Stochastic Processes and their Applications 194 (2026) 104848

Contents lists available at ScienceDirect

Stochastic Processes and their Applications

journal homepage: www.elsevier.com/locate/spa

Continuous time reinforcement learning: A random measure approach

Christian Bender [a,*], Nguyen Tran Thuan [a,b]

[a] Department of Mathematics, Saarland University, Campus E2 4, 66123 Saarbrücken, Germany
[b] University of Economics Ho Chi Minh City, Ho Chi Minh City, Vietnam

ARTICLE INFO

2000 MSC:
Primary: 60G57
Secondary: 28A33
60H10
93B52
93E35

Keywords:
Exploratory control
Orthogonal martingale measures
Poisson random measures
Reinforcement learning
Weak convergence.

ABSTRACT

We present a random measure approach for modeling exploration, i.e., the execution of measure-valued controls, in continuous-time reinforcement learning with controlled diffusion and jumps. We begin with the case when sampling the randomized control in continuous time takes place on a discrete-time grid and reformulate the resulting SDE as an equation driven by suitable random measures. Our main result is a limit theorem for these random measures as the mesh-size of the sampling grid goes to zero. The resulting limit SDE can be applied for the theoretical analysis of exploratory control problems and for the derivation of learning algorithms.

1. Introduction

Recent years have seen tremendous progress in the development of reinforcement learning (RL) for systems in continuous time and space, which are formulated in the language of stochastic differential equations (SDEs). The articles [34,35] constitute an important starting point for the modeling of exploration of the state space in such a framework. Roughly speaking, the exploration mechanism consists of first choosing a relaxed control (which is a policy with values in the set of probability distributions) and then executing the policy by drawing a sample from the chosen distribution. Based on a heuristic argument using law of large numbers, Wang et al. [34] identify the drift and diffusion coefficient, when averaging over many independent executions of the relaxed control, leading to the *exploratory SDE* in a diffusion setting. Regularizing the cost function by adding a running reward for exploration (e.g., in terms of Shannon entropy as in [34,35]), they come up with a formulation of *exploratory control problems*.

The exploratory control approach of [34] has been generalized in many directions, including a mean-field setting [8,12], regime-switching models [36], and models with jumps [1,10]. A significant part of the literature focuses on exploratory versions of linear-quadratic problems (which are no longer linear-quadratic due to the presence of the regularization term) and on applications to mean-variance portfolio selection, see, e.g., [1,6,12,34–36]. Moreover, alternatives to the Shannon entropy regularization term have been suggested, see [7,12,13,27]. More information about the recent progress in continuous-time RL can be found in the survey article by Zhou [40].

* Corresponding author.
E-mail addresses: bender@math.uni-saarland.de (C. Bender), nguyen@math.uni-saarland.de, thuannt@ueh.edu.vn (N.T. Thuan).

https://doi.org/10.1016/j.spa.2025.104848
Received 19 November 2024; Received in revised form 4 September 2025; Accepted 8 December 2025
Available online 11 December 2025

While the exploratory SDE is tailor-made to adapt the classical dynamic programming approach and to tackle exploratory control by means of a suitable variant of the Hamilton–Jacobi–Bellman (HJB) equation (see [33] for a detailed study of the exploratory HJB equation), it cannot be interpreted as the response of the system to a randomized control (i.e, a sample drawn from a given relaxed control), as highlighted in [18, p.9]. This is due to the averaging effect in its derivation. Hence, trajectories of the exploratory SDE cannot be regarded as observable and, thus, learning algorithms cannot be formulated in terms of (time-discretized) trajectories of the exploratory SDE, see, again [18, p.9].

As a way out, Jia and Zhou [18,19] introduce the *sample state process* as the solution to an SDE, which we call the *sample SDE* in this article, to model the dynamics of the system along a randomized control in continuous time. Based on this SDE and a martingale criterion for optimality in continuous time, they provide continuous-time versions of several learning algorithms (including temporal-difference learning and Q-learning), see also [31] for an overview on learning algorithms in the classical framework of Markov decision processes. In their construction of the sample SDE, an uncountable family $(Z_t)_{t\in[0,T]}$ of independent uniform random variables on the unit cube is employed for the randomization procedure. To avoid some measurability issues in the construction (see, e.g., [29, Proposition 2.1 and Corollary 4.3]), they exploit the theory of rich Fubini extensions [29,30] in the formulation of the sample SDE. However, we will argue in Section 3 under a simplified setting of drift control with additive noise that the sample state process in the framework of rich Fubini extensions is indistinguishable from the solution of the exploratory SDE. Hence, it is subjected to the same limitations for the design of learning algorithms as the exploratory SDE. The key issue here is that the sampling via an uncountable family of (essentially) pairwise independent random variables leads to an averaging effect by Sun's exact law of large numbers in [29].

In order to circumvent the measurability problems and to avoid the averaging effect, we exploit an idea in [32]. Namely, we sample the independent uniform random variables on a finite time-grid only and extend the randomization scheme piecewise constantly to a left-continuous process (which, consequently, becomes predictable). This approach leads to a well-defined SDE, which we call *grid-sampling SDE*. It has a sound interpretation as response of the system to the grid-randomization of a relaxed control. Technically, this is an SDE with random coefficients.

We are mainly interested in the limit dynamics of this grid-sampling SDE, as the mesh-size of the grid tends to zero. To this end, we reformulate it as an SDE with deterministic coefficients driven by appropriate random measures which depend on the grid-sampling randomization process. In this way, the additional randomness for policy execution is moved from the integrand to the integrator. Our main result (Theorem 3 below) implies vague convergence of these grid-dependent random measures, as the grid-size converges to zero. Replacing the grid-dependent random measures by their limit measures, we arrive at the *grid-sampling limit SDE*, which we consider as a natural SDE formulation for RL with state space exploration in continuous time.

Note that we work in a framework with controlled diffusion and controlled jumps in which the SDE under a classical control is driven by a multidimensional Brownian motion and a Poisson random measure. In the "control randomization limit", i.e. in our formulation of the grid-sampling limit SDE, the Brownian motion is replaced by a family of independent white noise martingale measures (in the sense of [23,37]) and the limit Poisson random measures is defined on an extended measurable space to account for the randomization.

Our weak convergence approach extends the derivation of the exploratory dynamics for mean-variance portfolio selection with jumps in [1]. Due to the linear dependence of the diffusion coefficient on the control, the white noise martingale measures do not show up there but are replaced by a high-dimensional Brownian motion (which features additional components to model the control randomization) in the context of [1], see also Example 6.3. However, the limit Poisson random measure is essentially the same as in [1] in our more general situation.

We also mention that recently the framework of Zhou and coauthors [18,19,34] has been extended to the jump-diffusion case by Gao et al. [10]. They derive in [10] the infinitesimal generator of the averaged (over independent policy executions) dynamics heuristically by extending the law of large numbers argument from [34] in order to define an exploratory SDE with jumps. While the jump part features the same structure as in our grid-sampling limit SDE and as in [1], the diffusion part of their exploratory SDE with jumps is driven by a Brownian motion, which can be lower-dimensional than the Brownian motion that drives the original SDE without control randomization.

We finally remark that the grid-sampling limit SDE resembles the classical formulation of relaxed control, see, e.g., [26] for the case of diffusion control or Chapter 13 in [25]. We emphasize, however, that relaxed controls have been introduced as a technical tool for compactification of the control space in the framework of classical control, while the importance of the grid-sampling limit SDE is in its interpretation as limit to the response of the system to randomized controls.

The article is structured as follows: Section 2 presents a classical setting for controlled SDEs with jumps and randomized policies. In Section 3, we discuss the approach using the idealized sampling for randomization in a rich Fubini extension framework. In Section 4, we introduce the grid-sampling SDE and construct in Proposition 4.2 some random measures related to grid sampling and reformulate the grid-sampling SDE as an SDE driven by these random measures. The main limit theorem is stated in Section 5.1, leading to the definition of the grid-sampling limit SDE, which is shown in Section 5.2 to be well-posed under standard Lipschitz conditions.

In Section 6.1, we show how to simplify the grid-sampling limit SDE in the case of coefficients that depend linearly on the control, while, in Section 6.2, we compare the exploratory SDE of [34] and the grid-sampling limit SDE. It turns out that the solutions to both SDEs share the same probability law, although one is derived by averaging out the policy randomization a-priori, while the other one is obtained in a limit, when one adds more and more randomization noise. A main difference is that our limit theorem combined with stability results for SDEs driven by martingale measures (e.g., Chapter 13 in [25]) suggests a joint convergence of SDE and integrator for the grid-sampling limit SDE, while such a result cannot hold for the exploratory SDE.

This difference plays a key role in Section 6.3, where we re-derive the temporal difference TD(0)-algorithm of [17,18] for policy evaluation in continuous time based on the grid-sampling limit SDE. In doing so, we avoid reference to any kind of idealized sampling that requires independent, identically distributed families of random variables indexed by continuous time for control randomization.

The proof of the main result, Theorem 3, will be given in Section 7 and relies on a limit theorem for triangular arrays by Jacod and Shiryaev [15]. The key step of the proof is contained in Proposition 7.4, which implies convergence of the (modified) semimartingale characteristics of the grid-sampling random measures (integrated against a sufficiently large class of test integrands) to the semimartingale characteristics of the limit random measures.

Proofs of the well-posedness of grid-sampling SDEs and of grid-sampling limit SDEs are given in Section 8. Section 9 concludes.

2. Preliminaries

2.1. Notations

Let $\mathbb{N} := \{1, 2, \dots\}$ and $\mathbb{R}^m_0 := \mathbb{R}^m \backslash \{0\}$. For $a, b \in \mathbb{R}$, denote $a \vee b := \max\{a, b\}$ and $a \wedge b := \min\{a, b\}$ as usual. We let $\int_a^b := \int_{(a,b]}$ and $\int_\emptyset = \sum_{i \in \emptyset} := 0$ by convention. Notation log stands for the natural logarithm.

In this article, all vectors are interpreted as column matrices. For a vector x we use $x^{(i)}$ to denote its i-th component. For a matrix A, the entry in the i-th row and j-th column is $A^{(i,j)}$. Notation $A^\top$ stands for the transpose of A. The collection of real matrices of size $m \times p$ is denoted by $\mathbb{R}^{m \times p}$ which is equipped with the Euclidean/Frobenius norm $\|A\|_F := \sqrt{\text{trace}[A^\top A]}$. For $m \in \mathbb{N}$, we denote by I_m the identity matrix of the size $m \times m$.

Let $|\cdot|$ be the Euclidean norm in $\mathbb{R}^m$. The open ball in $\mathbb{R}^m$ centered at 0 with radius $r > 0$ is $B_m(r) := \{x \in \mathbb{R}^m : |x| < r\}$. In $\mathbb{R}^m$ we always use the Borel σ-field $\mathcal{B}(\mathbb{R}^m)$ induced by the Euclidean norm. For $A \in \mathcal{B}(\mathbb{R})$, $\boldsymbol{\lambda}_A$ means the 1-dimensional Lebesgue measure restricted on A.

Let $U \in \mathcal{B}(\mathbb{R}^d)$. Denote by $B_b(U; \mathbb{R}^m)$ the family of all Borel measurable functions $f : U \to \mathbb{R}^m$ satisfying $\|f\|_{B_b(U;\mathbb{R}^m)} := \sup_{u \in U} |f(u)| < \infty$. For $m = 1$, we simply write $B_b(U) := B_b(U; \mathbb{R})$.

Notations $\partial_k f$, $\partial^2_{k,l} f$ stand for usual partial derivatives of f with respect to scalar components. Let ∇f and $\nabla^2 f$ denote the gradient and the Hessian of f respectively. The family $C^2_b(\mathbb{R}^m)$ consists of all twice continuously differentiable and bounded functions $f : \mathbb{R}^m \to \mathbb{R}$ with bounded gradient and Hessian. $C^2_c(\mathbb{R}^m)$ contains all $f \in C^2_b(\mathbb{R}^m)$ with compact support. We let $f \in C^{1,2}([0,T] \times \mathbb{R}^m)$ if f is (resp. twice) continuously differentiable with respect to $t \in [0,T]$ (resp. to $y \in \mathbb{R}^m$) and its partial derivatives are jointly continuous.

Stochastic basis

Let $T \in (0, \infty)$. Assume that $(\Omega, \mathcal{A}, \mathbb{A}, \mathbb{P})$ satisfies the usual conditions, which means that $(\Omega, \mathcal{A}, \mathbb{P})$ is a complete probability space, the filtration $\mathbb{A} = (\mathcal{A}_t)_{t \in [0,T]}$ is right-continuous and $\mathcal{A}_0$ contains all $\mathbb{P}$-null sets. This allows us to assume that every $\mathbb{A}$-adapted local martingale has *càdlàg* (right-continuous with finite left limits) paths. For a random variable ξ, the expectation and conditional expectation given a sub-σ-algebra $\mathcal{G} \subseteq \mathcal{A}$, if it exists under $\mathbb{P}$, is respectively denoted by $\mathbb{E}[\xi]$ and $\mathbb{E}[\xi|\mathcal{G}]$. We also use the notation $\mathbf{L}^p(\mathbb{P}) := \mathbf{L}^p(\Omega, \mathcal{A}, \mathbb{P})$.

We write $\mathcal{P}_\mathbb{A}$ for the predictable σ-field on $\Omega \times [0,T]$ with respect to the filtration $\mathbb{A}$ and say that an $\mathbb{R}^d$-valued stochastic process $X = (X_t)_{t \in [0,T]}$ is $\mathbb{A}$-predictable, if the map $X : \Omega \times [0,T] \to \mathbb{R}^d$ is $\mathcal{P}_\mathbb{A}/\mathcal{B}(\mathbb{R}^d)$-measurable.

For a càdlàg process $X = (X_t)_{t \in [0,T]}$, set $\Delta X_t := X_t - X_{t-}$ for $t \in [0,T]$, where $X_{0-} := X_0$ and $X_{t-} := \lim_{t > s \uparrow t} X_s$ for $t \in (0,T]$. For processes $X = (X_t)_{t \in [0,T]}$, $Y = (Y_t)_{t \in [0,T]}$, we write $X = Y$ to indicate that $X_t = Y_t$ for all $t \in [0,T]$ a.s., and the same meaning applied when the relation "$=$" is replaced by some other relations such as "$\leq$", "$>$", etc. Similarly, for a process $(X_t)_{t \in [0,T]}$ with finite right-limit paths, we write $X_{t+} := \lim_{t < s \downarrow t} X_s$ for $t \in [0,T)$ and set $X_{T+} := X_T$.

We refer to [15] for unexplained notions such as semimartingales, (optional) quadratic covariation $[X, Y]$ and predictable quadratic covariation $\langle X, Y \rangle$ of semimartingales X, Y.

2.2. Controlled SDEs with jumps

We think of the model dynamics as a system with input coefficients (a, b, γ below) that depend on a control (policy) h in feedback form. The output of the system is influenced by the random noise generated by a multivariate Brownian motion B and an independent Poisson random measure N. Here, we assume that (B, N) is defined on a filtered probability space $(\Omega, \mathcal{F}, \bar{\mathbb{F}}, \mathbb{P})$, which satisfies the usual conditions, and note that the filtration $\bar{\mathbb{F}}$ may be larger than the one generated by (B, N). Thus, for a classical (non-randomized) policy h, we end up with the dynamics, for $t \in [0,T]$,

$$\begin{aligned} \mathrm{d}X^h_t &= b(t, X^h_{t-}, h(t, X^h_{t-}))\mathrm{d}t + a(t, X^h_{t-}, h(t, X^h_{t-}))\mathrm{d}B_t \\ &\quad + \int_{0<|z|\leq \mathfrak{r}} \gamma(t, X^h_{t-}, z, h(t, X^h_{t-}))\tilde{N}(\mathrm{d}t, \mathrm{d}z) + \int_{|z|>\mathfrak{r}} \gamma(t, X^h_{t-}, z, h(t, X^h_{t-}))N(\mathrm{d}t, \mathrm{d}z), \end{aligned} \tag{2.1}$$

with initial condition $X^h_0 = x_0 \in \mathbb{R}^m$. The coefficients $b : [0,T] \times \mathbb{R}^m \times \mathbb{R}^d \to \mathbb{R}^m$, $a : [0,T] \times \mathbb{R}^m \times \mathbb{R}^d \to \mathbb{R}^{m \times p}$ and $\gamma : [0,T] \times \mathbb{R}^m \times \mathbb{R}^q_0 \times \mathbb{R}^d \to \mathbb{R}^m$ and the feedback policy $h : [0,T] \times \mathbb{R}^m \to \mathbb{R}^d$ are measurable and assumed to be sufficiently regular to guarantee existence of a unique strong solution. Moreover, $B = (B_t)_{t \in [0,T]}$ is a standard p-dimensional Brownian motion, $N(\mathrm{d}t, \mathrm{d}z)$ is a (possibly inhomogeneous) Poisson random measure independent of B with intensity $\nu(\mathrm{d}t, \mathrm{d}z) = \nu_t(\mathrm{d}z)\mathrm{d}t$ where $(\nu_t(\mathrm{d}z))_{t \in [0,T]}$ is a transition kernel consisting of Lévy measures on $\mathbb{R}^q_0$ (i.e., ν_t is a Borel measure with $\int_{\mathbb{R}^q_0} (|z|^2 \wedge 1)\nu_t(\mathrm{d}z) < \infty$ for all $t \in [0,T]$).

The following Assumption 1 is imposed throughout this article:

Assumption 1. One has for some fixed $\mathfrak{r} \in [0,\infty]$ that

$$\int_0^T \int_{\mathbb{R}_0^q} (|z|^2 \mathbb{1}_{\{0<|z|\le \mathfrak{r}\}} + \mathbb{1}_{\{|z|>\mathfrak{r}\}}) \nu_t(\mathrm{d}z)\mathrm{d}t < \infty. \tag{2.2}$$

The parameter $\mathfrak{r}$ is regarded as the threshold to distinguish between small jumps and large jumps – and, as usual, the small jumps are integrated with respect to the compensated random measure $\tilde{N}(\mathrm{d}t,\mathrm{d}z) := N(\mathrm{d}t,\mathrm{d}z) - \nu_t(\mathrm{d}z)\mathrm{d}t$.

Remark 2.1.

(1) One typically takes $\mathfrak{r} = 1$ which corresponds to the canonical truncation function $z\mathbb{1}_{\{0<|z|\le 1\}}$. However, since the random measures introduced below are handled differently between the "compensated jump part" and the "finite activity jump part", we include here the case $\mathfrak{r} = 0$, which means that the jump part $\int_0^\cdot \int_{\mathbb{R}_0^q} zN(\mathrm{d}t,\mathrm{d}z)$ of the driving (inhomogeneous) Lévy process is of finite activity, and the case $\mathfrak{r} = \infty$ which means that the jump part $\int_0^\cdot \int_{\mathbb{R}_0^q} z\tilde{N}(\mathrm{d}t,\mathrm{d}z)$ is a square integrable martingale.

(2) Note that (2.2) holds for some $\mathfrak{r} \in (0,\infty)$ if and only if (2.2) holds for all $\mathfrak{r} \in (0,\infty)$.

2.3. Randomized policies

A *relaxed* (or, *measure-valued*) *control* in feedback form is a mapping $h\colon [0,T]\times\mathbb{R}^m \to \mathcal{P}r(\mathcal{B}(\mathbb{R}^d))$, where $\mathcal{P}r(\mathcal{B}(\mathbb{R}^d))$ denotes the space of probability measures on the Borel field $\mathcal{B}(\mathbb{R}^d)$. For the execution of a relaxed control, we consider an $\bar{\mathbb{F}}$-predictable stochastic process $\xi = (\xi_t)_{t\in[0,T]}$ independent of (B,N), whose marginal distribution ξ_t is a uniform distribution on $[0,1]^d$ for every $t \in [0,T]$. Such a ξ is called a *randomization process*. We think of a Borel measurable function $\mathbf{h}\colon [0,T]\times\mathbb{R}^m\times[0,1]^d \to \mathbb{R}^d$ as a *randomized control* in feedback form. The actual randomization is performed by plugging a randomization process in the last variable of $\mathbf{h}$. Adapting the terminology in [32] to our setting, we say that a randomized control $\mathbf{h}$ *executes* a relaxed control h, if the random variable $\mathbf{h}(t,x,\xi_t)$ has the distribution $h(t,x)$ for every $t\in[0,T]$ and $x\in\mathbb{R}^m$ (for some, and then for any, randomization process ξ). For a given randomization process ξ, the random field $(\mathbf{h}(t,x,\xi_t))_{t\in[0,T],\,x\in\mathbb{R}^m}$ is called a *ξ-randomized policy* in feedback form.

The crucial property of the randomization process ξ is its predictability which necessarily implies the predictability of the random field $(\mathbf{h}(t,x,\xi_t))_{t\in[0,T],\,x\in\mathbb{R}^m}$. Hence, for a randomized control $\mathbf{h}$ and a randomization process ξ, it makes sense to consider the random coefficient SDE

$$\begin{aligned}\mathrm{d}X_t^{\xi,\mathbf{h}} &= b(t,X_{t-}^{\xi,\mathbf{h}},\mathbf{h}(t,X_{t-}^{\xi,\mathbf{h}},\xi_t))\mathrm{d}t + a(t,X_{t-}^{\xi,\mathbf{h}},\mathbf{h}(t,X_{t-}^{\xi,\mathbf{h}},\xi_t))\mathrm{d}B_t + \int_{0<|z|\le\mathfrak{r}} \gamma(t,X_{t-}^{\xi,\mathbf{h}},z,\mathbf{h}(t,X_{t-}^{\xi,\mathbf{h}},\xi_t))\tilde{N}(\mathrm{d}t,\mathrm{d}z)\\ &\quad + \int_{|z|>\mathfrak{r}} \gamma(t,X_{t-}^{\xi,\mathbf{h}},z,\mathbf{h}(t,X_{t-}^{\xi,\mathbf{h}},\xi_t))N(\mathrm{d}t,\mathrm{d}z).\end{aligned} \tag{2.3}$$

Motivated by the terminology in [18,19], we call the process $(\mathbf{h}(t,X_{t-}^{\xi,\mathbf{h}},\xi_t))_{t\in[0,T]}$ the *action process* of the randomized control $\mathbf{h}$ under ξ-randomization, provided that the SDE (2.3) admits a unique strong solution. The action process is, then, the control which is actually applied by the actor and the solution $X^{\xi,\mathbf{h}}$ of (2.3) can be considered as the response of the system to the ξ-randomized policy $(\mathbf{h}(t,x,\xi_t))_{t\in[0,T],\,x\in\mathbb{R}^m}$ in the sense that $X^{\xi,\mathbf{h}}$ is the observable state variable of the corresponding action process $(\mathbf{h}(t,X_{t-}^{\xi,\mathbf{h}},\xi_t))_{t\in[0,T]}$.

Remark 2.2.

1. We have only fixed the marginal distribution of the randomization process ξ, but not the joint distribution. In particular, ξ_s and ξ_t are, for the moment, not supposed to be independent for $s\neq t$. Two approaches for ξ will be discussed in Section 3 (idealized sampling) and Section 4 (grid-sampling) below.
2. It is well known that for every distribution P on $\mathcal{B}(\mathbb{R}^d)$, there is a measurable function H such that $H(\eta)$ is P-distributed for any uniform random variable η on $[0,1]^d$, see, e.g., the construction in [3, pp. 491–492]. This is one motivation to assume that the marginals of ξ are uniformly distributed. Note, however, that for any vector $(\eta_1,\dots,\eta_d)$ of independent standard Gaussian random variables, the vector $(\Phi(\eta_1),\dots,\Phi(\eta_d))$ is uniformly distributed on $[0,1]^d$. Here, Φ denotes the cumulative distribution function of a standard Gaussian. Hence, changing the marginal distribution of $(\xi_t)_{t\in[0,T]}$, e.g., to a multivariate Gaussian as in [1] does not make any essential difference in the constructions to come.

3. Idealized sampling and rich Fubini extensions

Ideally, the randomization procedure would be performed at each time point t independently of the other time points, leading to the requirement that the family $\xi^* = (\xi_t^*)_{t\in[0,T]}$ consists of independent random variables. Although there is no problem to construct the triplet (B,N,ξ^*) on an appropriate product space, it is known that a family of non-constant independent identically distributed random variables $(\xi_t^*)_{t\in[0,T]}$ cannot be realized in a jointly measurable way with respect to the standard product σ-field. Namely, the map $\xi^*\colon \Omega\times[0,T]\to[0,1]^d$ cannot be $\mathcal{F}\otimes\mathcal{B}([0,T])/\mathcal{B}([0,1]^d)$-measurable, see, e.g., [29, Proposition 2.1] and the detailed discussion on the relevance of the results in [29] for policy execution in [32]. In particular, with this type of idealized sampling, we can never obtain the crucial predictability property of ξ^*, and, hence, it is not clear how to make any good sense of the SDE (2.3) (in the classical way) for a sufficiently large class of ξ^*-randomized policies.

To overcome this measurability issue, several authors, e.g., [9,19], have pointed to the framework of rich Fubini extensions introduced in [29] for defining the sample state process as the solution of a suitable reformulation of (2.3). However, we will discuss in this section that, even with the rich Fubini extension framework, the use of the sample state process may lead to interpretability issues caused by an averaging effect.

Our discussion below elaborates the one in [9]. For simplicity, we consider the one-dimensional case and let $T = 1$ but still write $[0,T]$ and $\int_0^T$ instead of $[0,1]$ and $\int_0^1$, respectively, to distinguish the time- and space-variables. As stochastic integration under the rich Fubini extension setting is beyond the scope of the classical Itô calculus, we only consider the case of drift control with additive noise given in terms of a one-dimensional Brownian motion B. We now discuss how to make sense of an SDE of the form

$$Y_t = y_0 + \int_0^t b(s, Y_s, \mathbf{h}(s, Y_s, \xi_s^*))\mathrm{d}s + \sigma B_t, \quad t \in [0,T], \tag{3.1}$$

where, ideally, $\xi^* = (\xi_t^*)_{t\in[0,T]}$ is a family of independent random variables, which are uniformly distributed on $[0,1]$. Moreover, ξ^* is assumed to be independent of the Brownian motion B. In (3.1), the functions $b\colon [0,T]\times\mathbb{R}\times\mathbb{R}\to\mathbb{R}$ and $\mathbf{h}\colon [0,T]\times\mathbb{R}\times[0,1]\to\mathbb{R}$ are measurable with respect to the standard Borel σ-fields, and $\sigma \geq 0$, $y_0 \in \mathbb{R}$ are constants.

According to [30, Theorem 1], there exist an extension $([0,T],\Lambda,\rho)$ of the Lebesgue probability space $([0,T],\mathcal{B}([0,T]),\boldsymbol{\lambda}_{[0,T]})$ and some probability space $(\Omega_1,\mathcal{F}_1,\mathbb{P}_1)$ such that the product space $([0,T]\times\Omega_1,\Lambda\otimes\mathcal{F}_1,\rho\otimes\mathbb{P}_1)$ has a rich Fubini extension $([0,T]\times\Omega_1,\Lambda\boxtimes\mathcal{F}_1,\rho\boxtimes\mathbb{P}_1)$, i.e., the following properties hold:

(1) There exists a $\Lambda\boxtimes\mathcal{F}_1/\mathcal{B}(\mathbb{R})$-measurable process $\xi^*\colon [0,T]\times\Omega_1\to\mathbb{R}$ such that, for ρ-a.e. $t\in[0,T]$, ξ_t^* is uniformly distributed on $[0,1]$ and independent of ξ_s^* for ρ-a.e. $s\in[0,T]$.
(2) For any $\rho\boxtimes\mathbb{P}_1$-integrable function F, iterated integration is meaningful and

$$\int_{[0,T]\times\Omega_1} F(t,\omega_1)(\rho\boxtimes\mathbb{P}_1)(\mathrm{d}t,\mathrm{d}\omega_1) = \int_{\Omega_1}\left(\int_0^T F(t,\omega_1)\rho(\mathrm{d}t)\right)\mathbb{P}_1(\mathrm{d}\omega_1) = \int_0^T\left(\int_{\Omega_1} F(t,\omega_1)\mathbb{P}_1(\mathrm{d}\omega_1)\right)\rho(\mathrm{d}t),$$

see [30, Definition 3] or [29, Definition 2.2] for the complete statement of (2).

Moreover, we let $(\Omega_2,\mathcal{F}_2,\mathbb{P}_2)$ be a probability space, which carries a one-dimensional Brownian motion B with respect to its own filtration. We consider the usual product space $(\Omega,\mathcal{F},\mathbb{P}) := (\Omega_1\times\Omega_2,\mathcal{F}_1\otimes\mathcal{F}_2,\mathbb{P}_1\otimes\mathbb{P}_2)$ and extend ξ^* and B to mappings on $[0,T]\times\Omega_1\times\Omega_2$ by setting $\xi_t^*(\omega_1,\omega_2) = \xi_t^*(\omega_1)$ and $B_t(\omega_1,\omega_2) = B_t(\omega_2)$, respectively. Then, there exists a ρ-null set $N_\rho\in\Lambda$ such that for every $t\in[0,T]\backslash N_\rho$, ξ_t^* is a random variable (i.e., $\mathcal{F}/\mathcal{B}(\mathbb{R})$-measurable) and, by the product construction, the families $(\xi_t^*)_{t\in[0,T]\backslash N_\rho}$ and $(B_t)_{t\in[0,T]}$ are independent. We note that ξ^* "almost" satisfies the properties required for the ideal sampling procedure mentioned above. However, ξ^* is not $\mathcal{B}([0,T])\otimes\mathcal{F}/\mathcal{B}(\mathbb{R})$-measurable and, hence, not predictable in the usual sense. Instead, ξ^* only satisfies the weaker measurability property with respect to the larger σ-field $(\Lambda\boxtimes\mathcal{F}_1)\otimes\mathcal{F}_2$. While this ξ^* does not qualify as a randomization process in the sense of our definition, SDE (3.1) can be given a rigorous meaning in the framework of rich Fubini extensions, replacing the Lebesgue measure by its extension ρ:

$$Y_t = y_0 + \int_0^t b(s, Y_s, \mathbf{h}(s, Y_s, \xi_s^*))\rho(\mathrm{d}s) + \sigma B_t, \quad t\in[0,T]. \tag{3.2}$$

We first motivate the notion of a solution to this equation, which we call a *sample SDE* (or sample state process in the terminology of [19]). Since $\rho(\{s\}) = \boldsymbol{\lambda}_{[0,T]}(\{s\}) = 0$ for every $s\in[0,T]$, integrals with respect to ρ are continuous as functions in the upper integration limit. Hence, a solution Y to (3.2) should have continuous paths and, then, $t\mapsto Y_t(\omega)$ is $\mathcal{B}([0,T])/\mathcal{B}(\mathbb{R})$-measurable for every $\omega\in\Omega$. Moreover, the function $t\mapsto\xi_t^*(\omega)$ is $\Lambda/\mathcal{B}(\mathbb{R})$-measurable for $\mathbb{P}$-almost every $\omega\in\Omega$ by the definition of the Fubini extension and, thus, $s\mapsto b(s,Y_s(\omega),\mathbf{h}(s,Y_s(\omega),\xi_s^*(\omega)))$ is $\Lambda/\mathcal{B}(\mathbb{R})$-measurable for $\mathbb{P}$-almost every $\omega\in\Omega$. Consequently, the integral in (3.2) "makes sense" pathwise.

