Chapter 6

THE CHOICES OF THE PARAMETERS IN EFBS METHODS

6.1. The Computational Experiments and Step Sizes

We have randomly chosen several discrete-time open-loop control problems which are special cases of discrete-time multistage problems in Section 3.1. We only consider the *box-diagonal* case here for simplicity. All examples are time-variant problems. The symbols in the following examples correspond to those in Section 3.1, and we actually solve the saddle point problem associated with the Lagrangian $L(u,v)$ on $U \times V$ in (3.1.5). We choose T_1 and T_2 such that the BFBS method becomes the projection method and we let the computational stop criterion is $\|z^{k+1} - z^k\| \le \varepsilon$.

We have computed 6 examples. We took all of them with $N = 4$, $\varepsilon = 10^{-8}$. The dimensionality of control and state vectors do not depend on time for Example 1, while the dimensionality of control and state vectors do change with time for the next five examples. In fact, Example 4 is a LQR (linear-quadratic regulator) problem. We chose all of P_i and Q_i as positive definite for convenience. We also took $\lambda_k \equiv \lambda$.

Table 6.1. Example Information

Ex.	$u \in$	$v \in$	$x, y \in$	U	V
1	$\mathbb{R}^{1+1+1+1+1}$	$\mathbb{R}^{1+1+1+1+1}$	$\mathbb{R}^{2+2+2+2+2}$	bounded	bounded
2	$\mathbb{R}^{3+3+2+2+3}$	$\mathbb{R}^{2+3+3+2+2}$	$\mathbb{R}^{4+3+2+3+4}$	bounded	bounded
3	same	same	same	bounded	$\mathbb{R}^{12}$
4	same	same	same	$\mathbb{R}^{13}$	$\mathbb{R}^{12}$
5	same	same	same	bounded	bounded
6	same	same	same	bounded	bounded

Let N^i be the stop iteration numbers. The computational time is too short to be counted by the computer so that we could not give it. The results are in the following.

Table 6.2. Computation Results			
Ex.	λ	N^i	$\Vert x^N - x^{N-1}\Vert$
1	0.5	28	5.59×10^{-9}
2	0.5	26	7.45×10^{-9}
3	0.5	27	5.27×10^{-9}
4	0.5	27	5.27×10^{-9}
5	0.5	40	0
6	0.5	162	0

Table 6.3. Influence of λ on Convergence			
λ	N_1^i	N_6^i	t_6
0.05	–	1602	9.7s
0.1	158	802	5s
0.15	106	536	3.3s
0.25	62	322	1.9s
0.35	–	231	**
0.5	28	162	**
0.75	15	109	**
0.85	11	96	**
0.95	8	87	**
1.0	3	82	**
1.05	8	78	**
1.15	11	72	**
1.25	15	66	**
1.5	29	56	**
1.6	–	52	**
1.65	–	51	**
1.7	–	53	**
1.75	69	65	**
1.85	121	114	**
1.9	186	176	**
1.95	382	360	2.2s
2.0	diverge(bdd)	diverge(bdd)	∞

The influence of λ on the convergence of $\{z^k\}$ seems evident. We computed Example 1 and Example 6 with different λ values. In the table 6.3, N_1^i and N_6^i represent the iteration times for Example 1 and Example 6 respectively and t_6 stands for the computational time of Example 6. While ** means that the computational time is too short to be counted. The computational time of Example

1 is ignored since it is too short. Symbol '–' means that we did not compute it.

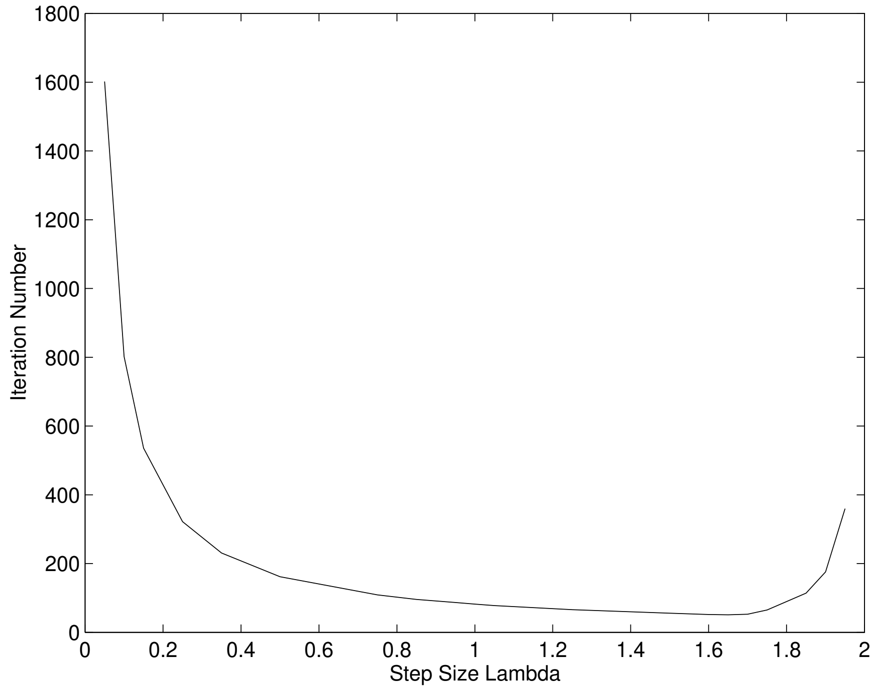

Figure 6.1: Influence of Step Sizes on Convergence for Example 6

When $\lambda = 2$, both diverge because z^k began to jump on two points after certain iterations. From the above table, it is easy to see that there is a 'best' λ in both examples. It looks interesting that the iteration numbers of Example 1 are almost symmetric to $\lambda = 1$. We shall do further analysis in our next section.

