Erratum to "q-Learning in Continuous Time"

Yanwei Jia YANWEIJIA@CUHK.EDU.HK
Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong
Shatin, NT, Hong Kong

Xun Yu Zhou XZ2574@COLUMBIA.EDU
Department of Industrial Engineering and Operations Research &
The Data Science Institute
Columbia University
New York, NY 10027, USA

Abstract

Jia and Zhou (2023, *Journal of Machine Learning Research*, 24(161), 1-61) introduce the notion of "q-function" for continuous-time reinforcement learning and characterize q-function and value function by martingale conditions involving action processes that are sampled from the underlying stochastic policy continuously. However, there is a subtle measurability issue in such a continuum independent sampling. This erratum resolves this measure-theoretical issue and provides corrected statements and proofs of the main results in Jia and Zhou (2023). The algorithms and numerical studies in the original paper are not impacted.

Keywords: continuous-time reinforcement learning, q-function, martingale, discrete sampling

1 Introduction

Jia and Zhou (2023, *Journal of Machine Learning Research*, 24(161), 1-61) introduce the q-function for continuous-time reinforcement learning (RL) with controlled diffusion processes, and provide martingale characterizations for learning the q-function and the value function in a data-driven fashion. An implicit assumption in Jia and Zhou (2023) is the possibility of continuum independent sampling from a given admissible feedback policy $\boldsymbol{\pi}$. More precisely, at any time–state pair (t, x), the agent generates an action $a_t \sim \boldsymbol{\pi}(\cdot|t, x)$, and then applies this action to the environment instantaneously. This procedure leads to the time–state–action–reward sequences (all continuous-time processes) $\{s, X_s, a_s, r_s : 0 \leq s \leq T\}$ that satisfy

$$\mathrm{d}X_s = b(s, X_s, a_s)\mathrm{d}s + \sigma(s, X_s, a_s)\mathrm{d}W_s,$$

where

$$a_s \sim \boldsymbol{\pi}(\cdot|s, X_s),\ r_s = r(s, X_s, a_s),\ \forall s \in [0, T].$$

1.1 Measure-Theoretical Issue with Continuum Sampling

The above sampling procedure requires continuum independent draws from a non-degenerate distribution, for which Jia and Zhou (2023) refer to the Fubini extension framework of Sun (2006) that shows it is possible to extend the Lebesgue measure (in t) to accommodate "essentially pairwise independent" continuum random variables. However, there is a gap

in this treatment. Theoretically, the resulting action process $\{a_s : 0 \leq s \leq T\}$ needs to be progressively measurable for the integral $\int_0^T b(t, X_t, a_t)\mathrm{d}t$ and the stochastic integral $\int_0^T \sigma(t, X_t, a_t)\mathrm{d}W_t$ to be well defined. However, Szpruch et al. (2024, Remark 2.1) and Bender and Thuan (2024, Section 3) point out that it is not the case in general.

While this represents a very delicate technical gap, the theoretical results in Jia and Zhou (2023) are so important that we believe an erratum is warranted.

1.2 Discretely Sampled Processes

Szpruch et al. (2024); Bender and Thuan (2024); Jia et al. (2025) all propose using (different versions of) time-discretely sampled action processes to overcome the measurability issue. In this erratum, we take the recent framework of Jia et al. (2025).

Consider another probability space $(\Omega^\xi, \mathcal{F}^\xi, \mathbb{P}^\xi)$ and a measurable function $\phi : [0,T] \times \mathbb{R}^d \times \Omega^\xi \to \mathcal{A}$ such that for all $(t,x) \in [0,T] \times \mathbb{R}^d$, the $\mathcal{A}$-valued random variable $\phi(t,x,\xi)$ has the distribution $\boldsymbol{\pi}(\cdot|t,x)$. Let $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$ and let $(\Omega^{\xi_n}, \mathcal{F}^{\xi_n}, \mathbb{P}^{\xi_n}, \xi_n)_{n\in\mathbb{N}_0}$ be independent copies of $(\Omega^\xi, \mathcal{F}^\xi, \mathbb{P}^\xi, \xi)$. Consider a probability space of the following form:

$$(\Omega, \mathcal{F}, \mathbb{P}) := \left(\Omega^W \times \prod_{n=0}^{\infty} \Omega^{\xi_n}, \mathcal{F}^W \otimes \bigotimes_{n=0}^{\infty} \mathcal{F}^{\xi_n}, \mathbb{P}^W \otimes \bigotimes_{n=0}^{\infty} \mathbb{P}^{\xi_n}\right), \tag{1}$$

where $(\Omega^W, \mathcal{F}^W, \mathbb{P}^W)$ is the probability space where the Brownian motion (representing the environmental noises) lives, and for each $n \in \mathbb{N}_0$, $(\Omega^{\xi_n}, \mathcal{F}^{\xi_n}, \mathbb{P}^{\xi_n})$ supports the random variable ξ_n used to generate the random actions. Moreover, we define the filtration $\mathcal{F}_t := \sigma\{(W_s)_{s\leq t}, (\xi_i)_{t_i \leq t}\}$, which is right continuous and satisfies the usual condition.

