Adaptive Spectral Trust Gate for Physics-Constrained Operator Learning

Soham Chakraborty

Adaptive Spectral Trust Gate for Physics-Constrained Operator Learning

DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Technology
in
Computer Science

by

Soham Chakraborty

[Roll No: CS-2432]

under the guidance of

Prof. Swagatam Das

Professor
Electronics and Communication Sciences Unit

Indian Statistical Institute
Kolkata-700108, India

July 2026

To my parents and my guide

"Imagination is more important than Knowledge"

- Albert Einstein

CERTIFICATE

This is to certify that the dissertation entitled **"Adaptive Spectral Trust Gate for Physics-Constrained Operator Learning"** submitted by **Soham Chakraborty** to Indian Statistical Institute, Kolkata, in partial fulfillment for the award of the degree of **Master of Technology in Computer Science** is a bonafide record of work carried out by him under my supervision and guidance. The dissertation has fulfilled all the requirements as per the regulations of this institute and, in my opinion, has reached the standard needed for submission.

Prof. Swagatam Das
Professor,
Electronics and Communication Sciences Unit,
Indian Statistical Institute,
Kolkata-700108, INDIA.

Acknowledgments

I would like to show my highest gratitude to my advisor, *Prof. Swagatam Das*, Electronics and Communication Sciences Unit, Indian Statistical Institute, Kolkata, along with Mr. Dhiraj Nagaraja Hegde, Senior Researcher at Ericsson Research for his guidance and continuous support and encouragement. Both of them have literally taught me how to do good research, and motivated me with great insights and innovative ideas.

My deepest thanks to all the teachers of Indian Statistical Institute, for their valuable suggestions and discussions which added an important dimension to my research work.

Finally, I am very much thankful to my parents and family for their everlasting supports.

Last but not the least, I would like to thank all of my friends for their help and support. I thank all those, whom I have missed out from the above list.

Soham Chakraborty
Indian Statistical Institute
Kolkata - 700108 , India.

Abstract

Physics-informed machine learning improves the plausibility, data-efficiency and generalization of surrogate models by injecting prior physical knowledge into the learning process. The current approaches can be broadly divided into two main categories: soft constraints, which add a physics residual to the training loss but guarantee nothing at inference time, and hard constraints, which project the model output onto the constraint set exactly but apply the projection uniformly to every part of the signal — including parts that are dominated by noise, discretization error, or model mismatch, where the idealized physics is not actually trustworthy.

This dissertation proposes the Adaptive Spectral Trust Gate (ASPINO), a mechanism that learns where to trust the physics. Operating in the Fourier domain on top of any surrogate model, a small gating network forms a per-mode convex combination of a data-driven soft path and a physics hard path. The gate is driven by features of the spectral coordinate and the spectral amplitude, so that it can apply the hard constraint in well-conditioned spectral regions and defer to the data-driven operator in regions corrupted by noise or aliasing. A single gate serves two very different hard paths - the linear Leray projection (incompressible flow) and a nonlinear rank-r SVD projection (massive-MIMO channel estimation).

On the theoretical side, we give an empirical-Rademacher-complexity analysis: an unconditional safety floor — the gated class never exceeds the soft path it wraps — and, under a stated low-rank-transfer assumption, a capacity-reduction factor of $1-\bar{\alpha}\,(1-\sqrt{\rho_r})$, where $\bar{\alpha}$ is the fraction of capacity routed through the hard path. On Kolmogorov-flow denoising ASPINO is simultaneously the most accurate and near-physical, dominating the unconstrained, hard and soft baselines; on ray-traced MIMO it improves a strong physics-informed baseline across all pilot budgets and SNRs without ever regressing. A third study, zero-shot super-resolution on the Poisson equation, confirms the discretization invariance of the gated construction. ASPINO is discretization-invariant and "plug-and-play" over the underlying operator.

Keywords: *Physics-Informed learning, Neural Operators, Hard Constraints, Fourier projection, Spectral gating, Rademacher Complexity, Channel Estimation.*

Contents

List of Figures

List of Tables

Table 1: List of principal notation used in this thesis.

$f \in \mathcal{F}$	data-driven soft path and its hypothesis class
$\mathcal{P}_r$	exact physics hard path (Leray / rank-r SVD)
$g \in \mathcal{G}$	spectral gate, $g(k) \in [0,1]$, and its class
k	Fourier mode index
$\hat{f}(k)$	Fourier coefficient of the soft path at mode k
$\mathcal{H}_{\text{gate}}$	gated hypothesis class
$\bar{g}(k) = \mathbb{E}[g(k)]$	mean gate value at mode k
$\bar{\alpha}$	capacity fraction routed through the hard path
$\mathcal{V}_r$	variety of rank-$\leq r$ matrices
$d_r = 2r(m+n-r)$	real dimension of $\mathcal{V}_r$
$\rho_r = d_r/d_{\text{full}}$	capacity ratio; $\sqrt{\rho_r}$ the reduction factor
$\widehat{\mathfrak{R}}_n(\cdot)$	empirical Rademacher complexity
$\|\cdot\|_F,\ \|\cdot\|_{\text{op}}$	Frobenius and operator norms
$\zeta,\ \mu_\beta$	soft-penalty and gate-regularizer weights

Chapter 1

Introduction

1.1 Introduction

Partial differential equations (PDEs) are the language in which much of physics, engineering and the applied sciences are written. They model the flow of fluids, the propagation of electromagnetic waves, the diffusion of heat and the dynamics of countless other systems. Classical numerical solvers for these equations — finite differences, finite elements, spectral methods — are accurate and well understood, but they become prohibitively expensive on large-scale problems, and they must be re-run from scratch for every new input. This computational bottleneck has motivated a large body of work on learning solution operators of PDEs directly from data, so that the expensive solve is amortized into a single training phase and inference becomes a cheap forward pass through a network [1, 3, 4].

Among the data-driven approaches, neural operators [3, 4] have emerged as a particularly powerful tool. Unlike ordinary neural networks, which learn a map between finite-dimensional vectors tied to a fixed grid, neural operators learn a map between function spaces. As a consequence they are discretization invariant: a single trained model can accept inputs sampled at one resolution and produce outputs queried at another, and can even perform "zero-shot" super-resolution. The Fourier Neural Operator (FNO) [3] realizes this idea efficiently by parameterizing the integral kernel in the Fourier domain, where convolution becomes pointwise multiplication.

Purely data-driven surrogates, however, are only as good as their training data. In the low-data or low-resolution regime they may produce predictions that fit the data statistically yet violate the very physics that generated it [6]. The natural remedy is to supplement the data with prior physical knowledge — conservation laws, symmetries, governing equations. This is the central idea of physics-informed machine learning [1, 2]. Physics-Informed Neural Networks (PINNs) [1] were the landmark proposal: they add the PDE residual, evaluated at a set of collocation points, as an extra term in the training loss. The same idea carries over to operator learning, where a physics residual regularizes the learned operator [6, 7].

Constraints of this kind are called *soft*, because they are imposed as a penalty rather than enforced exactly. Soft constraints are simple and architecture-agnostic, but they come with well-documented drawbacks. The composite loss can be harder to optimize and worse conditioned; the physics residual involves derivatives that are expensive and sometimes inaccurate to approximate; the two loss terms can conflict; and, crucially, a soft constraint provides no guarantee that the physics holds at inference time [17, 18, 8]. For high-stakes applications, where physical plausibility must be guaranteed rather than encouraged, this is a serious limitation.

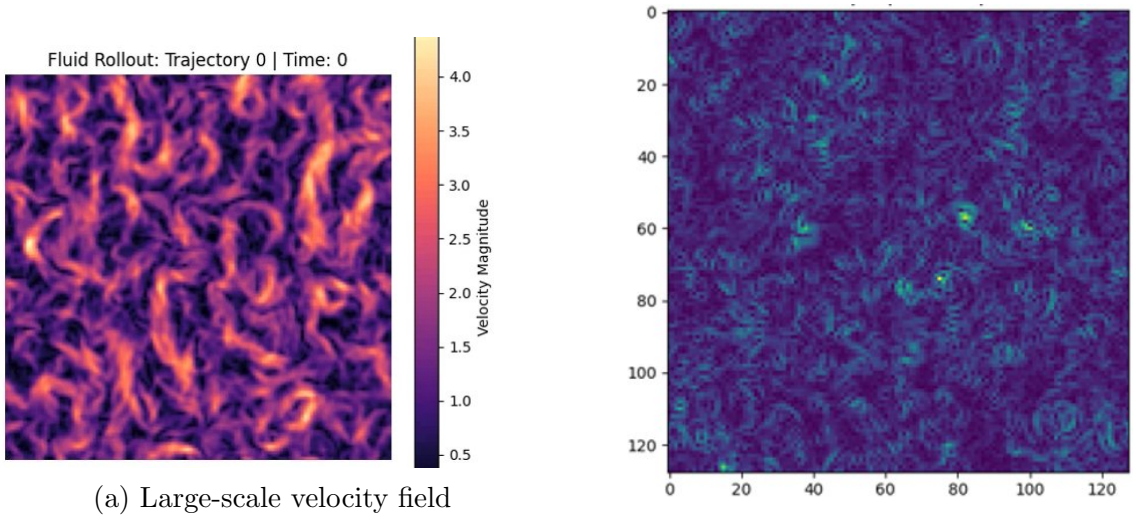
These shortcomings motivate *hard* constraints, which modify the model so that its output satisfies the physics exactly. A clean and general way to do this is to project the surrogate's output onto the constraint set. Duruisseaux et al. [8] show that for linear differential constraints this projection can be computed very efficiently in Fourier space: enforcing the divergence-free condition of the incompressible Navier–Stokes equations reduces to a least-norm problem with a linear constraint, whose closed-form solution is the Leray projection. Their projection layer enforces the constraint at every point of the domain to arbitrary accuracy, is compatible with any operator architecture, and costs only a small constant factor over the unconstrained model.

There is, however, a subtlety that the hard-constraint literature has largely left implicit. A hard projection is applied *uniformly* to the entire signal. It treats every spectral mode as equally trustworthy and forces all of them onto the idealized constraint manifold. But the idealized physics is not equally trustworthy everywhere. In practice the data are corrupted by measurement noise, by numerical and discretization error, and by the inevitable mismatch between a clean mathematical model and a messy real environment. In the spectral regions dominated by these effects — typically the high-frequency and Nyquist modes — forcing the output onto the constraint manifold can inject bias rather than remove it, degrading accuracy even while the residual looks perfect. Concretely, additive noise has a divergence-free component that the projection cannot remove; at high modes, where little true signal remains, enforcing the constraint mainly re-introduces structured noise that a good data-driven estimator would otherwise suppress. A purely hard model is, in this sense, "stiff": it is perfectly physical but brittle against noise. A purely soft model is flexible but unconstrained. Neither is uniformly correct.

This dissertation argues that the right object to learn is not *whether* to apply the physics, but *where*. We propose a mechanism that decides, mode by mode in the Fourier domain, how much to trust the hard constraint — applying it where the physics is reliable and deferring to the data-driven operator where it is not.

1.2 Motivating Examples

Two concrete settings, drawn from very different parts of physics, will run through this dissertation and illustrate why an adaptive treatment of physics is needed.



(a) Large-scale velocity field

(b) Small-scale fluctuation

Figure 1.1: Two scales of an incompressible turbulent flow. The large, energy-bearing scales (a) are well resolved and the divergence-free physics is reliable there; the fine-scale fluctuations and residuals (b) are where measurement noise and dealiasing artefacts concentrate and the idealized physics is least trustworthy.

Incompressible fluid flow. The velocity field of an incompressible fluid is divergence-free, $\nabla \cdot u = 0$. This is a linear differential constraint, and the Leray projection enforces it exactly in Fourier space [8]. When the observations are clean this is exactly the right thing to do. But consider a denoising task in which the input flow field has been corrupted by additive noise. The large-scale, low-wavenumber modes of a turbulent field carry most of the energy and are well resolved; the physics there is reliable. The high-wavenumber and Nyquist modes, on the other hand, are where the noise and the 2/3-dealiasing artefacts concentrate. A uniform Leray projection cannot tell these regions apart; an adaptive gate can learn to enforce the constraint on the trustworthy large scales and to ignore it on the noise-dominated modes.

Massive-MIMO channel estimation. In next-generation wireless systems a base station with a large antenna array must estimate the propagation channel from a small number of pilot signals. The physics here is Maxwell's equations, and a physics-informed network can exploit a received signal-strength (RSS) map computed by ray tracing as a prior [22]. The antenna-aperture response of a sparse multipath channel is approximately low-rank, so a natural hard constraint is to project the aperture onto the set of rank-r matrices via a truncated SVD. Again this is helpful where the low-rank model is accurate and harmful where it is not: at very low pilot counts, or in spectral regions dominated by interference, the rank-r assumption is only approximately true, and forcing it everywhere costs accuracy. An adaptive gate can recover the benefit of the low-rank physics without paying its bias in the unreliable

regions.

The two hard projections could hardly be more different — one is a fixed linear operator, the other a nonlinear, data-dependent projection onto a curved variety — yet, as we will show, the same gating principle applies to both, and the same generalization analysis explains both.

1.3 Our Contribution

The contributions of this dissertation are as follows.

- **A spectral trust gate (ASPINO).** We propose a mechanism that learns, per Fourier mode, a convex combination of a data-driven soft path and a physics hard path. The gate is a small network driven by features of the spectral coordinate and the spectral amplitude, so that it can localize trust in well-conditioned spectral regions. The mechanism is architecture-agnostic, operates on top of any neural operator, and inherits the discretization invariance of the underlying operator. A Wasserstein gate regularizer encourages the per-mode trust weights to specialize across the spectrum rather than collapse to a single global value; we find it essential for the gate to localize trust in the denoising experiment of Chapter 6.

- **A unified treatment of two hard paths.** We instantiate the hard path both as the linear Leray projection (incompressible flow) and as the nonlinear rank-r truncated-SVD projection (MIMO channel estimation), and show that a single gating design serves both.

- **A generalization analysis.** We give an empirical-Rademacher-complexity analysis of the gated hypothesis class. Its unconditional core is a *mixing inequality*: the gated class is never more complex than the gate-weighted average of the soft and hard paths, plus a vanishing gate term. From it we derive a *safety floor* (the gate does not increase complexity beyond the soft path on the clamp-inactive regime) and, under one explicitly stated low-rank-transfer assumption, a *capacity-reduction bound* showing the complexity contracts by a factor $1 - \bar{\alpha}(1 - \sqrt{\rho_r})$.

- **Empirical validation in two domains.** On Kolmogorov-flow denoising, ASPINO attains the best test accuracy and a near-physical PDE residual, dominating the unconstrained, hard and soft baselines at once. On realistic ray-traced MIMO channel estimation, it improves over a strong physics-informed baseline across all pilot budgets and SNRs while never regressing below it, recovering a consistent fraction of the oracle gain.

1.4 Dissertation Outline

The remainder of this dissertation is organized as follows: Chapter 2 collects the preliminaries: notation, PINNs and neural operators, the Fourier projection/ Leray projection, the SVD and low-rank geometry, the Wasserstein distance, and the empirical Rademacher complexity toolbox. Chapter 3 surveys the related literature on soft and hard physics constraints and on physics-informed channel estimation, and locates our contribution within it. Chapter 4 describes the proposed Adaptive Spectral Trust Gate in detail, including the gate architecture, the two hard paths, the loss and the gate regularizer, and the full training algorithm. Chapter 5 develops the generalization analysis: the model and notation, the supporting lemmas, the mixing inequality, the safety floor and capacity-reduction bound, a four-regime comparison, and a reconciliation with the measured numbers. Chapter 6 presents the experiments on Kolmogorov-flow denoising and on MIMO channel estimation, together with a zero-shot super-resolution study on the Poisson equation, with ablations and a visualization of the learned gate. Chapter 7 concludes and discusses directions for future work.

Chapter 2

Preliminaries

This chapter fixes notation and collects the background needed in the rest of the dissertation: physics-informed neural networks, neural operators and the Fourier neural operator, the Fourier projection that yields the Leray operator, the singular-value decomposition and the geometry of low-rank matrices, the Wasserstein distance, and the empirical Rademacher complexity together with the three classical results we use.

2.1 Notation

For a matrix $A \in \mathbb{C}^{m\times n}$, $\|A\|_F$ denotes the Frobenius norm, $\|A\|_{\mathrm{op}}$ the operator (spectral) norm, $\sigma_i(A)$ the i-th largest singular value, and $\operatorname{rank}(A)$ the rank. The discrete Fourier transform (DFT) is written F (or F_m when the size m matters); on a uniform grid it is realized by the fast Fourier transform (FFT). A sample is $S = \{x_i\}_{i=1}^n$; $\sigma_i \in \{\pm 1\}$ are independent Rademacher signs; $\mathbb{E}[\cdot]$ is expectation. We write $\widehat{\mathfrak{R}}_n(\mathcal{H})$ for the empirical Rademacher complexity of a class $\mathcal{H}$ (Section 2.8). A wireless channel is represented as a complex tensor; the per-tap antenna aperture is a matrix $A \in \mathbb{C}^{m\times n}$. A velocity field is $u(x,t)$ on a periodic spatial domain, and $\nabla \cdot u$ is its divergence.

2.2 Physics-Informed Neural Networks

Consider a PDE on a domain Ω with boundary/initial operators, $\mathcal{N}[u](x) = 0$ in Ω and $\mathcal{B}[u](x) = 0$ on $\partial\Omega$, whose solution u we approximate by a network u_θ. The physics-informed approach of Raissi, Perdikaris and Karniadakis [1] trains u_θ by minimizing a composite loss

$$\mathcal{L}(\theta) = \frac{1}{N_d}\sum_i \big\|u_\theta(x_i) - u_i\big\|^2 + \zeta\,\frac{1}{N_c}\sum_j \big\|\mathcal{N}[u_\theta](x_j)\big\|^2, \tag{2.1}$$

where the residual is evaluated at collocation points $\{x_j\}$ (typically by automatic differentiation) and $\zeta > 0$ weights the physics term. The penalty encodes the governing equation as a *soft* constraint: it pushes the network toward physically consistent solutions but does not guarantee $\mathcal{N}[u_\theta] = 0$ at inference. A body of work documents PINN training pathologies: stiff and conflicting gradients, poor conditioning, sensitivity to ζ, and spurious minimizers that satisfy the residual at the collocation points without solving the PDE elsewhere [17, 18, 19].

A bias–variance reading. The benefit of the physics term in the data-scarce regime admits a simple bias–variance reading [22]. Without constraints a data-driven estimator searches a high-dimensional hypothesis space, so its variance scales like $\mathrm{Var}[u_\theta] \propto d_{\text{eff}}/M$ with d_{eff} the effective model dimension and M the number of observations. Enforcing $\mathcal{N}[u_\theta] \approx 0$ restricts the search to a lower-dimensional manifold, reducing d_{eff} and hence the variance, at the price of a small bias when the physical model is imperfect. Quantifying this trade-off rigorously, in the harder operator-learning setting, is the subject of Chapter 5.

2.3 Neural Operators and the Fourier Neural Operator

Let $\mathcal{A}$ and $\mathcal{U}$ be Banach spaces of functions. The operator-learning problem is to learn the solution operator $\mathcal{G} : \mathcal{A} \to \mathcal{U}$ of a parametric PDE, mapping an input function a to the solution u. A neural operator [4] composes pointwise lifting and projection maps P, Q with L layers that interleave a pointwise linear map, an integral kernel operator $\mathcal{K}_\ell$, a bias, and a nonlinearity σ:

$$\mathcal{G}_\theta = Q \circ (W_L + \mathcal{K}_L + b_L) \circ \cdots \circ \sigma(W_1 + \mathcal{K}_1 + b_1) \circ P. \tag{2.2}$$

Each $\mathcal{K}_\ell$ acts on functions, not on grid vectors, so the model is discretization invariant: it can be evaluated at any resolution and queried at arbitrary points. The Fourier Neural Operator (FNO) [3] chooses the kernel to be a convolution and computes it in the spectral domain,

$$(\mathcal{K}(\phi)v)(x) = F^{-1}\big(R_\phi \cdot (Fv)\big)(x), \tag{2.3}$$

where R_ϕ is a learned complex multiplier applied to the lowest Fourier modes and F is realized by the FFT. Because the FNO already works mode-by-mode in Fourier space, it is the natural backbone for a spectral gate: the place where the operator combines information across modes is exactly the place where we will insert the trust gate.

