Finance and Stochastics
https://doi.org/10.1007/s00780-024-00549-x

Importance sampling for option pricing with feedforward neural networks

Aleksandar Arandjelović[1] · Thorsten Rheinländer[1] · Pavel V. Shevchenko[2]

Received: 6 January 2022 / Accepted: 24 September 2023

Abstract
We study the problem of reducing the variance of Monte Carlo estimators through performing suitable changes of the sampling measure computed by feedforward neural networks. To this end, building on the concept of vector stochastic integration, we characterise the Cameron–Martin spaces of a large class of Gaussian measures induced by vector-valued continuous local martingales with deterministic covariation. We prove that feedforward neural networks enjoy, up to an isometry, the universal approximation property in these topological spaces. We then prove that sampling measures generated by feedforward neural networks can approximate the optimal sampling measure arbitrarily well. We conclude with a comprehensive numerical study pricing path-dependent European options for asset price models that incorporate factors such as changing business activity, knock-out barriers, dynamic correlations and high-dimensional baskets.

Keywords Cameron–Martin space · Doléans–Dade exponential · Feedforward neural networks · Importance sampling · Universal approximation

Mathematics Subject Classification 60G15 · 65B99 · 65C05 · 68T07 · 91G20 · 91G60

JEL Classification C15 · C45 · C63

✉ A. Arandjelović
aleksandar.arandjelovic@fam.tuwien.ac.at

T. Rheinländer
thorsten.rheinlander@fam.tuwien.ac.at

P.V. Shevchenko
pavel.shevchenko@mq.edu.au

[1] Institute of Statistics and Mathematical Methods in Economics, TU Wien, Vienna, Austria

[2] Department of Actuarial Studies and Business Analytics, Macquarie University, Sydney, Australia

Published online: 11 November 2024

1 Introduction

Monte Carlo methods are amongst the most essential tools for the numerical valuation of financial derivatives. Classical asset pricing theory often calls for the computation of expectations of the form

$$\mathbb{E}_{\mathbb{P}}[F(X)] = \int_{\Omega} F\big(X(\omega)\big)\mathbb{P}(\mathrm{d}\omega),$$

where F is a payoff functional, X is an asset price process solving a stochastic differential equation (SDE) of the form $\mathrm{d}X_t = a(X)\,\mathrm{d}C_t + b(X)\,\mathrm{d}M_t$ for $t \in [0, T]$ with a finite time horizon $T > 0$ for some potentially path-dependent coefficients a, b, a process C of locally finite variation, and a local martingale M, and $\mathbb{P}$ is a probability measure on a measurable space $(\Omega, \mathcal{F})$. By averaging the payoffs over randomly sampled trajectories of X, one can estimate the price in many cases where no analytic solution for $\mathbb{E}_{\mathbb{P}}[F(X)]$ is available.

The variance of the Monte Carlo estimator is inversely proportional to the number of trajectories simulated and proportional to the variance of the option payoff. The square root of this variance is referred to as the standard error and, in principle, it can be made as small as needed by simulating a sufficiently high number of trajectories. However, given limitations on computational time, the error can still be too large to be acceptable, especially for further calculations of option price derivatives (the so-called Greeks) required for hedging and risk management.

The usage of variance reduction methods can drastically reduce this error. There are many different methods to reduce the variance of Monte Carlo estimators, one of which is importance sampling. This method is based on changing the sampling measure from which the trajectories are generated from $\mathbb{P}$ to some equivalent measure $\mathbb{P}_h$, thereby overweighting important scenarios to increase the numerical efficiency of the estimates. Due to Girsanov's theorem, this corresponds to adding a drift $h \in H$ to the process M, where H denotes a prescribed space of functions or processes from which the drift adjustment is chosen. One then writes

$$\mathbb{E}_{\mathbb{P}}[F(X)] = \mathbb{E}_{\mathbb{P}_h}[F(X)Z_h^{-1}],$$

where Z_h denotes a Radon–Nikodým density of $\mathbb{P}_h$ with respect to $\mathbb{P}$ which also depends on the time horizon T. Instead of simulating realisations of $F(X)$ with respect to $\mathbb{P}$, one then simulates realisations of $F(X)Z_h^{-1}$ with respect to $\mathbb{P}_h$ and chooses $h \in H$ such that the variance of $F(X)Z_h^{-1}$ is minimised.

While this method usually requires a lot of specific knowledge about the model at hand, it has the potential to drastically reduce the variance of the corresponding Monte Carlo estimator. In other words, importance sampling is a powerful method that involves the complex optimisation problem of choosing an appropriate sampling measure which minimises the variance of the Monte Carlo estimators.

Neural networks provide an algorithmically generated class of functions which, on the one hand, enjoy the universal approximation property in many different topological spaces, meaning that they are dense in these spaces, and, on the other hand, can be trained in a numerically efficient way. Having recently entered the realm of

mathematical finance, neural networks are successfully used e.g. for model calibration, hedging and pricing. This paper develops a method that uses feedforward neural networks to perform importance sampling for complex stochastic models, which applies in particular to the valuation of path-dependent derivatives. By optimising over drifts from a dense subspace $H(D) \subseteq H$ generated by a set D of feedforward neural networks, we obtain a tractable problem where the usage of feedforward neural networks can be theoretically justified, and where the optimisation can be carried out in a numerically efficient way.

1.1 Outline of the paper and main results

In Sect. 2, we characterise tractable spaces H from which the drift adjustments may be chosen, and study their analytic properties. Whenever M is a vector-valued continuous local martingale with deterministic covariation, it induces a Gaussian measure, to which one can assign a Hilbert space H, the Cameron–Martin space. Due to the multivariate nature of our study, we recall in Lemma 2.10 some concepts which originate from the theory of stochastic integration with respect to vector-valued semimartingales. A detailed characterisation of the corresponding Cameron–Martin space H that is induced by M is provided in Proposition 2.18, where the general formulation allows us to specifically incorporate intricate and in particular time-inhomogeneous covariance patterns for M into our models. Theorem 2.24 then yields the essential approximation result that characterises dense subspaces $H(D)$ of H which are generated by prescribed sets of functions D in an abstract and general setting, and in particular applies to sets D of feedforward neural networks as a special case.

In Sect. 3, we focus our attention on feedforward neural networks, where we distinguish between neural networks of deep, narrow and shallow kind. Propositions 3.4 and 3.5 yield two approximation results which provide a theoretical justification for considering sets D that consist of feedforward neural networks. Example 3.7 then shows in a classical setting that the set $H(D)$ which is generated by a set D of feedforward neural networks has an explicit and tractable characterisation. As a direct consequence of Theorem 2.24, Sect. 3.1 then discusses a result which in particular implies that every smooth function can, up to an isometry, be approximated by feedforward neural networks arbitrarily well with respect to Hölder-type topologies, which are stronger than the topology of uniform convergence.

Section 4 contains a detailed study of the importance sampling problem, where we aim to minimise the variance of $F(X)Z_h^{-1}$ with respect to $\mathbb{P}_h$ by approximating the optimal drift h with a feedforward neural network. Theorem 4.6 proves that the functional $V\colon H \to \mathbb{R}_+$ which needs to be minimised is, under suitable generic assumptions, continuous and admitting a minimiser $h^* \in H$, which can be approximated, up to an isometry, arbitrarily well by feedforward neural networks. Here, we not only prove convergence to the optimal drift h^*, but also show that the corresponding Radon–Nikodým densities converge to Z_{h^*}. To this end, we prove that feedforward neural networks induce equivalent probability measures whose densities with respect to the original measure converge in L^p-spaces; see Lemma 4.5. Moreover, Sect. 4.1 contains a discussion of a classical importance sampling approach that utilises results

from the theory of large deviations, where we show that feedforward neural networks can be employed to solve the corresponding variational problem which appears in this approach. Let us note that while the results from Sects. 2 and 3 are applied to importance sampling in Sect. 4, they are also of independent interest.

Section 5 contains a comprehensive numerical study pricing path-dependent European options for asset price models that incorporate factors such as changing business activity, knock-out barriers, dynamic correlations and high-dimensional baskets. To conclude, we summarise our findings and give an outlook on future work in Sect. 6. The Appendix contains a brief glimpse at the theory of Gaussian measures and collects the proofs of all results.

1.2 Related literature

The line of research which eventually led up to the present work originates from Glasserman et al. [21]. The authors study the problem of pricing path-dependent options by using techniques from the theory of large deviations to perform a change of sampling measure that reduces the standard error of the Monte Carlo estimator. Moreover, they use stratified sampling in order to further improve their simulations.

The main motivation for this work was provided by Guasoni and Robertson [24]. As in [21], the authors employ methods from the theory of large deviations to obtain a variational problem whose solution yields an asymptotically optimal drift adjustment. The main difference to the present work is that we do not pass to a small noise limit. However, as it turns out, our method also complements the method presented in [24]; see Sect. 4.1 for further details.

An extension of the methods used in [24] to the study of importance sampling for stochastic volatility models has been provided in Robertson [50]. Note that our method applies to these types of models as well; see Example 4.2 in Sect. 4 and Sect. 5, where we provide simulation results for several stochastic volatility models.

Another interesting contribution is by dos Reis et al. [48] who study importance sampling for McKean–Vlasov SDEs. Similarly as in [24, 50], methods from the theory of large deviations yield an asymptotically optimal drift adjustment, and the authors discuss two different methods for the simulation of the solution to the McKean–Vlasov SDE under a change of measure.

The idea to use methods from the theory of stochastic approximation for the purpose of importance sampling has been studied extensively in Lemaire and Pagès [35]. This paper heavily influenced Sect. 4; especially the proof of Theorem 4.6 relies partially on a straightforward extension of the proof of [35, Proposition 4]. Let us also note that while the setting of [35, Sect. 3] could be extended to our setting below, it might be of particular interest to understand how the algorithm proposed in [35, Theorem 4] could be adapted to the setting of Sect. 4 below in order to yield convergence of the stochastic gradient descent algorithm when training feedforward neural networks.

Finally, one very original contribution that studies measure changes which are induced by neural networks for the purpose of Monte Carlo simulations is Müller et al. [44]. The authors also study importance sampling and apply their results to light-transport simulations. The main difference to [44] is that our method applies to the

pricing of financial derivatives in a mathematically more natural way by using methods from the theory of stochastic calculus. Here, we focus on neural networks that are of feedforward type. For more details on the studied neural network architectures and related literature, see Sect. 3.

Notation Unless stated otherwise, we endow $\mathbb{R}^d$ for each $d \in \mathbb{N}$ with the corresponding Euclidean norm $|\cdot|$; I_d denotes the identity matrix in $\mathbb{R}^{d\times d}$, and we write $\overline{\mathbb{R}}_+ = \mathbb{R}_+ \cup \{+\infty\}$. Given two vectors x, y of the same dimension, $x \odot y$ denotes their Hadamard (i.e., coordinatewise) product and $\langle x, y\rangle$ their inner (scalar) product. If Σ is a matrix, $\Sigma^\top$ is its transpose. For $x \in \mathbb{R}^d_+$ and $p > 0$, we understand x^p to hold componentwise and write $\sqrt{x}$ if $p = 1/2$. We also convene that $\inf \emptyset = \infty$. For each continuous linear operator A between normed spaces, $\|A\|_{\mathrm{op}}$ denotes its operator norm. If E_1, E_2 are two metric spaces and D is a subset of E_1, we say that D is dense in E_2 up to an isometry if there exists an isometry $J\colon E_1 \to E_2$ such that $J(D)$ is dense in E_2. If H is a Hilbert space, the notation $H^* \cong H$ indicates that we identify H^* with H via the isometric isomorphism given by the Fréchet–Riesz representation theorem.

Given a topological space $(S, \mathcal{T})$, we denote by S^* and S' the topological and algebraic dual spaces, respectively, and write $\mathcal{B}_S$ for the Borel σ-algebra on S. For $F \in \mathcal{T}$, we denote by F° and $\overline{F}$ the interior and closure of F, respectively. Whenever ν is a Borel measure on S, we say that $f\colon S \to \mathbb{R}^d$ is locally ν-essentially bounded if $(\nu\text{-})\operatorname{ess\,sup}_{x\in K} |f(x)| < \infty$ for each compact $K \subseteq S$. For an interval $[0, T]$, the space $C_0([0, T]; \mathbb{R}^d)$ consists of all $\mathbb{R}^d$-valued continuous functions on $[0, T]$ that vanish at the origin.

For a measurable space $(S, \mathcal{S})$, we denote by $\mathcal{L}^0(\mathcal{S}; \mathbb{R}^d)$ the space of $\mathbb{R}^d$-valued $\mathcal{S}$-measurable functions on S. If μ is a measure on $\mathcal{S}$ and $f \in \mathcal{L}^0(\mathcal{S}) = \mathcal{L}^0(\mathcal{S}; \mathbb{R})$, we denote by $f \cdot \mu$ the Lebesgue integral of f with respect to μ, provided it exists. For $p > 0$, we further denote by $L^p(\mu)$ the space of equivalence classes of p-integrable functions from $\mathcal{L}^0(\mathcal{S})$. The law of a random variable Z is denoted by $\mathcal{L}(Z)$. If M is an $\mathbb{R}^d$-valued semimartingale and $H \in L(M)$, we denote by $H \bullet M$ the stochastic integral of H with respect to M. We denote by $\mathcal{N}(m, \Sigma)$ the normal distribution with expected value $m \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$. Finally, we denote for each $p \geq 1$ by $\mathcal{H}^p$ the Banach space of continuous L^p-integrable martingales $M = (M_t)_{t\in[0,T]}$, where the dependency on the underlying filtered probability space is implicit.

2 Universal approximation in Cameron–Martin space

In this section, we study a tractable space H whose elements will be used to adjust the drift of M for the purpose of importance sampling in Sects. 4 and 5 below. Moreover, we identify dense linear subspaces of H and obtain an explicit characterisation of the Cameron–Martin spaces of a large class of Gaussian measures, which is of independent interest. For details about Gaussian measures, we refer to Appendix A.

Let $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ with $\mathbb{F} = (\mathcal{F}_t)_{t\in[0,T]}$ denote a filtered probability space such that $\mathbb{F}$ is right-continuous and $\mathcal{F}_0$ contains all $\mathbb{P}$-nullsets of $\mathcal{F}$. As index set for the time

parameter, we consider $[0, T]$ with a finite time horizon $T > 0$. Without loss of generality, we may assume that $\mathcal{F}_T = \mathcal{F}$. We denote by λ the restriction of the Lebesgue–Borel measure to $[0, T]$ and fix two dimensions $d, n \in \mathbb{N}$. Let $M = (M_t)_{t\in[0,T]}$ be an $\mathbb{R}^d$-valued continuous local martingale with $M_0 = 0$. Unless stated otherwise, we assume all stochastic processes to be $\mathbb{F}$-adapted.

Let us start with a classical example that highlights the main concepts which are of importance in this section, before extending the study to a more general setting.

Example 2.1 Let (E, H, γ) denote the classical Wiener space; this means that we take $E = C_0([0, T]; \mathbb{R}^d)$, $H = \{h(t) = (\mathbb{1}_{[0,t]} f_h) \cdot \lambda, t \in [0, T] : f_h \in L^2(\lambda; \mathbb{R}^d)\}$ is the space of $\mathbb{R}^d$-valued absolutely continuous functions on $[0, T]$ that admit a square-integrable density with respect to λ, and γ is the classical Wiener measure on E, which is the Borel probability measure on E induced by an $\mathbb{R}^d$-valued standard Brownian motion $B = (B_t)_{t\in[0,T]}$.

One can show that with the inner product $\langle g, h\rangle_H = \langle f_g, f_h\rangle_{L^2(\lambda;\mathbb{R}^d)}$, the space $(H, \langle \cdot, \cdot \rangle_H)$ is a real separable Hilbert space continuously embedded into E as a dense linear subspace. The operator $J\colon L^2(\lambda; \mathbb{R}^d) \to H$, $f_h \mapsto h(\cdot) = (\mathbb{1}_{[0,\cdot]} f_h) \cdot \lambda$ is a linear isometry by construction and thus continuous. Whenever D is a dense linear subspace of $L^2(\lambda; \mathbb{R}^d)$, it follows that $J(D)$ is a dense linear subspace of H and thus densely embedded into E. In other words, $\overline{J(D)} = H$ and $\overline{J(D)} = E$, where the closure of $J(D)$ is taken in H and E, respectively.

Section 2 is dedicated to a refined study of the identity $\overline{J(D)} = H$ in a generalised setting. To this end, let us state an assumption which allows us to study the process M as a Gaussian process and simplifies the proofs of Sect. 4. Moreover, it leads to a natural candidate for the space H of drift adjustments, which consists of deterministic functions (see Definition 2.13 below).

Assumption 2.2 The covariation process $[M]$ is up to indistinguishability deterministic, and $\operatorname{tr}([M])_T > 0$ outside a $\mathbb{P}$-nullset.

In what follows, we disregard the evanescent set and $\mathbb{P}$-nullset on which the two conditions from Assumption 2.2 are violated and consider equalities between stochastic processes and (in)equalities between random variables to hold up to indistinguishability and $\mathbb{P}$-almost surely, respectively.

Definition 2.3 The quadratic variation process $C := \operatorname{tr}([M])$, being increasing and of finite variation, induces a finite Lebesgue–Stieltjes measure on $([0, T], \mathcal{B}_{[0,T]})$, which we denote by μ.

Remark 2.4 Due to (the proof of) Lévy's characterisation theorem, the increments $M_t - M_s$ are independent of $\mathcal{F}_s$ with $\mathcal{L}(M_t - M_s) = \mathcal{N}(0, [M]_t - [M]_s)$ for all $s < t$ in $[0, T]$. Therefore M is a centered Gaussian process, and Kallenberg [30, Theorem 11.5] shows that M is an $\mathbb{F}$-Markov process. Moreover, M is a martingale because $\mathbb{E}[M_t - M_s|\mathcal{F}_s] = \mathbb{E}[M_t - M_s] = 0$ for $s < t$ in $[0, T]$. Note that

$\mathrm{Cov}(M_t, M_s) = [M]_{s\wedge t}$ for $s, t \in [0, T]$ since, assuming without loss of generality that $s < t$,

$$\mathrm{Cov}(M_t, M_s) = \mathbb{E}[M_t M_s^\top] = \mathbb{E}[(M_t - M_s)M_s^\top] + \mathrm{Cov}(M_s, M_s) = [M]_s.$$

Remark 2.5 Assumption 2.2 implies that $\mu([0, T]) > 0$.

We write $\mu = \mu_\mathrm{a} + \mu_\mathrm{s}$ for the Lebesgue decomposition of μ with respect to λ into an absolutely continuous measure $\mu_\mathrm{a} = f_\lambda \cdot \lambda$ and a singular measure μ_s, where f_λ denotes a Radon–Nikodým density of μ_a with respect to λ, and both μ_a and μ_s are finite measures. Note that μ has no atoms since $[M]$ and therefore also C are continuous.

We now proceed in line with Cohen and Elliott [12, Sect. 12.5]. The covariation process $[M^i, M^j]$, being continuous and of finite variation, induces a finite signed and atomless measure $\mu_{i,j}$ on $([0, T], \mathcal{B}_{[0,T]})$ for all $i, j \in \{1, 2, \ldots, d\}$. It follows from the Kunita–Watanabe inequality for Lebesgue–Stieltjes integrals that the total variation measure $|\mu_{i,j}|$ is absolutely continuous with respect to μ. Hence an application of the Radon–Nikodým theorem for signed measures (cf. [12, Sect. 1.7.14]) yields the existence of a real-valued density $\mathrm{d}\mu_{i,j}/\mathrm{d}\mu =: \pi_{i,j} \in L^1(\mu)$ as $\mu_{i,j}$ is finite.

