

Rectified deep neural networks overcome the curse of dimensionality for nonsmooth value functions in zero-sum games of nonlinear stiff systems

Christoph Reisinger* Yufei Zhang†

Abstract. In this paper, we establish that for a wide class of controlled stochastic differential equations (SDEs) with stiff coefficients, the value functions of corresponding zero-sum games can be represented by a deep artificial neural network (DNN), whose complexity grows at most polynomially in both the dimension of the state equation and the reciprocal of the required accuracy. Such nonlinear stiff systems may arise, for example, from Galerkin approximations of controlled stochastic partial differential equations (SPDEs), or controlled PDEs with uncertain initial conditions and source terms. This implies that DNNs can break the curse of dimensionality in numerical approximations and optimal control of PDEs and SPDEs. The main ingredient of our proof is to construct a suitable discrete-time system to effectively approximate the evolution of the underlying stochastic dynamics. Similar ideas can also be applied to obtain expression rates of DNNs for value functions induced by stiff systems with regime switching coefficients and driven by general Lévy noise.

Key words. Deep neural networks, approximation theory, curse of dimensionality, optimal control, stochastic partial differential equation.

AMS subject classifications. 82C32, 41A25, 35R60

1 Introduction

In this paper, we study the expressive power of deep artificial neural networks (DNNs), and demonstrate that one can construct DNNs with polynomial complexity to approximate nonsmooth value functions associated with stiff stochastic differential equations (SDEs).

More precisely, for each $d \in \mathbb{N}$, we consider the value function $v_d : \mathbb{R}^d \rightarrow \mathbb{R}$ of the following d -dimensional zero-sum stochastic differential game on a finite time horizon $[0, T]$:

$$v_d(x) := \inf_{u_1 \in \mathcal{U}_{1,d}} \sup_{u_2 \in \mathcal{U}_{2,d}} \mathbb{E} \left[f_d(Y_T^{x,d,u_1,u_2}) + g_d(u_1, u_2) \right], \quad x \in \mathbb{R}^d,$$

where $\mathcal{U}_{i,d}$, $i = 1, 2$, are sets of admissible open-loop control strategies (see Section 2.2 for a precise definition), $f_d : \mathbb{R}^d \rightarrow \mathbb{R}$ is a (possibly nonsmooth) terminal cost function with at most quadratic growth at infinity, and for each $x \in \mathbb{R}^d$, $u_i \in \mathcal{U}_{i,d}$, $i = 1, 2$, $(Y_t^{x,d,u_1,u_2})_{t \in [0,T]}$ is the solution to the following d -dimensional controlled SDE:

$$dY_t = (-A_d Y_t + \mu_d(t, Y_t, u_1, u_2)) dt + \sigma_d(t, Y_t, u_1, u_2) dB_t, \quad t \in (0, T]; \quad Y_0 = x, \quad (1.1)$$

*Mathematical Institute, University of Oxford, United Kingdom (christoph.reisinger@maths.ox.ac.uk, yufei.zhang@maths.ox.ac.uk)

where A_d is a $d \times d$ matrix, μ_d and σ_d are respectively \mathbb{R}^d and $\mathbb{R}^{d \times d}$ -valued functions, and $(B_t)_{t \in [0, T]}$ is a d -dimensional Brownian motion defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

The above problem is called a zero-sum stochastic differential game since the underlying SDE (1.1) is controlled by two players with opposite objectives, i.e., the “inf-player” aims to minimize the associated cost function over all strategies $u_1 \in \mathcal{U}_{1,d}$, while the “sup-player” aims to maximize the same cost function over all strategies $u_2 \in \mathcal{U}_{2,d}$. The admissible controls $\mathcal{U}_{i,d}$, $i = 1, 2$, are called open-loop controls since they are deterministic processes; see page 23 of [51] for different types of strategies. In the case with $\sigma_d \equiv 0$, (1.1) degenerates to a controlled ordinary differential equation. Moreover, if one of the sets $\mathcal{U}_{1,d}$ and $\mathcal{U}_{2,d}$ is singleton, the zero-sum game reduces to an optimal control problem.

In this work, we shall allow the coefficients A_d , μ_d and σ_d to be stiff in the sense that they are Lipschitz continuous (with respect to the Euclidean norm on \mathbb{R}^d) but the Lipschitz constants grow polynomially in the dimension d . Such stiff SDEs arise naturally from spatial discretizations of stochastic partial differential equations (SPDEs) by using spectral methods (see e.g. [30, 19, 33, 37, 36]), or finite difference/element methods (see e.g. [19, 2, 15]).

For simplicity, let us consider the following uncontrolled SPDE as motivating example, but similar arguments also apply to controlled SPDEs. Let $B = (B(t))_{t \geq 0}$ be an m -dimensional Brownian motion on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, $H = L^2(\mathbb{R}^p)$, $V = H^1(\mathbb{R}^p)$, and V^* denotes the strong dual space of V . Here H^* is identified with H so that $V \subset H = H^* \subset V^*$. Then, for given mappings $A : V \mapsto V^*$, $U : [0, T] \times V \mapsto V^*$, and $G : [0, T] \times V \mapsto H^m$, it has been shown in [30, 19, 33] that the following semilinear SPDE (with a spatial domain in \mathbb{R}^p):

$$dy(t) + Ay(t) dt = U(t, y(t)) dt + G(t, y(t)) dB(t), \quad t \in (0, T], \quad y(0) = y_0 \in H, \quad (1.2)$$

admits a solution y under the following strong monotonicity condition¹: there exist some $\lambda, \beta > 0$, such that for all $t \in [0, T]$, $u, v \in V$,

$$2\langle u - v, -A(u - v) + U(t, u) - U(t, v) \rangle_{V \times V^*} + \|G(t, u) - G(t, v)\|_H^2 \leq -\lambda \|u - v\|_V^2 + \beta \|u - v\|_H^2, \quad (1.3)$$

where $\langle \cdot, \cdot \rangle_{V \times V^*}$ denotes the duality product of $V \times V^*$ (see e.g. Assumption 2.1(i) in [19]). Important special cases of (1.2) include suitable semilinear parabolic PDEs with (additive or multiplicative) noise and the Zakai equation from nonlinear filtering (see e.g. [30]) or from a large pool limit of interacting particles (see e.g. [10, 15]).

We are interested in the value functional associated with the SPDE (1.2):

$$\mathcal{V} : y(0) \in H \mapsto \mathbb{E}[f(y(T))] \in \mathbb{R}, \quad (1.4)$$

where $y(0)$ is taken within a neighbourhood of the initial condition y_0 in (1.2), and $f : V \rightarrow \mathbb{R}$ is a given locally Lipschitz cost functional. This is practically important if the exact dynamics of (1.2) is only known subject to uncertain initial conditions, or if we would like to compare the optimal cost of a control problem among all initial states (see e.g. [14, 24]). An accurate representation of the value functional is also crucial for the control design in reinforcement learning (see [4]).

In practice, we shall consider a finite-dimensional version of the (infinite-dimensional) value functional \mathcal{V} . Let $\{e_k\}_{k \geq 1}$ be an orthonormal basis of H , made of elements in V , and $H_d = \text{span}\{e_k \mid k = 1, \dots, d\}$ for all $d \in \mathbb{N}$. Then for each $d \geq 1$, we can project the SPDE (1.2) onto

¹ In general, the coefficients U and G need to satisfy other technical assumptions, such as continuity, coercivity and growth conditions, to ensure the well-posedness of (1.2) in $L^2(\Omega \times [0, T]; V)$; see e.g. [19]. However, since we only use (1.2) to motivate the high-dimensional stiff SDE (1.5) and shall establish approximation results for the corresponding high-dimensional value functions, we omit other technical assumptions on the coefficients U and G here and introduce the precise conditions for the finite-dimensional SDEs in Sections 2.1 and 2.2.

the subspace H_d and consider a d -dimensional Itô-Galerkin approximation of (1.2) in \mathbb{R}^d of the form:

$$dy_d(t) + A_d y_d(t) dt = U_d(t, y_d(t)) dt + G_d(t, y_d(t)) dB(t), \quad t \in (0, T], \quad y_d(0) = y_{0,d}, \quad (1.5)$$

where the discrete operators A_d, U_d, G_d satisfy a monotonicity condition similar to (1.3). Then, under suitable regularity assumptions, one can show the well-posedness of a solution y_d to the finite-dimensional SDE (1.5), and estimate the rate of convergence in terms of the dimension d . The convergence of $y_d(T)$ to $y(T)$ as $d \rightarrow \infty$ suggests us to approximate the functional \mathcal{V} by the d -dimensional value function

$$v_d : y_d(0) \in H_d \mapsto \mathbb{E}[f(y_d(T))] \in \mathbb{R}$$

with a sufficiently large $d \in \mathbb{N}$. Note that for SPDEs driven by H -valued random fields, one can consider similar Itô-Galerkin SDEs with finite-dimensional noises by truncating the series representation of the (space-time) random process (see [2] for sufficient conditions under which this extra approximation of the noise preserves the overall convergence order in d).

However, we face several difficulties in approximating the d -dimensional value function v_d . Recall that the errors of the Galerkin approximations for the SPDE (1.2) (with a spatial domain in \mathbb{R}^p) are in general of the magnitude $\mathcal{O}(d^{-\gamma/p})$ for some $\gamma > 0$ (see e.g. [19, 2, 15]). Thus, the local Lipschitz continuity of the cost function in (1.4) suggests that, to achieve an accuracy ε in representing the value functional $\mathcal{V} : H = L^2(\mathbb{R}^p) \rightarrow \mathbb{R}$, we need to approximate the value function $v_d : H_d \rightarrow \mathbb{R}$, where H_d can be identified as a d -dimensional Euclidean space with $d = \mathcal{O}(\varepsilon^{-p/\gamma})$. Since many classical function approximation methods, such as piecewise constant and piecewise linear approximations, require a complexity of $\mathcal{O}(\varepsilon^{-d})$ to approximate a d -dimensional function within an accuracy ε , we see the total complexity for classical methods to represent the value functional (1.4) within the accuracy ε is of the magnitude $\mathcal{O}(\varepsilon^{-\varepsilon^{-p/\gamma}})$, which suffers from the so-called Bellman's curse of dimensionality.

Moreover, the control processes and the nonsmoothness of the terminal costs imply that the value function v_d typically has weak regularity, e.g. v_d is merely locally Lipschitz continuous and could grow quadratically at infinity. This prevents us from approximating the value function by using sparse grid approximations [7, 49], or high-order polynomial expansions [9]. Finally, since the mappings A, U and G in (1.2) could involve differential operators, the Lipschitz constants (with respect to the Euclidean norm) of A_d, U_d, G_d in (1.5) will in general grow polynomially in dimension d . This stiffness of coefficients creates a difficulty in constructing efficient discrete-time dynamics to approximate the time evolution of the Itô-Galerkin SDE (1.5).

In recent years, DNNs have achieved remarkable performance in representing high-dimensional mappings in a wide range of applications (see e.g. [37, 35, 36, 43, 20, 3, 25, 31, 48] and the references therein for applications in optimal control and numerical simulation of PDEs), and it seems that DNNs admit the flexibility to overcome the curse of dimensionality. However, even though there is a vast literature on the approximation theory of artificial neural networks (see e.g. [23, 38, 44, 47, 52, 1, 11, 12, 21, 27, 34, 50, 5, 28, 29, 17, 18, 45]), to the best of our knowledge, only [12, 16, 27, 34, 28, 29] established DNNs' expression rates for approximating nonsmooth value functions (associated with d -dimensional SDEs whose diffusion coefficients are affine with respect to the state variable and both the drift and diffusion coefficients are Lipschitz continuous with a constant *independent of the dimension d*).

In this work, we shall extend their results by giving a rigorous proof of the fact that DNNs do overcome the curse of dimensionality for approximating (nonsmooth) value functions of zero-sum games of *controlled* SDEs with *stiff, time-inhomogeneous, nonlinear* coefficients. More precisely,

we shall establish that for a wide class of controlled stiff SDEs, to represent the corresponding value functions with accuracy ε , the number of parameters in the employed DNNs grows at most polynomially in both the dimension of the state equation and the reciprocal of the accuracy ε (see Theorems 2.1 and 2.3). As a direct consequence of these expression rates, we show that one can approximate the viscosity solution to a Kolmogorov backward PDE with stiff coefficients by DNNs with polynomial complexity (see Corollary 2.2). In particular, if one further assumes that the Galerkin approximation of a controlled SPDE has a convergence rate $\mathcal{O}(d^{-\gamma})$ for some $\gamma > 0$, our result indicates that we can represent the nonlinear value functional \mathcal{V} without the curse of dimensionality.

The approach we take here is to first describe the evolution of a d -dimensional controlled SDE (1.1) by using a suitable discrete-time dynamical system, and then constructing the desired DNN by a specific realization of the discrete-time dynamics. This is of the same spirit as [34], where the authors represent an uncontrolled SDE with constant diffusion and nonlinear drift coefficients by its explicit Euler discretization. However, due to the stiffness of the Itô-Galerkin SDEs considered in this paper, such an explicit time discretization will in fact lead to an approximation error depending exponentially on the dimension d (cf. [34, Proposition 4.4]), and hence it cannot be used in our construction. We shall overcome this difficulty by approximating the underlying dynamics with its partial-implicit Euler discretization, whose error depends polynomially on the dimension d and the (time) stepsize. We also adopt a two-step approximation of the terminal cost function involving truncation and extrapolation, which allows us to construct rectified neural networks for quadratically growing terminal costs; see the discussion below (H.1) for details.

The rest of this paper is structured as follows. Section 2 states the assumptions and presents the main theoretical results of the expression rates. We discuss several fundamental operations of DNNs in Section 3, and analyze a perturbed linear-implicit Euler discretization of SDEs in Section 4. Based on these estimates, we establish the expression rates of rectified neural networks for uncontrolled systems in Section 6, and controlled systems in Section 7. Section 8 offers possible extensions and directions for further research.

2 Main results

In this section, we shall recall the notion of DNN, and state our main results on the expression rates of DNNs for approximating value functions associated with controlled SDEs with stiff coefficients.

We start with some notation which is needed frequently throughout this work. For any given $d \in \mathbb{N}$, we denote by $\|\cdot\|$ the Euclidean norm of a vector in \mathbb{R}^d , by $\langle \cdot, \cdot \rangle$ the canonical Euclidean inner product, and by I_d the $d \times d$ identity matrix. For a given matrix $A \in \mathbb{R}^{d_1 \times d_2}$, we denote by $\|A\|$ the Frobenius norm of A , and by $\|A\|_{\text{op}}$ the matrix norm induced by Euclidean vector norms. We shall also denote by C a generic constant, which may take a different value at each occurrence. Dependence of C on parameters will be indicated explicitly by $C_{(\cdot)}$, e.g. $C_{(\alpha, \beta)}$.

Now we introduce the basic concepts of DNNs. By following the notation in [47, 12, 16] (up to some minor changes), we shall distinguish between a deep artificial neural network, represented as a structured set of weights, and its realization, a multi-valued function on \mathbb{R}^d . This enables us to construct complex neural networks from simple ones in an explicit and unambiguous way, and further analyze the complexity of DNNs.

Definition 2.1 (Deep artificial neural networks). Let \mathcal{N} be the set of DNNs given by

$$\mathcal{N} = \bigcup_{L \in \mathbb{N}} \bigcup_{(N_0, N_1, \dots, N_L) \in \mathbb{N}^{L+1}} \mathcal{N}_L^{N_0, N_1, \dots, N_L}, \quad \text{where } \mathcal{N}_L^{N_0, N_1, \dots, N_L} = \bigtimes_{l=1}^L (\mathbb{R}^{N_l \times N_{l-1}} \times \mathbb{R}^{N_l}).$$

Let $\mathcal{C}, \mathcal{L}, \dim_{\text{in}}, \dim_{\text{out}} : \mathcal{N} \rightarrow \mathbb{N}$ and $\dim : \mathcal{N} \rightarrow \cup_{L \in \mathbb{N}} \mathbb{N}^{L+1}$ be functions such that for any given $\phi \in \mathcal{N}_L^{N_0, N_1, \dots, N_L}$, we have $\mathcal{C}(\phi) = \sum_{l=1}^L N_l(N_{l-1} + 1)$, $\mathcal{L}(\phi) = L$, $\dim_{\text{in}}(\phi) = N_0$, $\dim_{\text{out}}(\phi) = N_L$ and $\dim(\phi) = (N_0, N_1, \dots, N_L)$. We shall refer to the quantities $\mathcal{C}(\phi)$, $\mathcal{L}(\phi)$, $\dim_{\text{in}}(\phi)$ and $\dim_{\text{out}}(\phi)$ as the size, depth, input dimension and output dimension of the DNN ϕ , respectively.

For any given activation function $\varrho \in C(\mathbb{R}; \mathbb{R})$, let $\varrho^* : \cup_{d \in \mathbb{N}} \mathbb{R}^d \rightarrow \cup_{d \in \mathbb{N}} \mathbb{R}^d$ be the function which satisfies for all $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ that $\varrho^*(x) = (\varrho^*(x_1), \dots, \varrho^*(x_d))$, and let $\mathcal{R}_\varrho : \mathcal{N} \rightarrow \cup_{a, b \in \mathbb{N}} C(\mathbb{R}^a; \mathbb{R}^b)$ be the realization operator such that for any given $x_0 \in \mathbb{R}^{N_0}$ and

$$\phi = ((W_1, b_1), (W_2, b_2), \dots, (W_L, b_L)) \in \mathcal{N}_L^{N_0, N_1, \dots, N_L}, \quad \text{with } L \in \mathbb{N} \text{ and } (N_0, N_1, \dots, N_L) \in \mathbb{N}^{L+1},$$

we have $\mathcal{R}_\varrho(\phi) \in C(\mathbb{R}^{N_0}; \mathbb{R}^{N_L})$ defined recursively as follows: let $x_l = \varrho^*(W_l x_{l-1} + b_l)$ for all $l = 1, \dots, L-1$, and let

$$[\mathcal{R}_\varrho(\phi)](x_0) = W_L x_{L-1} + b_L.$$

Roughly speaking, one can describe a DNN by its *architecture*, that is the number of layers L and the dimensions of all layers N_0, N_1, \dots, N_L , together with the coefficients of the affine functions used to compute each layer from the previous one. Note that Definition 2.1 does not specify a fixed nonlinear activation function in the architecture of a DNN, but instead considers the realization of a DNN with respect to a given activation function, which allows us to study the approximation capacity of DNNs with arbitrary activation functions (see e.g. Lemma A.1).

To simplify the presentation, in the work we shall mainly focus on DNNs with the commonly used Rectified Linear Unit (ReLU) activation function, i.e., $\varrho(x) = \max(0, x)$, due to its representation flexibility. Moreover, we allow the weights of a DNN (i.e., the coefficients of the affine functions) to take arbitrary real numbers when approximating a given function. A similar analysis can be carried out for networks with quantization (i.e., the maximal magnitude of weights in the network is *a priori* fixed), by allowing the *a priori* bound of the weights to increase in a controlled way (see e.g. [47]).

For any given DNN $\phi \in \mathcal{N}$, the quantity $\mathcal{C}(\phi) \in \mathbb{N}$ represents the number of all real parameters, including zeros, used to describe the DNN. We remark that one can also consider the number of non-zero entries of the DNN ϕ as in [12]. However, since it is in general difficult to build a sparse architecture with pre-allocated zero entries to approximate a desired value function, we choose to adopt the notation of ‘size’ by considering all parameters and quantify the complexity of the DNN in a conservative manner.

Motivated by the application to optimal control problems of SPDEs, in the remaining part of this section, we shall construct a sequence of DNNs $(\psi_{\varepsilon, d})_{\varepsilon, d}$, such that for each $\varepsilon \in (0, 1)$, $d \in \mathbb{N}$, $\psi_{\varepsilon, d}$ represents the value function v_d induced by a d -dimensional stiff SDE with the accuracy ε on \mathbb{R}^d . We shall demonstrate that under a monotonicity condition similar to (1.3), the complexity of the constructed DNN $\psi_{\varepsilon, d}$ depends polynomially on both d and ε^{-1} , i.e., the DNNs $(\psi_{\varepsilon, d})_{\varepsilon, d}$ overcome the curse of dimensionality. We first give the results for uncontrolled SDEs with stiff coefficients in Section 2.1, and then extend the results to controlled SDEs with piecewise-constant strategies in Section 2.2.

2.1 Expression rate for SDEs and Kolmogorov PDEs with stiff coefficients

In this section, we present the expression rate of DNNs for approximating value functions induced by nonlinear SDEs with stiff coefficients.

We start by introducing the value functions of interest. For each $d \in \mathbb{N}$, we consider the following value function:

$$v_d : x \in \mathbb{R}^d \mapsto \mathbb{E}[f_d(Y_T^{x,d})] \in \mathbb{R}, \quad (2.1)$$

where $Y^{x,d} = (Y_t^{x,d})_{t \in [0,T]}$ is the strong solution to the following d -dimensional SDE:

$$dY_t^{x,d} = (-A_d Y_t^{x,d} + \mu_d(t, Y_t^{x,d})) dt + \sigma_d(t, Y_t^{x,d}) dB_t, \quad t \in (0, T]; \quad Y_0^{x,d} = x, \quad (2.2)$$

with a d -dimensional Brownian motion $(B_t)_{t \in [0,T]}$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

We now list the main assumptions on the coefficients.

H.1. Let $\beta, \kappa_0, T \geq 0$ and $\eta > 0$ be fixed constants. For all $d \in \mathbb{N}$ and $D > 0$, let $A_d \in \mathbb{R}^{d \times d}$, and $\mu_d : [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}^d$, $\sigma_d : [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d}$, $f_d, f_{d,D} : \mathbb{R}^d \rightarrow \mathbb{R}$ be measurable functions satisfying the following conditions:

(a) The matrix A_d and the functions (μ_d, σ_d) satisfy for all $t \in [0, T]$ and $x, y \in \mathbb{R}^d$ that:

$$\begin{aligned} \langle x - y, \mu_d(t, x) - \mu_d(t, y) \rangle + \eta \|\mu_d(t, x) - \mu_d(t, y)\|^2 + \frac{1+\eta}{2} \|\sigma_d(t, x) - \sigma_d(t, y)\|^2 \\ \leq \beta \|x - y\|^2 + \langle x - y, A_d(x - y) \rangle. \end{aligned} \quad (2.3)$$

(b) $\|A_d\|_{\text{op}} \leq \kappa_0 d^{\kappa_0}$, and $\langle x, A_d x \rangle \geq 0$ for all $x \in \mathbb{R}^d$.

(c) There exist constants $C_{d,0}^\mu, C_{d,1}^\mu, C_{d,0}^\sigma, C_{d,1}^\sigma \in [0, \kappa_0 d^{\kappa_0}]$ such that for all $t, s \in [0, T]$, $x, y \in \mathbb{R}^d$ it holds that

$$\begin{aligned} \|\mu_d(t, x) - \mu_d(s, y)\| &\leq C_{d,1}^\mu (\sqrt{t-s} + \|x - y\|), \quad \|\mu_d(t, 0)\| \leq C_{d,0}^\mu; \\ \|\sigma_d(t, x) - \sigma_d(s, y)\| &\leq C_{d,1}^\sigma (\sqrt{t-s} + \|x - y\|), \quad \|\sigma_d(t, 0)\| \leq C_{d,0}^\sigma. \end{aligned}$$

(d) There exists a constant $C_d^f \in [0, \kappa_0 d^{\kappa_0}]$ such that for all $x \in \mathbb{R}^d$ and $D > 0$ it holds that

$$|f_d(0)| \leq C_d^f, \quad |f_d(x) - f_{d,D}(x)| \leq C_d^f \|x\|^2 1_{B_\infty(D)^c}(x), \quad |f_{d,D}(x) - f_{d,D}(y)| \leq C_d^f D \|x - y\|,$$

where $B_\infty(D) := \{x \in \mathbb{R}^d \mid x_i \in [-D, D], \forall i\}$.

Let us briefly discuss the importance of the above assumptions. The monotonicity condition (2.3) in (H.1(a)) is weaker than the finite-dimensional analogue of the strong monotonicity condition (1.3), in the sense that (2.3) involves only the standard Euclidean norm instead of discrete Sobolev norms. The monotonicity, along with the Lipschitz continuity in (H.1(c)), ensures the well-posedness of (2.2) (see e.g. [39]), and allows us to derive precise regularity estimates (in L^p -norms for $p \in [2, 2 + \eta)$) of the solution $Y^{x,d}$ to the SDE (2.2) with respect to the coefficients and the initial condition.

