

Additive Operator Decomposition and Optimization–Based Reconnection with Applications

Pavel Bochev¹ and Denis Ridzal²

¹ Applied Mathematics and Applications,

² Optimization and Uncertainty Quantification,
Sandia National Laboratories³, Albuquerque, NM 87185-1320, USA,
{pboche, dridzal}@sandia.gov

Abstract. We develop an optimization-based approach for additive decomposition and reconnection of algebraic problems arising from discretization of partial differential equations (PDEs). Application to a scalar advection-diffusion PDE illustrates the new approach. In particular, we obtain a robust iterative solver for advection-dominated problems using standard multi-level solvers for the Poisson equation.

1 Introduction

Decomposition of PDEs into component problems that are easier to solve is at the heart of many numerical procedures: operator split [1] and domain decomposition [2] are two classical examples. The use of optimization and control ideas to this end is another possibility that has yet to receive proper attention despite its many promises and excellent theoretical foundations. For previous work in this direction we refer to [3–5] and the references cited therein.

In this paper we develop an optimization-based approach for additive decomposition and reconnection of algebraic equations that is appropriate for problems associated with discretized PDEs. Our approach differs from the ideas in [3–5] in several important ways. The main focus of [4, 5] is on formulation of non-overlapping domain-decomposition via optimization, whereas our approach targets the complementary single domain/multiple physics setting.

Our ideas are closer to the decomposition framework in [3]. Nevertheless, there are crucial differences in the definition of the objectives, controls, and the constraints. The approach in [3] retains the original equation, albeit written in a form that includes the controls, and the constraints impose equality of states and controls. In other words, reconnection in [3] is effected through the constraints rather than the objective.

³ Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under contract DE-AC04-94-AL85000.

In contrast, we replace the original problem by component problems that, by themselves, are not equivalent to the original problem. These problems define the constraints, while reconnection is effected via the objective functional. Consequently, in our approach the states are different from the controls and the objective functional is critical for “closing” the formulation.

2 Additive Operator Decomposition

For clarity we present the approach in an algebraic setting, i.e., we consider the solution of the linear system

$$Ax = b, \quad (1)$$

where $A \in \mathbb{R}^{n \times n}$ is a nonsingular matrix and x and b are vectors in \mathbb{R}^n . We present a method suitable for the scenario in which the matrix A comprises multiple operators with fundamentally different mathematical properties, e.g. resulting from an all-at-once discretization of a multiphysics problem, in which case the linear system (1) requires nonstandard, highly specialized solution techniques. For a concrete example and discussion, see Section 4.

The proposed optimization-based approach for the solution of (1) rests on the assumption that the matrix A can be written as the sum of two *component* matrices

$$A = A_1 + A_2 \quad (2)$$

for which robust solvers are readily available. We note that A_1 and A_2 can represent the original operator components of a multi-operator equation, however, other, nontrivial decompositions are oftentimes more useful, see Section 4.

We assume that A_1 and A_2 are nonsingular. To motivate the optimization formulation, we consider an equivalent formulation of (1) in terms the component matrices,

$$A_1x - u - b + A_2x + u = 0,$$

where $u \in \mathbb{R}^n$ is an arbitrary *coupling* vector. As the overall intent is to make use of robust solvers available for the solution of linear systems involving A_1 and A_2 , we aim to develop a procedure that would allow us to repeatedly solve linear systems of the type

$$A_1x = u + b \quad \text{and} \quad A_2x = -u,$$

instead of the original problem. Our approach, based on ideas from optimization and control, is presented next.

