Enlarged Krylov Subspace Conjugate Gradient Methods for Reducing Communication
Laura Grigori, Sophie Moufawad, Frédéric Nataf

To cite this version:
Laura Grigori, Sophie Moufawad, Frédéric Nataf. Enlarged Krylov Subspace Conjugate Gradient Methods for Reducing Communication. [Research Report] RR-8597, INRIA. 2014. hal-01065985

HAL Id: hal-01065985
https://inria.hal.science/hal-01065985
Submitted on 18 Sep 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Enlarged Krylov Subspace Conjugate Gradient Methods for Reducing Communication

Laura Grigori, Sophie Moufawad, Frederic Nataf
Enlarged Krylov Subspace Conjugate Gradient Methods for Reducing Communication

rapport de recherche

Inria

Laura Grigori, Sophie Moufawad, Frederic Nataf

Équipes-Projets
ALPINES

Rapport de recherche n° 8597 — September 2014 — 46 pages

**Résumé** : Dans cet article, nous présentons deux nouveaux méthodes itératives pour la résolution des systèmes linéaires d’équations de très grande taille en minimisant les communications. Ces deux méthodes sont basées sur l’enrichissement de l’espace de Krylov en décomposant le domaine de $A$.

**Mots-clés** : algèbre linéaire, méthodes itératives
Enlarged Krylov Subspace Conjugate Gradient Methods for Reducing Communication

Abstract: In this paper we introduce a new approach for reducing communication in Krylov subspace methods that consists of enlarging the Krylov subspace by a maximum of $t$ vectors per iteration, based on the domain decomposition of the graph of $A$. The obtained enlarged Krylov subspace $\mathcal{K}_{t,k}(A, r_0)$ is a superset of the Krylov subspace $\mathcal{K}_k(A, r_0)$. Thus it is possible to search for the solution of the system $Ax = b$ in $\mathcal{K}_{t,k}(A, r_0)$ instead of $\mathcal{K}_k(A, r_0)$. Moreover, we show in this paper that the enlarged Krylov projection subspace methods lead to faster convergence in terms of iterations and parallelizable algorithms with less communication, with respect to Krylov methods.

In this paper we focus on Conjugate Gradient (CG) [16], a Krylov projection method for symmetric (Hermitian) positive definite matrices. We discuss two new versions of Conjugate Gradient (section 3). The first method, multiple search direction with orthogonalization CG (MSDO-CG), is an adapted version of MSD-CG [14] with the A-orthonormalization of the search directions to obtain a projection method that guarantees convergence at least as fast as CG. The second projection method that we propose here, long recurrence enlarged CG (LRE-CG), is similar to GMRES in that we build an orthonormal basis for the enlarged Krylov subspace rather than finding search directions. Then, we use the whole basis to update the solution and the residual. Both methods converge faster than CG in terms of iterations, but LRE-CG converges faster than MSDO-CG since it uses the whole basis to update the solution rather than only $t$ search directions. And the more subdomains are introduced or the larger $t$ is, the faster is the convergence of both methods with respect to CG in terms of iterations. For example, for $t = 64$ the MSDO-CG and LRE-CG methods converge in 75% up to 98% less iteration with respect to CG for the different test matrices. But increasing $t$ also means increasing the memory requirements. Thus, in practice, $t$ should be relatively small, depending on the available memory, on the size of the matrix, and on the number of iterations needed for convergence, as explained in section 4. We also present the parallel algorithms along with their expected performance based on the estimated run times, and the preconditioned versions with their convergence behavior.

Key-words: minimizing communication, linear algebra, iterative methods
1 Introduction

Krylov subspace methods are among the most practical and popular iterative methods today. They are polynomial iterative methods that aim to solve systems of linear equations \((Ax = b)\) by finding a sequence of vectors \(x_1, x_2, x_3, x_4, ..., x_k\) that minimizes some measure of error over the corresponding spaces

\[ x_0 + \mathcal{K}_i(A, r_0), \quad i = 1, ..., k \]

where \(x_0\) is the initial iterate, \(r_0\) is the initial residual, and \(\mathcal{K}_i(A, r_0) = \text{span}\{r_0, A_1r_0, A^2r_0, ..., A^{i-1}r_0\}\) is the Krylov subspace of dimension \(i\). Conjugate Gradient (CG) [16], Generalized Minimal Residual (GMRES) [25], bi-Conjugate Gradient [19, 7], and bi-Conjugate Gradient Stabilized [27] are some of the most used Krylov subspace methods.

These methods are governed by Blas1 and Blas2 operations as dot products and sparse matrix vector multiplications. Parallelizing dot products is constrained by communication since the performed computation is negligible. If the dot products are performed by one processor, then there is a need for a communication before and after the computation. In both cases, communication is a bottleneck. This problem has been tackled by different approaches. First, block methods that solve system with multiple right-hand sides \(AX = B\) were introduced, as block CG [22]. Then, s-step methods that compute \(s\) basis vectors per iteration were proposed, examples are s-step CG [28, 2] and s-step GMRES [29, 6]. Both methods, block and s-step, use Blas2 and Blas3 operations. Recently, communication avoiding methods, based on s-step methods, that aim at avoiding communication at the expense of performing some redundant flops were introduced, as CA-CG, CA-GMRES [21, 17] and CA-ILU0 preconditioner [11]. Another approach is to hide the cost of communication by overlapping it with other computation, like pipelined CG [5, 13] and pipelined GMRES [8].

In this paper we introduce a new approach that consists of enlarging the Krylov subspace by a maximum of \(t\) vectors per iteration. First, the input matrix is partitioned into \(t\) sub-domains by using a graph partitioning algorithm. At the beginning of the iterative method, the residual is split into \(t\) vectors corresponding to the \(t\) sub-domains. Then, the obtained \(t\) vectors are multiplied by \(A\) at each iteration to generate \(t\) new basis vectors. The obtained enlarged Krylov subspace \(\mathcal{K}_{t,k}(A, r_0)\) is a superset of the Krylov subspace \(\mathcal{K}_k(A, r_0)\). Thus it is possible to search for the solution of the system \(Ax = b\) in \(\mathcal{K}_{t,k}(A, r_0)\) instead of \(\mathcal{K}_k(A, r_0)\). Moreover, we show in this paper that the enlarged Krylov projection subspace methods lead to faster convergence in terms of iterations and parallelizable algorithms with less communication, with respect to Krylov methods.

In this paper we focus on Conjugate Gradient (CG) [16], a Krylov projection method for symmetric (Hermitian) positive definite matrices, which was introduced by Hestenes and Stiefel in 1952 (section 2.1). After giving a brief overview of related existing CG methods (section 2) such as block-CG [22], coop-CG [1], and MSD-CG [14], we discuss two new versions of Conjugate Gradient (section 3). The first method, multiple search direction with orthogonalization CG (MSDO-CG), is an adapted version of MSD-CG [14]. MSD-CG has the same structure as the classical conjugate gradient method where first new search directions are defined on the \(t\) subdomains, then the \(t\) step lengths are obtained by solving a \(t \times t\) system, and finally the solution and the residual are updated. But unlike CG, the search directions are not \(A\)-orthogonal. Thus, in MSD-CG we \(A\)-orthonormalize the search directions, to obtain a projection method that guarantees convergence at least as fast as CG. The idea of using more than one search direction was also exploited in Rixen’s thesis [23] for two subdomains in the context of domain decomposition methods, and further developed in [10]. The second method that we propose here, long recurrence enlarged CG (LRE-CG), is similar to GMRES in that we build an orthonormal basis for the enlarged Krylov subspace rather than finding search directions. Then, we use the whole basis to update the solution and the residual. We show that this method is a projection method and hence should converge at least as fast as CG. We compare the convergence behavior of both methods using different \(A\)-orthonormalization and orthonormalization methods and then we compare the most stable versions with CG and other related methods (section 4).
We have tested our methods on matrices arising from the discretization of 2D poisson equations (POISSON2D), 3D elasticity equations (ELASTICITY3D), and 2D and 3D convection-diffusion equations such as Nh2D, SKY2D, SKY3D, and ANI3D as discussed in section 4. Both methods converge faster than CG in terms of iterations, but LRE-CG converges faster than MSDO-CG since it uses the whole basis to update the solution rather than only \( t \) search directions. And the more subdomains are introduced or the larger \( t \) is, the faster is the convergence of both methods with respect to CG in terms of iterations. For example, for \( t = 64 \) the MSDO-CG and LRE-CG methods converge in 75% to 82% less iteration with respect to CG for the matrices Nh2D, POISSON2D, and ELASTICITY3D, and 95% to 98% less iteration with respect to CG for the matrices SKY2D, SKY3D, and ANI3D. But increasing \( t \) also means increasing the memory requirements. Thus, in practice, \( t \) should be relatively small, depending on the available memory, on the size of the matrix, and on the number of iterations needed for convergence, as explained in section 4. We present the parallel algorithms along with their expected performance based on the estimated run times in section 5. And in section 6 we introduce the preconditioned version with its convergence behavior.

2 Overview of Existing Conjugate Gradient (CG) Methods

The Krylov projection methods find a sequence of approximate solutions \( x_k \) \((k > 0)\) of the system \( Ax = b \), and are defined by the following two conditions:

1. Subspace condition: \( x_k \in x_0 + K_k(A, r_0) \)

2. Petrov-Galerkin condition: \( r_k \perp \mathcal{L}_k \) \( \iff (r_k)^Ty = 0, \forall y \in \mathcal{L}_k \)

where \( x_0 \) is the initial iterate, \( r_0 \) is the initial residual, \( K_k(A, r_0) = \text{span}\{r_0, Ar_0, A^2r_0, ..., A^{k-1}r_0\} \) is the Krylov subspace of dimension \( k \), and \( \mathcal{L}_k \) is a well-defined subspace of dimension \( k \). The classical conjugate gradient is a Krylov projection method where \( \mathcal{L}_k = K_k(A, r_0) \).

In this section we briefly introduce the conjugate gradient versions related to our MSDO-CG and LRE-CG versions, starting with the 1952 Hestenes and Stiefel version (section 2.1). Since then, many different versions of CG have been introduced (refer to [2] for a historical overview of CG till 1976). In 1980 O’Leary introduced a Block CG version [22] that solves a system with multiple right-hand sides \( AX = B \) (section 2.2). The cooperative-CG [11] which was recently introduced, solves the system \( AX = b \) by starting with \( t \) distinct initial guesses. This is equivalent to solving the system \( AX = b + 1 \) (algorithmically very similar to Block CG ) where \( 1 \) is a vector of ones of size \( t \). The authors also present a parallel implementation that needs 2 to 3 synchronizations per iteration. This method has faster convergence than CG (section 2.3). The multiple search directions CG (MSD-CG) [14] solves \( Ax = b \) by decomposing \( A \)’s domain into \( t \) subdomains and defining a search direction on each of the \( t \) subdomains. Then \( x_k = x_{k-1} + P_k\alpha_k \), where \( P_k \) is a matrix containing all the \( t \) search directions and \( \alpha_k \) is a vector of size \( t \) (section 2.4). Unlike CG, block CG and coop CG, MSD-CG does not have the A-orthogonality condition of the search directions, i.e. \( P_k^tAP_k \) is not equal to zero for all \( i \) not equal to \( j \). Hence it is not a projection method. This causes MSD-CG to have slower convergence than CG, and in some cases not to converge at all. That is why in multiple search directions with orthogonalization CG (MSDO-CG), after defining a search direction on each of the \( t \) subdomains, we A-orthonormalize the search directions and this leads to better convergence than CG (section 5.2).
2.1 Conjugate Gradient (CG) Method

Conjugate Gradient [16] is an iterative Krylov projection method for symmetric (Hermitian) positive definite (SPD) matrices of the form

\[
\begin{align*}
Ax &= b, \\
A &= A', \\
x^T Ax &> 0, \forall x \\
\end{align*}
\]

Given an initial guess or iterate \(x_0\), at the \(k\)th iteration CG finds the new approximate solution \(x_k = x_{k-1} + \alpha_k r_k\) that minimizes \(\phi(x) = \frac{1}{2}(x)^T Ax - b'x\) over the corresponding space \(x_0 + K_k(A, r_0)\), where \(k > 0\), \(p_k \in K_k(A, r_0)\) is the \(k\)th search direction, and \(\alpha_k\) is the step along the search direction.

The minimum of \(\phi(x)\) is given by \(\nabla \phi(x) = Ax - b = 0\). Thus, by minimizing \(\phi(x)\) we are solving the system \(Ax = b\). As the name of the method indicates, the gradients \(\nabla \phi(x_i)\) for all \(i\) should be conjugate. And since CG is a Krylov projection method, the residual \(r_k = b - Ax_k\) should respect the Petrov-Galerkin condition

\[r_k \perp L_k,\]

where \(r_k\) is orthogonal to some well-defined subspace \(L_k \subseteq \mathbb{R}^n\) (or \(\subseteq \mathbb{C}^n\)) of dimension \(k\). In CG, the subspace \(L_k\) is the same as the Krylov subspace \(K_k\). Hence, the residuals form an orthogonal set, \((r_k)^T r_i = 0\), for all \(i < k\).

Moreover, the Petrov-Galerkin condition \(r_k \perp K_k(A, r_0)\) is equivalent to the conjugacy of the gradients \(\nabla \phi(x_k)^T \nabla \phi(x_i) = 0\), for all \(i \neq k\). Once \(x_k\) has been chosen, either \(x_k\) is the required approximate solution of \(Ax = b\) or a new search direction \(p_{k+1} 

This procedure is repeated until convergence or until the maximum number of allowed iterations has been reached without convergence. The convergence criterion is set as

\[||r_k||_2 \leq \epsilon ||b||_2, \text{ for some } \epsilon \in \mathbb{R},\]

where \(r_k = b - Ax_k \in K_{k+1}(A, r_0)\) is the \(k\)th residual.

**Theorem 2.1.** The Petrov-Galerkin condition \((r_k)^T r_i = 0\), for all \(y \in K_k\) implies the A-orthogonality of the search directions \(p_i^T A p_j = 0\), for all \(i \neq j \) and \(i, j \leq k\).

**Proof.** By definition, \(p_i \in K_i\) and \(K_i \subseteq K_{i+1}\). Thus \(p_i \in K_{i+c}\) for \(c \geq 0\). By the Petrov-Galerkin condition \(r_{k-1} p_i = 0\) and \(r_i p_k = 0\) for \(i < k - 1\). Thus, \(r_i^T p_k = r_i^T r_{k-1} p_i - \alpha_k p_k^T A p_i = 0\) for \(i \leq k - 1\). This implies that \(p_k^T A p_i = 0\) for \(i < k - 1\) since \(\alpha \neq 0\). Therefore, the A-orthogonality of the search directions.

This theorem means that the A-orthogonality of the search directions has to be ensured or else the Petrov-Galerkin condition won’t be respected. On the other hand, the search direction \(p_k \in K_k\) is chosen according to the following recursion relation:

\[
\begin{align*}
p_0 &= r_0, \\
p_k &= r_{k-1} + \beta_k p_{k-1}
\end{align*}
\]

where \(p_0\) is set equal to \(r_0\) since the initial residual is equal to negative the gradient \(- \nabla \phi(x_0)\) which is the steepest descent from \(x_0\). But \(p_0\) is not set to \(r_{k-1}\), the steepest descent from \(x_{k-1}\) for \(k > 1\), since the residuals are not A-orthogonal. It can be shown that the search directions defined in (2) are A-orthogonal i.e. \(p_i^T A p_i = 0\) for all \(i \leq k - 1\). For \(i < k - 1\), we have

\[p_i^T A p_i = r_{i-1}^T A p_i + \beta_k p_{k-1}^T A p_i = \beta_k p_{k-1}^T A p_i (3)\]

since \(r_{i-1}^T A p_i = 0\) by Petrov-Galerkin condition. In addition, \(r_i^T p_k = r_{i-1}^T r_{k-1} p_i - \alpha_k p_k^T A p_i = 0\) with \(r_{i-1}^T r_{k-1} p_i = 0\) since \(i < k - 2\). Thus, \(p_k^T A p_i = 0\) for \(i < k - 1\).

As for \(i = k - 1\), \(r_i^T r_{k-1} p_{k-1} = 0\) and \(p_k^T A p_{k-1} = 0\). Thus, \(\beta_k = \frac{(r_{k-1})^T A p_{k-1}}{p_{k-1}^T A p_{k-1}}\) is chosen so that \(p_k^T A p_{k-1} = 0\)

At each iteration, the step \(\alpha_k = (r_{k-1})^T r_{k-1}/(p_{k-1})^T A p_{k-1}\) is chosen such that,

\[\phi(x_k) = \min \{\phi(x_{k-1} + \alpha p_k), \forall \alpha \in \mathbb{R}\}.\]

Using the definition of \(\alpha_k, \beta_k = \frac{(r_{k-1})^T A p_{k-1}}{(p_{k-1})^T A p_{k-1}}\).
2.2 Block Conjugate Gradient (B-CG) Method

In 1980 O’Leary introduced a Block CG version [22] that solves an SPD system with multiple right-hand sides

\[
\begin{align*}
AX &= B, \\
A &= A^t, \\
x^tAx &> 0, \forall x \neq 0
\end{align*}
\]

where \(A\) is an \(n \times n\) matrix, \(X \in \mathbb{R}^{n \times t}\) is a block vector, and \(B\) is a block vector of size \(n \times t\) containing the multiple right-hand sides.

Starting with an initial guess \(X_0 \in \mathbb{R}^{n \times t}\), initial residual \(R_0 = B - AX_0\), \(P_1 = R_0\gamma_1\) with \(\gamma_1\) a \(t \times t\) full rank freely chosen matrix, the B-CG searches for an approximate solution \(X_{k+1} \in X_0 + \mathcal{K}_{k+1}(A, R_0)\) where

\[
\mathcal{K}_{k+1}(A, R_0) = \text{block span}\{R_0, AR_0, A^2R_0, \ldots, A^kR_0\}
\]

is the block Krylov subspace. Every \(n \times t\) block \(Z \in \mathcal{K}_{k+1}(A, R_0)\) is defined as \(Z = \sum_{i=1}^{k} A^iR_0\zeta_i\) where \(\zeta_i\) is a \(t \times t\) matrix. By the Petrov-Galerkin condition we have that \(R_{k+1}^tY = 0\) for all \(Y \in \mathcal{K}_{k+1}(A, R_0)\), which implies that \(R_{k+1}^tR_i = 0\) and \(R_{k+1}^tAP_i = 0\) for all \(i < k + 1\).

Then, for \(k \geq 0\) the iterates are defined similarly to CG:

\[
\begin{align*}
X_{k+1} &= X_k + P_{k+1}\alpha_{k+1} \quad \in \mathcal{K}_{k+1}(A, R_0) \\
R_{k+1} &= R_k - AP_{k+1}\alpha_{k+1} \quad \in \mathcal{K}_{k+2}(A, R_0) \\
P_{k+2} &= (R_{k+1} + P_{k+1}\beta_{k+2})\gamma_{k+2} \quad \in \mathcal{K}_{k+2}(A, R_0)
\end{align*}
\]

where

\[
\begin{align*}
\alpha_{k+1} &= (P_{k+1}^tAP_{k+1})^{-1}\gamma_k^t(R_k^tR_k) \\
\beta_{k+2} &= \gamma_{k+1}^{-1}(R_k^tR_k)^{-1}(R_{k+1}^tR_{k+1})
\end{align*}
\]

Note that \(\alpha_{k+1}\) is chosen such that \(\phi(X_{k+1}) = \min_{\alpha \in \mathbb{R}^{t \times t}} \phi(X_k + P_{k+1}\alpha)\), for all \(\alpha \in \mathbb{R}^{t \times t}\). As for \(\beta_{k+1}\), it is chosen to ensure the A-orthogonality of the \(P_{k+1}\) and \(P_k\) \((P_{k+1}^tAP_k = 0)\). Whereas \(\gamma_k\) is a \(t \times t\) full rank matrix that can be chosen freely to decrease roundoff errors in the implementation. Moreover, the search direction \(P_{k+1} \in \mathcal{K}_{k+1}(A, R_0)\) of the block conjugate gradient method is A-conjugate, \((P_{k+1})^tAY = 0\), for all \(Y \in \mathcal{K}_{k}(A, R_0)\). This leads to the A-orthogonality of the search direction \(\implies (P_{k+1})^tAP_i = 0\) for all \(i < k + 1\). We present the Block-CG algorithm in Appendix A (Algorithm ??).

2.3 Cooperative Conjugate Gradient (coop-CG) Method

Recently, in 2012, Bhaya et al. presented a new version of conjugate gradient which is similar in structure to the Block conjugate gradient method. The coop-CG [1] solves the system \(AX = b\) by starting with \(t\) different initial guesses and solving the same system \(t\) times in parallel, where \(t\) threads/agents cooperate to find the solution. This is equivalent to solving the system \(AX = b \star \text{ones}(1, t)\) where \(X_0\) is a block-vector containing the \(t\) initial guesses, \(R_0 = AX_0 - b \star \Pi_1\) is the block residual, \(P_1 = R_0\) is the initial block search direction. Then the derivations and the algorithm of the coop-CG (Algorithm ??, Appendix A) are the same as the Block-CG with \(\gamma_k = I\).

2.4 Multiple search direction Conjugate Gradient (MSD-CG) Method

The multiple search directions CG (MSD-CG), introduced by Gu et al. [14], solves the system \(AX=b\), and starts by having a decomposed domain and by defining at each iteration \(k\) a search direction \(p_k^j\) on each of the \(t\) subdomains \((\delta_i, i = 1, 2, ..., t)\) such that \(p_k^j(\delta_j) = 0\) for all \(j \neq i\). Then, the approximate solution at the \(k^{th}\) iteration is defined as \(x_k = x_{k-1} + P_k\alpha_k\) where \(P_k = [p_1^k, p_2^k, p_3^k ... p_t^k]\) is a matrix containing all the \(k^{th}\) search directions and \(\alpha_k\) is a vector of size \(t\).
Given an initial guess $x_0$, the residual is defined as $r_k = b - Ax_k$ for $k \geq 0$. The first set of domain search directions is defined by the initial residual $r_0$, such that $p_1^k(\delta_1) = r_0(\delta_1)$ for $i = 1, 2, ..., t$ and zero otherwise. Then, for $k > 1$ the domain search directions are defined as follows, $p_k^k = T_k(r_{k-1}) + \beta_k^k p_{k-1}^{k-1}$ for $i = 1, 2, ..., t$ where $\beta_k^k$ is a scalar and $T_k$ is an operator that projects a vector onto the subdomain $\delta_i ([T_i(x)](\delta_j) = 0$ for $j \neq i$ and $[T_i(x)](\delta_i) = x(\delta_i)$). The search directions block has the following sparsity pattern for all $k$,

$$P_k = \begin{pmatrix}
\ast & \ast & \ast & \ast \\
\ast & \ast & \ast & \ast \\
\vdots & \vdots & \vdots & \vdots \\
\ast & \ast & \ast & \ast \\
\end{pmatrix}_{n \times t}$$

As for $\alpha_k = (P_k^t A P_k)^{-1} P_k^t r_{k-1}$, it is chosen such that it minimizes $\phi(x_k) = \min \{ \phi(x_{k-1} + P_k \alpha), \forall \alpha \in \mathbb{R}^t \}$. Unlike CG, block CG and coop CG, MSD-CG does not have the A-orthogonality condition of the search directions, i.e. $P_k^t A P_k$ is not equal to zero for all $i$ not equal to $k$. Thus $\beta_k = (P_{k-1}^t A P_{k-1})^{-1} P_{k-1}^t r_{k-1}$ is chosen so that the global search direction $p_k = \sum_{i=1}^t p_i^k$ is A-orthogonal to the previous domain search direction $p_i^{k-1}$, i.e. $(p_i^k)^t A p_{k-1} = 0$, for $i = 1, 2, .., t$. As for the convergence, it is shown that the rate of convergence of MSD-CG is at least as fast as that of the steepest descent method. Yet, steepest descent is known for its slow “zig-zagging” convergence. This causes the MSD-CG to have slower convergence than CG, and in some cases it does not converge at all with respect to the given stopping criteria as shown in section 3.

3 The New Conjugate Gradients

We will introduce two new conjugate gradient methods, MSDO-CG and LRE-CG, which are based on replacing the Krylov subspace $K_k$ with a larger subspace leading to better convergence. Thus we will first introduce the new enlarged Krylov subspace and its properties in the context of conjugate gradient methods in section 3.1. Then in section 3.2 and section 3.3 we introduce the multiple search direction conjugate gradient with orthogonalization (MSDO-CG) and the long recurrence enlarged conjugate gradient (LRE-CG).

As previously mentioned, MSDO-CG is an adapted version of MSD-CG, where the $t$ newly defined search directions are A-orthonormalized against previous search directions and against each others. This A-orthonormalization guarantees a convergence behavior at least as good as CG. As for the LRE-CG, at each iteration, $t$ new basis vectors are computed for the enlarged Krylov subspace. Then, rather than having short recurrences, $x_k$ is defined by all the basis vectors as in GMRES, where the basis vectors are orthonormalized. Both methods, MSD-CG and LRE-CG, require saving at most $tk$ vectors versus one search direction in CG. Yet LRE-CG converges faster than MSDO-CG (section 4) at the expense of solving growing systems of size $tk$. Several remedies to this problem are discussed in section 5.

