z) \\ &+ \frac{1}{2}\mathsf{Tr}\left(\boldsymbol{\sigma}_\pi(x)^\top \mathsf{H}_x F_{\eta^\pi}(x, z)\boldsymbol{\sigma}_\pi(x)\right) \end{aligned} \tag{9}$$

*Proof.* This result follows directly from Theorem 2, since the infinitesimal generator $\mathscr{L}_X$ of an Itô diffusion is given by (5). □

Note that the term $\frac{\partial}{\partial z}F_{\eta^\pi}(x, \cdot)$ is the density of the return distribution at $x$. When the policy and environment dynamics are deterministic, we can relate Equation (9) to (HJB) by interpreting the derivatives using what is called the theory of distributions (an unfortunately-named class of objects that are usually not probability distributions; see Appendix G), and setting $\pi = \pi^*$.

### 3.3. Finitely-Parametrized Return Distributions

We now turn our attention to approximating the infinite-dimensional CDF $F_{\eta^\pi}(x, \cdot)$. Specifically, we consider what happens to Corollary 1 when return distributions are represented by a *statistics function* $\vec{\mathsf{s}} : \mathcal{X} \to \mathcal{S}_\Phi$, the statistical functional analogue of a value function $V : \mathcal{X} \to \mathbf{R}$. This statistics function corresponds to the values of $N$ statistical functionals, which can be transformed into a probability distribution by means of the imputation strategy $\Phi$. Consequently, we make the approximation

$$\eta^\pi(x) \approx \Phi(\vec{\mathsf{s}}(x)).$$

With this approximation, each return distribution can be represented in memory. The approximate distributional policy evaluation problem is then to determine a statistics function $\vec{\mathsf{s}}$ that satisfies the Itô Diffusion HJB. In order to derive a robust characterization of the return distribution function in the proposed manner, we will require a mild regularity condition on the imputation strategy.

**Definition 6** (Statistical Smoothness)**.** *An imputation strategy* $\Phi : \mathcal{S}_\Phi \to \mathcal{P}_p(\mathcal{R})$ *is said to be* statistically smooth *if* $\Phi(s)$ *is a tempered distribution (see Appendix G) for each* $s \in \mathcal{S}_\Phi$. *Likewise, a return distribution function* $\eta$ *is said to be statistically smooth if* $F_\eta(x, \cdot)$ *is a tempered distribution for each* $x \in \mathcal{X}$ *and* $F_\eta(\cdot, z)$ *is twice continuously differentiable almost everywhere for each* $z \in \mathcal{R}$.

**Definition 7** (Spatial Diffusivity)**.** *Let* $\Phi : \mathcal{S}_\Phi \to \mathcal{P}(\mathcal{R})$ *be a statistically smooth imputation strategy, let* $\vec{\mathsf{s}}(x)$ *be a statistics function, and suppose that* $(X_t)_{t\geq 0}$ *is governed by the Itô diffusion of* (4)*. The* spatial diffusivity *of the random return under the imputation strategy* $\Phi$ *is defined as the mapping* $\mathbf{K}^x_\Phi : \mathcal{X} \times \mathcal{R} \to \mathbf{R}^{d\times d}$ *given by*

$$\mathbf{K}^x_\Phi(x, z) = \sum_{k=1}^N \frac{\partial}{\partial \iota_k \vec{\mathsf{s}}(x)} \Phi(\vec{\mathsf{s}}(x), [V_{\min}, z]) \mathsf{H}_x \iota_k \vec{\mathsf{s}}(x)$$

*where* $\mathsf{H}_x$ *is the Hessian operator with respect to* $x$.

Spatial diffusivity relates the stochasticity of the approximate return distribution to the stochasticity of the state process. We will also identify a similar term relating the stochasticity of the return to the variability of the statistics as a result of the stochasticity in the state process.

**Definition 8** (Statistical Diffusivity)**.** *Let* $\Phi : \mathcal{S}_\Phi \to \mathcal{P}(\mathcal{R})$ *be a statistically smooth imputation strategy and* $(X_t)_{t\geq 0} : \mathbf{R}_+ \to \mathcal{X} \subset \mathbf{R}^d$ *the Itô diffusion* (4)*. The* statistical diffusivity *of the random return under the imputation strategy* $\Phi$ *is defined as the mapping* $\mathbf{K}^s_\Phi : \mathcal{X} \times \mathcal{R} \to \mathbf{R}^{d\times d}$ *given by*

$$\mathbf{K}^s_\Phi(x, z) = \mathsf{J}_x \vec{\mathsf{s}}(x)^\top \left(\mathsf{H}_{\vec{\mathsf{s}}(x)} \Phi(\vec{\mathsf{s}}(x), [V_{\min}, z])\right) \mathsf{J}_x \vec{\mathsf{s}}(x)$$

We can now characterize the return distribution function as a PDE with respect to the statistics function. Notably, this generalizes to all return distribution parameterizations that can be expressed by statistical functionals and imputation strategies, such as those employed by categorical (Bellemare et al., 2017), quantile (Dabney et al., 2018b), and expectile (Rowland et al., 2019) TD-learning.

**Theorem 3** (The Statistical HJB Loss for Policy Evaluation)**.** *Let* $\Phi$ *be a statistically smooth imputation strategy with a corresponding set of admissible statistics* $\mathcal{S}_\Phi$, *and let* $\vec{\mathsf{s}} : \mathcal{X} \to \mathcal{S}_\Phi$ *be a statistics function. We define the mapping* $\Psi(\vec{\mathsf{s}}(x), z) = \Phi(\vec{\mathsf{s}}(x), [V_{\min}, z])$*. The* Statistical HJB Loss $\mathcal{L}_S$ *is defined as*

$$\begin{aligned}
&\mathcal{L}_S(\vec{\mathsf{s}}, \Psi) = \\
&\quad \Bigg[\nabla_{\vec{\mathsf{s}}(x)} \Psi(\vec{\mathsf{s}}(x), z)^\top \vec{\mathsf{s}}_x(x) \mu_\pi(x) \\
&\quad - (r(x) + \log \gamma z) \frac{\partial}{\partial z} \Psi(\vec{\mathsf{s}}(x), z) \\
&\quad + \frac{1}{2} \mathsf{Tr}\left(\boldsymbol{\sigma}_\pi(x)^\top \left(\mathbf{K}^x_\Phi(x, z) + \mathbf{K}^s_\Phi(x, z)\right) \boldsymbol{\sigma}_\pi(x)\right)\Bigg]^2
\end{aligned} \tag{10}$$

*where* $\vec{\mathsf{s}}_x \triangleq \mathsf{J}_x \vec{\mathsf{s}}$. *Let the assumptions of Corollary 1 hold. Then if* $F_\eta$ *satisfies* (9) *and* $F_{\eta(x)} = \Phi(\vec{\mathsf{s}}(x))$ *for each* $x \in \mathcal{X}$, *we have*

$$\mathcal{L}_S(\vec{\mathsf{s}}, \Phi) = 0 \tag{11}$$

We define a statistical HJB *loss*, as opposed to a PDE, since by restricting the return distribution function to a class that can be imputed by a given set of statistical functionals, (8) will not generally have a solution. However, analysis of (10) can reveal a lower bound on the approximation error, which can be useful when designing DRL algorithms in practice. Corollary 2 will demonstrate this.

While (11) looks daunting, for certain imputation strategies it can be simplified drastically. Imputation strategies that construct quantile distributions are particularly well-behaved in this regard, however, they necessitate a weakened interpretation of differentiability in order to make sense of the statistical HJB equation. Recall that quantile distributions are finite convex combinations of Dirac measures, so their CDFs are finite convex combinations of Heaviside functions. While these functions are in fact differentiable almost everywhere, their derivatives are zero, so all information about the distribution is lost under differentiation. When the return distribution function is statistically smooth, however, we can reason about solutions to distributional HJB equations in *the distributional sense*. For the purpose of the following results, $\psi'$ is said to be a *distributional derivative* of the tempered distribution $\psi : \mathbf{R} \to \mathbf{R}$ if for every smooth and rapidly-decaying function $\rho : \mathbf{R} \to \mathbf{R}$, we have

$$\int_{\mathbf{R}} \rho(z)\psi'(z)dz = -\int_{\mathbf{R}} \rho'(z)\psi(z)dz$$

Intuitively, a distributional solution to a differential equation is a mapping that satisfies the equation upon convolution with every "reasonable" smoothing kernel. This concept is discussed with more rigor in Appendix G.

**Corollary 2** (The Quantile HJB Equation for Policy Evaluation)**.** *Let $\eta^\pi$ be statistically smooth, and let $\mathrm{s}$ be the sketch that maps $\eta(x)$ to a quantile distribution for each $x \in \mathcal{X}$.*

*If $\vec{\mathrm{s}}(x) = \mathrm{s}(\eta^\pi(x))$ for each $x \in \mathcal{X}$ and $F_{\eta^\pi}$ is a distributional solution to (9), then sketch $\vec{\mathrm{s}}$ of the statistical functionals $\{s_k\}_{k=1}^N$ is a distributional solution to the following system of PDEs,*

$$\begin{cases} \langle \nabla_x \iota_k \vec{\mathrm{s}}(x), \mu_\pi(x)\rangle + r(x) + \log\gamma\iota_k\vec{\mathrm{s}}(x) \\ \qquad +\frac{1}{2}\operatorname{Tr}\left(\boldsymbol{\sigma}_\pi(x)^\top \mathsf{H}_x\iota_k\vec{\mathrm{s}}(x)\boldsymbol{\sigma}_\pi(x)\right) = 0 \\ \iota_k\vec{\mathrm{s}}(x) = s_k(\eta(x)) \\ \qquad k = 1,\dots,N \end{cases} \tag{12}$$

Remarkably, this shows that the statistical diffusivity present in (11) vanishes under the quantile imputation strategy. The significance of this corollary is twofold. Firstly, it demonstrates that under the quantile representation, distributional dynamic programming reduces to solving a system of HJB equations, so existing HJB solving methods can be leveraged (such as that of Munos & Bourgine (1997)) for continuous-time distributional RL. Moreover, comparing (12) and (10), it is clear that such a reduction *is not possible* in general – in particular, to adapt categorical or expectile TD-learning algorithms to the continuous-time setting, one must take extra care to account for the spatial and statistical diffusivity due to the corresponding imputation strategies.

## 4. A Reinforcement Learning Algorithm

We propose a model-based DRL algorithm for jointly learning the return distribution function and optimizing the policy. Our algorithm is akin to the Quantile Regression TD-Learning (QTD) algorithm (Dabney et al., 2018b) with two important differences: *(a)* we update return distributions according to the Quantile HJB equation (12) as opposed to the distributional Bellman equation (Bellemare et al., 2017), and *(b)* we employ a differential updating scheme that converges in the limit of continuous time updates as opposed to a simple gradient descent.

In order to adapt a TD-learning algorithm like QTD to the continuous-time setting, we must note that TD updates may occur at arbitrarily high frequencies – as such, return distributions can evolve continuously in time. We must ensure that our update scheme is well-defined in this limit. To do so, we model a gradient *flow* as opposed to a sequence of gradient updates. Jordan et al. (2002) presents the *JKO scheme* to accomplish this in the 2-Wasserstein space when the loss functional has the form $\mathscr{F}(\eta) = \int_{\mathcal{R}} U\,d\eta - \frac{1}{\beta}\mathcal{H}(\eta)$, where $\mathcal{H}$ denotes entropy. We must derive the function $U$ such that the loss is minimized at the return distribution function. This is precisely what is done by (Martin et al., 2020) to minimize the distributional Bellman error.

In order to adapt the JKO scheme of Martin et al. (2020), we replace the distributional Bellman error with a signal that we refer to as the *kinetic energy of returns*. We use a quantile imputation strategy $\Phi$ to approximate return distribution functions, so return distributions can be interpreted as finite sets of $N$ return "particles" each having equal mass. For a set of particles distributed by $\eta(x) \in \mathcal{P}(\mathcal{R})$ for $x \in \mathcal{X}$, let $\Psi(x,z) = \Phi(\vec{\mathrm{s}}(x), z) = F_\eta(x,z)$. Denoting by $\mathscr{L}$ the infinitesimal generator of the conditional backward return process, we define the kinetic energy according to

$$U(z) = \frac{1}{2}\left(\mathscr{L}\Psi(x,z)\right)^2 \tag{13}$$

This results in the loss $\mathscr{F}_\beta : \mathcal{P}(\mathcal{R}) \to \mathbf{R}$ given by

$$\mathscr{F}_\beta(\eta(x)) = \int_{\mathcal{R}} \frac{1}{2}\left(\mathscr{L}\Psi(x,z)\right)^2 \eta(x,dz) - \frac{1}{\beta}\mathcal{H}(\eta) \tag{14}$$

Under the quantile distribution representation, we see in (12) that $\mathscr{L}\Psi(x,\cdot)$ is affine. Therefore, the kinetic energy is convex, and then it is a well established result that $\mathscr{F}_\beta$ is convex (Ambrosio et al., 2008). Therefore, $\mathscr{F}_\beta$ has a unique (global) minimum. We consider the gradient flow of (14):

$$\eta_s(x) = -\nabla\mathscr{F}_\beta(\eta_s(x,\cdot)) \tag{15}$$

where $s$ is a continuous time parameter.[4] Remarkably, Jordan et al. (2002) shows that (15) in the 2-Wasserstein space is equivalent to the *Fokker-Planck equation*, which is a well-known PDE in various scientific disciplines. As a result of this, it is well known that (14) is minimized when the density $\varrho$ of $\eta$ satisfies $\eta \propto \exp(-\beta U)$.

