SIAM J. OPTIMIZATION
Vol. 7, No. 2, pp. xxx–xxx, 1997

CONVERGENCE RATES IN FORWARD-BACKWARD SPLITTING*

GEORGE H-G. CHEN† AND R. T. ROCKAFELLAR‡

Abstract. Forward-backward splitting methods provide a range of approaches to solving large-scale optimization problems and variational inequalities in which structure conducive to decomposition can be utilized. Apart from special cases where the forward step is absent and a version of the proximal point algorithm comes out, efforts at evaluating the convergence potential of such methods have so far relied on Lipschitz properties and strong monotonicity, or inverse strong monotonicity, of the mapping involved in the forward step, the perspective mainly being that of projection algorithms. Here convergence is analyzed by a technique that allows properties of the mapping in the backward step to be brought in as well. For the first time in such a general setting, global and local contraction rates are derived, moreover in a form making it possible to determine the optimal step size relative to certain constants associated with the given problem. Insights are thereby gained into the effects of shifting strong monotonicity between the forward and backward mappings when a splitting is selected.

Key words. Forward-backward splitting, numerical optimization, variational inequalities, projection algorithms, matrix splitting, operator splitting, convex programming

AMS subject classifications. 49R40, 49M27, 90C25, 90C06

1. INTRODUCTION.

This paper concerns a class of numerical methods for finding solutions to variational inequalities and other "generalized equations," especially in circumstances where a need for decomposition into simpler subproblems is apparent. Optimization problems fit the framework of these methods through the ways that variational inequalities can express first-order optimality conditions in primal, dual, or primal-dual form. Variational inequalities serve also in models of equilibrium and a diversity of other applications.

In general, the *variational inequality* problem for a closed, convex set $C \subset I\!R^n$ and a continuous mapping $F : C \to I\!R^n$ looks for a vector $\bar{x}$ such that

$$0 \in T(\bar{x}) \quad \text{for} \quad T(x) = F(x) + N_C(x), \tag{1.1}$$

where $N_C(x)$ is the set-valued normal cone mapping associated with C:

$$N_C(x) = \begin{cases} \{ w \in I\!R^n \mid \langle w, x' - x \rangle \le 0 \text{ for all } x' \in C \} & \text{when } x \in C, \\ \emptyset & \text{when } x \notin C, \end{cases} \tag{1.2}$$

with $\langle \cdot, \cdot \rangle$ denoting the canonical scalar product of vectors. The variational inequality problem is a *complementarity* problem when $C = I\!R^n_+$. Especially important is the case where F is *monotone* on C, in the sense that

$$\big\langle F(x') - F(x), x' - x \big\rangle \ge 0 \quad \text{for all} \quad x, x' \in C, \tag{1.3}$$

*Received by the editors July, 1995; accepted for publication (in revised form) October 2, 1995. This work was supported in part by NSF grants DMS–9200303 and DMS–9500957..

†Aero-Geo-Astro, Ltd. 15631 King Place, Lynnwood, WA 98037.

‡Dept. of Mathematics, Box 354350, University of Washington, Seattle, WA 98195-4350.

which in the optimization setting characterizes problems of convex type. Then the set-valued mapping T is itself monotone,

$$\big\langle w' - w,\, x' - x \big\rangle \geq 0 \quad \text{whenever} \quad w \in T(x),\ w' \in T(x'), \tag{1.4}$$

in fact *maximal* monotone: its graph set $\big\{(x, w) \,\big|\, w \in T(x)\big\}$ can't be enlarged without destroying monotonicity.

Forward-backward splitting methods are versatile in offering ways of exploiting the special structure of variational inequality problems. Following Lions and Mercier [1], such methods can be posed broadly in terms of solving

$$0 \in T(\bar{x}) \quad \text{when} \quad T(x) = T_1(x) + T_2(x) \tag{1.5}$$

for any mapping T that associates with each $x \in I\!R^n$ a (possibly empty) set $T(x) \subset I\!R^n$, a situation we symbolize by $T : I\!R^n \rightrightarrows I\!R^n$, and any representation of T as a sum of two other such mappings T_1 and T_2. The representation $T = T_1 + T_2$, which might be set up in a multitude of different ways, is called a *splitting* of T. From an initial point x_0, a point x_k is generated in each iteration k for $k = 1, 2, \ldots$ by solving the subproblem

$$0 \ \in \ (T_{1k} + T_2)(x_k) \quad \text{with} \quad T_{1k}(x) = T_1(x_{k-1}) + \frac{1}{\lambda_k} H_k[x - x_{k-1}] \tag{1.6}$$

for a *step size* value $\lambda_k > 0$ and an *implementation matrix* $H_k \in I\!R^{n \times n}$. Under the license of denoting the linear mapping $x \mapsto H_k x$ by the same symbol H_k, the iterations can be written in the form

$$x_k \in S_k(x_{k-1}) \quad \text{for} \quad S_k = \big(H_k + \lambda_k T_2\big)^{-1}\big(H_k - \lambda_k T_1\big). \tag{1.7}$$

The *forward-backward* name comes from the fact that (as long as H_k is nonsingular) the iteration mapping S_k has the equivalent expression

$$S_k = \big(I + \lambda_k H_k^{-1} T_2\big)^{-1}\big(I - \lambda_k H_k^{-1} T_1\big).$$

In the language of numerical analysis, $I - \lambda_k H_k^{-1} T_1$ gives a *forward* step with step size λ_k and direction vector $d_k = -H_k^{-1} u_k$, $u_k \in T_1(x_k)$ (or $u_k = T_1(x_k)$ when T_1 is single-valued), whereas $(I + \lambda_k H_k^{-1} T_2)^{-1}$ gives a *backward* step. Implementations where H_k is symmetric and positive definite are central, but weaker requirements are of interest in some situations.

For the purpose of solving a variational inequality (1.1), forward-backward splitting methods can be applied to

$$T = F + N_C, \qquad T_1 = F_1, \qquad T_2 = F_2 + N_C, \qquad \text{with } F = F_1 + F_2 \tag{1.8}$$

for a choice of continuous mappings $F_1 : C \to I\!R^n$ and $F_2 : C \to I\!R^n$. The iterations mean then that x_k is determined by solving

$$\begin{aligned} 0 \in T_k(x_k) \quad \text{for} \quad & T_k(x) = (F_{1k} + F_2)(x) + N_C(x) \quad \text{with} \\ & F_{1k}(x) = F_1(x_{k-1}) + \frac{1}{\lambda_k} H_k[x - x_{k-1}]. \end{aligned} \tag{1.9}$$

This covers many numerical procedures, the most familiar among them being ones that correspond to the splitting choices where either $F_1 = F$, $F_2 = 0$, or at the other extreme, $F_1 = 0$, $F_2 = F$.

For the splitting where $F_1 = F$ and $F_2 = 0$ in (1.8), so that $T_1 = F$ and $T_2 = N_C$, the forward-backward iterations with symmetric, positive definite H_k give a *projection algorithm* (of possibly "variable metric" type): x_k is the point of C nearest to

$$x'_k = x_{k-1} - \lambda_k H_k^{-1} F(x_{k-1}) \tag{1.10}$$

with respect to the norm induced by H_k. Indeed, (1.9) can be written in terms of (1.10) as the relation $-H_k[x_k - x'_k] \in N_C(x_k)$, which is necessary and sufficient for having

$$x_k = \operatorname*{argmin}_{x \in C} \Big\langle [x - x'_k], H_k[x - x'_k] \Big\rangle. \tag{1.11}$$

Of course, if $C = I\!R^n$ the projection trivializes and there's no backward step, just a forward step: one has $x_k = x_{k-1} - \lambda_k H_k^{-1} F(x_{k-1})$.

Among projection algorithms (1.9)–(1.10) the *gradient* case $F = \nabla f$ is the best known. If $H_k = I$ a variant of Cauchy's method is obtained, whereas if H_k is taken to be an approximation to $\nabla F(\bar{x}) = \nabla^2 f(\bar{x})$ a form of Newton's method comes out. Gradient projection algorithms were first studied in the Cauchy form by Goldstein [2] and in the Newton form by Levitin and Polyak [3], and they have since generated a large literature in optimization. For general variational inequalities, projection algorithms go back to Brézis and Sibony [4]; see also Sibony [5], Gajewski and Kluge [6], and for early developments attuned to mathematical programming, especially Dafermos [7].

For the other extreme splitting in (1.8), where $F_1 = F$ and $F_2 = 0$ so that $T_1 = 0$ and $T_2 = F + N_C$, the forward-backward procedure specializes to backward steps only and thus turns into (a "variable metric" form of) the *proximal point algorithm* for the mapping $T = F + N_C$. The proximal point algorithm was developed as a numerical method by Rockafellar [8], [9], in the case of $H_k \equiv I$, or equivalently $H_k \equiv H$ symmetric and positive definite, since that differs only in the designation of the norm (the context being one of a Hilbert space anyway). This algorithm is known to include, through various special choices, many other schemes such as generalized Douglas-Rachford splitting, cf. Eckstein and Bertsekas [10], and Spingarn splitting [11], which apply to maximal monotone mappings T not just of the variational inequality type in (1.1). An illuminating overview of splitting methods of all kinds has been provided by Eckstein [12].

Forward-backward splitting is closely related to an algorithmic approach introduced by Cohen as the "auxiliary problem principle" for problems of optimization in [13], [14], and variational inequalities in [15]. Cohen's formulation allows for the replacement of the linear implementation mapping $x \mapsto H_k x$ by a kind of nonlinear mapping, an idea treated also by Pang and Chan [16], among others. Patriksson [17] has explored this possibility broadly, showing how a vast array of known procedures can thereby be put into the framework of forward-backward methods.

Our focus in this paper is on the general iterations (1.7) for splittings $T = T_1 + T_2$ with T_1 single-valued in which T, T_1 and T_2 are monotone and both $T_1 \not\equiv 0$ and $T_2 \not\equiv 0$, so that nontrivial forward steps as well as nontrivial backward steps can be expected. In the variational inequality context this corresponds to splittings of type (1.8) in which F, F_1 and F_2 are monotone, and $F_1 \not\equiv 0$. We aim in particular at an understanding of convergence in cases where $F_2 \not\equiv 0$ too, so that more than a projection algorithm is involved. Such forms of forward-backward splitting methods are suggested by the decomposition needs of large-scale optimization problems with dynamic or stochastic structure [18], [19], [20], or PDE structure [21], but they haven't previously received much attention.

Except in connection with a weak ergodic type of convergence, cf. Passty [22], most of the research on general forward-backward splitting methods has relied on assumptions of strong monotonicity. Recall that a mapping $T : I\!R^n \rightrightarrows I\!R^n$ is *strongly monotone* if there is a constant $\mu > 0$ such that

$$\langle w' - w,\, x' - x \rangle \geq \mu \|x' - x\|^2 \quad \text{whenever} \quad w \in T(x),\ w' \in T(x'), \tag{1.12}$$

or equivalently, the mapping $T - \mu I$ is monotone. By the same token the inverse mapping T^{-1}, defined by taking $x \in T^{-1}(w)$ to mean that $w \in T(x)$, is strongly monotone if there is a constant $\nu > 0$ such that

$$\langle w' - w,\, x' - x \rangle \geq \nu \|w' - w\|^2 \quad \text{whenever} \quad w \in T(x),\ w' \in T(x'). \tag{1.13}$$

The strong monotonicity of T^{-1} is sometimes called the *Dunn* property or the *co-coercivity* of T. If T is single-valued and Lipschitz continuous with constant κ and strongly monotone with constant μ, then T^{-1} is strongly monotone with constant $\nu = \mu/\kappa^2$.

For implementations with $H_k \equiv I$ and $\lambda_k \equiv \lambda$, Gabay [23] showed that if T_1 is single-valued and maximal monotone with constant μ_1 as well as Lipschitz continuous with constant κ_1 the sequence of iterates x_k generated from any starting point x_0 converges to the unique solution $\bar{x}$ to (1.1), as long as $0 < \lambda < 2\mu_1/\kappa_1^2$. Alternatively he obtained convergence by assuming that a solution exists and T_1^{-1} is strongly monotone with constant ν_1 (which entails T_1 being Lipschitz continuous with constant $1/\nu_1$), and taking $0 < \lambda < 2\nu_1$. Tseng [24] extended the latter result to nonconstant step sizes λ_k and used it in that paper and in [25] to verify convergence for some schemes of problem decomposition. Further work in this vein, allowing for nonlinear implementation mappings and even for the approximation of T_1 and T_2 by mappings T_1^k and T_2^k in iteration k, was carried out to a certain degree by Mouallif, Nguyen and Strodiot [26] and Makler-Scheimberg, Nguyen and Strodiot [27].

