Partial Differential Equations in Machine Learning and AI: Naturality, Well-Posedness, and Spectral Rates

Anonymous

Higher Categorical Semantics · Version 1.0, 18 November 2026

Abstract. Convolutional layers are difference operators, depth is a time variable, generative diffusion is the heat equation, message passing is a graph Laplacian, and a kernel is a Green's function. We organize these identifications by three questions that the theory of partial differential equations asks of any construction. *Is it natural*, that is, does it commute with the maps of domains, resolutions and symmetries along which one wants to move it? *Is it well-posed*, that is, does the output depend continuously on the input? *At what rate* does it act on each mode? Each qualitative answer is paired with a computed one. Naturality: half-sample reflection models the Neumann condition. Circular layers preserve the reflected subspace exactly when the periodized kernels are symmetric. zero padding is the Dirichlet condition only for three-point stencils; Fourier multipliers commute with resampling, stencils do so to second order, squaring needs more than $3K$ points for band K, and a rectifier is aliased at order N^{-2} on every grid. Well-posedness: reversing diffusion amplifies mode 10 by 5×10^{21}, while the reverse flow driven by the score moves samples by 0.26 per unit of score error, the difficulty reappearing as a score of slope $1/t^2$. Forward Euler amplifies nonzero oscillatory modes. In the stable range $0 < h\omega < 2$, the leapfrog condition number is at most $(1 + h\omega/2)/(1 - h\omega/2)$. A strong-form loss squares the condition number of the energy loss, 414 against 171,682 on 31 unknowns. Rates: oversmoothing proceeds at $1 - \tau\lambda_2$ per layer, with depth scales 415, 5 and 1 on a path, a hypercube and a complete graph of 32 nodes; sheaf diffusion converges to global sections, which exist for 16 of the 128 sign sheaves on a small graph; kernel regression learns mode k at rate $1/(1 + k^2)$ plus aliases; and the spectral gap of Langevin dynamics in a double well approaches the Kramers rate at low noise. All numerical statements are recomputed by a companion script with 42 checks. No new theorem is claimed; the contribution is the organization, the exact constants, and the three questions asked uniformly.

1 Introduction

The companion volume [PDEHC] retells an introductory course on partial differential equations [Str08] around one question: which constructions are natural for maps of domains, of equations and of function spaces. This paper asks the same question of machine learning, where the same equations have been rediscovered piecemeal. A convolutional layer with a three-point kernel is a difference operator, and its padding rule is a boundary condition [Str99]. A residual network is a difference scheme in a time variable called depth [E17, HaRu17, CRBD18]. A diffusion model noises data with the heat equation and samples by reversing it [SWMG15, HJA20, SSKKEP21]. Message passing on a graph is a discrete diffusion [CRGBB21], or a discrete wave equation [RCRMB22], or a diffusion of sections of a sheaf [HaGh19, BDCLB22]. Kernel regression inverts an elliptic operator, physics-informed networks minimize the residual of an equation [RPK19], and operator learning approximates its solution map [LKALBSA21, KLLABSA23, LJPZK21]. A recent survey organizes the last two families [DCWJXKY26].

Three questions organize the theory of the equations themselves, and we ask each of them of each construction.

Naturality. A domain can be reflected, translated, refined or glued, and a construction is natural if it commutes with the corresponding maps. Equivariance [CoWe16, BBCV21] is naturality for symmetries; discretization invariance and the absence of aliasing [KALHHLA21, BBRMMA23] are naturality for changes of resolution.

Well-posedness. Hadamard's criterion asks that a solution exist, be unique, and depend continuously on the data. Learning problems fail the third condition in characteristic ways: by running a parabolic equation backwards, by discretizing a conservative system with a dissipative scheme or the reverse, or by squaring an operator in the loss.

Rates. For linear constructions the answer to "how fast" is a spectrum. Oversmoothing, spectral bias, mixing of stochastic gradient descent and the rank of a learned operator are all statements about eigenvalues of a Laplacian.

What is and is not claimed. Every statement below carries a status tag: classical, literature, reformulation, observation or programme. Almost all are classical or known; where the form or the constant seems not to be standard we say "observation". The checks, tagged like [N3], are performed by the companion script `verify-pde-learning.py`, which uses only the Python standard library and a fixed seed, and are listed in Appendix A. They recompute instances; they are evidence and not proofs. Most models are linear and one-dimensional, so that every number is exact or accurate to many digits; Section 14 says what this leaves out. The references were assembled from memory, except for [DCWJXKY26, BBRMMA23, BDCLB22], which were checked online on 17 November 2026.

2 Setting

Let Dom be a category of domains: intervals, circles, grids, graphs, with the maps of interest between them (inclusions, coverings, symmetries, refinements). A ***signal space*** is a functor $F\colon \mathrm{Dom}^{\mathrm{op}} \to \mathrm{Vect}$, such as functions on the domain with restriction or pullback. A ***layer*** is a family of maps $T_D\colon F(D) \to F(D)$, not necessarily linear.

Definition 2.1. The layer T is ***natural*** for a map $f\colon D \to D'$ if the square

$$\begin{array}{ccc} F(D') & \xrightarrow{T_{D'}} & F(D') \\ {\scriptstyle f^*}\downarrow & & \downarrow{\scriptstyle f^*} \\ F(D) & \xrightarrow[T_D]{} & F(D) \end{array}$$

commutes. Its ***defect*** on an input x is $\|f^* T_{D'} x - T_D f^* x\|$, relative to the norm of either term.

For f a symmetry of D this is equivariance. For f the inclusion of a coarse grid in a fine one, with F restricted to band-limited signals, it is discretization invariance. For f a covering map and F replaced by its invariant part, it is the method of images, with which we begin. A ***depth semigroup*** is a family T^t with $T^{t+s} = T^t T^s$, and a network of N layers of step h is a discretization of T^{Nh}; naturality in t is consistency of a scheme, and well-posedness is its stability [PDEHC, Ch. 12].

3 Padding is a boundary condition

Let $x \in \mathbb{R}^n$ and let $w = (w_{-r}, \dots, w_r)$ be a kernel. The layer is $(C_w x)_i = \sum_j w_j x_{i+j}$, and a ***padding*** rule supplies the values x_i for i outside $0, \dots, n-1$: *zero* ($x_i = 0$), *circular* ($x_{i+n} = x_i$) or *reflect* ($x_{-1-i} = x_i$ and $x_{n+i} = x_{n-1-i}$).

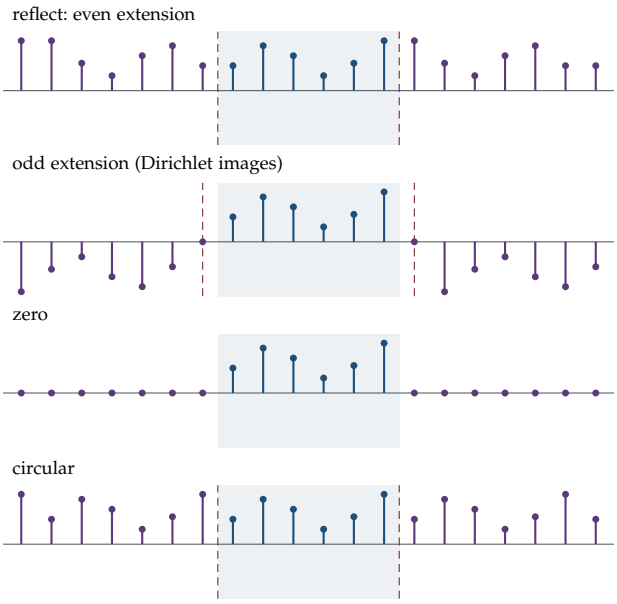


Figure 1: Four ways of continuing a signal of six samples (shaded) beyond its ends. Reflect padding is the even extension about the half-sample points (dashed), which is the Neumann condition. The odd extension about the neighbouring grid points is the Dirichlet condition; zero padding agrees with it only on the first sample outside. Circular padding is the periodic extension.