3 Reconnection via an Optimization Formulation

We propose to replace (1) by the following constrained optimization problem:

$$\begin{cases} \text{minimize} & J^\varepsilon(x_1, x_2, u) = \frac{1}{2} \left(\|x_1 - x_2\|_2^2 + \varepsilon \|u\|_2^2 \right) \\ \text{subject to} & \begin{cases} A_1x_1 - u = b \\ A_2x_2 + u = 0, \end{cases} \end{cases} \quad (3)$$

where $\varepsilon > 0$ is a regularization parameter and $\|\cdot\|_2$ denotes the Euclidean 2-norm. In the language of optimization and control (see, e.g., [6]), x_1 and x_2 are the *state* variables and u is the *distributed control* variable. We first show that (3) is well-posed, i.e., that it has a unique solution $\{x_1^\varepsilon, x_2^\varepsilon, u^\varepsilon\}$. Then we examine the connection between $\{x_1^\varepsilon, x_2^\varepsilon\}$ and the solution x of the original equation (1).

3.1 Existence of Optimal Solutions

The first-order necessary conditions for $\{x_1^\varepsilon, x_2^\varepsilon, u^\varepsilon\}$ to be a solution of (3) state that there is a Lagrange multiplier vector pair $\{\lambda_1^\varepsilon, \lambda_2^\varepsilon\}$ such that the *KKT system* of equations is satisfied,

$$\begin{pmatrix} I & -I & 0 & A_1^T & 0 \\ -I & I & 0 & 0 & A_2^T \\ 0 & 0 & \varepsilon I & -I & I \\ A_1 & 0 & -I & 0 & 0 \\ 0 & A_2 & I & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1^\varepsilon \\ x_2^\varepsilon \\ u^\varepsilon \\ \lambda_1^\varepsilon \\ \lambda_2^\varepsilon \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ b \\ 0 \end{pmatrix}, \quad (4)$$

where I is the $n \times n$ identity matrix. In the following, we let H denote the $3n \times 3n$ (full) Hessian matrix and Z the $3n \times n$ null-space matrix, respectively,

$$H = \begin{pmatrix} I & -I & 0 \\ -I & I & 0 \\ 0 & 0 & \varepsilon I \end{pmatrix}, \quad Z = \begin{pmatrix} A_1^{-1} \\ -A_2^{-1} \\ I \end{pmatrix},$$

with the property

$$\begin{pmatrix} A_1 & 0 & -I \\ 0 & A_2 & I \end{pmatrix} Z = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Lemma 1. *The matrix $\widehat{H} = Z^T H Z$, known as the reduced Hessian, is symmetric positive definite.*

Proof. A straightforward calculation yields $\widehat{H} = (A_1^{-1} + A_2^{-1})^T (A_1^{-1} + A_2^{-1}) + \varepsilon I$. The matrix $(A_1^{-1} + A_2^{-1})^T (A_1^{-1} + A_2^{-1})$ is (at least) symmetric positive semidefinite, while εI is symmetric positive definite. The claim follows.

Corollary 1. *Invertibility of A_1 and A_2 and Lemma 1 directly imply that the KKT matrix defined in (4) is nonsingular, and that there is a unique pair $\{\{x_1^\varepsilon, x_2^\varepsilon, u^\varepsilon\}, \{\lambda_1^\varepsilon, \lambda_2^\varepsilon\}\}$ satisfying (4). Furthermore, the vector triple $\{x_1^\varepsilon, x_2^\varepsilon, u^\varepsilon\}$ is the unique global solution of the optimization problem (3). For a proof, see [7, p.444-447].*

Remark 1. As we will show below, the matrix $(A_1^{-1} + A_2^{-1})^T (A_1^{-1} + A_2^{-1})$ is in fact symmetric positive definite. Hence the regularization term $\varepsilon \|u\|_2^2$ in (3), which is typically needed to guarantee existence and uniqueness of optimal solutions, is seemingly superfluous. However, one should not forget that we think of (1) as

resulting from discretization of a PDE, i.e., that we are dealing with a family of linear systems parametrized by some measure h of the mesh size, instead of with a single linear system. In this case the regularization term is needed to guarantee the uniform in h invertibility of A ; see [8].

3.2 Reformulation Error

In general, as $\varepsilon > 0$, the state solutions x_1^ε and x_2^ε of (3) will differ from the solution x of the original problem (1). In this section we calculate the error induced by ε .

