AFNO: Universal Certifiably-Arbitrage-Free Operator Learning Framework for European Call-Price Surfaces

Anastasis Kratsios[3,4*] **Yannick Limmer**[5] **Hans Buehler**[1]
Blanka Horvath[1,2] **Raeid Saqur**[1,4]

[1]Mathematical Institute, University of Oxford
[2]Oxford-Man Institute for Quantitative Finance, University of Oxford
[3]Department of Mathematics, McMaster University
[4]Vector Institute
[5]DRW

Abstract

Neural operators (NO) have been introduces as deep learning-powered solutions to the computational bottleneck of resolving families of problems; since, once trained an NO can rapidly output approximate solutions, thus avoiding the need to re-run classical solvers. While standard NO architectures resolve this computation trouble they are inadequate for predicting European call-price surfaces from market or model features cannot guarantee static no-arbitrage conditions and, as we verify experimentally, can yield negative PnL in downstream trading tasks.

We introduce a certifiably arbitrage-free (AF) and universal encoder-decoder NO architectural framework—*continuously* mapping inputs from a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ over any fixed subset of $[0,\infty)^2$ to a convex combination of *arbitrage-free* European call-price surfaces; by continuously resolving an approximate nearest neighbour problem. We show that a nearest arbitrage-free correction of an unconstrained predicted surface need not exist, implying that no-arbitrage cannot, in general, be restored through post-processing. Further, we show that standard NOs such as the FNO and certain encoder-decoder NOs admit parameter configurations whose outputs violate static no-arbitrage. Thus, out-of-the-box NOs, in general, cannot be certifiably restricted to arbitrage-free configurations, but may only be guided towards them through PINN-type regularization penalization during training.

Keywords: Certifiable Machine Learning, European Call Price Surfaces, Operator Learning, No Static Arbitrage, Universal Approximation, Conformal Prediction.

1 Arbitrage-Free Operator Learning

In computational mathematics, eg. computational finance, one often must resolve *large families* of problem *fast*. *Neural operators* (NO) offer a way out to the computational burden involved in resolving large families of such problems, wherein classical numerical solver can become too costly to re-run several times (even if individual runs are fast, their totality may become unrealistic). NOs however, can be nearly instantaneously queried, after one costly training round, offering rapid approximate solutions to massive numbers of classical computational problems. Briefly, the idea behind NOs is to view the relevant families of problems as input-outputs of some (nonlinear) *operator* between infinite-dimensional topological vector spaces; often Hilbert spaces.

Many problems in mathematical finance are naturally formulated as nonlinear operators between infinite-dimensional Hilbert spaces. Operator learning extends classical deep learning to this setting, replacing repeated numerical solution by the rapid evaluation of a learned map. Applications include simultaneously solving families of SDEs and BSDEs [GLK22, FK24, FKPR25], modelling term structures and volatility surfaces [BDL21, BDG24, GT21, WJG25a, CJB24, NJZB23], learning pricing and optimal-stopping maps [CLZZ23, BFZZ25], solving stochastic games and execution problems [AEKY25, FKY25, MVD+26], and learning risk measures or transition kernels [DLDN26,

*Corresponding author.

AKP24] both of which obtain exceptionally-accurate fits discouraging static arbitrage but not exactly prohibiting it.

For financial prediction, however, numerical accuracy alone is not sufficient; as the NO predictions must also satisfy the structural conditions required by economic theory. In particular, a predicted European call-price surface may be economically meaningless if it violates static no-arbitrage, even when its pointwise prediction error is small, and defective predictions may be erroneously exploited by downstream AI agents which converge to trading strategies generating negative PnL (cf. Figure 8 below). This is problematic since standard neural-operator architectures do not enforce these conditions uniformly across all parameter configurations; and the state-of-the-art at-best guides it towards no arbitrage by PINN-type penalization; e.g. [WJG25b] for neural operator-based and [CJB24, NJZB23] for neural SDE-based solutions.

This paper addresses this by introducing an *arbitrage-free neural operator* (AFNO): an encoder-decoder neural operator whose decoder maps directly into the nonlinear space of arbitrage-free European call-price surfaces. The resulting architecture is certifiable, since every parameter configuration produces an arbitrage-free surface, and universal, since it can approximate every Lipschitz operator with the corresponding domain and codomain to arbitrary precision. Thus, structural certification is achieved without sacrificing expressivity by building on the geometric deep learning results of [KZLD22, KLL+23b] which can handle approximation into non-linear co-domains such as the space of arbitrage-free European call-price surfaces.

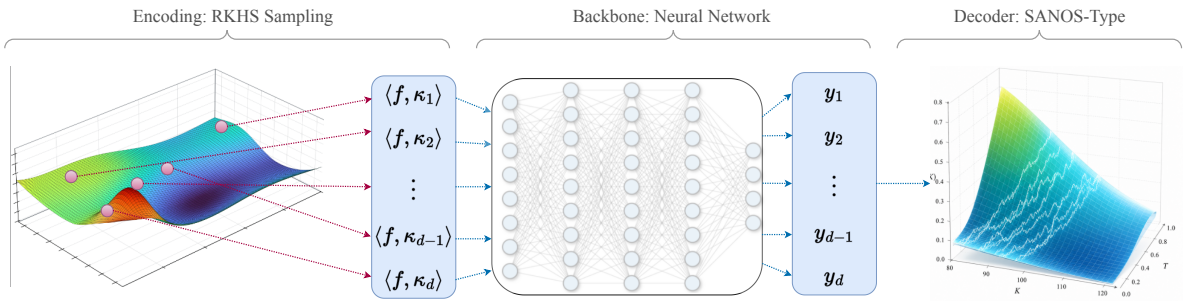


Figure 1: **AFNO architecture:** During the *encoding phase* (cf. § 3.1), input features are encoded from an RKHS over some $\mathcal{D} \subseteq [0, \infty)^2$ with kernel κ by projecting them onto a finite collection of, possibly trainable, basis vectors or sampling functionals. The resulting finite-dimensional representation is then passed through a transformer *backbone* (cf. § 3.2); we use a ReLU-MLP, since this class is analytically well understood. Finally, during the *decoding phase* (cf. § 3.3), the backbone output is mapped into the space of arbitrage-free volatility surfaces using any decoder satisfying the structural properties of the SANOS model; cf. [BHK+26]. Here, $\kappa_n \overset{\text{def.}}{=} \kappa(\cdot, x_n)$ for some *sample points* $x_1, \ldots, x_d \in \mathcal{D}$. Thus, $\langle f, \kappa_n \rangle$ represents "smoothed sampling".

Our construction, summarized by Figure 1, parallels the encoder-decoder NO paradigm cf. [KLL+23a, KLM21, GLK22] encompassing PCA-Net [BHKS21, Lan23], DeepONet [LJP+21], but not the Fourier neural operator (FNO) [LKA+21]. Rather, we use a new formulation of the geometric attention mechanism of [KZLD22, AKP24], inspired by the classical attention mechanism of transformers [VSP+17], but tailored to the space of arbitrage-free European call-price surfaces. The main example of our decoder is the SANOS model, recently introduced in [BHK+26].

1.1 Contributions

Our contributions are both positive and negative.

1.1.1 Primary Contribution

On the *positive* side, we prove that every parameter configuration of an AFNO produces an arbitrage-free European call-price surface with controlled right-wing decay (Theorem 4.1). Moreover, AFNOs are universal: every Lipschitz map from $\mathbb{R}^d$ into the corresponding class of arbitrage-free call

surfaces can be approximated to arbitrary precision (Theorem C.11). In particular, this applies to parameter-to-surface maps generated by parametric families of SDEs (Corollary 4.5).

Why this is challenging. Existing universal approximation theorems for neural operators treat operators taking values in linear spaces, typically Banach spaces [Lan23, LLS25, SSZ26, ST26] or, more generally, Fréchet spaces [GLK22, BDG23]. These results do not enforce membership in nonlinear, nonclosed subsets of such spaces. Consequently, they cannot simultaneously guarantee universal approximation of maps into the space of arbitrage-free call surfaces $\mathcal{C}\mathrm{AF}$ and certify that every output of the approximating neural operator remains in $\mathcal{C}\mathrm{AF}$.

A SANOS-type decoder approximates individual arbitrage-free surfaces. Our contribution is to construct one continuous *rapid to evaluate* input-to-surface map that approximates a target operator uniformly on a compact input set, under the hypotheses of Theorem 4.2, while remaining arbitrage-free at every input. In this way, one does not need to re-fit a surface to any dataset (either real-world or synthetic); e.g. avoiding re-running a linear program as with [BHK+26] or calibrating an SDE-type model, but they simply can simply evaluate a forward pass of our AFNO model.

1.1.2 Secondary Contributions

On the *negative* side, a predicted surface need not admit a nearest arbitrage-free approximation (Theorem 5.1); thus, no-arbitrage cannot, in general, be restored by post-processing. Moreover, standard encoder-decoder neural operators and Fourier neural operators admit parameter configurations whose outputs violate static no-arbitrage (Propositions 5.2 and 5.4). Penalty-based methods may therefore discourage arbitrage during training, but cannot certify its absence. This motivates enforcing no-arbitrage structurally, at the level of the architecture.

Why this is important. These results distinguish empirical constraint enforcement from architectural certification. Penalties may reduce arbitrage violations, and post-processing may repair particular outputs, but neither approach, by itself, guarantees that the resulting operator maps every input into the space of arbitrage-free call surfaces. This motivates enforcing no-arbitrage directly through the architecture. Practically, our experiments demonstrate the consequences of forgoing such a certificate: a downstream learning agent may focus on spurious arbitrage opportunities created by defects in the predicted call-price surfaces rather than learn genuinely profitable trading strategies.

1.2 Organization of the Paper

Section 2 introduces the input and output spaces. Section 3 develops the AFNO framework and its arbitrage-free decoders. Sections 4 and 5 present our positive and negative theoretical results, respectively. Section 6 provides numerical illustrations, while the supplementary diagnostics and proofs are deferred to the appendices.

2 Preliminaries

We now overview the necessary background and elementary preliminary findings required to formulate our main results.

2.1 Input Spaces: Reproducing Kernel Hilbert Spaces with Riesz Bases

In practice, market data rarely arrives on a fixed and perfectly regular grid. Available call-price quotes may be irregularly distributed across strikes and maturities, their number may change from one observation time to another, and in many applications one may wish to work not only with individual quotes but also with interpolated slices or entire observed surfaces. It is therefore useful to work with an input space which is flexible enough to accommodate all of these situations within a single framework.

Reproducing kernel Hilbert spaces provide a natural setting for this purpose. Point evaluations are continuous, so individual market quotes can be used directly as features, while suitable sampling and interpolation properties allow finite collections of such observations to represent increasingly rich functional inputs. At the same time, the infinite-dimensional formulation allows us to regard a complete call-price slice or surface as a single input object. The sampling and interpolation theory introduced below is precisely what allows us to move consistently between these finite observations and their underlying functional representation.

We briefly review the relevant theory of reproducing kernel Hilbert spaces; which have already been extensively used in mathematical finance [Fil01, Fil09, BK23, Aro50, BTA04, Wen05] and implicitly via Gaussian process regression techniques; e.g. [Lud18, LR25]. Let $\mathcal{D} \subseteq [0,\infty)^2$ be non-empty, and let $\mathcal{H}$ be a reproducing kernel Hilbert space (RKHS) over $\mathcal{D}$, with reproducing kernel $\kappa : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$; cf. [Aro50]. A countable sequence $\{z_\lambda\}_{\lambda\in\Lambda} \subset \mathcal{D}$, indexed by $\Lambda \subseteq \mathbb{N}$, is called a *sampling sequence* for $\mathcal{H}$ if there exists a constant $C \geq 1$ such that, for every $f \in \mathcal{H}$,

$$\frac{1}{C}\,\|f\|_{\mathcal{H}} \leq \|f\|_{\Lambda} \overset{\text{def.}}{=} \left(\sum_{\lambda\in\Lambda} |f(z_\lambda)|^2\right)^{1/2} \leq C\,\|f\|_{\mathcal{H}}. \tag{1}$$

It is called an *interpolating sequence* if, for every square-summable sequence $v. \overset{\text{def.}}{=} (v_\lambda)_{\lambda\in\Lambda} \in \ell^2(\Lambda)$, there exists an $f \in \mathcal{H}$ such that: for each $\lambda \in \Lambda$ we have

$$f(z_\lambda) = v_\lambda. \tag{2}$$

If $\{z_\lambda\}_{\lambda\in\Lambda}$ is both sampling and interpolating, then we call it a *complete interpolating sampling sequence* (CIS); since that will allow us the feature maps $\{\langle \kappa(z,\cdot),\cdot,\rangle : z \in \mathcal{D}\}$ to be trainable (in z) set of continuous features which can encode a Riesz basis for $\mathcal{H}$. Throughout the sequel, we assume that $\mathcal{H}$ admits a CIS.

Riesz bases of reproducing kernels, sampling, and interpolation in Fock-type RKHSs have been studied in depth; see, for example, [BL10] and [BDHK15]. In statistics they are usually referred to as kernel-regression. Landau-type necessary density conditions in general reproducing kernel Hilbert spaces were developed [FGH$^+$17], extending the classical density results of [Lan67].

As a first example we demonstrate sampling and interpolating sequences in the context of quotes on a call price surface. Though samples (quotes) don't arrive in strictly ordered grids, many popular modelling techniques [BR15, BR18] rely on grid-based sampling of surfaces, which makes them our first relevant example:

Example 2.1 (Finite-Dimensional RKHS: Evaluations on a Fixed Grid on the Call Surface)**.** Let $d \in \mathbb{N}_+$ and let $\mathcal{D} = \{x_n\}_{n=1}^d = \{(K_n, T_n)\}_{n=1}^d \subset [0,\infty)^2$. Since $\mathcal{D}$ is finite, a kernel on $\mathcal{D}$ is equivalently a *fixed* $d \times d$ symmetric positive semidefinite matrix $\mathbf{K} = \big(\kappa(x_i,x_j)\big)_{i,j=1}^d$. If $\mathbf{K}$ is strictly positive definite, then the associated RKHS $\mathcal{H}$ is d-dimensional and consists of *all* functions $f : \mathcal{D} \to \mathbb{R}$. Intuitively, this RKHS consists of all call and put features observed on a finite pre-determined set of points, e.g. possibly but not necessarily a finite grid. Moreover, every $f \in \mathcal{H}$ has a unique representation

$$f = \sum_{n=1}^{d} \beta_n\, \kappa(\cdot, x_n) \tag{3}$$

and the map sending any such f to its vector of coefficients $(\beta_n)_{n=1}^d \in \mathbb{R}^d$ is a vector-space isomorphism given by the generalized regression formula

$$\beta = (\kappa(x_i,x_j)_{i,j})^{-1}(f(x_1,\dots,x_d))^T\ .$$

Thus, one may directly plug in the output of the feature map (3) into a downstream standard deep learning model; e.g. MLP, GNN, or a transformer.
If $f = \sum_{n=1}^d \beta_n \kappa(\cdot,x_n)$ and $g = \sum_{n=1}^d \gamma_n \kappa(\cdot,x_n)$, then the RKHS inner product on $\mathcal{H}$ is given by

$$\langle f, g\rangle_{\mathcal{H}} = \beta^\top \mathbf{K} \gamma.$$

Thus, the isomorphism mapping any f in (3) to $\beta = (\beta_n)_{n=1}^d \in \mathbb{R}^d$ is bi-Lipschitz, and it is an isometry in the special case where $\mathbf{K} = I_d$. ◁

The next example of an RKHS abstracts the idea of having an arbitrarily large, but finite, number of quoted call prices at our disposal. This removes the fixed finite-grid construction of Example 2.1 by extending that example's simple RKHS, which was effectively $\mathbb{R}^d$, to a genuinely infinite-dimensional RKHS. Previously, d represented the fixed number of quotes that could be observed at any given time in the market; in the next example, the infinite-dimensionality instead provides the flexibility to accommodate an arbitrary finite number of quoted prices. Moreover, the infinite-dimensional setting allows one to regard the entire continuous call surface as a feature, rather than merely finitely many quoted prices[1].

We present this RKHS in two steps. First we introduce a slice-wise RKHS and then we pass to one which can encode both strike and maturity data via tensorization of the following example:

Example 2.2 (An RKHS for European Call-price slices)**.** Let $\mathcal{H}_0$ denote the anchored Sobolev-type RKHS on $(0,\infty)$ with kernel $\kappa_0 : (0,\infty)^2 \to \mathbb{R}$ given for every $x, y > 0$ by

$$\kappa_0(x,y) = x \wedge y, \tag{4}$$

and, as in [SS16, Theorem 1.6], this kernel induces the following RKHS on $(0,\infty)$

$$\mathcal{H}_0 = \left\{h \in AC_{\mathrm{loc}}([0,\infty)) : h(0) = 0,\ h' \in L^2(0,\infty)\right\} \quad \text{normed by} \quad \|h\|^2_{\mathcal{H}_0} = \int_0^\infty |h'(u)|^2\, du.$$

To impose the normalized boundary condition $f(0) = 1$, we use the affine translate

$$\mathcal{A}_1 = 1 + \mathcal{H}_0 = \{f : [0,\infty) \to \mathbb{R} : f - 1 \in \mathcal{H}_0\}. \tag{5}$$

Then every $f \in \mathcal{A}_1$ satisfies $f(0) = 1$. The set $\mathcal{A}_1$ is not a linear RKHS, but it is an affine Hilbert space canonically modelled on the RKHS $\mathcal{H}_0$. ◁

To consider European call-price *surfaces*, rather than only one-dimensional slices, as inputs, a convenient construction is to take the Hilbert tensor product of the RKHS from the preceding example with itself. By the tensor-product theorem for RKHSs [SS16, Theorem 2.20], this again yields an RKHS, now over $[0,\infty)^2$, whose kernel is the product of the kernels in (4).

Example 2.3 (Tensor-product anchored Sobolev RKHS on the call surface domain)**.** In the setting of Example 2.2, consider $\mathcal{D} = [0,\infty)^2$. Define the tensor-product kernel $\kappa_\otimes : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ for each $(K_1,T_1),(K_2,T_2) \in [0,\infty)^2$ by

$$\kappa_\otimes\big((K_1,T_1),(K_2,T_2)\big) = (K_1 \wedge K_2)\,(T_1 \wedge T_2),. \tag{6}$$

Then $\kappa_\otimes$ is positive definite and its associated RKHS is the tensor product $\mathcal{H}_\otimes = \mathcal{H}_0 \widehat{\otimes} \mathcal{H}_0$; where $\mathcal{H}_0$ is as in Example 2.2. Thus, $\mathcal{H}_\otimes$ is the mixed-smoothness Sobolev-type RKHS

$$\mathcal{H}_\otimes = \left\{ f : \mathcal{D} \to \mathbb{R} : \text{there exists } g \in L^2(\mathcal{D}) \text{ such that } f(K,T) = \int_0^K \int_0^T g(k,t)\, dt\, dk \right\}.$$

Every $f \in \mathcal{H}_\otimes$ satisfies the anchored boundary conditions: $f(K,0) = 0$ and $f(0,T) = 0$ for each $K, T \geq 0$. The inner product for any $f, g \in \mathcal{H}$ is

$$\langle f, h\rangle_{\mathcal{H}_\otimes} = \int_0^\infty \int_0^\infty \partial_K \partial_T f(K,T)\, \partial_K \partial_T h(K,T)\, dT\, dK \qquad \triangleleft.$$

Having established a convenient unified framework for representing both finitely many quotes and the entire call surface as features to be used as inputs to our deep learning models, we are now ready to turn to the structure of the relevant output space; namely, the space of arbitrage-free call surfaces.

[1] Interestingly, a variant classical RKHS considered for studying forward rate curves, in term structure modelling cf. [Fil01], can be used to obtain a natural RKHS backdrop for European call surface-price modelling.

2.2 Output Spaces: Spaces of Arbitrage-Free Call Surfaces

We now discuss the space of arbitrage-free call surfaces. While input spaces may vary, one point on which there is broad agreement is the value of generating surface-valued outputs that are guaranteed to be arbitrage-free by construction.

While the broad structural properties of arbitrage-free European call-price surfaces are by now well understood in mathematical finance, the space of all such objects is surprisingly less well studied. This is likely because it does not arise naturally in many classical problem formulations in mathematical finance and, moreover, because it is not a particularly tame space. We begin with the following classical result.

Definition 2.4 (Arbitrage-Free Call-Price Surfaces)**.** We call a function $C : [0,\infty)^2 \longrightarrow [0,\infty)$ an *arbitrage-free European call-price surface* if there exist a probability measure $\mathbb{P}$ and a non-negative $\mathbb{P}$-martingale $Z_\cdot \overset{\text{def.}}{=} (Z_t)_{t\in[0,\infty)}$ satisfying $Z_0 = 1$ $\mathbb{P}$-a.s. such that: for every $T.K \geq 0$

$$C(T,K) = \mathbb{E}_{\mathbb{P}}\left[(Z_T - K)^+\right].$$

In this case, $Z_\cdot$ is said to be a *martingale representation* of C.

It is convenient to give an analytic characterization of the arbitrage-free call-price surfaces introduced in Definition 2.4.

Proposition 2.5 (Analytic Characterization of Arbitrage-Free Call Surfaces)**.** *A continuous map $C : [0,\infty)^2 \to [0,\infty)$ is an arbitrage-free call surface (cf. Definition 2.4) if and only if:*

(i) $C(T,0) = 1$ *for every* $T \geq 0$.

(ii) $\partial_K^+ C(T,0) \overset{\text{def.}}{=} \lim_{\delta\downarrow 0} \frac{C(T,\delta)-C(T,0)}{\delta} \geq -1$ *for every* $T \geq 0$.

(iii) $\lim_{K\uparrow\infty} C(T,K) = 0$ *for every* $T \geq 0$.

(iv) For every $T \geq 0$, *the map* $K \mapsto C(T,K)$ *is convex.*

(v) For every $K \geq 0$, *the map* $T \mapsto C(T,K)$ *is non-decreasing.*

(vi) For every $K \geq 0$, $C(0,K) = (1-K)^+$.

If C satisfies (i)–(vi) *then, for each* $T \geq 0$, *every martingale representation $Z_\cdot$ of C satisfies*

$$\partial_K^+ C(T,0) = -1 + \mathbb{P}\left(Z_T = 0\right). \tag{7}$$

Proof. See Supplement S1, for a proof of this standard result. □

Remark 2.6 (Default-Free Models). Let C be an arbitrage-free call surface and let $Z_\cdot$ be any martingale representation of C. Then, for every fixed $T \geq 0$, (7) implies that

$$\partial_K^+ C(T,0) = -1 \text{ if and only if } Z_T > 0 \quad \mathbb{P}\text{-}a.s. \qquad \triangleleft($$

Let $\mathcal{C}_{\text{AF}} \subseteq C([0,\infty)^2, [0,\infty))$ consist of all continuous functions satisfying Definition 2.4; i.e. the space of **arbitrage-free call surfaces**. Since $\mathcal{C}_{\text{AF}}$ does not contain zero it is by no means a vector space (at-least under the familiar pointwise operation on functions), nor is it even a group under addition; moreover, the set has a natural attainable lower bound $(1-K)^+$ but no closed upper bound. Thus, it will be far beyond the limited reach of operator learning [LSYZ21, LKA+21, Lan23, LS23][2]. This means, we will have to rely on the more sophisticated tools of infinite-dimensional *geometric deep learning*; e.g. [KZLD22, KLL+23c, AKP24].

Proposition 2.7 (Arbitrage-Free Call Surfaces are Not A Vector Space)**.** *The space $\mathcal{C}_{\text{AF}}$ is not closed under addition, scalar multiplication, nor under pointwise products.*

[2]Whose objective is to learn, or at-times approximate, non-linear operators between separable topological vector spaces; e.g. Banach spaces.

The class $\mathcal{C}_{\mathrm{AF}}$ is convex but is not closed for uniform convergence on compact subsets of $[0,\infty)^2$; see Proposition C.13[3]. This is important, as convexity sets up the approximately simplicial property, cf. [AKP24, Definition 3.1 and Example 5.1] and first part of a QAS space [AKP24, Definition 3.4], required for that paper's main "static" approximation theorem in non-linear spaces. Which, as far as we know is the only available universal approximation theorem in non-linear space holding in that level of generality.

Proposition 2.8. *The space $\mathcal{C}_{\mathrm{AF}}$ is convex in $C([0,\infty)^2,\mathbb{R})$.*

It is already clear that one must be careful here, since in the topology of uniform convergence on compact sets the set $\mathcal{C}_{AF}$ **fails** to be closed. This is problematic, since limit points (and thus approximation) may not stay within $\mathcal{C}_{AF}$; cf. Proposition C.13 in Appendix C.5. However, closedness is not needed for QAS spaces as completeness is "local" on the range of the target function's image of a compact set of inputs; which is itself compact, and thus complete, by continuity of the target function. Both the lack of closeness, and the lack of linearity, means that the classical "universal" approximation theorems for neural operators fail.

2.2.1 Spaces of Call Surfaces with Uniform Right-Wing and Maturity Decay-Rate

As a prototype, we ask: what is the *right-wing decay* of the Black-Scholes formula? For any given maturity $T>0$, it is given by the following. Fix $T>0$ and let

$$C_{\mathrm{BS}}(T,K) \overset{\text{def.}}{=} e^{-rT}\,\mathbb{E}\big[(S_T-K)^+\big] \tag{8}$$

be the Black-Scholes call price with strike $K>0$, where $S \overset{\text{def.}}{=} (S_t)_{t\geq 0}$ follows a geometric Brownian motion with log-mean $\mu\in\mathbb{R}$ and log-volatility $\sigma>0$; i.e.

$$S_T = S_0 \exp\Big(\big(r-\tfrac{1}{2}\sigma^2\big)T+\sigma W_T\Big).$$

We thus have the basic right-wing estimate, which shows that for any fixed maturity the right wing $K\mapsto C(T,K)$ decays faster than any power of the reciprocal strike K^{-1}. It is straightforward to see that: for every $T,p>0$ there is a constant $C_{T,p}>0$ such that for each strike $K>0$ the Black-Scholes call price decays like $C_{\mathrm{BS}}(T,K)\leq \frac{C_{T,p}}{(1+K)^p}$. This motivates the following power-law decay rate condition: For a fixed $p>0$ we write $\mathcal{C}^p_{\mathrm{AF}}\subseteq\mathcal{C}_{\mathrm{AF}}$ to consist of all arbitrage-free call surfaces satisfying the following right-wing asymptotic: for every $T\geq 0$, there exists a constant $C_{T,p}>0$ such that

$$C(T,K)\leq C_{T,p}\,K^{-p}. \tag{9}$$

We will consider several examples of families of arbitrage-free call surfaces satisfying this decay condition in § 2.2.2, below. Let $\mathcal{C}_{AF}\subseteq C([0,\infty)^2,[0,\infty))$ denote the *metric space* of arbitrage-free call surfaces equipped with the topology of *uniform* convergence (not uniform convergence on compact sets). Fix $p,\lambda>0$ and consider the norm[4] $\|C\|_{p,\lambda}\overset{\text{def.}}{=}\sup_{T,K\geq 0} e^{-\lambda T}(1+K)^p|C(T,K)|$ on

$$\mathcal{X}^p_\lambda \overset{\text{def.}}{=} \Big\{C\in C([0,\infty)^2,\mathbb{R}):\ \|C\|_{p,\lambda}<\infty\Big\}.$$

Then every $C\in\mathcal{X}^p_\lambda$ automatically satisfies the weighted right-wing bound $C(T,K)\leq e^{\lambda T}\|C\|_{p,\lambda}(1+K)^{-p}$ for all $T,K\geq 0$. In particular, $\mathcal{C}_{AF}\cap\mathcal{X}^p_\lambda\subseteq\mathcal{C}^p_{AF}$.

We will typically consider "uniformly controlled" classes of call surfaces, if it has uniform right-wing decay rate, uniform maturity growth rate, and it satisfies "compactness-type conditions" akin to those in the Arzela-Ascoli theorems; cf. [CZ07, Xia09]. This is formalized as follows:

Definition 2.9 (Uniformly Controlled Call-Surface Class)**.** A class of (arbitrage-free) call surfaces $\mathcal{Y}\subseteq\mathcal{C}^{p,\lambda}_{AF}\overset{\text{def.}}{=}\mathcal{C}_{AF}\cap\mathcal{X}^p_\lambda$ is called uniformly controlled if there exists a constant $B>0$ and a modulus of continuity ω such that

[3] By contrast, it is closed in the (global) uniform norm. Indeed, if $C_n\in\mathcal{C}_{\mathrm{AF}}$ and $\|C_n-C\|_\infty\to 0$, then continuity, non-negativity, the boundary values, strike convexity, maturity monotonicity, and the strike Lipschitz bound pass to C. For each fixed $T\geq 0$, $0\leq \limsup_{K\to\infty} C(T,K)\leq\|C-C_n\|_\infty$, so the right-wing limit is zero. Proposition 2.5 yields $C\in\mathcal{C}_{\mathrm{AF}}$.

[4] See [ST26] for results on weighted spaces.

(a) **Uniform weighted bound:** $\sup_{C\in\mathcal{Y}} \|C\|_{p,\lambda} \le B$,

(b) **Equi Uniform Continuity:** $\mathcal{Y}$ is ω-equi-uniformly continuous; i.e. for all $T, \tilde{T}, K, \tilde{K} \ge 0$

$$|C(T,K) - C(\tilde{T},\tilde{K})| \le \omega\big(|T-\tilde{T}| + |K-\tilde{K}|\big).$$

When needed, we call (p, λ, ω, B) the regularity-data of $\mathcal{Y}$.

2.2.2 Examples Families of Surfaces Satisfying the Decay Condition

We begin with the Black-Scholes prototype example.

Example 2.10 (Black–Scholes Call Surfaces)**.** Let $Z_t \overset{\text{def.}}{=} \exp\big(\sigma W_t - \frac{1}{2}\sigma^2 t\big)$ for some $\sigma > 0$, and define the corresponding normalized Black–Scholes call surface by $C_{\mathrm{BS}}(T,K) \overset{\text{def.}}{=} \mathbb{E}[(Z_T - K)_+]$. For $T > 0$ and $K > 0$, the Black–Scholes formula gives

$$C_{\mathrm{BS}}(T,K) = \mathcal{N}(d_+(T,K)) - K\mathcal{N}(d_-(T,K)),$$

where $d_\pm(T,K) \overset{\text{def.}}{=} -\frac{\log K}{\sigma\sqrt{T}} \pm \frac{\sigma\sqrt{T}}{2}$. In particular, $C_{\mathrm{BS}}(T,K) \le \mathcal{N}(d_+(T,K))$. Therefore, for any fixed $p > 0$. If $K \ge e^{\sigma^2 T/2}$, then $d_+(T,K) \le 0$ and the Gaussian tail bound $\mathcal{N}(-x) \le e^{-x^2/2}$ for $x \ge 0$ yields $C_{\mathrm{BS}}(T,K) \le \exp\left(-\frac{(\log K - \sigma^2 T/2)^2}{2\sigma^2 T}\right)$ and noting that $-\frac{(\log K - \sigma^2 T/2)^2}{2\sigma^2 T} \le -p\log K + \frac{p(p+1)}{2}\sigma^2 T$ is enough to deduce that

$$C_{\mathrm{BS}}(T,K) \le e^{\frac{1}{2}p(p+1)\sigma^2 T} K^{-p}.$$

If instead $0 < K < e^{\sigma^2 T/2}$, then $C_{\mathrm{BS}}(T,K) \le 1$ and $e^{\frac{1}{2}p(p+1)\sigma^2 T} K^{-p} \ge 1$. The same bound also holds at $T = 0$, since $C_{\mathrm{BS}}(0,K) = (1-K)_+ \le K^{-p}$. Therefore, for every $p > 0$, $C_{\mathrm{BS}} \in \mathcal{C}^p_{\mathrm{AF}}$ for every $p > 0$. ◁

We now provide examples of popular families of call-price surfaces satisfying condition (9).

Proposition 2.11 (Polynomial Decay of Call Payoffs)**.** *Let $S_\cdot \overset{\text{def.}}{=} (S_t)_{t\ge 0}$ be a real-valued stochastic process such that there exists a $q > 1$ satisfying $\mathbb{E}[|S_t|^q] < \infty$ for every $t \ge 0$. Then, for every $T \ge 0$ and $K > 0$,*

$$\mathbb{E}\big[(S_T - K)_+\big] \le \frac{(q-1)^{q-1}}{q^q}\,\mathbb{E}\big[(S_T)_+^q\big] K^{-(q-1)}. \tag{10}$$

In particular, (9) holds for every $T \ge 0$, with $p = q-1$ and $C_{T,p} = \frac{(q-1)^{q-1}}{q^q}\mathbb{E}[(S_T)_+^q]$.

Proof. See Section C.1. □

As a simple illustration, the Heston model[5] satisfies the sufficient condition of Proposition 2.11.

Example 2.12 (Lognormal Stochastic Volatility)**.** Let

$$dZ_t = Z_t\sigma_t\,dW_t, \qquad d\sigma_t = \eta\sigma_t\,dB_t, \qquad d\langle W,B\rangle_t = \rho\,dt,$$

where $\eta > 0$, $\sigma_0 > 0$, and $\rho \in (-1,0)$. By [LM07, Theorem 2.3], for every $q > 1$, $\sup_{t\in[0,T]}\mathbb{E}[Z_t^q] < \infty$ for every $T > 0$ if and only if $\rho \le -\sqrt{\frac{q-1}{q}}$; whence, for every $0 < p \le \frac{\rho^2}{1-\rho^2}$, if we set $q \overset{\text{def.}}{=} p+1$ then Proposition 2.11 implies that: for every $T \ge 0$ and $K > 0$

$$C(T,K) \overset{\text{def.}}{=} \mathbb{E}\big[(Z_T - K)_+\big] \le \frac{p^p}{(p+1)^{p+1}}\,\mathbb{E}\big[Z_T^{p+1}\big] K^{-p}.$$

Thus, (9) holds and $C \in \mathcal{C}^p_{\mathrm{AF}}$. ◁

[5]The following can be extended to the rough Heston model by [GGP19, Theorem 2.4].

Example 2.13 (Heston Call Surfaces)**.** Consider the discounted Heston model

$$dS_t = S_t\sqrt{V_t}\,dW_t, \qquad dV_t = \kappa(\theta - V_t)\,dt + \sigma\sqrt{V_t}\,dZ_t, \qquad d\langle W, Z\rangle_t = \rho\,dt,$$

where $\kappa, \theta, \sigma > 0$ and $\rho \in (-1, 1)$. If $\kappa - 3\rho\sigma \geq \sqrt{6}\,\sigma$, then the third-moment explosion time is infinite, and hence $\mathbb{E}[S_T^3] < \infty$ for every $T \geq 0$; cf. [GGP19, Equation (2.13) and case (C)]. Proposition 2.11, applied with $q = 3$, therefore gives

$$C_{\text{Heston}}(T, K) \overset{\text{def.}}{=} \mathbb{E}\big[(S_T - K)_+\big] \leq \frac{4}{27}\,\frac{\mathbb{E}[S_T^3]}{K^2},$$

for each $T \geq 0$ and $K > 0$. Thus, (9) holds with $p = 2$. ◁

Example 2.14 (Rough Bergomi)**.** Fix a Hurst parameter $H \in (0, \frac{1}{2})$ and a correlation $\rho \in (-1, 0)$. We consider the associated fractional rough Bergomi-type model in [BJ26, Equation (1.1)] with log-normal response $\sigma(t, x) = m(t)e^x$, where $m \in C^1_{\text{loc}}([0, \infty); (0, \infty))$, and fractional kernel $K_H(t) \overset{\text{def.}}{=} \frac{t^{H-\frac{1}{2}}}{\Gamma(H+\frac{1}{2})}$; that is, $S_\cdot \overset{\text{def.}}{=} (S_t)_{t\geq 0}$ is given by

$$S_t = S_0 + \int_0^t S_s m(s) e^{Y_s}\,dW_s, \qquad Y_t = y_0 + \nu\int_0^t K_H(t-s)\,dB_s, \qquad W_t = \rho B_t + \sqrt{1-\rho^2}\,B_t^{\perp}, \tag{11}$$

where $S_0 > 0$, $y_0 \in \mathbb{R}$, $\nu > 0$, and B and $B^{\perp}$ are independent Brownian motions. If K_H and $\sigma(t, x)$ satisfy the technical conditions in[BJ26, Assumptions 1 and 2] then [BJ26, Theorem 1] implies that $\mathbb{E}[S_T^q] < \infty$ for every $0 < q < \frac{1}{1-\rho^2}$ and for every $T > 0$. Consequently, for every

$$0 < p < \frac{\rho^2}{1-\rho^2}, \tag{12}$$

if we set $q = p + 1$ then Proposition 2.11 holds, which yields

$$C_{\text{rBergomi}}(T, K) \overset{\text{def.}}{=} \mathbb{E}\big[\max\{S_T - K, 0\}\big] \leq \frac{p^p}{(p+1)^{p+1}}\mathbb{E}[S_T^{p+1}]K^{-p}$$

implying that $C_{\text{rBergomi}} \in \mathcal{C}^p_{\text{AF}}$ for every $0 < p < \rho^2/(1-\rho^2)$. ◁

Example 2.15 (Exponential Lévy; CGMY)**.** Let $Z_T = e^{X_T}$ for a Lévy process X satisfying $\mathbb{E}[Z_T] = 1$, and consider the associated call surface $C(T, K) \overset{\text{def.}}{=} \mathbb{E}[(Z_T - K)_+]$. For the CGMY model [CGMY02], the Lévy measure has right tail $\nu(dx) = Ce^{-Mx}x^{-1-Y}\,dx$ for $x > 0$ *for some* $M > 1$. Hence, for every $q < M$, $\mathbb{E}[Z_T^q] = \mathbb{E}[e^{qX_T}] < \infty$ for every $T \geq 0$. Thus for any $0 < p < M - 1$, we may set $q \overset{\text{def.}}{=} p + 1$ and apply Proposition 2.11 to conclude that $C \in \mathcal{C}^p_{\text{AF}}$; again for every $0 < p < M - 1$. ◁

Examples in the SSVI Framework We note that condition (9) need *not* be directly interpreted and explicit moment condition. Rather, it is simply an analytic decay condition on the call surface. We illustrate it by showing that it holds for the SSVI model of [GJ13], for which a stochastic representation is not known to the authors.

