

RESEARCH ARTICLE

Disentangling Sampling from Loss Weighting in Physics-Informed Neural Networks

Rasım Çekik^{a,*}^a Computer Engineering Department, Sırnak University, 73000 Sırnak, Türkiye* Corresponding author: rasimcekik@sirnak.edu.tr · ORCID: 0000-0002-7820-413X**ARTICLE INFO****Article history**

Received 24 June 2026

Revised 24 June 2026

Accepted 25 June 2026

Available online 26 June 2026

Keywords

Physics-informed neural networks

Collocation sampling

Loss weighting

Stochastic gradient noise

Experimental attribution

Reproducibility

Abstract

Interventions on the collocation set enlarging it, or drawing it from a residual-based adaptive distribution are among the most widely adopted means of improving physics-informed neural networks, and the resulting gains are routinely attributed to *where* the points are placed. We observe that such interventions cannot change the sampling geometry alone: because the residual loss is an empirical mean, the magnitude of its stochastic gradient carries an inflation term $\text{tr}(\Sigma_r)/N_r$, so any change to the number or the distribution of collocation points silently rescales the residual term relative to the boundary and initial-condition terms. Reported gains therefore confound two mechanisms that the literature treats as distinct: sampling and loss weighting. We measure the confound directly and find it to be real but moderate, shifting the interior-to-constraint gradient balance by 20 to 40 per cent when the interior batch is enlarged sixteenfold. We then separate the two mechanisms with a matched-control design in which a control arm reproduces, to machine precision, the effective-balance trajectory of an intervention while leaving the collocation set untouched. On the one benchmark where enlarging the collocation set clearly helps viscous Burgers, where it reduces the error fivefold the control recovers none of the gain, giving an attribution ratio of $A = -0.10$ with a bootstrap interval of $[-0.90, 0.22]$ that excludes the reweighting explanation. Matching the balance through the weights alone is in fact harmful: across problems the control arms are 1.1 to 1.7 times worse than the baseline they start from. Sampling gains in this setting are therefore geometric, and an alternative explanation that had never been excluded can now be set aside. We also report that seed variance in this regime is large enough to require tens of seeds per arm, and we release a verified implementation with the pre-registered analysis plan.

1. Introduction

A physics-informed neural network (PINN) is trained by minimising a weighted sum of a partial differential equation (PDE) residual, evaluated at a set of collocation points, and a set of constraint terms enforcing boundary and initial data [1, 2]. Two families of modification dominate the practical literature. The first alters the collocation set: more points, resampling schedules, or non-uniform distributions concentrated where the residual is large [3, 4]. The second alters the weights that multiply the loss terms, either by hand or through adaptive schemes driven by gradient statistics or by the neural tangent kernel [5–8]. Both families report substantial

accuracy improvements, and both are usually presented as addressing a distinct deficiency: the first, that the residual is sampled in the wrong places; the second, that the loss terms are balanced wrongly.

This paper starts from an observation that makes the separation less clean than it appears. The residual loss is an empirical mean over N_r points, so the gradient the optimiser receives is a stochastic estimate whose squared magnitude satisfies

$$\mathbb{E}\|\hat{g}_r\|^2 = \|g_r\|^2 + \frac{\text{tr}(\Sigma_r)}{N_r}, \quad (1)$$

with g_r the exact gradient and Σ_r the covariance of the per-point gradient [9, 10]. Enlarging the collocation set leaves g_r untouched but shrinks the inflation term; drawing the points from a different distribution changes g_r as well. In either case the *length* of the residual contribution changes, while the constraint term, evaluated on an unchanged set, does not. Since the optimiser only ever sees $\lambda_r \hat{g}_r$ next to $\lambda_c \hat{g}_c$, an intervention on the collocation set performs an implicit retuning of λ_r the very quantity the second family of methods tunes explicitly.

The consequence is an attribution problem. When a sampling scheme improves accuracy, part of the improvement may be a reweighting effect that could have been obtained, at no cost, by changing a scalar. Nothing in the standard ablation protocol distinguishes the two, because the weights are held fixed at their nominal values and the induced change in the effective balance is never measured.

Against this background, the contributions of this paper are as follows.

- (i) We identify and quantify a confound between collocation sampling and loss weighting in PINN training, and give a cheap online estimator of the effective interior-to-constraint balance (Section 3).
- (ii) We propose a matched-control experimental design that reproduces the effective balance of an intervention without touching the collocation set, together with an attribution ratio that reports what fraction of a reported gain survives the control (Section 4).
- (iii) We report a pre-registered power analysis showing that seed variance in this regime is of the same order as the effect being measured, and we derive the number of seeds required (Section 5). We regard this as a contribution in its own right: several published comparisons in this literature use fewer.
- (iv) We evaluate the design on three benchmarks and find that the reweighting explanation does not survive the control: where the intervention helps, the control recovers none of the gain, and matching the balance through the weights alone degrades accuracy (Section 7).
- (v) We release a verified implementation whose residual operators are checked against closed-form solutions and whose reference solutions are cross-checked by two independent numerical routes (Section 6, Appendix A).

