A CONVERGENCE FRAMEWORK FOR ENERGY MINIMISATION OF LINEAR SELF-ADJOINT ELLIPTIC PDES IN NONLINEAR APPROXIMATION SPACES

ALEXANDRE MAGUERESSE^{†*} AND SANTIAGO BADIA[†]

ABSTRACT. Recent years have seen the emergence of nonlinear methods for solving partial differential equations (PDEs), such as physics-informed neural networks (PINNs). While these approaches often perform well in practice, their theoretical analysis remains limited, especially regarding convergence guarantees. This work develops a general optimisation framework for energy minimisation problems arising from linear self-adjoint elliptic PDEs, formulated over nonlinear but analytically tractable approximation spaces. The framework accommodates a natural split between linear and nonlinear parameters and supports hybrid optimisation strategies: linear variables are updated via linear solves or steepest descent, while nonlinear variables are handled using constrained projected descent. We establish both local and global convergence of the resulting algorithm under modular structural assumptions on the discrete energy functional, including differentiability, boundedness, regularity, and directional convexity. These assumptions are stated in an abstract form, allowing the framework to apply to a broad class of nonlinear approximation manifolds. In a companion paper [Magueresse, Badia (2025, arXiv:2508.17705)], we introduce a concrete instance of such a space based on overlapping free-knot tensor-product B-splines, which satisfies the required assumptions and enables geometrically adaptive solvers with rigorous convergence guarantees.

1. Introduction

Recent years have seen a growing interest in solving variational problems and partial differential equations (PDEs) using nonlinear approximation spaces, where tunable parameters control the basis functions, allowing the discretisation to adapt dynamically to the solution. Unlike classical linear methods, such as the finite element method (FEM), which relies on fixed basis functions, nonlinear approximation offers the potential to capture sharp gradients, layers, and singularities with significantly fewer degrees of freedom by concentrating resolution where needed. However, this flexibility comes at the cost of increased algorithmic and analytical complexity. The optimisation problems that arise are often non-convex, the approximation spaces may lack linear structure, and enforcing constraints, such as boundary conditions, mesh regularity, or boundedness, becomes nontrivial.

One of the most influential developments in this context is the rise of physics-informed neural networks (PINNs) [1]. By representing PDE solutions as neural networks trained to minimise residual-based losses, PINNs bypass mesh generation and operate on highly expressive, nonlinear function spaces. The Deep Ritz Method [2] exemplifies another branch of nonlinear approximation for variational problems, where neural networks parameterise trial functions minimising energy functionals directly. This framework has successfully tackled problems in diverse fields, yet it also exposes some of the central challenges of nonlinear approximation: expensive numerical integration, difficulties in enforcing boundary conditions, sensitivity to hyperparameters, and highly non-convex optimisation landscapes. Despite growing empirical evidence, rigorous convergence analyses for PINNs remain scarce and typically apply only in asymptotic or simplified regimes [3, 4]. This gap highlights a broader lack of theoretical foundations for nonlinear approximation methods in variational settings.

A related line of work explores free-knot B-splines, which introduce nonlinearity through adaptive knot placement while retaining the structure and locality of classical spline bases. Originating in one-dimensional data fitting [5–7], these methods have been extended to higher dimensions via tensor-product constructions [8–10], often guided by heuristic adaptivity or constrained optimisation to maintain mesh quality. The approximation properties of linear free-knot splines are well understood through their equivalence with ReLU neural networks [11, 12], which have been extensively analysed in terms of expressive power and optimal approximation rates [13–15]. However, the practical feasibility of achieving these rates—namely, how to compute or approximate the best representations numerically—remains poorly understood. Moreover, the use of free-knot spline spaces for the discretisation of PDEs has received little attention and remains an open area of study.

[†]SCHOOL OF MATHEMATICS, MONASH UNIVERSITY, CLAYTON, VICTORIA 3800, AUSTRALIA *E-mail addresses*: alexandre.magueresse@monash.edu, santiago.badia@monash.edu.

Date: August 27, 2025.

Key words and phrases. Energy minimisation, Nonlinear approximation, Nonlinear Céa's lemma.

^{*}Corresponding author.

In contrast, classical adaptive methods, such as h-, p-, and r-adaptivity, have long provided reliable strategies for controlling approximation error and distributing computational resources efficiently. These methods refine, enrich, or reposition discretisation elements based on error estimators, while maintaining robust mathematical properties like stability and convergence [16, 17]. Among them, r-adaptivity stands out for its conceptual proximity to nonlinear approximation: by relocating mesh nodes according to a parametric transformation [18, 19], it introduces a nonlinear dependence on discretisation parameters, echoing modern approaches based on parametric basis functions.

The convergence of first-order methods in unconstrained settings is by now classical. For smooth, non-convex problems, it is well-known that gradient descent converges to a quasi-stationary point in $O(\varepsilon^{-2})$ iterates, where ε is the target gradient norm. Global convergence guarantees can be obtained under some kind of convexity or gradient growth assumption, for example the Polyak–Łojasiewicz (PL) inequality guarantee global convergence to critical points and even linear rates, without requiring convexity [20]. Recent work has advanced the understanding of gradient descent dynamics for structured non-convex functionals, with a focus on characterising attraction regions under weak regularity conditions [21]. Their insights on basin geometry and convergence pathways directly inspire key aspects of our analysis in the nonlinear approximation setting.

In constrained optimisation, projected gradient descent and its extensions, such as mirror descent, are standard tools for handling inequality, geometric, or manifold constraints. However, most theoretical results for these methods assume convexity of the objective or the feasible set, or rely on strong regularity conditions [22]. For structured non-convex problems, particularly those arising from variational formulations of PDEs, such assumptions may fail, and convergence guarantees are much less mature. Designing and analysing algorithms that can robustly handle constraints in non-convex, parameter-dependent settings remains a central open challenge.

Contributions. We develop a general analytical framework for variational problems posed over nonlinear approximation spaces, aiming to address these theoretical gaps. Our setting is abstract and encompasses a wide class of parameter-dependent spaces, independent of any specific choice of basis functions or representation. Under minimal structural assumptions—uniform coercivity, differentiability, boundedness, and a directional convexity condition on the discrete energy functional—we establish both local and global convergence results for alternating minimisation schemes coupling linear and nonlinear parameters.

We propose a general two-step optimisation strategy: the linear parameters must satisfy a sufficient energy decrease condition, which, for instance, can be realised via an exact linear solve, an inexact conjugate gradient (CG) step, or a simple gradient update. The idea of eliminating linear parameters to simplify nonlinear optimisation problems is classical, tracing back to [23] in the context of least-squares curve fitting. To take constraints into account, the nonlinear parameters are updated via mirror descent, a generalisation of projected gradient descent, though our analysis can accommodate broader update rules. This leads to a nonlinear analogue of Céa's lemma, quantifying the gap between the iterates and a best approximation up to an optimisation error. Our results highlight the importance of structural properties in establishing convergence guarantees, an aspect often lacking in the analysis of PINNs and related methods. Beyond its theoretical interest, our framework provides algorithmic insights that can inform the design of robust numerical methods for constrained variational problems on nonlinear approximation spaces.

The theoretical framework we develop in this work has been largely motivated by our companion paper [24], where we apply it to tensor-product free-knot B-spline spaces. In that context, the interplay between the linear spline coefficients and the nonlinear knot positions naturally leads to alternating minimisation algorithms of the type studied here. Several modelling choices and algorithmic observations made in the companion paper have directly influenced the assumptions we adopt in this work. In particular, the geometric constraints on knot placement and the lack of global convexity prompted us to seek convergence guarantees under weaker conditions, such as directional convexity and boundedness of derivatives, rather than relying on stronger, but less applicable, global regularity or convexity assumptions. A key example is the differentiability of the discrete energy functional: while differentiability of the basis functions in the underlying Hilbert space is not strictly necessary, it simplifies the analysis and provides a practical guideline. We strike a balance by assuming differentiability of the energy while also giving a sufficient condition that is often easy to verify in applications. This analytical work thus provides a rigorous foundation for the algorithms explored in the companion study while extending beyond it to a more general class of nonlinear approximation spaces.

2. Abstract setting

2.1. Continuous problem. We consider a self-adjoint elliptic PDE posed in weak (variational) form, via the minimisation of an associated energy functional. Let $a: U \times U \to \mathbb{R}$ be a symmetric, coercive, and continuous

bilinear form, and let $\ell:U\to\mathbb{R}$ be a continuous linear form, where U is a suitable Hilbert space. We seek $u^\star\in U$ such that

$$a(u^*, v) = \ell(v), \quad \forall v \in U.$$

The well-posedness of this problem relies on the coercivity and continuity of the forms: for all $u, v \in U$,

$$a(u,u) \geq \alpha \|u\|_U^2, \qquad |a(u,v)| \leq \|a\|_{U \times U} \|u\|_U \|v\|_U, \qquad |\ell(v)| \leq \|\ell\|_U \|v\|_U,$$

for some coercivity constant $\alpha > 0$. The continuity constants $0 \le \|a\|_{U \times U}, \|\ell\|_U < \infty$ coincide with the operator norm of a and ℓ , defined respectively as

$$\|a\|_{U\times U} \doteq \sup_{u\in U} \sup_{v\in U} \frac{|a(u,v)|}{\|u\|_{U}\|v\|_{U}}, \qquad \|\ell\|_{U} \doteq \sup_{v\in U} \frac{|\ell(v)|}{\|v\|_{U}}.$$

The Lax-Milgram lemma then ensures the existence and uniqueness of the solution. Moreover, u^* is the minimiser of the energy functional $\mathcal{J}:U\to\mathbb{R}$ defined by

$$\mathcal{J}(u) \doteq \frac{1}{2}a(u,u) - \ell(u).$$

To quantify approximation quality, we introduce the energy norm $\|u\|_a^2 \doteq a(u,u)$, which reflects the natural topology induced by the variational problem. Although the exact energy $\mathcal{J}(u^*)$ is typically unknown, the identity

(1)
$$\mathcal{J}(u) - \mathcal{J}(u^*) = \frac{1}{2} \|u - u^*\|_a^2$$

for all $u \in U$ shows that, up to an additive constant, the energy functional provides a direct measure of proximity to the global minimiser in the energy norm. This observation motivates its use not only as an optimisation target but also as a surrogate error indicator, which is particularly relevant in adaptive settings where refinement is guided by energetic considerations.

While the proposed methodology applies broadly to such variational problems, we illustrate it using two model cases: a function approximation problem and a diffusion-reaction (Poisson-type) equation. We will use these examples to highlight regularity and differentiability aspects of the nonlinear approximation spaces.

- 2.1.1. Notations. Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with Lipschitz boundary $\Gamma = \partial \Omega$. Define the Sobolev spaces $H^0(\Omega) = L^2(\Omega)$, consisting of square-integrable functions on Ω ; $H^1(\Omega)$, consisting of functions in $L^2(\Omega)$ with square-integrable weak derivatives; and $H^1_0(\Omega) = \{v \in H^1(\Omega) : v|_{\Gamma} = 0\}$, consisting of functions in $H^1(\Omega)$ with zero trace on Γ . Let also $H^{-1}(\Omega)$ denote the (topological) dual of $H^1_0(\Omega)$.
- 2.1.2. Function approximation problem. Given $f \in L^2(\Omega)$, the function approximation problem consists of finding $u \in L^2(\Omega)$ such that u = f in Ω . Its variational formulation corresponds to

$$a(u,v) \doteq \int_{\Omega} uv \, d\Omega, \qquad \ell(v) \doteq \int_{\Omega} fv \, d\Omega.$$

2.1.3. Diffusion-reaction problem. In strong form, the diffusion-reaction problem is: find $u \in H^1(\Omega)$ such that

$$-\nabla \cdot (\mathbf{K}\nabla u) + \sigma u = f \text{ in } \Omega, \qquad u = g \text{ on } \Gamma,$$

where $K \in L^{\infty}(\Omega, \mathbb{R}^{d \times d})$ is a symmetric, positive-definite diffusivity matrix, $\sigma \in L^{\infty}(\Omega)$ with $\sigma \geq 0$ is the reaction coefficient, $f \in H^{-1}(\Omega)$ is a source term, and $g \in H^{1/2}(\Gamma)$ is the boundary condition. For simplicity, we consider only Dirichlet boundary conditions, but Neumann or Robin boundary conditions could also be treated. Let $\bar{u} \in H^1(\Omega)$ be a lifting of the Dirichlet boundary conditions such that $\bar{u} = g$ on Γ . The weak form of this problem corresponds to

$$a(u,v) \doteq \int_{\Omega} (\mathbf{K} \nabla u \cdot \nabla v + \sigma u v) \, d\Omega, \qquad \ell(v) \doteq \int_{\Omega} (f + \nabla \cdot (\mathbf{K} \nabla \bar{u}) - \sigma \bar{u}) v \, d\Omega,$$

where the integral $\int_{\Omega} fv \ d\Omega$ is understood as a duality pairing between $H^{-1}(\Omega)$ and $H^1_0(\Omega)$.

2.2. Energy minimisation in nonlinear approximation spaces. We aim to approximate the exact solution u^* in a finite-dimensional space $V \subset U$ by minimising the energy $\mathcal J$ over V. Unlike traditional methods where V is a fixed linear subspace, we assume V is a smoothly parameterised manifold of functions, defined as the image of a realisation map $\mathcal R:\Theta\to U$ from a finite-dimensional parameter space Θ . The corresponding discrete problem becomes the minimisation of the discrete energy functional

$$\mathcal{K} \doteq \mathcal{J} \circ \mathcal{R} : \Theta \to \mathbb{R}.$$

Not all approximation spaces are suitable for energy minimisation. To ensure that the problem is well-posed, amenable to gradient-based methods, and numerically tractable, we impose structural assumptions on the parameter space and the realisation map.

- Existence of minimisers: Any continuous, coercive, and lower-bounded function defined on a compact set admits global minimisers [25, Chapter 2]. Since \mathcal{J} satisfies these properties on U, they are automatically transferred to the restriction of \mathcal{J} to V, and it suffices to ensure that V is closed in U. Assuming that the realisation map is continuous, V is closed if Θ is compact.
- **Approximability:** The nonlinear space V must have strong approximation properties in U. Ideally, the decay rate of the approximation error with respect to the number of parameters should match or exceed that of classical schemes.
- **Differentiability and regularity:** Gradient-based optimisation requires that the discrete energy \mathcal{K} be differentiable, with a uniformly continuous gradient on Θ , ensuring convergence of standard optimisation algorithms [26, Chapter 1].
- Well-conditioned parameterisation: As pointed out in [27], another key factor in ensuring reliable convergence is the comparability of the function norm in U and the parameter norm in Θ . This requires the realisation map \mathcal{R} to be a uniformly continuous homeomorphism with a uniformly continuous inverse, thereby ensuring a well-conditioned parameterisation.
- **Computational feasibility:** Efficient and reliable evaluation of both \mathcal{K} and its gradient is crucial, especially when \mathcal{J} contains integrals approximated by quadrature. Controlling the accuracy of these approximations is essential to prevent artificial minima or spurious oscillations in the optimised solution [28].

We now examine in more detail the differentiability of the energy functional and the approximation properties of nonlinear spaces.

2.2.1. Differentiability of the discrete energy. The differentiability of the energy functional with respect to the nonlinear parameters is a delicate matter that cannot be answered by a naive application of the chain rule. While the realisation map \mathcal{R} is often weakly differentiable, it may fail to be differentiable into the function space U on which the energy \mathcal{J} is defined. As a result, the variations generated by the differential $D\mathcal{R}$ along parameter directions may not belong to the tangent space of U, making the composition $D\mathcal{J} \circ D\mathcal{R}$ ill-defined.

The core difficulty stems from the fact that differentiation with respect to parameters does not, in general, commute with spatial integration. This mismatch often manifests as a loss of regularity: the derivative of $\mathcal R$ may produce variations of lower smoothness than required for admissible test directions in the domain of $D\mathcal J$. Geometrically, this reflects a failure of the tangent space to the approximation manifold V to embed within the tangent space of the ambient function space U. The following example illustrates that the inclusion $D\mathcal R \subset U$ is not necessary for the differentiability, or even continuous differentiability, of the discrete energy.

Example. Let $\Omega \subset \mathbb{R}$ be a bounded interval. Given $I \subset \mathbb{R}$, let $\chi_I : \mathbb{R} \to \{0,1\}$ denote the indicator function of I. Consider the realisation

$$\mathcal{R}(\theta) = w_1 \chi_{(a,b)} + w_2 \chi_{(b,c)},$$

for $\theta = (w_1, w_2, a, b, c) \in \Theta$, where $\Theta \subset \mathbb{R}^5$ is the subset enforcing the constraints $a \leq b \leq c \in \Omega$. Since Ω is bounded, it is easy to see that $\mathcal{R}(\theta) \in L^2(\mathbb{R})$ for all $\theta \in \Theta$. Still, \mathcal{R} is not differentiable in $L^2(\mathbb{R})$, but only in $H^{-1}(\mathbb{R})$, with

$$D\mathcal{R}(\theta) = \chi_{(a,b)} \, dw_1 + \chi_{(b,c)} \, dw_2 - w_1 \delta_a \, da + (w_1 - w_2) \delta_b \, db + w_2 \delta_c \, dc.$$

Here $\delta_z \in H^{-1}(\mathbb{R})$ denotes the Dirac delta centred at $z \in \mathbb{R}$. For the function approximation problem, we compute

$$a(\mathcal{R}(\theta), \mathcal{R}(\theta)) = w_1^2(b-a) + w_2^2(c-b), \qquad \ell(\mathcal{R}(\theta)) = w_1 \int_{(a,b)\cap\Omega} f \, d\Omega + w_2 \int_{(b,c)\cap\Omega} f \, d\Omega.$$

In particular, $a(\mathcal{R}(\theta), \mathcal{R}(\theta))$ is infinitely differentiable in θ , and $\ell(\mathcal{R}(\theta))$ has the same regularity as f in Ω .

Whether or not this type of regularity issue arises depends on the properties of both the realisation map and the structure of the approximation space. As such, establishing the differentiability of the discrete energy must be addressed on a case-by-case basis.

- 2.2.2. Nonlinear approximation rates. The effectiveness of a nonlinear space V hinges on its approximation power. Ideally, when V belongs to a nested sequence of approximation spaces, the associated approximation error should decrease at least as rapidly as in standard methods. In standard discretisations such as finite elements, finite volumes or finite differences, convergence rates are expressed in terms of mesh size assuming quasi-uniformity. However, more general approximation spaces may lack a natural length scale, making it more appropriate to measure convergence relative to the number of parameters. In this context, classical convergence rates must be interpreted carefully, as they depend on the spatial dimension.
- 2.3. **Separation of linear and nonlinear parameters.** In many cases, it is natural for the realisation map to depend linearly on some parameters and nonlinearly on others. This separation gives rise to additional structure that can be leveraged in both the representation of the energy functional and the design of efficient optimisation algorithms, as we now describe.
- 2.3.1. Linear parameters and parametric basis functions. To formalise this, we assume without loss of generality that the parameter space Θ decomposes as a product $\mathbb{W} \times \mathbb{X}$, where $\mathbb{W} = \mathbb{R}^{n_L}$ for some $n_L \geq 1$ represents a set of linear parameters and $\mathbb{X} \subset \mathbb{R}^{n_{NL}}$, for some $n_{NL} \geq 0$, is a closed set of nonlinear parameters. The realisation map then takes the form

$$\mathcal{R}(oldsymbol{w},oldsymbol{\xi}) = \sum_{k=1}^{n_{
m L}} oldsymbol{w}_k oldsymbol{arphi}_k(oldsymbol{\xi}) = oldsymbol{w}^* oldsymbol{arphi}(oldsymbol{\xi}),$$

where each $\varphi_k : \mathbb{X} \to U$ is a parametric basis function, gathered in the vector-valued map $\varphi : \mathbb{X} \to U^{n_L}$. The vector $\mathbf{w} \in \mathbb{W}$ thus plays the role of degrees of freedom. Here $\mathbf{w}^* \varphi(\boldsymbol{\xi})$ denotes the inner product of two vectors.

This decomposition includes, as special cases, both purely linear and purely nonlinear parameterisations. Taking $n_{\rm NL}=0~(\mathbb{X}=\emptyset)$ recovers the classical setting of linear discretisation in a fixed basis. At the other extreme, a fully nonlinear model—where no linear parameter is separated—can be represented by setting $n_{\rm L}=1$ and constraining $\boldsymbol{w}_1=1$, so that the realisation becomes $\varphi_1(\boldsymbol{\xi})=\mathcal{R}(\boldsymbol{\xi})$.

