Article

Deep Arbitrage-Free Learning in a Generalized HJM Framework via Arbitrage-Regularization

Anastasis Kratsios [1,†,‡] and Cody Hyndman [1,‡*]

[1] ETH Zürich; eanastasis.kratsios@math.ethz.ch
[2] Concordia University; cody.hyndman@concordia.ca

* Correspondence: anastasis.kratsios@math.ethz.ch; Tel.: +41-44-632-3751

Received: February 28, 2020; Accepted: NA; Published: NA

Abstract: A regularization approach to model selection, within a generalized HJM framework, is introduced which learns the closest arbitrage-free model to a prespecified factor model. This optimization problem is represented as the limit of a one-parameter family of computationally tractable penalized model selection tasks. General theoretical results are derived and then specialized to affine term-structure models where new types of arbitrage-free machine learning models for the forward-rate curve are estimated numerically and compared to classical short-rate and the dynamic Nelson-Siegel factor models.

Keywords: Arbitrage-Regularization; Bond Pricing; Model Selection; Deep Learning; Dynamic PCA.

1. Introduction

The compatibility of penalized regularization with machine learning approaches allows for the successful treatment of various challenges in learning theory such as variable selection (see Tibshirani (1996)) and dimension reduction (see Zou et al. (2006)). The objective of many machine learning models used in mathematical finance is to predict asset prices by learning functions depending on stochastic inputs. In general there is no guarantee that these stochastic factor models are consistent with no-arbitrage conditions. This paper introduces a novel penalized regularization approach to address this modelling difficulty in a manner consistent with financial theory. The incorporation of an arbitrage-penalty term allows various machine learning methods to be directly and coherently integrated into mathematical finance applications. We focuse on regression-type model selection tasks, however the penalty developed here can also be applied to other types of machine learning algorithms with financial applications.

To motivate our approach we first consider informally, similar to (Björk 2009, Chapter 10), the following simple situation that will later be made more precise. Let $(\Omega, \mathscr{F}, \{\mathscr{F}_t\}_{t\geq 0}, \mathbb{P})$ be a filtered probability space satisfying the usual conditions. Let $\{X_t\}$ be an $\mathscr{F}_t$-adapted real-valued stochastic process with continuous paths representing the price of a financial asset. Let r be the constant risk-free interest rate and assume a fixed time interval $[0, T]$. Under the absence of arbitrage there exists a risk-neutral probability measure $\mathbb{Q}$ that is equivalent to the real-world probability measure $\mathbb{P}$. The price

at time $t \in [0, T]$ of a derivative security with continuous bounded payoff function $f(X_T)$ at time T is given by the risk-neutral pricing formula

$$\mathbb{E}_{\mathbb{Q}}\left[e^{-r(T-t)} f(X_T) \middle| \mathscr{F}_t\right]. \tag{1}$$

With $L_t = \frac{d\mathbb{Q}}{d\mathbb{P}}\big|_{\mathscr{F}_t}$ we may express Equation (1) under the real-world probability measure $\mathbb{P}$ as

$$\mathbb{E}_{\mathbb{P}}\left[e^{-r(T-t)} \frac{L_T}{L_t} f(X_T) \middle| \mathscr{F}_t\right]. \tag{2}$$

Equivalently, the price given by Equation (2) can be expressed under $\mathbb{P}$ by defining the state-price density process $Z_t = e^{-rt} L_t$. If $\mathbb{Q}$ is the minimal martingale measure of Schweizer (1995) then the transformation

$$f(X_t) \mapsto Z_t f(X_t) \tag{3}$$

can be interpreted as finding the closest process to $f(X_t)$ which is a (local) martingale under $\mathbb{P}$. The purpose of this paper is to find an analogue of the transformation (3) in this setting when X_t is described by a stochastic factor model, as is the case with most machine learning approaches to mathematical finance. For example, X_t may be described by a deep neural network with stochastic inputs.

The above ignores well known results regarding the uniqueness of $\mathbb{Q}$ (see Schweizer (1995)) and other important generalizations of the martingale approach to arbitrage theory. In particular, the more general setting for the fundamental theorem of asset pricing of Delbaen and Schachermayer (1998) implies that if "arbitrage", in the sense of no-free lunch with vanishing risk, exists then the transformation (3) is undefined. However, many machine learning approaches to mathematical finance may admit arbitrage so it is necessary to consider the general case. The arbitrage-regularization framework introduced in this paper integrates machine learning methodologies with the general martingale approach to arbitrage theory.

We consider a general framework for learning the most similar arbitrage-free factor model to a factor model within a prespecified class of alternative factor models. This search is optimized by minimizing a loss-function measuring the distance of the alternative model to the original factor model with the additional constraint that the market described by the alternative model is a local martingale under a reference probability measure.

The main theoretical results rely on asymptotics for the arbitrage-regularization penalty for selecting the optimal arbitrage-free model from a class of stochastic factor models. Relaxation of the asymptotic results necessary for practical implementation are presented. Throughout this paper, the bond market will serve as the primary example of our methods since no-arbitrage conditions for factor models are well understood, see Filipović (2001) and the references therein. Numerical results applying the arbitrage-regularization methodology are implemented using real data.

The remainder of this paper is organized as follows. Section 2 states the arbitrage-regularization problem and overviews relevant background on bond markets. Section 3 develops the arbitrage-penalty and establishes the main asymptotic optimality results. Non-asymptotic relaxations of these results are also considered and linked with transaction costs. Section 4 specializes the general results to bond markets and where a simplified expression for the arbitrage-penalty is obtained. Numerical implementations of the results are considered and the arbitrage-regularization methodology is used to generate new machine learning based models consistent with no free lunch with vanishing risk (NFLVR) and the results are

benchmarked against classical term-structure models. The article concludes with Section 6. Appendix A contains supplementary material primarily required for the proofs, such as functional Itô calculus and Γ-convergence results. Proofs of the main theorems of the paper are included in Appendix B.

2. The Arbitrage-Regularization Problem

For the remainder of the paper all stochastic processes described in this paper are defined on a common stochastic base $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbb{P})$. Let $\mathbb{P}^\star$ be probability measure equivalent to the reference probability measure $\mathbb{P}$ and r_t denotes the risk-free rate in effect at time $t \geq 0$. Assume that there exists an asset whose prices process, denoted by N_t, is a strictly-positive $\mathbb{P}^\star$-martingale which serve as numéraire. Unless otherwise specified all the processes in this paper will be described under the martingale-measure for N_t, denoted $\mathbb{P}_N$ is defined by

$$\frac{d\mathbb{P}_N^\star}{d\mathbb{P}^\star} = \exp\left(\int_0^T -r_s ds\right)\frac{N_T}{N_0}.$$

The choice of the numéraire can be used to encode or remove any trend from the the price processes being modelled. Price processes which are local martingales under $\mathbb{P}_N^\star$ or $\mathbb{P}^\star$ are usually only semi-martingales under the objective measure $\mathbb{P}$. Further details on numéraires can be found in Shreve (2004).

We consider a large financial market $\{X_t(u)\}_{u\in\mathcal{U}}$, indexed by a non-empty Borel subset $\mathcal{U} \subseteq \mathbb{R}^D$, were D is a positive integer. For example, $\{X_t(u)\}_{u\in\mathcal{U}}$ may be used to represent a bond market where, using the parameterization of Musiela (1993), $\mathcal{U} = [0,\infty)$ represents the collection of all possible times to maturity and $X_t(u)$ represents the time t price of a zero-coupon bond with maturity $T \triangleq u + t$. As observed in Musiela (1993), the choice of parameterizing with respect to time to maturity removes the dependence on time in $\mathcal{U}$.

For each $u \in \mathcal{U}$, the process $X_t(u)$ will be driven by a latent factor process. In the case of the bond market, this latent process will be the forward-rate curve. Write

$$X_t(u) \triangleq S_t\left(\phi_t^u, [[\phi^u]]_t; u\right) \tag{4}$$

where $\phi_t^u \triangleq \phi(t,\beta_t,u)$, $\{S_t(\cdot,\cdot;u)\}_{u\in\mathcal{U}}$ is a family of path-dependent functionals encoding the latent process into the asset price $X_t(u)$, ϕ_t^u is the factor model for the latent process, and β_t are the $\mathbb{R}^d$-valued stochastic factors driving the latent process. Following Fournie (2010), S_t will be allowed to depend on the local quadratic-variation of the factor process $\phi(t,\beta_t,u)$, denoted by $[[\phi^u]]_t$ and defined by

$$[\phi(\cdot,\beta_\cdot,u)]_t = \int_0^t [[\phi^u]]_s ds,$$

where $[\phi(\cdot,\beta_\cdot,u)]_t$ denotes the usual quadratic-variation of the factor process $[\phi(\cdot,\beta_\cdot,u)]_t$.

In the case of the bond market, S_t will be the map taking a forward-rate curve to the price of a zero-coupon bond with time to maturity u defined by

$$S_t\left(\phi_t^u, [[\phi^u]]_t; u\right) \triangleq \exp\left(\int_t^{u+t} \phi(t,\beta_t,v)dv\right). \tag{5}$$

In general, S_t will be allowed to depend on the path of $\phi(t,\beta_t,u)$. Thus S_t will be a path-dependent functional of regularity $\mathbb{C}_b^{1,2}$ in the sense of Fournie (2010) as discussed in Appendix A.2. However, as in

the bond market, if S_t depends only on the current value of $\phi(t, \beta_t, u)$ then the requirement that S_t be of class $\mathbb{C}_b^{1,2}$, in the sense of Fournie (2010), is equivalent to it being of regularity $C^{1,2}(I \times \mathbb{R}^d)$ in the classical sense; where $I \triangleq [0, \infty)$. Therefore, the classical Itô-calculus would apply to S_t.

Analogously to Björk and Christensen (1999), the factor model ϕ for the latent process will always be suitably integrable and suitably differentiable. Specifically, ϕ will belong to a Banach subspace $\mathcal{X}$ of $L^p_{\nu\otimes\mu}\left(I \times \mathbb{R}^d \times \mathcal{U}\right)$ which can be continuously embedded within the Fréchet space $C^{1,2,2}(I \times \mathbb{R}^d \times \mathcal{U})$; where ν is a Borel probability measure supported on I, μ is a Borel probability measure supported on $\mathbb{R}^d \times \mathcal{U}$, and both ν and μ are equivalent to the corresponding Lebesgue measures restricted to their supports. Here, $1 \leq p < \infty$ is kept fixed.

An example from the bond modelling literature is the Nelson-Siegel model (see Nelson and Siegel (1987) and Diebold and Rudebusch (2013)), which expresses the forward-rate curve as a function of its level, slope, and curvature through the factor model. The Nelson-Siegel family is part of a larger class of affine term-structure models, in which, at any given time, the forward-rate curve is described in terms of a set of market factors as

$$\varphi(t, \beta, u) \triangleq \varphi_0(u + t) + \sum_{i=1}^{d} \beta^i \varphi_i(u + t), \tag{6}$$

where d is a positive integer and $\varphi_i \in \mathcal{X}$ and φ_0 is a forward-rate curve typically calibrated to the data available at time $t = 0$. However, as shown in Filipović (2001), the Nelson-Siegel model is typically not arbitrage-free therefore we would like to learn the closest arbitrage-free factor model, driven by the same stochastic factors. Therefore, given hypothesis class $\mathscr{H} \subseteq \mathcal{X}$ of plausible alternative models, we optimize

$$\begin{aligned} &\operatorname{argmin}_{\phi\in\mathscr{H}} \ell\left(\varphi - \phi\right) \\ &\text{subject to: } S_t(\phi_t^u, [[\phi^u]]_t; u) \text{ is a } \mathbb{P}_N^\star\text{-local martingale for all } u \in \mathcal{U}; \end{aligned} \tag{7}$$

where $\mathscr{H}$ is required to contain the (naive) factor model φ and $\ell : \mathcal{X} \to [0, \infty)$ is continuous and coercive loss function. For example, ℓ may be taken to be the norm on $\mathcal{X}$. Geometrically, 7 describes a projection of φ onto the (possibly non-convex) subset of $\mathscr{H}$ of factor models making each $S_t(\phi_t^u, [[\phi^u]]_t; u)$ into a $\mathbb{P}_N^\star$-local martingale for every $u \in \mathcal{U}$. The requirement that $\mathscr{H}$ contains the (naive) factor model φ is for consistency, in order to ensure that for any arbitrage-free factor model φ the solution to problem (7) is itself.

In general, the problem described by problem (7) may be challenging to implement as projections onto non-convex sets are intractable. In analogy with regularization literature, such as Hastie et al. (2015), instead we consider the following relaxation of problem (7) which is more amenable to numerical implementation

$$\operatorname{argmin}_{\phi\in\mathscr{H}} \ell\left(\varphi - \phi\right) + \mathrm{AF}^\lambda(\phi); \tag{8}$$

where $\{\mathrm{AF}^\lambda\}_{2\leq\lambda<\infty}$ is a family of functions from $\mathscr{H}$ to $[0, \infty]$ taking value 0 if each $S_t(\phi_t^u, [[\phi^u]]_t; u)$ is a $\mathbb{P}_N^\star$-local martingale simultaneously for every value of u and λ is a meta-parameter determining the amount of emphasis placed on the penalizing factor models which fail to meet this requirement. Problem (8) is called the *arbitrage-regularization* problem. Before presenting the main results we first state necessary assumptions and notation.

Notation 2.1. The following notation will be maintained throughout this paper.

(i) $I \triangleq [0, \infty)$.

