MULTISCALE REGULARIZATION OF PHYSICS-INFORMED NEURAL NETWORKS BASED ON A MAXIMAL LOCAL OSCILLATION OPERATOR FOR PROBLEMS WITH SHARP GRADIENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper proposes a multiscale regularizer for physics-informed neural networks (PINNs) designed for differential equation problems with sharp spatial gradients. The regularizer is based on a maximal local oscillation operator that measures the deviation of a function from its local mean across several scales. Unlike standard penalties on high-order derivatives, the proposed approach does not require their explicit computation and naturally accounts for the multiscale structure of the solution. For practical use, we introduce a differentiable discrete implementation of the operator on a finite set of radii using one-dimensional convolutions, compatible with automatic differentiation in PyTorch. The method is integrated into the PINN loss function as an additional regularizer that suppresses local nonphysical oscillations. Using the viscous Burgers equation as an example, we compare the method against a standard PINN and several classical smoothness penalties at several viscosity values; the exact solution obtained via the Cole–Hopf transformation serves as the reference. Numerical results show that, in the front-formation regime, the proposed approach achieves the best mean approximation accuracy among the methods considered while preserving the sharpness of the front and avoiding the over-smoothing characteristic of penalties on high-order derivatives; in the extremely smooth and extremely stiff regimes the method is on par with the alternatives.

1 INTRODUCTION

Physics-informed neural networks (PINNs) have become one of the most prominent directions in scientific machine learning following the works of Raissi, Perdikaris, and Karniadakis (Raissi et al., 2019; 2020). The idea of the method is to train a neural network u_θ by minimizing not only the data misfit but also the residual of the differential equation, as well as the residuals of the initial and boundary conditions. This approach makes it possible to construct approximate solutions for a broad class of problems in mathematical physics without explicitly relying on classical numerical schemes.

Despite the appeal of PINNs, standard formulations often turn out to be unstable for equations with sharp gradients and fronts (Wang et al., 2021; Krishnapriyan et al., 2021). In particular, one may observe:

1. local nonphysical oscillations near fronts;
2. excessive smoothing of the solution;
3. poor reproduction of sharp transitions as the diffusion coefficient or viscosity is decreased.

Such effects are well known for convection–diffusion problems and conservation laws. In the case of the viscous Burgers equation, small values of the viscosity ν produce steep spatial profiles on which a standard PINN either exhibits spurious oscillations or replaces the front with an overly smooth transition.

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.

One way to improve the stability of PINNs is to introduce additional regularizers. The literature includes penalties on $\|u_x\|_{L^2}$ and $\|u_{xx}\|_{L^2}$, TV-type functionals, weak-form formulations, and various adaptive strategies (Jagtap et al., 2020; Kharazmi et al., 2021; Czarnecki et al., 2017; Yu et al., 2022; Rudin et al., 1992; Vogel, 2002). However, many of these methods are either tailored to a single local scale or require the computation of high-order derivatives, which can be computationally expensive and numerically sensitive.

In this work, we propose a multiscale local oscillation regularizer rooted in ideas from harmonic analysis. Instead of directly penalizing high-order derivatives, we consider the deviation of the solution from its local mean at several spatial scales, followed by aggregation over the radii. This approach:

- measures the local inhomogeneity of the solution;
- does not require the explicit computation of high-order derivatives;
- naturally accounts for several characteristic spatial scales.

From a theoretical standpoint, the method is inspired by constructions related to local means, sharp maximal functions, and oscillation-based characterizations of regularity (Stein, 1970; Adams & Fournier, 2003; Triebel, 1983; Bennett & Sharpley, 1988). At the same time, the present work is applied in nature: we do not formulate a new general theorem characterizing Sobolev spaces, but instead use these ideas to build a computationally efficient regularizer within the PINN framework.

Main contributions.

1. We propose a multiscale regularizer for PINNs based on a maximal local oscillation operator.
2. We develop a differentiable discrete implementation of the operator on a finite set of scales using one-dimensional convolutions.
3. We compare the method against several standard regularizers on the viscous Burgers equation using an exact reference solution.
4. We study the method at several viscosity values corresponding to different degrees of gradient sharpness.

2 Related Work

The basic concept of physics-informed neural networks was systematically formulated in (Raissi et al., 2019; 2020). Subsequent studies showed that, for stiff problems and problems with sharp local structures, the standard PINN can exhibit a number of characteristic issues related to gradient flow pathologies, poor balancing of the loss terms, and difficulties in approximating high-frequency components (Wang et al., 2021; Krishnapriyan et al., 2021).

An important line of research concerns conservative and weak-form PINNs, including variational formulations such as VPINNs (Jagtap et al., 2020; Kharazmi et al., 2021). These methods often prove more robust on problems with low solution regularity or when dealing with conservation laws.

There is also a family of methods that use derivatives of the solution during training. These include Sobolev training (Czarnecki et al., 2017) and gradient-enhanced PINNs (Yu et al., 2022), where additional derivative information is incorporated into the loss function. Such approaches provide better control over the smoothness of the approximation but typically require the computation of higher-order derivatives and can be sensitive to weight tuning.

