RAMPING UP PINN TRAINING SIGNALS VIA PARAMETER-SENSITIVITY-GUIDED COLLOCATION REFINEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

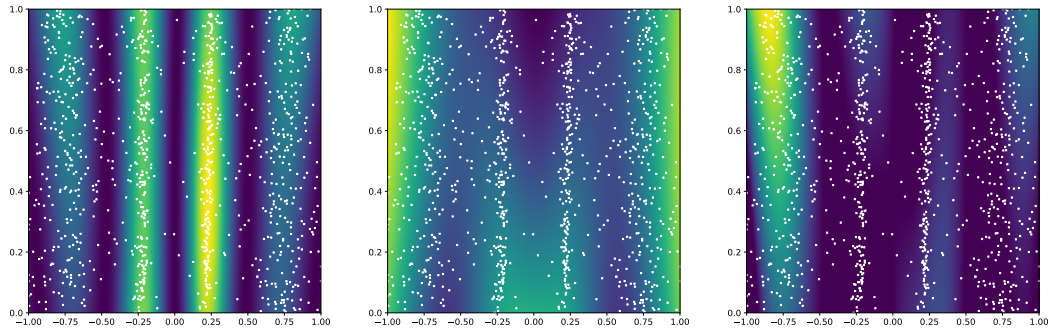
Collocation point selection plays a crucial role in training physics-informed neural networks (PINNs), as physics constraints are enforced only at sampled locations. Most existing adaptive sampling methods prioritize points with large constraint violations, typically measured by PDE residual magnitude, thereby implicitly treating residual magnitude as a sufficient pointwise indicator of training utility. We revisit adaptive collocation from a parameter-space perspective and introduce *pointwise parameter sensitivity*, defined as the norm of the gradient of the pointwise PDE loss with respect to the trainable parameters. We show that sampling in proportion to this quantity directly amplifies the induced parameter-space training signal, up to a directional coherence factor. Building on this criterion, we propose **R**adius-bounded **A**scent for **M**aximizing **P**arameter-gradient norm (**RAMP**), a parameter-sensitivity-guided collocation refinement strategy that adapts sampled points to the current parameter state. Given candidates drawn from a prescribed base sampler, RAMP independently refines each point by projected unit-gradient ascent on pointwise parameter sensitivity within a bounded radius. This local refinement preserves the spatial coverage induced by the base sampler while steering collocation points toward regions that provide stronger parameter-space signals. We further analyze magnitude-based sampling criteria and show that they can be viewed as residual-only approximations of parameter sensitivity, which neglect the local parameter responsiveness of the PDE residual. Across benchmark PDEs and multiple PINN optimizers, RAMP achieves consistently larger aggregate gains over the base sampler than recent adaptive sampling methods, while retaining parallelizable point updates.

1 INTRODUCTION

Partial differential equations (PDEs) are fundamental tools for describing physical phenomena in science and engineering (Morton and Mayers, 2005; Haghighat et al., 2021). Analytic solutions are rarely available in practice, particularly for nonlinear or high-dimensional PDEs (Zhu et al., 2019; Bu et al., 2026). Classical mesh-based solvers discretize the equations (Bathe, 2006; Grossmann et al., 2007), inevitably causing mesh-dependent numerical errors (Papež et al., 2014). Physics-informed neural networks (PINNs) provide a complementary mesh-free approach by penalizing violations of the PDEs at collocation points sampled from the continuous domain (Raissi et al., 2019). Therefore, the selected collocation points are central to approximating physical consistency over the full domain (Wu et al., 2024). Standard domain samplers, such as uniform sampling, Latin hypercube sampling, and low-discrepancy sequences, are widely used in collocation sampling to achieve broad coverage of the continuous domain (Fang et al., 2023).

While these samplers are simple and ensure global coverage, they do not respond to the evolving PINN training state. This limitation has motivated adaptive collocation sampling, which updates collocation points during training using model-dependent information evaluated at the current training state (Lu et al., 2021). A common approach uses the current PDE residual magnitude to refine collocation points, as they indicate severe violations and relate to controlling worst-case errors (Wang et al., 2022a). Accordingly, residual magnitude has been widely adopted as a pointwise sampling criterion (Nabian et al., 2021; Wu et al., 2023; Daw et al., 2023). The empirical success of magnitude-based sampling naturally raises a fundamental question: *does residual magnitude sufficiently characterize*

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.



(a) (b) (c)

Figure 1: **1D Burgers snapshot during PINN training.** All panels are evaluated on the same spatio-temporal grid. White points are collocation points sampled by residual magnitude, yielding dense samples in high-residual regions and sparse samples elsewhere. (a) PDE residual magnitude. (b) Parameter-gradient norm of the signed PDE residual. (c) Immediate reduction in PDE residual magnitude after one descent step using the shown points. **Yellow** indicates larger values, and **purple** indicates values near zero. Further details are provided in Appendix A.

the pointwise training state for adaptive collocation sampling, or should adaptive sampling directly target parameter-space training signals?

To empirically examine this, Figure 1 compares (a) the residual magnitude before the update, (b) the parameter-gradient norm of the signed PDE residual, and (c) the immediate reduction in residual magnitude. Specifically, (c) is computed as the residual magnitude in (a) minus the residual magnitude after one descent step using the displayed collocation points, and therefore measures the local improvement in PDE-constraint satisfaction induced by that update. Although residual-based sampling concentrates points in high-residual regions, (c) shows that the subsequent update does not necessarily reduce the residual most in those heavily sampled areas. Instead, locations with moderate or small residuals can exhibit comparable or even larger immediate reductions, and such regions are often associated with large parameter-gradient norms. This empirical pattern is consistent with recent adaptive training strategies indicating that residual magnitude alone may be an incomplete training-state signal. Jeong et al. (2025) shows that excessive concentration on high-residual points can degrade stability by neglecting regions with moderate residual magnitude. Furthermore, Duan et al. (2025) show that going beyond residual magnitude by estimating sample difficulty from spatio-temporal residual variations and adjusting the priorities throughout training can be beneficial. Together, these observations suggest that there is still no established pointwise criterion that captures how a collocation point affects PINN training.

To address this gap, we view collocation sampling from a parameter-space perspective, where each point is evaluated by the sensitivity of its residual to the trainable parameters. We instantiate this criterion with the parameter-gradient norm of the pointwise PDE residual loss, which indicates the strength of the parameter-space training signal induced by that point. In this context, we propose **R**adius-bounded **A**scent for **M**aximizing **P**arameter-gradient norm (**RAMP**). RAMP first samples initial candidates from a standard domain sampler and then applies a radius-bounded ascent step to each candidate, moving it toward a nearby location with a larger parameter-space training signal. The radius bound retains the spatial coverage inherited from the sampler, while the ascent step makes the final collocation points responsive to the current parameter state.

The main contributions are summarized as follows. We first introduce a parameter-space criterion for adaptive collocation sampling, using the parameter-gradient norm of the pointwise residual loss to characterize the training signal induced by each collocation point. We provide a theoretical analysis of magnitude-based sampling, showing that residual magnitude is a tractable but incomplete proxy for parameter sensitivity. We then develop RAMP as a computationally efficient realization of this criterion, ramping up parameter-space training signals while retaining the coverage and parallelism of standard sampling. Finally, we evaluate RAMP on benchmark PDEs against recent adaptive sampling baselines across three PINN training optimizers, demonstrating improved empirical performance.

2 PRELIMINARIES AND BACKGROUNDS

Collocation Sampling in PINNs Let $\Omega \subset \mathbb{R}^p$ denote a bounded spatio-temporal domain, and let $\mathbf{z} \in \Omega$ be a collocation point. For time-dependent PDEs, we write $\mathbf{z} = (\mathbf{x}, t)$, where $\mathbf{x}$ denotes the spatial coordinate and t denotes time. We consider a neural network $u_\theta : \mathbb{R}^p \to \mathbb{R}^m$ parameterized by $\boldsymbol{\theta} \in \mathbb{R}^d$. We assume that the target solution $\mathbf{u}$ satisfies a governing PDE of the form

$$\mathcal{F}[\mathbf{u}](\mathbf{z}) := \frac{\partial \mathbf{u}}{\partial t}(\mathbf{z}) + \mathcal{N}_{\mathbf{x}}[\mathbf{u}](\mathbf{z}) = 0, \qquad \mathbf{z} \in \Omega, \tag{1}$$

where $\mathcal{N}_{\mathbf{x}}$ denotes a spatial differential operator. Following the PINN formulation (Raissi et al., 2019), we obtain the PDE residual by substituting the neural approximation into the governing equation:

$$R_\theta(\mathbf{z}) := \mathcal{F}[u_\theta](\mathbf{z}) \in \mathbb{R}, \qquad \mathbf{z} \in \Omega. \tag{2}$$

The pointwise PDE residual loss and the corresponding full-domain PDE objective are defined as

$$\ell(\mathbf{z}; \boldsymbol{\theta}) := \frac{1}{2}|R_{\boldsymbol{\theta}}(\mathbf{z})|^2, \qquad \mathcal{L}_Q(\boldsymbol{\theta}) := \mathbb{E}_{\mathbf{z}\sim Q}\big[\ell(\mathbf{z}; \boldsymbol{\theta})\big], \tag{3}$$

where Q is a reference distribution on Ω, typically the uniform distribution. In practice, this full-domain objective is approximated using a finite collocation set $\mathbf{Z}^{(k)} = \{\mathbf{z}_i^{(k)}\}_{i=1}^N \subset \Omega$ with training iteration k, and the parameters are updated with learning rate $\eta^{(k)} > 0$ by

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \eta^{(k)} \nabla_{\boldsymbol{\theta}} \mathcal{L}(\mathbf{Z}^{(k)}; \boldsymbol{\theta}^{(k)}), \quad \text{where } \mathcal{L}(\mathbf{Z}^{(k)}; \boldsymbol{\theta}^{(k)}) := \frac{1}{N}\sum_{i=1}^{N} \ell(\mathbf{z}_i^{(k)}; \boldsymbol{\theta}^{(k)}). \tag{4}$$

Thus, collocation sampling determines the finite set of points through which the PDE constraint is imposed during training. Boundary and initial conditions can also be incorporated through additional residual losses; in this work, we focus only on the PDE residual.

Training State Guided Collocation Sampling Adaptive collocation methods update training points using model-dependent signals that evolve with the PINN training state, most commonly the pointwise PDE residual magnitude. L^∞-based method draws points from a standard domain sampler and adversarially perturbs them in input space toward regions with larger residual magnitudes (Wang et al., 2022a). Retain-Resample-Release (R3) retains points whose residual magnitudes exceed the average and replaces the remaining points with newly drawn samples from the standard domain sampler (Daw et al., 2023). Residual-based adaptive distribution (RAD) evaluates residual magnitudes on candidates drawn from the standard domain sampler and resamples points with probabilities determined by those magnitudes (Wu et al., 2023). Langevin dynamics-based adaptive sampling (LAS) evolves points according to the current residual landscape, using residual-based drift to steer them toward regions with larger residual magnitude (Jeong et al., 2025). Despite their different adaptation mechanisms, these methods primarily characterize the current training state through residual magnitude and do not explicitly account for the parameter-space training signal induced by each collocation point.

3 METHODOLOGY

3.1 PARAMETER SENSITIVITY AS A MEASURE OF THE TRAINING STATE

We now view the training state through the sensitivity of the loss with respect to the model parameters. Specifically, under the reference distribution Q, we define the population parameter gradient and its norm by

$$\mathbf{G}_Q(\boldsymbol{\theta}) := \nabla_{\boldsymbol{\theta}} \mathcal{L}_Q(\boldsymbol{\theta}), \qquad S_Q(\boldsymbol{\theta}) := \|\mathbf{G}_Q(\boldsymbol{\theta})\|_2. \tag{5}$$

We refer to $S_Q(\boldsymbol{\theta})$ as the parameter sensitivity of the PDE residual loss. It measures how strongly the loss can change under an infinitesimal perturbation of the parameters.

Theorem 1 (One-step descent lower bound via parameter sensitivity)**.** *Assume that $\mathcal{L}_Q$ is β-smooth with respect to $\boldsymbol{\theta}$. For any $0 < \eta \le 1/\beta$, let $\boldsymbol{\theta}^+ = \boldsymbol{\theta} - \eta \nabla_{\boldsymbol{\theta}} \mathcal{L}_Q(\boldsymbol{\theta})$. Then*

$$\mathcal{L}_Q(\boldsymbol{\theta}) - \mathcal{L}_Q(\boldsymbol{\theta}^+) \ge \frac{\eta}{2} S_Q(\boldsymbol{\theta})^2 \quad \textit{and} \quad \|\boldsymbol{\theta}^+ - \boldsymbol{\theta}\|_2 = \eta S_Q(\boldsymbol{\theta}). \tag{6}$$

Theorem 1 shows that $S_Q(\boldsymbol{\theta})$ quantifies both the magnitude of the parameter update and a guaranteed decrease in the full-domain loss under gradient descent. Therefore, the parameter-gradient norm is a natural measure of the training state: a large value indicates that the loss still admits a strong descent signal, while a small value indicates that the parameters are close to being stationary for the full-domain objective.

Parameter Sensitivity in Existing Training Criteria The parameter-gradient norm as a measure of the training state is consistent with a parameter-sensitivity view of training dynamics. This view appears in several training criteria that use parameter-space responses to assess either parameter states or training samples. In studies of parameter symmetry, loss-preserving transformations are used to move a model to an equivalent parameter configuration with stronger sensitivity, showing that parameter configurations with equal loss values can differ in their usefulness for subsequent optimization (Zhao et al., 2022; 2024). In importance sampling, samples that induce larger parameter-space responses, such as larger gradient norms, are assigned higher priority during training (Katharopoulos and Fleuret, 2018; Paul et al., 2021). In PINNs, NTK-based approaches further characterize training through parameter-Jacobian interactions and their spectral modes, typically analyzing the dynamics under an approximately fixed empirical NTK over a local training phase (Wang et al., 2022b; Lau et al., 2024). Together, these lines motivate a parameter-space view of training, while suggesting a complementary pointwise criterion that directly reflects the evolving training state at the current parameter configuration, without relying on a locally fixed kernel representation.

3.2 Pointwise Parameter Sensitivity and Proportional Sampling

Pointwise Parameter Sensitivity The parameter sensitivity in Theorem 1 is defined at the level of the full-domain loss. Ideally, an adaptive collocation sampler should select a finite collocation set $\mathbf{Z} = \{\mathbf{z}_i\}_{i=1}^N$ that yields a large parameter sensitivity. The induced parameter sensitivity for $\mathbf{Z}$ is

$$S_{\mathbf{Z}}(\boldsymbol{\theta}) = \left\| \frac{1}{N} \sum_{i=1}^{N} \nabla_{\boldsymbol{\theta}} \ell(\mathbf{z}_i; \boldsymbol{\theta}) \right\| = \frac{1}{N} \sqrt{\sum_{i=1}^{N} \sum_{j=1}^{N} \langle \nabla_{\boldsymbol{\theta}} \ell(\mathbf{z}_i; \boldsymbol{\theta}), \nabla_{\boldsymbol{\theta}} \ell(\mathbf{z}_j; \boldsymbol{\theta}) \rangle}. \tag{7}$$

This expression contains all pairwise interactions among the selected samples. Thus, the contribution of each point cannot be evaluated independently, because the norm depends on the vector-valued sum of all selected pointwise gradients. Therefore, directly maximizing $S_{\mathbf{Z}}(\boldsymbol{\theta})$ requires accounting for coupled effects among samples, making the sampling decision difficult to parallelize and computationally expensive when the candidate pool is large. To obtain an efficient pointwise surrogate, we consider the parameter-space signal induced by a single collocation point. This leads to a criterion that can be locally improved independently for each candidate, enabling parallel computation.

For fixed parameters $\boldsymbol{\theta}$, we define the pointwise parameter gradient and its norm as

$$\mathbf{g}_{\boldsymbol{\theta}}(\mathbf{z}) := \nabla_{\boldsymbol{\theta}} \ell(\mathbf{z}; \boldsymbol{\theta}) = R_{\boldsymbol{\theta}}(\mathbf{z}) \nabla_{\boldsymbol{\theta}} R_{\boldsymbol{\theta}}(\mathbf{z}), \qquad \psi_{\boldsymbol{\theta}}(\mathbf{z}) := \|\mathbf{g}_{\boldsymbol{\theta}}(\mathbf{z})\| = |R_{\boldsymbol{\theta}}(\mathbf{z})| \|\nabla_{\boldsymbol{\theta}} R_{\boldsymbol{\theta}}(\mathbf{z})\|. \tag{8}$$

We call $\psi_{\boldsymbol{\theta}}(\mathbf{z})$ the pointwise parameter sensitivity. This decomposition shows that residual magnitude is only one factor of the pointwise training signal; the other factor is the sensitivity of the residual with respect to the parameters.

The connection between pointwise parameter sensitivity and finite-set parameter sensitivity can be made explicit by decomposing each pointwise gradient into its norm and direction. For $\psi_i = \psi_{\boldsymbol{\theta}}(\mathbf{z}_i)$, define

$$\mathbf{v}_i := \begin{cases} \mathbf{g}_{\boldsymbol{\theta}}(\mathbf{z}_i)/\psi_i, & \psi_i > 0, \\ \mathbf{0}, & \psi_i = 0. \end{cases} \qquad S_{\mathbf{Z}}(\boldsymbol{\theta}) = \frac{1}{N} \left\| \sum_{i=1}^{N} \psi_i \mathbf{v}_i \right\| = \frac{1}{N} \sqrt{\sum_{i=1}^{N} \sum_{j=1}^{N} \psi_i \psi_j \langle \mathbf{v}_i, \mathbf{v}_j \rangle}. \tag{9}$$

Thus, ψ_i determines the magnitude of each pointwise contribution, while $\mathbf{v}_i$ determines whether the contributions reinforce or cancel each other after aggregation.