Definition 3.1. We say, a map $Y\colon [0,T]\times\Omega\to\mathbb{R}$ is a *solution* to (3.2), if

(i) Y has continuous paths;
(ii) There is a $\mathbb{P}$-null set $\mathcal{N}_0$ such that for every $\omega\in\Omega\backslash\mathcal{N}_0$, the map

$$[0,T]\ni s\mapsto b(s,Y_s(\omega),\mathbf{h}(s,Y_s(\omega),\xi_s^*(\omega)))$$

is ρ-integrable and Eq. (3.2) is satisfied for every $(t,\omega)\in[0,T]\times(\Omega\backslash\mathcal{N}_0)$.

Recalling that the function $t\mapsto\xi_t^*(\omega)$ is $\Lambda/\mathcal{B}(\mathbb{R})$-measurable for $\mathbb{P}$-almost every $\omega\in\Omega$, we can introduce the measures

$$\hat{\rho}(\omega;\mathrm{d}t,\mathrm{d}u) = \delta_{\xi_t^*(\omega)}(\mathrm{d}u)\rho(\mathrm{d}t)$$

on $\Lambda\otimes\mathcal{B}(\mathbb{R})$ for $\mathbb{P}$-almost every $\omega\in\Omega$. If Y is a solution to (3.2), then, by the (classical) Fubini theorem, Y satisfies

$$Y_t = y_0 + \int_{(0,t]\times[0,1]} b(s,Y_s,\mathbf{h}(s,Y_s,u))\hat{\rho}(\mathrm{d}s,\mathrm{d}u) + \sigma B_t, \quad t\in[0,T] \tag{3.3}$$

outside a $\mathbb{P}$-null set.

The next theorem shows that the restriction of the measures $\hat{\rho}(\omega,\cdot)$ to $\mathcal{B}([0,T])\otimes\mathcal{B}(\mathbb{R})$ is nothing but the Lebesgue measure on $[0,T]\times[0,1]$.

Theorem 1. *There is a $\mathbb{P}$-null set $\mathcal{N}$ such that for every $\omega \in \Omega\backslash\mathcal{N}$ and $A \in \mathcal{B}([0,T]) \otimes \mathcal{B}(\mathbb{R})$,*

$$\hat{\rho}(\omega; A) = (\boldsymbol{\lambda}_{[0,T]} \otimes \boldsymbol{\lambda}_{[0,1]})(A),$$

where in slight abuse of notation we write $\boldsymbol{\lambda}_{[0,1]}(B) = \boldsymbol{\lambda}_{\mathbb{R}}(B \cap [0,1])$ for $B \in \mathcal{B}(\mathbb{R})$.

Proof. Let $A = B \times C \in \mathcal{B}([0,T]) \otimes \mathcal{B}(\mathbb{R})$. If $\boldsymbol{\lambda}_{[0,T]}(B) > 0$, then, by Sun's exact law of large numbers [29, Theorem 2.6], for $\mathbb{P}$-almost every $\omega \in \Omega$,

$$\hat{\rho}(\omega; A) = \int_B \mathbb{1}_{\{\xi_t^*(\omega) \in C\}} \rho(\mathrm{d}t) = \int_B \mathbb{E}[\mathbb{1}_{\{\xi_t^* \in C\}}] \rho(\mathrm{d}t) = \boldsymbol{\lambda}_{[0,1]}(C)\rho(B) = (\boldsymbol{\lambda}_{[0,T]} \otimes \boldsymbol{\lambda}_{[0,1]})(A).$$

If $\boldsymbol{\lambda}_{[0,T]}(B) = 0$, then obviously both sides of the previous equation are zero. Thus, we find a $\mathbb{P}$-null set $\mathcal{N}$ such that for every $\omega \in \Omega\backslash\mathcal{N}$, the measures $\hat{\rho}(\omega; \cdot)$ and $\boldsymbol{\lambda}_{[0,T]} \otimes \boldsymbol{\lambda}_{[0,1]}$ coincide on all Cartesian products of subintervals with rational endpoints. Now, Dynkin's π-λ theorem completes the proof. □

Since $(s,u) \mapsto b(s, Y_s(\omega), \mathbf{h}(s, Y_s(\omega), u))$ is $\mathcal{B}([0,T]) \otimes \mathcal{B}(\mathbb{R})$-measurable for any $\omega \in \Omega$, Theorem 1 and (3.3) imply that every solution Y to (3.2) solves

$$Y_t = y_0 + \int_0^t \int_0^1 b(s, Y_s, \mathbf{h}(s, Y_s, u))\mathrm{d}u\mathrm{d}s + \sigma B_t, \quad t \in [0,T], \tag{3.4}$$

$\mathbb{P}$-almost surely. Note that (3.4) coincides with the *exploratory SDE* introduced in [34]. Indeed, if $\mathbf{h}$ executes the relaxed control h and $h(t,x)$ has a density $\dot{h}(t,x,\cdot)$ with respect to the Lebesgue measure for every pair $(t,x) \in [0,T] \times \mathbb{R}$, then (3.4) becomes

$$Y_t = y_0 + \int_0^t \int_{\mathbb{R}} b(s, Y_s, a)\dot{h}(s, Y_s, a)\mathrm{d}a\mathrm{d}s + \sigma B_t, \quad t \in [0,T];$$

cp. Eqs. (6)–(8) in [34]. Summarizing, we have shown the following result.

Theorem 2. *Any solution $(Y_t)_{t\in[0,T]}$ to the sample SDE (3.2) also solves the exploratory SDE (3.4) $\mathbb{P}$-almost surely.*

There are (at least) two ways to interpret this result:

(1) If one interprets the abstract constructions underlying the rich Fubini extension as a genuine sampling mechanism in continuous time, then Theorem 2 can be seen as a justification for working with the exploratory SDE.
(2) If one interprets these abstract constructions, in view of Theorem 1, just as a technically and notationally involved way to re-write integration with respect to the Lebesgue measure, then the sample SDE formulation in the framework of rich Fubini extensions does not provide any additional benefit beyond the exploratory SDE formulation. With this interpretation, the disadvantage of the solution to the exploratory SDE that it "is the *average* of the sample trajectories […] and is in itself *not* a sample trajectory *nor* observable" [18, p. 9] extends to the sample SDE in the rich Fubini framework.

In any case, under standard Lipschitz assumptions on b and $\mathbf{h}$, the exploratory SDE (3.4) will have a unique solution (up to indistinguishability) and this solution will be adapted to the augmented filtration generated by the Brownian motion B (or, even be deterministic, if $\sigma = 0$). Therefore, the solution Y to the sample SDE (3.2) will be stochastically independent of the information generated by $(\xi_t^*)_{t\in[0,T]}$, i.e., the realization of the state process of a randomized control does not depend on the realization of the sampling mechanism in the rich Fubini framework.

We close this section with an illustrative example.

Example 3.2. Consider the special case $b(t,x,u) = u$ and $\mathbf{h}(t,x,u) = \Phi^{-1}(u)$ (where Φ is the cumulative distribution function of a standard Gaussian), i.e., the measure-valued control $h(t,x)$ is (independent of time and state) a standard Gaussian distribution. Since $\int_0^1 \Phi^{-1}(u)\mathrm{d}u = 0$ (the integral is just the expectation of a standard Gaussian), the solution of the exploratory SDE (3.4) is $Y_t^{\exp} = y_0 + \sigma B_t$. We first show that it is not possible to recover (the distribution of) $Y^{\exp}$ as the response of the system to the ξ-randomized policy $\mathbf{h}(t,x,\xi_t)$, for any *predictable* randomization process ξ. Precisely, we consider the SDE (2.3) starting in y_0 with the corresponding coefficient functions $a(t,x,u) = \sigma$, $\gamma(t,x,z,u) = 0$ and b and $\mathbf{h}$ as specified above, i.e.,

$$X_t^{\xi,\mathbf{h}} = y_0 + \int_0^t \Phi^{-1}(\xi_s)\mathrm{d}s + \sigma B_t = Y_t^{\exp} + \int_0^t \Phi^{-1}(\xi_s)\mathrm{d}s.$$

We show that $X^{\xi,\mathbf{h}}$ and $Y^{\exp}$ cannot even be equal in one-dimensional marginal distributions. Otherwise, computing

$$\mathbb{E}[|Y_t^{\exp}|^2] = y_0^2 + \sigma^2 t < \infty, \quad t \in [0,T],$$

and (using that $Y_t^{\exp}$ and $\int_0^t \Phi^{-1}(\xi_s)\mathrm{d}s$ are independent and that $\mathbb{E}[\int_0^t \Phi^{-1}(\xi_s)\mathrm{d}s] = 0$)

$$\mathbb{E}[|X_t^{\xi,\mathbf{h}}|^2] = \mathbb{E}[|Y_t^{\exp}|^2] + \mathbb{E}\left[\left|\int_0^t \Phi^{-1}(\xi_s)\mathrm{d}s\right|^2\right],$$

we conclude that

$$\mathbb{E}\left[\left|\int_0^t \Phi^{-1}(\xi_s)\mathrm{d}s\right|^2\right] = 0, \quad t \in [0,T].$$

Therefore, we find a null set $\mathcal{N}_\Phi \in \mathcal{F}$ such that $\int_0^t \Phi^{-1}(\xi_s)(\omega)\mathrm{d}s = 0$ for every $\omega \notin \mathcal{N}_\Phi$ and $t \in [0,T] \cap \mathbb{Q}$. By continuity, this identity extends to every $\omega \notin \mathcal{N}_\Phi$ and $t \in [0,T]$. Hence, by Lebesgue's differentiation theorem, we obtain for every $\omega \notin \mathcal{N}_\Phi$: $\Phi^{-1}(\xi_s(\omega)) = 0$ for $\boldsymbol{\lambda}_{[0,T]}$-almost every $s \in [0,T]$. Applying the classical Fubini theorem, we conclude that, for $\boldsymbol{\lambda}_{[0,T]}$-almost every $s \in [0,T]$, one has $\Phi^{-1}(\xi_s) = 0$ $\mathbb{P}$-a.s. This clearly contradicts the property that $\Phi^{-1}(\xi_s)$ is a standard Gaussian random variable for every $s \in [0,T]$. Note that this argument does not necessarily require ξ to be predictable, but remains valid under the weaker assumption that $\Phi^{-1} \circ \xi$ is $\mathcal{B}([0,T]) \otimes \mathcal{F}/\mathcal{B}(\mathbb{R})$-measurable.

If we write, however, $X^{\xi^*,\mathbf{h}}$ for the solution of the sample SDE (3.2), then, in the context of this example, by Theorem 2,

$$X_t^{\xi^*,\mathbf{h}} = y_0 + \int_0^t \Phi^{-1}(\xi_s^*)\rho(\mathrm{d}s) + \sigma B_t = Y_t^{\exp}, \quad t \in [0,T],$$

$\mathbb{P}$-almost surely.

We note that the approach in [18,19] crucially relies on the property that the exploratory SDE and the sample state process have the same one-dimensional marginal distributions. This is required to conclude that both SDEs lead to the same expected cost and, thus, to justify that theoretical questions can be studied through the exploratory SDE, see [19, p. 8]. As illustrated by the example, this approach can only work, if one abandons the established predictability assumption on the controls and replaces the Lebesgue measure by its abstract extension ρ, which is part of the rich Fubini construction. This makes an extension beyond the drift control case impossible in the classical framework of stochastic calculus, which relies on joint measurability assumptions of the integrands such as predictability or progressive measurability.

4. The grid-sampling SDE

The problems discussed in Section 3 suggest that the randomization procedure based on (essentially pairwise) independent families $\xi^* = (\xi_t^*)_{t\in[0,T]}$ of uniformly distributed random variables may not be suitable in continuous time. Motivated by [32], who have also pointed to the measurability problem arising from the use of idealized sampling, we replace ξ^* by a piecewise constant interpolation of finitely many independent uniform random variables on a finite grid of $[0,T]$.

Let Π be a partition of $[0,T]$ with grid points $0 = t_0 < t_1 < \cdots < t_n = T$, $n \in \mathbb{N}$. We denote the mesh-size of Π by $|\Pi| := \max_{1\le i\le n} |t_i - t_{i-1}|$ and suppose that the underlying probability space carries an independent family $(\xi_1, \dots, \xi_n)$ of uniforms on $[0,1]^d$ independent of (B, N). For the randomization on the grid Π, we define the *grid-sampling process* $\xi^\Pi = (\xi_t^\Pi)_{t\in[0,T]}$ by

$$\xi_t^\Pi := \sum_{j=1}^n \xi_j \mathbf{1}_{(t_{j-1},t_j]}(t), \quad t \in [0,T].$$

We emphasize that the authors in [32] and [11] have already applied this type of grid-sampling as a substitution for the infeasible idealized sampling when executing Gaussian relaxed policies in the context of linear-quadratic control.

Denote by $\mathbb{F}^\Pi = (\mathcal{F}_t^\Pi)_{t\in[0,T]}$ the right-continuous, augmented version of the filtration generated by (B, N, ξ^Π). Precisely, $\mathbb{F}^\Pi$ is constructed via

$$\mathcal{F}_t^\Pi := \bigcap_{t<s\le T} \hat{\mathcal{F}}_s^\Pi \quad \text{if } t < T, \quad \text{and } \mathcal{F}_T^\Pi := \hat{\mathcal{F}}_T^\Pi,$$

where, for $s \in [0,T]$,

$$\hat{\mathcal{F}}_s^\Pi := \sigma\{B_r, \xi_r^\Pi, N((0,r]\times A) : A \in \mathcal{B}(\mathbb{R}_0^q), r \in [0,s]\} \vee \mathcal{N},$$

and $\mathcal{N}$ is the collection of all $\mathbb{P}$-null sets. Then, ξ^Π is left-continuous and adapted and, thus, is $\mathbb{F}^\Pi$-predictable. Note that $\xi_{t_i}^\Pi = \xi_i$ is $\mathcal{F}_{t_{i-1}}^\Pi$-measurable, but independent of $\mathcal{F}_{(t_{i-1})-}^\Pi$. Moreover, B and $N(\mathrm{d}t, \mathrm{d}z)$ are still a Brownian motion and a Poisson random measure with intensity $\nu_t(\mathrm{d}z)\mathrm{d}t$ with respect to $\mathbb{F}^\Pi$.

By the $\mathbb{F}^\Pi$-predictability of the grid-sampling process ξ^Π, we may consider the SDE (2.3) with $\xi = \xi^\Pi$, which is given for $t \in (t_{i-1}, t_i]$, $i = 1, \dots, n$ by

$$\begin{aligned} X_t^{\Pi,\mathbf{h}} = X_{t_{i-1}}^{\Pi,\mathbf{h}} &+ \int_{t_{i-1}}^t b(s, X_{s-}^{\Pi,\mathbf{h}}, \mathbf{h}(s, X_{s-}^{\Pi,\mathbf{h}}, \xi_{t_i}^\Pi))\mathrm{d}s + \int_{t_{i-1}}^t a(s, X_{s-}^{\Pi,\mathbf{h}}, \mathbf{h}(s, X_{s-}^{\Pi,\mathbf{h}}, \xi_{t_i}^\Pi))\mathrm{d}B_s \\ &+ \int_{(t_{i-1},t]} \int_{0<|z|\le\mathfrak{r}} \gamma(s, X_{s-}^{\Pi,\mathbf{h}}, z, \mathbf{h}(s, X_{s-}^{\Pi,\mathbf{h}}, \xi_{t_i}^\Pi))\tilde{N}(\mathrm{d}s, \mathrm{d}z) + \int_{(t_{i-1},t]} \int_{|z|>\mathfrak{r}} \gamma(s, X_{s-}^{\Pi,\mathbf{h}}, z, \mathbf{h}(s, X_{s-}^{\Pi,\mathbf{h}}, \xi_{t_i}^\Pi))N(\mathrm{d}s, \mathrm{d}z). \end{aligned} \tag{4.1}$$

We call (4.1) the *grid-sampling SDE* for policy $\mathbf{h}$ along the randomization process ξ^Π.

Remark 4.1. Suppose that the randomized control $\mathbf{h}$ is continuous and executes a relaxed control h and that the sampling grid Π is "sufficiently fine". Then, we may consider

$$\mathbf{h}(t_{i-1}, X_{t_{i-1}}^{\Pi,\mathbf{h}}, \xi_{t_i}^\Pi) = \lim_{s\searrow t_{i-1}} \mathbf{h}(s, X_{s-}^{\Pi,\mathbf{h}}, \xi_{t_i}^\Pi)$$

as a "good" approximation to $\mathbf{h}(s, X_{s-}^{\Pi,\mathbf{h}}, \xi_{t_i}^\Pi)$ for $s \in (t_{i-1}, t_i]$. Note that $X_{t_{i-1}}^{\Pi,\mathbf{h}} = X_{(t_{i-1})-}^{\Pi,\mathbf{h}}$ a.s. Thus, $X_{t_{i-1}}^{\Pi,\mathbf{h}}$ is $\mathcal{F}_{(t_{i-1})-}^\Pi$-measurable and, consequently, independent of $\xi_{t_i}^\Pi$. Therefore, we can interpret the approximation $\mathbf{h}(t_{i-1}, X_{t_{i-1}}^{\Pi,\mathbf{h}}, \xi_{t_i}^\Pi)$ in the following way: The actor first chooses the distribution $h(t_{i-1}, X_{t_{i-1}}^{\Pi,\mathbf{h}})$ and, then, the independent uniform random variable $\xi_{t_i}^\Pi$ is generated to sample from this distribution.

For the coefficients b, a, γ in Section 2.2 and a randomized control $\mathbf{h}\colon [0,T]\times\mathbb{R}^m\times[0,1]^d \to \mathbb{R}^d$, we define the Borel functions $b_{\mathbf{h}}\colon [0,T]\times\mathbb{R}^m\times[0,1]^d \to \mathbb{R}^m$, $a_{\mathbf{h}}\colon [0,T]\times\mathbb{R}^m\times[0,1]^d \to \mathbb{R}^{m\times p}$, and $\gamma_{\mathbf{h}}\colon [0,T]\times\mathbb{R}^m\times\mathbb{R}_0^q\times[0,1]^d \to \mathbb{R}^m$ via

$$b_{\mathbf{h}}(s,x,u) := b(s,x,\mathbf{h}(s,x,u)), \quad a_{\mathbf{h}}(s,x,u) := a(s,x,\mathbf{h}(s,x,u)), \quad \gamma_{\mathbf{h}}(s,x,z,u) := \gamma(s,x,z,\mathbf{h}(s,x,u)).$$

Assumption 2. The coefficients $b_{\mathbf{h}}$, $a_{\mathbf{h}}$, $\gamma_{\mathbf{h}}$ satisfy the following integrability condition: The function

$$G_0(s) := \int_{[0,1]^d}\left[|b_{\mathbf{h}}(s,0,u)|^2 + \|a_{\mathbf{h}}(s,0,u)\|_{\mathrm{F}}^2 + \int_{0<|z|\le\mathfrak{r}} |\gamma_{\mathbf{h}}(s,0,z,u)|^2\nu_s(\mathrm{d}z)\right]\mathrm{d}u$$

takes finite values for all $s\in[0,T]$ and $G_0\in\mathbf{L}^1([0,T],\boldsymbol{\lambda}_{[0,T]})$. Moreover, $b_{\mathbf{h}}$, $a_{\mathbf{h}}$, $\gamma_{\mathbf{h}}$ are Lipschitz continuous with respect to the space variable x in the following sense: There exists a constant $K_{\mathrm{Lip}}\ge 0$ independent of s and u (but may depend on $\mathfrak{r}$) such that the following condition holds for any $s\in[0,T]$, $u\in[0,1]^d$, and $x_1,x_2\in\mathbb{R}^m$:

$$|b_{\mathbf{h}}(s,x_1,u)-b_{\mathbf{h}}(s,x_2,u)|^2 + \|a_{\mathbf{h}}(s,x_1,u)-a_{\mathbf{h}}(s,x_2,u)\|_{\mathrm{F}}^2 + \int_{0<|z|\le\mathfrak{r}} |\gamma_{\mathbf{h}}(s,x_1,z,u)-\gamma_{\mathbf{h}}(s,x_2,z,u)|^2\nu_s(\mathrm{d}z) \le K_{\mathrm{Lip}}^2|x_1-x_2|^2. \tag{4.2}$$

Proposition 4.2. *Under Assumption 2, the grid-sampling SDE (4.1) for policy* $\mathbf{h}$ *with initial condition* $x_0\in\mathbb{R}^m$ *has a unique (up to an indistinguishability) strong solution* $X^{\Pi,\mathbf{h}}$*, for any choice of the partition* Π*. Moreover, the strong solution* $X^{\Pi,\mathbf{h}}$ *also solves the SDE*

$$\begin{aligned} X_t^{\Pi,\mathbf{h}} = x_0 &+ \int_{(0,t]\times[0,1]^d} b_{\mathbf{h}}(s,X_{s-}^{\Pi,\mathbf{h}},u)M_D^{\Pi}(\mathrm{d}s,\mathrm{d}u) + \sum_{l=1}^p \int_{(0,t]\times[0,1]^d} a_{\mathbf{h}}^{(\cdot,l)}(s,X_{s-}^{\Pi,\mathbf{h}},u)M_{B^{(l)}}^{\Pi}(\mathrm{d}s,\mathrm{d}u)\\ &+ \int_{(0,t]\times\{0<|z|\le\mathfrak{r}\}\times[0,1]^d} \gamma_{\mathbf{h}}(s,X_{s-}^{\Pi,\mathbf{h}},z,u)\tilde{M}_J^{\Pi}(\mathrm{d}s,\mathrm{d}z,\mathrm{d}u)\\ &+ \int_{(0,t]\times\{|z|>\mathfrak{r}\}\times[0,1]^d} \gamma_{\mathbf{h}}(s,X_{s-}^{\Pi,\mathbf{h}},z,u)M_J^{\Pi}(\mathrm{d}s,\mathrm{d}z,\mathrm{d}u), \quad t\in[0,T], \end{aligned} \tag{4.3}$$

driven by the following random measures:

$$\begin{aligned} M_D^{\Pi}(\omega,\mathrm{d}t,\mathrm{d}u) &:= \sum_{i=1}^n \mathbb{1}_{(t_{i-1},t_i]}(t)\delta_{\xi_{t_i}^{\Pi}(\omega)}(\mathrm{d}u)\mathrm{d}t,\\ M_{B^{(l)}}^{\Pi}(\omega,t,A) &:= \left(\int_0^t \sum_{i=1}^n \mathbb{1}_{(t_{i-1},t_i]}(s)\mathbb{1}_A(\xi_{t_i}^{\Pi})\,\mathrm{d}B_s^{(l)}\right)(\omega), \quad A\in\mathcal{B}([0,1]^d),\ t\in[0,T],\ l=1,\dots,p,\\ M_J^{\Pi}(\omega,\mathrm{d}t,\mathrm{d}z,\mathrm{d}u) &:= \sum_{i=1}^n \sum_{t\in(t_{i-1},t_i]} \mathbb{1}_{\{\Delta L_t(\omega)\neq 0\}}\delta_{(t,\Delta L_t(\omega),\xi_{t_i}^{\Pi}(\omega))}(\mathrm{d}t,\mathrm{d}z,\mathrm{d}u), \end{aligned}$$

where $L_t := \int_0^t\int_{0<|z|\le\mathfrak{r}} z\tilde{N}(\mathrm{d}s,\mathrm{d}z) + \int_0^t\int_{|z|>\mathfrak{r}} zN(\mathrm{d}s,\mathrm{d}z)$ *and* δ_y *is the Dirac distribution on the point* y*; here,* $M_{B^{(l)}}^{\Pi}$ *are orthogonal martingale measures with intensity measure* M_D^{Π}*, and* M_J^{Π} *is an integer-valued random measure with* $\mathbb{F}^{\Pi}$*-predictable compensator measure*

$$\mu_J^{\Pi}(\omega,\mathrm{d}t,\mathrm{d}z,\mathrm{d}u) := \sum_{i=1}^n \mathbb{1}_{(t_{i-1},t_i]}(t)\delta_{\xi_{t_i}^{\Pi}(\omega)}(\mathrm{d}u)\nu_t(\mathrm{d}z)\mathrm{d}t$$

and corresponding compensated measure $\tilde{M}_J^{\Pi} := M_J^{\Pi} - \mu_J^{\Pi}$.

For integration with respect to compensated integer-valued random measures and to orthogonal martingale measures, we refer to [15] and [23], respectively. The proof of Proposition 4.2 is presented in Section 8.1.

Remark 4.3. We note that the solution $X^{\Pi,\mathbf{h}}$ to the grid-sampling SDE fails to be Markovian with respect to $\mathbb{F}^{\Pi}$, in general. For a concrete counterexample, take $m=d=1$, $x_0=0$, $a_{\mathbf{h}}\equiv\gamma_{\mathbf{h}}\equiv 0$, and $b_{\mathbf{h}}(s,x,u)=u$. Then, for $t\in[0,t_1]$, $X_t^{\Pi,\mathbf{h}} = t\xi_1$. Since $\xi_1=\xi_t^{\Pi}$ for every $t\in(0,t_1]$ and the filtration $\mathbb{F}^{\Pi}$ is right-continuous, we conclude that ξ_1 is $\mathcal{F}_0^{\Pi}$-measurable. Hence, $\mathbb{E}[X_{t_1}^{\Pi,\mathbf{h}}|\mathcal{F}_0^{\Pi}] = \xi_1$, while $\mathbb{E}[X_{t_1}^{\Pi,\mathbf{h}}|X_0^{\Pi,\mathbf{h}}] = \mathbb{E}[\xi_1] = 1/2$, because $X_0^{\Pi,\mathbf{h}}$ is constant.

One can show, however, that the pair of processes $(X_t^{\Pi,\mathbf{h}},\xi_{t+}^{\Pi})_{t\in[0,T]}$ is Markovian with respect to $\mathbb{F}^{\Pi}$ under the assumptions of Proposition 4.2, as sketched in Appendix A.3. Since $\xi_0^{\Pi}=0$ and $\xi_{0+}^{\Pi}=\xi_1$, the same counterexample as above shows that ξ_{t+}^{Π} cannot be replaced by ξ_t^{Π}.

Remark 4.4. Suppose $(\Pi_n)_{n\in\mathbb{N}}$ is a sequence of sampling partitions with mesh-size converging to zero. Fix any $k\in\mathbb{N}$ time points $s_1,\dots,s_k$. Then, for sufficiently large $n\ge N_0$, these k points will be in different sub-intervals of the partition Π_n and, therefore, the random vector $(\xi_{s_1}^{\Pi_n},\dots,\xi_{s_k}^{\Pi_n})$ consists of independent uniform random variables. Consequently, $(\xi_t^{\Pi_n})_{t\in[0,T]}$ converges in finite-dimensional distributions to an independent, identically distributed family $(\xi_t^*)_{t\in[0,T]}$ of uniform random variables, i.e., to idealized sampling. Due to the issues of idealized sampling discussed in Section 3, it is not promising to pass to the limit on the level of the integrands in the formulation (4.1) of the grid-sampling SDE. Therefore, we switch to the formulation (4.3) of the grid-sampling SDE, where the influence of the randomization process $(\xi_t^{\Pi_n})_{t\in[0,T]}$ has been moved to the integrator, before passing to the limit.

5. The grid-sampling limit SDE

5.1. Limit theorem and grid-sampling limit SDE

In this subsection, we establish a limit theorem for the grid-sampling random measures $(M_D^\Pi, M_{B^{(1)}}^\Pi, \dots, M_{B^{(p)}}^\Pi, M_J^\Pi)$ defined in Proposition 4.2, which drive the grid-sampling SDE (4.3), as the mesh-size of Π goes to zero. This limit theorem suggests a formulation for the *grid-sampling limit SDE* in which the grid-sampling random measures in (4.3) are replaced by the limit random measures $(M_D, M_{B^{(1)}}, \dots, M_{B^{(p)}}, M_J)$.

We define

$$M_D(A) := \boldsymbol{\lambda}_{[0,T]} \otimes \boldsymbol{\lambda}_{[0,1]}^{\otimes d}(A), \quad A \in \mathcal{B}([0,T]) \otimes \mathcal{B}([0,1]^d),$$

where $\boldsymbol{\lambda}_U$ is the restriction of the 1-dimensional Lebesgue measure to a Borel set $U \in \mathcal{B}(\mathbb{R})$. Moreover, we let $(M_{B^{(1)}}, \dots, M_{B^{(p)}})$ denote p independent martingale measures with continuous paths and intensity measure M_D. Continuous martingale measures with deterministic intensities are also called white noise martingale measures, and we refer to [23] for a construction of such martingale measures and more background information. Also, Lemma 6.1 below provides some information on their relation to Brownian motion.

Finally, M_J denotes a Poisson random measure on $[0,T] \times \mathbb{R}_0^q \times [0,1]^d$ with intensity measure

$$\mu_J(\mathrm{d}t, \mathrm{d}z, \mathrm{d}u) := \nu_t(\mathrm{d}z)\mathrm{d}u\mathrm{d}t.$$

An explicit construction of M_J can be found in [2]. As usual $\tilde{M}_J := M_J - \mu_J$ stands for the compensated Poisson random measure.