3.1 The Enlarged Krylov Subspace

The enlarged Krylov subspace and methods are based on a partition of the unknowns, or alternatively the rows of the $n \times n$ matrix $A$. Assume that the index domain $\delta = \{1, 2, .., n\}$ is divided into $t$ distinct
subdomains $\delta_i$, where $\delta = \cup_{i=1}^{t} \delta_i$.

We define $T_i(x)$ to be the operator that projects the vector $x$ onto the subdomain $\delta_i$. Let $y = T_i(x)$, then $y(\delta_i) = x(\delta_i)$ and zero elsewhere. Then, we define $T(x)$ to be an operator that transforms the $n \times 1$ vector $x$ into $t$ vectors of size $n \times 1$ that correspond to the projection of $x$ onto the subdomains $\delta_i$ for $i = 1, 2, \ldots, t$. If the obtained $t$ vectors are assembled in increasing order into a block vector $X$, then we have $X(\delta_i, i) = x(\delta_i)$ for all $i$ and zero elsewhere. We will refer to $R_0$ as the block containing the $t$ vectors obtained from $T(r_0)$. Note that $R_0 \neq T(r_0)$ since $R_0$ is a matrix, whereas $T(r_0) = \{T_1(r_0), T_2(r_0), \ldots, T_t(r_0)\}$ is a set of vectors. But $R_0 = [T_1(r_0) \ T_2(r_0) \ldots \ T_t(r_0)]$, where the brackets $[\ldots]$ denote a matrix format.

**Definition 3.1.** Let

$$\mathcal{K}_{t,k} = \text{span}\{T(r_0), AT(r_0), A^2T(r_0), \ldots, A^{k-1}T(r_0)\}$$

be an enlarged Krylov subspace of dimension $k \leq z \leq tk$ generated by the matrix $A$ and the vector $r_0$, and associated to a given partition defined by $\delta_i$ for $i = 1, 2, \ldots, t$.

The enlarged Krylov subspaces $\mathcal{K}_{t,k}(A, r_0)$ are increasing subspaces, yet bounded. We denote by $k_{max}$ the upper bound $k$ for which the dimension of the enlarged Krylov subspace $\mathcal{K}_{t,k}(A, r_0)$ stops increasing. For simplicity, we will denote the enlarged Krylov subspace generated by $A$ and $r_0, \mathcal{K}_{t,k}(A, r_0)$, by $\mathcal{K}_{t,k}$, and the Krylov subspace generated by $A$ and $r_0, \mathcal{K}(A, r_0)$ by $\mathcal{K}$.

**Theorem 3.2.** The Krylov subspace $\mathcal{K}$ is a subset of the enlarged Krylov subspace $\mathcal{K}_{t,k}$ ($\mathcal{K} \subset \mathcal{K}_{t,k}$).

**Proof.** Let $y \in \mathcal{K}_k$ where $\mathcal{K}_k = \text{span}\{r_0, Ar_0, \ldots, A^{k-1}r_0\}$. Then

$$y = \sum_{j=0}^{k-1} a_j A^j r_0 = \sum_{j=0}^{k-1} a_j A^j R_0 \ast 1_t = \sum_{j=0}^{t-1} a_j A^j T_i(r_0) \in \mathcal{K}_{t,k}$$

since $r_0 = R_0 \ast 1_t = [T_1(r_0) \ T_2(r_0) \ldots \ T_t(r_0)] \ast 1_t$. $\square$

Krylov subspace methods search for an approximate solution $x_k \in x_0 + \mathcal{K}_k$. A corollary of theorem 3.2 is that we can search for an approximate solution $x_k$ in $x_0 + \mathcal{K}_{t,k}$ instead, since $\mathcal{K}_k \subset \mathcal{K}_{t,k}$.

In theorem 3.3 we do not use the direct sum $\oplus$ since it is not guaranteed that the intersection of the two subspaces, $\mathcal{K}_{t,k}$ and $\text{span}\{A^k T_1(r_0), A^k T_2(r_0), \ldots, A^k T_t(r_0)\}$, is empty.

**Theorem 3.3.** By definition 3.1 of the enlarged Krylov subspace, $\mathcal{K}_{t,k+1} = \mathcal{K}_{t,k} + \text{span}\{A^k T_1(r_0), A^k T_2(r_0), \ldots, A^k T_t(r_0)\}$

If $A^k T_v(r_0) \in \mathcal{K}_{t,k}$ for all $1 \leq v \leq t$, then $A^{k+q} T_i(r_0) \in \mathcal{K}_{t,k}$ for some $1 \leq i \leq t$ and for some $q > 0$.

**Proof.** We prove this by induction.

**Base Case:**

Given that $A^k T_v(r_0) \in \mathcal{K}_{t,k}$ for all $1 \leq v \leq t$, we show that $A^{k+1} T_i(r_0) \in \mathcal{K}_{t,k}$, where $1 \leq i \leq t$.

$$A^{k+1} T_i(r_0) = \sum_{u=0}^{k-1} \sum_{v=1}^{t} \alpha_{u,v} A^u T_v(r_0)$$

Since $A^k T_v(r_0) \in \mathcal{K}_{t,k}$. Then

$$A^{k+1} T_i(r_0) = \sum_{u=0}^{k-1} \sum_{v=1}^{t} \alpha_{u,v} A^u T_v(r_0) + \sum_{u=0}^{k-1} \sum_{v=1}^{t} \alpha_{u,v} A^u T_v(r_0)$$

$$= \sum_{u=0}^{k-1} \sum_{v=1}^{t} \alpha_{u,v} A^u T_v(r_0) + \sum_{v=1}^{t} \alpha_{k-1,v} A^k T_v(r_0)$$

$$= \sum_{v=1}^{t} \sum_{u=0}^{k-1} \alpha_{u,v} A^u T_v(r_0) \in \mathcal{K}_{t,k}$$

Inria
Assume true for $q$
Assume that $A^{k+q}T_i(r_0) \in K_{t,k}$ where $1 \leq i \leq t$, that is $A^{k+q}T_i(r_0) = \sum_{u=0}^{k-1} \sum_{v=1}^\infty \alpha_{u,v} A^u T_v(r_0)$
Prove true for $q+1$
Show that $A^{k+q+1}T_i(r_0) \in K_{t,k}$

$$A^{k+q+1}T_i(r_0) = \sum_{u=0}^{k-1} \sum_{v=1}^\infty \alpha_{u,v} A^{u+1}T_v(r_0) = \sum_{u=0}^{k-2} \sum_{v=1}^\infty \alpha_{u,v} A^{u+1}T_v(r_0) + \sum_{v=1}^t \alpha_{k-1,v} A^{k}T_v(r_0)$$

$$= \sum_{u=0}^{k-2} \sum_{v=1}^\infty \alpha_{u,v} A^{u+1}T_v(r_0) + \sum_{v=1}^t \alpha_{k-1,v} \sum_{u=0}^\infty \sum_{y=1}^\infty \beta_{u,y} A^u T_y(r_0))$$

$$= \sum_{u=0}^{k-1} \sum_{v=1}^\infty \gamma_{u,v} A^u T_v(r_0) \in K_{t,k}$$

Given that $K_{t,k} \neq K_{t,k-1}$, then a corollary of Theorem 3.3 is that $K_{t,k} = K_{t,k+q}$ for all $q > 0$, where $k_{\text{max}} = k$ is the upper bound for which the dimension of the enlarged Krylov subspace stops increasing.
Assume that $A^kT_i(r_0) \in K_{t,k}$ for all $1 \leq i \leq t$, then by Theorem 3.3 $A^{k+q}T_i(r_0) \in K_{t,k}$ for all $q > 0$ and for some $1 \leq i \leq t$. Then for all $1 \leq i \leq t$ and for all $q > 0$, $A^{k+q}T_i(r_0) \in K_{t,k}$. Thus no new vector is added to the basis of $K_{t,k+q}$ for all $q > 0$ and $K_{t,k} = K_{t,k+q}$. Moreover, since $K_{t,k} \neq K_{t,k-1}$ then $k_{\text{max}} = k$, the upper bound for which the dimension of enlarged Krylov subspace $K_{t,k}(A, r_0)$ stops increasing.

**Theorem 3.4.** If $A^kT_i(r_0) \in K_{t,k} + \text{span}\{A^kT_1(r_0), \ldots, A^{k}T_i-1(r_0), A^{k+1}T_i(r_0), \ldots, A^{k+q}T_i(r_0)\}$, then $A^{k+q}T_i(r_0) \in K_{t,k+q} + \text{span}\{A^{k+q}T_1(r_0), \ldots, A^{k+q}T_i(r_0), \ldots, A^{k+q}T_i(r_0)\}$ for all $1 \leq i \leq t$.

**Proof.** If $A^kT_i(r_0) \in K_{t,k} + \text{span}\{A^kT_1(r_0), \ldots, A^{k}T_i-1(r_0), A^{k+1}T_i(r_0), \ldots, A^{k+q}T_i(r_0)\}$, then $A^kT_i(r_0) = \sum_{u=0}^\infty \sum_{v=1}^\infty \alpha_{u,v} A^u T_v(r_0) + \sum_{v=1}^t \alpha_{k,v} A^k T_v(r_0)$. Thus,

$$A^{k+q}T_i(r_0) = \sum_{u=0}^k \sum_{v=1}^\infty \alpha_{u,v} A^{u+q}T_v(r_0) + \sum_{v=1}^t \alpha_{j,v} A^{k+q}T_v(r_0)$$

$$\in K_{t,k+q} + \text{span}\{A^{k+q}T_1(r_0), \ldots, A^{k+q}T_i(r_0), \ldots, A^{k+q}T_i(r_0)\}$$

A corollary of Theorem 3.4 is that if $t-i_k$ vectors of the form $A^kT_y(r_0)$ with $y = i_k + 1, \ldots, t$ belong to the subspace $K_{t,k} + \text{span}\{A^kT_1(r_0), A^{k}T_2(r_0), \ldots, A^{k}T_i(r_0)\}$, then the $t-i_k$ vectors of the form $A^{k+q}T_y(r_0)$ belong to the subspace $K_{t,k+q} + \text{span}\{A^{k+q}T_1(r_0), A^{k+q}T_2(r_0), \ldots, A^{k+q}T_i(r_0)\}$.

**Theorem 3.5.** Let $k_{\text{max}}$ be the smallest integer such that $K_{t,k_{\text{max}}} = K_{t,k_{\text{max}}+q}$ for all $q > 0$. Then, for all $k < k_{\text{max}}$ the dimension of the enlarged Krylov subspaces $K_{t,k}$ and $K_{t,k+1}$ is strictly increasing by some number $i_k$ and $i_{k+1}$ respectively, where $1 \leq i_k+1 \leq i_k \leq t$.

**Proof.** By definition of $k_{\text{max}}$, we have that for all $q > 0$

$$K_{t,1} \subseteq \ldots \subseteq K_{t,k_{\text{max}}-1} \subseteq K_{t,k_{\text{max}}} = K_{t,k_{\text{max}}+q}$$

Then for all $k < k_{\text{max}}$, the dimension of the enlarged Krylov subspaces $K_{t,k}$ is strictly increasing by some number $i_k \neq 0$ with respect to the dimension of $K_{t,k-1}$.

If the $t$ new vectors are linearly independent and none of them belongs to $K_{t,k-1}$, then the $t$ vectors are added to the basis of $K_{t,k}$ and $\text{dim}(K_{t,k}) = \text{dim}(K_{t,k-1}) + t$, where $i_k = t$ and $\text{dim}(\cdot)$ is the dimension of a subspace. In case the $t$ new vectors are linearly dependent and none of them belongs to $K_{t,k-1}$, then only one vector is added to the basis of $K_{t,k}$ and $\text{dim}(K_{t,k}) = \text{dim}(K_{t,k-1}) + 1$, that
is $i_k = 1$. There are many other cases where $1 < t - i_k < t$ of the $t$ vectors belong to $\mathcal{K}_{t,k-1}$ or are linearly dependant on the other $i_k$ vectors and $\mathcal{K}_{t,k-1}$. Then $i_k$ vectors are added to the basis of $\mathcal{K}_{t,k}$ and $\dim(\mathcal{K}_{t,k}) = (\mathcal{K}_{t,k-1} + i_k$, where $1 < i_k < t$. In general, $\dim(\mathcal{K}_{t,k}) = \dim(\mathcal{K}_{t,k-1}) + i_k$, where $1 \leq i_k \leq t$. Similarly, $\dim(\mathcal{K}_{t,k+1}) = \dim(\mathcal{K}_{t,k}) + i_{k+1}$, where $1 \leq i_{k+1} \leq t$. Moreover, in $\mathcal{K}_{t,k}$’s basis we added $i_k$ new vectors of the form $A^{k-1}T_i(r_0)$ while the other $t - i_k$ either belong to $\mathcal{K}_{t,k-1}$ or are linearly dependant on the $i_k$ vectors and $\mathcal{K}_{t,k-1}$. In both cases, the $t - i_k$ vectors of the form $A^{k-1}T_i(r_0)$ belong to the subspace $\mathcal{K}_{t,k-1} + \text{span}\{A^{k-1}T_i(r_0), \ldots, A^{k-1}T_i(r_0)\}$. Then by Theorem 3.4 and its corollary, the $t - i_k$ vectors of the form $A^{k+q}T_i(r_0)$ belong to the subspace $\mathcal{K}_{t,k+q} + \text{span}\{A^{k+q}T_i(r_0), A^{k+q}T_i(r_0), \ldots, A^{k+q}T_i(r_0)\}$ for $q > 0$. Therefore, we have at least $t - i_k$ linearly dependent vectors added to $\mathcal{K}_{t,k+1}$, hence $i_{k+1}$ can never be greater than $i_k$.

**Theorem 3.6.** Let $p_{\text{max}}$ and $k_{\text{max}}$ be such that $\mathcal{K}_{p_{\text{max}}} = \mathcal{K}_{p_{\text{max}}} + q$ and $\mathcal{K}_{t,k_{\text{max}}} = \mathcal{K}_{t,k_{\text{max}}} + q$ for $q > 0$. Then $k_{\text{max}} \leq p_{\text{max}}$.

**Proof.** Let $\mathcal{K}_{p_{\text{max}}} = \mathcal{K}_{p_{\text{max}}} + q$ and $A^{p_{\text{max}}+q-1}r_0 \in \mathcal{K}_{p_{\text{max}}} + q$ where $q > 0$. Then $A^{p_{\text{max}}+q-1}r_0 \in \mathcal{K}_{p_{\text{max}}} \subset \mathcal{K}_{t,p_{\text{max}}}$, and $A^{p_{\text{max}}+q-1}r_0 = \sum_{j=1}^{p_{\text{max}}} \sum_{i=1}^{t} \alpha_{j,i} A^{j-1}T_i(r_0)$. Thus $A^{p_{\text{max}}+q-1} = \sum_{j=1}^{p_{\text{max}}} \sum_{i=1}^{t} \alpha_{j,i} A^{j-1}T_i(r_0) = \sum_{j=1}^{p_{\text{max}}} \sum_{i=1}^{t} \alpha_{j,i} A^{j-1}T_i(r_0)$.

Suppose that $A^{p_{\text{max}}+q-1}T_i(r_0) \notin \mathcal{K}_{p_{\text{max}}}$ for all $1 \leq i \leq t$. Then $A^{p_{\text{max}}+q-1}T_i(r_0) = \sum_{j=1}^{p_{\text{max}}} \sum_{i=1}^{t} \alpha_{j,i} A^{j-1}T_i(r_0)$. We may assume that there exists at least one $\alpha_{j,i} \neq 0$ for $j > p_{\text{max}}$, then this leads to a contradiction. This implies that $A^{p_{\text{max}}+q-1}T_i(r_0) \in \mathcal{K}_{t,p_{\text{max}}}$ for all $1 \leq i \leq t$.

Thus by definition of the $T()$ operator and since $\mathcal{K}_p$ is a subset of $\mathcal{K}_{t,p}$, if $\mathcal{K}_{p_{\text{max}}} = \mathcal{K}_{p_{\text{max}}} + q$, then $\mathcal{K}_{p_{\text{max}}} = \mathcal{K}_{t,p_{\text{max}}} + q$. However, if $\mathcal{K}_{t,k_{\text{max}}} = \mathcal{K}_{t,k_{\text{max}}} + q$ this does not imply that $\mathcal{K}_{k_{\text{max}}} = \mathcal{K}_{k_{\text{max}}} + q$. Therefore, since $\mathcal{K}_{t,k}$ is a much larger subspace than $\mathcal{K}_k$, it is possible to reach stagnation earlier. Therefore $k_{\text{max}} \leq p_{\text{max}}$.

**Theorem 3.7.** The solution of the system $Ax = b$ belongs to the subspace $x_0 \mathcal{K}_{t,k_{\text{max}}},$ where $\mathcal{K}_{t,k_{\text{max}}} + q = \mathcal{K}_{t,k_{\text{max}}} + q$ for $q > 0$.

**Proof.** The solution $x_{\text{sol}} \in x_0 + \mathcal{K}_{p_{\text{max}}}$, where $\mathcal{K}_{p_{\text{max}}} = \text{span}\{r_0, A^r_0, \ldots, A^{p_{\text{max}}-1}r_0\}$ and $\mathcal{K}_{p_{\text{max}}} = \mathcal{K}_{p_{\text{max}}} + q$ for $q > 0$. Since $\mathcal{K}_{p_{\text{max}}} \subset \mathcal{K}_{t,p_{\text{max}}}$, the solution $x_{\text{sol}} \in x_0 + \mathcal{K}_{t,p_{\text{max}}}$, where $p_{\text{max}} \geq k_{\text{max}}$ by Theorem 3.6.

Suppose that $x_{\text{sol}} \in x_0 + \mathcal{K}_{t,p_{\text{max}}}$, but $x_{\text{sol}} \notin x_0 + \mathcal{K}_{t,k_{\text{max}}}$. This implies that $\mathcal{K}_{t,k_{\text{max}}} \neq \mathcal{K}_{t,p_{\text{max}}}$. However, by definition of $k_{\text{max}}$ and since $k_{\text{max}} \leq p_{\text{max}}$, we have that $\mathcal{K}_{t,k_{\text{max}}} = \mathcal{K}_{t,p_{\text{max}}}$. This is a contradiction.

### 3.1.1 Krylov projection methods

The Krylov projection methods find a sequence of approximate solutions $x_k$ ($k > 0$) of the system $Ax = b$ from the subspace $x_0 + \mathcal{K} \subseteq \mathbb{C}^n$ (or $\subseteq \mathbb{C}^n$) by imposing the Petrov-Galerkin constraint on the $k$th residual $r_k = b - Ax_k$, that is $r_k$ is orthogonal to some well-defined subspace of dimension $k$.

We define our new enlarged Krylov projection methods based on CG by the subspace $\mathcal{K}_{t,k}$ and the following two conditions:

1. Subspace condition: $x_k \in x_0 + \mathcal{K}_{t,k}$

2. Orthogonality condition: $r_k \perp \mathcal{K}_{t,k}$

$$
\iff \langle r_k, y \rangle = 0, \quad \text{for all } y \in \mathcal{K}_{t,k}
$$

where $\mathcal{K}_{t,k}$ is a well-defined subspace of dimension $k \leq z \leq tk$.

Inria
3.1.2 The minimization property

The new enlarged CG methods find the new approximate solution by minimizing the function $\phi(x)$ over the subspace $x_0 + \mathcal{K}_{t,k}$.

**Theorem 3.8.** If $r_k \perp \mathcal{K}_{t,k}$, then $\phi(x_k) = \min \{ \phi(x), \forall x \in x_0 + \mathcal{K}_{t,k} \}$.

**Proof.** By the Petrov-Galerkin condition we have that $r_k \perp \mathcal{K}_{t,k}$

\[
\Rightarrow (r_k)^t y = 0, \forall y \in \mathcal{K}_{t,k}
\]

\[
(b - Ax_k)^t y = 0, \forall y \in \mathcal{K}_{t,k}
\]

\[
b^t y - (x_k)^t Ay = 0, \forall y \in \mathcal{K}_{t,k}
\]

Let $y = x_k - x_0 \in \mathcal{K}_{t,k}$

\[
\Rightarrow (x_k)^t A(x_k - x_0) - b^t (x_k - x_0) = 0
\]

\[
\Rightarrow (x_k)^t A x_k - b^t x_k = (x_k)^t A x_0 - b^t x_0
\]

\[
\Rightarrow \phi(x_k) = \frac{1}{2} (x_k)^t A x_k - b^t x_k = -\frac{1}{2} (x_k)^t A x_k + (x_k)^t A x_0 - b^t x_0
\]

By showing that $\phi(x) \geq \phi(x_k)$, for all $x \in x_0 + \mathcal{K}_{t,k}$ then we have proven that

\[
\phi(x_k) = \min \{ \phi(x), \forall x \in x_0 + \mathcal{K}_{k} \}. \tag{5}
\]

**Theorem 3.9.** $\phi(x_k) = \min \{ \phi(x), \forall x \in x_0 + \mathcal{K}_{t,k} \}$ if and only if $\|x^* - x_k\|_A = \min \{ \|x^* - x\|_A, \forall x \in x_0 + \mathcal{K}_{t,k} \}$, where $x^*$ is the exact solution of (1).

**Proof.** $f(x) = \|x^* - x\|_A = (x^*)^t A x^* - 2(x^*)^t A x + x^t A x = b^t x^* - 2b^t x + x^t A x = b^t x^* + 2\phi(x)$. The minimum of $f(x)$ is given by $f'(x) = \nabla \phi(x) = 0$.

3.1.3 Convergence analysis

The conjugate gradient method of Hestenes and Stiefel is known to converge in $K$ iterations where $K \leq n$, if the matrix $A \in \mathbb{R}^{n,n}$ is SPD. Moreover, the $k^{th}$ error of CG $\tau_k = \|x^* - \tau_k\| \leq 2\left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^k \|\tau_0\|_A$ where $\kappa = \|A\|_2 \|A^{-1}\|_2 = \frac{\lambda_{\max}}{\lambda_{\min}}$ is the L2-condition number of the matrix $A$.

Assuming that the $k^{th}$ residual of the new conjugate gradient methods $r_k \perp \mathcal{K}_{t,k}$, then by Theorem 3.8 and Theorem 3.9 we have that

\[
\|e_k\|_A = \|x^* - x_k\|_A = \min \{ \|x^* - x\|_A, \forall x \in x_0 + \mathcal{K}_{t,k} \} \tag{6}
\]

\[
\leq \min \{ \|x^* - \tau\|_A, \forall \tau \in x_0 + \mathcal{K}_{t,k} \} \text{ since } \mathcal{K}_k \subset \mathcal{K}_{t,k} \tag{7}
\]

\[
\leq \|\tau_k\|_A \tag{8}
\]

Therefore, our methods converge at least as fast as the Classical Conjugate Gradient method, assuming that the Petrov Galerikin condition is respected ($r_k \perp \mathcal{K}_{t,k}$). Hence, the enlarged Krylov subspace CG methods will converge in $K$ iterations, where $K \leq K \leq n$. 

RR n° 8597
3.2 Multiple search direction with orthogonalization Conjugate Gradient (MSDO-CG) Method

The MSD-CG method introduced in Gu et al [14], can be viewed as an enlarged Krylov method where $P_0 = [T(r_0)]$, and at the $k^{th}$ iteration $p_i^k = T_i(r_{k-1}) + \beta_i^{-1} p_i^{k-1}$ for $i = 1, 2, ..., t$, $P_k = [p_1^k, p_2^k, ..., p_t^k]$, $x_k = x_{k-1} + P_k \alpha_k$ and $r_k = r_{k-1} - A P_k \alpha_k$ with $\alpha_k = (P_k^T A P_k)^{-1} P_k^T r_{k-1}$ and $\beta_k = [\beta_1^k, \beta_2^k, ..., \beta_t^k] = (P_{k-1}^T A P_{k-1})^{-1} P_{k-1}^T A r_{k-1}$. However, the $P_k$’s are not A-orthogonal implying that $r_k \notin \mathcal{K}_{t,k}$. Thus, MSD-CG is not a projection method.

The multiple search directions with orthogonalization CG (MSDO-CG) is an enlarged Krylov projection method that solves the system (1) (Ax=b), by approximating the solution at the $k^{th}$ iteration with the vector $x_k = x_{k-1} + P_k \alpha_k$ such that

$$\phi(x_k) = \min \{ \phi(x), \forall x \in \mathcal{K}_{t,k} \}$$

where $P_k \alpha_k \in \mathcal{K}_{t,k}$, $P_k$ is an $n \times t$ block vector containing the $t$ subdomain search directions, and $\alpha_k$ is a vector of size $t$.