It is worth emphasizing that (H.1) allows the operator norm of A_d and the Lipschitz constants of the nonlinear functions μ_d and σ_d to grow with respect to the dimension d , which is crucial for applications to stiff SDEs arising from Galerkin approximations of (controlled) SPDEs. In fact, most existing results on overcoming the curse of dimensionality with DNNs (see e.g. [12, 16, 27, 34, 28, 29]) are for value functions associated with high-dimensional SDEs whose diffusion

coefficients are affine with respect to the state variable and both drift and diffusion coefficients are Lipschitz continuous uniformly with respect to the dimensions. Note that it is easy to check that if μ_d, σ_d satisfy (H.1(c)) with a Lipschitz constant independent of the dimension d , then the coefficients satisfy (H.1(a)). In particular, our setting includes the representation result in [34] as a special case.

We remark that both the monotonicity condition (2.3) and the Lipschitz continuity of μ_d are crucial for constructing networks with polynomial complexity to approximate the desired value functions. With the help of the monotonicity condition (H.1(a)), we can demonstrate that both the regularity of the solution $Y^{x,d}$ to (2.2) and the error estimates of a corresponding partial-implicit Euler scheme depend *polynomially* on $\|A_d\|_{\text{op}}$, $[\mu_d]_1$ and $[\sigma_d]_1$, i.e., the Lipschitz constants of the coefficients (see Section 4 for details; see also [34] for SDEs with merely Lipschitz continuous coefficients, for which the corresponding estimates depend exponentially on the Lipschitz constants of the coefficients). These polynomial dependence results subsequently enable us to construct DNNs with polynomial complexities to approximate the value functions induced by stiff SDEs, including those arising from Galerkin approximations of SPDEs.

On the other hand, the Lipschitz continuity of μ_d allows us to construct the desired DNNs through a linear-implicit Euler scheme of (2.2), which is implicit in the linear part of the drift and remains explicit for the nonlinear part of the drift. In fact, to the best of our knowledge, if the function μ_d is not globally Lipschitz continuous, then one needs to adopt a fully-implicit scheme, a tamed explicit scheme or an adaptive Euler scheme to obtain a convergent approximation of (2.2) in the L^2 -norm. These schemes in general involve of nonlinear mappings that are difficult to represent by ReLU networks; in particular, the fully-implicit scheme involves of the inverse of the mapping $y \mapsto y + \Delta t(A_d y - \mu_d(t, y))$ (see e.g. [40]), the tamed explicit scheme involves of the mapping $y \mapsto \frac{\mu_d(t, y)}{1 + \Delta t \|\mu_d(t, y)\|}$ (see e.g. [26]), while the adaptive Euler scheme involves a non-uniform random stepsize which varies for different realisations of the Brownian motion and needs to be constructed in a problem-dependent way (see e.g. [13]).

Finally, instead of approximating directly f_d on \mathbb{R}^d , (H.1(d)) allows us to focus on approximating f_d on a hypercube, and then extend the approximation linearly outside the domain. This is motivated by the fact that approximating a function by neural networks on a prescribed compact set has been better understood than approximating the function globally on \mathbb{R}^d (see e.g. [44, 47, 52, 11, 5, 17, 18, 45]). In particular, since f_d can admit quadratic growth at infinity and ReLU networks can only generate piecewise linear functions, for a given small enough ε , there exists no ReLU network ϕ such that the inequality $|f_d(x) - [\mathcal{R}_\phi(\phi)](x)| \leq \varepsilon(1 + \|x\|^2)$ holds for all $x \in \mathbb{R}^d$. Therefore, we adopt a two-step approximation by first approximating f_d with a suitable Lipschitz continuous function $f_{d,D}$, and then representing $f_{d,D}$ by a ReLU network on \mathbb{R}^d with a desired accuracy; see Proposition 3.1 for the representation results for weighted square functions, which are the commonly used cost functions for PDE-constrained optimal control problems.

To construct neural networks with the desired complexities, we shall assume that the family of functions $(\mu_d, \sigma_d)_{d \in \mathbb{N}}$ and $(f_{d,D})_{d \in \mathbb{N}, D > 0}$ can be approximated by ReLU networks without curse of dimensionality.

H.2. Assume the notation of (H.1). Let $\kappa_1 \geq 0$ and $\varrho : \mathbb{R} \rightarrow \mathbb{R}$ be the function satisfying $\varrho(x) = \max(0, x)$ for all $x \in \mathbb{R}$. Let $(\phi_{\varepsilon,d}^\mu, \phi_{\varepsilon,d}^{\sigma,i}, \phi_{\varepsilon,d,D}^f)_{\varepsilon,d,D,i} \subset \mathcal{N}$, $\varepsilon \in (0, 1]$, $d \in \mathbb{N}$, $D > 0$, $i = 1, \dots, d$, be a family of DNNs with the following properties, for any given $d \in \mathbb{N}$, $\varepsilon \in (0, 1]$ and $D > 0$:

(a) The DNNs $(\phi_{\varepsilon,d}^\mu, \phi_{\varepsilon,d}^{\sigma,i})_i$ have the same architecture, i.e.,

$$\phi_{\varepsilon,d}^\mu, \phi_{\varepsilon,d}^{\sigma,i} \in \mathcal{N}_{L_{\varepsilon,d}}^{d+1, N_1^{\varepsilon,d}, \dots, N_{L_{\varepsilon,d}-1}^{\varepsilon,d}}, \quad i = 1, \dots, d,$$

for some integers $L_{\varepsilon,d}, N_1^{\varepsilon,d}, \dots, N_{L_{\varepsilon,d}-1}^{\varepsilon,d} \in \mathbb{N}$, depending on d and ε .

(b) The DNNs $(\phi_{\varepsilon,d}^\mu, \phi_{\varepsilon,d}^{\sigma,i}, \phi_{\varepsilon,d,D}^f)$ admit the following complexity estimates:

$$\mathcal{C}(\phi_{\varepsilon,d}^\mu) + \sum_{i=1}^d \mathcal{C}(\phi_{\varepsilon,d}^{\sigma,i}) \leq \kappa_1 d^{\kappa_1} \varepsilon^{-\kappa_1}, \quad \mathcal{C}(\phi_{\varepsilon,d,D}^f) \leq \kappa_1 d^{\kappa_1} D^{\kappa_1} \varepsilon^{-\kappa_1}.$$

(c) The realizations $\mu_d^\varepsilon = \mathcal{R}_\varrho(\phi_{\varepsilon,d}^\mu)$, $\sigma_d^\varepsilon = (\mathcal{R}_\varrho(\phi_{\varepsilon,d}^{\sigma,1}), \dots, \mathcal{R}_\varrho(\phi_{\varepsilon,d}^{\sigma,d}))$, and $f_{d,D}^\varepsilon = \mathcal{R}_\varrho(\phi_{\varepsilon,d,D}^f)$ admit the following approximation properties: for all $t \in [0, T]$ and $x \in \mathbb{R}^d$,

$$\|\mu_d(t, x) - \mu_d^\varepsilon(t, x)\| + \|\sigma_d(t, x) - \sigma_d^\varepsilon(t, x)\| \leq \varepsilon \kappa_1 d^{\kappa_1}, \quad |f_{d,D}(x) - f_{d,D}^\varepsilon(x)| \leq \varepsilon \kappa_1 d^{\kappa_1} D^{\kappa_1}.$$

Since a ReLU network can be extended to an arbitrary depth and width without changing its realization (Lemma A.3), we assume without loss of generality in (H.2(a)) that $(\phi_{\varepsilon,d}^\mu, \phi_{\varepsilon,d}^{\sigma,i})_{i=1,\dots,d}$ have the same architecture to simplify our analysis.

The conditions (H.2(b),(c)) imply the function $(\mu_d, \sigma_d)_{d \in \mathbb{N}}$ and $(f_{d,D})_{d \in \mathbb{N}, D > 0}$ can be approximated by ReLU networks with polynomial complexity in ε , d and D . These conditions clearly hold for most sensible discretizations of linear SPDEs, such as the Zakai equation (see e.g. [30, 15]):

$$dy(t) + Ay(t) dt = Gy(t) dB(t), \quad t \in (0, T], \quad y(0) = y_0,$$

where A and G are second-order and first-order linear differential operators, respectively. Moreover, by virtue of the fact that ReLU networks can efficiently represent the pointwise maximum/minimum operations (see Proposition 3.3), one can see (H.2(b),(c)) also hold for the discretizations of the following Hamilton–Jacobi–Bellman–Isaacs equation, since the (discretized) Hamiltonian can be *exactly* expressed by ReLU networks:

$$dy(t, x) + (-\nu(\Delta y)(t, x) - H(t, x, y(t, x), (\nabla_x y)(t, x))) dt = 0, \quad (t, x) \in (0, T] \times \mathcal{D},$$

where $\nu > 0$, \mathcal{D} is a bounded open set in \mathbb{R}^m , and the Hamiltonian $H : [0, T] \times \mathcal{D} \times \mathbb{R} \times \mathbb{R}^m \rightarrow \mathbb{R}$ is given by:

$$H(t, x, u, p) = \inf_{\alpha \in \mathbf{A}} \sup_{\beta \in \mathbf{B}} [b(t, x, \alpha, \beta)^T p + c(t, x, \alpha, \beta)u + \ell(t, x, \alpha, \beta)],$$

and \mathbf{A}, \mathbf{B} are two given finite sets. Finally, for general semilinear PDEs with bounded solutions, one may consider an equivalent semilinear PDE by truncating the nonlinearity outside a compact set, and approximate the truncated coefficients by DNNs.

Finally, we remark that (H.1(d)) and (H.2(c)) essentially assume that for any given $D > 0$, there exists a deep ReLU network approximating the terminal function $f_d|_{B_\infty(D)}$ with polynomial complexity, and the difference between the terminal function f_d and the deep ReLU network can be controlled by the quadratic growth of f_d outside the hypercube $B_\infty(D)$. We refer the reader to Proposition 3.1, where we verify (H.1(d)) and (H.2(c)) for a class of quadratic cost functions.

Now we are ready to state one of the main results of this paper, which shows that one can construct DNNs with polynomial complexity to approximate the value functions induced by nonlinear stiff SDEs. Similar representation results have been shown in [16] for SDEs with affine drift and diffusion coefficients, and in [34] for SDEs with nonlinear drift and constant diffusion coefficients. Our results extend these results to SDEs with time-inhomogeneous nonlinear drift and diffusion coefficients. Moreover, we allow the Lipschitz constants of the coefficients to grow with the dimension d , which is crucial for the application to SPDE-constrained optimal control problems. The proof of this theorem is given in Section 6.

Theorem 2.1. Suppose (H.1) and (H.2) hold. For each $d \in \mathbb{N}$, let $v_d : \mathbb{R}^d \rightarrow \mathbb{R}$ be the function satisfying for all $x \in \mathbb{R}^d$ that $v_d(x) = \mathbb{E}[f_d(Y_T^{x,d})]$ with $(Y_t^{x,d})_{t \in [0,T]}$ being the solution to (2.2), and let ν_d be a probability measure on \mathbb{R}^d satisfying $\int_{\mathbb{R}^d} \|x\|^{4+\eta} \nu_d(dx) \leq \tau d^\tau$, with the same constant η as in (H.1), and some constant $\tau > 0$ independent of d .

Then there exists a family of DNNs $(\psi_{\varepsilon,d})_{\varepsilon \in (0,1], d \in \mathbb{N}}$ and a constant $c > 0$, depending only on $\beta, \eta, \kappa_1, \kappa_2, \tau$ and T , such that for all $d \in \mathbb{N}$, $\varepsilon \in (0, 1]$, we have $\mathcal{C}(\psi_{\varepsilon,d}) \leq cd^c \varepsilon^{-c}$, $\mathcal{R}_\varrho(\psi_{\varepsilon,d}^\varepsilon) \in C(\mathbb{R}^d; \mathbb{R})$ and

$$\left(\int_{\mathbb{R}^d} |v_d(x) - [\mathcal{R}_\varrho(\psi_{\varepsilon,d})](x)|^2 \nu_d(dx) \right)^{1/2} < \varepsilon.$$

The following result is a direct consequence of Theorem 2.1 and the Feynman-Kac formula in [46, Theorem 2.2], which shows one can approximate the viscosity solution to a Kolmogorov backward PDE with stiff coefficients on a bounded domain without curse of dimensionality. The proof will be postponed to Section 6.

Corollary 2.2. Suppose (H.1) and (H.2) hold. For each $d \in \mathbb{N}$, let u_d be the unique continuous viscosity solution to the following PDE with at most quadratic growth at infinity:

$$\frac{\partial u_d}{\partial t}(t, x) + \frac{1}{2} \text{tr}(\sigma_d(t, x) \sigma_d^T(t, x) (\text{Hess}_x u_d)(t, x)) + (-A_d x + \mu_d(t, x))^T (\nabla_x u_d)(t, x) = 0 \quad (2.4)$$

for all $(t, x) \in [0, T] \times \mathbb{R}^d$, and $u_d(T, x) = f_d(x)$ for $x \in \mathbb{R}^d$. Then there exists a family of DNNs $(\psi_{\varepsilon,d})_{\varepsilon \in (0,1], d \in \mathbb{N}}$ and a constant $c > 0$, depending only on $\beta, \eta, \kappa_1, \kappa_2$ and T , such that for all $d \in \mathbb{N}$, $\varepsilon \in (0, 1]$, we have $\mathcal{C}(\psi_{\varepsilon,d}) \leq cd^c \varepsilon^{-c}$, $\mathcal{R}_\varrho(\psi_{\varepsilon,d}^\varepsilon) \in C(\mathbb{R}^d; \mathbb{R})$ and

$$\left(\int_{[0,1]^d} |u_d(0, x) - [\mathcal{R}_\varrho(\psi_{\varepsilon,d})](x)|^2 dx \right)^{1/2} < \varepsilon.$$

2.2 Expression rate for controlled SDEs with stiff coefficients

In this section, we extend the expression rates in Section 2.1, and construct DNNs with polynomial complexity to approximate value functions associated with a sequence of controlled SDEs with stiff coefficients.

We start by introducing the set of admissible strategies. Let $M \in \mathbb{N}$ and \mathcal{T}_M be the set of intervention times defined as:

$$\mathcal{T}_M = \{\bar{t}_k \in [0, T] \mid \bar{t}_k = kT/M, k = 0, \dots, M\}. \quad (2.5)$$

For each $d \in \mathbb{N}$, we consider the following piecewise-constant, deterministic strategies: for $i = 1, 2$,

$$u_i \in \mathcal{U}_{i,d} := \{u_i : [0, T] \rightarrow U_{i,d} \mid u_i(t) = u_i(\bar{t}_k) \in U_{i,d}, \forall t \in [\bar{t}_k, \bar{t}_{k+1}), k = 0, \dots, M-1\}, \quad (2.6)$$

where $U_{i,d}$, $i = 1, 2$, are given nonempty finite subsets of \mathbb{R}^{m_d} for some $m_d \in \mathbb{N}$. Note that u_i can be the coefficients of a parameterized control policy in the sense that if $u_i(t) = (u_{i,j}(\bar{t}_k))_{j=1}^{m_d}$ on $[t_k, t_{k+1})$, the state equation is controlled by a policy $\tilde{u}_i(t, x) = \sum_{j=1}^{m_d} u_{i,j}(\bar{t}_k) e_j(t, x)$ on $[t_k, t_{k+1})$, where $\{e_j(t, x)\}_{j=1}^{m_d}$ are some prescribed basis functions.

Now for each $d \in \mathbb{N}$, we consider a two-player zero-sum stochastic differential game, where the “inf-player” aims to minimize a particular function over all strategies $u_1 \in \mathcal{U}_{1,d}$, while the “sup-player” aims to maximize it over all strategies $u_2 \in \mathcal{U}_{2,d}$. The value function $v_d : \mathbb{R}^d \rightarrow \mathbb{R}$ is given by:

$$v_d(x) := \inf_{u_1 \in \mathcal{U}_{1,d}} \sup_{u_2 \in \mathcal{U}_{2,d}} \mathbb{E} \left[f_d(Y_T^{x,d,u_1,u_2}) + g_d(u_1, u_2) \right], \quad x \in \mathbb{R}^d, \quad (2.7)$$

where for each $x \in \mathbb{R}^d$, $u_i \in \mathcal{U}_{i,d}$ (see (2.6)), $i = 1, 2$, $Y^{x,d,u_1,u_2} = (Y_t^{x,d,u_1,u_2})_{t \in [0,T]}$ is the strong solution to the following d -dimensional controlled SDE:

$$dY_t = (-A_d Y_t + \mu_d(t, Y_t, u_1, u_2)) dt + \sigma_d(t, Y_t, u_1, u_2) dB_t, \quad t \in (0, T]; \quad Y_0 = x, \quad (2.8)$$

with a d -dimensional Brownian motion $(B_t)_{t \in [0,T]}$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For simplicity, here we do not take into account any running costs of the state process Y^{x,d,u_1,u_2} , but it is straightforward to extend our results to control problems with running costs. Moreover, the result would not change if the linear part of the drift is also controlled.

Note that, if the coefficients μ_d and σ_d of (2.8) are independent of the control parameters and the cost function $g_d : \mathcal{U}_{1,d} \times \mathcal{U}_{1,d} \rightarrow \mathbb{R}$ in (2.7) is the zero function, then the value function defined in (2.7) reduces to the function defined in (2.1).

We then state the assumptions on the coefficients of (2.8) for deriving the expression rates of DNNs. Roughly speaking, we assume (H.1) and (H.2) hold uniformly in terms of the control parameters. However, we would like to point out that even though the functions μ_d, σ_d are continuous in time, the controlled drift and diffusion of (2.8) are discontinuous in time due to the jumps in the control processes.

H.3. Let $\beta, \kappa_0, T \geq 0$, $\eta > 0$ and $M \in \mathbb{N}$ be fixed constants. Let the set \mathcal{T}_M be defined as in (2.5). For all $d \in \mathbb{N}$ and $D > 0$, let $A_d \in \mathbb{R}^{d \times d}$, $U_d = U_{1,d} \times U_{2,d}$ be a *nonempty* subset of \mathbb{R}^{m_d} for some $m_d \in \mathbb{N}$, and $\mu_d : [0, T] \times \mathbb{R}^d \times U_d \rightarrow \mathbb{R}^d$, $\sigma_d : [0, T] \times \mathbb{R}^d \times U_d \rightarrow \mathbb{R}^{d \times d}$, $f_d, f_{d,D} : \mathbb{R}^d \rightarrow \mathbb{R}$, $g_d : U_d \rightarrow \mathbb{R}$ be measurable functions with the following properties:

- (a) For all $u \in U_d$, the matrix A_d and the functions $\mu_d(\cdot, \cdot, u) : [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}^d$, $\sigma_d(\cdot, \cdot, u) : [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d}$ satisfy (H.1(a),(b),(c)) with the constants β, κ_0, η .
- (b) The functions f_d and $f_{d,D}$ satisfy (H.1(d)).
- (c) The cardinality of the set U_d satisfies $|U_d| \leq \kappa_0 d^{\kappa_0}$.

H.4. Assume the notation of (H.3). Let $\kappa_1 \geq 0$ and $\varrho : \mathbb{R} \rightarrow \mathbb{R}$ be the function satisfying $\varrho(x) = \max(0, x)$ for all $x \in \mathbb{R}$. Let $(\phi_{\varepsilon,d}^\mu, \phi_{\varepsilon,d}^{\sigma,i}, \phi_{\varepsilon,d,D}^f)_{\varepsilon,d,D,i} \subset \mathcal{N}$, $\varepsilon \in (0, 1]$, $d \in \mathbb{N}$, $D > 0$, $i = 1, \dots, d$, be a family of DNNs with the following properties, for any given $d \in \mathbb{N}$, $\varepsilon \in (0, 1]$ and $D > 0$:

- (a) The DNNs $(\phi_{\varepsilon,d}^\mu, \phi_{\varepsilon,d}^{\sigma,i})_i$ have the same architecture with the input dimension $d + m_d + 1$.
- (b) The complexities of the DNNs $(\phi_{\varepsilon,d}^\mu, \phi_{\varepsilon,d}^{\sigma,i}, \phi_{\varepsilon,d,D}^f)$ satisfy (H.2(b)) with the constant κ_1 .
- (c) The realizations $\mu_d^\varepsilon = \mathcal{R}_\varrho(\phi_{\varepsilon,d}^\mu)$, $\sigma_d^\varepsilon = (\mathcal{R}_\varrho(\phi_{\varepsilon,d}^{\sigma,1}), \dots, \mathcal{R}_\varrho(\phi_{\varepsilon,d}^{\sigma,d}))$, and $f_{d,D}^\varepsilon = \mathcal{R}_\varrho(\phi_{\varepsilon,d,D}^f)$ admit the following approximation properties: for all $t \in [0, T]$, $x \in \mathbb{R}^d$ and $u \in U_d$,

$$\|\mu_d(t, x, u) - \mu_d^\varepsilon(t, x, u)\| + \|\sigma_d(t, x, u) - \sigma_d^\varepsilon(t, x, u)\| \leq \varepsilon \kappa_1 d^{\kappa_1}, \quad |f_{d,D}(x) - f_{d,D}^\varepsilon(x)| \leq \varepsilon \kappa_1 d^{\kappa_1} D^{\kappa_1}.$$

We remark that (H.3(a),(c)) are natural assumptions if the controlled stiff SDEs (2.8) arise from a discretization of controlled SPDEs. For example, in the deterministic setting, one can consider the following optimal control problems within a spatial domain $\Omega \subset \mathbb{R}^p$:

$$\mathcal{V}(y_0) = \min_{u \in \mathcal{U}} \mathcal{J}(u), \quad \mathcal{J}(u) := \|y(T) - \bar{y}\|_{L^2(\Omega)}^2 + \|u\|_{\mathbb{R}^m}^2, \quad (2.9)$$

where the set of admissible controls \mathcal{U} is a compact subset of \mathbb{R}^m , $\bar{y} \in L^2(\Omega)$ is the desired terminal state, and y is governed by a controlled semilinear parabolic PDE with initial condition $y_0 \in L^2(\Omega)$:

$$\frac{\partial}{\partial t}y - \Delta y + G(y) = \sum_{i=1}^m u_i e_i(t, x) \text{ in } \Omega \times (0, T), \quad y = 0 \text{ on } \partial\Omega \times (0, T), \quad y = y_0 \text{ in } \Omega \times \{t = 0\}$$

with given Lipschitz function G and sufficiently regular basis functions $\{e_i\}_{i=1}^M$. Note that finitely many control parameters appear frequently in practical applications of optimal control theory, since it is difficult to implement control strategies that vary arbitrarily in time and space; see e.g. [41, 42] for elliptic optimal control problems with finite dimensional control spaces. Then it is clear that the discrete version of (2.9) (in both the space and control variables) is a special case of the zero-sum game (2.7) whose coefficients satisfy (H.3) (with $M = 1$ in (2.6)).

Now we are ready to present the main theorem in this section, which shows one can represent the value function (2.7) by DNNs without curse of dimensionality, whose proof will be deferred to Section 7.

Note that here the value function (2.7) is induced by optimizing the cost functional over deterministic control strategies (i.e., open-loop controls). For general stochastic games with adapted stochastic strategies (i.e., closed-loop controls), the value function can be identified as the solution of a d -dimensional fully-nonlinear HJBI equation (see e.g. [51]), for which the analysis of DNN approximation rates is more involved (see [27, 28, 29] for some results on overcoming the curse of dimensionality with DNNs for some semilinear PDEs).

Theorem 2.3. *Suppose (H.3) and (H.4) hold. For each $d \in \mathbb{N}$, let $v_d : \mathbb{R}^d \rightarrow \mathbb{R}$ be the value function defined in (2.7), and let ν_d be a probability measure on \mathbb{R}^d satisfying $\int_{\mathbb{R}^d} \|x\|^{4+\eta} \nu_d(dx) \leq \tau d^\tau$, with the same constant η as in (H.1), and some constant $\tau > 0$ independent of d .*

Then there exists a family of DNNs $(\psi_{\varepsilon,d})_{\varepsilon \in (0,1], d \in \mathbb{N}}$ and a constant $c > 0$, depending only on $\beta, \eta, \kappa_1, \kappa_2, \tau, M$ and T , such that for all $d \in \mathbb{N}$, $\varepsilon \in (0, 1]$, we have $\mathcal{C}(\psi_{\varepsilon,d}) \leq cd^c \varepsilon^{-c}$, $\mathcal{R}_\varrho(\psi_{\varepsilon,d}^\varepsilon) \in C(\mathbb{R}^d; \mathbb{R})$ and

$$\left(\int_{\mathbb{R}^d} |v_d(x) - [\mathcal{R}_\varrho(\psi_{\varepsilon,d})](x)|^2 \nu_d(dx) \right)^{1/2} < \varepsilon.$$

3 ReLU network calculus

In this section, we shall discuss several basic operations to construct new DNNs from existing ones. We shall also establish some fundamental results on the representation flexibility of DNNs by following the setting of Definition 2.1, which are essential for our subsequent analysis.

