

Multi-Preconditioned LBFGS for Training Finite-Basis PINNs

Marc Salvadó-Benasco^[0009-0001-8647-5159],
Aymane Kssim^[0009-0007-1814-655X],
Alexander Heinlein^[0000-0003-1578-8104],
Rolf Krause^[0000-0001-5408-5271],
Serge Gratton^[0000-0002-5021-2357],
Alena Kopaničáková^[0000-0001-8388-5518]

1 Introduction

Many scientific and engineering applications require the solution of partial differential equations (PDEs). Classical numerical methods, such as finite element (FE) discretizations, are accurate but can become computationally prohibitive for high-dimensional, multiscale, or data-driven problems. Neural network-based discretization methods, such as Physics-informed neural networks (PINNs) [12, 17], offer a mesh-free alternative that can naturally incorporate observational data by enforcing physical constraints through the training loss. This comes at the cost of a challenging, non-convex optimization problem for training the PINN. While standard PINNs struggle with multiscale or highly oscillatory solutions, finite-basis PINNs (FBPINNs) [15, 4] mitigate these difficulties through an additive, domain-decomposition (DD) inspired architecture defined on overlapping subdomains, in which collocation points are split and shared across neighboring subdomains.

The training of PINNs, including FBPINNs, has predominantly been carried out using stochastic gradient descent (SGD) and its variants. Alternatively, quasi-Newton

Marc Salvadó-Benasco
Università della Svizzera Italiana, Switzerland, e-mail: marc.salvado@usi.ch

Ayman Kssim
Toulouse INP-ENSEEIH, IRIT, ANITI, France, e-mail: aymane.kssim@toulouse-inp.fr

Alexander Heinlein
Delft Institute of Applied Mathematics, TU Delft, The Netherlands, e-mail: a.heinlein@tudelft.nl

Rolf Krause
King Abdullah University of Science and Technology, Saudi Arabia, e-mail: rolf.krause@kaust.edu.sa

Serge Gratton
Toulouse INP-ENSEEIH, IRIT, ANITI, France, e-mail: serge.gratton@toulouse-inp.fr

Alena Kopaničáková
Toulouse INP-ENSEEIH, IRIT, ANITI, France, e-mail: alena.kopanicakova@toulouse-inp.fr

methods, such as the limited-memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) algorithm [14], have been considered due to their ability to incorporate curvature information [9]. In FBPINNs, the additive structure of the model allows the forward and backward passes to be computed in parallel across subdomains. However, a parallel implementation of the LBFGS optimizer still requires global synchronization at each iteration, which can lead to communication-dominated execution and limit overall parallel scalability. In this work, we propose to address this difficulty by introducing a multi-preconditioned LBFGS (MP-LBFGS) method that enables multiple local optimization steps prior to global synchronization, thereby increasing local work while reducing the communication overhead.

The proposed MP-LBFGS algorithm is motivated by the nonlinear additive Schwarz method [2] and explicitly exploits the FBPINN decomposition of both the computational domain and the model parameters. The method performs a sequence of local LBFGS iterations on each subdomain, followed by a global LBFGS synchronization step. To aggregate the resulting local corrections, we introduce a subspace minimization strategy for scaling and combining subdomain updates. This is particularly essential in the FBPINN setting, where standard scaling strategies from classical DD may be ineffective due to fundamental structural differences between neural network models and FE discretizations. As demonstrated by our numerical experiments, this novel scaling mechanism enables stable aggregation of local updates while preserving the benefits of localized desynchronized optimization.

This manuscript is organized as follows. Sect. 2 introduces the FBPINN architecture, Sect. 3 presents the multi-preconditioned LBFGS algorithm, and numerical results and conclusions are given in Sect. 4 and Sect. 5, respectively.

2 Finite Basis PINNs (FBPINNs)

PINNs are neural network (NN) models that approximate solutions of differential equations. To this aim, let $\Omega \subset \mathbb{R}^d$ be a bounded domain with boundary $\partial\Omega$. We then consider the following abstract boundary value problem:

$$\begin{aligned} \mathcal{P}[u](\mathbf{x}) &= f(\mathbf{x}), & \forall \mathbf{x} \in \Omega, \\ u(\mathbf{x}) &= g(\mathbf{x}), & \forall \mathbf{x} \in \partial\Omega, \end{aligned} \tag{1}$$

where \mathcal{P} denotes a differential operator, and f is a forcing term. Note that other boundary conditions can be treated analogously.

We aim to approximate the unknown solution $u : \Omega \rightarrow \mathbb{R}$ using an NN model $\mathcal{N} : \mathbb{R}^p \times \Omega \rightarrow \mathbb{R}$, which is parameterized by a set of weights $\theta \in \mathbb{R}^p$, where p denotes the number of trainable parameters. To determine suitable network parameters, we sample a set of n collocation points $\mathcal{D} := \{\mathbf{x}_i\}_{i=1}^n \subset \Omega$ located in the interior of Ω . In addition, we consider a boundary dataset \mathcal{D}_{bc} , with n^{BC} collocation points sampled on the boundary $\partial\Omega$. The optimal network parameters are obtained by solving the following nonlinear minimization problem:

$$\boldsymbol{\theta}^* := \arg \min_{\boldsymbol{\theta}} \lambda_{\text{phy}} \mathcal{L}_{\text{phy}}(\boldsymbol{\theta}; \mathcal{D}) + \lambda_{\text{bc}} \mathcal{L}_{\text{bc}}(\boldsymbol{\theta}; \mathcal{D}_{\text{bc}}), \quad (2)$$

