CURVATURE-HARMONIZED METRIC DESCENT FOR MULTI-LOSS OPTIMIZATION IN PINNS

Anonymous authors
Paper under double-blind review

ABSTRACT

Training physics-informed neural networks (PINNs) requires jointly reducing multiple physics-based objectives despite gradient imbalance, directional conflicts, and ill-conditioned parameter-space geometry. Objective-aware methods can remain slow along ill-conditioned directions, whereas curvature-aware methods alone do not guarantee simultaneous descent across objectives. We propose Curvature-HarmonIzed MEtric Descent (CHIME), which selects multi-objective directions in a shared curvature metric. CHIME maximizes the worst relative first-order reduction in residual norm under a step budget induced by this metric. The resulting direction lies in a curvature-harmonized hull, coupling a first-order non-conflict guarantee across objectives with curvature-aware preconditioning. A local linearized analysis shows that worst-objective progress and residual conditioning jointly control contraction of positive-eigenvalue residual modes. A practical SOAP-style realization estimates the shared metric and performs direction selection efficiently in objective space. Across four forward PINN benchmarks, CHIME reduces error by geometric mean factors of 5.85 and 5.45 relative to the best objective-aware and curvature-aware baseline for each benchmark, respectively. On the multi-task learning benchmark, it also achieves the best aggregate performance among the compared methods.

1 INTRODUCTION

Partial differential equations (PDEs) are widely used to describe physical, chemical, and biological phenomena (Morton and Mayers, 2005). Forward problems aim to predict solutions to specified PDEs under given initial and boundary conditions (Lau et al., 2024). Physics-informed neural networks (PINNs) offer a mesh-free approach to these problems, using neural-network outputs to approximate solutions at spatial or spatio-temporal coordinates (Lu et al., 2021). For such neural representations, automatic differentiation evaluates derivatives with respect to the input coordinates (Baydin et al., 2018). Using these derivatives, standard PINN training minimizes a composite loss that penalizes PDE residuals and violations of initial and boundary conditions (Raissi et al., 2019). This differentiable representation also supports derivative-based computations, including strains from displacement gradients in solid mechanics (Haghighat et al., 2021) and policy improvement using value-function gradients in optimal control (Meng et al., 2024). Despite this promise, the difficulty of minimizing the composite loss can limit the accuracy of PINN solutions (Rathore et al., 2024).

Accurate forward solutions can require all loss terms to approach zero (Rathore et al., 2024). One difficulty is an imbalance in the magnitudes of loss-specific gradients, which can make the combined update dominated by one loss while neglecting others (Wang et al., 2021). Directional conflicts among these gradients can also cause the aggregate update to increase some loss terms (Hwang and Lim, 2024; Liu et al., 2025). Beyond these objective-level difficulties, differential operators in PDE residuals can induce ill-conditioned parameter-space geometry, making first-order training slow or impractical (Ryck et al., 2024; Rathore et al., 2024). In the neural tangent kernel (NTK) regime, error modes associated with small eigenvalues decay slowly under gradient descent, linking this bottleneck to spectral bias in fully connected PINNs (Wang et al., 2022).

Objective-aware methods address gradient imbalance via adaptive loss weighting, including LRA and ReLoBRaLo (Wang et al., 2021; Bischof and Kraus, 2025). Beyond loss rescaling, DCGD and ConFIG use multi-objective optimization (MOO) to construct non-conflicting directions from loss-specific gradients (Hwang and Lim, 2024; Liu et al., 2025). HARMONIC further combines nonnegative aggregation and non-conflict through a harmonized cone and analyzes convergence to

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

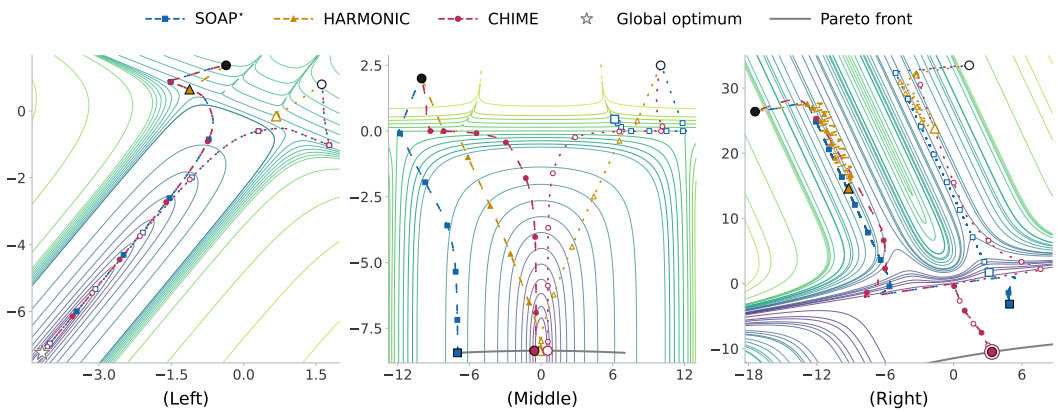


Figure 1: **Optimization trajectories across three illustrative cases. (Left)** An ill-conditioned single-objective problem. **(Middle)** A well-conditioned bi-objective problem with severe scale imbalance. **(Right)** An ill-conditioned bi-objective problem with scale imbalance. The problems are variants of the CAGrad toy example (Liu et al., 2021a). Each method is initialized from ● and ○. The global optimum and Pareto front are indicated by ☆ and ——, respectively. Colored trajectories correspond to -■- SOAP*, -▲- HARMONIC, and -●- CHIME. SOAP* reaches the corresponding target from both initializations only in the left panel, and HARMONIC only in the middle panel; neither does so in the right panel. CHIME reaches the corresponding target from both initializations in all three cases. See App. A for detailed experimental settings.

Pareto-stationary points (Bu et al., 2026). However, these first-order direction-selection rules do not explicitly account for the curvature underlying ill-conditioning. Curvature-aware methods instead use natural-gradient or Newton-type preconditioning for the scalar composite loss (Müller and Zeinhofer, 2023; Rathore et al., 2024). For scalable PINN training, SOAP (Vyas et al., 2025) with LRA-style loss balancing has been shown to improve gradient alignment through preconditioning (Wang et al., 2025). Yet curvature-aware methods do not explicitly account for simultaneous descent across objectives when selecting the preconditioned update.

Either approach alone can be insufficient when gradient imbalance, directional conflict, and ill-conditioning coexist. Fig. 1 illustrates these limitations using HARMONIC and SOAP with LRA-style loss balancing (SOAP*), recent representatives of objective-aware and curvature-aware PINN optimization, respectively. HARMONIC struggles in the ill-conditioned single-objective case (Left), whereas SOAP* can miss the Pareto front even in the well-conditioned bi-objective case with severe scale imbalance (Middle). The ill-conditioned bi-objective case (Right) brings these effects together, reflecting the objective-level and conditioning difficulties identified in prior analyses of PINN training (Wang et al., 2021; 2022; Rathore et al., 2024); neither baseline consistently reaches the target. These observations motivate selecting loss-gradient combinations by their objective-wise progress in the geometry of the preconditioned update.

To this end, we introduce **C**urvature-**H**armon**I**zed **ME**tric descent (**CHIME**) for training PINNs. We formulate direction selection as a max–min problem over relative first-order reduction in residual norm under a shared curvature-metric step budget. To the best of our knowledge, no prior PINN optimization method directly optimizes this criterion. Our geometric analysis characterizes the selected direction as a metric projection onto a curvature-harmonized hull, yielding a first-order non-conflict guarantee across objectives. We further derive a local linearized contraction bound for positive-eigenvalue residual modes that depends jointly on worst-objective progress and residual conditioning. A practical SOAP (Vyas et al., 2025)-style realization estimates the shared metric and selects directions efficiently in objective space. Across four forward benchmarks from PINNacle (Hao et al., 2024) and the A-PINN study (Yuan et al., 2022), CHIME achieves geometric mean error-reduction factors of 5.85 and 5.45 over the best objective-aware and curvature-aware baseline on each benchmark, respectively. Beyond PINNs, CHIME also achieves the best aggregate performance among the compared methods on the multi-task learning (MTL) benchmark.

2 Related Works

Objective-aware optimization. Objective-aware optimization has been extensively studied in MTL through adaptive reweighting and gradient aggregation. GradNorm (Chen et al., 2018) adjusts gradient

Table 1: Comparison of representative PINN optimization methods. **Objective-aware**: explicitly uses objective-specific information; **Conflict-free**: guarantees a nonnegative inner product between the selected direction and every objective gradient; **Curvature-aware**: applies non-diagonal curvature-based preconditioning to the parameter update. ✓/✗ indicate whether each property is provided.

Method	Objective-aware	Conflict-free	Curvature-aware
LRA, NTK, ReLoBRaLo, MultiAdam	✓	✗	✗
DCGD, ConFIG, HARMONIC	✓	✓	✗
ENGD, SNGD, KFAC*	✗	✗	✓
Muon*, SOAP*	✓	✗	✓
CHIME	✓	✓	✓

norms according to relative training rates, while NTKMTL (Qin et al., 2025) reweights tasks using the maximum eigenvalue of each task-specific NTK block. MGDA (Sener and Koltun, 2018) selects the minimum-norm convex combination of task gradients, while PCGrad (Yu et al., 2020) removes conflicting components through pairwise projections. CAGrad (Liu et al., 2021a) maximizes worst-task first-order improvement within a Euclidean ball around the mean gradient. Nash-MTL (Navon et al., 2022) selects a norm-bounded direction by maximizing a Nash bargaining objective over first-order task improvements. Aligned-MTL (Senushkin et al., 2023) equalizes the nonzero singular values of the task-gradient matrix before aggregation. In PINNs, adaptive reweighting uses gradient statistics in LRA (Wang et al., 2021), kernel traces in NTK reweighting (Wang et al., 2022), and relative loss progress in ReLoBRaLo (Bischof and Kraus, 2025). MultiAdam (Yao et al., 2023) maintains per-parameter first- and second-moment estimates for each loss group's gradients. For PINN-specific MOO, DCGD (Hwang and Lim, 2024) uses a dual gradient cone for two loss groups, ConFIG (Liu et al., 2025) equalizes positive normalized-gradient projections, and HARMONIC (Bu et al., 2026) intersects primal and dual cones to combine nonnegative aggregation with non-conflict.

Curvature-aware optimization. Curvature-aware optimization has evolved toward scalable approximations suitable for neural-network training, including limited-memory quasi-Newton updates in L-BFGS (Liu and Nocedal, 1989) and layerwise Kronecker-factored Fisher approximations in K-FAC (Martens and Grosse, 2015). Shampoo (Gupta et al., 2018) uses tensor-structured gradient statistics, Muon orthogonalizes matrix-valued momentum (Wang et al., 2025), and SOAP (Vyas et al., 2025) applies Adam in Shampoo's eigenbasis. In PINNs, Adam is often followed by L-BFGS (Lu et al., 2021); NNCG further refines such solutions using Hessian-vector products and Nyström-preconditioned conjugate gradients (Rathore et al., 2024). Energy natural gradient descent (ENGD) (Müller and Zeinhofer, 2023) uses a Hessian-induced function-space metric, while sketchy natural gradient descent (SNGD) (Mckay et al., 2025) reduces natural-gradient computation and storage through randomized sketching. KFAC* (Dangel et al., 2024) adapts Kronecker factors to differential-operator residuals through Taylor-mode automatic differentiation and weight sharing. Muon* and SOAP* pair the respective optimizers with LRA-style reweighting, combining objective-aware loss adaptation with curvature-aware updates (Wang et al., 2025).

Remaining gap. Table 1 highlights a complementary gap among the PINN optimizers. DCGD, ConFIG, and HARMONIC guarantee conflict-free directions but do not incorporate curvature-aware preconditioning, whereas Muon* and SOAP* combine objective-aware loss adaptation with curvature-aware updates but do not guarantee non-conflict across objectives. The toy trajectories in Fig. 1 provide an empirical illustration of this gap when objective-level difficulty and ill-conditioning coexist. This leaves open how to select updates that are simultaneously objective-aware, conflict-free, and curvature-aware.

3 PRELIMINARIES

Let $u_{\boldsymbol{\theta}}(\mathbf{x})$ denote a PINN with parameters $\boldsymbol{\theta} \in \mathbb{R}^D$ and spatial or spatio-temporal inputs $\mathbf{x} \in \mathbb{R}^p$. For each residual type $i \in \{1, \dots, m\}$, let $\mathbf{X}_i = \{\mathbf{x}_i^n\}_{n=1}^{N_i}$ be its collocation set and $\mathbf{r}_i(\mathbf{X}_i, \boldsymbol{\theta}) \in \mathbb{R}^{N_i}$ collect PDE residuals or initial/boundary-condition violations (Raissi et al., 2019). At a fixed iteration, we omit $\mathbf{X}_i$ and $\boldsymbol{\theta}$ from the notation and define

$$\mathbf{J}_i = \frac{\partial \mathbf{r}_i}{\partial \boldsymbol{\theta}} \in \mathbb{R}^{N_i \times D}, \quad L_i = \frac{1}{N_i} \|\mathbf{r}_i\|_2^2, \quad \mathbf{g}_i = \nabla_{\boldsymbol{\theta}} L_i, \quad \mathbf{G} = [\mathbf{g}_1, \dots, \mathbf{g}_m] \in \mathbb{R}^{D \times m}. \tag{1}$$

For forward problems, the target is $L_i = 0$ for every i. With update direction $\mathbf{d} \in \mathbb{R}^D$ and step size $\eta > 0$, we write $\boldsymbol{\theta}^+ = \boldsymbol{\theta} - \eta\mathbf{d}$.

3.1 First-Order Multi-Objective Optimization

Since $L_i \geq 0$, requiring $L_i \leq 0$ is equivalent to $L_i = 0$, giving the feasibility problem

$$\min_{\boldsymbol{\theta}} \ 0 \quad \text{s.t.} \quad L_i(\boldsymbol{\theta}) \leq 0, \qquad i = 1, \ldots, m. \tag{2}$$

To obtain a local directional interpretation, linearize each loss as $L_i(\boldsymbol{\theta} - \eta\mathbf{d}) \approx L_i - \eta\mathbf{g}_i^\top\mathbf{d} \leq 0$. Enforcing the constraints in this linear model yields $\mathbf{G}^\top\mathbf{d} \geq \mathbf{b}$, where $\mathbf{b} = [L_1/\eta, \ldots, L_m/\eta]^\top \geq \mathbf{0}$. Whenever this system is strictly feasible, consider the local direction-selection problem under the Euclidean metric, which treats all parameter directions equally:

$$\min_{\mathbf{d}} \ \frac{1}{2}\|\mathbf{d}\|_2^2 \quad \text{s.t.} \quad \mathbf{G}^\top\mathbf{d} \geq \mathbf{b}. \tag{3}$$

Its Lagrangian is

$$\mathcal{L}(\mathbf{d}, \boldsymbol{\lambda}) = \frac{1}{2}\|\mathbf{d}\|_2^2 + \boldsymbol{\lambda}^\top(\mathbf{b} - \mathbf{G}^\top\mathbf{d}). \tag{4}$$

Stationarity, primal feasibility, and dual feasibility give

$$\mathbf{d}^\star = \mathbf{G}\boldsymbol{\lambda}^\star, \qquad \mathbf{G}^\top\mathbf{d}^\star \geq \mathbf{b} \geq \mathbf{0}, \qquad \boldsymbol{\lambda}^\star \geq \mathbf{0}. \tag{5}$$

Stationarity together with dual feasibility yields a nonnegative aggregation of the loss gradients, while primal feasibility gives a first-order non-conflicting direction. These conditions correspond to membership in the primal gradient cone $\mathcal{K}$ and dual gradient cone $\mathcal{K}^*$, respectively, whose intersection recovers the *harmonized cone* $\mathcal{H}$ defined by Bu et al. (2026):

$$\mathcal{K} = \{\mathbf{G}\boldsymbol{\lambda} : \boldsymbol{\lambda} \in \mathbb{R}_+^m\}, \quad \mathcal{K}^* = \{\mathbf{d} \in \mathbb{R}^D : \mathbf{G}^\top\mathbf{d} \geq \mathbf{0}\}, \quad \mathcal{H} = \mathcal{K} \cap \mathcal{K}^*. \tag{6}$$

Thus, harmonized-cone membership emerges as the geometric condition induced by the KKT system of Euclidean first-order multi-objective direction selection.

3.2 Second-Order Single-Objective Optimization

For a fixed objective L_i, define its Gauss–Newton (GN) Hessian approximation (Bonfanti et al., 2024) and residual tangent kernel (Wang et al., 2022) by

$$\mathbf{B}_{i,\mathrm{GN}} = \frac{2}{N_i}\mathbf{J}_i^\top\mathbf{J}_i, \qquad \mathbf{K}_i = \frac{2}{N_i}\mathbf{J}_i\mathbf{J}_i^\top. \tag{7}$$

These matrices share their nonzero eigenvalues. Let $\mathbf{K}_i = \mathbf{U}_i \operatorname{Diag}(\lambda_{i1}, \ldots, \lambda_{iN_i})\mathbf{U}_i^\top$, where $\mathbf{U}_i = [\mathbf{u}_{i1}, \ldots, \mathbf{u}_{iN_i}]$ is orthogonal and positive eigenvalues are ordered first. The residual component along $\mathbf{u}_{ik}$ is the kth *residual mode*, with amplitude $a_{ik} = \mathbf{u}_{ik}^\top\mathbf{r}_i$. Since $\mathbf{g}_i = (2/N_i)\mathbf{J}_i^\top\mathbf{r}_i$, the step $\mathbf{d} = \mathbf{g}_i$ gives, in the local linearized residual model,

$$\mathbf{r}_i^+ \approx (\mathbf{I}_{N_i} - \eta\mathbf{K}_i)\mathbf{r}_i, \qquad a_{ik}^+ \approx (1 - \eta\lambda_{ik})a_{ik}. \tag{8}$$

For $0 < \eta \leq 1/\lambda_{\max}(\mathbf{K}_i)$, small positive eigenvalues give slowly contracting *weak modes*. The GN direction $\mathbf{d} = \mathbf{B}_{i,\mathrm{GN}}^-\mathbf{g}_i$, where $^-$ denotes the Moore–Penrose pseudoinverse, replaces $\mathbf{K}_i$ in the residual dynamics of Eq. (8) with

$$\mathbf{K}_{i,\mathrm{GN}} = \frac{2}{N_i}\mathbf{J}_i\mathbf{B}_{i,\mathrm{GN}}^-\mathbf{J}_i^\top = \mathbf{U}_i \begin{pmatrix} \mathbf{I}_{s_i} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}_{N_i - s_i} \end{pmatrix} \mathbf{U}_i^\top, \qquad s_i = \operatorname{rank}(\mathbf{J}_i). \tag{9}$$

This *whitening* equalizes the positive eigenvalues to one, giving the corresponding modes a common contraction factor $1 - \eta$ for $0 < \eta \leq 1$ and removing the eigenvalue-dependent slowdown of weak modes in the local model (Bonfanti et al., 2024; Jiang et al., 2026).

SOAP (Vyas et al., 2025) applies Adam (Kingma and Ba, 2014) in the eigenbasis of Shampoo's tensor-structured gradient statistics (Gupta et al., 2018). For a vectorized layer ℓ with D_ℓ parameters, its preconditioning operator is

$$\mathbf{P}_{i,\mathrm{SOAP}}^{(\ell)} = \mathbf{Q}_{i,\ell} \operatorname{Diag}\left(\frac{1}{\sqrt{\widehat{v}_{i,\ell,1}}}, \ldots, \frac{1}{\sqrt{\widehat{v}_{i,\ell,D_\ell}}}\right) \mathbf{Q}_{i,\ell}^\top. \tag{10}$$

Here, $\mathbf{Q}_{i,\ell}$ is an orthogonal Kronecker basis, and each diagonal entry is the reciprocal square root of the bias-corrected exponential moving average $\widehat{v}_{i,\ell,k} > 0$ of the corresponding squared gradient coordinate. Under local Hessian–GN agreement, a Kronecker curvature model, and isotropic parameter deviations near a minimum, SOAP can be interpreted as a layerwise inverse-GN approximation (Wang et al., 2025). For a positive-definite residual GN block $\mathbf{B}_{i,\mathrm{GN}}^{(\ell)}$ satisfying the corresponding diagonalization and second-moment assumptions, this gives $\mathbf{P}_{i,\mathrm{SOAP}}^{(\ell)} \approx c_{i,\ell}(\mathbf{B}_{i,\mathrm{GN}}^{(\ell)})^{-1}$ with $c_{i,\ell} > 0$, providing a structured approximation to the inverse-curvature scaling underlying GN whitening.

3.3 Second-Order Multi-Objective Optimization

Multi-objective Newton methods incorporate curvature by minimizing the worst predicted change among quadratic objective models (Fliege et al., 2009). For the trial step $\boldsymbol{\theta} - \mathbf{d}$, residual linearization yields the following quadratic model of the loss change using the GN matrix in Eq. (7),

$$\frac{1}{N_i}\|\mathbf{r}_i - \mathbf{J}_i\mathbf{d}\|_2^2 - L_i = -\mathbf{g}_i^\top \mathbf{d} + \frac{1}{2}\mathbf{d}^\top \mathbf{B}_{i,\mathrm{GN}}\mathbf{d}. \tag{11}$$

Using positive-definite curvature models $\mathbf{B}_i$, the common direction solves

$$\mathbf{d}_{\mathrm{Q}} = \arg\min_{\mathbf{d}\in\mathbb{R}^D} \max_{i=1,\dots,m} \left\{ -\mathbf{g}_i^\top \mathbf{d} + \frac{1}{2}\mathbf{d}^\top \mathbf{B}_i \mathbf{d} \right\}. \tag{12}$$

Exact Newton uses $\mathbf{B}_i = \nabla^2_{\boldsymbol{\theta}} L_i$ when these Hessians are positive definite, whereas $\mathbf{B}_i = \mathbf{B}_{i,\mathrm{GN}} + \mu_i \mathbf{I}$ with $\mu_i > 0$ gives a regularized GN-type model.

A shared-metric specialization of Eq. (12), introduced by Chen et al. (2023), sets $\mathbf{B}_i = \mathbf{B} \succ 0$ for all i, yielding

$$\mathbf{d}_{\mathrm{VM}} = \arg\min_{\mathbf{d}\in\mathbb{R}^D} \max_{i=1,\dots,m} \left\{ -\mathbf{g}_i^\top \mathbf{d} + \frac{1}{2}\mathbf{d}^\top \mathbf{B}\mathbf{d} \right\}. \tag{13}$$

Here, $\|\mathbf{d}\|_{\mathbf{B}}^2 = \mathbf{d}^\top \mathbf{B}\mathbf{d}$ is the squared step norm in the shared curvature metric, replacing the Euclidean measure in Eq. (3). For $m = 1$, the solution reduces to $\mathbf{B}^{-1}\mathbf{g}_1$, recovering the single-objective preconditioned direction. For multiple objectives, Eq. (13) is a regularized max–min problem that balances the worst first-order improvement against the metric-norm penalty.

4 Methodology

We select updates by maximizing the worst relative first-order reduction in residual norm under a shared curvature-metric step budget. We characterize the solution through a *curvature-harmonized hull* that combines non-conflict geometry with curvature-aware preconditioning, connect objective progress to residual conditioning through a local mode analysis, and realize the formulation with a shared SOAP-style metric and objective-space QP.

4.1 Curvature-Harmonized Direction

We measure each objective's first-order progress relative to its current residual norm. For $L_i > 0$, define the residual-normalized gradients and relative progress by

$$\widetilde{\mathbf{g}}_i = \frac{\mathbf{g}_i}{2L_i} = \frac{\mathbf{J}_i^\top \mathbf{r}_i}{\|\mathbf{r}_i\|_2^2} \in \mathbb{R}^D, \quad \widetilde{\mathbf{G}} = [\widetilde{\mathbf{g}}_1, \dots, \widetilde{\mathbf{g}}_m] \in \mathbb{R}^{D\times m}, \quad p_i(\mathbf{d}) = \widetilde{\mathbf{g}}_i^\top \mathbf{d}. \tag{14}$$

For a small step size η, the relative residual norm satisfies $\|\mathbf{r}_i(\boldsymbol{\theta} - \eta\mathbf{d})\|_2 / \|\mathbf{r}_i\|_2 \approx 1 - \eta p_i(\mathbf{d})$. Following Eq. (13), fix a symmetric positive-definite shared curvature metric $\mathbf{B}$. For a prescribed squared step budget $q > 0$, we maximize the common progress bound γ:

$$\max_{\mathbf{d},\gamma} \ \gamma \qquad \text{s.t.} \quad \widetilde{\mathbf{G}}^\top \mathbf{d} \geq \gamma \mathbf{1}, \quad \|\mathbf{d}\|_{\mathbf{B}}^2 \leq q. \tag{15}$$

With multipliers $\boldsymbol{\alpha} \in \mathbb{R}_+^m$ and $\nu \geq 0$ for the progress and budget constraints, the Lagrangian is

$$\mathcal{L}(\mathbf{d}, \gamma, \boldsymbol{\alpha}, \nu) = \gamma + \boldsymbol{\alpha}^\top(\widetilde{\mathbf{G}}^\top \mathbf{d} - \gamma \mathbf{1}) + \frac{\nu}{2}(q - \|\mathbf{d}\|_{\mathbf{B}}^2). \tag{16}$$

Its supremum over γ is finite only if $\mathbf{1}^\top \boldsymbol{\alpha} = 1$, yielding the simplex $\Delta_m = \{\boldsymbol{\alpha} \in \mathbb{R}_+^m : \mathbf{1}^\top \boldsymbol{\alpha} = 1\}$. With $\mathbf{P} = \mathbf{B}^{-1}$, eliminating $\mathbf{d}$ and ν reduces the dual, up to a monotone transformation, to

$$\boldsymbol{\alpha}^\star \in \arg\min_{\boldsymbol{\alpha}\in\Delta_m} \frac{1}{2}\boldsymbol{\alpha}^\top \mathbf{C}\boldsymbol{\alpha}, \quad \text{where } \mathbf{C} = \widetilde{\mathbf{G}}^\top \mathbf{P} \widetilde{\mathbf{G}} \in \mathbb{R}^{m\times m}. \tag{17}$$

The metric Gram matrix $\mathbf{C}$ is positive semidefinite, making Eq. (17) a convex QP. Expressing the same combination in the original-gradient coordinates via $\lambda_i = \alpha_i/(2L_i)$, the simplex constraint becomes $\sum_i 2L_i\lambda_i = 1$.

Proposition 1 (Non-conflict at dual optimality)**.** *For any minimizer $\boldsymbol{\alpha}^\star$ of Eq.* (17)*, the corresponding combination $\mathbf{PG}\boldsymbol{\lambda}^\star$ belongs to the dual gradient cone $\mathcal{K}^* = \{\mathbf{d} \in \mathbb{R}^D : \mathbf{G}^\top \mathbf{d} \geq \mathbf{0}\}$.*

See App. B.2 for a detailed proof.

The coefficient constraint and non-conflict condition define the following curvature-aware geometry.