Proposition 2.16 (SSVI Call Surfaces Satisfy (9) for every $p > 0$)**.** *Fix smooth functions $\varphi, \theta \in C((0, \infty), (0, \infty))$ with $\theta_\cdot$ increasing, and let $C_{\text{SSVI}} \in \mathcal{C}_{\text{AF}}$ be an arbitrage-free call surface induced by the SSVI total implied variance*

$$w(k, \theta_T) = \frac{\theta_T}{2}\left\{1 + \rho\varphi(\theta_T)k + \sqrt{\big(\varphi(\theta_T)k + \rho\big)^2 + 1 - \rho^2}\right\},$$

where $\rho \in (-1, 1)$. For any fixed $p > 0$, $C_{\text{SSVI}} \in \mathcal{C}^p_{\text{AF}}$ holds if

$$\sup_{T>0}\,\frac{1+\rho}{2}\theta_T\varphi(\theta_T) \leq 2\big(\sqrt{p+1} - \sqrt{p}\big)^2. \tag{13}$$

Proof. See Section C.1. □

Example 2.17 (Heston-like SSVI)**.** In the context of Proposition 2.16, consider the Heston-like SSVI model in [GJ14, Example 4.1]; i.e. let $\lambda > 0$ and for each $\theta > 0$

$$\varphi(\theta) \overset{\text{def.}}{=} \frac{1}{\lambda\theta}\left(1 - \frac{1 - e^{-\lambda\theta}}{\lambda\theta}\right).$$

Then $\theta \mapsto \theta\varphi(\theta)$ is increasing and $\sup_{\theta>0} \theta\varphi(\theta) = \lim_{\theta\to\infty} \theta\varphi(\theta) = \frac{1}{\lambda}$. Consequently, condition (13) holds if $\frac{1+\rho}{2}\frac{1}{\lambda} \leq 2(\sqrt{p+1} - \sqrt{p})^2$; or simply

$$\lambda \geq \frac{1+\rho}{4\left(\sqrt{p+1} - \sqrt{p}\right)^2}. \tag{14}$$

Note that, if $\rho \geq 0$ then (14) implies the no butterfly arbitrage condition in [GJ14, Theorem 4.2], namely $(1 + |\rho|)\theta_T\varphi(\Theta_T) < 4$, since if $p > 0$ then $4\left(\sqrt{p+1} - \sqrt{p}\right)^2 < 4$. In either case, if (14) holds then $C_{\text{SSVI}} \in \mathcal{C}^p_{\text{AF}}$. ◁

We wrap up this set of examples with a modified power-law decay SSVI model since, as noted in [GJ14, Remark 4.4], power law decay SSVI model of [GJ14, Example 4.3] may not be arbitrage-free for unbounded strikes. Instead, we consider their "modified power-law" decaying SSVI.

Example 2.18 (Modified power-law SSVI)**.** In the context of Proposition 2.16, consider the modified power-law SSVI specification in [GJ14, Remark 4.4, Equation (4.5)]; i.e. for $\eta > 0$ and $\gamma \in (0, \frac{1}{2}]$, for every $\theta > 0$ let

$$\varphi(\theta) \overset{\text{def.}}{=} \frac{\eta}{\theta^\gamma(1+\theta)^{1-\gamma}}.$$

Then $\theta\varphi(\theta) = \eta\left(\frac{\theta}{1+\theta}\right)^{1-\gamma}$ and therefore $\sup_{\theta>0} \theta\varphi(\theta) = \eta$. Condition (13) thus holds whenever

$$\eta \leq \frac{4}{1+\rho}\left(\sqrt{p+1} - \sqrt{p}\right)^2. \tag{15}$$

Moreover, [GJ14, Remark 4.4] shows that the corresponding SSVI surface is free of static arbitrage whenever $\eta(1 + |\rho|) \leq 2$. Combining this with (15) we deduce that $C_{\text{SSVI}} \in \mathcal{C}^p_{\text{AF}}$ if

$$\eta \leq \min\left\{\frac{2}{1+|\rho|}, \frac{4}{1+\rho}\left(\sqrt{p+1} - \sqrt{p}\right)^2\right\}. \tag{16}$$

◁

We close our examples section by noting that satisfying condition (9) does not imply being free of static arbitrage.

Example 2.19 (Condition (9) does not imply no arbitrage)**.** Fix $\beta > 0$ and set $C_\beta : [0,\infty)^2 \to [0,\infty)$ by

$$C_\beta(T, K) \overset{\text{def.}}{=} (1 - \beta K)_+.$$

Then C_β satisfies (9) for every $p > 0$. Indeed, $C_\beta(T, K) = 0$ whenever $K \geq \beta^{-1}$, while for $0 < K < \beta^{-1}$, $C_\beta(T, K) \leq 1 \leq \beta^{-p}K^{-p}$. We note that: if $\beta > 1$, then $\partial_K^+ C_\beta(T, 0) = -\beta < -1$, so C_β violates condition (ii) of Proposition 2.5. ◁

We now present our AFNO framework, and provide several examples of neural operator architecture variants which are covered by our main guarantees.

3 The Arbitrage-Free Neural Operator Framework

We will consider encoder-processor-decoder, sometimes called encoder-decoder, neural network models which are informed with the SANOS : $\mathbb{R}^{F_{\text{out}}} \to C_{\text{AF}}$, cf. [BHK+26], as decoding head. We first start by abstracting this into a class of "good decoders".

3.1 Phase 1: Sample-Based Encoder on RKHS

Unlike neural operators built from fixed encoders, our model allows the sampling locations and the subsequent linear mixing of the sampled values to be trained, thereby permitting the encoder to adapt to the geometry relevant to the RKHS $\mathcal{H}$. At an architectural level, this is closest in spirit to DeepONets [LJP+21], whose trunk network learns a basis for the output functions, and to PCA-Net [BHKS21], whose linear input and output encoders are constructed using functional PCA. A related trainable sampling encoder was recently considered in the sample neural operator (SNO) for deep functional clustering [LBH+26], while related trainable decoder constructions were also noted in [AKP24, Example 5.2 and Remark 5.4].

For every $K \in \mathbb{N}_+$ satisfying $K \leq \max\{1, \dim(\mathcal{H})\}$, we define the trainable encoder

$$\begin{aligned} \mathcal{E}_K : \mathcal{H} \times \mathcal{D}^K \times \mathbb{R}^{K\times K} &\longrightarrow \mathbb{R}^K, \\ (C, (x_k)_{k=1}^K, A) &\longmapsto A\Big(\big\langle C, \kappa(\cdot, x_k)\big\rangle_{\mathcal{H}}\Big)_{k=1}^K \end{aligned} \tag{17}$$

Here $C \in \mathcal{H}$ denotes the input feature function, and κ is the reproducing kernel of the input RKHS. Thus,

$$\mathcal{E}_K(C, (x_k)_{k=1}^K, A) = A\big(C(x_k)\big)_{k=1}^K.$$

These input coordinates are unrestricted real numbers, not convex weights. The kernel sections used for input reconstruction are distinct from the arbitrage-free surfaces used in the output decoder.

The trainable matrix A allows the encoder to linearly mix, and in particular to precondition, the sampled values. The following result shows that suitable choices of the sampling locations and of A yield reconstructions converging uniformly to the identity on compact subsets of $\mathcal{H}$.

Proposition 3.1 (Consistent Reconstruction)**.** *Suppose that $\mathcal{H}$ admits a complete interpolating sequence $\{x_i\}_{i\in I}$ (for some non-empty countable indexing set I). Then, for every non-empty compact set $\mathcal{K} \subseteq \mathcal{H}$ and every $\varepsilon > 0$, there exist $K_\varepsilon \in \mathbb{N}_+$, $(x_k)_{k=1}^{K_\varepsilon} \in \mathcal{D}^{K_\varepsilon}$, and $A_\varepsilon \in \mathbb{R}^{K_\varepsilon \times K_\varepsilon}$ such that*

$$\sup_{C\in\mathcal{K}} \left\| C - \sum_{k=1}^{K_\varepsilon} \mathcal{E}_{K_\varepsilon}\big(C, (x_j)_{j=1}^{K_\varepsilon}, A_\varepsilon\big)_k \kappa(\cdot, x_k) \right\|_{\mathcal{H}} \leq \varepsilon. \tag{18}$$

Proof. See Section C.4. □

While the above statement is straightforward from the *bounded approximation property* (BAP) for any Riesz basis, it has only shown that each C in H can be written as linear combination of basis functions with weights in $\tilde{\beta} \in \mathbb{R}^{K_\epsilon}$. it is *not* evident that the same statement remains true for convex weights $\beta \in [0,1]^{K_\epsilon}$ such that $|\beta| = 1$; which we use below.

3.2 Phase 2: Transformer Backbone

The backbone of our model will be the standard transformer architecture.

Definition 3.2 (Transformers with multi-head attention)**.** Let $\lambda > 0$, and let $d_{\text{in}}, d_{\text{key}}, d_{\text{out}}, N \in \mathbb{N}$. Let $\mathbf{Q}, \mathbf{K} \in \mathbb{R}^{d_{\text{key}}\times d_{\text{in}}}$ and $\mathbf{V} \in \mathbb{R}^{d_{\text{out}}\times d_{\text{in}}}$. We define the associated attention mechanism $\text{Attn}(\cdot|\mathbf{Q}, \mathbf{K}, \mathbf{V}, \lambda) : \mathbb{R}^{N\times d_{\text{in}}} \to \mathbb{R}^{N\times d_{\text{out}}}$ by

$$[\text{Attn}(\mathbf{X}|\mathbf{Q}, \mathbf{K}, \mathbf{V}, \lambda)]_n \overset{\text{def.}}{=} \sum_{m=1}^{N} \frac{e^{\lambda\langle \mathbf{Q}[\mathbf{X}]_n, \mathbf{K}[\mathbf{X}]_m\rangle/\sqrt{d_{\text{key}}}}}{\sum_{l=1}^{N} e^{\lambda\langle \mathbf{Q}[\mathbf{X}]_n, \mathbf{K}[\mathbf{X}]_l\rangle/\sqrt{d_{\text{key}}}}} \mathbf{V}[\mathbf{X}]_m, \qquad n = 1, \ldots, N. \tag{19}$$

Here $[\mathbf{X}]_j \in \mathbb{R}^{d_{\text{in}}}$ denotes the jth row of $\mathbf{X}$, interpreted as a column vector when multiplied by $\mathbf{Q}, \mathbf{K}, \mathbf{V}$. A transformer network operates by iterating layers, each consisting of a row-wise activation–bias map followed by an attention mechanism. Concretely, let $d, D, L \in \mathbb{N}$, and let $d_0, d_1, \ldots, d_L, H_0, H_1, \ldots, H_L \in \mathbb{N}$ such that $d_0 = d$ and $d_L = D$. For $j = 0, 1, \ldots, L-2$ and $\mathtt{h} = 1, \ldots, H_j$, let $d_{\text{key},\mathtt{h}}^{(j)}, d_{\text{value},\mathtt{h}}^{(j)} \in \mathbb{N}$ such that $d_{j+1} = \sum_{\mathtt{h}=1}^{H_j} d_{\text{value},\mathtt{h}}^{(j)}$. A jth *transformer layer with*

multi-head attention, for $j = 0, 1, \ldots, L-2$, is a map $\mathcal{T}_j : \mathbb{R}^{N \times d_j} \to \mathbb{R}^{N \times d_{j+1}}$ that sends an input matrix $\mathbf{X}^{(j)} \in \mathbb{R}^{N \times d_j}$ to $\mathcal{T}_j(\mathbf{X}^{(j)}) \overset{\text{def.}}{=} \mathbf{X}^{(j+1)} \in \mathbb{R}^{N \times d_{j+1}}$, where

$$\mathbf{Z}^{(j)} \overset{\text{def.}}{=} \underbrace{\bigoplus_{\mathtt{h}=1}^{H_j} \mathrm{Attn}\left(\mathbf{X}^{(j)} | \mathbf{Q}_{\mathtt{h}}^{(j)}, \mathbf{K}_{\mathtt{h}}^{(j)}, \mathbf{V}_{\mathtt{h}}^{(j)}, \lambda\right)}_{\text{multi-head attention}}$$

$$[\mathbf{X}^{(j+1)}]_n \overset{\text{def.}}{=} \sigma \bullet \left([\mathbf{Z}^{(j)}]_n + b^{(j)}\right), \qquad n = 1, \ldots, N. \tag{20}$$

Here $\bigoplus$ denotes concatenation along the second (column) dimension, $\mathbf{Q}_{\mathtt{h}}^{(j)}, \mathbf{K}_{\mathtt{h}}^{(j)} \in \mathbb{R}^{d_{\text{key},\mathtt{h}}^{(j)} \times d_j}$ and $\mathbf{V}_{\mathtt{h}}^{(j)} \in \mathbb{R}^{d_{\text{value},\mathtt{h}}^{(j)} \times d_j}$, and $b^{(j)} \in \mathbb{R}^{d_{j+1}}$. Further, the addition in (20) is applied row-wise. A *transformer with multi-head attention* with depth L, obtained by composing the layers $\mathcal{T}_0, \mathcal{T}_1, \ldots, \mathcal{T}_{L-2}$, is a map $\mathcal{T} : \mathbb{R}^{N \times d} \to \mathbb{R}^{N \times D}$ sending an input matrix $\mathbf{X}^{(0)} \overset{\text{def.}}{=} \mathbf{X} \in \mathbb{R}^{N \times d}$ to $\mathcal{T}(\mathbf{X}) \in \mathbb{R}^{N \times D}$, defined row-wise by $[\mathcal{T}(\mathbf{X})]_n \overset{\text{def.}}{=} \mathbf{A}^{(L-1)}[\mathbf{X}^{(L-1)}]_n + b^{(L-1)}$ for $n = 1, \ldots, N$, for some $\mathbf{A}^{(L-1)} \in \mathbb{R}^{D \times d_{L-1}}$ and $b^{(L-1)} \in \mathbb{R}^D$.

We will derive all our results for ReLU-multilayer perceptrons (MLPs) backbones and then convert them into transformer backbones using the transfer principle of [KNP26, Proposition A.1]; which follows in similar suit to the conversion results for CNNs in [PV20] or for transformers with probabilistic formulations of context [KF25, Proposition 11]. For completeness, we record the definition of the standard MLP model in Appendix A.

3.3 Phase 3: Arbitrage-Free Decoders

We will consider pointwise decoders, mapping latent parameters to call surfaces of similar type to the SANOS model of [BHK+26]. We refer to these as arbitrage-free decoders, and defined them as follows. We will consider pointwise decoders which map finite-dimensional latent parameters to arbitrage-free call surfaces. These decoders are modelled on the SANOS architecture of [BHK+26], but the following definition isolates only the structural properties needed in the sequel.

We first concisely present the abstract definitions which are enough for our framework. Shortly after we provide piecewise linear and smooth examples of arbitrage-free decoders in § 3.5 towards the end of this section before passing to our main results.

Definition 3.3 (Arbitrage-Free (AF) Surface-Decoder)**.** Let $\mathcal{Y} \subseteq \mathcal{C}_{AF}^{p,\lambda}$ be a uniformly controlled call-surface class with regularity-data (p, λ, B, ω). We say that $\mathcal{Y}$ is admitted by an arbitrage-free decoder if, for every $R, S > 0$ and every $N, M \in \mathbb{N}_+$, there exist $D_{N,M,R,S} \in \mathbb{N}_+$ and a map

$$\phi^{(N,M,R,S)} : \mathbb{R}^{D_{N,M,R,S}} \times [0,\infty)^2 \to [0,\infty)$$

such that, for every $C \in \mathcal{C}_{AF}^{p,\lambda}$ and $\varepsilon > 0$ there is a $\theta_C \in \mathbb{R}^{D_{N,M,R,S}}$ such that $\phi_{\theta_C} \overset{\text{def.}}{=} \phi(\theta_C, \cdot)$ satisfies:

(i) **Approximate-interpolation on the rectangular grid:** on the uniform grid $\{(t_m, K_n)\}_{m,n=0}^{M,N}$, where $t_m \overset{\text{def.}}{=} \frac{mS}{M}$ and $K_n \overset{\text{def.}}{=} \frac{nR}{N}$, one has: for each $m \in [M]$ and $n \in [N]$

$$\left|\phi_{\theta_C}^{(N,M,R,S)}(t_m, K_n) - C(t_m, K_n)\right| < \varepsilon$$

(ii) **No-arbitrage:** for every parameter $\theta \in \mathbb{R}^{D_{N,M,R,S}}$, $\phi_\theta^{(N,M,R,S)} \in \mathcal{C}_{AF}$,

(iii) **Locally uniform right-wing control:** For every $A > 0$ and every compact maturity interval $[T_{\min}, T_{\max}] \subseteq [0,\infty)$, there exist $\rho > p$ and $B_{\rho,A,R,S,N,M,T_{\min},T_{\max}} > 0$ such that, for every $\theta \in [-A, A]^{D_{N,M,R,S}}$, every $T \in [T_{\min}, T_{\max}]$, and every $K \geq R + 1$,

$$\phi_\theta^{(N,M,R,S)}(T, K) \leq B_{\rho,A,R,S,N,M,T_{\min},T_{\max}} (1 + K)^{-\rho}. \tag{21}$$

When clear from the context, we abbreviate $\Phi_{\theta_C} \overset{\text{def.}}{=} \phi_{\theta_C}^{(N,M,R,S)}$ for notational convenience.

The arbitrage-free decoder should be understood as finitely parameterized individual arbitrage-free call surfaces which are extremely flexible; i.e. non-parametric. These are then convexly combined together to construct our arbitrage-free decoder[6].

Definition 3.4 (Arbitrage-Free Decoder)**.** An *arbitrage-free decoder* combines arbitrage-free surface-decoder outputs using nonnegative weights summing to one. For each input, these weights are independent of strike and maturity.

In (22), the surface parameters may depend on the input. In the fixed-dictionary specialization (23), the parameters θ_i are shared across inputs; they may be pre-fitted or trained jointly with the backbone. We propose here to learn the convex combinations with the attention mechanism introduced earlier, i.e.:

$$\begin{aligned} \Phi : \mathbb{R}^{F_{\mathrm{mkt}}} &\to C([0,\infty)^2, [0,\infty)) \\ \Phi(\theta) &\mapsto \mathrm{Softmax}\left(\tfrac{1}{\lambda}\alpha(\vartheta)\right)^\top [\phi(v(\vartheta)_k)]_{k=1}^K, \end{aligned} \tag{22}$$

where $F_{\mathrm{mkt}}, K \in \mathbb{N}_+$, $\alpha : \mathbb{R}^{F_{\mathrm{mkt}}} \to \mathbb{R}^K$ is a transformer with multi-head attention an ReLU-activation, in (20), $\lambda > 0$, $v : \mathbb{R}^{F_{\mathrm{mkt}}} \to \mathbb{R}^{K\times F}$ is another ς-MLP, the product in (22) is applied point-wise in x, and ϕ is a AF-Decoder model[7].

In Definition 3.4, α represents adaptively computed alignment scores, testing how similar the input market features are to each of the arbitrage-free call surfaces generated by feeding the rows of $v(x)$ into the AF-Decoder model Φ.

3.4 The End-to-End Framework: Arbitrage-Free Neural Operator

We may now close this section by formally defining the class of architectures illustrated in Figure 1. We will show that every such architecture is compatible with risk-neutral pricing theory, uniformly over all parameter specifications, and remains expressive: suitable parameter choices can approximate any sufficiently regular map from an RKHS into the space of arbitrage-free European call-price surfaces.

We emphasize that the Arbitrage-Free Neural Operator (AFNO) architecture is best viewed as an architectural framework. The encoder extracts finitely many real-valued input features. The backbone then computes simplex weights by softmax, and the decoder combines a separate dictionary of arbitrage-free output surfaces using those weights. Convexity of $\mathcal{C}_{AF}$ provides the no-arbitrage certificate. The output dictionary need not be a basis and does not determine the input RKHS.

We can *model* the RKHS used to model the input data, as well as the arbitrage-free decoder most appropriate for the particular use case or downstream application, provided that it satisfies the basic requirements of Definition 3.3. Examples include the SANOS model of [BHK+26] and the simpler piecewise-linear decoder introduced in Definition 3.7.

Definition 3.5 (Arbitrage-Free Neural Operator)**.** Let $\mathcal{H}$ be an RKHS over a non-empty set $\mathcal{D} \subseteq [0,\infty)^2$. A map $\Psi : \mathcal{H} \to C\big([0,\infty)^2, [0,\infty)\big)$ is called an *arbitrage-free neural operator* (AFNO) if it admits a representation of the form

$$\begin{aligned} \Psi : \mathcal{H} &\longrightarrow C\big([0,\infty)^2, [0,\infty)\big), \\ C &\longmapsto \sum_{i=1}^{I} \mathrm{Softmax}\left(\frac{1}{\tau}\alpha\left(\mathcal{E}_K(C, \mathbf{x}, A)\right)\right)_i \phi_{\theta_i}, \end{aligned} \tag{23}$$

where $K, I \in \mathbb{N}_+$, $\mathbf{x} \overset{\text{def.}}{=} (x_k)_{k=1}^K \in \mathcal{D}^K$, $A \in \mathbb{R}^{K\times K}$, and $\alpha : \mathbb{R}^K \to \mathbb{R}^I$ is either a transformer with multi-head attention or a ReLU-MLP. Moreover, $\tau > 0$, and $\phi_{\theta_1}, \ldots, \phi_{\theta_I} \in \mathcal{C}_{AF}$ are outputs of an arbitrage-free decoder in the sense of Definition 3.3.

In what follows, we refer to both *AF surface-decoders* and *AF decoders* simply as *AF decoders* to keep terminology light, as it will be conceptually clear.

[6] Each individual call price surfaces can, but need not, be trained.
[7] Cf. Definition 3.3.

3.5 Examples of AF-Decoders

We now provide two examples of AF-decoders showing the reader how they can build their own AFNO. The first is an intuitive model which is piecewise linear in each strike and along maturities for a fixed strike. The second is the SANOS model of [BHK+26] which is effectively a shallow neural network with Black-Scholes European call-price as is activation function.

3.5.1 A Simple Example of an Arbitrage-Free Decoder

While one may consider more sophisticated arbitrage-free decoders, such as the C^∞-regularity SANOS model in [BHK+26], we find it most illustrative to focus here on piecewise-linear examples, since our aim is not to advocate for a particular decoder, but rather to demonstrate how such decoders can be used to construct a certifiably arbitrage-free neural operator. We also note that our chosen decoder admits a tensorized ReLU-MLP representation, thereby further connecting our construction to a "structured version" of DeepONets; cf. [LSYZ21]. One could also build an arbitrage-free decoder using the $C^{1,1}$-regularity reconstruction formula of [CN17].

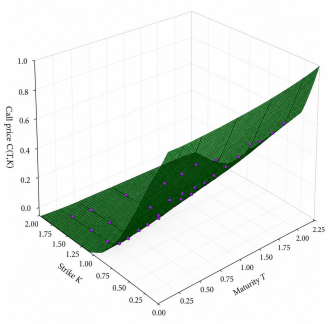


Figure 2: Illustration of the structured ReLU-MLP arbitrage-free decoder with $M = 5$ maturities and $N = 8$ strikes; the dots indicate the interpolated grid quotes.

The following is the "Discrete Local Vol" interpolation scheme for homogeneous strikes described in [BR15]. This amounts to linear interpolation between strikes:

Proposition 3.6 (Arbitrage-Free Discrete Polytope)**.** *Let $M, N \in \mathbb{N}_+$. Fix $0 < T_1 < \cdots < T_M$ and $0 < K_1 < \cdots < K_N$. Fix, moreover, two virtual strikes $K_{\mathrm{L}}, K_{\mathrm{R}} > 0$ satisfying $0 < K_{\mathrm{L}} < \min\{K_1, 1\}$ and $K_{\mathrm{R}} > \max\{K_N, 1\}$. Write $\mathbf{T} \overset{\text{def.}}{=} (T_m)_{m=1}^M$, $\mathbf{K} \overset{\text{def.}}{=} (K_n)_{n=1}^N$, and set*

$$\widehat{K}_0 \overset{\text{def.}}{=} 0, \qquad \widehat{K}_1 \overset{\text{def.}}{=} K_{\mathrm{L}}, \qquad \widehat{K}_{n+1} \overset{\text{def.}}{=} K_n \quad (n \in [N]_+), \qquad \widehat{K}_{N+2} \overset{\text{def.}}{=} K_{\mathrm{R}}.$$

Given any matrix of quotes $y = (y_{m,n})_{m,n=1}^{M,N} \in \mathbb{R}^{M\times N}$, we defined the extended matrix of quotes by

$$\widehat{y}_{m,0} \overset{\text{def.}}{=} 1, \qquad \widehat{y}_{m,1} \overset{\text{def.}}{=} 1 + \frac{K_{\mathrm{L}}}{K_1}(y_{m,1} - 1) \qquad \widehat{y}_{m,n+1} \overset{\text{def.}}{=} y_{m,n} \quad (n \in [N]_+), \qquad \widehat{y}_{m,N+2} \overset{\text{def.}}{=} 0$$

and for each $j \in [N+1]$, define $s_{m,j}(y) \overset{\text{def.}}{=} \frac{\widehat{y}_{m,j+1} - \widehat{y}_{m,j}}{\widehat{K}_{j+1} - \widehat{K}_j}$.
Then, for each $y \in \mathbb{R}^{M\times N}$, the following are equivalent:

(i) ***Existence of an Arbitrage-Free Interpolator:*** *There exists a continuous map $\mathfrak{I} : [0,\infty)^2 \to [0,\infty)$ such that: $\mathfrak{I}(T,\cdot)$ and $\mathfrak{I}(\cdot,K)$ are piecewise affine for each $T, K \geq 0$ such that:*

$$\mathfrak{I}(T_m, \widehat{K}_j) = \widehat{y}_{m,j} \qquad \text{for every } m \in [M]_+ \text{ and } j \in [N+2].$$

and such that $\mathfrak{I}$ is an arbitrage-free call-price surface (cf. Proposition 2.5),

(ii) ***Polytope Membership:*** *y belongs to the compact and convex polytope*

$$\mathfrak{C}_{\mathbf{T},\mathbf{K}} \overset{\text{def.}}{=} \Big\{ y \in \mathbb{R}^{M\times N} : -1 \leq s_{m,0}(y) = s_{m,1}(y) \leq \cdots \leq s_{m,N+1}(y) \leq 0, \qquad m \in [M]_+ \\ y_{m,n} \leq y_{m+1,n}, \quad m \in [M-1]_+, \quad n \in [N]_+ \Big\}. \tag{24}$$

Moreover, $\mathfrak{C}_{\mathbf{T},\mathbf{K}}$ is non-empty and there is a unique orthogonal projection $P_{\mathbf{T},\mathbf{K}} : \mathbb{R}^{M\times N} \to \mathfrak{C}_{\mathbf{T},\mathbf{K}}$.

Proof. See Section A.1. □

The polytope in (24) characterizes no-arbitrage for 'low-regularity" European call-price interpolators, by which we mean interpolators that are piecewise linear in each argument separately. This characterization allows us to define the following arbitrage-free decoder on a single parameter space.

Definition 3.7 (Structured ReLU-MLP Arbitrage-Free Decoder)**.** Fix $\delta > 0$ and $M, N \in \mathbb{N}_2$. Consider the decoder

$$\begin{aligned} \phi_\delta^{(M,N)} : \mathbb{R}^M \times \mathbb{R}^N \times \mathbb{R}^{M\times N} &\longrightarrow C\big([0,\infty)^2, [0,\infty)\big), \\ (\mathbf{t}, \mathbf{k}, \mathbf{z}) &\longmapsto \Psi_{\mathbf{T},\mathbf{K},\mathbf{y}}, \end{aligned} \tag{25}$$

where $\Psi_{\mathbf{T},\mathbf{K},\mathbf{y}}$ is defined as follows: Write $\mathbf{t} \overset{\text{def.}}{=} (t_m)_{m=1}^M$ and $\mathbf{k} \overset{\text{def.}}{=} (k_n)_{n=1}^N$, and set $t_0 \overset{\text{def.}}{=} k_0 \overset{\text{def.}}{=} 0$. Define the ordered grids

$$(T_1, \dots, T_M) \overset{\text{def.}}{=} \left(\sum_{j=1}^m \max\left\{|t_j - t_{j-1}|, \delta\right\} \right)_{m=1}^M \quad \text{and} \quad (K_1, \dots, K_N) \overset{\text{def.}}{=} \left(\sum_{j=1}^n \max\left\{|k_j - k_{j-1}|, \delta\right\} \right)_{n=1}^N$$

Define $K_{\mathrm{L}} \overset{\text{def.}}{=} \frac{1}{2}\min\{\delta, 1\}$, $K_{\mathrm{R}} \overset{\text{def.}}{=} K_N + 1 + \delta^{-1}$, and define the corrected quote data by $\mathbf{y} = (y_{m,n})_{m,n=1}^{M,N} \overset{\text{def.}}{=} P_{\mathbf{T},\mathbf{K},\delta}(\mathbf{z})$ where $P_{\mathbf{T},\mathbf{K},\delta} : \mathbb{R}^{M\times N} \mapsto \mathfrak{C}_{\mathbf{T},\mathbf{K},\delta}$ is the projection onto the polytope $\mathfrak{C}_{\mathbf{T},\mathbf{K},\delta}$, defined as in Proposition 3.6[8]. Next, for each $m \in [M]_+$, define the augmented strike and quote data, where $n \in [N]_+$, by

$$\begin{aligned} (\widehat{K}_0, \widehat{y}_{m,0}) &\overset{\text{def.}}{=} (0, 1) \\ (\widehat{K}_1, \widehat{y}_{m,1}) &\overset{\text{def.}}{=} \left(K_{\mathrm{L}}, 1 + \frac{K_{\mathrm{L}}}{K_1}(y_{m,1} - 1) \right) \\ (\widehat{K}_{n+1}, \widehat{y}_{m,n+1}) &\overset{\text{def.}}{=} (K_n, y_{m,n}) \\ (\widehat{K}_{N+2}, \widehat{y}_{m,N+2}) &\overset{\text{def.}}{=} (K_{\mathrm{R}}, 0). \end{aligned}$$

For each $j \in [N+2]_+$, define $\Delta\widehat{K}_j \overset{\text{def.}}{=} \widehat{K}_j - \widehat{K}_{j-1}$, and the decreasing secant slopes $S_{m,j} \overset{\text{def.}}{=} \frac{\widehat{y}_{m,j-1} - \widehat{y}_{m,j}}{\Delta\widehat{K}_j}$. Set $S_{m,N+3} \overset{\text{def.}}{=} 0$. Set $T_0 \overset{\text{def.}}{=} 0$, and let $e_0, \dots, e_M$ denote the canonical basis of $\mathbb{R}^{M+1}$, indexed from 0 to M. Consider the ReLU-MLPs $\mathbf{\Psi}, \Lambda : \mathbb{R} \to \mathbb{R}^{M+1}$ defined, for each $K, T \in \mathbb{R}$, by

$$\begin{aligned} \mathbf{\Psi}(K)_0 &\overset{\text{def.}}{=} \mathrm{ReLU}(1 - K), \\ \mathbf{\Psi}(K)_m &\overset{\text{def.}}{=} \mathrm{ReLU}\left(\sum_{j=1}^{N+2} \left(S_{m,j} - S_{m,j+1}\right) \mathrm{ReLU}\left(\widehat{K}_j - K\right) \right) \qquad m \in [M]_+, \\ \Lambda(T) &\overset{\text{def.}}{=} \mathrm{ReLU}\left(e_0 + \sum_{m=0}^{M-1} \frac{\mathrm{ReLU}(T - T_m) - \mathrm{ReLU}(T - T_{m+1})}{T_{m+1} - T_m} (e_{m+1} - e_m) \right). \end{aligned} \tag{26}$$

Finally, define $\Psi_{\mathbf{T},\mathbf{K},\mathbf{y}} : [0,\infty)^2 \to [0,\infty)$ by $\Psi_{\mathbf{T},\mathbf{K},\mathbf{y}}(T, K) \overset{\text{def.}}{=} \Lambda(T)^\top \mathbf{\Psi}(K)$.

The next proposition shows that, for sufficiently small ordering tolerance, the decoder exactly interpolates any arbitrage-free call surface on the prescribed grid while remaining arbitrage-free globally.

Proposition 3.8 (Universal Arbitrage-Free Interpolation)**.** *Let $C : [0,\infty)^2 \to [0,\infty)$ be an arbitrage-free call surface, let $M, N \in \mathbb{N}_2$, and fix grids $0 < T_1 < \cdots < T_M$ and $0 < K_1 < \cdots < K_N$. Set $z_{m,n} \overset{\text{def.}}{=} C(T_m, K_n)$ for each $m \in [M]_+$ and $n \in [N]_+$. Then there exists $\delta_0 > 0$ such that, for every $0 < \delta \le \delta_0$,*

$$\phi_\delta^{(M,N)}(\mathbf{T}, \mathbf{K}, \mathbf{z})(T_m, K_n) = C(T_m, K_n)$$

for each $m \in [M]_+$ and $n \in [N]_+$. Moreover, $\phi_\delta^{(M,N)}(\mathbf{T}, \mathbf{K}, \mathbf{z})$ is an arbitrage-free call surface.

Proof. See Section A.1. □

[8] Using the fixed virtual strikes K_{L} and K_{R}.

The following proposition records that the resulting ReLU construction is an arbitrage-free decoder in the sense of Definition 3.3.

Proposition 3.9 (The Structured ReLU-MLP Family Induce Arbitrage-Free Decoders)**.** *For $R, S > 0$ and $N, M \in \mathbb{N}_+$, set $\widehat{N} \overset{\text{def.}}{=} \max\{N, 2\}$ and $\widehat{M} \overset{\text{def.}}{=} \max\{M, 2\}$. The family*

$$\phi_\theta^{(N,M,R,S)} \overset{\text{def.}}{=} \phi_{e^{-\xi}}^{(\widehat{M},\widehat{N})}(\mathbf{t}, \mathbf{k}, \mathbf{z}), \qquad \theta = (\xi, \mathbf{t}, \mathbf{k}, \mathbf{z}) \in \mathbb{R}^{1+\widehat{M}+\widehat{N}+\widehat{M}\widehat{N}},$$

satisfies Definition 3.3 for every $p, \lambda > 0$. In particular, the ordering tolerance is a selectable parameter.

Proof. See Section A.1. □

Next, we show that the SANOS model is a smooth decoder and the approximate interpolation property gives us just enough slack to obtain *closed-form* parameter specifications certifying no static arbitrage.

3.5.2 SANOS is an Arbitrage-Free Decoder

We now construct a smooth counterpart of the structured ReLU decoder in Definition 3.7. The construction is inspired by the Black-Scholes smoothing mechanism underlying the SANOS parameterization of [BHK+26]. A direct replacement of each hinge $(\widehat{K}_j - K)^+$ by a smooth approximation of $\mathrm{ReLU}(\widehat{K}_j - K)$ does not, in general, preserve the two left-wing conditions $C(T, 0) = 1$ and $\partial_K^+ C(T, 0) = -1$. We therefore first introduce a smooth Black-Scholes approximation of ReLU function, defined for every $\nu > 0$ and each $x \in \mathbb{R}$ by $\mathrm{BS}_\nu : \mathbb{R} \to [0, \infty)$ by

$$\mathrm{BS}_\nu(x) \overset{\text{def.}}{=} \begin{cases} (1+x)\Phi\left(\dfrac{\log(1+x) + \frac{1}{2}\nu^2}{\nu}\right) - \Phi\left(\dfrac{\log(1+x) - \frac{1}{2}\nu^2}{\nu}\right) & \text{if: } x > -1 \\ 0 & \text{if: } x \le -1. \end{cases} \tag{27}$$

We extend (27) it to values of $\nu \le 0$ by setting $\mathrm{BS}_\nu \overset{\text{def.}}{=} \mathrm{ReLU}$ for every $\nu \le 0$. It can be shown (Proposition A.4) that doing so makes BS continuous in both ν and x on all of $\mathbb{R}^2$ since BS_ν converges to ReLU uniformly, over all of $\mathbb{R}$, as ν tends to 0 from the right. We highlight that, for $\nu > 0$ and $x > -1$, $\mathrm{BS}_\nu(x)$ is the Black-Scholes put price with spot and maturity equal to 1, zero interest rate, strike $1+x$, and volatility ν and that BS_ν is smooth on $(0, \infty)$ whenever $\nu > 0$ and it is in $C^2(\mathbb{R})$.

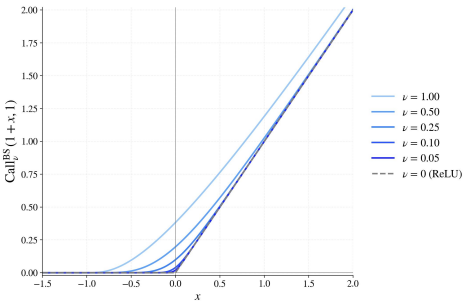


Figure 3: Illustration of the normalized Black–Scholes activation $\mathrm{BS}_\nu(x) = \mathrm{Call}_\nu^{\mathrm{BS}}(1+x, 1)$, corresponding to the homogeneous Black–Scholes call hinge (28). As ν decreases towards 0 $\mathrm{Call}_\nu^{\mathrm{BS}}$ converges uniformly to ReLU.

The strike-side smoothing must respect the homogeneity of call prices. For $\nu \ge 0$, $s > 0$, and $K \ge 0$, define the Black-Scholes call hinge by

$$\mathrm{Call}_\nu^{\mathrm{BS}}(s, K) \overset{\text{def.}}{=} s - K + s\ \mathrm{BS}_\nu\left(\frac{K}{s} - 1\right) \tag{28}$$

Moreover, the zero-volatility limit recovers the intrinsic-value; i.e. $\mathrm{Call}_0^{\mathrm{BS}}(s, K) = (s - K)^+$.

Definition 3.10 (SANOS Decoder)**.** Fix $\delta, \overline{\nu} > 0$ and $M, N \in \mathbb{N}_2$. Retain the grid correction, polytope projection, augmented strikes, augmented quotes, and secant slopes from Definition 3.7. Define

$$\begin{aligned} \phi_{\delta,\overline{\nu}}^{(M,N)} : \mathbb{R}^M \times \mathbb{R}^N \times \mathbb{R}^{M\times N} \times \mathbb{R} &\longrightarrow C\big([0,\infty)^2, [0,\infty)\big) \cap C^\infty\big((0,\infty)^2\big), \\ (\mathbf{t}, \mathbf{k}, \mathbf{z}, \eta) &\longmapsto \Psi_{\mathbf{T},\mathbf{K},\mathbf{y},\nu}, \end{aligned} \tag{29}$$

where $\nu \overset{\text{def.}}{=} \nu_{\overline{\nu}}(\eta)$ and the surface on the right-hand side is defined as follows. Here $e_1, \dots, e_M$ denote the canonical basis of $\mathbb{R}^M$. For each $K, T \geq 0$, set

$$\mathbf{\Psi}_\nu(K)_m \overset{\text{def.}}{=} \sum_{j=1}^{N+2} q_{m,j} \, \mathrm{Call}_\nu^{\mathrm{BS}}\left(\widehat{K}_j, K\right) \qquad \text{for } m \in [M]_+ \tag{30}$$

$$a_{m,\nu}(T) \overset{\text{def.}}{=} \frac{\mathrm{BS}_\nu(T - T_m) - \mathrm{BS}_\nu(T - T_{m+1})}{T_{m+1} - T_m} \qquad \text{for } m \in [M-1]_+ \tag{31}$$

$$\Lambda_\nu(T) \overset{\text{def.}}{=} e_1 + \sum_{m=1}^{M-1} a_{m,\nu}(T)\big(e_{m+1} - e_m\big) \tag{32}$$

where, for every $m \in [M]_+$ and $j \in [N+2]_+$, set $q_{m,j} \overset{\text{def.}}{=} S_{m,j} - S_{m,j+1}$. For every $\nu, T \geq 0$ set $b_\nu(T) \overset{\text{def.}}{=} \frac{\mathrm{BS}_\nu\left(\frac{T}{T_1}\right) - \mathrm{BS}_\nu\left(\frac{T}{T_1} - 1\right) - \mathrm{BS}_\nu(0)}{1 - \mathrm{BS}_\nu(0)}$[9], $B_\nu(T,K) \overset{\text{def.}}{=} \mathrm{Call}^{\mathrm{BS}}_{\nu b_\nu(T)}(1,K)$, $\Gamma_{\mathbf{T},\mathbf{K},\mathbf{y},\nu}(T,K) \overset{\text{def.}}{=} \Lambda_\nu(T)^\top \mathbf{\Psi}_\nu(K)$ and define $\Psi_{\mathbf{T},\mathbf{K},\mathbf{y},\nu} : [0,\infty)^2 \to [0,\infty)$ by

$$\Psi_{\mathbf{T},\mathbf{K},\mathbf{y},\nu}(T,K) \overset{\text{def.}}{=} \big(1 - b_\nu(T)\big) B_\nu(T,K) + b_\nu(T) \Gamma_{\mathbf{T},\mathbf{K},\mathbf{y},\nu}(T,K). \tag{33}$$

Proposition 3.11 (Smooth arbitrage-free approximation)**.** *In the setting of Definition 3.10, let* $\Psi_{\mathbf{T},\mathbf{K},\mathbf{y},0}$ *denote the structured* ReLU *surface obtained at* $\nu = 0$*, and set* $\Delta_T^{\min} \overset{\text{def.}}{=} \min_{m \in [M-1]_+} \big(T_{m+1} - T_m\big) > 0$*. Then, for every* $\nu \in (0, \overline{\nu})$*, the following assertions hold:*

(i) ***Regularity:*** $\Psi_{\mathbf{T},\mathbf{K},\mathbf{y},\nu} \in C^\infty\big((0,\infty)^2\big) \cap C\big([0,\infty)^2, [0,\infty)\big)$.

