A new fast method to compute saddle-points in constrained optimization and applications

Philippe Angot, Jean-Paul Caltagirone, Pierre Fabrie

Accepted Manuscript

A new fast method to compute saddle-points in constrained optimization and applications

Philippe Angot, Jean-Paul Caltagirone, Pierre Fabrie

PII: S0893-9659(11)00378-8
DOI: 10.1016/j.aml.2011.08.015
Reference: AML 3762

To appear in: *Applied Mathematics Letters*

Received date: 1 January 2011
Revised date: 21 April 2011
Accepted date: 11 August 2011

Please cite this article as: P. Angot, J.-P. Caltagirone, P. Fabrie, A new fast method to compute saddle-points in constrained optimization and applications, *Appl. Math. Lett.* (2011), doi:10.1016/j.aml.2011.08.015

Applied Mathematics Letters, 2011 (revised version).

A new fast method to compute saddle-points in constrained optimization and applications

Philippe Angot[a], Jean-Paul Caltagirone[b], and Pierre Fabrie[c]

[a]Aix-Marseille Université, LATP - CMI UMR CNRS 6632, 39 rue F. Joliot Curie, 13453 Marseille Cedex 13 - France.
[b]Université de Bordeaux & IPB, TREFLE UMR CNRS 8508, ENSEIRB-MATMECA, Talence - France.
[c]Université de Bordeaux & IPB, Institut Mathématiques de Bordeaux UMR CNRS 5251, ENSEIRB-MATMECA, Talence - France.

Abstract

The solution of the augmented Lagrangian related system $(A + r\,B^T B)\,u = f$ is a key ingredient of many iterative algorithms for the solution of saddle-point problems in constrained optimization with quasi-Newton methods. However, such problems are ill-conditioned when the penalty parameter $\varepsilon = 1/r > 0$ tends to zero, whereas the error vanishes as $O(\varepsilon)$. We present a new fast method based on a *splitting penalty scheme* to solve such problems with a judicious prediction-correction. We prove that, due to the *adapted right-hand side*, the solution of the correction step only requires the approximation of operators independent on ε, when ε is taken sufficiently small. Hence, the proposed method is all the cheaper as ε tends to zero. We apply the two-step scheme to efficiently solve the saddle-point problem with a penalty method. Indeed, that fully justifies the interest of the *vector penalty-projection methods* recently proposed in [1] to solve the unsteady incompressible Navier-Stokes equations, for which we give the stability result and some quasi-optimal error estimates. Moreover, the numerical experiments confirm both the theoretical analysis and the efficiency of the proposed method which produces a fast splitting solution to augmented Lagrangian or penalty problems, possibly used as a suitable preconditioner to the fully coupled system.

Keywords: Constrained optimization, Saddle-point problems, Augmented Lagrangian, Penalty method, Splitting prediction-correction scheme, Vector penalty-projection methods
2010 MSC: 49K35, 65F05, 65F08, 65F10, 65F35, 65K05, 65K10, 65K15, 65N22, 76D05, 76D07, 90C25, 90C26

1. The prediction-correction method for augmented Lagrangian problems

By using the ordinary Lagrangian, a nonconvex optimization problem often has a duality gap and the value of the dual problem is strictly less than the value of the primal problem. A common strategy for bridging this gap is to augment the ordinary Lagrangian with a penalty term, see [7, 20]. Besides, Bertsekas [7] observes that such a result also applies to inequality constrained problems, since an inequality can be made equivalent to an equality. Then, by using quasi-Newton schemes, Daniel, Fletcher-Reeves or Polak-Ribière formulations of conjugate gradient algorithm, the solution to a linear equality constrained problem with the augmented Lagrangian proves to be a fundamental ingredient of locally quadratically convergent methods for optimization problems with both equality or inequality constraints even in the nonconvex case; see also [12, 14, 20] and barrier or interior point methods in [22].

We now focus on the augmented Lagrangian problem for a linear equality constraint. Let us address the solution to the linear equality constrained problem in the quadratic convex case for sake of simplicity. Let A be an $n \times n$ symmetric positive definite matrix and $b \in \mathbb{R}^n$ defining the strictly convex functional $F(x) = x^T A\,x - 2b^T x$. Let B be an $m \times n$ matrix and $d \in \mathbb{R}^m$ a vector in $\mathrm{Im}(B)$, the range of B, which define the closed convex subset of constraint:

Email address: angot@cmi.univ-mrs.fr (**Philippe Angot**)
URL: http://www.latp.univ-mrs.fr/~angot (**Philippe Angot**)

$K = \{x \in \mathbb{R}^n;\ G(x) := Bx - d = 0\}$. Then, we recall that there exists a unique global minimizer $x^* \in K$ for F on K, solution to: $F(x^*) = \min_{x \in K} F(x)$ and $(x^*, \lambda^*) \in \mathbb{R}^n \times \mathbb{R}^m$ is solution to the linear block-system of order $n + m$ below [6]:

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} x^* \\ \lambda^* \end{pmatrix} = \begin{pmatrix} b \\ d \end{pmatrix}. \tag{1}$$

Moreover, if B is surjective, *i.e.* B has the maximal rank m, then λ^* solution to (1) is unique. The fully coupled solvers of (1) require Krylov gradient methods with suitable preconditioners, e.g. [9]. Then, the usual approximate solution to (1) with the Uzawa augmented Lagrangian iterations, *i.e.* a gradient type method, see [14] or [19] for applications with finite volume methods, reads with λ^0 and $0 < \rho_k < 2r_k$ given:

$$\left(A + r_k B^T B\right) x^{k+1} = b + r_k B^T d - B^T \lambda^k, \tag{2}$$

$$\lambda^{k+1} = \lambda^k + \rho_k (B x^{k+1} - d), \quad \forall k \in \mathbb{N}. \tag{3}$$