We assume that the original filtered probability space $(\Omega, \mathcal{F}, \bar{\mathbb{F}}, \mathbb{P})$ has been chosen sufficiently large to carry $(M_{B^{(1)}}, \dots, M_{B^{(p)}})$ and M_J. Then, $(M_{B^{(1)}}, \dots, M_{B^{(p)}})$ and M_J are automatically independent (see the argument in [2]). Denote by $\mathbb{F}$ the right-continuous, augmented version of the filtration generated by $(M_{B^{(1)}}, \dots, M_{B^{(p)}}, M_J)$.

Theorem 3. *Let $(\Pi_n)_{n\in\mathbb{N}}$ be a sequence of finite partitions of $[0,T]$ with $\lim_{n\to\infty} |\Pi_n| = 0$. For any $m \in \mathbb{N}$, $R \in (0,\infty) \cup \{\mathfrak{r}\}$, and for any bounded measurable functions $f_l^{(k)} : [0,T] \times [0,1]^d \to \mathbb{R}$ $(l = 0, \dots, p;\ k = 1, \dots, m)$, $f_l^{(k)} : [0,T] \times \mathbb{R}_0^q \times [0,1]^d \to \mathbb{R}$ $(l = p+1, p+2;\ k = 1, \dots, m)$, consider the sequence of $\mathbb{R}^m$-valued processes $\mathcal{X}^n = (\mathcal{X}^{n,(1)}, \dots, \mathcal{X}^{n,(m)})$ defined via*

$$\begin{aligned}\mathcal{X}_t^{n,(k)} &= \int_{(0,t]\times[0,1]^d} f_0^{(k)}(s,u) M_D^{\Pi_n}(\mathrm{d}s,\mathrm{d}u) + \sum_{l=1}^p \int_{(0,t]\times[0,1]^d} f_l^{(k)}(s,u) M_{B^{(l)}}^{\Pi_n}(\mathrm{d}s,\mathrm{d}u)\\ &\quad + \int_{(0,t]\times\{0<|z|\le R\}\times[0,1]^d} f_{p+1}^{(k)}(s,z,u)|z|\tilde{M}_J^{\Pi_n}(\mathrm{d}s,\mathrm{d}z,\mathrm{d}u)\\ &\quad + \int_{(0,t]\times\{|z|>R\}\times[0,1]^d} f_{p+2}^{(k)}(s,z,u) M_J^{\Pi_n}(\mathrm{d}s,\mathrm{d}z,\mathrm{d}u), \quad t\in[0,T],\ k = 1,\dots,m.\end{aligned}$$

Then, $(\mathcal{X}^n)_{n\in\mathbb{N}}$ converges weakly in the Skorokhod topology on the space $\mathbb{D}_T(\mathbb{R}^m)$ of $\mathbb{R}^m$-valued, càdlàg functions to $\mathcal{X} = (\mathcal{X}^{(1)}, \dots, \mathcal{X}^{(m)})$, where

$$\begin{aligned}\mathcal{X}_t^{(k)} &= \int_{(0,t]\times[0,1]^d} f_0^{(k)}(s,u) M_D(\mathrm{d}s,\mathrm{d}u) + \sum_{l=1}^p \int_{(0,t]\times[0,1]^d} f_l^{(k)}(s,u) M_{B^{(l)}}(\mathrm{d}s,\mathrm{d}u)\\ &\quad + \int_{(0,t]\times\{0<|z|\le R\}\times[0,1]^d} f_{p+1}^{(k)}(s,z,u)|z|\tilde{M}_J(\mathrm{d}s,\mathrm{d}z,\mathrm{d}u)\\ &\quad + \int_{(0,t]\times\{|z|>R\}\times[0,1]^d} f_{p+2}^{(k)}(s,z,u) M_J(\mathrm{d}s,\mathrm{d}z,\mathrm{d}u), \quad t\in[0,T],\ k = 1,\dots,m.\end{aligned}$$

The proof of Theorem 3 is provided in Section 7.

Remark 5.1. As a consequence of Theorem 3, $(M_D^{\Pi_n}, M_{B^{(1)}}^{\Pi_n}, \dots, M_{B^{(p)}}^{\Pi_n}, M_J^{\Pi_n})$ vaguely converges to $(M_D, M_{B^{(1)}}, \dots, M_{B^{(p)}}, M_J)$ in the following sense: For any $m \in \mathbb{N}$, and for any continuous functions with compact support $f_l^{(k)} : [0,T] \times [0,1]^d \to \mathbb{R}$ $(l = 0, \dots, p;\ k = 1, \dots, m)$, $f_{p+2}^{(k)} : [0,T] \times \mathbb{R}_0^q \times [0,1]^d \to \mathbb{R}$ $(k = 1, \dots, m)$, the sequence of $\mathbb{R}^m$-valued processes $\mathcal{X}^n = (\mathcal{X}^{n,(1)}, \dots, \mathcal{X}^{n,(m)})$ defined via

$$\begin{aligned}\mathcal{X}_t^{n,(k)} &= \int_{(0,t]\times[0,1]^d} f_0^{(k)}(s,u) M_D^{\Pi_n}(\mathrm{d}s,\mathrm{d}u) + \sum_{l=1}^p \int_{(0,t]\times[0,1]^d} f_l^{(k)}(s,u) M_{B^{(l)}}^{\Pi_n}(\mathrm{d}s,\mathrm{d}u)\\ &\quad + \int_{(0,t]\times\mathbb{R}_0^q\times[0,1]^d} f_{p+2}^{(k)}(s,z,u) M_J^{\Pi_n}(\mathrm{d}s,\mathrm{d}z,\mathrm{d}u), \quad t\in[0,T],\ k = 1,\dots,m,\end{aligned}$$

weakly converges in the Skorokhod topology on $\mathbb{D}_T(\mathbb{R}^m)$ to $\mathcal{X} = (\mathcal{X}^{(1)}, \dots, \mathcal{X}^{(m)})$, where

$$\begin{aligned}\mathcal{X}_t^{(k)} &= \int_{(0,t]\times[0,1]^d} f_0^{(k)}(s,u) M_D(\mathrm{d}s,\mathrm{d}u) + \sum_{l=1}^p \int_{(0,t]\times[0,1]^d} f_l^{(k)}(s,u) M_{B^{(l)}}(\mathrm{d}s,\mathrm{d}u)\\ &\quad + \int_{(0,t]\times\mathbb{R}_0^q\times[0,1]^d} f_{p+2}^{(k)}(s,z,u) M_J(\mathrm{d}s,\mathrm{d}z,\mathrm{d}u), \quad t\in[0,T],\ k = 1,\dots,m.\end{aligned}$$

Indeed, if the $f_l^{(k)}$'s $(l = p+1, p+2;\ k = 1, \dots, m)$ in Theorem 3 have compact support, then there is an $\varepsilon > 0$ (independent of k, l, t, u) such that $f_l^{(k)} = 0$, if $0 < |z| \le \varepsilon$. Hence, we can apply Theorem 3 with $R = \varepsilon$. We also refer to [22] for background information on the general theory of vague convergence of random measures and to [38,39] for the case of martingale measures.

In view of Theorem 3, the random measure formulation (4.3) of the grid-sampling SDE (4.1), and the definition of M_D, a natural limit formulation of the grid-sampling SDE for a given randomized policy $\mathbf{h}: [0,T]\times\mathbb{R}^m\times[0,1]^d \to \mathbb{R}^d$ is

$$X_t^{\mathbf{h}} = x_0 + \int_0^t \int_{[0,1]^d} b(s, X_{s-}^{\mathbf{h}}, \mathbf{h}(s, X_{s-}^{\mathbf{h}}, u))\mathrm{d}u\mathrm{d}s + \sum_{l=1}^{p} \int_{(0,t]\times[0,1]^d} a^{(\cdot,l)}(s, X_{s-}^{\mathbf{h}}, \mathbf{h}(s, X_{s-}^{\mathbf{h}}, u)) M_{B^{(l)}}(\mathrm{d}s, \mathrm{d}u)$$
$$+ \int_{(0,t]\times\{0<|z|\le\mathfrak{r}\}\times[0,1]^d} \gamma(s, X_{s-}^{\mathbf{h}}, z, \mathbf{h}(s, X_{s-}^{\mathbf{h}}, u))\tilde{M}_J(\mathrm{d}s, \mathrm{d}z, \mathrm{d}u) + \int_{(0,t]\times\{|z|>\mathfrak{r}\}\times[0,1]^d} \gamma(s, X_{s-}^{\mathbf{h}}, z, \mathbf{h}(s, X_{s-}^{\mathbf{h}}, u)) M_J(\mathrm{d}s, \mathrm{d}z, \mathrm{d}u). \tag{5.1}$$

We call this SDE the *grid-sampling limit SDE* for policy $\mathbf{h}$.

Remark 5.2. We emphasize that the random measures $(M_D, M_{B^{(1)}}, \dots, M_{B^{(p)}}, M_J)$ appearing in the limit are independent, whereas the pre-limit random measures $(M_D^{\Pi}, M_{B^{(1)}}^{\Pi}, \dots, M_{B^{(p)}}^{\Pi}, M_J^{\Pi})$ are jointly constructed in terms of the randomization process ξ^{Π} and are, thus, dependent. Hence, it is intuitively clear that a solution $X^{\mathbf{h}}$ of the grid-sampling limit SDE (5.1) cannot be interpreted as the model dynamics evaluated along a $(\xi_t)_{t\in[0,T]}$-randomized policy, i.e., it cannot be reformulated in the form (2.3) for some randomization process ξ in general. More formally, Example 3.2 serves as a simple counterexample. Nonetheless, we think that the limit SDE (5.1) is practically relevant as the limit dynamics of the observable state variables of action processes under ξ^{Π}-randomization and can be applied for justifying learning algorithms derived by the first-optimize-then-discretize approach. This aspect will be briefly sketched in Section 6.3 below.

5.2. Well-posedness of grid-sampling limit SDE

We illustrate in Proposition 5.3 below the existence and uniqueness of strong solutions to the SDE (5.1). Its proof is given in Section 8.2.

Proposition 5.3. *Let $\mathfrak{r}\in[0,\infty]$. Under Assumption 2, the grid-sampling limit SDE (5.1) for policy $\mathbf{h}$ with initial data $x_0\in\mathbb{R}^m$ has a unique (up to an indistinguishability) strong solution $X^{\mathbf{h}}$. Moreover, the law of $X^{\mathbf{h}}$ solves the martingale problem for the operator $\mathcal{L}_{\mathbf{h}}$ defined for $f\in C_c^2(\mathbb{R}^m)$ by*

$$(\mathcal{L}_{\mathbf{h}} f)(s,x) := \int_{[0,1]^d} \Bigg[\sum_{i=1}^m b_{\mathbf{h}}^{(i)}(s,x,u)\frac{\partial f}{\partial x_i}(x) + \frac{1}{2}\sum_{i,j=1}^m A_{\mathbf{h}}^{(i,j)}(s,x,u)\frac{\partial^2 f}{\partial x_i \partial x_j}(x)$$
$$+ \int_{\mathbb{R}_0^q} \left(f(x+\gamma_{\mathbf{h}}(s,x,z,u)) - f(x) - \sum_{i=1}^m \mathbb{1}_{\{0<|z|\le\mathfrak{r}\}}\gamma_{\mathbf{h}}^{(i)}(s,x,z,u)\frac{\partial f}{\partial x_i}(x)\right)\nu_s(\mathrm{d}z)\Bigg]\mathrm{d}u,$$

with initial distribution δ_{x_0}, where $A_{\mathbf{h}} := a_{\mathbf{h}} a_{\mathbf{h}}^{\top}$.

Remark 5.4. Under the assumptions of Proposition 5.3, the solution $(X_t^{\mathbf{h}})_{\in[0,T]}$ of the grid-sampling limit SDE is Markovian with respect ot $\mathbb{F}$, see Appendix A.3 below. This important property in the context of reinforcement learning and stochastic control explains, why we focus on randomized policies in *feedback* form.

Remark 5.5. Suppose that we are in the no-jump case, i.e, $\gamma\equiv 0$.

(1) Well-posedness of SDEs driven by white noise martingale measures is studied, e.g., in [23, Proposition IV-1], and in that case Proposition 5.3 can be seen as a variant of their result.
(2) By combining Theorem 3 with the stability results for SDEs driven by continuous orthogonal martingale measures in [25, p.354], we observe that under at most technical assumptions the following limit theorem is valid: If $\mathbf{h}_1,\dots,\mathbf{h}_K: [0,T]\times\mathbb{R}^m\times[0,1]^d\to\mathbb{R}^d$ are randomized policies, then one obtains the joint weak convergence
$$(X^{\Pi_n,\mathbf{h}_1},\dots,X^{\Pi_n,\mathbf{h}_K}, M_D^{\Pi_n}, M_{B^{(1)}}^{\Pi_n},\dots,M_{B^{(p)}}^{\Pi_n}) \to (X^{\mathbf{h}_1},\dots,X^{\mathbf{h}_K}, M_D, M_{B^{(1)}},\dots,M_{B^{(p)}}).$$
This result serves as another justification for using the grid-sampling limit SDE (5.1). We leave a detailed study of this aspect in the general case with jumps to future research.

Remark 5.6.

(1) The proof of Proposition 5.3 reveals that the conclusion in Proposition 5.3 still holds true when the Lipschitz condition (4.2) in Assumption 2 is weakened to: There is a constant $K\ge 0$ independent of s (but may depend on $\mathfrak{r}$) such that for any $s\in[0,T]$, $x_1,x_2\in\mathbb{R}^m$,
$$\left(\int_{[0,1]^d} |b_{\mathbf{h}}(s,x_1,u) - b_{\mathbf{h}}(s,x_2,u)|\mathrm{d}u\right)^2 + \int_{[0,1]^d} \|a_{\mathbf{h}}(s,x_1,u) - a_{\mathbf{h}}(s,x_2,u)\|_{\mathrm{F}}^2\,\mathrm{d}u$$
$$+ \int_{\{0<|z|\le\mathfrak{r}\}\times[0,1]^d} |\gamma_{\mathbf{h}}(s,x_1,z,u) - \gamma_{\mathbf{h}}(s,x_2,z,u)|^2\,\nu_s(\mathrm{d}z)\mathrm{d}u \le K^2|x_1-x_2|^2.$$
(2) If $\mathfrak{r}=\infty$, then there exists a constant $\tilde{K}\ge 0$ not depending on x_0 such that the strong solution $X^{\mathbf{h}}$ to (5.1) satisfies (see Remark A.4)
$$\mathbb{E}\left[\sup_{0\le t\le T}|X_t^{\mathbf{h}}|^2\right] \le \tilde{K}^2(1+|x_0|^2).$$

6. Examples and discussion

6.1. Examples

We first discuss two examples in which the grid-sampling limit SDE (5.1) is simplified. They rely on the following elementary lemma, whose proof is given in [2].

Lemma 6.1. *Suppose that $\eta : \Omega \times [0,T] \times [0,1]^d \to \mathbb{R}^m$ is an $\mathbb{F}$-predictable random field satisfying*

$$\int_{[0,1]^d} (\eta_t \eta_t^\mathsf{T})(u) \mathrm{d}u = I_m \quad \mathbb{P} \otimes \boldsymbol{\lambda}_{[0,T]}\text{-a.e. } (\omega, t) \in \Omega \times [0,T].$$

Define

$$B_t^{\eta,(k,l)} = \int_0^t \int_{[0,1]^d} \eta_s^{(k)}(u)\, M_{B^{(l)}}(\mathrm{d}s, \mathrm{d}u), \quad t \in [0,T],\ l = 1, \dots, p,\ k = 1, \dots, m.$$

Then, $B^\eta = (B^{\eta,(k,l)} : l = 1, \dots, p,\ k = 1, \dots, m)$ is an mp-dimensional Brownian motion.

Example 6.2. Suppose that $\mathbf{h}$ is a classical, non-randomized control in feedback form, i.e., $\mathbf{h}$ does not depend on u. By Lemma 6.1 (with η being the $\mathbb{R}$-valued function which is constant 1),

$$B^{\mathbf{1}} = \left(\int_0^\cdot \int_{[0,1]^d} M_{B^{(1)}}(\mathrm{d}s, \mathrm{d}u), \dots, \int_0^\cdot \int_{[0,1]^d} M_{B^{(p)}}(\mathrm{d}s, \mathrm{d}u) \right)^\mathsf{T}$$

is a p-dimensional Brownian motion. Moreover,

$$N^{\mathbf{1}}(\mathrm{d}t, \mathrm{d}z) = \int_{[0,1]^d} M_J(\mathrm{d}t, \mathrm{d}z, \mathrm{d}u)$$

is a Poisson random measure independent of B^η with intensity $\nu_t(\mathrm{d}z)\mathrm{d}t$. Then, SDE (5.1) can be re-written as

$$\begin{aligned} X_t^{\mathbf{h}} = x_0 &+ \int_0^t b(s, X_{s-}^{\mathbf{h}}, \mathbf{h}(s, X_{s-}^{\mathbf{h}}))\mathrm{d}s + \int_0^t a(s, X_{s-}^{\mathbf{h}}, \mathbf{h}(s, X_{s-}^{\mathbf{h}}))\mathrm{d}B_s^{\mathbf{1}} \\ &+ \int_{(0,t]\times\{0<|z|\le \mathfrak{r}\}} \gamma(s, X_{s-}^{\mathbf{h}}, z, \mathbf{h}(s, X_{s-}^{\mathbf{h}}))\tilde{N}^{\mathbf{1}}(\mathrm{d}s, \mathrm{d}z) + \int_{(0,t]\times\{|z|>\mathfrak{r}\}} \gamma(s, X_{s-}^{\mathbf{h}}, z, \mathbf{h}(s, X_{s-}^{\mathbf{h}}))N^{\mathbf{1}}(\mathrm{d}s, \mathrm{d}z), \end{aligned}$$

i.e., we recover the dynamics (2.1), as it should be.

Example 6.3. We now assume the drift coefficient b and the diffusion coefficient a are affine-linear in the control, i.e.,

$$a(t,x,y) = a_0(t,x) + \sum_{j=1}^d y^{(j)} a_j(t,x), \quad b(t,x,y) = b_0(t,x) + \sum_{j=1}^d y^{(j)} b_j(t,x)$$

for measurable functions $a_j : [0,T] \times \mathbb{R}^m \to \mathbb{R}^{m\times p}$ and $b_j : [0,T] \times \mathbb{R}^m \to \mathbb{R}^m$. The randomized control is given in terms of the measurable function $\mathbf{h} : [0,T] \times \mathbb{R}^m \times [0,1]^d \to \mathbb{R}^d$. We assume that the coefficients are sufficiently regular to guarantee that a solution $X^{\mathbf{h}}$ to (5.1) exists. Supposing that $\mathbf{h}$ is square integrable with respect to the uniform distribution in the u-variable, we then consider the mean vector and covariance matrix

$$\mu_{\mathbf{h}}(t,x) = \int_{[0,1]^d} \mathbf{h}(t,x,u)\mathrm{d}u, \quad \Theta_{\mathbf{h}}(t,x) = \int_{[0,1]^d} (\mathbf{h}(t,x,u) - \mu_{\mathbf{h}}(t,x))(\mathbf{h}(t,x,u) - \mu_{\mathbf{h}}(t,x))^\mathsf{T}\mathrm{d}u$$

as a function of (t,x). Assuming that $\Theta_{\mathbf{h}}(t,x)$ is positive definite for every $(t,x) \in [0,T] \times \mathbb{R}^m$, we write $\vartheta_{\mathbf{h}}(t,x)$ for the positive definite matrix root of $\Theta_{\mathbf{h}}(t,x)$ and define

$$\eta_{\mathbf{h}} : [0,T] \times \mathbb{R}^m \times [0,1]^d \to \mathbb{R}^d, (t,x,u) \mapsto \vartheta_{\mathbf{h}}(t,x)^{-1}(\mathbf{h}(t,x,u) - \mu_{\mathbf{h}}(t,x)).$$

Note that for every $(t,x) \in [0,T] \times \mathbb{R}^m$,

$$\int_{[0,1]^d} \eta_{\mathbf{h}}(t,x,u)\mathrm{d}u = 0, \quad \int_{[0,1]^d} (\eta_{\mathbf{h}} \eta_{\mathbf{h}}^\mathsf{T})(t,x,u)\mathrm{d}u = I_d.$$

Thus, the $\mathbb{R}^{d+1}$-valued random field

$$\eta_t(u) = (\eta_{\mathbf{h}}^{(1)}(t, X_{t-}^{\mathbf{h}}, u), \dots, \eta_{\mathbf{h}}^{(d)}(t, X_{t-}^{\mathbf{h}}, u), 1)^\mathsf{T}$$

satisfies the assumptions of Lemma 6.1 and we denote the corresponding Brownian motion by $B^\eta = (B^{\eta,(i,l)})_{i=1,\dots,d+1,\ l=1,\dots,p}$. Then, the white noise measures can be replaced by the $(d+1)p$-dimensional Brownian motion B^η and (5.1) becomes

$$\begin{aligned} X_t^{\mathbf{h}} = x_0 &+ \int_0^t \left(b_0(s, X_{s-}^{\mathbf{h}}) + \sum_{j=1}^d b_j(s, X_{s-}^{\mathbf{h}}) \mu_{\mathbf{h}}^{(j)}(s, X_{s-}^{\mathbf{h}}) \right) \mathrm{d}s \\ &+ \sum_{l=1}^p \int_0^t \left(a_0^{(\cdot,l)}(s, X_{s-}^{\mathbf{h}}) + \sum_{j=1}^d a_j^{(\cdot,l)}(s, X_{s-}^{\mathbf{h}}) \mu_{\mathbf{h}}^{(j)}(s, X_{s-}^{\mathbf{h}}) \right) \mathrm{d}B_s^{\eta,(d+1,l)} + \sum_{l=1}^p \sum_{i=1}^d \int_0^t \left(\sum_{j=1}^d a_j^{(\cdot,l)}(s, X_{s-}^{\mathbf{h}}) \vartheta_{\mathbf{h}}^{(j,i)}(s, X_{s-}^{\mathbf{h}}) \right) \mathrm{d}B_s^{\eta,(i,l)} \end{aligned}$$

$$+\int_{(0,t]\times\{0<|z|\le\mathfrak{r}\}\times[0,1]^d}\gamma(s,X^{\mathbf{h}}_{s-},z,\mathbf{h}(s,X^{\mathbf{h}}_{s-},u))\tilde{M}_J(\mathrm{d}s,\mathrm{d}z,\mathrm{d}u)+\int_{(0,t]\times\{|z|>\mathfrak{r}\}\times[0,1]^d}\gamma(s,X^{\mathbf{h}}_{s-},z,\mathbf{h}(s,X^{\mathbf{h}}_{s-},u))M_J(\mathrm{d}s,\mathrm{d}z,\mathrm{d}u).$$

This example extends the analogous SDE formulation for entropy-regularized mean-variance portfolio optimization with jumps derived in [1]. Note, however, that the white noise measure approach clarifies that (and how exactly) the driving Brownian motion depends on the choice of the randomized control $\mathbf{h}$.

6.2. Comparison to the exploratory SDE of [34]

In this subsection, we briefly compare the grid-sampling limit SDE (5.1) to the exploratory SDE introduced in [34]. In order to keep the notation simple, we confine ourselves to the one-dimensional case ($m = p = d = 1$) without jumps $\gamma = 0$, compare [34]. We note, however, that the multivariate case of the exploratory SDE is covered in [18] and, recently, a setting with jumps has been developed in [10]. In any of these cases, the derivation of the exploratory SDE relies on a heuristic law of large numbers argument to extract the semimartingale characteristics when averaging over independent executions of a relaxed control.

Given a relaxed control $h\colon [0,T]\times\mathbb{R}\to\mathcal{P}r(\mathcal{B}(\mathbb{R}))$ with Lebesgue density $\dot{h}(t,x,\cdot)$, the exploratory SDE takes the form

$$\tilde{X}^h_t = x_0+\int_0^t\int_{\mathbb{R}} b(s,\tilde{X}^h_s,y)\dot{h}(s,\tilde{X}^h_s,y)\mathrm{d}y\mathrm{d}s+\int_0^t\sqrt{\int_{\mathbb{R}} a(s,\tilde{X}^h_s,y)^2\dot{h}(s,\tilde{X}^h_s,y)\mathrm{d}y}\,\mathrm{d}W_s$$

for some 1-dimensional Brownian motion W. Lemma 2 in [18] states sufficient conditions on b, a, and $\dot{h}$ for existence and uniqueness of a strong solution. Note that the law of $\tilde{X}^h$ then solves the martingale problem for the operator

$$(\mathcal{L}_h f)(t,x) := \int_{\mathbb{R}}\left(\frac{1}{2}a(t,x,y)^2 f''(x)+b(t,x,y)f'(x)\right)\dot{h}(t,x,y)\mathrm{d}y. \tag{6.1}$$

We suppose for the rest of this subsection that $\mathbf{h}$ is a randomized control, which executes h, and that the assumptions of Proposition 5.3 are satisfied. Note that, by a change of variables, the operators $\mathcal{L}_h$ and $\mathcal{L}_{\mathbf{h}}$ coincide. Hence, by Proposition 5.3, the law of the unique solution $X^{\mathbf{h}}$ to the grid-sampling limit SDE solves the martingale problem for the same operator $\mathcal{L}_h(=\mathcal{L}_{\mathbf{h}})$. Moreover, uniqueness of the martingale problem for the operator $\mathcal{L}_{\mathbf{h}}$ is established under the Lipschitz assumptions in Proposition IV.1 in [23]. Hence, by Corollaries 5.4.8–5.4.9 in [21], the exploratory SDE has a weak solution $\tilde{X}^h$, and, moreover, $\tilde{X}^h$ and $X^{\mathbf{h}}$ have the same probability law. Hence, in a stochastic control framework (e.g., to compute the expected cost of a given relaxed/randomized control pair h, $\mathbf{h}$ or for the derivation of an HJB equation), the grid-sampling limit SDE $X^{\mathbf{h}}$ and the exploratory SDE $\tilde{X}^h$ will lead to the same result – and it is a matter of taste which one to use. In the former SDE the white noise martingale measure comes up, while, in the latter SDE, one has to deal with the square-root in the diffusion coefficient, compare the Remarks in [25, pp. 350–351].

However, if one considers several controls at the same time, then their joint distributions, for example the distributions of $(\tilde{X}^{h_1},\tilde{X}^{h_2})$ and $(X^{\mathbf{h}_1},X^{\mathbf{h}_2})$, may differ as illustrated by the following example.

Example 6.4. (1) Suppose $T=1$, $b=0$ and $a(t,x,u)=u$. We apply the randomized controls $\mathbf{h}_j(t,x,u)=\mu_j+\sigma_j\Phi^{-1}(u)$, ($\mu_j\in\mathbb{R}$, $\sigma_j>0$, $j=1,2$), which execute a Gaussian law $h_j(t,x)$ with mean μ_j and variance σ_j^2 independent of the time and state of the system. For a fixed sampling partition Π, the predictable covariation of the model dynamics along the ξ^Π-randomized controls satisfies

$$\langle X^{\Pi,\mathbf{h}_1},X^{\Pi,\mathbf{h}_2}\rangle_1=\sum_{i=1}^n(t_i-t_{i-1})(\mu_1+\sigma_1\Phi^{-1}(\xi^\Pi_{t_i}))(\mu_2+\sigma_2\Phi^{-1}(\xi^\Pi_{t_i})).$$

If, e.g., Π_n is the equidistant partition of the unit interval into n subintervals and η denotes any uniform random variable on $[0,1]$, then a straightforward application of the strong law of large numbers implies, a.s.,

$$\langle X^{\Pi_n,\mathbf{h}_1},X^{\Pi_n,\mathbf{h}_2}\rangle_1\to\mathbb{E}\Big[(\mu_1+\sigma_1\Phi^{-1}(\eta))(\mu_2+\sigma_2\Phi^{-1}(\eta))\Big]=\mu_1\mu_2+\sigma_1\sigma_2.$$

This limit coincides with the predictable covariation of the grid-sampling limit SDEs. Indeed, we first note that by Example 6.3, due to the linear dependence on the control,

$$X^{\mathbf{h}_j}_t = X^{\mathbf{h}_j}_0+\mu_j B^{\mathbf{1}}_t+\sigma_j B^{\Phi^{-1}}_t,\ j=1,2,$$

for the independent Brownian motions $B^{\mathbf{1}}_t=M_B([0,t]\times[0,1])$, $B^{\Phi^{-1}}_t=\int_0^t\int_0^1\Phi^{-1}(u)M_B(\mathrm{d}s,\mathrm{d}u)$; and therefore

$$\langle X^{\mathbf{h}_1},X^{\mathbf{h}_2}\rangle_1=\mu_1\mu_2+\sigma_1\sigma_2.$$

However, the predictable covariation of the corresponding exploratory SDEs differs from this expression and can be computed as

$$\begin{aligned}&\langle\tilde{X}^{h_1},\tilde{X}^{h_2}\rangle_1\\&=\left\langle\int_0^\cdot\sqrt{\int_{\mathbb{R}}y^2\frac{1}{\sqrt{2\pi\sigma_1^2}}\mathrm{e}^{-(y-\mu_1)^2/(2\sigma_1^2)}\mathrm{d}y}\,\mathrm{d}W_s,\int_0^\cdot\sqrt{\int_{\mathbb{R}}y^2\frac{1}{\sqrt{2\pi\sigma_2^2}}\mathrm{e}^{-(y-\mu_2)^2/(2\sigma_2^2)}\mathrm{d}y}\,\mathrm{d}W_s\right\rangle_1\\&=\sqrt{(\mu_1^2+\sigma_1^2)(\mu_2^2+\sigma_2^2)}.\end{aligned} \tag{6.2}$$

(2) In part (1) of the example, the execution of the two Gaussian control laws h_1 and h_2 is performed at the grid point t_i of the grid Π via the perfectly correlated random variables $(\mu_j + \sigma_j \Phi^{-1}(\xi^{\Pi}_{t_i}))$, $j = 1, 2$. Our setting allows for more general correlation structures. Choose two Borel-measurable functions $\mathbf{u}_1$, $\mathbf{u}_2 : [0,1] \to [0,1]$ such that, for every uniform random variable η on $[0,1]$, the random variables $\mathbf{u}_1(\eta)$ and $\mathbf{u}_2(\eta)$ are independent uniform random variables on $[0,1]$, see the proof of Theorem 20.4 in [3] for a construction of such functions. Then, for every $\rho \in [-1,1]$, we may also model the execution of the Gaussian control laws h_j, $j = 1, 2$, in part (1) via the randomized controls $\bar{\mathbf{h}}_1(t,x,u) = \mu_1 + \sigma_1(\Phi^{-1} \circ \mathbf{u}_1)(u)$ and $\bar{\mathbf{h}}_2(t,x,u) = \mu_2 + \sigma_2(\rho\Phi^{-1} \circ \mathbf{u}_1 + \sqrt{1-\rho^2}\Phi^{-1} \circ \mathbf{u}_2)(u)$. Leaving the setting of part (1) unchanged except for the different choices of the randomized controls, we now compute the limiting predictable covariation of the grid-sampling SDEs based on the equidistant partitions Π_n. By the same law-of-large-number argument as above, we obtain, as n tends to infinity,

$$\langle X^{\Pi_n,\bar{\mathbf{h}}_1}, X^{\Pi_n,\bar{\mathbf{h}}_2}\rangle_1 \to \mathbb{E}\Big[(\mu_1 + \sigma_1(\Phi^{-1} \circ \mathbf{u}_1)(\eta))(\mu_2 + \sigma_2(\rho\Phi^{-1} \circ \mathbf{u}_1 + \sqrt{1-\rho^2}\Phi^{-1} \circ \mathbf{u}_2)(\eta))\Big] = \mu_1\mu_2 + \rho\sigma_1\sigma_2, \quad \text{a.s.}$$

Again, this limit of the predictable covariation is not in line with predictable covariation of the exploratory SDEs for h_1 and h_2 in (6.2), but it coincides with the predictable covariation of the grid-sampling limit SDEs $X^{\bar{\mathbf{h}}_1}$ and $X^{\bar{\mathbf{h}}_2}$ at time $t = 1$. This is, because the latter SDEs can, in view of Example 6.3, be re-written in the form

$$X^{\bar{\mathbf{h}}_1}_t = X^{\bar{\mathbf{h}}_1}_0 + \mu_1 B^{\mathbf{1}}_t + \sigma_1 B^{\mathbf{2}}_t, \quad X^{\bar{\mathbf{h}}_2}_t = X^{\bar{\mathbf{h}}_2}_0 + \mu_2 B^{\mathbf{1}}_t + \sigma_2(\rho B^{\mathbf{2}}_t + \sqrt{1-\rho^2} B^{\mathbf{3}}_t)$$

for the three independent Brownian motions $B^{\mathbf{1}}_t = M_B([0,t] \times [0,1])$,

$$B^{\mathbf{2}}_t = \int_0^t \int_0^1 \Phi^{-1}(\mathbf{u}_1(u)) M_B(\mathrm{d}s, \mathrm{d}u), \quad B^{\mathbf{3}}_t = \int_0^t \int_0^1 \Phi^{-1}(\mathbf{u}_2(u)) M_B(\mathrm{d}s, \mathrm{d}u).$$

Thus, this part of the example illustrates that our framework allows a flexible dependency structure for the joint execution of several relaxed controls, leading to different joint distributions of the resulting grid-sampling limit SDEs.

