UNROLLED OPTIMIZED RESTRICTED ADDITIVE SCHWARZ METHODS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large linear systems obtained from PDE discretizations are typically solved iteratively, and their runtime depends heavily on the quality of the preconditioner. Domain-decomposition preconditioners are attractive because they split the global problem into parallel local solves, but their convergence depends on how information is exchanged across subdomain interfaces. Classical optimized Schwarz methods prescribe these interface parameters analytically, using Fourier minimax arguments that are exact only under idealized assumptions. We introduce UORAS, an unrolled optimized Schwarz preconditioner that keeps the classical analytical parameter rule as a backbone and learns only a multiplicative correction. We also introduce a residual-conditioned non-stationary variant that learns an iteration-dependent schedule through a shared neural network. On the theory side, we show that minimax tuning can be strictly suboptimal for finite-depth distributional losses with positive overlap, and we connect this gap to the stationary UORAS hypothesis class. Empirically, UORAS improves on both classical optimized Schwarz baselines and learning-based interface models.

1 INTRODUCTION

Discretized partial differential equations (PDEs) generate large sparse linear systems in scientific computing, with applications spanning fluid dynamics, electromagnetics, structural mechanics, and geophysical modeling. Direct factorization is rarely viable at scale, so practical solvers iterate on the system

$$\mathbf{A}\mathbf{u} = \mathbf{b}, \tag{1}$$

with $\mathbf{A} \in \mathbb{R}^{N\times N}$, $\mathbf{b} \in \mathbb{R}^N$, $\mathbf{u} \in \mathbb{R}^N$, and a preconditioner $\mathbf{M}^{-1}$ that controls the rate at which the error contracts. The runtime of such a solver is dominated by the choice of $\mathbf{M}^{-1}$, and designing one that is both cheap to apply and effective on the spectrum of $\mathbf{A}$ is the central question of preconditioning.

For PDE-structured systems, the additive Schwarz family is a workhorse parallel preconditioner. Restricted additive Schwarz (RAS, Cai & Sarkis (1999)) and its optimized variant (ORAS, St-Cyr et al. (2007)) decompose the computational domain into overlapping subdomains, solve the smaller subproblems in parallel, and combine the local updates through subdomain restriction operators. The convergence rate of the resulting iteration is governed by the interface conditions exchanged between subdomains; in ORAS, these conditions are encoded by transmission parameters (p, q) that modify the subdomain block matrices.

Classical optimized Schwarz methods (OSM) Gander (2006) prescribe analytical transmission parameters p_{OO2}, q_{OO2} obtained from a Fourier-domain minimax analysis. The derivation rests on three assumptions that rarely hold in practice: (i) the decomposition has exactly two subdomains, (ii) the subdomains extend infinitely along the interface normal, and (iii) the spectral radius is minimized in the worst-case sense across all admissible frequencies. Real decompositions involve many subdomains, finite domains, and problem-dependent spectra, so the analytical OO2 values are routinely applied outside their regime of validity. Lifting any of these assumptions analytically is technically delicate beyond the standard two-subdomain, stationary, semi-infinite setting Gander (2006).

Recent work has explored learning the interface conditions instead. MLORAS Taghibakhshi et al. (2022) bypasses the analytical (p, q) parametrization altogether and uses a graph neural network

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

to predict per-edge interface entries directly from the system matrix. This flexible parametrization can exploit local matrix and graph structure, but it also moves far from the analytical OO2 formula and requires a relatively large architecture with roughly 1.3×10^6 parameters. This motivates a complementary question: whether a much smaller learned correction, applied directly to the analytical OO2 backbone, can recover the part of the transmission rule left unresolved by the classical derivation.

We study this constrained alternative. Our contributions are as follows.

- We put forth UORAS, an unrolled ORAS preconditioner in which a small MLP ($\sim$2300 parameters) predicts a multiplicative correction to the analytical OO2 transmission parameters, together with a non-stationary variant whose correction is conditioned on the residual norm $\|\mathbf{r}^k\|$ at each unrolled iteration.
- We establish a strict-gap result showing that classical minimax tuning is not population-optimal for finite-depth losses once overlap biases the modal decay, with an explicit two-mode counterexample at critical weight $\alpha^\star < 1/2$, and we connect this gap to the stationary UORAS hypothesis class through containment and a universal-approximation argument.
- Empirically, UORAS improves on T2 and OO2 on modified Helmholtz in the regime where the two-subdomain assumption breaks down, generalizes across decompositions, and matches or exceeds MLORAS at roughly $600\times$ fewer parameters.

2 Related Work

Algorithm unrolling. Algorithm unrolling reinterprets the iterations of a classical solver as the layers of a deep network whose parameters are learned by backpropagation through the unrolled iteration Monga et al. (2021); Shlezinger et al. (2025). The technique was introduced for sparse coding (LISTA, Gregor & LeCun (2010)) and has since been applied to linear inverse problems, compressive sensing, and convex optimization solvers Yang et al. (2016); Zhang & Ghanem (2018); Chowdhury et al. (2021). UORAS instantiates this idea for a numerical-PDE preconditioner: each unrolled layer applies a Schwarz-style preconditioned step in which the interface block depends on a learned correction to the analytical OO2 parameters, and the parameters of the correction network are shared across all unrolled layers.

Learning-based domain decomposition. Within the domain-decomposition literature, the closest predecessor is MLORAS Taghibakhshi et al. (2022), which bypasses the analytical (p, q) parametrization of optimized Schwarz and uses a graph neural network to predict the per-edge entries of the interface perturbation matrices. MLORAS is substantially more expressive than the parametrization we study: it conditions on the system matrix, learns the modified subdomain blocks directly, and uses several layers of topology-adaptive graph convolution, totaling roughly 1.3×10^6 parameters. UORAS makes a complementary design choice. Rather than learning interface entries directly, it preserves the analytical OO2 formula and learns only a two-scalar log-correction with $\sim$2300 parameters. This restriction makes the learned component easier to interpret as a correction to a known optimized Schwarz rule, while the empirical section compares whether the reduced parametrization retains the convergence benefits of MLORAS. A related body of work learns adaptive coarse spaces for two-level Schwarz preconditioners via neural-network classifiers Heinlein et al. (2019); a recent review of the DD-ML intersection is Heinlein et al. (2021). Beyond Schwarz-style domain decomposition, the closest learning-based numerical-solver precedent is Luz et al. (2020), which learns the prolongation operator of algebraic multigrid via a graph neural network operating on the system matrix.

Optimized Schwarz theory. The classical theory of (optimized) Schwarz methods is developed in Lions (1989); Efstathiou & Gander (2003); Gander (2006); St-Cyr et al. (2007), with non-overlapping extensions analyzed in Gander & Golub (2002); Gander et al. (2002). These works derive analytical transmission rules in their respective regimes; extending the closed-form analysis beyond the two-subdomain, stationary, semi-infinite setting remains technically delicate Gander (2006), and comprehensive treatments are available in the textbooks of Toselli & Widlund (2005); Dolean et al. (2015). UORAS complements rather than replaces this body of work: the analytical OO2 backbone is preserved, and the learned multiplicative correction adapts to regimes (multi-subdomain, bounded geometry, biased spectra) outside the assumptions under which the analytical theory was derived, with an explicit strict-gap counterexample for positive overlap under generic mode weights (Section 5).

3 BACKGROUND

3.1 PRECONDITIONED ITERATIVE SOLVERS

We consider linear systems equation 1 in which $\mathbf{A}$ is sparse, structured by the underlying PDE discretization, and (in the cases of interest here) symmetric positive definite. Direct factorization is impractical at the scales we target, so we adopt a preconditioned Richardson iteration

$$\mathbf{u}^{k+1} = \mathbf{u}^k + \mathbf{M}^{-1}(\mathbf{b} - \mathbf{A}\mathbf{u}^k), \qquad \mathbf{e}^{k+1} = (\mathbf{I} - \mathbf{M}^{-1}\mathbf{A})\,\mathbf{e}^k, \tag{2}$$

with $\mathbf{e}^k = \mathbf{u}^k - \mathbf{u}$ being the error and $\mathbf{M}^{-1}$ a preconditioner. The convergence rate of equation 2 is governed by the spectral radius of $\mathbf{I} - \mathbf{M}^{-1}\mathbf{A}$, and equation 2 admits a non-stationary extension in which the preconditioner depends on the iteration index k. Designing a preconditioner $\mathbf{M}^{-1}$ (or its index-dependent counterpart) that combines a low per-iteration cost with a small spectral radius is the core design problem of preconditioning; the additive Schwarz family is a standard answer when $\mathbf{A}$ originates from a PDE discretization.

3.2 RESTRICTED AND OPTIMIZED ADDITIVE SCHWARZ

We focus on linear systems equation 1 arising from the discretization of a second-order elliptic PDE $\mathcal{L}u = f$ on a domain $\Omega \subset \mathbb{R}^2$. The additive Schwarz family partitions the index set $\{1, \dots, N\}$ into J subdomains $\mathcal{S}_j$ with possibly overlapping supports. Let $\mathbf{R}_j \in \{0,1\}^{N_j \times N}$ select the entries of an N-vector lying in $\mathcal{S}_j$, and let $\mathbf{A}_j = \mathbf{R}_j \mathbf{A} \mathbf{R}_j^\top$ denote the restriction of $\mathbf{A}$ to subdomain j. Plain additive Schwarz, $\sum_j \mathbf{R}_j^\top \mathbf{A}_j^{-1} \mathbf{R}_j$, double-counts contributions on the overlap and can suffer from poor convergence without restriction or weighting. The restricted additive Schwarz (RAS) preconditioner Cai & Sarkis (1999) addresses this overlap-induced double counting by replacing the overlapping prolongation $\mathbf{R}_j^\top$ with a non-overlapping counterpart $\tilde{\mathbf{R}}_j^\top$, whose columns index a disjoint partition $\tilde{\mathcal{S}}_j \subseteq \mathcal{S}_j$.