Notation. $\|\cdot\|$ is the Euclidean norm on $\mathbb{R}^d$; $\theta \in \mathbb{R}^d$ collects the network parameters; $\odot$ and $\oslash$ denote element-wise multiplication and division. Expectations are with respect to the sampling of collocation points. N_r and N_c are the sizes of the interior and constraint batches.

2. Related Work

On collocation sampling, residual-based adaptive refinement was introduced alongside DeepXDE [3], and Wu et al. [4] conducted the reference comparison of the field, evaluating six uniform and several adaptive schemes over more than six thousand PINN runs and recommending residual-based adaptive distribution (RAD) as the most robust. Subsequent work has refined the acquisition rule using gradients of the solution, failure probabilities, or hybrid criteria: Daw et al. [11] retain, resample, and release points according to how persistently the residual stays large at each location; Nabian et al. [12] draw the collocation set by importance sampling from a distribution built on the residual norm; and Tang et al. [13] couple a deep generative model to the residual so that new points are drawn directly from the distribution it induces, extending the acquisition rule well beyond simple residual-power weighting. Across this literature the reported control is the baseline uniform sampler at fixed loss weights; the induced change in the effective weighting is not reported.

On the loss-weighting side, Wang et al. [5] traced the failure of stiff PINNs to disparities between the gradient

magnitudes of the loss terms and proposed a rebalancing rule built from those magnitudes; Wang et al. [6] derived a related scheme from the neural tangent kernel; McClenny and Braga-Neto [7] attached trainable weights to individual points, and Bischof and Kraus [8] treated the problem as multi-objective balancing. Xiang et al. [14] proposed a self-adaptive scheme that updates each loss term's weight from its own gradient statistics as training proceeds, and Anagnostopoulos et al. [15] connect a residual-based, point-wise attention mechanism to the information bottleneck principle, weighting each collocation point by how much it currently constrains the network. These methods measure precisely the quantity we argue is perturbed by sampling interventions, but they measure it in order to control it, not in order to attribute gains.

As for the interaction between the two, the closest work to ours is VW-PINNs [16], which analyses the ill-conditioning of the PDE loss under non-uniform collocation points and weights each residual by the volume its point occupies. That analysis concerns weighting within the residual term, across regions of the domain; ours concerns the balance between the residual term and the constraint terms, which is the quantity the adaptive-weighting literature tunes. Several papers combine adaptive sampling with adaptive weighting and report the combination as superior; none, to our knowledge, runs the experiment in the other direction and asks how much of a sampling gain is reweighting in disguise.

On experimental practice more broadly, Krishnapriyan et al. [17] showed that apparent optimiser failures in PINNs are often properties of the loss landscape rather than of the optimiser, an early example of the kind of re-attribution we pursue here. Wang et al. [18] later compiled a practical guide to training PINNs that catalogues many of the failure modes and mitigations scattered across this literature, sampling and weighting alike, but does not attempt to separate their individual contributions to a reported gain. Our statistical protocol follows standard practice in empirical machine learning: multiple seeds, a variance-aware metric, and an explicit power calculation..

3. The Confound

Let $\Omega \subset \mathbb{R}^n$ be a bounded domain and $\mathcal{N}$ a differential operator. A network u_θ is trained by minimising

$$F(\theta) = \lambda_r \mathcal{L}_r(\theta) + \lambda_c \mathcal{L}_c(\theta), \quad (2)$$

where the residual term is an empirical mean over the interior collocation set,

$$\mathcal{L}_r(\theta) = \frac{1}{N_r} \sum_{i=1}^{N_r} |\mathcal{N}[u_\theta](x_i) - f(x_i)|^2, \quad (3)$$

and $\mathcal{L}_c$ pools the boundary and initial-condition residuals in the same way. Grouping the constraints into a single term makes the balance between the two contributions one-dimensional, which is what the argument below requires; nothing prevents the analysis from being carried out with the terms kept separate, at the cost of a vector-valued balance.

This raises the question of why an intervention on the collocation set is also an intervention on the weights. Write $\hat{g}_r$ for the stochastic gradient of $\mathcal{L}_r$ computed on a batch of N_r points. Under uniform resampling, $\hat{g}_r$ is unbiased for the population gradient g_r , and its second moment obeys (1). Two observations follow.

First, enlarging the collocation set does not change what the residual term asks the network to do; it changes how long the vector is that asks for it. The optimiser step is driven by $\lambda_r \hat{g}_r + \lambda_c \hat{g}_c$, so shrinking $\|\hat{g}_r\|$ by reducing the inflation term is, from the optimiser's point of view, indistinguishable from reducing λ_r .

Second, adaptive sampling changes the sampling distribution and therefore changes g_r itself, not only its noise. Both the direction and the magnitude move, and the magnitude change again acts as an implicit reweighting.