Definition 1 (Curvature-harmonized hull)**.** *For fixed $\mathbf{P} \succ 0$, define*

$$\mathcal{H}_{\mathbf{P}}^{\Delta} = \left\{ \mathbf{PG}\boldsymbol{\lambda} : \boldsymbol{\lambda} \in \mathbb{R}_+^m,\ \sum_{i=1}^m 2L_i\lambda_i = 1,\ \mathbf{G}^\top\mathbf{PG}\boldsymbol{\lambda} \geq \mathbf{0} \right\}. \tag{18}$$

This is the intersection of the image of the loss-weighted simplex under $\mathbf{PG}$ with $\mathcal{K}^*$. For $\mathbf{P} = \mathbf{I}$, its conic hull recovers the harmonized cone $\mathcal{H}$ in Eq. (6).

Stationarity in $\mathbf{d}$ gives $\nu^\star\mathbf{d}^\star = \mathbf{PG}\boldsymbol{\lambda}^\star$. Since the dual minimizers are independent of q, we can choose the budget so that this combination is itself the optimal direction.

Proposition 2 (Metric projection and optimal direction)**.** *Let $\boldsymbol{\alpha}^\star$ solve Eq.* (17)*. With budget $q = \boldsymbol{\alpha}^{\star\top}\mathbf{C}\boldsymbol{\alpha}^\star > 0$, the unique optimal direction of Eq.* (15) *is the $\mathbf{B}$-metric projection of the origin onto $\mathcal{H}_{\mathbf{P}}^{\Delta}$:*

$$\mathbf{d}^\star = \mathbf{PG}\boldsymbol{\lambda}^\star = \arg\min_{\mathbf{d} \in \mathcal{H}_{\mathbf{P}}^{\Delta}} \frac{1}{2}\|\mathbf{d}\|_{\mathbf{B}}^2, \qquad \gamma^\star = q. \tag{19}$$

See App. B.3 for a detailed proof.

4.2 Objective Progress and Residual-Mode Contraction

To connect the worst-objective progress in Proposition 2 to residual conditioning, we use the optimal coefficients $\boldsymbol{\alpha}^\star$ to form the joint residual and Jacobian

$$\mathbf{e}_{\boldsymbol{\alpha}} = \begin{bmatrix} \sqrt{\alpha_1^\star}\mathbf{r}_1/\|\mathbf{r}_1\|_2 \\ \vdots \\ \sqrt{\alpha_m^\star}\mathbf{r}_m/\|\mathbf{r}_m\|_2 \end{bmatrix} \in \mathbb{R}^N, \quad \mathbf{J}_{\boldsymbol{\alpha}} = \begin{bmatrix} \sqrt{\alpha_1^\star}\mathbf{J}_1/\|\mathbf{r}_1\|_2 \\ \vdots \\ \sqrt{\alpha_m^\star}\mathbf{J}_m/\|\mathbf{r}_m\|_2 \end{bmatrix} \in \mathbb{R}^{N\times D}, \quad \mathbf{K}_{\boldsymbol{\alpha},\mathbf{P}} = \mathbf{J}_{\boldsymbol{\alpha}}\mathbf{P}\mathbf{J}_{\boldsymbol{\alpha}}^\top, \tag{20}$$

where $N = \sum_i N_i$. The joint residual kernel $\mathbf{K}_{\boldsymbol{\alpha},\mathbf{P}} \in \mathbb{R}^{N\times N}$ couples the objectives through its off-diagonal blocks, extending the single-objective kernels in Sec. 3.2.

Since $\|\mathbf{e}_{\boldsymbol{\alpha}}\|_2^2 = 1$ and $\mathbf{J}_{\boldsymbol{\alpha}}^\top\mathbf{e}_{\boldsymbol{\alpha}} = \widetilde{\mathbf{G}}\boldsymbol{\alpha}^\star$, the selected direction is $\mathbf{d}^\star = \mathbf{P}\mathbf{J}_{\boldsymbol{\alpha}}^\top\mathbf{e}_{\boldsymbol{\alpha}}$. Writing $\mathbf{e}_{\boldsymbol{\alpha}}^+$ for the same rescaled residual after the parameter step gives

$$\mathbf{e}_{\boldsymbol{\alpha}}^+ \approx (\mathbf{I} - \eta\mathbf{K}_{\boldsymbol{\alpha},\mathbf{P}})\mathbf{e}_{\boldsymbol{\alpha}}, \qquad \frac{\mathbf{e}_{\boldsymbol{\alpha}}^\top\mathbf{K}_{\boldsymbol{\alpha},\mathbf{P}}\mathbf{e}_{\boldsymbol{\alpha}}}{\mathbf{e}_{\boldsymbol{\alpha}}^\top\mathbf{e}_{\boldsymbol{\alpha}}} = q = \gamma^\star. \tag{21}$$

Here, $\gamma^\star = \min_i p_i(\mathbf{d}^\star)$ is the worst relative first-order residual reduction defined by Eqs. (14)–(15). Thus, under the matched budget of Proposition 2, the worst-objective progress is exactly a Rayleigh quotient of the joint residual kernel.

Theorem 1 (Objective progress and local residual-mode contraction)**.** *Let $\mathbf{P} \succ 0$ and $\boldsymbol{\alpha}^\star$ solve Eq.* (17) *with the matched budget $q > 0$ in Proposition 2. Denote the largest and smallest positive eigenvalues of $\mathbf{K}_{\boldsymbol{\alpha},\mathbf{P}}$ by $\lambda_{\max}$ and $\lambda_{\min}^+$, and define its effective condition number as $\kappa_{\text{eff}} = \lambda_{\max}/\lambda_{\min}^+$. For a unit eigenvector $\mathbf{u}_k$ of $\mathbf{K}_{\boldsymbol{\alpha},\mathbf{P}}$ with eigenvalue $\lambda_k > 0$, let $a_k = \mathbf{u}_k^\top\mathbf{e}_{\boldsymbol{\alpha}}$ and $a_k^+ = \mathbf{u}_k^\top\mathbf{e}_{\boldsymbol{\alpha}}^+$. In the local linearized model of Eq.* (21)*, for $0 < \eta \leq 1/\lambda_{\max}$,*

$$|a_k^+| \leq \left(1 - \eta\frac{q}{\kappa_{\text{eff}}}\right)|a_k|. \tag{22}$$

See App. B.4 for a detailed proof.

The bound links the worst-objective progress $q = \gamma^\star$ with the residual conditioning κ_{eff} studied from the single-objective perspective in Sec. 3.2. A positive q ensures first-order progress across objectives, whereas a large κ_{eff} weakens the modewise contraction bound; objective-wise progress and residual conditioning must therefore be considered jointly. Complementary PINN diagnostics in App. B.5 examine both effects empirically. Metric-aware selection avoids post-hoc first-order conflicts, while the shared metric reduces residual-kernel spectral disparity.

4.3 CURVATURE-HARMONIZED METRIC DESCENT

Having derived the curvature-harmonized direction and its residual-mode interpretation, we now realize it with a shared SOAP-style metric (Vyas et al., 2025). Algorithm 1 summarizes a CHIME update after initialization. For layer ℓ, write its parameter block as $\mathbf{\Theta}_t^{(\ell)} \in \mathbb{R}^{d_\ell^{\mathrm{r}} \times d_\ell^{\mathrm{c}}}$, with $D_\ell = d_\ell^{\mathrm{r}} d_\ell^{\mathrm{c}}$ and $\sum_\ell D_\ell = D$. For thresholds $\tau_i \geq 0$, let $\mathcal{A}_t = \{i : L_i(\boldsymbol{\theta}_t) > \tau_i\}$ denote the active-objective index set and define

$$\mathbf{H}_{i,t}^{(\ell)} = \frac{\nabla_{\mathbf{\Theta}^{(\ell)}} L_i(\boldsymbol{\theta}_t)}{2\big(L_i(\boldsymbol{\theta}_t) - \tau_i\big)}, \qquad i \in \mathcal{A}_t. \quad (23)$$

For $\tau_i = 0$, these are the layer blocks of $\widetilde{\mathbf{g}}_{i,t}$.

For each layer, $\mathbf{F}_{\mathrm{L}} \in \mathbb{R}^{d_\ell^{\mathrm{r}} \times d_\ell^{\mathrm{r}}}$ and $\mathbf{F}_{\mathrm{R}} \in \mathbb{R}^{d_\ell^{\mathrm{c}} \times d_\ell^{\mathrm{c}}}$ are the shared row and column Gram statistics, with orthogonal eigenvector bases $\mathbf{Q}_{\mathrm{L}}$ and $\mathbf{Q}_{\mathrm{R}}$, respectively. The first- and second-moment estimates $\mathbf{M}, \mathbf{V} \in \mathbb{R}^{d_\ell^{\mathrm{r}} \times d_\ell^{\mathrm{c}}}$ are maintained in this shared eigenbasis. To avoid a circular dependence between metric estimation and current coefficient selection, CHIME constructs the shared metric using lagged weights $\boldsymbol{\omega}_t$. These weights are obtained by restricting the objective coefficients $\boldsymbol{\alpha}_{t-1}$ selected at the previous iteration to $\mathcal{A}_t$ and renormalizing them. Their weighted aggregate $\mathbf{A} = \sum_{i \in \mathcal{A}_t} \omega_{t,i} \mathbf{H}_{i,t}^{(\ell)}$ updates $\mathbf{F}_{\mathrm{L}}$ and $\mathbf{F}_{\mathrm{R}}$. Its coordinates in the current shared eigenbasis, $\mathbf{Z}_{\mathrm{lag}} = \mathbf{Q}_{\mathrm{L}}^\top \mathbf{A} \mathbf{Q}_{\mathrm{R}}$, update the second moment $\mathbf{V}$ through $\mathbf{Z}_{\mathrm{lag}} \odot \mathbf{Z}_{\mathrm{lag}}$, where $\odot$ denotes entrywise multiplication.

Algorithm 1 CHIME

Input: $\{\mathbf{\Theta}_t^{(\ell)}\}_\ell$, $\{\mathbf{H}_{i,t}^{(\ell)}\}_{i \in \mathcal{A}_t, \ell}$, $\boldsymbol{\alpha}_{t-1}$.
Output: $\{\mathbf{\Theta}_{t+1}^{(\ell)}\}_\ell$, $\boldsymbol{\alpha}_t$.

$\omega_{t,i} \leftarrow \dfrac{\alpha_{t-1,i}}{\sum_{j \in \mathcal{A}_t} \alpha_{t-1,j}}, \quad i \in \mathcal{A}_t.$
for each layer block ℓ **do**
  $\mathbf{A} \leftarrow \sum_{i \in \mathcal{A}_t} \omega_{t,i} \mathbf{H}_{i,t}^{(\ell)}.$
  $\mathbf{F}_{\mathrm{L}} \leftarrow \beta_3 \mathbf{F}_{\mathrm{L}} + (1-\beta_3) \mathbf{A}\mathbf{A}^\top.$
  $\mathbf{F}_{\mathrm{R}} \leftarrow \beta_3 \mathbf{F}_{\mathrm{R}} + (1-\beta_3) \mathbf{A}^\top \mathbf{A}.$
  $\mathbf{Z}_{\mathrm{lag}} \leftarrow \mathbf{Q}_{\mathrm{L}}^\top \mathbf{A} \mathbf{Q}_{\mathrm{R}}.$
  $\mathbf{V} \leftarrow \beta_2 \mathbf{V} + (1-\beta_2)(\mathbf{Z}_{\mathrm{lag}} \odot \mathbf{Z}_{\mathrm{lag}}).$
  if $t \bmod f = 0$ **then**
    $\mathbf{Q}_{\mathrm{L}} \leftarrow \mathrm{Eigenvectors}(\mathbf{F}_{\mathrm{L}}, \mathbf{Q}_{\mathrm{L}}).$
    $\mathbf{Q}_{\mathrm{R}} \leftarrow \mathrm{Eigenvectors}(\mathbf{F}_{\mathrm{R}}, \mathbf{Q}_{\mathrm{R}}).$
    Transport $\mathbf{M}$ and $\mathbf{V}$ using Eq. (95).
  end if
  $\mathbf{Z}_{i,t}^{(\ell)} \leftarrow \mathbf{Q}_{\mathrm{L}}^\top \mathbf{H}_{i,t}^{(\ell)} \mathbf{Q}_{\mathrm{R}}, \; i \in \mathcal{A}_t.$
end for
Assemble $\mathbf{C}_t \in \mathbb{R}^{|\mathcal{A}_t| \times |\mathcal{A}_t|}$ using Eq. (24).
$\boldsymbol{\alpha}_{t,\mathcal{A}_t} \leftarrow \arg\min_{\boldsymbol{\alpha} \in \Delta_{|\mathcal{A}_t|}} \dfrac{1}{2} \boldsymbol{\alpha}^\top \mathbf{C}_t \boldsymbol{\alpha}.$
for each layer block ℓ **do**
  $\mathbf{M} \leftarrow \beta_1 \mathbf{M} + (1-\beta_1) \sum_{i \in \mathcal{A}_t} \alpha_{t,i} \mathbf{Z}_{i,t}^{(\ell)}.$
  $\widehat{\mathbf{M}} \leftarrow \mathbf{M}/(1-\beta_1^t), \; \widehat{\mathbf{V}} \leftarrow \mathbf{V}/(1-\beta_2^t).$
  $\mathbf{\Theta}_{t+1}^{(\ell)} \leftarrow \mathbf{\Theta}_t^{(\ell)} - \eta_t \mathbf{Q}_{\mathrm{L}} \left[\widehat{\mathbf{M}}/\sqrt{\widehat{\mathbf{V}}}\right] \mathbf{Q}_{\mathrm{R}}^\top.$
end for

Every f iterations, $\mathbf{Q}_{\mathrm{L}}$ and $\mathbf{Q}_{\mathrm{R}}$ are refreshed from $\mathbf{F}_{\mathrm{L}}$ and $\mathbf{F}_{\mathrm{R}}$, respectively. With the shared metric fixed, the current active-objective blocks are separately rotated as

$$\mathbf{Z}_{i,t}^{(\ell)} = \mathbf{Q}_{\mathrm{L}}^\top \mathbf{H}_{i,t}^{(\ell)} \mathbf{Q}_{\mathrm{R}}, \qquad i \in \mathcal{A}_t.$$

The blocks $\mathbf{Z}_{i,t}^{(\ell)}$ retain the individual objectives for current coefficient selection. Let $\mathbf{C}_t \in \mathbb{R}^{|\mathcal{A}_t| \times |\mathcal{A}_t|}$ denote the resulting active-objective Gram, with entries

$$(C_t)_{ij} = \sum_\ell \left\langle \mathbf{Z}_{i,t}^{(\ell)}, \frac{\mathbf{Z}_{j,t}^{(\ell)}}{\sqrt{\mathbf{V}_t^{(\ell)}/(1-\beta_2^t)}} \right\rangle, \qquad i, j \in \mathcal{A}_t, \quad (24)$$

where $\langle \cdot, \cdot \rangle$ is the Frobenius inner product, and the square root and division are entrywise. When $\tau_i = 0$ for all i, $\mathbf{C}_t$ is the structured realization of $\widetilde{\mathbf{G}}_t^\top \mathbf{P}_t \widetilde{\mathbf{G}}_t$ in Eq. (17); more generally, the same shared metric acts on the active normalized blocks in Eq. (23). Solving the simplex QP over $\mathbf{C}_t$ yields $\boldsymbol{\alpha}_{t,\mathcal{A}_t}$, which is shared across all layer blocks.

Finally, the selected coefficients aggregate the rotated objective gradients into the first moment $\mathbf{M}$, and the parameter update applies the bias-corrected first and second moments in the same shared eigenbasis. With one active objective, CHIME reduces to SOAP applied to its normalized gradient; with multiple active objectives, the same metric additionally determines their objective-space combination. Further implementation details are provided in App. B.6.

5 EXPERIMENTS

We primarily evaluate CHIME on PINN benchmarks and additionally examine whether its benefits extend to standard MTL benchmarks. For PINNs, we use three forward PDE benchmarks (Wave1d-C,

Table 2: Relative L^2 errors on four forward problems and the inverse problem HInv; lower is better. Entries are means (standard deviations) over five random seeds. Best and second-best displayed means are shown in **bold** and underline.

	Forward				Inverse
Method	**Wave1d-C**	**Poisson2d-C**	**HNd**	**Volterra1d**	**HInv**
MultiAdam	1.0958 (0.1004)	0.6733 (0.0820)	0.0040 (0.0028)	0.0004 (0.0002)	0.1446 (0.0245)
LRA	0.2608 (0.0542)	0.5712 (0.1595)	0.0020 (0.0011)	0.0070 (0.0065)	0.0546 (0.0131)
ReLoBRaLo	0.3542 (0.0089)	0.6602 (0.0221)	0.0004 (0.0000)	0.0002 (0.0000)	1.5999 (0.0576)
MGDA	0.3995 (0.0135)	0.6777 (0.0239)	0.0009 (0.0001)	0.0011 (0.0019)	2.4224 (1.6765)
PCGrad	0.2256 (0.0701)	0.6918 (0.0190)	0.0005 (0.0002)	0.0004 (0.0004)	0.0660 (0.0152)
CAGrad	0.1186 (0.0071)	0.6969 (0.0197)	0.0005 (0.0001)	0.0014 (0.0024)	0.6435 (1.0862)
IMTL-G	0.2724 (0.3372)	0.5039 (0.0684)	0.0006 (0.0001)	0.0735 (0.0989)	0.0498 (0.0075)
Aligned-MTL	0.2302 (0.0514)	0.2847 (0.2363)	0.0006 (0.0002)	0.0037 (0.0044)	0.0804 (0.0457)
DCGD	0.4022 (0.0309)	0.4227 (0.3284)	0.0004 (0.0000)	0.0007 (0.0008)	0.0466 (0.0112)
ConFIG	0.0668 (0.0279)	0.6856 (0.0274)	0.0006 (0.0001)	0.0018 (0.0024)	0.0466 (0.0068)
HARMONIC	0.0655 (0.0293)	0.0214 (0.0179)	0.0005 (0.0000)	0.0003 (0.0001)	**0.0461** (**0.0098**)
KFAC*	0.5886 (0.0221)	0.6898 (0.0186)	0.0006 (0.0001)	0.0002 (0.0000)	0.3573 (0.0188)
Muon*	0.0229 (0.0011)	0.6627 (0.0200)	0.0002 (0.0000)	0.0006 (0.0003)	0.0633 (0.0121)
SOAP*	0.0064 (0.0015)	0.6592 (0.0170)	**0.0001** (**0.0000**)	0.0002 (0.0002)	0.0837 (0.0144)
CHIME	**0.0033** (**0.0010**)	**0.0029** (**0.0006**)	**0.0001** (**0.0001**)	**0.0001** (**0.0001**)	0.0636 (0.0167)

Poisson2d-C, and HNd) and one inverse PDE benchmark (HInv) from PINNacle (Hao et al., 2024), together with the forward integro-differential equation benchmark Volterra1d from A-PINN (Yuan et al., 2022). Wave1d-C and Volterra1d each contain four loss terms, while the remaining problems contain three. We follow the benchmark protocol of Bu et al. (2026), including the network architectures, training budget, and the same five seeds. Where available, objective-aware baseline results are taken from that study; the remaining evaluations use the same benchmark protocol. Objective-aware baselines include MultiAdam (Yao et al., 2023), LRA (Wang et al., 2021), ReLoBRaLo (Bischof and Kraus, 2025), MGDA (Sener and Koltun, 2018), PCGrad (Yu et al., 2020), CAGrad (Liu et al., 2021a), IMTL-G (Liu et al., 2021b), Aligned-MTL (Senushkin et al., 2023), DCGD (Hwang and Lim, 2024), ConFIG (Liu et al., 2025), and HARMONIC (Bu et al., 2026); curvature-aware baselines are KFAC* (Dangel et al., 2024), Muon*, and SOAP* (Wang et al., 2025). For CHIME, all forward objectives use $\tau_i = 0$, while the noisy observation objective in HInv uses $\tau_{\text{data}} = 0.01$. We report relative L^2 errors over the five seeds; detailed settings are given in App. C.

Table 2 shows that CHIME attains the best or tied-best mean error on all four forward problems. The two most diagnostic cases reflect the complementary strengths targeted by our formulation. On Wave1d-C, SOAP* is the strongest baseline at 0.0064, while CHIME further reduces the error to 0.0033; on Poisson2d-C, HARMONIC is the strongest baseline at 0.0214, while CHIME reaches 0.0029. Across the four forward problems, the means correspond to geometric-mean error-reduction factors of $5.85\times$ and $5.45\times$ relative to the best objective-aware and curvature-aware baseline on each problem, respectively. On HInv, DCGD, ConFIG, and HARMONIC achieve lower mean coefficient-recovery errors than CHIME. App. C.5.1 shows that the curvature-aware baselines Muon* and SOAP* exhibit overfitting to noisy observations, with decreasing data loss despite worsening inverse accuracy. A mini-batch ablation in App. C.5.2 further shows that using smaller observation batches mitigates this late-stage degradation. These results suggest that curvature-aware optimization for noisy inverse problems benefits from training strategies that control observation fitting. Additional PINN benchmarks, equal-wall-clock and computational costs, and learning-rate sensitivity are reported in Apps. C–E.

Multi-task learning. To assess applicability beyond PINNs, we use NYU-v2 (Silberman et al., 2012) for joint semantic segmentation, depth estimation, and surface-normal prediction. This benchmark exhibits substantial task imbalance, which Qin et al. (2025) analyze through task-specific NTK spectra and convergence rates. We follow their NYU-v2 training protocol and use their reported baseline results, with all methods averaged over three seeds. The comparison includes LS and SI; RLW (Lin et al., 2022), DWA (Liu et al., 2019), and UW (Kendall et al., 2018); MGDA (Sener and Koltun, 2018), PCGrad (Yu et al., 2020), GradDrop (Chen et al., 2020), CAGrad (Liu et al., 2021a), IMTL-G (Liu et al., 2021b), MoCo (Fernando et al., 2023), Nash-MTL (Navon et al., 2022), Aligned-MTL (Senushkin et al., 2023), FAMO (Liu et al., 2023), SDMGrad (Xiao et al., 2023), DB-MTL (Lin et al., 2026), STCH (Lin et al., 2024), FairGrad (Ban and Ji, 2024), GO4Align (Shen et al., 2024), and NTKMTL (Qin et al., 2025). Following Qin et al. (2025), for metric j we measure

Table 3: Results on NYU-v2 (Silberman et al., 2012) (three tasks), averaged over three random seeds. Baseline results follow Qin et al. (2025). The best scores are highlighted in **gray**, and the second-best scores are underlined.

Method	Segmentation		Depth		Surface Normal					MR ↓	$\Delta_m\%$ ↓
					Angle Distance ↓		Within t° ↑				
	mIoU ↑	Pix Acc ↑	Abs Err ↓	Rel Err ↓	Mean	Median	11.25	22.5	30		
STL	38.30	63.76	0.6754	0.2780	25.01	19.21	30.14	57.20	69.15	–	–
LS	39.29	65.33	0.5493	0.2263	28.15	23.96	22.09	47.50	61.08	17.56	5.59
SI	38.45	64.27	0.5354	0.2201	27.60	23.37	22.53	48.57	62.32	16.22	4.39
RLW	37.17	63.77	0.5759	0.2410	28.27	24.18	22.26	47.05	60.62	20.22	7.78
DWA	39.11	65.31	0.5510	0.2285	27.61	23.18	24.17	50.18	62.39	16.33	3.57
UW	36.87	63.17	0.5446	0.2260	27.04	22.61	23.54	49.05	63.65	16.00	4.05
MGDA	30.47	59.90	0.6070	0.2555	24.88	19.45	29.18	56.88	69.36	11.89	1.38
PCGrad	38.06	64.64	0.5550	0.2325	27.41	22.80	23.86	49.83	63.14	16.89	3.97
GradDrop	39.39	65.12	0.5455	0.2279	27.48	22.96	23.38	49.44	62.87	15.56	3.58
CAGrad	39.79	65.49	0.5486	0.2250	26.31	21.58	25.61	52.36	65.58	11.61	0.20
IMTL-G	39.35	65.60	0.5426	0.2256	26.02	21.19	26.20	53.13	66.24	11.00	-0.76
MoCo	40.30	66.07	0.5575	0.2135	26.67	21.83	25.61	51.78	64.85	10.89	0.16
Nash-MTL	40.13	65.93	0.5261	0.2171	25.26	20.08	28.40	55.47	68.15	8.00	-4.04
Aligned-MTL	40.15	66.05	0.5520	0.2291	25.37	19.89	28.30	55.29	67.95	10.33	-3.12
FAMO	38.88	64.90	0.5474	0.2194	25.06	19.57	29.21	56.61	68.98	9.11	-4.10
SDMGrad	40.47	65.90	0.5225	0.2084	25.07	19.99	28.54	55.74	68.53	6.11	-4.84
DB-MTL	**41.42**	**66.45**	0.5251	0.2160	25.03	19.50	28.72	56.17	68.73	4.78	-5.36
STCH	41.35	66.07	**0.4965**	**0.2010**	26.55	21.81	24.84	51.39	64.86	8.17	-1.35
FairGrad	39.74	66.01	0.5377	0.2236	24.84	19.60	29.26	56.58	69.16	6.56	-4.66
GO4Align	40.42	65.37	0.5492	0.2167	24.76	18.94	30.54	57.87	69.84	5.44	-6.08
NTKMTL	39.68	65.43	0.5296	0.2168	**24.24**	**18.63**	30.74	**58.72**	**70.78**	4.56	-6.99
CHIME	40.12	65.71	0.5240	0.2088	24.87	18.70	**31.76**	58.19	69.76	**3.78**	**-7.37**

the relative change from the single-task reference and average over the $S = 9$ reported metrics,

$$\Delta_m\% = \frac{100}{S}\sum_{j=1}^{S}(-1)^{\delta_j}\frac{M_j - M_{\mathrm{STL},j}}{M_{\mathrm{STL},j}}, \tag{25}$$

where $\delta_j = 1$ for higher-is-better metrics and 0 otherwise; lower $\Delta_m\%$ is better. Mean rank (MR) averages the metric-wise ranks, with lower values again better. Table 3 shows that CHIME achieves the best aggregate performance among the compared methods, with $\mathrm{MR} = 3.78$ and $\Delta_m\% = -7.37$. We additionally evaluate scalability to 40 objectives on CelebA (Liu et al., 2015); detailed results and seed variability for both MTL benchmarks are provided in App. F.

6 Conclusion

We introduced CHIME, which selects objective combinations in the same curvature geometry used for the parameter update. Its max–min formulation yields a curvature-harmonized direction with a first-order non-conflict guarantee, while the local analysis connects worst-objective progress, residual conditioning, and modewise contraction. A shared SOAP-style realization reduces direction selection to a small objective-space QP. Empirically, CHIME improves accuracy across the four forward PINN benchmarks and achieves the best aggregate performance in the MTL comparisons.