(ii) ***No static arbitrage:*** $\Psi_{\mathbf{T},\mathbf{K},\mathbf{y},\nu} \in \mathcal{C}_{AF}$.

(iii) ***Uniform perturbation of PWL Surface:*** *For any* $\varepsilon > 0$*, if* $0 < \nu \leq \min\left\{\frac{\sqrt{2\pi}}{2}, \frac{\sqrt{2\pi}\,\Delta_T^{\min}}{10\Delta_T^{\min} + 4(M-1)}\,\varepsilon\right\}$

$$\sup_{T,K \geq 0} \left|\Psi_{\mathbf{T},\mathbf{K},\mathbf{y},\nu}(T,K) - \Psi_{\mathbf{T},\mathbf{K},\mathbf{y},0}(T,K)\right| \leq \varepsilon. \tag{34}$$

Moreover, fix $\overline{\nu} > 0$*. For* $R, S > 0$ *and* $N, M \in \mathbb{N}_+$*, set* $\widehat{N} \overset{\text{def.}}{=} \max\{N, 2\}$ *and* $\widehat{M} \overset{\text{def.}}{=} \max\{M, 2\}$*. The family*

$$\phi_\theta^{(N,M,R,S)} \overset{\text{def.}}{=} \phi_{e^{-\xi},\overline{\nu}}^{(\widehat{M},\widehat{N})}(\mathbf{t},\mathbf{k},\mathbf{z},\zeta), \qquad \theta = (\xi,\mathbf{t},\mathbf{k},\mathbf{z},\zeta) \in \mathbb{R}^{2+\widehat{M}+\widehat{N}+\widehat{M}\widehat{N}},$$

is an arbitrage-free surface-decoder in the sense of Definition 3.3.

Proof. See Section A.1.1. □

Remark 3.12. The main difference between the formula in (3.10) and the formulation of the SANOS model in [BHK+26] is that here we choose the parameters for approximate interpolation, cf. (34), which allow us enough flexibility access to the closed-form expression in (30)-(31); while in [BHK+26] exact interpolation is achieved which is solved implicitly via a linear program. ◁

We are now ready to showcase our positive results first.

4 Positive Results: AF-NO

We present our first positive result for our AFNO model. Namely, we show that it necessarily always produces arbitrage-free European call-price surfaces.

Theorem 4.1 (AF-NO is Certifiably Arbitrage-Free)**.** *Every AF-NO maps into* $\mathcal{C}_{AF}$*; i.e. for every input* x *in its domain,* $(T,K) \mapsto \Phi(x)(T,K)$ *satisfies the conditions in Proposition 2.5.*

Proof. See Appendix C.3.1. □

[9] Since $\mathrm{BS}_\nu(-1) = 0$, we have $b_\nu(0) = 0$ and at $\nu = 0$, $b_0(T) = \min\{\frac{T}{T_1}, 1\}$.

The approximation proof chooses finitely many anchor surfaces for the target operator and tolerance, then combines their decoder approximations using continuous softmax weights. Unlike hard nearest-neighbour selection, this yields a continuous map while preserving no-arbitrage. The theorem does not assert a smaller dictionary or a better worst-case approximation rate than nearest-neighbour approximation.

Setting 4.1. Fix an RKHS $\mathcal{H}$ over a non-empty set $\mathcal{D} \subseteq [0,\infty)^2$ and suppose that $\mathcal{H}$ admits a complete interpolating sequence. Fix $0 < q < p$ and let $l, \lambda > 0$ satisfy $\frac{q}{p}\lambda < l < \lambda$. Let $\mathcal{K} \subseteq \mathcal{H}$ be non-empty and compact (set if inputs on which we will approximate), and let $\mathcal{Y} \subseteq \mathcal{C}_{AF}^{p,\lambda}$ be a uniformly controlled call-surface class admitted by the arbitrage-free decoder of Definition 3.3, and assume that this class and decoder satisfy Assumption C.1. Let $\mathcal{Q} \subset \mathcal{C}_{AF}^{p,\lambda}$ be compact and non-empty. Let $\mathcal{F} : \mathcal{H} \to \mathcal{Y}$ be L-Lipschitz as a map from $(\mathcal{H}, \|\,\|_{\mathcal{H}})$ into $(\mathcal{C}_{AF}^{q,l}, \|\,\|_{q,l})$.

Theorem 4.2 (Universal Arbitrage-Free Approximation)**.** *Under Setting 4.1, for every $0 < \varepsilon < 1$, there exist $I, K \in \mathbb{N}_+$, points $x_1, \dots, x_K \in \mathcal{D}$, a matrix $A \in \mathbb{R}^{K\times K}$, numbers $R, S > 0$, integers $N_0, M_0 \in \mathbb{N}_+$, parameters $\theta_1, \dots, \theta_I \in \mathbb{R}^{D_{N_0,M_0,R,S}}$, a temperature $\tau > 0$, and a* ReLU*-MLP $\alpha_{\mathbf{X}} : \mathbb{R}^K \to \mathbb{R}^I$ such that the map $\Phi : \mathcal{H} \to \mathcal{C}_{AF}$ defined by*

$$\Phi(C) \overset{\text{def.}}{=} \sum_{i=1}^{I} \text{Softmax}\left(\frac{1}{\tau}\alpha_{\mathbf{X}}\left(\mathcal{E}_K\big(C, (x_k)_{k=1}^K, A\big)\right)\right)_i \widetilde{C}_i \quad \textit{where} \quad \widetilde{C}_i \overset{\text{def.}}{=} \phi_{\theta_i}^{(N_0,M_0,R,S)} \tag{35}$$

satisfies:

(i) ***Uniform Approximation:*** $\|\Phi(C) - \mathcal{F}(C)\|_{q,l} \le \varepsilon$ *for all* $C \in \mathcal{K}$,

(ii) ***Interpretable Range:*** $\Phi(\mathcal{H}) \subseteq \mathcal{C}_{AF} \cap \text{co}\{\widetilde{C}_1, \dots, \widetilde{C}_I\}$.

Proof Sketch. For full details, see Appendix C.4; at a high-level, the idea of the proof is the following:

Step 1 : For $\delta > 0$, compactness of $\mathcal{K}$ gives a finite δ-net $\{x_n\}_{n=1}^N \subseteq \mathcal{K}$. Setting $y_n \overset{\text{def.}}{=} \mathcal{F}(x_n)$, Lipschitz continuity gives

$$\sup_{x\in\mathcal{K}} \min_{1\le n\le N} \|\mathcal{F}(x) - y_n\|_{q,l} \le L \sup_{x\in\mathcal{K}} \min_{1\le n\le N} \|x - x_n\|_{\mathcal{H}} \le L\delta.$$

Proposition C.2 then provides decoder approximations of these finitely many output surfaces with common decoder hyperparameters.

Step 2 : The next step is to interpolate the a *continuous* approximate relaxation of the *nearest neighbourhoods problem*[10]; i.e. the map sending any $x \in \mathcal{K}$ to a choice of closest x_n and then, assigning x to $\mathcal{F}(x_n)$.

□

4.1 Illustration: Simultaneous Call Surface Prediction

We now apply our theory to show how one can learn the *parameter-to-European Call-price surface* map for a broad class of parametric families of SDEs; possibly with non-Brownian drivers. We recall that any finite-dimensional RKHS is isometrically isomorphic to $\mathbb{R}^d$, and we treat it as such henceforth.

The basic idea here is that of a pre-trained model for call prices: we use sufficiently parameterized general diffusion processes to generate a large number of $\tilde{C}$ basis call price surfaces. We then show, first, that a surface generated by any new set of diffusion parameters is approximated to precision.

Assumption 4.3 (Latent Process)**.** Let $d_\nu, F_{\text{mkt}} \in \mathbb{N}_+$, $r \ge 0$, and let $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{Q})$ be a filtered probability space. Let $M = (M_t)_{t\ge 0}$ be a continuous square-integrable $(\mathbb{Q}, \mathbb{F})$-martingale generating $\mathbb{F}$, and assume that $d\langle M\rangle_t = \chi_t\, dt$ for a predictable process χ satisfying $0 \le \chi_t \le \bar{c} < \infty$. Let

[10] Which itself is typically discontinuous, and thus not desirable as it would yield a network not trainable by standard gradient-based optimization algorithms in practice.

ν be a fixed $\mathbb{R}^{d_\nu}$-valued predictable process, let $\Theta \subseteq \mathbb{R}^{F_{\text{mkt}}}$ be non-empty and compact, and let $\sigma : [0,\infty) \times \mathbb{R}^{d_\nu} \times \Theta \times \mathbb{R} \to \mathbb{R}$ be Borel measurable. Assume that there exist finite constants $\overline{\sigma}, L_x, L_\theta \geq 0$ such that, for all $t \geq 0$, $v \in \mathbb{R}^{d_\nu}$, $\theta, \theta' \in \Theta$, and $x, y \in \mathbb{R}$,

$$\begin{aligned} |\sigma(t,v,\theta,x)| &\leq \overline{\sigma}, \\ |x\sigma(t,v,\theta,x) - y\sigma(t,v,\theta,y)| &\leq L_x|x-y|, \\ |\sigma(t,v,\theta,x) - \sigma(t,v,\theta',x)| &\leq L_\theta \|\theta - \theta'\|_2. \end{aligned} \tag{36}$$

and for each $\theta \in \Theta$, let X^θ be the unique adapted solution of

$$X_t^\theta = 1 + \int_0^t r X_s^\theta \, ds + \int_0^t X_s^\theta \sigma(s, \nu_s, \theta, X_s^\theta) \, dM_s, \qquad t \geq 0. \tag{37}$$

Assume, moreover, that, for every $\theta \in \Theta$, $\inf_{0 \leq t \leq T} X_t^\theta > 0$ $\mathbb{Q}$-a.s.

Our objective is to approximate the solution operator corresponding to the family of latent price generating process X above; namely, the map

$$\text{Sol} : \Theta \longrightarrow C([0,\infty)^2, [0,\infty)), \qquad \text{Sol}(\theta)(T,K) \overset{\text{def.}}{=} \mathbb{E}_{\mathbb{Q}}\big[(e^{-rT} X_T^\theta - K)_+\big]. \tag{38}$$

Let us first note that our problem is well-posed before approximately solving it.

Proposition 4.4 (Well-posedness)**.** *Under Assumption 4.3, put* $a \overset{\text{def.}}{=} \overline{c}\,\overline{\sigma}^2$, $b \overset{\text{def.}}{=} \max\{2\overline{c}L_x^2, a\}$, $\kappa_u \overset{\text{def.}}{=} \frac{a}{2}u(u-1)^+$ *for* $u > 0$, *and* $\gamma_q \overset{\text{def.}}{=} \frac{1}{2}(b + \kappa_{2q})$ *for* $q > 0$. *The solution operator in (38) is well-defined and takes values in* $\mathcal{C}_{AF}$. *For every* $p, \lambda > 0$ *satisfying* $\lambda \geq \kappa_{p+1}$, *its image* $\mathcal{Y} \overset{\text{def.}}{=} \text{Sol}(\Theta)$ *is uniformly controlled with* $B = 2^p$ *and common modulus* $\omega(h) = h + \max\{1, \sqrt{e^a - 1}\}\sqrt{h}$. *For every* $q > 0$ *and* $l > \gamma_q$, *for each* $\theta, \theta' \in \Theta$ *we have*

$$\|\,\text{Sol}(\theta) - \text{Sol}(\theta')\|_{q,l} \leq \frac{2^{q+1}\sqrt{\overline{c}}\,L_\theta}{\sqrt{2e(l-\gamma_q)}} \|\theta - \theta'\|_2.$$

Upon specializing to Proposition 4.3, we obtain the following *universal approximation* result with no-arbitrage constraints. Briefly, the result says that under mild conditions, for a parametric family of price processes, the parameter-to-call surface map can be approximated to arbitrary precision without ever violating the static no-arbitrage conditions.

Corollary 4.5 (Universal Arbitrage-Free Pricing for Families of Pricing Models of the Form (37))**.** *Under Assumption 4.3, choose* $0 < q < p$, $\lambda > \max\{\kappa_{p+1}, \gamma_q\}$, *and* $\max\{q\lambda/p, \gamma_q\} < l < \lambda$, *with the constants from Proposition 4.4. Let* $\mathcal{Y} \overset{\text{def.}}{=} \text{Sol}(\Theta)$ *and use an AF surface-decoder*[11] *for* $\mathcal{Y}$ *then, for every* $0 < \varepsilon < 1$ *and* $H > 0$, *there exists an AF-NO* $\Phi : \mathbb{R}^{F_{\text{mkt}}} \to \mathcal{C}_{AF}$ *satisfying*

$$\sup_{\theta \in \Theta} \max_{0 \leq T,K \leq H} \left| \mathbb{E}_{\mathbb{Q}}\big[(e^{-rT} X_T^\theta - K)_+\big] - \Phi(\theta)(T,K) \right| \leq \varepsilon.$$

We now showcase our negative results.

5 Negative Results: *Un*-certifiability for Projections and Standard NOs do not Have No-Arbitrage

In this section, we show that the natural alternative tools available either fail all-together to be arbitrage-free in general, or at-the very least, they cannot be **certified** to always yield **exactly** arbitrage-free call surfaces.

[11] Satisfying Definition 3.3 and Assumption C.1; e.g. the PWL and SANOS families satisfy these requirements by Proposition A.3 and Corollary A.6.

5.1 No Arbitrage-Free Projections

We first show that one simply cannot learn a surface which fits market data well and then "project" it onto the space of arbitrage-free surfaces; where by projection we mean minimizing *any* metric which is similar to an $\|\cdot\|_{p,\lambda}$ metric on the natural space $\mathcal{X}_p^\lambda \subseteq C([0,\infty)^2,[0,\infty))$. I.e. we show that there is some surface which is "un-projectable" onto $\mathcal{C}_{AF} \cap \mathcal{X}_p^\lambda$.

Theorem 5.1 (No *Nearest* Arbitrage-Free Projection)**.** *Fix $p,\lambda>0$, and let ρ be any metric on $C([0,\infty)^2,[0,\infty))$ which induces a topology which is no stronger than the uniform convergence on compact sets topology on $C([0,\infty)^2,[0,\infty))$. Then, there exists surface $C_{\text{bad}} \in C([0,\infty)^2,[0,\infty))$ such that there is no $\Pi_C \in \mathcal{C}_{AF}$ satisfying*

$$\rho(\Pi_C, C_{\text{bad}}) = \inf_{\tilde{C}\in\mathcal{C}_{AF}} \rho(\tilde{C}, C_{\text{bad}}).$$

Proof. See Section C.5.1. □

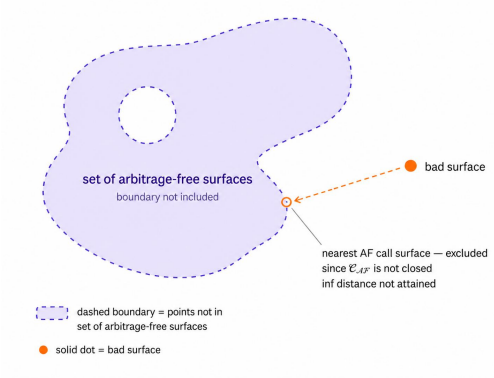


Figure 4: **No Nearest Arbitrage-Free Projection:** Theorem 5.1 states that there exists a candidate European call-price surface $C_{\text{bad}} : [0,\infty)^2 \to [0,\infty)$ (red dot) which admits no best approximation in $\mathcal{C}_{AF}$. Indeed, C_{bad} lies in the topological closure of $\mathcal{C}_{AF}$, represented by the dashed green boundary of the green region, but does not itself belong to $\mathcal{C}_{AF}$. Consequently, its distance from $\mathcal{C}_{AF}$ is zero, while this infimum is not attained by any arbitrage-free call surface. This is possible because $\mathcal{C}_{AF}$ is not closed in the topology of uniform convergence on compact subsets, and therefore is not closed in any weaker metrizable topology considered in Theorem 5.1.

5.2 Neural Operators Admit Arbitrage

Our last set of results show that standard NO builds cannot generally be certified to be free of arbitrage. We consider encoder-decoder type builds where decoding is done by projecting onto basis functions; and then we consider the FNO build.

Encoder-Decoder Architectures We consider encoder-decoder NO builds; cf.[GKL26, Lan23].

Proposition 5.2 (Failure of strike-convexity certificate for decoder-type NOs)**.** *Fix $F, M \in \mathbb{N}_+$, fix pointwise-defined representatives $(e_n)_{n=1}^\infty$ of an orthonormal basis of $L^2([0,1])$, and define*

$$E_{n,m}(T,K) \overset{\text{def.}}{=} e_n(K)e_m(T),$$

for any $(T,K) \in [0,1]^2$. Fix any activation function $\sigma \in C(\mathbb{R})$, fix pairwise distinct indices $n_1,n_2,n_3 \in \mathbb{N}$ and some $m_\star \in \mathbb{N}$, and consider the decoder

$$f_{D\text{-}NO}(x) \overset{\text{def.}}{=} \sum_{k=1}^{3} [f_{NN}(x)]_k \, E_{n_k,m_\star}. \tag{39}$$

Then there exists a σ-MLP $f_{NN} : \mathbb{R}^F \to \mathbb{R}^3$ such that: for every $x \in \mathbb{R}^F$

$$f_{D\text{-}NO}(x) \notin \{C|_{[0,1]^2} : C \in \mathcal{C}_{AF}\}.$$

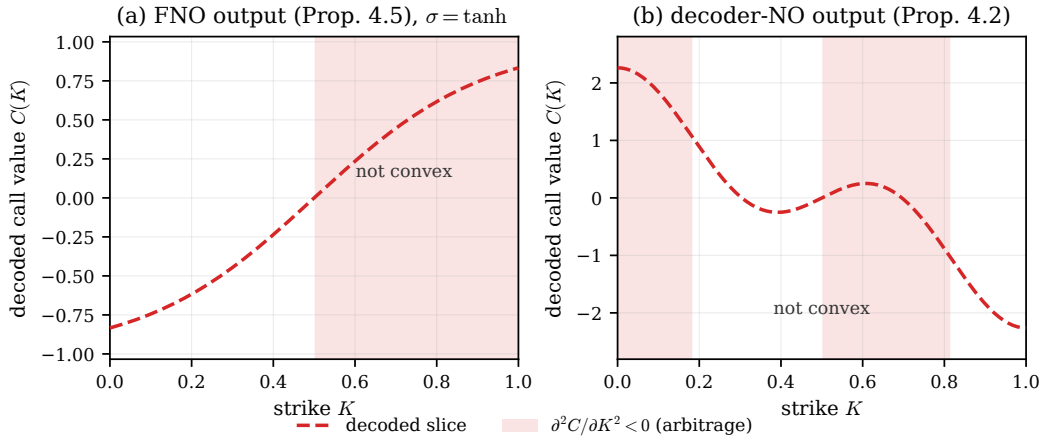


Figure 5: Standard neural operators are not certifiable. Left: a scalar Fourier neural operator with an activation whose second derivative changes sign, as in Proposition 5.4, produces a non-convex strike slice. Right: a decoder-type neural operator on a tensorized basis, as in Proposition 5.2, does the same. The shaded regions mark violations of strike convexity.

Fourier Neural Operators (FNOs) We also consider the mainstream Fourier Neural Operator (FNO) architecture [LKA+21, LHLA23], e.g. considered in [WJG25b] for call surface smoothing, are *not certifiably arbitrage-free*: their parameter spaces contain configurations whose outputs violate strike convexity. Thus, without an additional architectural restriction or certified decoder, the FNO architecture itself cannot guarantee no-arbitrage. Before stating our *un-certifiability* result, we recall the definition of the FNO architecture.

Definition 5.3 (Fourier Neural Operator)**.** Let $D \overset{\text{def.}}{=} [0,1]^2 \subseteq \mathbb{R}^2$ be a rectangular domain, let $d_a, d_u, m \in \mathbb{N}_+$, and let $L \in \mathbb{N}_+$. Fix a pointwise *lifting* map $Q : \mathbb{R}^{d_a} \to \mathbb{R}^m$, a pointwise *projection* map $P : \mathbb{R}^m \to \mathbb{R}^{d_u}$, linear maps $W_\ell : \mathbb{R}^m \to \mathbb{R}^m$ for $\ell \in \{0, \dots, L-1\}$, Fourier multipliers R_ℓ acting on finitely many Fourier modes, and an activation function $\sigma : \mathbb{R} \to \mathbb{R}$ acting componentwise on $\mathbb{R}^m$. Given an input field $a : D \to \mathbb{R}^{d_a}$, define the latent fields $(v_\ell)_{\ell=0}^{L}$ recursively by

$$v_0(x) \overset{\text{def.}}{=} Q(a(x)) \tag{40}$$

and, for $\ell \in \{0, \dots, L-1\}$, for each $x \in D$ by

$$v_{\ell+1}(x) \overset{\text{def.}}{=} \sigma\Big(W_\ell v_\ell(x) + \mathcal{F}^{-1}\big(R_\ell \cdot \mathcal{F}(v_\ell)\big)(x) + b_\ell(x)\Big), \tag{41}$$

where $\mathcal{F}$ denotes the Fourier transform on D and $b_\ell : D \to \mathbb{R}^m$ is a bias field. The associated Fourier Neural Operator is the map $\Phi : [L^2(D)]^{d_a} \to [L^2(D)]^{d_u}$ sending any input field $a \in [L^2(D)]^{d_a}$ to $\Phi(a)(x) \overset{\text{def.}}{=} P(v_L(x))$; where $x \in D$.

The limitations of the FNO for predictions of European call-price surfaces which are strictly free of arbitrage is illustrated by Figure 5 plots the explicit parameter configurations used in Propositions 5.2 and 5.4. These are not training failures. They are fixed admissible parameter choices of the corresponding architectures whose decoded strike slices are non-convex, and therefore contain butterfly arbitrage for every input.

Proposition 5.4 (FNO architecture is not certifiably arbitrage-free)**.** *Let $\sigma \in C^2(\mathbb{R})$ and suppose that there exist $\xi \in \mathbb{R}$ and $\eta > 0$ such that σ'' takes both positive and negative values on $(\xi-\eta, \xi+\eta)$. Consider the scalar FNO architecture of Definition 5.3 on $D = [0,1]^2$ with $d_a = d_u = m = 1$ and pointwise lifting and projection $Q = P = \mathrm{Id}$. Let $Q^\uparrow : \mathbb{R}^F \to L^2([0,1]^2)$ be any injective linear map.*

*If $L \geq 1$, then there is a choice of parameters $(W_\ell, R_\ell, b_\ell)_{\ell=0}^{L-1}$ such that, for every $a \in \mathbb{R}^F$, the continuous representative of $\Phi(Q^\uparrow(a))$ is **not** the restriction to $[0,1]^2$ of an element of $\mathcal{C}_{AF}$.*

Table 1: Accuracy and static-arbitrage compliance on the rough-Bergomi test set. The first two rows are AFNO instances and are arbitrage-free by construction. Some direct baselines improve pointwise implied-volatility accuracy, but their price surfaces are not certified and become arbitrageable under the off-grid and LP diagnostics of Figures 7-8. LP profit is the mean over the 256-surface dense-grid diagnostic; definitions and fuller summaries are in Appendix B.

Model	IV MAE (bps)	IV p95 (bps)	Arb. surf. (%)	Max. arb.	LP profit
SANOS-DLV (ours)	14.2	59.6	0.00	0	4.80×10^{-3}
AFNO mixture	35.9	162	0.00	0	4.89×10^{-3}
Direct-IV	5.58	17.7	0.0[†]	8.91×10^{-9}	1.26×10^{-2}
Direct-Price	161	2.05×10^{3}	100	4.51	2.64×10^{-1}
Direct-Price+arb	259	2.51×10^{3}	99.8	3.43	2.86×10^{-1}
Per-surface SANOS fit	40.1	143	0.00	0	-
rBergomi reference	-	-	-	-	5.23×10^{-3}

[†]Canonical-grid diagnostic only; the dense-grid violations of the Direct-IV head are shown in Figure 7. LP profit denotes the per-surface extractable profit of Figure 8, in units of forward notional.

6 Numerical Illustration

We illustrate the AFNO construction on the parameter-to-surface maps generated by two stochastic-volatility families, Heston and rough Bergomi with $H < 1/2$, both viewed as instances of the solution operator (38). For each family, parameters are sampled from a compact set, a reference European-call surface is computed, and the learned operator maps model parameters to the corresponding call-price surface. We compare two instances of Definition 3.4: the single-decoder SANOS-DLV instance, for which $K = 1$ and the network predicts the decoder parameters, and a fixed-anchor AFNO mixture, for which the network learns only the softmax weights over a finite dictionary of arbitrage-free decoder outputs. The baselines are standard decoder-type neural operators with, respectively, a direct implied-volatility head, a direct price head, and a direct price head trained with a soft no-arbitrage penalty. Appendix B gives the complete data-generation, pricing, architecture, optimization, diagnostic, and reproducibility protocol, together with the Heston results.

Accuracy is reported as implied-volatility error in basis points on the liquid 10Δ-90Δ band. Static-arbitrage diagnostics are computed on call prices. On a non-uniform strike grid $K_1 < \cdots < K_L$, we use the grid-aware butterfly value

$$b_\ell(T) \overset{\text{def.}}{=} \frac{K_{\ell+1} - K_\ell}{K_{\ell+1} - K_{\ell-1}} C(T, K_{\ell-1}) - C(T, K_\ell) + \frac{K_\ell - K_{\ell-1}}{K_{\ell+1} - K_{\ell-1}} C(T, K_{\ell+1}),$$

for $2 \le \ell \le L - 1$; non-negativity of all $b_\ell(T)$ is precisely the discrete convexity condition on this grid. This avoids the equal-spacing second difference, which is not a valid arbitrage diagnostic on delta-based strike grids.

Learning from irregular observation grids. Figure 6 tests the universality claim of Theorem 4.2 and Corollary 4.5 in the data regime most relevant for market surfaces: each training sample is observed on its own random strike grid. The SANOS-DLV instance accepts these grids directly and its canonical-grid median error is essentially unchanged, increasing by 1.6%. Models with a fixed output lattice, including the fixed-anchor AFNO mixture and the direct heads, must first interpolate the targets onto a canonical grid and incur a 1.7–6.3 times increase in median error.

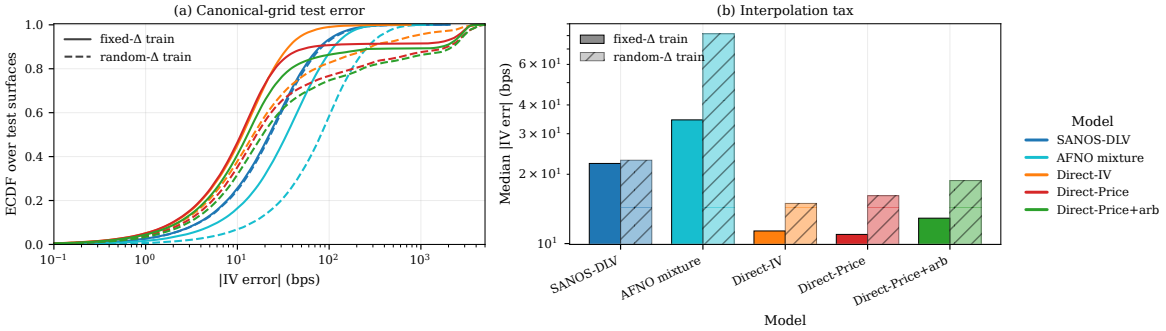
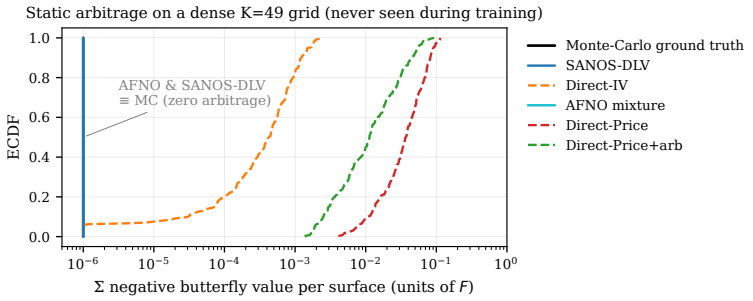


Figure 6: Grid-agnostic training. Each model is trained once on a fixed canonical strike grid and once on per-sample random strike grids, then evaluated on the same canonical test set. Panel (a) reports the empirical distribution of absolute implied-volatility errors; panel (b) reports the median error. The single-decoder AFNO, SANOS-DLV, ingests random grids without first imposing a lattice, while fixed-output models incur the interpolation cost.

Off-grid emission and the arbitrage certificate. Figure 7 queries the trained operators on a dense strike grid not used during training. The two AFNO instances retain non-negative butterfly values, as required by Theorem 4.1, because the decoder produces an element of $\mathcal{C}_{AF}$ at every query point. By contrast, the direct price heads develop negative butterfly values between training strikes. The Direct-IV head has no canonical-grid violations in Table 1, but the dense-grid diagnostic exposes small and persistent off-grid convexity violations.



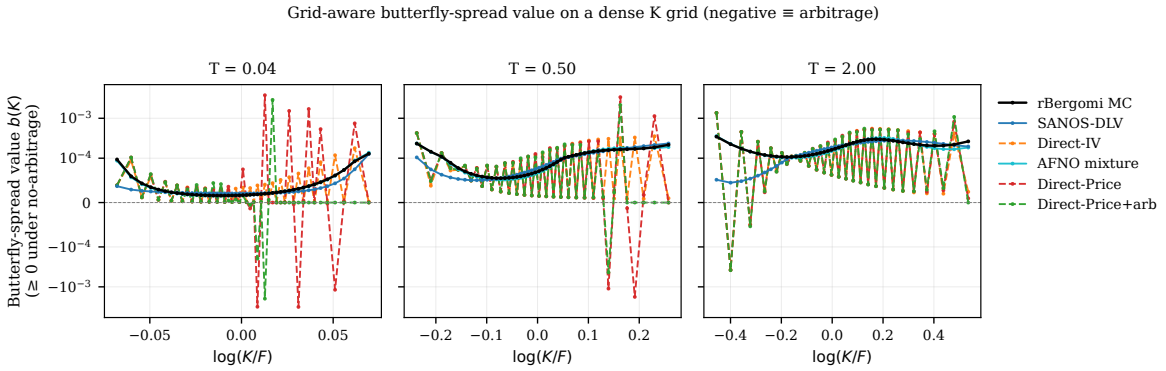
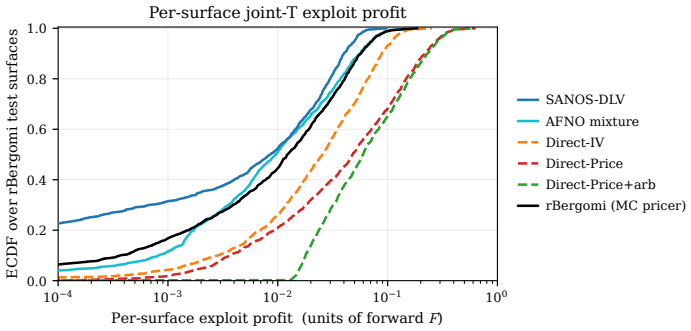


Figure 7: Emitting irregular data while preserving no-arbitrage. All operators are queried on a dense $K = 49$ grid unseen during training and scored by the grid-aware butterfly value $b_\ell(T)$. Top: at three maturities, SANOS-DLV and the AFNO mixture track the reference pricer through non-negative butterfly curves, whereas the direct price heads produce negative values. Bottom: empirical distribution of the total negative butterfly value per surface; the AFNO instances and the reference pricer concentrate at zero.

Figure 8: Economic magnitude of static-arbitrage violations. Each rough-Bergomi test surface is passed to a bounded scenario LP; positive values indicate negative-cost portfolios with nonnegative payoff on every tested path. The AFNO instances are comparable with the finite-scenario Monte Carlo reference level, whereas the direct heads are shifted to substantially larger profits. Figure 9 gives the corresponding Heston and rough-Bergomi grid of diagnostics.



Economic size of residual violations. Figure 8 measures the economic value of the remaining violations. For each quoted surface, we solve a bounded scenario-based static-arbitrage linear program; the objective value is the riskless profit extractable from that surface on the tested scenarios, in units of forward notional. The AFNO outputs remain close to the reference-pricer range, while the direct price heads have mean profits of 26%–29% and maxima near 70% of forward notional. The particular soft lattice penalty used here does not remove this defect. This is consistent with, but is not by itself a proof of, the structural distinction in Theorem 5.1; the exact LP and penalty are specified in Appendix B.

Comparison scope. The controlled baselines share the same input normalization and MLP backbone, isolating the effect of the output parameterization and the soft penalty. We do not include SVI or related per-surface smile calibrators: those methods solve a different, non-amortized calibration/interpolation problem, and their no-arbitrage guarantees depend on additional cross-slice constraints. The per-surface SANOS fit in Table 1 provides the closest like-for-like calibration baseline. We therefore make no claim here that AFNO dominates all classical surface calibrators; the experiment tests whether an amortized parameter-to-surface operator retains accuracy while carrying a global architectural certificate.

7 Conclusion

We introduced AFNO, an encoder-processor-decoder framework for learning maps into the space of arbitrage-free European call-price surfaces. No-arbitrage is imposed structurally by the decoder rather than encouraged through training penalties or repaired after prediction. Consequently, every parameter configuration produces a globally arbitrage-free call surface (Theorem 4.1). This guarantee is available with both an explicit piecewise-linear decoder (Proposition 3.9) and a smooth Black–Scholes-based decoder (Proposition 3.11).

From a financial perspective, the resulting model may be queried at maturities and strikes that were not present during training without losing the static no-arbitrage conditions required for consistent pricing, interpolation, and downstream risk calculations. At the same time, architectural certification does not sacrifice expressivity: AFNOs approximate Lipschitz maps into uniformly controlled classes of arbitrage-free call surfaces (Theorem 4.2). In particular, they approximate parameter-to-call-surface maps generated by broad parametric families of risk-neutral martingale models (Corollary 4.5).

The negative results explain why this structural approach is necessary. A closest arbitrage-free correction need not exist for a surface produced by an unconstrained model (Theorem 5.1), while standard decoder-type neural operators (Proposition 5.2) and Fourier neural operators (Proposition 5.4) admit parameter configurations producing butterfly arbitrage. The numerical results in Section 6 reflect this distinction: AFNO retains its certificate on unseen grids, whereas unconstrained price and volatility heads may exhibit economically exploitable off-grid violations.

Taken together, the results support a simple modelling principle for financial operator learning: when every predicted surface must remain economically admissible, no-arbitrage should be encoded

in the range of the architecture itself. AFNO provides such a certificate while retaining the flexibility needed for accurate and computationally efficient surface prediction.

Acknowledgements and funding

A. Kratsios acknowledges financial support from the Natural Sciences and Engineering Research Council of Canada (NSERC) through Discovery Grant Nos. RGPIN-2023-04482 and DGECR-2023-00230. We further acknowledge that resources used in the preparation of this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and the industry sponsors of the Vector Institute[12].

A Notation

We use the following notation throughout our manuscipt.

- $\mathbb{N} \overset{\text{def.}}{=} \{0, 1, 2, \dots\}$, $\mathbb{N}_+ \overset{\text{def.}}{=} \{1, 2, \dots\}$, and $\mathbb{N}_2 \overset{\text{def.}}{=} \{n \in \mathbb{N} : n \geq 2\}$.
- For each $n \in \mathbb{N}_+$, $[n] \overset{\text{def.}}{=} \{0, \dots, n\}$ and $[n]_+ \overset{\text{def.}}{=} \{1, \dots, n\}$.
- For each $M \in \mathbb{N}_2$, $\mathbb{R}^M_{\leq} \overset{\text{def.}}{=} \big\{(x_1, \dots, x_M) \in \mathbb{R}^M : x_1 < \cdots < x_M\big\}$.
- For each $n \in \mathbb{N}_+$, $\{e_i\}_{i=1}^n$ denotes the standard orthonormal basis of $\mathbb{R}^n$, $\mathbf{1}_n \overset{\text{def.}}{=} (1, \dots, 1) \in \mathbb{R}^n$, and $I_n \in \mathbb{R}^{n \times n}$ denotes the identity matrix.
- For $x \in \mathbb{R}$, $x^+ \overset{\text{def.}}{=} \max\{x, 0\}$, and for $x, y \in \mathbb{R}$, $x \wedge y \overset{\text{def.}}{=} \min\{x, y\}$.
- If $\mathcal{H}$ is a Hilbert space, then $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and $\| \cdot \|_{\mathcal{H}}$ denote its inner product and norm, respectively.
- Throughout, $\mathcal{D} \subseteq [0, \infty)^2$ is the input domain, $\mathcal{H}$ is an RKHS over $\mathcal{D}$, and $\kappa : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ is its reproducing kernel.
- The variables $T \geq 0$ and $K \geq 0$ denote maturity and strike, respectively, and $C(T, K)$ denotes a European call price.
- $\mathcal{C}_{\text{AF}}$ denotes the space of continuous arbitrage-free European call-price surfaces.
- For $p, \lambda > 0$, $\mathcal{C}_{\text{AF}}^{p,\lambda} \overset{\text{def.}}{=} \mathcal{C}_{\text{AF}} \cap \mathcal{X}_\lambda^p$ is normed by

$$\|C\|_{p,\lambda} \overset{\text{def.}}{=} \sup_{T,K \geq 0} e^{-\lambda T}(1 + K)^p |C(T, K)|$$

  where $\mathcal{X}_\lambda^p \overset{\text{def.}}{=} \big\{C \in C([0, \infty)^2, \mathbb{R}) : \|C\|_{p,\lambda} < \infty\big\}$.
- For $N \in \mathbb{N}_+$, $\Delta_N \overset{\text{def.}}{=} \big\{w \in [0, 1]^N : \sum_{n=1}^N w_n = 1\big\}$ denotes the probability simplex, and $\text{co}(A)$ denotes the convex hull of a set A.
- For a Lipschitz map f, $\text{Lip}(f)$ denotes its optimal Lipschitz constant, and $\text{diam}(A)$ denotes the diameter of a non-empty set A.
- $\text{ReLU}(x) \overset{\text{def.}}{=} \max\{x, 0\}$, $\sigma \bullet x$ denotes component-wise application of σ, and Softmax denotes the usual softmax map.
- For an MLP, $\text{depth}(\cdot)$, $\text{width}(\cdot)$, and $\text{size}(\cdot)$ denote its number of layers, maximal hidden-layer width, and total number of affine parameters, respectively.
- We write $a \lesssim b$ if $a \leq cb$ for a constant $c > 0$ independent of the quantities under consideration.

[12] https://vectorinstitute.ai/partnerships/current-partners/

Table 2: Parameter distributions for the numerical study. Every coordinate is sampled independently and uniformly on the displayed interval.