Given an admissible feedback policy $\boldsymbol{\pi}$ (see Jia and Zhou 2023, Definition 1 for the precise definition), denoted by $\boldsymbol{\pi} \in \boldsymbol{\Pi}$, and $(t,x) \in [0,T) \times \mathbb{R}^d$, consider a time grid $\mathscr{G}_{t:T} = \{t = s_0 < s_1 < \ldots < s_n = T\}$ of $[t,T]$. We sample actions from $\boldsymbol{\pi}$ *only* at the grid points in $\mathscr{G}_{t:T}$. The corresponding state process satisfies, for all $i = 0, \ldots, n-1$ and all $s \in [s_i, s_{i+1}]$,

$$X_s = X_{s_i} + \int_{s_i}^{s} b(u, X_u, a_{s_i})\mathrm{d}u + \int_{s_i}^{s} \sigma(u, X_u, a_{s_i})\mathrm{d}W_u, \quad \text{with } a_{s_i} = \phi(s_i, X_{s_i}, \xi_i), \tag{2}$$

which will be referred henceforth to as the *discretely sampled state process*.[1] Jia et al. (2025, Lemma 3.1) show that (2) is a well-posed SDE whose solution has a continuous trajectory and is adapted to a smaller filtration $\mathcal{G}_s := \sigma\{(W_u)_{u\leq s}, (\xi_i)_{s_i < s}\}$. In addition, the action process $a_s = \sum_{i=0}^{n-1} \mathbb{1}_{\{s\in[s_i,s_{i+1})\}} a_{s_i}$ is a simple process that is adapted to $\mathcal{F}_s$.

In the following, we denote by $a^{\mathscr{G},\boldsymbol{\pi}}$ the resulting action process and by $X^{\mathscr{G},\boldsymbol{\pi}}$ the solution to (2), given $X_t = x$, associated with the policy $\boldsymbol{\pi}$ and the grid $\mathscr{G}_{t:T}$. For simplicity, we may also rewrite (2) as

$$\mathrm{d}X_s = b(s, X_s, a_{\delta(s)})\mathrm{d}s + \sigma(s, X_s, a_{\delta(s)})\mathrm{d}W_s, \quad s \in [t,T]; \quad X_t = x \tag{3}$$

with $\delta(s) = s_i$ for $s \in [s_i, s_{i+1})$, and $a_s = a_{\delta(s)}$ given in (2).

1. The term "discretely" here is slightly misleading as the state process $\{X_s, t \leq s \leq T\}$ itself is still *continuous* in time s. It is the action that is sampled discretely in time from the policy $\boldsymbol{\pi}$.

2 Martingale Characterizations for q-Learning with Discretely Sampled Processes

We will now state and prove the revised martingale characterizations for q-learning, originally presented in Jia and Zhou (2023), in terms of the discretely sampled state–action processes defined in (3). Note that the definition of the q-function is solely based on the "exploratory problem" (the equations (8) and (9) in Jia and Zhou 2023) and, hence, is independent of any discrete sampling. Moreover, the value function, $J(\cdot,\cdot;\boldsymbol{\pi})$, of a policy $\boldsymbol{\pi}\in\boldsymbol{\Pi}$ is now also based on the exploratory problem, i.e. the equation (9) in Jia and Zhou (2023).[2] However, we will prove that Theorems 6, 7, 9, and 12 in Jia and Zhou (2023) are all valid when "$(X^{\boldsymbol{\pi}},a^{\boldsymbol{\pi}})$" therein (which are not rigorously defined due to the measurability issue) is replaced by "$(X^{\mathscr{G},\boldsymbol{\pi}},a^{\mathscr{G},\boldsymbol{\pi}})$" with any given time grid.

In the following, for reader's convenience, we label the theorems with the same numbers corresponding to those in the original paper Jia and Zhou (2023). For example, Theorem 6 here is the revision of Theorem 6 therein.

The first theorem characterizes the q-function of a given admissible policy, assuming its value function is accessed.

Theorem 6 *Let a policy $\boldsymbol{\pi}\in\boldsymbol{\Pi}$, its value function J and a continuous function $\hat{q}:[0,T]\times\mathbb{R}^d\times\mathcal{A}\to\mathbb{R}$ be given. Then*

(i) *$\hat{q}(t,x,a)=q(t,x,a;\boldsymbol{\pi})$ for all $(t,x,a)\in[0,T]\times\mathbb{R}^d\times\mathcal{A}$ if and only if for all $(t,x)\in[0,T]\times\mathbb{R}^d$ and any time grid $\mathscr{G}_{t:T}$ on $[t,T]$, the following process*

$$e^{-\beta s}J(s,X_s^{\mathscr{G},\boldsymbol{\pi}};\boldsymbol{\pi})+\int_t^s e^{-\beta u}[r(u,X_u^{\mathscr{G},\boldsymbol{\pi}},a_u^{\mathscr{G},\boldsymbol{\pi}})-\hat{q}(u,X_u^{\mathscr{G},\boldsymbol{\pi}},a_u^{\mathscr{G},\boldsymbol{\pi}})]\mathrm{d}u \tag{4}$$

is an $(\{\mathcal{F}_s\}_{s\geq0},\mathbb{P})$-martingale, where $\{X_s^{\mathscr{G},\boldsymbol{\pi}},t\leq s\leq T\}$ is the solution to (3) under $\boldsymbol{\pi}$ with $X_t^{\mathscr{G},\boldsymbol{\pi}}=x$.

