Multilevel Optimization Methods: Convergence and Problem Structure

Chin Pang Ho*, Panos Parpas†

Department of Computing, Imperial College London, United Kingdom

October 18, 2016

Abstract

Building upon multigrid methods, the framework of multilevel optimization methods was developed to solve structured optimization problems, including problems in optimal control [13], image processing [29], etc. In this paper, we give a broader view of the multilevel framework and establish some connections between multilevel algorithms and the other approaches. An interesting case of the so called Galerkin model is further studied. By studying three different case studies of the Galerkin model, we take the first step to show how the structure of optimization problems could improve the convergence of multilevel algorithms.

1 Introduction

Multigrid methods are considered as the standard approach in solving differential equations [3, 15, 17, 34, 37, 40]. When solving a differential equation using numerical methods, an approximation of the solution is obtained on a mesh via discretization. The computational cost of solving the discretized problem, however, varies and it depends on the choice of the mesh size used. Therefore, by considering different mesh sizes, a hierarchy of discretized models can be defined. In general, a more accurate solution can be obtained with a smaller mesh size chosen, which results in a discretized problem in higher dimensions. We shall follow the traditional terminologies in the multigrid community and call a *fine model* to be the discretization in which its solution is sufficiently close to the solution of the original differential equation; otherwise we call it *coarse model* [3]. The main idea of multigrid methods is to make use of the geometric similarity between different discretizations. In particular, during the iterative process of computing solution of the fine model, one replaces part of the computations with the information from coarse models. The advantages of using multigrid methods are twofold. Firstly, coarse models are in the lower dimensions compared to the fine model, and so the computational cost is reduced. Secondly and interestingly, the directions generated by coarse model and fine model are in fact complementary. It has been shown that using the fine model is effective in reducing the high frequency components of the residual (error) but ineffective in reducing and

*c.ho12@imperial.ac.uk

†p.parpas@imperial.ac.uk

alternating the low frequency components. Those low frequency components, however, will become high frequency after dimensional reduction. Thus, they could be eliminated effectively using coarse models [3, 34].

This idea of multigrid was extended to optimization. Nash [27] proposed a multigrid framework for unconstrained infinite-dimensional convex optimization problems. Examples of such problems could be found in the area of optimal control. Following the idea of Nash, many multigrid optimization methods were further developed [27, 28, 25, 24, 22, 39, 14]. In particular, Wen and Goldfarb [39] provided a line search-based multigrid optimization algorithm under the framework in [27], and further extended the framework to nonconvex problems. Gratton et al [14] provided a sophisticated trust-region version of multigrid optimization algorithms, in which they called it multiscale algorithm, and in the later developments [39], the name multilevel algorithm is used. In this paper, we will consistently use the name *multilevel algorithms* for all these optimization algorithms, but we emphasize that the terms multilevel, multigrid, and multiscale were used interchangeably in different literatures. On the other hand, we keep the name *multigrid methods* for the conventional multigrid methods that solve linear or nonlinear equations that are discretizations arising from partial differential equations (PDEs).

It is worth mentioning that different multilevel algorithms were developed beyond infinite-dimensional problems, see for example Markov decision processes [18], image deblurring [29], and face recognition [19]. The above algorithms all have the same aim: to speed up the computations by making use of the geometric similarity between different models in the hierarchy.

The numerical performance of multilevel algorithms has been satisfying. In particular, both of the line-search based [39] and trust-region based [13] algorithms outperform standard methods when solving infinite-dimensional problems. Numerical results show that multilevel algorithms can take the advantage of the geometric similarity between different discretizations just as the original multigrid methods.

However, to the best of our knowledge, no theoretical result is able to show the advantages of using multilevel optimization algorithms. For the line-search based algorithm, Wen and Goldfarb [39] proved a sublinear convergence rate for strongly convex problems and convergence for nonconvex problems. Gratton et al [14] proved that their trust-region based multilevel algorithm requires the same order of number of iterations as compared to the gradient method.

Building upon the above developments, in this paper, we aim to address three fundamental issues with the current multilevel optimization framework. Firstly, under the general framework of multilevel optimization, could we connect classical optimization algorithms with the recently developed multilevel optimization algorithms? Secondly, could we extend the current analysis and explain why multilevel optimization algorithms outperform standard methods for some classes of problems (e.g. infinite-dimensional problems)? Thirdly, how do we construct a coarse model when the hierarchy is not obvious?

The contributions of this paper are:

- We provide a more complete view of line search multilevel algorithm, and in particular, we connect the general framework of the multilevel algorithm with classical optimization algorithms, such as variable metric methods and block-coordinate type methods. We also make a connection with the algorithm stochastic variance reduced gradient (SVRG) [20].

- We analyze the multilevel algorithm with the Galerkin model. The key feature of the Galerkin model is that a coarse model is created from the first and second order information of the fine model. The name "Galerkin model" is given in [14] since this is related to the Galerkin approximation in algebraic multigrid methods [35]. We will call this algorithm

the **G**alerkin-based **A**lgebraic **M**ultilevel **A**lgorithm (**GAMA**). A global convergence analysis of GAMA is provided.

- We propose to use the composite rate for analysis of the local convergence of GAMA. As we will show later, neither linear convergence nor quadratic convergence is suitable when studying the local convergence due to the broadness of GAMA.
- We study the composite rate of GAMA in a case study of infinite dimensional optimization problems. We show that the linear component of the composite rate is inversely proportional to the smoothness of the residual, which agrees with the findings in conventional multigrid methods.
- We show that GAMA can be set up as Newton's method in lower dimensions with low rank approximation to Hessians. This is done by a low rank approximation method called the navïe Nyström method. We show how the dimensions of the coarse model and the spectrum of the eigenvalues would affect the composite rate.
- GAMA can also be set up as Newton's method with block-diagonal approximation of the Hessians. We define a class of objective functions with weakly-connected Hessians. That is, the Hessians of the function have the form of a linear combination of a block-diagonal matrix and a general matrix which its entries are in $\mathcal{O}(\delta)$, for $\delta << 1$. We show how δ would vary the composite rate, and at the limit $\delta \to 0$, GAMA would achieve the quadratic rate of convergence.

The rest of this paper is structured as follows: In Section 2 we provide background material and introduce different variants of multilevel algorithms. We also show that several existing optimization algorithms are in fact special cases under the general framework of multilevel algorithm. In Section 3, we study the convergence of GAMA. We first derive the global convergence rate of GAMA, and then show that GAMA exhibits composite convergence when the current incumbent is sufficiently close to the optimum. Composite convergence rate is defined as a linear combination of linear convergence and quadratic convergence, and we denote r_1 and r_2 as the coefficient of linear rate and quadratic rate, respectively. Using these results, in Section 4 we derive the complexity of both GAMA and Newton's method. When r_1 is sufficiently small, we show that GAMA has less complexity compared to Newton's method. In Section 5-7, three special cases of GAMA are considered. We compute r_1 in each case and show the relationship between r_1 and the structure of the problem. In Section 5, we study problems arising from discretizations of one-dimensional PDE problems; in Section 6 we study problems where low rank approximation of Hessians is sufficiently accurate; in Section 7 we study the problems where the Hessians of the objective function are nearly block-diagonal. In Section 8 we illustrate the convergence of GAMA using several numerical examples, including variational optimization problems and machine learning problems.

2 Multilevel Models

In this section a broad view of the general multilevel framework will be provided. We start with basic settings and the core idea of multilevel algorithms in [14, 24, 39], then we show that the general multilevel framework covers several optimization algorithms, including the variable metric methods, block-coordinate descent, and stochastic variance reduced gradient. At the end of this section we provide the settings and details of the core topic of this paper - Galerkin model.

2.1 Basic Settings

In this paper we are interested in solving,

$$\min_{\mathbf{x}_h \in \mathbb{R}^N} f_h(\mathbf{x}_h), \tag{1}$$

where $\mathbf{x}_h \in \mathbb{R}^N$, and function $f_h : \mathbb{R}^N \to \mathbb{R}$ is continuous, differentiable, and strongly convex.

We first clarify the use of the subscript h. Throughout this paper, the lower case h represents that this is associated with the ***fine*** (exact) model. To use multilevel methods, one needs to formulate a hierarchy of models, and models with lower dimensions (resolutions) called the ***coarse*** models. To avoid the unnecessary complications, in this paper we consider only two models in the hierarchy: fine and coarse. In the same manner of using subscript h, we assign the upper case H to represent the association with coarse model. We assign N and n ($n \leq N$) to be the dimensions of fine model and coarse model, respectively. For instance, any vector that is within the space $\mathbb{R}^N$ is denoted with subscript h, and similarly, any vector with subscript H is within the space $\mathbb{R}^n$.

Assumption 1. *There exists constants μ_h, L_h, and M_h such that*

$$\mu_h \mathbf{I} \preccurlyeq \nabla^2 f_h(\mathbf{x}) \preccurlyeq L_h \mathbf{I}, \quad \forall \mathbf{x}_h \in \mathbb{R}^n, \tag{2}$$

and

$$\|\nabla^2 f_h(\mathbf{x}) - \nabla^2 f_h(\mathbf{y})\| \leq M_h \|\mathbf{x} - \mathbf{y}\|. \tag{3}$$

Equation (2) implies

$$\|\nabla f_h(\mathbf{x}_h) - \nabla f_h(\mathbf{y}_h)\| \leq L_h \|\mathbf{x}_h - \mathbf{y}_h\|.$$

The above assumption of the objective function will be used throughout this paper, and it is common when studying second order algorithms.

Multilevel methods require mapping information across different dimensions. To this end, we define a matrix $\mathbf{P} \in \mathbb{R}^{N \times n}$ to be the prolongation operator which maps information from coarse to fine, and we define a matrix $\mathbf{R} \in \mathbb{R}^{n \times N}$ to be the restriction operator which maps information from fine to coarse. We make the following assumption on $\mathbf{P}$ and $\mathbf{R}$.

Assumption 2. *The restriction operator $\mathbf{R}$ is the transpose of the prolongation operator $\mathbf{P}$ up to a constant c. That is,*

$$\mathbf{P} = c\mathbf{R}^T, \qquad c > 0.$$

Without loss of generality, we take $c = 1$ throughout this paper to simplify the use of notation for the analysis. We also assume any useful (non-zero) information in the coarse model will not become zero after prolongation and make the following assumption.

Assumption 3. *The prolongation operator $\mathbf{P}$ has full column rank, and so*

$$\mathit{rank}(\mathbf{P}) = n.$$

Notice that Assumption 2 and 3 are standard assumptions for multilevel methods [3, 16, 39]. Since $\mathbf{P}$ has full column rank, we define the pseudoinverse and its norm

$$\mathbf{P}^+ = (\mathbf{R}\mathbf{P})^{-1}\mathbf{R}, \quad \text{and} \quad \xi = \|\mathbf{P}^+\|. \tag{4}$$

The coarse model is constructed in the following manner. Suppose in the k^{th} iterations we have an incumbent solution $\mathbf{x}_{h,k}$ and gradient $\nabla f_{h,k} \triangleq \nabla f_h(\mathbf{x}_{h,k})$, then the corresponding coarse model is,

$$\min_{\mathbf{x}_H \in \mathbf{R}^n} \phi_H(\mathbf{x}_H) \triangleq f_H(\mathbf{x}_H) + \langle \mathbf{v}_H, \mathbf{x}_H - \mathbf{x}_{H,0} \rangle, \tag{5}$$

where,

$$\mathbf{v}_H \triangleq -\nabla f_{H,0} + \mathbf{R}\nabla f_{h,k},$$

$\mathbf{x}_{H,0} = \mathbf{R}\mathbf{x}_{h,k}$, and $f_H : \mathbb{R}^n \to \mathbb{R}$. Similar to $\nabla f_{h,k}$, we denote $\nabla^2 f_{H,0} \triangleq \nabla^2 f_h(\mathbf{x}_{H,0})$ and $\nabla \phi_{H,0} \triangleq \nabla \phi_H(\mathbf{x}_{H,0})$ to simplify notation. Similar notation will be used consistently unless it is specified otherwise. We emphasize the construction of coarse model (5) is common in the line of multilevel optimization research and it is not original in this paper. See for example [14, 24, 39]. Note that when constructing the coarse model (5), one needs to add an additional linear term on $f_H(\mathbf{x}_H)$. This linear term ensures the following is satisfied,

$$\nabla \phi_{H,0} = \mathbf{R}\nabla f_{h,k}. \tag{6}$$

For infinite-dimensional optimization problems, one can define f_h and f_H using discretization with different mesh sizes. In general, f_h is the function that is sufficiently close to the original problem, and that can be achieved using small mesh sizes. Based on geometric similarity between discretizations with different meshes, $f_h \approx f_H$ even though $n \leq N$.

However, we want to emphasize $f_h \approx f_H$ is not a necessary requirement when using multilevel methods. In principle, $f_H(\mathbf{x}_H)$ can be any function. Galerkin model, as we will show later, is a quadratic model where f_H is chosen to be an approximation of the Hessian of f_h.

2.2 The General Multilevel Algorithm

The main idea of multilevel algorithms is to use the coarse model to compute search directions. We call such direction the ***coarse correction step***. When using coarse correction step, we compute the direction by solving the corresponding coarse model (5) and perform the update,

$$\mathbf{x}_{h,k+1} = \mathbf{x}_{h,k} + \alpha_{h,k}\hat{\mathbf{d}}_{h,k},$$

with

$$\hat{\mathbf{d}}_{h,k} \triangleq \mathbf{P}(\mathbf{x}_{H,\star} - \mathbf{x}_{H,0}), \tag{7}$$

where $\mathbf{x}_{H,\star}$ is the solution of the coarse model, and $\alpha_{h,k} \in \mathbb{R}^+$ is the stepsize. We clarify that the "hat" in $\hat{\mathbf{d}}_{h,k}$ is used to identify a coarse correction step. The subscript h in $\hat{\mathbf{d}}_{h,k}$ is used because $\hat{\mathbf{d}}_{h,k} \in \mathbb{R}^N$.

We should emphasize that $\mathbf{x}_{H,\star}$ in (7) can be replaced by $\mathbf{x}_{H,r}$ for $r = 1, 2, \ldots$, i.e. the incumbent solution of the coarse mode (5) after r^{th} iterations. However, for the purpose of this paper and simplicity, we ignore this case unless there is extra specification, and we let (7) be the coarse correction step.

It is known that the coarse correction step $\hat{\mathbf{d}}_{h,k}$ is a descent direction if f_H is convex. The following lemma states this argument rigorously. Even though the proof is provided in [39], we provide it with our notation for the completeness of this paper.

Lemma 4 ([39]). *If f_H is a convex function, then the coarse correction step is a descent direction. In particular, in the k^{th} iteration,*

$$\nabla f_{h,k}^T \hat{\mathbf{d}}_{h,k} \leq \phi_{H,\star} - \phi_{H,0} \leq 0.$$

Proof.

$$\begin{aligned}\nabla f_{h,k}^T \hat{\mathbf{d}}_{h,k} &= \nabla f_{h,k}^T \mathbf{R}^T \left(\mathbf{x}_{H,\star} - \mathbf{x}_{H,0}\right), \\ &= \left(\mathbf{R}\nabla f_{h,k}\right)^T \left(\mathbf{x}_{H,\star} - \mathbf{x}_{H,0}\right), \\ &= \nabla \phi_{H,0}^T \left(\mathbf{x}_{H,\star} - \mathbf{x}_{H,0}\right), \\ &\leq \phi_{H,\star} - \phi_{H,0}.\end{aligned}$$

as required. ■

The last inequality holds because ϕ_H is a convex function. Even though Lemma 4 states that $\hat{\mathbf{d}}_{h,k}$ is a descent direction, using coarse correction step solely is not sufficient to solve the fine model (1).

Proposition 5. *Suppose $\nabla f_{h,k} \neq 0$ and $\nabla f_{h,k} \in$ null$(\mathbf{R})$, then the coarse correction step*

$$\hat{\mathbf{d}}_{h,k} = 0.$$

Proof. From (6), $\mathbf{x}_{H,\star} = \mathbf{x}_{H,0}$ when $\mathbf{R}\nabla f_{h,k} = 0$. Thus, $\hat{\mathbf{d}}_{h,k} = \mathbf{P}(\mathbf{x}_{H,\star} - \mathbf{x}_{H,0}) = 0$. ■

Recall that $\mathbf{R} \in \mathbb{R}^{n \times N}$, and so for $n < N$, a coarse correction step could be zero and make no progress even when the first order necessary condition $\nabla f_h = 0$ has not been satisfied.

2.2.1 Fine Correction Step

Two approaches can be used when coarse correction step is not progressing nor effective. The first approach is to compute directions using standard optimization methods. We call such step the ***fine correction step***. As opposed to coarse correction step $\hat{\mathbf{d}}_{h,k}$, we abandon the use of "hat" for all fine correction steps and denote them as $\mathbf{d}_{h,k}$'s.

Classical examples of $\mathbf{d}_{h,k}$'s are steps that are computed by standard methods such as gradient descent method, quasi-Newton method, etc. We perform fine correction step when coarse correction step is not effective. That is,

$$\|\mathbf{R}\nabla f_{h,k}\| < \kappa \|\nabla f_{h,k}\| \quad \text{or} \quad \|\mathbf{R}\nabla f_{h,k}\| < \epsilon, \tag{8}$$

where $\kappa \in (0, \min(1, \|\mathbf{R}\|))$, and $\epsilon \in (0, 1)$. The above criteria prevent using coarse model when $\mathbf{x}_{H,0} \approx \mathbf{x}_{H,\star}$, i.e. the coarse correction step $\hat{\mathbf{d}}_{h,k}$ is close to $\mathbf{0}$. We point out that these criteria were also proposed in [39]. We also make the following assumption on the fine correction step throughout this paper.

Assumption 6. *There exists strictly positive constants $\nu_h, \zeta_h > 0$ such that*

$$\|\mathbf{d}_{h,k}\| \leq \nu_h \|\nabla f_{h,k}\|, \quad \textit{and} \quad -\nabla f_{h,k}^T \mathbf{d}_{h,k} \geq \zeta_h \|\nabla f_{h,k}\|^2,$$

where $\mathbf{d}_{h,k}$ is a fine correction step. As a consequence, there exists a constant $\Lambda_h > 0$ such that

$$f_{h,k} - f_{h,k+1} \geq \Lambda_h \|\nabla f_{h,k}\|^2,$$

where $f_{h,k+1}$ is updated using a fine correction step.

As we will show later, Assumption 6 is not restrictive, and Λ_h is known for well-known cases like gradient descent, Newton method, etc. Using the combination of fine and coarse correction steps is the standard approach in multilevel methods, especially for PDE-based optimization problems [14, 24, 39].

2.2.2 Multiple P's and R's

The second approach to overcome issue of ineffective coarse correction step is by creating multiple coarse models with different $\mathbf{P}$'s and $\mathbf{R}$'s.

Proposition 7. *Suppose $\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_p$ are all restriction operators that satisfy Assumption 2 and 3, where $\mathbf{R}_i \in \mathbb{R}^{n_i \times N}$ for $i = 1, 2, \ldots, p$. Denote $\mathcal{S}$ to be a set that contains the rows of $\mathbf{R}_i$'s in $\mathbb{R}^N$, for $i = 1, 2, \ldots, p$. If*

$$\textit{span}(\mathcal{S}) = \mathbb{R}^N,$$

then for $\nabla f_{h,k} \neq 0$ there exists at least one $\mathbf{R}_j \in \{\mathbf{R}_i\}_{i=1}^p$ such that

$$\hat{\mathbf{d}}_{h,k} \neq 0 \quad \textit{and} \quad \nabla f_{h,k}^T \hat{\mathbf{d}}_{h,k} < 0,$$

where $\hat{\mathbf{d}}_{h,k}$ is the coarse correction step computed using $\mathbf{R}_j$.

Proof. Since $\text{span}(\mathcal{S}) = \mathbb{R}^N$, then for $\nabla f_{h,k} \neq 0$, there exists one $\mathbf{R}_j$ such that $\mathbf{R}_j \nabla f_{h,k} \neq 0$. So the corresponding coarse model would have $\mathbf{x}_{H,\star} \neq \mathbf{x}_{H,0}$, and thus $\hat{\mathbf{d}}_{h,k_j} \neq 0$. ■

Proposition 7 shows that if the rows of restriction operators $\mathbf{R}_i$'s span $\mathbb{R}^N$, then at least one coarse correction step from these restriction operators would be nonzero and thus effective. In each iteration, one could use the similar idea as in (8) to rule out ineffective coarse models. However, this checking process could be expensive for large scale problems with large p (number of restriction operators). To omit this checking process, one could choose the following alternatives.

i. **Cyclical approach**: choose $\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_p$ in order at each iteration, and choose $\mathbf{R}_1$ after $\mathbf{R}_p$.

ii. **Probabilistic approach**: assign a probability mass function with $\{\mathbf{R}_i\}_{i=1}^p$ as a sample space, and choose the coarse model randomly based on the mass function. The mass function has to be strictly positive for each $\mathbf{R}_i$'s.

We point out that this idea of using multiple coarse models is related to domain decomposition methods, which solve (non-)linear equations arising from PDEs. Domain decomposition methods partition the problem domain into several sub-domains, and thus decompose the original problem into several smaller problems. We refer the readers to [5] for more details about domain decomposition methods.

In Section 2.3, we will show that using multiple $\mathbf{P}$'s and $\mathbf{R}$'s is not new in the optimization research community. Using the above multilevel framework, one can re-generate the block-coordinate descent.

2.3 Connection with Variable Metric Methods

Using the above multilevel framework, in the rest of this section we will introduce different versions of multilevel algorithms: variable metric methods, block-coordinate descent, and stochastic variance reduced gradient. At the end of this section we will introduce the Galerkin model, which is an interesting case of the multilevel framework.

Recall that for variable metric methods, the direction $\mathbf{d}_{h,k}$ is computed by solving

$$\begin{aligned} \mathbf{d}_{h,k} &= \arg\min_{\mathbf{d}} \frac{1}{2}\langle \mathbf{d}, \mathbf{Q}\mathbf{d}\rangle + \langle \nabla f_{h,k}, \mathbf{d}\rangle, \\ &= -\mathbf{Q}^{-1}\nabla f_{h,k}. \end{aligned} \tag{9}$$

where $\mathbf{Q} \in \mathbb{R}^{N\times N}$ is a positive definite matrix. When $\mathbf{Q} = \mathbf{I}$, $\mathbf{d}_{h,k}$ is the steepest descent search direction. When $\mathbf{Q} = \nabla^2 f_{h,k}$, $\mathbf{d}_{h,k}$ is the search direction by Newton's method. When $\mathbf{Q}$ is an approximation of the Hessian, then $\mathbf{d}_{h,k}$ is the quasi-Newton search direction.

To show the connections between multilevel methods and variable metric methods, consider the following f_H.

$$f_H(\mathbf{x}_H) = \frac{1}{2}\langle \mathbf{x}_H - \mathbf{x}_{H,0}, \mathbf{Q}_H(\mathbf{x}_H - \mathbf{x}_{H,0})\rangle, \tag{10}$$

where $\mathbf{Q}_H \in \mathbb{R}^{n\times n}$, and $\mathbf{x}_{H,0} = \mathbf{R}\mathbf{x}_{h,k}$ as defined in (5). Applying the definition of the coarse model (5), we obtain,

$$\min_{\mathbf{x}_H\in\mathbf{R}^n} \phi_H(\mathbf{x}_H) = \frac{1}{2}\langle \mathbf{x}_H - \mathbf{x}_{H,0}, \mathbf{Q}_H(\mathbf{x}_H - \mathbf{x}_{H,0})\rangle + \langle \mathbf{R}\nabla f_{h,k}, \mathbf{x}_H - \mathbf{x}_{H,0}\rangle. \tag{11}$$

Thus from the definition in (7), the associated coarse correction step is,

$$\hat{\mathbf{d}}_{h,k} = \mathbf{P}\left(\arg\min_{\mathbf{d}_H\in\mathbf{R}^n} \underbrace{\frac{1}{2}\langle \mathbf{d}_H, \mathbf{Q}_H\mathbf{d}_H\rangle + \langle \mathbf{R}\nabla f_{h,k}, \mathbf{d}_H\rangle}_{\mathbf{d}_H = \mathbf{x}_H - \mathbf{x}_{H,0}}\right) = -\mathbf{P}\mathbf{Q}_H^{-1}\mathbf{R}\nabla f_{h,k}. \tag{12}$$

Therefore, with this specific f_H in (10), the resulting coarse model (11) is analogous to variable metric methods. In a naive case where $n = N$ and $\mathbf{P} = \mathbf{R} = \mathbf{I}$, the corresponding coarse correction step (12) would be the same as steepest descent direction, Newton direction, and quasi-Newton direction for $\mathbf{Q}_H$ that is identity matrix, Hessian, and approximation of Hessian, respectively.

2.4 Connection with Block-coordinate Descent

Interestingly, the coarse model (11) is also related to block-coordinate type methods. Suppose we have p coarse models with prolongation and restriction operators, $\{\mathbf{P}_i\}_{i=1}^p$ and $\{\mathbf{R}_i\}_{i=1}^p$, respectively. For each coarse model, we let (10) be the corresponding f_H with $\mathbf{Q}_H = \mathbf{I}$, and we further restrict our setting with the following properties.

1. $\mathbf{P}_i \in \mathbb{R}^{N\times n_i}$, $\forall i = 1, 2, \ldots, p$.
2. $\mathbf{P}_i = \mathbf{R}_i^T$, $\forall i = 1, 2, \ldots, p$.
3. $[\mathbf{P}_1\ \mathbf{P}_2 \ldots \mathbf{P}_p] = \mathbf{I}$.

From (12), the above setting results in $\hat{\mathbf{d}}_{h,k_i} = -\mathbf{P}_i\mathbf{R}_i\nabla f_{h,k}$, where $\hat{\mathbf{d}}_{h,k_i}$ is the coarse correction step for the i^{th} model. Notice that

$$(\mathbf{P}_i\mathbf{R}_i\nabla f_{h,k})_j = \begin{cases} (\nabla f_{h,k})_j & \text{if } \sum_{q=1}^{i-1} n_q < j \le \sum_{q=1}^{i} n_q, \\ 0 & \text{otherwise .} \end{cases}$$

Therefore, $\hat{\mathbf{d}}_{h,k_i}$ is equivalent to a block-coordinate descent update [1]. When $n_i = 1$, for $i = 1, 2, \ldots, p$, it becomes a coordinate descent method. When $1 < n_i < N$, for $i = 1, 2, \ldots, p$, it becomes a block-coordinate descent. When $\mathbf{P}_i$'s and $\mathbf{R}_i$'s are chosen using the cyclical approach, then it would be a cyclical (block)-coordinate descent. When $\mathbf{P}_i$'s and $\mathbf{R}_i$'s are chosen using the probabilistic approach, then it would be a randomized (block)-coordinate descent method.

