AMORTIZED SHAPLEY INTERACTIONS FOR NEURAL OPERATORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Explaining model predictions can increase trust in AI systems and provide valuable insights to developers. Feature-based attribution methods, such as Shapley values, are a popular approach to model interpretability. In the physical sciences, however, model inputs often represent continuous domains rather than discrete features. Additionally, their mesh-based discretizations may vary between samples, making fast amortized explanations difficult. Infinite-player Aumann–Shapley values have recently been proposed as an alternative for such continuous inputs, but the existing framework is limited to additive attributions and therefore cannot capture interactions between input regions. In this work, we propose *OperatorSHAP-IQ*, which extends the Aumann–Shapley framework to feature interactions. We further introduce an amortization scheme that enables efficient inference of interaction maps and show how different classes of partial differential equations exhibit different levels of interactions.

1 INTRODUCTION

Partial differential equations (PDEs) are widely used to model physical processes. Solving them numerically allows for accurate predictions, but is computationally heavy and tedious to set up. *Neural operators* are an AI-based alternative for predicting PDE solutions and have shown success in various areas, like weather forecasting (Kurth et al., 2023), plasma physics (Gopakumar et al., 2024), or medicine (Jatyani et al., 2025). Efficiently explaining their output is difficult, however. As samples in physical applications are often discretized on heterogeneous grids (e.g., aerodynamics around different car shapes), amortized explainers must be able to produce explanations for grid-nodes that may not have been present in any of the training samples.

A common approach to explainablity is feature attribution, with *Shapley values* as the arguably most popular representative (Shapley, 1988; Lundberg & Lee, 2017). As an alternative to classical finite-player Shapley values, the infinite player *Aumann-Shapley values* have recently been introduced to explain Neural Operators (Stiller et al., 2026). There, instead of computing an attribution value per input feature, an attribution **density** over the input **region** is fitted. In this work, we extend this idea to allow for interactions between regions, which is not covered by classical Aumann-Shapley theory. Interactions allow us to see if two regions have a synergetic or redundant effect, which previously would have been split up between the *first-order attributions* of the regions (Tsai et al., 2023; Muschalik et al., 2024). Unlike linear PDEs, nonlinear PDEs do produce such joint effects, for example when two waves are merging (cf. Figure 1).

We provide the following contributions:

(i) We extend the Aumann-Shapley theory to interactions between input regions. To this end, we show that the Faithful Shapley Interaction Index (FSII) is a mixture of discrete derivatives, a property that Tsai et al. (2023) only showed for the highest interaction order.

(ii) Extending OperatorSHAP, we develop an algorithm that yields amortized interaction attributions for Neural Operators.

(iii) In an extensive benchmark, we show that our method is on par with existing state-of-the-art Shapley interaction methods and can predict instantly for any sample and grid, whereas classical methods need costly computation for each new sample.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

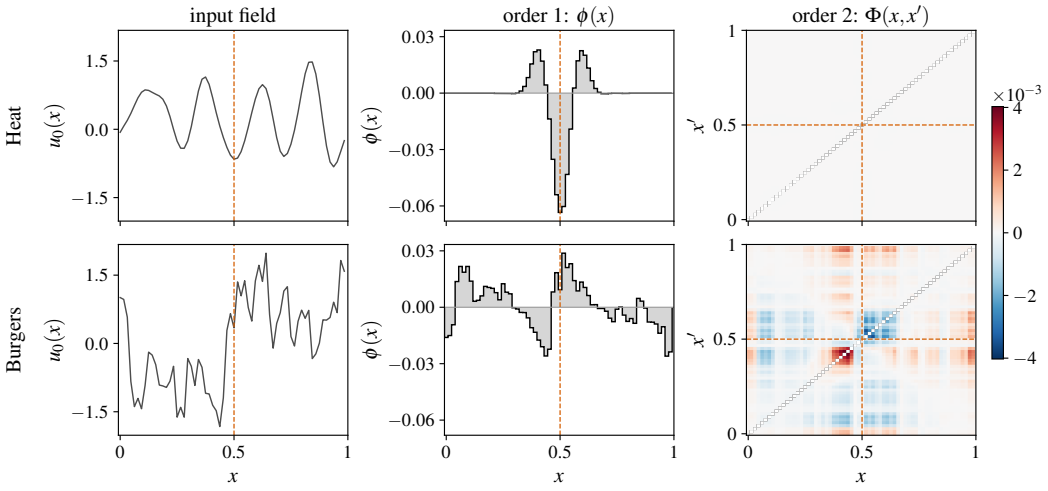


Figure 1: Attributions for models of a linear (top row: heat equation) and a nonlinear (bottom row: Burgers equation) PDE at $x^\star = 0.5$ (dashed line). Left: the input state. Center: the first-order attributions $\phi(x)$. Right: the second-order attributions $\Phi(x, x')$ between the regions around x and x'. The linear PDE shows no interaction, as expected from its additive structure, while the nonlinear PDE does. Second-order carries 1.6% of the attribution mass $\sum_i |\phi_i| + \sum_{i<j} |\Phi_{ij}|$ for the heat operator and 48.9% for Burgers. Section C gives the setup.

2 Related Work

Shapley values and interactions. The Shapley value (Shapley, 1988) has gained increasing attention in machine learning (Lundberg & Lee, 2017) as a measure for feature attribution. Grounded in cooperative game theory, the Shapley values is axiomatically justified, however, its computation requires an exponential number of model evaluations. Therefore, approximation methods are needed even for relatively small input dimensions (Kolpaczki & Hüllermeier, 2026). Beyond single-feature effects, the Shapley *interaction index* (Grabisch & Roubens, 1999; Fujimoto et al., 2006), the Shapley–Taylor index (Sundararajan et al., 2020), k-Shapley Interaction Index (Bordt & von Luxburg, 2023), and the Faithful Shapley Interaction Index (FSII) (Tsai et al., 2023) have been proposed, each fulfilling different axiomatic properties. We base our theory on FSII, as it fulfills practically relevant axioms such as efficiency and faithfulness. As with Shapley values, in many practical domains estimation is necessary, either via Monte-Carlo sampling (Fumagalli et al., 2023), least squares approximation (Fumagalli et al., 2024), or surrogate models (Thies et al., 2026).

Amortized Shapley values and interactions. FastSHAP (Jethani et al., 2022) first amortized Shapley value estimation by training one explainer network that returns the Shapley values of any input in a single forward pass. InstaSHAP (Enouen & Liu, 2025) carries this idea over to Shapley interactions. Stiller et al. (2026) introduce OperatorSHAP for learning the Aumann–Shapley density to amortize Shapley values for heterogeneous grid inputs. In this paper, we extend this grid-agnostic amortization to interactions.

Neural operators. Neural operators (Kovachki et al., 2023) learn the solution operator of a PDE from data. Fourier neural operators (FNOs) (Li et al., 2021) are a simple and efficient version that parameterize the operator by a multiplication in Fourier space, GINO (Li et al., 2023) combines graph layers with an FNO to handle irregular meshes, and further variants replace the Fourier transform by wavelets (Gupta et al., 2021) or restrict it to local patches (Liu-Schiaffini et al., 2024). More powerful transformer-based Neural Operators have since been published (Hao et al., 2023; Wen et al., 2025; Herde et al., 2024), but the simpler models suffice for our demonstrations.

3 BACKGROUND

We start by introducing finite player Shapley values and interactions. Then, after briefly introducing our PDE setting, we will extend them to infinite player Aumann-Shapley values and their amortization with OperatorSHAP.

3.1 ATTRIBUTIONS AND INTERACTIONS FOR FINITELY MANY PLAYERS

A *game* on the player set $[n] := \{1, \dots, n\}$ is a map $v : 2^{[n]} \to \mathbb{R}$, and a coalition $S \subseteq [n]$ is a subset of players. In machine learning, usually players are input features to a model and $v(S)$ is the model output when the features in S are revealed and the others are removed (through masking or replacement by a default value). Throughout, $\mathbf{1} := (1, \dots, 1)$, $\mathbb{1}[\cdot]$ is the indicator of a statement, and $|x|_1 := \sum_i |x_i|$, $|x|_\infty := \max_i |x_i|$ for $x \in \mathbb{R}^m$.

An *attribution* is a real number that a method assigns to a player or to a set of players. A *value* maps a game to one attribution per player, and an ℓ*-interaction index* maps it to one attribution per set of at most ℓ players. The Shapley value (Shapley, 1988) $\phi(v) = (\phi_i(v))_{i\in[n]}$ is the *unique* value satisfying *efficiency* ($\sum_{i\in[n]} \phi_i(v) = v([n]) - v(\emptyset)$), *symmetry* (players i, j with $v(S \cup \{i\}) = v(S \cup \{j\})$ for all $S \subseteq [n] \setminus \{i, j\}$ receive the same value), *null-player property* (a player that never changes the game value has value 0), and *linearity* ($\phi(av + bw) = a\phi(v) + b\phi(w)$).

In the interaction case, *multiple* interaction indices satisfy the interaction version of these axioms (Grabisch & Roubens, 1999; Fujimoto et al., 2006). Tsai et al. (2023) additionally introduce *faithfulness* as a desirable property, meaning that the attributions c_T of an interaction index approximate the game by

$$v(S) \approx \sum_{T\subseteq S,\, |T|\le \ell} c_T \quad \forall S \subseteq [n].$$

This is only achievable approximately, which is why they consider the following least-squares problem to define the Faithful Shapley Interaction Index (FSII).

Definition 1 (FSII, (Tsai et al., 2023))**.** Let v be a game on $[n]$ and $\ell \in [n]$ the maximum interaction order we consider. The *FSII of maximum order* ℓ, $\mathcal{E}(v, \ell) = (\mathcal{E}_T)_{|T|\le\ell}$, is the minimizer of

$$\sum_{S\subseteq[n],\, 0<|S|<n} \mu_n(S)\Big(v(S) - \sum_{T\subseteq S,\, |T|\le\ell} \mathcal{E}_T\Big)^2, \qquad \mu_n(S) = \frac{n-1}{\binom{n}{|S|}\,|S|\,(n-|S|)}, \tag{1}$$

subject to $\mathcal{E}_\emptyset = v(\emptyset)$ and $\sum_{|T|\le\ell} \mathcal{E}_T = v([n])$. We write $\mathcal{E}_T(v, \ell)$ for the component of $\mathcal{E}(v, \ell)$ at T. The minimizer exists, is unique, and has a closed form (Tsai et al., 2023, Thm. 19).

The FSII maximizes faithfulness, while preserving the above axioms. It will be the finite-player basis for OperatorSHAP-IQ. We will show that, with an increasing number of players, we will converge to a limit object that gives us attributions on a continuous domain.

For this, we will rewrite FSII as a w*-mixture index*, following Fujimoto et al. (2006), which frames FSII via *discrete derivatives* over random coalitions. The discrete derivative $(\Delta_T v)(S)$ below is the joint effect of adding the players of T to a coalition S. The mixture index averages the discrete derivative over random coalitions in which every other player is present with probability t, and weights the levels t by w.

For $T \subseteq [n]$ and $S \subseteq [n] \setminus T$, the *discrete derivative* is

$$(\Delta_T v)(S) := \sum_{L\subseteq T} (-1)^{|T|-|L|}\, v(S \cup L). \tag{2}$$

Definition 2 (Signed mixture index, after Fujimoto et al. (2006))**.** Let w be a finite signed Borel measure on $[0, 1]$, let v be a game on $[n]$ and $T \subseteq [n]$ with $|T| = k \ge 1$. The w*-mixture index of order* k is

$$I_k^w(v; T) := \int_0^1 \mathbb{E}\big[(\Delta_T v)(S_t)\big]\, dw(t), \qquad S_t \sim \bigotimes_{c\in[n]\setminus T} \text{Bernoulli}(t), \tag{3}$$

where S_t contains each player of $[n] \setminus T$ independently with probability t.

For $k = 1$ and w the uniform measure, $I_1^w(v; \{i\})$ is the Shapley value of player i.

3.2 PDEs and the Operator Setting

Partial differential equations (PDEs) describe physical states, usually dependent on time and space. On a spatial domain $\Omega \subseteq \mathbb{R}^d$, a time-dependent state $u : [0,T] \times \Omega \to \mathbb{R}^{c_{\text{out}}}$ can be described abstractly by $\partial_t u = \mathcal{L}_a(u), u(0) = u_0$, together with suitable boundary conditions. The differential operator $\mathcal{L}_a$ acts on the spatial variables of u and may contain derivatives, coefficient functions a, and nonlinear terms. Time-independent PDEs, like the Darcy equation, define the state directly as $u : \Omega \to \mathbb{R}^{c_{\text{out}}}$. In both cases, the PDE induces a solution operator $\mathcal{F}$ that maps input functions, such as u_0, a, or f, to the solution u. A neural operator $\hat{\mathcal{F}}$ learns an approximation of $\mathcal{F}$ from data.

3.3 Continuum player space and operator games

As we have seen in the previous sections, our data lies in a **continuous** domain $\Omega \subseteq \mathbb{R}^d$, but Shapley values are only defined for **finitely** many players. To overcome this, we leverage Aumann-Shapley theory (Aumann & Shapley, 2015), which extends Shapley values to a continuum of players.

A *player space* $(\Omega, \mathcal{C})$ is a measurable space isomorphic to $([0,1], \mathcal{B}([0,1]))$, and we denote by NA^1 the set of its non-atomic probability measures (every set of positive mass has a measurable subset of strictly smaller positive mass). A *smooth game of class* C^k is a set function $v : \mathcal{C} \to \mathbb{R}$ of the form

$$v(S) = f\big(\mu_1(S), \dots, \mu_m(S)\big), \qquad S \in \mathcal{C}, \tag{4}$$

for some $m \in \mathbb{N}$, $\mu = (\mu_1, \dots, \mu_m) \in (\mathrm{NA}^1)^m$ and $f \in C^k([0,1]^m)$ with $f(0) = 0$. The pair (f, μ) is a *representation*. $D^k f(x)$ denotes the symmetric tensor of k-th partial derivatives, acting on vectors by $D^k f(x)[q_1, \dots, q_k] = \sum_{i_1,\dots,i_k} \partial^k_{i_1 \dots i_k} f(x)\, q_{1,i_1} \dots q_{k,i_k}$, and $\|D^k f(x)\|_\infty$ is its largest entry in absolute value. A *partition* $\pi = \{B_1, \dots, B_n\} \subset \mathcal{C}$ of Ω has *cell masses* $p_c := (\mu_1(B_c), \dots, \mu_m(B_c)) \in [0,1]^m$, so that $\sum_c p_c = \mathbf{1}$. The *restriction* of v to π is the finite game $v_\pi(A) := v\big(\bigcup_{c \in A} B_c\big) = f\big(\sum_{c \in A} p_c\big)$, $A \subseteq [n]$. The *mesh* of π relative to μ is $\mathrm{mesh}_\mu(\pi) := \max_{i,c} \mu_i(B_c) = \max_c |p_c|_\infty$.

Concretely, in our case we have a physical domain Ω and we define the coalition space $\mathcal{C}$ as a suitable set of regions S, with each $S \subset \Omega$. A game that maps a region S to a real number can then be defined, e.g., by evaluating a model's prediction at a specific output location $x \in \Omega$ when the input is activated in the region S (and set to a baseline at the rest of the input space $\Omega \setminus S$). The definition via f and μ helps us manage the infinite players by routing their attribution through finite μ. We are looking for an interaction index that assigns an order-1 attribution to every region S and an order-2 attribution to every tuple of disjoint regions (S, T).

OperatorSHAP (Stiller et al., 2026) defines a game that fits the above description. It divides the input region Ω with smooth bump function $\eta_i(x)$ that regulate how much of a subregion is activated by the coalition S (see an example in Fig. 2). OperatorSHAP's theory requires the games to be smooth. For this to hold in a function space setting, some technical requirements are necessary. Stiller et al. (2026) define the domain as a flat torus $\Omega = \mathbb{T}^d$ (though this can be relaxed) and restrict to input functions from Sobolev spaces $H^s(\Omega), s > d/2$ (which are certain function spaces used in PDE theory), and $\hat{\mathcal{F}}$ to be differentiable (in particular Fréchet differentiable). We will use the same setup, but additionally require $\hat{\mathcal{F}}$ to be ℓ-times continuously differentiable and omit these specifications to keep the definitions and explanations clean. We refer to Section A.4 for a short introduction and justification of this setup.

Definition 3 (Operator game, adapted from Stiller et al. (2026))**.** Let $\eta_1, \dots, \eta_m \in C^\infty(\Omega)$ be smooth masks on the domain that satisfy $\eta_i \geq 0$, $\int_\Omega \eta_i \, d\lambda > 0$ and $\sum_i \eta_i \equiv 1$. Let $\mu_i(S) := \int_S \eta_i \, d\lambda / \int_\Omega \eta_i \, d\lambda$ be the normalized integral of η_i over S. Let $\bar{u}, u_0$ be the baseline and the model input, $h := (u_0 - \bar{u})$, and let $\hat{\mathcal{F}}_x : H^s(\Omega) \to \mathbb{R}$, the model read out at x, be Fréchet-C^ℓ. The corresponding *operator game* is

$$v(S) := \hat{\mathcal{F}}_x\left(\bar{u} + \sum_{i=1}^{m} \mu_i(S) \eta_i h\right) - \hat{\mathcal{F}}_x(\bar{u}).$$

The game then tells us how much the prediction at x differs, if we change the input in a region S from baseline to u_0; see Fig. 2 for a visualization. The operator game is a smooth game of class

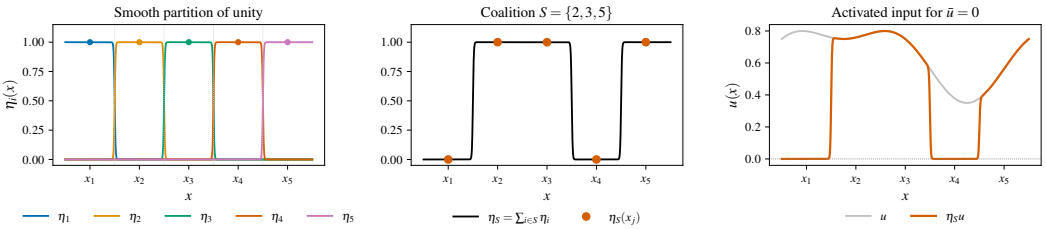


Figure 2: Construction of an operator game. Left: Smooth masks forming a partition of unity. Center: The mask selected by a coalition S. Right: An example input u and the activated input $\bar{u} + \eta_S(u - \bar{u})$ for the zero baseline $\bar{u} = 0$. Adapted from Stiller et al. (2026).

C^ℓ for most Neural Operators with smooth activations (Lemma 23). The baseline $\bar{u}$ decides what a masked region is compared against. We use zero, or the node-wise training mean where zero is not an admissible state. Section B.11 discusses the choice.

4 OperatorSHAP-IQ

This section introduces the theory behind OperatorSHAP-IQ. We will start from an FSII on a partition π and show that the FSII converges as the partition refines. For this, we will first show that the FSII can be represented as a w-mixture of discrete derivatives. As the partition refines, this mixture of discrete derivatives converges to an integral over the actual derivative of our game.

4.1 FSII heads are signed mixture indices

We now construct the measures $w_k^{(\ell)}$ that define the FSII as a $w_k^{(\ell)}$-mixture of discrete derivatives.

Definition 4 (Per-order measures)**.** For integers $1 \le k \le \ell$, the *per-order measure* $w_k^{(\ell)}$ is the finite signed measure on $[0, 1]$ with the density

$$w_k^{(\ell)}(t) := \sum_{s=k}^{k+\ell-1} b_{k,\ell,s}\, t^{s-1}, \qquad b_{k,\ell,s} = (-1)^{s-k} \frac{k\,(\ell+k-1)!\,(s+\ell-k)!}{k!\,(\ell-k)!\,s!\,(s-k)!\,(k+\ell-1-s)!}. \quad (5)$$

In particular, $w_1^{(1)} = 1$, $w_1^{(2)}(t) = 4 - 6t$ and $w_2^{(2)}(t) = 6t(1-t)$. We write $w_k^{(\ell)}$ for both the measure and its density.

Theorem 5 (Exact mixture representation)**.** *Let $n \ge 2$, $\ell \in [n]$, v a game on $[n]$, and $T \subseteq [n]$ with $1 \le |T| = k \le \ell$. Then, the FSII is a $w_k^{(\ell)}$-mixture of discrete derivatives*

$$\mathcal{E}_T(v, \ell) = I_k^{w_k^{(\ell)}}(v; T). \quad (6)$$

Proof sketch (full proof in Section A.1). Both sides are linear in v, so it suffices to compare them on the unanimity games $v_R(S) = \mathbb{1}[S \supseteq R]$, which form a basis of the space of games. There, the mixture index is a moment of w and the closed form of Tsai et al. (2023, Thm. 19) is a rational function of the coalition size, and the densities (Eq. (5)) are the partial fractions of that rational function. □

Remark 6. At $k = \ell$ (the highest interaction order), the result was shown by Tsai et al. (2023, Props. 21–22). The case $k < \ell$ is, to the best of our knowledge, new. For $k < \ell$ and $n > \ell$, no probability measure w satisfies $\mathcal{E}_T(v, \ell) = I_k^w(v; T)$ for all games v. By Lemma 10, such a w has the same moments $m_r(w) = \int_0^1 t^r\, dw(t)$ as $w_k^{(\ell)}$ for all $r \le n-k$. By Lemma 11, these are $m_0 = 1$, $m_1 = 0$ and $m_{\ell-k+1} = R_{k,\ell}(\ell - k + 1) \ne 0$. The only probability measure with $m_1(w) = 0$ is δ_0, and $m_{\ell-k+1}(\delta_0) = 0$.

4.2 CONTINUUM LIMIT

The following theorem builds the centerpiece of this paper. It splits into three parts: (i) shows that the difference between FSII and the integral over the derivative is upper-bounded by a function of the mesh size (but could go to 0 as the mesh refines); (ii) tells us that the sum over all cell-attributions in a region converge to the integral over the derivative of said region (i.e., even though the single interactions converge to 0, their sum may converge to something bigger than zero); (iii) tells us when we can approximate the continuum interaction with a density (that can then be learned).

Theorem 7 (FSII fields). *Let $\ell \geq 1$ and let $v = f \circ \mu$ be a smooth game of class C^ℓ with $\mu \in (\mathrm{NA}^1)^m$. For $k \leq \ell$ let $L_k := \sup_x \|D^k f(x)\|_\infty$, let $\omega_k(r) := \sup\{\|D^k f(x) - D^k f(y)\|_\infty : |x-y|_\infty \leq r\}$ be the modulus of continuity of $D^k f$, let $\|w\|_{\mathrm{TV}}$ be the total variation of a signed measure w[1], and*

$$\rho_k(\delta) := \omega_k\big(2k\delta + \delta^{1/3}\big) + \frac{mL_k}{2}\,\delta^{1/3} \xrightarrow[\delta\to 0]{} 0, \qquad \rho_k(\delta) = O(\delta^{1/3}) \text{ if } D^k f \text{ is Lipschitz.} \tag{7}$$

Then:

(i) For every partition π with at least $\max\{2,\ell\}$ cells and $\mathrm{mesh}_\mu(\pi) = \delta$, and all distinct cells $a_1, \ldots, a_k$,

$$\Big|\mathcal{E}_{\{a_1,\ldots,a_k\}}(v_\pi, \ell) - \int_0^1 D^k f(t\mathbf{1})[p_{a_1}, \ldots, p_{a_k}]\, w_k^{(\ell)}(t)\, dt\Big| \leq \tag{8}$$
$$\leq \|w_k^{(\ell)}\|_{\mathrm{TV}} \Big(\prod_{l=1}^{k} |p_{a_l}|_1\Big)\rho_k(\delta).$$

(ii) Let $S_1, \ldots, S_k \in \mathcal{C}$ be pairwise disjoint and let $(\pi_r)_{r\geq 0}$ be partitions, each refining the previous, in which every S_l is a union of cells and $\mathrm{mesh}_\mu(\pi_r) \to 0$. Then,

$$\sum_{B_{a_1}\subseteq S_1} \cdots \sum_{B_{a_k}\subseteq S_k} \mathcal{E}_{\{a_1,\ldots,a_k\}}(v_{\pi_r}, \ell) \longrightarrow \mathcal{I}_k^{(\ell)}(S_1, \ldots, S_k),$$
$$\mathcal{I}_k^{(\ell)}(S_1, \ldots, S_k) := \int_0^1 D^k f(t\mathbf{1})\big[\mu(S_1), \ldots, \mu(S_k)\big]\, w_k^{(\ell)}(t)\, dt, \tag{9}$$

and the limit depends neither on the sequence nor on the C^ℓ representation (f, μ) of v.