Proposition 3.1. [classical; [Str99]] *With the kernel* $(-1, 2, -1)$ *the three rules give the Dirichlet, periodic and Neumann difference Laplacians. They are diagonalized by the sine transform* $\sin \frac{\pi k(j+1)}{n+1}$*, the discrete Fourier transform, and the cosine transform* $\cos \frac{\pi k(j+1/2)}{n}$*, with eigenvalues* $4\sin^2 \frac{\pi k}{2(n+1)}$, $4\sin^2 \frac{\pi k}{n}$ *and* $4\sin^2 \frac{\pi k}{2n}$. [N1, N2]

The method of images [PDEHC, Thm. 3.5] explains this construction. In Figure 1, the padding reflects about half-sample points, repeating the end samples. Let $E\colon \mathbb{R}^n \to \mathbb{R}^{2n}$ be the even extension $Ex = (x_0, \dots, x_{n-1}, x_{n-1}, \dots, x_0)$, regarded as a signal on the cycle $\mathbb{Z}/2n$, and let R be the restriction to the first n entries. The image of E is the space of signals invariant under the reflection $\rho\colon i \mapsto 2n - 1 - i$.

Theorem 3.2 (Reflect padding is the Neumann condition). [reformulation; cf. [Str99]; [PDEHC], Thm. 3.5] *Write* C_w^{refl} *and* C_w^{circ} *for the layer with reflect padding on* $\mathbb{R}^n$ *and circular padding on* $\mathbb{R}^{2n}$.

(a) $C_w^{\mathrm{refl}} = R\, C_w^{\mathrm{circ}} E$ *for every kernel.*

(b) $\rho\, C_w^{\mathrm{circ}} \rho = C_{\check{w}}^{\mathrm{circ}}$ *with* $\check{w}_j = w_{-j}$*. Hence if* w *is symmetric,* $C_w^{\mathrm{circ}} E = E\, C_w^{\mathrm{refl}}$*, and a network of symmetric kernels, reflect padding and pointwise nonlinearities is the restriction to invariant signals of a circular network on the doubled cycle.*

(c) The circular layer preserves the reflected subspace exactly when the kernel, reduced modulo $2n$ *by summing*

coefficients with the same residue, is symmetric. An asymmetric periodized kernel sends some even input outside that subspace.

Proof. (a) The reflect rule is the definition of E read at indices outside $0, \dots, n-1$. (b) Substituting $i \mapsto 2n-1-i$ in the sum reverses the kernel. Pointwise maps commute with E and R. For (c), work with the periodized kernel and set $a = w - \check{w}$. If the circular layer preserves the reflected subspace, then C_a^{circ} annihilates the even input $\delta_0 + \delta_{-1}$, by (b). Its output at i is $a_i + a_{i+1} = 0$. Since $a_0 = 0$, all a_i vanish. □

For a random input and the kernels $(0.2, -0.5, 1, -0.5, 0.2)$ and $(0.2, -0.5, 1, 0.7, -0.3)$: identity (a) holds to rounding for both; the output on the double is invariant to rounding for the first and has an anti-invariant part of relative size 1.18 for the second; and two layers on the double differ from two reflect-padded layers by 0 when the symmetric kernel comes first and by 0.100 when the other does. [N3] The circular model with symmetric kernels is equivariant under the dihedral group of the doubled cycle. The half-sample construction restricts it to signals fixed by the chosen reflection. The Dirichlet construction uses the odd extension about the neighbouring grid points.

Proposition 3.3 (Zero padding is Dirichlet only for short kernels)**.** [observation] *Let $O\colon \mathbb{R}^n \to \mathbb{R}^{2n+2}$ be the odd extension $(0, x, 0, -x^{\mathrm{rev}})$. For a symmetric kernel of radius one, zero padding equals circular convolution on Ox restricted to the original positions. For radius two or more it does not: the image construction supplies $-x_0$ beyond the boundary where zero padding supplies 0. Pointwise maps commute with O only if they are odd.*

The relative difference is 0 for $(0.3, 1, 0.3)$ and 0.160 for the five-point symmetric kernel above. [N4] The Dirichlet image construction can differ from zero padding for wider kernels. A rectifier can also destroy the odd symmetry required by the construction.

Proposition 3.4 (What padding breaks and keeps)**.** [classical] (a) *Circular convolution commutes with translation. With zero padding the commutator with a shift is supported within the kernel radius of the boundary.* (b) *A layer with a non-negative symmetric kernel of sum one and reflect padding is doubly stochastic: it conserves the mean and obeys the maximum principle. With zero padding it is substochastic, and the mass remaining after m layers is the probability that a random walk with steps distributed by the kernel has not left the interval.*

For a five-point kernel on 20 positions the commutator is supported on positions 0, 18 and 19. [N5] Position information therefore enters a zero-padded network at the boundary and spreads inwards by one kernel radius per layer, which is the mechanism behind the observation that such networks learn absolute position

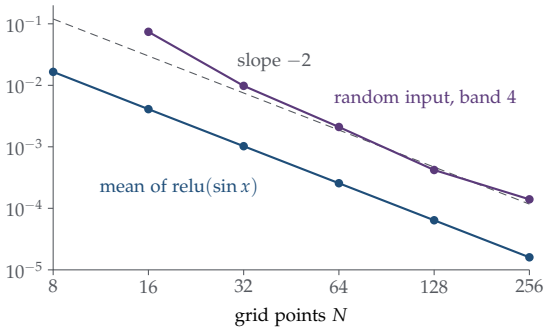


Figure 2: Aliasing error of a rectifier against the number of grid points, on logarithmic scales: the closed-form case of Theorem 4.1(d) and the largest in-band coefficient error for a random input of band 4. Both follow the slope −2.

[ILB20]. For the kernel $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$ on 32 positions, 200 layers with reflect padding change the total mass by 2×10^{-15}, and with zero padding leave 0.5320 of it, in agreement with the sum over the sine basis. [N6]

4 Resolution: what commutes with resampling

Let B_K be the trigonometric polynomials of degree at most K on the circle and $S_N\colon B_K \to \mathbb{R}^N$ the sampling map on N equally spaced points, injective for $N > 2K$. An operator T on functions with discretizations T_N is natural for refinement (Definition 2.1) when $S_N T = T_N S_N$ on B_K; this is the requirement of representation equivalence in [BBRMMA23].

Theorem 4.1 (Three kinds of layer)**.** [reformulation; cf. [BBRMMA23, KALHHLA21]] *On B_K and for $N > 2K$:*

(a) A Fourier multiplier $m(D)$, discretized by the discrete Fourier transform, is natural exactly.

(b) The three-point stencil for $-\partial_x^2$ has symbol $k^2 \operatorname{sinc}^2(kh/2)$, $h = 2\pi/N$: it is natural to second order.