The inverse-PDE results further indicate that curvature-aware optimization can be vulnerable to noisy objectives, where faster loss reduction need not translate into better recovery. Our HInv mini-batch ablation shows that stochastic observation sampling can delay and mitigate this degradation. Recent analyses of preconditioned training further suggest that reducing spectral bias can accelerate modes more uniformly while altering the implicit bias relevant to generalization (Jiang et al., 2026), motivating noise-aware training strategies for curvature-aware optimization. Complementarily, Adam's gradient mean–variance interpretation suggests that gradient statistics may provide useful signals for such noise-aware adaptation (Orvieto and Gower, 2025). Another promising direction is to develop curvature approximations that are both more faithful and computationally efficient, capturing objective-specific or cross-layer structure without sacrificing the scalability of the shared SOAP-style metric.

REFERENCES

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

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.

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.

Atilim Gunes Baydin, Barak A Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. Automatic differentiation in machine learning: a survey. *Journal of machine learning research*, 18 (153):1–43, 2018.

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.

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.

Yiming Meng, Ruikun Zhou, Amartya Mukherjee, Maxwell Fitzsimmons, Christopher Song, and Jun Liu. Physics-informed neural network policy iteration: Algorithms, convergence, and verification. In *Forty-first International Conference on Machine Learning*, 2024.

Pratik Rathore, Weimu Lei, Zachary Frangella, Lu Lu, and Madeleine Udell. Challenges in training PINNs: A loss landscape perspective. In *Forty-first International Conference on Machine Learning*, 2024.

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.

Youngsik Hwang and Dongyoung Lim. Dual cone gradient descent for training physics-informed neural networks. *Advances in Neural Information Processing Systems*, 37:98563–98595, 2024.

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.

Tim De Ryck, Florent Bonnet, Siddhartha Mishra, and Emmanuel de Bezenac. An operator preconditioning perspective on training in physics-informed machine learning. In *The Twelfth International Conference on Learning Representations*, 2024.

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, 2022.

Rafael Bischof and Michael A Kraus. Multi-objective loss balancing for physics-informed deep learning. *Computer Methods in Applied Mechanics and Engineering*, 439:117914, 2025.

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.

Johannes Müller and Marius Zeinhofer. Achieving high accuracy with PINNs via energy natural gradient descent. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 25471–25485. PMLR, 23–29 Jul 2023.

Nikhil Vyas, Depen Morwani, Rosie Zhao, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. Soap: Improving and stabilizing shampoo using adam for language modeling. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu, editors, *International Conference on Learning Representations*, volume 2025, pages 93423–93444, 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*, 2025.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. *Advances in Neural Information Processing Systems*, 34:18878–18890, 2021a.

Zhongkai Hao, Jiachen Yao, Chang Su, Hang Su, Ziao Wang, Fanzhi Lu, Zeyu Xia, Yichi Zhang, Songming Liu, Lu Lu, and Jun Zhu. Pinnacle: A comprehensive benchmark of physics-informed neural networks for solving pdes. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 76721–76774. Curran Associates, Inc., 2024. doi: 10.52202/079017-2442.

Lei Yuan, Yi-Qing Ni, Xiang-Yun Deng, and Shuo Hao. A-pinn: Auxiliary physics informed neural networks for forward and inverse problems of nonlinear integro-differential equations. *Journal of Computational Physics*, 462:111260, 2022.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, pages 794–803. PMLR, 2018.

Xiaohan Qin, Xiaoxing Wang, Ning Liao, and Junchi Yan. NTKMTL: Mitigating task imbalance in multi-task learning from neural tangent kernel perspective. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1612.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33: 5824–5836, 2020.

Aviv Navon, Aviv Shamsian, Idan Achituve, Haggai Maron, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. Multi-task learning as a bargaining game. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 16428–16446. PMLR, 2022.

Dmitry Senushkin, Nikolay Patakin, Arseny Kuznetsov, and Anton Konushin. Independent component alignment for multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20083–20093, 2023.

Jiachen Yao, Chang Su, Zhongkai Hao, Songming Liu, Hang Su, and Jun Zhu. Multiadam: Parameter-wise scale-invariant optimizer for multiscale training of physics-informed neural networks. In *International Conference on Machine Learning*, pages 39702–39721. PMLR, 2023.

Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1):503–528, 1989.

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International conference on machine learning*, pages 2408–2417. PMLR, 2015.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1842–1850. PMLR, 10–15 Jul 2018.

Maricela Best Mckay, Avleen Kaur, Chen Greif, and Brian Wetton. Near-optimal sketchy natural gradients for physics-informed neural networks. In *Forty-second International Conference on Machine Learning*, 2025.

Felix Dangel, Johannes Müller, and Marius Zeinhofer. Kronecker-factored approximate curvature for physics-informed neural networks. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Andrea Bonfanti, Giuseppe Bruno, and Cristina Cipriani. The challenges of the nonlinear regime for physics-informed neural networks. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Shuai Jiang, Eric C Cyr, Ben S. Southworth, and Alexey Voronin. On the convergence behavior of preconditioned gradient descent toward the rich learning regime. In *The Fourteenth International Conference on Learning Representations*, 2026.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Joerg Fliege, LM Grana Drummond, and Benar Fux Svaiter. Newton's method for multiobjective optimization. *SIAM Journal on Optimization*, 20(2):602–626, 2009.

Jian Chen, Gao-Xi Li, and Xin-Min Yang. Variable metric method for unconstrained multiobjective optimization problems. *Journal of the Operations Research Society of China*, 11(3):409–438, Sep 2023. ISSN 2194-6698. doi: 10.1007/s40305-022-00447-z.

Liyang Liu, Yi Li, Zhanghui Kuang, Jing-Hao Xue, Yimin Chen, Wenming Yang, Qingmin Liao, and Wayne Zhang. Towards impartial multi-task learning. In *International Conference on Learning Representations*, 2021b.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *Computer Vision – ECCV 2012*, volume 7576 of *Lecture Notes in Computer Science*, pages 746–760. Springer, 2012. doi: 10.1007/978-3-642-33715-4_54.

Baijiong Lin, Feiyang Ye, and Yu Zhang. A closer look at loss weighting in multi-task learning, 2022.

Shikun Liu, Edward Johns, and Andrew J Davison. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1871–1880, 2019.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7482–7491, 2018. doi: 10.1109/CVPR.2018.00781.

Zhao Chen, Jiquan Ngiam, Yanping Huang, Thang Luong, Henrik Kretzschmar, Yuning Chai, and Dragomir Anguelov. Just pick a sign: Optimizing deep multitask models with gradient sign dropout. In *Advances in Neural Information Processing Systems*, volume 33, pages 2039–2050, 2020.

Heshan Devaka Fernando, Han Shen, Miao Liu, Subhajit Chaudhury, Keerthiram Murugesan, and Tianyi Chen. Mitigating gradient bias in multi-objective learning: A provably convergent approach. In *The Eleventh International Conference on Learning Representations*, 2023.

Bo Liu, Yihao Feng, Peter Stone, and Qiang Liu. FAMO: Fast adaptive multitask optimization. In *Advances in Neural Information Processing Systems*, volume 36, pages 57226–57243, 2023. doi: 10.52202/075280-2500.

Peiyao Xiao, Hao Ban, and Kaiyi Ji. Direction-oriented multi-objective learning: Simple and provable stochastic algorithms. In *Advances in Neural Information Processing Systems*, volume 36, pages 4509–4533, 2023. doi: 10.52202/075280-0200.

Baijiong Lin, Weisen Jiang, Feiyang Ye, Yu Zhang, Pengguang Chen, Ying-Cong Chen, Shu Liu, Ivor W. Tsang, and James T. Kwok. Dual-balancing for multi-task learning. *Neural Networks*, 195: 108317, 2026. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2025.108317.

Xi Lin, Xiaoyuan Zhang, Zhiyuan Yang, Fei Liu, Zhenkun Wang, and Qingfu Zhang. Smooth tchebycheff scalarization for multi-objective optimization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 30479–30509. PMLR, 2024.

Hao Ban and Kaiyi Ji. Fair resource allocation in multi-task learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 2715–2731. PMLR, 2024.

Jiayi Shen, Qi (Cheems) Wang, Zehao Xiao, Nanne Van Noord, and Marcel Worring. GO4Align: Group optimization for multi-task alignment. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3537.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738, 2015. doi: 10.1109/ICCV.2015.425.

Antonio Orvieto and Robert M. Gower. In search of adam's secret sauce. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

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.

Davide Buffelli, Jamie McGowan, Wangkun Xu, Alexandru Cioba, Da-shan Shiu, Guillaume Hennequin, and Alberto Bernacchia. Exact, tractable gauss-newton optimization in deep reversible architectures reveal poor generalization. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 133541–133570. Curran Associates, Inc., 2024. doi: 10.52202/079017-4245.

Semih Cayci. Non-asymptotic generalization and optimization bounds for stochastic gauss-newton in deep neural networks. In Emtiyaz Khan, Yingzhen Li, Arno Solin, and Aaditya Ramdas, editors, *Proceedings of The 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pages 1549–1557. PMLR, 02–05 May 2026.

A TOY EXPERIMENTS FOR SEC. 1

The three examples in Fig. 1 modify the two-objective landscape used by CAGrad and MoCo (Liu et al., 2021a; Fernando et al., 2023). We give the objective construction, three-dimensional landscapes, optimization trajectories, and distance and mode diagnostics for each case.

A.1 OBJECTIVE CONSTRUCTION AND SETTINGS

Let $\boldsymbol{q} = (x, y)^\top$ and define

$$b_1 = -\tfrac{1}{2}(x+7) + \tanh y, \qquad b_2 = -\tfrac{1}{2}x + \tfrac{7}{2} - \tanh y,$$
$$c_1 = \tfrac{1}{10}\big[(x-7)^2 + 0.1(y+8)^2\big] - 20, \qquad c_2 = \tfrac{1}{10}\big[(x+7)^2 + 0.1(y+8)^2\big] - 20. \tag{26}$$

For transition width $w > 0$, floor $\delta > 0$, smoothing $\rho \geq 0$, and quadratic weight $\kappa \geq 0$, set

$$h_\rho(b) = \begin{cases} \log\max\{|b|, \delta\}, & \rho = 0, \\ \frac{1}{2}\log(b^2 + \rho^2), & \rho > 0, \end{cases} \tag{27}$$

$$f_i(\boldsymbol{q}) = \big[h_\rho(b_i) + 6 + \kappa b_i^2\big]\max\{\tanh(y/w), 0\} + c_i \max\{-\tanh(y/w), 0\}. \tag{28}$$

For a global scale $\gamma > 0$, anisotropy factor $s > 0$, and rotation angle φ, the optimized coordinates $\boldsymbol{\theta}$ satisfy

$$\boldsymbol{q} = T\boldsymbol{\theta}, \qquad T = \gamma\,\mathrm{diag}(s, 1)R(\varphi), \qquad R(\varphi) = \begin{bmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{bmatrix}. \tag{29}$$

Unequal axis scaling introduces anisotropy; rotation changes its orientation.

Optimization. All methods are run on CPU using 64-bit floating-point arithmetic and a learning rate of 10^{-4}. All runs are performed for $170{,}000$ steps. SOAP* follows the default SOAP-with-LRA setting of Wang et al. (2025): the LRA loss weights are updated every $1{,}000$ steps with momentum 0.9, while the SOAP basis is updated every 2 steps. HARMONIC first constructs its multi-objective direction and then applies Adam-style preconditioning (Bu et al., 2026); in the single-objective case, HARMONIC reduces to Adam and SOAP* reduces to SOAP.

Reference geometry and diagnostics. Let $\mathcal{P}_w$ denote the lower-region Pareto curve of the base objectives at transition width w, obtained by minimizing their weighted sums in $y < 0$. The local Hessian condition number κ_{ref} is the ratio of the largest to the smallest eigenvalue at the scalar optimum in Case 1 or the symmetric Pareto reference point in Cases 2–3. The Hessian is evaluated in optimization coordinates, using the scalar objective or the unweighted component mean, respectively.

Distance curves measure Euclidean distance to the optimum or the numerical Pareto curve in the original $\boldsymbol{q} = (x, y)^\top$ coordinates. For the mode diagnostics, let v_{weak} and v_{strong} be fixed eigenvectors of the reference Hessian associated with its smaller and larger eigenvalues. At each iterate, we project the parameter error relative to the nearest target point in optimization coordinates onto these two vectors. For a mode v, the displayed error for initialization j is

$$E_v^{(j)}(t) = \frac{|v^\top(\boldsymbol{\theta}_t^{(j)} - \boldsymbol{\theta}_{\mathrm{tar},t}^{(j)})|}{\max_{k\in\{1,2\}} |v^\top(\boldsymbol{\theta}_0^{(k)} - \boldsymbol{\theta}_{\mathrm{tar},0}^{(k)})|}. \tag{30}$$

These fixed-reference projections distinguish persistent directional errors. Filled and open markers distinguish the first and second initializations.

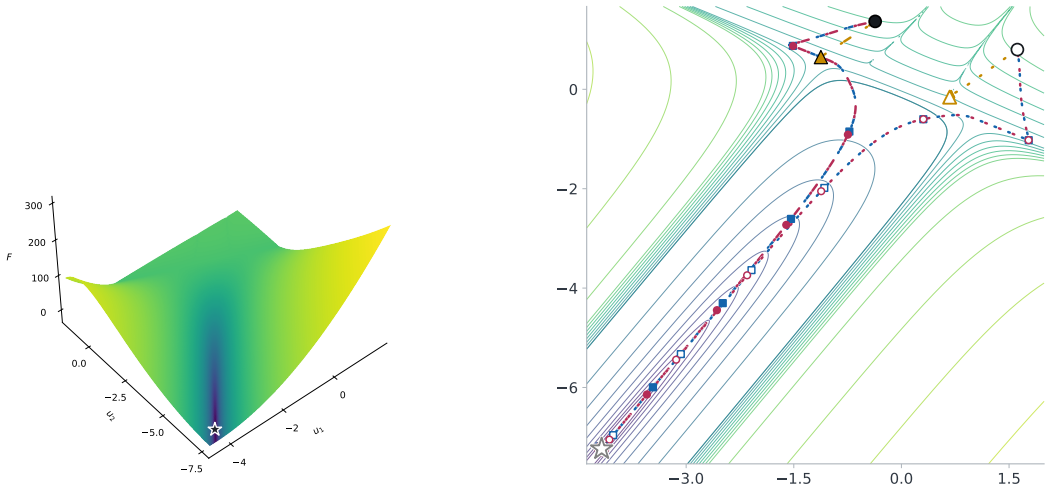
A.2 Case 1: Ill-conditioned single-objective optimization

Use $(w, \delta, \rho, \kappa) = (2, 5 \times 10^{-6}, 0, 0)$ and $T_1 = \mathrm{diag}(10, 1)R(30^\circ)$. Define the scalar landscape F and training objective L by

$$F(\boldsymbol{\theta}) = \frac{f_1(T_1\boldsymbol{\theta}) + f_2(T_1\boldsymbol{\theta})}{2}, \qquad L(\boldsymbol{\theta}) = F(\boldsymbol{\theta}) + 20. \tag{31}$$

The first and second initializations in original coordinates are $(-10, 1)^\top$ and $(10, 1.5)^\top$; their optimization coordinates are obtained by multiplying by T_1^{-1}. The scalar optimum in original coordinates is $(0, -8.35510978)^\top$, and the local condition number of F is $\kappa_{\mathrm{ref}} \approx 737.86$.

Figs. 2 and 3 show the landscape, trajectories, and error curves. SOAP* and CHIME approach the optimum from both starts, with final distances below 4×10^{-7}. HARMONIC remains at distances 15.42 and 10.81, respectively. The three-dimensional surface and contour trajectories show the narrow, rotated valley induced by T_1. The mode curves explain why a small displacement along one visible axis is not sufficient evidence of convergence: SOAP* and CHIME reduce both directional errors, whereas HARMONIC's Adam counterpart retains substantial error in both reference modes. For the curvature-aware trajectories, the strong-mode error reaches the display floor before the weak-mode error, which remains the later bottleneck. Since this case has only one objective, the contrast isolates a conditioning difficulty.



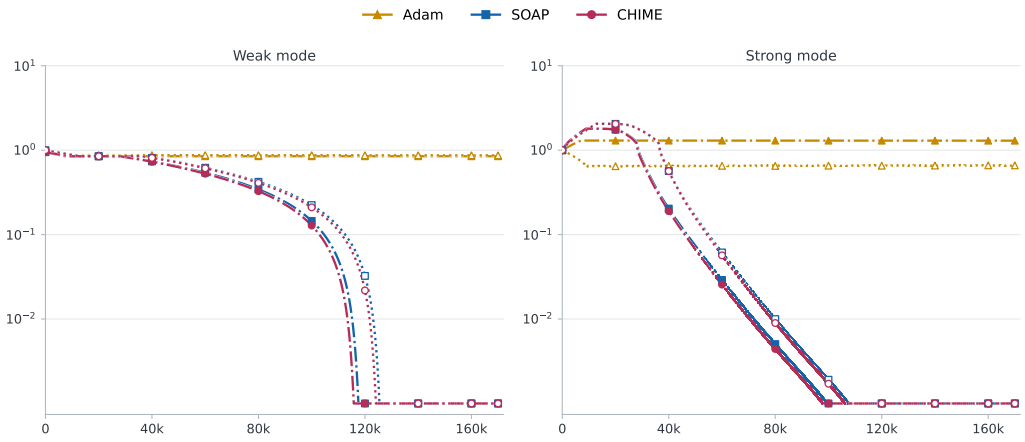
(a) Scalar landscape F.

(b) Optimization trajectories.

Figure 2: Case 1: the scalar landscape and trajectories in optimization coordinates. The plotted axes (u_1, u_2) denote (θ_1, θ_2), and the star marks the optimum.



(a) Distance to the optimum from the first (left) and second (right) starts.



(b) Normalized weak- and strong-mode errors, with both starts in each panel.

Figure 3: Case 1 error curves through 170,000 steps. The SOAP/Adam curve labels denote the single-objective counterparts of SOAP*/HARMONIC.

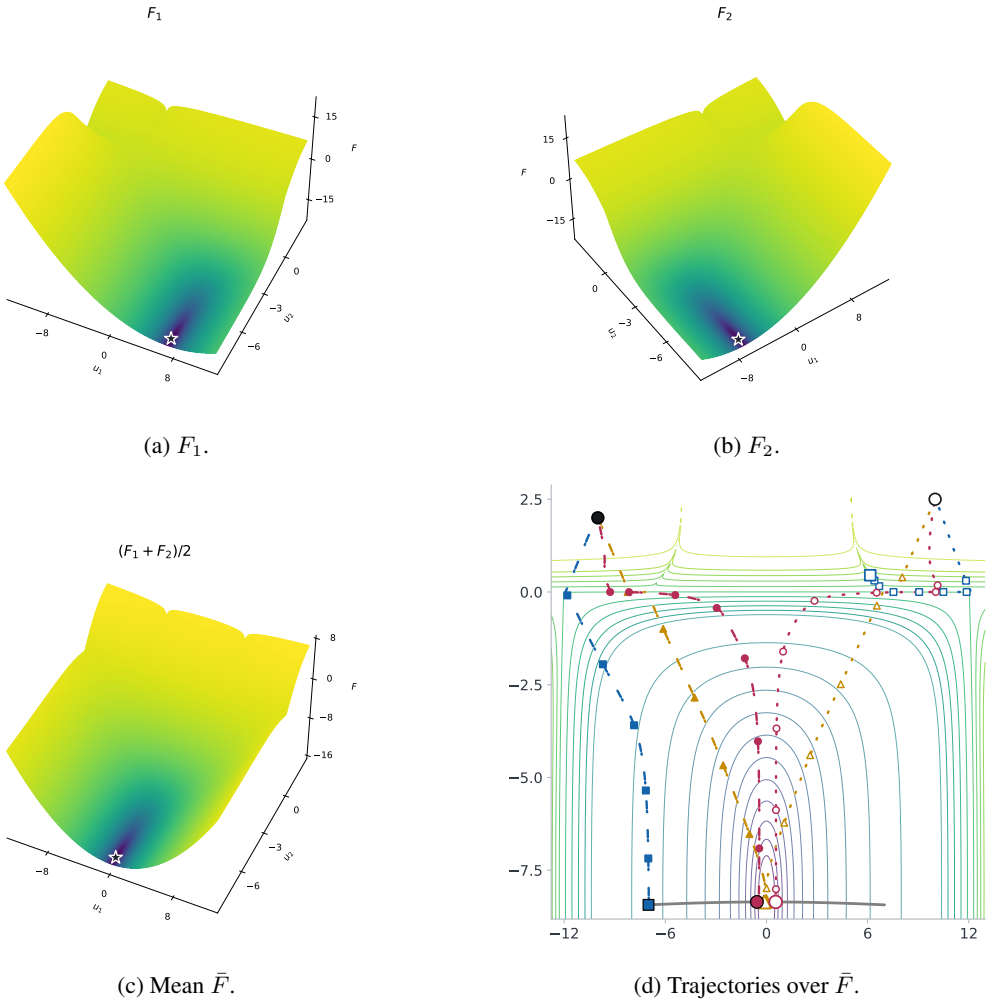
A.3 Case 2: Fixed scale imbalance and multi-objective interaction

Use $(w, \delta, \rho, \kappa) = (2, 5 \times 10^{-6}, 0, 0)$ in the original coordinates, so $\boldsymbol{\theta} = \boldsymbol{q}$. For $i \in \{1, 2\}$, define $F_i(\boldsymbol{\theta}) = f_i(\boldsymbol{\theta})$ and their mean $\bar{F} = (F_1 + F_2)/2$. The training objectives are

$$L_1(\boldsymbol{\theta}) = 10^{-8}\big(F_1(\boldsymbol{\theta}) + 20\big), \qquad L_2(\boldsymbol{\theta}) = F_2(\boldsymbol{\theta}) + 20. \tag{32}$$

The first and second initializations are $(-10, 2)^\top$ and $(10, 2.5)^\top$. The target here is $\mathcal{P}_2$, whose symmetric point is $(0, -8.35510978)^\top$; the local mean-Hessian condition number is $\kappa_{\mathrm{ref}} \approx 7.38$.

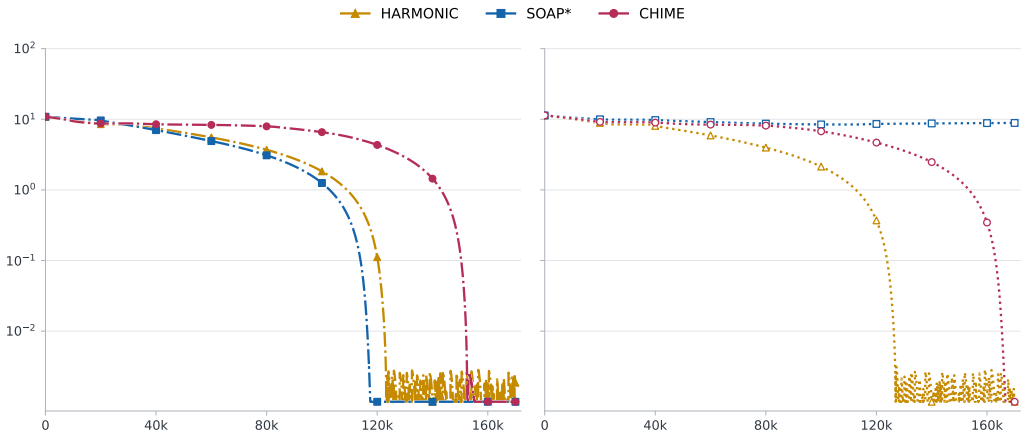
Figs. 4 and 5 show the objectives and optimization results. HARMONIC and CHIME approach the Pareto curve from both starts: their final distances are $(1.86 \times 10^{-3}, 6.42 \times 10^{-4})$ and $(1.31 \times 10^{-7}, 1.09 \times 10^{-4})$, respectively. SOAP* approaches the curve from the first start but remains at distance 8.87 from the second. The component surfaces show that the two objectives have distinct minimizers. A small change in the scalarized loss, or progress toward one component's local basin, does not by itself certify proximity to the lower-region Pareto target. For the unsuccessful SOAP* initialization, both mode errors plateau, with the weak-mode projection remaining especially large. HARMONIC and CHIME reduce both projections from both starts.



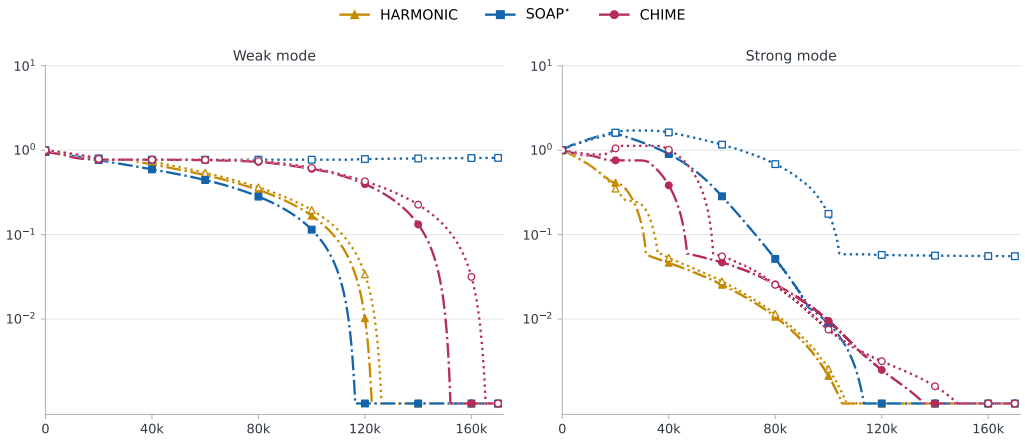
(a) F_1. (b) F_2.

(c) Mean $\bar{F}$. (d) Trajectories over $\bar{F}$.

Figure 4: Case 2: component landscapes, their mean, and optimization trajectories. Here $(u_1, u_2) = (\theta_1, \theta_2) = (x, y)$. Stars mark scalar minimizers; the gray curve marks the Pareto target.



(a) Distance to the Pareto curve from the first (left) and second (right) starts.



(b) Normalized weak- and strong-mode errors.

Figure 5: Case 2 error curves through 170,000 steps. Filled/open markers indicate the first/second initialization.