(ii) *If $\hat{q}(t,x,a)=q(t,x,a;\boldsymbol{\pi})$ for all $(t,x,a)\in[0,T]\times\mathbb{R}^d\times\mathcal{A}$, then given any $\boldsymbol{\pi}'\in\boldsymbol{\Pi}$, for all $(t,x)\in[0,T]\times\mathbb{R}^d$ and any time grid $\mathscr{G}_{t:T}$ on $[t,T]$, the following process*

$$e^{-\beta s}J(s,X_s^{\mathscr{G},\boldsymbol{\pi}'};\boldsymbol{\pi})+\int_t^s e^{-\beta u}[r(u,X_u^{\mathscr{G},\boldsymbol{\pi}'},a_u^{\mathscr{G},\boldsymbol{\pi}'})-\hat{q}(u,X_u^{\mathscr{G},\boldsymbol{\pi}'},a_u^{\mathscr{G},\boldsymbol{\pi}'})]\mathrm{d}u \tag{5}$$

is an $(\{\mathcal{F}_s\}_{s\geq0},\mathbb{P})$-martingale, where $\{X^{\mathscr{G},\boldsymbol{\pi}'},t\leq s\leq T\}$ is the solution to (3) under $\boldsymbol{\pi}'$ with initial condition $X_t^{\mathscr{G},\boldsymbol{\pi}'}=x$.

(iii) *If there exists $\boldsymbol{\pi}'\in\boldsymbol{\Pi}$ such that for all $(t,x)\in[0,T]\times\mathbb{R}^d$ and any time grid $\mathscr{G}_{t:T}$ on $[t,T]$, (5) is an $(\{\mathcal{F}_s\}_{s\geq0},\mathbb{P})$-martingale with initial condition $X_t^{\mathscr{G},\boldsymbol{\pi}'}=x$, then $\hat{q}(t,x,a)=q(t,x,a;\boldsymbol{\pi})$ for all $(t,x,a)\in[0,T]\times\mathbb{R}^d\times\mathcal{A}$.*

2. In Jia and Zhou (2023), the value function is first defined on the continuously sampled control and state processes (see (7) therein), and then argued to be equivalent to the one based on the exploratory problem. The equation (7) in Jia and Zhou (2023) has the same measurability issue, but can be replaced by discretely sampled processes. Jia et al. (2025) show that the total expected reward under the discretely sampled policy converges to the value function (defined based on the exploratory problem) as the grid size tends to zero.

Moreover, in any of the three cases above, the q-function satisfies

$$\int_{\mathcal{A}} \big[q(t,x,a;\boldsymbol{\pi}) - \gamma \log \boldsymbol{\pi}(a|t,x)\big]\boldsymbol{\pi}(a|t,x)\mathrm{d}a = 0, \ \ \forall (t,x) \in [0,T]\times\mathbb{R}^d. \tag{6}$$

Proof The proof of (6) is the same as that in Jia and Zhou (2023). It suffices to show (i) to (iii). For simplicity, denote

$$\mathcal{L}^a V(t,x) := \frac{\partial V}{\partial t}(t,x) + b(t,x,a)\circ\frac{\partial V}{\partial x}(t,x) + \frac{1}{2}\sigma\sigma^\top(t,x,a)\circ\frac{\partial^2 V}{\partial x^2}(t,x).$$

(i) First of all, we apply Itô's lemma to (3) to obtain

$$\begin{aligned}
&e^{-\beta s}J(s,X_s^{\mathscr{G},\boldsymbol{\pi}};\boldsymbol{\pi}) + \int_t^s e^{-\beta u}\left[r(u,X_u^{\mathscr{G},\boldsymbol{\pi}},a_u^{\mathscr{G},\boldsymbol{\pi}}) - \hat{q}(u,X_u^{\mathscr{G},\boldsymbol{\pi}},a_u^{\mathscr{G},\boldsymbol{\pi}})\right]\mathrm{d}u\\
=&e^{-\beta t}J(t,x) + \int_t^s e^{-\beta u}\left[\mathcal{L}^{a_{\delta(u)}^{\mathscr{G},\boldsymbol{\pi}}}J(u,X_u^{\mathscr{G},\boldsymbol{\pi}}) - \beta J(u,X_u^{\mathscr{G},\boldsymbol{\pi}}) + r(u,X_u^{\mathscr{G},\boldsymbol{\pi}},a_{\delta(u)}^{\mathscr{G},\boldsymbol{\pi}}) - \hat{q}(u,X_u^{\mathscr{G},\boldsymbol{\pi}},a_{\delta(u)}^{\mathscr{G},\boldsymbol{\pi}})\right]\mathrm{d}u\\
&+\int_t^s e^{-\beta u}\frac{\partial J}{\partial x}(u,X_u^{\mathscr{G},\boldsymbol{\pi}})\sigma(u,X_u^{\mathscr{G},\boldsymbol{\pi}},a_{\delta(u)}^{\mathscr{G},\boldsymbol{\pi}})\mathrm{d}W_u\\
=&e^{-\beta t}J(t,x) + \int_t^s e^{-\beta u}\left[q(u,X_u^{\mathscr{G},\boldsymbol{\pi}},a_{\delta(u)}^{\mathscr{G},\boldsymbol{\pi}};\boldsymbol{\pi}) - \hat{q}(u,X_u^{\mathscr{G},\boldsymbol{\pi}},a_{\delta(u)}^{\mathscr{G},\boldsymbol{\pi}})\right]\mathrm{d}u\\
&+\int_t^s e^{-\beta u}\frac{\partial J}{\partial x}(u,X_u^{\mathscr{G},\boldsymbol{\pi}})\sigma(u,X_u^{\mathscr{G},\boldsymbol{\pi}},a_{\delta(u)}^{\mathscr{G},\boldsymbol{\pi}})\mathrm{d}W_u.
\end{aligned}$$