To suppress local artifacts, H^1- and H^2-type penalties, as well as TV-type regularizers, are widely used (Rudin et al., 1992; Vogel, 2002). Their limitation is that they are generally geared toward a single local mechanism of regularity control and are not always flexible enough for multiscale solutions.

The mathematical motivation for our approach goes back to ideas from harmonic analysis and the theory of function spaces. The classical monographs by Stein (Stein, 1970), Adams and Fournier (Adams & Fournier, 2003), Triebel (Triebel, 1983), and Bennett–Sharpley (Bennett & Sharpley,

1988) show that local means, local oscillations, and maximal constructions play an important role in describing the smoothness of functions.

3 Multiscale Local Oscillation Operator

Let $u : \Omega \to \mathbb{R}$, where $\Omega \subset \mathbb{R}$ is the spatial domain. For a point $x \in \Omega$ and a radius $r > 0$, denote by

$$u_B(x,r) = \fint_{B(x,r)} u(y)\,dy = \frac{1}{|B(x,r)|}\int_{B(x,r)} u(y)\,dy \tag{1}$$

the local mean of the function over the neighborhood $B(x,r) = (x-r, x+r)\cap\Omega$. Near the boundary of the domain, the integral is defined by extending the function (in the practical implementation, padding is used).

To quantify local nonsmoothness, we introduce the local oscillation operator:

$$M_{\mathrm{osc}}u(x) = \max_{r\in R}\left(\fint_{B(x,r)} |u(y) - u_B(x,r)|^q\,dy\right)^{1/q}, \qquad q \geq 1, \tag{2}$$

where R is a finite set of radii.

Unlike the classical Hardy–Littlewood maximal operator, which measures the local magnitude of a function, the operator M_{osc} measures the deviation of the function from its local mean and thus serves as an indicator of local inhomogeneity. If u is nearly constant in a neighborhood of a point x, the value of $M_{\mathrm{osc}}u(x)$ is small. If, on the other hand, sharp transitions, fronts, or spurious oscillations are present near x, the value of the operator increases.

In the practical implementation, we use a finite set of physical radii $R = \{r_1, \ldots, r_K\}$ chosen according to the dyadic rule $r_{k+1} = 2r_k$; the specific values are given in the hyperparameter table (Table 1).

4 Discretization and Implementation

4.1 Discrete Local Mean

Suppose that, on a fixed time slice t_j, the values of the neural network solution are computed on a uniform grid $x_1, \ldots, x_{N_x}$. For each radius $r \in R$, we introduce the discrete local mean:

$$\bar{u}_r(t_j, x_i) = \frac{1}{N_r}\sum_{x_m \in B_r(x_i)} u_\theta(t_j, x_m), \tag{3}$$

where $B_r(x_i)$ is the discrete neighborhood of the point x_i of radius r, and N_r is the number of points in this neighborhood.

4.2 Discrete Local Oscillation

At the same scale, we define the discrete oscillation for $q = 2$:

$$\mathrm{Osc}_r(u_\theta)(t_j, x_i) = \left(\frac{1}{N_r}\sum_{x_m \in B_r(x_i)} \big(u_\theta(t_j, x_m) - \bar{u}_r(t_j, x_i)\big)^2\right)^{1/2}. \tag{4}$$

The choice $q = 2$ makes it possible to compute the quantity (4) exactly with two convolution operations, thanks to the variance identity:

$$\mathrm{Osc}_r^2(u_\theta)(t_j, x_i) = \overline{u_\theta^2}_r(t_j, x_i) - \big(\bar{u}_r(t_j, x_i)\big)^2, \tag{5}$$

where $\overline{u_\theta^2}_r$ is the local mean of the squared solution over the same window. Thus, each scale requires only two one-dimensional convolutions with a constant normalized kernel and no derivative computations.

The discrete multiscale operator then takes the form:

$$\widetilde{M}_{\mathrm{osc}} u_\theta(t_j, x_i) = \max_{r \in R} \mathrm{Osc}_r(u_\theta)(t_j, x_i). \tag{6}$$

To obtain a smoother computational scheme, the maximum is approximated by the log-sum-exp function:

$$\widetilde{M}^{(\tau)}_{\mathrm{osc}} u_\theta(t_j, x_i) = \tau \log \left(\sum_{r \in R} \exp \left(\frac{\mathrm{Osc}_r(u_\theta)(t_j, x_i)}{\tau} \right) \right), \qquad \tau > 0. \tag{7}$$

4.3 Connection to the Software Implementation

In the code, the local means $\bar{u}_r$ and $\overline{u^2_{\theta_r}}$ are implemented via one-dimensional convolutions (`Conv1d`) with fixed normalized kernels, after which the oscillation is computed via the identity (5). Boundary effects are handled correctly using the replicate padding mode. For each time slice, the profile $u_\theta(t_j, \cdot)$ is treated as a one-dimensional signal. All operations used are compatible with automatic differentiation in PyTorch, which allows the regularizer to be included in the loss function without modifying the backpropagation algorithm.