(c) A polynomial nonlinearity of degree d followed by projection to B_K is natural if $N > (d+1)K$, and for generic inputs only then.

(d) A rectifier followed by projection to B_K is natural for no N; the defect is of order N^{-2}.

Proof. (a) Sampling intertwines the characters e^{ikx}, $|k| \le K$, with the discrete ones. (b) Apply the stencil to e^{ikx}. (c) The power has band dK, and a frequency k' is sampled onto $k \neq k'$ when N divides $k' - k$; the largest $|k' - k|$ with $|k| \le K$ is $(d+1)K$. (d) A function with a corner has Fourier coefficients of order k^{-2}, and the sum of those at $k + mN$, $m \neq 0$, is of order N^{-2}. □

For a random input of band 5, the multiplier $(1 + k^2)^{-1}$ computed on 16 and on 64 points agrees with the ex-

act result to 10^{-14}; the stencil is in relative error by 0.221 and 0.0153, its symbol at $k = 5$ being 0.7173 and 0.9801 of the true one. [A1] For band 4 the in-band coefficients of the square are exact on 13 points and wrong by 1.4 on 12: the "two-thirds rule" of spectral methods [Ors71]. [A2] For the rectifier one case is closed-form: the mean of $\mathrm{relu}(\sin x)$ sampled on N points is $\frac{1}{N}\cot\frac{\pi}{N}$ instead of $\frac{1}{\pi}$, an error of $-\pi/3N^2$ to leading order; N^2 times the error is -1.058, -1.050, -1.048, -1.047 for $N = 8, \dots, 64$. [A3] For a random input of band 4 the largest in-band error of the rectified signal is 7.4×10^{-2}, 9.8×10^{-3}, 2.1×10^{-3}, 4.2×10^{-4}, 1.4×10^{-4} on $16, \dots, 256$ points (Figure 2). [A4] This is the quantitative content of the practice of applying nonlinearities on an oversampled grid and filtering [KALHHLA21]: oversampling by a factor M buys M^2, and no finite factor buys exactness. An architecture built from (a) alone is resolution-independent [LKALBSA21]; one with stencils is so to the order of the stencil; one with rectifiers is so only approximately, at every resolution.

5 Diffusion models

Let data $X_0 \sim p_0$ on the line be noised by $X_t = X_0 + W_t$, so that the density solves the heat equation $\partial_t p = \frac{1}{2}\partial_x^2 p$, and write $s = \partial_x \log p_t$ for the ***score***. Our example is a mixture of two Gaussians, for which everything is closed-form. [S1]

Proposition 5.1. [classical; [Efr11, Hyv05, Vin11]] (a) $\mathbb{E}[X_0 \mid X_t = x] = x + t\,s(x,t)$. (b) *For every candidate* σ,

$$\mathbb{E}\big|\sigma(X_t) + \tfrac{X_t - X_0}{t}\big|^2 = \mathbb{E}|\sigma(X_t) - s(X_t,t)|^2 + \tfrac{1}{t} - J(p_t),$$

where J is the Fisher information: denoising and explicit score matching differ by a constant.

Both hold to rounding in the example; for two-point data at $t = \frac{1}{2}$ the constant is 0.924073 for each of two test functions. [S2, S8] Since $\frac{1}{2}\partial_x^2 p = -\partial_x\big(p\cdot(-\frac{1}{2}s)\big)$, the heat equation is also a continuity equation, and its characteristics are the ***probability flow*** $\dot{x} = -\frac{1}{2}s(x,t)$ [SSKKEP21].

Theorem 5.2 (Where the ill-posedness goes). [reformulation; cf. [PDEHC], Ex. 2.5; [And82, SSKKEP21]]

(a) The backward heat equation multiplies the k-th Fourier mode of p_T by $e^{k^2T/2}$: recovering p_0 from p_T is ill-posed in every Sobolev norm.

(b) Given the score on $[t_0, T]$ with $t_0 > 0$, the probability flow is an ordinary differential equation with a Lipschitz field; it transports p_T to p_{t_0}, and an error ϵ in the score, in the supremum norm, moves each sample by at most $C(t_0)\epsilon$.

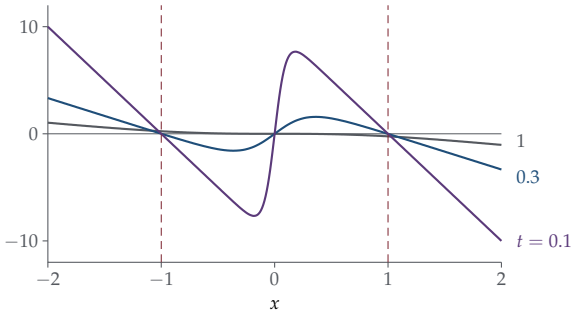


Figure 3: The score $s(x,t) = -x/t + \tanh(x/t)/t$ of data on the two points ±1 (dashed), at three noise levels. As $t \to 0$ it becomes a restoring force towards each point, joined by a transition of slope $1/t^2$ at the midpoint.

(c) For data on the two points $\pm a$, $\partial_x s(0,t) = -1/t + a^2/t^2$. The Lipschitz constant of the field is not integrable at $t = 0$, and $C(t_0) \to \infty$ as $t_0 \to 0$.

Proof. (a) is Fourier analysis. (b) is the uniqueness theorem for characteristics together with Grönwall's inequality. (c) Here $s = -x/t + (a/t)\tanh(ax/t)$. □

For $T = 1$ and $k = 10$ the factor in (a) is 5.2×10^{21}. In the example the flow from $T = 1$ to 0 preserves the value of the distribution function along five trajectories to 10^{-14}, and a perturbation $\epsilon\sin 3x$ of the score displaces the samples by $0.264\,\epsilon$. [S3, S4] The slope in (c) is 90 at $t = 0.1$ and 9900 at $t = 0.01$ (Figure 3). [S5] The three statements together say that a diffusion model does not solve the backward heat equation. It replaces an ill-posed linear problem for the density by a well-posed nonlinear transport problem for samples, whose coefficient, the score, carries the whole difficulty, concentrated at small times and near the boundaries between modes. Statement (c) is the sample-level form of Hadamard's example.

Proposition 5.3 (Entropy). [classical; [Sta59, BaEm85]] (a) *Under the heat flow $\frac{d}{dt}h(p_t) = \frac{1}{2}J(p_t)$.* (b) *Under the Ornstein–Uhlenbeck flow $dX = -X\,dt + \sqrt{2}\,dW$, with Gaussian limit γ, $D(p_t\,\|\,\gamma) \le e^{-2t}D(p_0\,\|\,\gamma)$, and $e^{2t}D(p_t\,\|\,\gamma) \to \frac{1}{2}(\mathbb{E}X_0)^2$.*

In the example $dh/dt = 0.535607$ and $J/2 = 0.535606$ at $t = 0.3$. [S6] The ratio $D(p_t\,\|\,\gamma)/e^{-2t}D(p_0\,\|\,\gamma)$ is 0.353, 0.204, 0.119, 0.088 at $t = \frac{1}{4}, \frac{1}{2}, 1, 2$, and $e^{2t}D(p_t\,\|\,\gamma) = 0.03128$ at $t = 4$ against the limit 0.03125. [S7] The rate 2 is twice the spectral gap of the generator, and it fixes the noising time needed to make the terminal law Gaussian to a given accuracy: $\frac{1}{2}\log(D(p_0\,\|\,\gamma)/\delta)$.