The minimum of $\phi(x)$ is given by $\nabla \phi(x) = 0$, which is equivalent to $Ax = b = 0$. Thus, by minimizing $\phi(x)$, we are solving the system (1). Note that since $\phi(x_k) = \min \{ \phi(x), \forall x \in x_0 + \mathcal{K}_{t,k} \}$, then

$$\phi(x_k) = \phi(x_{k-1} + P_k \alpha_k) = \min \{ \phi(x_{k-1} + P_k \alpha), \forall \alpha \in \mathbb{R}^t \}.$$ (9)

Once $x_k$ has been chosen, either $x_k = \text{the desired solution of } Ax = b$, or $t$ new domain search direction vectors $P_{k+1}$ and a new approximation $x_{k+1} = x_k + P_{k+1} \alpha_{k+1}$ are computed. Similarly to MSD-CG, $P_{k+1} = [p_1^{k+1}, p_2^{k+1}, ..., p_t^{k+1}]$, where $p_1^k = T_i(r_0)$ and $p_t^k = T_i(r_{k-1}) + \beta_i^{-1} p_i^{k-1}$ for $i = 1, 2, ..., t$. But unlike MSD-CG, MSDO-CG is a projection method. Hence, we A-orthonormalize all the search directions, $P_{k+1}$, to ensure that $r_{k+1} \perp \mathcal{K}_{t,k+1}$ as discussed in section 3.2.1.3 By imposing the orthogonality condition, $r_{k+1} \perp \mathcal{K}_{t,k+1}$, it is guaranteed that MSDO-CG converges at least as fast as CG as proven in section 3.1.1.3.

This procedure is repeated until convergence. Thus, we need to find the recursion relations of $r_k$, $P_k$, $\alpha_k$, and $\beta_k = [\beta_1^k, \beta_2^k, ..., \beta_t^k]^T$.

3.2.1 The residual $r_k$

By definition, the residual $r_k = b - Ax_k$, where $x_k \in \mathcal{K}_{t,k}$. Thus $r_k \in \mathcal{K}_{t,k+1}$. As for the recursion relation of $r_k$, we simply replace $x_k$ by its expression and obtain the following:

$$r_k = b - Ax_k$$
$$r_k = b - A(x_{k-1} + P_k \alpha_k)$$
$$r_k = r_{k-1} - A P_k \alpha_k$$

Moreover, if the orthogonality condition, $r_k \perp \mathcal{K}_{t,k}$, is ensured, then $(r_k)^T r_i = 0$, for all $i < k$. Hence, the residuals form an orthogonal set.

Theorem 3.10. The orthogonality condition $(r_k)^T y = 0$ for all $y \in \mathcal{K}_{t,k}$, implies the A-orthogonality of the block search directions $P_k^T A P_k = 0$, for all $i \neq j$, and $i, j \leq k$.

Proof. By definition, $P_i \in \mathcal{K}_{t,i}$ and $\mathcal{K}_{t,1} \subset \mathcal{K}_{t,i+1}$. Thus $P_i \subset \mathcal{K}_{t,i+c}$ for $c \geq 0$. By the Petrov-Galerkin condition $r_i^{k-1} P_i = 0$ for $i \leq k-1$ and $r_i^0 P_i = 0$. Thus, $r_i^T P_i = r_i^{k-1} P_i - \alpha_k P_i A P_i = 0$ for $i \leq k-1$. This implies that $P_k^T A P_i = 0$ for $i \leq k-1$ since $\alpha \neq 0$. Therefore, the A-orthogonality of the search directions. \qed

3.2.2 The domain search direction $P_k$

By definition, the domain search direction is $P_k = [p_1^k \ p_2^k \ ... \ p_t^k]$ where $p_i^k = T_i(r_0)$ and $p_i^k = T_i(r_{k-1}) + \beta_i^{-1} p_i^{k-1}$ for $i = 1, 2, ..., t$, $p_i^k \in \mathcal{K}_{t,k}$ for $i = 1, 2, ..., t$ and $P_k \alpha_k \in \mathcal{K}_{t,k}$.\^{Inria}
The recursion relation of $P_k$ is defined as follows
\[ P_k = T(r_{k-1}) + P_{k-1} \text{diag}(\beta_k), \]
where $\text{diag}(\beta_k)$ is a $t \times t$ matrix with the vector $\beta_k$ on the diagonal.

The domain search directions defined in (10) are not $A$-orthogonal to each others. To ensure that the orthogonality condition is valid, at each iteration $k$ the block vector $P_k$ is $A$-orthonormalized against all the previous $P_i$, where $i = 1, 2, \ldots, k - 1$. Then the column vectors of $P_k$ are $A$-orthonormalized against each others. Thus, the obtained search directions $\tilde{P}_k$ satisfy $(\tilde{P}_k)^\top \tilde{A} \tilde{P}_k = 0$ for all $i \neq k$. Moreover, $(\tilde{P}_k)^\top \tilde{A} \tilde{P}_k = I$, where $I$ is the identity matrix, assuming that the column vectors of $P_k$ are linearly independent with respect to each others and the previous directions, or alternatively none of the column vectors of $\tilde{P}_k$ is zero. Note that, once $P_k = T(r_{k-1}) + P_{k-1} \text{diag}(\beta_k)$ is defined, it is directly $A$-orthonormalized. Thus, in the sections that follow, we denote by $P_k$ the $A$-orthonormalized search directions and we do not use the $\tilde{P}_k$ notation to be consistent with the initial definitions in the previous sections.

There are several $A$-orthonormalization methods. First, for $A$-orthonormalizing $P_k$ against all the previous $P_i$, where $i = 1, 2, \ldots, k - 1$, one can use classical Gram Schmidt (CGS), modified Gram Schmidt (MGS), or classical Gram Schmidt with reorthogonalization (CGS2) where we apply the CGS algorithm twice for numerical stability reasons. As for $A$-orthonormalizing $P_k$, there are many methods that are discussed in [20] [24], but we will only refer to CGS, CGS2, MGS, A-CholQR and Pre-CholQR. We seek a combination of both $A$-orthonormalizations that is stable and parallelizable with reduced communication. For that reason, in section [4] we test the MSDO-CG method with the different combinations of the $A$-orthonormalization methods and we conclude that the MSDO-CG is numerically most stable when we use MGS, CGS2+A-CholQR, or CGS2+Pre-CholQR. In section [5] we discuss the parallelization of the MSDO-CG algorithm with the stable $A$-orthonormalization methods.

Note that in Appendix B, we discuss the $A$-orthonormalization using modified Gram Schmidt and classical Gram Schmidt. We also present versions of the algorithms that reduce communication along with their parallelizations. For example, Algorithm [17] is a block Gram Schmidt $A$-orthonormalization based on classical Gram Schmidt that $A$-orthonormalizes $P_k$ against previous vectors. And Algorithm [20] $A$-orthonormalizes $P_k$’s vectors against each others using a classical Gram Schmidt.

### 3.2.3 Finding the expression of $\alpha_{k+1}$ and $\beta_{k+1}$

At each iteration the step $\alpha_{k+1}$ is chosen such that
\[ \phi(x_{k+1}) = \min \{ \phi(x_k + P_{k+1} \alpha), \forall \alpha \in \mathbb{R}^t \} \]
Let $F(\alpha) = \phi(x_k + P_{k+1} \alpha)$ where $\phi(x) = \frac{1}{2} x^\top A x - x^\top b$.

Then,
\[ F(\alpha) = \frac{1}{2} (x_k + P_{k+1} \alpha)^\top A (x_k + P_{k+1} \alpha) - (x_k + P_{k+1} \alpha)^\top b \]
\[ = \phi(x_k) + \frac{1}{2} [(x_k)^\top A P_{k+1} \alpha + \alpha^\top (P_{k+1})^\top A x_k + \alpha^\top (P_{k+1})^\top A P_{k+1} \alpha] - \alpha^\top (P_{k+1})^\top b \]
\[ = \phi(x_k) + \frac{1}{2} [(x_k)^\top A P_{k+1} \alpha - \alpha^\top (P_{k+1})^\top A x_k] + \frac{1}{2} \alpha^\top (P_{k+1})^\top A P_{k+1} \alpha - \alpha^\top (P_{k+1})^\top r_k \]
\[ = \phi(x_k) + \frac{1}{2} \alpha^\top (P_{k+1})^\top A P_{k+1} \alpha - \alpha^\top (P_{k+1})^\top r_k, \text{ since } A \text{ is SPD} \]

The minimum of $F(\alpha)$ is given by $F'(\alpha) = 0$.
\[ \Rightarrow F'(\alpha) = (P_{k+1})^\top A P_{k+1} \alpha - (P_{k+1})^\top r_k = 0 \]

Therefore,
\[ \alpha_{k+1} = (P_{k+1}^\top A P_{k+1})^{-1} (P_{k+1}^\top r_k) \]

As for $\beta_{k+1}$, it should be chosen to ensure that $P_{k+1}$ is $A$-orthogonal to $P_k$. $P_{k+1} = T(r_k) + P_k \text{diag}(\beta_{k+1})$ and $P_{k+1}^\top A P_{k+1} = P_k^\top A T(r_k) + P_k^\top A P_k \text{diag}(\beta_{k+1})$. Since $P_k$ is an $A$-orthonormal matrix, $P_k^\top A P_k = I$, $\text{diag}(\beta_{k+1})$ should be equal to $-P_k^\top A T(r_k)$. But nothing guarantees that $P_k^\top A T(r_k)$ is
a diagonal matrix. So we choose \( \beta_{k+1} = (P_k^t A P_k)^{-1} P_k^t A r_k \) which guarantees that \( P_{k+1} \ast \mathbb{I} \) is A-orthogonal to \( P_k \), similarly to MSD-CG. Moreover, in case \( P_k^t A T (r_k) \) is a diagonal matrix, then our choice of \( \beta_{k+1} \) implies that \( P_{k+1} \) is A-orthogonal to \( P_k \). If \( t = 1 \), then MSDO-CG is reduced to the classical conjugate gradient.

Note that, since the vectors of \( P_{k+1} \) are A-orthonormalized \( (P_{k+1}^t A P_{k+1} = I) \), then \( \alpha_{k+1} \) and \( \beta_{k+1} \) systems are reduced to \( \alpha_{k+1} = P_{k+1}^t r_k \) and \( \beta_{k+1} = -P_k^t A r_k \).

### 3.2.4 The MSDO-CG Algorithm

After deriving the recurrence relations of \( x_k, r_k, P_k, \alpha_k, \) and \( \beta_k \), we present the MSDO-CG algorithm in Algorithm 1. We do not specify the A-orthonormalization methods, since this choice will be based first on the numerical stability of the method (section 4), then on its parallelization with the least communication possible (section 5.1).

Thus we present the MSDO-CG algorithm (Algorithm 1) and the computed flops per iteration except for the A-orthonormalizations. To reduce communication and computation in the A-orthonormalizations, be it MGS (Algorithms 14 and 15), CGS (Algorithms 18 and 21), A-CholQR (Algorithm 25), or Pre-CholQR (Algorithm 27), we replace \( W_{k+1} = A P_{k+1} \) by

\[
\begin{align*}
W_1 & = A P_1 \\
W_{k+1} & = A T (r_k) + A P_k \text{diag}(\beta) \quad \forall k \geq 1 \\
\end{align*}
\]

**Algorithm 1 MSDO-CG algorithm**

**Input:** \( A \), the \( n \times n \) symmetric positive definite matrix  
**Input:** \( b \), the \( n \times 1 \) right-hand side; \( x_0 \), the initial guess or iterate  
**Input:** \( \varepsilon \), the stopping tolerance; \( k_{\text{max}} \), the maximum allowed iterations  
**Output:** \( x_k \), the approximate solution of the system \( A x = b \)

1: \( r = b - A x_0\), \( \rho = ||r||^2_2, k = 1 \) \hspace{1cm} 2nnz + 2n - 1  
2: Let \( P_1 = T (r_0) \) and \( W_1 = A P_1 \) \hspace{1cm} 2nnz - (t - 1)n  
3: A-orthonormalize \( P_1 \) \hspace{1cm} not included here  
4: while \( (\sqrt{\rho} > \varepsilon ||b||_2 \text{ and } k < k_{\text{max}}) \) do \hspace{1cm} 2n  
5: \( \alpha = (P_k^t W_k)^{-1} (P_k^t r) = P_k^t r \) \hspace{1cm} \((2n - 1)t\)  
6: \( x = x + P_k \alpha \) \hspace{1cm} \((2t - 1)n + n\)  
7: \( r = r - W_k \alpha \) \hspace{1cm} \((2t - 1)n + n\)  
8: \( \rho = ||r||_2^2 \) \hspace{1cm} 2n - 1  
9: \( \beta = (P_k^t W_k)^{-1} (W_k^t r) = -W_k^t r \) \hspace{1cm} \((2n - 1)t\)  
10: \( P_{k+1} = T (r) + P_k \text{diag}(\beta) \) \hspace{1cm} 2nt  
11: \( W_{k+1} = A T (r_k) + W_k \text{diag}(\beta) \) \hspace{1cm} \((2n - (t - 1)n + 2nt\)  
12: A-orthonormalize \( P_{k+1} \) against all \( P_i \)'s for \( i \leq k \) \hspace{1cm} not included here  
13: A-orthonormalize \( P_{k+1} \) \hspace{1cm} not included here  
14: \( k = k + 1 \) \hspace{1cm} 1  
15: end while

The total number of flops computed sequentially after \( k_c \) iterations, except for the A-orthonormalizations, is

\[
\text{Total Flops} = 4\text{nnz} - nt + 5n - 1 + k_c [11nt - 2t + 2n - 1 + 2\text{nnz} + n + 1] \\
= 4\text{nnz} - nt + 5n - 1 + k_c [11nt + 3n - 2t + 2\text{nnz}] \\
\approx 4\text{nnz} + 5n + k_c [11nt + 2\text{nnz}],
\]

which is of the order of \( n \text{nnz} k_c + ntk_c \) flops, where \( \text{nnz} \) is the number of nonzero entries in the \( n \times n \) matrix \( A \) and \( t \) is the number of search directions computed at each iteration.
3.3 Long Recurrence Enlarged Conjugate Gradient (LRE-CG) Method

In this section, we introduce the long recurrence enlarged conjugate gradient (LRE-CG) method which is an enlarged Krylov projection method that solves the system $Ax = b$ by approximating the solution at the $k^{th}$ iteration with the vector $x_k = x_{k-1} + Q_k \alpha_k \in x_0 + \mathcal{K}_{t,k}$ such that

$$\phi(x_k) = \min \{ \phi(x), \forall x \in x_0 + \mathcal{K}_{t,k} \},$$

where $Q_k \alpha_k \in \mathcal{K}_{t,k}$ and $Q_k$ is an $n \times tk$ matrix containing the orthonormal basis vectors of $\mathcal{K}_{t,k}$ and $\phi(x) = \frac{1}{2}x^tAx - x^tb$. The LRE-CG method does not have short recurrences as MSDO-CG, but it has similarities with GMRES in that the whole basis is used to define the new approximate solution rather than $t$ search directions. As mentioned earlier, the minimum of $\phi(x)$ is given by $\nabla \phi(x) = 0$ which is equivalent to $Ax - b = 0$. Thus, by minimizing $\phi(x)$ we are solving the system $Ax = b$. Since $\phi(x_k) = \min \{ \phi(x), \forall x \in x_0 + \mathcal{K}_{t,k} \}$, then

$$\phi(x_k) = \phi(x_{k-1} + Q_k \alpha_k) = \min \{ \phi(x_{k-1} + Q_k \alpha), \forall \alpha \in \mathbb{R}^{tk} \}. \quad (11)$$

Once $x_k$ has been chosen, either $x_k$ is the exact solution of $Ax = b$, or $t$ new basis vectors and the new approximation $x_{k+1} = x_k + Q_{k+1} \alpha_{k+1}$ are computed. This procedure is repeated until convergence.

Thus, we need to find the recursion relations of $r_k$ and $\alpha_k$. By definition, the residual $r_k = b - Ax_k$ where $x_k \in x_0 + \mathcal{K}_{t,k}$. Thus $r_k \in \mathcal{K}_{t,k+1}$. The recursion relation of $r_k$ can be simply obtained by replacing $x_k$ by its expression as follows:

$$r_k = b - Ax_k$$

$$= b - A(x_{k-1} + Q_k \alpha_k)$$

$$= r_{k-1} - A \tilde{Q}_k \alpha_k.$$

At each iteration the step $\alpha_{k+1}$ is chosen such that

$$\phi(x_{k+1}) = \min \{ \phi(x_k + Q_{k+1} \alpha), \forall \alpha \in \mathbb{R}^t(k+1) \}.$$

Let $F(\alpha) = \phi(x_k + Q_{k+1} \alpha)$ where $\phi(x) = \frac{1}{2}x^tAx - x^tb$. Then,

$$F(\alpha) = \frac{1}{2} (x_k + Q_{k+1} \alpha)^t A (x_k + Q_{k+1} \alpha) - (x_k + Q_{k+1} \alpha)^t b$$

$$= \phi(x_k) + \frac{1}{2} [(x_k)^t A Q_{k+1} \alpha + \alpha^t (Q_{k+1})^t A x_k + \alpha^t (Q_{k+1})^t A Q_{k+1} \alpha - \alpha^t (Q_{k+1})^t b$$

$$= \phi(x_k) + \frac{1}{2} [(x_k)^t A Q_{k+1} \alpha - \alpha^t (Q_{k+1})^t A x_k] + \frac{1}{2} \alpha^t (Q_{k+1})^t A Q_{k+1} \alpha - \alpha^t (Q_{k+1})^t r_k$$

$$= \phi(x_k) + \frac{1}{2} \alpha^t (Q_{k+1})^t A Q_{k+1} \alpha - \alpha^t (Q_{k+1})^t r_k, \quad \text{since } A \text{ is SPD.}$$

The minimum of $F(\alpha)$ is given by $F'(\alpha) = 0$

Therefore, $\alpha_{k+1} = (Q_{k+1}^t A Q_{k+1})^{-1} (Q_{k+1}^t r_k)$.

By minimizing $\phi(x)$, the Petrov-Galerkin condition, $r_k \perp \mathcal{K}_{t,k}$, is ensured (Theorem 3.11). Therefore, $(r_k)^t r_i = 0$, for all $i < k$, and the residuals form an orthogonal set.

**Theorem 3.11.** The Petrov-Galerkin condition in LRE-CG, $r_k \perp \mathcal{K}_{t,k}$, is equivalent to $x_k$ being the minimum of $\phi(x)$ in $x_0 + \mathcal{K}_{t,k}$.

**Proof.**
1. \( x_k \) is the minimum of \( \phi(x) \) in \( x_0 + \mathcal{K}_{t,k} \) implies \( r_k \perp \mathcal{K}_{t,k} \). 
The minimum of \( F(\alpha) = \phi(x_k) = \phi(x_k - 1 + Q_k \alpha) \) is given by
\[
F'(\alpha) = (Q_k)^T A Q_k \alpha - (Q_k)^T r_{k-1} = 0.
\] 
Since \( x_k \) is the minimum, then \( \alpha = \alpha_k \) and \( F'(\alpha) =
- Q_k^T r_k = 0 \). Thus \( r_k \perp \mathcal{K}_{t,k} \).

2. \( r_k \perp \mathcal{K}_{t,k} \) implies \( x_k \) is the minimum of \( \phi(x) \) in \( x_0 + \mathcal{K}_{t,k} \) (Proof by contradiction)
Assume that \( r_k \perp \mathcal{K}_{t,k} \) and \( x_k \) is not the minimum of \( \phi(x) \) in \( x_0 + \mathcal{K}_{t,k} \). Then \( F'(\alpha_k) \neq 0 \).
Hence \( Q_k^T r_k \neq 0 \) and \( r_k \) is not orthogonal to \( \mathcal{K}_{t,k} \). This contradicts our assumption. Thus \( x_k \) is the minimum of \( \phi(x) \).

3.3.1 The LRE-CG Algorithm
After deriving the expressions and the recursion relations of
\[
\begin{align*}
 x_k &= x_{k-1} + Q_k \alpha_k, \\
 r_k &= r_{k-1} - A Q_k \alpha_k, \\
 \alpha_k &= (Q_k^T A Q_k)^{-1} (Q_k^T r_{k-1}),
\end{align*}
\]
we present in Algorithm 2 the LRE-CG algorithm and the performed flops, except for the orthonormalization. We refer to the cost of solving the \( tk \times tk \) linear system from step 5 in Algorithm 2 as Solve_{\alpha}(tk).

Algorithm 2 The LRE-CG algorithm

<table>
<thead>
<tr>
<th>Flops</th>
</tr>
</thead>
</table>
| Input: \( A \), the \( n \times n \) symmetric positive definite matrix
| Input: \( b \), the \( n \times 1 \) right-hand side; \( x_0 \), the initial guess or iterate
| Input: \( \epsilon \), the stopping tolerance; \( k_{\text{max}} \), the maximum allowed iterations
| Output: \( x_k \), the approximate solution of the system \( Ax = b \) |
| 1: \( r_0 = b - A x_0, \rho_0 = ||r_0||^2, k = 1 \) |
| 2: Let \( W = T(r_0) \), normalise its vectors and then let \( Q = W \) |
| 3: while (\( \sqrt{\rho_{k-1}} > \epsilon ||b||_2 \) and \( k < k_{\text{max}} \)) do |
| 4: \( G = (Q^T A Q) \) |
| 5: \( \alpha = G^{-1} (Q^T r) \) |
| 6: \( x = x + Q \alpha \) |
| 7: \( r = r - A Q \alpha \) |
| 8: \( \rho_k = ||r||^2 \) |
| 9: Let \( W = AW \) |
| 10: Orthonormalise the vectors of \( W \) against the vectors of \( Q \) |
| 11: Orthonormalise the vectors of \( W \) and let \( Q = [Q \ W] \) |
| 12: \( k = k+1 \) |
| 13: end while |

The cost of the LRE-CG, using Algorithm 2 except for the orthonormalization in steps 10 and 11, is
EKS-CG methods for reducing communication

where the first term $\text{nnz} k_c^2$ corresponds to the multiplication $AQ$ and the second term $nt^2 k_c^2$ corresponds to the orthonormalization with respect to previous vectors. As for the memory requirements, we have to store the matrix $A$ and $tk_c + 2$ vectors of size $n \times 1$. And there should be enough memory for the $tk_c \times tk_c$ matrix $Q^t AQ$.

However, the multiplication $Q^t_k AQ_k$ can be reduced since $Q_k = [Q_{k-1} W_k]$. Let $Z_k = AW_k$, then $D_k = AQ_k = [AQ_{k-1} Z_k]$. At iteration $k - 1$, $D_{k-1} = AQ_{k-1}$ is computed. Thus at iteration $k$, only $Z_k = AW_k$ is computed. As for $G_k = Q^t_k AQ_k$, it is equal to

$$G_k = Q^t_k AQ_k = Q^t_k D_k = \left( \begin{array}{cc} Q^t_{k-1} D_{k-1} & Q^t_{k-1} Z_k \\ W^t_k D_{k-1} & W^t_k Z_k \end{array} \right) = \left( \begin{array}{cc} G_{k-1} & Q^t_{k-1} Z_k \\ Z_k^t Q_{k-1} & W^t_k Z_k \end{array} \right) = \left( \begin{array}{cc} G_{k-1} F_k & F_k \\ Z_k^t E_k & E_k \end{array} \right),$$

where $G_{k-1}$ is computed at iteration $k - 1$, $F_k = Q^t_{k-1} Z_k$, and $E_k = W^t_k Z_k$. Thus computing $G_k = Q^t_k AQ_k$ can be reduced to computing $F_k$ and $E_k$.

Algorithm 3 The LRE-CG Algorithm

<table>
<thead>
<tr>
<th>Flops</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input:</strong> $A$, the $n \times n$ symmetric positive definite matrix</td>
</tr>
<tr>
<td><strong>Input:</strong> $b$, the $n \times 1$ right-hand side; $x_0$, the initial guess or iterate</td>
</tr>
<tr>
<td><strong>Input:</strong> $\epsilon$, the stopping tolerance; $k_{\text{max}}$, the maximum allowed iterations</td>
</tr>
<tr>
<td><strong>Output:</strong> $x_k$, the approximate solution of the system $Ax = b$</td>
</tr>
<tr>
<td>$r_0 = b - Ax_0, \rho_0 = |r_0|^2_2, k = 1$</td>
</tr>
<tr>
<td>$2\text{nnz} + 2n - 1$</td>
</tr>
<tr>
<td>Let $W = T(r_0)$, normalise its vectors and then let $Q = W$</td>
</tr>
<tr>
<td>$3n$</td>
</tr>
<tr>
<td>while $(\sqrt{\rho_{k-1}} &gt; \epsilon</td>
</tr>
<tr>
<td>$2n$</td>
</tr>
<tr>
<td>$Z = AW$</td>
</tr>
<tr>
<td>$(2nnz - nt)t$</td>
</tr>
<tr>
<td>$E = W^t Z$</td>
</tr>
<tr>
<td>$(2n-1)t^2$</td>
</tr>
<tr>
<td>if $k == 1$ then</td>
</tr>
<tr>
<td>$D = Z$ and $G = E$</td>
</tr>
<tr>
<td>else</td>
</tr>
<tr>
<td>$F = Q(:, 1 : t(k-1))^t Z$</td>
</tr>
<tr>
<td>$(2n-1)t^2(k-1)$</td>
</tr>
<tr>
<td>$G = \left( \begin{array}{cc} G &amp; F \ F^t &amp; E \end{array} \right)$ and $D = [D \ Z]$</td>
</tr>
<tr>
<td>$2nnz + 2n - 1$</td>
</tr>
<tr>
<td>end if</td>
</tr>
<tr>
<td>$\alpha = G^{-1}(Q^t r)$</td>
</tr>
<tr>
<td>$(2n-1)tk + \text{Solve}_{\alpha}(tk)$</td>
</tr>
<tr>
<td>$x = x + Q\alpha$</td>
</tr>
<tr>
<td>$2tkn$</td>
</tr>
<tr>
<td>$r = r - D\alpha$</td>
</tr>
<tr>
<td>$2tkn$</td>
</tr>
<tr>
<td>$\rho_k =</td>
</tr>
<tr>
<td>$2n - 1$</td>
</tr>
<tr>
<td>Let $W = Z$</td>
</tr>
<tr>
<td>not included in here</td>
</tr>
<tr>
<td>Orthonormalise the columns of $W$ against those of $Q$</td>
</tr>
<tr>
<td>not included in here</td>
</tr>
<tr>
<td>Orthonormalise the vectors of $W$ and let $Q = [Q \ W]$</td>
</tr>
<tr>
<td>$k = k + 1$</td>
</tr>
<tr>
<td>end while</td>
</tr>
</tbody>
</table>

RR n° 8597
Then, the cost of $k_c$ iterations of LRE-CG using Algorithm\(^3\) except for the orthonormalization, is

\[
\text{Total Flops} = 2nnz + 7n - 1 + k_c[(2nnz - nt - (6n - 1)t + (2n - 1)t^2)k^{k-1} + 2n] + \sum_{k=1}^{k_c} \text{Solve}_o(tk)
\]

\[
= 2nnz + 7n - 1 + k_c[(2nnz + 2n - \frac{1}{2})t + (n - \frac{1}{2})t^2 + (6n - 1)t^2 k_c^{k-1} + (2n - 1)t^2 k_c^{k-2} + 2n] + \sum_{k=1}^{k_c} \text{Solve}_o(tk)
\]

\[
\approx 3ntk_c^2 + nt^2k_c^n + nt^2k_c + \sum_{k=1}^{k_c} \text{Solve}_o(tk)
\]

Note that $tk_c$ should be much smaller than $n$, or otherwise the cost of Algorithm\(^3\) would be $O(n^3 + \text{Solve}_o(n))$, and $n$ vectors of size $n$ have to be stored in addition to the $n \times n$ system $Q_k^tAQ_k$.