We collect $(\pi_{i,j})_{i,j=1,\ldots,d}$ into a measurable function π that assumes by the symmetry of $[M]$ values in the space of symmetric matrices in $\mathbb{R}^{d\times d}$, and we write

$$[M]_t = (\pi \bullet C)_t = (\mathbb{1}_{[0,t]}\pi) \cdot \mu, \qquad t \in [0, T], \tag{2.1}$$

where the notation is to be understood componentwise. Let $(\eta_k)_{k\in\mathbb{N}}$ be dense in $\mathbb{R}^d$. For each $k \in \mathbb{N}$, set $A_k = \{s \in [0, T] : \eta_k^\top \pi(s)\eta_k \geq 0\}$ and moreover $A = \bigcap_{k\in\mathbb{N}} A_k$. Note that $A = \{s \in [0, T] : \eta^\top \pi(s)\eta \geq 0, \forall \eta \in \mathbb{R}^d\}$ and

$$\int_0^t \eta_k^\top \pi(s)\eta_k \mu(\mathrm{d}s) = \big((\eta_k^\top \pi \eta_k) \bullet C\big)_t = [\eta_k^\top M]_t \geq 0, \qquad k \in \mathbb{N}, t \in [0, T],$$

which implies that each A_k^c is a μ-nullset, and therefore so is A^c as the countable union of all sets A_k^c. We conclude that π is positive semidefinite μ-almost everywhere. Note that we could, in the spirit of Cherny and Shiryaev [8, Sect. 3] and without loss of generality, replace π by $\tilde{\pi} = \pi\mathbb{1}_A$ and thus assume that π is positive semidefinite for each $t \in [0, T]$. For the purpose of this paper, this step is not necessary though.

Remark 2.6 The decomposition of $[M]$ into a matrix-valued function π and an increasing process C is not unique. For example, take $\tilde{C} = \sum_{i=1}^d \eta_i [M^i]$, where $\eta \in \mathbb{R}^d$ is chosen such that $\eta_i > 0$ for all $i \in \{1, 2, \ldots, d\}$. More generally, take $\tilde{C} = \sum_{i=1}^d f_i \bullet [M^i]$ with $f_i \colon [0, T] \to \mathbb{R}_+ \setminus \{0\}$ in $L^1(\mu_{i,i})$ for $i \in \{1, 2, \ldots, d\}$. In both cases, the corresponding function $\tilde{\pi}$ is then constructed as in Remark 2.5 and in general differs from π. Lemma 2.10(e) below shows that the non-uniqueness of (π, C) is not a problem though.

Example 2.7 If $\pi \equiv I_d$ and $\mu = \lambda$, then M is by Lévy's characterisation an $\mathbb{R}^d$-valued standard Brownian motion.

Example 2.8 An example of relevance for practitioners is the multivariate Heston model which we now briefly describe.

Let $d = 2n$ for some $n \in \mathbb{N}$. We consider a dynamic diffusion matrix given by $[0, T] \ni t \mapsto \Sigma(t) \in \mathbb{R}^{d\times d}$, a vector of appreciation rates $r \in \mathbb{R}^n$, an n-dimensional vector of positive mean-reversion levels m and an $(n \times n)$-dimensional diagonal matrix Θ with positive entries representing mean-reversion speeds. To avoid degeneracy, we assume that for each $k \in \{1, 2, \dots, n\}$ and $t \in [0, T]$, $(\Sigma_{k,\cdot}(t))^\top$ is not the zero vector. Let $M_t = \Sigma(t)B_t$, where B denotes a standard Brownian motion with values in $\mathbb{R}^d$. Note that $[M]_t = \mathrm{Cov}(M_t, M_t) = \int_0^t \Sigma(s)\Sigma^\top(s)\,\mathrm{d}s$ for each $t \in [0, T]$. Hence in view of Remark 2.5, we may choose $\mu = \lambda$ and $\pi(t) = \Sigma(t)\Sigma^\top(t)$.

Fix two n-dimensional initial value vectors s, v with positive entries. For simplicity, we write $M^{(1)} = (M^1, M^2, \dots, M^n)^\top$ and $M^{(2)} = (M^{n+1}, M^{n+2}, \dots, M^{2n})^\top$ so that $M = (M^{(1)}, M^{(2)})^\top$. Let $X = (S, V)^\top$ and let the asset price follow the SDE $\mathrm{d}S_t = (r \odot S_t)\,\mathrm{d}t + (S_t \odot \sqrt{V_t}) \odot \mathrm{d}M_t^{(1)}$ with $S_0 \equiv s$. The n-dimensional instantaneous variance process V follows the Cox–Ingersoll–Ross (CIR)-type SDE $\mathrm{d}V_t = \Theta(m - V_t)\,\mathrm{d}t + \sqrt{V_t} \odot \mathrm{d}M_t^{(2)}$ with $V_0 \equiv v$. Here we see that asset price models whose dynamics are driven by multivariate Brownian motions with dynamic variance–covariance matrices fall within the scope of our setting. More generally, one could think of replacing (B_t) by a time-changed Brownian motion $(B_{f(t)})$ for a given deterministic time-change f.

In Sect. 5, we study special cases of this model, with either a time-change to model changing levels of business activity, or a dynamic correlation structure.

Example 2.9 Set $T = 1$ and let B be a standard $\mathbb{R}^d$-valued $(\mathbb{G}, \mathbb{P})$-Brownian motion, where $\mathbb{G} = (\mathcal{G}_t)_{t\in[0,T]}$ denotes a filtration of $\mathcal{F}$. Let $f\colon [0, T] \to [0, T]$ be either Cantor's ternary function or Minkowski's question-mark function, and let $\Sigma \in \mathbb{R}^{d\times d}$ be a diffusion matrix. Recall that Cantor's ternary function is continuous, monotonically increasing, has derivative zero on a set of Lebesgue measure one, but is not absolutely continuous. Likewise, Minkowski's question-mark function (see Salem [52, Sect. 4]) has the same properties, while being even strictly increasing. Set $M_t := \Sigma B_{f(t)}$ for $t \in [0, T]$ and note that M is an $\mathbb{R}^d$-valued continuous $(\mathbb{F}, \mathbb{P})$-martingale with $M_0 = 0$ and $[M]_t = f(t)\Sigma\Sigma^\top$ for $t \in [0, T]$, where the filtration $\mathbb{F} = (\mathcal{F}_t)_{t\in[0,T]}$ is given by $\mathcal{F}_t = \mathcal{G}_{f(t)}$ for $t \in [0, T]$. The corresponding Lebesgue–Stieltjes measure μ is singular with respect to λ, and $\mathrm{tr}([M])$ is increasing when f is Cantor's ternary function, and even strictly increasing when f is Minkowski's question-mark function. See [52] for further examples of functions f that can be used for constructions of this kind.

Based on the pair (π, μ), we define a weighted L^2-space, which we denote by Λ^2, which is a generalisation of the space $L^2(\lambda; \mathbb{R}^d)$ in the context of Example 2.1, and we recall some elementary properties. In Sect. 4 where we study importance sampling, functions f from Λ^2 are used to construct equivalent measures via the Doléans–Dade exponential $\mathcal{E}(f \bullet M)$. As we argue in Sect. 3, feedforward neural networks are dense in Λ^2 under suitable assumptions, which due to Theorem 4.6 provides a theoretical justification for using feedforward neural networks in order to calibrate an optimal sampling measure that minimises the variance of the Monte

Carlo estimators in Sects. 4 and 5. The definition of the space Λ^2 is inspired by the concept of vector stochastic integration; see Jacod [28, Chap. IV] for further details and generalisations.

Lemma 2.10 *Let Λ^2 denote the set of all $f \in \mathcal{L}^0(\mathcal{B}_{[0,T]}; \mathbb{R}^d)$ with*

$$\|f\|_{\Lambda^2} := \left(\int_0^T f^\top(s)\pi(s)f(s)\mu(\mathrm{d}s) \right)^{1/2} < \infty,$$

where we identify $f, g \in \Lambda^2$ if $(f-g)^\top \pi (f-g) = 0$ μ-almost everywhere and write $f \sim g$ in this case. We further set $\langle f, g \rangle_{\Lambda^2} := \int_0^T f^\top(s)\pi(s)g(s)\mu(\mathrm{d}s)$ for $f, g \in \Lambda^2$. Then:

(a) *$(\Lambda^2, \langle \cdot, \cdot \rangle_{\Lambda^2})$ is a real separable Hilbert space.*

(b) *To each $F \in (\Lambda^2)^*$, there corresponds a unique function $g \in \Lambda^2$ such that*

$$F(f) = \int_0^T g^\top(s)\pi(s)f(s)\mu(\mathrm{d}s), \qquad f \in \Lambda^2,$$

and $\|F\|_{\mathrm{op}} = \|g\|_{\Lambda^2}$. Therefore $(\Lambda^2)^*$ is isometrically isomorphic to Λ^2.

(c) *We denote by $\Lambda^{2,0}$ the set of all $f \in \mathcal{L}^0(\mathcal{B}_{[0,T]}; \mathbb{R}^d)$ that satisfy $f_i \in L^2(\mu_{i,i})$ for each $i \in \{1, 2, \ldots, d\}$, where we identify functions in the same manner as above. Then*:

1) *$(\Lambda^{2,0}, \langle \cdot, \cdot \rangle_{\Lambda^2})$ is a separable inner product space with $\Lambda^{2,0} \subseteq \Lambda^2$.*

2) *$\Lambda^{2,0}$ is dense in Λ^2; hence Λ^2 is the completion of $\Lambda^{2,0}$ with respect to $\|\cdot\|_{\Lambda^2}$.*

(d) *$(C([0,T]; \mathbb{R}^d), \|\cdot\|_\infty)$ is continuously embedded into $(\Lambda^{2,0}, \|\cdot\|_{\Lambda^2})$ as a dense linear subspace, where $\|f\|_\infty := \sup_{t\in[0,T]} |f(t)|$ for $f \in C([0,T]; \mathbb{R}^d)$.*

(e) *Λ^2 and $\Lambda^{2,0}$ do not depend on the specific choice of (π, μ) that satisfy (2.1).*

Example 2.11 According to [28, Lemme 4.30] and the discussion thereafter, a sufficient condition for $\Lambda^{2,0} = \Lambda^2$ to hold is that there exists a constant $c > 0$ such that $\sum_{i=1}^d \pi_{i,i} f_i^2 \le c f^\top \pi f$ holds μ-almost everywhere for all $f \in \Lambda^2$. Examples where this applies are when π is a diagonal or uniformly strictly elliptic matrix, where the latter condition means that there is a constant $c > 0$ with $c|\eta|^2 \le \eta^\top \pi \eta$ for all $\eta \in \mathbb{R}^d$.

Example 2.12 Let us state one example, which is a deterministic version of Cohen and Elliott [12, Example 12.5.1], where $\Lambda^{2,0} \neq \Lambda^2$. To this end, let B denote a real-valued standard Brownian motion. Set $M = (B, -B)^\top$ and note that M is an $\mathbb{R}^2$-valued continuous martingale with covariation

$$[M]_t = \begin{pmatrix} t & -t \\ -t & t \end{pmatrix}, \qquad \text{hence } \pi \equiv \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix},$$

where we choose $\mu = \lambda$. Then π is positive semidefinite since $\eta^\top \pi \eta = (\eta_1 - \eta_2)^2$ for each $\eta \in \mathbb{R}^2$, and this is zero precisely when $\eta_1 = \eta_2$ and positive otherwise. Let $f\colon [0,T] \to \mathbb{R}$ be measurable and such that $f \notin L^2(\lambda)$. Consider the function $g\colon [0,T] \to \mathbb{R}^2$ given by $g = (f, f)^\top$. By construction, we then have $g \notin \Lambda^{2,0}$, but since $\|g\|_{\Lambda^2} = 0$, we have $g \in \Lambda^2$.

For $f \in \Lambda^2$, we have $f \in L^2(M)$ in the sense of vector stochastic integration (see also Lemma B.2 below). Since $[f \bullet M]_T = \|f\|^2_{\Lambda^2}$ is deterministic and finite, Novikov's criterion shows that $Z = \mathcal{E}(f \bullet M)$ is a strictly positive martingale. Girsanov's theorem shows that under the measure $\mathbb{Q}$ with $\mathrm{d}\mathbb{Q}/\mathrm{d}\mathbb{P} = Z_T$ on $\mathcal{F}_T$, the finite-variation part in the semimartingale decomposition of M is given by $[f \bullet M, M] = h$, where $h(t) = (\mathbb{1}_{[0,t]}\pi f) \cdot \mu$ for $t \in [0, T]$. These considerations motivate the following definition.

Definition 2.13 We denote by H the set of all $h\colon [0, T] \to \mathbb{R}^d$ with the representation

$$h(t) = J(f_h)(t) := \int_0^t \pi(s) f_h(s) \mu(\mathrm{d}s), \qquad t \in [0, T], \tag{2.2}$$

for some $f_h \in \Lambda^2$, where the integral in (2.2) is to be understood componentwise as a Lebesgue–Stieltjes integral.

As Proposition 2.18 below will show, when endowed with an appropriate inner product, H becomes the Cameron–Martin space of the Gaussian measure γ_M which is induced by M on $C_0([0, T]; \mathbb{R}^d)$. To the best of our knowledge, there exists no explicit characterisation of the Cameron–Martin space of γ_M at the present level of generality in the literature so far, as one usually assumes M to be a Brownian motion, which is a special case of our setting (see Example 2.7).

Example 2.14 In the context of Example 2.7, H coincides with the set of absolutely continuous functions whose densities are square-integrable with respect to λ.

Remark 2.15 The Cameron–Martin space of a fractional Brownian motion is not contained in our framework, except for the special case of a Brownian motion. The matrix-valued function π is not to be confused with the square-integrable but singular kernel which appears in integral representations of the fractional Brownian motion and more generally Volterra-type Gaussian processes. However, our framework can be extended to multivariate versions of these processes with representations of the form $\tilde{M}_t = \int_0^T k(t, s)\, \mathrm{d}M_s$, where k denotes an $\mathbb{R}^{d\times d}$-valued kernel function, for which, under suitable assumptions on k, the corresponding Cameron–Martin space consists of functions of the form

$$\tilde{h}(t) = \int_0^T k(t, s)\pi(s) f_h(s) \mu(\mathrm{d}s), \qquad t \in [0, T],$$

for some $f_h \in \Lambda^2$. This formulation gives rise to the study of refined versions of multivariate Volterra-type Gaussian processes as well as multivariate fractional stochastic volatility models, as one can now distinguish more explicitly between time-inhomogeneous volatility patterns which are induced by μ (or, equivalently, by the quadratic variation C), the dependency structure of the components of M which is modelled by the function π, and the path irregularities of $\tilde{M}$ which are induced by the matrix-valued kernel k.

Remark 2.16 Equation (2.2) suggests a generalisation, where the functions f_h assume values in a (possibly infinite-dimensional) Hilbert space $\tilde{H}$ and π assumes μ-almost everywhere values in the set of positive semidefinite operators on $\tilde{H}$. In this case, the integral in (2.2) is to be understood as a Lebesgue–Stieltjes–Bochner integral.

Example 2.17 Let $w\colon [0,T] \to [1,\infty)$ be a function in $L^1(\lambda;\mathbb{R})$ that is nondecreasing, set $\Sigma = I_d$ and define the measure μ through $\mu(A) = \int_A w(s)\lambda(\mathrm{d}s)$ for $A \in \mathcal{B}_{[0,T]}$. In line with Example 2.9, we can then construct a process M by means of the increasing and continuous function $f\colon [0,T] \to \mathbb{R}_+$ given by $f(t) := \mu([0,t])$ for $t \in [0,T]$. In the context of Definition 2.13, the corresponding space H then has similarities to the forward curve space H_w (cf. Filipović [15, Chap. 5]) used in interest rate modelling.

Since elements from H will be precisely those which we consider for the drift adjustment of M in Sects. 4 and 5, we need to collect some useful properties needed later on (in particular for Theorem 2.24). The following proposition collects these properties and further deepens the connections to the process M. As it turns out, when endowed with a suitable inner product, H is not only the isometric image of the space Λ^2 whose definition was inspired by the representation (2.1) of $[M]$, but also the Cameron–Martin space of the Gaussian measure γ_M induced by M on $C_0([0,T];\mathbb{R}^d)$.

Proposition 2.18 *Consider the mapping $\langle\cdot\,,\,\cdot\rangle_H$ given by $\langle g,h\rangle_H := \langle f_g, f_h\rangle_{\Lambda^2}$ for $g,h \in H$. Then:*

(a) *The integral in (2.2) is well defined for all $f_h \in \Lambda^2$ and $t \in [0,T]$.*

(b) *$(H, \langle\cdot\,,\,\cdot\rangle_H)$ is a real separable Hilbert space.*

(c) *$J\colon \Lambda^2 \to H$ is a linear isometry and $(H^0, \langle\cdot\,,\,\cdot\rangle_H)$ is an inner product space whose completion is $(H, \langle\cdot\,,\,\cdot\rangle_H)$, where we set $H^0 := J(\Lambda^{2,0}) \subseteq H$.*

(d) *To each $F \in H^*$, there corresponds a unique function $g_F \in H$ such that*

$$F(h) = \langle g_F, h\rangle_H = \int_0^T f_{g_F}^\top(s)\pi(s)f_h(s)\mu(\mathrm{d}s), \qquad h \in H,$$

and $\|F\|_{\mathrm{op}} = \|g_F\|_H$. Therefore H^* is isometrically isomorphic to H.

(e) *H is the Cameron–Martin space of the centered Gaussian measure γ_M that is induced by M on $E = C_0([0,T];\mathbb{R}^d)$.*

Remark 2.19 Unless $\overline{H} = E$, where the closure is taken in E, the measure γ_M is degenerate (see Remark A.3). The identity $\overline{H} = E$ holds in some special cases, e.g. if $\mu = \lambda$ and $\pi \equiv I_d$, where the proof builds on the fact that continuous functions can be uniformly approximated by piecewise linear functions.

For the purpose of the next result, we introduce the function $I\colon E \to \overline{\mathbb{R}}_+$ by

$$I(g) = \begin{cases} \frac{1}{2}\int_0^T f_g^\top(s)\pi(s)f_g(s)\mu(\mathrm{d}s) & \text{for } g \in H, \\ \infty, & \text{otherwise.} \end{cases} \tag{2.3}$$

Example 2.20 In the context of Example 2.8, π takes the form $\pi(t) = \Sigma(t)\Sigma^\top(t)$. On the other hand, in the context of Example 2.9, the measure μ could be the Lebesgue–Stieltjes measure that is induced by Cantor's ternary function and thus singular with respect to λ. The setting typically discussed in the literature, where M is a standard Brownian motion, does not encompass either of these examples.

In Sect. 4.1, we discuss an importance sampling method that uses methods from the theory of large deviations. To this end, one needs to understand the asymptotic behaviour of the scaled process $\sqrt{\varepsilon}M$ as $\varepsilon \searrow 0$. If M is a Brownian motion, the corresponding result is referred to as Schilder's theorem (cf. Bogachev [5, Corollary 4.9.3]). As a consequence of Propositions 2.18(e) and A.7, we obtain the following result, whose novelty is the explicit characterisation of the function I (also referred to as rate function) in (2.3) at the present level of generality.

Proposition 2.21 *In the context of Proposition 2.18(e), we have for each $F \in \mathcal{B}_E$ that*

$$-\inf_{g\in F^\circ} I(g) \le \liminf_{\varepsilon\searrow 0} \varepsilon \log \mathbb{P}[\sqrt{\varepsilon}M \in F]$$
$$\le \limsup_{\varepsilon\searrow 0} \varepsilon \log \mathbb{P}[\sqrt{\varepsilon}M \in F] \le -\inf_{g\in \bar{F}} I(g),$$

where the function $I\colon E \to \overline{\mathbb{R}}_+$ is specified in (2.3).

For notational convenience, we introduce the following condition.

Standing Assumption 2.22 Henceforth we denote by D a dense subset of Λ^2.

Example 2.23 We have already encountered two admissible candidates for the set D, namely $\Lambda^{2,0}$ and $C([0,T];\mathbb{R}^d)$; see Lemma 2.10(c) and 2.10(d).