Furthermore, Martin et al. (2020) shows that as $\beta \to \infty$, the minimizer of $\mathscr{F}_\beta$ coincides with $U \equiv 0$. With $U$ given by (13), this occurs when $\eta(x)$ satisfies the Kolmogorov backward equation for $\mathscr{L}$. By Theorem 3, we see that the loss is minimized by the return distribution function.

To construct a reinforcement learning algorithm, we must discretize the gradient flow (15) in time. The JKO scheme for (15) consists of computing the sequence of iterates $\{\widetilde{\eta}_k\}_{k=1}^\infty$ given by

$$\widetilde{\eta}_{k+1} \in \arg\min_\eta \left[2\tau\int_{\mathcal{R}} U\,d\eta(x) + W_2^\beta(\eta, \widetilde{\eta}_k)\right] \tag{16}$$

where $W_2^\beta$ is the entropically-regularized 2-Wasserstein distance (Cuturi, 2013) with inverse temperature $\beta$. Computation of this distance is tractable for quantile distributions via the *Sinkhorn algorithm* (Cuturi, 2013; Martin et al., 2020). Remarkably, the Sinkhorn algorithm is differentiable (Peyré

[4] This is not necessarily equivalent to the time parameter in the MDP.

et al., 2019), which allows us to incorporate it with gradient-based optimization schemes.

Under a continuous time interpolation of $\{\widetilde{\eta}_k\}_{k=1}^\infty$ given by Jordan et al. (2002), the interpolated curve converges to (24) as $\tau \to 0$ in (16).

### 4.1. Control

In continuous time, individual actions have negligible effects on the return, and consequently the action-value function cannot be used to infer optimal actions (Baird III, 1993). This concept is formalized by Bellemare et al. (2016) and Tallec et al. (2019). To account for this, *advantage-updating* (Baird III, 1993) and similar schemes (Bellemare et al., 2016) introduce alternative notions of action values that are meaningful in the continuous-time limit. However, to the best of our knowledge, such concepts have not been studied in a distributional framework. Since the theory that was presented in this paper is concerned only with policy evaluation, such developments are out of scope, but are certainly interesting avenues for future work.

In order to perform simulations, we must discretize time. When time is discretized, individual actions are no longer negligible[5], so state-action pairs will not be completely invariant to the action.

We associate $|\mathcal{A}|$ return distributions to each state (one per action), and henceforth we use the notation $\eta^\pi(x, a)$ to denote the return distribution associated to the policy $\pi$ corresponding to the state-action pair $(x, a)$. Likewise, statistics functions are indexed by actions, so we now write $\Phi(\vec{\mathsf{s}}(x, a), z) = \eta(x, a)$.

In order to infer an optimal policy given a return distribution function, we must impose an ordering among return distributions. We simply order return distributions by their expected values, akin to many common DRL algorithms (Bellemare et al., 2017; Dabney et al., 2018b). Subsequently, we deem a policy $\pi^\star$ optimal if, for every state-action pair $(x, a)$, $\eta^{\pi^\star}(x, a)$ has greater expectation than $\eta^\pi(x, a)$ for any other policy $\pi$.

### 4.2. Approximating Solutions to the DHJB Equation

In order to maintain estimates of the return distribution function at each state-action pair, we must discretize the state space to a finite collection of points. Consequently, we must derive approximations of the differential terms in (12). We will write $f^\pm(x) = \max(\pm f(x), 0)$, and we will approximate the drift and variance of the dynamics by $\widehat{\mu} : \mathcal{X}_\varepsilon \times \mathcal{A} \to \mathcal{X} \approx \mu_\pi$ and $\widehat{\mathbf{\Sigma}} : \mathcal{X}_\varepsilon \times \mathcal{A} \to \mathbf{R}^{d\times d} \approx \boldsymbol{\sigma}_\pi \boldsymbol{\sigma}_\pi^\top$.

Fortunately, (12) has a very special form: it is simply a system of HJB equations. Due to the prevalence of HJB equations in control and continuous-time RL research, there are myriad established methods for solving them. We will make use of the finite-differences scheme that was introduced by Munos & Bourgine (1997), which approximates solutions to HJB equations driven by Itô diffusions.

For some $\varepsilon > 0$, we discretize $\mathcal{X}$ to a lattice $\mathcal{X}_\varepsilon = \{\sum_{n=1}^d i_n \varepsilon \vec{e}_n : i_n \in \mathbf{Z}\} \cap \mathcal{X}$ where $\{\vec{e}_i\}_{i=1}^d$ is the standard basis of $\mathbf{R}^d$. The *neighbors* of each state $\xi \in \mathcal{X}_\varepsilon$ are points adjacent to $\xi$ in the lattice. The mapping $\mathsf{A}_\varepsilon : \mathcal{X}_\varepsilon \to 2^{\mathcal{X}_\varepsilon}$ maps each state to the set of its neighbors, given as follows,

$$\begin{aligned}\mathsf{A}_\varepsilon(\xi) = \{\xi' \in \mathcal{X}_\varepsilon \setminus \{\xi\} : \xi' = \xi + a\varepsilon\vec{e}_i + b\varepsilon\vec{e}_j,\\ i, j \in [d],\ i \neq j,\\ a, b \in \{0, \pm 1\}\}\end{aligned}$$

Since the state space is divided into a finite collection of "cells", all states within a given cell are indistinguishable from one another in our approximation. As such, at any given cell, even if the dynamics are deterministic, the agent's future states are randomly distributed. This phenomenon is depicted in Figure 1. Suppose the agent has velocity $\vec{v}$ with speed $v$. For the timestep $\Delta$ for which $\Delta v = 1 \cdot \varepsilon$, the components of $\vec{v}$ can be interpreted as probabilities $p\varepsilon$ as shown in Figure 1. When the dynamics are driven by an Itô diffusion, Munos & Bourgine (1997) shows that this timestep is given by

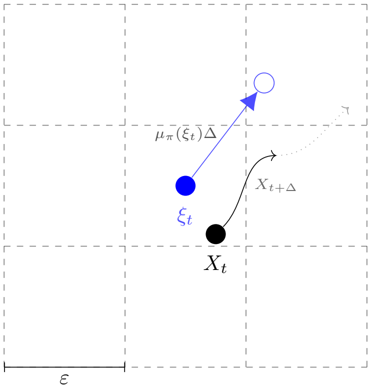

*Figure 1.* Finite-differences approximate trajectory (blue) relative to real trajectory (black).

$$\Delta_{\xi,a} = \frac{\varepsilon^2}{\varepsilon\,\|\widehat{\mu}(\xi, a)\|_1 + \mathsf{Tr}(\widehat{\mathbf{\Sigma}}(\xi, a)) - \frac{1}{2}\sum_{j\neq i} |\widehat{\mathbf{\Sigma}}(\xi, a)_{ij}|}$$

Subsequently, the transition probabilities are given by

$$p(\xi, a, \xi \pm \varepsilon\vec{e}_i) =$$

[5] Note that, while individual actions may have discernible influence on the return in this setting, their influence is still small. Consequently, due to noise in the training process, convergence to an optimal policy can be quite slow as reported by Baird III (1993), so it would still behoove us to study a distributional analogue to advantage updating even in the time-discretized setting.

$$\frac{\Delta_{\xi,a}}{2\varepsilon^2}\left[2|(\iota_i\widehat{\mu})^{\pm}(\xi,a)| + \widehat{\mathbf{\Sigma}}(\xi,a)_{ii} - \sum_{j\neq i}|\widehat{\mathbf{\Sigma}}(\xi,a)_{ij}|\right]$$

$$p(\xi,a,\xi+\varepsilon(\vec{e}_i \pm \vec{e}_j)) = \frac{\Delta_{\xi,a}}{2\varepsilon^2}\widehat{\mathbf{\Sigma}}^{\pm}(\xi,a)_{ij} \quad i \neq j$$

$$p(\xi,a,\xi-\varepsilon(\vec{e}_i \pm \vec{e}_j)) = \frac{\Delta_{\xi,a}}{2\varepsilon^2}\widehat{\mathbf{\Sigma}}^{\pm}(\xi,a)_{ij} \quad i \neq j$$

$$p(\xi,a,\xi') = 0 \quad \text{otherwise}$$

We define the finite differences distributional Bellman operator by $\mathcal{T}_\triangle$ where

$$\begin{aligned} &\pi^\star_\xi \leftarrow \arg\max_{a'\in\mathcal{A}}\mathbf{E}\left[\Phi(\xi,a')\right] \qquad \forall\xi\in\mathcal{X}_\varepsilon \\ &\mathsf{b}^{\Delta_{\xi,a}}_{r,\gamma} : \mathcal{R}\to\mathcal{R} : z \mapsto \Delta_{\xi,a}r + \gamma^{\Delta_{\xi,a}}z \\ &\qquad\qquad i\neq j\in[d],\ a,b\in\{0,\pm1\}\} \\ &\mathcal{T}_\triangle\Phi(\xi,a) \leftarrow \sum_{\xi'\in\mathsf{A}_\varepsilon(\xi)} p(\xi,a,\xi')\left(\mathsf{b}^{\Delta_{\xi,a}}_{r,\gamma}\right)_\sharp\Phi(\xi',\pi^\star_{\xi'}) \end{aligned} \tag{17}$$

where $\sharp$ denotes the pushforward operation, defined by $f_\sharp\mu = \mu\circ f^{-1}$ for a measure $\mu$ and a measurable function $f$. Finally, the finite differences approximation of (12) is the fixed point equation

$$\mathcal{T}_\triangle\Phi(\xi,a) = \Phi(\xi,a) \qquad \xi\in\mathcal{X}_\varepsilon,\ a\in\mathcal{A} \tag{18}$$

We derive an algorithm based on these principles as an iterative method for solving (18). Notably, with the quantile representation, our algorithm is tractable relative to a HJB-solving oracle, such as the algorithm proposed by Munos (1997). The learning update is summarized in Algorithm 1, which can be applied in both online and offline settings. When the dynamics $\mu_\pi, \boldsymbol{\sigma}_\pi$ are unknown, which is usually the case in reinforcement learning, they can be estimated by the sample mean and sample covariance of observed transitions, respectively (Munos & Bourgine, 1997). The algorithm has access to a mapping $\mathsf{Enc} : \mathcal{X} \to \mathcal{X}_\varepsilon$ which maps states to their closest point in the lattice $\mathcal{X}_\varepsilon$. For the purpose of exploration, we simply employ a $\varepsilon$-greedy policy.

## 5. A Qualitative Demonstration

We simulate the performance of the FD-WGF $Q$-learning algorithm on a simple task based on a continuous MDP suggested by Munos (2004) as an example of an MDP whose value function does not satisfy the HJB equation in the usual sense. In this environment, we control a particle on $\mathcal{X} = [0,1]$ with actions among $\mathcal{A} = \{-1,1\}$. The dynamics of the particle are given by $\dot{x}(t) = a(t)$.

Rewards are zero in the interior of $\mathcal{X}$, and are otherwise sampled from $\mathcal{N}(2,2)$ and $\mathcal{N}(1,1)$ at states $1, 0$ respectively. The discount factor is $\gamma = 0.3$, and observations occur at a frequency $\omega = 1$kHz. We observe the performance of FD-WGF $Q$-learning relative to the Quantile Regression TD-learning algorithm (QTD) proposed by Dabney et al. (2018b). Figure 2 depicts an overview of the return distribution functions learned by both algorithms. In Figure 2, the darker blue regions represent larger probability mass of the return distribution. The dashed blue line is the analytical value function. We observe that our proposed algorithm learns a good representation of the value function, whereas the QTD algorithm tends to fail near the point of non-differentiability of the value function. Consequently, we see that QTD overestimates the value function over much of the state space. Figure 3 shows the return distributions learned by each algorithm near the boundaries of the state space, where the return variance is greatest.

**Algorithm 1** Continuous-time distributional RL update

```
Require: WGF time parameter $\tau$
Require: Learning rate $\alpha$
Require: State transition $(x, a, r, x')$
Require: Duration of transition $\delta$
  $\xi \leftarrow \mathsf{Enc}(x)$
  $\{$Update model$\}$
  $\widehat{\mu}(\xi,a) \leftarrow (1-\alpha)\widehat{\mu}(\xi,a) + \alpha(x'-x)/\delta$
  $\sigma \leftarrow x' - x - \Delta\widehat{\mu}(\xi)$
  $\widehat{\mathbf{\Sigma}}(\xi,a) \leftarrow (1-\alpha)\widehat{\mathbf{\Sigma}}(\xi,a) + \alpha\Delta^{-1}\sigma\sigma^\top$
  $\{$Compute mixture of target quantiles$\}$
  for $y \in \mathsf{A}_\varepsilon(\xi)$ do
    $(\mathbf{T}_{\Delta_{\xi,a}})_y \leftarrow \Delta_{\xi,a}r + \gamma^{\Delta_{\xi,a}}\vec{\mathsf{s}}(y,\pi^\star_y)$
    $\mathbf{p}_y \leftarrow p(\xi,a,y)$
  end for
  $\widehat{\eta} \leftarrow \frac{1}{N}\sum_{y\in\mathsf{A}_\varepsilon(\xi)}\mathbf{p}_y\sum_{k=1}^N \delta_{(\mathbf{T}_{\Delta_{\xi,a}})_{y,k}}$
  $\{$Update quantiles$\}$
  $\eta_0 \leftarrow \frac{1}{N}\sum_{k=1}^N \delta_{\vec{\mathsf{s}}(\xi,a)_k}$
  $\eta \leftarrow \arg\min_{\nu\in\mathcal{P}(\mathcal{R})}\left[2\tau\mathop{\mathbf{E}}_{Z'\sim\widehat{\eta},Z\sim\nu}\left[(Z-Z')^2\right] + W_2^\beta(\nu,\eta_0)\right]$
  $\vec{\mathsf{s}}(\xi,a) \leftarrow \mathbf{s}(\eta)\{$Extract quantiles of return distribution$\}$
```

We observe that FD-WGF $Q$-learning represents the true return distribution far more accurately near $\partial\mathcal{X}$, while both algorithms tend to “lose” variance further away from the boundaries. That said, especially when $x > 0.5$, we see that QTD has substantially more difficulty learning the variance of the return distributions than FD-WGF $Q$-learning.