We emphasize, that the exploratory SDE has been introduced in [34] as a tool for studying the expected cost for a fixed relaxed control h and, for this purpose, the joint dynamics of the controlled states of several relaxed controls is irrelevant. In contrast, our grid-sampling limit SDE comes up as the limit dynamics of sample trajectories, when executing randomly drawn realizations of a relaxed control. And from this modeling perspective, the joint distribution, when executing several controls, matters.

Let us summarize: By the considerations at the beginning of this subsection $\bar{X}^h$ and $X^{\mathbf{h}}$ have the same probability law, if $\mathbf{h}$ executes h. The SDEs governing these two processes cannot be interpreted as dynamics of the system along a ξ-randomized control. One way to justify these SDEs is to view them as the limit dynamics of the grid-sampling SDE, which has a sound interpretation in terms of ξ^{Π}-randomized controls. By Remark 5.5(2), we observe that the law of $X^{\Pi_n,\mathbf{h}}$ converges to the law of $\bar{X}^h$ under at most technical conditions for one fixed control pair h, $\mathbf{h}$. However, as illustrated by Example 6.4, one cannot hope that the joint convergence result to the grid-sampling limit SDEs indicated in Remark 5.5(2) carries over to the exploratory SDE. We will illustrate in the next subsection that this difference can be essential for the justification of learning algorithms.

6.3. Outlook: Towards learning

In this part, we give a glimpse of how our random measure approach can be applied for supporting the use of some learning algorithms which have been recently suggested in the continuous time RL literature [17–19], but a more thorough study of this aspect is beyond the scope of this paper.

We note that there are two possible ways to come up with learning algorithms in continuous time. The first one is to discretize time first and then to apply one of the established learning algorithms for Markov decision processes [31] to the time-discretized problem. The second way is to identify optimality conditions for the learning task directly in continuous time and to discretize time in a second step. This "first-optimize-then-discretize" approach is pursued by Jia and Zhou [17–19] for the design of policy gradient algorithms and q-learning algorithms in continuous time. The numerical test cases in [19] suggest that the new algorithms based on the continuous-time optimality conditions are very promising, outperforming, e.g., the application of the classical Q-learning to the time-discretized problem. In view of the discussion in Section 4.2 in [19] one reason could be that a stochastic gradient descent implementation of the (discretized versions of the) continuous-time optimality conditions benefits from a variance reduction and, hence, converges in a less noisy way.

The derivation of the learning algorithms of Jia and Zhou via the "first-optimize-then-discretize" approach relies, however, on the idealized randomization mechanism via the sample state process. In view of the issues concerning idealized sampling discussed in Section 3 above, we now explain how to re-derive such algorithms based on the grid-sampling limit SDE in place of the sample state process. For sake of exposition, we here only discuss the policy evaluation stage, which is a building block for the actor-critic policy gradient algorithms [18], in a simplified setting. We expect, however, that the policy optimization step of these algorithms as well as the q-learning algorithms in [19] can be theoretically supported in a similar way; see the general discussion of our approach at the end of this subsection.

We fix a randomized control $\mathbf{h}$ and restrict ourselves to the no-jump case in dimension one ($m = d = p = 1$). Assuming that the conditions in Proposition 5.3 are satisfied, the unique solution of the grid-sampling limit SDE takes the form

$$X^{\mathbf{h}}_t = x_0 + \int_0^t \int_0^1 b(s, X^{\mathbf{h}}_s, \mathbf{h}(s, X^{\mathbf{h}}_s, u))\mathrm{d}u\mathrm{d}s + \int_{(0,t]\times[0,1]} a(s, X^{\mathbf{h}}_s, \mathbf{h}(s, X^{\mathbf{h}}_s, u)) M_B(\mathrm{d}s, \mathrm{d}u).$$

We suppose that the law of $\mathbf{h}(t,x,\eta)$ (where η is a uniform random variable on $[0,1]$) is absolutely continuous with respect to the Lebesgue measure with density $\dot{h}(t,x,\cdot)$ for every $(t,x)\in[0,T]\times\mathbb{R}$ and that its Shannon entropy

$$-\int_{\mathbb{R}}\dot{h}(t,x,y)\log\dot{h}(t,x,y)\mathrm{d}y$$

exists in $\mathbb{R}$ and is measurable and bounded as a function in (t,x). We consider the problem of evaluating the expected terminal cost g and running cost r with a running entropy-regularization term, which rewards exploration, as suggested in [34]. The corresponding cost process is given by

$$\mathcal{J}_t^{\mathbf{h}}=\mathbb{E}\left[g(X_T^{\mathbf{h}})+\int_t^T\int_0^1 r(s,X_s^{\mathbf{h}},\mathbf{h}(s,X_s^{\mathbf{h}},u))\mathrm{d}u\mathrm{d}s+\lambda\int_t^T\int_{\mathbb{R}}\dot{h}(s,X_s^{\mathbf{h}},y)\log\dot{h}(s,X_s^{\mathbf{h}},y)\mathrm{d}y\mathrm{d}s\,\middle|\,\mathcal{F}_t\right]$$

for some fixed temperature parameter $\lambda>0$. We additionally assume, for the sake of simplicity, that the terminal cost g and the running cost r are bounded, and, consequently, the process $\mathcal{J}^{\mathbf{h}}$ is bounded as well. We say that a measurable function $J^{\mathbf{h}}\colon[0,T]\times\mathbb{R}\to\mathbb{R}$ is a version of the *value function* of $\mathbf{h}$, if

$$J^{\mathbf{h}}(t,X_t^{\mathbf{h}})=\mathcal{J}_t^{\mathbf{h}}\quad\mathbb{P}\text{-a.s.},\quad t\in[0,T].$$

The aim of policy evaluation is to learn the value function $J^{\mathbf{h}}$ from observations of the system $X^{\xi,\mathbf{h}}$, when feeding in the ξ-randomized policy $\mathbf{h}(t,x,\xi_t)$ for some randomization process ξ (see Section 2.3), without knowing the true model parameters b,a. Recall that in the simplified setting of this subsection

$$\mathrm{d}X_t^{\xi,\mathbf{h}}=b(t,X_t^{\xi,\mathbf{h}},\mathbf{h}(t,X_t^{\xi,\mathbf{h}},\xi_t))\mathrm{d}t+a(t,X_t^{\xi,\mathbf{h}},\mathbf{h}(t,X_t^{\xi,\mathbf{h}},\xi_t))\mathrm{d}B_t,\quad X_0^{\xi,\mathbf{h}}=x_0.\tag{6.3}$$

Here, $(\mathbf{h}(t,X_t^{\xi,\mathbf{h}},\xi_t))_{t\in[0,T]}$ is the action process of $\mathbf{h}$ under ξ-randomization, i.e., it is the actual control fed into the system by the actor, leading to the observed state variable $(X_t^{\xi,\mathbf{h}})_{t\in[0,T]}$.

The algorithms for policy evaluation derived in [17,18] rely on the martingale characterization of the value function $J^{\mathbf{h}}$, which can be formulated for the grid-sampling limit SDE in the following way (see [2] for the routine proof).

Proposition 6.5.

(1) Suppose that the following partial differential equation has a bounded solution $J\in C^{1,2}([0,T]\times\mathbb{R})$:

$$\frac{\partial J}{\partial t}(t,x)+(\mathcal{L}_hJ(t,\cdot))(t,x)+\int_0^1 r(t,x,\mathbf{h}(t,x,u))\mathrm{d}u+\lambda\int_{\mathbb{R}}\dot{h}(t,x,y)\log\dot{h}(t,x,y)\mathrm{d}y=0,$$

for $(t,x)\in[0,T)\times\mathbb{R}$, with the terminal condition $J(T,\cdot)=g$ (where the differential operator $\mathcal{L}_h$ is defined in (6.1)). Then, J is a version of the value function of $\mathbf{h}$.

(2) Assume that $\tilde{J}\colon[0,T]\times\mathbb{R}\to\mathbb{R}$ is measurable with $\tilde{J}(T,\cdot)=g$. Then, $\tilde{J}$ is a version of the value function of $\mathbf{h}$, if and only if

$$\tilde{J}(t,X_t^{\mathbf{h}})+\int_0^t\int_0^1 r(s,X_s^{\mathbf{h}},\mathbf{h}(s,X_s^{\mathbf{h}},u))\mathrm{d}u\mathrm{d}s+\lambda\int_0^t\int_{\mathbb{R}}\dot{h}(s,X_s^{\mathbf{h}},y)\log\dot{h}(s,X_s^{\mathbf{h}},y)\mathrm{d}y\mathrm{d}s,$$

$0\le t\le T$, is an $\mathbb{F}$-martingale.

We now provide an alternative derivation of the offline variant of the continuous-time TD(0)-algorithm in [17,18] for policy evaluation: To this end, fix a parametric class of functions $\{J_\vartheta:\vartheta\in\Theta\}$ for some open parameter set $\Theta\subseteq\mathbb{R}^L$. We will implicitly assume that the function

$$\mathbf{J}_\Theta\colon[0,T]\times\mathbb{R}\times\Theta\to\mathbb{R},\quad(t,x,\vartheta)\mapsto J_\vartheta(t,x)$$

satisfies sufficient smoothness and boundedness assumptions to justify the manipulations below. Moreover, we postulate that $J_\vartheta(T,\cdot)=g$ for every $\vartheta\in\Theta$. We aim at finding a parameter $\vartheta^*\in\Theta$ such that J_{ϑ^*} is a good approximation to the value function $J^{\mathbf{h}}$ of the randomized control $\mathbf{h}$. Since integrals of sufficiently good integrands with respect to a martingale have zero expectation, the martingale characterization of the value function in Proposition 6.5 motivates to search for a parameter ϑ^* such that

$$\mathbb{E}\left[\int_0^T\nabla_\vartheta\mathbf{J}_\Theta(s,X_s^{\mathbf{h}},\vartheta^*)\left(\mathrm{d}J_{\vartheta^*}(s,X_s^{\mathbf{h}})+\int_0^1 r(s,X_s^{\mathbf{h}},\mathbf{h}(s,X_s^{\mathbf{h}},u))\mathrm{d}u\mathrm{d}s+\lambda\int_{\mathbb{R}}\dot{h}(s,X_s^{\mathbf{h}},y)\log\dot{h}(s,X_s^{\mathbf{h}},y)\mathrm{d}y\mathrm{d}s\right)\right]=0,$$

compare [17]. Here, ∇_ϑ stands for the gradient in the ϑ-variable. Then, the stochastic approximation algorithm of Robbins and Monro [28] suggests to consider the update step

$$\vartheta\leftarrow\vartheta+\alpha\int_0^T\nabla_\vartheta\mathbf{J}_\Theta(s,X_s^{\mathbf{h}},\vartheta)\left(\mathrm{d}J_\vartheta(s,X_s^{\mathbf{h}})+\int_0^1 r(s,X_s^{\mathbf{h}},\mathbf{h}(s,X_s^{\mathbf{h}},u))\mathrm{d}u\mathrm{d}s+\lambda\int_{\mathbb{R}}\dot{h}(s,X_s^{\mathbf{h}},y)\log\dot{h}(s,X_s^{\mathbf{h}},y)\mathrm{d}y\mathrm{d}s\right)\tag{6.4}$$

for some step-size $\alpha>0$. Up to here, the derivation follows the one in [17,18] with the grid-sampling limit SDE in place of the sample SDE of [18]. Note that, although the unknown coefficients b and a do not show up in (6.4), its implementation is infeasible, because $X^{\mathbf{h}}$ is not observable (it is not the state variable of an action process in general; cp. Remark 5.2). We view (6.4) as an idealized continuous-limit update step, which will be discretized next. By Itô's formula, recalling that $\mathcal{L}_h$ in (6.1) is the infinitesimal generator of $X^{\mathbf{h}}$, we obtain

$$\int_0^T\nabla_\vartheta\mathbf{J}_\Theta(s,X_s^{\mathbf{h}},\vartheta)\left(\mathrm{d}J_\vartheta(s,X_s^{\mathbf{h}})+\int_0^1 r(s,X_s^{\mathbf{h}},\mathbf{h}(s,X_s^{\mathbf{h}},u))\mathrm{d}u\mathrm{d}s+\lambda\int_{\mathbb{R}}\dot{h}(s,X_s^{\mathbf{h}},y)\log\dot{h}(s,X_s^{\mathbf{h}},y)\mathrm{d}y\mathrm{d}s\right)$$

$$= \int_0^T \nabla_\vartheta \mathbf{J}_\Theta(s, X_s^{\mathbf{h}}, \vartheta) \frac{\partial J_\vartheta}{\partial t}(s, X_s^{\mathbf{h}}) \mathrm{d}s$$
$$+ \int_0^T \nabla_\vartheta \mathbf{J}_\Theta(s, X_s^{\mathbf{h}}, \vartheta) \Big((\mathcal{L}_h J_\vartheta(s,\cdot))(s, X_s^{\mathbf{h}}) + \int_0^1 r(s, X_s^{\mathbf{h}}, \mathbf{h}(s, X_s^{\mathbf{h}}, u)) \mathrm{d}u + \lambda \int_{\mathbb{R}} \dot{h}(s, X_s^{\mathbf{h}}, y) \log \dot{h}(s, X_s^{\mathbf{h}}, y) \mathrm{d}y \Big) \mathrm{d}s$$
$$+ \int_{(0,T]\times[0,1]} \nabla_\vartheta \mathbf{J}_\Theta(s, X_s^{\mathbf{h}}, \vartheta) a(s, X_s^{\mathbf{h}}, \mathbf{h}(s, X_s^{\mathbf{h}}, u)) \frac{\partial J_\vartheta}{\partial x}(s, X_s^{\mathbf{h}}) M_B(\mathrm{d}s, \mathrm{d}u). \tag{6.5}$$

By change of variables and applying the notation introduced in Proposition 5.3, the second integral on the right-hand side of (6.5) becomes

$$\int_0^T \int_0^1 \nabla_\vartheta \mathbf{J}_\Theta(s, X_s^{\mathbf{h}}, \vartheta) \Big(\frac{1}{2} a_{\mathbf{h}}(s, X_s^{\mathbf{h}}, u)^2 \frac{\partial^2 J_\vartheta}{\partial x^2}(s, X_s^{\mathbf{h}}) + b_{\mathbf{h}}(s, X_s^{\mathbf{h}}, u) \frac{\partial J_\vartheta}{\partial x}(s, X_s^{\mathbf{h}}) + r(s, X_s^{\mathbf{h}}, \mathbf{h}(s, X_s^{\mathbf{h}}, u)) + \lambda \log \dot{h}(s, X_s^{\mathbf{h}}, \mathbf{h}(s, X_s^{\mathbf{h}}, u)) \Big) \mathrm{d}u \mathrm{d}s,$$

which, in fact, is an integral with respect to the limit drift measure M_D. Thus, approximating $(X^{\mathbf{h}}, M_D, M_B)$ by $(X^{\Pi,\mathbf{h}}, M_D^\Pi, M_B^\Pi)$, where $X^{\Pi,\mathbf{h}} := X^{\xi^\Pi,\mathbf{h}}$ is given in (6.3) (i.e., $X^{\Pi,\mathbf{h}}$ solves the grid-sampling SDE), the joint convergence in Remark 5.5(2) suggests that

$$\int_0^T \nabla_\vartheta \mathbf{J}_\Theta(s, X_s^{\mathbf{h}}, \vartheta) \Big(\mathrm{d}J_\vartheta(s, X_s^{\mathbf{h}}) + \int_0^1 r(s, X_s^{\mathbf{h}}, \mathbf{h}(s, X_s^{\mathbf{h}}, u)) \mathrm{d}u \mathrm{d}s + \lambda \int_{\mathbb{R}} \dot{h}(s, X_s^{\mathbf{h}}, y) \log \dot{h}(s, X_s^{\mathbf{h}}, y) \mathrm{d}y \mathrm{d}s \Big)$$

can be approximated in law by

$$\int_{(0,T]\times[0,1]} \nabla_\vartheta \mathbf{J}_\Theta(s, X_s^{\Pi,\mathbf{h}}, \vartheta) \Big(\frac{\partial J_\vartheta}{\partial t}(s, X_s^{\Pi,\mathbf{h}}) + (r + \lambda \log \dot{h})(s, X_s^{\Pi,\mathbf{h}}, \mathbf{h}(s, X_s^{\Pi,\mathbf{h}}, u))$$
$$+ \frac{1}{2} a_{\mathbf{h}}(s, X_s^{\Pi,\mathbf{h}}, u)^2 \frac{\partial^2 J_\vartheta}{\partial x^2}(s, X_s^{\Pi,\mathbf{h}}) + b_{\mathbf{h}}(s, X_s^{\Pi,\mathbf{h}}, u) \frac{\partial J_\vartheta}{\partial x}(s, X_s^{\Pi,\mathbf{h}}) \Big) M_D^\Pi(\mathrm{d}s, \mathrm{d}u)$$
$$+ \int_{(0,T]\times[0,1]} \nabla_\vartheta \mathbf{J}_\Theta(s, X_s^{\Pi,\mathbf{h}}, \vartheta) a_{\mathbf{h}}(s, X_s^{\Pi,\mathbf{h}}, u) \frac{\partial J_\vartheta}{\partial x}(s, X_s^{\Pi,\mathbf{h}}) M_B^\Pi(\mathrm{d}s, \mathrm{d}u)$$

for a sufficiently fine sampling grid Π. In view of Proposition 4.2, and applying Itô's formula once more, this expression equals

$$\sum_{i=1}^n \int_{t_{i-1}}^{t_i} \nabla_\vartheta \mathbf{J}_\Theta(s, X_s^{\Pi,\mathbf{h}}, \vartheta) \Big(\mathrm{d}J_\vartheta(s, X_s^{\Pi,\mathbf{h}}) + (r + \lambda \log \dot{h})(s, X_s^{\Pi,\mathbf{h}}, \mathbf{h}(s, X_s^{\Pi,\mathbf{h}}, \xi_{t_i}^\Pi)) \mathrm{d}s \Big),$$

leading to the modified update step

$$\vartheta \leftarrow \vartheta + \alpha \sum_{i=1}^n \int_{t_{i-1}}^{t_i} \nabla_\vartheta \mathbf{J}_\Theta(s, X_s^{\Pi,\mathbf{h}}, \vartheta) \Big(\mathrm{d}J_\vartheta(s, X_s^{\Pi,\mathbf{h}}) + (r + \lambda \log \dot{h})(s, X_s^{\Pi,\mathbf{h}}, \mathbf{h}(s, X_s^{\Pi,\mathbf{h}}, \xi_{t_i}^\Pi)) \mathrm{d}s \Big). \tag{6.6}$$

Here, the t_i's are, of course, the grid points of the sampling grid Π. We emphasize that the update step (6.6) is independent of the unknown parameters b and a and only depends on observables, namely the action process $\mathbf{a}_t^{\Pi,\mathbf{h}} := \mathbf{h}(t, X_t^{\Pi,\mathbf{h}}, \xi_t^\Pi)$, $t \in [0,T]$ under grid sampling and its state variable $(X_t^{\Pi,\mathbf{h}})_{t\in[0,T]}$. We note that the update-step (6.6) is still formulated in continuous time. For the actual implementation, it is, in view of Remark 4.1, natural to consider the time-discretization relative to the grid Π given by

$$\vartheta \leftarrow \vartheta + \alpha \sum_{i=1}^n \nabla_\vartheta \mathbf{J}_\Theta(t_{i-1}, x_{t_{i-1}}^{\Pi,\mathbf{h}}, \vartheta) \Big[J_\vartheta(t_i, x_{t_i}^{\Pi,\mathbf{h}}) - J_\vartheta(t_{i-1}, x_{t_{i-1}}^{\Pi,\mathbf{h}}) + (t_i - t_{i-1})(r + \lambda \log \dot{h})(t_{i-1}, x_{t_{i-1}}^{\Pi,\mathbf{h}}, \mathbf{h}(t_{i-1}, x_{t_{i-1}}^{\Pi,\mathbf{h}}, \xi_{t_i}^\Pi)) \Big]. \tag{6.7}$$

Here, $x_{t_i}^{\Pi,\mathbf{h}}$ is recursively observed as the state variable of the time-discretized action process $\mathbf{h}(t_{i-1}, x_{t_{i-1}}^{\Pi,\mathbf{h}}, \xi_{t_i}^\Pi)$, which is applied on the time interval $(t_{i-1}, t_i]$ based on a realization of the uniformly (on $[0,1]$) distributed random variable $\xi_{t_i}^\Pi$, which is drawn at time t_{i-1} independently of the past up to time $(t_{i-1})-$. This expression coincides with the policy evaluation update step in Algorithm 1 of [18] for the TD(0) test function $\nabla_\vartheta \mathbf{J}_\Theta(t_{i-1}, x_{t_{i-1}}^{\Pi,\mathbf{h}}, \vartheta)$ in the case of a zero interest rate for discounting. Hence, we have provided a new justification of the continuous-time TD(0)-algorithm for policy evaluation, which avoids making use of idealized sampling.

The derivation of the TD(0) policy evaluation algorithm above is based on the following generic scheme for designing learning algorithms (which we now formulate in a general framework including jumps):

Step I: Derivation of a martingale criterion (for characterizing the value function of a policy, or for optimality) in continuous time in terms of the grid-sampling limit SDE $X^{\mathbf{h}}$ of a randomized control $\mathbf{h}$ in feedback form. [Cp. Proposition 6.5 for policy evaluation]

Step II: Iterative updating (in continuous time) for enforcing the martingale condition via the Robbins-Monro algorithm/stochastic gradient descent based on the pair $(X^{\mathbf{h}}, \mathbf{h})$. [Cp. (6.4)]

Step III: Approximation of the iterative update rule in Step II via grid sampling: replace the grid-sampling limit SDE $X^{\mathbf{h}}$ and the limit random measures of Theorem 3 by the grid sampling SDE $X^{\Pi,\mathbf{h}}$ and the pre-limit random measures of Proposition 4.2 for a sufficiently fine grid Π. Write $\mathbf{a}_t^{\Pi,\mathbf{h}} := \mathbf{h}(t, X_{t-}^{\Pi,\mathbf{h}}, \xi_t^\Pi)$, $t \in [0,T]$, for the action process of $\mathbf{h}$ under grid sampling. [Cp. (6.6)]

Step IV: Time-discretization (piecewise constant in time on Π) of $\mathbf{a}^{\Pi,\mathbf{h}}$ via Remark 4.1 and of all integrals by Riemann sums to come up with an implementable update rule based on observables only. [Cp. (6.7)]

We emphasize that the random measure approach and our main convergence theorem (Theorem 3) are utilized (heuristically) in the crucial Step III: it turns the idealized, non-implementable update rule (Step II) based on the continuous-time martingale criterion

for the unobservable grid-sampling limit SDE (Step I) into an approximate update rule which is formulated in terms of observables only (the action process under grid sampling and its state variable). Once, this step is achieved, a standard quadrature of the integrals (Step IV) is performed to come up with the fully implementable update rule.

Note that the derivations of the learning algorithms by Jia and Zhou [18,19] apply the sample state process (based on idealized sampling) in Steps I and II. As the latter one is considered as an observable, Step III is not required in their approach. However, due to the measurability issues of idealized sampling, Steps I and II with the sample state process cannot be performed mathematically rigorously. In contrast, in our alternative approach, only well-defined objects are applied in each step. In order to better justify the algorithms of [18,19] by re-deriving them via our approach, one basically has to turn the heuristic martingale conditions of [18,19] for the sample state process into rigorously provable martingale conditions for the grid-sampling limit SDE (Step I). Once this as been achieved, it is expected that Steps II–IV can be performed in a routine way.

7. Proof of Theorem 3

7.1. Preliminaries

To avoid double-indexing, we assume that Π_n partitions $[0,T]$ into n subintervals and write $0 = t_0^n < \cdots < t_n^n = T$ for the grid points of Π_n. We emphasize that the same proof also works, even if Π_n decomposes $[0,T]$ into $k(n) \in \mathbb{N}$, which is not necessarily equal to n, subintervals. Denote

$$\mathbf{U} := [0,T] \times [0,1]^d, \quad \mathbf{V} := [0,T] \times \mathbb{R}_0^q \times [0,1]^d.$$

The assumptions imply that $f_l \in B_b(\mathbf{U}; \mathbb{R}^m)$ for $l = 0, \dots, p$ and $f_l \in B_b(\mathbf{V}; \mathbb{R}^m)$ for $l = p+1, p+2$. Moreover, by Remark 2.1,

$$\int_0^T \int_{\mathbb{R}_0^q} (|z|^2 \mathbb{1}_{\{0<|z|\le R\}} + \mathbb{1}_{\{|z|>R\}}) \nu_s(\mathrm{d}z)\mathrm{d}s < \infty, \quad \forall R \in (0,\infty) \cup \{\mathfrak{r}\}. \tag{7.1}$$

In view of Proposition 4.2, we have the representation

$$\begin{aligned} \mathcal{X}_t^n = \sum_{i=1}^n \Bigg[& \int_0^t f_0(s,\xi_i^n) \mathbb{1}_{(t_{i-1}^n, t_i^n]}(s)\mathrm{d}s + \sum_{l=1}^p \int_0^t f_l(s,\xi_i^n) \mathbb{1}_{(t_{i-1}^n, t_i^n]}(s)\mathrm{d}B_s^{(l)} \\ & + \int_0^t \int_{0<|z|\le R} f_{p+1}(s,z,\xi_i^n) \mathbb{1}_{(t_{i-1}^n, t_i^n]}(s)|z|\tilde{N}(\mathrm{d}s,\mathrm{d}z) + \int_0^t \int_{|z|>R} f_{p+2}(s,z,\xi_i^n) \mathbb{1}_{(t_{i-1}^n, t_i^n]}(s) N(\mathrm{d}s,\mathrm{d}z) \Bigg]. \end{aligned} \tag{7.2}$$

We will also consider the piecewise constant interpolation of $\mathcal{X}^n$ between the grid points of Π_n. Introducing the notation

$$\rho_n(t) := \sup\{t_i^n : t_i^n \le t\}, \quad t \in [0,T],$$

it can be written as $\mathcal{X}^n_{\rho_n(t)}$, $t \in [0,T]$.

By Theorem 3.1 in [4], it suffices to show that, as $n \to \infty$,

$$\tilde{d}_T^m(\mathcal{X}^n, \mathcal{X}^n_{\rho_n}) \xrightarrow{\mathbb{P}} 0, \tag{7.3}$$

and

$$\mathcal{X}^n_{\rho_n} \xrightarrow{\mathscr{D}_T} \mathcal{X}, \tag{7.4}$$

where the metric $\tilde{d}_T^m$, which is defined by

$$\tilde{d}_T^m(x,y) := \inf_{\lambda \in \Lambda_T} \max \left\{ \sup_{0\le t\le T} |\lambda(t) - t|, \sup_{0\le t\le T} |x(t) - y(\lambda(t))| \right\},$$

induces the Skorokhod topology on the space $\mathbb{D}_T(\mathbb{R}^m)$ of càdlàg functions $F : [0,T] \to \mathbb{R}^m$, and $\xrightarrow{\mathscr{D}_T}$ stands for convergence in distribution in the Skorokhod space $(\mathbb{D}_T(\mathbb{R}^m), \tilde{d}_T^m)$. Here, Λ_T consists of all strictly increasing and continuous functions $\lambda : [0,T] \to [0,T]$ with $\lambda(0) = 0$, $\lambda(T) = T$. For further details, we refer to [4], or to [2] for a short recap on the Skorokhod space.

The proof of assertions (7.3) and (7.4) will be provided in Section 7.2 and Section 7.3, respectively.