A.4 CASE 3: ILL-CONDITIONED MULTI-OBJECTIVE OPTIMIZATION

Use $(w, \rho, \kappa) = (4, 1, 1)$ and $T_3 = 2\,\mathrm{diag}(30, 1)R(30^\circ)$. With the common function

$$B(\boldsymbol{q}) = 40 \exp\left(-\frac{x^2}{2(1.5)^2}\right) \max\{\tanh(y/4), 0\}, \tag{33}$$

define the two components, for $i \in \{1, 2\}$, and their mean by

$$F_i(\boldsymbol{\theta}) = \exp\left(\frac{f_i(T_3\boldsymbol{\theta}) + B(T_3\boldsymbol{\theta})}{5}\right), \qquad \bar{F}(\boldsymbol{\theta}) = \frac{F_1(\boldsymbol{\theta}) + F_2(\boldsymbol{\theta})}{2}. \tag{34}$$

The training objectives are

$$L_1(\boldsymbol{\theta}) = F_1(\boldsymbol{\theta}), \qquad L_2(\boldsymbol{\theta}) = 10^{-6} F_2(\boldsymbol{\theta}). \tag{35}$$

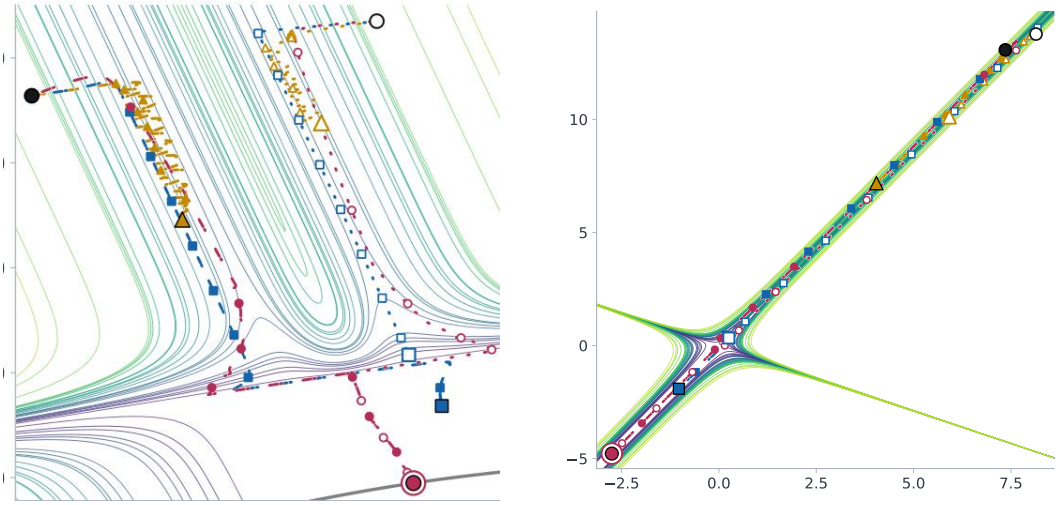
The first and second initializations in original coordinates are $(-10, 30)^\top$ and $(10, 32)^\top$; their optimization coordinates are obtained using T_3^{-1}. The target is $\mathcal{P}_4$, with symmetric point $(0, -11.02649122)^\top$ in original coordinates, and $\kappa_{\mathrm{ref}} \approx 1.05 \times 10^4$ for $\bar{F}$. To limit large initial steps in this ill-conditioned problem, all methods linearly warm up the learning rate from 10^{-7} to 10^{-4} over the first 1,000 steps.

Visualization of Case 3. Fig. 6 compares the Case 3 panel shown in Fig. 1 (right) with the same trajectories plotted directly in the optimization coordinates $\boldsymbol{\theta}$. Without the display adjustment, the strong anisotropy compresses the contours and trajectories into a narrow band, making the characteristic differences between methods much harder to distinguish. For the Introduction panel, we therefore use display coordinates $\boldsymbol{u} = (u_1, u_2)^\top$ defined by

$$\boldsymbol{q} = R(-15^\circ)\boldsymbol{u}, \qquad \boldsymbol{u} = R(15^\circ)T_3\boldsymbol{\theta}. \tag{36}$$

This presents the landscape without the anisotropic coordinate stretching, with a rotation chosen for readability. Contours evaluate $\bar{F}(T_3^{-1}R(-15^\circ)\boldsymbol{u})$; all six trajectories, their initial and final points, and the Pareto curve are mapped to the same display coordinates. The mapping is applied only for visualization: the optimization still uses $\boldsymbol{\theta}$, and the recorded iterates and error diagnostics are unchanged.

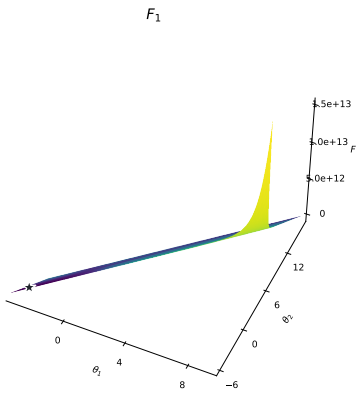
Results. At 170,000 steps, SOAP* remains at distances 6.80 and 11.94, and HARMONIC at 27.65 and 34.55. CHIME approaches the Pareto curve from both starts, reaching distances 1.39×10^{-7} and 4.27×10^{-4}; Fig. 8 gives the corresponding error curves. The three-dimensional component surfaces in Fig. 7 expose the stretched geometry that is visually removed by the display transformation. The mode curves show that HARMONIC makes little progress in the weak reference direction, while SOAP* reduces it more slowly and retains a nonzero strong-mode error. CHIME ultimately reduces both projections and the distance to the Pareto curve from both starts. Transient nonmonotonicity of the strong-mode projection is consistent with a nonlinear trajectory measured against fixed reference directions and a moving nearest Pareto point. Together, the three cases motivate selecting objective-aware directions in the same curvature geometry used for the update, rather than relying on objective balancing or preconditioning alone.



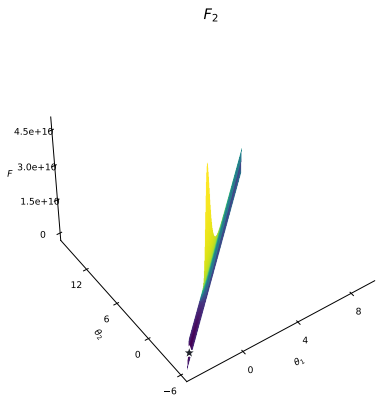
(a) Display coordinates used in Fig. 1 (right).

(b) Coordinates without the display adjustment.

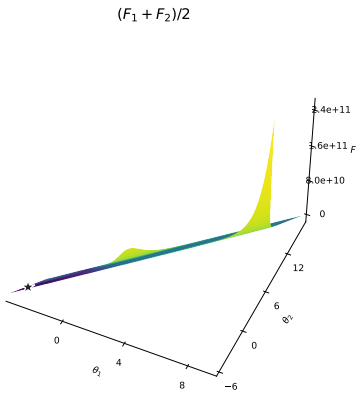
Figure 6: Case 3 visualization comparison. Subfigure (a) reproduces the right panel used in the Introduction. Subfigure (b) shows the same contours, trajectories, initial and final points, and Pareto curve directly in optimization coordinates. The display map in Eq. (36) removes the anisotropic stretching and rotates the plot for readability, making the characteristic optimization behaviors more visible.



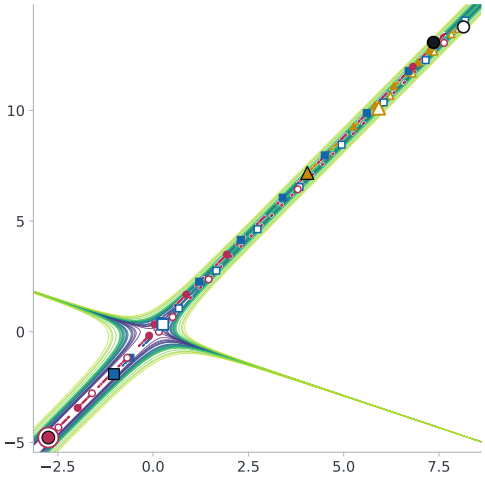
(a) F_1.



(b) F_2.

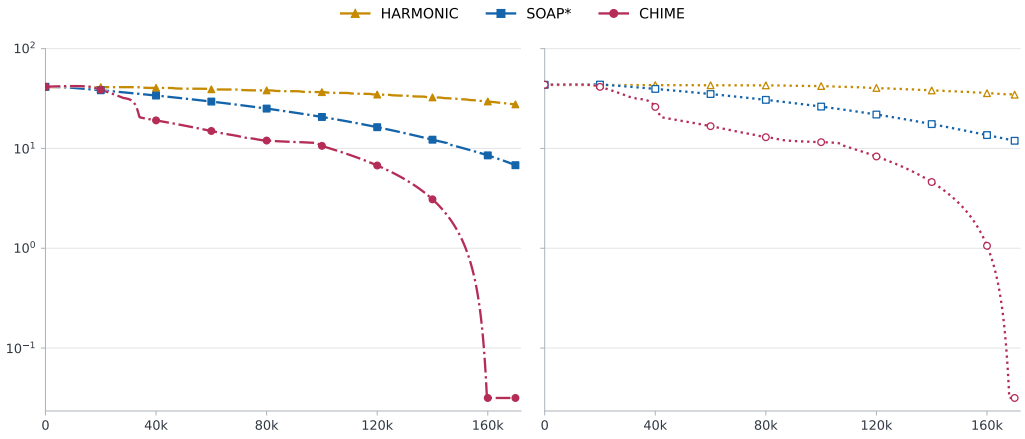


(c) Mean $\bar{F}$.

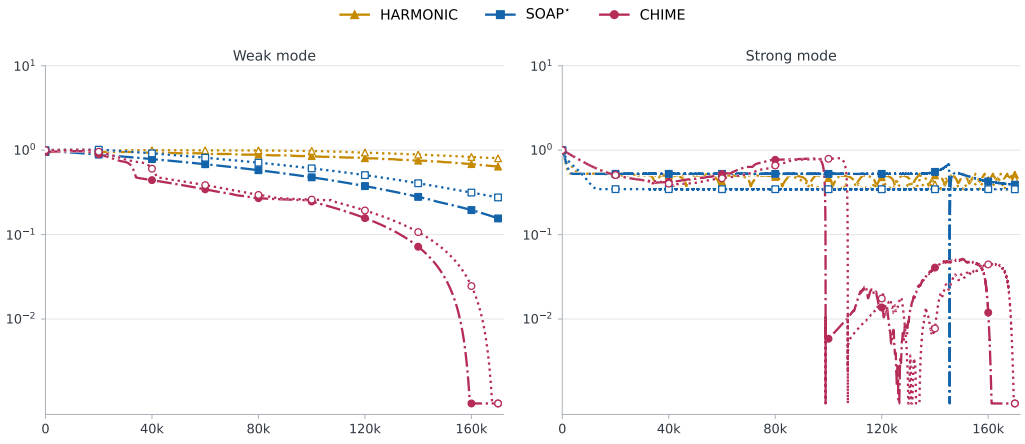


(d) Trajectories in $\boldsymbol{\theta}$.

Figure 7: Case 3 component landscapes and their mean, together with the raw trajectory plot in optimization coordinates. Panel (d) shows the same trajectories as Fig. 1 (right) without the display adjustment in Eq. (36). Stars mark scalar minimizers, and the gray curve marks the Pareto target.



(a) Distance to the Pareto curve from the first (left) and second (right) starts.



(b) Normalized weak- and strong-mode errors.

Figure 8: Case 3 error curves through 170,000 steps, using the distance and mode definitions given above. Filled/open markers distinguish the two initializations.

B Additional Details for Sec. 4

Throughout the proofs, $L_i > 0$, $\mathbf{B} \succ 0$, and $\mathbf{P} = \mathbf{B}^{-1}$. All quantities are evaluated at the current iterate unless stated otherwise.

B.1 Dual formulation

For a prescribed budget $q > 0$, the supremum of Eq. (16) over γ is finite only when $\mathbf{1}^\top \boldsymbol{\alpha} = 1$. For $\boldsymbol{\alpha} \in \Delta_m$ and $\nu > 0$, completing the square gives

$$\mathcal{L} = -\frac{\nu}{2} \left\| \mathbf{d} - \nu^{-1} \mathbf{P} \widetilde{\mathbf{G}} \boldsymbol{\alpha} \right\|_{\mathbf{B}}^2 + \frac{\boldsymbol{\alpha}^\top \mathbf{C} \boldsymbol{\alpha}}{2\nu} + \frac{\nu q}{2}, \tag{37}$$

$$\sup_{\mathbf{d},\gamma} \mathcal{L} = \frac{\boldsymbol{\alpha}^\top \mathbf{C} \boldsymbol{\alpha}}{2\nu} + \frac{\nu q}{2}. \tag{38}$$

If $\boldsymbol{\alpha}^\top \mathbf{C} \boldsymbol{\alpha} > 0$, minimizing over ν yields

$$\nu = \sqrt{\frac{\boldsymbol{\alpha}^\top \mathbf{C} \boldsymbol{\alpha}}{q}}, \qquad \inf_{\nu > 0} \sup_{\mathbf{d},\gamma} \mathcal{L} = \sqrt{q\, \boldsymbol{\alpha}^\top \mathbf{C} \boldsymbol{\alpha}}. \tag{39}$$

If the quadratic form is zero, $\widetilde{\mathbf{G}} \boldsymbol{\alpha} = \mathbf{0}$ because $\mathbf{P} \succ 0$, and the same value is attained at $\nu = 0$. The pair $(\mathbf{d}, \gamma) = (\mathbf{0}, -1)$ is strictly feasible, so Slater's condition gives the exact dual

$$\min_{\boldsymbol{\alpha} \in \Delta_m} \sqrt{q\, \boldsymbol{\alpha}^\top \mathbf{C} \boldsymbol{\alpha}}. \tag{40}$$

Its minimizers coincide with those of Eq. (17), independently of q.

B.2 Proof of Proposition 1

Proposition 1 (Non-conflict at dual optimality)**.** *For any minimizer $\boldsymbol{\alpha}^\star$ of Eq. (17), the corresponding combination $\mathbf{P}\mathbf{G}\boldsymbol{\lambda}^\star$ belongs to the dual gradient cone $\mathcal{K}^* = \{\mathbf{d} \in \mathbb{R}^D : \mathbf{G}^\top \mathbf{d} \geq \mathbf{0}\}$.*

Proof.
For every $\mathbf{z} \in \mathbb{R}^m$,

$$\mathbf{z}^\top \mathbf{C} \mathbf{z} = \|\mathbf{P}^{1/2} \widetilde{\mathbf{G}} \mathbf{z}\|_2^2 \geq 0, \tag{41}$$

so $\mathbf{C} \succeq 0$ and the objective in Eq. (17) is convex.

Since $\boldsymbol{\alpha}^\star$ minimizes this objective over the simplex, the first-order optimality condition gives, for any feasible $\boldsymbol{\beta}$,

$$(\boldsymbol{\beta} - \boldsymbol{\alpha}^\star)^\top \mathbf{C} \boldsymbol{\alpha}^\star \geq 0. \tag{42}$$

Taking $\boldsymbol{\beta} = \mathbf{e}_i$, where $\mathbf{e}_i$ is the ith simplex vertex, yields

$$(\mathbf{C}\boldsymbol{\alpha}^\star)_i \geq \boldsymbol{\alpha}^{\star\top} \mathbf{C} \boldsymbol{\alpha}^\star = c_\star \geq 0 \qquad \text{for all } i. \tag{43}$$

Hence

$$\mathbf{C}\boldsymbol{\alpha}^\star \geq c_\star \mathbf{1} \geq \mathbf{0}. \tag{44}$$

Let

$$\mathbf{D}_L := \operatorname{diag}(2L_1, \ldots, 2L_m).$$

By $\widetilde{\mathbf{g}}_i = \mathbf{g}_i/(2L_i)$ and $\lambda_i^\star = \alpha_i^\star/(2L_i)$,

$$\mathbf{G} = \widetilde{\mathbf{G}} \mathbf{D}_L, \qquad \mathbf{G}\boldsymbol{\lambda}^\star = \widetilde{\mathbf{G}} \boldsymbol{\alpha}^\star.$$

Therefore, for $\mathbf{d}^\star = \mathbf{P}\mathbf{G}\boldsymbol{\lambda}^\star$,

$$\begin{aligned} \mathbf{G}^\top \mathbf{d}^\star &= \mathbf{G}^\top \mathbf{P} \widetilde{\mathbf{G}} \boldsymbol{\alpha}^\star \\ &= \mathbf{D}_L \widetilde{\mathbf{G}}^\top \mathbf{P} \widetilde{\mathbf{G}} \boldsymbol{\alpha}^\star \\ &= \mathbf{D}_L \mathbf{C} \boldsymbol{\alpha}^\star \geq \mathbf{0}. \end{aligned} \tag{45}$$

Thus $\mathbf{d}^\star \in \mathcal{K}^*$. □

B.3 Proof of Proposition 2

Proposition 2 (Metric projection and optimal direction)**.** *Let $\boldsymbol{\alpha}^\star$ solve Eq.* (17)*. With budget $q = \boldsymbol{\alpha}^{\star\top}\mathbf{C}\boldsymbol{\alpha}^\star > 0$, the unique optimal direction of Eq.* (15) *is the $\mathbf{B}$-metric projection of the origin onto $\mathcal{H}^{\Delta}_{\mathbf{P}}$:*

$$\mathbf{d}^\star = \mathbf{P}\mathbf{G}\boldsymbol{\lambda}^\star = \arg\min_{\mathbf{d}\in\mathcal{H}^{\Delta}_{\mathbf{P}}} \frac{1}{2}\|\mathbf{d}\|^2_{\mathbf{B}}, \qquad \gamma^\star = q.$$

Proof.
Set

$$\mathbf{d}^\star = \mathbf{P}\widetilde{\mathbf{G}}\boldsymbol{\alpha}^\star = \mathbf{P}\mathbf{G}\boldsymbol{\lambda}^\star, \qquad q = \boldsymbol{\alpha}^{\star\top}\mathbf{C}\boldsymbol{\alpha}^\star. \tag{46}$$

Since $\mathbf{B} = \mathbf{P}^{-1}$,

$$\|\mathbf{d}^\star\|^2_{\mathbf{B}} = \boldsymbol{\alpha}^{\star\top}\widetilde{\mathbf{G}}^\top\mathbf{P}\widetilde{\mathbf{G}}\boldsymbol{\alpha}^\star = \boldsymbol{\alpha}^{\star\top}\mathbf{C}\boldsymbol{\alpha}^\star = q. \tag{47}$$

Moreover, Eq. (43) gives, for every objective i,

$$p_i(\mathbf{d}^\star) = \widetilde{\mathbf{g}}_i^\top\mathbf{P}\widetilde{\mathbf{G}}\boldsymbol{\alpha}^\star = (\mathbf{C}\boldsymbol{\alpha}^\star)_i \geq q. \tag{48}$$

Because $\boldsymbol{\alpha}^\star \in \Delta_m$,

$$\sum_{i=1}^m \alpha_i^\star p_i(\mathbf{d}^\star) = \boldsymbol{\alpha}^{\star\top}\mathbf{C}\boldsymbol{\alpha}^\star = q. \tag{49}$$

Thus every $p_i(\mathbf{d}^\star)$ is at least q, while their $\boldsymbol{\alpha}^\star$-weighted average equals q. Hence $p_i(\mathbf{d}^\star) = q$ for every i with $\alpha_i^\star > 0$. Since at least one coefficient is positive,

$$\min_i p_i(\mathbf{d}^\star) = q. \tag{50}$$

Together with Eq. (47), this shows that $(\mathbf{d}^\star, q)$ is feasible for Eq. (15).

Using $\alpha_i = 2L_i\lambda_i$, the curvature-harmonized hull can equivalently be written as

$$\begin{aligned}\mathcal{H}^{\Delta}_{\mathbf{P}} &= \left\{\mathbf{P}\mathbf{G}\boldsymbol{\lambda} : \boldsymbol{\lambda} \geq \mathbf{0},\ \sum_{i=1}^m 2L_i\lambda_i = 1,\ \mathbf{G}^\top\mathbf{P}\mathbf{G}\boldsymbol{\lambda} \geq \mathbf{0}\right\} \\ &= \left\{\mathbf{P}\widetilde{\mathbf{G}}\boldsymbol{\alpha} : \boldsymbol{\alpha} \in \Delta_m,\ \mathbf{C}\boldsymbol{\alpha} \geq \mathbf{0}\right\}.\end{aligned} \tag{51}$$

Equation (44) gives $\mathbf{C}\boldsymbol{\alpha}^\star \geq \mathbf{0}$. Therefore

$$\mathbf{d}^\star = \mathbf{P}\widetilde{\mathbf{G}}\boldsymbol{\alpha}^\star \in \mathcal{H}^{\Delta}_{\mathbf{P}}. \tag{52}$$

For any feasible $(\mathbf{d}, \gamma)$,

$$\begin{aligned}\gamma \leq \min_i p_i(\mathbf{d}) &\leq \sum_{i=1}^m \alpha_i^\star p_i(\mathbf{d}) \\ &= \boldsymbol{\alpha}^{\star\top}\widetilde{\mathbf{G}}^\top\mathbf{d} \\ &= (\mathbf{d}^\star)^\top\mathbf{B}\mathbf{d}.\end{aligned} \tag{53}$$

Because $\mathbf{B} \succ 0$, $\langle \mathbf{u}, \mathbf{v}\rangle_{\mathbf{B}} := \mathbf{u}^\top\mathbf{B}\mathbf{v}$ defines an inner product, and its Cauchy–Schwarz inequality gives

$$(\mathbf{d}^\star)^\top\mathbf{B}\mathbf{d} \leq \|\mathbf{d}^\star\|_{\mathbf{B}}\,\|\mathbf{d}\|_{\mathbf{B}}. \tag{54}$$

Using $\|\mathbf{d}^\star\|_{\mathbf{B}} = \sqrt{q}$ and the feasible budget $\|\mathbf{d}\|^2_{\mathbf{B}} \leq q$, we obtain

$$\gamma \leq (\mathbf{d}^\star)^\top\mathbf{B}\mathbf{d} \leq q. \tag{55}$$

Since $(\mathbf{d}^\star, q)$ attains this upper bound,

$$\gamma^\star = q. \tag{56}$$

Let $\widehat{\mathbf{d}}$ be any other optimal direction. Then $\gamma^\star = q$ forces equality throughout Eq. (55), and hence

$$(\mathbf{d}^\star)^\top\mathbf{B}\widehat{\mathbf{d}} = q, \qquad \|\widehat{\mathbf{d}}\|^2_{\mathbf{B}} = \|\mathbf{d}^\star\|^2_{\mathbf{B}} = q. \tag{57}$$

Therefore

$$\begin{aligned}\|\widehat{\mathbf{d}} - \mathbf{d}^\star\|_{\mathbf{B}}^2 &= \|\widehat{\mathbf{d}}\|_{\mathbf{B}}^2 + \|\mathbf{d}^\star\|_{\mathbf{B}}^2 - 2(\mathbf{d}^\star)^\top \mathbf{B}\widehat{\mathbf{d}} \\ &= q + q - 2q = 0. \end{aligned} \tag{58}$$

Since $\mathbf{B} \succ 0$, this implies $\widehat{\mathbf{d}} = \mathbf{d}^\star$. Thus $\mathbf{d}^\star$ is the unique optimal direction of Eq. (15).

It remains to establish the metric-projection characterization. For any $\mathbf{v} \in \mathcal{H}_{\mathbf{P}}^{\Delta}$, Eq. (51) gives some $\boldsymbol{\alpha} \in \Delta_m$ such that

$$\mathbf{v} = \mathbf{P}\widetilde{\mathbf{G}}\boldsymbol{\alpha}.$$

Hence

$$\|\mathbf{v}\|_{\mathbf{B}}^2 = \boldsymbol{\alpha}^\top \mathbf{C}\boldsymbol{\alpha} \geq \boldsymbol{\alpha}^{\star\top} \mathbf{C}\boldsymbol{\alpha}^\star = q = \|\mathbf{d}^\star\|_{\mathbf{B}}^2, \tag{59}$$

where the inequality follows from the optimality of $\boldsymbol{\alpha}^\star$ in Eq. (17). Since $\mathbf{d}^\star \in \mathcal{H}_{\mathbf{P}}^{\Delta}$, it therefore minimizes the $\mathbf{B}$-norm over the hull.

The coefficient set

$$\{\boldsymbol{\alpha} \in \Delta_m : \mathbf{C}\boldsymbol{\alpha} \geq \mathbf{0}\}$$

is convex, and so is its linear image $\mathcal{H}_{\mathbf{P}}^{\Delta}$. Because $\mathbf{B} \succ 0$, $\frac{1}{2}\|\mathbf{d}\|_{\mathbf{B}}^2$ is strictly convex. Thus this minimum is unique, and

$$\mathbf{d}^\star = \arg\min_{\mathbf{d} \in \mathcal{H}_{\mathbf{P}}^{\Delta}} \frac{1}{2}\|\mathbf{d}\|_{\mathbf{B}}^2. \tag{60}$$

This is exactly the $\mathbf{B}$-metric projection of the origin onto $\mathcal{H}_{\mathbf{P}}^{\Delta}$. □

Zero dual value and Pareto stationarity. Let $c_\star = \min_{\boldsymbol{\alpha} \in \Delta_m} \boldsymbol{\alpha}^\top \mathbf{C}\boldsymbol{\alpha}$. Since $\mathbf{P} \succ 0$,

$$c_\star = 0 \quad \Longleftrightarrow \quad \exists\, \boldsymbol{\alpha} \in \Delta_m : \widetilde{\mathbf{G}}\boldsymbol{\alpha} = \mathbf{0}. \tag{61}$$

For such an $\boldsymbol{\alpha}$, define

$$\beta_i = \frac{\alpha_i/(2L_i)}{\sum_{j=1}^m \alpha_j/(2L_j)}.$$

Then $\boldsymbol{\beta} \in \Delta_m$ and $\mathbf{G}\boldsymbol{\beta} = \mathbf{0}$. Conversely, if $\boldsymbol{\beta} \in \Delta_m$ satisfies $\mathbf{G}\boldsymbol{\beta} = \mathbf{0}$, then

$$\alpha_i = \frac{2L_i\beta_i}{\sum_{j=1}^m 2L_j\beta_j}$$

belongs to Δ_m and satisfies $\widetilde{\mathbf{G}}\boldsymbol{\alpha} = \mathbf{0}$. Therefore,

$$c_\star = 0 \quad \Longleftrightarrow \quad \exists\, \boldsymbol{\beta} \in \Delta_m : \mathbf{G}\boldsymbol{\beta} = \mathbf{0}, \tag{62}$$

which is precisely first-order Pareto stationarity. Proposition 2 therefore characterizes the nonstationary case $c_\star = q > 0$.

Connections to the shared-metric formulation and the Euclidean harmonized cone. Replacing $\mathbf{g}_i$ by $\widetilde{\mathbf{g}}_i$ in Eq. (13) gives

$$\min_{\mathbf{d} \in \mathbb{R}^D} \left\{ \frac{1}{2}\|\mathbf{d}\|_{\mathbf{B}}^2 - \min_i p_i(\mathbf{d}) \right\}. \tag{63}$$

For any $\mathbf{d}$,

$$\begin{aligned}\frac{1}{2}\|\mathbf{d}\|_{\mathbf{B}}^2 - \min_i p_i(\mathbf{d}) &\geq \frac{1}{2}\|\mathbf{d}\|_{\mathbf{B}}^2 - \sum_{i=1}^m \alpha_i^\star p_i(\mathbf{d}) \\ &= \frac{1}{2}\|\mathbf{d}\|_{\mathbf{B}}^2 - (\mathbf{d}^\star)^\top \mathbf{B}\mathbf{d} \\ &= \frac{1}{2}\|\mathbf{d} - \mathbf{d}^\star\|_{\mathbf{B}}^2 - \frac{q}{2}. \end{aligned} \tag{64}$$

At $\mathbf{d} = \mathbf{d}^\star$, Eqs. (50) and (47) make both sides equal to $-q/2$. For every $\mathbf{d} \neq \mathbf{d}^\star$, the final expression is strictly larger than $-q/2$. Therefore $\mathbf{d}^\star$ is also the unique minimizer of Eq. (63).