The optimized variant additionally modifies the subdomain blocks $\mathbf{A}_j \to \tilde{\mathbf{A}}_j$ to encode richer interface conditions: optimized Schwarz methods (OSM) Gander (2006) form the continuous domain-decomposition framework in which these interface conditions are optimized. In this work, OO2 denotes the corresponding analytical second-order transmission rule, ORAS denotes the discrete restricted additive Schwarz preconditioner that uses such transmission conditions, and UORAS denotes ORAS with learned multiplicative corrections to OO2. Concretely, OSM replaces the Dirichlet boundary exchange of additive Schwarz with Robin-type transmission conditions parametrized by an interface operator $\mathcal{S}_j$, and absorbs $\mathcal{S}_j$ into modified subdomain blocks $\tilde{\mathbf{A}}_j$. For the modified Helmholtz operator $\mathcal{L} = \eta - \Delta$ with shift $\eta > 0$ on two semi-infinite subdomains, the Fourier analysis of Gander (2006) reduces $\mathcal{S}_j$ to a local Robin operator $\pm(p - q\,\partial_{\mathbf{tt}})$ along the interface tangent, parametrized by two scalars $s = (p, q) \in \mathbb{R}^2$ that minimize the spectral radius of the iteration; we write $\tilde{\mathbf{A}}_j(s)$ to make the dependence on these parameters explicit. The resulting discrete preconditioner is the optimized restricted additive Schwarz operator St-Cyr et al. (2007),

$$\mathbf{M}_{\text{ORAS}}^{-1}(s) = \sum_{j=1}^{J} \tilde{\mathbf{R}}_j^\top \tilde{\mathbf{A}}_j(s)^{-1} \mathbf{R}_j. \tag{3}$$

The uniformly optimized choice (OO2) is the analytical pair $s_{\text{OO2}}(h, \eta) = (p_{\text{OO2}}, q_{\text{OO2}})$ given by

$$p_{\text{OO2}} = 2^{-3/5}(\xi_{\min}^2 + \eta)^{2/5} h^{-1/5}, \qquad q_{\text{OO2}} = (2(\xi_{\min}^2 + \eta))^{-1/5} h^{3/5}, \tag{4}$$

with $\xi_{\min}$ the minimum tangential frequency on the grid and h the mesh size; a second-order Taylor approximation (T2), $p = \sqrt{\eta}$ and $q = 1/(2\sqrt{\eta})$, is used as a baseline in Section 6.

Three limitations of OO2. The OO2 values in equation 4 optimize the spectral radius under three structural assumptions: (i) exactly two subdomains, (ii) semi-infinite subdomains, and (iii) worst-case minimax over all admissible frequencies. Real decompositions involve many subdomains, the subdomains are bounded, and the data distribution induces a biased spectrum, so each violation moves the practical problem away from the setting in which the analytical guarantee is exact; extending the closed-form analysis beyond the standard two-subdomain, stationary, semi-infinite setting remains technically delicate Gander (2006) (cf. Section 5). UORAS keeps the analytical rule equation 4 as a backbone and learns a multiplicative correction for the residual mismatch.

4 UNROLLED OPTIMIZED RESTRICTED ADDITIVE SCHWARZ

UORAS unrolls the preconditioned iteration equation 2 for a fixed depth K and replaces the analytical OO2 transmission parameters $(p_{\text{OO2}}, q_{\text{OO2}})$ in equation 4 by a small learned multiplicative correction. The analytical backbone is preserved: the network output rescales OO2 in log-space, so the unrolled solver recovers classical OSM at zero correction (cf. Remark 1 in Section 5). We describe the parametrization, training loss, and non-stationary extension.

Multiplicative-correction parametrization. We parametrize the UORAS transmission tuple as the analytical OO2 pair s_{OO2} in equation 4 multiplied entrywise by an exponentiated log-correction,

$$s(\boldsymbol{\theta}) = \big(p_{\text{OO2}} \exp(p_{\text{c}}(\boldsymbol{\theta})),\, q_{\text{OO2}} \exp(q_{\text{c}}(\boldsymbol{\theta}))\big), \qquad (p_{\text{c}}(\boldsymbol{\theta}),\, q_{\text{c}}(\boldsymbol{\theta})) \in \mathbb{R}^2, \tag{5}$$

where the log-correction $(p_{\text{c}}(\boldsymbol{\theta}), q_{\text{c}}(\boldsymbol{\theta}))$ is produced by a multi-layer perceptron (MLP) $\Psi(\cdot;\boldsymbol{\theta})$,

$$\big(p_{\text{c}}(\boldsymbol{\theta}),\, q_{\text{c}}(\boldsymbol{\theta})\big) = \Psi(h, \eta; \boldsymbol{\theta}). \tag{6}$$

Exponentiating the learned correction keeps the actual transmission parameters positive, as required by OSM (Section 3.2), and $\Psi \equiv 0$ recovers the OO2 parameters exactly. The unrolled UORAS preconditioner is then ORAS evaluated at the learned tuple,

$$\mathbf{M}^{-1}_{\text{UORAS}}(\boldsymbol{\theta}) = \mathbf{M}^{-1}_{\text{ORAS}}\big(s(\boldsymbol{\theta})\big) = \sum_{j=1}^{J} \tilde{\mathbf{R}}_j^{\top} \big(\tilde{\mathbf{A}}_j(s(\boldsymbol{\theta}))\big)^{-1} \mathbf{R}_j, \tag{7}$$

and the unrolled iteration substitutes $\mathbf{M}^{-1} \leftarrow \mathbf{M}^{-1}_{\text{UORAS}}(\boldsymbol{\theta})$ in equation 2. The subdomain decomposition, overlap, restriction operators, and local discretized operators are fixed by the problem instance; only the two transmission scalars (p, q) are learned, through the descriptor-dependent log-corrections above.

The MLP is small by design. Empirically (Section 6), a shallow network with $\sim$2300 parameters suffices, in contrast to the $\sim 1.3 \times 10^6$ parameters of MLORAS. Specifically, the stationary model consists of 3 linear layers, taking a 2-dimensional input (h, η) and producing the 2-dimensional log-correction output $(p_{\text{c}}, q_{\text{c}})$; the dimensions of the two hidden layers are 64 and 32. The non-stationary variant uses the same architecture but extends the input dimension to 3 to admit the running residual norm $\|\mathbf{r}^k\|$. As a result, in all our experiments, the total number of parameters is 2240 and 2304 for stationary and non-stationary architectures, respectively. Each layer except the last is followed by a sigmoid nonlinearity. All layers use standard random initialization except the last, whose weights and biases are initialized at 0. Consequently, the multiplicative correction factor, the exponential of the neural network's output, is 1, which promotes initialization informed by analytical OO2 parameters. Conditioning the correction on (h, η) alone rather than on the full system matrix is what makes this size feasible: the OO2 backbone already absorbs the instance descriptors for which an analytical answer exists, and the network only has to model the residual gap left by that backbone.

Unrolled training loss. Given a depth $K \geq 1$ and a distribution ν over problem instances $z = (h, \eta, J, \dots)$, we train $\boldsymbol{\theta}$ by minimizing the expected squared K-step error norm

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{z \sim \nu} \left[\|\mathbf{e}^K(z; \boldsymbol{\theta})\|_2^2 \right], \tag{8}$$

where $\mathbf{e}^K(z;\boldsymbol{\theta}) = \mathbf{u}^K(z;\boldsymbol{\theta}) - \mathbf{u}(z)$ is the error after K unrolled applications of equation 2 with $\mathbf{M}^{-1}_{\text{UORAS}}(\boldsymbol{\theta})$ from equation 7, starting from a random initial guess $\mathbf{u}^0$. The loss is differentiable in $\boldsymbol{\theta}$ through the K-fold composition of the error iteration matrix $\mathbf{E} = \mathbf{I} - \mathbf{M}^{-1}_{\text{UORAS}}(\boldsymbol{\theta})\,\mathbf{A}$, since $\tilde{\mathbf{A}}_j$ depends analytically on $s(\boldsymbol{\theta})$. The training is unsupervised: no reference solver trajectory is required, only the discretized operator and the random initial guess.

Non-stationary extension. The same parametrization extends to a non-stationary iteration (cf. Section 3.1): the correction at iteration k is now informed by the running residual norm,

$$\big(p_{\text{c}}^k(\boldsymbol{\theta}),\, q_{\text{c}}^k(\boldsymbol{\theta})\big) = \Psi\big(h, \eta, \|\mathbf{r}^k\|;\, \boldsymbol{\theta}\big), \tag{9}$$

giving the layerwise transmission tuple $s^k(\boldsymbol{\theta}) = \big(p_{\text{OO2}} \exp(p_{\text{c}}^k(\boldsymbol{\theta})),\, q_{\text{OO2}} \exp(q_{\text{c}}^k(\boldsymbol{\theta}))\big)$. The layer-$k$ preconditioner is then ORAS evaluated at $s^k(\boldsymbol{\theta})$,

$$\mathbf{M}^{-1}_{\text{ORAS}}\big(s^k(\boldsymbol{\theta})\big) = \sum_{j=1}^{J} \tilde{\mathbf{R}}_j^{\top} \big(\tilde{\mathbf{A}}_j(s^k(\boldsymbol{\theta}))\big)^{-1} \mathbf{R}_j. \tag{10}$$

The iteration index enters the correction only through $\|\mathbf{r}^k\|$, so the parameters $\boldsymbol{\theta}$ remain shared across the unrolled depth. This gives the solver a Chebyshev-style schedule: just as fixed-coefficient Richardson is accelerated by a non-stationary ω_k with optimal Chebyshev coefficients, the residual norm provides UORAS with an iteration-dependent state variable from which it can learn a schedule correlated with the stage of convergence. Deriving an analytical layerwise schedule is technically delicate beyond the standard stationary setting Gander (2006), which is exactly the regime where a learned schedule pays off. As $\|\mathbf{r}^k\| \to 0$, the input $(h, \eta, 0)$ to Ψ stabilizes and the correction converges to a stationary value.