where $\lambda_{\text{phy}}, \lambda_{\text{bc}} \in \mathbb{R}^+$ are weighting parameters. The physics loss $\mathcal{L}_{\text{phy}} : \mathbb{R}^P \times \mathcal{D} \rightarrow \mathbb{R}$ quantifies the violation of the governing differential equation, i.e.,

$$\mathcal{L}_{\text{phy}}(\boldsymbol{\theta}; \mathcal{D}) := \frac{1}{n} \sum_{i=1}^n \left(\mathcal{P}[\hat{u}(\boldsymbol{\theta}; \mathbf{x}_i)] - f(\mathbf{x}_i) \right)^2,$$

where \hat{u} denotes the NN approximation of the solution. Similarly, each boundary loss $\mathcal{L}_{\text{bc}} : \mathbb{R}^P \times \mathcal{D}_{\text{bc}} \rightarrow \mathbb{R}$ enforces the boundary conditions and is defined as

$$\mathcal{L}_{\text{bc}}(\boldsymbol{\theta}; \mathcal{D}_{\text{bc}}) := \frac{1}{n_{\text{BC}}} \sum_j^{n_{\text{BC}}} \left(\hat{u}(\boldsymbol{\theta}; \mathbf{x}_j) - g(\mathbf{x}_j) \right)^2.$$

The weights λ_{phy} and λ_{bc} in (2) are used to balance the contributions of the different loss terms. In practice, however, it can be challenging to choose these parameters so that the boundary conditions (BCs) are enforced with sufficient accuracy while simultaneously ensuring a satisfactory decrease of the physics loss [12]. To overcome this difficulty, we enforce BCs by designing the solution ansatz such that they are automatically satisfied [12]. By choosing a function $\ell : \Omega \rightarrow \mathbb{R}$ which vanishes on the boundary $\partial\Omega$, we define the network solution approximation as

$$\hat{u}(\boldsymbol{\theta}; \mathbf{x}) := C(\mathbf{x}) + \ell(\mathbf{x}) \mathcal{N}(\boldsymbol{\theta}; \mathbf{x}), \quad (3)$$

where the function $C(\mathbf{x})$ encodes the prescribed BC through data g . With this construction, the BC are satisfied by design,

and (2) can be reformulated as

$$\boldsymbol{\theta}^* := \arg \min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}; \mathcal{D}) := \mathcal{L}_{\text{phy}}(\boldsymbol{\theta}; \mathcal{D}). \quad (4)$$

2.1 FBPINN architecture

Similarly to standard neural networks, PINNs also suffer from spectral bias [15], meaning that they tend to learn low-frequency components efficiently while struggling to capture high-frequency features. To address this limitation, a domain-decomposition-motivated architecture, termed the finite-basis PINN (FBPINN), was introduced in [15]. FBPINNs decompose the computational domain into several overlapping subdomains, within which high-frequency components are effectively rescaled to lower frequencies, thereby improving their resolution during training.

Let us decompose the domain Ω into n_s subdomains, such that $\bigcup_{j=1}^{n_s} \Omega_j = \Omega$, and the width of the overlap between neighboring subdomains is denoted by the parameter $\delta > 0$; see also [5, 15].

For each subdomain Ω_j , we introduce a dataset \mathcal{D}_j of collocation points, with $\mathcal{D} = \bigcup_{j=1}^{n_s} \mathcal{D}_j$.

For each subdomain Ω_j , we define a corresponding space of functions as

$$\mathcal{V}_j := \{\mathcal{N}_j \mid \mathcal{N}_j : \mathbb{R}^{P_j} \times \Omega \rightarrow \mathbb{R}\}, \quad (5)$$

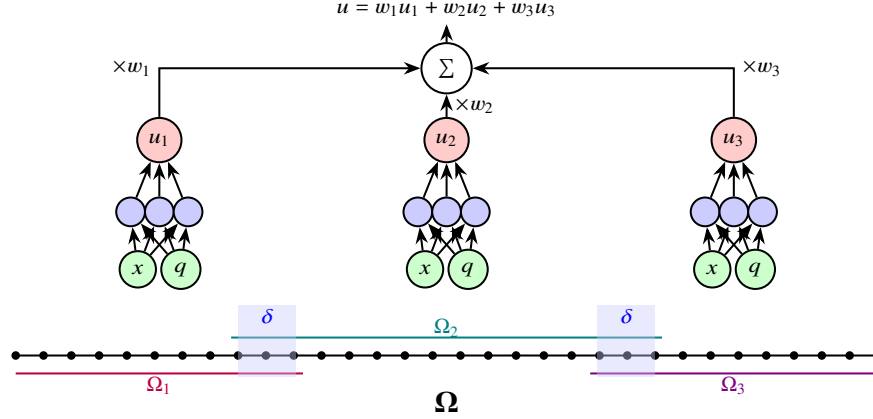


Fig. 1 FBPINN with collocation points decomposed into three overlapping subdomains. Each subdomain is associated with a subnetwork, and each subnetwork takes as input the spatial coordinates x and, possibly, additional physical features q .

where \mathcal{N}_j denotes a DNN associated with the j -th subdomain and parameterized by p_j trainable parameters. To restrict the support of each local network \mathcal{N}_j to its corresponding subdomain Ω_j and to appropriately weight the overlap between subdomains, we introduce a collection of “window” functions $\{w_j\}_{j=1}^{n_s} : \Omega \rightarrow \mathbb{R}$ such that $\text{supp}(w_j) \subset \Omega_j$ for all $j \in \{1, \dots, n_s\}$ and $\sum_{j=1}^{n_s} w_j \equiv 1$ on Ω ; this means that they form a partition of unity. Then, we define the global approximation space as $\mathcal{V} := \sum_{j=1}^{n_s} w_j \mathcal{V}_j$. The solution u is then approximated by a global network \mathcal{N} , obtained by combining the predictions of all subnetworks, i.e.,

$$\mathcal{N}(\boldsymbol{\theta}; \mathbf{x}) = \sum_{j=1}^{n_s} w_j(\mathbf{x}) \mathcal{N}_j(\boldsymbol{\theta}_j; \text{norm}_j(\mathbf{x})). \quad (6)$$