The following theorem helps us to identify dense subsets and subspaces of H and shows how these relate to the topological support (see Remark A.3 in Appendix A) of the measure γ_M which we encountered in Proposition 2.18.

Theorem 2.24 *Let $H(D)$ denote the set of all $h \in H$ with $f_h \in D$. Then:*

(a) *$H(D)$ is a dense subset of H which is separable when endowed with the subspace topology.*

(b) *If D is also a linear subspace of Λ^2, then:*

1) *$(H(D), \langle\cdot\,,\,\cdot\rangle_H)$ is an inner product space, whose completion is H.*

2) *There exists a countable orthonormal basis of H which consist of elements from $H(D)$.*

(c) *In the context of Proposition 2.18(e), the topological support of γ_M coincides with $\overline{H(D)}$, where the closure is taken in E. In other words,*

$$\gamma_M\big(C_0([0,T];\mathbb{R}^d) \setminus \overline{H(D)}\big) = \mathbb{P}\big[M \in C_0([0,T];\mathbb{R}^d) \setminus \overline{H(D)}\big] = 0.$$

Hence outside of a $\mathbb{P}$-nullset, paths of M can be uniformly approximated by sequences from $H(D)$.

In Sect. 3, we discuss classes of feedforward neural networks that are also dense subsets of Λ^2, thereby satisfying the Standing Assumption 2.22. Together with Theorem 2.24, this shows that we can approximate any element from H, up to the isometry J, by feedforward neural networks. This is essential for Sects. 4 and 5, where we approximate with feedforward neural networks the drift adjustment of M which minimises the variance of the Monte Carlo estimator.

3 Approximation capabilities of feedforward neural networks

In this section, we study feedforward neural networks as elements of the space Λ^2 and show how they generate, under suitable assumptions on the activation function, dense subspaces of H, thereby providing a first theoretical justification for approximating the optimal drift adjustment with feedforward neural networks when studying importance sampling in Sects. 4 and 5 below.

We know from Lemma 2.10(d) that $C([0,T];\mathbb{R}^d)$ is continuously embedded into $\Lambda^{2,0}$ as a dense linear subspace. Moreover, Lemma 2.10(c) shows that $\Lambda^{2,0}$ is dense in Λ^2. Consequently, every dense linear subspace D of $C([0,T];\mathbb{R}^d)$ is densely embedded into Λ^2, thereby satisfying the Standing Assumption 2.22, in which case $H(D)$ is dense in H by Theorem 2.24(a). For this reason, we first focus our attention on finding dense linear subspaces of $C([0,T];\mathbb{R}^d)$ in Proposition 3.4, before relaxing the continuity assumption in Proposition 3.5 below.

Neural networks are one particular class of functions of interest to us. On the one hand, it satisfies the required density in $C([0,T];\mathbb{R}^d)$ and, on the other hand, it gives rise to efficient numerical optimisation procedures which have led to fascinating results in the domain of financial and actuarial mathematics in recent years. When studying neural networks, the property of a set of them to be dense in a topological space is referred to as the universal approximation property (UAP); cf. Kratsios [34, Definition 2]. Theorems which establish denseness of sets of neural networks in a topological space are referred to as universal approximation theorems (UATs).

There are many different neural network architectures for which the UAP has been shown to hold in various topological spaces (cf. Cybenko [13], Funahashi [18], Hornik [26], Hornik et al. [27], Kidger and Lyons [32], Leshno et al. [36], Liao et al. [38], Mhaskar and Micchelli [43], Park and Sandberg [45, 46], Pinkus [47] and Zhang et al. [55]). For ease of presentation, we discuss the class of feedforward neural networks, also called multilayer perceptrons or multilayer feedforward neural networks. Note that our discussion is limited to architectures that yield UATs in $C([0,T];\mathbb{R}^d)$ and Λ^2 (see also Sect. 3.1 for a UAT with respect to Hölder-norms). There are many architectures for which the UAP has been established in other topological spaces, but which we do not discuss in this paper.

Definition 3.1 Given $k, \ell \in \mathbb{N}$ and $\psi\colon \mathbb{R} \to \mathbb{R}$, we denote by $\mathcal{NN}^d_{k,\ell}(\psi)$ the set of feedforward neural networks with one neuron in the input layer, d neurons with identity activation function in the output layer, k hidden layers, and at most ℓ hidden nodes with ψ as activation function in each hidden layer (cf. Kidger and Lyons [32, Definition 3.1]).

Remark 3.2 The set $\mathcal{NN}_{k,\ell}^d(\psi)$ consists of all functions of the form $t \mapsto W_{k+1} \circ F_k \circ \cdots \circ F_1(t)$, where $k \in \mathbb{N}$, $W_1 \colon \mathbb{R} \to \mathbb{R}^\ell$, $W_{k+1} \colon \mathbb{R}^\ell \to \mathbb{R}^d$ and $W_2, \ldots, W_k \colon \mathbb{R}^\ell \to \mathbb{R}^\ell$ are affine functions and $F_i = \psi \circ W_i$ for $i = 1, 2, \ldots, k$, where the activation function ψ is applied componentwise.

If the number of nodes in the hidden layers can be arbitrarily large, we write $\mathcal{NN}_{k,\infty}^d(\psi)$. Likewise, we write $\mathcal{NN}_{\infty,\ell}^d(\psi)$ if the number of hidden layers can be arbitrarily large. Finally, the notation

$$\mathcal{NN}_{\infty,\infty}^d(\psi) = \bigcup_{k \in \mathbb{N}} \mathcal{NN}_{k,\infty}^d(\psi) = \bigcup_{\ell \in \mathbb{N}} \mathcal{NN}_{\infty,\ell}^d(\psi)$$

is to be understood in an analogous way. For the purpose of Example 3.7 below, we also introduce the following notation: If A is a finite set of functions $f \colon \mathbb{R} \to \mathbb{R}$, then $\mathcal{NN}_{k,\ell}^d(A)$ denotes the set of feedforward neural networks where the hidden nodes are endowed with any of the functions from A. As a special case, we then have $\mathcal{NN}_{k,\ell}^d(A) = \mathcal{NN}_{k,\ell}^d(\psi)$ for $A = \{\psi\}$.

Functions in $\mathcal{NN}_{1,\infty}^d(\psi)$ are called shallow feedforward neural networks, while functions in $\mathcal{NN}_{\infty,\infty}^d(\psi)$ are generally referred to as deep feedforward neural networks. The set $\mathcal{NN}_{\infty,\ell}^d(\psi) \subseteq \mathcal{NN}_{\infty,\infty}^d(\psi)$ of deep narrow networks, where $\ell \in \mathbb{N}$ is fixed, is also of special interest (cf. Kidger and Lyons [32]).

In the context of Definition 3.1, the function ψ is sometimes also called squashing function, sigmoid function or ridge activation function. Different terms have been chosen based on the properties of ψ, which in general differ based on which topological spaces we are studying the UAP in. For the sake of simplicity, we call ψ an *activation function* **throughout this paper**, and impose properties on ψ wherever needed. In view of Lemma 2.10(d), we are particularly interested in the UAP in $C([0, T]; \mathbb{R}^d)$. At this point, however, we need to discuss a technicality first.

For a Borel measure ν on $[0, T]$ and $f, g \in C([0, T]; \mathbb{R}^d)$, we write $f \sim_\nu g$ if $f = g$ outside a ν-nullset. Because $\sim_\nu$ is an equivalence relation on $C([0, T]; \mathbb{R}^d)$, we can therefore consider the quotient space $C_\nu([0, T]; \mathbb{R}^d)$ of $C([0, T]; \mathbb{R}^d)$ under $\sim_\nu$, on which the ν-essential supremum $\| \cdot \|_{L^\infty([0,T],\nu)}$ is a norm, making $(C_\nu([0, T]; \mathbb{R}^d), \| \cdot \|_{L^\infty([0,T],\nu)})$ a normed vector space. Then a small modification of Lemma 2.10(d) shows that $C_\nu([0, T]; \mathbb{R}^d)$ is continuously embedded into $\Lambda^{2,0}$ as a dense linear subspace, provided that μ is absolutely continuous with respect to ν.

Let us collect classical versions of the universal approximation theorem which are concerned with the (almost everywhere) uniform approximation of continuous functions (cf. [26, 32, 36]), as we refer to them in the proofs of the subsequent results.

Theorem 3.3 *Given $\psi \colon \mathbb{R} \to \mathbb{R}$, consider the following assumptions*:

1) *ψ is continuous, bounded and nonconstant.*

2) *ψ is continuous and nonaffine, and there exists a point $x \in \mathbb{R}$ at which ψ is continuously differentiable with $\psi'(x) \neq 0$.*

3) *ψ is locally λ-essentially bounded. Moreover, ψ is λ-almost everywhere not an algebraic polynomial, and the set of points of discontinuity of ψ is a λ-nullset.*

Then:
(a) *If* 1) *holds, then* $\mathcal{NN}^d_{1,\infty}(\psi)$ *is dense in* $(C([0,T];\mathbb{R}^d), \|\cdot\|_\infty)$.
(b) *If* 2) *holds, then* $\mathcal{NN}^d_{\infty,d+3}(\psi)$ *is dense in* $(C([0,T];\mathbb{R}^d), \|\cdot\|_\infty)$.
(c) *If* 3) *holds, then* $\mathcal{NN}^d_{1,\infty}(\psi)$ *is dense in* $(C_\lambda([0,T];\mathbb{R}^d), \|\cdot\|_{L^\infty([0,T],\nu)})$.

The following two results yield dense subsets of Λ^2 which consist of feedforward neural networks. We can therefore consider these sets as admissible for the set D in the context of the Standing Assumption 2.22. Consequently, due to Theorem 2.24 and under suitable assumptions on ψ, feedforward neural networks are, up to the isometry J, dense in H.

Proposition 3.4 below is a consequence of Theorem 3.3 and Lemma 2.10(d). For simplicity, we only formulate it for $\mathcal{NN}^d_{1,\infty}(\psi)$, where the case for $\mathcal{NN}^d_{\infty,d+3}(\psi)$ can be argued analogously. Since $\mathcal{NN}^d_{1,\infty}(\psi)$ is a subset of $\mathcal{NN}^d_{k,\infty}(\psi)$ for every $k \in \mathbb{N}$, Propositions 3.4 and 3.5 below hold for $\mathcal{NN}^d_{k,\infty}(\psi)$, $k \in \mathbb{N}$, too.

Proposition 3.4 *In the context of Theorem* 3.3, *assume either that Condition* 1) *holds, or that Condition* 3) *holds and* μ *is absolutely continuous with respect to* λ. *Then* $\mathcal{NN}^d_{1,\infty}(\psi)$ *is a dense linear subspace of* $\Lambda^{2,0}$.

By looking at the proof of Proposition 3.4 (which is presented in Appendix B), it becomes clear that we cannot impose Assumption 3) from Theorem 3.3 if μ is not absolutely continuous with respect to λ. This is relevant in particular for Example 2.9, where we need to impose either Assumption 1) or 2) from Theorem 3.3. Moreover, assuming ψ to be (λ-almost everywhere) continuous is also rather restrictive, given that functions in Λ^2 need not be continuous. By arguing along the lines of Cybenko [13] and Hornik [26], we can actually drop the continuity assumption on ψ at the cost of requiring boundedness, which is not required in Assumptions 2) and 3) from Theorem 3.3.

Proposition 3.5 *If* ψ *is bounded, measurable and nonconstant, then* $\mathcal{NN}^d_{1,\infty}(\psi)$ *is a dense linear subspace of* $\Lambda^{2,0}$.

For notational simplicity, let us assume that the activation function ψ satisfies sufficient conditions such that either Proposition 3.4 or 3.5 is applicable.

Standing Assumption 3.6 Henceforth we assume one of the following:
(i) Condition 1) from Theorem 3.3 holds so that ψ is continuous.
(ii) The assumptions from Proposition 3.5 hold so that ψ is bounded, but not necessarily continuous.
(iii) μ is absolutely continuous with respect to λ and Condition 3) from Theorem 3.3 holds so that ψ is λ-almost everywhere continuous.

Example 3.7 Fix $d = 1$ as well as $\mu = \lambda$. Note that in this case, we have $\pi \equiv 1$. Set $D = \mathcal{NN}^1_{1,\infty}(\psi) = \text{span}\{[0,T] \ni t \mapsto \psi(\alpha t + \eta) : \alpha, \eta \in \mathbb{R}\}$, where $\psi = \tanh$. Since every $f \in D$ is continuous and thus bounded on $[0,T]$, we may replace the Lebesgue by the Riemann integral.

If $\alpha = 0$ and $\eta \in \mathbb{R}$, then $\int_0^t \psi(\eta)\,\mathrm{d}s = \psi(\eta)t$ for $t \in [0, T]$. On the other hand, if $\alpha \neq 0$ and $\eta \in \mathbb{R}$, then by substitution,

$$\int_0^t \psi(\alpha s + \eta)\,\mathrm{d}s = \frac{1}{\alpha}\big(\tilde{\psi}(\alpha t + \eta) - \tilde{\psi}(\eta)\big), \qquad t \in [0, T],$$

where $\tilde{\psi}(\,\cdot\,) = \log\cosh(\,\cdot\,)$. Similarly, if ψ is the standard sigmoid (logistic) function, the same applies with $\tilde{\psi}(\,\cdot\,) = \log(1+\exp(\,\cdot\,))$, which is also called softplus function. To sum up, we see that

$$H(D) = \operatorname{span}\{\mathrm{id}\colon [0, T] \ni t \mapsto t, \mathcal{NN}^1_{1,\infty}(\tilde{\psi})\} = \mathcal{NN}^1_{1,\infty}(\{\mathrm{id}, \tilde{\psi}\}).$$

Provided that ψ is Riemann-integrable, Example 3.7 shows that compared to the set $D = \mathcal{NN}^1_{1,\infty}(\psi)$, the set $H(D)$ can be obtained by modifying the activation function and adding the identity function into the set of admissible activation functions. This can be helpful when optimising over functions in $H(D)$, because one avoids having to implement integral operations. What is more, functions in $H(D)$ enjoy the property of being absolutely continuous, provided that μ is absolutely continuous with respect to λ, while this is not always the case for functions from $\mathcal{NN}^1_{1,\infty}(\psi)$, e.g. if ψ is not continuous.

Note that we formulated Propositions 3.4 and 3.5 for shallow feedforward neural networks. However, as already mentioned above, they hold for the set of deep neural networks, too. For a discussion on the topic of depth versus width, see for example Lu et al. [41] and Ronen and Ohad [51].

3.1 Interlude: universal approximation in a Hölder-norm

In Theorem 3.3, we cited classical versions of the universal approximation theorem which are concerned with (almost everywhere) uniform approximation of continuous functions. Note that this is in essence a topological statement, and we may seek for refined approximation results that hold with respect to stricter topologies. Natural candidate topologies with respect to which we may seek to derive a universal approximation theorem are Hölder-type topologies.

Given $\alpha \in (0, 1)$, we denote by $E^\alpha = C_0^\alpha([0, T]; \mathbb{R}^d) \subseteq C_0([0, T]; \mathbb{R}^d)$ the vector space of $\mathbb{R}^d$-valued, α-Hölder-continuous functions on $[0, T]$ that are zero at the origin. The space E^α is also referred to as α-Hölder space. We endow this space with the topology which is induced by the norm

$$E^\alpha \ni f \mapsto \|f\|_\alpha := \sup_{\substack{s,t\in[0,T]\\ 0<t-s\leq 1}} \frac{|f(t) - f(s)|}{(t-s)^\alpha}.$$

The Kolmogorov–Chentsov continuity theorem shows that all paths of $\mathbb{R}^d$-valued standard Brownian motion are α-Hölder-continuous for every $\alpha \in (0, 1/2)$; hence it would be desirable to use E^α as the space on which to consider the restriction of the classical Wiener measure (see Example 2.1 and Definition A.1). Although $(E^\alpha, \|\cdot\|_\alpha)$ is indeed a real Banach space, it does not contain a countable dense subset and is thus not separable (cf. Friz and Victoir [17, Theorem 5.25]).

We need to pass to the little α-Hölder space, i.e., the subspace $E^{\alpha,0}$ of all $f \in E^\alpha$ that satisfy $|f(t) - f(s)| = o(|t-s|^\alpha)$ as $|t-s| \searrow 0$. Then $(E^{\alpha,0}, \|\cdot\|_\alpha)$ is a real Banach space. The space $E^{\alpha,0}$ is also referred to as the space of α-Hölder paths with vanishing Hölder oscillation (cf. Friz and Hairer [16, Exercise 2.12]). Note that $E^{\alpha,0}$ has a very useful characterisation: It is the closure of $C_0^\infty([0,T];\mathbb{R}^d)$, the vector space of $\mathbb{R}^d$-valued smooth functions on $[0,T]$ that are zero at the origin, where the closure is taken with respect to the topology induced by $\|\cdot\|_\alpha$. Moreover, we have the inclusion $E^\beta \subseteq E^{\alpha,0}$ for all $0 < \alpha < \beta < 1$.

In the context of Gaussian measures and large deviations theory, the space $E^{\alpha,0}$ has been studied in great detail; cf. Andresen et al. [1], Baldi et al. [3] and Ciesielski [9]. In particular, the following important property has been shown in [1] to hold: If we fix $\pi = I_d$ and $\mu = \lambda$, then H is continuously embedded into $E^{\alpha,0}$ for each $\alpha \in (0, 1/2)$, and there exists a countable family of functions in H (the Faber–Schauder system) that constitutes a Schauder basis of $(E^{\alpha,0}, \|\cdot\|_\alpha)$. While on the one hand this implies the separability of $E^{\alpha,0}$, more importantly, we see that H is not only continuously, but also densely embedded into $E^{\alpha,0}$. In conjunction with Theorem 2.24 and as a direct consequence of this observation, we obtain the following result.

Proposition 3.8 *Fix $\mu = \lambda$, $\pi = I_d$ and $\alpha, \beta \in (0, 1/2)$ with $\alpha < \beta$. Then in the context of Propositions 3.4 and 3.5, for each $f \in E^{\alpha,0}$, there exists a sequence $(f_n)_{n\in\mathbb{N}}$ in $\mathcal{NN}^d_{1,\infty}(\psi)$ such that*

$$\lim_{n\to\infty} \|f - J(f_n)\|_\alpha = 0.$$

In particular, every smooth function $f \in C_0^\infty([0,T];\mathbb{R}^d) \subset E^{\alpha,0}$ and every β-Hölder-continuous function $f \in E^\beta \subseteq E^{\alpha,0}$ can be approximated, up to the linear isometry J (see Definition 2.13), by sequences from $\mathcal{NN}^d_{1,\infty}(\psi)$ with respect to $\|\cdot\|_\alpha$.

Let us conclude this subsection with several remarks. First note that based on Ciesielski et al. [10], Proposition 3.8 should extend to certain Besov–Orlicz-type norms, which induce stricter topologies than the Hölder-norms. Moreover, it should be possible to relax the assumption $\mu = \lambda$ by considering a modified Hölder-norm with denominator $\mu((s,t])^\alpha$ instead of $(t-s)^\alpha$. Finally, note that the universal approximation property of neural networks in topological spaces with topologies that are stricter than that induced by the uniform norm have already been studied in the literature. See e.g. Gühring et al. [25] who study the UAP in Sobolev spaces.

4 Importance sampling with feedforward neural networks

Having studied the tractable space H of drift adjustments which coincides with the Cameron–Martin space of the Gaussian measure γ_M, and having proved that feedforward neural networks are, up to the isometry J, dense in H by combining Theorem 2.24 and Propositions 3.4 and 3.5, we now turn our attention to importance sampling. To this end, we first write down the basic setting. In Sect. 4.1, we then study how our method complements a classical approach which employs ideas from

large deviations theory, before finally studying the full problem in Sect. 4.2. Most notably, Theorem 4.6 below provides a theoretical justification for our simulations in Sect. 5; see also Remark 4.7.