Recall that it has been shown in [12, 16] that linear combination and composition of a finite number of ReLU DNNs can be realized by a ReLU DNN with polynomial complexity. Moreover, the identity function can be implemented as a ReLU network with one hidden layer. The precise statements of these results will be given in Appendix A for completeness.

The following proposition presents a *global* approximation result for the weighted square function on \mathbb{R}^d . Since the quadratic growth of the weighted square function prevents us to directly approximate it by a ReLU neural network, we shall employ a two-step approximation by first approximating the function on a prescribed compact set and then linearly extending it outside the bounded domain.

Proposition 3.1. *Let $\varrho : \mathbb{R} \rightarrow \mathbb{R}$ be the activation function defined as $\varrho(x) = \max(0, x)$ for all $x \in \mathbb{R}$. Let $d \in \mathbb{N}$, $\beta = (\beta_m)_{m=1}^d \in \mathbb{R}$, and $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be the weighted square function defined*

by $f(x) := \sum_{m=1}^d \beta_m x_m^2$ for all $x = (x_1, \dots, x_d)^T \in \mathbb{R}^d$. Then for any $D > 0, \varepsilon \in (0, 1/2)$, there exists a function $f_{d,D} : \mathbb{R}^d \rightarrow \mathbb{R}$ and a DNN $\phi_{\varepsilon,d,D}$ satisfying

$$\begin{aligned} |f(x) - f_{d,D}(x)| &\leq \|\beta\|_\infty \|x\|^2 1_{B_\infty(D)^c}(x), & |f_{d,D}(x) - f_{d,D}(y)| &\leq 2\|\beta\|_\infty d^{1/2} D \|x - y\|, \\ |f_{d,D}(x) - [\mathcal{R}_\varrho(\phi_{d,\varepsilon,D})](x)| &\leq \|\beta\|_\infty d D^2 \varepsilon, & \mathcal{C}(\phi_{d,\varepsilon,D}) &\leq C d^2 \log(\varepsilon^{-1}) + d + 1, \end{aligned}$$

where $B_\infty(D) := \{x \in \mathbb{R}^d \mid |x_m| < D, \forall m\}$, $\|\beta\|_\infty = \sup_m |\beta_m|$, and C is a constant independent of β, D, d and ε .

Proof. We start by constructing a *global* approximation of the (one-dimensional) squaring operation $x \mapsto x^2$. Since this construction is similar to that in [17, Proposition III.2], we only repeat the main steps for reader's convenience. Recall that it has been shown in [17, Proposition III.1] (see also [52]) that for any $\varepsilon \in (0, 1/2)$, one can use the “tooth” (or “mirror”) function $g : \mathbb{R} \rightarrow [0, 1]$:

$$g(x) = \begin{cases} 2x, & 0 \leq x < 1/2; \\ 2(1-x), & 1/2 \leq x < 1; \\ 0, & \text{otherwise,} \end{cases}$$

and its s -fold composition (i.e., the “sawtooth” function) to construct a DNN $\phi_{1,\varepsilon}$ satisfying $\mathcal{L}(\phi_{1,\varepsilon}) \leq C \log(\varepsilon^{-1})$, $\mathcal{C}(\phi_{1,\varepsilon}) \leq C \log(\varepsilon^{-1})$, $[\mathcal{R}_\varrho(\phi_{1,\varepsilon})](0) = 0$, $[\mathcal{R}_\varrho(\phi_{1,\varepsilon})](x) = x$ for $x \notin [0, 1]$, and $||[\mathcal{R}_\varrho(\phi_{1,\varepsilon})](x) - x^2|_{L^\infty[0,1]} \leq \varepsilon$ for some constant C independent of ε . Then for any given $D > 0$, we shall consider the function $f_{\varepsilon,1,D} : x \mapsto D^2 [\mathcal{R}_\varrho(\phi_{1,\varepsilon})](|x|/D)$, $x \in \mathbb{R}$, and approximate the square function by the following function:

$$f_{1,D}(x) = \begin{cases} x^2, & |x| \leq D; \\ D|x|, & \text{otherwise.} \end{cases} \quad (3.1)$$

It follows directly from the properties of $\phi_{1,\varepsilon}$ that $f_{\varepsilon,1,D}(x) = f_{1,D}(x)$ for $|x| > D$ and $|f_{\varepsilon,1,D}(x) - f_{1,D}(x)|_{L^\infty[-D,D]} \leq D^2 \varepsilon$. Since $f_{\varepsilon,1,D}$ is a composition of the functions $x \mapsto [\mathcal{R}_\varrho(\phi_{1,\varepsilon})](x)$ and $x \rightarrow |x| = \varrho(x) + \varrho(-x)$, we know it is the realization of a ReLU network $\phi_{\varepsilon,1,D}$ with complexity $\mathcal{C}(\phi_{\varepsilon,1,D}) \leq C \log(\varepsilon^{-1})$, for some constant C independent of ε and D .

Now for any given $(\beta_m)_{m=1}^d \in \mathbb{R}$, we consider the functions $f_{d,D}(x) = \sum_{m=1}^d \beta_m f_{1,D}(x_m)$ and $f_{\varepsilon,d,D}(x) = \sum_{m=1}^d \beta_m f_{\varepsilon,1,D}(x_m)$ for all $x = (x_1, \dots, x_d)^T \in \mathbb{R}^d$. It is straightforward to verify that it holds for all $x \in B_\infty(D)$ that $f_{d,D}(x) - f(x) = 0$, and for all $x \in B_\infty(D)^c$ that

$$|f_{d,D}(x) - f(x)| \leq \sum_{m=1}^d |\beta_m| |f_{1,D}(x_m) - x_m^2| \leq \sum_{m=1}^d |\beta_m| (x_m^2 - D|x_m|) 1_{\{|x_m| > D\}} \leq \|\beta\|_\infty \|x\|^2,$$

where we denote $\|\beta\|_\infty = \sup_m |\beta_m|$. Also $f_{d,D}$ is Lipschitz continuous, i.e., for all $x, y \in \mathbb{R}^d$,

$$\begin{aligned} |f_{d,D}(x) - f_{d,D}(y)| &\leq \sum_{m=1}^d |\beta_m| |f_{1,D}(x_m) - f_{1,D}(y_m)| \\ &\leq 2D \|\beta\|_\infty \sum_{m=1}^d |x_m - y_m| \leq 2D \|\beta\|_\infty d^{1/2} \|x - y\|. \end{aligned}$$

Moreover, the following approximation property holds:

$$\begin{aligned} |f_{d,D}(x) - f_{\varepsilon,d,D}(x)| &\leq \sum_{m=1}^d |\beta_m| |f_{1,D}(x_m) - f_{\varepsilon,1,D}(x_m)| \\ &= \sum_{m=1}^d |\beta_m| |f_{1,D}(x_m) - f_{\varepsilon,1,D}(x_m)| 1_{\{|x_m| \leq D\}} \leq \|\beta\|_{\infty} d D^2 \varepsilon. \end{aligned}$$

Therefore, it remains to show $f_{\varepsilon,d,D}$ is the realization of a ReLU network and estimate its complexity. The main tool to construct the desired ReLU network is a “parallelization” of the network $\phi_{1,\varepsilon,D}$ (see [12]). Suppose that the network $\phi_{1,\varepsilon,D}$ is given by $\phi_{1,\varepsilon,D} = ((W_1, b_1), (W_2, b_2), \dots, (W_L, b_L))$, and the dimension is $\dim(\phi_{1,\varepsilon,D}) = (N_0, N_1, \dots, N_{L-1}, N_L)$. Then we consider the DNN $\phi_{d,\varepsilon,D} = ((W'_1, b'_1), (W'_2, b'_2), \dots, (W'_{L+1}, b'_{L+1}))$, where we have for all $i = 1, \dots, L$,

$$W_i = \begin{pmatrix} W_i & & & \\ & W_i & & \\ & & \ddots & \\ & & & W_i \end{pmatrix} \in \mathbb{R}^{(dN_i) \times (dN_{i-1})}, \quad b'_i = \begin{pmatrix} b_i \\ b_i \\ \vdots \\ b_i \end{pmatrix},$$

and $W'_{L+1} = (\beta_1, \dots, \beta_d) \in \mathbb{R}^{1 \times dN_L}$ and $b'_{L+1} = 0$. Then it is easy to see that $f_{\varepsilon,d,D}(x) = [\mathcal{R}_{\varrho}(\phi_{d,\varepsilon,D})](x)$ for all $x \in \mathbb{R}^d$, and the complexity of $\phi_{d,\varepsilon,D}$ is given by

$$\mathcal{C}(\phi_{d,\varepsilon,D}) = \sum_{l=1}^L (dN_l)(dN_{l-1} + 1) + (d+1) \leq d^2 \left(\sum_{l=1}^L N_l(N_{l-1} + 1) \right) + d+1 \leq C d^2 \log(\varepsilon^{-1}) + d+1,$$

some constant C independent of ε, d and D . \square

The next proposition presents an operation involving linear combination and compositions of networks with different architectures, which extends Proposition 5.2 in [16] for two networks with the same input-output dimensions (i.e., $M = 2$ and $d' = 0$) to multiple networks with different input-output dimensions (i.e., $M \geq 2$ and $d' \in \mathbb{N}$). Such an extension is essential for the subsequent analysis of controlled SDEs with time-inhomogeneous coefficients and a nonlinear diffusion term.

Proposition 3.2. *Let $\varrho : \mathbb{R} \rightarrow \mathbb{R}$ be the activation function defined as $\varrho(x) = \max(0, x)$ for all $x \in \mathbb{R}$. Let $M, L, L' \in \mathbb{N}$, $d \in \mathbb{N}$, $d' \in \mathbb{N} \cup \{0\}$, $u \in \mathbb{R}^{d'}$, and $(\phi_m)_{m=1}^M \in \mathcal{N}$ be DNNs such that*

$$\begin{aligned} \mathcal{L}(\phi_1) &= L, \quad \dim(\phi_1) = (d, N_1^{(1)}, N_2^{(1)}, \dots, N_{L-1}^{(1)}, d), \\ \mathcal{L}(\phi_m) &= L', \quad \dim(\phi_m) = (d + d', N_1^{(m)}, N_2^{(m)}, \dots, N_{L'-1}^{(m)}, d), \quad m \in \{2, \dots, M\}, \end{aligned}$$

for some $N_l^{(1)}, N_{l'}^{(m)} \in \mathbb{N}$, $l = 1, \dots, L-1$, $l' = 1, \dots, L'-1$, $m = 2, \dots, M$.

Then there exists a DNN $\psi \in \mathcal{N}$ such that the depth $\mathcal{L}(\psi) = L + L' - 1$, the dimension

$$\dim(\psi) = \begin{cases} (d, N_1^{(1)}, N_2^{(1)}, \dots, N_{L-1}^{(1)}, d) & L' = 1, \\ (d, N_1^{(1)}, N_2^{(1)}, \dots, N_{L-1}^{(1)}, 2d + \sum_{m=2}^M N_1^{(m)}, \dots, 2d + \sum_{m=2}^M N_{L'-1}^{(m)}, d), & L' \geq 2, \end{cases} \quad (3.2)$$

and satisfies the following identity:

$$[\mathcal{R}_{\varrho}(\psi)](x) = [\mathcal{R}_{\varrho}(\phi_1)](x) + \sum_{m=2}^M [\mathcal{R}_{\varrho}(\phi_m)]([\mathcal{R}_{\varrho}(\phi_1)](x), u), \quad x \in \mathbb{R}^d. \quad (3.3)$$

Assume in addition for the case $L' \geq 2$ that, $N_{L-1}^{(1)} \leq 2d + \sum_{m=2}^M N_{L'-1}^{(m)}$, and there exists $m_0 \in \{2, \dots, M\}$ such that we have for all $i \in \{2, \dots, L' - 1\}$, $m \in \{2, \dots, M\}$ that $N_i^{(m)} \leq N_i^{(m_0)}$. Then it holds that

$$\mathcal{C}(\psi) \leq \mathcal{C}(\phi_1) + (M-1)^2 \left(\sup_{m \in \{2, \dots, M\}} \mathcal{C}(\phi_m) + \mathcal{C}(\phi_{d,2}^{\text{Id}}) \right)^3, \quad (3.4)$$

where $\phi_{d,2}^{\text{Id}}$ is the two-layer representation of the d -dimensional identity function defined as in (A.1).

Proof. We shall assume the networks $(\phi_m)_{m=1}^M$ are given as follows, and construct the desired network ψ differently based on whether $L' = 1$ or $L' \geq 2$:

$$\begin{aligned} \phi_1 &= ((W_1^{(1)}, b_1^{(1)}), (W_2^{(1)}, b_2^{(1)}), \dots, (W_L^{(1)}, b_L^{(1)})) \in \mathcal{N}_L^{d, N_1^{(1)}, N_2^{(1)}, \dots, d}, \\ \phi_m &= ((W_1^{(m)}, b_1^{(m)}), (W_2^{(m)}, b_2^{(m)}), \dots, (W_{L'}^{(m)}, b_{L'}^{(m)})) \in \mathcal{N}_{L'}^{d+d', N_1^{(m)}, N_2^{(m)}, \dots, d}, \quad m \in \{2, \dots, M\}. \end{aligned}$$

Suppose $L' = 1$, we shall consider the DNN $\psi = ((W_1, b_1), (W_2, b_2), \dots, (W_L, b_L))$, where $(W_i, b_i) = (W_i^{(1)}, b_i^{(1)})$ for $i \in \{1, \dots, L-1\}$ and

$$W_L = W_L^{(1)} + \sum_{m=2}^M W_L^{(m)} \begin{pmatrix} W_L^{(1)} \\ \mathbf{0} \end{pmatrix} \in \mathbb{R}^{d \times N_{L-1}^{(1)}}, \quad b_L = b_L^{(1)} + \sum_{m=2}^M \left(W_L^{(m)} \begin{pmatrix} b_L^{(1)} \\ u \end{pmatrix} + b_L^{(m)} \right),$$

with $\mathbf{0} \in \mathbb{R}^{d' \times N_{L-1}^{(1)}}$. Then we have for all $x \in \mathbb{R}^{N_{L-1}^{(1)}}$ that

$$W_L x + b_L = W_L^{(1)} x + b_L^{(1)} + \sum_{m=2}^M \left(W_L^{(m)} \begin{pmatrix} W_L^{(1)} x + b_L^{(1)} \\ u \end{pmatrix} + b_L^{(m)} \right),$$

which implies (3.3). Also it is clear that $\mathcal{C}(\psi) = \mathcal{C}(\phi_1)$.

Now let $L' \geq 2$, $\phi_{d,2}^{\text{Id}} = ((W_1^{\text{Id}}, 0), (W_2^{\text{Id}}, 0)) \in \mathcal{N}_2^{d, 2d, d}$ be the DNN representation of the d -dimensional identity function defined as in (A.1). We shall construct the desired DNN as follows:

$$\psi = ((W_1, b_1), (W_2, b_2), \dots, (W_{L+L'-1}, b_{L+L'-1})),$$

where for $i \in \{1, \dots, L-1\}$, we have $(W_i, b_i) = (W_i^{(1)}, b_i^{(1)})$; for $i = L$, we have

$$W_i = \begin{pmatrix} W_1^{\text{Id}} W_L^{(1)} \\ W_1^{(2)} \begin{pmatrix} W_L^{(1)} \\ \mathbf{0} \end{pmatrix} \\ \vdots \\ W_1^{(M)} \begin{pmatrix} W_L^{(1)} \\ \mathbf{0} \end{pmatrix} \end{pmatrix} \in \mathbb{R}^{(2d + \sum_{m=2}^M N_1^{(m)}) \times N_{L-1}^{(1)}}, \quad b_i = \begin{pmatrix} W_1^{\text{Id}} b_L^{(1)} \\ W_1^{(2)} \begin{pmatrix} b_L^{(1)} \\ u \end{pmatrix} + b_1^{(2)} \\ \vdots \\ W_1^{(M)} \begin{pmatrix} b_L^{(1)} \\ u \end{pmatrix} + b_1^{(M)} \end{pmatrix},$$

with $\mathbf{0} \in \mathbb{R}^{d' \times N_{L-1}^{(1)}}$; for $i \in \{L+1, \dots, L+L'-2\}$, we have

$$W_i = \begin{pmatrix} W_1^{\text{Id}} W_2^{\text{Id}} & & & \\ & W_{i-L+1}^{(2)} & & \\ & & \ddots & \\ & & & W_{i-L+1}^{(M)} \end{pmatrix} \in \mathbb{R}^{(2d + \sum_{m=2}^M N_{i-L+1}^{(m)}) \times (2d + \sum_{m=2}^M N_{i-L}^{(m)})}, \quad b_i = \begin{pmatrix} 0 \\ b_{i-L+1}^{(2)} \\ \vdots \\ b_{i-L+1}^{(M)} \end{pmatrix};$$

and for $i = L + L' - 1$ we have

$$W_i = \begin{pmatrix} W_2^{\text{Id}} & W_{L'}^{(2)} & \dots & W_{L'}^{(M)} \end{pmatrix} \in \mathbb{R}^{d \times (2d + \sum_{m=2}^M N_{L'-1}^{(m)})}, \quad b_i = \sum_{m=2}^M b_{L'}^{(m)}.$$

Note that we have for all $x \in \mathbb{R}^{N_{L-1}^{(1)}}$ that

$$W_L x + b_L = \begin{pmatrix} W_1^{\text{Id}}(W_L^{(1)}x + b_L^{(1)}) \\ W_1^{(2)} \begin{pmatrix} W_L^{(1)}x + b_L^{(1)} \\ u \end{pmatrix} + b_1^{(2)} \\ \vdots \\ W_1^{(M)} \begin{pmatrix} W_L^{(1)}x + b_L^{(1)} \\ u \end{pmatrix} + b_1^{(M)} \end{pmatrix},$$

for all $i \in \{L + 1, \dots, L + L' - 2\}$, $x \in \mathbb{R}^{2d}$, $y^{(m)} \in \mathbb{R}^{N_{i-L}^{(m)}}$, $m \in \{2, \dots, M\}$ that

$$W_i \begin{pmatrix} x \\ y^{(2)} \\ \vdots \\ y^{(M)} \end{pmatrix} + b_i = \begin{pmatrix} W_1^{\text{Id}}(W_2^{\text{Id}}x) \\ W_{i-L+1}^{(2)}y^{(2)} + b_{i-L+1}^{(2)} \\ \vdots \\ W_{i-L+1}^{(M)}y^{(M)} + b_{i-L+1}^{(M)} \end{pmatrix},$$

and for all $i = L + L' - 1$, $x \in \mathbb{R}^{2d}$, $y^{(m)} \in \mathbb{R}^{N_{i-L}^{(m)}}$, $m \in \{2, \dots, M\}$ that

$$W_i \begin{pmatrix} x \\ y^{(2)} \\ \vdots \\ y^{(M)} \end{pmatrix} + b_i = W_2^{\text{Id}}x + \sum_{m=2}^M (W_{L'}^{(m)}y^{(m)} + b_{L'}^{(m)}).$$

The above identities together with the fact that the DNN $\phi_{d,2}^{\text{Id}}$ represents the d -dimensional identity function, i.e., $W_2^{\text{Id}}\varrho^*(W_1^{\text{Id}}x) = x$ for all $x \in \mathbb{R}^d$, enable us to conclude (3.3) by induction on i .

Now we turn to estimate the complexity of ψ , which is given by

$$\begin{aligned} \mathcal{C}(\psi) &= \sum_{i=1}^{L-1} N_i^{(1)}(N_{i-1}^{(1)} + 1) + (2d + \sum_{m=2}^M N_1^{(m)})(N_{L-1}^{(1)} + 1) \\ &\quad + \sum_{i=L+1}^{L+L'-2} \left(2d + \sum_{m=2}^M N_{i-L+1}^{(m)} \right) \left(2d + \sum_{m=2}^M N_{i-L}^{(m)} + 1 \right) + d(2d + \sum_{m=2}^M N_{L'-1}^{(m)} + 1) \\ &= \sum_{i=1}^{L-1} N_i^{(1)}(N_{i-1}^{(1)} + 1) + (\mathbf{i} + \sum_{m=2}^M N_1^{(m)})(N_{L-1}^{(1)} + 1) \\ &\quad + \sum_{i=1}^{L'-2} \left(\mathbf{i} + \sum_{m=2}^M N_{i+1}^{(m)} \right) \left(\mathbf{i} + \sum_{m=2}^M N_i^{(m)} + 1 \right) + d(\mathbf{i} + \sum_{m=2}^M N_{L'-1}^{(m)} + 1), \end{aligned}$$

where we denote $\mathbf{i} = 2d$. Now using the assumptions that $N_{L-1}^{(1)} \leq \mathbf{i} + \sum_{m=2}^M N_{L'-1}^{(m)}$, and $N_i^{(m)} \leq N_i^{(m_0)}$ for all $i \in \{2, \dots, L' - 1\}$, $m \in \{2, \dots, M\}$, we have

$$\begin{aligned} \mathcal{C}(\psi) &\leq \sum_{i=1}^{L-1} N_i^{(1)} (N_{i-1}^{(1)} + 1) + (\mathbf{i} + \sum_{m=2}^M N_1^{(m)}) (\mathbf{i} + \sum_{m=2}^M N_{L'-1}^{(m)} + 1) \\ &\quad + \sum_{i=1}^{L'-2} \left(\mathbf{i} + \sum_{m=2}^M N_{i+1}^{(m)} \right) \left(\mathbf{i} + \sum_{m=2}^M N_i^{(m)} + 1 \right) + d(\mathbf{i} + \sum_{m=2}^M N_{L'-1}^{(m)} + 1) \\ &\leq \mathcal{C}(\phi_1) + (M-1)^2 \left[(\mathbf{i} + N_1^{(m_0)}) (\mathbf{i} + N_{L'-1}^{(m_0)} + 1) \right. \\ &\quad \left. + \sum_{i=1}^{L'-2} \left(\mathbf{i} + N_{i+1}^{(m_0)} \right) \left(\mathbf{i} + N_i^{(m_0)} + 1 \right) + N_{L'}^{(m_0)} (\mathbf{i} + N_{L'-1}^{(m_0)} + 1) \right]. \end{aligned}$$

Then from the same arguments as [16, Proposition 5.3] (c.f. equation (124) in [16]), we can bound the terms in the square bracket and deduce that

$$\mathcal{C}(\psi) \leq \mathcal{C}(\phi_1) + (M-1)^2 [(\mathcal{C}(\phi_{m_0}) + \mathcal{C}(\phi_{d,2}^{\text{Id}}))^3],$$

which leads to the complexity estimate (3.4) by using $\mathcal{C}(\phi_{m_0}) \leq \sup_{m \in \{2, \dots, M\}} \mathcal{C}(\phi_m)$. \square

Remark 3.1. Note that the complexity of the resulting network ψ is additive to that of the network ϕ_1 . Moreover, for fixed networks $(\phi_m)_{m=2}^M$, if we start with a network ϕ_1 whose the last hidden layer's dimension satisfies $N_{L-1}^{(1)} \leq 2d + \sum_{m=2}^M N_{L'-1}^{(m)}$, our construction ensures that the dimension of the last hidden layer of the resulting network ψ also enjoys the same property. These two important observations enable us to iteratively apply Proposition 3.2, and construct a network with desired complexity in Sections 6 and 7.

We end this section with the fact that taking pointwise maximum or minimum preserves the property of being represented by a ReLU DNN. One can find similar results in [1, Lemma A.3], where the authors adopt a different notation of neural network by allowing connections between nodes in non-consecutive layers.

Proposition 3.3. *Let $\varrho : \mathbb{R} \rightarrow \mathbb{R}$ be the activation function defined as $\varrho(x) = \max(0, x)$ for all $x \in \mathbb{R}$. Let $(\phi_m)_{m=1}^\infty \in \mathcal{N}$ by a family of DNNs with the same architecture, i.e., $\phi_m \in \mathcal{N}_L^{N_0, N_1, \dots, N_{L-1}, N_L}$ for all m , where $L, N_1, \dots, N_{L-1} \in \mathbb{N}$, $N_0 = d$ and $N_L = 1$. Then for each $n \in \mathbb{N} \cup \{0\}$, there exists a DNN $\psi_n \in \mathcal{N}$ such that*

$$[\mathcal{R}_\varrho(\psi_n)](x) = \max_{m=1, \dots, 2^n} [\mathcal{R}_\varrho(\phi_m)](x), \quad x \in \mathbb{R}^d.$$

and $\mathcal{C}(\psi_n) \leq 8^n (\mathcal{C}(\phi) + \frac{34}{7}) - \frac{34}{7}$, where $\mathcal{C}(\phi)$ denotes the complexity of ϕ_m , $m \in \mathbb{N}$. The same result hold for the pointwise minimum operation.