6 Flows and characteristics

A continuous normalizing flow [CRBD18] moves samples by $\dot{x} = v(x,t)$; the density then solves the continu-

ity equation $\partial_t p + \partial_x(pv) = 0$, of which the trajectories are the characteristics [PDEHC, Prop. 1.8].

Proposition 6.1. [classical; [CRBD18]] *Along a trajectory* $\frac{d}{dt}\log p(x(t),t) = -\partial_x v(x(t),t)$.

For $v = 0.8\tanh(1.3x + 0.2)$ the exponential of the integrated divergence equals the derivative of the flow map to 10^{-8}. [F1] ***Flow matching*** [LCBNL23, LGL23] chooses a coupling of $X_0 \sim p_0$ and $X_1 \sim p_1$, sets $X_t = (1-t)X_0 + tX_1$, and regresses onto $v(x,t) = \mathbb{E}[X_1 - X_0 \mid X_t = x]$.

Theorem 6.2 (One dimension). [observation] *On the line, if v is regular enough for trajectories to be unique, the flow of the flow-matching field is the increasing rearrangement of p_0 onto p_1, whatever the coupling.*

Proof. The marginal field transports p_0 to p_1 by construction. Trajectories of a scalar equation cannot cross, so the flow map is increasing, and an increasing map carrying one atomless law to another is unique. □

For $p_0 = N(0,1)$, $p_1 = N(2, \frac{1}{4})$ and the independent coupling the field is linear in x with coefficients rational in t, and its flow is $x \mapsto 2 + \frac{1}{2}x$ to 10^{-13}. [F2] The conditional paths are straight lines that cross; the marginal characteristics do not. In higher dimensions non-crossing no longer determines the map, and the flow depends on the coupling.

Proposition 6.3 (A pitchfork in time). [observation] *For $p_0 = N(0,1)$ and p_1 uniform on ± 1, with the independent coupling,*

$$v(x,t) = \frac{1}{1-t}\left[\tanh\frac{tx}{(1-t)^2} - x\right].$$

The midpoint $x = 0$ is an attracting equilibrium for $t < t^* = \frac{3-\sqrt{5}}{2} = 0.382$ and a repelling one afterwards, and $\partial_x v(0,t) \sim (1-t)^{-3}$ as $t \to 1$.

The field satisfies the continuity equation with p_t to 10^{-13}; the slope at the midpoint changes sign at 0.381966 and is 890 at $t = 0.9$ (Figure 4). [F3] This is the transport-side picture of Theorem 5.2(c): mass begins to be committed to one of the two points at a definite time t^*, by the bifurcation that [PDEHC, Thm. 14.7] describes for a buckling rod, and the field then steepens without bound as the characteristics focus. Such transition times have been studied for diffusion models in high dimension [BBDM24].

7 Graph networks

Let L be the Laplacian of a connected graph with eigenvalues $0 = \lambda_1 < \lambda_2 \le \cdots \le \lambda_n$. The linear part of a message-passing layer is $X \mapsto (I - \tau L)X$, one explicit Euler step of the heat equation on the graph [CRGBB21].

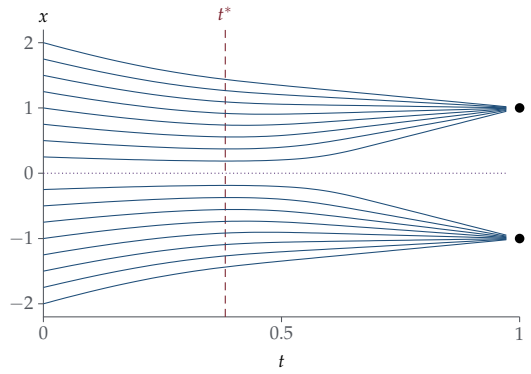


Figure 4: Characteristics of the flow-matching field from $N(0,1)$ to the two points ± 1 (Proposition 6.3). The midpoint is attracting until $t^* = 0.382$ and repelling afterwards; the dotted line is the unstable trajectory. No two characteristics cross.

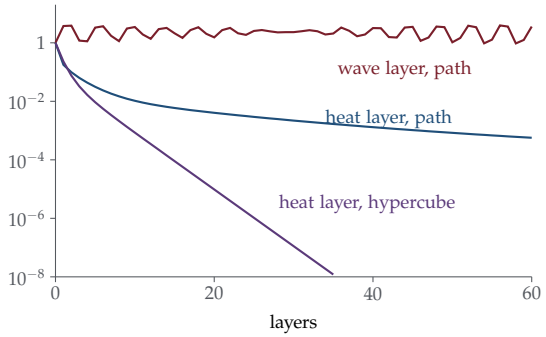


Figure 5: Dirichlet energy of a random signal on 32 nodes against depth, relative to its initial value. Heat-type layers lose it at a rate set by the spectral gap, slowly on a path and quickly on a hypercube; a wave-type layer on the path (Theorem 7.2) keeps it.

Theorem 7.1 (Oversmoothing is the spectral gap). [classical; [LHW18, OoSu20]] *For $0 < \tau \le 1/\lambda_n$ the non-constant part of X is multiplied by at most $1 - \tau\lambda_2$ per layer, with equality in the limit, and the Dirichlet energy $X^{\mathsf{T}}LX$ by at most $(1-\tau\lambda_2)^2$. The depth at which features become indistinguishable is of order $1/\tau\lambda_2$.*

On a random connected graph with 12 nodes, $\lambda_2 = 0.6251$, $\lambda_n = 8.2655$ and the measured factor is 0.92437, as predicted; 99% of the Dirichlet energy of a random input is gone after 15 layers, the bound from λ_2 alone being 29. [G1] The depth scale varies by orders of magnitude with the graph: on 32 nodes, with τ at its largest stable value, it is 415 layers for a path ($\lambda_2 = 0.00963$), 5 for the hypercube ($\lambda_2 = 2$, $\lambda_n = 10$) and 1 for the complete graph, where a single layer returns the mean exactly (Figure 5). [G2] Graphs that are good for propagating information are the ones that oversmooth fastest; the two are the same eigenvalue.

A layer of second order in depth behaves as a wave equation [RCRMB22].

Theorem 7.2 (Wave layers). [reformulation; cf. [PDEHC], Thm. 12.3] *The recursion* $X_{m+1} = 2X_m - X_{m-1} - \tau^2 L X_m$ *conserves*

$$\tau^{-2}\|X_{m+1} - X_m\|^2 + X_{m+1}^{\mathsf{T}} L X_m,$$

which is positive definite on non-constant signals for $\tau^2\lambda_n < 4$. *For* $\tau^2\lambda_n > 4$ *the top mode grows per layer by the modulus of the larger root of* $\xi^2 + (\tau^2\lambda_n - 2)\xi + 1 = 0$.

Over 500 layers at $\tau^2\lambda_n = 3.6$ the invariant drifts by 2×10^{-15} and the mean Dirichlet energy of the last hundred layers is 1.001 times that of the first hundred: nothing is smoothed. At $\tau^2\lambda_n = 4.4$ the growth factor is 1.8633. [G3] The bound is the Courant–Friedrichs–Lewy condition, and the dichotomy between the two theorems is that of [PDEHC, Ch. 4]: a Markov semigroup forgets, a symplectic group does not.