2.5 Connection with SVRG

The multilevel framework is also related to the Stochastic Variance Reduced Gradient (SVRG) and its variants [12, 20, 26], which is a state-of-the-art algorithm for structured machine learning problems. Suppose the fine model has the following form

$$\min_{\mathbf{x}_h \in \mathbb{R}^N} f_h(\mathbf{x}_h) = \frac{1}{M} \sum_{i=1}^{M} f_i(\mathbf{x}_h).$$

We denote a set, $\mathcal{S}_H \subseteq \{1, 2, \ldots, M\}$ with $|\mathcal{S}_H| = m$, and construct the following coarse model

$$\min_{\mathbf{x}_H \in \mathbb{R}^N} f_H(\mathbf{x}_H) = \frac{1}{m} \sum_{i \in \mathcal{S}_H} f_i(\mathbf{x}_H).$$

In this particular case where $\mathbf{x}_h, \mathbf{x}_H \in \mathbb{R}^N$, no dimension is reduced, and we let $\mathbf{P} = \mathbf{R} = \mathbf{I}$. In the k^{th} iteration with incumbent $\mathbf{x}_k$, the coarse model is

$$\min_{\mathbf{x}_H \in \mathbb{R}^N} \frac{1}{m} \sum_{i \in \mathcal{S}_H} f_i(\mathbf{x}_H) + \left\langle -\frac{1}{m} \sum_{i \in \mathcal{S}_H} \nabla f_i(\mathbf{x}_{h,k}) + \frac{1}{M} \sum_{i=1}^{M} \nabla f_i(\mathbf{x}_{h,k}), \mathbf{x}_H - \mathbf{x}_{h,k} \right\rangle.$$

Suppose steepest descent is applied for K steps to solve the above coarse model, then

$$\mathbf{x}_{H,j} = \mathbf{x}_{H,j-1} - \alpha_{H,j} \left(\frac{1}{m} \sum_{i \in \mathcal{S}_H} \nabla f_i(\mathbf{x}_{H,j-1}) - \frac{1}{m} \sum_{i \in \mathcal{S}_H} \nabla f_i(\mathbf{x}_{h,k}) + \frac{1}{M} \sum_{i=1}^{M} \nabla f_i(\mathbf{x}_{h,k}) \right),$$

for $j = 1, 2, \ldots, K$. The above update is the key step in SVRG and its variants. In particular, when $m = K_d = 1$, the above setting is the same as the original SVRG in [20] with 1 inner iteration. Even though the coarse model is in the same dimension as the fine model, the cost of computing function values and gradients is much cheaper when $m << M$.

2.6 The Galerkin Model

We end this section with the core topic of this paper - the Galerkin model. The Galerkin coarse model is a special case of (11) where,

$$\mathbf{Q}_H = \nabla_H^2 f_{h,k} \triangleq \mathbf{R} \nabla^2 f_{h,k} \mathbf{P}, \tag{13}$$

and so the Galerkin (coarse) model is,

$$\min_{\mathbf{x}_H \in \mathbf{R}^n} \phi_H(\mathbf{x}_H) = \frac{1}{2} \langle \mathbf{x}_H - \mathbf{x}_{H,0}, \nabla_H^2 f_{h,k} (\mathbf{x}_H - \mathbf{x}_{H,0}) \rangle + \langle \mathbf{R} \nabla f_{h,k}, \mathbf{x}_H - \mathbf{x}_{H,0} \rangle. \tag{14}$$

According to (12), the corresponding coarse correction step is

$$\hat{\mathbf{d}}_{h,k} = -\mathbf{P} [\mathbf{R} \nabla^2 f_{h,k} \mathbf{P}]^{-1} \mathbf{R} \nabla f_{h,k} = -\mathbf{P} [\nabla_H^2 f_{h,k}]^{-1} \mathbf{R} \nabla f_{h,k}. \tag{15}$$

The Galerkin model is closely related to algebraic multigrid methods which solve (non-)linear equations arising from PDEs. Algebraic multigrid methods are used when computation or implementation of f_H is difficult (see e.g. [35]). In the context of multilevel optimization, to the best of our knowledge, this is first mentioned in [14] by Gratton, Sartenaer, and Toint. In [14] a trust-region type multilevel method is proposed to solve PDE-based optimization problems, and

the Galerkin model is described as a "radical strategy". In a later paper from Gratton et al. [13], the trust-region type multilevel method is tested numerically, and Galerkin model provides good numerical results.

It is worth mentioning that the above coarse correction step is equivalent to the solution of the system of linear equations,

$$\mathbf{R}\nabla^2 f_{h,k}\mathbf{P}\mathbf{d}_H = -\mathbf{R}\nabla f_{h,k}. \tag{16}$$

which is the general case of the Newton's method in which $\mathbf{P} = \mathbf{R} = \mathbf{I}$. Using Assumption 3, we can show that $\nabla_H^2 f_{h,k}$ is positive definite, and so equation (16) has a unique solution.

Proposition 8. $\mathbf{R}\nabla^2 f_h(\mathbf{x}_h)\mathbf{P}$ *is positive definite, and in particular,*

$$\mu_h \xi^{-2}\mathbf{I} \preceq \mathbf{R}\nabla^2 f_h(\mathbf{x}_h)\mathbf{P} \preceq L_h\omega^2\mathbf{I}$$

where $\omega = \max\{\|\mathbf{P}\|, \|\mathbf{R}\|\}$ *and* $\xi = \|\mathbf{P}^+\|$.

Proof.

$$\mathbf{x}^T\left(\mathbf{R}\nabla^2 f_h(\mathbf{x}_h)\mathbf{P}\right)\mathbf{x} = (\mathbf{P}\mathbf{x})^T\nabla^2 f_h(\mathbf{x}_h)(\mathbf{P}\mathbf{x}) \leq L_h\|\mathbf{P}\mathbf{x}\|^2 \leq L_h\omega^2\|\mathbf{x}\|^2.$$

Also,

$$\mathbf{x}^T\left(\mathbf{R}\nabla^2 f_h(\mathbf{x}_h)\mathbf{P}\right)\mathbf{x} = (\mathbf{P}\mathbf{x})^T\nabla^2 f_h(\mathbf{x}_h)(\mathbf{P}\mathbf{x}) \geq \mu_h\|\mathbf{P}\mathbf{x}\|^2 \geq \frac{\mu_h}{\|\mathbf{P}^+\|^2}\|\mathbf{x}\|^2 = \frac{\mu_h}{\xi^2}\|\mathbf{x}\|^2.$$

So we obtain the desired result. ■

3 Convergence of GAMA

In this section we will analyze GAMA that is stated as Algorithm 1. The fine correction steps in Algorithm 1 are deployed by variable metric methods, and an Armijo rule is used as stepsize strategy for both fine and coarse correction steps. We emphasize that Algorithm 1 is the basic version of GAMA, but the general techniques of analysis in this section could be applied to its variants which we introduced in Section 2. The results in this section will be used in Section 3 to compare the complexity between GAMA and Newton's method.

We will first show that Algorithm 1 achieves a sublinear rate of convergence. We then analyze the maximum number of coarse correction steps that would be taken by Algorithm 1, and the condition that when the coarse correction steps yield quadratic reduction in the gradients in the subspace. At the end of this section, we will provide the composite convergence rate for the coarse correction steps.

To provide convergence properties when coarse correction step is used, the following quantity will be used

$$\chi_{H,k} \triangleq [(\mathbf{R}\nabla f_{h,k})^T[\nabla_H^2 f_{h,k}]^{-1}\mathbf{R}\nabla f_{h,k}]^{1/2}.$$

Notice that $\chi_{H,k}$ is analogous to the Newton decrement, which is used to study the convergence of Newton method [2]. In particular, the defined $\chi_{H,k}$ has the following properties.

1. $\nabla f_{h,k}^T\hat{\mathbf{d}}_{h,k} = -\chi_{H,k}^2$.
2. $\hat{\mathbf{d}}_{h,k}^T\nabla^2 f_{h,k}\hat{\mathbf{d}}_{h,k} = \chi_{H,k}^2$.

We omit the proofs of the above properies since these can be done by using direct computation and the definition of $\chi_{H,k}$.

Algorithm 1 GAMA

Input: κ, ϵ, $\rho_1 \in (0, 0.5)$, $\beta_{ls} \in (0, 1)$,
$\mathbf{P} \in \mathbb{R}^{N\times n}$ and $\mathbf{R} \in \mathbb{R}^{N\times n}$ which satisfy Assumption 2 and 3.
Initialization: $\mathbf{x}_{h,0} \in \mathbb{R}^N$
for $k = 0, 1, 2, \ldots$ **do**
Compute the direction

$$\mathbf{d} = \begin{cases} \hat{\mathbf{d}}_{h,k} \text{ in (15)} & \text{if } \|\mathbf{R}\nabla f_{h,k}\| > \kappa\|\nabla f_{h,k}\| \text{ and } \|\mathbf{R}\nabla f_{h,k}\| > \epsilon, \\ \mathbf{d}_{h,k} \text{ in (9)} & \text{otherwise.} \end{cases}$$

Find the smallest $q \in \mathbb{N}$ such that for stepsize $\alpha_{h,k} = \beta_{ls}^q$,

$$f_h(\mathbf{x}_{h,k} + \alpha_{h,k}\mathbf{d}) \leq f_{h,k} + \rho_1\alpha_{h,k}\nabla^T f_{h,k}\mathbf{d}.$$

Update

$$\mathbf{x}_{h,k+1} \triangleq \mathbf{x}_{h,k} + \alpha_{h,k}\mathbf{d}.$$

end for

3.1 The worse case $\mathcal{O}(1/k)$ Convergence

We will show that Algorithm 1 will achieve a sublinear rate of convergence. We will deploy the techniques from [1] and [2]. Starting with the following lemma, we state reduction in function value using coarse correction steps. We would like to clarify that even though GAMA is considered as a special case in [39], we take advantage of this simplification and specification to provide analysis with results that are easier to interpret. In particular, the analysis of stepsizes $\alpha_{h,k}$'s in [39] relies on the maximum number of iterations taken. This result is unfavourable and unnecessary for the settings we consider.

Lemma 9. *The coarse correction step $\hat{\mathbf{d}}_{h,k}$ in Algorithm 1 will lead to reduction in function value*

$$f_{h,k} - f_h(\mathbf{x}_{h,k} + \alpha_{h,k}\hat{\mathbf{d}}_{h,k}) \geq \frac{\rho_1\kappa^2\beta_{ls}\mu_h}{L_h^2}\|\nabla f_{h,k}\|^2,$$

where ρ_1, κ, and β_{ls} are user-defined parameters in Algorithm 1. $_h$ and μ_h are defined in Assumption 1.

Proof. By convexity,

$$\begin{aligned} f(\mathbf{x}_{h,k} + \alpha\hat{\mathbf{d}}_{h,k}) &\leq f_{h,k} + \alpha\langle\nabla f_{h,k}, \hat{\mathbf{d}}_{h,k}\rangle + \frac{L_h}{2}\alpha^2\|\hat{\mathbf{d}}_{h,k}\|^2, \\ &\leq f_{h,k} - \alpha\chi_{H,k}^2 + \frac{L_h}{2\mu_h}\alpha^2\chi_{H,k}^2, \end{aligned}$$

since

$$\mu_h\|\hat{\mathbf{d}}_{h,k}\|^2 \leq \hat{\mathbf{d}}_{h,k}^T\nabla^2 f(x_k)\hat{\mathbf{d}}_{h,k} = \chi_{H,k}^2.$$

Notice that $\hat{\alpha} = \mu_h/L_h$, we have

$$-\hat{\alpha} + \frac{L_h}{2\mu_h}\hat{\alpha}^2 = -\hat{\alpha} + \frac{L_h}{2\mu_h}\frac{\mu_h}{L_h}\hat{\alpha} = -\frac{1}{2}\hat{\alpha},$$

and

$$\begin{aligned} f(\mathbf{x}_{h,k}+\hat{\alpha}\hat{\mathbf{d}}_{h,k}) &\leq f_{h,k}-\frac{\hat{\alpha}}{2}\chi^2_{H,k}, \\ &\leq f_{h,k}+\frac{\hat{\alpha}}{2}\nabla f_{h,k}^T\hat{\mathbf{d}}_{h,k}, \\ &< f_{h,k}+\rho_1\hat{\alpha}\nabla f_{h,k}^T\hat{\mathbf{d}}_{h,k}, \end{aligned}$$

which satisfies the Armijo condition. Therefore, line search will return stepsize $\alpha_{h,k} \geq \hat{\alpha} = (\beta_{ls}\mu_h)/L_h$. Using the fact that

$$\frac{1}{L_h}\|\mathbf{R}\nabla f(x_k)\|^2 \leq (\mathbf{R}\nabla f(x_k))^T[\nabla^2_H f(x_k)]^{-1}\mathbf{R}\nabla f(x_k) = \chi^2_{H,k},$$

we obtain

$$\begin{aligned} f(\mathbf{x}_{h,k}+\alpha_{h,k}\hat{\mathbf{d}}_{h,k})-f_{h,k} &\leq \rho_1\alpha_{h,k}\nabla f_{h,k}^T\hat{\mathbf{d}}_{h,k}, \\ &\leq -\rho_1\hat{\alpha}\chi^2_{H,k}, \\ &\leq -\rho_1\frac{\beta_{ls}\mu_h}{L_h^2}\|\mathbf{R}\nabla f_{h,k}\|^2, \\ &\leq -\frac{\rho_1\kappa^2\beta_{ls}\mu_h}{L_h^2}\|\nabla f_{h,k}\|^2, \end{aligned}$$

as required. ■

Using the result in Lemma 9, we derive the guaranteed reduction in function value in the following two lemmas.

Lemma 10. *Let* $\Lambda \triangleq \min\left\{\Lambda_h, \frac{\rho_1\kappa^2\beta_{ls}\mu_h}{L_h^2}\right\}$*, then the step* $\mathbf{d}$ *in Algorithm 1 will lead to*

$$f_{h,k}-f_{h,k+1} \geq \Lambda\|\nabla f_{h,k}\|^2,$$

where ρ_1*,* κ*, and* β_{ls} *are user-defined parameters in Algorithm 1.* $_h$ *and* μ_h *are defined in Assumption 1.* Λ_h *is defined in Assumption 6.*

Proof. This is a direct result from Lemma 9 and Assumption 6. ■

Lemma 11. *Suppose*

$$\mathcal{R}(\mathbf{x}_{h,0}) \triangleq \max_{\mathbf{x}_h\in\mathbb{R}^N}\{\|\mathbf{x}_h-\mathbf{x}_{h,\star}\| : f_h(\mathbf{x}_h) \leq f_h(\mathbf{x}_{h,0})\},$$

the step in Algorithm 1 will guarantee

$$f_{h,k}-f_{h,k+1} \geq \frac{\Lambda}{\mathcal{R}^2(\mathbf{x}_{h,0})}\left(f_{h,k}-f_{h,\star}\right)^2,$$

where Λ *is defined in Lemma 10.*

Proof. By convexity, for $k = 0, 1, 2, \ldots,$

$$\begin{aligned} f_{h,k} - f_{h,\star} &\leq \langle \nabla f_{h,k}, \mathbf{x}_{h,k} - \mathbf{x}_{h,\star} \rangle, \\ &\leq \|\nabla f_{h,k}\| \, \|\mathbf{x}_{h,k} - \mathbf{x}_{h,\star}\|, \\ &\leq \mathcal{R}(\mathbf{x}_{h,0}) \|\nabla f_{h,k}\|. \end{aligned}$$

Using Lemma 10, we have

$$\begin{aligned} f_{h,k} - f_{h,\star} &\leq \mathcal{R}(\mathbf{x}_{h,0}) \sqrt{\Lambda^{-1} \left(f_{h,k} - f_{h,k+1} \right)}, \\ \left(\frac{f_{h,k} - f_{h,\star}}{\mathcal{R}(\mathbf{x}_{h,0})} \right)^2 &\leq \Lambda^{-1} \left(f_{h,k} - f_{h,k+1} \right), \\ \Lambda \left(\frac{f_{h,k} - f_{h,\star}}{\mathcal{R}(\mathbf{x}_{h,0})} \right)^2 &\leq f_{h,k} - f_{h,k+1}, \end{aligned}$$

as required. ■

The constant Λ in Lemma 11 depends on Λ_h, which is introduced in Assumption 6. This constant depends on both fine correction step chosen and the user-defined parameter ρ_1 in Armijo rule. For instance,

$$\Lambda_h = \begin{cases} \dfrac{\rho_1 \mu_h}{L_h^2} & \text{if} \quad \mathbf{d}_{h,k} = -[\nabla^2 f_{h,k}]^{-1} \nabla f_{h,k}, \\ \dfrac{\rho_1}{L_h} & \text{if} \quad \mathbf{d}_{h,k} = -\nabla f_{h,k}. \end{cases}$$

In order to derive the convergence rate in this section, we use the following lemma on nonnegative scalar sequences.

Lemma 12. *[1] Let $\{A_k\}_{k \geq 0}$ be a nonnegative sequence of the real numbers satisfying*

$$A_k - A_{k+1} \geq \gamma A_k^2, \quad k = 0, 1, 2, \ldots,$$

and

$$A_0 \leq \frac{1}{q\gamma}$$

for some positive γ and q. Then

$$A_k \leq \frac{1}{\gamma(k+q)}, \quad k = 0, 1, 2, \ldots,$$

and so

$$A_k \leq \frac{1}{\gamma k}, \quad k = 0, 1, 2, \ldots.$$

Proof. see Lemma 3.5 in [1]. ■

Combining the above results, we obtain the rate of convergence.

Theorem 13. *Let $\{\mathbf{x}_k\}_{k \geq 0}$ be the sequence that is generated by Algorithm 1. Then,*

$$f_{h,k} - f_{h,\star} \leq \frac{\mathcal{R}^2(\mathbf{x}_{h,0})}{\Lambda} \frac{1}{2+k},$$

where Λ and $\mathcal{R}(\cdot)$ are defined as in Lemma 10 and 11, respectively.

Proof. Notice that

$$f_{h,k} - f_{h,k+1} \geq \frac{\Lambda}{\mathcal{R}^2(\mathbf{x}_{h,0})} \left(f_{h,k} - f_{h,\star}\right)^2.$$

and so

$$(f_{h,k} - f_{h,\star}) - (f_{h,k+1} - f_{h,\star}) \geq \frac{\Lambda}{\mathcal{R}^2(\mathbf{x}_{h,0})} \left(f_{h,k} - f_{h,\star}\right)^2.$$

Also, we have

$$\begin{aligned} f_{h,0} - f_{h,\star} \leq \frac{L_h}{2}\|\mathbf{x}_{h,0} - \mathbf{x}_{h,\star}\|^2 \leq \frac{L_h}{2}\mathcal{R}^2(\mathbf{x}_{h,0}) \leq \frac{L_h^2\mathcal{R}^2(\mathbf{x}_{h,0})}{2\mu_h} &\leq \frac{L_h^2\mathcal{R}^2(\mathbf{x}_{h,0})}{2\mu_h\beta_{ls}\kappa^2\rho_1}, \\ &\leq \frac{\mathcal{R}^2(\mathbf{x}_{h,0})}{2\Lambda}. \end{aligned}$$

Let's $A_k \triangleq f_{h,k} - f_{h,\star}$, $\gamma \triangleq \dfrac{\Lambda}{R^2(\mathbf{x}_{h,0})}$, and $q \triangleq 2$. By applying Lemma 12, we have

$$f_{h,k} - f_{h,\star} \leq \frac{\mathcal{R}^2(\mathbf{x}_{h,0})}{\Lambda}\frac{1}{2+k},$$

as required. ■

Theorem 13 provides the sublinear convergence of Algorithm 1. We emphasize that the rate is inversely proportional to $\Lambda = \min\{\Lambda_h, \rho_1\kappa^2\mu_h/L_h^2\}$, and so small κ would result in low convergence. Therefore, even though κ could be arbitrary small, it is not desirable in terms of worse case complexity. Note that κ is a user-defined parameter for determining whether coarse correction step would be used. If κ is chosen to be too large, then it is less likely that the coarse correction step would be used. In the extreme case where $\kappa \geq \|\mathbf{R}\|$, coarse correction step would not be deployed because

$$\|\mathbf{R}\nabla f_{h,k}\| \leq \|\mathbf{R}\|\|\nabla f_{h,k}\|,$$

and so Algorithm 1 reduces to the standard variable metric method. Therefore, there is a trade-off between the worse case complexity and the likelihood that coarse correction step would be deployed.

Bear in mind that one can deploy GAMA without using any fine correction step, as stated in Section 2.2. In this case the criterion (8) would not be used, but we clarify that the analysis in this section is still valid as long as we assume there are constants κ, ϵ such that criterion (8) is always satisfied.

3.2 Maximum Number of Iterations of Coarse Correction Step

We now discuss the maximum number of coarse correction steps in Algorithm 1. The following lemma will state the sufficient conditions for not taking any coarse correction step.

Lemma 14. *No coarse correction step in Algorithm 1 will be taken when*

$$\|\nabla f_{h,k}\| \leq \frac{\epsilon}{\omega},$$

where $\omega = \max\{\|\mathbf{P}\|, \|\mathbf{R}\|\}$*, and* ϵ *is a user-defined parameter in Algorithm 1.*

Proof. Recall that in Algorithm 1, the coarse step is only taken when $\|\mathbf{R}\nabla f_{h,k}\| > \epsilon$. We have,

$$\|\mathbf{R}\nabla f_{h,k}\| \leq \omega\|\nabla f_{h,k}\| \leq \omega\frac{\epsilon}{\omega} = \epsilon,$$

and so no coarse correction step will be taken. ■

The above lemma states the condition when the coarse correction step would not be performed. We then investigate the maximum number of iterations to achieve that sufficient condition.

Lemma 15. *Let $\{\mathbf{x}_k\}_{k\geq 0}$ be a sequence generated by Algorithm 1. Then, $\forall\bar{\epsilon}, \bar{k} > 0$ such that,*

$$\bar{k} \geq \left(\frac{1}{\bar{\epsilon}}\right)^2 \frac{\mathcal{R}^2(\mathbf{x}_{h,0})}{\Lambda^2} - 2,$$

we obtain

$$\|\nabla f_h(\mathbf{x}_{h,\bar{k}})\| \leq \bar{\epsilon},$$

where Λ and $\mathcal{R}(\cdot)$ are defined as in Lemma 10 and 11, respectively.

Proof. We know that

$$\Lambda\|\nabla f_{h,k}\|^2 \quad \leq \quad f_{h,k} - f_{h,k+1}.$$

Also, we have,

$$f_{h,k} - f_{h,\star} \leq \frac{\mathcal{R}^2(\mathbf{x}_{h,0})}{\Lambda}\frac{1}{2+k}.$$

Therefore,

$$\begin{aligned}\|\nabla f_{h,k}\|^2 &\leq \frac{1}{\Lambda}\left(f_{h,k} - f_{h,k+1}\right), \\ &\leq \frac{1}{\Lambda}\left(f_{h,k} - f_{h,\star}\right), \\ &\leq \frac{\mathcal{R}^2(\mathbf{x}_{h,0})}{\Lambda^2}\frac{1}{2+k}.\end{aligned}$$

For

$$k = \left(\frac{1}{\bar{\epsilon}}\right)^2 \frac{\mathcal{R}^2(\mathbf{x}_{h,0})}{\Lambda^2} - 2,$$

we have

$$\|\nabla f_{h,k}\| \leq \sqrt{\frac{\mathcal{R}^2(\mathbf{x}_{h,0})}{\Lambda^2}\frac{1}{2+k}} \leq \sqrt{\frac{\mathcal{R}^2(\mathbf{x}_{h,0})}{\Lambda^2}\left(\bar{\epsilon}\right)^2\frac{\Lambda^2}{\mathcal{R}^2(\mathbf{x}_{h,0})}} = \bar{\epsilon},$$

as required. ■

By integrating the above results, we obtain the maximum number of iterations to achieve $\|\nabla f_{h,k}\| \leq \epsilon/\omega$. That is, no coarse correction step will be taken after

$$\left(\frac{\omega}{\epsilon}\right)^2 \frac{\mathcal{R}^2(\mathbf{x}_{h,0})}{\Lambda^2} - 2 \quad \text{iterations.}$$

Notice that the smaller ϵ, the more coarse correction step will be taken. Depending on the choice of $\mathbf{d}_{h,k}$, the choice of ϵ could be different. For example, if $\mathbf{d}_{h,k}$ is chosen as the Newton step where $\mathbf{d}_{h,k} = -[\nabla^2 f_{h,k}]^{-1}\nabla f_{h,k}$, one good choice of ϵ could be $3\omega(1-2\rho_1)\mu_h^2/L_h$ if μ_h and L_h are known. This is because Newton's method achieves quadratic rate of convergence when $\|\nabla f_{h,k}\| \leq 3(1-2\rho_1)\mu_h^2/L_h$ [2]. Therefore, for such ϵ, no coarse correction step would be taken when the Newton method performs in its quadratically convergent phase.

3.3 Quadratic Phase in Subspace

We now state the required condition for stepsize $\alpha_{h,k} = 1$, and then we will show that when $\|\mathbf{R}\nabla f_{h,k}\|$ is sufficiently small, the coarse correction step would reduce $\|\mathbf{R}\nabla f_{h,k}\|$ quadratically. The results below are analogous to the analysis of the Newton's method in [2].

Lemma 16. *Suppose coarse correction step $\hat{\mathbf{d}}_{h,k}$ in Algorithm 1 is taken, then $\alpha_{h,k} = 1$ when*

$$\|\mathbf{R}\nabla f_{h,k}\| \leq \eta = \frac{3\mu_h^2}{M_h}(1 - 2\rho_1),$$

where ρ_1 is an user-defined parameter in Algorithm 1. M_h and μ_h are defined in Assumption 1.