Lemma 2. *The matrix $(A_1^{-1} + A_2^{-1})$ is nonsingular, with the inverse given by $A_1 - A_1 A^{-1} A_1$. (It follows trivially that the matrix $(A_1^{-1} + A_2^{-1})^T (A_1^{-1} + A_2^{-1})$ is symmetric positive definite.)*

Proof. By inspection,

$$\begin{aligned} (A_1^{-1} + A_2^{-1})(A_1 - A_1 A^{-1} A_1) &= I + A_2^{-1} A_1 - A^{-1} A_1 - A_2^{-1} A_1 A^{-1} A_1 \\ &= I + [A_2^{-1} - (A_1 + A_2)^{-1} - A_2^{-1} A_1 (A_1 + A_2)^{-1}] A_1 \\ &= I + [A_2^{-1} (A_1 + A_2) - I - A_2^{-1} A_1] (A_1 + A_2)^{-1} A_1 = I. \end{aligned}$$

Lemma 3. *The following statements hold:*

$$x_1^\varepsilon - x = \varepsilon A_1^{-1} \widehat{H}^{-1} (I - A_1 A^{-1}) b, \quad (5)$$

and

$$x_2^\varepsilon - x = -\varepsilon A_2^{-1} \widehat{H}^{-1} A_2 A^{-1} b. \quad (6)$$

Proof. We present a proof for (6), statement (5) can be verified analogously. The KKT system (4) yields expressions for the states,

$$x_1^\varepsilon = A_1^{-1} (u^\varepsilon + b), \quad x_2^\varepsilon = -A_2^{-1} u^\varepsilon, \quad (7)$$

the Lagrange multipliers,

$$\lambda_1^\varepsilon = -A_1^{-T} (x_1^\varepsilon - x_2^\varepsilon), \quad \lambda_2^\varepsilon = A_2^{-T} (x_1^\varepsilon - x_2^\varepsilon),$$

and the controls,

$$u^\varepsilon = (1/\varepsilon)(\lambda_1^\varepsilon - \lambda_2^\varepsilon).$$

By substitution, we obtain the so-called *reduced system* for the controls,

$$((A_1^{-1} + A_2^{-1})^T (A_1^{-1} + A_2^{-1}) + \varepsilon I) u^\varepsilon = -(A_1^{-1} + A_2^{-1})^T A_1^{-1} b, \quad (8)$$

which using the notation for the reduced Hessian \widehat{H} implies

$$\begin{aligned} x_2^\varepsilon - x &= [A_2^{-1} \widehat{H}^{-1} (A_1^{-1} + A_2^{-1})^T A_1^{-1} - A^{-1}] b \\ &= A_2^{-1} \widehat{H}^{-1} [(A_1^{-1} + A_2^{-1})^T A_1^{-1} \\ &\quad - ((A_1^{-1} + A_2^{-1})^T (A_1^{-1} + A_2^{-1}) + \varepsilon I) A_2 A^{-1}] b \\ &= A_2^{-1} \widehat{H}^{-1} [(A_1^{-1} + A_2^{-1})^T (A_1^{-1} - (A_1^{-1} + A_2^{-1}) A_2 A^{-1}) - \varepsilon A_2 A^{-1}] b \\ &= A_2^{-1} \widehat{H}^{-1} [(A_1^{-1} + A_2^{-1})^T (A_1^{-1} (A_1 + A_2) - (A_1^{-1} + A_2^{-1}) A_2) A^{-1} \\ &\quad - \varepsilon A_2 A^{-1}] b = -\varepsilon A_2^{-1} \widehat{H}^{-1} A_2 A^{-1} b. \end{aligned}$$