5 THEORY

We analyze why a learned correction to the analytical OO2 rule can improve finite-depth performance, even when the OO2 rule remains the backbone of the parametrization. Let $z \in \mathcal{Z}$ denote a problem instance, including the mesh size h, shift η, subdomain count J, overlap L, geometry, and initial error. Let a transmission law ϕ map z to the layerwise transmission parameters used across the K unrolled steps, e.g., $((p_1, q_1), \ldots, (p_K, q_K))$ in the second-order ORAS case and $(\sigma_1, \ldots, \sigma_K)$ in the zeroth-order model below. For such a law, let $\ell_K(z; \phi) = \|\mathrm{e}^K(z; \phi)\|^2$ be the K-step error loss and let

$$\mathcal{R}_\nu(\phi) = \mathbb{E}_{z\sim\nu}\big[\ell_K(z; \phi)\big] \tag{11}$$

be the population risk over a problem distribution ν. We write ϕ_{OSM} for the OO2 law applied identically at every unrolled layer, and $\phi_{\mathrm{UORAS}}(\boldsymbol{\theta})$ for the law produced by the UORAS parametrization equation 5. The theory has the following structure. First, UORAS contains classical OSM exactly (Remark 1). Second, minimax tuning can be strictly suboptimal for this finite-depth population risk, already in a two-mode overlapping Schwarz model (Proposition 1). Third, this oracle-level gap is realized by the stationary UORAS network class under standard regularity conditions (Proposition 2). The full formal setup, auxiliary distributional statements (cf. Proposition 3), and proofs are deferred to Appendix A.

Remark 1 (Containment)**.** *The UORAS parametrization equation 5 contains ϕ_{OSM} as the special case $p_c \equiv q_c \equiv 0$. If the network class admits $\boldsymbol{\theta}_0$ with $\Psi(\cdot; \boldsymbol{\theta}_0) = 0$, so that $p_c(\boldsymbol{\theta}_0) = q_c(\boldsymbol{\theta}_0) = 0$, then $\phi_{UORAS}(\boldsymbol{\theta}_0) = \phi_{OSM}$ and $\mathcal{R}_\nu(\phi_{UORAS}(\boldsymbol{\theta}_0)) = \mathcal{R}_\nu(\phi_{OSM})$. In particular, the hypothesis class cannot have worse optimal population risk than classical OSM, although a particular trained model may still perform worse if optimization fails or the training sample is unrepresentative.*

A two-mode strict-gap counterexample. The construction is on the continuous two-subdomain modified Helmholtz model $(\eta - \Delta)u = f$ on $\mathbb{R}^2$, decomposed into $\Omega_1 = (-\infty, \beta) \times \mathbb{R}$ and $\Omega_2 = (\alpha, +\infty) \times \mathbb{R}$ with overlap $L = \beta - \alpha > 0$. Following Gander (2006), applying a tangential Fourier transform with frequency ξ yields the per-mode convergence factor

$$\rho(\xi; \sigma) = \left|\frac{\sigma - \lambda(\xi)}{\sigma + \lambda(\xi)}\right| \mathrm{e}^{-\lambda(\xi)L}, \qquad \lambda(\xi) = \sqrt{\eta + \xi^2}, \tag{12}$$

where $\sigma > 0$ is the (zeroth-order) transmission parameter. Classical OSM derives σ by minimizing the worst-case rational factor across all admissible frequencies; for a two-mode model with extreme frequencies ξ_1, ξ_2 and $\lambda_i = \lambda(\xi_i)$, this yields $\sigma_{\mathrm{mm}} = \sqrt{\lambda_1 \lambda_2}$. The next proposition is intentionally minimal: it shows that even in a two-mode, two-subdomain, zeroth-order setting, σ_{mm} is not the population-optimal choice for generic finite-depth losses.

Proposition 1 (Strict distributional gap)**.** *Fix $K \geq 1$, $L > 0$, and two frequencies ξ_1, ξ_2 with $\lambda_1 < \lambda_2$. Consider initial errors with spectral energy supported on these two modes with relative weight $\alpha \in (0, 1)$ on the low-frequency mode. Define the K-step weighted loss*

$$J_\alpha(\sigma) = \alpha\, \rho_1(\sigma)^{2K} + (1 - \alpha)\, \rho_2(\sigma)^{2K}, \qquad \rho_i(\sigma) = \frac{|\sigma - \lambda_i|}{\sigma + \lambda_i}\, \mathrm{e}^{-\lambda_i L}, \tag{13}$$

and the critical weight

$$\alpha^\star = \frac{\mathrm{e}^{-2K\lambda_2 L}}{\mathrm{e}^{-2K\lambda_1 L} + \mathrm{e}^{-2K\lambda_2 L}} \in \big(0, \tfrac{1}{2}\big). \tag{14}$$

Then for every $\alpha \neq \alpha^\star$, the minimax-optimal $\sigma_{\mathrm{mm}} = \sqrt{\lambda_1 \lambda_2}$ is strictly suboptimal:

$$\min_{\sigma > 0} J_\alpha(\sigma) < J_\alpha(\sigma_{\mathrm{mm}}). \tag{15}$$

The proof is in Appendix C. The critical weight $\alpha^\star$ falls below $1/2$ because the overlap exponential $\mathrm{e}^{-\lambda_i L}$ already damps the high-frequency mode, biasing the loss toward the slowly decaying low-frequency mode. Even uniform mode weighting ($\alpha = 1/2$) therefore makes minimax tuning suboptimal as soon as $L > 0$, and the range of mode weights for which the minimax rule is suboptimal widens with the overlap, the spectral ratio $\kappa = \lambda_2/\lambda_1$, and the unrolling depth K. The construction should therefore be read as a counterexample to the distributional optimality of minimax tuning, not as a complete theory for the multi-subdomain, second-order UORAS setting. The second-order case and residual-conditioned schedules are discussed in Appendix G.

From oracle corrections to UORAS. Proposition 1 constructs an oracle-level improvement. The next statement connects such a stationary descriptor-based oracle to the implemented UORAS parametrization: a single correction is produced from $d(z) = (h, \eta)$, exponentiated, and reused at every unrolled layer.

Proposition 2 (Stationary approximation-to-oracle)**.** *Let $\mathcal{D}_{\mathrm{stat}}$ denote the descriptor space for $d(z) = (h, \eta)$, and let Φ_{stat} be the class of stationary descriptor laws*

$$\phi_s(d) = (s(d), \dots, s(d)), \qquad s\colon \mathcal{D}_{\mathrm{stat}} \to (0, \infty)^2.$$

Let $\mathcal{R}^\star_{\nu,\mathrm{stat}} = \inf_{\phi\in\Phi_{\mathrm{stat}}} \mathcal{R}_\nu(\phi)$. Assume: (i) *the stationary strict gap $\Delta_{\mathrm{stat}} = \mathcal{R}_\nu(\phi_{OSM}) - \mathcal{R}^\star_{\nu,\mathrm{stat}} > 0$ holds;* (ii) *an oracle stationary transmission rule $s^\star_{\mathrm{stat}}\colon \mathcal{D}_{\mathrm{stat}} \to (0,\infty)^2$ attaining $\mathcal{R}^\star_{\nu,\mathrm{stat}}$ is Lipschitz continuous on a compact descriptor set $\mathcal{D}_{\mathrm{stat},0} \subseteq \mathcal{D}_{\mathrm{stat}}$ with induced descriptor measure one, and its image lies in a compact subset of $(0,\infty)^2$;* (iii) *the loss is uniformly continuous in the stationary transmission parameters on this compact range and is dominated by an integrable envelope; and* (iv) *the stationary UORAS log-correction class is a universal approximator on Lipschitz maps $\mathcal{D}_{\mathrm{stat},0} \to \mathbb{R}^2$. Then there exists $\boldsymbol{\theta}^\star \in \Theta$ with*

$$\mathcal{R}_\nu\big(\phi_{UORAS}(\boldsymbol{\theta}^\star)\big) < \mathcal{R}_\nu(\phi_{OSM}). \tag{16}$$

The proof is in Appendix E. Hypothesis (iv) is satisfied by sufficiently wide feedforward networks with continuous activation functions Hornik (1991); Cybenko (1989). The bounded-subdomain calculation in Appendix F identifies a second source of mismatch: the OO2 coefficients are derived in a semi-infinite two-subdomain geometry, whereas the discrete problems solved in Section 6 use bounded subdomains. Together, these facts justify learning a small correction while retaining the analytical OO2 rule as the default solution contained inside the hypothesis class.

6 Numerical Results

We consider a modified Helmholtz equation on a unit square with Dirichlet boundary conditions.[1] We apply finite difference discretization with the standard five-point stencil for the Laplacian and the uniform mesh of size $h = \frac{1}{\sqrt{N}+1}$. The condition number of each discretized grid is inversely proportional to h^2. For each problem, we average over 5 homogeneous error equations, $\mathbf{f} = \mathbf{0}$, with random initial guesses and report iterations to reach error 10^{-6}.