Proof. By Lipschitz continuity (3),

$$\|\nabla^2 f_h(\mathbf{x}_{h,k} + \alpha\hat{\mathbf{d}}_{h,k}) - \nabla^2 f_{h,k}\| \leq \alpha M_h \|\hat{\mathbf{d}}_{h,k}\|,$$

which implies

$$\|\hat{\mathbf{d}}_{h,k}^T(\nabla^2 f_h(\mathbf{x}_{h,k} + \alpha\hat{\mathbf{d}}_{h,k}) - \nabla^2 f_{h,k})\hat{\mathbf{d}}_{h,k}\| \leq \alpha M_h \|\hat{\mathbf{d}}_{h,k}\|^3.$$

Let $\tilde{f}(\alpha) = f_h(\mathbf{x}_{h,k} + \alpha\hat{\mathbf{d}}_{h,k})$, then the above inequality can be rewritten as

$$|\tilde{f}''(\alpha) - \tilde{f}''(0)| \leq \alpha M_h \|\hat{\mathbf{d}}_{h,k}\|^3,$$

and so

$$\tilde{f}''(\alpha) \leq \tilde{f}''(0) + \alpha M_h \|\hat{\mathbf{d}}_{h,k}\|^3.$$

Since $\tilde{f}''(0) = \hat{\mathbf{d}}_{h,k}^T \nabla^2 f_{h,k} \hat{\mathbf{d}}_{h,k} = \chi_{H,k}^2$,

$$\tilde{f}''(\alpha) \leq \chi_{H,k}^2 + \alpha M_h \|\hat{\mathbf{d}}_{h,k}\|^3.$$

By integration,

$$\tilde{f}'(\alpha) \leq \tilde{f}'(0) + \alpha\chi_{H,k}^2 + (\alpha^2/2) M_h \|\hat{\mathbf{d}}_{h,k}\|^3.$$

Similarly, $\tilde{f}'(0) = \nabla f_{h,k}^T \hat{\mathbf{d}}_{h,k} = -\chi_{H,k}^2$, and so

$$\tilde{f}'(\alpha) \leq -\chi_{H,k}^2 + \alpha\chi_{H,k}^2 + (\alpha^2/2) M_h \|\hat{\mathbf{d}}_{h,k}\|^3.$$

Integrating the above inequality, we obtain

$$\tilde{f}(\alpha) \leq \tilde{f}(0) - \alpha\chi_{H,k}^2 + (\alpha^2/2)\chi_{H,k}^2 + (\alpha^3/6) M_h \|\hat{\mathbf{d}}_{h,k}\|^3.$$

Recall that $\mu_h \|\hat{\mathbf{d}}_{h,k}\|^2 \leq \hat{\mathbf{d}}_{h,k}^T \nabla^2 f_{h,k} \hat{\mathbf{d}}_{h,k} = \chi_{H,k}^2$; thus,

$$\tilde{f}(\alpha) \leq \tilde{f}(0) - \alpha\chi_{H,k}^2 + \frac{\alpha^2}{2}\chi_{H,k}^2 + \frac{\alpha^3 M_h}{6\mu_h^{3/2}}\chi_{H,k}^3.$$

Let $\alpha = 1$,

$$\begin{aligned}
\tilde{f}(1) - \tilde{f}(0) &\leq -\chi_{H,k}^2 + \frac{1}{2}\chi_{H,k}^2 + \frac{M_h}{6\mu_h^{3/2}}\chi_{H,k}^3, \\
&\leq -\left(\frac{1}{2} - \frac{M_h}{6\mu_h^{3/2}}\chi_{H,k}\right)\chi_{H,k}^2.
\end{aligned}$$

Using the fact that

$$\|\mathbf{R}\nabla f_{h,k}\| \leq \eta = \frac{3\mu_h^2}{M_h}(1-2\rho_1),$$

and

$$\chi_{H,k} = ((\mathbf{R}\nabla f_{h,k})^T[\nabla_H^2 f_{h,k}]^{-1}\mathbf{R}\nabla f_{h,k})^{1/2} \leq \frac{1}{\sqrt{\mu_h}}\|\mathbf{R}\nabla f_{h,k}\|,$$

we have

$$\chi_{H,k} \leq \frac{3\mu_h^{3/2}}{M_h}(1-2\rho_1) \quad \Longleftrightarrow \quad \rho_1 \leq \frac{1}{2} - \frac{M_h}{6\mu_h^{3/2}}\chi_{H,k}.$$

Therefore,

$$\tilde{f}(1) - \tilde{f}(0) \leq -\rho_1\chi_{H,k}^2 = \rho_1\nabla f_{h,k}^T\hat{\mathbf{d}}_{h,k},$$

and we have $\alpha_{h,k} = 1$ when $\|\mathbf{R}\nabla f_{h,k}\| \leq \eta$. ■

The above lemma yields the following theorem.

Theorem 17. *Suppose the coarse correction step $\hat{\mathbf{d}}_{h,k}$ in Algorithm 1 is taken and $\alpha_{h,k} = 1$, then*

$$\|\mathbf{R}\nabla f_{h,k+1}\| \leq \frac{\omega^3\xi^4 M_h}{2\mu_h^2}\|\mathbf{R}\nabla f_{h,k}\|^2,$$

where M_h and μ_h are defined in Assumption 1, $\omega = \max\{\|\mathbf{P}\|, \|\mathbf{R}\|\}$ and $\xi = \|\mathbf{P}^+\|$.

Proof. Since $\alpha_{h,k} = 1$, we have

$$\begin{aligned}
\|\mathbf{R}\nabla f_{h,k+1}\| &= \|\mathbf{R}\nabla f_h(\mathbf{x}_{h,k} + \hat{\mathbf{d}}_{h,k}) - \mathbf{R}\nabla f_{h,k} - \mathbf{R}\nabla^2 f_{h,k}\mathbf{P}\tilde{\mathbf{d}}_{H,i^\star}\| \\
&\leq \|\mathbf{R}\|\,\|\nabla f_h(\mathbf{x}_{h,k} + \hat{\mathbf{d}}_{h,k}) - \nabla f_{h,k} - \nabla^2 f_{h,k}\hat{\mathbf{d}}_{h,k}\| \\
&\leq \omega\left\|\int_0^1 (\nabla^2 f_h(\mathbf{x}_{h,k} + t\hat{\mathbf{d}}_{h,k}) - \nabla^2 f_{h,k})\hat{\mathbf{d}}_{h,k}\,dt\right\| \\
&\leq \omega\frac{M_h}{2}\|\hat{\mathbf{d}}_{h,k}\|^2.
\end{aligned}$$

Notice that

$$\begin{aligned}
\|\hat{\mathbf{d}}_{h,k}\| &= \|\mathbf{P}[\mathbf{R}\nabla^2 f_{h,k}\mathbf{P}]^{-1}\mathbf{R}\nabla f_{h,k}\| \\
&\leq \|\mathbf{P}\|\,\|[\mathbf{R}\nabla^2 f_{h,k}\mathbf{P}]^{-1}\|\,\|\mathbf{R}\nabla f_{h,k}\| \\
&\leq \frac{\omega\xi^2}{\mu_h}\|\mathbf{R}\nabla f_{h,k}\|.
\end{aligned}$$

Thus,

$$\|\mathbf{R}\nabla f_{h,k+1}\| \leq \frac{\omega^3\xi^4 M_h}{2\mu_h^2}\|\mathbf{R}\nabla f_{h,k}\|^2,$$

as required. ■

The above theorem states the quadratic convergence of $\|\nabla f_{h,k}\|$ within the subspace range$(\mathbf{R})$. However, it does not give insight on the convergence behaviour on the full space $\mathbb{R}^N$. To address this, we study the composite rate of convergence in the next section.

3.4 Composite Convergence Rate

At the end of this section, we study the convergence properties of the coarse correction step when incumbent is sufficiently close to the solution. In particular, we deploy the idea of composite convergence rate in [7], and consider the convergence of coarse correction step as a combination of linear and quadratic convergence.

The reason of proving composite convergence is due to the broadness of GAMA. Suppose in the naive case when $\mathbf{P} = \mathbf{R} = \mathbf{I}$, then the coarse correction step in GAMA becomes Newton's method. In such case we expect quadratic convergence when incumbent is sufficiently close to the solution. On the other hand, suppose $\mathbf{P}$ is any column of $\mathbf{I}$ and $\mathbf{R} = \mathbf{P}^T$, then the coarse correction step is a (weighted) coordinate descent direction, as described in Section 2.4. One should expect not more than linear convergence in that case. Therefore, both quadratic convergence and linear convergence are not suitable for GAMA, and one needs the combination of them. In this paper, we propose to use composite convergence, and show that it can better explain the convergence of different variants of GAMA.

We would like to emphasize the difference between our setting compared to [7]. To the best of our knowledge, composite convergence rate was used in [7] to study subsample Newton methods for machine learning problems without dimensional reduction. In this paper, the class of problems that we consider is not restricted to machine learning, and we focus on the Galerkin model, which is a reduced dimension model. The results presented in this section are not direct results of the approach in [7]. In particular, if the exact analysis of [7] is taken, the derived composite rate would not be useful in our setting, because the coefficient of the linear component would be greater than 1.

Theorem 18. *Suppose the coarse correction step $\hat{\mathbf{d}}_{h,k}$ in Algorithm 1 is taken and $\alpha_{h,k} = 1$, then*

$$\|\mathbf{x}_{h,k+1} - \mathbf{x}_{h,\star}\| \leq \|\mathbf{I} - \mathbf{P}[\nabla_H^2 f_{h,k}]^{-1}\mathbf{R}\nabla^2 f_{h,k}\|\|(\mathbf{I} - \mathbf{PR})(\mathbf{x}_{h,k} - \mathbf{x}_{h,\star})\| + \frac{M_h\omega^2\xi^2}{2\mu_h}\|\mathbf{x}_{h,k} - \mathbf{x}_{h,\star}\|^2, \quad (17)$$

where M_h and μ_h are defined in Assumption 1, $\omega = \max\{\|\mathbf{P}\|, \|\mathbf{R}\|\}$ and $\xi = \|\mathbf{P}^+\|$. The operator ∇_H^2 is defined in (13).

Proof. Denote

$$\tilde{\mathbf{Q}} = \int_0^1 \nabla^2 f(\mathbf{x}_{h,\star} - t(\mathbf{x}_{h,k} - \mathbf{x}_{h,\star}))\mathrm{d}t,$$

we have

$$\begin{aligned}
\mathbf{x}_{h,k+1} - \mathbf{x}_{h,\star} &= \mathbf{x}_{h,k} - \mathbf{x}_{h,\star} - \mathbf{P}[\nabla_H^2 f_{h,k}]^{-1}\mathbf{R}\nabla f_{h,k}, \\
&= \mathbf{x}_{h,k} - \mathbf{x}_{h,\star} - \mathbf{P}[\nabla_H^2 f_{h,k}]^{-1}\mathbf{R}\tilde{\mathbf{Q}}(\mathbf{x}_{h,k} - \mathbf{x}_{h,\star}), \\
&= \left(\mathbf{I} - \mathbf{P}[\nabla_H^2 f_{h,k}]^{-1}\mathbf{R}\tilde{\mathbf{Q}}\right)(\mathbf{x}_{h,k} - \mathbf{x}_{h,\star}), \\
&= \left(\mathbf{I} - \mathbf{P}[\nabla_H^2 f_{h,k}]^{-1}\mathbf{R}\nabla^2 f_{h,k}\right)(\mathbf{x}_{h,k} - \mathbf{x}_{h,\star}) \\
&\quad + \left(\mathbf{P}[\nabla_H^2 f_{h,k}]^{-1}\mathbf{R}\nabla^2 f_{h,k} - \mathbf{P}[\nabla_H^2 f_{h,k}]^{-1}\mathbf{R}\tilde{\mathbf{Q}}\right)(\mathbf{x}_{h,k} - \mathbf{x}_{h,\star}), \\
&= \left(\mathbf{I} - \mathbf{P}[\nabla_H^2 f_{h,k}]^{-1}\mathbf{R}\nabla^2 f_{h,k}\right)(\mathbf{I} - \mathbf{PR})(\mathbf{x}_{h,k} - \mathbf{x}_{h,\star}) \\
&\quad + \mathbf{P}[\nabla_H^2 f_{h,k}]^{-1}\mathbf{R}\left(\nabla^2 f_{h,k} - \tilde{\mathbf{Q}}\right)(\mathbf{x}_{h,k} - \mathbf{x}_{h,\star}).
\end{aligned}$$

Note that

$$\|\nabla^2 f_{h,k} - \tilde{\mathbf{Q}}\| = \left\| \nabla^2 f_{h,k} - \int_0^1 \nabla^2 f(\mathbf{x}_{h,\star} - t(\mathbf{x}_{h,k} - \mathbf{x}_{h,\star}))\mathrm{d}t \right\| \leq \frac{M_h}{2}\|\mathbf{x}_{h,k} - \mathbf{x}_{h,\star}\|.$$

Therefore,

$$\begin{aligned}
\|\mathbf{x}_{h,k+1} - \mathbf{x}_{h,\star}\| &\leq \|\mathbf{I} - \mathbf{P}[\nabla_H^2 f_{h,k}]^{-1}\mathbf{R}\nabla^2 f_{h,k}\|\|(\mathbf{I} - \mathbf{P}\mathbf{R})(\mathbf{x}_{h,k} - \mathbf{x}_{h,\star})\| \\
&\quad + \|\mathbf{P}[\nabla_H^2 f_{h,k}]^{-1}\mathbf{R}\|\frac{M_h}{2}\|\mathbf{x}_{h,k} - \mathbf{x}_{h,\star}\|^2, \\
&\leq \|\mathbf{I} - \mathbf{P}[\nabla_H^2 f_{h,k}]^{-1}\mathbf{R}\nabla^2 f_{h,k}\|\|(\mathbf{I} - \mathbf{P}\mathbf{R})(\mathbf{x}_{h,k} - \mathbf{x}_{h,\star})\| \\
&\quad + \frac{M_h\omega^2\xi^2}{2\mu_h}\|\mathbf{x}_{h,k} - \mathbf{x}_{h,\star}\|^2,
\end{aligned}$$

as required. ■

Theorem 18 provides the composite convergence rate for the coarse correction step. However, some terms remain unclear, and in particular $\|\mathbf{I} - \mathbf{P}[\nabla_H^2 f_{h,k}]^{-1}\mathbf{R}\nabla^2 f_{h,k}\|$. Notice that in the case when $\mathrm{rank}(\mathbf{P}) = N$ (i.e. $\mathbf{P}$ is invertible),

$$\begin{aligned}
\|\mathbf{I} - \mathbf{P}[\nabla_H^2 f_{h,k}]^{-1}\mathbf{R}\nabla^2 f_{h,k}\| &= \|\mathbf{I} - \mathbf{P}[\mathbf{R}\nabla^2 f_{h,k}\mathbf{P}]^{-1}\mathbf{R}\nabla^2 f_{h,k}\|, \\
&= \|\mathbf{I} - \mathbf{P}\mathbf{P}^{-1}[\nabla^2 f_{h,k}]^{-1}\mathbf{R}^{-1}\mathbf{R}\nabla^2 f_{h,k}\|, \\
&= 0.
\end{aligned}$$

It is intuitive to consider that $\|\mathbf{I} - \mathbf{P}[\nabla_H^2 f_{h,k}]^{-1}\mathbf{R}\nabla^2 f_{h,k}\|$ should be small and less than 1 when $\mathrm{rank}(\mathbf{P})$ is close to but not equal to N. However, the above intuition is not true, and we prove this in the following lemma.

Lemma 19. *Suppose rank*$(\mathbf{P}) \neq N$*, then*

$$1 \leq \|\mathbf{I} - \mathbf{P}[\nabla_H^2 f_{h,k}]^{-1}\mathbf{R}\nabla^2 f_{h,k}\| \leq \sqrt{\frac{L_h}{\mu_h}},$$

where L_h and μ_h are defined in Assumption 1. The operator ∇_H^2 is defined in (13).

Proof. Since $\nabla^2 f_{h,k}$ is a positive definite matrix, consider the eigendecomposition of $\nabla^2 f_{h,k}$,

$$\nabla^2 f_{h,k} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^T,$$

where $\mathbf{\Sigma}$ is a diagonal matrix containing the eigenvalues of $\nabla^2 f_{h,k}$, and $\mathbf{U}$ is a orthogonal matrix where its columns are eigenvectors of $\nabla^2 f_{h,k}$. We then have

$$\begin{aligned}
& \mathbf{I} - \mathbf{P}[\nabla_H^2 f_{h,k}]^{-1}\mathbf{R}\nabla^2 f_{h,k} \\
= \; & \mathbf{I} - \mathbf{P}[\mathbf{R}\nabla^2 f_{h,k}\mathbf{P}]^{-1}\mathbf{R}\nabla^2 f_{h,k}, \\
= \; & \mathbf{U}\mathbf{\Sigma}^{-1/2}\mathbf{\Sigma}^{1/2}\mathbf{U}^T - \mathbf{U}\mathbf{\Sigma}^{-1/2}\mathbf{\Sigma}^{1/2}\mathbf{U}^T\mathbf{P}[\mathbf{R}\mathbf{U}\mathbf{\Sigma}^{1/2}\mathbf{\Sigma}^{1/2}\mathbf{U}^T\mathbf{P}]^{-1}\mathbf{R}\mathbf{U}\mathbf{\Sigma}^{1/2}\mathbf{\Sigma}^{1/2}\mathbf{U}^T, \\
= \; & \mathbf{U}\mathbf{\Sigma}^{-1/2}\mathbf{\Sigma}^{1/2}\mathbf{U}^T \\
& -\mathbf{U}\mathbf{\Sigma}^{-1/2}(\mathbf{\Sigma}^{1/2}\mathbf{U}^T\mathbf{P})[(\mathbf{\Sigma}^{1/2}\mathbf{U}^T\mathbf{P})^T(\mathbf{\Sigma}^{1/2}\mathbf{U}^T\mathbf{P})]^{-1}(\mathbf{\Sigma}^{1/2}\mathbf{U}^T\mathbf{P})^T\mathbf{\Sigma}^{1/2}\mathbf{U}^T, \\
= \; & \mathbf{U}\mathbf{\Sigma}^{-1/2}(\mathbf{I} - \mathbf{\Gamma}_{\mathbf{\Sigma}^{1/2}\mathbf{U}^T\mathbf{P}})\mathbf{\Sigma}^{1/2}\mathbf{U}^T,
\end{aligned}$$

where $\mathbf{\Gamma}_{\mathbf{\Sigma}^{1/2}\mathbf{U}^T\mathbf{P}}$ is the orthogonal projection operator onto the range of $\mathbf{\Sigma}^{1/2}\mathbf{U}^T\mathbf{P}$, and so

$$\begin{aligned}\|\mathbf{I}-\mathbf{P}[\nabla^2_H f_{h,k}]^{-1}\mathbf{R}\nabla^2 f_{h,k}\| &= \|\mathbf{U}\mathbf{\Sigma}^{-1/2}(\mathbf{I}-\mathbf{\Gamma}_{\mathbf{\Sigma}^{1/2}\mathbf{U}^T\mathbf{P}})\mathbf{\Sigma}^{1/2}\mathbf{U}^T\|,\\ &= \|\mathbf{\Sigma}^{-1/2}(\mathbf{I}-\mathbf{\Gamma}_{\mathbf{\Sigma}^{1/2}\mathbf{U}^T\mathbf{P}})\mathbf{\Sigma}^{1/2}\|.\end{aligned}$$

For the upper bound, we have

$$\|\mathbf{\Sigma}^{-1/2}(\mathbf{I}-\mathbf{\Gamma}_{\mathbf{\Sigma}^{1/2}\mathbf{U}^T\mathbf{P}})\mathbf{\Sigma}^{1/2}\| \leq \|\mathbf{\Sigma}^{-1/2}\|\|(\mathbf{I}-\mathbf{\Gamma}_{\mathbf{\Sigma}^{1/2}\mathbf{U}^T\mathbf{P}})\|\|\mathbf{\Sigma}^{1/2}\| \leq \sqrt{\frac{L_h}{\mu_h}},$$

since $\mathbf{I}-\mathbf{\Gamma}_{\mathbf{\Sigma}^{1/2}\mathbf{U}^T\mathbf{P}}$ is an orthogonal projector and $\|(\mathbf{I}-\mathbf{\Gamma}_{\mathbf{\Sigma}^{1/2}\mathbf{U}^T\mathbf{P}})\| \leq 1$. For the lower bound, we have

$$\begin{aligned}\|\mathbf{\Sigma}^{-1/2}(\mathbf{I}-\mathbf{\Gamma}_{\mathbf{\Sigma}^{1/2}\mathbf{U}^T\mathbf{P}})\mathbf{\Sigma}^{1/2}\| &= \|\mathbf{\Sigma}^{-1/2}(\mathbf{I}-\mathbf{\Gamma}_{\mathbf{\Sigma}^{1/2}\mathbf{U}^T\mathbf{P}})(\mathbf{I}-\mathbf{\Gamma}_{\mathbf{\Sigma}^{1/2}\mathbf{U}^T\mathbf{P}})\mathbf{\Sigma}^{1/2}\|,\\ &= \|\mathbf{\Sigma}^{-1/2}(\mathbf{I}-\mathbf{\Gamma}_{\mathbf{\Sigma}^{1/2}\mathbf{U}^T\mathbf{P}})\mathbf{\Sigma}^{1/2}\mathbf{\Sigma}^{-1/2}(\mathbf{I}-\mathbf{\Gamma}_{\mathbf{\Sigma}^{1/2}\mathbf{U}^T\mathbf{P}})\mathbf{\Sigma}^{1/2}\|,\\ &\leq \|\mathbf{\Sigma}^{-1/2}(\mathbf{I}-\mathbf{\Gamma}_{\mathbf{\Sigma}^{1/2}\mathbf{U}^T\mathbf{P}})\mathbf{\Sigma}^{1/2}\|\|\mathbf{\Sigma}^{-1/2}(\mathbf{I}-\mathbf{\Gamma}_{\mathbf{\Sigma}^{1/2}\mathbf{U}^T\mathbf{P}})\mathbf{\Sigma}^{1/2}\|,\\ &= \|\mathbf{\Sigma}^{-1/2}(\mathbf{I}-\mathbf{\Gamma}_{\mathbf{\Sigma}^{1/2}\mathbf{U}^T\mathbf{P}})\mathbf{\Sigma}^{1/2}\|^2.\end{aligned}$$

The assumption rank$(\mathbf{P}) \neq N$ implies

$$\mathbf{I} \neq \mathbf{\Gamma}_{\mathbf{\Sigma}^{1/2}\mathbf{U}^T\mathbf{P}} \quad \text{and} \quad \|\mathbf{\Sigma}^{-1/2}(\mathbf{I}-\mathbf{\Gamma}_{\mathbf{\Sigma}^{1/2}\mathbf{U}^T\mathbf{P}})\mathbf{\Sigma}^{1/2}\| \neq 0.$$

Therefore, $1 \leq \|\mathbf{\Sigma}^{-1/2}(\mathbf{I}-\mathbf{\Gamma}_{\mathbf{\Sigma}^{1/2}\mathbf{U}^T\mathbf{P}})\mathbf{\Sigma}^{1/2}\|$, as required. ■

Lemma 19 clarifies the fact that the term $\|\mathbf{I}-\mathbf{P}[\nabla^2_H f_{h,k}]^{-1}\mathbf{R}\nabla^2 f_{h,k}\|$ is at least 1 when $n < N$. This fact reduces the usefulness of the composite convergence rate in Theorem 18. In Section 5-7, we will investigate different Galerkin models, and show that $\|(\mathbf{I}-\mathbf{PR})(\mathbf{x}_{h,k}-\mathbf{x}_{h,\star})\|$ is sufficiently small in those cases.

4 Complexity Analysis

In this section we will perform the complexity analysis for both the Newton's method and GAMA. Our complexity analysis for Newton's method is a variant of the results in [2, 21, 30]. The main difference is that in this paper we focus on the complexity that yield $\|\mathbf{x}_{h,k}-\mathbf{x}_{h,\star}\| \leq \epsilon_h$ accuracy instead of $\|\nabla f_{h,k}\| \leq \epsilon_h$. This choice is made for simpler comparison with GAMA. At the end of this section, we compare the complexity of Newton's method and GAMA, and we will state the condition for which GAMA has lower complexity.

4.1 Complexity Analysis: Newton's Method

It is known that for Newton's method, the algorithm enters its quadratic convergence phase when $\alpha_{h,k} = 1$, with

$$\|\mathbf{x}_{h,k+1}-\mathbf{x}_{h,\star}\| \leq \frac{M_h}{2\mu_h}\|\mathbf{x}_{h,k}-\mathbf{x}_{h,\star}\|^2.$$

The above equation, however, does not guarantee that $\|\mathbf{x}_{h,k+1}-\mathbf{x}_{h,\star}\|$ is a contraction. To obtain this guarantee, it requires

$$\frac{M_h}{2\mu_h}\|\mathbf{x}_{h,k}-\mathbf{x}_{h,\star}\| < 1 \quad \Longleftrightarrow \|\mathbf{x}_{h,k}-\mathbf{x}_{h,\star}\| < \frac{2\mu_h}{M_h}. \tag{18}$$

Moreover, $\alpha_{h,k} = 1$ when

$$\|\nabla f_{h,k}\| \leq 3(1-2\rho_1)\frac{\mu_h^2}{L_h}. \tag{19}$$

In what follows we will first prove the number of iterations needed to satisfy condition (18)-(19) (called the damped Newton phase), and we will then compute the number of iterations needed in the quadratically convergent phase. To this end, we define the following two variables:

- k_d: The number of iterations in the **d**amped Newton phase.
- k_q: The number of iterations in the **q**uadratically convergent phase.

Thus, the total number of iterations needed is $k_d + k_q$.

Lemma 20. *Suppose Newton's method is performed, the conditions (18)-(19) are satisfied after*

$$k_d \geq \left(\frac{1}{\epsilon_N}\right)^2 \frac{\mathcal{R}^2(\mathbf{x}_{h,0})}{\Lambda_N^2} - 2$$

iterations, where

$$\epsilon_N \triangleq \underbrace{\min\left\{\frac{3}{2}(1-2\rho_1), \delta\right\}}_{\triangleq \eta_N} \frac{2\mu_h^2}{M_h},\ \forall \delta \in (0,1), \qquad \Lambda_N \triangleq \frac{\rho_1 \beta_{ls} \mu_h}{L_h^2}.$$

Note that ρ_1 and β_{ls} are user-defined parameters in Armijo rule as Algorithm 1; M_h, $_h$, and μ_h are defined in Assumption 1; $\mathcal{R}(\cdot)$ is defined in Lemma 11.

Proof. It is known that for Newton's method

$$f_{h,k+1} - f_{h,k} \leq -\Lambda_N \|\nabla f_{h,k}\|^2.$$

Using the above equation together with the proofs of Lemma 11 and Theorem 13, we obtain

$$f_{h,k} - f_{h,\star} \leq \frac{\mathcal{R}^2(\mathbf{x}_{h,0})}{\Lambda_N} \frac{1}{2+k}.$$

Therefore, using the proof of Lemma 15, it takes a finite number of iterations, k_d, to achieve $\|\nabla f_{h,k_d}\| \leq \epsilon_N$ for $\epsilon_N > 0$, and

$$k_d \leq \left(\frac{1}{\epsilon_N}\right)^2 \frac{\mathcal{R}^2(\mathbf{x}_{h,0})}{\Lambda_N^2} - 2.$$

By convexity and the definition of ϵ_N, we obtain

$$\|\mathbf{x}_{h,k_d} - \mathbf{x}_{h,\star}\| \leq \frac{1}{\mu_h}\|\nabla f_{h,k_d}\| \quad \leq \quad \frac{1}{\mu_h}\epsilon_N = \frac{1}{\mu_h}\min\left\{\frac{3}{2}(1-2\rho_1), \delta\right\}\frac{2\mu_h^2}{M_h} < \frac{2\mu_h}{M_h}.$$

So we obtain the desired result. ■

Lemma 20 gives, k_d,the number of iterations required in order to enter the quadratic phase. In the following lemma we derive k_q.