To quantify this aggregation effect, let

$$\bar{\psi}_{\mathbf{Z}} := \frac{1}{N} \sum_{i=1}^{N} \psi_i, \qquad \kappa_{\mathbf{Z}} := \frac{\left\| \sum_{i=1}^{N} \psi_i \mathbf{v}_i \right\|}{\sum_{i=1}^{N} \psi_i} \in [0, 1], \qquad \left(\sum_{i=1}^{N} \psi_i > 0 \right). \tag{10}$$

Then $S_{\mathbf{Z}}(\boldsymbol{\theta}) = \kappa_{\mathbf{Z}} \bar{\psi}_{\mathbf{Z}}$ where the coefficient $\kappa_{\mathbf{Z}}$ measures the directional coherence of the selected pointwise gradients. Hence, the finite-set parameter sensitivity equals the average pointwise parameter sensitivity up to the multiplicative factor $\kappa_{\mathbf{Z}}$. In particular, if $\kappa_{\mathbf{Z}} \geq \rho$ for some $\rho \in [0, 1]$, then

$$S_{\mathbf{Z}}(\boldsymbol{\theta}) \geq \rho \bar{\psi}_{\mathbf{Z}}. \tag{11}$$

Thus, increasing ψ_i improves a lower bound on the induced parameter sensitivity as long as the selected gradients maintain nontrivial directional coherence.

Sampling Proportional to Pointwise Sensitivity We next consider the distribution that samples points proportionally to their pointwise parameter sensitivity. For any distribution P on Ω, define

$$\mathbf{G}_P(\boldsymbol{\theta}) := \mathbb{E}_{\mathbf{z}\sim P}[\mathbf{g}_{\boldsymbol{\theta}}(\mathbf{z})], \qquad S_P(\boldsymbol{\theta}) := \|\mathbf{G}_P(\boldsymbol{\theta})\|. \tag{12}$$

Let P_ψ be the distribution whose density with respect to the reference distribution Q is proportional to $\psi_{\boldsymbol{\theta}}(\mathbf{z})$:

$$\frac{dP_\psi}{dQ}(\mathbf{z}) := \frac{\psi_{\boldsymbol{\theta}}(\mathbf{z})}{\mathbb{E}_{\mathbf{z}\sim Q}[\psi_{\boldsymbol{\theta}}(\mathbf{z})]}. \tag{13}$$

Theorem 2 (Parameter sensitivity under sampling proportional to pointwise sensitivity)**.** *Fix $\boldsymbol{\theta}$, and write $\mathbf{g}_{\boldsymbol{\theta}}(\mathbf{z}) = \psi_{\boldsymbol{\theta}}(\mathbf{z})\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{z})$, where $\|\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{z})\| = 1$ whenever $\psi_{\boldsymbol{\theta}}(\mathbf{z}) > 0$, and $\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{z}) = \mathbf{0}$ otherwise. Assume $0 < \mathbb{E}_Q[\psi_{\boldsymbol{\theta}}(\mathbf{z})] < \infty$ and $\mathbb{E}_Q[\psi_{\boldsymbol{\theta}}(\mathbf{z})^2] < \infty$. Then*

$$S_{P_\psi}(\boldsymbol{\theta}) = \frac{\left\|\mathbb{E}_Q[\psi_{\boldsymbol{\theta}}(\mathbf{z})^2\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{z})]\right\|}{\mathbb{E}_Q[\psi_{\boldsymbol{\theta}}(\mathbf{z})]}. \tag{14}$$

Consequently, if for some $\rho \in [0,1]$,

$$\left\|\mathbb{E}_Q[\psi_{\boldsymbol{\theta}}(\mathbf{z})^2\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{z})]\right\| \geq \rho\, \mathbb{E}_Q[\psi_{\boldsymbol{\theta}}(\mathbf{z})^2],$$

then

$$S_{P_\psi}(\boldsymbol{\theta}) \geq \rho\, \frac{\mathbb{E}_Q[\psi_{\boldsymbol{\theta}}(\mathbf{z})^2]}{\mathbb{E}_Q[\psi_{\boldsymbol{\theta}}(\mathbf{z})]} \geq \rho\, \mathbb{E}_Q[\psi_{\boldsymbol{\theta}}(\mathbf{z})]. \tag{15}$$

Theorem 2 shows that sampling in proportion to pointwise sensitivity amplifies the induced parameter sensitivity through the ratio of the second moment to the first moment, up to directional coherence. Thus, regions with large $\psi_{\boldsymbol{\theta}}(\mathbf{z})$ receive stronger influence than under the reference distribution.

The same finite-set decomposition gives an implication for local pointwise maximization. Suppose that a set of initial candidates $\mathbf{Z}$ is moved to $\widetilde{\mathbf{Z}} = \{\widetilde{\mathbf{z}}_i\}_{i=1}^N$, and that each move increases the pointwise parameter sensitivity by at least $\Delta_i \geq 0$, i.e., $\psi_{\boldsymbol{\theta}}(\widetilde{\mathbf{z}}_i) \geq \psi_{\boldsymbol{\theta}}(\mathbf{z}_i) + \Delta_i$. If the moved set satisfies $\kappa_{\widetilde{\mathbf{Z}}} \geq \rho$, then

$$S_{\widetilde{\mathbf{Z}}}(\boldsymbol{\theta}) = \kappa_{\widetilde{\mathbf{Z}}}\bar{\psi}_{\widetilde{\mathbf{Z}}} \geq \rho\left(\bar{\psi}_{\mathbf{Z}} + \frac{1}{N}\sum_{i=1}^N \Delta_i\right). \tag{16}$$

Thus, local maximization of $\psi_{\boldsymbol{\theta}}(\mathbf{z})$ at individual points improves a lower bound on the finite-set parameter sensitivity, as long as the aggregated pointwise gradients retain nontrivial directional coherence. Empirical diagnostics in Appendix A further indicate that directional coherence remains nontrivial and that pointwise sensitivity captures training-state information beyond residual magnitude across different training states, supporting the practical relevance of this analysis.

3.3 Radius-Bounded Ascent for Maximizing Parameter-Gradient Norm

The analysis in Section 3.2 suggests that an adaptive sampler should favor points with large pointwise parameter sensitivity $\psi_{\boldsymbol{\theta}}(\mathbf{z})$, since increasing this quantity improves a lower bound on the finite-set parameter sensitivity when the aggregated pointwise gradients retain nontrivial directional coherence. At the same time, because the finite-set sensitivity also depends on the directions of the selected pointwise gradients, increasing $\psi_{\boldsymbol{\theta}}(\mathbf{z})$ should not be done by allowing the collocation set to drift arbitrarily away from the reference samples drawn from Q. This motivates a local adaptation strategy that increases $\psi_{\boldsymbol{\theta}}(\mathbf{z})$ while keeping each point close to its initial Q-sample. However, directly sampling from P_ψ requires estimating a global normalizer and drawing samples from the resulting density over Ω, while directly maximizing $S_{\mathbf{Z}}(\boldsymbol{\theta})$ requires accounting for interactions among all selected samples. We therefore propose RAMP as an efficient realization of this local adaptation strategy. Starting from candidates drawn from Q, RAMP independently moves each point along an ascent direction of $\psi_{\boldsymbol{\theta}}(\mathbf{z})$, while restricting the displacement to a radius-ε neighborhood of its initial location. This radius-bounded ascent preserves the coverage inherited from the reference sampler and makes the collocation points responsive to the current parameter state, as outlined in Algorithm 1.

At training iteration k, RAMP first draws initial candidates $\mathbf{Z}^{(0)} = \{\mathbf{z}_i^{(0)}\}_{i=1}^N$ from a standard domain sampler Q. For each candidate $\mathbf{z}_i^{(0)}$, the ideal local target is

$$\mathbf{z}_i^\star \in \operatorname*{argmax}_{\mathbf{z}\in\Omega\cap B_i} \psi_{\boldsymbol{\theta}^{(k)}}(\mathbf{z}), \qquad B_i := \{\mathbf{z} : \|\mathbf{z} - \mathbf{z}_i^{(0)}\| \leq \varepsilon\}. \tag{17}$$

Algorithm 1 Radius-Bounded Ascent for Maximizing Parameter-Gradient Norm (RAMP)

Require: parameter $\boldsymbol{\theta}^{(k)}$, sampler Q, radius ε, steps K, points N
Ensure: RAMP collocation set $\widetilde{\mathbf{Z}}^{(k)}$
1: Draw $\mathbf{z}_{1:N}^{(0)} \sim Q^{\otimes N}$, set $\alpha = \varepsilon/K$.
2: Define $\mathrm{unit}(\mathbf{d}) := \mathbf{d}/\|\mathbf{d}\|$ if $\|\mathbf{d}\| > 0$, and $\mathrm{unit}(\mathbf{d}) := \mathbf{0}$ otherwise
3: **for** $s = 0, \ldots, K-1$ **do**
4: **Parallel over** $i = 1, \ldots, N$**:**
5: $\mathbf{d}_i^{(s)} \leftarrow \nabla_{\mathbf{z}} \psi_{\boldsymbol{\theta}^{(k)}}(\mathbf{z}_i^{(s)}), \qquad \mathbf{z}_i^{(s+1)} \leftarrow \Pi_\Omega\Big(\mathbf{z}_i^{(s)} + \alpha\ \mathrm{unit}(\mathbf{d}_i^{(s)})\Big)$
6: **end for**
7: $\widetilde{\mathbf{Z}}^{(k)} \leftarrow \{\mathbf{z}_i^{(K)}\}_{i=1}^N$
8: **return** $\widetilde{\mathbf{Z}}^{(k)}$

Solving equation 17 exactly would require a costly inner optimization. Instead, RAMP performs a fixed number of projected unit-gradient ascent steps. Let K be the number of ascent steps and set $\alpha = \varepsilon/K$. For $s = 0, \ldots, K-1$, define

$$\mathbf{d}_i^{(s)} := \nabla_{\mathbf{z}} \psi_{\boldsymbol{\theta}^{(k)}}(\mathbf{z}_i^{(s)}), \qquad \mathbf{u}_i^{(s)} := \begin{cases} \mathbf{d}_i^{(s)}/\|\mathbf{d}_i^{(s)}\|, & \|\mathbf{d}_i^{(s)}\| > 0, \\ \mathbf{0}, & \|\mathbf{d}_i^{(s)}\| = 0, \end{cases} \tag{18}$$

where $\boldsymbol{\theta}^{(k)}$ is fixed when computing the input-space gradient. The candidate is updated by

$$\mathbf{z}_i^{(s+1)} = \Pi_\Omega\left(\mathbf{z}_i^{(s)} + \alpha \mathbf{u}_i^{(s)}\right), \qquad i = 1, \ldots, N. \tag{19}$$

The projection keeps the moved point inside the PDE domain and within radius ε of its initial candidate. As a result, RAMP has only two sampling hyperparameters: the radius bound ε and the number of ascent steps K. Analogously to the descent bound in Theorem 1, a smoothness argument gives a local ascent guarantee for the pointwise sensitivity. When the projection is inactive and $\psi_{\boldsymbol{\theta}}$ is $L_{\mathbf{z}}$-smooth in $\mathbf{z}$, one ascent step satisfies

$$\psi_{\boldsymbol{\theta}}(\mathbf{z}_i^{(s+1)}) \geq \psi_{\boldsymbol{\theta}}(\mathbf{z}_i^{(s)}) + \alpha \|\nabla_{\mathbf{z}} \psi_{\boldsymbol{\theta}}(\mathbf{z}_i^{(s)})\| - \frac{L_{\mathbf{z}} \alpha^2}{2}. \tag{20}$$

Hence, for a sufficiently small local step, the update increases the same pointwise quantity that appears in the parameter-sensitivity lower bounds.

The output of RAMP is the collocation set $\widetilde{\mathbf{Z}}^{(k)} = \{\mathbf{z}_i^{(K)}\}_{i=1}^N$. This set is used to evaluate the PDE residual loss, while the PINN parameters can be updated with the full training objective including boundary and initial condition losses. For example, the total objective at iteration k can be written as

$$\mathcal{L}_{\mathrm{total}}^{(k)}(\boldsymbol{\theta}) = \lambda_{\mathrm{pde}} \mathcal{L}(\widetilde{\mathbf{Z}}^{(k)}; \boldsymbol{\theta}) + \lambda_{\mathrm{ic}} \mathcal{L}_{\mathrm{IC}}(\mathbf{Z}_{\mathrm{IC}}^{(k)}; \boldsymbol{\theta}) + \lambda_{\mathrm{bc}} \mathcal{L}_{\mathrm{BC}}(\mathbf{Z}_{\mathrm{BC}}^{(k)}; \boldsymbol{\theta}), \tag{21}$$

and the parameter update is

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \eta^{(k)} \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{total}}^{(k)}(\boldsymbol{\theta}^{(k)}). \tag{22}$$

Here, RAMP modifies only the PDE collocation set $\widetilde{\mathbf{Z}}^{(k)}$; the boundary and initial condition points are handled by the underlying PINN training procedure. During the parameter update, all sampled points are treated as fixed inputs.

4 Evaluations

4.1 Parameter-Sensitivity Interpretation of Residual-Based Samplers

We reinterpret existing residual-based adaptive samplers as restricted surrogates of pointwise parameter-sensitivity sampling. This perspective identifies the assumptions under which residual magnitude is informative, while explaining why residual-only criteria can miss points that provide strong parameter-space training signals.

Residual-Only Approximation For brevity, write

$$r_{\boldsymbol{\theta}}(\mathbf{z}) := |R_{\boldsymbol{\theta}}(\mathbf{z})|, \qquad a_{\boldsymbol{\theta}}(\mathbf{z}) := \|\nabla_{\boldsymbol{\theta}} R_{\boldsymbol{\theta}}(\mathbf{z})\|, \qquad \psi_{\boldsymbol{\theta}}(\mathbf{z}) = r_{\boldsymbol{\theta}}(\mathbf{z}) a_{\boldsymbol{\theta}}(\mathbf{z}). \tag{23}$$

Existing magnitude-based samplers construct their adaptive scores from $r_{\boldsymbol{\theta}}(\mathbf{z})$ alone. Thus, they can be viewed as using a residual-only score $b(r_{\boldsymbol{\theta}}(\mathbf{z}))$, while pointwise parameter-sensitivity sampling uses the full score $r_{\boldsymbol{\theta}}(\mathbf{z})a_{\boldsymbol{\theta}}(\mathbf{z})$. For a nonnegative residual-only score $b(r_{\boldsymbol{\theta}}(\mathbf{z}))$, define

$$\frac{dP_b}{dQ}(\mathbf{z}) := \frac{b(r_{\boldsymbol{\theta}}(\mathbf{z}))}{\mathbb{E}_Q[b(r_{\boldsymbol{\theta}}(\mathbf{z}))]}. \tag{24}$$

Theorem 3 (Residual-only sampling as a restricted sensitivity surrogate)**.** *Fix $\boldsymbol{\theta}$, and write $\mathbf{g}_{\boldsymbol{\theta}}(\mathbf{z}) = \psi_{\boldsymbol{\theta}}(\mathbf{z})\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{z})$, where $\|\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{z})\| = 1$ whenever $\psi_{\boldsymbol{\theta}}(\mathbf{z}) > 0$, and $\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{z}) = \mathbf{0}$ otherwise. Let $b(r_{\boldsymbol{\theta}}) \geq 0$ satisfy $0 < \mathbb{E}_Q[b(r_{\boldsymbol{\theta}})], \mathbb{E}_Q[b(r_{\boldsymbol{\theta}})\psi_{\boldsymbol{\theta}}] < \infty$. Define*

$$\kappa_b(\boldsymbol{\theta}) := \frac{\|\mathbb{E}_Q[b(r_{\boldsymbol{\theta}}(\mathbf{z}))\,\psi_{\boldsymbol{\theta}}(\mathbf{z})\,\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{z})]\|}{\mathbb{E}_Q[b(r_{\boldsymbol{\theta}}(\mathbf{z}))\,\psi_{\boldsymbol{\theta}}(\mathbf{z})]} \in [0,1]. \tag{25}$$

Then the parameter sensitivity induced by P_b is

$$S_{P_b}(\boldsymbol{\theta}) = \kappa_b(\boldsymbol{\theta}) \frac{\mathbb{E}_Q[b(r_{\boldsymbol{\theta}}(\mathbf{z}))\,\psi_{\boldsymbol{\theta}}(\mathbf{z})]}{\mathbb{E}_Q[b(r_{\boldsymbol{\theta}}(\mathbf{z}))]}. \tag{26}$$

Moreover, among all square-integrable residual-only functions $h(r_{\boldsymbol{\theta}}(\mathbf{z}))$, the best $L^2(Q)$ approximation of $\psi_{\boldsymbol{\theta}}(\mathbf{z})$ is

$$h^*(r_{\boldsymbol{\theta}}(\mathbf{z})) = \mathbb{E}_Q[\psi_{\boldsymbol{\theta}}(\mathbf{z}) \mid r_{\boldsymbol{\theta}}(\mathbf{z})] = r_{\boldsymbol{\theta}}(\mathbf{z})\mathbb{E}_Q[a_{\boldsymbol{\theta}}(\mathbf{z}) \mid r_{\boldsymbol{\theta}}(\mathbf{z})]. \tag{27}$$

Theorem 3 shows that magnitude-based sampling is a residual-only restriction of pointwise parameter-sensitivity sampling. The restriction is exact only when $a_{\boldsymbol{\theta}}(\mathbf{z})$ is determined by $r_{\boldsymbol{\theta}}(\mathbf{z})$.