(ii) ν and μ are Borel probability measures on I and on $\mathbb{R}^d \times \mathcal{U}$ respectively, (with respect to their respective relative topologies),

(iii) $C^{1,2,2}(I \times \mathbb{R}^d \times \mathcal{U})$ is the space of functions from $I \times \mathbb{R}^d \times \mathcal{U}$ to $\mathbb{R}$ admitting one derivative in the first input and two derivatives in its other inputs; it will be topologized by the following family of semi-norms

$$p_{\gamma,\alpha,\delta,K}(\phi) \triangleq \sup_{|\gamma|\leq 1,|\alpha|\leq 2,|\delta|\leq 2} \sup_{(t,\beta,u)\in K} \left| \frac{\partial^{\gamma}}{\partial s^{\gamma}} \frac{\partial^{|\alpha|}}{\partial \beta_i^{\alpha_i} \dots \beta_j^{\alpha_j}} \frac{\partial^{|\delta|}}{\partial u_{\tilde{i}}^{\delta_l} \dots \partial u_{\tilde{j}}^{\delta_k}} \phi\,(t,\beta,u) \right|;$$

where K is a non-empty compact subset of $I \times \mathbb{R}^d \times \mathcal{U}$ and γ, α, and δ are (multi-)indices.

Assumption 2.2. The following assumptions will be maintained throughout this paper.

(i) β_t is an $\mathbb{R}^d$-valued diffusion process which is the unique strong solution to

$$\beta_t = \beta_0 + \int_0^t \mu(s,\beta_s)ds + \int_0^t \sigma(s,\beta_s)dW_s,$$

where $\beta_0 \in \mathbb{R}^d$, W_t is an $\mathbb{R}^d$-valued Brownian motion, the components $\mu^i : \mathbb{R}^{1+d} \to \mathbb{R}$ are globally Lipschitz, the components $\left(\sigma^{i,j} : \mathbb{R}^{1+d} \to \mathbb{R}^{d\times d}\right)_{i,j=1}^{d}$ are globally Lipschitz.

(ii) $\phi \in \mathcal{X}$, where $\mathcal{X}$ is a Banach subspace of $L^p_{\nu\otimes\mu}\left(I \times \mathbb{R}^d \times \mathcal{U}\right)$ which admits a continuous embedding into $C^{1,2,2}(I \times \mathbb{R}^d \times \mathcal{U})$. Furthermore, $\mathcal{X}$ will always be viewed as continuously embedded within this space.

(iii) For every $u \in \mathcal{U}$, $\{S_t(\cdot,\cdot;u)\}_{t\in[0,\infty)}$ is a non-anticipative functional in $\mathbb{C}_b^{1,2}$ verifying the following "predictable-dependence" condition of Fournie (2010):

$$S_t(x_t, x_t; u) = S_t(x_t, x_{t-}; u),$$

for all $t \in [0,\infty)$ and all $(x,v) \in D([0,t];\mathbb{R}^d) \times D([0,t];S_+^d)$, where S_+^d is the set of $d \times d$-dimensional positive semi-definite matrices with real-coefficients,

(iv) The hypothesis class $\mathscr{H} \subseteq \mathcal{X}$ is a non-empty and unbounded.

The central problem of the paper will be addressed in full generality before turning to applications in term-structure models, in the next Section.

3. Main Results

In this section, we show the asymptotic equivalence of problems (7) and (8) for general asset classes. This requires the construction of the penalty term AF^{λ} measuring how far a given factor model is from being a $\mathbb{P}_N^{\star}$-local martingale. The construction of AF^{λ} is made in two steps. First a drift condition ensuring that each $\{S_t(\phi_t^u, [[\phi^u]]_t; u)\}_{u\in\mathcal{U}}$ is simultaneously a $\mathbb{P}_N^{\star}$-local martingale, generalizing the drift condition of Heath et al. (1992), it provides an analogue to the consistency condition of Filipović and Teichmann (2004). Second, the drift condition is used to build the penalty term in (8). Subsequently, the optimizers of (8) will be used to asymptotically solve problem (7).

Proposition 3.1 (Drift Condition). The processes $S_t(\phi_t^u, [[\phi^u]]_t; u)$ are $\mathbb{P}_N^\star$-local-martingales, for each $u \in \mathcal{U}$ simultaneously, if and only if

$$\begin{aligned} -\mathcal{D}S_s(\phi_s^u, [[\phi^u]]_s; u) =& \Delta S_s(\phi_s^u, [[\phi^u]]_s; u) \left[\frac{\partial \phi}{\partial t}(s, \beta_s, u) + \sum_{i=1}^{d} \frac{\partial \phi}{\partial \beta^i}(s, \beta_s, u) \mu^i(s, \beta_s) \right. \\ & \left. + \frac{1}{2} \sum_{i,j=1}^{d} \left(\frac{\partial^2 \phi}{\partial \beta^i \partial \beta^j}(s, \beta_s, u) \right) \sigma^i(s, \beta_s) \sigma^j(s, \beta_s) \right] \\ & + \frac{1}{2} [\Delta^2 S_s(\phi_s^u, [[\phi^u]]_s; u)] \left(\sum_{i=1}^{d} \frac{\partial \phi}{\partial \beta^i}(s, \beta_s, u) \sigma^i(s, \beta_s) \right)^2. \end{aligned} \tag{9}$$

is satisfied $\mathbb{P}_N^\star$-a.s. for every $t \in [0, \infty)$ and every $u \in \mathcal{U}$

The drift condition in Proposition 3.1 implies that if ϕ is such that the difference of the left and right-hand sides of (9) is equal to 0, $\mathbb{P}_N^\star$-a.s. for all $u \in \mathcal{U}$ then $S_t(\phi_t^u, [[\phi^u]]_t; u)$ is a $\mathbb{P}_N^\star$-local martingale simultaneously for all $u \in \mathcal{U}$. Thus, $S_t(\phi_t^u, [[\phi^u]]_t; u)$ is simultaneously a $\mathbb{P}_N^\star$-local martingale for all $u \in \mathcal{U}$ if for every $u \in \mathcal{U}$ the $[0, \infty)$-valued process $\underline{\Lambda}_t^u(\phi)$ is equal to 0 $\mathbb{P}_N^\star$-a.s, where $\underline{\Lambda}_t^u(\phi)$ is defined using (9) by

$$\begin{aligned} \underline{\Lambda}_t^u(\phi) \triangleq \Big| & \mathcal{D}S_s(\phi_s^u, [[\phi^u]]_s; u) + \Delta S_s(\phi_s^u, [[\phi^u]]_s; u) \left[\frac{\partial \phi}{\partial t}(s, \beta_s, u) + \sum_{i=1}^{d} \frac{\partial \phi}{\partial \beta^i}(s, \beta_s, u) \mu^i(s, \beta_s) \right. \\ & \left. + \frac{1}{2} \sum_{i,j=1}^{d} \left(\frac{\partial^2 \phi}{\partial \beta^i \partial \beta^j}(s, \beta_s, u) \right) \sigma^i(s, \beta_s) \sigma^j(s, \beta_s) \right] \\ & + \frac{1}{2} [\Delta^2 S_s(\phi_s^u, [[\phi^u]]_s; u)] \left(\sum_{i=1}^{d} \frac{\partial \phi}{\partial \beta^i}(s, \beta_s, u) \sigma^i(s, \beta_s) \right)^2 \Big|, \end{aligned} \tag{10}$$

where $\phi_t^u - \phi(t, \beta_t, u)$.

The processes $\underline{\Lambda}_t^u(\phi)$ can be used to define the penalty in (8), by integrating over all values of t and u. However, in certain applications, such as with bond-markets, it is more convenient to instead build AF^λ using $\underline{\Lambda}_t^u(\phi)$.

Definition 3.2 (Arbitrage-Penalty). *Let $\{\Lambda_t^u(\phi)\}_{\phi \in \mathscr{H}; u \in \mathcal{U}}$ be a family of $\mathscr{F}_t$-adapted $[0, \infty)$-valued stochastic processes for which*

$$\Lambda_t^u(\phi)(\omega) = 0 \Leftrightarrow \underline{\Lambda}_t^u(\phi)(\omega) = 0, \tag{11}$$

holds for all $\phi \in \mathscr{H}$, $t \in I$, $u \in \mathcal{U}$, and $\mathbb{P}_N^\star$-almost every $\omega \in \Omega$. Then, for every $\lambda \geq 0$, the family $\{\mathrm{AF}^\lambda\}_{\lambda \geq 0}$ of functions

$$\begin{aligned} \mathrm{AF}^\lambda : \mathcal{X} &\to [0, \infty] \\ \phi &\mapsto \lambda \mathbb{E}_{\mathbb{P}} \left[\sqrt[\lambda]{\int_{(t,u) \in I \times \mathcal{U}} |\Lambda_t^u(\phi)|^\lambda d\mu(t, u)} \right]. \end{aligned} \tag{12}$$

is said to define an arbitrage-penalty.

Remark 3.3. Whenever $|\Lambda_t^u(\phi)|^\lambda$ fails to be integrable, we make the convention that $\mathbb{E}_{\mathbb{P}} \left[\sqrt[\lambda]{\int_{(t,u) \in I \times \mathcal{U}} |\Lambda_t^u(\phi)|^\lambda d\mu(t, u)} \right] = \infty$.

The convergence of (8) to (7) is demonstrated in the next theorem. The proof relies on the theory of Γ-convergence, which is useful for interchanging the limit and an arginf operations.

Assumption 3.4. Assume that

(i) For every $\phi \in \mathscr{H}$ and $\mathbb{P}_N^\star$-a.e. $\omega \in \Omega$, the function $(t,u) \mapsto \Lambda_t^u(\phi)(\omega)$ is continuous on $\mathscr{H}$,
(ii) $\left\{\phi \in \mathscr{H} : (\forall u \in \mathcal{U})\, S_t(\phi_t^u, [[\phi^u]]_t; u) \text{ is a } \mathbb{P}_N^\star\text{-local-martingale}\right\} \subseteq \mathscr{H}$ is closed and non-empty.

Note that both statements (i) and (ii) are with respect to the relative topology on $\mathscr{H}$.

Theorem 3.5. Under Assumption 3.4 the following hold:

(i) Equation (7) admits a minimizer on $\mathscr{H}$,
(ii) $\lim_{\lambda\uparrow\infty;\lambda\geq 2} \inf_{\phi\in\mathscr{H}} \ell(\varphi-\phi) + \mathrm{AF}^\lambda(\phi) = \min_{\phi\in\mathscr{H}} \ell(\varphi-\phi) + \iota_{\mathscr{H}}(\phi)$,
(iii) If for every $\lambda \geq 2$ AF^λ is lower-semi-continuous on $\mathscr{H}$ then

$$\lim_{\lambda\uparrow\infty;\lambda\geq 2} \operatorname*{argmin}_{\phi\in\mathscr{H}} \ell(\varphi-\phi) + \mathrm{AF}^\lambda(\phi) \in \operatorname*{argmin}_{\phi\in\mathscr{H}} \ell(\varphi-\phi) + \iota_{\mathscr{H}}(\phi), \tag{13}$$

where $\iota_{\mathscr{H}}$ is defined on $\mathscr{H}$ as

$$\iota_{\mathscr{H}}(\phi) \triangleq \begin{cases} 0 & \text{if } (\forall u \in \mathcal{U})\, S_t(\phi_t^u, [[\phi^u]]_t; u) \text{ is a } \mathbb{P}_N^\star\text{-local-martingale} \\ \infty & \text{otherwise.} \end{cases} \tag{14}$$

Theorem 3.5 provides a theoretical means of asymptotically computing the optimizer $\hat{\phi}$ of problem (7). In practice this limit cannot always be computed and only very large values of λ can be used. However, in reality trading does not occur in a friction-less market but every transaction placed at time t incurs a cost $0 < k_t$. Moreover, only a finite number of assets are traded.

Following Guasoni (2006), the liquidation value of a portfolio determined by finitely many assets in a market with frictions is defined as follows. First, recall an admissible strategy is an adapted, left-continuous of finite-variation process $\theta_t \in \mathbb{R}^n$ which is $\mathbb{P}$-a.s. bounded below. Following (Guasoni 2006, Equation 2.2) and (Guasoni 2006, Remark 2.4), this means that any portfolio determined by an admissible strategy θ_t defined on a finite sub-market $\left\{S_t(\phi_t^{u_i}, [\phi^{u_i}]_t; u_i)\right\}_{i=1}^n$ has liquidation value:

$$V(\theta_t) = \sum_{i=1}^n \int_0^t \theta_s^i dS_s(\phi_s^{u_i}, [[\phi^{u_i}]]_s; u_i) - \int_0^t k_s S_s(\phi_t^{u_i}, [[\phi^{u_i}]]_s; u_i) d|D\theta_s^i| - k_t S_s(\phi_t^{u_i}, [[\phi^{u_i}]]_t; u_i)|\theta_t^i|, \tag{15}$$

where ϕ denote the optimizer of 8 for a fixed value of $2 \leq \lambda < \infty$, $D\theta^i$ denotes the weak derivative of θ^i in the sense of measures, and $|D\theta^i|$ denotes its variation. The first term on the right-hand side of (15) is the capital gains from trading, second represents the cost incurred from various transaction costs, and the last term represents the cost of instantaneous liquidation at time t.

The next result guarantees that the market model ϕ is arbitrage-free, granted that k_t is large enough to cover the spread between $S_t(\phi_t^{u_i}, [\phi^{u_i}]_t; u_i)$ and $S_t(\hat{\phi}_t^{u_i}, [[\hat{\phi}^{u_i}]]_t; u_i)$. The following assumption quantifies the requirement that λ be taken to be sufficiently large.

Assumption 3.6. There exists some $0 < m < M$ and some $2 < \lambda^\star$ such that for every $0 \leq t$, positive integer n, and every $u_1, \ldots, u_n \in U$ the following holds:

(i) $\sup_{0\leq t} \max_{i=1,\ldots,n} \text{ess-sup} \left|S_t(\hat{\phi}_t^{u_i}, [[\hat{\phi}^{u_i}]]_t; u_i) - S_t(\phi_t^{u_i}, [[\phi^{u_i}]]_t; u_i)\right| < M$,

(ii) $m < \inf_{0\le t} \inf_{i=1,\dots,n} \text{ess-inf} \left| S_t(\phi_t^{u_i}, [[\phi^{u_i}]]_t; u_i) \right|$.