Proof. We first prove the result for the pointwise maximum operation by an induction on n . It is clear the statement holds for $n = 0$ with $\psi = \phi_1$. Now for any given $k \in \mathbb{N} \cup \{0\}$, we shall consider the case $2l = 2^{k+1}$ by assuming the statement holds for $l = 2^k$. Note that we have

$$g_{2l}(x) := \max_{m=1, \dots, 2l} f_m(x) = \max(g_l^{(1)}(x), g_l^{(2)}(x)), \quad x \in \mathbb{R}^d, \quad (3.5)$$

where we denote $f_m = \mathcal{R}_\varrho(\phi_m)$ for all m , $g_l^{(1)} = \max_{m=1,\dots,l} f_m$ and $g_l^{(2)} = \max_{m=l+1,\dots,2l} f_m$. Then since ϕ_m has the same architecture, by induction hypothesis, we know $g_l^{(1)}$ and $g_l^{(2)}$ can be represented by networks $\psi^{(1)}$ and $\psi^{(2)}$, respectively, with the same architecture:

$$\psi_l^{(i)} = ((W_1^{(m)}, b_1^{(m)}), (W_2^{(m)}, b_2^{(m)}), \dots, (W_{L'}^{(m)}, b_{L'}^{(m)})) \in \mathcal{N}_{L'}^{N'_0, N'_1, \dots, N'_{L'}}, \quad m = 1, 2,$$

for some integers $L', N'_0, N'_1, \dots, N'_{L'} \in \mathbb{N}$ with $N'_0 = d$ and $N'_{L'} = 1$. Then one can construct the following network:

$$\begin{aligned} & \mathcal{P}(\psi_l^{(1)}, \psi_l^{(2)}) \\ &= \left(\begin{pmatrix} W_1^{(1)} \\ W_1^{(2)} \end{pmatrix}, \begin{pmatrix} b_1^{(1)} \\ b_1^{(2)} \end{pmatrix} \right), \left(\begin{pmatrix} W_2^{(1)} & & \\ & W_2^{(2)} & \\ & & \ddots \end{pmatrix}, \begin{pmatrix} b_2^{(1)} \\ b_2^{(2)} \\ \vdots \end{pmatrix} \right), \dots, \left(\begin{pmatrix} W_{L'}^{(1)} & & \\ & W_{L'}^{(2)} & \\ & & \ddots \end{pmatrix}, \begin{pmatrix} b_{L'}^{(1)} \\ b_{L'}^{(2)} \\ \vdots \end{pmatrix} \right), \end{aligned}$$

and verify that it represents the parallelization of $\psi^{(1)}$ and $\psi^{(2)}$:

$$[\mathcal{R}_\varrho(\mathcal{P}(\psi_l^{(1)}, \psi_l^{(2)}))](x) = ([\mathcal{R}_\varrho(\psi_l^{(1)})](x), [\mathcal{R}_\varrho(\psi_l^{(2)})](x)) = (g_l^{(1)}(x), g_l^{(2)}(x)), \quad x \in \mathbb{R}^d.$$

Moreover, we have

$$\begin{aligned} \mathcal{C}(\mathcal{P}(\psi^{(1)}, \psi^{(2)})) &= 2N'_1(N'_0 + 1) + \sum_{i=2}^{L'} (2N'_i)(2N'_{i-1} + 1) \\ &\leq 4 \left(N'_1(N'_0 + 1) + \sum_{i=2}^{L'} N'_i(N'_{i-1} + 1) \right) = 2(\mathcal{C}(\psi^{(1)}) + \mathcal{C}(\psi^{(2)})). \end{aligned}$$

Note that the following identity for the max function $(x, y) \in \mathbb{R}^2 \rightarrow \max(x, y) \in \mathbb{R}$:

$$\max(x, y) = 0.5(\max(x - y, 0) + \max(y - x, 0) + \max(x + y, 0) - \max(-x - y, 0)),$$

implies that the max function can be represented by the following 2-layer ReLU network:

$$\psi_{\max} = \left(\left(\begin{pmatrix} 1 & -1 \\ -1 & 1 \\ 1 & 1 \\ -1 & -1 \end{pmatrix}, 0 \right), \left((0.5 \quad 0.5 \quad 0.5 \quad -0.5), 0 \right) \right).$$

Therefore, by using (3.5) and Lemma A.4, we deduce that there exists a DNN $\psi_{2l} \in \mathcal{N}$ representing g_{2l} with the complexity

$$\mathcal{C}(\psi_{2l}) \leq 2(\mathcal{C}(\psi_{\max}) + \mathcal{C}(\mathcal{P}(\psi^{(1)}, \psi^{(2)}))) = 2(17 + 2(\mathcal{C}(\psi_l^{(1)}) + \mathcal{C}(\psi_l^{(2)}))).$$

Then by using the hypothesis on $\mathcal{C}(\psi_l^{(1)})$ and $\mathcal{C}(\psi_l^{(2)})$, we obtain that

$$\mathcal{C}(\psi_{2l}) \leq 34 + 8 \left(8^k(\mathcal{C}(\phi) + \frac{34}{7}) - \frac{34}{7} \right) = 8^{k+1} \left(\mathcal{C}(\phi) + \frac{34}{7} \right) - \frac{34}{7},$$

which completes our proof for the pointwise maximum operation.

Finally, by observing the simple identity

$$\min_{m=1,\dots,2^n} \mathcal{R}_\varrho(\phi_m) = - \max_{m=1,\dots,2^n} -\mathcal{R}_\varrho(\phi_m)$$

and the fact that scaling a function can be achieved by adjusting the weights in the output layer of its DNN representation without change its architecture, we can conclude the same result for the pointwise minimum operation. \square

4 Linear-implicit Euler discretizations for SDEs

In this section, we shall derive precise error estimates of linear-implicit Euler discretization for a finite-dimensional SDE. In particular, we shall demonstrate that under the monotonicity condition in (H.1(a)), the approximation error of the linear-implicit Euler scheme depends polynomially on the Lipschitz constants of the coefficients, which is crucial for our analysis on the DNN expression rates in Section 6.

Let $d \in \mathbb{N}$ and $x_0 \in \mathbb{R}^d$ be fixed throughout this section. We consider the following SDE:

$$dY_t = (-AY_t + \mu(t, Y_t)) dt + \sigma(t, Y_t) dB_t, \quad t \in (0, T]; \quad Y_0 = x_0, \quad (4.1)$$

where $(B_t)_{t \in [0, T]}$ is a d -dimensional Brownian motion on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

We now introduce two linear-implicit Euler discretizations for (4.1). For any given $N \in \mathbb{N}$, we shall consider the family of random variables $(Y_n^\pi)_{n=0}^N$ defined as follows: $Y_0^\pi = (I_d + hA)^{-1}x_0$, and for all $n = 0, \dots, N-1$,

$$Y_{n+1}^\pi - Y_n^\pi + hAY_{n+1}^\pi = h\mu(t_n, Y_n^\pi) + \sigma(t_n, Y_n^\pi)\Delta B_{n+1}, \quad (4.2)$$

where $h = T/N$ and $\Delta B_{n+1} = B_{(n+1)h} - B_{nh}$. We shall also consider the family of random variables $(\tilde{Y}_n^\pi)_{n=0}^N$ defined by the following perturbed Euler scheme: $\tilde{Y}_0^\pi = (I_d + hA)^{-1}x_0$, and for all $n = 0, \dots, N-1$,

$$\tilde{Y}_{n+1}^\pi - \tilde{Y}_n^\pi + hA\tilde{Y}_{n+1}^\pi = h\tilde{\mu}(t_n, \tilde{Y}_n^\pi) + \tilde{\sigma}(t_n, \tilde{Y}_n^\pi)\Delta B_{n+1}. \quad (4.3)$$

In the sequel, we shall simply refer (4.2) and (4.3) as ES and PES, respectively.

We shall make the following assumptions on the coefficients of the SDE (4.1) and the Euler schemes, which are analogues of (H.1) and (H.2) for the fixed d -dimensional problem.

H.5. Let $x_0 \in \mathbb{R}^d$, $A \in \mathbb{R}^{d \times d}$, $\eta, D > 0$ and $\beta, \gamma, \theta, C_{\mu,0}, C_{\mu,1}, C_{\sigma,0}, C_{\sigma,1}, C_f \geq 0$ be given constants. Let $\mu, \tilde{\mu} : [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}^d$, $\sigma, \tilde{\sigma} : [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d}$, $f, f_D, \tilde{f}_D : \mathbb{R}^d \rightarrow \mathbb{R}$ be measurable functions satisfying the following conditions, for all $t, s \in [0, T]$ and $x, y \in \mathbb{R}^d$:

(a) The matrix A and the functions μ, σ satisfy the following monotonicity condition:

$$\begin{aligned} \langle x - y, \mu(t, x) - \mu(t, y) \rangle + \eta \|\mu(t, x) - \mu(t, y)\|^2 + \frac{1 + \eta}{2} \|\sigma(t, x) - \sigma(t, y)\|^2 \\ \leq \beta \|x - y\|^2 + \langle x - y, A(x - y) \rangle. \end{aligned}$$

(b) $\langle x, Ax \rangle \geq 0$ for all $x \in \mathbb{R}^d$.

(c) μ and σ admit the following regularity:

$$\begin{aligned} \|\mu(t, x) - \mu(s, y)\| &\leq C_{\mu,1}(\sqrt{t-s} + \|x - y\|), \quad \|\mu(t, 0)\| \leq C_{\mu,0}; \\ \|\sigma(t, x) - \sigma(s, y)\| &\leq C_{\sigma,1}(\sqrt{t-s} + \|x - y\|), \quad \|\sigma(t, 0)\| \leq C_{\sigma,0}. \end{aligned}$$

(d) (μ, σ) and $(\tilde{\mu}, \tilde{\sigma})$ satisfy the following estimate:

$$\|\mu(t, x) - \tilde{\mu}(t, x)\| + \|\sigma(t, x) - \tilde{\sigma}(t, x)\| \leq \gamma.$$

(e) f, f_D, \tilde{f}_D satisfy the following estimates:

$$|f(x) - f_D(x)| \leq C_f \|x\|^2 1_{B_\infty(D)^c}(x), \quad |f_D(x) - f_D(y)| \leq C_f D \|x - y\|, \quad |f_D(x) - \tilde{f}_D(x)| \leq \theta,$$

where $B_\infty(D) := \{x \in \mathbb{R}^d \mid x_i \in [-D, D], \forall i\}$.

Remark 4.1. Throughout this section, we shall assume without loss of generality that $\eta \in (0, 1)$ in (H.5(a)). Moreover, for any given $h > 0$, we can directly deduce from (H.5(b)) that $(I_d + hA)x \neq 0$ for all $x \neq 0$, which implies that the matrix $I_d + hA$ is nonsingular, and satisfies the estimates $\|(I_d + hA)^{-1}\|_{\text{op}} \leq 1$ and $\|hA(I_d + hA)^{-1}\|_{\text{op}} \leq 1$ (see e.g. [6, Proposition 7.2]).

The following lemma estimates the growth rates of the coefficients (μ, σ) and $(\tilde{\mu}, \tilde{\sigma})$.

Lemma 4.1. *Suppose (H.5) holds. Then the functions μ, σ satisfy the following growth condition: for all $\eta' \in [0, \eta]$ and $(t, x) \in [0, T] \times \mathbb{R}^d$, we have*

$$\langle x, \mu(t, x) \rangle + \eta' \|\mu(t, x)\|^2 + \frac{1 + \eta'}{2} \|\sigma(t, x)\|^2 \leq \alpha' + \left(\beta + \frac{1}{2} \right) \|x\|^2 + \langle x, Ax \rangle, \quad (4.4)$$

with the constant

$$\alpha' = \left(\frac{1}{2} + \frac{\eta \eta'}{\eta - \eta'} \right) C_{\mu,0}^2 + \frac{(1 + \eta)(1 + \eta')}{2(\eta - \eta')} C_{\sigma,0}^2.$$

Similarly, the functions $\tilde{\mu}, \tilde{\sigma}$ satisfy the following growth condition: for all $(t, x) \in [0, T] \times \mathbb{R}^d$,

$$\langle x, \tilde{\mu}(t, x) \rangle + \frac{\eta}{2} \|\tilde{\mu}(t, x)\|^2 + \frac{1}{2} \|\tilde{\sigma}(t, x)\|^2 \leq \frac{2(1 + \eta)^2}{\eta} (C_{\mu,0}^2 + C_{\sigma,0}^2 + \gamma^2) + (\beta + 1) \|x\|^2 + \langle x, Ax \rangle. \quad (4.5)$$

Proof. Note that for any given matrices $A, B \in \mathbb{R}^{d_1 \times d_2}$, Young's inequality implies that $\|A + B\|^2 \leq (1 + \tau) \|A\|^2 + (1 + 1/\tau) \|B\|^2$ for all $\tau > 0$. Then for any fixed $\eta' \in (0, \eta)$, we can deduce from (H.5(a)) and (H.5(c)) that

$$\begin{aligned} & \langle x, \mu(t, x) \rangle + \eta' \|\mu(t, x)\|^2 + \frac{1 + \eta'}{2} \|\sigma(t, x)\|^2 \\ &= \langle x, \mu(t, x) - \mu(t, 0) + \mu(t, 0) \rangle + \eta' \|\mu(t, x) - \mu(t, 0) + \mu(t, 0)\|^2 + \frac{1 + \eta'}{2} \|\sigma(t, x) - \sigma(t, 0) + \sigma(t, 0)\|^2 \\ &\leq \langle x, \mu(t, x) - \mu(t, 0) \rangle + \frac{1}{2} (\|x\|^2 + \|\mu(t, 0)\|^2) + \eta' \left(\frac{\eta}{\eta'} \|\mu(t, x) - \mu(t, 0)\|^2 + \frac{\eta}{\eta - \eta'} \|\mu(t, 0)\|^2 \right) \\ &\quad + \frac{1 + \eta'}{2} \left(\frac{1 + \eta}{1 + \eta'} \|\sigma(t, x) - \sigma(t, 0)\|^2 + \frac{1 + \eta}{\eta - \eta'} \|\sigma(t, 0)\|^2 \right) \\ &\leq \beta \|x\|^2 + \langle x, Ax \rangle + \frac{1}{2} \|x\|^2 + \left(\frac{1}{2} + \frac{\eta \eta'}{\eta - \eta'} \right) C_{\mu,0}^2 + \frac{(1 + \eta)(1 + \eta')}{2(\eta - \eta')} C_{\sigma,0}^2. \end{aligned} \quad (4.6)$$

It is straightforward to verify that the above inequality also holds for $\eta' = 0$.

On the other hand, by using (H.5(d)), we can obtain

$$\begin{aligned} & \langle x, \tilde{\mu}(t, x) \rangle + \frac{\eta}{2} \|\tilde{\mu}(t, x)\|^2 + \frac{1}{2} \|\tilde{\sigma}(t, x)\|^2 \\ &= \langle x, \tilde{\mu}(t, x) - \mu(t, x) + \mu(t, x) \rangle + \frac{\eta}{2} \|\tilde{\mu}(t, x) - \mu(t, x) + \mu(t, x)\|^2 + \frac{1}{2} \|\tilde{\sigma}(t, x) - \sigma(t, x) + \sigma(t, x)\|^2 \\ &\leq \langle x, \mu(t, x) \rangle + \frac{1}{2} (\|x\|^2 + \|\tilde{\mu}(t, x) - \mu(t, x)\|^2) + \frac{\eta}{2} \left(\frac{3}{2} \|\mu(t, x)\|^2 + 3 \|\tilde{\mu}(t, x) - \mu(t, x)\|^2 \right) \\ &\quad + \frac{1}{2} \left(\left(1 + \frac{3\eta}{4} \right) \|\sigma(t, x)\|^2 + \left(1 + \frac{4}{3\eta} \right) \|\tilde{\sigma}(t, x) - \sigma(t, x)\|^2 \right) \\ &\leq \langle x, \mu(t, x) \rangle + \frac{3\eta}{4} \|\mu(t, x)\|^2 + \frac{1 + 3\eta/4}{2} \|\sigma(t, x)\|^2 + \frac{1}{2} \|x\|^2 + \left(1 + \frac{3\eta}{2} + \frac{2}{3\eta} \right) \gamma^2, \end{aligned}$$

which, together with the estimate (4.6) (with $\eta' = 3\eta/4$), gives us that

$$\begin{aligned}
& \langle x, \tilde{\mu}(t, x) \rangle + \frac{\eta}{2} \|\tilde{\mu}(t, x)\|^2 + \frac{1}{2} \|\tilde{\sigma}(t, x)\|^2 \\
& \leq \beta \|x\|^2 + \langle x, Ax \rangle + \frac{1}{2} \|x\|^2 + \left(\frac{1}{2} + 3\eta \right) C_{\mu,0}^2 + \frac{2(1+\eta)^2}{\eta} C_{\sigma,0}^2 + \frac{1}{2} \|x\|^2 + \left(1 + \frac{3\eta}{2} + \frac{2}{3\eta} \right) \gamma^2 \\
& \leq (\beta + 1) \|x\|^2 + \langle x, Ax \rangle + \frac{2(1+\eta)^2}{\eta} \left(C_{\mu,0}^2 + C_{\sigma,0}^2 + \gamma^2 \right),
\end{aligned}$$

where we used the assumption that $\eta < 1$ in the last inequality. \square

With Lemma 4.1 in hand, we now present the following moment estimate and time regularity result for the strong solutions to (4.1). Note that both the moment estimate and the regularity estimate depend polynomially on the parameters $\|A\|_{\text{op}}$, $C_{\mu,l}$, $C_{\sigma,l}$, $l = 0, 1$. This observation plays a crucial role in our subsequent analysis.

Lemma 4.2. *Suppose (H.5) holds. Then the SDE (4.1) admits a unique strong solution $(Y_t)_{t \in [0, T]}$, which admits the following a priori estimate: for all $p \in [2, 2 + \eta]$ and $t \in [0, T]$,*

$$\mathbb{E}[\|Y_t\|^p] \leq 2^{\frac{p-2}{2}} (\alpha_p + \|x_0\|^p) e^{p(\beta+1/2)T},$$

with the constant α_p defined as:

$$\alpha_p = \left(\frac{1}{2} + \frac{\eta(p-2)}{\eta+2-p} \right) C_{\mu,0}^2 + \frac{(1+\eta)(p-1)}{2(\eta+2-p)} C_{\sigma,0}^2, \quad \forall p \in [2, 2 + \eta], \quad (4.7)$$

and the following time regularity: for all $t, s \in [0, T]$,

$$\mathbb{E}[\|Y_t - Y_s\|^2] \leq 8(1 + \alpha + \|x_0\|^2) e^{(2\beta+1)T} \left((t-s)^2 (\|A\|_{\text{op}}^2 + C_{\mu,0}^2 + C_{\mu,1}^2) + (t-s)(C_{\sigma,0}^2 + C_{\sigma,1}^2) \right),$$

with $\alpha = \frac{1}{2} C_{\mu,0}^2 + \frac{1+\eta}{2\eta} C_{\sigma,0}^2$.

Proof. The *a priori* estimate follows precisely the steps in the arguments for [39, Theorem 4.1 pp. 59] by applying Itô's formula to the quantity $(\alpha + \|Y_t\|^2)^{\frac{p}{2}}$ and using the growth condition (4.4) in Lemma 4.1. Then one can deduce from (H.5(a)) that

$$\begin{aligned}
\mathbb{E}[\|Y_t - Y_s\|^2] & \leq 2 \left((t-s) \int_s^t \mathbb{E}[\| -AY_r + \mu(s, Y_r) \|^2] dr + \int_s^t \mathbb{E}[\|\sigma(r, Y_r)\|^2] dr \right) \\
& \leq 2 \left\{ 2(t-s) \int_s^t \left(\|A\|_{\text{op}}^2 \mathbb{E}[\|Y_r\|^2] + 2(C_{\mu,0}^2 + C_{\mu,1}^2 \mathbb{E}[\|Y_r\|^2]) \right) dr \right. \\
& \quad \left. + 2 \int_s^t \left(C_{\sigma,0}^2 + C_{\sigma,1}^2 \mathbb{E}[\|Y_r\|^2] \right) dr \right\} \\
& \leq 8 \left((t-s)^2 (\|A\|_{\text{op}}^2 + C_{\mu,0}^2 + C_{\mu,1}^2) + (t-s)(C_{\sigma,0}^2 + C_{\sigma,1}^2) \right) \left(1 + \sup_{t \in [0, T]} \mathbb{E}[\|Y_t\|^2] \right),
\end{aligned}$$

which, along with the estimate of $\mathbb{E}[\|Y_t\|^2]$, implies the desired modulus of continuity in time. \square

Now we proceed to study the linear-implicit Euler schemes (4.2) and (4.3). The following proposition shows the stability of the linear-implicit Euler scheme.

Proposition 4.3. Suppose (H.5(b)) holds. Let $h > 0$ and $t \in [0, T]$ be given constants, $\mu_i : [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}^d$, $\sigma_i : [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d}$, $i = 1, 2$, be Borel measurable functions, Z be a \mathbb{R}^d -valued random variable with mean 0 and variance h , and $(Y^i)_{i=1,2}$ be \mathbb{R}^d -valued integrable random variables independent of Z . For each $i = 1, 2$, let X^i be the random variable defined by the following one-step linear-implicit Euler scheme:

$$X^i - Y^i + hAX^i = h\mu_i(t, Y^i) + \sigma_i(t, Y^i)Z. \quad (4.8)$$

Then we have the following stability estimate:

$$\begin{aligned} & \frac{1}{2}\mathbb{E}[\|X^1 - X^2\|^2] + h\mathbb{E}[\langle X^1 - X^2, A(X^1 - X^2) \rangle] \\ & \leq \frac{1}{2}\mathbb{E}[\|Y^1 - Y^2\|^2] + h\mathbb{E}\left[\langle Y^1 - Y^2, \mu_1(t, Y^1) - \mu_2(t, Y^2) \rangle + \frac{1}{2}h\|\mu_1(t, Y^1) - \mu_2(t, Y^2)\|^2 \right. \\ & \quad \left. + \frac{1}{2}\|\sigma_1(t, Y^1) - \sigma_2(t, Y^2)\|^2\right]. \end{aligned}$$

Proof. For notational simplicity, we introduce the following terms: $\delta X = X^1 - X^2$, $\delta Y = Y^1 - Y^2$, $\delta\mu = \mu_1(t, Y^1) - \mu_2(t, Y^2)$, and $\delta\sigma = \sigma_1(t, Y^1) - \sigma_2(t, Y^2)$. Then we can deduce from (4.8) that

$$\delta X - \delta Y + hA(\delta X) - h(\delta\mu) = (\delta\sigma)Z.$$

Multiplying the above identity by δX , we obtain that

$$\langle \delta X, \delta X - \delta Y - h(\delta\mu) \rangle + h\langle \delta X, A(\delta X) \rangle = \langle \delta X, (\delta\sigma)Z \rangle.$$

from which, by completing the square, one can deduce that

$$\frac{1}{2}\|\delta X\|^2 + \frac{1}{2}\|\delta X - \delta Y - h(\delta\mu)\|^2 - \frac{1}{2}\|\delta Y + h(\delta\mu)\|^2 + h\langle \delta X, A(\delta X) \rangle = \langle \delta X, (\delta\sigma)Z \rangle. \quad (4.9)$$

Then, by using the following inequality:

$$\begin{aligned} 0 & \leq \|\delta X - \delta Y - h(\delta\mu) - (\delta\sigma)Z\|^2 \\ & \leq \|\delta X - \delta Y - h(\delta\mu)\|^2 - 2\langle \delta X - \delta Y - h(\delta\mu), (\delta\sigma)Z \rangle + \|(\delta\sigma)Z\|^2, \end{aligned}$$

and the fact that Z is independent of δY , we can obtain that

$$\begin{aligned} \frac{1}{2}\mathbb{E}[\|\delta X\|^2] + h\mathbb{E}[\langle \delta X, A(\delta X) \rangle] & \leq \frac{1}{2}\mathbb{E}[\|\delta Y + h(\delta\mu)\|^2] + \frac{1}{2}h\mathbb{E}[\|(\delta\sigma)\|^2] \\ & \leq \frac{1}{2}\mathbb{E}[\|\delta Y\|^2] + h\mathbb{E}\left[\langle \delta Y, (\delta\mu) \rangle + \frac{1}{2}h\|(\delta\mu)\|^2 + \frac{1}{2}\|(\delta\sigma)\|^2\right], \end{aligned}$$

which completes the proof of the desired stability estimates. \square

The next two corollaries follow directly from Proposition 4.3, which give an L^2 -estimate of the numerical solutions to ES (4.2) and PES (4.3), and establish an upper bound of the difference between these two solutions.