Lemma 4. Let $M = (A_1^{-1} + A_2^{-1})^T (A_1^{-1} + A_2^{-1})$ and let Λ_{\min} denote the smallest eigenvalue of M . Let $\varepsilon < (\Lambda_{\min}/2)$. Then,

$$\|\widehat{H}^{-1}\|_2 \leq \frac{2}{\Lambda_{\min}}.$$

Proof. We have $\widehat{H} = M(I - (-\varepsilon M^{-1}))$. Due to $\varepsilon < (\Lambda_{\min}/2)$ and e.g. [9, Lemma 2.3.3], the matrix $(I - (-\varepsilon M^{-1}))$ is nonsingular, hence we can write $\widehat{H}^{-1} = (I - (-\varepsilon M^{-1}))^{-1} M^{-1}$, which implies

$$\begin{aligned} \|\widehat{H}^{-1}\|_2 &\leq \frac{1}{\Lambda_{\min}} \left\| (I - (-\varepsilon M^{-1}))^{-1} \right\|_2 \leq \frac{1}{\Lambda_{\min}(1 - \|\varepsilon M^{-1}\|_2)} \\ &= \frac{1}{\Lambda_{\min}(1 - (\varepsilon/\Lambda_{\min}))} = \frac{1}{\Lambda_{\min} - \varepsilon} \leq \frac{2}{\Lambda_{\min}}. \end{aligned}$$

Theorem 1. Let the assumptions of Lemma 4 be satisfied. There exists a constant C , independent of ε , such that

$$\|x_1^\varepsilon - x\|_2 + \|x_2^\varepsilon - x\|_2 \leq \varepsilon C \|b\|_2.$$

Proof. Lemmas 3 and 4 yield the claim directly, with

$$C = \frac{2}{\Lambda_{\min}} \left(\|A_1^{-1}\|_2 \|(I - A_1 A^{-1})\|_2 + \|A_2^{-1}\|_2 \|A_2\|_2 \|A^{-1}\|_2 \right).$$

Remark 2. For parametrized linear systems (1) corresponding to discretized PDEs the constant C may depend on the mesh size h .

3.3 A Solution Algorithm

We now exploit the structure of the reformulated problem to develop robust solution methods for (1). Our approach is based on the premise that robust solution methods for linear systems involving A_1 and A_2 are readily available.

We focus on a solution method known as the *null-space* or *reduced-space* approach, which effectively decouples the component equations in (4). In this approach one typically solves the reduced system (8) iteratively, using a Krylov subspace method. The main computational cost of such an approach is in the repeated application of the reduced Hessian operator, which we specify below.

A similar procedure can be derived for the (one-time) computation of the right-hand side in (8). Optimal states can be recovered by solving either of the equations given in (7).

4 Numerical Results

In this section we apply the optimization-based reformulation to an SUPG-stabilized [10] discretization of the scalar convection-diffusion elliptic problem

$$-\nu \Delta u + \mathbf{c} \cdot \nabla u = f \text{ in } \Omega, \quad u = u_D \text{ on } \Gamma_D, \quad \nabla u \cdot \mathbf{n} = 0 \text{ on } \Gamma_N. \quad (9)$$

Algorithm 1: Application of reduced Hessian \widehat{H} to vector u .

input : vector u
output: vector $\widehat{H}u$

- 1 Solve: $A_1 y_1 = u, A_2 y_2 = u$ (state equations)
- 2 Compute: $y_3 = y_1 + y_2$
- 3 Solve: $A_1^T y_4 = y_3, A_2^T y_5 = y_3$ (adjoint equations)
- 4 Compute: $\widehat{H}u = y_4 + y_5 + \varepsilon u$

We assume that Ω is a bounded open domain in \mathbb{R}^d , $d = 2, 3$, with the Lipschitz continuous boundary $\partial\Omega = \Gamma_D \cup \Gamma_N$, \mathbf{c} is a given velocity field with $\nabla \cdot \mathbf{c} = 0$, u_D is a given Dirichlet boundary function, \mathbf{n} is the outward unit normal, and $\nu > 0$ is a constant diffusion coefficient. We focus on settings where ν is small compared to $|\mathbf{c}|$, i.e., when (9) is convection-dominated.