(iii) Let Ω carry a metric whose Borel σ-algebra is $\mathcal{C}$, let λ be a finite Borel measure on Ω, and let $\mu_i = \tilde{\eta}_i \lambda$ for some bounded continuous densities $\tilde{\eta}_i$. Then Eq. (9) has the continuous density

$$\iota_k^{(\ell)}(x_1, \ldots, x_k) = \int_0^1 D^k f(t\mathbf{1})\big[\tilde{\eta}(x_1), \ldots, \tilde{\eta}(x_k)\big]\, w_k^{(\ell)}(t)\, dt, \qquad \tilde{\eta} := (\tilde{\eta}_1, \ldots, \tilde{\eta}_m), \tag{10}$$

with respect to $\lambda^{\otimes k}$ on $S_1 \times \cdots \times S_k$, and along partitions with $\mathrm{mesh}_\mu(\pi) \to 0$ and cells $B_{a_l} \ni x_l$ of positive λ-mass and vanishing diameter, for distinct $x_1, \ldots, x_k$,

$$\frac{\mathcal{E}_{\{a_1,\ldots,a_k\}}(v_\pi, \ell)}{\prod_{l=1}^{k} \lambda(B_{a_l})} \longrightarrow \iota_k^{(\ell)}(x_1, \ldots, x_k). \tag{11}$$

Proof sketch (full proof in Section A.2). By Theorem 5 every head is a mixture index (Eq. (3)). As the mesh refines, the cell masses of a random coalition concentrate at $t\mathbf{1}$ by Chebyshev's inequality, and the discrete derivative is an integral of $D^k f$ close to that point, which gives (i). Summing (i) over the cells of the regions gives (ii) by multilinearity, and dividing by the cell volumes gives (iii). □

At $k = \ell = 1$, (ii) is the asymptotic-value theorem for games of the form $f \circ \mu$ (Aumann & Shapley, 2015; Neyman, 2002) and Eq. (9) is the Aumann–Shapley diagonal formula.

[1] $\|w\|_{\mathrm{TV}} := |w|([0,1])$ with $|w| := w^+ + w^-$ for the Jordan decomposition $w = w^+ - w^-$ into non-negative measures.

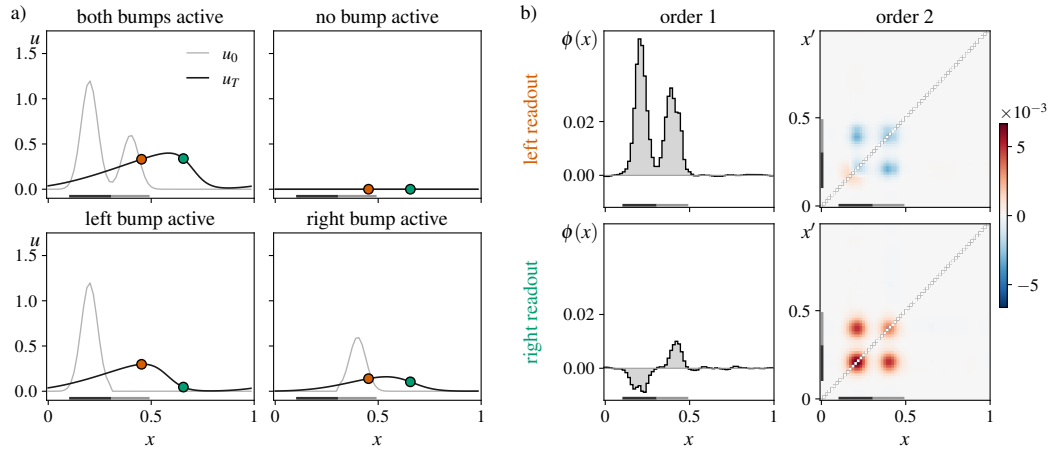


Figure 3: Interactions in the Burgers equation for an input with two bumps travelling to the right, read out at $x^\star = 0.45$ (orange) or $x^\star = 0.65$ (green). Bars mark the bumps. a) Input and prediction with none, one or both bumps active. b) Explanation at each readout. At the orange readout both bumps together give less than the sum of their single values (negative interaction). At the green readout the left bump amplifies the right one (positive interaction). Section C gives the setup.

For the operator game of Definition 3, the assumptions of Theorem 7 hold with the densities $\tilde{\eta}_i = \eta_i / \int_\Omega \eta_i \, d\lambda$ of the μ_i (Lemma 23), and Proposition 24 in the appendix writes its fields in terms of the derivatives of $\hat{\mathcal{F}}_x$.

Remark 8 (Exact null for affine operators). If $\hat{\mathcal{F}}_x$ is affine, then f is affine and $D^k f \equiv 0$ for $k \geq 2$, so $\mathcal{E}_T(v_\pi, \ell) = 0$, showing why an accurate model of a linear PDE has very little interaction.

4.3 Algorithm

We will now show how we can learn the densities from Theorem 7 (iii) for a maximum interaction level of $\ell = 2$. For this, we have to learn the 1-interaction density $\varphi(x) := \iota_1^{(2)}(x)$ and the 2-interaction density field $\Phi(x, y) := \iota_2^{(2)}(x, y)$. With these, we will get instantaneous explanations for the game v from Definition 3 for any input u_0 and query point $x^\star$. We train one explainer Neural Operator $E_\theta(u_0, x^\star)$ that returns both densities (Algorithm 1), where u_0 is represented on a mesh and $x^\star$ is represented as a Gaussian bump on the same mesh as an additional channel. $E_\theta(u_0, x^\star)$ returns features at every node, from which one linear head gives $\hat{\varphi}_\theta(x)$ and a second gives $g(x) \in \mathbb{R}^R$, $R \in \mathbb{N}$, with $\hat{\Phi}_\theta(x, y) := \sum_{r=1}^R \beta_r g_r(x) g_r(y)$ and trainable $\beta_1, \ldots, \beta_R \in \mathbb{R}$.

We assign each node of the mesh a cell $B_i \subset \Omega$ (e.g., with a Voronoi partition) and fix $\eta_1, \ldots, \eta_m$, which appear as hard masks on the grid nodes (see Section B.10 for an extended reasoning). We then train E_θ with the FastSHAP-like loss

$$\mathcal{L}(\theta) = \mathbb{E}_{u_0, x^\star, A} \left[\left(v_\pi(A) - \hat{v}_\theta(A) \right)^2 \right], \tag{12}$$

where u_0 is drawn uniformly from the training set and $x^\star$ uniformly from Ω. To draw A, we draw its size k with probability proportional to $1/(k(n-k))$ and then A uniformly among the sets of size k, which makes its probability proportional to $\mu_n(A)$. We add the complement $[n] \setminus A$ as a second sample. $v_\pi(A)$ denotes the operator game for coalition A on the partition defined by u_0 ($x^\star$, $\bar{u}$, and u_0 are implicit in v) and the explainer estimates it by $\hat{v}_\theta(A) := \sum_{a \in A} \hat{E}_a + \sum_{a < b \in A} \hat{E}_{ab}$, where the cell coefficients $\hat{E}_T$ estimate the FSII $\mathcal{E}_T(v_\pi, 2)$ and are

$$\hat{E}_a = \int_{B_a} \hat{\varphi}_\theta \, d\lambda + \tfrac{1}{2} \int_{B_a \times B_a} \hat{\Phi}_\theta \, d\lambda^{\otimes 2}, \qquad \hat{E}_{ab} = \int_{B_a \times B_b} \hat{\Phi}_\theta \, d\lambda^{\otimes 2}. \tag{13}$$

We enforce efficiency by projecting $\hat{E} = (\hat{E}_T)_{1 \leq |T| \leq 2}$ onto $\sum_T \hat{E}_T = v_\pi([n]) - v_\pi(\emptyset)$ along $z = B^{-1}\mathbf{1} / (\mathbf{1}^\top B^{-1} \mathbf{1})$ with B the Hessian of the μ_n-weighted objective. Section A.10 reduces z

Table 1: Faithfulness R^2 and pair agreement (Pearson r of the pair coefficients with the labels) of ProxySHAP at three budgets, the limit fields (Limit), OperatorSHAP-IQ (Ours) and the reference (Ref.), mean over three query points, all grids of Tables 4 and 5 and 30 test samples, with R^2 floored at 0. The pair coefficients of the linear PDEs are near zero (Section B.7), so their agreement with the reference pairs is marked n/a. The meshes use 32×32 box players, where the reference is ProxySHAP at 50,000 coalitions without labels, and have no limit fields (Section B.6). Bold marks the best method.

	Faithfulness R^2						Order-2 Pearson r to the reference				
	ProxySHAP at budget			Limit	Ours	Ref.	ProxySHAP at budget			Limit	Ours
Dataset	1k	4k	16k				1k	4k	16k		
Heat 1D	0.990	0.996	0.997	0.998	**0.999**	0.997	n/a	n/a	n/a	n/a	n/a
Advection 1D	0.959	**0.962**	**0.962**	0.150	0.956	0.964	n/a	n/a	n/a	n/a	n/a
Burgers 1D	0.883	0.902	0.900	0.914	**0.931**	0.901	0.504	0.623	0.688	0.684	**0.700**
Allen–Cahn 1D	0.967	0.979	0.982	0.998	**0.999**	0.984	0.353	0.447	0.515	**0.633**	0.592
Diff.–sorption 1D	0.987	0.991	**0.992**	0.985	0.980	0.992	0.587	0.708	**0.774**	0.751	0.499
Heat 2D	0.938	0.961	0.968	0.995	**0.997**	0.969	n/a	n/a	n/a	n/a	n/a
Shallow water 2D	0.938	0.961	0.968	**1.000**	0.997	0.969	0.092	0.168	0.262	**0.386**	0.233
Gray–Scott 2D	0.227	0.282	0.296	0.005	**0.472**	0.311	0.308	0.536	**0.713**	0.169	0.574
Darcy 2D	0.961	0.975	0.980	**0.988**	0.962	0.981	0.429	0.592	**0.691**	**0.691**	0.308
Heat 3D	0.878	0.922	0.934	**0.992**	0.989	0.935	n/a	n/a	n/a	n/a	n/a
Airfoil mesh	0.982	0.987	**0.989**	–	0.842	0.990	n/a	n/a	n/a	n/a	n/a
Cylinder mesh	0.970	0.975	**0.976**	–	0.914	0.977	n/a	n/a	n/a	n/a	n/a

to a 2×2 solve and gives it in closed form. We apply the projection to the cell coefficients of one mesh, in every training step and at inference (Algorithm 1).

5 Experiments

We explain FNOs on ten grid datasets in 1D, 2D, and 3D, and GINOs on the irregular airfoil and cylinder meshes of MeshGraphNets (Pfaff et al., 2021) (Section B.6), under the point readout $\hat{\mathcal{F}}_{x^\star}(u) = \hat{\mathcal{F}}(u)(x^\star)$ with the zero baseline (the node-wise training mean for shallow water), and one player per node, at three query points and on the grids the explainer was trained on (16 to 256 nodes in 1D, 8^2 and 16^2 in 2D, 8^3 in 3D). The baseline is ProxySHAP (Thies et al., 2026) from shapiq (Muschalik et al., 2024) at budgets $b = 1{,}024$, 4,096 and 16,384. We ablated other choices of shapiq's FSII estimators, but none yielded competitive results (Section B.8). The reference labels are ProxySHAP at $b = 50{,}000$, and on the 16-node grids the exact order-2 FSII over all 2^{16} coalitions. We also evaluate the fields of Theorem 7 directly, by quadrature of the backbone derivatives along the path from the baseline (Limit, Section B.2). Per test sample, we report the faithfulness R^2 of the degree-2 model on 4,096 independently sampled coalitions from the FSII measure, which needs no reference, and the Pearson correlation of the pair coefficients with the reference. Section B gives the protocol and every detailed results.

Algorithm 1: OperatorSHAP-IQ training

Input: Dataset $\mathcal{X}$, backbone $\hat{\mathcal{F}}$, baseline $\bar{u}$, learning rate α.
Output: explainer E_θ.

initialise E_θ
while *not converged* **do**
- sample $u \sim \mathcal{U}(\mathcal{X})$ with cells $B_1, \dots, B_n$, $x^\star \sim \mathcal{U}(\Omega)$, A with $\mathbb{P}(A) \propto \mu_n(A)$
- $(\hat{\varphi}, \hat{\Phi}) \leftarrow E_\theta(u, x^\star)$
- $\hat{E}_a, \hat{E}_{ab} \leftarrow$ quadratures Eq. (13)
- $v(S) \leftarrow \hat{\mathcal{F}}_{x^\star}(u_S) - \hat{\mathcal{F}}_{x^\star}(\bar{u})$ for $S = A$ and $S = [n]$
- project $\hat{E}$ onto $\sum_T \hat{E}_T = v([n])$ along z
- $\hat{v}(A) \leftarrow \sum_{a \in A} \hat{E}_a + \sum_{a<b \in A} \hat{E}_{ab}$
- $\mathcal{L} \leftarrow \big(v(A) - \hat{v}(A)\big)^2$
- $\theta \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}$

Attribution quality. OperatorSHAP-IQ shows competitive performance to both ProxySHAP at budget 16,384 and the direct evaluation of the limit fields (Table 1). While the limit field has similarly quick inference, for reasons of non-differentiability it fully fails on some PDEs like Gray-Scott

or Advection, as well as some models, such as the GINO model of the MGN datasets or the ReLU-backbones (Section B.3). Pair agreement with the reference drops on fine grids for every method, ProxySHAP at $16{,}384$ coalitions included, because the ProxySHAP labels are noisy on fine grids (Section B.7). This shows that agreement with a baseline is not a suitable measure. We still include it to show our reasoning, as it was part of OperatorSHAP. On the MGN airfoil, ProxySHAP is noticably more faithful than OperatorSHAP-IQ, which can be explained by poor training results of the explainer, yielding a limitaiton of OperatorSHAP-IQ (Section B.6). The coefficients of the explainer satisfy efficiency to 10^{-8}, against a median gap of $2 \cdot 10^{-3}$ for ProxySHAP (Table 6).

Amortization and speed. ProxySHAP costs b backbone calls per explanation. The explainer costs $2.7 \cdot 10^8$ calls once for training and one per explanation, so at the reference budget it is cheaper from the $5{,}470$th explanation on. One explanation takes the explainer 7 to 14 ms against 1.5 to 11.5 s for ProxySHAP at $16{,}384$ coalitions on the same CPU (Section B.9).

Analytic attribution. For the heat semigroup, the fields of Theorem 7 have a closed form. Figure 6 compares it with trained explainers and shows that their explanations agree with what we expect analytically from the heat kernel. Figure 3 shows a concrete use case with a learned explainer on Burgers equation (Section C). The example shows two bumps traveling to the right. Depending on the location and time of the readout, the faster bump on the back has surpassed to slower one on the front and their contribution adds up in the result (i.e., they show positive interaction) or not (i.e., they show negative interaction). This shows how interactions can be used to interpret the behaviour of the predictive model.

6 Conclusion and Discussion

We derived an infinite player attribution method for Shapley interactions, which can be used for data on continuous domains or with irregular grids. We constructed this as a limit of the FSII on refining meshes and developed an amortization method with which attribution densities over continuous domains can be efficiently evaluated. We demonstrated its usefulness on a large benchmark, including selected use cases, where interactions could be directly linked to PDE behaviour.

Limitations. Amortization usually pays off only in a production setting, where deployed models are monitored, or customers are supplied with an explanation along the prediction. If just a few explanations are needed, a sampling estimator is cheaper. Neural operators often work at high resolutions, which increases the computational complexity or necessitates coarser attribution fidelity. Moreover, the explainer is training dependent, i.e., little data or incorrect models can reduce performance as we saw in our airfoil example. Regarding the theory, we need smooth masks and a Fréchet-C^ℓ backbone, though in practice the explainer works as well with a ReLU backbone (Section B.3). While higher order interactions are covered by the theory, the algorithm is written for $\ell = 2$, and all of our experiments only show $\ell = 2$. Higher orders are computationally heavier and their practical use is limited, as interpreting them is complex.

For **future work**, we see three directions. First, our construction covers every signed mixture index, so other interaction indices (e.g., Banzhaf indices) can be amortized with the same explainer once their measures are known. Second, Shapley values and interactions are an active research field and many engineering improvements that have been done for the finite case might be transferable to the infinite player case. Third, the theory could be extended to reduce the necessary assumptions, to make the choice of non-atomic measures more dynamic, or to connect it to the existing Aumann-Shapley theory.

AI use statement

We used LLM assistants for the six purposes disclosed on the OpenReview submission form. Writing: polishing the text. Retrieval and discovery: searching for related work; the authors read and verified every cited reference. Research ideation and execution: exploring candidate methods and experiments, and writing the experiment code; the authors reviewed and tested all code. Drafting: drafting sections of the paper from our notes. Synthetic datasets: the scripts that generate our PDE datasets were written with AI assistance; the data itself comes from numerical solvers, not from a

language model. Mathematical claims: drafting and checking proof steps; the authors verified every proof in the appendix. We have reviewed all AI-assisted output and take responsibility for the final content, including text, claims, and artifacts.

Ethics statement

This work studies attribution methods for neural PDE surrogates on synthetic physical data and involves no human subjects, personal data, or foreseeable harmful applications, so no ethical concerns apply.

Reproducibility statement

Source code for data generation, training, and evaluation, with all configuration files, is provided as supplementary material. Proofs of all theoretical claims, including the assumptions they require, are in the appendix. The datasets are generated by scripts in the repository, so all experiments can be rerun from scratch.

References

Robert J. Aumann and Lloyd S. Shapley. *Values of non-atomic games*. Princeton University Press, December 2015. ISBN 978-1-4008-6708-0. doi: 10.1515/9781400867080. URL https://www.degruyterbrill.com/document/doi/10.1515/9781400867080/html.

Sebastian Bordt and Ulrike von Luxburg. From Shapley Values to Generalized Additive Models and back. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 709–745. PMLR, April 2023. URL https://proceedings.mlr.press/v206/bordt23a.html.

James Enouen and Yan Liu. InstaSHAP: interpretable additive models explain shapley values instantly. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 22720–22765, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/38fc6254f73450813db3b3e04397a9fc-Paper-Conference.pdf.

Matthias Fey and Jan Eric Lenssen. Fast Graph Representation Learning with PyTorch Geometric, April 2019. URL http://arxiv.org/abs/1903.02428. arXiv:1903.02428 [cs.LG].

Katsushige Fujimoto, Ivan Kojadinovic, and Jean-Luc Marichal. Axiomatic characterizations of probabilistic and cardinal-probabilistic interaction indices. *Games and Economic Behavior*, 55 (1):72–99, April 2006. ISSN 08998256. doi: 10.1016/j.geb.2005.03.002. URL https://linkinghub.elsevier.com/retrieve/pii/S0899825605000278.

Fabian Fumagalli, Maximilian Muschalik, Patrick Kolpaczki, Eyke Hüllermeier, and Barbara Hammer. SHAP-IQ: unified approximation of any-order shapley interactions. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 11515–11551. Curran Associates, Inc., 2023. doi: 10.52202/075280-0508. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/264f2e10479c9370972847e96107db7f-Paper-Conference.pdf.

Fabian Fumagalli, Maximilian Muschalik, Patrick Kolpaczki, Eyke Hüllermeier, and Barbara Hammer. KernelSHAP-IQ: weighted least square optimization for shapley interactions. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 14308–14342. PMLR, July 2024. URL https://proceedings.mlr.press/v235/fumagalli24a.html.

Vignesh Gopakumar, Stanislas Pamela, Lorenzo Zanisi, Zongyi Li, Ander Gray, Daniel Brennand, Nitesh Bhatia, Gregory Stathopoulos, Matt Kusner, Marc Peter Deisenroth, Anima Anandkumar,

JOREK Team, and MAST Team. Plasma surrogate modelling using fourier neural operators. *Nuclear Fusion*, 64(5):056025, May 2024. ISSN 0029-5515, 1741-4326. doi: 10.1088/1741-4326/ad313a. URL http://arxiv.org/abs/2311.05967. arXiv:2311.05967 [physics].

Michel Grabisch and Marc Roubens. An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of Game Theory*, 28(4):547–565, November 1999. ISSN 0020-7276, 1432-1270. doi: 10.1007/s001820050125. URL http://link.springer.com/10.1007/s001820050125.

Gaurav Gupta, Xiongye Xiao, and Paul Bogdan. Multiwavelet-based operator learning for differential equations. In *Advances in Neural Information Processing Systems*, pp. 24048–24062, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/c9e5c2b59d98488fe1070e744041ea0e-Abstract.html.

Zhongkai Hao, Zhengyi Wang, Hang Su, Chengyang Ying, Yinpeng Dong, Songming Liu, Ze Cheng, Jian Song, and Jun Zhu. GNOT: a general neural operator transformer for operator learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 12556–12569. PMLR, July 2023. URL https://proceedings.mlr.press/v202/hao23c.html.

Dan Hendrycks and Kevin Gimpel. Gaussian Error Linear Units (GELUs), June 2023. URL http://arxiv.org/abs/1606.08415. arXiv:1606.08415 [cs.LG].

Maximilian Herde, Bogdan Raonić, Tobias Rohner, Roger Käppeli, Roberto Molinaro, Emmanuel de Bézenac, and Siddhartha Mishra. Poseidon: Efficient Foundation Models for PDEs. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 72525–72624. Curran Associates, Inc., 2024. doi: 10.52202/079017-2311. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/84e1b1ec17bb11c57234e96433022a9a-Paper-Conference.pdf.

Armeet Singh Jatyani, Jiayun Wang, Aditi Chandrashekar, Zihui Wu, Miguel Liu-Schiaffini, Bahareh Tolooshams, and Anima Anandkumar. A unified model for compressed sensing MRI across undersampling patterns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26004–26013, June 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Jatyani_A_Unified_Model_for_Compressed_Sensing_MRI_Across_Undersampling_Patterns_CVPR_2025_paper.html.

Neil Jethani, Mukund Sudarshan, Ian Covert, Su-In Lee, and Rajesh Ranganath. FastSHAP: real-time shapley value estimation. In *International Conference on Learning Representations*, volume abs/2107.07436. arXiv, March 2022. doi: 10.48550/arXiv.2107.07436. URL http://arxiv.org/abs/2107.07436. arXiv:2107.07436 [stat].

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: a highly efficient gradient boosting decision tree. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

Alexander S. Kechris. *Classical Descriptive Set Theory*, volume 156 of *Graduate Texts in Mathematics*. Springer New York, New York, NY, 1995. ISBN 978-1-4612-8692-9 978-1-4612-4190-4. doi: 10.1007/978-1-4612-4190-4. URL http://link.springer.com/10.1007/978-1-4612-4190-4.

Patrick Kolpaczki and Eyke Hüllermeier. Approximation Algorithms for the Shapley Value: Taxonomy and Properties. In Diego Calvanese (ed.), *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence, IJCAI-26*, pp. 7873–7881. International Joint Conferences on Artificial Intelligence Organization, August 2026. doi: 10.24963/ijcai.2026/874. URL https://doi.org/10.24963/ijcai.2026/874.

Patrick Kolpaczki, Maximilian Muschalik, Fabian Fumagalli, Barbara Hammer, and Eyke Hüllermeier. SVARM-IQ: Efficient Approximation of Any-order Shapley Interactions through Stratification. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 3520–3528. PMLR, May 2024. URL https://proceedings.mlr.press/v238/kolpaczki24a.html.

Jean Kossaifi, Nikola Kovachki, Zongyi Li, David Pitt, Miguel Liu-Schiaffini, Robert J. George, Boris Bonev, Kamyar Azizzadenesheli, Julius Berner, Valentin Duruisseaux, and Anima Anandkumar. A Library for Learning Neural Operators. *Journal of Machine Learning Research*, 27 (205):1–6, 2026. URL http://jmlr.org/papers/v27/26-0434.html.

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. URL http://jmlr.org/papers/v24/21-1524.html.

Thorsten Kurth, Shashank Subramanian, Peter Harrington, Jaideep Pathak, Morteza Mardani, David Hall, Andrea Miele, Karthik Kashinath, and Anima Anandkumar. FourCastNet: accelerating global high-resolution weather forecasting using adaptive fourier neural operators. In Axel Huebl, Cristina Silvano, and Timothy Robinson (eds.), *Proceedings of the Platform for Advanced Scientific Computing Conference*, pp. 13:1–13:11. ACM, 2023. doi: 10.1145/3592979.3593412. URL https://doi.org/10.1145/3592979.3593412.