Proposition 3.7. If $mk_t \le M$ for all times $0 \le t$ then for any admissible strategy θ trading $S_t(\phi_t^{u_1}, [\phi^{u_1}]_t; u_1), \dots, S_t(\phi_t^{u_n}, [\phi^{u_n}]_t; u_n)$, then $\mathbb{P}(0 \le V_T(\theta))$ implies that $\mathbb{P}(V_T(\theta) = 0)$.

In the next section we apply Theorem 3.5 and the arbitrage-regularization (8) to the bond market.

4. Arbitrage-Regularization for Bond Pricing

As discussed in Diebold and Rudebusch (2013), affine term-structure models are commonly used in forward-rate curve modelling due to their tractability and the interpretability. In the formulation of Björk and Christensen (1999), as further developed in Filipović (2000); Filipović et al. (2010), affine term-structure models are characterized by (6) together with the additional requirement that its stochastic factor process β_t follows an affine diffusion. By Cuchiero (2011) this means that the dynamics of β_t are given by

$$\begin{aligned} \mu^i(t,\beta) &\triangleq \gamma^i + \sum_{j=1}^{d} \gamma_{i,j}\beta^i \\ \left[\sigma^i(t,\beta)\right]^T \sigma^j(t,\beta) &\triangleq \alpha^{i,j} + \sum_{k,j=1}^{d} \alpha_{k;i,j}\beta^k, \end{aligned} \tag{16}$$

where $\gamma^i, \gamma_{i,j}, \alpha^{i,j}, \alpha_{k;i,j} \in \mathbb{R}$ and $i, j = 1, \dots, d$.

Fix meta-parameters $p, \kappa \ge 1$. For the next result, all the factor models will be taken as belonging to the weighted Sobolev space $W_w^{p,k}(I \times \mathbb{R}^d \times \mathcal{U})$ with weight function

$$w(t,\beta,u) \triangleq Ce^{-|t|-\|\beta\|^\kappa - |u|^\kappa}, \tag{17}$$

where C is a unique constant ensuring that $1 \in W_w^{p,k}$ and its weighted integral is equal to 1. The space $W_w^{p,k}(I \times \mathbb{R}^d \times \mathcal{U})$ is defined of all $\nu \otimes \mu$-locally-integrable, k-times weakly differentiable functions $f : I \times \mathbb{R}^d \times \mathcal{U} \to \mathbb{R}$ equipped with the norm:

$$\|f\| \triangleq \int_0^\infty \int_{\beta\in\mathbb{R}^d} e^{-|u|^\kappa - \|\beta\|^\kappa} |f|^p (t,u,\beta) d\beta du + \sum_{|a|=k} \int_0^\infty \int_{\beta\in\mathbb{R}^d} e^{-|u|^\kappa - \|\beta\|^\kappa} |D^\alpha f|^p (t,u,\beta) d\beta du,$$

where $\alpha \triangleq (\alpha_1, \dots, \alpha_d)$ is a multi-index, $\alpha_i = 0, 1, \dots$, $|\alpha| = \sum_{i=1}^d \alpha_i$, and $D^\alpha f$ is the weak derivative of f of order α defined by

$$\int_0^\infty \int_{\beta\in\mathbb{R}^d} f(D^\alpha g) d\beta du = (-1)^{|\alpha|} \int_0^\infty \int_{\beta\in\mathbb{R}^d} (D^\alpha f) g d\beta du \qquad \left(\forall g \in C_0^\infty(I \times \mathbb{R}^d \times U)\right).$$

Here, $C_0^\infty(I \times \mathbb{R}^d \times U)$ is the space of all compactly-supported functions with infinitely many derivatives. Furthermore, k is required to satisfy

$$k \ge \frac{1 + d + D}{p} + 2. \tag{18}$$

Analytic tractability is ensured by requiring that the factor models considered for the arbitrage-regularization (8) belong to the class $\mathscr{H}$ defined by

$$\phi(t,\beta,u) = \phi_0(u+t) + \sum_{i=1}^{d} \beta^i \phi_i(u+t). \tag{19}$$

Under these conditions the following theorem characterizes the asymptotic behavior of (8) in λ as solving problem (7), given fixed meta-parameters $p, \kappa \geq 1$. Following Filipović (2001), it will be convenient to denote

$$\Phi^i(u) = \int_0^u \phi^i(s)ds. \tag{20}$$

We also require the regularity of the factors $\{\phi_i\}_{i=0}^n$. Therefore, it will further be assumed that, for each $i = 0, \ldots, n$ $\phi_i \in W_w^{p,k}$.

Theorem 4.1. Let φ be in $\mathscr{H}$ and fix $p, \kappa \geq 1$. Then

(i) For every $\lambda \geq 2$ there exists an element ϕ^λ in $\mathscr{H}$ minimizing

$$\int_0^\infty \int_{\beta\in\mathbb{R}^d} e^{-|u|^\kappa - \|\beta\|^\kappa} \left(\varphi(u,\beta) - \phi(u,\beta)\right)^p d\beta du + \frac{\lambda}{\Gamma(1+\frac{1}{\kappa})^{\frac{1}{\lambda}}} \sqrt[\lambda]{\int_0^\infty e^{-|u|^\kappa} \left|\Lambda^u(\phi)\right|^\lambda du},$$

where $\Lambda_t^u(\phi)$ is defined by

$$\begin{aligned} \Lambda_t^u(\phi) \triangleq & \left| c_0 - \frac{\partial \Phi^0}{\partial u}(u) + \sum_{i=1}^d \gamma^i \Phi^i(u) - \frac{1}{2} \sum_{i,j=1}^d \alpha^{i,j} \Phi^i(u)^\star \Phi^j(u) \right|^p \\ & + \sum_{k=1}^d \left| c_k - \frac{\partial \Phi^k}{\partial u}(u) + \sum_{i=1}^d \gamma^{k,i} \Phi^i(u) - \frac{1}{2} \sum_{i,j=1}^d \alpha_{k;i,j} \Phi^i(u) \Phi^j(u) \right|^p. \end{aligned} \tag{21}$$

(ii) The following inclusion holds

$$\lim_{\lambda\uparrow\infty;\lambda\geq 2} \phi^\lambda \in \operatorname*{argmin}_{\phi\in\mathscr{H}} \int_0^\infty \int_{\beta\in\mathbb{R}^d} e^{-|u|^\kappa - \|\beta\|^\kappa} \left(\varphi(u,\beta) - \phi(u,\beta)\right)^p d\beta du + \iota_{\mathscr{H}}(\phi), \tag{22}$$

where $\iota_{\mathscr{H}}$ is as in (14).

Next, the arbitrage-regularization of forward-rate curves will be considered using deep learning methods.

4.1. A Deep Learning Approach to Arbitrage-Regularization

The flexibility of feed-forward artificial neural networks (ANNs), as described in the universal approximation theorems of Hornik (1991); Kratsios (2019b), makes the collection of ANNs a well-suited class of alternative models for the arbitrage-regularization problem. In the context of this paper an ANN is any function from $\mathbb{R}^d$ to $\mathbb{R}^d$ of the form

$$W_{N+1} \circ \rho \bullet W_N \circ \ldots \rho \bullet W_1, \tag{23}$$

where $\{W_i\}_{i=1}^{N+1}$ are affine functions from $\mathbb{R}^{d_i}$ to $\mathbb{R}^{d_{i+1}}$ where $d_1 = d$ and $d_{N+1} = n+1$, ρ is a continuous activation-function, and $\bullet$ denotes component-wise composition. Fix integers $N > 1, h > 1$, and $n > 0$.

The set of all feed-forward neural networks with $d_i = h$ for $1 < i \leq N$, $d_{N+1} = n+1$, and fixed activation function ρ will be denoted by $\mathcal{NN}^{\rho}_{N,h,n+1}$.

In order to maintain analytic tractability, it will be required that the class of alternative models still be of affine type; thus, in our analysis, $\mathscr{H}$ will consist of all functions in $W^{p,k}_{w}(I \times \mathbb{R}^d \times \mathcal{U})$ of the form

$$\phi(t,\beta,u) = \underline{\beta}^{\star}\left(\rho \bullet W_N \circ \cdots \circ \rho \bullet W_1(u+t)\right), \tag{24}$$

where $\underline{\beta}_1 = 1$, $\underline{\beta}_{i+1} = \beta^i$ for all $i > 1$, and where $\beta^{\star}$ denotes the transpose of β.

It has been shown in Rahimi and Recht (2008), amongst others, that if a network is appropriately designed, then training only the final layer and suitably initializing the matrices $W_N, \ldots, W_1$ performs comparably well to networks with all the layers trained. This phenomenon has been observed in numerous numerical studies, such as Jaeger and Haas (2004), where the entries of the matrices $W_N, \ldots, W_1$ are chosen entirely randomly. This practice has also become fundamental to feasible implementations of recurrent neural network (RNN) theory and reservoir computing, as studied in Gelenbe (1989), where training speed becomes a key factor in determining the feasibility of the RNN and reservoir computing paradigms.

The hypothesis class of alternative factor models to be considered in the arbitrage-regularization problem effectively reduces from (24) to

$$\phi(t,\beta,u) = \underline{\beta}^{\star} f(u+t), \tag{25}$$

where $\beta \in \mathbb{R}^{n+1}$ and $f \in \mathcal{NN}^{\rho}_{N,h,n+1}$ is initialized through by

$$(\beta, f) \in \underset{\beta \in \mathbb{R}^{n+1}, f \in \mathcal{NN}^{\rho}_{N,h,n+1}}{\operatorname{argmin}} \sum_{j=1}^{J} \sum_{i} \left(\beta^T f(u_j) - \phi_i(u_j)\right)^p e^{-|u_j|^{\kappa}}, \tag{26}$$

and $\{u_j\}_{j=1}^{J}$ is a uniform random sample on a non-empty compact subset of $\mathcal{U}$; $J > 0$. Thus, the optimization problem (26) is random since it relies on randomly generated data points $\{u_j\}_{j=1}^{J}$. However, instead of initializing $\hat{f}$ in an ad-hoc random manner, the initialization (26) guarantees that the shapes generated by (25) are close to those produced by the naive factor model (19). In this case, a brief computation shows that $\Lambda^u_t(\phi)$ simplifies to

$$\begin{aligned} \Lambda^u(\beta) \triangleq & \left| \beta^0 f(0) - \beta^0 f(u) + \sum_{i=1}^{d} \gamma^i \beta^i F(u) - \frac{1}{2} \sum_{i,j=1}^{d} \alpha^{i,j} F(u)^{\star} (\beta^i)^{\star} \beta^j F(u) \right|^p \\ & + \sum_{k=1}^{d} \left| \beta^k f(0) - \beta^k f(u) + \sum_{i=1}^{d} \gamma^{k,i} \beta^i F(u) - \frac{1}{2} \sum_{i,j=1}^{d} \alpha_{k;i,j} F(u)^{\star} (\beta^i)^{\star} \beta^j F(u) \right|^p, \end{aligned} \tag{27}$$

where $F(u) \triangleq \int_0^u f(s)ds$ with the integration is defined component-wise and β^i denotes the i^{th} entry of the vector β.

4.2. Numerical Implementations

The performance of the arbitrage-regularization methodology will now be applied to two factor models of affine type and its performance will be evaluated numerically. The first factor model is the

commonly used dynamic Nelson-Siegel model of Diebold and Rudebusch (2013) and the second is a machine learning extension of the classical PCA approach to term-structure modeling. The performance of the arbitrage-regularization for each model will be benchmarked against both the original factor models and against the HJM-extension of the Vasiček model. The Vasiček model is a natural benchmark since, as shown in Björk and Christensen (1999), it is consistent with a low-dimensional factor model. Therefore, each of the factor models contains roughly the same number of driving factors which ensures that the comparisons are fair. Moreover, the numéraire process N_t will be taken to be the money-market and we take $\mathbb{P}^\star = \mathbb{P}$.

The data-set for this implementation consists of German bond data for 31 maturities with observations obtained on 1273 trading days from January 4^{th} 2010 to December 30^{th} 2014. As is common practice in machine learning, further details of our code and implementation can be found on Kratsios (2019a).

As described in (25)-(27), the solution to the arbitrage-regularization (8), will be numerically approximated using randomly initialized deep feed-forward neural networks. The initialization network f of (25) is selected to have fixed depth $N = 5$, fixed height $d = d_i = 10^2$ for all but the consider and last layers, and its weights are learned using the ADAM algorithm. The meta-parameters $p = 2$ and $\kappa = 1$ are chosen empirically, and the parameters of the Ornstein-Uhlenbeck process are estimated using the maximum-likelihood. Once the model parameters have been learned, and the factor model optimizing (8) has been learned, the day ahead predictions of the stochastic factors are obtained through Kalman filter estimates of the hidden parameters β_t for each of the factor models. In the case of the Vasiček model the unobservable short-rate parameter is also estimated using the Kalman filter (see Bain and Crisan (2009)). These day-ahead predictions are then fed into the factor model and used to compute the next-day bond prices. These predictions are then compared to the realized next-day bond prices.

4.2.1. Model 1: The Dynamic Nelson-Siegel Model (Practitioner Model)

The Nelson-Siegel family is a low-dimensional family of forward-rate curve models used by various central banks to produce forward-rate or yield curves. As discussed in Carmona (2014), Finland, Italy, and Spain are such examples with other countries such as Canada, Belgium, and France relying on a slight extension of this model. The Nelson-Siegel model's popularity is largely due to its interpretable factors and satisfactory empirical performance. It is defined by

$$\varphi(t,\beta,u) \triangleq \beta^1 + \beta^2 e^{-(u+t)\tau} + \beta^3 (u+t) e^{-(u+t)\tau}, \tag{28}$$

where, as discussed in Diebold and Rudebusch (2013), the first factor represents the long-term level of the forward-rate curve, the second represents its shape, the third represents its curvature, and τ is a shape parameter; typically kept fixed.