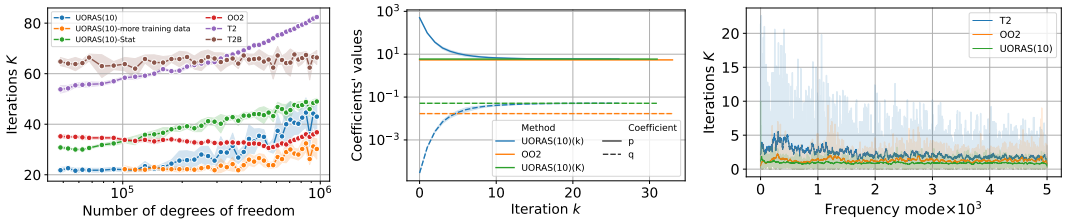
As the MLORAS model poorly scales in the number of degrees of freedom N, in the first experiment, we train both models on small problems given by the modified Helmholtz equations on a unit square of the size $\{36, ...144\}$ DoFs. The domains are decomposed into $J = 2$ overlapping subdomains with the overlap size h. We test performance against the modified Helmholtz equation on unit-square grids of larger sizes, $\{169, ..., 1936\}$. The shift parameter $\eta = 1$ throughout all experiments.

We train the models for 500 epochs with early stopping. We use Adam optimizer with standard $\beta_1 = 0.9$, $\beta_2 = 0.999$, and the learning rate being 10^{-4} for MLORAS and 10^{-3} for our non-stationary unrolled model. A mini-batch consists of a single training grid and the associated 5 error equations. The input to our model is a vector $[\eta, h, \|\mathbf{r}^k\|]$, where the residual $\mathbf{r}^k = \mathbf{b} - \mathbf{A}\mathbf{u}^k$. The loss is given by the error norm $\|\mathbf{e}^K\|$. The number of iterations to reach the tolerance threshold (mean and standard deviation) averaged over 5 instances are reported in Table 1 for analytical OO2 parameters and deep learning-based models. Both MLORAS and our UORAS(2), which stands for unrolled non-stationary model trained on $J = 2$ subdomains, were trained on exactly the same data split.

[1]The code is available at https://github.com/uoras-anonymous/uoras-anonymous

Table 1: Number of iterations until prescribed error tolerance. Comparison against MLORAS on smaller grids.

N	144	400	784	1296	1936
OO2	6.0 ± 0.00	6.0 ± 0.00	6.0 ± 0.00	6.6 ± 0.49	7.0 ± 0.00
MLORAS	5.0 ± 0.00	7.0 ± 0.00	9.0 ± 0.64	11.8 ± 0.40	15.2 ± 1.33
UORAS(2)	5.0 ± 0.00	5.0 ± 0.00	6.0 ± 0.00	6.0 ± 0.00	6.6 ± 0.49



(a) (b) (c)

Figure 1: Performance of UORAS analyzed from different perspectives: (a) the number of iterations required to reach the error tolerance of 10^{-6} for ORAS with analytical T2, T2B, and OO2 transmission parameters, and for unrolled ORAS trained on 10 subdomains with stationary, UORAS(10)-Stat, and non-stationary, UORAS(10), correction coefficients; (b) the evolution of the non-stationary transmission parameters of UORAS(10) on $N = 102400$ and $J = 10$. Vertical lines are the values of transmission parameters for OO2 and the UORAS(10) obtained at the last iteration K; (c) the number of iterations required to eliminate 5000 low-frequency modes in the error equations from the grids with $N = 102400$ and $J = 10$.

For a grid from the training set, $N = 144$, OO2 converges in 6 iterations, while MLORAS and our unrolled non-stationary model require 1 iteration less. As the grid size increases, the convergence of MLORAS deteriorates significantly, while OO2 and our model remain competitive throughout. This can be attributed to the fact that the number of predicted parameters of MLORAS scales directly with the problem size. By contrast, UORAS(2) only predicts two scalars, which are also rescaled by analytical OO2 parameters. In addition to that, the number of trainable parameters of MLORAS is around $|\theta| = 1.3 \cdot 10^6$, while our model has only 2304.

Next, we evaluate the performance on larger grids. We train our model on the grids of sizes spanning 400 to 44100 degrees of freedom and decomposed into 10 subdomains. The batch size, h, η, and the loss function remain the same as in the previous experiment. We train the model until convergence and manually interrupt the training when the validation performance metric, which is the number of iterations until convergence, stops improving for a large number of iterations. We perform validation every 10 epochs. In this setting, the training for 2300 epochs was sufficient with the best model achieved at the epoch 1340.

During training[2], we unroll the stationary model for 10 iterations. We find that the non-stationary model benefits from the varying number of unrolled iterations during training. Because gradients in unrolled architectures tend to vanish, a fixed-length non-stationary model typically concentrates its greatest error reduction at the final iteration, K. By sampling K uniformly at random, we train the model to minimize the error at every step. We sample K from the interval $[3, 15]$ in our experiments.

For testing, we decompose the domains of sizes 48400 to 960400 DoFs into $J = \{2, 10, 20\}$ overlapping stripes (subdomains) with the overlap size h. The hyperparameters of Adam are $\beta_1 = 0.9$, $\beta_2 = 0.999$ and the learning rate $5 \cdot 10^{-3}$. As before, we measure the number of iterations required to achieve the error of 10^{-6}. The error bands indicate the standard deviation over 5 error equations. For 10 subdomains, we plot this metric as a function of N in Fig. 1a.

We use four classical baselines: RAS, OO2 and T2, and T2B whose parameters are obtained analytically for each grid. We then compare these against the stationary and non-stationary versions

[2]Ran on an RTX 2080 Ti GPU. The training took around 24 hours.

of UORAS, trained on the same data. The range of grid sizes in the training set determines the condition numbers a model is exposed to, and therefore how far the unrolled models can be expected to extrapolate. To measure this effect, we further conduct an experiment in which the training set of UORAS contains 10 additional grids ranging from 48400 to 96100 DoFs.

Among the baselines, OO2 achieves the lowest number of iterations until convergence across the whole range of grid sizes. The unrolled stationary model generalizes to the grids beyond its training range, remaining competitive with OO2 up to 10^5 DoFs and requiring on average around 6% fewer iterations for grids of small size. UORAS extends this range considerably, outperforming OO2 across unseen problems spanning an order of magnitude in grid size and requiring around 40% fewer iterations for small and medium-sized problems. Only for the larger problems does the performance become comparable, with OO2 taking over for very large grids.

The effect of the larger grids in the training set is shown as the orange trace: UORAS remains strictly better than OO2 for grids up to 10^6 DoFs and exhibits a smaller variance across error equations. This confirms that UORAS' advantage over OO2 is a function of the grid condition numbers seen during training. Furthermore, it indicates that its range of applicability moves towards larger grids as the range of condition numbers seen during training grows.

In addition, we compare our approach against the T2 and T2B baselines, where T2B uses the bounded-subdomain transmission coefficients from Proposition 4 rather than the semi-infinite approximation. As the grid size grows, the approximated T2 coefficients become inaccurate, due to the underlying assumption on semi-infinite subdomains. Meanwhile, T2B, which was derived for 2 bounded subdomains, converges faster on larger grids. As for T2 ORAS, the p and q values for T2B are derived from a Taylor expansion of the exact optimal symbol presented in Proposition 4 around $\xi = 0$, yielding

$$p = \frac{\sqrt{\eta}}{\tanh(M\sqrt{\eta})}, \quad \text{and} \quad q = \frac{1}{2}\left(\frac{1}{\sqrt{\eta}\tanh(M\sqrt{\eta})} - \frac{M}{\sinh^2(M\sqrt{\eta})}\right), \tag{17}$$

where M denotes the interior width of the respective bounded subdomain.

In Fig. 1b, we analyze the behavior of non-stationary coefficients on a single problem from the validation set with $N = 102400$. During the initial iterations, the values of p^k and q^k deviate significantly from the stationary constants predicted by the OO2 baseline. Specifically, larger values of p^k and smaller values of q^k correspond to a more accurate approximation of the low-frequency components of the error, which is the most challenging to eliminate for smoother methods, such as additive Schwarz. As the iterations progress, p^k converges very closely to the values of p predicted by OO2, while q^k asymptotes to a slightly larger magnitude. Interestingly, if we reuse the last coefficients' values predicted by our unrolled models, we obtain faster convergence than that obtained by the OO2 method.

To evaluate the convergence properties of the proposed methods in the frequency domain, we project the iterative error onto the orthogonal eigenbasis of the SPD system matrix $\mathbf{A} \in \mathbb{R}^{102400\times 102400}$. This projection ensures a consistent evaluation of the spatial frequencies of all analyzed preconditioners. As the computation of all eigenvectors is computationally intractable due to the problem size, we limit the eigenbasis to the first 5000 eigenvectors corresponding to the low-frequency modes. We focus on the low-frequency components, as these are most difficult to eliminate compared to their high-frequency counterparts. We track the number of iterations required to reach the tolerance threshold of 10^{-6} for each frequency mode. We average the results across five error equations and apply a moving average filter with a window length of 40 to visualize the global trend. The baseline T2 method requires most iterations to eliminate smooth, low-frequency error modes, which is a well-known bottleneck in the classical iterative solvers. The analytically optimized OO2 method can eliminate the low-frequency modes much more efficiently, but introduces new vulnerabilities, requiring high iteration counts to clear the modes of higher frequencies.