The conjugate gradient can be used as well instead of the above steepest descent, see [14]. However, a common criticism of the augmented Lagrangian approach is that the system (2) can be ill-conditioned due to the penalty term and iterative algorithms for solving the problem converge very slowly as the augmentation parameter r_k tends to infinity. Indeed, the convergence rate of gradient methods for solving (2) is governed by the condition number $\mathrm{cond}(A_r) = \mu_n/\mu_1$, the ratio between the largest eigenvalue μ_n and the smallest one μ_1 of the matrix $A_r := A + r B^T B$. For example, the error $x_l - \bar{x}$ of the conjugate gradient iteration x_l satisfies the energy inequality below, e.g. [17]:

$$E(x_{l+1} - \bar{x}) \leq \left(\frac{\sqrt{\mathrm{cond}(A_r)} - 1}{\sqrt{\mathrm{cond}(A_r)} + 1}\right)^2 E(x_l - \bar{x}),\ \forall l \in \mathbb{N} \quad \text{where} \quad E(x) = x^T (A + r B^T B)\, x. \tag{4}$$

It is shown in [14, 20] that $\mathrm{cond}(A_r) = O(r)$, which is observed numerically [11], and the iterative algorithm converges arbitrarily slowly as r tends to infinity. Thus, techniques for handling the instabilities are developed: preconditioning and/or adjustments with a variable parameter r with criterion for deciding when to increase it, e.g. [12, 16, 14, 20] or also [15, Chap. 4].

Here, we present a new and simple splitting solution to the augmented Lagrangian system (AL) or (2) with $f = b - B^T \lambda^k$:

$$(AL) \quad \left(A + \frac{1}{\varepsilon} B^T B\right) u_\varepsilon = f + \frac{1}{\varepsilon} B^T d, \quad f \in \mathbb{R}^n \qquad \text{with} \quad 0 < \varepsilon = \frac{1}{r} \ll 1 \tag{5}$$

with the following prediction-correction scheme ($AL2$) where the constraint is handled only in the correction step:

$$A \tilde{u} = f, \tag{6}$$

$$(AL2) \quad \left(A + \frac{1}{\varepsilon} B^T B\right) \hat{u}_\varepsilon = -\frac{1}{\varepsilon} B^T (B \tilde{u} - d), \quad \text{and} \quad u_\varepsilon = \tilde{u} + \hat{u}_\varepsilon. \tag{7}$$

We observe that (6,7) is equivalent to (5) and we prove the crucial result below due to the *adapted right-hand side* in the correction step (7) which lies in the range of the operator B^T or of the limit operator $B^T B$. Indeed, (7) can be viewed as a singular perturbation problem with well-suited data in the right-hand side. More precisely, we give in Theorem 1.1 the asymptotic expansion of the solution $\hat{u}_\varepsilon$ to (7):

$$\hat{u}_\varepsilon = -\frac{1}{\varepsilon} \left(A + \frac{1}{\varepsilon} B^T B\right)^{-1} B^T (B \tilde{u} - d) \tag{8}$$

when the penalty parameter ε is chosen sufficiently small. We denote by $\|.\|$ the Euclidean norm.

Theorem 1.1 (Solution of the splitting augmented Lagrangian system)**.** *Let A be an $n \times n$ positive definite matrix and B an $m \times n$ matrix. If the rows of B are linearly independent, $\mathrm{rank}(B) = m$, then for ε small enough, $0 < \varepsilon < 1/\|S^{-1}\|$ where $S = BA^{-1}B^T$ ($-S$ being the Schur complement of A), there exists an $n \times m$ matrix E given in (12) and bounded independently on ε such that the solution of the correction step (8) writes with $\tilde{u} = A^{-1} f$:*

$$\hat{u}_\varepsilon = (C + \varepsilon E)(B \tilde{u} - d) = C_0 \tilde{u} - C d + \varepsilon E (B \tilde{u} - d) \quad \text{where} \quad C = -A^{-1} B^T S^{-1}, \quad C_0 = -A^{-1} B^T S^{-1} B. \tag{9}$$

If $\mathrm{rank}(B) = p < m$, there exists a surjective $p \times n$ matrix T such that $B^T B = T^T T$ and the similar result holds replacing B by T.

Proof. We first remove redundant rows from B. If the rank of B is $p < m$, then by the QR factorization [17], there exists an orthogonal $m \times m$ matrix Q such that $B = QR$, where the first p rows of R are linearly independent and the next $m - p$ rows are completely zero. Letting T be the submatrix of R formed by the first p rows, we have: $B^T B = R^T Q^T QR = R^T R = T^T T$. Since $B^T B = T^T T$ and the rows of T are linearly independent, there is no loss of generality in assuming that the rows of B are linearly independent.

Let us now prove the main result. By using the Woodbury formula [24, 21], a generalization of the Sherman-Morrison formula [17, Chap. 2], we can express $(A + r\,B^T B)^{-1}$ as:

$$\left(A + \frac{1}{\varepsilon} B^T B\right)^{-1} = A^{-1} - A^{-1} B^T \left(\varepsilon\, I + B A^{-1} B^T\right)^{-1} B A^{-1}, \quad \text{for all } \varepsilon = \frac{1}{r} > 0. \tag{10}$$

Since A is positive definite, A is nonsingular and A^{-1} also positive definite; since $\mathrm{rank}(B^T) = \mathrm{rank}(B) = m$, we have $\ker(B^T) = \{0\}$. Thus, the Lagrange multiplier operator $S = BA^{-1}B^T$ is nonsingular.