Both layers assume that neighbouring nodes should agree. A ***cellular sheaf*** [HaGh19] replaces agreement by agreement after transport: each edge carries linear maps from the data at its ends, and the sheaf Laplacian $L_{\mathcal{F}}$ penalizes the mismatch. Its kernel is the space H^0 of global sections. The simplest case is a sign $\sigma_e = \pm 1$ on each edge, with $x^{\mathsf{T}} L_{\mathcal{F}} x = \sum_{e=uv}(x_u - \sigma_e x_v)^2$.

Theorem 7.3 (Sheaf diffusion). [literature; [HaGh19, BDCLB22]] *Diffusion by* $L_{\mathcal{F}}$ *converges to the orthogonal projection of the input onto* H^0. *For a sign sheaf on a connected graph,* H^0 *has dimension one if the signs are a coboundary,* $\sigma_{uv} = \ell_u \ell_v$, *in which case it is spanned by* ℓ; *otherwise* $H^0 = 0$. *The obstruction is the class of* σ *in* $H^1(G; \mathbb{Z}/2)$, *a group of order* 2^{E-V+1}.

On a random graph with ten nodes, two classes, and signs $+$ within and $-$ between classes, sheaf diffusion converges to the class labelling times $\langle \ell, X_0 \rangle / n$, to rounding, while ordinary diffusion converges to a constant and loses the classes. Reversing one edge that lies on a cycle makes the smallest eigenvalue 0.1606 and the limit zero; for a cycle of length 7 with one negative edge it is $4\sin^2 \frac{\pi}{14} = 0.19806$, the antiperiodic Laplacian. [G4] On a graph with 5 nodes and 7 edges, exactly $16 = 2^{V-1}$ of the 128 sign assignments have a non-zero section, and they are exactly those whose Laplacian is singular over the rationals; the 128 assignments fall into $8 = 2^{E-V+1}$ classes. [G5] This is the same first cohomology that classifies flat gauge fields in [PDEHC, Prop. 13.1]: choosing the transport maps of a sheaf network is choosing a connection, and heterophily is non-trivial holonomy. A sign sheaf whose class is non-zero cannot be made consistent by relabelling nodes, and its diffusion has no non-trivial limit.

8 Depth as time

A residual layer $x \mapsto x + hf(x)$ is one step of forward Euler for $\dot{x} = f(x)$ [E17, HaRu17]. The type of the linearization of f decides which schemes are admissible, as for partial differential equations [PDEHC, Ch. 12].

Theorem 8.1 (Layers by type). [reformulation; cf. [HaRu17, CMHRBH18]; [PDEHC], Thm. 12.2, 12.3]

(a) Parabolic, $f = -Ax$ *with* A *positive definite: forward Euler is stable if and only if* $h \le 2/\lambda_{\max}(A)$.

(b) Hyperbolic, $f = Jx$ *with* J *antisymmetric with eigenvalues* $\pm i\omega$: *forward Euler multiplies the norm by* $\sqrt{1 + h^2\omega^2}$ *per layer, for every* h.

(c) *For* $h, \omega > 0$, *the leapfrog layer* $q \mapsto q + hp$ *followed by* $p \mapsto p - h\omega^2 q$, *with the new* q, *has determinant one, conserves* $p^2 + \omega^2 q^2 + h\omega^2 pq$, *and has bounded powers exactly when* $h\omega < 2$.

(d) *In this stable range, the Jacobian of* N *leapfrog layers in coordinates* $(\omega q, p)$ *has condition number at most* $(1 + \frac{h\omega}{2})/(1 - \frac{h\omega}{2})$, *for every* N. *The Jacobian of* N *parabolic layers with* $h\lambda_{\max} < 1$ *has condition number* $\big((1 - h\lambda_{\min})/(1 - h\lambda_{\max})\big)^N$.

Proof. (a) and (b) follow from the eigenvalues of $I + hDf$. For (c), direct computation gives the invariant, which is positive definite when $h\omega < 2$. The update matrix has trace $2 - (h\omega)^2$ and determinant one. At $h\omega = 2$ it has a nontrivial Jordan block at -1, giving linear growth for some data. For $h\omega > 2$ an eigenvalue has modulus greater than one. (d): a map that preserves a positive definite quadratic form Q is conjugate to an orthogonal one by $Q^{1/2}$, so its powers have condition number at most that of Q. □

For a random positive definite A of size 6 the per-layer factor is 0.9814 at $h = 1.9/\lambda_{\max}$ and 1.1000 at $2.1/\lambda_{\max}$. [R1] Forward Euler on a rotation grows by 1.6446 over time 10 with $h = 0.1$ and by 1.0513 with $h = 0.01$, that is by $e^{h\omega^2 T/2}$: the instability vanishes only in the limit of infinitely many layers. [R2] Leapfrog conserves its invariant to 10^{-14} at $h\omega = 1.9$ and grows by 1.8773 per layer at 2.1. [R3] With $h\omega = \frac{1}{2}$ the condition number of the product of 50, 200 and 800 leapfrog Jacobians is 1.07, 1.30 and 1.53, below the bound 1.667; for fifty Euler layers on eigenvalues 1 and 10 with $h = 0.05$ it is 8.7×10^{13}. [R4] Since back-propagation multiplies by the transposed Jacobians, (d) is a statement about vanishing and exploding gradients: a symplectic architecture has neither, by the conservation law that a wave equation has and a heat equation lacks. The price is Theorem 7.2: it does not forget either.

9 Kernels and losses

In the kernel regime, gradient descent on the squared loss moves the function by $f \mapsto f - \eta K(f - y)$ for a positive kernel operator K [JGH18]. For a translation-invariant kernel whose Fourier coefficients decay like

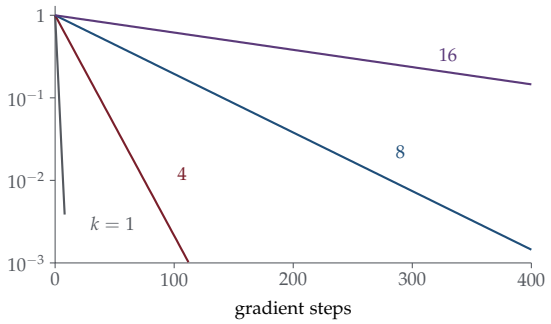


Figure 6: Spectral bias in the kernel regime: the remaining fraction of mode k of the target against the number of gradient steps, for the sampled H^1 kernel on 64 points (Theorem 9.1).

k^{-2}, as for the simplest rectifier networks on the circle [RBADLHBC19], the model case is $K = (1 - \partial_x^2)^{-1}$, the Green's operator of an elliptic equation.