In contrast, the MLP-driven approach has a much flatter spectrum and requires fewer iterations to eliminate the error modes across the majority of the spectrum. Its advantage is most pronounced in the low-frequency modes. By mapping the current residual to non-stationary transmission parameters, the MLP estimates and eliminates the frequency modes that dominate the current iteration, resulting in more robust and efficient convergence.

Lastly, we report the test set results as a mean and standard deviation in Table 2, averaged over the mesh sizes h and five error equations. For two subdomains, OO2 outperforms our unrolled method that was trained on 10 subdomains. This behavior is expected, as the OO2 coefficients are explicitly derived under a two-subdomain assumption. At the same time, the deviation of our model from this theoretical baseline remains small. For 10 subdomains, our method requires fewer iterations to converge than OO2, and at $J = 20$ the gap widens further. This shows that UORAS, despite being trained on $J = 10$, generalizes to configurations with more subdomains. As J increases and the assumptions underlying the OO2 derivation hold less frequently, our model provides a more accurate approximation of the coefficients.

Table 2: Number of iterations until error tolerance of 10^{-6}, averaged over mesh sizes h and five error equations

J_{test}	T2	OO2	UORAS(10)	UORAS(10)-Stat
2	59.79 ± 9.31	12.56 ± 1.18	13.69 ± 1.62	16.93 ± 0.93
10	69.13 ± 7.82	33.29 ± 1.55	31.38 ± 8.26	41.83 ± 5.13
20	111.19 ± 4.61	70.82 ± 2.18	59.20 ± 10.67	75.88 ± 9.16

We also verify how the analyzed models behave when applied as preconditioners to standard GMRES Saad & Schultz (1986) without restarts, Table 3. Although flexible GMRES (fGMRES) could be combined with UORAS(10), notice that the latter is trained to minimize the residual at each iteration. Therefore, inserting the fGMRES into this sequence will interrupt the sequence of operations that UORAS(10) was trained on, which might potentially degrade the performance. For the stationary unrolled algorithm and the classical ORAS methods we see a similar trend as in the case of stationary solvers: for 2 subdomains, the OO2 preconditioning achieves the lowest number of iterations. However, as J increases, the OO2 loses its optimality and significantly benefits from the correction predicted by our UORAS.

Table 3: Number of GMRES-preconditioned iterations until error tolerance of 10^{-6}, averaged over mesh sizes h and five error equations

J_{test}	T2	OO2	UORAS(10)-Stat
2	18.71 ± 1.35	8.77 ± 0.65	11.22 ± 0.46
10	30.44 ± 1.43	25.58 ± 0.60	23.83 ± 0.47
20	55.13 ± 1.05	51.57 ± 1.17	48.03 ± 0.94

7 Conclusions and limitations

We introduced UORAS, a constrained learned extension of optimized restricted additive Schwarz method that keeps the analytical OO2 transmission rule as a backbone and learns only two descriptor-dependent log-corrections to its parameters, with a residual-conditioned variant for non-stationary transmission schedules. The theory shows that minimax-tuned interface parameters can be suboptimal for finite-depth distributional losses with positive overlap, and that stationary and residual-conditioned UORAS can represent useful corrections to the classical rule. The numerical results support the main design choice: UORAS improves on classical optimized Schwarz baselines and on learning-based preconditioners with far fewer parameters and better scalability.

The main limitations are scope and theory. The experiments focus on modified Helmholtz-type problems with regular decompositions and controlled variation in mesh size and problem parameters; nonsymmetric, indefinite, strongly heterogeneous, nonlinear, or highly irregular geometric problems may require richer descriptors or retraining. The theory isolates why learned corrections can help, but it does not provide a complete convergence guarantee for the full multi-subdomain, second-order, residual-conditioned method after gradient-based training. Because UORAS targets numerical PDE solvers and does not release human data or user-facing predictive models, we do not anticipate direct societal harms beyond the usual computational and energy costs of training and evaluation.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools to improve the readability and comprehensiveness of the manuscript. We also used AI to facilitate understanding of baseline methods. Lastly, AI tools were used to implement a part of the codebase. We have not used generative AI tools for drafting sections of the paper, generating synthetic datasets, and proving mathematical claims. We have reviewed all AI-assisted work: the AI-generated text was verified by the authors. The AI-generated code was verified by an author. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Xiao-Chuan Cai and Marcus Sarkis. A restricted additive Schwarz preconditioner for general sparse linear systems. *SIAM J. Sci. Comput.*, 21:792–797, 1999. doi: 10.1137/S106482759732678X.

Arindam Chowdhury, Gunjan Verma, Chirag Rao, Ananthram Swami, and Santiago Segarra. Unfolding wmmse using graph neural networks for efficient power allocation. *IEEE Transactions on Wireless Communications*, 20(9):6004–6017, 2021.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2(4):303–314, 1989. doi: 10.1007/BF02551274.

Victorita Dolean, Pierre Jolivet, and Frédéric Nataf. *An Introduction to Domain Decomposition Methods: Algorithms, Theory, and Parallel Implementation*. SIAM, Philadelphia, PA, 2015. ISBN 978-1-611974-05-8. doi: 10.1137/1.9781611974065.

Evridiki Efstathiou and Martin J. Gander. Why restricted additive Schwarz converges faster than additive Schwarz. *BIT Numerical Mathematics*, 43(5):945–959, Dec 2003. ISSN 1572-9125. doi: 10.1023/B:BITN.0000014563.33622.1d.

Martin J. Gander. Optimized Schwarz methods. *SIAM Journal on Numerical Analysis*, 44(2):699–731, 2006. doi: 10.1137/S0036142903425409.

Martin J Gander and Gene H Golub. A non-overlapping optimized Schwarz method which converges with arbitrarily weak dependence on h. In *International Conference on Domain Decomposition Methods*, 2002.

Martin J. Gander, Frédéric Magoulès, and Frédéric Nataf. Optimized Schwarz methods without overlap for the Helmholtz equation. *SIAM J. Sci. Comput.*, 24(1):38–60, 2002. doi: 10.1137/S1064827501387012.

Karol Gregor and Yann LeCun. Learning fast approximations of sparse coding. In *Int. Conf. Mach. Learn. (ICML)*, pp. 399–406, 2010.

Alexander Heinlein, Axel Klawonn, Martin Lanser, and Janine Weber. Machine learning in adaptive domain decomposition methods — predicting the geometric location of constraints. *SIAM J. Sci. Comput.*, 41(6):A3887–A3912, 2019. doi: 10.1137/18M1205364.

Alexander Heinlein, Axel Klawonn, Martin Lanser, and Janine Weber. Combining machine learning and domain decomposition methods for the solution of partial differential equations — a review. *GAMM-Mitteilungen*, 44(1):e202100001, 2021. doi: 10.1002/gamm.202100001.

Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2): 251–257, 1991. doi: 10.1016/0893-6080(91)90009-T.

P.-L. Lions. On the Schwarz alternating method II: Stochastic interpretation and order properties. In *Domain Decomposition Methods*, pp. 47–70. SIAM, Philadelphia, 1989.

Ilay Luz, Meirav Galun, Haggai Maron, Ronen Basri, and Irad Yavneh. Learning algebraic multigrid using graph neural networks. In *Int. Conf. Mach. Learn. (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6489–6499. PMLR, 13–18 Jul 2020.

Vishal Monga, Yuelong Li, and Yonina C. Eldar. Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. *IEEE Signal Process. Mag.*, 38(2):18–44, 2021. doi: 10.1109/MSP.2020.3016905.

Youcef Saad and Martin H. Schultz. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM Journal on Scientific and Statistical Computing*, 7(3):856–869, 1986. doi: 10.1137/0907058. URL https://doi.org/10.1137/0907058.

Nir Shlezinger, Santiago Segarra, Yi Zhang, Dvir Avrahami, Zohar Davidov, Tirza Routtenberg, and Yonina C Eldar. Deep unfolding: Recent developments, theory, and design guidelines. *arXiv preprint arXiv:2512.03768*, 2025.

Amik St-Cyr, Martin J. Gander, and Stephen J. Thomas. Optimized restricted additive Schwarz methods. In *Domain Decomposition Methods in Science and Engineering XVI*, pp. 213–220. Springer Berlin Heidelberg, 2007. ISBN 978-3-540-34469-8.

Ali Taghibakhshi, Nicolas Nytko, Tareq Uz Zaman, Scott MacLachlan, Luke Olson, and Matthew West. Learning interface conditions in domain decomposition solvers. In *Int. Conf. Neural Inf. Process. Syst. (NeurIPS)*, volume 35, pp. 7222–7235. Curran Associates, Inc., 2022.

Andrea Toselli and Olof B. Widlund. *Domain Decomposition Methods — Algorithms and Theory*, volume 34 of *Springer Series in Computational Mathematics*. Springer-Verlag Berlin Heidelberg, 2005. ISBN 978-3-540-26662-3. doi: 10.1007/b137868.

Yan Yang, Jian Sun, Huibin Li, and Zongben Xu. Deep ADMM-Net for compressive sensing MRI. In *Int. Conf. Neural Inf. Process. Syst. (NeurIPS)*, volume 29, 2016.

Jian Zhang and Bernard Ghanem. ISTA-Net: Interpretable optimization-inspired deep network for image compressive sensing. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 1828–1837, 2018.

A Appendix

This appendix collects the proofs of Propositions 1 and 2, an auxiliary positive-measure improvement lemma, the bounded-subdomain calculation, and longer remarks on residual-conditioned schedules, optimization landscape, and generalization. The convergence-factor setup needed for the proof of Proposition 1 is recorded first.