When the approximation space separates linear parameters, it naturally acquires the structure of a vector bundle, in which each parameter corresponds to a vector space within the total space, known as its fiber. Fig. 1 illustrates the key components of this structure: the base space, the total space, individual fibers, and sections that select one representative vector from each fiber.

2.3.2. Parametric quadratic energy. This structure leads to a convenient form for the energy functional. Substituting the expression of \mathcal{R} into the definition of \mathcal{K} , and using the bilinearity of a and linearity of ℓ , we obtain

$$\mathcal{K}(\boldsymbol{w}, \boldsymbol{\xi}) = \frac{1}{2} \boldsymbol{w}^* \boldsymbol{A}(\boldsymbol{\xi}) \boldsymbol{w} - \boldsymbol{w}^* \boldsymbol{\ell}(\boldsymbol{\xi}),$$

where the stiffness matrix $A: \mathbb{X} \to \mathbb{R}^{n_L \times n_L}$ and the load vector $\ell: \mathbb{X} \to \mathbb{R}^{n_L}$ are defined componentwise by

$$A(\xi)_{ij} = a(\varphi_i(\xi), \varphi_i(\xi)), \qquad \ell(\xi)_j = \ell(\varphi_i(\xi)),$$

for all $i, j \in \{1 : n_L\}$, respectively. Here the notation $\{a : b\}$, for $a \le b \in \mathbb{N}$, refers to the set of integers $\{a, \ldots, b\}$. By symmetry and coercivity of the bilinear form a, the matrix $A(\boldsymbol{\xi})$ is symmetric and positive semi-definite for all $\boldsymbol{\xi} \in \mathbb{X}$. It may fail to be positive definite if the basis functions $(\varphi_k(\boldsymbol{\xi}))_{k \in \{1 : n_L\}}$ are linearly dependent

For a fixed nonlinear parameter ξ , the map $w \mapsto \mathcal{K}(w, \xi)$ is a convex quadratic function. More precisely, it is minimal when w solves the linear system

$$A(\xi)w = \ell(\xi).$$

This observation motivates treating the linear and nonlinear parameters differently in the optimisation process: the convexity in w can be exploited using solvers tailored to quadratic minimisation or even linear solvers to eliminate the linear parameters, while ξ may be updated using general-purpose nonlinear optimisation methods.

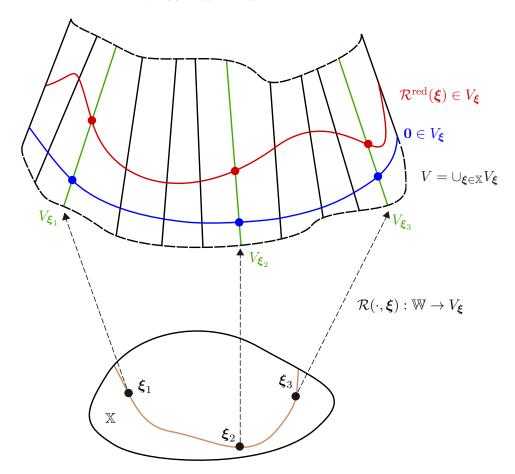


FIGURE 1. Visualisation of a parameter-dependent approximation space with vector bundle structure. The base space \mathbb{X} (nonlinear parameter space) is shown at the bottom. To each parameter $\boldsymbol{\xi} \in \mathbb{X}$ corresponds a fiber $V_{\boldsymbol{\xi}} = \operatorname{Span}(\varphi(\boldsymbol{\xi})) = \mathcal{R}(\mathbb{W}, \boldsymbol{\xi})$. The figure depicts the fibers lying above parameters along the brown path in \mathbb{X} , and highlights three fibers in green. Collectively, the fibers form the total space V, illustrating how the local linear spaces vary continuously with the parameters to assemble into a nonlinear approximation space. A section of this bundle is a map that assigns to every parameter $\boldsymbol{\xi}$ a vector in the corresponding fiber $V_{\boldsymbol{\xi}}$. Two sections are shown: $\mathcal{R}^{\mathrm{red}}$, in red, which selects the energy minimiser in each fiber (see (3)), and the zero section, in blue.

2.3.3. Best linear parameters and reduced energy. The following lemma shows that the linear system defining the best linear parameters remains consistent even when the basis functions are linearly dependent. Here, consistent means that the linear system admits at least one solution; equivalently, $\ell(\xi) \in \text{Im}(A(\xi))$.

Lemma 1. The linear system (2) is consistent for all $\boldsymbol{\xi} \in \mathbb{X}$. Moreover, any two solutions $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{W}$ define the same realisation in V; that is, $\mathcal{R}(\boldsymbol{w}_1, \boldsymbol{\xi}) = \mathcal{R}(\boldsymbol{w}_2, \boldsymbol{\xi})$.

In other words, Lemma 1 ensures that given $\xi \in \mathbb{X}$, the energy functional \mathcal{J} has a unique minimiser over the fibre $V_{\xi} \doteq \{\mathcal{R}(\boldsymbol{w}, \xi), \boldsymbol{w} \in \mathbb{W}\}$. The minimisation of \mathcal{J} is even well-posed in V_{ξ} , as

$$\alpha\|\mathcal{R}(\boldsymbol{w},\boldsymbol{\xi})\|_U^2 \leq a(\mathcal{R}(\boldsymbol{w},\boldsymbol{\xi}),\mathcal{R}(\boldsymbol{w},\boldsymbol{\xi})) = \ell(\mathcal{R}(\boldsymbol{w},\boldsymbol{\xi})) \leq \|\ell\|_U \|\mathcal{R}(\boldsymbol{w},\boldsymbol{\xi})\|_U,$$

so $\|\mathcal{R}(\boldsymbol{w}, \boldsymbol{\xi})\|_U \le \alpha^{-1} \|\ell\|_U$ for all \boldsymbol{w} solving the linear system (2).

The ambiguity arising from a nontrivial kernel of $A(\xi)$ can be resolved by selecting the unique solution of minimal Euclidean norm, which solves a convex quadratic problem: minimise the Euclidean norm $\|w\|_2$ over the affine solution space $w_p + \operatorname{Ker}(A(\xi))$, where w_p is any particular solution to the linear system (2). This problem has a unique solution, which corresponds to the Euclidean projection of w_p onto $\operatorname{Ker}(A(\xi))$. However, when $A(\xi)$ is ill-conditioned, this projection may be numerically unstable, as its sensitivity increases with the inverse of the smallest nonzero eigenvalue of $A(\xi)$; a quantity over which we have no direct control.

To circumvent this issue, we adopt a stronger assumption and require a uniform lower bound on the smallest eigenvalue of $A(\xi)$. For clarity, we separate the conditioning of the bilinear form a and that of the parametric basis $\varphi(\xi)$. To that aim, we introduce the Gram matrix $G(\xi) \in \mathbb{R}^{n_{\rm L} \times n_{\rm L}}$ with entries

$$G(\boldsymbol{\xi})_{ij} \doteq (\boldsymbol{\varphi}_i(\boldsymbol{\xi}), \boldsymbol{\varphi}_i(\boldsymbol{\xi}))_U,$$

for all $i, j \in \{1 : n_L\}$. This matrix satisfies $(\mathcal{R}(\boldsymbol{v}, \boldsymbol{\xi}), \mathcal{R}(\boldsymbol{w}, \boldsymbol{\xi}))_U = \boldsymbol{v}^* \boldsymbol{G}(\boldsymbol{\xi}) \boldsymbol{w}$ for all $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{W}$ and $\boldsymbol{\xi} \in \mathbb{X}$. Let $\omega(\boldsymbol{\xi})$ denote the smallest eigenvalue of $\boldsymbol{G}(\boldsymbol{\xi})$. Then, the coercivity of a implies

$$\|\boldsymbol{w}^*\boldsymbol{A}(\boldsymbol{\xi})\boldsymbol{w} = a(\mathcal{R}(\boldsymbol{w},\boldsymbol{\xi}),\mathcal{R}(\boldsymbol{w},\boldsymbol{\xi})) \ge \alpha \|\mathcal{R}(\boldsymbol{w},\boldsymbol{\xi})\|_U^2 = \alpha \boldsymbol{w}^*\boldsymbol{G}(\boldsymbol{\xi})\boldsymbol{w} \ge \alpha \omega(\boldsymbol{\xi}) \|\boldsymbol{w}\|_2^2,$$

showing that the smallest eigenvalue of $A(\xi)$ is greater than $\alpha\omega(\xi)$. Therefore, a uniform lower bound on $\omega(\xi)$ is sufficient to ensure the uniform positive definiteness of $A(\xi)$.

Assumption 1. There exists $\omega_{\min} > 0$ such that $\omega(\xi) \geq \omega_{\min}$ for all $\xi \in \mathbb{X}$.

Under Assumption 1, the linear system defining the best linear parameters can be solved uniquely at each nonlinear parameter, thereby defining the best linear parameters map $w^{\text{best}}: \mathbb{X} \to \mathbb{W}$ by

$$\boldsymbol{w}^{\mathrm{best}}(\boldsymbol{\xi}) = \boldsymbol{A}(\boldsymbol{\xi})^{-1} \boldsymbol{\ell}(\boldsymbol{\xi}).$$

This induces the reduced realisation

(3)
$$\mathcal{R}^{\text{red}}: \mathbb{X} \to V, \quad \boldsymbol{\xi} \mapsto \mathcal{R}(\boldsymbol{w}^{\text{best}}(\boldsymbol{\xi}), \boldsymbol{\xi}) = \boldsymbol{\ell}(\boldsymbol{\xi})^* \boldsymbol{A}(\boldsymbol{\xi})^{-1} \boldsymbol{\varphi}(\boldsymbol{\xi}),$$

and the associated reduced energy

$$\mathcal{K}^{\mathrm{red}}: \mathbb{X} o \mathbb{R}, \quad oldsymbol{\xi} \mapsto \mathcal{K}(oldsymbol{w}^{\mathrm{best}}(oldsymbol{\xi}), oldsymbol{\xi}) = -rac{1}{2}oldsymbol{\ell}(oldsymbol{\xi})^* oldsymbol{A}(oldsymbol{\xi})^{-1} oldsymbol{\ell}(oldsymbol{\xi}).$$

Thus, minimising K over $\mathbb{W} \times \mathbb{X}$ is equivalent to minimising K^{red} over \mathbb{X} .

The elimination the linear parameters from the optimisation problem reduces it to a purely nonlinear minimisation over $\mathbb X$. Importantly, the gradient of the reduced energy can be computed without differentiating the best linear parameter map $\boldsymbol{w}^{\text{best}}$. Let $\nabla_{\mathbb W}$ and $\nabla_{\mathbb X}$ denote the gradients with respect to the linear and nonlinear parameters, respectively. Then, by construction,

$$abla_{\mathbb{W}}\mathcal{K}(oldsymbol{w}^{ ext{best}}(oldsymbol{\xi}),oldsymbol{\xi})=\mathbf{0}.$$

Assuming that $\mathcal K$ is differentiable with respect to $\mathbb X$, the chain rule implies that $\mathcal K^{\mathrm{red}}$ is differentiable and

$$\begin{split} \nabla \mathcal{K}^{\text{red}}(\boldsymbol{\xi}) &= \nabla_{\mathbb{W}} \mathcal{K}(\boldsymbol{w}^{\text{best}}(\boldsymbol{\xi}), \boldsymbol{\xi}) \nabla_{\mathbb{X}} \boldsymbol{w}^{\text{best}}(\boldsymbol{\xi}) + \nabla_{\mathbb{X}} \mathcal{K}(\boldsymbol{w}, \boldsymbol{\xi})|_{\boldsymbol{w} = \boldsymbol{w}^{\text{best}}(\boldsymbol{\xi})} \\ &= \nabla_{\mathbb{X}} \mathcal{K}(\boldsymbol{w}, \boldsymbol{\xi})|_{\boldsymbol{w} = \boldsymbol{w}^{\text{best}}(\boldsymbol{\xi})}. \end{split}$$

The derivative of the map w^{best} does not appear in this expression. This simplification makes the reduced formulation especially attractive, as it enables a computationally efficient alternating optimisation strategy.

3. ALGORITHM

Building on the structure of the abstract framework, we introduce a hybrid optimisation algorithm tailored to approximation spaces with both linear and nonlinear parameters.

3.1. **Two-step algorithm.** We propose a minimisation algorithm that generates a sequence of parameters $(w_k, \xi_k)_{k\geq 0}$ by alternating updates of the linear and nonlinear parameters, leveraging convexity in w and accommodating the generally nonconvex nature in ξ . A generic version of the alternating scheme is summarised in Algorithm 1.

The algorithm begins by updating the linear parameters based on the initial nonlinear ones (line 1), ensuring that the linear variables are optimally (or approximately) set before any nonlinear updates. At each iteration, the nonlinear parameters are first updated using a general optimisation step (line 4). We present and analyse the mirror descent update (6), but other optimisers could be employed. After this, the linear parameters are updated based on the new nonlinear parameters (line 5). We consider two strategies for this step: either an exact solve of the linear system (4), or an approximate update that guarantees a decrease in energy (5). The algorithm concludes with a final linear solve (line 13), typically carried out with higher accuracy than the updates performed during the optimisation loop, to ensure that the linear parameters are close to optimal.

Owing to the conformity of the approximation, making use of (1), the discrete energy may be expressed as

$$\mathcal{K}(\boldsymbol{w}, \boldsymbol{\xi}) = \mathcal{J}(u^{\star}) + \frac{1}{2} \|\mathcal{R}(\boldsymbol{w}, \boldsymbol{\xi}) - u^{\star}\|_{a}^{2}.$$

Thus, minimising $\mathcal{K}(\boldsymbol{w}, \boldsymbol{\xi})$ is equivalent to minimising distance between the approximation $\mathcal{R}(\boldsymbol{w}, \boldsymbol{\xi})$ and the exact solution u^* in the energy norm. This interpretation motivates tracking the parameters $(\boldsymbol{w}_{\min}, \boldsymbol{\xi}_{\min})$ corresponding to the lowest energy observed during the optimisation process (lines 6–8), and returning the associated approximation $\mathcal{R}(\boldsymbol{w}_{\min}, \boldsymbol{\xi}_{\min})$ at the end of the training (line 14). This mechanism also offers worst-case bounds by guaranteeing that the algorithm produces a solution at least as good as the initial one.

The algorithm terminates after a fixed number of iteration, or when an early-stopping criterion is met (lines 9–11). Common criteria include the stabilisation of nonlinear parameters, indicated by

$$\|\boldsymbol{\xi}_{k+1} - \boldsymbol{\xi}_k\|_{\mathbb{X}} \le \varepsilon_{\mathbb{X}}$$

for some threshold $\varepsilon_{\mathbb{X}} > 0$, or the plateauing of the energy values, detected when

$$|\mathcal{K}(\boldsymbol{w}_{k+1}, \boldsymbol{\xi}_{k+1}) - \mathcal{K}(\boldsymbol{w}_k, \boldsymbol{\xi}_k)| \leq \varepsilon_{\mathcal{K}}$$

for some threshold $\varepsilon_{\mathcal{K}}>0.$ Here $\|\cdot\|_{\mathbb{X}}$ denotes a norm on $\mathbb{R}^{n_{\mathrm{NL}}}.$

As demonstrated in Subsection 4.1, the distance between successive parameters provides an upper bound on the norm of the gradient of the energy functional. This makes it a suitable criterion for identifying a quasi-minimiser. Other termination criteria could be used, such as requiring the slope of the energy values over a fixed window to fall below a prescribed threshold, indicating that further progress is negligible.

Algorithm 1 Two-step alternating minimisation of the energy functional. Given initial parameters $(\boldsymbol{w}_0, \boldsymbol{\xi}_0) \in \mathbb{W} \times \mathbb{X}$, number of epochs $E \geq 0$, linear and nonlinear update routines, step sizes $(\gamma_k)_{k \geq 0}$ and tolerances $\varepsilon_{\mathbb{X}}, \varepsilon_{\mathcal{K}} > 0$.

```
1: \boldsymbol{w}_0 \leftarrow \texttt{UpdateLinear}(\boldsymbol{w}_0, \boldsymbol{\xi}_0)
                                                                                                                                                                       ▶ Initial update of linear parameters
 2: (\boldsymbol{w}_{\min}, \boldsymbol{\xi}_{\min}, \mathcal{K}_{\min}) \leftarrow (\boldsymbol{w}_0, \boldsymbol{\xi}_0, \mathcal{K}(\boldsymbol{w}_0, \boldsymbol{\xi}_0))
  3: for k = 0, \dots, E - 1 do
                  \begin{aligned} & \pmb{\xi}_{k+1} \leftarrow \texttt{UpdateNonlinear}(\pmb{w}_k, \pmb{\xi}_k, \gamma_k) \\ & \pmb{w}_{k+1} \leftarrow \texttt{UpdateLinear}(\pmb{w}_k, \pmb{\xi}_{k+1}) \end{aligned} 
                                                                                                                                                                                   ▶ Update nonlinear parameters
  4:
  5:
                                                                                                                                                                                           if \mathcal{K}(\boldsymbol{w}_{k+1}, \boldsymbol{\xi}_{k+1}) < \mathcal{K}_{\min} then
  6:
                                                                                                                                                                                                  (oldsymbol{w}_{\min}, oldsymbol{\xi}_{\min}, \mathcal{K}_{\min}) \leftarrow (oldsymbol{w}_{k+1}, oldsymbol{\xi}_{k+1}, \mathcal{K}(oldsymbol{w}_{k+1}, oldsymbol{\xi}_{k+1}))
  7:
  8:
                 \text{if } \|\boldsymbol{\xi}_{k+1} - \boldsymbol{\xi}_k\|_{\mathbb{X}} \leq \varepsilon_{\mathbb{X}} \text{ or } \left|\mathcal{K}(\boldsymbol{w}_{k+1}, \boldsymbol{\xi}_{k+1}) - \mathcal{K}(\boldsymbol{w}_k, \boldsymbol{\xi}_k)\right| \leq \varepsilon_{\mathcal{K}} \text{ then}
                                                                                                                                                                                                         9:
10:
                 end if
11.
12: end for
                                                                                                                                                                         ⊳ Final update of linear parameters
13: oldsymbol{w}_{\min} \leftarrow 	exttt{UpdateLinear}(oldsymbol{w}_{\min}, oldsymbol{\xi}_{\min})
14: return \mathcal{R}(\boldsymbol{w}_{\min}, \boldsymbol{\xi}_{\min})
```

3.2. Update of the linear parameters. Since the matrix $A(\xi)$ is symmetric and positive definite under Assumption 1, the best linear parameters can be efficiently computed using the CG method [29]. A natural update for the linear component in Algorithm 1 is therefore

(4) UpdateLinear
$$(w, \xi) = A(\xi)^{-1} \ell(\xi)$$
,

with the inverse implicitly computed via CG. This is particularly advantageous when $A(\xi)$ is large and sparse. When a full linear solve is too expensive, it can be replaced with an inexact update. For instance, one may perform a fixed number of CG steps starting from the previous iterate w_k , or take a few gradient descent steps to reduce the energy. A simple instance of the latter is the steepest descent update:

(5) UpdateLinear(
$$\mathbf{w}, \boldsymbol{\xi}$$
) = $\mathbf{w} + \beta(\boldsymbol{\xi}) \mathbf{r}(\boldsymbol{\xi})$,

where $r(\xi) \doteq \ell(\xi) - A(\xi)w \in \mathbb{W}^*$ (the dual of \mathbb{W}) and the step size is given by

$$\beta(\xi) = \frac{(r(\xi), r(\xi))_2}{(r(\xi), A(\xi)r(\xi))_2} \ge 0.$$

Here $(\cdot,\cdot)_2$ denotes the Euclidean inner product in \mathbb{W} . Strictly speaking, the update $w+\beta(\xi)r(\xi)$ is ill-posed at the continuous level, since $r(\xi) \in \mathbb{W}^*$ and $w \in \mathbb{W}$. A proper formulation would require applying a Riesz isomorphism to map $r(\xi)$ back into the primal space. However, in this finite-dimensional setting, we identify $\mathbb{W} \cong \mathbb{W}^*$ via the Euclidean structure and take the Riesz map to be the identity.