Since market conditions are continually changing, the Nelson-Siegel model is typically extended from a static model to a dynamic model by replacing the static choice of β with a three-dimensional Ornstein-Uhlenbeck process and fixing the shape parameter $\tau > 0$ as in Diebold and Rudebusch (2013). However, as demonstrated in Filipović (2001), the dynamic Nelson-Siegel model does not admit an equivalent measure to $\mathbb{P}^\star_N$, that makes the entire bond market simultaneously into local martingales. It was then shown in Christensen et al. (2011) that a specific additive perturbation of the Nelson-Siegel family circumvents this problem, but empirically this is observed to come at the cost of reduced predictive accuracy. In our implementation, the parameters of the Ornstein-Uhlenbeck process driving β^i_t will be estimated using the maximum likelihood method described in Meucci (2005).

4.2.2. Model 2: dPCA (Machine-Learning Model)

The dynamic Nelson-Siegel model's shape has been developed through practitioner experience. The second factor model considered here will be of a different type, with its factors learned algorithmically. Similarly to (28), consider a static three-factor model for the forward-rate curve of the form

$$\varphi(t, \beta, u) \triangleq \sum_{i=1}^{3} \beta^i \phi_i(u + t) \tag{29}$$

where each $\phi_1, \ldots, \phi_3$ are the first three principal components of the forward-rate curve calibrated on the first 100 days of data.

Subsequently, a time-series for the β^i parameters is generated, using the first 100 days of data, where on each day the β^i are optimizes according to the Elastic-Net (ENET) regression problem of Hastie et al. (2015) defined by

$$\beta_t^{ENET} = \underset{\beta \in \mathbb{R}^3}{\operatorname{argmin}} \sum_{t=t_{j-100}}^{t_j} \left(\sum_{i=1}^{3} \beta^i \phi_i(u + t) \right)^2 + \gamma \sum_{i=1}^{3} |\beta_i| + \gamma_2 \sum_{i=1}^{3} |\beta_i|^2, \tag{30}$$

on rolling windows consisting of 100 data points. The meta-parameters γ_1 and γ_2 are chosen by cross-validation on the first 100 training days and then fixed.

The ENET regression is used due to its factor selection abilities and computational efficiency. Next, analogously to the dynamic Nelson-Siegel model and $\mathbb{R}^d$-valued Ornstein-Uhlenbeck process $\hat{\beta}_t$ is calibrated, using the maximum likelihood methodology outlined in Meucci (2005) to the time-series $\{\beta_t^{ENET}\}$. These will provide the hidden stochastic factors in the dynamic PCA model (29). Thus, the dPCA model is the factor model with stochastic inputs defined by

$$\sum_{i=1}^{3} \hat{\beta}_t^i \phi_i(u + t). \tag{31}$$

The resulting model differs from the dynamic Nelson-Siegel model in that its factors and dynamics are not chosen by practitioner experience but learned through the data and implicitly encode some path-dependence. However, like the dynamic Nelson-Siegel model it falls within the scope of Theorem 4.1.

5. Discussion

The predictive performance of the Vasiček (Vasiček), dPCA, A-Reg(dPCA), the dynamic Nelson-Siegel Model (dNS), the arbitrage-free Nelson-Siegel model of Christensen et al. (2011) (AFNS), and the arbitrage-regularization of the dynamic Nelson-Siegel Model (A-Reg(NS)) is reported in the following tables. The predictive quality is quantified by the estimated mean-squared errors when making day-ahead predictions of the bond price for each maturity, for all but the first days in our data-set. The lowest estimated mean-squared error recorded are highlighted using bold font and the second lowest estimated mean-squared error on each maturity are emphasized using italics.

Model \ Maturity	0.5	1	2	3	4
Vasiček	3.155e-01	4.323e-01	3.622e-01	1.950e-01	**5.730e-02**
dPCA	*2.526e-01*	4.349e-01	4.176e-01	2.526e-01	9.261e-02
A-Reg(dPCA)	8.066e-01	6.943e-01	5.110e-01	2.755e-01	9.588e-02
NS	4.513e-02	*1.479e-01*	*2.134e-01*	*1.477e-01*	*5.968e-02*
AFNS	4.513e-02	*1.479e-01*	*2.134e-01*	*1.477e-01*	*5.968e-02*
A-Reg(NS)	**2.903e-02**	**9.514e-02**	**1.601e-01**	**1.235e-01**	6.482e-02
Model \ Maturity	5	6	7	8	9
Vasiček	**7.735e-03**	**1.996e-04**	1.024e-03	1.480e-03	1.348e-03
dPCA	2.193e-02	3.326e-03	**3.119e-04**	**1.897e-05**	**8.097e-07**
A-Reg(dPCA)	2.221e-02	3.340e-03	*3.123e-04*	*1.898e-05*	*8.099e-07*
NS	*1.972e-02*	*8.313e-03*	5.323e-03	3.925e-03	2.998e-03
AFNS	*1.972e-02*	*8.313e-03*	5.323e-03	3.925e-03	2.998e-03
A-Reg(NS)	3.579e-02	2.236e-02	1.523e-02	1.050e-02	7.308e-03

Table 1. (Short): MSE Comparisons for 1-day ahead bond-price predictions.

Table 1 evaluates the performance of the considered models on the short-mid end of the curve. Overall, the performance of all the models are generally comparable at the very short end but rapidly after the dPCA model begins to outperform the rest. The accuracy of the Vasiček model on small maturities is likely to it being a short-rate model.

In Table, 2 the dPCA model outperforms the rest by progressively larger margins. Most notably, in Tables 3, 4 which summarize the performance of the models for very long bong maturities the A-Reg(dPCA) model shows very low predictive error for a low number of factors while simultaneously being consistent with no-arbitrage conditions.

Model \ Maturity	10	11	12	13	14
Vasiček	1.108e-03	9.002e-04	7.382e-04	6.125e-04	5.135e-04
dPCA	**2.578e-08**	**6.328e-10**	**1.433e-11**	**2.607e-13**	**4.179e-15**
A-Reg(dPCA)	*2.579e-08*	*6.328e-10*	*1.433e-11*	*2.607e-13*	*4.179e-15*
NS	2.381e-03	1.969e-03	1.686e-03	1.484e-03	1.337e-03
AFNS	2.381e-03	1.969e-03	1.686e-03	1.484e-03	1.337e-03
A-Reg(NS)	5.215e-03	3.827e-03	2.885e-03	2.229e-03	1.761e-03
Model \ Maturity	15	16	17	18	19
Vasiček	4.342e-04	3.698e-04	3.169e-04	2.729e-04	2.360e-04
dPCA	**6.714e-17**	**9.566e-19**	**1.426e-20**	**1.819e-22**	**2.749e-24**
A-Reg(dPCA)	*6.714e-17*	*9.566e-19*	*1.426e-20*	*1.818e-22*	*2.746e-24*
NS	1.225e-03	1.138e-03	1.069e-03	1.012e-03	9.639e-04
AFNS	1.225e-03	1.138e-03	1.069e-03	1.012e-03	9.639e-04
A-Reg(NS)	1.422e-03	1.171e-03	9.831e-04	8.406e-04	7.316e-04

Table 2. (Mid): MSE Comparisons for 1-day ahead bond-price predictions.

Even though arbitrage-free regularization does slightly reduce its accuracy, which is natural since it adds a constraint into an otherwise purely predictive process, the arbitrage-regularized dPCA model is still much more accurate than the rest.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2020 doi:10.20944/preprints202003.0022.v1

Peer-reviewed version available at Risks 2020, 8, 40; doi:10.3390/risks8020040

Model \ Maturity	20	21	22	23	24
Vasiček	2.049e-04	1.784e-04	1.558e-04	1.364e-04	1.196e-04
dPCA	**3.816e-26**	**5.254e-28**	**8.047e-30**	**9.958e-32**	**1.336e-33**
A-Reg(dPCA)	*3.781e-26*	*4.847e-28*	*3.015e-30*	*2.684e-30*	*1.452e-29*
NS	9.228e-04	8.866e-04	8.542e-04	8.247e-04	7.976e-04
AFNS	9.228e-04	8.866e-04	8.542e-04	8.247e-04	7.976e-04
A-Reg(NS)	6.480e-04	5.838e-04	5.349e-04	4.984e-04	4.814e-04
Model \ Maturity	25	26	27	28	29
Vasiček	1.051e-04	9.254e-05	8.160e-05	7.205e-05	6.371e-05
dPCA	**2.067e-35**	**2.814e-37**	**3.639e-39**	**5.371e-41**	**7.459e-43**
A-Reg(dPCA)	*9.846e-29*	*1.102e-27*	*2.108e-26*	*6.986e-25*	*3.979e-23*
NS	7.722e-04	7.484e-04	7.257e-04	7.041e-04	6.835e-04
AFNS	7.722e-04	7.484e-04	7.257e-04	7.041e-04	6.835e-04
A-Reg(NS)	4.911e-04	5.288e-04	6.011e-04	7.214e-04	9.138e-04

Table 3. (Long): MSE Comparisons for 1-day ahead bond-price predictions.

A highlight of the A-Reg(dPCA) model is that it can accurately model the long-end of the forward-rate curve in an arbitrage-free manner. This is due to the dynamic factor selection properties of the dPCA model which otherwise could not have been used in a consistent manner if it were not for Theorem 4.1.

	30
Vasiček	6.371e-05
dPCA	**7.459e-43**
A-Reg(dPCA)	*3.979e-23*
NS	6.835e-04
AFNS	6.835e-04
A-Reg(NS)	9.138e-04

Table 4. (30 Year): MSE Comparisons for 1-day ahead bond-price predictions.

The numerical implementation highlight a few key facts about the arbitrage-regularization methodology. First, for nearly every maturity, the empirical performance of the arbitrage-regularization of a factor model is comparable to the original factor model. An analogous phenomenon was observed in Devin et al. (2010) when projecting infinite-dimensional arbitrage-free HJM models onto the finite-dimensional manifold of Nelson-Siegel curves. Therefore, correcting for arbitrage does not come at a significant predictive cost. However, it does come with the benefit of making the model theoretically sound and compatible with the techniques of arbitrage-pricing theory.

Second, since (8) incorporates an additional constraint into the modeling procedure the arbitrage-regularization of a factor model has a reduction in performance as compared to the initial factor model. This phenomenon, has also been observed empirically in Christensen et al. (2011) for the arbitrage-free Nelson-Siegel correction of the dynamic Nelson-Siegel model. Therefore, one should not expect to improve on the predictive performance of the initial factor model by correcting for the existence of arbitrage.

Third, the empirical performance of A-Reg(dPCA) was significantly better than the empirical performance of the other arbitrage-free models, namely AFNS, A-Reg(NS), and the Vasiček model, across nearly all maturities. This was especially true for mid and long maturity zero-coupon bonds. Moreover, the performance of A-Reg(dPCA) and dPCA were comparable. Similarly, for most maturities,

Peer-reviewed version available at Risks 2020, 8, 40; doi:10.3390/risks8020040

the empirical performance of the AFNS, dNS, and A-Reg(NS) models were all similar and notably lower than the performance of the A-Reg(dPCA), dPCA, and Vasiček models. This emphasizes the fact that arbitrage-regularization methodology produces performant models only if the original model itself produces accurate predictions. Therefore, it is up to the practitioner to make an appropriate choice of model. However, the methodology used to develop dPCA and A-Reg(dPCA) could be used as a generic starting point.

Since the arbitrage-regularization methodology applies to nearly any factor model, one may use any methodology to produce a accurate reference factor model and then apply arbitrage-regularization to make it theoretically consistent at a small cost in performance. This opens the possibility to applying machine learning models, such as dPCA, to finance without the worry that they are not arbitrage-free since their asymptotic arbitrage-regularization is well-defined. Furthermore, the flexibility of deep feed-forward neural networks allows for the efficient implementation of implement (8).

6. Conclusion

This paper introduced a novel model-selection problem and provided an asymptotic solution in the form of the penalized optimization given by problem (8). The problem was posed and solved in a generalized HJM-type setting, within Theorem 3.5 and specialized to the term-structure of interest setting in Theorem 4.1 where simple expressions for the penalty term were derived.

The key innovation of the paper was the construction of the penalty term AF^{λ} defining the arbitrage-regularization problem (8). The construction of this term in Proposition 3.1 relied on the structure of the generalized HJM-type setting proposed in Heath et al. (1992) and generalized in (4) which allowed one to encode the dynamics of any factor model with stochastic inputs into the specific structure of any asset class.

The numerical feasibility of the proposed method was made possible by the flexibility of feed-forward neural networks, as demonstrated in Hornik (1991); Kratsios (2019), which allowed the optimizer of the arbitrage-regularization problem (8) to be approximated to arbitrary precision. In the numerics section of this paper, it was found that the arbitrage-regularization of a factor model does not heavily impact its predictive performance but does make it approximately consistent with no-arbitrage requirements.

In particular, the compatibility of the proposed approach with generic factor models with stochastic inputs allowed for the consistent use of factor models generated from machine learning methods. The dPCA model is a novel example of such a model where the dynamics and factors were generated algorithmically instead of through practitioner experience. Applying arbitrage-regularization to dPCA drastically out-performance the classical benchmark models while being approximately consistent with no-arbitrage requirements.

The precise quantification of approximately arbitrage-free were made in Proposition 3.7. Thus approximately arbitrage-free factor models under the stylized assumption of no transaction costs were indeed arbitrage-free when proportional transaction costs are in place, which is a more realistic assumption.

Finally, the arbitrage-regularization approach introduced in this paper opens the door to the compatible use of predictive machine-learning factor models with the no-free lunch with vanishing risk condition. The general treatment in Theorem 3.5 can be transferred to other asset classes and models generated from other learning algorithms. This approach can be an important new avenue of research

lying at the junction of predictive machine learning and mathematical finance and can be an natural tools these types of practical applications.

Author Contributions: For research articles with several authors, a short paragraph specifying their individual contributions must be provided. The following statements should be used "conceptualization, X.X. and Y.Y.; methodology, X.X.; software, X.X.; validation, X.X., Y.Y. and Z.Z.; formal analysis, X.X.; investigation, X.X.; resources, X.X.; data curation, X.X.; writing–original draft preparation, X.X.; writing–review and editing, X.X.; visualization, X.X.; supervision, X.X.; project administration, X.X.; funding acquisition, Y.Y.", please turn to the CRediT taxonomy for the term explanation. Authorship must be limited to those who have contributed substantially to the work reported.