One remedy to this problem, that we do not address in this paper, is to restart LRE-CG after some iterations. But this restart might have an effect on convergence as in restarted GMRES. Another alternative is to choose a linearly independent subset of the $t$ computed vectors at each iteration $i$. This reduces the size of the system solved at each iteration. A third alternative is to compute at each iteration $i$, $t_i$ vectors and then choose a linearly independent set of cardinality $\tilde{t}_i$, where $t_0 = t$, $t_i \leq t$, $\tilde{t}_i = t_{i+1}$. This reduces not only the size of the system solved at each iteration, but also the memory requirements and the number of computed vectors per iteration. In exact precision, the second and third alternatives are equivalent by Theorem\(^3\) since if a vector $T_j(r_0)$ is linearly dependent on $\{T_1(r_0), \ldots, T_{j-1}(r_0), T_{j+1}(r_0), \ldots, T_t(r_0)\}$ then $AT_j(r_0)$ is linearly dependent on $\{AT_1(r_0), \ldots, AT_{j-1}(r_0), AT_{j+1}(r_0), \ldots, AT_t(r_0)\}$. However, this has to be tested in finite precision. Note that there is an additional cost for choosing a linearly independent subset of the $t$ or $t_i$ vectors.

The $tk \times tk$ $\alpha$ system can be solved using iterative methods like Jacobi method or Krylov subspace methods. Moreover, the $s$-step or communication avoiding Krylov subspace methods can be used to reduce communication. We use Matlab’s backslash to solve the $\alpha$ systems in the convergence tests that follow.

### 4 Convergence Results

After introducing the new CG methods, MSDO-CG and LRE-CG, we compare their convergence behavior with respect to different A-orthonormalization and orthonormalization schemes respectively. Then we compare the convergence behavior of both methods with respect to CG, Coop-CG, MSD-CG on several matrices for different number of partitions (2, 4, 8, 16, 32 and 64 partitions) or number of initial guesses (for Coop-CG only). The matrices are first reordered using Metis’s k-way partitioning\(^18\) that defines the subdomains $\delta_i$. Then $x$ is chosen randomly using Matlab’s rand function and $b = A \cdot x$, except for the Elasticity3Dmatrix where $A$ and $b$ are available. In tables\(^2\) and \(^3\) “Iter” is the number of iterations, $k_c$, needed for convergence and “Err” is the relative error $\frac{\|b - Ax\|}{\|b\|}$ at convergence.

The first matrix POISSON2Dis a block tridiagonal matrix obtained from Poisson’s equation (sparse) using Matlab’s gallery(’poisson’,100) function. The matrices refered to as NH2D, SKY2D, SKY3D, and ANI3D, arise from boundary value problem of the convection diffusion equations

\[
\eta(x)u + \text{div}(a(x)u) - \text{div}(\kappa(x)\nabla u) = f \quad \text{in} \quad \Omega
\]

\[
u = 0 \quad \text{on} \quad \partial \Omega_D
\]

\[
\frac{\partial u}{\partial n} = 0 \quad \text{on} \quad \partial \Omega_N
\]

where $\Omega = [0, 1]^n$, $n = 2$, or 3 and $\partial \Omega_N = \partial \Omega \setminus \partial \Omega_D$. The function $\eta$, the vector field $a$, and the tensor $\kappa$ are the given coefficients of the partial differential operator. In the 2D case, we have $\partial \Omega_D = [0, 1] \times [0, 1]$, and in the 3D case, we have $\partial \Omega_D = [0, 1] \times [0, 1] \times [0, 1]$. We focus on the following cases:

- **NH2D**: A non-homogeneous problem with large jumps in the coefficients. The coefficient $\eta$ and $a$ are both zero. The tensor $\kappa$ is isotropic and discontinuous. It jumps from the constant value $10^3$ in the ring $\frac{1}{2} \leq |x - c| \leq 1$, $c = (\frac{1}{2}, \frac{1}{2})^T$, to 1 outside.

Inria
SKY2D and SKY3D Skyscraper problems: The tensor \( \kappa \) is isotropic and discontinuous. The domain contains many zones of high permeability which are isolated from each other

\[
\kappa(x) = \begin{cases} 
10^3 \times \lceil 10 \times x_2 \rceil + 1 & \text{if } \lceil 10x_1 \rceil \text{ is odd, } i = 1, 2 \\
1, & \text{otherwise.}
\end{cases}
\]

where we note \( \lceil x \rceil \) as the integer value of \( x \). SKY2D and SKY3D are discretized on a 2D and 3D cartesian grids respectively.

Anisotropic layers: the domain is made of 10 anisotropic layers with jumps of up to four orders of magnitude and an anisotropy ratio of up to \( 10^3 \) in each layer. The domain is divided into 10 layers parallel to \( z = 0 \), of size 0.1, in which the coefficients are constant. We have \( \kappa_y = 10\kappa_x \) and \( \kappa_z = 1000\kappa_x \). The velocity field is zero.

POISSON2D, NH2D and SKY2D are discretized on a \( 100 \times 100 \) 2D cartesian grid. SKY3D and ANI3D are discretized on a \( 20 \times 20 \times 20 \) grid.

As for the ELASTICITY3D matrix, it arises from the linear elasticity problem with Dirichlet and Neumann boundary conditions, defined as follows:

\[
\begin{align*}
\text{div}(\sigma(u)) + f &= 0 & \text{on } \Omega, \\
\sigma(u) \cdot n &= g & \text{on } \partial \Omega_N, \\
u &= u_D & \text{on } \partial \Omega_D,
\end{align*}
\]

where \( \Omega \) is a 3D \( 30 \times 10 \times 10 \) parallelepiped, \( \Omega_D \) is the Dirichlet boundary, \( \Omega_N \) is the Neumann boundary, \( u \) is the unknown displacement field, \( f \) is some body force, \( \sigma(u) \) is the Cauchy stress tensor given by Hooke’s law. The ELASTICITY3D matrix was discretized with P1 finite elements and a triangular mesh using FreeFem++ [15]. For a detailed description of the problem refer to [12]. Table 1 briefly describes the test matrices.

<table>
<thead>
<tr>
<th>Matrix</th>
<th>Size</th>
<th>Nonzeros</th>
<th>Symmetric</th>
<th>2D/3D</th>
<th>Problem</th>
</tr>
</thead>
<tbody>
<tr>
<td>POISSON2D</td>
<td>10000</td>
<td>49600</td>
<td>Yes</td>
<td>2D</td>
<td>Poisson equations</td>
</tr>
<tr>
<td>NH2D</td>
<td>10000</td>
<td>49600</td>
<td>Yes</td>
<td>2D</td>
<td>Boundary value</td>
</tr>
<tr>
<td>SKY2D</td>
<td>10000</td>
<td>49600</td>
<td>Yes</td>
<td>2D</td>
<td>Boundary value</td>
</tr>
<tr>
<td>SKY3D</td>
<td>8000</td>
<td>53600</td>
<td>Yes</td>
<td>3D</td>
<td>Skyscraper</td>
</tr>
<tr>
<td>ANI3D</td>
<td>8000</td>
<td>53600</td>
<td>Yes</td>
<td>3D</td>
<td>Anisotropic Layers</td>
</tr>
<tr>
<td>ELASTICITY3D</td>
<td>11253</td>
<td>373647</td>
<td>Yes</td>
<td>3D</td>
<td>Linear Elasticity P1 FE</td>
</tr>
</tbody>
</table>

In table 2 we compare the convergence behavior of the MSDO-CG method (Algorithm 1) with different A-orthonormalization schemes for A-orthonormalizing \( P_k \) against previous \( P_i \)'s (MGS, CGS, CGS2) and then A-orthonormalizing \( P_k \) against itself (MGS, CGS, CGS2, A-CholQR, Pre-CholQR) and for different number of partitions \( t = 2, 4, 8, 16, 32, 64 \) that correspond to the maximum number of vectors added at each iteration to the enlarged Krylov subspace, \( \mathcal{K}_{t,k} \). We have tested different combinations of A-orthonormalization, but we only show MGS (MGS+MGS), CGS+A-CholQR, CGS+Pre-CholQR, CGS2+PRE-CholQR, and CGS2+PRE-CholQR. Note that MSDO-CG with CGS A-orthonormalization (CGS+CGS) did not converge neither with CGS2 A-orthonormalization (CGS2+CGS2) nor with CGS2+CGS or CGS+CGS2 A-orthonormalization. The reason is that the search directions are not A-orthogonal to satisfactory precision. And by Theorem 3.10, this implies that \( r_k \not\perp \mathcal{K}_{t,k} \). Thus, nothing guarantees convergence since we have shown in section 3.1.3 that MSDO-CG converges faster than CG if \( r_k \perp \mathcal{K}_{t,k} \). Moreover, we did not test combinations of MGS and QR factorizations since MGS is expensive in terms of communication compared to the other methods (section 5.1). But we tested MSDO-CG with MGS for comparison purposes since MGS is known for its numerical stability.
Table 2: Comparison of the convergence of MSDO-CG with different A-orthonormalization schemes, with respect to number of partitions (Pa) with $x_0 = 0$.

<table>
<thead>
<tr>
<th>MSDO-CG with different A-orthonormalization Methods</th>
<th>MGS</th>
<th>CGS+A-CholQR</th>
<th>CGS+Pre-CholQR</th>
<th>CGS2+A-CholQR</th>
<th>CGS2+Pre-CholQR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pa</td>
<td>Iter</td>
<td>Err</td>
<td>Iter</td>
<td>Err</td>
<td>Iter</td>
</tr>
<tr>
<td>---</td>
<td>------</td>
<td>-----</td>
<td>------</td>
<td>-----</td>
<td>------</td>
</tr>
<tr>
<td>2</td>
<td>200</td>
<td>4E-5</td>
<td>204</td>
<td>3E-5</td>
<td>204</td>
</tr>
<tr>
<td>4</td>
<td>167</td>
<td>2E-5</td>
<td>167</td>
<td>2E-5</td>
<td>167</td>
</tr>
<tr>
<td>8</td>
<td>139</td>
<td>1E-5</td>
<td>139</td>
<td>1E-5</td>
<td>139</td>
</tr>
<tr>
<td>16</td>
<td>121</td>
<td>5E-6</td>
<td>121</td>
<td>5E-6</td>
<td>121</td>
</tr>
<tr>
<td>32</td>
<td>94</td>
<td>2E-6</td>
<td>94</td>
<td>2E-6</td>
<td>94</td>
</tr>
<tr>
<td>64</td>
<td>69</td>
<td>2E-6</td>
<td>69</td>
<td>2E-6</td>
<td>69</td>
</tr>
</tbody>
</table>

As shown in Table 2, MSDO-CG with MGS A-orthonormalization converges for all the tested matrices and as we increase $t$, the number of iterations needed for convergence decreases. As we mentioned earlier, MSDO-CG with CGS A-orthonormalization did not converge. Therefore, we replaced CGS with CGS+A-CholQR and with CGS+Pre-CholQR A-orthonormalization. We notice that MSDO-CG with CGS+A-CholQR A-orthonormalization and MSDO-CG with CGS+Pre-CholQR A-orthonormalization have the same convergence behavior. For the matrices Poisson2D and NH2D, both methods converge with the same number of iterations as MSDO-CG with MGS A-orthonormalization. However, for the matrix Sky2D, both methods did not converge. As for the matrices Sky3D and AN13D, both methods converged only for $t = 2$ partitions, and $t = 2, 4, 8$ partitions respectively. The reason for this difference in behavior for different matrices is the condition number ($\text{cond}_2 = ||A||_2||A^{-1}||_2$). The condition number of the matrices Poisson2D and NH2D is $6 \times 10^3$, whereas that of the matrices Sky3D, AN13D
and Sky2D is $1 \times 10^6$, $2 \times 10^6$, and $3 \times 10^7$ respectively. Although it was shown in [20] that Pre-CholQR A-orthonormalization is more stable than A-CholQR, however MSDO-CG with CGS+A-CholQR A-orthonormalization and MSDO-CG with CGS+Pre-CholQR A-orthonormalization are both numerically unstable.

Thus, we replace CGS with CGS2 where the A-orthonormalization is performed twice for numerical stability. Then, the MSDO-CG with CGS2+A-CholQR A-orthonormalization and MSDO-CG with CGS2+Pre-CholQR A-orthonormalization converge as fast as MSDO-CG with MGS A-orthonormalization for all $t$ and all the tested matrices. Hence, we conclude that MGS, CGS2+A-CholQR, and CGS2+Pre-CholQR A-orthonormalizations are stable enough to be used in the MSDO-CG method (Algorithm 1).

Table 3: Comparison of the convergence of the LRE-CG method with different orthonormalization schemes, with respect to number of partitions $P_a$, with $x_0 = 0$.

<table>
<thead>
<tr>
<th>$P_a$</th>
<th>Iter</th>
<th>Err</th>
<th>Iter</th>
<th>Err</th>
<th>Iter</th>
<th>Err</th>
</tr>
</thead>
<tbody>
<tr>
<td>POISSON2D</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tol = $10^{-6}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>193</td>
<td>2E-5</td>
<td>193</td>
<td>2E-5</td>
<td>193</td>
<td>2E-5</td>
</tr>
<tr>
<td>4</td>
<td>153</td>
<td>1E-5</td>
<td>153</td>
<td>1E-5</td>
<td>153</td>
<td>1E-5</td>
</tr>
<tr>
<td>8</td>
<td>123</td>
<td>8E-6</td>
<td>123</td>
<td>8E-6</td>
<td>123</td>
<td>8E-6</td>
</tr>
<tr>
<td>16</td>
<td>95</td>
<td>4E-6</td>
<td>95</td>
<td>4E-6</td>
<td>95</td>
<td>4E-6</td>
</tr>
<tr>
<td>32</td>
<td>70</td>
<td>2E-6</td>
<td>70</td>
<td>2E-6</td>
<td>70</td>
<td>2E-6</td>
</tr>
<tr>
<td>64</td>
<td>52</td>
<td>1E-6</td>
<td>52</td>
<td>1E-6</td>
<td>52</td>
<td>1E-6</td>
</tr>
<tr>
<td>NH2D</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tol = $10^{-8}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>245</td>
<td>1E-7</td>
<td>245</td>
<td>1E-7</td>
<td>245</td>
<td>1E-7</td>
</tr>
<tr>
<td>4</td>
<td>188</td>
<td>1E-7</td>
<td>188</td>
<td>1E-7</td>
<td>188</td>
<td>1E-7</td>
</tr>
<tr>
<td>8</td>
<td>149</td>
<td>5E-8</td>
<td>149</td>
<td>5E-8</td>
<td>149</td>
<td>5E-8</td>
</tr>
<tr>
<td>16</td>
<td>112</td>
<td>3E-8</td>
<td>112</td>
<td>3E-8</td>
<td>112</td>
<td>3E-8</td>
</tr>
<tr>
<td>32</td>
<td>82</td>
<td>2E-8</td>
<td>82</td>
<td>2E-8</td>
<td>82</td>
<td>2E-8</td>
</tr>
<tr>
<td>64</td>
<td>60</td>
<td>1E-8</td>
<td>60</td>
<td>1E-8</td>
<td>60</td>
<td>1E-8</td>
</tr>
<tr>
<td>SKY2D</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tol = $10^{-8}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1415</td>
<td>5E-04</td>
<td>1415</td>
<td>5E-04</td>
<td>1415</td>
<td>5E-04</td>
</tr>
<tr>
<td>4</td>
<td>757</td>
<td>1E-4</td>
<td>757</td>
<td>1E-4</td>
<td>757</td>
<td>1E-4</td>
</tr>
<tr>
<td>8</td>
<td>398</td>
<td>1E-4</td>
<td>398</td>
<td>1E-4</td>
<td>398</td>
<td>1E-4</td>
</tr>
<tr>
<td>16</td>
<td>220</td>
<td>9E-5</td>
<td>220</td>
<td>9E-5</td>
<td>220</td>
<td>9E-5</td>
</tr>
<tr>
<td>32</td>
<td>126</td>
<td>5E-5</td>
<td>126</td>
<td>5E-5</td>
<td>126</td>
<td>5E-5</td>
</tr>
<tr>
<td>64</td>
<td>75</td>
<td>3E-5</td>
<td>75</td>
<td>3E-5</td>
<td>75</td>
<td>3E-5</td>
</tr>
<tr>
<td>SKY3D</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tol = $10^{-8}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>557</td>
<td>2E-5</td>
<td>570</td>
<td>1E-5</td>
<td>563</td>
<td>1E-5</td>
</tr>
<tr>
<td>4</td>
<td>373</td>
<td>2E-5</td>
<td>373</td>
<td>2E-5</td>
<td>377</td>
<td>1E-5</td>
</tr>
<tr>
<td>8</td>
<td>211</td>
<td>1E-5</td>
<td>211</td>
<td>1E-5</td>
<td>211</td>
<td>1E-5</td>
</tr>
<tr>
<td>16</td>
<td>119</td>
<td>9E-6</td>
<td>119</td>
<td>9E-6</td>
<td>119</td>
<td>9E-6</td>
</tr>
<tr>
<td>32</td>
<td>69</td>
<td>9E-6</td>
<td>69</td>
<td>9E-6</td>
<td>69</td>
<td>9E-6</td>
</tr>
<tr>
<td>64</td>
<td>43</td>
<td>4E-6</td>
<td>42</td>
<td>1E-5</td>
<td>42</td>
<td>1E-5</td>
</tr>
<tr>
<td>ANI3D</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tol = $10^{-8}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>875</td>
<td>7e-5</td>
<td>875</td>
<td>7E-5</td>
<td>875</td>
<td>7e-5</td>
</tr>
<tr>
<td>4</td>
<td>673</td>
<td>8e-5</td>
<td>673</td>
<td>8e-5</td>
<td>673</td>
<td>8e-5</td>
</tr>
<tr>
<td>8</td>
<td>449</td>
<td>1e-4</td>
<td>449</td>
<td>1e-4</td>
<td>449</td>
<td>1e-4</td>
</tr>
<tr>
<td>16</td>
<td>253</td>
<td>2e-4</td>
<td>253</td>
<td>2e-4</td>
<td>253</td>
<td>2e-4</td>
</tr>
<tr>
<td>32</td>
<td>148</td>
<td>2e-4</td>
<td>148</td>
<td>2e-4</td>
<td>148</td>
<td>2e-4</td>
</tr>
<tr>
<td>64</td>
<td>92</td>
<td>1e-4</td>
<td>92</td>
<td>1e-4</td>
<td>92</td>
<td>1e-4</td>
</tr>
</tbody>
</table>

In table 3 we compare the convergence behavior of the LRE-CG method (Algorithm 3) with dif-
Table 4: Comparison between the convergence of the different CG versions with respect to number of partitions or initial guesses for Coop-CG with $x_0 = 0$.

<table>
<thead>
<tr>
<th></th>
<th>CG</th>
<th>Coop-CG</th>
<th>MSD-CG</th>
<th>MSDO-CG</th>
<th>LRE-CG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pa</td>
<td>Iter</td>
<td>Err</td>
<td>Iter</td>
<td>Err</td>
<td>Iter</td>
</tr>
<tr>
<td>POISSON2D</td>
<td>2</td>
<td>195</td>
<td>2E-5</td>
<td>206</td>
<td>2E-7</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>195</td>
<td>2E-5</td>
<td>171</td>
<td>1E-7</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>195</td>
<td>2E-5</td>
<td>137</td>
<td>1E-7</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>195</td>
<td>2E-5</td>
<td>106</td>
<td>3E-8</td>
</tr>
<tr>
<td></td>
<td>32</td>
<td>195</td>
<td>2E-5</td>
<td>80</td>
<td>1E-8</td>
</tr>
<tr>
<td></td>
<td>64</td>
<td>195</td>
<td>2E-5</td>
<td>59</td>
<td>1E-8</td>
</tr>
<tr>
<td>NH2D</td>
<td>2</td>
<td>259</td>
<td>4E-7</td>
<td>206</td>
<td>2E-7</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>259</td>
<td>4E-7</td>
<td>179</td>
<td>1E-7</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>259</td>
<td>4E-7</td>
<td>157</td>
<td>2.0E-5</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>259</td>
<td>4E-7</td>
<td>107</td>
<td>2E-8</td>
</tr>
<tr>
<td></td>
<td>32</td>
<td>259</td>
<td>4E-7</td>
<td>81</td>
<td>2E-8</td>
</tr>
<tr>
<td></td>
<td>64</td>
<td>259</td>
<td>4E-7</td>
<td>59</td>
<td>1E-8</td>
</tr>
<tr>
<td>SKY2D</td>
<td>2</td>
<td>5951</td>
<td>4E-4</td>
<td>4893</td>
<td>2E-4</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>5951</td>
<td>4E-4</td>
<td>3737</td>
<td>9E-5</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>5951</td>
<td>4E-4</td>
<td>3391</td>
<td>1E-5</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>5951</td>
<td>4E-4</td>
<td>2437</td>
<td>9E-6</td>
</tr>
<tr>
<td></td>
<td>32</td>
<td>5951</td>
<td>4E-4</td>
<td>1406</td>
<td>4E-6</td>
</tr>
<tr>
<td></td>
<td>64</td>
<td>5951</td>
<td>4E-4</td>
<td>802</td>
<td>2E-6</td>
</tr>
<tr>
<td>SKY3D</td>
<td>2</td>
<td>902</td>
<td>1E-5</td>
<td>795</td>
<td>8E-6</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>902</td>
<td>1E-5</td>
<td>627</td>
<td>1E-5</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>902</td>
<td>1E-5</td>
<td>542</td>
<td>4E-6</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>902</td>
<td>1E-5</td>
<td>414</td>
<td>3E-6</td>
</tr>
<tr>
<td></td>
<td>32</td>
<td>902</td>
<td>1E-5</td>
<td>290</td>
<td>1E-6</td>
</tr>
<tr>
<td></td>
<td>64</td>
<td>902</td>
<td>1E-5</td>
<td>183</td>
<td>8E-7</td>
</tr>
<tr>
<td>AN13D</td>
<td>2</td>
<td>4187</td>
<td>4E-5</td>
<td>3584</td>
<td>5e-5</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>4146</td>
<td>4E-5</td>
<td>3371</td>
<td>4e-5</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>4146</td>
<td>4e-5</td>
<td>2865</td>
<td>4e-5</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>4146</td>
<td>4e-5</td>
<td>2314</td>
<td>3e-5</td>
</tr>
<tr>
<td></td>
<td>32</td>
<td>4146</td>
<td>4e-5</td>
<td>1615</td>
<td>2e-5</td>
</tr>
<tr>
<td></td>
<td>64</td>
<td>4146</td>
<td>4e-5</td>
<td>1002</td>
<td>1e-5</td>
</tr>
<tr>
<td>ELASTICITY3D</td>
<td>2</td>
<td>987</td>
<td>4e-12</td>
<td>718</td>
<td>3e-12</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>987</td>
<td>4e-12</td>
<td>534</td>
<td>8e-12</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>987</td>
<td>4e-12</td>
<td>425</td>
<td>6e-11</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>987</td>
<td>4e-12</td>
<td>348</td>
<td>6e-11</td>
</tr>
<tr>
<td></td>
<td>32</td>
<td>987</td>
<td>4e-12</td>
<td>294</td>
<td>9e-12</td>
</tr>
<tr>
<td></td>
<td>64</td>
<td>987</td>
<td>4e-12</td>
<td>235</td>
<td>1e-11</td>
</tr>
</tbody>
</table>

Different orthonormalization schemes for orthonormalizing $W$ against the $n \times tl$ matrix $Q$ (MGS, CGS) and then orthonormalizing $W$ against itself (MGS, CGS, TSQR) and for different number of partitions $t = 2, 4, 8, 16, 32, 64$ that correspond to the maximum number of vectors added at each iteration to the enlarged Krylov subspace, $\mathcal{K}_{t,k}$. We start by testing the convergence of LRE-CG with MGS (MGS+MGS) orthonormalization. It converges for all the tested matrices since it is numerically stable, and the number
of iterations needed for convergence decreases when increasing the number of partitions \( t \). However, as mentioned in section 5.1, MGS is expensive in terms of communication \( O(tk\log(t)) \) messages per iteration, where \( t \) processors A-orthonormalized \( t \) vectors against \( tk \) vectors. Thus, we tested the LRE-CG method with Classical Gram Schmidt (CGS) orthogonization which requires sending \( O(tk\log(t)) \) messages per iteration. The LRE-CG with CGS converges in the same number of iterations as LRE-CG with MGS for the matrices POISSON2D and NH2D. However, for the other matrices, it does not converge for the given stopping criteria except for \( t = 2 \) as shown in Table 5. The reason is that the matrix \( C = Q^tAQ \) is becoming close to singular, with \( \text{rank}(C) < tk \), as the iterations proceed due to the loss of orthogonality in the CGS orthogonalization. The number of iterations in parentheses in Table 5 is not the number of iterations for convergence but it denotes the iteration at which the \( C \) matrix becomes close to singular.