3.3. **Update of the nonlinear parameters.** A variety of methods are available for updating the nonlinear parameters, including momentum-based schemes such as ADAM [30] or RMSProp, quasi-Newton methods like BFGS and SR1 [31, Chapter 6], and metric-aware approaches such as natural gradients [32, 33]. Each offers different trade-offs in terms of convergence speed, robustness, and computational cost. For clarity of presentation and ease of analysis, we focus here on mirror gradient descent, a flexible framework that naturally accommodates constraints and includes projected gradient descent as a special case.

In our setting, the parameter set \mathbb{X} is constrained, and the optimisation step must incorporate a projection back onto \mathbb{X} . Depending on the structure of the feasible set \mathbb{X} , computing the Euclidean projection onto \mathbb{X} required by projected gradient descent can be computationally demanding or even intractable. Mirror descent offers a principled generalisation by replacing the Euclidean distance in the proximal step with a Bregman divergence generated by a strictly convex distance-generating function [22].

Let $\psi: \mathbb{X} \to \mathbb{R}$ be a twice differentiable, μ -strongly convex function with respect to $\|\cdot\|_{\mathbb{X}}$, meaning that

$$\psi(\boldsymbol{\eta}) \ge \psi(\boldsymbol{\xi}) + \langle \nabla \psi(\boldsymbol{\xi}), \boldsymbol{\eta} - \boldsymbol{\xi} \rangle_{\mathbb{X}} + \frac{1}{2} \mu \| \boldsymbol{\eta} - \boldsymbol{\xi} \|_{\mathbb{X}}^{2},$$

for all $\xi, \eta \in \mathbb{X}$. Here $\langle \cdot, \cdot \rangle_{\mathbb{X}}$ denotes the duality pairing between \mathbb{X} (seen as a subset of $\mathbb{R}^{n_{\rm NL}}$) and the dual of $\mathbb{R}^{n_{\rm NL}}$. The associated *Bregman divergence* centred at $\xi \in \mathbb{X}$ is given by

$$D_{\psi}(\boldsymbol{\eta};\boldsymbol{\xi}) \doteq \psi(\boldsymbol{\eta}) - \psi(\boldsymbol{\xi}) - \langle \nabla \psi(\boldsymbol{\xi}), \boldsymbol{\eta} - \boldsymbol{\xi} \rangle_{\mathbb{X}}.$$

It can be shown that the function $D_{\psi}(\cdot; \boldsymbol{\xi})$ is nonnegative and μ -strongly convex for every fixed $\boldsymbol{\xi} \in \mathbb{X}$. The *proximal map* $\operatorname{Prox} : \mathbb{W} \times \mathbb{X} \times \mathbb{R}_+ \rightrightarrows \mathbb{X}$ is defined by

$$\operatorname{Prox}_{\psi}(\boldsymbol{w}, \boldsymbol{\xi}, \gamma) = \arg \min_{\boldsymbol{\eta} \in \mathbb{X}} \gamma \langle \nabla_{\mathbb{X}} \mathcal{K}(\boldsymbol{w}, \boldsymbol{\xi}), \boldsymbol{\eta} \rangle_{\mathbb{X}} + D_{\psi}(\boldsymbol{\eta}; \boldsymbol{\xi}).$$

The proximal map may be set-valued if \mathbb{X} is not convex. The *mirror descent update* is then obtained by choosing a suitable step size $\gamma > 0$ and setting

(6) UpdateNonlinear(
$$\boldsymbol{w}, \boldsymbol{\xi}, \gamma$$
) $\in \operatorname{Prox}_{\boldsymbol{\psi}}(\boldsymbol{w}, \boldsymbol{\xi}, \gamma)$.

This step balances descent in the direction $-\nabla_{\mathbb{X}}\mathcal{K}(\boldsymbol{w},\boldsymbol{\xi})$ with proximity to the current iterate $\boldsymbol{\xi}$, measured in the geometry induced by ψ .

The mirror descent update requires minimising a strongly convex function over the nonlinear parameter space. The existence readily follows from the closedness of \mathbb{X} , but the uniqueness requires that \mathbb{X} be closed. Altogether, we make the following assumption to ensure the existence of the mirror descent step. We also include the compactness of the parameter space to ensure the existence and numerical reachability of global minimisers, by continuity of the discrete energy and the extreme value theorem.

Assumption 2. The nonlinear parameter space X is convex and compact.

A notable special case of mirror descent is when $\psi(\boldsymbol{\xi}) = \|\boldsymbol{\xi}\|_{\mathbb{X}}^2/2$, which yields $D_{\psi}(\boldsymbol{\eta};\boldsymbol{\xi}) = \|\boldsymbol{\eta} - \boldsymbol{\xi}\|_{\mathbb{X}}^2/2$ and the proximal map $\operatorname{Prox}_{\psi}(\boldsymbol{w},\boldsymbol{\xi},\gamma)$ reduces to the standard Euclidean projection onto \mathbb{X} , recovering projected gradient descent.

4. Convergence analysis

We now analyse the convergence properties of the hybrid optimisation algorithm introduced above. Our results are divided into local and global components, depending on the strength of the structural assumptions placed on the reduced energy functional.

Local convergence is established under mild regularity conditions, such as Hölder continuity of the gradient, which are satisfied in many practical settings. Global convergence requires stronger assumptions, notably directional convexity of the reduced energy in a neighbourhood of the global minimisers. Under these conditions, we prove that the algorithm converges to a quasi-stationary point within the basin of attraction of a global minimum.

The resulting bounds yield a nonlinear analogue of Céa's lemma, showing that the computed solution approximates the best possible within the chosen approximation space, up to a quantifiable optimisation error.

To clarify the logical structure of our analysis and the dependencies between assumptions and results, we conclude this section with a diagram outlining the main implications, see Fig. 2. Each node represents a key assumption or result, and directed edges indicate logical dependencies. This visual summary provides a high-level overview of the proof strategy and serves as a reference for verifying which hypotheses are required for each convergence result.

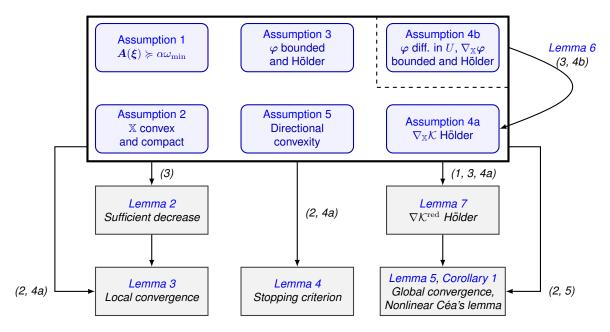


FIGURE 2. Logical dependency graph between assumptions and results in our abstract framework. Nodes represent assumptions (blue) and results (gray). In situations where Assumption 4b is not satisfied, Assumption 4a may be checked independently.

The diagram also highlights two layers of assumptions in our analysis. In practice, Assumption 1 and Assumption 3 are typically straightforward to verify, as they reflect standard properties of well-posed discretisations. Likewise, Assumption 2 holds for many parameter spaces. In contrast, Assumption 4b may fail in certain approximation spaces, especially when nonlinear parametrisations yield irregular basis functions. Even then, it is often possible to verify Assumption 4a directly. This regularity property is the key requirement for establishing the local convergence of our algorithm and can often be checked independently of the more structural assumptions. The example discussed in Subsection 2.2.1 illustrates such a situation.

- 4.1. **Local convergence of mirror descent.** The local convergence of Algorithm 1 relies on a guaranteed energy decrease at each step: for the linear parameters, this follows from classical properties of quadratic minimisation; for the nonlinear parameters, it is ensured by regularity of the energy gradient with respect to the nonlinear variables.
- 4.1.1. *Decrease condition for the linear parameters*. Both the full linear solve (4) and the steepest descent update (5) satisfy a decrease condition, as formulated in the following lemma.

Lemma 2. Let $w \in \mathbb{W}$ and $\xi \in \mathbb{X}$, and define $w_+ = UpdateLinear(w, \xi)$ via the update rule (4) or (5). It holds

(7)
$$\mathcal{K}(\boldsymbol{w}_{+},\boldsymbol{\xi}) \leq \mathcal{K}(\boldsymbol{w},\boldsymbol{\xi}) - \frac{1}{2}\lambda_{\max}(\boldsymbol{\xi})^{-1} \|\nabla_{\mathbb{W}}\mathcal{K}(\boldsymbol{w},\boldsymbol{\xi})\|_{2,*}^{2},$$

where $\lambda_{\max}(\mathbf{A}(\boldsymbol{\xi})) \leq \|a\|_{U \times U} \|\varphi(\boldsymbol{\xi})\|_{U,2}^2$ is the largest eigenvalue of $\mathbf{A}(\boldsymbol{\xi})$.

In the upper bound of the largest eigenvalue, the norm $\|\cdot\|_{2,*}$ denotes dual norm associated to $\|\cdot\|_2$, and $\|\cdot\|_{U,2}$ denotes the norm on U^{n_L} defined by

$$\|oldsymbol{arphi}(oldsymbol{\xi})\|_{U,2}^2 \doteq \sum_{i=1}^{n_{
m L}} \|oldsymbol{arphi}_i(oldsymbol{\xi})\|_U^2.$$

The local convergence of our algorithm will require the uniform boundedness of the constant in front of $\|\nabla_{\mathbb{W}}\mathcal{K}(\boldsymbol{w},\boldsymbol{\xi})\|_{2,*}^2$ in the energy decrease after the update of the linear parameters. In other words, it requires the uniform boundedness of $\|\varphi(\boldsymbol{\xi})\|_{U,2}$, as formalised in the following assumption. For later use, we also include the regularity of the basis functions in the same assumption.

Assumption 3. The basis functions are uniformly bounded and Hölder continuous in X; that is,

(1)
$$\|\varphi\|_{U,2,\infty} \doteq \sup_{\xi \in \mathbb{X}} \|\varphi(\xi)\|_{U,2} < \infty$$
,

(2) there exist
$$L_{\varphi} > 0$$
 and $\nu \in (0,1]$ such that $\|\varphi(\xi) - \varphi(\eta)\|_{U,2} \le L_{\varphi} \|\xi - \eta\|_{\mathbb{X}}^{\nu}$ for all $\xi, \eta \in \mathbb{X}$.

Under Assumption 3, the decrease condition (7) holds uniformly over \mathbb{X} with constant $\lambda_{\max}(A(\xi))^{-1} \geq \|a\|_{U\times U}^{-1}\|\varphi\|_{U,2,\infty}^{-2}$. This ensures that each update of the linear parameters yields an energy decrease uniformly proportional to the squared norm of the energy gradient.

Remark. Assumption 3 does not generally hold for neural networks with standard activation functions such as \tanh or ReLU. For example, the H^1 semi-norm of the function $x \mapsto \tanh(ax+b)$ on a bounded interval grows proportionally with $|a|^{1/2}$, so the assumption can only hold if the parameter a is uniformly bounded. This observation would motivate bounding the parameter space of a neural network.

Remark. More generally, the local convergence of Algorithm 1 can be established for any update rule of the linear parameters that satisfies the following energy decrease condition: for all $\xi \in \mathbb{X}$, there exists $0 < \beta(\xi) < \infty$ such that

(8)
$$\mathcal{K}(\boldsymbol{w}_{+},\boldsymbol{\xi}) \leq \mathcal{K}(\boldsymbol{w},\boldsymbol{\xi}) - \frac{1}{2}\beta(\boldsymbol{\xi})^{-1} \|\nabla_{\mathbb{W}}\mathcal{K}(\boldsymbol{w},\boldsymbol{\xi})\|_{2,*}^{2}$$

for all $w \in \mathbb{W}$ and $w_+ = \text{UpdateLinear}(w, \xi)$. This condition is stated in such a way that $\beta(\xi)$ has the same physical dimension as an eigenvalue of $A(\xi)$. A uniform lower bound for $\beta(\xi)$ is easily obtained from Assumption 1: the optimal update being $w^{\text{best}}(\xi)$, we obtain

$$\mathcal{K}(\boldsymbol{w}_{+},\boldsymbol{\xi}) - \mathcal{K}(\boldsymbol{w},\boldsymbol{\xi}) \geq \mathcal{K}(\boldsymbol{w}^{\text{best}}(\boldsymbol{\xi}),\boldsymbol{\xi}) - \mathcal{K}(\boldsymbol{w},\boldsymbol{\xi})$$

$$= -\frac{1}{2}(\boldsymbol{w}^{\text{best}}(\boldsymbol{\xi}) - \boldsymbol{w})^{*}\boldsymbol{A}(\boldsymbol{\xi})(\boldsymbol{w}^{\text{best}}(\boldsymbol{\xi}) - \boldsymbol{w})$$

$$= -\frac{1}{2}\nabla_{\mathbb{W}}\mathcal{K}(\boldsymbol{w},\boldsymbol{\xi})^{*}\boldsymbol{A}(\boldsymbol{\xi})^{-1}\nabla_{\mathbb{W}}\mathcal{K}(\boldsymbol{w},\boldsymbol{\xi})$$

$$\geq -\frac{1}{2}\lambda_{\min}(\boldsymbol{A}(\boldsymbol{\xi}))^{-1}\|\nabla_{\mathbb{W}}\mathcal{K}(\boldsymbol{w},\boldsymbol{\xi})\|_{2,*}^{2}.$$

Combining this fact with (8), we find $\beta(\xi) \geq \lambda_{\min}(A(\xi))$. Using Assumption 1, we reach $\beta(\xi) \geq \alpha \omega_{\min}$. As Lemma 2 showed, obtaining a uniform upper bound for $\beta(\xi)$ typically requires the uniform boundedness of $\varphi(\xi)$ in U.

4.1.2. Local convergence. We recall the convergence properties of mirror descent. Although the results are classical, we state them under slightly relaxed assumptions to account for the convexity of the energy functional in the linear parameters and for the possibility that the set of global minimisers has positive measure (it may not consist of isolated points). Our analysis builds on and extends the basin-based convergence framework developed in [21], which provides guarantees for gradient descent in non-convex optimisation. We adapt these ideas to the constrained setting, where the energy is minimised via mirror descent rather than standard gradient descent, by incorporating tools from constrained optimisation.

Standard analyses of gradient descent schemes typically require some form of gradient regularity, such as Lipschitz continuity [34, Chapter 1] or Hölder continuity [35]. Since the discrete energy is quadratic in the linear parameters, we only need to assume locally Hölder gradients with respect to the nonlinear parameters.

Assumption 4a. For all $\mathbf{w} \in \mathbb{W}$, the map $\mathbb{X} \ni \boldsymbol{\xi} \mapsto \mathcal{K}(\mathbf{w}, \boldsymbol{\xi})$ has Hölder continuous gradients; that is, there exists $\nu \in (0, 1]$ such that for each $\mathbf{w} \in \mathbb{W}$, there exists $L(\mathbf{w}) > 0$ satisfying

$$\|\nabla_{\mathbb{X}}\mathcal{K}(\boldsymbol{w},\boldsymbol{\xi}) - \nabla_{\mathbb{X}}\mathcal{K}(\boldsymbol{w},\boldsymbol{\eta})\|_{\mathbb{X},*} \leq L(\boldsymbol{w})\|\boldsymbol{\xi} - \boldsymbol{\eta}\|_{\mathbb{X}}^{\nu}$$

for all $\boldsymbol{\xi}, \boldsymbol{\eta} \in \mathbb{X}$.

Here $\|\cdot\|_{\mathbb{X}_{*}}$ denotes the dual norm induced by $\|\cdot\|_{\mathbb{X}}$.

Remark. More general moduli of gradient continuity could be considered, such as log-Lipschitz regularity, Dini-continuity or other concave modulus functions. We restrict our analysis to Hölder conditions for simplicity, which cover most cases of practical interest.

To preserve the clarity of the exposition, we defer the discussion of sufficient conditions for Assumption 4a to the end of the section. As shown in Lemma 6, the local Hölder constant L(w) needs to depend on w, unless w is uniformly bounded.

In particular, Assumption 4a implies that $\mathcal K$ is continuous. Since $\mathbb X$ is closed under Assumption 2 and $\mathbb W$ is closed, the product space $\mathbb W \times \mathbb X$ is also closed. Given that $\mathcal J$ is also coercive, the Weierstrass theorem

guarantees the existence of a global minimum of \mathcal{K} in $\mathbb{W} \times \mathbb{X}$ [25]. Let $\mathcal{K}^* \doteq \min_{\boldsymbol{w} \in \mathbb{W}, \boldsymbol{\xi} \in \mathbb{X}} \mathcal{K}(\boldsymbol{w}, \boldsymbol{\xi})$ denote the value of that minimum.

We now introduce common tools for constrained optimisation on convex sets. The *normal cone* to \mathbb{X} at $\boldsymbol{\xi} \in \mathbb{X}$ is defined as

$$N_{\mathbb{X}}(\boldsymbol{\xi}) \doteq \{ \boldsymbol{v} \in (\mathbb{R}^{n_{\mathrm{NL}}})^* : \forall \boldsymbol{\eta} \in \mathbb{X}, \langle \boldsymbol{v}, \boldsymbol{\eta} - \boldsymbol{\xi} \rangle_{\mathbb{X}} \leq 0 \}.$$

Normal cones are instrumental in expressing first-order necessary optimality conditions over convex sets: if $(\boldsymbol{w}, \boldsymbol{\xi}) \in \mathbb{W} \times \mathbb{X}$ is a local minimum of \mathcal{K} , then $\nabla_{\mathbb{W}} \mathcal{K}(\boldsymbol{w}, \boldsymbol{\xi}) = \mathbf{0}$ and $-\nabla_{\mathbb{X}} \mathcal{K}(\boldsymbol{w}, \boldsymbol{\xi}) \in N_{\mathbb{X}}(\boldsymbol{\xi})$. This condition, however, is not sufficient for local minimality, since it also characterises saddle points and local maxima. Nevertheless, since D_{ψ} is strongly convex, it does not have saddle point or local maxima so any $\boldsymbol{\xi}_+ \in \operatorname{Prox}_{\psi}(\boldsymbol{w}, \boldsymbol{\xi}, \gamma)$ is characterised by the first-order optimality condition

(9)
$$-\gamma \nabla_{\mathbb{X}} \mathcal{K}(\boldsymbol{w}, \boldsymbol{\xi}) - \nabla \psi(\boldsymbol{\xi}_{+}) + \nabla \psi(\boldsymbol{\xi}) \in N_{\mathbb{X}}(\boldsymbol{\xi}_{+}).$$

For convenience, we also define the gradient mapping $G_{\psi}: \mathbb{W} \times \mathbb{X} \times \mathbb{R}_+ \to (\mathbb{R}^{n_{\rm NL}})^*$ via

$$G_{\psi}(\boldsymbol{w},\boldsymbol{\xi},\gamma) \doteq \gamma^{-1}R_{\mathbb{X}}(\boldsymbol{\xi}-\texttt{UpdateNonlinear}(\boldsymbol{w},\boldsymbol{\xi},\gamma)),$$

where $R_{\mathbb{X}}: \mathbb{R}^{n_{\mathrm{NL}}} \to (\mathbb{R}^{n_{\mathrm{NL}}})^*$ is the Riesz map defined by $\langle R_{\mathbb{X}} \boldsymbol{\xi}, \boldsymbol{\eta} \rangle_{\mathbb{X}} = (\boldsymbol{\xi}, \boldsymbol{\eta})_{\mathbb{X}}$ for all $\boldsymbol{\xi}, \boldsymbol{\eta} \in \mathbb{R}^{n_{\mathrm{NL}}}$. In this way, one can write

$$\texttt{UpdateNonlinear}(\boldsymbol{w},\boldsymbol{\xi},\gamma) = \boldsymbol{\xi} - \gamma R_{\mathbb{X}}^{-1}G_{\psi}(\boldsymbol{w},\boldsymbol{\xi},\gamma),$$

where $R_{\mathbb{X}}^{-1}:(\mathbb{R}^{n_{\mathrm{NL}}})^*\to\mathbb{R}^{n_{\mathrm{NL}}}$ is the inverse Riesz map. This formulation provides a consistent way to interpret the gradient mapping as a dual vector.

In particular, in the case of projected gradient descent, when the unconstrained step $\boldsymbol{\xi} - \gamma R_{\mathbb{X}}^{-1} \nabla_{\mathbb{X}} \mathcal{K}(\boldsymbol{w}, \boldsymbol{\xi})$ remains inside \mathbb{X} , the proximal map $\operatorname{Prox}_{\psi}(\boldsymbol{w}, \boldsymbol{\xi}, \gamma)$ reduces to $\boldsymbol{\xi} - \gamma R_{\mathbb{X}}^{-1} \nabla_{\mathbb{X}} \mathcal{K}(\boldsymbol{w}, \boldsymbol{\xi})$, and $G_{\psi}(\boldsymbol{w}, \boldsymbol{\xi}, \gamma) = \nabla_{\mathbb{X}} \mathcal{K}(\boldsymbol{w}, \boldsymbol{\xi})$. Thus, G_{ψ} plays the role of an effective gradient, incorporating both the projection and step size effects.