2.4 Hard Constraints by Fourier Projection: the Leray Operator

Given a surrogate $\tilde{\mathcal{G}}$ whose output need not satisfy a constraint $\mathcal{C}(u) = 0$, one can define a constrained operator by composing it with a projection onto the constraint set, $\mathcal{G}^*(a) = (\mathrm{pr}_{\mathcal{C}} \circ \tilde{\mathcal{G}})(a)$. For a linear differential constraint, Duruisseaux et al. [8] show the projection is cheapest in Fourier space, where differentiation becomes pointwise multiplication and the constraint becomes a linear system on the Fourier coefficients. For the incompressible Navier–Stokes equations the constraint is the divergence-free condition $\nabla \cdot u = 0$, and the resulting projection is the classical Leray projection. On the 2-D torus, taking the Fourier transform of $\nabla \cdot u = 0$ gives

$$\xi_1 \, \hat{u}_1(\xi) + \xi_2 \, \hat{u}_2(\xi) = 0 \quad \text{for all } \xi = (\xi_1, \xi_2), \tag{2.4}$$

i.e. the Fourier coefficient of the velocity must be orthogonal to the wavevector ξ. Enforcing this is the per-mode orthogonal projection onto the plane perpendicular to ξ,

$$P(\xi) = I - \frac{\xi \xi^\top}{|\xi|^2}, \qquad \hat{u}(\xi) \;\mapsto\; P(\xi)\, \hat{u}(\xi), \tag{2.5}$$

which removes the component of $\hat{u}(\xi)$ along ξ (and we set $P(0) = I$ for the mean mode). This is the classical Leray projection: a fixed, linear, per-mode operator that deletes one of the D velocity directions, so $\operatorname{tr} P(\xi) = D - 1$ and the per-mode degrees of freedom contract by $(D-1)/D = \frac{1}{2}$ in two dimensions. This linearity is what lets us describe its effect by a per-mode contraction factor in Chapter 5.

2.5 The Wireless MIMO Channel Model

Signal propagation is governed by Maxwell's equations; under the time-harmonic assumption these reduce to an inhomogeneous wave equation for the electric field, whose Green's-function solution expresses the field as a superposition of contributions from source currents and medium inhomogeneities [22]. The total received power at a location is proportional to the squared magnitude of the total field, and this quantity, precomputed by a ray tracer, is the received-signal-strength (RSS) map that the physics-informed estimator uses as a prior. The frequency-selective MIMO channel is a 3-tensor $H \in \mathbb{C}^{D \times N_r \times N_t}$, one matrix slice H_d per delay tap, each a superposition of steering-vector outer products:

$$H_d = \sum_{\ell=1}^{P} \alpha_\ell \, f_p\big(dT_s - (t_\ell - t_{\mathrm{off}})\big) \, a_r(\theta_\ell) \, a_t(\phi_\ell)^*, \tag{2.6}$$

with α_ℓ the complex path gain, t_ℓ the time-of-arrival, and a_r, a_t the receive/transmit array responses. Each term $a_r(\theta_\ell)\, a_t(\phi_\ell)^*$ is an outer product of two vectors and is

therefore a rank-one matrix. With only P dominant paths, the aperture H_d is a sum of P such rank-one terms, so $\operatorname{rank}(H_d) \le P$; when the number of resolvable paths is small, $P \ll \min(N_r, N_t)$, the aperture is accordingly approximately low rank. This is the structural fact the rank-r SVD hard path exploits: projecting the estimated aperture onto the rank-r variety enforces the multipath-sparsity physics, just as the Leray projection enforces incompressibility. The estimation problem is to recover H from a small number N_p of pilot observations, which in the pilot-limited regime is severely under-determined — exactly the regime in which a physics prior pays off.

2.6 The Singular Value Decomposition and Low-Rank Geometry

Every matrix $A \in \mathbb{C}^{m\times n}$ admits a singular value decomposition $A = \sum_{i=1}^{\min(m,n)} \sigma_i u_i v_i^*$ with $\sigma_1 \ge \sigma_2 \ge \cdots \ge 0$. The Eckart–Young theorem states that the best rank-r approximation in Frobenius (or operator) norm is the truncation

$$P_r^{(1)}(A) = \sum_{i=1}^{r} \sigma_i u_i v_i^*, \qquad \|A\|_F^2 - \|P_r^{(1)}(A)\|_F^2 = \sum_{i>r} \sigma_i^2 \ge 0. \tag{2.7}$$

Truncation therefore never increases energy. Unlike the Leray projection, $P_r^{(1)}$ is a nonlinear, data-dependent map: its singular vectors move with the input. The image of $P_r^{(1)}$ lives on the set of rank-$\le r$ matrices, $\mathcal{V}_r = \{A \in \mathbb{C}^{m\times n} : \operatorname{rank}(A) \le r\}$. This determinantal variety is not a linear subspace but a curved manifold (away from its singular locus). Writing a rank-r matrix as $A = UV^*$ with $U \in \mathbb{C}^{m\times r}$, $V \in \mathbb{C}^{n\times r}$ uses $2mr + 2nr$ real parameters, but this factorization is redundant: for any invertible $G \in \mathrm{GL}_r(\mathbb{C})$ the replacement $U \mapsto UG^{-1}$, $V \mapsto VG^*$ leaves the product UV^* unchanged. Modding out this gauge freedom—the action of $\mathrm{GL}_r(\mathbb{C})$, which has real dimension $2r^2$—gives the real dimension of the rank-r variety on its smooth stratum,

$$d_r = 2(mr + nr) - 2r^2 = 2r(m + n - r). \tag{2.8}$$

For comparison the full aperture has real dimension $d_{\text{full}} = 2mn$. The dimension ratio $\rho_r = d_r/d_{\text{full}} = r(m + n - r)/mn$ will control the capacity saved by routing a signal through the low-rank path. For a 24×24 aperture with $r = 3$, $d_r = 270$, $d_{\text{full}} = 1152$, so $\rho_3 = 0.234$ and $\sqrt{\rho_3} \approx 0.484$.

Two further facts about SVD truncation will be needed. First, on its smooth stratum the factored parametrization admits, within a Frobenius ball of radius B, a covering of size $(9B/\epsilon)^{d_r}$, so $\log N(\epsilon, \mathcal{V}_r \cap B\mathbb{B}, \|\cdot\|_F) \le d_r \log(9B/\epsilon)$ [32]. Second, by the Wedin / Davis–Kahan $\sin\Theta$ theorem, if the spectral gap $\sigma_r(A) - \sigma_{r+1}(A) \ge \delta > 0$ then $P_r^{(1)}$ is Lipschitz with constant $1 + 2\|A\|_{\text{op}}/\delta$ on that region. A small jitter added before the SVD makes the singular values distinct with probability one, so the gap holds almost surely and SVD back-propagation is well defined.

2.7 The Wasserstein Distance

To regularize the distribution of gate values we use the 1-Wasserstein (earth-mover) distance. For probability measures μ, ν on $\mathbb{R}$,

$$W_1(\mu,\nu) = \inf_{\gamma\in\Pi(\mu,\nu)} \mathbb{E}_{(x,y)\sim\gamma}|x-y| = \sup_{\|h\|_{\mathrm{Lip}}\le 1} \mathbb{E}_\mu[h] - \mathbb{E}_\nu[h], \tag{2.9}$$

the second equality being Kantorovich–Rubinstein duality. The dual form is the one used in Wasserstein GANs [33], where the supremum is taken over a 1-Lipschitz critic enforced by spectral normalization. The relevance of spectral normalization is that it bounds the operator norms of the critic's weights, which feeds directly into the Rademacher size bound of the next section.

2.8 Empirical Rademacher Complexity

For a real-valued function class $\mathcal{H}$ and a sample $S = \{x_i\}_{i=1}^n$, the empirical Rademacher complexity is

$$\widehat{\mathfrak{R}}_n(\mathcal{H}) = \frac{1}{n}\,\mathbb{E}_\sigma\Big[\sup_{h\in\mathcal{H}} \sum_{i=1}^n \sigma_i\, h(x_i)\Big], \tag{2.10}$$

with σ_i i.i.d. Rademacher signs. It measures how well the class can correlate with random noise; smaller complexity means tighter generalization. We use three classical results.

(A) Contraction lemma (Talagrand).
If $\psi_1,\dots,\psi_n$ are μ-Lipschitz, then $\widehat{\mathfrak{R}}_n(\psi\circ\mathcal{H}) \le \mu\,\widehat{\mathfrak{R}}_n(\mathcal{H})$ [30]. In particular a convex combination of two classes is 1-Lipschitz and contracts.

(B) Spectral-norm (size) bound.
For an L-layer network with weight matrices $\{W_\ell\}$, ρ-Lipschitz activations and inputs bounded by B_x,

$$\widehat{\mathfrak{R}}_n(\mathcal{H}) \le \frac{c\,B_x\,\rho^L\sqrt{L\log(2d)}\,\prod_{\ell=1}^L \|W_\ell\|_{\mathrm{op}}}{\sqrt{n}} = O(n^{-1/2}) \tag{2.11}$$

for fixed trained weights [28, 29]. Spectral normalization controls the product $\prod_\ell \|W_\ell\|_{\mathrm{op}}$ directly.

(C) Dudley's entropy integral.
Let $\mathcal{H}|_S = \{(h(x_1),\dots,h(x_n)) : h\in\mathcal{H}\}$ be the class *restricted to the sample*, measured in the empirical L_2 metric $\|\cdot\|_n$. If $\mathcal{H}|_S$ has radius B in this metric, then

$$\widehat{\mathfrak{R}}_n(\mathcal{H}) \le \frac{12}{\sqrt{n}} \int_0^B \sqrt{\log N(\epsilon, \mathcal{H}|_S, \|\cdot\|_n)}\, d\epsilon. \tag{2.12}$$

This is the tool for the *nonlinear* low-rank path. We emphasize that the covering must be of the class on the sample, not of the range of a single output; the distinction is taken up carefully in Chapter 5.

A vector-valued refinement [31] lets us treat the real and imaginary parts of complex Fourier outputs as separate real coordinates and sum their contributions; we use this implicitly when we decompose complexity mode-by-mode in Chapter 5.

Chapter 3

Related Work and Our Contribution

This chapter surveys the work most directly related to the Adaptive Spectral Trust Gate. We organize it around the central distinction of this dissertation — soft versus hard physics constraints — and then turn to the two application domains that motivate our experiments. Throughout, we highlight the gap our method fills: existing methods decide *whether* to enforce a constraint globally, whereas we learn *where* to enforce it.

3.1 Physics-Informed Neural Networks

Physics-Informed Neural Networks [1] are the canonical soft-constraint method: they introduce physics into a neural network through a residual penalty in the loss. The framework has been enormously influential and is surveyed in [2]. Subsequent work has both extended PINNs [20] and dissected their failure modes. Krishnapriyan et al. [17] characterize how the physics loss can create optimization landscapes that gradient descent cannot navigate; Wang et al. [18, 19] analyze gradient pathologies and the neural-tangent-kernel behaviour that makes some PINNs fail to train. Structural modifications enforce certain boundary conditions exactly by construction [16], an early form of hard constraint. The common thread is that a soft penalty regularizes the loss, not the hypothesis class; we make this precise in Chapter 5, where the soft penalty leaves the Rademacher complexity unchanged.

3.2 Neural Operators

Neural operators learn maps between function spaces and are discretization invariant [3, 4]. The Fourier Neural Operator [3] is the most widely used instance and the backbone for our experiments; DeepONet [5] is an alternative family. Physics has been

added to operators in two ways. The Physics-Informed Neural Operator (PINO) [6] adds a PDE residual to the operator-learning loss — the soft route, lifted to function spaces. Physics-informed DeepONets [7] do the same for DeepONet. These inherit both the benefits and the inference-time non-guarantees of soft constraints.

3.3 Hard Physics Constraints in Operator Learning

A growing line of work enforces constraints exactly. ProbConserv enforces conservation laws through a Bayesian update on a predictive distribution [10]. Constraining layers enforce mass and energy conservation in climate downscaling [11]. BOON modifies integral kernels to enforce boundary conditions [12]. Differentiable-optimization layers find linear combinations of learned bases that satisfy a PDE-constrained problem [13]. For the divergence-free constraint, one can reparameterize the field as a curl [14], use a Helmholtz decomposition [15], or apply a spectral projection layer [9]. The most directly relevant work is Duruisseaux et al. [8], who project the operator output onto the constraint set in Fourier space; for incompressible Navier–Stokes this is the Leray projection, and they show it enforces the divergence-free condition at every point without harming fidelity to the data, at a small computational overhead.

The gap. Every hard-constraint method above applies the projection uniformly. Our work addresses the complementary question of *when* to apply the projection. We retain the exact, efficient Fourier projection as the hard path, but place a learned gate in front of it that trusts the constraint in well-conditioned spectral regions and defers to the data-driven operator where the physics is unreliable. To our knowledge, this is the first method to make the soft/hard choice adaptive and per-mode: a learned spectral gate that interpolates not between two trained components but against an *exact* physics projection, so the gate weight reads as how much the idealized physics is trusted at each mode, and the hard path *reduces* rather than adds model capacity (Chapter 5).

3.4 Physics-Informed Wireless Channel Estimation

Accurate channel estimation is critical for massive-MIMO systems, and is hard when the number of pilots is small. Classical model-based methods rely on the angular-domain sparsity of the channel and use greedy compressed-sensing algorithms such as orthogonal matching pursuit and its variants [23, 24, 25]; they degrade under pilot and overhead constraints. Purely data-driven estimators based on CNNs or diffusion models [26, 27] can be accurate when pilots are abundant but lack interpretability, need large datasets, and generalize poorly across sites. Javid and González-Prelcic [22]

propose a physics-informed estimator that fuses a least-squares pilot estimate with a ray-traced received-signal-strength (RSS) map through a U-Net with transformer and cross-attention modules, adding a Maxwell-based power-consistency penalty as a soft constraint. It achieves strong performance in the pilot-limited regime and is the soft-path baseline (and the source of the physics prior) for our MIMO experiments.

Our position. We take the physics-informed channel estimator as the soft path and add a hard path — the rank-r truncated-SVD projection of the antenna aperture — gated per Fourier mode. This converts a globally soft estimator into one that selectively trusts the low-rank physics where it is reliable. The generalization analysis of Chapter 5 quantifies the capacity saved.

3.5 Generalization Theory for Constrained Models

Our analysis rests on the empirical Rademacher complexity [30] and three standard estimates: the contraction lemma [30], the spectral-norm size bound for neural networks [28, 29], and Dudley's entropy integral together with covering numbers for low-rank matrices [32]. The vector-contraction inequality of Maurer [31] handles multi-output (here, complex) classes. The novelty in Chapter 5 is not these tools individually but their combination: we treat a gated class whose hard path is a projection onto a curved variety, decompose the complexity per Fourier mode, and obtain an interpolation between the soft and hard regimes through a single learned quantity $\bar{\alpha}$. We are careful to separate the unconditional part of the argument (a mixing inequality and a safety floor) from the multiplicative capacity reduction, which we state under one explicit low-rank-transfer assumption. A soft penalty, by contrast, does not reduce the Rademacher complexity at all — it acts on the loss, not the class — which is exactly why the analysis distinguishes the gate from a soft constraint.

Chapter 4

Adaptive Spectral Physics-INformed Operator

This chapter presents the proposed method. We first state the design philosophy (Section 4.1), then describe the spectral gate and the per-mode mixing (Section 4.2), the soft and hard paths (Sections 4.3–4.4), the training objective including a gate regularizer and when it is needed (Section 4.5), and finally the full algorithm (Section 4.6).

4.1 Overview and Design Philosophy

The Adaptive Spectral Trust Gate — which we abbreviate ASPINO, for Adaptive Spectral Physics-INformed Operator — rests on a single observation: a hard physics constraint is helpful exactly where the idealized physics is trustworthy, and harmful where it is not. The trustworthiness of the physics is not uniform across the signal. It is high on the large, well-resolved scales and low on the modes dominated by noise, aliasing and discretization error. Since these regions are naturally separated in the frequency domain, the right place to make the soft/hard decision is the Fourier domain, and the right granularity is per mode.

Concretely, let f be a data-driven neural-operator surrogate (the soft path), let $\mathcal{P}_r$ be an exact physics projection (the hard path), and let g be a learned gate producing a value in $[0, 1]$ for every Fourier mode. ASPINO forms the per-mode convex combination

$$\hat{h}_{\text{gate}}(k) = g(k)\,\big(\widehat{\mathcal{P}_r f}\big)(k) + \big(1 - g(k)\big)\,\hat{f}(k), \qquad \mathcal{H}_{\text{gate}} = \big\{x \mapsto h_{\text{gate}} : f \in \mathcal{F},\ g \in \mathcal{G}\big\}. \tag{4.1}$$

When $g(k) = 1$ the mode is taken entirely from the physics-projected path; when $g(k) = 0$ it is taken entirely from the soft path; intermediate values blend the two. The mixing happens in the same Fourier domain in which an FNO already operates, so the gate slots naturally into the architecture. Figure 4.1 shows the data flow.

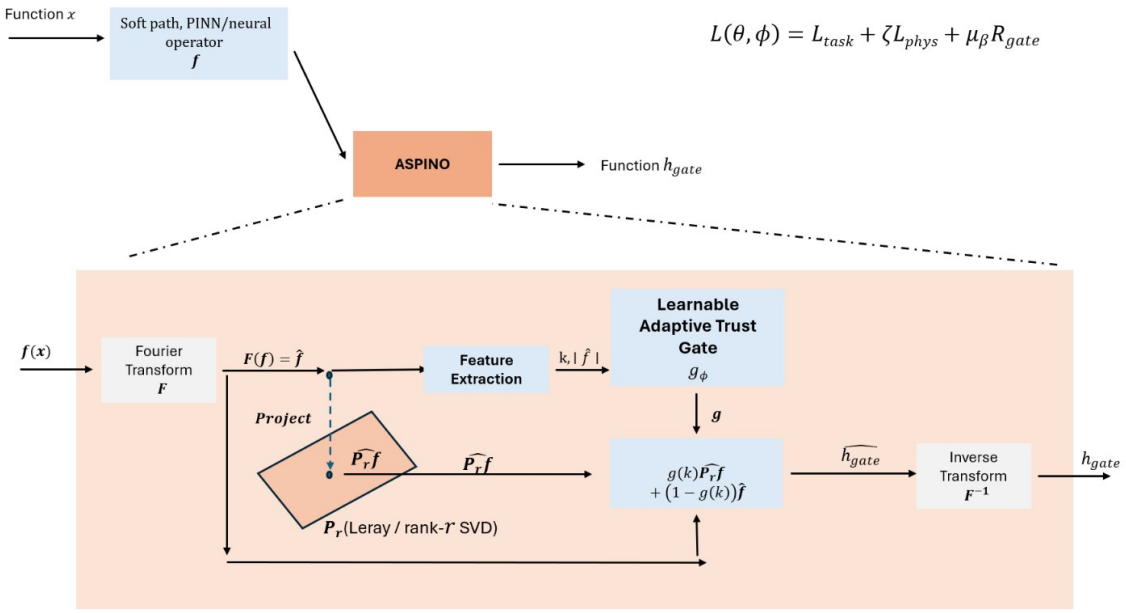


Figure 4.1: The ASPINO architecture. A neural-operator soft path f produces a candidate field; its Fourier transform is fed both to an exact physics projection $\mathcal{P}_r$ (the hard path) and to the mixer. A spectral gate g, driven by the spectral coordinate k and the spectral amplitude $|\hat{f}(k)|$, forms the per-mode convex combination of the projected and unprojected paths (Eq. (4.1)). An inverse transform returns the physical-domain output. The gate is the only new trainable component.

Table 4.1: Notation used in the construction.

Symbol	Meaning
$f \in \mathcal{F}$	data-driven soft path (neural operator / PINN)
$\mathcal{P}_r$	exact physics projection (hard path)
$g \in \mathcal{G}$	spectral gate, $g(k) \in [0,1]$
k	Fourier mode index ($k = 1, \ldots, K$)
$\hat{f}(k)$	Fourier coefficient of the soft-path output at mode k
$\mathcal{H}_{\text{gate}}$	gated hypothesis class, Eq. (4.1)
$\bar{g}(k) = \mathbb{E}[g(k)]$	mean gate value at mode k
ζ	weight of the soft physics residual in the loss
μ_β	weight of the gate (Wasserstein/Beta) regularizer
r	truncation rank of the MIMO hard path (Eq. (2.7))
c	per-sample energy-rescaling factor, clamped to $[c_{\min}, c_{\max}]$
P_{RSS}	RSS-predicted received power (energy target for c)

4.2 The Spectral Gate

The gate g_ϕ is a small network that outputs one value per spectral bin through a sigmoid, $g_\phi(\cdot) \in [0,1]^K$. Its realization adapts to the symmetry of each problem. In the fluid experiments the trust profile is essentially isotropic in wavenumber, so the gate is a per-mode multilayer perceptron ($2 \to 32 \to 16 \to 1$) with LeakyReLU activations, applied pointwise to each mode's two scalar features (the squared radial wavenumber $|k|^2$ and the log soft-path amplitude $\log|\hat{f}(k)|$). In the MIMO experiments the trust profile is directional in the aperture spectrum, so the gate is a small one-dimensional convolutional network over the spectral bins. In both cases the gate emits one value in $[0,1]$ per mode and the mixing is Eq. (4.1); only the network realizing the per-mode map differs. Two design choices are central.