5 Regularized PINN

Consider the viscous Burgers equation:

$$u_t + u u_x = \nu u_{xx}, \qquad (t, x) \in [0, T] \times [-1, 1], \tag{8}$$

with the initial condition $u(0, x) = -\sin(\pi x)$ and Dirichlet boundary conditions $u(t, -1) = u(t, 1) = 0$.

Let $u_\theta(t, x)$ be the neural network approximation of the solution. Training minimizes the functional:

$$L(\theta) = \lambda_{PDE} L_{PDE} + \lambda_{IC} L_{IC} + \lambda_{BC} L_{BC} + \lambda_{reg} L_{reg}, \tag{9}$$

where the equation residual penalty is computed on a set of collocation points (t_i, x_i):

$$L_{PDE} = \frac{1}{N_f} \sum_{i=1}^{N_f} \left(\frac{\partial u_\theta}{\partial t}(t_i, x_i) + u_\theta(t_i, x_i) \frac{\partial u_\theta}{\partial x}(t_i, x_i) - \nu \frac{\partial^2 u_\theta}{\partial x^2}(t_i, x_i) \right)^2. \tag{10}$$

The corresponding terms for the initial and boundary conditions are:

$$L_{IC} = \frac{1}{N_0} \sum_{i=1}^{N_0} \left(u_\theta(0, x_i) - u_0(x_i) \right)^2, \tag{11}$$

$$L_{BC} = \frac{1}{N_b} \sum_{i=1}^{N_b} \left(u_\theta(t_i, -1)^2 + u_\theta(t_i, 1)^2 \right). \tag{12}$$

The local oscillation regularizer is applied along the spatial coordinate on a uniform grid of size $N_t \times N_x$:

$$L_{reg} = \frac{1}{N_t N_x} \sum_{j=1}^{N_t} \sum_{k=1}^{N_x} \left(\widetilde{M}^{(\tau)}_{\mathrm{osc}} u_\theta(t_j, x_k) \right)^2. \tag{13}$$

In this work, the regularization is applied specifically along the spatial coordinate x, since the main goal is to suppress spatial oscillations near sharp fronts.

6 Alternative Regularizers

For the comparative analysis, we consider the following standard smoothness penalties (computed on the same collocation set):

Table 1: Experiment hyperparameters.

Parameter	**Value**
Network architecture	MLP, 3 hidden layers of 64 neurons, tanh
Optimizer	Adam, lr $= 10^{-3}$, 3000 epochs
Collocation grid (PDE)	$N_t \times N_x = 50 \times 128$
IC / BC points	$N_0 = 128$, $N_b = 128$
Loss term weights	$\lambda_{PDE} = 1$, $\lambda_{IC} = 10$, $\lambda_{BC} = 1$
Radii R	$\{0.02, 0.04, 0.08, 0.16\}$
Operator parameters	$q = 2$, log-sum-exp with $\tau = 0.05$
λ_{reg} (our method)	10^{-4} (chosen via the ablation study, Sec. 7.5)
λ_{reg} (Grad / Hessian / TV)	10^{-4} / 10^{-6} / 10^{-4}
Number of independent runs	5 (seeds 17, 42, 123, 256, 777)

1. Gradient penalty (control of the H^1-seminorm):

$$L_{\nabla} = \frac{1}{N_f} \sum_{i=1}^{N_f} \left(\frac{\partial u_\theta}{\partial x}(t_i, x_i) \right)^2. \tag{14}$$

2. Second-derivative penalty (control of the H^2-seminorm):

$$L_{H^2} = \frac{1}{N_f} \sum_{i=1}^{N_f} \left(\frac{\partial^2 u_\theta}{\partial x^2}(t_i, x_i) \right)^2. \tag{15}$$

3. TV-type penalty (Total Variation):

$$L_{TV} = \frac{1}{N_f} \sum_{i=1}^{N_f} \sqrt{\left(\frac{\partial u_\theta}{\partial x}(t_i, x_i) \right)^2 + \varepsilon}, \qquad \varepsilon > 0. \tag{16}$$

These regularizers are applied within the same overall training scheme as the proposed method and serve as baselines for comparison.

7 Numerical Experiments

7.1 Setup

The numerical experiments are carried out for the viscous Burgers equation on the domain $(t, x) \in [0, 1] \times [-1, 1]$ at several viscosity values: $\nu \in \{0.1/\pi,\ 0.01/\pi,\ 0.005/\pi\}$. As ν decreases, the solution develops increasingly sharp spatial profiles, which makes it possible to assess the robustness of the method across regimes of varying stiffness. The hyperparameters of all experiments are listed in Table 1.

7.2 Methods Compared

We compare: (1) a standard PINN without additional regularization; (2) a PINN with L_{∇}; (3) a PINN with L_{H^2}; (4) a PINN with L_{TV}; and (5) a PINN with the proposed regularizer L_{reg}.