Residual Density Methods RAD constructs its adaptive sampling probabilities from a residual-dependent density (Wu et al., 2023). In our notation, its adaptive score is

$$b_{\mathrm{RAD}}(r_{\boldsymbol{\theta}}(\mathbf{z})) = \frac{r_{\boldsymbol{\theta}}(\mathbf{z})^k}{\mathbb{E}_Q[r_{\boldsymbol{\theta}}(\mathbf{z})^k]} + c, \qquad k \geq 0, \quad c \geq 0. \tag{28}$$

Therefore, by equation 26,

$$S_{P_{\mathrm{RAD}}}(\boldsymbol{\theta}) = \kappa_{\mathrm{RAD}}(\boldsymbol{\theta}) \frac{\mathbb{E}_Q[(r_{\boldsymbol{\theta}}(\mathbf{z})^k/\mathbb{E}_Q[r_{\boldsymbol{\theta}}(\mathbf{z})^k] + c)\psi_{\boldsymbol{\theta}}(\mathbf{z})]}{\mathbb{E}_Q[r_{\boldsymbol{\theta}}(\mathbf{z})^k/\mathbb{E}_Q[r_{\boldsymbol{\theta}}(\mathbf{z})^k] + c]}. \tag{29}$$

Since the full pointwise sensitivity is $r_{\boldsymbol{\theta}}(\mathbf{z})a_{\boldsymbol{\theta}}(\mathbf{z})$, a residual-only score $b(r_{\boldsymbol{\theta}}(\mathbf{z}))$ implicitly approximates $a_{\boldsymbol{\theta}}(\mathbf{z})$ by $b(r_{\boldsymbol{\theta}}(\mathbf{z}))/r_{\boldsymbol{\theta}}(\mathbf{z})$ on $r_{\boldsymbol{\theta}}(\mathbf{z}) > 0$. Thus RAD corresponds to

$$\widehat{a}_{\mathrm{RAD}}(\mathbf{z}) \propto \frac{1}{r_{\boldsymbol{\theta}}(\mathbf{z})} \left(\frac{r_{\boldsymbol{\theta}}(\mathbf{z})^k}{\mathbb{E}_Q[r_{\boldsymbol{\theta}}(\mathbf{z})^k]} + c \right), \qquad r_{\boldsymbol{\theta}}(\mathbf{z}) > 0. \tag{30}$$

When $c = 0$, this reduces to the residual-only assumption $a_{\boldsymbol{\theta}}(\mathbf{z}) \propto r_{\boldsymbol{\theta}}(\mathbf{z})^{k-1}$. Hence, RAD is a faithful sensitivity surrogate only when $a_{\boldsymbol{\theta}}(\mathbf{z})$ follows this residual-only relationship and has small variation within residual level sets.

LAS moves collocation points by Langevin dynamics driven by the input gradient of the squared residual, whose stationary density is proportional to $\exp(r_{\boldsymbol{\theta}}(\mathbf{z})^2/\beta_{\mathrm{LAS}}^2)$ for fixed $\boldsymbol{\theta}$ (Jeong et al., 2025). Thus, LAS corresponds to the residual-only score

$$b_{\mathrm{LAS}}(r_{\boldsymbol{\theta}}(\mathbf{z})) = \exp(r_{\boldsymbol{\theta}}(\mathbf{z})^2/\beta_{\mathrm{LAS}}^2), \qquad S_{P_{\mathrm{LAS}}}(\boldsymbol{\theta}) = \kappa_{\mathrm{LAS}}(\boldsymbol{\theta}) \frac{\mathbb{E}_Q[\exp(r_{\boldsymbol{\theta}}(\mathbf{z})^2/\beta_{\mathrm{LAS}}^2)\psi_{\boldsymbol{\theta}}(\mathbf{z})]}{\mathbb{E}_Q[\exp(r_{\boldsymbol{\theta}}(\mathbf{z})^2/\beta_{\mathrm{LAS}}^2)]}. \tag{31}$$

The Langevin noise can improve diversity and stability, but the stationary score is still a function of $r_{\boldsymbol{\theta}}(\mathbf{z})$ alone. Therefore, LAS also approximates the missing factor $a_{\boldsymbol{\theta}}(\mathbf{z})$ by a residual-only function.

Extreme Residual and Threshold Mixture Methods L^∞-based method also belongs to this extreme residual family, because its inner maximization moves points using input-space gradients of $r_{\boldsymbol{\theta}}(\mathbf{z})^2$, rather than gradients of $\psi_{\boldsymbol{\theta}}(\mathbf{z})$ (Wang et al., 2022a). Extreme residual methods are accurate sensitivity surrogates only when residual ranking agrees with pointwise sensitivity ranking:

$$r_{\boldsymbol{\theta}}(\mathbf{z}_i) \geq r_{\boldsymbol{\theta}}(\mathbf{z}_j) \quad \Longrightarrow \quad r_{\boldsymbol{\theta}}(\mathbf{z}_i)a_{\boldsymbol{\theta}}(\mathbf{z}_i) \geq r_{\boldsymbol{\theta}}(\mathbf{z}_j)a_{\boldsymbol{\theta}}(\mathbf{z}_j). \tag{32}$$

If this condition fails, a larger-residual point can have smaller pointwise parameter sensitivity than a smaller-residual point.

R3 retains points whose residual magnitudes exceed the average residual and replaces the remaining points with uniformly sampled points (Daw et al., 2023). A one-step population approximation from Q is obtained by defining

$$H := \{\mathbf{z} : r_{\boldsymbol{\theta}}(\mathbf{z}) > \mathbb{E}_Q[r_{\boldsymbol{\theta}}(\mathbf{z})]\}, \qquad q := Q(H), \qquad \frac{dP_{\mathrm{R3}}}{dQ}(\mathbf{z}) = \mathbf{1}\{\mathbf{z} \in H\} + 1 - q. \tag{33}$$

Thus,

$$S_{P_{\mathrm{R3}}}(\boldsymbol{\theta}) = \kappa_{\mathrm{R3}}(\boldsymbol{\theta})\mathbb{E}_Q[(\mathbf{1}\{\mathbf{z} \in H\} + 1 - q)\psi_{\boldsymbol{\theta}}(\mathbf{z})]. \tag{34}$$

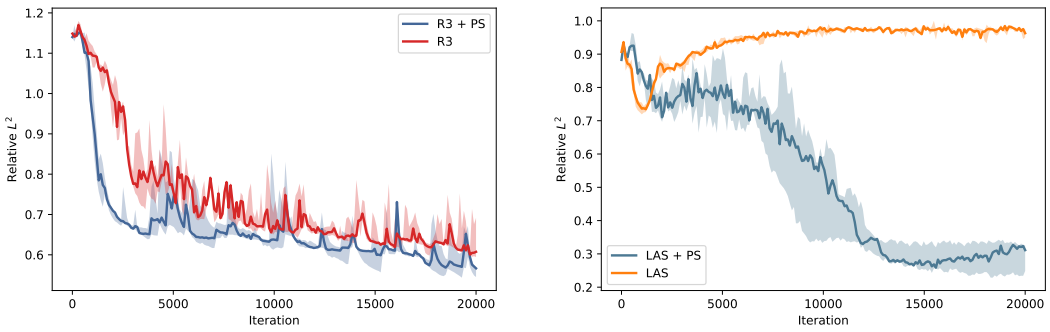
The uniform component preserves nonzero domain coverage, but the adaptive component is still a residual-only threshold. Consequently, R3 matches pointwise parameter sensitivity only when $r_{\boldsymbol{\theta}}(\mathbf{z})a_{\boldsymbol{\theta}}(\mathbf{z})$ is well represented by this threshold mixture score.

4.2 Empirical Evaluation on PDE Benchmarks

Comparison on Benchmark Equations We evaluate RAMP on the PINNacle benchmark (Zhongkai et al., 2024) using 12 PDEs. We report the mean rank (**MR**) across all PDEs to assess overall performance, where a lower value indicates better performance. In addition, we report the aggregate relative performance improvement over Rand-R, denoted by $\Delta\%$, to quantify the benefit of adaptive sampling across the benchmark. $\Delta\% = \left[1 - \prod_{i=1}^{12}\left(\frac{\bar{E}_{m,i}}{\bar{E}_{\text{Rand-R},i}}\right)^{\frac{1}{12}}\right] \times 100$, where i indexes the 12 PDEs, and $\bar{E}_{m,i}$ and $\bar{E}_{\text{Rand-R},i}$ denote the mean relative L^2 errors of sampling method m and Rand-R on PDE i, respectively. A larger $\Delta\%$ indicates a greater overall improvement over Rand-R. We compare RAMP with adaptive collocation sampling methods that dynamically adjust their sampling decisions to the current training state, including L^∞ (Wang et al., 2022a), R3 (Daw et al., 2023), RAD (Wu et al., 2023), and LAS (Jeong et al., 2025), as well as PINNACLE (Lau et al., 2024), which periodically updates its sampling decisions during training. For each baseline, we use the hyperparameter settings reported in the corresponding paper, as summarized in Appendix D.1. Unless otherwise stated, we follow the PINNacle experimental protocol while adopting a unified setup across all PDEs: a fully connected network with 3 hidden layers and 50 neurons per layer, $20{,}000$ training iterations, and $8{,}192$ PDE collocation points. We evaluate all sampling methods under three optimization settings, Adam (Kingma and Ba, 2015) with equal loss weights, loss-reweighted LRA (Wang et al., 2021), and gradient-adjusted HARMONIC (Bu et al., 2026), to compare their performance across different optimization strategies.

Table 1: Results on the PINNacle benchmark. Each experiment is repeated 5 times with different random seeds, and the results are reported as mean (standard deviation). The best scores are reported in **gray**.

Method	Adam		LRA		HARMONIC	
	MR ↓	$\Delta\%$ ↑	**MR** ↓	$\Delta\%$ ↑	**MR** ↓	$\Delta\%$ ↑
L^∞	3.54 (1.55)	-68.04 (3.42)	4.12 (1.63)	-89.45 (2.28)	4.33 (1.11)	-110.53 (3.42)
R3	4.17 (1.77)	-51.18 (1.84)	4.83 (1.40)	-73.34 (1.89)	4.83 (1.34)	-193.15 (3.05)
RAD	3.08 (1.50)	-9.85 (1.22)	3.83 (0.99)	-34.23 (1.64)	3.00 (0.82)	+5.73 (2.18)
LAS	3.58 (1.50)	-31.32 (1.95)	3.50 (1.66)	-48.16 (2.16)	3.17 (1.86)	-31.88 (2.22)
PINNACLE	3.79 (1.68)	-22.05 (1.28)	2.50 (1.54)	+0.20 (1.18)	3.46 (2.10)	-110.66 (5.14)
RAMP	**2.83 (1.86)**	**+6.32 (1.17)**	**2.21 (1.31)**	**+2.81 (1.37)**	**2.21 (1.18)**	**+19.57 (2.13)**



(a) R3 vs. R3+PS thresholding (b) LAS vs. LAS+PS drift

Figure 2: Replacing residual-based signals with pointwise parameter sensitivity (+PS) on Wave1d-C (left) and Heat2d-CG (right). Solid curves show medians over five runs; shaded bands show interquartile ranges.

Benchmark Results Table 1 shows that RAMP achieves the lowest **MR** under all optimization settings. RAMP also attains the largest improvement in $\Delta\%$, demonstrating that its sampling strategy provides substantial performance gains over Rand-R. These results suggest that parameter-sensitivity-guided refinement serves as an effective and competitive sampling criterion across PDEs that reflects the training state at each stage of optimization. The detailed relative L^2 errors for all methods on each of the 12 PDEs, together with the results of RAMP under different hyperparameter settings, are provided in Appendix D.7.

Pointwise Parameter Sensitivity as a Sampling Criterion We further investigate whether pointwise parameter sensitivity can serve as a replacement for residual magnitude in existing adaptive samplers. For R3, we retain or release points using pointwise parameter sensitivity (R3+PS), while for LAS, we replace the residual-gradient drift with the input-space gradient of pointwise parameter sensitivity (LAS+PS). Figure 2a shows that R3+PS produces a much steeper initial error decay, consistent with prioritizing points that induce stronger parameter-space training signals, and maintains this advantage throughout training. Figure 2b shows that replacing the residual-based drift with parameter sensitivity enables LAS+PS to sustain error reduction throughout training and converge to a stable regime. These controlled replacements indicate that pointwise parameter sensitivity provides a transferable signal of the evolving PINN training state for adaptive sampling.

5 Conclusions

In this work, we revisited adaptive collocation sampling for PINNs from a parameter-space perspective. While existing methods primarily rely on residual magnitude as a proxy for training utility, we showed that this criterion is incomplete, as it ignores how strongly each collocation point influences the model parameters. To address this limitation, we introduced pointwise parameter sensitivity, defined as the norm of the gradient of the pointwise PDE loss with respect to the parameters, as a direct measure of the induced training signal. Building on this criterion, we proposed RAMP, a simple and efficient refinement strategy that locally adjusts collocation points to increase parameter-space training signals while preserving the locality and coverage inherited from the base sampler. Our theoretical analysis clarified that residual-based methods can be interpreted as restricted approximations of parameter sensitivity, highlighting when and why they may fail. Empirically, we demonstrated that parameter sensitivity correlates better with actual training progress, and that RAMP consistently achieves competitive or improved performance across diverse PDE benchmarks. Overall, our results suggest that incorporating parameter-space information provides a principled and effective direction for adaptive sampling in PINNs, and we hope this perspective inspires further research on training dynamics-aware sampling strategies.

A current limitation is that RAMP measures the training state without explicitly resolving gradient conflicts among the losses; efficiently coupling it with recent PINN optimizers (Liu et al., 2025; Wang et al., 2026) is a promising direction. Also, RAMP does not explicitly model temporal causality, leaving integration with causality-aware PINN training as future work (Wang et al., 2024). Finally, RAMP refines points independently through local gradient ascent, so distributional refinements that exploit spatial coherence in sensitivity fields remain unexplored.

REFERENCES

Keith W Morton and David Francis Mayers. *Numerical solution of partial differential equations: an introduction*. Cambridge university press, 2005.

Ehsan Haghighat, Maziar Raissi, Adrian Moure, Hector Gomez, and Ruben Juanes. A physics-informed deep learning framework for inversion and surrogate modeling in solid mechanics. *Computer Methods in Applied Mechanics and Engineering*, 379:113741, 2021.

Yinhao Zhu, Nicholas Zabaras, Phaedon-Stelios Koutsourelakis, and Paris Perdikaris. Physics-constrained deep learning for high-dimensional surrogate modeling and uncertainty quantification without labeled data. *Journal of Computational Physics*, 394:56–81, 2019.

Dohyun Bu, Yujung Byun, and Jong-Seok Lee. Harmonized cone for feasible and non-conflict directions in training physics-informed neural networks. In *The Fourteenth International Conference on Learning Representations*, 2026.

Klaus-Jürgen Bathe. *Finite element procedures*. Klaus-Jurgen Bathe, 2006.

Christian Grossmann, Hans-Gorg Roos, and Martin Stynes. *Numerical Treatment of Partial Differential Equations: translated and revised by Martin Stynes*. Springer, 2007.

J Papež, J Liesen, and Z Strakoš. Distribution of the discretization and algebraic error in numerical solution of partial differential equations. *Linear Algebra and its Applications*, 449:89–114, 2014.

Maziar Raissi, Paris Perdikaris, and George 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.

Haixu Wu, Huakun Luo, Yuezhou Ma, Jianmin Wang, and Mingsheng Long. Ropinn: Region optimized physics-informed neural networks. *Advances in Neural Information Processing Systems*, 37:110494–110532, 2024.

Qian Fang, Xuankang Mou, and Shiben Li. A physics-informed neural network based on mixed data sampling for solving modified diffusion equations. *Scientific Reports*, 13(1):2491, 2023.

Lu Lu, Xuhui Meng, Zhiping Mao, and George Em Karniadakis. Deepxde: A deep learning library for solving differential equations. *SIAM review*, 63(1):208–228, 2021.

Chuwei Wang, Shanda Li, Di He, and Liwei Wang. Is L^2 physics informed loss always suitable for training physics informed neural network? *Advances in Neural Information Processing Systems*, 35:8278–8290, 2022a.

Mohammad Amin Nabian, Rini Jasmine Gladstone, and Hadi Meidani. Efficient training of physics-informed neural networks via importance sampling. *Computer-Aided Civil and Infrastructure Engineering*, 36(8):962–977, 2021.

Chenxi Wu, Min Zhu, Qinyang Tan, Yadhu Kartha, and Lu Lu. A comprehensive study of non-adaptive and residual-based adaptive sampling for physics-informed neural networks. *Computer Methods in Applied Mechanics and Engineering*, 403:115671, 2023.

Arka Daw, Jie Bu, Sifan Wang, Paris Perdikaris, and Anuj Karpatne. Mitigating propagation failures in physics-informed neural networks using retain-resample-release (r3) sampling. In *International Conference on Machine Learning*, pages 7264–7302. PMLR, 2023.

Minseok Jeong, Giup Seo, and Euiseok Hwang. Mitigating instability in high residual adaptive sampling for pinns via langevin dynamics. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Siyuan Duan, Wenyuan Wu, Peng Hu, Zhenwen Ren, Dezhong Peng, and Yuan Sun. Copinn: Cognitive physics-informed neural networks. In *Forty-second International Conference on Machine Learning*, 2025.

Bo Zhao, Nima Dehmamy, Robin Walters, and Rose Yu. Symmetry teleportation for accelerated optimization. *Advances in neural information processing systems*, 35:16679–16690, 2022.