In the special case of projection algorithms, Dafermos in [7] obtained Q-linear convergence as a consequence of deriving a global contraction rate for the iterations (1.10)–(1.11). She did this for a fixed matrix $H_k \equiv H$, possibly different from I, employing the H-norm

$$\|x\|_H = \sqrt{\langle x, Hx \rangle} \tag{1.14}$$

and its dual instead of the canonical norm $\|x\|$. She determined the fixed step size $\lambda_k \equiv \lambda$ for which the contraction rate would be optimal relative to constants of Lipschitz continuity and strong monotonicity for F when estimated in a certain way. These results were sharpened for affine variational inequalities by Dupuis and Darveau [28]. Bertsekas and Gafni [29] demonstrated R-linear convergence, i.e.,

$$\limsup_{k \to \infty} \|x_k - \bar{x}\|^{1/k} < 1, \tag{1.15}$$

for the case where C is polyhedral but F is not itself strongly monotone, rather just of the form $A^\top F_0 A$ for a strongly monotone mapping F_0 and a matrix A. Zanni [30] showed that the rate estimates of Dafermos and of Bertsekas and Gafni could not be expected to support rapid convergence; as an alternative he developed for the affine case in [31] a change of variables which offers a substantial improvement. Renaud in his thesis [32] got a contraction rate based on strong monotonicity constants for both F and F^{-1}. Marcotte and Wu [33], in proceeding from Tseng [25] and Luo and Tseng [34], proved linear convergence when C is polyhedral and F is affine with F^{-1} strongly monotone. Tseng in [35] developed broad conditions for Q-linear convergence of iterative methods which he applied to projection methods for affine variational inequalities, without however dealing explicitly with rate estimates or step sizes. For a survey of solution methods for finite-dimensional variational inequalities more generally, see Harker and Pang [36].

Little was known until recently about linear rates of convergence in the general setting of forward-backward methods. Renaud [32] succeeded in demonstrating R-linear convergence (1.15), although not actual contraction, in circumstances where T_1^{-1} is strongly monotone while T exhibits strong monotonicity relative to a unique solution $\bar{x}$. In Chen's thesis [37], contraction rates were developed under a variety of hypotheses entailing strong monotonicity of T, and step size optimization relative to those rate estimates was carried out.

Our efforts here take off from [37] in directions pioneered by Dafermos [7], going further than her and through territory encompassing much more than just projection algorithms. We reach conclusions significantly stronger than those of Chen [37] in some respects.

For simplicity at the start, we concentrate in Section 2 on a constant step size $\lambda_k \equiv \lambda$ and a constant matrix $H_k \equiv H$, which we allow to differ from I but assume to be symmetric and positive definite. We work at establishing linear convergence in the strong sense of global contractivity of the mapping

$$S_\lambda \;=\; \big(H + \lambda T_2\big)^{-1}\big(H - \lambda T_1\big) \;=\; \big(I + \lambda H^{-1} T_2\big)^{-1}\big(I - \lambda H^{-1} T_1\big) \tag{1.16}$$

with respect to the norm $\|\cdot\|_H$. Thus, we seek $\theta_\lambda \in [0,1)$ such that $\|S_\lambda(x') - S_\lambda(x)\|_H \leq \theta_\lambda \|x' - x\|_H$ for all x and x', hence in particular

$$\|S_\lambda(x) - \bar{x}\|_H \;\leq\; \theta_\lambda \|x - \bar{x}\|_H \quad \text{for all} \quad x. \tag{1.17}$$

We try to do this in such a manner that θ_λ can be expressed in terms of estimated properties of the given problem, thereby opening the way to optimizing θ_λ with respect to the choice of λ and obtaining some guidance on how λ might be selected in practice.

Obviously $\alpha_{\min}\|x\| \;\leq\; \|x\|_H \;\leq\; \alpha_{\max}\|x\|$ for the lowest and highest eigenvalues $\alpha_{\min}$ and $\alpha_{\max}$ of H, so that linear convergence with respect to $\|\cdot\|_H$ is equivalent to linear convergence with respect to $\|\cdot\|$. But the *rate* of linear convergence, as quantified by the size of the contraction factor, which is the crucial measure for numerical purposes, could be quite different in the two cases. By working with $\|\cdot\|_H$ we are able to capture a better rate through finer tuning. This corresponds essentially to a change of variables in which we look at behavior in $u = H^{-1/2}x$ instead of x, but our pattern is to proceed with the analysis directly in terms of x. More consistently than Dafermos and others in this subject, we avoid reference to the canonical norm $\|\cdot\|$ so as to keep our results close to the natural geometry of the method and away from extraneous dependence on the condition number of H through appeal to the eigenvalues $\alpha_{\min}$ and $\alpha_{\max}$. The philosophy is that if the condition number is to have any role at all, it should only be relative to a *one-time* change of variables, not a change to another norm and back again in every iteration, which is the unfortunate effect of bringing $\alpha_{\min}$ and $\alpha_{\max}$ into estimates of a contraction rate.

We utilize Lipschitz properties of T_1 but, in contrast to all previous research, we base the constant on a *residual* part of T_1, obtained by subtracting off the strong monotonicity that has been identified. We refer the Lipschitz constant to $\|\cdot\|_H$ and the corresponding dual norm $\|\cdot\|_{H^{-1}}$. Likewise, we adapt our estimates of strong monotonicity to $\|\cdot\|_H$ instead of $\|\cdot\|$.

Especially to be noted is that we don't insist on strong monotonicity of either T_1 or T_1^{-1}. This is motivated by prospective applications to the large-scale problems cited in [18]–[20]. Roughly, such problems follow the lines of minimizing $f(x) + g\big(D(x)\big)$ for proper, lsc, convex functions f and g and a mapping D like a discrete differential operator, integration operator or expectation operator. The subgradient condition for $\bar{x}$ to be optimal involves a dual element $\bar{y}$ such that $-D^\top\bar{y} \in \partial f(\bar{x})$ and $D\bar{x} \in \partial g^*(\bar{y})$, where g^* is the convex function conjugate to g. This condition can be written as

$$(0,0) \in (T_1 + T_2)(\bar{x}, \bar{y}) \quad \text{for} \quad \begin{cases} T_1(x,y) = (D^\top y, -Dx), \\ T_2(x,y) = \big(\partial f(x), \partial g^*(y)\big), \end{cases} \tag{1.18}$$

and it thus corresponds to a problem in $z = (x, y)$ that consists of solving $0 \in T(\bar{z})$ in the presence of a splitting $T = T_1 + T_2$ with T_1 and T_2 maximal monotone. Separability properties of f and g, reflected in a parallel choice of H, typically make it easy to iterate with $(x_k, y_k) = S_\lambda(x_{k-1}, y_{k-1})$, but T_1 is an *antisymmetric* linear mapping, so that neither T_1 nor T_1^{-1} can be strongly monotone. No results prior to ours could say anything substantial about convergence in this instance of a forward-backward splitting method. Note that (1.18) also gives incentive for not stopping at variational inequality models (1.8) in the treatment of such methods.

In Section 3 we study the implications of our basic results for the ways that a splitting $T = T_1 + T_2$ might be set up most advantageously. Applications are made to procedures for solving variational inequalities, in particular projection algorithms. We show a better contraction rate than that of Dafermos [7] or the one of Dupuis and Darveau [28] for affine variational inequalities; the result resembles a recent one of Zanni [31], but goes further. The step size associated with our contraction rate has the remarkable property of automatically optimizing performance with respect to the possible shifts of strong monotonicity between T_1 and T_2. The surprising result is thus achieved that, as long as our step size prescription is followed, any forward-backward method in the variational inequality case (1.8)–(1.9) can equally well be executed as a projection algorithm.

The global analysis of Section 3 is supplemented in Section 4 by a local analysis of convergence. Variable step sizes λ_k and implementation matrices H_k are taken up in Section 5, and methods with asymmetric implementation matrices in Section 6. For the literature on asymmetric implementations in solving variational inequalities; see Pang and Chan [16], Dafermos [38], Tseng [25], and Patriksson [17].

Because we are concerned with broad theoretical issues, we omit from the present study a number of refinements that could be pursued. The question of what happens when the subproblems in (1.6) or (1.9) are solved only approximately is not dealt with here, nor is the question of improvements based on augmenting the procedure with line search relative to some merit function. On the other hand, because we put our energy into the task of solving $0 \in T(\bar{x})$ for mappings T not necessarily of the variational inequality form (1.1), we get results that apply equally well to problems where, for example as in (1.18), the normal cone mapping N_C in (1.1) may be replaced by the subgradient mapping associated with a possibly nonsmooth convex function.

2. GLOBAL CONVERGENCE ANALYSIS.

A mapping T that assigns to each $x \in I\!R^n$ a set $T(x) \subset I\!R^n$ (perhaps a singleton) is indicated by $T : I\!R^n \rightrightarrows I\!R^n$. The effective domain of such a mapping is the set $\operatorname{dom} T = \{x \mid T(x) \neq \emptyset\}$. When T is maximal monotone, $\operatorname{dom} T$ is almost convex, in the sense that $\operatorname{cl}(\operatorname{dom} T)$ is a convex set whose relative interior lies within $\operatorname{dom} T$; cf. Minty [39]. The graph of T is considered to be the set of pairs (x, w) such that $w \in T(x)$, and the graph of T^{-1} consists therefore of the reversals (w, x) of all such pairs. The set of solutions $\bar{x}$ to $0 \in T(\bar{x})$ is $T^{-1}(0)$.

We investigate the feasibility of determining a solution $\bar{x}$ through iterations $x_k \in S_\lambda(x_{k-1})$ of the mapping in (1.16), as dictated by a choice of a splitting $T = T_1 + T_2$, a step size $\lambda > 0$ and an implementation matrix H. We don't suppose necessarily that T takes the variational inequality form in (1.1), but we do, for now, make the following assumptions.

Basic Assumptions (A). *The mapping $T_2 : I\!R^n \rightrightarrows I\!R^n$ is maximal monotone, and the set $\operatorname{dom} T_2$, denoted for simplicity by D, contains more than just one point (to avoid trivialities). The mapping $T_1 : D \to I\!R^n$ is single-valued, monotone and Lipschitz continuous, so in particular the mapping $T = T_1 + T_2$ has effective domain D, like T_2. The matrix $H \in I\!R^{n\times n}$ is symmetric and positive definite (hence nonsingular with H^{-1} symmetric and positive definite), while μ_1 and μ_2 denote constants such that*

$$\begin{cases} \text{the mappings } \widetilde{T}_1 = T_1 - \mu_1 H \text{ and } \widetilde{T}_2 = T_2 - \mu_2 H \text{ are} \\ \text{monotone on } D \text{ with } \mu_1 \geq 0,\ \mu_2 \geq 0,\ \mu_1 + \mu_2 > 0. \end{cases} \tag{2.1}$$

Furthermore, $\tilde{\kappa}_1$ is a Lipschitz constant for $\widetilde{T}_1$ on D from $\|\cdot\|_H$ to $\|\cdot\|_{H^{-1}}$:

$$\|\widetilde{T}_1(x') - \widetilde{T}_1(x)\|_{H^{-1}} \leq \tilde{\kappa}_1 \|x' - x\|_H \quad \text{for all} \quad x', x \in D. \tag{2.2}$$

Here in parallel to (1.14) we use the notation $\|w\|_{H^{-1}} = \sqrt{\langle w, H^{-1}w\rangle}$. The norm $\|w\|_{H^{-1}}$ is dual to the norm $\|\cdot\|_H$; one has

$$\langle x, w\rangle \leq \|x\|_H \|w\|_{H^{-1}} \quad \text{for all} \quad x,\, w \in I\!R^n. \tag{2.3}$$

Because monotone mappings must be interpreted technically as going from a vector space to its dual, it's natural in (2.2) in taking the H-metric on the domain of $\widetilde{T}_1$ to match it with the H^{-1}-metric on the range of $\widetilde{T}_1$.

The monotonicity assumptions in (2.1) correspond (in the face of T_1 being single-valued on $D = \operatorname{dom} T_2$) to requiring that

$$\langle T_1(x') - T_1(x),\, x' - x\rangle \geq \mu_1 \langle x' - x, H[x' - x]\rangle \quad \text{for all} \quad x,\, x' \in D,$$

$$\langle w' - w,\, x' - x\rangle \geq \mu_2 \langle x' - x, H[x' - x]\rangle \quad \text{whenever} \quad w \in T_2(x),\ w' \in T_2(x').$$

Because $\mu_1 + \mu_2 > 0$, these inequalities combine to imply that T is strongly monotone with constant $(\mu_1 + \mu_2)\alpha_{\min}$, where $\alpha_{\min}$ stands again for the lowest eigenvalue of H. But this constant of strong monotonicity won't itself come into play. We'll stay entirely with μ_1 and μ_2 as measures of monotonicity adapted to $\|\cdot\|_H$ rather than to $\|\cdot\|$.

Assumptions (A) in the variational inequality case (1.1) (for a closed, convex set C with more than one point, and continuous mappings F_1 and F_2 from C to $I\!R^n$) have $D = C$ and mean that $F_1 - \mu_1 H$ and $F_2 - \mu_2 H$ are monotone on C, or equivalently for $i = 1, 2$, that

$$\langle F_i(x') - F_i(x),\, x' - x\rangle \geq \mu_i \langle x' - x, H[x' - x]\rangle \quad \text{when} \quad x,\, x' \in C,$$

while $\widetilde{F}_1 = F_1 - \mu_1 H$ is Lipschitz continuous on C with constant $\tilde{\kappa}_1$ from $\|\cdot\|_H$ to $\|\cdot\|_{H^{-1}}$. For the maximal monotonicity of $T_2 = F_2 + N_C$, see Rockafellar [40, Thm. 3].