If $\hat{q}(t,x,a) = q(t,x,a;\boldsymbol{\pi})$, then it follows from the moment estimates in Jia et al. (2025, Lemma 3.1) for the dicretely sampled state process $X^{\mathscr{G},\boldsymbol{\pi}}$ that (4) is a martingale.

Conversely, if (4) is a martingale, then

$$\int_t^s e^{-\beta u}\left[q(u,X_u^{\mathscr{G},\boldsymbol{\pi}},a_{\delta(u)}^{\mathscr{G},\boldsymbol{\pi}};\boldsymbol{\pi}) - \hat{q}(u,X_u^{\mathscr{G},\boldsymbol{\pi}},a_{\delta(u)}^{\mathscr{G},\boldsymbol{\pi}})\right]\mathrm{d}u$$

is a martingale for all initial (t,x) and any given time grid $\mathscr{G}_{t:T}$. The same argument in Jia and Zhou (2023), i.e., a martingale with zero quadratic variation has to be a constant, yields that $\mathbb{P}$-almost surely,

$$\int_t^s e^{-\beta u}\left[q(u,X_u^{\mathscr{G},\boldsymbol{\pi}},a_{\delta(u)}^{\mathscr{G},\boldsymbol{\pi}};\boldsymbol{\pi}) - \hat{q}(u,X_u^{\mathscr{G},\boldsymbol{\pi}},a_{\delta(u)}^{\mathscr{G},\boldsymbol{\pi}})\right]\mathrm{d}u = 0$$

for all $s\in[t,T]$. Denote $f(t,x,a) = q(t,x,a;\boldsymbol{\pi}) - \hat{q}(t,x,a)$. We prove $f\equiv 0$ by contradiction by assuming that there exists a triple $(t^*,x^*,a^*)\in[0,T)\times\mathbb{R}^d\times\mathcal{A}$ and $\epsilon>0$ such that $f(t^*,x^*,a^*)>\epsilon$. Because f is continuous, there exists $\delta>0$ such that $f(u,x',a')>\epsilon/2$ for all (u,x',a') with $|u-t^*|\vee|x'-x^*|\vee|a'-a^*|<\delta$. Here "$\vee$" is the maximum operator, i.e., $u\vee v = \max\{u,v\}$.

Now consider a discretely sampled state process, still denoted by $X^{\mathscr{G},\boldsymbol{\pi}}$, starting from (t^*,x^*) with a time grid $\mathscr{G}_{t:T}$ satisfying $t^* = t_0 < t^*+\delta < t_1 < \cdots$. Define

$$\tau = \inf\{u\ge t^* : |u-t^*|>\delta \text{ or } |X_u^{\mathscr{G},\boldsymbol{\pi}}-x^*|>\delta\} = \inf\{u\ge t^* : |X_u^{\mathscr{G},\boldsymbol{\pi}}-x^*|>\delta\}\wedge(t^*+\delta),$$

where "$\wedge$" denotes the minimum operator, i.e., $u \wedge v = \min\{u, v\}$. The continuity of $X^{\mathscr{G},\boldsymbol{\pi}}$ implies that $\tau > t^*$, $\mathbb{P}$-almost surely.