Bo Zhao, Robert M. Gower, Robin Walters, and Rose Yu. Improving convergence and generalization using parameter symmetries. In *The Twelfth International Conference on Learning Representations*, 2024.

Angelos Katharopoulos and François Fleuret. Not all samples are created equal: Deep learning with importance sampling. In *International conference on machine learning*, pages 2525–2534. PMLR, 2018.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. *Advances in neural information processing systems*, 34: 20596–20607, 2021.

Sifan Wang, Xinling Yu, and Paris Perdikaris. When and why pinns fail to train: A neural tangent kernel perspective. *Journal of Computational Physics*, 449:110768, 2022b.

Gregory Kang Ruey Lau, Apivich Hemachandra, See-Kiong Ng, and Bryan Kian Hsiang Low. PINNACLE: PINN adaptive collocation and experimental points selection. In *The Twelfth International Conference on Learning Representations*, 2024.

Hao Zhongkai, Jiachen Yao, Chang Su, Hang Su, Ziao Wang, Fanzhi Lu, Zeyu Xia, Yichi Zhang, Songming Liu, Lu Lu, et al. Pinnacle: A comprehensive benchmark of physics-informed neural networks for solving pdes. *Advances in Neural Information Processing Systems*, 37:76721–76774, 2024.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *The Third International Conference on Learning Representations*, 2015.

Sifan Wang, Yujun Teng, and Paris Perdikaris. Understanding and mitigating gradient flow pathologies in physics-informed neural networks. *SIAM Journal on Scientific Computing*, 43(5):A3055–A3081, 2021.

Qiang Liu, Mengyu Chu, and Nils Thuerey. ConFIG: Towards conflict-free training of physics informed neural networks. In *The Thirteenth International Conference on Learning Representations*, 2025.

Sifan Wang, Ananyae Kumar bhartari, Bowen Li, and Paris Perdikaris. Gradient alignment in physics-informed neural networks: A second-order optimization perspective. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.

Sifan Wang, Shyam Sankaran, and Paris Perdikaris. Respecting causality for training physics-informed neural networks. *Computer Methods in Applied Mechanics and Engineering*, 421: 116813, 2024.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

A Additional Analysis for Section 1

In this section, we provide detailed information for Figure 1. For the 1D Burgers equation (Burgers1d), we use a fully connected neural network with three hidden layers of 50 neurons each, trained using the HARMONIC optimizer with a learning rate of 1×10^{-3}. We employ the `tanh` activation function with Glorot normal initialization (Glorot and Bengio, 2010). A total of 1024 collocation points are used for sampling. For the L^∞ method, we use 10 steps with a step size of 0.2. The results shown in Figure 1 correspond to the 71st training epoch. Additional snapshots at different training epochs are provided below.

Empirical Justification For the five frozen states shown in Figure 3, we evaluate 65,536 uniformly sampled Sobol points and compute

$$r(\mathbf{z}) = |R_{\boldsymbol{\theta}}(\mathbf{z})|, \qquad a(z) = \|\nabla_{\boldsymbol{\theta}} R_{\boldsymbol{\theta}}(\mathbf{z})\|, \qquad \psi(\mathbf{z}) = r(\mathbf{z})a(\mathbf{z}). \tag{35}$$

For the finite probe set Z, we measure

$$\kappa_Z = \frac{S_Z}{\bar{\psi}_Z}, \qquad S_Z = \left| \frac{1}{N} \sum_i R_i \nabla_{\boldsymbol{\theta}} R_i \right|. \tag{36}$$

We report the q_{95}/q_{05} spread of a, the overlap between the points ranked in the top 10% by r and by ψ, a two-fold cross-fitted residual-only approximation gap, obtained by fitting

$$\widehat{\psi}(r) = r\,\widehat{\mathbb{E}}\left[a \mid \mathrm{bin}(r)\right] \tag{37}$$

with 32 equal-frequency residual bins and reporting

$$100\sqrt{\frac{\mathbb{E}\left[(\psi - \widehat{\psi})^2\right]}{\mathbb{E}[\psi^2]}} \tag{38}$$

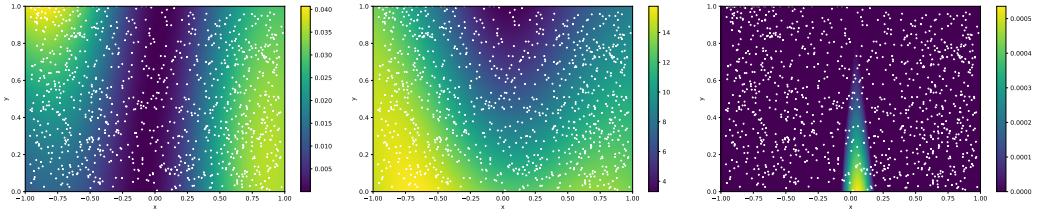
as a finite-bin proxy for the conditional-variance term, and the optimistic residual-bound tightness

$$T_Z^{\mathrm{opt}} = \frac{\kappa_Z c_Z \bar{r}_Z}{S_Z} = \frac{c_Z \bar{r}_Z}{\bar{\psi}_Z}, \qquad c_Z = \min_{z \in Z} a(z). \tag{39}$$

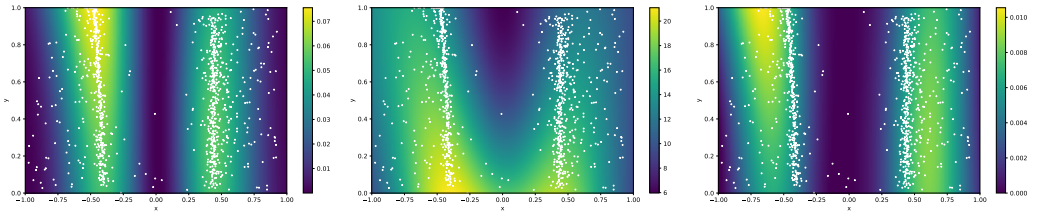
Table 2: Empirical diagnostics evaluated at five frozen Burgers1D training states.

Epoch	**Mean κ_Z**	$q_{95}(a)/q_{05}(a)$	**Top-10% overlap**	**Residual-only gap**	T_Z^{opt}
11	0.661	1.84	57.6%	17.8%	54.0%
41	0.638	1.50	42.0%	14.8%	66.7%
71	0.440	1.73	59.1%	17.9%	66.7%
101	0.488	1.56	65.1%	20.3%	70.9%
131	0.451	1.65	70.8%	19.8%	73.2%

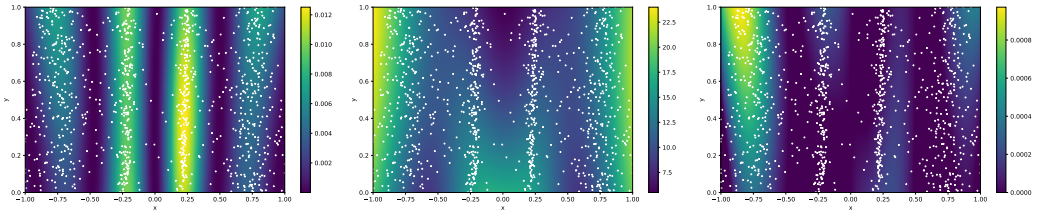
First, the base candidate sets have nontrivial directional coherence, with $\kappa_{\mathrm{base}} = 0.440$–$0.661$, where κ_{base} is computed from the 1,024 candidates before refinement as $\kappa_Z = \frac{\left\|\sum_i g_i\right\|}{\sum_i \|g_i\|}$. Using the same candidates and frozen parameter state, κ_{RAMP} denotes the coherence after one RAMP refinement step at the main radius $\varepsilon = 0.002$. Across 16 matched candidate sets per snapshot, the mean ratio $\kappa_{\mathrm{RAMP}}/\kappa_{\mathrm{base}} = 0.9990$–$1.0012$, showing that the local sensitivity ascent essentially preserves the original gradient alignment rather than increasing pointwise magnitudes at the expense of cancellation. Second, $a(z)$ varies across the domain, and a two-fold cross-fitted 32-bin residual-only predictor still leaves a discrepancy of 14.8%–20.3% in estimating $\psi(\mathbf{z}) = r(\mathbf{z})a(\mathbf{z})$. Although residual magnitude is globally informative, it does not preserve the high-signal set: the overlap between the top 10% of points ranked by $r(\mathbf{z})$ and the top 10% ranked by $\psi(\mathbf{z})$ is only 42.0%–70.8%, and changes substantially across training snapshots.



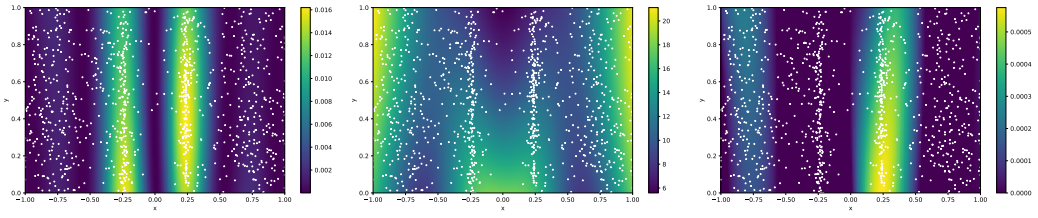
(a) Epoch 11



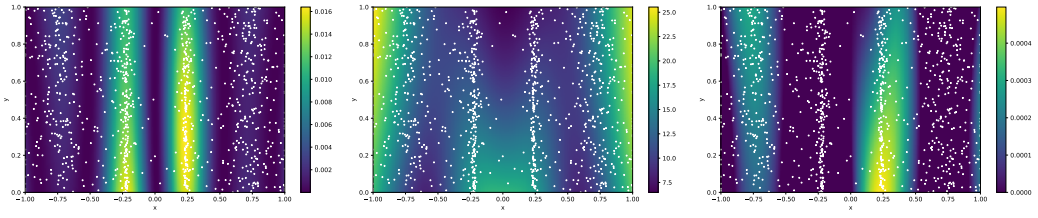
(b) Epoch 41



(c) Epoch 71



(d) Epoch 101



(e) Epoch 131

Figure 3: Snapshots at different training epochs for the 1D Burgers equation.

B Detail proof for Section 3

B.1 Detail proof for Theorem 1

Theorem 1 (One-step descent lower bound via parameter sensitivity)**.** *Assume that $\mathcal{L}_Q$ is β-smooth with respect to $\boldsymbol{\theta}$. For any $0 < \eta \leq 1/\beta$, let $\boldsymbol{\theta}^+ = \boldsymbol{\theta} - \eta \nabla_{\boldsymbol{\theta}} \mathcal{L}_Q(\boldsymbol{\theta})$. Then*

$$\mathcal{L}_Q(\boldsymbol{\theta}) - \mathcal{L}_Q(\boldsymbol{\theta}^+) \geq \frac{\eta}{2} S_Q(\boldsymbol{\theta})^2 \quad \textit{and} \quad \|\boldsymbol{\theta}^+ - \boldsymbol{\theta}\|_2 = \eta S_Q(\boldsymbol{\theta}). \tag{40}$$

Proof.
Let

$$\mathbf{G}_Q(\boldsymbol{\theta}) := \nabla_{\boldsymbol{\theta}} \mathcal{L}_Q(\boldsymbol{\theta}), \qquad S_Q(\boldsymbol{\theta}) := \|\mathbf{G}_Q(\boldsymbol{\theta})\|_2. \tag{41}$$

By β-smoothness, the gradient $\nabla_{\boldsymbol{\theta}} \mathcal{L}_Q$ is β-Lipschitz, and hence for any perturbation $\boldsymbol{\Delta} \in \mathbb{R}^d$,

$$\mathcal{L}_Q(\boldsymbol{\theta} + \boldsymbol{\Delta}) - \mathcal{L}_Q(\boldsymbol{\theta}) = \int_0^1 \langle \nabla_{\boldsymbol{\theta}} \mathcal{L}_Q(\boldsymbol{\theta} + t\boldsymbol{\Delta}), \boldsymbol{\Delta} \rangle \, dt. \tag{42}$$

Adding and subtracting $\nabla_{\boldsymbol{\theta}} \mathcal{L}_Q(\boldsymbol{\theta})$ inside the inner product gives

$$\begin{aligned} \mathcal{L}_Q(\boldsymbol{\theta} + \boldsymbol{\Delta}) - \mathcal{L}_Q(\boldsymbol{\theta}) &= \langle \nabla_{\boldsymbol{\theta}} \mathcal{L}_Q(\boldsymbol{\theta}), \boldsymbol{\Delta} \rangle \\ &\quad + \int_0^1 \langle \nabla_{\boldsymbol{\theta}} \mathcal{L}_Q(\boldsymbol{\theta} + t\boldsymbol{\Delta}) - \nabla_{\boldsymbol{\theta}} \mathcal{L}_Q(\boldsymbol{\theta}), \boldsymbol{\Delta} \rangle \, dt. \end{aligned} \tag{43}$$

Using the Cauchy–Schwarz inequality and the β-Lipschitz continuity of the gradient, the second term is bounded as

$$\begin{aligned} \int_0^1 \langle \nabla_{\boldsymbol{\theta}} \mathcal{L}_Q(\boldsymbol{\theta} + t\boldsymbol{\Delta}) - \nabla_{\boldsymbol{\theta}} \mathcal{L}_Q(\boldsymbol{\theta}), \boldsymbol{\Delta} \rangle \, dt &\leq \int_0^1 \|\nabla_{\boldsymbol{\theta}} \mathcal{L}_Q(\boldsymbol{\theta} + t\boldsymbol{\Delta}) - \nabla_{\boldsymbol{\theta}} \mathcal{L}_Q(\boldsymbol{\theta})\|_2 \, \|\boldsymbol{\Delta}\|_2 dt \\ &\leq \int_0^1 \beta t \|\boldsymbol{\Delta}\|_2^2 dt = \frac{\beta}{2} \|\boldsymbol{\Delta}\|_2^2. \end{aligned} \tag{44}$$

Therefore, for every $\boldsymbol{\Delta}$, we have the standard descent inequality

$$\mathcal{L}_Q(\boldsymbol{\theta} + \boldsymbol{\Delta}) \leq \mathcal{L}_Q(\boldsymbol{\theta}) + \langle \nabla_{\boldsymbol{\theta}} \mathcal{L}_Q(\boldsymbol{\theta}), \boldsymbol{\Delta} \rangle + \frac{\beta}{2} \|\boldsymbol{\Delta}\|_2^2. \tag{45}$$

Now set

$$\boldsymbol{\Delta} := \boldsymbol{\theta}^+ - \boldsymbol{\theta} = -\eta \nabla_{\boldsymbol{\theta}} \mathcal{L}_Q(\boldsymbol{\theta}) = -\eta \mathbf{G}_Q(\boldsymbol{\theta}). \tag{46}$$

Substituting this choice of $\boldsymbol{\Delta}$ into equation 45 yields

$$\begin{aligned} \mathcal{L}_Q(\boldsymbol{\theta}^+) &\leq \mathcal{L}_Q(\boldsymbol{\theta}) + \langle \mathbf{G}_Q(\boldsymbol{\theta}), -\eta \mathbf{G}_Q(\boldsymbol{\theta}) \rangle + \frac{\beta}{2} \|-\eta \mathbf{G}_Q(\boldsymbol{\theta})\|_2^2 \\ &= \mathcal{L}_Q(\boldsymbol{\theta}) - \eta \|\mathbf{G}_Q(\boldsymbol{\theta})\|_2^2 + \frac{\beta \eta^2}{2} \|\mathbf{G}_Q(\boldsymbol{\theta})\|_2^2 \\ &= \mathcal{L}_Q(\boldsymbol{\theta}) - \eta \left(1 - \frac{\beta \eta}{2}\right) \|\mathbf{G}_Q(\boldsymbol{\theta})\|_2^2. \end{aligned} \tag{47}$$

Rearranging the above inequality gives

$$\mathcal{L}_Q(\boldsymbol{\theta}) - \mathcal{L}_Q(\boldsymbol{\theta}^+) \geq \eta \left(1 - \frac{\beta \eta}{2}\right) \|\mathbf{G}_Q(\boldsymbol{\theta})\|_2^2. \tag{48}$$

Since $0 < \eta \leq 1/\beta$, we have $\beta \eta \leq 1$, and therefore

$$1 - \frac{\beta \eta}{2} \geq \frac{1}{2}. \tag{49}$$

Combining this with equation 48 and using $S_Q(\boldsymbol{\theta}) = \|\mathbf{G}_Q(\boldsymbol{\theta})\|_2$, we obtain

$$\mathcal{L}_Q(\boldsymbol{\theta}) - \mathcal{L}_Q(\boldsymbol{\theta}^+) \geq \frac{\eta}{2} S_Q(\boldsymbol{\theta})^2. \tag{50}$$

From the definition of $\boldsymbol{\theta}^+$,

$$\boldsymbol{\theta}^+ - \boldsymbol{\theta} = -\eta \mathbf{G}_Q(\boldsymbol{\theta}). \tag{51}$$

Taking the norm on both sides gives

$$\|\boldsymbol{\theta}^+ - \boldsymbol{\theta}\|_2 = \eta \|\mathbf{G}_Q(\boldsymbol{\theta})\|_2 = \eta S_Q(\boldsymbol{\theta}). \tag{52}$$