The introduction in (A) of a Lipschitz constant not for T_1 but the residual mapping $\widetilde{T}_1 = T_1 - \mu_1 H$ may seem odd, but it's crucial to our strategy of trying to separate the convergence analysis of forward-backward splitting methods from certain "unessential" features of the splitting. This will be clarified in Section 3. For practical purposes there's no disadvantage, at least, by virtue of the following fall-back estimate.

Proposition 2.1 (Lipschitz estimate). *Suppose κ_1 is a Lipschitz constant for T_1 itself on D from the norm $\|\cdot\|_H$ to the norm $\|\cdot\|_{H^{-1}}$:*

$$\|T_1(x') - T_1(x)\|_{H^{-1}} \leq \kappa_1 \|x' - x\|_H \quad \text{for all} \quad x', x \in D.$$

Then $\kappa_1 \geq \mu_1$, and the value $\sqrt{\kappa_1^2 - \mu_1^2}$ serves as a Lipschitz constant for $\widetilde{T}_1 = T_1 - \mu_1 H$ on D with respect to the same norms. Thus, one can always take $\tilde{\kappa}_1 = \sqrt{\kappa_1^2 - \mu_1^2}$ in the absence of anything better.

Proof. Squaring both sides of the Lipschitz inequality given by κ_1, we can write it as

$$\begin{aligned}
\kappa_1^2 \big\|x' - x\big\|_H^2 \;\geq\; & \big\|T_1(x') - T_1(x)\big\|_{H^{-1}}^2 \;=\; \big\|(\widetilde{T}_1 + \mu_1 H)(x') - (\widetilde{T}_1 + \mu_1 H)(x)\big\|_{H^{-1}}^2 \\
=\; & \big\|\,[\widetilde{T}_1(x') - \widetilde{T}_1(x)] + \mu_1 H[x' - x]\,\big\|_{H^{-1}}^2 \\
=\; & \big\|\widetilde{T}_1(x') - \widetilde{T}_1(x)\big\|_{H^{-1}}^2 + 2\mu_1 \big\langle H[x' - x],\, H^{-1}[\widetilde{T}_1(x') - \widetilde{T}_1(x)]\big\rangle \\
& \quad + \mu_1^2 \big\langle H[x' - x],\, H^{-1} H[x' - x]\big\rangle \\
=\; & \big\|\widetilde{T}_1(x') - \widetilde{T}_1(x)\big\|_{H^{-1}}^2 + 2\mu_1 \big\langle x' - x,\, \widetilde{T}_1(x') - \widetilde{T}_1(x)\big\rangle + \mu_1^2 \big\|x' - x\big\|_H^2.
\end{aligned}$$

Here $\big\langle x' - x,\, \widetilde{T}_1(x') - \widetilde{T}_1(x)\big\rangle \geq 0$ because $\widetilde{T}_1$ is monotone by assumption. Hence

$$\big\|\widetilde{T}_1(x') - \widetilde{T}_1(x)\big\|_{H^{-1}}^2 \;\leq\; (\kappa_1^2 - \mu_1^2)\big\|x' - x\big\|_H^2.$$

Because this holds for all x and x' in D, and D has more than one point, it's apparent that $\kappa_1 \geq \mu_1$, and that $\sqrt{\kappa_1^2 - \mu_1^2}$ serves as a Lipschitz constant $\tilde{\kappa}_1$ for $\widetilde{T}_1$ on D. □

We develop next a technical fact which will repeatedly be brought into play.

Proposition 2.2 (inverse Lipschitz continuity from strong monotonicity). *If a mapping $T_0 : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is maximal monotone, and $T_0 - \mu_0 H$ is monotone for some $\mu_0 > 0$, where H is symmetric and positive definite, then T_0^{-1} is single-valued and Lipschitz continuous, with μ_0^{-1} serving as a Lipschitz constant from the $\|\cdot\|_{H^{-1}}$ metric to the $\|\cdot\|_H$ metric.*

Proof. Whenever $w \in T_0(x)$ and $w' \in T_0(x')$ we have by assumption that

$$\begin{aligned}
0 \;\leq\; & \big\langle [w' - \mu_0 H x'] - [w - \mu_0 H x],\, x' - x\big\rangle \\
=\; & \big\langle w' - w,\, x' - x]\big\rangle - \mu_0 \big\langle H[x' - x], [x' - x]\big\rangle \\
\leq\; & \|x' - x\|_H \|w' - w\|_{H^{-1}} - \mu_0 \|x' - x\|_H^2.
\end{aligned}$$

Thus, $\|x' - x\|_H \leq \mu_0^{-1} \|w' - w\|_{H^{-1}}$ whenever $x' \in T_0(w')$ and $x \in T_0^{-1}(w)$, so that $T_0^{-1}(w)$ can't contain more than one point, and T_0^{-1} is Lipschitz continuous on its effective domain with constant μ_0^{-1} and in particular is locally bounded everywhere. But T_0^{-1} inherits maximal monotonicity from T_0, so the latter necessitates T_0^{-1} being nonempty-valued everywhere, cf. Rockafellar [41]. □

Theorem 2.3 (algorithmic background)**.** *Under (A) the mapping $T = T_1 + T_2$ is maximal monotone and also strongly monotone. There is a unique solution $\bar{x}$ to $0 \in T(\bar{x})$, and for any $\lambda > 0$ the iteration mapping S_λ is single-valued and Lipschitz continuous from the set $D = \operatorname{dom} T$ into itself, with unique fixed point $\bar{x}$.*

Proof. Although the single-valued mapping T_1 need not be defined outside of D, it at least has through Lipschitz continuity a unique continuous extension T_1' to the closed, convex set $C = \operatorname{cl} D$, this extension being monotone and having the same Lipschitz constant as T_1. We can enlarge T_1' to a maximal monotone mapping $T_1'' : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ by defining $T_1''(x) = T_1'(x) + N_C(x)$ when $x \in C$ but $T_1''(x) = \emptyset$ when $x \notin C$, cf. Rockafellar [40, Thm. 3]. Since $\operatorname{dom} T_1'' = C = \operatorname{cl}(\operatorname{dom} T_2)$, the relative interiors of $\operatorname{dom} T_1''$ and $\operatorname{dom} T_2$ have nonempty intersection (they actually coincide). Then, because T_2 like T_1'' is maximal monotone, it follows that $T_1'' + T_2$ is maximal monotone, cf. Rockafellar [40, Thm. 2]. To deduce that T is maximal monotone, it suffices therefore to demonstrate that $T_1''(x) + T_2(x) = T_1(x) + T_2(x)$ for all $x \in C$, or in other words that $T_2(x) + N_C(x) \subset T_2(x)$ for all $x \in C$. Unless actually $x \in D$, this holds trivially with both sides empty.

For any $x \in D$ and $w \in T_2(x)$ we have $\langle x' - x,\, w' - w\rangle \geq 0$ whenever $w' \in T_2(x')$; also, for any $u \in N_C(\hat{x})$ we have $\langle x' - x,\, u\rangle \leq 0$ for all $x' \in C$. Consequently, we have $\langle x' - x,\, w' - (w + u)\rangle \geq 0$ whenever $w' \in T_2(x')$. The maximal monotonicity of T_2 then implies $w + u \in T_2(x)$; for if not, the pair $(x, w + u)$ could be added to the graph of T_2 to get a properly larger mapping that is still monotone. Therefore, $T_2(x) + N_C(x) \subset T_2(x)$ for all $x \in D$, as required. Thus, T is maximal monotone.

From the representation $T = (T_1 - \mu_1 H) + (T_2 - \mu_2 H) + (\mu_1 + \mu_2)H$ with $\mu_1 + \mu_2 > 0$, where the first two terms are monotone by assumption, we have $T - (\mu_1 + \mu_2)H$ monotone. Because H is itself strongly monotone, as a consequence of being positive definite, T likewise is strongly monotone. Hence T^{-1} is single-valued and Lipschitz continuous by Proposition 2.2. In particular the set $T^{-1}(0)$, which consists of the solutions $\bar{x}$ to $0 \in T(\bar{x})$, has to be a singleton.

Turning now to the properties of S_λ, we observe first that the mapping λT_2, like T_2 itself, is maximal monotone and has the same effective domain as T_2, the relative interior of which meets that of the mapping $x \mapsto Hx$, namely $\mathbb{R}^n$. Furthermore the latter mapping, by virtue of linearity and positive definiteness, is maximal monotone, even strongly monotone. It follows through [40, Thm. 2] that $H + \lambda T_2$ is maximal monotone. Moreover, the mapping $[H + \lambda T_2] - (1 + \lambda\mu_2)H$ is monotone, so by Proposition 2.2 the mapping $(H + \lambda T_2)^{-1}$ must be single-valued everywhere and Lipschitz continuous, in fact with constant $(1 + \lambda\mu_2)^{-1}$ from $\|\cdot\|_{H^{-1}}$ to $\|\cdot\|_H$. At the same time the mapping $H - \lambda T_1$ is single-valued and Lipschitz continuous on D under (A), and therefore S_λ, the composite of these two mappings, is of such type as well.

The condition $x = S_\lambda(x)$ corresponds to having $[H - \lambda T_1](x) \in [H + \lambda T_2](x)$ and hence to having $-T_1(x) \in T_2(x)$, which is the same as $0 \in T(x)$. Therefore, the unique fixed point of S_λ on D is the unique $\bar{x}$ with $0 \in T(\bar{x})$. □

Theorem 2.4 (global contraction rate)**.** *Under (A) and for any $\lambda > 0$, the value*

$$\theta_\lambda = \begin{cases} \dfrac{\sqrt{(1-\lambda\mu_1)^2 + \lambda^2\tilde{\kappa}_1^2}}{1+\lambda\mu_2} & \text{when } \lambda^{-1} \geq \mu_1, \\[2ex] \dfrac{\lambda(\tilde{\kappa}_1 + \mu_1) - 1}{1+\lambda\mu_2} & \text{when } \lambda^{-1} \leq \mu_1, \end{cases} \tag{2.4}$$

which depends continuously on λ, is a Lipschitz constant for $S_\lambda : D \to D$ as a mapping *from the* $\|\cdot\|_H$ *metric to the* $\|\cdot\|_H$ *metric. In particular*

$$\|S_\lambda(x) - \bar{x}\|_H \;\le\; \theta_\lambda \|x - \bar{x}\|_H \quad \text{for all} \quad x \in D,$$

so that S_λ *is globally contractive to* $\bar{x}$ *on* D *when* $\theta_\lambda < 1$*, which is true for all* $\lambda > 0$ *sufficiently small, specifically if and only if* λ *is chosen small enough that*

$$\lambda^{-1} \;>\; \frac{\mu_1 - \mu_2}{2} + \frac{\tilde{\kappa}_1}{2} \max\left\{1, \frac{\tilde{\kappa}_1}{\mu_1 + \mu_2}\right\}. \tag{2.5}$$

The best such estimated contraction rate θ_λ*, as* λ *ranges over these choices, is*

$$\overline{\theta} = \theta_{\overline{\lambda}} = \frac{1}{\sqrt{1 + \left(\dfrac{\mu_1 + \mu_2}{\tilde{\kappa}_1}\right)^2}}, \qquad \text{for } \overline{\lambda} = \frac{1}{\left(\dfrac{\tilde{\kappa}_1^2}{\mu_1 + \mu_2}\right) + \mu_1}. \tag{2.6}$$

Proof. As already argued in the proof of Proposition 2.2, our assumptions on T_2 in (A) ensure that $(H+\lambda T_2)^{-1}$ is single-valued and Lipschitz continuous with constant $(1+\lambda\mu_2)^{-1}$ from $\|\cdot\|_{H^{-1}}$ to $\|\cdot\|_H$. Since $S_\lambda = (H+\lambda T_2)^{-1}(H - \lambda T_1)$, our task in establishing the Lipschitz constant θ_λ for S_λ comes down to showing that the second factor in the formula for θ_λ serves as a Lipschitz constant for $H - \lambda T_1$ on D from $\|\cdot\|_H$ to $\|\cdot\|_{H^{-1}}$. Fix any points x and x' in D. In terms of having $T_1 = \widetilde{T}_1 + \mu_1 H$, we expand

$$\begin{aligned}
&\big\|(H - \lambda T_1)(x') - (H - \lambda T_1)(x)\big\|^2_{H^{-1}} \\
&\qquad = \big\|\big[(1-\lambda\mu_1)H - \lambda\widetilde{T}_1\big](x') - \big[(1-\lambda\mu_1)H - \lambda T_1\big](x)\big\|^2_{H^{-1}} \\
&\qquad = \big\|(1-\lambda\mu_1)H[x'-x] - \lambda\big[\widetilde{T}_1(x') - \widetilde{T}_1(x)\big]\,\big\|^2_{H^{-1}} \\
&\qquad = (1-\lambda\mu_1)^2\big\langle H[x'-x], H^{-1}H[x'-x]\big\rangle \\
&\qquad\qquad - 2\lambda(1-\lambda\mu_1)\big\langle H[x'-x], H^{-1}\big[\widetilde{T}_1(x') - \widetilde{T}_1(x)\big]\big\rangle \\
&\qquad\qquad\qquad + \lambda^2\big\langle\big[\widetilde{T}_1(x') - \widetilde{T}_1(x)\big], H^{-1}\big[\widetilde{T}_1(x') - \widetilde{T}_1(x)\big]\big\rangle \\
&\qquad = (1-\lambda\mu_1)^2\big\|x'-x\big\|^2_H + \lambda^2\big\|\widetilde{T}_1(x') - \widetilde{T}_1(x)\big\|^2_{H^{-1}} \\
&\qquad\qquad - 2\lambda(1-\lambda\mu_1)\big\langle x'-x,\, \widetilde{T}_1(x') - \widetilde{T}_1(x)\big\rangle.
\end{aligned} \tag{2.7}$$

At this stage our analysis divides into the cases where $1 - \lambda\mu_1 \ge 0$ or $1 - \lambda\mu_1 \le 0$, which correspond to $\lambda^{-1} \ge \mu_1$ or $\lambda^{-1} \le \mu_1$. (When equality holds in these relations the two paths of argument will lead to the same thing.)