Family	Parameter	Sampling interval
rough Bergomi	ξ_0	$[0.015, 0.12]$
	H	$[0.05, 0.20]$
	η	$[0.8, 2.5]$
	ρ	$[-0.95, -0.10]$
Heston	v_0	$[0.005, 0.16]$
	κ	$[0.3, 4.0]$
	ϑ	$[0.005, 0.16]$
	σ	$[0.10, 0.80]$
	ρ	$[-0.90, -0.10]$

B Numerical Protocol and Supplementary Results

This appendix specifies the full-profile experiment implemented in the companion repository[13]. All reported quantities below are computed from the saved prediction arrays rather than transcribed into the experiment code. Prices are undiscounted forward prices: the forward and money-market account are both normalized to one, rates are zero, and strikes are reported as K/F. Thus, a call price lies between $(1-K)^+$ and one. Total variance, rather than instantaneous variance or volatility, is the variance argument of every Black–Scholes map.

B.1 Data-generating distributions and grids

For each model family we draw parameters independently and uniformly over the boxes in Table 2. The rough-Bergomi parameter vector is (ξ_0, H, η, ρ), where ξ_0 is the flat forward-variance level, H is the Hurst parameter, η is vol-of-vol, and ρ is spot–volatility correlation. The Heston vector is $(v_0, \kappa, \vartheta, \sigma, \rho)$, with long-run variance denoted by ϑ to distinguish it from the full parameter vector. The Heston box is not restricted by the Feller condition. Each family has 60,000 training, 4,000 validation, and 8,000 test surfaces; no sampled surface was removed by the implemented hygiene filter.

Every surface contains the seven maturities

$$\mathcal{T} = \{2/52, 1/12, 1/4, 1/2, 1, 3/2, 2\} \text{ years}$$

and 25 strikes per maturity. Given a reference volatility σ_{ref} and call delta Δ, the strike is

$$K(T, \Delta; \sigma_{\text{ref}}) = \exp\Big(-\Phi^{-1}(\Delta)\sigma_{\text{ref}}\sqrt{T} + \tfrac{1}{2}\sigma_{\text{ref}}^2 T\Big).$$

The canonical grid uses 25 linearly spaced deltas from 0.98 down to 0.02 and $\sigma_{\text{ref}} = 0.20$; across all maturities its strikes range from 0.5822 to 1.8606. The irregular-grid experiment draws one σ_{ref} per surface log-uniformly on $[0.08, 0.45]$ and uses it at every maturity of that surface. This gives sample-dependent strikes ranging from 0.3314 to 4.5242 in the saved training set. SANOS-DLV is trained directly on these grids. For every fixed-output model, irregular targets are first linearly interpolated in log-strike onto the canonical grid. The fixed- and random-grid rough-Bergomi training sets use the same parameter draws, isolating the effect of the observation grid.

The parameter/data seeds are $0, 1, 2$ for rough-Bergomi train/validation/test splits and $10, 11, 12$ for Heston. The irregular rough-Bergomi train/validation sets again use seeds $0, 1$; the canonical test set with seed 2 is shared by both training regimes. The 256 frozen mixture anchors are selected without replacement from the relevant training set using seed 17.

B.2 Reference prices and numerical inversion

Heston. Reference calls are computed by a Fang–Oosterlee COS expansion for put prices followed by put–call parity. The expansion uses 256 cosine terms and truncation interval $[c_1 - 12\sqrt{c_2}, c_1 +$

[13] `arbitrage-free-neural-operator`

Table 3: Output parameterization and trainable parameter count. The small difference between families is due to the four-dimensional rough-Bergomi versus five-dimensional Heston input.

Model	Output head	rBergomi	Heston
SANOS-DLV	217 positive interior DLV values and seven positive total variances; 33 decoder anchors	322,016	322,272
AFNO mixture	softmax weights over 256 frozen training-surface anchors; temperature one	330,240	330,496
Direct-IV	175 positive implied volatilities, converted pointwise to Black–Scholes prices	309,423	309,679
Direct-Price	175 prices, clipped to $[0, 1]$	309,423	309,679
Direct-Price+arb	same as Direct-Price	309,423	309,679

$12\sqrt{c_2}]$, where the first two log-forward cumulants are recovered from five-point characteristic-function differences with step 10^{-3}. Computations use double precision, and final prices are clipped to the model-independent interval $[(1-K)^+, 1]$.

Rough Bergomi. The Volterra driver is simulated with the $\kappa = 1$ hybrid discretization on 256 uniform time steps over $[0, 2]$. Each surface uses 32,768 paths, processed in four chunks of 8,192, with antithetic Gaussian pairs. Calls are evaluated by conditional Monte Carlo after conditioning on the Brownian driver correlated with spot. Each surface has its own saved Monte Carlo seed. For numerical stability, the implementation clips the log-variance ratio to $[-8, 8]$ and the conditional log-forward to $[-4, 4]$, then clips call prices to $[(1-K)^+, 1]$. The same specification is used when repricing the off-grid targets and mixture anchors.

Implied total variance is recovered from price by 60 bisection steps on log-variance over $[10^{-10}, 30]$, followed by eight safeguarded Newton steps; implied volatility is $\sqrt{v/T}$. Pricing and inversion are performed in double precision, after which the dataset's prices, total variances, and parameters are stored in single precision. The rough-Bergomi prices are Monte Carlo labels, not exact ground truth. We did not attach Monte Carlo confidence intervals to individual quotes; their residual noise is visible in the nonzero LP profit of the reference row in Tables 1 and 6. Likewise, the saved study contains no separate COS- or time-step-convergence sweep, so the stated discretizations are part of the definition of the numerical targets.

B.3 Models, objectives, and optimization

All five operators first map every input coordinate affinely to $[-1, 1]$ and then use a five-hidden-layer, width-256 SiLU MLP. Their output heads differ as follows.

SANOS-DLV uses the strike anchors

$$\{0\} \cup \{\exp(x) : x \text{ is one of 31 equally spaced points in } [-0.975, 0.975]\} \cup \{8\},$$

which include the forward strike exactly. Softplus maps impose a floor of 10^{-5} on interior DLV values and 10^{-6} on total variances. The 256-anchor AFNO mixture learns only simplex weights. Its frozen dictionary consists of training surfaces selected with seed 17; at a new query grid, the anchors are repriced on that grid before applying the learned weights. Direct-IV applies softplus plus 10^{-4} to each output volatility. The direct price heads are clipped pointwise to $[0, 1]$ but have no structural strike or maturity constraint.

The basic loss is the equally weighted price mean-square error over all 7×25 quotes. The penalized baseline minimizes

$$\mathcal{L}_{\text{price}} + 0.003\,\mathcal{P}_{\text{lattice}},$$

where, writing $\langle\cdot\rangle$ for the mean over every valid batch and lattice index,

$$\begin{aligned}\mathcal{P}_{\text{lattice}} = {} & \langle [C_{j,\ell+1} - C_{j,\ell}]_+ \rangle + \langle [-(C_{j,\ell+2} - 2C_{j,\ell+1} + C_{j,\ell})]_+ \rangle \\ & + \langle [C_{j,\ell} - C_{j+1,\ell}]_+ \rangle + \left\langle [(1-K_{j,\ell})^+ - C_{j,\ell}]_+ \right\rangle + \langle [C_{j,\ell} - 1]_+ \rangle .\end{aligned}$$

This is deliberately a soft, lattice-local baseline. In particular, its convexity term is an unweighted second difference rather than the non-uniform-grid butterfly used for evaluation, and equal strike indices at adjacent maturities need not represent equal strikes on a delta grid. The weight 0.003 is

the fixed experiment configuration; the code base does not record a penalty-weight sweep or select it on the test set.

Every model is trained for exactly 80 epochs with AdamW, learning rate 2×10^{-3}, zero weight decay, batch size 512, and gradient-norm clipping at one. A cosine schedule decreases the learning rate to 10^{-4} at epoch 80. Validation metrics are recorded after every epoch, but there is no early stopping or best-epoch selection: evaluation uses the final checkpoint. Network initialization and batch shuffling use seed 2026. Thus the reported tables are one deterministic training run, not a mean over repeated initializations; no training-seed error bars should be inferred from them. The per-surface SANOS fit in Table 1 independently optimizes a 49-anchor decoder on the first 256 rough-Bergomi test surfaces for 1,500 Adam iterations at learning rate 0.05 and seed 77. It is an optimization baseline, not a certified global optimum or a statistical oracle.

B.4 Metrics and arbitrage diagnostics

Price RMSE is computed over all test quotes. Implied-volatility MAE, median, p95, and maximum are computed only where the *target-price-implied* Black–Scholes delta lies in $[0.10, 0.90]$; one volatility basis point is 10^{-4}. For each surface, the canonical-grid severity is the maximum of: a positive adjacent price increment, a decrease in adjacent strike slopes, an intrinsic-value shortfall, and an upper-bound excess. A surface is counted as arbitrageable when severity exceeds 10^{-4}. Because the delta grid has maturity-dependent strikes, the canonical table does not compare calendar prices at unequal strikes; cross-maturity effects are instead probed by the joint LP below. Off-grid butterfly counts use the grid-aware value in Section 6 and tolerance 10^{-6}.

The off-grid sample comprises 256 test parameters drawn without replacement with seed 202604. In addition to the canonical 25-strike grid, the saved evaluation contains: 49 deltas from 0.98 to 0.02 (the dense grid used in Figures 7 and 8); 25 deltas from 0.995 to 0.005; and a 25-strike grid constructed with $\sigma_{\text{ref}} = 0.35$. The dense grid contains $256 \times 7 \times 47 = 84{,}224$ interior butterflies. Neither AFNO instance nor the reference pricer has a violation below -10^{-6} among them; the minimum butterfly values are respectively 4.86×10^{-7}, 5.18×10^{-6}, and 5.00×10^{-7}. Direct-IV has 701 violating butterflies across 73.8% of surfaces, while both direct-price variants violate at least one butterfly on every dense-grid surface. SANOS-DLV and the mixture are evaluated natively on the new grid; direct heads are linearly interpolated in log-strike from their canonical price outputs.

Bounded scenario LP. For a quoted surface, let $z = (z_0, z_S, (z_{j\ell}))$ hold cash, the terminal underlying, and the calls. With quote vector $q = (1, 1, (C(T_j, K_{j\ell})))$, the diagnostic is

$$\Pi(C) = \max_{-1 \le z_i \le 1} \{-q^\top z : Az \ge 0\}, \qquad (Az)_r = z_0 + z_S S_{T_M}^{(r)} + \sum_{j,\ell} z_{j\ell} (S_{T_j}^{(r)} - K_{j\ell})^+.$$

Hence Π is the largest negative initial cost among portfolios whose aggregate payoff is nonnegative on every tested path. The box $[-1, 1]$ applies separately to every instrument, so the reported number is in units of forward notional and scales with this portfolio bound.

The joint test uses 800 as its nominal scenario count. Its deterministic additions give 954 actual paths for seven maturities: 154 one-maturity spot scans, 160 constant paths, 160 one-jump paths, and 480 drift-corrected GBM paths with volatility uniform on $[0.08, 0.65]$. Spot scans cover $\exp([-1.5, 1.5])$; the single-maturity version additionally includes zero, 100, and every strike kink. Scenario randomness uses seed 202604. We solve with SciPy/HiGHS, the solver's default feasibility tolerances, position bound one, and a four-second limit per LP. The implementation records zero if HiGHS does not return success. Rough-Bergomi LP summaries use all 256 dense off-grid surfaces; Heston summaries use the first 1,000 canonical-grid test surfaces. This finite-scenario bounded LP measures economic severity but is not itself a global no-arbitrage certificate.

B.5 Complete tabulations

Table 4 supplies the Heston counterpart of Table 1. The ordering is consistent across families: Direct-IV is most accurate pointwise, the structured models have no detected grid arbitrage, and the unconstrained price heads have large tail errors and economically material LP values.

Table 4: Complete Heston test accuracy and compliance results on 8,000 surfaces. IV errors use the liquid band; LP mean is the joint-maturity bounded profit over 1,000 surfaces.

Model	Price RMSE	IV MAE	IV median (bps)	IV p95	IV max	Arb. surf. (%)	Max. arb.	LP mean
SANOS-DLV (ours)	2.10×10^{-4}	8.39	2.97	35.6	392	0	4.57×10^{-7}	8.81×10^{-4}
AFNO mixture	9.78×10^{-4}	30.5	14.3	112	856	0	0	6.79×10^{-4}
Direct-IV	9.43×10^{-5}	3.05	1.67	10.4	169	0	3.50×10^{-5}	9.36×10^{-4}
Direct-Price	3.63×10^{-3}	191	2.06	2462	3987	99.1	5.31	1.14×10^{-1}
Direct-Price+arb	3.61×10^{-3}	234	3.35	2667	3988	97.5	4.19	1.03×10^{-1}
Heston reference	–	–	–	–	–	–	–	1.02×10^{-3}

Table 5: Effect of per-sample irregular training grids on canonical-grid test accuracy.

Model	Fixed-grid train	Random-grid train	Ratio
	median absolute IV error (bps)		
SANOS-DLV (ours)	5.79	5.88	1.02
AFNO mixture	14.9	92.9	6.25
Direct-IV	3.55	6.53	1.84
Direct-Price	3.93	8.36	2.13
Direct-Price+arb	6.76	11.5	1.70

Table 6: Joint-maturity bounded LP profit in units of forward notional. Rough-Bergomi uses 256 dense-grid surfaces; Heston uses 1,000 canonical-grid surfaces.

	rough Bergomi				Heston			
Model	Mean	Median	p95	Max	Mean	Median	p95	Max
SANOS-DLV (ours)	.00480	.00299	.0122	.0211	.000881	0	.00524	.0240
AFNO mixture	.00489	.00287	.0135	.0219	.000679	0	.00370	.0245
Direct-IV	.0126	.0114	.0204	.0308	.000936	0	.00536	.0283
Direct-Price	.264	.236	.539	.698	.114	.109	.240	.314
Direct-Price+arb	.286	.266	.575	.709	.103	.0916	.213	.278
Numerical reference	.00523	.00295	.0136	.0278	.00102	0	.00576	.0397

The irregular-input comparison is shown numerically in Table 5. These are median absolute implied-volatility errors on the same canonical rough-Bergomi test set, so the ratio isolates the training-grid change. Only SANOS-DLV consumes the random grids natively.

Table 6 reports the full joint-LP distribution underlying the mean column in the main text. The nonzero reference levels quantify the diagnostic's numerical/scenario floor, especially for Monte Carlo rough-Bergomi prices.

Figure 9 complements Figure 8 with the maturity-by-maturity and joint distributions. In the Heston per-maturity LP, SANOS-DLV, the mixture, and the Heston reference are zero to reported precision; Direct-IV has mean 5.31×10^{-11} and maximum 4.17×10^{-8}. The corresponding means for Direct-Price and Direct-Price+arb are 0.0875 and 0.0632.

B.6 Runtime and reproducibility

The saved manifest records Python 3.12.0 on macOS 15.2 arm64, with NumPy 2.5.1, SciPy 1.18.0, PyTorch 2.13.0, Matplotlib 3.11.1, pandas 3.0.3, and joblib 1.5.3. It does not record the processor model, so the timings below are indicative rather than hardware-normalized. Table 7 reports wall time from each final training history. The entire 15-model training stage (fixed-grid rough Bergomi, random-grid rough Bergomi, and Heston) took 11,795 seconds. For inference, we loaded the saved fixed-grid rough-Bergomi checkpoints, used one PyTorch CPU thread and batches of 1,024, warmed up once, timed five passes over all 8,000 test surfaces, and report the median.

From a clean `afno-code` checkout, the full pipeline is invoked by

```
uv run afno-reproduce --profile full.
```

It generates data and anchors, trains all models, evaluates canonical and off-grid predictions, fits the per-surface decoder, solves the LPs, renders figures, and finally writes `artifacts/manifest.json`.

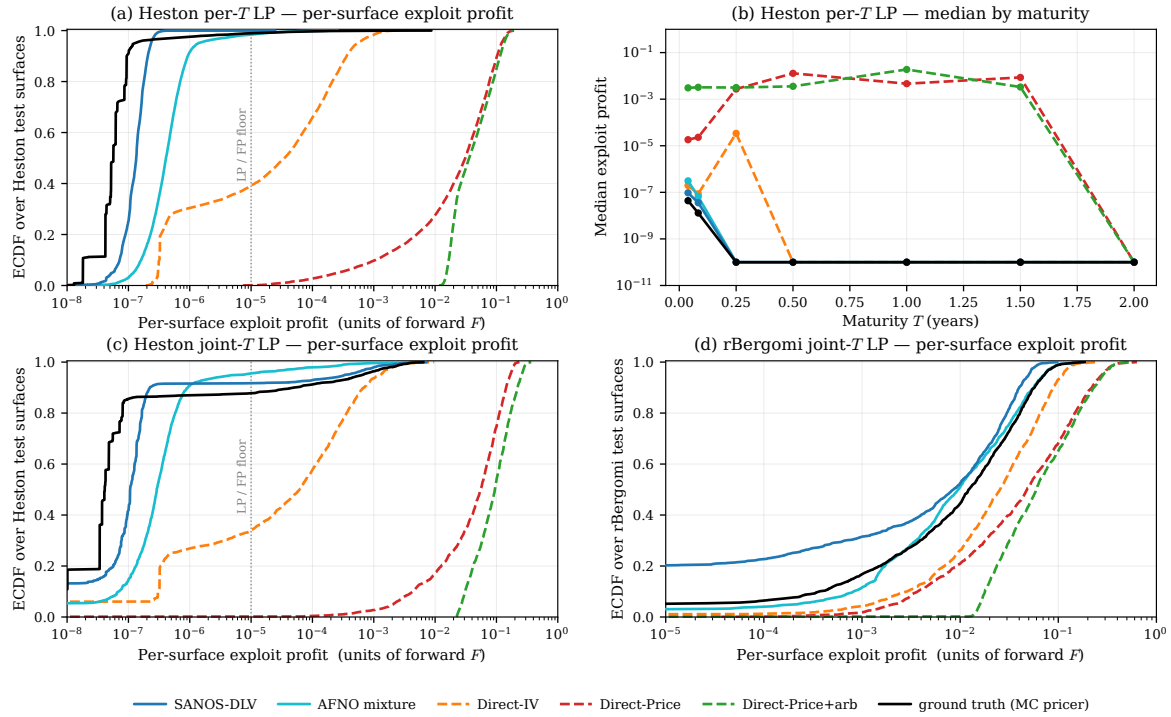


Figure 9: Full static-arbitrage LP diagnostic. The panels report per-surface exploit-profit distributions and maturity-wise medians for Heston and rough-Bergomi test sets. AFNO instances stay close to the numerical reference level; direct decoder heads remain materially arbitrageable.

Table 7: Recorded 80-epoch training time and batched CPU inference time. Inference produces all 7×25 quotes for one surface.

Model	rBergomi train (min)	Heston train (min)	Inference (μs/surface)
SANOS-DLV (ours)	34.9	35.9	89.2
AFNO mixture	6.91	6.95	2.24
Direct-IV	8.06	7.91	3.39
Direct-Price	6.82	6.93	1.78
Direct-Price+arb	7.93	8.06	1.84

The manifest stores the full profile, global seed, package versions, stage times, SHA-256 hashes of every artifact, and the minimum dense-grid butterfly of each certified model. Pipeline stages are cached and resumable; rough-Bergomi generation checkpoints the completed-surface mask and rejects a checkpoint whose simulation configuration differs. The included test suite is run with `uv run pytest`. A reduced `--profile quick` is provided only as a smoke test and is not the source of any number in this paper.

C Proofs

We now prove our main results.

C.1 Auxiliary Results

Proof of Proposition 2.11. Fix $T \geq 0$ and $K > 0$. Since $\mathbb{E}[|S_T|^q] < \infty$, we also have $\mathbb{E}[(S_T)_+^q] < \infty$. We first establish the required estimate pointwise. For $x \leq K$, one has $(x-K)_+ = 0$. Suppose therefore that $x > K$. In particular, $x > 0$, and, upon setting $y \overset{\text{def.}}{=} x/K > 1$, we obtain $(x-K)_+ = K(y-1) = K^{-(q-1)} x^q \frac{y-1}{y^q}$. Consider $g : (1,\infty) \to \mathbb{R}$ given by $g(y) \overset{\text{def.}}{=} (y-1)y^{-q}$. Its derivative is $g'(y) = y^{-q-1}\big(q - (q-1)y\big)$. Since $q > 1$, it follows that g is increasing on

$(1, q/(q-1))$ and decreasing on $(q/(q-1), \infty)$. Hence g attains its maximum at $y = q/(q-1)$, where $g\left(\frac{q}{q-1}\right) = \frac{(q-1)^{q-1}}{q^q}$. Consequently, for every $x \in \mathbb{R}$, $(x-K)_+ \leq \frac{(q-1)^{q-1}}{q^q} K^{-(q-1)}(x_+)^q$. Applying this deterministic inequality with $x = S_T(\omega)$ gives $(S_T - K)_+ \leq \frac{(q-1)^{q-1}}{q^q} K^{-(q-1)}(S_T)_+^q$ $\mathbb{P}$-a.s. The random variable on the right-hand side is integrable because $(S_T)_+^q \leq |S_T|^q$ and $\mathbb{E}[|S_T|^q] < \infty$. Taking expectations therefore yields (10). Finally, set $p \overset{\text{def.}}{=} q - 1 > 0$. Then (10) becomes $\mathbb{E}\big[(S_T - K)_+\big] \leq C_{T,p} K^{-p}$ and $C_{T,p} \overset{\text{def.}}{=} \frac{(q-1)^{q-1}}{q^q} \mathbb{E}\big[(S_T)_+^q\big] < \infty$; which is precisely (9). □

Proof of Proposition 2.7. Consider the "bad" call surface $C^\star \in C([0,\infty)^2, [0,\infty))$ defined for every $T, K \geq 0$ by

$$C^\star(T,K) \overset{\text{def.}}{=} (1-K)_+.$$

First, observe that call surface generated by the deterministic strictly-positive martingale $Z_T \equiv 1$. Hence $C^\star \in \mathcal{C}_{\mathrm{AF}}$. However, $\widetilde{C} \overset{\text{def.}}{=} C^\star + C^\star = 2(1-K)_+.$, and so, $\widetilde{C}(T,0) = 2 \neq 1$ for all $T \geq 0.$; violating Theorem 2.5 (i). Let $\lambda = 0$ and define $\widehat{C} \overset{\text{def.}}{=} 0C^\star = 0$; which violates Theorem 2.5 (i). Finally, consider $\overline{C} \overset{\text{def.}}{=} C^\star C^\star = (1-K)_+^2$ and observe that, for $K \in [0,1)$, $\overline{C}(T,K) = (1-K)^2,$. Whence $\partial_K \overline{C}(T,0) = -2.$ and Theorem 2.5 (ii) is violated. □

Proof of Proposition 2.8. Let $C_1, C_2 \in \mathcal{C}_{\mathrm{AF}}$ and let $\alpha \in [0,1]$. We prove that $\alpha C_1 + (1-\alpha)C_2 \in \mathcal{C}_{\mathrm{AF}}$ directly from Definition 2.4. For each $i \in \{1,2\}$, since $C_i \in \mathcal{C}_{\mathrm{AF}}$, there exists a filtered probability space $(\Omega_i, \mathcal{F}^i, (\mathcal{F}_T^i)_{T\geq 0}, \mathbb{P}_i)$ and a non-negative $(\mathbb{P}_i, \mathcal{F}_T^i)$-martingale $Z^i = (Z_T^i)_{T\geq 0}$ with $Z_0^i = 1$ a.s. such that $C_i(T,K) = \mathbb{E}_{\mathbb{P}_i}\big[(Z_T^i - K)^+\big]$ for every $T, K \geq 0$. We now construct a filtered probability space which randomly selects one of these two martingale models. Consider the disjoint union $\Omega \overset{\text{def.}}{=} \big(\{1\} \times \Omega_1\big) \sqcup \big(\{2\} \times \Omega_2\big)$, define the probability measure $\mathbb{P}$ on Ω by

$$\mathbb{P}\big(\{1\} \times A_1\big) \overset{\text{def.}}{=} \alpha\, \mathbb{P}_1(A_1) \text{ and } \mathbb{P}\big(\{2\} \times A_2\big) \overset{\text{def.}}{=} (1-\alpha)\, \mathbb{P}_2(A_2),$$

for $A_i \in \mathcal{F}^i$, and for every $T \geq 0$, define the terminal time σ-algebra by

$$\mathcal{F}_T \overset{\text{def.}}{=} \Big\{ \big(\{1\} \times A_1\big) \cup \big(\{2\} \times A_2\big) : A_i \in \mathcal{F}_T^i,\ i = 1,2 \Big\}.$$

Then $(\mathcal{F}_T)_{T\geq 0}$ is a filtration on Ω. Define $Z = (Z_T)_{T\geq 0}$ by $Z_T(i,\omega_i) \overset{\text{def.}}{=} Z_T^i(\omega_i)$ for each $i \in \{1,2\}$. The process Z is nonnegative and satisfies $Z_0 = 1$ a.s. We verify that Z is a $(\mathbb{P}, \mathcal{F}_T)$-martingale. Let $0 \leq S \leq T$ and let $A = \big(\{1\} \times A_1\big) \cup \big(\{2\} \times A_2\big) \in \mathcal{F}_S$, where $A_i \in \mathcal{F}_S^i$. Since each Z^i is a martingale, we find that

$$\begin{aligned}
\int_A Z_T \, d\mathbb{P} &= \alpha \int_{A_1} Z_T^1 \, d\mathbb{P}_1 + (1-\alpha) \int_{A_2} Z_T^2 \, d\mathbb{P}_2 \\
&= \alpha \int_{A_1} Z_S^1 \, d\mathbb{P}_1 + (1-\alpha) \int_{A_2} Z_S^2 \, d\mathbb{P}_2 \\
&= \int_A Z_S \, d\mathbb{P}.
\end{aligned}$$

Hence $\mathbb{E}_{\mathbb{P}}[Z_T \mid \mathcal{F}_S] = Z_S$; whence, Z is a nonnegative martingale.
It only remains to compute the expectation of $(Z_t - K)^+$ under the mixture laws; indeed, for every $T, K \geq 0$, we compute

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}}\big[(Z_T - K)^+\big] &= \alpha \mathbb{E}_{\mathbb{P}_1}\big[(Z_T^1 - K)^+\big] + (1-\alpha)\mathbb{E}_{\mathbb{P}_2}\big[(Z_T^2 - K)^+\big] \\
&= \alpha C_1(T,K) + (1-\alpha) C_2(T,K).
\end{aligned}$$

Thus $\alpha C_1 + (1-\alpha)C_2$ admits a martingale representation in the sense of Definition 2.4. Hence $\alpha C_1 + (1-\alpha)C_2 \in \mathcal{C}_{\mathrm{AF}}$, proving that $\mathcal{C}_{\mathrm{AF}}$ is convex. □

Proof of Proposition 2.16. Use normalized spot $S_0 = 1$, fix $T > 0$, abbreviate $\phi_T \overset{\text{def.}}{=} \varphi(\theta_T)$, and set $\beta_T \overset{\text{def.}}{=} \frac{1+\rho}{2}\theta_T \phi_T$. Since $\theta_T > 0$, $\phi_T > 0$, and $\rho \in (-1,1)$, we have $\beta_T > 0$. Moreover, since $p > 0$, $2(\sqrt{p+1} - \sqrt{p})^2 < 2$, and hence the assumed inequality $\beta_T \leq 2(\sqrt{p+1} - \sqrt{p})^2$ implies $\beta_T < 2$.
We first identify the right-wing asymptotics of the SSVI total variance. Recall that

$$w(k, \theta_T) = \frac{\theta_T}{2}\left(1 + \rho\phi_T k + \sqrt{(\phi_T k + \rho)^2 + 1 - \rho^2}\right).$$

This is consistent with the asymptotic statement in [GJ14, Remark 4.3], but we record the required remainder explicitly. Since $\phi_T > 0$, there exists $k_T^{(0)} > 0$ such that $\phi_T k + \rho > 0$ whenever $k \geq k_T^{(0)}$. For such k, rationalization gives

$$\sqrt{(\phi_T k + \rho)^2 + 1 - \rho^2} - (\phi_T k + \rho) = \frac{1 - \rho^2}{\sqrt{(\phi_T k + \rho)^2 + 1 - \rho^2} + \phi_T k + \rho}.$$

The denominator on the right-hand side is of order k and is bounded from below by a positive multiple of k for all sufficiently large k. Consequently, $\sqrt{(\phi_T k + \rho)^2 + 1 - \rho^2} = \phi_T k + \rho + O(k^{-1})$ as $k \to \infty$. Substituting this identity into the SSVI formula yields

$$w(k, \theta_T) = \beta_T k + \frac{1 + \rho}{2}\theta_T + O(k^{-1}), \qquad k \to \infty.$$

In particular, $w(k, \theta_T) = \beta_T k + O(1)$ and $w(k, \theta_T)/k \to \beta_T$ as $k \to \infty$. Write, as usual, $d_\pm(k) = -k/\sqrt{w(k, \theta_T)} \pm \sqrt{w(k, \theta_T)}/2$. For every k for which $w(k, \theta_T) > 0$, direct algebra gives

$$\frac{d_+(k)^2}{2} = \frac{(2k - w(k, \theta_T))^2}{8w(k, \theta_T)} = \frac{k^2}{2w(k, \theta_T)} - \frac{k}{2} + \frac{w(k, \theta_T)}{8}.$$

Since $w(k, \theta_T) = \beta_T k + O(1)$ and $\beta_T > 0$, we also have $w(k, \theta_T)^{-1} = (\beta_T k)^{-1} + O(k^{-2})$ as $k \to \infty$. Therefore,

$$\frac{d_+(k)^2}{2} = \left(\frac{1}{2\beta_T} - \frac{1}{2} + \frac{\beta_T}{8}\right) k + O(1) = \frac{(2 - \beta_T)^2}{8\beta_T}\, k + O(1).$$

Define $q_T \overset{\text{def.}}{=} (2 - \beta_T)^2/(8\beta_T)$. Thus there exist constants $M_T > 0$ and $k_T^{(1)} > 0$ such that

$$\left| \frac{d_+(k)^2}{2} - q_T k \right| \leq M_T, \qquad k \geq k_T^{(1)}.$$

Notice also that the corresponding expression involving d_- is not merely asymptotically equal to the preceding one: it is identical to it. Indeed,

$$\frac{d_-(k)^2}{2} - k = \frac{(2k + w(k, \theta_T))^2}{8w(k, \theta_T)} - k = \frac{(2k - w(k, \theta_T))^2}{8w(k, \theta_T)} = \frac{d_+(k)^2}{2}.$$

Furthermore,

$$d_+(k) = \frac{w(k, \theta_T) - 2k}{2\sqrt{w(k, \theta_T)}} = -\frac{2 - \beta_T}{2\sqrt{\beta_T}}\sqrt{k} + O(k^{-1/2}), \qquad k \to \infty.$$

Since $\beta_T < 2$, it follows that $d_+(k) \to -\infty$. Hence, after increasing $k_T^{(1)}$ if necessary, we may assume that $d_+(k) \leq -1$ for every $k \geq k_T^{(1)}$.
Let $\mathcal{N}$ denote the standard Gaussian distribution function. For every $x > 0$, the elementary Gaussian tail estimate

$$\mathcal{N}(-x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-u^2/2}\, du \leq \frac{1}{x\sqrt{2\pi}} \int_x^\infty u e^{-u^2/2}\, du = \frac{1}{x\sqrt{2\pi}} e^{-x^2/2}$$

holds. In particular, if $x \geq 1$, then $\mathcal{N}(-x) \leq (2\pi)^{-1/2} e^{-x^2/2}$.
Now let $K = S_0 e^k$. By the Black–Scholes representation of the SSVI call price,

$$C_{\mathrm{SSVI}}(T, K) = S_0 \left(\mathcal{N}(d_+(k)) - e^k \mathcal{N}(d_-(k)) \right) \leq S_0 \mathcal{N}(d_+(k)).$$

Therefore, for every $k \geq k_T^{(1)}$, $C_{\mathrm{SSVI}}(T, S_0 e^k) \leq \frac{S_0}{\sqrt{2\pi}} \exp\left(-\frac{d_+(k)^2}{2}\right) \leq \frac{S_0 e^{M_T}}{\sqrt{2\pi}} e^{-q_T k}$.. Thus the SSVI call price has right-wing decay of order at least $e^{-q_T k}$.
It remains to compare q_T with p. Since $0 < \beta_T < 2$,

$$q_T \geq p \quad \Longleftrightarrow \quad (2 - \beta_T)^2 \geq 8p\beta_T \quad \Longleftrightarrow \quad \beta_T^2 - 4(2p + 1)\beta_T + 4 \geq 0.$$

The two roots of the polynomial on the right are $2(\sqrt{p+1}-\sqrt{p})^2$ and $2(\sqrt{p+1}+\sqrt{p})^2$. The second root is strictly larger than 2, while $\beta_T < 2$. Consequently,

$$q_T \geq p \quad \Longleftrightarrow \quad \beta_T \leq 2(\sqrt{p+1}-\sqrt{p})^2.$$

Hence the assumption of the proposition implies $q_T \geq p$. Set $k_T^* \overset{\text{def.}}{=} \max\{k_T^{(1)}, 0\}$ and $K_T^* \overset{\text{def.}}{=} S_0 e^{k_T^*}$. For $K \geq K_T^*$, writing $k = \log(K/S_0) \geq 0$ and using $q_T \geq p$ gives

$$C_{\mathrm{SSVI}}(T,K) \leq \frac{S_0 e^{M_T}}{\sqrt{2\pi}} e^{-q_T k} \leq \frac{S_0 e^{M_T}}{\sqrt{2\pi}} e^{-pk} = \frac{S_0^{p+1} e^{M_T}}{\sqrt{2\pi}} K^{-p}.$$

For the remaining strikes $0 < K < K_T^*$, the Black–Scholes call price satisfies $0 \leq C_{\mathrm{SSVI}}(T,K) \leq S_0$. Since $K < K_T^*$ and $p > 0$,

$$C_{\mathrm{SSVI}}(T,K) \leq S_0 \leq S_0 (K_T^*)^p K^{-p}.$$

Consequently, with

$$C_{T,p} \overset{\text{def.}}{=} \max\left\{ \frac{S_0^{p+1} e^{M_T}}{\sqrt{2\pi}},\, S_0 (K_T^*)^p \right\},$$

we obtain

$$C_{\mathrm{SSVI}}(T,K) \leq C_{T,p} K^{-p}, \qquad K > 0.$$

Since $T > 0$ was arbitrary, (9) holds at every positive maturity. Finally, at $T = 0$, $C_{\mathrm{SSVI}}(0,K) = (S_0 - K)^+$. For $K \geq S_0$ this quantity vanishes, whereas for $0 < K < S_0$ we have $(S_0 - K)^+ \leq S_0 \leq S_0^{p+1} K^{-p}$. Thus (9) also holds at $T = 0$, which completes the proof. □

C.2 Well-Posedness of the Solution Map and Universal Arbitrage-Free Pricing

Proof of Proposition 4.4. For each $\theta \in \Theta$, put $Z_t^\theta = e^{-rt} X_t^\theta$ and $g_t^\theta = X_t^\theta \sigma(t, \nu_t, \theta, X_t^\theta)$. Integration by parts gives

$$Z_t^\theta = 1 + \int_0^t Z_s^\theta \sigma(s, \nu_s, \theta, X_s^\theta)\, dM_s.$$

For $\theta_1, \theta_2 \in \Theta$, write $D_t = X_t^{\theta_2} - X_t^{\theta_1}$. Then

$$D_t = \int_0^t r D_s\, ds + \int_0^t (g_s^{\theta_2} - g_s^{\theta_1})\, dM_s, \qquad |g_s^{\theta_2} - g_s^{\theta_1}| \leq L_x |D_s| + L_\theta |X_s^{\theta_1}| \|\theta_2 - \theta_1\|_2.$$

The first identity makes Z^θ a positive local martingale, with the stochastic integrals understood in the predictable sense required for the stated equation. Square integrability of M alone does not ensure that Z^θ is a true martingale, justify the required BDG–Grönwall moment and stability estimates, or ensure continuity of the call prices in maturity. Thus the stated hypotheses do not complete this proof; the well-posedness and pricing conclusions remain pending the separately proposed assumption repair. □

Proof of Corollary 4.5. Fix $0 < \varepsilon < 1$ and $H > 0$, and set $\eta \overset{\text{def.}}{=} \varepsilon e^{-lH} \in (0,1)$. By Proposition 4.4, the class $\mathcal{Y} = \mathrm{Sol}(\Theta)$ is uniformly controlled with $B = 2^p$ and common modulus $\omega(h) = h + \max\{1, \sqrt{e^a - 1}\}\sqrt{h}$. Moreover, $l > q\lambda/p > \frac{a}{2} q(p+1) \geq \kappa_{q+1}$, so the same proposition gives $\mathcal{Y} \subseteq \mathcal{C}_{AF}^{q,l}$. Its parameter estimate, together with $\|\theta - \theta'\|_2 \leq \|\theta - \theta'\|_1$, shows that $\mathrm{Sol} : (\Theta, \|\cdot\|_1) \to (\mathcal{C}_{AF}^{q,l}, \|\cdot\|_{q,l})$ is Lipschitz.

Since Θ is non-empty and compact, $q\lambda/p < l < \lambda$, and the decoder satisfies Definition 3.3 and Assumption C.1, Proposition C.11, applied with $f = \mathrm{Sol}$ and tolerance η, yields an AF-NO $\Phi : \mathbb{R}^{F_{\mathrm{mkt}}} \to \mathcal{C}_{AF}$ with the asserted range property and

$$\sup_{\theta \in \Theta} \|\mathrm{Sol}(\theta) - \Phi(\theta)\|_{q,l} \leq \eta.$$

For every $\theta \in \Theta$ and $0 \leq T, K \leq H$, the definition of the weighted norm gives

$$\begin{aligned} \left| \mathbb{E}_{\mathbb{Q}}[(e^{-rT} X_T^\theta - K)_+] - \Phi(\theta)(T,K) \right| &\leq e^{lT} (1+K)^{-q} \|\mathrm{Sol}(\theta) - \Phi(\theta)\|_{q,l} \\ &\leq e^{lH} \eta = \varepsilon. \end{aligned}$$

Taking the maximum over $(T,K) \in [0,H]^2$ and the supremum over $\theta \in \Theta$ proves the claim. □

C.3 Proof of Positive Results

C.3.1 Proof of Certifiability

We may now deduce our arbitrage-free certificate; i.e. Theorem 4.1.

Proof. For each x, the softmax weights belong to the simplex. By Definition 3.3(ii), each decoder output belongs to $\mathcal{C}_{AF}$. Since $\mathcal{C}_{AF}$ is convex by Proposition 2.8, their convex combination also belongs to $\mathcal{C}_{AF}$. □

The remainder of our analysis will be to establish the corresponding universality guarantee.