Of course, the different choices of H_k, splitting $T_1 + T_2$, and starting point z^0 will also influence the convergence of $\{z^k\}$. Some numerical comparisons of these choices and related figures can be found in Section 7.9 of next chapter. We will leave the theoretical discussion to the later sections of this chapter.

6.2. The Suboptimal Choice of $\{\lambda_k\}$ for BFBS Methods

In the last section, we give some experiments for influence of different $\{\lambda_k\}$ on the convergence of the BFBS methods and pointed out that there should be a sequence of optimal λ_k such that the BFBS method converges fastest. Here we try to give some discussion.

Definition 6.2.1. *Suppose $\{z^k\}$ is from (2.10.1) and $z^* \in zer(T)$ is its limit point. If there is a function $c_k : \mathbb{R} \to \mathbb{R}$ for every $k \geq 0$ such that*

$$\|z^{k+1} - z^*\|^2 \leq c_k(\lambda_k)\|z^k - z^*\|^2, \tag{6.2.1}$$

Or

$$\|z^{k+1} - z^*\|^2 \leq \|z^k - z^*\|^2 + c_k(\lambda_k), \tag{6.2.2}$$

then we call $\bar{\lambda}_k$ and $\bar{c}_k$ as suboptimal parameter and value of (2.10.1) from (6.2.1) or (6.2.2) respectively for every $k \geq 0$. Where

$$\bar{\lambda}_k \in \operatorname{argmin} c_k(\lambda), \qquad \bar{c}_k := c_k(\bar{\lambda}_k). \tag{6.2.3}$$

Proposition 6.2.1. *Under the Condition C of Section 4.3, a suboptimal $\bar{\lambda}_k$ and a corresponding $\bar{c}_k$ of (2.10.1) from (6.2.1) are*

$$\bar{\lambda}_k = \frac{\alpha}{M^2}, \qquad \bar{c}_k = 1 - (\frac{\alpha}{M})^2. \tag{6.2.4}$$

Proof. From Theorem 4.3.4, we may take

$$c_k(\lambda) \equiv \lambda^2 M^2 + 1 - 2\alpha\lambda.$$

It is simple to verify (6.2.4) from Definition 6.2.1. □

Proposition 6.2.2. *Under the Condition A in Section 4.3, a suboptimal $\bar{\lambda}_k$ and a corresponding $\bar{c}_k$ of (2.10.1) from (6.2.1) are*

$$\bar{\lambda}_k = \frac{\alpha}{M^2}, \qquad \bar{c}_k = \frac{M^2}{\alpha^2 + M^2} \tag{6.2.5}$$

Proof. From Theorem 4.3.5, we may take

$$c_k(\lambda) \equiv \frac{1 + M^2\lambda^2}{(1 + \lambda\alpha)^2}.$$

Minimizing $c_k(\lambda)$ gives (6.2.5). □

Proposition 6.2.3. *Under the Condition B in Section 4.3, a suboptimal $\bar{\lambda}_k$ and a corresponding $\bar{c}_k$ of (2.10.1) from (6.2.1) are*

$$\bar{\lambda}_k = \frac{\alpha_1 + \alpha_2}{M^2 + \alpha_1\alpha_2}, \qquad \bar{c}_k = \frac{M^2 - \alpha_1^2}{M^2 + 2\alpha_1\alpha_2 + \alpha_2^2}. \tag{6.2.6}$$

Proof. From Theorem 4.3.6, we may take

$$c_k(\lambda) \equiv \frac{1 + M^2\lambda^2 - 2\lambda\alpha_1}{(1 + \lambda\alpha_2)^2}.$$

Minimizing $c_k(\lambda)$ gives (6.2.6). □

Proposition 6.2.4. *Under the Condition D of Section 4.3, a suboptimal $\bar{\lambda}_k$ and a corresponding $\bar{c}_k$ of (2.10.1) from (6.2.2) are*

$$\bar{\lambda}_k = \beta, \qquad \bar{c}_k = \beta^2 \|T_1(z^k) - T_1(z^*)\|^2. \tag{6.2.7}$$

Proof. From Theorem 4.3.7, we can choose

$$c_k(\lambda) = \lambda(2\beta - \lambda)\|T_1(z^k) - T_1(z^*)\|^2$$

Minimizing it gives (6.2.7). □

Theorem 6.2.5(Finite termination). *If $T = T_1 + T_2$ is strongly monotone with α, then, when we take $T_1(z) = \alpha z + a$ with $a \in \mathcal{H}$ and $\lambda_k = \alpha^{-1}$, the BFBS scheme (2.10.1) will terminate at one iteration for a zero point of T.*

Proof. At this situation, $(I - \lambda_k T_1)(z) \equiv a$ for all $z \in \mathcal{H}$ and hence $z^2 = z^1$. □

6.3. The Suboptimal Step Sizes for Extended FBS Methods

In the last section, we defined the suboptimal choice of $\{\lambda_k\}$ for BFBS methods and theoretically got some approximation for the optimal (or best) step sizes. Here we shall extend the concept to EFBS methods.