Zongyi Li, Nikola Borislavov Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew M. Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*. OpenReview.net, 2021. URL https://openreview.net/forum?id=c8P9NQVtmnO.

Zongyi Li, Nikola Kovachki, Chris Choy, Boyi Li, Jean Kossaifi, Shourya Otta, Mohammad Amin Nabian, Maximilian Stadler, Christian Hundt, Kamyar Azizzadenesheli, and Animashree Anandkumar. Geometry-informed neural operator for large-scale 3D PDEs. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 35836–35854. Curran Associates, Inc., 2023. doi: 10.48550/arXiv.2309.00583. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/70518ea42831f02afc3a2828993935ad-Paper-Conference.pdf.

AA Liapounoff. Sur les fonctions-vecteurs completement additives. *Izvestiya Rossiiskoi Akademii Nauk. Seriya Matematicheskaya*, 4(6):465–478, 1940.

Miguel Liu-Schiaffini, Julius Berner, Boris Bonev, Thorsten Kurth, Kamyar Azizzadenesheli, and Anima Anandkumar. Neural operators with localized integral and differential kernels. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32576–32594. PMLR, July 2024. URL https://proceedings.mlr.press/v235/liu-schiaffini24a.html.

Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*. OpenReview.net, 2017. URL https://openreview.net/forum?id=Skq89Scxx.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30, pp. 4765–4774, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

Maximilian Muschalik, Hubert Baniecki, Fabian Fumagalli, Patrick Kolpaczki, Barbara Hammer, and Eyke Hüllermeier. shapiq: Shapley Interactions for Machine Learning. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=knxGmi6SJi.

Abraham Neyman. Values of games with infinitely many players. In *Handbook of Game Theory with Economic Applications*, volume 3, pp. 2121–2167. Elsevier, 2002. ISBN 978-0-444-89428-1. doi: 10.1016/S1574-0005(02)03019-9. URL https://linkinghub.elsevier.com/retrieve/pii/S1574000502030199.

Tobias Pfaff, Meire Fortunato, Alvaro Sanchez-Gonzalez, and Peter Battaglia. Learning Mesh-Based Simulation with Graph Networks. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=roNqYL0_XP.

David A. Ross. An Elementary Proof of Lyapunov's Theorem. *The American Mathematical Monthly*, 112(7):651–653, 2005. ISSN 00029890, 19300972. URL http://www.jstor.org/stable/30037551.

Thomas Runst and Winfried Sickel. *Sobolev spaces of fractional order, nemytskij operators, and nonlinear partial differential equations*. Number 3 in De Gruyter series in nonlinear analysis and applications. Walter de Gruyter, Berlin ; New York, 1996. ISBN 978-3-11-015113-8.

Lloyd S. Shapley. A value for *n* -person games. In Alvin E. Roth (ed.), *The Shapley Value*, pp. 31–40. Cambridge University Press, 1 edition, October 1988. ISBN 978-0-521-36177-4 978-0-521-02133-3 978-0-511-52844-6. doi: 10.1017/CBO9780511528446.003. URL https://www.cambridge.org/core/product/identifier/CBO9780511528446A008/type/book_part.

Joshua Stiller, Santo M. A. R. Thies, Felix Czaja, and Eyke Hüllermeier. OperatorSHAP: fast and accurate shapley value estimation for neural operators, June 2026. URL http://arxiv.org/abs/2606.28065. arXiv:2606.28065 [cs.LG].

Mukund Sundararajan, Kedar Dhamdhere, and Ashish Agarwal. The shapley taylor interaction index. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9259–9268. PMLR, July 2020. URL https://proceedings.mlr.press/v119/sundararajan20a.html.

Santo M. A. R. Thies, Hubert Baniecki, R. Teal Witter, Eyke Hüllermeier, Maximilian Muschalik, and Fabian Fumagalli. Proxy-based approximation of shapley and banzhaf interactions, May 2026. URL http://arxiv.org/abs/2605.22738. arXiv:2605.22738 [cs.LG].

Che-Ping Tsai, Chih-Kuan Yeh, and Pradeep Ravikumar. Faith-Shap: the faithful shapley interaction index. *Journal of Machine Learning Research*, 24(94):1–42, 2023. URL http://jmlr.org/papers/v24/22-0202.html.

Shizheng Wen, Arsh Kumbhat, Levi Lingsch, Sepehr Mousavi, Yizhou Zhao, Praveen Chandrashekar, and Siddhartha Mishra. Geometry aware operator transformer as an efficient and accurate neural surrogate for PDEs on arbitrary domains. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 155423–155501. Curran Associates, Inc., 2025. doi: 10.52202/085713-5197. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/e45a448dfa778f6d62729a7bc8633c06-Paper-Conference.pdf.

A Proofs

A.1 Proof of Theorem 5 (exact mixture representation)

For a game v on $[n]$, its *Möbius transform* is $a(v,S) = \sum_{T\subseteq S}(-1)^{|S|-|T|}v(T)$, and the *unanimity game* of $R \subseteq [n]$ is $v_R(S) := \mathbb{1}[S \supseteq R]$. Möbius inversion gives $v = \sum_{R\subseteq[n]} a(v,R)\, v_R$, and

$\{v_R\}_R$ is a basis of the space of games on $[n]$ (Tsai et al., 2023, Eq. (4)). In this basis, the FSII of Definition 1 has the closed form (Tsai et al., 2023, Thm. 19, Eq. (16))

$$\mathcal{E}_T(v,\ell) = a(v,T) + (-1)^{\ell-|T|}\,\frac{|T|}{\ell+|T|}\binom{\ell}{|T|}\sum_{R\supsetneq T,\ |R|>\ell}\frac{\binom{|R|-1}{\ell}}{\binom{|R|+\ell-1}{\ell+|T|}}\,a(v,R), \qquad 1\le|T|\le\ell. \tag{14}$$

The proof of Theorem 5 compares the coefficients of $a(v,R)$ on the two sides of Eq. (6). Both are values of one rational function. For integers $1\le k\le\ell$ define

$$R_{k,\ell}(r) := (-1)^{\ell-k}\,\frac{k\,(\ell+k-1)!}{k!\,(\ell-k)!}\cdot\frac{\prod_{i=1}^{\ell-k}(r-i)}{\prod_{j=k}^{k+\ell-1}(r+j)}, \tag{15}$$

with empty products equal to 1. Since $\prod_{i=1}^{\ell-k}(-i) = (-1)^{\ell-k}(\ell-k)!$ and $\prod_{j=k}^{k+\ell-1} j = (k+\ell-1)!/(k-1)!$, and since the numerator vanishes at $r=1,\dots,\ell-k$,

$$R_{k,\ell}(0) = \frac{k\,(\ell+k-1)!}{k!\,(\ell-k)!}\cdot\frac{(\ell-k)!\,(k-1)!}{(k+\ell-1)!} = 1, \qquad R_{k,\ell}(r) = 0 \quad \text{for } 1\le r\le\ell-k. \tag{16}$$

Lemma 9 (Coefficients of the closed form)**.** *Let $1\le k\le\ell\le n$ and $T\subseteq[n]$ with $|T|=k$. For every $R\supseteq T$ with $|R|=\tau$, the coefficient of $a(v,R)$ in Eq.* (14) *equals $R_{k,\ell}(\tau-k)$.*

Proof. For $\tau=k$ the coefficient is 1 and for $k<\tau\le\ell$ it is 0, matching Eq. (16). For $\tau>\ell$, expanding the binomial coefficients in Eq. (14),

$$\begin{aligned}\frac{k}{\ell+k}\binom{\ell}{k}\frac{\binom{\tau-1}{\ell}}{\binom{\tau+\ell-1}{\ell+k}} &= \frac{k\,\ell!}{(\ell+k)\,k!\,(\ell-k)!}\cdot\frac{(\tau-1)!}{\ell!\,(\tau-1-\ell)!}\cdot\frac{(\ell+k)!\,(\tau-1-k)!}{(\tau+\ell-1)!}\\ &= \frac{k\,(\ell+k-1)!}{k!\,(\ell-k)!}\cdot\frac{(\tau-1)!\,(\tau-1-k)!}{(\tau-1-\ell)!\,(\tau+\ell-1)!}.\end{aligned}$$

Put $r:=\tau-k>\ell-k\ge0$. The ℓ consecutive integers $\tau-\ell,\dots,\tau-1$ are $r-(\ell-k),\dots,r-1$ followed by $r,\dots,r+k-1$, and the integers $\tau-k,\dots,\tau+\ell-1$ are $r,\dots,r+k+\ell-1$, so

$$\frac{(\tau-1)!}{(\tau-1-\ell)!} = \prod_{i=1}^{\ell-k}(r-i)\prod_{j=0}^{k-1}(r+j), \qquad \frac{(\tau-1-k)!}{(\tau+\ell-1)!} = \Big[\prod_{j=0}^{k+\ell-1}(r+j)\Big]^{-1}.$$

The factors $\prod_{j=0}^{k-1}(r+j)$ cancel, leaving $\prod_{i=1}^{\ell-k}(r-i)\big/\prod_{j=k}^{k+\ell-1}(r+j)$; the sign $(-1)^{\ell-k}$ of Eq. (14) completes $R_{k,\ell}(r)$. □

The mixture index I_k^w is linear in v and in w, since Δ_T and the expectation in Eq. (3) are linear. Its Möbius expansion follows from its values on unanimity games.

Lemma 10 (Möbius expansion of mixture indices)**.** *For every finite signed Borel measure w on $[0,1]$, every game v on $[n]$ and every $T\subseteq[n]$ with $|T|=k\ge1$,*

$$I_k^w(v;T) = \sum_{R\supseteq T} a(v,R)\,m_{|R|-k}(w), \qquad m_r(w) := \int_0^1 t^r\,dw(t).$$

Proof. By Möbius inversion, $v=\sum_{R\subseteq[n]} a(v,R)\,v_R$, and $I_k^w(\,\cdot\,;T)$ is linear in the game, so

$$I_k^w(v;T) = \sum_{R\subseteq[n]} a(v,R)\,I_k^w(v_R;T).$$

Hence the lemma follows once we show, for every $R\subseteq[n]$,

$$I_k^w(v_R;T) = \begin{cases} m_{|R|-k}(w) & R\supseteq T,\\ 0 & \text{otherwise.}\end{cases} \tag{17}$$

The terms with $R\not\supseteq T$ then vanish and the remaining ones are the claimed sum.

To compute the discrete derivative of v_R, fix R and a coalition $S \subseteq [n] \setminus T$. By Eq. (2),

$$(\Delta_T v_R)(S) = \sum_{L \subseteq T} (-1)^{k-|L|}\, v_R(S \cup L) = \sum_{L \subseteq T} (-1)^{k-|L|}\, \mathbb{1}[S \cup L \supseteq R].$$

Since $S \cap T = \emptyset$ and $L \subseteq T$, the players of $R \setminus T$ can only be covered by S and the players of $R \cap T$ can only be covered by L. Therefore $S \cup L \supseteq R$ holds iff $S \supseteq R \setminus T$ and $L \supseteq R \cap T$, and

$$(\Delta_T v_R)(S) = \mathbb{1}[S \supseteq R \setminus T] \sum_{L:\, R \cap T \subseteq L \subseteq T} (-1)^{k-|L|}.$$

Every L with $R \cap T \subseteq L \subseteq T$ is uniquely $L = (R \cap T) \cup M$ with $M \subseteq T \setminus R$, and then $k - |L| = |T \setminus R| - |M|$. With $N := |T \setminus R|$,

$$\sum_{L:\, R \cap T \subseteq L \subseteq T} (-1)^{k-|L|} = \sum_{M \subseteq T \setminus R} (-1)^{N-|M|} = \sum_{j=0}^{N} \binom{N}{j} (-1)^{N-j} = (1-1)^N,$$

which equals 1 if $N = 0$ and 0 if $N \geq 1$. As $N = 0$ means $T \subseteq R$, we have shown

$$(\Delta_T v_R)(S) = \mathbb{1}[R \supseteq T]\, \mathbb{1}[S \supseteq R \setminus T] \qquad \text{for all } S \subseteq [n] \setminus T.$$

If $R \not\supseteq T$, then $\Delta_T v_R \equiv 0$ and both sides of Eq. (17) are 0. If $R \supseteq T$, then $|R \setminus T| = |R| - k$, and $S_t \supseteq R \setminus T$ means that S_t contains each of these $|R| - k$ players. S_t contains each player of $[n] \setminus T$ independently with probability t, so $\mathbb{E}[(\Delta_T v_R)(S_t)] = \mathbb{P}(S_t \supseteq R \setminus T) = t^{|R|-k}$.

Integrating against w as in Eq. (3), $I_k^w(v_R; T) = \int_0^1 t^{|R|-k}\, dw(t) = m_{|R|-k}(w)$ for $R \supseteq T$, which is Eq. (17). □

For a probability measure w, Lemma 10 is Fujimoto et al. (2006, Eq. (6)), and I_k^w is their cardinal-probabilistic interaction index with mixing law w.

Lemma 11 (Moments of the per-order measures)**.** *For all integers $1 \leq k \leq \ell$ and $r \geq 0$, $m_r\big(w_k^{(\ell)}\big) = R_{k,\ell}(r)$. In particular, by Eq. (16), $m_0\big(w_k^{(\ell)}\big) = 1$ and $m_r\big(w_k^{(\ell)}\big) = 0$ for $1 \leq r \leq \ell - k$.*

Proof. As a rational function of r, $R_{k,\ell}$ in Eq. (15) has a numerator of degree $\ell - k$ and a denominator of degree ℓ with the simple zeros $r = -s$, $s = k, \ldots, k + \ell - 1$. None of these is a zero of the numerator, whose zeros $r = 1, \ldots, \ell - k$ are positive. A rational function whose numerator degree is smaller than its denominator degree and whose poles are simple is the sum of its principal parts, so

$$R_{k,\ell}(r) = \sum_{s=k}^{k+\ell-1} \frac{c_s}{r+s}, \qquad c_s = \lim_{r \to -s} (r+s) R_{k,\ell}(r).$$

To compute c_s, note that multiplying by $(r+s)$ removes the factor $(r+s)$ from the denominator of Eq. (15); the remaining expression is evaluated at $r = -s$. The numerator becomes

$$\prod_{i=1}^{\ell-k} (-s-i) = (-1)^{\ell-k} \prod_{i=1}^{\ell-k} (s+i) = (-1)^{\ell-k} \frac{(s+\ell-k)!}{s!},$$

and the remaining denominator becomes, splitting at $j = s$,

$$\prod_{\substack{j=k \\ j \neq s}}^{k+\ell-1} (j-s) = \prod_{j=k}^{s-1} (j-s) \prod_{j=s+1}^{k+\ell-1} (j-s) = (-1)^{s-k} (s-k)!\, (k+\ell-1-s)!.$$

With the prefactor of Eq. (15),

$$\begin{aligned} c_s &= (-1)^{\ell-k} \frac{k\,(\ell+k-1)!}{k!\,(\ell-k)!} \cdot \frac{(-1)^{\ell-k}(s+\ell-k)!/s!}{(-1)^{s-k}(s-k)!\,(k+\ell-1-s)!} \\ &= (-1)^{s-k} \frac{k\,(\ell+k-1)!\,(s+\ell-k)!}{k!\,(\ell-k)!\,s!\,(s-k)!\,(k+\ell-1-s)!} = b_{k,\ell,s}. \end{aligned}$$

Now, for $r \geq 0$ and $s \geq 1$, $\int_0^1 t^r \cdot t^{s-1}\,dt = 1/(r+s)$. Since m_r is linear in the measure and $w_k^{(\ell)}(t)\,dt = \sum_s b_{k,\ell,s}\,t^{s-1}\,dt$ by Eq. (5),

$$m_r\big(w_k^{(\ell)}\big) = \sum_{s=k}^{k+\ell-1} b_{k,\ell,s} \int_0^1 t^{r+s-1}\,dt = \sum_{s=k}^{k+\ell-1} \frac{b_{k,\ell,s}}{r+s} = \sum_{s=k}^{k+\ell-1} \frac{c_s}{r+s} = R_{k,\ell}(r).$$

□

Theorem 5 (Exact mixture representation). *Let $n \geq 2$, $\ell \in [n]$, v a game on $[n]$, and $T \subseteq [n]$ with $1 \leq |T| = k \leq \ell$. Then, the FSII is a $w_k^{(\ell)}$-mixture of discrete derivatives*

$$\mathcal{E}_T(v,\ell) = I_k^{w_k^{(\ell)}}(v;T). \tag{6}$$

Proof. Both sides of Eq. (6) are linear combinations of the Möbius coefficients of v. By Eq. (14) and Lemma 9, the left side of Eq. (6) is

$$\mathcal{E}_T(v,\ell) = \sum_{T \subseteq R \subseteq [n]} R_{k,\ell}(|R|-k)\,a(v,R),$$

where only $R \supseteq T$ occur in Eq. (14). By Lemma 10 with $w = w_k^{(\ell)}\,dt$ and then Lemma 11, the right side is

$$I_k^{w_k^{(\ell)}}(v;T) = \sum_{T \subseteq R \subseteq [n]} m_{|R|-k}\big(w_k^{(\ell)}\big)\,a(v,R) = \sum_{T \subseteq R \subseteq [n]} R_{k,\ell}(|R|-k)\,a(v,R).$$

The two expansions have the same coefficient in front of every $a(v,T)$, so they are equal for every game v. □

A.2 Proof of Theorem 7 (FSII fields)

Partitions of arbitrarily small mesh relative to μ in which every set of a given finite collection is a union of cells exist; the convexity of Lyapunov's theorem (Liapounoff, 1940) supplies the halving step, and Lemma 13 the sequence. The mesh is measured by μ, not by the diameter of the cells, which Theorem 7(iii) assumes separately. Ross (2005) gives a short proof of the theorem.

Theorem 12 (Lyapunov). *Let $\nu_1, \ldots, \nu_d$ be finite non-atomic measures on $(\Omega, \mathcal{C})$ and $\nu := (\nu_1, \ldots, \nu_d)$. Then the range $\{\nu(A) : A \in \mathcal{C}\}$ is a compact convex subset of $\mathbb{R}^d$.*

Applied on the measurable space $(B, \{A \in \mathcal{C} : A \subseteq B\})$ for $B \in \mathcal{C}$, on which the ν_i remain finite and non-atomic, the theorem shows that $\{\nu(A) : A \in \mathcal{C},\ A \subseteq B\}$ is compact and convex.

Lemma 13 (Halving; admissible sequences). *Let $\nu_1, \ldots, \nu_d$ be finite non-atomic measures on $(\Omega, \mathcal{C})$ and $\nu := (\nu_1, \ldots, \nu_d)$. (i) Every $B \in \mathcal{C}$ splits as $B = B' \sqcup B''$ with $\nu(B') = \nu(B'') = \frac{1}{2}\nu(B)$. (ii) For every finite collection $S_1, \ldots, S_r \in \mathcal{C}$ there is a sequence of partitions $\pi_0 \preceq \pi_1 \preceq \cdots$, each refining the previous, such that every S_l is a union of cells of π_0 and $\max_{i,c} \nu_i(B_c) \leq 2^{-j} \max_i \nu_i(\Omega)$ for every cell B_c of π_j.*

Proof. (i) Let $K := \{\nu(A) : A \in \mathcal{C},\ A \subseteq B\}$. K contains $\nu(\emptyset) = 0$ and $\nu(B)$, and K is convex by Theorem 12, so K contains the midpoint $\frac{1}{2}\nu(B)$. Choose $B' \subseteq B$ with $\nu(B') = \frac{1}{2}\nu(B)$ and put $B'' := B \setminus B'$. By additivity, $\nu(B'') = \nu(B) - \nu(B') = \frac{1}{2}\nu(B)$.

(ii) Let π_0 consist of the non-empty sets $\bigcap_{l=1}^r C_l$ with $C_l \in \{S_l, \Omega \setminus S_l\}$. These sets are pairwise disjoint and cover Ω, and S_l is the union of those cells with $C_l = S_l$.

Given π_j, let π_{j+1} consist of the non-empty ones among the sets B', B'' obtained from every cell B of π_j by (i); a cell with $\nu(B) = 0$ may produce an empty child, which is discarded. Then π_{j+1} is a partition refining π_j, so every union of cells of π_j, in particular every S_l, is a union of cells of π_{j+1}.

The mesh bound follows by induction on j: for $j = 0$, $\nu_i(B) \leq \nu_i(\Omega)$ for every cell by monotonicity; and if $\nu_i(B) \leq 2^{-j}\nu_i(\Omega)$ for every cell B of π_j, then every cell of π_{j+1} has half that mass in every component by (i), hence at most $2^{-(j+1)}\nu_i(\Omega)$. □

Tensor notation. For $f \in C^k([0,1]^m)$, $D^k f(x)$ is the k-linear form on $\mathbb{R}^m$ of Section 3.3; it is symmetric by Schwarz's theorem, and by the chain rule $D^k f(x)[q_1, \dots, q_k]$ is the mixed derivative $\partial_{s_1} \cdots \partial_{s_k}$ of $s \mapsto f(x + \sum_l s_l q_l)$ at $s = 0$. For a k-linear form A with entries $A_{i_1 \cdots i_k}$, $\|A\|_\infty := \max |A_{i_1 \cdots i_k}|$. The two bounds below are the operator-norm estimates for k-linear forms when vectors carry $|\cdot|_1$ and forms carry $\|\cdot\|_\infty$.

Lemma 14 (Multilinear perturbation bounds)**.** *For k-linear forms A, B on $\mathbb{R}^m$ and $q_1, \dots, q_k, q'_1, \dots, q'_k \in \mathbb{R}^m$,*

$$\big|A[q_1, \dots, q_k] - B[q_1, \dots, q_k]\big| \le \|A - B\|_\infty \prod_{l=1}^{k} |q_l|_1,$$

$$\big|A[q_1, \dots, q_k] - A[q'_1, \dots, q'_k]\big| \le \|A\|_\infty \sum_{l=1}^{k} |q_l - q'_l|_1 \prod_{j<l} |q_j|_1 \prod_{j>l} |q'_j|_1.$$

Proof. First bound. By multilinearity, $A[q_1, \dots, q_k] - B[q_1, \dots, q_k] = \sum_{i_1, \dots, i_k} (A - B)_{i_1 \cdots i_k} \prod_l q_{l, i_l}$. The triangle inequality and $|(A - B)_{i_1 \cdots i_k}| \le \|A - B\|_\infty$ give the bound $\|A - B\|_\infty \sum_{i_1, \dots, i_k} \prod_l |q_{l, i_l}|$, and the sum factorizes: $\sum_{i_1, \dots, i_k} \prod_l |q_{l, i_l}| = \prod_l \sum_i |q_{l,i}| = \prod_l |q_l|_1$.

Second bound. Replace the arguments one slot at a time. With $A_l := A[q_1, \dots, q_l, q'_{l+1}, \dots, q'_k]$ for $l = 0, \dots, k$, we have $A_k = A[q_1, \dots, q_k]$, $A_0 = A[q'_1, \dots, q'_k]$ and the telescoping sum $A_k - A_0 = \sum_{l=1}^{k} (A_l - A_{l-1})$. The terms A_l and A_{l-1} differ only in slot l, so by linearity in that slot $A_l - A_{l-1} = A[q_1, \dots, q_{l-1}, q_l - q'_l, q'_{l+1}, \dots, q'_k]$. The first bound with $B = 0$ gives $|A_l - A_{l-1}| \le \|A\|_\infty \prod_{j<l} |q_j|_1 \cdot |q_l - q'_l|_1 \cdot \prod_{j>l} |q'_j|_1$; sum over l. □

Lemma 15 (Integral representation of k-th differences)**.** *Let $f \in C^k([0,1]^m)$, $x \in [0,1]^m$ and $q_1, \dots, q_k \in [0, \infty)^m$ with $x + \sum_{l=1}^{k} q_l \in [0,1]^m$. Then*

$$\sum_{L \subseteq [k]} (-1)^{k - |L|} f\Big(x + \sum_{l \in L} q_l\Big) = \int_{[0,1]^k} D^k f\Big(x + \sum_{l=1}^{k} s_l q_l\Big)[q_1, \dots, q_k]\, ds,$$

and all arguments of f and $D^k f$ lie in $[0,1]^m$.

Proof. Domain. For $s \in [0,1]^k$ and $L \subseteq [k]$, coordinatewise $0 \le x_i \le x_i + \sum_l s_l q_{l,i} \le x_i + \sum_l q_{l,i} \le 1$ because all $q_{l,i} \ge 0$.