Here, the global parameter vector is given by $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{n_s})$. We use restriction operator $\mathbf{R}_j : \mathbb{R}^p \rightarrow \mathbb{R}^{p_j}$ to extract $\boldsymbol{\theta}_j$ from $\boldsymbol{\theta}$, i.e., $\boldsymbol{\theta}_j = \mathbf{R}_j \boldsymbol{\theta}$. Conversely, \mathbf{R}_j^\top is used to assign local parameters to the global network, i.e., $\boldsymbol{\theta} = \sum_j \mathbf{R}_j^\top \boldsymbol{\theta}_j$. Note, there is no overlap between the parameters of different subnetworks, see also Fig. 1. Furthermore, to mitigate the spectral bias, the normalization function $\text{norm}_j : \Omega \rightarrow (-1, 1)^d$ is used to rescale the input coordinates on Ω_j in order to map high-frequency features in the global domain to lower-frequency representations on each subdomain. The FBPINN is then trained by inserting (6) into the loss function (4); cf. [15, 4].

3 Multi-preconditioned LBFGS (MP-LBFGS)

Quasi-Newton (QN) methods are widely used for training PINNs, as they exploit curvature information without requiring the explicit computation or storage of the Hessian matrix. At each iteration k , the network parameters are updated as

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \alpha^{(k)} (\mathbf{B}^{(k)})^{-1} \nabla \mathcal{L}(\boldsymbol{\theta}^{(k)}), \quad (7)$$

where $\mathbf{B}^{(k)}$ is a low-rank Hessian approximation and $\alpha^{(k)}$ is obtained by line-search method with strong Wolfe's conditions [18].

Traditionally, approximation $\mathbf{B}^{(k)}$ is constructed to satisfy the secant equation:

$$\mathbf{B}^{(k+1)} \mathbf{s}^{(k)} = \mathbf{y}^{(k)},$$

where $\mathbf{s}^{(k)} := \boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(k)}$ and $\mathbf{y}^{(k)} := \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}^{(k+1)}) - \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}^{(k)})$. Since the secant equation alone does not uniquely determine $\mathbf{B}^{(k+1)}$, additional conditions must be imposed, leading to different variants of QN methods. In this work, we focus on the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method [16], as it is the most widely used in the context of PINNs [9].

In particular, we consider the limited-memory variant (LBFGS) [16], which approximates the Hessian using only the last Q secant pairs $(\mathbf{s}_{k-q}, \mathbf{y}_{k-q})_{q=1}^Q$. A standard compact representation of the LBFGS is given as

$$\mathbf{B}^{(k+1)} = \mathbf{B}^{(0)} - [\mathbf{B}^{(0)} \mathbf{S}^{(k)} \mathbf{Y}^{(k)}] \begin{bmatrix} (\mathbf{S}^{(k)})^{\top} \mathbf{B}^{(0)} \mathbf{S}^{(k)} & \mathbf{L}^{(k)} \\ (\mathbf{L}^{(k)})^{\top} & -\mathbf{D}^{(k)} \end{bmatrix}^{-1} \begin{bmatrix} (\mathbf{S}^{(k)})^{\top} \mathbf{B}^{(0)} \\ (\mathbf{Y}^{(k)})^{\top} \end{bmatrix},$$

where $\mathbf{S}^{(k)} := [\mathbf{s}^{(k-Q+1)}, \dots, \mathbf{s}^{(k)}]$ and $\mathbf{Y}^{(k)} := [\mathbf{y}^{(k-Q+1)}, \dots, \mathbf{y}^{(k)}]$, for $k > 1$. Note, that the secant pair $(\mathbf{s}^{(k)}, \mathbf{y}^{(k)})$ is added to the L-BFGS memory only if $(\mathbf{y}^{(k)})^{\top} \mathbf{s}^{(k)} > 0$. Otherwise, the pair is discarded to preserve positive definiteness of the Hessian approximation [16]. The symbols $\mathbf{L}^{(k)}$ and $\mathbf{D}^{(k)}$ denote the strictly lower triangular and diagonal parts of $(\mathbf{Y}^{(k)})^{\top} \mathbf{S}^{(k)}$, while $\mathbf{B}^{(0)}$ is an initial approximation, e.g., $\mathbf{B}^{(0)} = \gamma \mathbf{I}$, where $\gamma > 0$.

Note, to compute the search direction in (7), $(\mathbf{B}^{(k)})^{-1}$ is required. Rather than solving the associated linear system, the product $(\mathbf{B}^{(k)})^{-1} \nabla \mathcal{L}(\boldsymbol{\theta}^{(k)})$ can be computed efficiently using, for example, the well-known two-loop recursion algorithm [16].

3.1 Nonlinearly preconditioned LBFGS

Minimizing the loss \mathcal{L} leads to the following first-order optimality condition:

$$\nabla \mathcal{L}(\boldsymbol{\theta}; \mathcal{D}) = \mathbf{0}. \quad (8)$$

To accelerate convergence of LBFGS, we introduce a nonlinear preconditioning operator $P : \mathbb{R}^P \rightarrow \mathbb{R}^P$ that approximates the inverse of the gradient mapping, i.e., $P \approx \nabla \mathcal{L}^{-1}$.