Lemma 21. *Suppose Newton's method is performed and $\|\nabla f_{h,0}\| \leq \epsilon_N$, where ϵ_N is defined in Lemma 20. Then, for ϵ_h and k_q such that*

$$\epsilon_h \in (0,1), \quad \textit{and} \quad k_q \geq \frac{1}{\log 2} \log \left(\frac{\log\left(\frac{M_h \epsilon_h}{2\mu_h}\right)}{\log \eta_N} \right) - 1,$$

we obtain $\|\mathbf{x}_{h,k_q} - \mathbf{x}_{h,\star}\| \leq \epsilon_h$. Note that M_h and η_N are defined in Assumption 1 and Lemma 20, respectively.

Proof. Given that

$$\|\mathbf{x}_{h,0} - \mathbf{x}_{h,\star}\| \leq \frac{1}{\mu_h} \|\nabla f_{h,0}\| \leq \frac{\epsilon_N}{\mu_h} \leq \eta_N \frac{2\mu_h}{M_h},$$

we have

$$\begin{aligned}
\|\mathbf{x}_{h,k+1} - \mathbf{x}_{h,\star}\| &\leq \frac{M_h}{2\mu_h} \|\mathbf{x}_{h,k} - \mathbf{x}_{h,\star}\|^2, \\
&\leq \left(\frac{M_h}{2\mu_h}\right) \left(\frac{M_h}{2\mu_h}\right)^2 \|\mathbf{x}_{h,k-1} - \mathbf{x}_{h,\star}\|^4, \\
&\leq \left(\frac{M_h}{2\mu_h}\right)^{\sum_{j=0}^{k} 2^j} \|\mathbf{x}_{h,0} - \mathbf{x}_{h,\star}\|^{2^{k+1}}, \\
&= \left(\frac{M_h}{2\mu_h}\right)^{2^{k+1}-1} \left(\eta_N \frac{2\mu_h}{M_h}\right)^{2^{k+1}}, \\
&= \frac{2\mu_h}{M_h} \eta_N^{2^{k+1}}.
\end{aligned}$$

To achieve the desired accuracy, we require

$$\begin{aligned}
\frac{2\mu_h}{M_h} \eta_N^{2^{k_q+1}} &\leq \epsilon_h, \\
2^{k_q+1} \log \eta_N &\leq \log\left(\frac{M_h \epsilon_h}{2\mu_h}\right), \\
(k_q+1) \log 2 &\geq \log\left(\frac{\log\left(\frac{M_h \epsilon_h}{2\mu_h}\right)}{\log \eta_N} \right), \\
k_q &\geq \frac{1}{\log 2} \log\left(\frac{\log\left(\frac{M_h \epsilon_h}{2\mu_h}\right)}{\log \eta_N} \right) - 1.
\end{aligned}$$

So we obtain the desired result. ■

Combining the results in Lemma 20 and Lemma 21, we obtain the complexity of Newton's method.

Theorem 22. *Suppose Newton's method is performed and $k = k_d + k_q$, where k_d and k_q are defined as in Lemma 20 and Lemma 21, respectively. Then we obtain the ϵ_h-accuracy $\|\mathbf{x}_{h,k} - \mathbf{x}_{h,\star}\| \leq \epsilon_h$ with the complexity*

$$\mathcal{O}\left((k_d + k_q) N^3\right).$$

Proof. The total complexity is the number of iterations, $k_d + k_q$, multiply by the cost per iteration, which is $\mathcal{O}(N^3)$. ■

4.2 Complexity Analysis: GAMA

We follow the same strategy to compute the complexity of GAMA. In order to avoid unnecessary complications in notations, in the section, we let r_1 and r_2 to be the composite rate in which

$$\|\mathbf{x}_{h,k+1} - \mathbf{x}_{h,\star}\| \leq r_1\|\mathbf{x}_{h,k} - \mathbf{x}_{h,\star}\| + r_2\|\mathbf{x}_{h,k} - \mathbf{x}_{h,\star}\|^2, \tag{20}$$

when

$$\|\mathbf{R}\nabla f_{h,k}\| \leq \frac{3\mu_h^2}{M_h}(1 - 2\rho_1). \tag{21}$$

For $\|\mathbf{x}_{h,k+1} - \mathbf{x}_{h,\star}\|$ in (20) to be a contraction, we need $r_1 < 1$ and

$$\|\mathbf{x}_{h,k} - \mathbf{x}_{h,\star}\| < \frac{1 - r_1}{r_2}. \tag{22}$$

We clarify that the above form in (20) is not exactly in the same form of the composite rate in Section 3.4, where $\|(\mathbf{I} - \mathbf{PR})(\mathbf{x}_{h,k} - \mathbf{x}_{h,\star})\|$ is used instead of $\|\mathbf{x}_{h,k} - \mathbf{x}_{h,\star}\|$. The latter case is used solely for simpler analysis, and does not contradict with the results presented in Section 3.4; in particular, one can simply let

$$r_1 = \|\mathbf{I} - \mathbf{P}[\nabla_H^2 f_{h,k}]^{-1}\mathbf{R}\nabla^2 f_{h,k}\|\frac{\|(\mathbf{I} - \mathbf{PR})(\mathbf{x}_{h,k} - \mathbf{x}_{h,\star})\|}{\|\mathbf{x}_{h,k} - \mathbf{x}_{h,\star}\|}. \tag{23}$$

In order to guarantee convergence, we simply assume 1 fine correction step is taken after a fixed number of coarse correction steps. For the purpose of simplifying analysis, we make the following assumptions on the fine correction step taken.

Assumption 23. *The coarse correction step of Algorithm 1 has the following properties:*

1. 1 *fine correction step is taken for every* K_H *coarse correction steps.*
2. *The computational cost of each fine correction step is* $\mathcal{O}(C_h)$.
3. *When the composite rate (20) applies for the coarse correction steps, the fine correction step satisfies*
$$\|\mathbf{x}_{h,k+1} - \mathbf{x}_{h,\star}\| \leq \|\mathbf{x}_{h,k} - \mathbf{x}_{h,\star}\|.$$

Recall that GAMA only achieves the composite rate when condition (21) is satisfied, as stated in Lemma 16 and Theorem 18. When (21) does not hold, a global sublinear rate of convergence is still guaranteed, as concluded in Theorem 13. We shall call the former case and the latter case as composite convergent phase and sublinear convergent phase, respectively.

In the following lemma, we compute the number of iterations needed for both composite convergent phase and sublinear convergent phase. Similar to the case of Newton's method, we define the following notation:

- k_s: The number of iterations in the **s**ublinear convergent phase.
- k_c: The number of iterations in the **c**omposite convergent phase.

Thus, the total number of iterations of GAMA would be $k_s + k_c$.

Lemma 24. *Suppose Algorithm 1 is performed and Assumption 23 holds, the conditions (21)-(22) are satisfied after*

$$k_s \geq \left(\frac{1}{\epsilon_G}\right)^2 \frac{\mathcal{R}^2(\mathbf{x}_{h,0})}{\Lambda^2} - 2$$

iterations, where

$$\epsilon_G \triangleq \min\left\{\frac{3\mu_h^2}{\omega M_h}(1-2\rho_1), \delta\right\},\ \forall \delta \in \left(0, \frac{\mu_h(1-r_1)}{r_2}\right).$$

Note that ρ_1 is a user-defined parameter in Algorithm 1; $\omega = \max\{\|\mathbf{P}\|, \|\mathbf{R}\|\}$; M_h and μ_h are defined in Assumption 1; Λ and $\mathcal{R}(\cdot)$ are defined in Lemma 10 and 11, respectively; r_1 and r_2 are defined in (20).

Proof. Using the result in Lemma 15, we obtain

$$\|\nabla f_{h,k_s}\| \leq \epsilon_G.$$

We then show that the above condition is sufficient for $\alpha_{h,k_s} = 1$. By definitions,

$$\|\mathbf{R}\nabla f_{h,k_s}\| \leq \omega\|\nabla f_{h,k_s}\| \leq \omega\epsilon_G \leq \omega\frac{3\mu_h^2}{\omega M_h}(1-2\rho_1) = \frac{3\mu_h^2}{M_h}(1-2\rho_1).$$

By Lemma 16, $\alpha_{h,k_s} = 1$. On the other hand,

$$\|\mathbf{x}_{h,k_s} - \mathbf{x}_{h,\star}\| \leq \frac{1}{\mu_h}\|\nabla f_{h,k_s}\| \quad < \quad \frac{1}{\mu_h}\frac{\mu_h(1-r_1)}{r_2} = \frac{1-r_1}{r_2}.$$

Therefore, we obtain the desired result. ■

Lemma 24 gives the number of iterations required in the sublinear convergent phase, k_s. In the following lemma, we derive k_c.

Lemma 25. *Suppose Algorithm 1 is performed, Assumption 23 holds, and*

$$\|\nabla f_{h,0}\| \leq \epsilon_G,$$

where ϵ_G is defined in Lemma 24. Then for ϵ_h and k_c such that

$$\epsilon_h \in (0,1), \quad \textit{and}$$

$$k_c \geq \frac{1+1/K_H}{r_1-1}\left(\log\left(\frac{\mu_h\epsilon_h}{\epsilon_G}\right) - \log\left(r_1 + r_2\frac{\epsilon_G}{\mu_h}\right) + \frac{r_2\epsilon_G}{\mu_h \log\left(r_1 + r_2\frac{\epsilon_G}{\mu_h}\right)}\right),$$

we obtain $\|\mathbf{x}_{h,k_c} - \mathbf{x}_{h,\star}\| \leq \epsilon_h$. Note that μ_h and K_H are defined in Assumption 1 and Assumption 23, respectively; r_1 and r_2 are defined in (20).

Proof. Based on Assumption 23, if k coarse correction steps are needed, k/K_H fine correction steps would be taken. The total number of searches would then be $k(1+1/K_H)$. Therefore, we first neglect the use of fine correction steps, and consider this factor at the end of the proof by multiplying $(1+1/K_H)$.

We obtain

$$\|\mathbf{x}_{h,0} - \mathbf{x}_{h,\star}\| \leq \frac{1}{\mu_h}\|\nabla f_{h,0}\| \leq \frac{\epsilon_G}{\mu_h}.$$

and $\frac{\epsilon_G}{\mu_h} < \frac{1-r_1}{r_2}$, based on the definition of ϵ_G. Since $\|\mathbf{x}_{h,k} - \mathbf{x}_{h,\star}\|$ is a contraction,

$$\|\mathbf{x}_{h,0} - \mathbf{x}_{h,\star}\| \geq \|\mathbf{x}_{h,1} - \mathbf{x}_{h,\star}\| \geq \|\mathbf{x}_{h,2} - \mathbf{x}_{h,\star}\| \geq \dots$$

Based on the above notations, observations, and the fact that composite rate holds, we obtain

$$\begin{aligned}
\|\mathbf{x}_{h,k} - \mathbf{x}_{h,\star}\| &\leq (r_1 + r_2\|\mathbf{x}_{h,k-1} - \mathbf{x}_{h,\star}\|)\, \|\mathbf{x}_{h,k-1} - \mathbf{x}_{h,\star}\|, \\
&\leq \left(r_1 + r_2\frac{\epsilon_G}{\mu_h}\right) \|\mathbf{x}_{h,k-1} - \mathbf{x}_{h,\star}\|, \\
&\leq \left(r_1 + r_2\frac{\epsilon_G}{\mu_h}\right)^k \|\mathbf{x}_{h,0} - \mathbf{x}_{h,\star}\|, \\
&= \left(r_1 + r_2\frac{\epsilon_G}{\mu_h}\right)^k \frac{\epsilon_G}{\mu_h}.
\end{aligned}$$

We denote $r(k) \triangleq r_1 + r_2\left(r_1 + r_2\frac{\epsilon_G}{\mu_h}\right)^k \frac{\epsilon_G}{\mu_h}$ and we obtain

$$\begin{aligned}
\|\mathbf{x}_{h,k+1} - \mathbf{x}_{h,\star}\| &\leq r_1\|\mathbf{x}_{h,k} - \mathbf{x}_{h,\star}\| + r_2\|\mathbf{x}_{h,k} - \mathbf{x}_{h,\star}\|^2, \\
&\leq (r_1 + r_2\|\mathbf{x}_{h,k} - \mathbf{x}_{h,\star}\|)\, \|\mathbf{x}_{h,k} - \mathbf{x}_{h,\star}\|, \\
&\leq r(k)\|\mathbf{x}_{h,k} - \mathbf{x}_{h,\star}\|, \\
&\leq \left(\prod_{j=0}^{k} r(j)\right) \|\mathbf{x}_{h,0} - \mathbf{x}_{h,\star}\|, \\
&\leq \left(\prod_{j=0}^{k} r(j)\right) \frac{\epsilon_G}{\mu_h}.
\end{aligned}$$

Therefore, it is sufficient to achieve ϵ_h-accuracy when

$$\begin{aligned}
\left(\prod_{j=0}^{k} r(j)\right) \frac{\epsilon_G}{\mu_h} &\leq \epsilon_h, \\
\prod_{j=0}^{k} \left(r_1 + r_2\left(r_1 + r_2\frac{\epsilon_G}{\mu_h}\right)^j \frac{\epsilon_G}{\mu_h}\right) &\leq \frac{\mu_h\epsilon_h}{\epsilon_G}, \\
\sum_{j=0}^{k} \log\left(r_1 + r_2\left(r_1 + r_2\frac{\epsilon_G}{\mu_h}\right)^j \frac{\epsilon_G}{\mu_h}\right) &\leq \log\left(\frac{\mu_h\epsilon_h}{\epsilon_G}\right), \\
\sum_{j=1}^{k+1} \log\left(r_1 + r_2\left(r_1 + r_2\frac{\epsilon_G}{\mu_h}\right)^{j-1} \frac{\epsilon_G}{\mu_h}\right) &\leq \log\left(\frac{\mu_h\epsilon_h}{\epsilon_G}\right).
\end{aligned}$$

Using calculus, we know that

$$\begin{aligned}
&\sum_{j=1}^{k+1} \log\left(r_1 + r_2\left(r_1 + r_2\frac{\epsilon_G}{\mu_h}\right)^{j-1}\frac{\epsilon_G}{\mu_h}\right)\\
\leq\ & \log\left(r_1 + r_2\frac{\epsilon_G}{\mu_h}\right) + \int_1^{k+1} \log\left(r_1 + r_2\frac{\epsilon_G}{\mu_h}\left(r_1 + r_2\frac{\epsilon_G}{\mu_h}\right)^{x-1}\right)\mathrm{d}x,\\
\leq\ & \log\left(r_1 + r_2\frac{\epsilon_G}{\mu_h}\right) + \int_1^{k+1} (r_1 - 1) + r_2\frac{\epsilon_G}{\mu_h}\left(r_1 + r_2\frac{\epsilon_G}{\mu_h}\right)^{x-1}\mathrm{d}x,\\
\leq\ & \log\left(r_1 + r_2\frac{\epsilon_G}{\mu_h}\right) + k(r_1 - 1) + \frac{r_2\frac{\epsilon_G}{\mu_h}}{r_1 + r_2\frac{\epsilon_G}{\mu_h}}\int_1^{k+1}\left(r_1 + r_2\frac{\epsilon_G}{\mu_h}\right)^{x}\mathrm{d}x,\\
\leq\ & \log\left(r_1 + r_2\frac{\epsilon_G}{\mu_h}\right) + k(r_1 - 1) + \frac{r_2\frac{\epsilon_G}{\mu_h}}{r_1 + r_2\frac{\epsilon_G}{\mu_h}}\left(\frac{\left(r_1 + r_2\frac{\epsilon_G}{\mu_h}\right)^{x}}{\log\left(r_1 + r_2\frac{\epsilon_G}{\mu_h}\right)}\right|_{x=1}^{x=k+1},\\
\leq\ & \log\left(r_1 + r_2\frac{\epsilon_G}{\mu_h}\right) + k(r_1 - 1) + \frac{r_2\frac{\epsilon_G}{\mu_h}}{\log\left(r_1 + r_2\frac{\epsilon_G}{\mu_h}\right)}\left(r_1 + r_2\frac{\epsilon_G}{\mu_h}\right)^{k} - \frac{r_2\frac{\epsilon_G}{\mu_h}}{\log\left(r_1 + r_2\frac{\epsilon_G}{\mu_h}\right)},\\
\leq\ & \log\left(r_1 + r_2\frac{\epsilon_G}{\mu_h}\right) + k(r_1 - 1) - \frac{r_2\epsilon_G}{\mu_h\log\left(r_1 + r_2\frac{\epsilon_G}{\mu_h}\right)}.
\end{aligned}$$

So, it is sufficient to achieve ϵ_h-accuracy if

$$\begin{aligned}
\log\left(\frac{\mu_h\epsilon_h}{\epsilon_G}\right) &\geq \log\left(r_1 + r_2\frac{\epsilon_G}{\mu_h}\right) + k(r_1 - 1) - \frac{r_2\epsilon_G}{\mu_h\log\left(r_1 + r_2\frac{\epsilon_G}{\mu_h}\right)},\\
k(r_1 - 1) &\leq \log\left(\frac{\mu_h\epsilon_h}{\epsilon_G}\right) - \log\left(r_1 + r_2\frac{\epsilon_G}{\mu_h}\right) + \frac{r_2\epsilon_G}{\mu_h\log\left(r_1 + r_2\frac{\epsilon_G}{\mu_h}\right)},\\
k &\geq \frac{1}{r_1 - 1}\left(\log\left(\frac{\mu_h\epsilon_h}{\epsilon_G}\right) - \log\left(r_1 + r_2\frac{\epsilon_G}{\mu_h}\right) + \frac{r_2\epsilon_G}{\mu_h\log\left(r_1 + r_2\frac{\epsilon_G}{\mu_h}\right)}\right).
\end{aligned}$$

So we obtain the desired result. ■

Although the result of Lemma 25 states the number of iterations needed for composite convergent phase. The derived result is, however, difficult to interpret. To this end, in the following lemma, we study a special case of Lemma 25 .

Lemma 26. *Consider the setting as in Lemma 25 with*

$$\epsilon_G = \min\left\{\frac{3\mu_h^2}{\omega M_h}(1 - 2\rho_1), \frac{\mu_h(1 - r_1)}{2r_2}\right\}.$$

Then for ϵ_h and k_c such that

$$\epsilon_h \in (0, 1), \quad \textit{and} \quad k_c \geq \frac{1 + 1/K_H}{1 - r_1}\left(\log\left(\frac{3\mu_h}{\omega M_h\epsilon_h}\right) + 1\right),$$

we obtain $\|\mathbf{x}_{h,k_c} - \mathbf{x}_{h,\star}\| \leq \epsilon_h$. Note that M_h and μ_h are defined in Assumption 1; K_H is defined in Assumption 23; r_1 is defined in (20).

Proof. By definition,

$$\epsilon_G \leq \frac{\mu_h(1-r_1)}{2r_2} \Rightarrow r_2\frac{\epsilon_G}{\mu_h} \leq \frac{1-r_1}{2}, \quad \text{and} \quad r_1 + r_2\frac{\epsilon_G}{\mu_h} \leq \frac{1+r_1}{2}.$$

Also,

$$\epsilon_G \leq \frac{3\mu_h^2}{\omega M_h}(1-2\rho_1) \Rightarrow \frac{\epsilon_G}{\mu_h} \leq \frac{3\mu_h}{\omega M_h}(1-2\rho_1) \leq \frac{3\mu_h}{\omega M_h}.$$

Thus, using the results in Lemma 25, it is sufficient when

$$\begin{aligned} k_c &\geq \frac{1+1/K_H}{r_1-1}\left(\log\left(\frac{\omega M_h \epsilon_h}{3\mu_h}\right) - \log\left(\frac{1+r_1}{2}\right) + \frac{1-r_1}{2\log\left(\frac{1+r_1}{2}\right)}\right), \\ &= \frac{1+1/K_H}{1-r_1}\left(\log\left(\frac{3\mu_h}{\omega M_h \epsilon_h}\right) + \log\left(\frac{1+r_1}{2}\right) + \frac{r_1-1}{2\log\left(\frac{1+r_1}{2}\right)}\right). \end{aligned}$$

Since

$$\frac{r_1-1}{2\log\left(\frac{1+r_1}{2}\right)} < 1, \quad \text{and} \quad \log\left(\frac{1+r_1}{2}\right) < 0, \quad \text{for} \quad 0 < r_1 < 1,$$

It is sufficient when

$$k_c \geq \frac{1+1/K_H}{1-r_1}\left(\log\left(\frac{3\mu_h}{\omega M_h \epsilon_h}\right) + 1\right).$$

So we obtain the desired result. ■

Lemma 26 provides a better picture of the convergence when composite rate holds. One can see that the number of iterations required, k_c, is clearly inverse proportional to $1 - r_1$.

Theorem 27. *Suppose Algorithm 1 is perform, Assumption 23 holds, and $k = k_s + k_c$, where k_s and k_c are defined in Lemma 24 and 25. Then we obtain the ϵ_h-accuracy $\|\mathbf{x}_{h,k} - \mathbf{x}_{h,\star}\| \leq \epsilon_h$ with complexity*

$$\mathcal{O}\left(\frac{k_s+k_c}{1+1/K_H}n^3 + \frac{1/K_H(k_s+k_c)}{1+1/K_H}C_h\right),$$

where K_H and C_h are defined in Assumption 23.

Proof. The total complexity is the number of iterations, k_s+k_c, multiply by the cost per iteration. Based on Assumption 23, $\frac{k_s+k_c}{1+1/K_H}$ coarse correction steps and $\frac{1/K_H(k_s+k_c)}{1+1/K_H}$ fine correction steps are taken. The computational cost of each coarse correction step and fine correction step is $\mathcal{O}(n^3)$ and $\mathcal{O}(C_h)$, respectively. ■

4.3 Comparison: Newton v.s. Multilevel

Using the derived complexity results, we now compare the complexity of Newton's method and GAMA. We conclude this section by stating the condition for which GAMA has lower complexity.

Theorem 28. *Suppose Assumption 23 holds, then for sufficiently large enough N, the complexity of Algorithm 1 is lower than the complexity of Newton's method. In particular, if ϵ_G in Lemma 24 is chosen to be*

$$\epsilon_G \triangleq \min\left\{\frac{3\mu_h^2}{\omega M_h}(1-2\rho_1), \frac{\mu_h(1-r_1)}{2r_2}\right\},$$

then the complexity of Algorithm 1 is lower than the complexity of Newton's method when

$$r_1 \leq 1 - \frac{(K_H n^3 + C_h)(1 + 1/K_H)\left(\log\left(\frac{3\mu_h}{\omega M_h \epsilon_h}\right) + 1\right)}{N^3(K_H + 1)(k_d + k_q) - (K_H n^3 + C_h)k_s}, \tag{24}$$

for

$$N^3(K_H + 1)(k_d + k_q) - (K_H n^3 + C_h)k_s > 0.$$

Note that ρ_1 is a user-defined parameter in Algorithm 1; $\omega = \max\{\|\mathbf{P}\|, \|\mathbf{R}\|\}$; M_h and μ_h are defined in Assumption 1; K_H and C_h are defined in Assumption 23; r_1 and r_2 are defined in (20); k_d, k_q, and k_s are defined in Lemma 20, 21, and 24, respectively.

Proof. When the complexity of Algorithm 1 is less than Newton's method, we have

$$\begin{aligned}
\frac{k_s + k_c}{1 + 1/K_H} n^3 + \frac{1/K_H(k_s + k_c)}{1 + 1/K_H} C_h &\leq (k_d + k_q)N^3, \\
k_s n^3 + k_c n^3 + \frac{1}{K_H}(k_s + k_c)C_h &\leq \left(1 + \frac{1}{K_H}\right)(k_d + k_q)N^3, \\
k_s\left(n^3 + \frac{C_h}{K_H}\right) + k_c\left(n^3 + \frac{C_h}{K_H}\right) &\leq \left(1 + \frac{1}{K_H}\right)(k_d + k_q)N^3, \\
\left(1 + \frac{1}{K_H}\right)(k_d + k_q)N^3 - k_s\left(n^3 + \frac{C_h}{K_H}\right) &\geq k_c\left(n^3 + \frac{C_h}{K_H}\right).
\end{aligned}$$

From the first inequality we can see that it satisfies when N is sufficiently large. Using the definition of k_c in Lemma 26, we obtain

$$\begin{aligned}
\frac{1 + 1/K_H}{1 - r_1}\left(\log\left(\frac{3\mu_h}{\omega M_h \epsilon_h}\right) + 1\right) &\leq \left(\frac{(K_H + 1)N^3}{K_H n^3 + C_h}\right)(k_d + k_q) - k_s, \\
\frac{1}{1 - r_1} &\leq \frac{\left(\frac{(K_H+1)N^3}{K_H n^3 + C_h}\right)(k_d + k_q) - k_s}{(1 + 1/K_H)\left(\log\left(\frac{3\mu_h}{\omega M_h \epsilon_h}\right) + 1\right)}, \\
1 - r_1 &\geq \frac{(1 + 1/K_H)\left(\log\left(\frac{3\mu_h}{\omega M_h \epsilon_h}\right) + 1\right)}{\left(\frac{(K_H+1)N^3}{K_H n^3 + C_h}\right)(k_d + k_q) - k_s}, \\
r_1 &\leq 1 - \frac{(1 + 1/K_H)\left(\log\left(\frac{3\mu_h}{\omega M_h \epsilon_h}\right) + 1\right)}{\left(\frac{(K_H+1)N^3}{K_H n^3 + C_h}\right)(k_d + k_q) - k_s},
\end{aligned}$$

as required. ■

Theorem 28 shows that when the dimension of the fine model, N, is sufficiently large, GAMA has lower computational complexity. The condition (24) requires a sufficiently small r_1 in the composite rate (20). This result agrees with the intuition with the following reasoning: when $r_1 << 1$, it implies that GAMA converges with very fast linear rate, which could outperform Newton's method because of the cheaper per-iteration cost.