Corollary 4.4. Suppose (H.5) holds. Let $N \in \mathbb{N}$ with $N \geq 2T/\eta$, and $(Y_n^\pi)_{n=0}^N$ (resp. $(\tilde{Y}_n^\pi)_{n=0}^N$) be the solution to ES (4.2) (resp. PES (4.3)). Then the following estimate holds:

$$\begin{aligned} \max_{n=0, \dots, N} \left(\mathbb{E}[\|Y_n^\pi\|^2] + 2h\mathbb{E}[\langle Y_n^\pi, AY_n^\pi \rangle] \right) & \leq 3e^{(2\beta+1)T}(\|x_0\|^2 + \alpha_1 T), \\ \max_{n=0, \dots, N} \left(\mathbb{E}[\|\tilde{Y}_n^\pi\|^2] + 2h\mathbb{E}[\langle \tilde{Y}_n^\pi, A\tilde{Y}_n^\pi \rangle] \right) & \leq 3e^{(2\beta+1)T}(\|x_0\|^2 + \alpha_2 T), \end{aligned}$$

where $\alpha_1 = \frac{(1+\eta)^2}{\eta}(C_{\mu,0}^2 + C_{\sigma,0}^2)$ and $\alpha_2 = \frac{2(1+\eta)^2}{\eta}(C_{\mu,0}^2 + C_{\sigma,0}^2 + \gamma^2)$.

Proof. We shall only estimate $\mathbb{E}[\|Y_n^\pi\|^2]$, since the *a priori* bound of \tilde{Y}_n^π can be established by a similar approach. For any given $n = 0, \dots, N-1$, by setting $(\mu_1, \sigma_1) = (\mu, \sigma)$, $(\mu_2, \sigma_2) = (0, 0)$ and $Y^2 = 0$ in Proposition 4.3, we obtain from Lemma 4.1 that: for all $h < \eta' := \eta/2$,

$$\begin{aligned} & \mathbb{E}[\|Y_{n+1}^\pi\|^2] + 2h\mathbb{E}[\langle Y_{n+1}^\pi, AY_{n+1}^\pi \rangle] \\ & \leq \mathbb{E}[\|Y_n^\pi\|^2] + 2h\mathbb{E}\left[\eta'\|\mu(t_n, Y_n^\pi)\|^2 + \frac{1}{2}\|\sigma(t_n, Y_n^\pi)\|^2 + \langle Y_n^\pi, \mu(t_n, Y_n^\pi) \rangle\right] \\ & \leq \mathbb{E}[\|Y_n^\pi\|^2] + 2h\mathbb{E}\left[\langle Y_n^\pi, AY_n^\pi \rangle + \left(\frac{1}{2} + \eta\right)C_{\mu,0}^2 + \frac{(1+\eta)(1+\eta/2)}{\eta}C_{\sigma,0}^2 + \left(\beta + \frac{1}{2}\right)\|Y_n^\pi\|^2\right] \\ & \leq (1 + 2\beta'h)\left(\mathbb{E}[\|Y_n^\pi\|^2] + 2h\mathbb{E}[\langle Y_n^\pi, AY_n^\pi \rangle]\right) + 2\alpha'h, \end{aligned}$$

with $\alpha_1 = \frac{(1+\eta)^2}{\eta}(C_{\mu,0}^2 + C_{\sigma,0}^2)$, $\beta' = \beta + 1/2$, where we have used the fact $\mathbb{E}[\langle Y_n^\pi, AY_n^\pi \rangle] \geq 0$. Then, by multiplying the above inequality with $(1 + 2\beta'h)^{-(n+1)}$, we can deduce that

$$\begin{aligned} & (1 + 2\beta'h)^{-(n+1)}\left(\mathbb{E}[\|Y_{n+1}^\pi\|^2] + 2h\mathbb{E}[\langle Y_{n+1}^\pi, AY_{n+1}^\pi \rangle]\right) \\ & \leq (1 + 2\beta'h)^{-n}\left(\mathbb{E}[\|Y_n^\pi\|^2] + 2h\mathbb{E}[\langle Y_n^\pi, AY_n^\pi \rangle]\right) + (1 + 2\beta'h)^{-n-1}2h\alpha', \end{aligned}$$

which leads to the following estimate: for all $n = 0, \dots, N-1$,

$$\begin{aligned} \mathbb{E}[\|Y_{n+1}^\pi\|^2] + 2h\mathbb{E}[\langle Y_{n+1}^\pi, AY_{n+1}^\pi \rangle] & \leq (1 + 2\beta'h)^N\left(\mathbb{E}[\|Y_0^\pi\|^2] + 2h\mathbb{E}[\langle Y_0^\pi, AY_0^\pi \rangle] + 2\alpha'T\right) \\ & \leq e^{2\beta'T}\left(\|Y_0^\pi\|^2 + 2h\langle Y_0^\pi, AY_0^\pi \rangle + 2\alpha'T\right). \end{aligned}$$

We can then conclude the desired result from the Cauchy-Schwarz inequality and Remark 4.1. \square

Corollary 4.5. *Suppose (H.5) holds. Let $N \in \mathbb{N}$ with $N \geq T \max(2\eta, 1/\eta)$, and $(Y_n^\pi)_{n=0}^N$ (resp. $(\tilde{Y}_n^\pi)_{n=0}^N$) be the solution to ES (4.2) (resp. PES (4.3)). Then we have the following error estimate:*

$$\max_{n=0,\dots,N} \left(\mathbb{E}[\|Y_n^\pi - \tilde{Y}_n^\pi\|^2] + 2h\mathbb{E}[\langle Y_n^\pi - \tilde{Y}_n^\pi, A(Y_n^\pi - \tilde{Y}_n^\pi) \rangle] \right) \leq e^{(2\beta+1)T}T(1+\eta)\gamma^2/\eta.$$

Proof. Let us define the random variable $\delta Y_n = Y_n^\pi - \tilde{Y}_n^\pi$ for all n . For any given $n = 0, \dots, N-1$, we deduce from Proposition 4.3 that

$$\begin{aligned} & \mathbb{E}[\|\delta Y_{n+1}\|^2] + 2h\mathbb{E}[\langle \delta Y_{n+1}, A(\delta Y_{n+1}) \rangle] \leq \mathbb{E}[\|\delta Y_n\|^2] + 2h\mathbb{E}\left[\langle \delta Y_n, \mu(t_n, Y_n^\pi) - \mu(t_n, \tilde{Y}_n^\pi) \rangle \right. \\ & \quad \left. + \langle \delta Y_n, \mu(t_n, \tilde{Y}_n^\pi) - \tilde{\mu}(t_n, \tilde{Y}_n^\pi) \rangle + \frac{1}{2}h\|\mu(t_n, Y_n^\pi) - \tilde{\mu}(t_n, \tilde{Y}_n^\pi)\|^2 + \frac{1}{2}\|\sigma(t_n, Y_n^\pi) - \tilde{\sigma}(t_n, \tilde{Y}_n^\pi)\|^2\right]. \end{aligned} \quad (4.10)$$

Now we estimate the last three terms in the above inequality. It is clear that the Cauchy-Schwarz inequality gives us that

$$\begin{aligned} \langle \delta Y_n, \mu(t_n, \tilde{Y}_n^\pi) - \tilde{\mu}(t_n, \tilde{Y}_n^\pi) \rangle & \leq \frac{1}{2}(\|\delta Y_n\|^2 + \|\mu(t_n, \tilde{Y}_n^\pi) - \tilde{\mu}(t_n, \tilde{Y}_n^\pi)\|^2), \\ \|\mu(t_n, Y_n^\pi) - \tilde{\mu}(t_n, \tilde{Y}_n^\pi)\|^2 & = \|\mu(t_n, Y_n^\pi) - \mu(t_n, \tilde{Y}_n^\pi) + \mu(t_n, \tilde{Y}_n^\pi) - \tilde{\mu}(t_n, \tilde{Y}_n^\pi)\|^2 \\ & \leq 2\|\mu(t_n, Y_n^\pi) - \mu(t_n, \tilde{Y}_n^\pi)\|^2 + 2\|\mu(t_n, \tilde{Y}_n^\pi) - \tilde{\mu}(t_n, \tilde{Y}_n^\pi)\|^2. \end{aligned}$$

Moreover, by applying the Cauchy-Schwarz inequality to Frobenius inner product of matrices, we obtain that

$$\begin{aligned}\|\sigma(t_n, Y_n^\pi) - \tilde{\sigma}(t_n, \tilde{Y}_n^\pi)\|^2 &= \|\sigma(t_n, Y_n^\pi) - \sigma(t_n, \tilde{Y}_n^\pi) + \sigma(t_n, \tilde{Y}_n^\pi) - \tilde{\sigma}(t_n, \tilde{Y}_n^\pi)\|^2 \\ &\leq (1 + \eta)\|\sigma(t_n, Y_n^\pi) - \sigma(t_n, \tilde{Y}_n^\pi)\|^2 + (1 + \frac{1}{\eta})\|\sigma(t_n, \tilde{Y}_n^\pi) - \tilde{\sigma}(t_n, \tilde{Y}_n^\pi)\|^2.\end{aligned}$$

Hence, by substituting the above estimates into (4.10) and rearranging the terms, we deduce that

$$\begin{aligned}&\mathbb{E}[\|\delta Y_{n+1}\|^2] + 2h\mathbb{E}[\langle \delta Y_{n+1}, A(\delta Y_{n+1}) \rangle] \\ &\leq (1 + h)\mathbb{E}[\|\delta Y_n\|^2] + 2h\mathbb{E}\left[\langle \delta Y_n, \mu(t_n, Y_n^\pi) - \mu(t_n, \tilde{Y}_n^\pi) \rangle + h\|\mu(t_n, Y_n^\pi) - \mu(t_n, \tilde{Y}_n^\pi)\|^2\right. \\ &\quad \left. + \frac{1 + \eta}{2}\|\sigma(t_n, Y_n^\pi) - \sigma(t_n, \tilde{Y}_n^\pi)\|^2\right] \\ &\quad + h\mathbb{E}\left[(1 + 2h)\|\mu(t_n, \tilde{Y}_n^\pi) - \tilde{\mu}(t_n, \tilde{Y}_n^\pi)\|^2 + (1 + \frac{1}{\eta})\|\sigma(t_n, \tilde{Y}_n^\pi) - \tilde{\sigma}(t_n, \tilde{Y}_n^\pi)\|^2\right].\end{aligned}$$

Hence for all $N \in \mathbb{N}$ such that $T/N \leq \min(1/(2\eta), \eta)$, (H.5(a)) and (H.5(d)) imply that

$$\begin{aligned}&\mathbb{E}[\|\delta Y_{n+1}\|^2] + 2h\mathbb{E}[\langle \delta Y_{n+1}, A(\delta Y_{n+1}) \rangle] \\ &\leq (1 + h)\mathbb{E}[\|\delta Y_n\|^2] + 2h\mathbb{E}[\beta\|\delta Y_n\|^2 + \langle \delta Y_n, A(\delta Y_n) \rangle] + h(1 + \eta)\gamma^2/\eta \\ &\leq (1 + (2\beta + 1)h)\mathbb{E}[\|\delta Y_n\|^2] + 2h\mathbb{E}[\langle \delta Y_n, A(\delta Y_n) \rangle] + h(1 + \eta)\gamma^2/\eta.\end{aligned}$$

Thus, following similar arguments as those for Corollary 4.4, we can conclude the desired estimate by using the fact that $\delta Y_0 = 0$. \square

Now we proceed to derive precise error estimates of the linear-implicit Euler schemes (4.2) and (4.3). The following proposition shows the overall approximation error can be bounded by the one-step local truncation errors.

Proposition 4.6. *Suppose (H.5) holds. Let $N \in \mathbb{N}$ with $N \geq T \max\left(\frac{2\beta + 2^{1/T}}{2^{1/T} - 1}, \frac{1}{\eta}\right)$, $(Y_t)_{t \in [0, T]}$ be the solution to SDE (4.1), and $(Y_n^\pi)_{n=0}^N$ be the solution to ES (4.2). Then we have the following error estimate:*

$$\begin{aligned}&\max_{n=0, \dots, N} \mathbb{E}[\|Y_n^\pi - Y_{t_n}\|^2] + 2h\mathbb{E}[\langle Y_n^\pi - Y_{t_n}, A(Y_n^\pi - Y_{t_n}) \rangle] \\ &\leq 2e^{(2\beta + 1)T} \left(\|\delta Y_0\|^2 + 2h\langle \delta Y_0, A\delta Y_0 \rangle + \frac{1}{h} \sum_{n=0}^{N-1} \mathbb{E}[\|e_{n+1}^1\|^2] + (1 + \frac{1}{\eta}) \sum_{n=0}^{N-1} \mathbb{E}[\|e_{n+1}^2\|^2] \right),\end{aligned}$$

with $\delta Y_0 = ((I_d + hA)^{-1} - I_d)x_0$, and the truncation errors e_{n+1}^1, e_{n+1}^2 defined as:

$$e_{n+1}^1 := \int_{t_n}^{t_{n+1}} (-A(Y_s - Y_{t_{n+1}}) + \mu(s, Y_s) - \mu(t_n, Y_{t_n})) ds, \quad e_{n+1}^2 := \int_{t_n}^{t_{n+1}} (\sigma(s, Y_s) - \sigma(t_n, Y_{t_n})) dB_s. \quad (4.11)$$

Proof. For any $n = 0, \dots, N$, we define the random variables $\delta Y_n = Y_n^\pi - Y_{t_n}$, $\delta \mu_n = \mu(t_n, Y_n^\pi) - \mu(t_n, Y_{t_n})$, and $\delta \sigma_n = \sigma(t_n, Y_n^\pi) - \sigma(t_n, Y_{t_n})$. Note that we have

$$Y_{t_{n+1}} - Y_{t_n} + hAY_{t_{n+1}} = h\mu(t_n, Y_{t_n}) + \sigma(t_n, Y_{t_n})\Delta B_{n+1} + e_{n+1}^1 + e_{n+1}^2, \quad n = 0, \dots, N-1. \quad (4.12)$$

Then, by subtracting (4.12) from (4.2), multiplying the resulting equation with δY_{n+1} , and completing the square (cf. (4.9)), we can deduce the following identity:

$$\begin{aligned} \frac{1}{2}\|\delta Y_{n+1}\|^2 + \frac{1}{2}\|\delta Y_{n+1} - \delta Y_n - h(\delta \mu_n)\|^2 - \frac{1}{2}\|\delta Y_n + h(\delta \mu_n)\|^2 \\ + h\langle \delta Y_{n+1}, A(\delta Y_{n+1}) \rangle = \langle \delta Y_{n+1}, (\delta \sigma_n) \Delta B_{n+1} - e_{n+1}^1 - e_{n+1}^2 \rangle. \end{aligned}$$

which, together with the following inequality:

$$\begin{aligned} 0 &\leq \|\delta Y_{n+1} - \delta Y_n - h(\delta \mu_n) - (\delta \sigma_n) \Delta B_{n+1} + e_{n+1}^2\|^2 \\ &= \|\delta Y_{n+1} - \delta Y_n - h(\delta \mu_n)\|^2 - 2\langle \delta Y_{n+1} - \delta Y_n - h(\delta \mu_n), (\delta \sigma_n) \Delta B_{n+1} - e_{n+1}^2 \rangle \\ &\quad + \|(\delta \sigma_n) \Delta B_{n+1}\|^2 - 2\langle (\delta \sigma_n) \Delta B_{n+1}, e_{n+1}^2 \rangle + \|e_{n+1}^2\|^2, \end{aligned}$$

and the fact that $\mathbb{E}[\langle \delta Y_n + h(\delta \mu_n), (\delta \sigma_n) \Delta B_{n+1} - e_{n+1}^2 \rangle] = 0$, lead us to the estimate:

$$\begin{aligned} &\mathbb{E}[\|\delta Y_{n+1}\|^2] + 2h\mathbb{E}[\langle \delta Y_{n+1}, A(\delta Y_{n+1}) \rangle] \\ &\leq \mathbb{E}[\|\delta Y_n + h(\delta \mu_n)\|^2] - 2\mathbb{E}[\langle \delta Y_{n+1}, e_{n+1}^1 \rangle] \\ &\quad + \mathbb{E}[\|(\delta \sigma_n) \Delta B_{n+1}\|^2] - 2\mathbb{E}[\langle (\delta \sigma_n) \Delta B_{n+1}, e_{n+1}^2 \rangle] + \mathbb{E}[\|e_{n+1}^2\|^2] \\ &\leq \mathbb{E}[\|\delta Y_n\|^2] + 2h\langle \delta Y_n, \delta \mu_n \rangle + h^2\|\delta \mu_n\|^2 + h\mathbb{E}[\|\delta Y_{n+1}\|^2] + \frac{1}{h}\mathbb{E}[\|e_{n+1}^1\|^2] \\ &\quad + \mathbb{E}[\|(\delta \sigma_n) \Delta B_{n+1}\|^2] + \eta\mathbb{E}[\|(\delta \sigma_n) \Delta B_{n+1}\|^2] + \frac{1}{\eta}\mathbb{E}[\|e_{n+1}^2\|^2] + \mathbb{E}[\|e_{n+1}^2\|^2]. \end{aligned}$$

Consequently, for any $h \leq \eta$, we have

$$\begin{aligned} &(1-h)\mathbb{E}[\|\delta Y_{n+1}\|^2] + 2h\mathbb{E}[\langle \delta Y_{n+1}, A(\delta Y_{n+1}) \rangle] \\ &\leq \mathbb{E}[\|\delta Y_n\|^2] + 2h\mathbb{E}\left[\langle \delta Y_n, \delta \mu_n \rangle + \eta\|\delta \mu_n\|^2 + \frac{1+\eta}{2}\|\delta \sigma_n\|^2\right] + \frac{1}{h}\mathbb{E}[\|e_{n+1}^1\|^2] + (1+\frac{1}{\eta})\mathbb{E}[\|e_{n+1}^2\|^2] \\ &\leq (1+2\beta h)\mathbb{E}[\|\delta Y_n\|^2] + 2h\mathbb{E}[\langle \delta Y_n, A\delta Y_n \rangle] + \frac{1}{h}\mathbb{E}[\|e_{n+1}^1\|^2] + (1+\frac{1}{\eta})\mathbb{E}[\|e_{n+1}^2\|^2], \end{aligned}$$

from which, one can deduce by induction that

$$\begin{aligned} &\mathbb{E}[\|\delta Y_{n+1}\|^2] + 2h\mathbb{E}[\langle \delta Y_{n+1}, A(\delta Y_{n+1}) \rangle] \\ &\leq \left(\frac{1+2\beta h}{1-h}\right)^N \left(\mathbb{E}[\|\delta Y_0\|^2] + 2h\mathbb{E}[\langle \delta Y_0, A\delta Y_0 \rangle] + \frac{1}{h} \sum_{n=0}^{N-1} \mathbb{E}[\|e_{n+1}^1\|^2] + (1+\frac{1}{\eta}) \sum_{n=0}^{N-1} \mathbb{E}[\|e_{n+1}^2\|^2]\right), \end{aligned}$$

which leads to the desired statement for all large enough N such that $(\frac{N+2\beta T}{N-T})^T \leq 2$. \square

Now we are ready to present the strong convergence result of the perturbed Euler scheme.

Theorem 4.7. *Suppose (H.5) holds. Let $N \in \mathbb{N}$ with $N \geq T \max\left(\frac{2\beta+2^{1/T}}{2^{1/T}-1}, \frac{1}{\eta}, 2\eta\right)$, $(Y_t)_{t \in [0, T]}$ be the solution to SDE (4.1), and $(\tilde{Y}_n^\pi)_{n=0}^N$ be the solution to PES (4.3). Then we have the following error estimate:*

$$\begin{aligned} &\max_{n=0, \dots, N} \mathbb{E}[\|\tilde{Y}_n^\pi - Y_{t_n}\|^2] \\ &\leq C_{(\beta, \eta, T)} h \left\{ \left(1 + \|A\|_{\text{op}}^2 + \sum_{l=0}^1 (C_{\mu, l}^2 + C_{\sigma, l}^2)\right)^2 \|x_0\|^2 + \left(1 + \|A\|_{\text{op}}^2 + \sum_{l=0}^1 (C_{\mu, l}^2 + C_{\sigma, l}^2)\right)^3 + \frac{\gamma^2}{h} \right\}, \end{aligned}$$

Proof. For any given $n = 0, \dots, N-1$, by using (H.5(c)) and the Cauchy-Schwarz inequality, we can estimate the truncation error e_{n+1}^1 defined by (4.11) as follows:

$$\begin{aligned}
\mathbb{E}[\|e_{n+1}^1\|^2] &\leq \mathbb{E}\left[\left(\int_{t_n}^{t_{n+1}} \|\cdot - A(Y_s - Y_{t_{n+1}}) + \mu(s, Y_s) - \mu(t_n, Y_{t_n})\| ds\right)^2\right] \\
&\leq 2h\mathbb{E}\left[\int_{t_n}^{t_{n+1}} \left(\|A(Y_s - Y_{t_{n+1}})\|^2 + \|\mu(s, Y_s) - \mu(t_n, Y_{t_n})\|^2\right) ds\right] \\
&\leq 2h\mathbb{E}\left[\int_{t_n}^{t_{n+1}} \left(\|A\|_{\text{op}}^2 \|Y_{t_{n+1}} - Y_s\|^2 + C_{\mu,1}^2 (\sqrt{s - t_n} + \|Y_s - Y_{t_n}\|)^2\right) ds\right] \\
&\leq 2h^2 \left[(\|A\|_{\text{op}}^2 + 2C_{\mu,1}^2) \sup_{s \in [t_n, t_{n+1}]} \left(\mathbb{E}[\|Y_{t_{n+1}} - Y_s\|^2] + \mathbb{E}[\|Y_s - Y_{t_n}\|^2] \right) + 2C_{\mu,1}^2 h \right].
\end{aligned}$$

Similarly, one can obtain the following upper bound of the truncation error e_{n+1}^2 :

$$\begin{aligned}
\mathbb{E}[\|e_{n+1}^2\|^2] &= \mathbb{E}\left[\int_{t_n}^{t_{n+1}} \|\sigma(s, Y_s) - \sigma(t_n, Y_{t_n})\|^2 ds\right] \leq \mathbb{E}\left[\int_{t_n}^{t_{n+1}} C_{\sigma,1}^2 (\sqrt{s - t_n} + \|Y_s - Y_{t_n}\|)^2 ds\right] \\
&\leq 2hC_{\sigma,1}^2 \left[h + \sup_{s \in [t_n, t_{n+1}]} \mathbb{E}[\|Y_s - Y_{t_n}\|^2] \right].
\end{aligned}$$