□

B.2 Detail proof for Theorem 2

Theorem 2 (Parameter sensitivity under sampling proportional to pointwise sensitivity)**.** *Fix $\boldsymbol{\theta}$, and write $\mathbf{g}_{\boldsymbol{\theta}}(\mathbf{z}) = \psi_{\boldsymbol{\theta}}(\mathbf{z})\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{z})$, where $\|\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{z})\| = 1$ whenever $\psi_{\boldsymbol{\theta}}(\mathbf{z}) > 0$, and $\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{z}) = \mathbf{0}$ otherwise. Assume $0 < \mathbb{E}_Q[\psi_{\boldsymbol{\theta}}(\mathbf{z})] < \infty$ and $\mathbb{E}_Q[\psi_{\boldsymbol{\theta}}(\mathbf{z})^2] < \infty$. Then*

$$S_{P_\psi}(\boldsymbol{\theta}) = \frac{\left\|\mathbb{E}_Q[\psi_{\boldsymbol{\theta}}(\mathbf{z})^2\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{z})]\right\|}{\mathbb{E}_Q[\psi_{\boldsymbol{\theta}}(\mathbf{z})]}. \tag{53}$$

Consequently, if for some $\rho \in [0,1]$,

$$\left\|\mathbb{E}_Q[\psi_{\boldsymbol{\theta}}(\mathbf{z})^2\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{z})]\right\| \geq \rho\, \mathbb{E}_Q[\psi_{\boldsymbol{\theta}}(\mathbf{z})^2],$$

then

$$S_{P_\psi}(\boldsymbol{\theta}) \geq \rho\, \frac{\mathbb{E}_Q[\psi_{\boldsymbol{\theta}}(\mathbf{z})^2]}{\mathbb{E}_Q[\psi_{\boldsymbol{\theta}}(\mathbf{z})]} \geq \rho\, \mathbb{E}_Q[\psi_{\boldsymbol{\theta}}(\mathbf{z})]. \tag{54}$$

Proof.
Fix $\boldsymbol{\theta}$ and define

$$M_1 := \mathbb{E}_{\mathbf{z}\sim Q}[\psi_{\boldsymbol{\theta}}(\mathbf{z})], \qquad M_2 := \mathbb{E}_{\mathbf{z}\sim Q}[\psi_{\boldsymbol{\theta}}(\mathbf{z})^2]. \tag{55}$$

By assumption, $0 < M_1 < \infty$ and $M_2 < \infty$. Since $\psi_{\boldsymbol{\theta}}(\mathbf{z}) = \|\mathbf{g}_{\boldsymbol{\theta}}(\mathbf{z})\| \geq 0$, the density

$$\frac{dP_\psi}{dQ}(\mathbf{z}) = \frac{\psi_{\boldsymbol{\theta}}(\mathbf{z})}{M_1} \tag{56}$$

is nonnegative and integrates to one:

$$\int_\Omega \frac{dP_\psi}{dQ}(\mathbf{z})dQ(\mathbf{z}) = \frac{1}{M_1}\mathbb{E}_{\mathbf{z}\sim Q}[\psi_{\boldsymbol{\theta}}(\mathbf{z})] = 1. \tag{57}$$

Therefore, P_ψ is a valid probability distribution on Ω.

Since

$$\mathbf{g}_{\boldsymbol{\theta}}(\mathbf{z}) = \psi_{\boldsymbol{\theta}}(\mathbf{z})\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{z}),$$

and $\|\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{z})\| \leq 1$ for all $\mathbf{z}$, we have

$$\mathbb{E}_{\mathbf{z}\sim P_\psi}[\|\mathbf{g}_{\boldsymbol{\theta}}(\mathbf{z})\|] = \mathbb{E}_{\mathbf{z}\sim Q}\left[\frac{\psi_{\boldsymbol{\theta}}(\mathbf{z})}{M_1}\|\mathbf{g}_{\boldsymbol{\theta}}(\mathbf{z})\|\right] = \frac{\mathbb{E}_{\mathbf{z}\sim Q}[\psi_{\boldsymbol{\theta}}(\mathbf{z})^2]}{M_1} = \frac{M_2}{M_1} < \infty. \tag{58}$$

Thus, $\mathbf{G}_{P_\psi}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z}\sim P_\psi}[\mathbf{g}_{\boldsymbol{\theta}}(\mathbf{z})]$ is well-defined.

Using the Radon–Nikodym derivative of P_ψ with respect to Q, we obtain

$$\begin{aligned}
\mathbf{G}_{P_\psi}(\boldsymbol{\theta}) &= \mathbb{E}_{\mathbf{z}\sim P_\psi}[\mathbf{g}_{\boldsymbol{\theta}}(\mathbf{z})] \\
&= \mathbb{E}_{\mathbf{z}\sim Q}\left[\frac{dP_\psi}{dQ}(\mathbf{z})\mathbf{g}_{\boldsymbol{\theta}}(\mathbf{z})\right] \\
&= \mathbb{E}_{\mathbf{z}\sim Q}\left[\frac{\psi_{\boldsymbol{\theta}}(\mathbf{z})}{M_1}\mathbf{g}_{\boldsymbol{\theta}}(\mathbf{z})\right].
\end{aligned} \tag{59}$$

Substituting $\mathbf{g}_{\boldsymbol{\theta}}(\mathbf{z}) = \psi_{\boldsymbol{\theta}}(\mathbf{z})\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{z})$ gives

$$\begin{aligned}
\mathbf{G}_{P_\psi}(\boldsymbol{\theta}) &= \frac{1}{M_1}\mathbb{E}_{\mathbf{z}\sim Q}[\psi_{\boldsymbol{\theta}}(\mathbf{z})\psi_{\boldsymbol{\theta}}(\mathbf{z})\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{z})] \\
&= \frac{\mathbb{E}_{\mathbf{z}\sim Q}[\psi_{\boldsymbol{\theta}}(\mathbf{z})^2\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{z})]}{\mathbb{E}_{\mathbf{z}\sim Q}[\psi_{\boldsymbol{\theta}}(\mathbf{z})]}.
\end{aligned} \tag{60}$$

Taking the norm of both sides yields

$$S_{P_\psi}(\boldsymbol{\theta}) = \|\mathbf{G}_{P_\psi}(\boldsymbol{\theta})\| = \frac{\left\|\mathbb{E}_{\mathbf{z}\sim Q}[\psi_{\boldsymbol{\theta}}(\mathbf{z})^2\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{z})]\right\|}{\mathbb{E}_{\mathbf{z}\sim Q}[\psi_{\boldsymbol{\theta}}(\mathbf{z})]}. \tag{61}$$

Assume that for some $\rho \in [0,1]$,

$$\left\|\mathbb{E}_{\mathbf{z}\sim Q}[\psi_{\boldsymbol{\theta}}(\mathbf{z})^2\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{z})]\right\| \geq \rho\mathbb{E}_{\mathbf{z}\sim Q}[\psi_{\boldsymbol{\theta}}(\mathbf{z})^2]. \tag{62}$$

Substituting this assumption into the identity above gives

$$S_{P_\psi}(\boldsymbol{\theta}) \geq \rho \frac{\mathbb{E}_{\mathbf{z}\sim Q}[\psi_{\boldsymbol{\theta}}(\mathbf{z})^2]}{\mathbb{E}_{\mathbf{z}\sim Q}[\psi_{\boldsymbol{\theta}}(\mathbf{z})]}. \tag{63}$$

It remains to show that the second-to-first moment ratio is at least the first moment. Since $\psi_{\boldsymbol{\theta}}(\mathbf{z}) \geq 0$ and $\mathbb{E}_Q[\psi_{\boldsymbol{\theta}}(\mathbf{z})^2] < \infty$, Jensen's inequality, or equivalently the nonnegativity of variance, gives

$$\mathbb{E}_{\mathbf{z}\sim Q}[\psi_{\boldsymbol{\theta}}(\mathbf{z})^2] \geq \left(\mathbb{E}_{\mathbf{z}\sim Q}[\psi_{\boldsymbol{\theta}}(\mathbf{z})]\right)^2. \tag{64}$$

Because $\mathbb{E}_{\mathbf{z}\sim Q}[\psi_{\boldsymbol{\theta}}(\mathbf{z})] > 0$, we can divide both sides by $\mathbb{E}_{\mathbf{z}\sim Q}[\psi_{\boldsymbol{\theta}}(\mathbf{z})]$ to obtain

$$\frac{\mathbb{E}_{\mathbf{z}\sim Q}[\psi_{\boldsymbol{\theta}}(\mathbf{z})^2]}{\mathbb{E}_{\mathbf{z}\sim Q}[\psi_{\boldsymbol{\theta}}(\mathbf{z})]} \geq \mathbb{E}_{\mathbf{z}\sim Q}[\psi_{\boldsymbol{\theta}}(\mathbf{z})]. \tag{65}$$

Combining the two inequalities yields

$$S_{P_\psi}(\boldsymbol{\theta}) \geq \rho \frac{\mathbb{E}_{\mathbf{z}\sim Q}[\psi_{\boldsymbol{\theta}}(\mathbf{z})^2]}{\mathbb{E}_{\mathbf{z}\sim Q}[\psi_{\boldsymbol{\theta}}(\mathbf{z})]} \geq \rho \mathbb{E}_{\mathbf{z}\sim Q}[\psi_{\boldsymbol{\theta}}(\mathbf{z})]. \tag{66}$$

Finally, the coherence assumption is naturally bounded. Indeed, by the triangle inequality and $\|\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{z})\| \leq 1$,

$$\left\|\mathbb{E}_{\mathbf{z}\sim Q}[\psi_{\boldsymbol{\theta}}(\mathbf{z})^2 \mathbf{v}_{\boldsymbol{\theta}}(\mathbf{z})]\right\| \leq \mathbb{E}_{\mathbf{z}\sim Q}[\psi_{\boldsymbol{\theta}}(\mathbf{z})^2 \|\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{z})\|] \leq \mathbb{E}_{\mathbf{z}\sim Q}[\psi_{\boldsymbol{\theta}}(\mathbf{z})^2]. \tag{67}$$

Thus the ratio

$$\frac{\left\|\mathbb{E}_{\mathbf{z}\sim Q}[\psi_{\boldsymbol{\theta}}(\mathbf{z})^2 \mathbf{v}_{\boldsymbol{\theta}}(\mathbf{z})]\right\|}{\mathbb{E}_{\mathbf{z}\sim Q}[\psi_{\boldsymbol{\theta}}(\mathbf{z})^2]}$$

lies in $[0, 1]$ whenever $\mathbb{E}_Q[\psi_{\boldsymbol{\theta}}(\mathbf{z})^2] > 0$. □

C Detail proof for Section 4.1

C.1 Detail proof for Theorem 3

Theorem 3 (Residual-only sampling as a restricted sensitivity surrogate)**.** *Fix $\boldsymbol{\theta}$, and write $\mathbf{g}_{\boldsymbol{\theta}}(\mathbf{z}) = \psi_{\boldsymbol{\theta}}(\mathbf{z})\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{z})$, where $\|\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{z})\| = 1$ whenever $\psi_{\boldsymbol{\theta}}(\mathbf{z}) > 0$, and $\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{z}) = \mathbf{0}$ otherwise. Let $b(r_{\boldsymbol{\theta}}) \geq 0$ satisfy $0 < \mathbb{E}_Q[b(r_{\boldsymbol{\theta}})], \mathbb{E}_Q[b(r_{\boldsymbol{\theta}})\psi_{\boldsymbol{\theta}}] < \infty$. Define*

$$\kappa_b(\boldsymbol{\theta}) := \frac{\|\mathbb{E}_Q[b(r_{\boldsymbol{\theta}}(\mathbf{z}))\,\psi_{\boldsymbol{\theta}}(\mathbf{z})\,\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{z})]\|}{\mathbb{E}_Q[b(r_{\boldsymbol{\theta}}(\mathbf{z}))\,\psi_{\boldsymbol{\theta}}(\mathbf{z})]} \in [0,1]. \tag{68}$$

Then the parameter sensitivity induced by P_b is

$$S_{P_b}(\boldsymbol{\theta}) = \kappa_b(\boldsymbol{\theta})\frac{\mathbb{E}_Q[b(r_{\boldsymbol{\theta}}(\mathbf{z}))\,\psi_{\boldsymbol{\theta}}(\mathbf{z})]}{\mathbb{E}_Q[b(r_{\boldsymbol{\theta}}(\mathbf{z}))]}. \tag{69}$$

Moreover, among all square-integrable residual-only functions $h(r_{\boldsymbol{\theta}}(\mathbf{z}))$, the best $L^2(Q)$ approximation of $\psi_{\boldsymbol{\theta}}(\mathbf{z})$ is

$$h^*(r_{\boldsymbol{\theta}}(\mathbf{z})) = \mathbb{E}_Q[\psi_{\boldsymbol{\theta}}(\mathbf{z}) \mid r_{\boldsymbol{\theta}}(\mathbf{z})] = r_{\boldsymbol{\theta}}(\mathbf{z})\mathbb{E}_Q[a_{\boldsymbol{\theta}}(\mathbf{z}) \mid r_{\boldsymbol{\theta}}(\mathbf{z})]. \tag{70}$$

Proof.
Fix $\boldsymbol{\theta}$ and define

$$b_{\boldsymbol{\theta}}(\mathbf{z}) := b(r_{\boldsymbol{\theta}}(\mathbf{z})), \qquad Z_b := \mathbb{E}_Q[b_{\boldsymbol{\theta}}(\mathbf{z})]. \tag{71}$$

By assumption, $0 < Z_b < \infty$. Since $b_{\boldsymbol{\theta}}(\mathbf{z}) \geq 0$, the density

$$\frac{dP_b}{dQ}(\mathbf{z}) = \frac{b_{\boldsymbol{\theta}}(\mathbf{z})}{Z_b}$$

is nonnegative and satisfies

$$\int_\Omega \frac{dP_b}{dQ}(\mathbf{z})dQ(\mathbf{z}) = \frac{\mathbb{E}_Q[b_{\boldsymbol{\theta}}(\mathbf{z})]}{Z_b} = 1. \tag{72}$$

Therefore, P_b is a valid probability distribution on Ω.

Since

$$\mathbf{g}_{\boldsymbol{\theta}}(\mathbf{z}) = \psi_{\boldsymbol{\theta}}(\mathbf{z})\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{z}),$$

we have

$$\|\mathbf{g}_{\boldsymbol{\theta}}(\mathbf{z})\| = \psi_{\boldsymbol{\theta}}(\mathbf{z}), \qquad \|\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{z})\| \leq 1.$$

Moreover, by the assumption

$$\mathbb{E}_Q[b_{\boldsymbol{\theta}}(\mathbf{z})\psi_{\boldsymbol{\theta}}(\mathbf{z})] < \infty,$$

the parameter gradient is integrable under P_b, because

$$\mathbb{E}_{P_b}[\|\mathbf{g}_{\boldsymbol{\theta}}(\mathbf{z})\|] = \mathbb{E}_Q\left[\frac{b_{\boldsymbol{\theta}}(\mathbf{z})}{Z_b}\psi_{\boldsymbol{\theta}}(\mathbf{z})\right] = \frac{\mathbb{E}_Q[b_{\boldsymbol{\theta}}(\mathbf{z})\psi_{\boldsymbol{\theta}}(\mathbf{z})]}{\mathbb{E}_Q[b_{\boldsymbol{\theta}}(\mathbf{z})]} < \infty. \tag{73}$$

Thus, $\mathbf{G}_{P_b}(\boldsymbol{\theta}) = \mathbb{E}_{P_b}[\mathbf{g}_{\boldsymbol{\theta}}(\mathbf{z})]$ is well-defined.

Using the Radon–Nikodym derivative of P_b with respect to Q, we obtain

$$\begin{aligned}
\mathbf{G}_{P_b}(\boldsymbol{\theta}) &= \mathbb{E}_{P_b}[\mathbf{g}_{\boldsymbol{\theta}}(\mathbf{z})] \\
&= \mathbb{E}_Q\left[\frac{dP_b}{dQ}(\mathbf{z})\mathbf{g}_{\boldsymbol{\theta}}(\mathbf{z})\right] \\
&= \frac{\mathbb{E}_Q[b_{\boldsymbol{\theta}}(\mathbf{z})\mathbf{g}_{\boldsymbol{\theta}}(\mathbf{z})]}{\mathbb{E}_Q[b_{\boldsymbol{\theta}}(\mathbf{z})]} \\
&= \frac{\mathbb{E}_Q[b_{\boldsymbol{\theta}}(\mathbf{z})\psi_{\boldsymbol{\theta}}(\mathbf{z})\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{z})]}{\mathbb{E}_Q[b_{\boldsymbol{\theta}}(\mathbf{z})]}.
\end{aligned} \tag{74}$$

Taking the norm of both sides gives

$$S_{P_b}(\boldsymbol{\theta}) = \|\mathbf{G}_{P_b}(\boldsymbol{\theta})\| = \frac{\|\mathbb{E}_Q[b_{\boldsymbol{\theta}}(\mathbf{z})\psi_{\boldsymbol{\theta}}(\mathbf{z})\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{z})]\|}{\mathbb{E}_Q[b_{\boldsymbol{\theta}}(\mathbf{z})]}. \tag{75}$$

If

$$\mathbb{E}_Q[b_{\boldsymbol{\theta}}(\mathbf{z})\psi_{\boldsymbol{\theta}}(\mathbf{z})] = 0,$$

then $b_{\boldsymbol{\theta}}(\mathbf{z})\psi_{\boldsymbol{\theta}}(\mathbf{z}) = 0$ (Q-almost surely), because the integrand is nonnegative. Hence the numerator in equation 75 is zero, and $S_{P_b}(\boldsymbol{\theta}) = 0$. In this degenerate case, the stated identity holds by setting $\kappa_b(\boldsymbol{\theta}) = 0$.