## 6. Conclusion

Our work demonstrates that extra care should be taken when designing distributional RL algorithms for continuous-time problems. Notably, we have shown that the approximation of return distributions as empirical distributions is particularly well suited to continuous-time problems, as these representations eliminate the *statistical diffusivity* of the return due to the stochasticity of the system. Through our simulated experiments, we confirmed the hypothesis that

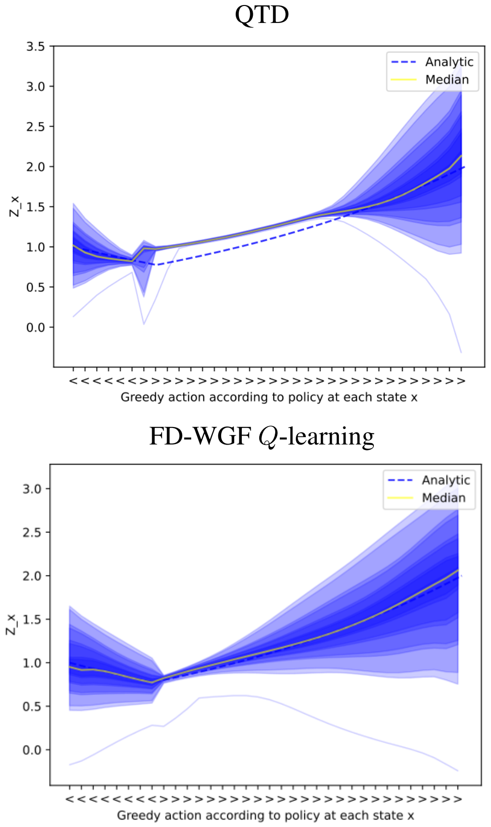

*Figure 2.* Return distribution functions and policies learned by FD-WGF $Q$-learning and QTD

accounting for continuous time aids DRL algorithms to preserve the return distribution entropy.

The algorithm presented in this work, as a finite-differences based scheme, becomes intractable as the dimension $d$ of the state space grows. However, we note that function approximation can be integrated without much difficulty to account for these cases. Since the loss function is differentiable, we can envision an algorithm similar to Algorithm 1 with $\vec{s}$, $\mu_\pi$, and $\boldsymbol{\sigma}_\pi$ parameterized by neural networks, with the gradient and Hessian of $\vec{s}$ computed via automatic differentiation and parameters trained via gradient descent. This algorithm would be similar to *Online WGF Fitted $Q$-iteration* (Martin et al., 2020), which demonstrates promising results. Such extensions are left for future work.

## Acknowledgements

The authors thank Julie Alhosh, Matthieu Geist, and Mark Rowland for their detailed and constructive feedback on previous drafts, as well as the anonymous reviewers whose insights helped improve the paper. This work was supported by the National Sciences and Engineering Research Council (NSERC) Canadian Robotics Network (NCRN), as well as the Canada CIFAR AI Chair program. Harley Wiltzer's research was supported by a NSERC CGS-M grant.

## References

Ambrosio, L., Gigli, N., and Savaré, G. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

Baird III, L. C. Advantage updating. Technical report, Wright Lab Wright-Patterson AFB OH, 1993.

Bellemare, M. G., Ostrovski, G., Guez, A., Thomas, P., and Munos, R. Increasing the action gap: New operators for reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

Bellemare, M. G., Dabney, W., and Munos, R. A distributional perspective on reinforcement learning. In *ICML*, 2017.

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

Bellman, R. A markovian decision process. *Journal of mathematics and mechanics*, 6(5):679–684, 1957.

Chizat, L. and Bach, F. On the global convergence of gradient descent for over-parameterized models using optimal transport. *arXiv preprint arXiv:1805.09545*, 2018.

Chow, Y. and Ghavamzadeh, M. Algorithms for CVaR optimization in MDPs. In *Advances in Neural Information Processing Systems*, 2014.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transportation distances, 2013.

Dabney, W., Ostrovski, G., Silver, D., and Munos, R. Implicit quantile networks for distributional reinforcement learning. In *International conference on machine learning*, pp. 1096–1105. PMLR, 2018a.

Dabney, W., Rowland, M., Bellemare, M. G., and Munos, R. Distributional reinforcement learning with quantile regression. In *AAAI*, 2018b.

De Giorgi, E. New problems on minimizing movements. *Ennio de Giorgi: Selected Papers*, pp. 699–713, 1993.

De Giorgi, E., Marino, A., and Tosques, M. Problems of evolution in metric spaces and maximal decreasing curve. *Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur.(8)*, 68(3):180–187, 1980.

Fleming, W. H. and Soner, H. M. *Controlled Markov processes and viscosity solutions*, volume 25. Springer Science & Business Media, 2006.

Halperin, I. Distributional offline continuous-time reinforcement learning with neural physics-informed pdes (sciphy rl for doctr-l). *arXiv preprint arXiv:2104.01040*, 2021.

Hessel, M., Modayil, J., Van Hasselt, H., Schaul, T., Ostrovski, G., Dabney, W., Horgan, D., Piot, B., Azar, M., and Silver, D. Rainbow: Combining improvements in deep reinforcement learning. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

Jordan, R., Kinderlehrer, D., and Otto, F. The variational formulation of the fokker-planck equation. *SIAM J. Math. Anal*, 29:1–17, 2002.

Kolmogorov, A. Über die analytischen methoden in der wahrscheinlichkeitsrechnung. *Mathematische Annalen*, 104(1):415–458, 1931.

Krylov, N. V. Controlled diffusion processes. 1980.

Lax, P. and Sons, J. W. . *Functional Analysis*. Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts. Wiley, 2002. ISBN 9780471556046. URL https://books.google.ca/books?id=-jbvAAAAMAAJ.

Le Gall, J.-F. *Brownian motion, martingales, and stochastic calculus*. Springer, 2016.

Martin, J. D., Lyskawinski, M., Li, X., and Englot, B. Stochastically dominant distributional reinforcement learning, 2020.

Mavrin, B., Yao, H., Kong, L., Wu, K., and Yu, Y. Distributional reinforcement learning for efficient exploration. In *International conference on machine learning*, pp. 4424–4434. PMLR, 2019.

Munos, R. A convergent reinforcement learning algorithm in the continuous case based on a finite difference method. In *IJCAI (2)*, pp. 826–831, 1997.

Munos, R. A study of reinforcement learning in the continuous case by the means of viscosity solutions. *Machine Learning*, 40:265–299, 2004.

Munos, R. and Bourgine, P. Reinforcement learning for continuous stochastic control problems. In *NIPS*, pp. 1029–1035, 1997.

Muratori, M. and Savaré, G. Gradient flows and evolution variational inequalities in metric spaces. i: structural properties, 2018.

Peyré, G., Cuturi, M., et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Prashanth, L. and Fu, M. Risk-sensitive reinforcement learning. *arXiv preprint arXiv:1810.09126*, 2021.

Prashanth, L. and Ghavamzadeh, M. Actor-critic algorithms for risk-sensitive MDPs. In *Advances in Neural Information Processing Systems*, 2013.

Puterman, M. L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Rogers, L. C. G. and Williams, D. Diffusions, markov processes and martingales, volume 1: Foundations. *John Wiley & Sons, Ltd., Chichester*, 7, 1994.

Rowland, M., Dadashi, R., Kumar, S., Munos, R., Bellemare, M., and Dabney, W. Statistics and samples in distributional reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 5528–5536, 2019. URL http://proceedings.mlr.press/v97/rowland19a/rowland19a.pdf.

Santambrogio, F. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7:87–154, 2016.

Tallec, C., Blier, L., and Ollivier, Y. Making deep q-learning methods robust to time discretization. In *International Conference on Machine Learning*, pp. 6096–6104. PMLR, 2019.

Tamar, A., Glassner, Y., and Mannor, S. Optimizing the CVaR via sampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2015.

Williams, C. K. and Rasmussen, C. E. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

Yang, D., Zhao, L., Lin, Z., Qin, T., Bian, J., and Liu, T.-Y. Fully parameterized quantile function for distributional reinforcement learning. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/f471223d1a1614b58a7dc45c9d01df19-Paper.pdf.

Zhang, R., Chen, C., Li, C., and Carin, L. Policy optimization as wasserstein gradient flows. In *ICML*, 2018.

# A. Proofs

## A.1. Proofs of Results in §3

*Proof of Proposition 1.* Let $\psi \in C(\mathcal{X} \times \mathcal{R}; \mathbf{R})$ and $h > 0$. As usual, we denote the canonical filtration by $(\mathcal{F}_t)_{t \geq 0}$. By the definition of the truncated return process,

$$\begin{aligned}
\mathbf{E}\left[\psi(J_{t+h}) \mid \mathcal{F}_t\right] &= \mathbf{E}\left[\psi(X_{t+h}, \overline{G}_{t+h}) \;\middle|\; \mathcal{F}_t\right] \\
&= \mathbf{E}\left[\psi\left(X_{t+h}, \overline{G}_t + \int_t^{t+h} \gamma^s r(X_s) ds\right) \;\middle|\; \mathcal{F}_t\right] \\
&= \mathbf{E}\left[\psi\left(X_{t+h}, \overline{G}_t + \int_t^{t+h} \gamma^s r(X_s) ds\right) \;\middle|\; J_t\right]
\end{aligned}$$

where the final step holds since the process $(X_t)_{t \geq 0}$ is assumed to be Markovian. Thus, we've shown that for any $\psi \in C(\mathcal{X} \times \mathcal{R}; \mathbf{R})$, there exists a function $m : \mathcal{X} \times \mathcal{R} \to \mathbf{R}$ where

$$\mathbf{E}\left[\psi(J_{t+h}) \mid \mathcal{F}_t\right] = m(X_t, \overline{G}_t)$$

Therefore, the process $(J_t)_{t \geq 0}$ is Markovian. □

**Lemma 2.** *Let $(J_t)_{t \geq 0} = (X_t, \overline{G}_t)_{t \geq 0}$ be the truncated return process defined in Theorem 2. Then $\left(\overline{G}_t\right)_{t \geq 0}$ is a finite variation process.*

In order to determine the infinitesimal generator of the truncated return process, it will be necessary to estimate its quadratic variation and the bracket $([X, \overline{G}]_t)_{t \geq 0}$. Establishing $\left(\overline{G}_t\right)_{t \geq 0}$ as a finite variation process will greatly simplify this estimate.

*Proof of Lemma 2.* By definition, we have

$$\overline{G}_t = \int_0^t \gamma^s r(X_s) ds$$

Consider the measurable space $(\mathbf{R}_+, \Sigma)$ where $\Sigma$ is the $\sigma$-algebra of Lebesgue-measurable subsets of the nonnegative reals, and let $\Lambda$ denote the Lebesgue measure. We will use $(\mathbf{R}_+, \Sigma)$ to measure *time*. By the Radon-Nikodym theorem, for each sample path $\omega \in \Omega$, the function $\mu_\omega : \Sigma \to \mathbf{R}$ shown below is a signed measure on this measurable space,

$$\mu_\omega(A) = \int_A \gamma^{s \wedge T(\omega)} r(X_{s \wedge T(\omega)}(\omega)) \Lambda(ds) \qquad A \in \Sigma$$

Then, for any $\omega \in \Omega$, the mapping $t \mapsto G_t(\omega) = \mu_\omega([0, t])$. This shows that each sample path is a function $a : t \mapsto \mu_\omega([0, t])$ for the measure $\mu_\omega$, so every sample path is a finite variation function by definition. □

**Lemma 3.** *The truncated return process $(J_t)_{t \geq 0}$ as defined in Theorem 2 is a Feller-Dynkin process.*

*Proof.* Consider the filtered probability space $\mathsf{P} = (\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, \Pr)$ defined previously. Proposition 1 shows that $(J_t)_{t \geq 0}$ is a Markov process. It remains to show that it is a Feller-Dynkin process. First, we must show that its transition semigroup maps $(P_t)_{t \geq 0}$ are endomorphisms on $C_0(\mathbf{R}_+ \times \mathcal{X} \times \mathcal{R})$. Let $\psi \in C_0(\mathbf{R}_+ \times \mathcal{X} \times \mathcal{R})$.