C.3.2 Proof of Quantizability: Pointwise Approximate Interpolation

Next, we verify the quantizability property of any separable and closed (metric) subspace of $\mathcal{C}_{AF}^{p,\lambda} \overset{\text{def.}}{=} \mathcal{C}_{AF} \cap \mathcal{X}_\lambda^p$ metrized by $\|\|_{p,\lambda}$. Our analysis holds for any SANOS-type model, in that the model satisfies the following interpolation property:

Assumption C.1 (Uniformly bounded weighted interpolation)**.** Let $\mathcal{Y} \subseteq \mathcal{C}_{AF}^{p,\lambda}$ be admitted by a decoder as in Definition 3.3. There exists $1 \leq B_{\text{dec}} < \infty$ such that, for every $R, S, \eta > 0$, there are $N, M \in \mathbb{N}_+$ with $R/N + S/M \leq \eta$ for which every $C \in \mathcal{Y}$ admits $\theta_C \in \mathbb{R}^{D_{N,M,R,S}}$ satisfying

$$\max_{\substack{0\leq m\leq M\\ 0\leq n\leq N}} \left|\phi_{\theta_C}^{(N,M,R,S)}(mS/M, nR/N) - C(mS/M, nR/N)\right| \leq \eta, \qquad \left\|\phi_{\theta_C}^{(N,M,R,S)}\right\|_{p,\lambda} \leq B_{\text{dec}}. \quad (42)$$

The integers N, M are common to all $C \in \mathcal{Y}$; the parameters θ_C may depend on C.

Proposition C.2 (Quantizability of SANOS-Type Parameterizations)**.** *Fix $0 < q < p$ and let $l, \lambda > 0$ satisfy: $\frac{q}{p}\lambda < l < \lambda$. Let $\mathcal{Y} \subseteq \mathcal{C}_{AF}^{p,\lambda}$ be a uniformly controlled call-surface class with regularity-data (p, λ, ω, B), where ω is strictly increasing, and suppose that $\mathcal{Y}$ is admitted by a decoder satisfying Definition 3.3.*

Moreover, under Assumption C.1. Then for every $\varepsilon > 0$ there exist constants [14] *$R, S > 0$ and $N, M \in \mathbb{N}_+$, such that for every $\Phi \in \mathcal{Y}$ there exists θ_Φ with*

$$\left\|\phi_{\theta_\Phi}^{(N,M,R,S)} - \Phi\right\|_{q,l} < \varepsilon. \quad (43)$$

Moreover, $\frac{R}{N} + \frac{S}{M} \lesssim \omega^{-1}\big(\varepsilon(1+R)^{-q}\big)$ and if $\omega(r) \leq L_\omega r$ for every $r \geq 0$, with a common constant $L_\omega < \infty$, then $(1+R)^q\,(\frac{R}{N} + \frac{S}{M}) \lesssim \varepsilon$.

Proof of Proposition C.2. Fix $\varepsilon > 0$. Since $\frac{q}{p}\lambda < l$ then we may fix a constant $a > 0$ such that

$$\frac{\lambda - l}{p - q} < a < \frac{l}{q}. \quad (44)$$

We first control the target surface C. Since $C \in \mathcal{C}_{AF}$, we have $0 \leq C(T,K) \leq C(T,0) = 1$ for all $(T,K) \in [0,\infty)^2$. Also, by the weighted bound, $C(T,K) \leq B\,e^{\lambda T}(1+K)^{-p}$. Hence, for every $(T,K) \in [0,\infty)^2$, $C(T,K) \leq \min\left\{1,\, Be^{\lambda T}(1+K)^{-p}\right\}$. We claim that this implies

$$e^{-lT}(1+K)^q C(T,K) \leq B^{q/p}\, e^{-(l-\frac{q}{p}\lambda)T}. \quad (45)$$

Indeed, set $A_T \overset{\text{def.}}{=} Be^{\lambda T}$. If $(1+K) \leq A_T^{1/p}$, then $C(T,K) \leq 1$, so

$$e^{-lT}(1+K)^q C(T,K) \leq e^{-lT}(1+K)^q \leq e^{-lT} A_T^{q/p} = B^{q/p} e^{-(l-\frac{q}{p}\lambda)T}.$$

If instead $(1+K) \geq A_T^{1/p}$, then $C(T,K) \leq A_T(1+K)^{-p}$, and thus

$$e^{-lT}(1+K)^q C(T,K) \leq e^{-lT} A_T (1+K)^{q-p} \leq e^{-lT} A_T^{q/p} = B^{q/p} e^{-(l-\frac{q}{p}\lambda)T}.$$

[14] The quantities R, S may be chosen depending only on $\varepsilon, p, q, l, \lambda, B, B_{\text{dec}}, \omega$. The integers N, M may additionally depend on the fixed class $\mathcal{Y}$ and the decoder in Assumption C.1, but not on the individual target $\Phi \in \mathcal{Y}$.

This proves (45). Choose $S > 0$ so large that, with $R \overset{\text{def.}}{=} e^{aS} - 1$, one has $\varepsilon(1+R)^{-q} < \omega(1)$ and

$$\big(B^{q/p} + B_{\text{dec}}^{q/p}\big)e^{-(l-\frac{q}{p}\lambda)S} + (B + B_{\text{dec}})e^{[(\lambda-l)+a(q-p)]S} < \frac{\varepsilon}{3}. \tag{46}$$

Such a choice exists because $l - \frac{q}{p}\lambda > 0$, $(\lambda - l) + a(q-p) < 0$, and $a, q, \omega(1) > 0$. For every $C \in \mathcal{Y}$ and every decoder output $\phi_{\theta_C}^{(N,M,R,S)}$ satisfying $\|\phi_{\theta_C}^{(N,M,R,S)}\|_{p,\lambda} \le B_{\text{dec}}$, Definition 3.3(ii) gives $0 \le \phi_{\theta_C}^{(N,M,R,S)} \le 1$. Applying the calculation establishing (45) with B replaced by B_{dec} yields

$$e^{-lT}(1+K)^q \phi_{\theta_C}^{(N,M,R,S)}(T,K) \le B_{\text{dec}}^{q/p} e^{-(l-\frac{q}{p}\lambda)T}, \qquad T, K \ge 0.$$

Consequently,

$$\sup_{\substack{T>S\\K\ge 0}} e^{-lT}(1+K)^q \big|\phi_{\theta_C}^{(N,M,R,S)}(T,K) - C(T,K)\big| \le \big(B^{q/p} + B_{\text{dec}}^{q/p}\big)e^{-(l-\frac{q}{p}\lambda)S} < \frac{\varepsilon}{3}. \tag{47}$$

For $0 \le T \le S$ and $K > R$, the two weighted bounds give

$$\begin{aligned} e^{-lT}(1+K)^q \big|\phi_{\theta_C}^{(N,M,R,S)}(T,K) - C(T,K)\big| &\le (B + B_{\text{dec}})e^{(\lambda-l)T}(1+K)^{q-p} \\ &\le (B + B_{\text{dec}})e^{(\lambda-l)S}(1+R)^{q-p} \\ &= (B + B_{\text{dec}})e^{[(\lambda-l)+a(q-p)]S} < \frac{\varepsilon}{3}. \end{aligned}$$

Thus

$$\sup_{\substack{0\le T\le S\\K>R}} e^{-lT}(1+K)^q \big|\phi_{\theta_C}^{(N,M,R,S)}(T,K) - C(T,K)\big| < \frac{\varepsilon}{3}. \tag{48}$$

Choose $\eta > 0$ such that

$$(1+R)^q\big(\eta + \omega(\eta)\big) < \frac{\varepsilon}{3}. \tag{49}$$

Assumption C.1 gives common integers $N, M \in \mathbb{N}_+$ with $R/N + S/M \le \eta$ and, for every $C \in \mathcal{Y}$, a parameter θ_C satisfying (42). Set $t_m \overset{\text{def.}}{=} mS/M$, $K_n \overset{\text{def.}}{=} nR/N$, $\Delta_T \overset{\text{def.}}{=} S/M$, and $\Delta_K \overset{\text{def.}}{=} R/N$. Fix any cell $[t_m, t_{m+1}] \times [K_n, K_{n+1}] \subseteq [0,S] \times [0,R]$ and any (T,K) in that cell. Since both C and $\phi_{\theta_C}^{(N,M,R,S)}$ belong to $\mathcal{C}_{AF}$, they are increasing in T and decreasing in K. Therefore

$$C(t_m, K_{n+1}) \le C(T,K) \le C(t_{m+1}, K_n),$$

and likewise $\phi_{\theta_C}^{(N,M,R,S)}(t_m, K_{n+1}) \le \phi_{\theta_C}^{(N,M,R,S)}(T,K) \le \phi_{\theta_C}^{(N,M,R,S)}(t_{m+1}, K_n)$. By (42),

$$C(t_m, K_{n+1}) - \eta \le \phi_{\theta_C}^{(N,M,R,S)}(T,K) \le C(t_{m+1}, K_n) + \eta.$$

The preceding bounds on $C(T,K)$ therefore give

$$\begin{aligned} \big|\phi_{\theta_C}^{(N,M,R,S)}(T,K) - C(T,K)\big| &\le \eta + C(t_{m+1}, K_n) - C(t_m, K_{n+1}) \\ &\le \eta + \omega\big((t_{m+1} - t_m) + (K_{n+1} - K_n)\big) \\ &= \eta + \omega(\Delta_T + \Delta_K). \end{aligned}$$

Since $[0,S] \times [0,R] = \bigcup_{m=0}^{M-1} \bigcup_{n=0}^{N-1} [t_m, t_{m+1}] \times [K_n, K_{n+1}]$ and $e^{-lT}(1+K)^q \le (1+R)^q$ on this rectangle,

$$\sup_{(T,K)\in[0,S]\times[0,R]} e^{-lT}(1+K)^q \big|\phi_{\theta_C}^{(N,M,R,S)}(T,K) - C(T,K)\big| \le (1+R)^q\big(\eta + \omega(\Delta_T + \Delta_K)\big). \tag{50}$$

Using $\Delta_T + \Delta_K \le \eta$ and (49), we obtain

$$\sup_{(T,K)\in[0,S]\times[0,R]} e^{-lT}(1+K)^q \big|\phi_{\theta_C}^{(N,M,R,S)}(T,K) - C(T,K)\big| < \frac{\varepsilon}{3}. \tag{51}$$

Wrapping things up, by combining (47), (48), and (51), we find that

$$\big\|\phi_{\theta_C}^{(N,M,R,S)} - C\big\|_{q,l} < \varepsilon.$$

Since R, S, N, M were chosen independently of $C \in \mathcal{Y}$, this proves the claim. Moreover, $R/N + S/M \le \eta < \omega^{-1}\big(\varepsilon(1+R)^{-q}/3\big) \le \omega^{-1}\big(\varepsilon(1+R)^{-q}\big)$ by (49). If $\omega(r) \le L_\omega r$ for every $r \ge 0$, one may choose $0 < \eta < \varepsilon/[3(1+L_\omega)(1+R)^q]$, giving $(1+R)^q(R/N + S/M) < \varepsilon/[3(1+L_\omega)]$. □

C.3.3 Clustering Mechanism

To solve the general non-injective case, we will next create a clustering mechanism; as in [KCR25]. We with relative to the *intrinsic* dimension of the compact set of features $\Theta \subseteq \mathbb{R}^{F_{\mathrm{mkt}}}$ as formalized by an d-Ahlfors-regular[15] Borel probability measure μ supported on Θ.

Lemma C.3 (Softmax net selector)**.** *Let $\Theta \subseteq \mathbb{R}^{F_{\mathrm{mkt}}}$ be non-empty and compact, and let $\mathcal{Y}$ be a convex subset of a normed vector space, whose norm is denoted by $\|\,\|_{\mathcal{Y}}$. Let $f : \Theta \to \mathcal{Y}$ be L-Lipschitz with respect to $\|\,\|_1$ on Θ and $\|\,\|_{\mathcal{Y}}$ on $\mathcal{Y}$. Fix $r, s > 0$, and let $\{x_n\}_{n=1}^N \subseteq \Theta$ be a maximal r-packing of Θ in $\|\,\|_1$. Write $C_n \overset{\text{def.}}{=} f(x_n)$ for each $n \in [N]_+$ and $D_f \overset{\text{def.}}{=} \operatorname{diam}_{\mathcal{Y}}(f(\Theta))$. Define the alignment score map $\alpha_{\mathbf{X}} : \mathbb{R}^{F_{\mathrm{mkt}}} \to \mathbb{R}^N$ by $\alpha_{\mathbf{X}}(x)_n \overset{\text{def.}}{=} -\|x - x_n\|_1$ for every $n \in [N]_+$. For $\tau > 0$, define*

$$\Phi_{f,r,s,\tau}(x) \overset{\text{def.}}{=} \sum_{n=1}^{N} \operatorname{Softmax}\left(\tfrac{1}{\tau}\alpha_{\mathbf{X}}(x)\right)_n C_n.$$

Then the following hold:

(i) ***Range Control:*** $\Phi_{f,r,s,\tau}(\mathbb{R}^{F_{\mathrm{mkt}}}) \subseteq \operatorname{co}\{C_1, \ldots, C_N\}$.

(ii) ***Approximation:*** *When $D_f > 0$, fix $\beta > 0$. If $\beta \geq ND_f$, let $\tau > 0$ be arbitrary; if $0 < \beta < ND_f$, let $0 < \tau \leq s/\log(ND_f/\beta)$. Then*

we have

$$\sup_{\vartheta \in \Theta} \|\Phi_{f,r,s,\tau}(\vartheta) - f(\vartheta)\|_{\mathcal{Y}} \leq L(r+s) + \beta. \tag{52}$$

When $D_f = 0$, then (52) holds with $\beta = 0$,

(iii) ***ReLU realization of the alignment scores:*** *The alignment-score map $\alpha_{\mathbf{X}}$ is exactly realizable by an N-Lipschitz* ReLU*-MLP of depth 2, width at most $2F_{\mathrm{mkt}}N$, and at most $6F_{\mathrm{mkt}}N$ size*[16]*.*

Proof of Lemma C.3. Since $\{x_n\}_{n=1}^N$ is a maximal r-packing of Θ, it is also an r-net. Hence, for every $\vartheta \in \Theta$, there exists $n_\vartheta \in [N]_+$ such that $\|\vartheta - x_{n_\vartheta}\|_1 \leq r$. For each $\vartheta \in \mathbb{R}^{F_{\mathrm{mkt}}}$ and each $n \in [N]_+$, write $w_n(\vartheta) \overset{\text{def.}}{=} \operatorname{Softmax}\left(\frac{1}{\tau}\alpha_{\mathbf{X}}(\vartheta)\right)_n$. Then $w(\vartheta) \in \Delta_N$, and therefore $\Phi_{f,r,s,\tau}(\vartheta) = \sum_{n=1}^N w_n(\vartheta) C_n$ belongs to $\operatorname{co}\{C_1, \ldots, C_N\}$; proving (i).
We next prove item (ii). Fix $\vartheta \in \Theta$ and define the indexing sets $A_\vartheta \overset{\text{def.}}{=} \{n \in [N]_+ : \|\vartheta - x_n\|_1 \leq r+s\}$ and $A_\vartheta^c \overset{\text{def.}}{=} [N]_+ \setminus A_\vartheta$. Now, if $n \in A_\vartheta$, then the Lipschitzness of f gives

$$\|C_n - f(\vartheta)\|_{\mathcal{Y}} = \|f(x_n) - f(\vartheta)\|_{\mathcal{Y}} \leq L\|\vartheta - x_n\|_1 \leq L(r+s). \tag{53}$$

On the other hand, if $n \in A_\vartheta^c$, then $\|\vartheta - x_n\|_1 > r + s$, while $\|\vartheta - x_{n_\vartheta}\|_1 \leq r$. Therefore

$$\alpha_{\mathbf{X}}(\vartheta)_n - \alpha_{\mathbf{X}}(\vartheta)_{n_\vartheta} = -\|\vartheta - x_n\|_1 + \|\vartheta - x_{n_\vartheta}\|_1 < -s. \tag{54}$$

Consequently, together (53) and (54) imply that

$$w_n(\vartheta) = \frac{\exp\left(\alpha_{\mathbf{X}}(\vartheta)_n/\tau\right)}{\sum_{j=1}^N \exp\left(\alpha_{\mathbf{X}}(\vartheta)_j/\tau\right)} \leq \exp\left(\frac{\alpha_{\mathbf{X}}(\vartheta)_n - \alpha_{\mathbf{X}}(\vartheta)_{n_\vartheta}}{\tau}\right) \leq e^{-s/\tau}$$

and so $\sum_{n \in A_\vartheta^c} w_n(\vartheta) \leq Ne^{-s/\tau}$.

[15] That is, there are constant $0 < c \leq C < \infty$ such that for every $x \in \mathcal{K}$ and each $0 < r \leq \operatorname{diam}(\mathcal{K}, \ell^1)$ we have $c\, r^d \leq \mu(\mathcal{K} \cap \operatorname{Ball}_1(x, r)) \leq C\, r^d$. The prototypical such measures are the Lebesgue measure Riemannian on $[0,1]^d$ when $d = F_{\mathrm{mkt}}$, Riemannian volume measures on compact Riemannian manifolds (when d is an integer), and the certain self-similar measures on fractals (when d is not an integer).

[16] Following [PZ24], the size of a neural network refers to the number of its non-zero real parameters.

We are now all set to prove claim (ii). Let's do it case by case. First, suppose that $D_f > 0$; then, $C_n \in f(\Theta)$ and $f(\vartheta) \in f(\Theta)$, we have $\|C_n - f(\vartheta)\|_{\mathcal{Y}} \leq D_f$ for every $n \in [N]_+$. Thus

$$\begin{aligned}\|\Phi_{f,r,s,\tau}(\vartheta) - f(\vartheta)\|_{\mathcal{Y}} &= \left\| \sum_{n=1}^{N} w_n(\vartheta)\big(C_n - f(\vartheta)\big) \right\|_{\mathcal{Y}} \\ &\leq \sum_{n=1}^{N} w_n(\vartheta)\|C_n - f(\vartheta)\|_{\mathcal{Y}} \\ &\leq \sum_{n\in A_\vartheta} w_n(\vartheta)L(r+s) + \sum_{n\in A_\vartheta^c} w_n(\vartheta)D_f \\ &\leq L(r+s) + D_f N e^{-s/\tau}. \end{aligned} \tag{55}$$

If $0 < \beta < ND_f$ and $0 < \tau \leq s/\log(ND_f/\beta)$, then $s/\tau \geq \log(ND_f/\beta)$ and $ND_f e^{-s/\tau} \leq \beta$. If $\beta \geq ND_f$, then $ND_f e^{-s/\tau} \leq ND_f \leq \beta$ for every $\tau > 0$. Consequently, (55) may be controlled as

$$\|\Phi_{f,r,s,\tau}(\vartheta) - f(\vartheta)\|_{\mathcal{Y}} \leq L(r+s) + \beta. \tag{56}$$

Upon taking the supremum in (56) over $\vartheta \in \Theta$ we deduce (52) in the case where $D_f > 0$.
In the simple case where $D_f = 0$, then f is constant on Θ. Therefore $C_1 = \cdots = C_N = f(\vartheta)$ for every $\vartheta \in \Theta$, and hence $\Phi_{f,r,s,\tau}(\vartheta) = f(\vartheta)$ for every $\vartheta \in \Theta$. Thus (52) holds with $\beta = 0$; and (ii) is verified.

It remain to justify (iii); namely, the ReLU realization of $\alpha_{\mathbf{X}}$. For each $n \in [N]_+$ and each $j \in \{1, \ldots, F_{\text{mkt}}\}$, the maps $x \mapsto \text{ReLU}(x_j - (x_n)_j)$ and $x \mapsto \text{ReLU}((x_n)_j - x_j)$ are realized by one ReLU layer. Hence $\|x - x_n\|_1 = \sum_{j=1}^{F_{\text{mkt}}} \big(\text{ReLU}(x_j - (x_n)_j) + \text{ReLU}((x_n)_j - x_j)\big)$. Applying an affine output layer with coefficients -1 on the corresponding hidden units realizes $\alpha_{\mathbf{X}}(x)_n = -\|x - x_n\|_1$ for every $n \in [N]_+$. Computing all N coordinates in parallel gives depth 2, width at most $2F_{\text{mkt}}N$, and at most $6F_{\text{mkt}}N$ size. Finally, for every $x, y \in \mathbb{R}^{F_{\text{mkt}}}$ and every $n \in [N]_+$, $|\alpha_{\mathbf{X}}(x)_n - \alpha_{\mathbf{X}}(y)_n| = \big|\|y - x_n\|_1 - \|x - x_n\|_1\big| \leq \|x - y\|_1$. Summing over $n \in [N]_+$ gives $\|\alpha_{\mathbf{X}}(x) - \alpha_{\mathbf{X}}(y)\|_1 \leq N\|x - y\|_1$. □

C.3.4 Clustering to Interpolation

Lemma C.4 (K-Point Lipschitz Interpolator)**.** *Let $F, K \in \mathbb{N}_+$ and let $\mathbf{X} = (x_1, \ldots, x_K)^\top \in \mathbb{R}^{K\times F}$ have distinct rows. For $K \geq 2$, set $\text{sep}(X) = \min_{j\neq k} \|x_j - x_k\|_1$; for $K = 1$, set $\text{sep}(X) = 2$. Put $\delta = \text{sep}(X)/2 > 0$. Then there exists an $\frac{2}{\text{sep}(X)}$-Lipschitz ReLU-MLP $\alpha_{\mathbf{X}} : \mathbb{R}^F \to \mathbb{R}^K$ of depth 3, width at most $2FK$, and size at most $K(6F+2)$ satisfying: for every $k \in [K]_+$*

$$\alpha_{\mathbf{X}}(x) = e_k \text{ if and only if } x = x_k. \tag{57}$$

Proof of Lemma C.4. The construction below proves the interpolation and Lipschitz claims when $K \geq 2$. Under the MLP definition in Appendix A, it has depth 3 and size at most $K(6F+2)$; it does not establish the smaller bounds in the unchanged statement. For $K = 1$, the construction works for any specified $\delta > 0$, but the definition of $\text{sep}(X)$ requires a separate convention.

Set $f_1(x) \overset{\text{def.}}{=} (\|x - x_k\|_1)_{k=1}^K$, $f_2(z) \overset{\text{def.}}{=} \text{ReLU}(\mathbf{1}_K - z)$, and $\alpha_{\mathbf{X}} \overset{\text{def.}}{=} f_2 \circ (\delta^{-1} f_1)$; motivated by the observation that the only norm-zero vector is the zero-vector itself. Explicitly, consider the ReLU-MLP

$$\alpha_{\mathbf{X}}(x) = \text{ReLU}\left(\mathbf{1}_K - \tfrac{1}{\delta} f_1(x)\right) = \text{ReLU}\left(\mathbf{1}_K - \tfrac{1}{\delta}\left((I_F, I_F)\,\text{ReLU}\left((I_F, -I_F)^\top (x - x_k)\right)\right)_{k=1}^K\right), \tag{58}$$

with two ReLU layers followed by an identity affine output layer, hence depth 3, width $2FK$, and size at most $K(6F+2)$.

We first verify the interpolation property in (58). Fix $k \in [K]_+$. Then $f_1(x_k)_k = \|x_k - x_k\|_1 = 0$, so $\alpha_{\mathbf{X}}(x_k)_k = \text{ReLU}(1-0) = 1$. For $j \neq k$, the separation condition gives $f_1(x_k)_j = \|x_k - x_j\|_1 \geq 2\delta$, hence $\alpha_{\mathbf{X}}(x_k)_j = \text{ReLU}(1 - \frac{1}{\delta}\|x_k - x_j\|_1) \leq \text{ReLU}(1-2) = 0$. Thus $\alpha_{\mathbf{X}}(x_k) = e_k$. Conversely, if $\alpha_{\mathbf{X}}(x) = e_k$ then in particular $\alpha_{\mathbf{X}}(x)_k = 1$, which forces $\frac{1}{\delta}\|x - x_k\|_1 = 0$, i.e. $x = x_k$.

To prove the ℓ^1 Lipschitz bound, note that at most one tent is positive at any point, since the open balls of radius δ about the centers are disjoint. If x, y lie in the same tent support, the reverse triangle inequality gives $\|\alpha_{\mathbf{X}}(x) - \alpha_{\mathbf{X}}(y)\|_1 \le \|x - y\|_1/\delta$. If only one point lies in a support, the same bound follows because the other point has distance at least δ from that center. If x lies in support i and y in a different support j, then

$$\|\alpha_{\mathbf{X}}(x) - \alpha_{\mathbf{X}}(y)\|_1 = 2 - \frac{\|x - x_i\|_1 + \|y - x_j\|_1}{\delta} \le \frac{\|x - y\|_1}{\delta},$$

using $\|x_i - x_j\|_1 \ge 2\delta$ and the triangle inequality. If both points are outside all supports, the difference is zero. Thus $\alpha_{\mathbf{X}}$ is δ^{-1}-Lipschitz. □

Lemma C.5 (Key Properties of Alignment Scores)**.** *Let $F, K \in \mathbb{N}_+$ and $\mathbf{X} \overset{\text{def.}}{=} (x_1, \dots, x_K)^\top \in \mathbb{R}^{K\times F}$ with $x_1, \dots, x_K$ distinct. For $K = 1$, use $\operatorname{sep}(X) = 2$; assertion (ii) is required only when $K \ge 2$. Fix $0 < \varepsilon < 1$. Then there exists a $\frac{2}{\operatorname{sep}(X)}$-Lipschitz* ReLU*-MLP $\alpha_{\mathbf{X}} : \mathbb{R}^F \to \mathbb{R}^K$ such that, for every temperature parameter $0 < \tau \le \big(\log\big(\frac{(2-\varepsilon)(\max\{2,K\}-1)}{\varepsilon}\big)\big)^{-1}$:*

(i) ***Approximation:*** $\max_{k\in[K]_+} \big\|e_k - \operatorname{Softmax}\big(\frac{1}{\tau}\alpha_{\mathbf{X}}(x_k)\big)\big\|_1 \le \varepsilon$.

(ii) ***Separation:*** $\min_{k\in[K]_+} \min_{j\ne k} \big\|e_j - \operatorname{Softmax}\big(\frac{1}{\tau}\alpha_{\mathbf{X}}(x_k)\big)\big\|_1 \ge 2 - \frac{\varepsilon}{\max\{K,2\}-1} > 1$.

(iii) ***Moderate Regularity:*** *The map $\operatorname{Softmax}\big(\frac{1}{\tau}\alpha_{\mathbf{X}}(\,)\big)\colon \mathbb{R}^F \to \operatorname{relint}(\Delta_K)$ is $\frac{1}{\tau \operatorname{sep}(X)}$-Lipschitz.*

Moreover, $\alpha_{\mathbf{X}}$ has depth 3, width $2FK$, and at most $K(6F + 2)$ size.

Lemma C.6 (Smooth and Lipschitz-Regular Approximate Projection of Extremal Simplicial Points)**.** *Let $K \in \mathbb{N}_+$ and $0 < \varepsilon < 1$. For each $k \in [K]_+$, let $e_k \overset{\text{def.}}{=} (I_{j=k})_{j=1}^K$ be the k^{th} standard basis vector. If $0 < \tau \le \big(\log\big(\frac{(2-\varepsilon)(\max\{2,K\}-1)}{\varepsilon}\big)\big)^{-1}$, then*

$$\max_{k\in[K]_+} \big\|e_k - \operatorname{Softmax}\big(\tfrac{1}{\tau}e_k\big)\big\|_1 \le \varepsilon \quad \text{and} \quad \min_{k\in[K]_+} \min_{j\ne k} \big\|e_j - \operatorname{Softmax}\big(\tfrac{1}{\tau}e_k\big)\big\|_1 \ge 2 - \tfrac{\varepsilon}{\max\{K,2\}-1}. \tag{59}$$

The separation inequality is asserted only for $K \ge 2$; for $K = 1$, the softmax is constant and the approximation error is zero. Moreover, the map $\mathbb{R}^K \ni x \mapsto \operatorname{Softmax}(\frac{1}{\tau}x)$ is $\frac{1}{2\tau}$-Lipschitz with respect to $\|\,\|_1$.

Proof of Lemma C.6. If $K = 1$ the statement is trivial; assume therefore $K \ge 2$. Fix $k \in [K]_+$. Since $(e_k)_j = I_{j=k}$, a direct computation gives

$$\operatorname{Softmax}\big(\tfrac{1}{\tau}e_k\big)_j = \begin{cases} \dfrac{e^{1/\tau}}{e^{1/\tau} + K - 1} & j = k, \\ \dfrac{1}{e^{1/\tau} + K - 1} & j \ne k. \end{cases}$$

Hence $\|e_k - \operatorname{Softmax}(\frac{1}{\tau}e_k)\|_1 = \frac{2(K-1)}{e^{1/\tau}+K-1}$, which is independent of k. The bound $\tau \le \big(\log\frac{(2-\varepsilon)(K-1)}{\varepsilon}\big)^{-1}$ gives $e^{1/\tau} \ge \frac{(2-\varepsilon)(K-1)}{\varepsilon}$, and therefore

$$\frac{2(K-1)}{e^{1/\tau} + K - 1} \le \frac{2(K-1)}{\frac{(2-\varepsilon)(K-1)}{\varepsilon} + K - 1} = \frac{2(K-1)}{(K-1)\frac{2}{\varepsilon}} = \varepsilon,$$

which proves the approximation bound in (59).

For the separation bound, fix $k \in [K]_+$ and $j \ne k$, and write $s^{(k)} \overset{\text{def.}}{=} \operatorname{Softmax}(\frac{1}{\tau}e_k)$. Then

$$\|e_j - s^{(k)}\|_1 = \left(1 - \frac{1}{e^{1/\tau} + K - 1}\right) + \frac{e^{1/\tau}}{e^{1/\tau} + K - 1} + \frac{K-2}{e^{1/\tau} + K - 1} = 2 - \frac{2}{e^{1/\tau} + K - 1}.$$

Since $e^{1/\tau} + K - 1 \ge \frac{2(K-1)}{\varepsilon}$, we obtain $\frac{2}{e^{1/\tau}+K-1} \le \frac{\varepsilon}{K-1} = \frac{\varepsilon}{\max\{K,2\}-1}$, and hence $\|e_j - s^{(k)}\|_1 \ge 2 - \frac{\varepsilon}{\max\{K,2\}-1}$, which proves the separation bound in (59).

Finally, the Jacobian of Softmax at $p = \operatorname{Softmax}(z)$ is $J(p) = \operatorname{diag}(p) - pp^\top$, whose $\ell^1 \to \ell^1$ operator norm satisfies $\|J(p)\|_{\ell^1\to\ell^1} = \max_j \sum_i |J_{ij}| = \max_j 2p_j(1 - p_j) \le \frac{1}{2}$, so Softmax is $\frac{1}{2}$-Lipschitz, and the additional factor $\frac{1}{\tau}$ from the chain rule gives the $\frac{1}{2\tau}$-Lipschitz claim. □

We now return to the proof of Lemma C.5.

Proof of Lemma C.5. For $K \geq 2$, the argument below proves the interpolation, separation, and Lipschitz estimates using the displayed tent construction. Its depth is 3 and its size is at most $K(6F+2)$. Its softmax range lies in the relative interior of Δ_K, whose ambient interior is empty. The smaller complexity bounds, the ambient-interior notation, and the singleton convention in the unchanged statement remain pending.

Apply Lemma C.4 to obtain a $\frac{2}{\operatorname{sep}(X)}$-Lipschitz ReLU-MLP $\alpha_{\mathbf{X}} : \mathbb{R}^F \to \mathbb{R}^K$ satisfying $\alpha_{\mathbf{X}}(x_k) = e_k$ for every $k \in [K]_+$. Hence, for every $k \in [K]_+$, we have

$$\operatorname{Softmax}\left(\tfrac{1}{\tau}\,\alpha_{\mathbf{X}}(x_k)\right) = \operatorname{Softmax}\left(\tfrac{1}{\tau}\,e_k\right).$$

For the proofs of (i) and of (ii). Since $\alpha_{\mathbf{X}}(x_k) = e_k$ for every $k \in [K]_+$, both claims follow immediately from Lemma C.6 applied with the given ε and τ.

For the proof of (iii). By Lemma C.6, the map $z \mapsto \operatorname{Softmax}(\frac{1}{\tau}z)$ is $\frac{1}{2\tau}$-Lipschitz with respect to $\|\,\|_1$. By Lemma C.4, $\alpha_{\mathbf{X}}$ is $\frac{2}{\operatorname{sep}(X)}$-Lipschitz. Hence, by composition,

$$\operatorname{Lip}\Big(\operatorname{Softmax}\left(\tfrac{1}{\tau}\,\alpha_{\mathbf{X}}(\cdot)\right)\Big) \leq \frac{1}{2\tau}\,\frac{2}{\operatorname{sep}(X)} = \frac{1}{\tau\,\operatorname{sep}(X)}.$$

Finally, since $\operatorname{Softmax}(z) \in \operatorname{relint}(\Delta_K)$ for every $z \in \mathbb{R}^K$, the composition takes values in $\operatorname{relint}(\Delta_K)$. □

From Alignment Scores to Arbitrage-Free Surfaces via Geometric Attention Geometric Attention via Approximate Simplicial Structure of $\mathcal{C}_{AF}$ Having setup a suitable metric structure, we can now argue similarly to [AKP24, Example 5.1]. We now deduce approximate simplicial structure. However, we first recall the definition.

Definition C.7 (Approximately Simplicial)**.** A metric space $(\mathcal{Y}, d_{\mathcal{Y}})$ is said to be approximately simplicial if there is a function

$$\eta : \bigcup_{N \in \mathbb{N}_+} (\Delta_N \times \mathcal{Y}^N) \to \mathcal{Y},$$

called a mixing function, and constants $C_\eta \geq 1$ and $p \in \mathbb{N}_+$, such that for every $N \in \mathbb{N}_+$, $w = (w_1, \ldots, w_N) \in \Delta_N$, and $\mathbf{y} = (y_1, \ldots, y_N) \in \mathcal{Y}^N$, one has

$$d_{\mathcal{Y}}\left(\eta\left(w, \mathbf{y}\right), C_i\right) \leq C_\eta \left(\sum_{j=1}^{N} d_{\mathcal{Y}}(C_i, y_j)^p w_j\right)^{1/p}.$$

Note that in particular the function η satisfies $\eta(e_i, \mathbf{y}) = C_i$, $i = 1, \ldots, N$.

We now have a suitable metric structure for our approximately simplicial structure.

Lemma C.8 (Approximately Simplicial)**.** *Fix $p, \lambda > 0$, and suppose that $\mathcal{C}_{AF} \cap \mathcal{X}_\lambda^p$ is convex. Define the mixing function, for every $N \in \mathbb{N}_+$, each $w \in \Delta_N$, and every $C^{(1)}, \ldots, C^{(N)} \in \mathcal{C}_{AF} \cap \mathcal{X}_\lambda^p$ by*

$$\eta\big(w, (C^{(1)}, \ldots, C^{(N)})\big) \overset{\text{def.}}{=} \sum_{k=1}^{N} w_k C^{(k)}. \tag{60}$$

Then, η satisfies Definition C.7 with constants $C_\eta = 1$ and exponent 1.

Proof of Lemma C.8. Let $N \in \mathbb{N}_+$, let $w = (w_1, \ldots, w_N) \in \Delta_N$, and let $\mathcal{C} = (C^{(1)}, \ldots, C^{(N)}) \in (\mathcal{C}_{AF} \cap \mathcal{X}_\lambda^p)^N$. Since $\mathcal{C}_{AF} \cap \mathcal{X}_\lambda^p$ is convex, we have $\eta(w, \mathcal{C}) \in \mathcal{C}_{AF} \cap \mathcal{X}_\lambda^p$. Fix $i \in \{1, \ldots, N\}$. Then

$$\eta(w, \mathcal{C}) - C^{(i)} = \sum_{j=1}^{N} w_j \big(C^{(j)} - C^{(i)}\big).$$

Hence, by the triangle inequality and absolute homogeneity of the norm,

$$\delta_{p,\lambda}\big(\eta(w,\mathcal{C}),C^{(i)}\big)=\Big\|\sum_{j=1}^{N}w_j\big(C^{(j)}-C^{(i)}\big)\Big\|_{p,\lambda}\le\sum_{j=1}^{N}w_j\,\|C^{(j)}-C^{(i)}\|_{p,\lambda}.$$

That is, $\delta_{p,\lambda}\big(\eta(w,\mathcal{C}),C^{(i)}\big)\le\sum_{j=1}^{N}w_j\,\delta_{p,\lambda}\big(C^{(i)},C^{(j)}\big)$. Therefore Definition C.7 holds with $C_\eta=1$ and exponent 1. Moreover, if $w=e_i$, then $\eta(e_i,\mathcal{C})=C^{(i)}$. □

We first solve the case, effectively corresponding to the target function is injective.

Lemma C.9 (Quantized geometric attention on $\mathcal{C}_{AF}^{p,\lambda}$ - Injective Case)**.** *Let $p,\lambda>0$, $F_{\mathrm{mkt}}\in\mathbb{N}_+$, and let $\mathcal{K}\subseteq\mathcal{C}_{AF}^{p,\lambda}$ be non-empty and compact. Fix $\varepsilon>0$, and let $\mathbf{C}\overset{\text{def.}}{=}\{C_n\}_{n=1}^{N}\subseteq\mathcal{C}_{AF}^{p,\lambda}$ be a maximal ε-packing for $\mathcal{K}$ in $(\mathcal{C}_{AF}^{p,\lambda},\|\,\|_{p,\lambda})$. Let $\mathbf{X}\overset{\text{def.}}{=}(x_1,\dots,x_N)^\top\in\mathbb{R}^{N\times F_{\mathrm{mkt}}}$ have pairwise distinct rows, and fix $0<\delta<1$ and set the temperature to $0<\tau\le\left(\log\left(\frac{(2-\delta)(\max\{2,N\}-1)}{\delta}\right)\right)^{-1}$. There exists an alignment score map*

$$\begin{aligned}\alpha_{\mathbf{X}:\mathbf{C}}:\mathbb{R}^{F_{\mathrm{mkt}}}&\to\mathcal{C}_{AF}^{p,\lambda}\\ x&\mapsto\sum_{n=1}^{N}\mathrm{Softmax}\left(\tfrac{1}{\tau}\alpha_{\mathbf{X}}(x)\right)_n C_n,\end{aligned}\tag{61}$$

where $\alpha_{\mathbf{X}}:\mathbb{R}^{F_{\mathrm{mkt}}}\to\mathbb{R}^N$ is the network from Lemma C.5, such that:

(i) ***Approximate ε-Net Interpolation:*** $\max_{n\in[N]_+}\big\|\alpha_{\mathbf{X}:\mathbf{C}}(x_n)-C_n\big\|_{p,\lambda}\le\frac{\delta}{2}\,\mathrm{diam}_{p,\lambda}(\mathbf{C})$,

(ii) ***Range Control:***

(ii-a) No-Arbitrage: $\alpha_{\mathbf{X}:\mathbf{C}}(\mathbb{R}^{F_{\mathrm{mkt}}})\subseteq\mathcal{C}_{AF}^{p,\lambda}$,

(ii-b) In-Context Range: $\alpha_{\mathbf{X}:\mathbf{C}}(\mathbb{R}^{F_{\mathrm{mkt}}})\subseteq\mathrm{co}\{C_1,\dots,C_N\}$.

Moreover, the finite-dimensional encoder $\alpha_{\mathbf{X}}$ has depth 3, width $2F_{\mathrm{mkt}}N$, and no more than $N(6F_{\mathrm{mkt}}+2)$ size. Moreover $\alpha_{\mathbf{X}:\mathbf{C}}$ is $\frac{\|\mathbf{C}\|_{\ell^1\to(\mathcal{X}_\lambda^p,\|\,\|_{p,\lambda})}}{\tau\ \mathrm{sep}(X)}$-Lipschitz
Where $\mathrm{diam}_{p,\lambda}(\mathbf{C})\overset{\text{def.}}{=}\max_{i,j\in[N]_+}\|C_i-C_j\|_{p,\lambda}$ *and* $\|\mathbf{C}\|_{\ell^1\to(\mathcal{X}_\lambda^p,\|\,\|_{p,\lambda})}\overset{\text{def.}}{=}\sup_{\|w\|_1\le1}\|\sum_{n=1}^{N}w_nC_n\|_{p,\lambda}$.