We therefore define the *effective balance*

$$\rho_{\text{eff}} = \frac{\lambda_c \|\hat{g}_c\|}{\lambda_r \|\hat{g}_r\|}, \quad (4)$$

estimated online every P iterations at the cost of two extra backward passes. It is the quantity a sampling intervention perturbs without saying so.

Figure 1 reports ρ_{eff} and the noise-to-signal ratio $\text{tr}(\Sigma_r)/(N_r \|g_r\|^2)$ over the course of a single training run on the viscous Burgers problem, with $\|g_r\|^2$ and $\text{tr}(\Sigma_r)$ obtained by fitting (1) across batch sizes $N_r \in \{500, \dots, 8000\}$ with 25 draws each. The fit is accurate to better than two per cent throughout.

Three features matter. At initialisation the noise term is negligible, so the confound does not exist: any intervention at that stage is a pure sampling intervention. Once training is under way the noise-to-signal ratio settles between 0.13 and 0.28, and enlarging the interior batch from 500 to 8000 shifts ρ_{eff} by a factor of 1.2 to 1.4. The effect is real and measurable, but it is a tens-of-per-cent effect, not an order of magnitude, and any claim to the contrary would overstate what we observe.

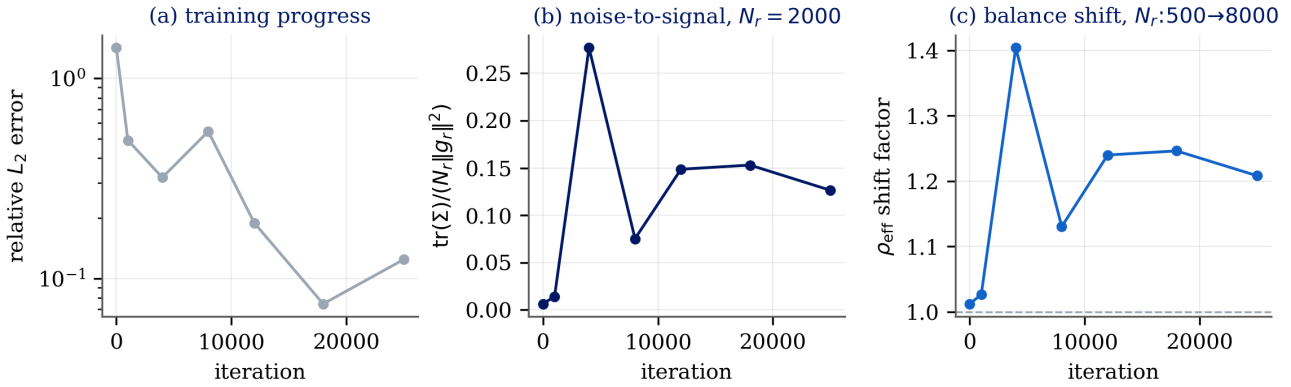


Figure 1. Pilot measurement on the Burgers problem: one training run of a network with four hidden layers of 48 units. (a) Training progress. (b) Noise-to-signal ratio of the residual gradient at $N_r = 2000$, from a fit of (1). (c) Factor by which the effective balance ρ_{eff} shifts when the interior batch is enlarged from 500 to 8000 points. The confound is absent at initialisation and settles at tens of per cent once training is under way.

4. Matched-Control Design

For each problem and each seed we run five arms. The baseline uses N_0 uniformly drawn interior points and nominal weights $(\lambda_r, \lambda_c) = (1, 1)$. Two intervention arms modify the collocation set at nominal weights: MORE-POINTS uses $4N_0$ uniform points, and RAD uses N_0 points drawn from the residual-based adaptive distribution of Wu et al. [4], $p(x) \propto |r(x)|^k / \mathbb{E}[|r|^k] + c$ with $k = c = 1$. Each intervention has a matched control arm, which uses the baseline collocation set and adjusts λ_r so that the measured effective balance follows the trajectory recorded in the intervention run.

To construct the control, the intervention run stores $\rho_{\text{eff}}(k)$ at every probe iteration. In the control run, at the same iterations, we measure the unweighted norms $\|\hat{g}_r\|$ and $\|\hat{g}_c\|$ on the baseline batches and solve (4) for the weight,

$$\lambda_r(k) = \frac{\lambda_c \|\hat{g}_c\|}{\rho_{\text{eff}}^{\text{target}}(k) \|\hat{g}_r\|}, \quad (5)$$

after which the pair is rescaled so that $\lambda_r + \lambda_c = 2$. The rescaling preserves the ratio, which is the quantity being matched, while holding the overall gradient scale fixed so that the meaning of the learning rate does not drift between arms. In our implementation the reproduced trajectory agrees with the target to 2.2×10^{-16} , that is to machine precision; Figure 2 shows the check.