We now show the local convergence of the mirror descent iterates to a quasi-minimiser.

Lemma 3 (Local convergence). Let $w, \xi \in \mathbb{W} \times \mathbb{X}$ and $(w_+, \xi_+) \in \mathbb{W} \times \mathbb{X}$ denote the next iterate produced by Algorithm 1 with step size $\gamma > 0$. Under Assumption 2 and Assumption 4a, for all $\varepsilon > 0$, or $\varepsilon = 0$ if $\nu = 1$, it holds

$$\mathcal{K}(\boldsymbol{w}_{+},\boldsymbol{\xi}_{+}) \leq \mathcal{K}(\boldsymbol{w},\boldsymbol{\xi}) - \frac{1}{2}\gamma(2\mu - \gamma L_{\nu,\varepsilon}(\boldsymbol{w})) \|G_{\psi}(\boldsymbol{w},\boldsymbol{\xi},\gamma)\|_{\mathbb{X},*}^{2} - \frac{1}{2}\beta(\boldsymbol{\xi})^{-1} \|\nabla_{\mathbb{W}}\mathcal{K}(\boldsymbol{w},\boldsymbol{\xi}_{+})\|_{\mathbb{W},*}^{2} + \varepsilon,$$

where
$$L_{\nu,\varepsilon}(\boldsymbol{w}) \doteq \left(\frac{1}{2\varepsilon} \frac{1-\nu}{1+\nu}\right)^{\frac{1-\nu}{1+\nu}} L(\boldsymbol{w})^{\frac{2}{1+\nu}}$$
 and $\beta(\boldsymbol{\xi}) \doteq \lambda_{\max}(\boldsymbol{A}(\boldsymbol{\xi}))$.

Let $(\mathbf{w}_0, \boldsymbol{\xi}_0) \in \mathring{\mathbb{W}} \times \mathbb{X}$ denote some initial parameters and $(\mathbf{w}_k, \boldsymbol{\xi}_k)_{k \geq 0}$ denote the sequence of parameters produced by Algorithm 1 with step sizes $\gamma_k > 0$. Under the same assumptions, for all $n \geq 1$, it holds

$$\min_{0 \leq k \leq n-1} \left\{ \left\| G_{\psi}(\boldsymbol{w}_{k}, \boldsymbol{\xi}_{k}, \gamma_{k}) \right\|_{\mathbb{X}, *}^{2} + \left\| \nabla_{\mathbb{W}} \mathcal{K}(\boldsymbol{w}_{k}, \boldsymbol{\xi}_{k+1}) \right\|_{\mathbb{W}, *}^{2} \right\} \leq \frac{2(\mathcal{K}(\boldsymbol{w}_{0}, \boldsymbol{\xi}_{0}) - \mathcal{K}^{*} + n\varepsilon)}{S_{n}},$$

where $S_n \doteq \sum_{k=0}^{n-1} \min(\gamma_k(2\mu - \gamma_k L_{\nu,\varepsilon}(\boldsymbol{w}_k)), \beta(\boldsymbol{\xi}_k)^{-1})$ provided $S_n > 0$.

In the case of Lipschitz gradients ($\nu=1$), taking $\varepsilon=0$ ($L_{\nu,\varepsilon}=L$) in Lemma 3 shows that under a suitable step size, the value of the energy functional is non-increasing along the trajectory of Algorithm 1. In the more general setting where the gradient is only Hölder continuous, the energy functional need not decrease monotonically during Algorithm 1. Under Assumption 3.1, we have $\beta(\xi) \leq \beta_{\max} \doteq \|a\|_{U \times U} \|\varphi\|_{U,2,\infty}$. Using a step size $\gamma_k = \zeta \mu / L_{\nu,\varepsilon}(\boldsymbol{w}_k)$ for some $\zeta \in (0,2)$ and assuming $L(\boldsymbol{w}) \leq L_{\max}$, Lemma 3 still provides

$$\min_{0 \leq k \leq n-1} \left\{ \|G_{\psi}(\boldsymbol{w}_{k}, \boldsymbol{\xi}_{k}, \gamma_{k})\|_{\mathbb{X}, *}^{2} + \|\nabla_{\mathbb{W}} \mathcal{K}(\boldsymbol{w}_{k}, \boldsymbol{\xi}_{k+1})\|_{\mathbb{W}, *}^{2} \right\} \leq \frac{2(\mathcal{K}(\boldsymbol{w}_{0}, \boldsymbol{\xi}_{0}) - \mathcal{K}^{*} + n\varepsilon)}{n \min(\mu^{2} \zeta(2 - \zeta) L_{\nu, \varepsilon, \max}^{-1}, \beta_{\max}^{-1})},$$

where $L_{\nu,\varepsilon,\max} = \left(\frac{1}{2\varepsilon}\frac{1-\nu}{1+\nu}\right)^{\frac{1-\nu}{1+\nu}} L_{\max}^{\frac{2}{1+\nu}}$. As $n \to \infty$, this upper bound converges to

$$\frac{2\left(\frac{1}{2}\frac{1-\nu}{1+\nu}\right)^{\frac{1-\nu}{1+\nu}}L_{\max}^{\frac{2}{1+\nu}}\varepsilon^{\frac{2\nu}{1+\nu}}}{\min\left(\mu^2\zeta(2-\zeta),\left(\frac{1}{2\varepsilon}\frac{1-\nu}{1+\nu}\right)^{\frac{1-\nu}{1+\nu}}L_{\max}^{\frac{2}{1+\nu}}\beta_{\max}^{-1}\right)},$$

where we expanded the expression of $L_{\nu,\varepsilon,\rm max}$ to show the dependence of this bound on ε . As $\varepsilon \to 0$, the denominator becomes $\mu^2 \zeta(2-\zeta)$, while the numerator goes to 0, so this upper bound can be made arbitrarily small by decreasing ε .

Importantly, the estimate of Lemma 3 shows that it is not necessary to solve the linear system exactly at each iteration for Algorithm 1 to converge. Any linear update that ensures the energy decrease condition in (8) suffices to guarantee convergence.

Remark. The proof of Lemma 3 fundamentally relies on the fact that a gradient descent step decreases the energy whenever the gradient is nonzero. More precisely, the mirror descent update (6) satisfies

$$\mathcal{K}(\boldsymbol{w},\boldsymbol{\xi}_{+}) \leq \mathcal{K}(\boldsymbol{w},\boldsymbol{\xi}) - \frac{1}{2}\gamma(2\mu - \gamma L_{\nu,\varepsilon}(\boldsymbol{w})) \|G_{\psi}(\boldsymbol{w},\boldsymbol{\xi},\gamma)\|_{\mathbb{X},*}^{2} + \varepsilon.$$

The convergence proof extends directly to any update rule for the nonlinear parameters that guarantees a similar energy decrease.

4.2. **Identification of quasi-stationary points.** In practical settings, reaching an exact stationary point is often too costly or unnecessary in light of numerical error and diminishing returns, and it is often more reasonable to look for a quasi-stationary point.

Given $\varepsilon > 0$, we say that $(\boldsymbol{w}, \boldsymbol{\xi}) \in \mathbb{W} \times \mathbb{X}$ is an ε -quasi-stationary point if $\|\nabla_{\mathbb{W}} \mathcal{K}(\boldsymbol{w}, \boldsymbol{\xi})\|_{2,*} \leq \varepsilon$ and $-\nabla_{\mathbb{X}} \mathcal{K}(\boldsymbol{w}, \boldsymbol{\xi}) \in N_{\mathbb{X}}(\boldsymbol{\xi}) + \overline{B}_{\mathbb{X},*}(0, \varepsilon)$, where

$$N_{\mathbb{X}}(\boldsymbol{\xi}) + \overline{B}_{\mathbb{X},*}(0,\varepsilon) \doteq \{\boldsymbol{v} \in (\mathbb{R}^p)^* \ : \ \exists \boldsymbol{w} \in N_{\mathbb{X}}(\boldsymbol{\xi}), \|\boldsymbol{v} - \boldsymbol{w}\|_{\mathbb{X},*} \leq \varepsilon\}$$

is an ε -neighbourhood of $N_{\mathbb{X}}(\xi)$. The following result shows that the gradient map acts as a surrogate for quasi-stationarity.

Lemma 4 (Surrogate for quasi-stationarity). Under Assumption 2, let $(\boldsymbol{w}, \boldsymbol{\xi}) \in \mathbb{W} \times \mathbb{X}$, and $\boldsymbol{\xi}_+$ denote the next iterate produced by Algorithm 1 with step $\gamma > 0$. Under Assumption 4a, $(\boldsymbol{w}, \boldsymbol{\xi}_+)$ is a $(L(\gamma c)^{\nu} + \mu c)$ -quasi stationary point, where $c^2 = \|G_{\psi}(\boldsymbol{w}, \boldsymbol{\xi}, \gamma)\|_{\mathbb{X}_*}^2 + \|\nabla_{\mathbb{W}} \mathcal{K}(\boldsymbol{w}, \boldsymbol{\xi}_+)\|_{\mathbb{W}_*}^2$.

In other words, Lemma 4 states that the condition

$$\|G_{\psi}(\boldsymbol{w}, \boldsymbol{\xi}, \gamma)\|_{\mathbb{X}_{*}}^{2} + \|\nabla_{\mathbb{W}}\mathcal{K}(\boldsymbol{w}, \boldsymbol{\xi}_{+})\|_{\mathbb{W}_{*}}^{2} \leq \varepsilon^{2}$$

can serve as a termination criterion to find a quasi-stationary point during the minimisation process. This termination criterion is equivalent to

$$\gamma^{-2} \|\boldsymbol{\xi}_{+} - \boldsymbol{\xi}\|_{\mathbb{X}}^{2} + \|\nabla_{\mathbb{W}} \mathcal{K}(\boldsymbol{w}, \boldsymbol{\xi}_{+})\|_{\mathbb{W}, *}^{2} \leq \varepsilon^{2}.$$

Since a final linear solve is performed at the end of the optimisation process (Algorithm 1, line 13), we do not need to check the stabilisation of the linear parameters. In that case, the stopping criterion becomes $\|\boldsymbol{\xi}_+ - \boldsymbol{\xi}\|_{\mathbb{X}} \leq \varepsilon \gamma$, as implemented in Algorithm 1 with $\varepsilon_{\mathbb{X}} = \varepsilon \gamma$.

We now combine Lemma 3 and Lemma 4 to estimate the number of steps to find a quasi-stationary point in the Lipschitz case ($\nu=1, \varepsilon=0$). Running Algorithm 1 with step size $\gamma_k=\zeta\mu/L_{\nu,\varepsilon}(\boldsymbol{w}_k)$ for some $\zeta\in(0,2)$ and assuming $L(\boldsymbol{w})\leq L_{\max}$ and $\beta(\boldsymbol{\xi})\leq\beta_{\max}$ (Assumption 3.1) will produce an iterate $(\boldsymbol{w},\boldsymbol{\xi})$ such that $\|G_{\psi}(\boldsymbol{w},\boldsymbol{\xi},\gamma)\|_{\mathbb{X},*}^2+\|\nabla_{\mathbb{W}}\mathcal{K}(\boldsymbol{w},\boldsymbol{\xi}_+)\|_{\mathbb{W},*}^2\leq c^2$ after at most $2(\mathcal{K}(\boldsymbol{w}_0,\boldsymbol{\xi}_0)-\mathcal{K}^\star)/[c^2\min(\mu^2\zeta(2-\zeta)L_{\max}^{-1},\beta_{\max}^{-1})]$ iterations. Given a tolerance $\tau>0$, choose c>0 such that $\gamma L_{\max}c+\mu c=\tau$. Then a τ -quasi-stationary point is guaranteed after at most

$$E(\tau) = \frac{2\mu^2 (1+\zeta)^2 (\mathcal{K}(\boldsymbol{w}_0, \boldsymbol{\xi}_0) - \mathcal{K}^*)}{\tau^2 \min(\mu^2 \zeta (2-\zeta) L_{\text{max}}^{-1}, \beta_{\text{max}}^{-1})}$$

iterations. This number of iterations is proportional to $2(\mathcal{K}(\boldsymbol{w}_0, \boldsymbol{\xi}_0) - \mathcal{K}^{\star})$, which is equal to $\|\mathcal{R}(\boldsymbol{w}_0, \boldsymbol{\xi}_0) - \mathcal{R}^{\star}\|_a^2$, the squared energy distance between the initial realisation and a global minimiser \mathcal{R}^{\star} of \mathcal{J} in V. The prefactor depending on ζ is minimal for the choice

$$\zeta = \begin{cases} 1 - \sqrt{1 - \frac{L_{\max}}{\mu^2 \beta_{\max}}} & \text{if } \frac{L_{\max}}{\mu^2 \beta_{\max}} \leq \frac{3}{4}, \\ \frac{1}{2} & \text{otherwise.} \end{cases}$$

Remark. Our analysis highlights the critical role of the Lipschitz constant of the gradient of the discrete energy functional, as it directly constrains the maximum step size in Algorithm 1. A large Lipschitz constant necessitates smaller steps, potentially slowing down progress towards the minimiser.

This limitation is structural to first-order methods based on fixed metrics. However, it may be alleviated through more advanced optimisation techniques that incorporate curvature information. For instance, natural gradient methods [32, 36] or its low-rank approximations [37] adapt the step direction using the inverse of the metric tensor induced by an inner product of the embedding space U (for example, the one induced by the bilinear form of the variational problem), while quasi-Newton approaches approximate second-order information to adjust step sizes dynamically. Such strategies could lead to stronger convergence guarantees under less restrictive conditions on the step size, especially in settings where the gradient of the discrete energy functional has a large Lipschitz constant.

4.3. **Global convergence of projected gradient descent.** We now turn to global convergence guarantees for Algorithm 1, focusing on the case where the reduced energy functional admits Lipschitz continuous gradients. This is the case whenever both Assumption 1, Assumption 3 and Assumption 4a hold. For clarity of exposition, we postpone the verification of this sufficient condition to the next subsection.

A notable difficulty in the analysis of the basins of attraction of discrete energies defined on nonlinear approximation spaces is related to the lack of injectivity of the realisation map. Indeed, there may exist $v \in V$ such that the inverse image $\mathcal{R}^{-1}(\{v\})$ is not a set of isolated points. Consequently, the set of global minimisers of \mathcal{K} may have a positive measure in $\mathbb{W} \times \mathbb{X}$, and \mathcal{K} may fail to be convex in a neighbourhood of one of its global minimisers.

We will prove the existence of basins of attraction for the reduced energy functional under the assumption of directional convexity in a neighbourhood of its global minimisers.

To define directional convexity, we introduce a few notations and definitions. Using Assumption 1, any pair $(\boldsymbol{w}, \boldsymbol{\xi}) \in \mathbb{W} \times \mathbb{X}$ satisfying $\mathcal{K}(\boldsymbol{w}, \boldsymbol{\xi}) = \mathcal{K}^{\star}$ must be such that $\boldsymbol{w} = \boldsymbol{w}^{\text{best}}(\boldsymbol{\xi})$. It follows that the reduced energy \mathcal{K}^{red} also attains its minimum on \mathbb{X} , and that this minimum agrees with \mathcal{K}^{\star} . We define the set of *best nonlinear parameters* as

$$\mathbb{X}^{\star} \doteq \{ \boldsymbol{\xi} \in \mathbb{X} : \mathcal{K}^{\mathrm{red}}(\boldsymbol{\xi}) = \mathcal{K}^{\star} \}.$$

the distance to the best nonlinear parameters,

$$\delta_{\psi}^{\star}: \mathbb{X} \to \mathbb{R}_{+}, \qquad \delta_{\psi}^{\star}(\boldsymbol{\xi}) \doteq \inf_{\boldsymbol{\xi}^{\star} \in \mathbb{X}^{\star}} D_{\psi}(\boldsymbol{\xi}^{\star}; \boldsymbol{\xi}),$$

and the set-valued *projection onto* \mathbb{X}^* ,

$$\Pi_{\psi}^{\star}: \mathbb{X} \rightrightarrows \mathbb{X}^{\star}, \qquad \Pi_{\psi}^{\star}(\boldsymbol{\xi}) \doteq \arg\inf_{\boldsymbol{\xi}^{\star} \in \mathbb{X}^{\star}} D_{\psi}(\boldsymbol{\xi}^{\star}; \boldsymbol{\xi}).$$

Directional convexity is defined as follows.

Definition 1 (Directional convexity). We say that $\mathcal{K}^{\mathrm{red}}$ is directionally convex in a subset $S \subset \mathbb{X}$ if for all $\boldsymbol{\xi} \in S$, there exists $\boldsymbol{\xi}^{\star} \in \Pi_{\psi}^{\star}(\boldsymbol{\xi})$ such that $\mathcal{K}^{\mathrm{red}}$ is convex on the segment $[\boldsymbol{\xi}, \boldsymbol{\xi}^{\star}]$.

Directional convexity is weaker than standard convexity. For example, the function $f: \mathbb{R}^2 \ni (x,y) \mapsto (x^2+y^2-1)^2$ reaches its minimum on the unit circle. For the generating function $\psi(x,y) = \|(x,y)\|_2^2/2$, one can check that $\Pi_{\psi}^{\star}(x,y) = (x,y)/\|(x,y)\|_2$, and f is directionally convex in the region $\{(x,y) \in \mathbb{R}^2, x^2+y^2 \geq 1/3\}$. However, f is convex only in the region $\{(x,y) \in \mathbb{R}^2, x^2+y^2 \geq 1\}$.

We state a global convergence result for the iterates of Algorithm 1 under the following directional convexity assumption.

Assumption 5. There exists $\rho > 0$ such that K^{red} is directionally convex in the neighbourhood

$$\Lambda_{\psi}^{\star}(\rho) \doteq \{ \boldsymbol{\xi} \in \mathbb{X} : \delta_{\psi}^{\star}(\boldsymbol{\xi}) \leq \rho \}.$$

Noticing that $D_{\psi}^{\star}(\boldsymbol{\xi})=0$ if and only if $\boldsymbol{\xi}\in\mathbb{X}^{\star}$, the neighbourhood $\Lambda_{\psi}^{\star}(\rho)$ corresponds to the subset of nonlinear parameters that are at distance at most ρ from \mathbb{X}^{\star} .

Lemma 5 (Global convergence). Under Assumption 2, Assumption 5 and assuming that K^{red} is Lipschitz continuous with constant $L^{\text{red}} > 0$, let $\boldsymbol{\xi}_0 \in \Lambda_{\psi}^{\star}(\rho)$ and define the sequence $(\boldsymbol{\xi}_k)_{k \geq 0}$ by running Algorithm 1 with exact updates of the linear parameters, and step sizes $(\gamma_k)_{k \geq 0}$ satisfying $\gamma_k \leq \mu/L^{\text{red}}$ for all $k \geq 0$. Then

- The distance to the best nonlinear parameters is non-increasing: $\delta_{\psi}^{\star}(\boldsymbol{\xi}_{k+1}) \leq \delta_{\psi}^{\star}(\boldsymbol{\xi}_{k})$ for all $k \geq 0$ so that $\boldsymbol{\xi}_{k} \in \Lambda_{\psi}^{\star}(\delta_{\psi}^{\star}(\boldsymbol{\xi}_{0}))$ for all $k \geq 0$, and
- The reduced energy satisfies the bound

$$\mathcal{K}^{\mathrm{red}}(\boldsymbol{\xi}_n) \leq \mathcal{K}^{\star} + \frac{\delta_{\psi}^{\star}(\boldsymbol{\xi}_0)}{\sum_{k=0}^{n-1} \gamma_k}$$

for all n > 1.

Proof. See Appendix A.6.

Under suitable assumptions on the energy functional, Lemma 5 guarantees the stability of mirror descent iterates within a neighbourhood of the global minimisers. In particular, the distance to the best nonlinear parameters decreases monotonically along the trajectory, and the reduced energy converges to the global minimum. These properties define a *basin of attraction* for the global minimisers under the mirror descent dynamics.