While such an operator cannot be constructed explicitly, it can be defined implicitly via a nonlinear fixed-point map of the form $\boldsymbol{\theta} = P(\boldsymbol{\theta}; \mathcal{D})$.

Following a right-preconditioning strategy [10, 11], we reformulate the optimality system by introducing the preconditioned gradient as $\mathcal{F}(\boldsymbol{\theta}; \mathcal{D}) := \nabla \mathcal{L}(P(\boldsymbol{\theta}; \mathcal{D}); \mathcal{D})$. The original nonlinear problem (8) is thus replaced by $\mathcal{F}(\boldsymbol{\theta}; \mathcal{D}) = \mathbf{0}$, which we solve using an LBFGS method. Note, the preconditioner P induces a nonlinear change of coordinates in the parameter space, while the quasi-Newton method operates on \mathcal{F} .

Algorithm 1 Multi-Preconditioned LBFGS (MP-LBFGS)

Require: $\theta^{(0)} \in \mathbb{R}^P$, \mathcal{D} , $\mathcal{N} : \mathbb{R}^P \times \Omega \rightarrow \mathbb{R}$, $k_{\max} \in \mathbb{N}$

- 1: **for** $k = 0, \dots, k_{\max}$ **do**
- 2: **for** $j \in \{1, \dots, n_s\}$ **do** ▷ Parallel preconditioning step
- 3: $\theta_j^{(k+1/2)} \leftarrow \text{LBFGS}(\mathcal{N}_j, \mathbf{R}_j \theta^{(k)}, \eta)$ ▷ Perform η steps of local LBFGS
- 4: $\mathbf{c}_j \leftarrow \theta_j^{(k+1/2)} - \mathbf{R}_j \theta^{(k)}$
- 5: **end for**
- 6: $\theta^{(k+1/2)} \leftarrow \theta^{(k)} + \sum_{j=1}^{n_s} \beta_j^{(k)} \mathbf{R}_j^\top \mathbf{c}_j$ ▷ Compute $\beta^{(k)}$ using approaches from Sect. 3.2
- 7: $\theta^{(k+1)} = \text{LBFGS}(\mathcal{N}, \theta^{(k+1/2)}, 1)$ ▷ Perform one step of global LBFGS
- 8: **end for**

At the k -th iteration, the right-preconditioned LBFGS method proceeds in two steps. First, we perform a nonlinear change of variables, thus the current iterate $\theta^{(k)}$ is mapped through the nonlinear preconditioner, which defines an intermediate (pre-conditioned) iterate, i.e., $\theta^{(k+1/2)} = P(\theta^{(k)}; \mathcal{D})$. Second, a quasi-Newton correction is applied to the preconditioned gradient, yielding

$$\theta^{(k+1)} = \theta^{(k+1/2)} - \alpha^{(k)} (\mathbf{B}_{\mathcal{F}}^{(k)})^{-1} \mathcal{F}(\theta^{(k+1/2)}; \mathcal{D}). \quad (9)$$

Here, the update is taken around the preconditioned iterate $\theta^{(k+1/2)}$ and the LBFGS correction is applied in the preconditioned coordinates associated with \mathcal{F} . The approximate Hessian $\mathbf{B}_{\mathcal{F}}^{(k)}$ is constructed using the standard LBFGS update with secant pairs $\mathbf{s}^{(k)} = \theta^{(k+1)} - \theta^{(k+1/2)}$, and $\mathbf{y}^{(k)} = \mathcal{F}(\theta^{(k+1)}) - \mathcal{F}(\theta^{(k+1/2)})$.

To define the fixed-point operator P , we exploit the FBPINN structure, i.e.,

$$P(\theta^{(k)}; \mathcal{D}) = \theta^{(k)} + \beta^{(k)} \sum_{j=1}^{n_s} \mathbf{R}_j^\top (\theta_j^{(*)} - \mathbf{R}_j \theta^{(k)}), \quad (10)$$

where $\beta^{(k)} \in \mathbb{R}^+$ and $\theta_j^{(*)}$ is a solution of the local minimization problem

$$\theta_j^{(*)} = \arg \min_{\theta_j} \mathcal{L}_j(\theta_j; \mathcal{D}_j), \quad (11)$$

with the local loss function

$$\mathcal{L}_j(\theta_j; \mathcal{D}_j) := \frac{1}{|\mathcal{D}_j|} \sum_{\mathbf{x}_i \in \mathcal{D}_j} \left(\mathcal{P}[w_j(\mathbf{x}) \mathcal{N}_j(\theta_j; \text{norm}_j(\mathbf{x}_j))] - f(\mathbf{x}_i) \right)^2,$$

where w_j corresponds to hard enforcement of homogeneous Dirichlet boundary conditions on $\partial\Omega_j$. The set of interior collocation points is denoted by $\mathcal{D}_j \subset \Omega_j$.

Eq. (10) corresponds to a nonlinear additive Schwarz preconditioner in parameter space. In practice, the local problem (11) does not need to be solved exactly. Instead, an approximate solution suffices and can be obtained, for instance, by performing a fixed number η of L-BFGS iterations initialized at $\theta_j^{(k)} = \mathbf{R}_j \theta^{(k)}$. Here, we remark that local and global LBFGS maintain separate secant memories. The resulting right-preconditioned LBFGS algorithm is summarized in Alg. 1.