Gate inputs: frequency and amplitude. The gate decides trust from two cues. The first is the spectral coordinate itself: physics reliability has a systematic frequency structure — it is high on the large-scale, low-wavenumber modes that carry the energy of the field and low on the noise- and aliasing-dominated bands (the Nyquist strip $k_y \approx$ Nyquist and the high-k_x band) — and the coordinate lets the gate localize this structure. In the fluid setting this structure is to good approximation isotropic, so the coordinate enters through the squared radial wavenumber $|k|^2 = k_x^2 + k_y^2$, a single scalar per mode; in the MIMO setting the aperture spectrum is directional and the gate is given the full two-dimensional spectral coordinate. In practice the learned gate trusts the physics ($g \to 1$) on the low-wavenumber modes and withdraws it ($g \to 0$) on those noise bands, with the withdrawn region growing as the input noise increases (Section 6.1.5). The second is the spectral amplitude $|\hat{f}(k)|$ of the soft-path output: a mode that the data-driven model has populated with substantial, structured energy is more likely to be signal than a near-empty mode that is mostly noise. Driving the gate by a function of frequency and a function of amplitude lets it identify both the systematically reliable bands and the modes where the data themselves vouch for the content. The gate profile is shared across the per-block / per-tap structure of the problem, so a single learned $g(k)$ applies to all blocks.

The isotropy of the fluid trust profile is also why a two-input pointwise MLP suffices there: with no anisotropic structure to exploit, the per-mode map needs only $|k|^2$ and the log amplitude, and a convolution over the spectral grid would add capacity without adding information. The spectral coordinate and the soft-path amplitude are thus the minimal cues the gate requires, and they are exactly the inputs used in the fluid experiments. In the MIMO experiments the gate is given a strictly richer view of the same two cues: in place of the magnitude alone it receives the phase-aware spectrum (real and imaginary parts) of both the pilot estimate $\hat{x}$ and the soft-path output $\hat{f}$, together with a global power prior obtained from the RSS map. Phase is informative here because the aperture spectrum is complex and its directivity — not only its magnitude — distinguishes the signal subspace from interference, and

the RSS prior supplies a per-sample power scale that the pilot spectrum alone does not fix. These remain functions of the spectral coordinate and amplitude: they do not alter what the gate decides, only how sharply it can localise the reliable bands. The construction is otherwise unchanged — the gate still emits one value in $[0, 1]$ per mode and the mixing is still Eq. (4.1).

Spectral normalization. The gate's layers are spectrally normalized in both regimes — the LeakyReLU MLP in the fluid setting and the convolutional stack in the MIMO setting — which bounds the product of operator norms $\prod_\ell \|W_\ell\|_{\mathrm{op}}$. The primary role of this constraint is capacity control: it makes the size bound (B) of Section 2.8 directly applicable, so that the gate's own Rademacher complexity is provably $O(n^{-1/2})$ and therefore asymptotically negligible (Lemma 5). Because LeakyReLU and the convolution are both 1-Lipschitz up to the bounded layer norms, the same Talagrand contraction applies to either realization. This holds independently of the training objective and in particular does not require the gate regularizer; it therefore applies to the MIMO gate, for which $\mu_\beta = 0$ (Section 4.5), exactly as it does to the fluid gate. In the idealized regime the same normalization additionally supports the Wasserstein regularizer (Section 4.5), which is defined with respect to a controlled Lipschitz constant, but this is a secondary benefit rather than the reason for the constraint.

4.3 The Soft Path

The soft path f is any neural-operator surrogate, optionally trained with a soft physics penalty. In the fluid experiments it is a Fourier Neural Operator [3]; in the channel-estimation experiments it is the physics-informed estimator of Javid and González-Prelcic [22], which fuses a least-squares pilot estimate with a ray-traced RSS map through a U-Net with transformer and cross-attention modules. In both cases f is the unconstrained hypothesis class: it has full expressive power and full flexibility, and it is the object the gate falls back to wherever the physics is not to be trusted. A key property of ASPINO is that this path is treated as a black box — the gate sits on top of it and the construction is agnostic to its internals.

4.4 The Hard Path

The hard path $\mathcal{P}_r$ is an exact physics projection applied in the Fourier domain. We use two instantiations.

Leray projection (incompressible flow). For the divergence-free constraint $\nabla \cdot u = 0$, $\mathcal{P}_r$ is the Leray projection of Section 2: at each mode it removes the component of $\hat{u}(\xi)$ along the wavevector ξ. This is a fixed, linear per-mode operator. Its effect at a mode is a contraction by a constant factor: in a D-dimensional velocity field it

deletes one of D directions, contracting the per-mode degrees of freedom by $(D-1)/D$ (which is $\frac{1}{2}$ in two dimensions).

Rank-r truncated SVD (MIMO aperture). For the antenna aperture, $\mathcal{P}_r$ is the rank-r Eckart–Young truncation (Eq. (2.7)), followed by a per-sample energy rescaling c to match the RSS-predicted power, clamped to $[c_{\min}, c_{\max}]$:

$$\mathcal{P}_r(A) = c\sum_{i=1}^{r}\sigma_i u_i v_i^*, \qquad c = \sqrt{P_{\mathrm{RSS}}/\|P_r^{(1)}(A)\|_F^2} \text{ clamped to } [c_{\min}, c_{\max}]. \tag{4.2}$$

A small jitter (10^{-7}) is added before the SVD so that the spectral gap is positive almost surely and SVD back-propagation is well defined (cf. Lemma 4). Unlike the Leray projection this map is nonlinear and data-dependent — its singular vectors move with the input — so its capacity cannot be summarized by a per-mode contraction factor and instead requires the covering-number argument of Chapter 5.

We deliberately use a low-rank (Eckart–Young) projection here rather than a hard projection onto an exact electromagnetic constraint such as a Maxwell / Helmholtz boundary condition. In the fluid setting the idealized physics (incompressibility) is a mode-wise linear constraint that is exact on the resolved grid, so a hard projection is both well-defined and trustworthy on the large scales. In massive MIMO the analogous "exact physics" constraint would have to be imposed on the sampled aperture, but the antenna spacing places the physically meaningful structure near and beyond the spatial Nyquist limit: the propagation environment is rich and the array is electrically large, so the angular content that a Maxwell-based projector would enforce is precisely the content that is under-resolved and aliased by the finite aperture. Imposing it as a hard constraint would therefore inject the discretization and aliasing error it is meant to remove. The low-rank structure, by contrast, is the property that survives this sampling: the dominant propagation paths concentrate the aperture energy in a few singular directions regardless of the Nyquist limit, which is exactly why the rank-r truncation is the appropriate trustworthy core for the gate to fall back on.

Why one gate serves both. The two hard paths are mathematically very different, yet both are computed in the Fourier domain and both are rank/structure constraints that are invariant under the unitary DFT (Lemma 2). The gate does not need to know which projection it is mixing — it operates only on the spectral coordinate and the soft-path amplitude (with the optional phase/RSS context of Section 4.2 where available), never on the identity or internals of $\mathcal{P}_r$. This is what makes ASPINO a single, reusable mechanism across domains.

4.5 Training Objective

The base objective is the task loss optionally augmented with the soft physics residual of the underlying surrogate,

$$\mathcal{L}(\theta, \phi) = \mathcal{L}_{\text{task}}(h_{\text{gate}}) + \zeta\, \mathcal{L}_{\text{phys}}(h_{\text{gate}}) + \mu_\beta\, \mathcal{R}_{\text{gate}}(\phi), \tag{4.3}$$

where $\mathcal{L}_{\text{task}}$ is the NMSE / relative-L^2 error, $\mathcal{L}_{\text{phys}}$ is the PDE (or Maxwell power) residual, and $\mathcal{R}_{\text{gate}}$ is an optional regularizer on the distribution of gate values.

A gate regularizer: when it is and is not needed. The gate can collapse to a degenerate, uniformly-hard profile ($g \equiv 1$) when the data provide little pressure to differentiate trust across modes. This is precisely what happens in the idealized Kolmogorov setting, where the physics is trustworthy across most of the spectrum: with nothing in the task loss penalizing a uniformly-hard gate, the gate drifts toward $g \equiv 1$ and loses the selective behaviour that is the point of the method. To prevent this we regularize the gate's output distribution toward a unimodal Beta$(2, 2)$ prior using a Wasserstein distance (Section 2), with a small weight $\mu_\beta > 0$; this keeps the gate differentiated and preserves the high-wavenumber soft region. In the MIMO setting the situation is reversed: the data are intrinsically unreliable in parts of the aperture spectrum (interference, imperfect low-rank structure, pilot noise), so the task loss itself supplies the pressure to keep the gate differentiated, and no distributional regularizer is needed ($\mu_\beta = 0$).

We also tried a bimodal Beta$(0.5, 0.5)$ (arcsine) prior, reasoning that the gate should be decisive at the extremes. This failed: a U-shaped prior rewards undifferentiated extremity without specifying *which* modes should be hard and which soft, so the optimizer satisfies it the cheap way — sending all modes to one extreme — which is the very collapse it was meant to prevent. The correct frequency-ordered structure (low-wavenumber modes hard, noisy high-wavenumber modes soft) must be earned per mode from the data, not imposed on the output marginal.

A note on the theory. The gate regularizer is an *optimization* device: it changes which solution the optimizer reaches, not the capacity of the hypothesis class. By the analysis of Chapter 5 a soft loss term does not reduce the Rademacher complexity (it acts on the loss, not the class), and the gate's own capacity is already controlled to $o(1)$ by spectral normalization and its bounded $[0, 1]$ output (Lemma 5). The regularizer should therefore be read as preventing gate collapse where the data underdetermine the gate, not as a generalization mechanism.

On the soft penalty and the gate. A second cautionary interaction is worth recording. When an energy-conserving hard path is combined with a soft physics penalty ($\zeta > 0$), the hard path trivially minimizes $\mathcal{L}_{\text{phys}}$, so the penalty actively rewards $g \to 1$ everywhere and can hurt task accuracy. The theory predicts this: a soft penalty does not reduce the hypothesis-class complexity, it only reshapes the

loss, and here it reshapes it in a way that biases the gate. Setting $\zeta = 0$ in the gated MIMO model removes both the (predicted) null effect on capacity and the optimization pathology.

4.6 Algorithm

Algorithm 1 summarizes one training step. The soft path may be pretrained and frozen ("plug-and-play") or trained jointly; in our experiments we use a frozen or lightly fine-tuned soft path and train only the gate, which is the cheapest and most robust option and the one for which the safety floor of Chapter 5 applies most cleanly.

Algorithm 1 ASPINO — one training step

Require: batch $\{x_i\}$, soft path f, hard projection $\mathcal{P}_r$, gate g_ϕ, weights ζ, μ_β

for each x_i in the batch **do**
  $u \leftarrow f(x_i)$ ▷ soft-path output (physical domain)
  $\hat{u} \leftarrow F(u)$ ▷ to Fourier domain
  $\hat{u}_{\mathcal{P}_r} \leftarrow \mathcal{P}_r(\hat{u})$ ▷ exact projection (Leray / rank-r SVD)
  $g \leftarrow g_\phi\big(|k|^2, \log|\hat{u}|; \theta_{\text{ctx}}\big)$ ▷ per-mode trust in $[0,1]$; fluid: MLP on the two scalars; MIMO: conv with phase/RSS context θ_{ctx}
  $\hat{h} \leftarrow g \odot \hat{u}_{\mathcal{P}_r} + (1-g) \odot \hat{u}$ ▷ per-mode mix, Eq. (4.1)
  $h_i \leftarrow F^{-1}(\hat{h})$ ▷ back to physical domain
end for
$\mathcal{L} \leftarrow \mathcal{L}_{\text{task}}(\{h_i\}) + \zeta\, \mathcal{L}_{\text{phys}}(\{h_i\}) + \mu_\beta\, \mathcal{R}_{\text{gate}}(\phi)$
update ϕ (and optionally f) by a gradient step on $\mathcal{L}$
return updated parameters

The forward pass adds, over the bare soft path, one Fourier transform, one inverse transform, one projection, and one small gate evaluation. For the Leray path the projection is two matrix–vector products with a precomputed pseudo-inverse; for the SVD path it is a per-block truncated SVD. In both cases the overhead is a small constant factor, consistent with the 1.06×–1.16× cost reported for the uniform projection alone [8].

Chapter 5

Generalization Analysis

This chapter analyses the generalization of the gated hypothesis class through its empirical Rademacher complexity. The results are deliberately layered by how much they assume, and we have been careful to separate what is *proved* from what is *assumed*.

The unconditional core is a mixing inequality (Proposition 1), stated for a fixed per-mode trust profile: the complexity of the gated class is at most the gate-weighted average of the complexities of the soft and hard paths. From it we obtain two qualitatively different conclusions depending on the hard path.

For the *linear Leray* hard path the reduction is fully proved: the projection is a fixed orthogonal operator, so it yields an *unconditional* safety floor (Corollary 1) and, under a mild and explicitly named direction-isotropy hypothesis on the soft class, an *exact* multiplicative reduction by the factor $\frac{D-1}{D}$ (Theorem 1). This is the cleanest result of the chapter and the one we recommend the reader keep as the anchor.

For the *nonlinear rank-r SVD* hard path the projection couples all modes and is data-dependent, so a genuine per-mode multiplier does not exist. There the multiplicative reduction $1-\bar{\alpha}(1-\sqrt{\rho_r})$ rests on one explicitly stated hypothesis, Assumption 2, which we motivate but do not prove from first principles; the safety floor for the SVD path is correspondingly conditional. Section 5.10 states precisely what is and is not established.

5.1 Model and Notation

We instantiate the analysis on the MIMO channel-estimation model and recover the fluid (Leray) model as the simpler special case of Section 5.5. The transmit array is a 24×24 planar aperture, so each (delay tap d, receive antenna ν) slice of the channel is a complex matrix $A_{d,\nu} \in \mathbb{C}^{m\times n}$ with $m = n = 24$. There are $D_{\text{tap}} = 16$ complex delay taps and $N_{\text{rx}} = 4$ receive antennas, i.e. $B_{\text{blk}} = 64$ aperture blocks per sample. (We write D_{tap} for the delay count to free the symbol D for the velocity dimension of the Navier–Stokes example.)

Soft path $\mathcal{F}$. The PINN mapping an input x (least-squares pilot estimate + RSS map) to a full-rank aperture $f(x)$. We stress that $\mathcal{F}$ is a *class*, not the single trained network: it is the set of maps the fixed architecture can represent as its weights range over a ball of bounded spectral norm $\prod_\ell \|W_\ell\|_{\mathrm{op}} \le S$. This is the object the generalization gap depends on, and it is what makes the analysis non-vacuous, since the empirical Rademacher complexity of a *singleton* $\{f\}$ is identically zero (the Rademacher signs are mean-zero). The spectral-norm bound is precisely what enters the size bound (B) of Section 2.8: it controls the constant in

$$\widehat{\mathfrak{R}}_n(\mathcal{F}) \;\le\; \frac{c\, B_x\, \rho^L \sqrt{L \log(2d)}\; \prod_\ell \|W_\ell\|_{\mathrm{op}}}{\sqrt{n}} \;=\; O(n^{-1/2}),$$

so a finite spectral norm yields a finite constant and the $n^{-1/2}$ rate. In experiments the soft path is frozen; that is for *attribution* (any gain is the gate's) and does not change the capacity of the class the bound reasons about.

Hard path $\mathcal{P}_r$ (SVD). For each of the $B_{\mathrm{blk}} = 64$ aperture blocks, the rank-r truncated SVD (Eckart–Young projection, Eq. (2.7)); the truncated blocks are then rescaled by a single *per-sample* scalar c that matches the channel's *total* energy — summed over all delay taps and receive antennas — to the RSS-predicted power up to a learned calibration constant $\kappa_0 = e^\theta$,

$$c = \sqrt{\frac{\kappa_0\, \overline{\mathrm{RSS}}}{\sum_{d,\nu} \|P_r^{(1)}(A_{d,\nu})\|_F^2}}, \qquad c \text{ clamped to } [c_{\min}, c_{\max}] = [0.5, 2.5],$$

with $\overline{\mathrm{RSS}}$ the spatial mean of the RSS map and $r = 3$ in code. A 10^{-7} jitter is added before the SVD (Remark 5). The SVD preserves the power-delay shape across taps; c only fixes the overall scale.

Hard path (Leray). On the periodic torus the divergence-free constraint is the fixed, linear, per-mode orthogonal projection $P_{\mathrm{L}}(\xi) = I - \xi\xi^*/|\xi|^2$ acting on the D-component velocity Fourier coefficient $\hat{u}(\xi) \in \mathbb{C}^D$ (with $P_{\mathrm{L}}(0) = I$). It deletes the component of $\hat{u}(\xi)$ along the wavevector ξ, so its image is the $(D-1)$-dimensional plane $\xi^\perp$ and $\|P_{\mathrm{L}}(\xi)\|_{\mathrm{op}} = 1$. This linearity is what makes the Leray reduction provable in Section 5.5.

Gate $\mathcal{G}$. A small trained network $g_\phi : x \mapsto [0,1]^K$ with sigmoid output ($K = 312$ for the MIMO model, the number of `rfft2` bins of a 24×24 block). In the analysis we write $g(k)$ for its per-mode trust value and $\bar{g}(k)$ for the fixed per-mode profile we analyse (see Remark 7). Both gate realizations are spectrally normalized (except a final zero-initialized 1×1 readout, a fixed bounded linear map after training), so Lemma 5 bounds either one.

Mixing. The hard path truncates rank in the *spatial* aperture; both paths are then transformed to the 2-D Fourier domain, where the gate mixes them, $\hat{h}_{\text{gate}}(k) = g(k)\,\widehat{\mathcal{P}_r f}(k) + (1 - g(k))\,\hat{f}(k)$, with $\mathcal{H}_{\text{gate}}$ as in Eq. (4.1); an inverse transform returns the output. By the commutation property of Lemma 2 the spatial-then-Fourier order is equivalent to projecting in the Fourier domain. Crucially, $\widehat{\mathcal{P}_r f}$ and $\hat{f}$ are *coupled*: both are deterministic functions of the same $f \in \mathcal{F}$.

5.2 Per-Mode Rademacher Complexity

Definition 1 (Per-mode and per-coordinate complexity) *Fix an orthonormal output basis. For a class $\mathcal{H}$ mapping x to Fourier-domain outputs $\hat{h}(k) \in \mathbb{C}^{D'}$, write the (real and imaginary parts of the) D' output components at mode k as real coordinates c. The per-coordinate empirical Rademacher complexity is*

$$\widehat{\mathfrak{R}}_n(\mathcal{H}, k, c) = \tfrac{1}{n}\mathbb{E}_\sigma\Big[\sup_{h\in\mathcal{H}} \sum_i \sigma_i \left(\hat{h}(x_i)(k)\right)_c\Big],$$

the per-mode complexity is $\widehat{\mathfrak{R}}_n(\mathcal{H}, k) = \sum_c \widehat{\mathfrak{R}}_n(\mathcal{H}, k, c)$ (vector contraction, [31]), and the total budget is $\widehat{\mathfrak{R}}_n^{\Sigma}(\mathcal{H}) := \sum_k \widehat{\mathfrak{R}}_n(\mathcal{H}, k) = \sum_{k,c} \widehat{\mathfrak{R}}_n(\mathcal{H}, k, c)$.

Lemma 1 (Subadditive decomposition; Parseval radius invariance) *The empirical Rademacher complexity is bounded by the budget,*

$$\widehat{\mathfrak{R}}_n(\mathcal{H}) \;\le\; \sum_k \widehat{\mathfrak{R}}_n(\mathcal{H}, k) \;=\; \widehat{\mathfrak{R}}_n^{\Sigma}(\mathcal{H}), \tag{5.1}$$

in any orthonormal output basis. Moreover, by Parseval's identity the total output energy is the same in the spatial and Fourier bases, so the radius B of any Frobenius ball used in the covering arguments of Section 5.3 is basis-independent; we therefore compute all per-mode quantities in the Fourier basis (or, in Section 5.5, in the ξ-adapted basis) without changing B.