7.3 Reference Solution and Metrics

The reference u_{ref} is the exact solution of the problem obtained via the Cole–Hopf transformation. For the initial and boundary conditions under consideration, the solution is expressed as a ratio of integrals

$$u_{ref}(t, x) = -\frac{\displaystyle\int_{-\infty}^{\infty} \sin\big(\pi(x-\eta)\big)\, f(x-\eta)\, e^{-\eta^2/(4\nu t)}\, d\eta}{\displaystyle\int_{-\infty}^{\infty} f(x-\eta)\, e^{-\eta^2/(4\nu t)}\, d\eta}, \qquad f(y) = \exp\left(-\frac{\cos(\pi y)}{2\pi\nu}\right), \tag{17}$$

Table 2: Comparison of accuracy and equation residual for the different methods at various viscosities (averaged over 5 independent runs; the reference is the exact Cole–Hopf solution).

Viscosity ν	**Method**	**Rel. L_2**	L_∞	**MSE ($\times 10^{-3}$)**	**PDE Res. ($\times 10^{-2}$)**
$0.1/\pi$	Baseline	0.004 ± 0.002	0.010 ± 0.005	0.01 ± 0.00	0.06 ± 0.04
	Grad	0.006 ± 0.003	0.012 ± 0.005	0.01 ± 0.01	0.07 ± 0.04
	Hessian	0.005 ± 0.002	0.011 ± 0.005	0.01 ± 0.01	0.07 ± 0.04
	TV	0.004 ± 0.002	0.010 ± 0.005	0.01 ± 0.00	0.06 ± 0.04
	Ours (Osc)	0.015 ± 0.018	0.018 ± 0.016	0.15 ± 0.26	0.19 ± 0.25
$0.01/\pi$	Baseline	0.185 ± 0.077	1.194 ± 0.518	14.60 ± 9.60	6.85 ± 1.25
	Grad	0.189 ± 0.080	1.179 ± 0.537	15.31 ± 10.31	6.49 ± 1.83
	Hessian	0.174 ± 0.032	0.775 ± 0.033	11.68 ± 4.46	9.30 ± 2.44
	TV	0.158 ± 0.071	0.980 ± 0.540	10.89 ± 8.37	6.60 ± 1.63
	Ours (Osc)	$\mathbf{0.138 \pm 0.052}$	0.881 ± 0.422	$\mathbf{8.01 \pm 5.07}$	6.62 ± 1.60
$0.005/\pi$	Baseline	0.291 ± 0.033	0.922 ± 0.017	32.59 ± 7.01	15.26 ± 1.11
	Grad	0.293 ± 0.033	0.913 ± 0.021	32.99 ± 7.04	15.27 ± 1.15
	Hessian	0.307 ± 0.017	0.929 ± 0.021	36.01 ± 3.93	15.64 ± 1.21
	TV	0.291 ± 0.034	0.921 ± 0.026	32.56 ± 7.22	15.48 ± 1.22
	Ours (Osc)	0.291 ± 0.033	0.920 ± 0.016	32.52 ± 7.05	15.32 ± 1.10

which are evaluated by Gauss–Hermite quadrature. Using the exact solution eliminates the numerical error of the reference, which is critical at small viscosity values.

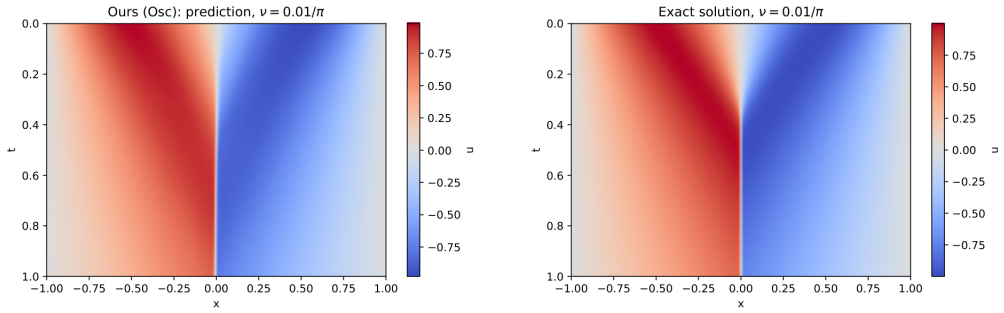
The comparison is carried out using the following metrics: relative error in the L_2 norm, $\mathrm{Rel}_{L_2} = \|u_\theta - u_{ref}\|_{L_2} / \|u_{ref}\|_{L_2}$; maximum error, $\mathrm{Err}_\infty = \|u_\theta - u_{ref}\|_{L_\infty}$; mean squared error, $\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N} |u_\theta - u_{ref}|^2$; and mean squared PDE residual on the test grid.

7.4 Results and Discussion

The summary results of the computational experiments, averaged over 5 independent runs with different random initializations, are presented in Table 2.