Now, writing $(\varepsilon\, I + S)^{-1} = (I + \varepsilon\, S^{-1})^{-1} S^{-1}$, we can expand this inverse matrix in the Neumann geometric series if ε is sufficiently small, e.g. $\varepsilon < 1/\|S^{-1}\|$. Thus, with $\varepsilon \le (1 - \xi)/\|S^{-1}\|$ for any $\xi > 0$, we have:

$$\left(\varepsilon\, I + B A^{-1} B^T\right)^{-1} = S^{-1} - \varepsilon\, S^{-2} + \sum_{k=2}^{\infty} (-1)^k\, \varepsilon^k\, S^{-k-1} \quad \text{for } \varepsilon \le (1 - \xi)/\|S^{-1}\|,\ \forall \xi > 0.$$

Combining with (10), we get by a simple calculation the asymptotic expansion below:

$$\left(A + \frac{1}{\varepsilon} B^T B\right)^{-1} = A^{-1} - A^{-1} B^T S^{-1} B A^{-1} + \varepsilon\, A^{-1} B^T S^{-2} B A^{-1} - \sum_{k=2}^{\infty} (-1)^k\, \varepsilon^k\, A^{-1} B^T S^{-k-1} B A^{-1}. \tag{11}$$

Then, multiplying (11) by the right-hand side in (7), observing that the coefficient of the $1/\varepsilon$ term is zero and the coefficient of the ε^0 term is $C(B\,\tilde{u} - d)$ with $C = -A^{-1} B^T S^{-1}$, it yields (9) where

$$E = A^{-1} B^T S^{-2} \sum_{k=0}^{\infty} (-1)^k\, \varepsilon^k\, S^{-k} = A^{-1} B^T S^{-2} \left(I + \varepsilon\, S^{-1}\right)^{-1} \quad \text{with} \quad \|E\| \le \|A^{-1} B^T S^{-2}\|\, \xi^{-1} \tag{12}$$

which completes the proof since: $\|(I + \varepsilon\, S^{-1})^{-1}\| \le (1 - \varepsilon\, \|S^{-1}\|)^{-1} \le \xi^{-1}$. □

Hence, for ε small enough, the computational effort required to solve (7) amounts to approximate the matrices C_0 or C. Moreover, the following corollaries can be easily proved, showing that the estimate (4) is far from being optimal as far as the right-hand side in (7) is adapted to the left-hand side operator.

Corollary 1.2 (Adapted conditioning property)**.** *In the solution procedure for (7), assume that some perturbations exist: either $\tilde{u} + \delta\tilde{u}$ in $\tilde{u} \neq 0$ or $C_0 + \delta C_0$ in C_0 from Theorem 1.1. Then, the perturbed solution $\hat{u}_\varepsilon + \delta\hat{u}_\varepsilon$ respectively satisfies for ε sufficiently small, with $H = E\,B$:*

$$\frac{\|\delta\hat{u}_\varepsilon\|}{\|\hat{u}_\varepsilon\|} \le \|C_0 + \varepsilon\, H\|\, \|(C_0 + \varepsilon\, H)^{-1}\|\, \frac{\|\delta\tilde{u}\|}{\|\tilde{u}\|} \quad \text{or} \quad \frac{\|\delta\hat{u}_\varepsilon\|}{\|\hat{u}_\varepsilon\|} \le \|C_0\|\, \|(C_0 + \varepsilon\, H)^{-1}\|\, \frac{\|\delta C_0\|}{\|C_0\|}. \tag{13}$$

This defines the effective condition number for the linear system (7) by $\mathrm{cond}_e = \|C_0\|\, \|(C_0 + \varepsilon\, H)^{-1}\|$ for ε small enough.

Corollary 1.3 (Fast solution for $A = I \pm \varepsilon\, M$ or $A = BB^T$ and $d = 0$)**.** *Assume the framework of Theorem 1.1 with $A = I$ the $n \times n$ Identity matrix and $d = 0$. Then for all ε small enough: $0 < \varepsilon < 1/\|S^{-1}\|$ where $S = BB^T$, we have:*

$$\hat{u}_\varepsilon = C_0\, \tilde{u} + \varepsilon\, C_1\, \tilde{u} \quad \text{where} \quad C_0 = -B^T S^{-1} B = -B^T (BB^T)^{-1} B, \quad C_1 = EB = B^T (BB^T)^{-2} \left(I + \varepsilon\, (BB^T)^{-1}\right)^{-1} B. \tag{14}$$

Moreover, if $\mathrm{rank}(B) = p \le m \le n$, the zero-order solution $\hat{u} = C_0\, \tilde{u}$ in (14) is the solution of minimal Euclidean norm in $\mathbb{R}^n$ to the linear system: $B\,\hat{u} = -B\,\tilde{u}$ by the least-squares method, and the matrix $B^\dagger = B^T (BB^T)^{-1}$ is the Moore-Penrose pseudo-inverse of B such that $C_0 = -B^\dagger B$. Indeed, a singular value decomposition (SVD) or a QR factorization of B yields: $C_0 = -I_0$ where I_0 is the $n \times n$ diagonal matrix having only 1 or 0 coefficients, the zero entries in the diagonal being the $n - p$ null eigenvalues of the operator $B^T B$.

The same result also holds with any perturbation of Identity $A = I \pm \varepsilon\, M$, whatever the $n \times n$ matrix M, if $0 < \varepsilon < \min(\|M\|,\ 1/\|S^{-1}\|)$.

If $A = BB^T$ with $m = n$ and B nonsingular, a similar result also holds, i.e. *we get: $C_0 = -I$.*

2. The splitting penalty method for saddle-point problems

We now illustrate the splitting augmented Lagrangian method by applying the two-step scheme to solve saddle-point problems for continuous or discrete operators with a penalty method. For sake of simplicity here, we restrict ourselves to the Hilbertian framework although the result can be extended to reflexive Banach spaces.