In CA-GMRES [21], the authors use a parallelizable tall and skinny QR (TSQR) factorization [4] for orthonormalizing the \( n \times t \) tall and skinny matrix instead of CGS. They have shown that the combination of CGS for orthonormalization \( W \) against \( Q \) and TSQR for orthonormalizing \( W \) is stable. We have tested LRE-CG with CGS and TSQR (CGS+TSQR) orthonormalization, and it has the same convergence behavior as LRE-CG with MGS (MGS+MGS) orthonormalization (Table 3). Thus, we conclude that MGS, and CGS+TSQR orthonormalizations are stable enough to be used in the LRE-CG method from Algorithm 5.

In Table 5, we compare the convergence behavior of MSDO-CG with MGS A-orthonormalization, LRE-CG with MGS orthonormalization, Coop-CG and MSD-CG with respect to CG for several matrices with different number of partitions \( t = 2, 4, 8, 16, 32, 64 \). The MSDO-CG, COOP-CG and LRE-CG have better convergence than CG, and LRE-CG has the best convergence. The MSD-CG converges, but requires more iterations than CG, three times more iterations for the matrices Sky2D, Sky3D, Ani3D, and ELASTICITY3D. As for Coop-CG, which starts with \( tk \) different initial guesses and solves two systems of fixed size \( t \times t \), its convergence is slightly better than MSDO-CG for the matrices POISSON2D, NH2D, and ELASTICITY3D. But it requires much more iterations than both MSDO-CG and LRE-CG for the other matrices (Sky2D, Sky3D, Ani3D). Moreover, the results may vary depending on the \( t \) initial guesses that are used for the different matrices.

Table 5: The memory requirements in words, for MSDO-CG and LRE-CG with a varying number of partitions (\( Pa \)) or \( t \) and for different matrices.

<table>
<thead>
<tr>
<th>Pa</th>
<th>Iter</th>
<th>Mem</th>
<th>Iter</th>
<th>Mem</th>
<th>Iter</th>
<th>Mem</th>
<th>Iter</th>
<th>Mem</th>
<th>Iter</th>
<th>Mem</th>
<th>Iter</th>
<th>Mem</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSDO-CG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>200</td>
<td>4040004</td>
<td>256</td>
<td>5160004</td>
<td>1559</td>
<td>31220004</td>
<td>610</td>
<td>9792004</td>
<td>893</td>
<td>14320004</td>
<td>764</td>
<td>17239060</td>
</tr>
<tr>
<td>4</td>
<td>167</td>
<td>6740016</td>
<td>208</td>
<td>8380016</td>
<td>917</td>
<td>36740016</td>
<td>420</td>
<td>13488016</td>
<td>749</td>
<td>24016016</td>
<td>622</td>
<td>28064999</td>
</tr>
<tr>
<td>8</td>
<td>139</td>
<td>11220064</td>
<td>169</td>
<td>13620064</td>
<td>532</td>
<td>42660064</td>
<td>228</td>
<td>14672064</td>
<td>498</td>
<td>31952064</td>
<td>742</td>
<td>42609322</td>
</tr>
<tr>
<td>16</td>
<td>121</td>
<td>19540256</td>
<td>138</td>
<td>22260256</td>
<td>307</td>
<td>49300256</td>
<td>134</td>
<td>17296256</td>
<td>328</td>
<td>42128256</td>
<td>343</td>
<td>61959274</td>
</tr>
<tr>
<td>32</td>
<td>94</td>
<td>30421024</td>
<td>107</td>
<td>34581024</td>
<td>178</td>
<td>57301024</td>
<td>87</td>
<td>22545024</td>
<td>192</td>
<td>49425024</td>
<td>234</td>
<td>84646090</td>
</tr>
<tr>
<td>64</td>
<td>69</td>
<td>44824096</td>
<td>77</td>
<td>49944096</td>
<td>126</td>
<td>81304096</td>
<td>53</td>
<td>27668096</td>
<td>122</td>
<td>62996096</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| LRE-CG | | | | | | | | | | | | |
| 2 | 193  | 4028996 | 245  | 5160100 | 1415 | 36328900 | 557  | 10168996 | 875  | 17078996 | 634  | 15899134 |
| 4 | 155  | 6514544 | 188  | 8105504 | 757  | 39468784 | 373  | 14178064 | 673  | 28798864 | 480  | 25314666 |
| 8 | 123  | 10828256 | 149  | 13360864 | 398  | 41997856 | 211  | 16369344 | 449  | 41654464 | 334  | 37230106 |
| 16 | 95   | 17530400 | 112  | 21515264 | 220  | 47610400 | 119  | 18873216 | 253  | 48786304 | 235  | 56471386 |
| 32 | 70   | 27437600 | 82   | 33145376 | 126  | 56597024 | 69   | 22555264 | 148  | 60333696 | 170  | 90832426 |
| 64 | 52   | 44375584 | 60   | 53165600 | 75   | 71060000 | 43   | 29605504 | 92   | 81788544 | 117  | 140355114 |

For the tested matrices, LRE-CG has slightly better convergence than MSDO-CG, since it uses the whole basis to define the new approximate solution rather than \( t \) search directions. For the matrices
POISSON2D and Nh2D, LRE-CG and MSDO-CG have almost the same convergence as CG for \( t = 2 \), and then as \( t \) is doubled the iterations needed for convergence is decreased by 20% to 30%. For \( t = 2 \), LRE-CG requires 35% and 40% less iterations than CG for the matrices ELASTICITY3D and SKY3D respectively. And as \( t \) is doubled the number of iterations needed for convergence is decreased by 25% to 30%, and 32% to 45% respectively. For \( t = 2 \), LRE-CG requires 60% and 80% less iterations than CG for the matrices SKY2D and ANt3D respectively. And as \( t \) is doubled, the number of iterations needed for convergence is decreased by 45% to 50% and 25% to 40% respectively.

As it is clear from the convergence tests, by doubling \( t \) the number of iterations needed for convergence is not always reduced by 50% for all the matrices. However, as shown in the previous sections the memory requirements for MSDO-CG and LRE-CG, except for the matrix \( A \), is \((tk_c + 2)n + (tk_c)^2\) and \((tk_c + t + 2)n + t^2\) respectively, where \( n \) is the size of the matrix, \( k_c \) is the number of computed iterations. Thus by doubling \( t \), the memory requirements for MSDO-CG and LRE-CG for performing \( k \) iterations is at least doubled. But, when \( t \) is doubled \( k_c \) is decreased. Thus, the memory requirements increase and at most double, when \( t \) is doubled as shown in Table 5. Hence, \( t \) should be relatively small depending on the size of the matrix, on the performed iterations and on the available memory.

5 Parallel Model and Expected Performance

In this section we describe the parallelization of the MSDO-CG method (section 5.1) and the LRE-CG method (section 5.2) with computed flops, number of messages and words sent and the estimated parallel runtime. The estimated time for computing \( z \) flops is \( \gamma_c z \), where \( \gamma_c \) is the inverse floating-point rate, also called the floating-point throughput (seconds per floating-point operation). The estimated time for sending a messages of size \( k \) is \( \alpha_c + \beta_c k \), where \( \alpha_c \) is the latency (with units of seconds) and \( \beta_c \) is the inverse bandwidth (seconds per word). Hence, the estimated runtime of an algorithm with a total of \( z \) computed flops and \( s \) sent messages each of size \( k \) is the sum of their corresponding estimated times \( \gamma_c z + \alpha_c s + \beta_c s \).

For simplicity, we assume that the algorithms are executed on a distributed memory machine formed by \( t \) processors, where \( t \) corresponds to the number of vectors computed at each iteration. We partition the graph of \( A \) into \( t \) subdomains using \( k \)-way partitioning or another graph partitioning. We denote by \( \delta_i \), for \( i = 1, 2, \ldots, t \) the subsets of indices obtained from the partitioning. That is \( \delta_i \cap \delta_h = \emptyset \) for all \( i \neq h \), \( \cup_{i=1}^{t} \delta_i = \{1, 2, 3, \ldots, n\} \), and \( |\delta_i| \approx \frac{n}{t} \). Then each processor \( i \) is assigned the \( \frac{n}{t} \times n \) rowwise part of the matrix \( A (A(\delta_i, :) = A(:, \delta_i) \) since \( A \) is SPD), the \( \frac{n}{t} \times 1 \) rowwise part of the vector \( b (b(\delta_i)) \), and the vector \( x_0(\delta_i) \), where \( \delta_i = \text{Adjacent}(G(A), \delta_i) \) is the adjacent of \( \delta_i \) in the graph of \( A \). Processor \( i \) computes \( x_k(\delta_i) \).

However, for performance reasons and due to the multicore nature of most architectures, it is possible to use a number of processors greater than \( t \), preferably a multiple of \( t \). In this case, we start by partitioning the graph of \( A \) into \( t \) subdomains using \( k \)-way partitioning or another graph partitioning, where \( \delta_i \) for \( i = 1, 2, \ldots, t \) are the subsets of indices obtained from the partitioning. This partitioning is used to define the \( T(\cdot) \) operator and eventually the enlarged Krylov subspace. Assuming that we have \( ct \) processors, then every \( c \) processors are assigned an \( \frac{n}{t} \times n \) rowwise part of the matrix \( A, \ A((\delta_i, :) = \frac{n}{t} \times 1 \) rowwise part of the vector \( b (b(\delta_i)) \) and the vector \( x_0(\delta_i) \), and should output \( x_k(\delta_i) \). In other words, we partition each of our \( t \) subdomains into \( c \) non-overlapping subdomains to obtain a total of \( ct \) subdomains with set of indices \( \delta_{i,j} \), where \( i = 1, 2, \ldots, t, \ j = 1, 2, \ldots, c \), and \( \delta_{i,j} = \cup_{j=1}^{c} \delta_{i,j} \). Then, in Algorithms 4 and 5, \( \log(t) \) is replaced by \( \log(ct) \), and \( \frac{n}{t} \) is replaced by \( \frac{n}{ct} \).
5.1 MSDO-CG

In this section we describe the parallelization of the MSDO-CG algorithm and we estimate its runtime in terms of flops, number of messages, and words sent. As mentioned in section 2, MGS, CGS2+A-CholQR, and CGS2+Pre-CholQR A-orthonormalizations are numerically the most stable and allow the convergence of MSDO-CG for the matrices in our test set. As discussed in Appendix B, the most parallelizable versions of MGS, Algorithms 14 and 15, require sending \((tk + 1) \log(t)\) and \(2(t - 1) \log(t)\) messages respectively. Whereas CGS2, Algorithm 22, requires sending \(4 \log(t)\) messages. On the other hand, Algorithm 25 of A-CholQR requires sending \(\log(t)\) messages, and Pre-CholQR Algorithm 27 requires sending \(3 \log(t)\) messages. The CGS2-A-CholQR and CGS2+Pre-CholQR A-orthonormalizations can be called communication avoiding since they require sending \(5 \log(t)\) and \(7 \log(t)\) messages respectively, unlike the MGS A-orthonormalization. Since both methods are stable and CGS2+A-CholQR requires less communication, we present the Parallel MSDO-CG with CGS2+A-CholQR A-orthonormalization in Algorithm 4.

### Algorithm 4 Parallel MSDO-CG with CGS2+A-CholQR A-orthonormalization

<table>
<thead>
<tr>
<th>Input:</th>
<th>(\delta_i), the set of indices assigned to processor (i)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input:</td>
<td>(A(\delta_i,:)), the (\frac{n}{p} \times n) row part of (A)</td>
</tr>
<tr>
<td>Input:</td>
<td>(b(\delta_i),) the (\frac{n}{p} \times 1) row part of (b); (x(\delta_i),) the (</td>
</tr>
<tr>
<td>Input:</td>
<td>(\epsilon), the stopping tolerance; (k_{max}), the maximum allowed iterations</td>
</tr>
<tr>
<td>Output:</td>
<td>(x_k(\delta_i),) the row part of the approximate solution of (Ax = b)</td>
</tr>
</tbody>
</table>

1. for each processor \(i = 1 : t\) in parallel do
2. Processor \(i\) computes \(r(\delta_i) = b(\delta_i) - A(\delta_i, \delta_i)x(\delta_i)\), and let \(k = 1\) \(\gamma_c(2\frac{nnz}{t})\)
3. Processor \(i\) computes \(r(\delta_i)^T r(\delta_i)\) and receives the full \(\rho = ||r||^2\) \(\gamma_c(2\frac{p}{t})\)
4. Processor \(i\) sends \(P_1(\delta_i,:)[T(\delta_i)](\delta_i,:) = [0, \ldots, 0, r(\delta_i), 0, \ldots, 0]\) to its \(m_{MB}\) neighboring processors and receives from them the corresponding blocks to obtain \(P_1(\delta_i,:).\) Then it computes \(W_1(\delta_i) = A(\delta_i, \delta_i)P_1(\delta_i,:).\) \(\gamma_c(2\frac{nnz}{t})\)
5. Call A-CholQR algorithm 25 to A-orthonormalize \(P_1\) \(\gamma_c4nt + (\alpha_c + \beta_c t^2)\log(t)\)
6. while \((\sqrt{t} > \epsilon ||b||_2\) and \(k < k_{max}\) ) do
7. Processor \(i\) computes \(P_k(\delta_i,:)r(\delta_i)\) and receives the full \(\alpha = P_k^T r\) via an all reduce \(\gamma_c(2\frac{p}{t} - 1)t\) \(+ (\alpha_c + t\beta_c)\log(t)\)
8. Processor \(i\) computes \(x(\delta_i) = x(\delta_i) + P_k(\delta_i,:)\alpha\) \(4\gamma_c\frac{t}{p}\)
9. Processor \(i\) computes \(r(\delta_i) = r(\delta_i) - W_k(\delta_i,:)\alpha\) \(\gamma_c(2\frac{p}{t} - 1)\)
10. Processor \(i\) computes \(r(\delta_i)^T r(\delta_i)\) and receives \(\rho = ||r||^2\) \(\gamma_c(2\frac{p}{t} - 1)t\)
11. Processor \(i\) computes \(\beta = -W_k(\delta_i,:)r(\delta_i)\) and receives the full \(\gamma_c(2\frac{p}{t} - 1)\)
12. Processor \(i\) sends \([T(\delta_i)](\delta_i,:) + P_k(\delta_i,:)\text{diag}(\beta)\) \(2\gamma_c\frac{t}{p}\)
13. Processor \(i\) computes \(W_{k+1}(\delta_i,:) = Z(\delta_i,:) + W_k(\delta_i,:)\text{diag}(\beta)\) \(2\gamma_c\frac{t}{p}\)
14. Call CGS2 Algorithm 22 to A-orthonormalize \(P_{k+1}\) against \(P_1\) \(12\gamma_c nt\kappa\)
15. Call A-CholQR Algorithm 25 to A-orthonormalize \(P_{k+1}\) \(\gamma_c4nt + (\alpha_c + \beta_c t^2)\log(t)\)
16. \(k = k + 1\)
17. end while
18. end for
In Algorithm 3, we have two types of communication. The first is an “all reduce” communication that requires synchronization between all the processors and is equivalent to \( \log(t) \) messages, each of the same size (refer to [26]). For example, in line 10 of Algorithm 4, the “all reduce” is equivalent to \( \log(t) \) messages each of size \( t \) words, since \( \beta \) is a vector of size \( t \).

The second type of communication is a point-to-point communication between each processor \( i \) and its \( m_i \) neighboring processors for computing a matrix - block of vectors multiplication, specifically \( A[T(r)] \). We denote by \( m_{MB} = \max(m_i \mid i = 1, 2, \ldots, t) \) the largest number of neighboring processors, where \( m_i \leq m_{MB} \leq (t - 1) \) for all \( i \). Note that processor \( i \) has to compute \( A(\delta_i, \delta_i)\hat{V}(r)\hat{W}(r) \) where \( \delta_i = \text{Adjacent}(G(A), \delta_i) \). Then, the neighboring processors of a given processor \( i \) are defined as all the processors \( j \) from which processor \( i \) needs some rows of \( [T(r)] \) to compute its part of \( A[T(r)] \). In other words, neighboring processors are all the processors \( j \) for which \( \delta_i \cap \delta_j \neq \emptyset \). Moreover, \( [T(r)](\delta_i, :) \) is all zeros except for the \( \hat{V}^{(:, u)} \) column which is equal to \( r(\delta_i) \). Thus, processor \( i \) sends \( r(\delta_i) \) of size \( \frac{t}{2} \times 1 \) to its neighboring processors once \( r(\delta_i) \) is computed at step 8. Since \( r(\delta_i) \) is used in the computation at step 12, this communication is overlapped with the computations from step 9 to 11. Simultaneously, processor \( i \) receives \( r(\delta_j) \) from all its neighboring processors \( j \) for \( j = 1, 2, \ldots, m_i \). Then it computes \( A(\delta_i, \delta_i)[T(r)](\delta_i, :) \) by performing approximately \( \frac{m_{MB} - n}{t} \) flops.

In summary, without the A-orthonormalization at steps 14 and 15, the estimated time of \( k_c \) iterations of Algorithm 4, where we ignore lower order terms, is

\[
\gamma_c(11n + 2n\frac{m_{MB}}{t})k_c + \alpha_c(2\log(t) + m_{MB})k_c + \beta_c(\frac{n}{t}m_{MB} + 2\log(t))k_c.
\]

At iteration \( k \), the CGS2+A-CholQR A-orthonormalization requires sending \( 5\log(t) \) messages with \( (t + 2k + 2)\log(t) \) words and performing approximately \( 12ntk + 4nt + 6n \) flops. After \( k_c \) iterations the estimated time for the A-orthonormalization is

\[
\gamma_c((12ntk_c + 16nt + 6n)k_c + \alpha_c(5\log(t))k_c + \beta_c(t + 2(k_c+1) + 2)\log(t))k_c.
\]

Thus, the estimated time of \( k_c \) iterations of Algorithm 4 is

\[
\text{Time}_{MGSDT}(k_c) \approx \gamma_c(2n\frac{m_{MB}}{t} + 12ntk_c + 10nt + 17n)k_c + \alpha_c(7\log(t) + m_{MB})k_c + \beta_c(\frac{2}{t}m_{MB} + t^2k_c\log(t))k_c.
\]

### 5.2 LRE-CG

In this section we describe the parallelization of the LRE-CG algorithm and we estimate its runtime in terms of flops, number of messages, and words sent. As mentioned in section 3, MGS and the CGS+TSQR orthonormalizations are numerically the most stable and allow the convergence of LRE-CG for the matrices in our test set. The parallel version of MGS orthonormalization is similar to that of the A-orthonormalization discussed in Appendix B. Algorithms 14 and 15 and requires sending \( (tk + 1)\log(t) \) and \( 2(t - 1)\log(t) \) messages respectively. Whereas the CGS orthonormalization can be parallelized in a block format like Algorithm 18 and requires sending \( 2\log(t) \) messages. On the other hand, the TSQR orthonormalization using binary trees as discussed in 14 requires sending \( \log(t) \) messages. The combination of BCGS and TSQR was discussed in 21 and it requires sending only \( 3\log(t) \) messages as compared to the \( (tk + 2t - 1)\log(t) \) messages of MGS. We present the Parallel LRE-CG with BCGS+TSQR orthonormalization in Algorithm 5.

In Algorithm 5 there are two types of communication, similarly to Algorithm 4. The first type of communication is a point-to-point communication between each processor \( i \) and its \( m_i \) neighboring processors for computing the matrix - block of vectors multiplication \( Z = AW \) in line 7. We denote by \( m_{MB} = \max(m_i \mid i = 1, 2, \ldots, t) \) the largest number of neighboring processors, where \( m_i \leq m_{MB} \leq (t - 1) \) for all \( i \). Note that processor \( i \) has to compute \( A(\delta_i, \delta_i)W(\delta_i, :) \), where \( \delta_i = \text{Adjacent}(G(A), \delta_i) \). Then, the neighboring processors of a given processor \( i \) are defined as all the processors \( j \) from which processor \( i \) needs some rows of \( W \) to compute its part of \( A^\prime \). In other words, neighboring processors are all the processors \( j \) for which \( \delta_i \cap \delta_j \neq \emptyset \). Note that for the first iteration, \( W(\delta_i, :) = [T(r)](\delta_i, :) \) is all zeros except for the \( \hat{V}^{(:, u)} \) column which is equal to \( r(\delta_i) \). Thus, processor \( i \) sends \( r(\delta_i) \) of size \( \frac{t}{2} \times 1 \) to
**Algorithm 5** Parallel LRE-CG with BCGS+TSQR Algorithm

<table>
<thead>
<tr>
<th>Line</th>
<th>Description</th>
<th>Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>for each processor $i = 1 : t$ in parallel do</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Processor $i$ computes $r(\delta_i) = b(\delta_i) - A(\delta_i, \overline{\partial}_i)x_0(\overline{\partial}_i)$</td>
<td>$\gamma_c\frac{2nnz}{t}$</td>
</tr>
<tr>
<td>3</td>
<td>Processor $i$ computes $r(\delta_i)^T r(\delta_i)$ and receives the full $\rho =</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Let $Q(\delta_i, :) = W(\delta_i, :) = [T(r)](\delta_i, :)$, and normalise its vectors</td>
<td>$(2\frac{t}{\beta_c})$</td>
</tr>
<tr>
<td>5</td>
<td>Processor $i$ sends $W(\delta_i, :) = [T(r)](\delta_i, :)$ to its $m_i$ neighboring processors and receives from them the corresponding blocks</td>
<td>$\alpha_c m_{MB}$</td>
</tr>
<tr>
<td></td>
<td>to obtain $W(\overline{\partial}_i, :)$</td>
<td>$\beta_c \frac{n}{t} m_{MB}$</td>
</tr>
<tr>
<td>6</td>
<td>while $(\sqrt{pk-1} &gt; \epsilon |b|<em>2$ and $k &lt; k</em>{\text{max}}$) do</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Processor $i$ computes $Z(\delta_i, :) = A(\delta_i, \overline{\partial}_i)W(\overline{\partial}_i, :)$</td>
<td>$2\gamma_c \frac{nnz}{t} t$</td>
</tr>
<tr>
<td>8</td>
<td>Processor $i$ computes $W(\delta_i, :)^T Z(\delta_i, :)$ and receives $E = W^T Z$ via an “all reduce”</td>
<td>$(\alpha_c + t^2\beta_c)\log(t)$</td>
</tr>
<tr>
<td>9</td>
<td>if $k = 1$ then</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>$D = Z$ and $G = E$</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>else</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Processor $i$ computes $Q(\delta_i, 1 : t(k-1))^T Z(\delta_i, :)$ and receives $F = Q(:, 1 : t(k-1))^T Z$ via an “all reduce”</td>
<td>$\gamma_c 2\frac{t^2}{\beta_c} k$</td>
</tr>
<tr>
<td>13</td>
<td>$G = \begin{pmatrix} F^T &amp; E \end{pmatrix}$ and $D = [D \quad Z]$</td>
<td>$+\alpha_c \log(t)$</td>
</tr>
<tr>
<td></td>
<td>end if</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Processor $i$ computes $Q(\delta_i, :)^T r(\delta_i)$ and receives $Q^T r$ via an “all reduce”</td>
<td>$\gamma_c 2\frac{t^2}{\beta_c} k t k$</td>
</tr>
<tr>
<td></td>
<td>$\alpha = G^{-1}(Q^T r)$</td>
<td>$+(\alpha_c + k^2\beta_c)\log(t)$</td>
</tr>
<tr>
<td>17</td>
<td>Processor $i$ computes $x(\delta_i) = x(\delta_i) + Q(\delta_i, :)\alpha$</td>
<td>$\gamma_c (2tk^2 \frac{t}{\beta_c})$</td>
</tr>
<tr>
<td>18</td>
<td>Processor $i$ computes $r(\delta_i) = r(\delta_i) - D(\delta_i, :)\alpha$</td>
<td>$\gamma_c (2tk^2 \frac{t}{\beta_c})$</td>
</tr>
<tr>
<td>19</td>
<td>Processor $i$ computes $r(\delta_i)^T r(\delta_i)$ and receives $\rho_k =</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Let $W = Z$ and orthogonalise the columns of $W$ against those of $Q$ using BCGS</td>
<td>$\gamma_c \text{Antk} + 2\alpha_c \log(t)$</td>
</tr>
<tr>
<td></td>
<td>$+t^2k^2\beta_c \log(t)$</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Orthonormalise the vectors of $W$ using TSQR and let $Q = [Q \quad W]$</td>
<td>$+(\alpha_c + \beta_c \frac{n}{t}) \log(t)$</td>
</tr>
<tr>
<td>22</td>
<td>Processor $i$ sends $W(\delta_i, :)$ to its $m_i$ neighboring processors and receives from them the corresponding blocks to obtain $W(\overline{\partial}_i, :)$</td>
<td>$+\alpha_c m_{MB}$</td>
</tr>
<tr>
<td></td>
<td>$+\beta_c \frac{n}{t} m_{MB}$</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>$k = k+1$</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>end for</td>
<td></td>
</tr>
</tbody>
</table>

its neighboring processors once $r(\delta_i)$ is computed. However, for the next iteration the $W$ is no longer sparse, therefore $W(\overline{\partial}_i, :)$ of size $\frac{p}{t} \times t$ is sent.