When $\mathbf{P} = \mathbf{I}$, the curvature-harmonized hull becomes

$$\mathcal{H}_{\mathbf{I}}^{\Delta} = \left\{ \mathbf{G}\boldsymbol{\lambda} : \boldsymbol{\lambda} \geq \mathbf{0},\ \sum_{i=1}^{m} 2L_i\lambda_i = 1,\ \mathbf{G}^\top \mathbf{G}\boldsymbol{\lambda} \geq \mathbf{0} \right\}, \tag{65}$$

whereas the harmonized cone in Eq. (6) is

$$\mathcal{H} = \left\{ \mathbf{G}\boldsymbol{\lambda} : \boldsymbol{\lambda} \geq \mathbf{0},\ \mathbf{G}^\top \mathbf{G}\boldsymbol{\lambda} \geq \mathbf{0} \right\}. \tag{66}$$

Hence $\mathcal{H}_{\mathbf{I}}^{\Delta} \subseteq \mathcal{H}$, and therefore

$$\mathrm{cone}(\mathcal{H}_{\mathbf{I}}^{\Delta}) \subseteq \mathcal{H}. \tag{67}$$

To show that $\mathcal{H}_{\mathbf{I}}^{\Delta}$ intersects every nonzero ray of $\mathcal{H}$, take any nonzero $\mathbf{v} = \mathbf{G}\boldsymbol{\lambda} \in \mathcal{H}$ and define

$$s = \sum_{i=1}^{m} 2L_i\lambda_i > 0, \qquad \overline{\boldsymbol{\lambda}} = \frac{\boldsymbol{\lambda}}{s}. \tag{68}$$

Then

$$\sum_{i=1}^{m} 2L_i\overline{\lambda}_i = 1, \qquad \mathbf{G}^\top \mathbf{G}\overline{\boldsymbol{\lambda}} = \frac{1}{s}\mathbf{G}^\top \mathbf{G}\boldsymbol{\lambda} \geq \mathbf{0},$$

so

$$\frac{\mathbf{v}}{s} = \mathbf{G}\overline{\boldsymbol{\lambda}} \in \mathcal{H}_{\mathbf{I}}^{\Delta}. \tag{69}$$

Therefore $\mathbf{v} \in \mathrm{cone}(\mathcal{H}_{\mathbf{I}}^{\Delta})$, which gives the reverse inclusion and hence

$$\mathrm{cone}(\mathcal{H}_{\mathbf{I}}^{\Delta}) = \mathcal{H}. \tag{70}$$

Thus the residual-normalized shared-metric formulation selects the same direction $\mathbf{d}^\star$, while in the Euclidean case $\mathcal{H}_{\mathbf{I}}^{\Delta}$ is a normalized cross-section of the harmonized cone $\mathcal{H}$.

B.4 Proof of Theorem 1

Theorem 1 (Objective progress and local residual-mode contraction)**.** *Let $\mathbf{P} \succ 0$ and $\boldsymbol{\alpha}^\star$ solve Eq.* (17) *with the matched budget $q > 0$ in Proposition 2. Denote the largest and smallest positive eigenvalues of $\mathbf{K}_{\boldsymbol{\alpha},\mathbf{P}}$ by $\lambda_{\max}$ and $\lambda_{\min}^{+}$, and define its effective condition number as $\kappa_{\mathrm{eff}} = \lambda_{\max}/\lambda_{\min}^{+}$. For a unit eigenvector $\mathbf{u}_k$ of $\mathbf{K}_{\boldsymbol{\alpha},\mathbf{P}}$ with eigenvalue $\lambda_k > 0$, let $a_k = \mathbf{u}_k^\top \mathbf{e}_{\boldsymbol{\alpha}}$ and $a_k^{+} = \mathbf{u}_k^\top \mathbf{e}_{\boldsymbol{\alpha}}^{+}$. In the local linearized model of Eq.* (21)*, for $0 < \eta \leq 1/\lambda_{\max}$,*

$$|a_k^{+}| \leq \left(1 - \eta \frac{q}{\kappa_{\mathrm{eff}}}\right) |a_k|.$$

Proof.
Fix the current parameters $\boldsymbol{\theta}$, coefficients $\boldsymbol{\alpha}^\star$, metric $\mathbf{P}$, and residual normalizers. For a nearby parameter vector $\boldsymbol{\vartheta}$, define the frozen-normalization residual map

$$\mathbf{e}_{\boldsymbol{\alpha}}(\boldsymbol{\vartheta}) = \begin{bmatrix} \dfrac{\sqrt{\alpha_1^\star}\mathbf{r}_1(\boldsymbol{\vartheta})}{\|\mathbf{r}_1(\boldsymbol{\theta})\|_2} \\ \vdots \\ \dfrac{\sqrt{\alpha_m^\star}\mathbf{r}_m(\boldsymbol{\vartheta})}{\|\mathbf{r}_m(\boldsymbol{\theta})\|_2} \end{bmatrix}. \tag{71}$$

Evaluating this map and its Jacobian at $\boldsymbol{\theta}$ gives $\mathbf{e}_{\boldsymbol{\alpha}}$ and $\mathbf{J}_{\boldsymbol{\alpha}}$ in Eq. (20). Since $\boldsymbol{\alpha}^\star \in \Delta_m$,

$$\|\mathbf{e}_{\boldsymbol{\alpha}}\|_2^2 = \sum_{i=1}^{m} \alpha_i^\star \frac{\|\mathbf{r}_i\|_2^2}{\|\mathbf{r}_i\|_2^2} = \sum_{i=1}^{m} \alpha_i^\star = 1. \tag{72}$$

Moreover,

$$\begin{aligned}\mathbf{J}_{\boldsymbol{\alpha}}^{\top}\mathbf{e}_{\boldsymbol{\alpha}} &= \sum_{i=1}^{m} \alpha_i^{\star} \frac{\mathbf{J}_i^{\top}\mathbf{r}_i}{\|\mathbf{r}_i\|_2^2} \\ &= \sum_{i=1}^{m} \alpha_i^{\star}\widetilde{\mathbf{g}}_i = \widetilde{\mathbf{G}}\boldsymbol{\alpha}^{\star}. \end{aligned} \tag{73}$$

Hence

$$\mathbf{d}^{\star} = \mathbf{P}\widetilde{\mathbf{G}}\boldsymbol{\alpha}^{\star} = \mathbf{P}\mathbf{J}_{\boldsymbol{\alpha}}^{\top}\mathbf{e}_{\boldsymbol{\alpha}}. \tag{74}$$

By definition, the local linearized model uses the first-order expansion of the frozen residual map around $\boldsymbol{\theta}$. For the update $\boldsymbol{\theta}^{+} = \boldsymbol{\theta} - \eta\mathbf{d}^{\star}$, its residual dynamics are therefore

$$\begin{aligned}\mathbf{e}_{\boldsymbol{\alpha}}^{+} &= \mathbf{e}_{\boldsymbol{\alpha}} - \eta\mathbf{J}_{\boldsymbol{\alpha}}\mathbf{d}^{\star} \\ &= \mathbf{e}_{\boldsymbol{\alpha}} - \eta\mathbf{J}_{\boldsymbol{\alpha}}\mathbf{P}\mathbf{J}_{\boldsymbol{\alpha}}^{\top}\mathbf{e}_{\boldsymbol{\alpha}} \\ &= \left(\mathbf{I} - \eta\mathbf{K}_{\boldsymbol{\alpha},\mathbf{P}}\right)\mathbf{e}_{\boldsymbol{\alpha}}. \end{aligned} \tag{75}$$

At the same state,

$$\begin{aligned}\mathbf{e}_{\boldsymbol{\alpha}}^{\top}\mathbf{K}_{\boldsymbol{\alpha},\mathbf{P}}\mathbf{e}_{\boldsymbol{\alpha}} &= \left(\mathbf{J}_{\boldsymbol{\alpha}}^{\top}\mathbf{e}_{\boldsymbol{\alpha}}\right)^{\top}\mathbf{P}\left(\mathbf{J}_{\boldsymbol{\alpha}}^{\top}\mathbf{e}_{\boldsymbol{\alpha}}\right) \\ &= \left(\widetilde{\mathbf{G}}\boldsymbol{\alpha}^{\star}\right)^{\top}\mathbf{P}\left(\widetilde{\mathbf{G}}\boldsymbol{\alpha}^{\star}\right) \\ &= \boldsymbol{\alpha}^{\star\top}\mathbf{C}\boldsymbol{\alpha}^{\star} = q. \end{aligned} \tag{76}$$

The kernel is positive semidefinite because

$$\mathbf{K}_{\boldsymbol{\alpha},\mathbf{P}} = \left(\mathbf{J}_{\boldsymbol{\alpha}}\mathbf{P}^{1/2}\right)\left(\mathbf{J}_{\boldsymbol{\alpha}}\mathbf{P}^{1/2}\right)^{\top} \succeq 0. \tag{77}$$

Let $\{\mathbf{u}_j\}$ be an orthonormal eigenbasis of $\mathbf{K}_{\boldsymbol{\alpha},\mathbf{P}}$ with eigenvalues $\lambda_j \geq 0$, and write

$$\mathbf{e}_{\boldsymbol{\alpha}} = \sum_j a_j\mathbf{u}_j, \qquad a_j = \mathbf{u}_j^{\top}\mathbf{e}_{\boldsymbol{\alpha}}. \tag{78}$$

Equation (72) gives

$$\sum_j a_j^2 = \|\mathbf{e}_{\boldsymbol{\alpha}}\|_2^2 = 1, \tag{79}$$

and substituting the eigendecomposition into Eq. (76) gives

$$q = \sum_j \lambda_j a_j^2. \tag{80}$$

Therefore,

$$q = \sum_j \lambda_j a_j^2 \leq \lambda_{\max}\sum_j a_j^2 = \lambda_{\max}. \tag{81}$$

Since $\kappa_{\text{eff}} = \lambda_{\max}/\lambda_{\min}^{+}$,

$$\lambda_{\min}^{+} = \frac{\lambda_{\max}}{\kappa_{\text{eff}}} \geq \frac{q}{\kappa_{\text{eff}}}. \tag{82}$$

Projecting Eq. (75) onto an eigenvector $\mathbf{u}_k$ with $\lambda_k > 0$ gives

$$a_k^{+} = (1 - \eta\lambda_k)\, a_k. \tag{83}$$

For $0 < \eta \leq 1/\lambda_{\max}$, $0 \leq 1 - \eta\lambda_k \leq 1$, and $\lambda_k \geq \lambda_{\min}^{+} \geq q/\kappa_{\text{eff}}$. Hence

$$\begin{aligned}|a_k^{+}| &= (1 - \eta\lambda_k)\,|a_k| \\ &\leq \left(1 - \eta\lambda_{\min}^{+}\right)|a_k| \\ &\leq \left(1 - \eta\frac{q}{\kappa_{\text{eff}}}\right)|a_k|. \end{aligned} \tag{84}$$

□

Positive-spectrum contraction. Let $\mathbf{\Pi}_+$ and $\mathbf{\Pi}_0$ denote the orthogonal projectors onto the positive eigenspace and nullspace of $\mathbf{K}_{\boldsymbol{\alpha},\mathbf{P}}$, respectively. Within the local linearized model of Theorem 1, the null component is unchanged, whereas every positive mode satisfies the contraction bound in Eq. (84). Hence

$$\|\mathbf{e}_{\boldsymbol{\alpha}}^{+}\|_2^2 \le \|\mathbf{\Pi}_0 \mathbf{e}_{\boldsymbol{\alpha}}\|_2^2 + \left(1 - \eta \frac{q}{\kappa_{\text{eff}}}\right)^2 \|\mathbf{\Pi}_+ \mathbf{e}_{\boldsymbol{\alpha}}\|_2^2. \tag{85}$$

If $\alpha_i^\star = 0$, the corresponding residual and Jacobian blocks are zero and therefore contribute only zero eigenvalues; removing such blocks leaves the positive spectrum unchanged. The objective-wise non-conflict guarantee of Proposition 1 nevertheless continues to hold for every objective.

B.5 Empirical validation

We empirically examine the two quantities highlighted by the analysis in Sec. 4.2: objective-wise progress under the update metric and conditioning of the corresponding residual dynamics. The validation is conducted on the 1D wave-equation benchmark (Wave1d-C), whose governing equation, initial and boundary conditions, and reference solution are given in App. C.2. Fig. 9 summarizes both diagnostics along training.

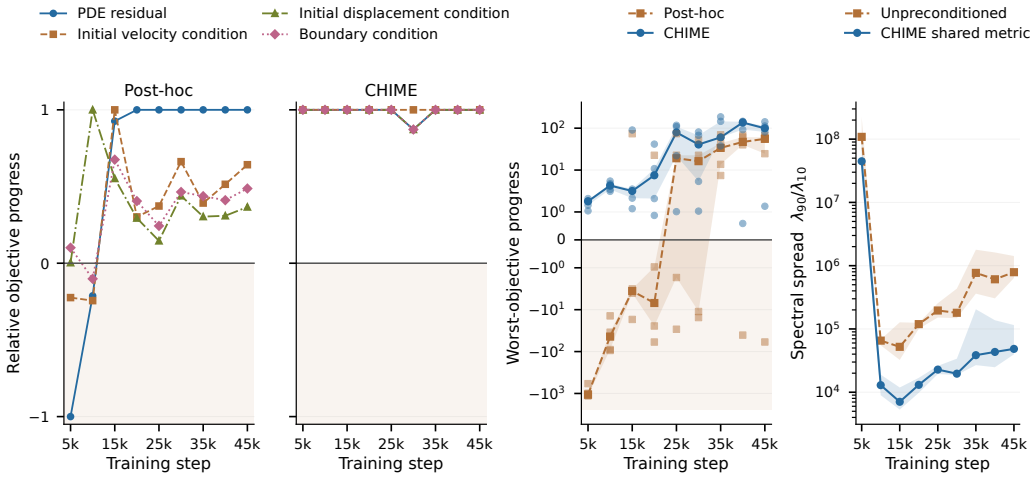


Figure 9: **Objective progress and residual spectral disparity at stored Wave1d-C states.** *Left:* Sign-preserving normalized objective progress. *Center:* Unscaled worst-objective progress for each seed. Both selections use the same preconditioner and unit metric step norm. *Right:* Eigenvalue-quantile ratio on a common subspace retaining 99% of the diagnostic joint residual squared norm. Center/right lines and bands show five-seed medians and interquartile ranges. Protocols are given in Apps. B.5.2–B.5.3.

B.5.1 Experimental setup

We repeat Wave1d-C over five random seeds for 50,000 training steps. The PINN has three hidden layers with 50 hidden units per layer and $\tanh$ activations, and is trained with learning rate 10^{-4} using the SOAP-style shared preconditioner in CHIME, refreshed every two steps. We evaluate the diagnostics at $5{,}000, 10{,}000, \ldots, 45{,}000$ training steps. Objective-wise progress is computed using the full training residual set. For the spectral calculation, we use a fixed subset of 128 residual rows from each objective throughout a run, so that the same residual locations are compared across training states.

B.5.2 Metric-consistent objective selection

At each training state, we compute L_i, $\mathbf{g}_i$, and the residual-normalized gradients $\widetilde{\mathbf{g}}_i = \mathbf{g}_i/(2L_i)$. We compare two coefficient selections while keeping the same shared metric $\mathbf{P}$ for the resulting

parameter update. The post-hoc variant selects coefficients in Euclidean geometry,

$$\boldsymbol{\alpha}_{\mathrm{E}} \in \arg\min_{\boldsymbol{\alpha}\in\Delta_m} \frac{1}{2}\boldsymbol{\alpha}^\top \widetilde{\mathbf{G}}^\top \widetilde{\mathbf{G}}\boldsymbol{\alpha}, \tag{86}$$

and then applies $\mathbf{P}$, whereas CHIME selects coefficients directly in the same metric used for the update through Eq. (17). Both simplex problems are solved with ProxQP. To compare directions at the same metric step size, we use

$$\widehat{\mathbf{d}}(\boldsymbol{\alpha}) = \frac{\mathbf{P}\widetilde{\mathbf{G}}\boldsymbol{\alpha}}{\sqrt{\boldsymbol{\alpha}^\top \mathbf{C}\boldsymbol{\alpha}}}, \qquad \|\widehat{\mathbf{d}}(\boldsymbol{\alpha})\|_{\mathbf{B}} = 1, \quad \boldsymbol{\alpha} \in \{\boldsymbol{\alpha}_{\mathrm{E}}, \boldsymbol{\alpha}^\star\}. \tag{87}$$

We then evaluate the objective-wise first-order progress $p_i(\widehat{\mathbf{d}}) = \widetilde{\mathbf{g}}_i^\top \widehat{\mathbf{d}}$. This comparison isolates the effect of choosing the objective combination before or within the curvature geometry.

For the left region of Fig. 9, we show a seed and normalize the displayed progress within each method and training state as

$$\bar{p}_i(\widehat{\mathbf{d}}) = \frac{p_i(\widehat{\mathbf{d}})}{\max_j |p_j(\widehat{\mathbf{d}})|}. \tag{88}$$

The center region reports the worst-objective progress $\min_i p_i(\widehat{\mathbf{d}})$ across all five seeds, with median and interquartile range at each training step. Post-hoc selection yields a negative worst-objective progress in 24 of the 45 analyzed states, whereas the metric-consistent CHIME direction has positive and larger worst-objective progress in all 45 states. These observations empirically support Proposition 1 and the metric-consistent direction selection characterized in Proposition 2.

B.5.3 RESIDUAL-SPECTRUM CONDITIONING

We next examine the conditioning term appearing in Theorem 1. Using the CHIME-selected mixture $\boldsymbol{\alpha}^\star$, let $\mathbf{r}_i^{\mathrm{sub}}$ and $\mathbf{J}_i^{\mathrm{sub}}$ denote the residual vector and Jacobian on the fixed spectral subset. For objectives with $\alpha_i^\star > 0$, define

$$\mathbf{e}_{\boldsymbol{\alpha}}^{\mathrm{sub}} = \begin{bmatrix} \frac{\sqrt{\alpha_1^\star}\mathbf{r}_1^{\mathrm{sub}}}{\|\mathbf{r}_1^{\mathrm{sub}}\|_2} \\ \vdots \\ \frac{\sqrt{\alpha_m^\star}\mathbf{r}_m^{\mathrm{sub}}}{\|\mathbf{r}_m^{\mathrm{sub}}\|_2} \end{bmatrix}, \qquad \mathbf{J}_{\boldsymbol{\alpha}}^{\mathrm{sub}} = \begin{bmatrix} \frac{\sqrt{\alpha_1^\star}\mathbf{J}_1^{\mathrm{sub}}}{\|\mathbf{r}_1^{\mathrm{sub}}\|_2} \\ \vdots \\ \frac{\sqrt{\alpha_m^\star}\mathbf{J}_m^{\mathrm{sub}}}{\|\mathbf{r}_m^{\mathrm{sub}}\|_2} \end{bmatrix}. \tag{89}$$

Zero-weight blocks contribute neither residual energy nor positive eigenvalues and are omitted in computation. We compare the unpreconditioned and shared-metric residual kernels

$$\mathbf{K}_{\mathrm{I}} = \mathbf{J}_{\boldsymbol{\alpha}}^{\mathrm{sub}}(\mathbf{J}_{\boldsymbol{\alpha}}^{\mathrm{sub}})^\top, \qquad \mathbf{K}_{\mathrm{P}} = \mathbf{J}_{\boldsymbol{\alpha}}^{\mathrm{sub}}\mathbf{P}(\mathbf{J}_{\boldsymbol{\alpha}}^{\mathrm{sub}})^\top. \tag{90}$$

Thus the objective mixture and residual samples are identical, and the only difference is whether the shared metric $\mathbf{P}$ is applied.

To compare both spectra on the same residual subspace, we eigendecompose $\mathbf{K}_{\mathrm{I}}$ and select the smallest set of its eigenvectors whose squared projections of $\mathbf{e}_{\boldsymbol{\alpha}}^{\mathrm{sub}}$ account for at least 99% of the residual energy. Let $\mathbf{U}_A$ collect these eigenvectors and define

$$\mathbf{K}_{\mathrm{I}}^A = \mathbf{U}_A^\top \mathbf{K}_{\mathrm{I}} \mathbf{U}_A, \qquad \mathbf{K}_{\mathrm{P}}^A = \mathbf{U}_A^\top \mathbf{K}_{\mathrm{P}} \mathbf{U}_A. \tag{91}$$

We summarize spectral disparity by

$$S_{90/10}(\mathbf{K}^A) = \frac{Q_{0.90}(\operatorname{eig}\mathbf{K}^A)}{Q_{0.10}(\operatorname{eig}\mathbf{K}^A)}, \tag{92}$$

where Q_p denotes the pth eigenvalue quantile. This robust central-spectrum ratio serves as an empirical conditioning diagnostic.

The right region of Fig. 9 shows the median and interquartile range of $S_{90/10}$ across the five seeds. The shared metric reduces the spectral spread in all 45 analyzed states, with a median paired reduction factor of $8.31\times$. Together with the objective-progress diagnostic, this supports the two factors in Theorem 1: CHIME selects a direction with positive worst-objective progress while its shared curvature metric reduces disparity among the active residual modes.

B.6 IMPLEMENTATION DETAILS

Active objectives and initialization. Algorithm 1 shows the regular update after initialization. Inactive objectives are masked before the normalization in Eq. (23) and receive zero coefficients. If the previous coefficients have no mass on the current active set, the lagged weights are uniform on that set; an empty active set gives zero current coefficients. At initialization, the same simplex QP is solved using the Euclidean gradient Gram. Its aggregate initializes the shared Gram statistics and eigenbases, while both moments are set to zero. The counter t then counts post-initialization updates. On each such update, all blocks accumulate the lagged aggregate in their gradient statistics and second moment in the old basis, then perform any scheduled basis refresh. Only after these operations do they contribute to the global objective Gram. A single coefficient vector is solved for the entire network, and every block uses it for the current numerator and parameter update.

Shared operator and stabilization. Each matrix block maintains one left/right pair of gradient Gram statistics and one pair of moments, shared across objectives. The statistics accumulate outer products of the lagged aggregate $\mathbf{A}$, not separate objective outer products, and therefore retain cross terms between objective gradients. With column-major vectorization, the layerwise operator underlying Eq. (24) is

$$\mathbf{P}_t^{(\ell)} = \mathbf{Q}_{t,\ell}\,\mathrm{Diag}(\mathrm{vec}\,\mathbf{S}_t^{(\ell)})\mathbf{Q}_{t,\ell}^\top, \qquad \mathbf{Q}_{t,\ell} = \mathbf{Q}_{\mathrm{R},t}^{(\ell)} \otimes \mathbf{Q}_{\mathrm{L},t}^{(\ell)}, \qquad \mathbf{S}_t = \frac{\sqrt{1-\beta_2^t}}{\sqrt{\mathbf{V}_t}+\epsilon}. \tag{93}$$

The network operator is block diagonal, and the same multiplier is used in the objective Gram and parameter update. The implementation stabilizes divisions with $\epsilon = 10^{-8}$ and safeguards vanishing bias-correction factors. The second moment is stored without bias correction; the stabilizer is added to $\sqrt{\mathbf{V}_t}$ before the multiplicative correction shown above. The numerator uses the bias-corrected first moment when momentum is enabled and the current aggregate otherwise. The class defaults are $(\beta_1, \beta_2) = (0.9, 0.999)$, $\beta_3 = \beta_2$, refresh frequency $f = 2$, and zero Gram damping; benchmark protocols are reported separately in Apps. C and F.

Basis refresh. After updating the statistics and second moment in the old basis, each scheduled refresh performs a SOAP-style QR step (Vyas et al., 2025). For $a \in \{\mathrm{L}, \mathrm{R}\}$, let π_a sort the diagonal of $(\mathbf{Q}_a^{\mathrm{old}})^\top \mathbf{F}_a \mathbf{Q}_a^{\mathrm{old}}$ in descending order and set

$$\mathbf{Q}_a^{\mathrm{new}} = \mathrm{qf}\big(\mathbf{F}_a \mathbf{Q}_a^{\mathrm{old}}[:, \pi_a]\big)\,, \tag{94}$$

where qf returns the orthogonal factor of a QR decomposition. The stored moments are updated by

$$\begin{aligned}(\mathbf{V}^{\mathrm{new}})_{jk} &= (\mathbf{V}^{\mathrm{old}})_{\pi_{\mathrm{L}}(j),\pi_{\mathrm{R}}(k)},\\ \mathbf{M}^{\mathrm{new}} &= (\mathbf{Q}_{\mathrm{L}}^{\mathrm{new}})^\top \mathbf{Q}_{\mathrm{L}}^{\mathrm{old}} \mathbf{M}^{\mathrm{old}} (\mathbf{Q}_{\mathrm{R}}^{\mathrm{old}})^\top \mathbf{Q}_{\mathrm{R}}^{\mathrm{new}}.\end{aligned} \tag{95}$$

The first moment therefore represents the same parameter-space vector after refresh; the diagonal second moment is only permuted. Higher-order tensors use the same operation axis by axis, with identity rotations on unpreconditioned axes.

Coefficient selection and guarantees. The active-objective Gram is symmetrized and its simplex QP is solved by warm-started ProxQP, with feasibility and optimality checks. The resulting vector is shared across all layer blocks and extended by zero outside the active set. For the exact undamped QP, Eq. (43) gives nonnegative first-order progress for every active objective, including those with zero coefficients. The non-conflict result in Proposition 1 applies to the instantaneous metric-selected direction; when $\beta_1 > 0$, Algorithm 1 additionally performs SOAP-style temporal averaging, which preserves non-conflict whenever the preconditioned transported-history contribution also lies in the current dual cone $\mathcal{K}^*$.

Computational Cost. For matrix blocks, persistent curvature and moment storage is $O\big(\sum_\ell[(d_\ell^{\mathrm{r}})^2 + (d_\ell^{c})^2 + D_\ell]\big)$. Objective gradients require $O(mD)$ storage, and the Gram requires $O(m^2)$ storage and $O(m^2 D)$ contractions after rotation. The only coefficient optimization is a dense QP with at most m variables, independent of D.