B Convergence-factor setup for the two-subdomain Helmholtz model

We expand on the Fourier-domain setup outlined in Section 5. On a finite mesh with mesh size h, the tangential frequency takes values in a bounded discrete set; let $\xi_{\min}$ and $\xi_{\max}$ denote the extreme frequencies, with characteristic exponents $\lambda_{\min} = \lambda(\xi_{\min})$ and $\lambda_{\max} = \lambda(\xi_{\max})$, $0 < \lambda_{\min} < \lambda_{\max} < \infty$.

Minimax-optimal transmission parameter. The classical OSM derives the transmission parameter by minimizing the worst-case rational convergence factor over all frequencies:

$$\sigma_{\mathrm{mm}} \in \arg\min_{\sigma>0} \max_{\xi\in[\xi_{\min},\xi_{\max}]} \frac{|\sigma-\lambda(\xi)|}{\sigma+\lambda(\xi)}. \tag{18}$$

Since the rational factor is decreasing in σ for $\sigma < \lambda$ and increasing for $\sigma > \lambda$, the minimax solution equioscillates between the two extreme modes, and a direct calculation yields

$$\sigma_{\mathrm{mm}} = \sqrt{\lambda_{\min}\lambda_{\max}}, \tag{19}$$

with rational equioscillation value

$$r_{\mathrm{mm}} = \frac{|\sigma_{\mathrm{mm}}-\lambda_{\min}|}{\sigma_{\mathrm{mm}}+\lambda_{\min}} = \frac{|\sigma_{\mathrm{mm}}-\lambda_{\max}|}{\sigma_{\mathrm{mm}}+\lambda_{\max}} = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}, \tag{20}$$

where $\kappa = \lambda_{\max}/\lambda_{\min} > 1$. Note that this equioscillation holds only for the rational part of the convergence factor; the full factor $\rho(\xi;\sigma) = r(\xi;\sigma)\,\mathrm{e}^{-\lambda(\xi)L}$ does not equioscillate because $\mathrm{e}^{-\lambda_{\min}L} > \mathrm{e}^{-\lambda_{\max}L}$. This asymmetry is the key driver of the strict-gap result.

Remark 2 (Second-order transmission conditions)**.** *When $q \neq 0$, the transmission parameter becomes frequency-dependent: $\sigma(\xi) = p + q\xi^2$, and the convergence factor is*

$$\rho(\xi; p, q) = \left|\frac{(p + q\xi^2) - \lambda(\xi)}{(p + q\xi^2) + \lambda(\xi)}\right| \mathrm{e}^{-\lambda(\xi)L}.$$

The additional degree of freedom in q allows $\sigma(\xi)$ to better track $\lambda(\xi) = \sqrt{\eta + \xi^2}$ across the frequency range, which shrinks the minimax spectral radius. The same mechanism is expected to appear in this setting as well, since the suboptimality comes from minimax tuning under a biased frequency distribution rather than from the number of tunable parameters. We use the zeroth-order ($q = 0$) case for an explicit counterexample.

C Proof of Proposition 1

We work in the convergence-factor setup of Appendix B, with two tangential frequencies ξ_1, ξ_2, $\lambda_i = \lambda(\xi_i)$, $\lambda_1 < \lambda_2$, and the per-mode factors ρ_1, ρ_2 from equation 13. For brevity, write $r_i(\sigma) = |\sigma - \lambda_i|/(\sigma + \lambda_i)$ for the rational parts of ρ_i. The minimax-optimal $\sigma_{\mathrm{mm}} = \sqrt{\lambda_1\lambda_2}$ comes from equation 19, and $r_{\mathrm{mm}} = (\sqrt{\kappa} - 1)/(\sqrt{\kappa} + 1)$ with $\kappa = \lambda_2/\lambda_1$ comes from equation 20.

We organize the proof in three steps: **(1)** monotonicity of ρ_1, ρ_2 on the interval (λ_1, λ_2); **(2)** the first-order optimality condition at σ_{mm}, which gives the critical-weight characterization in equation 14 and the strict-gap claim in equation 15; and **(3)** the direction of the optimal shift relative to σ_{mm}.

Proof of Proposition 1. **Part 1.** Write $\rho_i(\sigma) = r_i(\sigma)\,\mathrm{e}^{-\lambda_i L}$. For $\sigma \in (\lambda_1, \lambda_2)$, we have $\sigma > \lambda_1$ and $\sigma < \lambda_2$, so

$$r_1(\sigma) = \frac{\sigma - \lambda_1}{\sigma + \lambda_1}, \qquad r_2(\sigma) = \frac{\lambda_2 - \sigma}{\sigma + \lambda_2}.$$

Differentiating,

$$r_1'(\sigma) = \frac{2\lambda_1}{(\sigma + \lambda_1)^2} > 0, \qquad r_2'(\sigma) = \frac{-2\lambda_2}{(\sigma + \lambda_2)^2} < 0, \tag{21}$$

which proves the strict monotonicity of the rational parts; since the exponentials $\mathrm{e}^{-\lambda_i L}$ are positive constants, ρ_1 is strictly increasing and ρ_2 is strictly decreasing in σ on this interval.

Part 2 (first-order optimality and critical weight). Differentiating J_α from equation 13,

$$\frac{dJ_\alpha}{d\sigma} = 2K\big[\alpha\,\rho_1^{2K-1}\,\rho_1' + (1-\alpha)\,\rho_2^{2K-1}\,\rho_2'\big]. \tag{22}$$

At $\sigma = \sigma_{\mathrm{mm}}$, the rational parts equioscillate, $r_1(\sigma_{\mathrm{mm}}) = r_2(\sigma_{\mathrm{mm}}) = r_{\mathrm{mm}}$, so $\rho_i(\sigma_{\mathrm{mm}}) = r_{\mathrm{mm}}\,\mathrm{e}^{-\lambda_i L}$ and equation 21 gives

$$\left.\frac{dJ_\alpha}{d\sigma}\right|_{\sigma_{\mathrm{mm}}} = 2K\,r_{\mathrm{mm}}^{2K-1}\left[\alpha\,\mathrm{e}^{-2K\lambda_1 L}\,\frac{2\lambda_1}{(\sigma_{\mathrm{mm}} + \lambda_1)^2} - (1-\alpha)\,\mathrm{e}^{-2K\lambda_2 L}\,\frac{2\lambda_2}{(\sigma_{\mathrm{mm}} + \lambda_2)^2}\right]. \tag{23}$$

We show that the two coefficients $\lambda_i/(\sigma_{\mathrm{mm}} + \lambda_i)^2$ are equal. Let $t = \sqrt{\kappa}$, so $\lambda_2 = t^2\lambda_1$ and $\sigma_{\mathrm{mm}} = t\lambda_1$. Then

$$\frac{\lambda_1}{(\sigma_{\mathrm{mm}} + \lambda_1)^2} = \frac{1}{\lambda_1(t+1)^2}, \qquad \frac{\lambda_2}{(\sigma_{\mathrm{mm}} + \lambda_2)^2} = \frac{t^2\lambda_1}{\lambda_1^2 t^2(1+t)^2} = \frac{1}{\lambda_1(1+t)^2}.$$

Denoting this common value by $C = 1/(\lambda_1(1+t)^2)$,

$$\left.\frac{dJ_\alpha}{d\sigma}\right|_{\sigma_{\mathrm{mm}}} = 2K\,r_{\mathrm{mm}}^{2K-1}\,2C\left(\alpha\,\mathrm{e}^{-2K\lambda_1 L} - (1-\alpha)\,\mathrm{e}^{-2K\lambda_2 L}\right). \tag{24}$$

Since $K \geq 1$, $r_{\mathrm{mm}} > 0$, and $C > 0$, the derivative vanishes if and only if

$$\frac{\alpha}{1-\alpha} = \mathrm{e}^{-2K(\lambda_2 - \lambda_1)L},$$

which gives $\alpha = \alpha^\star$ as defined in equation 14. Since $\lambda_2 > \lambda_1$ and $L > 0$, the right-hand side is strictly less than 1, so $\alpha^\star < 1/2$.

For $\alpha \neq \alpha^\star$, the derivative is nonzero, so σ_{mm} is not a critical point of J_α. Since J_α is smooth and the derivative at σ_{mm} is nonzero, there exists σ' in a neighborhood of σ_{mm} with $J_\alpha(\sigma') < J_\alpha(\sigma_{\mathrm{mm}})$. Moreover, J_α is continuous on $(0, \infty)$ and $J_\alpha(\sigma) \to \alpha\, \mathrm{e}^{-2K\lambda_1 L} + (1-\alpha)\, \mathrm{e}^{-2K\lambda_2 L}$ as $\sigma \to 0^+$ or $\sigma \to +\infty$ (since the rational parts $r_i \to 1$ in these limits). This endpoint limit is strictly larger than $J_\alpha(\sigma_{\mathrm{mm}})$, because $r_{\mathrm{mm}} < 1$. Therefore the strict local improvement $J_\alpha(\sigma') < J_\alpha(\sigma_{\mathrm{mm}})$ lies strictly below the endpoint limit, and the infimum is attained at some $\sigma_\alpha^\star \in (0, \infty)$. Hence $J_\alpha(\sigma_\alpha^\star) \leq J_\alpha(\sigma') < J_\alpha(\sigma_{\mathrm{mm}})$, establishing equation 15.