The second type of communication is an “all reduce” that requires synchronization between all the processors, and it is equivalent to $\log(t)$ messages each of the same size (refer to [26]). For example, in lines 3 and 19 of Algorithm 5, the “all reduce” is equivalent to $\log(t)$ messages each of size 1 word.
As mentioned, this communication can be overlapped with the next computation. The reception of $E = W^tZ$, $F = Q^1, 1 : t(k - 1)^tZ$ and $Q^tr$ via an “all reduce” in lines 8, 12 and 15 of Algorithm 5 is equivalent to $log(t)$ messages each of size $t^2$ words, $log(t)$ messages each of size $t^2(k - 1)$ words, and $log(t)$ messages each of size $tk$ words respectively. However, the three computations are independent. Thus, each processor can compute its part of the three aforementioned computations and then receive the full matrices and vectors via $log(t)$ messages each of size $kt(t + 1)$ words, assuming that it is possible to send $t^2k$ words in one message. Another alternative is to compute $Q(\delta_1, 1 : t(k - 1)^tZ(\delta_1,:))$ in several steps and overlap the communication with the next computation. The number of steps depends on the machine’s architecture and on the values of $t$ and $k$.

In Algorithm 5 we show the estimated time for each computation and communication, where $Time_{\alpha}(tk)$ is the estimated time for solving the $tk \times tk$ $\alpha$ system in line 16. At the $k^{th}$ iteration of Algorithm 5, the total flops, except for the $\alpha$ system, is $2mnt + (6nt - 2t^2 + 6n - t)k + 2nt + \frac{2}{3}t^3log(t) + 2n + 2\frac{t}{3} - 1$. And $4log(t) + m_{MB}$ messages are sent with $((2t^2 + t)k + \frac{t^2}{2} + tlog(t) + m_{MB})$ words.

Then, by ignoring the lower order terms the estimated time of $k_c$ iterations of Algorithm 5 where $t > 1, k_c > 1$, is

$$Time_{\alpha}(k_c) \approx \gamma_c(2mnt + 3ntk_c + \frac{2}{3}t^3log(t))k_c + \alpha_c(4log(t) + m_{MB})k_c + \beta_c(t^2log(t) + \frac{2}{3}t^2log(t) + m_{MB})k_c + \sum_{k=1}^{n} Time_{\alpha}(tk)$$

### 6 Preconditioned enlarged Krylov subspace methods

After introducing the enlarged Krylov subspace methods and proving, theoretically and numerically, that these methods converge, we describe the preconditioned enlarged Krylov methods. A system $Ax = b$ can be left, right, or split preconditioned. In the case of conjugate gradient methods, the matrix $A$ is symmetric positive definite (SPD). Hence, the preconditioned matrix should also be SPD. For left and right preconditioning, it is not easy to find some matrix $M$ such that $M^{-1}A$ or $AM^{-1}$ is SPD. But assuming that $M = LL^t$, then the split preconditioned matrix $L^{-1}AL^{-t}$ is SPD with $L^{-t} = (L^t)^{-1}$.

Given an $n \times n$ SPD matrix $A$, $n \times 1$ vector $b$ and some preconditioner $M = LL^t$, then the split preconditioned enlarged Krylov subspace corresponding to the system $L^{-1}AL^{-t}y = L^{-1}b$ with $y = L^tx$ and $M = LL^t$, is defined by

$$\mathcal{K}_{L,k}(L^{-1}AL^{-t}, r_0) = \text{span}\{T(r_0), L^{-1}AL^{-t}T(r_0), (L^{-1}AL^{-t})^2T(r_0), ... , (L^{-1}AL^{-t})^{k-1}T(r_0)\}$$

$$= \text{span}\{T_1(r_0), T_2(r_0), ... , T_k(r_0), L^{-1}AL^{-t}T_1(r_0), L^{-1}AL^{-t}T_2(r_0), ... , L^{-1}AL^{-t}T_k(r_0), ... , (L^{-1}AL^{-t})^{k-1}T_1(r_0), ... , (L^{-1}AL^{-t})^{k-1}T_k(r_0)\},$$

where $r_0 = L^{-1}(b - AL^{-t}y_0) = L^{-1}(b - Ax_0), y_0 = L^tx_0$, and $x_0$ is the initial guess.

Consequently, the split preconditioned enlarged conjugate gradient methods are defined by the orthogonality condition and the subspace condition associated with preconditioned enlarged Krylov subspace. For example, given a split preconditioned system $L^{-1}AL^{-t}y = L^{-1}b$ with $y = L^tx$ and $M = LL^t$, the the enlarged CG Krylov projection methods are defined by $y_k \in y_0 + \mathcal{K}_{L,k}(L^{-1}AL^{-t}, r_0)$ (the subspace condition), and $r_k \perp \mathcal{K}_{L,k}(L^{-1}AL^{-t}, r_0)$ (orthogonality condition), where $r_k = L^{-1}(b - AL^{-t}y_k) = L^{-1}(b - Ax_k)$. Assuming $\hat{A} = L^{-1}AL^{-t}, \hat{b} = L^{-1}b_2$ and $\hat{z} = y$, then all the theorems and properties discussed in section 3.1 are valid for the system $\hat{A}\hat{x} = \hat{b}$.

Given an SPD matrix $M$, then the Cholesky factorization $M = LL^t$ can be used for split preconditioning the system $Ax = b$, where the matrix $L^{-1}AL^{-t}$ is SPD. As the Cholesky factorization of an $n \times n$ matrix can be expensive, another alternative is to use block Jacobi preconditioner with Cholesky factorization of the diagonal blocks. A four blocks Jacobi preconditioner with Cholesky decomposition
has the following form.

\[
M = \begin{pmatrix}
M_{1,1} & 0 & 0 & 0 \\
0 & M_{2,2} & 0 & 0 \\
0 & 0 & M_{3,3} & 0 \\
0 & 0 & 0 & M_{4,4}
\end{pmatrix} = \begin{pmatrix}
L_{1,1} & 0 & 0 & 0 \\
0 & L_{2,2} & 0 & 0 \\
0 & 0 & L_{3,3} & 0 \\
0 & 0 & 0 & L_{4,4}
\end{pmatrix}
\begin{pmatrix}
L'_{1,1} & 0 & 0 & 0 \\
0 & L'_{2,2} & 0 & 0 \\
0 & 0 & L'_{3,3} & 0 \\
0 & 0 & 0 & L'_{4,4}
\end{pmatrix}
\]

We present in Algorithm 6 the split preconditioned MSDO-CG with CGS2+ \( \hat{A} \)-CholQR \( \hat{A} \)-orthonormalization of the system \( \hat{A}x = \hat{b} \), where \( \hat{A} = L^{-1}AL^{-t} \), \( \hat{b} = L^{-1}b \), \( \hat{x} = y \), and \( y = L'x \). We omit the \( W \) recursion due to numerical errors since \( W = \hat{A}P = L^{-1}AL^{-t}P \) consists of performing backward and forward substitution in addition to the matrix vector multiplication. Thus, we use a version of the CGS2 \( \hat{A} \)-orthonormalization (Algorithm 23) that computes \( W = \hat{A}P = L^{-1}AL^{-t}P \) and outputs it. As for the \( \hat{A} \)-CholQR, by assuming that \( W = \hat{A}P = L^{-1}AL^{-t}P \) is computed, then we can use Algorithm 26 with input \( W = \hat{A}P \). The additional cost of preconditioning is computing at each iteration \( k \), four times \( W_{k+1} = L^{-1}AL^{-t}P_{k+1} \) which is equivalent to a backward and forward substitution with \( t \) right hand sides and a matrix vector multiplication. The difference between the preconditioned and unpreconditioned MSDO-CG is in the \( \hat{A} \)-orthonormalization, the computation of \( W \), and the backward substitution \( L'x = y \). Note that the split preconditioned MSDO-CG with MGS \( \hat{A} \)-orthonormalization did not converge. This might be due to numerical errors in solving the \( tk \) backward and forward substitutions in the MGS \( \hat{A} \)-orthonormalization. However, the MSDO-CG with CGS2+\( \hat{A} \)-CholQR \( \hat{A} \)-orthonormalization converges very well as shown in section 6.1.

**Algorithm 6** Split preconditioned MSDO-CG algorithm with CGS2+\( \hat{A} \)-CholQR \( \hat{A} \)-orthonormalization

Input: \( A \), the \( n \times n \) symmetric positive definite matrix; \( L \), \( n \times n \) split preconditioner

Input: \( b \), the \( n \times 1 \) right-hand side; \( x_0 \), the initial guess or iterate

Input: \( \epsilon \), the stopping tolerance; \( k_{\text{max}} \), the maximum allowed iterations

Output: \( x \), the approximate solution of \( Ax = b \) by solving for \( L^{-1}AL^{-t}y = L^{-1}b \) and \( y = L'x \)

1: \( y = L'x_0 \), \( r = L^{-1}(b - Ax_0) \), \( \rho = ||r||^2_2 \), \( nb = ||L^{-1}b||_2 \), \( k = 1 \)

2: Let \( P_1 = T(r) \) and \( L^{-1}AL^{-t} \)-orthonormalize it using Algorithm 26 which outputs \( W_1 = L^{-1}AL^{-t}P_1 \)

3: while \( (\sqrt{\beta} > \epsilon(nb) \) and \( k < k_{\text{max}} ) \) do

4: \( \alpha = (P_1^*W_k)^{-1}(P_1^*r) = P_k^*r \)

5: \( y = y + P_k \alpha \) and \( r = r - W_k \alpha \)

6: \( \rho = ||r||^2_2 \)

7: \( \beta = -(P_1^*W_k)^{-1}(W_k^*r) = -W_k^*r \)

8: \( P_{k+1} = T(r) + P_k \text{diag}(\beta) \)

9: \( L^{-1}AL^{-t} \)-orthonormalize \( P_{k+1} \) against all \( P_i \)’s for \( i \leq k \) using Algorithm 23

10: \( L^{-1}AL^{-t} \)-orthonormalize \( P_{k+1} \) using Algorithm 26 which outputs \( W_{k+1} = L^{-1}AL^{-t}P_{k+1} \)

11: \( k = k + 1 \)

12: end while

13: Solve \( L'x = y \)

In Algorithm 7 we present the split preconditioned LRE-CG algorithm with BCGS+TSQR orthonormalization of the system \( \hat{A}x = \hat{b} \), where \( \hat{A} = L^{-1}AL^{-t} \), \( \hat{b} = L^{-1}b \), \( \hat{x} = y \), and \( y = L'x \). At first glance, it might appear to the reader that the additional cost at iteration \( k \) in Algorithm 7 is solving a forward and backward substitution with \( tk \) right hand sides \( (L^{-1}AL^{-t}Q) \) and a forward and backward substitution with \( t \) right hand sides \( (L^{-1}AL^{-t}W) \). However, by taking a quick look at Algorithm 8 it is clear that the additional cost of preconditioning at iteration \( k \) is solving only a forward and backward substitution.
Algorithm 7 Split preconditioned LRE-CG Algorithm with BCGS+TSQR orthonormalization

\begin{itemize}
\item \textbf{Input:} $A$, the $n \times n$ symmetric positive definite matrix; $L$, $n \times n$ split preconditioner
\item \textbf{Input:} $b$, the $n \times 1$ right-hand side; $x_0$, the initial guess or iterate
\item \textbf{Input:} $\epsilon$, the stopping tolerance; $k_{max}$, the maximum allowed iterations
\item \textbf{Output:} $x$, the approximate solution of $Ax = b$ by solving for $L^{-1}A L^{-\dagger} y = L^{-1}b$ and $y = L^t x$
\end{itemize}

1: $y = L^t x_0$, $r = L^{-1}(b - A x_0)$, $\rho_0 = ||r||_2^2$, $nb = ||L^{-1}b||_2, k = 1$
2: Let $W = T(r_0)$, normalise its vectors and then let $Q = W$
3: \textbf{while} ($\sqrt{\rho_{k-1}} > \epsilon \langle nb \rangle$ and $k < k_{max}$) \textbf{do}
4: \hspace{1em} $\alpha = (Q^t L^{-1} A L^{-\dagger} Q)^{-1} (Q^t r)$
5: \hspace{1em} $y = y + Q \alpha$
6: \hspace{1em} $r = r - L^{-1} A L^{-\dagger} Q \alpha$ and $\rho_k = ||r||_2^2$
7: \hspace{1em} Let $W = L^{-1} A L^{-\dagger} W$
8: \hspace{1em} Orthonormalise the vectors of $W$ against the vectors of $Q$
9: \hspace{1em} Orthonormalise the vectors of $W$ and let $Q = [Q \ W]$ and $k = k + 1$
10: \textbf{end while}
11: Solve $L^t x = y$

Algorithm 8 Split preconditioned LRE-CG Pseudo Code with BCGS+TSQR orthonormalization

\begin{itemize}
\item \textbf{Input:} $A$, the $n \times n$ symmetric positive definite matrix
\item \textbf{Input:} $b$, the $n \times 1$ right-hand side; $x_0$, the initial guess or iterate
\item \textbf{Input:} $\epsilon$, the stopping tolerance; $k_{max}$, the maximum allowed iterations
\item \textbf{Output:} $x_k$, the approximate solution of the system $Ax = b$
\end{itemize}

1: $y = L^t x_0$, $r = L^{-1}(b - A x_0)$, $\rho_0 = ||r||_2^2$, $nb = ||L^{-1}b||_2, k = 1$
2: Let $W = T(r_0)$, normalise its vectors and then let $Q = W$
3: \textbf{while} ($\sqrt{\rho_{k-1}} > \epsilon \langle nb \rangle$ and $k < k_{max}$) \textbf{do}
4: \hspace{1em} $Z = L^{-1} A L^{-\dagger} W$, $E = W^t Z$
5: \hspace{1em} if $k == 1$ then
6: \hspace{2em} $D = Z$ and $G = E$
7: \hspace{1em} else
8: \hspace{2em} $D = [D \ Z]$, $F = Q(:,1:t(k-1))^t Z$, and $G = \begin{pmatrix} G \\ F^t \\ E \end{pmatrix}$
9: \hspace{1em} \textbf{end if}
10: \hspace{1em} $\alpha = G^{-1}(Q^t r)$
11: \hspace{1em} $y = y + Q \alpha$
12: \hspace{1em} $r = r - D \alpha$ and $\rho_k = ||r||_2^2$
13: \hspace{1em} Let $W = Z$
14: \hspace{1em} Orthonormalise the columns of $W$ against those of $Q$ using BCGS
15: \hspace{1em} Orthonormalise the vectors of $W$ using TSQR and let $Q = [Q \ W]$ and $k = k + 1$
16: \textbf{end while}
17: Solve $L^t x = y$

with $t$ right hand sides ($L^{-1} A L^{-\dagger} W$). And this is the only difference with the unpreconditioned version in addition to the backward substitution $L^t x = y$ at the end.
6.1 Convergence

We compare the convergence of split preconditioned MSDO-CG with CGS2+CholQR orthonormalization and split preconditioned LRE-CG with CGS+TSQR orthonormalization to CG and split preconditioned CG (PCG). We use Block Jacobi with Cholesky factorization of the block diagonals as a preconditioner.

Table 6: Comparison of the convergence of the split preconditioned CG, MSDO-CG with CGS2+CholQR orthonormalization, and LRE-CG with CGS+TSQR orthonormalization method with varying Block Jacobi preconditioners, with respect to number of partitions $Pa$, with $x_0 = 0$.

<table>
<thead>
<tr>
<th>Split Preconditioned Methods</th>
<th>CG</th>
<th>PCG</th>
<th>MSDO-CG</th>
<th>LRE-CG</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Pa$</td>
<td>Iter</td>
<td>Err</td>
<td>Iter</td>
<td>Err</td>
</tr>
<tr>
<td>2</td>
<td>35</td>
<td>1E-5</td>
<td>30</td>
<td>2E-6</td>
</tr>
<tr>
<td>4</td>
<td>40</td>
<td>1E-5</td>
<td>28</td>
<td>4E-6</td>
</tr>
<tr>
<td>8</td>
<td>44</td>
<td>2E-5</td>
<td>30</td>
<td>6E-6</td>
</tr>
<tr>
<td>16</td>
<td>50</td>
<td>1E-5</td>
<td>28</td>
<td>1E-6</td>
</tr>
<tr>
<td>32</td>
<td>57</td>
<td>2E-5</td>
<td>26</td>
<td>8E-7</td>
</tr>
<tr>
<td>64</td>
<td>66</td>
<td>2E-5</td>
<td>23</td>
<td>1E-6</td>
</tr>
<tr>
<td>$t_{ol} = 10^{-6}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>47</td>
<td>3E-8</td>
<td>37</td>
<td>6E-9</td>
</tr>
<tr>
<td>4</td>
<td>55</td>
<td>7E-8</td>
<td>34</td>
<td>2E-8</td>
</tr>
<tr>
<td>8</td>
<td>65</td>
<td>1E-7</td>
<td>36</td>
<td>1E-8</td>
</tr>
<tr>
<td>16</td>
<td>71</td>
<td>3E-7</td>
<td>33</td>
<td>1E-8</td>
</tr>
<tr>
<td>32</td>
<td>83</td>
<td>1E-7</td>
<td>29</td>
<td>1E-8</td>
</tr>
<tr>
<td>64</td>
<td>88</td>
<td>5E-7</td>
<td>26</td>
<td>5E-9</td>
</tr>
<tr>
<td>$t_{ol} = 10^{-8}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>74</td>
<td>3E-7</td>
<td>40</td>
<td>4E-7</td>
</tr>
<tr>
<td>4</td>
<td>80</td>
<td>2E-6</td>
<td>43</td>
<td>1E-7</td>
</tr>
<tr>
<td>8</td>
<td>144</td>
<td>2E-5</td>
<td>48</td>
<td>3E-7</td>
</tr>
<tr>
<td>16</td>
<td>162</td>
<td>1E-4</td>
<td>46</td>
<td>1E-7</td>
</tr>
<tr>
<td>32</td>
<td>210</td>
<td>3E-4</td>
<td>39</td>
<td>1E-7</td>
</tr>
<tr>
<td>64</td>
<td>260</td>
<td>2E-7</td>
<td>31</td>
<td>8E-8</td>
</tr>
<tr>
<td>$t_{ol} = 10^{-8}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>37</td>
<td>2E-6</td>
<td>24</td>
<td>2E-7</td>
</tr>
<tr>
<td>4</td>
<td>113</td>
<td>2E-5</td>
<td>54</td>
<td>1E-7</td>
</tr>
<tr>
<td>8</td>
<td>120</td>
<td>8E-6</td>
<td>54</td>
<td>7E-8</td>
</tr>
<tr>
<td>16</td>
<td>154</td>
<td>1E-5</td>
<td>49</td>
<td>1E-7</td>
</tr>
<tr>
<td>32</td>
<td>208</td>
<td>1E-5</td>
<td>60</td>
<td>2E-8</td>
</tr>
<tr>
<td>64</td>
<td>213</td>
<td>1E-5</td>
<td>46</td>
<td>1E-8</td>
</tr>
<tr>
<td>$t_{ol} = 10^{-8}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>26</td>
<td>1E-5</td>
<td>31</td>
<td>3E-7</td>
</tr>
<tr>
<td>4</td>
<td>43</td>
<td>4E-6</td>
<td>39</td>
<td>5E-7</td>
</tr>
<tr>
<td>8</td>
<td>47</td>
<td>5E-7</td>
<td>39</td>
<td>6E-7</td>
</tr>
<tr>
<td>16</td>
<td>54</td>
<td>7E-7</td>
<td>43</td>
<td>1E-6</td>
</tr>
<tr>
<td>32</td>
<td>61</td>
<td>2E-7</td>
<td>47</td>
<td>4E-7</td>
</tr>
<tr>
<td>64</td>
<td>66</td>
<td>8E-7</td>
<td>46</td>
<td>2E-7</td>
</tr>
</tbody>
</table>

In table [6] we use a different Block Jacobi preconditioner for the different partitions. First, the graph of $\tilde{A}$ is partitioned into $t$ parts that define the enlarged Krylov subspace using Metis’s kway edge separator where $t = 2, 4, 8, 16, 32, 64$. Then the Block Jacobi preconditioner $M$ is defined as the $t$ diagonal blocks of the permuted matrix $\tilde{A}$. Each of the $t$ blocks is factorized using Cholesky decomposition to obtain a block $L$. The preconditioned LRE-CG converges faster than the preconditioned MSDO-CG and PCG.
for the different configurations. As the number of partitions or the maximum basis vectors added at each iteration is doubled, the Block Jacobi preconditioned CG needs more iterations to converge. However, for the matrices $\text{POISSON2D, NH2D, and SKY2D}$, the number of iteration of the preconditioned LRE-CG and MSDO-CG decreases. As for the matrices $\text{SKY3D and AN13D}$, the number of iterations of LRE-CG increases then decreases back to the same number of iterations for $t = 2$, unlike preconditioned MSDO-CG.

Table 7: Comparison of the convergence of the split preconditioned CG, MSDO-CG with CGS2+CholQR $\hat{A}$-orthonormalization, and LRE-CG with CGS+TSQR orthonormalization method with a fixed Block Jacobi preconditioner, with respect to number of partitions $P_a$, with $x_0 = 0$.