Remark. Our proof of Lemma 5 remains valid for more general star-shaped neighbourhoods S of \mathbb{X}^* satisfying the following property: if $\boldsymbol{\xi} \in S$ and $\boldsymbol{\eta} \in \mathbb{X}$ is such that $\delta_{\psi}^*(\boldsymbol{\eta}) \leq \delta_{\psi}^*(\boldsymbol{\xi})$, then $\boldsymbol{\eta} \in S$. In particular, S can be of the type $S = \{\boldsymbol{\xi} \in \mathbb{X} : \exists \boldsymbol{\xi}^* \in \mathbb{X}^*, D_{\psi}(\boldsymbol{\xi}^*; \boldsymbol{\xi}) \leq \rho(\boldsymbol{\xi}^*)\}$, where $\boldsymbol{\xi}^* \mapsto \rho(\boldsymbol{\xi}^*) > 0$ is some parameter-dependent radius.

Remark. Various generalisations of convexity can be used to derive global convergence guarantees for mirror descent. We refer to [20] for a comprehensive discussion and precise definitions of conditions such as (weak) strong convexity, the restricted secant inequality, the Polyak–Łojasiewicz (PL) inequality, and quadratic growth. Among these, the PL condition is particularly appealing in practice, as it guarantees global convergence without requiring convexity, by ensuring that the gradient norm controls the suboptimality; that is, the difference between the current energy value and the global minimum. In this work, we focus on directional convexity as a structural assumption tailored to our problem setting, but similar arguments could be developed under other gradient growth conditions like PL.

Remark. Even when the initialisation lies outside the basin of attraction, global convergence may still be achieved if the iterates eventually enter the basin. Once this occurs, the stability properties of the basin guarantee that all subsequent iterates remain within it, leading to global convergence. For instance, it is shown in [38] that under the assumption of a Lipschitz-continuous Hessian, a perturbed gradient descent algorithm can escape strict saddle points and enter a basin of attraction within $O(\varepsilon^{-2})$ iterations, matching the order of complexity we derived above.

In our case, the energy functional is directly related to the distance to the global minimiser u^* of \mathcal{J} (in the energy norm) via the relation

$$\mathcal{K}^{\mathrm{red}}(\boldsymbol{\xi}) = \mathcal{J}(\mathcal{R}^{\mathrm{red}}(\boldsymbol{\xi})) = \mathcal{J}(u^{\star}) + \frac{1}{2} \|\mathcal{R}^{\mathrm{red}}(\boldsymbol{\xi}) - u^{\star}\|_{a}^{2}.$$

The following corollary is a direct consequence of Lemma 5.

Corollary 1. With the same assumptions and notations as in Lemma 5, choosing step sizes $\gamma_k = \zeta \mu / L^{\rm red}$ for some $\zeta \in (0,1]$, it holds

$$\left\|\mathcal{R}^{\mathrm{red}}(\boldsymbol{\xi}_n) - u^{\star}\right\|_a^2 \le \inf_{v \in V} \left\|v - u^{\star}\right\|_a^2 + \frac{2L^{\mathrm{red}}\delta_{\psi}^{\star}(\boldsymbol{\xi}_0)}{\zeta \mu n},$$

for all $n \geq 1$.

Proof. See Appendix A.7.

Corollary 1 can be understood as a nonlinear version of Céa's lemma: $\mathcal{R}^{\mathrm{red}}(\boldsymbol{\xi}_n)$ approaches an optimal solution in the approximation space V at a rate of $O(n^{-1/2})$, where n denotes the iteration count. Crucially, this provides a quantitative and rigorous guarantee that the discrete solution generated by Algorithm 1 approaches the optimal energy-minimising solution within the nonlinear approximation space.

- 4.4. **Sufficient conditions for convergence.** To simplify the application of our framework, we provide sufficient conditions for the key assumptions, formulated in terms of boundedness and regularity properties of the parametric basis functions.
- 4.4.1. Regularity of the energy. We show the regularity of the gradient of the discrete energy with respect to \mathbb{W} and \mathbb{X} when the realisation is differentiable in U, as formalised in the following assumption. For convenience, we include the boundedness and regularity of the gradient of the realisation in this assumption. We equip $U^{n_{\rm NL}\times n_{\rm L}}$ with the norm $\|\boldsymbol{U}\|_{U,2,2}^2 \doteq \sum_{i=1}^{n_{\rm NL}} \sum_{j=1}^{n_{\rm L}} \|\boldsymbol{U}_{i,j}\|_U^2$ for all $\boldsymbol{U} \in U^{n_{\rm NL}\times n_{\rm L}}$.

Assumption 4b. The basis functions are differentiable in U and their gradient is uniformly bounded and Hölder continuous in X; that is,

- (1) $\partial_i \varphi_k(\boldsymbol{\xi}) \in U$, for all $\boldsymbol{\xi} \in \mathbb{X}$, $k \in \{1 : n_L\}$ and $i \in \{1 : n_{NL}\}$,
- (2) $\|\nabla_{\mathbb{X}}\varphi\|_{U,2,2,\infty} \doteq \sup_{\boldsymbol{\xi} \in \mathbb{X}} \|\nabla_{\mathbb{X}}\varphi(\boldsymbol{\xi})\|_{U,2,2} < \infty$, (3) there exist $L_{\nabla_{\mathbb{X}}\varphi} > 0$ and $\nu \in (0,1]$ such that $\|\nabla_{\mathbb{X}}\varphi(\boldsymbol{\xi}) \nabla_{\mathbb{X}}\varphi(\boldsymbol{\eta})\|_{U,2,2} \leq L_{\nabla_{\mathbb{X}}\varphi}\|\boldsymbol{\xi} \boldsymbol{\eta}\|_{\mathbb{X}}^{\nu}$ for all

Lemma 6. For all $v, w \in \mathbb{W}$, $\xi, \eta \in \mathbb{X}$,

$$\|\nabla_{\mathbb{W}}\mathcal{K}(\boldsymbol{v},\boldsymbol{\xi}) - \nabla_{\mathbb{W}}\mathcal{K}(\boldsymbol{w},\boldsymbol{\eta})\|_{2} \leq \|a\|_{U\times U}M_{\boldsymbol{\varphi}}(\boldsymbol{\xi},\boldsymbol{\eta})^{2}\|\boldsymbol{v} - \boldsymbol{w}\|_{2} + (2\|a\|_{U\times U}M_{\mathbb{W}}(\boldsymbol{v},\boldsymbol{w})M_{\boldsymbol{\varphi}}(\boldsymbol{\xi},\boldsymbol{\eta}) + \|\ell\|_{U})\|\boldsymbol{\varphi}(\boldsymbol{\xi}) - \boldsymbol{\varphi}(\boldsymbol{\eta})\|_{U2},$$

where $M_{\varphi}(\xi, \eta) = \max(\|\varphi(\xi)\|_{U,2}, \|\varphi(\eta)\|_{U,2})$ and $M_{\mathbb{W}}(v, w) = \max(\|v\|_{2}, \|w\|_{2})$. *Under Assumption 4b.1, for all* $v, w \in \mathbb{W}$, $\boldsymbol{\xi}, \boldsymbol{\eta} \in \mathbb{X}$,

$$\begin{split} &\|\nabla_{\mathbb{X}}\mathcal{K}(\boldsymbol{v},\boldsymbol{\xi}) - \nabla_{\mathbb{X}}\mathcal{K}(\boldsymbol{w},\boldsymbol{\eta})\|_{2} \\ &\leq (2\|a\|_{U\times U}M_{\mathbb{W}}(\boldsymbol{v},\boldsymbol{w})M_{\boldsymbol{\varphi}}(\boldsymbol{\xi},\boldsymbol{\eta}) + \|\ell\|_{U})M_{\nabla_{\mathbb{X}}\boldsymbol{\varphi}}(\boldsymbol{\xi},\boldsymbol{\eta})\|\boldsymbol{v} - \boldsymbol{w}\|_{2} \\ &+ \|a\|_{U\times U}M_{\mathbb{W}}(\boldsymbol{v},\boldsymbol{w})^{2}M_{\nabla_{\mathbb{X}}\boldsymbol{\varphi}}(\boldsymbol{\xi},\boldsymbol{\eta})\|\boldsymbol{\varphi}(\boldsymbol{\xi}) - \boldsymbol{\varphi}(\boldsymbol{\eta})\|_{U,2} \\ &+ M_{\mathbb{W}}(\boldsymbol{v},\boldsymbol{w})(\|a\|_{U\times U}M_{\mathbb{W}}(\boldsymbol{v},\boldsymbol{w})M_{\boldsymbol{\varphi}}(\boldsymbol{\xi},\boldsymbol{\eta}) + \|\ell\|_{U})\|\nabla_{\mathbb{X}}\boldsymbol{\varphi}(\boldsymbol{\xi}) - \nabla_{\mathbb{X}}\boldsymbol{\varphi}(\boldsymbol{\eta})\|_{U,2,2}, \end{split}$$

where $M_{\nabla_{\mathbb{X}}\boldsymbol{\varphi}}(\boldsymbol{\xi},\boldsymbol{\eta}) = \max(\|\nabla_{\mathbb{X}}\boldsymbol{\varphi}(\boldsymbol{\xi})\|_{U_{2,2}}, \|\nabla_{\mathbb{X}}\boldsymbol{\varphi}(\boldsymbol{\eta})\|_{U_{2,2}}).$

Proof. See Appendix A.8.

In particular, the map $\nabla_{\mathbb{W}}\mathcal{K}$ is Lipschitz continuous with respect to the linear parameters provided that the basis function map φ is uniformly bounded in U (Assumption 3.1). With respect to the nonlinear parameters, $\nabla_{\mathbb{W}}\mathcal{K}$ inherits the modulus of continuity of φ , but this continuity holds only locally in the linear parameters (on bounded subsets). The Lipschitz regularity of $\nabla_{\mathbb{W}}\mathcal{K}$ in the linear parameters is only needed when the linear parameters are not separated from the nonlinear parameters. If they are separated, only the regularity in the nonlinear parameters is needed, at every fixed linear parameter (see Assumption 4a).

For the derivative $\nabla_{\mathbb{X}}\mathcal{K}$, the continuity properties follow from those of φ and its parametric derivative $\nabla_{\mathbb{X}}\varphi$, again only locally in the linear parameters. This observation motivated the formulation of Assumption 4a allowing the continuity constant to depend on the norm of the linear parameters.

It is clear from Lemma 6 that Assumption 3 and Assumption 4b together imply Assumption 4a. However, when the basis functions fail to be differentiable in U, this implication breaks down; yet Assumption 4a may still hold and can often be verified directly.

4.4.2. Regularity of the reduced energy. It was shown in Subsection 2.3.3 that the lowest eigenvalue of $A(\xi)$ is larger than $\alpha\omega(\xi)$. Besides, Lemma 2 shows that the largest eigenvalue of $A(\xi)$ is smaller than $||a||_{U\times U}||\varphi(\xi)||_{U,2}^2$. Under Assumption 1 and Assumption 3.1, the condition number of $A(\xi)$ is thus uniformly bounded by

$$\kappa_{\max} \doteq \frac{\|a\|_{U \times U}}{\alpha} \frac{\|\varphi\|_{U,2,\infty}}{\omega_{\min}}.$$

With this notation, we can establish the boundedness of the best linear parameters and the regularity of the reduced energy.

Lemma 7. Under Assumption 1 and Assumption 3.1, it holds

$$\sup_{\boldsymbol{\xi} \in \mathbb{X}} \|\boldsymbol{w}^{\mathrm{best}}(\boldsymbol{\xi})\|_2 \leq \frac{\|\ell\|_U \|\boldsymbol{\varphi}\|_{U,2,\infty}}{\alpha \omega_{\min}},$$

and for all $\boldsymbol{\xi}, \boldsymbol{\eta} \in \mathbb{X}$, we have the bound

$$\|\boldsymbol{w}^{\mathrm{best}}(\boldsymbol{\xi}) - \boldsymbol{w}^{\mathrm{best}}(\boldsymbol{\eta})\|_{2} \leq (1 + 2\kappa_{\mathrm{max}}) \frac{\|\ell\|_{U}}{\alpha\omega_{\mathrm{min}}} \|\boldsymbol{\varphi}(\boldsymbol{\xi}) - \boldsymbol{\varphi}(\boldsymbol{\eta})\|_{U,2}.$$

Under the additional Assumption 4b.1 and Assumption 4b.2, it also holds

$$\begin{split} \|\nabla \mathcal{K}^{\mathrm{red}}(\boldsymbol{\xi}) - \nabla \mathcal{K}(\boldsymbol{\eta})^{\mathrm{red}}\|_{2} &\leq [\kappa_{\mathrm{max}} + (1 + 2\kappa_{\mathrm{max}})^{2}] \frac{\|\ell\|_{U}^{2}}{\alpha\omega_{\mathrm{min}}} \|\nabla_{\mathbb{X}}\boldsymbol{\varphi}\|_{U,2,2,\infty} \|\boldsymbol{\varphi}(\boldsymbol{\xi}) - \boldsymbol{\varphi}(\boldsymbol{\eta})\|_{U,2} \\ &+ (1 + \kappa_{\mathrm{max}}) \frac{\|\ell\|_{U}^{2}}{\alpha\omega_{\mathrm{min}}} \|\boldsymbol{\varphi}\|_{U,2,\infty} \|\nabla_{\mathbb{X}}\boldsymbol{\varphi}(\boldsymbol{\xi}) - \nabla_{\mathbb{X}}\boldsymbol{\varphi}(\boldsymbol{\eta})\|_{U,2,2}. \end{split}$$

Proof. See Appendix A.9.

Lemma 7 establishes that the best linear parameters are uniformly bounded and inherit the same regularity as the basis functions. Consequently, the reduced energy functional inherits the continuity properties of the basis functions and their gradients. Whenever $\nabla_{\mathbb{X}}\mathcal{K}$ is Lipschitz (via Assumption 4a or Lemma 6), the fact that $\boldsymbol{w}^{\text{best}}$ is Lipschitz and the multilinearity of $\nabla_{\mathbb{X}}\mathcal{K}$ in \boldsymbol{w} ensure that $\nabla\mathcal{K}^{\text{red}}$ is Lipschitz. When Assumption 4b holds, Lemma 7 provides tighter constants.

4.4.3. *Directional convexity*. We conclude our analysis with a lower bound on the Hessian of \mathcal{K}^{red} , which can be leveraged to show Assumption 5.

Lemma 8. Suppose $\varphi(\xi)$ is twice differentiable in U for all $\xi \in \mathbb{X}$. For all $\xi \in \mathbb{X}$, $v \in \mathbb{R}^{n_{\mathrm{NL}}}$, and $\xi^{\star} \in \mathbb{X}^{\star}$,

$$\nabla^2 \mathcal{K}^{\mathrm{red}}(\boldsymbol{\xi})(\boldsymbol{v},\boldsymbol{v}) \geq \left\| \nabla \mathcal{R}^{\mathrm{red}}(\boldsymbol{\xi}) \boldsymbol{v} \right\|_a^2 - (\left\| \mathcal{R}^{\mathrm{red}}(\boldsymbol{\xi}) - \mathcal{R}^{\mathrm{red}}(\boldsymbol{\xi}^{\star}) \right\|_a + \inf_{\boldsymbol{v} \in V} \left\| \boldsymbol{u}^{\star} - \boldsymbol{v} \right\|_a) \left\| \nabla^2 \mathcal{R}^{\mathrm{red}}(\boldsymbol{\xi})(\boldsymbol{v},\boldsymbol{v}) \right\|_a.$$

Proof. See Appendix A.10.

We now show how Lemma 8 can serve to prove that $\mathcal{K}^{\mathrm{red}}$ is directionally convex in a neighbourhood of its global minimisers. First, by Lemma 7 under Assumption 1 and Assumption 3.1, we infer

$$\begin{split} \|\mathcal{R}^{\mathrm{red}}(\boldsymbol{\xi}) - \mathcal{R}^{\mathrm{red}}(\boldsymbol{\eta})\|_{U} &= \|\boldsymbol{w}^{\mathrm{best}}(\boldsymbol{\xi})^{*}\boldsymbol{\varphi}(\boldsymbol{\xi}) - \boldsymbol{w}^{\mathrm{best}}(\boldsymbol{\xi})^{*}\boldsymbol{\varphi}(\boldsymbol{\eta})\|_{U} \\ &\leq \|[\boldsymbol{w}^{\mathrm{best}}(\boldsymbol{\xi}) - \boldsymbol{w}^{\mathrm{best}}(\boldsymbol{\eta})]^{*}\boldsymbol{\varphi}(\boldsymbol{\xi})\|_{U} + \|\boldsymbol{w}^{\mathrm{best}}(\boldsymbol{\eta})^{*}[\boldsymbol{\varphi}(\boldsymbol{\xi}) - \boldsymbol{\varphi}(\boldsymbol{\eta})]\|_{U} \\ &\leq 2(1 + \kappa_{\max}) \frac{\|\ell\|_{U}}{\alpha\omega_{\min}} \|\boldsymbol{\varphi}\|_{U,2,\infty} \|\boldsymbol{\varphi}(\boldsymbol{\xi}) - \boldsymbol{\varphi}(\boldsymbol{\eta})\|_{U,2}. \end{split}$$

From the coercivity and boundedness of a, we obtain $\|\cdot\|_a \leq \alpha^{-1} \|a\|_{U \times U} \|\cdot\|_U$ and therefore

$$\nabla^2 \mathcal{K}^{\mathrm{red}}(\boldsymbol{\xi})(\boldsymbol{v},\boldsymbol{v}) \geq \left\| \nabla \mathcal{R}^{\mathrm{red}}(\boldsymbol{\xi}) \boldsymbol{v} \right\|_a^2 - \left(C \| \boldsymbol{\varphi}(\boldsymbol{\xi}) - \boldsymbol{\varphi}(\boldsymbol{\xi}^\star) \|_{U,2} + \inf_{v \in V} \| \boldsymbol{u}^\star - \boldsymbol{v} \|_a \right) \| \nabla^2 \mathcal{R}^{\mathrm{red}}(\boldsymbol{\xi})(\boldsymbol{v},\boldsymbol{v}) \|_a,$$

where $C=2\kappa_{\max}(1+\kappa_{\max})\|\ell\|_U/\alpha$. We would like this lower bound to be non-negative in the direction $v=\xi^\star-\xi$ pointing to a global minimum that is closest to ξ , that is, for some $\xi^\star\in\Pi_\psi^\star(\xi)$. This condition is equivalent to

$$\|\nabla \mathcal{R}^{\mathrm{red}}(\boldsymbol{\eta})(\boldsymbol{\xi}^{\star} - \boldsymbol{\xi})\|_{a}^{2} \ge (C\|\boldsymbol{\varphi}(\boldsymbol{\xi}) - \boldsymbol{\varphi}(\boldsymbol{\xi}^{\star})\|_{U,2} + \inf_{v \in V} \|u^{\star} - v\|_{a})\|\nabla^{2}\mathcal{R}^{\mathrm{red}}(\boldsymbol{\eta})(\boldsymbol{\xi}^{\star} - \boldsymbol{\xi}, \boldsymbol{\xi}^{\star} - \boldsymbol{\xi})\|_{a},$$

for all $\eta \in [\xi, \xi^{\star}]$. We are interested in satisfying directional convexity in a neighbourhood of the type $\Lambda_{\psi}^{\star}(\rho)$ for some $\rho > 0$. Assuming that φ is Lipschitz (Assumption 3.2), a stronger condition for directional convexity in $\Lambda_{\psi}^{\star}(\rho)$ is the following: for all $\xi \in \Lambda_{\psi}^{\star}(\rho)$, there exists $\xi^{\star} \in \Pi_{\psi}^{\star}(\xi)$, such that for all $\eta \in [\xi, \xi^{\star}]$, it holds

$$\|\nabla \mathcal{R}^{\text{red}}(\boldsymbol{\eta})(\boldsymbol{\xi}^{\star} - \boldsymbol{\xi})\|_{a}^{2} \ge (CL_{\varphi}\rho + \inf_{v \in V} \|u^{\star} - v\|_{a})\|\nabla^{2}\mathcal{R}^{\text{red}}(\boldsymbol{\eta})(\boldsymbol{\xi}^{\star} - \boldsymbol{\xi}, \boldsymbol{\xi}^{\star} - \boldsymbol{\xi})\|_{a},$$

where L_{φ} is the Lipschitz constant of φ .

Intuitively, this condition characterises a form of quantitative directional convexity in directions pointing towards global minimisers. It ensures that, in a sufficiently small neighbourhood of the global minimisers, the directional derivative of $\mathcal{K}^{\rm red}$ in the descent direction $\boldsymbol{\xi}^{\star}-\boldsymbol{\xi}$ dominates the directional curvature, with a bound that scales linearly in the distance to the minimisers. This provides a measure of how sharply the function descends towards its minima, and can be interpreted as a localised, geometric analogue of gradient dominance. This condition has been introduced in [21] in the context of non-convex inverse problems, where it plays a central role in quantifying the size of the basins of attraction around global minimisers.