In the case where $1-\lambda\mu_1 \ge 0$, we can invoke the fact that $\big\langle x'-x,\, \widetilde{T}_1(x')-\widetilde{T}_1(x)\big\rangle \ge 0$ because $\widetilde{T}_1$ is monotone on D. We get then from (2.7) and the specification of $\tilde{\kappa}_1$ that

$$\big\|(H - \lambda T_1)(x') - (H - \lambda T_1)(x)\big\|^2_{H^{-1}} \;\le\; (1-\lambda\mu_1)^2\big\|x'-x\big\|^2_H + \lambda^2\tilde{\kappa}_1^2\big\|x'-x\big\|^2_H,$$

hence $\|(H - \lambda T_1)(x') - (H - \lambda T_1)(x)\|_{H^{-1}} \;\le\; [(1-\lambda\mu_1)^2 + \lambda^2\tilde{\kappa}_1^2]^{1/2}\|x'-x\|_H$ in accordance with the first version of θ_λ. In the case where $1 - \lambda\mu_1 \le 0$ instead, we use the inequality

$$\big\langle x'-x,\, \widetilde{T}_1(x') - \widetilde{T}_1(x)\big\rangle \;\le\; \|x'-x\|_H\|\widetilde{T}_1(x') - \widetilde{T}_1(x)\|_{H^{-1}} \;\le\; \tilde{\kappa}_1\|x'-x\|^2_H$$

from (2.3) to argue through (2.7) that

$$\begin{aligned}
&\big\|(H-\lambda T_1)(x')-(H-\lambda T_1)(x)\big\|_{H^{-1}}^2 \\
&\quad\leq\ (1-\lambda\mu_1)^2\big\|x'-x\big\|_H^2+\lambda^2\tilde{\kappa}_1^2\big\|x'-x\big\|_H^2+2\lambda(\lambda\mu_1-1)\tilde{\kappa}_1\big\|x'-x\big\|_H^2 \\
&\quad\leq\ \big[(1-\lambda\mu_1)^2+\lambda^2\tilde{\kappa}_1^2+2\lambda(\lambda\mu_1-1)\tilde{\kappa}_1\big]\big\|x'-x\big\|_H^2 \\
&\quad=\ \big[\lambda(\tilde{\kappa}_1+\mu_1)-1\big]^2\big\|x'-x\big\|_H^2.
\end{aligned}$$

We obtain $\|(H-\lambda T_1)(x')-(H-\lambda T_1)(x)\|_{H^{-1}}\leq[\lambda(\tilde{\kappa}_1+\mu_1)-1]\|x'-x\|_H$ in accordance with the second version of θ_λ.

In order to understand the nature of the factor θ_λ better, we begin with observation that for λ large enough that $\lambda^{-1}\leq\mu_1$ the function $\phi(\lambda)=\theta_\lambda=[\lambda(\tilde{\kappa}_1+\mu+1)-1]/(1+\lambda\mu_2)$ has $\phi'(\lambda)=[\tilde{\kappa}_1+\mu_1+\mu_2]/(1+\lambda\mu_2)^2>0$ and hence is an increasing function. In seeking low values of θ_λ we therefore aren't interested in λ with $\lambda^{-1}<\mu_1$ and can concentrate on the case of $\lambda^{-1}\geq\mu_1$, where the other formula holds for θ_λ. Note, though, that

$$\theta_\lambda<1 \iff \lambda^{-1}>(\mu_1-\mu_2+\tilde{\kappa}_1)/2 \quad \text{when } \lambda^{-1}<\mu_1. \tag{2.8}$$

The analysis of θ_λ when $\lambda^{-1}\geq\mu_1$ is simplified by passing temporarily from λ to the parameter

$$\tau=(\lambda^{-1}+\mu_2)^{-1}, \quad \text{which gives} \quad \tau^{-1}=\lambda^{-1}+\mu_2, \quad \lambda^{-1}=\tau^{-1}-\mu_2.$$

The condition $\lambda^{-1}\geq\mu_1$ means $\tau^{-1}\geq\mu$, where we introduce the notation $\mu=\mu_1+\mu_2$ for simplicity. We get

$$\theta_\lambda^2\ =\ \frac{(\lambda^{-1}-\mu_1)^2+\tilde{\kappa}_1^2}{(\lambda^{-1}+\mu_2)^2}\ =\ \tau^2\big[(\tau^{-1}-\mu)^2+\tilde{\kappa}_1^2\big]\ =\ 1-2\mu\tau+(\tilde{\kappa}_1^2+\mu^2)\tau^2. \tag{2.9}$$

From this expression it's obvious that $\theta_\lambda<1$ if and only if $(\tilde{\kappa}_1^2+\mu^2)\tau^2<2\mu\tau$, or in other words $\tau^{-1}>(\tilde{\kappa}_1^2+\mu^2)/2\mu$. This condition translates back to $\lambda^{-1}>\big[(\tilde{\kappa}_1^2+\mu^2)/2\mu\big]-\mu_2=\big[\tilde{\kappa}_1^2/2\mu\big]+(\mu_1-\mu_2)/2$, because $\mu^2-2\mu\mu_2=\mu(\mu_1+\mu_2-2\mu_2)=\mu(\mu_1-\mu_2)$. Thus,

$$\theta_\lambda<1 \iff \lambda^{-1}>\frac{\mu_1-\mu_2}{2}+\frac{\tilde{\kappa}_1^2}{2(\mu_1+\mu_2)} \quad \text{when } \lambda^{-1}\geq\mu_1. \tag{2.10}$$

The union of (2.8) with (2.10) furnishes the condition claimed in (2.5) for having $\theta_\lambda<1$.

The expression in (2.9) is a strictly convex function of τ which achieves its minimum uniquely when $-2\mu+2(\tilde{\kappa}_1^2+\mu^2)\tau=0$, or in other words for the value $\bar{\tau}=\mu/(\tilde{\kappa}_1^2+\mu^2)$. This does have the property that $\bar{\tau}^{-1}\geq\mu$, so the associated step size $\overline{\lambda}$ satisfies $\overline{\lambda}^{-1}\geq\mu_1$. The corresponding minimum value for the expression in (2.9) is $\tilde{\kappa}_1^2/(\tilde{\kappa}_1^2+\mu^2)$. Therefore, the lowest achievable value for θ_λ is $\theta_{\overline{\lambda}}=\tilde{\kappa}_1/\sqrt{\tilde{\kappa}_1^2+\mu^2}$ for

$$\overline{\lambda}=1/(\bar{\tau}^{-1}-\mu_2)=\mu/[(\tilde{\kappa}_1^2+\mu^2)-\mu\mu_2],$$

which works out to the value claimed for $\overline{\lambda}$ in the theorem. □

Corollary 2.5 (special rate estimates). *When the estimate $\tilde{\kappa}_1 = \sqrt{\kappa_1^2 - \mu_1^2}$ is used in accordance with Proposition 2.1, the corresponding best contraction rate that can be guaranteed is*

$$\theta_{\overline{\lambda}} = \frac{1}{\sqrt{1 + \dfrac{(\mu_1+\mu_2)^2}{\kappa_1^2 - \mu_1^2}}}, \qquad \textit{for } \overline{\lambda} = \frac{\mu_1+\mu_2}{\kappa_1^2 + \mu_1\mu_2}. \tag{2.11}$$

In the case of $\mu_2 = 0$ this reduces to

$$\theta_{\overline{\lambda}} = \sqrt{1 - \left(\frac{\mu_1}{\kappa_1}\right)^2}, \qquad \textit{for } \overline{\lambda} = \frac{\mu_1}{\kappa_1^2}. \tag{2.12}$$

Proof. The case in (2.11) is obvious from Theorem 2.4, and the one in (2.12) then follows by elementary algebra in replacing μ_2 by 0. □

The convergence result in Corollary 2.5 was developed in Chen's thesis [37], but Theorem 2.4 itself, with its emphasis on $\tilde{\kappa}_1$ instead of κ_1, appears here for the first time.

An alternative result of Renaud [32, Prop. VI.25] under the assumption that T and T_1^{-1} are strongly monotone gives R-linear convergence, the convergence factor (not necessarily a contraction factor as above) being

$$\frac{1}{\sqrt{1 + \dfrac{\mu\nu_1}{\alpha_{\min}/\alpha_{\max}}}} \qquad \text{for } \lambda = \nu_1\alpha_{\min}, \tag{2.13}$$

where μ and ν_1 are strong monotonicity constants for T and T_1^{-1} in the sense of (1.12) and (1.13) (i.e., calibrated by I instead of H), and $\alpha_{\min}$ and $\alpha_{\max}$ are the smallest and biggest eigenvalues of H. (Here we specialize to $\mathbb{R}^n$; Renaud operated in the context of a possibly infinite-dimensional Hilbert space.) Renaud didn't actually require μ to be a strong monotonicity constant in the full sense of (1.12), but just a value satisfying

$$\langle w - \bar{w}, x - \bar{x}\rangle \geq \mu\|x - \bar{x}\|^2 \quad \text{if} \quad w \in T(x), \quad \text{where} \quad \bar{w} = 0 \in T(\bar{x}). \tag{2.14}$$

Likewise this would suffice in Theorem 2.4 if we aimed at Q-linear convergence to $\bar{x}$ instead of insisting that S_λ be a contraction mapping; see Section 4.

The dependence of Renaud's factor in (2.13) on $\alpha_{\min}/\alpha_{\max}$, which is the condition number of H, should be noted. This is disadvantageous unless $H = I$ so the condition number is 1; see Section 3. When $H = I$ and T_1 is strongly monotone, it's possible under our assumptions to take $\nu_1 = \mu_1/\kappa_1^2$. Then Renaud's factor in (2.13) becomes

$$\frac{1}{\sqrt{1 + \dfrac{\mu_1(\mu_1+\mu_2)}{\kappa_1^2}}} \qquad \text{for } \lambda = \frac{\mu_1}{\kappa_1^2},$$

which isn't as sharp as our factor in Corollary 2.5. On the other hand, if $\nu_1 > 0$ is known directly one can take $\kappa_1 = 1/\nu_1$ and get $1/\sqrt{1 + (\mu_1+\mu_2)\nu_1}$ in (2.13) in comparison to $1/\sqrt{1 + [(\mu_1+\mu_2)\nu_1]^2}$ in Corollary 2.5, where $\mu_1\nu_1 \leq 1$ but perhaps $(\mu_1 + \mu_2)\nu_1 > 1$.

3. UTILIZATION OF STRONG MONOTONICITY.

A major purpose of our analysis has been to gain insight into how a splitting can be set up advantageously. In expressing T as a sum $T_1 + T_2$, there may be terms that could be assigned either to T_1 or to T_2 without creating an obstacle to the implementation of the forward-backward method. What approach is best in enhancing convergence?

Let's focus on shifts of positive monotonicity. On the basis of (A) we can write $T = \widetilde{T}_1 + \widetilde{T}_2 + \mu H$ for $\widetilde{T}_1 = T_1 - \mu_1 H$, $\widetilde{T}_2 = T_2 - \mu_2 H$, and $\mu = \mu_1 + \mu_2$. Here $\widetilde{T}_1$ and $\widetilde{T}_2$ are maximal monotone (for if not, that would mean the graph of one of them, say $\widetilde{T}_1$, could be enlarged without destroying monotonicity, in which case the same would be true for $\widetilde{T}_1 + \mu_1 H = T_1$, contrary to the maximality of T_1).

Suppose we were to divide up μ in a different way, $\mu = \mu_1' + \mu_2'$ with $\mu_1' \geq 0$ and $\mu_2' \geq 0$, and set $T_1' = \widetilde{T}_1 + \mu_1' H$ and $T_2' = \widetilde{T}_2 + \mu_2' H$. This would give a different splitting, $T = T_1' + T_2'$, in which T_1' and T_2' are again maximal monotone. Could there be any advantage in this for the algorithm's performance when implemented with the matrix H?