We summarise the comparison with an attribution ratio. Let E_{base} , E_{int} and E_{ctrl} be the median relative L_2 errors of the baseline, the intervention and its matched control. Improvements in this field are multiplicative, so gains are measured on a logarithmic scale,

$$g_{\text{int}} = \log_{10} \frac{E_{\text{base}}}{E_{\text{int}}}, \quad g_{\text{ctrl}} = \log_{10} \frac{E_{\text{base}}}{E_{\text{ctrl}}}, \quad A = \frac{g_{\text{ctrl}}}{g_{\text{int}}}. \quad (6)$$

A near one means the intervention bought nothing that a change of scalar weights would not have bought; A near zero means the gain is genuinely attributable to where the points lie. A is undefined when the intervention does not help, so we report it only when g_{int} is clearly positive, and we accompany it with a bootstrap interval over seeds.

The design is also paired: Arms sharing a seed share their initial parameters and their constraint batches, drawn from a separate random stream, so that the comparison is paired. Denoting by r the correlation between paired arms on the logarithmic scale, the variance of the paired difference is $2\sigma^2(1-r)$, so the required number of seeds falls by the factor $(1-r)$.

5. Statistical Protocol

The effect under study is a shift of tens of per cent, and PINN training is noisy, so the design must be justified before it is run. We measure the seed-to-seed standard deviation σ of $\log_{10}$ error for a single arm at the full budget and obtain the number of seeds per arm from the two-sample normal approximation

$$n = \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2 \sigma^2}{\delta^2} (1-r), \quad \delta = \log_{10} \frac{1}{1-\varepsilon}, \quad (7)$$

for a target relative improvement ε at significance α and power $1-\beta$.

Measured on the baseline arm at the full budget over eight seeds, the Burgers problem gives $\sigma = 0.149$, a one-sigma swing of a factor of 1.41; the MORE-POINTS arm is noisier still, at $\sigma = 0.242$. The paired correlation between the two arms is $r = 0.359$, so pairing removes roughly a third of the variance but no more. Table 1 converts these figures into seed requirements.

The requirement is severe. Detecting a twenty per cent improvement at eighty per cent power needs between 38 and 99 seeds per arm depending on which arm's variance is used; thirty per cent needs between 15 and 39. We ran ten seeds per arm (eight on Burgers), which supports only effects of roughly a factor of two or larger. This is the central methodological finding of the present work, and it applies well beyond it: a large part of the PINN sampling literature reports comparisons at five seeds or fewer, on differences far smaller than the ones resolvable here. We return to the consequences in Section 8.

A second pilot varied λ_r directly at a reduced budget and found the error spread across λ_r to be $2.8\times$ within a seed against $1.8\times$ across seeds at fixed λ_r , the same order, confirming that the weight effect is not separable from seed noise without many repetitions. As the metric we therefore use the median of the evaluations in the final quarter of training rather than the final value, and we aggregate across seeds by the median.

Table 1. Seeds required per arm to detect a relative improvement ε at $\alpha = 0.05$ and 80% power, from (7), using the measured σ of the baseline and of the MORE-POINTS arm on Burgers and the measured paired correlation $r = 0.359$.

Target improvement ε	δ	$\sigma = 0.149$	$\sigma = 0.242$	paired ($r = 0.359$)
10%	0.046	167	441	107–282
20%	0.097	38	99	24–63
30%	0.155	15	39	10–25
50%	0.301	4	11	3–7

6. Experimental Setup

We use three problems spanning a smooth elliptic case, a case with a sharp interior layer, and a case with a thin metastable interface.

Viscous Burgers. $u_t + uu_x - \nu u_{xx} = 0$ with $\nu = 0.01/\pi$ on $x \in [-1, 1]$, $t \in [0, 1]$, initial datum $u(0, x) = -\sin(\pi x)$ and $u(t, \pm 1) = 0$ [19]. The initial datum is odd about $x = 0$ and about $x = \pm 1$, and both the advective and the diffusive operator preserve that symmetry, so the periodic solution satisfies the homogeneous boundary conditions identically; we verify this numerically to 1.2×10^{-16} .

Allen–Cahn. $u_t - d u_{xx} + 5u^3 - 5u = 0$ with $d = 10^{-4}$, periodic in x , $u(0, x) = x^2 \cos(\pi x)$.

High-frequency Poisson. $\Delta u + f = 0$ on $[0, 1]^2$ with homogeneous Dirichlet data and the manufactured solution $u = \sin(4\pi x) \sin(4\pi y)$, whence $f = 32\pi^2 \sin(4\pi x) \sin(4\pi y)$.

Turning to how the reference solutions were obtained and checked, the Poisson solution is exact. The two time-dependent references are computed with a Fourier pseudo-spectral discretisation and explicit RK4 in time, with the nonlinear products de-aliased by the two-thirds rule, at $N = 2048$ (Burgers) and $N = 1024$ (Allen–Cahn). Grid refinement gives relative differences of 1.4×10^{-6} and 2.7×10^{-6} against twice the resolution.

Because an error in a reference solution is invisible in the training curves, each reference is also checked against an independent numerical route. The Burgers reference is compared with a Cole–Hopf evaluation, in which the exact heat-kernel integral is computed in the logarithmic domain and no time stepping occurs at all; agreement is between 2.4×10^{-6} and 3.5×10^{-6} at $t \in \{0.25, 0.5, 0.75, 1\}$. The Allen–Cahn reference is compared with a second-order finite-difference solver, whose error against the spectral solution falls by exactly the expected factor of four under grid doubling. Appendix A collects the figures.