5. CONCLUSION

We develop a general optimisation framework for solving variational PDEs over nonlinear approximation spaces. The method combines energy minimisation in linear parameters with constrained mirror descent for nonlinear parameters. We provide a theoretical foundation ensuring local and global convergence under structural assumptions, including differentiability, boundedness, and directional convexity of the discrete energy functional.

These assumptions are stated in a modular fashion, enabling applicability to a broad class of nonlinear approximation manifolds. Examples include adaptive bases built from piecewise polynomials, wavelets, or radial basis functions. The framework also naturally extends to sparse grid methods, where hierarchical decompositions and selective tensor-product combinations replace full tensor-product spaces [39]. In a companion paper [24], we explore its application to overlapping tensor-product free-knot B-spline spaces. A detailed study of other constructions is left for future work.

Another interesting direction is the preconditioning of the optimisation process through Hessian-based techniques. In particular, preconditioning the gradient using (an approximation of) the Hessian matrix—as in second-order methods such as Newton or quasi-Newton algorithms, or in natural gradient descent [33]—could help mitigate ill-conditioning in the nonlinear optimisation problem. This may accelerate convergence and improve robustness by incorporating curvature information, accounting for the geometry of the parameter space, and better balancing the scales of different optimisation variables.

ACKNOWLEDGEMENTS

This research was partially funded by the Australian Government through the Australian Research Council (project DP220103160). A. Magueresse gratefully acknowledges the Monash Graduate Scholarship from Monash University.

APPENDIX A. PROOFS

A.1. Proof of Lemma 1.

Proof. Let $w \in \text{Ker}(A(\xi))$. By coercivity of the continuous bilinear form a, we have

$$\alpha \|\mathcal{R}(\boldsymbol{w}, \boldsymbol{\xi})\|_{U}^{2} \leq a(\mathcal{R}(\boldsymbol{w}, \boldsymbol{\xi}), \mathcal{R}(\boldsymbol{w}, \boldsymbol{\xi})) = \boldsymbol{w}^{*} \boldsymbol{A}(\boldsymbol{\xi}) \boldsymbol{w} = 0,$$

and thus $\mathcal{R}(\boldsymbol{w},\boldsymbol{\xi}) = 0$. In particular, this implies

$$\ell(\mathcal{R}(\boldsymbol{w},\boldsymbol{\xi})) = \boldsymbol{w}^* \boldsymbol{\ell}(\boldsymbol{\xi}) = 0.$$

Let us now show that $\ell(\xi) \in \operatorname{Im}(A(\xi))$. Since $A(\xi)$ is symmetric with real coefficients, it admits an orthogonal diagonalisation $A(\xi) = Q(\xi)^* \Lambda(\xi) Q(\xi)$, where $Q(\xi) \in \mathbb{R}^{n_L \times n_L}$ is orthogonal and $\Lambda(\xi) \in \mathbb{R}^{n_L \times n_L}$ is diagonal, with diagonal entries $(\lambda_k)_{k \in \{1:n_L\}}$, the eigenvalues of $A(\xi)$. Let $(q_k)_{k \in \{1:n_L\}}$ denote the columns of $Q(\xi)$, so that q_k is an eigenvector associated with λ_k . In particular, $(q_k)_{k \in \{1:n_L\}}$ is a basis so there exist $(c_k)_{k \in \{1:n_L\}}$ such that $\ell(\xi) = \sum_{k=1}^{n_L} c_k q_k$. Let $\ell \in \{1:n_L\}$ such that $\ell \in \{1:n_L\}$ such that $\ell \in \{1:n_L\}$ and by what is above, we infer that $\ell \in \{1:n_L\}$ such that $\ell \in \{1:n_L\}$ su

If $w_1, w_2 \in \mathbb{W}$ are two solutions of the system, then $w_1 - w_2 \in \operatorname{Ker}(A(\xi))$, and therefore $\mathcal{R}(w_1, \xi) = \mathcal{R}(w_2, \xi)$. This shows that the realisation is independent of the choice of solution.

A.2. Proof of Lemma 2.

Proof. Let $w \in \mathbb{W}$ and $\xi \in \mathbb{X}$, and w_+ correspond to the full linear solve. Note that $\nabla_{\mathbb{W}} \mathcal{K}(w, \xi) = A(\xi)w - \ell = A(\xi)(w - w_+)$. We thus express

$$\mathcal{K}(\boldsymbol{w}_{+},\boldsymbol{\xi}) = \mathcal{K}(\boldsymbol{w},\boldsymbol{\xi}) - \frac{1}{2}(\boldsymbol{w} - \boldsymbol{w}_{+})^{*}\boldsymbol{A}(\boldsymbol{\xi})(\boldsymbol{w} - \boldsymbol{w}_{+})$$

$$= \mathcal{K}(\boldsymbol{w},\boldsymbol{\xi}) - \frac{1}{2}\nabla_{\mathbb{W}}\mathcal{K}(\boldsymbol{w},\boldsymbol{\xi})^{*}\boldsymbol{A}(\boldsymbol{\xi})^{-1}\nabla_{\mathbb{W}}\mathcal{K}(\boldsymbol{w},\boldsymbol{\xi})$$

$$\leq \mathcal{K}(\boldsymbol{w},\boldsymbol{\xi}) - \frac{1}{2}\lambda_{\max}(\boldsymbol{A}(\boldsymbol{\xi}))^{-1}\|\nabla_{\mathbb{W}}\mathcal{K}(\boldsymbol{w},\boldsymbol{\xi})\|_{2,*}$$

If w_+ corresponds to the steepest descent update, we find

$$\mathcal{K}(\boldsymbol{w}_{+},\boldsymbol{\xi}) = \mathcal{K}(\boldsymbol{w},\boldsymbol{\xi}) + \frac{1}{2}\beta^{2}\boldsymbol{r}^{*}\boldsymbol{A}(\boldsymbol{\xi})\boldsymbol{r} + \beta\boldsymbol{r}^{*}\boldsymbol{A}(\boldsymbol{\xi})\boldsymbol{w} - \beta\boldsymbol{r}^{*}\boldsymbol{\ell}(\boldsymbol{\xi})$$
$$= \mathcal{K}(\boldsymbol{w},\boldsymbol{\xi}) - \frac{1}{2}\frac{(\boldsymbol{r},\boldsymbol{r})_{2}^{2}}{(\boldsymbol{r},\boldsymbol{A}(\boldsymbol{\xi})\boldsymbol{r})_{2}}.$$

Here again we conclude using the largest eigenvalue of $A(\xi)$ and the fact that $\nabla_{\mathbb{W}}\mathcal{K}(w, \xi) = r$. To bound the largest eigenvalue of $A(\xi)$, the continuity of a gives

$$\boldsymbol{w}^* \boldsymbol{A}(\boldsymbol{\xi}) \boldsymbol{w} = a(\mathcal{R}(\boldsymbol{w}, \boldsymbol{\xi}), \mathcal{R}(\boldsymbol{w}, \boldsymbol{\xi})) \le \|a\|_{U \times U} \|\mathcal{R}(\boldsymbol{w}, \boldsymbol{\xi})\|_U^2.$$

Now, the Gram matrix $G(\xi)$ allows one to write

$$\|\mathcal{R}(w, \xi)\|_U^2 = w^* G(\xi) w \le \lambda_{\max}(G(\xi)) \|w\|_2^2.$$

Since the Frobenius norm of a symmetric matrix is equal to the sum the squares of its eigenvalues, one can bound $\lambda_{\max}(G(\xi))^2$ by the Frobenius norm of $G(\xi)$. Using the Cauchy-Schwarz inequality, we obtain

$$\lambda_{\max}(\boldsymbol{G}(\boldsymbol{\xi}))^2 \leq \sum_{i,j=1}^{n_{\mathrm{L}}} (\boldsymbol{\varphi}_i(\boldsymbol{\xi}), \boldsymbol{\varphi}_j(\boldsymbol{\xi}))_U^2 \leq \sum_{i,j=1}^{n_{\mathrm{L}}} \|\boldsymbol{\varphi}_i(\boldsymbol{\xi})\|_U^2 \|\boldsymbol{\varphi}_j(\boldsymbol{\xi})\|_U^2 = \|\boldsymbol{\varphi}(\boldsymbol{\xi})\|_{U,2}^4.$$

We conclude that $\lambda_{\max}(G(\boldsymbol{\xi})) \leq \|\varphi(\boldsymbol{\xi})\|_{U,2}^2$ and $\lambda_{\max}(A(\boldsymbol{\xi})) \leq \|a\|_{U \times U} \|\varphi(\boldsymbol{\xi})\|_{U,2}^2$.

A.3. Tools for Lemma 3, Lemma 4 and Lemma 5. Let $w \in \mathbb{W}$. The function $\mathcal{K}(w,\cdot)$ has Hölder gradients with exponent ν and constant L(w). It is well-known that the fundamental theorem of calculus then yields

$$\mathcal{K}(\boldsymbol{w},\boldsymbol{\eta}) \leq \mathcal{K}(\boldsymbol{w},\boldsymbol{\xi}) + \langle \nabla_{\mathbb{X}}\mathcal{K}(\boldsymbol{w},\boldsymbol{\xi}), \boldsymbol{\eta} - \boldsymbol{\xi} \rangle_{\mathbb{X}} + \frac{L(\boldsymbol{w})}{1+\nu} \|\boldsymbol{\eta} - \boldsymbol{\xi}\|_{\mathbb{X}}^{1+\nu}$$

for all $\xi, \eta \in \mathbb{X}$. We now repeat an argument used in [40] to convert Hölder continuity into Lipschitz continuity up to an arbitrary small constant: by Young's inequality, it holds

$$ab \le \frac{a^p}{p} + \frac{b^q}{q},$$

for all $a, b \ge 0$ and $p, q \ge 1$ satisfying 1/p + 1/q = 1. Taking $a = \|\boldsymbol{\xi} - \boldsymbol{\eta}\|_{\mathbb{X}}^{1+\nu}$, $p = 2/(1+\nu)$ and therefore $q = 2/(1-\nu)$, we find

$$\frac{L(\boldsymbol{w})}{1+\nu}\|\boldsymbol{\xi}-\boldsymbol{\eta}\|_{\mathbb{X}}^{1+\nu} \leq \frac{L(\boldsymbol{w})}{2b}\|\boldsymbol{\xi}-\boldsymbol{\eta}\|_{\mathbb{X}}^{2} + \frac{1}{2}\frac{1-\nu}{1+\nu}b^{\frac{1+\nu}{1-\nu}}L(\boldsymbol{w}).$$

Letting $\varepsilon=\frac{1}{2}\frac{1-\nu}{1+\nu}b^{\frac{1+\nu}{1-\nu}}L(\boldsymbol{w})>0$ and solving for b in terms of ε , we obtain

$$\mathcal{K}(\boldsymbol{w}, \boldsymbol{\eta}) \leq \mathcal{K}(\boldsymbol{w}, \boldsymbol{\xi}) + \langle \nabla_{\mathbb{X}} \mathcal{K}(\boldsymbol{w}, \boldsymbol{\xi}), \boldsymbol{\eta} - \boldsymbol{\xi} \rangle_{\mathbb{X}} + \frac{1}{2} L_{\nu, \varepsilon}(\boldsymbol{w}) \| \boldsymbol{\eta} - \boldsymbol{\xi} \|_{\mathbb{X}}^{1+\nu} + \varepsilon$$

where we have introduced the Lipschitz constant

$$L_{
u,arepsilon}(oldsymbol{w}) \doteq \left(rac{1}{2arepsilon}rac{1-
u}{1+
u}
ight)^{rac{1-
u}{1+
u}}L(oldsymbol{w})^{rac{2}{1+
u}}.$$

If $\nu = 1$, one can also take $\varepsilon = 0$, and in that case $L_{\nu,\varepsilon}(\boldsymbol{w}) = L(\boldsymbol{w})$.

A.4. Proof of Lemma 3.

Proof. Let $(w, \xi) \in \mathbb{W} \times \mathbb{X}$. The Hölder continuity of $\nabla_{\mathbb{X}} \mathcal{K}$ at w implies

$$\mathcal{K}(\boldsymbol{w}, \boldsymbol{\xi}_{+}) \leq \mathcal{K}(\boldsymbol{w}, \boldsymbol{\xi}) + \left\langle \nabla_{\mathbb{X}} \mathcal{K}(\boldsymbol{w}, \boldsymbol{\xi}), \boldsymbol{\xi}_{+} - \boldsymbol{\xi} \right\rangle_{\mathbb{X}} + \frac{1}{2} L_{\nu, \varepsilon}(\boldsymbol{w}) \|\boldsymbol{\xi}_{+} - \boldsymbol{\xi}\|_{\mathbb{X}}^{2} + \varepsilon,$$

for any $\varepsilon > 0$. The first-order optimality condition characterising ξ_+ (9) states

$$-\gamma \nabla_{\mathbb{X}} \mathcal{K}(\boldsymbol{w}, \boldsymbol{\xi}) - \nabla \psi(\boldsymbol{\xi}_{+}) + \nabla \psi(\boldsymbol{\xi}) \in N_{\mathbb{X}}(\boldsymbol{\xi}_{+}).$$

In particular,

$$\langle -\gamma \nabla_{\mathbb{X}} \mathcal{K}(\boldsymbol{w}, \boldsymbol{\xi}) - \nabla \psi(\boldsymbol{\xi}_{+}) + \nabla \psi(\boldsymbol{\xi}), \boldsymbol{\xi} - \boldsymbol{\xi}_{+} \rangle_{\mathbb{X}} \leq 0,$$

which is equivalent to

$$\gamma \left\langle \nabla_{\mathbb{X}} \mathcal{K}(\boldsymbol{w}, \boldsymbol{\xi}), \boldsymbol{\xi}_{+} - \boldsymbol{\xi} \right\rangle_{\mathbb{X}} \leq - \left\langle \nabla \psi(\boldsymbol{\xi}_{+}) - \nabla \psi(\boldsymbol{\xi}), \boldsymbol{\xi}_{+} - \boldsymbol{\xi} \right\rangle_{\mathbb{X}}.$$

The strong convexity of ψ gives the upper bound $-\mu \|\xi_+ - \xi\|_{\mathbb{X}}^2$ for the right-hand side. Combining these inequalities and the definition of the gradient mapping, we reach

$$\mathcal{K}(\boldsymbol{w},\boldsymbol{\xi}_{+}) \leq \mathcal{K}(\boldsymbol{w},\boldsymbol{\xi}) - \frac{1}{2}\gamma(2\mu - \gamma L_{\nu,\varepsilon}(\boldsymbol{w})) \|G_{\psi}(\boldsymbol{w},\boldsymbol{\xi},\gamma)\|_{\mathbb{X},*}^{2} + \varepsilon.$$

Using the energy decrease condition (8), we arrive at

$$(10) \quad \mathcal{K}(\boldsymbol{w}_{+},\boldsymbol{\xi}_{+}) \leq \mathcal{K}(\boldsymbol{w},\boldsymbol{\xi}) - \frac{1}{2}\gamma(2\mu - \gamma L_{\nu,\varepsilon}(\boldsymbol{w})) \|G_{\psi}(\boldsymbol{w},\boldsymbol{\xi},\gamma)\|_{\mathbb{X},*}^{2} - \frac{1}{2}\beta(\boldsymbol{\xi})^{-1} \|\nabla_{\mathbb{W}}\mathcal{K}(\boldsymbol{w},\boldsymbol{\xi}_{+})\|_{\mathbb{W},*}^{2} + \varepsilon.$$

Let now $(\boldsymbol{w}_0, \boldsymbol{\xi}_0) \in \mathbb{W} \times \mathbb{X}$ and $(\boldsymbol{w}_k, \boldsymbol{\xi}_k)_{k \geq 0}$ denote the iterates produced by Algorithm 1 with step sizes $\gamma_k > 0$. The estimate (10) holds at each iteration. Summing this inequality and recognising a telescoping sum, we obtain

$$\sum_{k=0}^{n-1} A_k \|G_{\psi}(\boldsymbol{w}_k, \boldsymbol{\xi}_k, \gamma_k)\|_{\mathbb{X},*}^2 + B_k \|\nabla_{\mathbb{W}} \mathcal{K}(\boldsymbol{w}_k, \boldsymbol{\xi}_{k+1})\|_{\mathbb{W},*}^2 \leq 2(\mathcal{K}(\boldsymbol{w}_0, \boldsymbol{\xi}_0) - \mathcal{K}(\boldsymbol{w}_n, \boldsymbol{\xi}_n) + n\varepsilon),$$

where $A_k \doteq \gamma_k (2\mu - \gamma_k L_{\nu,\varepsilon}(\boldsymbol{w}_k))$ and $B_k \doteq \beta(\boldsymbol{\xi}_k)^{-1}$ Assuming $\sum_{k=0}^{n-1} \min(A_k, B_k) > 0$ and dividing by that sum to form a weighted mean, we conclude

$$\min_{0 \le k \le n-1} \left\{ \|G_{\psi}(\boldsymbol{w}_{k}, \boldsymbol{\xi}_{k}, \gamma_{k})\|_{\mathbb{X}, *}^{2} + \|\nabla_{\mathbb{W}} \mathcal{K}(\boldsymbol{w}_{k}, \boldsymbol{\xi}_{k+1})\|_{\mathbb{W}, *}^{2} \right\} \le \frac{2(\mathcal{K}(\boldsymbol{w}_{0}, \boldsymbol{\xi}_{0}) - \mathcal{K}^{*} + n\varepsilon)}{\sum_{k=0}^{n-1} \min(A_{k}, B_{k})},$$

where we also used $\mathcal{K}(\boldsymbol{w}_n, \boldsymbol{\xi}_n) \geq \mathcal{K}^*$.

A.5. Proof of Lemma 4.

Proof. The optimality condition defining $\boldsymbol{\xi}_+ \in \operatorname{Prox}_{\psi}(\boldsymbol{w}, \boldsymbol{\xi}, \gamma)$ (9) states $\nabla \psi(\boldsymbol{\xi}) - \nabla \psi(\boldsymbol{\xi}_+) - \gamma \nabla_{\mathbb{X}} \mathcal{K}(\boldsymbol{w}, \boldsymbol{\xi}) \in N_{\mathbb{X}}(\boldsymbol{\xi}_+)$. Dividing by γ and adding and subtracting $\nabla_{\mathbb{X}} \mathcal{K}(\boldsymbol{w}, \boldsymbol{\xi}_+)$, this is equivalent to

$$-\nabla_{\mathbb{X}}\mathcal{K}(\boldsymbol{w},\boldsymbol{\xi}_{+}) + \underbrace{\left[\nabla_{\mathbb{X}}\mathcal{K}(\boldsymbol{w},\boldsymbol{\xi}_{+}) - \nabla_{\mathbb{X}}\mathcal{K}(\boldsymbol{w},\boldsymbol{\xi}) + \gamma^{-1}(\nabla\psi(\boldsymbol{\xi}) - \nabla\psi(\boldsymbol{\xi}_{+}))\right]}_{\doteq \boldsymbol{v}} \in N_{\mathbb{X}}(\boldsymbol{\xi}_{+}).$$

Using the Hölder continuity of $\nabla_{\mathbb{X}} \mathcal{K}$, we compute

$$\|\boldsymbol{v}\|_{\mathbb{X},*} \leq \|\nabla_{\mathbb{X}}\mathcal{K}(\boldsymbol{w},\boldsymbol{\xi}_{+}) - \nabla_{\mathbb{X}}\mathcal{K}(\boldsymbol{w},\boldsymbol{\xi})\|_{\mathbb{X},*} + \gamma^{-1}\|\nabla\psi(\boldsymbol{\xi}_{+}) - \nabla\psi(\boldsymbol{\xi})\|_{\mathbb{X},*}$$
$$\leq L\|\boldsymbol{\xi}_{+} - \boldsymbol{\xi}\|_{\mathbb{X}}^{\nu} + \mu\gamma^{-1}\|\boldsymbol{\xi}_{+} - \boldsymbol{\xi}\|_{\mathbb{X},*}.$$

This shows $-\nabla_{\mathbb{X}}\mathcal{K}(\boldsymbol{w},\boldsymbol{\xi}_{+})\in N_{\mathbb{X}}(\boldsymbol{\xi}_{+})+\overline{B}_{\mathbb{X},*}(0,Lc^{\nu}+\mu\gamma^{-1}c),$ where $c=\|\boldsymbol{\xi}_{+}-\boldsymbol{\xi}\|_{\mathbb{X}}.$

Let now $h^2 = \|G_{\psi}(\boldsymbol{w}, \boldsymbol{\xi}, \gamma)\|_{\mathbb{X}, *}^2 + \|\nabla_{\mathbb{W}} \mathcal{K}(\boldsymbol{w}, \boldsymbol{\xi}_+)\|_{\mathbb{W}, *}^2$. In particular, the two terms are smaller than h. Therefore $\|\boldsymbol{\xi}_+ - \boldsymbol{\xi}\| \leq \gamma h$ and thus $-\nabla_{\mathbb{X}} \mathcal{K}(\boldsymbol{w}, \boldsymbol{\xi}_+) \in N_{\mathbb{X}}(\boldsymbol{\xi}_+) + \overline{B}_{\mathbb{X}, *}(0, L(\gamma h)^{\nu} + \mu h)$. We conclude that $(\boldsymbol{w}, \boldsymbol{\xi}_+)$ is a $(L(\gamma h)^{\nu} + \mu h)$ -quasi-stationary point.