The answer is *no*—as long as the optimal step size prescription of Theorem 2.4 is employed. This is clear from the fact that the optimal contraction rate $\overline{\theta}$ in (2.6) depends only on $\tilde{\kappa}_1$ and the sum $\mu_1 + \mu_2$ and therefore would be the same under the different splitting, since $\mu_1' + \mu_2' = \mu_1 + \mu_2$ and even $T_1' - \mu_1' H = \widetilde{T}_1 = T_1 - \mu_1 H$ (so κ is unaffected). Indeed, the contraction rate has been optimized in Theorem 2.4 with respect to the whole range of splittings that we are looking at. In using the step size $\overline{\lambda}$ prescribed for the splitting $T = T_1 + T_2$, one is able *automatically* to capture whatever algorithmic advantages may lie in this direction. Although the step sizes for the splittings $T = T_1 + T_2$ and $T = T_1' + T_2'$ are given differently as

$$\overline{\lambda} = \frac{1}{\left(\dfrac{\tilde{\kappa}_1^2}{\mu_1 + \mu_2}\right) + \mu_1}, \qquad \overline{\lambda}' = \frac{1}{\left(\dfrac{\tilde{\kappa}_1^2}{\mu_1' + \mu_2'}\right) + \mu_1'},$$

and may not themselves be the same, they necessarily result in the same optimal rate $\overline{\theta}$.

But a subtle distinction must be noted between Theorem 2.4 and Corollary 2.5. If the tactic in developing a Lipschitz constant for $\widetilde{T}_1$ were to use an estimate based on Proposition 2.1, the answer to the question posed would instead be *yes*!

The reason is that in passing from $T = T_1 + T_2$ to $T = T_1' + T_2'$ such an estimate $\tilde{\kappa}_1 = \sqrt{\kappa_1^2 - \mu_1^2}$, where κ_1 is a Lipschitz constant for T_1, would be replaced by a *different* value $\tilde{\kappa}_1' = \sqrt{{\kappa_1'}^2 - {\mu_1'}^2}$, where κ_1' is a Lipschitz constant for T_1' (relative to the specified norms). Then not only would the corresponding step sizes $\overline{\lambda}$ and $\overline{\lambda}'$, as dictated by (2.11), be different, but they would result in different contraction rates: $\theta_{\overline{\lambda}} \neq \theta_{\overline{\lambda}'}$ in (2.11). The issue would arise of determining which splitting $T = T_1' + T_2'$ minimizes $\sqrt{{\kappa_1'}^2 - {\mu_1'}^2}$ and thus furnishes the best contraction rate. Actually, we know from Proposition 2.1 that the minimum is achieved when $T_1' = \widetilde{T}_1 = T_1 - \mu_1 H$, $T_2' = \widetilde{T}_2 + \mu H = T_2 + \mu_1 H$. Thus, if we were to rely on the result in Corollary 2.5 rather than the one in Theorem 2.4, as for instance in [37], the optimal splitting would be obtained by extracting all possible strong monotonicity from T_1 and reassigning it to T_2, a qualitatively very different conclusion.

This highlights the contrast between the technique adopted here and previous research, which has utilized a Lipschitz constant for T_1 itself (moreover one in terms of the canonical norm only), not to speak of concentrating on strong monotonicity of T_1. Through Theorem 2.4 we can optimally exploit strong monotonicity of T_1 or T_2 or both, without in the end having to switch any terms in the splitting.

The idea is illustrated by its application to solving variational inequalities.

Theorem 3.1 (application to projection algorithms)**.** *Consider the variational inequality problem (1.1) in the case of a nonempty, closed, convex set $C \subset I\!R^n$ and a continuous, single-valued mapping $F : C \to I\!R^n$. Let H be a symmetric, positive definite matrix, and let $\mu > 0$ be a constant such that F satisfies the strong monotonicity condition*

$$\big\langle F(x') - F(x),\, x' - x\big\rangle \;\geq\; \mu\,\|x' - x\|_H^2 \quad \text{for all} \quad x, x' \in C. \tag{3.1}$$

Let $\tilde{\kappa} \geq 0$ be a Lipschitz constant for $\widetilde{F} = F - \mu H$ on C from the norm $\|\cdot\|_H$ to the norm $\|\cdot\|_{H^{-1}}$. Then in applying Theorem 2.4 to the splitting $T = T_1 + T_2$ for $T_1 = F$, $T_2 = N_C$, and with $\mu_1 = \mu$, $\mu_2 = 0$, the optimal contraction rate is

$$\bar{\theta} = \theta_{\overline{\lambda}} = \frac{1}{\sqrt{1 + (\mu/\tilde{\kappa})^2}}, \qquad \text{for } \overline{\lambda} = \frac{\mu}{\tilde{\kappa}^2 + \mu^2}. \tag{3.2}$$

No alternative splitting $T = T_1' + T_2'$ in the mode of $T_1' = F - \tau H$ and $T_2' = \tau H + N_C$ for some $\tau \in (0, \mu]$ can provide a better contraction rate through Theorem 2.4.

Proof. This is evident from the preceding remarks. The assumptions furnish a specialization of the conditions in (A) to the special case in question. □

The fact that, under the circumstances described, execution of the forward-backward splitting method as a projection method is just as good as any alternative execution obtainable by shifting the strong monotonicity from the "forward" part to the "backward" part of the iteration mapping, is perhaps surprising. But again, it must be remembered that this result depends on utilizing a Lipschitz constant $\tilde{\kappa}$ for $\widetilde{F} = F - \mu H$ rather than a constant κ attached directly to F itself.

Corollary 3.2. *When the estimate $\tilde{\kappa} = \sqrt{\kappa^2 - \mu^2}$ is used in Corollary 3.2 in accordance with Proposition 2.1, κ being a Lipschitz constant for F on C from $\|\cdot\|_H$ to $\|\cdot\|_{H^{-1}}$, the corresponding best contraction rate that can be guaranteed for the projection algorithm is*

$$\theta_{\overline{\lambda}} = \sqrt{1 - \left(\frac{\mu}{\kappa}\right)^2}, \qquad \text{for } \overline{\lambda} = \frac{\mu}{\kappa^2}. \tag{3.3}$$

Proof. This applies the second part of Corollary 2.5. □

For the case of $H = I$, results related to Corollary 3.2 were obtained recently by Renaud. He noted in [32, p. 143] the contraction rate in (3.3) and went on to demonstrate Q-linear convergence, although not the full contraction property, under the assumption that F^{-1} is strongly monotone with constant $\nu > 0$, the factor then being

$$\frac{1}{\sqrt{1 + \mu\nu}}, \qquad \text{for } \lambda = \nu,$$

cf. [32, Prop. VI.2]. This alternative assumption is satisfied when F is Lipschitz continuous with constant κ (from $\|\cdot\|$ to $\|\cdot\|$), namely with $\nu = \mu/\kappa^2$, and Renaud's factor reduces then to ours.

For the general case where $H \neq I$, the contraction rate in Corollary 3.2 may be compared for the one derived for projection algorithms by Dafermos [7]. In effect she got

$$\sqrt{1-\frac{\mu^2}{\beta_1^2\beta_2^2\operatorname{cond}(H)}}, \tag{3.4}$$

where $\operatorname{cond}(H)$ is the condition number of H (its highest eigenvalue divided by its lowest eigenvalue), β_1 is a conversion factor from $\|\cdot\|_H$ to $\|\cdot\|$, and β_2 is a Lipschitz constant for F from $\|\cdot\|$ to $\|\cdot\|_{H^{-1}}$, so that $\beta_1\beta_2$ is an (upper) estimate for the Lipschitz constant κ in Corollary 3.2. Unless $H = I$, Dafermos' denominator in (3.4) has to be greater than ours in (3.3), and her contraction factor accordingly has to be nearer to 1, thus not as good. The dependence of (3.4) on the condition number for H illustrates very well the unwarranted consequences of bringing in the canonical norm $\|\cdot\|$ instead of sticking consistently with the method's intrinsic geometry. The canonical norm is irrelevant in this appraisal of algorithmic performance.

Theorem 3.3 (application to affine variational inequalities)**.** *Consider the variational inequality problem (1.1) in the case of a nonempty, closed, convex set $C \subset \mathbb{R}^n$ and an affine mapping $F(x) = Mx + q$. Let $M_s = \frac{1}{2}(M + M^\top)$ and $M_a = \frac{1}{2}(M - M^\top)$ be the symmetric and antisymmetric parts of the matrix M, and suppose M_s is positive definite. Take $H = M_s$ and define (with the canonical matrix norm)*

$$\operatorname{skew}(M) \;=\; \|M_s^{-1/2} M_a M_s^{-1/2}\|. \tag{3.5}$$

Then Theorem 3.1 applies with $\mu = 1$ and $\tilde{\kappa}_1 = \operatorname{skew}(M)$, which is the minimal Lipschitz constant for this case. The projection algorithm thus attains the global contraction rate

$$\overline{\theta} = \theta_{\overline{\lambda}} = \frac{1}{\sqrt{1+\dfrac{1}{\operatorname{skew}(M)^2}}}, \qquad \textit{for } \overline{\lambda} = \frac{1}{1+\operatorname{skew}(M)^2}. \tag{3.6}$$

Proof. Here $\widetilde{T}_1(x) = (F - \mu H)(x) = M_a x + q$, an affine monotone mapping devoid of strong monotonicity. We must verify that the specified value of $\tilde{\kappa}_1$ serves as the minimal Lipschitz constant for this mapping from the norm $\|\cdot\|_H$ to the norm $\|\cdot\|_{H^{-1}}$. The square of the required constant is the supremum of the quotient

$$\frac{\big\|\widetilde{T}_1(x') - \widetilde{T}_1(x)\big\|_{H^{-1}}^2}{\|x'-x\|_H^2} = \frac{\big\langle M_a[x'-x], M_s^{-1} M_a[x'-x]\big\rangle}{\big\langle [x'-x], M_s[x'-x]\big\rangle} = \frac{\big\|M_s^{-1/2} M_a[x'-x]\big\|^2}{\big\|M_s^{1/2}[x'-x]\big\|^2}$$

over all x and x' with $x' \neq 0$. Through the change of variables $u = M_s^{1/2}[x'-x]$, giving $[x'-x] = M_s^{-1/2}u$, we see that the constant is the supremum of the expression $\big\|M_s^{-1/2} M_a M_s^{-1/2} u\big\| / \|u\|$ over all $u \neq 0$, and this is $\|M_s^{-1/2} M_a M_s^{-1/2}\|$. □

The value $\operatorname{skew}(M) \in (0, \infty)$ in (3.5) intrinsically measures the *skewness* of the matrix M. Obviously

$$\operatorname{skew}(M) \leq \|M_a\| \big/ \|M_s\| \tag{3.7}$$

in particular, but the right side of this inequality is dependent on the "conditioning" of M with respect to the canonical norm, whereas $\operatorname{skew}(M)$ itself isn't. The smaller $\operatorname{skew}(M)$ is, the nearer M is to being symmetric and the better the rate of convergence that is assured for the solution method addressed by Theorem 3.3. Of course, this realization of forward-backward splitting is practical only when it's easy to project onto C with respect to the norm induced by M_s as H, but that does cover many applications in which C has a product structure matched by a box-diagonal pattern of M_s, as in [20].

Dupuis and Darveau [28], in building on the result of Dafermos [7], likewise obtained for the affine variational inequality case of projection algorithms a contraction factor incorporating the value $\|M_s^{-1/2} M_a M_s^{-1/2}\|$. But the factor they got resembles the one in (3.4) in being the square root of an expression that depends in part on the condition number of H. In contrast to our contraction factor in (3.3), it doesn't tend to 0 as M approaches symmetry and the implementation matrix $H = M_s$ coalesces with M. Again, the cost of deviating from the underlying geometry is evident.

The result in Theorem 3.3 can best be compared with a recent result of Zanni [31] for the same method. He obtains the rate

$$\sqrt{1 - \frac{1}{\|M_s^{-1/2} M M_s^{-1/2}\|^2}} \quad \text{for } \lambda = \frac{1}{\|M_s^{-1/2} M M_s^{-1/2}\|^2}, \tag{3.8}$$

which he elaborates by the estimate

$$\|M_s^{-1/2} M M_s^{-1/2}\| \leq 1 + \operatorname{cond}(M_s) \frac{\|M_a\|}{\|M_s\|}, \tag{3.9}$$

taking the ratio $\|M_a\|/\|M_s\|$ as a measure of skewness. The appearance of M instead of M_a in (3.8) can be seen as reflecting a reliance on a Lipschitz constant for M instead of for M_a; this parallels the difference between Corollary 3.2 and Theorem 3.1. The estimate in (3.9) suffers from dependence on translation to the canonical norm, but to avoid this it could be replaced by

$$\begin{aligned} \|M_s^{-1/2} M M_s^{-1/2}\| &= \|M_s^{-1/2}(M_s + M_a) M_s^{-1/2}\| \\ &\leq \|M_s^{-1/2} M_s M_s^{-1/2}\| + \|M_s^{-1/2} M_a M_s^{-1/2}\| = 1 + \operatorname{skew}(M). \end{aligned}$$

Yet even so it wouldn't yield the lower contraction factor in Theorem 3.3.

Yet another measure of skewness was introduced by Marcotte and Guélat [42] for the special context of solving problems of traffic equilibrium. This differs from ours in being localized to the solution point $\bar{x}$ and dependent on the vector q as well as on the submatrices M_s and M_a. These authors nonetheless demonstrate through numerical testing of several algorithms an empirical relationship between skewness and difficulty of solvability such as appears in Theorem 3.3.

For projected *gradient* algorithms, where $F = \nabla f$ for a $\mathcal{C}^2$ function f with bounded Hessians $\nabla^2 f(x)$, better contraction estimates can be given than are obtainable by specializing the ones here; see Polyak [43].