2.1. The two-step penalty augmented Lagrangian method

Let V and X be two Hilbert spaces and V', X' the dual spaces with $\langle .,. \rangle$ denoting the duality pairing. Introduce the linear and continuous (bounded) operators A and B such that $A:\ V \to V'$, $B:\ V \to X'$ and thus $B^T:\ X \simeq X'' \to V'$. For $f \in V'$ and $g \in X'$, we consider the abstract saddle-point problem:

$$\text{seek } (u,p) \in V \times X \quad \text{such that} \quad \left\{ \begin{array}{ll} A\,u + B^T p & = f, \\ B\,u & = g. \end{array} \right. \tag{15}$$

Assume that the operator A is coercive on V and that the *inf-sup* condition holds, *i.e.*

$$(i) \quad \exists\, \alpha > 0,\ \langle Au, u \rangle_{V',V} \geq \alpha\, \|u\|_V^2,\ \forall u \in V \qquad (ii) \quad \exists \beta > 0,\ \sup_{w \in V} \frac{\langle Bw, q \rangle_{X',X}}{\|w\|_V} \geq \beta\, \|q\|_X,\ \forall q \in X. \tag{16}$$

Then, it is well-known that the problem (15) is well-posed with (16): there exists a unique solution $(u,p) \in V \times X$ which continuously depends on the data f and g, see [13].

Let us now consider the penalty method, originally introduced by Courant [10] in the context of constrained optimization, for the approximate solution of problem (15) where X and X' are identified using the Riesz-Fréchet representation theorem. For all $\varepsilon > 0$, seek $u_\varepsilon \in V$ and $p_\varepsilon \in X$ such that:

$$(AL) \quad \left\{ \begin{array}{l} A\,u_\varepsilon + B^T p_\varepsilon = f, \\ p_\varepsilon = \dfrac{1}{\varepsilon}(B\,u_\varepsilon - g) \end{array} \right. \quad \Leftrightarrow \quad \left\{ \begin{array}{l} \left(A + \dfrac{1}{\varepsilon} B^T B\right) u_\varepsilon = f + \dfrac{1}{\varepsilon} B^T g, \\ p_\varepsilon = \dfrac{1}{\varepsilon}(B\,u_\varepsilon - g). \end{array} \right. \tag{17}$$

The corresponding augmented Lagrangian problem can be efficiently solved by the prediction-correction scheme for the penalty augmented Lagrangian method to get the solution of (17) with $0 < \varepsilon \ll 1$:

$$(AL2) \quad \left\{ \begin{array}{l} A\,\tilde{u} = f, \\ \left(A + \dfrac{1}{\varepsilon} B^T B\right) \hat{u}_\varepsilon = -\dfrac{1}{\varepsilon} B^T (B\,\tilde{u} - g), \\ u_\varepsilon = \tilde{u} + \hat{u}_\varepsilon \quad \text{and} \quad p_\varepsilon = \dfrac{1}{\varepsilon}(B\,u_\varepsilon - g). \end{array} \right. \tag{18}$$

For numerical applications, V and X are finite-dimensional spaces and the two-step scheme $(AL2)$ is all the cheaper as the penalty parameter ε tends to zero, as proved in Section 1. Moreover, (18) yields an $O(\varepsilon)$ accurate approximation of the saddle-point solution, as stated below where the proof, slightly different from [13], does not require the smallness of ε; see also the case of the Stokes problem in [23].

Theorem 2.1 (Error estimate of the penalty method for saddle-point problems)**.** *Under the above framework, there exists $c = c(\|A\|, \alpha, \beta) > 0$ such that the following error estimate holds for all $\varepsilon > 0$:*

$$\|u_\varepsilon - u\|_V + \|p_\varepsilon - p\|_X + \|B(u_\varepsilon - u)\|_X \leq c(\|A\|, \alpha, \beta)\ (\|f\|_{V'} + \|g\|_X)\ \varepsilon. \tag{19}$$

Sketch of proof. Indeed, (18) and (17) are equivalent and the error equation with (15) writes:

$$A(u_\varepsilon - u) + B^T(p_\varepsilon - p) = 0 \ \text{ or } \quad A(u_\varepsilon - u) + \frac{1}{\varepsilon} B^T B(u_\varepsilon - u) = B^T p = f - Au \quad \text{and} \quad p_\varepsilon = \frac{1}{\varepsilon} B(u_\varepsilon - u).$$

Taking the duality brackets with $u_\varepsilon - u$ and using $B(u_\varepsilon - u) = \varepsilon\, p_\varepsilon = \varepsilon\,(p_\varepsilon - p) + \varepsilon\, p$, we have:

$$\langle A(u_\varepsilon - u), u_\varepsilon - u \rangle_{V',V} + \varepsilon\, \|p_\varepsilon - p\|_X^2 = -\varepsilon\,(p, p_\varepsilon - p)_X \leq \|p\|_X\, \|p_\varepsilon - p\|_X\, \varepsilon.$$

From the *inf-sup* condition (16(*ii*)), we deduce that $\beta\,\|p\|_X \le \|B^T p\|_{V'} \le \|A\|\,\|u\|_V + \|f\|_{V'}$, giving the bound on p from the bound on u, and similarly for the bound on $p_\varepsilon - p$. Thus, we have:

$$\|p\|_X \le c_0(\|A\|,\alpha,\beta)\,(\|f\|_{V'} + \|g\|_X) \quad \text{and} \quad \beta\,\|p_\varepsilon - p\|_X \le \|B^T(p_\varepsilon - p)\|_{V'} \le \|A\|\,\|u_\varepsilon - u\|_V.$$