Funding: This research was funded by the ETH Zürich Foundation and by the Natural Sciences and Engineering Research Council of Canada (NSERC).

Acknowledgments: The authors thank Alina Stancu for many helpful discussions.

Abbreviations

The following abbreviations are used in this manuscript:

A-Reg	Arbitrage-Regularization
AFNS	Arbitrage-Free Nelson Siegel Model
A-Reg(dPCA)	Arbitrage-Regularized Dynamic Principal Component Analysis Model A-Reg(NS)
Arbitrage-Regularized Nelson-Siegel Model	
dPCA	Dynamic Principal Component Analysis Model NFLVR
No Free Lunch with Vanishing Risk	
NS	Nelson Siegel Model

Appendix A Background

Some relevant technical background is briefly discussed within this appendix. These topics include related aspects of the functional Itô calculus introduced in Dupire (2009) and developed in Fournie (2010), as well as pertinent stochastic differential geometric considerations; as developed in Elworthy (1982). Some elements from arbitrage-theory are also discussed concisely.

Next some background on arbitrage theory in large financial markets is discussed.

Appendix A.1 Arbitrage-Theory

The efficient market hypothesis, introduced in Bachelier (1900), states that the typical market participant cannot earn a risk-less profit. The efficient market hypothesis has found several mathematical formulations, as summarized in Fontana (2014). The most commonly used form is No Free Lunch with Vanishing Risk (NFLVR) as formulated in the sequence of papers Delbaen and Schachermayer (1998) which builds on the ideas of Harrison and Kreps (1979). Essentially, in the case of locally bounded processes, NFLVR expresses the non-existence of arbitrage-strategies as the existence of an equivalent local martingale measure (ELMM); that is, a probability measure which is equivalent to the reference probability measure and which simultaneously makes the price process of each market asset a local martingale.

However, mathematically bond markets are unlike traditional financial markets in that they are comprised of an uncountable number of assets, one for each potential maturity; thus, the results of

Delbaen and Schachermayer (1998) no longer apply since their formulation requires that only a finite number of assets be tradeable. Instead, in the setting of such a *large financial market*, a satisfactory and economically meaningful no-arbitrage condition is obtained in Cuchiero et al. (2016) by considering strategies which can be described by limits of classical strategies written on a finite number of market assets. It is shown in Cuchiero et al. (2016), that when each asset in the market is locally bounded, as in Delbaen and Schachermayer (1998), then the no-arbitrage condition derived in Cuchiero et al. (2016) reduces to the existence of an equivalent local martingale measure. However, if the local-boundedness assumption is dropped, then the existence of an equivalent local martingale measure remains sufficient for precluding no-arbitrage but it no longer necessary.

Appendix A.2 Functional Itô Calculus

In what follows, the set of $d \times d$ symmetric positive definite matrices will be denoted by S_d^+. Moreover, the Skorohod space of cádlág paths in $\mathbb{R}^d$ and S_t^+ will be respectively denoted by $D([0,T];\mathbb{R}^d)$ and $D([0,T];S_d^+)$. Moreover, for any $\mathbb{R}^d$-valued semi-martingale X_t, one associates to it an S_d^+-valued process $[[X]]_t$ defined by

$$[X]_t = \int_0^t [[X]]_u du.$$

Here $[[X]]_t$ is interpreted as the local quadratic variation of X_t.

The functional Itô calculus of Dupire (2009) and Fournie (2010), has found many applications in mathematical finance. Applications range from, but are not limited to, computational methods for the Greeks of path-dependent options in Jazaerli and Saporito (2017) to portfolio theory in Pang and Hussain (2015). For fixed $T > 0$, the basic concept of the Functional Itô Calculus, relies on non-anticipative and path-dependent extensions of the time and spatial derivative operators. Both these extensions defined on any cádlág path in

$$\Lambda_d \triangleq \bigcup_{t\in[0,T]} D([0,t];\mathbb{R}^d) \times D([0,t];S_d^+),$$

by artificially extending its endpoint either vertically or horizontally. For any fixed $0 \leq t \leq s \leq T$, the horizontal extension of a path $x_t \in \Lambda_d$ is defined by

$$x_{t,s-t}(u) \triangleq \begin{cases} x_u : & 0 \leq u \leq t \\ x_t : & t \leq u \leq s \end{cases}, \tag{A1}$$

and its height $h > 0$ vertical extension is defined by

$$x^{t,h}(u) \triangleq \begin{cases} x_u : & 0 \leq u < t \\ x_t + h : & u = t \end{cases}. \tag{A2}$$

For a functional from Λ_d to $\mathbb{R}$, it's vertical and horizontal derivatives on the path $x_t \in \Lambda_d$ are defined by infinitesimally extending the path $S(x_t, v_t)$ either vertically or horizontally using (A1) and (A2). However, since the calculus should not look into the future, instead only non-anticipative functionals are to be considered.

Definition A.1 (Non-Anticipative Functional; (Fournie 2010, Definition 2.1)). *A non-anticipative functional is a family of functionals* $S \triangleq \{S_t\}_{t\in[0,T]}$ *where*

$$\begin{aligned} S_t : D([0,t];\mathbb{R}^d) \times D([0,t];S_d^+) &\to \mathbb{R} \\ (x_t, v_t) &\mapsto S_t(x_t, v_t) \end{aligned}$$

is measurable to the Borel σ*-algebra on* $D([0,t];\mathbb{R}^d) \times D([0,t];S_d^+)$.

Analogously to classical calculus, the limiting ratio between the difference of $S(x_t, v_t)$ and its extensions define its horizontal and vertical at any given time $0 \leq t$, respectively, by

$$\mathcal{D}S(x_t, v_t) = \lim_{\Delta\downarrow 0} \frac{S(x_{t,\Delta+t}, v_{t,\Delta+t}) - S(x_t, v_t)}{\Delta} \quad \Delta S_t(x_t, v_t) = \lim_{h\downarrow 0} \frac{S(x^{t,h}, v^{t,h}) - S(x_t, v_t)}{h}; \tag{A3}$$

where the limits defined in (A3) are taken in Λ_d with respect to metric

$$d(x_t, y_s) \triangleq \sup_{u\in[0,s]} \|x_{t,s-t}(u) - y_u\| + |s-t|,$$

and for any $x \in \mathbb{R}^d, v \in S_d^+$ one has $\|(x,v)\| = \sqrt{\|x\|^2 + \|v\|_F^2}$ where $\|x\|$ is the usual Euclidean norm and $\|v\|_F$ is the Fröbenius norm. As it will be seen shortly, the horizontal and vertical derivatives extend the time and spacial derivatives from ordinary calculus. However, some technical remarks must first be addressed.

In general, these path derivatives are not defined on any non-anticipative functional $S : \Lambda_d \to \mathbb{R}$ moreover even if it is, analogously to the classical calculus, there is no guarantee that its vertical (resp. horizontal) derivative is continuous with respect to d. Analogously with the traditional Itô calculus, the collection of paths for which one can derive a useful Itô formula are those which admit one continuous horizontal derivative and for which $\mathcal{D}S(x_t, v_t)$ and two vertical derivatives; i.e. $\mathcal{D}^2S(x_t, v_t) \triangleq \mathcal{D}(\mathcal{D}S(x_t, v_t))$ are both continuous.

However, for an tractable extension of the Itô formula with respect to S to be possible it is additionally required that S be boundedness-preserving. Here, a functional $S : \Lambda_d \to \mathbb{R}$ is said to be boundedness-preserving if, for every non-empty compact subset $K \subseteq \mathbb{R}$, there exists some $C_K > 0$ such that $|f(z_t)| \leq C_K$ if $z_t \in \Lambda_d$ satisfies

$$\{y \in \mathbb{R} : z_t(s) = y \text{ for some } 0 \leq s \leq t\} \subseteq K. \tag{A4}$$

The collection subset of all functional $S : \Lambda_d \to \mathbb{R}$ which are boundedness-preserving and have continuous boundedness-preserving derivatives $\Delta S(x_t, v_t)$, $\mathcal{D}S(x_t, v_t)$, and $\mathcal{D}^2S(x_t, v_t)$ at every path $x_t \in \Lambda_d$ is denoted by $\mathbb{C}^{1,2}$.

For any functional $S \in \mathbb{C}^{1,2}$ and any $\mathbb{R}^d$-valued semi-martingale, if in addition it satisfies the predictable-dependence condition

$$(\forall t \in [0,T])(\forall (x,v) \in D([0,t];\mathbb{R}^d) \times S_d^+) \qquad S_t(x_t, v_t) = S_t(x_t, v_t). \tag{A5}$$

Theorem A.2 (Functional Itô Formula (Fournie 2010, Theorem 4.1)). For any non-anticipative functional $S \in \mathbb{C}^{1,2}$ satisfying (A5) and any $\mathbb{R}^d$-valued semi-martingale the following holds

$$\begin{aligned} S_t(X_t, [[X]]_t) - S_t(X_t, [[X]]_t) = &\int_0^t \mathcal{D}S(X_u, [[X]]_u)du + \int_0^t \Delta S_u(X_u, [[X]]_u)dX_u \\ &+\frac{1}{2}\int_0^t \operatorname{tr}\left(\Delta^2 S_u(X_u, [[X]]_u)d[X]_u\right). \end{aligned} \tag{A6}$$

This section closes by noting that Theorem A.2 is a strict generalization of the classical Itô formula. This is because, the vertical and horizontal derivatives reduce to the familiar spacial and time derivatives when S_t does not depend on any path-data as formalized by the following result.

Proposition A.3 ((Fournie 2010, Example 1)). If $S(x_t, v_t) = f(t, x_t(t))$ for some $f \in C^{1,2}([0,t] \times \mathbb{R}^d; \mathbb{R})$ then

$$\mathcal{D}S(x_t, v_t) = \partial_t f(t, x_t) \text{ and } \Delta^i S(x_t, v_t) = \partial_x^i f(t, x_t),$$

for $i = 1, 2$.

Appendix A.3 Background on Γ*-Convergence*

Pioneered in De Giorgi (1975), the theory of Γ-convergence describes the precise conditions required for the optimizers of a sequence of loss-functions $\{\ell_n\}_{n\in\mathbb{N}}$ to converge to the optimizer of a limiting loss-function ℓ. The entire theory of Γ-convergence can be seen as a sequential generalization of the Weierstrass's theorem, a fundamental existence result from non-convex optimization theory. Geometrically, Weierstrass's theorem states that it possible to continuously descent along the epigraph of ℓ (lower semi-continuity), if the set all small values of ℓ is compact (coercivity) then ℓ can be minimized; granted that ℓ does not only take infinite-values (proper).

Theorem A.4 (Weierstrass' Theorem; (Focardi 2012, Theorem 2.2)). Let (X, d) be a metric space and $\ell : X \to \mathbb{R}$ be a lower semi-continuous function which is mildly coercive, that is there exists a sequentially compact subset K of X such that

$$\inf_{x\in K} \ell(x) = \inf_{x\in X} \ell(x). \tag{A7}$$

Then ℓ admits a minimizer on X if in addition $\inf_{x\in X} \ell(x)$ is finite.

Γ-limits provide precise conditions ensuring that any sequence of optimizers of $\{\ell_n\}_{n\in\mathbb{N}}$ converges to an optimizer of ℓ if ℓ is the Γ-limit of $\{\ell_n\}_{n\in\mathbb{N}}$, written $\Gamma\text{-}\lim_{n\to\infty} \ell_n = \ell$. Before providing a precise definition of Γ-limits, a few of its properties are discussed.

Theorem A.5 (Properties of Γ-convergence; (Focardi 2012, Theorem 2.8)). Let (X, d) be a metric space and $\{\ell_n\}_{n\in\mathbb{N}}$ be a sequence of functions from (X, d) to $\mathbb{R} \cup \{\infty\}$. If $\Gamma\text{-}\lim_{n\to\infty} \ell_n$ exists, then

(i) (Lower Semicontinuity): $\Gamma\text{-}\lim_{n\to\infty} \ell_n$ is lower semicontinuous on X,
(ii) (Stability Under Continuous Perturbation): If $g : X \to \mathbb{R}$ is continuous, then

$$\Gamma\text{-}\lim_{n\to\infty} (\ell_n + g) = \left(\Gamma\text{-}\lim_{n\to\infty} \ell_n\right) + g,$$

(iii) (Stability Under Relaxation): For every $n \in \mathbb{N}$ let $\{\tilde{\ell}_n\}_{n\in\mathbb{N}}$ be a sequence of functions from X to $\mathbb{R}\cup\{\infty\}$ satisfying $\ell_n^{lsc} \leq \tilde{\ell}_n \leq \ell_n$. Then

$$\Gamma\text{-}\lim_{n\to\infty} \tilde{\ell}_n = \Gamma\text{-}\lim_{n\to\infty} \ell_n,$$

where ℓ^{lsc} is the largest lower semi-continuous function dominated by ℓ, point-wise.

The first of the two critical ingredients in Theorem A.4 was the lower semi-continuity of ℓ and the second was its coerciveness. Analogously to the definition of equi-continuity, in general, when working with a sequence of functions, to be able to apply the analogous machinery to Theorem A.4 we require that there exists a non-empty compact subset of K satisfying

$$\inf_{x\in X} \ell_n(x) = \inf_{x\in X} \ell(x);\ (\forall n \in \mathbb{N}). \tag{A8}$$

The property described by (A8) is called *mild equi-coerciveness*. A stronger condition, that we will make use of is *equi-coerciveness*, which states that for every $t > 0$, there exists a compact subset K_t of (X, d) satisfying

$$\bigcup_{n\in\mathbb{N}} \{x \in X : \ell_n(x) \leq t\} \subseteq K_t.$$

The central result in the Theory of Γ-converges is the following extension of Theorem A.4.