4. LOCAL CONVERGENCE ANALYSIS.

Our efforts so far have gone into identifying a rate of linear convergence that's effective immediately from any starting point x_0 for a forward-backward splitting method. There is interest too, of course, in knowing what might be possible with convergence as the solution $\bar{x}$ is neared. For this purpose we don't have to start building up a broader theory but can make use of the results we already have. Although Theorem 2.4 presents a contraction rate relative to the entire set $D = \operatorname{dom} T$, its formulation already allows us to deduce local contraction rates in a neighborhood of $\bar{x}$.

Theorem 4.1 (local contraction rates)**.** *Let U be an open ball around $\bar{x}$ with respect to the norm $\|\cdot\|_H$, and let $\hat{\mu}_1$, $\hat{\mu}_2$, and $\hat{\kappa}_1$ be constants as in (A) but relative to $D\cap U$ in place of D. Then, as long as $\lambda > 0$ is small enough that*

$$\lambda^{-1} > \frac{\hat{\mu}_1 - \hat{\mu}_2}{2} + \frac{\hat{\kappa}_1}{2}\max\left\{1, \frac{\hat{\kappa}_1}{\hat{\mu}_1 + \hat{\mu}_2}\right\}, \tag{4.1}$$

the mapping S_λ carries $D\cap U$ into $D\cap U$, and the conclusions of Theorem 2.4 hold for this localization of S_λ, but with $\hat{\mu}_1$, $\hat{\mu}_2$, and $\hat{\kappa}_1$ in place of μ_1, μ_2, and $\tilde{\kappa}_1$.

Proof. Taking $C = \operatorname{cl} U$, define $\widehat{T}_2 = T_2 + N_C$. This mapping, like T_2, is maximal monotone; cf. [39, Thm. 2]. Proposition 2.1 and Theorem 2.4 are applicable to $\widehat{T} = T_1 + \widehat{T}_2$ with respect to the constants $\hat{\mu}_1$, $\hat{\mu}_2$ and $\hat{\kappa}_1$ on $\widehat{D} = \operatorname{dom}\widehat{T}_2 = D\cap C$. In particular, $\widehat{T}^{-1}(0)$ must be a singleton, but because $\bar{x}$ belongs to the interior of C, we have $N_C(\bar{x}) = \{0\}$ and $\widehat{T}(\bar{x}) = T(\bar{x})$. Hence $\widehat{T}^{-1}(0) = \{\bar{x}\}$, and the contraction properties given by Theorem 2.4 for the mapping $\widehat{S}_\lambda = (H + \lambda\widehat{T}_2)^{-1}(H - \lambda T_1)$ must refer to this same $\bar{x}$. Distances from $\bar{x}$ can then only be decreased under $\widehat{S}_\lambda$, so $\widehat{S}_\lambda$ must carry $D\cap U$ into itself.

Consider now any $x \in D\cap U$ and let $w = \widehat{S}_\lambda(x)$. As just seen, we have $w \in D\cap U$, which implies that w belongs to the interior of C, so $N_C(w) = \{0\}$. From the definition of $\widehat{S}_\lambda$ we see that

$$(H - \lambda T_1)(x) \in (H + \lambda\widehat{T}_2)(w) = (H + \lambda T_2)(w) + N_C(w) = (H + \lambda T_2)(w),$$

hence in fact $w = (H + \lambda T_2)(H - \lambda T_1)(x) = S_\lambda(x)$. This shows that $\widehat{S}_\lambda$ agrees with S_λ on $D\cap U$. The conclusions about the behavior of $\widehat{S}_\lambda$ on $D\cap U$ therefore translate to ones about S_λ. □

The proof of Theorem 2.4 reveals a way of refining that result, and with it Corollary 2.5 and Theorem 4.1. Although the monotonicity of $T_2 - \mu_2 H$ is fully utilized in obtaining a Lipschitz constant for the factor $(H+\lambda T_2)^{-1}$ of S_λ, the assumptions in (A) about μ_1 and $\tilde{\kappa}_1$ could be weakened if instead of asking for S_λ to be contractive on D we merely asked for a bound in $[0,1)$ on the ratios $\|S_\lambda(x) - S_\lambda(\bar{x})\|_H / \|x - \bar{x}\|_H$. The key is just to observe that if the argument for estimating $\|(H - \lambda T_1)(x') - (H - \lambda T_1)(x)\|_{H^{-1}}$ is applied only to $\|(H - \lambda T_1)(x) - (H - \lambda T_1)(\bar{x})\|_{H^{-1}}$, all that one needs from μ_1 and $\tilde{\kappa}_1$ is that the mapping $\widetilde{T}_1 = T_1 - \mu_1 H$ satisfies

$$\big\langle x - \bar{x}, \widetilde{T}_1(x) - \widetilde{T}_1(\bar{x})\big\rangle \geq 0, \qquad \big\|\widetilde{T}_1(x) - \widetilde{T}_1(\bar{x})\big\|_{H^{-1}} \leq \tilde{\kappa}_1\big\|x - \bar{x}\big\|_H.$$

Likewise, under these inequalities the estimate in Proposition 2.1 remains valid with respect to a constant κ_1 merely satisfying

$$\big\|T_1(x) - T_1(\bar{x})\big\|_{H^{-1}} \leq \kappa_1\big\|x - \bar{x}\big\|_H.$$

This refinement appears to offer little advantage in general over the global picture in Theorem 2.4, inasmuch as special properties of T_1 and T_2 around the solution point $\bar{x}$, in contrast to other points, can hardly be available in advance of calculating $\bar{x}$, which threatens a kind of circularity. Indeed, if the assumption on μ_1 is weakened the very existence and uniqueness of $\bar{x}$ could be thrown into doubt, because Theorem 2.3 might no longer be applicable. Yet in the localized context of Theorem 4.1 the refinement does at least furnish insights into what might be expected of the rate of in the tail of a forward-backward sequence, as x_k nears $\bar{x}$. The following is what we get.

Theorem 4.2 (ultimate linear convergence rate)**.** *Assuming (A), define the constants* $\bar\mu_1 \geq 0$, $\bar\mu_2 \geq 0$, *and* $\bar\kappa_1 \geq 0$ *at the unique solution point* $\bar x$ *by*

$$\bar\mu_2 = \limsup_{\substack{w\in T_2(x),\, w'\in T_2(x')\\ x,x'\to\bar x,\, x'\neq x}} \frac{\langle x'-x,\, w'-w\rangle}{\|x'-x\|_H^2},$$

$$\bar\mu_1 = \limsup_{\substack{x\to\bar x\\ x\in D,\, x\neq\bar x}} \frac{\langle x-\bar x,\, T_1(x)-T_1(\bar x)\rangle}{\|x-\bar x\|_H^2},$$

$$\bar\kappa_1 = \limsup_{\substack{x\to\bar x\\ x\in D,\, x\neq\bar x}} \frac{\|\overline{T}_1(x)-\overline{T}_1(\bar x)\|_{H^{-1}}}{\|x-\bar x\|_H} \quad \textit{for} \quad \overline{T}_1 = T_1 - \bar\mu_1 H,$$

necessarily obtaining $\bar\mu_1 \geq \mu_1$, $\bar\mu_2 \geq \mu_2$, *and* $\bar\kappa_1 \leq \tilde\kappa_1$, *in fact* $\bar\kappa_1 \leq \sqrt{\tilde\kappa_1^2 - (\bar\mu_1-\mu_1)^2}$ *(hence* $\bar\mu_1 \leq \mu_1 + \tilde\kappa_1$*). Also define*

$$\gamma = \liminf_{\substack{(x,u)\to(\bar x, T_1(\bar x))\\ -u\in T_2(x)}} \frac{\|u-T_1(\bar x)\|_{H^{-1}}}{\|x-\bar x\|_H},$$

necessarily obtaining $\gamma \geq \bar\mu_2$. *Then for any step size* $\lambda > 0$ *the sequence of points* x_k *generated by* $x_k = S_\lambda(x_{k-1})$ *from any starting point* $x_0 \in D$ *will satisfy*

$$\limsup_{k\to\infty} \frac{\|x_k-\bar x\|_H}{\|x_{k-1}-\bar x\|_H} \leq \begin{cases} \dfrac{\sqrt{(1-\lambda\bar\mu_1)^2+\lambda^2\bar\kappa_1^2}}{\sqrt{1+2\lambda\bar\mu_2+\lambda^2\gamma^2}} & \textit{when } \lambda^{-1}\geq\bar\mu_1, \\[2ex] \dfrac{\lambda(\bar\kappa_1+\bar\mu_1)-1}{\sqrt{1+2\lambda\bar\mu_2+\lambda^2\gamma^2}} & \textit{when } \lambda^{-1}\leq\bar\mu_1. \end{cases} \tag{4.2}$$

In particular this holds for the step size $\bar\lambda$ *identified in (2.6) as optimal relative to the globally estimated constants* μ_1, μ_2, *and* $\tilde\kappa_1$.

Proof. It's clear from (A) that $\bar\mu_1 \geq \mu_1$ and $\bar\mu_2 \geq \mu_2$, since the monotonicity of $T_i - \mu_i H$ on D corresponds to having $\langle x'-x, T_i(x')-T_i(x)\rangle \geq \mu_i\|x'-x\|_H^2$ for $x, x' \in D$. The verification that $\bar\kappa_1 \leq \tilde\kappa_1$ takes more effort. It relies indirectly on the observation above that Proposition 2.1 stays valid when the context is that of points x compared to $\bar x$ rather than general pairs x and x'. If actually $\bar\mu_1 = \mu_1$, we have $\overline{T}_1 = \widetilde{T}_1$ and the inequality $\bar\kappa_1 \leq \tilde\kappa_1$ is elementary from the definitions, so we can concentrate on the case where $\bar\mu_1 > \mu_1$.

Consider any $\delta \in (0, \bar\mu_1 - \mu_1)$. From the definition of $\bar\mu_1$ there's a neighborhood Z of $\bar x$ consisting of points x for which $\langle x-\bar x,\, T_1(x)-T_1(\bar x)\rangle \geq (\bar\mu_1-\delta)\|x-\bar x\|_H^2$. This inequality means for the mapping $\overline{T}_1^\delta = T_1 - (\bar\mu_1-\delta)H = \overline{T}_1 - \delta H$ that

$$\langle x-\bar x,\, \overline{T}_1^\delta(x)-\overline{T}_1^\delta(\bar x)\rangle \geq 0 \quad \text{for } x\in D\cap Z. \tag{4.3}$$

But $\overline{T}_1^\delta = \widetilde{T}_1 - \tau H$ for $\tau = \bar\mu_1 - \mu_1 - \delta > 0$. It follows from applying the extended version of Proposition 2.1 to this relation in light of (4.3) that

$$\|\overline{T}_1^\delta(x)-\overline{T}_1^\delta(\bar x)\|_{H^{-1}} \leq \sqrt{\tilde\kappa_1^2-\tau^2}\,\|x-\bar x\|_H \quad \text{for } x\in D\cap Z.$$

On the other hand, since $\overline{T}_1^\delta = \overline{T}_1 - \delta H$ and $\|H[x-\bar{x}]\|_{H^{-1}} = \|H(x-\bar{x})\|_{H^{-1}} = \|x'-x\|_H$, we know that $\big\|\overline{T}_1(x)-\overline{T}_1(\bar{x})\big\|_{H^{-1}} \leq \big\|\overline{T}_1^\delta(x)-\overline{T}_1^\delta(\bar{x})\big\|_{H^{-1}} + \delta\big\|x-\bar{x}\big\|_H$. This tells us that

$$\big\|\overline{T}_1(x)-\overline{T}_1(\bar{x})\big\|_{H^{-1}} \leq \left(\delta + \sqrt{\tilde{\kappa}_1^2 - (\bar{\mu}_1-\mu_1-\delta)^2}\right)\big\|x-\bar{x}\big\|_H \quad \text{for } x \in D\cap Z.$$

Taking the limit in the definition of $\bar{\kappa}_1$ and using the fact that a neighborhood Z like this exists for any $\delta > 0$, we obtain $\bar{\kappa}_1 \leq \sqrt{\tilde{\kappa}_1^2 - (\bar{\mu}_1-\mu_1)^2}$, hence $\bar{\kappa}_1 \leq \tilde{\kappa}_1$ because $\bar{\mu}_1 \geq \mu_1$.

We look next at the claims about γ and θ_λ^*, the latter being the symbol by which we'll denote the right side of (4.2). For any $\epsilon > 0$, let $\hat{\mu}_i = \max\{\mu_i, \bar{\mu}_i - \epsilon\}$ for $i = 1,2$ and $\hat{\kappa}_1 = \min\{\bar{\kappa}_1, \bar{\kappa}_1+\epsilon\}$. On the basis of the definitions we know there's a ball U around $\bar{x}$ with respect to the norm $\|\cdot\|_H$ such that

$$\begin{cases} \langle x'-x,\, w'-w\rangle \geq \hat{\mu}_2\|x'-x\|_H & \text{if } x,x' \in D\cap U,\ w\in T_2(w),\ w'\in T_2(x'),\\ \langle x-\bar{x},\, T_1(x)-T_1(\bar{x})\rangle \geq \hat{\mu}_1\|x-\bar{x}\|_H & \text{if } x\in D\cap U,\\ \|\overline{T}_1(x)-\overline{T}_1(\bar{x})\|_{H^{-1}} \leq \hat{\kappa}_1\|x-\bar{x}\|_H & \text{if } x\in D\cap U. \end{cases}$$

We are then in the framework of the extended version of Theorem 4.1 and are able to see that $\limsup_k \|x_k-\bar{x}\|_H \big/ \|x_{k-1}-\bar{x}\|_H \leq \widehat{\theta}_\lambda$, the latter being the same as θ_λ except that $\hat{\mu}_1$, $\hat{\mu}_2$ and $\hat{\kappa}_1$ replace μ_1, μ_2, and $\tilde{\kappa}_1$. An improvement can be made, however, in taking advantage of the constant γ.