Proof of Proposition C.9. For $N\ge2$, the following argument gives the approximation, convex-hull, and Lipschitz estimates with the tent encoder of depth 3 and size at most $N(6F_{\mathrm{mkt}}+2)$. The linear dictionary map is interpreted in the ambient weighted vector space. The smaller complexity bounds, the stated operator-norm codomain, and the singleton convention in the unchanged statement remain pending.

For each $x\in\mathbb{R}^{F_{\mathrm{mkt}}}$, write $w(x)\overset{\text{def.}}{=}\mathrm{Softmax}(\frac{1}{\tau}\alpha_{\mathbf{X}}(x))\in\Delta_N$, so that $\alpha_{\mathbf{X}:\mathbf{C}}(x)=\sum_{n=1}^{N}w_n(x)\,C_n$.

We first prove (i). Fix $n\in[N]_+$. By Lemma C.5 and the assumed upper bound on τ, we have $\|w(x_n)-e_n\|_1\le\delta$. Then, $\alpha_{\mathbf{X}:\mathbf{C}}(x_n)-C_n=\sum_{j=1}^{N}w_j(x_n)(C_j-C_n),$, so by the triangle inequality and the definition of $\mathrm{diam}_{p,\lambda}(\mathbf{C})$,

$$\big\|\alpha_{\mathbf{X}:\mathbf{C}}(x_n)-C_n\big\|_{p,\lambda}\le\mathrm{diam}_{p,\lambda}(\mathbf{C})\sum_{j\ne n}w_j(x_n).$$

Since $w(x_n)\in\Delta_N$, we have $\sum_{j\ne n}w_j(x_n)=1-w_n(x_n)=\frac{1}{2}\|w(x_n)-e_n\|_1\le\frac{\delta}{2}$, and taking the maximum over n proves (i).

We prove (ii-b). For every $x\in\mathbb{R}^{F_{\mathrm{mkt}}}$, $\alpha_{\mathbf{X}:\mathbf{C}}(x)=\sum_{n=1}^{N}w_n(x)C_n$ is a convex combination of $C_1,\dots,C_N$ (since $w(x)\in\Delta_N$), so $\alpha_{\mathbf{X}:\mathbf{C}}(\mathbb{R}^{F_{\mathrm{mkt}}})\subseteq\mathrm{co}\{C_1,\dots,C_N\}$.

We now show (ii-a). Since each $C_n\in\mathcal{C}_{AF}^{p,\lambda}=\mathcal{C}_{AF}\cap\mathcal{X}_\lambda^p$, and both $\mathcal{C}_{AF}$ and $\mathcal{X}_\lambda^p$ are convex, (ii-b) gives $\alpha_{\mathbf{X}:\mathbf{C}}(\mathbb{R}^{F_{\mathrm{mkt}}})\subseteq\mathrm{co}\{C_1,\dots,C_N\}\subseteq\mathcal{C}_{AF}^{p,\lambda}$.

Tally the complexity of $\alpha_{\mathbf{X}}$, we see that it has depth 3, width $2F_{\mathrm{mkt}}N$, and at most $N(6F_{\mathrm{mkt}}+2)$ size by Lemma C.5. We now estimate the Lipschitzness of our ReLU-MLP. Writing $\alpha_{\mathbf{X}:\mathbf{C}}=$

$T \circ \mathrm{Softmax}(\frac{1}{\tau}\alpha_{\mathbf{X}}(\cdot))$, where $T(w) \overset{\text{def.}}{=} \sum_{n=1}^{N} w_n C_n$, Lemma C.6 gives that $z \mapsto \mathrm{Softmax}(\frac{1}{\tau}z)$ is $\frac{1}{2\tau}$-Lipschitz and Lemma C.5 gives that $\alpha_{\mathbf{X}}$ is $\frac{2}{\mathrm{sep}(X)}$-Lipschitz, so by composition

$$\mathrm{Lip}(\alpha_{\mathbf{X};\mathbf{C}}) \leq \|\mathbf{C}\|_{\ell^1 \to (\mathcal{X}_\lambda^p, \|\,\|_{p,\lambda})} \frac{1}{2\tau} \frac{2}{\mathrm{sep}(X)} = \frac{\|\mathbf{C}\|_{\ell^1 \to (\mathcal{X}_\lambda^p, \|\,\|_{p,\lambda})}}{\tau\ \mathrm{sep}(X)}.$$

□

We are now prepare to prove the main approximation result.

Proposition C.10 (Uniform softmax geometric-attention approximation)**.** *Let $\Theta \subseteq \mathbb{R}^{F_{\mathrm{mkt}}}$ be non-empty and compact, and let $\mathcal{Y}$ be a convex subset of a normed vector space, whose norm is denoted by $\|\,\|_{\mathcal{Y}}$. Let $f : \Theta \to \mathcal{Y}$ be L-Lipschitz with respect to $\|\,\|_1$ on Θ and $\|\,\|_{\mathcal{Y}}$ on $\mathcal{Y}$. Then, for every $0 < \varepsilon < 1$, there exist $N \in \mathbb{N}_+$, points $x_1, \dots, x_N \in \Theta$, elements $C_n \overset{\text{def.}}{=} f(x_n)$, a* ReLU*-MLP alignment-score map $\alpha_{\mathbf{X}} : \mathbb{R}^{F_{\mathrm{mkt}}} \to \mathbb{R}^N$, and a temperature $\tau > 0$ such that the map $\Phi_{f,\varepsilon}(x) \overset{\text{def.}}{=} \sum_{n=1}^{N} \mathrm{Softmax}\left(\frac{1}{\tau}\alpha_{\mathbf{X}}(x)\right)_n C_n$ satisfies the uniform estimate*

$$\sup_{x \in \Theta} \|\Phi_{f,\varepsilon}(x) - f(x)\|_{\mathcal{Y}} \leq \varepsilon. \tag{62}$$

Moreover, $\Phi_{f,\varepsilon}(\mathbb{R}^{F_{\mathrm{mkt}}}) \subseteq \mathrm{co}\{C_1, \dots, C_N\}$.
Furthermore, $\alpha_{\mathbf{X}}$ can be realized by a ReLU*-MLP of depth 2, width at most $2F_{\mathrm{mkt}}N$, and at most $6F_{\mathrm{mkt}}N$ size.*

Proof. Fix $0 < \varepsilon < 1$. Choose $r, s > 0$ such that $L(r+s) < \varepsilon$. Let $\{x_n\}_{n=1}^N \subseteq \Theta$ be a maximal r-packing of Θ in $\|\,\|_1$. Since Θ is compact, this packing is finite. Write $C_n \overset{\text{def.}}{=} f(x_n)$ for $n \in [N]_+$ and set $D_f \overset{\text{def.}}{=} \mathrm{diam}_{\mathcal{Y}}(f(\Theta))$. If $D_f > 0$, choose $0 < \beta < \varepsilon - L(r+s)$ with $\beta < ND_f$, and then choose $0 < \tau \leq \frac{s}{\log\left(\frac{ND_f}{\beta}\right)}$; while if $D_f = 0$ then choose $\tau = 1$. Define $\alpha_{\mathbf{X}}$ and $\Phi_{f,\varepsilon}$ as in Lemma C.3.
The range-control statement follows directly from Lemma C.3 (i). If $D_f > 0$, Lemma C.3 (ii) gives

$$\sup_{x \in \Theta} \|\Phi_{f,\varepsilon}(x) - f(x)\|_{\mathcal{Y}} \leq L(r+s) + \beta \leq \varepsilon. \tag{63}$$

If $D_f = 0$, then (63) holds with $\beta = 0$. The ReLU-MLP realization of $\alpha_{\mathbf{X}}$ follows from Lemma C.3 (iii). □

It only remains to replace the $C_1, \dots, C_N$ in (61) with $\frac{\varepsilon}{2}$ approximations of an ε-packing of $f(\Theta)$ by SANOS, and to take f to be as in Proposition 4.4. Upon doing so, we obtain the following *finite dimensional* analogue of our result.

Proposition C.11 (Universal Arbitrage-Free Approximation: Finite-Dimensional Version)**.** *Fix $0 < q < p$ and let $l, \lambda > 0$ satisfy $\frac{q}{p}\lambda < l < \lambda$. Let $\Theta \subseteq \mathbb{R}^{F_{\mathrm{mkt}}}$ be non-empty and compact. Let $\mathcal{Y} \subseteq \mathcal{C}_{AF}^{p,\lambda}$ be a uniformly controlled call-surface class admitted by a decoder satisfying Definition 3.3 and Assumption C.1. Let $f : \Theta \to \mathcal{Y}$ be L-Lipschitz as a map from $(\Theta, \|\,\|_1)$ into $(\mathcal{C}_{AF}^{q,l}, \|\,\|_{q,l})$.*
For every $0 < \varepsilon < 1$, there exist $I \in \mathbb{N}_+$, points $x_1, \dots, x_I \in \Theta$, $R, S > 0$, $N_0, M_0 \in \mathbb{N}_+$, $\theta_1, \dots, \theta_I \in \mathbb{R}^{D_{N_0,M_0,R,S}}$, $\tau > 0$, and a ReLU*-MLP $\alpha_{\mathbf{X}} : \mathbb{R}^{F_{\mathrm{mkt}}} \to \mathbb{R}^I$ such that*

$$\Phi(x) \overset{\text{def.}}{=} \sum_{i=1}^{I} \mathrm{Softmax}\left(\tfrac{1}{\tau}\alpha_{\mathbf{X}}(x)\right)_i \widetilde{C}_i, \ \text{where } \widetilde{C}_i \overset{\text{def.}}{=} \phi_{\theta_i}^{(N_0,M_0,R,S)}, \tag{64}$$

satisfies:

(i) ***Approximation Guarantee:*** $\sup_{x \in \Theta} \|\Phi(x) - f(x)\|_{q,l} \leq \varepsilon$

(ii) ***Interpretable Range:*** $\Phi(\mathbb{R}^{F_{\mathrm{mkt}}}) \subseteq \mathcal{C}_{AF} \bigcap \mathrm{co}\{\widetilde{C}_1, \dots, \widetilde{C}_I\}$.

Moreover, $\alpha_{\mathbf{X}}$ is realizable by a ReLU*-MLP of depth 2, width at most $2F_{\mathrm{mkt}}I$, and size at most $6F_{\mathrm{mkt}}I$.*

Proof of Proposition C.11. Apply Proposition C.10 with $(\mathcal{Y}, \|\,\|_{\mathcal{Y}}) = (\mathcal{C}_{AF}^{q,l}, \|\,\|_{q,l})$ and approximation tolerance $\varepsilon/2$. This gives points $x_1, \dots, x_I \in \Theta$, surfaces $C_i \overset{\text{def.}}{=} f(x_i)$, a ReLU-MLP alignment-score map $\alpha_{\mathbf{X}} : \mathbb{R}^{F_{\text{mkt}}} \to \mathbb{R}^I$, and a temperature $\tau > 0$ such that

$$\sup_{x\in\Theta} \left\| \sum_{i=1}^{I} \text{Softmax}\left(\tfrac{1}{\tau}\alpha_{\mathbf{X}}(x)\right)_i C_i - f(x) \right\|_{q,l} \leq \frac{\varepsilon}{2}.$$

Moreover, by Proposition C.10, $\alpha_{\mathbf{X}}$ may be chosen to be exactly realizable by a ReLU-MLP of depth 2, width at most $2F_{\text{mkt}}I$, and at most $6F_{\text{mkt}}I$ size. Since $\{C_1, \dots, C_I\} \subseteq f(\Theta) \subseteq \mathcal{Y}$, Proposition C.2, applied with tolerance $\varepsilon/2$, gives common decoder hyperparameters $R, S > 0$ and $N_0, M_0 \in \mathbb{N}_+$, together with parameters $\theta_1, \dots, \theta_I \in \mathbb{R}^{D_{N_0,M_0,R,S}}$, such that, with $\widetilde{C}_i \overset{\text{def.}}{=} \phi_{\theta_i}^{(N_0,M_0,R,S)}$; then, one has $\|\widetilde{C}_i - C_i\|_{q,l} \leq \frac{\varepsilon}{2}$ holding for every $i \in [I]_+$.
Next, define $\Phi(x) \overset{\text{def.}}{=} \sum_{i=1}^{I} \text{Softmax}\left(\frac{1}{\tau}\alpha_{\mathbf{X}}(x)\right)_i \widetilde{C}_i$. For every $x \in \Theta$, write $w_i(x) \overset{\text{def.}}{=} \text{Softmax}\left(\frac{1}{\tau}\alpha_{\mathbf{X}}(x)\right)_i$. Since $w(x) \in \Delta_I$, we have

$$\begin{aligned} \|\Phi(x) - f(x)\|_{q,l} &\leq \left\| \sum_{i=1}^{I} w_i(x)(\widetilde{C}_i - C_i) \right\|_{q,l} + \left\| \sum_{i=1}^{I} w_i(x) C_i - f(x) \right\|_{q,l} \\ &\leq \sum_{i=1}^{I} w_i(x) \|\widetilde{C}_i - C_i\|_{q,l} + \frac{\varepsilon}{2} \\ &\leq \varepsilon. \end{aligned} \tag{65}$$

Taking the supremum across (65) over $x \in \Theta$ gives the uniform approximation estimate. Finally, by Definition 3.3(ii), each $\widetilde{C}_i$ belongs to $\mathcal{C}_{AF}$. Since $w(x) \in \Delta_I$ for every x and $\mathcal{C}_{AF}$ is convex by Proposition 2.8, we obtain $\Phi(\mathbb{R}^{F_{\text{mkt}}}) \subseteq \mathcal{C}_{AF} \cap \text{co}\{\widetilde{C}_1, \dots, \widetilde{C}_I\}$. □

C.4 Extension to Infinite-Dimensional RKHS Input

We now extend Proposition C.11 to inputs in a separable RKHS over $\mathcal{D} \subseteq [0, \infty)^2$ which admits a sampling and interpolating sequence $\{z_\lambda\}_{\lambda\in\Lambda}$. The first step is to prove Proposition 3.1 which will allow us to approximately reduce our finite-dimensional problem to a finite-dimensional one; thus bridging our setting to one compatible with Proposition C.11.

Proof of Proposition 3.1. If $\mathcal{H} = \{0\}$, choose $K_\varepsilon = 1$, any $x_1 \in \mathcal{D}$, and $A_\varepsilon = (0)$; then (18) is immediate. Otherwise, enumerate the complete interpolating sampling sequence as (x_n) and put $\phi_n \overset{\text{def.}}{=} \kappa(\cdot, x_n)$. For each index j, interpolation supplies $f_j \in \mathcal{H}$ with $f_j(x_n) = \mathbf{1}_{\{n=j\}}$. Thus every finite relation $\sum_{n=1}^{K} a_n \phi_n = 0$ gives

$$a_j = \left\langle \sum_{n=1}^{K} a_n \phi_n, f_j \right\rangle_{\mathcal{H}} = 0, \qquad 1 \leq j \leq K.$$

The sections are therefore linearly independent. If $h \perp \phi_n$ for every n, then $h(x_n) = 0$ for every n and the sampling inequality (1) gives $\|h\|_{\mathcal{H}} = 0$. Hence their span is dense. In finite dimension the sequence consequently has exactly $\dim(\mathcal{H})$ terms. For $K \leq \dim(\mathcal{H})$, set $V_K \overset{\text{def.}}{=} \text{span}\{\phi_1, \dots, \phi_K\}$ and define the associated Gram matrix by $\mathbf{G}_K \overset{\text{def.}}{=} \left(\langle \phi_j, \phi_i \rangle_{\mathcal{H}}\right)_{i,j=1}^{K} = \left(\kappa(x_i, x_j)\right)_{i,j=1}^{K}$. Since $\phi_1, \dots, \phi_K$ are linearly independent, $\mathbf{G}_K$ is positive definite and hence invertible. Choose $A_K \overset{\text{def.}}{=} \mathbf{G}_K^{-1}$ and define $\Pi_K C \overset{\text{def.}}{=} \sum_{k=1}^{K} \mathcal{E}_K\left(C, (x_j)_{j=1}^{K}, A_K\right)_k \phi_k$. For every $i \in \{1, \dots, K\}$,

$$\langle C - \Pi_K C, \phi_i \rangle_{\mathcal{H}} = \langle C, \phi_i \rangle_{\mathcal{H}} - \left[\mathbf{G}_K \mathbf{G}_K^{-1} \Big(\langle C, \phi_j \rangle_{\mathcal{H}} \Big)_{j=1}^{K} \right]_i = 0.$$

Thus Π_K is the orthogonal projection onto V_K. If $d \overset{\text{def.}}{=} \dim(\mathcal{H}) < \infty$, then $V_d = \mathcal{H}$ and $\Pi_d = I_{\mathcal{H}}$; taking $K_\varepsilon = d$ and $A_\varepsilon = \mathbf{G}_d^{-1}$ proves (18). Henceforth suppose that $\mathcal{H}$ is infinite-dimensional, with the selected kernel sections indexed by $\mathbb{N}_+$. Completeness gives $\overline{\bigcup_{K\geq 1} V_K} = \mathcal{H}$. For each $C \in \mathcal{H}$

and $\eta > 0$, choose $K_0 \in \mathbb{N}_+$ and $v \in V_{K_0}$ such that $\|C - v\|_{\mathcal{H}} < \eta$. Since $V_{K_0} \subseteq V_K$ for $K \geq K_0$, the best-approximation property yields

$$\|C - \Pi_K C\|_{\mathcal{H}} \leq \|C - v\|_{\mathcal{H}} < \eta.$$

Thus $\Pi_K C \to C$ in $\mathcal{H}$. It remains to show that the convergence is uniform on $\mathcal{K}$. Fix $\varepsilon > 0$. By compactness, there exist $C_1, \dots, C_M \in \mathcal{K}$ such that

$$\mathcal{K} \subseteq \bigcup_{m=1}^{M} B_{\mathcal{H}}\left(C_m, \frac{\varepsilon}{2}\right).$$

For each $m = 1, \dots, M$, choose $K_m \in \mathbb{N}_+$ such that $\|C_m - \Pi_{K_m} C_m\|_{\mathcal{H}} < \varepsilon/2$, and set $K_\varepsilon \stackrel{\text{def.}}{=} \max_{1 \leq m \leq M} K_m$. Since $V_{K_m} \subseteq V_{K_\varepsilon}$, the best-approximation property gives

$$\|C_m - \Pi_{K_\varepsilon} C_m\|_{\mathcal{H}} \leq \|C_m - \Pi_{K_m} C_m\|_{\mathcal{H}} < \varepsilon/2$$

for every $m = 1, \dots, M$. For any $C \in \mathcal{K}$, choose m satisfying $\|C - C_m\|_{\mathcal{H}} < \varepsilon/2$. Since $I_{\mathcal{H}} - \Pi_{K_\varepsilon}$ is an orthogonal projection and therefore has operator norm at most one,

$$\begin{aligned} \|C - \Pi_{K_\varepsilon} C\|_{\mathcal{H}} &\leq \left\|(I_{\mathcal{H}} - \Pi_{K_\varepsilon})(C - C_m)\right\|_{\mathcal{H}} + \|C_m - \Pi_{K_\varepsilon} C_m\|_{\mathcal{H}} \\ &\leq \|C - C_m\|_{\mathcal{H}} + \|C_m - \Pi_{K_\varepsilon} C_m\|_{\mathcal{H}} \\ &< \varepsilon. \end{aligned} \tag{66}$$

Taking the supremum in (66) over all $C \in \mathcal{K}$ and then setting $A_\varepsilon \stackrel{\text{def.}}{=} \mathbf{G}_{K_\varepsilon}^{-1}$ yields (18). □

It is time to prove our main approximation theorem; i.e. Theorem 4.2.

Proof of Theorem 4.2. The surfaces $\widetilde{C}_i$ used below will be decoder outputs of the form $\widetilde{C}_i = \phi_{\theta_i}^{(N_0, M_0, R, S)}$, with common decoder hyperparameters N_0, M_0, R, S and parameters θ_i to be chosen below. Once chosen, these surfaces are fixed independently of the input C; only their softmax mixing weights depend on C. Fix $\varepsilon_{\text{DR}} \stackrel{\text{def.}}{=} \frac{\varepsilon}{2}$, $\varepsilon_{\text{FD-Aprx}} \stackrel{\text{def.}}{=} \frac{\varepsilon}{2}$, and set $\delta \stackrel{\text{def.}}{=} \frac{\varepsilon_{\text{DR}}}{\max\{1, L\}}$. By Proposition 3.1, there exist $K \in \mathbb{N}_+$, $(x_k)_{k=1}^K \in \mathcal{D}^K$, and $A \in \mathbb{R}^{K \times K}$ such that

$$\sup_{C \in \mathcal{K}} \left\| C - \sum_{k=1}^{K} \mathcal{E}_K\big(C, (x_j)_{j=1}^K, A\big)_k \kappa(\cdot, x_k) \right\|_{\mathcal{H}} \leq \delta. \tag{67}$$

Define the linear reconstruction map $\mathcal{R}_K : \mathbb{R}^K \to \mathcal{H}$ given for every $z \in \mathbb{R}^K$ by $\mathcal{R}_K(z) \stackrel{\text{def.}}{=} \sum_{k=1}^K z_k \kappa(\cdot, x_k)$ and abbreviate $\mathcal{E} \stackrel{\text{def.}}{=} \mathcal{E}_K\big(\,\cdot\,, (x_k)_{k=1}^K, A\big)$. Thus, the finite-rank reconstruction operator appearing in Proposition 3.1 is $P_K \stackrel{\text{def.}}{=} \mathcal{R}_K \circ \mathcal{E}$. Consequently, (67) reads

$$\sup_{C \in \mathcal{K}} \|C - P_K C\|_{\mathcal{H}} \leq \delta. \tag{68}$$

Since $\mathcal{E} : \mathcal{H} \to \mathbb{R}^K$ is continuous and $\mathcal{K}$ is compact, the set $\Theta_K \stackrel{\text{def.}}{=} \mathcal{E}(\mathcal{K}) \subseteq \mathbb{R}^K$ is also non-empty and compact. Consider the finite-dimensional surrogate

$$\begin{aligned} \mathcal{F}_K : \Theta_K &\to \mathcal{Y}, \\ z &\mapsto \mathcal{F}\big(\mathcal{R}_K(z)\big). \end{aligned} \tag{69}$$

The linear map $\mathcal{R}_K$ is continuous from $(\mathbb{R}^K, \|\,\|_1)$ into $\mathcal{H}$. Therefore, for every $z, \tilde{z} \in \Theta_K$,

$$\|\mathcal{F}_K(z) - \mathcal{F}_K(\tilde{z})\|_{q,l} \leq L \|\mathcal{R}_K(z - \tilde{z})\|_{\mathcal{H}} \leq L \|\mathcal{R}_K\|_{(\mathbb{R}^K, \|\,\|_1) \to \mathcal{H}} \|z - \tilde{z}\|_1.$$

Hence, $\mathcal{F}_K$ is Lipschitz from $(\Theta_K, \|\,\|_1)$ into $(\mathcal{C}_{AF}^{q,l}, \|\,\|_{q,l})$. Since $\mathcal{F}$ is defined on all of $\mathcal{H}$ with values in $\mathcal{Y}$, one has $\mathcal{F}_K(z) = \mathcal{F}(\mathcal{R}_K z) \in \mathcal{Y}$ for every $z \in \Theta_K$. In particular, the required call surface is $\mathcal{F}(\mathcal{R}_K z)$; no call-surface condition is imposed on $\mathcal{R}_K z$. With $c_K \stackrel{\text{def.}}{=} \max_{1 \leq k \leq K} \sqrt{\kappa(x_k, x_k)}$, one has $\|\mathcal{R}_K z\|_{\mathcal{H}} \leq c_K \|z\|_1$ for every $z \in \mathbb{R}^K$, and hence $\|\mathcal{F}_K(z) - \mathcal{F}_K(z')\|_{q,l} \leq L c_K \|z - z'\|_1$ for every $z, z' \in \Theta_K$. We may therefore apply Proposition C.11 to $\mathcal{F}_K$ on Θ_K with approximation tolerance

$\varepsilon_{\text{FD-Aprx}}$. It yields $I \in \mathbb{N}_+$, $R, S > 0$, $N_0, M_0 \in \mathbb{N}_+$, parameters $\theta_1, \dots, \theta_I \in \mathbb{R}^{D_{N_0,M_0,R,S}}$, a temperature $\tau > 0$, and a ReLU-MLP $\alpha_{\mathbf{X}} : \mathbb{R}^K \to \mathbb{R}^I$ with the following properties. For $i = 1, \dots, I$, define $\widetilde{C}_i \overset{\text{def.}}{=} \phi_{\theta_i}^{(N_0,M_0,R,S)}$, where $\phi^{(N_0,M_0,R,S)}$ is the surface decoder from Definition 3.3.[17] In the proof of the finite-dimensional proposition, the parameters θ_i are chosen so that $\|\widetilde{C}_i - \mathcal{F}_K(z_i)\|_{q,l} \leq \varepsilon_{\text{FD-Aprx}}/2$ for suitable anchor points $z_1, \dots, z_I \in \Theta_K$. Moreover, $\widetilde{C}_i \in \mathcal{C}_{AF}$ by Definition 3.3(ii). The map $\widehat{\Phi}(z) \overset{\text{def.}}{=} \sum_{i=1}^I \text{Softmax}\left(\frac{1}{\tau}\alpha_{\mathbf{X}}(z)\right)_i \widetilde{C}_i$ satisfies

$$\sup_{z \in \Theta_K} \|\widehat{\Phi}(z) - \mathcal{F}_K(z)\|_{q,l} \leq \varepsilon_{\text{FD-Aprx}}, \tag{70}$$

as well as

$$\widehat{\Phi}(\mathbb{R}^K) \subseteq \mathcal{C}_{AF} \cap \text{co}\{\widetilde{C}_1, \dots, \widetilde{C}_I\}. \tag{71}$$

Write $\alpha_{\mathbf{X}}(z) = A_1 \,\text{ReLU}(A_0 z + b_0) + b_1$, where $h \in \mathbb{N}_+$, $A_0 \in \mathbb{R}^{h \times K}$, $b_0 \in \mathbb{R}^h$, $A_1 \in \mathbb{R}^{I \times h}$, and $b_1 \in \mathbb{R}^I$. For the one-token input $z^\top \in \mathbb{R}^{1 \times K}$, take one attention head with key dimension 1, value dimension h, $\mathbf{Q} = \mathbf{K} = 0_{1\times K}$, $\mathbf{V} = A_0$, and attention scale 1. Use a transformer of depth 2 with $d_0 = K$, $d_1 = h$, $d_2 = I$, activation ReLU, $b^{(0)} = b_0$, and final readout $(\mathbf{A}^{(1)}, b^{(1)}) = (A_1, b_1)$. Then, applying A.2 we deduce that $\alpha_{\mathbf{X}}$ can be realized by a multi-head transformer:

$$\left[\,\text{Attn}(z^\top \mid \mathbf{Q}, \mathbf{K}, \mathbf{V}, 1)\right]_1 = \frac{e^0}{e^0} A_0 z = A_0 z, \qquad [\mathcal{T}(z^\top)]_1 = A_1 \,\text{ReLU}(A_0 z + b_0) + b_1 = \alpha_{\mathbf{X}}(z), \quad z \in \mathbb{R}^K.$$

Next, define $\Phi \overset{\text{def.}}{=} \widehat{\Phi} \circ \mathcal{E}$. This is precisely the map in (35). For every $C \in \mathcal{K}$, we have

$$\mathcal{F}_K\big(\mathcal{E}(C)\big) = \mathcal{F}\big(\mathcal{R}_K \circ \mathcal{E}(C)\big) = \mathcal{F}(P_K C).$$

Consequently, by (70), the Lipschitz continuity of $\mathcal{F}$, and (68),

$$\begin{aligned} \|\Phi(C) - \mathcal{F}(C)\|_{q,l} &\leq \left\|\widehat{\Phi}\big(\mathcal{E}(C)\big) - \mathcal{F}_K\big(\mathcal{E}(C)\big)\right\|_{q,l} \\ &\quad + \|\mathcal{F}(P_K C) - \mathcal{F}(C)\|_{q,l} \\ &\leq \varepsilon_{\text{FD-Aprx}} + L\|P_K C - C\|_{\mathcal{H}} \\ &\leq \varepsilon_{\text{FD-Aprx}} + L\delta \\ &\leq \varepsilon_{\text{FD-Aprx}} + \varepsilon_{\text{DR}} \\ &= \varepsilon. \end{aligned} \tag{72}$$

Taking the supremum in (72) over $C \in \mathcal{K}$ proves (i). Finally, since $\mathcal{E}(\mathcal{H}) \subseteq \mathbb{R}^K$, the range estimate (71) gives $\Phi(\mathcal{H}) = \widehat{\Phi}\big(\mathcal{E}(\mathcal{H})\big) \subseteq \widehat{\Phi}(\mathbb{R}^K) \subseteq \mathcal{C}_{AF} \cap \text{co}\{\widetilde{C}_1, \dots, \widetilde{C}_I\}$; proving (ii). □

Theorem C.12 (Certifiably Arbitrage-Free)**.** *Any map of the form* (22) *satisfies: For every* $\vartheta \in \mathbb{R}^{F_{\text{mkt}}}$ *the predicted call surface*

$$\mathcal{AF}\text{-}\mathcal{NO}(\vartheta) : [0, \infty)^2 \to [0, \infty)$$

is arbitrage-free; i.e. it satisfies the conditions of Theorem 2.5.

Proof. For every input, the component surfaces in (22) belong to $\mathcal{C}_{AF}$ and their weights form a simplex vector independent of strike and maturity. Proposition 2.8 therefore proves the claim for every parameter choice. □

C.5 Proof of Negative Results

We now prove our negative results.

[17] E.g. as in Definition 3.10.

C.5.1 Proof of No Arbitrage-Free Projection Theorem

For every $\lambda, p > 0$ define

$$\mathcal{X}_\lambda^p \overset{\text{def.}}{=} \Big\{ C : [0,\infty)^2 \to \mathbb{R} \,:\, \|C\|_{p,\lambda} < \infty \Big\}.$$

Since finer topologies have more closed sets than weaker ones, it suffices to prove the result in the topology of uniform convergence on compact sets on $C([0,\infty)^2,[0,\infty))$. Indeed, if τ is any topology on this space which is no stronger than the compact-open topology, then every τ-closed set is also closed in the compact-open topology. Thus, non-closedness in the compact-open topology immediately implies non-closedness in τ. Equivalently, the same compact-open convergent sequence witnesses non-closedness for every such weaker topology.

Proposition C.13 (Non-Closedness of $\mathcal{C}_{AF}$ in the Compact-Open Topology)**.** *The set $\mathcal{C}_{AF}$ is not closed in $C\big([0,\infty)^2,[0,\infty)\big)$ equipped with the topology of uniform convergence on compact sets. Consequently, for every topology τ on $C\big([0,\infty)^2,[0,\infty)\big)$ which is no stronger than the topology of uniform convergence on compact sets, the set $\mathcal{C}_{AF}$ is not τ-closed.*

Proof of Proposition C.13. Define $a(T) \overset{\text{def.}}{=} \frac{T}{1+T}$ for every $T \geq 0$. For each $n \in \mathbb{N}_2$, define

$$C_n(T,K) \overset{\text{def.}}{=} \big(1-a(T)\big)(1-K)^+ + a(T)\left(1-\frac{K}{n}\right)^+$$

for every $T, K \geq 0$. We first verify that $C_n \in \mathcal{C}_{AF}$. Clearly, C_n is continuous and $C_n(T,0) = 1$ for every $T \geq 0$. Moreover,

$$\partial_K^+ C_n(T,0) = -\big(1-a(T)\big) - \frac{a(T)}{n} = -1 + a(T)\left(1-\frac{1}{n}\right) \geq -1.$$

For every fixed $T \geq 0$, the map $K \mapsto C_n(T,K)$ is convex, being a convex combination of convex call-payoff functions, and $\lim_{K\uparrow\infty} C_n(T,K) = 0$. Furthermore,

$$C_n(T,K) = (1-K)^+ + a(T)\left[\left(1-\frac{K}{n}\right)^+ - (1-K)^+\right].$$

Since a is non-decreasing and $\big(1-\frac{K}{n}\big)^+ \geq (1-K)^+$ for every $K \geq 0$, the map $T \mapsto C_n(T,K)$ is non-decreasing for every $K \geq 0$. Finally, since $a(0) = 0$, $C_n(0,K) = (1-K)^+$ for every $K \geq 0$. Proposition 2.5 therefore implies that $C_n \in \mathcal{C}_{AF}$ for every $n \in \mathbb{N}_2$. Define $C_\infty(T,K) \overset{\text{def.}}{=} \big(1-a(T)\big)(1-K)^+ + a(T)$. Let $T_0, K_0 > 0$. For every $n > K_0$ and every $(T,K) \in [0,T_0]\times[0,K_0]$, we have

$$\big|C_n(T,K) - C_\infty(T,K)\big| = a(T)\left|\left(1-\frac{K}{n}\right)^+ - 1\right| = a(T)\frac{K}{n} \leq \frac{K_0}{n}.$$

Consequently, $\sup_{(T,K)\in[0,T_0]\times[0,K_0]} \big|C_n(T,K) - C_\infty(T,K)\big| \leq \frac{K_0}{n}$, and hence $C_n \to C_\infty$ uniformly on compact subsets of $[0,\infty)^2$. However, for every $T > 0$, $\lim_{K\uparrow\infty} C_\infty(T,K) = a(T) > 0$.. Thus, C_∞ violates condition (iii) of Proposition 2.5, and therefore $C_\infty \notin \mathcal{C}_{AF}$. Hence $\mathcal{C}_{AF}$ is not closed in the topology of uniform convergence on compact sets. Finally, let τ be any topology on the same ambient space which is no stronger than the compact-open topology. Since $C_n \to C_\infty$ in the compact-open topology, we also have $C_n \to C_\infty$ in τ. Since every C_n belongs to $\mathcal{C}_{AF}$ while C_∞ does not, the set $\mathcal{C}_{AF}$ cannot be τ-closed. □

Now, the following non-existence of metric projections onto non-closed subsets in metric spaces; which follows readily from the definitions of a metric projection (not considering continuity or Borel-measurability of such an object) and elementary characterizations of closed sets in metric spaces.

Lemma C.14 (Non-closed sets do not admit nearest points everywhere)**.** *Let (X,ρ) be a metric space and let $A \subseteq X$ be non-empty and not closed. Then, there exists some $x^\star \in X$ such that*

$$\inf_{a\in A} \rho(x^\star, a) = 0, \tag{73}$$

but there is no $a^\star \in A$ satisfying

$$\rho(x^\star, a^\star) = \inf_{a \in A} \rho(x^\star, a). \tag{74}$$

If for every $x \in X$ there exists $a_x \in A$ such that $\rho(x, a_x) = \inf_{a \in A} \rho(x, a)$, then A is closed.

Prof of Lemma C.14. Since A is not closed, there exists some $x^\star \in \overline{A} \setminus A$. Because $x^\star \in \overline{A}$, there exists a sequence $(a_n)_{n=1}^\infty \subseteq A$ with $\rho(x^\star, a_n) \to 0$. Hence, $\inf_{a \in A} \rho(x^\star, a) = 0$, which proves (73). Assume now, for contradiction, that there exists some $a^\star \in A$ such that $\rho(x^\star, a^\star) = \inf_{a \in A} \rho(x^\star, a)$. Then $\rho(x^\star, a^\star) = 0$, and therefore $x^\star = a^\star$. But this contradicts the fact that $x^\star \notin A$. Thus, no minimizer in A can exist for $x^\star$, proving (74). The final claim follows from contraposition. □

We now have the necessary tools to prove our "no arbitrage-free projections" theorem.

Proof of Theorem 5.1. By Proposition C.13, the set $\mathcal{C}_{AF}$ is not closed in the topology induced by any metric ρ on $C([0,\infty)^2, [0,\infty))$ which is no stronger than the topology of uniform convergence on compact sets. Applying Lemma C.14 with $X = C([0,\infty)^2, [0,\infty))$ and $A = \mathcal{C}_{AF}$ yields the conclusion. □

D Proofs of Certificate Failures for Vanilla Arbitrage-Free Decoder-Neural Operators

D.1 Proofs for (Encoder) Decoder-Type NOs

Lemma D.1 (Affine obstruction for decoder-type neural operators)**.** *Fix a continuous orthonormal system $(e_n)_{n=1}^\infty \subset C([0,1])$ in $L^2([0,1])$, and consider the tensorized orthonormal system $(E_{n,m})_{n,m=1}^\infty$ of $L^2([0,1]^2)$ defined for each $n, m \in \mathbb{N}$ by $E_{n,m}(s,t) \overset{\text{def.}}{=} e_n(s)e_m(t)$. Fix $K \in \mathbb{N}_+$ and pairwise distinct indices $(n_1, m_1), \dots, (n_K, m_K) \in \mathbb{N}^2$, and set*

$$V_K \overset{\text{def.}}{=} \operatorname{span}\{E_{n_k,m_k} : 1 \le k \le K\}.$$

Then the following are equivalent:

(i) Every $g \in V_K$ is convex on $[0,1]^2$.

(ii) Every $g \in V_K$ is affine on $[0,1]^2$.

(iii) $V_K \subseteq \operatorname{Aff}([0,1]^2)$.

In particular, if $K \ge 4$, then $V_K \not\subseteq \operatorname{Aff}([0,1]^2)$, and therefore there exists some $a^\star \in \mathbb{R}^K$ such that $D_{a^\star}(x) \overset{\text{def.}}{=} \sum_{k=1}^K a_k^\star E_{n_k,m_k}(x)$ is not convex on $[0,1]^2$.

Proof of Lemma D.1. The equivalence of (ii) and (iii) is immediate, and $(iii) \Rightarrow (i)$ holds since every affine function is convex. It remains to prove $(i) \Rightarrow (ii)$. Fix any $g \in V_K$. Since V_K is a linear subspace, we also have $-g \in V_K$. By (i), both g and $-g$ are convex on $[0,1]^2$. Hence g is both convex and concave, and therefore affine on $[0,1]^2$. Since $g \in V_K$ was arbitrary, every element of V_K is affine. For the final claim, note that the tensorized system $(E_{n,m})_{n,m \ge 1}$ is orthonormal in $L^2([0,1]^2)$, so the pairwise distinct functions $E_{n_1,m_1}, \dots, E_{n_K,m_K}$ are linearly independent. Thus $\dim(V_K) = K$. Since $\operatorname{Aff}([0,1]^2)$ has dimension 3, it follows that $V_K \not\subseteq \operatorname{Aff}([0,1]^2)$ whenever $K \ge 4$. The existence of a non-convex element of V_K then follows from the equivalence already proved. □

We may now prove the non-certifiability of (encoder) decoder-type networks in $L^2([0,1])$; for arbitrage-free certificates.