We shall further study the condition (24). Assume the cost of fine correction step is at most in the same order of the coarse correction step, i.e. $C_h = \mathcal{O}(n^3)$. Then, by fixing all the quantities except N, the condition (24) can be recognized asymptotically as

$$r_1 \leq 1 - \mathcal{O}\left(\frac{1}{N^3}\right).$$

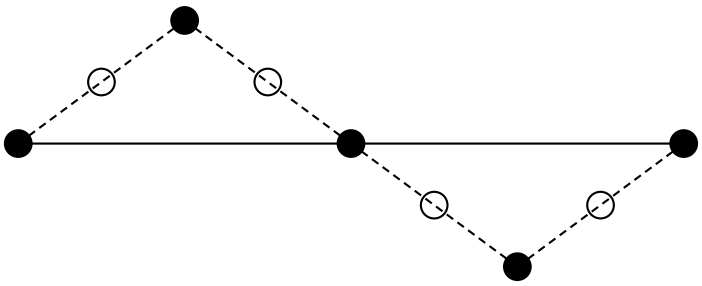

Figure 1: $\mathbf{P}$ in (25)

Thus, as $N \to \infty$, the above condition holds even when $r_1 \approx 1$. This condition is relaxed quickly because N grows in cube.

From equation (23), recall that $r_1 << 1$ is equivalent to

$$\|\mathbf{I} - \mathbf{P}[\nabla_H^2 f_{h,k}]^{-1}\mathbf{R}\nabla^2 f_{h,k}\| \, \|(\mathbf{I} - \mathbf{PR})(\mathbf{x}_{h,k} - \mathbf{x}_{h,\star})\| << \|\mathbf{x}_{h,k} - \mathbf{x}_{h,\star}\|.$$

From the above expression, one can see that a small $\|(\mathbf{I} - \mathbf{PR})(\mathbf{x}_{h,k} - \mathbf{x}_{h,\star})\|$ is equivalent to small r_1. In the following three sections, we will consider three cases of GAMA and derive the $\|(\mathbf{I} - \mathbf{PR})(\mathbf{x}_{h,k} - \mathbf{x}_{h,\star})\|$ for each case. In particular, we show how the magnitude of $\|(\mathbf{I} - \mathbf{PR})(\mathbf{x}_{h,k} - \mathbf{x}_{h,\star})\|$ varies depending on the structure of the problems and the parameters chosen.

5 PDE-based Problems: One-dimensional Case

In this section, we study the Galerkin model that arises from PDE-based problems. We begin with introducing the basic setting, and then we analyze the coarse correction step in this specific case. Building upon the composite rate in Section 3.4, at the end of this section we re-derive the composite rate with a more insightful bound of $\|(\mathbf{I} - \mathbf{PR})(\mathbf{x}_{h,k} - \mathbf{x}_{h,\star})\|$. As mentioned in Section 3, this quantity is critical in analyzing the performance and complexity of GAMA.

Since the conventional multigrid methods were originally developed for solving (non-)linear equations arising from PDEs, most research on multilevel optimization algorithms have been focusing on solving the discretizations of infinite dimensional problems [13, 14, 22, 24, 27, 39]. As mentioned before, the Galerkin model in optimization was first mentioned in [14] and later tested numerically in [13]. We point out that in the theoretical perspective, the Galerkin model has been only considered as one special case of the general multilevel framework, and it has not been shown to have any particular advantage. For the trust-region based multilevel algorithm in [14], it has the same order of complexity bound as pure gradient method. For the line-search based multilevel algorithm in [39], the convergence rate was proven to be sublinear for strongly convex problems, which agrees with our results in Section 3.

For the simplicity of the analysis, we consider specifically the one-dimensional case, i.e. the decision variable of the infinite dimensional problems is a functional in $\mathbb{R}$. We further assume that the decision variable is discretized uniformly over $[0, 1]$ with value 0 on the boundary. We could like to clarify that the approach of analysis in this section could be applied to more general and high-dimensional settings.

5.1 Galerkin Model by One-dimensional Interpolations

For one dimensional problems, we consider the standard linear prolongation operator and restriction operator. Based on the traditional setting in multigrid research, we define the

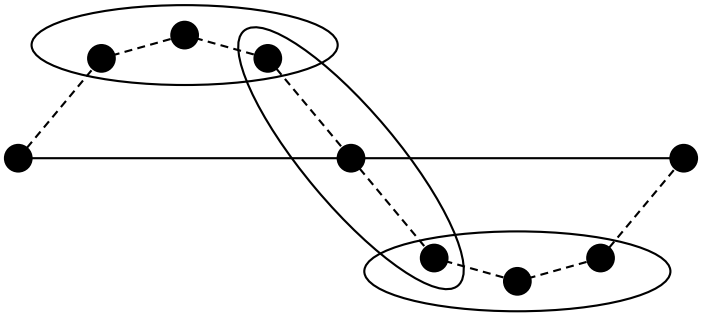

Figure 2: $\mathbf{R}$ in (26)

following Galerkin model.

- N is an even number,
- the (fine) discretized decision variable is in $\mathbf{R}^{N-1}$, and
- the coarse model is in $\mathbf{R}^{N/2-1}$.

For interpolation operator $\mathbf{P} \in \mathbb{R}^{(N-1)\times(N/2-1)}$, we consider

$$\mathbf{P} = \frac{1}{2}\begin{pmatrix} 1 & & & & \\ 2 & & & & \\ 1 & 1 & & & \\ & 2 & & & \\ & 1 & & & \\ & & \ddots & 1 & \\ & & & 2 & \\ & & & 1 & \end{pmatrix}, \tag{25}$$

and the restriction operator

$$\mathbf{R} = \frac{1}{2}\mathbf{P}^T \in \mathbb{R}^{(N/2-1)\times(N-1)}. \tag{26}$$

Notice that the $\mathbf{P}$ and $\mathbf{R}$ in (25) and (26) have geometric meanings, and they are one of the standard pairs of operators in multilevel and multigrid methods [3]. As shown in Figure 1, $\mathbf{P}$ is an interpolation operator such that one point is interpolated linearly between every two points. On the other hand, from Figure 2, $\mathbf{R}$ performs restriction by doing weighted average onto every three points. These two operators assume the boundary condition is zero for both end points. We emphasize that the approach of convergence analysis in this section is not restricted for this specific pair of $\mathbf{P}$ and $\mathbf{R}$. We believe the general approach could be applied to interpolation type operators, especially operators that are designed for PDE-based optimization problems.

5.2 Analysis

With the definitions (25) and (26), we investigate the convergence behaviour of the coarse correction step. The analytical tool we used in this section is Taylor's expansion. To deploy this technique, we consider interpolations over the elements of vectors. In particular, we consider interpolations that are twice differentiable with the following definition.

Definition 29. *For any vector* $\mathbf{r} \in \mathbb{R}^{N-1}$*, we denote* $\mathcal{F}_{\mathbf{r}}^{N-1}$ *to be the set of twice differentiable functions such that* $\forall w \in \mathcal{F}_{\mathbf{r}}^{N-1}$*,*

$$w(0) = w(1) = 0, \quad \textit{and} \quad w_i = w(y_i) = (\mathbf{r})_i,$$

where $y_i = i/N$ *for* $i = 1, 2, \ldots, N-1$*.*

Using the definitions (25) and (26), we can estimate the "information loss" via interpolations using the following proposition.

Proposition 30. *Suppose* $\mathbf{P}$ *and* $\mathbf{R}$ *are defined in (25) and (26), respectively. For any vector* $\mathbf{r}_h \in \mathbb{R}^{N-1}$*, we denote* $(\mathbf{r}_h)_0 = (\mathbf{r}_h)_N = 0$ *and obtain*

$$(\mathbf{PR}\mathbf{r}_h)_j = \begin{cases} \frac{1}{4}((\mathbf{r}_h)_{j-1} + 2(\mathbf{r}_h)_j + (\mathbf{r}_h)_{j+1}) & \textit{if } j \textit{ is even,} \\ \frac{1}{8}((\mathbf{r}_h)_{j-2} + 2(\mathbf{r}_h)_{j-1} + 2(\mathbf{r}_h)_j + 2(\mathbf{r}_h)_{j+1} + (\mathbf{r}_h)_{j+2}) & \textit{if } j \textit{ is odd,} \end{cases}$$

for $j = 1, 2, \ldots, N-1$*.*

Proof. By the definition of $\mathbf{R}$ and $\mathbf{P}$, we have

$$(\mathbf{R}\mathbf{r}_h)_j = \frac{1}{4}((\mathbf{r}_h)_{2j-1} + 2(\mathbf{r}_h)_{2j} + (\mathbf{r}_h)_{2j+1}), \quad 1 \le j \le \frac{n}{2} - 1.$$

So

$$(\mathbf{PR}\mathbf{r}_h)_j = (\mathbf{R}\mathbf{r}_h)_{j/2} = \frac{1}{4}((\mathbf{r}_h)_{j-1} + 2(\mathbf{r}_h)_j + (\mathbf{r}_h)_{j+1}) \quad \text{if } j \text{ is even,}$$

and

$$\begin{aligned} (\mathbf{PR}\mathbf{r}_h)_j &= \frac{1}{2}\left((\mathbf{R}\mathbf{r}_h)_{(j-1)/2} + (\mathbf{R}\mathbf{r}_h)_{(j+1)/2}\right), \\ &= \frac{1}{8}((\mathbf{r}_h)_{j-2} + 2(\mathbf{r}_h)_{j-1} + 2(\mathbf{r}_h)_j + 2(\mathbf{r}_h)_{j+1} + (\mathbf{r}_h)_{j+2}) \quad \text{if } j \text{ is odd.} \end{aligned}$$

So we obtain the desired result. ■

Using the above proposition and Taylor's expansion, we obtain the following lemma.

Lemma 31. *Suppose* $\mathbf{P}$ *and* $\mathbf{R}$ *are defined in (25) and (26), respectively. For any vector* $\mathbf{r}_h \in \mathbb{R}^{N-1}$*,*

$$\|(\mathbf{I} - \mathbf{PR})\mathbf{r}_h\|_\infty \le \min_{w \in \mathcal{F}_{\mathbf{r}_h}^{N-1}} \max_{y \in [0,1]} |w''(y)| \frac{3}{4N^2}.$$

Note that the definition of $\mathcal{F}_{\mathbf{r}_h}^{N-1}$ *follows from Definition 29.*

Proof. Using Proposition 30 and Taylor's Theorem, in the case that j is even, we obtain

$$\begin{aligned} \frac{1}{4}((\mathbf{r}_h)_{j-1} + 2(\mathbf{r}_h)_j + (\mathbf{r}_h)_{j+1}) &= \frac{1}{4}\left(w\left(y_{j-1}\right) + 2w\left(y_j\right) + w\left(y_{j+1}\right)\right), \\ &= w\left(y_j\right) + \frac{w''(y_{c1})}{8}\frac{1}{N^2} + \frac{w''(y_{c2})}{8}\frac{1}{N^2}, \\ &= (\mathbf{r}_h)_j + \frac{w''(y_{c1}) + w''(y_{c2})}{8}\frac{1}{N^2}, \end{aligned}$$

where $w(\cdot) \in \mathcal{F}_{\mathbf{r}_h}^{N-1}$, $y_{j-1} \le y_{c1} \le y_j$, and $y_j \le y_{c2} \le y_{j+1}$. Similarly, in the case that j is odd, we have

$$\frac{1}{8}((\mathbf{r}_h)_{j-2} + 2(\mathbf{r}_h)_{j-1} + 2(\mathbf{r}_h)_j + 2(\mathbf{r}_h)_{j+1} + (\mathbf{r}_h)_{j+2}) \\ = (\mathbf{r}_h)_j + \frac{4w''(y_{c3}) + 2w''(y_{c4}) + 2w''(y_{c5}) + 4w''(y_{c6})}{16}\frac{1}{N^2}, \quad (27)$$

where $y_{j-2} \le y_{c3} \le y_j$, $y_{j-1} \le y_{c4} \le y_j$, $y_j \le y_{c5} \le y_{j+1}$, and $y_j \le y_{c6} \le y_{j+2}$. Therefore,

$$\|(\mathbf{I} - \mathbf{PR})\mathbf{r}_h\|_\infty \le \max_{y\in[0,1]} |w''(y)|\frac{3}{4N^2} \quad \text{for} \quad \forall w(\cdot) \in \mathcal{F}_{\mathbf{r}_h}^{N-1}.$$

So we obtain the desired result. ■

Lemma 31 provides upper bound of $\|(\mathbf{I} - \mathbf{PR})\mathbf{r}_h\|_\infty$, for any $\mathbf{r}_h \in \mathbf{R}^{N-1}$. This result can be used to derive the upper bound of $\|(\mathbf{I} - \mathbf{PR})(\mathbf{x}_{h,k} - \mathbf{x}_{h,\star})\|$, where $\mathbf{r}_h = \mathbf{x}_{h,k} - \mathbf{x}_{h,\star}$. As we can see, if $|w''(y)| = \mathcal{O}(1)$, where $w \in \mathcal{F}_{\mathbf{r}_h}^{N-1}$, then $\|(\mathbf{I} - \mathbf{PR})\mathbf{r}_h\|_\infty = \mathcal{O}(N^{-2})$. This can be explained by the fact that when the mesh size is fine enough (i.e. large N), linear interpolation and restriction provide very good estimations of the fine model.

In the following lemma, we provide an upper bound of $|w''|$ in terms of the original vector $\mathbf{r}_h$. The idea is to specify the interpolation method in which we construct w, and we will use cubic spline in particular. Cubic spline is one of the standard interpolation methods, and the output interpolated function w satisfies the setting in Definition 29 and Lemma 31.

Lemma 32. *Suppose* $\mathbf{P}$ *and* $\mathbf{R}$ *are defined in (25) and (26), respectively. For any vector* $\mathbf{r}_h \in \mathbb{R}^{N-1}$*, we obtain*

$$\|(\mathbf{I} - \mathbf{PR})\mathbf{r}_h\|_\infty \le \frac{9}{4N^2}\|\mathbf{A}\mathbf{r}_h\|_\infty,$$

where

$$\mathbf{A} = N^2 \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & \ddots & \ddots & \\ & & \ddots & 2 & -1 \\ & & & -1 & 2 \end{pmatrix}.$$

Proof. We follow the notation in Definition 29. For $w \in \mathcal{F}_{\mathbf{r}_h}^{N-1}$ that is constructed via cubic spline, in the interval (y_i, y_{i+1}), we have

$$w(y) = Aw_i + Bw_{i+1} + Cw_i'' + Dw_{i+1}'',$$

where

$$\begin{aligned} A &= \frac{y_{i+1} - y}{y_{i+1} - y_i}, \\ B &= \frac{y - y_i}{y_{i+1} - y_i}, \\ C &= \frac{1}{6}(A^3 - A)(y_{i+1} - y_i)^2, \\ D &= \frac{1}{6}(B^3 - B)(y_{i+1} - y_i)^2. \end{aligned}$$

It is known from [31] that

$$\frac{\mathrm{d}^2 w}{\mathrm{d}y^2} = Aw_i'' + Bw_{i+1}'', \tag{28}$$

and

$$\frac{y_i - y_{i-1}}{6} w_{i-1}'' + \frac{y_{i+1} - y_{i-1}}{3} w_i'' + \frac{y_{i+1} - y_i}{6} w_{i+1}'' = \frac{w_{i+1} - w_i}{y_{i+1} - y_i} - \frac{w_i - w_{i-1}}{y_i - y_{i-1}}, \tag{29}$$

and for $i = 1, 2, \ldots, N-1$. Using the above equation (28), at the interval (y_i, y_{i+1}), we obtain

$$\begin{aligned}
\left|\frac{\mathrm{d}^2 w}{\mathrm{d}y^2}\right| = \left|Aw_i'' + Bw_{i+1}''\right| &= \left|\frac{y_{i+1} - y}{y_{i+1} - y_i} w_i'' + \frac{y - y_i}{y_{i+1} - y_i} w_{i+1}''\right|, \\
&\leq \left|\frac{y_{i+1} - y}{y_{i+1} - y_i}\right| |w_i''| + \left|\frac{y - y_i}{y_{i+1} - y_i}\right| |w_{i+1}''|, \\
&\leq \max\{|w_i''|, |w_{i+1}''|\}.
\end{aligned}$$

Suppose $j \in \arg\max_i \{|w_i''|\}_i$, then from (29) and the fact that $y_{j+1} - y_j = 1/N$,

$$\begin{aligned}
\frac{y_{j+1} - y_{j-1}}{3} w_j'' &= \frac{w_{j+1} - w_j}{y_{j+1} - y_j} - \frac{w_j - w_{j-1}}{y_j - y_{j-1}} - \frac{y_j - y_{j-1}}{6} w_{j-1}'' - \frac{y_{j+1} - y_j}{6} w_{j+1}'', \\
\frac{2}{3N} w_j'' &= N(w_{j+1} - w_j) - N(w_j - w_{j-1}) - \frac{1}{6N} w_{j-1}'' - \frac{1}{6N} w_{j+1}'', \\
2w_j'' &= 3N^2(w_{j+1} - 2w_j + w_{j-1}) - \frac{1}{2} w_{j-1}'' - \frac{1}{2} w_{j+1}''.
\end{aligned}$$

Thus,

$$\begin{aligned}
|2w_j''| &\leq 3N^2 |w_{j+1} - 2w_j + w_{j-1}| + \frac{1}{2}|w_{j-1}''| + \frac{1}{2}|w_{j+1}''|, \\
2|w_j''| &\leq 3N^2 |w_{j+1} - 2w_j + w_{j-1}| + \frac{1}{2}|w_j''| + \frac{1}{2}|w_j''|, \\
|w_j''| &\leq 3N^2 |w_{j+1} - 2w_j + w_{j-1}|.
\end{aligned}$$

Therefore,

$$|w_i''| \leq \max_i 3N^2 |w_{i+1} - 2w_i + w_{i-1}|,$$

and so,

$$\|(\mathbf{I} - \mathbf{PR})\mathbf{r}_h\|_\infty \leq \max_{y \in [0,1]} |w''(y)| \frac{3}{4N^2} \leq \max_i \frac{9|w_{i+1} - 2w_i + w_{i-1}|}{4} = \frac{9}{4N^2} \|\mathbf{A}\mathbf{r}_h\|_\infty,$$

as required. ■

Lemma 32 provides the discrete version of the result presented in Lemma 31. The matrix $\mathbf{A}$ is the discretized Laplacian operator, which is equivalent to twice differentiation using finite difference with a uniform mesh.

5.3 Convergence

With all the results, we revisit the composite convergence rate with the following Corollary.

Corollary 33. *Suppose* $\mathbf{P}$ *and* $\mathbf{R}$ *are defined in (25) and (26), respectively. If the coarse correction step* $\hat{\mathbf{d}}_{h,k}$ *in (15) is taken with* $\alpha_{h,k} = 1$*, then*

$$\begin{aligned}\|\mathbf{x}_{h,k+1} - \mathbf{x}_{h,\star}\| &\leq \sqrt{\frac{L_h}{\mu_h}} \min_{w \in \mathcal{F}^{N-1}_{\mathbf{x}_{h,k}-\mathbf{x}_{h,\star}}} \max_{y\in[0,1]} |w''(y)| \frac{3}{4N^{3/2}} + \frac{M_h\omega^2\xi^2}{2\mu_h}\|\mathbf{x}_{h,k} - \mathbf{x}_{h,\star}\|^2, \\ &\leq \frac{9}{4N^{3/2}}\sqrt{\frac{L_h}{\mu_h}}\|\mathbf{A}(\mathbf{x}_{h,k} - \mathbf{x}_{h,\star})\| + \frac{M_h\omega^2\xi^2}{2\mu_h}\|\mathbf{x}_{h,k} - \mathbf{x}_{h,\star}\|^2,\end{aligned}$$

where $\mathbf{A}$ *is defined in Lemma 32. Note that* M_h*,* L_h*, and* μ_h *are defined in Assumption 1,* $\omega = \max\{\|\mathbf{P}\|, \|\mathbf{R}\|\}$*, and* $\xi = \|\mathbf{P}^+\|$*.*

Proof.

$$\begin{aligned}\|\mathbf{x}_{h,k+1} - \mathbf{x}_{h,\star}\| &\leq \|\mathbf{I} - \mathbf{P}[\nabla^2_H f_{h,k}]^{-1}\mathbf{R}\nabla^2 f_{h,k}\|\|(\mathbf{I} - \mathbf{PR})(\mathbf{x}_{h,k} - \mathbf{x}_{h,\star})\| \\ &\quad + \frac{M_h\omega^2\xi^2}{2\mu_h}\|\mathbf{x}_{h,k} - \mathbf{x}_{h,\star}\|^2, \\ &\leq \sqrt{\frac{L_h}{\mu_h}} \min_{w \in \mathcal{F}^{N-1}_{\mathbf{x}_{h,k}-\mathbf{x}_{h,\star}}} \max_{y\in[0,1]} |w''(y)| \frac{3}{4N^{3/2}} + \frac{M_h\omega^2\xi^2}{2\mu_h}\|\mathbf{x}_{h,k} - \mathbf{x}_{h,\star}\|^2, \\ &\leq \frac{9}{4N^{3/2}}\sqrt{\frac{L_h}{\mu_h}}\|\mathbf{A}(\mathbf{x}_{h,k} - \mathbf{x}_{h,\star})\| + \frac{M_h\omega^2\xi^2}{2\mu_h}\|\mathbf{x}_{h,k} - \mathbf{x}_{h,\star}\|^2,\end{aligned}$$

as required. ■

Corollary 33 provides the convergence of using Galerkin model for PDE-based problems that we considered. This result shows the complementary of fine correction step and coarse correction step. Suppose the fine correction step can effectively reduce $\|\mathbf{A}(\mathbf{x}_{h,k} - \mathbf{x}_{h,\star})\|$, then the coarse correction step could yield major reduction based on the result shown in Corollary 33.

6 Low Rank Approximation using Nyström Method

In this section, we focus on the Galerkin model that is based on low rank approximation of the Hessian matrix. We begin with an introduction of low rank approximation and the Nyström method. Then we make the connection between the Nyström method and the coarse correction step in (15). Finally, we re-derive the composite rate with a more insightful bounds of both $\|\mathbf{I} - \mathbf{P}[\nabla^2_H f_{h,k}]^{-1}\mathbf{R}\nabla^2 f_{h,k}\|$ and $\|(\mathbf{I} - \mathbf{PR})(\mathbf{x}_{h,k} - \mathbf{x}_{h,\star})\|$.

Before introducing the obscure connection between low rank approximation and Galerkin model, let's start with the setting and consider a symmetric positive semi-definite matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$. The best low rank approximation of $\mathbf{A}$ with rank q can be obtained by solving the following optimization problem

$$\min_{\mathbf{A}_q \in \mathbb{R}^{N\times N}} \|\mathbf{A} - \mathbf{A}_q\|, \quad \text{s.t. } \operatorname{rank}(\mathbf{A}_q) = q. \tag{30}$$

It is known that the above problem can be solved via eigendecomposition. However, eigendecomposition is computationally expensive. In the context of optimization, the cost for each

iteration of Newton's method is not more expensive than performing eigendecomposition. If a Galerkin model is constructed via eigendecomposition, one could apply Newton's method instead.

Although computing the exact solution of (30) is unfavorable, we could seek for its approximation. Nyström method was originally developed to numerically approximate eigenfunctions, and the idea was applied later in the machine learning community for the low rank optimization problem [41]. It provides a suboptimal solution of the low rank approximation with cheaper computational cost.

Nyström method is performed by the column selection procedure. Consider a set $\mathcal{Q} = \{1, 2, \ldots, N\}$, and suppose a subset $\mathcal{Q}_1 \subseteq \mathcal{Q}$ with n elements. We denote q_i as the i^{th} element of $\mathcal{Q}_1$, for $i = 1, 2, \ldots, n$. Then one can approximate $\mathbf{A} \in \mathbb{R}^{N\times N}$ using the following procedures.

1. Define a matrix $\mathbf{A}_1 \in \mathbb{R}^{n\times N}$ such that the i^{th} row of $\mathbf{A}_1$ is the q_i^{th} row of $\mathbf{A}$.
2. Define a matrix $\mathbf{A}_2 \in \mathbb{R}^{N\times n}$ such that the i^{th} column of $\mathbf{A}_1$ is the q_i^{th} column of $\mathbf{A}$.
3. Define a matrix $\mathbf{A}_3 \in \mathbb{R}^{n\times n}$ such that $(\mathbf{A}_3)_{i,j}$ is the element of $\mathbf{A}$ in q_i^{th} row and q_j^{th} column.
4. Compute the pseudo-inverse $\mathbf{A}_3^+$.
5. Compute the low rank approximation of $\mathbf{A}$ by $\mathbf{A}_2\mathbf{A}_3^+\mathbf{A}_1$.

Equivalently, the above procedure can be described by using a matrix $\mathbf{S} \in \mathbb{R}^{N\times n}$ such that the i^{th} column of $\mathbf{S}$ is the q_i^{th} column of the identity matrix $\mathbf{I}$. The output of the above procedure is the same as

$$\mathbf{A} \approx \mathbf{A}_2\mathbf{A}_3^+\mathbf{A}_1 = \mathbf{AS}[\mathbf{S}^T\mathbf{AS}]^+\mathbf{S}^T\mathbf{A}. \tag{31}$$

Much research have been focused on developing Nyström based on different methods on selecting the subset $\mathcal{Q}_1$ [6, 11, 33, 41]. In this paper, we consider the naïve Nyström method in which elements in $\mathcal{Q}_1$ are selected uniformly without replacement from $\mathcal{Q}$.

6.1 Galerkin Model by Naïve Nyström Method

Now we are in the position to show how Nyström method can be used to develop Galerkin model. The approximation (31) is highly similar to the coarse correction step in multilevel algorithm.

Definition 34. *Consider a set $\mathcal{Q} = \{1, 2, \ldots, N\}$, and an n elements subset $\mathcal{Q}_1$ in which elements are selected randomly, and uniformly without replacement from $\mathcal{Q}$. Denote q_i as the i^{th} element of $\mathcal{Q}_1$. Then the prolongation operator, $\mathbf{P} \in \mathbb{R}^{N\times n}$, and restriction operator, $\mathbf{R} \in \mathbb{R}^{n\times N}$, are generated using naïve Nyström method if*

i. The i^{th} column of $\mathbf{P}$ is the q_i^{th} column of the identity matrix $\mathbf{I}$.

ii. $\mathbf{R} = \mathbf{P}^T$.

Definition 34 defines the prolongation and restriction operators that are based on naïve Nyström method. One can see the analogy by substituting $\mathbf{S} = \mathbf{P}$, $\mathbf{S}^T = \mathbf{R}$, and $\mathbf{A} = \nabla^2 f_{h,k}$ in equation (31). Under the setting of naïve Nyström method, $\mathbf{P}$ is full column rank, and so Assumption 3 is satisfied. Moreover, different from the assumption that $\mathbf{A}$ is positive semi-definite in (31), $\nabla^2 f_{h,k}$ is positive definite as stated in Assumption 1, and so it is guaranteed to

be invertible. Consider the low rank approximation (31) with $\mathbf{P}$, $\mathbf{R}$, and $\nabla^2 f_{h,k}$. Multiplying $\nabla^2 f_{h,k}^{-1}$ from both left and right yields,

$$\nabla^2 f_{h,k}^{-1} \approx \mathbf{P}[\mathbf{R}\nabla^2 f_{h,k}\mathbf{P}]^{+}\mathbf{R} = \mathbf{P}[\mathbf{R}\nabla^2 f_{h,k}\mathbf{P}]^{-1}\mathbf{R},$$

and so

$$-\nabla^2 f_{h,k}^{-1}\nabla f_{h,k} \approx -\mathbf{P}[\mathbf{R}\nabla^2 f_{h,k}\mathbf{P}]^{-1}\mathbf{R}\nabla f_{h,k} = \hat{\mathbf{d}}_{h,k}.$$

Thus, the coarse correction step $\hat{\mathbf{d}}_{h,k}$ is an approximation of Newton step. We emphasize that naïve Nyström method is effective in practice, and computationally inexpensive to perform (uniform sampling without replacement).