The analysis of the results reveals three characteristic operating regimes of the methods depending on the viscosity ν:

1. **Smooth regime ($\nu = 0.1/\pi$).** In the high-diffusion case, the solution contains no sharp fronts, and all the methods under study solve the problem with high accuracy (a relative L_2 error of about 0.4–1.5%). The proposed regularizer does not disturb the natural smoothness of the solution; the somewhat larger spread of its results (0.015 ± 0.018) is due to a few unlucky runs and remains within the range of small errors.
2. **Front-formation regime ($\nu = 0.01/\pi$).** In this regime, the differences between the methods become substantive (Figs. 1, 2). The proposed method shows the lowest mean L_2 error (0.138 ± 0.052) and the lowest MSE (8.01×10^{-3}), ahead of TV (0.158), Hessian (0.174), the baseline PINN (0.185), and Grad (0.189); note, however, that the confidence intervals of the methods overlap. The behavior of the equation residual is telling. The second-derivative penalty achieves the lowest maximum error ($L_\infty = 0.775$) at the cost of nonphysical smoothing of the front: the smoothed profile produces no peak error spike, but this is reflected in the worst equation residual (9.30×10^{-2}). The proposed method ranks second in L_∞ (0.881) while preserving the sharpness of the front (Fig. 2) and keeping the residual on par with the best methods ($6.62 \pm 1.60 \times 10^{-2}$ versus 6.49 for Grad and 6.85 for the baseline PINN). Thus, local oscillation regularization improves the approximation of the front without imposing excessive smoothness on the solution; the value $\lambda_{reg} = 10^{-4}$ was chosen based on the ablation study (Sec. 7.5).
3. **Stiff regime ($\nu = 0.005/\pi$).** At extremely small viscosity, all five methods produce statistically indistinguishable results (a relative L_2 error of about 29%, with differences within the standard deviations). This means that the limiting factor in this regime is not the type of regularization but the ability of the basic multilayer perceptron architecture, under the given training budget (3000 epochs, Adam), to approximate the extremely sharp front — a manifestation of spectral bias (Fig. 6). Modifying the loss function alone is insufficient



(a) Prediction (Ours, Osc) (b) Exact solution

Figure 1: Solution of the Burgers equation at $\nu = 0.01/\pi$: the prediction of the PINN with the proposed regularizer (left) and the exact Cole–Hopf solution (right). The method correctly reproduces the formation of the sharp front near $x = 0$.

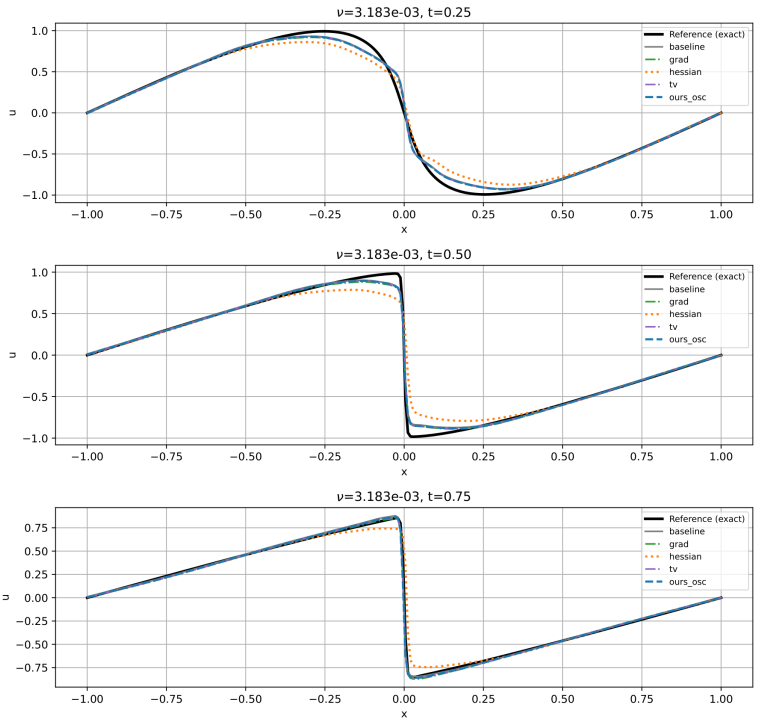


Figure 2: Solution slices $u(t, x)$ at $t = 0.25, 0.5, 0.75$ for $\nu = 0.01/\pi$: the exact solution and all five compared methods. The methods exhibit different error characteristics: the second-derivative penalty smooths the front region, whereas the remaining methods, including the proposed one, preserve a steeper transition profile.

here: the local oscillation regularization method should be combined with adaptive collocation point sampling or architectural improvements (e.g., Fourier features), which is a subject of future research.

Additional diagnostics — the absolute error map, training convergence curves, and the stiff-regime solution slices — are given in Appendix B (Figs. 4–6).

7.5 Ablation Study

To assess the sensitivity of the method to its hyperparameters, we conduct two ablation experiments in the front-formation regime ($\nu = 0.01/\pi$), where the differences between the methods are most substantive: a sweep over the regularization coefficient λ_{reg}, and a sweep over the set of scales R. As λ_{reg} grows, the mean equation residual increases monotonically while the maximum error L_∞ decreases, i.e. the regularizer progressively smooths the solution, approaching the behavior of penalties on high-order derivatives; the best balance is achieved at $\lambda_{reg} = 10^{-4}$, which is adopted as the primary value in the experiments of Table 2. We note that the coefficient was selected on the same problem and by the same metrics as the final comparison, so the transferability of this specific value to other equations and regimes requires separate verification. Varying the set of radii R (with $\lambda_{reg} = 10^{-3}$ fixed) shows that the configurations are statistically indistinguishable in accuracy; in this one-dimensional problem with a single front, one correctly chosen scale is already sufficient, while the advantages of the multiscale construction are expected to emerge in problems with several characteristic spatial scales. Full tables and the corresponding figure are given in Appendix A.