Combining with the previous inequality and using the coercivity (16(*i*)), it yields the error estimates:

$$\|u_\varepsilon - u\|_V \le c_1(\|A\|,\alpha,\beta)\,(\|f\|_{V'} + \|g\|_X)\,\varepsilon \quad \text{and then} \quad \|p_\varepsilon - p\|_X \le c_2(\|A\|,\alpha,\beta)\,(\|f\|_{V'} + \|g\|_X)\,\varepsilon.$$

We conclude the proof of (19) with: $\|B(u_\varepsilon - u)\|_X \le \varepsilon\,(\|p_\varepsilon - p\|_X + \|p\|_X)$, and the previous bounds. A refined result can be also derived with an asymptotic expansion of $(u_\varepsilon, p_\varepsilon)$ in powers of ε, see [13, 23]. □

We now observe that Theorem 1.1 can be generalized to some continuous problems with assumptions allowing to write the asymptotic expansion (11). It is the case in the following result.

Corollary 2.2 (Generalization of Theorem 1.1 for the Stokes problem)**.** *Let the domain $\Omega \subset \mathbb{R}^d$ ($d = 2$ or 3 in practice) be an open bounded and connected set with a Lipschitz continuous boundary $\Gamma = \partial\Omega$. We consider the following Stokes problem where the viscosity $\mu > 0$ and $\mathbf{f} \in H^{-1}(\Omega)^d$ are given:*

$$-\mu\,\boldsymbol{\Delta}\mathbf{v} + \boldsymbol{\nabla} p = \mathbf{f}, \quad \textit{with} \quad \boldsymbol{\nabla}\cdot\mathbf{v} = 0 \quad \textit{in } \Omega, \quad \textit{and } \mathbf{v}_{|\Gamma} = 0. \tag{20}$$

Then, Theorem 1.1 holds and the solution $\hat{\mathbf{v}}_\varepsilon$ in (8) satisfies the asymptotic expansion given in (9,12).

Sketch of proof. In the case of the Stokes problem, the velocity correction $\hat{\mathbf{v}}_\varepsilon$ in the splitting augmented Lagrangian problem (6,7) satisfies an homogeneous Dirichlet boundary condition on Γ. Thus, the concerned Hilbert spaces are: $V = H^1_0(\Omega)^d$, $V' = H^{-1}(\Omega)^d$ and $X = X' = L^2_0(\Omega)$. The operators are now: $A = -\mu\,\boldsymbol{\Delta} = \mu\,BB^T$, a self-adjoint coercive operator of compact inverse, $B^T = \nabla$ and $B = -\mathrm{div}$, which is a surjective operator onto $L^2_0(\Omega)$, see [8].

Then, the "pressure" operator $S = BA^{-1}B^T$ is a coercive, self-adjoint isomorphism from $L^2_0(\Omega)$ onto $L^2_0(\Omega)$, see [14, Theorem 5.10], *i.e.* a zero-order operator. Hence, the asymptotic expansion (11) is valid and Theorem 1.1 also holds for the continuous Stokes problem where the operator $C_0 = -A^{-1}B^T S^{-1} B$ to calculate in (9) is only of zero-order. □

Remark 1 (Navier-Stokes problem with periodic boundary conditions). *In the case of periodic boundary conditions for the Navier-Stokes equations, the operators B and A^{-1} commute and we have $S = I$ and C_0 reduces to $C_0 = -A^{-1}B^T B$. That can be used for the numerical simulation of turbulence.*

2.2. Vector penalty-projection methods (VPP$_{r,\varepsilon}$) for unsteady incompressible Navier-Stokes problems

We use below the usual functionnal setting for the unsteady Navier-Stokes equations, see [23, 13, 8]. Let the domain $\Omega \subset \mathbb{R}^d$ ($d = 2$ or 3 in practice) be an open bounded and connected set with a Lipschitz continuous boundary $\Gamma = \partial\Omega$. For $T > 0$, we consider the following unsteady Navier-Stokes problem governing incompressible flows at a given Reynolds number Re where Dirichlet boundary conditions for the velocity $\mathbf{v}_{|\Gamma} = \mathbf{v}_D$ on Γ, $\mathbf{f}$ and an initial data $\mathbf{v}(t=0) = \mathbf{v}_0$ are given:

$$\partial_t\mathbf{v} + (\mathbf{v}\cdot\boldsymbol{\nabla})\mathbf{v} - \frac{1}{\mathrm{Re}}\boldsymbol{\Delta}\mathbf{v} + \boldsymbol{\nabla} p = \mathbf{f} \qquad \text{with} \quad \boldsymbol{\nabla}\cdot\mathbf{v} = 0 \quad \text{in } \Omega\times(0,T). \tag{21}$$

The following family of *vector penalty-projection* methods recently proposed in [1] is indeed designed on the basis of the previous two-step augmented Lagrangian method with: $r = r_0 + 1/\varepsilon$, yielding a correction step for the velocity vector at each time step. We also refer to [18, 5, 11] for the interest of the augmented term to drastically reduce the splitting error. We describe hereafter the two-step vector penalty-projection (VPP$_{r,\varepsilon}$) methods with an augmentation parameter $r_0 \ge 0$ and a penalty parameter $0 < \varepsilon \le 1$. For $\tilde{\mathbf{v}}^0, \mathbf{v}^0, \hat{\mathbf{v}}^0 = \mathbf{v}^0 - \tilde{\mathbf{v}}^0 \in L^2(\Omega)^d$ and $p^0 \in L^2_0(\Omega)$ given, they read as below with usual notations for the semi-discrete setting in time, $\delta t > 0$ being the time step. For all $n \in \mathbb{N}$ such that $(n+1)\delta t \le T$, find $\mathbf{v}^{n+1}$ and p^{n+1} satisfying $\mathbf{v}^{n+1}_{|\Gamma} = \mathbf{v}_D$, with $\tilde{\mathbf{v}}^{n+1}_{|\Gamma} = \mathbf{v}_D$ and $\hat{\mathbf{v}}^{n+1}_{|\Gamma} = 0$, such that:

$$\frac{\tilde{\mathbf{v}}^{n+1} - \tilde{\mathbf{v}}^n}{\delta t} + (\mathbf{v}^n\cdot\boldsymbol{\nabla})\tilde{\mathbf{v}}^{n+1} - \frac{1}{\mathrm{Re}}\boldsymbol{\Delta}\tilde{\mathbf{v}}^{n+1} - r_0\boldsymbol{\nabla}\left(\boldsymbol{\nabla}\cdot\tilde{\mathbf{v}}^{n+1}\right) + \boldsymbol{\nabla} p^n = \mathbf{f}^{n+1} \qquad \text{in } \Omega, \tag{22}$$

$$\frac{\hat{\mathbf{v}}^{n+1} - \hat{\mathbf{v}}^n}{\delta t} + (\mathbf{v}^n\cdot\boldsymbol{\nabla})\hat{\mathbf{v}}^{n+1} - \frac{1}{\mathrm{Re}}\boldsymbol{\Delta}\hat{\mathbf{v}}^{n+1} - \frac{1}{\varepsilon}\boldsymbol{\nabla}\left(\boldsymbol{\nabla}\cdot\hat{\mathbf{v}}^{n+1}\right) = \frac{1}{\varepsilon}\boldsymbol{\nabla}\left(\boldsymbol{\nabla}\cdot\tilde{\mathbf{v}}^{n+1}\right) \qquad \text{in } \Omega, \tag{23}$$

$$\mathbf{v}^{n+1} = \tilde{\mathbf{v}}^{n+1} + \hat{\mathbf{v}}^{n+1}, \quad \text{and} \quad p^{n+1} = p^n - r_0\boldsymbol{\nabla}\cdot\tilde{\mathbf{v}}^{n+1} - \frac{1}{\varepsilon}\boldsymbol{\nabla}\cdot\mathbf{v}^{n+1} \qquad \text{in } \Omega \tag{24}$$

Let us notice that in the (VPP$_{r,\varepsilon}$) method (22-24), the operator A in (17) also includes the discrete time derivative and the linearized convection term in addition to the diffusion term. A slightly modified version of the present (VPP$_{r,\varepsilon}$) scheme giving similar results was early presented in [1], as well as preliminary theoretical and numerical results. The complete analysis of the (VPP$_{r,\varepsilon}$) methods is carried out in [4], but the present study fully justifies the interest of such methods. We conclude by giving below the stability result of the (VPP$_{r,\varepsilon}$) method for the Navier-Stokes equations and some quasi-optimal error estimates for smooth solutions of the Stokes problem, see [4] for the details.

Theorem 2.3 (Stability of (VPP$_{r,\varepsilon}$) for Navier-Stokes problem with $\mathbf{v}_D = 0$)**.** *For $\mathbf{f} \in L^2(0,T;H^{-1}(\Omega)^d)$, $\mathbf{v}^0 \in L^2(\Omega)^d$ and $p^0 \in L^2_0(\Omega)$, there exists $C = C\left(\Omega, T, Re, \|\mathbf{f}\|_{L^2(0,T;H^{-1})}, \|\mathbf{v}^0\|_0, \|p^0\|_0\right) > 0$ such that, for all $r_0 \geq 0$, $0 < \varepsilon \leq 1$ and $0 < \delta t \leq T$, the solution $(\mathbf{v}^n, p^n)$ of the (VPP$_{r,\varepsilon}$) method (22-24) satisfies: for all $n \in \mathbb{N}$ with $(n+1)\delta t \leq T$,*

$$(i) \quad \|\mathbf{v}^{n+1}\|_0^2 + \varepsilon\delta t\,\|p^{n+1}\|_0^2 + \frac{1}{Re}\sum_{k=0}^{n}\delta t\,\|\boldsymbol{\nabla}\mathbf{v}^{k+1}\|_0^2 + \sum_{k=0}^{n}\left(\|\mathbf{v}^{k+1}-\mathbf{v}^k\|_0^2 + \frac{\varepsilon\delta t}{3}\|p^{k+1}-p^k\|_0^2\right) \leq C$$

$$(ii) \quad \sum_{k=0}^{n}\delta t\,\|\boldsymbol{\nabla}\cdot\mathbf{v}^{k+1}\|_0^2 \leq C\,\varepsilon \quad \text{and} \quad \sum_{k=0}^{n}\delta t\,\|\pi^{k+1}\|_0^{\frac{4}{3}} \leq C \quad \text{with } \pi^{k+1} = \sqrt{\delta t}\,p^{k+1} \quad \text{for } d = 2.$$

Theorem 2.4 (Error estimates of (VPP$_{r,\varepsilon}$) for the Stokes problem)**.** *Assume the solution $(\mathbf{v}, p)$ of the Stokes-Dirichlet problem smooth enough in time and space, well-prepared initial conditions and sufficiently small parameters such that: $4r_0(Re+\varepsilon) \leq 1$ and $4c(\Omega)\sqrt{Re}\,r_0\varepsilon \leq \sqrt{\delta t}$, $c(\Omega)$ being the Poincaré constant. Then, there exists $C = C(\Omega, T, Re, \mathbf{f}, \mathbf{v}_0, \mathbf{e}^0, q^0) > 0$ such that the velocity error $\mathbf{e}^n = \mathbf{v}^n - \mathbf{v}(t_n)$ and the pressure error $q^n = p^n - p(t_n)$ of the (VPP$_{r,\varepsilon}$) method (22-24) satisfy: for all $n \in \mathbb{N}$ with $(n+1)\delta t \leq T$,*