Theorem 9.1 (Spectral bias, with aliases)**.** [reformulation; cf. [RBADLHBC19, JGH18]] (a) *The residual in mode k after m steps is $(1 - \eta\mu_k)^m$ times its initial value, with $\mu_k = (1 + k^2)^{-1}$.* (b) *On N equally spaced samples the eigenvalue is*

$$\mu_k^N = \sum_{m \in \mathbb{Z}} \frac{1}{1 + (k + mN)^2} = \frac{\frac{\pi}{N} \sinh \frac{2\pi}{N}}{\cosh \frac{2\pi}{N} - \cos \frac{2\pi k}{N}}.$$

(c) *The interpolant $u = \sum_i c_i K(\cdot - x_i)$ solves $u - u'' = \sum_i c_i \delta_{x_i}$.*

Statement (b) is the Poisson summation formula of [PDEHC, Thm. 4.6]: sampling in space is summation over the dual lattice. On 64 points μ_k^N / μ_k is 1.0016, 1.0852 and 1.2346 at $k = 1, 10, 16$. [K1] With $\eta = 1$ and target $\cos x + \cos 8x$, after 200 steps mode 1 is fitted to rounding and 0.0380 of mode 8 remains; the steps needed to reach one per cent are 6.6, 282 and 956 for $k = 1, 8, 16$ (Figure 6). [K2] For (c), with five data points the interpolant satisfies the homogeneous equation away from the data to 10^{-9} and its derivative jumps by $-c_i$ at x_i: the representer theorem is the statement that the minimizer is a combination of fundamental solutions. [K5]

A physics-informed loss [RPK19] for $-u'' = f$ is the squared residual $\|Au - f\|^2$ with $A = -\partial_x^2$; the energy, or Ritz, loss [EYu18] is $\frac{1}{2}\langle u, Au\rangle - \langle f, u\rangle$.

Theorem 9.2 (Squaring the operator)**.** [classical; cf. [KGZKM21, WTP21]] (a) *The Hessian of the strong-form loss is $A^{\mathsf{T}}A$ and that of the energy loss is A: the condition numbers are κ^2 and κ, with $\kappa \sim 4N^2/\pi^2$ for N unknowns.* (b) *Composed with the kernel of Theorem 9.1, the strong-form loss has mode rates $\mu_k k^4 = k^4/(1 + k^2)$, increasing in k.*

On 31 unknowns $\kappa = 414.3$ and $\kappa^2 = 171{,}682$; gradient descent with the optimal step needs 2862 and 1,185,936 iterations to gain six digits. [K3] On 64 points the regression rates at $k = 1$ and 16 are 0.5008 and 0.0048, a ratio of 1/104, and the strong-form rates are 0.5000 and 206.8, a ratio of 414. [K4] The bias is reversed and the spread of rates is four times as large: the low modes, which carry the solution, are now the slow ones relative to the step that the high modes allow. This is a linear model of the training difficulties reported for physics-informed networks [KGZKM21, WTP21], and it is the reason the weak formulation is the standard one in numerical analysis [PDEHC, Ch. 12].

10 Noise

Gradient descent with small isotropic noise on a loss V is modelled by the Langevin equation $dX = -V'(X)\,dt + \sqrt{2/\beta}\,dW$, whose density solves a Fokker–Planck equation with stationary solution proportional to $e^{-\beta V}$ [MHB17].

Proposition 10.1. [classical; [Kra40, BaEm85]] (a) *For $V = \frac{1}{2}\lambda x^2$ the spectral gap is λ for every β.* (b) *For a symmetric double well with minima $\pm a$ and barrier ΔV at 0, the gap is $\frac{1}{\pi}\sqrt{V''(a)|V''(0)|}\,e^{-\beta\Delta V}(1 + O(\beta^{-1}))$.* (c) *Constant-step stochastic gradient descent on $\frac{1}{2}\lambda x^2$ with gradient noise of variance σ^2 has stationary variance $\eta\sigma^2/\lambda(2 - \eta\lambda)$.*

A reversible discretization has the Gibbs density as its exact stationary vector, and for $\lambda = 1$ its gap is 0.9997 at $\beta = 1$ and at $\beta = 4$. [L1] For $V = (x^2 - 1)^2$ the gap divided by $\frac{\sqrt{32}}{\pi}e^{-\beta}$ is 0.903, 0.928, 0.947, 0.959 at $\beta = 4, 6, 8, 10$, and the gap at $\beta = 10$ is 7.8×10^{-5}. [L2] The recursion for the variance converges to 0.0625 for $\eta = 0.1$, $\lambda = 2$, $\sigma = 1.5$; the first-order value $\eta\sigma^2/2\lambda$ is 0.05625, corresponding to the temperature $\beta^{-1} = \eta\sigma^2/2$. [L3] Within a basin, mixing is as fast as the curvature and does not depend on the noise; between basins it is exponentially slow in the ratio of barrier height to noise. The exponent in (b) is where a claim that noise selects wide minima has to be made quantitative.

11 Control

Proposition 11.1. [classical; [CrLi83, BaSo91]] *The Bellman operator $TV = \min_a(c_a + \gamma P_a V)$ is monotone and a contraction of factor γ in the supremum norm. A monotone, stable and consistent scheme for a Hamilton–Jacobi equation converges to its viscosity solution.*

For a random problem with 6 states and 3 actions, monotonicity and the contraction hold on 300 random pairs, with largest ratio 0.7072 at $\gamma = 0.9$, and value iteration obeys the geometric bound. [Q1] For $|u'| = 1$ on the unit interval with zero boundary values the scheme $u_i = \min(u_{i-1}, u_{i+1}) + h$ converges to the distance to the boundary exactly, and has residual $2h$ on a sawtooth and on minus the distance, which satisfy

the equation almost everywhere. [Q2] Monotonicity is the discrete maximum principle: the Bellman operator is to control what a Markov kernel is to diffusion, and the selection of the viscosity solution is the selection of the entropy solution of a conservation law [PDEHC, Thm. 14.3].

12 How many numbers an operator needs

Operator learning approximates a solution map by a finite-rank one, composed with nonlinear maps [LJPZK21, KLLABSA23]. For linear problems the best rank-r error is the $(r+1)$-st singular value.

Theorem 12.1 (Rank by type). [classical] *On the unit interval with Dirichlet conditions:*

(a) Elliptic. $(-\partial_x^2)^{-1}$ *has singular values* $(\pi k)^{-2}$*: relative accuracy* ϵ *needs* $\epsilon^{-1/2}$ *modes.*

(b) Parabolic. $e^{t\partial_x^2}$ *has singular values* $e^{-\pi^2 k^2 t}$*: accuracy* ϵ *needs* $\sqrt{1+\log(1/\epsilon)/\pi^2 t}$ *modes.*

(c) Hyperbolic. *The wave propagator is an isometry of the energy: every mode is needed.*

(d) Translates of a step function have covariance eigenvalues $((k-\frac{1}{2})\pi)^{-2}$*, so the best r-dimensional linear approximation of the family has relative error* $\sim (\pi^2 r/2)^{-1/2}$.

At $\epsilon = 10^{-6}$ this is 1000 modes for the elliptic inverse and 11 for the heat flow at $t = 0.01$ (12 on a grid of 63 points, whose stencil underestimates the higher eigenvalues as in Theorem 4.1(b)); the discrete wave propagator preserves the energy form to 10^{-15}. [O1] For the steps, the first three eigenvalues are 0.40531, 0.04505, 0.01623 on 60 points, and the relative error is 0.142 with ten modes and 0.045 with a hundred (Figure 7). [O2] A linear-reduction architecture is therefore well matched to parabolic problems, adequate for elliptic ones, and badly matched to transport, where the natural coordinates are the characteristics of Section 6 and not a fixed basis. This is the Kolmogorov barrier of reduced-order modelling [OhRa16], read off the type of the equation.

13 Summary of results

Table 1 pairs the statements. Read by column of the title, the entries on padding, resolution, flow matching and sheaves are about naturality; those on diffusion models, residual layers, physics-informed losses and control are about well-posedness; and those on message passing, kernels, noise and rank are about rates. The divisions are not sharp, because in each case the three questions are answered by the same object: a group acting on a domain and the spectrum of an operator that commutes with it.