8 Limitations of the Method

The proposed approach is primarily aimed at problems with sharp but finite gradients. In the case of genuine discontinuities of the solution, overly strong regularization can lead to excessive smearing of the front. In addition, the method is sensitive to the choice of the coefficient λ_{reg} and the set of radii R. A λ_{reg} that is too small makes the contribution of the regularization negligible, whereas one that is too large leads to over-smoothing. The ablation study conducted in Appendix A outlines the working range of the hyperparameters for the problem considered; however, when transferring the method to other equations and regimes, their values may require re-tuning.

9 Conclusion

This paper proposes a multiscale regularizer for physics-informed neural networks based on a maximal local oscillation operator. Unlike standard penalties on high-order derivatives, this approach measures the deviation of the solution from its local mean at several scales and is therefore better suited to problems with sharp spatial gradients.

The discrete implementation developed here uses one-dimensional convolutions, computes the oscillation exactly via the variance identity, and is compatible with automatic differentiation in PyTorch. Using the viscous Burgers equation at several viscosity values, with the exact Cole–Hopf reference solution, we show that in the front-formation regime the method achieves the best mean approximation accuracy among the approaches considered, preserving the sharpness of the front and exhibiting none of the over-smoothing characteristic of penalties on high-order derivatives. In the regime of extremely small viscosity, all the methods considered run into the limitations of the basic architecture, which delineates the boundaries of applicability of the regularization approach as such.

Promising directions for future work include extending the method to multidimensional problems, adaptive scale selection, combining it with weak-form PINNs, and a more rigorous theoretical analysis of its connection to functional characterizations of smoothness.

AI use statement

Generative AI tools (a large language model, Claude) were used to assist in converting the manuscript from its original PDF/typeset form into this LaTeX source (equation, table, and figure formatting), in preparing the submission checklist against ICLR 2027 formatting requirements, and in translating the code comments, docstrings, and documentation of the accompanying code repository from Russian into English. The AI tool was not used to generate research ideas, run experiments, or produce numerical results; all reported results derive from the code and experiments described in the paper. All AI-assisted formatting and translation was reviewed by the authors against the original manuscript and code for correctness. The authors take responsibility for the final content of this work, including text, claims, and artifacts.

Reproducibility statement

The full source code of the software implementation of the proposed method, along with scripts for reproducing all numerical experiments, is available in an anonymized public repository at `https://anonymous.4open.science/r/Multiscale-MLO-PINNs-0FBC/README.md`. The LaTeX source of this manuscript is included as supplementary material. Training hyperparameters, the set of random seeds, and the exact form of all loss terms are given in Section 7 and Table 1. The exact reference solution used for evaluation is given in Eq. (17) and computed via Gauss–Hermite quadrature.

References

Robert A. Adams and John J. F. Fournier. *Sobolev Spaces*. Academic Press, 2nd edition, 2003.

Colin Bennett and Robert Sharpley. *Interpolation of Operators*. Academic Press, 1988.

Wojciech M. Czarnecki, Simon Osindero, Max Jaderberg, Grzegorz Swirszcz, and Razvan Pascanu. Sobolev training for neural networks. *Advances in Neural Information Processing Systems*, 30, 2017.

Ameya D. Jagtap, Ehsan Kharazmi, and George E. Karniadakis. Conservative physics-informed neural networks on discrete domains for conservation laws: Applications to forward and inverse problems. *Computer Methods in Applied Mechanics and Engineering*, 365:113028, 2020.

Ehsan Kharazmi, Zhongqiang Zhang, and George E. Karniadakis. hp-vpinns: Variational physics-informed neural networks with domain decomposition. *Computer Methods in Applied Mechanics and Engineering*, 374:113547, 2021.

Aditi Krishnapriyan, Amir Gholami, Shandian Zhe, Robert Kirby, and Michael W. Mahoney. Characterizing possible failure modes in physics-informed neural networks. *Advances in Neural Information Processing Systems*, 34:26548–26560, 2021.

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.

Maziar Raissi, Alireza Yazdani, and George E. Karniadakis. Hidden fluid mechanics: Learning velocity and pressure fields from flow visualizations. *Science*, 367(6481):1026–1030, 2020.

Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4):259–268, 1992.

Elias M. Stein. *Singular Integrals and Differentiability Properties of Functions*. Princeton University Press, 1970.

Hans Triebel. *Theory of Function Spaces*. Birkhäuser, 1983.

Curtis R. Vogel. *Computational Methods for Inverse Problems*. SIAM, 2002.

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.

Jeremy Yu, Lu Lu, Xuhui Meng, and George E. Karniadakis. Gradient-enhanced physics-informed neural networks for forward and inverse pde problems. *Computer Methods in Applied Mechanics and Engineering*, 393:114823, 2022.