7.2. Proof of assertion (7.3)

Let $\kappa \in (0,\infty) \cap (0,R]$ and let $\kappa = 0$ if $R = 0$. We define the process $\mathcal{X}^{n,\kappa}$ by setting

$$\mathcal{X}^{n,\kappa} := \mathcal{X}^n - \sum_{i=1}^n \int_0^{\cdot} \int_{0<|z|\le\kappa} f_{p+1}(s,z,\xi_i^n) \mathbb{1}_{(t_{i-1}^n, t_i^n]}(s)|z|\tilde{N}(\mathrm{d}s,\mathrm{d}z).$$

By separating $\tilde{N} = N - \nu$ on $[0,T] \times \{\kappa < |z| \le R\}$, which is possible as $\int_0^T \int_{\kappa<|z|\le R} |z|\nu_s(\mathrm{d}z)\mathrm{d}s < \infty$ and f_{p+1} is bounded, and then rearranging terms we get

$$\mathcal{X}^{n,\kappa} = \sum_{i=1}^n \int_0^{\cdot} f_0(s,\xi_i^n) \mathbb{1}_{(t_{i-1}^n, t_i^n]}(s)\mathrm{d}s - \sum_{i=1}^n \int_0^{\cdot} \int_{\kappa<|z|\le R} f_{p+1}(s,z,\xi_i^n) \mathbb{1}_{(t_{i-1}^n, t_i^n]}(s)|z|\nu_s(\mathrm{d}z)\mathrm{d}s$$

$$
\begin{aligned}
&+\sum_{i=1}^{n}\sum_{l=1}^{p}\int_0^{\cdot} f_l(s,\xi_i^n)\mathbb{1}_{(t_{i-1}^n,t_i^n]}(s)\mathrm{d}B_s^{(l)}\\
&+\sum_{i=1}^{n}\int_0^{\cdot}\int_{|z|>\kappa}[f_{p+1}(s,z,\xi_i^n)|z|\mathbb{1}_{\{0<|z|\le R\}}+f_{p+2}(s,z,\xi_i^n)\mathbb{1}_{\{|z|>R\}}]\mathbb{1}_{(t_{i-1}^n,t_i^n]}(s)N(\mathrm{d}s,\mathrm{d}z)\\
=:&\ (\mathcal{X}_D^{n,\kappa}-\mathcal{X}_v^{n,\kappa}+\mathcal{X}_B^{n,\kappa})+\mathcal{X}_J^{n,\kappa}\\
=:&\ \mathcal{X}_C^{n,\kappa}+\mathcal{X}_J^{n,\kappa}.
\end{aligned}
$$

Using the triangle inequality we obtain

$$
\begin{aligned}
\tilde{d}_T^m(\mathcal{X}^n,\mathcal{X}_{\rho_n}^n)&\le \tilde{d}_T^m(\mathcal{X}^n,\mathcal{X}^{n,\kappa})+\tilde{d}_T^m(\mathcal{X}^{n,\kappa},\mathcal{X}_{\rho_n}^{n,\kappa})+\tilde{d}_T^m(\mathcal{X}_{\rho_n}^{n,\kappa},\mathcal{X}_{\rho_n}^n)\\
&\le \sup_{t\in[0,T]}|\mathcal{X}_t^n-\mathcal{X}_t^{n,\kappa}|+\tilde{d}_T^m(\mathcal{X}^{n,\kappa},\mathcal{X}_{\rho_n}^{n,\kappa})+\sup_{t\in[0,T]}|\mathcal{X}_{\rho_n(t)}^{n,\kappa}-\mathcal{X}_{\rho_n(t)}^n|\\
&\le 2\sup_{t\in[0,T]}|\mathcal{X}_t^n-\mathcal{X}_t^{n,\kappa}|+\tilde{d}_T^m(\mathcal{X}^{n,\kappa},\mathcal{X}_{\rho_n}^{n,\kappa}).
\end{aligned}
\tag{7.5}
$$

For $\varepsilon>0$, since $\mathcal{X}^n-\mathcal{X}^{n,\kappa}$ is an $\mathbb{F}^{\Pi_n}$-martingale, applying Doob's maximal inequality yields

$$
\begin{aligned}
\mathbb{P}\left(\left\{\sup_{t\in[0,T]}|\mathcal{X}_t^n-\mathcal{X}_t^{n,\kappa}|>\varepsilon\right\}\right)&\le 4\varepsilon^{-2}\mathbb{E}\left[\int_0^T\int_{0<|z|\le\kappa}\sum_{i=1}^{n}|f_{p+1}(s,z,\xi_i^n)|^2\mathbb{1}_{(t_{i-1}^n,t_i^n]}(s)|z|^2\nu_s(\mathrm{d}z)\mathrm{d}s\right]\\
&\le 4\varepsilon^{-2}\|f_{p+1}\|^2_{B_b(\mathbf{V};\mathbb{R}^m)}\int_0^T\int_{0<|z|\le\kappa}|z|^2\nu_s(\mathrm{d}z)\mathrm{d}s.
\end{aligned}
\tag{7.6}
$$

We now deal with the term $\tilde{d}_T^m(\mathcal{X}^{n,\kappa},\mathcal{X}_{\rho_n}^{n,\kappa})$. Set $\tau_0^n:=0$ and

$$
\tau_i^n:=\inf\{t\in(t_{i-1}^n,t_i^n]:|\Delta L_t|>\kappa\}\wedge t_i^n,\quad i=1,\dots,n,
$$

with the convention $\inf\emptyset:=\infty$, and denote the events $A_i^{n,\kappa}$ by

$$
\begin{aligned}
A_i^{n,\kappa}&:=\left\{\int_{(t_{i-1}^n,t_i^n]\times\{|z|>\kappa\}}N(\mathrm{d}s,\mathrm{d}z)\le 1\right\},\quad i=1,\dots,n-1,\\
A_n^{n,\kappa}&:=\left\{\int_{(t_{n-1}^n,T]\times\{|z|>\kappa\}}N(\mathrm{d}s,\mathrm{d}z)=0\right\}.
\end{aligned}
$$

Then $t_{i-1}^n<\tau_i^n\le t_i^n$ on $A_i^{n,\kappa}$ and $\tau_n^n=T$ on $A_n^{n,\kappa}$. Now, for $\omega\in\cap_{i=1}^n A_i^{n,\kappa}$, we define the function $\lambda=\lambda_{\omega,n,\kappa}:[0,T]\to[0,T]$ which piecewise linearly interpolates the points $(0,0),(\tau_1^n,t_1^n),\dots,(\tau_{n-1}^n,t_{n-1}^n),(\tau_n^n,T)$. Namely,

$$
\lambda(t)=t_{i-1}^n+(t_i^n-t_{i-1}^n)\frac{t-\tau_{i-1}^n}{\tau_i^n-\tau_{i-1}^n},\quad t\in(\tau_{i-1}^n,\tau_i^n],\quad i=1,\dots,n.
$$

Then, λ is a strictly increasing and continuous function with $\lambda(0)=0$, $\lambda(T)=T$. It is clear that, for all $t\in(\tau_{i-1}^n,\tau_i^n]$, $i=1,\dots,n$,

$$
|\lambda(t)-t|\le\max\{t_{i-1}^n-\tau_{i-1}^n,\tau_i^n-t_{i-1}^n\}+(t_i^n-t_{i-1}^n)\frac{t-\tau_{i-1}^n}{\tau_i^n-\tau_{i-1}^n}\le 2|\Pi_n|.
$$

Hence, on $\cap_{i=1}^n A_i^{n,\kappa}$ and for such a choice of λ as above, it follows from the definition of $\tilde{d}_T^m$ and the triangle inequality that

$$
\begin{aligned}
\tilde{d}_T^m(\mathcal{X}^{n,\kappa},\mathcal{X}_{\rho_n}^{n,\kappa})&\le\sup_{t\in[0,T]}|\lambda(t)-t|+\sup_{t\in[0,T]}|\mathcal{X}_t^{n,\kappa}-\mathcal{X}_{\rho_n(\lambda(t))}^{n,\kappa}|\\
&\le 2|\Pi_n|+\sup_{t\in[0,T]}|\mathcal{X}_{C,t}^{n,\kappa}-\mathcal{X}_{C,\rho_n(\lambda(t))}^{n,\kappa}|+\sup_{t\in[0,T]}|\mathcal{X}_{J,t}^{n,\kappa}-\mathcal{X}_{J,\rho_n(\lambda(t))}^{n,\kappa}|\\
&=2|\Pi_n|+\max_{1\le i\le n}\sup_{t\in(t_{i-1}^n,t_i^n]}|\mathcal{X}_{C,t}^{n,\kappa}-\mathcal{X}_{C,\rho_n(\lambda(t))}^{n,\kappa}|+\max_{1\le i\le n}\sup_{t\in[\tau_{i-1}^n,\tau_i^n)}|\mathcal{X}_{J,t}^{n,\kappa}-\mathcal{X}_{J,\rho_n(\lambda(t))}^{n,\kappa}|.
\end{aligned}
$$

Notice that $t_{i-1}^n\in[\tau_{i-1}^n,\tau_i^n)$, $\lambda(t)\in[t_{i-1}^n,t_i^n)$ for $t\in[\tau_{i-1}^n,\tau_i^n)$, and on the event $\cap_{i=1}^n A_i^{n,\kappa}$, $\mathcal{X}_J^{n,\kappa}$ is constant on $[\tau_{i-1}^n,\tau_i^n)$ as it does not have jumps on (τ_{i-1}^n,τ_i^n), it thus implies that

$$
\mathcal{X}_{J,t}^{n,\kappa}=\mathcal{X}_{J,t_{i-1}^n}^{n,\kappa}=\mathcal{X}_{J,\rho_n(\lambda(t))}^{n,\kappa},\quad t\in[\tau_{i-1}^n,\tau_i^n).
$$

Moreover, for $i=1,\dots,n$ and $t\in(t_{i-1}^n,t_i^n]$, we observe that

$$
\begin{aligned}
&\text{for } t\in(t_{i-1}^n,\tau_i^n):\quad t_{i-1}^n<\lambda(t)<t_i^n,\\
&\text{for } t\in[\tau_i^n,t_i^n]:\quad t_i^n\le\lambda(t)\le\lambda(t_i^n)\begin{cases}<\lambda(\tau_{i+1}^n)=t_{i+1}^n & \text{if } i\le n-1\\ =t_i^n & \text{if } i=n,\end{cases}
\end{aligned}
$$

which implies $\rho_n(\lambda(t))\in\{t_{i-1}^n,t_i^n\}$ for $t\in(t_{i-1}^n,t_i^n]$. Summarizing those arguments, on $\cap_{i=1}^n A_i^{n,\kappa}$ we have

$$
\tilde{d}_T^m(\mathcal{X}^{n,\kappa},\mathcal{X}_{\rho_n}^{n,\kappa})\le 2|\Pi_n|+2\max_{1\le i\le n}\sup_{t\in(t_{i-1}^n,t_i^n]}|\mathcal{X}_{C,t}^{n,\kappa}-\mathcal{X}_{C,t_{i-1}^n}^{n,\kappa}|
$$

$$\leq 2\left[|\Pi_n| + \max_{1\leq i\leq n}\sup_{t\in(t_{i-1}^n,t_i^n]}|\mathcal{X}_{D,t}^{n,\kappa}-\mathcal{X}_{D,t_{i-1}^n}^{n,\kappa}| + \max_{1\leq i\leq n}\sup_{t\in(t_{i-1}^n,t_i^n]}|\mathcal{X}_{v,t}^{n,\kappa}-\mathcal{X}_{v,t_{i-1}^n}^{n,\kappa}| + \max_{1\leq i\leq n}\sup_{t\in(t_{i-1}^n,t_i^n]}|\mathcal{X}_{B,t}^{n,\kappa}-\mathcal{X}_{B,t_{i-1}^n}^{n,\kappa}|\right]$$

$$\leq 2\left[|\Pi_n| + \|f_0\|_{B_b(\mathbf{U};\mathbb{R}^m)}|\Pi_n| + \|f_{p+1}\|_{B_b(\mathbf{V};\mathbb{R}^m)}\max_{1\leq i\leq n}\int_{t_{i-1}^n}^{t_i^n}\int_{\kappa<|z|\leq R}|z|\nu_s(\mathrm{d}z)\mathrm{d}s + \sum_{l=1}^p\max_{1\leq i\leq n}\sup_{t\in(t_{i-1}^n,t_i^n]}\left|\int_{t_{i-1}^n}^t f_l(s,\xi_i^n)\mathrm{d}B_s^{(l)}\right|\right]. \tag{7.7}$$

For any $\varepsilon>0$,

$$\mathbb{P}(\{\tilde{d}_T^m(\mathcal{X}^{n,\kappa},\mathcal{X}_{\rho_n}^{n,\kappa})>\varepsilon\}) \leq \mathbb{P}\left(\bigcup_{i=1}^n(A_i^{n,\kappa})^c\right) + \mathbb{P}\left(\{\tilde{d}_T^m(\mathcal{X}^{n,\kappa},\mathcal{X}_{\rho_n}^{n,\kappa})>\varepsilon\}\cap\bigcap_{i=1}^n A_i^{n,\kappa}\right). \tag{7.8}$$

For the first term on the right-hand side, letting $x_i := \int_{t_{i-1}^n}^{t_i^n}\int_{|z|>\kappa}\nu_s(\mathrm{d}z)\mathrm{d}s$ and using the inequality $\mathrm{e}^x-1-x\leq\frac{1}{2}\mathrm{e}^K x^2$ for $x\in[0,K]$, we obtain

$$\mathbb{P}\left(\bigcup_{i=1}^n(A_i^{n,\kappa})^c\right) \leq \sum_{i=1}^n(1-\mathbb{P}(A_i^{n,\kappa})) = \sum_{i=1}^{n-1}(1-\mathrm{e}^{-x_i}-x_i\mathrm{e}^{-x_i})+1-\mathrm{e}^{-x_n}$$

$$\leq \frac{1}{2}\mathrm{e}^{\max_{1\leq i\leq n-1}x_i}\sum_{i=1}^{n-1}\mathrm{e}^{-x_i}x_i^2 + x_n \leq \frac{1}{2}\mathrm{e}^{\max_{1\leq i\leq n-1}x_i}\left(\max_{1\leq i\leq n-1}x_i\right)\sum_{i=1}^{n-1}x_i + x_n.$$

Since $\int_0^T\int_{|z|>\kappa}\nu_s(\mathrm{d}z)\mathrm{d}s<\infty$ which ensures the uniform continuity of $[0,T]\ni t\mapsto\int_0^t\int_{|z|>\kappa}\nu_s(\mathrm{d}z)\mathrm{d}s$, we deduce that $\max_{1\leq i\leq n}x_i\to 0$ as $n\to\infty$. Hence,

$$\mathbb{P}\left(\bigcup_{i=1}^n(A_i^{n,\kappa})^c\right)\to 0 \quad \text{as } n\to\infty. \tag{7.9}$$

For the second term, since $\max_{1\leq i\leq n}\int_{t_{i-1}^n}^{t_i^n}\int_{\kappa<|z|\leq R}|z|\nu_s(\mathrm{d}z)\mathrm{d}s\to 0$ as $n\to\infty$ due to the uniform continuity, we deduce from (7.7) that, when n is sufficiently large,

$$\mathbb{P}\left(\{\tilde{d}_T^m(\mathcal{X}^{n,\kappa},\mathcal{X}_{\rho_n}^{n,\kappa})>\varepsilon\}\cap\bigcap_{i=1}^n A_i^{n,\kappa}\right) \leq \mathbb{P}\left(\left\{\sum_{l=1}^p\max_{1\leq i\leq n}\sup_{t\in(t_{i-1}^n,t_i^n]}\left|\int_{t_{i-1}^n}^t f_l(s,\xi_i^n)\mathrm{d}B_s^{(l)}\right|>\frac{\varepsilon}{4}\right\}\right).$$

Applying the Burkholder–Davis–Gundy inequality with the exponent 4 yields

$$\begin{aligned}
&\mathbb{P}\left(\{\tilde{d}_T^m(\mathcal{X}^{n,\kappa},\mathcal{X}_{\rho_n}^{n,\kappa})>\varepsilon\}\cap\bigcap_{i=1}^n A_i^{n,\kappa}\right)\\
&\leq \sum_{l=1}^p\sum_{i=1}^n\mathbb{P}\left(\left\{\sup_{t\in(t_{i-1}^n,t_i^n]}\left|\int_{t_{i-1}^n}^t f_l(s,\xi_i^n)\mathrm{d}B_s^{(l)}\right|>\frac{\varepsilon}{4p}\right\}\right)\\
&\leq c\frac{256p^4}{\varepsilon^4}\sum_{l=1}^p\sum_{i=1}^n\mathbb{E}\left[\left|\int_{t_{i-1}^n}^{t_i^n}|f_l(s,\xi_i^n)|^2\mathrm{d}s\right|^2\right]\\
&\leq c\frac{256p^5}{\varepsilon^4}\max_{1\leq l\leq p}\|f_l\|_{B_b(\mathbf{U};\mathbb{R}^m)}^4\sum_{i=1}^n(t_i^n-t_{i-1}^n)^2\\
&\leq c\frac{256p^5T}{\varepsilon^4}\max_{1\leq l\leq p}\|f_l\|_{B_b(\mathbf{U};\mathbb{R}^m)}^4|\Pi_n|\xrightarrow{n\to\infty}0,
\end{aligned} \tag{7.10}$$

where $c>0$ is a constant independent of ε,n,p,T. Combining (7.9) and (7.10) with (7.8), and then plugging them together with (7.6) into (7.5) we arrive at

$$\limsup_{n\to\infty}\mathbb{P}(\{\tilde{d}_T^m(\mathcal{X}^n,\mathcal{X}_{\rho_n}^n)>3\varepsilon\}) \leq 4\varepsilon^{-2}\|f_{p+1}\|_{B_b(\mathbf{V};\mathbb{R}^m)}^2\int_0^T\int_{0<|z|\leq\kappa}|z|^2\nu_s(\mathrm{d}z)\mathrm{d}s.$$

Letting $\kappa\downarrow 0$ and exploiting (7.1) we eventually obtain

$$\limsup_{n\to\infty}\mathbb{P}(\{\tilde{d}_T^m(\mathcal{X}^n,\mathcal{X}_{\rho_n}^n)>3\varepsilon\})=0,$$

which then verifies (7.3). □

7.3. Proof of assertion (7.4)

For the proof of (7.4), we apply a limit theorem of Jacod and Shiryaev, which is briefly reviewed in [2]. It relies on verifying the convergence of the modified semimartingale characteristics of $\mathcal{X}_{\rho_n}^n$ to the modified semimartingale characteristics of the limit process $\mathcal{X}$. Here, "modified" is understood in the sense of [15, Definition II.2.16].

Let us fix a truncation function $\mathfrak{h}\colon \mathbb{R}^m \to \mathbb{R}^m$, see [15, Definition II.2.3], i.e. $\mathfrak{h}$ is bounded and $\mathfrak{h}(z) = z$ in a neighborhood of 0. It is convenient for us to assume furthermore that $\mathfrak{h}^{(k)} \in C_b^2(\mathbb{R}^m)$ for any $k = 1, \dots, m$.

The following lemma states the semimartingale characteristics of $\mathcal{X}$ with respect to the truncation function $\mathfrak{h}$, compare [15, Definition II.2.6]. Its proof follows routine arguments and can be found in [2].

Lemma 7.1. *$\mathcal{X}$ is an m-dimensional semimartingale whose characteristics $(\mathfrak{b}^{\mathcal{X}}, C^{\mathcal{X}}, \nu^{\mathcal{X}})$ with respect to the truncation function $\mathfrak{h}$ is given by*

$$\mathfrak{b}_t^{\mathcal{X}} = \int_0^t \left[\int_{[0,1]^d} f_0(s,u)\mathrm{d}u + \int_{\{|z|>R\}\times[0,1]^d} \mathfrak{h}(f_{p+2}(s,z,u))\nu_s(\mathrm{d}z)\mathrm{d}u + \int_{\{0<|z|\le R\}\times[0,1]^d} [\mathfrak{h}(f_{p+1}(s,z,u)|z|) - f_{p+1}(s,z,u)|z|]\nu_s(\mathrm{d}z)\mathrm{d}u \right] \mathrm{d}s,$$

$$C_t^{\mathcal{X}} = \left(\sum_{l=1}^p \int_0^t \int_{[0,1]^d} (f_l^{(k)} f_l^{(k')})(s,u)\mathrm{d}u\mathrm{d}s \right)_{k,k'} \in \mathbb{R}^{m\times m}, \quad 0 \le t \le T,$$

$$\nu^{\mathcal{X}}((s,t]\times A) = \int_s^t \int_{\{0<|z|\le R\}\times[0,1]^d} \mathbb{1}_A(f_{p+1}(r,z,u)|z|)\nu_r(\mathrm{d}z)\mathrm{d}u\mathrm{d}r + \int_s^t \int_{\{|z|>R\}\times[0,1]^d} \mathbb{1}_A(f_{p+2}(r,z,u))\nu_r(\mathrm{d}z)\mathrm{d}u\mathrm{d}r$$

for $0 \le s < t \le T$, $A \in \mathcal{B}(\mathbb{R}_0^m)$.

Remark 7.2. By a standard approximation argument, the measure $\nu^{\mathcal{X}}$ in Lemma 7.1 satisfies

$$\int_0^T \int_{\mathbb{R}_0^m} g(y)\nu^{\mathcal{X}}(\mathrm{d}s,\mathrm{d}y) = \int_0^T \int_{\mathbb{R}_0^q\times[0,1]^d} [g(f_{p+1}(s,z,u)|z|)\mathbb{1}_{\{0<|z|\le R\}} + g(f_{p+2}(s,z,u))\mathbb{1}_{\{|z|>R\}}]\nu_s(\mathrm{d}z)\mathrm{d}u\mathrm{d}s$$

for any measurable $g\colon \mathbb{R}_0^m \to \mathbb{R}$ which is non-negative or $g\mathbb{1}_{[0,T]}$ is $\nu^{\mathcal{X}}$-integrable. In particular, for $g(y) = \mathbb{1}_{\{|y|\ge\kappa\}}$ with some $\kappa > 0$ we get

$$\begin{aligned}
&\int_0^T \int_{|y|\ge\kappa} \nu^{\mathcal{X}}(\mathrm{d}s,\mathrm{d}y) \\
&= \int_0^T \int_{\mathbb{R}_0^q\times[0,1]^d} [\mathbb{1}_{\{|f_{p+1}(s,z,u)||z|\ge\kappa\}}\mathbb{1}_{\{0<|z|\le R\}} + \mathbb{1}_{\{|f_{p+2}(s,z,u)|\ge\kappa\}}\mathbb{1}_{\{|z|>R\}}]\nu_s(\mathrm{d}z)\mathrm{d}u\mathrm{d}s \\
&\le \frac{\|f_{p+1}\|^2_{B_b(\mathbf{V};\mathbb{R}^m)}}{\kappa^2} \int_0^T \int_{0<|z|\le R} |z|^2\nu_s(\mathrm{d}z)\mathrm{d}s + \frac{\|f_{p+2}\|_{B_b(\mathbf{V};\mathbb{R}^m)}}{\kappa} \int_0^T \int_{|z|>R} \nu_s(\mathrm{d}z)\mathrm{d}s \qquad (7.11) \\
&< \infty,
\end{aligned}$$

where the finiteness can be derived from (7.1) and the inequalities

$$\mathbb{1}_{\{|f_{p+1}(s,z,u)||z|\ge\kappa\}} \le \kappa^{-2}\|f_{p+1}\|^2_{B_b(\mathbf{V};\mathbb{R}^m)}|z|^2 \quad \text{and} \quad \mathbb{1}_{\{|f_{p+2}(s,z,u)|\ge\kappa\}} \le \kappa^{-1}\|f_{p+2}\|_{B_b(\mathbf{V};\mathbb{R}^m)}.$$

We now turn to $\mathcal{X}^n_{\rho_n}$, whose modified semimartingale characteristics will be computed in relation to a new filtration, which we construct next. To this end, we set

$$\sigma_n(t) := \sup\{i : t_i^n \le t\} \in \{0,1,\dots,n\}, \quad t \in [0,\infty).$$

Denote $\Delta_i^n\mathcal{X}^n := \mathcal{X}^n_{t_i^n} - \mathcal{X}^n_{t_{i-1}^n}$. Then

$$\mathcal{X}^n_{\rho_n(t)} = \sum_{i=1}^{\sigma_n(t)} \Delta_i^n\mathcal{X}^n, \quad t \in [0,T].$$

For $n \ge 1$, we define the discrete-time filtration $(\mathcal{G}_i^n)_{i=0}^n$ by

$$\mathcal{G}_0^n := \{\emptyset,\Omega\}, \quad \mathcal{G}_i^n := \sigma\{\Delta_j^n\mathcal{X}^n, j \le i\}, \quad i = 1,\dots,n.$$

Then $\{\Delta_i^n\mathcal{X}^n, \mathcal{G}_i^n : 1 \le i \le n, n \ge 1\}$ is an adapted triangular array. Since $\Delta_i^n\mathcal{X}^n$ is independent of $\mathcal{G}_{i-1}^n$, we get for any bounded measurable g and $t \in [0,\infty)$ that, a.s.,

$$\sum_{i=1}^{\sigma_n(t)} \mathbb{E}[g(\Delta_i^n\mathcal{X}^n)|\mathcal{G}_{i-1}^n] = \sum_{i=1}^{\sigma_n(t)} \mathbb{E}[g(\Delta_i^n\mathcal{X}^n)].$$

Remark 7.3.

(1) By [15, Ch.II, §3b], the modified semimartingale characteristics of $\mathcal{X}^n_{\rho_n}$ with respect to the filtration $\mathbb{G}^{\sigma_n} = (\mathcal{G}^n_{\sigma_n(t)})_{t\ge0}$ is the triplet (drift part, modified diffusion part, jump part) which is respectively described by

$$\sum_{i=1}^{\sigma_n} \mathbb{E}[\mathfrak{h}(\Delta_i^n\mathcal{X}^n)],$$

$$\left(\sum_{i=1}^{\sigma_n} \left(\mathbb{E}[(\mathfrak{h}^{(k)}\mathfrak{h}^{(k')})(\Delta_i^n\mathcal{X}^n)] - \mathbb{E}[\mathfrak{h}^{(k)}(\Delta_i^n\mathcal{X}^n)]\mathbb{E}[\mathfrak{h}^{(k')}(\Delta_i^n\mathcal{X}^n)] \right) \right)_{k,k'=1,\dots,m},$$

$$\sum_{i=1}^{\sigma_n} \mathbb{E}[g(\Delta_i^n \mathcal{X}^n)],$$

where g runs through a sufficiently large class of test functions vanishing around zero.

(2) A key difference between $\mathbb{G}^{\sigma_n}$ and $\mathbb{F}^{\Pi_n}$ is that information about the random variable $\xi^{\Pi_n}_{t_i^n}$, which is sampled for the randomization on the interval $(t_{i-1}^n, t_i^n]$, is only revealed at time t_i^n in the filtration $\mathbb{G}^{\sigma_n}$, whereas it is already known at time t_{i-1}^n in the filtration $\mathbb{F}^{\Pi_n}$.

The following proposition plays the key role for deriving the convergence of the semimartingale characteristics.