$$(i) \quad \|\mathbf{e}^{n+1}\|_0^2 + \varepsilon\delta t\,\|q^{n+1}\|_0^2 + \sum_{k=0}^{n}\frac{\delta t}{Re}\|\boldsymbol{\nabla}\mathbf{e}^{k+1}\|_0^2 \leq C\left(\delta t^2 + \varepsilon^2\delta t^{\frac{3}{2}}\right), \qquad \sum_{k=0}^{n}\delta t\,\|q^{k+1}\|_0^2 \leq C\left(\delta t^2 + \varepsilon^2\delta t\right)$$

$$(ii) \quad \sum_{k=0}^{n}\delta t\,\|\boldsymbol{\nabla}\cdot\mathbf{v}^{k+1}\|_0^2 = \sum_{k=0}^{n}\delta t\,\|\boldsymbol{\nabla}\cdot\mathbf{e}^{k+1}\|_0^2 \leq C\,(\delta t+\varepsilon)\,\varepsilon\delta t^2, \qquad \text{and} \quad \|\boldsymbol{\nabla}\mathbf{e}^{n+1}\|_0^2 \leq C\,Re^2\left(\delta t + \varepsilon^2\right).$$

With compactness arguments from Aubin-Lions-Simon, see e.g. [8], Theorem 2.3 allows us to prove the convergence of the (VPP$_{r,\varepsilon}$) solution of (22-24) to Navier-Stokes solutions of (21), when $\varepsilon = \delta t$ tends to zero, without additional regularity assumption; see [4] for the details.

3. Numerical experiments

The (VPP$_{r,\varepsilon}$) method is implemented with a Navier-Stokes finite volumes solver on the staggered uniform MAC mesh of size h issued from previous works; see [19]. The first test case is the unsteady Green-Taylor vortex such that the mean steady velocity field is of order 1 at Re = 100. The scheme is $O(\delta t)$ accurate in time for the velocity and pressure with $r_0 \geq 10^{-4}$, whereas it is $O(h^2)$ in space; see [1, 4] for additional results. We observe in Figure 1 (left) that the L^2-norm of the velocity divergence vanishes like $O(\varepsilon\,\delta t)$, as expected from Theorems 2.1 and 2.4 for $\varepsilon \leq \delta t$.

The second benchmark problem is the Rayleigh-Bénard thermal convection inside a square differentially heated vertical cavity at $Ra = 10^5$, the vertical walls being isothermal and the horizontal walls insulating. Here, we study the convergence properties of the velocity correction step (23) for this sharp test case. Again, we get the convergence of the velocity divergence as $O(\varepsilon\,\delta t)$, whatever the viscosity term included in the penalty-correction step and also for a viscosity coefficient $\mu = 0$, see [1]. We can reach the machine precision of 10^{-15} for double precision floating point computations. Besides, the solution of the *penalty-correction step* (23) proves to be all the cheaper as $\eta = \varepsilon/\delta t$ tends to zero, as expected from Theorem 1.1 and Corollary 1.2. Indeed, we dropped both the diffusion and convection terms in the correction step (23), *i.e.* we get: $A = 1/\delta t\, I$ as in Corollary 1.3 with the discrete operators in space: $B = -Div_h$, $B^T = Grad_h$ such that rank$(B) = m < n$. Then, we observe in Figure 1 (right) that, for $\eta = \varepsilon/\delta t \leq 10^{-4}$, only one iteration of the ILU(0)-BiCGStab2 preconditioned Krylov solver, is sufficient to get an accurate approximation of the operator $C_0 = -I_0$ in Corollary 1.3, and that independently on the mesh size h.

Hence, the key feature of the proposed splitting scheme is that the solution to the linear system associated with the vector projection step can be very fast and cheap because of the adapted form of the right-hand side. Indeed, we really take advantage of that feature to design the new fast vector penalty-projection method (VPP$_{\varepsilon}$) in [2, 3].

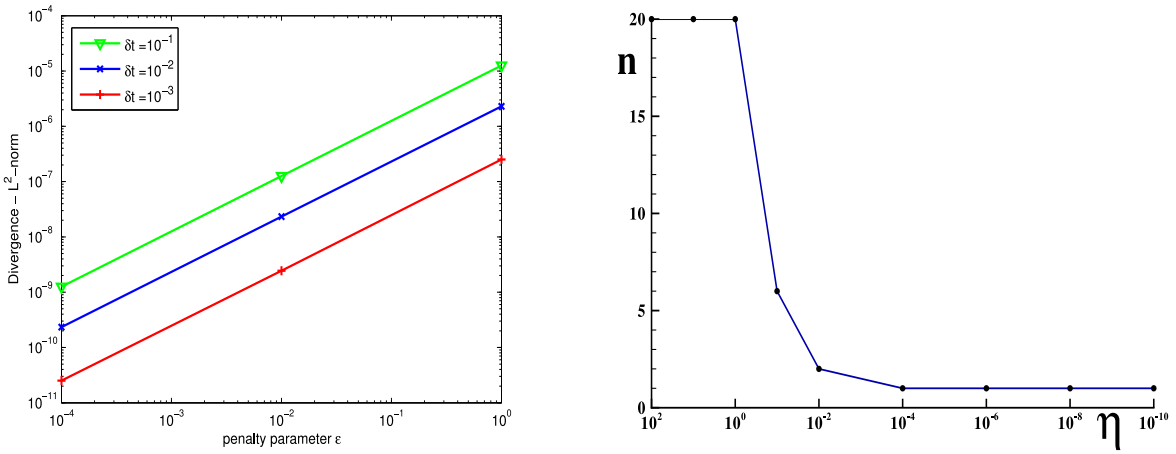

Figure 1: LEFT: ($VPP_{r,\varepsilon}$) velocity divergence versus penalty ε for the Green-Taylor vortex at Re = 100, $t = 10$ - $h = 1/512$, $r_0 = 1$, $\|res\|_2 \leq 10^{-10}$. RIGHT: number of ILU(0)-BiCGStab2 iterations versus $\eta = \varepsilon/\delta t$ for natural convection at $Ra = 10^5$ with $t = 2\delta t$, $\delta t = 1$, $h = 1/256$, $\|res\|_2 \leq 10^{-6}$.