C PINN BENCHMARKS AND ADDITIONAL RESULTS

C.1 EXPERIMENTAL PROTOCOL

Benchmark construction and training. We follow the problem definitions, network architectures, sampling configurations, learning rates, and random seeds of HARMONIC (Bu et al., 2026). The main suite contains three forward PDE problems (Wave1d-C, Poisson2d-C, and the five-dimensional heat equation HNd) and the inverse PDE problem HInv from PINNacle (Hao et al., 2024), together with the forward IDE problem Volterra1d from A-PINN (Yuan et al., 2022). Before baseline-specific grouping, Wave1d-C and Volterra1d have four loss terms each; Poisson2d-C, HNd, and HInv have three each. The five main benchmarks use fully connected networks with three hidden layers of 50 neurons, $\tanh$ activations, and Glorot normal initialization (Glorot and Bengio, 2010). The fixed-step budget is $50{,}000$ training steps, using five seeds. Architectural exceptions are summarized in Table 4, and each problem family below includes its sampling table. For the additional PINNacle and A-PINN experiments, we likewise adopt the settings used by Bu et al. (2026); the corresponding equations, domains, conditions, coefficient choices, reference solutions, and sampling configurations are reproduced below for completeness. All derivatives in the residual objectives are computed by automatic differentiation, and each listed physical condition defines its corresponding mean-squared residual objective before any baseline-specific grouping. The sampling tables retain the reference protocol's nominal collocation, initial, boundary, and test points. For stationary problems, the reference protocol may list initial points even though no initial-condition objective is defined; actual residual-row counts can also differ from nominal point counts when a sampled point contributes to multiple residual groups. CHIME-specific normalization and update details are given in App. B.6. The problem-specific confidence-interval tables below use the fixed learning rate 10^{-3}; App. E separately compares 10^{-3} and 10^{-4} on the five main benchmarks.

Baseline implementations. The objective-aware comparisons use the HARMONIC study's results where available, supplemented by the DCGD and curvature-aware evaluations in Table 2. For the forward problems, DCGD and KFAC* use two loss groups: one combines the governing-equation residuals, and the other combines initial and boundary conditions. Muon* and SOAP* denote the LRA-style loss-balanced variants (Wang et al., 2025). The learning-rate and wall-clock tables retain these same method names.

Error and uncertainty. For reference values $\mathbf{u}_{\mathrm{ref}}$ and predictions $\widehat{\mathbf{u}}$ at the test points, the relative L^2 error is

$$E = \frac{\|\widehat{\mathbf{u}} - \mathbf{u}_{\mathrm{ref}}\|_2}{\|\mathbf{u}_{\mathrm{ref}}\|_2}. \tag{96}$$

For inverse problems, the same metric is applied to the recovered coefficient field. Table 2 reports means and standard deviations, whereas the supplementary problem and learning-rate tables report 95% confidence intervals. Their † markers indicate interval overlap.

Forward improvement factors. For each baseline family $\mathcal{B}$ and forward problem p, let $E_{a,p}$ denote the displayed mean in Table 2 and define

$$R_{\mathcal{B},p} = \frac{\min_{a\in\mathcal{B}} E_{a,p}}{E_{\mathrm{CHIME},p}}, \qquad R_{\mathcal{B}} = \left(\prod_{p\in\mathcal{F}} R_{\mathcal{B},p}\right)^{1/4}, \tag{97}$$

where $\mathcal{F} = \{\mathrm{Wave1d\text{-}C}, \mathrm{Poisson2d\text{-}C}, \mathrm{HNd}, \mathrm{Volterra1d}\}$. The objective-aware family is Multi-Adam through HARMONIC in that table; the curvature-aware family is KFAC*, Muon*, and SOAP*. Equation (97) gives $R_{\mathcal{B}} \approx 5.85$ and 5.45, respectively. These are geometric means from table entries, selecting the best baseline separately for each problem.

C.2 WAVE EQUATION

1D Wave Equation (Wave1d-C) The governing equation is

$$u_{tt} - 4\,u_{xx} = 0. \tag{98}$$

Table 4: Network architectures for the PINN and A-PINN experiments, following Bu et al. (2026). All use $\tanh$ activations, Glorot normal initialization, and the common 50,000-step, five-seed protocol.

Problems	Hidden layers	Width
Wave1d-C, Poisson2d-C, Poisson2d-CG, HNd, HInv, Volterra1d, Burgers1d, Burgers2d, Heat2d-VC, Heat2d-CG	3	50
Heat2d-MS and the Volterra system	5	100
NS2d-C and NS-CG	3	100
PInv	1	100

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

$$u(x, 0) = \sin(\pi x) + \frac{1}{2}\sin(4\pi x), \tag{99}$$

$$u_t(x, 0) = 0. \tag{100}$$

The boundary conditions are

$$u(0, t) = u(1, t) = 0. \tag{101}$$

The analytical solution of this problem is

$$u(x, t) = \sin(\pi x)\cos(2\pi t) + \frac{1}{2}\sin(4\pi x)\cos(8\pi t). \tag{102}$$

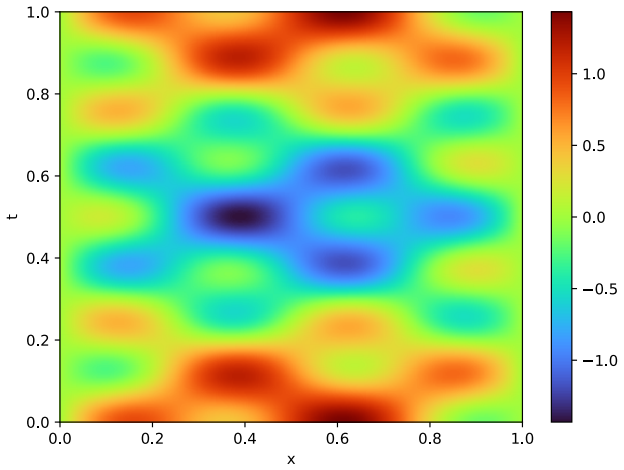


Figure 10: Reference solution of Wave1d-C using PINNacle reference data.

Table 5: Nominal sampling sizes from the HARMONIC/PINNacle protocol; see App. C.1 for the sampling-point convention.

	Wave1d-C
Collocation points	8192
Initial points	2048
Boundary points	2048
Test points	8192

C.3 Poisson equations

Poisson 2D Classic (Poisson2d-C) The PDE is given by

$$u_{xx} + u_{yy} = 0. \tag{103}$$

Table 6: 95% confidence intervals of relative L^2 errors over five random seeds at learning rate 10^{-3}. In each column, the lowest-mean result(s) are shown in **bold**, and † denotes a non-bold method whose 95% confidence interval overlaps with that of at least one bolded result.

Method	**Wave1d-C**
MultiAdam	$[1.0758, 1.2322]$
LRA	$[0.1935, 0.3281]$
ReLoBRaLo	$[0.3511, 0.4693]$
MGDA	$[0.2940, 0.6314]$
PCGrad	$[0.1386, 0.3126]$
CAGrad	$[0.1098, 0.1274]$
IMTL-G	$[0.1757, 1.3571]$
Aligned-MTL	$[0.1639, 0.5963]$
ConFIG	$[0.0322, 0.1014]$
HARMONIC	$[0.0291, 0.1019]$
CHIME	$[\mathbf{0.0042}, \mathbf{0.0064}]$

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:

$$R_1 = \{(x, y) : (x - 0.3)^2 + (y - 0.3)^2 \leq 0.1^2\}, \tag{104}$$

$$R_2 = \{(x, y) : (x + 0.3)^2 + (y - 0.3)^2 \leq 0.1^2\}, \tag{105}$$

$$R_3 = \{(x, y) : (x - 0.3)^2 + (y + 0.3)^2 \leq 0.1^2\}, \tag{106}$$

$$R_4 = \{(x, y) : (x + 0.3)^2 + (y + 0.3)^2 \leq 0.1^2\}. \tag{107}$$

The boundary condition is

$$u(x, y) = 0, (x, y) \in \partial R_i, \tag{108}$$

$$u(x, y) = 1, (x, y) \in \partial \Omega_{\text{rec}}. \tag{109}$$

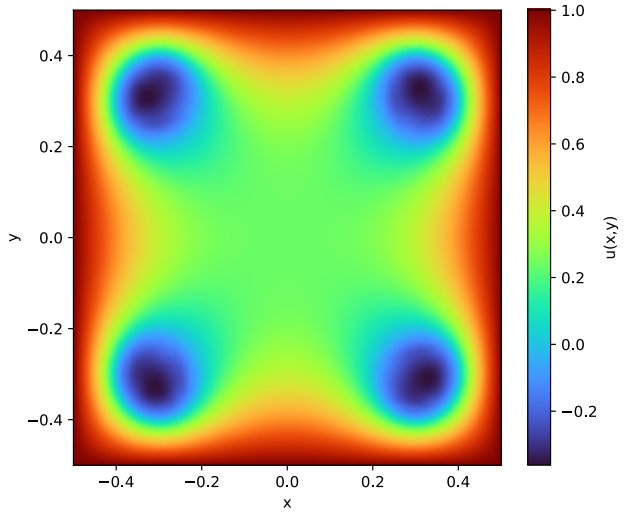


Figure 11: Reference solution of Poisson2d-C using PINNacle reference data.

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

$$-\Delta u + k^2 u = f(x, y), \tag{110}$$

where the prescribed forcing is

$$f(x, y) = A(\mu_1^2 + \mu_2^2 + x^2 + y^2) \sin(\mu_1 \pi x) \sin(\mu_2 \pi y). \tag{111}$$

The domain is $\Omega = \Omega_{\text{rec}} \setminus \bigcup_{i=1}^{4} R_i$, where $\Omega_{\text{rec}} = [-1, 1]^2$ and the removed disks are

$$R_1 = \{(x, y) : (x - 0.5)^2 + (y - 0.5)^2 \leq 0.2^2\}, \tag{112}$$

$$R_2 = \{(x, y) : (x - 0.4)^2 + (y + 0.4)^2 \leq 0.4^2\}, \tag{113}$$

$$R_3 = \{(x, y) : (x + 0.2)^2 + (y + 0.7)^2 \leq 0.1^2\}, \tag{114}$$

$$R_4 = \{(x, y) : (x + 0.6)^2 + (y - 0.5)^2 \leq 0.3^2\}. \tag{115}$$

The boundary conditions are

$$u(x, y) = 0.2, \quad (x, y) \in \partial\Omega_{\text{rec}}, \tag{116}$$

$$u(x, y) = 1, \quad (x, y) \in \bigcup_i \partial R_i. \tag{117}$$

The parameters are

$$\mu_1 = 1, \ \mu_2 = 4, \ k = 8, \ A = 10. \tag{118}$$

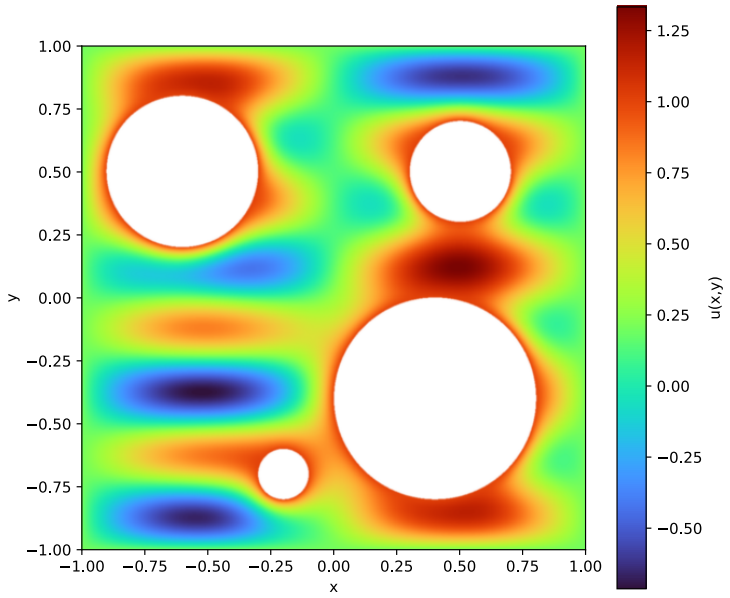


Figure 12: Reference solution of Poisson2d-CG using PINNacle reference data.

Table 7: Nominal sampling sizes from the HARMONIC/PINNacle protocol; see App. C.1 for the sampling-point convention.

	Poisson2d-C	**Poisson2d-CG**
Collocation points	8192	8192
Initial points	2048	2048
Boundary points	2048	2048
Test points	8192	8192

C.4 Heat Equations

N Dimensional Heat Equation (HNd) The governing equation is

$$u_t = k\Delta u + f(\boldsymbol{x}, t), \tag{119}$$

where

$$f(\boldsymbol{x}, t) = -\frac{1}{d} \|\boldsymbol{x}\|_2^2 \exp\left(\tfrac{1}{2}\|\boldsymbol{x}\|_2^2 + t\right). \tag{120}$$

Table 8: 95% confidence intervals of relative L^2 errors over five random seeds at learning rate 10^{-3}. In each column, the lowest-mean result(s) are shown in **bold**, and † denotes a non-bold method whose 95% confidence interval overlaps with that of at least one bolded result.

Method	**Poisson2d-C**	**Poisson2d-CG**
MultiAdam	[0.5715, 0.7751]	[0.7955, 0.8253]
LRA	[0.3732, 0.7692]	[0.2598, 0.3710]
ReLoBRaLo	[0.6328, 0.6876]	[0.6016, 0.6660]
MGDA	[0.6480, 0.7074]	[0.7137, 0.7835]
PCGrad	[0.6682, 0.7154]	[0.1355, 0.2473]
CAGrad	[0.5878, 0.9734]	[0.2844, 0.4386]
IMTL-G	[0.4190, 0.5888]	[0.0000, 2.4135]†
Aligned-MTL	[0.0000, 0.5781]†	[0.6507, 0.7541]
ConFIG	[0.1620, 1.2288]	[0.0000, 0.3662]†
HARMONIC	[0.0000, 0.0436]†	[0.0000, 0.3707]†
CHIME	[**0.0029**, **0.0053**]	[**0.0056**, **0.0094**]

The spatial domain is the five-dimensional unit ball, $\Omega = \{\boldsymbol{x} \in \mathbb{R}^5 : \|\boldsymbol{x}\|_2 < 1\}$, and $T = [0, 1]$.

The initial and boundary conditions are

$$\boldsymbol{n}\cdot\nabla u = g(\boldsymbol{x}, t),\ (\boldsymbol{x}, t) \in \partial\Omega \times [0, 1], \tag{121}$$

$$u(\boldsymbol{x}, 0) = g(\boldsymbol{x}, 0),\ \boldsymbol{x} \in \Omega, \tag{122}$$

where

$$g(\boldsymbol{x}, t) = \exp\left(\tfrac{1}{2}\|\boldsymbol{x}\|_2^2 + t\right). \tag{123}$$

The dimension and diffusion coefficient are

$$d = 5, \qquad k = \frac{1}{d}. \tag{124}$$

2D Heat with Varying Coefficients (Heat2d-VC) The governing equation is

$$u_t - \nabla\cdot(a(x, y)\nabla u) = f(x, y, t). \tag{125}$$

The domain is

$$(x, y, t) \in \Omega \times T\ =\ [0, 1]^2 \times [0, 5]. \tag{126}$$

The positive coefficient field is an exponential Gaussian random field, $a(x, y) = \exp(\xi(x, y))$. The registered benchmark realization is tabulated on a 256×256 grid in `ref/heat_2d_coef_256.dat`; the same prescribed field is used across methods rather than drawing a new PDE for each optimizer. The benchmark generator constructs a real Fourier Gaussian field with amplitudes proportional to $(k_x^2 + k_y^2)^{-1}$ for nonzero wavevectors, zero mean mode, and sample normalization to zero mean and unit standard deviation before exponentiation (Hao et al., 2024).

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

$$f(x, y, t) = A\ \sin(m_1\pi x)\ \sin(m_2\pi y)\ \sin(m_3\pi t). \tag{127}$$

The initial and boundary conditions are

$$u(x, y, 0) = 0,\ \ (x, y) \in \Omega, \tag{128}$$

$$u(x, y, t) = 0,\ \ (x, y) \in \partial\Omega. \tag{129}$$

The parameters are

$$A = 200,\ m_1 = 1,\ m_2 = 5,\ m_3 = 1. \tag{130}$$

2D Heat Multi-Scale (Heat2d-MS) The PDE is

$$u_t - \frac{1}{(500\pi)^2}u_{xx} - \frac{1}{\pi^2}u_{yy} = 0. \tag{131}$$

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

The initial and boundary conditions are

$$u(x, y, 0) = \sin(20\pi x)\ \sin(\pi y), \qquad (x, y) \in \Omega, \tag{132}$$

$$u(x, y, t) = 0, \qquad (x, y) \in \partial\Omega,\ t \in T. \tag{133}$$

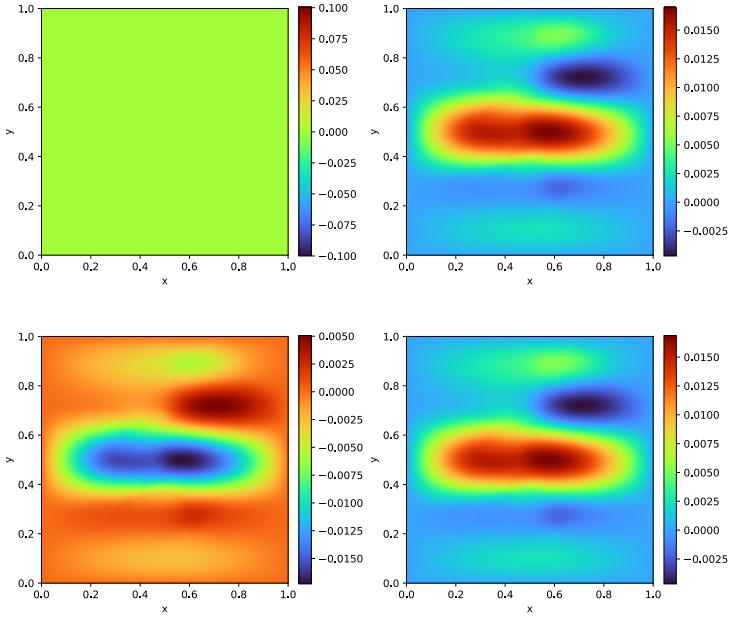


Figure 13: Reference solution of Heat2d-VC at times $t = 0, 0.3, 0.7, 1$ using PINNacle reference data.

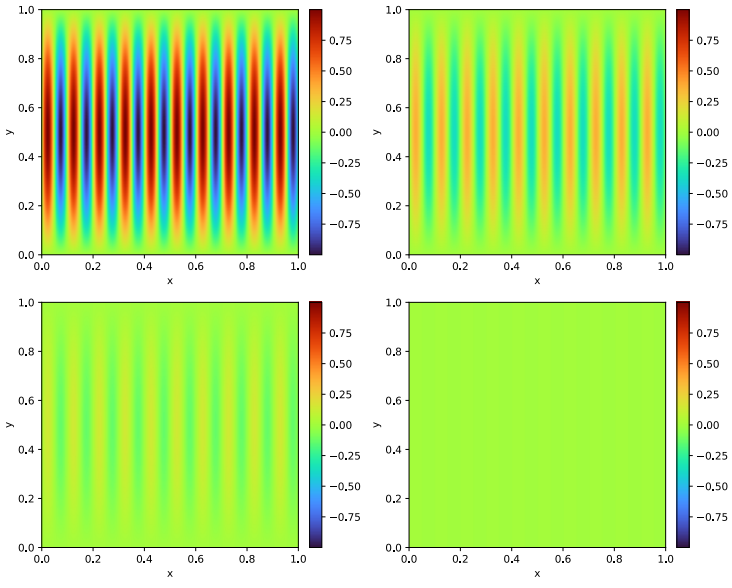


Figure 14: Reference solution of Heat2d-MS at times $t = 0, 1, 2, 5$ using PINNacle reference data.

2D Heat Complex Geometry (Heat2d-CG) The PDE is

$$u_t - \Delta u = 0. \tag{134}$$

The domain is defined as

$$(x, y, t) \in \Omega \times T = ([-8, 8] \times [-12, 12] \setminus \cup_i R_i) \times [0, 3]. \tag{135}$$

The initial condition is

$$u(x, y, 0) = 0. \tag{136}$$

The removed disks comprise radius-one disks centered at

$$\mathcal{C}_{\text{large}} = \{(\pm 4, \pm 3), (\pm 4, \pm 9), (0, 0), (0, 6), (0, -6)\}, \tag{137}$$

where the two signs vary independently, and radius-0.4 disks centered at

$$\mathcal{C}_{\text{small}} = \{(\pm 3.2, -6), (\pm 3.2, 0), (\pm 3.2, 6)\}. \tag{138}$$

With $\boldsymbol{n}$ the outward unit normal of the perforated domain, the Robin conditions are

$$\boldsymbol{n} \cdot \nabla u = \begin{cases} 5 - u, & \text{on the large circular boundaries,} \\ 1 - u, & \text{on the small circular boundaries,} \\ 0.1 - u, & \text{on the outer rectangular boundary.} \end{cases} \tag{139}$$

The disk locations and boundary assignments follow the registered PINNacle geometry (Hao et al., 2024).

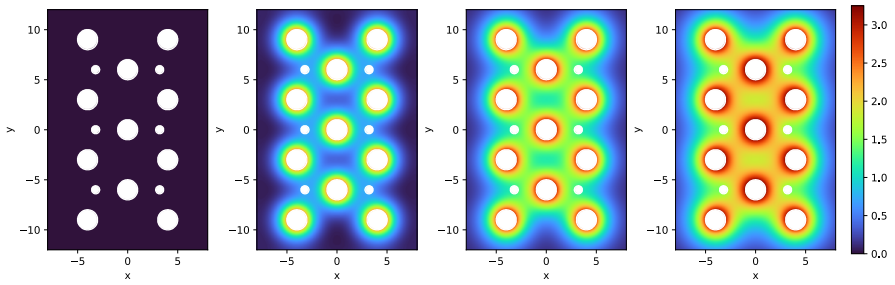


Figure 15: Reference solution of Heat2d-CG at times $t = 0, 1, 2, 3$ using PINNacle reference data.

Table 9: Nominal sampling sizes from the HARMONIC/PINNacle protocol; see App. C.1 for the sampling-point convention.

	Heat2d-VC	**Heat2d-MS**	**Heat2d-CG**	**HNd**
Collocation points	32768	32768	32768	32768
Initial points	8192	8192	8192	8192
Boundary points	8192	8192	8192	8192
Test points	32768	32768	32768	32768

C.5 Inverse problems

Heat (Diffusion) inverse problem (HInv) The PDE is given by

$$u_t - \nabla \cdot (a \nabla u) = f(x, y, t), \tag{140}$$

where the source function $f(x, y, t)$ is

$$f(x, y, t) = \Big((4\pi^2 - 1) \sin(\pi x) \sin(\pi y) + \pi^2 \big(2 \sin^2(\pi x) \sin^2(\pi y) - \cos^2(\pi x) \sin^2(\pi y) - \sin^2(\pi x) \cos^2(\pi y) \big) \Big) e^{-t}. \tag{141}$$

Table 10: 95% confidence intervals of relative L^2 errors over five random seeds at learning rate 10^{-3}. In each column, the lowest-mean result(s) are shown in **bold**, and † denotes a non-bold method whose 95% confidence interval overlaps with that of at least one bolded result.

Method	**Heat2d-VC**	**Heat2d-MS**	**Heat2d-CG**	**HNd**
MultiAdam	[1.0159, 1.1917]	[0.9961, 1.0013]	[0.6544, 0.7554]	[0.0000, 0.0742]†
LRA	[0.2395, 0.3573]	[**0.0191**, **0.0365**]	[0.1388, 0.4472]	[0.0006, 0.0034]†
ReLoBRaLo	[1.6805, 1.9805]	[0.0092, 0.0618]†	[0.1015, 0.1099]	[0.0000, 0.0117]†
MGDA	[0.9859, 1.4455]	[0.0000, 0.2582]†	[0.1127, 0.1283]	[0.0000, 0.0157]†
PCGrad	[0.9566, 1.2830]	[0.0124, 0.0524]†	[0.0913, 0.1069]	[0.0004, 0.0052]†
CAGrad	[0.2812, 0.3898]	[0.0223, 0.0661]†	[0.1157, 0.1189]	[**0.0004**, **0.0006**]
IMTL-G	[0.6623, 1.8237]	[0.9899, 0.9967]	[0.0943, 0.1035]	[0.0004, 0.0008]†
Aligned-MTL	[0.0757, 1.7403]†	[0.4496, 1.5202]	[0.1102, 0.1276]	[0.0004, 0.0008]†
ConFIG	[0.1162, 0.4462]†	[0.2892, 1.3350]	[0.1062, 0.1328]	[0.0005, 0.0007]†
HARMONIC	[**0.1839**, **0.2015**]	[0.0000, 0.0728]†	[0.0841, 0.0995]	[**0.0005**, **0.0005**]
CHIME	[0.2283, 0.2753]	[0.0000, 0.1558]†	[**0.0385**, **0.0481**]	[0.0002, 0.0014]†

The geometric domain is $(x, y, t) \in \Omega \times T = [-1, 1]^2 \times [0, 1]$, and

$$u = e^{-t} \sin(\pi x) \sin(\pi y). \tag{142}$$

We impose a boundary condition for the diffusion coefficient field as

$$a(x, y) = 2, \qquad (x, y) \in \partial\Omega. \tag{143}$$

We sample data of $u(x, y, t)$ randomly with 2500 points from the temporal domain $\Omega \times T$ and add independent Gaussian noise with standard deviation 0.1, i.e., $\mathcal{N}(0, 0.1^2)$ to it. The ground truth is

$$a(x, y) = 2 + \sin(\pi x) \sin(\pi y), \qquad (x, y) \in \Omega. \tag{144}$$

Poisson inverse problem (PInv) The governing PDE is

$$-\nabla \cdot (a \nabla u) = f(x, y) \tag{145}$$

The source term f is

$$f(x, y) = \frac{2\pi^2 \sin(\pi x) \sin(\pi y)}{1 + x^2 + y^2 + (x-1)^2 + (y-1)^2} + \frac{2\pi\big((2x-1)\cos(\pi x)\sin(\pi y) + (2y-1)\sin(\pi x)\cos(\pi y)\big)}{(1 + x^2 + y^2 + (x-1)^2 + (y-1)^2)^2}. \tag{146}$$

The geometric domain is $(x, y) \in \Omega = [0, 1]^2$, and

$$u(x, y) = \sin(\pi x) \sin(\pi y). \tag{147}$$

We prescribe the coefficient on the boundary,

$$a(x, y) = \frac{1}{1 + x^2 + y^2 + (x-1)^2 + (y-1)^2}, \qquad (x, y) \in \partial\Omega. \tag{148}$$

We sample data of $u(x, y)$ with 2500 uniformly distributed 50×50 points and add independent Gaussian noise with standard deviation 0.1, i.e., $\mathcal{N}(0, 0.1^2)$ to it. The ground truth of $a(x, y)$ is

$$a(x, y) = \frac{1}{1 + x^2 + y^2 + (x-1)^2 + (y-1)^2}, \qquad (x, y) \in \Omega. \tag{149}$$

C.5.1 EARLY CONVERGENCE AND LONG-HORIZON BEHAVIOR ON HINV

HInv combines noisy observations of the state u with deterministic PDE and coefficient-boundary constraints. Figures 16 and 17 show the first 2,000 steps and the complete 50,000-step trajectories for the same five methods. Both views are needed to distinguish fast initial recovery from the subsequent relationship between training losses and solution accuracy. The last panel reports the plotted relative-error diagnostic for u and a together; the benchmark table separately reports coefficient-recovery error.