Proposition 7.4. *For any $g \in C_b^2(\mathbb{R}^m)$, one has*

$$\sum_{i=1}^{n} \left| \mathbb{E}[g(\Delta_i^n \mathcal{X}^n)] - g(0) - \int_{t_{i-1}^n}^{t_i^n} \Psi_f(g)(s)\mathrm{d}s \right| \xrightarrow{n\to\infty} 0, \tag{7.12}$$

where the function $\Psi_f(g)\colon [0,T] \to \mathbb{R}$ is defined by

$$\begin{aligned}\Psi_f(g)(s) :=& \int_{[0,1]^d} \left(\nabla g(0)^\top f_0(s,u) + \frac{1}{2} \sum_{k,k'=1}^{m} \partial^2_{k,k'} g(0) \sum_{l=1}^{p} (f_l^{(k)} f_l^{(k')})(s,u) \right) \mathrm{d}u \\ &+ \int_{\{0<|z|\le R\}\times[0,1]^d} \Big[g(f_{p+1}(s,z,u)|z|) - g(0) - |z| \nabla g(0)^\top f_{p+1}(s,z,u) \Big] \nu_s(\mathrm{d}z)\mathrm{d}u \\ &+ \int_{\{|z|>R\}\times[0,1]^d} \Big[g(f_{p+2}(s,z,u)) - g(0) \Big] \nu_s(\mathrm{d}z)\mathrm{d}u. \end{aligned} \tag{7.13}$$

Consequently, for any $t \in [0,\infty)$,

$$\sum_{i=1}^{\sigma_n(t)} \mathbb{E}[g(\Delta_i^n \mathcal{X}^n)] \xrightarrow{n\to\infty} g(0) + \int_0^{t\wedge T} \Psi_f(g)(s)\mathrm{d}s.$$

Proof. **Step 1.** It is obvious that $\Psi_f(g)$ is measurable by Fubini's theorem, and moreover, there exists a constant $c_{T,m} > 0$ such that

$$\begin{aligned}\int_0^T |\Psi_f(g)(s)|\mathrm{d}s \le c_{T,m} \Bigg(& \|f_0\|_{B_b(\mathbf{U};\mathbb{R}^m)} |\nabla g(0)| + \sum_{k,k'=1}^{m} |\partial^2_{k,k'} g(0)| \sum_{l=1}^{p} \|f_l^{(k)} f_l^{(k')}\|_{B_b(\mathbf{U})} \\ &+ \|f_{p+1}\|^2_{B_b(\mathbf{V};\mathbb{R}^m)} \|\nabla^2 g\|_{B_b(\mathbb{R}^m;\mathbb{R}^{m\times m})} \int_0^T \int_{\{0<|z|\le R\}\times[0,1]^d} |z|^2 \nu_s(\mathrm{d}z)\mathrm{d}u\mathrm{d}s \\ &+ 2\|g\|_{B_b(\mathbb{R}^m)} \int_0^T \int_{\{|z|>R\}\times[0,1]^d} \nu_s(\mathrm{d}z)\mathrm{d}u\mathrm{d}s \Bigg) \\ < & \infty. \end{aligned}$$

Next, for $n \ge 1, i = 1, \dots, n$, we define the càdlàg and $\mathbb{F}^{\Pi_n}$-adapted process $F^{n,i} = (F_t^{n,i})_{t\in[t_{i-1}^n, t_i^n]}$ null at t_{i-1}^n by setting, for $t \in (t_{i-1}^n, t_i^n]$,

$$F_t^{n,i} := \int_{t_{i-1}^n}^t f_0(s,\xi_i^n)\mathrm{d}s + \sum_{l=1}^{p} \int_{t_{i-1}^n}^t f_l(s,\xi_i^n)\mathrm{d}B_s^{(l)} + \int_{t_{i-1}^n}^t \int_{0<|z|\le R} f_{p+1}(s,z,\xi_i^n)|z|\tilde{N}(\mathrm{d}s,\mathrm{d}z) + \int_{t_{i-1}^n}^t \int_{|z|>R} f_{p+2}(s,z,\xi_i^n) N(\mathrm{d}s,\mathrm{d}z).$$

Let $s \in (0,T]$ be now fixed. Then, for any $n \ge 1$, there exists uniquely $1 \le i(s,n) \le n$ such that

$$s \in (t_{i(s,n)-1}^n, t_{i(s,n)}^n] \quad \text{and} \quad \lim_{n\to\infty} t_{i(s,n)-1}^n = \lim_{n\to\infty} t_{i(s,n)}^n = s.$$

We claim that

$$F_s^{n,i(s,n)} \xrightarrow{\mathbf{L}^1(\mathbb{P})} 0 \quad \text{as } n \to \infty.$$

It is straightforward to check when $n \to \infty$ that, in the representation of $F_s^{n,i(s,n)}$, the Lebesgue integral part tends to 0 in $\mathbf{L}^2(\mathbb{P})$ as f_0 is bounded, the martingale part converges to 0 in $\mathbf{L}^2(\mathbb{P})$ by applying Itô's isometry and using the boundedness of f_l, $l = 1, \dots, p+1$. For the "large jump part", since $\nu_r(\mathrm{d}z)\mathrm{d}r$ is the predictable compensator of $N(\mathrm{d}r,\mathrm{d}z)$, together with (7.1), we get

$$\begin{aligned}&\mathbb{E}\left[\left| \int_{t_{i(s,n)-1}^n}^s \int_{|z|>R} f_{p+2}(r,z,\xi_{i(s,n)}^n) N(\mathrm{d}r,\mathrm{d}z) \right| \right] \le \|f_{p+2}\|_{B_b(\mathbf{V};\mathbb{R}^m)} \mathbb{E}\left[\int_{t_{i(s,n)-1}^n}^s \int_{|z|>R} N(\mathrm{d}r,\mathrm{d}z) \right] \\ &= \|f_{p+2}\|_{B_b(\mathbf{V};\mathbb{R}^m)} \int_{t_{i(s,n)-1}^n}^s \int_{|z|>R} \nu_r(\mathrm{d}z)\mathrm{d}r \xrightarrow{n\to\infty} 0, \end{aligned}$$

which verifies the claim. Since $\mathbb{E}[N(\{s\}\times\mathbb{R}_0^q)] = \nu(\{s\}\times\mathbb{R}_0^q) = 0$, it holds that $F_s^{n,i(s,n)} = F_{s-}^{n,i(s,n)}$ a.s., and hence,

$$F_{s-}^{n,i(s,n)} \xrightarrow{\mathbf{L}^1(\mathbb{P})} 0 \quad \text{as } n \to \infty. \tag{7.14}$$

Step 2. Using Itô's formula for $F^{n,i}$ and $g \in C_b^2(\mathbb{R}^m)$ (see, e.g., [24, Theorem 2.5]) we get, a.s.,

$$
\begin{aligned}
&g(\Delta_i^n \mathcal{X}^n) = g(F^{n,i}_{t_i^n}) \\
&= g(0) + \int_{t_{i-1}^n}^{t_i^n} \nabla g(F^{n,i}_{s-})^\top f_0(s,\xi_i^n)\mathrm{d}s \\
&\quad + \sum_{l=1}^{p} \int_{t_{i-1}^n}^{t_i^n} \nabla g(F^{n,i}_{s-})^\top f_l(s,\xi_i^n)\mathrm{d}B_s^{(l)} + \frac{1}{2}\sum_{k,k'=1}^{m}\sum_{l=1}^{p}\int_{t_{i-1}^n}^{t_i^n} \partial^2_{k,k'} g(F^{n,i}_{s-})(f_l^{(k)} f_l^{(k')})(s,\xi_i^n)\mathrm{d}s \\
&\quad + \int_{t_{i-1}^n}^{t_i^n}\int_{0<|z|\le R}\Big[g\Big(F^{n,i}_{s-} + f_{p+1}(s,z,\xi_i^n)|z|\Big) - g(F^{n,i}_{s-})\Big]\tilde{N}(\mathrm{d}s,\mathrm{d}z) \\
&\quad + \int_{t_{i-1}^n}^{t_i^n}\int_{0<|z|\le R}\Big[g\Big(F^{n,i}_{s-} + f_{p+1}(s,z,\xi_i^n)|z|\Big) - g(F^{n,i}_{s-}) - |z|\nabla g(F^{n,i}_{s-})^\top f_{p+1}(s,z,\xi_i^n)\Big]\nu_s(\mathrm{d}z)\mathrm{d}s \\
&\quad + \int_{t_{i-1}^n}^{t_i^n}\int_{|z|>R}\Big[g\Big(F^{n,i}_{s-} + f_{p+2}(s,z,\xi_i^n)\Big) - g(F^{n,i}_{s-})\Big]N(\mathrm{d}s,\mathrm{d}z).
\end{aligned}
$$

Since ∇g and f_l are bounded for any $l = 1,\dots,p+1$, the integrals with respect to the Brownian motions and the compensated random measure are square integrable martingales which vanish after taking the expectation $\mathbb{E}$. Let us now investigate the remaining parts.

- *The "drift part":* Using Fubini's theorem and the Cauchy–Schwarz inequality yields

$$
\begin{aligned}
&\sum_{i=1}^{n}\left|\mathbb{E}\left[\int_{t_{i-1}^n}^{t_i^n}\nabla g(F^{n,i}_{s-})^\top f_0(s,\xi_i^n)\mathrm{d}s\right] - \int_{t_{i-1}^n}^{t_i^n}\int_{[0,1]^d}\nabla g(0)^\top f_0(s,u)\mathrm{d}u\mathrm{d}s\right| \\
&= \sum_{i=1}^{n}\left|\mathbb{E}\left[\int_{t_{i-1}^n}^{t_i^n}\nabla g(F^{n,i}_{s-})^\top f_0(s,\xi_i^n)\mathrm{d}s - \int_{t_{i-1}^n}^{t_i^n}\nabla g(0)^\top f_0(s,\xi_i^n)\mathrm{d}s\right]\right| \\
&\le \|f_0\|_{B_b(\mathbf{U};\mathbb{R}^m)}\sum_{i=1}^{n}\int_{t_{i-1}^n}^{t_i^n}\mathbb{E}[|\nabla g(F^{n,i}_{s-}) - \nabla g(0)|]\mathrm{d}s \\
&= \|f_0\|_{B_b(\mathbf{U};\mathbb{R}^m)}\int_0^T \mathbb{E}\left[\sum_{i=1}^{n}|\nabla g(F^{n,i}_{s-}) - \nabla g(0)|\mathbb{1}_{(t_{i-1}^n,t_i^n]}(s)\right]\mathrm{d}s \\
&= \|f_0\|_{B_b(\mathbf{U};\mathbb{R}^m)}\int_0^T \mathbb{E}\left[\left|\nabla g\Big(F^{n,i(s,n)}_{s-}\Big) - \nabla g(0)\right|\right]\mathrm{d}s \\
&\xrightarrow{n\to\infty} 0,
\end{aligned}
$$

  where we apply the dominated convergence theorem using (7.14) together with the continuity and boundedness of ∇g. Analogously, for $k,k' = 1,\dots,m$ and $l = 1,\dots,p$,

$$
\sum_{i=1}^{n}\left|\mathbb{E}\left[\int_{t_{i-1}^n}^{t_i^n}\partial^2_{k,k'}g(F^{n,i}_{s-})(f_l^{(k)}f_l^{(k')})(s,\xi_i^n)\mathrm{d}s\right] - \int_{t_{i-1}^n}^{t_i^n}\int_{[0,1]^d}\partial^2_{k,k'}g(0)(f_l^{(k)}f_l^{(k')})(s,u)\mathrm{d}u\mathrm{d}s\right| \xrightarrow{n\to\infty} 0.
$$

- *The "small jump part":* For $i(n,s)$ introduced in **Step 1** one has

$$
\begin{aligned}
&\sum_{i=1}^{n}\left|\mathbb{E}\left[\int_{t_{i-1}^n}^{t_i^n}\int_{0<|z|\le R}\Big[g\Big(F^{n,i}_{s-} + f_{p+1}(s,z,\xi_i^n)|z|\Big) - g(F^{n,i}_{s-}) - |z|\nabla g(F^{n,i}_{s-})^\top f_{p+1}(s,z,\xi_i^n)\Big]\nu_s(\mathrm{d}z)\mathrm{d}s\right]\right. \\
&\qquad \left. - \int_{t_{i-1}^n}^{t_i^n}\int_{\{0<|z|\le R\}\times[0,1]^d}\Big[g(f_{p+1}(s,z,u)|z|) - g(0) - |z|\nabla g(0)^\top f_{p+1}(s,z,u)\Big]\nu_s(\mathrm{d}z)\mathrm{d}u\mathrm{d}s\right| \\
&\le \int_0^T\int_{0<|z|\le R}\mathbb{E}\left[\sum_{i=1}^{n}\left|g\Big(F^{n,i}_{s-} + f_{p+1}(s,z,\xi_i^n)|z|\Big) - g(F^{n,i}_{s-}) - |z|\nabla g(F^{n,i}_{s-})^\top f_{p+1}(s,z,\xi_i^n)\right.\right. \\
&\qquad \left.\left. - g(f_{p+1}(s,z,\xi_i^n)|z|) + g(0) + |z|\nabla g(0)^\top f_{p+1}(s,z,\xi_i^n)\right|\mathbb{1}_{(t_{i-1}^n,t_i^n]}(s)\right]\nu_s(\mathrm{d}z)\mathrm{d}s \\
&= \int_0^T\int_{0<|z|\le R}\mathbb{E}\left[\left|g\Big(F^{n,i(s,n)}_{s-} + f_{p+1}(s,z,\xi^n_{i(s,n)})|z|\Big) - g\Big(F^{n,i(s,n)}_{s-}\Big) - |z|\nabla g\Big(F^{n,i(s,n)}_{s-}\Big)^\top f_{p+1}(s,z,\xi^n_{i(s,n)})\right.\right. \\
&\qquad \left.\left. - g(f_{p+1}(s,z,\xi^n_{i(s,n)})|z|) + g(0) + |z|\nabla g(0)^\top f_{p+1}(s,z,\xi^n_{i(s,n)})\right|\right]\nu_s(\mathrm{d}z)\mathrm{d}s \\
&=: \int_0^T\int_{0<|z|\le R}\mathbb{E}[G_n^{\mathrm{S}}(s,z)]\nu_s(\mathrm{d}z)\mathrm{d}s.
\end{aligned}
$$

  Using Taylor's expansion we obtain a constant $c_m > 0$ depending only on m such that

$$
G_n^{\mathrm{S}}(s,z) \le c_m\|\nabla^2 g\|_{B_b(\mathbb{R}^m;\mathbb{R}^{m\times m})}\|f_{p+1}\|^2_{B_b(\mathbf{V};\mathbb{R}^m)}|z|^2.
$$

Hence, it is easy to check using (7.14) and dominated convergence that $\mathbb{E}[G_n^{\mathrm{S}}(s,z)] \to 0$ as $n \to \infty$ for any s, z. Due to (7.1), dominated convergence also yields

$$\int_0^T \int_{0<|z|\le R} \mathbb{E}[G_n^{\mathrm{S}}(s,z)] \nu_s(\mathrm{d}z)\mathrm{d}s \xrightarrow{n\to\infty} 0.$$

- *The "large jump part":* Since $\nu_s(\mathrm{d}z)\mathrm{d}s$ is the predictable compensator of $N(\mathrm{d}s,\mathrm{d}z)$, using Fubini's theorem, again, for interchanging integrals we get

$$\begin{aligned}
&\sum_{i=1}^n \left| \mathbb{E}\left[\int_{t_{i-1}^n}^{t_i^n} \int_{|z|>R} \left[g\Big(F_{s-}^{n,i} + f_{p+2}(s,z,\xi_i^n) \Big) - g(F_{s-}^{n,i}) \right] N(\mathrm{d}s,\mathrm{d}z) \right] - \int_{t_{i-1}^n}^{t_i^n} \int_{\{|z|>R\}\times[0,1]^d} \Big[g(f_{p+2}(s,z,u)) - g(0) \Big] \nu_s(\mathrm{d}z)\mathrm{d}u\mathrm{d}s \right] \right| \\
&= \sum_{i=1}^n \left| \mathbb{E}\left[\int_{t_{i-1}^n}^{t_i^n} \int_{|z|>R} \left[g\Big(F_{s-}^{n,i} + f_{p+2}(s,z,\xi_i^n) \Big) - g(F_{s-}^{n,i}) \right] \nu_s(\mathrm{d}z)\mathrm{d}s \right] - \mathbb{E}\left[\int_{t_{i-1}^n}^{t_i^n} \int_{|z|>R} \Big[g(f_{p+2}(s,z,\xi_i^n)) - g(0) \Big] \nu_s(\mathrm{d}z)\mathrm{d}s \right] \right| \\
&\le \int_0^T \int_{|z|>R} \mathbb{E}\left[\left| g\Big(F_{s-}^{n,i(s,n)} + f_{p+2}(s,z,\xi_{i(s,n)}^n) \Big) - g\Big(F_{s-}^{n,i(s,n)} \Big) - g(f_{p+2}(s,z,\xi_{i(s,n)}^n)) + g(0) \right| \right] \nu_s(\mathrm{d}z)\mathrm{d}s \\
&=: \int_0^T \int_{|z|>R} \mathbb{E}[G_n^{\mathrm{L}}(s,z)] \nu_s(\mathrm{d}z)\mathrm{d}s.
\end{aligned}$$

It is clear that G_n^{L} is uniformly bounded by $4\|g\|_{B_b(\mathbb{R}^m)}$. Moreover, using the Lipschitzian of g, the boundedness of f_{p+2} and (7.14) and (7.1), we may apply the dominated convergence theorem to obtain

$$\int_0^T \int_{|z|>R} \mathbb{E}[G_n^{\mathrm{L}}(s,z)] \nu_s(\mathrm{d}z)\mathrm{d}s \xrightarrow{n\to\infty} 0.$$

Combining the arguments above yields (7.12). The consequence follows from $\int_{\rho_n(t)}^t |\Psi_f(g)(s)|\mathrm{d}s \to 0$ as $n \to \infty$. □

We apply Proposition 7.4 in the next three lemmas, to prove convergence of the drift part, the modified diffusion part, and the jump part of the semimartingale characteristics as aforementioned in Remark 7.3(1).

Lemma 7.5. *For $\mathfrak{b}^{\mathcal{X}}$ in Lemma 7.1 and any $t \in [0,\infty)$,*

$$I_{(7.15)} := \sup_{0\le s\le t} \left| \sum_{i=1}^{\sigma_n(s)} \mathbb{E}[\mathfrak{h}(\Delta_i^n \mathcal{X}^n)] - \mathfrak{b}_{s\wedge T}^{\mathcal{X}} \right| \xrightarrow{n\to\infty} 0. \tag{7.15}$$

Proof. It is sufficient to verify the convergence for any k-th coordinate, $k = 1,\dots,m$ and $t \in [0,T]$. Observe that

$$\mathfrak{b}_t^{\mathcal{X},(k)} = \int_0^t \Psi_f(\mathfrak{h}^{(k)})(s)\mathrm{d}s$$

for $\Psi_f(\mathfrak{h}^{(k)})$ associated with $\mathfrak{h}^{(k)}$ introduced in Proposition 7.4. Then we get

$$\begin{aligned}
&\sup_{0\le s\le t} \left| \sum_{i=1}^{\sigma_n(s)} \mathbb{E}[\mathfrak{h}^{(k)}(\Delta_i^n \mathcal{X}^n)] - \mathfrak{b}_s^{\mathcal{X},(k)} \right| \\
&\le \sup_{0\le s\le t} \left| \sum_{i=1}^{\sigma_n(s)} \mathbb{E}[\mathfrak{h}^{(k)}(\Delta_i^n \mathcal{X}^n)] - \sum_{i=1}^{\sigma_n(s)} \int_{t_{i-1}^n}^{t_i^n} \Psi_f(\mathfrak{h}^{(k)})(r)\mathrm{d}r \right| + \sup_{0\le s\le t} \left| \int_{\rho_n(s)}^s \Psi_f(\mathfrak{h}^{(k)})(r)\mathrm{d}r \right| \\
&\le \sum_{i=1}^n \left| \mathbb{E}[\mathfrak{h}^{(k)}(\Delta_i^n \mathcal{X}^n)] - \int_{t_{i-1}^n}^{t_i^n} \Psi_f(\mathfrak{h}^{(k)})(r)\mathrm{d}r \right| + \max_{1\le i\le n} \int_{t_{i-1}^n}^{t_i^n} |\Psi_f(\mathfrak{h}^{(k)})(r)|\mathrm{d}r.
\end{aligned}$$

The first term on the right-hand side above converges to 0 by applying Proposition 7.4 for $\mathfrak{h}^{(k)} \in C_b^2(\mathbb{R}^m)$. For the second term, since $t \mapsto \int_0^t |\Psi_f(\mathfrak{h}^{(k)})(r)|\mathrm{d}r$ is uniformly continuous on $[0,T]$ and $\max_{1\le i\le n} |t_i^n - t_{i-1}^n| \to 0$, it implies that

$$\max_{1\le i\le n} \int_{t_{i-1}^n}^{t_i^n} |\Psi_f(\mathfrak{h}^{(k)})(r)|\mathrm{d}r \xrightarrow{n\to\infty} 0.$$

Therefore, $I_{(7.15)} \to 0$ as $n \to \infty$. □

Lemma 7.6. *For $C^{\mathcal{X}}$ given in Lemma 7.1, for any $t \in [0,\infty)$ and $k, k' = 1,\dots,m$, one has*

$$I_{(7.16)} := \sum_{i=1}^{\sigma_n(t)} \mathbb{E}[\mathfrak{h}^{(k)}(\Delta_i^n \mathcal{X}^n)] \mathbb{E}[\mathfrak{h}^{(k')}(\Delta_i^n \mathcal{X}^n)] \xrightarrow{n\to\infty} 0, \tag{7.16}$$

$$I_{(7.17)} := \sum_{i=1}^{\sigma_n(t)} \mathbb{E}[(\mathfrak{h}^{(k)}\mathfrak{h}^{(k')})(\Delta_i^n \mathcal{X}^n)] \xrightarrow{n\to\infty} C_{t\wedge T}^{\mathcal{X},(k,k')} + \int_0^{t\wedge T} \int_{\mathbb{R}_0^m} (\mathfrak{h}^{(k)}\mathfrak{h}^{(k')})(y) \nu^{\mathcal{X}}(\mathrm{d}s,\mathrm{d}y). \tag{7.17}$$

Proof. It suffices to show the convergences for $t \in [0,T]$. For $I_{(7.16)}$, we first express

$$I_{(7.16)} = \sum_{i=1}^{\sigma_n(t)} \left(\mathbb{E}[\mathfrak{h}^{(k)}(\Delta_i^n \mathcal{X}^n)] - \int_{t_{i-1}^n}^{t_i^n} \Psi_f(\mathfrak{h}^{(k)})(s)\mathrm{d}s \right) \mathbb{E}[\mathfrak{h}^{(k')}(\Delta_i^n \mathcal{X}^n)]$$
$$+ \sum_{i=1}^{\sigma_n(t)} \left(\mathbb{E}[\mathfrak{h}^{(k')}(\Delta_i^n \mathcal{X}^n)] - \int_{t_{i-1}^n}^{t_i^n} \Psi_f(\mathfrak{h}^{(k')})(s)\mathrm{d}s \right) \int_{t_{i-1}^n}^{t_i^n} \Psi_f(\mathfrak{h}^{(k)})(s)\mathrm{d}s + \sum_{i=1}^{\sigma_n(t)} \left(\int_{t_{i-1}^n}^{t_i^n} \Psi_f(\mathfrak{h}^{(k)})(s)\mathrm{d}s \right) \left(\int_{t_{i-1}^n}^{t_i^n} \Psi_f(\mathfrak{h}^{(k')})(s)\mathrm{d}s \right).$$

Hence, the triangle inequality yields

$$|I_{(7.16)}| \le \|\mathfrak{h}^{(k')}\|_{B_b(\mathbb{R}^m)} \sum_{i=1}^{n} \left| \mathbb{E}[\mathfrak{h}^{(k)}(\Delta_i^n \mathcal{X}^n)] - \int_{t_{i-1}^n}^{t_i^n} \Psi_f(\mathfrak{h}^{(k)})(s)\mathrm{d}s \right|$$
$$+ \left(\int_0^T |\Psi_f(\mathfrak{h}^{(k)})(s)|\mathrm{d}s \right) \sum_{i=1}^{n} \left| \mathbb{E}[\mathfrak{h}^{(k')}(\Delta_i^n \mathcal{X}^n)] - \int_{t_{i-1}^n}^{t_i^n} \Psi_f(\mathfrak{h}^{(k')})(s)\mathrm{d}s \right| + \left(\int_0^T |\Psi_f(\mathfrak{h}^{(k)})(s)|\mathrm{d}s \right) \max_{1\le i\le n} \int_{t_{i-1}^n}^{t_i^n} |\Psi_f(\mathfrak{h}^{(k')})(s)|\mathrm{d}s.$$

Applying Proposition 7.4 for $\mathfrak{h}^{(k)}, \mathfrak{h}^{(k')} \in C_b^2(\mathbb{R}^m)$, we obtain that the sums $\sum_{i=1}^n$ in the first two terms on the right-hand side converge to 0 as $n \to \infty$. Since $\max_{1\le i\le n} \int_{t_{i-1}^n}^{t_i^n} |\Psi_f(\mathfrak{h}^{(k')})(s)|\mathrm{d}s \to 0$, we derive $I_{(7.16)} \to 0$ as desired.

For $I_{(7.17)}$, since $\mathfrak{h}^{(k)}\mathfrak{h}^{(k')} \in C_b^2(\mathbb{R}^m)$ and $\mathfrak{h}^{(k)}\mathfrak{h}^{(k')}(z) = z^{(k)}z^{(k')}$ around 0, the function $\Psi_f(\mathfrak{h}^{(k)}\mathfrak{h}^{(k')})$ given in (7.13) can be explicitly written as

$$\Psi_f(\mathfrak{h}^{(k)}\mathfrak{h}^{(k')})(s) = \sum_{l=1}^{p} \int_{[0,1]^d} (f_l^{(k)} f_l^{(k')})(s,u)\mathrm{d}u$$
$$+ \int_{\{0<|z|\le R\}\times[0,1]^d} (\mathfrak{h}^{(k)}\mathfrak{h}^{(k')})(f_{p+1}(s,z,u)|z|)\nu_s(\mathrm{d}z)\mathrm{d}u + \int_{\{|z|>R\}\times[0,1]^d} (\mathfrak{h}^{(k)}\mathfrak{h}^{(k')})(f_{p+2}(s,z,u))\nu_s(\mathrm{d}z)\mathrm{d}u$$

so that

$$\int_0^t \Psi_f(\mathfrak{h}^{(k)}\mathfrak{h}^{(k')})(s)\mathrm{d}s = C_t^{\mathcal{X},(k,k')} + \int_0^t \int_{\mathbb{R}_0^m} (\mathfrak{h}^{(k)}\mathfrak{h}^{(k')})(y)\nu^{\mathcal{X}}(\mathrm{d}s,\mathrm{d}y)$$

where we apply Remark 7.2 for the $\nu^{\mathcal{X}}$-integrable function $\mathfrak{h}^{(k)}\mathfrak{h}^{(k')}\mathbb{1}_{[0,T]}$. Hence, (7.17) follows directly from the consequence in Proposition 7.4. □

To investigate the jump part of the limiting process, we recall from [15, p.395] the family $C_2(\mathbb{R}^m)$ of bounded and continuous functions $g : \mathbb{R}^m \to \mathbb{R}$ with $g(0) = 0$ around 0.

Lemma 7.7. *For $\nu^{\mathcal{X}}$ in Lemma 7.1 and for any $g \in C_2(\mathbb{R}^m)$, $t \in [0,\infty)$, one has*

$$I_{(7.18)}^g := \left| \sum_{i=1}^{\sigma_n(t)} \mathbb{E}[g(\Delta_i^n \mathcal{X}^n)] - \int_0^{t\wedge T} \int_{\mathbb{R}_0^m} g(y)\nu^{\mathcal{X}}(\mathrm{d}s,\mathrm{d}y) \right| \xrightarrow{n\to\infty} 0. \tag{7.18}$$

Proof. We only need to prove for $t \in [0,T]$.
Step 1. Recall $\Delta_i^n \mathcal{X}^n$ from (7.2). We show that for any $\kappa > 0$,

$$\sum_{i=1}^{n} \mathbb{P}(\{|\Delta_i^n \mathcal{X}^n| \ge \kappa\}) \le \frac{9}{\kappa^2} \left(pT \max_{1\le l\le p} \|f_l\|_{B_b(\mathbf{U};\mathbb{R}^m)}^2 + \|f_{p+1}\|_{B_b(\mathbf{V};\mathbb{R}^m)}^2 \int_0^T \int_{0<|z|\le R} |z|^2 \nu_s(\mathrm{d}z)\mathrm{d}s \right)$$
$$+ \frac{3T}{\kappa} \|f_0\|_{B_b(\mathbf{U};\mathbb{R}^m)} + \frac{3}{\kappa} \|f_{p+2}\|_{B_b(\mathbf{V};\mathbb{R}^m)} \int_0^T \int_{|z|>R} \nu_s(\mathrm{d}z)\mathrm{d}s. \tag{7.19}$$

Indeed, by the triangle inequality we get

$$\sum_{i=1}^{n} \mathbb{P}(\{|\Delta_i^n \mathcal{X}^n| \ge \kappa\})$$
$$\le \sum_{i=1}^{n} \mathbb{P}\left(\left\{ \left| \int_{t_{i-1}^n}^{t_i^n} f_0(s,\xi_i^n)\mathrm{d}s \right| \ge \frac{\kappa}{3} \right\} \right)$$
$$+ \sum_{i=1}^{n} \mathbb{P}\left(\left\{ \left| \sum_{l=1}^{p} \int_{t_{i-1}^n}^{t_i^n} f_l(s,\xi_i^n)\mathrm{d}B_s^{(l)} + \int_{t_{i-1}^n}^{t_i^n} \int_{0<|z|\le R} f_{p+1}(s,z,\xi_i^n)|z|\tilde{N}(\mathrm{d}s,\mathrm{d}z) \right| \ge \frac{\kappa}{3} \right\} \right)$$
$$+ \sum_{i=1}^{n} \mathbb{P}\left(\left\{ \left| \int_{t_{i-1}^n}^{t_i^n} \int_{|z|>R} f_{p+2}(s,z,\xi_i^n)N(\mathrm{d}s,\mathrm{d}z) \right| \ge \frac{\kappa}{3} \right\} \right)$$
$$=: I_{(7.20)} + II_{(7.20)} + III_{(7.20)}. \tag{7.20}$$

For the first term, Markov's inequality yields

$$I_{(7.20)} \leq \frac{3}{\kappa}\sum_{i=1}^{n}\mathbb{E}\left[\left|\int_{t_{i-1}^n}^{t_i^n} f_0(s,\xi_i^n)\mathrm{d}s\right|\right] \leq \frac{3T}{\kappa}\|f_0\|_{B_b(\mathbf{U};\mathbb{R}^m)}.$$

For the second term, applying the Markov's inequality and Itô's isometry we get

$$\begin{aligned} II_{(7.20)} &\leq \frac{9}{\kappa^2}\sum_{i=1}^{n}\mathbb{E}\left[\left|\sum_{l=1}^{p}\int_{t_{i-1}^n}^{t_i^n} f_l(s,\xi_i^n)\mathrm{d}B_s^{(l)} + \int_{t_{i-1}^n}^{t_i^n}\int_{0<|z|\leq R} f_{p+1}(s,z,\xi_i^n)|z|\tilde{N}(\mathrm{d}s,\mathrm{d}z)\right|^2\right] \\ &= \frac{9}{\kappa^2}\sum_{i=1}^{n}\mathbb{E}\left[\sum_{l=1}^{p}\int_{t_{i-1}^n}^{t_i^n}|f_l(s,\xi_i^n)|^2\mathrm{d}s + \int_{t_{i-1}^n}^{t_i^n}\int_{0<|z|\leq R}|f_{p+1}(s,z,\xi_i^n)|^2|z|^2\nu_s(\mathrm{d}z)\mathrm{d}s\right] \\ &\leq \frac{9}{\kappa^2}\left(pT\max_{1\leq l\leq p}\|f_l\|^2_{B_b(\mathbf{U};\mathbb{R}^m)} + \|f_{p+1}\|^2_{B_b(\mathbf{V};\mathbb{R}^m)}\int_0^T\int_{0<|z|\leq R}|z|^2\nu_s(\mathrm{d}z)\mathrm{d}s\right). \end{aligned}$$

For the third term, using Markov's inequality we obtain

$$\begin{aligned} III_{(7.20)} &\leq \frac{3}{\kappa}\sum_{i=1}^{n}\mathbb{E}\left[\left|\int_{t_{i-1}^n}^{t_i^n}\int_{|z|>R} f_{p+2}(s,z,\xi_i^n)N(\mathrm{d}s,\mathrm{d}z)\right|\right] \\ &\leq \frac{3}{\kappa}\|f_{p+2}\|_{B_b(\mathbf{V};\mathbb{R}^m)}\sum_{i=1}^{n}\mathbb{E}\left[\int_{t_{i-1}^n}^{t_i^n}\int_{|z|>R}N(\mathrm{d}s,\mathrm{d}z)\right] \\ &= \frac{3}{\kappa}\|f_{p+2}\|_{B_b(\mathbf{V};\mathbb{R}^m)}\int_0^T\int_{|z|>R}\nu_s(\mathrm{d}z)\mathrm{d}s. \end{aligned}$$

Hence, combining those four estimates yields (7.19).