Thus, if we denote by $(Y_n^\pi)_{n=0}^N$ the solution to ES (4.2), then for all $N \in \mathbb{N}$ with $N \geq T \max\left(\frac{2\beta+2^{1/T}}{2^{1/T}-1}, \frac{1}{\eta}\right)$, we can infer from Proposition 4.6 and the assumption $\eta < 1$ that

$$\begin{aligned}
&\max_{n=0, \dots, N} \mathbb{E}[\|Y_n^\pi - Y_{t_n}\|^2] + 2h\mathbb{E}[\langle Y_n^\pi - Y_{t_n}, A(Y_n^\pi - Y_{t_n}) \rangle] \\
&\leq 2e^{(2\beta+1)T} \left(\|\delta Y_0\|^2 + 2h\langle \delta Y_0, A\delta Y_0 \rangle + \frac{1}{h} \sum_{n=0}^{N-1} \mathbb{E}[\|e_{n+1}^1\|^2] + \left(1 + \frac{1}{\eta}\right) \sum_{n=0}^{N-1} \mathbb{E}[\|e_{n+1}^2\|^2] \right) \\
&\leq 2e^{(2\beta+1)T} \left(\|\delta Y_0\|^2 + 2h\langle \delta Y_0, A\delta Y_0 \rangle \right. \\
&\quad \left. + 4T \left[C_{\mu,1}^2 h + (\|A\|_{\text{op}}^2 + C_{\mu,1}^2) \sup_{s \in [t_n, t_{n+1}]} \left(\mathbb{E}[\|Y_{t_{n+1}} - Y_s\|^2] + \mathbb{E}[\|Y_s - Y_{t_n}\|^2] \right) \right] \right. \\
&\quad \left. + \left(1 + \frac{1}{\eta}\right) 2TC_{\sigma,1}^2 \left[h + \sup_{s \in [t_n, t_{n+1}]} \mathbb{E}[\|Y_s - Y_{t_n}\|^2] \right] \right) \\
&\leq C_{(\beta, \eta, T)} \left[\langle \delta Y_0, (I_d + hA)\delta Y_0 \rangle + (C_{\mu,1}^2 + C_{\sigma,1}^2)h \right. \\
&\quad \left. + (\|A\|_{\text{op}}^2 + C_{\mu,1}^2 + C_{\sigma,1}^2) \sup_{s \in [t_n, t_{n+1}]} \left(\mathbb{E}[\|Y_{t_{n+1}} - Y_s\|^2] + \mathbb{E}[\|Y_s - Y_{t_n}\|^2] \right) \right],
\end{aligned}$$

where $\delta Y_0 = ((I_d + hA)^{-1} - I_d)x_0$. Then, we can directly deduce the following inequality from Remark 4.1:

$$\langle \delta Y_0, (I_d + hA)\delta Y_0 \rangle = \langle -hA(I_d + hA)^{-1}x_0, -hAx_0 \rangle \leq \|hA(I_d + hA)^{-1}x_0\| \|hAx_0\| \leq h\|A\|_{\text{op}}\|x_0\|^2,$$

which, together with $h < 1$ and the time regularity of the solution $(Y_t)_t$ (Lemma 4.2), leads us to

$$\begin{aligned}
\max_{n=0,\dots,N} \mathbb{E}[\|Y_n^\pi - Y_{t_n}\|^2] &\leq C_{(\beta,\eta,T)} \left[h\|A\|_{\text{op}}\|x_0\|^2 + (C_{\mu,1}^2 + C_{\sigma,1}^2)h \right. \\
&\quad \left. + (\|A\|_{\text{op}}^2 + C_{\mu,1}^2 + C_{\sigma,1}^2)(1 + C_{\mu,0}^2 + C_{\sigma,0}^2 + \|x_0\|^2) \left(h^2(\|A\|_{\text{op}}^2 + C_{\mu,0}^2 + C_{\sigma,1}^2) + h(C_{\sigma,0}^2 + C_{\sigma,1}^2) \right) \right] \\
&\leq C_{(\beta,\eta,T)} h \left\{ \left[\|A\|_{\text{op}} + \left(\|A\|_{\text{op}}^2 + \sum_{l=0}^1 (C_{\mu,l}^2 + C_{\sigma,l}^2) \right) \right] \|x_0\|^2 \right. \\
&\quad \left. + \left[C_{\mu,1}^2 + C_{\sigma,1}^2 + \left(1 + C_{\mu,0}^2 + C_{\sigma,0}^2 \right) \left(\|A\|_{\text{op}}^2 + \sum_{l=0}^1 (C_{\mu,l}^2 + C_{\sigma,l}^2) \right) \right] \right\} \\
&\leq C_{(\beta,\eta,T)} h \left\{ \left(1 + \|A\|_{\text{op}}^2 + \sum_{l=0}^1 (C_{\mu,l}^2 + C_{\sigma,l}^2) \right)^2 \|x_0\|^2 + \left(1 + \|A\|_{\text{op}}^2 + \sum_{l=0}^1 (C_{\mu,l}^2 + C_{\sigma,l}^2) \right)^3 \right\}.
\end{aligned}$$

Finally, by further assuming $N \geq T \max(2\eta, 1/\eta)$, and using Corollary 4.5, we can conclude that:

$$\begin{aligned}
\max_{n=0,\dots,N} \mathbb{E}[\|\tilde{Y}_n^\pi - Y_{t_n}\|^2] &\leq 2 \max_{n=0,\dots,N} \left(\mathbb{E}[\|Y_n^\pi - Y_{t_n}\|^2] + \mathbb{E}[\|\tilde{Y}_n^\pi - Y_{t_n}^\pi\|^2] \right) \\
&\leq C_{(\beta,\eta,T)} h \left\{ \left(1 + \|A\|_{\text{op}}^2 + \sum_{l=0}^1 (C_{\mu,l}^2 + C_{\sigma,l}^2) \right)^2 \|x_0\|^2 + \left(1 + \|A\|_{\text{op}}^2 + \sum_{l=0}^1 (C_{\mu,l}^2 + C_{\sigma,l}^2) \right)^3 + \frac{\gamma^2}{h} \right\},
\end{aligned}$$

which completes the proof of the desired error estimate. \square

We end this section with the following weak convergence rate of the perturbed Euler scheme (4.3) with a perturbed terminal cost.

Theorem 4.8. *Suppose (H.5) holds. Let $N \in \mathbb{N}$ with $N \geq T \max\left(\frac{2\beta+2^{1/T}}{2^{1/T}-1}, \frac{1}{\eta}, 2\eta\right)$, $(Y_t)_{t \in [0,T]}$ be the solution to SDE (4.1), and $(\tilde{Y}_n^\pi)_{n=0}^N$ be the solution to PES (4.3). Then we have the following error estimate:*

$$\begin{aligned}
\max_{n=0,\dots,N} \|\mathbb{E}[f(Y_{t_n}) - \mathbb{E}[\tilde{f}_D(\tilde{Y}_n^\pi)]]\| &\leq C_{(\beta,\eta,T)} C_f \left\{ D^{\frac{-2\eta}{\eta+4}} (\|x_0\|^{2+\eta/2} + \|x_0\|^2 + C_{\mu,0}^2 + C_{\sigma,0}^2) \right. \\
&\quad \left. + Dh^{\frac{1}{2}} \left[\left(1 + \|A\|_{\text{op}}^2 + \sum_{l=0}^1 (C_{\mu,l}^2 + C_{\sigma,l}^2) \right) \|x_0\| + \left(1 + \|A\|_{\text{op}}^2 + \sum_{l=0}^1 (C_{\mu,l}^2 + C_{\sigma,l}^2) \right)^{\frac{3}{2}} + \gamma h^{-\frac{1}{2}} \right] \right\} + \theta.
\end{aligned}$$

Proof. The assumption (H.5(e)) implies that for all $n = 0, \dots, N$,

$$\begin{aligned}
\|\mathbb{E}[f(Y_{t_n}) - \mathbb{E}[\tilde{f}_D(\tilde{Y}_n^\pi)]]\| &\leq \mathbb{E}[|f(Y_{t_n}) - f_D(Y_{t_n})|] + \mathbb{E}[|f_D(Y_{t_n}) - f_D(\tilde{Y}_n^\pi)|] + \mathbb{E}[|f_D(\tilde{Y}_n^\pi) - \tilde{f}_D(\tilde{Y}_n^\pi)|] \\
&\leq C_f \mathbb{E}[\|Y_{t_n}\|^2 1_{B_\infty(D)^c}(Y_{t_n})] + C_f D \mathbb{E}[\|Y_{t_n} - \tilde{Y}_n^\pi\|] + \theta.
\end{aligned} \tag{4.13}$$

Now we bound the term $\mathbb{E}[\|Y_{t_n}\|^2 1_{B_\infty(D)^c}(Y_{t_n})]$. Since $\{\|x\| \leq D\} \subset B_\infty(D)$, we have $1_{B_\infty(D)^c}(Y_{t_n}) \leq 1_{\{\|x\| \geq D\}}(Y_{t_n})$. Thus for any given $p' \in (1, 1 + \eta/2)$, by using Hölder's inequality and Lemma 4.2, we obtain that

$$\begin{aligned}
\mathbb{E}[\|Y_{t_n}\|^2 1_{B_\infty(D)^c}(Y_{t_n})] &\leq \mathbb{E}[\|Y_{t_n}\|^{2p'}]^{\frac{1}{p'}} \mathbb{E}[1_{\{\|x\| \geq D\}}(Y_{t_n})]^{\frac{p'-1}{p'}} \leq \mathbb{E}[\|Y_{t_n}\|^{2p'}]^{\frac{1}{p'}} \left(\frac{\mathbb{E}[\|Y_{t_n}\|^2]}{D^2} \right)^{\frac{p'-1}{p'}} \\
&\leq \left(2^{p'-1} (\alpha_{2p'} + \|x_0\|^{2p'}) e^{2p'(\beta+1/2)T} \right)^{\frac{1}{p'}} D^{\frac{-2(p'-1)}{p'}} \left((\alpha_2 + \|x_0\|^2) e^{2(\beta+1/2)T} \right)^{\frac{p'-1}{p'}}
\end{aligned}$$

with the constant α_p defined as in (4.7) for all $p \in [2, 2 + \eta)$. Thus by choosing $p' = 1 + \eta/4$, we deduce from Young's inequality $xy \leq \frac{1}{p}x^p + \frac{1}{q}y^q$, $x, y \geq 0$, $p > 1$, $q = p/(p-1)$, that

$$\begin{aligned}\mathbb{E}[\|Y_{t_n}\|^2 1_{B_\infty(D)^c}(Y_{t_n})] &\leq C_{(\beta, \eta, T)} D^{-\frac{2(p'-1)}{p'}} (\alpha_{2p'} + \|x_0\|^{2p'})^{\frac{1}{p'}} (\alpha_2 + \|x_0\|^2)^{\frac{p'-1}{p'}} \\ &\leq C_{(\beta, \eta, T)} D^{-\frac{2(p'-1)}{p'}} \left(\alpha_{2p'} + \|x_0\|^{2p'} + \alpha_2 + \|x_0\|^2 \right) \\ &\leq C_{(\beta, \eta, T)} D^{-\frac{2(p'-1)}{p'}} (C_{\mu,0}^2 + C_{\sigma,0}^2 + \|x_0\|^{2p'} + \|x_0\|^2).\end{aligned}$$

Thus, by using (4.13) and Theorem 4.7, we obtain that

$$\begin{aligned}\|\mathbb{E}[f(Y_{t_n}) - \mathbb{E}[\tilde{f}_D(\tilde{Y}_n^\pi)]]\| &\leq C_f \mathbb{E}[\|Y_{t_n}\|^2 1_{B_\infty(D)^c}(Y_{t_n})] + C_f D \mathbb{E}[\|Y_{t_n} - \tilde{Y}_n^\pi\|^2]^{1/2} + \theta \\ &\leq C_{(\beta, \eta, T)} C_f \left\{ D^{\frac{-2\eta}{\eta+4}} (\|x_0\|^{2+\eta/2} + \|x_0\|^2 + C_{\mu,0}^2 + C_{\sigma,0}^2) + Dh^{\frac{1}{2}} \left[\left(1 + \|A\|_{\text{op}}^2 + \sum_{l=0}^1 (C_{\mu,l}^2 + C_{\sigma,l}^2) \right)^2 \|x_0\|^2 \right. \right. \\ &\quad \left. \left. + \left(1 + \|A\|_{\text{op}}^2 + \sum_{l=0}^1 (C_{\mu,l}^2 + C_{\sigma,l}^2) \right)^3 + \frac{\gamma^2}{h} \right]^{\frac{1}{2}} \right\} + \theta,\end{aligned}$$

which, along with the fact that $\psi(x) = x^{1/2}$ is subadditive on $[0, \infty)$, completes our proof. \square

5 Linear-implicit Euler discretizations for controlled SDEs

In this section, we extend the convergence analysis in Section 4 to SDEs controlled by a piecewise-constant deterministic strategy, whose coefficients are merely piecewise Hölder continuous in time. We shall establish that, similar to Theorem 4.8, the approximation error of the perturbed Euler scheme depends polynomially on the Lipschitz constant of the coefficients. Such error estimate will be used in Section 7 to establish the expression rate of DNN for value functions of zero-sum games.

We start by introducing the controlled SDE and its Euler approximations. Let $d, m, M \in \mathbb{N}$ and $x_0 \in \mathbb{R}^d$ be fixed. We consider the following SDE: $Y_0 = x_0$,

$$dY_t = (-AY_t + \mu(t, Y_t, u_k)) dt + \sigma(t, Y_t, u_k) dB_t, \quad t \in [\bar{t}_k, \bar{t}_{k+1}), \quad k = 0, \dots, M-1, \quad (5.1)$$

where $\bar{t}_k = kT/M$, $k = 0, \dots, M$, $U = \{u_k\}_{k=0}^{M-1} \subset \mathbb{R}^m$ is the set of control parameters, $\mu : [0, T] \times \mathbb{R}^d \times U \rightarrow \mathbb{R}^d$ and $\sigma : [0, T] \times \mathbb{R}^d \times U \rightarrow \mathbb{R}^{d \times d}$ are given functions, and $(B_t)_{t \in [0, T]}$ is a d -dimensional Brownian motion on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Now we introduce two Euler schemes for (5.1). For any given $N \in \mathbb{N}$, we shall consider the time stepsize $h = T/(NM)$, $t_n = nh$ for all $n = 0, \dots, NM$. We then consider the families of random variables $(Y_n^\pi)_{n=0}^{NM}$ and $(\tilde{Y}_n^\pi)_{n=0}^N$ defined as follows: $Y_0^\pi = \tilde{Y}_0^\pi = (I_d + hA)^{-1}x_0$, and for all $n = 0, \dots, NM-1$,

$$Y_{n+1}^\pi - Y_n^\pi + hAY_{n+1}^\pi = h\mu(t_n, Y_n^\pi, u_{\lfloor n/N \rfloor}) + \sigma(t_n, Y_n^\pi, u_{\lfloor n/N \rfloor})\Delta B_{n+1}, \quad (5.2)$$

$$\tilde{Y}_{n+1}^\pi - \tilde{Y}_n^\pi + hA\tilde{Y}_{n+1}^\pi = h\tilde{\mu}(t_n, \tilde{Y}_n^\pi, u_{\lfloor n/N \rfloor}) + \tilde{\sigma}(t_n, \tilde{Y}_n^\pi, u_{\lfloor n/N \rfloor})\Delta B_{n+1}, \quad (5.3)$$

where $\Delta B_{n+1} = B_{t_{n+1}} - B_{t_n}$ and the functions $\tilde{\mu}$ and $\tilde{\sigma}$ approximate μ and σ , respectively.

We shall assume the coefficients of the SDE (5.1) and the Euler schemes satisfy (H.5) uniformly with respect to the control parameter $(u_k)_{k=0}^{M-1}$, which is an analogue of (H.3) and (H.4) for the fixed d -dimensional problem.

H.6. Let $x_0 \in \mathbb{R}^d$, $U = \{u_k\}_{k=0}^{M-1}$, $A \in \mathbb{R}^{d \times d}$, $\eta, D > 0$ and $\beta, \gamma, \theta, C_{\mu,0}, C_{\mu,1}, C_{\sigma,0}, C_{\sigma,1}, C_f \geq 0$ be given constants. Let $\mu, \tilde{\mu} : [0, T] \times \mathbb{R}^d \times U \rightarrow \mathbb{R}^d$, $\sigma, \tilde{\sigma} : [0, T] \times \mathbb{R}^d \times U \rightarrow \mathbb{R}^{d \times d}$, $f, f_D, \tilde{f}_D : \mathbb{R}^d \rightarrow \mathbb{R}$ be measurable functions with the following properties:

- (a) For all $u \in U$, the matrix A and the functions $\mu(\cdot, \cdot, u), \tilde{\mu}(\cdot, \cdot, u) : [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}^d$, $\sigma(\cdot, \cdot, u), \tilde{\sigma}(\cdot, \cdot, u) : [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d}$ satisfy (H.5(a),(b),(c),(d)) with the constants $\eta, \beta, C_{\mu,0}, C_{\mu,1}, C_{\sigma,0}, C_{\sigma,1}, \gamma$.
- (b) The functions f, f_D, \tilde{f}_D satisfy (H.5(e)) with the constants C_f, D, θ .

Note that, by introducing the piecewise-constant function $\tilde{u} : [0, T] \rightarrow U$ such that $\tilde{u}(t) = u_{\lfloor tM/T \rfloor}$ for all $t \in [0, T]$, we can view (5.1)-(5.3) as (4.1)-(4.3) with coefficients:

$$\begin{aligned} \mu^u : (t, x) \in [0, T] \times \mathbb{R}^d &\mapsto \mu(t, x, \tilde{u}(t)) \in \mathbb{R}^d, & \sigma^u : (t, x) \in [0, T] \times \mathbb{R}^d &\mapsto \sigma(t, x, \tilde{u}(t)) \in \mathbb{R}^{d \times d}, \\ \tilde{\mu}^u : (t, x) \in [0, T] \times \mathbb{R}^d &\mapsto \tilde{\mu}(t, x, \tilde{u}(t)) \in \mathbb{R}^d, & \tilde{\sigma}^u : (t, x) \in [0, T] \times \mathbb{R}^d &\mapsto \tilde{\sigma}(t, x, \tilde{u}(t)) \in \mathbb{R}^{d \times d}, \end{aligned}$$

which, under (H.6), satisfy all conditions in (H.5) (with the same constants) except the 1/2-Hölder continuous in time on $[0, T]$. Consequently, we can deduce that Lemmas 4.1 and 4.2, Proposition 4.3, Corollaries 4.4 and 4.5 and Proposition 4.6 (with N replaced by NM in the statements) also hold for the solutions to (5.1)-(5.3), whose proofs do not rely on the time regularity of coefficients.

Now we extend Theorems 4.7 and 4.8 to establish strong and weak convergence rates for the perturbed Euler scheme (5.3).

Theorem 5.1. Suppose (H.6) holds. Let $N \in \mathbb{N}$ with $N \geq \frac{T}{M} \max\left(\frac{2\beta+2^{1/T}}{2^{1/T}-1}, \frac{1}{\eta}, 2\eta\right)$, $(Y_t)_{t \in [0, T]}$ be the solution to SDE (5.1), and $(\tilde{Y}_n^\pi)_{n=0}^{NM}$ be the solution to PES (5.3). Then $\max_{n=0, \dots, NM} \mathbb{E}[\|\tilde{Y}_n^\pi - Y_{t_n}\|^2]$ and $\max_{n=0, \dots, NM} |\mathbb{E}[f(Y_{t_n}) - \mathbb{E}[\tilde{f}_D(\tilde{Y}_n^\pi)]]|$ satisfy the same estimates as in Theorems 4.7 and 4.8, respectively.

Proof. Let $N \in \mathbb{N}$ with $N \geq \frac{T}{M} \max\left(\frac{2\beta+2^{1/T}}{2^{1/T}-1}, \frac{1}{\eta}, 2\eta\right)$ be fixed. Since Proposition 4.6 also holds for $(Y_t)_{t \in [0, T]}$ and $(\tilde{Y}_n^\pi)_{n=0}^{NM}$, a careful examination of the proofs of Theorems 4.7 and 4.8 shows that it suffices to estimate the local truncation errors e_{n+1}^1 and e_{n+1}^2 on $[t_n, t_{n+1}]$ for all $n = 0, \dots, NM-1$. Note that, for all $n = 0, \dots, NM-1$, the definitions of $(\bar{t}_k)_{k=0}^M$ and $(t_n)_{n=0}^{NM}$ ensure that $[t_n, t_{n+1}] \subset [\bar{t}_{\lfloor n/N \rfloor}, \bar{t}_{\lfloor n/N \rfloor + 1}]$, hence we have $\tilde{u}(t) = u_{\lfloor n/N \rfloor}$ for $t \in [t_n, t_{n+1}]$, which together with (H.6) implies that the (controlled) coefficients μ^u and σ^u of (5.1) are 1/2-Hölder continuous in time on each subinterval $[t_n, t_{n+1}]$ uniformly with respect to n . Consequently, it follows from the same arguments as in the proof of Theorem 4.7 that

$$\begin{aligned} \mathbb{E}[\|e_{n+1}^1\|^2] &\leq \mathbb{E}\left[\left(\int_{t_n}^{t_{n+1}} \|\mu(s, Y_s, u_{\lfloor n/N \rfloor}) - \mu(t_n, Y_{t_n}, u_{\lfloor n/N \rfloor})\| ds\right)^2\right] \\ &\leq 2h^2 \left[(\|A\|_{\text{op}}^2 + 2C_{\mu,1}^2) \sup_{s \in [t_n, t_{n+1}]} \left(\mathbb{E}[\|Y_{t_{n+1}} - Y_s\|^2] + \mathbb{E}[\|Y_s - Y_{t_n}\|^2] \right) + 2C_{\mu,1}^2 h \right], \\ \mathbb{E}[\|e_{n+1}^2\|^2] &= \mathbb{E}\left[\int_{t_n}^{t_{n+1}} \|\sigma(s, Y_s, u_{\lfloor n/N \rfloor}) - \sigma(t_n, Y_{t_n}, u_{\lfloor n/N \rfloor})\|^2 ds\right] \\ &\leq 2hC_{\sigma,1}^2 \left[h + \sup_{s \in [t_n, t_{n+1}]} \mathbb{E}[\|Y_s - Y_{t_n}\|^2] \right]. \end{aligned}$$

We can then proceed along the lines of Theorems 4.7 and 4.8 to obtain the desired error estimates. \square

6 Proofs of Theorem 2.1 and Corollary 2.2

This section is devoted to the proofs of Theorem 2.1 and Corollary 2.2.

We first prove Theorem 2.1. Given $d \in \mathbb{N}$ and $\varepsilon \in (0, 1]$, we consider $\delta \in (0, 1)$, $D > 1$, $N, M \in \mathbb{N}$, whose precise values will be specified later. Let $\kappa = \max(1, \kappa_0, \kappa_1)$ and $(B^m)_{m=1}^M$ be M independent copies of d -dimensional Brownian motions defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ supporting the Brownian motion $(B_t)_t$ in (2.2).² Then, for any given $x \in \mathbb{R}^d$, $m = 1, \dots, M$, let $(Y_n^{x,d,m,\pi})_{n=0}^N$ be the family of random variables defined by the following linear-implicit Euler scheme with perturbed coefficients $(\mu_d^\varepsilon, \sigma_d^\varepsilon)$ and the m -th Brownian motion $(B_t^m)_t$: $Y_0^{x,d,m,\pi} = (I_d + hA_d)^{-1}x$, and

$$Y_{n+1}^{x,d,m,\pi} - Y_n^{x,d,m,\pi} + hA_d Y_{n+1}^{x,d,m,\pi} = h\mu_d^\varepsilon(t_n, Y_n^{x,d,m,\pi}) + \sigma_d^\varepsilon(t_n, Y_n^{x,d,m,\pi})\Delta B_{n+1}^m, \quad (6.1)$$

where $h = T/N$ and $\Delta B_{n+1}^m = B_{(n+1)h}^m - B_{nh}^m$, for all $n = 0, \dots, N-1$.

The following lemma demonstrates that there exists a realization of the perturbed Euler scheme approximating the value function v_d globally with the desired accuracy.

Lemma 6.1. *Suppose the same assumptions of Theorem 2.1 hold. Then it holds for some constant $c > 0$, depending only on $\beta, \eta, \kappa_1, \kappa_2, \tau$ and T that: for any given $\varepsilon \in (0, 1)$, $d \in \mathbb{N}$, and for any $D, N, M = \mathcal{O}(\varepsilon^{-c}d^c)$, $\delta = \mathcal{O}(\varepsilon^c d^{-c})$, there exists a realization $\omega_{\varepsilon,d} \in \Omega$, such that*

$$\left(\int_{\mathbb{R}^d} \left| v_d(x) - \frac{1}{M} \sum_{m=1}^M f_{d,D}^\delta(Y_N^{x,d,m,\pi})(\omega_{\varepsilon,d}) \right|^2 \nu_d(dx) \right)^{1/2} \leq \varepsilon. \quad (6.2)$$

Proof of Lemma 6.1. Note that $(f_{d,D}^\delta(Y_N^{x,d,m,\pi}))_{m=1}^M$ are independent and identically distributed random variables. Hence, by using the definition of v_d and the weak uniqueness of the SDE (2.2), we can obtain that

$$\begin{aligned} & \int_{\mathbb{R}^d} \mathbb{E} \left[\left| \mathbb{E}[f_d(Y_T^{x,d})] - \frac{1}{M} \sum_{m=1}^M f_{d,D}^\delta(Y_N^{x,d,m,\pi}) \right|^2 \right] \nu_d(dx) \\ &= \int_{\mathbb{R}^d} \left| \mathbb{E}[f_d(Y_T^{x,d})] - \mathbb{E}[f_{d,D}^\delta(Y_N^{x,d,1,\pi})] \right|^2 + \mathbb{E} \left[\left| \mathbb{E}[f_{d,D}^\delta(Y_N^{x,d,1,\pi})] - \frac{1}{M} \sum_{m=1}^M f_{d,D}^\delta(Y_N^{x,d,m,\pi}) \right|^2 \right] \nu_d(dx) \\ &\leq \int_{\mathbb{R}^d} \left(\left| \mathbb{E}[f_d(Y_T^{x,d,1})] - \mathbb{E}[f_{d,D}^\delta(Y_N^{x,d,1,\pi})] \right|^2 + \frac{1}{M} \mathbb{E}[|f_{d,D}^\delta(Y_N^{x,d,1,\pi})|^2] \right) \nu_d(dx), \end{aligned} \quad (6.3)$$

where $(Y_t^{x,d,1})_{t \in [0,T]}$ is the solution to the SDE (2.2) driven by the Brownian motion $(B_t^1)_{t \in [0,T]}$.