Definition 6.3.1. *Suppose $\{z^k\}$ is from (5.2.7) and $z^* \in zer(T)$ is its limit point. If there is a function $c_k : \mathbb{R} \to \mathbb{R}$ for every $k \geq 0$ such that*

$$\|\bar{P}_k^{\frac{1}{2}}(z^{k+1} - z^*)\|^2 \leq c_k(\lambda_k)\|\bar{P}_k^{\frac{1}{2}}(z^k - z^*)\|^2, \tag{6.3.1}$$

Or

$$\|\bar{P}_k^{\frac{1}{2}}(z^{k+1} - z^*)\|^2 \leq \|\bar{P}_k^{\frac{1}{2}}(z^k - z^*)\|^2 + c_k(\lambda_k), \tag{6.3.2}$$

then we call $\bar{\lambda}_k$ and $\bar{c}_k$ as suboptimal step size and value of (5.2.7) from (6.3.1) or (6.3.2) respectively for every $k \geq 0$. Where

$$\bar{\lambda}_k \in \operatorname{argmin} c_k(\lambda), \qquad \bar{c}_k := c_k(\bar{\lambda}_k). \tag{6.3.3}$$

Proposition 6.3.1. *Under the conditions of Theorem 5.4.2, the suboptimal step sizes $\bar{\lambda}_k$ and corresponding values $\bar{c}_k$ (linear convergence rate) of (5.2.2) are respectively as follows.*

(a) If β_k is the modulus of $S(T_1^k)^{-1}S$, then from (6.3.2)

$$\bar{\lambda}_k = \beta_k, \qquad \bar{c}_k = \beta_k^2\|T_1^k(x^k) - T_1^k(x^*)\|^2. \tag{6.3.4}$$

(b.1) If M_k and α_k are the Lipschitz constant and modulus of $S^{-1}T_1^kS^{-1}$, then from (6.3.1)

$$\bar{\lambda}_k = \frac{\alpha_k}{M_k^2}, \qquad \bar{c}_k = 1 - (\frac{\alpha_k}{M_k})^2, \tag{6.3.5}$$

(b.2) If M_k and α_k are the Lipschitz constant of $S^{-1}T_1^kS^{-1}$ and modulus of $S^{-1}T_2^kS^{-1}$, then from (6.3.1)

$$\bar{\lambda}_k = \frac{\alpha_k}{M_k^2}, \qquad \bar{c}_k = \frac{M_k^2}{\alpha_k^2 + M_k^2}, \tag{6.3.6}$$

(b.3) If M_k, α_k' and α_k'' are the Lipschitz constant of $S^{-1}T_1^kS^{-1}$, the moduli of $S^{-1}T_1^kS^{-1}$ and $S^{-1}T_2^kS^{-1}$, then from (6.3.1)

$$\bar{\lambda}_k = \frac{\alpha_k' + \alpha_k''}{M_k^2 + \alpha_k'\alpha_k''}, \qquad \bar{c}_k = \frac{M_k^2 - (\alpha_k')^2}{M_k^2 + 2\alpha_k'\alpha_k'' + (\alpha_k'')^2}. \tag{6.3.7}$$

Proof. These are respectively similar to the proofs of Proposition 6.2.4, Proposition 6.2.1, Proposition 6.2.2 and Proposition 6.2.3 in Section 6.2. □

Proposition 6.3.2. *Under the conditions of Theorem 5.4.3, the suboptimal step sizes $\bar{\lambda}_k$ and corresponding values $\bar{c}_k$ (linear convergence rate) of (5.2.7) from (6.3.1) are respectively as follows.*

(a) If M_k and α_k are the Lipschitz constant and modulus of $\bar{P}_k^{-\frac{1}{2}} T_1^k \bar{P}_k^{-\frac{1}{2}}$, then (6.3.5) is true.

(b) If M_k and α_k are the Lipschitz constant of $\bar{P}_k^{-\frac{1}{2}} T_1^k \bar{P}_k^{-\frac{1}{2}}$ and modulus of $\bar{P}_k^{-\frac{1}{2}} T_2^k \bar{P}_k^{-\frac{1}{2}}$, then (6.3.6) is true.

(c) If M_k, α'_k and α''_k are the Lipschitz constant of $\bar{P}_k^{-\frac{1}{2}} T_1^k \bar{P}_k^{-\frac{1}{2}}$, the moduli of $\bar{P}_k^{-\frac{1}{2}} T_1^k \bar{P}_k^{-\frac{1}{2}}$ and $\bar{P}_k^{-\frac{1}{2}} T_2^k \bar{P}_k^{-\frac{1}{2}}$, then (6.3.7) is true.

Proof. Consider

$$G_k^{-1} x^{k+1} \in (I + \lambda_k G_k \bar{T}_2^k G_k)^{-1} (I - \lambda_k G_k \bar{T}_1^k G_k)(G_k^{-1} x^k), \tag{6.3.8}$$

where $G_k = \bar{P}_k^{-\frac{1}{2}}$. Then from Proposition 6.3.1 we can easily deduce that (6.3.5), (6.3.6) and (6.3.7) are the suboptimal step sizes and values of (6.3.8) respectively. And hence Definition 6.2.1 and Definition 6.3.1 imply the conclusion. □

6.4. Estimating the Lipschitz Constants and Strong Monotone Moduli

As we have seen before, the Lipschitz constants and the strong monotone moduli are important for the convergence analysis of FBS methods or EFBS methods and choices of step sizes. In this section, we shall consider how to estimate these constants in practice.

Let $T = T_1 + T_2$. In this section, we consider the case of $\mathcal{H} = \mathbb{R}^n$. We shall estimate the Lipschitz constant and strongly monotone modulus of T_1 with $\mathrm{dom}(T_1) = \mathbb{R}^n$.