We have already proved that there exists $\Omega_0 \in \mathcal{F}$ with $\mathbb{P}(\Omega_0) = 0$ such that for all $\omega \in \Omega \setminus \Omega_0$, $\int_{t^*}^s e^{-\beta u} f(u, X_u^{\mathscr{G},\boldsymbol{\pi}}(\omega), a_{\delta(u)}^{\mathscr{G},\boldsymbol{\pi}}(\omega))\mathrm{d}u = 0$ for all $s \in [t^*, T]$. It follows from Lebesgue's differentiation theorem that for any $\omega \in \Omega \setminus \Omega_0$,

$$f(s, X_s^{\mathscr{G},\boldsymbol{\pi}}(\omega), a_{\delta(s)}^{\mathscr{G},\boldsymbol{\pi}}(\omega)) = 0, \text{ a.e. } s \in [t^*, \tau(\omega)]. \tag{7}$$

On the other hand, for the grid chosen above, for any $s \in [t^*, \tau(\omega)] \subset [t^*, t^* + \delta]$, $a_{\delta(s)}^{\mathscr{G},\boldsymbol{\pi}}(\omega) = a_{t^*}^{\mathscr{G},\boldsymbol{\pi}}(\omega) = \phi(t^*, x^*, \xi_0(\omega))$. Recall the definition of the admissible policy (Definition 1-(i) in Jia and Zhou 2023), we have

$$\mathbb{P}(\phi(t^*, x^*, \xi_0(\omega)) \in \mathcal{B}_\delta(a^*)) = \int_{\mathcal{B}_\delta(a^*)} \boldsymbol{\pi}(a|t^*, x^*)\mathrm{d}a > 0,$$

where $\mathcal{B}_\delta(a^*) = \{a' \in \mathcal{A} : |a' - a^*| < \delta\}$ is a neighborhood of a^*. Hence there exists $\omega \in \Omega \setminus \Omega_0$ such that for every $s \in [t^*, \tau(\omega)]$,

$$f(s, X_s^{\mathscr{G},\boldsymbol{\pi}}(\omega), a_{\delta(s)}^{\mathscr{G},\boldsymbol{\pi}}(\omega)) = f(s, X_s^{\mathscr{G},\boldsymbol{\pi}}(\omega), \phi(t^*, x^*, \xi_0(\omega))) > \frac{\epsilon}{2} > 0,$$

contradicting (7). This proves that $q(t, x, a; \boldsymbol{\pi}) = \hat{q}(t, x, a)$ for every (t, x, a).

(ii) The proof is parallel to the first part of the proof of (i).

(iii) The proof is parallel to the second part of the proof of (i).

■

The next result underpins both on-policy and off-policy RL algorithms for learning the value function and the q-function *jointly*.

Theorem 7 *Let a policy $\boldsymbol{\pi} \in \boldsymbol{\Pi}$, a function $\hat{J} \in C^{1,2}\big([0,T) \times \mathbb{R}^d\big) \cap C\big([0,T] \times \mathbb{R}^d\big)$ with polynomial growth, and a continuous function $\hat{q} : [0,T] \times \mathbb{R}^d \times \mathcal{A} \to \mathbb{R}$ be given satisfying*

$$\hat{J}(T, x) = h(x), \quad \int_{\mathcal{A}} \big[\hat{q}(t, x, a) - \gamma \log \boldsymbol{\pi}(a|t, x)\big]\boldsymbol{\pi}(a|t, x)\mathrm{d}a = 0, \ \ \forall (t, x) \in [0,T] \times \mathbb{R}^d. \tag{8}$$

Then

(i) $\hat{J}$ and $\hat{q}$ are respectively the value function and the q-function associated with $\boldsymbol{\pi}$ if and only if for all $(t, x) \in [0,T] \times \mathbb{R}^d$ and any time grid $\mathscr{G}_{t:T}$ on $[t, T]$, the following process

$$e^{-\beta s}\hat{J}(s, X_s^{\mathscr{G},\boldsymbol{\pi}}) + \int_t^s e^{-\beta u}[r(u, X_u^{\mathscr{G},\boldsymbol{\pi}}, a_u^{\mathscr{G},\boldsymbol{\pi}}) - \hat{q}(u, X_u^{\mathscr{G},\boldsymbol{\pi}}, a_u^{\mathscr{G},\boldsymbol{\pi}})]\mathrm{d}u \tag{9}$$

is an $(\{\mathcal{F}_s\}_{s\geq 0}, \mathbb{P})$-martingale, where $\{X_s^{\mathscr{G},\boldsymbol{\pi}}, t \leq s \leq T\}$ is the solution to (3) under $\boldsymbol{\pi}$ with $X_t^{\mathscr{G},\boldsymbol{\pi}} = x$.

(ii) *If $\hat{J}$ and $\hat{q}$ are respectively the value function and the q-function associated with $\boldsymbol{\pi}$, then given any $\boldsymbol{\pi}' \in \mathbf{\Pi}$, for all $(t,x) \in [0,T] \times \mathbb{R}^d$ and any time grid $\mathscr{G}_{t:T}$ on $[t,T]$, the following process*

$$e^{-\beta s}\hat{J}(s, X_s^{\mathscr{G},\boldsymbol{\pi}'}) + \int_t^s e^{-\beta u}[r(u, X_u^{\mathscr{G},\boldsymbol{\pi}'}, a_u^{\mathscr{G},\boldsymbol{\pi}'}) - \hat{q}(u, X_u^{\mathscr{G},\boldsymbol{\pi}'}, a_u^{\mathscr{G},\boldsymbol{\pi}'})]\mathrm{d}u \tag{10}$$

is an $(\{\mathcal{F}_s\}_{s\geq 0}, \mathbb{P})$-martingale, where $\{X_s^{\mathscr{G},\boldsymbol{\pi}'}, t \leq s \leq T\}$ is the solution to (3) under $\boldsymbol{\pi}'$ with $X_t^{\mathscr{G},\boldsymbol{\pi}'} = x$.