Table 1 Estimates of the parallel number of loss function evaluations ($\#\mathcal{L}_e$), gradient evaluations ($\#g_e$), update cost (UC), and memory cost (MC) per iteration.

| Method | $\#\mathcal{L}_e$ | $\#g_e$ | UC | MC |
|-----------------|---|--|---|--|
| LBFGS | $1 + \text{its}_{ls}$ | 1 | $p + 4Qp$ | $p + Qp$ |
| MP-LBFGS (UniS) | $1 + \text{its}_{ls} + \eta(\#\mathcal{L}_{e_j})$ | $2 + \eta(\#g_{e_j})$ | $2p + 4Qp + \eta\text{UC}_j$ | $2p + 2Qp + \text{MC}_j$ |
| MP-LBFGS (SPM) | $1 + \#\text{its}_{ls} + \eta\#\mathcal{L}_{e_j} + 2 + \#\text{its}_{\text{Newton}}(1 + \#\text{its}_{ls})$ | $2 + \eta\#g_{e_j} + 2 + \#\text{its}_{\text{Newton}}$ | $2p + 4Qp + \eta\text{UC}_j + 2\frac{p}{n_s} + \#\text{its}_{\text{Newton}}(\frac{p}{n_s} + \#\text{its}_{ls})$ | $2p + 2Qp + \text{MC}_j + 2p + 2n_s + n_s^2$ |

3.2 Scaling of Subdomain Corrections

The preconditioning step (10) yields a set of directions $\{\mathbf{c}_j\}_{j=1}^{n_s}$, where each $\mathbf{c}_j := \boldsymbol{\theta}_j^{(*)} - \mathbf{R}_j \boldsymbol{\theta}^{(k)}$. In this section, we discuss how to optimally update the global parameters using these search directions. Therefore, at each iteration k , we are looking for scaling parameters $\boldsymbol{\beta} = [\beta_1, \beta_2, \dots, \beta_{n_s}]^\top$, such that

$$\mathcal{L}\left(\boldsymbol{\theta} + \sum_{j=1}^{n_s} \beta_j \mathbf{R}_j^\top \mathbf{c}_j\right) < \mathcal{L}(\boldsymbol{\theta}). \quad (12)$$

We emphasize that the choice of the scaling parameters $\boldsymbol{\beta}$ is critical due to the intrinsic structure of FBPINNs and their differences from classical DD methods. In particular, classical DD methods, originally proposed for elliptic PDEs, exploit structural properties of finite-element/difference discretizations, such as locality of basis functions and sparse, spatially localized coupling between degrees of freedom. This ensures that subdomain corrections mainly affect local error components, while global modes are handled through coarse-grid or global synchronization steps.

In contrast, machine-learning problems involve nonconvex minimization, for which only a few domain-decomposition methods have been developed, e.g., [6]. Moreover, locally defined physical degrees of freedom are replaced by globally coupled parameters, fundamentally altering error-propagation mechanisms. As a result, classical DD assumptions, such as convexity, locality, and spectral separation, no longer hold, which often leads to degraded convergence of conventional DD methods.

For the FBPINNs considered in this work, spatial overlap induces indirect coupling, leading to desynchronization of parameters during the preconditioning step and motivating the design of novel scaling strategies for the subdomain corrections. In particular, we investigate three such strategies:

- **Uniform scaling (UniS):** As a first approach, we consider uniform scaling by a constant β_0 , i.e., $\beta_j = \beta_0$, for all $j = 1, \dots, n_s$. The value of β_0 is typically set to 1 in non-overlapping DD methods, but it can be fine-tuned. This approach has a low computational cost and is naturally parallel. However, it has been shown in [13] that it does not yield an efficient preconditioner for NN problems.
- **Multiplicative line search scaling (LSS):** Following [13], we also investigate a line search approach, where we start with a search direction \mathbf{c}_1 on the first subdomain, and look for the biggest value β_1 that satisfies $\mathcal{L}(\boldsymbol{\theta} + \beta_1 \mathbf{R}_1^\top \mathbf{c}_1) < \mathcal{L}(\boldsymbol{\theta})$. We then iterate through all search directions (subdomains) and seek their

corresponding scaling factors such that

$$\mathcal{L}\left(\boldsymbol{\theta} + \sum_{j=1}^{i-1} \beta_j \mathbf{R}_j^\top \mathbf{c}_j + \beta_i \mathbf{R}_i^\top \mathbf{c}_i\right) < \mathcal{L}\left(\boldsymbol{\theta} + \sum_{j=1}^{i-1} \beta_j \mathbf{R}_j^\top \mathbf{c}_j\right), \quad \forall i = 2, \dots, n_s. \quad (13)$$

In practice, to ensure a sufficient decrease of the loss, we employ a line search [18]. The LSS scheme introduces two challenges: first, its inherently sequential, and second, the scheme implicitly depends on the ordering of the search directions.

- **Subspace minimization (SPM):** Given the limitations of the previous approaches, we propose determining the subdomain scaling parameters $\boldsymbol{\beta}$ by solving a low-dimensional subspace minimization problem. Motivated by linear multi-preconditioning strategies [1], we define $\mathbf{C} := [\mathbf{R}_1^\top \mathbf{c}_1 \ \mathbf{R}_2^\top \mathbf{c}_2 \ \dots \ \mathbf{R}_{n_s}^\top \mathbf{c}_{n_s}]$. The scaling vector $\boldsymbol{\beta} \in \mathbb{R}^{n_s}$ can now be obtained by solving the subspace problem

$$\min_{\boldsymbol{\beta}} \phi(\boldsymbol{\beta}) := \mathcal{L}(\boldsymbol{\theta} + \mathbf{C}\boldsymbol{\beta}). \quad (14)$$