Part 3 (direction of the shift). From equation 24: if $\alpha > \alpha^\star$, the derivative at σ_{mm} is positive, so J_α is locally decreasing to the left of σ_{mm}, and there exists $\sigma' < \sigma_{\mathrm{mm}}$ with $J_\alpha(\sigma') < J_\alpha(\sigma_{\mathrm{mm}})$; this σ' is closer to λ_1 than σ_{mm}, meaning the optimal parameter shifts toward the slowly decaying low-frequency mode. In particular, since $\alpha^\star < 1/2$, the uniform weight $\alpha = 1/2$ falls in this regime: even without any distributional bias, the overlap makes the minimax rule suboptimal. Symmetrically, if $\alpha < \alpha^\star$, the derivative is negative and there exists $\sigma' > \sigma_{\mathrm{mm}}$ with $J_\alpha(\sigma') < J_\alpha(\sigma_{\mathrm{mm}})$, shifting toward λ_2. □

Remark 3 (Interpretation of the gap)**.** *The proposition reveals two distinct sources of suboptimality in classical minimax tuning. First, the overlap itself introduces an asymmetry: low-frequency modes decay slowly ($\mathrm{e}^{-\lambda_1 L}$ close to 1) while high-frequency modes are already damped efficiently ($\mathrm{e}^{-\lambda_2 L}$ small). This makes the loss dominated by the low-frequency mode, so the optimal σ shifts toward λ_1, while the minimax rule, which balances the rational factors equally, ignores this. The strict gap therefore exists even for $\alpha = 1/2$ (uniform mode weighting) whenever $L > 0$. Second, any additional distributional bias ($\alpha \neq 1/2$) compounds with the overlap-induced asymmetry: the critical weight $\alpha^\star$ is already below $1/2$, so biasing toward the low-frequency mode ($\alpha > 1/2$) widens the gap further.*

The critical weight $\alpha^\star$ decreases with the overlap L, the spectral ratio κ, and the unrolling depth K, widening the range of distributions for which the minimax rule is suboptimal. The asymmetry $|\alpha - \alpha^\star|$ controls the magnitude of the derivative at σ_{mm} and therefore the local improvement available by shifting σ. For problems with large condition numbers and spectrally biased data, the potential improvement over classical OSM can be significant.

D Proof of Proposition 3

Proposition 3 (Strict improvement from a positive-measure advantage set)**.** *Let ϕ_{base} and $\widetilde{\phi}$ be two admissible transmission laws. Suppose there exist a measurable set $U \subseteq \mathcal{Z}$ with $\nu(U) > 0$ and constants $\delta > 0$, $\Gamma \geq 0$ such that*

$$\ell_K(z; \widetilde{\phi}) \leq \ell_K(z; \phi_{base}) - \delta, \quad z \in U, \tag{25}$$

$$\ell_K(z; \widetilde{\phi}) \leq \ell_K(z; \phi_{base}) + \Gamma, \quad z \in \mathcal{Z} \setminus U, \tag{26}$$

and the gap dominates the degradation,

$$\delta\, \nu(U) > \Gamma\, \nu(\mathcal{Z} \setminus U). \tag{27}$$

Then $\mathcal{R}_\nu(\widetilde{\phi}) < \mathcal{R}_\nu(\phi_{base})$.

Proof of Proposition 3. By the population-risk definition in equation 11,

$$\mathcal{R}_\nu(\widetilde{\phi}) - \mathcal{R}_\nu(\phi_{\mathrm{base}}) = \int_{\mathcal{Z}} \big(\ell_K(z; \widetilde{\phi}) - \ell_K(z; \phi_{\mathrm{base}})\big)\, d\nu(z).$$

Splitting the integral over U and $\mathcal{Z} \setminus U$ and applying equation 25 and equation 26,

$$\begin{aligned} \mathcal{R}_\nu(\widetilde{\phi}) - \mathcal{R}_\nu(\phi_{\mathrm{base}}) &\leq \int_U (-\delta)\, d\nu(z) + \int_{\mathcal{Z}\setminus U} \Gamma\, d\nu(z) \\ &= -\delta\, \nu(U) + \Gamma\, \nu(\mathcal{Z} \setminus U). \end{aligned}$$

The conclusion follows from equation 27. □

Remark 4. *The bounded degradation Γ in equation 26 is essential: it relaxes the "strictly better everywhere" requirement to "better on a positive-measure set, by enough to outweigh degradations elsewhere." The structural parallel with Proposition 1 is direct: there, the asymmetry $|\alpha - \alpha^\star|$ played the role that $\delta\,\nu(U) - \Gamma\,\nu(\mathcal{Z} \setminus U)$ plays here.*

E Proof of Proposition 2

Proof of Proposition 2. Let $s_{\mathrm{OO2}}(d)$ denote the analytical OO2 tuple and define the oracle log-correction

$$c^\star(d) = \log s^\star_{\mathrm{stat}}(d) - \log s_{\mathrm{OO2}}(d),$$

with the logarithm applied componentwise. Since $s^\star_{\mathrm{stat}}$ has compact image in $(0,\infty)^2$ and s_{OO2} is positive and continuous on the compact descriptor set, $c^\star$ is Lipschitz on $\mathcal{D}_{\mathrm{stat},0}$. By assumption (iv), for every $\delta > 0$ there exists $\boldsymbol{\theta}_\delta \in \Theta$ whose log-correction $c(\cdot;\boldsymbol{\theta}_\delta)$ satisfies

$$\sup_{d \in \mathcal{D}_{\mathrm{stat},0}} \|c^\star(d) - c(d;\boldsymbol{\theta}_\delta)\| < \delta.$$

The represented UORAS transmission rule is $s(d;\boldsymbol{\theta}_\delta) = s_{\mathrm{OO2}}(d) \odot \exp(c(d;\boldsymbol{\theta}_\delta))$, so compactness and continuity of the exponential imply $\sup_d \|s^\star_{\mathrm{stat}}(d) - s(d;\boldsymbol{\theta}_\delta)\| \to 0$ as $\delta \to 0$. Let $\phi_{\boldsymbol{\theta}_\delta}(d) = (s(d;\boldsymbol{\theta}_\delta), \ldots, s(d;\boldsymbol{\theta}_\delta))$ be the stationary law represented by this network. By uniform continuity of the loss on the compact transmission range and the integrable envelope in assumption (iii), dominated convergence gives $\mathcal{R}_\nu(\phi_{\boldsymbol{\theta}_\delta}) \to \mathcal{R}^\star_{\nu,\mathrm{stat}}$ as $\delta \to 0$. Thus for every $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\mathcal{R}_\nu\big(\phi_{\boldsymbol{\theta}_\delta}\big) \le \mathcal{R}^\star_{\nu,\mathrm{stat}} + \varepsilon. \tag{28}$$

Choosing $0 < \varepsilon < \Delta_{\mathrm{stat}}$ (with $\Delta_{\mathrm{stat}} = \mathcal{R}_\nu(\phi_{\mathrm{OSM}}) - \mathcal{R}^\star_{\nu,\mathrm{stat}} > 0$ from assumption (i)) and setting $\boldsymbol{\theta}^\star = \boldsymbol{\theta}_\delta$,

$$\mathcal{R}_\nu\big(\phi_{\boldsymbol{\theta}^\star}\big) \le \mathcal{R}^\star_{\nu,\mathrm{stat}} + \varepsilon < \mathcal{R}^\star_{\nu,\mathrm{stat}} + \Delta_{\mathrm{stat}} = \mathcal{R}_\nu(\phi_{\mathrm{OSM}}),$$

which establishes equation 16. □

Remark 5 (On the Lipschitz hypothesis)**.** *Assumption (ii) of Proposition 2 is expected to hold in the Schwarz setting when the descriptors range over a compact, regular family: the eigenvalues of the modified Helmholtz operator depend continuously on the underlying problem parameters, and the best stationary descriptor-based transmission parameter inherits this regularity when the non-observed decomposition variables are fixed or coupled to the descriptors by the problem distribution. The compact positive-image condition avoids degenerate transmission parameters and makes the log-correction representation in equation 5 equivalent to approximating the transmission rule itself. Assumption (iv) is satisfied by sufficiently wide feedforward networks with continuous activation functions Hornik (1991); Cybenko (1989).*

F Convergence rate over bounded subdomains: full statement and proof

We now state the bounded-subdomain calculation with the full closed-form modal multiplier in cosh / sinh form, and give the Fourier-domain proof.

Proposition 4 (Convergence rate, bounded subdomains; formal)**.** *Let $\Omega = (\gamma,\zeta) \times \mathbb{R}$ be a spatial domain decomposed into two overlapping subdomains $\Omega_1 = (\gamma,\beta) \times \mathbb{R}$ and $\Omega_2 = (\alpha,\zeta) \times \mathbb{R}$, with $\gamma < \alpha < \beta < \zeta$. Consider the parallel Schwarz algorithm for the modified Helmholtz problem $(\eta - \Delta)u = f$. By linearity, the convergence analysis reduces to the homogeneous case $f = 0$. The iteration is given by*

$$(\eta - \Delta)u_1^n = 0 \text{ in } \Omega_1, \qquad (\eta - \Delta)u_2^n = 0 \text{ in } \Omega_2, \tag{29}$$

with homogeneous Dirichlet boundary conditions on the outer boundaries

$$u_1^n(\gamma, y) = 0, \qquad u_2^n(\zeta, y) = 0, \qquad y \in \mathbb{R}, \tag{30}$$

and Robin-type interface conditions

$$(\partial_x + \mathcal{S}_1)u_1^n(\beta,\cdot) = (\partial_x + \mathcal{S}_1)u_2^{n-1}(\beta,\cdot), \quad (-\partial_x + \mathcal{S}_2)u_2^n(\alpha,\cdot) = (-\partial_x + \mathcal{S}_2)u_1^{n-1}(\alpha,\cdot), \tag{31}$$

where $\mathcal{S}_j$ are linear operators along the interface in the y direction. Let $k \in \mathbb{R}$ be the spatial frequency variable in the y-direction, $\sigma_j(k) = \mathcal{F}(\mathcal{S}_j)$ the symbols of the transmission operators, and $r_1 = \sqrt{\eta + k^2}$ the principal root of the characteristic equation. Then the signed modal error multiplier at frequency k is

$$\rho_{opt}(k,\gamma,\alpha,\beta,\zeta) = \frac{r_1 \frac{\cosh(r_1(\beta-\zeta))}{\sinh(r_1(\alpha-\zeta))} + \sigma_1(k)\frac{\sinh(r_1(\beta-\zeta))}{\sinh(r_1(\alpha-\zeta))}}{r_1 \coth(r_1(\beta-\gamma)) + \sigma_1(k)} \cdot \frac{r_1 \frac{\cosh(r_1(\gamma-\alpha))}{\sinh(r_1(\gamma-\beta))} + \sigma_2(k)\frac{\sinh(r_1(\gamma-\alpha))}{\sinh(r_1(\gamma-\beta))}}{r_1 \coth(r_1(\zeta-\alpha)) + \sigma_2(k)}. \tag{32}$$