Otherwise, assume

$$\mathbb{E}_Q[b_{\boldsymbol{\theta}}(\mathbf{z})\psi_{\boldsymbol{\theta}}(\mathbf{z})] > 0.$$

Then

$$\kappa_b(\boldsymbol{\theta}) := \frac{\|\mathbb{E}_Q[b_{\boldsymbol{\theta}}(\mathbf{z})\psi_{\boldsymbol{\theta}}(\mathbf{z})\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{z})]\|}{\mathbb{E}_Q[b_{\boldsymbol{\theta}}(\mathbf{z})\psi_{\boldsymbol{\theta}}(\mathbf{z})]}$$

is well-defined. It is nonnegative by definition. Moreover, by the triangle inequality and $\|\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{z})\| \leq 1$,

$$\begin{aligned}\|\mathbb{E}_Q[b_{\boldsymbol{\theta}}(\mathbf{z})\psi_{\boldsymbol{\theta}}(\mathbf{z})\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{z})]\| &\leq \mathbb{E}_Q[b_{\boldsymbol{\theta}}(\mathbf{z})\psi_{\boldsymbol{\theta}}(\mathbf{z})\|\mathbf{v}_{\boldsymbol{\theta}}(\mathbf{z})\|] \\ &\leq \mathbb{E}_Q[b_{\boldsymbol{\theta}}(\mathbf{z})\psi_{\boldsymbol{\theta}}(\mathbf{z})]. \end{aligned} \tag{76}$$

Therefore, $\kappa_b(\boldsymbol{\theta}) \in [0, 1]$. Multiplying and dividing equation 75 by $\mathbb{E}_Q[b_{\boldsymbol{\theta}}(\mathbf{z})\psi_{\boldsymbol{\theta}}(\mathbf{z})]$ yields

$$S_{P_b}(\boldsymbol{\theta}) = \kappa_b(\boldsymbol{\theta}) \frac{\mathbb{E}_Q[b_{\boldsymbol{\theta}}(\mathbf{z})\psi_{\boldsymbol{\theta}}(\mathbf{z})]}{\mathbb{E}_Q[b_{\boldsymbol{\theta}}(\mathbf{z})]}. \tag{77}$$

Assume that

$$\mathbb{E}_Q[\psi_{\boldsymbol{\theta}}(\mathbf{z})^2] < \infty,$$

so that $\psi_{\boldsymbol{\theta}}(\mathbf{z}) \in L^2(Q)$. Let

$$\mathcal{H}_r := \{h(r_{\boldsymbol{\theta}}(\mathbf{z})) : h(r_{\boldsymbol{\theta}}(\mathbf{z})) \in L^2(Q)\}$$

be the set of square-integrable residual-only functions.

Define

$$m_{\boldsymbol{\theta}}(\mathbf{z}) := \mathbb{E}_Q[\psi_{\boldsymbol{\theta}}(\mathbf{z}) \mid r_{\boldsymbol{\theta}}(\mathbf{z})]. \tag{78}$$

By the property of conditional expectation, $m_{\boldsymbol{\theta}}(\mathbf{z}) \in \mathcal{H}_r$, and for every $h(r_{\boldsymbol{\theta}}(\mathbf{z})) \in \mathcal{H}_r$,

$$\mathbb{E}_Q\left[(\psi_{\boldsymbol{\theta}}(\mathbf{z}) - m_{\boldsymbol{\theta}}(\mathbf{z}))\, h(r_{\boldsymbol{\theta}}(\mathbf{z}))\right] = 0. \tag{79}$$

That is, the residual $\psi_{\boldsymbol{\theta}}(\mathbf{z}) - m_{\boldsymbol{\theta}}(\mathbf{z})$ is orthogonal in $L^2(Q)$ to every residual-only function.

Now take any $h(r_{\boldsymbol{\theta}}(\mathbf{z})) \in \mathcal{H}_r$. We decompose the squared error as

$$\begin{aligned}&\mathbb{E}_Q\left[(\psi_{\boldsymbol{\theta}}(\mathbf{z}) - h(r_{\boldsymbol{\theta}}(\mathbf{z})))^2\right] \\ &= \mathbb{E}_Q\left[(\psi_{\boldsymbol{\theta}}(\mathbf{z}) - m_{\boldsymbol{\theta}}(\mathbf{z}) + m_{\boldsymbol{\theta}}(\mathbf{z}) - h(r_{\boldsymbol{\theta}}(\mathbf{z})))^2\right] \\ &= \mathbb{E}_Q\left[(\psi_{\boldsymbol{\theta}}(\mathbf{z}) - m_{\boldsymbol{\theta}}(\mathbf{z}))^2\right] + \mathbb{E}_Q\left[(m_{\boldsymbol{\theta}}(\mathbf{z}) - h(r_{\boldsymbol{\theta}}(\mathbf{z})))^2\right] \\ &\quad + 2\mathbb{E}_Q\left[(\psi_{\boldsymbol{\theta}}(\mathbf{z}) - m_{\boldsymbol{\theta}}(\mathbf{z}))\,(m_{\boldsymbol{\theta}}(\mathbf{z}) - h(r_{\boldsymbol{\theta}}(\mathbf{z})))\right]. \end{aligned} \tag{80}$$

The last term is zero by equation 79, because $m_{\boldsymbol{\theta}}(\mathbf{z}) - h(r_{\boldsymbol{\theta}}(\mathbf{z})) \in \mathcal{H}_r$. Therefore,

$$\mathbb{E}_Q\left[(\psi_{\boldsymbol{\theta}}(\mathbf{z}) - h(r_{\boldsymbol{\theta}}(\mathbf{z})))^2\right] = \mathbb{E}_Q\left[(\psi_{\boldsymbol{\theta}}(\mathbf{z}) - m_{\boldsymbol{\theta}}(\mathbf{z}))^2\right] + \mathbb{E}_Q\left[(m_{\boldsymbol{\theta}}(\mathbf{z}) - h(r_{\boldsymbol{\theta}}(\mathbf{z})))^2\right]. \tag{81}$$

The second term on the right-hand side is nonnegative and is minimized if and only if

$$h(r_{\boldsymbol{\theta}}(\mathbf{z})) = m_{\boldsymbol{\theta}}(\mathbf{z}) \quad Q\text{-almost surely}.$$

Thus the unique best residual-only $L^2(Q)$ approximation, up to Q-almost sure equivalence, is

$$h^*(r_{\boldsymbol{\theta}}(\mathbf{z})) = \mathbb{E}_Q[\psi_{\boldsymbol{\theta}}(\mathbf{z}) \mid r_{\boldsymbol{\theta}}(\mathbf{z})]. \tag{82}$$

It remains to express this conditional expectation in terms of $a_{\boldsymbol{\theta}}(\mathbf{z})$. Since

$$\psi_{\boldsymbol{\theta}}(\mathbf{z}) = r_{\boldsymbol{\theta}}(\mathbf{z})a_{\boldsymbol{\theta}}(\mathbf{z}),$$

and $r_{\boldsymbol{\theta}}(\mathbf{z})$ is measurable with respect to $\sigma(r_{\boldsymbol{\theta}})$, we can pull it outside the conditional expectation:

$$\begin{aligned} h^*(r_{\boldsymbol{\theta}}(\mathbf{z})) &= \mathbb{E}_Q[r_{\boldsymbol{\theta}}(\mathbf{z})a_{\boldsymbol{\theta}}(\mathbf{z}) \mid r_{\boldsymbol{\theta}}(\mathbf{z})] \\ &= r_{\boldsymbol{\theta}}(\mathbf{z})\mathbb{E}_Q[a_{\boldsymbol{\theta}}(\mathbf{z}) \mid r_{\boldsymbol{\theta}}(\mathbf{z})]. \end{aligned} \tag{83}$$

Finally, we compute the irreducible approximation error. Taking the infimum over h in equation 81 gives

$$\begin{aligned} \inf_h \mathbb{E}_Q\left[\left(\psi_{\boldsymbol{\theta}}(\mathbf{z}) - h(r_{\boldsymbol{\theta}}(\mathbf{z}))\right)^2\right] &= \mathbb{E}_Q\left[\left(\psi_{\boldsymbol{\theta}}(\mathbf{z}) - \mathbb{E}_Q[\psi_{\boldsymbol{\theta}}(\mathbf{z}) \mid r_{\boldsymbol{\theta}}(\mathbf{z})]\right)^2\right] \\ &= \mathbb{E}_Q\left[\mathrm{Var}_Q\left(\psi_{\boldsymbol{\theta}}(\mathbf{z}) \mid r_{\boldsymbol{\theta}}(\mathbf{z})\right)\right]. \end{aligned} \tag{84}$$

Using again $\psi_{\boldsymbol{\theta}}(\mathbf{z}) = r_{\boldsymbol{\theta}}(\mathbf{z})a_{\boldsymbol{\theta}}(\mathbf{z})$, and noting that $r_{\boldsymbol{\theta}}(\mathbf{z})$ is fixed after conditioning on itself, we obtain

$$\begin{aligned} \mathrm{Var}_Q\left(\psi_{\boldsymbol{\theta}}(\mathbf{z}) \mid r_{\boldsymbol{\theta}}(\mathbf{z})\right) &= \mathrm{Var}_Q\left(r_{\boldsymbol{\theta}}(\mathbf{z})a_{\boldsymbol{\theta}}(\mathbf{z}) \mid r_{\boldsymbol{\theta}}(\mathbf{z})\right) \\ &= r_{\boldsymbol{\theta}}(\mathbf{z})^2\,\mathrm{Var}_Q\left(a_{\boldsymbol{\theta}}(\mathbf{z}) \mid r_{\boldsymbol{\theta}}(\mathbf{z})\right). \end{aligned} \tag{85}$$

Substituting this identity into equation 84 yields

$$\inf_h \mathbb{E}_Q\left[\left(\psi_{\boldsymbol{\theta}}(\mathbf{z}) - h(r_{\boldsymbol{\theta}}(\mathbf{z}))\right)^2\right] = \mathbb{E}_Q\left[r_{\boldsymbol{\theta}}(\mathbf{z})^2\,\mathrm{Var}_Q\left(a_{\boldsymbol{\theta}}(\mathbf{z}) \mid r_{\boldsymbol{\theta}}(\mathbf{z})\right)\right]. \tag{86}$$

Equation equation 86 also characterizes when residual-only approximation is exact. The error is zero if and only if

$$r_{\boldsymbol{\theta}}(\mathbf{z})^2\,\mathrm{Var}_Q\left(a_{\boldsymbol{\theta}}(\mathbf{z}) \mid r_{\boldsymbol{\theta}}(\mathbf{z})\right) = 0 \quad Q\text{-almost surely.}$$

Equivalently, on the set where $r_{\boldsymbol{\theta}}(\mathbf{z}) > 0$, the residual parameter-gradient norm $a_{\boldsymbol{\theta}}(\mathbf{z})$ must be determined by $r_{\boldsymbol{\theta}}(\mathbf{z})$ up to Q-almost sure equivalence. On the set where $r_{\boldsymbol{\theta}}(\mathbf{z}) = 0$, variation in $a_{\boldsymbol{\theta}}(\mathbf{z})$ does not affect $\psi_{\boldsymbol{\theta}}(\mathbf{z})$, because $\psi_{\boldsymbol{\theta}}(\mathbf{z}) = 0$ there. □

D Implementation Details of Section 4

D.1 Hyperparameter Settings

The hyperparameter settings used in Table 1 are summarized. These configurations follow the default settings reported in the respective papers.

Table 3: Hyperparameter settings of baseline methods.

Method	Hyperparameters
L^∞	$K = 20,\ \eta = 0.05$
RAD	$k = 1,\ c = 1$
LAS	$l_L = 1,\ \beta = 0.2,\ \tau = 0.002$
RAMP	$K = 1,\ \varepsilon = \{2 \times 10^{-5}, 2 \times 10^{-4}, 2 \times 10^{-3}\}$

D.2 Burgers' equations

1D Burgers Equation (Burgers1d)

The PDE is given by

$$u_t + u\, u_x = \nu\, u_{xx}. \tag{87}$$

The domain is defined as

$$(x, t) \in \Omega \times T = [-1, 1] \times [0, 1].$$

The initial and boundary conditions are

$$u(x, 0) = -\sin(\pi x), \tag{88}$$

$$u(-1, t) = u(1, t) = 0. \tag{89}$$

The parameter is

$$\nu = \frac{0.01}{\pi}. \tag{90}$$

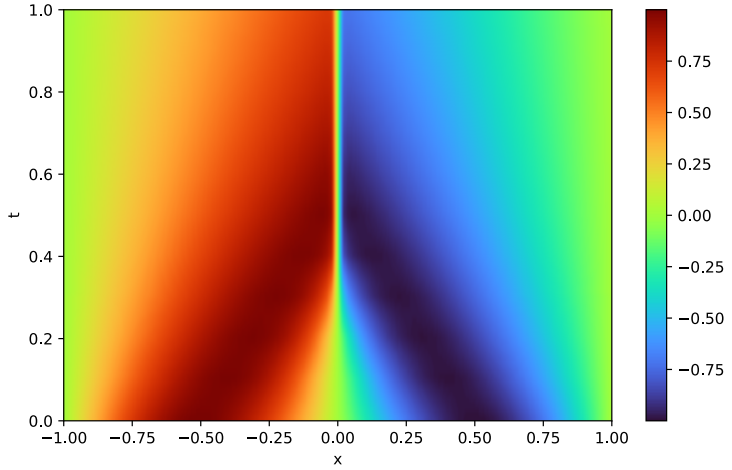


Figure 4: Reference solution of Burgers1d using PINNacle reference data.

D.3 Poisson equations

Poisson 2D Classic (Poisson2d-C)

The PDE is given by

$$u_{xx} + u_{yy} = 0. \tag{91}$$

The domain is a rectangle minus four circles $(x, y) \in \Omega = \Omega_{\text{rec}} \setminus R_i$, where $\Omega_{\text{rec}} = [-0.5, 0.5]^2$ is the rectangle and R_i denotes four circle areas:

$$R_1 = \{(x, y) : (x - 0.3)^2 + (y - 0.3)^2 \leq 0.1^2\}, \tag{92}$$
$$R_2 = \{(x, y) : (x + 0.3)^2 + (y - 0.3)^2 \leq 0.1^2\}, \tag{93}$$
$$R_3 = \{(x, y) : (x - 0.3)^2 + (y + 0.3)^2 \leq 0.1^2\}, \tag{94}$$
$$R_4 = \{(x, y) : (x + 0.3)^2 + (y + 0.3)^2 \leq 0.1^2\}. \tag{95}$$

The boundary condition is

$$u(x, y) = 0, (x, y) \in \partial R_i, \tag{96}$$
$$u(x, y) = 1, (x, y) \in \partial\Omega_{\text{rec}}. \tag{97}$$

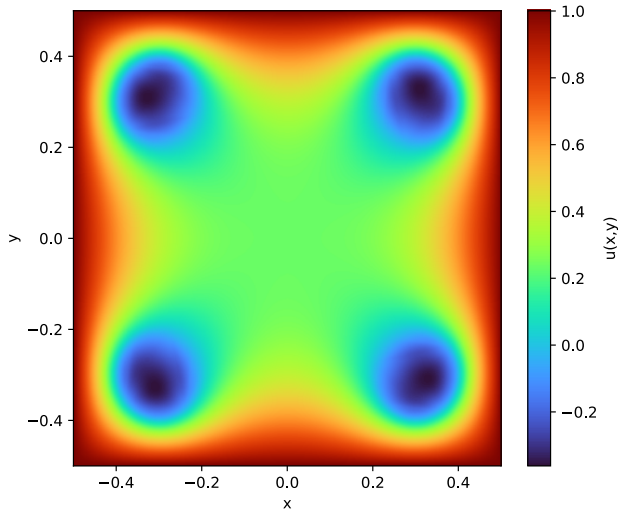


Figure 5: Reference solution of Poisson2d-C using PINNacle reference data.

Poisson-Boltzmann (Helmholtz) 2D Irregular Geometry (Poisson2d-CG)

The PDE is

$$-\nabla u + k^2 u = f(x, y), \tag{98}$$

where $f(x)$ is defined as

$$f(x) = A \cdot (\Sigma_i \mu_i^2 + x_i^2) \cdot \sin(\mu_1 \pi x_1) \sin(\mu_2 \pi x_2). \tag{99}$$

The domain is $(x, y) \in \Omega = [-1, 1]^2$ with several circles removed. The circles $\Omega_{\text{rec}} = \cup_{i=1}^4 R_i$ are

$$R_1 = \{(x, y) : (x - 0.5)^2 + (y - 0.5)^2 \leq 0.2^2\}, \tag{100}$$
$$R_2 = \{(x, y) : (x - 0.4)^2 + (y + 0.4)^2 \leq 0.4^2\}, \tag{101}$$
$$R_3 = \{(x, y) : (x + 0.2)^2 + (y + 0.7)^2 \leq 0.1^2\}, \tag{102}$$
$$R_4 = \{(x, y) : (x + 0.6)^2 + (y - 0.5)^2 \leq 0.3^2\}. \tag{103}$$

The boundary conditions are

$$u(x, y) = 0.2, \quad (x, y) \in \partial\Omega_{\text{rec}}, \tag{104}$$
$$u(x, y) = 1, \quad (x, y) \in \partial\Omega_{\text{circle}}. \tag{105}$$

The parameters are

$$\mu_1 = 1,\ \mu_2 = 4,\ k = 8,\ A = 10. \tag{106}$$

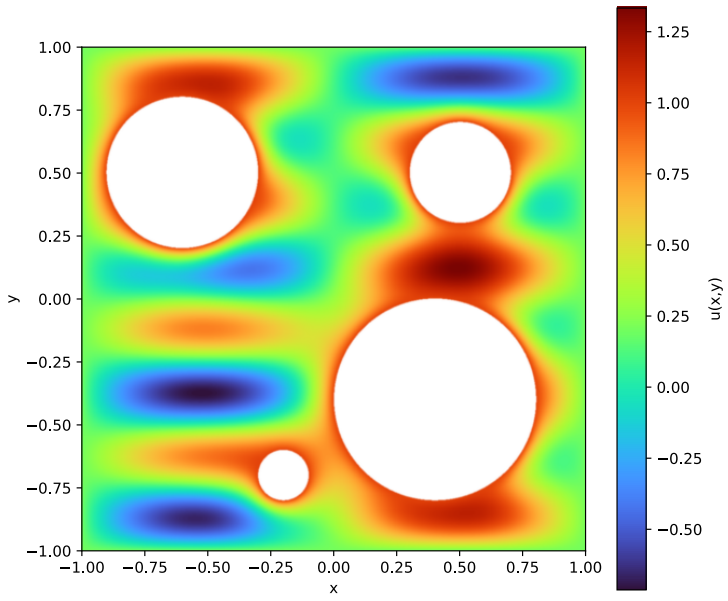


Figure 6: Reference solution of Poisson2d-CG using PINNacle reference data.