On the network, optimiser, and metric, all runs use a fully connected network with four hidden layers of 64 units and tanh activations, Glorot initialisation, inputs mapped affinely to $[-1, 1]^n$, and Adam at a fixed learning rate of 10^{-3} for 20,000 iterations. Batches are $N_0 = 2000$ interior and $N_c = 400$ constraint points, resampled every 100 iterations, and the effective balance is probed every 50 iterations. All arithmetic is in double precision: in single precision the second derivatives required by the residual lose enough significant digits to mask the effect under study. The metric is the relative L_2 error against the reference on a fixed grid, reported as the median over the evaluations in the final quarter of training and then over seeds: ten seeds per arm on Allen–Cahn and Poisson, eight on Burgers.

Finally, to verify the residual operators themselves, each residual is checked twice, against a closed-form solution of its own equation and against finite differences of the network itself. For Burgers the travelling wave $u = c - a \tanh(a(x - ct)/2v)$ is an exact solution, and our operator returns 8.9×10^{-16} on it. The manufactured Allen–Cahn residual agrees with the hand-derived expression exactly, and the Poisson residual vanishes on its manufactured solution to 5.7×10^{-14} . Automatic differentiation agrees with central differences to 2.5×10^{-8} on all three problems. These checks are the reason we trust the negative results below as much as the positive ones.

7. Results

Table 2 reports the accuracy of every arm. Three observations organise the rest of the discussion.

First, only one intervention on one problem produces a gain large enough to resolve at this sample size. Enlarging the collocation set fourfold on Burgers reduces the median relative L_2 error from 1.88×10^{-1} to 3.70×10^{-2} , a factor of 5.1. Everything else is small: on Poisson the two interventions improve the baseline by ten and eleven per cent, well inside the noise band established in Section 5, and on Burgers residual-based adaptive sampling is slightly *worse* than uniform sampling.

Second, the matched controls do not reproduce the gain. On Burgers the control that tracks the MORE-POINTS balance ends at 2.07×10^{-1} , marginally worse than the baseline it started from and a factor of 5.6 behind the intervention it was built to mimic.

Third, and unexpectedly, matching the balance through the weights is actively harmful. On Poisson the two control arms reach 3.31×10^{-1} and 3.49×10^{-1} against a baseline of 2.10×10^{-1} , that is 1.6 and 1.7 times worse. The same sign appears on Burgers. Reproducing the trajectory of ρ_{eff} without the mechanism that generated it degrades training rather than reproducing its benefit.

Turning to attribution, Table 3 gives the attribution ratio. It is interpretable in exactly one cell of the design, the one in which the intervention itself clearly helps, and there it is decisive: $A = -0.10$ with a bootstrap interval of $[-0.90, 0.22]$. The interval excludes $A = 1$ by a wide margin and is consistent with zero. On this benchmark the fivefold improvement from enlarging the collocation set is not reweighting in disguise; it is attributable to the sampling itself.

Elsewhere A is not defined, because g_{int} is either negative or too small to divide by, and we report the values only to show that the pre-registered rule was applied rather than chosen after the fact. The negative A values on Poisson reflect controls that are worse than the baseline, not a reweighting effect of the opposite sign.

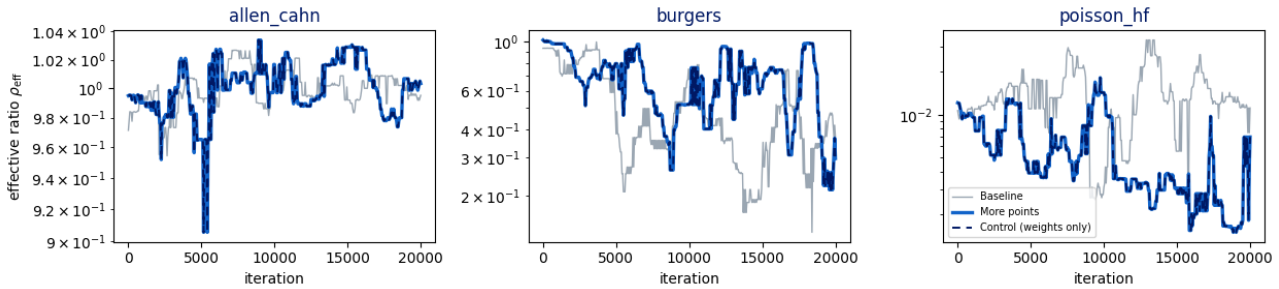
Table 2. Relative L_2 error against the reference solution: median over seeds of the median of the evaluations in the final quarter of training. Ranges over seeds are given in brackets. Lower is better.