To solve (14), we perform a few iterations of a simplified Newton method [3], updating $\boldsymbol{\beta}$ at iteration m using the following update rule:

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} - t^{(m)} (\nabla^2 \phi(\boldsymbol{\beta}^{(0)}))^{-1} \nabla \phi(\boldsymbol{\beta}^{(m)}),$$

where $t^{(m)}$ is a damping parameter chosen by line search [18]. The gradient and Hessian of ϕ are given by $\nabla \phi(\boldsymbol{\beta}) = \mathbf{C}^\top \nabla \mathcal{L}(\boldsymbol{\theta} + \mathbf{C}\boldsymbol{\beta})$, and $\nabla^2 \phi(\boldsymbol{\beta}) = \mathbf{C}^\top \nabla^2 \mathcal{L}(\boldsymbol{\theta} + \mathbf{C}\boldsymbol{\beta}) \mathbf{C}$, respectively. To avoid explicit Hessian computations, we approximate the action of $\nabla^2 \mathcal{L}(\boldsymbol{\theta})$ on \mathbf{C} using finite differences and precompute $(\nabla^2 \phi(\boldsymbol{\beta}^{(0)}))^{-1}$ at $m = 0$. Since $\nabla^2 \phi$ has size $n_s \times n_s$, the resulting system remains inexpensive to solve. Moreover, although each Newton step requires multiple forward and backward evaluations, these operations are fully parallel across subdomains.

Tab. 1 compares the estimated parallel computational costs of LBFSGS and MP-LBFSGS. Here, the word parallel refers to the per-device computational cost assuming a parallel implementation of FBPINN (one subdomain, one subnetwork per device), LBFSGS, and MP-LBFSGS. Notably, communication cost is not included in this estimate. For LBFSGS, each epoch involves one forward and one backward pass to evaluate the gradient. In addition, the line search in (7) requires $\#\text{its}_{\text{ls}}$ forward evaluations. The computational cost of the LBFSGS update scales with the number of stored secant pairs Q as $2p + 4Qp$, while the memory requirements accounting for the parameters, secant pairs, and momentum terms scale as $2p + 2Qp$.

The cost of MP-LBFSGS depends on the choice of scaling strategy. In all cases, η forward and backward passes are required for each subnetwork, denoted by $\#\mathcal{L}e_j$ and $\#g_{e_j}$, respectively. Since the training of each subnetwork can be performed in parallel, the local L-BFGS update cost is denoted by UC_j , and the corresponding memory requirements by MC_j . The cost of global L-BFGS step corresponds to that of LBFSGS, with an addition of computing and storing the gradient at $\boldsymbol{\theta}^{(k+1/2)}$.

For SPM, two additional loss and gradient evaluations are needed to compute $\nabla^2 \mathcal{L}(\boldsymbol{\theta}) \mathbf{C}$. Moreover, at each Newton iteration, we require gradient evaluation to compute $\mathbf{C}^\top \nabla \mathcal{L}(\boldsymbol{\theta} + \mathbf{C}\boldsymbol{\beta})$ as well as $\#\text{its}_{\text{ls}}$ loss calls to find $t^{(m)}$. Storing the quantity $\nabla^2 \mathcal{L}(\boldsymbol{\theta}) \mathbf{C}$ requires storing p parameters per compute node. Each product

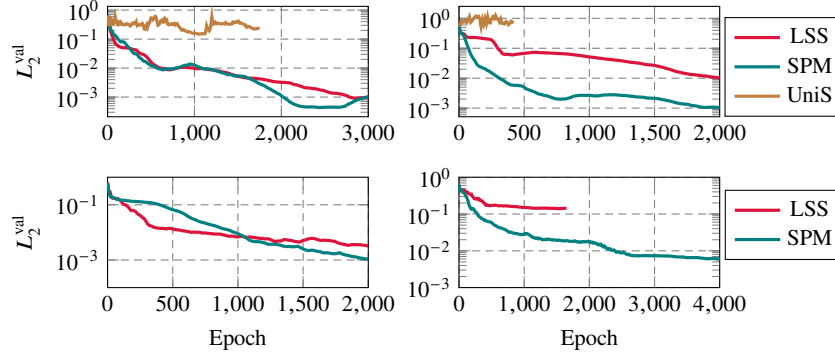


Fig. 2 Convergence of relative error L_2^{val} for Burger's equation. Problem is trained using MP-LBFGS with different subdomains scaling strategies. We consider 4×2 (top left), 4×4 (top right), 2×2 (bottom left), and 4×1 (bottom right) subdomains.

$\nabla^2 \mathcal{L}(\theta) \mathbf{R}_j^\top \mathbf{c}_j$ can be computed and stored locally, after which $\nabla^2 \phi$ is assembled by multiplying with \mathbf{C}^\top and concatenating the resulting columns. In addition, the vectors $\beta^{(m)}$ and Newton's search direction, as well as the Hessian matrix $\nabla^2 \phi \in \mathbb{R}^{n_s \times n_s}$, must be stored. The UC further includes cost associated with Hessian evaluation.

4 Numerical results

In this section, we investigate the numerical performance of our MP-LBFGS algorithm. To this aim, we consider two benchmark problems, namely:

- **Poisson's equation:**

We consider one- and two-dimensional Poisson equations with homogeneous Dirichlet boundary conditions, i.e.,

$$\begin{aligned} -\Delta u &= f, & \forall \mathbf{x} \in \Omega, \\ u &= 0, & \forall \mathbf{x} \in \partial\Omega. \end{aligned} \quad (15)$$

For $\Omega = (0, 1)^2$, f is chosen such that the exact solution is $u_{\text{true}}(\mathbf{x}) = \sin(4\pi x_1) \sin(4\pi x_2)$. For $\Omega = (0, 1)$, the exact solution is $u_{\text{true}}(\mathbf{x}) = \sin(20\pi x_1)$. Validation is performed using the relative L_2 error with respect to u_{true} .