Identity, base case $k = 1$. The map $s \mapsto f(x + s q_1)$ is C^1 on $[0,1]$ with derivative $Df(x + s q_1)[q_1]$ by the chain rule, so the fundamental theorem of calculus gives $f(x + q_1) - f(x) = \int_0^1 Df(x + s q_1)[q_1]\, ds$.

Identity, induction step $k - 1 \to k$. Let $k \ge 2$ and assume the identity for order $k - 1$, for all base points and vectors that satisfy the domain condition. For $y \in [0,1]^m$ with $y + \sum_{l<k} q_l \in [0,1]^m$ define

$$g(y) := \int_{[0,1]^{k-1}} D^{k-1} f\Big(y + \sum_{l<k} s_l q_l\Big)[q_1, \dots, q_{k-1}]\, ds_1 \cdots ds_{k-1}.$$

Split the sum over $L \subseteq [k]$ according to whether $k \in L$, writing $L = L' \cup \{k\}$ or $L = L'$ with $L' \subseteq [k-1]$. Since $(-1)^{k - |L' \cup \{k\}|} = (-1)^{k-1-|L'|}$ and $(-1)^{k-|L'|} = -(-1)^{k-1-|L'|}$,

$$\begin{aligned}\sum_{L \subseteq [k]} (-1)^{k-|L|} f\Big(x + \sum_{l \in L} q_l\Big) = &\sum_{L' \subseteq [k-1]} (-1)^{k-1-|L'|} f\Big(x + q_k + \sum_{l \in L'} q_l\Big) \\ &- \sum_{L' \subseteq [k-1]} (-1)^{k-1-|L'|} f\Big(x + \sum_{l \in L'} q_l\Big).\end{aligned}$$

Both sums are $(k-1)$-th differences with vectors $q_1, \dots, q_{k-1}$, at the base points $x + q_k$ and x; both base points satisfy the domain condition because $x + q_k + \sum_{l<k} q_l = x + \sum_{l \le k} q_l \in [0,1]^m$. By the induction hypothesis the right side equals $g(x + q_k) - g(x)$.

It remains to write $g(x+q_k)-g(x)$ as an integral. Put $\gamma(s_k) := g(x+s_kq_k)$ for $s_k \in [0,1]$. The integrand of $g(x+s_kq_k)$ is $F(s_k,s') := D^{k-1}f\big(x+s_kq_k+\sum_{l<k} s_lq_l\big)[q_1,\dots,q_{k-1}]$ with $s' = (s_1,\dots,s_{k-1})$. Since $f \in C^k$, the entries of $D^{k-1}f$ are C^1, and the chain rule gives $\partial_{s_k}F(s_k,s') = D^kf\big(x+\sum_{l=1}^{k} s_lq_l\big)[q_1,\dots,q_{k-1},q_k]$, which is continuous in (s_k,s') on the compact set $[0,1]^k$ and hence bounded. Differentiation under the integral sign is therefore allowed and gives $\gamma'(s_k) = \int_{[0,1]^{k-1}} D^kf\big(x+\sum_{l=1}^{k} s_lq_l\big)[q_1,\dots,q_k]\,ds'$. By the fundamental theorem of calculus, $g(x+q_k)-g(x) = \gamma(1)-\gamma(0) = \int_0^1 \gamma'(s_k)\,ds_k$, and the resulting iterated integral over $[0,1]^{k-1}\times[0,1]$ is the integral over $[0,1]^k$ by Fubini's theorem, as the integrand is continuous. □

For the restricted game v_π of Section 3.3, $T=\{a_1,\dots,a_k\}\subseteq[n]$, $S\subseteq[n]\setminus T$ and $x_S := \sum_{c\in S} p_c$, the coalition values in Eq. (2) are $v_\pi(S\cup L) = f\big(x_S+\sum_{l\in L} p_{a_l}\big)$ for $L\subseteq T$, because S and T are disjoint. So Lemma 15 with $x=x_S$ and $q_l=p_{a_l}$ gives

$$(\Delta_T v_\pi)(S) = \int_{[0,1]^k} D^kf\Big(x_S+\sum_l s_lp_{a_l}\Big)[p_{a_1},\dots,p_{a_k}]\,ds, \tag{18}$$

where the domain condition holds because $x_S+\sum_l p_{a_l} \le \sum_{c\in[n]} p_c = \mathbf{1}$ componentwise.

Theorem 16 (Uniform local limit)**.** *Let $v=f\circ\mu$ be a smooth game of class C^k, π a partition with $\delta := \mathrm{mesh}_\mu(\pi)$, and w a finite signed Borel measure on $[0,1]$. Then for every set $T=\{a_1,\dots,a_k\}$ of k distinct cells of π,*

$$\Big|I_k^w(v_\pi;T) - \int_0^1 D^kf(t\mathbf{1})[p_{a_1},\dots,p_{a_k}]\,dw(t)\Big| \le \|w\|_{\mathrm{TV}}\Big(\prod_{l=1}^{k}|p_{a_l}|_1\Big)\rho(\delta),$$

with $\rho=\rho_k$ of Eq. (7), $L=L_k$ and $\omega=\omega_k$. The function ρ depends neither on w nor on T, and the bound is uniform over all partitions of mesh at most δ.

Proof. Fix T and write $q_l := p_{a_l}$; by definition of the mesh, $|q_l|_\infty \le \delta$ and $|p_c|_\infty \le \delta$ for every cell. Let $(\xi_c)_{c\in[n]\setminus T}$ be i.i.d. Bernoulli(t) and $X_t := \sum_{c\notin T}\xi_c p_c$. The random coalition S_t of Eq. (3) is $\{c\notin T : \xi_c=1\}$, so $x_{S_t}=X_t$. Inserting Eq. (18) into Eq. (3) gives

$$I_k^w(v_\pi;T) = \int_0^1 \mathbb{E}\Big[\int_{[0,1]^k} D^kf(Y_{t,s})[q_1,\dots,q_k]\,ds\Big]dw(t), \qquad Y_{t,s} := X_t+\sum_l s_lq_l,$$

and $Y_{t,s}\in[0,1]^m$ by the domain argument of Eq. (18).

The main term $\int_0^1 D^kf(t\mathbf{1})[q_1,\dots,q_k]\,dw(t)$ has the same form with the constant $D^kf(t\mathbf{1})$ in place of $D^kf(Y_{t,s})$, since the constant is unchanged by the expectation and by the integral over s. Subtracting,

$$I_k^w(v_\pi;T) - \int_0^1 D^kf(t\mathbf{1})[q_1,\dots,q_k]\,dw(t) = \int_0^1 h(t)\,dw(t),$$
$$h(t) := \mathbb{E}\Big[\int_{[0,1]^k}\big(D^kf(Y_{t,s})-D^kf(t\mathbf{1})\big)[q_1,\dots,q_k]\,ds\Big].$$

For the signed measure $w=w^+-w^-$ and bounded measurable h, $\big|\int_0^1 h\,dw\big| \le \int_0^1 |h|\,d|w| \le \|w\|_{\mathrm{TV}}\sup_t|h(t)|$. For h itself, the first bound of Lemma 14 gives pointwise $|(D^kf(Y_{t,s})-D^kf(t\mathbf{1}))[q_1,\dots,q_k]| \le \|D^kf(Y_{t,s})-D^kf(t\mathbf{1})\|_\infty\prod_l|q_l|_1$, and moving the absolute value inside the expectation and the s-integral,

$$|h(t)| \le \Big(\prod_l|q_l|_1\Big)\sup_{s\in[0,1]^k}\mathbb{E}\big\|D^kf(Y_{t,s})-D^kf(t\mathbf{1})\big\|_\infty.$$

Hence the theorem follows once we show

$$\sup_{t\in[0,1],\,s\in[0,1]^k}\mathbb{E}\big\|D^kf(Y_{t,s})-D^kf(t\mathbf{1})\big\|_\infty \le \rho(\delta). \tag{19}$$

We first bound the distance of $Y_{t,s}$ to $t\mathbf{1}$. Since $\mathbb{E}\xi_c = t$ and $\sum_{c\in[n]} p_c = \mathbf{1}$, $\mathbb{E}X_t = t\sum_{c\notin T} p_c = t\big(\mathbf{1} - \sum_l q_l\big)$, so $|\mathbb{E}X_t - t\mathbf{1}|_\infty = t\,|\sum_l q_l|_\infty \le \sum_l |q_l|_\infty \le k\delta$. Likewise $|\sum_l s_l q_l|_\infty \le \sum_l |q_l|_\infty \le k\delta$. By the triangle inequality,

$$|Y_{t,s} - t\mathbf{1}|_\infty \le |X_t - \mathbb{E}X_t|_\infty + |\mathbb{E}X_t - t\mathbf{1}|_\infty + \Big|\sum_l s_l q_l\Big|_\infty \le |X_t - \mathbb{E}X_t|_\infty + 2k\delta.$$

Next, X_t concentrates around its mean. Fix a coordinate i. $(X_t)_i = \sum_{c\notin T} \xi_c p_{ic}$ is a sum of independent terms with $\mathrm{Var}(\xi_c p_{ic}) = t(1-t)p_{ic}^2 \le \frac{1}{4}p_{ic}^2$, so, using $p_{ic} \le \delta$ and $\sum_c p_{ic} = \mu_i(\Omega) = 1$,

$$\mathrm{Var}\big((X_t)_i\big) \le \tfrac{1}{4}\sum_c p_{ic}^2 \le \tfrac{1}{4}\big(\max_c p_{ic}\big)\sum_c p_{ic} \le \tfrac{\delta}{4}.$$

Chebyshev's inequality gives $\mathbb{P}(|(X_t)_i - \mathbb{E}(X_t)_i| > \varepsilon) \le \delta/(4\varepsilon^2)$ for every $\varepsilon > 0$, and a union bound over the m coordinates gives $\mathbb{P}(|X_t - \mathbb{E}X_t|_\infty > \varepsilon) \le m\delta/(4\varepsilon^2)$, uniformly in t.

Now fix $\varepsilon > 0$, t and s, and let $E := \{|X_t - \mathbb{E}X_t|_\infty \le \varepsilon\}$. On E, the distance bound gives $|Y_{t,s} - t\mathbf{1}|_\infty \le 2k\delta + \varepsilon$, so $\|D^k f(Y_{t,s}) - D^k f(t\mathbf{1})\|_\infty \le \omega(2k\delta + \varepsilon)$ by the definition of the modulus of continuity. On the complement of E, $\|D^k f(Y_{t,s}) - D^k f(t\mathbf{1})\|_\infty \le 2L$ by the triangle inequality and the definition of L. Therefore

$$\mathbb{E}\big\|D^k f(Y_{t,s}) - D^k f(t\mathbf{1})\big\|_\infty \le \omega(2k\delta+\varepsilon)\,\mathbb{P}(E) + 2L\,\mathbb{P}(E^c) \le \omega(2k\delta+\varepsilon) + 2L\cdot\frac{m\delta}{4\varepsilon^2}.$$

The choice $\varepsilon = \delta^{1/3}$ turns the last term into $\frac{mL}{2}\delta^{1/3}$, and the right side is $\rho(\delta)$ of Eq. (7). As the bound does not depend on t and s, Eq. (19) holds.

Finally, $D^k f$ is continuous on the compact cube, hence uniformly continuous, so $\omega(r) \to 0$ as $r \to 0$ and $\rho(\delta) \to 0$. If $\omega(r) \le \Lambda r$, then $\omega(2k\delta + \delta^{1/3}) \le \Lambda(2k\delta + \delta^{1/3})$ and both terms of $\rho(\delta)$ are $O(\delta^{1/3})$. □

Corollary 17 (Aggregation over regions)**.** *Let $v = f \circ \mu$ be a smooth game of class C^k, w a finite signed Borel measure on $[0,1]$, $S_1, \dots, S_k \in \mathcal{C}$ pairwise disjoint, and π a partition of mesh δ in which each S_l is a union of cells. Then, with ρ as in Theorem 16,*

$$\Big|\sum_{B_{a_1}\subseteq S_1}\cdots\sum_{B_{a_k}\subseteq S_k} I_k^w\big(v_\pi;\{a_1,\dots,a_k\}\big) - \int_0^1 D^k f(t\mathbf{1})\big[\mu(S_1),\dots,\mu(S_k)\big]\,dw(t)\Big| \le \|w\|_{\mathrm{TV}}\, m^k \rho(\delta).$$

Proof. A cell $B_{a_l} \subseteq S_l$ and a cell $B_{a_{l'}} \subseteq S_{l'}$ with $l \ne l'$ are distinct because $S_l \cap S_{l'} = \emptyset$. So every tuple $(a_1,\dots,a_k)$ in the sum consists of k distinct cells, and Theorem 16 gives

$$I_k^w\big(v_\pi;\{a_1,\dots,a_k\}\big) = \int_0^1 D^k f(t\mathbf{1})[p_{a_1},\dots,p_{a_k}]\,dw(t) + e_{a_1\cdots a_k},$$
$$|e_{a_1\cdots a_k}| \le \|w\|_{\mathrm{TV}}\rho(\delta)\prod_l |p_{a_l}|_1.$$

Write $\sum_{a_1,\dots,a_k}$ for the sum over cells $B_{a_l} \subseteq S_l$, $l = 1,\dots,k$. Pulling each sum into its slot, by linearity of the integral and multilinearity of $D^k f(t\mathbf{1})[\cdot,\dots,\cdot]$,

$$\sum_{a_1,\dots,a_k}\int_0^1 D^k f(t\mathbf{1})[p_{a_1},\dots,p_{a_k}]\,dw(t) = \int_0^1 D^k f(t\mathbf{1})\Big[\sum_{B_{a_1}\subseteq S_1} p_{a_1},\dots,\sum_{B_{a_k}\subseteq S_k} p_{a_k}\Big]dw(t)$$
$$= \int_0^1 D^k f(t\mathbf{1})\big[\mu(S_1),\dots,\mu(S_k)\big]\,dw(t),$$

where the second equality holds because S_l is the disjoint union of its cells, so that finite additivity of each μ_i gives $\sum_{B_a\subseteq S_l} p_a = \mu(S_l)$.

The error terms satisfy

$$\sum_{B_{a_1}\subseteq S_1}\cdots\sum_{B_{a_k}\subseteq S_k}|e_{a_1\cdots a_k}|\le\|w\|_{\mathrm{TV}}\rho(\delta)\prod_{l=1}^{k}\Big(\sum_{B_a\subseteq S_l}|p_a|_1\Big)\le\|w\|_{\mathrm{TV}}\rho(\delta)\,m^k,$$

because $\sum_{B_a\subseteq S_l}|p_a|_1\le\sum_{a\in[n]}|p_a|_1=\sum_{i=1}^m\sum_{a\in[n]}\mu_i(B_a)=\sum_{i=1}^m\mu_i(\Omega)=m$. The triangle inequality combines the two displays into the claim. □

Definition 18 (Asymptotic interactions)**.** Let v be a game on $(\Omega,\mathcal{C})$ and $S_1,\dots,S_k\in\mathcal{C}$ pairwise disjoint. Relative to a representation (f,μ), a sequence of partitions $(\pi_r)_{r\ge0}$ is *admissible for* $(S_1,\dots,S_k)$ if every S_l is a union of cells of every π_r, π_{r+1} refines π_r, and $\mathrm{mesh}_\mu(\pi_r)\to0$. The game *admits the asymptotic w-interaction of order k at* $(S_1,\dots,S_k)$ if

$$\mathcal{I}_k^w(v)(S_1,\dots,S_k):=\lim_{r\to\infty}\sum_{B_{a_1}\subseteq S_1}\cdots\sum_{B_{a_k}\subseteq S_k}I_k^w\big(v_{\pi_r};\{a_1,\dots,a_k\}\big)\tag{20}$$

exists and takes the same value along every admissible sequence of every C^k representation of v.

Theorem 19 (Existence, formula, representation independence)**.** *Let w be a finite signed Borel measure on $[0,1]$. Every smooth game v of class C^k admits asymptotic w-interactions of all orders up to k, and for every representation (f,μ),*

$$\mathcal{I}_k^w(v)(S_1,\dots,S_k)=\int_0^1 D^kf(t\mathbf{1})\big[\mu(S_1),\dots,\mu(S_k)\big]\,dw(t).$$

In particular the right-hand side takes the same value for every C^k representation of v.

Proof. Write A_r for the aggregated sum in Eq. (20) at the partition π_r, and $J(f,\mu)$ for the integral in the statement.

Admissible sequences exist: Lemma 13(ii) with $\nu=\mu$ and the collection $S_1,\dots,S_k$ gives partitions $\pi_0\preceq\pi_1\preceq\cdots$ in which every S_l is a union of cells and $\mathrm{mesh}_\mu(\pi_j)\le2^{-j}$, since $\mu_i(\Omega)=1$. This sequence is admissible for (f,μ).

Along every admissible sequence the aggregated sums converge. Let (π_r) be admissible for (f,μ) and $\delta_r:=\mathrm{mesh}_\mu(\pi_r)$. Corollary 17 applies to every π_r and gives $|A_r-J(f,\mu)|\le\|w\|_{\mathrm{TV}}m^k\rho(\delta_r)$. As $\delta_r\to0$ and $\rho(\delta)\to0$ for $\delta\to0$ (Theorem 16), $A_r\to J(f,\mu)$.

Two representations give the same value. Let (f,μ) and (g,ν) both represent v, with $f\in C^k([0,1]^m)$ and $g\in C^k([0,1]^{m'})$. Lemma 13(ii), applied to the $m+m'$ non-atomic probability measures $\mu_1,\dots,\mu_m,\nu_1,\dots,\nu_{m'}$ and the collection $S_1,\dots,S_k$, gives one sequence (π_r) with $\mathrm{mesh}_\mu(\pi_r)\le2^{-r}$ and $\mathrm{mesh}_\nu(\pi_r)\le2^{-r}$; it is admissible for both representations. The restricted game $v_{\pi_r}(A)=v(\bigcup_{c\in A}B_c)$ depends only on v and π_r, so A_r is one and the same real number for both representations. By the previous paragraph, $A_r\to J(f,\mu)$ and $A_r\to J(g,\nu)$; limits of real sequences are unique, so $J(f,\mu)=J(g,\nu)$.

So the limit in Eq. (20) exists along every admissible sequence of every C^k representation and equals that representation's J, and all these values coincide. This is the definition of admitting the asymptotic w-interaction, with the stated formula. □

Proposition 20 (Density form)**.** *Let $v=f\circ\mu$ be a smooth game of class C^k and w a finite signed Borel measure on $[0,1]$. Let Ω carry a metric whose Borel σ-algebra is $\mathcal{C}$ and a finite Borel measure λ, and let $\mu_i=\tilde\eta_i\lambda$ with bounded continuous $\tilde\eta_i$. Then for all $S_1,\dots,S_k\in\mathcal{C}$,*

$$\int_0^1 D^kf(t\mathbf{1})\big[\mu(S_1),\dots,\mu(S_k)\big]dw(t)=\int_{S_1}\cdots\int_{S_k}\iota_k^w(x_1,\dots,x_k)\,d\lambda(x_1)\cdots d\lambda(x_k),$$

$$\iota_k^w(x_1,\dots,x_k):=\int_0^1 D^kf(t\mathbf{1})\big[\tilde\eta(x_1),\dots,\tilde\eta(x_k)\big]dw(t),$$

with ι_k^w continuous on Ω^k. Moreover, along partitions π with $\mathrm{mesh}_\mu(\pi)\to0$ and cells $B_{a_l}\ni x_l$ with $\lambda(B_{a_l})>0$ and $\mathrm{diam}(B_{a_l})\to0$, for distinct $x_1,\dots,x_k$,

$$\frac{I_k^w\big(v_\pi;\{a_1,\dots,a_k\}\big)}{\prod_l\lambda(B_{a_l})}\longrightarrow\iota_k^w(x_1,\dots,x_k).$$

Proof. Throughout, $L := \sup_x \|D^k f(x)\|_\infty$ and $C := \sum_i \sup_\Omega \tilde{\eta}_i < \infty$, so that $|\tilde{\eta}(x)|_1 \le C$ for all x.

Fix t. By $\mu_i(S_l) = \int_{S_l} \tilde{\eta}_i \, d\lambda$ and multilinearity,

$$D^k f(t\mathbf{1})\big[\mu(S_1), \dots, \mu(S_k)\big] = \sum_{i_1,\dots,i_k} D^k f(t\mathbf{1})_{i_1 \cdots i_k} \prod_{l=1}^{k} \int_{S_l} \tilde{\eta}_{i_l}(x_l) \, d\lambda(x_l).$$

A product of k integrals of bounded functions over the finite measure λ is the integral of the product over $S_1 \times \cdots \times S_k$ with respect to $\lambda^{\otimes k}$ (Fubini). Pulling the finite sum inside the integral and recognizing the multilinear form,

$$D^k f(t\mathbf{1})\big[\mu(S_1), \dots, \mu(S_k)\big] = \int_{S_1 \times \cdots \times S_k} D^k f(t\mathbf{1})\big[\tilde{\eta}(x_1), \dots, \tilde{\eta}(x_k)\big] \, d\lambda^{\otimes k}(x).$$

By the first bound of Lemma 14 with $B = 0$, the last integrand is bounded by $L \prod_l |\tilde{\eta}(x_l)|_1 \le LC^k$, and $|w| \otimes \lambda^{\otimes k}$ is a finite measure. Fubini's theorem therefore allows integrating the last display against w and moving the t-integral inside, which gives the first display of the statement with ι_k^w as defined there.

For the continuity of ι_k^w, let $x^{(j)} \to x$ in Ω^k. For every t, the second bound of Lemma 14 gives $|D^k f(t\mathbf{1})[\tilde{\eta}(x_1^{(j)}), \dots] - D^k f(t\mathbf{1})[\tilde{\eta}(x_1), \dots]| \le LC^{k-1} \sum_l |\tilde{\eta}(x_l^{(j)}) - \tilde{\eta}(x_l)|_1$, which tends to 0 by continuity of $\tilde{\eta}$. The integrands are bounded by LC^k and $|w|$ is finite, so dominated convergence gives $\iota_k^w(x^{(j)}) \to \iota_k^w(x)$.

For the pointwise limit, let $\delta := \mathrm{mesh}_\mu(\pi) \to 0$ along the partitions considered.

For $k = 1$ there is one cell. For $k \ge 2$ the cells are eventually distinct: let $d_{\min} := \min_{l \ne l'} d(x_l, x_{l'}) > 0$. Once all diameters are below $d_{\min}$, the cell containing x_l cannot contain $x_{l'}$ for $l' \ne l$, so $B_{a_1}, \dots, B_{a_k}$ are distinct and Theorem 16 applies.

The normalized cell masses converge to the densities. Put $u_l := p_{a_l}/\lambda(B_{a_l})$, with coordinates $u_{l,i} = \mu_i(B_{a_l})/\lambda(B_{a_l}) = \frac{1}{\lambda(B_{a_l})} \int_{B_{a_l}} \tilde{\eta}_i \, d\lambda$. Then

$$|u_{l,i} - \tilde{\eta}_i(x_l)| = \Big| \frac{1}{\lambda(B_{a_l})} \int_{B_{a_l}} \big(\tilde{\eta}_i(x) - \tilde{\eta}_i(x_l)\big) \, d\lambda(x) \Big| \le \sup_{x \in B_{a_l}} |\tilde{\eta}_i(x) - \tilde{\eta}_i(x_l)|,$$

which tends to 0 as $\mathrm{diam}(B_{a_l}) \to 0$ by continuity of $\tilde{\eta}_i$ at x_l. Hence $u_l \to \tilde{\eta}(x_l)$, and $|u_l|_1 = \sum_i u_{l,i} \le C$ for every partition.