Proof of Proposition 5.2. Since the functions $e_{n_1}, e_{n_2}, e_{n_3}$ are linearly independent, the space $U \overset{\text{def.}}{=} \operatorname{span}\{e_{n_1}, e_{n_2}, e_{n_3}\}$ has dimension 3. If all elements of U were convex, then for each $h \in U$ both h and $-h$ would be convex, forcing every element to be affine. On the other hand, the space of affine functions on $[0,1]$ has dimension 2. Therefore, $U \not\subseteq \operatorname{Aff}([0,1])$, so there exists

some $a^\star = (a_1^\star, a_2^\star, a_3^\star) \in \mathbb{R}^3$ such that $h(K) \overset{\text{def.}}{=} \sum_{k=1}^3 a_k^\star e_{n_k}(K)$ is not convex on $[0,1]$. Next, since $e_{m_\star} \not\equiv 0$ in $L^2([0,1])$, there exists some $T^\star \in [0,1]$ such that

$$e_{m_\star}(T^\star) \neq 0.$$

Replacing $a^\star$ by $\operatorname{sign}(e_{m_\star}(T^\star))\, a^\star$ if necessary, we may assume that $e_{m_\star}(T^\star)\, h$ is still non-convex on $[0,1]$. Now choose $f_{NN} : \mathbb{R}^F \to \mathbb{R}^3$ to be the constant map

$$f_{NN}(x) \equiv a^\star.$$

Then, for every $x \in \mathbb{R}^F$, $f_{D\text{-}NO}(x)(T,K) = \sum_{k=1}^3 a_k^\star e_{n_k}(K) e_{m_\star}(T)$. In particular, at $T = T^\star$ we obtain $K \mapsto f_{D\text{-}NO}(x)(T^\star, K) = e_{m_\star}(T^\star) \sum_{k=1}^3 a_k^\star e_{n_k}(K) = e_{m_\star}(T^\star)\, h(K)$, which is not convex on $[0,1]$. Therefore, for every $x \in \mathbb{R}^F$, the decoded surface $f_{D\text{-}NO}(x)$ fails the strike-convexity condition required for membership in $\mathcal{C}_{AF}$. Hence $f_{D\text{-}NO}(x) \notin \{C|_{[0,1]^2} : C \in \mathcal{C}_{AF}\}$ for every $x \in \mathbb{R}^F$. □

D.2 Proofs for FNOs

Proof of Proposition 5.4. Set $W_\ell = 0$ and $R_\ell = 0$ for every $\ell \in \{0, \dots, L-1\}$. If $L \geq 2$, set $b_\ell \equiv 0$ for every $\ell \in \{0, \dots, L-2\}$. For the final layer, define $b_{L-1}(T,K) \overset{\text{def.}}{=} \xi - \eta + 2\eta K$ for $(T,K) \in [0,1]^2$.

Since $m = 1$ and $P = \mathrm{Id}$, the corresponding FNO output satisfies, for every $a \in \mathbb{R}^F$,

$$\Phi(Q^\uparrow(a))(T,K) = \sigma\big(\xi - \eta + 2\eta K\big).$$

Indeed, the input field $Q^\uparrow(a)$ is killed by the choice $W_\ell = R_\ell = 0$, and all preceding biases vanish.

Fix $T \in [0,1]$. Then

$$K \mapsto \Phi(Q^\uparrow(a))(T,K) = \sigma\big(\xi - \eta + 2\eta K\big).$$

Differentiating twice in K gives

$$\partial_{KK}\Phi(Q^\uparrow(a))(T,K) = (2\eta)^2 \sigma''\big(\xi - \eta + 2\eta K\big).$$

Since $K \mapsto \xi - \eta + 2\eta K$ maps $[0,1]$ onto $[\xi - \eta, \xi + \eta]$, and σ'' takes both positive and negative values on $(\xi - \eta, \xi + \eta)$, it follows that $\partial_{KK}\Phi(Q^\uparrow(a))(T,K)$ takes both positive and negative values on $[0,1]$. Hence $K \mapsto \Phi(Q^\uparrow(a))(T,K)$ is not convex on $[0,1]$.

On the other hand, the restriction to $[0,1]^2$ of any element of $\mathcal{C}_{AF}$ is convex in the strike variable by Theorem 2.5. Therefore, for every $a \in \mathbb{R}^F$, the continuous representative of $\Phi(Q^\uparrow(a))$ is not the restriction to $[0,1]^2$ of an element of $\mathcal{C}_{AF}$. □

E Funding and Acknowledgements

A. Kratsios acknowledges financial support from an NSERC Discovery Grant No. RGPIN-2023-04482 and No. DGECR-2023-00230. and by the project Bando PRIN 2022 named "Qnt4Green - Quantitative Approaches for Green Bond Market: Risk Assessment, Agency Problems and Policy Incentives", codice 2022JRY7EF, CUP E53D23006330006, funded by European Union – NextGenerationEU, M4c2.
B. Horvath acknowledges financial support from UKRI1010: High order mathematical and computational infrastructure for streamed data that enhance contemporary generative and large language models.
A. Kratsios and R. Saqur both also acknowledge that resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute[18].

Acknowledgements A. Kratsios would also like to thank Frédéric Godin and Geneviève Gauthier for their helpful feedback while the authors were completing our manuscript. A. Kratsios would like to thank Mathieu Rosenbaum for his helpful discussions on static no-arbitrage vs. dynamic no-arbitrage right before this manuscript's release.

[18] https://vectorinstitute.ai/partnerships/current-partners/

A Additional Background on Deep Learning

We now record some key results in the literature which we will use now. We recall that the ReLU $\stackrel{\text{def.}}{=} \max\{0, \cdot\}$ activation function on $\mathbb{R}$.

Definition A.1 (Multilayer Perceptrons (MLPs))**.** Let $d, D, L \in \mathbb{N}$, and let $d_0, d_1, \ldots, d_L \in \mathbb{N}$ such that $d_0 = d$ and $d_L = D$. Let $\sigma \in \mathcal{C}(\mathbb{R})$ be an activation function. A fully-connected MLP with depth L and activation function σ is a map $f : \mathbb{R}^d \to \mathbb{R}^D$ with iterative representation:

$$X^{(0)} \stackrel{\text{def.}}{=} X,$$
$$X^{(j+1)} \stackrel{\text{def.}}{=} \sigma \bullet (\mathbf{A}^{(j)} X^{(j)} + b^{(j)}) \qquad j = 0, 1, \ldots, L-2, \qquad f(X) \stackrel{\text{def.}}{=} \mathbf{A}^{(L-1)} X^{(L-1)} + b^{(L-1)},$$

where $\bullet$ denotes component-wise application, $\mathbf{A}^{(j)} \in \mathbb{R}^{d_{j+1} \times d_j}$, and $b^{(j)} \in \mathbb{R}^{d_{j+1}}$.

Proposition A.2 (Transformerification of MLPs - [KNP26, Proposition A.1])**.** *Let $f : \mathbb{R}^d \to \mathbb{R}^D$ be a MLP with depth L. Then, for every $\lambda > 0$, f can be implemented exactly as a one-token transformer with the same depth as f, exactly one attention head in each hidden layer, and the same final affine readout.*

Proof. Choose one token, one attention head, key dimension 1, and value dimension d_{j+1} in hidden layer j. The attention softmax consists of the single weight 1. In hidden layer j, take $\mathbf{Q}^{(j)} = \mathbf{K}^{(j)} = 0$, $\mathbf{V}^{(j)} = \mathbf{A}^{(j)}$, and the same bias and activation as in the MLP. The layer is therefore exactly $x \mapsto \sigma(\mathbf{A}^{(j)} x + b^{(j)})$. Retain the MLP's final affine output layer. Their compositions agree for every input; when $L = 1$, only that affine layer remains. □

A.1 Proofs of Structured ReLU-MLP Arbitrage-Free Decoder in Section 3.3

We first prove the polytopal characterization of "good no-arbitrage quotes" over a fixed finite grid. The following proof is a largely case-by-case verification; which we encourage the reader to skip on a first read as it is tedious.

Proof of Proposition 3.6. Throughout the proof, the virtual strikes K_{L} and K_{R} are fixed as in the statement. In particular,

$$0 = \widehat{K}_0 < \widehat{K}_1 = K_{\mathrm{L}} < \widehat{K}_2 = K_1 < \cdots < \widehat{K}_{N+1} = K_N < \widehat{K}_{N+2} = K_{\mathrm{R}}. \tag{75}$$

For convenience, abbreviate $\Delta\widehat{K}_j \stackrel{\text{def.}}{=} \widehat{K}_{j+1} - \widehat{K}_j > 0$ for each $j \in [N+1]$.
We first make two simple observations which we repeatedly use. Let $f : [0, \infty) \to [0, \infty)$ be convex and suppose that

$$\lim_{K \uparrow \infty} f(K) = 0. \tag{76}$$

Then f is non-increasing. Indeed, suppose to the contrary that there exist $0 \le x < y$ such that $f(x) < f(y)$. Set $q \stackrel{\text{def.}}{=} \frac{f(y)-f(x)}{y-x} > 0$. Now, for every $z > y$, the monotonicity of secant slopes of a convex function gives $\frac{f(z)-f(y)}{z-y} \ge \frac{f(y)-f(x)}{y-x} = q$ and, consequently, we have $f(z) \ge f(y) + q(z-y)$; implying that f tends to ∞ as $z \uparrow \infty$, contradicting (76). Thus f is non-increasing.
We shall also use the following standard secant-slope inequality. If $x_0 < x_1 < x_2$, then

$$\frac{f(x_1) - f(x_0)}{x_1 - x_0} \le \frac{f(x_2) - f(x_0)}{x_2 - x_0} \le \frac{f(x_2) - f(x_1)}{x_2 - x_1}. \tag{77}$$

To verify this, write $x_1 = (1-\lambda)x_0 + \lambda x_2$ and $\lambda \stackrel{\text{def.}}{=} \frac{x_1 - x_0}{x_2 - x_0} \in (0, 1)$ and, note that by convexity we have that Convexity gives

$$f(x_1) \le (1-\lambda) f(x_0) + \lambda f(x_2). \tag{78}$$

Rearranging across (78) yields (77). In particular, if $x_0 < \cdots < x_J$, then the consecutive secant slopes satisfy

$$\frac{f(x_1) - f(x_0)}{x_1 - x_0} \le \cdots \le \frac{f(x_J) - f(x_{J-1})}{x_J - x_{J-1}}. \tag{79}$$

We are now set to prove that (i) implies (ii). Suppose that item (i) holds, and let $\mathfrak{I} : [0,\infty)^2 \to [0,\infty)$ be an interpolator satisfying the stated conditions. Now, fix $m \in [M]_+$ and define the corresponding maturity slice

$$f_m(K) \overset{\text{def.}}{=} \mathfrak{I}(T_m, K), \qquad K \geq 0. \tag{80}$$

Since $\mathfrak{I}$ is arbitrage-free, Proposition 2.5 implies that f_m is convex and satisfies

$$f_m(0) = 1, \qquad \lim_{\delta \downarrow 0} \frac{f_m(\delta) - f_m(0)}{\delta} \geq -1, \qquad \lim_{K \uparrow \infty} f_m(K) = 0. \tag{81}$$

By Step 1, f_m is non-increasing. The interpolation and virtual data-point conditions give $f_m(\widehat{K}_j) = \widehat{y}_{m,j}$ for each $j \in [N+2]$. Therefore, for every $j \in [N+1]$ we have that

$$s_{m,j}(y) = \frac{\widehat{y}_{m,j+1} - \widehat{y}_{m,j}}{\widehat{K}_{j+1} - \widehat{K}_j} = \frac{f_m(\widehat{K}_{j+1}) - f_m(\widehat{K}_j)}{\widehat{K}_{j+1} - \widehat{K}_j}. \tag{82}$$

Since f_m is convex, the consecutive secant-slope inequality (79) yields

$$s_{m,0}(y) \leq s_{m,1}(y) \leq \cdots \leq s_{m,N+1}(y). \tag{83}$$

Moreover, f_m is non-increasing, and hence $s_{m,j}(y) \leq 0$ for each $j \in [N+1]$. By convexity, the right derivative at 0 is bounded above by every secant slope issuing from 0. Hence, $-1 \leq \partial_K^+ f_m(0) \leq s_{m,0}(y)$. Moreover, the definition of $\widehat{y}_{m,1}$ gives $s_{m,0}(y) = s_{m,1}(y)$. Therefore, $-1 \leq s_{m,0}(y) = s_{m,1}(y) \leq \cdots \leq s_{m,N+1}(y) \leq 0$. It remains to verify the calendar inequalities. Fix $m \in [M-1]_+$ and $n \in [N]_+$. Since $T_m < T_{m+1}$ and $\mathfrak{I}$ is non-decreasing in maturity by Proposition 2.5, we have

$$y_{m,n} = \mathfrak{I}(T_m, K_n) \leq \mathfrak{I}(T_{m+1}, K_n) = y_{m+1,n}. \tag{84}$$

Thus all the inequalities defining $\mathfrak{C}_{\mathbf{T},\mathbf{K}}$ hold, and therefore $y \in \mathfrak{C}_{\mathbf{T},\mathbf{K}}$.

Now, we show that (ii) implies (i). Suppose that $y \in \mathfrak{C}_{\mathbf{T},\mathbf{K}}$. For each $m \in [M]_+$, let $\Phi_m^y : [0,\infty) \to \mathbb{R}$ be the continuous piecewise-affine interpolant of the augmented datapoints: $\big(\widehat{K}_j, \widehat{y}_{m,j}\big)$ for each $j \in [N+2]$ and extended constantly by zero to the right of K_{R}; i.e. for each $K \in [\widehat{K}_j, \widehat{K}_{j+1}]$ with $j = 0, \ldots, N+1$ we define

$$\Phi_m^y(K) \overset{\text{def.}}{=} \frac{\widehat{K}_{j+1} - K}{\widehat{K}_{j+1} - \widehat{K}_j} \widehat{y}_{m,j} + \frac{K - \widehat{K}_j}{\widehat{K}_{j+1} - \widehat{K}_j} \widehat{y}_{m,j+1}, \tag{85}$$

and, for $K \geq K_{\mathrm{R}}$, we set $\Phi_m^y(K) \overset{\text{def.}}{=} 0$. Then, by construction we have that: $\Phi_m^y(\widehat{K}_j) = \widehat{y}_{m,j}$ for each $j \in [N+2]$. On the open sub-interval $(\widehat{K}_j, \widehat{K}_{j+1})$, the derivative of Φ_m^y is $\partial_K \Phi_m^y(K) = s_{m,j}(y)$. Now, since $y \in \mathfrak{C}_{\mathbf{T},\mathbf{K}}$, these slopes satisfy

$$-1 \leq s_{m,0}(y) = s_{m,1}(y) \leq \cdots \leq s_{m,N+1}(y) \leq 0. \tag{86}$$

The slopes of Φ_m^y are therefore non-decreasing as K increases. In addition, the slope of the constant extension on (K_{R}, ∞) is zero, which is no smaller than $s_{m,N+1}(y)$. It follows that Φ_m^y is convex on all of $[0,\infty)$. Every affine-piece slope in (86) is non-positive; whence, Φ_m^y is non-increasing. Since $\Phi_m^y(K_{\mathrm{R}}) = 0$, we consequently obtain $\Phi_m^y(K) \geq 0$ for every $0 \leq K \leq K_{\mathrm{R}}$. Together with the zero extension, this shows that $\Phi_m^y : [0,\infty) \mapsto [0,\infty)$. By construction, $\Phi_m^y(0) = 1$, and Φ_m^y is affine on $[0, K_{\mathrm{L}}]$ with slope $s_{m,0}(y)$. Consequently, $\partial_K^+ \Phi_m^y(0) = s_{m,0}(y) \geq -1$. Moreover, convexity gives $\Phi_m^y(K) \geq 1 + s_{m,0}(y)K \geq 1 - K$ for every $K \geq 0$. Since Φ_m^y is non-negative,

$$(1-K)^+ \leq \Phi_m^y(K) \qquad \text{for every } K \geq 0. \tag{87}$$

Finally, the zero extension gives: $\lim_{K \uparrow \infty} \Phi_m^y(K) = 0$.

Next, we prove that the interpolated maturity slices are pointwise ordered. Fix $m \in [M-1]_+$ and define: $D_m^y(K) \overset{\text{def.}}{=} \Phi_{m+1}^y(K) - \Phi_m^y(K)$; then, the function D_m^y is affine on each interval $[\widehat{K}_j, \widehat{K}_{j+1}]$. At the boundary strikes, $D_m^y(\widehat{K}_0) = D_m^y(\widehat{K}_{N+2}) = 0$. At the auxiliary left strike, $D_m^y(K_{\mathrm{L}}) =$

$\frac{K_{\mathrm{L}}}{K_1}(y_{m+1,1} - y_{m,1}) \geq 0$. At every quoted strike, we have that $D^y_m(K_n) = y_{m+1,n} - y_{m,n} \geq 0$ for $n \in [N]_+$; where, the inequality follows from the calendar constraints defining $\mathfrak{C}_{\mathbf{T},\mathbf{K}}$. Thus, D^y_m is nonnegative at both endpoints of every interval $[\widehat{K}_j, \widehat{K}_{j+1}]$. Since D^y_m is affine on that interval, it follows that

$$\Phi^y_m(K) \leq \Phi^y_{m+1}(K) \qquad \text{for every } K \geq 0. \tag{88}$$

Set $T_0 \overset{\text{def.}}{=} 0$ and $\Phi^y_0(K) \overset{\text{def.}}{=} (1-K)^+$. By (87), $\Phi^y_0 \leq \Phi^y_1$ pointwise. For $m \in [M-1]$ and $T \in [T_m, T_{m+1}]$, set $\lambda_m(T) \overset{\text{def.}}{=} \frac{T-T_m}{T_{m+1}-T_m} \in [0,1]$. Define $\mathfrak{I}_y : [0,\infty)^2 \to [0,\infty)$ by

$$\mathfrak{I}_y(T,K) \overset{\text{def.}}{=} \begin{cases} \big(1-\lambda_m(T)\big)\Phi^y_m(K) + \lambda_m(T)\Phi^y_{m+1}(K), & T_m \leq T \leq T_{m+1}, \quad m \in [M-1], \\ \Phi^y_M(K), & T \geq T_M. \end{cases} \tag{89}$$

The definitions agree at every maturity knot. Hence $\mathfrak{I}_y$ is continuous. For every fixed K, the map $T \mapsto \mathfrak{I}_y(T,K)$ is piecewise affine. For every fixed T, the map $K \mapsto \mathfrak{I}_y(T,K)$ is a linear combination of two piecewise-affine functions having the same breakpoints and is therefore piecewise affine. At a quoted maturity T_m, $\mathfrak{I}_y(T_m,K) = \Phi^y_m(K)$. Consequently, for every $m \in [M]_+$ and $n \in [N]_+$, we have

$$\mathfrak{I}_y(T_m,K_n) = \Phi^y_m(K_n) = y_{m,n}. \tag{90}$$

Similarly: $\mathfrak{I}_y(T_m,0) = 1$, $\mathfrak{I}_y(T_m,K_{\mathrm{L}}) = 1 + (K_{\mathrm{L}}/K_1)(y_{m,1}-1)$, and $\mathfrak{I}_y(T_m,K_{\mathrm{R}}) = 0$. It remains to verify the six conditions in Proposition 2.5. First, every maturity slice appearing in (89) takes value 1 at $K=0$. Therefore,

$$\mathfrak{I}_y(T,0) = 1 \qquad \text{for every } T \geq 0. \tag{91}$$

Second, every learned slice has right derivative $s_{m,0}(y) \geq -1$ at strike zero, and the intrinsic slice has right derivative -1. Therefore each maturity interpolation has

$$\partial^+_K \mathfrak{I}_y(T,0) \geq -1, \qquad \mathfrak{I}_y(T,K) \geq (1-K)^+. \tag{92}$$

Thirdly, each Φ^y_m vanishes on $[K_{\mathrm{R}},\infty)$. Therefore, $\mathfrak{I}_y(T,K) = 0$ for every $T \geq 0$ and each $K \geq K_{\mathrm{R}}$; implying, in turn, that $\lim_{K\uparrow\infty} \mathfrak{I}_y(T,K) = 0$ for each $T \geq 0$. Fourth, for every fixed T, the map $K \mapsto \mathfrak{I}_y(T,K)$ is either one of the convex functions Φ^y_m or a convex combination of two such functions. It is therefore convex. Finally, fix $K \geq 0$. On every interval $(T_m, T_{m+1}),$, (88) implies that

$$\partial_T \mathfrak{I}_y(T,K) = \frac{\Phi^y_{m+1}(K) - \Phi^y_m(K)}{T_{m+1}-T_m} \geq 0 \tag{93}$$

On $[0,T_1]$, maturity monotonicity follows from $\Phi^y_0 \leq \Phi^y_1$, while the surface is constant in maturity on $[T_M,\infty)$. Moreover, $\mathfrak{I}_y(0,K) = \Phi^y_0(K) = (1-K)^+$, so condition (vi) holds . Hence $\mathfrak{I}_y(T,K) \leq \mathfrak{I}_y(S,K)$ for each $0 \leq T \leq S$. All six conditions in Proposition 2.5 are satisfied. Therefore, $\mathfrak{I}_y$ is an arbitrage-free call-price surface, and item (i) follows.

Having verified the equivalence of (i) and of (ii), it now only remains to verify our statement about the polytope $\mathfrak{C}_{\mathbf{T},\mathbf{K}}$. For each fixed $m \in [M]_+$ and $j = 0,\ldots,N+1$, the map $y \mapsto s_{m,j}(y)$ is affine in the entries of y, because the augmented strikes are fixed and all augmented quote maps are affine in y. Consequently, every inequality

$$s_{m,j}(y) \leq s_{m,j+1}(y) \tag{94}$$

and every calendar inequality

$$y_{m,n} \leq y_{m+1,n} \tag{95}$$

defines a closed affine half-space in $\mathbb{R}^{M\times N}$. It follows that $\mathfrak{C}_{\mathbf{T},\mathbf{K}}$ is an intersection of finitely many closed affine half-spaces. Therefore, it is closed and convex. The slope inequalities also imply boundedness. Indeed, since every $s_{m,j}(y) \leq 0$, we have

$$\widehat{y}_{m,j+1} - \widehat{y}_{m,j} = s_{m,j}(y)\big(\widehat{K}_{j+1} - \widehat{K}_j\big) \leq 0.$$

Hence, $1 = \widehat{y}_{m,0} \geq \widehat{y}_{m,1} \geq \cdots \geq \widehat{y}_{m,N+2} = 0$. In particular, $0 \leq y_{m,n} \leq 1$ holds for every $m \in [M]_+$ and $n \in [N]_+$. Thus $\mathfrak{C}_{\mathbf{T},\mathbf{K}}$ is bounded. It remains to show that the set is non-empty. Define

$$y^\star_{m,n} \stackrel{\text{def.}}{=} (1 - K_n)^+, \qquad m \in [M]_+, \quad n \in [N]_+. \tag{96}$$

The convex function $\varphi(K) \stackrel{\text{def.}}{=} (1-K)^+$ has nondecreasing secant slopes on the grid $0, K_1, \ldots, K_N, K_{\mathrm{R}}$, all in $[-1, 0]$. Inserting K_{L} with its prescribed quote subdivides the first secant without changing its slope. Hence

$$-1 \leq s_{m,0}(y^\star) = s_{m,1}(y^\star) \leq \cdots \leq s_{m,N+1}(y^\star) \leq 0. \tag{97}$$

The values $y^\star_{m,n}$ do not depend on m, and hence all calendar inequalities hold with equality. Thus, $y^\star \in \mathfrak{C}_{\mathbf{T},\mathbf{K}}$ implying that our polytope is non-empty. Compactness gives a minimizer of $y \mapsto \|z - y\|_2^2$ for every z. Strict convexity of this objective on the convex polytope gives uniqueness, proving the projection assertion. □

We now, nearly directly, obtain the validity of our arbitrage-free decoder.

Proof of Proposition 3.8. Set $\delta_{\mathrm{grid}} \stackrel{\text{def.}}{=} \frac{1}{2} \min\{T_1, K_1, \min_{m \in [M-1]_+} |T_{m+1} - T_m|, \min_{n \in [N-1]_+} |K_{n+1} - K_n|\}$. Then $\delta_{\mathrm{grid}} > 0$, and, for every $0 < \delta \leq \delta_{\mathrm{grid}}$, the grid correction in Definition 3.7 leaves the prescribed maturity and strike grids unchanged.

By Proposition 2.5, for every $m \in [M]_+$ the map $K \mapsto C(T_m, K)$ is convex and non-increasing, satisfies $C(T_m, 0) = 1$, and has right derivative at 0 bounded below by -1. Consequently,

$$-1 \leq \frac{C(T_m, K_1) - 1}{K_1},$$

and the consecutive secant slopes through $(0, 1), (K_1, z_{m,1}), \ldots, (K_N, z_{m,N})$ are non-decreasing and non-positive. By the definition of the auxiliary quote at K_{L}, the first two augmented secant slopes coincide. Let $d_m \stackrel{\text{def.}}{=} (z_{m,N} - z_{m,N-1})/(K_N - K_{N-1}) \leq 0$. If $z_{m,N} > 0$, then $d_m < 0$: otherwise convexity and monotonicity force the slice to remain constant and positive beyond K_N, contradicting its zero right-wing limit. If $z_{m,N} = 0$, the terminal extension has slope zero. Consequently, the virtual strike K_{R} may be chosen sufficiently large that the final secant slope from $(K_N, z_{m,N})$ to $(K_{\mathrm{R}}, 0)$ is no smaller than the preceding secant slope. Since there are only finitely many maturity slices and $K_{\mathrm{R}} = K_N + 1 + \delta^{-1} \uparrow \infty$ as $\delta \downarrow 0$, there exists $0 < \delta_0 \leq \delta_{\mathrm{grid}}$ such that $\mathbf{z} \in \mathfrak{C}_{\mathbf{T},\mathbf{K},\delta}$ for every $0 < \delta \leq \delta_0$. Hence $P_{\mathbf{T},\mathbf{K},\delta}(\mathbf{z}) = \mathbf{z}$. Each component $\mathbf{\Psi}(\cdot)_m$, $m \in [M]_+$, is precisely the piecewise-affine strike interpolant through the corresponding augmented quote data, while $\mathbf{\Psi}(\cdot)_0 = (1 - \cdot)^+$. Furthermore, $\Lambda(T_m) = e_m$ for every $m \in [M]$, where $T_0 = 0$, and Λ provides the linear barycentric interpolation between consecutive maturities. Consequently, $\Psi_{\mathbf{T},\mathbf{K},\mathbf{z}}$ coincides with the canonical piecewise-linear interpolator constructed in Proposition 3.6, including the interpolation from the intrinsic-value slice at $T = 0$ to the first learned slice at $T = T_1$. It follows that $\Psi_{\mathbf{T},\mathbf{K},\mathbf{z}}(T_m, K_n) = C(T_m, K_n)$ for every $m \in [M]_+$ and $n \in [N]_+$, and shows that $\Psi_{\mathbf{T},\mathbf{K},\mathbf{z}}$ is an arbitrage-free call surface. □

Proof of Proposition 3.9. The no-arbitrage and local right-wing arguments below apply for each fixed $\delta > 0$. The interpolation argument chooses δ after the target and grid, and therefore requires a family allowing δ to vary. The fixed-δ interpretation of the family assertion and its singleton-grid cases remain pending the separately proposed parameterization.

Fix $R, S > 0$ and integers $M, N \geq 2$. For each $C \in \mathcal{Y}$, choose the parameter θ_C consisting of the uniform maturity and strike grids together with the sampled quote matrix $\big(C(t_m, K_n)\big)_{m,n=1}^{M,N}$. Choose $\delta > 0$ sufficiently small for this target and grid as in Proposition 3.8. The resulting decoded surface then interpolates C at every prescribed positive-maturity grid point. Moreover, $\phi_{\theta_C}(0, K) = (1 - K)^+ = C(0, K)$ for every $K \geq 0$. Thus the decoder also interpolates the prescribed time-zero grid values, proving the interpolation assertion for this target and grid with the selected tolerance. This does not prove (i) for one fixed positive tolerance or supply the singleton-grid wrapper.

For an arbitrary parameter θ, the quote matrix is first projected onto $\mathfrak{C}_{\mathbf{T},\mathbf{K},\delta}$. Proposition 3.6 therefore implies that the learned strike slices are convex, non-increasing, ordered in maturity, and dominate the intrinsic-value slice. The decoder interpolates from $(1 - K)^+$ at $T = 0$ through these

ordered slices. Hence all six conditions of Proposition 2.5 hold, proving (ii) of Definition 3.3. Finally, fix $A > 0$ and let $\theta \in [-A, A]^{D_{N,M,R,S}}$. Since the ordered strike grid is obtained from finitely many increments of the form $\max\{|k_j - k_{j-1}|, \delta\}$, there exists a finite constant $K_{\mathrm{R}}^{A,R,S,N,M}$ such that $K_{\mathrm{R}}(\theta) \le K_{\mathrm{R}}^{A,R,S,N,M}$ for every such θ. Hence $\phi_\theta^{(N,M,R,S)}(T,K) = 0$ for every $T \ge 0$, every $K \ge K_{\mathrm{R}}^{A,R,S,N,M}$, and every $\theta \in [-A, A]^{D_{N,M,R,S}}$. Thus (21) holds for any $\rho > p$, after increasing the constant on the compact interval $[R+1, K_{\mathrm{R}}^{A,R,S,N,M}]$, which proves (iii) of Definition 3.3. □

We also prove the following uniform-controlled tailed behaviour of these PWL surfaces.

Proposition A.3 (Uniform Tail-Control for PWL AF-Decoder)**.** *Let $p, \lambda, B > 0$ and let $\mathcal{Y} \subseteq \mathcal{C}_{AF}^{p,\lambda}$ be uniformly controlled with $\sup_{C\in\mathcal{Y}} \|C\|_{p,\lambda} \le B$. For $R, S > 0$ and integers $N, M \ge 2$, and in the notation of Definition 3.7 set*

$$\phi_\theta^{(N,M,R,S)} \overset{\text{def.}}{=} \phi_{e^{-\xi}}^{(M,N)}(\mathbf{t}, \mathbf{k}, \mathbf{z}), \qquad \theta = (\xi, \mathbf{t}, \mathbf{k}, \mathbf{z}) \in \mathbb{R}^{1+M+N+MN}.$$

For every $R, S, \eta > 0$, there exist common integers $N, M \ge 2$ with $R/N + S/M \le \eta$ such that every $C \in \mathcal{Y}$ admits θ_C satisfying (42), with $B_{\mathrm{dec}} = \mathrm{e}\, 2^p \max\{1, B\}$.

Proof. Fix $R, S, \eta > 0$, and set $\overline{B} \overset{\text{def.}}{=} \max\{1, B\}$, $c \overset{\text{def.}}{=} \min\{1/2, \eta/2\}$, and $a \overset{\text{def.}}{=} 1 - c$. Choose $L \overset{\text{def.}}{=} \max\left\{R, 2, a^{-1}\left(\frac{\overline{B}\, e^{\lambda S}}{c}\right)^{1/p}\right\}$, $N \overset{\text{def.}}{=} \max\left\{2, \left\lceil \frac{L}{\min\{1,\eta/2\}} \right\rceil\right\}$, and $M \overset{\text{def.}}{=} \max\left\{2, \left\lceil \frac{S}{\min\{1/\lambda,\eta/2\}} \right\rceil\right\}$. Put $h \overset{\text{def.}}{=} L/N$, $\Delta \overset{\text{def.}}{=} S/M$, and choose the internal grids $T_m \overset{\text{def.}}{=} m\Delta$ for $0 \le m \le M$ and $K_n \overset{\text{def.}}{=} nh$ for $0 \le n \le N$. Thus $h \le \min\{1, \eta/2\}$, $\Delta \le \min\{1/\lambda, \eta/2\}$, and $R/N + S/M \le h + \Delta \le \eta$. For $C \in \mathcal{Y}$, define $H_C(T,K) \overset{\text{def.}}{=} a^{-1}(C(T, aK) - c)^+$. Convexity and monotonicity of $C(T, \cdot)$ give

$$C(T,K) \le C(T, aK) \le aC(T,K) + c \qquad \text{and} \qquad a^{-1}(C(T,K) - c)^+ \le H_C(T,K) \le C(T,K).$$

If $C(T,K) \ge c$, then $C(T,K) - H_C(T,K) \le c(1 - C(T,K))/a \le c$; otherwise $C(T,K) - H_C(T,K) \le C(T,K) < c$. Hence $0 \le C - H_C \le c$. Moreover, $H_C(T, \cdot)$ is convex and non-increasing, $H_C(\cdot, K)$ is non-decreasing, $H_C(T, 0) = 1$, $H_C(0, K) = (1 - K)^+$, and

$$|H_C(T,K) - H_C(T,K')| \le a^{-1}|C(T, aK) - C(T, aK')| \le |K - K'|.$$

For $0 \le T \le S$ and $K \ge L$,

$$C(T, aK) \le \overline{B} e^{\lambda S}(1 + aL)^{-p} \le c, \qquad H_C(T,K) = 0.$$

Set $\delta \overset{\text{def.}}{=} \frac{1}{2}\min\{h, \Delta\}$, $\xi \overset{\text{def.}}{=} -\log\delta$, $\mathbf{t} \overset{\text{def.}}{=} (T_m)_{m=1}^M$, $\mathbf{k} \overset{\text{def.}}{=} (K_n)_{n=1}^N$, and $z_{m,n} \overset{\text{def.}}{=} H_C(T_m, K_n)$. Since $\delta < h, \Delta$, the grid correction leaves these grids unchanged. The auxiliary left quote lies on the segment from $(0,1)$ to $(K_1, z_{m,1})$, and $z_{m,N} = 0$. Convexity and the strike Lipschitz bound therefore give

$$\begin{aligned} -1 \le s_{m,0}(\mathbf{z}) = s_{m,1}(\mathbf{z}) \le \cdots \le s_{m,N+1}(\mathbf{z}) = 0 \quad & (\text{for} 1 \le m \le M), \\ z_{m,n} \le z_{m+1,n} \quad & (\text{for} 1 \le m < M,\ 1 \le n \le N). \end{aligned}$$

These are the inequalities in (24), so $\mathbf{z} \in \mathfrak{C}_{\mathbf{T},\mathbf{K},\delta}$ and $P_{\mathbf{T},\mathbf{K},\delta}(\mathbf{z}) = \mathbf{z}$. Write $U_C \overset{\text{def.}}{=} \phi_{\theta_C}^{(N,M,R,S)}$ for $\theta_C \overset{\text{def.}}{=} (\xi, \mathbf{t}, \mathbf{k}, \mathbf{z})$, and put $u_m(K) \overset{\text{def.}}{=} U_C(T_m, K)$ for $0 \le m \le M$. For the augmented knots and quotes of Definition 3.7, $S_{m,j} - S_{m,j+1} \ge 0$ for $1 \le m \le M$ and $1 \le j \le N+2$. For such m and $K \in [\widehat{K}_r, \widehat{K}_{r+1}]$ with $0 \le r \le N+1$, telescoping gives

$$\sum_{j=r+1}^{N+2} (S_{m,j} - S_{m,j+1})(\widehat{K}_j - K) = \widehat{y}_{m,r+1} + S_{m,r+1}(\widehat{K}_{r+1} - K).$$

Thus (26) gives $u_0(K) = (1 - K)^+$; for $m \ge 1$, u_m is the linear interpolant of $H_C(T_m, \cdot)$ on $(K_n)_{n=0}^N$ and vanishes on $[L, \infty)$. In particular, the virtual segment $[L, K_{\mathrm{R}}]$ is identically zero.

For $1 \le m \le M$, $1 \le n \le N$, and $K \in [K_{n-1}, K_n]$, convexity and monotonicity yield $H_C(T_m, K) \le u_m(K) \le H_C(T_m, K_{n-1})$, whence

$$0 \le u_m(K) - H_C(T_m, K) \le H_C(T_m, K_{n-1}) - H_C(T_m, K) \le h.$$

Since $0 \le H_C \le C$ and $0 \le C - H_C \le c$, it follows that $|u_m(K) - C(T_m, K)| \le h + c \le \eta$ on $[0, L]$. At $m = 0$ the error is zero. Since $R \le L$,

$$\max_{\substack{0 \le m \le M \\ 0 \le n \le N}} |U_C(mS/M, nR/N) - C(mS/M, nR/N)| \le \eta.$$

For $1 \le m \le M$, $1 \le n \le N$, and $K \in [K_{n-1}, K_n]$, the weighted bound and $h \le 1$ give

$$(1+K)^p u_m(K) \le (1+K)^p H_C(T_m, K_{n-1}) \le \overline{B} e^{\lambda T_m} \left(\frac{1+K}{1+K_{n-1}} \right)^p \le 2^p \overline{B} e^{\lambda T_m}.$$

This also holds for $K \ge L$ because $u_m(K) = 0$, and for $m = 0$ because $u_0 = C(0, \cdot)$. For $T \in [T_m, T_{m+1}]$, $0 \le m < M$, put $b \overset{\text{def.}}{=} (T - T_m)/\Delta$. The maturity weights in (26) give $U_C(T, K) = (1-b)u_m(K) + b u_{m+1}(K)$, so

$$e^{-\lambda T}(1+K)^p U_C(T, K) \le 2^p \overline{B}\big((1-b)e^{\lambda(T_m - T)} + b e^{\lambda(T_{m+1} - T)}\big) \le 2^p \overline{B} e^{\lambda \Delta} \le \mathrm{e}\, 2^p \overline{B}.$$

For $T \ge S$, the same formula gives $U_C(T, K) = u_M(K)$ and

$$e^{-\lambda T}(1+K)^p U_C(T, K) \le 2^p \overline{B} e^{-\lambda(T-S)} \le 2^p \overline{B}.$$

Therefore $\|U_C\|_{p,\lambda} \le B_{\text{dec}}$. The integers N, M, the grids, and δ are independent of C, which proves the required simultaneous selection. □

A.1.1 Proofs for SANOS AF-Decoder

Proposition A.4 (Properties of the Black-Scholes activation)**.** *For every $\nu > 0$, $\mathrm{BS}_\nu \in C^\infty\big((0,\infty),[0,\infty)\big) \cap C\big([0,\infty),[0,\infty)\big)$ and BS_ν is increasing, convex, and 1-Lipschitz. Moreover,*

$$0 \le \mathrm{BS}_\nu(x) - \mathrm{ReLU}(x) \le \frac{\nu}{\sqrt{2\pi}}, \qquad x \in \mathbb{R}. \tag{98}$$

and the map $\mathbb{R}^2 \ni (\nu, x) \to \mathrm{BS}_\nu(x)$ is continuous.