Proof:

Writing h through its coordinates and using that the supremum of a sum is at most the sum of the suprema, $\sup_h \sum_i \sigma_i \sum_{k,c} \left(\hat{h}(x_i)(k)\right)_c \le \sum_{k,c} \sup_h \sum_i \sigma_i \left(\hat{h}(x_i)(k)\right)_c$; taking $\frac{1}{n}\mathbb{E}_\sigma$ gives (5.1). For the radius, the 2-D DFT is realized by a unitary U ($U^*U = I$), so $\|Uy\|_F = \|y\|_F$ for every output y; hence the Frobenius radius of the output set is identical in the two bases. □

Remark 1 (What we do *not* claim) *Unlike the radius, the budget $\widehat{\mathfrak{R}}_n^{\Sigma}$ is* not *in general invariant under a unitary change of output basis: each summand contains a supremum, and summing support-function-type terms over a rotated basis can change the total. We therefore never equate a spatial-basis budget with a Fourier-basis budget*

numerically; we simply fix a convenient basis at each mode and work in it. Equation (5.1) *is an inequality, not an equality: Rademacher complexity is not additive across orthogonal coordinates, and equality would require the supremising h to be attainable coordinate-by-coordinate simultaneously.*

5.3 Lemmas for the SVD Hard Path

Lemma 2 (Unitary invariance of rank and truncation–transform commutation) *Let F_m, F_n be the unitary DFT matrices. For any $A \in \mathbb{C}^{m\times n}$,* $\operatorname{rank}(F_m A F_n^*) = \operatorname{rank}(A)$*, and the rank-r truncation commutes with the transform:* $P_r^{(1)}(F_m A F_n^*) = F_m\, P_r^{(1)}(A)\, F_n^*$.

Proof:

F_m, F_n are invertible, and multiplying left/right by invertible matrices preserves rank, so $\operatorname{rank}(F_m A F_n^*) = \operatorname{rank}(A)$ and both the spatial and Fourier constraints define the same variety $\mathcal{V}_r$. Writing $A = U\Sigma V^*$, unitarity of F_m, F_n gives $F_m A F_n^* = (F_m U)\,\Sigma\,(F_n V)^*$, a valid SVD with the *same* singular values; Eckart–Young keeps the top r, so truncating before or after the transform coincides. The scalar rescaling commutes trivially. □

Remark 2 *This reconciles the implementation with Algorithm 1: the code truncates rank in the spatial aperture and only then transforms to the Fourier domain where the gate mixes, while the algorithm states the projection in the Fourier domain — and the two are identical.*

Lemma 3 (Energy contraction and per-sample pinning) *For $r \le \min(m,n)$ and every block, $\|P_r^{(1)}(A)\|_F \le \|A\|_F$, so block-wise $\|\mathcal{P}_r(A)\|_F \le c_{\max}\|A\|_F$. When the per-sample scalar c lies strictly inside the clamp band, the* total *output energy is pinned,*

$$\sum_{d,\nu} \|\mathcal{P}_r(A_{d,\nu})\|_F^2 = \kappa_0\, \overline{\mathrm{RSS}},$$

a per-sample constant independent of the gate, where $\kappa_0 = e^{\theta}$.

Proof:

By Eckart–Young, $\|A\|_F^2 - \|P_r^{(1)}(A)\|_F^2 = \sum_{i>r}\sigma_i^2 \ge 0$, hence $\|P_r^{(1)}(A)\|_F \le \|A\|_F$; the rescaling multiplies by $c \le c_{\max}$. With $c^2 = \kappa_0\,\overline{\mathrm{RSS}}/\sum_{d,\nu}\|P_r^{(1)}(A_{d,\nu})\|_F^2$ applied unclamped, $\sum_{d,\nu}\|\mathcal{P}_r(A_{d,\nu})\|_F^2 = c^2 \sum_{d,\nu}\|P_r^{(1)}(A_{d,\nu})\|_F^2 = \kappa_0\,\overline{\mathrm{RSS}}$. □

Remark 3 (The amplification caveat) *Because c can exceed* 1 *(up to $c_{\max} = 2.5$), the SVD hard path is* not *unconditionally non-expansive: even inside the clamp band the per-sample energy is pinned to $\kappa_0\,\overline{\mathrm{RSS}}$, which may exceed the soft-path energy,*

and when the clamp binds the enlargement is bounded only by $c_{\max}$*. Norm-pinning constrains output magnitude but not how the class correlates with noise, so it does not by itself imply the per-mode inequality* $\widehat{\mathfrak{R}}_n(\mathcal{P}_r \circ \mathcal{F}, k) \le \widehat{\mathfrak{R}}_n(\mathcal{F}, k)$*. The SVD safety floor (Corollary 2) is therefore stated conditionally and verified empirically. Contrast the Leray path, where* $\|P_{\mathrm{L}}\|_{\mathrm{op}} = 1$ *makes the floor unconditional (Corollary 1).*

Lemma 4 (Local spectral-gap stability) *Fix* $\delta_0 > 0$*,* $R > 0$ *and let* $\mathcal{G}_{\delta_0,R} = \{A : \sigma_r(A) - \sigma_{r+1}(A) \ge \delta_0,\ \|A\|_{\mathrm{op}} \le R\}$*. On* $\mathcal{G}_{\delta_0,R}$ *the rank-r truncation* $P_r^{(1)}$ *is uniformly Lipschitz in Frobenius norm: for all* $A, B \in \mathcal{G}_{\delta_0,R}$*,*

$$\|P_r^{(1)}(A) - P_r^{(1)}(B)\|_F \le L_{\delta_0,R}\,\|A - B\|_F, \qquad L_{\delta_0,R} = C\Big(1 + \tfrac{R}{\delta_0}\Big),$$

for an absolute constant C*. The constant is uniform over the region, depending only on the gap floor* δ_0 *and radius* R*, not on the individual matrix.*

Proof:

Write $P_r^{(1)}(A) = \Pi_L^A A\,\Pi_R^A$ with top-r spectral projectors $\Pi_L^A = U_r U_r^*$, $\Pi_R^A = V_r V_r^*$. Telescoping,

$$P_r^{(1)}(B) - P_r^{(1)}(A) = \Pi_L^B(B - A)\Pi_R^B + (\Pi_L^B - \Pi_L^A)A\,\Pi_R^B + \Pi_L^A A\,(\Pi_R^B - \Pi_R^A).$$

The first term is a two-sided projection of $B - A$, so its Frobenius norm is $\le \|A - B\|_F$ (the Mirsky / 1-Lipschitz singular-value contribution; this is the "+1"). For the second and third terms, $\|XYZ\|_F \le \|X\|_F \|Y\|_{\mathrm{op}} \|Z\|_{\mathrm{op}}$ with $\|A\|_{\mathrm{op}} \le R$ and projector op-norm 1 gives a factor $R\,\|\Pi_\bullet^B - \Pi_\bullet^A\|_F$. By the Davis–Kahan / Wedin $\sin\Theta$ theorem under the gap floor δ_0, $\|\Pi_\bullet^B - \Pi_\bullet^A\|_F \le \sqrt{2}\,\|A - B\|_F/\delta_0$ on each side. Collecting, $\|P_r^{(1)}(A) - P_r^{(1)}(B)\|_F \le (1 + 2\sqrt{2}\,R/\delta_0)\|A - B\|_F$, which is the stated bound with C absorbing the $\sin\Theta$ normalization. Using $\|A\|_{\mathrm{op}}$ rather than $\|A\|_F$ keeps C free of the ambient dimension. □

Remark 4 (Role of the lemma, and what it does *not* give) *The uniform constant is exactly what makes the composed class* $\mathcal{P}_r \circ \mathcal{F}$ *coverable on the sample, so that Dudley's integral (tool (C)) applies to it at all; this is the lemma's only role. Note the direction: a Lipschitz map gives* $N(\epsilon, \mathcal{P}_r \circ \mathcal{F}|_S) \le N(\epsilon/L, \mathcal{F}|_S)$*, an upper bound at a* finer *radius, so the lemma controls that the hard path is not much* more *complex than* $\mathcal{F}$*; it cannot by itself produce a multiplicative* reduction*, which (for SVD) is supplied by Assumption 2. The uniform gap* $\delta \ge \delta_0$ *is essential: pointwise distinctness of the singular values does not bound* $\sup_A L(A)$*, because* $\inf_A(\sigma_r - \sigma_{r+1})$ *may still be* 0*. We carry* $\delta \ge \delta_0$ *explicitly into the covering arguments and list it among the caveats of Section 5.10.*

Remark 5 (The jitter is a numerical device, not a complexity tool) *The* 10^{-7} *jitter perturbs the singular values to be distinct with probability one, which makes the singular vectors — and hence SVD back-propagation — well defined. This is a training-stability device; it is* not *the source of the uniform gap* δ_0 *above, which is a hypothesis on the data region, not a consequence of the jitter.*

Remark 6 (The capacity ratio, and why it is not yet a Rademacher inequality) *The complex rank-r variety in $\mathbb{C}^{m\times n}$ is, off its singular locus, a smooth manifold of real dimension $d_r = 2r(m+n-r)$. Relative to $d_{full} = 2mn$ this defines the capacity ratio*

$$\rho_r := \frac{d_r}{d_{full}} = \frac{r(m+n-r)}{mn}, \qquad \rho_3 = \tfrac{270}{1152} = 0.234, \quad \sqrt{\rho_3} = 0.484 \quad (m = n = 24,\ r = 3). \tag{5.2}$$

*On the smooth stratum, $\log N(\epsilon, \mathcal{V}_r \cap B\mathbb{B}, \|\cdot\|_F) \le d_r \log(9B/\epsilon)$ [32]. **This bounds the metric entropy of the** range **of a single hard-path output, not the empirical L_2 entropy of the function class $\mathcal{P}_r \circ \mathcal{F}$ on the n-sample design**, which is what Dudley's integral requires. The two coincide only when the hard-path outputs are tied across samples by a shared low-dimensional parameter; under per-sample truncation they are not. We therefore promote the desired SVD reduction to an explicit assumption rather than claim it as a theorem.*

Lemma 5 (Gate complexity) *Let g_ϕ be a fixed trained gate of depth L_g with sigmoid output and bounded weight spectral norms $\|W_\ell\|_{\mathrm{op}}$ (feature-mixing layers spectrally normalized; the final readout a fixed bounded linear map), and 1-Lipschitz hidden activations. Then*

$$\widehat{\mathfrak{R}}_n(\mathcal{G}) \le \tfrac{1}{4}\,\widehat{\mathfrak{R}}_n(\mathcal{G}^{\mathrm{pre}}) = O\Big(\tfrac{B_x\,\rho^{L_g}\sqrt{L_g \log d}\,\prod_\ell \|W_\ell\|_{\mathrm{op}}}{\sqrt{n}}\Big) = o(1). \tag{5.3}$$

Proof:

The sigmoid is $\frac{1}{4}$-Lipschitz, so Talagrand contraction (A) gives $\widehat{\mathfrak{R}}_n(\mathcal{G}) \le \frac{1}{4}\widehat{\mathfrak{R}}_n(\mathcal{G}^{\mathrm{pre}})$. Peeling each hidden layer multiplies the bound by its activation Lipschitz factor (≤ 1) and by $\|W_\ell\|_{\mathrm{op}}$ (for a convolution, the spectral norm of the reshaped doubly-block-circulant weight upper-bounds the true operator norm, which is all bound (B) needs; for a dense layer it is the matrix spectral norm directly). Collecting all L_g layers gives the displayed bound, $\propto n^{-1/2}$. The zero-initialized 1×1 readout, though not spectrally normalized, is a fixed bounded linear map of finite operator norm, absorbed into the constant; in particular the bound is independent of the gate-regularizer weight μ_β. □

5.4 The Mixing Inequality (Unconditional Core)

Remark 7 (We analyse a fixed per-mode profile) *The trained gate g_ϕ is input-dependent, so its per-sample values $g_\phi(x_i)(k)$ sit inside the supremum defining the Rademacher complexity, and one cannot simply factor out a mean weight. We therefore analyse the gated class with the gate fixed to a deterministic per-mode profile $\bar{g}(k) \in [0,1]$ — the trained gate's per-mode trust profile. This is* exact *when the gate is frozen and per-mode (the regime of the MIMO experiments, where the soft*

path is frozen and the gate trained); the residual sample-to-sample variation of the input-dependent gate enlarges the class to a product with $\mathcal{G}$ and contributes at most the term $C\,\widehat{\mathfrak{R}}_n(\mathcal{G}) = o(1)$ of Lemma 5. We make this idealization explicit because it is what licenses the per-mode weighting below; without it, bounding each weight by 1 *would only give the weaker* $\widehat{\mathfrak{R}}_n(\mathcal{H}_{gate}) \le \sum_k [\widehat{\mathfrak{R}}_n(\mathcal{P}_r \circ \mathcal{F}, k) + \widehat{\mathfrak{R}}_n(\mathcal{F}, k)]$.

Proposition 1 (Gated mixing inequality, fixed profile) *For a fixed per-mode profile* $\bar{g}(k) \in [0,1]$,

$$\widehat{\mathfrak{R}}_n(\mathcal{H}_{gate}) \;\le\; \sum_k \Big[\bar{g}(k)\,\widehat{\mathfrak{R}}_n(\mathcal{P}_r \circ \mathcal{F}, k) + (1-\bar{g}(k))\,\widehat{\mathfrak{R}}_n(\mathcal{F}, k)\Big]. \tag{5.4}$$

For the deployed input-dependent gate the same bound holds up to an additive $C\,\widehat{\mathfrak{R}}_n(\mathcal{G}) = o(1)$ *with* $C \propto (\|\mathcal{P}_r \circ \mathcal{F}\|_\infty + \|\mathcal{F}\|_\infty)$.

Proof:

Decompose by mode with Lemma 1: $\widehat{\mathfrak{R}}_n(\mathcal{H}_{\text{gate}}) \le \sum_k \widehat{\mathfrak{R}}_n(\mathcal{H}_{\text{gate}}, k)$. For fixed $\bar{g}(k) \in [0,1]$ the per-mode class is $\{x \mapsto \bar{g}(k)\widehat{\mathcal{P}_r f}(x)(k) + (1-\bar{g}(k))\hat{f}(x)(k) : f \in \mathcal{F}\}$, parametrized by the single f. Since $\bar{g}(k), 1-\bar{g}(k) \ge 0$, the supremum of the sum is at most the weighted sum of the suprema,

$$\widehat{\mathfrak{R}}_n(\mathcal{H}_{\text{gate}}, k) \le \bar{g}(k)\widehat{\mathfrak{R}}_n(\mathcal{P}_r \circ \mathcal{F}, k) + (1-\bar{g}(k))\widehat{\mathfrak{R}}_n(\mathcal{F}, k),$$

the coupling of $\widehat{\mathcal{P}_r f}$ and $\hat{f}$ through the same f making this a single supremum that the split only relaxes. Summing over k gives (5.4). For the learned gate, $\mathcal{H}_{\text{gate}}$ is the product of the fixed-profile mixing class with $\mathcal{G}$; the standard product bound for bounded classes adds at most $C\,\widehat{\mathfrak{R}}_n(\mathcal{G})$, which is $o(1)$ by Lemma 5. □

Proposition 1 uses only subadditivity and the convexity of the mixer; it is unconditional in the fixed-profile sense above. Everything below specialises it by relating $\widehat{\mathfrak{R}}_n(\mathcal{P}_r \circ \mathcal{F}, k)$ to $\widehat{\mathfrak{R}}_n(\mathcal{F}, k)$, which is exactly where the two hard paths diverge.

5.5 The Leray Hard Path: a Fully Proved Reduction

For the linear Leray projection the per-mode map is a fixed orthogonal operator, and the reduction can be established outright. This is the special case where Assumption 2 is *not* needed.

Lemma 6 (Per-mode contraction of a fixed orthogonal projection) *Let the Leray hard path act per mode as* $\hat{u}(\xi) \mapsto P_{\mathrm{L}}(\xi)\hat{u}(\xi)$ *with* $P_{\mathrm{L}}(\xi)$ *the orthogonal projection onto* $\xi^\perp$ *(rank* $D-1$*). Then, at every mode,*

$$\widehat{\mathfrak{R}}_n(P_{\mathrm{L}} \circ \mathcal{F}, \xi) \;\le\; \widehat{\mathfrak{R}}_n(\mathcal{F}, \xi),$$

unconditionally and with no data-dependence.

Proof:

Work in the ξ-adapted orthonormal basis $\{e_1 = \xi/|\xi|, e_2, \ldots, e_D\}$, in which $P_{\mathrm{L}}(\xi) = \mathrm{diag}(0, 1, \ldots, 1)$. Per coordinate (Def. 1), the projected class has $(P_{\mathrm{L}}\hat{u})_{e_1} \equiv 0$ and $(P_{\mathrm{L}}\hat{u})_{e_j} = \hat{u}_{e_j}$ for $j \geq 2$. Hence $\widehat{\mathfrak{R}}_n(P_{\mathrm{L}} \circ \mathcal{F}, \xi, e_1) = 0$ and $\widehat{\mathfrak{R}}_n(P_{\mathrm{L}} \circ \mathcal{F}, \xi, e_j) = \widehat{\mathfrak{R}}_n(\mathcal{F}, \xi, e_j)$ for $j \geq 2$. Summing, $\widehat{\mathfrak{R}}_n(P_{\mathrm{L}} \circ \mathcal{F}, \xi) = \sum_{j\geq 2} \widehat{\mathfrak{R}}_n(\mathcal{F}, \xi, e_j) \leq \sum_{j\geq 1} \widehat{\mathfrak{R}}_n(\mathcal{F}, \xi, e_j) = \widehat{\mathfrak{R}}_n(\mathcal{F}, \xi)$. □

Corollary 1 (Unconditional Leray safety floor) *For the Leray hard path and any fixed profile $\bar{g}$, $\widehat{\mathfrak{R}}_n(\mathcal{H}_{gate}) \leq \widehat{\mathfrak{R}}_n(\mathcal{F})$ (and $\widehat{\mathfrak{R}}_n(\mathcal{F}) + o(1)$ for the learned gate). The Leray-gated class is never more complex than the soft path it wraps — no clamp condition and no Assumption 2 are required.*

Proof:

By Lemma 6 each bracket in (5.4) is at most $\widehat{\mathfrak{R}}_n(\mathcal{F}, k)$; sum over k and apply Lemma 1 to $\mathcal{F}$. □

Assumption 1 (Direction isotropy of the soft class) *At each mode ξ the soft class has no preferred velocity direction: the per-direction complexities are equal, $\widehat{\mathfrak{R}}_n(\mathcal{F}, \xi, e_1) = \cdots = \widehat{\mathfrak{R}}_n(\mathcal{F}, \xi, e_D)$.*

This is a symmetry hypothesis on the *soft class*, natural for homogeneous, isotropic turbulence, and directly checkable from the per-direction complexities — unlike the entropy-transfer hypothesis the SVD path will require.

Theorem 1 (Exact Leray reduction under isotropy) *Under Assumption 1, $\widehat{\mathfrak{R}}_n(P_{\mathrm{L}} \circ \mathcal{F}, \xi) = \frac{D-1}{D}\widehat{\mathfrak{R}}_n(\mathcal{F}, \xi)$ at every mode. Consequently, with the capacity-weighted mean gate $\bar{\alpha} := \sum_k \bar{g}(k)\widehat{\mathfrak{R}}_n(\mathcal{F}, k) / \sum_k \widehat{\mathfrak{R}}_n(\mathcal{F}, k) \in [0, 1]$,*

$$\boxed{\widehat{\mathfrak{R}}_n(\mathcal{H}_{gate}) \;\leq\; \Big(1 - \tfrac{\bar{\alpha}}{D}\Big)\widehat{\mathfrak{R}}_n(\mathcal{F}) + o(1),} \tag{5.5}$$

which for $D = 2$ is $\big(1 - \frac{\bar{\alpha}}{2}\big)\widehat{\mathfrak{R}}_n(\mathcal{F}) + o(1)$, i.e. a per-mode reduction factor $\frac{D-1}{D} = \frac{1}{2}$.

Proof:

By Assumption 1 the D per-direction complexities at mode ξ are equal, so each equals $\frac{1}{D}\widehat{\mathfrak{R}}_n(\mathcal{F}, \xi)$. Lemma 6 removes exactly the e_1 direction, leaving $\frac{D-1}{D}\widehat{\mathfrak{R}}_n(\mathcal{F}, \xi)$. Substituting into the hard bracket of (5.4), $\widehat{\mathfrak{R}}_n(\mathcal{H}_{\mathrm{gate}}) \leq \sum_k [\bar{g}(k)\frac{D-1}{D} + (1 - \bar{g}(k))]\widehat{\mathfrak{R}}_n(\mathcal{F}, k) + o(1) = \widehat{\mathfrak{R}}_n(\mathcal{F}) - \frac{1}{D}\sum_k \bar{g}(k)\widehat{\mathfrak{R}}_n(\mathcal{F}, k) + o(1)$. Dividing the sum by $\sum_k \widehat{\mathfrak{R}}_n(\mathcal{F}, k)$ identifies $\bar{\alpha}$ and gives (5.5). □

5.6 The SVD Hard Path: Conditional Safety Floor and Reduction

Corollary 2 (Conditional SVD safety floor) *Suppose* $\widehat{\mathfrak{R}}_n(\mathcal{P}_r \circ \mathcal{F}, k) \le \widehat{\mathfrak{R}}_n(\mathcal{F}, k)$ *for every mode. Then, for any fixed profile* $\bar{g}$, $\widehat{\mathfrak{R}}_n(\mathcal{H}_{gate}) \le \widehat{\mathfrak{R}}_n(\mathcal{F}) + o(1)$.