(iii) *If there exists $\boldsymbol{\pi}' \in \mathbf{\Pi}$ such that for all $(t,x) \in [0,T] \times \mathbb{R}^d$ and any time grid $\mathscr{G}_{t:T}$ on $[t,T]$, (10) is an $(\{\mathcal{F}_s\}_{s\geq 0}, \mathbb{P})$-martingale, where $\{X_s^{\mathscr{G},\boldsymbol{\pi}'}, t \leq s \leq T\}$ is the solution to (3) under $\boldsymbol{\pi}'$ with $X_t^{\mathscr{G},\boldsymbol{\pi}'} = x$, then $\hat{J}$ and $\hat{q}$ are respectively the value function and the q-function associated with $\boldsymbol{\pi}$.*

Moreover, in any of the three cases above, if it holds further that $\boldsymbol{\pi}(a|t,x) = \frac{\exp\{\frac{1}{\gamma}\hat{q}(t,x,a)\}}{\int_{\mathcal{A}}\exp\{\frac{1}{\gamma}\hat{q}(t,x,a)\}\mathrm{d}a}$, then $\boldsymbol{\pi}$ is the optimal policy and $\hat{J}$ is the optimal value function.

Proof

(i) We only prove the "only if" part because the "if" part is straightforward following the same argument as in the proof of Theorem 6.

Define $\hat{r}(t,x,a) := \mathcal{L}^a\hat{J}(t,x) - \beta\hat{J}(t,x)$ and consider the process

$$M_s = e^{-\beta s}\hat{J}(s, X_s^{\mathscr{G},\boldsymbol{\pi}}) - \int_t^s e^{-\beta u}\hat{r}(u, X_u^{\mathscr{G},\boldsymbol{\pi}}, a_u^{\mathscr{G},\boldsymbol{\pi}})\mathrm{d}u.$$

By applying Itô's lemma and arguing similarly to the first part of the proof of Theorem 6-(i), we obtain that M_s is an $(\{\mathcal{F}_s\}_{s\geq 0}, \mathbb{P})$-martingale. As a result, $\int_t^s e^{-\beta u}[r(u, X_u^{\mathscr{G},\boldsymbol{\pi}}, a_u^{\mathscr{G},\boldsymbol{\pi}}) - \hat{q}(u, X_u^{\mathscr{G},\boldsymbol{\pi}}, a_u^{\mathscr{G},\boldsymbol{\pi}}) + \hat{r}(u, X_u^{\mathscr{G},\boldsymbol{\pi}}, a_u^{\mathscr{G},\boldsymbol{\pi}})]\mathrm{d}u$ is an $(\{\mathcal{F}_s\}_{s\geq 0}, \mathbb{P})$-martingale. The same argument as in the second part of the proof of Theorem 6-(i) applies, yielding

$$\begin{aligned}\hat{q}(t,x,a) =& r(t,x,a) + \hat{r}(t,x,a)\\ =& \mathcal{L}^a\hat{J} - \beta\hat{J}(t,x) + r(t,x,a)\\ =& \frac{\partial\hat{J}}{\partial t}(t,x) + H\big(t,x,a,\frac{\partial\hat{J}}{\partial x}(t,x),\frac{\partial^2\hat{J}}{\partial x^2}(t,x)\big) - \beta\hat{J}(t,x)\end{aligned}$$

for every (t,x,a). Now the constraint (8) reads

$$\int_{\mathcal{A}}\left[\frac{\partial\hat{J}}{\partial t}(t,x) + H\big(t,x,a,\frac{\partial\hat{J}}{\partial x}(t,x),\frac{\partial^2\hat{J}}{\partial x^2}(t,x)\big) - \beta\hat{J}(t,x) - \gamma\log\boldsymbol{\pi}(a|t,x)\right]\boldsymbol{\pi}(a|t,x)\mathrm{d}a = 0,$$

for all (t,x), which, together with the terminal condition $\hat{J}(T,x) = h(x)$, is the Feynman–Kac PDE that characterizes the value function under the policy $\boldsymbol{\pi}$ (equation (11) in Jia and Zhou 2023). Therefore, it follows from the uniqueness of the solution to the PDE to conclude that $\hat{J} \equiv J(\cdot,\cdot;\boldsymbol{\pi})$. Moreover, it follows from Theorem 6-(i) that $\hat{q} \equiv q(\cdot,\cdot,\cdot;\boldsymbol{\pi})$.

(ii) This follows immediately from Theorem 6-(ii).

(iii) The proof is parallel to the second part of the proof of (i).

The last conclusion follows from the same argument in Jia and Zhou (2023). ■

The following theorem concerns the optimal value function and optimal q-function.