<table>
<thead>
<tr>
<th>$P_a$</th>
<th>Iter</th>
<th>PCG</th>
<th>Iter</th>
<th>MSDO-CG</th>
<th>Iter</th>
<th>LRE-CG</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>195</td>
<td>2E-5</td>
<td>66</td>
<td>2E-5</td>
<td>62</td>
<td>2E-5</td>
</tr>
<tr>
<td>4</td>
<td>54</td>
<td>9E-6</td>
<td>50</td>
<td>8E-6</td>
<td>47</td>
<td>4E-6</td>
</tr>
<tr>
<td>8</td>
<td>39</td>
<td>3E-6</td>
<td>33</td>
<td>1E-6</td>
<td>31</td>
<td>2E-6</td>
</tr>
<tr>
<td>16</td>
<td>25</td>
<td>8E-7</td>
<td>20</td>
<td>3E-7</td>
<td>166</td>
<td>6E-7</td>
</tr>
<tr>
<td>32</td>
<td>64</td>
<td>3E-8</td>
<td>23</td>
<td>4E-9</td>
<td>64</td>
<td>3E-8</td>
</tr>
<tr>
<td>64</td>
<td>16</td>
<td>4E-8</td>
<td>20</td>
<td>2E-7</td>
<td>16</td>
<td>4E-8</td>
</tr>
<tr>
<td>2</td>
<td>259</td>
<td>4E-7</td>
<td>88</td>
<td>5E-7</td>
<td>82</td>
<td>1E-7</td>
</tr>
<tr>
<td>4</td>
<td>67</td>
<td>5E-8</td>
<td>63</td>
<td>5E-8</td>
<td>57</td>
<td>5E-8</td>
</tr>
<tr>
<td>8</td>
<td>46</td>
<td>1E-8</td>
<td>39</td>
<td>2E-8</td>
<td>36</td>
<td>2E-8</td>
</tr>
<tr>
<td>16</td>
<td>28</td>
<td>7E-9</td>
<td>23</td>
<td>4E-9</td>
<td>28</td>
<td>7E-9</td>
</tr>
<tr>
<td>32</td>
<td>223</td>
<td>2E-5</td>
<td>184</td>
<td>6E-7</td>
<td>223</td>
<td>2E-5</td>
</tr>
<tr>
<td>64</td>
<td>152</td>
<td>4E-7</td>
<td>99</td>
<td>3E-7</td>
<td>152</td>
<td>4E-7</td>
</tr>
<tr>
<td>2</td>
<td>5773</td>
<td>5E-04</td>
<td>261</td>
<td>2E-4</td>
<td>261</td>
<td>2E-4</td>
</tr>
<tr>
<td>4</td>
<td>109</td>
<td>2E-7</td>
<td>66</td>
<td>4E-7</td>
<td>109</td>
<td>2E-7</td>
</tr>
<tr>
<td>8</td>
<td>72</td>
<td>1E-7</td>
<td>44</td>
<td>4E-7</td>
<td>72</td>
<td>1E-7</td>
</tr>
<tr>
<td>16</td>
<td>52</td>
<td>5E-8</td>
<td>29</td>
<td>1E-7</td>
<td>52</td>
<td>5E-8</td>
</tr>
<tr>
<td>32</td>
<td>34</td>
<td>7E-8</td>
<td>20</td>
<td>2E-7</td>
<td>34</td>
<td>7E-8</td>
</tr>
<tr>
<td>64</td>
<td>233</td>
<td>1E-6</td>
<td>184</td>
<td>6E-7</td>
<td>233</td>
<td>1E-6</td>
</tr>
<tr>
<td>2</td>
<td>9673</td>
<td>5E-04</td>
<td>225</td>
<td>4E-6</td>
<td>225</td>
<td>4E-6</td>
</tr>
<tr>
<td>4</td>
<td>120</td>
<td>2E-6</td>
<td>78</td>
<td>1E-7</td>
<td>120</td>
<td>2E-6</td>
</tr>
<tr>
<td>8</td>
<td>73</td>
<td>1E-8</td>
<td>48</td>
<td>9E-8</td>
<td>73</td>
<td>1E-8</td>
</tr>
<tr>
<td>16</td>
<td>61</td>
<td>7E-8</td>
<td>28</td>
<td>1E-7</td>
<td>61</td>
<td>7E-8</td>
</tr>
<tr>
<td>32</td>
<td>47</td>
<td>3E-8</td>
<td>21</td>
<td>1E-7</td>
<td>47</td>
<td>3E-8</td>
</tr>
<tr>
<td>64</td>
<td>68</td>
<td>8E-7</td>
<td>66</td>
<td>7E-7</td>
<td>68</td>
<td>8E-7</td>
</tr>
<tr>
<td>2</td>
<td>4184</td>
<td>4E-5</td>
<td>69</td>
<td>8E-7</td>
<td>69</td>
<td>8E-7</td>
</tr>
<tr>
<td>4</td>
<td>66</td>
<td>4E-7</td>
<td>63</td>
<td>4E-7</td>
<td>66</td>
<td>4E-7</td>
</tr>
<tr>
<td>8</td>
<td>61</td>
<td>4E-7</td>
<td>57</td>
<td>3E-7</td>
<td>61</td>
<td>4E-7</td>
</tr>
<tr>
<td>16</td>
<td>58</td>
<td>5E-7</td>
<td>52</td>
<td>6E-7</td>
<td>58</td>
<td>5E-7</td>
</tr>
<tr>
<td>32</td>
<td>53</td>
<td>6E-7</td>
<td>46</td>
<td>1E-6</td>
<td>53</td>
<td>6E-7</td>
</tr>
<tr>
<td>64</td>
<td>45</td>
<td>3E-7</td>
<td>37</td>
<td>8E-7</td>
<td>45</td>
<td>3E-7</td>
</tr>
</tbody>
</table>

In table 7, we use a fixed Block Jacobi preconditioner for all the partitions to compare the convergence of the methods with respect to the doubling of the number of partitions $t$. First, the graph of $A$ is partitioned into 64 parts using Metis’s kway edge separator. Then the Block Jacobi preconditioner $M$ is defined as the 64 block diagonals of the permuted matrix $A$. Each of the 64 blocks is factorized using $\text{Inria}$
Cholesky decomposition to obtain a block $L$. Then the matrix $A$ is partitioned once again using kway into $t = 2, 4, 8, 16, 32$ or $64$ parts that define the enlarged Krylov subspace. The preconditioner is permuted accordingly. The preconditioned LRE-CG converges faster than the preconditioned MSDO-CG and PCG. As the number of partitions or the maximum basis vectors added at each iteration is doubled, the preconditioned LRE-CG and MSDO-CG converge faster. However, for some matrices, like P0ISSON2D, NH2D, and ANI3D, as the number of partitions is doubled, the number of iterations till convergence decreases by only $10\% - 20\%$. Thus, it is efficient to use a maximum of $t = 4$ partitions which correspond to adding at most $t$ vectors to the basis of the enlarged Krylov subspace. For the matrices SKY2D and SKY3D, the number of iterations decreases by $33\% - 45\%$ and $25\% - 45\%$ respectively. Hence it is possible to use a maximum of $t = 8$ partitions or at most $t = 16$.

7 Conclusion

In this paper we have introduced two new iterative methods, MSDO-CG and LRE-CG, which are based on the enlarged Krylov subspace. After introducing the related existing methods (Block-CG, Coop-Cg and MCD-CG), we have defined the properties of the enlarged Krylov subspace, derived the new methods in the context of projection CG versions, provided parallel versions that reduce communication, and shown that the methods converge at least as fast as Classical CG in exact precision arithmetic. The convergence results show that they also converge faster than CG in finite precision arithmetic. We have also presented the preconditioned versions and tested their convergence with block Jacobi preconditioner.

MSDO-CG is a variation of the MSD-CG version, where we A-orthonormalize the $t$ search directions against previous directions and against each others. Due to the A-orthonormalization, we lose the short recurrence property of CG and we are obliged to save all the $tk_c$ search directions, where $k_c$ is the number of iterations till convergence. In LRE-CG we start by building an orthonormal basis for the enlarged Krylov subspace, then we use the whole basis to update the solution. The main difference between both methods in terms of performance, is that at each iteration of MSDO-CG, we use $t$ search directions to update the new approximate solution. Whereas in LRE-CG, at each iteration $i$, we use the entire basis formed by $ti$ vectors, to update the approximate solution and we solve a $ti \times ti$ system. However, this use of the whole basis leads to a relatively faster convergence than MSDO-CG. One way to limit this increasing cost is by restarting LRE-CG after some iterations. Another alternative is to choose at each iteration $i$, a linearly independant subset of the $ti$ computed vectors. This adds an extra cost, but reduces the size of the system that has to be solved at each iteration. A third alternative is to compute $ti$ vectors at each iteration $i$, where $t_0 = t$, and $t_i \leq t$. Then choose $ti$ linearly independant vectors where $\tilde{t}_i \leq t_i$, and $t_{i+1} = \tilde{t}_i$.

Although each iteration of the MSDO-CG and LRE-CG methods is at least $t$ times more expensive than the CG iteration in terms of flops, as shown in section 5 both methods use less communication, and Blas2 and Blas3 operations that can be parallelized in a more efficient way than the dot products in CG. Moreover, the MSDO-CG and LRE-CG methods converge faster than CG in terms of iterations as shown in section 4.

Our future work will focus on testing the LRE-CG versions discussed above, that are less expensive in terms of flops and memory requirements than LRE-CG, like restarted LRE-CG or LRE-CG with selected basis vectors. Then, the most stable version will be implemented in a parallel environment. We will also test LRE-CG on other real applications’ matrices, and with different preconditioners. Moreover, we would also like to compare the runtime of the LRE-CG version with the MSDO-CG method on a parallel environment. We will also derive and test other enlarged Krylov methods, like enlarged GMRES which has been derived but not tested yet.
References


Inria


Appendix A  MCD-CG Algorithm

Algorithm 9 The MSD-CG Algorithm

Input: $A$, the $n \times n$ symmetric positive definite matrix
Input: $b$, the $n \times 1$ right-hand side; $x_0$, the initial guess or iterate
Input: $\epsilon$, the stopping tolerance; $k_{max}$, the maximum allowed iterations
Output: $x_k$, the approximate solution of the system $Ax = b$

1: $r_0 = b - Ax_0$, $\rho = ||r_0||_2$, $k = 1$
2: for $i = 1, \ldots, t$ do Let $P(:, i) = T_i(r_0)$
3: end for
4: while $(\sqrt{\rho} > \epsilon ||b||_2$ and $k < k_{max}$ ) do
5: \quad $\alpha = (P^tAP)^{-1}(P^tr)$
6: \quad $x = x + P\alpha$
7: \quad $\beta = (P^tAP)^{-1}(P^tAr)$
8: \quad for $i = 1, \ldots, t$ do Let $P(:, i) = T_i(r) + \beta(i)P(:, i)$
9: \quad end for
10: $\rho = ||r||_2^2$
11: $k = k + 1$
12: end while

Appendix B  The A-orthogonalization of search directions

In this section we describe the A-orthogonalization of the $t$ newly computed vectors $P_{k+1}$ against all the previous vectors $P_i$’s for $i < k + 1$, and then against each others. Here $A$ is a symmetric positive definite matrix. The A-orthogonalization is simply an orthogonalization with the A inner product ($\langle \cdot, \cdot \rangle_A = \langle \cdot, \cdot \rangle$) rather than the L2 inner product ($\langle \cdot, \cdot \rangle$). A-orthonormalizing a tall and skinny $n \times t$ matrix $P_{k+1}$, or alternatively computing the oblique QR factorization of $P_{k+1}$, has been discussed in [21] and [20] in terms of stability and ease of parallelization. The goal is to get a $\hat{P}_{k+1}$, such that $\hat{P}_{k+1}^tA\hat{P}_{k+1} = I$.

There are two main classes for computing this oblique QR factorization of $P_{k+1} = \hat{P}_{k+1}R$. The first class is to factorize the matrix $A = B^tB$ using Cholesky decomposition or eigenvalue decomposition, which is expensive. Then $P_{k+1}^tAP_{k+1} = (BP_{k+1})^t(BP_{k+1})$, where the oblique QR factorization of $P_{k+1}$ is transformed into a Euclidean QR factorization of the matrix $BP_{k+1} = Q_B R_B$ with $\hat{P}_{k+1} = B^{-1}Q_B$ and $R = R_B$. The second class consists of avoiding any factorization of $A$, like CGS, CGS2, MGS, and the Cholesky factorization of the $t \times t$ matrix $P_{k+1}^tAP_{k+1}$. For A-orthonormalizing $P_{k+1}$ against all the previous vectors $P_i$ with $i < k + 1$, it is possible to use CGS, CGS2, MGS and A-choleskyBGS which was discussed in Hoemmen’s thesis (17, page 115).

Thus, we start by discussing the A-orthonormalization using modified Gram Schmidt in section B.1. However, this version is not easily parallelized and requires a lot of communication ($(tk + 1)log(t) + 2(t - 1)log(t)$ messages) as compared to the classical Gram Schmidt version. Then, in section B.2, we adapt the A-orthonormalization of the vectors of $P_{k+1}$ against $P_i$’s for $i < k + 1$ using the classical Gram Schmidt (CGS) to obtain a Block Gram Schmidt (BGS) version (Algorithm 14) with a inner product that requires only $2log(t)$ messages. As for the A-orthonormalization of the $P_{k+1}$ vectors against each others, we introduce a parallelizable version with reduced communication ($(2t - 1)log(t)$ messages). Note that CGS2, section B.2.3, consists of calling the algorithm CGS twice. Thus its cost is twice the cost of CGS. In section B.3, we briefly discuss the A-orthonormalization of $P_{k+1}$ using the Cholesky factorization (CholQR) of the $t \times t$ matrix $P_{k+1}^tAP_{k+1}$ which is referred to as A-CholQR and requires...
only \( \log(t) \) messages. We also present the Pre-CholQR version that was introduced in [20] and requires \( 3\log(t) \) messages.

### B.1 Modified Gram Schmidt A-orthonormalization

We start by introducing A-orthonormalization of the vectors of \( P_{k+1} \) against all the previous \( P_i \)'s for \( i < k + 1 \), then against each others in the first section. In the second section, we discuss versions that save flops and reduce communication. Finally, in the last section the parallelization of both kernels is described.

#### B.1.1 The A-orthonormalization using MGS

Assuming that the vectors of \( P_i \) are A-normalized, i.e. \( P_i(:,j)^T A P_i(:,j) = 1 \) for all \( j = 1,2,..,t \) and \( i = 1,2,..,k \), then the A-orthonormalization of the vectors of \( P_{k+1} \) against all the previous \( P_i \)'s for \( i < k + 1 \) is defined as follows:

**Algorithm 10** A-orthonormalization against previous vectors with MGS

**Input:** \( P_k \), \( P_{k+1} \), the \( k \) \( \times \) \( n \) symmetric positive definite matrix

**Input:** \( P_k, P_{k+1}, \) the \( k \) \( \times \) \( n \) sets of search directions

**Output:** \( P_{k+1} \), the search directions A-orthonormalized against \( P_1, P_2,.., P_k \)

1: \( \text{for } i = 1 : t \text{ do} \) \( \% \) loop over the vectors of \( P_{k+1} \)
2: \( \text{for } j = 1 : k \text{ do} \) \( \% \) loop over the different \( P_i \)'s
3: \( P_{k+1}(;i) = P_k(;i) - \left( P_k(:,j)^T A P_{k+1}(;o) P_k(:,i) \right) P_k(:,j) \)
4: \text{end for}
5: \text{end for}
6: \( P_{k+1}(;o) = P_{k+1}(;o) / \left| |P_{k+1}(;o)|_A \right| \) \( \% \) A-normalize
7: \text{end for}

At each inner iteration, one matrix-vector multiplication has to be computed (\( A P_k(:,o) \)), 1 dot product, and 1 saxpy, which costs \( 2n \log_2(n) - n + (2n - 1) + 2n = 2n \log_2(n) + 3n - 1 \) flops. Then, at each outermost iteration, one matrix-vector multiplication is computed (\( A P_k(:,o) \)), 1 dot product, 1 square root and 1 division which costs \( 2n \log_2(n) - n + (2n - 1) + 2 = 2n \log_2(n) + n + 1 \). The total cost of Algorithm 10 is \( (2n \log_2(n) + 3n - 1) t^2 k + (2n \log_2(n) + n + 1) t \), which is of the order of \( n n z t^2 k + n t^2 k \).

For the A-orthonormalization of the vectors of \( P_{k+1} \) against each others, it is defined as follows:

**Algorithm 11** A-orthonormalization against each others using MGS

**Input:** \( A \), the \( n \times n \) symmetric positive definite matrix

**Input:** \( P_k \), the search directions to be A-orthonormalized

**Output:** \( P_{k+1} \), the A-orthonormalized search directions

1: \( \text{for } i = 1 : t \text{ do} \) \( \% \) loop over the vectors of \( P_{k+1} \)
2: \( \text{for } j = 1 : (i - 1) \text{ do} \) \( \% \) A-orthogonalize against the vectors \( P_{k+1}(;1:i-1) \)
3: \( P_{k+1}(;i) = P_{k+1}(;i) - (P_{k+1}(;j)^T A P_{k+1}(;i)) P_{k+1}(;j) \)
4: \text{end for}
5: \( P_{k+1}(;i) = P_{k+1}(;i) / \left| |P_{k+1}(;i)|_A \right| \) \( \% \) A-normalize
6: \text{end for}

Similarly, the cost of the inner loop is \( 2n \log_2(n) + 3n - 1 \) flops and that of the outer loop is \( 2n \log_2(n) + n + 1 \), but the total cost is \( (2n \log_2(n) + 3n - 1)(t - 1) t^2 k + (2n \log_2(n) + n + 1) t \) flops, which is of the order of \( n n z t^2 + n t^2 \).
B.1.2 Saving flops in the A-orthonormalization using MGS

Since the A-orthonormalizations are expensive in term of flops, we present another alternative for computing the A-orthonormalizations that reduces the computed flops at the expense of storing more vectors. In Algorithm 11 and Algorithm 10, some matrix vector multiplications are repeatedly computed. For example, in Algorithm 11, \( AP_{k+1}(:,1) \) is computed \( t - 1 \) times, \( AP_{k+1}(:,2) \) is computed \( t - 2 \) times and generally, \( AP_{k+1}(:,i) \) is computed \( t - i \) times, which means that the matrix \( A \) is accessed \( (t - 1) \frac{t}{2} \) times for every call of the algorithm. Thus, it is possible after A-orthogonalizing a vector \( P_{k+1}(:,i) \) to compute and store \( w_i = AP_{k+1}(:,i) \). This eliminates the redundant flops and reduces the number of accesses of \( A \) to \( t \) times, but there is a need to store \( t \) extra vectors \( (w_i) \).

Moreover, it is possible to further reduce the computations and the number of times \( A \) is accessed at every call of the algorithm. Thus, it is possible after A-orthogonalizing a vector \( P_{k+1}(:,i) \) to compute and store \( w_i = AP_{k+1}(:,i) \). This eliminates the redundant flops and reduces the number of accesses of \( A \) to \( t \) times, but there is a need to store \( t \) extra vectors \( (w_i) \).

The A-orthonormalization against previous vectors with flops reduction can be performed as follows:

Algorithm 12 A-orthonormalization against previous vectors with MGS

<table>
<thead>
<tr>
<th>Input:</th>
<th>Output:</th>
<th>Flops</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P_1, P_2, \ldots, P_{k+1} ), the ( k + 1 ) sets of search directions</td>
<td>( P_{k+1} ), the search directions A-orthonormalized against ( P_1, P_2, \ldots, P_k )</td>
<td>( 4n - 1 + 2n + 2 )</td>
</tr>
</tbody>
</table>

Then, the cost of the A-orthonormalization against previous vectors in Algorithm 12 is \((6n - 1)t^2k + (4n + 1)t\) of the order of \( 6nt^2k \) flops.

Algorithm 13 Flops reduction in A-orthonormalization against each others with MGS

<table>
<thead>
<tr>
<th>Input:</th>
<th>Output:</th>
<th>Flops</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P_{k+1} ), the search directions to be A-orthonormalized</td>
<td>( W_{k+1}, AP_{k+1} ) where ( P_{k+1} ) is the A-orthonormalized search directions</td>
<td>( 4n - 1 + 2n + 2 )</td>
</tr>
</tbody>
</table>

Inria
As for the A-orthonormalization of $P_{k+1}$’s vectors against each others using MGS with flops reductions, it can be performed as in Algorithm 13. The cost of this version of A-orthonormalization (Algorithm 13) is $(6n - 1)(t - 1)^2 + (4n + 1)t$, which is of the order of $3nt^2$.

### B.1.3 Parallelization of the A-orthonormalization using MGS

In Algorithm 12 at each inner iteration we are A-orthonormalizing the updated vectors $P_{k+1}(; o)$ against the vector $P_t(; j)$, where the vector $P_{k+1}(; o)$ is changed at each inner iteration. Thus it is not possible to have a block MGS by eliminating all the for loops. However, it is possible to eliminate one for loop in Algorithm 12 as shown in Algorithm 14 by A-orthonormalizing the whole block $P_{k+1}$ against the vector $P_t(; j)$, where $P_{k+1}(; o) = P_{k+1}(; o) - (P_k(; o) W_t(; j)) P_t(; j)$ for all $o = 1, 2, ..., t$. Let $[P_t(; j)]_1$ be an $n \times t$ block containing $t$ duplicates of the vector $P_t(; j)$. Then, $P_{k+1} = P_{k+1} - [P_i(:, j)]_1 \text{diag}(P_{k+1}^t W_t(; j))$.

**Algorithm 14** A-orthonormalization against previous vectors with MGS

<table>
<thead>
<tr>
<th>Flops</th>
</tr>
</thead>
<tbody>
<tr>
<td>$2n - 1$</td>
</tr>
</tbody>
</table>

In Algorithm 13 rather than A-orthonormalizing each vector $P_{k+1}(; i)$ against all previous vectors $P_{k+1}(; j)$, we can A-orthogonalize $P_{k+1}(; i+1 : t)$ against the A-normalised vector $P_{k+1}(; i)$ as shown in Algorithm 15. Let $[P_{k+1}(; j)]_{i-1}$ be an $n \times (t - i)$ block containing $t - i$ duplicates of the vector $P_{k+1}(; j)$. Then $P_{k+1}(; i+1 : t) = P_{k+1}(; i+1 : t) - [P_{k+1}(; j)]_{i-1} \text{diag}(W_{k+1}(; i) P_{k+1}(; i+1 : t))$

**Algorithm 15** A-orthonormalization against each others with MGS

<table>
<thead>
<tr>
<th>Flops</th>
</tr>
</thead>
<tbody>
<tr>
<td>$2n + 2t$</td>
</tr>
</tbody>
</table>

Then the parallelization of Algorithms 14 and 15 goes as follows. We assume that we have $t$ proces-
sors with distributed memory, and each processor $p_i$ is assigned a rowwise part of all $W_j$ ($W_j(\delta_{pi}, :)$) for $j = 1, 2, \ldots, k + 1$ and the same rowwise part of all $P_j$ ($P_j(\delta_{pi}, :)$) for $j = 1, 2, \ldots, k + 1$ where $\delta_{pi} \cap \delta_h = \emptyset$ for all $p_i \neq h$ and $\cup_{h=1}^{n} \delta_h = \{1, 2, 3, \ldots, n\}$.

At each inner iteration of Algorithm 14, each processor $p_i$ has to compute $P_{k+1}(\delta_{pi}, :) = P_{k+1}(\delta_{pi}, :) - [P_i(\delta_{pi}, :)\text{diag}(W_i(:, j)) P_{k+1}]$. First, each processor $p_i$ computes a part of the matrix vector multiplication $W_i(\delta_{pi}, :) P_{k+1}(\delta_{pi}, :)$. Then, a communication of the form “all reduce” is performed to send the $1 \times t W_i(:, j) P_{k+1}$’s value to all the processors. Finally, processor $p_i$ computes $P_{k+1}(\delta_{pi}, :)$ and $W_{k+1}(\delta_{pi}, :)$.

Finally, each processor $p_i$ computes its corresponding part of the dot product $W_{k+1}(\delta_{pi}, :) P_{k+1}(\delta_{pi}, :)$ for all $o = 1, 2, \ldots, t$ and an “all reduce” is used to send $p_{ap_{k+1}}$’s value to all the processors. Then, each processor $A$-normalizes $P_{k+1}(\delta_{pi}, :)$ and $W_{k+1}(\delta_{pi}, :)$.

Thus, at each iteration and then receives the whole $P_{k+1}(\delta_{pi}, :)$, $A$-orthonormalization.

As for the parallelization of Algorithm 15, it is similar to that of Algorithm 14 where at each inner iteration processor $p_i$ computes a part of the matrix vector multiplication $W_{k+1}(\delta_{pi}, :) P_{k+1}(\delta_{pi}, :)$ and then receives the whole $1 \times t$ vector, $W_{k+1}(\delta_{pi}, :) P_{k+1}(\delta_{pi}, :)$, using an “all reduce”. Then, it computes $P_{k+1}(\delta_{pi}, :) W_{k+1}(\delta_{pi}, :)$ and a part of the dot product $p_{ap_{k+1}}$ and receives the whole dot product by an “all reduce”. Finally, each processor $A$-normalizes its part of $P_{k+1}(\delta_{pi}, :)$ and $W_{k+1}(\delta_{pi}, :)$.

Thus, at each iteration “all reduce” communications are performed, where $t$ words are sent in the first and one word in second. So, in total $(tk + 1)(log(t) + t)log(t)$ words are sent in Algorithm 15, hence, by ignoring lower order terms we obtain

$Time_{MGS1Aort} \approx \gamma_0 (2n^2 + \alpha) t \log(t) + \beta t^2 \log(t)$

For the parallelization of Algorithm 15, it is similar to that of Algorithm 14 where at each inner iteration processor $p_i$ computes a part of the matrix vector multiplication $W_{k+1}(\delta_{pi}, :) P_{k+1}(\delta_{pi}, :)$ and then receives the whole $1 \times t$ vector, $W_{k+1}(\delta_{pi}, :) P_{k+1}(\delta_{pi}, :)$, using an “all reduce”. Then, it computes $P_{k+1}(\delta_{pi}, :) W_{k+1}(\delta_{pi}, :)$ and a part of the dot product $p_{ap_{k+1}}$ and receives the whole dot product by an “all reduce”. Finally, each processor $A$-normalizes its part of $P_{k+1}(\delta_{pi}, :)$ and $W_{k+1}(\delta_{pi}, :)$.

Thus, at each iteration “all reduce” communications are performed, where $t$ words are sent in the first and one word in second. So, in total $(tk + 1)(log(t) + t)log(t)$ words are sent in Algorithm 15, hence, by ignoring lower order terms we obtain

$Time_{MGS2Aort} \approx \gamma_0 (2n^2 + \alpha) t \log(t) + \beta t^2 \log(t)$

### B.2 Classical Gram Schmidt A-orthonormalization

Since the MGS A-orthonormalization is costly in terms of communication, we introduce the classical Gram Schmidt (CGS) A-orthonormalization and show that it is equivalent to a QR decomposition with A inner product rather than the usual L2 inner product. Then we present the parallelization of the introduced algorithms. In section B.2.1, the A-orthonormalization against previous vectors using CGS is discussed, whereas in section B.2.2 we discuss the A-orthonormalization of the vectors using CGS. Then in section B.2.3 we introduce the CGS A-orthonormalization with reorthogonalization.

#### B.2.1 A-orthonormalization against previous vectors using CGS

The A-orthonormalization of $P_{k+1}$ against the vectors of all the previous $P_i$’s for $i \leq k + 1$ is defined as in Algorithm 16.