Arm	Burgers ($n = 8$)	Allen–Cahn ($n = 10$)	Poisson ($n = 10$)
Baseline	1.88×10^{-1} [0.095–0.307]	9.96×10^{-1} [0.992–0.997]	2.10×10^{-1} [0.115–0.313]
More points ($4N_0$)	3.70×10^{-2} [0.022–0.106]	9.96×10^{-1} [0.991–0.998]	1.89×10^{-1} [0.119–0.238]
matched control	2.07×10^{-1} [0.092–0.638]	9.90×10^{-1} [0.943–1.000]	3.31×10^{-1} [0.193–0.522]
RAD sampling	2.25×10^{-1} [0.179–0.409]	9.99×10^{-1} [0.997–0.999]	1.87×10^{-1} [0.121–0.238]
matched control	2.01×10^{-1} [0.074–0.574]	9.99×10^{-1} [0.972–1.000]	3.49×10^{-1} [0.205–0.574]

Table 3. Attribution ratio A from (6) with a bootstrap interval over seeds. Following the pre-registered rule, A is interpreted only where $g_{\text{int}} > 0.05$.

Problem	Intervention	g_{int}	g_{ctrl}	A (95% CI)
Burgers	More points	0.707	−0.041	−0.10 [−0.90, 0.22]
Burgers	RAD	−0.077	−0.029	not defined
Allen–Cahn	More points	−0.000	0.003	not defined
Allen–Cahn	RAD	−0.001	−0.001	not defined
Poisson	More points	0.047	−0.197	not defined
Poisson	RAD	0.051	−0.220	not defined

The validity of this comparison rests on the control itself, and Figure 2 confirms that the control arm reproduces the effective balance of the intervention it mimics while sampling uniformly from the baseline collocation set, and that the baseline trajectory departs from both. The reproduction is exact to 2.2×10^{-16} by construction. Without this check the attribution ratio would be uninterpretable: a control that failed to match the balance would understate the reweighting contribution, and the conclusion of Section 7 would be an artefact of a badly built control rather than a property of the problem.

**Figure 2.** Effective balance ρ_{eff} over training, smoothed by a running median over 25 probes. The matched control (dashed, dark) lies on top of the intervention it mimics (solid, blue) while sampling uniformly from the baseline collocation set, whereas the baseline (grey) departs from both. The agreement is exact to 2.2×10^{-16} by construction; the figure confirms that the construction behaves as intended in every run.

One further result needs stating plainly before we draw conclusions: No arm learns anything on the Allen–Cahn problem: every configuration ends between 0.99 and 1.00 in relative L_2 error, which is the error of a network that has not left its initialisation in any useful direction. This is not a null result about sampling; it is a training failure, and it is a known one. The sharp interfaces of the Allen–Cahn solution defeat plain Adam-trained PINNs, which is precisely why the literature that treats this problem uses curriculum schemes, adaptive weighting, or a discrete-time formulation [7, 17]. We report the runs rather than removing them, and we exclude the problem from the attribution analysis: no intervention can be credited or discredited on a benchmark

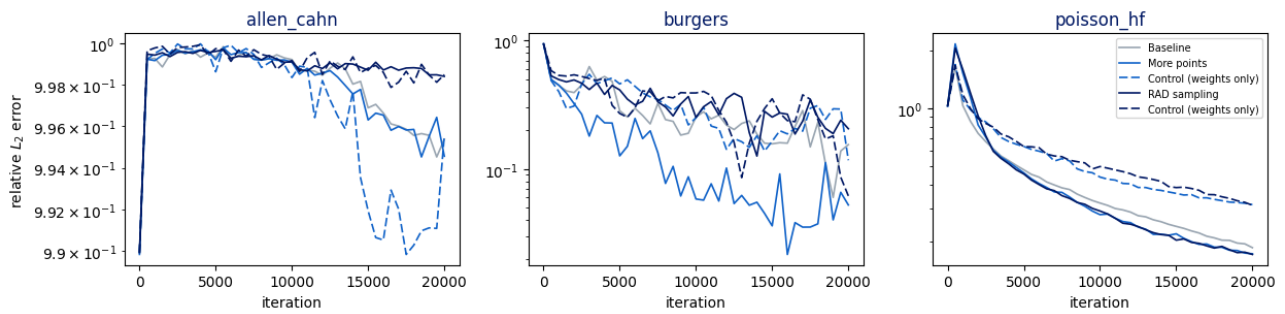


Figure 3. Median relative L_2 error against iteration for every arm. The gap between MORE-POINTS (solid, blue) and its matched control (dashed) on Burgers is the result that the attribution ratio summarises. On Poisson both controls sit above the baseline throughout training. On Allen–Cahn no arm departs from unit error note the vertical scale, which spans less than one per cent because training fails for all of them.

where the baseline itself does not train.

8. Discussion

On the one benchmark where enlarging the collocation set clearly helps, the gain is not reweighting. That is a negative answer to the question this paper set out to ask, and it is worth stating plainly: the confound we identified is real, measurable, and not the explanation for the improvement. The practical value lies in the exclusion. Adaptive and enlarged sampling schemes have been credited with geometric benefits on the basis of comparisons that never controlled for the induced change in loss weighting; that control has now been run, and the credit stands.