Theorem 9 *Let a function* $\widehat{J^*} \in C^{1,2}\big([0,T)\times\mathbb{R}^d\big)\cap C\big([0,T]\times\mathbb{R}^d\big)$ *with polynomial growth and a continuous function* $\widehat{q^*}:[0,T]\times\mathbb{R}^d\times\mathcal{A}\to\mathbb{R}$ *be given satisfying*

$$\widehat{J^*}(T,x)=h(x),\quad \int_{\mathcal{A}}\exp\{\frac{1}{\gamma}\widehat{q^*}(t,x,a)\}\mathrm{d}a=1,\ \ \forall(t,x)\in[0,T]\times\mathbb{R}^d. \tag{11}$$

Then

(i) If $\widehat{J^*}$ *and* $\widehat{q^*}$ *are respectively the optimal value function and the optimal q-function, then given any* $\boldsymbol{\pi}\in\mathbf{\Pi}$*, for all* $(t,x)\in[0,T]\times\mathbb{R}^d$ *and any time grid* $\mathscr{G}_{t:T}$ *on* $[t,T]$*, the following process*

$$e^{-\beta s}\widehat{J^*}(s,X_s^{\mathscr{G},\boldsymbol{\pi}})+\int_t^s e^{-\beta u}[r(u,X_u^{\mathscr{G},\boldsymbol{\pi}},a_u^{\mathscr{G},\boldsymbol{\pi}})-\widehat{q^*}(u,X_u^{\mathscr{G},\boldsymbol{\pi}},a_u^{\mathscr{G},\boldsymbol{\pi}})]\mathrm{d}u \tag{12}$$

is an $(\{\mathcal{F}_s\}_{s\geq0},\mathbb{P})$*-martingale, where* $\{X_s^{\mathscr{G},\boldsymbol{\pi}},t\leq s\leq T\}$ *is the solution to* (3) *under* $\boldsymbol{\pi}$ *with* $X_t^{\mathscr{G},\boldsymbol{\pi}}=x$*. Moreover, in this case,* $\widehat{\boldsymbol{\pi}^*}(a|t,x)=\exp\{\frac{1}{\gamma}\widehat{q^*}(t,x,a)\}$ *is the optimal policy.*

(ii) If there exists one $\boldsymbol{\pi}\in\mathbf{\Pi}$ *such that for all* $(t,x)\in[0,T]\times\mathbb{R}^d$ *and any time grid* $\mathscr{G}_{t:T}$ *on* $[t,T]$*,* (12) *is an* $(\{\mathcal{F}_s\}_{s\geq0},\mathbb{P})$*-martingale, then* $\widehat{J^*}$ *and* $\widehat{q^*}$ *are respectively the optimal value function and the optimal q-function.*

Proof

(i) The proof is parallel to the first part of Theorem 6-(i), while the optimality of $\widehat{\boldsymbol{\pi}^*}$ follows from Proposition 8 in Jia and Zhou (2023).

(ii) The second constraint in (11) implies that $\widehat{\boldsymbol{\pi}^*}(a|t,x):=\exp\{\frac{1}{\gamma}\widehat{q^*}(t,x,a)\}$ is a probability density function, and $\widehat{q^*}(t,x,a)=\gamma\log\widehat{\boldsymbol{\pi}^*}(a|t,x)$. So $\widehat{q^*}(t,x,a)$ satisfies the second constraint in (8) with respect to the policy $\widehat{\boldsymbol{\pi}^*}$. When (12) is an $(\{\mathcal{F}_s\}_{s\geq0},\mathbb{P})$-martingale under the given admissible policy $\boldsymbol{\pi}$, it follows from Theorem 7–(iii) that $\widehat{J^*}$ and $\widehat{q^*}$ are respectively the value function and the q-function associated with $\widehat{\boldsymbol{\pi}^*}$. Then the improved policy is $\mathcal{I}\widehat{\boldsymbol{\pi}^*}(a|t,x):=\frac{\exp\{\frac{1}{\gamma}\widehat{q^*}(t,x,a)\}}{\int_{\mathcal{A}}\exp\{\frac{1}{\gamma}\widehat{q^*}(t,x,a)\}\mathrm{d}a}=\exp\{\frac{1}{\gamma}\widehat{q^*}(t,x,a)\}=\widehat{\boldsymbol{\pi}^*}(a|t,x)$. However, Theorem 2 in Jia and Zhou (2023) yields that $\widehat{\boldsymbol{\pi}^*}$ is optimal, completing the proof.

■

The last result below deals with the case of ergodic tasks.

Theorem 12 *Let an admissible policy $\boldsymbol{\pi}$, a number $\hat{V}$, a function $\hat{J} \in C^2\big(\mathbb{R}^d\big)$ with polynomial growth, and a continuous function $\hat{q} : \mathbb{R}^d \times \mathcal{A} \to \mathbb{R}$ be given satisfying*

$$\lim_{T\to\infty} \frac{1}{T}\mathbb{E}[\hat{J}(\tilde{X}_T^{\boldsymbol{\pi}})] = 0, \quad \int_{\mathcal{A}} \big[\hat{q}(x,a) - \gamma \log \boldsymbol{\pi}(a|x)\big]\boldsymbol{\pi}(a|x)\mathrm{d}a = 0, \ \ \forall x \in \mathbb{R}^d, \tag{13}$$

where $\tilde{X}^{\boldsymbol{\pi}}$ follows the exploratory dynamic (the equation (8) in Jia and Zhou (2023)). Then