Proof:
Each bracket in (5.4) is then at most $\widehat{\mathfrak{R}}_n(\mathcal{F}, k)$; sum and apply Lemma 1. □

Remark 8 (When the SVD hypothesis holds) *The hypothesis removes the only mechanism by which the SVD hard path could increase capacity. It is guaranteed when the rescaling is non-amplifying (*$c \le 1$*, so* $\mathcal{P}_r$ *is non-expansive by Lemma 3). When the clamp is inactive the gross* $c_{\max}$*-enlargement is excluded; this does not formally imply the per-mode inequality (Remark 3), so the corollary stays conditional and is verified empirically in Chapter 6, where the gated error never exceeds the PINN.*

Assumption 2 (Low-rank entropy transfer, capacity-weighted form) *On the fixed design* $S = \{x_i\}$*, routing the gate-weighted output through the rank-r hard path contracts the relevant empirical* L_2 *covering entropy by the capacity ratio:*

$$\sum_k \bar{g}(k)\, \widehat{\mathfrak{R}}_n(\mathcal{P}_r \circ \mathcal{F}, k) \;\le\; \sqrt{\rho_r} \sum_k \bar{g}(k)\, \widehat{\mathfrak{R}}_n(\mathcal{F}, k). \tag{5.6}$$

Remark 9 (Status of Assumption 2) *This is the one substantive hypothesis of the SVD analysis and it is* not *proved here. It holds rigorously in the transductive / shared-low-rank setting (a single rank-r object observed across the sample, the matrix-completion regime where* $\sqrt{d_r/n}$ *scaling is classical), and is plausible when the soft network's reachable outputs on S already lie near a shared rank-r family. The range-covering argument of Remark 6 motivates* $\sqrt{\rho_r}$ *but does not establish the empirical-*L_2 *statement; closing this gap is the principal open problem (Section 5.10). Contrast the Leray path, where the analogous reduction is the proved Theorem 1 under a checkable symmetry hypothesis.*

Theorem 2 (Gated SVD bound under Assumption 2) *With* $\bar{\alpha}$ *as in Theorem 1, if Assumption 2 holds then*

$$\widehat{\mathfrak{R}}_n(\mathcal{H}_{gate}) \le \big(1 - \bar{\alpha}\,(1 - \sqrt{\rho_r})\big)\, \widehat{\mathfrak{R}}_n(\mathcal{F}) + o(1). \tag{5.7}$$

Proof:
From (5.4) and Assumption 2, $\widehat{\mathfrak{R}}_n(\mathcal{H}_{\text{gate}}) \le \widehat{\mathfrak{R}}_n(\mathcal{F}) - (1-\sqrt{\rho_r}) \sum_k \bar{g}(k) \widehat{\mathfrak{R}}_n(\mathcal{F}, k) + o(1)$; dividing the middle sum by $\sum_k \widehat{\mathfrak{R}}_n(\mathcal{F}, k)$ identifies $\bar{\alpha}$ and gives (5.7). □

Remark 10 (Why aggregate, not per-mode) *A per-mode multiplier* $\widehat{\mathfrak{R}}_n(\mathcal{P}_r \circ \mathcal{F}, k) \le \sqrt{\rho_r}\widehat{\mathfrak{R}}_n(\mathcal{F}, k)$ *cannot be justified: SVD truncation is nonlinear and couples modes, so no per-mode factor exists (contrast Lemma 6, where the linear projection* does *have one). The honest SVD reduction is aggregate, and* $\bar{\alpha}$ *is defined as a capacity-weighted average precisely so the aggregate inequality closes.*

5.7 Generalization Consequence

Corollary 3 (Generalization gap) *With probability at least $1-\delta$, for every $h \in \mathcal{H}_{gate}$ (M-bounded loss, standard Rademacher bound):*

- *Leray, unconditionally under Assumption 1:* $L(h) \leq \widehat{L}(h) + 2\big(1 - \frac{\bar{\alpha}}{D}\big)\widehat{\mathfrak{R}}_n(\mathcal{F}) + 3M\sqrt{\frac{\log(2/\delta)}{2n}}$.
- *SVD, under Assumption 2: the same with factor* $1 - \bar{\alpha}(1 - \sqrt{\rho_r})$*; absent it, the unconditional* $L(h) \leq \widehat{L}(h) + 2\widehat{\mathfrak{R}}_n(\mathcal{F}) + o(1) + 3M\sqrt{\log(2/\delta)/2n}$ *via Corollary 2.*

Proof:
Apply the standard Rademacher generalization bound to $\mathcal{H}_{\text{gate}}$ and substitute Theorem 1, Theorem 2, or Corollary 2. □

5.8 Four Regimes

Table 5.1 places the gate among the alternatives. The Leray-gated row is unconditional (under the symmetry Assumption 1); the SVD-gated row carries the qualifier "under Assumption 2."

Key insight. A soft physics penalty regularizes the loss, not the hypothesis class, so it leaves $\widehat{\mathfrak{R}}_n$ unchanged. A hard constraint can reduce $\widehat{\mathfrak{R}}_n$ but injects bias everywhere. The gate keeps the reduction where the physics is reliable and full flexibility elsewhere.

Remark 11 (Theory meets the optimizer) *With $\zeta > 0$ the energy-conserving hard path minimizes $\mathcal{L}_{phys}$, so the penalty biases $g \to 1$ and can hurt accuracy. Setting ζ to zero (or negligibly small) removes both the null effect on capacity and the optimization pathology.*

Table 5.1: Effect on hypothesis-class complexity. The Leray reduction is proved; the SVD reduction is conditional on Assumption 2.

Regime	Rademacher complexity	Bias / variance
Soft only ($g \equiv 0$)	$\widehat{\mathfrak{R}}_n(\mathcal{F})$	low bias, high variance
Hard only ($g \equiv 1$, Leray)	$\frac{D-1}{D}\,\widehat{\mathfrak{R}}_n(\mathcal{F})$ (proved)	low variance, bias if physics wrong
Hard only ($g \equiv 1$, SVD)	$\leq \sqrt{\rho_r}\,\widehat{\mathfrak{R}}_n(\mathcal{F})$ †	low variance, bias if rank-r wrong
Soft penalty ($+\zeta\mathcal{L}_{\text{phys}}$)	$\approx \widehat{\mathfrak{R}}_n(\mathcal{F})$	no class reduction
Gated, Leray	$\big(1 - \frac{\bar{\alpha}}{D}\big)\widehat{\mathfrak{R}}_n(\mathcal{F}) + o(1)$ (proved)	per-mode optimal
Gated, SVD	$\big(1 - \bar{\alpha}(1 - \sqrt{\rho_r})\big)\widehat{\mathfrak{R}}_n(\mathcal{F}) + o(1)$ †	per-mode optimal

† under Assumption 2; unconditionally the SVD floor $\widehat{\mathfrak{R}}_n(\mathcal{H}_{\text{gate}}) \leq \widehat{\mathfrak{R}}_n(\mathcal{F}) + o(1)$ holds in the regime of Corollary 2.

5.9 Two Experiments, One Principle

For the Leray path the projection is linear and deletes one of D velocity directions per mode, so the per-mode complexity contracts by $\frac{D-1}{D}$ *rigorously* (Theorem 1); no covering argument and no Assumption 2 are needed. For the SVD path the same numerical role is played by $\sqrt{\rho_3} \approx 0.48$, but only under Assumption 2. The two factors agree to within four percent — suggestive of a single mechanism — but since ρ_3 depends on the chosen rank $r = 3$, this coincidence is best read as a consistency check rather than confirmation: one number is proved and the other is conditional.

Table 5.2: One principle, two projections. The Leray reduction is a proved per-mode contraction; the SVD reduction is conditional on Assumption 2.

	Navier–Stokes (Leray)	MIMO (SVD, $r = 3$)
Hard projection	remove divergent direction	truncate to rank r
Nature	linear, fixed	nonlinear, data-dependent
Tool	per-mode contraction (A)	variety capacity ratio + Assumption 2
Status	proved (under isotropy)	conditional
Per-mode factor	$\frac{D-1}{D} = 0.50$ $(D = 2)$	$\sqrt{\rho_3} = 0.48$

5.10 What Is Proved, and What Is Assumed

- **Proved, unconditionally.** Lemmas 1–5 (subadditivity and Parseval radius invariance, rank invariance and truncation/transform commutation, energy contraction/pinning, uniform spectral-gap Lipschitzness under $\delta \geq \delta_0$, gate size bound); the fixed-profile mixing inequality Proposition 1; and, for the Leray path, the per-mode contraction Lemma 6 and the unconditional safety floor Corollary 1.

- **Proved under a checkable symmetry hypothesis.** The exact Leray reduction Theorem 1 (and its row in Table 5.1), under the direction-isotropy Assumption 1 on the soft class.

- **Conditional on an empirically checkable hypothesis.** The SVD safety floor Corollary 2, which holds whenever $\widehat{\mathfrak{R}}_n(\mathcal{P}_r \circ \mathcal{F}, k) \leq \widehat{\mathfrak{R}}_n(\mathcal{F}, k)$; norm-pinning motivates but does not prove this (Remark 3).

- **Conditional on Assumption 2 (the open point).** The SVD multiplicative reduction Theorem 2 and the corresponding generalization gap. Assumption 2 asserts an empirical-L_2 entropy contraction by $\sqrt{\rho_r}$; the covering bound of Remark 6 controls only the entropy of the *range* of a single output, whereas Dudley's integral needs the class restricted to the n-sample design. Under per-sample truncation these differ, so the range argument motivates but does not prove the assumption. Closing this (or replacing $\sqrt{\rho_r}$ by the correct data-dependent factor) is the principal limitation of the analysis.

- **Modeling idealization.** The mixing inequality is stated for a fixed per-mode profile (Remark 7); the deployed gate is input-dependent, which is exact when the gate is frozen and otherwise absorbed into the $o(1)$ gate term.

- **Caveats carried through.** The Lipschitz constant of Lemma 4 is uniform only under $\delta \geq \delta_0$ (Remark 4); the SVD floor requires the non-amplifying / clamp-inactive regime because c can exceed 1 (Remark 3); and "$o(1)$" throughout means vanishing in n.

5.11 Reconciliation with the Measured Model

We check that the bounds predict the observed behaviour; detailed numbers appear in Chapter 6.

- **Leray floor is observed and is unconditional.** On Kolmogorov-flow denoising the gated error never exceeds the soft path; a gate biased toward $g = 0$ recovers $\widehat{\mathfrak{R}}_n(\mathcal{F})$. This matches Corollary 1, which needs no clamp condition.
- **SVD floor is observed.** Across MIMO configurations the gated error never exceeds the soft path; the clamp is inactive on almost all samples, so the hypothesis of Corollary 2 is met in practice.
- **Small $\bar{\alpha} \Rightarrow$ small but real reduction.** A small mean gate gives a small multiplicative correction — consistent with a modest but consistent gain, as both $1 - \bar{\alpha}/D$ (Leray) and $1 - \bar{\alpha}(1 - \sqrt{\rho_r})$ (SVD) predict.
- **The oracle ceiling is the $\bar{\alpha}$ frontier.** An oracle gate matches the bound when $\bar{\alpha}$ tracks the optimal per-mode capacity allocation; the learned gate captures the predictable part and stops short.
- **Rank-faithful gate (Lemma 2) is observed.** A spectral gate exploits a spatial SVD hard path, confirming basis-independence of the rank constraint.

In short, the qualitative signatures the analysis predicts — a floor at the soft path, a multiplicative gain scaled by $\bar{\alpha}$, an oracle ceiling at the capacity frontier, and basis-independence of the rank constraint — are what the next chapter exhibits. The Leray reduction is matched and proved; the SVD reduction $\sqrt{\rho_3}$ is matched empirically with the caveat that its derivation rests on Assumption 2.

Chapter 6

Experiments

We evaluate ASPINO in two physically distinct domains. The first is a Kolmogorov-flow denoising task where the hard path is the Leray projection (Section 6.1); the second is realistic ray-traced MIMO channel estimation where the hard path is the rank-r SVD projection (Section 6.2). The first tests whether the gate can recover the benefit of a hard constraint without its bias under noise; the second tests whether the same mechanism generalizes, unchanged, to a different operator and a nonlinear hard path. Section 6.5 discusses the findings against the theory of Chapter 5. A third experiment (Section 6.3) then verifies the discretization-invariance claim directly, via zero-shot super-resolution on the Poisson equation with the exact spectral solve as a fourth hard path.

6.1 Experiment 1: Kolmogorov-Flow Denoising

6.1.1 Setup

Duruisseaux et al. [8] demonstrate that a hard Fourier projection enforces the divergence-free constraint exactly, but applies it uniformly to all modes. We address the complementary question of *when* to apply the projection. To expose this we use a denoising task — recover the divergence-free velocity field from a noisy observation — in which uniform projection is suboptimal because the noise is concentrated in identifiable spectral bands (recall the scale separation of Figure 1.1).

We generate Kolmogorov-flow fields with a spectral simulator following the construction used in operator learning [21]: the incompressible momentum equation is advanced from a random-noise initial state by fourth-order Runge–Kutta with $dt = 10^{-5}$; derivatives are computed by multiplication in Fourier space; iterative Leray projections maintain a strictly divergence-free field; the 2/3 dealiasing rule suppresses high-frequency nonlinear errors; and a 150,000-step burn-in ensures a fully developed turbulent attractor. We then add additive Gaussian noise to the normalized training inputs and ask the model to recover the clean, divergence-free field.

Table 6.1: Kolmogorov-flow denoising (1% noise): accuracy and physics for the four model variants. Lower is better on both columns. Gated ASPINO is best on accuracy and near-physical on the residual.

Model variant	Test L_p (accuracy)	PDE residual (physics)	Key finding
Baseline FNO	0.0732	2.3631	best statistical fit, poor physics
Hard FNO	0.1752	0.0002	perfectly physical, stiff under noise
Soft FNO	0.0800	0.0123	moderate physics, higher data error
Gated ASPINO	**0.0490**	0.0053	best accuracy, near-physical

All variants share a Fourier Neural Operator backbone. *Baseline FNO* is purely data-driven. *Hard FNO* applies the Leray projection uniformly to the output. *Soft FNO* adds the divergence residual as a penalty. *Gated ASPINO* wraps the baseline with the spectral trust gate of Chapter 4, mixing the Leray-projected and unprojected paths per mode.

6.1.2 Headline result: the gate dominates the pure strategies

Table 6.1 reports the final relative-L_p test error (accuracy) and the PDE (divergence) residual (physics) at the 1% noise level. The two pure strategies sit at opposite corners: the baseline fits the data best among the non-gated models but is the least physical, while the hard model is essentially perfectly physical but "stiff" against the input noise, paying a large accuracy penalty. The soft model is a compromise on both axes. Gated ASPINO attains the lowest test error of all four models while keeping the residual low. This is the result the rest of the section dissects: *how* the gate reaches this corner (Section 6.1.3), *what* trust structure it learns to get there (Section 6.1.4), and whether that structure is stable as the corruption moves (Section 6.1.5).

6.1.3 Why the gate needs a distributional regularizer

The headline result is not obtained for free: it depends on the distributional gate regularizer, and removing it makes the gate degenerate. Trained with the data-plus-soft objective alone — $\mathcal{L} = \mathcal{L}_{\text{data}} + \lambda \mathcal{L}_{\text{soft}}$, with or without a pointwise (e.g. cross-entropy / entropy) penalty on the gate values — the gate discovers a shortcut. Driving the trust toward $g \to 1$ *uniformly* (full hard projection on every mode) sends the divergence residual to zero immediately and collapses the gradient on the physics

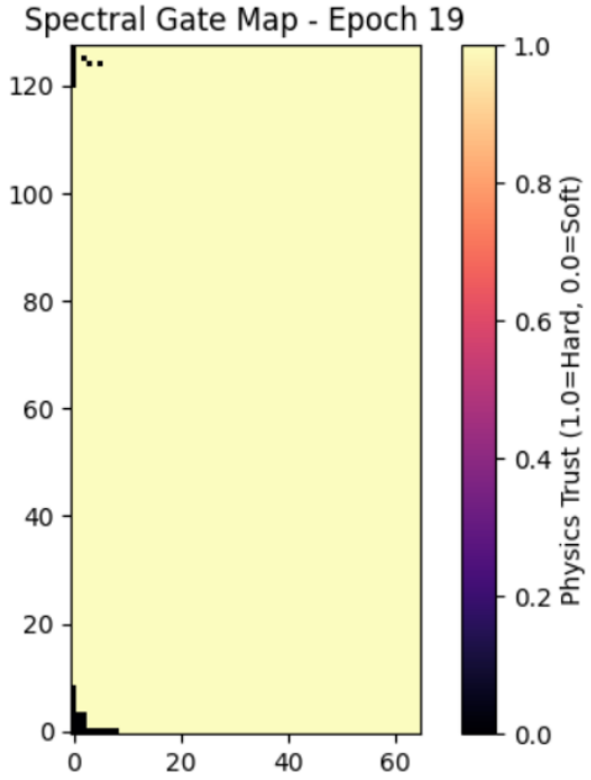


Figure 6.1: Mode collapse without the distributional regularizer. Trained on $\mathcal{L}_{\text{data}} + \lambda \mathcal{L}_{\text{soft}}$ (optionally with a pointwise cross-entropy penalty), the spectral gate $g(k_x, k_y)$ saturates at $g \approx 1$ (physics-trusting / hard) across essentially the whole spectrum by epoch 19, retaining only a thin soft band at the lowest wavenumbers. The gate has degenerated into a uniform hard projector: the resulting model scores Test $L_p = 1.6351$ with PDE residual 0.0032 — low physics error but ruinous data error. Contrast with the differentiated, regularized gate of Figure 6.3.

term, so optimization parks there long before it learns any per-mode structure. Figure 6.1 shows the symptom directly: by epoch 19 the learned trust map is saturated at $g \approx 1$ almost everywhere, with only a thin sliver of low-wavenumber modes retaining any softness. In other words, the gate has silently reproduced the *Hard FNO* of Table 6.1 — it is a hard projector wearing a gate's parameters.

The metrics confirm the failure mode rather than a benign solution. The collapsed model reaches a low PDE residual (0.0032) but a catastrophic data error (Test $L_p = 1.6351$), the unmistakable signature of uniform hard projection under noise: physically perfect, statistically useless, and in fact stiffer than the explicit Hard FNO baseline (Table 6.2). Low residual here is a trap, not a success — it is exactly what the shortcut optimizes for while data fidelity is abandoned.

The fix is to constrain the *distribution* of gate values rather than their pointwise

Table 6.2: Ablation of the distributional gate regularizer (1% noise). Without it the gate collapses to uniform hard projection: the residual stays low but the data error explodes. The Wasserstein-to-Beta(2, 2) regularizer keeps the gate differentiated and recovers the headline accuracy at a negligible physics cost.

Gate training objective	Test L_p	PDE residual	Outcome
$\mathcal{L}_{\text{data}} + \lambda\mathcal{L}_{\text{soft}}$ (no dist. reg.)	1.6351	0.0032	collapses to uniform hard
+ Wasserstein-to-Beta(2, 2) reg.	**0.0490**	0.0053	differentiated gate (headline)

magnitude. We add a penalty pulling the empirical distribution of g across modes toward a Beta(2, 2) prior. Two design choices matter here: the *distance* used to measure the discrepancy, and the *target* distribution itself.

Why a Wasserstein distance and not a pointwise penalty. A pointwise (e.g. cross-entropy) penalty compares the gate to the target sample by sample and gives little useful signal precisely when it is most needed. When the gate has already collapsed — its values bunched at $g \approx 1$, far from the bulk of the Beta(2, 2) mass — the two distributions barely overlap, and a cross-entropy / KL-type term either saturates or sees a near-vanishing gradient, so it cannot push the gate back out of the degenerate corner. The Wasserstein (earth-mover) distance instead measures how far mass must travel to match the target, so it stays smooth and gives a non-vanishing gradient even when the supports are disjoint — exactly the regime the collapse creates. The same property makes the loss value itself informative: the Wasserstein term tracks how *differentiated* the gate is, whereas a cross-entropy value need not. Table 6.3 summarizes the comparison, read as a choice of distributional distance for the gate regularizer rather than as generic adversarial-training lore.