It is worth mentioning that the coarse correction step is highly related block-coordinate descent algorithms. In fact, $\mathbf{P}$ and $\mathbf{R}$ from Definition 34 can be used to derive block-coordinate descent algorithms, as described in Section 2.4. The coarse correction step in this section is different from first order block-coordinate descent type methods because GAMA uses the Hessian $\nabla^2 f_{h,k}$ instead of identity matrix in the coarse model (11).

Interestingly, similar works have been done from the perspective of block coordinate methods for machine learning problems. In particular, Gower et al [12] recently developed a stochastic block BFGS for solving the sum of twice differentiable convex functions. The coarse correction step we study in this section is a special case of the stochastic block BFGS: when the previous approximated inverse Hessian is set to be zero and when all functions (in the summation) are used to compute Hessians. On the other hand, the proposed coarse correction step is also studied by Qu et al [32] for the dual formulation of empirical risk minimization. In both cases, they provided (expected) linear convergence rates. Moreoever, due to different sources of motivation, they did not mention that Nyström is used inherently within the search direction.

6.2 Analysis

We are now in the position to analyze the two important factors in the composition convergence rate, $\|\mathbf{I} - \mathbf{P}[\nabla_H^2 f_{h,k}]^{-1}\mathbf{R}\nabla^2 f_{h,k}\|$ and $\|(\mathbf{I} - \mathbf{PR})(\mathbf{x}_{h,k} - \mathbf{x}_{h,\star})\|$. The analytical tool we used is concentration inequality. The following Chernoff bounds will be used to analyze $\|(\mathbf{I} - \mathbf{PR})(\mathbf{x}_{h,k} - \mathbf{x}_{h,\star})\|$.

Theorem 35 ([36])**.** *Let $\mathcal{Q}$ be a finite set of positive numbers, and suppose*

$$\max_{q\in\mathcal{Q}} q \leq B.$$

Sample $\{q_1, q_2, \ldots, q_l\}$ uniformly at random from $\mathcal{Q}$ without replacement. Compute

$$s = l \cdot \mathbb{E}(q_1).$$

Then

$$\mathbb{P}\left\{\sum_j q_j \leq (1-\sigma)s\right\} \leq \left(\frac{e^{-\sigma}}{(1-\sigma)^{1-\sigma}}\right)^{s/B} \qquad \textit{for} \quad \sigma \in [0,1), \quad \textit{and}$$

$$\mathbb{P}\left\{\sum_j q_j \geq (1+\sigma)s\right\} \leq \left(\frac{e^{\sigma}}{(1+\sigma)^{1+\sigma}}\right)^{s/B} \qquad \textit{for} \quad \sigma \geq 0.$$

Proof. See Theorem 2.1 from Tropp [36]. ■

Theorem 35 is useful to derive statistical bounds for $\|(\mathbf{I}-\mathbf{PR})\mathbf{r}_h\|$, for any $\mathbf{r}_h \in \mathbb{R}^N$. The results are provided in the following lemma.

Lemma 36. *Suppose prolongation operator $\mathbf{P} \in \mathbb{R}^{N\times n}$ and restriction operator $\mathbf{R} \in \mathbb{R}^{n\times N}$ are generated using naïve Nyström method according to Definition 34 and $\mathbf{r}_h \in \mathbb{R}^N$. Then $\forall \sigma \in [0,1)$, we obtain*

$$\mathbb{P}\left\{\|(\mathbf{I}-\mathbf{PR})\mathbf{r}_h\| \leq \sqrt{(1-\sigma)\frac{N-n}{N}}\|\mathbf{r}_h\|\right\} \leq \left(\frac{e^{-\sigma}}{(1-\sigma)^{1-\sigma}}\right)^{\frac{N-n}{N}\|\mathbf{r}_h\|^2/\|\mathbf{r}_h\|_\infty^2},$$

and $\forall \sigma \geq 0$, we obtain

$$\mathbb{P}\left\{\|(\mathbf{I}-\mathbf{PR})\mathbf{r}_h\| \geq \sqrt{(1+\sigma)\frac{N-n}{N}}\|\mathbf{r}_h\|\right\} \leq \left(\frac{e^{\sigma}}{(1+\sigma)^{1+\sigma}}\right)^{\frac{N-n}{N}\|\mathbf{r}_h\|^2/\|\mathbf{r}_h\|_\infty^2}.$$

Proof. We denote $\mathcal{Q} = \{1, 2, \ldots, N\}$ to be a set of indices, a subset $\mathcal{Q}_1 \subset \mathcal{Q}$ such that

$$\text{range}(\mathbf{P}) = \text{span}\left(\{\mathbf{e}_j : j \in \mathcal{Q}_1\}\right),$$

and the complement $\mathcal{Q}_2 \subset \mathcal{Q}$ such that

$$\mathcal{Q}_2 \cup \mathcal{Q}_1 = \mathcal{Q}, \quad \text{and} \quad \mathcal{Q}_2 \cap \mathcal{Q}_1 = \emptyset.$$

These definitions lead to

$$\|(\mathbf{I}-\mathbf{PR})\mathbf{r}_h\|^2 = \sum_{j\in\mathcal{Q}_2}(\mathbf{r}_h)_j^2,$$

since $\mathcal{Q}_2$ is a set of indices that are associated with the selected coordinates in $\mathbf{I}-\mathbf{PR}$. Therefore, $\mathcal{Q}_2$ contains $N-n$ samples from $\mathcal{Q}$ that are distributed uniformly without replacement. By applying Theorem 35, we obtain

$$\max_{j\in\mathcal{Q}}(\mathbf{r}_h)_j^2 = \|\mathbf{r}_h\|_\infty^2,$$

and

$$s = (N-n)\frac{1}{N}\sum_{j\in\mathcal{Q}}(\mathbf{r}_h)_j^2 = \frac{N-n}{N}\|\mathbf{r}_h\|^2.$$

By direct substitutions, we obtain the desired result. ■

Lemma 36 provides bounds for $\|(\mathbf{I}-\mathbf{PR})\mathbf{r}_h\|$, for any $\mathbf{r}_h \in \mathbb{R}^N$. On the other hand, we bear in mind that Nyström method is a method of computing low rank approximations. In the following lemma, we will show that this feature is shown in the bound of $\|\mathbf{I}-\mathbf{P}[\nabla_H^2 f_{h,k}]^{-1}\mathbf{R}\nabla^2 f_{h,k}\|$.

Lemma 37. *Suppose prolongation operator $\mathbf{P} \in \mathbb{R}^{N\times n}$ and restriction operator $\mathbf{R} \in \mathbb{R}^{n\times N}$ are generated using naïve Nyström method according to Definition 34. For $p \in \{1, 2, \ldots, N\}$, let the eigendecomposition of $\nabla^2 f_{h,k}$ has the following form*

$$\nabla^2 f_{h,k} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^T = \begin{pmatrix}\mathbf{U}_1 & \mathbf{U}_2\end{pmatrix}\begin{pmatrix}\mathbf{\Sigma}_1 & \\ & \mathbf{\Sigma}_2\end{pmatrix}\begin{pmatrix}\mathbf{U}_1^T \\ \mathbf{U}_2^T\end{pmatrix},$$

where $\mathbf{\Sigma}_1 \in \mathbb{R}^{p\times p}$, $\mathbf{\Sigma}_2 \in \mathbb{R}^{(N-p)\times(N-p)}$, $\mathbf{U}_1 \in \mathbb{R}^{N\times p}$, and $\mathbf{U}_2 \in \mathbb{R}^{N\times(N-p)}$ are the sub-matrices of $\mathbf{\Sigma}$ and $\mathbf{U}$. Denote τ as the coherence of $\mathbf{U}_1$,

$$\tau \triangleq \frac{N}{p}\max_i(\mathbf{U}_1\mathbf{U}_1^T)_{ii}.$$

Then, for β, σ and n such that

$$\beta, \sigma \in (0,1), \quad \textit{and} \quad n \geq \frac{2\tau p \log\left(\frac{p}{\beta}\right)}{(1-\sigma)^2},$$

we obtain

$$\|\mathbf{I} - \mathbf{P}[\nabla^2_H f_{h,k}]^{-1}\mathbf{R}\nabla^2 f_{h,k}\| \leq \sqrt{\frac{\lambda_{p+1}(\nabla^2 f_{h,k})}{\mu_h}\left(1 + \frac{N}{n\sigma}\right)},$$

with probability at least $1-\beta$. Note that $\lambda_{p+1}(\nabla^2 f_{h,k})$ is the $p+1^{th}$ largest eigenvalue of $\nabla^2 f_{h,k}$.

Proof. Following from Lemma 19, we have

$$\begin{aligned}
\|\mathbf{I} - \mathbf{P}[\nabla^2_H f_{h,k}]^{-1}\mathbf{R}\nabla^2 f_{h,k}\| &= \|\mathbf{U}\mathbf{\Sigma}^{-1/2}(\mathbf{I} - \mathbf{\Gamma}_{\mathbf{\Sigma}^{1/2}\mathbf{U}^T\mathbf{P}})\mathbf{\Sigma}^{1/2}\mathbf{U}^T\|, \\
&\leq \|\mathbf{U}\mathbf{\Sigma}^{-1/2}\|\|(\mathbf{I} - \mathbf{\Gamma}_{\mathbf{\Sigma}^{1/2}\mathbf{U}^T\mathbf{P}})\mathbf{\Sigma}^{1/2}\mathbf{U}^T\|, \\
&\leq \sqrt{\frac{1}{\mu_h}}\|(\mathbf{I} - \mathbf{\Gamma}_{\mathbf{\Sigma}^{1/2}\mathbf{U}^T\mathbf{P}})\mathbf{\Sigma}^{1/2}\mathbf{U}^T\|.
\end{aligned}$$

Using results from Gittens [10], Theorem 2,

$$\|(\mathbf{I} - \mathbf{\Gamma}_{\mathbf{\Sigma}^{1/2}\mathbf{U}^T\mathbf{P}})\mathbf{\Sigma}^{1/2}\mathbf{U}^T\| \leq \sqrt{\lambda_{p+1}(\nabla^2 f_{h,k})\left(1 + \frac{N}{n\sigma}\right)},$$

with probability at least $1-\beta$. ■

In addition to Lemma 19, Lemma 37 provides a new alternative on the bounding of the term $\|\mathbf{I} - \mathbf{P}[\nabla^2_H f_{h,k}]^{-1}\mathbf{R}\nabla^2 f_{h,k}\|$. This is a direct result from the fact that Nyström is used inherently with the $\mathbf{P}$ and $\mathbf{R}$ in Definition 34. As we will show later, this result would improve the convergence rate if the Hessian can be well-approximated using low rank approximation.

As mentioned in [10, 4], we would like to point out that the coherence τ defined in Lemma 19 ranges from 1 to N/p. For an $N \times p$ random orthogonal matrix in which its columns are selected uniformly among all families of p orthonormal vectors, its coherence is bounded by $\mathcal{O}(\max\{p, \log N\}/p)$ with high probability [4].

6.3 Convergence

Using the above results, we obtain the following corollaries.

Corollary 38. *Suppose $\mathbf{P} \in \mathbb{R}^{N\times n}$ and $\mathbf{R} \in \mathbb{R}^{n\times N}$ are generated using naïve Nyström method according to Definition 34, and τ is the coherence as defined in Lemma 37. If the coarse correction step $\hat{\mathbf{d}}_{h,k}$ is taken with $\alpha_{h,k} = 1$, then $\forall \sigma_2 \geq 0$,*

$$\|\mathbf{x}_{h,k+1} - \mathbf{x}_{h,\star}\| \leq \sqrt{\frac{L_h}{\mu_h}(1+\sigma_2)\frac{N-n}{N}}\|\mathbf{x}_{h,k} - \mathbf{x}_{h,\star}\| + \frac{M_h\omega^2\xi^2}{2\mu_h}\|\mathbf{x}_{h,k} - \mathbf{x}_{h,\star}\|^2,$$

with probability at least

$$1 - (\Phi(\sigma_2))^{\frac{N-n}{N}\|\mathbf{x}_{h,k}-\mathbf{x}_{h,\star}\|^2/\|\mathbf{x}_{h,k}-\mathbf{x}_{h,\star}\|^2_\infty} \quad \textit{for} \quad \Phi(\sigma_2) = \frac{e^{\sigma_2}}{(1+\sigma_2)^{1+\sigma_2}}. \tag{32}$$

Note that L_h, M_h, and μ_h are defined in Assumption 1, $\omega = \max\{\|\mathbf{P}\|, \|\mathbf{R}\|\}$ and $\xi = \|\mathbf{P}^+\|$.

Proof. The result could be obtained by combining results from Lemma 36 with $\mathbf{r}_h = \mathbf{x}_{h,k} - \mathbf{x}_{h,\star}$, Lemma 19, and Theorem 18. ■

Corollary 38 provides the probabilistic composite convergence rate. As expected, the coefficient of the linear component goes to 0 as $n \to N$. We point out that when $n = N$, the probability in (32) is equal to zero since $(N - n)/N = 0$. Thus, Corollary 38 is not meaningful at the exact limit of $n = N$. However, in this case, no dimension is reduced, and so based on Theorem 18, the quadratic convergence is obtained.

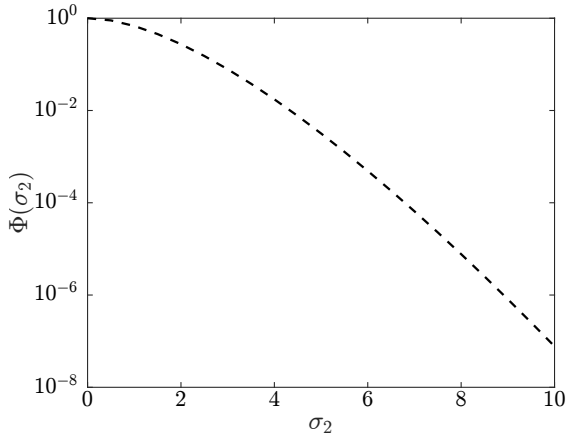

Figure 3: $\Phi(\sigma_2)$ in (32)

Figure 3 shows the value of $\Phi(\sigma_2)$ in (32), and one can see that with reasonably small σ_2, $\Phi(\sigma_2) << 1$. Also, since $\|\mathbf{x}_{h,k} - \mathbf{x}_{h,\star}\|^2$ is the sum of squares of the error in each dimension, it is reasonable to expect that it is in $\mathcal{O}(N)$. Therefore, one could expect that

$$\frac{N-n}{N}\frac{\|\mathbf{x}_{h,k} - \mathbf{x}_{h,\star}\|^2}{\|\mathbf{x}_{h,k} - \mathbf{x}_{h,\star}\|_\infty^2} \sim \mathcal{O}(N-n),$$

and so for $n < N$, the power coefficient above should reduce $\Phi(\sigma_2)$ further.

While Corollary 38 illustrates how $\|(\mathbf{I} - \mathbf{PR})(\mathbf{x}_{h,k} - \mathbf{x}_{h,\star})\|$ varies with respect to n, it does not show that using the prolongation and restriction operators that are inspired by Nyström method has any advantage when Hessians have the low rank structure. By combining result in Lemma 37, we obtain the following corollary.

Corollary 39. *Suppose $\mathbf{P} \in \mathbb{R}^{N\times n}$ and $\mathbf{R} \in \mathbb{R}^{n\times N}$ are generated using naïve Nyström method according to Definition 34, and τ is the coherence as defined in Lemma 37. If the coarse correction step $\hat{\mathbf{d}}_{h,k}$ is taken with $\alpha_{h,k} = 1$, then $\forall \beta, \sigma_1\sigma_2, p, n$ such that*

$$\beta, \sigma_1 \in (0,1), \quad \sigma_2 \geq 0, \quad p \in \{1, 2, \ldots, N\}, \quad \textit{and} \quad n \geq \frac{2\tau p \log\left(\frac{p}{\beta}\right)}{(1-\sigma_1)^2},$$

we obtain

$$\|\mathbf{x}_{h,k+1} - \mathbf{x}_{h,\star}\| \leq \sqrt{\frac{\lambda_{p+1}(\nabla^2 f_{h,k})}{\mu_h}\left(1 + \frac{N}{n\sigma_1}\right)(1+\sigma_2)\frac{N-n}{N}}\|\mathbf{x}_{h,k} - \mathbf{x}_{h,\star}\| + \frac{M_h\omega^2\xi^2}{2\mu_h}\|\mathbf{x}_{h,k} - \mathbf{x}_{h,\star}\|^2,$$

with probability at least

$$(1-\beta)\left(1-(\Phi(\sigma_2))^{\frac{N-n}{N}\|\mathbf{x}_{h,k}-\mathbf{x}_{h,\star}\|^2/\|\mathbf{x}_{h,k}-\mathbf{x}_{h,\star}\|_\infty^2}\right) \quad \textit{for} \quad \Phi(\sigma_2)=\frac{e^{\sigma_2}}{(1+\sigma_2)^{1+\sigma_2}}.$$

Note that L_h, M_h, and μ_h are defined in Assumption 1; $\omega=\max\{\|\mathbf{P}\|,\|\mathbf{R}\|\}$ and $\xi=\|\mathbf{P}^+\|$; $\lambda_{p+1}(\nabla^2 f_{h,k})$ is the $p+1^{th}$ largest eigenvalue of $\nabla^2 f_{h,k}$.

Proof. The result could be obtained by combining results from Lemma 36 with $\mathbf{r}_h=\mathbf{x}_{h,k}-\mathbf{x}_{h,\star}$, Lemma 37, and Theorem 18. ■

Compared to Corollary 38, Corollary 39 replaces the largest eigenvalue of $\nabla^2 f_{h,k}$, L_h, with the scaled $p+1^{\text{th}}$ largest eigenvalue, $\lambda_{p+1}(\nabla^2 f_{h,k})$, with high probability. It provides a clear advantage when there is a large gap between the p^{th} and $p+1^{\text{th}}$ eigenvalue. In particular, when

$$\mu_h \le \lambda_N(\nabla^2 f_{h,k}) \le \cdots \le \lambda_{p+1}(\nabla^2 f_{h,k}) << \lambda_p(\nabla^2 f_{h,k}) \le \lambda_1(\nabla^2 f_{h,k}) \le L_h.$$

We point out that concentration inequality is not only useful for getting composite convergence rate, but also useful for bounding the parameter κ in Algorithm 1.

Lemma 40. *Suppose prolongation operator $\mathbf{P}\in\mathbb{R}^{N\times n}$ and restriction operator $\mathbf{R}\in\mathbb{R}^{n\times N}$ are generated using naïve Nyström method according to Definition 34. Then $\forall \mathbf{r}_h\in\mathbb{R}^N, \forall\sigma\in[0,1)$, we obtain*

$$\mathbb{P}\left\{\|\mathbf{PR}\mathbf{r}_{h,k}\| \le \sqrt{(1-\sigma)\frac{n}{N}}\|\mathbf{r}_{h,k}\|\right\} \le \left(\frac{e^{-\sigma}}{(1-\sigma)^{1-\sigma}}\right)^{\frac{n}{N}\|\mathbf{r}_{h,k}\|^2/\|\mathbf{r}_{h,k}\|_\infty^2},$$

and $\forall\sigma\ge 0$, we have

$$\mathbb{P}\left\{\|\mathbf{PR}\mathbf{r}_{h,k}\| \ge \sqrt{(1+\sigma)\frac{n}{N}}\|\mathbf{r}_{h,k}\|\right\} \le \left(\frac{e^{\sigma}}{(1+\sigma)^{1+\sigma}}\right)^{\frac{n}{N}\|\mathbf{r}_{h,k}\|^2/\|\mathbf{r}_{h,k}\|_\infty^2}.$$

Proof. The proof is exactly the same as in Lemma 36 with consideration of $\mathcal{Q}_1$ as a sample set instead. ■

Lemma 40 provides the fact that with high probability

$$\|\mathbf{R}\nabla f_{h,k}\| = \|\mathbf{PR}\nabla f_{h,k}\| \ge \mathcal{O}\left(\sqrt{\frac{n}{N}}\right)\|\nabla f_{h,k}\|.$$

Note that in the analysis in Section 3, when the coarse correction step is taken, we assume $\|\mathbf{R}\nabla f_{h,k}\| > \kappa\|\nabla f_{h,k}\|$ for some constant κ. As stated in Lemma 10 and Theorem 13, the square of this kappa is proportional to Λ, which is inversely proportional to the rate of convergence. Therefore, we shall conclude that in the setting considered in this section, with high probability the rate of convergence is inversely proportional to $\mathcal{O}(n/N)$, or equivalently, proportional to $\mathcal{O}(N/n)$.

7 Block Diagonal Approximation

In this section, we focus on the case that the Hessian $\nabla^2 f_{h,k}$ is approximated by block diagonal approximation. The structure of this section is similar to the last two sections: we introduce and formally define block diagonal approximation, perform analysis, and finally re-derive the composite rate in this setting.

Definition 41. *Suppose* $\nabla^2 f_{h,k} \in \mathbb{R}^{N\times N}$ *and* $n_1, n_2, \ldots, n_q \in \mathbb{N}$ *such that* $n_1+n_2+\cdots+n_q = N$. *Then the* q*-block diagonal approximation of* $\nabla^2 f_{h,k}$ *is defined as* $\nabla^2_B f_{h,k}$ *where*

$$(\nabla^2_B f_{h,k})_{i,j} = \begin{cases} (\nabla^2 f_{h,k})_{i,j} & \text{if} \quad \sum_{p=1}^{m-1} n_p < i,j \leq \sum_{p=1}^{m} n_p, \quad \text{for any } \ m = 1,2,\ldots,q, \\ 0 & \text{otherwise}. \end{cases}$$

Definition 41 states the formal definition of block diagonal approximation of a Hessian. That is, we only preserve the elements which are located in block diagonal positions, and set all the other elements to zeros. Recall that even though Newton's method is one of the best algorithms with quadratic convergence rate, the trade-off, however, goes into the high computational cost at each iteration: solving an N-by-N system of linear equations. By replacing the Hessian with its q-block diagonal approximation, the corresponding N-by-N system of linear equations can be decomposed by q smaller systems of linear equations, and thus lower computational cost is required.

The above block diagonal approximation approach is a special case of the incomplete Hessian Newton minimization method proposed by Xie and Ni [42]. In the case where $n_1 = n_2 = \cdots = n_N = 1$, this diagonal approximation is also considered in [9]. While it is clear that the block diagonal approximation contains partial second order information and one should expect that it performs better than first order algorithms, no theoretical indication has pointed in this direction.

7.1 Multiple Galerkin Models

We will show that q-block diagonal approximation from Definition 41 could be formulated using multiple Galerkin models. We denote prolongation operators $\mathbf{P}_i \in \mathbb{R}^{N\times n_i}$, for $i = 1,2,\ldots,q$. Notice that $n_1 + n_2 + \cdots + n_q = N$, and we assume

$$[\mathbf{P}_1\ \mathbf{P}_2 \ldots \mathbf{P}_q] = \mathbf{I}.$$

We also denote the corresponding restriction operators $\mathbf{R}_i = \mathbf{P}_i^T$, for $i = 1,2,\ldots,q$. Then, block diagonal approximation can be expressed as

$$\nabla^2_{\mathrm{B}} f_{h,k} = \mathrm{diag}(\mathbf{R}_1\nabla^2 f_{h,k}\mathbf{P}_1, \mathbf{R}_2\nabla^2 f_{h,k}\mathbf{P}_2, \ldots, \mathbf{R}_q\nabla^2 f_{h,k}\mathbf{P}_q),$$

and the corresponding coarse correction step is defined as

$$\hat{\mathbf{d}}_{h,k} = -[\nabla^2_{\mathrm{B}} f_{h,k}]^{-1}\nabla f_{h,k} = \sum_{i=1}^{q} -\mathbf{P}_i[\mathbf{R}_i\nabla^2 f_{h,k}\mathbf{P}_i]\mathbf{R}_i\nabla f_{h,k}. \tag{33}$$

7.2 General functions? A counterexample

We start with a counterexample to show that it is impossible to be as good as the classical Newton's method for general functions in term of convergence. Suppose we have the following problem

$$\min_{\mathbf{x}_h\in\mathbb{R}^2} f_h(\mathbf{x}_h) \triangleq \frac{1}{2}\mathbf{x}_h^T \begin{pmatrix} 1 & -\sqrt{0.5} \\ -\sqrt{0.5} & 1 \end{pmatrix} \mathbf{x}_h + \mathbf{x}_h^T \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

The above quadratic program (QP) has positive definite Hessian

$$\begin{pmatrix} 1 & -\sqrt{0.5} \\ -\sqrt{0.5} & 1 \end{pmatrix} = \begin{pmatrix} 1 \\ -\sqrt{0.5} \end{pmatrix} \begin{pmatrix} 1 & -\sqrt{0.5} \end{pmatrix} + \begin{pmatrix} 0 \\ \sqrt{0.5} \end{pmatrix} \begin{pmatrix} 0 & \sqrt{0.5} \end{pmatrix}.$$

Therefore, the above function is a strongly convex function. In this example we assume 2-blocks approximation is performed, with $n_1 = n_2 = 1$. Notice that the classical Newton's method solves the above QP in 1 iteration. The coarse correction step, on the other hand, fail to do so; in fact, the diagonal of the Hessian has only 1's, which implies that in this particular example, the coarse correction step is equivalent to gradient descent.

7.3 Weakly connected Hessian

We now introduce specific class of problems in which the coarse correction step could be as good as Newton's method at the limit.