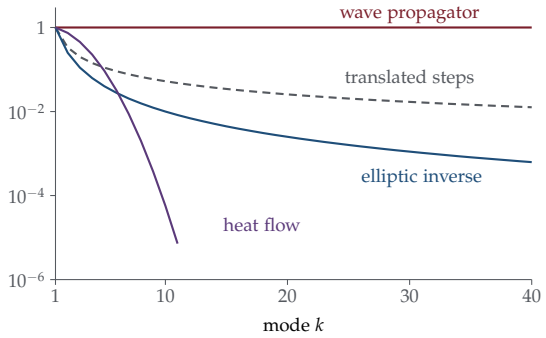


Figure 7: Singular values of solution operators on the unit interval, relative to the first, by type (Theorem 12.1): none decay for the wave equation, they decay algebraically for the elliptic inverse and for translates of a step, and faster than exponentially for the heat flow at $t = 0.01$.

14 Limits

Linearity. Most results concern the linear part of a layer or the kernel regime of training. Nonlinear layers enter only where they can be treated exactly: pointwise maps commute with even extension (Theorem 3.2), and their aliasing is computed (Theorem 4.1). Nothing is proved about the training dynamics of a finite network.

Dimension. The examples are one-dimensional or on small graphs. Theorem 3.2 extends to boxes and to the other chambers of reflection groups [PDEHC, Rem. 3.6]; Theorem 6.2 does not extend, and says so.

The score is assumed. Theorem 5.2 takes the score as given with a uniform error. How well a network estimates it, and in which norm, is the statistical question that the theorem isolates and does not answer.

Models of noise. Proposition 10.1 models stochastic gradient descent by isotropic Langevin dynamics, which is known to be a rough approximation: real gradient noise is anisotropic and state-dependent.

Checks. The script verifies instances with a fixed seed. A check that passes on a random instance supports a statement and does not prove it; the proofs are those given or cited.

Novelty. We know of no result here that a specialist in the relevant area would find new. The statements tagged as observations (Proposition 3.3, Theorem 6.2, Proposition 6.3) are elementary and may well be in the literature in other words.

What is higher. As in [PDEHC], the categorical content is modest: naturality squares, a group action with its characters, and cohomology in degree one for sheaves on graphs. The one place where a higher structure is forced, gauge fields forming a groupoid [PDEHC, Rem. 13.2], has a counterpart here in Theorem 7.3,

Topic	Qualitative statement	Quantitative statement	Checks
Padding	reflect = Neumann by images, iff kernels symmetric; zero = Dirichlet only for radius one	defects 1.18, 0.100, 0.160; mass 0.532 left after 200 zero-padded layers	N1–N6
Resolution	multipliers natural; stencils to second order; polynomials iff $N > (d+1)K$; rectifiers never	symbol 0.717, 0.980 of k^2; rectifier error $-\pi/3N^2$	A1–A4
Diffusion models	reversal is ill-posed for densities, well-posed for samples given the score	5.2×10^{21} against 0.264; score slope $-1/t + a^2/t^2$	S3–S5
Score matching, entropy	denoising = explicit + constant; de Bruijn; contraction	constant $1/t - J = 0.924$; rate 2, limit $\frac{1}{2}(\mathbb{E}X_0)^2$	S2, S6–S8
Flow matching	in one dimension, the monotone map for any coupling; commitment by a pitchfork	error 10^{-13}; $t^* = 0.382$; slope 890 at $t = 0.9$	F1–F3
Message passing	oversmoothing is the spectral gap	factor 0.92437; depth scales 415, 5, 1	G1, G2
Wave layers	conserve an energy; CFL bound	drift 2×10^{-15}; growth 1.8633 beyond	G3
Sheaf diffusion	limit is the projection to H^0; obstruction in $H^1(G;\mathbb{Z}/2)$	16 of 128; gaps 0.1606, 0.19806	G4, G5
Residual layers	Euler unstable on antisymmetric layers; leapfrog bounded gradients	1.6446, 1.0513; condition ≤ 1.667 against 8.7×10^{13}	R1–R4
Kernel regression	low modes first; sampled spectrum is a Poisson sum	ratios 1.0016, 1.0852, 1.2346; steps 6.6, 282, 956	K1, K2, K5
Physics-informed loss	squares the condition number and reverses the bias	414 against 171,682; rate ratios 1/104 and 414	K3, K4
Noise	gap = curvature in a basin, Kramers between basins	0.9997; ratios $0.903, \ldots, 0.959$	L1–L3
Control	Bellman operator monotone contraction; monotone scheme selects the viscosity solution	ratio ≤ 0.7072; residual $2h$ on the wrong solutions	Q1, Q2
Operator rank	algebraic, exponential, none, by type	1000, 11, all; steps: 0.142 with ten modes	O1, O2

Table 1: Qualitative statements with their quantitative counterparts. Every number is recomputed by the companion script.

where relabelling nodes is a gauge transformation; we have not developed it.

A The companion script

`verify-pde-learning.py` runs in about a second and prints one line per check. The checks are: **N1–N6** padding rules, their eigenbases, the image constructions and their failures, mass and the maximum principle; **A1–A4** resampling of multipliers, stencils, squares and rectifiers; **S1–S8** the heat equation for a Gaussian mixture, Tweedie's formula, the probability flow, sensitivity to the score, the two-point score, de Bruijn's identity, contraction of relative entropy, and the score-matching constant; **F1–F3** change of variables along characteristics, flow matching between Gaussians, and the two-point field with its bifurcation time; **G1–G5** oversmoothing rates, depth scales on three graphs, the wave layer, sheaf diffusion, and the count of balanced sign sheaves; **R1–R4** stability of Euler and leapfrog layers and the conditioning of their Jacobians; **K1–K5** the sampled kernel spectrum, kernel gradient descent, the two condition numbers, the reversed rates, and the interpolant as a sum of fundamental solutions; **L1–L3** Fokker–Planck gaps and the variance of stochastic gradient descent; **Q1–Q2** the Bellman operator and the monotone scheme for the eikonal equation; **O1–O2** ranks of solution operators and the spectrum of translated steps. Random inputs are drawn with seed 20261117. Tolerances are stated in the script next to each check. The curves in the figures were computed from the formulas of the text by a separate plotting script and are embedded in the source as coordinates; the points of Figure 2 for the random input are the values printed by check A4, and the signal of Figure 5 is a random one drawn for the figure.

References

[And82] B. D. O. Anderson, Reverse-time diffusion equation models, *Stochastic Process. Appl.* 12 (1982), 313–326.

[BaEm85] D. Bakry, M. Émery, Diffusions hypercontractives, in *Séminaire de Probabilités XIX*, Lecture Notes in Math. 1123, Springer, 1985, 177–206.

[BaSo91] G. Barles, P. E. Souganidis, Convergence of approximation schemes for fully nonlinear second order equations, *Asymptotic Anal.* 4 (1991), 271–283.

[BBCV21] M. M. Bronstein, J. Bruna, T. Cohen, P. Veličković, Geometric deep learning: grids, groups, graphs, geodesics, and gauges, arXiv:2104.13478, 2021.

[BBDM24] G. Biroli, T. Bonnaire, V. de Bortoli, M. Mézard, Dynamical regimes of diffusion models, *Nature Commun.* 15 (2024).