A.6. Proof of Lemma 5.

Proof. Let $\boldsymbol{\xi} \in \Lambda_{\psi}^{\star}(\rho)$ and $\boldsymbol{\xi}_{+} \in \operatorname{Prox}_{\psi}(\boldsymbol{w}^{\operatorname{best}}(\boldsymbol{\xi}), \boldsymbol{\xi}, \gamma)$ for some $\gamma > 0$. By the directional convexity assumption, there exists $\boldsymbol{\xi}^{\star} \in \Pi_{\psi}^{\star}(\boldsymbol{\xi})$ such that $\mathcal{K}^{\operatorname{red}}$ is convex on the segment $[\boldsymbol{\xi}, \boldsymbol{\xi}^{\star}]$. Therefore

$$\mathcal{K}^{\text{red}}(\boldsymbol{\xi}) \leq \mathcal{K}^{\text{red}}(\boldsymbol{\xi}^{\star}) + \left\langle \nabla \mathcal{K}^{\text{red}}(\boldsymbol{\xi}), \boldsymbol{\xi} - \boldsymbol{\xi}^{\star} \right\rangle_{\mathbb{X}} \\
= \mathcal{K}^{\star} + \left\langle \nabla \mathcal{K}^{\text{red}}(\boldsymbol{\xi}), \boldsymbol{\xi} - \boldsymbol{\xi}_{+} \right\rangle_{\mathbb{X}} + \left\langle \nabla \mathcal{K}^{\text{red}}(\boldsymbol{\xi}), \boldsymbol{\xi}_{+} - \boldsymbol{\xi}^{\star} \right\rangle_{\mathbb{X}} \\
\leq \mathcal{K}^{\star} + \mathcal{K}^{\text{red}}(\boldsymbol{\xi}) - \mathcal{K}^{\text{red}}(\boldsymbol{\xi}_{+}) + \frac{1}{2} L^{\text{red}} \|\boldsymbol{\xi}_{+} - \boldsymbol{\xi}\|_{\mathbb{X}}^{2} + \left\langle \nabla \mathcal{K}^{\text{red}}(\boldsymbol{\xi}), \boldsymbol{\xi}_{+} - \boldsymbol{\xi}^{\star} \right\rangle_{\mathbb{X}},$$
(11)

where we used the Lipschitz continuity of $\nabla \mathcal{K}^{\mathrm{red}}$ to bound $\langle \nabla \mathcal{K}^{\mathrm{red}}(\boldsymbol{\xi}), \boldsymbol{\xi} - \boldsymbol{\xi}_+ \rangle_{\mathbb{X}}$. The first-order optimality condition (9) defining $\boldsymbol{\xi}_+$ at $\boldsymbol{\xi}^*$ yields

$$\left\langle -\gamma \nabla \mathcal{K}^{\text{red}}(\boldsymbol{\xi}) - \nabla \psi(\boldsymbol{\xi}_{+}) + \nabla \psi(\boldsymbol{\xi}) \in N_{\mathbb{X}}(\boldsymbol{\xi}_{+}), \boldsymbol{\xi}^{\star} - \boldsymbol{\xi}_{+} \right\rangle_{\mathbb{X}} \leq 0,$$

which is equivalent to

$$\gamma \left\langle \nabla \mathcal{K}^{\mathrm{red}}(\boldsymbol{\xi}), \boldsymbol{\xi}_{+} - \boldsymbol{\xi}^{\star} \right\rangle_{\mathbb{X}} \leq \left\langle \nabla \psi(\boldsymbol{\xi}) - \nabla \psi(\boldsymbol{\xi}_{+}), \boldsymbol{\xi}_{+} - \boldsymbol{\xi}^{\star} \right\rangle_{\mathbb{X}}.$$

From the definition of the Bregman divergence, we rewrite

$$\langle \nabla \psi(\boldsymbol{\xi}) - \nabla \psi(\boldsymbol{\xi}_+), \boldsymbol{\xi}_+ - \boldsymbol{\xi}^* \rangle_{\mathbb{X}} = D_{\psi}(\boldsymbol{\xi}^*; \boldsymbol{\xi}) - D_{\psi}(\boldsymbol{\xi}^*; \boldsymbol{\xi}_+) - D_{\psi}(\boldsymbol{\xi}_+; \boldsymbol{\xi}).$$

Plugging this inequality in (11) and simplifying, we reach

$$\mathcal{K}^{\text{red}}(\boldsymbol{\xi}_{+}) \leq \mathcal{K}^{\star} - \gamma^{-1} D_{\psi}(\boldsymbol{\xi}_{+}; \boldsymbol{\xi}) + \frac{1}{2} L^{\text{red}} \|\boldsymbol{\xi}_{+} - \boldsymbol{\xi}\|_{\mathbb{X}}^{2} + \gamma^{-1} (D_{\psi}(\boldsymbol{\xi}^{\star}; \boldsymbol{\xi}) - D_{\psi}(\boldsymbol{\xi}^{\star}; \boldsymbol{\xi}_{+})).$$

Now, using the fact that $\delta_{\psi}^{\star}(\boldsymbol{\xi}) = D_{\psi}(\boldsymbol{\xi}^{\star};\boldsymbol{\xi})$, the inequality $\delta_{\psi}^{\star}(\boldsymbol{\xi}_{+}) = D_{\psi}(\boldsymbol{\xi}^{\star};\boldsymbol{\xi}_{+})$ and the strong convexity inequality $D_{\psi}(\boldsymbol{\xi};\boldsymbol{\eta}) \geq \mu \|\boldsymbol{\xi} - \boldsymbol{\eta}\|_{\mathbb{X}}^{2}/2$, we finally obtain

(12)
$$\mathcal{K}^{\text{red}}(\boldsymbol{\xi}_{+}) \leq \mathcal{K}^{\star} - \frac{1}{2}(\mu \gamma^{-1} - L^{\text{red}}) \|\boldsymbol{\xi}_{+} - \boldsymbol{\xi}\|_{\mathbb{X}}^{2} + \gamma^{-1}(\delta_{\psi}^{\star}(\boldsymbol{\xi}) - \delta_{\psi}^{\star}(\boldsymbol{\xi}_{+})).$$

Assuming $\gamma L^{\mathrm{red}} \leq \mu$, we find that $\delta_{\psi}^{\star}(\boldsymbol{\xi}_{+}) \leq \delta_{\psi}^{\star}(\boldsymbol{\xi}) \leq \rho$, and therefore $\boldsymbol{\xi}_{+} \in \Lambda_{\psi}^{\star}(\rho)$.

A quick induction shows that if $\xi_0 \in \Lambda_\psi^\star(\rho)$ and $\gamma_k L^{\mathrm{red}} \leq \mu$ for all $k \geq 0$, then the iterates of Algorithm 1 remain in $\Lambda_\psi^\star(\rho)$, and that the estimate (12) holds at each step. According to Lemma 3, the condition $\gamma_k L^{\mathrm{red}} \leq \mu$

also ensures that $(\mathcal{K}^{\text{red}}(\boldsymbol{\xi}_k))_{k\geq 0}$ is non-increasing. We apply (12) at each step and multiply it by γ_k in view of telescoping $\delta_{\psi}^{\star}(\boldsymbol{\xi}_k)$. We recognise a weighted average of $(\mathcal{K}^{\text{red}}(\boldsymbol{\xi}_k) - \mathcal{K}^{\star})$ and find

$$\mathcal{K}^{\mathrm{red}}(\boldsymbol{\xi}_n) - \mathcal{K}^{\star} \leq \frac{\sum_{k=0}^{n-1} \gamma_k (\mathcal{K}^{\mathrm{red}}(\boldsymbol{\xi}_{k+1}) - \mathcal{K}^{\star})}{\sum_{k=0}^{n-1} \gamma_k} \leq \frac{\delta_{\psi}^{\star}(\boldsymbol{\xi}_0) - \delta_{\psi}^{\star}(\boldsymbol{\xi}_n)}{\sum_{k=0}^{n-1} \gamma_k} \leq \frac{\delta_{\psi}^{\star}(\boldsymbol{\xi}_0)}{\sum_{k=0}^{n-1} \gamma_k},$$

where the first inequality comes from the monotonicity of $(\mathcal{K}^{\text{red}}(\boldsymbol{\xi}_k))_{k>0}$.

A.7. Proof of Corollary 1.

Proof. Since \mathcal{K}^{red} satisfies the assumptions of Lemma 5, for all $n \geq 1$ and $v \in V$, we have

$$\begin{split} \frac{1}{2} \| \mathcal{R}^{\text{red}}(\boldsymbol{\xi}_n) - u^{\star} \|_a^2 &= \mathcal{K}^{\text{red}}(\boldsymbol{\xi}_n) - \mathcal{J}(u^{\star}) \\ &= \mathcal{J}(v) - \mathcal{J}(u^{\star}) + \mathcal{K}^{\text{red}}(\boldsymbol{\xi}_n) - \mathcal{J}(v) \\ &\leq \frac{1}{2} \| v - u^{\star} \|_a^2 + \mathcal{K}^{\text{red}}(\boldsymbol{\xi}_n) - \mathcal{K}^{\star} \\ &\leq \frac{1}{2} \| v - u^{\star} \|_a^2 + \frac{L^{\text{red}} \delta_{\psi}^{\star}(\boldsymbol{\xi}_0)}{\zeta \mu n}. \end{split}$$

Taking the infimum on $v \in V$ yields the result of the corollary.

A.8. Proof of Lemma 6.

Proof. Let $v, w \in \mathbb{W}$, and $\xi, \eta \in \mathbb{X}$. Let $i \in \{1 : n_L\}$ and ∂_i denote the derivative with respect to the *i*-th linear parameter. We express

$$\partial_i \mathcal{K}(\boldsymbol{v}, \boldsymbol{\xi}) = a(\boldsymbol{v}^* \boldsymbol{\varphi}(\boldsymbol{\xi}), \boldsymbol{\varphi}_i(\boldsymbol{\xi})) - \ell(\boldsymbol{\varphi}_i(\boldsymbol{\xi})).$$

Using the fact that x = c + d and y = c - d, where c = (x + y)/2 and d = (x - y)/2, to symmetrise $\partial_i \mathcal{K}(\boldsymbol{v}, \boldsymbol{\xi}) - \partial_i \mathcal{K}(\boldsymbol{w}, \boldsymbol{\eta})$, we find

$$\begin{aligned} |\partial_{i}\mathcal{K}(\boldsymbol{v},\boldsymbol{\xi}) - \partial_{i}\mathcal{K}(\boldsymbol{w},\boldsymbol{\eta})| &\leq \frac{1}{4}|a([\boldsymbol{v}-\boldsymbol{w}]^{*}[\boldsymbol{\varphi}(\boldsymbol{\xi}) + \boldsymbol{\varphi}(\boldsymbol{\eta})], \boldsymbol{\varphi}_{i}(\boldsymbol{\xi}) + \boldsymbol{\varphi}_{i}(\boldsymbol{\eta}))| \\ &+ \frac{1}{4}|a([\boldsymbol{v}+\boldsymbol{w}]^{*}[\boldsymbol{\varphi}(\boldsymbol{\xi}) - \boldsymbol{\varphi}(\boldsymbol{\eta})], \boldsymbol{\varphi}_{i}(\boldsymbol{\xi}) + \boldsymbol{\varphi}_{i}(\boldsymbol{\eta}))| \\ &+ \frac{1}{2}|a(\boldsymbol{v}^{*}\boldsymbol{\varphi}(\boldsymbol{\xi}) + \boldsymbol{w}^{*}\boldsymbol{\varphi}(\boldsymbol{\eta}), \boldsymbol{\varphi}_{i}(\boldsymbol{\xi}) - \boldsymbol{\varphi}_{i}(\boldsymbol{\eta}))| \\ &+ |\ell(\boldsymbol{\varphi}_{i}(\boldsymbol{\xi}) - \boldsymbol{\varphi}_{i}(\boldsymbol{\eta}))|. \end{aligned}$$

The inequality of the lemma follows from the boundedness of a and ℓ , the triangle inequality, the Cauchy-Schwarz inequality $\|\boldsymbol{v}^*\boldsymbol{\varphi}(\boldsymbol{\xi})\|_U \leq \|\boldsymbol{v}\|_2 \|\boldsymbol{\varphi}(\boldsymbol{\xi})\|_{U,2}$, the fact that $x+y \leq 2\max(x,y)$ and $\max(xy,zt) \leq \max(x,z)\max(y,t)$ for all $x,y,z,t\geq 0$.

For all $i \in \{1 : n_{\rm NL}\}$, let ∂_i now denote the derivative with respect to the i-th nonlinear parameter. Since $\partial_i \varphi_k(\xi) \in U$, we have

$$\partial_i \mathcal{K}(\boldsymbol{v}, \boldsymbol{\xi}) = a(\boldsymbol{v}^* \boldsymbol{\varphi}(\boldsymbol{\xi}), \boldsymbol{v}^* \partial_i \boldsymbol{\varphi}(\boldsymbol{\xi})) - \ell(\boldsymbol{v}^* \partial_i \boldsymbol{\varphi}(\boldsymbol{\xi})).$$

Applying the same symmetrisation technique as above twice to isolate each term with a single subtraction, we find

$$\begin{aligned} |\partial_{i}\mathcal{K}(\boldsymbol{v},\boldsymbol{\xi}) - \partial_{i}\mathcal{K}(\boldsymbol{w},\boldsymbol{\eta})| &\leq \frac{1}{4}|a(\boldsymbol{v}^{*}\boldsymbol{\varphi}(\boldsymbol{\xi}) + \boldsymbol{w}^{*}\boldsymbol{\varphi}(\boldsymbol{\eta}), [\boldsymbol{v} - \boldsymbol{w}]^{*}[\partial_{i}\boldsymbol{\varphi}(\boldsymbol{\xi}) + \partial_{i}\boldsymbol{\varphi}(\boldsymbol{\eta})])| \\ &+ \frac{1}{4}|a(\boldsymbol{v}^{*}\boldsymbol{\varphi}(\boldsymbol{\xi}) + \boldsymbol{w}^{*}\boldsymbol{\varphi}(\boldsymbol{\eta}), [\boldsymbol{v} + \boldsymbol{w}]^{*}[\partial_{i}\boldsymbol{\varphi}(\boldsymbol{\xi}) - \partial_{i}\boldsymbol{\varphi}(\boldsymbol{\eta})])| \\ &+ \frac{1}{4}|a([\boldsymbol{v} - \boldsymbol{w}]^{*}[\boldsymbol{\varphi}(\boldsymbol{\xi}) + \boldsymbol{\varphi}(\boldsymbol{\eta})], \boldsymbol{v}^{*}\partial_{i}\boldsymbol{\varphi}(\boldsymbol{\xi}) + \boldsymbol{w}^{*}\partial_{i}\boldsymbol{\varphi}(\boldsymbol{\eta}))| \\ &+ \frac{1}{4}|a([\boldsymbol{v} + \boldsymbol{w}]^{*}[\boldsymbol{\varphi}(\boldsymbol{\xi}) - \boldsymbol{\varphi}(\boldsymbol{\eta})], \boldsymbol{v}^{*}\partial_{i}\boldsymbol{\varphi}(\boldsymbol{\xi}) + \boldsymbol{w}^{*}\partial_{i}\boldsymbol{\varphi}(\boldsymbol{\eta}))| \\ &+ \frac{1}{2}|\ell([\boldsymbol{v} - \boldsymbol{w}]^{*}[\partial_{i}\boldsymbol{\varphi}(\boldsymbol{\xi}) + \partial_{i}\boldsymbol{\varphi}(\boldsymbol{\eta})])| \\ &+ \frac{1}{2}|\ell([\boldsymbol{v} + \boldsymbol{w}]^{*}[\partial_{i}\boldsymbol{\varphi}(\boldsymbol{\xi}) - \partial_{i}\boldsymbol{\varphi}(\boldsymbol{\eta})])|. \end{aligned}$$

We conclude using the same arguments as above.

A.9. **Proof of Lemma 7.**

Proof. Let $\xi \in \mathbb{X}$. The definition of $w^{\text{best}}(\xi)$, Assumption 1 and the Cauchy-Schwarz inequality yield

$$\begin{split} \alpha \omega_{\min} \| \boldsymbol{w}^{\text{best}}(\boldsymbol{\xi}) \|_2^2 &\leq \boldsymbol{w}^{\text{best}}(\boldsymbol{\xi})^* \boldsymbol{A}(\boldsymbol{\xi}) \boldsymbol{w}^{\text{best}}(\boldsymbol{\xi}) \\ &= \boldsymbol{w}^{\text{best}}(\boldsymbol{\xi})^* \boldsymbol{\ell}(\boldsymbol{\xi}) \\ &= \ell(\boldsymbol{w}^{\text{best}}(\boldsymbol{\xi})^* \boldsymbol{\varphi}(\boldsymbol{\xi})) \\ &\leq \| \ell \|_U \| \boldsymbol{w}^{\text{best}}(\boldsymbol{\xi}) \|_2 \| \boldsymbol{\varphi}(\boldsymbol{\xi}) \|_{U.2}. \end{split}$$

The first inequality of the lemma is obtained by dividing this inequality by $\alpha \omega_{\min} \| \boldsymbol{w}^{\text{best}}(\boldsymbol{\xi}) \|_2$ and invoking Assumption 3.1 to bound $\| \boldsymbol{\varphi}(\boldsymbol{\xi}) \|_{U,2}$.