Let $\mathcal{C}$ denote the vector space of $\mathbb{R}^n$-valued, continuous and $\mathbb{F}$-adapted processes. Let $a\colon \Omega \times [0,T] \times \mathcal{C} \to \mathbb{R}^n$ and $b\colon \Omega \times [0,T] \times \mathcal{C} \to \mathbb{R}^{n\times d}$ be non-anticipative coefficients (cf. Cohen and Elliott [12, Definition 16.0.3]) with the property that the stochastic differential equation

$$\mathrm{d}X_t = a_t(X)\,\mathrm{d}C_t + b_t(X)\,\mathrm{d}M_t, \qquad t \in [0,T], \tag{4.1}$$

with $X_0 \equiv x \in \mathbb{R}^n$ admits a unique weak solution. For ease of notation, we write $a_t(X), b_t(X)$ instead of $a(\omega,t,X), b(\omega,t,X)$ and interpret (4.1) to hold componentwise, i.e., $X_t^i = x^i + (a(X)^i \bullet C)_t + (b(X)^{i,\cdot} \bullet M)_t$ for $i \in \{1,2,\ldots,n\}$ and $t \in [0,T]$.

Let F be a real-valued random functional on $\Omega \times C([0,T];\mathbb{R}^n)$ with the property that the mapping $\Omega \ni \omega \mapsto F(\omega, X(\omega))$ is $\mathcal{F}_T$-measurable. For simplicity, we write $F(\cdot, X) = F(X)$ and call $F(X)$ a random payoff. We are interested in obtaining a Monte Carlo estimate of its expectation under $\mathbb{P}$,

$$\mathbb{E}_{\mathbb{P}}[F(X)] = \int_\Omega F\big(\omega, X(\omega)\big)\mathbb{P}(\mathrm{d}\omega), \tag{4.2}$$

provided that (4.2) is well defined in $\mathbb{R}$.

Remark 4.1 If $\mathbb{E}_{\mathbb{P}}[F(X)]$ is to denote an option price, then we should require $\mathbb{P}$ to be a risk-neutral measure. However, the results in this section do not require $\mathbb{P}$ to be risk-neutral. Actually, we do not need to assume $\mathbb{E}_{\mathbb{P}}[F(X)]$ to be an option price as long as X follows the SDE (4.1) and $F(X)$ is $\mathcal{F}_T$-measurable.

Example 4.2 The SDE (4.1) can model the evolution of asset prices within stochastic volatility models. Thus our method complements Robertson [50] who uses methods from the theory of large deviations to derive asymptotically optimal drift adjustments (see Sect. 4.1, where we discuss asymptotic optimality in the context of Guasoni and Robertson [24]) for pricing stochastic volatility models, very much in the spirit of [24].

Let $(X^i)_{i\in\mathbb{N}}$ denote a sequence of independent copies of solutions to (4.1). By the strong law of large numbers, the sample means $Z_k = \sum_{i=1}^k F(X^i)/k$ converge $\mathbb{P}$-almost surely to $m = \mathbb{E}_{\mathbb{P}}[F(X)]$. Moreover, if $F(X^1)$ has a finite variance $\sigma^2 > 0$, then according to the central limit theorem, as $k \to \infty$, the law of $\sqrt{k}(Z_k - m)$ converges weakly to $\mathcal{N}(0,\sigma^2)$. We therefore see that $Z_k - m$ is approximately normally distributed with mean zero and standard deviation $\sigma/\sqrt{k}$. In practice, the standard error $\sigma/\sqrt{k}$ can be large when the random payoff $F(X)$ has a high variance, even for large sample sizes k, which calls for the application of variance reduction methods.

Remark 4.3 For each $f \in \Lambda^2$, we have that $f \bullet M$ is a real-valued continuous local martingale with $[f \bullet M]_t = \int_0^t f^\top(s)\pi(s)f(s)\mu(\mathrm{d}s)$ for $t \in [0,T]$. As a

consequence, similarly as in Remark 2.4, $f \bullet M$ is a Gaussian $\mathbb{F}$-Markov process with

$$\mathcal{L}\big((f \bullet M)_t - (f \bullet M)_s\big) = \mathcal{N}\left(0, \int_s^t f^\top(u)\pi(u)f(u)\mu(\mathrm{d}u)\right)$$

for $s < t$ in $[0, T]$, which shows how one can simulate increments of $f \bullet M$, provided that the integrals $\int_s^t f^\top(u)\pi(u)f(u)\mu(\mathrm{d}u)$ can be explicitly computed.

Recall that for any real-valued continuous semimartingale Y, the Doléans–Dade exponential $\mathcal{E}(Y)$ is the strictly positive continuous semimartingale that is, up to indistinguishability, the unique solution to the stochastic integral equation

$$\mathcal{E}(Y) = \exp(Y_0) + \int_0^\cdot \mathcal{E}(Y)_s \, \mathrm{d}Y_s$$

and is given by $\mathcal{E}(Y) = \exp(Y - [Y]/2)$. A direct computation also gives $\mathcal{E}(Y)^{-1} = \mathcal{E}(-Y + [Y])$.

For each $h \in H$, the process $N = f_h \bullet M$ is a Gaussian process whose quadratic variation $[N]$ is deterministic. Therefore N_T has all exponential moments, $\mathcal{E}(N)_T$ is in L^p for every finite p, and an application of Novikov's criterion shows that $\mathcal{E}(N)$ is uniformly integrable. By a change of measure, (4.2) can now be rewritten as

$$\mathbb{E}_{\mathbb{P}}[F(X)] = \mathbb{E}_{\mathbb{P}_h}\big[F(X)\big(\mathcal{E}(f_h \bullet M)_T\big)^{-1}\big], \tag{4.3}$$

where $\mathbb{P}_h$ is defined by $\mathrm{d}\mathbb{P}_h = \mathcal{E}(f_h \bullet M)_T \, \mathrm{d}\mathbb{P}$ on $\mathcal{F}_T$. If we denote the modified random payoff by $F_h(X) := F(X)(\mathcal{E}(f_h \bullet M)_T)^{-1}$, then the $\mathbb{P}$-expectation of $F(X)$ and the $\mathbb{P}_h$-expectation of $F_h(X)$ are identical. If $F(X)$ has a finite second moment with respect to $\mathbb{P}$, the variance of $F_h(X)$ under $\mathbb{P}_h$ is given by

$$\mathbb{E}_{\mathbb{P}_h}[F_h^2(X)] - \mathbb{E}_{\mathbb{P}_h}[F_h(X)]^2 = \mathbb{E}_{\mathbb{P}}\big[F^2(X)\big(\mathcal{E}(f_h \bullet M)_T\big)^{-1}\big] - \mathbb{E}_{\mathbb{P}}[F(X)]^2. \tag{4.4}$$

Therefore we can compute (4.3) under the measure $\mathbb{P}_h$ and try to find $h \in H$ such that (4.4) is minimised. Note that the second term on the right-hand side of (4.4) does not depend on h; so we focus on minimising the first term, which for each $h \in H$ is given by

$$V(h) := \mathbb{E}_{\mathbb{P}}\big[F^2(X)\big(\mathcal{E}(f_h \bullet M)_T\big)^{-1}\big] = \mathbb{E}_{\mathbb{P}}\big[F^2(X)\exp\big(-(f_h \bullet M)_T + \|h\|_H^2/2\big)\big].$$

To sum up, our problem reads

$$\min_{h \in H} V(h), \tag{4.5}$$

provided that a minimiser of V exists; see Theorem 4.6 for sufficient conditions.

4.1 Approximating the asymptotically optimal sampling measure

Before we turn our attention to solving (4.5), we first discuss the classical approach presented in Guasoni and Robertson [24] that uses methods from the theory of large deviations, and show how our method complements it. Note that the setting of [24]

is a special case of the setting of Sect. 4. To see this, set $d = 1$, let M be a standard Brownian motion and $\mathbb{F}$ the augmented natural filtration of M (which satisfies the usual hypotheses). Set $X = M$ and assume that the payoff $F\colon C_0([0,T];\mathbb{R}) \to \mathbb{R}_+$ is continuous, where $C_0([0,T];\mathbb{R})$ has the topology of uniform convergence.

In [24], the authors argue that (4.5) is in general intractable. Rather than solving (4.5), they consider for each $h \in H$ the small-noise limit

$$L(h) := \limsup_{\epsilon \searrow 0} \epsilon \log \mathbb{E}_{\mathbb{P}}\left[\exp\left(\frac{1}{\epsilon}\left(2\tilde{F}(\sqrt{\epsilon}M) - \left((\sqrt{\epsilon} f_h) \bullet M\right)_T + \frac{\|h\|_H^2}{2} \right)\right)\right], \quad (4.6)$$

where $\tilde{F} := \log F$. The limit (4.6) corresponds to approximating $V(h) \approx \exp(L(h))$.

Assume that $\tilde{F}\colon C_0([0,T];\mathbb{R}) \to \mathbb{R} \cup \{-\infty\}$ is continuous. Moreover, assume that there exist constants $K_1, K_2 > 0$ as well as $\alpha \in (0,2)$ such that $\tilde{F}(x) \le K_1 + K_2 \|x\|_\infty^\alpha$ for each $x \in C_0([0,T];\mathbb{R})$. As [24, Theorem 3.6] shows, one can invoke a version of Varadhan's integral lemma to rewrite (4.6) as a variational problem, provided that h is an element of H_{bv}, the space of all $h \in H$ such that $f_h \in \Lambda^2 = L^2(\lambda)$ is of bounded variation, and aim to solve $\min_{h \in H_{\mathrm{bv}}} L(h)$, provided that a minimiser exists.

For the proof of the central result in [24, Theorem 3.6], the following functional is important: For $c > 0$ and $h \in H$, let $\tilde{F}_{h,c}\colon H \ni g \mapsto 2\tilde{F}(g) - c\|g+h\|_H^2 + \|h\|_H^2$. By [24, Lemma 7.1], there exists a maximiser $g_{h,c} \in H$ of $\tilde{F}_{h,c}$. Together with Proposition 3.4, we then obtain

Proposition 4.4 *Assume that $\psi\colon \mathbb{R} \to \mathbb{R}$ is continuously differentiable, bounded and nonconstant, and set $D = \mathcal{NN}^1_{1,\infty}(\psi)$. Then $H(D) \subseteq H_{\mathrm{bv}}$, and for each $c > 0$ and $h \in H$, there exists a sequence $(h_n)_{n\in\mathbb{N}}$ in $H(D)$ such that*

$$\lim_{n\to\infty} \tilde{F}_{h,c}(h_n) = \tilde{F}_{h,c}(g_{h,c}) = \max_{g\in H}\left(2\tilde{F}(g) - c\|g+h\|_H^2 + \|h\|_H^2\right).$$

According to [24, Theorem 3.6] and the discussion thereafter, the strategy for finding a minimiser of L is as follows: Find a maximiser $g_{h,1}$ to $\tilde{F}_{h,1}$ for $h \equiv 0$. Check whether $g_{h,1}$ is actually an element of H_{bv}, and if this is the case, then $g_{h,1}$ minimises L provided that $L(g_{h,1}) = \tilde{F}_{h,1}(g_{h,1})$ holds true. Since the evaluation of L at $g_{h,1}$ involves having to find a maximiser of $\tilde{F}_{g_{h,1},1/2}$, checking whether $L(g_{h,1}) = \tilde{F}_{h,1}(g_{h,1})$ might only be feasible by a numerical approximation which introduces an error. However, if one could establish said identity, then $g_{h,1}$ would be a minimiser of L, in which case we say that $g_{h,1}$ is asymptotically optimal.

Let us assume that there exists a maximiser $g_{h,1}$ to $\tilde{F}_{h,1}$ for $h \equiv 0$ that is indeed asymptotically optimal. In view of Proposition 4.4, we can then approximate $g_{h,1}$ by a sequence $(h_n)_{n\in\mathbb{N}}$ from $H(\mathcal{NN}^1_{1,\infty}(\psi))$ such that $\tilde{F}_{h,1}(h_n)$ converges to $\tilde{F}_{h,1}(g_{h,1})$. Theorem 4.6(b) below then implies that $V(h_n)$ converges to $V(g_{h,1})$. To sum up, rather than trying to find a minimiser of V, one might instead study

$$\max_{g\in H} \tilde{F}_{h,1}(g) = \max_{g\in H}\left(2\tilde{F}(g) - \|g\|_H^2\right) \quad (4.7)$$

and solve the modified problem (4.7) with feedforward neural networks.

4.2 Approximating the optimal sampling measure

In what follows, we consider the full problem (4.5) and propose to solve it with feedforward neural networks. Moreover, Theorem 4.6(d) and Remark 4.7 provide a theoretical justification for employing the tractable class of shallow feedforward neural networks for this optimisation problem. The numerical simulations in Sect. 5 will demonstrate that we obtain indeed substantial reductions in the variance of the Monte Carlo estimators for several multivariate asset price processes and path-dependent payoff functionals.

In the following result, we consider a nonlinear operator which maps elements from the Cameron–Martin space to probability densities. This result is essential as it implies in Theorem 4.6 below that the optimal sampling measure can be approximated by measures which are generated by feedforward neural networks.

Lemma 4.5 *The operator* $A_p \colon H \ni h \mapsto (\mathcal{E}(f_h \bullet M)_T)^{-1} \in L^p(\mathbb{P})$ *is continuous for each* $p \in [1, \infty)$. *Moreover,* A_p *is not quasi-bounded, meaning that*

$$\limsup_{\|h\|_H \to \infty} \frac{\|A_p(h)\|_{L^p(\mathbb{P})}}{\|h\|_H} = \infty.$$

Finally, we formulate Theorem 4.6. The proof (presented in Appendix B) complements Lemaire and Pagès [35, Proposition 4] and not only shows under rather weak assumptions that the functional V does indeed admit a minimiser. Theorem 4.6(d) applied to a dense subset D of Λ^2 which consists of feedforward neural networks also provides the theoretical justification for the simulations in Sect. 5; see also Remark 4.7 below.

Theorem 4.6 *Assume that* $\mathbb{P}[F^2(X) > 0] > 0$ *and that there exists some* $\varepsilon > 0$ *such that* $F(X) \in L^{2+\epsilon}(\mathbb{P})$. *Then*:

(a) V *is* $\mathbb{R}_+$*-valued.*

(b) V *is continuous.*

(c) *There exists a minimiser of* V, *i.e.*,

$$\arg\min_{h \in H} V(h) = \{g \in H : V(g) \leq V(h), \forall h \in H\} \neq \emptyset.$$

(d) *There exists a sequence* $(h_n)_{n \in \mathbb{N}}$ *in* $H(D)$ *such that*

$$\lim_{n \to \infty} V(h_n) = \min_{h \in H} V(h).$$

Remark 4.7 In the context of Theorem 4.6(d), we may seek to find a minimiser of V by performing measure changes induced by Doléans–Dade exponentials of the form $\mathcal{E}(f \bullet M)$, where $f \in \mathcal{NN}^d_{1,\infty}(\psi)$. In Sect. 5, we pursue this approach for several different asset price models, achieving substantial reductions in the variance of the corresponding Monte Carlo estimators.

Remark 4.8 Theorem 4.6(d) shows that neural-network-induced changes of the sampling measure can approximate the optimal sampling measure arbitrarily well in the sense that the second moment of the modified payoff under the optimal measure can be approximated up to an arbitrarily small $\epsilon > 0$. However, the proof is not constructive; it does not deliver a recipe how to actually obtain such a sequence (h_n) of neural-network-induced elements from the Cameron–Martin space that converges to the optimum. In Sect. 5 below, we use stochastic gradient descent to train our neural networks. This procedure builds on the method of stochastic approximation, which was pioneered in 1951 by Robbins and Monro [49]. Stochastic approximation for importance sampling for option pricing in continuous-time models has been studied by Lemaire and Pagès [35]. We refer to [35, Sect. 3] for details on how to construct convergent sequences of functions based on the method of stochastic approximation.

Remark 4.9 Let us assume that the SDE (4.1) depends on a set of parameters $\alpha \in \mathbb{R}^m$ for some $m \in \mathbb{N}$. Fix $i \in \{1, 2, \dots, m\}$ and further assume that we can exchange the order of differentiation and integration, i.e., $\frac{\partial}{\partial \alpha_i}\mathbb{E}_{\mathbb{P}}[F(X)] = \mathbb{E}_{\mathbb{P}}[\frac{\partial}{\partial \alpha_i}F(X)]$. If we want to jointly reduce the standard error of the Monte Carlo estimators of the expected random payoff and of its sensitivity with respect to α_i, we could modify the definition of V to

$$\tilde{V}(h) = \mathbb{E}_{\mathbb{P}}\left[\left(w_1 F^2(X) + w_2\Big(\frac{\partial}{\partial \alpha_i}F(X)\Big)^2\right)\big(\mathcal{E}(f_h \bullet M)_T\big)^{-1}\right], \qquad h \in H,$$

where $w_1, w_2 \in (0, 1)$ are weights that sum up to 1. If there exists some $\varepsilon > 0$ with $w_1 F^2(X) + w_2(\frac{\partial}{\partial \alpha_i}F(X))^2 \in L^{1+\varepsilon}(\mathbb{P})$ and $\mathbb{P}[w_1 F^2(X) + w_2(\frac{\partial}{\partial \alpha_i}F(X))^2 > 0] > 0$, then Theorem 4.6 applies correspondingly. Analogous considerations hold for higher-order sensitivities as well as for the joint reduction of standard errors for more than one sensitivity. We refer the reader to Glasserman [20, Sect. 7.2] for details on the computation of pathwise derivatives for some classical models and payoffs.

5 Numerical study

In this section, we provide a range of carefully chosen numerical examples to showcase the various strengths of our method. Additionally, we compare our approach to other methods that have been proposed in the literature. All computational tasks were performed using Python, leveraging the Keras deep learning API for the construction and training of our neural networks. All codes that were used for the simulations are available on Github; see https://github.com/aarandjel/importance-sampling-with-feedforward-neural-networks.

Let us provide a brief overview of the examples appearing in the subsequent subsections. In Sect. 5.2, we explore a time-change instance that deviates from the conventional assumption of $\mu = \lambda$ to better represent phases of changing business activity. Section 5.3 considers a knock-out option and discusses the occurrence of multiple rare events. Moving on to Sect. 5.4, we examine a stochastic volatility model with an imposed dynamic correlation structure, which directly influences the norm on the

Cameron–Martin space. Lastly, in Sect. 5.5, we investigate the feasibility of utilising neural networks for importance sampling in a high-dimensional model. **In all our examples**, we consider arithmetic Asian (basket) call options with strike K and basket weights w as the chosen payoffs, i.e.,

$$F(X) = \left(\frac{1}{T} \int_0^T \langle w, X_t \rangle \, \mathrm{d}t - K \right)^+,$$

while Sect. 5.3 additionally incorporates knock-out barriers for further analysis.

To establish a solid basis for comparison, we have selected the methodologies proposed by Glasserman et al. [21], Guasoni and Robertson [24], Capriotti [7], Arouna [2], Su and Fu [54] as well as Jourdain and Lelong [35]. To underscore the versatility of our approach in handling more general models than those presented in the literature, we initially present results for the models discussed in the previous paragraph. Subsequently, we report results from simulations performed for the models studied in the literature mentioned above.

To train a feedforward neural network, our approach is as follows. First, we simulate N trajectories X^i, $i = 1, \ldots, N$, of the asset price using the Euler–Maruyama method, based on a pre-defined time-grid. Then we decide on a set $\mathcal{NN}^d_{k,\ell}(\psi)$ from which we seek to identify the optimal function by selecting the number of hidden layers k, the number of hidden nodes ℓ and the activation function ψ. The output dimension d of the neural networks aligns with the dimension of the process M. We approximate V by computing an average over the N trajectories,

$$V(\theta) = \frac{1}{N} \sum_{i=1}^{N} F^2(X^i) \exp\big(- (f_\theta \bullet M^i)_T + \| f_\theta \|^2_{\Lambda^2}/2 \big), \tag{5.1}$$

where θ represents the vector encompassing all trainable parameters of the neural network f_θ and all quantities on the right-hand side of (5.1) are appropriately discretised. We therefore consider V as a function of the finite parameter vector θ and aim to find the optimal θ^* and thus the optimal element f_{θ^*} from $\mathcal{NN}^d_{k,\ell}(\psi)$.