Step 2. Since $g \in C_2(\mathbb{R}^m)$, there is an $r_g > 0$ such that $g = 0$ on the open ball $B_m(r_g)$. Then, we use Remark 7.2 to obtain that

$$\int_0^T\int_{|y|\geq r_g}|g(y)|\nu^{\mathcal{X}}(\mathrm{d}s,\mathrm{d}y) \leq \|g\|_{B_b(\mathbb{R}^m)}\int_0^T\int_{|y|\geq r_g}\nu^{\mathcal{X}}(\mathrm{d}s,\mathrm{d}y) < \infty.$$

Hence, the integral on the right-hand side of (7.18) finitely exists.

We now only prove (7.18) in the case $0 \leq R < \infty$ as the case $R = \infty$ is analogous. Let $\varepsilon > 0$ and $\theta > r_g \vee R^2$. Since g is continuous and bounded, there exists a continuous function g_θ with compact support such that

$$\|g_\theta\|_{B_b(\mathbb{R}^m)} \leq \|g\|_{B_b(\mathbb{R}^m)} \quad \text{and} \quad g_\theta = g \ \text{ on } B_m(\theta).$$

Moreover, by convolution approximation, we can find a $g_{\varepsilon,\theta} \in C_2(\mathbb{R}^m) \cap C_c^2(\mathbb{R}^m)$ such that

$$g_{\varepsilon,\theta} = g_\theta = 0 \ \text{ on } B_m(r_g/2), \quad \text{and} \quad \|g_{\varepsilon,\theta} - g_\theta\|_{B_b(\mathbb{R}^m)} \leq \varepsilon.$$

It follows from the linearity and the triangle inequality that

$$I^g_{(7.18)} \leq I^{g-g_\theta}_{(7.18)} + I^{g_\theta - g_{\varepsilon,\theta}}_{(7.18)} + I^{g_{\varepsilon,\theta}}_{(7.18)}. \tag{7.21}$$

Since $g_{\varepsilon,\theta} \in C_c^2(\mathbb{R}^m)$ takes value 0 in a neighborhood of 0, Remark 7.2 implies

$$\int_0^t \Psi_f(g_{\varepsilon,\theta})(s)\mathrm{d}s = \int_0^t\int_{\mathbb{R}_0^q\times[0,1]^d}[g(f_{p+1}(s,z,u)|z|)\mathbb{1}_{\{0<|z|\leq R\}} + g(f_{p+2}(s,z,u))\mathbb{1}_{\{0<|z|\leq R\}}]\nu_s(\mathrm{d}z)\mathrm{d}u\mathrm{d}s = \int_0^t\int_{\mathbb{R}_0^m} g(y)\nu^{\mathcal{X}}(\mathrm{d}s,\mathrm{d}y)$$

so that the consequence in Proposition 7.4 verifies

$$I^{g_{\varepsilon,\theta}}_{(7.18)} \xrightarrow{n\to\infty} 0.$$

For $I^{g-g_\theta}_{(7.18)}$, one has

$$I^{g-g_\theta}_{(7.18)} \leq \sum_{i=1}^{n}\mathbb{E}[|(g-g_\theta)(\Delta_i^n\mathcal{X}^n)|] + \int_0^T\int_{\mathbb{R}_0^m}|g(y)-g_\theta(y)|\nu^{\mathcal{X}}(\mathrm{d}s,\mathrm{d}y) \leq \|g-g_\theta\|_{B_b(\mathbb{R}^m)}\left(\sum_{i=1}^{n}\mathbb{P}(\{|\Delta_i^n\mathcal{X}^n| \geq \theta\}) + \int_0^T\int_{|y|\geq\theta}\nu^{\mathcal{X}}(\mathrm{d}s,\mathrm{d}y)\right).$$

We let $\kappa = \theta$ in (7.19) to find that $\sum_{i=1}^n \mathbb{P}(\{|\Delta_i^n\mathcal{X}^n| \geq \theta\}) \to 0$ uniformly in n as $\theta \to \infty$. Moreover, it follows from (7.11) that $\int_0^T\int_{|y|\geq\theta}\nu^{\mathcal{X}}(\mathrm{d}s,\mathrm{d}y) \to 0$ as $\theta \to \infty$ which thus yields

$$I^{g-g_\theta}_{(7.18)} \to 0 \quad \text{uniformly in } n \text{ as } \theta \to \infty.$$

For $I^{g_\theta - g_{\varepsilon,\theta}}_{(7.18)}$, one has

$$I^{g_\theta-g_{\varepsilon,\theta}}_{(7.18)} \leq \sum_{i=1}^{n}\mathbb{E}[|(g_\theta-g_{\varepsilon,\theta})(\Delta_i^n\mathcal{X}^n)|] + \int_0^T\int_{\mathbb{R}_0^m}|g_\theta(y)-g_{\varepsilon,\theta}(y)|\nu^{\mathcal{X}}(\mathrm{d}s,\mathrm{d}y)$$

$$\leq \|g_\theta - g_{\varepsilon,\theta}\|_{B_b(\mathbb{R}^m)} \left(\sum_{i=1}^n \mathbb{P}(\{|\Delta_i^n \mathcal{X}^n| \geq r_g/2\}) + \int_0^T \int_{|y|\geq r_g/2} \nu^{\mathcal{X}}(\mathrm{d}s, \mathrm{d}y) \right)$$

$$\leq \varepsilon \left(\sum_{i=1}^n \mathbb{P}(\{|\Delta_i^n \mathcal{X}^n| \geq r_g/2\}) + \int_0^T \int_{|y|\geq r_g/2} \nu^{\mathcal{X}}(\mathrm{d}s, \mathrm{d}y) \right).$$

Choosing $\kappa = r_g/2$ in (7.19) and using (7.11) we obtain

$$I^{g_\theta - g_{\varepsilon,\theta}}_{(7.18)} \to 0 \quad \text{uniformly over } n \text{ as } \varepsilon \to 0.$$

Since θ can be chosen arbitrarily large and $\varepsilon > 0$ arbitrarily small, we derive from (7.21) the desired conclusion. □

We can now finalize the *proof of assertion* (7.4). Combining Lemmas 7.5, 7.6, 7.7 with Lemma 7.1, together with applying [15, Theorem VII.2.29] (see also [2] for a recap of this result), we get that $(\sum_{i=1}^{\sigma_n(t)} \Delta_i^n \mathcal{X}^n)_{t\in[0,\infty)} \to (\mathcal{X}_{t\wedge T})_{t\in[0,\infty)}$ as $n \to \infty$ weakly in the Skorokhod topology on the space $\mathbb{D}_\infty(\mathbb{R}^m)$ of càdlàg functions $F : [0,\infty) \to \mathbb{R}^m$ (see [4,15] for $\mathbb{D}_\infty(\mathbb{R}^m)$). Since $\mathcal{X}$ has no fixed time of discontinuity, we use [4, Theorem 16.7] to infer that $\mathcal{X}^n_{\rho_n} = (\sum_{i=1}^{\sigma_n(t)} \Delta_i^n \mathcal{X}^n)_{t\in[0,T]} \xrightarrow{\mathscr{D}_T} (\mathcal{X}_t)_{t\in[0,T]}$ as $n \to \infty$. □

8. Proofs of Proposition 4.2 and Proposition 5.3

8.1. Proof of Proposition 4.2

As the grid-sampling SDE (4.1) is an SDE driven by a Brownian motion and a Poisson random measure, the existence and uniqueness of strong solution to (4.1) can be achieved by a routine argument combined with the interlacing technique to handle the large jump part. We refer, for example, to [14, Theorem IV.9.1] for SDEs driven by a finite dimensional Brownian motion and homogeneous Poisson random measure, and to [5, Section 4.2] for the infinite dimensional case.

We now deal with the random measures $M^\Pi_{B^{(l)}}$ and M^Π_J. By the definition of $M^\Pi_{B^{(l)}}$, for any $A \in \mathcal{B}([0,1]^d)$ one has $M^\Pi_{B^{(l)}}(0, A) = 0$, and for $t \in (0, T]$, we can write

$$M^\Pi_{B^{(l)}}(t, A) = \int_0^t \mathbb{1}_A \left(\sum_{i=1}^n \mathbb{1}_{(t_{i-1},t_i]}(s) \xi^\Pi_{t_i} \right) \mathrm{d}B^{(l)}_s.$$

Then, due to [23, Proposition II-1], $M^\Pi_{B^{(l)}}$ is an orthogonal $(\mathbb{F}^\Pi, \mathbb{P})$-martingale measure on $[0,T] \times \mathcal{B}([0,1]^d)$ with intensity $\mu^\Pi_{B^{(l)}}(\mathrm{d}s, \mathrm{d}x) = \delta_{\sum_{i=1}^n \mathbb{1}_{(t_{i-1},t_i]}(s)\xi^\Pi_{t_i}}(\mathrm{d}x)\mathrm{d}s$. It is clear that $\mu^\Pi_{B^{(l)}} = M^\Pi_D$.

By the definition of M^Π_J, for any $\mathbb{F}^\Pi$-predictable $Y \geq 0$ and for $(Y \cdot M^\Pi_J) = ((Y \cdot M^\Pi_J)_t)_{t\in[0,T]}$ defined by $(Y \cdot M^\Pi_J)_t := \int_{(0,t]\times\mathbb{R}^q_0\times[0,1]^d} Y_s(z,u) M^\Pi_J(\mathrm{d}s, \mathrm{d}z, \mathrm{d}u)$ one has, a.s.,

$$(Y \cdot M^\Pi_J)_T = \sum_{i=1}^n \sum_{s\in(t_{i-1},t_i]} \mathbb{1}_{\{\Delta L_s \neq 0\}} Y_s(\Delta L_s, \xi^\Pi_{t_i}) = \sum_{i=1}^n \int_{(0,T]\times\mathbb{R}^q_0} \mathbb{1}_{(t_{i-1},t_i]}(s) Y_s(z, \xi^\Pi_{t_i}) N(\mathrm{d}s, \mathrm{d}z).$$

As $\nu_s(\mathrm{d}z)\mathrm{d}s$ is the $(\mathbb{F}^\Pi, \mathbb{P})$-predictable compensator of $N(\mathrm{d}s, \mathrm{d}z)$ (see [15, Proposition II.1.21]), we get

$$\begin{aligned}
\mathbb{E}[(Y \cdot M^\Pi_J)_T] &= \mathbb{E}\left[\sum_{i=1}^n \int_{(0,T]\times\mathbb{R}^q_0} \mathbb{1}_{(t_{i-1},t_i]}(s) Y_s(z, \xi^\Pi_{t_i}) \nu_s(\mathrm{d}z)\mathrm{d}s \right] \\
&= \mathbb{E}\left[\sum_{i=1}^n \int_{(0,T]\times\mathbb{R}^q_0\times[0,1]^d} Y_s(z,u) \mathbb{1}_{(t_{i-1},t_i]}(s) \delta_{\xi^\Pi_{t_i}}(\mathrm{d}u) \nu_s(\mathrm{d}z)\mathrm{d}s \right] \\
&= \mathbb{E}\left[\int_{(0,T]\times\mathbb{R}^q_0\times[0,1]^d} Y_s(z,u) \mu^\Pi_J(\mathrm{d}s, \mathrm{d}z, \mathrm{d}u) \right] \\
&= \mathbb{E}[(Y \cdot \mu^\Pi_J)_T].
\end{aligned}$$

We note that $(Y \cdot \mu^\Pi_J)$ is $\mathbb{F}^\Pi$-predictable as the pointwise limit of the continuous and $\mathbb{F}^\Pi$-adapted processes $(Y^n \cdot \mu^\Pi_J)$ as $n \to \infty$ where $Y^n := (Y \wedge n)\mathbb{1}_{\{|z|>1/n\}}$. Hence, μ^Π_J is an $\mathbb{F}^\Pi$-predictable random measure in the sense of [15, Definition II.1.6(a)]. By [15, Theorem II.1.8(i)], we conclude that μ^Π_J is the $(\mathbb{F}^\Pi, \mathbb{P})$-predictable compensator of M^Π_J.

To show that the strong solution to the grid-sampling SDE (4.1) also solves the SDE (4.3), we need the following lemma whose standard proof via approximation is provided in [2].

Lemma 8.1.

(1) For any $\mathcal{F} \otimes \mathcal{B}([0,T]) \otimes \mathcal{B}([0,1]^d)/\mathcal{B}(\mathbb{R})$-measurable random field $Y : \Omega \times [0,T] \times [0,1]^d \to \mathbb{R}$ satisfying $\int_0^T |Y_s(\xi^\Pi_s)|\mathrm{d}s < \infty$ a.s, one has

$$\int_{(0,T]\times[0,1]^d} Y_s(u) M^\Pi_D(\mathrm{d}s, \mathrm{d}u) = \int_0^T Y_s(\xi^\Pi_s)\mathrm{d}s \quad \mathbb{P}\text{-a.s.}$$

(2) If a $\mathcal{P}_{\mathbb{F}^\Pi} \otimes \mathcal{B}([0,1]^d)/\mathcal{B}(\mathbb{R})$-measurable random field $Y : \Omega \times [0,T] \times [0,1]^d \to \mathbb{R}$ satisfies $\int_0^T |Y_s(\xi^\Pi_s)|^2\mathrm{d}s < \infty$ a.s., then for any $l = 1, \ldots, p$ one has

$$\int_{(0,T]\times[0,1]^d} Y_s(u) M^\Pi_{B^{(l)}}(\mathrm{d}s, \mathrm{d}u) = \int_0^T Y_s(\xi^\Pi_s)\mathrm{d}B^{(l)}_s \quad \mathbb{P}\text{-a.s.}$$

(3) Suppose that $Y : \Omega \times [0,T] \times \mathbb{R}_0^q \times [0,1]^d \to \mathbb{R}$ *is* $\mathcal{P}_{\mathbb{F}^\Pi} \otimes \mathcal{B}(\mathbb{R}_0^q) \otimes \mathcal{B}([0,1]^d)/\mathcal{B}(\mathbb{R})$*-measurable. If* $\int_{(0,T]\times\mathbb{R}_0^q} |Y_s(z,\xi_s^\Pi)| N(\mathrm{d}s,\mathrm{d}z) < \infty$ *a.s., then*

$$\int_{(0,T]\times\mathbb{R}_0^q\times[0,1]^d} Y_s(z,u) M_J^\Pi(\mathrm{d}s,\mathrm{d}z,\mathrm{d}u) = \int_{(0,T]\times\mathbb{R}_0^q} Y_s(z,\xi_s^\Pi) N(\mathrm{d}s,\mathrm{d}z) \quad \mathbb{P}\text{-a.s.}$$

Moreover, if $\int_0^T \int_{\mathbb{R}_0^q} |Y_s(z,\xi_s^\Pi)|^2 \nu_s(\mathrm{d}z)\mathrm{d}s < \infty$ *a.s., then*

$$\int_{(0,T]\times\mathbb{R}_0^q\times[0,1]^d} Y_s(z,u) \tilde{M}_J^\Pi(\mathrm{d}s,\mathrm{d}z,\mathrm{d}u) = \int_{(0,T]\times\mathbb{R}_0^q} Y_s(z,\xi_s^\Pi) \tilde{N}(\mathrm{d}s,\mathrm{d}z) \quad \mathbb{P}\text{-a.s.}$$

8.2. Proof of Proposition 5.3

For the well-posedness, we aim to combine Proposition A.3 with the interlacing technique used in the proof of [14, Theorem IV.9.1]. To do that, following the notation of Proposition A.3, we take $E = \mathbb{R}^q \times [0,1]^d$, which is equipped with the Euclidean norm, and let $\ell = p+1$. For $t \in [0,T]$ and $A \in \mathcal{B}(\mathbb{R}^q \times [0,1]^d)$, we define

$$\mathcal{M}^{(j)}(t,A) := \begin{cases} \int_{(0,t]\times\{u\in[0,1]^d\,:\,(0,u)\in A\}} M_{B^{(j)}}(\mathrm{d}s,\mathrm{d}u) & \text{if } j = 1,\dots,p, \\ \int_{(0,t]\times A} \mathbb{1}_{\{0<|z|\le\mathfrak{r}\}} |z| \tilde{M}_J(\mathrm{d}s,\mathrm{d}z,\mathrm{d}u) & \text{if } j = p+1. \end{cases}$$

It is easy to check that $\mathcal{M}^{(j)}$ is an orthogonal $(\mathbb{F},\mathbb{P})$-martingale measures on $[0,T]\times\mathcal{B}(\mathbb{R}^q\times[0,1]^d)$ with the (deterministic) intensity

$$\mu^{(j)}(\mathrm{d}s,\mathrm{d}z,\mathrm{d}u) := \mu_s^{(j)}(\mathrm{d}z,\mathrm{d}u)\mathrm{d}s := \begin{cases} \delta_0(\mathrm{d}z)\mathrm{d}u\mathrm{d}s & \text{if } j = 1,\dots,p, \\ \mathbb{1}_{\{0<|z|\le\mathfrak{r}\}} |z|^2 \nu_s(\mathrm{d}z)\mathrm{d}u\mathrm{d}s & \text{if } j = p+1. \end{cases}$$

For the $\mathbb{R}^m$-valued function $\hat{b}_{\mathbf{h}}$ and $\mathbb{R}^{m\times(p+1)}$-valued function $\hat{a}_{\mathbf{h}}$ defined by

$$\hat{b}_{\mathbf{h}}(s,x) := \int_{[0,1]^d} b_{\mathbf{h}}(s,x,u)\mathrm{d}u,$$

$$\hat{a}_{\mathbf{h}}^{(i,j)}(s,x,z,u) := \begin{cases} \mathbb{1}_{\{z=0\}} a_{\mathbf{h}}^{(i,j)}(s,x,u) & \text{if } 1 \le i \le m,\ 1 \le j \le p, \\ \mathbb{1}_{\{0<|z|\le\mathfrak{r}\}} |z|^{-1} \gamma_{\mathbf{h}}^{(i)}(s,x,z,u) & \text{if } 1 \le i \le m,\ j = p+1, \end{cases}$$

and for $\mathcal{M} = (\mathcal{M}^{(1)},\dots,\mathcal{M}^{(p+1)})^\top$, the SDE (5.1) can be re-written (we omit the superscript $\mathbf{h}$ of $X^{\mathbf{h}}$) as

$$X_t = x_0 + \int_0^t \hat{b}_{\mathbf{h}}(s,X_{s-})\mathrm{d}s + \int_{(0,t]\times\mathbb{R}^q\times[0,1]^d} \hat{a}_{\mathbf{h}}(s,X_{s-},z,u)\mathcal{M}(\mathrm{d}s,\mathrm{d}z,\mathrm{d}u) + \int_{(0,t]\times\{|z|>\mathfrak{r}\}\times[0,1]^d} \gamma_{\mathbf{h}}(s,X_{s-},z,u) M_J(\mathrm{d}s,\mathrm{d}z,\mathrm{d}u). \tag{8.1}$$

It follows from the condition (2.2) that

$$\mathbb{E}\left[\int_{(0,T]\times\{|z|>\mathfrak{r}\}\times[0,1]^d} M_J(\mathrm{d}s,\mathrm{d}z,\mathrm{d}u)\right] = \int_{(0,T]\times\{|z|>\mathfrak{r}\}\times[0,1]^d} \nu_s(\mathrm{d}z)\mathrm{d}u\mathrm{d}s < \infty,$$

there exists a sequence of $\mathbb{F}$-stopping times $0 < \tau_1 < \cdots$ with values on $[0,T]$ capturing the jump times of the Poisson point process $[0,T] \ni t \mapsto M_J((0,t]\times\{|z|>\mathfrak{r}\}\times[0,1]^d)$, where we set $\tau_0 := 0$ and $\tau_j := T$ if there is no jumps on $(\tau_{j-1},T]$. Then, the large jump part in (8.1) is re-written as

$$\int_{(0,t]\times\{|z|>\mathfrak{r}\}\times[0,1]^d} \gamma_{\mathbf{h}}(s,X_{s-},z,u) M_J(\mathrm{d}s,\mathrm{d}z,\mathrm{d}u) = \sum_{\tau_j\le t} \gamma_{\mathbf{h}}(\tau_j, X_{\tau_j-}, \Delta L_{\tau_j}^{\mathfrak{r}}), \quad t \in [0,T] \quad \mathbb{P}\text{-a.s.},$$

where

$$L_t^{\mathfrak{r}} := \int_{(0,t]\times\{|z|>\mathfrak{r}\}\times[0,1]^d} (z,u)^\top M_J(\mathrm{d}s,\mathrm{d}z,\mathrm{d}u), \quad t \in [0,T].$$

Step 1. We first construct a solution to (8.1) on $[0,\tau_1]$. Consider the following SDE on $[0,T]$,

$$Y_t = c_0 + \int_0^t \hat{b}_{\mathbf{h}}(s,Y_{s-})\mathrm{d}s + \int_{(0,t]\times\mathbb{R}^q\times[0,1]^d} \hat{a}_{\mathbf{h}}(s,Y_{s-},z,u)\mathcal{M}(\mathrm{d}s,\mathrm{d}z,\mathrm{d}u). \tag{8.2}$$

In Proposition A.3, we let $\eta = x_0$, $\beta(\omega,s,y) := \hat{b}_{\mathbf{h}}(s,y)$, $\alpha(\omega,s,y,z,u) := \hat{a}_{\mathbf{h}}(s,y,z,u)$ and $M := \mathcal{M}$, and note that all assumptions there are fulfilled so that the SDE (8.2) with initial condition x_0 has a strong unique solution which is denoted by $Y^{\tau_0} = (Y_t^{\tau_0})_{t\in[0,T]}$. Define

$$X_t := \begin{cases} Y_t^{\tau_0} & \text{if } t \in [0,\tau_1) \\ Y_{\tau_1-}^{\tau_0} + \gamma_{\mathbf{h}}(\tau_1, Y_{\tau_1-}^{\tau_0}, \Delta L_{\tau_1}^{\mathfrak{r}}) & \text{if } t = \tau_1. \end{cases}$$

Then, $(X_t)_{t\in[0,\tau_1]}$ is a unique strong solution to (8.1) on $[0,\tau_1]$.

Step 2. Construction of a solution to (8.1) on any $[\tau_{j-1},\tau_j]$, $j \ge 2$. Consider the interval $[\tau_1,\tau_2]$. We now need to shift the entire dynamic of (8.2) by the $\mathbb{F}$-stopping time τ_1. Define the filtration $\mathbb{F}^{\tau_1} = (\mathcal{F}_t^{\tau_1})_{t\in[0,T]}$ with $\mathcal{F}_t^{\tau_1} := \mathcal{F}_{(\tau_1+t)\wedge T}$, which satisfies the usual conditions. For $j = 1,\dots,p+1$ and $(t,A) \in [0,T]\times\mathcal{B}(\mathbb{R}^q\times[0,1]^d)$, we set

$$\mathcal{M}_{\tau_1}^{(j)}(t,A) := \mathcal{M}^{(j)}((\tau_1+t)\wedge T, A) - \mathcal{M}^{(j)}(\tau_1,A).$$

According to Lemma A.2, $\mathcal{M}^{(j)}_{\tau_1}$ is an orthogonal $(\mathbb{F}^{\tau_1},\mathbb{P})$-martingale measure with ($\mathbb{F}^{\tau_1}$-predictable) intensity $\mu_{\mathcal{M}^{(j)}_{\tau_1}}$ given by

$$\mu_{\mathcal{M}^{(j)}_{\tau_1}}(\mathrm{d}s,\mathrm{d}z,\mathrm{d}u) = \mathbb{1}_{(0,T-\tau_1]}(s)\mu^{(j)}_{\tau_1+s}(\mathrm{d}z,\mathrm{d}u)\mathrm{d}s.$$

Consider the following SDE on $(\Omega,\mathcal{F},\mathbb{F}^{\tau_1},\mathbb{P})$ with initial condition X_{τ_1},

$$Y_t = X_{\tau_1} + \int_0^t \mathbb{1}_{(0,T-\tau_1]}(s)\,\hat{b}_{\mathbf{h}}(\tau_1+s,Y_{s-})\mathrm{d}s + \int_{(0,t]\times\mathbb{R}^q\times[0,1]^d} \mathbb{1}_{(0,T-\tau_1]}(s)\,\hat{a}_{\mathbf{h}}(\tau_1+s,Y_{s-},z,u)\mathcal{M}_{\tau_1}(\mathrm{d}s,\mathrm{d}z,\mathrm{d}u) \tag{8.3}$$

where $\mathcal{M}_{\tau_1} = (\mathcal{M}^{(1)}_{\tau_1},\dots,\mathcal{M}^{(p+1)}_{\tau_1})^\intercal$. In Proposition A.3, we let $\eta = X_{\tau_1}$, $M := \mathcal{M}_{\tau_1}$, and

$$\beta(\omega,s,y) := \mathbb{1}_{(0,T-\tau_1(\omega)]}(s)\,\hat{b}_{\mathbf{h}}(\tau_1(\omega)+s,y),$$
$$\alpha(\omega,s,y,z,u) := \mathbb{1}_{(0,T-\tau_1(\omega)]}(s)\,\hat{a}_{\mathbf{h}}(\tau_1(\omega)+s,y,z,u).$$

Then, α is $\mathcal{P}_{\mathbb{F}^{\tau_1}}\otimes\mathcal{B}(\mathbb{R}^m)\otimes\mathcal{B}(\mathbb{R}^q\times[0,1]^d)/\mathcal{B}(\mathbb{R}^{m\times(p+1)})$-measurable and β is $\mathcal{P}_{\mathbb{F}^{\tau_1}}\otimes\mathcal{B}(\mathbb{R}^m)/\mathcal{B}(\mathbb{R}^m)$-measurable. Moreover, for any (ω,s,y_1,y_2),

$$\begin{aligned}
&|\beta(\omega,s,y_1)-\beta(\omega,s,y_2)| = \mathbb{1}_{(0,T-\tau_1(\omega)]}(s)\,|\hat{b}_{\mathbf{h}}(\tau_1(\omega)+s,y_1)-\hat{b}_{\mathbf{h}}(\tau_1(\omega)+s,y_2)| \le K_{\mathrm{Lip}}|y_1-y_2|,\\
&\int_{\mathbb{R}^q\times[0,1]^d} |\alpha^{(i,j)}(\omega,s,y_1,z,u)-\alpha^{(i,j)}(\omega,s,y_2,z,u)|^2\mathbb{1}_{(0,T-\tau_1(\omega)]}(s)\mu^{(j)}_{\tau_1(\omega)+s}(\mathrm{d}z,\mathrm{d}u)\\
&= \int_{\mathbb{R}^q\times[0,1]^d} |\hat{a}^{(i,j)}_{\mathbf{h}}(\tau_1(\omega)+s,y_1,z,u)-\hat{a}^{(i,j)}_{\mathbf{h}}(\tau_1(\omega)+s,y_2,z,u)|^2\mathbb{1}_{(0,T-\tau_1(\omega)]}(s)\mu^{(j)}_{\tau_1(\omega)+s}(\mathrm{d}z,\mathrm{d}u)\\
&\le K^2_{\mathrm{Lip}}|y_1-y_2|^2,
\end{aligned}$$

and

$$\begin{aligned}
&\int_0^T \mathbb{1}_{(0,T-\tau_1(\omega)]}(s)\,|\hat{b}_{\mathbf{h}}(\tau_1(\omega)+s,0)|^2\mathrm{d}s = \int_0^{T-\tau_1(\omega)} |\hat{b}_{\mathbf{h}}(\tau_1(\omega)+s,0)|^2\mathrm{d}s \le \int_0^T |\hat{b}_{\mathbf{h}}(s,0)|^2\mathrm{d}s,\\
&\int_0^T\int_{\mathbb{R}^q\times[0,1]^d} |\hat{a}^{(i,j)}_{\mathbf{h}}(\tau_1(\omega)+s,0,z,u)|^2\mathbb{1}_{(0,T-\tau_1(\omega)]}(s)\mu^{(j)}_{\tau_1(\omega)+s}(\mathrm{d}z,\mathrm{d}u)\mathrm{d}s \le \int_0^T\int_{\mathbb{R}^q\times[0,1]^d} |\hat{a}^{(i,j)}_{\mathbf{h}}(s,0,z,u)|^2\mu^{(j)}_s(\mathrm{d}z,\mathrm{d}u)\mathrm{d}s.
\end{aligned}$$

It thus follows from Proposition A.3 that (8.3) admits a unique strong solution $Y^{\tau_1} = (Y^{\tau_1}_t)_{t\in[0,T]}$ on $(\Omega,\mathcal{F},\mathbb{F}^{\tau_1},\mathbb{P})$ with initial condition X_{τ_1}. Set $\tilde{Y}^{\tau_1}_t := \mathbb{1}_{[\tau_1,T]}(t)Y^{\tau_1}_{t-\tau_1}$. Then, by approximating τ_1 from the right by discrete $\mathbb{F}$-stopping times, we infer that $\tilde{Y}^{\tau_1}_t$ is $\mathcal{F}_t$-measurable and we derive from (8.3) and Lemma A.2 that, for $t\in[\tau_1,T]$,

$$\tilde{Y}^{\tau_1}_t = X_{\tau_1} + \int_{\tau_1}^t \hat{b}_{\mathbf{h}}(s,\tilde{Y}^{\tau_1}_{s-})\mathrm{d}s + \int_{(\tau_1,t]\times\mathbb{R}^q\times[0,1]^d} \mathbb{1}_{(\tau_1,T]}(s)\,\hat{a}_{\mathbf{h}}(s,\tilde{Y}^{\tau_1}_{s-},z,u)\mathcal{M}(\mathrm{d}s,\mathrm{d}z,\mathrm{d}u).$$

We define

$$X_t := \begin{cases} \tilde{Y}^{\tau_1}_t & \text{if } t\in[\tau_1,\tau_2)\\ \tilde{Y}^{\tau_1}_{\tau_2-} + \gamma_{\mathbf{h}}(\tau_2,\tilde{Y}^{\tau_1}_{\tau_2-},\Delta L^{\mathfrak{r}}_{\tau_2}) & \text{if } t=\tau_2. \end{cases}$$

Then, X solves (8.1) on $[\tau_1,\tau_2]$.