Note that since $(X_t)_{t \geq 0}$ has continuous sample paths, $\left(\overline{G}_t\right)_{t \geq 0}$ has absolutely continuous sample paths since

$$\overline{G}_t(\omega) = \int_0^t \gamma^s r(X_s(\omega)) ds \qquad \omega \in \Omega$$

so it is bounded by the integral of a bounded function. Therefore $P_\delta \psi$ can be expressed as

$$P_\delta \psi = \int \psi \circ (t+\delta, X_{t+\delta}, \overline{G}_{t+\delta}) d \operatorname{Pr}$$

Since the sample paths $X_{t+\delta}, \overline{G}_{t+\delta}$ are continuous, the integrand above is a continuous function. Additionally, since $\psi, \mathcal{X}, \mathcal{R}$ are all compactly supported, we see that $P_\delta \psi$ is as well. Therefore $P_\delta \psi \in C_0(\mathbf{R}_+ \times \mathcal{X} \times \mathcal{R})$.

It is easy to check that $P_0 \psi = \mathsf{id}$. This follows simply from the fact that $(X_t)_{t \geq 0}$ is a Feller-Dynkin process (so its semigroup has an identity) and $\left(\overline{G}_t\right)_{t \geq 0}$ is deterministic given $(X_t)_{t \geq 0}$. For the same reason, it follows that $P_t P_s = P_{t+s}$.

It remains to show that $\|P_\delta \psi - P_0 \psi\|_\infty \overset{\delta \downarrow 0}{\longrightarrow} 0$. We have

$$\begin{aligned}
\|P_\delta \psi - P_0 \psi\|_\infty &= \|P_\delta \psi - \psi\|_\infty \\
&= \left\| \int_{\mathcal{X} \times \mathcal{R}} \left( \psi \circ (t+\delta, X_{t+\delta}, \overline{G}_{t+\delta}) - \psi(t, X_t, \overline{G}_t) \right) d\mathsf{P} \right\|_\infty \\
&= \left\| \int_{\mathcal{X} \times \mathcal{R}} \psi \circ (t+\delta, X_{t+\delta}, \overline{G}_{t+\delta}) d\mathsf{P} - \psi(t, X_t, \overline{G}_t) \right\|_\infty
\end{aligned}$$

Since $\psi$ is supported on a compact finite-dimensional set and it is continuous, it follows that it is bounded. Therefore, it follows by the dominated convergence theorem that

$$\begin{aligned}
\lim_{\delta \to 0} \int \psi \circ (t+\delta, X_{t+\delta}, \overline{G}_{t+\delta}) d \operatorname{Pr} &= \int \psi \circ \lim_{\delta \to 0} (t+\delta, X_{t+\delta}, \overline{G}_{t+\delta}) d \operatorname{Pr} \\
&= \int \psi(t, X_t, \overline{G}_t) d \operatorname{Pr} \\
&= \psi(t, X_t, \overline{G}_t)
\end{aligned}$$

This proves the claim. □

**Lemma 4.** *The truncated return process $(J_t)_{t \geq 0}$ defined in Theorem 2 has an infinitesimal generator $\mathscr{L} : C_0(\mathbf{R}_+ \times \mathcal{X} \times \mathcal{R}) \to C_0(\mathbf{R}_+ \times \mathcal{X} \times \mathcal{R})$ given by*

$$\mathscr{L}\psi(t, x, \overline{g}) = (\mathscr{L}_X \psi(t, \cdot, \overline{g}))(x) + \gamma^t r(x) \frac{\partial}{\partial \overline{g}} \psi(t, x, \overline{g}) + \frac{\partial}{\partial t} \psi(t, x, \overline{g}) \tag{19}$$

*where $\mathscr{L}_X$ is the infinitesimal generator of the process $(\iota_2 J_t)_{t \geq 0} = (X_t)_{t \geq 0}$.*

*Proof.* Since Lemma 3 shows that $(J_t)_{t \geq 0}$ is a Feller-Dynkin process, the existence of an infinitesimal generator driving this process is guaranteed. Let $\psi \in C_0^2(\mathbf{R}_+ \times \mathcal{X} \times \mathcal{R})$ and denote $j = (t, x, \overline{g})$. Then

$$\begin{aligned}
\frac{P_\delta \psi(j) - \psi(j)}{\delta} &= \frac{1}{\delta} \Big( \mathbf{E} \left[ \psi(J_{t+\delta}) \mid J_t = j \right] - \psi(j) \Big) \\
&= \mathbf{E} \left[ \frac{1}{\delta} \left( \psi(J_{t+\delta}) - \psi(J_t) \right) \,\middle|\, J_t = j \right]
\end{aligned} \tag{$*$}$$

We will proceed by applying Itô's Lemma to this expectation. However, we must first verify that $(J_t)_{t \geq 0}$ satisfies the hypotheses of Itô's Lemma, namely, it must be a semimartingale. It is easy to verify that this is the case. We will express the tuples $J_t = (t, X_t, \overline{G}_t) \in \mathbf{R}_+ \times \mathcal{X} \times \mathcal{R}$ as $d+2$-dimensional vectors (since $\mathcal{X} \subset \mathbf{R}^d$), where the first $d$ dimensions encode the state $X_t$, the $d+1$th dimension encodes the truncated return $\overline{G}_t$, and the last dimension encodes time. We have

$$M_t \triangleq \begin{bmatrix} X_t - \mathbf{E}\left[X_t\right] \\ 0 \\ 0 \end{bmatrix}$$

$$A_t \triangleq \begin{bmatrix} \mathbf{E}\left[X_t\right] \\ \overline{G}_t \\ t \end{bmatrix}$$

$$J_t = M_t + A_t$$

It follows immediately from Lemma 2 that $(A_t)_{t\geq 0}$ is a finite variation process. Furthermore, since $(X_t)_{t\geq 0}$ is a Feller-Dynkin process, we know from Lemma 5 that $(X_t - \mathbf{E}\left[X_t\right])_{t\geq 0}$ is a martingale. Thus, $(J_t)_{t\geq 0}$ can be expressed as a sum of a local martingale[6] and a finite variation process, making it a semimartingale by definition.

Since $(J_t)_{t\geq 0}$ is a semimartingale and $\psi \in C_0^2(\mathbf{R}_+ \times \mathcal{X} \times \mathcal{R})$, we may apply Itô's lemma to expand $(*)$ as follows, where all expectations are conditioned on $J_t = j$,

$$\begin{aligned}
\frac{(P_\delta - \mathsf{id})\psi(j)}{\delta} &= \frac{1}{\delta}\mathbf{E}\left[\int_t^{t+\delta}\sum_{i=1}^{d+2}\frac{\partial\psi(J_s)}{\partial j^i}dJ_s^i + \frac{1}{2}\int_t^{t+\delta}\sum_{i=1}^{d+2}\sum_{k=1}^{d+2}\frac{\partial^2\psi(J_s)}{\partial j^i\partial j^k}d[J^i, J^k]_s\right] \\
&= \frac{\partial}{\partial t}\psi(j) + \overbrace{\frac{1}{\delta}\mathbf{E}\left[\int_t^{t+\delta}\sum_{i=1}^{d}\frac{\partial\psi(J_s)}{\partial j^i}dJ_s^i + \frac{1}{2}\int_t^{t+\delta}\sum_{i=1}^{d}\sum_{k=1}^{d}\frac{\partial^2\psi(J_s)}{\partial j^i\partial j^k}d[J^i, J^k]_s\right]}^{a} \\
&\quad + \overbrace{\frac{1}{\delta}\mathbf{E}\left[\int_t^{t+\delta}\frac{\partial\psi(J_s)}{\partial j^{d+1}}dJ_s^{d+1} + \frac{1}{2}\frac{\partial^2\psi(J_s)}{\partial (j^{d+1})^2}d[J^{d+1}, J^{d+1}]_s\right]}^{b} \\
&\quad + \overbrace{\frac{1}{2\delta}\mathbf{E}\left[\int_t^{t+\delta}\sum_{i=1}^{d}\left(\frac{\partial^2\psi(J_s)}{\partial j^i\partial j^{d+1}}d[J^i, J^{d+1}]_s + \frac{\partial^2\psi(J_s)}{\partial j^i\partial j^{d+2}}d[J^i, J^{d+2}]_s\right)\right]}^{c}
\end{aligned}$$

Recall that $J_t^{1:d} = \iota_1 J_t = X_t$, and $J_t^{d+1} = \iota_2 J_t = \overline{G}_t$. In the limit as $\delta \downarrow 0$, the term $a$ above therefore is simply the generator of the process $(X_t)_{t\geq 0}$ applied to $\psi$. Moreover, since it was shown that $\left(\overline{G}_t\right)_{t\geq 0}$ is a finite variation process in Lemma 2, it follows that $[J^i, J^{d+1}] = [J^i, J^{d+2}] \equiv 0$ for any $i \in \{1, \dots, d+1\}$ (Le Gall, 2016). Consequently, we have $c \equiv 0$. Simplifying,

$$\begin{aligned}
\lim_{\delta\to 0}\frac{P_\delta\psi(j) - \psi(j)}{\delta} &= \mathscr{L}_X\psi(j) + \lim_{\delta\to 0}\frac{1}{\delta}\mathbf{E}\left[\int_t^{t+\delta}\frac{\partial\psi(J_s)}{\partial\overline{g}}d\overline{G}_s \;\middle|\; J_t = j\right] + \frac{\partial\psi(j)}{\partial t} \\
&= \mathscr{L}_X\psi(j) + \frac{\partial\psi(j)}{\partial\overline{g}}\gamma^t r(x) + \frac{\partial}{\partial t}\psi(j)
\end{aligned}$$

This completes the proof. □

**Lemma 1.** *Let $z \in \mathcal{R}$ be a desired return, and suppose that $\left(\overleftarrow{J}(z)_t\right)_{t\geq 0}$ is a Feller-Dynkin process with infinitesimal generator $\mathscr{L}_J$. Then at each $x \in \mathcal{O}$ and $z' \in \mathcal{R}$, $\eta^\pi$ satisfies*

$$\mathscr{L}_J F_{\eta^\pi}(x, z') = 0 \tag{7}$$

[6] By the definition of a local martingale, given in Appendix C.1.2, it is clear that all martingales are local martingales.

*Proof.* Let $z \in \mathcal{R}$. Let $\phi : (\mathcal{O} \times \mathcal{R}) \to \mathbf{R}$ be given by $\phi((x, z')) = \mathbf{1}_{[z' \geq 0]}$. Then define the function $u : \mathbf{R}_+ \times (\mathcal{X} \times \mathcal{R}) \to \mathbf{R}$ according to

$$\begin{aligned}
u(t, (x, z')) &= \mathbf{E}\left[\phi\big(x, \overleftarrow{G}(z)_T\big) \;\Big|\; \overleftarrow{J}_t(z) = (x, z')\right] \\
&= \Pr\left(\gamma^{-T}(z - \overline{G}_T) \geq 0 \middle| \overleftarrow{J}_t(z) = (x, z')\right) \\
&= \Pr\left(z \geq \overline{G}_T \middle| \overleftarrow{J}_t(z) = (x, z')\right) \\
&= \Pr\left(\gamma^t z' \geq \overline{G}_T - \overline{G}_t \middle| X_t = x\right) \\
&= \Pr\left(z' \geq \int_0^{T-t} \gamma^s r(X_{(t+s)\wedge T}) ds \middle| X_t = x\right) \\
&= \Pr\left(z' \geq \int_0^{T} \gamma^s r(X_{(t+s)\wedge T}) ds \middle| X_t = x\right) \\
&= \Pr(G^\pi(x) \leq z') \\
&= F_{\eta^\pi}(x, z')
\end{aligned}$$

The conditional expectation and probabilities are well-defined by Assumption 1. Note that $u$ has precisely the form of the solution to the Kolmogorov backward equation in Theorem 1. Thus, Theorem 1 establishes that $F_{\eta^\pi}(x, \cdot)$ satisfies (3) with the infinitesimal generator $\mathscr{L}_J$ of the conditional backward return process. Finally, since $F_{\eta^\pi}$ is time-homogeneous, its time derivative vanishes, and we are left with (7). □

**Theorem 2** (Distributional HJB Equation for Policy Evaluation)**.** *Denote by $\mathscr{L}_X$ the infinitesimal generator of the process $(X_t)_{t\geq 0} = \left(\iota_1 \overleftarrow{J}(z)_t\right)_{t \geq 0}$. Moreover, suppose Assumptions 1 and 2 hold. Then $F_{\eta^\pi}$ satisfies*

$$\begin{aligned}
(\mathscr{L}_X F_{\eta^\pi}(\cdot, z))(x) - (r(x) + z \log \gamma)\frac{\partial}{\partial z} F_{\eta^\pi}(x, z) &= 0 \\
\mathsf{P} - \textit{almost surely}
\end{aligned} \tag{8}$$

*Proof.* Note that the term $\frac{\partial}{\partial z} F_{\eta^\pi}(x, z)$ is the Radon-Nikodym derivative of $\eta^\pi(x)$ with respect to the Lebesgue measures. This derivative exists by Assumption 1. We have, for any $z \in \mathcal{R}$, $\overline{G}_t = z - \gamma^t \overleftarrow{G}(z)_t$. Since $\overleftarrow{G}(z)_t$ can be computed by a deterministic, differentiable transformation of $\overline{G}_t$ for any given $z \in \mathcal{R}$, it follows that $\left(\overleftarrow{J}(z)_t\right)_{t \geq 0}$ is a Feller-Dynkin process for each $z \in \mathcal{R}$.