Dividing the estimate of Theorem 16 by $\prod_l \lambda(B_{a_l}) > 0$ and using multilinearity in the main term,

$$\begin{aligned} \Big| \frac{I_k^w(v_\pi; \{a_1, \dots, a_k\})}{\prod_l \lambda(B_{a_l})} - \int_0^1 D^k f(t\mathbf{1})[u_1, \dots, u_k] \, dw(t) \Big| &\le \|w\|_{\mathrm{TV}} \, \rho(\delta) \prod_l |u_l|_1 \\ &\le \|w\|_{\mathrm{TV}} \, C^k \rho(\delta) \xrightarrow[\delta \to 0]{} 0. \end{aligned}$$

The main term converges to $\iota_k^w(x_1, \dots, x_k)$: for every t, the second bound of Lemma 14 with all vectors bounded by C in $|\cdot|_1$ gives $|D^k f(t\mathbf{1})[u_1, \dots, u_k] - D^k f(t\mathbf{1})[\tilde{\eta}(x_1), \dots, \tilde{\eta}(x_k)]| \le LC^{k-1} \sum_l |u_l - \tilde{\eta}(x_l)|_1$, a bound that does not depend on t. Integrating against w as in the proof of Theorem 16,

$$\Big| \int_0^1 D^k f(t\mathbf{1})[u_1, \dots, u_k] \, dw(t) - \iota_k^w(x_1, \dots, x_k) \Big| \le \|w\|_{\mathrm{TV}} \, LC^{k-1} \sum_l |u_l - \tilde{\eta}(x_l)|_1 \longrightarrow 0$$

by the convergence $u_l \to \tilde{\eta}(x_l)$. By the triangle inequality, the distance between $I_k^w(v_\pi; \{a_1, \dots, a_k\}) / \prod_l \lambda(B_{a_l})$ and $\iota_k^w(x_1, \dots, x_k)$ is at most the sum of the last two bounds, which tends to 0. □

Theorem 7 (FSII fields). *Let $\ell \ge 1$ and let $v = f \circ \mu$ be a smooth game of class C^ℓ with $\mu \in (\mathrm{NA}^1)^m$. For $k \le \ell$ let $L_k := \sup_x \|D^k f(x)\|_\infty$, let $\omega_k(r) := \sup\{\|D^k f(x) - D^k f(y)\|_\infty :$*

$|x-y|_\infty \le r\}$ *be the modulus of continuity of* $D^k f$*, let* $\|w\|_{\mathrm{TV}}$ *be the total variation of a signed measure* w[2]*, and*

$$\rho_k(\delta) := \omega_k\big(2k\delta + \delta^{1/3}\big) + \frac{mL_k}{2}\,\delta^{1/3} \xrightarrow[\delta\to 0]{} 0, \qquad \rho_k(\delta) = O(\delta^{1/3}) \text{ if } D^k f \text{ is Lipschitz.} \tag{7}$$

Then:

(i) For every partition π *with at least* $\max\{2,\ell\}$ *cells and* $\mathrm{mesh}_\mu(\pi)=\delta$*, and all distinct cells* $a_1,\dots,a_k$*,*

$$\Big|\mathcal{E}_{\{a_1,\dots,a_k\}}(v_\pi,\ell) - \int_0^1 D^k f(t\mathbf{1})[p_{a_1},\dots,p_{a_k}]\, w_k^{(\ell)}(t)\,dt\Big| \le \tag{8}$$

$$\le \|w_k^{(\ell)}\|_{\mathrm{TV}}\Big(\prod_{l=1}^{k}|p_{a_l}|_1\Big)\rho_k(\delta).$$

(ii) Let $S_1,\dots,S_k \in \mathcal{C}$ *be pairwise disjoint and let* $(\pi_r)_{r\ge 0}$ *be partitions, each refining the previous, in which every* S_l *is a union of cells and* $\mathrm{mesh}_\mu(\pi_r)\to 0$*. Then,*

$$\sum_{B_{a_1}\subseteq S_1}\cdots\sum_{B_{a_k}\subseteq S_k}\mathcal{E}_{\{a_1,\dots,a_k\}}(v_{\pi_r},\ell) \;\longrightarrow\; \mathcal{I}_k^{(\ell)}(S_1,\dots,S_k),$$
$$\mathcal{I}_k^{(\ell)}(S_1,\dots,S_k) := \int_0^1 D^k f(t\mathbf{1})\big[\mu(S_1),\dots,\mu(S_k)\big]\, w_k^{(\ell)}(t)\,dt, \tag{9}$$

and the limit depends neither on the sequence nor on the C^ℓ *representation* (f,μ) *of* v*.*

(iii) Let Ω *carry a metric whose Borel* σ*-algebra is* $\mathcal{C}$*, let* λ *be a finite Borel measure on* Ω*, and let* $\mu_i = \tilde\eta_i\lambda$ *for some bounded continuous densities* $\tilde\eta_i$*. Then Eq.* (9) *has the continuous density*

$$\iota_k^{(\ell)}(x_1,\dots,x_k) = \int_0^1 D^k f(t\mathbf{1})\big[\tilde\eta(x_1),\dots,\tilde\eta(x_k)\big]\, w_k^{(\ell)}(t)\,dt, \qquad \tilde\eta := (\tilde\eta_1,\dots,\tilde\eta_m), \tag{10}$$

with respect to $\lambda^{\otimes k}$ *on* $S_1\times\cdots\times S_k$*, and along partitions with* $\mathrm{mesh}_\mu(\pi)\to 0$ *and cells* $B_{a_l}\ni x_l$ *of positive* λ*-mass and vanishing diameter, for distinct* $x_1,\dots,x_k$*,*

$$\frac{\mathcal{E}_{\{a_1,\dots,a_k\}}(v_\pi,\ell)}{\prod_{l=1}^k\lambda(B_{a_l})} \longrightarrow \iota_k^{(\ell)}(x_1,\dots,x_k). \tag{11}$$

Proof. A game of class C^ℓ is of class C^k for every $k\le\ell$. For a partition π with $n\ge\max\{2,\ell\}$ cells, Theorem 5 applies to the finite game v_π and gives $\mathcal{E}_{\{a_1,\dots,a_k\}}(v_\pi,\ell) = I_k^{w_k^{(\ell)}}(v_\pi;\{a_1,\dots,a_k\})$, where $w_k^{(\ell)}(u)\,du$ is a finite signed Borel measure on $[0,1]$ because $w_k^{(\ell)}$ is a polynomial. Along partitions with $\mathrm{mesh}_\mu(\pi)\to 0$ the number of cells satisfies $n\ge 1/\mathrm{mesh}_\mu(\pi)$, because the n numbers $\mu_1(B_c)\le\mathrm{mesh}_\mu(\pi)$ sum to 1; hence $n\ge\max\{2,\ell\}$ for all but finitely many partitions of such a sequence, and Theorem 5 applies to the rest.

With $w = w_k^{(\ell)}\,du$: (i) is Theorem 16, whose ρ is ρ_k of Eq. (7). In (ii) the sequence (π_r) is admissible for $(S_1,\dots,S_k)$ in the sense of Definition 18, so the convergence, the formula Eq. (9) and the independence of sequence and representation are Theorem 19. (iii) is Proposition 20: its first display is the density statement, and its last display combined with Theorem 5 is Eq. (11). □

A.3 Efficiency

Theorem 21 (Interaction efficiency)**.** *For every smooth game* $v = f\circ\mu$ *of class* C^ℓ*,*

$$\sum_{k=1}^{\ell}\frac{1}{k!}\int_0^1 D^k f(t\mathbf{1})[\mathbf{1},\dots,\mathbf{1}]\, w_k^{(\ell)}(t)\,dt = v(\Omega)-v(\emptyset), \tag{21}$$

[2] $\|w\|_{\mathrm{TV}} := |w|([0,1])$ with $|w| := w^+ + w^-$ for the Jordan decomposition $w = w^+ - w^-$ into non-negative measures.

and under the assumptions of Theorem 7(iii),

$$\sum_{k=1}^{\ell} \frac{1}{k!} \int_{\Omega^k} \iota_k^{(\ell)} \, d\lambda^{\otimes k} = v(\Omega) - v(\emptyset). \tag{22}$$

Proof sketch. The FSII satisfies the efficiency constraint of Definition 1 at every partition. Along a refining sequence, Theorem 7(i) and $\sum_a p_a = \mathbf{1}$ turn the order-k sum over all cells into the k-th term of Eq. (21), while the right-hand side stays $v(\Omega) - v(\emptyset)$. The full proof is in Section A.7. □

In Theorem 16 we saw that the continuum limit of our FSII can be represented by learnable densities, one for each degree of interaction. For example, the 2-interaction between regions B_1 and B_2 can be calculated by integrating $\int_{B_1 \times B_2} \iota_2^{(\ell)} \, d\lambda^{\otimes 2}$. Therefore also, when the cells are infinitesimal the 2-interaction of a single cell with itself vanishes, because its integration region $B \times B$ shrinks to the diagonal $\{(x, x)\}$, a nullset in the plane. A region of positive size can still interact with itself. However, our data resides on a finite dimensional grid and while these self-interactions are a nullset in the limit, they may contribute more than zero in the finite case. The following theorem shows us how we can mitigate this issue in the algorithm, by adding higher order self-interactions to the lower-order interactions.

Proposition 22 (Cell coefficients of continuum heads)**.** *Let λ be a finite measure on $(\Omega, \mathcal{C})$, $\pi = \{B_1, \ldots, B_n\}$ a partition, and for $k = 1, \ldots, \ell$ let $h_k : \Omega^k \to \mathbb{R}$ be bounded, measurable and symmetric. For $A \subseteq [n]$ put $S_A := \bigcup_{a \in A} B_a$ and*

$$F(A) := \sum_{k=1}^{\ell} \frac{1}{k!} \int_{S_A^k} h_k \, d\lambda^{\otimes k}. \tag{23}$$

Then $F(A) = \sum_{T \subseteq A,\, 1 \le |T| \le \ell} \hat{E}_T$ for every $A \subseteq [n]$, where, for $T = \{c_1, \ldots, c_r\}$,

$$\hat{E}_T := \sum_{k=r}^{\ell} \sum_{\substack{n_1 + \cdots + n_r = k \\ n_j \ge 1}} \frac{1}{n_1! \cdots n_r!} \int_{B_{c_1}^{n_1} \times \cdots \times B_{c_r}^{n_r}} h_k \, d\lambda^{\otimes k}. \tag{24}$$

In particular $\sum_{1 \le |T| \le \ell} \hat{E}_T = \sum_{k=1}^{\ell} \frac{1}{k!} \int_{\Omega^k} h_k \, d\lambda^{\otimes k}$.

Proof in Section A.8. At $\ell = 2$, Eq. (24) reads $\hat{E}_a = \int_{B_a} h_1 \, d\lambda + \frac{1}{2} \int_{B_a \times B_a} h_2 \, d\lambda^{\otimes 2}$ and $\hat{E}_{ab} = \int_{B_a \times B_b} h_2 \, d\lambda^{\otimes 2}$. If h_k are estimates of $\iota_k^{(\ell)}$ with $\sum_k \frac{1}{k!} \int_{\Omega^k} h_k \, d\lambda^{\otimes k} = v(\Omega) - v(\emptyset)$, then the coefficients Eq. (24) satisfy finite-game efficiency exactly at every partition.

A.4 Function-space setup

We use the flat torus $\mathbb{T}^d$ as a domain, which can be visualized as a donut shape. A donut has no boundaries, so we can get around the technicalities of boundary conditions for PDEs. The results also hold on a bounded smooth domain Ω. Stiller et al. (2026) show this for FNOs by extending the input from Ω to the torus, applying the FNO, and restricting the output back to Ω. Our proofs only need that λ is non-atomic, that the product of a smooth mask and a function in $H^s(\Omega)$ stays in $H^s(\Omega)$, and the chain rule. All three hold on Ω. The boundary conditions of the PDE do not enter, because the game evaluates the backbone and not the PDE.

Neural operators take functions as inputs and return functions as outputs. We therefore need a space that describes which functions are valid inputs and in which sense two inputs are close. We use the Sobolev space $H^s(\mathbb{T}^d)$. Informally, H^s contains functions whose derivatives up to order s are square-integrable. The H^s norm measures both the size of a function and the size of these derivatives. The Sobolev space contains also some functions that are not differentiable on few points of the space. Usually in practice, we do not care about single points on a domain, as long as the function is well-behaved on the rest of the domain.

The condition $s > d/2$ gives us two properties used in our construction. First, functions in $H^s(\mathbb{T}^d)$ can be represented with a continuous function, so values such as $u(x_0)$ are well-defined. Second,

products of functions in $H^s(\mathbb{T}^d)$ remain in $H^s(\mathbb{T}^d)$. Consequently, multiplying a perturbation h by a smooth mask η_i produces another valid input. This ensures that all masked inputs considered by the operator game remain in the same function space.

Finally, Fréchet differentiability is the extension of ordinary differentiability to maps between function spaces. A Fréchet derivative gives a bounded linear approximation of how the output changes under a small perturbation of the input. Higher derivatives describe how several perturbations interact. Since we consider interactions up to order ℓ, we require the scalar-valued operator $\hat{\mathcal{F}}_x$ to be Fréchet-C^ℓ.

A.5 Operator games

Lemma 23 (Operator games are smooth games)**.** *For the operator game in Definition 3, $(\Omega, \mathcal{B}(\mathbb{T}^d))$ is a player space, each μ_i is in NA^1 with bounded continuous density $\tilde{\eta}_i$ with respect to λ, and the operator game is $v = f \circ \mu$ with $f = \hat{\mathcal{F}}_x \circ \Lambda - \hat{\mathcal{F}}_x(\bar{u}) \in C^\ell([0,1]^m)$ for $\Lambda(a) := \bar{u} + \sum_{i=1}^m a_i \eta_i h$, $f(0) = 0$, and $\partial^k_{i_1 \cdots i_k} f(a) = D^k \hat{\mathcal{F}}_x(\Lambda(a))[\eta_{i_1} h, \ldots, \eta_{i_k} h]$ for $k \le \ell$.*

Proof. $\mathbb{T}^d$ is an uncountable complete separable metric space, so by the Borel isomorphism theorem (Kechris, 1995, Thm. 15.6) $(\mathbb{T}^d, \mathcal{B}(\mathbb{T}^d))$ is isomorphic to $([0,1], \mathcal{B}([0,1]))$.

$\tilde{\eta}_i = \eta_i / \int_\Omega \eta_i \, d\lambda$ is continuous, bounded and non-negative with $\int_\Omega \tilde{\eta}_i \, d\lambda = 1$, so $\mu_i = \tilde{\eta}_i \lambda$ is a probability measure. For non-atomicity let $\mu_i(A) > 0$; identify $\mathbb{T} = [0,1)$ and put $\phi(s) := \mu_i\big(A \cap ([0,s] \times \mathbb{T}^{d-1})\big)$. Then $\phi(0) = 0$ because $\{0\} \times \mathbb{T}^{d-1}$ is λ-null, $\phi(1) = \mu_i(A)$, and for $s < s'$ $0 \le \phi(s') - \phi(s) = \mu_i\big(A \cap ((s,s'] \times \mathbb{T}^{d-1})\big) \le \sup \tilde{\eta}_i \cdot \lambda\big((s,s'] \times \mathbb{T}^{d-1}\big) = \sup \tilde{\eta}_i \cdot (s' - s)$, so ϕ is continuous. The intermediate value theorem gives s^* with $\phi(s^*) = \mu_i(A)/2$, and $B := A \cap ([0,s^*] \times \mathbb{T}^{d-1}) \subseteq A$ has $0 < \mu_i(B) < \mu_i(A)$.

By the definition of the operator game and of Λ, $v(S) = \hat{\mathcal{F}}_x(\Lambda(\mu(S))) - \hat{\mathcal{F}}_x(\bar{u}) = f(\mu(S))$, and $f(0) = \hat{\mathcal{F}}_x(\bar{u}) - \hat{\mathcal{F}}_x(\bar{u}) = 0$. f is defined on all of $\mathbb{R}^m$, in particular on an open neighbourhood of $[0,1]^m$ as Section 3.3 requires.

$\eta_i h = \Lambda(e_i) - \Lambda(0) \in H^s(\Omega)$, so $\Lambda_0 b := \sum_i b_i \eta_i h$ is a linear map $\mathbb{R}^m \to H^s(\Omega)$. It is bounded because its domain is finite-dimensional: $\|\Lambda_0 b\|_{H^s} \le \sum_i |b_i| \, \|\eta_i h\|_{H^s} \le \big(\max_i \|\eta_i h\|_{H^s}\big) |b|_1$. Boundedness is needed because a Fréchet derivative between normed spaces is by definition a bounded linear map, and the chain rule $D(\hat{\mathcal{F}}_x \circ \Lambda)(a) = D\hat{\mathcal{F}}_x(\Lambda(a)) \circ \Lambda_0$ composes $D\hat{\mathcal{F}}_x(\Lambda(a))$, which is continuous only for the norm of H^s, with Λ_0. So $\Lambda = \Lambda(0) + \Lambda_0$ is affine with constant Fréchet derivative Λ_0. Since $\hat{\mathcal{F}}_x$ is C^ℓ on $H^s(\Omega)$, the chain rule in Banach spaces gives that $f = \hat{\mathcal{F}}_x \circ \Lambda - \hat{\mathcal{F}}_x(\bar{u})$ is C^ℓ with $D^k f(a)[b_1, \ldots, b_k] = D^k \hat{\mathcal{F}}_x(\Lambda(a))[\Lambda_0 b_1, \ldots, \Lambda_0 b_k]$ for $k \le \ell$; the higher derivatives of Λ vanish, so no further terms appear. At $b_l = e_{i_l}$, where $\Lambda_0 e_i = \eta_i h$, this is the formula for $\partial^k_{i_1 \cdots i_k} f$. □

Proposition 24 (Fields of operator games)**.** *For the operator game in Definition 3, Theorem 7 applies with $\tilde{\eta}_i = \eta_i / \int_\Omega \eta_i \, d\lambda$, and its fields are*

$$\iota^{(\ell)}_k(x_1, \ldots, x_k) = \sum_{i_1, \ldots, i_k = 1}^m \tilde{\eta}_{i_1}(x_1) \cdots \tilde{\eta}_{i_k}(x_k) \int_0^1 D^k \hat{\mathcal{F}}_x(\bar{u} + th)\big[\eta_{i_1} h, \ldots, \eta_{i_k} h\big] \, w^{(\ell)}_k(t) \, dt. \tag{25}$$

Proof. By Lemma 23 the operator game satisfies the hypotheses of Theorem 7, including (iii): it is a smooth game of class C^ℓ on the player space $(\mathbb{T}^d, \mathcal{B}(\mathbb{T}^d))$, and the μ_i have the bounded continuous densities $\tilde{\eta}_i$ with respect to the finite Borel measure λ of the metric space $\mathbb{T}^d$.

Since $\sum_i \eta_i \equiv 1$, $\Lambda(t\mathbf{1}) = \bar{u} + t \sum_i \eta_i h = \bar{u} + th$. So multilinearity and Lemma 23 at $a = t\mathbf{1}$ give

$$\begin{aligned} D^k f(t\mathbf{1})[\tilde{\eta}(x_1), \ldots, \tilde{\eta}(x_k)] &= \sum_{i_1, \ldots, i_k} \partial^k_{i_1 \cdots i_k} f(t\mathbf{1}) \prod_l \tilde{\eta}_{i_l}(x_l) \\ &= \sum_{i_1, \ldots, i_k} \Big(\prod_l \tilde{\eta}_{i_l}(x_l) \Big) D^k \hat{\mathcal{F}}_x(\bar{u} + th)[\eta_{i_1} h, \ldots, \eta_{i_k} h]. \end{aligned}$$

Inserting this into Eq. (10) and moving the finite sum and the t-independent factors $\tilde{\eta}_{i_l}(x_l)$ out of the t-integral gives Eq. (25). □

A.6 SMOOTH FNO READOUTS

Throughout, $H^s := H^s(\mathbb{T}^d)$ with $s > d/2$, and maps acting componentwise on $\mathbb{R}^c$-valued functions are treated one component at a time. For $s > d/2$, H^s is a Banach algebra, i.e., for $g, z \in H^s$ $\|gz\|_{H^s} \le C_s \|g\|_{H^s} \|z\|_{H^s}$, and embeds continuously into $C^0(\mathbb{T}^d)$ (Runst & Sickel, 1996, Thm. 4.6.4/1).

In the standard FNO definition (Li et al., 2021), a Fourier layer combines a pointwise channel matrix W_l with the truncated Fourier multiplier

$$\widehat{K_l v}(k) = \begin{cases} R_{l,k} \hat{v}(k), & k \in \mathcal{K}_l, \\ 0, & k \notin \mathcal{K}_l, \end{cases}$$

where $\mathcal{K}_l \subset \mathbb{Z}^d$ is finite. Hence

$$\|K_l v\|_{H^s} \le \max_{k \in \mathcal{K}_l} \|R_{l,k}\| \, \|v\|_{H^s},$$

so boundedness of the Fourier operator is automatic. The finite channel matrix W_l is likewise bounded on H^s, and the standard channelwise biases are constant functions in H^s. Thus, the affine part of every Fourier layer is a bounded affine map. The remaining ingredient is the smoothness of the pointwise nonlinearities. For $\tau \in C^\infty(\mathbb{R})$ with $\tau(0) = 0$, the superposition map $S_\tau(w) := \tau \circ w$ is Fréchet-C^1 on H^s with $DS_\tau(w)[z] = \tau'(w) z$ (Runst & Sickel, 1996, Thm. 5.5.3/2). Only the higher orders are needed below.

Lemma 25 (Superposition operators on H^s are C^∞)**.** *For $\tau \in C^\infty(\mathbb{R})$, $S_\tau : H^s \to H^s$ is C^ℓ for every ℓ, with $DS_\tau(w)[z] = \tau'(w)\, z$.*

Proof. Induction on ℓ, for all $\tau \in C^\infty(\mathbb{R})$ at once. For $\ell = 1$, $S_\tau = S_{\tau - \tau(0)} + \tau(0) \cdot 1$ with the constant function $1 \in H^s$, and $\tau - \tau(0)$ vanishes at 0, so the C^1 statement above applies. For the step $\ell - 1 \to \ell$, let $M : H^s \to L(H^s, H^s)$, $M(g)z := gz$, which is bounded linear by the Banach algebra property. The derivative formula reads $DS_\tau = M \circ S_{\tau'}$; the induction hypothesis for $\tau' \in C^\infty(\mathbb{R})$ gives $S_{\tau'} \in C^{\ell-1}$, so $DS_\tau \in C^{\ell-1}$, that is, $S_\tau \in C^\ell$. □

Proposition 26 (Smooth FNO readouts are C^∞)**.** *Let $s > d/2$, and let $\hat{\mathcal{F}} : H^s(\mathbb{T}^d; \mathbb{R}^{c_{\text{in}}}) \to H^s(\mathbb{T}^d; \mathbb{R}^{c_{\text{out}}})$ be a standard FNO (Li et al., 2021) whose pointwise nonlinearities, including any used in its lifting and projection maps, are C^∞. Let $R : H^s(\mathbb{T}^d; \mathbb{R}^{c_{\text{out}}}) \to \mathbb{R}$ be Fréchet-C^∞, for instance $R(y) = \int_\Omega y_c^p \, d\lambda$ or $R(y) = y_c(x_0)$ for a channel c, $p \in \mathbb{N}$ and $x_0 \in \Omega$. Then $\hat{\mathcal{F}}_x := R \circ \hat{\mathcal{F}}$ is Fréchet-C^∞ on $H^s(\mathbb{T}^d; \mathbb{R}^{c_{\text{in}}})$, so the conditions of Definition 3 hold for every ℓ.*

Proof. Every Fourier layer is a bounded affine map followed by componentwise maps of Lemma 25, so it is C^ℓ for every ℓ by the chain rule. Standard affine lifting and projection maps are bounded for the same reason as the channel matrices. If they are nonlinear finite-dimensional networks, their affine layers are bounded and their smooth pointwise nonlinearities satisfy Lemma 25. Hence the lifting, Fourier layers and projection are all C^∞, and so are $\hat{\mathcal{F}}$ and $\hat{\mathcal{F}}_x = R \circ \hat{\mathcal{F}}$.

For the two readouts: $y \mapsto y_c(x_0)$ is bounded linear by the embedding $H^s \hookrightarrow C^0(\mathbb{T}^d)$, hence C^∞. The form $(y_1, \dots, y_p) \mapsto \int_\Omega y_1 \cdots y_p \, d\lambda$ is p-linear and bounded, since $|\int_\Omega y_1 \cdots y_p \, d\lambda| \le \lambda(\Omega)^{1/2} \|y_1 \cdots y_p\|_{H^s} \le \lambda(\Omega)^{1/2} C_s^{p-1} \prod_j \|y_j\|_{H^s}$, so $y \mapsto \int_\Omega y_c^p \, d\lambda$ is a continuous homogeneous polynomial on H^s, hence C^∞. So $\hat{\mathcal{F}}_x$ is Fréchet-C^ℓ on $H^s(\mathbb{T}^d; \mathbb{R}^{c_{\text{in}}})$ for every ℓ, as Definition 3 requires. □

Remark 27. The same conclusion holds for other neural operators that are finite compositions of bounded affine maps between the respective H^s spaces and pointwise C^∞ maps. For these architectures, only boundedness of their nonlocal layers has to be checked separately.