More precisely, $P_{k+1}(\delta_{pi}, :) = P_{k+1}(\delta_{pi}, :) - \sum_{i=1}^{k} \sum_{j=1}^{t} (P_i(\delta_{pi}, :) A P_{k+1}(\delta_{pi}, :) P_i(\delta_{pi}, :) )$

If we let $W_{k+1} = A P_{k+1}$, then $P_{k+1}(\delta_{pi}, :) = P_{k+1}(\delta_{pi}, :) - \sum_{i=1}^{k} P_i(\delta_{pi}, :) W_{k+1}(\delta_{pi}, :o)$. Moreover, $P_{k+1} = P_{k+1} - \sum_{i=1}^{k} P_i(\delta_{pi}, :) W_{k+1}(\delta_{pi}, :o)$. Let $Q_k = [P_1, P_2, \ldots, P_k]$, then $P_{k+1} = P_{k+1} - Q_k(Q_k^T W_{k+1})$.

This represents a Block classical gram schmidt (BCGS) version of the A-orthonormalization.

The total flops performed in Algorithm 17 is $2(2n^2 - n) t + (2n - 1)t^2 k + 2t^2 kn + 3nt = 4nzt + nt + [4n^2 t^2 - 2k] t k \approx 4nzt + 4nt^2 k$.

As for the parallelization of Algorithm 17, it is straightforward due to the block format. Assuming that we have $t$ processors with distributed memory, and each processor $p_i$ is assigned a rowwise part of $A(\delta_{pi}, :)$, a rowwise part of $Q_k(Q_k(\delta_{pi}, :))$ and a rowwise part of $P_{k+1}$, where $\delta_{pi} \cap \delta_h = \emptyset$ for all $p_i \neq h$ and $\cup_{h=1}^{n} \delta_h = \{1, 2, 3, \ldots, n\}$. Inria
obtains the full dot products via an all reduce. Finally each processor $A$-normalizes its part of $P$.

Another “all reduce” is needed so that processor $A$ gets

message of size $Q$. The search directions $A$-orthonormalized against previous vectors with BCGS Flops

Algorithm 16 A-orthonormalization against previous vectors with CGS

**Input:** $A$, the $n \times n$ symmetric positive definite matrix

**Input:** $P_1, P_2, ..., P_k$, the $k$ sets of search directions

**Output:** $\tilde{P}_{k+1}$, the search directions A-orthonormalized against $P_1, P_2, ..., P_k$

1: Let $\tilde{P}_{k+1} = P_{k+1}$
2: for $i = 1 : t$ do %loop over the vectors of $P_{k+1}$
3:     for $j = 1 : k$ do %loop over the different $P_i$'s
4:         $\tilde{P}_{k+1}(i, o) = \tilde{P}_{k+1}(i, o) - (P_i(i, j)^t A P_{k+1}(i, o)) P_i(i, j)$
5:     end for
6: end for
7: $\tilde{P}_{k+1}(i, o) = \frac{\tilde{P}_{k+1}(i, o)}{||P_{k+1}(i, o)||_A} = \frac{\tilde{P}_{k+1}(i, o)}{\sqrt{P_{k+1}(i, o)^t A P_{k+1}(i, o)}}$ %A-normalize
8: end for

Algorithm 17 A-orthonormalization against previous vectors with BCGS

**Input:** $A$, the $n \times n$ symmetric positive definite matrix

**Input:** $P_1, P_2, ..., P_k$, the $k$ sets of search directions

**Output:** $\tilde{P}_{k+1}$, the search directions to be A-orthonormalized

1: $W_{k+1} = A \tilde{P}_{k+1}$
2: $\tilde{P}_{k+1} = P_{k+1} - Q_k Q_k^t W_{k+1}$
3: $\tilde{W}_{k+1} = A \tilde{P}_{k+1}$
4: for $i = 1 : t$ do %loop over the vectors of $P_{k+1}$ and A-normalize
5:     $\tilde{P}_{k+1}(i, o) = \frac{\tilde{P}_{k+1}(i, o)}{||P_{k+1}(i, o)||_A} = \frac{\tilde{P}_{k+1}(i, o)}{\sqrt{P_{k+1}(i, o)^t W_{k+1}(i, o)}}$
6: end for

First, processor $pi$ computes $A(\cdot, \delta_{pi}) P_{k+1}(\delta_{pi}, \cdot)$ and receives the full $n \times t$ matrix $W_{k+1}$ via an “all reduce”. Then it computes $Q_k(\delta_{pi}, \cdot)^t W_{k+1}(\delta_{pi}, \cdot)$ and obtains the full $t \times t$ matrix $Q_k^t W_{k+1}$ using an “all reduce”. Then, processor $pi$ computes $\tilde{P}_{k+1}(\delta_{pi}, \cdot) = P_{k+1}(\delta_{pi}, \cdot) - Q_k(\delta_{pi}, \cdot)(Q_k^t W_{k+1})$. Another “all reduce” is needed so that processor $pi$ has the full $\tilde{P}_{k+1}$ needed to compute $\tilde{W}_{k+1}(\delta_{pi}, \cdot) = A(\delta_{pi}, \cdot) \tilde{P}_{k+1}$. Processor $pi$ computes $t$ partial dot products of the form $\tilde{P}_{k+1}(\delta_{pi}, o)^t W_{k+1}(\delta_{pi}, o)$ and obtains the full dot products via an all reduce. Finally each processor $A$-normalizes its part of $P_{k+1}$, i.e $\tilde{P}_{k+1}(\delta_{pi}, i) = \frac{P_{k+1}(\delta_{pi}, i)}{\tilde{W}_{k+1}(\delta_{pi}, i)}$ for all $i = 1, 2, ..., t$. So in total there is a need to perform 4 all reduce for parallelizing algorithm [17].

It is possible to reduce the communication to only two by assuming that $W_{k+1} = A P_{k+1}$ has already been computed and it is an input to Algorithm [18] along with $W_k = A Q_k = [W_1, W_2, ..., W_k]$. The only communication is an “all reduce” of the $t \times t$ matrix $Q_k^t W_{k+1}$ and another “all reduce” of the vector of size $t$ containing the norms of the columns of $\tilde{P}_{k+1}$. We assume that it is possible to send a message of size $t^2 k$ words at once. Thus, $2log(t)$ messages are sent with $(tk + 1)log(t)$ words where

$$\frac{(6n-1)t^2k+3n}{t} = (6n - 1)tk + 3n$$

flops are performed in parallel. Hence, by ignoring lower order terms we obtain

$$Time_{BCGS_{Aort}} = \gamma_6ntk + \alpha_22log(t) + \beta_4t^2klog(t)$$

RR n° 8597
Algorithm 18 A-orthonormalization against previous vectors with BCGS

\textbf{Input:} $Q_k = [P_1, P_2, \ldots, P_k]$, the $tk$ search directions  \\
\textbf{Input:} $P_{k+1}$, the $t$ search directions to be A-orthonormalized  \\
\textbf{Input:} $W_{k+1} = AP_{k+1}$; $\hat{W}_k = A\hat{P}_k$  \\
\textbf{Output:} $\hat{P}_{k+1}$, the search directions A-orthonormalized against $P_1, P_2, \ldots, P_k$; $\hat{W}_{k+1} = A\hat{P}_{k+1}$

1: $\hat{P}_{k+1} = P_{k+1} - Q_k (Q_k^T W_{k+1})$  
2: $\hat{W}_{k+1} = W_{k+1} - W_k (Q_k^T W_{k+1})$
3: for $i = 1 : t$ do  
4: \hspace{1em} $\hat{P}_{k+1}(:, i) = \frac{\hat{P}_{k+1}(:, i) - \hat{P}_{k+1}(:, i)^T A \hat{P}_{k+1}(:, i) \hat{P}_{k+1}(:, i)}{||\hat{P}_{k+1}(:, i)||_2 A \hat{P}_{k+1}(:, i)}$  
5: \hspace{1em} end for
6: $\hat{W}_{k+1}(:, i) = \frac{\hat{W}_{k+1}(:, i) - \hat{W}_{k+1}(:, i)^T A \hat{W}_{k+1}(:, i) \hat{W}_{k+1}(:, i)}{||\hat{W}_{k+1}(:, i)||_2 A \hat{W}_{k+1}(:, i)}$  
7: end for
Flops $(2n - 1)t^2k + (2tk - 1)nt + nt$

\[ 2nt^2k \]

B.2.2 A-orthonormalization of a set of vectors using CGS

Given a set of vectors $P_{k+1}$ that are A-normalized, i.e. the diagonal of $P_{k+1}^T A P_{k+1}$ is equal to ones, we A-orthonormalize it ($P_{k+1}^T A P_{k+1} = I$) using a classical Gram Schmidt procedure as in algorithm 19.

Algorithm 19 A-orthonormalization against each others using CGS

\textbf{Input:} $A$, the $n \times n$ symmetric positive definite matrix  \\
\textbf{Input:} $P_{k+1}$, the search directions to be A-orthonormalized  \\
\textbf{Output:} $\hat{P}_{k+1}$, the A-orthonormalized search directions

1: Let $\hat{P}_{k+1} = P_{k+1}$  
2: for $i = 1 : t$ do  
3: \hspace{1em} for $j = 1 : (i - 1)$ do  
4: \hspace{2em} $\hat{P}_{k+1}(:, i) = \hat{P}_{k+1}(:, i) - (\hat{P}_{k+1}(:, j) A \hat{P}_{k+1}(:, i)) \hat{P}_{k+1}(:, j)$  
5: \hspace{1em} end for  
6: \hspace{1em} $\hat{P}_{k+1}(:, i) = \frac{\hat{P}_{k+1}(:, i) - \hat{P}_{k+1}(:, i)^T A \hat{P}_{k+1}(:, i) \hat{P}_{k+1}(:, i)}{||\hat{P}_{k+1}(:, i)||_2 A \hat{P}_{k+1}(:, i)}$  
7: \hspace{1em} end for$

The CGS A-orthonormalization can be reformulated as a QR factorization

\[ P_{k+1} = \hat{P}_{k+1} R \]

where $\hat{P}_{k+1}$ is an A-orthonormal matrix, and $R$ is a $t \times t$ upper triangular matrix defined by the entries $r_{j,i}$ for all $j = 1, 2, \ldots, i$ and $i = 1, 2, \ldots, t$.

\[
R = \begin{pmatrix}
    r_{1,1} & r_{1,2} & r_{1,3} & \cdots & r_{1,t} \\
    r_{2,2} & r_{2,3} & \cdots & \cdots & r_{2,t} \\
    r_{3,3} & \cdots & r_{3,t} \\
    \cdots & \cdots & \cdots & \cdots & \cdots \\
    r_{t,t} & \cdots & \cdots & \cdots & \cdots
\end{pmatrix} = \begin{pmatrix}
    ||p_1||_A & <\bar{p}_1, p_2>_A & <\bar{p}_1, p_1>_A & \cdots & <\bar{p}_1, p_t>_A \\
    r_{2,2} & <\bar{p}_2, p_1>_A & \cdots & \cdots & <\bar{p}_2, p_t>_A \\
    r_{3,3} & \cdots & \cdots & \cdots & \cdots \\
    \cdots & \cdots & \cdots & \cdots & \cdots \\
    r_{t,t} & \cdots & \cdots & \cdots & \cdots
\end{pmatrix}
\]

Although the CGS A-orthonormalization is equivalent to a QR factorization with the A inner product, we were not able parallelize it using reduction trees with the same communication pattern as in TSQR [3]. But we can optimize the communication in algorithm 19 by noticing that once a vector $p_i$ is orthonormalized, we can compute the corresponding entries of the matrix $R$, i.e $R(i, i + 1 : t)$. By taking this into consideration, algorithm 19 can be restructured as in algorithm 20.

Inria
Algorithm 20 QR factorization with A inner product using CGS

**Input:** A, the \( n \times n \) symmetric positive definite matrix

**Input:** \( P_{k+1} \), the search directions to be A-orthonormalized

**Output:** \( \hat{P}_{k+1} \), the A-orthonormalized search directions \( \hat{P}_{k+1}^T A \hat{P}_{k+1} = I \)

**Output:** \( R \), the upper triangular matrix such that \( \hat{P}_{k+1} = \hat{P}_{k+1} R \)

1. \( W_{k+1} = AP_{k+1} \)
2. \( R(1, 1) = \sqrt{\hat{P}_{k+1}^T (1, 1)^T W (1, 1)} \)
3. \( \hat{P}_{k+1} (1, :) = \hat{P}_{k+1} (1, :) / R(1, 1) \)
4. for \( i = 2 : t \) do
   5. \( R(i - 1, i : t) = \hat{P}_{k+1}(:, i - 1)^T W_{k+1}(:, i : t) \)
   6. \( \hat{P}_{k+1}(:, i) = \hat{P}_{k+1}(:, i) - \hat{P}_{k+1}(:, 1 : i - 1) R(1 : i - 1, i) \)
   7. \( R(i, i) = \sqrt{\hat{P}_{k+1}(:, i)^T A \hat{P}_{k+1}(:, i)} \)
   8. \( \hat{P}_{k+1}(:, i) = \hat{P}_{k+1}(:, i) / R(i, i) \)
5. end for

The total flops of Algorithm 20 is of the order of \( mnzt + nt^2 \).

\[
\text{Total} = 2mnzt - nt + 3n + \sum_{i=2}^{t} [(2n - 1)(t - i + 1) + 2(i - 1)n + (2nzn + 2n)]
\]
\[
= 2mnzt - nt + 3n + \sum_{i=2}^{t} [(2n - 1)(t + 1) - (2n - 1)i + 2ni - 2n + 2nzn + 2n]
\]
\[
= 2mnzt - nt + 3n + \sum_{i=2}^{t} [(2n - 1)(t + 1) + i + 2nzn]
\]
\[
= 2mnzt - nt + 3n + [(2n - 1)(t + 1) + 2nzn](t - 1) + \frac{(t+1)(t+2)}{2} - 1
\]
\[
= 4mnzt - 2nzt - nt + n + 2nt^2 - t^2 + \frac{t^2 + t}{2}
\]
\[
= 4mnzt - 2nzt - nt + n + 2nt^2 - t^2 + \frac{t^2 + t}{2}
\]

The parallelization of Algorithm 20 starts by distributing the data similarly to Algorithm 17. Processor \( pi \) computes \( A(:, \delta_{pi}) P_{k+1} (\delta_{pi}, :) \) and receives \( W_{k+1} \) via an “all reduce,” and computes \( P_{k+1} (\delta_{pi}, 1)^T W (\delta_{pi}, 1) \), and receives the full dot product \( P_{k+1} (1, :) W (1, :) \), needed to compute \( R(1, 1) \), via an “all reduce.”

Then, it computes \( \hat{P}_{k+1} (\delta_{pi}, 1) = \frac{P_{k+1} (\delta_{pi}, 1)}{R(1, 1)} \).

At each iteration, processor \( pi \) computes \( \hat{P}_{k+1} (\delta_{pi}, i - 1)^T W_{k+1} (\delta_{pi}, i : t) \) and receives the full \( R(i - 1, i : t) \) by an all reduce. Then, it computes \( \hat{P}_{k+1} (\delta_{pi}, i) = P_{k+1} (\delta_{pi}, i) - \hat{P}_{k+1} (\delta_{pi}, 1 : i - 1) R(1 : i - 1, i) \)

and receives the \( P_{k+1} (\delta_{pi}, i) \) from \( m_{MB} \) adjacent processors where \( \delta_{pi} = \text{Adjacent} (G(A), \delta_{pi}) \). Then it computes \( W_{k+1} (\delta_{pi}, i) = A (\delta_{pi}, \beta_{pi}) \hat{P}_{k+1} (\delta_{pi}, i) \) and \( \hat{P}_{k+1} (\delta_{pi}, i)^T W_{k+1} (\delta_{pi}, i) \) and receives the full \( \hat{P}_{k+1} (\delta_{pi}, i)^T A \hat{P}_{k+1} (\delta_{pi}, i) \), needed to compute \( R(i, i) \), via an all reduce. Finally, it computes \( \hat{P}_{k+1} (\delta_{pi}, i) = \frac{\hat{P}_{k+1} (\delta_{pi}, i)}{R(i, i)} \). So, there is a need for a total of \( 2t - 1 \) “all reduce” and \( t \) communications with the \( m_{MB} \) neighboring processors.

It is possible to reduce the communication to only \( 2t - 1 \) “all reduce” by assuming that \( W_{k+1} = A P_{k+1} \) has already been computed and it is an input to Algorithm 21. Then, at each iteration \( i \), an “all reduce” of the vector \( R(i - 1, i : t) \) of size \( t - i + 1 \) is performed and another “all reduce” of the entry \( R(i, i) \) is performed. Thus, a total of \( (2t - 1) \log(t) \) messages are sent with \( (1 + \sum_{i=2}^{t} 2 - 2 - i) \log(t) = \frac{t(t+1)}{2} \log(t) \) words where \( \frac{3nt^2 + nt + \frac{R(1, 1)}{2}}{4} = 3nt + n + \frac{(1-t)}{2} \) flops are performed in parallel. Hence, by ignoring lower order terms we obtain

\[
\text{Time}_{QRCSG_{sort}} \approx \gamma_c 3nt + \alpha_c 2t \log(t) + \beta_c t^2 \log(t)
\]
B.3 Cholesky QR A-orthonormalization

Algorithm 21 QR factorization with A inner product using CGS

| Input: | $P_{k+1}$, the search directions to be A-orthonormalized; $W_{k+1} = AP_{k+1}$ |
| Output: | $R$, the upper triangular matrix such that $P_{k+1} = \tilde{R}_{k+1}R$ |

Algorithm 22 A-orthonormalization of $P_{k+1}$ against previous vectors of $Q_k$ using CGS2

| Input: | $Q_k$, the $t$th search directions |
| Input: | $P_{k+1}$, the $t$ search directions to be A-orthonormalized |
| Input: | $W_{k+1} = AP_{k+1}$; $W_k = AQ_k$ |

Algorithm 23 Cholesky QR A-orthonormalization

A-orthonormalizing the $n \times t$ full rank matrix $P_{k+1}$ is equivalent to a QR factorization $P_{k+1} = \tilde{R}_{k+1}R$ where $P_{k+1}^T \tilde{R}_{k+1} = I$. Thus, $P_{k+1}^T AP_{k+1} = (\tilde{R}_{k+1}R)^T \tilde{R}_{k+1}R = R^T R$ and $R$ can be obtained by performing a Cholesky factorization of the SPD matrix $P_{k+1}^T AP_{k+1}$. Then, $\tilde{P}_{k+1} = P_{k+1}R^{-1}$ is obtained by solving the lower triangular system $R^T \tilde{P}_{k+1} = P_{k+1}$ with multiple right-hand sides. This procedure is called A-CholQR and summarized in Algorithm 24. Similarly to the other A-orthonormalization methods, we may assume that $W_{k+1}$ is already computed, then the obtained A-CholQR is described in Algorithm 25.

The parallelization of Algorithm 25 assumes that we have $t$ processors and each is assigned a rowwise part of $P_{k+1}$ and $W_{k+1}$ corresponding to the $\delta_i$ subset of indices defined previously, $P_{k+1}^T (\delta_i, :)$ and $W_{k+1}^T (\delta_i, :).$ And each processor $i$ should compute $\tilde{P}_{k+1}^T (\delta_i, :)$ and $\tilde{W}_{k+1}^T (\delta_i, :).$ Then each processor $i$ computes $W_{k+1}^T (\delta_i, :)^T \tilde{P}_{k+1}^T (\delta_i, :)$ and receives the $t \times t$ matrix $C$ via an “all reduce” or equivalently...
Algorithm 23 $L^{-1}AL^{-1}$-orthonormalization against previous vectors of $Q_k$ with CGS2

**Input:** $A$, the $n \times n$ symmetric positive definite matrix; $L$, $n \times n$ preconditioner

**Input:** $P_{k+1}$, the $t_k$ search directions

**Output:** $\tilde{P}_{k+1}$, the search directions to be $L^{-1}A(L')^{-1}$-orthonormalized against $Q_k$

**Output:** $\tilde{W}_{k+1} = L^{-1}AL^{-1}P_{k+1}$

1: $\tilde{P}_{k+1} = L^{-1}AL^{-1}P_{k+1}$
2: $\tilde{W}_{k+1} = L^{-1}AL^{-1}P_{k+1}$
3: $\tilde{W}_{k+1} = L^{-1}AL^{-1}P_{k+1}$

4: \textbf{for} $i = 1 : t_k$ \textbf{do}
5: \texttt{\% loop over the vectors of } $P_{k+1}$ and $L^{-1}AL^{-1}$, normalize
6: \texttt{\% loop over the vectors of } $P_{k+1}$ and $L^{-1}AL^{-1}$, normalize
7: \texttt{\% loop over the vectors of } $P_{k+1}$ and $L^{-1}AL^{-1}$, normalize

8: $\tilde{P}_{k+1} = \tilde{P}_{k+1} - Q_k(Q_k^{T}\tilde{W}_{k+1})$
9: $\tilde{W}_{k+1} = L^{-1}AL^{-1}P_{k+1}$
10: $\tilde{P}_{k+1} = \tilde{P}_{k+1} - Q_k(Q_k^{T}\tilde{W}_{k+1})$
11: $\tilde{W}_{k+1} = L^{-1}AL^{-1}P_{k+1}$

12: \textbf{end for}
13: \textbf{end for}
14: \textbf{end for}

Algorithm 24 A-CholQR

**Input:** $A$, the $n \times n$ symmetric positive definite matrix

**Input:** $P_{k+1}$, the search directions to be $A$-orthonormalized

**Output:** $\tilde{P}_{k+1}$, the $A$-orthonormalized search directions

1: Compute $W_{k+1} = AP_{k+1}$
2: Compute $C = W_{k+1}^{T}P_{k+1}$
3: Compute the Cholesky factorization of $C = R^{T}R$ to obtain $R$
4: Solve $R^{T}\tilde{P}_{k+1} = P_{k+1}$
5: \textbf{for} $i = 1 : t_k$ \textbf{do}
6: $\tilde{P}_{k+1} = \frac{P_{k+1}(.;i)}{|P_{k+1}(.;i)|A} = \frac{P_{k+1}(.;i)}{\sqrt{P_{k+1}(.;i)^{T}AP_{k+1}(.;i)}}$
7: \textbf{end for}

Algorithm 25 A-CholQR

**Input:** $P_{k+1}$, the search directions to be $A$-orthonormalized, $W_{k+1} = AP_{k+1}$

**Output:** $\tilde{P}_{k+1}$, the $A$-orthonormalized search directions; $\tilde{W}_{k+1} = AP_{k+1}$

1: Compute $C = W_{k+1}^{T}P_{k+1}$
2: Compute the Cholesky factorization of $C = R^{T}R$ to obtain $R$
3: Solve $R^{T}\tilde{P}_{k+1} = P_{k+1}$
4: Solve $R^{T}\tilde{W}_{k+1} = W_{k+1}$
5: \textbf{for} $i = 1 : t_k$ \textbf{do}
6: $\tilde{P}_{k+1} = \frac{P_{k+1}(.;i)}{|P_{k+1}(.;i)|A} = \frac{P_{k+1}(.;i)}{\sqrt{P_{k+1}(.;i)^{T}W_{k+1}(.;i)}}$
7: \textbf{end for}

\texttt{Flops}

- \texttt{Computations:} $2 \times n^2 - n$ for $L^{-1}AL^{-1}$
- \texttt{Solve:} $2 \times n^2$ for $R^{T}R$ and $A^{-1}R^{T}R$ in Cholesky
- \texttt{Orthonormalization:} $3 \times n$ for $S$ normalization
- \texttt{Total:} $4 \times n^2$ for $L^{-1}AL^{-1}$, $R^{T}R$, and $A^{-1}R^{T}R$
Algorithm 27 Pre-CholQR

Input: $A$, the $n \times n$ symmetric positive definite matrix; $L$, an $n \times n$ preconditioner

Output: $\tilde{P}_{k+1}$, the A-orthonormalized search directions

1: Compute $C = W_{k+1}^T P_{k+1}$
2: Compute the Cholesky QR factorization of $C$ to obtain $R$
3: Solve $R^T \tilde{P}_{k+1} = P_{k+1}$
4: Compute $W_{k+1} = L^{-1} A L^{-1} \tilde{P}_{k+1}$
5: for $i = 1 : t$ do
6: $\tilde{p}_{k+1}(:, i) = \frac{p_{k+1}(:, i)}{\|p_{k+1}(:, i)\|_A}$ and $w_{k+1}(:, i) = \frac{w_{k+1}(:, i)}{\|w_{k+1}(:, i)\|_A}$
7: end for

Recently, Lowery and Langou presented a new version of A-CholQR in [20], which they call Pre-CholQR (Algorithm 27). It consists in performing a Euclidean QR factorization with L2 before calling the A-CholQR A-orthonormalization, Algorithm 24. The QR factorization of $P_{k+1}$ can be done using the TSQR [4], which requires sending $\log(t)$ messages, each of size $\frac{t^2}{2}$ words and computing $2nt + \frac{2}{3} t^3 \log(t)$. Then, parallelizing Algorithm 24 requires performing two “all reduce” or $2 \log(t)$ messages with $(nt + t^2) \log(t)$ words. In total, parallelizing Algorithm 27 requires sending $3 \log(t)$ messages with $(nt + 1.5t^2) \log(t)$ words. Hence, by ignoring lower order terms we obtain

$$\text{Time}_{\text{PreCholQR}} \approx \gamma c (6nt + \frac{2}{3} t^3 \log(t)) + \alpha \gamma 2t \log(t) + \beta \gamma (nt + \frac{3}{2} t^2 \log(t))$$