Poisson 3D Complex Geometry with Two Domains (Poisson3d-CG)

The Poisson 3D equation with two domains is given by

$$-\mu_i \Delta u + k_i^2 u = f(x,y,z), \quad i = 1,2. \tag{107}$$

The function $f(x,y,z)$ is defined as

$$\begin{aligned} f(x,y,z) = A_1 \frac{\exp(\sin m_1\pi x + \sin m_2\pi y + \sin m_3\pi z)}{x^2+y^2+z^2+1}(x^2+y^2+z^2-1) \\ + A_2 \sin(m_1\pi x)\sin(m_2\pi y)\sin(m_3\pi z). \end{aligned} \tag{108}$$

The coefficients are defined as

$$\begin{cases} \mu = \mu_1, & k = k_1, \quad x \in \Omega_1, \\ \mu = \mu_2, & k = k_2, \quad x \in \Omega_2. \end{cases} \tag{109}$$

The boundary condition is

$$\frac{\partial u}{\partial n} = 0, \quad x \in \partial\Omega. \tag{110}$$

The domains and other parameters are defined as follows:

$$\Omega_1 = [0,1] \times [0,1] \times [0,0.5] \setminus \bigcup_{i=1}^{4} R_i, \tag{111}$$

$$\Omega_2 = [0,1] \times [0,1] \times [0.5,1] \setminus \bigcup_{i=1}^{4} R_i. \tag{112}$$

The circular regions R_i are

$$R_1 = \{(x,y,z) : (x-0.4)^2 + (y-0.3)^2 + (z-0.6)^2 \le 0.2^2\}, \tag{113}$$

$$R_2 = \{(x,y,z) : (x-0.6)^2 + (y-0.7)^2 + (z-0.6)^2 \le 0.2^2\}, \tag{114}$$

$$R_3 = \{(x,y,z) : (x-0.2)^2 + (y-0.8)^2 + (z-0.7)^2 \le 0.1^2\}, \tag{115}$$

$$R_4 = \{(x,y,z) : (x-0.6)^2 + (y-0.2)^2 + (z-0.3)^2 \le 0.1^2\}. \tag{116}$$

Other parameters are

$$m_1 = 1,\ \ m_2 = 10,\ \ m_3 = 5,\ \ \mu_1 = 1,\ \ \mu_2 = 1,\ \ k_1 = 8,\ \ k_2 = 10,\ \ A_1 = 20,\ \ A_2 = 100. \tag{117}$$

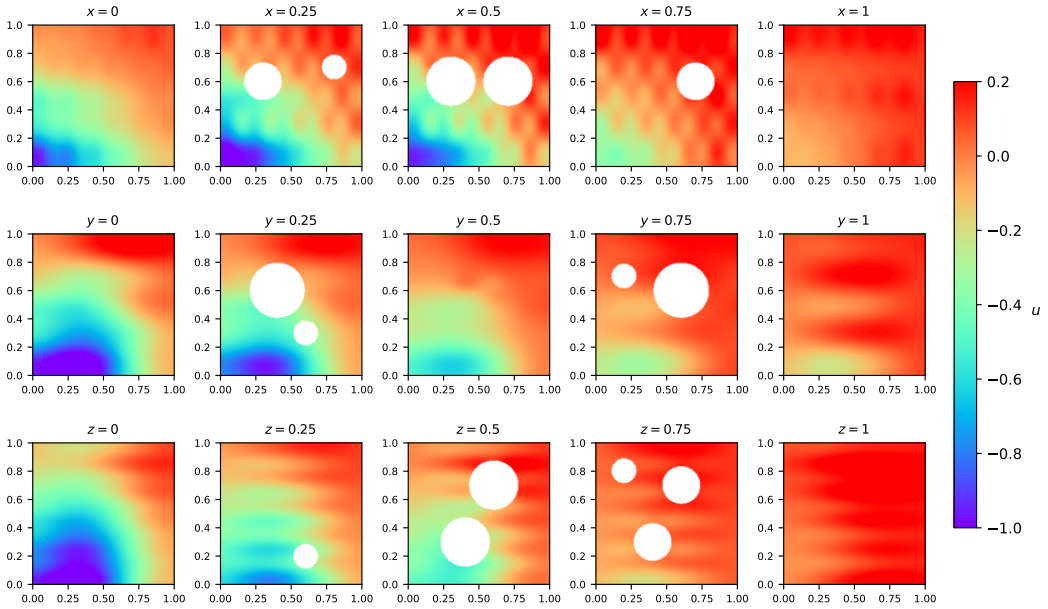


Figure 7: Reference solution of Poisson3d-CG on YZ, XZ, and XY planes at $x = 0, 0.25, 0.5, 0.75, 1.0$, $y = 0.0, 0.25, 0.5, 0.75, 1.0$, and $z = 0.0, 0.25, 0.5, 0.75, 1.0$.

N-Dimensional Poisson equation (PNd) The governing PDE is

$$-\Delta u = \frac{\pi^2}{4} \sum_{i=1}^{n} \sin\left(\frac{\pi}{2} x_i\right). \tag{118}$$

The domain is defined by $\Omega = [0, 1]^n$. The exact solution is

$$u = \sum_{i=1}^{n} \sin\left(\frac{\pi}{2} x_i\right). \tag{119}$$

We choose $n = 5$ in our experiments.

D.4 Heat equations

2D Heat with Varying Coefficients (Heat2d-VC)

The governing equation is

$$u_t - \nabla \cdot (a(x, y) \nabla u) = f(x, y, t). \tag{120}$$

The domain is

$$(x, y) \in \Omega \times T \ = \ [0, 1]^2 \times [0, 5]. \tag{121}$$

The function $a(x, y)$ is chosen similarly to Darcy flow but with an exponential GRF.

The function $f(x, y, t)$ is defined as

$$f(x, y, t) = A \sin(m_1 \pi x) \sin(m_2 \pi y) \sin(m_3 \pi t). \tag{122}$$

The initial and boundary conditions are

$$u(x, y, 0) = 0, \quad (x, y) \in \Omega, \tag{123}$$
$$u(x, y, t) = 0, \quad (x, y) \in \partial\Omega. \tag{124}$$

The parameters are

$$A = 200, \ m_1 = 1, \ m_2 = 5, \ m_3 = 1. \tag{125}$$

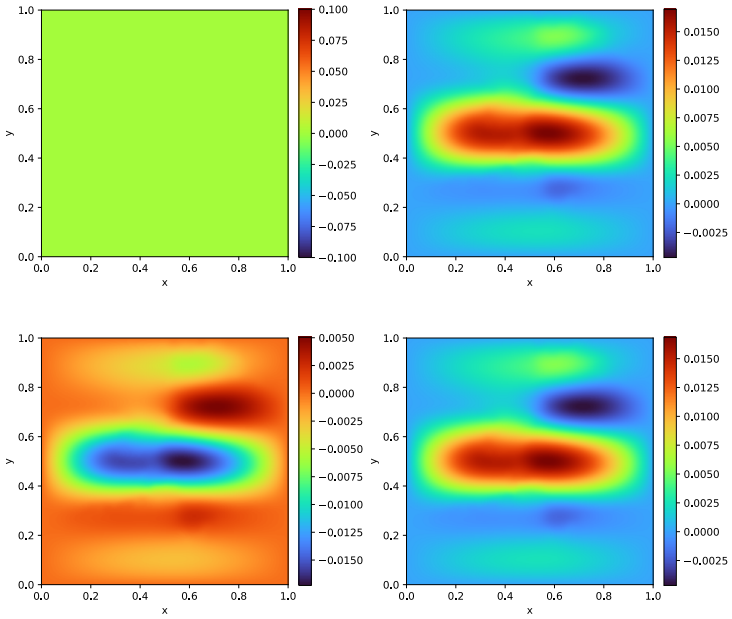


Figure 8: Reference solution of Heat2d-VC at timesteps $t = 0, 0.3, 0.7, 1$ using PINNacle reference data.

2D Heat Multi-Scale (Heat2d-MS)

The PDE equation is given by

$$u_t - \frac{1}{(500\pi)^2} u_{xx} - \frac{1}{\pi^2} u_{yy} = 0. \tag{126}$$

The domain is $(x, y) \in \Omega \times T = [0, 1]^2 \times [0, 5]$.

The initial and boundary conditions are

$$u(x, y, 0) = \sin(20\pi x) \sin(\pi y), \qquad (x, y) \in \Omega, \tag{127}$$
$$u(x, y, t) = 0, \qquad (x, y) \in \partial\Omega, \ t \in T. \tag{128}$$

2D Heat Complex Geometry (Heat2d-CG)

The PDE is

$$u_t - \nabla u = 0. \tag{129}$$

The domain is defined as

$$(x, y, t) \in \Omega \times T = ([-8, 8] \times [-12, 12] \setminus \cup_i R_i) \times [0, 3]. \tag{130}$$

The initial condition is

$$u(x, y, 0) = 0. \tag{131}$$

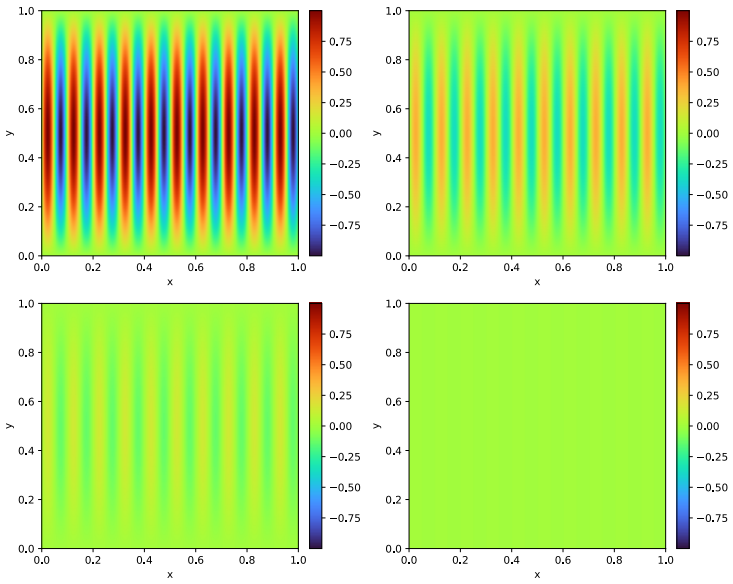


Figure 9: Reference solution of Heat2d-MS at timesteps $t = 0, 1, 2, 5$ using PINNacle reference data.

The boundary conditions are

$$-n\cdot(-\nabla u) = 5 - u, \tag{132}$$
$$-n\cdot(-\nabla u) = 1 - u, \tag{133}$$
$$-n\cdot(-\nabla u) = 0.1 - u. \tag{134}$$

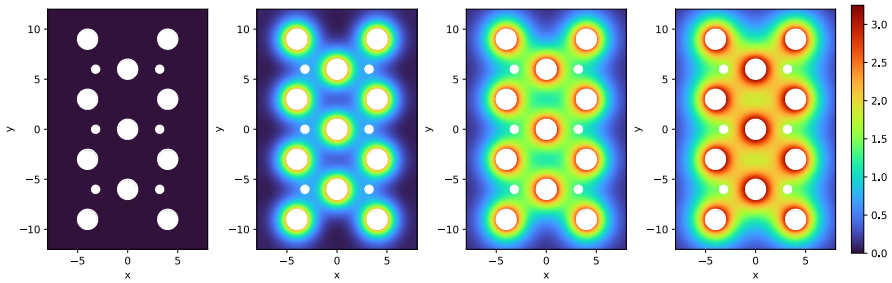


Figure 10: Reference solution of Heat2d-CG at timesteps $t = 0, 1, 2, 3$ using PINNacle reference data.

N Dimensional Heat Equation (HNd)

The governing equation is

$$u_t = k\Delta u + f(\boldsymbol{x}, t), \tag{135}$$

where

$$f(\boldsymbol{x}, t) = -\frac{1}{d}\,\|\boldsymbol{x}\|_2^2\,\exp\left(\tfrac{1}{2}\|\boldsymbol{x}\|_2^2 + t\right). \tag{136}$$

The domain is $(\boldsymbol{x}, t) \in \Omega \times T$. The geometric domain Ω is a unit sphere in d-dimensional space and $T = [0, 1]$. We choose d as 5.

The initial and boundary conditions are

$$\boldsymbol{n}\cdot\nabla u = g(\boldsymbol{x}, t),\ (\boldsymbol{x}, t) \in \partial\Omega \times [0, 1], \tag{137}$$
$$u(\boldsymbol{x}, 0) = g(\boldsymbol{x}, 0),\ \boldsymbol{x} \in \Omega, \tag{138}$$

where

$$g(\boldsymbol{x}, t) = \exp\left(\tfrac{1}{2}\|\boldsymbol{x}\|_2^2 + t\right). \tag{139}$$

We choose dimension $d = 5$ and the parameter is

$$k = \frac{1}{d}. \tag{140}$$

D.5 WAVE EQUATION

1D Wave Equation (Wave1d-C)

The governing equation is

$$u_{tt} - 4\,u_{xx} = 0. \tag{141}$$

The domain is $(x, t) \in \Omega \times T = [0, 1] \times [0, 1]$.
The initial condition is

$$u(x, 0) = \sin(\pi x) + \frac{1}{2}\sin(4\pi x), \tag{142}$$

$$u_t(x, 0) = 0. \tag{143}$$

The boundary conditions are

$$u(0, t) = u(1, t) = 0. \tag{144}$$

The analytical solution of this problem is

$$u(x, t) = \sin(\pi x)\cos(2\pi t) + \frac{1}{2}\sin(4\pi x)\cos(8\pi t). \tag{145}$$

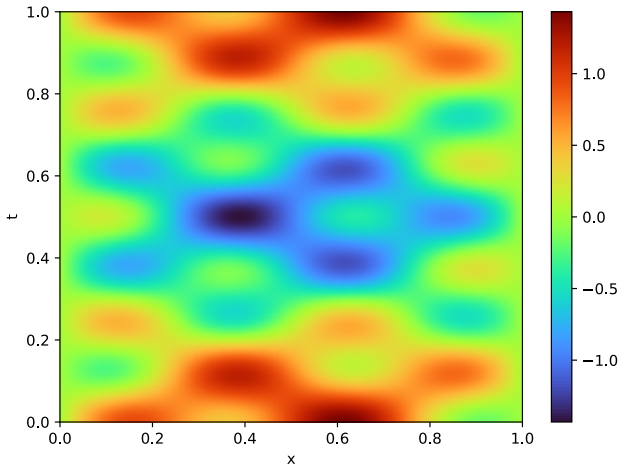


Figure 11: Reference solution of Wave1d-C using PINNacle reference data.

D.6 NAVIER-STOKES EQUATIONS

2D NS lid-driven flow (NS2d-C)

The PDE is given by

$$uu_x + vu_y + p_x = \frac{1}{Re}(u_{xx} + u_{yy}), \tag{146}$$

$$uv_x + vv_y + p_y = \frac{1}{Re}(v_{xx} + v_{yy}), \tag{147}$$

$$u_x + v_y = 0. \tag{148}$$

The domain is $(x, y) \in \Omega = [0, 1]^2$, the top boundary is Γ_1, the left, right and bottom boundary is Γ_2. The boundary conditions are

$$u(x, y) = 4x(1 - x), \qquad y \in \Gamma_1, \tag{149}$$
$$v(x, y) = 0, \qquad y \in \Gamma_1, \tag{150}$$
$$u(x, y) = 0, v(x, y) = 0, \qquad y \in \Gamma_2, \tag{151}$$
$$p(0, 0) = 0. \tag{152}$$

The Reynolds number $Re = 100$.

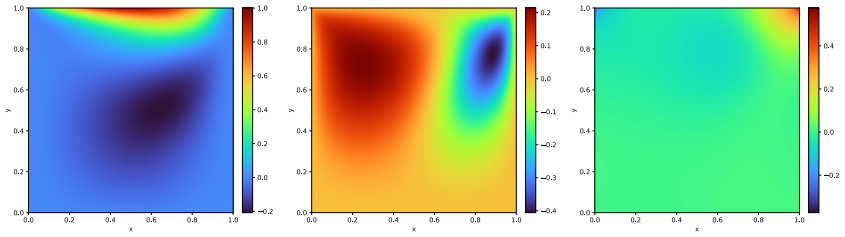


Figure 12: Reference solution of NS2d-C, where the left panel shows u, the middle pannel shows v, and the right panel shows p from the PINNacle reference data.

2D Back Step Flow (NS-CG)

The PDE is given by

$$uu_x + vu_y + p_x = \frac{1}{Re}(u_{xx} + u_{yy}), \tag{153}$$
$$uv_x + vv_y + p_y = \frac{1}{Re}(v_{xx} + v_{yy}), \tag{154}$$
$$u_x + v_y = 0. \tag{155}$$

The domain is defined as $(x, y) \in \Omega = [0, 4] \times [0, 2] \setminus ([0, 2] \times [1, 2] \cup R_i)$ (excluding the top-left quarter). The inlet boundary is Γ_{in} and the outlet boundary is Γ_{out}.