Let now $\eta \in \mathbb{X}$ such that $\omega(\eta) > 0$. We first show that the matrix $(1 - t)A(\xi) + tA(\eta)$ is invertible for all $t \in [0, 1]$. Let $v \in \mathbb{W}$. We compute

$$\mathbf{v}^*[(1-t)\mathbf{A}(\boldsymbol{\xi}) + t\mathbf{A}(\boldsymbol{\eta})]\mathbf{v} = (1-t)\mathbf{v}^*\mathbf{A}(\boldsymbol{\xi})\mathbf{v} + t\mathbf{v}^*\mathbf{A}(\boldsymbol{\eta})\mathbf{v}$$

$$\geq (1-t)\alpha\omega_{\min}\|\mathbf{v}\|_2^2 + t\alpha\omega_{\min}\|\mathbf{v}\|_2^2$$

$$= \alpha\omega_{\min}\|\mathbf{v}\|_2^2.$$

This shows that the smallest eigenvalue of $(1-t)\mathbf{A}(\boldsymbol{\xi})+t\mathbf{A}(\boldsymbol{\eta})$ is greater than $\alpha\omega_{\min}$. This enables the definition of the map

$$h: [0,1] \to \mathbb{W}, \quad t \mapsto [(1-t)\mathbf{A}(\boldsymbol{\xi}) + t\mathbf{A}(\boldsymbol{\eta})]^{-1}[(1-t)\boldsymbol{\ell}(\boldsymbol{\xi}) + t\boldsymbol{\ell}(\boldsymbol{\eta})],$$

such that $\boldsymbol{w}^{\text{best}}(\boldsymbol{\xi}) - \boldsymbol{w}^{\text{best}}(\boldsymbol{\eta}) = h(0) - h(1)$. This map is differentiable and its derivative is

$$h'(t) = [(1-t)\mathbf{A}(\xi) + t\mathbf{A}(\eta)]^{-1}[\mathbf{A}(\xi) - \mathbf{A}(\eta)][(1-t)\mathbf{A}(\xi) + t\mathbf{A}(\eta)]^{-1}[(1-t)\ell(\xi) + t\ell(\eta)]$$
$$+ [(1-t)\mathbf{A}(\xi) + t\mathbf{A}(\eta)]^{-1}[\ell(\eta) - \ell(\xi)].$$

By the mean value theorem there exists $t \in (0,1)$ such that h(1) - h(0) = h'(t). Taking the norm of that equality, we infer

$$\begin{split} \|\boldsymbol{w}^{\text{best}}(\boldsymbol{\xi}) - \boldsymbol{w}^{\text{best}}(\boldsymbol{\eta})\|_{2} &\leq \sigma_{\text{max}}([(1-t)\boldsymbol{A}(\boldsymbol{\xi}) + t\boldsymbol{A}(\boldsymbol{\eta})]^{-1})^{2}\sigma_{\text{max}}(\boldsymbol{A}(\boldsymbol{\xi}) - \boldsymbol{A}(\boldsymbol{\eta}))\|(1-t)\boldsymbol{\ell}(\boldsymbol{\xi}) + t\boldsymbol{\ell}(\boldsymbol{\eta})\|_{2} \\ &+ \sigma_{\text{max}}([(1-t)\boldsymbol{A}(\boldsymbol{\xi}) + t\boldsymbol{A}(\boldsymbol{\eta})]^{-1})\|\boldsymbol{\ell}(\boldsymbol{\varphi}(\boldsymbol{\eta}) - \boldsymbol{\varphi}(\boldsymbol{\xi}))\|_{2} \\ &\leq \alpha^{-2}\omega_{\text{min}}^{-2}\|\boldsymbol{\ell}\|_{U}M_{\boldsymbol{\varphi}}\sigma_{\text{max}}(\boldsymbol{A}(\boldsymbol{\xi}) - \boldsymbol{A}(\boldsymbol{\eta})) \\ &+ \alpha^{-1}\omega_{\text{min}}^{-1}\|\boldsymbol{\ell}\|_{U}\|\boldsymbol{\varphi}(\boldsymbol{\xi}) - \boldsymbol{\varphi}(\boldsymbol{\eta})\|_{U2}, \end{split}$$

where we used Assumption 3.1 to bound $\max(\|\varphi(\xi)\|_{U,2}, \|\varphi(\eta)\|_{U,2})$. We also used the fact that $\|\boldsymbol{M}\boldsymbol{x}\|_2 \leq \sigma_{\max}(\boldsymbol{M})\|\boldsymbol{x}\|_2$ for all $\boldsymbol{M} \in \mathbb{R}^{n_{\mathrm{L}} \times n_{\mathrm{L}}}$ and $\boldsymbol{x} \in \mathbb{R}^{n_{\mathrm{L}}}$. Here $\sigma_{\max}(\boldsymbol{M})$ denotes the largest singular value of \boldsymbol{M} . Since the largest singular value of a matrix is bounded by its Frobenius norm, we obtain

$$\begin{split} \sigma_{\max}(\boldsymbol{A}(\boldsymbol{\xi}) - \boldsymbol{A}(\boldsymbol{\eta}))^2 &\leq \sum_{i,j=1}^{n_{\mathrm{L}}} (\boldsymbol{A}(\boldsymbol{\xi})_{ij} - \boldsymbol{A}(\boldsymbol{\eta})_{ij})^2 \\ &= \sum_{i,j=1}^{n_{\mathrm{L}}} \left[\frac{1}{2} a(\boldsymbol{\varphi}_i(\boldsymbol{\xi}) - \boldsymbol{\varphi}_i(\boldsymbol{\eta}), \boldsymbol{\varphi}_j(\boldsymbol{\xi}) + \boldsymbol{\varphi}_j(\boldsymbol{\eta})) + \frac{1}{2} a(\boldsymbol{\varphi}_i(\boldsymbol{\xi}) + \boldsymbol{\varphi}_i(\boldsymbol{\eta}), \boldsymbol{\varphi}_j(\boldsymbol{\xi}) - \boldsymbol{\varphi}_j(\boldsymbol{\eta})) \right]^2. \end{split}$$

The inequality $(x+y)^2 \le 2(x^2+y^2)$ and the symmetry of a provide

$$\sigma_{\max}(\boldsymbol{A}(\boldsymbol{\xi}) - \boldsymbol{A}(\boldsymbol{\eta}))^{2} \leq \frac{1}{2} \sum_{i,j=1}^{n_{L}} \left[a(\boldsymbol{\varphi}_{i}(\boldsymbol{\xi}) - \boldsymbol{\varphi}_{i}(\boldsymbol{\eta}), \boldsymbol{\varphi}_{j}(\boldsymbol{\xi}) + \boldsymbol{\varphi}_{j}(\boldsymbol{\eta}))^{2} + a(\boldsymbol{\varphi}_{i}(\boldsymbol{\xi}) + \boldsymbol{\varphi}_{i}(\boldsymbol{\eta}), \boldsymbol{\varphi}_{j}(\boldsymbol{\xi}) - \boldsymbol{\varphi}_{j}(\boldsymbol{\eta}))^{2} \right]$$

$$= \sum_{i,j=1}^{n_{L}} a(\boldsymbol{\varphi}_{i}(\boldsymbol{\xi}) - \boldsymbol{\varphi}_{i}(\boldsymbol{\eta}), \boldsymbol{\varphi}_{j}(\boldsymbol{\xi}) + \boldsymbol{\varphi}_{j}(\boldsymbol{\eta}))^{2}.$$

Now using the boundedness of a and the definition of the $\|\cdot\|_{U_2}$ norm, we obtain

$$\begin{split} \sigma_{\max}(\boldsymbol{A}(\boldsymbol{\xi}) - \boldsymbol{A}(\boldsymbol{\eta}))^2 &\leq \sum_{i,j=1}^{n_{\mathrm{L}}} \|\boldsymbol{a}\|_{U \times U}^2 \|\boldsymbol{\varphi}_i(\boldsymbol{\xi}) - \boldsymbol{\varphi}_i(\boldsymbol{\eta})\|_U^2 \|\boldsymbol{\varphi}_j(\boldsymbol{\xi}) + \boldsymbol{\varphi}_j(\boldsymbol{\eta})\|_U^2 \\ &= \|\boldsymbol{a}\|_{U \times U}^2 \|\boldsymbol{\varphi}(\boldsymbol{\xi}) - \boldsymbol{\varphi}(\boldsymbol{\eta})\|_{U,2}^2 \|\boldsymbol{\varphi}(\boldsymbol{\xi}) + \boldsymbol{\varphi}(\boldsymbol{\eta})\|_{U,2}^2. \end{split}$$

REFERENCES 23

Combining this bound of $\sigma_{\max}(A(\xi) - A(\eta))$ and the upper bound of $\|w^{\text{best}}(\xi) - w^{\text{best}}(\eta)\|_2$ above shows the second inequality of the lemma.

The last inequality of the lemma is obtained by combining Lemma 6 with the first two inequalities. \Box

A.10. Proof of Lemma 8.

Proof. Let $\xi \in \mathbb{X}$ and $v \in \mathbb{R}^{n_{\mathrm{NL}}}$. Since φ is twice differentiable in U, we infer that

$$\nabla^2 \mathcal{K}^{\text{red}}(\boldsymbol{\xi}) = \nabla_{\mathbb{W}} \nabla_{\mathbb{X}} \mathcal{K}(\boldsymbol{w}, \boldsymbol{\xi})|_{\boldsymbol{w} = \boldsymbol{w}^{\text{best}}(\boldsymbol{\xi})} + \nabla_{\mathbb{X}} \nabla_{\mathbb{X}} \mathcal{K}(\boldsymbol{w}, \boldsymbol{\xi})|_{\boldsymbol{w} = \boldsymbol{w}^{\text{best}}(\boldsymbol{\xi})},$$

where we used the fact that $\nabla_{\mathbb{W}}\nabla_{\mathbb{X}}=\nabla_{\mathbb{W}}\nabla_{\mathbb{X}}$ and $\nabla_{\mathbb{W}}\mathcal{K}(\boldsymbol{w}^{\mathrm{best}}(\boldsymbol{\xi}),\boldsymbol{\xi})=0$. Therefore, it is enough to look at the hessian of \mathcal{K} with respect to \mathbb{X} . Recalling the expression of the hessian of the linear and bilinear forms, we write

$$\nabla_{\mathbb{X}}^{2}\mathcal{K}(\boldsymbol{w},\boldsymbol{\xi})(\boldsymbol{v},\boldsymbol{v}) = a(\nabla_{\mathbb{X}}\mathcal{R}(\boldsymbol{w},\boldsymbol{\xi})\boldsymbol{v}, \nabla_{\mathbb{X}}\mathcal{R}(\boldsymbol{w},\boldsymbol{\xi})\boldsymbol{v})$$

$$+ a(\mathcal{R}(\boldsymbol{w},\boldsymbol{\xi}), \nabla_{\mathbb{X}}^{2}\mathcal{R}(\boldsymbol{w},\boldsymbol{\xi})(\boldsymbol{v},\boldsymbol{v})) - \ell(\nabla_{\mathbb{X}}^{2}\mathcal{R}(\boldsymbol{w},\boldsymbol{\xi})(\boldsymbol{v},\boldsymbol{v}))$$

$$= a(\nabla_{\mathbb{X}}\mathcal{R}(\boldsymbol{w},\boldsymbol{\xi})\boldsymbol{v}, \nabla_{\mathbb{X}}\mathcal{R}(\boldsymbol{w},\boldsymbol{\xi})\boldsymbol{v})$$

$$+ a(\mathcal{R}(\boldsymbol{w},\boldsymbol{\xi}) - u^{\star}, \nabla_{\mathbb{X}}^{2}\mathcal{R}(\boldsymbol{w},\boldsymbol{\xi})(\boldsymbol{v},\boldsymbol{v})),$$

where we used the second-order conformity and the Galerkin orthogonality for the continuous solution. The Cauchy-Schwarz inequality then yields

$$\nabla_{\mathbb{X}}^{2}\mathcal{K}(\boldsymbol{w},\boldsymbol{\xi})(\boldsymbol{v},\boldsymbol{v}) \geq \|\nabla_{\mathbb{X}}\mathcal{R}(\boldsymbol{w},\boldsymbol{\xi})\boldsymbol{v}\|_{a}^{2} - \|\mathcal{R}(\boldsymbol{w},\boldsymbol{\xi}) - \boldsymbol{u}^{\star}\|_{a}\|\nabla_{\mathbb{X}}^{2}\mathcal{R}(\boldsymbol{w},\boldsymbol{\xi})(\boldsymbol{v},\boldsymbol{v})\|_{a}.$$

Next, we decompose

$$\|\mathcal{R}(\boldsymbol{w}, \boldsymbol{\xi}) - u^{\star}\|_{a} \le \|\mathcal{R}(\boldsymbol{w}, \boldsymbol{\xi}) - v\|_{a} + \|v - u^{\star}\|_{a}$$

for all $v \in V$. Choosing $v = \mathcal{R}^{\text{red}}(\boldsymbol{\xi}^*)$ for some $\boldsymbol{\xi}^* \in \mathbb{X}^*$, the second term is equal to the distance from u^* to V. Evaluating the expression above at $\boldsymbol{w} = \boldsymbol{w}^{\text{best}}(\boldsymbol{\xi})$, we conclude

$$\nabla^2 \mathcal{K}^{\mathrm{red}}(\boldsymbol{\xi})(\boldsymbol{v},\boldsymbol{v}) \geq \left\| \nabla \mathcal{R}^{\mathrm{red}}(\boldsymbol{\xi}) \boldsymbol{v} \right\|_a^2 - (\left\| \mathcal{R}^{\mathrm{red}}(\boldsymbol{\xi}) - \mathcal{R}^{\mathrm{red}}(\boldsymbol{\xi}^\star) \right\|_a + \inf_{\boldsymbol{v} \in V} \left\| \boldsymbol{u}^\star - \boldsymbol{v} \right\|_a) \left\| \nabla^2 \mathcal{R}^{\mathrm{red}}(\boldsymbol{\xi})(\boldsymbol{v},\boldsymbol{v}) \right\|_a,$$
 for all $\boldsymbol{\xi}^\star \in \mathbb{X}^\star$.

REFERENCES

- [1] M. Raissi, P. Perdikaris, and G. E. Karniadakis. "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations". In: *Journal of Computational physics* 378 (2019), pp. 686–707. DOI: 10.1016/j.jcp.2018.10.045.
- [2] W. E and B. Yu. "The Deep Ritz Method: A Deep Learning-Based Numerical Algorithm for Solving Variational Problems". In: *Communications in Mathematics and Statistics* 6.1 (2018). DOI: 10.1007/s40304-018-0127-z.
- [3] T. De Ryck and S. Mishra. "Numerical analysis of physics-informed neural networks and related models in physics-informed machine learning". In: *Acta Numerica* 33 (2024), pp. 633–713. DOI: 10.1017/S0962492923000089.
- [4] C. Beck, A. Jentzen, and B. Kuckuck. "Full error analysis for the training of deep neural networks". In: *Infinite Dimensional Analysis, Quantum Probability and Related Topics* 25.02 (2022), p. 2150020. DOI: 10.1142/S021902572150020X.
- [5] D. L. Jupp. "Approximation to data by splines with free knots". In: *SIAM Journal on Numerical Analysis* 15.2 (1978), pp. 328–343. DOI: 10.1137/0715022.
- [6] G. Beliakov. "Least squares splines with free knots: global optimization approach". In: *Applied mathematics and computation* 149.3 (2004), pp. 783–798. DOI: 10.1016/S0096-3003 (03) 00179-6.
- [7] P. Kovács and A. M. Fekete. "Nonlinear least-squares spline fitting with variable knots". In: *Applied Mathematics and Computation* 354 (2019), pp. 490–501. DOI: 10.1016/j.amc.2019.02.051.
- [8] T. Schütze and H. Schwetlick. "Bivariate free knot splines". In: *BIT Numerical Mathematics* 43 (2003), pp. 153–178. DOI: 10.1023/A:1023609324173.
- [9] X. Deng and T. S. Denney Jr. "On optimizing knot positions for multidimensional B-spline models". In: *Computational Imaging II*. Vol. 5299. SPIE. 2004, pp. 175–186. DOI: 10.1117/12.527245.
- [10] Y. Zhang, J. Cao, Z. Chen, X. Li, and X.-M. Zeng. "B-spline surface fitting with knot position optimization". In: *Computers & Graphics* 58 (2016), pp. 73–83. DOI: 10.1016/j.cag.2016.05.010.
- [11] J. He, L. Li, J. Xu, and C. Zheng. "Relu deep neural networks and linear finite elements". In: *Journal of Computational Mathematics* 38.3 (2020), pp. 502–527. DOI: 10.4208/jcm.1901-m2018-0160.

24 REFERENCES

[12] J. A. Opschoor, P. C. Petersen, and C. Schwab. "Deep ReLU networks and high-order finite element methods". In: *Analysis and Applications* 18.05 (2020), pp. 715–770. DOI: 10.1142/S0219530519410136.

- [13] D. Yarotsky. "Error bounds for approximations with deep ReLU networks". In: *Neural networks* 94 (2017), pp. 103–114. DOI: 10.1016/j.neunet.2017.07.002.
- P. Petersen and F. Voigtlaender. "Optimal approximation of piecewise smooth functions using deep ReLU neural networks". In: *Neural Networks* 108 (2018), pp. 296–330. DOI: 10.1016/j.neunet.2018.08.019.
- [15] I. Daubechies, R. DeVore, S. Foucart, B. Hanin, and G. Petrova. "Nonlinear approximation and (deep) ReLU networks". In: *Constructive Approximation* 55.1 (2022), pp. 127–172. DOI: 10.1007/s00365-021-09548-z.
- [16] M. Ainsworth and J. T. Oden. "A posteriori error estimation in finite element analysis". In: *Computer methods in applied mechanics and engineering* 142.1-2 (1997), pp. 1–88. DOI: 10.1016/S0045-7825 (96) 01107-3.
- [17] I. Babuška and M. Suri. "The p- and hp- versions of the finite element method, basic principles and properties". In: SIAM review 36.4 (1994), pp. 578–632. DOI: 10.1137/1036141.
- [18] W. Huang and R. D. Russell. *Adaptive moving mesh methods*. Vol. 174. Springer Science & Business Media, 2010. DOI: 10.1007/978-1-4419-7916-2.
- [19] C. J. Budd, W. Huang, and R. D. Russell. "Adaptivity with moving grids". In: *Acta Numerica* 18 (2009), pp. 111–241. DOI: 10.1017/S0962492906400015.
- [20] H. Karimi, J. Nutini, and M. Schmidt. "Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition". In: *Joint European conference on machine learning and knowledge discovery in databases*. Springer. 2016, pp. 795–811. DOI: 10.1007/978-3-319-46128-1 50.
- [21] Y. Traonmilin, J.-F. Aujol, and A. Leclaire. "The basins of attraction of the global minimizers of non-convex inverse problems with low-dimensional models in infinite dimension". In: *Information and Inference: A Journal of the IMA* 12.1 (2023), pp. 113–156. DOI: 10.1093/imaiai/iaac011.
- [22] S. Bubeck et al. "Convex optimization: Algorithms and complexity". In: *Foundations and Trends*® *in Machine Learning* 8.3-4 (2015), pp. 231–357. DOI: 10.1561/2200000050.
- [23] W. H. Lawton and E. A. Sylvestre. "Elimination of linear parameters in nonlinear regression". In: *Technometrics* 13.3 (1971), pp. 461–467. DOI: 10.2307/1267160.
- [24] A. Magueresse and S. Badia. *Energy minimisation using overlapping tensor-product free-knot B-splines*. 2025. arXiv: 2508.17705.
- [25] I. Ekeland and R. Témam. *Convex analysis and variational problems*. SIAM, 1999. DOI: 10.1137/1.9781611971088.
- [26] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*. Vol. 87. Springer Science & Business Media, 2013. DOI: 10.1007/978-1-4419-8853-9.
- [27] P. Petersen, M. Raslan, and F. Voigtlaender. "Topological properties of the set of functions generated by neural networks of fixed size". In: *Foundations of computational mathematics* 21 (2021), pp. 375–444. DOI: 10.1007/s10208-020-09461-0.
- [28] S. Mishra and R. Molinaro. "Estimates on the generalization error of physics-informed neural networks for approximating PDEs". In: *IMA Journal of Numerical Analysis* 43.1 (2023), pp. 1–43. DOI: 10.1093/imanum/drab032.
- [29] Y. Saad. Iterative methods for sparse linear systems. SIAM, 2003. DOI: 10.1137/1.9780898718003.
- [30] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. 2014. arXiv: 1412.6980.
- [31] J. Nocedal and S. J. Wright. *Numerical optimization*. Springer, 1999. DOI: 10.1007/978-0-387-40065-5.
- [32] S.-I. Amari. "Natural gradient works efficiently in learning". In: *Neural computation* 10.2 (1998), pp. 251–276. DOI: 10.1162/089976698300017746.
- [33] J. Martens. "New insights and perspectives on the natural gradient method". In: *Journal of Machine Learning Research* 21.146 (2020), pp. 1–76.
- [34] B. T. Polyak. *Introduction to optimization*. New York, Optimization Software, 1987.
- [35] M. Yashtini. "On the global convergence rate of the gradient descent method for functions with Hölder continuous gradients". In: *Optimization letters* 10 (2016), pp. 1361–1370. DOI: 10.1007/s11590–015-0936-x.

REFERENCES 25

- [36] L. Nurbekyan, W. Lei, and Y. Yang. "Efficient natural gradient descent methods for large-scale PDE-based optimization problems". In: *SIAM Journal on Scientific Computing* 45.4 (2023), A1621–A1655. DOI: 10.1137/22M147780.
- [37] I. Bioli, C. Marcati, and G. Sangalli. "Accelerating Natural Gradient Descent for PINNs with Randomized Numerical Linear Algebra". In: (2025). arXiv: 2505.11638.
- [38] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan. "How to escape saddle points efficiently". In: *International conference on machine learning*. PMLR. 2017, pp. 1724–1732.
- [39] H.-J. Bungartz and M. Griebel. "Sparse grids". In: *Acta numerica* 13 (2004), pp. 147–269. DOI: 10. 1017/S0962492904000182.
- [40] Y. Nesterov. "Universal gradient methods for convex optimization problems". In: *Mathematical Programming* 152.1 (2015), pp. 381–404. DOI: 10.1007/s10107-014-0790-0.