References

[1] PH. ANGOT, J.-P. CALTAGIRONE AND P. FABRIE, Vector penalty-projection methods for the solution of unsteady incompressible flows, in *Finite Volumes for Complex Applications V*, R. Eymard and J.-M. Hérard (Eds), pp. 169-176, ISTE Ltd and J. Wiley & Sons, 2008.

[2] PH. ANGOT, J.-P. CALTAGIRONE AND P. FABRIE, A spectacular vector penalty-projection method for Darcy and Navier-Stokes problems, in *Finite Volumes for Complex Applications VI - Problems & Perspectives*, J. Fořt et al. (Eds), International Symposium FVCA6 in Prague, June 6-10, Springer Proceedings in Mathematics **4**, Vol. 1, pp. 39-47, Springer-Verlag (Berlin), 2011.

[3] PH. ANGOT, J.-P. CALTAGIRONE AND P. FABRIE, A fast vector penalty-projection method for incompressible non-homogeneous or multiphase Navier-Stokes problems, Applied Mathematics Letters, 2011 (submitted).

[4] PH. ANGOT, J.-P. CALTAGIRONE AND P. FABRIE, Convergence analysis and error estimates of vector penalty-projection methods for unsteady incompressible Darcy and Navier-Stokes problems, (preprint in preparation).

[5] PH. ANGOT, M. JOBELIN AND J.-C. LATCHÉ, Error analysis of the penalty-projection method for the time-dependent Stokes equations, Int. J. on Finite Volumes (IJFV) **6**(1), 1-26, 2009.

[6] M. BENZI, G.H. GOLUB AND J. LIESEN, Numerical solution of saddle point problems, Acta Numerica **14**, 1-137, Cambridge University Press, 2005.

[7] D.P. BERTSEKAS, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press (New York), 1982.

[8] F. BOYER, P. FABRIE, *Éléments d'analyse pour l'étude de quelques modèles d'écoulements de fluides visqueux incompressibles*, Mathématiques & Applications **52**, Springer-Verlag, 2006.

[9] J. CIHLÁŘ AND PH. ANGOT, Numerical solution of Navier-Stokes systems, Numer. Linear Algebra with Appl. **6**(1), 17-27, 1999.

[10] R. COURANT, Variational methods for the solution of problems of equilibrium and vibrations, Bull. Amer. Math. Soc. **49**, 1-23, 1943.

[11] C. FÉVRIÈRE, J. LAMINIE, P. POULLET AND PH. ANGOT, On the penalty-projection method for the Navier-Stokes equations with the MAC mesh, J. Comput. Appl. Math. (JCAM) **226**(2), 228-245, 2009.

[12] M. FORTIN AND R. GLOWINSKI, *Augmented Lagrangians: Application to the numerical solution of boundary value problems*, North-Holland (Amsterdam), 1983.

[13] V. GIRAULT AND P.A. RAVIART, *Finite Element Methods for the Navier-Stokes Equations*, Springer Series in Comput. Math., **5**, Springer-Verlag, 2nd ed. 1986.

[14] R. GLOWINSKI, *Numerical Methods for Non-linear Variational problems*, Springer Series in Comput. Phys., Springer-Verlag (New York), 1984.

[15] R. GLOWINSKI, Finite element methods for incompressible viscous flow, in *Handbook of Numerical Analysis*, vol. **IX**, P.G. Ciarlet and J.-L. Lions (Eds), pp. 3-1176, North-Holland (Amsterdam), 2003.

[16] R. GLOWINSKI AND P. LE TALLEC, *Augmented Lagrangians and Operator-Splitting Methods in Nonlinear Mechanics: Application to the numerical solution of boundary value problems*, North-Holland (Amsterdam), 1983.

[17] G.H. GOLUB AND C.F. VAN LOAN, *Matrix Computations*, Johns Hopkins Series in Math. Sci., **3**, The Johns Hopkins University Press (Baltimore and London), 2nd ed. 1989.

[18] M. JOBELIN, C. LAPUERTA, J.-C. LATCHÉ, PH. ANGOT AND B. PIAR, A finite element penalty-projection method for incompressible flows, J. Comput. Phys. **217**(2), 502-518, 2006.

[19] K. KHADRA, PH. ANGOT, S. PARNEIX, J.-P. CALTAGIRONE, Fictitious domain approach for numerical modelling of Navier-Stokes equations, Int. J. Numer. Meth. in Fluids, **34**(8), 651-684, 2000.

[20] W.W. HAGER, Dual techniques for constrained optimization, J. Optimization Theory and Appl., **55**(1), 37-71, 1987.

[21] W.W. HAGER, Updating the inverse of a matrix, SIAM Review **31**(2), 221-239, 1989.

[22] J. NOCEDAL AND S.J. WRIGHT, *Numerical Optimization*, Springer Series in Operations Research, Springer (Berlin), 1999.

[23] R. TEMAM, *Navier-Stokes Equations; Theory and Numerical Analysis*, North-Holland (Amsterdam), 4th ed. 1986.

[24] M.A. WOODBURY, The stability of out-input matrices, Chicago, III., 5 pp., 1949.