- **Burgers' equation:** Let $\Omega = (0, 1) \times (-1, 1)$, we consider the Burgers' equation:

$$\begin{aligned} \frac{\partial u}{\partial t} + u \nabla u - \nu \nabla^2 u &= 0, & \forall (t, x) \in (0, 1] \times (-1, 1), \\ u(0, x) &= -\sin(\pi x) & \forall x \in [-1, 1], \\ u(t, 1) = u(t, -1) &= 0 & \forall t \in (0, 1], \end{aligned} \quad (16)$$

where $\nu = 0.01/\pi$ is the kinematic viscosity. The accuracy of the FBPINN solution is assessed by comparison with a finite element reference solution obtained on a mesh with 25,600 degrees of freedom, using the relative L_2 error as the metric.

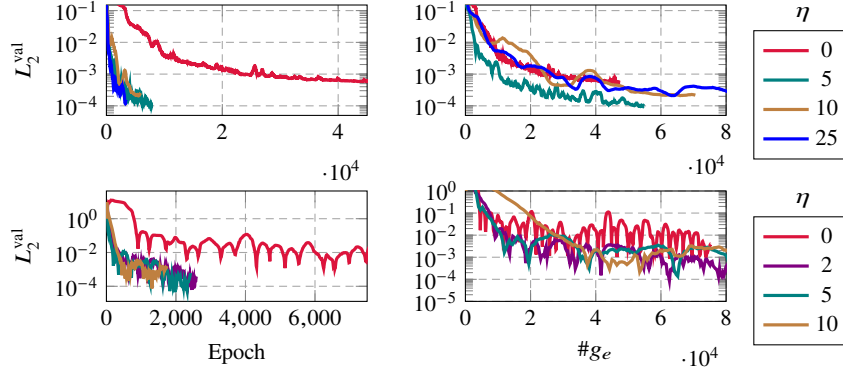


Fig. 3 Convergence of MP-LBFGS method with respect to varying number of local LBFGS iterations (η) for Burgers (top, 8 subdomains) and 1D Poisson (bottom, 20 subdomains), where the value $\eta = 0$ corresponds to standard LBFGS method.

All problems are posed on uniform domains, with 2D decompositions performed uniformly in both directions. The weighting functions are $w_j(\mathbf{x}) := \prod_{k=1}^d \left(1 + \cos(\pi \text{norm}_j(\mathbf{x})|_k)\right)^2$. Each subnetwork is a four-layer ResNet [8] of width 20 with tanh activation. The datasets consist of 20,000 collocation points for 2D problems and 3,000 for 1D, generated using Hammersley sampling [7].

4.1 Convergence properties of multi-preconditioned LBFGS

First, we investigate the impact of the scaling strategies discussed in Sect. 3.2 on the performance of the MP-LBFGS method. Fig. 2 shows that, independently of the chosen decomposition, the UniS strategy leads to unstable training. We further observe that as the number of subdomains increases (e.g., from 2×2 to 4×4), the SPM strategy significantly outperforms the LSS approach. This behavior can be attributed to the increasing distance between successive iterates after each line-search step. We also observe that, even for a fixed number of subdomains, the choice of decomposition strongly affects MP-LBFGS performance. For example, moving from a 2×2 to a 4×1 decomposition significantly degrades the LSS method, whereas the SPM strategy achieves an error reduction of more than one order of magnitude.

Next, we compare the performance of MP-LBFGS with standard LBFGS. Results are reported in terms of epochs and effective parallel gradient evaluations ($\#g_e$). The metric $\#g_e$ approximately reflects the per-device computational work, while excluding communication overhead, costs of parameter updates and loss evaluations (see Tab. 1). The number of epochs serves as a proxy for communication cost. Here, we also note that, for all presented experiments, the number of Newton iterations in the SPM method, denoted as $\#\text{its}_{\text{Newton}}$, is on average equal to two.

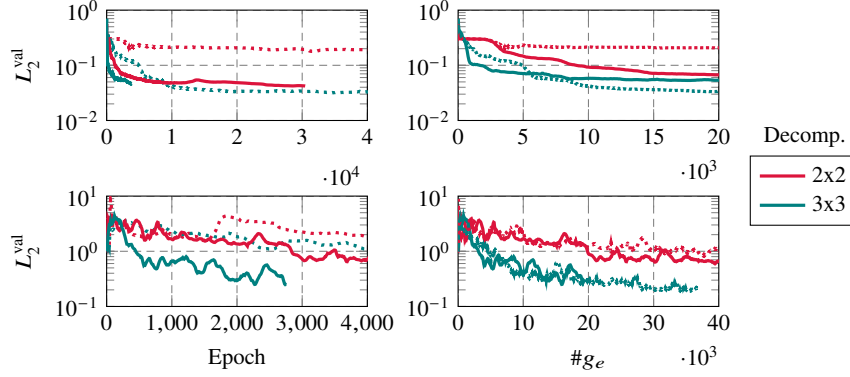


Fig. 4 Convergence of MP-LBFGS (solid lines, $\eta = 5$) and L-BFGS (dotted lines) for Burgers (top) and Poisson (bottom) problem. Experiments terminate after $\#g_e$ of 20,000 or 40,000 is reached.