Proof. Fix $\nu > 0$. For $x > -1$, set $k \overset{\text{def.}}{=} 1 + x$ and define

$$d_1(x,\nu) \overset{\text{def.}}{=} \frac{-\log k + \frac{1}{2}\nu^2}{\nu} \qquad \text{and} \qquad d_2(x,\nu) \overset{\text{def.}}{=} d_1(x,\nu) - \nu = \frac{-\log k - \frac{1}{2}\nu^2}{\nu}.$$

Then $\mathrm{BS}_\nu(x) = k\Phi\big(-d_2(x,\nu)\big) - \Phi\big(-d_1(x,\nu)\big)$. Recall the Black-Scholes strike derivatives

$$\partial_x \mathrm{BS}_\nu(x) = \Phi\big(-d_2(x,\nu)\big) = \Phi\left(\frac{\log(1+x) + \frac{1}{2}\nu^2}{\nu}\right) \in (0,1), \tag{99}$$

$$\partial_{xx} \mathrm{BS}_\nu(x) = \frac{\varphi\big(d_2(x,\nu)\big)}{(1+x)\nu} > 0, \tag{100}$$

where φ denotes the standard normal density. Hence BS_ν is increasing, convex, and 1-Lipschitz on $(-1,\infty)$. As $x \downarrow -1$, $0 \le \mathrm{BS}_\nu(x) \le 1 + x$ and $\partial_x \mathrm{BS}_\nu(x) \to 0$. Every derivative of order at least two on $(-1,\infty)$ is a finite linear combination of terms of the form

$$(1+x)^{-a} p\big(\log(1+x)\big) \exp\left(-\frac{\big(\log(1+x) + b\big)^2}{2\nu^2}\right),$$

where $a \ge 0$, $b \in \mathbb{R}$, and p is a polynomial. Since the Gaussian factor decays faster than every negative power of $1+x$, each such term converges to 0 as $x \downarrow -1$. The extension by 0 on $(-\infty, -1]$ therefore belongs to $C^\infty(\mathbb{R})$. For $x \le -1$, both terms $\mathrm{BS}_\nu(x)$ and $\mathrm{ReLU}(x)$ vanish. For $x > -1$, the Black-Scholes vega identity gives, for every $u > 0$,

$$\partial_u \mathrm{BS}_u(x) = \varphi\big(d_1(x,u)\big) \le \frac{1}{\sqrt{2\pi}}.$$

Since $\mathrm{BS}_0(x) = \big((1+x)-1\big)^+ = \mathrm{ReLU}(x)$ then, the fundamental theorem of calculus implies

$$0 \le \mathrm{BS}_\nu(x) - \mathrm{ReLU}(x) = \int_0^\nu \varphi\big(d_1(x,u)\big)\,du \le \frac{\nu}{\sqrt{2\pi}}.$$

This proves (98). Lastly, the continuity at $\nu = 0$ immediately follows from the uniform convergence; and so does the continuity on $\mathbb{R}\times\mathbb{R}$ follows from (27). □

Corollary A.5 (Properties of the Black-Scholes Activation $\mathrm{Call}_\nu^{\mathrm{BS}}$)**.** *For every $\nu > 0$ and $s > 0$, the map $[0,\infty) \ni K \mapsto \mathrm{Call}_\nu^{\mathrm{BS}}(s,K)$ is non-negative, decreasing, convex, and in $C^\infty((0,\infty)) \cap C([0,\infty),[0,\infty))$. Moreover,*

$$\mathrm{Call}_\nu^{\mathrm{BS}}(s,0) = s \tag{101}$$

$$\partial_K^+ \mathrm{Call}_\nu^{\mathrm{BS}}(s,0) = -1 \tag{102}$$

$$\lim_{K\uparrow\infty} \mathrm{Call}_\nu^{\mathrm{BS}}(s,K) = 0, \tag{103}$$

and, for each $K \ge 0$, we have $0 \le \mathrm{Call}_\nu^{\mathrm{BS}}(s,K) - (s-K)^+ \le \frac{s\nu}{\sqrt{2\pi}}$.

Proof. By (28), we have $\mathrm{Call}_\nu^{\mathrm{BS}}(s,K) - (s-K)^+ = s\left[\mathrm{BS}_\nu\left(\frac{K}{s}-1\right) - \mathrm{ReLU}\left(\frac{K}{s}-1\right)\right]$. Hence the estimate $0 \le \mathrm{Call}_\nu^{\mathrm{BS}}(s,K) - (s-K)^+ \le \frac{s\nu}{\sqrt{2\pi}}$ follows from Proposition A.4. The remaining assertions are the standard shape and boundary properties of the zero-rate Black-Scholes call price with spot s, strike K, unit maturity, and volatility ν. □

For each $m \in [M]_+$ and $j \in [N+2]_+$, set $q_{m,j} \overset{\text{def.}}{=} S_{m,j} - S_{m,j+1}$. The slope inequalities defining $\mathfrak{C}_{\mathbf{T},\mathbf{K},\delta}$ imply

$$q_{m,j} \ge 0, \qquad \sum_{j=1}^{N+2} q_{m,j} = S_{m,1} \le 1, \qquad \sum_{j=1}^{N+2} q_{m,j}\widehat{K}_j = 1. \tag{104}$$

Indeed, $S_{m,N+3} = 0$, so the first sum telescopes to $S_{m,1}$, while the corrected left-slope constraint gives $0 \le S_{m,1} \le 1$. Moreover,

$$\sum_{j=1}^{N+2} q_{m,j}\widehat{K}_j = \sum_{j=1}^{N+2} (S_{m,j} - S_{m,j+1})\widehat{K}_j = \sum_{j=1}^{N+2} S_{m,j}\big(\widehat{K}_j - \widehat{K}_{j-1}\big) = \sum_{j=1}^{N+2} \big(\widehat{y}_{m,j-1} - \widehat{y}_{m,j}\big) = 1.$$

Fix $\overline{\nu} > 0$ and for every $\eta \in \mathbb{R}$ define $\nu_{\overline{\nu}}(\eta) \overset{\text{def.}}{=} \frac{\overline{\nu}}{1+e^{-\eta}}$. Thus, $0 < \nu_{\overline{\nu}}(\eta) < \overline{\nu}$ for every $\eta \in \mathbb{R}$ and $\lim_{\eta\downarrow-\infty} \nu_{\overline{\nu}}(\eta) = 0$. We next verify Proposition 3.11.

Proof of Proposition 3.11. The following argument proves (i)–(iii), interpreting the smooth maturity weights by the M coordinates displayed below, and proves the locally uniform right-wing estimate. The final surface-decoder family assertion additionally requires the separately proposed parameterization allowing δ to vary.

Write Ψ_ν for the full surface and Γ_ν for the raw mixture. By (104), the strike slices are nonnegative combinations of Black–Scholes calls, with

$$\mathbf{\Psi}_\nu(0)_m = \sum_j q_{m,j}\widehat{K}_j = 1, \tag{105}$$

$$\partial_K^+ \mathbf{\Psi}_\nu(0)_m = -\sum_j q_{m,j} = -S_{m,1} \ge -1. \tag{106}$$

They are convex, non-increasing, smooth for $K > 0$, and tend to zero as $K \to \infty$. For $Z \sim \mathcal{N}(0,1)$ put $X_\nu = e^{\nu Z - \nu^2/2}$. The call representation gives

$$\mathbf{\Psi}_\nu(K)_m = \mathbb{E}[X_\nu \mathbf{\Psi}_0(K/X_\nu)_m]. \tag{107}$$

The PWL slices are ordered by Proposition 3.6, so positivity of this operator yields

$$\mathbf{\Psi}_\nu(K)_m \le \mathbf{\Psi}_\nu(K)_{m+1}. \tag{108}$$

Corollary A.5 also gives

$$0 \le \mathbf{\Psi}_\nu(K)_m - \mathbf{\Psi}_0(K)_m \le \epsilon_\nu, \qquad \epsilon_\nu \overset{\text{def.}}{=} \nu/\sqrt{2\pi}. \tag{109}$$

Let $h_\nu = \partial_x \mathrm{BS}_\nu$ and $\Delta_m = T_{m+1} - T_m$. Then

$$a_{m,\nu}(T) = \Delta_m^{-1} \int_{T-T_{m+1}}^{T-T_m} h_\nu(u)\, du. \tag{110}$$

Since h_ν is non-decreasing with values in $[0,1]$, adjacent ordered intervals give $1 \ge a_{1,\nu} \ge \cdots \ge a_{M-1,\nu} \ge 0$. Thus

$$\Lambda_\nu = \big(1 - a_{1,\nu}, (a_{m-1,\nu} - a_{m,\nu})_{m=2}^{M-1}, a_{M-1,\nu}\big)^\top \in \Delta_M. \tag{111}$$

For $M = 2$, this vector is $(1 - a_{1,\nu}, a_{1,\nu})^\top$. Moreover,

$$\partial_T a_{m,\nu}(T) = \Delta_m^{-1}[h_\nu(T - T_m) - h_\nu(T - T_{m+1})] \ge 0. \tag{112}$$

The raw mixture satisfies

$$\Gamma_\nu(T,K) = \mathbf{\Psi}_\nu(K)_1 + \sum_{m=1}^{M-1} a_{m,\nu}(T)[\mathbf{\Psi}_\nu(K)_{m+1} - \mathbf{\Psi}_\nu(K)_m]. \tag{113}$$

Hence it is non-decreasing in maturity. Its strike slices are convex and non-increasing, with $\Gamma_\nu(T,0) = 1$, $\partial_K^+ \Gamma_\nu(T,0) = -\sum_m \Lambda_\nu(T)_m S_{m,1} \ge -1$, and zero right-wing limit.

We next justify the blend. Put $d_\nu = \mathrm{BS}_\nu(0)$ and $r_\nu(u) = \mathrm{BS}_\nu(u) - \mathrm{BS}_\nu(u-1)$ for $u \ge 0$. The put representation of the activation gives

$$r_\nu(u) = \mathbb{E}[\min\{(1 + u - X_\nu)^+, 1\}], \qquad r_\nu(0) = d_\nu, \qquad 1 - d_\nu = \mathbb{E}[\min\{X_\nu, 1\}] > 0.$$

Thus r_ν is non-decreasing, $d_\nu \le r_\nu(u) \le 1$, and $b_\nu(T) = [r_\nu(T/T_1) - d_\nu]/(1 - d_\nu)$ is non-decreasing with values in $[0,1]$ and $b_\nu(0) = 0$. For $\nu > 0$, the strict increase of h_ν on $(-1,\infty)$ gives $b_\nu(T) > 0$ whenever $T > 0$. Since each PWL slice dominates $(1-K)^+$, the positive operator and volatility monotonicity give

$$B_\nu(T,K) \le \mathrm{Call}_\nu^{\mathrm{BS}}(1,K) \le \Gamma_\nu(T,K). \tag{114}$$

Both B_ν and Γ_ν are non-decreasing in maturity. For $0 \le T \le S$,

$$\begin{aligned}\Psi_\nu(S,K) - \Psi_\nu(T,K) = {} & (1 - b_\nu(S))[B_\nu(S,K) - B_\nu(T,K)] \\ & + b_\nu(S)[\Gamma_\nu(S,K) - \Gamma_\nu(T,K)] \\ & + [b_\nu(S) - b_\nu(T)][\Gamma_\nu(T,K) - B_\nu(T,K)] \ge 0.\end{aligned}$$

For every maturity the blend is a convex combination of normalized, convex, non-increasing strike slices with right derivative at zero at least -1 and zero right-wing limit. Also $\Psi_\nu(0,K) = (1-K)^+$. The activation is smooth on all of $\mathbb{R}$; the baseline call has positive volatility for $T > 0$, and the Black–Scholes formula is jointly smooth when volatility and strike are positive. This proves interior smoothness. Continuity at $T = 0$ follows from $b_\nu(T) \to 0$ and $|\mathrm{Call}_{\nu b_\nu(T)}^{\mathrm{BS}}(1,K) - (1-K)^+| \le \epsilon_\nu b_\nu(T)$, uniformly in K. Continuity at strike zero follows from the common strike Lipschitz bound. Proposition 2.5 now proves (i) and (ii).

For (iii), assume $\epsilon_\nu \le 1/2$. At $\nu = 0$, $b_0(T) = \min\{T/T_1, 1\}$ and the full surface is exactly the PWL interpolator. Equation (98) gives $|r_\nu - r_0| \le \epsilon_\nu$ and $d_\nu \le \epsilon_\nu$, hence

$$|b_\nu - b_0| \le \frac{2\epsilon_\nu}{1 - d_\nu} \le 4\epsilon_\nu.$$

Also $|a_{m,\nu} - a_{m,0}| \le 2\epsilon_\nu/\Delta_m$, so

$$\|\Lambda_\nu - \Lambda_0\|_1 \le 4\epsilon_\nu \sum_{m=1}^{M-1} \Delta_m^{-1}. \tag{115}$$

Since the maturity weights lie in the simplex and $0 \le \mathbf{\Psi}_0 \le 1$, it follows that $\|\Gamma_\nu - \Gamma_0\|_\infty \le \epsilon_\nu(1 + 4\sum_m \Delta_m^{-1})$. Put $I(K) = (1-K)^+$. The blend is controlled by the identity

$$\Psi_\nu - \Psi_0 = (1-b_\nu)(B_\nu - I) + b_\nu(\Gamma_\nu - \Gamma_0) + (b_\nu - b_0)(\Gamma_0 - I).$$

Here $\|B_\nu - I\|_\infty \le \epsilon_\nu$ and $0 \le \Gamma_0 - I \le 1$. Consequently,

$$\|\Psi_\nu - \Psi_0\|_\infty \le \epsilon_\nu \left(6 + 4\sum_{m=1}^{M-1} \Delta_m^{-1}\right) \le \epsilon_\nu \left(10 + \frac{4(M-1)}{\Delta_T^{\min}}\right).$$

In particular, the sufficient choice

$$0 < \nu < \overline{\nu}, \qquad \nu \le \min\left\{\frac{\sqrt{2\pi}}{2}, \frac{\sqrt{2\pi}\,\Delta_T^{\min}}{10\Delta_T^{\min} + 4(M-1)}\varepsilon\right\} \tag{116}$$

proves (34).

For the locally uniform right-wing bound, fix $\delta > 0$ and $A > 0$, and restrict $(\mathbf{t}, \mathbf{k}, \mathbf{z}, \zeta)$ to the parameter box $[-A, A]^{M+N+MN+1}$. The corrected strikes satisfy

$$\widehat{K}_j \le K_{\mathrm{R}} \le N \max\{2A, \delta\} + 1 + \delta^{-1} \stackrel{\text{def.}}{=} H_{A,\delta}.$$

For $\rho > p$, $s, K > 0$ and $0 < \nu < \overline{\nu}$, the moment formula for X_ν gives

$$\mathrm{Call}_\nu^{\mathrm{BS}}(s, K) \le K^{-\rho} s^{\rho+1} \mathbb{E}[X_\nu^{\rho+1}] \le K^{-\rho} s^{\rho+1} e^{\rho(\rho+1)\overline{\nu}^2/2}.$$

Since $\sum_j q_{m,j}\widehat{K}_j = 1$, one has $\sum_j q_{m,j}\widehat{K}_j^{\rho+1} \le H_{A,\delta}^\rho$. Using $\Lambda_\nu(T) \in \Delta_M$ and $0 \le \Psi_\nu \le \Gamma_\nu$ therefore yields

$$\Psi_\nu(T, K) \le H_{A,\delta}^\rho e^{\rho(\rho+1)\overline{\nu}^2/2} K^{-\rho} \le 2^\rho H_{A,\delta}^\rho e^{\rho(\rho+1)\overline{\nu}^2/2}(1+K)^{-\rho}, \qquad T \ge 0, \quad K \ge R+1.$$

This proves the locally uniform right-wing estimate at each fixed $\delta > 0$. □

We now prove the analogous result to Proposition A.3 for the smooth SANOS-type decoders.

Corollary A.6 (Uniform Right-Tail Control for SANOS-type decoders)**.** *Under the hypotheses of Proposition A.3, fix $\overline{\nu} > 0$. Then the SANOS-type family parameterized by $\theta = (\xi, \mathbf{t}, \mathbf{k}, \mathbf{z}, \zeta)$, with $\delta = e^{-\xi}$ as in Proposition 3.11, satisfies Assumption C.1, with $B_{\mathrm{dec}}^{\mathrm{SANOS}} = B_{\mathrm{dec}}^{\mathrm{PWL}} + 1$, where $B_{\mathrm{dec}}^{\mathrm{PWL}} \stackrel{\text{def.}}{=} \mathrm{e}\, 2^p \max\{1, B\}$.*

Proof. The argument below chooses δ after R, S, η and establishes simultaneous selection for that choice. It does not establish one fixed positive δ valid for all R, S, η; the quantifier in the unchanged statement remains pending.

Fix $R, S, \eta > 0$ and apply the construction of Proposition A.3 with accuracy $\eta/2$. Let U_C denote the resulting PWL approximants, with common grids, tolerance δ, and terminal internal strike $L \ge 2$. Thus $U_C(T, K) = 0$ for $K \ge L$, $\|U_C\|_{p,\lambda} \le B_{\mathrm{dec}}^{\mathrm{PWL}}$, and the grid error is at most $\eta/2$. Let $\Psi_{\nu,C}$ be the SANOS-type surface obtained from these same grids and quote arrays with smoothing parameter $0 < \nu \le \min\{1, \overline{\nu}/2\}$. Put $\Delta \stackrel{\text{def.}}{=} S/M$ and $c_{M,\Delta} \stackrel{\text{def.}}{=} (10 + 4(M-1)/\Delta)/\sqrt{2\pi}$. Proposition 3.11(iii) gives

$$\sup_{T,K \ge 0} |\Psi_{\nu,C}(T, K) - U_C(T, K)| \le c_{M,\Delta}\nu,$$

uniformly in $C \in \mathcal{Y}$.

Set $r \stackrel{\text{def.}}{=} p + 1$ and $D \stackrel{\text{def.}}{=} L^r e^{r(r+1)/2}$. For $Z \sim \mathcal{N}(0,1)$, $X_v \stackrel{\text{def.}}{=} e^{vZ - v^2/2}$, $s, K > 0$, and $0 \le v \le 1$, the Black–Scholes representation gives

$$\mathrm{Call}_v^{\mathrm{BS}}(s, K) = \mathbb{E}[(sX_v - K)^+] \le s^{r+1} K^{-r} \mathbb{E}[X_v^{r+1}] \le s^{r+1} K^{-r} e^{r(r+1)/2}.$$

In the selected PWL construction, $q_{m,j} > 0$ implies $\widehat{K}_j \le L$, and (104) gives $\sum_j q_{m,j}\widehat{K}_j = 1$. Hence $\sum_j q_{m,j}\widehat{K}_j^{r+1} \le L^r$. Since the maturity weights belong to the simplex, $0 \le b_\nu \le 1$, and the remaining Black–Scholes term has spot $1 \le L$ and volatility $\nu b_\nu(T) \le 1$, (33) yields

$$0 \le \Psi_{\nu,C}(T, K) \le DK^{-r}, \qquad T \ge 0, \quad K > 0.$$

For every $A \geq L$, splitting the strike domain at A therefore gives

$$\|\Psi_{\nu,C} - U_C\|_{p,\lambda} \leq \max\left\{(1+A)^p c_{M,\Delta}\nu,\ 2^p D A^{p-r}\right\}.$$

All constants here are independent of C. Put $\kappa \stackrel{\text{def.}}{=} \min\{1, \eta e^{-\lambda S}/2\}$, choose $A \geq \max\{L, 2^p D/\kappa\}$, and then choose

$$0 < \nu \leq \min\left\{1, \frac{\overline{\nu}}{2}, \frac{\kappa}{(1+A)^p c_{M,\Delta}}\right\}.$$

Since $p - r = -1$, the same choice gives $\|\Psi_{\nu,C} - U_C\|_{p,\lambda} \leq \kappa$ for every $C \in \mathcal{Y}$. Consequently,

$$\begin{aligned} \|\Psi_{\nu,C}\|_{p,\lambda} &\leq B_{\text{dec}}^{\text{PWL}} + 1, \\ \max_{\substack{0\leq m\leq M \\ 0\leq n\leq N}} |\Psi_{\nu,C}(mS/M, nR/N) - C(mS/M, nR/N)| &\leq \eta/2 + e^{\lambda S}\kappa \leq \eta. \end{aligned}$$

The PWL mesh already satisfies $R/N + S/M \leq \eta/2$. Finally, the common smoothing logit $\zeta \stackrel{\text{def.}}{=} \log(\nu/(\overline{\nu} - \nu))$ realizes the selected ν through $\nu_{\overline{\nu}}(\zeta) = \nu$. Thus the same selected parameters satisfy both inequalities in (42). □

S1 Supplementary Material: Proof of Proposition 2.5

The following result seems to be known but difficult, at-least for us, to track down over continuous domains. We therefore present a short proof here to self-containedness.

Proof of Proposition 2.5. Suppose first that C is an arbitrage-free call surface, and let $Z_\cdot$ be a non-negative martingale representation of C. Since $Z_\cdot$ is a martingale and $Z_0 = 1$ $\mathbb{P}$-almost surely,

$$C(T,0) = \mathbb{E}_{\mathbb{P}}[Z_T] = \mathbb{E}_{\mathbb{P}}[Z_0] = 1,$$

which proves (i). For every $T, K \geq 0$ and $\delta > 0$, we have

$$-1 \leq \frac{(Z_T - K - \delta)^+ - (Z_T - K)^+}{\delta} \leq 0. \tag{117}$$

Taking expectations in (117), letting δ tend to zero from the right, and applying the dominated convergence theorem, where $\lim_{\delta \downarrow 0} \frac{(Z_T-K-\delta)^+-(Z_T-K)^+}{\delta} = -\mathbb{1}_{\{Z_T>K\}}$ pointwise, therefore yields

$$\partial_K^+ C(T,K) = -\mathbb{P}\left(Z_T > K\right). \tag{118}$$

In particular, $\partial_K^+ C(T,0) = -\mathbb{P}(Z_T > 0) \geq -1$ which proves (ii). Equation (118) also yields $\partial_K^+ C(T,0) = -1 + \mathbb{P}(Z_T = 0)$, which proves the final assertion. Since $\lim_{K\uparrow\infty} (Z_T - K)^+ = 0$ and $(Z_T - K)^+ \leq Z_T$, a similar application of the dominated convergence theorem proves (iii). For every fixed $z \geq 0$, the map $K \mapsto (z-K)^+$ is convex. Taking expectations therefore proves (iv). Let $0 \leq T_1 \leq T_2$. Since $z \mapsto (z-K)^+$ is convex, conditional Jensen's inequality and the martingale property give

$$C(T_2,K) = \mathbb{E}_{\mathbb{P}}\left[(Z_{T_2} - K)^+\right] \geq \mathbb{E}_{\mathbb{P}}\left[\left(\mathbb{E}_{\mathbb{P}}\left[Z_{T_2} \,|\, \mathcal{F}_{T_1}\right] - K\right)^+\right] = \mathbb{E}_{\mathbb{P}}\left[(Z_{T_1} - K)^+\right] = C(T_1,K).$$

Thus (v) holds. Finally, $C(0,K) = \mathbb{E}_{\mathbb{P}}\left[(Z_0 - K)^+\right] = (1-K)^+$, which proves (vi). Conversely, suppose that C satisfies (i)–(vi). For every $T \geq 0$, define $K \mapsto c_T(K) \overset{\text{def.}}{=} C(T,K)$ on $[0,\infty)$. Conditions (iii) and (iv) imply that c_T is non-increasing. Indeed, if $c_T(K_2) > c_T(K_1)$ for some $K_1 < K_2$, then convexity would imply that every subsequent secant slope is bounded below by the strictly positive number

$$\frac{c_T(K_2) - c_T(K_1)}{K_2 - K_1},$$

contradicting $\lim_{K\uparrow\infty} c_T(K) = 0$. Since c_T is convex, its right derivative $c_T'^{,+}$ exists everywhere, is right-continuous and non-decreasing. By (ii), convexity, and the monotonicity of c_T,

$$-1 \leq c_T'^{,+}(0) \leq c_T'^{,+}(K) \leq 0$$

for every $K \geq 0$. Moreover, $\lim_{K\uparrow\infty} c_T'^{,+}(K) = 0$, since a strictly negative limit would contradict $\lim_{K\uparrow\infty} c_T(K) = 0$. Define

$$F_T(x) \overset{\text{def.}}{=} \begin{cases} 0, & x < 0, \\ 1 + c_T'^{,+}(x), & x \geq 0. \end{cases}$$

Then F_T is right-continuous and non-decreasing, with $\lim_{x\downarrow-\infty} F_T(x) = 0$ and $\lim_{x\uparrow\infty} F_T(x) = 1$. Hence, F_T is the distribution function of a probability measure μ_T on $[0,\infty)$. By construction, $\mu_T((K,\infty)) = -c_T'^{,+}(K)$. Since a convex function is absolutely continuous on every compact interval, for every $0 \leq K < R$ we have $c_T(R) - c_T(K) = \int_K^R c_T'^{,+}(u)\,du$. Letting $R \uparrow \infty$ and using condition (iii), followed by Tonelli's theorem, gives

$$c_T(K) = -\int_K^\infty c_T'^{,+}(u)\,du = \int_K^\infty \mu_T((u,\infty))\,du = \int_{[0,\infty)} (x-K)^+\,\mu_T(dx).$$

In particular, condition (i) implies $\int_{[0,\infty)} x\,\mu_T(dx) = c_T(0) = 1$. Thus, every μ_T is a probability measure on $[0,\infty)$ with unit mean whose call-price function is $C(T,\cdot)$. Condition (vi) implies that

$\mu_0 = \delta_1$. Indeed, the call-price function $K \mapsto (1-K)^+$ uniquely determines the Dirac measure at 1. Now let $0 \leq T_1 \leq T_2$. By (v), we have that: for every $K \geq 0$

$$\int_{[0,\infty)} (x-K)^+ \, \mu_{T_1}(dx) \leq \int_{[0,\infty)} (x-K)^+ \, \mu_{T_2}(dx).$$

Since μ_{T_1} and μ_{T_2} have the same mean, the call-function characterization of convex order in [HR12, Proposition 2.4] yields $\mu_{T_1} \preceq_{\text{cx}} \mu_{T_2}$. Consequently, $(\mu_T)_{T \geq 0}$ is increasing in convex order. By [HR12, Theorem 4.1], this family admits a martingale $Z_\cdot = (Z_T)_{T \geq 0}$ such that $Z_T \sim \mu_T$ for every $T \geq 0$. Since every μ_T is supported on $[0,\infty)$, the martingale $Z_\cdot$ is non-negative. Moreover, $\mu_0 = \delta_1$, so $Z_0 = 1$ $\mathbb{P}$-almost surely. Finally, for every $T, K \geq 0$ we have

$$\mathbb{E}_{\mathbb{P}}\left[(Z_T - K)^+\right] = \int_{[0,\infty)} (x-K)^+ \, \mu_T(dx) = C(T,K).$$

Thus, $Z_\cdot$ is a martingale representation of C, and C is an arbitrage-free call surface. □

References

[AEKY25] Guillermo A. Alvarez, Ibrahim Ekren, Anastasis Kratsios, and Xuwei Yang. Neural operators can play dynamic Stackelberg games. *Journal of Machine Learning Research*, 26(303):1–61, 2025.

[AKP24] Beatrice Acciaio, Anastasis Kratsios, and Gudmund Pammer. Designing universal causal deep learning models: The geometric (Hyper)Transformer. *Mathematical Finance*, 34(2):671–735, 2024. doi:10.1111/mafi.12389.

[Aro50] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950. doi:10.1090/S0002-9947-1950-0051437-7.

[BDG23] Fred Espen Benth, Nils Detering, and Luca Galimberti. Neural networks in Fréchet spaces. *Annals of Mathematics and Artificial Intelligence*, 91:75–103, 2023. doi:10.1007/s10472-022-09824-z.

[BDG24] Fred Espen Benth, Nils Detering, and Luca Galimberti. Pricing options on flow forwards by neural networks in a Hilbert space. *Finance and Stochastics*, 28(1):81–121, 2024. doi:10.1007/s00780-023-00520-2.

[BDHK15] Anton Baranov, André Dumont, Andreas Hartmann, and Karim Kellay. Sampling, interpolation and Riesz bases in small Fock spaces. *Journal de Mathématiques Pures et Appliquées*, 103(6):1358–1389, 2015. doi:10.1016/j.matpur.2014.11.002.

[BDL21] Fred Espen Benth, Nils Detering, and Silvia Lavagnini. Accuracy of deep learning in calibrating HJM forward curves. *Digital Finance*, 3(3):209–248, 2021. doi:10.1007/s42521-021-00030-w.

[BFZZ25] Erhan Bayraktar, Qi Feng, Zecheng Zhang, and Zhaoyu Zhang. Deep neural operator learning for probabilistic models, 2025. arXiv:2511.07235, doi:10.48550/arXiv.2511.07235.

[BHK+26] Hans Buehler, Blanka Horvath, Anastasis Kratsios, Yannick Limmer, and Raeid Saqur. Sanos–smooth strictly arbitrage-free non-parametric option surfaces. *arXiv preprint arXiv:2601.11209*, 2026.

[BHKS21] Kaushik Bhattacharya, Bamdad Hosseini, Nikola B. Kovachki, and Andrew M. Stuart. Model reduction and neural networks for parametric PDEs. *The SMAI Journal of Computational Mathematics*, 7:121–157, 2021. doi:10.5802/smai-jcm.74.

[BJ26] Arthur Bourdon and Thibault Jeannin. Moments in rough bergomi and boundary attainment in rough heston. *arXiv preprint arXiv:2606.07482*, 2026.

[BK23] Fred Espen Benth and Paul Krühner. The filipović space and operators. In *Stochastic Models for Prices Dynamics in Energy and Commodity Markets*, Springer Finance, pages 61–91. Springer, Cham, 2023. doi:10.1007/978-3-031-40367-5_3.

[BL10] Alexander Borichev and Yurii Lyubarskii. Riesz bases of reproducing kernels in Fock type spaces. *Journal of the Institute of Mathematics of Jussieu*, 9(3):449–461, 2010. doi:10.1017/S147474800900019X.

[BR15] Hans Buehler and Evgeny Ryskin. Discrete local volatility for large time steps (extended version). *SSRN Electronic Journal*, 2015. Version dated January 5, 2017. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2642630.

[BR18] Hans Buehler and Evgeny Ryskin. Discrete local volatility for large time steps (short version). *Available at SSRN 2783409*, 2018.

[BTA04] Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, Boston, 2004. doi:10.1007/978-1-4419-9096-9.

[CGMY02] Peter Carr, Hélyette Geman, Dilip B. Madan, and Marc Yor. The fine structure of asset returns: An empirical investigation. *The Journal of Business*, 75(2):305–332, 2002. doi:10.1086/338705.

[CJB24] Vedant Choudhary, Sebastian Jaimungal, and Maxime Bergeron. Funvol: multi-asset implied volatility market simulator using functional principal components and neural sdes. *Quantitative Finance*, 24(8):1077–1103, 2024.

[CLZZ23] Yeda Cui, Lingfei Li, Gongqiu Zhang, and Wenyong Zhang. Learning pricing maps for financial derivatives with deep operator networks. *SSRN Electronic Journal*, 2023. Available at SSRN 4670716. doi:10.2139/ssrn.4670716.

[CN17] Peter Carr and Sergey Nadtochiy. Local variance gamma and explicit calibration to option prices. *Mathematical Finance*, 27(1):151–193, 2017. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/mafi.12086, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/mafi.12086, doi:10.1111/mafi.12086.

[CZ07] Julia Collins and Johannes Zimmer. An asymmetric arzela–ascoli theorem. *Topology and its Applications*, 154(11):2312–2322, 2007.

[DLDN26] Pere Díaz Lozano and Giulia Di Nunno. Deep operator BSDE: A numerical scheme to approximate solution operators. *The Annals of Applied Probability*, 36(3):2594–2625, 2026. doi:10.1214/25-AAP2289.

[FGH+17] Hartmut Führ, Karlheinz Gröchenig, Antti Haimi, Andreas Klotz, and José Luis Romero. Density of sampling and interpolation in reproducing kernel Hilbert spaces. *Journal of the London Mathematical Society*, 96(3):663–686, 2017. doi:10.1112/jlms.12083.

[Fil01] Damir Filipović. *Consistency problems for Heath-Jarrow-Morton interest rate models*, volume 1760 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2001. doi:10.1007/b76888.

[Fil09] Damir Filipović. *Term-Structure Models: A Graduate Course*. Springer Finance. Springer, Berlin, 2009. doi:10.1007/978-3-540-68015-4.

[FK24] Takashi Furuya and Anastasis Kratsios. Simultaneously solving FBSDEs with neural operators of logarithmic depth, constant width, and sub-linear rank, 2024. arXiv:2410.14788.

[FKPR25] Takashi Furuya, Anastasis Kratsios, Dylan Possamaï, and Bogdan Raonić. One model to solve them all: 2BSDE families via neural operators, 2025. arXiv:2511.01125.

[FKY25] Dena Firoozi, Anastasis Kratsios, and Xuwei Yang. Simultaneously solving infinitely many LQ mean field games in Hilbert spaces: The power of neural operators, 2025. arXiv:2510.20017, doi:10.48550/arXiv.2510.20017.

[GGP19] Stefan Gerhold, Christoph Gerstenecker, and Arpad Pinter. Moment explosions in the rough heston model. *Decisions in Economics and Finance*, 42(2):575–608, 2019.

[GJ13] Jim Gatheral and Antoine Jacquier. Arbitrage-free svi volatility surfaces, 2013. URL: https://arxiv.org/abs/1204.0646, arXiv:1204.0646.

[GJ14] Jim Gatheral and Antoine Jacquier. Arbitrage-free svi volatility surfaces. *Quantitative Finance*, 14(1):59–71, 2014.

[GKL26] Luca Galimberti, Anastasis Kratsios, and Giulia Livieri. Designing universal causal deep learning models: The case of infinite-dimensional dynamical systems from stochastic analysis. *Constructive Approximation*, pages 1–67, 2026. doi:10.1007/s00365-026-09745-8.

[GLK22] Luca Galimberti, Giulia Livieri, and Anastasis Kratsios. Designing universal causal deep learning models: The case of infinite-dimensional dynamical systems from stochastic analysis. *arXiv preprint arXiv:2210.13300*, 2022.

[GT21] Matteo Gambara and Josef Teichmann. Consistent recalibration models and deep calibration, 2021. arXiv:2006.09455.

[HR12] Francis Hirsch and Bernard Roynette. A new proof of Kellerer's theorem. *ESAIM: Probability and Statistics*, 16:48–60, 2012. URL: https://www.numdam.org/articles/10.1051/ps/2011164/, doi:10.1051/ps/2011164.

[KCR25] Anastasis Kratsios, Tin Sum Cheng, and Daniel Roy. Beyond universal approximation theorems: Algorithmic uniform approximation by neural networks trained with noisy data. *arXiv preprint arXiv:2509.00924*, 2025.

[KF25] Anastasis Kratsios and Takashi Furuya. Is in-context universality enough? MLPs are also universal in-context. *arXiv preprint arXiv:2502.03327*, 2025. doi:10.48550/arXiv.2502.03327.

[KLL+23a] Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to pdes. *Journal of Machine Learning Research*, 24(89):1–97, 2023.

[KLL+23b] Anastasis Kratsios, Chong Liu, Matti Lassas, Maarten V. de Hoop, and Ivan Dokmanić. An approximation theory for metric space-valued functions with a view towards deep learning, 2023. arXiv:2304.12231, doi:10.48550/arXiv.2304.12231.

[KLL+23c] Anastasis Kratsios, Chong Liu, Matti Lassas, Maarten V de Hoop, and Ivan Dokmanić. An approximation theory for metric space-valued functions with a view towards deep learning. *arXiv preprint arXiv:2304.12231*, 2023.

[KLM21] Nikola B. Kovachki, Samuel Lanthaler, and Siddhartha Mishra. On universal approximation and error bounds for fourier neural operators. *Journal of Machine Learning Research*, 22(290):1–76, 2021.

[KNP26] Anastasis Kratsios, A Martina Neuman, and Philipp Petersen. Adaptivity under realizability constraints: Comparing in-context and agentic learning. *arXiv preprint arXiv:2605.04995*, 2026.

[KZLD22] Anastasis Kratsios, Behnoosh Zamanlooy, Tianlin Liu, and Ivan Dokmanić. Universal approximation under constraints is possible with transformers. In *International Conference on Learning Representations*, 2022. URL: https://openreview.net/forum?id=JGO8CvG5S9.

[Lan67] Henry J. Landau. Necessary density conditions for sampling and interpolation of certain entire functions. *Acta Mathematica*, 117:37–52, 1967. doi:10.1007/BF02395039.

[Lan23] Samuel Lanthaler. Operator learning with pca-net: upper and lower complexity bounds. *Journal of Machine Learning Research*, 24(318):1–67, 2023.

[LBH+26] Yicen Li, Jose Antonio Lara Benitez, Ruiyang Hong, Anastasis Kratsios, Paul David McNicholas, and Maarten Valentijn de Hoop. Neural operators can discover functional clusters. *arXiv preprint arXiv:2602.23528*, 2026.

[LHLA23] Zongyi Li, Daniel Zhengyu Huang, Burigede Liu, and Anima Anandkumar. Fourier neural operator with learned deformations for pdes on general geometries. *Journal of Machine Learning Research*, 24(388):1–26, 2023.

[LJP+21] Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3):218–229, 2021. doi:10.1038/s42256-021-00302-5.

[LKA+21] Zongyi Li, Nikola Borislavov Kovachki, Kamyar Azizzadenesheli, Burigede liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021. URL: https://openreview.net/forum?id=c8P9NQVtmnO.

[LLS25] Samuel Lanthaler, Zongyi Li, and Andrew M Stuart. Nonlocality and nonlinearity implies universality in operator learning. *Constructive Approximation*, 62(2):261–303, 2025.

[LM07] Pierre-Louis Lions and Marek Musiela. Correlations and bounds for stochastic volatility models. *Annales de l'Institut Henri Poincaré C, Analyse Non Linéaire*, 24(1):1–16, 2007. doi:10.1016/j.anihpc.2005.05.007.

[LR25] Michael Ludkovski and Jimmy Risk. *Gaussian Process Models for Quantitative Finance*. SpringerBriefs in Quantitative Finance. Springer, Cham, 2025. doi:10.1007/978-3-031-80874-6.

[LS23] Samuel Lanthaler and Andrew M Stuart. The curse of dimensionality in operator learning. *arXiv preprint arXiv:2306.15924*, 2023.

[LSYZ21] Jianfeng Lu, Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation for smooth functions. *SIAM Journal on Mathematical Analysis*, 53(5):5465–5506, 2021. doi:10.1137/20M134695X.

[Lud18] Mike Ludkovski. Kriging metamodels and experimental design for bermudan option pricing. *Journal of Computational Finance*, 22(1):37–77, 2018. doi:10.21314/JCF.2018.347.

[MVD+26] Tingwei Meng, Moritz Voß, Nils Detering, Giulio Farolfi, Stanley Osher, and Georg Menz. Solving optimal execution problems via in-context operator networks, 2026. arXiv:2501.15106, doi:10.48550/arXiv.2501.15106.

[NJZB23] Brian Ning, Sebastian Jaimungal, Xiaorong Zhang, and Maxime Bergeron. Arbitrage-free implied volatility surface generation with variational autoencoders. *SIAM Journal on Financial Mathematics*, 14(4):1004–1027, 2023.

[PV20] Philipp Petersen and Felix Voigtlaender. Equivalence of approximation by convolutional neural networks and fully-connected networks. *Proceedings of the American Mathematical Society*, 148(4):1567–1581, 2020. doi:10.1090/proc/14789.

[PZ24] Philipp Petersen and Jakob Zech. Mathematical theory of deep learning. *arXiv preprint arXiv:2407.18384*, 2024.

[SS16] Saburou Saitoh and Yoshihiro Sawano. *Theory of Reproducing Kernels and Applications*, volume 44 of *Developments in Mathematics*. Springer, Singapore, 2016. doi:10.1007/978-981-10-0530-5.

[SSZ26] Christoph Schwab, Andreas Stein, and Jakob Zech. Deep operator network approximation rates for lipschitz operators. *Analysis and Applications*, 24(01):199–239, 2026.

[ST26] Philipp Schmocker and Josef Teichmann. Weighted universal approximation of differentiable maps on infinite-dimensional manifolds. *arXiv preprint arXiv:2606.09820*, 2026.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.

[Wen05] Holger Wendland. *Scattered Data Approximation*, volume 17 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, Cambridge, 2005. doi:10.1017/CBO9780511617539.

[WJG25a] Ruben Wiedemann, Antoine Jacquier, and Lukas Gonon. Operator deep smoothing for implied volatility. In *The Thirteenth International Conference on Learning Representations*, 2025.

[WJG25b] Ruben Wiedemann, Antoine Jacquier, and Lukas Gonon. Operator deep smoothing for implied volatility. In *The Thirteenth International Conference on Learning Representations*, 2025. URL: https://openreview.net/forum?id=DPlUWG4WMw.

[Xia09] Qinglan Xia. The geodesic problem in quasimetric spaces. *Journal of Geometric Analysis*, 19(2):452–479, 2009.