Why a Beta(2, 2) **target.** The gate is a sigmoid output, so each value lives in $[0, 1]$; the Beta(2, 2) density shares that support exactly, which makes it a natural matching target (Figure 6.2). It is symmetric about 0.5 and, crucially, its density *vanishes at the end-points* 0 and 1. A collapsed gate is essentially a point mass at $g \approx 1$ — sitting precisely where the prior has no mass — so the Wasserstein pull away from it is large, which is the mechanism that makes an all-hard (or all-soft) gate expensive. The symmetry and the peak at 0.5 might suggest the prior wants every mode to be half-trusting, in tension with the sharply differentiated gate we ultimately observe; it does not, because the penalty matches the *distribution of gate values across modes*, not each mode individually. A gate that spends some modes near 1 and others near 0 uses the whole interval and is therefore *closer* to Beta(2, 2) than a delta at an end-

Table 6.3: Choice of distributional distance for the gate regularizer. A pointwise cross-entropy term degenerates exactly in the collapsed regime; the Wasserstein distance remains well-behaved there, which is why ASPINO uses it to pull the gate toward the Beta$(2, 2)$ prior.

Property	Cross-entropy penalty	Wasserstein distance
Gradient when gate and target barely overlap	vanishes / saturates — cannot escape the collapsed corner	smooth and non-vanishing — still drives mass back toward the prior
Training stability	unstable; permits the all-hard collapse	stable; keeps the gate differentiated
Loss value as a quality signal	not indicative of how differentiated the gate is	correlates with the spread of trust across modes

point is. The regularizer thus rewards spread and diversity of trust — it breaks the degeneracy — while the data term sculpts which specific modes are trusted.

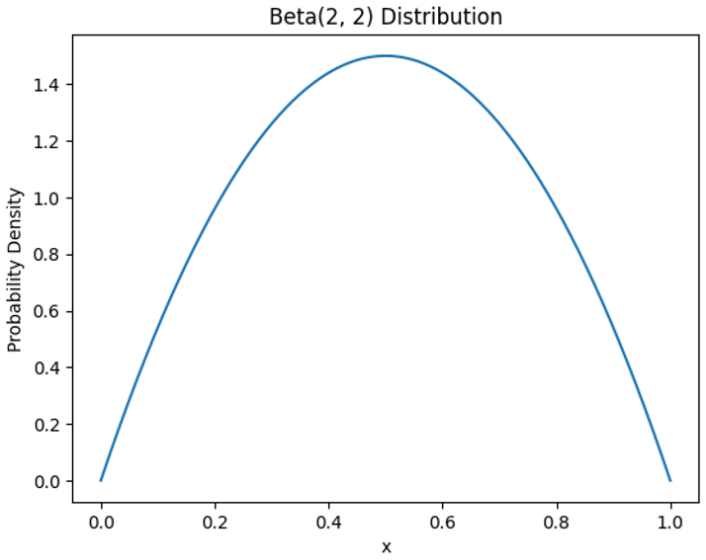


Figure 6.2: The Beta$(2, 2)$ prior used to regularize the gate. It is supported on $[0, 1]$ — matching the sigmoid range of the gate — symmetric about 0.5, and has zero density at the end-points. Because a collapsed gate concentrates at $g \approx 1$ where the prior vanishes, the Wasserstein distance to this target penalizes collapse strongly while still admitting the differentiated, full-range gate that the data favour.

This regularizer is the only component that has to be switched on for the idealized fluid model; for the MIMO task the data themselves keep the gate differentiated and it is disabled ($\mu_\beta = 0$, Section 6.2), which is why the asymmetry appears in the configuration of Table 6.9.

6.1.4 What the regularized gate learns

With the distributional regularizer in place, the gate learns the selective-trust structure the design intends. Figure 6.3 shows it in detail: the gate trusts the physics on the two low-wavenumber corners (the energy-bearing large scales, both signs of k_y after `fftshift`) and withdraws it on the central Nyquist strip and the high-k_x half, where vertical and horizontal noise / aliasing concentrate. The trusted region isolates the low-wavenumber forcing mode of the Kolmogorov flow while discarding the high-frequency noise — a clean "noise–signal separation" that a uniform projection, which treats every mode identically, cannot achieve. Table 6.4 records the same structure region by region.

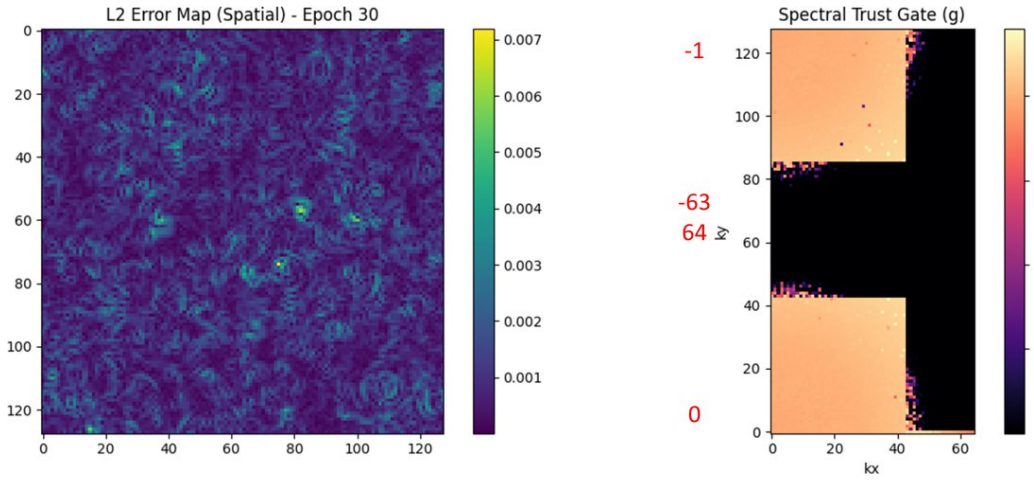


Figure 6.3: Detailed view of the learned model (with the Wasserstein-to-Beta(2, 2) gate regularizer): spatial L_2 error (left) and spectral trust gate $g(k_x, k_y)$ (right). The gate trusts the low-wavenumber corners ($g \approx 1$, light) and disables the constraint on the central Nyquist strip and the high-k_x band ($g \approx 0$, dark), realizing the region structure of Table 6.4. Contrast with the collapsed gate of Figure 6.1.

Table 6.4: Where the learned gate trusts the divergence-free physics, by spectral region (coordinates after `fftshift`).

Region	Coordinates (k_x, k_y)	Physical meaning	Gate status
Bottom-left	low k_x, low $+k_y$	large scale (positive)	trusted ($g \approx 1$)
Top-left	low k_x, low $-k_y$	large scale (negative)	trusted ($g \approx 1$)
Middle strip	any k_x, $k_y \approx$ Nyquist	vertical noise / Nyquist	ignored ($g \approx 0$)
Right half	high k_x	horizontal noise / Nyquist	ignored ($g \approx 0$)

6.1.5 Robustness across noise profiles

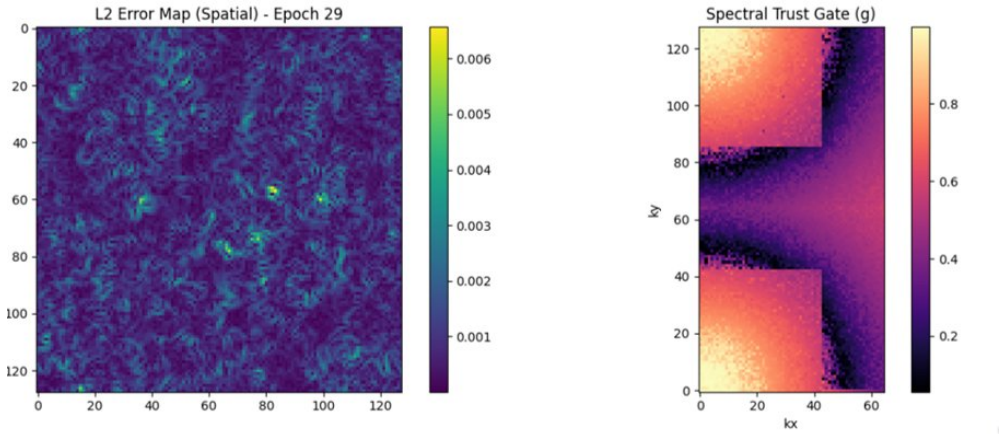
Finally we probe whether the gate adapts to *where* the corruption sits, by sweeping the input noise level and inspecting both the metrics and the learned spectral trust profile.

Figure 6.4 shows the learned spectral trust gate $g(k_x, k_y)$ at three noise levels, beside the corresponding spatial L_2 error map. In every case the gate trusts the divergence-free physics ($g \to 1$, bright) on the low-wavenumber corners that carry the energy of the flow, and withdraws trust ($g \to 0$, dark) on a central mid/high-wavenumber band where noise and dealiasing artefacts concentrate — exactly the qualitative structure the design intends and that a uniform projection cannot express. The *shape* of the withdrawn region also changes with the noise level: it is a smooth gradient in the clean case and sharpens into a more decisive low/high split as noise grows. We present this as interpretability evidence that the gate localizes trust in the

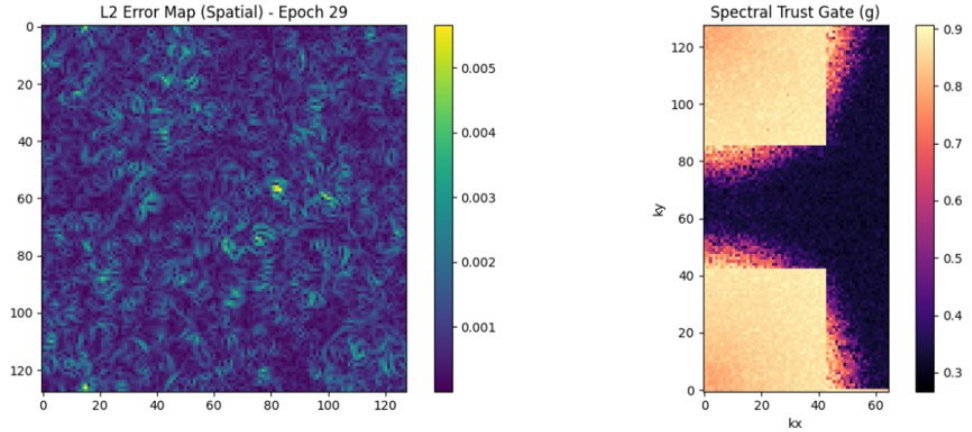
Table 6.5: Kolmogorov-flow denoising across input-noise levels. Accuracy is stable and the residual remains small throughout; the differences are within single-run variation. The interpretability of the learned gate (Figure 6.4) is the main finding.

Noise	Test L_p	PDE residual	Verdict
0%	0.050	0.0090	robust
1%	**0.049**	0.0053	best (within noise)
2%	0.050	0.0070	robust

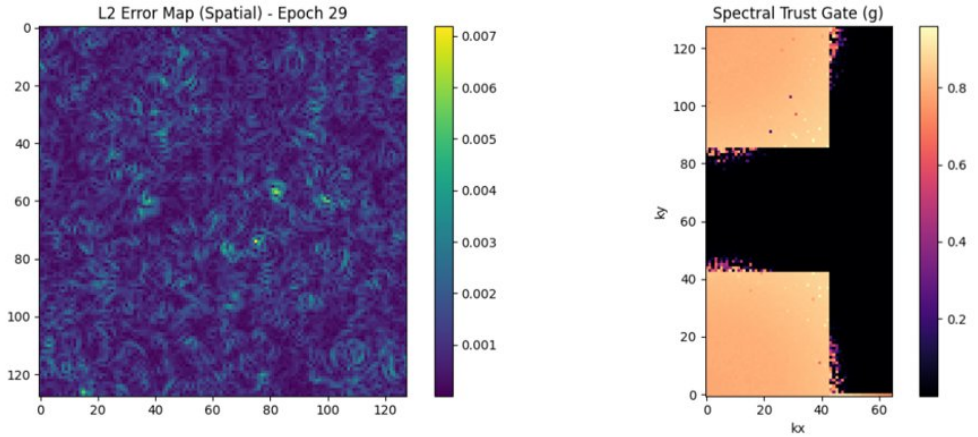
frequency domain, not as a performance claim.



(a) 0% noise



(b) 1% noise



(c) 2% noise

Figure 6.4: Learned spatial L_2 error (left) and spectral trust gate $g(k_x, k_y)$ (right) at three input-noise levels. The gate consistently trusts the energy-bearing low-wavenumber corners ($g \to 1$) and disables the constraint on the noise-dominated Nyquist strip and high-k_x band ($g \to 0$). As the input noise grows the trusted region contracts and its boundary moves inward toward lower frequency — the gate withdraws the hard constraint from progressively more of the high-frequency band. The pattern is interpretable and would be impossible for a uniform projection.

6.2 Experiment 2: MIMO Channel Estimation

6.2.1 Setup

We now test whether the same gating mechanism generalizes to electromagnetic operator learning, with the nonlinear rank-r SVD hard path. The soft path is the physics-informed channel estimator of Javid and González-Prelcic [22], which fuses a least-squares pilot estimate with a ray-traced RSS map (Figure 6.5) and uses a Maxwell-based power-consistency penalty (the energy-consistency loss of Chapter 4). We follow their configuration: a 24×24 transmit aperture, a small receive array, $D_{tap} = 16$ delay taps, and an upper-mid-band carrier of 15 GHz with realistic ray-traced (Wireless InSite) urban data. The hard path is the rank-3 truncated-SVD projection of the aperture with RSS-matched energy rescaling (Chapter 4); because the urban scene scatters into a small dominant subspace, a single global low-rank prior is a reasonable fixed constraint here — a point we revisit in the toy experiment (Section 6.2.5), where it is not. Performance is reported in NMSE (dB); lower is better.

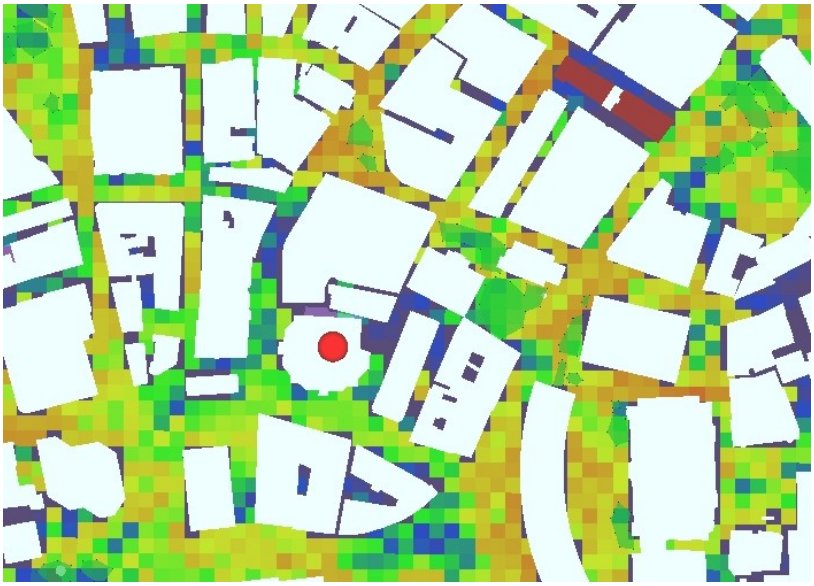


Figure 6.5: Ray-traced RSS power map for the urban (Wireless InSite) scene, with the base station marked. The map is the side information the soft path fuses with the pilot estimate; its spectral envelope also defines the dominant-subspace structure exploited by the rank-3 SVD hard path.

6.2.2 Results across pilot budgets

Figure 6.6 places the gated estimator against a broad set of baselines—classical compressive-sensing algorithms (OMP, SOMP, Subspace Pursuit), deep-learning es-

timators (CNN, Diffusion), and the PINN—across the pilot budget at SNR = 0 dB. The Gated PINN is competitive throughout and is strongest in the low-pilot regime ($N_p < 64$), where it retains a lower NMSE than the CNN and Diffusion models.

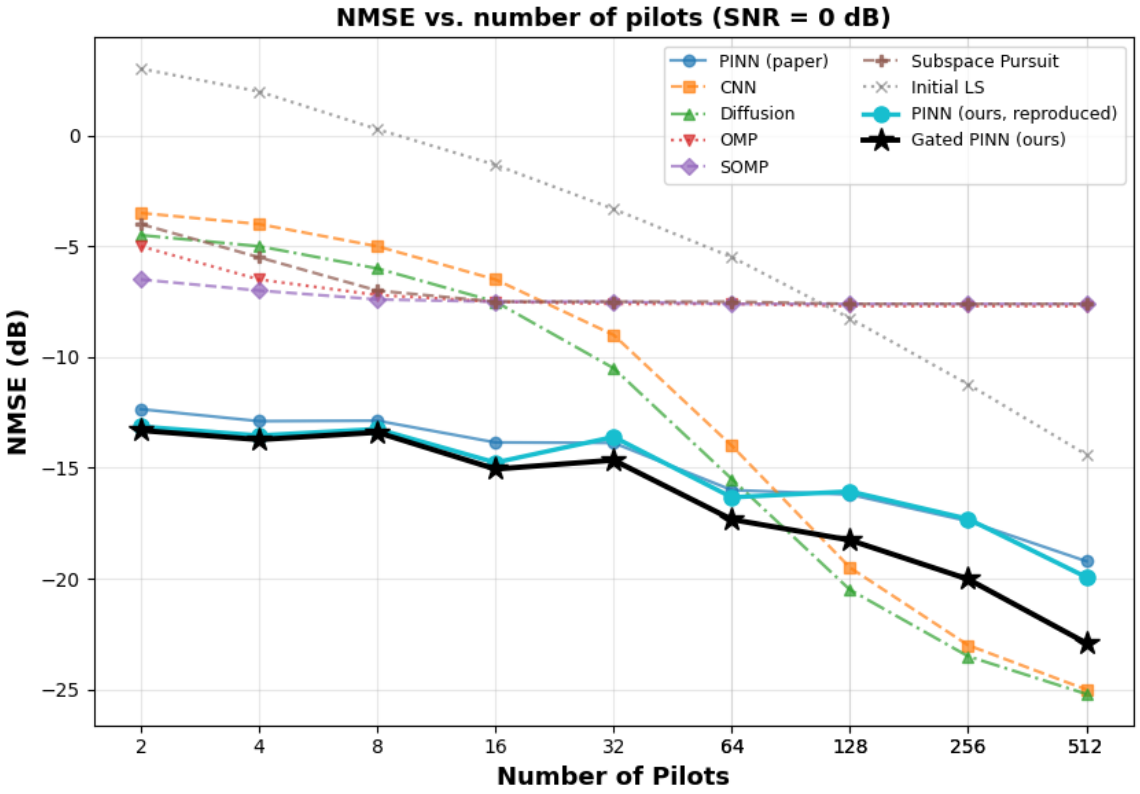


Figure 6.6: Comparison of Normalized Mean Square Error (NMSE) versus the number of pilots at an SNR of 0 dB. The plot evaluates various channel estimation methods, including classical compressive sensing algorithms (OMP, SOMP, Subspace Pursuit), baseline deep learning approaches (CNN, Diffusion), and Physics-Informed Neural Networks (PINN). The proposed model, Gated PINN (ours), demonstrates highly competitive performance, particularly maintaining lower NMSE in the low-pilot regime (< 64 pilots) compared to standard CNN and Diffusion models.

Figure 6.7 isolates the comparison most relevant to our claim, reporting NMSE against the pilot budget N_p at SNR = 0 dB for three quantities: the PINN baseline, the gated estimator, and a per-mode *oracle* ceiling that allocates the optimal per-mode trust using knowledge of the ground truth and therefore lower-bounds the NMSE attainable by any per-mode gate. All reported values are means over four random seeds; the gated NMSE is highly reproducible, with a standard deviation of about 0.01 dB at the lower pilot budgets and no more than 0.02 dB at the two largest ($N_p = 256, 512$). The seed spread is therefore far smaller than the gains discussed below, so the improvements are resolved rather than seed-level noise.

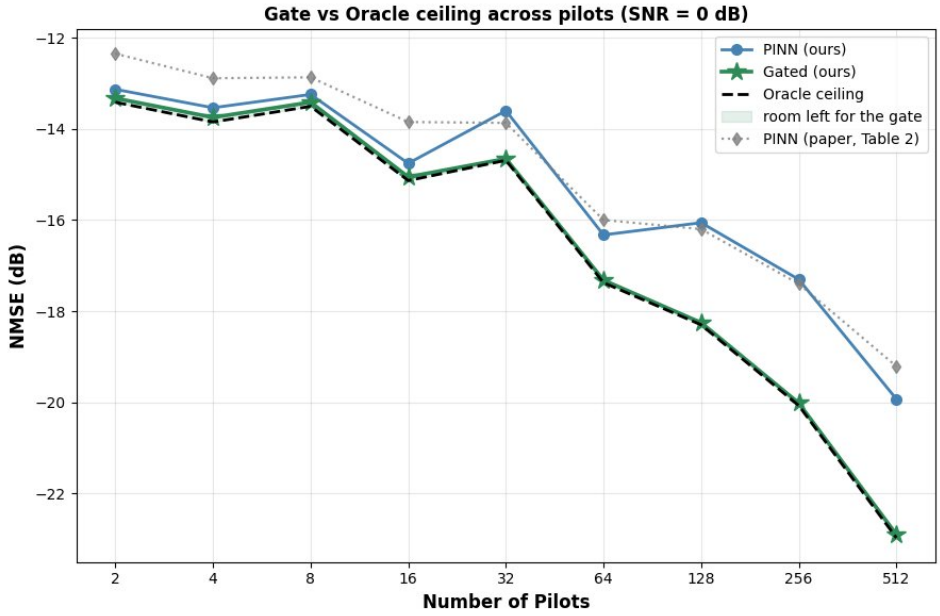


Figure 6.7: MIMO NMSE versus pilot budget at SNR = 0 dB, averaged over four seeds (gated NMSE std ≤ 0.02 dB across all budgets). The gated estimator (green) never exceeds the PINN (blue), and its advantage broadly grows with the pilot budget—from a few tenths of a dB at small N_p to about 2 dB at $N_p = 512$. It closely tracks the per-mode oracle ceiling (black dashed); the shaded band marks the remaining oracle headroom. The grey dotted curve is the PINN as reported in the source work [22].