Denote the infinitesimal generator of $\left(\overleftarrow{J}(z)_t\right)_{t\geq 0}$ by $\mathscr{L}_J$. The generator exists since the conditional backward return process is a Feller-Dynkin process, as previously mentioned. By a change of variables we immediately see that $\mathscr{L}_J = \mathscr{L}_G|_{t=0} - \log \gamma \iota_2 \frac{\partial}{\partial z}$, where $\mathscr{L}_G$ is the infinitesimal generator of the truncated return process.

By Lemma 3, we know that $F_{\eta^\pi}$ solves the Kolmogorov backward equation for the generator $\mathscr{L}_J$. Thus,

$$\begin{aligned}
0 &= \mathscr{L}_J F_{\eta^\pi}(x, z) \\
&= \mathscr{L}_G F_{\eta^\pi}(x, z) - z' \log \gamma \frac{\partial}{\partial z'} F_{\eta^\pi}(x, z) \\
&= \mathscr{L}_X F_{\eta^\pi}(x, z) - (r(x) + z \log \gamma)\frac{\partial}{\partial z} F_{\eta^\pi}(x, z)
\end{aligned}$$

□

**Theorem 3** (The Statistical HJB Loss for Policy Evaluation)**.** *Let $\Phi$ be a statistically smooth imputation strategy with a corresponding set of admissible statistics $\mathcal{S}_\Phi$, and let $\vec{\mathsf{s}} : \mathcal{X} \to \mathcal{S}_\Phi$ be a statistics function. We define the mapping $\Psi(\vec{\mathsf{s}}(x), z) = \Phi(\vec{\mathsf{s}}(x), [V_{\min}, z])$. The* Statistical HJB Loss *$\mathcal{L}_S$ is defined as*

$$\begin{aligned}
&\mathcal{L}_S(\vec{\mathsf{s}}, \Psi) = \\
&\quad\Bigg[\nabla_{\vec{\mathsf{s}}(x)}\Psi(\vec{\mathsf{s}}(x), z)^\top \vec{\mathsf{s}}_x(x)\mu_\pi(x) \\
&\quad - (r(x) + \log\gamma z)\frac{\partial}{\partial z}\Psi(\vec{\mathsf{s}}(x), z) \\
&\quad + \frac{1}{2}\,\mathsf{Tr}\left(\boldsymbol{\sigma}_\pi(x)^\top\left(\mathbf{K}^x_\Phi(x,z) + \mathbf{K}^s_\Phi(x,z)\right)\boldsymbol{\sigma}_\pi(x)\right)\Bigg]^2
\end{aligned} \tag{10}$$

*where $\vec{\mathsf{s}}_x \triangleq \mathsf{J}_x\vec{\mathsf{s}}$. Let the assumptions of Corollary 1 hold. Then if $F_\eta$ satisfies* (9) *and $F_{\eta(x)} = \Phi(\vec{\mathsf{s}}(x))$ for each $x \in \mathcal{X}$, we have*

$$\mathcal{L}_S(\vec{\mathsf{s}}, \Phi) = 0 \tag{11}$$

*Proof.* Suppose $F_\eta$ satisfies (9). Then, making the substitution $F_\eta(x, z) = \Psi(\vec{\mathsf{s}}(x), z)$ in (9), we have

$$\begin{aligned}
0 &= \langle\nabla_x\Psi(\vec{\mathsf{s}}(x), z), \mu_\pi(x)\rangle - (r(x) + z\log\gamma)\frac{\partial}{\partial z}\Psi(\vec{\mathsf{s}}(x), z) + \mathsf{Tr}\left(\boldsymbol{\sigma}_\pi(x)^\top\mathsf{H}_x\Psi(\vec{\mathsf{s}}(x), z)\boldsymbol{\sigma}_\pi(x)\right) \\
&= \nabla_{\vec{\mathsf{s}}(x)}\Psi(\vec{\mathsf{s}}(x), z)^\top\left(\mathsf{J}_x\vec{\mathsf{s}}(x)\right)\mu_\pi(x) - (r(x) + z\log\gamma)\frac{\partial}{\partial z}\Psi(\vec{\mathsf{s}}(x), z) \\
&\quad + \mathsf{Tr}\left[\boldsymbol{\sigma}_\pi(x)^\top\overbrace{\mathsf{J}_x\left(\nabla_{\vec{\mathsf{s}}(x)}\Psi(\vec{\mathsf{s}}(x), z)^\top\mathsf{J}_x\vec{\mathsf{s}}(x)\right)}^{(a)}\boldsymbol{\sigma}_\pi(x)\right]
\end{aligned}$$

All of the differential quantities above exist almost everywhere due to Assumption 2 and the hypothesis that $\Phi$ is statistically smooth. It remains only to compute $(a)$. We have

$$\begin{aligned}
(a) &= \mathsf{J}_x\left(\nabla_{\vec{\mathsf{s}}(x)}\Psi(\vec{\mathsf{s}}(x), z)^\top\mathsf{J}_x\vec{\mathsf{s}}(x)\right) \\
&= \mathsf{J}_x\vec{\mathsf{s}}(x)^\top\left(\mathsf{H}_{\vec{\mathsf{s}}(x)}\Psi(\vec{\mathsf{s}}(x), z)\right)\mathsf{J}_x\vec{\mathsf{s}}(x) + \nabla_{\vec{\mathsf{s}}(x)}\Psi(\vec{\mathsf{s}}(x), z)^\top\mathsf{J}_x\mathsf{J}_x\vec{\mathsf{s}}(x) \\
&= \mathsf{J}_x\vec{\mathsf{s}}(x)^\top\mathsf{H}_{\vec{\mathsf{s}}(x)}\Psi(\vec{\mathsf{s}}(x), z)\mathsf{J}_x\vec{\mathsf{s}}(x) + \sum_{k=1}^N\frac{\partial}{\partial\iota_k\vec{\mathsf{s}}(x)}\Psi(\vec{\mathsf{s}}(x), z)\mathsf{H}_x\iota_k\vec{\mathsf{s}}(x) \\
&= \mathsf{J}_x\vec{\mathsf{s}}(x)^\top\mathsf{H}_{\vec{\mathsf{s}}(x)}\Phi(\vec{\mathsf{s}}(x), [V_{\min}, z])\mathsf{J}_x\vec{\mathsf{s}}(x) + \sum_{k=1}^N\frac{\partial}{\partial\iota_k\vec{\mathsf{s}}(x)}\Phi(\vec{\mathsf{s}}(x), [V_{\min}, z])\mathsf{H}_x\iota_k\vec{\mathsf{s}}(x) \\
&= \mathbf{K}^s_\Phi(x, z) + \mathbf{K}^x_\Phi(x, z)
\end{aligned}$$

Substituting this into $(a)$ above, we arrive at the desired result. □

**Corollary 2** (The Quantile HJB Equation for Policy Evaluation)**.** *Let $\eta^\pi$ be statistically smooth, and let $\mathsf{s}$ be the sketch that maps $\eta(x)$ to a quantile distribution for each $x \in \mathcal{X}$.*

*If $\vec{\mathsf{s}}(x) = \mathsf{s}(\eta^\pi(x))$ for each $x \in \mathcal{X}$ and $F_{\eta^\pi}$ is a distributional solution to* (9)*, then sketch $\vec{\mathsf{s}}$ of the statistical functionals $\{s_k\}_{k=1}^N$ is a distributional solution to the following system of PDEs,*

$$\begin{cases}
\langle\nabla_x\iota_k\vec{\mathsf{s}}(x), \mu_\pi(x)\rangle + r(x) + \log\gamma\iota_k\vec{\mathsf{s}}(x) \\
\qquad + \frac{1}{2}\,\mathsf{Tr}\left(\boldsymbol{\sigma}_\pi(x)^\top\mathsf{H}_x\iota_k\vec{\mathsf{s}}(x)\boldsymbol{\sigma}_\pi(x)\right) = 0 \\
\iota_k\vec{\mathsf{s}}(x) = s_k(\eta(x)) \\
\qquad k = 1, \ldots, N
\end{cases} \tag{12}$$

*Proof.* We consider the case where $\Phi$ imputes the statistics $\vec{\mathsf{s}}(x)$ to a quantile distribution. Let $\phi : \mathcal{X} \times \mathcal{R} \to \mathbf{R}$ be an arbitrary test function in the Schwartz class $\mathscr{S}$, and let $\eta = \Phi(\vec{\mathsf{s}}(x))$ such that $F_\eta$ is a distributional solution to (9). For brevity, denote $\mathcal{Y} = \mathcal{X} \times \mathcal{R}$. Denote by $\vartheta : \mathbf{R} \to [0, 1]$ the Heaviside step function $\vartheta(z) = \mathbf{1}_{[z>0]}$. Then, we have that

$$\begin{aligned}
0 &= \int_{\mathcal{Y}} \Bigg[\phi(x,z) \left\langle \nabla_x \sum_{k=1}^{N} \vartheta(z - \iota_k \vec{\mathsf{s}}(x)), \mu_\pi(x) \right\rangle - \phi(x,z)(r(x) + z\log\gamma)\frac{\partial}{\partial z}\sum_{k=1}^{N}\vartheta(z - \iota_k\vec{\mathsf{s}}(x)) \\
&\qquad + \frac{1}{2}\phi(x,z)\,\mathsf{Tr}\left(\boldsymbol{\sigma}_\pi(x)^\top \left(\mathsf{H}_x \sum_{k=1}^{N}\vartheta(z - \iota_k\vec{\mathsf{s}}(x))\right)\boldsymbol{\sigma}_\pi(x)\right)\Bigg] dzdx \\
&= \int_{\mathcal{Y}} \Bigg[ \left\langle \phi(x,z)\nabla_x \sum_{k=1}^{N} \vartheta(z - \iota_k \vec{\mathsf{s}}(x)), \mu_\pi(x) \right\rangle - \phi(x,z)(r(x) + z\log\gamma)\frac{\partial}{\partial z}\sum_{k=1}^{N}\vartheta(z - \iota_k\vec{\mathsf{s}}(x)) \\
&\qquad + \frac{1}{2}\,\mathsf{Tr}\left(\boldsymbol{\sigma}_\pi(x)^\top \phi(x,z)\left(\mathsf{H}_x \sum_{k=1}^{N}\vartheta(z - \iota_k\vec{\mathsf{s}}(x))\right)\boldsymbol{\sigma}_\pi(x)\right)\Bigg] dzdx
\end{aligned}$$

Taking distributional derivatives once, the Heaviside step functions are transformed into Dirac distributions, yielding

$$\begin{aligned}
0 = \int_{\mathcal{Y}} \Bigg[ &\left\langle -\phi(x,z)\sum_{k=1}^{N}\delta_{\iota_k\vec{\mathsf{s}}(x)}(z)\nabla_x\iota_k\vec{\mathsf{s}}(x), \mu_\pi(x)\right\rangle - \phi(x,z)(r(x) + z\log\gamma)\sum_{k=1}^{N}\delta_{\iota_k\vec{\mathsf{s}}(x)}(z) \\
&- \frac{1}{2}\,\mathsf{Tr}\left(\boldsymbol{\sigma}_\pi(x)^\top\phi(x,z)\left(\nabla_x\sum_{k=1}^{N}\delta_{\iota_k\vec{\mathsf{s}}(x)}(z)\right)\nabla_x\iota_k\vec{\mathsf{s}}(x)\boldsymbol{\sigma}_\pi(x)\right)\Bigg] dzdx
\end{aligned}$$

Next, we carry out the second spatial derivative.

$$\begin{aligned}
0 &= \int_{\mathcal{Y}} \Bigg[ \left\langle -\phi(x,z)\sum_{k=1}^{N}\delta_{\iota_k\vec{\mathsf{s}}(x)}(z)\nabla_x\iota_k\vec{\mathsf{s}}(x), \mu_\pi(x)\right\rangle - \phi(x,z)(r(x) + z\log\gamma)\sum_{k=1}^{N}\delta_{\iota_k\vec{\mathsf{s}}(x)}(z) \\
&\qquad - \frac{1}{2}\,\mathsf{Tr}\left(\boldsymbol{\sigma}_\pi(x)^\top\phi(x,z)\left(\nabla_x\sum_{k=1}^{N}\delta_{\iota_k\vec{\mathsf{s}}(x)}(z)\nabla_x\iota_k\vec{\mathsf{s}}(x)\right)\boldsymbol{\sigma}_\pi(x)\right)\Bigg] dzdx \\
&= \int_{\mathcal{Y}} \Bigg[ \left\langle -\phi(x,z)\sum_{k=1}^{N}\delta_{\iota_k\vec{\mathsf{s}}(x)}(z)\nabla_x\iota_k\vec{\mathsf{s}}(x), \mu_\pi(x)\right\rangle - \phi(x,z)(r(x) + z\log\gamma)\sum_{k=1}^{N}\delta_{\iota_k\vec{\mathsf{s}}(x)}(z) \\
&\qquad - \frac{1}{2}\,\mathsf{Tr}\left(\boldsymbol{\sigma}_\pi(x)^\top\phi(x,z)\sum_{k=1}^{N}\Big[\nabla_x\delta_{\iota_k\vec{\mathsf{s}}(x)}(z)\nabla_x\iota_k\vec{\mathsf{s}}(x) + \delta_{\iota_k\vec{\mathsf{s}}(x)}(z)\mathsf{H}_x\iota_k\vec{\mathsf{s}}(x)\Big]\boldsymbol{\sigma}_\pi(x)\right)\Bigg] dzdx \\
&= \int_{\mathcal{Y}} \phi(x,z)\Bigg[ \left\langle \sum_{k=1}^{N}\delta_{\iota_k\vec{\mathsf{s}}(x)}(z)\nabla_x\iota_k\vec{\mathsf{s}}(x), \mu_\pi(x)\right\rangle + (r(x) + z\log\gamma)\sum_{k=1}^{N}\delta_{\iota_k\vec{\mathsf{s}}(x)}(z) \\
&\qquad + \frac{1}{2}\,\mathsf{Tr}\left(\boldsymbol{\sigma}_\pi(x)^\top\sum_{k=1}^{N}\delta_{\iota_k\vec{\mathsf{s}}(x)}(z)\mathsf{H}_x\iota_k\vec{\mathsf{s}}(x)\boldsymbol{\sigma}_\pi(x)\right)\Bigg] dzdx \\
&\qquad + \frac{1}{2}\overbrace{\int_{\mathcal{Y}}\mathsf{Tr}\left(\boldsymbol{\sigma}_\pi(x)^\top\phi(x,z)\nabla_x\delta_{\iota_k\vec{\mathsf{s}}(x)}(z)\nabla_x\vec{\mathsf{s}}(x)\boldsymbol{\sigma}_\pi(x)\right)dzdx}^{(a)}
\end{aligned}$$

We isolate the term $(a)$ as it involves the (distributional) derivative of the Dirac distribution, which is a strange object. However, since our equation holds for any test function $\phi$, we will show that, with the right choice of test function, $(a) = 0$.