Let $\hat{\gamma} = \max\{0, \gamma-\epsilon\}$. From the definition of γ there's a neighborhood V of $T_1(\bar{x})$ with respect to the norm $\|\cdot\|_{H-1}$ such that, when the ball U is small enough, we have

$$\big\|u - T_1(\bar{x})\|_{H^{-1}} \geq \hat{\gamma}\|x-\bar{x}\|_H^2 \quad \text{when} \quad x\in D\cap U,\ u\in V,\ -u\in T_2(x). \tag{4.4}$$

But also, the point $\bar{u} = T_1(\bar{x})$ satisfies $-\bar{u}\in T_2(\bar{x})$, because $0\in T(\bar{x}) = T_1(\bar{x})+T_2(\bar{x})$. This implies from earlier that

$$\hat{\mu}_2\|x-\bar{x}\|_H^2 \;\leq\; \big\langle -u+\bar{u},\, x-\bar{x}\big\rangle \;\leq\; \|x-\bar{x}\|_H\|u-\bar{u}\|_{H^{-1}},$$

so that $\hat{\mu}_2\|x-\bar{x}\|_H \leq \|u-\bar{u}\|_{H^{-1}}$. Therefore $\hat{\mu}_2 \leq \hat{\gamma}$, which establishes $\bar{\mu}_2 \leq \gamma$ through the arbitrariness of ϵ in the definition of $\hat{\mu}_2$ and $\hat{\gamma}$.

Let $\bar{w} = (H-\lambda T_1)(\bar{x}) = H\bar{x} - \lambda T_1(\bar{x})$. Since

$$\bar{x} = S_\lambda(\bar{x}) = (H+\lambda T_2)^{-1}(H-\lambda T_1)(\bar{x}),$$

we have $\bar{x} = (H+\lambda T_2)^{-1}(\bar{w})$. Consider along with this any elements w and x with $x = (H+\lambda T_2)^{-1}(w)$. For these the set $T_2^{-1}(x)$ contains $\lambda^{-1}[w-Hx]$, whereas $T_2^{-1}(\bar{x})$ contains $\lambda^{-1}[\bar{w}-H\bar{x}]$, the latter being just $-T_1(\bar{x})$. When w is close to $\bar{w}$, not only does x lie in the ball U around $\bar{x}$, due to the continuity of $(H+\lambda T_2)^{-1}$, but also the vector $u = -\lambda^{-1}[w-Hx]$ lies in the neighborhood V of $\bar{u} = -T_1(\bar{x})$. Then by (4.4),

$$\begin{aligned} \hat{\gamma}^2\|x-\bar{x}\|_H^2 \;&\leq\; \big\|u-T_1(\bar{x})\|_{H^{-1}}^2 \;=\; \big\|-\lambda^{-1}[w-Hx]+\lambda^{-1}[\bar{w}-H\bar{x}]\big\|_{H^{-1}}^2 \\ &=\; \lambda^{-2}\|w-\bar{w}\|_{H^{-1}}^2 - 2\lambda^{-2}\big\langle w-\bar{w},\, x-\bar{x}\big\rangle + \|x-\bar{x}\|_H^2 \\ &\leq\; \lambda^{-2}\|w-\bar{w}\|_{H^{-1}}^2 - 2\lambda^{-2}\hat{\mu}_2\|x-\bar{x}\|_H^2 + \lambda^{-2}\|x-\bar{x}\|_H^2, \end{aligned}$$

where the last inequality invokes the property arranged for $\hat{\mu}_2$. Rearranging, we obtain $\|x-\bar{x}\|_H^2 \le [1+\lambda\hat{\mu}_2+\lambda^2\hat{\gamma}^2]\|w-\bar{w}\|_{H^{-1}}^2$. This shows that the factor $(1+\lambda\hat{\mu}_2)^{-1}$ in $\widehat{\theta}_\lambda$ can be replaced by $(1+\lambda\hat{\mu}_2+\lambda^2\hat{\gamma}^2)^{-1/2}$, which if anything is lower.

It remains only to observe that, having demonstrated that this modified factor $\widehat{\theta}_\lambda$ operates in terms of $\hat{\mu}_1$, $\hat{\mu}_2$, $\hat{\kappa}_1$ and $\hat{\gamma}$ as defined for arbitrary $\epsilon > 0$, we must in the limit as $\epsilon \searrow 0$ get the factor θ_λ^* corresponding to $\bar{\mu}_1$, $\bar{\mu}_2$, $\bar{\kappa}_1$ and γ. □

5. VARIABLE STEP SIZES AND MATRICES.

In the introduction, forward-backward splitting methods were described with variable step sizes λ_k and matrices H_k. We now look at what can be said about such methods on the basis of our contraction results for fixed λ and H. The easier case of variable λ_k with a fixed H has broader significance, so we deal with it first.

Theorem 5.1 (convergence with variable step sizes)**.** *Under assumptions (A), consider any step size interval* $[\lambda_-, \lambda_+] \subset (0,\infty)$ *with* λ_+ *small enough that*

$$\lambda_+^{-1} \;>\; \frac{\mu_1-\mu_2}{2} + \frac{\tilde{\kappa}_1}{2}\max\left\{1, \frac{\tilde{\kappa}_1}{\mu_1+\mu_2}\right\}. \tag{5.1}$$

Let $\theta(\lambda_-,\lambda_+) = \max\{\theta_{\lambda_-},\theta_{\lambda_+}\}$ *for* θ_λ *defined as in (2.4). Then* $\theta(\lambda_-,\lambda_+) < 1$*, and for any sequence of step sizes* $\lambda_k \in [\lambda_-,\lambda_+]$ *all the iteration mappings*

$$S_k = (H+\lambda_k T_2)^{-1}(H-\lambda_k T_1) = (I+\lambda_k H^{-1}T_2)^{-1}(I-\lambda_k H^{-1}T_1) \tag{5.2}$$

are contractions from $D = \operatorname{dom} T$ *into itself with fixed point* $\bar{x}$ *and contraction factor* $\theta(\lambda_-,\lambda_+)$*. In particular, the iterates* $x_k = S_k(x_{k-1})$ *from any starting point* $x_0 \in D$ *converge linearly to* $\bar{x}$ *at a rate no worse than* $\theta(\lambda_-,\lambda_+)$*. Indeed,*

$$\limsup_{k\to\infty} \frac{\|x_k-\bar{x}\|_H}{\|x_{k-1}-\bar{x}\|_H} \;\le\; \min\big\{\theta(\lambda_-,\lambda_+), \theta^*(\lambda_-,\lambda_+)\big\}, \tag{5.3}$$

where $\theta^*(\lambda_-,\lambda_+) = \min\{\theta_{\lambda_-}^*,\theta_{\lambda_+}^*\}$ *with* θ_λ^* *denoting the right side of (4.2).*

Proof. The justification of this lies in the proof of Theorem 2.4. It was demonstrated there that θ_λ is a increasing function of λ on the interval of λ values satisfying $\lambda^{-1} < \mu_1$, which includes all λ sufficiently large. On the other hand, it was observed that on the complementary interval, where $\lambda^{-1} \ge \mu_1$, the expression θ_λ^2 is convex as a function of τ under the change of variables induced by taking $\tau^{-1} = \lambda^{-1}+\mu_2$. This implies that θ_λ^2 is unimodal on that interval with respect to λ, and the same then holds for θ_λ. Indeed, we saw for the value $\overline{\lambda}$ defined in (2.6) that θ_λ is a continuous, decreasing function of λ on $\big(0,\overline{\lambda}\,\big]$ but a continuous, increasing function of λ on $\big[\,\overline{\lambda},\infty\big)$.

It follows that the max of θ_λ over any interval $[\lambda_-,\lambda_+] \subset (0,\infty)$ is $\theta(\lambda_-,\lambda_+)$. As long as this value doesn't exceed 1, as guaranteed by (5.1) through Theorem 2.4, we get contraction at the claimed rate $\theta(\lambda_-,\lambda_+)$. An appeal to the ultimate convergence property in Theorem 4.2 then justifies the assertion in (5.3). □

For the case of variable implementation matrices, we won't attempt to prove a result along the lines of a Newton or quasi-Newton method. That would anyway be incompatible with most applications of forward-backward splitting to problem decomposition, where the need to preserve a degree of separability, in order to facilitate computation of the backward steps, is paramount. Also, such applications tend to demand a global statement rather than a local one. For literature on Newton-like results in the context of variational inequalities, see Pang and Chen [16] and Patriksson [17].

Theorem 5.2 (convergence with variable matrices). *Under (A), suppose the iterates* $x_k = S_k(x_{k-1})$ *are generated from any* $x_0 \in D$ *by the mappings*

$$S_k = (H_k + \lambda_k T_2)^{-1}(H_k - \lambda_k T_1) = (I + \lambda_k H_k^{-1} T_2)^{-1}(I - \lambda_k H_k^{-1} T_1) \tag{5.4}$$

through a sequence of step sizes $\lambda_k > 0$ *and symmetric, positive definite matrices* H_k *converging to* H. *Let* $\lambda_- = \liminf_k \lambda_k$ *and* $\lambda_+ = \limsup_k \lambda_k$, *and suppose that* $\lambda_- > 0$ *while* λ_+ *satisfies (5.1). Then (5.3) holds for these values* λ_- *and* λ_+.

Proof. The convergence of H_k to H implies the existence of values $0 < \alpha_k \nearrow 1$ and $0 < \beta_k \nearrow 1$ such that $H - \alpha_k H_k$ and $H_k - \beta_k H$ are positive definite. Through this, the monotonicity of $T_1 - \mu_1 H$ and $T_2 - \mu_2 H$ in condition (2.1) of (A) yields the monotonicity of $T_1 - \mu_{1k} H_k$ and $T_2 - \mu_{2k} H_k$ for the values $\mu_{1k} = \mu_1 \alpha_k \nearrow \mu_1$ and $\mu_{2k} = \mu_2 \alpha_k \nearrow \mu_2$.

We develop now a Lipschitz constant for $\widetilde{T}_{1k} = T_1 - \mu_{1k} H_k = \widetilde{T}_1 + \mu_1(H - \alpha_k H_k)$ from the norm $\|\cdot\|_{H_k}$ to the norm $\|\cdot\|_{H_k^{-1}}$. First, let η_k be such a constant for $H - \alpha_k H_k$; then $\eta_k \to 0$. Next, observe that $\|\cdot\|_{H_k} \geq \sqrt{\beta_k}\,\|\cdot\|_H$, which for the corresponding dual norms, given by the inverse matrices, means that $\sqrt{\beta_k}\,\|\cdot\|_{H_k^{-1}} \leq \|\cdot\|_{H^{-1}}$. By these estimates, the Lipschitz inequality in condition (2.2) of (A) gives us

$$\sqrt{\beta_k}\,\|\widetilde{T}_1(x') - \widetilde{T}_1(x)\|_{H_k^{-1}} \leq \tilde{\kappa}_1(1/\sqrt{\beta_k})\|x' - x\|_{H_k} \quad \text{for all } x', x \in D.$$

Hence $\tilde{\kappa}_1/\beta_k$ serves as a Lipschitz constant for $\widetilde{T}_1$ on D from $\|\cdot\|_{H_k}$ to $\|\cdot\|_{H_k^{-1}}$. Since $\widetilde{T}_{1k} = \widetilde{T}_1 + \mu_1(H - \alpha_k H_k)$, we conclude that the constant $\tilde{\kappa}_{1k} = (\tilde{\kappa}_1/\beta_k) + \mu_1 \eta_k \searrow \tilde{\kappa}_1$ has this property for $\widetilde{T}_{1k}$.

It follows that the splitting $T = T_{1k} + T_{2k}$ with implementation matrix H_k satisfies (A_k), the version of (A) in which where μ_1, μ_2 and $\tilde{\kappa}_1$ are replaced by μ_{1k}, μ_{2k} and $\tilde{\kappa}_{1k}$. Now let ϕ stand for the value on the right side of (2.5) and ϕ_k for the corresponding value under this same replacement of constants. Obviously $\phi_k \to \phi$.