In general, the least Lipschitz constant M and the biggest monotone modulus α of T_1 or β of T_1^{-1} are respectively

$$M = \sup \frac{\|T_1(z') - T_1(z'')\|}{\|z' - z''\|} \qquad \forall \quad z', z'' \in \mathbb{R}^n, \quad z' \neq z'' \tag{6.4.1}$$

and

$$\alpha = \inf \frac{\langle T_1(z') - T_1(z''), z' - z'' \rangle}{\|z' - z''\|^2} \quad \forall \quad z', z'' \in \mathbb{R}^n, \quad z' \neq z'', \tag{6.4.2}$$

or for any $z', z'' \in \mathbb{R}^n$, $T_1(z') \neq T_1(z'')$,

$$\beta = \inf \frac{\langle T_1(z') - T_1(z''), z' - z'' \rangle}{\|T_1(z') - T_1(z'')\|^2}. \tag{6.4.3}$$

However, in our previous convergence proofs in Chapter 4, we actually only need the following property: finding positive scalars M and α or β such that

$$\|T_1(z^k) - T_1(z^*)\| \leq M\|z^k - z^*\|, \quad z^* \in \mathrm{zer}(T), \quad \forall \ k \geq 0 \tag{6.4.4}$$

and

$$\langle T_1(z^k) - T_1(z^*), z^k - z^* \rangle \geq \alpha\|z^k - z^*\|^2, \quad z^* \in \mathrm{zer}(T), \quad \forall \ k \geq 0, \tag{6.4.5}$$

or for a $z^* \in \mathrm{zer}(T)$,

$$\langle T_1(z^k) - T_1(z^*), z^k - z^* \rangle \geq \beta\|T_1(z^k) - T_1(z^*)\|^2, \quad \forall \quad k \geq 0. \tag{6.4.6}$$

Hence, if there is a set $\mathcal{S} \in \mathbb{R}^n$ with $\{z^k\} \subset \mathcal{S}$ and $\mathrm{zer}(T) \subset \mathcal{S}$ (such as $\mathcal{S} = \mathrm{cl}[\mathrm{dom}\, T_2]$), then we can use the following M and α to replace those in (6.4.1) and (6.4.2):

$$M = \sup_{z' \neq z''} \frac{\|T_1(z') - T_1(z'')\|}{\|z' - z''\|}, \quad \alpha = \inf_{z' \neq z''} \frac{\langle T_1(z') - T_1(z''), z' - z'' \rangle}{\|z' - z''\|^2}, \tag{6.4.7}$$

where z' and z'' are any points in $\mathcal{S}$.

Now let us consider some practical estimates of the M, α and β for the case of T_1 is affine. That is,

$$T_1(z) = \Phi z + \varphi \quad \forall z \in \mathbb{R}^n, \quad \text{where} \quad \Phi \in \mathbb{R}^{n \times n}, \quad \varphi \in \mathbb{R}^n \tag{6.4.8}$$

and Φ is nonzero and positive semidefinite. At this case, (6.4.1), (6.4.2) and (6.4.3) become

$$M = \sup_{\|z\|=1} \|\Phi z\| = \sup_{\|z\|=1} \sqrt{z^T \Phi^T \Phi z} = \|\Phi\|, \qquad z \in \mathbb{R}^n \tag{6.4.9}$$

and

$$\alpha = \inf_{\|z\|=1} \langle z, \Phi z \rangle \qquad \beta = \inf_{\|\Phi z\|=1} \langle \Phi z, z \rangle, \qquad z \in \mathbb{R}^n. \tag{6.4.10}$$

Recall that the norm $\|A\|$ of a matrix $A \in \mathbb{R}^{n\times m}$ is defined by

$$\|A\|^2 := \sup_{\|z\|=1} \|Az\|^2 = \sup_{\|z\|=1} z^T A^T A z = \|A^T A\| \qquad z \in \mathbb{R}^n. \tag{6.4.11}$$

Hence, (6.4.9) says that $M = \|\Phi\|$.

From the functional analysis theory, we know that $\|A^T\| = \|A\|$ and there are vectors $\bar{z} \in \mathbb{R}^n$ and $\bar{w} \in \mathbb{R}^m$ such that

$$\|A\| = \|A\bar{z}\| = \|A^T \bar{w}\| \qquad \text{with} \qquad \|\bar{z}\| = \|\bar{w}\| = 1. \tag{6.4.12}$$

Theorem 6.4.1. *Let I be the identity matrix and $\gamma(\Phi)$ be the biggest eigenvalue of $\Phi^T\Phi$. Then*

(a) (6.4.9) is equivalent to $M = \sqrt{\gamma(\Phi)} = \|\Phi\|$;

(b) T_1 in (6.4.8) is monotone or strongly monotone if and only if Φ is semi-positive definite or positive definite;

(c) T_1^{-1} defined by (6.4.8) is strongly monotone with modulus β if and only if $\|I - 2\beta\Phi\| \leq 1$;

(d) if Φ is invertible, then finding β is equivalent to finding the strong monotone modulus of Φ^{-1};

(e) if Φ is symmetric and semi-positive definite, then T_1^{-1} defined by (6.4.8) is strongly monotone with modulus β which satisfies

$$1 - 2\beta\|\Phi\| = -1, \qquad \text{i.e.} \quad \beta = \|\Phi\|^{-1} = \sqrt{\gamma(\Phi)}^{-1} \tag{6.4.13}$$

and this β is the best modulus we can get;

(f) if Φ is symmetric and positive definite, the α in (6.4.10) is the smallest eigenvalue of Φ.

(g) if Φ is positive definite, the moduli in (6.4.10) are $\alpha = 1/\|\bar{\Phi}^{-1}\|$ and $\beta = 1/\|\hat{\Phi}\|$. Where $\bar{\Phi} = (\Phi + \Phi^T)/2$ and $2\hat{\Phi}^{-1} = \Phi^{-1} + (\Phi^T)^{-1}$.