Initial 2,000 steps. CHIME reduces the PDE residual by several orders of magnitude while also decreasing the observation and coefficient-boundary losses. Its solution-and-coefficient error falls to approximately 6×10^{-2} within the first 1,000 steps and remains substantially below the other displayed methods for the rest of the 2,000-step window. ConFIG and HARMONIC reduce the data and boundary losses rapidly, but their PDE residual and solution error improve more slowly. Thus, the early advantage is not simply a smaller value of one training term: it appears jointly in the physics residual and the recovery diagnostic.

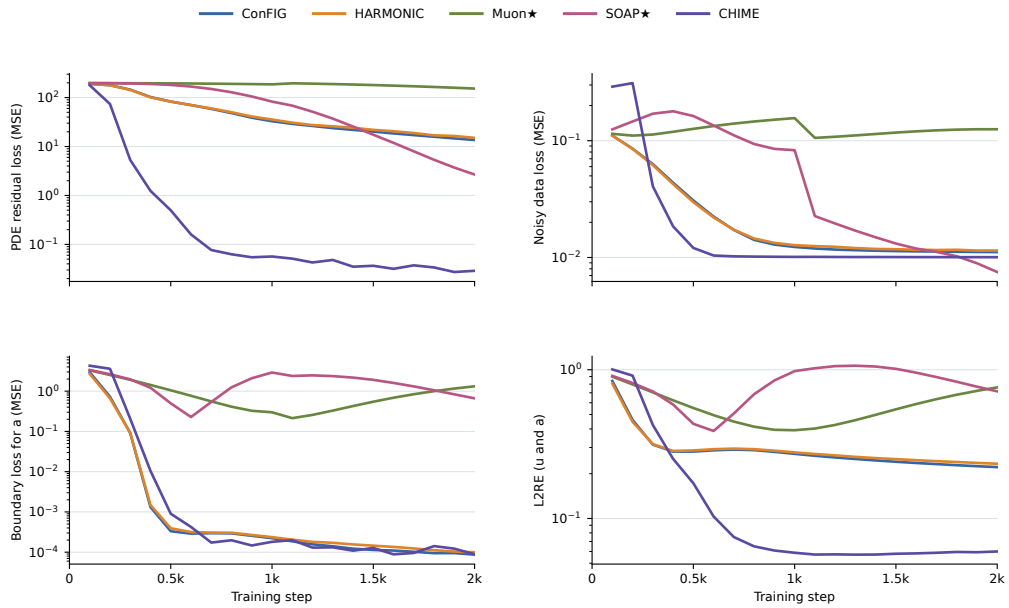


Figure 16: HInv over the first 2,000 training steps. The panels show the PDE residual loss, noisy-data loss, boundary loss for a, and the plotted relative L^2 error for u and a. CHIME rapidly reduces the PDE residual and recovery error, revealing an initial convergence advantage that is difficult to distinguish on the full-horizon axes.

Full 50,000 steps. The longer trajectories reveal a different tradeoff. Muon* and SOAP* continue to reduce the noisy-data and coefficient-boundary losses, and SOAP* also attains a smaller PDE residual over a substantial part of training. Nevertheless, their recovery errors rise after reaching intermediate minima even while several empirical objectives continue to decrease. This loss–recovery separation indicates that rapid optimization of the empirical objectives need not translate into stable recovery when the observation objective is noisy. ConFIG and HARMONIC instead reach a comparatively stable lower-error regime, whereas CHIME retains its rapid initial improvement but subsequently plateaus. Related analyses of Gauss–Newton optimization likewise show that curvature-aware training can have delicate generalization behavior: exact Gauss–Newton may fit individual mini-batches aggressively without transferring that progress to other data (Buffelli et al., 2024), while a recent stability analysis of stochastic Gauss–Newton makes batch size an explicit factor and obtains tighter stability bounds for smaller batches in its favorable regime (Cayci, 2026). These results motivate treating observation sampling as part of the training design rather than attributing the HInv behavior to curvature information alone.

C.5.2 EFFECT OF OBSERVATION MINI-BATCHING ON HINV

The reference HInv protocol evaluates the noisy-data objective using all 2,500 observations in a single batch. To isolate the effect of observation sampling from the loss threshold used by the main HInv configuration, we set $\tau_{\text{data}} = 0$ in this ablation and mini-batch only the noisy-data objective. Specifically, the observations are divided into $N_{\text{mb}} \in \{2, 8, 32, 128\}$ mini-batches per sampling cycle, so larger N_{mb} corresponds to smaller observation mini-batches. The PDE residual

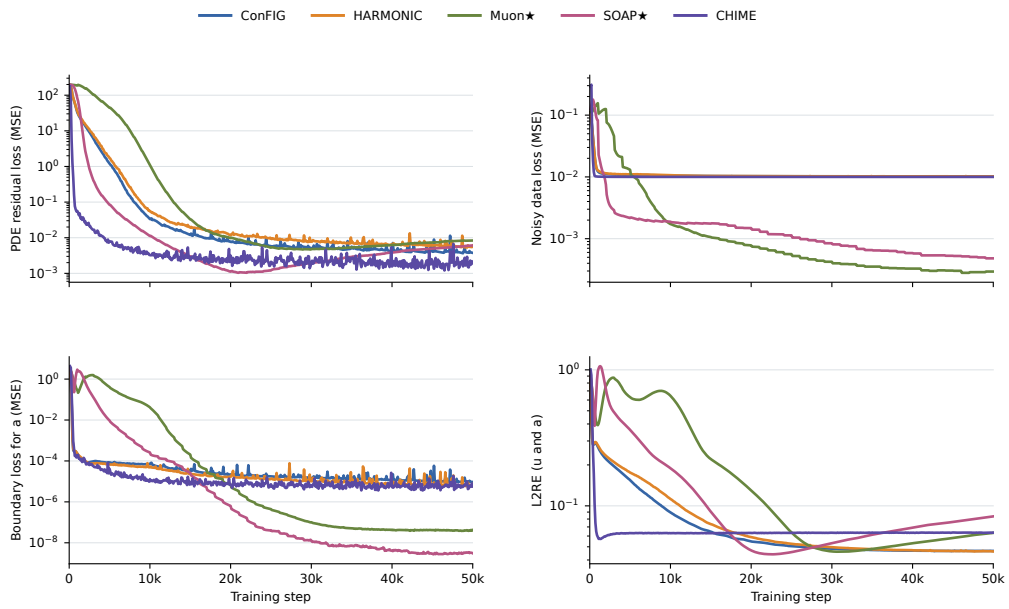


Figure 17: HInv over the complete 50,000-step horizon, using the same methods and quantities as Fig. 16. Muon* and SOAP* attain smaller empirical losses for several objectives, yet their recovery errors eventually increase. ConFIG and HARMONIC converge more stably to lower final errors, while the early advantage and later plateau of CHIME are visible together in this longer view.

and coefficient-boundary objectives retain their original full point sets. All settings use the same 50×3 network, 50,000 training steps, learning rate 10^{-4}, and five seeds.

Figure 18 shows that smaller observation mini-batches progressively delay and attenuate the late-stage recovery degradation. The final mean relative L^2 error decreases from 0.133 with 2 mini-batches to 0.0869, 0.0606, and 0.0431 with 8, 32, and 128 mini-batches, respectively. Moreover, the minimum of the mean recovery curve moves from approximately step 9,200 with 2 mini-batches to step 38,700 with 128, after which the latter remains close to its minimum through the end of training. The PDE and coefficient-boundary trajectories exhibit the same qualitative ordering in their late-stage deterioration, whereas the noisy-data loss remains on a similar 10^{-2} scale across settings.

Implication for noisy inverse training. For compatible, noise-free forward constraints, reducing the governing-equation and initial/boundary residuals is aligned with the target solution because those constraints can be satisfied simultaneously. In HInv, by contrast, the observation objective contains measurement noise, so increasingly aggressive empirical fitting need not improve recovery of the underlying state and coefficient. The full-horizon comparison and the mini-batch ablation therefore suggest that curvature-aware optimization for noisy inverse problems should be paired with a training strategy that controls observation fitting, such as appropriately stochastic observation sampling and, when necessary, complementary regularization or stopping criteria.

Table 11: Nominal sampling sizes from the HARMONIC/PINNacle protocol; see App. C.1 for the sampling-point convention.

	PInv	**HInv**
Collocation points	8192	32768
Initial points	2048	8192
Boundary points	2048	8192
Test points	8192	32768

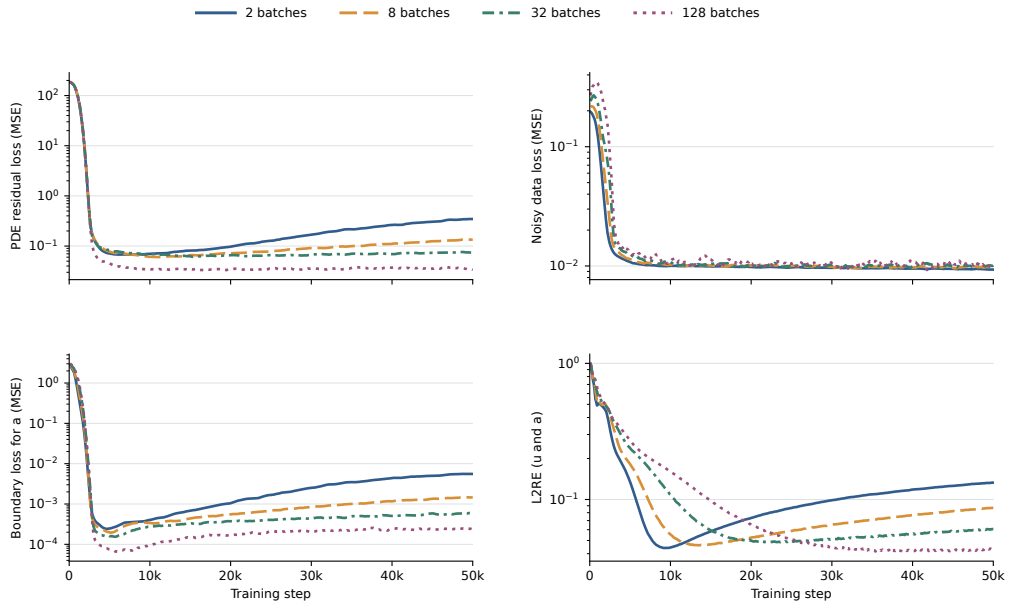


Figure 18: Effect of observation mini-batching on HInv with $\tau_{\text{data}} = 0$. Only the noisy-data objective is mini-batched, using $N_{\text{mb}} \in \{2, 8, 32, 128\}$ mini-batches per sampling cycle; the PDE residual and coefficient-boundary objectives retain their full point sets. Curves are arithmetic means over five seeds. Increasing N_{mb}, and hence reducing the observation mini-batch size, delays and suppresses the late-stage growth in recovery error.

Table 12: 95% confidence intervals of relative L^2 errors over five random seeds at learning rate 10^{-3}. In each column, the lowest-mean result(s) are shown in **bold**, and † denotes a non-bold method whose 95% confidence interval overlaps with that of at least one bolded result.

Method	**PInv**	**HInv**
MultiAdam	[0.1148, 0.2946]†	[0.1047, 0.2667]
LRA	[0.0985, 0.1663]†	[0.0383, 0.0709]†
ReLoBRaLo	[0.0000, 1.6397]†	[1.5284, 1.6714]
MGDA	[0.1778, 0.3090]	[0.3408, 4.5040]
PCGrad	[**0.0641**, **0.1177**]	[0.0937, 0.1601]
CAGrad	[0.1479, 0.2681]	[0.0000, 1.9922]†
IMTL-G	[0.0228, 0.7322]†	[0.0385, 0.0645]†
Aligned-MTL	[0.0674, 0.1156]†	[0.0237, 0.1371]†
ConFIG	[0.0947, 0.1741]†	[0.0361, 0.0689]†
HARMONIC	[0.0847, 0.1519]†	[**0.0356**, **0.0656**]
CHIME	[0.1336, 0.1562]	[0.0429, 0.0843]†

C.6 VOLTERRA PROBLEMS

1D Volterra problem (Volterra1d) The IDE is expressed as

$$\frac{du(x)}{dx} + u(x) = \lambda \int_0^x e^{\,t-x}\, u(t)\, dt. \tag{150}$$

If we define

$$v(x) = \int_0^x e^{\,t-x}\, u(t)\, dt, \tag{151}$$

then the IDE can be written as

$$\frac{du(x)}{dx} + u(x) = \lambda\, v(x). \tag{152}$$

Also, $v(x)$ should satisfy

$$\frac{dv(x)}{dx} = u(x) - \int_0^x e^{\,t-x}\, u(t)\, dt = u(x) - v(x). \tag{153}$$

The domain is $x \in \Omega = [0, 5]$. The equation has an exact solution

$$u(x) = e^{-x} \cosh(x). \tag{154}$$

The initial condition is

$$u(0) = 1, \tag{155}$$

$$v(0) = 0. \tag{156}$$

The parameter is $\lambda = 1$.

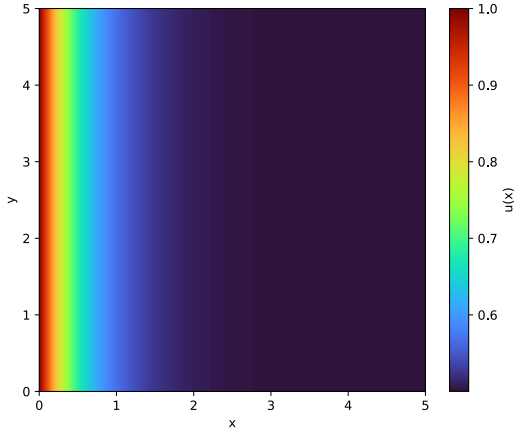


Figure 19: Reference solution of Volterra1d.

Volterra system The IDEs system is

$$\frac{d^2u(x)}{dx^2} = 1 - \frac{x^3}{3} - \frac{1}{2}\frac{dv(x)}{dx} + \lambda_1 \int_0^x \left(u^2(t) + v^2(t)\right) dt, \tag{157}$$

$$\frac{d^2v(x)}{dx^2} = -1 + x^2 - x\, u(x) + \lambda_2 \int_0^x \left(u^2(t) - v^2(t)\right) dt. \tag{158}$$

$$\tag{159}$$

As we define

$$w(x) = \int_0^x \left(u^2(t) + v^2(t)\right) dt, \tag{160}$$

$$p(x) = \int_0^x \left(u^2(t) - v^2(t)\right) dt, \tag{161}$$

we can rewrite IDEs as

$$\frac{d^2u(x)}{dx^2} = 1 - \frac{x^3}{3} - \frac{1}{2}\frac{dv(x)}{dx} + \lambda_1 w(x), \tag{162}$$

$$\frac{d^2v(x)}{dx^2} = -1 + x^2 - x\,u(x) + \lambda_2 p(x). \tag{163}$$

$$\tag{164}$$

As follows, auxiliary outputs $w(x)$ and $p(x)$ are governed by

$$\frac{dw(x)}{dx} = u^2(x) + v^2(x), \tag{165}$$

$$\frac{dp(x)}{dx} = u^2(x) - v^2(x). \tag{166}$$

The domain is $x \in \Omega = [0, 1]$. The exact solutions are

$$u(x) = x + e^x, \tag{167}$$

$$v(x) = x - e^x. \tag{168}$$

The initial conditions are

$$\frac{du(0)}{dx} = 2, \frac{dv(0)}{dx} = 0, \tag{169}$$

and

$$u(0) = 1,\ v(0) = -1,\ w(0) = 0,\ p(0) = 0. \tag{170}$$

The parameters are

$$\lambda_1 = \frac{1}{2},\ \ \lambda_2 = \frac{1}{4}. \tag{171}$$

Table 13: Sampling sizes used in our experiments.

	1D Volterra	**Volterra system**
Collocation points	8192	8192
Initial points	2048	2048
Boundary points	2048	2048
Test points	8192	8192

Table 14: 95% confidence intervals of relative L^2 errors over five random seeds at learning rate 10^{-3}. In each column, the lowest-mean result(s) are shown in **bold**, and † denotes a non-bold method whose 95% confidence interval overlaps with that of at least one bolded result.

Method	**Volterra1d**	**Volterra system**
MultiAdam	$[0.0002, 0.0012]^\dagger$	$[0.9998, 0.9998]$
LRA	$[0.0000, 0.0151]^\dagger$	$[0.0815, 0.5425]$
ReLoBRaLo	$[0.0000, 0.0091]^\dagger$	$[0.0003, 0.0015]^\dagger$
MGDA	$[0.1238, 0.4374]$	$[0.0000, 0.0303]^\dagger$
PCGrad	$[0.0000, 0.0024]^\dagger$	$[0.0000, 0.0143]^\dagger$
CAGrad	$[0.0000, 0.0044]^\dagger$	$[0.0000, 0.0173]^\dagger$
IMTL-G	$[0.0000, 2.9402]^\dagger$	$[0.5027, 0.8441]$
Aligned-MTL	$[0.0000, 0.0162]^\dagger$	$[0.0000, 0.0982]^\dagger$
ConFIG	$[0.0000, 0.0048]^\dagger$	$[0.0000, 0.2664]^\dagger$
HARMONIC	$[\mathbf{0.0002}, \mathbf{0.0004}]$	$[0.0000, 0.0017]^\dagger$
CHIME	$[0.0000, 0.0012]^\dagger$	$[\mathbf{0.0001}, \mathbf{0.0005}]$

C.7 BURGERS' EQUATIONS

1D Burgers Equation (Burgers1d) The PDE is given by

$$u_t + u\,u_x = \nu\,u_{xx}. \tag{172}$$

The domain is defined as

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

The initial and boundary conditions are

$$u(x,0) = -\sin(\pi x), \tag{173}$$
$$u(-1,t) = u(1,t) = 0. \tag{174}$$

The parameter is

$$\nu = \frac{0.01}{\pi}. \tag{175}$$

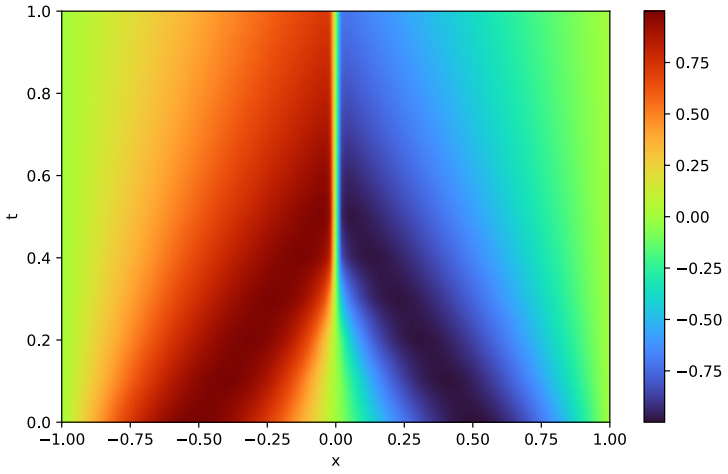


Figure 20: Reference solution of Burgers1d using PINNacle reference data.

2D Coupled Burgers Equation (Burgers2d) The governing equations are given by

$$u_t + uu_x + vu_y = \nu(u_{xx} + u_{yy}), \tag{176}$$
$$v_t + uv_x + vv_y = \nu(v_{xx} + v_{yy}). \tag{177}$$

The domain is defined as

$$(x,y,t) \in \Omega \times T = [0,L]^2 \times [0,1]. \tag{178}$$

The initial conditions are given by

$$\boldsymbol{w}(x,y) = \sum_{i=-L}^{L} \sum_{j=-L}^{L} \Big[\boldsymbol{a}_{ij} \sin\big(2\pi(ix+jy)\big) + \boldsymbol{b}_{ij} \cos\big(2\pi(ix+jy)\big)\Big], \tag{179}$$

$$[u(x,y,0),\ v(x,y,0)] = 2\,\boldsymbol{w}(x,y) + \boldsymbol{c}. \tag{180}$$

where $\boldsymbol{a}_{ij}, \boldsymbol{b}_{ij}, \boldsymbol{c} \sim \mathcal{N}(0,1) \in R^2$. The boundary conditions are given by

$$u(0,y,t) = u(L,y,t), \tag{181}$$
$$u(x,0,t) = u(x,L,t), \tag{182}$$

$$v(0,y,t) = v(L,y,t), \tag{183}$$
$$v(x,0,t) = v(x,L,t). \tag{184}$$

The parameters are

$$L = 4, \quad \nu = 0.001. \tag{185}$$

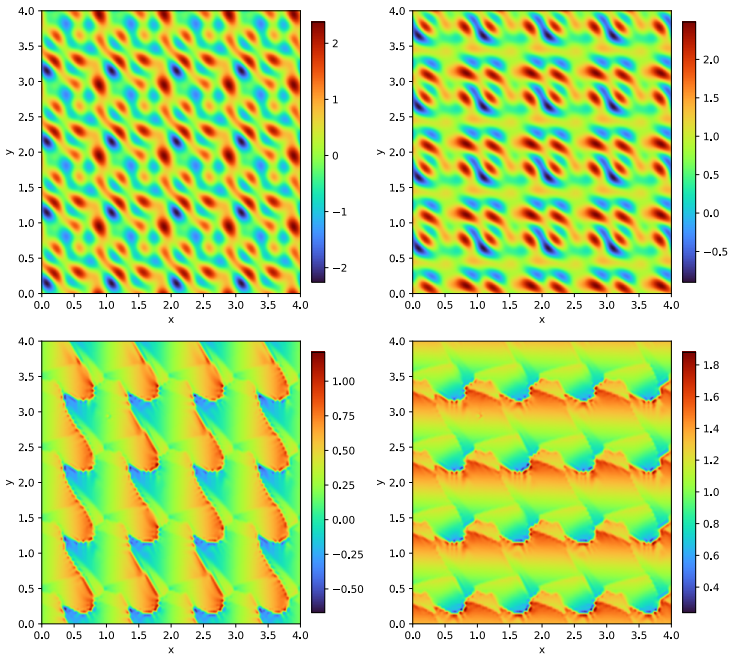


Figure 21: Reference solution of Burgers2d at times $t = 0, 1$, where the left panels show u and the right panels show v from the PINNacle reference data.

Table 15: Nominal sampling sizes from the HARMONIC/PINNacle protocol; see App. C.1 for the sampling-point convention.

	Burgers1d	**Burgers2d**
Collocation points	8192	32768
Initial points	2048	8192
Boundary points	2048	8192
Test points	8192	32768

Table 16: 95% confidence intervals of relative L^2 errors over five random seeds at learning rate 10^{-3}. In each column, the lowest-mean result(s) are shown in **bold**, and † denotes a non-bold method whose 95% confidence interval overlaps with that of at least one bolded result.

Method	**Burgers1d**	**Burgers2d**
MultiAdam	[0.6167, 1.0885]	[0.5491, 0.5491]
LRA	[0.0095, 0.0759]†	[0.5132, 0.5208]
ReLoBRaLo	[0.0138, 0.0180]	[0.5226, 0.5338]
MGDA	[0.0000, 0.1420]†	[0.5164, 0.5298]
PCGrad	[0.0093, 0.0343]†	[0.5217, 0.5239]
CAGrad	[0.0160, 0.0400]	[0.5223, 0.5265]
IMTL-G	[0.0699, 0.5743]	[0.0000, 3.0548]†
Aligned-MTL	[0.0160, 0.0400]	[0.5226, 0.5270]
ConFIG	[0.0130, 0.0398]†	[0.5029, 0.5461]
HARMONIC	[0.0135, 0.0193]†	[0.5283, 0.5595]
CHIME	[**0.0133**, **0.0137**]	[**0.4459**, **0.4631**]

C.8 NAVIER–STOKES EQUATIONS

2D NS lid-driven flow (NS2d-C) The PDE is given by

$$uu_x + vu_y + p_x = \frac{1}{Re}(u_{xx} + u_{yy}), \tag{186}$$

$$uv_x + vv_y + p_y = \frac{1}{Re}(v_{xx} + v_{yy}), \tag{187}$$

$$u_x + v_y = 0. \tag{188}$$

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

$$u(x, y) = 4x(1 - x), \qquad y \in \Gamma_1, \tag{189}$$

$$v(x, y) = 0, \qquad y \in \Gamma_1, \tag{190}$$

$$u(x, y) = 0, v(x, y) = 0, \qquad y \in \Gamma_2, \tag{191}$$

$$p(0, 0) = 0. \tag{192}$$

The Reynolds number $Re = 100$.

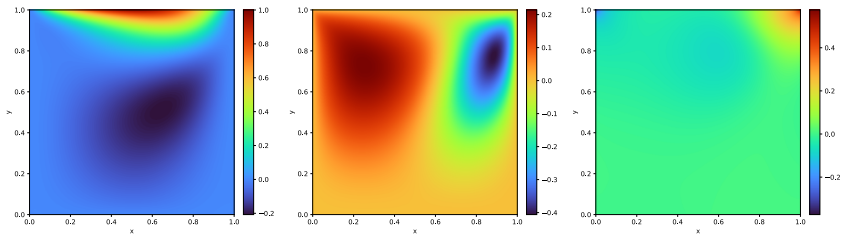


Figure 22: Reference solution of NS2d-C, where the left panel shows u, the middle panel shows v, and the right panel shows p from the PINNacle reference data.

2D Back Step Flow (NS-CG) The PDE is given by

$$uu_x + vu_y + p_x = \frac{1}{Re}(u_{xx} + u_{yy}), \tag{193}$$

$$uv_x + vv_y + p_y = \frac{1}{Re}(v_{xx} + v_{yy}), \tag{194}$$

$$u_x + v_y = 0. \tag{195}$$

The domain is $\Omega = ([0, 4] \times [0, 2]) \setminus ([0, 2] \times [1, 2])$, excluding the top-left quarter, with no additional obstacles in the reference configuration. The inlet is $\Gamma_{\text{in}} = \{0\} \times [0, 1]$ and the outlet is $\Gamma_{\text{out}} = \{4\} \times [0, 2]$.

The boundary conditions are

$$u(0, y) = 4y(1 - y), \tag{196}$$

$$v(0, y) = 0, \tag{197}$$

$$p(4, y) = 0, \tag{198}$$

$$u(x, y) = 0,\ v(x, y) = 0, \qquad (x, y) \in \partial\Omega \setminus (\Gamma_{\text{in}} \cup \Gamma_{\text{out}}). \tag{199}$$

The Reynolds number of $Re = 100$.

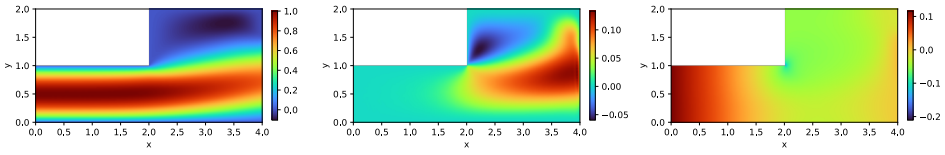


Figure 23: Reference solution of NS-CG, where the left panel shows u, the middle panel shows v, and the right panel shows p from the PINNacle reference data.

Table 17: Nominal sampling sizes from the HARMONIC/PINNacle protocol; see App. C.1 for the sampling-point convention.