To achieve this, we employ stochastic gradient descent, a technique originally pioneered by Robbins and Monro [49]. Specifically, we adopt the mini-batch variant of this method, which replaces the mean over all N trajectories with means over smaller sub-batches. Starting from an initial guess, the parameter vector θ is then iteratively updated with a scaled version of the gradient of V over those sub-batches, i.e., $\theta_{m+1} = \theta_m - \gamma_m \nabla_{\text{batch}} V(\theta_m)$ with learning rate γ_m and $\nabla_{\text{batch}} V$ denoting the gradient of V over one specific batch. Upon completing a full iteration through all batches, we consider the neural network to have completed one epoch of training. For each subsequent epoch, the trajectories contained in the individual batches can then be randomly shuffled around, and the parameter θ is updated until a stopping criterion is reached. One notable advantage of neural networks lies in their ability to efficiently compute gradients through the back-propagation method. Additionally, we utilise a popular modified version of this training routine known as Adam (cf. Kingma and Ba [33]), which incorporates the first and second moments of the gradient estimates to enhance performance.

In all our subsequent examples, we train the neural networks using 100 batches, each consisting of 1024 trajectories. For validation purposes, we employ an additional 100 batches, also comprising 1024 trajectories, and stop the training process when the loss $V(\theta)$ ceases to reduce on the validation set. The results presented in the following tables are derived from simulations performed on separate test datasets, each containing 10^5 trajectories. Throughout the training, validation and testing phases, we maintain a fixed learning rate of 10^{-3} for the stochastic gradient descent, and we fix the time horizon to $T = 1$ to consider the time interval $[0, 1]$. Unless otherwise specified, we utilise a step size of $\Delta t = 1/250$. However, in Sect. 5.5, we deviate from this convention. We employ a step size corresponding to $\Delta t = d/10^4$ during the training and validation process, where d denotes the dimension of the asset price process. For example, when $d = 200$, then $\Delta t = 1/50$. This adjustment is only implemented for dimensions ranging from $d = 100$ to $d = 1000$, while the step size always remains $\Delta t = 1/100$ for the testing dataset, as well as for the training and validation datasets in case $d < 100$. In all simulations described below, we train shallow feedforward neural networks with a single hidden layer, using $\psi(x) = \tanh(x)$ as activation function. The number of hidden nodes used for the various examples is reported beneath the tables. The tables below present results for different choices of model parameters, presenting mean estimates, standard errors, relative standard errors as a percentage of the mean and variance ratios. The variance ratios were obtained by comparing the variance of the mean estimate from both a Monte Carlo and a Monte Carlo with importance sampling run, dividing the former by the estimate of the latter.

5.1 Stratified sampling with feedforward neural networks

In addition to importance sampling, stratified sampling is a widely used variance reduction method. Stratified sampling involves constraining the fraction of trajectories sampled from specific subsets of the sample space. To implement this method effectively, suitable subsets of the sample space need to be chosen, covering the entire sample space, along with the desired fractions of the overall sample falling within each subset. It is important to note that stratification typically generates dependent sequences of random variables, which affects the calculation of the standard error and variance of the Monte Carlo estimator. For further information on this approach, we refer to Glasserman [20, Sect. 4.3].

In Glasserman et al. [21], the authors investigate importance sampling and stratification techniques for pricing path-dependent options. Similarly to Guasoni and Robertson [24], they employ large deviations techniques to determine asymptotically optimal drift adjustments in a discrete-time framework. In order to overcome the computational effort that might be required to perform optimal stratification, the authors of [24] propose utilising the drift identified for importance sampling to perform further stratification. In the following examples, we augment our results based on importance sampling with the stratified sampling approach.

More precisely, let us consider the estimation of $\mathbb{E}[F(X)]$. Having discretised the time interval into m points, assume that $F(X)$ can be expressed as a function of Z, with Z being an m-dimensional vector of independent standard normal variables.

If f denotes the optimal element from the Cameron–Martin space, sampled at t_i as a vector and appropriately rescaled so that adjusting the drift of M corresponds to adding f to Z in the discrete-time case, we want to sample Z conditionally on $\langle f, Z\rangle \in A_i$, where A_i denotes a stratum (a subset of the sample space). In our case, A_i is chosen to correspond to the interval between the $(i-1)/N$- and the i/N-quantile of the standard normal distribution, where N denotes the number of strata. We maintain an equal number of replications for each stratum. For further details on simulating Z conditionally on $\langle f, Z\rangle \in A_i$, we refer to [21, Sect. 4].

In Sects. 5.2–5.4, we extend our analysis beyond importance sampling by additionally using the trained neural networks to implement stratified sampling. By combining these two techniques, we demonstrate the significant potential for further variance reduction. It is crucial to emphasise that using the optimal importance sampling drift for stratification may not always result in optimal stratification in general. Furthermore, it is worth noting that the setting of [21] is in discrete time. There is ample scope to explore optimal stratified sampling in continuous time using neural networks.

5.2 Changing business activity

Methods typically employed for importance sampling based on continuous stochastic processes for asset prices often assume that the dynamics of the asset price are governed by an SDE driven by a Brownian motion. Here, we aim to deviate from the conventional framework where $\mu = \lambda$ and explore an example involving a time-changed Brownian motion. It is important to note that in this case, the time-change directly affects the definition of the Cameron–Martin norm through the Lebesgue–Stieltjes measure μ. The use of a deterministic time-change can be interpreted as a means of modelling periods characterised by varying business activity, thus incorporating effects such as seasonality. See Li et al. [37] for an example where this has been done.

Consider the asset X governed by the dynamics $\mathrm{d}X_t = rX_t\,\mathrm{d}[M]_t + \sigma X_t\,\mathrm{d}M_t$, where $X_0 = x$ and $M_t = B_{C_t}$ for B representing a standard Brownian motion. Motivated by [37], we make the assumption that $C_t = \int_0^t \nu(s)\,\mathrm{d}s$, where the activity rate function ν takes the form

$$\nu(s) = \begin{cases} 1 + \kappa(s - 0.2)/0.1, & s \in [0.2, 0.3), \\ 1 + \kappa(0.4 - s)/0.1, & s \in [0.3, 0.4), \\ 1 + 2\kappa(s - 0.6)/0.1, & s \in [0.6, 0.7), \\ 1 + 2\kappa(0.8 - s)/0.1, & s \in [0.7, 0.8), \\ 1, & \text{else}, \end{cases}$$

where κ denotes the level of business activity. Moreover, we normalise ν such that $C_1 = \int_0^1 \nu(s)\,\mathrm{d}s = 1$. In this case, μ is absolutely continuous with respect to λ with Radon–Nikodým density ν, and $[M] = C$.

Figure 1 illustrates a representative trajectory of X under the assumption of an activity rate function modelled by $\kappa = 10$. The trajectory exhibits two distinct phases

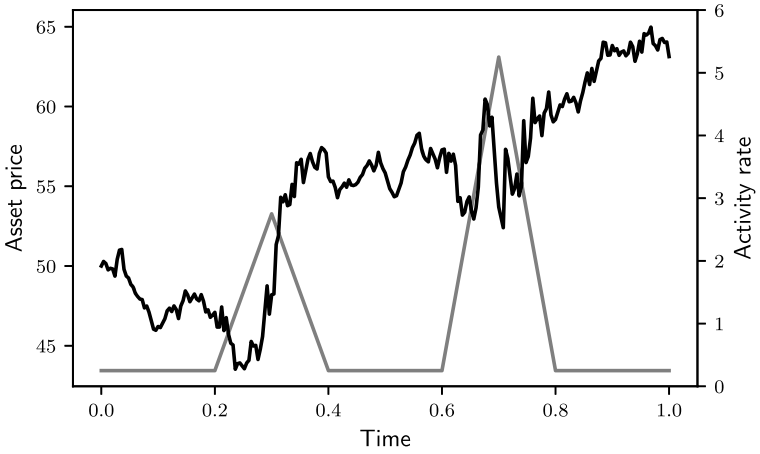

Fig. 1 Typical sample path for the model described above with $\kappa = 10$, along with the corresponding activity rate function ν. Other model parameters are $X_0 = 50$, $r = 0.05$ and $\sigma = 0.25$

Table 1 Variance ratios for different levels κ of business activity

Parameter	Importance sampling (IS)			IS and stratification		
κ	Mean	Std. err.	Var. ratio	Mean	Std. err.	Var. ratio
0	5.945	0.019 (0.32%)	129	5.9337	0.0018 (0.03%)	14,525
1	4.675	0.015 (0.32%)	154	4.6672	0.0023 (0.05%)	6,714
2	3.987	0.013 (0.33%)	167	3.9860	0.0026 (0.07%)	4,191
5	3.053	0.010 (0.33%)	207	3.0495	0.0016 (0.05%)	8,139
10	2.5286	0.0086 (0.34%)	235	2.5304	0.0016 (0.06%)	7,110

Note: Option prices and standard errors are quoted in cents. Only significant digits are reported. Number of hidden nodes is 2. Other model parameters are $X_0 = 50$, $r = 0.05$, $\sigma = 0.25$ and $K = 70$.

characterised by heightened volatility, which can be interpreted as periods of increased business activity. Table 1 presents results obtained for various values of κ. Note that the special case $\kappa = 0$ corresponds to the classical Black–Scholes model.

From Table 1, it is evident that both importance sampling and the combined approach of importance and stratified sampling exhibit substantial variance reduction across all values of κ. Notably, the combination of importance and stratified sampling demonstrates a remarkable enhancement in variance reduction compared to using importance sampling alone.

In Guasoni and Robertson [24], the authors study asymptotically optimal importance sampling in continuous time following a large deviations approach. Table 2 of [24] presents variance ratios for an arithmetic Asian call option within a Black–Scholes model across various values of volatility (σ) and strike (K). We refer to [24, Sect. 5] for details about the model and the selected parameters. We replicated [24, Table 2] using neural networks to induce optimal measure changes and subsequently compared the obtained variance ratios. On average, employing neural networks re-

sulted in a 20% increase in the variance ratio. For instance, when considering a volatility of 30% and a strike of 70, the authors of [24] report a variance ratio of 56 while our method yielded a variance ratio of 67.

Capriotti [7] studies importance sampling based on a least-squares optimisation procedure. Table 6 of [7] presents variance ratios for various combinations of σ and K, specifically for an arithmetic Asian call option within a Black–Scholes model. Additionally, the table includes variance ratios obtained using an adaptive Robbins–Monro procedure as proposed in Arouna [2] for the same set of model parameters. We replicated [7, Table 6] using our method and compared the resulting variance ratios. As it turns out, our method yields average variance ratios that are 10% and 95% higher than the values reported by [7] and [2], respectively.

Finally, [7, Table 7] provides the results for a partial average Asian call option as previously investigated in Su and Fu [54]. For detailed definitions of the models and parameters utilised in the simulations, we refer to [7, Sect. 5]. We implemented this particular model using our method. On average, our approach yielded variance ratios that were 10% and 50% higher than the values reported by [7] and [54], respectively.

5.3 Multiple rare events

In Glasserman and Wang [22], the authors emphasise that rare events often consist of unions of meaningful events that represent different ways in which the rare event can occur. In this context, we aim to examine an example where the rare event is formed by the intersection of two rare events. We also discuss the case of the union of rare events later on. An illustrative example is provided by knock-out call options, which exhibit a classical scenario where the payoff is discontinuous with respect to the asset price trajectory. In this case, two potentially rare events can arise: (1) the arithmetic average $\bar{X}_t = \int_0^t \langle w, X_s \rangle \, \mathrm{d}s$ must be above the strike at the terminal time, and (2) the option must not be knocked out.

Consider an asset price X that follows a classical Black–Scholes model, characterised by the SDE $\mathrm{d}X_t = r X_t \, \mathrm{d}t + \sigma X_t \, \mathrm{d}B_t$ with an initial value of $X_0 = x$, where B denotes a Brownian motion. In contrast to Sect. 5.2, we introduce knock-out barriers L, U that satisfy $0 < L < X_0 < K < U$. The option is considered knocked out if the arithmetic average $\bar{X}_t$ breaches either of the two barriers at any given point in time before or at maturity. In our example, there is a delicate balance which needs to be achieved between giving the asset a positive drift such that $\bar{X}_1 > K$ with sufficiently high probability, and making sure that the option is not knocked out.

In Fig. 2, we provide a graphical representation of the learning process of the neural network. On a fixed dataset, we calculate the probability of the arithmetic average ending up above the strike K, the probability of it remaining between the knock-out barriers at all times, as well as the variance ratio after each epoch that the neural network was trained. Table 2 provides a comprehensive overview of the variance ratios corresponding to different values of strikes K and upper knock-out barriers U.

Figure 2 highlights an interesting observation: increasing the variance ratio does not simply result from an indiscriminate rise in the probabilities of both rare events occurring. Instead, it becomes evident that a delicate balance between the occurrence

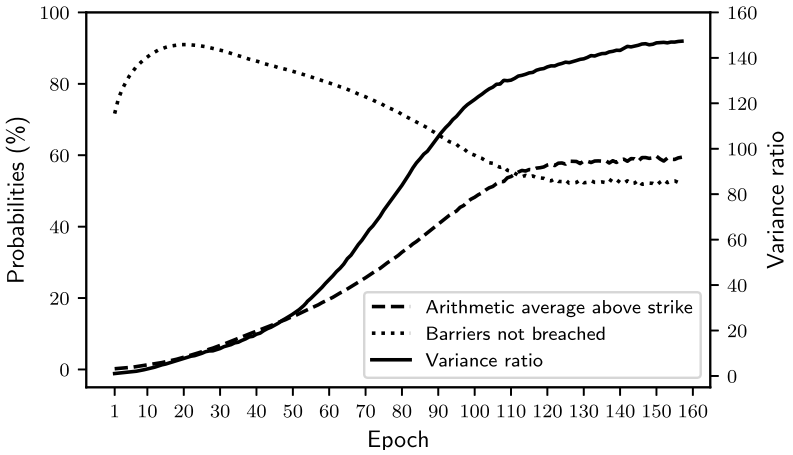

Fig. 2 A graphical representation of the learning process of the neural network

Table 2 Variance ratios for different values of strike K, volatility σ and upper knock-out barrier U

Parameters			Importance sampling (IS)			IS and stratification		
K	σ	U	Mean	Std. err.	Var. ratio	Mean	Std. err.	Var. ratio
60	0.2	70	0.763	0.013 (1.70%)	7	0.779	0.012 (1.54%)	7
		80	12.605	0.067 (0.53%)	10	12.588	0.049 (0.39%)	18
		90	22.607	0.078 (0.35%)	18	22.673	0.032 (0.14%)	112
	0.3	70	0.1826	0.0082 (4.49%)	3	0.1816	0.0080 (4.41%)	3
		80	13.65	0.12 (0.88%)	4	13.60	0.11 (0.81%)	4
		90	42.86	0.22 (0.51%)	5	42.89	0.16 (0.37%)	9
70	0.2	80	0.000775	0.000041 (5.29%)	356	0.000760	0.000041 (5.39%)	357
		90	0.1917	0.0018 (0.94%)	144	0.1921	0.0016 (0.83%)	189
		100	0.6473	0.0035 (0.54%)	203	0.6449	0.0021 (0.33%)	537
	0.3	80	0.00070	0.00014 (20%)	36	0.00068	0.00014 (20.59%)	37
		90	0.724	0.011 (1.52%)	17	0.733	0.010 (1.36%)	18
		100	4.513	0.034 (0.75%)	18	4.507	0.029 (0.64%)	25

Note: Option prices and standard errors are quoted in cents. Only significant digits are reported. Number of hidden nodes is 2. Other model parameters are $X_0 = 50$, $r = 0.05$ and $L = 40$.

of both rare events is crucial to increase the variance ratio. As demonstrated in Fig. 2, neural networks exhibit the capability to learn and navigate this balancing act. Table 2 shows again that the neural-network-induced change of measure is able to reduce the variance to varying degrees. We note that compared to the example of Sect. 5.2, adding stratification does not yield such a dramatic increase in variance ratio, but the improvement is still notable in most cases.

The model in this subsection has also been explored in Glasserman et al. [21], where [21, Table 5.2] reports variance ratios for different values of σ, K and the

knock-out barrier U (setting the lower knock-out barrier L to zero). In contrast to our model, the knock-out occurs if the asset price breaches the knock-out barrier U at the terminal time, i.e., if $X_1 > U$. We replicated their model using our methodology and compared the achieved variance ratios. Our method on average achieved 20% higher variance ratios for the case of importance sampling without stratification. However, when incorporating stratified sampling, our method on average achieved variance ratios that were 10% lower compared to those reported in [21, Table 5.2]. Note that the setting of [21] is in discrete time, and that the authors consider asymptotically optimal drift adjustments. These findings suggest that there might be ample scope to further investigate optimal neural-network-induced stratification for continuous-time models.

Let us now revisit the method proposed by Capriotti [7]. In [7, Sect. 5], there is an example of a European straddle with payoff $F(X) = (X_1 - K)^+ + (K - X_1)^+$. Capriotti [7] argues that in this case, the optimal sampling density would need to be bi-modal, a property that cannot be effectively captured by a normal distribution. As we attempted to implement this example, it became evident that the neural network struggled to determine the appropriate drift direction. This particular instance highlights the challenges associated with relying solely on drift adjustments for variance reduction. It serves as an example where the rare event can be characterised as the union of two events, shedding light on the limitations of such an approach.

5.4 Dynamic correlation

The generality of our paper builds on the decomposition $[M] = \int \pi(s)\mu(\mathrm{d}s)$. While Sect. 5.2 deviates from the conventional Brownian setting where $\mu = \lambda$, we also present an example that diverges from the typical scenario examined in the existing literature where $\pi \equiv \mathrm{id}$, representing the identity matrix. To this end, we consider a Heston model with a dynamic variance–covariance matrix.

We assume that the price process X follows the dynamics given by the SDE $\mathrm{d}X_t = rX_t\,\mathrm{d}t + \sqrt{V_t}X_t\,\mathrm{d}B_t$. The instantaneous variance V follows CIR-type dynamics described by $\mathrm{d}V_t = \kappa(\theta - V_t)\,\mathrm{d}t + \xi\sqrt{V_t}\,\mathrm{d}W_t$. Here, B and W are correlated Brownian motions related through $\mathrm{d}[B, W]_t = \rho(t)\,\mathrm{d}t$, where the correlation function takes the form $\rho(t) = \bar{\rho} + \tilde{\rho}A\sin(2\pi ft)$. In other words, we deviate from the constant correlations regime by means of the multiple of a sine wave with amplitude A and frequency f. We present the results for various combinations of amplitude and frequency choices in Table 3.

5.5 Basket option

So far, we have presented results in scenarios with low dimensions. However, the multidimensional formulation of our setting suggests investigating whether we can achieve satisfactory levels of variance reduction for higher-dimensional models. Inspired by Jourdain and Lelong [29], we study a multidimensional Black–Scholes model.