Proof. (6.4.9), (6.4.10) and Theorem 4.6.1 directly imply Part (a), Part (b) and Part (c) respectively. Part (d) comes from the following facts: $z \neq$

$0 \iff \Phi z \neq 0$ and

$$\frac{\langle \Phi z, z\rangle}{\|\Phi z\|^2} = \frac{\langle w, \Phi^{-1} w\rangle}{\|w\|^2}, \qquad w = \Phi z.$$

If Φ is symmetric, then there is a matrix A such that $\Phi = A^T A$. Thus,

$$\begin{aligned} \langle z, \Phi z\rangle &= \langle z, A^T A z\rangle = \|Az\|^2 \\ &\geq \|\Phi\|^{-1}\|A^T A z\|^2 = \|\Phi\|^{-1}\|\Phi z\|^2. \end{aligned}$$

Hence, $\beta = \|\Phi\|^{-1}$ is a strongly monotone modulus of T_1^{-1}. If there is another modulus of T_1^{-1} such that $\beta' > \beta$, then $\|I - 2\beta'\Phi\| \leq 1$ from Part (c). Since Φ is symmetric, the largest eigenvalue of Φ is $\|\Phi\|$ and the smallest eigenvalue of $I - 2\beta'\Phi$ is

$$1 - 2\beta'\|\Phi\| < 1 - 2\beta\|\Phi\| = -1.$$

The above inequality implies that $\|I - 2\beta'\Phi\| > 1$. This contradicts with Part (c). Therefore, the β in (6.4.13) is the biggest modulus of T_1^{-1}. Part (e) is hence proved.

Part (f) is elementary from (6.4.10) under given conditions.

As for Part (g), it is enough to prove that one of two moduli in Part (g) is true. From (6.4.10), since $\bar{\Phi}$ is symmetric and positive definite, we have

$$\begin{aligned} \alpha &= \inf_{\|z\|=1} z^T \Phi z \\ &= \inf_{\|z\|=1} \frac{z^T \Phi z + z^T \Phi^T z}{2} \\ &= \inf_{\|z\|=1} z^T \bar{\Phi} z \\ &= \inf_{z \neq 0} \frac{z^T \bar{\Phi} z}{z^T z} \\ &= \inf_{w \neq 0} \frac{w^T w}{w^T \bar{\Phi}^{-1} w} \\ &= \inf_{\|w\|=1} \frac{1}{w^T \bar{\Phi}^{-1} w} \\ &= \frac{1}{\|\bar{\Phi}\|}. \end{aligned}$$

Here $w = \sqrt{\bar{\Phi}} z$. $\square$

In general, the $\gamma(\Phi)$ or $\|\Phi\|$ is hard to estimate when n is large. However, the estimate is relatively easy for some special cases of Φ.

Example 6.4.1: $\Phi = \operatorname{diag}\{\varphi_1, \varphi_2, \cdots, \varphi_n\}$ with $\varphi_i \geq 0$. That is, Φ is diagonal and positive semidefinite. Then

$$M = \varphi^*, \qquad \beta = 1/\varphi^*, \qquad \text{where} \quad \varphi^* = \max\{\varphi_1, \varphi_2, \cdots, \varphi_n\} \neq 0. \tag{6.4.14}$$

If $\varphi_i > 0$ for $i = 1, 2, \cdots, n$, then $\alpha = \min\{\varphi_1, \cdots, \varphi_n\}$.

Example 6.4.2: Φ is block-diagonal, i.e., $\Phi = \operatorname{diag}\{\Phi_1, \cdots, \Phi_s\}$ with $s \leq n$. Then $M = \max\{\|\Phi_1\|, \cdots, \|\Phi_s\|\}$. When all Φ_i are symmetric, $\beta = 1/M$. Let α_i be the smallest eigenvalue of Φ_i for $i = 1, \cdots, s$ when all Φ_i are symmetric and positive definite. Then, $\alpha = \min\{\alpha_1, \cdots, \alpha_s\}$.

Example 6.4.3: Φ is the following block lower triangular matrix with $n = m + s$.

$$\Phi = \begin{pmatrix} I & \\ A & I \end{pmatrix}, \qquad A \in \mathbb{R}^{m\times s}. \tag{6.4.15}$$

Proposition 6.4.2. *Suppose Φ is defined by (6.4.15) and T_1 is given by (6.4.8). Then for a general matrix A,*

(a)

$$\|\Phi\| = \sqrt{1 + \frac{\|A\|^2}{2}\left(1 + \sqrt{1 + \frac{4}{\|A\|^2}}\right)}, \tag{6.4.16}$$

(b) T_1 is monotone if and only if $\|A\| \leq 2$. T_1 and T_1^{-1} are strongly monotone if and only if $\|A\| < 2$. At this case, (6.4.10) is equivalent to

$$\alpha = \beta = 1 - \frac{1}{2}\|A\|. \tag{6.4.17}$$

Proof. When $\|A\| = 0$, (6.4.16) and (6.4.17) are obviously true. So we suppose $\|A\| \neq 0$. Let $z^T = (u^T, v^T)$ and $u \in \mathbb{R}^s$, $v \in \mathbb{R}^m$. From (6.4.11) and (6.4.12),

$$\begin{aligned} \|\Phi\|^2 &= \max_{\|z\|=1} \|\Phi z\|^2 \\ &= \max_{\|z\|=1} 1 + 2v^T A u + \|Au\|^2 \\ &\leq \max_{\|z\|=1} 1 + 2\|A\|\|u\|\|v\| + \|A\|^2\|u\|^2. \end{aligned}$$