Choose any $\overline{x} \in \mathcal{X}$ and let $\varepsilon > 0$. Then let $\phi(x, z) = \varrho_\varepsilon(x)\psi(z)$ where $\varrho_\varepsilon : \mathcal{X} \to \mathbf{R}$ and $\psi : \mathcal{R} \to \mathbf{R}$ are members of the Schwartz class $\mathscr{S}$. We define $\varrho_\varepsilon(x)$ as follows,

$$\varrho_\varepsilon(x) = \frac{1}{\varepsilon\sqrt{\pi}} \exp\left(-\frac{\|x-\overline{x}\|^2}{\varepsilon^2}\right)$$

It is well known that $\varrho_\varepsilon$ is a Schwartz function (Lax & Sons, 2002). Moreover, since $\nabla_x \varrho_\varepsilon(\overline{x}) = 0$ and $\varrho_\varepsilon$ is smooth, we can find a neighborhood $B$ of $\overline{x}$ so small that $\sup_{x_1,x_2\in B}\|x_1-x_2\| \leq \varepsilon$. We are left with

$$\begin{aligned}
(a) &= \lim_{\varepsilon\to 0}\Bigg[\int_B \int_{\mathcal{R}} \mathsf{Tr}\left(\boldsymbol{\sigma}_\pi(x)^\top \phi(x,z) \nabla_x \delta_{\iota_k \vec{\mathsf{s}}(x)}(z) \nabla_x \iota_k \vec{\mathsf{s}}(x) \boldsymbol{\sigma}_\pi(x)\right) dz dx \\
&\qquad + \int_{\mathcal{X}\setminus B} \int_{\mathcal{R}} \mathsf{Tr}\left(\boldsymbol{\sigma}_\pi(x)^\top \phi(x,z) \nabla_x \delta_{\iota_k \vec{\mathsf{s}}(x)}(z) \nabla_x \iota_k \vec{\mathsf{s}}(x) \boldsymbol{\sigma}_\pi(x)\right) dz dx\Bigg] \\
&= \lim_{\varepsilon\to 0}\Bigg[-\overbrace{\int_B \int_{\mathcal{R}} \mathsf{Tr}\left(\boldsymbol{\sigma}_\pi(x)^\top \psi(z) \nabla_x \varrho_{\varepsilon(x)} \delta_{\iota_k \vec{\mathsf{s}}(x)}(z) \nabla_x \iota_k \vec{\mathsf{s}}(x) \boldsymbol{\sigma}_\pi(x)\right) dz dx}^{\mathcal{M}_\varepsilon} \\
&\qquad - \overbrace{\int_{\mathcal{X}\setminus B} \int_{\mathcal{R}} \mathsf{Tr}\left(\boldsymbol{\sigma}_\pi(x)^\top \psi(z) \nabla_x \varrho_\varepsilon(x) \delta_{\iota_k \vec{\mathsf{s}}(x)}(z) \nabla_x \iota_k \vec{\mathsf{s}}(x) \boldsymbol{\sigma}_\pi(x)\right) dz dx}^{\mathcal{E}_\varepsilon}\Bigg]
\end{aligned}$$

It is also well-known $\lim_{\varepsilon\to 0}\varrho_\varepsilon = \delta_{\overline{x}}$ (Lax & Sons, 2002). Since necessarily $\overline{x} \notin \mathcal{X}\setminus B$, the term $\mathcal{E}_\varepsilon$ vanishes. Given that $\sup_{x_1,x_2\in B}\|x_1-x_2\| \leq \varepsilon$, we have

$$\begin{aligned}
|\mathcal{M}_\varepsilon| &\leq \varepsilon \sup_{x\in B}\left|\int_{\mathcal{R}} \mathsf{Tr}\left(\boldsymbol{\sigma}_\pi(x)^\top \psi(z) \delta_{\iota_k \vec{\mathsf{s}}(x)}(z) \nabla_x \iota_k \vec{\mathsf{s}}(x) \boldsymbol{\sigma}_\pi(x)\right) dz\right| \\
&= \varepsilon \sup_{x\in B}\left|\mathsf{Tr}\left(\boldsymbol{\sigma}_\pi(x)^\top \psi(\iota_k \vec{\mathsf{s}}(x)) \nabla_x \iota_k \vec{\mathsf{s}}(x) \boldsymbol{\sigma}_\pi(x)\right) dz\right|
\end{aligned}$$

By the assumption that $\vec{\mathsf{s}}(x)$ is almost-everywhere differentiable, the supremum above is bounded for almost every $\overline{x}$, and it follows that $|\mathcal{M}_\varepsilon| \to 0$ almost surely.

We are left with the following equation:

$$\begin{aligned}
0 = \lim_{\varepsilon\to 0}\int_{\mathcal{X}}\int_{\mathcal{R}} \varrho_\varepsilon(x)\psi(z)\sum_{k=1}^N \delta_{\iota_k \vec{\mathsf{s}}(x)}(z)\Bigg[&\langle \nabla_x \iota_k \vec{\mathsf{s}}(x), \mu_\pi(x)\rangle + r(x) + z\log\gamma \\
&+\frac{1}{2}\mathsf{Tr}\left(\boldsymbol{\sigma}_\pi(x)^\top \mathsf{H}_x \iota_k \vec{\mathsf{s}}(x) \boldsymbol{\sigma}_\pi(x)\right)\Bigg] dz dx
\end{aligned}$$

Given that $\Phi(\vec{\mathsf{s}}(x))$ is statistically smooth, it is a tempered distribution, so this limit exists. We mentioned previously that $\varrho_\varepsilon \to \delta_{\overline{x}}$, so we have

$$\begin{aligned}
0 = \int_{\mathcal{R}} \psi(z)\sum_{k=1}^N \delta_{\iota_k \vec{\mathsf{s}}(\overline{x})}(z)\Bigg[&\langle \nabla_x \iota_k \vec{\mathsf{s}}(\overline{x}), \mu_\pi(\overline{x})\rangle + r(\overline{x}) + z\log\gamma \\
&+\frac{1}{2}\mathsf{Tr}\left(\boldsymbol{\sigma}_\pi(\overline{x})^\top \mathsf{H}_x \iota_k \vec{\mathsf{s}}(\overline{x}) \boldsymbol{\sigma}_\pi(\overline{x})\right)\Bigg] dz
\end{aligned}$$

It follows by definition that $\Phi(\vec{\mathsf{s}}(x))$ is a distributional solution to

$$0 = \sum_{k=1}^{N} \delta_{\iota_k \vec{\mathsf{s}}(\overline{x})}(z) \left[ \langle \nabla_x \iota_k \vec{\mathsf{s}}(\overline{x}), \mu_\pi(\overline{x}) \rangle + r(\overline{x}) + z \log \gamma + \frac{1}{2} \mathsf{Tr} \left( \boldsymbol{\sigma}_\pi(\overline{x})^\top \mathsf{H}_x \iota_k \vec{\mathsf{s}}(\overline{x}) \boldsymbol{\sigma}_\pi(\overline{x}) \right) \right]$$

Note that the equation above is a sum of weighted Diracs. Thus, the only way for it to be satisfied is if each of the terms in the sum individually vanishes. So, we have shown that for each $k \in [N]$ and almost every $x \in \mathcal{X}$, the statistics function $\iota_k \vec{\mathsf{s}}$ is a distributional solution of

$$0 = \langle \nabla_x \iota_k \vec{\mathsf{s}}(x), \mu_\pi(x) \rangle + r(x) + \iota_k \vec{\mathsf{s}}(x) \log \gamma + \frac{1}{2} \mathsf{Tr} \left( \boldsymbol{\sigma}_\pi(x)^\top \mathsf{H}_x \iota_k \vec{\mathsf{s}}(x) \boldsymbol{\sigma}_\pi(x) \right)$$

This completes the proof. □

### A.2. Solution of the Kolmogorov Backward Equation

Recall the identity presented about the solution of the Kolmogorov Backward Equation as an expectation,

**Theorem 1** (Kolmogorov Backward Equation)**.** *Let $(Y_t)_{t \geq 0} : \mathbf{R}_+ \to \mathcal{Y}$ be a Feller-Dynkin process in some space $\mathcal{Y}$, driven by an infinitesimal generator $\mathscr{L}$. Let $\mathcal{Y}^\circ \subset \mathcal{Y}$ be a measurable set with respect to the Borel $\sigma$-algebra on $\mathcal{Y}$, and let $S \in \mathbf{R}_+$ be the (random) exit time of $(\mathcal{Y}_t)_{t \geq 0}$ from $\mathcal{Y}^\circ$. It is assumed that $Y_0 \in \mathcal{Y}^\circ$ and $\mathsf{P}(S < \infty) = 1$. For any measurable function $\phi$ that is absolutely continuous and differentiable almost everywhere, $u(t, y) = \mathbf{E}[\phi(Y_S) \mid Y_{t \wedge S} = y]$ solves*

$$\frac{\partial u(t,y)}{\partial t} = -\mathscr{L} u(t, y) \tag{3}$$

*with the terminal condition $u(t, y) = \phi(y)$ for all $y \notin \mathcal{Y}^\circ$.*

In order to prove Theorem 1, the following lemma will be handy.

**Lemma 5** ((Le Gall, 2016), Theorem 6.14)**.** *Let $(X_t)_{t \geq 0}$ be a Feller-Dynkin process on a metric space $\mathcal{X}$, and consider functions $h, g \in C_0(\mathcal{X})$. The following two conditions are equivalent:*

1. *$h \in \mathscr{D}(\mathscr{L})$ and $\mathscr{L} h = g$;*
2. *For each $x \in \mathcal{X}$, the process*

$$h(X_t) - \int_0^t g(X_s) ds \,\bigg|\, X_0 = x$$

*is a martingale with respect to the filtration $(\mathcal{F}_t)$.*

*Proof of Theorem 1.* By Lemma 5, we know that the process $\Phi_t = \phi(X_t) - \int_0^t g(X_s) ds$ is a martingale with respect to $(\mathcal{F}_t)$. Let $s < t < T$. By the definition of a martingale, we have

$$\begin{aligned} 0 &= \mathbf{E} \left[ \Phi_T \mid \mathcal{F}_t \right] - \mathbf{E} \left[ \Phi_T \mid \mathcal{F}_s \right] \\ &= \mathbf{E} \left[ h(X_T) + \int_0^T g(X_r) dr \,\middle|\, \mathcal{F}_t \right] - \mathbf{E} \left[ h(X_T) + \int_0^T g(X_r) dr \,\middle|\, \mathcal{F}_s \right] \\ \mathbf{E} \left[ \int_s^t \mathscr{L} h(X_r) dr \,\middle|\, \mathcal{F}_t \right] &= \mathbf{E} \left[ h(X_T) \mid \mathcal{F}_t \right] - \mathbf{E} \left[ h(X_T) \mid \mathcal{F}_s \right] \end{aligned}$$

Dividing through by $t - s$ and taking the limit as $s \uparrow t$,

$$\begin{aligned} \frac{\partial}{\partial s} \mathbf{E} \left[ \phi(X_T) \mid \mathcal{F}_s \right] = \frac{\partial}{\partial s} u(x, s) &\overset{(a)}{=} \mathbf{E} \left[ \frac{\partial}{\partial s} \int_s^t \mathscr{L} \phi(X_r) dr \,\middle|\, \mathcal{F}_t \right] \\ &= -\mathbf{E} \left[ \mathscr{L} \phi(X_r) dr \mid \mathcal{F}_s \right] \\ &\overset{(b)}{=} -\mathscr{L} \mathbf{E} \left[ \phi(X_s) \mid \mathcal{F}_s \right] \\ &= -\mathscr{L} u(x, s) \end{aligned}$$

Step $(a)$ is allowed by the Leibniz integration rule since the infinitesimal generator preserves continuity and $\phi$ is absolutely continuous by assumption. Finally, step $(b)$ is allowed by the linearity of expectation, since $\mathscr{L}$ is a linear operator. □

## B. Further Experiment Details

Figure 2 below demonstrates that the continuous-time algorithm does indeed learn more accurate representations of the return distribution function than QTD.