Consider the d-dimensional asset price X governed by the SDE

$$\mathrm{d}X_t = r \odot X_t\,\mathrm{d}t + X_t \odot \mathrm{d}M_t,$$

Table 3 Variance ratios for different values of amplitude A and frequency f

Parameters		Importance sampling (IS)			IS and stratification		
A	f	Mean	Std. err.	Var. ratio	Mean	Std. err.	Var. ratio
0	0	2.2145	0.0085 (0.38%)	171	2.2308	0.0063 (0.28%)	311
0.2	1	1.9378	0.0076 (0.39%)	178	1.9517	0.0058 (0.30%)	304
	2	2.0807	0.0082 (0.39%)	171	2.0938	0.0062 (0.30%)	297
	4	2.1498	0.0085 (0.40%)	165	2.1651	0.0065 (0.30%)	283
0.5	1	1.5544	0.0062 (0.40%)	203	1.5666	0.0048 (0.31%)	338
	2	1.8926	0.0077 (0.41%)	173	1.9037	0.0059 (0.31%)	289
	4	2.0553	0.0081 (0.38%)	168	2.0691	0.0062 (0.30%)	289
1	1	1.0147	0.0041 (0.39%)	268	1.0239	0.0032 (0.31%)	443
	2	1.6117	0.0064 (0.40%)	202	1.6230	0.0047 (0.29%)	369
	4	1.9048	0.0078 (0.41%)	165	1.9160	0.0060 (0.31%)	277

Note: Option prices and standard errors are quoted in cents. Only significant digits are reported. Number of hidden nodes is 5. Other model parameters are $X_0 = 50$, $r = 0.05$, $V_0 = 0.04$, $\kappa = 2$, $\theta = 0.09$, $\xi = 0.2$, $\bar{\rho} = -0.5$ and $K = 70$.

where $M = \Sigma B$ represents a d-dimensional standard Brownian motion B with variance–covariance matrix $\Sigma\Sigma^\top$. We sample the initial value X_0 of X uniformly from the range of 10 to 200. Moreover, we sample the vector r of appreciation rates and the vector σ of volatilities uniformly between 1% and 9% as well as 10% and 30%, respectively. The weight vector w is then computed as $w_i = r_i/\sigma_i^2$ and further normalised to sum to 1.

To define the matrix $\Sigma\Sigma^\top$, it is necessary to specify the correlation matrix. In order to ensure a valid correlation matrix that remains positive definite even in high dimensions, we adopt the approach proposed by Davies and Higham [14]. First, we sample a d-dimensional vector y uniformly between 0 and 1. We then rescale the vector y so that the sum of its elements equals the dimension d. The algorithm proposed in [14] then generates a valid correlation matrix, whose eigenvalues correspond to the values in the rescaled vector y. Finally, we still need to specify the strike. To this end, we sample 10^4 observations of the arithmetic average $\bar{X}$ at maturity and then choose the strike K to approximately be above the 90th percentile of the distribution of $\bar{X}_1$. Note that the choice of K is highly dependent on the previously sampled parameters.

Table 4 presents variance ratios obtained for various dimensions d ranging from $d = 10$ up to $d = 1000$. Moreover, we also compared our method to the approach presented in [29]. In their study, the authors considered the 40-dimensional case, and all volatilities, appreciation rates and weights were chosen uniformly across all assets in the basket. We refer to [29, Sect. 3] for further details about the model as well as the model parameters in [29, Table 1]. As it turns out, our method achieves variance ratios that are on average comparable to those reported by [29]. It is important to note that the strikes which were chosen are relatively close to the initial value. In previous examples, we can observe that the obtained variance ratios tend to grow as the strike is increased. In contrast to [29], we present in Table 4 results for dimension up to $d = 1000$, which we believe is a distinctive aspect worth highlighting.

Table 4 Variance ratios for different dimensions d

Parameters		Importance sampling				
d	K	Mean	Std. err.	Var. ratio	$\mathbb{P}[F(X) > 0]$	$\mathbb{Q}[F(X) > 0]$
10	88	3.7557	0.0118 (0.31%)	62	3.24%	70.83%
20	115	1.3252	0.0049 (0.37%)	124	1.22%	65.42%
50	126	6.4457	0.0189 (2.93%)	28	6.96%	72.36%
100	106	1.6995	0.0056 (0.33%)	54	3.36%	70.69%
200	112	2.1373	0.0067 (0.31%)	36	5.17%	72.04%
500	110	1.1352	0.0042 (0.37%)	30	4.46%	74.99%
1000	110	2.327	0.011 (0.47%)	6	10.87%	84.72%

Note: Option prices and standard errors are quoted in cents. Only significant digits are reported. Number of hidden nodes corresponds to the dimension d. Also $\mathbb{P}[F(X) > 0]$ represents the proportion of trajectories in the test dataset where the payoff is positive, without incorporating a drift adjustment, while $\mathbb{Q}[F(X) > 0]$ denotes the proportion of trajectories in the test dataset where the payoff is positive under the drift adjustment.

6 Conclusions

In this paper, we presented a method that uses feedforward neural networks for the purpose of reducing the variance of Monte Carlo estimators. To this end, we studied the class of Gaussian measures which are induced by vector-valued continuous local martingales with deterministic covariation. Building on the theory of vector stochastic calculus, we identified the Cameron–Martin spaces of those measures and proved universal approximation theorems that establish, up to an isometry, topological density of feedforward neural networks in these spaces. We then applied our results to a classical importance sampling approach which seeks an optimal drift adjustment of the processes which are driving the asset prices. Finally, we presented the results of a numerical study which clearly indicate the potential of this approach.

We remark that our approach comes with several challenges. In principle, one needs to train separate feedforward networks for different models and model parameters. In view of Remark 4.9, one could train a feedforward network to minimise a weighted average standard error over several models or model parameters. Complex, high-dimensional models might call for the use of complex neural network architectures in order to achieve a sufficient variance reduction, which might lead to a considerable computational effort for training the feedforward networks. On the other hand, the competing approaches in Guasoni and Robertson [24] and Robertson [50] involve having to solve a potentially complex, high-dimensional variational problem, whose solution might involve a numerical procedure which might induce a considerable computational effort, too. Finally, while Theorem 4.6 and the simulations of Sect. 5 show that one can obtain a sufficient variance reduction with shallow feedforward networks, the model-dependent choice of optimal architecture has not been discussed at all, which highlights the potential for a further improvement of this method.

6.1 Outlook on further research

Throughout this paper, we assumed the process M to be a continuous local martingale with deterministic covariation, so that it is a Gaussian process and induces a Gaussian measure on path space. Clearly, there are Gaussian processes which cannot be local martingales, e.g. fractional Brownian motion with Hurst index $\neq 1/2$. In line with Remark 2.15, Sect. 2 can be extended to the study of multivariate Volterra-type Gaussian processes of the form $\tilde{M}_t = \int_0^T k(t,s)\,\mathrm{d}M_s$ with a matrix-valued kernel k. While Sect. 4 makes use of the semimartingale property of M by applying Girsanov's theorem and studying convergence of stochastic exponentials, the Cameron–Martin theorem (see Theorem A.5) can still be applied to the Gaussian measure that is induced by $\tilde{M}$ on path space. These considerations in particular motivate the study of a refined class of multivariate (fractional) stochastic volatility models, their small-time asymptotics as well as importance sampling methods for the numerical valuation of derivatives for these models, which is subject to a follow-up work.

In Sect. 4, we required $F(X)$ to be $\mathcal{F}_T$-measurable and L^p-integrable for some $p > 2$. However, the properties that we imposed on the process X were rather weak. In particular, Theorem 4.6 only considered $F(X)$ as a random variable, where we used the SDE for X only when performing a measure change and applying Girsanov's theorem in order to understand the semimartingale decomposition of X under a new sampling measure. Therefore, the methods from Sect. 4 should extend to the case where X is the solution to a McKean–Vlasov SDE, provided that we understand how the dynamics of the process change under a change of measure. We leave it to a follow-up work to combine our methods with ideas from dos Reis et al. [48], which should lead to a tractable importance sampling framework for the valuation of derivatives on solutions to McKean–Vlasov SDEs under weaker assumptions than those imposed in [48].

The setting of this paper naturally applies to the valuation of European options and asset price processes with continuous paths. More generally, reducing the standard error of Monte Carlo estimators with neural networks when pricing American options based on the popular algorithm proposed by Longstaff and Schwartz (cf. Clément et al. [11] and Longstaff and Schwartz [40]) and models with jumps, very much in the spirit of Genin and Tankov [19] as well as Kawai [31], provides another interesting challenge that is reserved for follow-up work.

Finally, the measure changes we studied in Sect. 4 were induced by density processes of the form $\mathcal{E}(f \bullet M)$, where $f \in \Lambda^2$ is a deterministic function. The reason why we did not consider the more general class of processes $U \in L^2(M)$ for which $\mathcal{E}(U \bullet M)$ is a martingale is twofold. While the proof of Theorem 4.6 would become more involved, one would need to use neural network architectures which are more complex than the ones discussed in Sect. 3. For this reason, we argue that the problem of considering deterministic functions $f \in \Lambda^2$ provides a tractable, numerically efficient method to reduce the variance in Monte Carlo simulations, and leave the extension to processes $U \in L^2(M)$ and their approximation with neural networks for future work.

Appendix A: Gaussian measures

In this appendix, we collect for the readers' convenience some classical definitions and results about Gaussian measures. Let $(E, \|\cdot\|_E)$ denote a real separable Banach space, γ a Borel probability measure on E and $M = (M_t)_{t\in[0,T]}$ an $\mathbb{R}^d$-valued process on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Given $h \in E$, we further denote by γ_h the measure on E induced by the translation $E \ni x \mapsto x + h$.

Definition A.1 The *measure* γ is called *Gaussian* if each $f \in E^*$ induces a Gaussian distribution on $(\mathbb{R}, \mathcal{B}_\mathbb{R})$. The *process* $M = (M_t)_{t\in[0,T]}$ is called *Gaussian* if $(M_{t_i})_{i=1}^n$ is jointly Gaussian for each $n \in \mathbb{N}$ and $0 \le t_1 < t_2 < \cdots < t_n \le T$.

A Gaussian measure γ is centered if each $f \in E^*$ induces a centered Gaussian distribution. Similarly, a Gaussian process $M = (M_t)_{t\in[0,T]}$ is centered if $(M_{t_i})_{i=1}^n$ is jointly centered Gaussian for each $n \in \mathbb{N}$ and $0 \le t_1 < t_2 < \cdots < t_n \le T$. Since Sect. 2 only considers centered Gaussian processes and measures, we restrict **from now on** to this special case.

In the context of Definition A.1, we have the natural embedding $j\colon E^* \to E^*_\gamma$, where E^*_γ denotes the reproducing kernel Hilbert space of γ, which is defined as the closure of E^* in $L^2(\gamma)$. We further define the covariance operator of γ by the map

$$R_\gamma\colon E^* \to (E^*)', \qquad f \mapsto \left(g \mapsto \int_E f(x)g(x)\gamma(\mathrm{d}x)\right),$$

and implicitly consider its extension to E^*_γ, i.e., $R_\gamma\colon E^*_\gamma \to (E^*)'$.

Given $f \in E^*_\gamma$, note that $R_\gamma(f)\colon E^* \to \mathbb{R}$ is a linear operator. If we endow E^* with the Mackey topology, then Bogachev [5, Lemma 3.2.1] shows that $R_\gamma(f)$ is continuous. Mackey's theorem (cf. [5, Theorem A 1.1]) yields the existence of $x_f \in E$ such that $R_\gamma(f)(g) = g(x_f)$ for each $g \in E^*$. We then also denote by R_γ the map $E^*_\gamma \ni f \mapsto x_f$.

Definition A.2 Given a centered Gaussian measure γ on E, the *Cameron–Martin space* $H(\gamma)$ of γ is defined as the range of R_γ in E, i.e., $H(\gamma) := R_\gamma(E^*_\gamma) \subseteq E$. We equip $H(\gamma)$ with the inner product

$$\langle h, k\rangle_{H(\gamma)} := \langle \hat{h}, \hat{k}\rangle_{L^2(\gamma)} = \int_E \hat{h}(x)\hat{k}(x)\gamma(\mathrm{d}x), \qquad h, k \in H(\gamma),$$

where $h = R_\gamma(\hat{h})$ and $k = R_\gamma(\hat{k})$ for some $\hat{h}, \hat{k} \in E^*_\gamma$.

The space $(H(\gamma), \langle\cdot,\cdot\rangle_{H(\gamma)})$ is a real separable Hilbert space that is continuously embedded into E (cf. [5, Proposition 2.4.6 and Theorem 3.2.7]). Moreover, [5, Theorem 2.4.7] shows that $H(\gamma)$ is of γ-measure zero whenever E^*_γ is infinite-dimensional.

Remark A.3 Given a centered Gaussian measure γ on E, the topological support of γ is defined as the smallest closed subset $S \subseteq E$ with $\gamma(E\setminus S) = 0$. It is given by $\overline{H(\gamma)}$,

where the closure is taken in E (cf. [5, Theorem 3.6.1]). We call γ *nondegenerate* if $\overline{H(\gamma)} = E$ or, equivalently, if $H(\gamma)$ is densely embedded into E. If $\overline{H(\gamma)}$ is a strict subspace of E, then we call γ *degenerate*.

Remark A.4 If γ and $\tilde{\gamma}$ are two centered Gaussian measures on E with $H(\gamma) = H(\tilde{\gamma})$ and $\|\cdot\|_{H(\gamma)} = \|\cdot\|_{H(\tilde{\gamma})}$, then γ and $\tilde{\gamma}$ coincide (cf. [5, Corollary 3.2.6]). Moreover, if E is continuously and linearly embedded into another real separable Banach space $\tilde{E}$ with embedding i and induced Gaussian measure $\nu = \gamma \circ i^{-1}$, then $\tilde{E} \supseteq H(\nu) = i(H(\gamma))$ (cf. [5, Lemma 3.2.2]).

The Cameron–Martin space has another useful characterisation, which is stated in the following theorem (cf. [5, Theorem 2.4.5] and [6, Theorem 1]).

Theorem A.5 *Given a centered Gaussian measure γ on E and $h \in E$, the measures γ and γ_h are equivalent precisely when $h \in H(\gamma)$, and singular otherwise. In particular,*

$$H(\gamma) = R_\gamma(E^*_\gamma) = \{h \in E : \gamma_h \approx \gamma\}.$$

Whenever γ is Gaussian, the measure γ_h is Gaussian for each $h \in E$ (cf. [5, Lemma 2.2.2]). Consequently, Theorem A.5 characterises a set of Gaussian measures which are equivalent to γ. The following theorem (cf. [5, Theorem 2.7.2]) is another central result, which in particular implies that γ_h and γ are singular whenever $h \in E \setminus H(\gamma)$.

Theorem A.6 *Any two Gaussian measures on E are either equivalent or mutually singular.*

In order to quantify the (exponential) decline of the probability of certain tail events, the following result is often useful (cf. [5, Corollary 4.9.3]).

Proposition A.7 *Let γ be a centered Gaussian measure on E and for $\varepsilon > 0$, denote by γ_ε the pushforward measure of γ under the map $E \ni f \mapsto \sqrt{\varepsilon} f$. Then $(\gamma_\varepsilon)_{\varepsilon>0}$ satisfies the large deviation principle with rate function $I_\gamma \colon E \to \overline{\mathbb{R}}_+$, where*

$$I_\gamma(f) = \begin{cases} \frac{1}{2}\|f\|^2_{H(\gamma)} & \text{for } f \in H(\gamma), \\ \infty, & \text{otherwise.} \end{cases}$$

In other words, for each $F \in \mathcal{B}_E$,

$$-\inf_{f \in F^\circ} I_\gamma(f) \le \liminf_{\varepsilon \searrow 0} \varepsilon \log \gamma_\varepsilon(F) \le \limsup_{\varepsilon \searrow 0} \varepsilon \log \gamma_\varepsilon(F) \le -\inf_{f \in \bar{F}} I_\gamma(f).$$

Before we finish this section, we state a result that allows us in many cases to obtain a tractable representation of $(H(\gamma), \langle \cdot, \cdot \rangle_{H(\gamma)})$ (cf. [5, Sect. 3.3]).

Theorem A.8 *Given a centered Gaussian measure γ on E, assume that there exist a Hilbert space $\tilde{H}$ and a continuous linear operator $J\colon \tilde{H} \to E$ such that R_γ admits the factorisation $R_\gamma = J \circ J^*$, where $J^*\colon E^* \to \tilde{H}^* \cong \tilde{H}$ denotes the adjoint of J. Then $H(\gamma)$ coincides with $J(\tilde{H})$. If J is moreover injective, then*

$$\langle f, g\rangle_{H(\gamma)} = \langle J^{-1}(f), J^{-1}(g)\rangle_{\tilde{H}}, \qquad f, g \in H(\gamma).$$

Appendix B: Proofs

Proof of Lemma 2.10 (a) The proof of Cherny and Shiryaev [8, Lemma 3.2] reveals that $\|\cdot\|_{\Lambda^2}$ satisfies the triangle inequality, which shows that Λ^2 is a real vector space. In order to see that $\langle\cdot,\cdot\rangle_{\Lambda^2}$ is an inner product on Λ^2, note that by construction, $\langle\cdot,\cdot\rangle_{\Lambda^2}$ is symmetric and linear in both arguments, and recall that π is positive semidefinite μ-almost everywhere, hence $f^\top \pi f \geq 0$ μ-almost everywhere and therefore $\int_0^T f^\top(s)\pi(s)f(s)\mu(\mathrm{d}s) \geq 0$ for each measurable $f\colon [0,T] \to \mathbb{R}^d$. If $\langle f, f\rangle_{\Lambda^2} = 0$ for some $f \in \Lambda^2$, then $f^\top \pi f = 0$ μ-almost everywhere, hence $f \sim 0$, which implies that $\langle\cdot,\cdot\rangle_{\Lambda^2}$ is positive definite.

Completeness of (Λ^2, ϱ_2), where ϱ_2 denotes the translation invariant metric induced by $\|\cdot\|_{\Lambda^2}$, follows from Jacod [28, Lemme 4.29], and separability can be argued by adapting the proof of [4, Theorem 19.2]. We conclude that $(\Lambda^2, \langle\cdot,\cdot\rangle_{\Lambda^2})$ is a real separable Hilbert space.

(b) This is a direct consequence of the Fréchet–Riesz representation theorem since we know by Lemma 2.10(a) that Λ^2 is a Hilbert space.

(c) $\Lambda^{2,0}$ is clearly a real vector space. Given $f \in \Lambda^{2,0}$ and $i, j \in \{1, 2, \ldots, d\}$, a version of the Kunita–Watanabe inequality for Lebesgue–Stieltjes integrals gives

$$\begin{aligned}\left(\left|\int_0^T f_i(s) f_j(s)\mu_{i,j}(\mathrm{d}s)\right|\right)^2 &\leq \left(\int_0^T |f_i(s) f_j(s)||\mu_{i,j}|(\mathrm{d}s)\right)^2 \\ &\leq \int_0^T f_i^2(s)\mu_{i,i}(\mathrm{d}s)\int_0^T f_j^2(s)\mu_{j,j}(\mathrm{d}s) < \infty, \quad \text{(B.1)}\end{aligned}$$

hence $\Lambda^{2,0} \subseteq \Lambda^2$, and $(\Lambda^{2,0}, \langle\cdot,\cdot\rangle_{\Lambda^2})$ is therefore an inner product space.

The fact that $\Lambda^{2,0}$ is dense in Λ^2 has been shown in [28, Lemme 4.29], which also implies the separability of $\Lambda^{2,0}$. From Lemma 2.10(a), we further know that $(\Lambda^2, \langle\cdot,\cdot\rangle_{\Lambda^2})$ is a Hilbert space and in particular complete.

(d) The continuity of the embedding follows from (B.1) and (B.4) below. The remaining assertion follows from a multivariate version of Kallenberg [30, Lemma 1.37].

(e) Let $(\tilde{\pi}, \tilde{\mu})$ be another pair that satisfies the representation (2.1) and take $f \in \Lambda^{2,0}$. Then $\mathrm{d}\mu_{i,j}/\mathrm{d}\mu = \pi_{i,j}$ as well as $\mathrm{d}\mu_{i,j}/\mathrm{d}\tilde{\mu} = \tilde{\pi}_{i,j}$ for $i, j \in \{1, 2, \ldots, d\}$.