Let $\tau = \|u\|$. Then $\tau \leq 1$. On the other hand, (6.4.12) tells us that there is a $\bar{u}$ with $\|\bar{u}\| = 1$ and $\|A\bar{u}\| = \|A\|$. Thus if

$$u = \tau\bar{u} \qquad \text{and} \qquad v = \sqrt{1-\tau^2}\|A\|^{-1}A\bar{u},$$

we have

$$\|z\|^2 = \|u\|^2 + \|v\|^2 = \tau^2 + 1 - \tau^2 = 1$$

and

$$\begin{aligned} 1 + 2v^T Au + \|Au\|^2 &= 1 + 2\|A\|\|u\|\|v\| + \|A\|^2\|u\|^2 \\ &= 1 + 2\tau\sqrt{1-\tau^2}\|A\| + \tau^2\|A\|^2. \end{aligned}$$

Therefore,

$$\|\Phi\|^2 = 1 + \|A\|(\max_{0\leq\tau\leq 1} 2\tau\sqrt{1-\tau^2} + \|A\|\tau^2).$$

Solving the maximization problem gives solution τ^*

$$\tau^* = \frac{1}{\sqrt{2}}\sqrt{1 + \frac{\|A\|}{\sqrt{4+\|A\|^2}}}$$

and

$$\|\Phi\|^2 = 1 + \frac{\|A\|}{2}(\|A\| + \sqrt{4+\|A\|^2}).$$

Therefore, (6.4.16) is true.

For the proof of Part (b), consider the following:

$$\begin{aligned} \inf_{\|z\|=1} z^T\Phi z &= \inf_{\|z\|=1} 1 + v^T Au \\ &\geq \inf_{\|z\|=1} 1 - \|A\|\|u\|\|v\| \end{aligned}$$

Similarly, let

$$u = \tau\bar{u} \qquad \text{and} \qquad v = -\sqrt{1-\tau^2}\|A\|^{-1}A\bar{u},$$

where $\|\bar{u}\| = 1$ and $\|A\bar{u}\| = \|A\|$. Then $\|z\|^2 = \|u\|^2 + \|v\|^2 = 1$ and

$$1 + v^T Au = 1 - \|v\|\|A\|\|u\| = 1 - \tau\sqrt{1-\tau^2}\|A\|.$$

Therefore,

$$\inf_{\|z\|=1} z^T \Phi z = \min_{0 \le \tau \le 1} 1 - \tau\sqrt{1-\tau^2}\|A\|.$$

Solving the minimization problem gives solution $\tau^* = 1/\sqrt{2}$ and

$$\inf_{\|z\|=1} z^T \Phi z = 1 - \frac{1}{2}\|A\|. \tag{6.4.18}$$

Hence, T_1 is monotone if and only if $1 - \frac{1}{2}\|A\| \ge 0$. That is, $\|A\| \le 2$. And T_1 is strongly monotone if and only if $1 - \frac{1}{2}\|A\| > 0$, i.e. $\|A\| < 2$. The same conclusion for T_1^{-1} is also true because

$$\Phi^{-1} = \begin{pmatrix} I & \\ -A & I \end{pmatrix}.$$

At the case of $\|A\| < 2$, (6.4.10) and (6.4.18) give (6.4.17). □

Example 6.4.4: Φ is the following block lower triangular matrix with $n = m + s$.

$$\Phi = \begin{pmatrix} \Phi_1 & \\ A & \Phi_2 \end{pmatrix}, \qquad A \in \mathbb{R}^{m \times s}. \tag{6.4.19}$$

Proposition 6.4.3. *Suppose Φ is defined by (6.4.19) and T_1 is given by (6.4.8). Then for general matrices A, Φ_1 and Φ_2*

(a) *the best estimate of the Lipschitz constant of T_1 in (6.4.8) we can get is*

$$M = \sqrt{\frac{1}{2}\left[a^2 + b^2 + c^2 + \sqrt{4b^2c^2 + (a^2 + c^2 - b^2)^2}\right]}, \tag{6.4.20}$$

here $M \ge \|\Phi\|$ in general and $a = \|\Phi_1\|$, $b = \|\Phi_2\|$, $c = \|A\|$;

(b) *A sufficient condition for T_1 monotone and strongly monotone are respectively*

$$\|A\| \le 2\sqrt{\alpha_1\alpha_2} \qquad \text{and} \qquad \|A\| < 2\sqrt{\alpha_1\alpha_2}, \tag{6.4.21}$$

where α_1 and α_2 are largest constants such that $u^T\Phi_1 u \ge \alpha_1 u^T u$ and $v^T \Phi_2 v \ge \alpha_2 v^T v$ for any $u \in \mathbb{R}^s$ and $v \in \mathbb{R}^m$.

(c) *when the second inequality of (6.4.21) is true, the best estimate of the α in (6.4.10) we can get are*

$$\alpha' = \frac{\alpha_1 + \alpha_2}{2} - \frac{\sqrt{\|A\|^2 + (\alpha_1 - \alpha_2)^2}}{2}. \tag{6.4.22}$$

Proof. When $\|A\| = 0$, (6.4.20) and (6.4.22) are coincident with the results in Example 6.4.2. So we suppose $c = \|A\| \neq 0$. Let $a = \|\Phi_1\|$, $b = \|\Phi_2\|$, $z^T = (u^T, v^T)$ and $u \in \mathbb{R}^s$, $v \in \mathbb{R}^m$. From (6.4.11) and (6.4.12),

$$\begin{aligned}\|\Phi\|^2 &= \max_{\|z\|=1} \|\Phi z\|^2 \\ &= \max_{\|z\|=1} \|\Phi_1 u\|^2 + \|Au + \Phi_2 v\|^2 \\ &\leq M^2 = \max_{0\leq\tau\leq 1} a^2\tau^2 + c^2\tau^2 + b^2(1-\tau^2) + 2bc\tau\sqrt{1-\tau^2}.\end{aligned}$$

Where $\tau = \|u\| \leq 1$. Solving the maximization problem gives solution τ^*

$$\tau^* = \frac{1}{\sqrt{2}}\sqrt{1 + \frac{a^2+c^2-b^2}{\sqrt{\eta}}}, \qquad \eta = (a^2+c^2-b^2)^2 + 4b^2c^2$$

and

$$M^2 = \frac{a^2+b^2+c^2}{2} + \frac{\sqrt{\eta}}{2}.$$

Therefore, (6.4.20) is true.