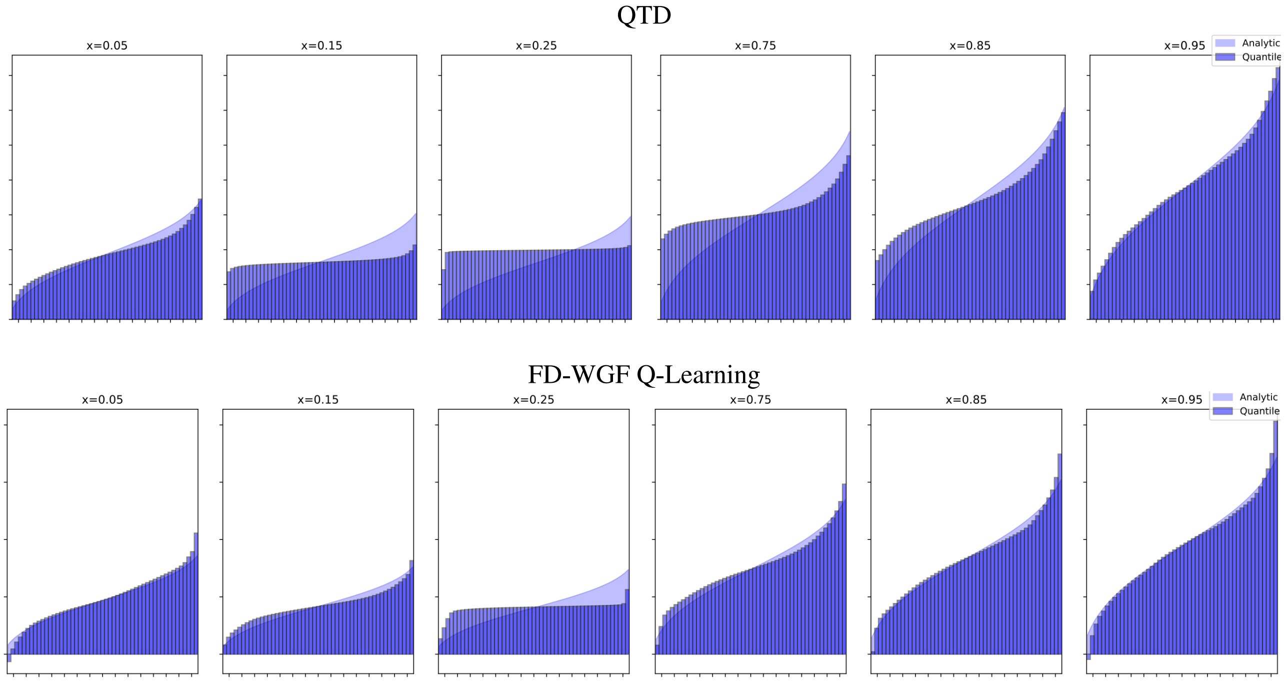

*Figure 3.* Quantile functions learned for the toy problem

## C. Tools from the Theory of Stochastic Processes

This appendix will survey some concepts from the theory of stochastic processes that are useful in the developments of this work.

### C.1. Some Special Classes of Stochastic Processes

#### C.1.1. Measurable, Adapted, and Progressive Processes

When dealing with stochastic processes, there are a few properties that we generally desire in order for us to be able to analyze them nicely. The most common examples will be summarized here. These definitions are due to Le Gall (2016).

For the following definitions, we will fix a probability space $(\Omega, \mathcal{F}, \Pr)$, and we will consider a stochastic process $(X_t)_{t\geq 0} \subset \mathcal{X}$, where $(\mathcal{X}, \Sigma)$ is a measurable space.

**Definition 9** (Measurable Process)**.** *The process $(X_t)_{t\geq 0} \subset \mathcal{X}$ is said to be* measurable *if $(\omega, t) \mapsto X_t(\omega)$ is a measurable map on $\Omega \times \mathbf{R}_+$ with respect to the smallest $\sigma$-algebra on $\mathscr{B}(\mathbf{R}_+) \times \mathcal{F}$.*

For the remainder of the definitions, we will also consider a filtration (see Definition 19) $(\mathcal{F}_t)_{t\geq 0}$ making $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \Pr)$ a filtered probability space.

**Definition 10** (Adapted Process)**.** *The process $(X_t)_{t\geq 0} \subset \mathcal{X}$ is* adapted *if $X_t$ is $\mathcal{F}_t$-measurable for every $t \geq 0$.*

**Definition 11** (Progressive Process)**.** *The process $(X_t)_{t\geq 0} \subset \mathcal{X}$ is* progressive *(or* progressively measurable*) if $(\omega, s) \mapsto X_t(\omega)$ is measurable on $\Omega \times [0, t]$ with respect to the smallest $\sigma$-algebra on $\mathcal{F}_t \times \mathscr{B}([0, t])$ for each $t \geq 0$.*

### C.1.2. Martingales

**Definition 12** (Martingales, (Rogers & Williams, 1994))**.** *A **martingale** (relative to a given filtration $(\mathcal{F}_t)_{t\geq 0}$) is a stochastic process $(M_t)_{t\geq 0}$ where $M_t \in L^1$ and*

$$M_s = \mathbf{E}\left[M_t \mid \mathcal{F}_s\right] \qquad 0 \leq s \leq t \tag{20}$$

*Equation (20) is referred to as "the martingale property". If the equality in (20) is instead $\geq$ (resp. $\leq$), $(M_t)_{t\geq 0}$ is called a **supermartingale** (resp. **submartingale**).*

**Definition 13** (Local Martingales, (Le Gall, 2016))**.** *A **local martingale** is a stochastic process $(M_t)_{t\geq 0}$ for which there exists a sequence of nondecreasing stopping times $(T_n)_{n=1}^{\infty}$ such that $M^{T_n} = (M_{t\wedge T_n})_{t\geq 0} \in L^1$ is a martingale.*

**Definition 14** (Semimartingales, (Le Gall, 2016))**.** *A **semimartingale** is a random process $(X_t)_{t\geq 0}$ such that $X_t = A_t + M_t$ for each $t \geq 0$, where $(A_t)_{t\geq 0}$ is a finite variation process and $(M_t)_{t\geq 0}$ is a local martingale.*

### C.1.3. Finite Variation Processes

**Definition 15** (Finite Variation Function, (Le Gall, 2016))**.** *Let $T \geq 0$. A continuous function $a : [0, T] \to \mathbf{R}$ with $a(0) = 0$ is said to have **finite variation** if there exists a signed measure $\mu$ on $[0, T]$ such that $a(t) = \mu([0, t])$ for any $t \in [0, T]$.*

A finite variation process is a process whose regularity is given by finite variation sample paths, as formalized in the next definition.

**Definition 16** (Finite Variation Process, (Le Gall, 2016))**.** *A process $(A_t)_{t\geq 0}$ is called a **finite variation process** if all of its sample paths are finite variation functions on $\mathbf{R}_+$.*

The following processes generalize the notion of covariance of random variables to stochastic processes, and appear frequently in important stochastic calculus theorems. Their definitions are given by Le Gall (2016).

**Definition 17** (Quadratic Variation)**.** *Let $(M_t)_{t\geq 0}$ be a local martingale. The* quadratic variation *of $(M_t)_{t\geq 0}$, denoted $([M, M]_t)_{t\geq 0}$, is the unique increasing process such that $(M_t^2 - [M, M]_t)_{t\geq 0}$ is a local martingale.*

**Remark 1.** *The existence and uniqueness of the quadratic variation is shown by Le Gall (2016, Theorem 4.9).*

**Definition 18** (The Bracket of Local Martingales)**.** *Let $(M_t)_{t\geq 0}, (N_t)_{t\geq 0}$ be local martingales. The* bracket *of $M, N$, denoted $([M, N]_t)_{t\geq 0}$ is the finite variation process $([M, N]_t)_{t\geq 0}$ given by*

$$[M, N]_t = \frac{1}{2}\bigg([M + N, M + N]_t - [M, M]_t - [N, N]_t\bigg)$$

### C.2. Itô's Lemma

Itô's Lemma is a very powerful tool in the analysis of stochastic processes. It can be thought of as a stochastic analog to Taylor's theorem.

**Theorem 4** (Itô's Lemma, (Le Gall, 2016))**.** *Let $(X^i)_{i=1}^p$ be real valued semimartingales and let $f \in C^2(\mathbf{R})$. Let $\mathbf{X}_t = (X_t^1, \dots, X_t^p)$. Then, for every $t \geq 0$,*

$$f(\mathbf{X}_t) = f(\mathbf{X}_0) + \sum_{i=1}^{p}\int_0^t \frac{\partial f}{\partial x^i}(\mathbf{X}_s)dX_s^i + \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}\int_0^t \frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{X}_s)d[X^i, X^j]_s \tag{21}$$

## D. Continuous-Time Markov Processes

While discrete-time Markov processes are common in the reinforcement learning literature, continuous-time Markov processes (particularly *stochastic* continuous-time Markov processes) are not a trivial extrapolation.

To begin, we recall the definition of a *filtration*, which extends the notion of a $\sigma$-algebra to time-dependent random variables (i.e., stochastic processes).

**Definition 19** (Filtration, (Le Gall, 2016))**.** *Let $(\Omega, \mathcal{F}, \Pr)$ be a probability space. A* filtration *of $\mathcal{F}$ is a collection $(\mathcal{F}_t)_{t\geq 0}$ of $\sigma$-algebras where $\mathcal{F}_t \subset \mathcal{F}$ for each $t$, and $\mathcal{F}_s \subset \mathcal{F}_t$ whenever $s < t$. A probability space associated with a filtration is called a* filtered probability space*, and is written as the 4-tuple $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \Pr)$.*

**Definition 20** (Canonical Filtration, (Le Gall, 2016))**.** *Let $(X_t)_{t\geq 0}$ be a stochastic process on a probability space $(\Omega, \mathcal{F}, \Pr)$. The* canonical filtration *is a filtration $(\mathcal{F}_t)_{t\geq 0}$ where $\mathcal{F}_t$ is the $\sigma$-algebra generated by all observations of the process $(X_t)_{t\geq 0}$ occuring at or before time $t$.*

A Markov process can then be defined as a stochastic process on a filtered probability space that satisfies a Markov property.

**Definition 21** (Markov Process, (Rogers & Williams, 1994))**.** *Let $(X_t)_{t\geq 0}$ be a stochastic process in the filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \Pr)$. A* Markovian transition kernel *$P_t : \Omega \times \mathcal{F} \to [0,1]$ is a transition kernel with a continuous parameter $t$, such that for any bounded $\mathscr{B}(\mathbf{R}_+) \otimes \mathcal{F}$-measurable function $f$, we have*

$$(P_t f)(s, X_s) = \mathbf{E}\left[f(s+t, X_{s+t}) \mid \mathcal{F}_s\right] \qquad \Pr -\textit{almost surely} \tag{22}$$

*A collection $(P_t)_{t\geq 0}$ of Markovian transition kernels is called a* transition semigroup[7] *when*

1. *For each $t \geq 0$ and $x \in \Omega$, $P_t(x, \cdot)$ is a measure on $\mathcal{F}$ and $P_t(x, \Omega) \leq 1$;*
2. *For each $t \geq 0$ and $\Gamma \in \mathcal{F}$, the mapping $P_t(\cdot, \Gamma)$ is $\mathcal{F}$-measurable; and*
3. *(The Chapman-Kolmogorov Identity) For each $s, t \geq 0$,each $x \in \Omega$, and each $\Gamma \in \mathcal{F}$, the collection satisfies*

$$P_{s+t}(x, \Gamma) = \int_{\Omega} P_s(x, dy) P_t(y, \Gamma)$$

*Then $P_t P_s = P_{t+s}$, so $(P_t)_{t\geq 0}$ is indeed a semigroup.*

*A* Markov process *is a stochastic process $(X_t)_{t\geq 0}$ together with a transition semigroup $(P_t)_{t\geq 0}$ such that (22) holds.*

Markov processes with smooth transition kernels are often desirable. This notion is formalized by the following concept.

**Definition 22** (Feller-Dynkin Process, Infinitesimal Generator, (Rogers & Williams, 1994))**.** *Consider a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \Pr)$ and let $\mathcal{X}$ be a Polish[8] space. A transition semigroup $(P_t)_{t\geq 0}$ is said to be a* Feller semigroup *if*

1. *$P_t : C_0(\mathcal{X}) \to C_0(\mathcal{X})$ for each $t \in \mathbf{R}_+$;*
2. *For any $f \in C_0(\mathcal{X})$ with $f \leq 1$, $P_t f \in [0,1]$;*
3. *$P_s P_t = P_{s+t}$ and $P_0 = \mathsf{id}$;*
4. *For any $f \in C_0(\mathcal{X})$, we have $\|P_t f - f\| \xrightarrow{t\downarrow 0} 0$.*

*A Markov process with a Feller semigroup is called a* Feller-Dynkin process.