Theorem A.6 (The Fundamental Theorem of Γ-Convergence; (Braides 2002, Theorem 2.10),(Focardi 2012, Theorem 2.1))**.** If $\{\ell_n\}_{n\in\mathbb{N}}$ is a mildly equi-coercive sequence of functions from X to $\mathbb{R}\cup\{\infty\}$ for which the Γ-limit exists in X, then

$$\lim_{k\uparrow\infty} \inf_{x\in X} \ell_{k_n}(x) = \inf_{x\in X} \Gamma\text{-}\lim_{n\to\infty} \ell_n(x).$$

If moreover, $\{\ell_n\}_{n\in\mathbb{N}}$ is equicoercive, then $\lim_{n\uparrow\infty} \operatorname{arginf}_{x\in X} \ell_n(x)$ exists in X and

$$\lim_{k\uparrow\infty} \operatorname{arginf}_{x\in X} \ell_{k_n}(x) \in \operatorname*{arginf}_{x\in X} \Gamma\text{-}\lim_{n\to\infty} \ell_n(x).$$

This section closes with the precise definition of convergence in the Γ-sense.

Definition A.7 ((Dal Maso 1993, Chapter 4))**.** *Let $\{\ell_n\}_{n\in\mathbb{N}}$ be a sequence of $\mathbb{R}\cup\{\infty\}$-valued functions on a metric space (X, d). A function ℓ is the Γ-limit of $\{\ell_n\}_{n\in\mathbb{N}}$ if and only if both*

(i) $\ell^{lsc}(x) \leq \text{lim-inf}_{n\in\mathbb{N}}\, \ell_n(x_n)$ *for every net $\{x_n\}_{n\in\mathbb{N}}$ converging to x in (X, d),*
(ii) $\ell^{lsc}(x) \geq \text{lim-inf}_{n\in\mathbb{N}}\, \ell_n(y_n)$ *for some net $\{y_n\}_{n\in\mathbb{N}}$ converging to x in (X, d)*

where ℓ^{lsc} is the largest lower semi-continuous function dominated by ℓ point-wise.

Appendix B Proofs

Proof of Proposition 3.1. For legibility, for each $u \in \mathcal{U}$, we represent the process $\phi(t, \beta_t, u)$ by

$$\phi_t^u = \phi_0^u + \int_0^t \alpha^u(s, \phi_s^u)ds + \int_0^t \gamma^u(s, \phi_s^u)dW_s.$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2020 doi:10.20944/preprints202003.0022.v1

Peer-reviewed version available at Risks 2020, 8, 40; doi:10.3390/risks8020040

By the Functional Itô Formula, (Cont and Fournié 2013, Theorem 4.1), if follows that, for every $u \in \mathcal{U}$

$$\begin{aligned} S_t(\phi_t^u, [[\phi^u]]_t; u) =& S_0(0, \phi_0^u, [[\phi^u]]_0; u) \\ &+ \int_0^t \left[\mathcal{D}S_s(\phi_s^u, [[\phi^u]]_s; u) + \Delta S_s(\phi_s^u, [[\phi^u]]_s; u)\alpha^u(s, \phi_s^u) \right. \\ &\left. + \frac{1}{2}[\Delta^2 S_s(\phi_s^u, [[\phi^u]]_s; u)](\gamma^u(s, \phi_s^u)) \right] ds \\ &+ \int_0^t \Delta S_s(\phi_s^u, [[\phi^u]]_s; u)\gamma^u(s, \phi_s^u) dW_s. \end{aligned} \tag{A9}$$

From (A9) it follows that, for each $u \in \mathcal{U}$, the price processes $X_t(u)$ are $\mathbb{P}_N^\star$-local-martingales if and only if

$$-\mathcal{D}S_s(\phi_s^u, [[\phi^u]]_s; u) = \Delta S_s(\phi_s^u, [[\phi^u]]_s; u)\alpha_s^u + \frac{1}{2}[\Delta^2 S_s(\phi_s^u, [[\phi^u]]_s; u)](\gamma_s^u) \tag{A10}$$

Next, the quantities a^u and γ^u are described. By the usual Itô formula, it follows that for each $u \in \mathcal{U}$

$$\begin{aligned} \phi(t, \beta_t, u) &= \int_0^t \frac{\partial \phi}{\partial t}(s, \beta_s, u) ds + \int_0^t \sum_{i=1}^d \frac{\partial \phi}{\partial \beta^i}(s, \beta_s, u) d\beta_s^i + \int_0^t \frac{1}{2} \sum_{i,j=1}^d \left(\frac{\partial^2 \phi}{\partial \beta^i \partial \beta^j}(s, \beta_s, u) \right) d[b]_s^{i,j} ds \\ &= \int_0^t \left[\frac{\partial \phi}{\partial t}(s, \beta_s, u) + \sum_{i=1}^d \frac{\partial \phi}{\partial \beta^i}(s, \beta_s, u)\mu^i(s, \beta_s) ds \right. \\ &\left. + \frac{1}{2} \sum_{i,j=1}^d \left(\frac{\partial^2 \phi}{\partial \beta^i \partial \beta^j}(s, \beta_s, u) \right) \sigma^i(s, \beta_s)\sigma^j(s, \beta_s) \right] ds \\ &+ \int_0^t \sum_{i=1}^d \frac{\partial \phi}{\partial \beta^i}(s, \beta_s, u)\sigma^i(s, \beta_s) dW_s^i. \end{aligned} \tag{A11}$$

Therefore, (A11) implies that

$$\begin{aligned} \alpha_s^u &= \frac{\partial \phi}{\partial t}(s, \beta_s, u) + \sum_{i=1}^d \frac{\partial \phi}{\partial \beta^i}(s, \beta_s, u)\mu^i(s, \beta_s) + \frac{1}{2} \sum_{i,j=1}^d \left(\frac{\partial^2 \phi}{\partial \beta^i \partial \beta^j}(s, \beta_s, u) \right) \sigma^i(s, \beta_s)\sigma^j(s, \beta_s) \\ \gamma_s^u &= \sum_{i=1}^d \frac{\partial \phi}{\partial \beta^i}(s, \beta_s, u)\sigma^i(s, \beta_s). \end{aligned} \tag{A12}$$

Incorporating (A12) into (A10) yields (9). Therefore, for $S_t(\phi_t^u, [[\phi^u]]_t, u)$ is a $\mathbb{P}_N^\star$-local-martingale, simultaneously for every $u \in \mathcal{U}$, if and only if $\mathbb{P}_N^\star$-a.s. (9) holds simultaneously for every $u \in \mathcal{U}$. □

Proof of Theorem 3.5. We begin by showing (ii) and (iii), simultaneously. Since $(I \times \mathcal{U}, \mathcal{B}(I \times \mathcal{U}), \mu)$ is a finite measure space then

$$\lim_{\lambda \uparrow \infty; \lambda \geq 2} \left(\int_{(t,u) \in I \times \mathcal{U}} |\Lambda_t^u(\phi)|^\lambda d\mu(t, u) \right)^{\frac{1}{\lambda}} = \operatorname*{esssup}_{(t,u) \in I \times \mathcal{U}} |\Lambda_t^u(\phi)|; \tag{A13}$$

By Assumption (3.4) (i), the map $(t, u) \mapsto \Lambda_t^u(\phi)$ is continuous, for each $\phi \in \mathscr{H}$, therefore

$$\operatorname*{esssup}_{(t,u) \in I \times \mathcal{U}} |\Lambda_t^u(\phi)| = \sup_{(t,u) \in I \times \mathcal{U}} |\Lambda_t^u(\phi)|. \tag{A14}$$

Since the product of limits is the limit of the product, then (A13) yields

$$\begin{aligned}\lim_{\lambda\uparrow\infty;\,\lambda\geq 2}\lambda\left(\int_{(t,u)\in I\times\mathcal{U}}|\Lambda_t^u(\phi)|^{\lambda}d\mu(t,u)\right)^{\frac{1}{\lambda}} &= \lim_{\lambda\uparrow\infty;\,\lambda\geq 2}\lambda\operatorname*{esssup}_{(t,u)\in I\times\mathcal{U}}|\Lambda_t^u(\phi)|\\ &= \lim_{\lambda\uparrow\infty;\,\lambda\geq 2}\lambda\sup_{(t,u)\in I\times\mathcal{U}}|\Lambda_t^u(\phi)|\\ &= \begin{cases}0 & : \sup\limits_{(t,u)\in I\times\mathcal{U}}|\Lambda_t^u(\phi)|=0\\ \infty & : \text{else.}\end{cases}\end{aligned} \tag{A15}$$

Since μ is a probability measure in $I\times\mathcal{U}$ then for every $1\leq\lambda_1\leq\lambda_2<\infty$, it follows that for every $\phi\in\mathscr{H}$

$$\left(\int_{(t,u)\in I\times\mathcal{U}}|\Lambda_t^u(\phi)|^{\lambda_1}d\mu(t,u)\right)^{\frac{1}{\lambda_1}}\leq\left(\int_{(t,u)\in I\times\mathcal{U}}|\Lambda_t^u(\phi)|^{\lambda_2}d\mu(t,u)\right)^{\frac{1}{\lambda_2}}\leq\operatorname*{esssup}_{(t,u)\in I\times\mathcal{U}}|\Lambda_t^u(\phi)|\,. \tag{A16}$$

Thus, for every $\phi\in\mathscr{H}$, the convergence described by (A13) is monotone (increasing) and non-negative; therefore by the monotone convergence theorem, it follows that

$$\lim_{\lambda\uparrow\infty;\,\lambda\geq 2}\lambda\mathbb{E}_{\mathbb{P}}\left[\left(\int_{(t,u)\in I\times\mathcal{U}}|\Lambda_t^u(\phi)|^{\lambda}d\mu(t,u)\right)^{\frac{1}{\lambda}}\right]=\begin{cases}0 & : \sup\limits_{(t,u)\in I\times\mathcal{U}}\Lambda_t^u(\phi)=0,\ \mathbb{P}_N^{\star}-a.s.\\ \infty & : \text{else.}\end{cases} \tag{A17}$$

Applying Proposition 3.1 to the right-hand side of (A17), it follows that

$$\iota_{\mathscr{H}}(\phi)=\begin{cases}0 & : \sup\limits_{(t,u)\in I\times\mathcal{U}}\Lambda_t^u(\phi),\ \mathbb{P}_N^{\star}-a.s.=0\\ \infty & : \text{else}\end{cases} \tag{A18}$$

where $\iota_{\mathscr{H}}$ is defined as in (14). Therefore, the following limit holds

$$\lim_{\lambda\uparrow\infty;\,\lambda\geq 2}\lambda\mathbb{E}_{\mathbb{P}}\left[\left(\int_{(t,u)\in I\times\mathcal{U}}|\Lambda_t^u(\phi)|^{\lambda}d\mu(t,u)\right)^{\frac{1}{\lambda}}\right]=\iota_{\mathscr{H}}(\phi)\qquad(\forall\phi\in\mathscr{H})\,. \tag{A19}$$

Thus (A19) establishes the convergence of the penalty functions AF^{λ} to $\iota_{\mathscr{H}}$, in $\mathscr{H}$. Next, their Γ-convergence is established and their Γ-convergence is used to deduce the Γ-convergence of the objective functions in (8) to the objective function of problem (7).

Applying (A16) and the monotonicity of integration, it follows that for every $\phi\in\mathscr{H}$

$$\lambda_1\mathbb{E}_{\mathbb{P}}\left[\left(\int_{(t,u)\in I\times\mathcal{U}}|\Lambda_t^u(\phi)|^{\lambda_1}d\mu(t,u)\right)^{\frac{1}{\lambda_1}}\right]\leq\lambda_2\mathbb{E}_{\mathbb{P}}\left[\left(\int_{(t,u)\in I\times\mathcal{U}}|\Lambda_t^u(\phi)|^{\lambda_2}d\mu(t,u)\right)^{\frac{1}{\lambda_2}}\right]\leq\iota_{\mathscr{H}}(\phi). \tag{A20}$$

Thus, (A20) together with (Dal Maso 1993, Proposition 5.4) and (Braides 2002, Remark 1.40 (ii)) imply that (on $\mathscr{H}$)

$$\Gamma\text{-}\lim_{\lambda\uparrow\infty;\,\lambda\geq 2}\lambda\mathbb{E}_{\mathbb{P}}\left[\left(\int_{(t,u)\in I\times\mathcal{U}}|\Lambda_t^u(\phi)|^{\lambda}d\mu(t,u)\right)^{\frac{1}{\lambda}}\right]=\iota_{\mathscr{H}}^{lsc}(\phi), \tag{A21}$$

where $\iota^{lsc}_{\mathscr{H}}$ is the lower-semi-continuous relaxation of $\iota_{\mathscr{H}}$ on $\mathscr{H}$; that is, the smallest lower-semi-continuous function dominating $\iota_{\mathscr{H}}$ on $\mathscr{H}$; a precise description can be found on (Focardi 2012, page 11). However, Assumption (3.4) (ii) implies that $\iota_{\mathscr{H}}$ is indeed lower-semi-continuous; thus $\iota^{lsc}_{\mathscr{H}} = \iota_{\mathscr{H}}$, on $\mathscr{H}$. Therefore (A21) simplifies (on $\mathscr{H}$) to

$$\Gamma\text{-}\lim_{\lambda\uparrow\infty;\lambda\geq 2} \lambda\mathbb{E}_{\mathbb{P}}\left[\left(\int_{(t,u)\in I\times\mathcal{U}} |\Lambda^u_t(\phi)|^\lambda d\mu(t,u)\right)^{\frac{1}{\lambda}}\right] = \iota_{\mathscr{H}}(\phi). \tag{A22}$$

Since Γ-limits are invariant under continuous perturbation, see (Focardi 2012, Theorem 2.8.), then (A22) and the continuity of $\ell(\varphi - \cdot)$ on $\mathscr{H}$ implies that (on $\mathscr{H}$)

$$\Gamma\text{-}\lim_{\lambda\uparrow\infty;\lambda\geq 2} \ell(\varphi - \cdot) + \lambda\mathbb{E}_{\mathbb{P}}\left[\left(\int_{(t,u)\in I\times\mathcal{U}} |\Lambda^u_t(\cdot)|^\lambda d\mu(t,u)\right)^{\frac{1}{\lambda}}\right] = \ell(\varphi-\phi) + \iota_{\mathscr{H}}(\cdot). \tag{A23}$$

In order to apply the Fundamental Theorem of Γ-convergence, the family of functions on the left-hand side of (A23) must be equicoercive. Since $\lambda\mathbb{E}_{\mathbb{P}}\left[\left(\int_{(t,u)\in I\times\mathcal{U}} |\Lambda^u_t(\cdot)|^\lambda d\mu(t,u)\right)^{\frac{1}{\lambda}}\right]$ is non-negative and since $\mathscr{H}$ is unbounded then $\ell(\varphi-\phi)$ is coercive on $\mathscr{H}$, ie:

$$\lim_{\lambda\uparrow\infty;\lambda\geq 2} \ell(\varphi-\phi) = \infty, \tag{A24}$$

thus, it follows that

$$\ell(\varphi-\phi) \leq \ell(\varphi-\phi) + \lambda\mathbb{E}_{\mathbb{P}}\left[\left(\int_{(t,u)\in I\times\mathcal{U}} |\Lambda^u_t(\cdot)|^\lambda d\mu(t,u)\right)^{\frac{1}{\lambda}}\right] \qquad (\forall\lambda\geq 2)\,; \tag{A25}$$

whence, by (Dal Maso 1993, Proposition 7.7) together with (A24) $\left\{\ell(\varphi-\phi) + \lambda\mathbb{E}_{\mathbb{P}}\left[\left(\int_{(t,u)\in I\times\mathcal{U}} |\Lambda^u_t(\cdot)|^\lambda d\mu(t,u)\right)^{\frac{1}{\lambda}}\right]\right\}_{\lambda\geq 2}$ forms an equicoercive family, on $\mathscr{H}$.