This is the best estimate for general Φ_1, Φ_2 and A because the above M can equal $\|\Phi\|$ for some Φ_1, Φ_2 and A. One example is that Φ is the special case in Example 6.4.3.

For the proofs of Part (b) and Part (c), consider the following:

$$\begin{aligned}\min_{\|z\|=1} z^T\Phi z &= \min_{\|z\|=1} u^T\Phi_1 u + v^T\Phi_2 v + v^T Au \\ &\geq \min_{\|z\|=1} \alpha_1\|u\|^2 + \alpha_2\|v\|^2 - \|A\|\|u\|\|v\| \\ &= \min_{0\leq\tau\leq 1} \alpha_1\tau^2 + \alpha_2(1-\tau^2) - c\tau\sqrt{1-\tau^2}\end{aligned}$$

Solving the minimization problem gives solution τ^*

$$\tau^* = \sqrt{\frac{1}{2} - \frac{\alpha_1-\alpha_2}{2\sqrt{c^2+(\alpha_1-\alpha_2)^2}}}$$

and

$$\inf_{\|z\|=1} z^T\Phi z \geq \frac{\alpha_1+\alpha_2}{2} - \frac{1}{2}\sqrt{c^2+(\alpha_1-\alpha_2)^2}. \tag{6.4.23}$$

Hence, a sufficient condition for T_1 monotone (strongly monotone) is

$$\frac{\alpha_1+\alpha_2}{2} - \frac{1}{2}\sqrt{c^2+(\alpha_1-\alpha_2)^2} \geq (>)0.$$

That is,

$$(\alpha_1+\alpha_2)^2 - (\alpha_1-\alpha_2)^2 \geq (>)c^2 = \|A\|^2. \tag{6.4.24}$$

(6.4.24) is just the same as (6.4.21). Part (b) is proved. However, (6.4.23) implies that a strongly monotone modulus is given by (6.4.22) under the conditions of Part (c). (6.4.22) is the best estimate for the α in (6.4.10) with general matrices A, Φ_1 and Φ_2 because the α' in (6.4.22) equals α in (6.4.10) for some special Φ such as the Φ at Example 6.4.3. Hence, Part (c) is also proved. □

The similar conclusion for T_1^{-1} can also be deduced because Φ^{-1} has a same structure as Φ in (6.4.19). Here

$$\Phi^{-1} = \begin{pmatrix} \Phi_1^{-1} & \\ -\Phi_2^{-1}A\Phi_1^{-1} & \Phi_2^{-1} \end{pmatrix}. \tag{6.4.25}$$

Example 6.4.5: Φ is an anti-symmetric matrix. Namely, there is a nonzero matrix A such that

$$\Phi = \begin{pmatrix} & A^T \\ -A & \end{pmatrix}, \qquad A \in \mathbb{R}^{m\times s}. \tag{6.4.26}$$

Proposition 6.4.4. *Suppose Φ is given by (6.4.26) and T_1 is defined by (6.4.8). Then*

(a) The smallest Lipschitz constant $\|\Phi\|$ of T_1 is $\|\Phi\| = \|A\|$.

(b) T_1 and $-T_1$ are monotone with

$$\langle T_1(z') - T_1(z''), z'-z''\rangle \equiv 0 \quad \forall\ z', z'' \in \mathbb{R}^n. \tag{6.4.27}$$

And hence neither T_1 nor T_1^{-1} is strongly monotone.

Proof. Similar to the proof of Proposition 6.4.2, let $z^T = (u^T, v^T)$ with $u \in \mathbb{R}^s$ and $v \in \mathbb{R}^m$. From (6.4.11),

$$\begin{aligned} \|\Phi\|^2 &= \max_{\|z\|=1} \|\Phi z\|^2 \\ &= \max_{\|z\|=1} \|A^T v\|^2 + \|Au\|^2 \qquad (6.4.28) \\ &\leq \|A\|^2. \end{aligned}$$

(6.4.12) points out that there are $\bar{u}$ and $\bar{v}$ such that

$$\|A\bar{u}\| = \|A\|, \quad \|A^T\bar{v}\| = \|A^T\| = \|A\| \quad \text{with} \quad \|\bar{u}\| = 1, \quad \|\bar{v}\| = 1.$$

Let $u' = \tau\bar{u}$, $v' = \sqrt{1-\tau^2}\bar{v}$ and $0 \le \tau \le 1$. Then $\|z'\|^2 = \|u'\|^2 + \|v'\|^2 = 1$ and (6.4.11), (6.4.26) or (6.4.28) imply that

$$\|\Phi\|^2 \ge \|A^T v'\|^2 + \|Au'\|^2 = \|A\|^2. \tag{6.4.29}$$

Combining (6.4.29) and (6.4.28) gives $\|\Phi\| = \|A\|$.