The linear system resulting from (9) is typically of the form

$$(\nu D + C)x = b, \quad (10)$$

where D is a (pure) diffusion matrix (discretization of the Laplace operator), C corresponds to the convection contribution (including any terms due to the SUPG stabilization), and b stands for a discretization of the source term f (plus stabilization terms). In order to apply our optimization-based approach to (10), we make the identification

$$A_1 = D + C, \quad \text{and} \quad A_2 = (\nu - 1)D. \quad (11)$$

The key property of this decomposition is that, unlike the original operator $(\nu D + C)$, the matrices A_1 and A_2 are *diffusion-dominated*.

The scalable iterative solution of linear systems arising from PDEs of the type (9) is a very active area of research. Algebraic multigrid methods (see [11] and references therein) work well for diffusion-dominated problems but their performance degrades in the convection-dominated case. The efforts to extend multigrid methods to such problems [12–15] have led to improvements, albeit at the expense of increased algorithm complexity. As we show below, widely used “off-the-shelf” multigrid solvers, such as *BoomerAMG* (*hypra* library) [16] or *ML* (Trilinos project) [17], can lack robustness when applied to problems with complex convection fields, particularly in the case of very large Péclet numbers.

We consider the so-called double-glazing problem, see [18, p.119]. The domain is given by $\Omega = [-1, 1]^2$, subject to a uniform triangular partition, generated by dividing each square of a uniform square partition of Ω into two triangles, diagonally from the bottom left to the top right corner. The convection field is given by $\mathbf{c} = (2y(1 - x^2), -2x(1 - y^2))$, and the diffusion constant is set to $\nu = 10^{-8}$. Boundary conditions are specified by

$$\Gamma_N = \emptyset, \quad \begin{cases} u_D = 0 & \text{on } \{-1, 1\} \times \{-1\} \cup \{-1\} \times (-1, 1) \cup \{-1, 1\} \times \{1\}, \\ u_D = 1 & \text{on } \{1\} \times [-1, 1]. \end{cases}$$

We compare the solution of linear systems arising from the full convection–diffusion problem (10), denoted the *full problem*, to the solution of the optimization reformulation (3) with A_1 and A_2 defined in (11). For the solution of (10) we use the multigrid solvers BoomerAMG and ML as preconditioners for the GMRES method with a maximum Krylov subspace size of 200, denoted by GMRES(200). Solver abbreviations, options, and failure codes used hereafter are summarized below. We note that the stated ML^{LU} and BAMG parameters reflect the best solver settings that we could find for the example problem.

ML^{LU}	ML: incomplete LU smoother (IFPACK, ILU, <code>threshold=1.05</code>), W cycle
BAMG	BoomerAMG: Falgout–CLJP coarsening (6), symmetric Gauss–Seidel / Jacobi hybrid relaxation (6), V cycle (1)
— ^{MX}	exceeded maximum number of GMRES iterations (2000)

The optimization reformulation is solved via the reduced–space approach, denoted here by OPT. The outer optimization loop amounts to solving the linear system (8) using unpreconditioned GMRES(200). Within every optimization iteration, we solve four linear systems, see Algorithm 1. The linear systems involving A_1 and A_2 are solved using GMRES(10), preconditioned with ML with a simple smoother, to a relative stopping tolerance of 10^{-10} . We remark that in either case only 5–8 iterations are required for the solution of each system. The regularization parameter ε is set to 10^{-12} .

Table 1. Number of **outer** GMRES(200) iterations for the optimization approach; **total** number of GMRES(200) iterations for multigrid applied to the full problem.