We shall then estimate the two terms in (6.3) separately. Note that Theorem 4.8 implies that

²In general, suppose that $(B^m)_{m=1}^M$ are defined on a probability space $(\Omega^{(M)}, \mathcal{F}^{(M)}, \mathbb{P}^{(M)})$, one can extend the original probability space $(\Omega, \mathcal{F}, \mathbb{P})$ into the product space $(\bar{\Omega}, \bar{\mathcal{F}}, \bar{\mathbb{P}}) = (\Omega, \mathcal{F}, \mathbb{P}) \otimes (\Omega^{(M)}, \mathcal{F}^{(M)}, \mathbb{P}^{(M)})$, and perform the subsequent analysis on the full probability space with the measure $\bar{\mathbb{P}}$.

for all $N \geq T \max\left(\frac{2\beta+2^{1/T}}{2^{1/T}-1}, \frac{1}{\eta}, 2\eta\right)$, we have

$$\begin{aligned}
& |\mathbb{E}[f_d(Y_T^{x,d,1})] - \mathbb{E}[f_{d,D}^\delta(Y_N^{x,d,1,\pi})]| \leq C_{(\beta,\eta,T)} \kappa d^\kappa \left\{ D^{\frac{-2\eta}{\eta+4}} (\|x_0\|^{2+\eta/2} + \|x_0\|^2 + \kappa^2 d^{2\kappa}) \right. \\
& \quad \left. + Dh^{\frac{1}{2}} \left[\kappa^2 d^{2\kappa} \|x_0\| + (\kappa^2 d^{2\kappa})^{3/2} + (\delta \kappa d^\kappa) h^{-\frac{1}{2}} \right] \right\} + \delta \kappa d^\kappa D^\kappa \\
& \leq C_{(\beta,\eta,\kappa,T)} d^\kappa \left\{ d^{2\kappa} (D^{\frac{-2\eta}{\eta+4}} + Dh^{\frac{1}{2}}) \|x_0\| 1_{\{\|x_0\| \leq 1\}} + d^{2\kappa} (D^{\frac{-2\eta}{\eta+4}} + Dh^{\frac{1}{2}}) \|x_0\|^{2+\eta/2} 1_{\{\|x_0\| > 1\}} \right. \\
& \quad \left. + d^{3\kappa} (D^{\frac{-2\eta}{\eta+4}} + Dh^{\frac{1}{2}}) + \delta d^\kappa D^\kappa \right\}.
\end{aligned}$$

Now by letting $D^{\frac{-2\eta}{\eta+4}} \leq Dh^{\frac{1}{2}}$, i.e., $D \geq h^{-\frac{\eta+4}{6\eta+8}}$, we can deduce from the above estimate that

$$|\mathbb{E}[f_d(Y_T^{x,d,1})] - \mathbb{E}[f_{d,D}^\delta(Y_N^{x,d,1,\pi})]| \leq C_{(\beta,\eta,\kappa,T)} \left(d^{3\kappa} h^{\frac{\eta}{3\eta+4}} \|x_0\|^{2+\frac{\eta}{2}} 1_{\{\|x_0\| > 1\}} + d^{4\kappa} h^{\frac{\eta}{3\eta+4}} + \delta d^{2\kappa} D^\kappa \right).$$

Therefore, by squaring the above inequality and using the integrability condition of the probability measure ν_d , we obtain the following estimate:

$$\begin{aligned}
& \int_{\mathbb{R}^d} |\mathbb{E}[f_d(Y_T^{x,d,1})] - \mathbb{E}[f_{d,D}^\delta(Y_N^{x,d,1,\pi})]|^2 \nu_d(dx) \\
& \leq C_{(\beta,\eta,\kappa,\tau,T)} \left(d^{6\kappa} h^{\frac{2\eta}{3\eta+4}} \int_{\mathbb{R}^d} \|x_0\|^{4+\eta} \nu_d(dx) + d^{8\kappa} h^{\frac{2\eta}{3\eta+4}} + \delta^2 d^{4\kappa} D^{2\kappa} \right) \\
& \leq C_{(\beta,\eta,\kappa,\tau,T)} \left(d^{6\kappa+\max(\tau,2\kappa)} h^{\frac{2\eta}{3\eta+4}} + \delta^2 d^{4\kappa} D^{2\kappa} \right). \tag{6.4}
\end{aligned}$$

We then proceed to obtain an upper bound of the second term in (6.3). Note that (H.1(d)) and (H.2(c)) lead to the following linear growth condition: for all $x \in \mathbb{R}^d$,

$$|f_{d,D}^\delta(x)| \leq |f_{d,D}^\delta(x) - f_{d,D}(x)| + |f_{d,D}(x) - f_{d,D}(0)| + |f_{d,D}(0)| \leq \delta \kappa d^\kappa D^\kappa + \kappa d^\kappa (D\|x\| + 1).$$

Thus, we can obtain from Corollary 4.4 that

$$\begin{aligned}
\mathbb{E}[|f_{d,D}^\delta(Y_N^{x,d,1,\pi})|^2] & \leq C_{(\kappa)} (\delta^2 d^{2\kappa} D^{2\kappa} + d^{2\kappa} + d^{2\kappa} D^2 \mathbb{E}[\|Y_N^{x,d,1,\pi}\|^2]) \\
& \leq C_{(\beta,\eta,\kappa,T)} (\delta^2 d^{2\kappa} D^{2\kappa} + d^{2\kappa} + d^{2\kappa} D^2 (\|x\|^2 + d^{2\kappa} + \delta^2 d^{2\kappa})),
\end{aligned}$$

which, along with the following estimate

$$\int_{\mathbb{R}^d} \|x\|^2 \nu_d(dx) \leq \left(\int_{\mathbb{R}^d} \|x\|^{4+\eta} \nu_d(dx) \right)^{2/(4+\eta)} \leq \tau^{1/2} d^{\tau/2},$$

enables us to bound the second term in (6.3) by

$$\begin{aligned}
\frac{1}{M} \int_{\mathbb{R}^d} \mathbb{E}[|f_{d,D}^\delta(Y_N^{x,d,1,\pi})|^2] \nu_d(dx) & \leq \frac{1}{M} C_{(\beta,\eta,\kappa,\tau,T)} (\delta^2 d^{2\kappa} D^{2\kappa} + d^{2\kappa} + d^{2\kappa} D^2 (d^{\tau/2} + d^{2\kappa})) \\
& \leq C_{(\beta,\eta,\kappa,\tau,T)} (\delta^2 d^{2\kappa} D^{2\kappa} + d^{2\kappa+\max(\tau/2,2\kappa)} D^2 / M). \tag{6.5}
\end{aligned}$$

Therefore, under the conditions $N \geq T \max\left(\frac{2\beta+2^{1/T}}{2^{1/T}-1}, \frac{1}{\eta}, 2\eta\right)$ and $D = \lceil h^{-\frac{\eta+4}{6\eta+8}} \rceil$, we can deduce from the estimates (6.3), (6.4) and (6.5) that

$$\begin{aligned} & \int_{\mathbb{R}^d} \mathbb{E} \left[\left| \mathbb{E}[f_d(Y_T^{x,d})] - \frac{1}{M} \sum_{m=1}^M f_{d,D}^\delta(Y_N^{x,d,m,\pi}) \right|^2 \right] \nu_d(dx) \\ & \leq C_{(\beta,\eta,\kappa,\tau,T)} \left(d^{6\kappa+\max(\tau,2\kappa)} h^{\frac{2\eta}{3\eta+4}} + \delta^2 d^{4\kappa} h^{-\frac{(\eta+4)\kappa}{3\eta+4}} + d^{2\kappa+\max(\tau/2,2\kappa)} h^{-\frac{\eta+4}{3\eta+4}} / M \right). \end{aligned}$$

Consequently, by further assuming that

$$d^{6\kappa+\max(\tau,2\kappa)} h^{\frac{2\eta}{3\eta+4}} \leq C\varepsilon^2, \quad \delta^2 d^{4\kappa} h^{-\frac{(\eta+4)\kappa}{3\eta+4}} \leq C\varepsilon^2, \quad d^{2\kappa+\max(\tau/2,2\kappa)} h^{-\frac{\eta+4}{3\eta+4}} / M \leq C\varepsilon^2, \quad (6.6)$$

with $C = 1/(3C_{(\beta,\eta,\kappa,\tau,T)})$, we have $\int_{\mathbb{R}^d} \mathbb{E}[|v_d(x) - \frac{1}{M} \sum_{m=1}^M f_{d,D}^\delta(Y_N^{x,d,m,\pi})|^2] \nu_d(dx) < \varepsilon^2$, which implies the existence of $\omega \in \Omega$ satisfying (6.2). Finally, we complete the proof by observing that there exists a constant $c > 0$, depending only on $\beta, \eta, \kappa, \tau$ and T , such that (6.6) holds for all $D, N, M = \mathcal{O}(\varepsilon^{-c} d^c)$ and $\delta = \mathcal{O}(\varepsilon^c d^{-c})$. \square

We now complete the proof of Theorem 2.1. By fixing the realization $\omega_{\varepsilon,d} \in \Omega$ in Lemma 6.1, we can see that it suffices to show that the map $x \mapsto \frac{1}{M} \sum_{m=1}^M f_{d,D}^\delta(Y_N^{x,d,m,\pi})(\omega_{\varepsilon,d})$ can be represented by a neural network with the desired complexity.

We start by constructing a network for $x \mapsto f_{d,D}^\delta(Y_N^{x,d,m,\pi})(\omega_{\varepsilon,d})$ with a fixed m . Without loss of generality, we shall assume the networks $\phi_{\delta,d}^\mu$ and $(\phi_{\delta,d}^{\sigma,i})_{i=1}^d$ have more than one hidden layers, i.e., $L_{\delta,d} \geq 2$. Note that due to the fixed realization $\omega_{\varepsilon,d}$, the mapping $t \in [0, T] \rightarrow B_t^m(\omega_{\varepsilon,d}) \in \mathbb{R}^d$ is a deterministic function. Hence, for any $n = 0, \dots, N-1$, by letting $\mathbf{b}_{n+1}^m = \Delta B_{n+1}^m(\omega_{\varepsilon,d}) \in \mathbb{R}^d$, we can obtain from Lemma A.1 and the fact that the networks $(\phi_{\delta,d}^{\sigma,i})_{i=1}^d$ have the same architecture (see H.2(a)) that there exists a DNN $\phi_{\delta,d}^{\sigma,n+1} \in \mathcal{N}$, such that $\mathcal{L}(\phi_{\delta,d}^{\sigma,n+1}) = L_{\delta,d}$, $\dim(\phi_{\delta,d}^{\sigma,n+1}) = (d+1, dN_1^{\delta,d}, \dots, dN_{L_{\delta,d}-1}^{\delta,d}, d)$, $\mathcal{C}(\phi_{\delta,d}^{\sigma,n+1}) \leq d^2 \mathcal{C}(\phi_{\delta,d}^{\sigma,1})$ and $[\mathcal{R}_\varrho(\phi_{\delta,d}^{\sigma,n+1})](t, x) = \sigma_d^\delta(t, x) \mathbf{b}_{n+1}^m$ for all $(t, x) \in [0, T] \times \mathbb{R}^d$. Note that for all $n = 0, \dots, N-1$, the networks $\phi_{\delta,d}^{\sigma,n+1}$ and $\phi_{\delta,d}^\mu$ have the same depth, and all hidden layers of $\phi_{\delta,d}^{\sigma,n+1}$ have higher dimensions than those of the hidden layers of $\phi_{\delta,d}^\mu$.

Now we consider the following inductive argument. Suppose for any given $n = 0, \dots, N-1$, the mapping $x \mapsto Y_n^{x,d,m,\pi}(\omega_{\varepsilon,d})$ is the realization of a ReLU network $\psi_n^m \in \mathcal{N}$ such that the dimension of its last hidden layer satisfies $N_{\mathcal{L}(\psi_n^m)-1} \leq 2d + (d+1)N_{L_{\delta,d}-1}^{\delta,d}$. Then we can obtain from Proposition 3.2 (by letting $\phi_1 = \psi_n^m, \phi_2 = \phi_{\delta,d}^\mu, \phi_3 = \phi_{\delta,d}^{\sigma,n+1}$) that there exists a network $\tilde{\psi}_n^m \in \mathcal{N}$ with depth $\mathcal{L}(\tilde{\psi}_n^m) = \mathcal{L}(\psi_n^m) + L_{\delta,d} - 1$, such that for all $x \in \mathbb{R}^d$,

$$\begin{aligned} [\mathcal{R}_\varrho(\tilde{\psi}_n^m)](x) &= [\mathcal{R}_\varrho(\psi_n^m)](x) + h[\mathcal{R}_\varrho(\phi_{\delta,d}^\mu)](t_n, [\mathcal{R}_\varrho(\psi_n^m)](x)) + [\mathcal{R}_\varrho(\phi_{\delta,d}^{\sigma,n+1})](t_n, [\mathcal{R}_\varrho(\psi_n^m)](x)) \\ &= (Y_n^{x,d,m,\pi} + h\mu_d^\delta(t_n, Y_n^{x,d,m,\pi}) + \sigma_d^\delta(t_n, Y_n^{x,d,m,\pi})\Delta B_{n+1}^m)(\omega_{\varepsilon,d}). \end{aligned}$$

Moreover, the dimension of the last hidden layer of $\tilde{\psi}_n^m$ is given by $N_{\mathcal{L}(\tilde{\psi}_n^m)-1} = 2d + (d+1)N_{L_{\delta,d}-1}^{\delta,d}$ (see Remark 3.1), and the complexity satisfies

$$\mathcal{C}(\tilde{\psi}_n^m) \leq \mathcal{C}(\psi_n^m) + 4 \left(\max(\mathcal{C}(\phi_{\delta,d}^\mu), \mathcal{C}(\phi_{\delta,d}^{\sigma,n+1})) + \mathcal{C}(\phi_{d,2}^{\text{Id}}) \right)^3 \leq \mathcal{C}(\psi_n^m) + 4 \left(d \sum_{i=1}^d \mathcal{C}(\phi_{\delta,d}^{\sigma,i}) + \mathcal{C}(\phi_{d,2}^{\text{Id}}) \right)^3,$$

where $\phi_{d,2}^{\text{Id}}$ is the two-layer representation of d -dimensional identity function defined as in (A.1). Then, by the definition of the linear-implicit Euler scheme (6.1), we know $Y_{n+1}^{x,d,m,\pi}(\omega_{\delta,d}) = (I_d + hA_d)^{-1}[\mathcal{R}_\varrho(\tilde{\psi}_n^m)](x)$ for all $x \in \mathbb{R}^d$. Since the architecture of a network is invariant under an affine transformation, we have shown the mapping $x \mapsto Y_{n+1}^{x,d,m,\pi}(\omega_{\delta,d})$ is the realization of a network $\psi_{n+1}^m \in \mathcal{N}$, which still satisfies the induction hypothesis. Hence, by observing that $Y_0^{x,d,m,\pi}(\omega_{\delta,d}) = (I_d + hA_d)^{-1}[\mathcal{R}_\varrho(\phi_{d,1}^{\text{Id}})](x)$, we can conclude that there exists a network $\psi_N^m \in \mathcal{N}$ representing the function $x \mapsto Y_N^{x,d,m,\pi}(\omega_{\delta,d})$ with the complexity

$$\mathcal{C}(\psi_N^m) \leq \mathcal{C}(\phi_{d,1}^{\text{Id}}) + 4 \left(d \sum_{i=1}^d \mathcal{C}(\phi_{\delta,d}^{\sigma,i}) + \mathcal{C}(\phi_{d,2}^{\text{Id}}) \right)^3 (N+1).$$

Consequently, we can infer from Lemma A.4 that there exists a network $\psi_N^{f,m} \in \mathcal{N}$ representing the function $x \mapsto f_{d,D}^\delta(Y_N^{x,d,m,\pi})(\omega_{\varepsilon,d})$ with the complexity $\mathcal{C}(\psi_N^{f,m}) \leq 2(\mathcal{C}(\phi_{\delta,d,D}^f) + \mathcal{C}(\psi_N^m))$.

Finally, we observe that the Brownian path $t \mapsto B_t^m(\omega_{\varepsilon,d})$ only affects the above construction through the vectors $(\mathbf{b}_n^m)_{n=1}^N$, hence the architecture of the network $\psi_N^{f,m}$ (i.e. the depth and the dimensions of all layers) remains the same for each m . Therefore, we can obtain from Lemma A.1 that there exists a network $\psi_{\varepsilon,d}$ with the realisation $[\mathbb{R}_\varrho(\psi_{\varepsilon,d})](x) = \frac{1}{M} \sum_{m=1}^M f_{d,D}^\delta(Y_N^{x,d,m,\pi})(\omega_{\varepsilon,d})$ for all $x \in \mathbb{R}^d$. Moreover, we can estimate the complexity of $\psi_{\varepsilon,d}$ by using the facts $\mathcal{C}(\phi_{d,1}^{\text{Id}}) = d^2 + d$ and $\mathcal{C}(\phi_{d,2}^{\text{Id}}) = 4d^2 + 3d$ (see Lemma A.2), and the polynomial dependence of D, N, M, δ on ε and d (see Lemma 6.1):

$$\begin{aligned} \mathcal{C}(\psi_{\varepsilon,d}) &\leq M^2 \mathcal{C}(\psi_N^{f,m}) \leq 2M^2 \left(\mathcal{C}(\phi_{\delta,d,D}^f) + \mathcal{C}(\phi_{d,1}^{\text{Id}}) + 4 \left(d \sum_{i=1}^d \mathcal{C}(\phi_{\delta,d}^{\sigma,i}) + \mathcal{C}(\phi_{d,2}^{\text{Id}}) \right)^3 (N+1) \right) \\ &\leq 8M^2 \left(\kappa \delta^{-\kappa} d^\kappa D^\kappa + d^2 + d + (\kappa d^{\kappa+1} \delta^{-\kappa} + 4d^2 + 3d)^3 N \right) \leq c\varepsilon^{-c} d^c, \end{aligned}$$

for some constant $c > 0$, depending only on $\beta, \eta, \kappa_1, \kappa_2, \tau$ and T . Hence the proof of Theorem 2.1 is finished.

In the remaining part of this section, we shall prove Corollary 2.2, which essentially follows from Theorem 2.1 and the Feynman-Kac formula in [46, Theorem 2.2].

Proof of Corollary 2.2. Throughout this proof, let $d \in \mathbb{N}$ be a fixed natural number and C be a generic constant, which depends on the dimension d and may take a different value at each occurrence.

We first show the value function $v_d : \mathbb{R}^d \rightarrow \mathbb{R}$ defined in (2.1) satisfies $v_d(x) = u_d(0, x)$ for all $x \in \mathbb{R}^d$. Note that (H.1(c)) implies the coefficients of (2.2) are Lipschitz continuous and satisfies the estimate $\|\mu_d(t, x)\| + \|\sigma_d(t, x)\| \leq C(1 + \|x\|)$ for all $(t, x) \in [0, T] \times \mathbb{R}^d$. For any $x, y \in \mathbb{R}^d$, we can also obtain from (H.1(d)) that

$$\begin{aligned} |f_d(x)| &= |f_d(x) - f_{d,1}(x) + f_{d,1}(x) - f_{d,1}(0) + f_{d,1}(0) - f_d(0) + f_d(0)| \\ &\leq |f_d(x) - f_{d,1}(x)| + |f_{d,1}(x) - f_{d,1}(0)| + |f_{d,1}(0) - f_d(0)| + |f_d(0)| \\ &\leq C_d^f \|x\|^2 1_{B_\infty(1)^c}(x) + C_d^f \|x\| + 0 + C_d^f \leq C(1 + \|x\|^2), \\ |f_d(x) - f_d(y)| &= |f_{d,1+\|x\|_\infty+\|y\|_\infty}(x) - f_{d,1+\|x\|_\infty+\|y\|_\infty}(y)| \\ &\leq C_d^f (1 + \|x\|_\infty + \|y\|_\infty) \|x - y\| \leq C_d^f (1 + \|x\| + \|y\|) \|x - y\|, \end{aligned}$$

which together with the moment estimate of the SDE (2.2) (see Lemma 4.2) and the definition of the value function $v_d : \mathbb{R}^d \rightarrow \mathbb{R}$ (see (2.1)) leads to the fact that $|v_d(x)| \leq C(1 + \|x\|^2)$ for all $x \in \mathbb{R}^d$. Consequently, the Feynman-Kac formula in [46, Theorem 2.2] and the uniqueness of continuous viscosity solution to (2.4) with at most polynomial growth (see e.g. [32, Theorem 4.3]) implies $v_d(x) = u_d(0, x)$ for all $x \in \mathbb{R}^d$.

Now let $\nu_d : \mathcal{B}(\mathbb{R}^d) \rightarrow [0, 1]$ be the probability measure on \mathbb{R}^d defined as $\nu_d(A) := \lambda_d(A \cap [0, 1]^d)$, where λ_d is the Lebesgue measure on \mathbb{R}^d . Then the desired result follows directly from Theorem 2.1, and the fact that $\int_{\mathbb{R}^d} \|x\|^{4+\eta} \nu_d(dx) \leq d^{2+\eta/2}$ (see [16, Lemma 3.15]). \square

7 Proof of Theorem 2.3

This section is devoted to the proof of Theorem 2.3. For each $d \in \mathbb{N}$, it is clear that $v_d(x) = \inf_{u_1 \in \mathcal{U}_{1,d}} \sup_{u_2 \in \mathcal{U}_{2,d}} w_d(x; u_1, u_2)$ for all $x \in \mathbb{R}^d$, where the function $w_d : \mathbb{R}^d \times \mathcal{U}_{1,d} \times \mathcal{U}_{2,d} \rightarrow \mathbb{R}$ is defined as:

$$w_d(x; u_1, u_2) = \mathbb{E} \left[f_d(Y_T^{x,d,u_1,u_2}) + g_d(u_1, u_2) \right], \quad x \in \mathbb{R}^d, u_1 \in \mathcal{U}_{1,d}, u_2 \in \mathcal{U}_{2,d},$$

and $Y^{x,d,u_1,u_2} = (Y_t^{x,d,u_1,u_2})_{t \in [0,T]}$ is the solution to the following d -dimensional controlled SDE:

$$dY_t = (-A_d Y_t + \mu_d(t, Y_t, u_1, u_2)) dt + \sigma_d(t, Y_t, u_1, u_2) dB_t, \quad t \in (0, T]; \quad Y_0 = x. \quad (7.1)$$

In the following we shall first extend Theorem 2.1 to construct DNNs for the function w_d , and then construct DNNs to represent the value function v_d .

Let $d \in \mathbb{N}$, $u_1 \in \mathcal{U}_{1,d}$, $u_2 \in \mathcal{U}_{2,d}$ be fixed, i.e., there exists $M \in \mathbb{N}$, independent of d (see (H.3)), such that for $i = 1, 2$, we have $u_i(t) = u_i(\bar{t}_k) \in U_{i,d}$ on $[\bar{t}_k, \bar{t}_{k+1})$, where $\bar{t}_k = kT/M$ for all $k = 0, \dots, M$. Note that the essential steps to prove Theorem 2.1 are to study the convergence order of the linear-implicit Euler scheme with perturbed coefficients (see Theorem 4.8). Similarly, under the assumptions (H.3) and (H.4), we see the coefficients of the controlled SDE (7.1) satisfies (H.6), hence we can conclude from Theorem 5.1 that the same weak convergence rate also holds for the perturbed Euler scheme of (7.1) with a perturbed terminal cost.