The boundary conditions are

$$u(0, y) = 4y(1 - y), \tag{156}$$
$$v(0, y) = 0, \tag{157}$$
$$p(4, y) = 0, \tag{158}$$
$$u(x, y) = 0, \ v(x, y) = 0, \qquad x \notin \Gamma_{in} \cup \Gamma_{out}. \tag{159}$$

The Reynolds number of $Re = 100$.

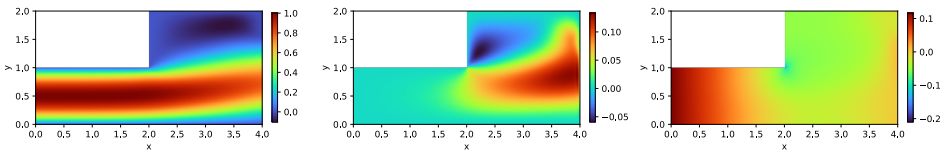


Figure 13: Reference solution of NS-CG, where the left panel shows u, the middle pannel shows v, and the right panel shows p from the PINNacle reference data.

D.7 Detailed Experiments Results

Experimental Setup All experimental settings follow those of PINNacle, except that we use a neural network architecture with 3 hidden layers and 50 neurons per layer. The models are trained for 20,000 iterations with a learning rate of 0.001, using a total of 8,192 collocation points across all experiments. We conduct experiments under three optimization settings: Adam (Kingma and Ba, 2015) with equal loss weights, LRA (Wang et al., 2021), and HARMONIC (Bu et al., 2026).

Results Table 1 reports the 95% confidence intervals of the relative L_2 errors over five independent runs on 12 PDEs for Rand-R, L^∞, R3, RAD, PINNACLE, LAS, and RAMP with $\varepsilon \in \{0.002, 0.0002, 0.00002\}$.

D.8 Computational Cost

Table 5 summarizes the computational cost of the compared methods across the PDEs. Using HARMONIC, we measure the wall-clock time required for 100 epochs over 10 repeated runs and report the corresponding 95% confidence intervals. Any negative lower confidence bound is truncated to zero to ensure physically meaningful estimates. All experiments were conducted on a single NVIDIA GeForce RTX 4090 GPU. R3 is the fastest method because it only computes the average sampled residual magnitude and compares the resulting values. RAD introduces additional overhead by sampling from a multinomial distribution over the discretized grid based on the sampled loss magnitudes. LAS further requires an automatic differentiation step with respect to the network input, while the L^∞-based method is the most expensive because it repeats this input-gradient computation 20 times. RAMP computes pointwise parameter sensitivity using a `vmap`-based parallelizable procedure, followed by an input-gradient computation. Although this introduces additional overhead compared with simpler sampling strategies such as R3, RAD, and LAS, the overall wall-clock cost remains moderate. In particular, RAMP is consistently much cheaper than the L^∞-based method, which repeatedly applies input-gradient computations 20 times. Using the midpoint of each confidence interval, RAMP is approximately $1.7\times$–$2.7\times$ faster than the L^∞-based method across the tested PDEs. This indicates that the `vmap`-based sensitivity computation provides an efficient way to incorporate parameter-sensitivity information without incurring the large computational overhead caused by repeated input-gradient evaluations.

Table 4: 95% confidence intervals of relative L_2 errors on 12 PINNacle PDEs under Adam, LRA, and HARMONIC, computed over five independent runs.

Method	Burgers1d	Poisson2d-C	Poisson2d-CG	Poisson3d-CG	PNd	Heat2d-VC	Heat2d-MS	Heat2d-CG	HNd	Wave1d-C	NS2d-C	NS-CG
Adam												
Rand-R	[0.0308, 0.1080]	[0.5648, 0.7630]	[0.1564, 0.3228]	[0.3700, 0.7254]	[0.0004, 0.0004]	[0.2296, 0.2544]	[0.0875, 0.2071]	[0.1109, 0.1373]	[0.0001, 0.0037]	[0.2601, 0.3775]	[0.4433, 0.5449]	[0.2891, 0.3175]
L^∞	[0.0925, 0.3943]	[0.6661, 0.7231]	[0.0910, 0.1520]	[0.2668, 0.5336]	[0.0000, 0.1014]	[0.2462, 0.2834]	[0.0819, 0.2273]	[0.4770, 0.5108]	[0.0005, 0.0041]	[0.3189, 0.3561]	[0.4208, 0.5024]	[0.2831, 0.3651]
R3	[0.0115, 0.3581]	[0.6644, 0.7458]	[0.2155, 0.4337]	[0.3958, 0.4960]	[0.0017, 0.0047]	[0.3091, 0.4409]	[0.1295, 0.3109]	[0.1617, 0.1819]	[0.0013, 0.0069]	[0.2576, 0.3364]	[0.3299, 0.4543]	[0.3212, 0.3562]
RAD	[0.0651, 0.1123]	[0.6804, 0.7138]	[0.3322, 0.4096]	[0.4692, 0.5928]	[0.0001, 0.0013]	[0.2233, 0.2871]	[0.1075, 0.1883]	[0.1131, 0.1289]	[0.0010, 0.0024]	[0.2927, 0.3599]	[0.4069, 0.4941]	[0.2973, 0.3193]
PINNACLE	[0.0370, 0.1938]	[0.6114, 0.7322]	[0.2219, 0.3747]	[0.4457, 0.6021]	[0.0004, 0.0006]	[0.2268, 0.4966]	[0.0000, 0.8560]	[0.1150, 0.1550]	[0.0001, 0.0045]	[0.2924, 0.3832]	[0.4231, 0.5269]	[0.2701, 0.3267]
LAS	[0.0228, 0.0612]	[0.6138, 0.7324]	[0.1017, 0.1763]	[0.3623, 0.5725]	[0.0000, 0.0045]	[0.2166, 0.3968]	[0.0199, 0.4439]	[0.4542, 0.6424]	[0.0000, 0.0072]	[0.2937, 0.3815]	[0.4053, 0.5181]	[0.2863, 0.3187]
RAMP ($\varepsilon = 0.002$)	[0.0827, 0.1157]	[0.6098, 0.7286]	[0.1177, 0.3967]	[0.3511, 0.6407]	[0.0004, 0.0020]	[0.2028, 0.2892]	[0.1057, 0.2049]	[0.1104, 0.1192]	[0.0013, 0.0015]	[0.2934, 0.3328]	[0.4639, 0.4851]	[0.2858, 0.3710]
RAMP ($\varepsilon = 0.0002$)	[0.0459, 0.1031]	[0.7107, 0.7107]	[0.2051, 0.2051]	[0.3951, 0.6363]	[0.0003, 0.0005]	[0.2243, 0.2653]	[0.0991, 0.1565]	[0.1147, 0.1171]	[0.0030, 0.0030]	[0.3433, 0.3433]	[0.4139, 0.6597]	[0.3279, 0.3431]
RAMP ($\varepsilon = 0.00002$)	[0.0288, 0.1154]	[0.7107, 0.7107]	[0.1712, 0.1712]	[0.4082, 0.6314]	[0.0004, 0.0004]	[0.2202, 0.2904]	[0.1047, 0.1649]	[0.1158, 0.1200]	[0.0012, 0.0012]	[0.3070, 0.3070]	[0.4301, 0.5379]	[0.3126, 0.3598]
LRA												
Rand-R	[0.0237, 0.0531]	[0.6819, 0.7307]	[0.7245, 0.7733]	[0.4517, 0.6075]	[0.0010, 0.0038]	[1.2968, 3.5754]	[0.1030, 0.1950]	[0.0652, 0.0810]	[0.0000, 0.0100]	[0.3907, 0.4693]	[0.0722, 0.0924]	[0.1591, 0.1747]
L^∞	[0.0000, 0.5065]	[0.6816, 0.7200]	[0.5883, 0.7113]	[0.4291, 0.5863]	[0.0015, 0.0021]	[2.5760, 3.7640]	[0.0007, 0.7873]	[0.6906, 0.7128]	[0.0041, 0.0097]	[0.4819, 0.5905]	[0.4143, 0.6175]	[0.2222, 0.2538]
R3	[0.1404, 0.2320]	[0.6907, 0.7171]	[0.7772, 0.8366]	[0.2170, 0.3520]	[0.0015, 0.0071]	[3.2827, 7.0135]	[0.1394, 0.6822]	[0.1208, 0.1306]	[0.0015, 0.0073]	[0.5644, 0.7190]	[0.3833, 0.5207]	[0.2674, 0.3024]
RAD	[0.1688, 0.2962]	[0.6807, 0.7149]	[0.7338, 0.8004]	[0.4724, 0.5726]	[0.0000, 0.0050]	[1.0227, 6.5215]	[0.1578, 0.2782]	[0.0757, 0.0935]	[0.0000, 0.0099]	[0.4145, 0.5547]	[0.1010, 0.1848]	[0.1867, 0.2039]
PINNACLE	[0.0244, 0.0852]	[0.6760, 0.7180]	[0.6830, 0.7502]	[0.4880, 0.7034]	[0.0011, 0.0025]	[1.6954, 2.1324]	[0.0788, 0.1820]	[0.0712, 0.0850]	[0.0000, 0.0091]	[0.3903, 0.4277]	[0.0714, 0.1062]	[0.1642, 0.1906]
LAS	[0.0139, 0.0599]	[0.6790, 0.7156]	[0.6211, 0.7279]	[0.4839, 0.6289]	[0.0000, 0.0102]	[1.8689, 3.8213]	[0.1784, 0.6878]	[0.9456, 0.9790]	[0.0005, 0.0049]	[0.4013, 0.5251]	[0.1359, 0.2047]	[0.1546, 0.1920]
RAMP ($\varepsilon = 0.002$)	[0.0307, 0.0435]	[0.6792, 0.7202]	[0.7187, 0.8081]	[0.3410, 0.6900]	[0.0006, 0.0090]	[1.0336, 2.2968]	[0.1685, 0.2271]	[0.0656, 0.0780]	[0.0010, 0.0040]	[0.3945, 0.4279]	[0.0786, 0.1008]	[0.1669, 0.1867]
RAMP ($\varepsilon = 0.0002$)	[0.0216, 0.0418]	[0.6740, 0.7140]	[0.7000, 0.7806]	[0.4931, 0.6527]	[0.0006, 0.0076]	[1.5439, 2.8103]	[0.0928, 0.2594]	[0.0639, 0.0801]	[0.0013, 0.0021]	[0.4035, 0.4595]	[0.0727, 0.1049]	[0.1616, 0.1894]
RAMP ($\varepsilon = 0.00002$)	[0.0238, 0.0416]	[0.6805, 0.7107]	[0.7421, 0.7521]	[0.4908, 0.6130]	[0.0023, 0.0047]	[1.5436, 2.0948]	[0.0922, 0.2236]	[0.0651, 0.0809]	[0.0000, 0.0174]	[0.4003, 0.4605]	[0.0770, 0.0922]	[0.1632, 0.1888]
HARMONIC												
Rand-R	[0.0178, 0.0566]	[0.0100, 0.0238]	[0.0249, 0.0323]	[0.1028, 0.2358]	[0.0000, 0.0069]	[0.1941, 0.2163]	[0.3428, 0.8168]	[0.0691, 0.0735]	[0.0000, 0.0517]	[0.0315, 0.0809]	[0.0955, 0.1449]	[0.1694, 0.1864]
L^∞	[0.0000, 0.5875]	[0.1310, 1.0126]	[0.0651, 0.0887]	[0.0918, 0.2178]	[0.0001, 0.0073]	[0.1851, 0.3469]	[0.2452, 0.6588]	[0.6078, 0.7600]	[0.0015, 0.0065]	[0.0443, 0.1263]	[0.1697, 0.3025]	[0.2541, 0.2829]
R3	[0.0272, 0.1410]	[0.8977, 1.4017]	[0.0320, 0.0934]	[0.2025, 0.4587]	[0.0027, 0.0129]	[0.9460, 1.6692]	[0.0000, 1.0702]	[0.0318, 0.0828]	[0.0000, 0.0956]	[0.2322, 0.3382]	[0.2181, 0.3891]	[0.2442, 0.2694]
RAD	[0.0464, 0.0974]	[0.0143, 0.0193]	[0.0261, 0.0441]	[0.1175, 0.2097]	[0.0020, 0.0050]	[0.1755, 0.3341]	[0.2882, 0.5958]	[0.0740, 0.0952]	[0.0009, 0.0015]	[0.0463, 0.1183]	[0.1169, 0.1735]	[0.1931, 0.2297]
PINNACLE	[0.0212, 0.0526]	[0.6701, 0.7187]	[0.6900, 0.7832]	[0.4294, 0.6590]	[0.0007, 0.0027]	[1.2617, 2.7941]	[0.0193, 0.3835]	[0.0717, 0.0889]	[0.0012, 0.0054]	[0.3926, 0.4192]	[0.0705, 0.1039]	[0.1459, 0.1795]
LAS	[0.0192, 0.0492]	[0.0073, 0.0331]	[0.0182, 0.0256]	[0.1084, 0.1884]	[0.0000, 0.0212]	[0.2083, 0.3129]	[0.4075, 0.7187]	[0.9325, 1.2195]	[0.0000, 0.0208]	[0.0376, 0.0892]	[0.1122, 0.1780]	[0.1655, 0.2297]
RAMP ($\varepsilon = 0.002$)	[0.0257, 0.0521]	[0.0119, 0.0169]	[0.0246, 0.0426]	[0.1092, 0.2214]	[0.0011, 0.0023]	[0.1890, 0.2570]	[0.4262, 0.6730]	[0.0650, 0.0862]	[0.0010, 0.0012]	[0.0619, 0.0929]	[0.0981, 0.1675]	[0.1586, 0.2180]
RAMP ($\varepsilon = 0.0002$)	[0.0180, 0.0472]	[0.0077, 0.0291]	[0.0215, 0.0373]	[0.1099, 0.1599]	[0.0008, 0.0016]	[0.1827, 0.2315]	[0.2419, 0.5751]	[0.0679, 0.0787]	[0.0000, 0.0122]	[0.0340, 0.0798]	[0.0970, 0.2524]	[0.1790, 0.1946]
RAMP ($\varepsilon = 0.00002$)	[0.0136, 0.0524]	[0.0124, 0.0224]	[0.0244, 0.0280]	[0.1141, 0.1455]	[0.0000, 0.0127]	[0.1229, 0.5233]	[0.2799, 0.5171]	[0.0674, 0.0852]	[0.0006, 0.0012]	[0.0418, 0.1106]	[0.1010, 0.2388]	[0.1683, 0.2119]

Table 5: Computational cost (in seconds) per 100 epochs, reported with 95% confidence intervals over 10 runs.

Method	Burgers1d	Poisson2d-C	Poisson2d-CG	Poisson3d-CG	PNd	Heat2d-VC	Heat2d-MS	Heat2d-CG	HNd	Wave1d-C	NS2d-C	NS-CG
L^∞	[7.4855, 7.7560]	[15.3250, 16.5106]	[17.1162, 17.8328]	[19.4301, 19.8897]	[19.1462, 19.5199]	[38.6965, 39.4866]	[9.7010, 9.8241]	[11.3025, 11.4883]	[19.8620, 20.2291]	[11.4605, 11.7190]	[24.2039, 24.5439]	[26.1724, 26.7299]
R3	[1.2955, 1.3102]	[2.0019, 2.0742]	[2.0618, 2.2622]	[2.2219, 2.5358]	[2.1342, 2.2079]	[5.5800, 5.7588]	[1.6723, 1.7649]	[2.9871, 3.1007]	[2.9502, 3.1026]	[2.0532, 2.1525]	[7.0039, 7.3435]	[7.0315, 7.1781]
RAD	[1.3929, 1.3999]	[2.8662, 3.0639]	[3.0454, 3.1854]	[3.1491, 3.2598]	[2.2248, 2.3363]	[5.3641, 5.5403]	[1.7424, 1.8150]	[3.6765, 3.8220]	[3.1049, 3.3217]	[2.0961, 2.1103]	[6.9716, 7.2137]	[7.0454, 7.2468]
LAS	[1.4143, 1.5107]	[1.8883, 1.9094]	[2.0879, 2.1485]	[2.1000, 2.1793]	[2.4409, 2.4666]	[5.6842, 5.8185]	[1.8057, 1.8294]	[2.7331, 2.8240]	[3.2274, 3.2744]	[2.2084, 2.2169]	[7.0781, 7.1464]	[7.2014, 7.3464]
PINNACLE	[0.0000, 5.4877]	[0.0000, 5.2560]	[0.0000, 5.6230]	[0.0000, 6.1158]	[0.0000, 5.7240]	[0.0000, 6.6048]	[0.0000, 6.2501]	[0.6715, 6.9403]	[0.4970, 6.8064]	[0.0895, 5.6425]	[5.4799, 12.7436]	[3.9582, 10.7901]
RAMP	[4.6373, 4.8101]	[5.1740, 5.3276]	[5.6756, 5.9160]	[6.5259, 6.9052]	[4.4788, 4.5957]	[7.5867, 7.9747]	[4.7648, 4.9963]	[6.5565, 6.6962]	[5.8880, 6.0355]	[4.2021, 4.3224]	[10.8024, 11.2617]	[10.6856, 10.8644]

E AI ASSISTANCE

Generative AI tools were used to check grammar and polish the writing for readability and clarity. All AI-assisted edits were reviewed by the authors, who take full responsibility for the final content of this work.