The result rests on a single benchmark, because it was the only one on which an intervention produced an effect large enough to be resolved at ten seeds. We regard the finding as solid for enlarging the collocation set on a problem with a moving interior layer, and as untested elsewhere.

The most surprising number in Table 2, however, is not the attribution ratio but the consistent degradation in the control arms: matching ρ_{eff} by adjusting λ_r makes training worse, by up to a factor of 1.7. We offer an interpretation rather than a demonstration. In the intervention arm the balance shifts because the residual gradient becomes a better estimate of its population value; the direction improves along with the magnitude. In the control arm only the magnitude is changed, and the change is applied to a noisier vector. Rescaling a noisy gradient upward amplifies the noise along with the signal, which is a plausible mechanism for the loss of accuracy. The same reasoning predicts that a control which matched both the magnitude and the variance would behave differently, and we return to this below.

It is also worth being precise about what the control does not equalise: in the MORE-POINTS arm the residual gradient is genuinely less noisy, whereas in the control it carries the baseline noise at a rescaled magnitude. The design therefore isolates the scale effect and leaves the variance effect inside what is attributed to sampling. Our A should accordingly be read as a lower bound on the share of the gain that is not geometric; the conclusion that sampling gains survive the control is, if anything, conservative. A sharper control would inject calibrated noise so that both the magnitude and the variance match, and we regard that as the natural next experiment.

Separately from the attribution question, the variance measurements deserve attention in their own right. With σ between 0.149 and 0.242 on the logarithmic scale, resolving a twenty per cent improvement requires tens of seeds per arm, and pairing on initialisation and constraint batches removes only about a third of the variance. Much of the sampling literature reports improvements of that order from far fewer runs. Our own Poisson numbers illustrate the trap: read without the variance analysis, the ten and eleven per cent gains there look like small confirmations of the sampling hypothesis, when the honest reading is that the experiment cannot see effects that size.

There is also a cost dimension worth flagging: the MORE-POINTS arm is four times more expensive per iteration, so the equal-iteration comparison in Table 2 flatters it. Wall-clock time is recorded for every run and

released with the code. The conclusion should be checked under an equal-time budget before it is generalised; a fivefold accuracy gain at fourfold cost per iteration is a genuine but narrower claim than the table alone suggests.

More broadly, the scope of what we have measured is limited: three low-dimensional problems, one architecture, and one optimiser, and one of the three problems failed to train at all. The mechanism in (1) is generic, but its magnitude depends on the heterogeneity of the per-point gradients: a stiff problem with a thin layer should show a larger $\text{tr}(\Sigma_r)$ than a smooth one, and therefore a larger confound. That prediction is untested here and is the obvious next measurement.

We close with a recommendation. Whatever the value of A on a given problem, the effective balance is cheap to measure two backward passes per probe and is perturbed by every intervention on the collocation set. We suggest that papers proposing such interventions report the induced change in ρ_{eff} alongside their accuracy tables, in the same way that learning-rate schedules are reported. Doing so removes an ambiguity that is otherwise invisible, and it costs almost nothing.

9. Conclusion

Interventions on the collocation set of a physics-informed neural network cannot alter the sampling geometry alone: because the residual loss is an empirical mean, they also rescale the residual gradient relative to the constraint terms and thereby retune the loss weights implicitly. We measured that rescaling and found it to be a tens-of-per-cent effect once training is under way, and we built a matched control that reproduces it to machine precision without touching the collocation set.

The control settles the attribution question in the direction that favours the existing literature. On the one benchmark where enlarging the collocation set clearly helps, reducing the error fivefold, the control recovers none of the gain: $A = -0.10$, with a bootstrap interval of $[-0.90, 0.22]$ that excludes the reweighting explanation. Matching the balance through the weights alone is not merely ineffective but harmful, degrading accuracy by up to a factor of 1.7. Sampling gains, at least in this setting, are geometric.

We also report a methodological result that we believe deserves wider attention: with a seed-to-seed standard deviation of 0.149 to 0.242 in $\log_{10}$ error, and a paired correlation of only 0.359, resolving a twenty per cent improvement in this setting requires tens of seeds per arm. Comparisons based on a handful of seeds and a final-iterate metric cannot support the conclusions they are routinely used to draw, and our own Poisson results are an example of differences that look meaningful until the variance is measured.

CRedit authorship contribution statement. Rasım ÇEKİK: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft, Writing – review & editing.

Funding. This research received no specific grant from any funding agency.

Declaration of competing interest. The author declares no competing financial or personal interest.

Data availability. Data available on request / reasonable request

Appendix A. Verification of the Implementation

Table 4 collects the checks described in Section 6. Every entry is produced by a test that is shipped with the code and runs in under a minute.

Appendix B. Pre-registered Analysis Plan

The following was fixed before the sweep was run.

1. The metric is the median of the relative L_2 evaluations in the final quarter of training, aggregated across seeds by the median.
2. The number of seeds follows (7) from a pilot estimate of σ : ten seeds per arm, reduced to eight on Burgers after two pilot-budget runs were excluded from the analysis directory.