*Define the set $\mathscr{D}(\mathscr{L})$ according to*

$$\mathscr{D}(\mathscr{L}) = \left\{ f \in C_0(\mathcal{X}) \mid \exists g \in C_0(\mathcal{X}) \quad \textit{such that} \right.$$
$$\left. \left\| \frac{P_\delta - f}{\delta} - g \right\| \xrightarrow{\delta\downarrow 0} 0 \right\}$$

*The* infinitesimal generator *of a Feller-Dynkin process is the operator $\mathscr{L} : \mathscr{D}(\mathscr{L}) \to C_0(\mathcal{X})$ where*

$$\mathscr{L} f = \lim_{\delta \to 0} \frac{P_\delta f - f}{\delta}$$

*and $\mathscr{D}(\mathscr{L})$ is called the* domain of the infinitesimal generator *$\mathscr{L}$.*

[7] This name emphasizes the semigroup nature of the collection of transition kernels. In the abstract algebra literature, a semigroup is a set of objects that is closed under an associative binary operation.

[8] A Polish space is a complete metric space that has a countable, dense subset.

To deal with non-deterministic times in the analysis of a continuous-time Markov process, we recall the formalism of a *stopping time*.

**Definition 23** (Stopping time, (Le Gall, 2016))**.** *Let* $(\Omega, \mathcal{F}, (\mathcal{F}_t))$ *be a measurable space with filtration* $(\mathcal{F}_t)$*. A random variable* $T : \Omega \to \mathbf{R}_+$ *is called a* stopping time *with respect to the filtration* $(\mathcal{F}_t)$ *if*

$$\{T \leq t\} \in \mathcal{F}_t \qquad t \geq 0$$

*We define the* $\sigma$-algebra of the past before $T$ *as the* $\sigma$*-algebra* $\mathcal{F}_T$ *given by*

$$\mathcal{F}_T = \{A \in \mathcal{F}_\infty : A \cap \{T \leq t\} \in \mathcal{F}_t\}$$

### D.1. Brownian Motion

Brownian motion is ubiquitous in the study of stochastic processes. The idea can be motivated as follows.

Let $X_0 \triangleq 0 \in \mathbf{R}$. Suppose we are modeling the trajectory of the random process $(X_t)_{t\geq 0}$, where $X$ is "continuously perturbed" by Gaussian noise with mean 0. What does it mean for something to be *continuously perturbed* by noise? A natural way to reason about this is to discretize time, and suppose that the variable at consecutive timesteps differs by a random quantity sampled independently from a Gaussian with zero mean. We want $X_1$ to have variance 1, and we want this variance to spread evenly through time in the sense that $X_t$ has variance $t$. We can begin with a very coarse discretization where the timestep $\tau$ has duration 1, which involves sampling $X_1 \sim \mathcal{N}(0, 1)$ and interpolate linearly form $t = 0$ to $t = 1$. Then we can study the behavior as $\tau \to 0$. For any $\tau > 0$, we simply sample $X_{t+\tau} \sim X_t + \mathcal{N}(0, \tau)$. Alternatively, we can sample $(X_{k\tau})_{k\in\mathbf{N}}$ via a Gaussian process with covariance kernel $K(X_s, X_t) = \min(s, t)$ (Williams & Rasmussen, 2006). Figure 4 illustrates some of these samples for various values of $\tau$.

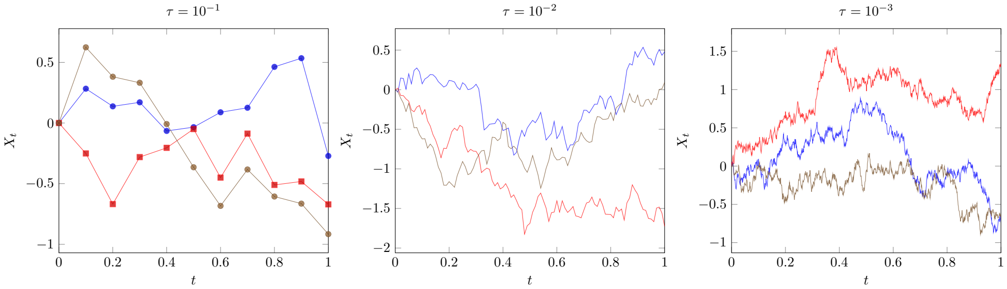

*Figure 4.* Discretized Brownian motion trajectories for various timesteps $\tau$

Considering once again the filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mu)$, the criteria for a Brownian motion $(B_t)_{t\geq 0}$ can be stated formally as

1. $B_0 = 0$, $\mu$-almost surely;
2. For any $0 \leq r < s < t$, the random variable $B_t - B_s$ is independent from $\mathcal{F}_r$ and is distributed according to $\mathcal{N}(0, t - s)$;
3. The *sample paths* of $(B_t)_{t\geq 0}$, defined as the mappings $t \mapsto B_t(\omega)$ for any fixed $\omega \in \mathcal{F}_t$, are continuous.

Proving that such a process exists is not trivial by any means. Fortunately, Brownian motion *does* exist, and Le Gall (2016) can be consulted for its construction.

## E. The Feynman-Kac Formula

We make use of the following formulation of the *Feynman-Kac formula*, as illustrated in Le Gall (2016, Exercise 6.26).

**Theorem 5.** *Let $(X_t)_{t\geq 0}$ be a Feller-Dynkin process in a space $\mathcal{X}$ and let $v \in C_0(\mathcal{X})$. Define for any $x \in \mathcal{X}$ and $\phi$ a bounded and measurable function over $\mathcal{X}$ the transition semigroup $(Q^\star_t)_{t\geq 0}$ where*

$$Q^\star_t \phi(x) = \mathbf{E}\left[\phi(X_t)\exp\left(-\int_0^t v(X_s)ds\right) \,\middle|\, X_0 = x\right]$$

*If $(X_t)_{t\geq 0}$ admits an infinitesimal generator $\mathscr{L}$ and $\phi \in \mathcal{D}(\mathscr{L})$, then*

$$\frac{d}{dt}Q^\star_t\phi|_{t=0} = \mathscr{L}\phi - v \otimes \phi \tag{23}$$

**Remark 2.** *The Feynman-Kac formula can be seen as the Kolmogorov Backward Equation with an "integrating factor". Effectively, the Feynman-Kac formula allows us to identify solutions of PDEs of the form*

$$\frac{\partial u}{\partial t} = -\mathscr{L}u + v \otimes \phi$$

*with conditional expectations of diffusion processes.*

## F. Wasserstein Gradient Flows

Recall that in continuous time, the value function is characterized by a PDE. We should therefore anticipate that the return distribution function will also be characterized by some differential equation. In discrete-time RL algorithms the value function is updated to minimize its difference to the fixed point of the Bellman operator. In the continuous-time limit, this is represented by a *gradient flow* (Santambrogio, 2016),

$$\frac{\partial}{\partial t}\eta_t = -\nabla\mathscr{G}\,\eta_t \tag{24}$$

where $\mathscr{G}$ is a "loss functional" that effectively computes the distance between $\eta_t$ and its fixed point. However, (24) has some glaring problems: the space of probability measures is not a vector space, so neither of the terms in (24) are meaningful. To cope with this, we will consider an alternate form of (24) that can be expressed entirely in terms of metric space properties, called the *Evolution Variational Inequality* (De Giorgi et al., 1980),

$$\frac{1}{2}\frac{\partial}{\partial t}d^2(\mu_t, \nu) \leq \mathscr{G}(\nu) - \mathscr{G}(\mu_t) + \frac{\lambda}{2}d^2(\mu_t, \nu) \tag{EVI$_\lambda$}$$

where $\lambda > 0$ and $\mu_t, \nu$ are elements of an abstract metric space with metric $d$. When the metric space is Euclidean, (EVI$_\lambda$) and (24) are equivalent (Santambrogio, 2016). This characterization of a gradient flow is much more attractive considering the following result.

**Theorem 6** (Muratori & Savaré (2018), Theorem 3.5)**.** *Let $(\mathcal{V}, d)$ be a metric space and suppose $\mathscr{G} : \mathcal{V} \to \mathbf{R}_+$ is $\lambda$-convex. If two curves $\mu, \nu : \mathbf{R}_+ \to \mathcal{V}$ satisfy (EVI$_\lambda$), then*

$$d(\mu_t, \nu_t) \leq e^{-\lambda t}d(\mu_0, \nu_0)$$

*Consequently, for any given initial data $\mu_0 = \varrho$, solutions to (EVI$_\lambda$) must be unique.*

The machinery of abstract gradient flows has been particularly fruitful in the analysis of curves in 2-Wasserstein space. The celebrated work of Jordan et al. (2002) establishes an equivalence between such a Wasserstein gradient flow (WGF) and the *Fokker-Planck equation*,

$$\frac{\partial}{\partial t}\varrho_t(x) = -\nabla\cdot(\varrho_t(x)f(x)) + \beta\Delta\varrho_t(x) \tag{FP}$$

whose solution is the density of the solution to the stochastic differential equation given by

$$dX_t = f_t(X_t)dt + \sqrt{\beta}dB_t$$

where $\beta \in \mathbf{R}_+$ and $(B_t)_{t\geq 0}$ is a Brownian motion (Ambrosio et al., 2008). Ultimately, Jordan et al. (2002) introduces a time-discretized scheme known as the *JKO scheme* for solving PDEs and optimization problems in 2-Wasserstein space. Remarkably, the JKO scheme takes the form of a regularized gradient descent algorithm of a tractable loss, and whose gradients can be estimated from samples without bias. The algorithm is the following generalized minimizing movements (De Giorgi, 1993) scheme:

$$\varrho_{k+1} \in \arg\min_{\varrho} \left\{ D_{\mathrm{KL}}\left(\varrho \parallel \mu\right) + \frac{1}{2\tau} W_2^2(\varrho, \varrho_k) \right\} \tag{JKO}$$

where $\mu(x) \propto \exp(-F_t(x))$, $f_t = \nabla F_t$, $\tau > 0$ is the discretized timestep, and $\varrho_k$ is short for $\varrho_{k\tau}$.

The JKO scheme has made several appearances in the ML literature. Entropically-regularized optimal transport methods, for instance, (Cuturi, 2013) are founded on the JKO scheme. Chizat & Bach (2018) uses the JKO scheme to guarantee convergence to a global optimum when training neural networks without convexity assumptions. In the RL literature, Zhang et al. (2018) employs a JKO scheme to learn a posterior distribution over optimal policies. More akin to the developments in this paper, Martin et al. (2020) presents a novel DRL algorithm where the return distributions are trained as a WGF.

## G. Tempered Distributions

A recurring concept in many areas of mathematics, physics, and engineering is that of *generalized functions*, known as *distributions*[9]. One such example is the Dirac delta. Distributions are particularly helpful at formally describing weakened solutions to PDEs by objects that may not be functions.

In this text, we will make use of the class of *tempered* distributions, which will be defined shortly. For more details, refer to Lax & Sons (2002).

**Definition 24** (Schwartz Class)**.** *Let $X$ be a normed space. A* Schwartz class *is a class $\mathcal{S}$ of rapidly decaying-smooth functions,*

$$\mathcal{S} = \left\{ f \in C^\infty(X; \mathbf{R}) : \sup_{x\in X} (1 + \|x\|^k)|f^{(m)}(x)| < \infty \quad \forall k, m \in \mathbf{N} \right\}$$

**Definition 25** (Tempered Distribution)**.** *A tempered distribution is an element of the topological dual*[10] *$\mathcal{S}'$ of the Schwartz class $\mathcal{S}$.*

**Remark 3.** *The Dirac delta is the operator $\delta$ such that $\langle \delta, \phi \rangle = \phi(0)$. Clearly $\delta$ is linear, and since it is bounded, it is continuous. Therefore $\delta$ is indeed a tempered distribution.*

Tempered distributions admit a notion of differentiability, which can be used to define "distributional" solutions to PDEs.

**Definition 26** (Distributional Derivative)**.** *Let $\mathcal{S}$ be a Schwartz class and $\psi \in \mathcal{S}'$ a tempered distribution. Then $\psi$ has a distributional derivative if there exists a tempered distribution $\psi'$ for which*

$$\langle \psi', \phi \rangle = -\langle \psi, \phi' \rangle \qquad \forall \phi \in \mathcal{S},$$

*and $\psi'$ is called the distributional derivative of $\psi$.*

**Definition 27** (Distributional Solutions of Hamilton-Jacobi PDEs)**.** *Consider the following PDE,*

$$\frac{\partial u}{\partial t} = f \circ u + \langle \nabla u, g \rangle + h^\top \mathsf{H}_y u h \tag{25}$$

*where $u \in C^2(\mathbf{R}_+ \times \mathcal{Y}; \mathbf{R})$ for a normed space $\mathcal{Y}$.*

*Then $\psi \in \mathcal{S}'$ is said to be a* distributional solution *to* (25) *if*

[9] Not to be confused with probability distributions.

[10] The dual of a normed space is the set of all continuous, linear functionals on that space.

$$\begin{aligned}&\int_0^\infty \int_{\mathcal{Y}} \phi(t,y)\left(f(\psi(y)) - \frac{\partial}{\partial t}\psi(y)\right) dydt \\ &\qquad = \int_0^\infty \int_{\mathcal{Y}} \left[\langle \psi(y)g(y), \nabla_y \phi(t,y)\rangle - h(y)^\top \psi(y)\mathsf{H}_y \phi(t,y) h(y)\right] dydt\end{aligned}$$

*for every test function $\phi \in \mathcal{S}$. This is justified by simply multiplying both sides of (25) by the test function, integrating over $\mathbf{R}_+ \times \mathcal{Y}$, and substituting gradient terms of $\psi$ with respect to its distributional derivative.*