Definition 42. *Consider a twice-differentiable strongly convex function f_h which satisfies Assumption 1. f_h is said to have (δ, q)-weakly connected Hessians if*

$$\nabla^2 f_h(\mathbf{x}_h) = \mathbf{Q}_h(\mathbf{x}_h) + \delta \hat{\mathbf{Q}}_h(\mathbf{x}_h), \tag{34}$$

where $\mathbf{Q}_h(\mathbf{x}_h) = diag(\mathbf{Q}_{h,1}(\mathbf{x}_h), \mathbf{Q}_{h,2}(\mathbf{x}_h), \ldots, \mathbf{Q}_{h,q}(\mathbf{x}_h))$ is a block diagonal matrix with q blocks, with $\mathbf{Q}_{h,i}(\mathbf{x}_h) \in \mathbb{R}^{n_i \times n_i}$ and $\sum_{j=1}^{q} n_j = N$ for $i = 1, 2, \ldots, q$. All $\mathbf{Q}_{h,i}(\mathbf{x}_h)$'s are positive definite, and there exists constants $\mu_{h,q}, \mu_{h,\hat{q}}, L_{h,q}, L_{h,\hat{q}}$ such that

$$\begin{aligned} \mu_{h,q}\mathbf{I} \preccurlyeq \quad & \mathbf{Q}_h(\mathbf{x}_h) \quad \preccurlyeq L_{h,q}\mathbf{I} \\ \mu_{h,\hat{q}}\mathbf{I} \preccurlyeq \quad & \hat{\mathbf{Q}}_h(\mathbf{x}_h) \quad \preccurlyeq L_{h,\hat{q}}\mathbf{I} \end{aligned}$$

Definition 42 defines the specific structure we consider in this section. The defined (δ, q)-weakly connected Hessian provides a connection between the block diagonal matrix and general positive definite matrix. Suppose when $\delta = 0$, then the (δ, q)-weakly connected Hessian is exactly a block diagonal matrix. Similarly, when $\delta = \mathcal{O}(1)$, then the (δ, q)-weakly connected Hessian is a general positive definite matrix.

Notice that when $\delta = 0$, the coarse correction step (33) is exactly same as Newton's method. In what follows, we will consider f_h which has (δ, q)-weakly connected Hessians and show how the performance of coarse correction step (33) converges to quadratic convergence when $\delta \to 0$.

7.4 Analysis

In order to analyze the convergence of coarse correction step (33), we relate it with the classical Newton's step and derive the difference using the following propositions.

Proposition 43 ([38])**.** *For matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$, suppose $\mathbf{A}$, $\mathbf{C}$, and $\mathbf{A}+\mathbf{BCD}$ are nonsingular, then*

$$(\mathbf{A} + \mathbf{BCD})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{C}^{-1} + \mathbf{DA}^{-1}\mathbf{B})^{-1}\mathbf{DA}^{-1}$$

Proof. See [38]. ■

Proposition 44. *Suppose the Hessian $\nabla^2 f_{h,k}$ is (δ, q)-weakly connected as defined in Definition 42, then*

$$\left(\frac{1}{L_{h,\hat{q}}} + \frac{\delta}{L_{h,q}}\right)\mathbf{I} \preccurlyeq \hat{\mathbf{Q}}^{-1} + \delta\mathbf{Q}^{-1} \preccurlyeq \left(\frac{1}{\mu_{h,\hat{q}}} + \frac{\delta}{\mu_{h,q}}\right)\mathbf{I},$$

and so

$$\left(\frac{1}{\mu_{h,\hat{q}}} + \frac{\delta}{\mu_{h,q}}\right)^{-1}\mathbf{I} \preccurlyeq (\hat{\mathbf{Q}}^{-1} + \delta\mathbf{Q}^{-1})^{-1} \preccurlyeq \left(\frac{1}{L_{h,\hat{q}}} + \frac{\delta}{L_{h,q}}\right)^{-1}\mathbf{I},$$

where

$$\left(\frac{1}{L_{h,\hat{q}}}+\frac{\delta}{L_{h,q}}\right)^{-1}=\frac{L_{h,\hat{q}}L_{h,q}}{L_{h,q}+\delta L_{h,\hat{q}}} \quad \textit{and} \quad \left(\frac{1}{\mu_{h,\hat{q}}}+\frac{\delta}{\mu_{h,q}}\right)^{-1}=\frac{\mu_{h,\hat{q}}\mu_{h,q}}{\mu_{h,q}+\delta\mu_{h,\hat{q}}}.$$

The constants $\mu_{h,q}, \mu_{h,\hat{q}}, L_{h,q}, L_{h,\hat{q}}$ are defined in Definition 42.

Proof. This can be obtained via direct computation. ■

Proposition 45. *Suppose the Hessian $\nabla^2 f_{h,k}$ is (δ, q)-weakly connected as defined in Definition 42, then*

$$\hat{\mathbf{Q}}_{h,k}^{-1}(\hat{\mathbf{Q}}_{h,k}^{-1}+\delta\mathbf{Q}_{h,k}^{-1})^{-1}=\mathbf{I}-\delta\mathbf{Q}_{h,k}^{-1}(\hat{\mathbf{Q}}_{h,k}^{-1}+\delta\mathbf{Q}_{h,k}^{-1})^{-1},$$

and for any $\mathbf{r}_h \in \mathbb{R}^N$

$$\frac{1}{L_{h,\hat{q}}}\frac{\mu_{h,\hat{q}}\mu_{h,q}}{\mu_{h,q}+\delta\mu_{h,\hat{q}}}\|\mathbf{r}_h\| \leq \|\hat{\mathbf{Q}}_{h,k}^{-1}(\hat{\mathbf{Q}}_{h,k}^{-1}+\delta\mathbf{Q}_{h,k}^{-1})^{-1}\mathbf{r}_h\|.$$

Proof.

$$\begin{aligned}\mathbf{I} &= (\hat{\mathbf{Q}}_{h,k}^{-1}+\delta\mathbf{Q}_{h,k}^{-1})(\hat{\mathbf{Q}}_{h,k}^{-1}+\delta\mathbf{Q}_{h,k}^{-1})^{-1},\\ &= \hat{\mathbf{Q}}_{h,k}^{-1}(\hat{\mathbf{Q}}_{h,k}^{-1}+\delta\mathbf{Q}_{h,k}^{-1})^{-1}+\delta\mathbf{Q}_{h,k}^{-1}(\hat{\mathbf{Q}}_{h,k}^{-1}+\delta\mathbf{Q}_{h,k}^{-1})^{-1},\end{aligned}$$

and thus,

$$\mathbf{I}-\delta\mathbf{Q}_{h,k}^{-1}(\hat{\mathbf{Q}}_{h,k}^{-1}+\delta\mathbf{Q}_{h,k}^{-1})^{-1}=\hat{\mathbf{Q}}_{h,k}^{-1}(\hat{\mathbf{Q}}_{h,k}^{-1}+\delta\mathbf{Q}_{h,k}^{-1})^{-1}.$$

For the second part,

$$\begin{aligned}\|\hat{\mathbf{Q}}_{h,k}^{-1}(\hat{\mathbf{Q}}_{h,k}^{-1}+\delta\mathbf{Q}_{h,k}^{-1})^{-1}\mathbf{r}_h\|^2 &= \mathbf{r}^T(\hat{\mathbf{Q}}_{h,k}^{-1}+\delta\mathbf{Q}_{h,k}^{-1})^{-1}\hat{\mathbf{Q}}_{h,k}^{-1}\hat{\mathbf{Q}}_{h,k}^{-1}(\hat{\mathbf{Q}}_{h,k}^{-1}+\delta\mathbf{Q}_{h,k}^{-1})^{-1}\mathbf{r}_h,\\ &\geq \frac{1}{L_{h,\hat{q}}^2}\mathbf{r}_h^T(\hat{\mathbf{Q}}_{h,k}^{-1}+\delta\mathbf{Q}_{h,k}^{-1})^{-1}(\hat{\mathbf{Q}}_{h,k}^{-1}+\delta\mathbf{Q}_{h,k}^{-1})^{-1}\mathbf{r}_h,\\ &\geq \frac{1}{L_{h,\hat{q}}^2}\left(\frac{\mu_{h,\hat{q}}\mu_{h,q}}{\mu_{h,q}+\delta\mu_{h,\hat{q}}}\right)^2\|\mathbf{r}_h\|^2.\end{aligned}$$

So we obtain the desired result. ■

We derive the difference between the classical Newton's step and coarse correction step in the following lemma.

Lemma 46. *Suppose the function $f_h(\mathbf{x}_h)$ has (δ, q)-weakly connected Hessians as defined in Definition 42. Let*

$$\begin{aligned}\mathbf{d}_{h,k}^N &= -[\nabla^2 f_{h,k}]^{-1}\nabla f_{h,k},\\ \mathbf{d}_{h,k}^B &= -[\mathbf{Q}_{h,k}]^{-1}\nabla f_{h,k}.\end{aligned} \tag{35}$$

Then

$$\mathbf{d}_{h,k}^N=\mathbf{d}_{h,k}^B-\delta\mathbf{Q}_{h,k}^{-1}(\hat{\mathbf{Q}}_{h,k}^{-1}+\delta\mathbf{Q}_{h,k}^{-1})^{-1}\mathbf{d}_{h,k}^B.$$

Proof. The Newton's step is

$$\mathbf{d}_{h,k}^{\mathrm{N}}=-[\nabla^2 f_{h,k}]^{-1}\nabla f_{h,k}=-[\mathbf{Q}_{h,k}+\delta\hat{\mathbf{Q}}_{h,k}]^{-1}\nabla f_{h,k}.$$

Using Proposition 43, we have

$$\begin{aligned}
[\mathbf{Q}_{h,k}+\delta\hat{\mathbf{Q}}_{h,k}]^{-1} &= [\mathbf{Q}_{h,k}+\mathbf{I}(\delta\hat{\mathbf{Q}}_{h,k})\mathbf{I}]^{-1},\\
&= \mathbf{Q}_{h,k}^{-1}-\mathbf{Q}_{h,k}^{-1}(\delta^{-1}\hat{\mathbf{Q}}_{h,k}^{-1}+\mathbf{Q}_{h,k}^{-1})^{-1}\mathbf{Q}_{h,k}^{-1},\\
&= \mathbf{Q}_{h,k}^{-1}-\mathbf{Q}_{h,k}^{-1}(\delta\mathbf{I})(\delta\mathbf{I})^{-1}(\delta^{-1}\hat{\mathbf{Q}}_{h,k}^{-1}+\mathbf{Q}_{h,k}^{-1})^{-1}\mathbf{Q}_{h,k}^{-1},\\
&= \mathbf{Q}_{h,k}^{-1}-\delta\mathbf{Q}_{h,k}^{-1}(\hat{\mathbf{Q}}_{h,k}^{-1}+\delta\mathbf{Q}_{h,k}^{-1})^{-1}\mathbf{Q}_{h,k}^{-1}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbf{d}_{h,k}^{\mathrm{N}} &= -\left(\mathbf{Q}_{h,k}^{-1}-\delta\mathbf{Q}_{h,k}^{-1}(\hat{\mathbf{Q}}_{h,k}^{-1}+\delta\mathbf{Q}_{h,k}^{-1})^{-1}\mathbf{Q}_{h,k}^{-1}\right)\nabla f_{h,k},\\
&= \mathbf{d}_{h,k}^{\mathrm{B}}-\delta\mathbf{Q}_{h,k}^{-1}(\hat{\mathbf{Q}}_{h,k}^{-1}+\delta\mathbf{Q}_{h,k}^{-1})^{-1}\mathbf{d}_{h,k}^{\mathrm{B}},
\end{aligned}$$

as required. ■

7.5 Convergence

Using Proposition 45 and Lemma 46, we derive the composite convergence rate.

Theorem 47. *Suppose the function $f_h(\mathbf{x}_h)$ has (δ,q)-weakly connected Hessians defined in Definition 42. Suppose $\mathbf{d}_{h,k}^{B}$ in (35) is taken and $\alpha_{h,k}=1$, then*

$$\begin{aligned}
\|\mathbf{x}_{h,k+1}-\mathbf{x}_{h,\star}\| \;\leq\; & \delta\frac{\mu_{h,q}+\delta\mu_{h,\hat{q}}}{\mu_{h,\hat{q}}\mu_{h,q}^2}\frac{L_{h,\hat{q}}^2L_{h,q}}{L_{h,q}+\delta L_{h,\hat{q}}}\|\mathbf{x}_{h,k+1}-\mathbf{x}_{h,k}\|\\
&+L_{h,\hat{q}}\frac{\mu_{h,q}+\delta\mu_{h,\hat{q}}}{\mu_{h,\hat{q}}\mu_{h,q}}\frac{M_h}{2\mu_h}\|\mathbf{x}_{h,k}-\mathbf{x}_{h,\star}\|^2.
\end{aligned}$$

Proof. Using Lemma 46, we obtain

$$\begin{aligned}
\mathbf{x}_{h,k+1}-\mathbf{x}_{h,\star} &= \mathbf{x}_{h,k}-\mathbf{x}_{h,\star}+\mathbf{d}_{h,k}^{\mathrm{B}},\\
&= \mathbf{x}_{h,k}-\mathbf{x}_{h,\star}+\mathbf{d}_{h,k}^{\mathrm{B}}+\mathbf{d}_{h,k}^{\mathrm{N}}-\mathbf{d}_{h,k}^{\mathrm{N}},\\
&= \left(\mathbf{x}_{h,k}-\mathbf{x}_{h,\star}+\mathbf{d}_{h,k}^{\mathrm{N}}\right)+\delta\mathbf{Q}_{h,k}^{-1}(\hat{\mathbf{Q}}_{h,k}^{-1}+\delta\mathbf{Q}_{h,k}^{-1})^{-1}\mathbf{d}_{h,k}^{\mathrm{B}},\\
&= \left(\mathbf{x}_{h,k}-\mathbf{x}_{h,\star}+\mathbf{d}_{h,k}^{\mathrm{N}}\right)+\delta\mathbf{Q}_{h,k}^{-1}(\hat{\mathbf{Q}}_{h,k}^{-1}+\delta\mathbf{Q}_{h,k}^{-1})^{-1}(\mathbf{x}_{h,k+1}-\mathbf{x}_{h,k}).
\end{aligned}$$

Using the fact that

$$\begin{aligned}
&\mathbf{Q}_{h,k}^{-1}(\hat{\mathbf{Q}}_{h,k}^{-1}+\delta\mathbf{Q}_{h,k}^{-1})^{-1}(\mathbf{x}_{h,k+1}-\mathbf{x}_{h,k})\\
&\qquad=\mathbf{Q}_{h,k}^{-1}(\hat{\mathbf{Q}}_{h,k}^{-1}+\delta\mathbf{Q}_{h,k}^{-1})^{-1}(\mathbf{x}_{h,k+1}-\mathbf{x}_{h,\star})-\mathbf{Q}_{h,k}^{-1}(\hat{\mathbf{Q}}_{h,k}^{-1}+\delta\mathbf{Q}_{h,k}^{-1})^{-1}(\mathbf{x}_{h,k}-\mathbf{x}_{h,\star}),
\end{aligned}$$

we have

$$\begin{aligned}
&(\mathbf{I}-\delta\mathbf{Q}_{h,k}^{-1}(\hat{\mathbf{Q}}_{h,k}^{-1}+\delta\mathbf{Q}_{h,k}^{-1})^{-1})(\mathbf{x}_{h,k+1}-\mathbf{x}_{h,\star})\\
&\qquad=\left(\mathbf{x}_{h,k}-\mathbf{x}_{h,\star}+\mathbf{d}_{h,k}^{\mathrm{N}}\right)-\delta\mathbf{Q}_{h,k}^{-1}(\hat{\mathbf{Q}}_{h,k}^{-1}+\delta\mathbf{Q}_{h,k}^{-1})^{-1}(\mathbf{x}_{h,k}-\mathbf{x}_{h,\star}). \quad (36)
\end{aligned}$$

Using Proposition 45, we have

$$\begin{aligned}\frac{1}{L_{h,\hat{q}}}\frac{\mu_{h,\hat{q}}\mu_{h,q}}{\mu_{h,q}+\delta\mu_{h,\hat{q}}}\|\mathbf{x}_{h,k+1}-\mathbf{x}_{h,\star}\| &\leq \|(\mathbf{I}-\delta\mathbf{Q}_{h,k}^{-1}(\hat{\mathbf{Q}}_{h,k}^{-1}+\delta\mathbf{Q}_{h,k}^{-1})^{-1})(\mathbf{x}_{h,k+1}-\mathbf{x}_{h,\star})\|,\\ &\leq \|\mathbf{x}_{h,k}-\mathbf{x}_{h,\star}+\mathbf{d}_{h,k}^{\mathrm{N}}\|\\ &\quad +\delta\|\mathbf{Q}_{h,k}^{-1}\|\|(\hat{\mathbf{Q}}_{h,k}^{-1}+\delta\mathbf{Q}_{h,k}^{-1})^{-1}\|\|\mathbf{x}_{h,k}-\mathbf{x}_{h,\star}\|,\\ &\leq \frac{M_h}{2\mu_h}\|\mathbf{x}_{h,k}-\mathbf{x}_{h,\star}\|^2\\ &\quad +\frac{\delta}{\mu_{h,q}}\frac{L_{h,\hat{q}}L_{h,q}}{L_{h,q}+\delta L_{h,\hat{q}}}\|\mathbf{x}_{h,k}-\mathbf{x}_{h,\star}\|.\end{aligned}$$

Therefore,

$$\begin{aligned}\|\mathbf{x}_{h,k+1}-\mathbf{x}_{h,\star}\| &\leq L_{h,\hat{q}}\frac{\mu_{h,q}+\delta\mu_{h,\hat{q}}}{\mu_{h,\hat{q}}\mu_{h,q}}\frac{M_h}{2\mu_h}\|\mathbf{x}_{h,k}-\mathbf{x}_{h,\star}\|^2\\ &\quad +L_{h,\hat{q}}\frac{\mu_{h,q}+\delta\mu_{h,\hat{q}}}{\mu_{h,\hat{q}}\mu_{h,q}}\frac{\delta}{\mu_{h,q}}\frac{L_{h,\hat{q}}L_{h,q}}{L_{h,q}+\delta L_{h,\hat{q}}}\|\mathbf{x}_{h,k}-\mathbf{x}_{h,\star}\|,\end{aligned}$$

as required. ■

Theorem 47 shows that the coefficient of $\|\mathbf{x}_{h,k}-\mathbf{x}_{h,\star}\|$ is in $\mathcal{O}(\delta)$. As expected, as $\delta\to 0$, the composite rate in Theorem 47 will recover the quadratic convergence, and the linear component of composite rate decays at least linearly with δ.

8 Numerical Experiments

In this section, we will first verify our convergence results with three numerical examples. Each example will correspond to each of the settings in Section 5-7. The first example corresponds to Section 5, and it is an one-dimensional Poisson's equation, which is a standard example in numerical analysis and multigrid algorithms. In the second example, we consider the case in Section 6, and we use regularized logistic problem to be the illustrative example. In the third example, we consider a synthetic example to study the case in Section 7. We investigate the convergence by varying the parameter δ.

In the second part of this section, we will compare GAMA with other algorithms. We emphasize that the goal of this paper is to gain understanding in Galerkin-based multilevel algorithm, which apparently is closely related to many existing algorithms: ranging from conventional multigrid algorithms to machine learning-driven algorithms. The use of this section is to show the potential of Galerkin model, and we are not trying to claim that GAMA outperforms the state-of-the-art algorithms, including variants of GAMA.

8.1 Poisson's Equation

We consider an one-dimensional Poisson's equation

$$-\frac{\mathrm{d}^2}{\mathrm{d}q^2}u = w(q) \quad \text{in } [0,1], \quad u(0)=u(1)=0,$$

where $w(q)$ is chosen as

$$w(q)=\sin(4\pi q)+8\sin(32\pi q)+16\sin(64\pi q).$$

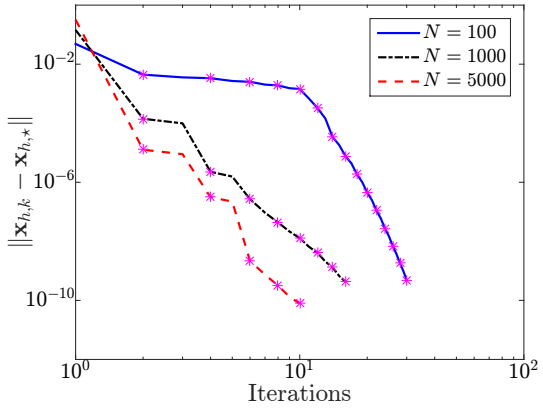

Figure 4: Convergence of solving Poisson's equation with different N's

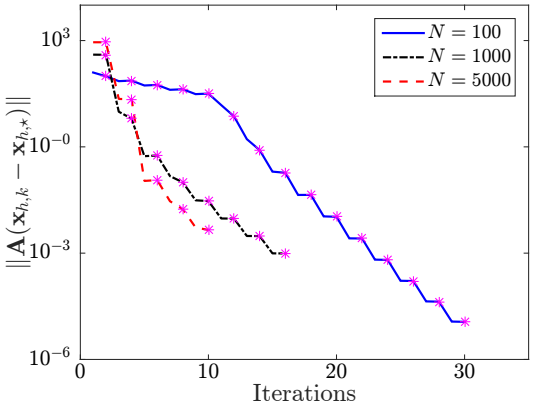

Figure 5: The smoothing effect with different N's

We discretize the above problem and denote $\mathbf{x}, \mathbf{b} \in \mathbb{R}^{N-1}$, where $(\mathbf{x})_i = u(i/N)$ and $(\mathbf{b})_i = w(i/N)$, for $i = 1, 2, \ldots, N-1$. By using finite difference, we approximate the above equation with

$$\min_{\mathbf{x} \in \mathbb{R}^{N-1}} \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} - \mathbf{b}^T\mathbf{x}, \tag{37}$$

where $\mathbf{A}$ is defined as in Lemma 32, which is a discretized Laplacian operator.

Figure 4 shows the convergence results of solving (37) with different N's. In this example we use the prolongation and restriction operators that are defined in (25) and (26). Since there is only one pair of $\mathbf{P}$ and $\mathbf{R}$, we follow the traditional multigrid approach in which we combine the coarse correction step with fine correction step. Steepest descent is used to compute the fine correction step. The pink stars in Figure 4 and Figure 5 indicate where coarse correction steps were used.

As expected from Corollary 33, the performance of convergence is inversely proportional to the discretization level N. More interestingly, one can see the complementary of fine correction step and coarse correction step. From Figure 4, fine correction steps are often deployed after coarse correction steps. Each pair of fine and coarse correction steps provides significant improvement in convergence. Figure 5 shows the smoothing effect of the fine correction step by looking at the quantity $\|\mathbf{A}(\mathbf{x}_{h,k} - \mathbf{x}_{h,\star})\|$, where $\mathbf{A}$ is the discretized Laplacian operator, as defined in Lemma 32. As opposed to coarse correction steps, fine correction steps are effective in reducing $\|\mathbf{A}(\mathbf{x}_{h,k} - \mathbf{x}_{h,\star})\|$. Once the error is smoothed, coarse correction steps provide large reduction in error, as shown in Figure 4.

8.2 Regularized Logistic Regression

We study the Galerkin model that is generated via navïe Nyström method and consider an example in ℓ_1 regularized logistic regression,

$$\min_{\mathbf{x} \in \mathbb{R}^N} \frac{1}{m} \sum_{i=1}^{m} \log(1 + \exp(-b_i \mathbf{x}^T \mathbf{a}_i)) + \omega_1 \|\mathbf{x}\|_1,$$

where $\omega_1 \in \mathbb{R}^+$, and $\{(\mathbf{a}_i, b_i)\}_{i=1}^m$ is a training set with m instances. For $i = 1, 2, \ldots, m$, $\mathbf{a}_i \in \mathbb{R}^N$ is an input and $b_i \in \mathbb{R}$ is the corresponding output.

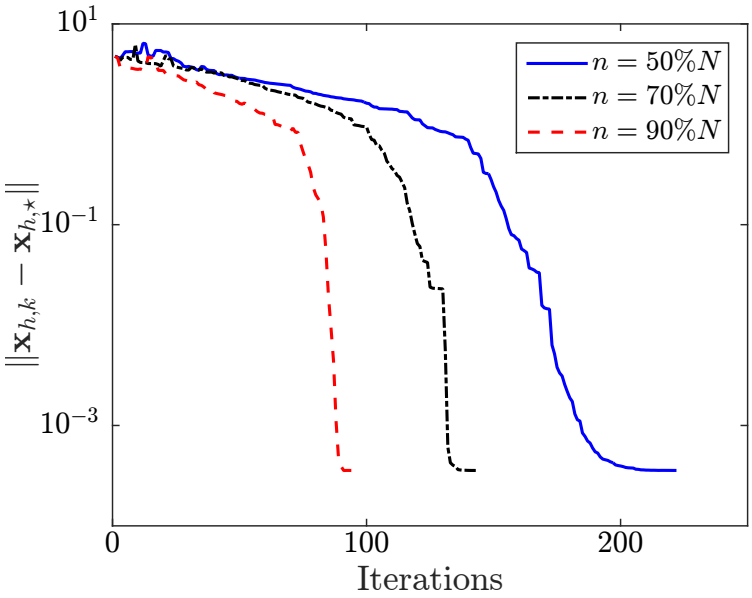

Figure 6: The ℓ_1 regularized logistic regression example.

Notice that the above formulation involves non-differentiable function $\|\mathbf{x}\|_1$, and so the above problem is beyond the scope of the setting in this paper. To overcome this issue, we replace the $\|\mathbf{x}\|_1$ with its approximation, the pseudo-Huber function [8], and solve the following formulation.

$$\min_{\mathbf{x}\in\mathbb{R}^N} \frac{1}{m}\sum_{i=1}^{m} \log(1+\exp(-b_i\mathbf{x}^T\mathbf{a}_i)) + \omega_1 \sum_{i=1}^{N} \left((\mu_r^2+\mathbf{x}_i^2)^{1/2} - \mu_r\right), \tag{38}$$

where $\mu_r \in \mathbb{R}^+$ is a parameter, and it provides good approximation of the ℓ_1 norm when μ_r is small.

The dataset *gisette* is used for $\{(\mathbf{a}_i, b_i)\}_{i=1}^m$. *Gisette* is a handwritten digits dataset from the NIPS 2003 feature selection challenge. In this example $N = 5000$, $m = 6000$, and we choose parameter ω from [23, 43] and $\mu_r = 0.001$. One can find and download *gisette* at *https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets*.

Notice that when $\mathbf{P} \in \mathbb{R}^{n\times N}$ and $\mathbf{R} \in \mathbb{R}^{n\times N}$ are generated using naïve Nyström method according to Definition 34, n is a user-defined parameter, and the probabilistic approach mentioned in Section 2 is used to generate multiple $\mathbf{P}$'s and $\mathbf{R}$'s. That is, a pair of $\mathbf{P}$ and $\mathbf{R}$ is sampled uniformly over $\binom{N}{n}$ possible coarse models. This setting satisfies the condition stated in Proposition 7, and so no fine correction step is needed.

Figure 6 shows the convergence results. As expected from Corollary 38 and 39, the performance of convergence is proportional to n.