Table 4. Verification of the residual operators and the reference solutions.

Check	Result
Burgers residual on the exact travelling wave	8.9×10^{-16}
Allen–Cahn residual against the hand-derived expression	0
Poisson residual on the manufactured solution	5.7×10^{-14}
Automatic differentiation against central differences	$\leq 2.5 \times 10^{-8}$
Spectral Burgers against Cole–Hopf quadrature	$2.4\text{--}3.5 \times 10^{-6}$
Spectral Allen–Cahn against second-order finite differences	1.3×10^{-6}
Reference data against the initial and boundary conditions	$\leq 10^{-16}$
Control arm against the target ρ_{eff} trajectory	2.2×10^{-16}

3. A is interpreted only when $g_{\text{int}} > 0.05$, that is when the intervention improves on the baseline by at least twelve per cent. Values falling below that threshold are tabulated but not read.
4. Hyperparameters are chosen on Burgers and Allen–Cahn and frozen before the Poisson runs.
5. No arm, problem, or seed is removed after the fact. Runs that fail to train are reported as such: the Allen–Cahn results are included in Table 2 and excluded only from the attribution analysis, for the stated reason that no baseline was established there.

References

- [1] 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. doi: 10.1016/j.jcp.2018.10.045.
- [2] George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021. doi: 10.1038/s42254-021-00314-5.
- [3] 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. doi: 10.1137/19M1274067.
- [4] Chenxi Wu, Min Zhu, Qinyang Tan, Yadhu Kartha, and Lu Lu. A comprehensive study of non-adaptive and residual-based adaptive sampling for physics-informed neural networks. *Computer Methods in Applied Mechanics and Engineering*, 403:115671, 2023. doi: 10.1016/j.cma.2022.115671.
- [5] 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. doi: 10.1137/20M1318043.
- [6] 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. doi: 10.1016/j.jcp.2021.110768.
- [7] Levi D. McClenny and Ulisses M. Braga-Neto. Self-adaptive physics-informed neural networks. *Journal of Computational Physics*, 474:111722, 2023. doi: 10.1016/j.jcp.2022.111722.
- [8] Rafael Bischof and Michael Kraus. Multi-objective loss balancing for physics-informed deep learning, 2021. URL <https://arxiv.org/abs/2110.09813>.
- [9] Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2): 223–311, 2018. doi: 10.1137/16M1080173.
- [10] Sam McCandlish, Jared Kaplan, Dario Amodei, and OpenAI Dota Team. An empirical model of large-batch training, 2018. URL <https://arxiv.org/abs/1812.06162>.
- [11] Arka Daw, Jie Bu, Sifan Wang, Paris Perdikaris, and Anuj Karpatne. Mitigating propagation failures in physics-informed neural networks using retain-resample-release (R3) sampling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *PMLR*, 2023.
- [12] Mohammad Amin Nabian, Rini Jasmine Gladstone, and Hadi Meidani. Efficient training of physics-informed neural networks via importance sampling. *Computer-Aided Civil and Infrastructure Engineering*, 36(8):962–977, 2021. doi: 10.1111/mice.12685.
- [13] Kejun Tang, Xiaoliang Wan, and Chao Yang. DAS-PINNs: A deep adaptive sampling method for solving high-dimensional partial differential equations. *Journal of Computational Physics*, 476:111868, 2023. doi: 10.1016/j.jcp.2022.111868.
- [14] Zixue Xiang, Wei Peng, Xu Liu, and Wen Yao. Self-adaptive loss balanced physics-informed neural networks. *Neurocomputing*, 496:11–34, 2022. doi: 10.1016/j.neucom.2022.05.015.
- [15] Sokratis J. Anagnostopoulos, Juan D. Toscano, Nikolaos Stergiopoulos, and George E. Karniadakis. Residual-based attention and connection to information bottleneck theory in PINNs, 2023. URL <https://arxiv.org/abs/2307.00379>.

- [16] Jiahao Song, Wenbo Cao, Fei Liao, and Weiwei Zhang. VW-PINNs: A volume weighting method for PDE residuals in physics-informed neural networks. *Acta Mechanica Sinica*, 41(3):324140, 2025. doi: 10.1007/s10409-024-24140-x.
- [17] Aditi S. Krishnapriyan, Amir Gholami, Shandian Zhe, Robert M. Kirby, and Michael W. Mahoney. Characterizing possible failure modes in physics-informed neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 26548–26560, 2021.
- [18] Sifan Wang, Shyam Sankaran, Hanwen Wang, and Paris Perdikaris. An expert’s guide to training physics-informed neural networks, 2023. URL <https://arxiv.org/abs/2308.08468>.
- [19] C. Basdevant, M. Deville, P. Haldenwang, J. M. Lacroix, J. Ouazzani, R. Peyret, P. Orlandi, and A. T. Patera. Spectral and finite difference solutions of the Burgers equation. *Computers & Fluids*, 14(1):23–41, 1986. doi: 10.1016/0045-7930(86)90036-8.