Consider any $\epsilon > 0$ small enough that the value $\lambda_-^\epsilon = \lambda_- - \epsilon$ is positive, while the value $\lambda_+^\epsilon = \lambda_+ + \epsilon$ satisfies (5.1), i.e., $(\lambda_+^\epsilon)^{-1} > \phi$. For all k sufficiently large we have $\lambda_k \in [\lambda_-^\epsilon, \lambda_+^\epsilon]$ and also that $(\lambda_+^\epsilon)^{-1} > \phi_k$. Then by Theorem 2.4 as applied under (A_k), the mapping S_k is a contraction from D into itself at the rate θ_{k,λ_k}, where $\theta_{k,\lambda}$ denotes the factor obtained from formula (2.4) with μ_{1k}, μ_{2k} and $\tilde{\kappa}_{1k}$ substituting for μ_1, μ_2 and $\tilde{\kappa}_1$. Furthermore, we have

$$\theta_{k,\lambda} \leq \theta_k(\lambda_-^\epsilon, \lambda_+^\epsilon) = \max\{\theta_{k,\lambda_-^\epsilon}, \theta_{k,\lambda_+^\epsilon}\}$$

for the reasons in the proof of Theorem 5.1 (when applied to $\theta_{k,\lambda}$ as a function of λ). The lim sup in (5.3) is bounded above, therefore by the limit of $\theta_k(\lambda_-^\epsilon, \lambda_+^\epsilon)$ as $k \to \infty$, which is $\theta(\lambda_-^\epsilon, \lambda_+^\epsilon)$. This being valid for all $\epsilon > 0$ sufficiently small, we can take the limit as $\epsilon \searrow 0$ and obtain the inequality in (5.3), as targeted. □

6. ASYMMETRIC IMPLEMENTATIONS.

Only symmetric implementation matrices H_k are covered directly by our results up to this stage, but what about the possibility of more general matrices that are not symmetric, although still positive definite? Such modes of implementation crop up for example in applications to variational inequality when H_k is taken to be an approximation to the Jacobian matrix $\nabla F(x_k)$ or some part of it. Aside from the gradient case where $F = \nabla f$ and $\nabla F(x_k) = \nabla^2 f(x_k)$, H_k may then lack symmetry.

Asymmetric implementation matrices can be incorporated into our theory by a simple device. This device has already used by others, e.g. Tseng in [25], but we go beyond previous instances because of the attention we pay to step sizes. To explain the idea we keep to the case of constant H for simplicity, and also, to avoid conflicts with our earlier statements, follow the notational strategy of replacing H by $H+K$ with K antisymmetric ($K^\top = -K$) and H still symmetric, rather than taking H itself to lack symmetry. This conforms to the fact that any positive definite matrix can be written as the sum of an antisymmetric matrix and a symmetric, positive definite matrix.

In this mode, the iteration mappings for the forward-backward method with respect to a splitting $T = T_1 + T_2$ take the form

$$\big([H+K] + \lambda T_2\big)^{-1}\big([H+K] - \lambda T_1\big). \tag{6.1}$$

Their practicality hinges on the ease of calculating images under the inverse mapping $\big([H+K]+\lambda T_2\big)^{-1}$. This has to be assumed for any analysis to be worthwhile, and it's true in applications such have been pinpointed by Pang and Chan [16] and Tseng [25].

For our purposes we'll make such practicality of backward step execution part of the framework by assuming that for any $\tau \in (-\infty,\infty)$ the inverse $\big([H+\tau K]+\lambda T_2\big)^{-1}$ can be handled just as readily as $\big([H+K]+\lambda T_2\big)^{-1}$. We put our focus therefore on *two-parameter* iteration mappings

$$S_{\lambda,\tau} = \big([H+\tau K] + \lambda T_2\big)^{-1}\big([H+\tau K] - \lambda T_1\big). \tag{6.2}$$

These mappings, like the earlier ones where K didn't appear, all have the unique solution $\bar{x}$ as their unique fixed point. We explore the relation between contraction properties of $S_{\lambda,\tau}$ and the values of both λ and τ.

Theorem 6.1 (reduction of asymmetric to symmetric implementations). *Assume (A) as before, except for what it says about $\tilde{\kappa}_1$; in place of that, consider a Lipschitz constant $\tilde{\kappa}_1(\sigma)$ for the mapping $\widetilde{T}_1 - \sigma K$ on D, with σ any value in $(-\infty,\infty)$. Let*

$$\lambda(\sigma) = \frac{1}{\left(\dfrac{\tilde{\kappa}_1(\sigma)^2}{\mu_1+\mu_2}\right) + \mu_1}, \qquad \tau(\sigma) = \sigma\lambda(\sigma). \tag{6.3}$$

Then the asymmetrically implemented iteration mapping

$$S_{\lambda(\sigma),\tau(\sigma)} \;=\; \big([H+\tau(\sigma)K] + \lambda(\sigma)T_2\big)^{-1}\big([H+\tau(\sigma)K] - \lambda(\sigma)T_1\big) \tag{6.4}$$

with respect to the splitting $T = T_1 + T_2$ is identical to the symmetrically implemented iteration mapping

$$S'_{\lambda(\sigma)} \;=\; \big(H + \lambda(\sigma)T_2'\big)^{-1}\big(H - \lambda(\sigma)T_1'\big) \tag{6.5}$$

with respect to the splitting $T = T_1' + T_2'$ where $T_1' = T_1 - \sigma K$ and $T_2' = T_2 + \sigma K$, and it is Lipschitz continuous on D with constant

$$\theta(\sigma) = \frac{1}{\sqrt{1 + \left(\dfrac{\mu_1+\mu_2}{\tilde{\kappa}_1(\sigma)}\right)^2}} \;<\; 1. \tag{6.6}$$

Proof. Iterations $x_k = S_{\lambda,\tau}(x_{k-1})$ have the meaning that

$$0 \in \frac{1}{\lambda}[H + \tau K][x_k - x_{k-1}] + T_1(x_{k-1}) + T_2(x_k).$$

This condition can equally well be written as

$$0 \in \frac{1}{\lambda} H[x_k - x_{k-1}] + [T_1 - \sigma K](x_{k-1}) + [T_2 + \sigma K](x_k) \tag{6.7}$$

under the correspondence $\sigma = \tau/\lambda$, $\tau = \sigma\lambda$. Thus, the same iterations can be written as $x_k = S'_\lambda(x_{k-1})$ for $S'_\lambda = (H + \lambda T'_2)^{-1}(H - \lambda T'_1)$. The splitting $T = T'_1 + T'_2$ satisfies (A) with Lipschitz constant $\tilde{\kappa}_1(\sigma)$, so Theorem 2.4 applies. The optimal step size coming out of that result is $\lambda(\sigma)$ as given by (6.3), and it yields for $S'_{\lambda(\sigma)}$ the contraction rate $\theta(\sigma)$ defined in (6.6). □

The observation to be made from Theorem 6.1 is that, instead of pursuing asymmetric implementations directly, a good strategy is to first subtract off from T_1 to get T'_1 whatever multiple σ of the asymmetric part K of the implementation matrix $H+K$ is appropriate in order to reduce the Lipschitz constant $\tilde{\kappa}_1(\sigma)$ as far as possible. This multiple is added to T_2 to get T'_2. Thereafter, it's just a matter of taking the optimal step size $\lambda(\sigma)$ for the altered splitting $T = T'_1 + T'_2$ with respect to the symmetric part H of the implementation matrix, in accordance with the earlier results. The net effect will be the same as the asymmetric iterations (6.4), but executed symmetrically and at an optimized rate.

REFERENCES

[1] P.-L. Lions and B. Mercier, "Splitting algorithms for the sum of two nonlinear operators," SIAM J. Numerical Analysis 16 (1979), 964–979.

[2] A. A. Goldstein, "Convex programming in Hilbert space," Bulletin Amer. Math. Soc. 70 (1964), 709–710.

[3] E. S. Levitin and B. T. Polyak, "Constrained minimization methods," U.S.S.R. Comput. Math. and Math. Physics 6 (1966), 1–50.

[4] H. Brézis and M. Sibony, "Méthodes d'approximation et d'itération pour les opérateurs monotones," Archive Rational Mech. Anal. 27 (1969), 59–82.

[5] M. Sibony, "Methodes iteratives pour les equations et inequations aux derivées partielles nonlinéaires de type monotone," Calcolo 7 (1970), 65–183.

[6] H. Gajewski and R. Kluge, "Projektionsverfahren für nichtlinearen variationsungleichungen," Math. Nachrichten 46 (1970), 363–373.

[7] S. Dafermos, "Traffic equilibrium and variational inequalities," Transportation Science 14 (1980), 42–54.

[8] R. T. Rockafellar, "Monotone mappings and the proximal point algorithm," SIAM J. Control and Optim. 14 (1976), 877–898.

[9] ———, "Augmented Lagrangians and applications of the proximal point algorithm in convex programming," Math. Oper. Res. 1 (1976), 97–116.

[10] J. Eckstein and D. P. Bertsekas, "On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone mappings," Math. Programming 55 (1992), 293–318.

[11] J. E. Spingarn, "Partial inverse of a monotone mapping," Applied Math. & Opt. 10 (1983), 247–265.

[12] J. Eckstein, *Splitting Methods for Monotone Operators with Applications to Parallel Optimization*, Ph.D. Thesis, CICS-TH-140, MIT, 1989.

[13] G. Cohen, "Optimization by decomposition and coordination: a unified approach," IEEE Transactions on Automatic Control AC-23 (1978), 222-232.

[14] ———, "Auxiliary problem principle and decomposition of optimization problems," J. Opt. Theory Appl. 32 (1980), 277–305.

[15] ———, "Auxiliary problem principle extended to variational inequalities," J. Opt. Theory Appl. 59 (1988), 369–390.

[16] J. S. Pang and D. Chan, "Iterative methods for variational and complementarity problems," Math. Programming 24 (1982), 284–313.

[17] M. Patriksson, *A Unified Framework of Descent Algorithms for Nonlinear Programs and Variational Inequalities*, doctoral thesis, Linköping Univ., 1993.

[18] R. T. Rockafellar, "Linear-quadratic programming and optimal control," SIAM J. Control Opt. 25 (1987), 781–814.

[19] ———, "Multistage convex programming and discrete-time optimal control," Control and Cybernetics 17 (1988), 225–246.

[20] R. T. Rockafellar and R. J-B Wets, "Generalized linear-quadratic problems of deterministic and stochastic optimal control in discrete time," SIAM J. Control Opt. 28 (1990), 810–822.

[21] R. Temam and I. Ekeland, *Analyse Convex et Problèmes Variationnels*, Dunod Gauthier-Villars, Paris, 1974.

[22] G. B. Passty, "Ergodic convergence to a zero of the sum of monotone mappings in Hilbert space," J. Math. Anal. Appl. 72 (1979), 383–390.

[23] D. Gabay, "Applications of the method of multipliers to variational inequalities," in: *Augmented Lagrangian Methods: Applications to the Solution of Boundary Value Problems*, M. Fortin and R. Glowinski (eds.), Amsterdam: North-Holland, 1983.

[24] P. Tseng, "Applications of a splitting algorithm to decomposition in convex programming and variational inequalities," SIAM J. Control Opt. 29 (1991), 119–138.

[25] ———, "Further applications of a splitting algorithm to decomposition in variational inequalities and convex programming," Math. Programming 48 (1990), 249–263.

[26] C. Dupuis and J.-M. Darveau, "The convergence conditions of diagonalization and projection methods for fixed demand asymmetric network equilibrium problems," Operations Research Letters 5 (1986), 149–155.

[27] D. P. Bertsekas and E. M. Gafni, "Projection methods for variational inequalities with application to the traffic assignment problem," Math. Programming Study 17 (1982), 139–159.

[28] L. Zanni, "On the convergence rate of two projection methods for variational inequalities in R^n," Calcolo 29 (1992), 193–212.

[29] ———, "Convergence properties of a projection method for affine variational inequality problems," Le Matematiche 49 (1994), 359–377.

[30] A. Renaud, *Algorithmes de Régularisation et Décomposition pour les Problèmes Variationnels Monotones*, doctoral thesis, L'École Nationale Supérieure des Mines de Paris, 1993.

[31] P. Marcotte and J. H. Wu, "On the convergence of projection methods: Application to the decomposition of affine variational inequalities," J. Optim. Theory Appl. 85 (1995), 347–362.

[32] Z. Q. Luo and P. Tseng, "Error bound and convergence analysis of matrix splitting algorithms for the affine variational inequality problem, SIAM J. Opt. 2 (1992), 43–54.

[33] P. Tseng, "On the linear convergence of iterative methods for the variational inequality problem," Computational & Appl. Math. 60 (1995).

[34] P. T. Harker and J.-S. Pang, "Finite-dimensional variational inequality and nonlinear complementarity problems: A survey of theory, algorithms and applications," Math. Programming 48 (1990), 161–220.

[35] G. H-G. Chen, *Forward-Backward Splitting Techniques: Theory and Applications*, doctoral thesis, University of Washington, 1994.

[36] S. Dafermos, "An iterative scheme for variational inequalities," Math. Programming 26 (1983), 40–47.

[37] G. J. Minty, "On the maximal domain o a 'monotone' function," Michigan Math. J. 8 (1961), 135–137.

[38] R. T. Rockafellar, "On the maximality of sums of nonlinear monotone mappings," Transactions of Amer. Math. Soc. 149 (1970), 75–88.

[39] ———, "Local boundedness of nonlinear monotone mappings," Michigan Math. J. 16 (1969), 397–407.

[40] P. Marcotte and F. Guélat, "Adaptation of a modified Newton method for solving the asymmetric traffic equilibrium problem," Transportation Science 22 (1988), 112–124.

[41] B. T. Polyak, *Introduction to Optimization*, New York: Optimization Software, 1987.