Then, by following the same arguments as those in Section 6, we can deduce that there exists a constant $c > 0$, depending only on $\beta, \eta, \kappa_1, \kappa_2, \tau$ and T , such that for any given $d \in \mathbb{N}$, $\delta > 0$, $u_1 \in \mathcal{U}_{1,d}$, $u_2 \in \mathcal{U}_{2,d}$, one can construct a DNN $\psi_{\delta,d}^{u_1,u_2} \in \mathcal{N}$ with $\mathcal{C}(\psi_{\delta,d}^{u_1,u_2}) \leq cd^c \delta^{-c}$, and

$$\left(\int_{\mathbb{R}^d} |w_d(x; u_1, u_2) - [\mathcal{R}_\varrho(\psi_{\delta,d}^{u_1,u_2})](x)|^2 \nu_d(dx) \right)^{1/2} < \delta. \quad (7.2)$$

Moreover, the family of DNNs $(\psi_{\delta,d}^{u_1,u_2})_{u_1 \in \mathcal{U}_{1,d}, u_2 \in \mathcal{U}_{2,d}}$ has the same architecture (see Proposition 3.2, where the architecture of the constructed network ψ does not depend on the value of u).

Now suppose that $\varepsilon > 0$, $d \in \mathbb{N}$ are given, we shall construct a DNN to represent the value function v_d with an accuracy ε . We consider $\delta > 0$, whose value will be specified later, and construct the family of DNNs $(\psi_{\delta,d}^{u_1,u_2})_{u_1 \in \mathcal{U}_{1,d}, u_2 \in \mathcal{U}_{2,d}}$ to represent the functions $(w_d(\cdot; u_1, u_2))_{u_1 \in \mathcal{U}_{1,d}, u_2 \in \mathcal{U}_{2,d}}$ such that (7.2) holds for each $u_1 \in \mathcal{U}_{1,d}$, $u_2 \in \mathcal{U}_{2,d}$. Note that the number of intervention times M is a constant independent of d , and the cardinality of the set $U_d = U_{1,d} \times U_{2,d}$ are bounded by $\kappa_0 d^{\kappa_0}$ (see (H.3)). Hence the fact that all admissible control strategies are piecewise-constant in time implies that $|\mathcal{U}_{1,d}| \leq (\kappa_0 d^{\kappa_0})^M$ and $|\mathcal{U}_{2,d}| \leq (\kappa_0 d^{\kappa_0})^M$. Since $(\psi_{\delta,d}^{u_1,u_2})_{u_1 \in \mathcal{U}_{1,d}, u_2 \in \mathcal{U}_{2,d}}$ have the same architecture, we can apply Proposition 3.3 twice (with $n \in \mathbb{N}$ such that $n \geq \log_2(|\mathcal{U}_{i,d}|)$, $i = 1, 2$) and

deduce that there exists a DNN $\psi_{\delta,d}$ such that $[\mathcal{R}(\psi_{\delta,d})](x) = \inf_{u_1 \in \mathcal{U}_{1,d}} \sup_{u_2 \in \mathcal{U}_{2,d}} [\mathcal{R}(\psi_{\delta,d}^{u_1,u_2})](x)$ for all $x \in \mathbb{R}^d$, and the complexity of $\psi_{\delta,d}$ is bounded by

$$\mathcal{C}(\psi_{\delta,d}) \leq c(|\mathcal{U}_{1,d}| |\mathcal{U}_{2,d}|)^3 \mathcal{C}(\psi_{\delta,d}^{u_1,u_2}) \leq cd^c \delta^{-c},$$

for some constant $c > 0$ independent of d and δ (note that we have put the constant $\frac{34}{7}$ from Proposition 3.3 in the constant c , which is possible due to the fact that $\mathcal{C}(\psi_{\delta,d}^{u_1,u_2}) \geq 1$).

Finally, we specify the dependence of δ on the desired accuracy ε . Note that the following inequality holds for all parametrized functions $(f^{\alpha,\beta}, g^{\alpha,\beta})_{\alpha \in \mathcal{A}, \beta \in \mathcal{B}}$:

$$\left| \inf_{\alpha \in \mathcal{A}} \sup_{\beta \in \mathcal{B}} f^{\alpha,\beta}(x) - \inf_{\alpha \in \mathcal{A}} \sup_{\beta \in \mathcal{B}} g^{\alpha,\beta}(x) \right| \leq \sup_{\alpha \in \mathcal{A}, \beta \in \mathcal{B}} \left| f^{\alpha,\beta}(x) - g^{\alpha,\beta}(x) \right|, \quad x \in \mathbb{R}^d.$$

Thus we have

$$\begin{aligned} & \int_{\mathbb{R}^d} |v_d(x) - [\mathcal{R}_\varrho(\psi_{\delta,d})](x)|^2 \nu_d(dx) \\ &= \int_{\mathbb{R}^d} \left| \inf_{u_1 \in \mathcal{U}_{1,d}} \sup_{u_2 \in \mathcal{U}_{2,d}} w_d(x; u_1, u_2) - \inf_{u_1 \in \mathcal{U}_{1,d}} \sup_{u_2 \in \mathcal{U}_{2,d}} [\mathcal{R}(\psi_{\delta,d}^{u_1,u_2})](x) \right|^2 \nu_d(dx) \\ &\leq \int_{\mathbb{R}^d} \left(\sup_{(u_1, u_2) \in \mathcal{U}_{1,d} \times \mathcal{U}_{2,d}} |w_d(x; u_1, u_2) - [\mathcal{R}(\psi_{\delta,d}^{u_1,u_2})](x)| \right)^2 \nu_d(dx) \\ &= \int_{\mathbb{R}^d} \left(\sup_{(u_1, u_2) \in \mathcal{U}_{1,d} \times \mathcal{U}_{2,d}} |w_d(x; u_1, u_2) - [\mathcal{R}(\psi_{\delta,d}^{u_1,u_2})](x)|^2 \right) \nu_d(dx) \\ &\leq \int_{\mathbb{R}^d} \left(\sum_{(u_1, u_2) \in \mathcal{U}_{1,d} \times \mathcal{U}_{2,d}} |w_d(x; u_1, u_2) - [\mathcal{R}(\psi_{\delta,d}^{u_1,u_2})](x)|^2 \right) \nu_d(dx) \leq |\mathcal{U}_{1,d} \times \mathcal{U}_{2,d}| \delta^2. \end{aligned}$$

Since $|\mathcal{U}_{1,d} \times \mathcal{U}_{2,d}| \leq (\kappa_0 d^{\kappa_0})^M$, by choosing $\delta \leq \varepsilon (\kappa_0 d^{\kappa_0})^{-M/2}$, one can construct a DNN $\psi_{\varepsilon,d}$ with the desired accuracy and complexity, and finish the proof of Theorem 2.3.

8 Conclusions

To the best of our knowledge, this is the first paper which rigorously explains the success of DNNs in high-dimensional control problems with stiff systems, which arise naturally from Galerkin approximations of controlled PDEs and SPDEs (see e.g. [14, 33, 37, 36]). The main ingredient of our proof for DNN's polynomial expression rate is that the underlying stochastic dynamics can be effectively described by a suitable discrete-time system, whose specific realization leads us to the desired DNNs. Similar ideas can be easily extended to study optimal control problems of controlled jump diffusion processes with regime switching (see e.g. [22, 53]), which enables us to conclude that DNNs can overcome the curse of dimensionality in numerical approximations of weakly coupled systems of nonlocal PDEs.

Natural next steps would be to derive optimal expression rates of DNNs for control problems, and to construct DNNs for approximating value functions in stronger norms, such as L^p norms with $p > 2$, or Sobolev norms.

A Basic operations of ReLU DNNs

In this section, we collect several well-known results on the representation flexibility of DNNs.

The following lemma shows a linear combination of realizations of DNNs of the same architecture is again a realization of a DNN with the same activation function, whose proof can be found in [34, Lemma 5.1]. The result has been generalized to the case where the DNNs have the same length but different hidden layer dimensions in [29, Lemma 3.9].

Lemma A.1. *Let $\varrho \in C(\mathbb{R}; \mathbb{R})$, $L \in \mathbb{N}$, $M, N_0, N_1, \dots, N_L \in \mathbb{N}$, $(\beta_m)_{m=1}^M \in \mathbb{R}$, and $(\phi_m)_{m=1}^M \in \mathcal{N}$ be DNNs such that $\mathcal{L}(\phi_m) = L$ and $\dim(\phi_m) = (N_0, N_1, \dots, N_{L-1}, N_L)$ for all m . Then there exists $\psi \in \mathcal{N}$, such that $\mathcal{L}(\psi) = L$, $\mathcal{C}(\psi) \leq M^2 \mathcal{C}(\phi_1)$, $\dim(\psi) = (N_0, MN_1, \dots, MN_{L-1}, N_L)$ and*

$$[\mathcal{R}_\varrho(\psi)](x) = \sum_{m=1}^M \beta_m [\mathcal{R}_\varrho(\phi_m)](x), \quad x \in \mathbb{R}^{N_0}.$$

The next result proves that the identity function can be represented by a ReLU network, which is proved in [12, Lemma 5.3].

Lemma A.2. *Let $\varrho : \mathbb{R} \rightarrow \mathbb{R}$ be the activation function defined as $\varrho(x) = \max(0, x)$ for all $x \in \mathbb{R}$. For any $d, L \in \mathbb{N}$, consider the DNN $\phi_{d,L}^{\text{Id}} \in \mathcal{N}$ given by:*

$$\phi_{d,L}^{\text{Id}} = \begin{cases} \left(\left(\begin{pmatrix} I_d \\ -I_d \end{pmatrix}, 0 \right), \underbrace{(I_{2d}, 0), \dots, (I_{2d}, 0)}_{L-2 \text{ times}}, \left(\begin{pmatrix} I_d & -I_d \end{pmatrix}, 0 \right) \right), & L \geq 2, \\ ((I_d, 0)), & L = 1. \end{cases} \quad (\text{A.1})$$

Then we have $[\mathcal{R}_\varrho(\phi_{d,L}^{\text{Id}})](x) = x$ for all $x \in \mathbb{R}^d$.

Using the above representation of the identity function, one can extend a ReLU network to a network with arbitrary depth and widths of hidden layers without changing its realization.

Lemma A.3. *Let $\varrho : \mathbb{R} \rightarrow \mathbb{R}$ be the activation function defined as $\varrho(x) = \max(0, x)$ for all $x \in \mathbb{R}$. Let $L \in \mathbb{N}$ and $\phi \in \mathcal{N}$ be a DNN with $\mathcal{L}(\phi) < L$. Then there exists a DNN $\mathcal{E}_L(\phi) \in \mathcal{N}$ such that $\mathcal{L}(\mathcal{E}_L(\phi)) = L$, $[\mathcal{R}_\varrho(\mathcal{E}_L(\phi))](x) = [\mathcal{R}_\varrho(\phi)](x)$ for all $x \in \mathbb{R}^{\dim_{\text{in}}(\phi)}$, and*

$$\mathcal{C}(\mathcal{E}_L(\phi)) \leq 2(\mathcal{C}(\phi_{\dim_{\text{out}}(\phi), L-\mathcal{L}(\phi)}^{\text{Id}}) + \mathcal{C}(\phi)).$$

Moreover, let $L \in \mathbb{N} \cap [2, \infty)$, $N_0, N_1, \dots, N_L \in \mathbb{N}$ and $\phi \in \mathcal{N}_L^{N_0, N_1, \dots, N_{L-1}, N_L}$. Then for all $l \in \{1, \dots, L-1\}$, there exists $\mathcal{W}_l(\phi) \in \mathcal{N}_L^{N_0, N'_1, \dots, N'_{L-1}, N_L}$ such that $N'_l = N_l + 1$, $N'_i = N_i$ for all $i \in \{1, \dots, L-1\} \setminus \{l\}$, and $\mathcal{R}_\varrho(\mathcal{W}_l(\phi)) = \mathcal{R}_\varrho(\phi)$.

Proof. The properties of $\mathcal{E}_L(\phi)$ have been proved in [12, Lemma 5.3]. Now we assume $\phi = ((W_i, b_i))_{i=1}^L \in \mathcal{N}$, and construct the network $\mathcal{W}(\phi) = ((W'_i, b'_i))_{i=1}^L \in \mathcal{N}$ by $(W'_i, b'_i) = (W_i, b_i)$ for all $i \in \{1, \dots, L-1\} \setminus \{l, l+1\}$, and

$$\begin{aligned} W'_l &= \begin{pmatrix} W_l \\ 0 \end{pmatrix} \in \mathbb{R}^{(N_l+1) \times N_{l-1}}, \quad b'_l = \begin{pmatrix} b_l \\ 0 \end{pmatrix}; \\ W'_{l+1} &= (W_{l+1} \quad 0) \in \mathbb{R}^{N_{l+1} \times (N_l+1)}, \quad b'_{l+1} = b_{l+1}. \end{aligned}$$

Note that for all $x \in \mathbb{R}^{N_{l-1}}$, we have $W'_l x + b'_l = \begin{pmatrix} W_l x + b_l \\ 0 \end{pmatrix}$, and for all $x \in \mathbb{R}^{N_l}$, $y \in \mathbb{R}$, we have

$$W'_{l+1} \begin{pmatrix} x \\ y \end{pmatrix} + b'_{l+1} = W_{l+1} x + b_{l+1},$$

which implies $[\mathcal{R}_\varrho(\mathcal{W}_l(\phi))](x) = [\mathcal{R}_\varrho(\phi)](x)$ for all $x \in \mathbb{R}^{N_0}$. \square

We then recall the composition of two DNNs and the complexity of the resulting network (see [12, Lemma 5.3]).

Lemma A.4. *Let $\varrho : \mathbb{R} \rightarrow \mathbb{R}$ be the activation function defined as $\varrho(x) = \max(0, x)$ for all $x \in \mathbb{R}$. Let $(\phi_m)_{m=1}^2 \in \mathcal{N}$ be two DNNs such that*

$$\dim(\phi_m) = (N_0^{(m)}, N_1^{(m)}, \dots, N_{\mathcal{L}(\phi_m)-1}^{(m)}, N_{\mathcal{L}(\phi_m)}^{(m)}), \quad m = 1, 2,$$

and $\dim_{\text{in}}(\phi_1) = \dim_{\text{out}}(\phi_2)$, i.e., $N_0^{(1)} = N_{\mathcal{L}(\phi_2)}^{(2)}$. Then there exists a DNN $\psi \in \mathcal{N}$ such that $\mathcal{L}(\psi) = \mathcal{L}(\phi_1) + \mathcal{L}(\phi_2)$, $\mathcal{C}(\psi) \leq 2(\mathcal{C}(\phi_1) + \mathcal{C}(\phi_2))$, and $\mathcal{R}_\varrho(\psi)(x) = [\mathcal{R}_\varrho(\phi_1)]([\mathcal{R}_\varrho(\phi_2)](x))$ for all $x \in \mathbb{R}^{\dim_{\text{in}}(\phi_2)}$.

References

- [1] R. Arora, A. Basu, P. Mianjy, and A. Mukherjee, *Understanding deep neural networks with rectified linear units*, in International Conference on Learning Representations, 2018.
- [2] A. Barth and A. Lang, *Simulation of stochastic partial differential equations using finite element methods*, Stochastics, 84 (2012), pp. 217–231.
- [3] C. Beck, W. E, and A. Jentzen, *Machine learning approximation algorithms for high-dimensional fully nonlinear partial differential equations and second order backward stochastic differential equations*, arXiv:1709.05963, 2017. Accepted in J. Nonlinear Sci.
- [4] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*, Athena Scientific, Belmont, MA, 1996.
- [5] H. Bölcskei, P. Grohs, G. Kutyniok, and P. Petersen, *Optimal approximation with sparsely connected deep neural networks*, SIAM J. Math. Data Sci., 1 (2019), pp. 8–45.
- [6] H. Brezis, *Functional Analysis, Sobolev Spaces and Partial Differential Equations*, Springer, New York, 2011.
- [7] H. Bungartz and M. Griebel, *Sparse grids*, Acta Numerica, 13 (2004), pp. 147–269.
- [8] N. Bush, B. Hambly, H. Haworth, L. Jin, and C. Reisinger, *Stochastic evolution equations in portfolio credit modelling*, SIAM J. Financ. Math., 2 (2011), 627–664.
- [9] D. Kalise and K. Kunisch, *Polynomial approximation of high-dimensional Hamilton–Jacobi–Bellman equations and applications to feedback control of semilinear parabolic PDEs*, SIAM J. Sci. Comput., 40 (2018), A629–A652.
- [10] T. G. Kurtz and J. Xiong, *Particle representations for a class of nonlinear SPDEs*, Stoch. Proc. Appl., 83 (1999), 103–126.
- [11] W. E and Q. Wang, *Exponential convergence of the deep neural network approximation for analytic functions*, Science China Mathematics (2018). Published online: September 6, 2018, <https://doi.org/10.1007/s11425-018-9387-x>.
- [12] D. Elbrächter, P. Grohs, A. Jentzen, and C. Schwab, *DNN expression rate analysis of high-dimensional PDEs: application to option pricing*, arXiv:1809.07669, 2018.

- [13] W. Fang and M.B. Giles, *Adaptive Euler-Maruyama method for SDEs with non-globally Lipschitz drift: Part I, finite time interval*, arXiv:1609.08101, 2016.
- [14] M. Farhood and G. E. Dullerud, *Control of systems with uncertain initial conditions*, IEEE Trans. Automat. Control, 53 (2008), pp. 2646–2651.
- [15] M. B. Giles and C. Reisinger, *Stochastic finite differences and multilevel Monte Carlo for a class of SPDEs in finance*, SIAM J. Financial Math., 3 (2012), pp. 572–592.
- [16] P. Grohs, F. Hornung, A. Jentzen, and P. von Wurstemberger, *A proof that artificial neural networks overcome the curse of dimensionality in the numerical approximation of Black-Scholes partial differential equations*, arXiv:1809.02362, 2018.
- [17] P. Grohs, D. Perekrestenko, D. Elbrächter, and H. Bölskei, *Deep neural network approximation theory*, Technical Report, 1901.02220, arXiv, 2019.
- [18] I. Gühring, G. Kutyniok, and P. Petersen, *Error bounds for approximations with deep ReLU neural networks in $W^{s,p}$ norms*, arXiv:1902.07896, 2019.
- [19] I. Gyöngy and A. Millet, *Rate of convergence of space time approximations for stochastic evolution equations*, Potential Anal., 30 (2009), pp. 29–64.
- [20] J. Han and W. E, *Deep learning approximation for stochastic control problems*, Deep Reinforcement Learning Workshop, NIPS 2016, arXiv:1611.07422, 2016.
- [21] J. He, L. Li, J. Xu, and C. Zheng, *ReLU deep neural networks and linear finite elements*, arXiv:1807.03973, 2018.
- [22] D. J. Higham and P. E. Kloeden, *Convergence and stability of implicit methods for jump-diffusion systems*, Int. J. Numer. Anal. Model., 3 (2006), pp. 125–140.
- [23] K. Hornik, M. Stinchcombe, and H. White, *Multilayer feedforward networks are universal approximators*, Neural Netw., 2 (1989), pp. 359–366.
- [24] J. Hu, S. Jin, and D. Xiu, *A stochastic Galerkin method for Hamilton–Jacobi equations with uncertainty*, SIAM J. Sci. Comput., 37 (2015), A2246–A2269.
- [25] C. Huré, H. Pham, and X. Warin, *Some machine learning schemes for high-dimensional nonlinear PDEs*. arXiv:1902.01599, 2019.
- [26] M. Hutzenthaler, A. Jentzen, and P. E. Kloeden, *Strong convergence of an explicit numerical method for SDEs with non-globally Lipschitz continuous coefficients*, Ann. Appl. Probab., 22 (2012), pp. 1611–1641.
- [27] M. Hutzenthaler, A. Jentzen, T. Kruse, T. A. Nguyen, and P. von Wurstemberger, *Overcoming the curse of dimensionality in the numerical approximation of semilinear parabolic partial differential equations*, arXiv:1807.01212, 2018.
- [28] M. Hutzenthaler, A. Jentzen, and T. Kruse, *Overcoming the curse of dimensionality in the numerical approximation of parabolic partial differential equations with gradient-dependent nonlinearities*, arXiv:1912.02571, (2019).
- [29] M. Hutzenthaler, A. Jentzen, T. Kruse, and T. A. Nguyen, *A proof that rectified deep neural networks overcome the curse of dimensionality in the numerical approximation of semilinear heat equations*, arXiv:1901.10854, 2019.

- [30] K. Ito, *Approximation of the Zakai equation for nonlinear filtering*, SIAM J. Control Optim., 34 (1996), pp. 620–634.
- [31] K. Ito, C. Reisinger, and Y. Zhang, *A neural network based policy iteration algorithm with global H^2 -superlinear convergence for stochastic games on domains*, arXiv:1906.02304, 2019.
- [32] E. R. Jakobsen and K. H. Karlsen, *Continuous dependence estimates for viscosity solutions of integro-PDEs*, J. Differential Equations, 212 (2005), pp. 278–318.
- [33] A. Jentzen and P. E. Kloeden, *Taylor Approximations for Stochastic Partial Differential Equations*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 83, SIAM, Philadelphia, 2011.
- [34] A. Jentzen, D. Salimova, and T. Welti, *A proof that deep artificial neural networks overcome the curse of dimensionality in the numerical approximation of Kolmogorov partial differential equations with constant diffusion and nonlinear drift coefficients*, arXiv:1809.07321, 2018.
- [35] Y. LeCun, Y. Bengio, and G. Hinton, *Deep learning*, Nature, 521 (2015), pp. 436–444.
- [36] B. Luo, H.-N. Wu, and H.-X. Li, *Adaptive optimal control of highly dissipative nonlinear spatially distributed processes with neuro-dynamic programming*, IEEE Trans. Neural Netw. Learn. Syst., 26 (2015), pp. 684–696.
- [37] B. Luo and H.-N. Wu, *Approximate optimal control design for nonlinear one-dimensional parabolic PDE systems using empirical eigenfunctions and neural network*, IEEE Trans. Syst., Man, Cybern. B, Cybern., 42 (2012), pp. 1538–1549.
- [38] H. N. Mhaskar and T. Poggio, *Deep vs. shallow networks: An approximation theory perspective*, Anal. Appl., 14 (2016), pp. 829–848.
- [39] X. Mao, *Stochastic Differential Equations and Applications*, Horwood, Chichester, 1997.
- [40] X. Mao and L. Szpruch, *Strong convergence and stability of implicit numerical methods for stochastic differential equations with non-globally Lipschitz continuous coefficients*, J. Comput. Appl. Math., 238 (2013), pp. 14–28.
- [41] P. Merino, I. Neitzel, and F. Tröltzsch, *Error estimates for the finite element discretization of semi-infinite elliptic optimal control problems*, Discuss. Math. Diff. Incl. Control Optim., 30 (2010), pp. 221–236.
- [42] P. Merino, F. Tröltzsch, and B. Vexler, *Error estimates for the finite element approximation of a semilinear elliptic control problem with state constraints and finite dimensional control space*, M2AN Math. Model. Numer. Anal., 44 (2010), pp. 167–188.
- [43] V. Mnih, K. Kavukcuoglu, D. Silver, A. Rusu, J. Veness, M. Bellemare, A. Graves, M. Riedmiller, A. Fidjeland, G. Ostrovski, and S. Petersen, *Human-level control through deep reinforcement learning*, Nature, 518 (2015), pp. 529–533.
- [44] H. Montanelli and Q. Du, *New error bounds for deep networks using sparse grids*, SIAM J. Math. Data Sci., 1 (2019), pp. 78–92.
- [45] J. Opschoor, P. Petersen, and C. Schwab, *Deep ReLU networks and high-order finite element methods*, SAM Report, ETH Zürich, 2019.

- [46] E. Pardoux, *Backward stochastic differential equations and viscosity solutions of systems of semilinear parabolic and elliptic PDEs of second order*, in L. Decreusefond et al. eds. *Stochastic Analysis and related Topics VI, The Geilo Workshop, 1996*, Birkhäuser 1998, pp. 79–127.
- [47] P. Petersen and F. Voigtlaender, *Optimal approximation of piecewise smooth functions using deep ReLU neural networks*, *Neural Netw.*, 108 (2018), pp. 296–330.
- [48] H. Pham and X. Warin, *Neural networks-based backward scheme for fully nonlinear PDEs*, arXiv:1908.00412, 2019.
- [49] C. Schwab and C. Gittelsohn, *Sparse tensor discretizations of high-dimensional parametric and stochastic PDEs*, *Acta Numerica*, 20 (2011), pp. 291–467.
- [50] C. Schwab and J. Zech, *Deep learning in high dimension: Neural network expression rates for generalized polynomial chaos expansions in UQ*, *Anal. Appl.*, 17 (2019), pp. 19–55.
- [51] N. Touzi, *Optimal Stochastic Control, Stochastic Target Problems, and Backward SDE*, Springer, New York, 2012.
- [52] D. Yarotsky, *Error bounds for approximations with deep ReLU networks*, *Neural Netw.*, 94 (2017), pp. 103–114.
- [53] G. Yin and C. Zhu, *Hybrid Switching Diffusions: Properties and Applications*, Springer, New York, 2010.