	NS2d-C	**NS-CG**
Collocation points	8192	8192
Initial points	2048	2048
Boundary points	2048	2048
Test points	8192	8192

Table 18: 95% confidence intervals of relative L^2 errors over five random seeds at learning rate 10^{-3}. In each column, the lowest-mean result(s) are shown in **bold**, and † denotes a non-bold method whose 95% confidence interval overlaps with that of at least one bolded result.

Method	**NS2d-C**	**NS-CG**
MultiAdam	[0.9105, 1.7223]	[0.7852, 0.8498]
LRA	[0.3541, 0.6513]	[0.2600, 0.2920]
ReLoBRaLo	[0.0500, 0.1408]	[0.1097, 0.1327]
MGDA	[0.0776, 0.7856]	[0.1248, 0.2750]
PCGrad	[0.0242, 0.0944]	[0.1017, 0.1441]
CAGrad	[0.0199, 0.2545]†	[0.1844, 0.2210]
IMTL-G	[0.4382, 0.9538]	[0.2241, 0.4685]
Aligned-MTL	[0.0098, 0.3862]†	[0.1411, 0.2463]
ConFIG	[0.3453, 1.5095]	[0.2058, 0.3720]
HARMONIC	[0.0000, 0.1560]†	[0.1368, 0.1726]
CHIME	[**0.0198**, **0.0216**]	[**0.0870**, **0.0934**]

D Computational Costs and Wall-Clock Comparisons

D.1 Per-step cost

Table 19 reports runtime per 100 training steps on the five main benchmarks. CHIME requires approximately 1.73–2.18 times the per-step cost of SOAP* and is less expensive than KFAC* on every problem.

Table 19: Time in seconds per 100 training steps, reported as mean (standard deviation). Lower is better; the largest cost in each column is shown in **bold**.

Method	**Wave1d-C**	**Poisson2d-C**	**HNd**	**Volterra1d**	**HInv**
LRA	1.80 (0.06)	1.43 (0.05)	4.47 (0.02)	1.08 (0.05)	2.41 (0.03)
ReLoBRaLo	0.69 (0.03)	0.63 (0.03)	1.57 (0.04)	0.50 (0.05)	0.98 (0.04)
PCGrad	1.64 (0.08)	1.28 (0.07)	3.40 (0.02)	1.18 (0.04)	1.73 (0.04)
CAGrad	2.69 (0.03)	2.34 (0.08)	4.62 (0.09)	2.33 (0.14)	2.97 (0.11)
DCGD	0.82 (0.02)	0.77 (0.02)	2.34 (0.01)	0.58 (0.00)	1.19 (0.04)
ConFIG	1.63 (0.09)	1.29 (0.05)	3.46 (0.04)	1.17 (0.03)	1.70 (0.03)
HARMONIC	1.55 (0.03)	1.26 (0.06)	3.43 (0.03)	1.08 (0.02)	1.88 (0.03)
KFAC*	**4.50 (0.15)**	**3.70 (0.11)**	**11.54 (0.20)**	**4.17 (0.21)**	**7.69 (0.12)**
Muon*	0.80 (0.07)	0.75 (0.04)	1.49 (0.04)	0.60 (0.04)	1.01 (0.10)
SOAP*	0.94 (0.06)	0.89 (0.04)	1.60 (0.06)	0.73 (0.04)	1.21 (0.06)
CHIME	2.05 (0.15)	1.63 (0.05)	3.37 (0.05)	1.55 (0.07)	2.09 (0.22)

D.2 Accuracy under equal wall-clock budgets

Table 20 compares the methods under a common time budget for each problem. Among these methods, CHIME has the lowest displayed mean error on Wave1d-C, Poisson2d-C, HNd, and HInv, and ties with SOAP* on Volterra1d.

Table 20: Relative L^2 errors under a common wall-clock budget for each problem. Entries are means (standard deviations) over five random seeds; lower is better. Best displayed means are in **bold**, including ties.

Benchmark	Muon*	SOAP*	CHIME
Wave1d-C	0.0154 (0.0021)	0.0042 (0.0010)	**0.0035 (0.0010)**
Poisson2d-C	0.6633 (0.0209)	0.6633 (0.0201)	**0.0029 (0.0006)**
HNd	0.0003 (0.0000)	0.0002 (0.0001)	**0.0001 (0.0000)**
Volterra1d	0.0008 (0.0003)	**0.0002 (0.0002)**	**0.0002 (0.0000)**
HInv	0.0996 (0.0206)	0.1127 (0.0173)	**0.0636 (0.0167)**

E LEARNING-RATE COMPARISONS

E.1 COMPARISON ACROSS LEARNING RATES

Tables 21–25 compare learning rates 10^{-3} and 10^{-4} under the protocol in App. C.1. For CHIME, 10^{-4} gives lower mean error on all four forward problems, while 10^{-3} is slightly better on HInv. Red bold marks the lowest reported mean across both rates; † marks overlap with a highlighted 95% confidence interval, regardless of rate.

Table 21: Wave1d-C: 95% confidence intervals of relative L^2 error over five seeds.

Method	10^{-3}	10^{-4}
MultiAdam	[1.0758, 1.2322]	[0.9711, 1.2205]
LRA	[0.1935, 0.3281]	[0.3041, 0.3729]
ReLoBRaLo	[0.3511, 0.4693]	[0.3431, 0.3653]
MGDA	[0.2940, 0.6314]	[0.3827, 0.4163]
PCGrad	[0.1386, 0.3126]	[0.2878, 0.3370]
CAGrad	[0.1098, 0.1274]	[0.2782, 0.3844]
IMTL-G	[0.1757, 1.3571]	[0.0000, 0.6911]†
Aligned-MTL	[0.1639, 0.5963]	[0.1664, 0.2940]
DCGD	[0.3638, 0.4406]	[0.4079, 0.4422]
ConFIG	[0.0322, 0.1014]	[0.1165, 0.2215]
HARMONIC	[0.0291, 0.1019]	[0.1933, 0.3065]
KFAC⋆	[0.5579, 0.6192]	[1.1123, 1.1659]
Muon⋆	[0.0215, 0.0243]	[0.4732, 0.5894]
SOAP⋆	[0.0045, 0.0083]†	[0.3254, 0.3446]
CHIME	[0.0042, 0.0064]†	[**0.0021**, **0.0045**]

Table 22: Poisson2d-C: 95% confidence intervals of relative L^2 error over five seeds.

Method	10^{-3}	10^{-4}
MultiAdam	[0.5715, 0.7751]	[0.9922, 1.4830]
LRA	[0.3732, 0.7692]	[0.6645, 0.7171]
ReLoBRaLo	[0.6328, 0.6876]	[0.6371, 0.6887]
MGDA	[0.6480, 0.7074]	[0.6602, 0.7078]
PCGrad	[0.6682, 0.7154]	[0.6718, 1.1854]
CAGrad	[0.5878, 0.9734]	[0.6724, 0.7214]
IMTL-G	[0.4190, 0.5888]	[0.1723, 1.7763]
Aligned-MTL	[0.0000, 0.5781]†	[0.6735, 0.7539]
DCGD	[0.0000, 0.8942]†	[0.0145, 0.8301]
ConFIG	[0.1620, 1.2288]	[0.6516, 0.7196]
HARMONIC	[0.0000, 0.0436]†	[0.0000, 0.9391]†
KFAC⋆	[0.6640, 0.7157]	[0.6855, 0.7276]
Muon⋆	[0.6404, 0.6928]	[0.6379, 0.6875]
SOAP⋆	[0.6381, 0.6803]	[0.6435, 0.6887]
CHIME	[0.0029, 0.0053]†	[**0.0022**, **0.0036**]

Table 23: HNd: 95% confidence intervals of relative L^2 error over five seeds.

Method	10^{-3}	10^{-4}
MultiAdam	[0.0000, 0.0742]†	[0.0005, 0.0075]
LRA	[0.0006, 0.0034]	[0.0000, 0.0107]†
ReLoBRaLo	[0.0000, 0.0117]†	[0.0004, 0.0004]
MGDA	[0.0000, 0.0157]†	[0.0008, 0.0010]
PCGrad	[0.0004, 0.0052]	[0.0003, 0.0007]
CAGrad	[0.0004, 0.0006]	[0.0006, 0.0006]
IMTL-G	[0.0004, 0.0008]	[0.0005, 0.0007]
Aligned-MTL	[0.0004, 0.0008]	[0.0000, 0.0033]†
DCGD	[0.0004, 0.0006]	[0.0004, 0.0005]
ConFIG	[0.0005, 0.0007]	[0.0000, 0.0018]†
HARMONIC	[0.0005, 0.0005]	[0.0000, 0.0132]†
KFAC⋆	[0.0005, 0.0007]	[0.0045, 0.0123]
Muon⋆	[0.0011, 0.0041]	[0.0002, 0.0002]†
SOAP⋆	[0.0000, 0.0007]†	**[0.0001, 0.0001]**
CHIME	[0.0002, 0.0014]†	**[0.0000, 0.0002]**

Table 24: Volterra1d: 95% confidence intervals of relative L^2 error over five seeds.

Method	10^{-3}	10^{-4}
MultiAdam	[0.0002, 0.0012]†	[0.0002, 0.0006]†
LRA	[0.0000, 0.0151]†	[0.0032, 0.0128]
ReLoBRaLo	[0.0000, 0.0091]†	[0.0002, 0.0002]†
MGDA	[0.1238, 0.4374]	[0.0000, 0.0035]†
PCGrad	[0.0000, 0.0024]†	[0.0000, 0.0009]†
CAGrad	[0.0000, 0.0044]†	[0.0000, 0.0051]†
IMTL-G	[0.0000, 2.9402]†	[0.0000, 0.1963]†
Aligned-MTL	[0.0000, 0.0162]†	[0.0000, 0.0092]†
DCGD	[0.0000, 0.0027]†	[0.0000, 0.0017]†
ConFIG	[0.0000, 0.0048]†	[0.0000, 0.0128]†
HARMONIC	[0.0002, 0.0004]†	[0.0003, 0.0005]
KFAC⋆	[0.0002, 0.0002]†	[0.0036, 0.0321]
Muon⋆	[0.0020, 0.0058]	[0.0002, 0.0010]
SOAP⋆	[0.0001, 0.0003]†	[0.0000, 0.0004]†
CHIME	[0.0000, 0.0012]†	**[0.0000, 0.0002]**

Table 25: HInv: 95% confidence intervals of relative L^2 error over five seeds.

Method	10^{-3}	10^{-4}
MultiAdam	[0.1047, 0.2667]	[0.1142, 0.1750]
LRA	[0.0383, 0.0709]†	[0.0550, 0.0702]†
ReLoBRaLo	[1.5284, 1.6714]	[1.6959, 1.8069]
MGDA	[0.3408, 4.5040]	[0.7959, 4.9871]
PCGrad	[0.0937, 0.1601]	[0.0471, 0.0849]†
CAGrad	[0.0000, 1.9922]†	[1.6544, 2.8434]
IMTL-G	[0.0385, 0.0645]†	[0.0405, 0.0591]†
Aligned-MTL	[0.0237, 0.1371]†	[0.1241, 0.2563]
DCGD	[0.0364, 0.0666]†	[0.0327, 0.0606]†
ConFIG	[0.0361, 0.0689]†	[0.0382, 0.0550]†
HARMONIC	[0.0356, 0.0656]†	**[0.0339, 0.0583]**
KFAC⋆	[0.3312, 0.3834]	[1.2157, 1.7623]
Muon⋆	[0.0629, 0.1187]	[0.0483, 0.0783]†
SOAP⋆	[0.1153, 0.1597]	[0.0658, 0.1016]
CHIME	[0.0429, 0.0843]†	[0.0425, 0.0899]†

F Multi-Task Learning Experiments

F.1 Benchmark protocols

We follow the NYU-v2 and CelebA experimental protocols of Qin et al. (2025), including the data processing and training schedules in their released implementation. Both benchmarks use three random seeds. CHIME uses the update in Algorithm 1, with $\tau_i = 0$ for every task on both datasets. The comparison methods and the definition of $\Delta_m\%$ are given in Sec. 5.

F.1.1 NYU-v2: three dense-prediction tasks

Each RGB image is paired with a 13-class semantic segmentation map, a depth map, and a three-component surface-normal map. We use the preprocessed training and evaluation splits released with MTAN and adopted by NTKMTL, with images at 288×384 resolution (Liu et al., 2019; Qin et al., 2025). The model is a SegNet-based Multi-Task Attention Network (MTAN): a shared encoder–decoder has five feature levels with channel widths $(64, 128, 256, 512, 512)$, and task-specific attention modules and prediction branches produce the three outputs. Segmentation is represented by per-pixel class probabilities, depth by one scalar per pixel, and surface normals by unit-normalized three-vectors.

Task objectives. Let V_s, V_d, and V_n denote the valid pixels for segmentation, depth, and surface normals, respectively. With segmentation probabilities $p_{v,c}$, depth predictions $\widehat{z}_v$, and predicted normals $\widehat{\boldsymbol{n}}_v$, the three objectives are

$$L_s = -\frac{1}{|V_s|} \sum_{v \in V_s} \log p_{v,y_v}, \tag{200}$$

$$L_d = \frac{1}{|V_d|} \sum_{v \in V_d} |\widehat{z}_v - z_v|, \tag{201}$$

$$L_n = 1 - \frac{1}{|V_n|} \sum_{v \in V_n} \widehat{\boldsymbol{n}}_v^\top \boldsymbol{n}_v. \tag{202}$$

The segmentation loss ignores labels equal to -1. For depth and normals, the reference implementation excludes a pixel when the sum of its target components is zero; the corresponding masks determine V_d and V_n.

Data augmentation and training. During training, a scale is sampled uniformly from $\{1.0, 1.2, 1.5\}$; a corresponding random crop is resized to the original input dimensions. Images and normal maps use bilinear interpolation, while segmentation and depth maps use nearest-neighbor interpolation. Depth values are divided by the sampled scale. A horizontal flip is applied with probability $1/2$ to all maps, with the horizontal normal component negated to preserve its orientation. Evaluation uses the unaugmented split. Training lasts 200 epochs with batch size 2. The learning rate is 10^{-4} for the first 100 epochs and 5×10^{-5} for the remaining 100 epochs.

F.1.2 CelebA: forty binary attributes

Each aligned RGB face image has 40 binary attribute labels, with one classification task per attribute. We use the official training, validation, and test partition and map the annotation values $\{-1, +1\}$ to $\{0, 1\}$. Images are resized to 64×64 and converted to tensors.

The nine-layer backbone described by Qin et al. (2025) contains seven 3×3 convolutional layers with channel widths $(64, 128, 128, 256, 256, 512, 512)$, followed by two fully connected layers of width 512. Each backbone layer is followed by batch normalization and ReLU. Convolutions use stride one and padding one; 2×2 max pooling follows convolutional layers 1, 3, and 5, and global average pooling follows layer 7. The resulting 512-dimensional shared representation feeds 40 separate linear heads, each followed by a sigmoid.

For a minibatch $\mathcal{B}$, attribute i uses binary cross-entropy

$$L_i = -\frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} \left[y_{b,i} \log p_{b,i} + (1 - y_{b,i}) \log(1 - p_{b,i}) \right], \qquad i = 1, \ldots, 40. \tag{203}$$

Training lasts 15 epochs with batch size 256 and initial learning rate 3×10^{-4}. A cosine schedule decreases the learning rate to zero over the 15 epochs. For each seed, the reported test scores correspond to the epoch with the best mean validation F1 across the 40 attributes. Attribute predictions are positive when their sigmoid probability exceeds 0.5; true positives, false positives, and false negatives are accumulated over the complete evaluation split before computing each attribute's F1.

F.2 METRICS AND AGGREGATION

NYU-v2 uses nine metrics across segmentation, depth, and surface-normal prediction. The three angular-threshold metrics count valid pixels with normal errors below $11.25°$, $22.5°$, and $30°$. CelebA uses one F1 score for each of its 40 binary attributes. Equation (25) applies with $S = 9$ for NYU-v2 and $S = 40$ for CelebA; all CelebA metrics are higher-is-better. The average weights metrics equally, and negative $\Delta_m\%$ indicates improvement over STL. Baseline $\Delta_m\%$ values are retained from the source comparison; CHIME's values follow its three-seed evaluations.

MR averages metric-wise ranks over the methods in each aggregate table, excluding STL and assigning average ranks to ties. Central MR values are recomputed from the displayed means, whereas the retained MR standard deviations describe seed-wise rank variability.

F.3 NYU-V2 SEED VARIABILITY

Table 26 provides the standard deviations omitted from Table 3. The aggregate advantage does not imply uniform per-metric dominance: the strongest individual results are distributed across methods.

Table 26: CHIME results on NYU-v2 corresponding to Table 3. Parentheses give standard deviations over three seeds; the MR convention is specified in App. F.2.

METHOD	mIoU ↑	Pix Acc ↑	Abs Err ↓	Rel Err ↓	Mean ↓	Median ↓	11.25 ↑	22.5 ↑	30 ↑	**MR** ↓	$\boldsymbol{\Delta_m\%}$ ↓
CHIME	40.12 (0.40)	65.71 (0.19)	0.5240 (0.0005)	0.2088 (0.0008)	24.87 (0.18)	18.70 (0.13)	31.76 (0.21)	58.19 (0.29)	69.76 (0.27)	3.78 (0.21)	-7.37 (0.27)

F.4 CELEBA: FORTY-TASK CLASSIFICATION

Table 27 reports the aggregate comparison, with the underlying 40 attribute-wise scores in Tables 28–29. CHIME attains the lowest MR and $\Delta_m\%$ among the displayed methods.

Table 27: Results on CelebA (40 tasks). MR is recomputed from the attribute-wise means; baseline $\Delta_m\%$ follows Qin et al. (2025). Parentheses give CHIME's seed variability as defined in App. F.2. The best aggregate scores are highlighted in **gray**.

METHOD	**MR** ↓	$\boldsymbol{\Delta_m\%}$ ↓
LS	7.05	4.15
SI	8.85	7.20
RLW	6.20	1.46
DWA	7.90	3.20
UW	6.53	3.23
MGDA	12.04	14.85
PCGrad	7.71	3.17
CAGrad	7.20	2.48
IMTL-G	5.51	0.84
Nash-MTL	5.74	2.84
FAMO	5.59	1.21
FairGrad	6.23	0.37
CHIME	**4.46 (0.41)**	**-0.24 (0.14)**

Table 28: Per-attribute F1 scores on CelebA (attributes 1–20). Entries are three-seed means, with CHIME's standard deviations in parentheses.

ATTRIBUTE	STL	LS	SI	RLW	DWA	UW	MGDA	PCGrad	CAGrad	IMTL-G	Nash-MTL	FAMO	FairGrad	**CHIME**
5_o_Clock_Shadow	0.6737	0.6745	0.6055	0.6605	0.6010	0.6808	0.5803	0.6336	0.6505	0.6624	0.6729	0.6539	0.6735	0.6696 (0.0126)
Arched_Eyebrows	0.6812	0.6292	0.6433	0.6996	0.6948	0.6828	0.5419	0.6777	0.6885	0.6658	0.6897	0.6777	0.6715	0.6825 (0.0103)
Attractive	0.8152	0.8053	0.8109	0.7842	0.8059	0.7930	0.7420	0.8025	0.8079	0.8188	0.8167	0.8159	0.8090	0.8116 (0.0028)
Bags_Under_Eyes	0.5760	0.4771	0.5513	0.6147	0.5265	0.5987	0.5140	0.5892	0.5596	0.5610	0.5492	0.5395	0.5913	0.5839 (0.0074)
Bald	0.7206	0.7077	0.7219	0.7104	0.6445	0.7199	0.6883	0.7058	0.7345	0.6829	0.7086	0.7208	0.7161	0.7109 (0.0121)
Bangs	0.8555	0.8535	0.8385	0.8527	0.8532	0.8542	0.8096	0.8525	0.8495	0.8532	0.8611	0.8411	0.8487	0.8516 (0.0011)
Big_Lips	0.3820	0.3797	0.3059	0.5259	0.4355	0.4168	0.2861	0.3908	0.3770	0.4211	0.3560	0.4497	0.5294	0.4536 (0.0124)
Big_Nose	0.5823	0.5878	0.5825	0.5907	0.5956	0.6028	0.5387	0.5912	0.5607	0.5745	0.5668	0.5829	0.6024	0.5865 (0.0042)
Black_Hair	0.7876	0.7926	0.7425	0.7811	0.7665	0.8021	0.7251	0.7778	0.7801	0.7847	0.7949	0.7489	0.7925	0.7881 (0.0104)
Blond_Hair	0.8321	0.8415	0.8393	0.8285	0.7937	0.8372	0.7349	0.8330	0.8129	0.8366	0.8410	0.8101	0.8233	0.8367 (0.0069)
Blurry	0.5030	0.3919	0.3616	0.4359	0.4831	0.4224	0.2682	0.4385	0.4256	0.5000	0.3552	0.4400	0.4916	0.4506 (0.0482)
Brown_Hair	0.6869	0.6039	0.6584	0.6496	0.5664	0.6452	0.5727	0.6337	0.6251	0.6030	0.6417	0.6205	0.6860	0.6831 (0.0057)
Bushy_Eyebrows	0.6781	0.6660	0.6498	0.6423	0.6612	0.6776	0.5449	0.6612	0.5869	0.6574	0.6249	0.6626	0.6678	0.6648 (0.0163)
Chubby	0.5240	0.5188	0.4996	0.4855	0.5018	0.5322	0.4834	0.5340	0.4798	0.5173	0.5266	0.5131	0.5008	0.5022 (0.0069)
Double_Chin	0.5162	0.5323	0.4878	0.4968	0.5290	0.5233	0.4552	0.5274	0.5060	0.5320	0.5279	0.5404	0.5353	0.5147 (0.0085)
Eyeglasses	0.9569	0.9277	0.9440	0.9370	0.9305	0.9125	0.9318	0.9153	0.9338	0.9221	0.9539	0.9528	0.9544	0.9389 (0.0055)
Goatee	0.6969	0.6418	0.6760	0.6926	0.6708	0.6663	0.6828	0.6776	0.6986	0.7018	0.6984	0.6919	0.6992	0.6810 (0.0053)
Gray_Hair	0.6798	0.6748	0.6295	0.6927	0.6740	0.6898	0.5835	0.6781	0.6621	0.6873	0.6970	0.6888	0.6246	0.7009 (0.0029)
Heavy_Makeup	0.8808	0.8802	0.8798	0.8632	0.8756	0.8672	0.8555	0.8788	0.8779	0.8826	0.8852	0.8848	0.8751	0.8871 (0.0017)
High_Cheekbones	0.8582	0.8551	0.8507	0.8637	0.8570	0.8597	0.8445	0.8537	0.8604	0.8647	0.8580	0.8587	0.8526	0.8622 (0.0014)

Table 29: Per-attribute F1 scores on CelebA (attributes 21–40), under the same reporting convention as Table 28.

ATTRIBUTE	STL	LS	SI	RLW	DWA	UW	MGDA	PCGrad	CAGrad	IMTL-G	Nash-MTL	FAMO	FairGrad	**CHIME**
Male	0.9703	0.9648	0.9691	0.9700	0.9698	0.9695	0.9488	0.9704	0.9669	0.9730	0.9715	0.9710	0.9680	0.9700 (0.0006)
Mouth_Slightly_Open	0.9327	0.9259	0.9230	0.9259	0.9169	0.9228	0.8421	0.9243	0.9241	0.9208	0.9249	0.9217	0.9130	0.9247 (0.0016)
Mustache	0.5058	0.4188	0.5256	0.5326	0.4724	0.3612	0.4868	0.4798	0.5129	0.5296	0.4897	0.4892	0.5394	0.5295 (0.0049)
Narrow_Eyes	0.4031	0.3813	0.3699	0.4391	0.3814	0.3773	0.3040	0.4527	0.4042	0.3693	0.3951	0.4166	0.4410	0.3971 (0.0138)
No_Beard	0.9704	0.9656	0.9696	0.9685	0.9718	0.9711	0.9674	0.9703	0.9712	0.9708	0.9717	0.9720	0.9698	0.9705 (0.0011)
Oval_Face	0.4864	0.4707	0.3547	0.2700	0.3882	0.3691	0.2819	0.4100	0.3688	0.4668	0.3933	0.4596	0.3088	0.5012 (0.0173)
Pale_Skin	0.6079	0.5538	0.3971	0.5368	0.5350	0.5588	0.2822	0.5169	0.5980	0.6053	0.5901	0.6097	0.5560	0.5764 (0.0263)
Pointy_Nose	0.5261	0.4894	0.4938	0.5184	0.5091	0.5303	0.4196	0.4492	0.5507	0.4668	0.4863	0.5062	0.5475	0.5487 (0.0217)
Receding_Hairline	0.5691	0.5574	0.5138	0.5857	0.5147	0.5367	0.4552	0.5078	0.5219	0.5529	0.5742	0.5827	0.5300	0.5617 (0.0030)
Rosy_Cheeks	0.5982	0.6345	0.6113	0.5730	0.6030	0.6233	0.4877	0.5627	0.6256	0.5941	0.5127	0.6320	0.6051	0.6167 (0.0037)
Sideburns	0.6858	0.6776	0.6336	0.6897	0.6761	0.6696	0.6736	0.6871	0.6998	0.7160	0.7214	0.6952	0.7211	0.6981 (0.0098)
Smiling	0.9241	0.9212	0.9109	0.9208	0.9188	0.9189	0.8942	0.9182	0.9141	0.9168	0.9181	0.9163	0.9100	0.9193 (0.0012)
Straight_Hair	0.5425	0.5425	0.4477	0.5066	0.5241	0.5297	0.2324	0.4622	0.5384	0.5166	0.4987	0.5267	0.4879	0.5283 (0.0148)
Wavy_Hair	0.7406	0.5473	0.6983	0.7441	0.6982	0.6179	0.6924	0.6747	0.6819	0.6914	0.6650	0.6864	0.6448	0.7268 (0.0035)
Wearing_Earrings	0.7102	0.6346	0.6355	0.7003	0.6053	0.6716	0.5239	0.6588	0.6732	0.6820	0.6777	0.6724	0.7013	0.6938 (0.0080)
Wearing_Hat	0.8737	0.8612	0.8249	0.8481	0.8171	0.8305	0.8384	0.8558	0.8170	0.8668	0.8783	0.8628	0.8597	0.8663 (0.0032)
Wearing_Lipstick	0.9306	0.9385	0.9403	0.9206	0.9331	0.9260	0.9193	0.9331	0.9368	0.9389	0.9378	0.9405	0.9294	0.9381 (0.0005)
Wearing_Necklace	0.3370	0.3016	0.1721	0.3350	0.4064	0.2920	0.1732	0.3352	0.3864	0.4086	0.3716	0.3204	0.4035	0.3676 (0.0211)
Wearing_Necktie	0.7648	0.7218	0.7397	0.7433	0.7340	0.7399	0.7218	0.7283	0.7362	0.7517	0.7056	0.7311	0.7262	0.7507 (0.0049)
Young	0.9191	0.9201	0.9072	0.9151	0.9156	0.9182	0.8936	0.9171	0.9104	0.9199	0.9156	0.9206	0.9127	0.9195 (0.0013)

G 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.