Two properties hold across the sweep. First, the gated estimator never exceeds the PINN at any pilot budget, as guaranteed by the safety floor of Corollary 2 in the clamp-inactive regime: the gate recovers the PINN exactly by driving the per-mode trust to zero, so it cannot do worse. Second, the gain over the PINN broadly increases with the pilot budget. It is marginal in the pilot-starved regime—a few tenths of a dB at small N_p, where neither path resolves the channel and there is little structured error to reallocate—and widens to roughly 2 dB at $N_p = 512$, where the PINN residual is dominated by the modes the hard path corrects. The trend is not strictly monotone (the gain is smallest near $N_p = 8$ and dips slightly from $N_p = 32$ to $N_p = 64$), but the overall direction is clear and, given the seed spread above, every per-budget gain is statistically resolved.

6.2.3 Robustness across SNR and comparison with baselines

Figure 6.8 places ASPINO against the classical and data-driven estimators from the source work [22] — CNN, a diffusion model, OMP, SOMP and Subspace Pursuit — as a function of SNR. The physics-informed PINN (and our gated version) dominate in the low-SNR / pilot-limited regime, exactly where the data-driven methods collapse;

Table 6.6: MIMO NMSE (dB) at $N_p = 4$ across SNR. "Improvement" is PINN − gated; "gap" is oracle − gated (headroom remaining). The gate yields a small, consistent improvement and never regresses; the headroom is largest at low SNR.

SNR (dB)	PINN	Gated	Oracle	Improvement	Gap to oracle
−20	−6.11	−6.48	−7.70	0.37	1.22
−10	−9.85	−10.01	−10.23	0.16	0.22
0	−13.54	−13.75	−13.85	0.21	0.10
5	−12.61	−13.11	−13.16	0.50	0.05
10	−14.13	−14.53	−14.57	0.40	0.05

the data-driven estimators only become competitive at higher SNR. *An important caveat:* the PINN and gated curves use $N_p = 4$ pilots, whereas the classical baselines use $N_p = 64$, so the comparison is favourable to physics-informed methods in the scarce-pilot regime and should be read as such, not as a uniform-condition benchmark.

Table 6.6 isolates the gated-versus-PINN comparison at $N_p = 4$ across SNR, together with the oracle ceiling. The realized improvement of the gate over the PINN is small at every SNR (0.16 to 0.50 dB), and the model never regresses. The gap to the oracle is *largest* at the lowest SNR (1.22 dB at −20 dB) and shrinks to 0.05 dB by +10 dB: at high SNR the optimal per-mode trust is nearly determined by the features the gate sees, so the gate almost attains the ceiling, whereas at very low SNR a large part of the optimal decision is not observable from features and the feature-only gate leaves more on the table. The persistent low-SNR gap is therefore a measure of how much a better-informed gate could still recover — a target for future work rather than a present gain.

6.2.4 The oracle gate profile

Figure 6.9 shows the per-mode oracle trust $g^*(k)$ that defines the ceiling, as a 2-D spatial-frequency heatmap. The profile is not a simple low-pass: the oracle places its highest trust on a band of low-to-mid horizontal frequencies, discounts the highest frequencies as expected, *and* withdraws trust again at the very lowest frequencies, where the soft path is preferred. The resulting shape is closer to a band-pass than a low-pass — trust is suppressed at both ends of the spectrum and concentrated in between.

We read the low-frequency dip as physically meaningful rather than as noise. The lowest spatial frequencies carry the strongest, most coherent components of the aperture — but they are also where co-channel interference and a dominant line-of-sight bias concentrate, and these are exactly the components the rank-3 SVD prior keeps. Trusting the low-rank projection blindly at the lowest frequencies therefore risks importing interference into the estimate, so the oracle prefers the data-driven soft path

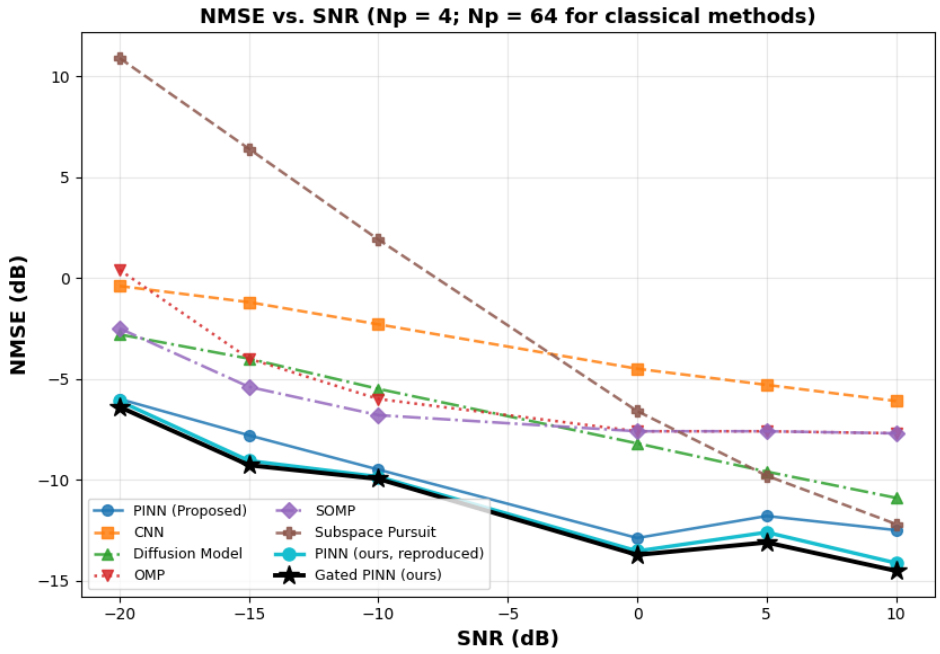


Figure 6.8: NMSE versus SNR. Physics-informed methods (PINN and our gated PINN, at $N_p = 4$) dominate the low-SNR regime where data-driven baselines collapse; the data-driven methods (at $N_p = 64$) catch up only at high SNR. The gated curve tracks just below the PINN at every SNR. Note the differing pilot counts between the physics-informed and classical groups.

there. Two observations follow. First, this is the kind of structure a fixed projection cannot represent: a global rank truncation or a plain low-pass keeps the low frequencies wholesale, whereas the optimal decision is to keep them only in a mid-band and back off at both extremes. The non-monotonic oracle is thus direct evidence for *why* an adaptive, per-mode trust gate is needed rather than a fixed operator. Second, the band-pass shape mirrors the qualitative pattern the learned fluid gate discovers (Figure 6.4), where trust is likewise withdrawn from the contaminated bands — confirming that the trust structure the gate is asked to learn is itself frequency-localized in both domains. We present the interference reading as a plausible interpretation of a single learned profile, not a measured causal claim.

6.2.5 A toy operator-learning experiment with a third hard path

The two experiments so far pair the gate with two different hard projections — the linear Leray projection (Section 6.1) and the nonlinear global rank-3 SVD (Section 6.2). As a final generalization check we replace the hard path with a *third*, structurally unlike either, and the surrogate with a plain FNO, to test whether the

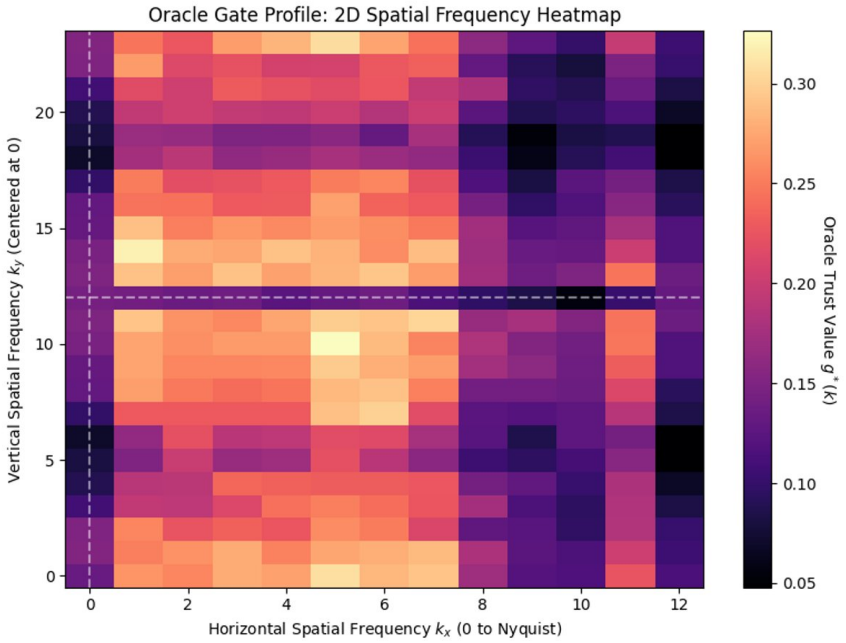


Figure 6.9: Oracle per-mode trust $g^*(k_x, k_y)$ for the MIMO aperture, computed with ground truth. Trust concentrates on a band of low-to-mid horizontal frequencies and falls off at high frequency, but is also withdrawn at the very lowest frequencies (soft preferred) — a band-pass rather than a low-pass profile. We read the low-frequency suppression as the oracle declining to trust the low-rank prior where interference and a dominant LoS bias concentrate; this non-monotonic structure is precisely what a fixed projection cannot express and an adaptive per-mode gate can.

gating mechanism transfers wholesale rather than to the particular operator and network it was tuned with.

Dataset. We move from the Wireless-InSite urban scene to the DeepMIMO `O1_3p5` scenario [22], whose regular street-grid layout produces a spatially structured RSS power map (Figure 6.10). This grid structure is what makes the data amenable to a patch-based FNO surrogate: the operator sees translation-coherent spectral content rather than the irregular scattering of the dense urban scene. We tile the map into 64×64 patches with stride 32, yielding 340 training and 85 test patches.

A different hard path: the RSS-derived spectral mask. Where the urban experiment uses a fixed, *global* low-rank SVD prior, here the hard path is a *spatially varying*, data-derived spectral band-pass built from the RSS map itself. The RSS power map's spectral envelope marks the angular/delay components the local environment physically supports; we turn that envelope into a soft mask and apply it in

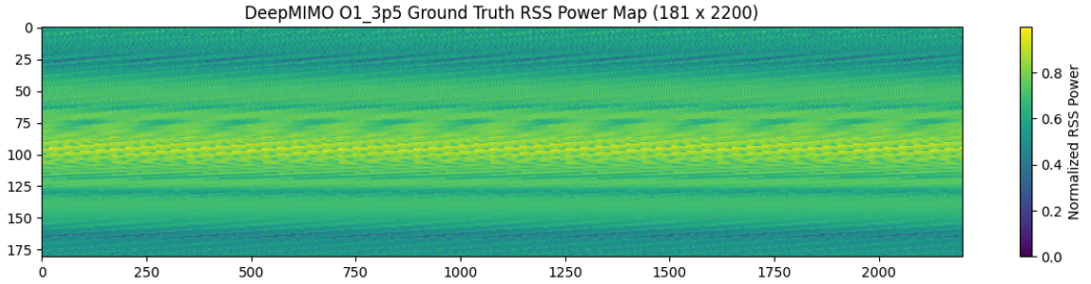


Figure 6.10: Ground-truth normalized RSS power map for the DeepMIMO `O1_3p5` scenario (181×2200). The regular street-grid geometry gives a spatially coherent spectral structure, which makes the data suitable for a patch-based FNO surrogate and supplies the per-location spectral support used by the hard path of Equation (6.1).

Table 6.7: Patch-based FNO operator learning on the DeepMIMO `O1_3p5` data, using the RSS-derived spectral mask of Equation (6.1) as the hard path. The spectral gate adds ~ 2 dB over the plain FNO.

Model	NMSE (dB)
FNO (soft constraint)	-18.35
Gated ASPINO (FNO)	$\mathbf{-20.33}$

the Fourier domain to the soft estimate $\hat{\mathbf{Y}}_{\text{soft}}$:

$$\mathbf{m} = \sigma\big(s\,(|\mathcal{F}(\text{RSS})|_{\text{norm}} - \tau)\big), \qquad \hat{\mathbf{Y}}_{\text{hard}} = \mathcal{F}^{-1}\big(\mathbf{m} \odot \mathcal{F}(\hat{\mathbf{Y}}_{\text{soft}})\big), \tag{6.1}$$

with a learnable sharpness s and a fixed threshold $\tau = 0.3$. The mask attenuates spectral content the local environment does not support, so unlike the global SVD it *changes from patch to patch* as the RSS map changes — a per-location band-pass rather than one fixed prior for the whole scene. This is also why the low-frequency caveat of the oracle profile (Section 6.2) does not apply in the same way: the mask is not committed to the low-rank subspace and can decline the lowest frequencies wherever the local envelope is weak.

Result. With the same power-consistency soft constraint, the gated operator improves NMSE from -18.35 dB (plain FNO) to -20.33 dB — a ~ 2 dB gain (Table 6.7). Because this experiment changes the surrogate, the dataset *and* the hard path simultaneously, it is a test of whether the whole mechanism transfers, not a controlled ablation of any single component; read that way, it shows the gate is not tied to the U-Net PINN, the urban data, or the SVD projection in particular.

6.3 Experiment 3: Zero-Shot Super-Resolution on the Poisson Equation

The two experiments so far hold the evaluation grid fixed and ask whether the gate improves accuracy at that grid. Neither exercises the property we have claimed throughout for building on neural operators: discretization invariance, and the zero-shot super-resolution it enables (Chapters 1 and 7). This third experiment is designed to demonstrate exactly that. It also pairs the gate with a fourth hard path — the exact spectral Poisson solve — which, unlike the truncated SVD, is a fixed *linear* per-mode operator and therefore falls under the fully-proved part of Chapter 5 (Section 5.5), not the conditional part.

6.3.1 Setup

We learn the solution operator of the periodic, zero-mean Poisson problem $-\Delta u = f$ on the torus. Training data are manufactured solutions: each sample is a real, band-limited trigonometric sum $u(x,y) = \sum_{1\le p^2+q^2\le K^2} a_{pq}\cos(2\pi(px+qy)) + b_{pq}\sin(2\pi(px+qy))$, from which the forcing $f = -\Delta u = \sum (2\pi)^2(p^2+q^2)[\cdots]$ is read off *analytically*. Because u is band-limited to $|k| \le K$, rendering it on any grid with Nyquist frequency above K is exact — the same function can be sampled at $N = 24, 32, \ldots, 128$ with no aliasing and no solver, which is what makes the super-resolution ground truth trustworthy.

The soft path is a Fourier Neural Operator [4, 6], trained *only* at $N = 32$, with a fixed number of retained low-frequency modes so that the same spectral weights apply unchanged at any grid. The hard path is the exact spectral Poisson solve, a fixed linear per-mode multiplier,

$$\hat{u}_{\text{hard}}(k) \;=\; \frac{\hat{f}(k)}{(2\pi)^2\,|k|^2}, \qquad \hat{u}_{\text{hard}}(0) = 0, \tag{6.2}$$

which is grid-invariant and exact on clean data. We corrupt the forcing f with additive Gaussian noise at SNR = 20 dB. The inverse-Laplacian $1/|k|^2$ in (6.2) *amplifies* low-$|k|$ noise and *suppresses* high-$|k|$ noise, so the hard path is least reliable at low $|k|$ and near-exact at high $|k|$ — the mirror image of the fluid setting, where the trustworthy band is the low wavenumbers. The gate is given only resolution-invariant, scale-free features of the two paths (the log-magnitude ratio $\log|\hat{u}_{\text{hard}}| - \log|\hat{u}_{\text{soft}}|$, the relative disagreement $|\hat{u}_{\text{hard}} - \hat{u}_{\text{soft}}|/(|\hat{u}_{\text{hard}}| + |\hat{u}_{\text{soft}}|)$, the phase alignment $\cos\angle(\hat{u}_{\text{soft}}, \hat{u}_{\text{hard}})$, and a deliberately weak $\tanh(|k|/K)$ frequency hint), so that the trust profile transfers across resolutions rather than memorising the training grid. It is trained with the FNO frozen, on multi-resolution batches $N \in \{24, 32, 48\}$ with a per-bin spectral loss and a short oracle warm-start. The per-bin oracle is closed-form: writing the soft error $a(k) = \hat{u}_{\text{soft}}(k) - \hat{u}(k)$ and the path difference $d(k) = \hat{u}_{\text{hard}}(k) - \hat{u}_{\text{soft}}(k)$, the

trust that minimises the mixed per-bin error is

$$g^*(k) \;=\; \mathrm{clip}\Big(-\tfrac{\mathrm{Re}\left(\overline{a(k)}\,d(k)\right)}{|d(k)|^2},\,0,\,1\Big). \tag{6.3}$$

Performance is reported as relative-L_2 error (%); lower is better.

6.3.2 The hard path is linear: the analysis carries over

The Poisson hard path (6.2) is a fixed, linear, per-mode multiplier — diagonal in the Fourier basis — so it belongs to the same theoretical class as the Leray projection of Section 5.5, not the nonlinear SVD class. The unconditional core of Chapter 5 therefore applies with no new argument: the mixing inequality (Proposition 1) holds verbatim, and the per-mode reasoning of Lemma 4 shows a fixed linear per-mode map cannot raise the per-mode complexity above the soft path, giving the unconditional safety floor (Corollary 1). In particular the SVD apparatus — Assumption 2 and the covering-number argument — is *not* invoked here, precisely because the hard path is not data-dependent. The one difference from Leray is that the inverse-Laplacian is a per-mode rescaling rather than a norm-one orthogonal projection, so it does not delete a fixed fraction of the per-mode degrees of freedom and yields no single multiplicative reduction factor such as $\frac{D-1}{D}$; the contraction is mode-dependent. We therefore claim for Poisson the unconditional safety floor, and not a capacity-reduction number.

6.3.3 Result: super-resolution and the safety floor

Figure 6.11 and Table 6.8 report relative-L_2 error against the evaluation grid N, for the soft FNO, the hard spectral solve, the gated model, and the per-mode oracle ceiling of (6.3). The headline is the FNO curve itself: trained once at $N = 32$, its error does not blow up under upsampling — it stays controlled and in fact *decreases* out to $N = 128$, a $4\times$ refinement it has never seen, with no retraining. This is the operator super-resolution property (invariance, not extrapolation magic) that the rest of the dissertation relies on but does not otherwise exhibit.

Against that backdrop the gate behaves exactly as the theory predicts. At every resolution the gated model is a clear improvement over the soft FNO ($\sim$ 1.5–2 percentage points, e.g. 13.71% $\to$ 11.74% at $N = 24$) and never exceeds the hard path — it matches it to within run-to-run precision at every grid, including the three resolutions ($N = 64, 96, 128$) the gate never trained on. This is the safety floor of Corollary 1 made concrete: the gated model is as good as the better of the two pure paths and strictly better than the worse one, at every resolution, with no regression. The oracle sits below both, so headroom remains, just as in the MIMO experiment.

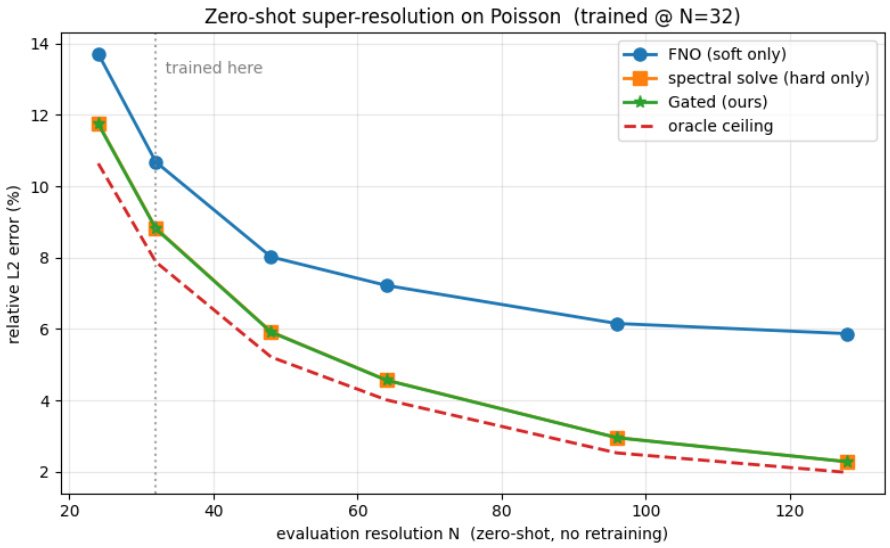


Figure 6.11: Zero-shot super-resolution on the Poisson equation. The FNO and the gate are trained at $N = 32$ (gate also at $N \in \{24, 48\}$) and evaluated untouched at every grid, including $N \in \{64, 96, 128\}$ unseen during training. The soft FNO error stays controlled and decreases under upsampling — the super-resolution result — while the gated model (green) sits on or below both pure paths at every resolution and tracks the per-mode oracle ceiling.