8.3 A Synthetic Example for Block Diagonal Approximation

To study the case of block diagonal approximation in Section 7, we construct an artificial example with weakly connected Hessian. In particular, we solve

$$\min_{\mathbf{x}\in\mathbb{R}^N} \frac{1}{2}\mathbf{x}^T\left(\mathbf{Q}_h + \delta\hat{\mathbf{Q}}_h\right)\mathbf{x} + \mathbf{b}^T\mathbf{x}, \tag{39}$$

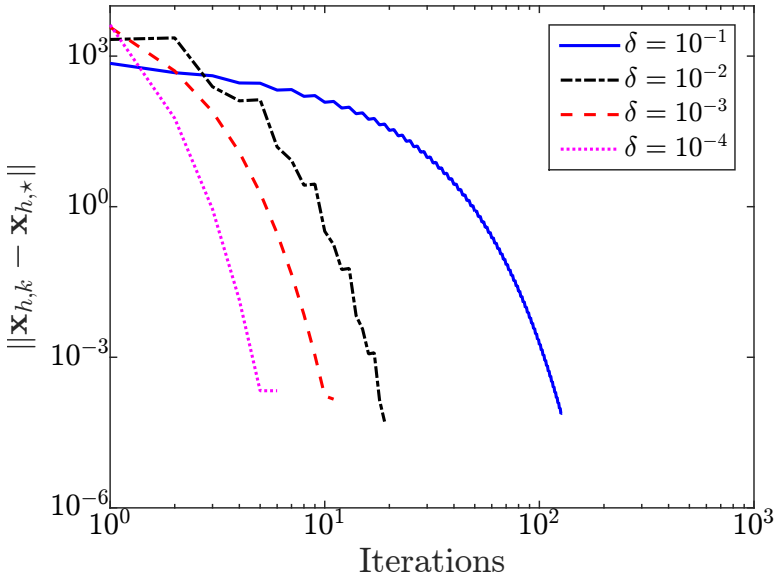

Figure 7: Block diagonal approximation.

where $\delta \in \mathbb{R}^+$, $\mathbf{Q}_h = \text{diag}(\mathbf{Q}_{h,1}, \mathbf{Q}_{h,2}, \ldots, \mathbf{Q}_{h,p})$ is a block diagonal matrix with p blocks, with $\mathbf{Q}_{h,i}(\mathbf{x}_h) \in \mathbb{R}^{n_i \times n_i}$ and $\sum_{i=1}^{p} n_i = n$ for $i = 1, 2, \ldots, p$. In this example, we have $N = 1000$, $p = 10$, $n_1 = n_2 = \cdots = n_{10} = 100$. We construct $\hat{\mathbf{Q}}_h$ via

$$\hat{\mathbf{Q}}_h = \sum_{j=1}^{N} v_j \mathbf{u}_j \mathbf{u}_j^T,$$

where $v_j \in \mathbb{R}^+$ is sampled uniformly from $[v_\delta, 1 + v_\delta]$, and $\mathbf{u}_j \in \mathbb{R}^N$ is a random orthonormal vectors, for $j = 1, 2, \ldots, N$. Each $\mathbf{Q}_{h,i}$ is also constructed similar to $\hat{\mathbf{Q}}_h$ but in the smaller dimension $\mathbb{R}^{n_i \times n_i}$, for $i = 1, 2, \ldots, p$. $v_\delta = 0.0001$ in this example.

We consider the optimization problem in (39) with different δ's. As expected from Theorem 47, Figure 7 shows that the performance of convergence is inversely proportional to δ.

8.4 Numerical Performance: PDE Test Cases

We now compare the numerical performance of GAMA with the conventional unconstrained optimization algorithms as well as conventional line search multilevel/multigrid algorithm in [39]. We focus on PDE-based optimization problems in this section.

We test algorithms on five examples from [39, 13], and all of them are discretized 2-dimensional variational problems over unit square $\mathcal{S}_2 \triangleq [0, 1] \times [0, 1]$. The decision variable, $u(x, y)$, obeys the boundary condition, $u = 0$ on $\partial\mathcal{S}_2$, for all problems. The five problems are listed in the following.

1. Problem DSSC:

$$\min_{u \in \mathcal{S}_2} \int_{\mathcal{S}_2} \frac{1}{2} \|\nabla u(x, y)\|^2 - \lambda \exp(u(x, y)), \quad \text{where} \quad \lambda = 6.$$

2. Problem WEN:

$$\min_{u\in\mathcal{S}_2}\int_{\mathcal{S}_2}\frac{1}{2}\|\nabla u(x,y)\|^2+\lambda\exp\left[u(x,y)\right](u(x,y)-1)-\gamma(x,y)u(x,y),$$

where $\lambda=6$ and

$$\gamma(x)=\left[\left(9\pi^2+\lambda\exp\left[\left(x^2-x^3\right)\sin(3\pi y)\right]\right)\left(x^2-x^3\right)+6x-2\right]\sin\left(3\pi y\right).$$

3. Problem BRATU:

$$\min_{u\in\mathcal{S}_2}\int_{\mathcal{S}_2}\|\Delta u(x,y)-\lambda\exp(u(x,y))\|^2,\quad\text{where}\quad\lambda=6.8.$$

4. Problem POSSION2D:

$$\min_{\mathbf{x}}\frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x}-\mathbf{b}^T\mathbf{x},$$

where $\mathbf{A}$ and $\mathbf{b}$ are the discretizations of the Laplacian and the function $\gamma(x,y)=2(y(1-y)+x(1-x))$, respectively.

5. Problem IGNISC:

$$\min_{u\in\mathcal{S}_2}\int_{\mathcal{S}_2}(u(x,y)-z)^2+\frac{\beta}{2}\int_{\mathcal{S}_2}(\exp[u(x,y)]-\exp[z])^2 + \frac{\nu}{2}\int_{\mathcal{S}_2}\|\Delta u(x,y)-\delta\exp(u(x,y))\|^2,$$

where $\delta=\beta=6.8$, $\nu=10^{-5}$, and $z=1/\pi^2$.

	DSSC		WEN		BRATU	
	Time	Accuracy	Time	Accuracy	Time	Accuracy
L-BFGS	5524.9	9.5164e-06	1048.7	9.9788e-06	44355	12.449
Newton	59.01	1.3351e-07	47.6	4.3493e-08	565.79	2.0853e-06
GAMA-NT	21.8	1.153e-13	21.26	4.6412e-12	180.19	1.9028e-06
COMA-NT	20.13	1.1531e-13	20.19	4.6412e-12	161.52	1.9027e-06
GAMA-qNT	13	7.2882e-06	5.1	7.9565e-06	840.47	0.0021644
COMA-qNT	12.52	9.4619e-06	6.43	7.3332e-06	860.4	37.708

	POSSION2D		IGNISC	
	Time	Accuracy	Time	Accuracy
L-BFGS	1105.7	7.815e-06	50274	0.00039108
Newton	15.99	7.2561e-15	124.93	2.1409e-06
GAMA-NT	20.93	0	77.58	2.0008e-11
COMA-NT	20.92	0	77.69	2.0008e-11
GAMA-qNT	1.28	8.1249e-06	62.81	9.0643e-06
COMA-qNT	1.46	8.1304e-06	43.13	9.1113e-06

Table 1: PDE-based text examples

Table 1 shows the numerical performance of different algorithms, i.e., the CPU time (Time) needed to achieve small $\|\nabla f_{h,k}\|$ (Accuracy). We denote the COnventional Multilevel Algorithm as *COMA*. For both GAMA and COMA, we denote "-NT" and "-qNT" when Newton's method and L-BFGS are used for fine correction steps, respectively. For all five examples, we choose the fine models to be the discretization with mesh size $\Delta x \times \Delta y$, where $\Delta x = \Delta x = 1/2^{10}$, and the standard five-point finite differences are used. We point out that all the algorithmic settings are the same as in [39], including line search strategy, stopping criteria, and choice of parameters.

For both GAMA and COMA, we follow the same strategy as in [39], and the standard full multilevel scheme is deployed. Suppose level j is denoted as the discretization with mesh size $\Delta x \times \Delta y$, where $\Delta x = \Delta y = 1/2^j$. For $j = 3, 4, \ldots, 9$, we compute the solution $\mathbf{x}_{j,\star}$ in level j, and use $\mathbf{P}_j^{j+1}\mathbf{x}_{j,\star}$ as the initial guess for level $j+1$. $\mathbf{P}_j^{j+1}$ is denoted as the prolongation operator from level j to level $j+1$.

From Table 1, we can see that the multilevel algorithms clearly outperform the conventional algorithms. The performance of GAMA is comparable with COMA and is more robust due to the use of second order information. In the problem BRATU, first order algorithms (i.e. L-BFGS, GAMA-qNT, and COMA-qNT) are not efficient, but GAMA-qNT is able to achieve much better accuracy. Therefore, GAMA is empirically competitive against the conventional multilevel algorithm, and yet more robust with a more understandable rate of convergence.

8.5 Numerical Performance: Machine Learning Test Cases

We now study the performance of GAMA that is generated by Nyström method. Suppose we have the training set $\{(\mathbf{a}_i, b_i)\}_{i=1}^m$, we use GAMA to solve the empirical risk minimization (ERM) problem

$$\min_{\mathbf{x}\in\mathbb{R}^N} \frac{1}{m}\sum_{i=1}^{m} f_i(\mathbf{a}_i^T\mathbf{x}) + \omega_1\|\mathbf{x}\|_1 + \frac{\omega_2}{2}\|\mathbf{x}\|_2^2,$$

where $\omega_1, \omega_2 \in \mathbb{R}$ and $\mathbf{a}_i \in \mathbb{R}^N$, for $i = 1, 2, \ldots, m$. Special cases of f_i include

1. Quadratic loss function: $f_i(x) = \frac{1}{2}(x - b_i)^2$.
2. Logistic loss function: $f_i(x) = \log(1 + \exp(-xb_i))$.

In the case where $w_1 = 0$ and f_i's are logistic loss functions, we yield to the ℓ_1 regularized logistic regression as in Section 8.2. Similar to Section 8.2, we replace the $\|\mathbf{x}\|_1$ with the pseudo-Huber function.

The numerical test is conducted over five examples. All dataset/training set can be download at *https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/*. Table 2 provides details of the test examples. We point out that for logistics regression, we select the choice of ω_1 and ω_2 based on [23, 12]. For linear regression, we simply select $\omega_1 = \omega_2 = 10^{-6}$, which is a commonly used value.

Figure 8-12 show the numerical performance of GAMA, compared to Newton's method and L-BFGS. Over these five examples, GAMA only performs coarse correction steps, and n is chosen to be $10\%N$, $20\%N$, and $30\%N$. An exception can be found in log1pE2006test because these choices of n's are too large to be traceable. The performance of Newton's method is also missing for log1pE2006test because computing its search direction is intractable due to the size of N. From Figure 8 and 10, we can see that when N is small, Newton's method outperforms the other methods. This is not surprising since the per-iteration cost is cheap for small N and yet Newton's method enjoys the quadratic convergence. When N is sufficiently large, as showed in Figure 9 and 11, GAMA is competitive compared to both Newton's method and L-BFGS.

	f_i's	N	m	ω_1	ω_2
YearPredictionMSDt	Quadratic	90	51630	10^{-6}	10^{-6}
log1pE2006test	Quadratic	4272226	3308	10^{-6}	10^{-6}
w8at	Logistic	300	14951	0	$1/m$
Gisette	Logistic	5000	6000	$1/(0.25m)$	0
epsilon_normalizedt	Logistic	2000	100000	0	$1/m$

Table 2: Details of ERM Test Examples

In Figure 12, we can see that the performance of GAMA's is better than Newton's method and similar to L-BFGS. From Table 2, $N = 2000$ and it is a reasonably good size for Newton's method. The poor performance of Newton's method is due to the large m, which is 100000. For large m, the evaluation of Hessians becomes the computational bottleneck. To further illustrate this, in Figure 13, we perform Newton's method and GAMA with sub-sampling. For subsample Newton's method, at each iterations, we evaluate Hessian based on $\sqrt{m}$ data points in the training set. Data points are sampled uniformly without replacement. For GAMA, we deploy the idea of SVRG, sample $\sqrt{m}$ data points at each coarse correction step, and create a coarse model with

$$f_H(\mathbf{x}) = \frac{1}{\sqrt{m}} \sum_{i \in \mathcal{S}_H} f_i(\mathbf{a}_i^T \mathbf{x}) + \omega_1 \sum_{i=1}^{N} \left((\mu_r^2 + \mathbf{x}_i^2)^{1/2} - \mu_r \right) + \frac{\omega_2}{2} \|\mathbf{x}\|_2^2,$$

where $\mu_r = 0.001$ and $\mathcal{S}_H$ is the set of the samples. We call the coarse model with above f_H as intermediate coarse model. When solving intermediate coarse model, we apply the Galerkin-model that is generated by Nyström method, and apply five coarse correction steps. The incumbent solution of intermediate coarse model is then prolonged to the fine model and results in a coarse correction step on the fine model. We clarify that this algorithmic procedure follows the idea of SVRG, as introduced in Section 2.5. As shown in Figure 13, great improvements are achieved for both (subsample) Newton's method and GAMA's. The computational bottleneck of evaluating Hessians is removed by subsampling data points. Since solving a system of 2000 linear equations can be managed easily, Newton's method outperforms all the other method in this case. Notice that since the Hessians are not evaluated exactly in this case, Newton's method and GAMA no longer enjoy quadratic rate and composite rate, respectively. The theoretical performance of these methods are beyond the scope of this paper.

9 Comments and Perspectives

We showed the connections between the general multilevel framework and the conventional optimization methods. The case of using Galerkin model (GAMA) is further studied, and the local composite rate of convergence is derived. When the coefficient of the linear component in composite rate is sufficiently small, then GAMA is superior to Newton's method in complexity. This linear component is then studied in three different cases, and we showed how the structure in each case would improve the rate of convergence.

This work advances research in multilevel optimization algorithms in several non-exploited directions. Firstly, the connections between multilevel framework and standard optimization methods would motivate systematic designs in optimization algorithms, and the multilevel framework could be used beyond the traditional linesearch multilevel method in [39].

Secondly, we take the first step in showing how the structure of problems could improve the convergence. We expect that similar manner of thinking could be applied beyond GAMA, and we believe this line of research could motivate more developments in multilevel algorithms when one tries to tackle problems with specific structure.

We believe the results presented in this paper can be generalized and refined. For example, the local composite rate of convergence when solving PDE-based optimization can be extended to cases beyond one-dimensional problems or uniform discretization. These extensions would require more careful and tedious algebra, but the general approach presented in Section 5 can be applied. On the other hand, one can extend results in Section 6 by considering different versions of Nyström method, or even different methods in low rank approximation in general. These generalizations could be done under the general approach of this paper.

References

[1] A. Beck and L. Tetruashvili. On the convergence of block coordinate descent type methods. *SIAM Journal on Optimization*, 23(4):2037–2060, 2013.

[2] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.

[3] W. L. Briggs, V. E. Henson, and S. F. McCormick. *A Multigrid Tutorial*. SIAM, 2000.

[4] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

[5] T. F. Chan and B. F. Smith. Domain decomposition and multigrid algorithms for elliptic problems on unstructured meshes. In *Domain decomposition methods in scientific and engineering computing (University Park, PA, 1993)*, volume 180 of *Contemp. Math.*, pages 175–189. Amer. Math. Soc., Providence, RI, 1994.

[6] P. Drineas and M. W. Mahoney. On the Nyström method for approximating a Gram matrix for improved kernel-based learning. *Journal of Machine Learning Research*, 6:2153–2175, 2005.

[7] M. A. Erdogdu and A. Montanari. Convergence rates of sub-sampled newton methods. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 3052–3060, 2015.

[8] K. Fountoulakis and J. Gondzio. A second-order method for strongly convex ℓ_1-regularization problems. *Mathematical Programming*, 156(1-2, Ser. A):189–219, 2016.

[9] K. Fountoulakis and R. Tappenden. A Flexible Coordinate Descent Method for Big Data Applications. *ArXiv e-prints*, July 2015.

[10] A. Gittens. The spectral norm error of the naive Nystrom extension. *ArXiv e-prints*, October 2011.

[11] A. Gittens. *Topics in Randomized Numerical Linear Algebra*. ProQuest LLC, Ann Arbor, MI, 2013. Thesis (Ph.D.)–California Institute of Technology.

[12] R. M. Gower, D. Goldfarb, and P. Richtárik. Stochastic Block BFGS: Squeezing More Curvature out of Data. *ArXiv e-prints*, March 2016.

[13] S. Gratton, M. Mouffe, A. Sartenaer, P. L. Toint, and D. Tomanos. Numerical experience with a recursive trust-region method for multilevel nonlinear bound-constrained optimization. *Optimization Methods and Software*, 25(3):359–386, 2010.

[14] S. Gratton, A. Sartenaer, and P. L. Tonint. Recursive trust-region methods for multiscale nonlinear optimization. *SIAM Journal on Optimization*, 19:414–444, 2008.

[15] W. Hackbusch. *Multigrid methods and applications*, volume 4 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 1985.

[16] W. Hackbusch. *Multi-Grid Methods and Applications*. Springer, 2003.

[17] J. Han, Y. Yang, and H. Bi. A new multigrid finite element method for the transmission eigenvalue problems. *Applied Mathematics and Computation*, 292:96–106, 2017.

[18] C. P. Ho and P. Parpas. Singularly perturbed Markov decision processes: a multiresolution algorithm. *SIAM J. Control Optim.*, 52(6):3854–3886, 2014.

[19] V. Hovhannisyan, P. Parpas, and S. Zafeiriou. MAGMA: Multi-level accelerated gradient mirror descent algorithm for large-scale convex composite minimization. *ArXiv e-prints*, September 2015.

[20] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 315–323. Curran Associates, Inc., 2013.

[21] L. V. Kantorovich. *Functional analysis and applied mathematics*. NBS Rep. 1509. U. S. Department of Commerce, National Bureau of Standards, Los Angeles, Calif., 1952. Translated by C. D. Benster.

[22] M. Kočvara and S. Mohammed. A first-order multigrid method for bound-constrained convex optimization. *Optimization Methods and Software*, 31(3):622–644, 2016.

[23] J. D. Lee, Y. Sun, and M. A. Saunders. Proximal Newton-type methods for minimizing composite functions. *SIAM Journal on Optimization*, 24(3):1420–1443, 2014.

[24] R. M. Lewis and S. G. Nash. Model problems for the multigrid optimization of systems governed by differential equations. *SIAM Journal on Scientific Computing*, 26(6):1811–1837 (electronic), 2005.

[25] R. M. Lewis and S. G. Nash. Using inexact gradients in a multilevel optimization algorithm. *Computational Optimization and Applications*, 56(1):39–61, 2013.

[26] P. Moritz, R. Nishihara, and M. I. Jordan. A linearly-convergent stochastic L-BFGS algorithm. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016, Cadiz, Spain, May 9-11, 2016*, pages 249–258, 2016.

[27] S. G. Nash. A multigrid approach to discretized optimization problems. *Optimization Methods and Software*, 14(1-2):99–116, 2000. International Conference on Nonlinear Programming and Variational Inequalities (Hong Kong, 1998).

[28] S. G. Nash. Properties of a class of multilevel optimization algorithms for equality-constrained problems. *Optimization Methods and Software*, 29(1):137–159, 2014.

[29] P. Parpas, D. V. N. Luong, D. Rueckert, and B. Rustem. A multilevel proximal algorithm for large scale composite convex optimization.

[30] B. T. Polyak. *Introduction to optimization*. Translations Series in Mathematics and Engineering. Optimization Software, Inc., Publications Division, New York, 1987. Translated from the Russian, With a foreword by Dimitri P. Bertsekas.

[31] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical recipes: the art of scientific computing. Code CD-ROM v 2.06 with Windows, DOS, or Macintosh single-screen license*. Cambridge University Press, Cambridge, 1996.

[32] Z. Qu, P. Richtárik, M. Takác, and O Fercoq. SDNA: stochastic dual newton ascent for empirical risk minimization. *CoRR*, abs/1502.02268, 2015.

[33] A. J. Smola and B. Schökopf. Sparse greedy matrix approximation for machine learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 911–918, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[34] G. Strang. *Computational science and engineering*. Wellesley-Cambridge Press, Wellesley, MA, 2007.

[35] K. Stüben. A review of algebraic multigrid. *Journal of Computational and Applied Mathematics*, 128(1-2):281–309, 2001. Numerical analysis 2000, Vol. VII, Partial differential equations.

[36] J. A. Tropp. Improved analysis of the subsampled randomized Hadamard transform. *Advances in Adaptive Data Analysis. Theory and Applications*, 3(1-2):115–126, 2011.

[37] U. Trottenberg, C. W. Oosterlee, and A. Schüller. *Multigrid*. Academic Press, Inc., San Diego, CA, 2001. With contributions by A. Brandt, P. Oswald and K. Stüben.

[38] D.J. Tylavsky and G.R.L. Sohie. Generalization of the matrix inversion lemma. *Proceedings of the IEEE*, 74(7):1050–1052, July 1986.

[39] Z. Wen and D. Goldfarb. A line search multigrid method for large-scale nonlinear optimization. *SIAM Journal on Optimization*, 20(3):1478–1503, 2009.

[40] P. Wesseling. *An introduction to multigrid methods*. Pure and Applied Mathematics (New York). John Wiley & Sons, Ltd., Chichester, 1992.

[41] C. K. I. Williams and M. Seeger. Using the nyström method to speed up kernel machines. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 682–688. MIT Press, 2001.

[42] D. Xie and Q. Ni. An incomplete Hessian Newton minimization method and its application in a chemical database problem. *Computational Optimization and Applications*, 44(3):467–485, 2009.

[43] G.-X. Yuan, C.-H. Ho, and C.-J. Lin. An improved glmnet for l1-regularized logistic regression. *Journal of Machine Learning Research*, 13(1):1999–2030, June 2012.

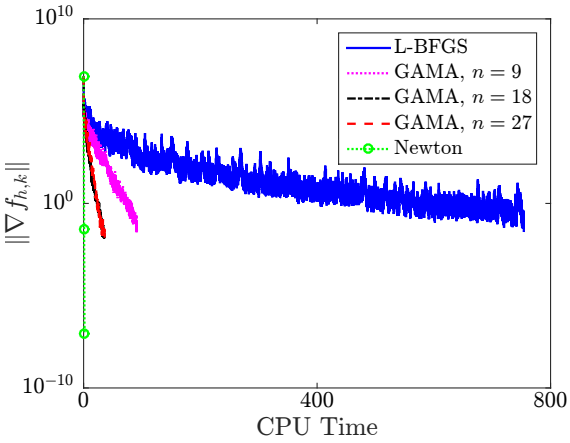

Figure 8: YearPredictionMSDt

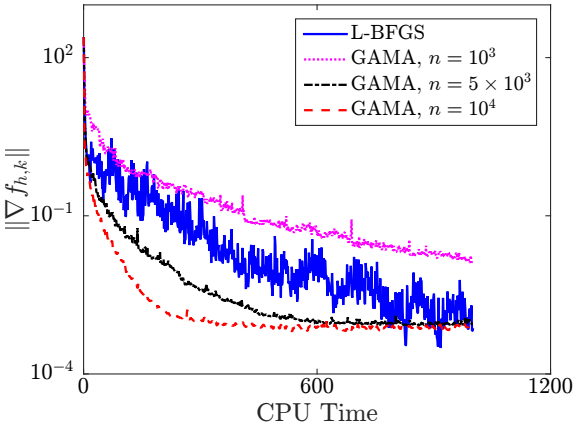

Figure 9: log1pE2006test

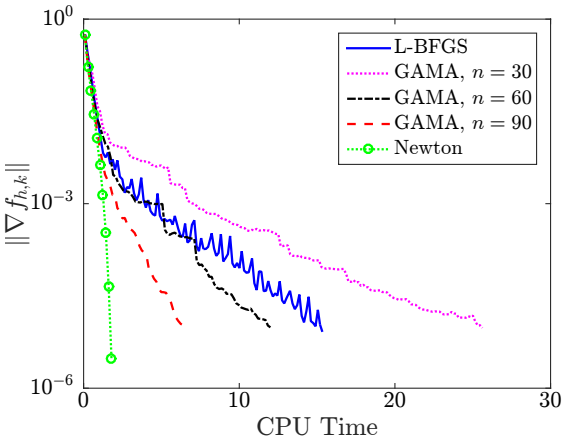

Figure 10: w8at

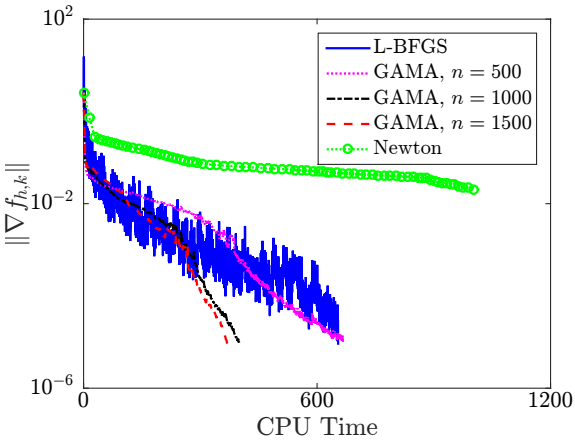

Figure 11: Gisette

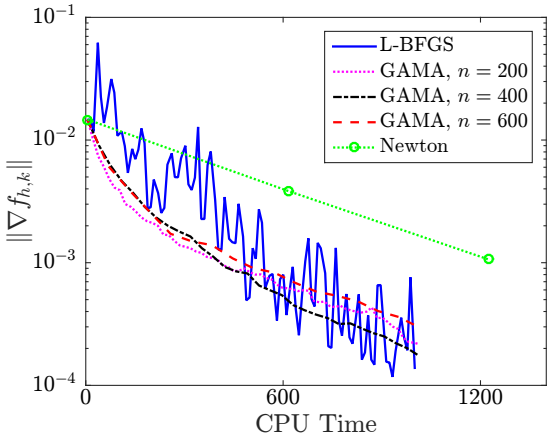

Figure 12: epsilon_normalizedt

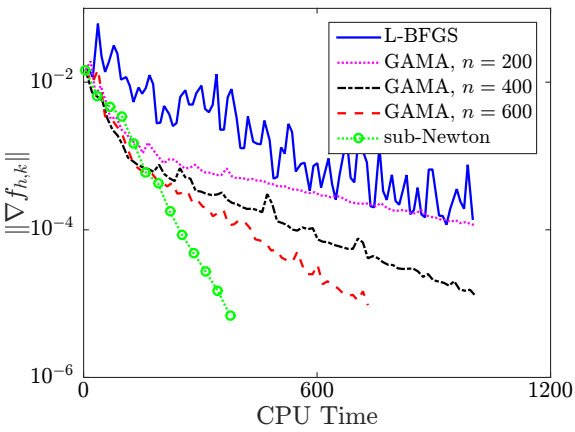

Figure 13: epsilon_normalizedt (subsample)