We first examine the performance of MP-LBFGS while varying the number of local LBFGS iterations η , with $\eta = 0$ corresponding to standard LBFGS. As shown in Fig. 3, all MP-LBFGS configurations converge significantly faster in terms of epochs, thereby reducing communication cost. Moreover, MP-LBFGS requires a comparable, and in several cases lower, number of effective gradient evaluations ($\#g_e$). We also highlight, that for several values of η , MP-LBFGS achieves up to an order-of-magnitude reduction in the validation error L_2^{val} .

Second, we compare LBFGS and MP-LBFGS with $\eta = 5$ on the 2D Poisson and Burgers problems using 2×2 and 3×3 decompositions, while keeping the number of collocation points per subdomain fixed. As shown in Fig. 4, MP-LBFGS consistently reduces the number of epochs while requiring comparable $\#g_e$ and achieving comparable accuracy across different decompositions. Note that, here, one epoch of the MP-LBFGS algorithm with $\eta = 5$ corresponds to seven gradient evaluations; see Tab. 1 for details. The only exception is the Burgers problem with a 3×3 decomposition, where L-BFGS attains higher accuracy. Our empirical evidence suggests that, for this particular example, the non-convexity of the problem causes MP-LBFGS to converge to a local minimum with a larger error than L-BFGS.

5 Conclusion

In this work, we developed a multi-preconditioned LBFGS (MP-LBFGS) framework for training FBPINNs. The proposed approach exploits the FBPINN architecture to compute local LBFGS search directions in parallel, while a global right-preconditioned LBFGS iteration ensures consistent convergence of the global minimization problem. A key component of the method is a novel subspace minimization strategy (SPM) for scaling locally computed search directions. Numerical experiments suggest that the proposed MP-LBFGS can achieve faster convergence, and higher accuracy than the standard LBFGS, while maintaining lower communication

overhead. Future work will focus on parallel implementation, allowing for wall-clock comparison and the incorporation of more complex examples as well as on incorporation of coarse spaces; and on reducing the cost of solving the SPM subproblem.

Acknowledgements The work of A.Ks, S.G. and A.Ko. benefited from ANITI, funded by the France 2030 program under Grant Agreement No. ANR-23-IACL-0002. The numerical results were carried out using HPC resources from GENCI-IDRIS (Grant No. AD011015766R1).

References

1. Bridson, R., Greif, C.: A multipreconditioned conjugate gradient algorithm. *SIAM J. Matrix Anal. Appl.* **27**(4), 1056–1068 (2006)
2. Cai, X.C., Keyes, D.E.: Nonlinearly preconditioned inexact Newton algorithms. *SIAM J. Sci. Comput.* **24**(1), 183–200 (2002)
3. Deuffhard, P.: Newton methods for nonlinear problems: affine invariance and adaptive algorithms, vol. 35. Springer Science & Business Media (2011)
4. Dolean, V., Heinlein, A., Mishra, S., Moseley, B.: Multilevel domain decomposition-based architectures for physics-informed neural networks. *Comput. Methods Appl. Mech. Eng.* **429**, 117116 (2024)
5. Dolean, V., Jolivet, P., Nataf, F.: An introduction to domain decomposition methods: algorithms, theory, and parallel implementation. SIAM (2015)
6. Gratton, S., Kopaničáková, A., Toint, P.: Recursive bound-constrained adagrad with applications to multilevel and domain decomposition minimization. [arXiv:2507.11513](https://arxiv.org/abs/2507.11513) (2025)
7. Hammersley, J.M.: Monte Carlo methods for solving multivariable problems. *Ann. N. Y. Acad. Sci.* **86**(3), 844–874 (1960)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 770–778 (2016)
9. Kiyani, E., Shukla, K., Urbán, J.F., Darbon, J., Karniadakis, G.E.: Which optimizer works best for physics-informed neural networks and kolmogorov-arnold networks? (2025). [ArXiv:2501.16371](https://arxiv.org/abs/2501.16371) [cs.LG]
10. Kopaničáková, A., Kothari, H., Karniadakis, G.E., Krause, R.: Enhancing training of physics-informed neural networks using domain decomposition-based preconditioning strategies. *SIAM J. Sci. Comput.* **46**(5), S46–S67 (2024)
11. Kothari, H.: Nonlinear Schwarz preconditioning for quasi-Newton methods. In: International Conference on Domain Decomposition Methods, pp. 311–318. Springer (2022)
12. Lagaris, I.E., Likas, A., Fotiadis, D.I.: Artificial neural networks for solving ordinary and partial differential equations. *IEEE Trans. Neural Networks* **9**(5), 987–1000 (1998)
13. Lee, Y., Kopaničáková, A., Karniadakis, G.E.: Two-level overlapping additive Schwarz preconditioner for training scientific machine learning applications. *Comput. Methods Appl. Mech. Eng.* **448**, 118400 (2026)
14. Liu, D.C., Nocedal, J.: On the limited memory BFGS method for large scale optimization. *Math. Program.* **45**(1), 503–528 (1989)
15. Moseley, B., Markham, A., Nissen-Meyer, T.: Finite basis physics-informed neural networks (FBPINNs): a scalable domain decomposition approach for solving differential equations. *Adv. Comput. Math.* **49**(4), 62 (2023)
16. Nocedal, J.: Updating quasi-Newton matrices with limited storage. *Math. Comput.* **35**(151), 773–782 (1980)
17. Raissi, M., Perdikaris, P., Karniadakis, G.E.: Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* **378**, 686–707 (2019)
18. Wolfe, P.: Convergence conditions for ascent methods. *SIAM Rev.* **11**(2), 226–235 (1969)