(i) $\hat{V}$, $\hat{J}$ and $\hat{q}$ are respectively the value, the value function and the q-function associated with $\boldsymbol{\pi}$ if and only if for all $x \in \mathbb{R}^d$ and any time grid $\mathcal{G}_{0:\infty}$, the following process

$$\hat{J}(X_t^{\mathcal{G},\boldsymbol{\pi}}) + \int_0^t [r(X_u^{\mathcal{G},\boldsymbol{\pi}}, a_u^{\mathcal{G},\boldsymbol{\pi}}) - \hat{q}(X_u^{\mathcal{G},\boldsymbol{\pi}}, a_u^{\mathcal{G},\boldsymbol{\pi}}) - \hat{V}]\mathrm{d}u \tag{14}$$

is an $(\{\mathcal{F}_t\}_{t\geq 0}, \mathbb{P})$-martingale, where $\{X_t^{\mathcal{G},\boldsymbol{\pi}}, 0 \leq t < \infty\}$ is the solution to (3) under $\boldsymbol{\pi}$ with $X_0^{\mathcal{G},\boldsymbol{\pi}} = x$.

(ii) If $\hat{V}$, $\hat{J}$ and $\hat{q}$ are respectively the value, value function and the q-function associated with $\boldsymbol{\pi}$, then given any admissible $\boldsymbol{\pi}'$, for all $x \in \mathbb{R}^d$ and any time grid $\mathcal{G}_{0:\infty}$, the following process

$$\hat{J}(X_t^{\mathcal{G},\boldsymbol{\pi}'}) + \int_0^t [r(X_u^{\mathcal{G},\boldsymbol{\pi}'}, a_u^{\mathcal{G},\boldsymbol{\pi}'}) - \hat{q}(X_u^{\mathcal{G},\boldsymbol{\pi}'}, a_u^{\mathcal{G},\boldsymbol{\pi}'}) - \hat{V}]\mathrm{d}u \tag{15}$$

is an $(\{\mathcal{F}_t\}_{t\geq 0}, \mathbb{P})$-martingale, where $\{X_t^{\mathcal{G},\boldsymbol{\pi}'}, 0 \leq t < \infty\}$ is the solution to (3) under $\boldsymbol{\pi}'$ with initial condition $X_0^{\mathcal{G},\boldsymbol{\pi}'} = x$.

(iii) If there exists an admissible $\boldsymbol{\pi}'$ such that for all $x \in \mathbb{R}^d$, (15) is an $(\{\mathcal{F}_t\}_{t\geq 0}, \mathbb{P})$-martingale where $X_0^{\boldsymbol{\pi}'} = x$, then $\hat{V}$, $\hat{J}$ and $\hat{q}$ are respectively the value, value function and the q-function associated with $\boldsymbol{\pi}$.

Moreover, in any of the three cases above, if it holds further that $\boldsymbol{\pi}(a|x) = \frac{\exp\{\frac{1}{\gamma}\hat{q}(x,a)\}}{\int_{\mathcal{A}} \exp\{\frac{1}{\gamma}\hat{q}(x,a)\}\mathrm{d}a}$, then $\boldsymbol{\pi}$ is the optimal policy and $\hat{V}$ is the optimal value.

Proof The proof is parallel to those of Theorems 6 and 7, and hence omitted. ■

Finally, an important remark is that the revision of the theoretical results in this erratum do not impact the algorithms and numerical experiments in Jia and Zhou (2023), because all the algorithms are naturally based on discretely sampled state processes.

Acknowledgments and Disclosure of Funding

The measurability issue was raised in Szpruch et al. (2024); Bender and Thuan (2024), and by conference participants at The First INFORMS Conference on Financial Engineering and FinTech. The proofs of the revised theorems are based on the discussions with Xuefeng Gao and Lingfei Li while they are working on a related paper Gao et al. (2024) that extends the q-learning theory to jump diffusions. All errors are ours.

References

C. Bender and N. T. Thuan. On the grid-sampling limit SDE. *arXiv preprint arXiv:2410.07778*, 2024.

X. Gao, L. Li, and X. Y. Zhou. Reinforcement learning for jump-diffusions, with financial applications. *arXiv preprint arXiv:2405.16449*, 2024.

Y. Jia and X. Y. Zhou. q-Learning in continuous time. *Journal of Machine Learning Research*, 24(161):1–61, 2023.

Y. Jia, D. Ouyang, and Y. Zhang. Accuracy of discretely sampled stochastic policies in continuous-time reinforcement learning. *arXiv preprint arXiv:2503.09981*, 2025.

Y. Sun. The exact law of large numbers via Fubini extension and characterization of insurable risks. *Journal of Economic Theory*, 126(1):31–69, 2006.

L. Szpruch, T. Treetanthiploet, and Y. Zhang. Optimal scheduling of entropy regularizer for continuous-time linear-quadratic reinforcement learning. *SIAM Journal on Control and Optimization*, 62(1):135–166, 2024.