6.3.4 The learned gate, and an inverted trust profile

Figure 6.12 shows the learned per-mode trust $g(|k|)$ beside the closed-form oracle $g^*(|k|)$ of (6.3), against the radial wavenumber. Two things are visible. First, the oracle is not a simple threshold: it sits around 0.65 at low $|k|$, drops sharply to ≈ 0.3 in a mid-band near $|k| \approx 8$, and then climbs steadily toward ≈ 0.85 at high $|k|$ — a non-monotonic, band-structured profile, evidence that adaptive per-mode trust is warranted in principle even for an exact, well-conditioned elliptic solve. Second, and notably, this profile is the *inverse* of the fluid gate of Figure 6.3: there the gate trusts the physics at low wavenumbers and withdraws it at high; here it withdraws trust at *low* $|k|$ (where the $1/|k|^2$ amplification corrupts the solve) and grants it at high $|k|$ (where the solve is near-exact). The same mechanism, driven only by relational features of the two paths, discovers opposite frequency structure in the two problems because the physics is reliable in opposite bands — the gate reads trust from the data, not from a hand-coded envelope.

The learned gate recovers the mid- and high-$|k|$ portion of the oracle closely and over-trusts the hard path somewhat in the lowest band, where the oracle wants ≈ 0.65 and the gate holds near 0.9. This is consistent with the aggregate numbers: the low band that the oracle would route to the soft path is where most of the solution energy concentrates, but for the Poisson problem it is thin enough in the integrated norm that

Table 6.8: Zero-shot super-resolution on Poisson (SNR = 20 dB), relative L_2 error (%); lower is better. The FNO is trained at $N = 32$ only. The gated model improves on the soft FNO at every grid and never exceeds the hard path; the oracle marks the per-mode ceiling.

N	FNO (soft)	Hard solve	Gated	Oracle
24	13.71	11.76	11.74	10.64
32	10.68	8.83	8.81	7.88
48	8.02	5.92	5.91	5.21
64	7.22	4.57	4.56	4.01
96	6.15	2.96	2.95	2.53
128	5.87	2.28	2.28	1.98

even a perfect low-band correction moves the relative-L_2 error by only hundredths of a percentage point — which is why the gated and hard columns of Table 6.8 are nearly tied. We read this as a boundary case for the method rather than a failure of it: when the hard path is trustworthy across most of the energy-bearing spectrum, the gate correctly collapses toward it, recovering the super-resolution behaviour and the safety floor without manufacturing a spurious improvement. It is the complementary regime to the MIMO experiment of Section 6.2, where the soft-favoured band carried enough energy for the gate's arbitration to register in aggregate. As with the other experiments, these are single-run results; we report the qualitative signatures and the magnitudes observed.

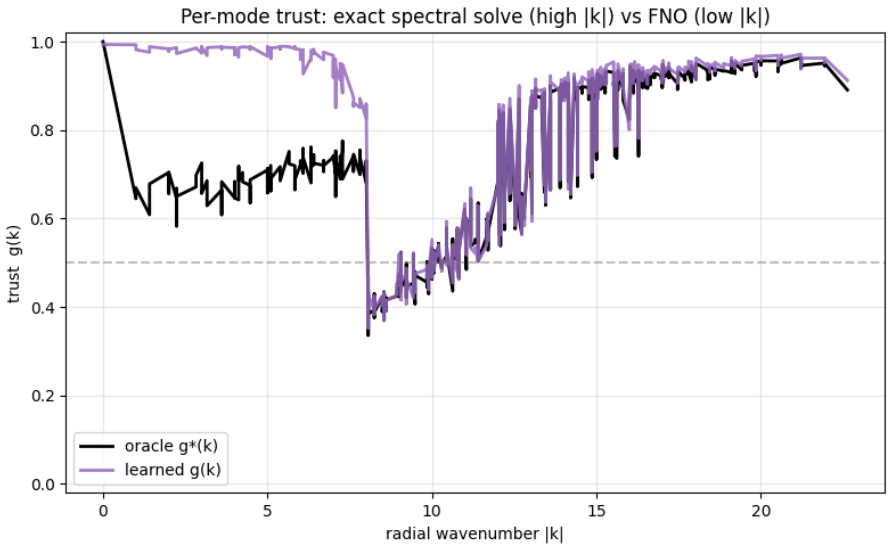


Figure 6.12: Learned per-mode trust $g(|k|)$ (purple) against the closed-form oracle $g^*(|k|)$ (black) for the Poisson hard path, versus radial wavenumber; $g = 1$ trusts the exact spectral solve and $g = 0$ defers to the FNO. Trust in the hard solve *rises* with $|k|$ — the inverse of the fluid profile of Figure 6.3 — because the $1/|k|^2$ amplification makes the solve unreliable at low $|k|$ and near-exact at high $|k|$. The learned gate tracks the oracle closely from the mid-band dip near $|k| \approx 8$ through the high-$|k|$ climb, and over-trusts the solve only in the lowest band ($|k| \lesssim 7$, oracle ≈ 0.65, gate ≈ 0.95); that residual low-band gap, in the energy-bearing modes, is why the aggregate gain over the hard path in Table 6.8 is small.

6.4 Reproducibility and Hyperparameters

For completeness we record the salient training configuration of the two experiments (Table 6.9). The gate is a per-mode spectral gate in both models, but its implementation differs: in the fluid model it is a small MLP whose inputs are the squared wavenumber $|k|^2$ and the log-amplitude $\log|\hat{u}|$ of each mode, while in the MIMO model it is a spectral CNN; both end in a sigmoid and have spectrally normalized weight layers.

In both models the gate is *initialized low*, and this is the mechanism behind the safety floor rather than an incidental choice. A low initial gate means the model starts by preferring its own data-driven output — the FNO operator in the fluid case, the PINN in the MIMO case — and only raises the gate, handing trust to the hard physics, on the modes where that physics proves reliable. Where the physics is unreliable the gate simply stays low and defers to the network output, so the gate can only leave the soft baseline unchanged or improve on it; it cannot drive the error up. This is exactly the clamp-inactive behaviour that Corollary 2 formalizes. The two models reach this regime slightly differently: in the fluid model the gate is

Table 6.9: Consolidated training configuration. The fluid and channel surrogates follow their respective source works; ASPINO adds only the gate.

Setting	Fluid (Kolmogorov)	Channel (MIMO)
Soft-path backbone	4-layer Fourier Neural Operator	physics-informed U-Net + transformer
Hard path	Leray projection	rank-3 truncated SVD
Gate	MLP on $(\lvert k\rvert^2, \log\lvert\hat{u}\rvert)$, sigmoid	spectral CNN, sigmoid
Gate initialization	low	low
Warmup	5 epochs, gate frozen at low init	— (soft path frozen)
Gate regularizer μ_β	Beta$(2,2)$, $\mu_\beta > 0$	0 (not needed)
Soft penalty ζ (gated)	1	10^{-5} (negligible)
Optimizer	Adam	Adam
Trained component	gate + FNO backbone	gate (soft path frozen)
Reported metric	rel-L_p error, PDE residual	NMSE (dB)

trained jointly with the FNO backbone after a 5-epoch warmup in which the gate is held frozen at its low initialization so the backbone settles first; in the MIMO model the soft PINN path is frozen outright and only the gate is trained, which keeps any improvement cleanly attributable to the gate.

The distributional gate regularizer is used only in the idealized fluid model ($\mu_\beta > 0$, a Beta$(2,2)$ prior), where it prevents the gate from collapsing to a uniformly-hard profile; it is switched off for MIMO ($\mu_\beta = 0$), where the data themselves keep the gate differentiated. The soft penalty weights reflect how trustworthy each physics signal is. In the fluid model the divergence residual is computed exactly from the simulator, so it is weighted at full strength ($\zeta = 1$); in the MIMO model the power-consistency term is a noisier, ray-tracing-derived estimate, so it is down-weighted to a near-zero $\zeta = 10^{-5}$ — present enough to exert a faint pull, small enough that the data term dominates and the safety floor applies essentially as in the $\zeta \to 0$ analysis of Section 4.5. A practical note: the rank-r SVD hard path adds a 10^{-7} jitter before the decomposition (Lemma 4), essential for stable back-propagation through the singular vectors.

6.5 Discussion

The two experiments tell a consistent story and match the qualitative predictions of Section 5.11.

The gate dominates the pure strategies on the fluid task. On Kolmogorov flow the gate is the most accurate model while remaining near-physical (Table 6.1), and it does so by being hard where the physics is reliable and soft where it is not (Figure 6.4) — behaviour impossible for any uniform-projection method.

The safety floor holds. On MIMO the gated model improves over the soft PINN at every pilot budget and every SNR and never regresses (Figure 6.7, Table 6.6), as Corollary 2 anticipates on the clamp-inactive regime.

The non-monotonic oracle motivates an adaptive gate. The MIMO oracle profile is band-pass, not low-pass: it withdraws trust at the highest frequencies (noise) and again at the lowest (where interference and a dominant LoS bias contaminate the very components the low-rank prior keeps), trusting a mid-band in between (Figure 6.9). No fixed projection can realize this shape; only a per-mode gate can keep the middle and back off at both ends. The oracle is in this sense the cleanest single piece of evidence in the chapter for why adaptive spectral trust is the right object to learn.

One mechanism, three hard operators. The same gate, unchanged in concept, works across two physical domains and three structurally different hard paths: the linear Leray projection on the fluid task, the nonlinear global rank-3 SVD on the urban channel, and a spatially-varying, data-derived RSS spectral mask on the `O1_3p5` channel — and with both a U-Net PINN and a plain FNO surrogate. That the mechanism survives swapping the operator, the surrogate and the dataset together is the strongest evidence we have for the discretization-invariant, plug-and-play behaviour promised by building on neural operators; it also means the gate is selecting *trust in a hard prior* generically, rather than exploiting a property of any one projection.

Chapter 7

Conclusion and Future Work

7.1 Summary

This dissertation introduced the Adaptive Spectral Trust Gate (ASPINO), a mechanism that learns *where* to trust a physics constraint rather than deciding *whether* to enforce it globally. Operating in the Fourier domain on top of any neural-operator surrogate, a small gating network forms a per-mode convex combination of a data-driven soft path and an exact physics hard path, driven by features of the spectral coordinate and amplitude so that it can apply the constraint in well-conditioned spectral regions and defer to the data elsewhere.

We showed that a single gating design serves two very different hard paths: the linear Leray (divergence-free) projection for incompressible flow, and the nonlinear rank-r truncated-SVD projection of the antenna aperture for MIMO channel estimation. On the theory side, an empirical-Rademacher-complexity analysis established an unconditional mixing inequality — the gated class is never more complex than the gate-weighted average of the two paths plus a vanishing gate term — from which we obtained a safety floor (the gate does not increase complexity beyond the soft path on the clamp-inactive regime) and, under one explicitly stated low-rank-transfer assumption, a capacity-reduction bound: the complexity contracts by $1 - \bar{\alpha}(1 - \sqrt{\rho_r})$, where the geometric factor $\sqrt{\rho_r} \approx 0.48$ for the rank-3 aperture coincides almost exactly with the Leray per-mode factor of 0.50. The analysis also explained two empirical observations: that a soft penalty does not reduce hypothesis-class complexity (it acts on the loss), and that an energy-conserving hard path combined with a soft penalty can bias the gate toward being uniformly hard — which is why the gated MIMO model sets the soft-penalty weight to zero. The distributional gate regularizer, in turn, is an optimization device rather than a capacity control: it is needed only in the idealized fluid regime, where the data underdetermine the gate and it would otherwise collapse to a uniformly-hard profile, and is switched off for MIMO, where the data keep the gate differentiated on their own. A zero-shot super-resolution study on the Poisson equation further confirmed the safety floor and the discretization invariance

at resolutions up to 4× the training grid. Empirically, ASPINO attained the best accuracy and a near-physical residual on Kolmogorov-flow denoising, dominating the unconstrained, hard and soft baselines at once; and it improved over a strong physics-informed baseline across all pilot budgets in MIMO channel estimation, recovering a consistent fraction of the oracle gain while never regressing below the baseline.

7.2 Advantages over Existing PINN-based Methods

Relative to existing physics-informed methods, the approach offers two main advantages. First, known constraints and physical laws are applied *selectively*, according to the situation: the method trusts the physics where it is reliable and ignores it where discretization error, numerical error or environmental noise make it unreliable, rather than enforcing it uniformly. Second, by building on neural operators it inherits their discretization invariance and "zero-shot" super-resolution, and it is plug-and-play over the underlying surrogate. We demonstrated effectiveness across two physically distinct domains, suggesting the mechanism is a general tool for physics-constrained operator learning rather than a single-problem trick.

7.3 Limitations

The method assumes that physics reliability has identifiable spectral structure — which holds for the noise/aliasing regimes studied here but may fail when unreliability is not spectrally localized. The hard-path projections we used (Leray and rank-r SVD) cover an important but limited class; nonlinear differential constraints, for which an efficient exact projection is not known, remain open. On the theoretical side, the multiplicative capacity-reduction bound rests on a low-rank-transfer assumption that we justify but do not prove from first principles; making it a theorem (or replacing the geometric factor by the correct data-dependent quantity) is the principal open problem of the analysis. The MIMO gains, while consistent and within the predicted envelope, are modest in absolute terms and are bounded by the oracle ceiling. Finally, the gate adds a constant-factor overhead (one transform pair, one projection, one small network) on top of the surrogate.

7.4 Future Directions

Several directions follow naturally from this work.

Proving the low-rank-transfer assumption. The cleanest next step is to establish, for the network-composed-with-per-sample-projection class, the empirical-

L_2 entropy contraction stated as an assumption in Chapter 5, either in the transductive / shared-low-rank regime or via a data-dependent factor.

Non-uniform geometries. Pairing the gate with graph neural operators and a non-uniform FFT would allow non-periodic fields and arbitrary geometries, retaining the function-space view in which the discretization error can be driven arbitrarily low.

Physics-informed world models. More broadly, physics-informed learning of the kind developed here can serve as a building block for better world models — in particular dynamic world models that evolve in time — where selectively trusting known physics could improve both sample efficiency and long-horizon stability.

In closing, the central message is simple: when combining data-driven models with known physics, the most useful thing to learn is often not the physics itself, nor whether to use it, but *where* to trust it.

Bibliography

[1] M. Raissi, P. Perdikaris, and G. E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

[2] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.

[3] Z. Li, N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, and A. Anandkumar. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020.

[4] N. Kovachki, Z. Li, B. Liu, K. Azizzadenesheli, K. Bhattacharya, A. Stuart, and A. Anandkumar. Neural operator: Learning maps between function spaces with applications to PDEs. *Journal of Machine Learning Research*, 24(89):1–97, 2023.

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

[6] Z. Li, H. Zheng, N. Kovachki, D. Jin, H. Chen, B. Liu, K. Azizzadenesheli, and A. Anandkumar. Physics-informed neural operator for learning partial differential equations. *ACM Journal of Data Science*, 2021.

[7] S. Wang, H. Wang, and P. Perdikaris. Learning the solution operator of parametric partial differential equations with physics-informed DeepONets. *Science Advances*, 7(40):eabi8605, 2021.

[8] V. Duruisseaux, M. Liu-Schiaffini, J. Berner, and A. Anandkumar. Towards enforcing hard physics constraints in operator learning frameworks. *ICML 2024 AI for Science Workshop*, 2024.

[9] C. M. Jiang, K. Kashinath, Prabhat, and P. Marcus. Enforcing physical constraints in CNNs through differentiable PDE layer. *ICLR 2020 Workshop on Integration of Deep Neural Models and Differential Equations*, 2020.

[10] D. Hansen, D. C. Maddix, S. Alizadeh, G. Gupta, and M. W. Mahoney. Learning physical models that can respect conservation laws. *Physica D: Nonlinear Phenomena*, 457:133952, 2024.

[11] P. Harder, A. Hernandez-Garcia, V. Ramesh, Q. Yang, P. Sattigeri, D. Szwarcman, C. Watson, and D. Rolnick. Hard-constrained deep learning for climate downscaling, 2024.

[12] N. Saad, G. Gupta, S. Alizadeh, and D. C. Maddix. Guiding continuous operator learning through physics-based boundary constraints. *International Conference on Learning Representations*, 2023.

[13] G. Negiar, M. W. Mahoney, and A. S. Krishnapriyan. Learning differentiable solvers for systems with hard constraints, 2022.

[14] A. T. Mohan, N. Lubbers, M. Chertkov, and D. Livescu. Embedding hard physical constraints in neural network coarse-graining of three-dimensional turbulence. *Physical Review Fluids*, 8:014604, 2023.

[15] L. Xing, H. Wu, Y. Ma, J. Wang, and M. Long. HelmFluid: Learning Helmholtz dynamics for interpretable fluid prediction. *International Conference on Machine Learning*, 2024.

[16] L. Lu, R. Pestourie, W. Yao, Z. Wang, F. Verdugo, and S. G. Johnson. Physics-informed neural networks with hard constraints for inverse design. *SIAM Journal on Scientific Computing*, 43(6):B1105–B1132, 2021.

[17] A. S. Krishnapriyan, A. Gholami, S. Zhe, R. M. Kirby, and M. W. Mahoney. Characterizing possible failure modes in physics-informed neural networks. *Advances in Neural Information Processing Systems*, 2021.

[18] S. Wang, Y. Teng, and P. Perdikaris. Understanding and mitigating gradient flow pathologies in physics-informed neural networks. *SIAM Journal on Scientific Computing*, 43(5):A3055–A3081, 2021.

[19] S. Wang, X. Yu, and P. Perdikaris. When and why PINNs fail to train: A neural tangent kernel perspective. *Journal of Computational Physics*, 449:110768, 2022.

[20] X. Jin, S. Cai, H. Li, and G. E. Karniadakis. NSFnets (Navier–Stokes flow nets): Physics-informed neural networks for the incompressible Navier–Stokes equations. *Journal of Computational Physics*, 426:109951, 2021.

[21] Z. Li, M. Liu-Schiaffini, N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, and A. Anandkumar. Learning chaotic dynamics in dissipative systems. *Advances in Neural Information Processing Systems*, 35:16768–16781, 2022.

[22] S. A. Javid and N. González-Prelcic. Physics-informed neural networks for wireless channel estimation with limited pilot signals. *NeurIPS 2025 Workshop on AI and ML for Next-Generation Wireless Communications and Networking (AI4NextG)*, 2025.

[23] J. A. Tropp and A. C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, 53(12):4655–4666, 2007.

[24] J. Lee, G.-T. Gil, and Y. H. Lee. Channel estimation via orthogonal matching pursuit for hybrid MIMO systems in millimeter wave communications. *IEEE Transactions on Communications*, 64(6):2370–2386, 2016.

[25] W. Dai and O. Milenkovic. Subspace pursuit for compressive sensing signal reconstruction. *IEEE Transactions on Information Theory*, 55(5):2230–2249, 2009.

[26] M. Sattari, H. Guo, D. Gündüz, A. Panahi, and T. Svensson. Full-duplex millimeter wave MIMO channel estimation: A neural network approach. *IEEE Transactions on Machine Learning in Communications and Networking*, 2024.

[27] X. Zhou, L. Liang, J. Zhang, P. Jiang, Y. Li, and S. Jin. Generative diffusion models for high dimensional channel estimation. *IEEE Transactions on Wireless Communications*, 2025.

[28] P. Bartlett, D. Foster, and M. Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in Neural Information Processing Systems*, 2017.

[29] N. Golowich, A. Rakhlin, and O. Shamir. Size-independent sample complexity of neural networks. *Conference on Learning Theory (COLT)*, 2018.

[30] M. Ledoux and M. Talagrand. *Probability in Banach Spaces*. Springer, 1991.

[31] A. Maurer. A vector-contraction inequality for Rademacher complexities. *Algorithmic Learning Theory (ALT)*, 2016.

[32] E. Candès and Y. Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 2011.

[33] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. *International Conference on Machine Learning (ICML)*, 2017.