	$\nu = 10^{-8}$			$\nu = 10^{-2}$	128×128	
	64×64	128×128	256×256		$\nu = 10^{-4}$	$\nu = 10^{-8}$
OPT	114	114	113	84	114	114
ML	71	196	— ^{MX}	9	96	196
BAMG	72	457	— ^{MX}	7	33	457

Table 1 presents a comparison of the number of outer GMRES(200) iterations for the optimization approach and the total number of GMRES(200) iterations for multigrid solvers applied to the full problem. Relative stopping tolerances are set to 10^{-6} . For our test problem, ML^{LU} and BAMG show very strong mesh dependence, and eventually fail to converge on the 256×256 mesh. OPT, on the other hand, is robust to mesh refinement, and successfully solves the problem for all mesh sizes. In addition, Table 1 clearly demonstrates that while ML^{LU} and BAMG are very sensitive to the size of the Péclet number, OPT’s performance is affected only mildly, and in fact does not change when the diffusion constant is lowered from $\nu = 10^{-4}$ to $\nu = 10^{-8}$. Overall, our optimization–based strat-

egy provides a robust solution alternative for problems on which widely used multigrid solvers struggle.

References

1. Strang, G.: On the construction and comparison of difference schemes. *SIAM J. Numer. Anal.* **5** (1968) 506–517
2. Toselli, A., Widlund, O.: *Domain Decomposition Methods - Algorithms and Theory*. Springer Verlag, New York (2005)
3. Lions, J.L.: Virtual and effective control for distributed systems and decomposition of everything. *Journal d'Analyse Mathématique* **80** (2000) 257–297
4. Du, Q., Gunzburger, M.: A gradient method approach to optimization-based multidisciplinary simulations and nonoverlapping domain decomposition algorithms. *SIAM J. Numer. Anal.* **37** (2000) 1513–1541
5. Gunzburger, M., Lee, H.K.: An optimization-based domain decomposition method for the Navier-Stokes equations. *SIAM J. Numer. Anal.* **37** (2000) 1455–1480
6. Gunzburger, M.D.: Perspectives in flow control and optimization. *Advances in Design and Control*. SIAM (2003)
7. Nocedal, J., Wright, S.J.: *Numerical Optimization*. Springer Verlag, Berlin, Heidelberg, New York (1999)
8. Bochev, P., Ridzal, D.: An optimization-based approach for the design of robust solution algorithms. *SIAM J. Numer. Anal.* (Submitted)
9. Golub, G.H., van Loan, C.F.: *Matrix Computations*. third edn. Johns Hopkins University Press, Baltimore, London (1996)
10. Hughes, T.J.R., Brooks, A.: A theoretical framework for Petrov-Galerkin methods with discontinuous weighting functions: Application to the streamline-upwind procedure. In et al, R.H.G., ed.: *Finite Elements in Fluids*. Volume 4., New York, J. Wiley & Sons (1982) 47–65
11. Stüben, K.: A review of algebraic multigrid. *J. Comp. Appl. Math.* **128**(1-2) (2001) 281–309
12. Bank, R.E., Wan, J.W.L., Qu, Z.: Kernel preserving multigrid methods for convection-diffusion equations. *SIAM J. Matrix Anal. Appl.* **27**(4) (2006) 1150–1171
13. Brezina, M., Falgout, R., MacLachlan, S., Manteuffel, T., McCormick, S., Ruge, J.: Adaptive smoothed aggregation (asa) multigrid. *SIAM Review* **47**(2) (2005) 317–346
14. Notay, Y.: Aggregation-based algebraic multilevel preconditioning. *SIAM J. Matrix Anal. Appl.* **27**(4) (2006) 998–1018
15. Sala, M., Tuminaro, R.S.: A new Petrov-Galerkin smoothed aggregation preconditioner for nonsymmetric linear systems. *SIAM J. Sci. Comp.* **31**(1) (2008) 143–166
16. Henson, V.E., Meier Yang, U.: Boomeramg: a parallel algebraic multigrid solver and preconditioner. *Appl. Numer. Math.* **41**(1) (April 2002) 155–177
17. Gee, M.W., Siefert, C.M., Hu, J.J., Tuminaro, R.S., Sala, M.G.: *ML 5.0 Smoothed Aggregation User's Guide*. Technical Report SAND2006-2649, Sandia National Laboratories, Albuquerque, NM (2006)
18. Elman, H., Silvester, D., Wathen, A.: *Finite Elements and Fast Iterative Solvers*. Oxford University Press (2005)