A.7 Proof of Theorem 21 (interaction efficiency)

Corollary 28 (Aggregation over all cells)**.** *Let w be a finite signed Borel measure on $[0,1]$, $v = f \circ \mu$ a smooth game of class C^k, π a partition of mesh δ, and $L = \sup_x \|D^k f(x)\|_\infty$. Then, the sum ranging over all k-element subsets of cells,*

$$\Big| \sum_{\{a_1,\dots,a_k\}\subseteq[n]} I_k^w\big(v_\pi;\{a_1,\dots,a_k\}\big) - \frac{1}{k!}\int_0^1 D^k f(t\mathbf{1})[\mathbf{1},\dots,\mathbf{1}]\,dw(t)\Big| \le \frac{\|w\|_{\mathrm{TV}}\, m^k}{k!}\Big(\rho(\delta) + \tbinom{k}{2} L\,\delta\Big).$$

Proof. Write $r_a := |p_a|_1 = \sum_i \mu_i(B_a)$, so that $\sum_a r_a = \sum_i \mu_i(\Omega) = m$ and $r_a \le m\delta$ by the definition of the mesh, and $J(a_1,\dots,a_k) := \int_0^1 D^k f(t\mathbf{1})[p_{a_1},\dots,p_{a_k}]\,dw(t)$ for ordered tuples of cells, distinct or not.

$I_k^w(v_\pi;\{a_1,\dots,a_k\})$ depends only on the set $\{a_1,\dots,a_k\}$, and every k-element set has $k!$ orderings, so the sum over subsets is $\frac{1}{k!}$ times the sum over ordered tuples of distinct cells. Summing the estimate of Theorem 16 over these,

$$\Big|\sum_{\text{distinct}} I_k^w\big(v_\pi;\{a_1,\dots,a_k\}\big) - \sum_{\text{distinct}} J(a_1,\dots,a_k)\Big| \le \|w\|_{\mathrm{TV}}\rho(\delta)\sum_{\text{distinct}}\prod_l r_{a_l} \le \|w\|_{\mathrm{TV}}\rho(\delta)\Big(\sum_a r_a\Big)^k = \|w\|_{\mathrm{TV}}\rho(\delta)\, m^k.$$

By multilinearity and $\sum_a p_a = \mathbf{1}$, the sum of J over all ordered tuples, repetitions allowed, is $\int_0^1 D^k f(t\mathbf{1})[\sum_a p_a,\dots,\sum_a p_a]\,dw(t) = \int_0^1 D^k f(t\mathbf{1})[\mathbf{1},\dots,\mathbf{1}]\,dw(t)$. Hence $\sum_{\text{distinct}} J = \int_0^1 D^k f(t\mathbf{1})[\mathbf{1},\dots,\mathbf{1}]\,dw(t) - \sum^* J$, where $\sum^*$ runs over the tuples with a repeated cell. For $k=1$ no tuple has a repeated cell, so $\sum^* J = 0$ and the claim follows from the first bound; assume $k \ge 2$. For every tuple, the first bound of Lemma 14 with $B=0$ gives $|J(a_1,\dots,a_k)| \le \|w\|_{\mathrm{TV}} L \prod_l r_{a_l}$. A tuple with a repetition has a pair of slots $l < l'$ with $a_l = a_{l'}$; summing the non-negative products first over the tuples with $a_l = a_{l'}$ and then over the $\binom{k}{2}$ pairs of slots, which counts tuples with several repetitions more than once and only increases the bound,

$$\sum^*\prod_l r_{a_l} \le \binom{k}{2}\Big(\sum_a r_a^2\Big)\Big(\sum_a r_a\Big)^{k-2} \le \binom{k}{2}\big(\max_a r_a\big)\Big(\sum_a r_a\Big)^{k-1} \le \binom{k}{2} m^k\delta,$$

so $|\sum^* J| \le \|w\|_{\mathrm{TV}} L\binom{k}{2} m^k\delta$. The triangle inequality applied to the two bounds gives $\big|\sum_{\text{distinct}} I_k^w(v_\pi;\{a_1,\dots,a_k\}) - \int_0^1 D^k f(t\mathbf{1})[\mathbf{1},\dots,\mathbf{1}]\,dw(t)\big| \le \|w\|_{\mathrm{TV}} m^k\big(\rho(\delta) + \binom{k}{2}L\delta\big)$, and dividing by $k!$ gives the claim. □

Theorem 21 (Interaction efficiency)**.** *For every smooth game $v = f \circ \mu$ of class C^ℓ,*

$$\sum_{k=1}^{\ell}\frac{1}{k!}\int_0^1 D^k f(t\mathbf{1})[\mathbf{1},\dots,\mathbf{1}]\, w_k^{(\ell)}(t)\,dt = v(\Omega) - v(\emptyset), \tag{21}$$

and under the assumptions of Theorem 7(iii),

$$\sum_{k=1}^{\ell}\frac{1}{k!}\int_{\Omega^k} \iota_k^{(\ell)}\, d\lambda^{\otimes k} = v(\Omega) - v(\emptyset). \tag{22}$$

Proof. For Eq. (21), Lemma 13(ii) with $\nu = \mu$ and the single set $S_1 = \Omega$ gives a refining sequence (π_r) with $\delta_r := \mathrm{mesh}_\mu(\pi_r) \le 2^{-r}$. As in the proof of Theorem 7, the number n_r of cells satisfies $n_r \ge 1/\delta_r$, so $n_r \ge \max\{2,\ell\}$ for all large r, which we assume. By Section 3.3, $v_{\pi_r}([n_r]) = v(\Omega)$ and $v_{\pi_r}(\emptyset) = v(\emptyset)$. The two constraints in Definition 1, with the sum split by $|T| = k$, therefore give

$$\sum_{k=1}^{\ell}\sum_{T\subseteq[n_r],\,|T|=k}\mathcal{E}_T(v_{\pi_r},\ell) = v(\Omega) - v(\emptyset).$$

By Theorem 5, the inner sum is $\sum_{|T|=k} I_k^{w_k^{(\ell)}}(v_{\pi_r}; T)$, and Corollary 28 with $w = w_k^{(\ell)}\,du$ bounds its distance to $\frac{1}{k!}\int_0^1 D^k f(t\mathbf{1})[\mathbf{1},\dots,\mathbf{1}]\, w_k^{(\ell)}(t)\,dt$ by $\frac{\|w_k^{(\ell)}\|_{\mathrm{TV}} m^k}{k!}\big(\rho_k(\delta_r) + \binom{k}{2} L_k \delta_r\big)$, which tends to 0 as $r \to \infty$. Letting $r \to \infty$ in the display, whose right side does not depend on r, gives Eq. (21).

For Eq. (22), Proposition 20 with $S_1 = \cdots = S_k = \Omega$ and $w = w_k^{(\ell)}\,du$ gives $\int_{\Omega^k} \iota_k^{(\ell)}\,d\lambda^{\otimes k} = \int_0^1 D^k f(t\mathbf{1})[\mu(\Omega),\dots,\mu(\Omega)]\, w_k^{(\ell)}(t)\,dt$, and $\mu(\Omega) = \mathbf{1}$ because each μ_i is a probability measure. So the k-th term of Eq. (22) is the k-th term of Eq. (21). □

A.8 Proof of Proposition 22 (cell coefficients)

Proposition 22 (Cell coefficients of continuum heads)**.** *Let λ be a finite measure on $(\Omega, \mathcal{C})$, $\pi = \{B_1,\dots,B_n\}$ a partition, and for $k = 1,\dots,\ell$ let $h_k : \Omega^k \to \mathbb{R}$ be bounded, measurable and symmetric. For $A \subseteq [n]$ put $S_A := \bigcup_{a\in A} B_a$ and*

$$F(A) := \sum_{k=1}^{\ell} \frac{1}{k!} \int_{S_A^k} h_k\,d\lambda^{\otimes k}. \tag{23}$$

Then $F(A) = \sum_{T\subseteq A,\, 1\le|T|\le\ell} \hat{E}_T$ for every $A \subseteq [n]$, where, for $T = \{c_1,\dots,c_r\}$,

$$\hat{E}_T := \sum_{k=r}^{\ell} \sum_{\substack{n_1+\cdots+n_r=k \\ n_j\ge 1}} \frac{1}{n_1!\cdots n_r!} \int_{B_{c_1}^{n_1}\times\cdots\times B_{c_r}^{n_r}} h_k\,d\lambda^{\otimes k}. \tag{24}$$

In particular $\sum_{1\le|T|\le\ell} \hat{E}_T = \sum_{k=1}^{\ell} \frac{1}{k!}\int_{\Omega^k} h_k\,d\lambda^{\otimes k}$.

Proof. Fix $A \subseteq [n]$ and $k \le \ell$. The cells are pairwise disjoint and cover S_A, so S_A^k is the disjoint union of the boxes $B_{a_1}\times\cdots\times B_{a_k}$ over the ordered tuples $(a_1,\dots,a_k) \in A^k$, and $\int_{S_A^k} h_k\,d\lambda^{\otimes k} = \sum_{(a_1,\dots,a_k)\in A^k} \int_{B_{a_1}\times\cdots\times B_{a_k}} h_k\,d\lambda^{\otimes k}$.

Group the tuples by the set $T := \{a_1,\dots,a_k\} \subseteq A$ of cells they use, of size $r \le k$. Fix such a set $T = \{c_1,\dots,c_r\}$ with $c_1 < \cdots < c_r$. A tuple with this set is the same as a surjection $\sigma : [k] \to T$, $\sigma(l) := a_l$, and its fibres $\sigma^{-1}(c_j) = \{l : a_l = c_j\}$ have sizes $n_j \ge 1$ with $\sum_j n_j = k$. The permutation of the coordinates of Ω^k that sorts the slots by their cell maps $B_{a_1}\times\cdots\times B_{a_k}$ onto $B_{c_1}^{n_1}\times\cdots\times B_{c_r}^{n_r}$, preserves $\lambda^{\otimes k}$ by Fubini's theorem, and leaves h_k unchanged by symmetry; so both boxes carry the same integral of h_k. The number of surjections with given fibre sizes $(n_1,\dots,n_r)$ is the multinomial coefficient $k!/(n_1!\cdots n_r!)$. Hence

$$\frac{1}{k!}\int_{S_A^k} h_k\,d\lambda^{\otimes k} = \sum_{T\subseteq A,\,1\le|T|\le k}\ \sum_{\substack{n_1+\cdots+n_r=k \\ n_j\ge 1}} \frac{1}{n_1!\cdots n_r!} \int_{B_{c_1}^{n_1}\times\cdots\times B_{c_r}^{n_r}} h_k\,d\lambda^{\otimes k}.$$

Summing over $k = 1,\dots,\ell$ and collecting, for each T with $1 \le |T| = r \le \ell$, the terms with $k \ge r$ gives $F(A) = \sum_{T\subseteq A,\,1\le|T|\le\ell} \hat{E}_T$ with $\hat{E}_T$ as in Eq. (24). The last claim is the case $A = [n]$, where $S_A = \Omega$.

□

A.9 The order-two case

Corollary 29 (Order two)**.** *Let $\ell = 2$ and let $v = f \circ \mu$ be a smooth game of class C^2 whose μ_i have bounded continuous densities $\tilde{\eta}_i$ with respect to λ as in Theorem 7(iii).*

(i) For every game v' on $[n]$, $n \ge 2$: $\mathcal{E}_i(v', 2) = \int_0^1 \mathbb{E}[(\Delta_i v')(S_t)]\,(4-6t)\,dt$ and $\mathcal{E}_{ij}(v', 2) = \int_0^1 \mathbb{E}[(\Delta_{ij} v')(S_t)]\,6t(1-t)\,dt$.

(ii) The fields are

$$\varphi(x) := \iota_1^{(2)}(x) = \int_0^1 \big\langle \nabla f(t\mathbf{1}), \tilde{\eta}(x) \big\rangle (4-6t)\, dt,$$
$$\Phi(x,y) := \iota_2^{(2)}(x,y) = \int_0^1 D^2 f(t\mathbf{1}) \big[\tilde{\eta}(x), \tilde{\eta}(y)\big]\, 6t(1-t)\, dt, \tag{26}$$

and Eq. (8) *holds with* $\|w_1^{(2)}\|_{\mathrm{TV}} = 5/3$ *and* $\|w_2^{(2)}\|_{\mathrm{TV}} = 1$.

(iii) $\int_\Omega \varphi\, d\lambda + \frac{1}{2}\int_{\Omega^2} \Phi\, d\lambda^{\otimes 2} = v(\Omega) - v(\emptyset)$.

(iv) For any partition $\pi = \{B_1, \dots, B_n\}$ *and any bounded measurable* $\hat{\varphi} : \Omega \to \mathbb{R}$ *and symmetric* $\hat{\Phi} : \Omega^2 \to \mathbb{R}$*, the cell values*

$$\hat{E}_a := \int_{B_a} \hat{\varphi}\, d\lambda + \frac{1}{2}\int_{B_a \times B_a} \hat{\Phi}\, d\lambda^{\otimes 2}, \qquad \hat{E}_{ab} := \int_{B_a \times B_b} \hat{\Phi}\, d\lambda^{\otimes 2} \quad (a<b), \tag{27}$$

satisfy, for every $A \subseteq [n]$ *with* $S_A = \bigcup_{a\in A} B_a$*,*

$$\sum_{a\in A} \hat{E}_a + \sum_{a<b,\ a,b\in A} \hat{E}_{ab} = \int_{S_A} \hat{\varphi}\, d\lambda + \frac{1}{2}\int_{S_A^2} \hat{\Phi}\, d\lambda^{\otimes 2}. \tag{28}$$

If $\int_\Omega \hat{\varphi}\, d\lambda + \frac{1}{2}\int_{\Omega^2} \hat{\Phi}\, d\lambda^{\otimes 2} = v(\Omega) - v(\emptyset)$*, then* $\sum_a \hat{E}_a + \sum_{a<b} \hat{E}_{ab} = v(\Omega) - v(\emptyset)$.

Proof. (i) Theorem 5 with $\ell = 2$ and $k \in \{1,2\}$ gives $\mathcal{E}_i(v',2) = I_1^{w_1^{(2)}}(v';\{i\})$ and $\mathcal{E}_{ij}(v',2) = I_2^{w_2^{(2)}}(v';\{i,j\})$, where $w_1^{(2)}(t) = 4-6t$ and $w_2^{(2)}(t) = 6t(1-t)$ by Definition 4; writing out Eq. (3) gives the two integrals.

(ii) Theorem 7(iii) with $\ell = 2$ gives Eq. (10) for $k = 1,2$, with integrands $D^1 f(t\mathbf{1})[\tilde{\eta}(x)] = \langle \nabla f(t\mathbf{1}), \tilde{\eta}(x)\rangle$ and $D^2 f(t\mathbf{1})[\tilde{\eta}(x), \tilde{\eta}(y)]$; this is Eq. (26). Since $w_2^{(2)} \geq 0$ on $[0,1]$, $\|w_2^{(2)}\|_{\mathrm{TV}} = \int_0^1 6t(1-t)\, dt = 1$. Since $w_1^{(2)}$ is positive on $[0, 2/3)$ and negative on $(2/3, 1]$, $\|w_1^{(2)}\|_{\mathrm{TV}} = \int_0^{2/3}(4-6t)\, dt + \int_{2/3}^1 (6t-4)\, dt = \frac{4}{3} + \frac{1}{3} = \frac{5}{3}$.

(iii) is Eq. (22) with $\ell = 2$, whose terms are $\int_\Omega \iota_1^{(2)}\, d\lambda = \int_\Omega \varphi\, d\lambda$ and $\frac{1}{2!}\int_{\Omega^2} \iota_2^{(2)}\, d\lambda^{\otimes 2} = \frac{1}{2}\int_{\Omega^2} \Phi\, d\lambda^{\otimes 2}$.

(iv) Apply Proposition 22 with $\ell = 2$, $h_1 = \hat{\varphi}$ and $h_2 = \hat{\Phi}$. In Eq. (24), a singleton $T = \{a\}$ collects $k = 1$ with $n_1 = 1$ and $k = 2$ with $n_1 = 2$, which gives $\hat{E}_a$ of Eq. (27); a pair $T = \{a,b\}$ collects $k = 2$ with $(n_1, n_2) = (1,1)$, which gives $\hat{E}_{ab}$. So $F(A)$ of Eq. (23) is the right side of Eq. (28) and $\sum_{T\subseteq A,\, 1\leq|T|\leq 2} \hat{E}_T$ is its left side. The last claim is Eq. (28) at $A = [n]$, where $S_A = \Omega$, combined with the assumption on $(\hat{\varphi}, \hat{\Phi})$. □

A.10 The efficiency projection direction

At $\ell = 2$ the objective Eq. (12) is quadratic in $\hat{E} = (\hat{E}_T)_{1\leq|T|\leq 2}$ with Hessian, up to a factor 2, $B_{T,T'} = \sum_{0<|A|<n} \mu_n(A)\mathbb{1}[T \subseteq A]\mathbb{1}[T' \subseteq A] = c_{|T\cup T'|}$, where, with $n^{\underline{m}} := n(n-1)\cdots(n-m+1)$,

$$c_m := \sum_{A\supseteq U,\ |U|=m} \mu_n(A) = \frac{n-1}{n^{\underline{m}}} \sum_{k=m}^{n-1} \frac{k^{\underline{m}}}{k\,(n-k)}, \tag{29}$$

by $|\{A : |A| = k,\ A \supseteq U\}| = \binom{n-m}{k-m}$ and $\binom{n-m}{k-m}/\binom{n}{k} = k^{\underline{m}}/n^{\underline{m}}$; each c_m is an $O(n)$ sum, and $c_m := 0$ for $m > n$. We take $n \geq 3$, where B is invertible. Since B is invariant under relabelling the players, $B^{-1}\mathbf{1}$ takes one value z_1 on singletons and one value z_2 on pairs, and $Bz = \mathbf{1}$ collapses to

$$\begin{pmatrix} c_1 + (n-1)c_2 & (n-1)c_2 + \binom{n-1}{2}c_3 \\ 2c_2 + (n-2)c_3 & c_2 + 2(n-2)c_3 + \binom{n-2}{2}c_4 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \tag{30}$$

Table 2: The datasets: equation, parameters, distribution of the input and native grid. Shallow water evolves (h, hu, hv) under $h_t + (hu)_x + (hv)_y = 0$, $(hu)_t + (hu^2 + gh^2/2)_x + (huv)_y = 0$, $(hv)_t + (huv)_x + (hv^2 + gh^2/2)_y = 0$ and is explained on the height channel. Gray–Scott is $u_t = D_u \Delta u - uv^2 + F(1-u)$, $v_t = D_v \Delta v + uv^2 - (F+k)v$ with $D_u = 2 \cdot 10^{-5}$, $D_v = 10^{-5}$, $F = 0.04$, $k = 0.06$ and is explained on u. The retardation factor of diffusion–sorption is $R(u) = 1 + (\rho_b/\phi)\, k_f n_f u^{n_f - 1}$ with $k_f = 3.5$, $n_f = 0.4$, $\rho_b = 1$, $\phi = 0.29$.

Dataset	Equation	Parameters	Input	Grid
Heat 1D	$u_t = \nu u_{xx}$	$\nu t = 2 \cdot 10^{-3}$	Fourier series	256
Advection 1D	$u_t + c u_x = 0$	$c = 1, T = 0.5$	$\mathrm{GRF}(2, 1)$	1024
Burgers 1D	$u_t + (u^2/2)_x = \nu u_{xx}$	$\nu = 0.01, T = 1$	$\mathrm{GRF}(2, 1)$	1024
Allen–Cahn 1D	$u_t = \varepsilon^2 u_{xx} + u - u^3$	$\varepsilon = 0.1, T = 0.5$	$\mathrm{GRF}(2, 0.5)$	1024
Diff.–sorption 1D	$R(u)\, u_t = D u_{xx}$	$D = 5 \cdot 10^{-4}, T = 500$	1 to 4 positive bumps	1024
Heat 2D	$u_t = \kappa \Delta u$	$\kappa = 0.01, T = 1$	$\mathrm{GRF}(2, 1)$	128^2
Shallow water 2D	see caption	$g = 9.81, \bar{h} = 1, T = 0.5$	sines	128^2
Gray–Scott 2D	see caption	$T = 5000, \Omega = [0, 2.5]^2$	1 to 4 bumps	128^2
Darcy 2D	$-\nabla \cdot (a \nabla u) = f$	$f = 1$	$\log a \sim \mathrm{GRF}(4, 1)$	128^2
Heat 3D	$u_t = \kappa \Delta u$	$\kappa = 0.01, T = 1$	$\mathrm{GRF}(2, 1)$	32^3

whose entries count, for a fixed T, the T' with each value of $|T \cup T'|$. Dividing by $\mathbf{1}^\top B^{-1} \mathbf{1} = n z_1 + \binom{n}{2} z_2$ gives, for every $n \geq 3$,

$$z_{\{i\}} = -\frac{2(n-2)}{n(n+1)}, \qquad z_{\{i,j\}} = \frac{6}{n(n+1)}, \tag{31}$$

so the projection never enumerates the 2^n coalitions. The c_m cancel in the normalized solution because z minimizes $\hat{E}^\top B \hat{E}$ subject to $\mathbf{1}^\top \hat{E} = 1$. This is the problem of Definition 1 for the unanimity game $v_{[n]}$, so z is the FSII of $v_{[n]}$, and Eq. (14) with its only nonzero Möbius coefficient $a(v_{[n]}, [n]) = 1$ gives Eq. (31).

B Experimental details and further results

Data. Table 2 lists the ten datasets. Each is simulated by our own solvers on the native grid of the table from random initial states, and the grids of the experiments are strided subsamples of it. Darcy flow uses the problem of Li et al. (2021) with a log-normal instead of a piecewise-constant permeability. The input of a sample is its initial state, or the log-permeability for Darcy flow, and the target is the state at time T. Boundaries are periodic except for diffusion–sorption ($u(0, t) = 1$ and zero flux at $x = 1$) and Darcy flow (zero pressure on the boundary). The heat, advection and Darcy data are solved exactly in Fourier space or by a direct sparse solve. The other equations use fourth order Runge-Kutta method in time with pseudo-spectral or finite-volume space discretizations. $\mathrm{GRF}(\alpha, \sigma)$ denotes a periodic Gaussian random field with spectrum $|k|^{-\alpha/2}$ scaled to standard deviation σ. The heat 1D inputs are periodic Fourier series with modes $k \leq 16$ and Gaussian coefficients of standard deviation $0.4\, e^{-(k/6)^2}$. Bumps are Gaussian bumps of random centre, sign and width (0.05 to 0.2 of the domain). For diffusion–sorption every bump is positive, because the solver clips the concentration at zero and a negative bump would leave an empty interior. For Gray–Scott the absolute bump field is clipped to $[0, 1]$, giving m, and the initial state is $u = 1 - 0.5m$ and $v = 0.25m$. The shallow water height is 1 plus a sum of sines $a_k \sin(2\pi k \cdot x + \varphi_k)$ over the wavevectors $k \in \{1, \ldots, 4\}^2$, with random phases and random amplitudes decaying as $|k|^{-1}$, scaled by 0.1. Its initial velocity is zero. Every dataset has 10,000 training, 1,000 validation and 30 test samples.

Backbones. FNOs of the neuraloperator library (Kossaifi et al., 2026) with four Fourier layers, 64 channels, GELU activations (Hendrycks & Gimpel, 2023) and 16 modes in 1D or 8 modes per axis in 2D and 3D, trained with AdamW at learning rate 10^{-3} and batches of 64 for up to 300 epochs with early stopping on the resolutions of each dataset (16 to 256 in 1D, 8 to 32 in 2D, 8 and 16 in 3D). The checkpoint with the best validation loss is used. Table 3 gives the relative L^2 error and the R^2 on the test split per grid. The 1D operators, shallow water and Darcy flow are accurate on every

Table 3: Backbone accuracy on the 30 test samples per grid of the backbone training ladder, which is 16 to 256 in 1D, 8^2 to 32^2 in 2D, and 8^3 and 16^3 in 3D. We report the relative L^2 error $\|\hat{\mathcal{F}}(u_0) - u_T\|_2/\|u_T\|_2$ averaged over samples and R^2 over all outputs. The explainers are trained and evaluated on the whole ladder in 1D and on its coarser part in 2D and 3D, 8^2 and 16^2 in 2D and 8^3 in 3D.