Thus, $\left\{\ell(\varphi-\phi) + \lambda\mathbb{E}_{\mathbb{P}}\left[\left(\int_{(t,u)\in I\times\mathcal{U}} |\Lambda^u_t(\cdot)|^\lambda d\mu(t,u)\right)^{\frac{1}{\lambda}}\right]\right\}_{\lambda\geq 2}$ defines an equicoercive family which Γ-converges to $\iota_{\mathscr{H}}$, on $\mathscr{H}$. Therefore, together (A23) and (A25) imply that the Fundamental Theorem of Γ-convergence, (Dal Maso 1993, Theorem 7.8), applies. Hence,

$$\lim_{\lambda\uparrow\infty;\lambda\geq 2} \inf_{\phi\in\mathscr{H}} \ell(\varphi-\phi) + \lambda\mathbb{E}_{\mathbb{P}}\left[\left(\int_{(t,u)\in I\times\mathcal{U}} |\Lambda^u_t(\phi)|^\lambda d\mu(t,u)\right)^{\frac{1}{\lambda}}\right] = \min_{\phi\in\mathscr{H}} \ell(\varphi-\phi) + \iota_{\mathscr{H}}(\phi). \tag{A26}$$

This shows both (ii) and (iii).

Lastly, for (i), (Dal Maso 1993, Theorem 7.8) also implies that $\ell(\varphi-\phi) + \iota_{\mathscr{H}}(\cdot)$ is coercive on $\mathscr{H}$. Hence, $\ell(\varphi-\phi) + \iota_{\mathscr{H}}(\cdot)$ is coercive, lower-semi-continuous, and bounded-below by 0. Therefore, by Weierstrass's Theorem, (Focardi 2012, Theorem 2.2), it follows that $\ell(\varphi-\phi) + \iota_{\mathscr{H}}(\cdot)$ admits a minimizer on $\mathscr{H}$. □

Proof of Proposition 3.7. For every $0 \leq t$ Assumption 3.6 implies that

$$\begin{aligned} \left|S_t(\hat{\phi}_t^{u_i}, [[\hat{\phi}^{u_i}]]_t; u_i) - S_t(\phi_t^{u_i}, [[\phi^{u_i}]]_t; u_i)\right| M &= \frac{M}{m} m \\ &< \frac{M}{m} \left|S_t(\phi_t^{u_i}, [[\phi^{u_i}]]_t; u_i)\right| \\ &< k_t \left|S_t(\phi_t^{u_i}, [[\phi^{u_i}]]_t; u_i)\right|. \end{aligned} \tag{A27}$$

Therefore, (Guasoni 2006, Lemma 2.1) and (Guasoni 2006, Remark 2.5) implies that for any admissible strategy θ

$$V_t(\theta) \leq \int_0^t S_s(\hat{\phi}_s^{u_i}, [[\hat{\phi}^{u_i}]]_s; u_i) ds. \tag{A28}$$

By Theorem 3.4, since $S_s(\hat{\phi}_s^{u_i}, [[\hat{\phi}^{u_i}]]_s; u_i)$ is a $\mathbb{P}_N$-local martingale and by construction $\mathbb{P}_N \sim \mathbb{P}$ then the (Delbaen and Schachermayer 1994, Fundamental Theorem of Asset Pricing) implies that if

$$\mathbb{P}\left(0 \leq \int_0^T S_s(\hat{\phi}_s^{u_i}, [[\hat{\phi}^{u_i}]]_s; u_i)\right) \Rightarrow \mathbb{P}\left(0 = \int_0^T S_s(\hat{\phi}_s^{u_i}, [[\hat{\phi}^{u_i}]]_s; u_i)\right). \tag{A29}$$

Combining (A28) and (A29) yields

$$\mathbb{P}\left(0 \leq V_T(\theta)\right) \Rightarrow \mathbb{P}\left(0 = V_T(\theta)\right),$$

for every $0 \leq T$ and every finite number of $u_1, \ldots, u_n \in U$. □

Proof of Theorem 4.1. By definition of $(\mathbb{R}^d, g_t)$, β_t, and $\mathscr{H}$ Assumptions 2.2 (i) and (iv) hold; thus only Assumptions 2.2 (ii) and (iii) must be verified, in order to ensure that the stated problem falls within the scope of this paper. Let $\mu = \tilde{\mu} \otimes \upsilon$ where $\frac{d\tilde{\mu}}{dm}(u) = \frac{e^{-|u|^\kappa}}{\Gamma\left(1+\frac{1}{\kappa}\right)} I_{[0,\infty)}$, υ is the unique probability measure with Lebesgue density proportional to $e^{-\|\beta\|^\kappa}$ on $\mathbb{R}^d$, and where $\frac{d\upsilon}{dm}(t) = e^{-|t|} 1_{[0,\infty)}$, $1_{[0,\infty)}$ is the (probabilistic not convex analytic) indicator function on the interval $[0,\infty)$, here m is the Lebesgue measure on $\mathbb{R}$. Therefore, the elements of $W_w^{p,k}(I \times \mathbb{R}^d \times \mathcal{U})$ are elements of $L^p_{\upsilon\otimes\mu}(I \times \mathbb{R}^d \times \mathcal{U})$. Since $[0,\infty) \times [0,\infty) \times \mathbb{R}^d$ has a smooth boundary and since k was assumed to satisfy (18), then the (weighted) Morrey-Sobolev Theorem of Brown and Opic (1992) applies. Therefore, $W_w^{p,k}(I \times \mathbb{R}^d \times \mathcal{U})$ can be continuously embedded within $C^{2,2,2}(I \times \mathbb{R}^d \times \mathcal{U})$ and therefore Assumption 2.2 (ii) holds. Furthermore, by (5) and together with (Cont and Fournié 2013, Example 1) each $S_t(\cdot,\cdot;u)$ satisfies Assumption 2.2 (iii); thus Assumption 2.2 is satisfied.

Next, (22) is reformulated in terms of Theorem 3.5 and Assumptions 3.4 are verified. Subsequently, the optimizers of the objective-function under the limit on the left-hand side of (22) are shown to exist for $\lambda \geq 2$.

In the case where each $S_t(\cdot,\cdot;u)$ is as in (5), it is shown in (Filipović 2009, Proposition 9.3) that for each $u \in \mathcal{U}$ the bond prices $S_t\left(\phi_t^u, [[\phi^u]]_t; u\right)$ are each $\mathbb{P}_N$-local martingales if and only if for every $u \in \mathcal{U}$, $t \in I$, and $\mathbb{P}_N$-a.e $\omega \in \Omega$, the following holds

$$\begin{aligned} 0 = &\left(\phi_0(0) - \phi_0(u) + \sum_{i=1}^d \gamma^i \Phi^i(u) - \frac{1}{2}\sum_{i,j=1}^d \alpha^{i,j}\Phi^i(u)\Phi^j(u)\right) \\ &+ \sum_{k=1}^d \beta_t^k \left(\phi_k(0) - \phi_k(u) + \sum_{i=1}^d \gamma^{k,i}\Phi^i(u) - \frac{1}{2}\sum_{i,j=1}^d \alpha_{k;i,j}\Phi^i(u)\Phi^j(u)\right) \end{aligned} \tag{A30}$$

Equation (A30) is satisfied for $\mathbb{P}^\star_N$-a.s every $\omega \in \Omega$, for every $t \in I$, and for every $u \in \mathcal{U}$ each $\Lambda^u_t(\phi) = 0$ are $\mathbb{P}^\star_N$-a.s; where $\Lambda^u_t(\phi)$ is defined by

$$\begin{aligned}\Lambda^u_t(\phi) \triangleq &\left| c_0 - \frac{\partial \Phi^0}{\partial u}(u) + \sum_{i=1}^d \gamma^i \Phi^i(u) - \frac{1}{2}\sum_{i,j=1}^d \alpha^{i,j}\Phi^i(u)\Phi^j(u) \right|^p \\ &+ \sum_{k=1}^d \left| c_k - \frac{\partial \Phi^k}{\partial u}(u) + \sum_{i=1}^d \gamma^{k,i}\Phi^i(u) - \frac{1}{2}\sum_{i,j=1}^d \alpha_{k;i,j}\Phi^i(u)\Phi^j(u) \right|^p .\end{aligned} \tag{A31}$$

Therefore, since $\Lambda^u_t(\phi)$ satisfies (11) then the family $\{AF^\lambda\}_{\lambda>0}$ of functions defined by

$$AF^\lambda(\phi) \triangleq \lambda \mathbb{E}_{\mathbb{P}} \left[\sqrt[\lambda]{\int_0^\infty \int_{\beta\in\mathbb{R}^d} \int_0^\infty \frac{1}{\Gamma(1+\frac{1}{\kappa})} e^{-|u|^\kappa} \left|\Lambda^u_s(\phi)\right|^\lambda ds d\beta du} \right], \tag{A32}$$

define an arbitrage-penalty in the sense of (12). Moreover, by (A31) Equation (A32) further simplifies to

$$AF^\lambda(\phi) = \frac{\lambda}{\Gamma(1+\frac{1}{\kappa})^{\frac{1}{\lambda}}} \sqrt[\lambda]{\int_0^\infty e^{-|u|^\kappa} \left|\Lambda^u_t(\phi)\right|^\lambda du} \tag{A33}$$

Since $W^{k,p}_w(I \times \mathbb{R}^d \times \mathcal{U})$ is continuously embedded in $C^2(I \times \mathbb{R}^d \times \mathcal{U})$, then each equivalence class $\phi \in W^{k,p}_w(I \times \mathbb{R}^d \times \mathcal{U})$ can be identified with a continuous function from $I \times \mathbb{R}^d \times \mathcal{U}$; therefore each of the functions

$$\begin{aligned} u &\mapsto \left| c_0 - \frac{\partial \Phi^0}{\partial u}(u) + \sum_{i=1}^d \gamma^i \Phi^i(u) - \frac{1}{2}\sum_{i,j=1}^d \alpha^{i,j}\Phi^i(u)\Phi^j(u) \right|^p \\ u &\mapsto \left| c_k - \frac{\partial \Phi^k}{\partial u}(u) + \sum_{i=1}^d \gamma^{k,i}\Phi^i(u) - \frac{1}{2}\sum_{i,j=1}^d \alpha_{k;i,j}\Phi^i(u)\Phi^j(u) \right|^p , \end{aligned} \tag{A34}$$

are continuous in u; moreover, they are continuous in t since they are constant in t. Therefore, $(t,u) \mapsto \Lambda^u_t(\phi)$ is continuous for every $\phi \in \mathscr{H}$; whence Assumption 3.4 (i) holds.