Hence (B.1) gives

$$\int_0^T f^\top(s)\pi(s)f(s)\mu(\mathrm{d}s) = \sum_{i,j=1}^d \int_0^T f_i(s)f_j(s)\mu_{i,j}(\mathrm{d}s)$$
$$= \int_0^T f^\top(s)\tilde{\pi}(s)f(s)\tilde{\mu}(\mathrm{d}s),$$

which extends to all $f \in \Lambda^2$ using the density of $\Lambda^{2,0}$ in Λ^2, and we see that $\|\cdot\|_{\Lambda^2}$ does not depend on the specific choice of (π, μ) satisfying (2.1). Moreover, for measurable $f, g\colon [0,T] \to \mathbb{R}^d$, $(f-g)^\top\pi(f-g) = 0$ μ-almost everywhere holds precisely when $\|f-g\|_{\Lambda^2} = 0$ which is equivalent to $(f-g)^\top\tilde{\pi}(f-g) = 0$ $\tilde{\mu}$-almost everywhere. □

Proof of Proposition 2.18 (a) By a variant of the Cauchy–Schwarz inequality, for $h \in H$, $i \in \{1, 2, \ldots, d\}$ and $t \in [0,T]$, it holds that

$$\int_0^t \left|\sum_{j=1}^d \pi_{i,j}(s)f_{h,j}(s)\right|\mu(\mathrm{d}s) \le \sqrt{\mu_{i,i}([0,t])}\|f_h\|_{\Lambda^2} \le \sqrt{\mu([0,T])}\|h\|_H, \quad \text{(B.2)}$$

which shows that the integral in (2.2) is well defined.

(b) H is clearly a real vector space and $\langle\cdot,\cdot\rangle_H$ is symmetric and linear in both arguments. To show that $\langle\cdot,\cdot\rangle_H$ is positive definite, note that $\langle h,h\rangle_H = \|f_h\|^2_{\Lambda^2} \ge 0$ for each $h \in H$. If $h \in H$ satisfies $\langle h,h\rangle_H = 0$, then $f_h^\top\pi f_h = 0$ μ-almost everywhere, hence $f_h \sim 0$. An application of (B.2) shows that $(\pi f_h)_i = 0$ μ-almost everywhere for $i \in \{1,2,\ldots,d\}$, hence $h = 0$. Thus $(H, \langle\cdot,\cdot\rangle_H)$ is an inner product space.

We obtain a norm $\|\cdot\|_H$ on H by setting $\|h\|_H := \sqrt{\langle h,h\rangle_H}$ and thus also a metric ϱ_H on H by setting $\varrho_H(f,g) := \|f-g\|_H$. In order to see that (H,ϱ_H) is complete, let $(h_n)_{n\in\mathbb{N}}$ be a Cauchy sequence in H. Then $(f_{h_n})_{n\in\mathbb{N}}$ is a Cauchy sequence in Λ^2. From Lemma 2.10(a), we know that Λ^2 is complete. Consequently, there exists some $f \in \Lambda^2$ such that $f_{h_n} \to f$ in Λ^2. If we set $h = J(f)$, then $h \in H$ and $h_n \to h$ in H.

Finally, to see that is separable, note first that Λ^2 is separable by Lemma 2.10(a). But then H is separable as well, because a countable dense subset of H is given by $\{h \in H\colon f_h \in B\}$, where B is a countable dense subset of Λ^2.

(c) By construction, $J\colon \Lambda^2 \to H$ is a linear isometry. Since $\Lambda^{2,0}$ is a linear subspace of Λ^2 by Lemma 2.10(c), we see that $(H^0, \langle\cdot,\cdot\rangle_H)$ is an inner product subspace of H. If $(h_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in H^0, then $(f_{h_n})_{n\in\mathbb{N}}$ is a Cauchy sequence in $\Lambda^{2,0}$. By Lemma 2.10(c), there exists an $f \in \Lambda^2$ such that $f_{h_n} \to f$ as $n \to \infty$. Denoting $h = J(f) \in H$, it follows that $h_n \to h$ in H as $n \to \infty$.

(d) This is a direct consequence of the Fréchet–Riesz representation theorem since we know by part (b) that H is a Hilbert space.

Remark B.1 For the proof of Proposition 2.18(e), we use a multivariate version of the Riesz–Markov–Kakutani representation theorem: Every $F \in (C([0,T];\mathbb{R}^d))^*$ can

be identified with an $\mathbb{R}^d$-valued function $\nu = (\nu_1, \nu_2, \dots, \nu_d)^\top$ on $\mathcal{B}_{[0,T]}$, where every entry is a signed Borel measure of finite total variation, such that

$$F(f) = \sum_{j=1}^{d} \int_0^T f_j(s)\nu_j(\mathrm{d}s) =: \int_0^T f^\top(s)\nu(\mathrm{d}s), \qquad f \in C([0,T];\mathbb{R}^d).$$

For $f\colon [0,T] \to \mathbb{R}^{n\times d}$ with $(f_{i,\cdot})^\top \in C([0,T];\mathbb{R}^d)$ for $i \in \{1,2,\dots,n\}$, we write

$$\int_0^T f(s)\nu(\mathrm{d}s) = \left(\int_0^T f_{1,\cdot}(s)\nu(\mathrm{d}s), \dots, \int_0^T f_{n,\cdot}(s)\nu(\mathrm{d}s) \right)^\top.$$

For generalisations to infinite-dimensional domains and image spaces, see for instance Gowurin [23] and Singer [53].

(e) We argue in line with Lifshits [39, Example 4.4] and use Theorem A.8. First we note that every $h \in H$ is continuous and satisfies $h(0) = 0$; hence $H \subseteq E$. Let us consider J as a linear operator $J\colon \Lambda^2 \to E$, which is continuous due to (B.2). In the context of Remark B.1, E^* is given as a quotient space, where we identify those $\nu \in (C([0,T];\mathbb{R}^d))^*$ that annihilate E.

For $f \sim \sigma$ and $g \sim \nu$ in E^*,

$$\begin{aligned}
&R_\gamma(f)(g)\\
&= \int_E f(x)g(x)\gamma_M(\mathrm{d}x) = \mathbb{E}[f(M)g(M)] = \mathbb{E}\left[\int_0^T M_s^\top \sigma(\mathrm{d}s) \int_0^T M_s^\top \nu(\mathrm{d}s) \right]\\
&= \sum_{i,j=1}^{d} \int_0^T \int_0^T \mathbb{E}[M_s^i M_t^j]\sigma_i(\mathrm{d}s)\nu_j(\mathrm{d}t) = \sum_{i,j=1}^{d} \int_0^T \int_0^T [M]_{s\wedge t}^{i,j}\sigma_i(\mathrm{d}s)\nu_j(\mathrm{d}t)\\
&= \sum_{j=1}^{d} \int_0^T \sum_{i=1}^{d} \int_0^T [M]_{s\wedge t}^{i,j}\sigma_i(\mathrm{d}s)\nu_j(\mathrm{d}t) = \sum_{j=1}^{d} \int_0^T \left(\int_0^T [M]_{s\wedge t}\sigma(\mathrm{d}s) \right)_j \nu_j(\mathrm{d}t)\\
&= \int_0^T \left(\int_0^T [M]_{s\wedge t}\sigma(\mathrm{d}s) \right)^\top \nu(\mathrm{d}t) = \left(g, \int_0^T [M]_{s\wedge \cdot}\sigma(\mathrm{d}s) \right).
\end{aligned}$$

We can therefore identify $R_\gamma(f)$ with $\int_0^T [M]_{s\wedge\cdot}\sigma(\mathrm{d}s)$.

Next, let us find the adjoint J^* of J. Given $f \in \Lambda^2$ and $g \sim \nu$ in E^*, we have

$$\begin{aligned} g\big(J(f)\big) &= \sum_{i=1}^{d} \int_0^T J_i(f)(t)\nu_i(\mathrm{d}t) = \sum_{i=1}^{d} \int_0^T \int_0^t \pi_{i,\cdot}(s) f(s)\mu(\mathrm{d}s)\nu_i(\mathrm{d}t) \\ &= \sum_{i=1}^{d} \int_0^T \int_0^T \mathbb{1}_{[0,t]}(s)\pi_{i,\cdot}(s) f(s)\mu(\mathrm{d}s)\nu_i(\mathrm{d}t) \\ &= \sum_{i=1}^{d} \int_0^T \int_0^T \mathbb{1}_{[0,t]}(s)\nu_i(\mathrm{d}t)\pi_{i,\cdot}(s) f(s)\mu(\mathrm{d}s) \\ &= \sum_{i=1}^{d} \int_0^T \nu_i([s,T])\pi_{i,\cdot}(s) f(s)\mu(\mathrm{d}s) \\ &= \int_0^T \nu([s,T])^\top \pi(s) f(s)\mu(\mathrm{d}s) = \langle \nu([\cdot,T]), f\rangle_{\Lambda^2}. \end{aligned}$$

So $J^*\colon E^* \to (\Lambda^2)^* \cong \Lambda^2$ is given by $g \sim \nu \mapsto ([0,T] \ni s \mapsto \nu([s,T]))$.

Finally, the covariance operator admits for $g \sim \nu$ in E^* the factorisation

$$\begin{aligned} R_\gamma(g)_i(t) &= \int_0^T [M]^{i,\cdot}_{s\wedge t}\nu(\mathrm{d}s) = \sum_{j=1}^{d} \int_0^T [M]^{i,j}_{s\wedge t}\nu_j(\mathrm{d}s) \\ &= \sum_{j=1}^{d} \int_0^T \int_0^{s\wedge t} \pi_{i,j}(w)\mu(\mathrm{d}w)\nu_j(\mathrm{d}s) \\ &= \sum_{j=1}^{d} \int_0^T \pi_{i,j}(w) \int_0^T \mathbb{1}_{[0,s\wedge t]}(w)\nu_j(\mathrm{d}s)\mu(\mathrm{d}w) \\ &= \sum_{j=1}^{d} \int_0^t \pi_{i,j}(w)\nu_j([w,T])\mu(\mathrm{d}w) \\ &= \int_0^t \pi_{i,\cdot}(w)\nu([w,T])\mu(\mathrm{d}w) \\ &= \int_0^t \pi_{i,\cdot}(w)J^*(g)(w)\mu(\mathrm{d}w) = (J\circ J^*)(g)_i(t), \end{aligned}$$

where $i \in \{1,2,\ldots,d\}$ and $t \in [0,T]$.

Let us show that J is injective. Let $f_1, f_2 \in \Lambda^2$ be such that $J(f_1) = J(f_2)$, i.e., $\|J(f_1) - J(f_2)\|_\infty = 0$, which implies in particular for $g = f_1 - f_2$ that

$$\int_A \pi(s)g(s)\mu(\mathrm{d}s) = 0 \in \mathbb{R}^d \tag{B.3}$$

for all $A \in \mathcal{B}_{[0,T]}$ of the form $A = (s,t]$ for $s < t$ in $[0,T]$. Since the half-open intervals generate $\mathcal{B}_{[0,T]}$, Dynkin's theorem shows that (B.3) extends to all $A \in \mathcal{B}_{[0,T]}$.

We now show that $g \sim 0$, i.e., $g^\top \pi g = 0$ μ-almost everywhere. If this were not the case, then we would have, without loss of generality, $\mu(\{g^\top \pi g > 0\}) > 0$. We claim that $\{g^\top \pi g > 0\} \subseteq \{\pi g \neq 0\}$. To see this, pick $s \in [0,T]$ such that $g^\top(s)\pi(s)g(s) > 0$ and assume that $\pi(s)g(s) = 0$. In other words, for each $i \in \{1,2,\ldots,d\}$, we should have $\sum_{j=1}^d \pi_{i,j}(s)g_j(s) = 0$. But this cannot be the case, since we then should have

$$0 < g^\top(s)\pi(s)g(s) = \sum_{i=1}^d g_i(s)\left(\sum_{j=1}^d \pi_{i,j}(s)g_j(s)\right) = 0.$$

Now $\mu(\{g^\top \pi g > 0\}) > 0$ implies that $\mu(\{\pi g \neq 0\}) > 0$. Because we clearly have $\{\pi g \neq 0\} = \bigcup_{i=1}^d \{(\pi g)_i \neq 0\}$, there is $i \in \{1,2,\ldots,d\}$ with $\mu(\{(\pi g)_i \neq 0\}) > 0$. Without loss of generality, we may assume that $\mu(\{(\pi g)_i > 0\}) > 0$. Note that the set $A = \{(\pi g)_i > 0\}$ can be written as

$$A = \bigcup_{n\in\mathbb{N}} \left\{(\pi g)_i \geq \frac{1}{n}\right\} = \bigcup_{n\in\mathbb{N}} \big((\pi g)_i\big)^{-1}\left(\left[\frac{1}{n},\infty\right)\right),$$

where every $A_n := ((\pi g)_i)^{-1}([\frac{1}{n},\infty))$ is $\mathcal{B}_{[0,T]}$-measurable, and thus so is A. Now $\mu(A) > 0$ implies $\mu(A_n) > 0$ for some $n \in \mathbb{N}$, hence $n\int_{A_n}(\pi g)_i(s)\mu(\mathrm{d}s) \geq \mu(A_n)$, which yields a contradiction to (B.3). We may therefore conclude that $f_1 - f_2 = g \sim 0$ in Λ^2, which shows that the operator $J\colon \Lambda^2 \to E$ is injective. Theorem A.8 now implies that the Cameron–Martin space of γ_M is given by $J(\Lambda^2) = H$. □

Proof of Theorem 2.24 That $H(D)$ is a dense subset of H follows from the Standing Assumption 2.22 and the definition of the norm on H induced by the inner product $\langle\cdot,\cdot\rangle_H$. Being a dense subset of a separable metric space implies the remaining assertion of part (a).

If D is also a linear subspace of Λ^2, then $H(D)$ is clearly an inner product space, whose completion is H by part (a). We now follow a standard argument, a version of which can be found e.g. in Mercer [42, Proposition 1]. Since $H(D)$ is dense in H by part (a) and H is separable due to Proposition 2.18(b), there exists a countable subset of $H(D)$ that is also dense in H. Upon applying the Gram–Schmidt process to this subset, one obtains a countable set of orthonormal vectors in $H(D)$ whose linear span is dense in H.

We know from Proposition 2.18(e) that H is the Cameron–Martin space of γ_M. By standard theory for Gaussian measures, we know that the topological support of γ_M then coincides with $\overline{H}$, where the closure is taken in E (see Remark A.3). But since $H(D)$ is dense in H by part (a), and the canonical injection from H to E is continuous by [5, Proposition 2.4.6], we have $\overline{H} = \overline{H(D)}$, which yields part (c). □

Proof of Proposition 3.4 For the purpose of the proof, we denote by $\|\cdot\|_{\infty;[0,T]}$ either the supremum or the λ-essential supremum over $[0, T]$, depending on which of the two conditions in the statement of Proposition 3.4 holds.

The affine functions $\mathbb{R} \ni x \mapsto \alpha x+\eta$ with $\alpha, \eta \in \mathbb{R}$ are continuous over $[0, T]$ and therefore bounded. Since ψ is (locally λ-essentially) bounded, each $f \in \mathcal{NN}^d_{1,\infty}(\psi)$ is (λ-essentially) bounded; hence

$$\int_0^T f_i^2(s)\mu_{i,i}(\mathrm{d}s) \le \|f_i\|^2_{\infty;[0,T]} \int_0^T \pi_{i,i}(s)\mu(\mathrm{d}s) < \infty \tag{B.4}$$

for each $i \in \{1, 2, \ldots, d\}$. Note that in (B.4), we implicitly used the fact that μ is absolutely continuous with respect to λ in the case of Condition 3) from Theorem 3.3 since in this case each, $f \in \mathcal{NN}^d_{1,\infty}(\psi)$ is μ-essentially bounded. We conclude that $\mathcal{NN}^d_{1,\infty}(\psi)$ is a linear subspace of $\Lambda^{2,0}$.

For $f \in \Lambda^{2,0}$, let $\epsilon > 0$. For each $\eta \in \mathbb{R}^d$ and $s \in [0, T]$, we have the inequality $\eta^\top \pi(s)\eta \le |\eta|^2 \operatorname{tr}(\pi(s))$ (cf. Cherny and Shiryaev [8, Sect. 3]). By Lemma 2.10(d), $C([0, T]; \mathbb{R}^d)$ is dense in $\Lambda^{2,0}$; hence there exists some $f_\epsilon \in C([0, T]; \mathbb{R}^d)$ such that $\|f - f_\epsilon\|_{\Lambda^2} < \epsilon/2$. By Theorem 3.3, there exists some $g \in \mathcal{NN}^d_{1,\infty}(\psi)$ such that $\|f_\epsilon - g\|_{\infty;[0,T]} < \epsilon/(2\sqrt{\|\operatorname{tr}(\pi)\|_{L^1(\mu)}})$, hence

$$\begin{aligned}\|f - g\|_{\Lambda^2} &\le \|f - f_\epsilon\|_{\Lambda^2} + \|f_\epsilon - g\|_{\Lambda^2} \\ &< \epsilon/2 + \|f_\epsilon - g\|_{\infty;[0,T]}\sqrt{\|\operatorname{tr}(\pi)\|_{L^1(\mu)}} < \epsilon,\end{aligned}$$

which concludes our proof. □

Proof of Proposition 3.5 Since ψ is bounded, one can show precisely as in the proof of Proposition 3.4 that $\mathcal{NN}^d_{1,\infty}(\psi)$ is a linear subspace of $\Lambda^{2,0}$. If $\mathcal{NN}^d_{1,\infty}(\psi)$ were not dense in $\Lambda^{2,0}$, there would exist by the geometric version of the Hahn–Banach theorem a functional $F \in (\Lambda^{2,0})^*$ such that $F \not\equiv 0$ and $F(f) = 0$ for each $f \in \mathcal{NN}^d_{1,\infty}(\psi)$. Let N denote the subspace of all $G \in (\Lambda^2)^*$ that annihilate $\Lambda^{2,0}$, i.e., for which $G(f) = 0$ for each $f \in \Lambda^{2,0}$ holds. The space $(\Lambda^{2,0})^*$ can then be identified with the quotient space $(\Lambda^2)^*/N$.

From Lemma 2.10(b), we know that there exists a function $g \in \Lambda^2$ such that $F(f) = \int_0^T f^\top(s)\pi(s)g(s)\mu(\mathrm{d}s)$ for each $f \in \Lambda^{2,0}$. The linearity of the Lebesgue–Stieltjes integral gives

$$F(f) = \sum_{i=1}^d \int_0^T f_i(s) \sum_{j=1}^d \pi_{i,j}(s)g_j(s)\mu(\mathrm{d}s) = 0, \qquad f \in \mathcal{NN}^d_{1,\infty}(\psi), \tag{B.5}$$

and by a variant of the Cauchy–Schwarz inequality (cf. [8, Lemma 4.17]), for each $i \in \{1, 2, \ldots, d\}$ and $A \in \mathcal{B}_{[0,T]}$, it holds that

$$\int_A \left| \sum_{j=1}^d \pi_{i,j}(s)g_j(s) \right| \mu(\mathrm{d}s) \le \sqrt{\mu_{i,i}(A)}\|g\|_{\Lambda^2} \le \sqrt{\mu(A)}\|g\|_{\Lambda^2}.$$

Hence $\nu_i(A) := \int_A \sum_{j=1}^d \pi_{i,j}(s) g_j(s) \mu(\mathrm{d}s)$ defines a signed Borel measure on $[0, T]$ that is of finite total variation.

Exactly as in Cybenko [13] and Hornik [26], we arrive at the question whether there exists a signed Borel measure $\nu \neq 0$ on $[0, T]$ of finite total variation such that $\int_0^T \psi(\alpha x + \eta)\nu(\mathrm{d}x) = 0$ holds for all $\alpha, \eta \in \mathbb{R}$. As we know from [13, Lemma 1], this is not the case if ψ is bounded, measurable and sigmoidal (meaning that $\psi(t) \to 0$ as $t \to -\infty$ and $\psi(t) \to 1$ as $t \to \infty$), and [26, Theorem 5] then generalised this finding to show that this is not the case if ψ is bounded, measurable and nonconstant. In other words, (B.5) implies that $F \equiv 0$, which yields a contradiction. □

Proof of Proposition 4.4 Since each $f \in D$ is a linear combination of compositions of ψ and affine functions, both of which are continuously differentiable, if follows that f is continuously differentiable, hence of bounded variation, which shows that $H(D)$ is a subspace of H_{bv}.