[BBRMMA23] F. Bartolucci, E. de Bézenac, B. Raonić, R. Molinaro, S. Mishra, R. Alaifari, Representation equivalent neural operators: a framework for alias-free operator learning, *Adv. Neural Inf. Process. Syst.* 36 (2023); arXiv:2305.19913. Checked.

[BDCLB22] C. Bodnar, F. Di Giovanni, B. P. Chamberlain, P. Liò, M. M. Bronstein, Neural sheaf diffusion: a topological perspective on heterophily and oversmoothing in GNNs, *Adv. Neural Inf. Process. Syst.* 35 (2022); arXiv:2202.04579. Checked.

[CMHRBH18] B. Chang, L. Meng, E. Haber, L. Ruthotto, D. Begert, E. Holtham, Reversible architectures for arbitrarily deep residual neural networks, *Proc. AAAI* 32 (2018).

[CoWe16] T. Cohen, M. Welling, Group equivariant convolutional networks, *Proc. ICML* 2016, 2990–2999.

[CRBD18] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, D. Duvenaud, Neural ordinary differential equations, *Adv. Neural Inf. Process. Syst.* 31 (2018).

[CRGBB21] B. P. Chamberlain, J. Rowbottom, M. Gorinova, S. Webb, E. Rossi, M. M. Bronstein, GRAND: graph neural diffusion, *Proc. ICML* 2021, 1407–1418.

[CrLi83] M. G. Crandall, P.-L. Lions, Viscosity solutions of Hamilton–Jacobi equations, *Trans. Amer. Math. Soc.* 277 (1983), 1–42.

[DCWJXKY26] Dai, Chen, Wang, Jia, Xie, Kumar, Yu, Learning PDE solvers with physics and data: a unifying view of physics-informed neural networks and neural operators, arXiv:2601.14517, 2026. Checked; surnames only were recorded.

[E17] W. E, A proposal on machine learning via dynamical systems, *Commun. Math. Stat.* 5 (2017), 1–11.

[Efr11] B. Efron, Tweedie's formula and selection bias, *J. Amer. Statist. Assoc.* 106 (2011), 1602–1614.

[EYu18] W. E, B. Yu, The deep Ritz method: a deep learning-based numerical algorithm for solving variational problems, *Commun. Math. Stat.* 6 (2018), 1–12.

[HaGh19] J. Hansen, R. Ghrist, Toward a spectral theory of cellular sheaves, *J. Appl. Comput. Topol.* 3 (2019), 315–358.

[HaRu17] E. Haber, L. Ruthotto, Stable architectures for deep neural networks, *Inverse Problems* 34 (2017), 014004.

[HJA20] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, *Adv. Neural Inf. Process. Syst.* 33 (2020).

[Hyv05] A. Hyvärinen, Estimation of non-normalized statistical models by score matching, *J. Mach. Learn. Res.* 6 (2005), 695–709.

[ILB20] M. A. Islam, S. Jia, N. D. B. Bruce, How much position information do convolutional neural networks encode?, *Proc. ICLR* 2020.

[JGH18] A. Jacot, F. Gabriel, C. Hongler, Neural tangent kernel: convergence and generalization in neural networks, *Adv. Neural Inf. Process. Syst.* 31 (2018).

[KALHHLA21] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, T. Aila, Alias-free generative adversarial networks, *Adv. Neural Inf. Process. Syst.* 34 (2021).

[KGZKM21] A. S. Krishnapriyan, A. Gholami, S. Zhe, R. M. Kirby, M. W. Mahoney, Characterizing possible failure modes in physics-informed neural networks, *Adv. Neural Inf. Process. Syst.* 34 (2021).

[KLLABSA23] N. Kovachki, Z. Li, B. Liu, K. Azizzadenesheli, K. Bhattacharya, A. Stuart, A. Anandkumar, Neural operator: learning maps between function spaces with applications to PDEs, *J. Mach. Learn. Res.* 24 (2023), no. 89.

[Kra40] H. A. Kramers, Brownian motion in a field of force and the diffusion model of chemical reactions, *Physica* 7 (1940), 284–304.

[LCBNL23] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, M. Le, Flow matching for generative modeling, *Proc. ICLR* 2023.

[LGL23] X. Liu, C. Gong, Q. Liu, Flow straight and fast: learning to generate and transfer data with rectified flow, *Proc. ICLR* 2023.

[LHW18] Q. Li, Z. Han, X.-M. Wu, Deeper insights into graph convolutional networks for semi-supervised learning, *Proc. AAAI* 32 (2018).

[LJPZK21] L. Lu, P. Jin, G. Pang, Z. Zhang, G. E. Karniadakis, Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators, *Nature Mach. Intell.* 3 (2021), 218–229.

[LKALBSA21] Z. Li, N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, A. Anandkumar, Fourier neural operator for parametric partial differential equations, *Proc. ICLR* 2021.

[MHB17] S. Mandt, M. D. Hoffman, D. M. Blei, Stochastic gradient descent as approximate Bayesian inference, *J. Mach. Learn. Res.* 18 (2017), no. 134.

[OhRa16] M. Ohlberger, S. Rave, Reduced basis methods: success, limitations and future challenges, *Proc. ALGORITMY* 2016, 1–12.

[OoSu20] K. Oono, T. Suzuki, Graph neural networks exponentially lose expressive power for node classification, *Proc. ICLR* 2020.

[Ors71] S. A. Orszag, On the elimination of aliasing in finite-difference schemes by filtering high-wavenumber components, *J. Atmos. Sci.* 28 (1971), 1074.

[PDEHC] Anonymous, *Partial Differential Equations in Higher Categories*, Higher Categorical Semantics, 2026. Companion volume.

[RBADLHBC19] N. Rahaman, A. Baratin, D. Arpit, F. Draxler, M. Lin, F. Hamprecht, Y. Bengio, A. Courville, On the spectral bias of neural networks, *Proc. ICML* 2019, 5301–5310.

[RCRMB22] T. K. Rusch, B. P. Chamberlain, J. Rowbottom, S. Mishra, M. M. Bronstein, Graph-coupled oscillator networks, *Proc. ICML* 2022, 18888–18909.

[RPK19] M. Raissi, P. Perdikaris, G. E. Karniadakis, Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations, *J. Comput. Phys.* 378 (2019), 686–707.

[SSKKEP21] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, B. Poole, Score-based generative modeling through stochastic differential equations, *Proc. ICLR* 2021.

[Sta59] A. J. Stam, Some inequalities satisfied by the quantities of information of Fisher and Shannon, *Inform. and Control* 2 (1959), 101–112.

[Str08] W. A. Strauss, *Partial Differential Equations: An Introduction*, 2nd ed., Wiley, 2008.

[Str99] G. Strang, The discrete cosine transform, *SIAM Rev.* 41 (1999), 135–147.

[SWMG15] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, S. Ganguli, Deep unsupervised learning using nonequilibrium thermodynamics, *Proc. ICML* 2015, 2256–2265.

[Vin11] P. Vincent, A connection between score matching and denoising autoencoders, *Neural Comput.* 23 (2011), 1661–1674.

[WTP21] S. Wang, Y. Teng, P. Perdikaris, Understanding and mitigating gradient flow pathologies in physics-informed neural networks, *SIAM J. Sci. Comput.* 43 (2021), A3055–A3081.