Proof of Proposition 4. The Fourier transform yields

$$\begin{aligned} (\eta + k^2 - \partial_{xx})\,\hat{u}_1^n &= 0, \quad x < \beta,\ k \in \mathbb{R}, \\ (\partial_x + \sigma_1(k))\,\hat{u}_1^n(\beta,k) &= (\partial_x + \sigma_1(k))\,\hat{u}_2^{n-1}(\beta,k), \end{aligned} \tag{33}$$

$$\begin{aligned} (\eta + k^2 - \partial_{xx})\,\hat{u}_2^n &= 0, \quad x > \alpha,\ k \in \mathbb{R}, \\ (-\partial_x + \sigma_2(k))\,\hat{u}_2^n(\alpha,k) &= (-\partial_x + \sigma_2(k))\,\hat{u}_1^{n-1}(\alpha,k). \end{aligned} \tag{34}$$

The subdomain solutions in the frequency domain are $\hat{u}_j^n(x,k) = A_j(k)e^{r_1 x} + B_j(k)e^{r_2 x}$ with $r_1 = \sqrt{\eta + k^2}$ and $r_2 = -\sqrt{\eta + k^2}$. Imposing homogeneous Dirichlet boundary conditions on the outer boundaries,

$$\hat{u}_1^n(\gamma,k) = A_1(k)e^{r_1\gamma} + B_1(k)e^{-r_1\gamma} = 0, \qquad \hat{u}_2^n(\zeta,k) = A_2(k)e^{r_1\zeta} + B_2(k)e^{-r_1\zeta} = 0,$$

yields

$$\hat{u}_1(x,k) = \frac{\sinh(r_1(\gamma-x))}{\sinh(r_1(\gamma-\beta))}\,\hat{u}_1^n(\beta,k), \qquad \hat{u}_2(x,k) = \frac{\sinh(r_1(x-\zeta))}{\sinh(r_1(\alpha-\zeta))}\,\hat{u}_2^n(\alpha,k). \tag{35}$$

Substituting into the Robin interface conditions gives

$$\begin{aligned} \big(r_1 \coth(r_1(\beta-\gamma)) + \sigma_1(k)\big)\,\hat{u}_1^n(\beta,k) = \left(r_1 \frac{\cosh(r_1(\beta-\zeta))}{\sinh(r_1(\alpha-\zeta))} + \sigma_1(k)\frac{\sinh(r_1(\beta-\zeta))}{\sinh(r_1(\alpha-\zeta))}\right) \\ \cdot\, \hat{u}_2^{n-1}(\alpha,k), \end{aligned} \tag{36}$$

and similarly,

$$\begin{aligned} \big(r_1 \coth(r_1(\zeta-\alpha)) + \sigma_2(k)\big)\,\hat{u}_2^n(\alpha,k) = \left(r_1 \frac{\cosh(r_1(\gamma-\alpha))}{\sinh(r_1(\gamma-\beta))} + \sigma_2(k)\frac{\sinh(r_1(\gamma-\alpha))}{\sinh(r_1(\gamma-\beta))}\right) \\ \cdot\, \hat{u}_1^{n-1}(\beta,k). \end{aligned} \tag{37}$$

Composing the two yields a two-step recursion for $\hat{u}_1^n(\beta,k)$ (resp. $\hat{u}_2^n(\alpha,k)$) whose signed modal multiplier is precisely equation 32; the corresponding contraction factor is its absolute value. □

Remark 6 (Recovery of the unbounded limit)**.** *In the limit $\gamma \to -\infty$, $\zeta \to +\infty$, the cosh / sinh ratios in equation 32 simplify and one recovers the classical convergence rate of the unbounded two-subdomain Schwarz iteration Gander (2006),*

$$\lim_{\substack{\gamma\to-\infty,\\ \zeta\to+\infty}} \rho_{opt}(k,\gamma,\alpha,\beta,\zeta) = \frac{-r_1 + \sigma_1(k)}{r_1 + \sigma_1(k)}\,\frac{-r_1 + \sigma_2(k)}{r_1 + \sigma_2(k)}\,\mathrm{e}^{-2r_1(\beta-\alpha)}. \tag{38}$$

For finite γ and ζ, the bounded-subdomain factor differs from this limit, which is the formal sense in which the OO2 parameters fixed under the unbounded analysis are not exactly tuned to bounded discretized problems.

G Additional remarks on UORAS

G.1 Residual-conditioned approximation

Remark 7 (Residual-conditioned approximation)**.** *The non-stationary UORAS variant uses a shared log-correction map $g(h,\eta,\rho;\boldsymbol{\theta}) \in \mathbb{R}^2$ with $\rho = \|\mathbf{r}^k\|$, rather than an arbitrary layerwise vector*

in $\mathbb{R}^{2K}$. Let Φ_{res} be the class of laws generated by such shared residual-conditioned maps, $s_k = s_{OO2}(h,\eta) \odot \exp(g(h,\eta,\|\mathbf{r}^k\|))$. If the best law in Φ_{res} has a positive population-risk gap over ϕ_{OSM}, if the corresponding oracle log-correction map is Lipschitz on the compact state-descriptor set visited by the unrolled dynamics, and if the MLP class universally approximates such maps with an integrable loss envelope, then the proof of Proposition 2 applies with the stationary log-correction replaced by the residual-conditioned one. Thus the residual-conditioned result matches the implemented non-stationary architecture: the network is shared across layers and produces two log-correction scalars from $(h,\eta,\|\mathbf{r}^k\|)$ at each layer.

G.2 Layer-wise schedule versus a shared parameter

Remark 8 (Layer-wise vs. shared schedule)**.** *The unrolled UORAS model allows different transmission parameters $s_k(z;\boldsymbol{\theta})$ at each layer k, whereas classical OSM applies a fixed parameter at every iteration. This additional freedom is analogous to the well-known advantage of Chebyshev iteration over Richardson iteration: by choosing different relaxation parameters at successive steps, Chebyshev iteration targets different spectral components at each step and achieves a convergence rate that improves exponentially with the number of steps, compared to the fixed-parameter rate. The unrolled Schwarz model can analogously learn an iteration-dependent schedule $(s_0, s_1, \ldots, s_{K-1})$ that addresses different parts of the error spectrum at each layer, with classical shared-parameter OSM as the special case $s_0 = s_1 = \cdots = s_{K-1}$.*

The propositions in Section 5 do not yet exploit this layer-wise freedom: Proposition 1 constructs a strict gap using a shared (but non-minimax) parameter. Extending the explicit gap analysis to layer-wise schedules (and quantifying the additional benefit) is an interesting direction for future work.

G.3 Optimization landscape

Remark 9 (Optimization landscape)**.** *The theoretical results in Section 5 concern the* expressivity *of the learned parametrization: they identify conditions under which good descriptor-based transmission laws exist within the hypothesis class. They do not address whether gradient-based training will* find *these laws.*

Two structural observations partially mitigate this concern. First, the loss $\ell_K(z;\boldsymbol{\theta})$ is a smooth function of the transmission parameters (p_k, q_k), because the Schwarz preconditioner $\mathbf{M}_k$ depends analytically on these parameters (the interface block $\widetilde{T}_k$ is affine in p_k and q_k), and matrix inversion and multiplication preserve smoothness. Second, the learned law has low dimensionality: $2K$ real parameters per problem instance, which limits the complexity of the loss landscape relative to typical deep learning settings.

That said, unrolled iterative methods are known to exhibit vanishing or exploding gradients as the depth K increases (see, e.g., Monga et al. (2021); Shlezinger et al. (2025)), and the loss may have spurious local minima. In practice, these issues can be addressed by keeping K moderate, using gradient clipping, or warm-starting with the classical OSM parameters (which is natural in the multiplicative parametrization equation 5: warm start at $p_c \equiv q_c \equiv 0$).

G.4 Generalization

Remark 10 (Generalization from finite samples)**.** *The population risk $\mathcal{R}_\nu(\phi)$ is defined as an expectation over the problem distribution ν, but the model is trained on a finite sample of problem instances and their descriptors. For the UORAS parametrization, the learned law is determined by the small MLP $\Psi(\cdot;\boldsymbol{\theta})$ in equation 6, whose output gives the log-correction scalars (p_c, q_c) that are exponentiated and multiplied by the OO2 parameters at each unrolled layer. Standard statistical learning theory bounds (e.g., via VC dimension or covering numbers) ensure that the generalization gap scales gracefully with the network's capacity. Because the required output dimension is exceptionally small ($2K$ scalars) and the network itself has only $\sim$2300 parameters, the sample complexity is well-controlled, and generalization is typically not the bottleneck in this setting.*