By Proposition 3.4, there exists a sequence $(h_n)_{n\in\mathbb{N}}$ in $H(D)$ that converges to $g_{h,c}$ in H. From the proof of Theorem 2.24(c), we know that the canonical injection from H to $C_0([0, T]; \mathbb{R})$ is continuous, which implies that $(h_n)_{n\in\mathbb{N}}$ converges to $g_{h,c}$ in $C_0([0, T]; \mathbb{R})$. Since $\tilde{F}\colon C_0([0, T]; \mathbb{R}) \to \mathbb{R} \cup \{-\infty\}$ is assumed to be continuous, it follows that $\tilde{F}(h_n)$ converges to $\tilde{F}(g_{h,c})$. Moreover, since $\|\cdot\|_H\colon H \to \mathbb{R}_+$ is Lipschitz-continuous, we can conclude that $\tilde{F}_{h,c}(h_n)$ converges to $\tilde{F}_{h,c}(g_{h,c})$. □

The following result follows from standard arguments. Recall that for $h \in H$, we denote by f_h the function in Λ^2 with $h(t) = J(f_h(t)$ for $t \in [0, T]$; see (2.2).

Lemma B.2 *For all $h \in H$ and $p \in [1, \infty)$, we have $f_h \in L^p(M)$, which implies that $f_h \bullet M \in \mathcal{H}^p$. Moreover, if $(h_n)_{n\in\mathbb{N}}$ denotes a sequence that converges to h in H, then $f_{h_n} \bullet M \to f_h \bullet M$ in $\mathcal{H}^p$ for each $p \in [1, \infty)$.*

Proof Since f_h is deterministic, it is predictable when viewed as a stochastic process. Moreover, since

$$\|f_h\|_{L^p(M)} = \mathbb{E}\big[[f_h \bullet M]_T^{p/2}\big]^{1/p} = \mathbb{E}\big[\big((f_h^\top \pi f_h) \bullet C\big)_T^{p/2}\big]^{1/p} = \|h\|_H < \infty, \quad \text{(B.6)}$$

we have $f_h \in L^p(M)$, and an application of the Burkholder–Davis–Gundy (BDG) inequality implies that $f_h \bullet M \in \mathcal{H}^p$. Keeping in mind (B.6), an application of the BDG inequality then yields the existence of a positive constant c_p such that

$$\|(f_{h_n} - f_h) \bullet M\|_{\mathcal{H}^p} \leq c_p \|h_n - h\|_H,$$

where the right-hand side converges to zero as $n \to \infty$. □

Proof of Lemma 4.5 Given $h \in H$, let $(h_n)_{n\in\mathbb{N}}$ be a sequence that converges to h in H. Set $A_p\colon H \ni h \mapsto (\mathcal{E}(f_h \bullet M)^{-1})_T$ and let $Y_n = A_p(h_n)$ for $n \in \mathbb{N}$, and $Y = A_p(h)$. By Lemma B.2, we have $f_{h_n} \bullet M \to f_h \bullet M$ in $\mathcal{H}^1$, hence $(f_{h_n} \bullet M)_T \to (f_h \bullet M)_T$ in $L^1(\mathbb{P})$ and thus also in probability. By the reverse triangle inequality, we have

$\|h_n\|_H \to \|h\|_H$ as $n \to \infty$, and in particular, the sequence $(\|h_n\|_H)_{n\in\mathbb{N}}$ is bounded. In consequence, Y_n converges to Y in probability.

For each $n \in \mathbb{N}$ and $p \in [1, \infty)$,

$$\begin{aligned}\mathbb{E}[Y_n^p] &= \mathbb{E}\big[\exp\big(-p(f_{h_n} \bullet M)_T - p^2\|h_n\|_H^2/2\big)\big]\exp\big((p+p^2)\|h_n\|_H^2/2\big)\\ &= \mathbb{E}[Z_T^n]\exp\big((p+p^2)\|h_n\|_H^2/2\big),\end{aligned} \tag{B.7}$$

where $Z^n := \mathcal{E}(-p(f_{h_n} \bullet M))$ is by Novikov's criterion a martingale. Thus we have $\mathbb{E}[Z_T^n] = \mathbb{E}[Z_0^n] = 1$, and so the sequence $(Y_n)_{n\in\mathbb{N}}$ is bounded in $L^p(\mathbb{P})$. This shows that for any $p \in [1, \infty)$, the sequence $(|Y_n - Y|^p)_{n\in\mathbb{N}}$ is uniformly integrable and hence converges to 0 in $L^1(\mathbb{P})$, which further implies that Y_n converges to Y in $L^p(\mathbb{P})$. Finally, (B.7) shows that $\|A_p(h)\|_{L^p(\mathbb{P})} = \exp((1+p)\|h\|_H^2/2)$, hence

$$\limsup_{\|h\|_H\to\infty} \frac{\|A_p(h)\|_{L^p(\mathbb{P})}}{\|h\|_H} = \lim_{\|h\|_H\to\infty} \frac{\exp\big((1+p)\|h\|_H^2/2\big)}{\|h\|_H} = \infty,$$

which yields the remaining assertion. □

Proof of Theorem 4.6 As in the proof of Lemaire and Pagès [35, Proposition 4], Hölder's inequality yields for $p = (2+\varepsilon)/2$ and $q = (2+\varepsilon)/\varepsilon$ that

$$\begin{aligned}V(h) &= \mathbb{E}_{\mathbb{P}}\big[F^2(X)\exp\big(-(f_h \bullet M)_T + \|h\|_H^2/2\big)\big]\\ &\le \mathbb{E}_{\mathbb{P}}[|F(X)|^{2+\varepsilon}]^{1/p}\mathbb{E}_{\mathbb{P}}\big[\exp\big(-q(f_h \bullet M)_T + q\|h\|_H^2/2\big)\big]^{1/q}\\ &= \mathbb{E}_{\mathbb{P}}[|F(X)|^{2+\varepsilon}]^{1/p}\mathbb{E}_{\mathbb{P}}\big[\mathcal{E}\big(-q(f_h \bullet M)\big)_T\big]^{1/q}\exp\big((1+q)\|h\|_H^2/2\big)\\ &= \mathbb{E}_{\mathbb{P}}[|F(X)|^{2+\varepsilon}]^{1/p}\exp\big((1+q)\|h\|_H^2/2\big), \quad h \in H,\end{aligned}$$

where the last equality follows because $\mathcal{E}(-q(f_h \bullet M))$ is a martingale so that

$$\mathbb{E}_{\mathbb{P}}\big[\mathcal{E}\big(-q(f_h \bullet M)\big)_T\big] = \mathbb{E}_{\mathbb{P}}\big[\mathcal{E}\big(-q(f_h \bullet M)\big)_0\big] = 1.$$

This shows that V is $\mathbb{R}_+$-valued.

Next, let us show that V is convex. To this end, pick $\eta \in (0,1)$ and $g, h \in H$ such that $g \neq h$. By the triangle inequality and positive homogeneity, we have the inequality $\|\eta g + (1-\eta)h\|_H \le \eta\|g\|_H + (1-\eta)\|h\|_H$. By the linearity of the vector stochastic integral and the convexity of $\mathbb{R} \ni x \mapsto x^2$, we thus have

$$\begin{aligned}&-\Big(\big(\eta f_g + (1-\eta)f_h\big) \bullet M\Big)_T + \|\eta g + (1-\eta)h\|_H^2/2\\ &\le \eta\big(-(f_g \bullet M)_T + \|g\|_H^2/2\big) + (1-\eta)\big(-(f_h \bullet M)_T + \|h\|_H^2/2\big).\end{aligned}$$

Together with the convexity and monotonicity of $\mathbb{R} \ni x \mapsto \exp x$, this shows that V is convex.

Given $h \in H$, let $(h_n)_{n\in\mathbb{N}}$ be a sequence that converges to h in H. Due to Lemma 4.5, $Z^n := A_q(h_n)$ converges to $Z := A_q(h)$ in $L^q(\mathbb{P})$. Note that

$F^2(X) \in L^p(\mathbb{P})$ by assumption. By Riesz's representation theorem, the topological dual of $L^q(\mathbb{P})$ is isometrically isomorphic to $L^p(\mathbb{P})$, where the isomorphism is given by

$$L^p(\mathbb{P}) \ni g \mapsto \left(L^q(\mathbb{P}) \ni f \mapsto \int_\Omega g(\omega) f(\omega) \mathbb{P}(\mathrm{d}\omega) \right).$$

Hence the map $L^q(\mathbb{P}) \ni Y \mapsto \mathbb{E}[F^2(X)Y] \in \mathbb{R}$ is continuous, which yields

$$\lim_{n\to\infty} V(h_n) = \lim_{n\to\infty} \mathbb{E}[F^2(X)Z^n] = \mathbb{E}[F^2(X)Z] = V(h).$$

Since H is in particular a metric space, continuity of V is equivalent to sequential continuity, which shows (b).

In order to prove existence of a minimiser of V, we borrow some tools from convex optimisation. First, we show that V is proper, meaning that $\{h \in H : V(h) < \infty\} \neq \emptyset$ and $V(h) > -\infty$ for all $h \in H$. The latter condition is clearly satisfied as V is nonnegative. For $h \equiv 0$, we further have $V(h) = \mathbb{E}[F^2(X)] < \infty$, which implies the former condition (which also follows from part (a)). Moreover, since V is continuous as argued above, it is in particular lower semicontinuous.

Let us show that V is coercive, i.e., that $V(h) \to \infty$ as $\|h\|_H \to \infty$. Since we assume that $\mathbb{P}[F^2(X) > 0] > 0$, there exists a constant $\delta > 0$ such that $\mathbb{P}[F^2(X) \geq \delta] > 0$. An application of the reverse Hölder inequality along the lines of the proof of [35, Proposition 4] yields the inequality

$$V(h) \geq \delta \mathbb{P}[F^2(X) \geq \delta]^3 \exp\left(\|h\|_H^2/4\right), \quad h \in H,$$

which shows that V is coercive. Consequently, Zălinescu [56, Proposition 2.5.6] shows that $\arg\min_{h\in H} V(h)$ is a convex set. Moreover, because H as a Hilbert space is reflexive, [56, Theorem 2.5.1] shows that $\arg\min_{h\in H} V(h)$ is not empty, which shows (c).

For $\delta > 0$, choose $h_\delta \in H$ such that $V(h_\delta) < \min_{h\in H} V(h) + \delta$. By Theorem 2.24(a), there exists a sequence $(h_n)_{n\in\mathbb{N}}$ in $H(D)$ that converges to h_δ in H. By part (b), we then obtain $\lim_{n\to\infty} V(h_n) = V(h_\delta) < \min_{h\in H} V(h) + \delta$. A diagonalisation argument yields (d). □

Acknowledgements We should like to thank Martin Schweizer for valuable feedback on our manuscript.

Funding Open access funding provided by TU Wien (TUW).

Declarations

Competing Interests The authors declare no competing interests.

References

1. Andresen, A., Imkeller, P., Perkowski, N.: Large deviations for Hilbert-space-valued Wiener processes: a sequence space approach. In: Viens, F., et al. (eds.) Malliavin Calculus and Stochastic Analysis: A Festschrift in Honor of David Nualart. Springer Proc. Math. Stat., vol. 34, pp. 115–138. Springer, New York (2013)
2. Arouna, B.: Robbins–Monro algorithms and variance reduction in finance. J. Comput. Finance **7**(2), 35–62 (2003)
3. Baldi, P., Ben Arous, G., Kerkyacharian, G.: Large deviations and the Strassen theorem in Hölder norm. Stoch. Process. Appl. **42**, 171–180 (1992)
4. Billingsley, P.: Probability and Measure, Anniversary edn. Wiley, Hoboken (2012)
5. Bogachev, V.: Gaussian Measures. Am. Math. Soc., Providence (1998)
6. Cameron, R., Martin, W.: Transformations of Wiener integrals under translations. Ann. Math. **45**, 386–396 (1944)
7. Capriotti, L.: Least-squares importance sampling for Monte Carlo security pricing. Quant. Finance **8**, 485–497 (2008)
8. Cherny, A., Shiryaev, A.: Vector stochastic integrals and the fundamental theorem of asset pricing. Proc. Steklov Inst. Math. **237**, 12–56 (2002)
9. Ciesielski, Z.: On the isomorphisms of the spaces H_α and m. Bull. Acad. Pol. Sci., Sér. Sci. Math. Astron. Phys. **8**, 217–222 (1960)
10. Ciesielski, Z., Kerkyacharian, G., Roynette, B.: Quelques espaces fonctionnels associés à des processus gaussiens. Stud. Math. **107**, 171–204 (1993)
11. Clément, E., Lamberton, D., Protter, P.: An analysis of a least squares regression method for American option pricing. Finance Stoch. **6**, 449–471 (2002)
12. Cohen, S., Elliott, R.: Stochastic Calculus and Applications, 2nd edn. Springer, New York (2015)
13. Cybenko, G.: Approximation by superpositions of a sigmoidal function. Math. Control Signals Syst. **2**, 303–314 (1989)
14. Davies, P., Higham, N.: Numerically stable generation of correlation matrices and their factors. BIT **40**, 640–651 (2000)
15. Filipović, D.: Consistency Problems for Heath–Jarrow–Morton Interest Rate Models. Lecture Notes in Mathematics, vol. 1760. Springer, Berlin (2001)
16. Friz, P., Hairer, M.: A Course on Rough Paths, 2nd edn. Springer, Cham (2020)
17. Friz, P., Victoir, N.B.: Multidimensional Stochastic Processes as Rough Paths: Theory and Applications. Cambridge University Press, Cambridge (2010)
18. Funahashi, K.: On the approximate realization of continuous mappings by neural networks. Neural Netw. **2**, 183–192 (1989)
19. Genin, A., Tankov, P.: Optimal importance sampling for Lévy processes. Stoch. Process. Appl. **130**, 20–46 (2020)
20. Glasserman, P.: Monte Carlo Methods in Financial Engineering. Springer, New York (2004)
21. Glasserman, P., Heidelberger, P., Shahabuddin, P.: Asymptotically optimal importance sampling and stratification for pricing path-dependent options. Math. Finance **9**, 117–152 (1999)
22. Glasserman, P., Wang, Y.: Counterexamples in importance sampling for large deviations probabilities. Ann. Appl. Probab. **7**, 731–746 (1997)
23. Gowurin, M.: Über die Stieltjessche Integration abstrakter Funktionen. Fundam. Math. **27**, 254–265 (1936)
24. Guasoni, P., Robertson, S.: Optimal importance sampling with explicit formulas in continuous time. Finance Stoch. **12**, 1–19 (2008)
25. Gühring, I., Kutyniok, G., Petersen, P.: Error bounds for approximations with deep ReLU neural networks in $W^{s,p}$ norms. Anal. Appl. **18**, 803–859 (2020)
26. Hornik, K.: Approximation capabilities of multilayer feedforward networks. Neural Netw. **4**, 251–257 (1991)
27. Hornik, K., Stinchcombe, M., White, H.: Multilayer feedforward networks are universal approximators. Neural Netw. **2**, 359–366 (1989)
28. Jacod, J.: Calcul Stochastique et Problèmes de Martingales. Lecture Notes in Mathematics, vol. 714. Springer, Berlin (1979)
29. Jourdain, B., Lelong, J.: Robust adaptive importance sampling for normal random vectors. Ann. Appl. Probab. **19**, 1687–1718 (2009)
30. Kallenberg, O.: Foundations of Modern Probability, 3rd edn. Springer, Cham (2021)

31. Kawai, R.: Optimal importance sampling parameter search for Lévy processes via stochastic approximation. SIAM J. Numer. Anal. **47**, 293–307 (2009)
32. Kidger, P., Lyons, T.: Universal approximation with deep narrow networks. In: Proceedings of Thirty Third Conference on Learning Theory. Proceedings of Machine Learning Research, vol. 125, pp. 2306–2327 (2020)
33. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings. (2015). Available online at https://arxiv.org/abs/1412.6980
34. Kratsios, A.: The universal approximation property. Ann. Math. Artif. Intell. **89**, 435–469 (2021)
35. Lemaire, V., Pagès, G.: Unconstrained recursive importance sampling. Ann. Appl. Probab. **20**, 1029–1067 (2010)
36. Leshno, M., Lin, V., Pinkus, A., Schocken, S.: Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. Neural Netw. **6**, 861–867 (1993)
37. Li, L., Mendoza-Arriaga, R., Mo, Z., Mitchell, D.: Modelling electricity prices: a time change approach. Quant. Finance **16**, 1089–1109 (2016)
38. Liao, Y., Fang, S.C., Nuttle, H.: Relaxed conditions for radial-basis function networks to be universal approximators. Neural Netw. **16**, 1019–1028 (2003)
39. Lifshits, M.: Lectures on Gaussian Processes. Springer, Heidelberg (2012)
40. Longstaff, F., Schwartz, E.: Valuing American options by simulation: a simple least-squares approach. Rev. Financ. Stud. **14**, 113–147 (2001)
41. Lu, Z., Pu, H., Wang, F., Hu, Z., Wang, L.: The expressive power of neural networks: a view from the width. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, pp. 6232–6240. Curran Associates Inc., Red Hook (2017)
42. Mercer, R.: Dense G_δ's contain orthonormal bases. Proc. Am. Math. Soc. **97**, 449–452 (1986)
43. Mhaskar, H., Micchelli, C.: Approximation by superposition of sigmoidal and radial basis functions. Adv. Appl. Math. **13**, 350–373 (1992)
44. Müller, T., McWilliams, B., Rousselle, F., Gross, M., Novák, J.: Neural importance sampling. ACM Trans. Graph. **38**(145), 1–19 (2019)
45. Park, J., Sandberg, I.: Universal approximation using radial-basis-function networks. Neural Comput. **3**, 246–257 (1991)
46. Park, J., Sandberg, I.: Approximation and radial-basis-function networks. Neural Comput. **5**, 305–316 (1993)
47. Pinkus, A.: Approximation theory of the MLP model in neural networks. Acta Numer. **8**, 143–195 (1999)
48. dos Reis, G., Smith, G., Tankov, P.: Importance sampling for McKean–Vlasov SDEs. Appl. Math. Comput. **453**, 1–31 (2023)
49. Robbins, H., Monro, S.: A stochastic approximation method. Ann. Math. Stat. **22**, 400–407 (1951)
50. Robertson, S.: Sample path large deviations and optimal importance sampling for stochastic volatility models. Stoch. Process. Appl. **120**, 66–83 (2010)
51. Ronen, E., Ohad, S.: The power of depth for feedforward neural networks. In: 29th Annual Conference on Learning Theory. Proceedings of Machine Learning Research, vol. 49, pp. 907–940 (2016)
52. Salem, R.: On some singular monotonic functions which are strictly increasing. Trans. Am. Math. Soc. **53**, 427–439 (1943)
53. Singer, I.: Linear functionals on the space of continuous mappings of a compact Hausdorff space into a Banach space. Rev. Math. Pures Appl. **2**, 301–315 (1957)
54. Su, Y., Fu, M.: Importance sampling in derivative securities pricing. In: Jeffrey, A., et al. (eds.) Proceedings of the 2000 Winter Simulation Conference, vol. 1, pp. 587–596. Institute of Electrical and Electronics Engineers (2000)
55. Zhang, J., Walter, G., Miao, Y., Lee, W.N.W.: Wavelet neural networks for function learning. IEEE Trans. Signal Process. **43**, 1485–1497 (1995)
56. Zălinescu, C.: Convex Analysis in General Vector Spaces. World Scientific Publishing Co., Inc., River Edge (2002)