Dataset	Grids	Relative L^2 error	R^2
Heat 1D	16/32/64/128/256	0.008/0.001/0.001/0.001/0.001	1.000/1.000/1.000/1.000/1.000
Advection 1D	16/32/64/128/256	0.002/0.006/0.011/0.022/0.083	1.000/1.000/1.000/0.999/0.993
Burgers 1D	16/32/64/128/256	0.221/0.129/0.070/0.051/0.050	0.934/0.978/0.995/0.997/0.998
Burgers 1D, ReLU	16/32/64/128/256	0.206/0.135/0.067/0.045/0.048	0.943/0.974/0.995/0.998/0.997
Allen–Cahn 1D	16/32/64/128/256	0.165/0.082/0.040/0.028/0.024	0.973/0.993/0.999/0.999/0.999
Diff.–sorption 1D	16/32/64/128/256	0.001/0.001/0.001/0.001/0.001	1.000/1.000/1.000/1.000/1.000
Heat 2D	8^2/16^2/32^2	0.542/0.277/0.134	0.740/0.935/0.985
Shallow water 2D	8^2/16^2/32^2	0.001/0.001/0.001	1.000/1.000/1.000
Gray–Scott 2D	8^2/16^2/32^2	0.312/0.316/0.318	0.664/0.654/0.649
Darcy 2D	8^2/16^2/32^2	0.042/0.019/0.016	0.995/0.999/0.999
Heat 3D	8^3/16^3	0.579/0.192	0.675/0.965

grid. Heat 2D and 3D lose accuracy on their coarsest grids (R^2 of 0.74 at 8^2 and 0.68 at 8^3), where the strided input undersamples the random field, and the Gray–Scott operator stays at $R^2 = 0.65$ on every grid, because after 5000 time units its patterns depend on details of the input that the strided grids do not carry.

Explainers. The explainer has the trunk of the backbone, a pair head of rank 24 and a query bump of radius 0.05. It is trained for 200 epochs with AdamW (Loshchilov & Hutter, 2019) at learning rate 10^{-3} and a cosine schedule (Loshchilov & Hutter, 2017), batches of 128 inputs in 1D, 64 in 2D and 32 in 3D, 128 coalitions per input and step, the efficiency projection in every step, and grids drawn per batch from 16, 32, 64, 128, 256 in 1D, 8^2, 16^2 in 2D and 8^3 in 3D. The checkpoint with the best validation loss is used. At inference the output is projected onto the efficiency constraint.

Rank of the pair head. Before the efficiency projection, the pair coefficients $\hat{E}_{ab}$ of a grid are the off-diagonal entries of the kernel $G\,\mathrm{diag}(\beta)G^\top$, where G holds the cell integrals of g. This kernel has rank at most $R = 24$, so the explainer represents at most R independent interaction patterns per query. We did not tune R. Interaction kernels with a slow spectral decay, for example sharp fronts in hyperbolic problems, may need a larger R on finer grids.

Reference and baselines. ProxySHAP draws coalitions of uniformly random size, fits a LightGBM (Ke et al., 2017) proxy of the game to them and extracts the FSII from the proxy. We use no residual adjustment, which overflows between 64 and 128 players. The labels use $50{,}000$ coalitions with seed 0. The baselines at $1{,}024$, $4{,}096$ and $16{,}384$ use seed 1, so their coalitions differ from the labels'.

Metrics. For every test sample, $4{,}096$ coalitions are drawn from the FSII measure μ_n and evaluated once. All methods of that sample are scored on this set, drawn independently of their fitting coalitions. Exact enumeration at 16 nodes contains every coalition. Faithfulness is $R^2 = 1 - \sum_A (v(A) - \hat{v}(A))^2 / \sum_A (v(A) - \bar{v})^2$ with $v(A) = v_\pi(A) - v_\pi(\emptyset)$ and $\bar{v}$ the mean of $v(A)$ over the $4{,}096$ coalitions, floored at 0 for the tables. Agreement is the Pearson correlation of the $n(n-1)/2$ pair coefficients with the labels.

Compute. The diffusion–sorption data were generated on a second, CPU-only cluster, and all other runs used one Slurm cluster. Backbones, explainers and label shards ran on one GPU with 16 CPU cores and 64 GB of memory, except the SHAP-IQ and SVARM-IQ runs, which ran on CPU. A backbone trains in 4 to 30 minutes. An explainer trains in about 30 minutes in 1D, 58 minutes in 2D and 199 minutes for heat 3D. On the meshes a backbone trains in 1.5 and 4 hours and an explainer in 7 and 16 hours, for the airfoil and the cylinder. A label shard of 30 samples at $50{,}000$ coalitions takes 7 minutes at 32 players and one hour at 256. The evaluations ran on CPU nodes (Intel Xeon

Platinum 8380, 8 cores and 32 GB per job), 3 to 57 minutes per job in 1D, 17 to 79 minutes in 2D and about 7 hours for heat 3D, for 30 samples, three ProxySHAP budgets, the limit fields and the explainer. The limit fields take most of the time on the larger grids.

B.1 Results per grid

Tables 4 and 5 report means and standard deviations over the 30 samples of each grid, aggregated over the three query points. OperatorSHAP-IQ or the limit fields are the most faithful on every grid of heat 1D, Burgers, Allen–Cahn, heat 2D, shallow water and heat 3D, and the explainer alone on Gray–Scott. ProxySHAP at 16,384 coalitions is ahead of the explainer on advection from 32 nodes on, by at most 0.024, on Darcy flow by 0.029 at 8^2 and 0.006 at 16^2, and on diffusion–sorption by 0.005 to 0.025. On diffusion–sorption the limit fields match ProxySHAP from 32 nodes on. For Gray–Scott the labels themselves have $R^2 = 0.10$ at 8^2 and 0.53 at 16^2 (Table 5), and its backbone has $R^2 = 0.65$ (Table 3).

B.2 Limit fields

The fields of Corollary 29 can also be evaluated directly. For the point readout with one hard mask η_a per node and $h = u_0 - \bar{u}$, the cell coefficients Eq. (13) of the fields Eq. (26), with half of each diagonal pair block folded into the cell (Proposition 22), are

$$E_a = \int_0^1 D\hat{\mathcal{F}}_{x^\star}(\bar{u} + th)[\eta_a h]\,(4 - 6t)\,dt + \frac{1}{2}\int_0^1 D^2\hat{\mathcal{F}}_{x^\star}(\bar{u} + th)[\eta_a h, \eta_a h]\,6t(1-t)\,dt,$$
$$E_{ab} = \int_0^1 D^2\hat{\mathcal{F}}_{x^\star}(\bar{u} + th)[\eta_a h, \eta_b h]\,6t(1-t)\,dt.$$

We compute the integrals by Gauss–Legendre quadrature with 16 nodes in t, with one gradient and one Hessian-vector product per player and node by automatic differentiation. One explanation therefore costs $16(n + 1)$ derivative evaluations of the backbone. It needs no training and no coalitions, but its cost grows with n and it needs a differentiable backbone. The tables list these coefficients as Limit.

On heat, Allen–Cahn, Burgers, shallow water, Darcy flow and heat 3D, the limit is about as faithful as the explainer, and its coefficients satisfy efficiency to 10^{-5} without any projection. At 16 nodes, where the labels are exact, its pair coefficients correlate with them at $r = 0.95$ on Burgers and 0.99 on Allen–Cahn. The limit fails on advection, with $R^2 = 0.73$ at 16 nodes and at most 0.01 on the finer grids, and with mean efficiency gaps of up to 4.6. The advection backbone causes this. Along the path $t \mapsto tu_0$ its readout grows by about one, but its derivative in t reaches 40 to 50 for $t < 0.025$ on three test samples. Inputs of one to two percent of the data amplitude lie outside the training distribution. The limit fields of this backbone are therefore not faithful to its finite game. The finite game only evaluates inputs that equal u_0 or the baseline at every node, so ProxySHAP and the explainer never see the region near the baseline. The limit therefore also requires the backbone to stay accurate along the whole path from the baseline. On Gray–Scott, which no degree-2 model describes well (the labels reach $R^2 = 0.10$ at 8^2 and 0.53 at 16^2), the limit reaches $R^2 = 0.01$.

B.3 ReLU backbone

Proposition 26 needs smooth activations, and every backbone of our experiments uses GELU. We therefore train the Burgers backbone a second time with ReLU and the same architecture, data and recipe, and train and evaluate its explainer in the same way. The ReLU backbone is as accurate as the GELU one, with R^2 from 0.943 at 16 points to 0.997 at 256 Table 3. Table 7 shows that the explainer of the ReLU backbone reaches a faithfulness of 0.89 to 0.92, against 0.93 on the GELU backbone, and it stays ahead of ProxySHAP at every budget and grid. Its pair coefficients agree with the exact labels at 16 points at $r = 0.94$. The limit fields of Section B.2 fail with R^2 below 0.08, because the second derivative of a ReLU network vanishes almost everywhere, so the path integral sees no interaction. The finite game uses no derivatives, so the explainer and ProxySHAP do not depend on smoothness.

Table 4: Faithfulness R^2, floored at 0, and pair agreement per grid of the 1D datasets, mean over three query points and 30 test samples each, standard deviation below. Bold marks the best method and every method within one standard deviation of it. Pair agreement is marked n/a for the linear PDEs, whose pair coefficients are near zero (Section B.7). The labels of the 16-node grids are exact.

		Faithfulness R^2						Order-2 Pearson r to the reference				
		ProxySHAP at budget			Limit	Ours	Ref.	ProxySHAP at budget			Limit	Ours
Dataset	Grid	1k	4k	16k				1k	4k	16k		
Heat 1D	16	**1.000** ±0.000	**1.000** ±0.000	**1.000** ±0.000	**1.000** ±0.000	**1.000** ±0.000	1.000 ±0.000	n/a	n/a	n/a	n/a	n/a
	32	**1.000** ±0.000	**1.000** ±0.000	**1.000** ±0.000	0.999 ±0.001	**1.000** ±0.001	1.000 ±0.000	n/a	n/a	n/a	n/a	n/a
	64	0.997 ±0.005	**0.999** ±0.001	**0.999** ±0.001	**0.998** ±0.003	**0.999** ±0.001	0.999 ±0.001	n/a	n/a	n/a	n/a	n/a
	128	0.986 ±0.025	0.995 ±0.009	**0.997** ±0.006	**0.997** ±0.004	**0.999** ±0.002	0.997 ±0.005	n/a	n/a	n/a	n/a	n/a
	256	0.970 ±0.065	0.986 ±0.031	0.990 ±0.023	**0.996** ±0.006	**0.998** ±0.002	0.991 ±0.021	n/a	n/a	n/a	n/a	n/a
Advection 1D	16	**0.991** ±0.034	**0.990** ±0.036	**0.989** ±0.042	0.727 ±0.397	**0.997** ±0.012	0.998 ±0.008	n/a	n/a	n/a	n/a	n/a
	32	**0.987** ±0.042	**0.988** ±0.039	**0.989** ±0.037	0.011 ±0.102	**0.981** ±0.074	0.989 ±0.038	n/a	n/a	n/a	n/a	n/a
	64	**0.967** ±0.101	**0.967** ±0.100	**0.968** ±0.098	0.010 ±0.099	**0.967** ±0.101	0.968 ±0.098	n/a	n/a	n/a	n/a	n/a
	128	**0.935** ±0.152	**0.942** ±0.130	**0.945** ±0.124	0.001 ±0.012	**0.921** ±0.185	0.945 ±0.123	n/a	n/a	n/a	n/a	n/a
	256	**0.915** ±0.184	**0.922** ±0.172	**0.921** ±0.177	0.000 ±0.000	**0.916** ±0.179	0.921 ±0.178	n/a	n/a	n/a	n/a	n/a
Burgers 1D	16	**0.922** ±0.047	**0.920** ±0.057	**0.914** ±0.061	0.854 ±0.133	**0.927** ±0.042	0.933 ±0.042	0.970 ±0.018	**0.994** ±0.003	**0.996** ±0.002	0.950 ±0.046	0.963 ±0.025
	32	**0.905** ±0.063	**0.912** ±0.069	**0.908** ±0.078	**0.912** ±0.065	**0.929** ±0.060	0.901 ±0.089	0.786 ±0.121	0.938 ±0.056	**0.980** ±0.018	0.925 ±0.040	0.935 ±0.032
	64	**0.879** ±0.101	**0.903** ±0.077	**0.901** ±0.080	**0.929** ±0.070	**0.932** ±0.070	0.899 ±0.083	0.489 ±0.157	0.688 ±0.147	**0.815** ±0.115	**0.785** ±0.084	**0.798** ±0.087
	128	**0.859** ±0.143	**0.890** ±0.105	**0.892** ±0.101	**0.935** ±0.080	**0.934** ±0.079	0.889 ±0.101	0.209 ±0.120	0.352 ±0.171	**0.453** ±0.191	**0.501** ±0.146	**0.524** ±0.149
	256	**0.852** ±0.180	**0.883** ±0.139	**0.887** ±0.131	**0.939** ±0.089	**0.935** ±0.091	0.883 ±0.133	0.064 ±0.056	0.141 ±0.113	**0.196** ±0.149	**0.261** ±0.128	**0.280** ±0.128
Allen–Cahn 1D	16	**0.997** ±0.002	**0.998** ±0.002	**0.997** ±0.003	0.995 ±0.004	**0.998** ±0.001	0.999 ±0.000	0.934 ±0.048	0.993 ±0.006	**0.998** ±0.002	0.994 ±0.003	0.928 ±0.048
	32	0.985 ±0.014	0.994 ±0.007	0.996 ±0.004	**0.998** ±0.001	**0.999** ±0.001	0.996 ±0.005	0.546 ±0.147	0.801 ±0.108	0.931 ±0.051	**0.964** ±0.024	0.886 ±0.046
	64	0.965 ±0.042	0.980 ±0.026	0.985 ±0.019	**0.999** ±0.001	**0.999** ±0.001	0.987 ±0.018	0.222 ±0.128	0.335 ±0.152	0.478 ±0.168	**0.717** ±0.121	**0.664** ±0.119
	128	0.948 ±0.078	0.966 ±0.052	0.971 ±0.045	**1.000** ±0.000	0.999 ±0.002	0.973 ±0.043	0.052 ±0.051	0.088 ±0.075	0.130 ±0.101	**0.340** ±0.125	**0.330** ±0.117
	256	0.940 ±0.110	0.957 ±0.081	0.963 ±0.071	**1.000** ±0.001	**0.999** ±0.003	0.964 ±0.069	0.014 ±0.024	0.021 ±0.035	0.039 ±0.038	**0.150** ±0.074	**0.151** ±0.065
Diff.–sorption 1D	16	**0.992** ±0.012	**0.993** ±0.011	**0.993** ±0.011	0.960 ±0.055	0.968 ±0.052	0.993 ±0.011	0.985 ±0.015	**0.996** ±0.004	**0.998** ±0.002	0.794 ±0.099	0.621 ±0.191
	32	**0.992** ±0.015	**0.994** ±0.012	**0.994** ±0.011	**0.986** ±0.028	0.977 ±0.041	0.994 ±0.011	0.882 ±0.080	0.974 ±0.027	**0.993** ±0.008	0.895 ±0.062	0.651 ±0.197
	64	**0.990** ±0.019	**0.994** ±0.013	**0.995** ±0.011	**0.993** ±0.015	**0.986** ±0.030	0.995 ±0.011	0.609 ±0.176	0.817 ±0.134	**0.919** ±0.088	**0.887** ±0.070	0.624 ±0.215
	128	**0.984** ±0.040	**0.990** ±0.027	**0.992** ±0.024	**0.993** ±0.018	**0.987** ±0.031	0.992 ±0.023	0.322 ±0.180	0.501 ±0.204	**0.633** ±0.209	**0.727** ±0.153	0.482 ±0.222
	256	**0.976** ±0.070	**0.984** ±0.049	**0.986** ±0.044	**0.991** ±0.029	**0.981** ±0.040	0.987 ±0.043	0.139 ±0.142	**0.250** ±0.194	**0.328** ±0.228	**0.450** ±0.202	0.119 ±0.142

Table 5: As Table 4, for the 2D and 3D datasets.

		Faithfulness R^2						Order-2 Pearson r to the reference				
		ProxySHAP at budget			Limit	Ours	Ref.	ProxySHAP at budget			Limit	Ours
Dataset	Grid	1k	4k	16k				1k	4k	16k		
Heat 2D	8^2	0.957 ±0.029	0.975 ±0.017	0.982 ±0.012	0.991 ±0.005	**0.995** ±0.003	0.983 ±0.012	n/a	n/a	n/a	n/a	n/a
	16^2	0.919 ±0.083	0.947 ±0.055	0.954 ±0.048	**0.999** ±0.001	**0.999** ±0.002	0.955 ±0.047	n/a	n/a	n/a	n/a	n/a
Shallow water 2D	8^2	0.956 ±0.038	0.976 ±0.021	0.983 ±0.015	**1.000** ±0.001	0.997 ±0.003	0.984 ±0.014	0.167 ±0.094	0.307 ±0.119	0.482 ±0.146	**0.671** ±0.106	0.376 ±0.153
	16^2	0.919 ±0.111	0.946 ±0.076	0.953 ±0.067	**1.000** ±0.001	0.998 ±0.003	0.954 ±0.066	0.017 ±0.027	0.030 ±0.043	**0.041** ±0.054	**0.101** ±0.061	**0.090** ±0.064
Gray–Scott 2D	8^2	0.059 ±0.082	0.087 ±0.118	0.087 ±0.122	0.000 ±0.000	**0.333** ±0.107	0.095 ±0.134	0.381 ±0.084	0.645 ±0.062	**0.815** ±0.043	0.138 ±0.097	0.687 ±0.035
	16^2	0.394 ±0.153	0.478 ±0.130	**0.506** ±0.127	0.010 ±0.052	**0.611** ±0.132	0.526 ±0.119	0.235 ±0.102	0.427 ±0.117	**0.610** ±0.097	0.199 ±0.107	0.461 ±0.085
Darcy 2D	8^2	0.968 ±0.034	**0.981** ±0.022	**0.985** ±0.017	**0.992** ±0.015	0.956 ±0.058	0.986 ±0.016	0.684 ±0.158	0.867 ±0.084	**0.946** ±0.036	0.897 ±0.043	0.405 ±0.165
	16^2	**0.954** ±0.066	**0.970** ±0.045	**0.975** ±0.039	**0.985** ±0.050	**0.969** ±0.056	0.976 ±0.038	0.174 ±0.126	**0.317** ±0.186	**0.437** ±0.223	**0.484** ±0.175	0.211 ±0.133
Heat 3D	8^3	0.878 ±0.096	0.922 ±0.063	0.934 ±0.053	**0.992** ±0.006	**0.989** ±0.009	0.935 ±0.053	n/a	n/a	n/a	n/a	n/a

Table 6: Median absolute efficiency gap $\left|\sum_T \hat{E}_T - (v([n]) - v(\emptyset))\right|$ over all grids, query points and test samples. The explainer is projected onto the efficiency constraint, so its gap is numerical round-off. The limit fields are not computed on the meshes.

	ProxySHAP at budget			Limit	Ours
Dataset	1k	4k	16k		
Heat 1D	$5.7 \cdot 10^{-4}$	$4.0 \cdot 10^{-4}$	$3.8 \cdot 10^{-4}$	$4.5 \cdot 10^{-8}$	$7.0 \cdot 10^{-11}$
Advection 1D	$8.7 \cdot 10^{-3}$	$9.2 \cdot 10^{-3}$	$9.4 \cdot 10^{-3}$	$9.5 \cdot 10^{-1}$	$4.5 \cdot 10^{-10}$
Burgers 1D	$3.4 \cdot 10^{-3}$	$3.0 \cdot 10^{-3}$	$3.9 \cdot 10^{-3}$	$4.9 \cdot 10^{-8}$	$1.5 \cdot 10^{-9}$
Allen–Cahn 1D	$2.1 \cdot 10^{-3}$	$2.1 \cdot 10^{-3}$	$2.5 \cdot 10^{-3}$	$4.1 \cdot 10^{-8}$	$6.8 \cdot 10^{-10}$
Diff.–sorption 1D	$1.3 \cdot 10^{-3}$	$1.8 \cdot 10^{-3}$	$2.1 \cdot 10^{-3}$	$1.7 \cdot 10^{-4}$	$1.5 \cdot 10^{-9}$
Heat 2D	$1.6 \cdot 10^{-3}$	$1.0 \cdot 10^{-3}$	$1.1 \cdot 10^{-3}$	$2.2 \cdot 10^{-8}$	$5.5 \cdot 10^{-10}$
Shallow water 2D	$6.5 \cdot 10^{-5}$	$4.0 \cdot 10^{-5}$	$4.3 \cdot 10^{-5}$	$1.4 \cdot 10^{-7}$	$1.0 \cdot 10^{-10}$
Gray–Scott 2D	$1.0 \cdot 10^{-2}$	$1.0 \cdot 10^{-2}$	$1.1 \cdot 10^{-2}$	$1.2 \cdot 10^{-4}$	$6.6 \cdot 10^{-10}$
Darcy 2D	$6.2 \cdot 10^{-5}$	$4.9 \cdot 10^{-5}$	$5.4 \cdot 10^{-5}$	$6.8 \cdot 10^{-9}$	$1.4 \cdot 10^{-10}$
Heat 3D	$8.4 \cdot 10^{-4}$	$5.3 \cdot 10^{-4}$	$3.9 \cdot 10^{-4}$	$1.5 \cdot 10^{-8}$	$8.2 \cdot 10^{-11}$
Airfoil mesh	$1.2 \cdot 10^{-3}$	$4.8 \cdot 10^{-4}$	$4.0 \cdot 10^{-4}$	–	$3.5 \cdot 10^{-9}$
Cylinder mesh	$7.2 \cdot 10^{-3}$	$1.3 \cdot 10^{-2}$	$1.6 \cdot 10^{-2}$	–	$1.6 \cdot 10^{-9}$

Table 7: Faithfulness R^2, floored at 0, per grid on the Burgers data for the GELU backbone of the benchmark and a ReLU backbone, mean over three query points and 30 test samples each, standard deviation below. Bold marks the best method and every method within one standard deviation of it.

		Faithfulness R^2				
		ProxySHAP at budget			Limit	Ours
Dataset	Grid	1k	4k	16k		
Burgers 1D	16	**0.922** ±0.047	**0.920** ±0.057	**0.914** ±0.061	0.854 ±0.133	**0.927** ±0.042
	32	**0.905** ±0.063	**0.912** ±0.069	**0.908** ±0.078	**0.912** ±0.065	**0.929** ±0.060
	64	**0.879** ±0.101	**0.903** ±0.077	**0.901** ±0.080	**0.929** ±0.070	**0.932** ±0.070
	128	**0.859** ±0.143	**0.890** ±0.105	**0.892** ±0.101	**0.935** ±0.080	**0.934** ±0.079
	256	**0.852** ±0.180	**0.883** ±0.139	**0.887** ±0.131	**0.939** ±0.089	**0.935** ±0.091
Burgers 1D, ReLU	16	**0.889** ±0.063	**0.884** ±0.077	**0.871** ±0.091	0.023 ±0.079	**0.892** ±0.051
	32	**0.877** ±0.072	**0.886** ±0.077	**0.879** ±0.095	0.043 ±0.121	**0.902** ±0.062
	64	**0.853** ±0.114	**0.885** ±0.077	**0.884** ±0.082	0.052 ±0.160	**0.911** ±0.068
	128	0.824 ±0.192	**0.874** ±0.114	**0.881** ±0.101	0.064 ±0.194	**0.915** ±0.080
	256	**0.827** ±0.210	**0.872** ±0.150	**0.881** ±0.133	0.072 ±0.218	**0.918** ±0.092

B.4 Training variants

Two changes to the training of Algorithm 1 help single datasets. The normalized loss divides the loss of every input by the variance of its coalition values under μ_n, so every input contributes $1 - R^2$ whatever the scale of its game. Soft efficiency drops the projection during training and adds the penalty $w_{\text{eff}}\big(v([n]) - \sum_T \hat{E}_T\big)^2$ with $w_{\text{eff}} = 1$ to the loss, and the projection is applied only at inference. Table 8 compares the plain recipe with the variant on the three datasets where the tables use it.

Shallow water needs the normalized loss. Its height sits at 1 and the node-wise mean baseline differs little from the inputs, so the game values have a standard deviation of 0.002 to 0.017 per sample, against about 0.1 for heat 2D. The plain loss then barely penalizes a constant output, and the explainer converges to one, with R^2 of 0.04 at 8^2 and 0.21 at 16^2. With the normalized loss it reaches 0.997 and 0.998. On Burgers the normalized explainer is as faithful as the plain one, 0.931 in both cases. On diffusion–sorption the standard deviation of the game values ranges from 0.001 to 0.5 between samples, so a few samples dominate the plain loss. The normalized loss raises the faithfulness from 0.92 to 0.95 to 0.97 to 0.99. Its pair coefficients agree less with the exact labels at 16 nodes, at $r = 0.62$ against 0.76. Heat 3D needs soft efficiency. With the projection in every step, training stopped at epoch 78 at a validation loss more than 100 times above the soft run, and the explainer reaches $R^2 = 0.15$ against 0.989 with soft efficiency.