Next, Assumption 3.4 (ii) will be verified. Given the dynamics of (16), (Filipović 2009, Proposition 9.3) characterizes all $\phi_0, \dots, \phi_d$ for which the forward-rate curve (19) corresponds to a bond market, through (5), in which each bond price is a $\mathbb{P}^\star_N$-local-martingale; all such $\phi_0, \dots, \phi_d$ are solutions to the differential Riccati system

$$\begin{aligned} \frac{\partial \Phi^0}{\partial u}(u) &= c_0 + \sum_{i=1}^d \gamma^i \Phi^i(u) - \frac{1}{2}\sum_{i,j=1}^d \alpha^{i,j}\Phi^i(u)\Phi^j(u) && \Phi^0(0) = 0 \\ \frac{\partial \Phi^k}{\partial u}(u) &= c_k + \sum_{i=1}^d \gamma^{k,i}\Phi^i(u) - \frac{1}{2}\sum_{i,j=1}^d \alpha_{k;i,j}\Phi^i(u)\Phi^j(u) && \Phi^k(0) = 0; \end{aligned} \tag{A35}$$

where $c_0, \dots, c_k$ are any elements of $\mathbb{R}$. Thus,

$$\begin{aligned} &\left\{\phi \in \mathscr{H} : (\forall u \in \mathcal{U})\, S_t(\phi^u_t, [[\phi^u]]_t; u) \text{ is a } \mathbb{P}^\star_N\text{-local-martingale}\right\} \\ &= \left\{\phi \in \mathscr{H} : (\exists c_0, \dots, c_d \in \mathbb{R})\, \{\Phi_i\}_{i=0}^d \text{ solves (A35)}\right\}, \end{aligned} \tag{A36}$$

where as before, $\{\Phi^i\}_{i=0}^d$ and ϕ are related through (19) and (20). Differentiating across the Riccatti system (A35) with respect to u yields an equivalent differential system of the form

$$\begin{aligned}
\frac{\partial^2\Phi^0}{\partial u^2}(u) = &\quad c_0 + \sum_{i=1}^d \gamma^i \frac{\partial \Phi^i}{\partial u} - \frac{1}{2}\sum_{i,j=1}^d \alpha^{i,j}\left[\frac{\partial\Phi^i}{\partial u}\Phi^j(u) + \Phi^i(u)\frac{\partial\Phi^j}{\partial u}\right],\\
\frac{\partial^2\Phi^k}{\partial u^2}(u) = &\quad c_k + \sum_{i=1}^d \gamma^{k,i} \frac{\partial \Phi^i}{\partial u} - \frac{1}{2}\sum_{i,j=1}^d \alpha_{k;i,j}\left[\frac{\partial\Phi^i}{\partial u}\Phi^j(u) + \Phi^i(u)\frac{\partial\Phi^j}{\partial u}\right],\\
\Phi^0(0) = 0, &\qquad \Phi^k(0) = 0,\\
\frac{\partial\Phi^0}{\partial u}(0) = c_0 &\qquad \frac{\partial\Phi^k}{\partial u}(0) = c_k;
\end{aligned} \tag{A37}$$

Therefore, (A36) can be rewritten as

$$\begin{aligned}
&\{\phi\in\mathscr{H}:\ (\forall u\in\mathcal{U})\ S_t(\phi_t^u, [[\phi^u]]_t; u) \text{ is a } \mathbb{P}_N^\star\text{-local-martingale}\}\\
=&\left\{\phi\in\mathscr{H}:\ \{\Phi_i\}_{i=0}^d \text{ solves (A37) for some } c_0,\dots,c_d\in\mathbb{R}\right\}.
\end{aligned} \tag{A38}$$

Since the $C^k(\mathbb{R};\mathbb{R}^{d+1})$ is equipped with its the topology of uniform convergence on compacts of functions and their first two derivatives, then it follows that the right-hand side of (A38) is closed in $C^2(\mathbb{R};\mathbb{R}^{d+1})$; whence it is closed in the relative topology on $\mathscr{H}\subseteq C^2(\mathbb{R};\mathbb{R}^{d+1})$. Thus, Assumption (ii) holds.

Lastly, since the loss function ℓ, defined by

$$\ell(\varphi-\phi)\triangleq\int_0^\infty\int_{\beta\in\mathbb{R}^d} e^{-|u|^\kappa-\|\beta\|^\kappa}\left(\varphi(u,\beta)-\phi(u,\beta)\right)^p du d\beta; \tag{A39}$$

is continuous on $\mathscr{H}$; then the conditions for Theorem 3.5 are all met. Therefore, (22) holds.

Since second-order differential operators are continuous from $C^2(I\times\mathbb{R}^d\times\mathcal{U})$ to $C^0(I\times\mathbb{R}^d\times\mathcal{U})$, where the latter is equipped with the convergence on compact topology, then functions

$$\begin{aligned}
\Phi&\mapsto\left|c_0-\frac{\partial\Phi^0}{\partial u}(u)+\sum_{i=1}^d\gamma^i\Phi^i(u)-\frac{1}{2}\sum_{i,j=1}^d\alpha^{i,j}\Phi^i(u)\Phi^j(u)\right|^p\\
\Phi&\mapsto\left|c_k-\frac{\partial\Phi^k}{\partial u}(u)+\sum_{i=1}^d\gamma^{k,i}\Phi^i(u)-\frac{1}{2}\sum_{i,j=1}^d\alpha_{k;i,j}\Phi^i(u)\Phi^j(u)\right|^p,
\end{aligned} \tag{A40}$$

are continuous from $C^2(I\times\mathbb{R}^d\times\mathcal{U})$ to $[0,\infty)$. Furthermore, since $W_w^{p,k}(I\times\mathbb{R}^d\times\mathcal{U})$ is continuously embedded within $C^2(I\times\mathbb{R}^d\times\mathcal{U})$ then the functions of (A40) are continuous from $W_w^{p,k}(I\times\mathbb{R}^d\times\mathcal{U})$ to $[0,\infty)$. The definition of the weight function in (17) implies that, for every $\phi\in\mathscr{H}$ and every $\lambda\geq 2$, the integral (A39) is finite. Thus, for every $\lambda\geq 2$, the map $\Phi\mapsto \mathrm{AF}^\lambda(\phi)$ is continuous from $W_w^{p,k}(I\times\mathbb{R}^d\times\mathcal{U})$ to $[0,\infty)$. Furthermore, since the loss-function ℓ is continuous, then, for every $\lambda\geq 2$ the function $\Phi\mapsto\ell(\varphi-\phi)+\mathrm{AF}^\lambda(\phi)$ is continuous. Furthermore, since both ℓ and AF^λ are bounded below by 0 and finite-valued then so is $\ell(\varphi-\cdot)+\mathrm{AF}^\lambda(\cdot)$. Lastly, since ℓ is coercive then by definition, for every $r\geq 0$, there exists a compact subset $K_r\subseteq\mathscr{H}$ satisfying

$$\{\phi\in\mathscr{H}:\ \ell(\varphi-\phi)\leq k\}\subseteq K_r. \tag{A41}$$

Therefore, the non-negativity of each AF^λ implies that for every $\lambda \geq 2$ and every $r \geq 0$,

$$\left\{\phi \in \mathscr{H} : \ell(\varphi - \phi) + \mathrm{AF}^\lambda(\phi) \leq k\right\} \subseteq \{\phi \in \mathscr{H} : \ell(\varphi - \phi) \leq k\} \subseteq K_r; \tag{A42}$$

thus (A42) implies that $\phi \mapsto \ell(\varphi - \cdot) + \mathrm{AF}^\lambda(\cdot)$ is coercive. Thus, for every $\lambda \geq 2$, the function $\phi \mapsto \ell(\varphi - \cdot) + \mathrm{AF}^\lambda(\cdot)$ is lower semi-continuous, bounded-below, proper, and coercive on $\mathscr{H}$; thus by Weirestrass's Theorem, (Focardi 2012, Theorem 2.2), it admits a minimizer on $\mathscr{H}$. □

References

Bachelier, L.. 1900. Théorie de la spéculation. *Ann. Sci. École Norm. Sup. (3) 17*, 21–86.

Bain, Alan and Dan Crisan. 2009. *Fundamentals of stochastic filtering*, Volume 60 of *Stochastic Modelling and Applied Probability*. Springer, New York.

Björk, Tomas. 2009. *Arbitrage theory in continuous time*. Oxford university press.

Björk, Tomas and Bent Jesper Christensen. 1999. Interest rate dynamics and consistent forward rate curves. *Math. Finance 9*(4), 323–348.

Braides, Andrea. 2002. Γ*-convergence for beginners*, Volume 22 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, Oxford.

Brown, R. C. and B. Opic. 1992. Embeddings of weighted Sobolev spaces into spaces of continuous functions. *Proc. Roy. Soc. London Ser. A 439*(1906), 279–296.

Carmona, René. 2014. *Statistical analysis of financial data in R* (Second ed.). Springer Texts in Statistics. Springer, New York.

Christensen, Jens H. E., Francis X. Diebold, and Glenn D. Rudebusch. 2011. The affine arbitrage-free class of Nelson-Siegel term structure models. *J. Econometrics 164*(1), 4–20.

Cont, Rama and David-Antoine Fournié. 2013. Functional Itô calculus and stochastic integral representation of martingales. *Ann. Probab. 41*.

Cuchiero, Christa. 2011. *Affine and polynomial processes*. Ph. D. thesis, ETH Zurich.

Cuchiero, C., I. Klein, and J. Teichmann. 2016. A new perspective on the fundamental theorem of asset pricing for large financial markets. *Theory Probab. Appl. 60*(4), 561–579.

Dal Maso, Gianni. 1993. *An introduction to* Γ*-convergence*, Volume 8 of *Progress in Nonlinear Differential Equations and their Applications*. Birkhäuser Boston, Inc., Boston, MA.

De Giorgi, Ennio. 1975. Sulla convergenza di alcune successioni d'integrali del tipo dell'area. *Rend. Mat. (6) 8*, 277–294. Collection of articles dedicated to Mauro Picone on the occasion of his ninetieth birthday.

Delbaen, Freddy and Walter Schachermayer. 1994. A general version of the fundamental theorem of asset pricing. *Math. Ann. 300*(3), 463–520.

Delbaen, F. and W. Schachermayer. 1998. The fundamental theorem of asset pricing for unbounded stochastic processes. *Math. Ann. 312*(2), 215–250.

Devin, Siobhán, Bernard Hanzon, and Thomas Ribarits. 2010. A Finite-Dimensional HJM Model: How Important is Arbitrage-Free Evolution? *Int. J. Theor. Appl. Finance 13*(8), 1241–1263.

Diebold, Francis X. and Glenn D. Rudebusch. 2013. *Yield curve modeling and forecasting*. Princeton University Press, Princeton, NJ.

Dupire, Bruno. 2009. Functional Itô calculus. *Bloomberg Portfolio Research Paper No. 2009-04-FRONTIERS*.

Elworthy, K. D.. 1982. *Stochastic differential equations on manifolds*, Volume 70 of *London Mathematical Society Lecture Note Series*. Cambridge University Press, Cambridge-New York.

Filipović, Damir. 2000. Exponential-polynomial families and the term structure of interest rates. *Bernoulli 6*(6), 1081–1107.

Filipović, Damir. 2001. *Consistency Problems for Heath-Jarrow-Morton Interest Rate Models*, Volume 1760 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin.

Filipović, Damir. 2009. *Term-structure models*. Springer Finance. Springer-Verlag, Berlin. A graduate course.

Filipović, Damir, Stefan Tappe, and Josef Teichmann. 2010. Term structure models driven by Wiener processes and Poisson measures: existence and positivity. *SIAM J. Financial Math. 1*(1), 523–554.

Filipović, Damir and Josef Teichmann. 2004. On the geometry of the term structure of interest rates. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, Volume 460, pp. 129–167. The Royal Society.

Focardi, Matteo. 2012. Γ-convergence: a tool to investigate physical phenomena across scales. *Math. Methods Appl. Sci. 35*(14), 1613–1658. doi:10.1002/mma.2551.

Fontana, Claudio. 2014. A note on arbitrage, approximate arbitrage and the fundamental theorem of asset pricing. *Stochastics An International Journal of Probability and Stochastic Processes 86*(6), 922–931.

Fournie, David-Antoine. 2010. *Functional Ito Calculus and Applications*. Ph. D. thesis. Thesis (Ph.D.)–Columbia University.

Gelenbe, Erol. 1989. Random neural networks with negative and positive signals and product form solution. *Neural computation 1*(4), 502–510.

Guasoni, Paolo. 2006. No arbitrage under transaction costs, with fractional brownian motion and beyond. *Mathematical Finance 16*(3), 569–582.

Harrison, J. Michael and David M. Kreps. 1979. Martingales and arbitrage in multiperiod securities markets. *J. Econom. Theory 20*(3), 381–408.

Hastie, Trevor, Robert Tibshirani, and Martin Wainwright. 2015. *Statistical learning with sparsity: the LASSO and generalizations*. CRC Press.

Heath, David, Robert Jarrow, and Andrew Morton. 1992. Bond Pricing and the Term Structure of Interest Rates: A New Methodology for Contingent Claims Valuation. *Econometrica*, 77–105.

Hornik, Kurt. 1991. Approximation capabilities of multilayer feedforward networks. *Neural networks 4*(2), 251–257.

Jaeger, Herbert and Harald Haas. 2004. Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *Science 304*(5667), 78–80.

Jazaerli, Samy and Yuri F. Saporito. 2017. Functional Itô calculus, Path-Dependence and the Computation of Greeks. *Stochastic Process. Appl. 127*(12), 3997–4028.

Kratsios, Anastasis. 2019a. Deep arbitrage-free regularization. https://github.com/AnastasisKratsios/Deep_Arbitrage_Free_Regularization/.

Kratsios, Anastasis. 2019b. The universal approximation property: Characterizations, existence, and a canonical topology for deep-learning.

Kratsios, Anastasis. 2019. Universal approximation theorems. *arXiv Mathematics e-prints*.

Meucci, Attilio. 2005. *Risk and asset allocation*. Springer Finance. Springer-Verlag, Berlin.

Musiela, Marek. 1993. Stochastic PDEs and term structure models. *Journées Internationales de Finance*.

Nelson, Charles R and Andrew F Siegel. 1987. Parsimonious modeling of yield curves. *Journal of business*, 473–489.

Pang, Tao and Azmat Hussain. 2015. An application of functional ito's formula to stochastic portfolio optimization with bounded memory. In *2015 Proceedings of the Conference on Control and its Applications*, pp. 159–166. SIAM.

Rahimi, Ali and Benjamin Recht. 2008. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pp. 1177–1184.

Schweizer, Martin. 1995. On the minimal martingale measure and the Föllmer-Schweizer decomposition. *Stochastic Anal. Appl. 13*(5), 573–599.

Shreve, Steven E. 2004. *Stochastic calculus for finance II: Continuous-time models*, Volume 11. Springer Science & Business Media.

Tibshirani, Robert. 1996. Regression shrinkage and selection via the LASSO. *J. R. Stat. Soc. Series B. Stat. Methodol*, 267–288.

Zou, Hui, Trevor Hastie, and Robert Tibshirani. 2006. Sparse principal component analysis. *J. Comput. Graph. Statist. 15*(2), 265–286.

Sample Availability: Code is available at: "https://github.com/AnastasisKratsios"