Iterating this procedure we obtain a strong solution to (8.1) on all intervals $[\tau_{j-1},\tau_j]$, $j\in\mathbb{N}$, and thus, on the entire $[0,T]$. The uniqueness of X on $[0,T]$ follows from its uniqueness on all $[\tau_{j-1},\tau_j]$, $j\in\mathbb{N}$.

Step 3. For the martingale problem, we first notice that $\mathcal{L}_{\mathbf{h}}f$ is finitely defined on $[0,T]\times\mathbb{R}^m$ due to the conditions imposed on the coefficients and the boundedness of the partial derivatives of f. Note that the continuous martingale part of $X^{\mathbf{h}}$ has the predictable quadratic variation

$$\int_0^\cdot \sum_{i,j=1}^m \left(\int_{[0,1]^d} A^{(i,j)}_{\mathbf{h}}(s,X^{\mathbf{h}}_{s-},u)\mathrm{d}u\right)\mathrm{d}s$$

by Proposition I-6(2) in [23], cp. the proof of Lemma 6.1 in [2] for more details. We may, thus, apply Itô's formula for $X^{\mathbf{h}}$ and $f\in C^2_c(\mathbb{R}^m)$ to get, a.s.,

$$\begin{aligned}
&f(X^{\mathbf{h}}_t)-f(x_0)\\
&= \int_0^t \sum_{i=1}^m \frac{\partial f}{\partial x_i}(X^{\mathbf{h}}_{s-})\left(\int_{[0,1]^d} b^{(i)}_{\mathbf{h}}(s,X^{\mathbf{h}}_{s-},u)\mathrm{d}u\right)\mathrm{d}s + \frac{1}{2}\int_0^t \sum_{i,j=1}^m \frac{\partial^2 f}{\partial x_i\partial x_j}(X^{\mathbf{h}}_{s-})\left(\int_{[0,1]^d} A^{(i,j)}_{\mathbf{h}}(s,X^{\mathbf{h}}_{s-},u)\mathrm{d}u\right)\mathrm{d}s\\
&\quad + \int_0^t\int_{\{0<|z|\le\mathfrak{r}\}\times[0,1]^d} \left(f(X^{\mathbf{h}}_{s-}+\gamma_{\mathbf{h}}(s,X^{\mathbf{h}}_{s-},z,u))-f(X^{\mathbf{h}}_{s-})-\sum_{i=1}^m \frac{\partial f}{\partial x_i}(X^{\mathbf{h}}_{s-})\gamma^{(i)}_{\mathbf{h}}(s,X^{\mathbf{h}}_{s-},z,u)\right)\nu_s(\mathrm{d}z)\mathrm{d}u\mathrm{d}s\\
&\quad + \int_0^t\int_{\{|z|>\mathfrak{r}\}\times[0,1]^d} (f(X^{\mathbf{h}}_{s-}+\gamma_{\mathbf{h}}(s,X^{\mathbf{h}}_{s-},z,u))-f(X^{\mathbf{h}}_{s-}))\nu_s(\mathrm{d}z)\mathrm{d}u\mathrm{d}s\\
&\quad + \text{local martingale terms}
\end{aligned}$$

$$= \int_0^t (\mathcal{L}_{\mathbf{h}} f)(s, X^{\mathbf{h}}_{s-})\mathrm{d}s + \text{local martingale terms.}$$

Since f has compact support, the local martingale terms become a proper bounded martingale. Hence, the law of $X^{\mathbf{h}}$ solves the said martingale problem above. □

9. Conclusion

The seminal development of reinforcement learning in continuous time in [18,19,34,35] has been based on a pair of stochastic processes: the exploratory SDE and the sample state process. Whereas the exploratory SDE is considered to be unobservable and is applied to study theoretical aspects, the sample state process is supposed to model the execution of randomized policies in continuous time. Due to the measurability issues with the sample state process, which have been noticed in [32] and which are further detailed in Section 3 above, a mathematical theory of modeling exploration via the sample state process cannot be developed rigorously. Building on an idea of [32], we study grid sampling as an alternative to the idealized sampling mechanism which underlies the definition of the sample state process. Compared to [32], who prove convergence of the expected cost when executing Gaussian controls on a grid for linear-quadratic problems in a diffusion setting, our main limit theorem (Theorem 3) is, to the best of our knowledge, the first result connecting policy execution via grid sampling to the exploratory dynamics in a general framework of nonlinear state dynamics and, moreover, includes controlled jumps. After the preprint version of our paper has been posted, several authors took up the approach of replacing the sample state process by grid sampling for modeling exploration in continuous-time reinforcement learning: The speed of convergence of the one-dimensional marginal distributions of (a variant of) the grid-sampling SDE to the exploratory SDE is derived in [16] in a diffusion setting. Moreover, in the erratum [20] to [19] and in the recent arXiv version[1] of [10], the sample state process has been replaced by grid sampling. These developments in the very recent literature indicate that grid sampling, which has been initiated in [32] and which is further advanced in the present paper, could become the accepted approach for modeling the execution of randomized policies in continuous-time reinforcement learning.

Acknowledgment

The paper benefited from the constructive comments of the reviewers. This paper is partially based on the preprints “A random measure approach to reinforcement learning in continuous time” (arXiv:2409.17200) and “On the grid-sampling limit SDE” (arXiv:2410.07778) by the same authors.

Appendix A. SDEs driven by orthogonal martingale measures

A.1. Martingale measures and integration

We briefly recall the notion of martingale measure initiated by Walsh [37]. We consider here the finite time interval $[0,T]$ but note that the discussion below can be readily extended for $[0,\infty)$. Let (E,d_E) be a complete and separable metric space equipped with its Borel σ-field $\mathcal{B}(E)$. A mapping $M : \Omega\times[0,T]\times\mathcal{B}(E)\to\mathbb{R}$ is called an $(\mathbb{F},\mathbb{P})$-*martingale measure* on $[0,T]\times\mathcal{B}(E)$ if:

(1) For $A\in\mathcal{B}(E)$, $(M(t,A))_{t\in[0,T]}$ is an $\mathbf{L}^2(\mathbb{P})$-martingale adapted with $\mathbb{F}$ and $M(0,A)=0$;
(2) For $t\in[0,T]$ and disjoint $A,B\in\mathcal{B}(E)$, one has $M(t,A\cup B)=M(t,A)+M(t,B)$ a.s.;
(3) There exists a non-decreasing sequence $(E_n)_{n\in\mathbb{N}}\subseteq\mathcal{B}(E)$ with $\cup_{n\in\mathbb{N}}E_n=E$ such that
  (a) For any $n\in\mathbb{N}$, $\sup_{A\in\mathcal{B}(E_n)}\|M(T,A)\|_{\mathbf{L}^2(\mathbb{P})}<\infty$;
  (b) For any $n\in\mathbb{N}$, one has $\|M(T,A_k)\|_{\mathbf{L}^2(\mathbb{P})}\to 0$ for all decreasing sequence $(A_k)_{k\in\mathbb{N}}\subseteq\mathcal{B}(E_n)$ with $\cap_{k\in\mathbb{N}}A_k=\emptyset$.

An $(\mathbb{F},\mathbb{P})$-martingale measure M is said to be *continuous* if $[0,T]\ni t\mapsto M(t,A)$ is continuous for all $A\in\mathcal{B}(E)$. Note that, due to the usual conditions, we always choose the càdlàg version of the martingale $M(\cdot,A)$ for any $A\in\mathcal{B}(E)$.

An $(\mathbb{F},\mathbb{P})$-martingale measure M is *orthogonal* if $M(\cdot,A)M(\cdot,B)$ is an $(\mathbb{F},\mathbb{P})$-martingale for any disjoint $A,B\in\mathcal{B}(E)$. It is indicated by Walsh [37] (see also [23, Theorem I-4]) that if an $(\mathbb{F},\mathbb{P})$-martingale measure M is orthogonal, then there is a random positive finite measure μ_M on $\mathcal{B}([0,T]\times E)$, which is $\mathbb{F}$-predictable (i.e. $(\mu_M((0,t]\times A))_{t\in[0,T]}$ is $\mathbb{F}$-predictable for all $A\in\mathcal{B}(E)$), such that

$$\mu_M((0,t]\times A)=\langle M(\cdot,A)\rangle_t \quad \mathbb{P}\text{-a.s.}, \quad \forall(t,A)\in[0,T]\times\mathcal{B}(E).$$

The measure μ_M is then called the *intensity measure* of M. Moreover, for $t\in[0,T]$, $A,B\in\mathcal{B}(E)$,

$$\langle M(\cdot,A),M(\cdot,B)\rangle_t=\langle M(\cdot,A\cap B)\rangle_t=\mu_M((0,t]\times(A\cap B)) \quad \mathbb{P}\text{-a.s.}$$

The stochastic integrals driven by an orthogonal martingale measure M can be constructed via the Itô's approach (see [23,37]) as follows: We first define the integrals for $\mathbb{F}$-predictable simple integrands H, and then, extend the integrals for $H\in\mathbf{L}^2(\mathbb{F},\mu_M)$ by the denseness, where $\mathbf{L}^2(\mathbb{F},\mu_M)$ is the collection of all $\mathbb{F}$-predictable H with $\mathbb{E}\left[\int_0^T\int_E H(t,x)^2\mu_M(\mathrm{d}t,\mathrm{d}x)\right]<\infty$.

[1] version 4, August 2025.

Assume that the intensity μ_M of M satisfies $\mu_M(\{t\}\times E)=0$ for all $t\in[0,T]$ a.s. Then, by a localization argument, one can extend the stochastic integrals driven by M for $H\in\mathbf{L}^2_{\mathrm{loc}}(\mathbb{F},\mu_M)$, where $\mathbf{L}^2_{\mathrm{loc}}(\mathbb{F},\mu_M)$ consists of all $\mathbb{F}$-predictable H with $\int_0^T\int_E H(t,x)^2\mu_M(\mathrm{d}t,\mathrm{d}x)<\infty$ a.s. (see, e.g., [25, Chapter 13] for continuous M). We refer to [2] for further details.

The following two lemmas are standard and their proofs can be found in [2].

Lemma A.1. *Let M be an orthogonal $(\mathbb{F},\mathbb{P})$-martingale measure with intensity μ_M. Then, for any $\mathbb{F}$-stopping times $\sigma,\tau:\Omega\to[0,T]$ with $\sigma\le\tau$, $A\in\mathcal{B}(E)$, and any bounded $\mathcal{F}_\sigma$-measurable $h:\Omega\to\mathbb{R}$, one has, a.s.,*

$$\int_{(0,T]\times E} h\mathbb{1}_{(\sigma,\tau]}(s)\mathbb{1}_A(e)M(\mathrm{d}s,\mathrm{d}e)=h[M(\tau,A)-M(\sigma,A)].$$

Lemma A.2. *Let M be an orthogonal $(\mathbb{F},\mathbb{P})$-martingale measure with intensity $\mu_M(\mathrm{d}s,\mathrm{d}e)=\mu_s(\mathrm{d}e)\mathrm{d}s$ for some transition kernel $\{(\omega,s,A)\mapsto\mu_s(\omega,A),(\omega,s)\in\Omega\times[0,T],A\in\mathcal{B}(E)\}$. For an $\mathbb{F}$-stopping time $\tau:\Omega\to[0,T]$, we define $\mathbb{F}^\tau=(\mathcal{F}^\tau_t)_{t\in[0,T]}$ with $\mathcal{F}^\tau_t:=\mathcal{F}_{(\tau+t)\wedge T}$ and*

$$M_\tau(t,A):=M((\tau+t)\wedge T,A)-M(\tau,A),\quad (t,A)\in[0,T]\times\mathcal{B}(E).$$

Then, M_τ is an orthogonal $(\mathbb{F}^\tau,\mathbb{P})$-martingale measure with ($\mathbb{F}^\tau$-predictable) intensity $\mu_{M_\tau}(\mathrm{d}s,\mathrm{d}e)=\mathbb{1}_{(0,T-\tau]}(s)\mu_{\tau+s}(\mathrm{d}e)\mathrm{d}s$. Moreover, for $g:\Omega\times[0,T]\times E\to\mathbb{R}$ with $\{(\omega,s,e)\mapsto\mathbb{1}_{(\tau(\omega),T]}(s)g(\omega,s,e)\}\in\mathbf{L}^2_{\mathrm{loc}}(\mathbb{F},\mu_M)$, one has $\{(\omega,s,e)\mapsto\mathbb{1}_{(0,T-\tau(\omega)]}(s)g(\omega,\tau(\omega)+s,e)\}\in\mathbf{L}^2_{\mathrm{loc}}(\mathbb{F}^\tau,\mu_{M_\tau})$ and, a.s.,

$$\int_{(0,T]\times E} g(s,e)\mathbb{1}_{(\tau,T]}(s)M(\mathrm{d}s,\mathrm{d}e)=\int_{(0,T]\times E} g(\tau+s,e)\mathbb{1}_{(0,T-\tau]}(s)M_\tau(\mathrm{d}s,\mathrm{d}e),$$

where the stochastic integrals on the left-hand side and on the right-hand side are constructed in relation to $\mathbb{F}$ and $\mathbb{F}^\tau$, respectively.

A.2. SDEs driven by orthogonal martingale measures

Let $\{M^{(1)},\dots,M^{(\ell)}\}$ be a collection of (càdlàg) $(\mathbb{F},\mathbb{P})$-martingale measures on $[0,T]\times\mathcal{B}(E)$. Assume that each $M^{(j)}$ is an orthogonal martingale measure with (random) intensity measure $\mu^{(j)}$ satisfying

$$\mu^{(j)}(\omega,\mathrm{d}s,\mathrm{d}e)=\mu^{(j)}_s(\omega,\mathrm{d}e)\mathrm{d}s\quad\mathbb{P}\text{-a.s. }\omega\in\Omega$$

for some transition kernel $\{(\omega,s,A)\mapsto\mu^{(j)}_s(\omega,A),(\omega,s)\in\Omega\times[0,T],A\in\mathcal{B}(E)\}$, $j=1,\dots,\ell$.

Let $\beta:\Omega\times[0,T]\times\mathbb{R}^m\to\mathbb{R}^m$ be $\mathcal{P}_\mathbb{F}\otimes\mathcal{B}(\mathbb{R}^m)/\mathcal{B}(\mathbb{R}^m)$-measurable, $\alpha:\Omega\times[0,T]\times\mathbb{R}^m\times E\to\mathbb{R}^{m\times\ell}$ be $\mathcal{P}_\mathbb{F}\otimes\mathcal{B}(\mathbb{R}^m)\otimes\mathcal{B}(E)/\mathcal{B}(\mathbb{R}^{m\times\ell})$-measurable and consider the following m-dimensional SDE

$$Y_t=Y_0+\int_0^t\beta(s,Y_{s-})\mathrm{d}s+\int_{(0,t]\times E}\alpha(s,Y_{s-},e)M(\mathrm{d}s,\mathrm{d}e),\quad t\in[0,T],\tag{A.1}$$

for some given $\mathcal{F}_0$-measurable $\mathbb{R}^m$-valued random variable Y_0, and for $M:=(M^{(1)},\dots,M^{(\ell)})^\top$.

Proposition A.3. *Assume that there exist constants $K_\beta,K_\alpha\ge0$ not depending on (ω,s,y_1,y_2) such that, for $\mathbb{P}$-a.s. $\omega\in\Omega$ and for all $s\in[0,T]$, $y_1,y_2\in\mathbb{R}^m$,*

$$|\beta(\omega,s,y_1)-\beta(\omega,s,y_2)|\le K_\beta|y_1-y_2|,$$
$$4\ell\sum_{i=1}^m\sum_{j=1}^\ell\int_E|\alpha^{(i,j)}(\omega,s,y_1,e)-\alpha^{(i,j)}(\omega,s,y_2,e)|^2\,\mu^{(j)}_s(\omega,\mathrm{d}e)\le K_\alpha^2|y_1-y_2|^2,$$

and that

$$K_0^2:=\mathbb{E}\left[T\int_0^T|\beta(s,0)|^2\mathrm{d}s+4\ell\sum_{i=1}^m\sum_{j=1}^\ell\int_0^T\int_E|\alpha^{(i,j)}(s,0,e)|^2\,\mu^{(j)}_s(\mathrm{d}e)\mathrm{d}s\right]<\infty.$$

Then, for any $\mathcal{F}_0$-measurable initial condition Y_0, the SDE (A.1) has a unique (up to an indistinguishability) strong solution Y.

Proof. See [2]. □

Remark A.4. The proof of Proposition A.3 reveals that, if in addition $Y_0\in\mathbf{L}^2(\mathbb{P})$ then the strong solution of (A.1) satisfies

$$\mathbb{E}\left[\sup_{0\le t\le T}|Y_t|^2\right]\le K(1+\mathbb{E}[|Y_0|^2])$$

for some constant $K\ge0$ depending only on K_α,K_β,K_0,T.

A.3. On the Markovianity of the grid-sampling (limit) SDE

We first argue that, under the assumptions of Proposition 5.3, the solution $X^{\mathbf{h}}$ to the grid-sampling limit SDE is Markovian with respect to $\mathbb{F}$; cp. Remark 5.4. To this end, fix $r \in [0,T)$ and consider the SDE

$$\begin{aligned} X_t^{\eta,r} = \eta &+ \int_r^t \int_{[0,1]^d} b_{\mathbf{h}}(s, X_{s-}^{\eta,r}, u)\mathrm{d}u\mathrm{d}s + \sum_{l=1}^p \int_{(r,t]\times[0,1]^d} a_{\mathbf{h}}^{(\cdot,l)}(s, X_{s-}^{\eta,r}, u) M_{B^{(l)}}(\mathrm{d}s, \mathrm{d}u) \\ &+ \int_{(r,t]\times\{0<|z|\le \mathfrak{r}\}\times[0,1]^d} \gamma_{\mathbf{h}}(s, X_{s-}^{\eta,r}, z, u)\tilde{M}_J(\mathrm{d}s, \mathrm{d}z, \mathrm{d}u) + \int_{(r,t]\times\{|z|>\mathfrak{r}\}\times[0,1]^d} \gamma_{\mathbf{h}}(s, X_{s-}^{\eta,r}, z, u) M_J(\mathrm{d}s, \mathrm{d}z, \mathrm{d}u), \end{aligned} \tag{A.2}$$

for some $\mathcal{F}_r$-measurable random variable η. This is the grid-sampling limit SDE restarted in η at time r. The proof of Proposition 5.3 applies in this situation as well and shows that a unique strong solution $(X_t^{\eta,r})_{t\in[r,T]}$ to SDE (A.2) exists. Moreover, the technique of proof via Picard iteration, iterated truncation of the initial condition, and the interlacing technique shows that, for every $t \in [r,T]$, $X_t^{\eta,r}$ is $\sigma(\eta) \vee \mathcal{F}_T^r$-measurable, where

$$\mathcal{F}_T^r = \sigma(\{M_{B^{(i)}}((r,u]\times A_i), M_J((r,u]\times A') : r < u \le T, A' \in \mathcal{B}(\mathbb{R}_0^d \times [0,1]^d),\ A_i \in \mathcal{B}([0,1]^d), i = 1,\dots,p\}) \vee \mathcal{N},$$

and, as before, $\mathcal{N}$ denotes the collection of all $\mathbb{P}$-null sets. We now observe that the solution $(X_t^{\mathbf{h}})_{t\in[r,T]}$ to the grid-sampling limit SDE (5.1) solves (A.2) with $\eta = X_r^{\mathbf{h}}$ and, thus, for every $t \in [r,T]$, $X_t^{\mathbf{h}}$ is $\sigma(X_r^{\mathbf{h}}) \vee \mathcal{F}_T^r$-measurable. Since $X_r^{\mathbf{h}}$ is $\mathcal{F}_r$-measurable and $\mathcal{F}_T^r$ is independent of $\mathcal{F}_r$, a standard argument based Dynkin's π-λ theorem implies that, for every $t \in [r,T]$ and $A \in \mathcal{B}(\mathbb{R}^m)$,

$$\mathbb{E}\Big[\mathbb{E}\Big[\mathbb{1}_{\{X_t^{\mathbf{h}}\in A\}} | \sigma(X_r^{\mathbf{h}}) \vee \mathcal{F}_T^r\Big] | \mathcal{F}_r\Big] = \mathbb{E}\Big[\mathbb{1}_{\{X_t^{\mathbf{h}}\in A\}} | \sigma(X_r^{\mathbf{h}})\Big].$$

Therefore,

$$\mathbb{P}(\{X_t^{\mathbf{h}} \in A\} | \mathcal{F}_r) = \mathbb{E}\Big[\mathbb{E}\Big[\mathbb{1}_{\{X_t^{\mathbf{h}}\in A\}} | \sigma(X_r^{\mathbf{h}}) \vee \mathcal{F}_T^r\Big] | \mathcal{F}_r\Big] = \mathbb{E}\Big[\mathbb{1}_{\{X_t^{\mathbf{h}}\in A\}} | \sigma(X_r^{\mathbf{h}})\Big] = \mathbb{P}(\{X_t^{\mathbf{h}} \in A\} | X_r^{\mathbf{h}}).$$

We now turn to the grid-sampling SDE (4.1) and sketch the analogous argument, which shows that the pair $(X_t^{\Pi,\mathbf{h}}, \xi_{t+}^{\Pi})_{t\in[0,T]}$ is Markovian with respect to $\mathbb{F}^{\Pi}$ under the assumptions of Proposition 4.2; cp. Remark 4.3. Fixing $r \in [0,T)$ and restarting (4.1) at time r, we observe, following an analogous argument as above that $X_t^{\Pi,\mathbf{h}}$ is $\sigma(X_r^{\Pi,\mathbf{h}}) \vee \mathcal{F}_T^{\Pi,r}$-measurable, where

$$\mathcal{F}_T^{\Pi,r} = \sigma(\{B_s - B_r, \xi_s^{\Pi}, N((r,s]\times A') : A' \in \mathcal{B}(\mathbb{R}_0^q), s \in (r,T]\}) \vee \mathcal{N}.$$

Now, let t_{i_0} denote the smallest grid point which is strictly larger than r. Then, for every $r < s \le t_{i_0}$, $\xi_s^{\Pi} = \xi_{r+}^{\Pi}$ is $\mathcal{F}_r^{\Pi}$-measurable, but $(\xi_s^{\Pi})_{t_{i_0}<s\le T}$ is independent of $\mathcal{F}_r^{\Pi}$. Therefore, $X_t^{\Pi,\mathbf{h}}$ is $\sigma(\{X_r^{\Pi,\mathbf{h}}, \xi_{r+}^{\Pi}\}) \vee \mathcal{F}_T^{\Pi,r,\times}$-measurable, where

$$\mathcal{F}_T^{\Pi,r,\times} = \sigma(\{B_s - B_r, \xi_{s'}^{\Pi}, N((r,s]\times A') : A' \in \mathcal{B}(\mathbb{R}_0^q), s \in (r,T], s' \in (t_{i_0},T]\}) \vee \mathcal{N},$$

is independent of $\mathcal{F}_r^{\Pi}$. As above, we may conclude that for every $t \in [r,T]$ and $A \in \mathcal{B}(\mathbb{R}^m)$,

$$\mathbb{E}\Big[\mathbb{E}\Big[\mathbb{1}_{\{X_t^{\Pi,\mathbf{h}}\in A\}} | \sigma(\{X_r^{\Pi,\mathbf{h}}, \xi_{r+}^{\Pi}\}) \vee \mathcal{F}_T^{\Pi,r,\times}\Big] | \mathcal{F}_r^{\Pi}\Big] = \mathbb{E}\Big[\mathbb{1}_{\{X_t^{\Pi,\mathbf{h}}\in A\}} | \sigma(\{X_r^{\Pi,\mathbf{h}}, \xi_{r+}^{\Pi}\})\Big],$$

leading to $\mathbb{P}(\{X_t^{\Pi,\mathbf{h}} \in A\} | \mathcal{F}_r^{\Pi}) = \mathbb{P}(\{X_t^{\Pi,\mathbf{h}} \in A\} | (X_r^{\Pi,\mathbf{h}}, \xi_{r+}^{\Pi}))$.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.spa.2025.104848

References

[1] C. Bender, N.T. Thuan, Entropy-regularized mean-variance portfolio optimization with jumps, arXiv:2312.13409, 2023.
[2] C. Bender, N.T. Thuan, Supplement to "Continuous time reinforcement learning: A random measure approach", 2024.
[3] P. Billingsley, Probability and measure, John Wiley & Sons, Inc., 3rd ed., 1995.
[4] P. Billingsley, Convergence of probability measures, John Wiley & Sons, Inc, 1999.
[5] Z. Brzeźniak, W. Liu, J. Zhu, Strong solutions for SPDE with locally monotone coefficients driven by Lévy noise, Nonlinear Anal. Real World Appl. 17 (2014) 283–310.
[6] M. Dai, Y. Dong, Y. Jia, Learning equilibrium mean-variance strategy, Math. Finance 33 (4) (2023) 1166–1212.
[7] R. Donnelly, S. Jaimungal, Exploratory control with Tsallis entropy for latent factor models, SIAM J. Financ. Math. 15 (1) (2024) 26–53.
[8] D. Firoozi, S. Jaimungal, Exploratory LQG mean field games with entropy regularization, Automatica 139 (2022) 110177.
[9] N. Frikha, M. Germain, M. Laurière, H. Pham, X. Song, Actor-critic learning for mean-field control in continuous time, J. Mach. Learn. Res. 26 (2025) 1–42.
[10] X. Gao, L. Li, X.Y. Zhou, Reinforcement learning for jump-diffusions, with financial applications, arXiv:2405.16449, 2024.
[11] M. Giegrich, C. Reisinger, Y. Zhang, Convergence of policy gradient methods for finite-horizon exploratory linear-quadratic control problems, SIAM J. Control Optim. 62 (2) (2024) 1060–1092.
[12] X. Guo, R. Xu, T. Zariphopoulou, Entropy regularization for mean field games with learning, Math. Oper. Res. 47 (2022) 3239–3260.
[13] X. Han, R. Wang, X.Y. Zhou, Choquet regularization for continuous-time reinforcement learning, SIAM J. Control Optim. 61 (5) (2023) 2777–2801.
[14] N. Ikeda, S. Watanabe, Stochastic differential equations and diffusion processes, North-Holland, 1989.
[15] J. Jacod, A. Shiryaev, Limit theorems for stochastic processes, Berlin; Heidelberg, Springer, 2003.
[16] Y. Jia, D. Ouyang, Y. Zhang, Accuracy of discretely sampled stochastic policies in continuous-time reinforcement learning, arXiv:2503.09981, 2025.
[17] Y. Jia, X.Y. Zhou, Policy evaluation and temporal difference learning in continuous time and space: a martingale approach, J. Mach. Learn. Res. 23 (2022a) 1–55.

[18] Y. Jia, X.Y. Zhou, Policy gradient and actor-critic learning in continuous time and space: theory and algorithms, J. Mach. Learn. Res. 23 (2022b) 1–50.
[19] Y. Jia, X.Y. Zhou, q-Learning in continuous time, J. Mach. Learn. Res. 24 (2023) 1–61.
[20] Y. Jia, X.Y. Zhou, Erratum to "q-Learning in continuous time", J. Mach. Learn. Res., 2025.
[21] I. Karatzas, S.E. Shreve, Brownian Motion and Stochastic Calculus, Springer, New York, 2nd ed., 1991.
[22] O. Kallenberg, Random measures, theory and applications, Springer, Cham, 2017.
[23] N.E. Karoui, S. Méléard, Martingale measures and stochastic calculus, Probab. Theory Rel. Field. 84 (1990) 83–101.
[24] H. Kunita, Stochastic differential equations based on Lévy processes and stochastic flows of diffeomorphisms, in: Real and Stochastic Analysis, Birkhäuser Boston, 2004.
[25] H.J. Kushner, P. Dupuis, Numerical methods for stochastic control problems in continuous time, Springer, New York, 2001. 2nd ed.
[26] S. Méléard, Representation and approximation of martingale measures, in: Stochastic Partial Differential Equations and Their Applications, Lecture Notes in Control and Information Sciences, 176 of Berlin Heidelberg, Springer, 1992.
[27] C. Reisinger, Y. Zhang, Regularity and stability of feedback relaxed controls, SIAM J. Control Optim. 59 (5) (2021) 3118–3151.
[28] H. Robbins, S. Monro, A stochastic approximation method, Ann. Math. Stat. 22 (1951) 400–407.
[29] Y. Sun, The exact law of large numbers via Fubini extension and characterization of insurable risks, J. Econ. Theory 126 (2006) 31–69.
[30] Y. Sun, Y. Zhang, Individual risk and Lebesgue extension without aggregate uncertainty, J. Econ. Theory 144 (2006) 432–443.
[31] R.S. Sutton, A.G. Barto, Reinforcement learning: An introduction, 2nd ed, MIT Press, Cambridge, MA, 2018.
[32] L. Szpruch, T. Treetanthiploet, Y. Zhang, Optimal scheduling of entropy regularizer for continuous-time linear-quadratic reinforcement learning, SIAM J. Control Optim. 62 (1) (2024) 135–166.
[33] W. Tang, Y.P. Zhang, X.Y. Zhou, Exploratory HJB equations and their convergence, SIAM J. Control Optim. 60 (6) (2022) 3191–3216.
[34] H. Wang, T. Zariphopoulou, X.Y. Zhou, Reinforcement learning in continuous time and space: a stochastic control approach, J. Mach. Learn. Res. 21 (2020) 1–34.
[35] H. Wang, X.Y. Zhou, Continuous-time mean-variance portfolio selection: a reinforcement learning framework, Math. Finance 30 (4) (2020) 1–36.
[36] B. Wu, L. Li, Reinforcement learning for continuous-time mean-variance portfolio selection in a regime-switching market, J. Econ. Dyn. Control 158 (2024) 104787.
[37] J. Walsh, An introduction to stochastic partial differential equations, Lect. Note. Maths 1180 (1986) 265–439.
[38] Y. Xie, Vague convergence of locally integrable martingale measures, Stochastic Process. Appl. 52 (1994) 211–227.
[39] Y. Xie, Limit theorems of Hilbert valued semimartingales and Hilbert valued martingale measures, Stochastic Process. Appl. 59 (2) (1995) 277–293.
[40] X.Y. Zhou, The curse of optimality, and how to break it?, in: Machine Learning and Data Sciences for Financial Markets. A Guide to Contemporary Practices, Cambridge, University Press, 2023, pp. 354–368.