A Ablation Study: Full Results

Regularization coefficient. In the first experiment, the coefficient λ_{reg} is varied from 10^{-4} to 10^{-1} with the set of scales R fixed; the point $\lambda_{reg} = 0$ corresponds to the baseline PINN. The results are given in Table 3 and Fig. 3. As λ_{reg} grows, the mean equation residual increases monotonically (from 6.62 to 8.01×10^{-2}) while the maximum error L_∞ simultaneously decreases (from 0.881 to 0.855): the regularizer smooths the solution progressively more strongly, approaching the behavior of penalties on high-order derivatives. The L_2 error is non-monotonic, yet for all $\lambda_{reg} > 0$ it does not exceed the baseline PINN level. The best balance is achieved at $\lambda_{reg} = 10^{-4}$: the accuracy improves over the baseline PINN (0.138 versus 0.185) without degradation of the equation residual; this value is adopted as the primary one in the experiments of Table 2.

Table 3: Ablation over the regularization coefficient λ_{reg} at $\nu = 0.01/\pi$ (5 independent runs).

λ_{reg}	**Rel.** L_2	L_∞	**PDE Residual** ($\times 10^{-2}$)
0 (Baseline)	0.185 ± 0.077	1.194 ± 0.518	6.85 ± 1.25
10^{-4}	0.138 ± 0.052	0.881 ± 0.422	6.62 ± 1.60
10^{-3}	0.194 ± 0.087	1.207 ± 0.516	6.65 ± 0.63
10^{-2}	0.152 ± 0.071	0.966 ± 0.516	7.16 ± 2.03
10^{-1}	0.135 ± 0.057	0.855 ± 0.399	8.01 ± 2.68

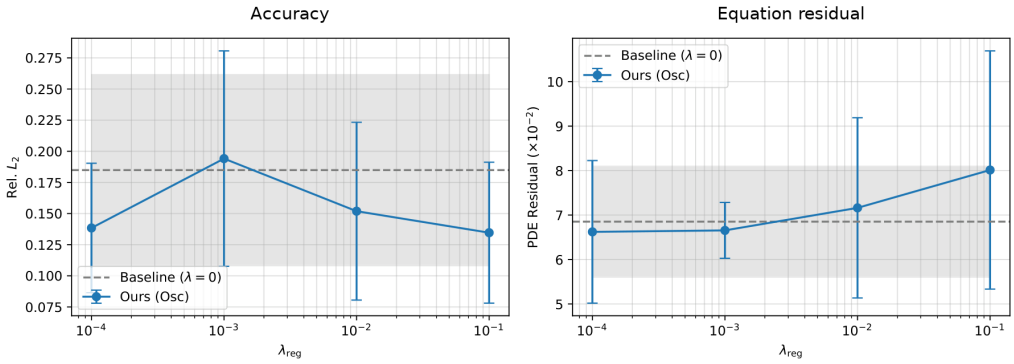


Figure 3: Dependence of the relative L_2 error (left) and the equation residual (right) on the regularization coefficient λ_{reg} at $\nu = 0.01/\pi$. The dashed line and the gray band show the mean and standard deviation of the baseline PINN ($\lambda_{reg} = 0$).

Set of scales. In the second experiment, with $\lambda_{reg} = 10^{-3}$ fixed, the set of radii R is varied: a single small scale, a single large scale, two scales, and the full dyadic set of four scales used in the main experiments. The results are given in Table 4 (the experiment was performed with the value $\lambda_{reg} = 10^{-3}$ that was in use before the final choice of the coefficient). In terms of accuracy, the configurations are statistically indistinguishable: the confidence intervals of all sets overlap. The two-scale set $\{0.02, 0.04\}$ exhibits signs of smoothing (the lowest $L_\infty = 0.799$ alongside the largest and most scattered residual, $9.61 \pm 5.57 \times 10^{-2}$), whereas the single scales deliver quality comparable to the full set. We conclude that in the one-dimensional problem at hand, with a single front, one correctly chosen scale is sufficient; the advantages of the multiscale construction are expected in problems with several characteristic spatial scales, while the full dyadic set requires no a priori knowledge of the solution scale.

B Additional Figures

Table 4: Ablation over the set of scales R at $\nu = 0.01/\pi$, $\lambda_{reg} = 10^{-3}$ (5 independent runs).