B.5 Zero-shot grids

The explainers are trained on grids of up to 256 nodes in 1D. Table 9 evaluates them on 512 nodes, a grid they never saw, against ProxySHAP labels at $50{,}000$ coalitions. Heat 1D is stored at 256 points and has no finer grid. The faithfulness of the explainer changes by at most 0.036 from 256 to 512 nodes. Increasing on Burgers (0.935 to 0.937) and Allen–Cahn (0.999 in both cases) and dropping on advection (0.916 to 0.907) and diffusion–sorption (0.981 to 0.945). It stays ahead of ProxySHAP at $16{,}384$ coalitions on Burgers, Allen–Cahn and advection, and behind it on diffusion–sorption.

Table 8: Faithfulness R^2, floored at 0, of the plain training recipe (Plain) and of the variant the main tables use (Ours), the normalized loss for shallow water and diffusion–sorption and soft efficiency for heat 3D, mean over three query points and 30 test samples each, standard deviation below.

		Faithfulness R^2	
		Plain	Ours
Dataset	Grid		
Diff.–sorption 1D	16	**0.920** ±0.225	**0.968** ±0.052
	32	0.932 ±0.206	**0.977** ±0.041
	64	0.936 ±0.211	**0.986** ±0.030
	128	0.942 ±0.195	**0.987** ±0.031
	256	**0.947** ±0.182	**0.981** ±0.040
Shallow water 2D	8^2	0.037 ±0.077	**0.997** ±0.003
	16^2	0.210 ±0.166	**0.998** ±0.003
Heat 3D	8^3	0.148 ±0.104	**0.989** ±0.009

Table 9: Faithfulness R^2, floored at 0, at 512 nodes, a grid outside every training ladder, mean over three query points and 30 test samples, standard deviation below. Bold marks the best method and every method within one standard deviation of it.

		Faithfulness R^2				
		ProxySHAP at budget			Limit	Ours
Dataset	Grid	1k	4k	16k		
Advection 1D	512	**0.897** ±0.218	**0.902** ±0.213	**0.899** ±0.224	0.035 ±0.155	**0.907** ±0.196
Burgers 1D	512	**0.858** ±0.194	**0.891** ±0.152	**0.893** ±0.146	**0.942** ±0.093	**0.937** ±0.097
Allen–Cahn 1D	512	0.941 ±0.135	0.956 ±0.102	0.960 ±0.093	**1.000** ±0.001	**0.999** ±0.004
Diff.–sorption 1D	512	**0.973** ±0.093	**0.979** ±0.072	**0.981** ±0.065	**0.989** ±0.039	0.945 ±0.053

B.6 Irregular meshes

We explain GINO backbones on the airfoil and cylinder-flow datasets of MeshGraphNets (Pfaff et al., 2021). Each sample is one trajectory. The input is the velocity field at a start time and the target is the velocity 50 steps later for the airfoil and 100 steps later for the cylinder, both on the native mesh. The airfoil mesh is cut to $[-2, 3] \times [-1, 1]$ and has 3,994 nodes. The cylinder meshes differ between trajectories and have 1,777 to 1,971 nodes. The players are the boxes of a 32×32 grid over the bounding box of the nodes, and a node belongs to the box that contains it. The query points are $x^\star = (0.67, -0.1)$ for the airfoil and $(0.4, 0.1)$ for the cylinder. The airfoil uses the node-wise mean of the training inputs as baseline and the cylinder the zero field. Both datasets have 1,000 training, 100 validation and 100 test trajectories.

The backbones are GINOs with a 64×64 latent grid, the FNO of the grid backbones on it, and a neighbour radius of 0.12 for the airfoil and 0.08 for the cylinder. On the 100 test trajectories they reach $R^2 = 0.991$ and 0.996 and a relative L^2 error of 0.053 and 0.042. The explainers are GINOs of the same form, trained with the soft efficiency penalty, 64 coalitions per input and step and learning rate 10^{-3} for up to 200 epochs. At 1,024 players there are no labels, so the reference is ProxySHAP at 50,000 coalitions on the first 30 test trajectories, and pair agreement is not scored.

Table 10: Faithfulness R^2, floored at 0, of shapiq's FSII estimators on Burgers at three budgets, mean over three query points and 10 test samples. SHAP-IQ and SVARM-IQ were run at 32 and 64 nodes only.

		ProxySHAP			Regression			SHAP-IQ			SVARM-IQ		
		1k	4k	16k	1k	4k	16k	1k	4k	16k	1k	4k	16k
Dataset	Grid												
Burgers 1D	32	0.912	0.919	0.917	0.816	0.929	**0.936**	0.000	0.006	0.009	0.000	0.000	0.009
	64	0.878	0.903	0.903	0.000	0.761	**0.922**	0.000	0.000	0.013	0.000	0.000	0.000
	128	0.851	0.883	**0.887**	0.000	0.000	0.709	–	–	–	–	–	–
	256	0.829	0.866	**0.871**	0.000	0.000	0.000	–	–	–	–	–	–

At 32×32 boxes ProxySHAP at 16,384 coalitions reaches $R^2 = 0.989$ on the airfoil and 0.976 on the cylinder, and the explainer 0.842 and 0.914 (Table 1). We do not report the limit fields on the meshes. GINO sums over neighbours with torch_scatter (Fey & Lenssen, 2019), whose backward pass autograd does not differentiate, so the Hessian-vector products vanish.

B.7 Reference noise

The reference labels are ProxySHAP estimates at 50,000 coalitions. On the 16-node grids we replace them by the exact order-2 FSII over all 2^{16} coalitions. There ProxySHAP at 16,384 coalitions agrees with the exact pair coefficients at $r = 0.996$ on Burgers and 0.998 on Allen–Cahn and diffusion–sorption. On finer grids ProxySHAP does not agree with itself across budgets. At 16,384 coalitions it agrees with the labels on Burgers at $r = 0.98$ for $n = 32$ and at 0.45 for $n = 128$ (Table 4). Pearson r also rewards errors that a method shares with the reference. The trained backbones of the linear PDEs are affine only approximately (Remark 8), so their pair coefficients are small but not zero. With the exact labels at 16 nodes, the mean absolute pair coefficient of advection is 0.6% of the mean absolute singleton, against 13% on Burgers. The pair agreement of the linear PDEs then measures estimator noise, so the tables mark it n/a.

B.8 Finite-player estimators

We compare ProxySHAP with the other FSII estimators of shapiq on Burgers, on the first 10 test samples at the three query points (Table 10). RegressionFSII solves the least-squares problem of Definition 1 on the sampled coalitions (Tsai et al., 2023). At 32 nodes it is more faithful than ProxySHAP from 4,096 coalitions on, with $R^2 = 0.936$ against 0.917 at 16,384 coalitions. Its R^2 is negative whenever the budget is below the number of coefficients, which is 528 at $n = 32$, 2,080 at 64, 8,256 at 128 and 32,896 at 256. ProxySHAP loses faithfulness slowly with n, from 0.917 at 32 nodes to 0.871 at 256 at 16,384 coalitions. For the FSII, shapiq's SHAP-IQ (Fumagalli et al., 2023) and SVARM-IQ (Kolpaczki et al., 2024) return only the pair coefficients and set every singleton to zero, so their R^2 stays below 0.02. We therefore use ProxySHAP as the baseline and the reference.

B.9 Amortization and inference speed

ProxySHAP evaluates the backbone on b coalitions and fits its proxy for every explanation. The explainer calls the backbone during training and once per explanation, on the grand coalition for the efficiency projection. A training step draws $k = 128$ coalitions per input from μ_n and also evaluates the grand coalition, so it costs $k + 1$ backbone calls per input. Over 200 epochs with the validation passes, this sums to $2.7 \cdot 10^8$ to $2.8 \cdot 10^8$ backbone calls per dataset. The explainer is therefore cheaper than ProxySHAP from $2.7 \cdot 10^8/(b-1)$ explanations on for Burgers, which is 5,470 at $b = 50{,}000$, 16,693 at $b = 16{,}384$ and 267,332 at $b = 1{,}024$ (Figure 4, right). On the three datasets of Figure 4 it is also more faithful than ProxySHAP at every budget. Figure 5 compares wall-clock times. One explanation takes the explainer 7 to 14 ms on every grid. ProxySHAP takes 0.2 to 0.9 s at $b = 1{,}024$ and 1.5 to 11.5 s at $b = 16{,}384$, and its time grows with n because every backbone call and the proxy fit grow with n. At $n = 256$ the explainer is 64 to 78 times faster than ProxySHAP at $b = 1{,}024$ and about 800 times faster at $b = 16{,}384$.

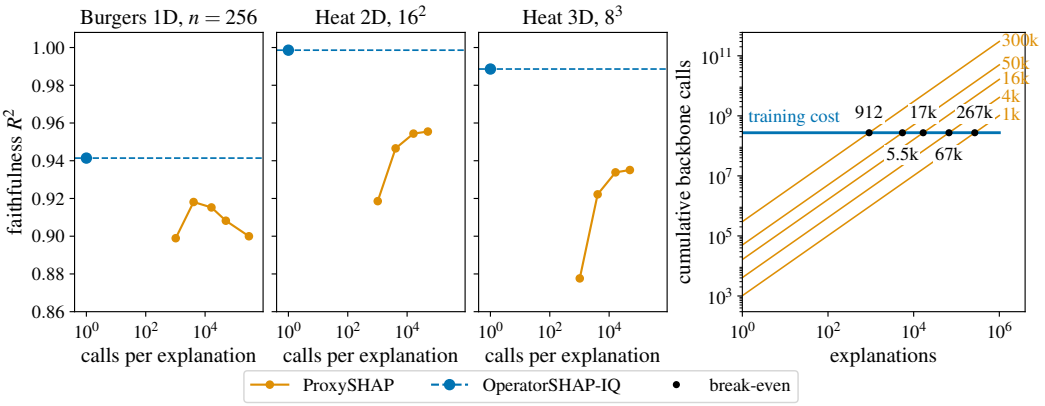


Figure 4: Amortization of OperatorSHAP-IQ. Left: mean faithfulness R^2 against backbone calls per explanation for ProxySHAP at budgets 1,024 to 300,000 and for OperatorSHAP-IQ, who calls the backbone once with the full coalition for the efficiency projection and the explainer itself once for the attribution fields. Burgers comes from a 10-sample run at $x^\star = 0.5$ that reaches 300,000 coalitions, heat 2D and heat 3D from 30 samples at each of three query points. Right: cumulative backbone calls against the number of explanations, one line per ProxySHAP budget, the Burgers training cost of $2.7 \cdot 10^8$ calls as the flat line with its range over the ten datasets as the band, and the break-even counts as dots.

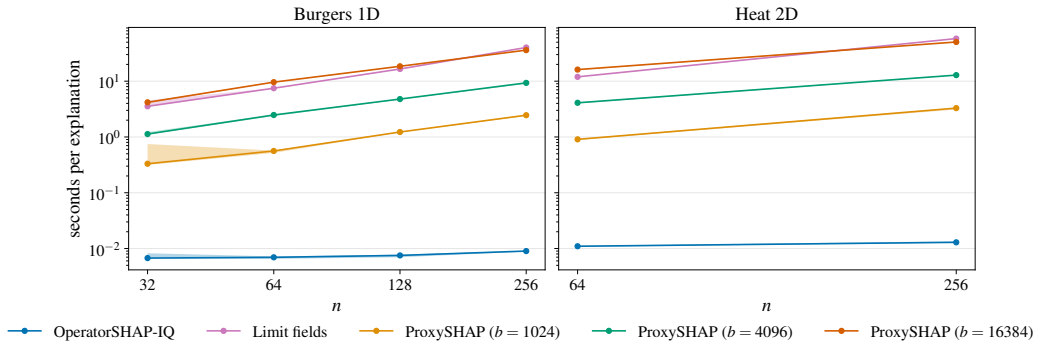


Figure 5: Wall-clock time of one degree-2 FSII explanation against the number of players n, on one CPU threads of one compute node. OperatorSHAP-IQ is one forward pass and the efficiency projection with its single backbone call. ProxySHAP evaluates the backbone once per coalition of its budget b and then fits its proxy. The limit fields take $16(n + 1)$ derivative evaluations of the backbone and are as slow as ProxySHAP at $b = 16{,}384$. Lines are medians over five test samples at $x^\star = 0.5$, bands their range.

B.10 Hard masks

Our theory uses smooth masks η_i with $\sum_i \eta_i \equiv 1$. Smooth masks with $\eta_i = 1$ at the node of B_i and 0 at the other nodes give the node input $\bar{u} + \mu_i(S)h$ at the node of B_i. Because the masks sum to one, η_i puts some mass into the neighbouring cells, so $\mu_i(S) < 1$ when S contains B_i but not a neighbour. As in Stiller et al. (2026), we ignore this spill and set a node to u_0 if its cell is in the coalition and to $\bar{u}$ otherwise. One masking routine then works for every mesh, the coalition values are the ones that the finite-player baselines evaluate, and we observe no degradation of the explanations.

As in Stiller et al. (2026), we assign one player per node when the mesh is small enough, so the players change with the mesh. The explainer is therefore trained on a family of finite games, one per mesh, and each target is the finite FSII of its game. Theorem 7 refines a single fixed game and does

not cover this family. It motivates the field parameterization, which lets one explainer represent the finite FSII of every mesh through the integrals Eq. (13).

B.11 Baseline choice

The baseline enters the operator game only through $h = u_0 - \bar{u}$ and the offset $\hat{\mathcal{F}}_x(\bar{u})$, so Theorem 7 holds for every baseline in $H^s(\Omega)$. However, the game, and with it the interpretation of the output, changes. As for finite player Shapley values, baseline selection is a modeling choice. We use two baselines.

- **Zero.** A masked region carries no initial heat, velocity or concentration, the attribution of a region is its contribution compared to the region being absent, and attributions are comparable across samples.
- **Node-wise mean.** Masked nodes are set to the mean of the training inputs at that node. Shallow water uses it, because its height channel sits around 1 and a depth of zero is not an admissible state of the equations. The same applies to permeability or density fields and to nodes next to a boundary condition that zero would violate.

When selecting a baseline, users should first pay attention that the baseline makes sense for the particular input data (e.g., zero permeability in Darcy's equation would greatly influence the prediction, which is why our Darcy input is the log-permeability, whose zero baseline is permeability one) and if the backbone was trained on similar states. Secondly, even though we interpret the underlying model and not the PDE itself, an integration of boundary values or similar might be preferred for the interpretation. Apart from that, the baseline is a modeling choice and should be declared during the interpretation of the outputs, but no general advice can be given.

Baseline ablation. We compare the two baselines on diffusion–sorption, whose node-wise training mean is 0.70 on average and whose inflow value is 1, so zero lies far from the data. Both explainers use the normalized loss of Section B.4, and each is scored against the labels of its own game (Table 11). Under the node-wise mean the explainer reaches 0.992 to 0.998, against 0.968 to 0.987 under zero, and the limit fields rise from 0.960–0.993 to 0.993–1.000. ProxySHAP at $16{,}384$ coalitions stays at 0.986 to 0.998 under both. On Burgers the node-wise mean is at most 0.08 at an input standard deviation of 1, so both baselines give nearly the same game and the same faithfulness.

C Setup of Figures 1, 3 and 6

Figure 1. Both rows explain a Fourier neural operator that maps the initial state u_0 on the periodic unit interval to the state at time $t = 1$. The heat operator has diffusivity times final time $\nu t = 2{\cdot}10^{-3}$, and its training data are Gaussian random fields with the exact solution. The Burgers operator has viscosity $\nu = 0.01$, and its training data are Gaussian random fields solved numerically on 1024 grid points. Both backbones have four Fourier layers with 64 channels and 16 modes and are trained on grids of 16 to 256 points.

The game is the operator game of Section 4.3 with the point readout $\hat{\mathcal{F}}_{x^\star}(u) = \hat{\mathcal{F}}(u)(x^\star)$ at $x^\star = 0.5$, the zero baseline and one player per node of the 64-point grid on which the test sample is shown. The explainer has the same trunk architecture as the backbones, a pair head of rank 24 and a query bump of radius 0.05. It is trained on grids of 16 to 256 points with 128 coalitions per input and step, with the efficiency projection of Section 4.3 in every step. At inference its output is projected onto the efficiency constraint. The center column shows the cell coefficients $\hat{E}_a$ of Eq. (13) as a step function over the cells, the right column the pair coefficients $\hat{E}_{ab}$ as a symmetric matrix whose diagonal is not defined and left blank. The attribution mass in the caption is $\sum_a |\hat{E}_a| + \sum_{a<b} |\hat{E}_{ab}|$ of the shown sample.

Figure 3. The Burgers example uses a backbone and an explainer trained on a separate dataset of one to three Gaussian bumps of width 0.03 to 0.10, because the Gaussian-random-field training set of Figure 1 does not cover pulse-shaped inputs. Architecture, training grids and inference projection are as above. The two-pulse state is constructed, not drawn from the test split. The finite-volume

Table 11: Faithfulness R^2, floored at 0, per grid of the explainer under the zero baseline (Zero) and the node-wise training mean (Node mean), both with the normalized loss, each against the labels of its own game, mean over three query points and 30 test samples, standard deviation below.

		Faithfulness R^2	
		Zero	Node mean
Dataset	Grid		
Burgers 1D	16	**0.926** ±0.041	**0.926** ±0.041
	32	**0.928** ±0.059	**0.929** ±0.059
	64	**0.932** ±0.069	**0.933** ±0.068
	128	**0.934** ±0.078	**0.935** ±0.076
	256	**0.936** ±0.090	**0.937** ±0.085
Diff.–sorption 1D	16	0.968 ±0.052	**0.992** ±0.005
	32	0.977 ±0.041	**0.995** ±0.004
	64	0.986 ±0.030	**0.997** ±0.002
	128	0.987 ±0.031	**0.998** ±0.002
	256	0.981 ±0.040	**0.998** ±0.002

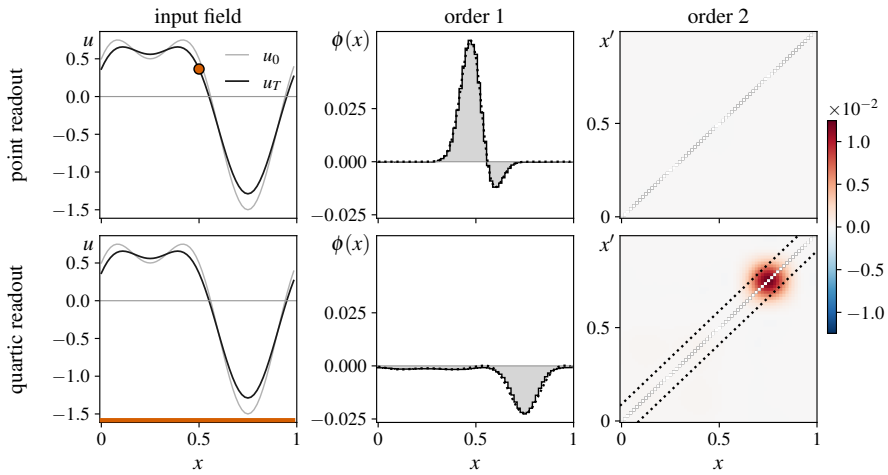


Figure 6: Trained explainers against the analytic heat fields, for a point readout at $x^\star = 0.5$ (top) and the quartic readout $\int_\Omega \hat{\mathcal{F}}(u_0)^4\,dx$ (bottom). The left column shows the input (gray) and the solution (black), the other columns the order-1 and order-2 fields of the explainers. The orange dot marks the point readout and the orange bar the domain of the integral readout. The dotted lines are the analytic fields of the exact heat operator. The linear point readout has no interactions, and the quartic readout has interactions in a band around the diagonal of width $\sqrt{4\nu t}$.

scheme that generated the training data, run at 1024 points on the two-pulse state and its region coalitions (split at $x = 0.3$) under the zero baseline and strided to the explained grid, is reproduced by the backbone to a grid RMSE below 0.001 against a signal of size 0.4.

Figure 6. The explainers are trained on the heat 1D data of Table 2 with the FNO backbone, one for the point readout and one for the quartic readout, with the recipe of Section B. Their order-1 fields correlate with the analytic ones at $r = 0.9999$ and 0.994, and the largest pair coefficient of

the quartic explainer is 0.0120 against 0.0125 analytically. The analytic fields use no network. The model is the exact heat solution operator S_T on the periodic unit interval with $\nu t = 2 \cdot 10^{-3}$. It acts by convolution, $S_T u = K_T * u$, with the heat kernel K_T, a periodic Gaussian of variance $2\nu t$. We explain the input $h(x) = \sin(2\pi x) + \frac{1}{2}\cos(4\pi x)$ with the zero baseline. The 64 grid nodes split the interval into cells $B_1, \dots, B_{64}$ of width $1/64$ with one player per cell, and the mask η_c is the indicator function of B_c (Section B.10). Coalition weights $x \in [0,1]^{64}$ give the input $\sum_c x_c \eta_c h$. The heat operator is linear, so the solution is

$$S_T\Big(\sum_c x_c \eta_c h\Big) = \sum_c x_c g_c, \qquad g_c := S_T(\eta_c h) = K_T * (\eta_c h),$$

where g_c is the part of the solution that comes from cell B_c. A readout R maps the solution to a number, so the game in the form of Theorem 7 is

$$f(x) = R\Big(\sum_c x_c g_c\Big).$$

On the diagonal $x = t\mathbf{1}$ the argument of R is $tS_T h$. The panels show the cell values of the fields of Theorem 7,

$$\int_0^1 D^k f(t\mathbf{1})[e_{a_1}, \dots, e_{a_k}]\, w_k^{(2)}(t)\, dt, \qquad k = 1, 2,$$

where e_a is the a-th unit vector, with half of each diagonal value folded into the order-1 coefficient (Proposition 22). These are the folded coefficients of the continuum fields, not the FSII of the 64-player game.

The point readout $R(u) = u(x^\star)$ is linear, so $f(x) = \sum_c x_c g_c(x^\star)$ and $Df(t\mathbf{1})[e_a] = g_a(x^\star)$ for every t. The second derivative vanishes, so all pair values are zero (Remark 8). Since $\int_0^1 w_1^{(2)}\, dt = 1$, the order-1 value of cell B_a is $g_a(x^\star) = \int_{B_a} K_T(x^\star - y) h(y)\, dy$, the integral over B_a of the density

$$\varphi(x) = K_T(x^\star - x)\, h(x).$$

The quartic readout $R(u) = \int_\Omega u^4\, d\lambda$ gives

$$Df(t\mathbf{1})[e_a] = 4t^3 \int_\Omega (S_T h)^3 g_a\, d\lambda, \qquad D^2 f(t\mathbf{1})[e_a, e_b] = 12t^2 \int_\Omega (S_T h)^2 g_a g_b\, d\lambda.$$

The weights $w_1^{(2)}(t) = 4 - 6t$ and $w_2^{(2)}(t) = 6t(1-t)$ contribute the constants

$$\int_0^1 4t^3(4 - 6t)\, dt = -\tfrac{4}{5}, \qquad \int_0^1 12t^2 \cdot 6t(1-t)\, dt = \tfrac{18}{5}.$$

We write out the convolutions in g_a and g_b and exchange the order of integration, using that K_T is even. The cell values then become integrals over B_a and $B_a \times B_b$ of the densities

$$\begin{aligned} \varphi(x) &= -\tfrac{4}{5}\, h(x)\, \big(K_T * (S_T h)^3\big)(x), \\ \Phi(x, x') &= \tfrac{18}{5}\, h(x) h(x') \int_\Omega (S_T h)^2(z)\, K_T(z - x) K_T(z - x')\, dz. \end{aligned} \tag{32}$$

The dotted curves evaluate these densities, with the folded half-diagonal in the order-1 panels. The cell values satisfy $\sum_a \hat{E}_a + \sum_{a<b} \hat{E}_{ab} = v(\Omega) - v(\emptyset)$ to machine precision, where $v(\Omega) - v(\emptyset)$ is 0.365 for the point and 0.450 for the quartic readout. If the weight $(S_T h)^2$ were constant, the semigroup property $\int_\Omega K_T(z - x) K_T(z - x')\, dz = K_{2T}(x - x')$ would make Φ a band around the diagonal of width $\sqrt{4\nu t} = 0.089$, which the dotted lines in the pair panel mark. The true weight concentrates the band near $x = 0.75$, where $|S_T h|$ is largest.