R	**Rel.** L_2	L_∞	**PDE Residual** ($\times 10^{-2}$)
$\{0.02\}$	0.168 ± 0.068	1.136 ± 0.452	6.15 ± 2.23
$\{0.16\}$	0.160 ± 0.056	1.099 ± 0.409	6.27 ± 1.51
$\{0.02, 0.04\}$	0.138 ± 0.051	0.799 ± 0.392	9.61 ± 5.57
$\{0.02, 0.04, 0.08, 0.16\}$	0.194 ± 0.087	1.207 ± 0.516	6.65 ± 0.63

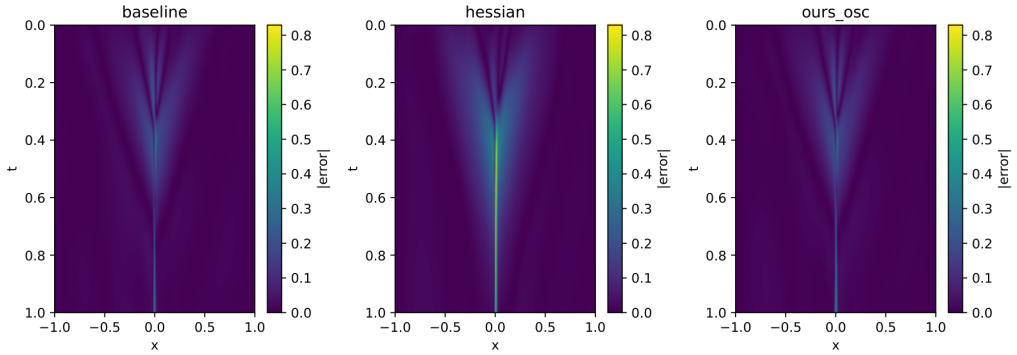


Figure 4: Absolute error $|u_\theta - u_{ref}|$ at $\nu = 0.01/\pi$ for the baseline PINN, the Hessian penalty, and the proposed method. The error is localized in the vicinity of the front.

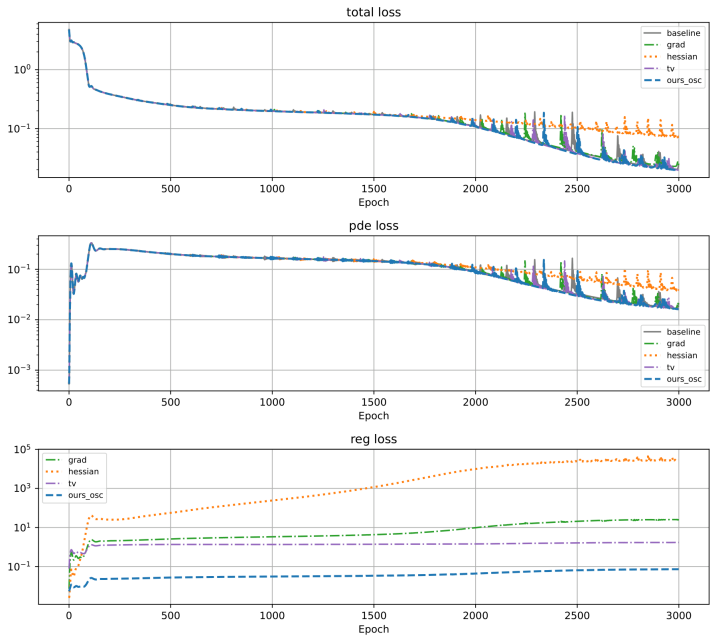


Figure 5: Convergence curves of the loss terms at $\nu = 0.01/\pi$ for all compared methods (logarithmic scale). Unweighted regularizer values are shown (before multiplication by λ_{reg}).

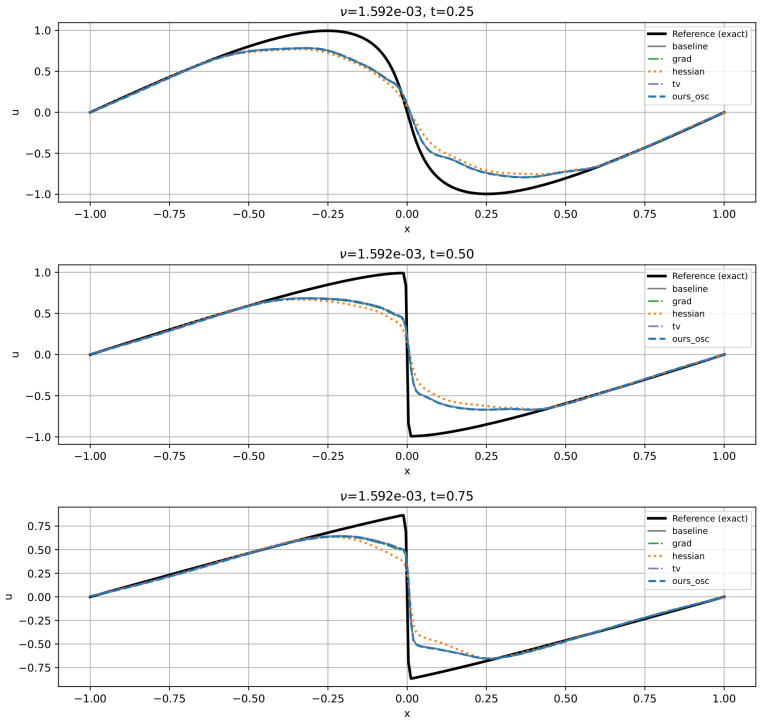


Figure 6: Solution slices at $\nu = 0.005/\pi$ (stiff regime). The basic MLP architecture, under the given training budget, fails to reproduce the extremely sharp front for any of the methods, illustrating the limitations discussed in Section 8.