Since for any $z' \in \mathbb{R}^n$ and $z'' \in \mathbb{R}^n$,

$$\langle T_1(z') - T_1(z''), z' - z''\rangle = (z' - z'')^T \Phi(z' - z''), \tag{6.4.30}$$

so (6.4.26) implies (6.4.27). Hence T_1 and $-T_1$ are monotone from (6.4.27) and (6.4.30), and so are T_1^{-1} and $-T_1^{-1}$. The last conclusion in Part (b) is also true from (6.4.27). $\square$

Example 6.4.6: Φ is the following symmetric matrix with $n = m + s$.

$$\Phi = \begin{pmatrix} \Phi_1 & A^T \\ A & \Phi_2 \end{pmatrix}, \qquad \begin{matrix} \Phi_1^T = \Phi_1 \\ \Phi_2^T = \Phi_2 \end{matrix}, \quad A \in \mathbb{R}^{m\times s}. \tag{6.4.31}$$

Proposition 6.4.5. *Suppose Φ is defined by (6.4.31) and T_1 is given by (6.4.8). Then for general matrices A, Φ_1 and Φ_2*

(a) *the best estimate of the Lipschitz constant $\|\Phi\|$ of T_1 we can get is*

$$M = \sqrt{c^2 + \frac{1}{2}\left[a^2 + b^2 + (a+b)\sqrt{4c^2 + (a-b)^2}\right]}, \tag{6.4.32}$$

here $M \ge \|\Phi\|$ in general and $a = \|\Phi_1\|$, $b = \|\Phi_2\|$, $c = \|A\|$. Hence an estimate for the β of (6.4.10) is $\beta' = 1/M$ with M given by (6.4.32).

(b) *A sufficient condition for T_1 monotone and strongly monotone are respectively*

$$\|A\| \le \sqrt{\alpha_1\alpha_2} \qquad \text{and} \qquad \|A\| < \sqrt{\alpha_1\alpha_2}, \tag{6.4.33}$$

where α_1 and α_2 are largest constants such that $u^T\Phi_1 u \ge \alpha_1 u^T u$ and $v^T\Phi_2 v \ge \alpha_2 v^T v$ for any $u \in \mathbb{R}^s$ and $v \in \mathbb{R}^m$ or $\alpha_i = 1/\|\Phi_i^{-1}\|$.

(c) when the second inequality of (6.4.33) is true, the best estimate of the α in (6.4.10) we can get are

$$\alpha' = \frac{\alpha_1 + \alpha_2}{2} - \frac{\sqrt{4\|A\|^2 + (\alpha_1 - \alpha_2)^2}}{2}. \tag{6.4.34}$$

Note: the α and β in (6.4.10) respectively equals $1/\|\Phi^{-1}\|$ and $1/\|\Phi\|$ from Theorem 6.4.1 if Φ is positive definite.

Proof. When $\|A\| = 0$, (6.4.32) and (6.4.34) are coincident with the results in Example 6.4.2. So we suppose $c = \|A\| \neq 0$. Let $a = \|\Phi_1\|$, $b = \|\Phi_2\|$, $z^T = (u^T, v^T)$ and $u \in \mathbb{R}^s$, $v \in \mathbb{R}^m$. From (6.4.11),

$$\begin{aligned}
\|\Phi\|^2 &= \max_{\|z\|=1} \|\Phi z\|^2 \\
&= \max_{\|z\|=1} \|\Phi_1 u + A^T v\|^2 + \|Au + \Phi_2 v\|^2 \\
&\leq \max_{0\leq\tau\leq 1} c^2 + b^2 + (a^2 - b^2)\tau^2 + 2(a+b)c\tau\sqrt{1-\tau^2} \\
&= c^2 + b^2 + (a+b) \max_{0\leq\tau\leq 1} (a-b)\tau^2 + 2c\tau\sqrt{1-\tau^2} \\
&= c^2 + b^2 + (a+b) \max_{0\leq\tau\leq 1} g(\tau) \equiv M^2.
\end{aligned}$$

Where $\tau = \|u\| \leq 1$. Solving the maximization problem gives solution τ^*

$$\tau^* = \frac{1}{\sqrt{2}}\sqrt{1 + \frac{a-b}{\sqrt{4c^2 + (a-b)^2}}}$$

and

$$g(\tau^*) = \frac{a-b}{2} + \frac{\sqrt{4c^2 + (a-b)^2}}{2}.$$

And hence,

$$M^2 = c^2 + \frac{a^2+b^2}{2} + \frac{a+b}{2}\sqrt{4c^2 + (a-b)^2}.$$

Therefore, (6.4.32) is true.

This is the best estimate for general Φ_1, Φ_2 and A because the above M can equal $\|\Phi\|$ for some Φ_1, Φ_2 and A. One example is that $M = \|\Phi\| = 1 + \|A\|$ when $\Phi_1 = I$ and $\Phi_2 = I$, where I is an appropriate dimension identity matrix.

For the proofs of Part (b) and Part (c), consider the following:

$$\begin{aligned}
\min_{\|z\|=1} z^T\Phi z &= \min_{\|z\|=1} u^T\Phi_1 u + v^T\Phi_2 v + 2v^T A u \\
&\geq \min_{\|z\|=1} \alpha_1\|u\|^2 + \alpha_2\|v\|^2 - 2\|A\|\|u\|\|v\| \\
&= \min_{0\leq\tau\leq 1} \alpha_1\tau^2 + \alpha_2(1-\tau^2) - 2c\tau\sqrt{1-\tau^2}
\end{aligned}$$

Solving the minimization problem gives solution τ^*

$$\tau^* = \sqrt{\frac{1}{2} - \frac{\alpha_1-\alpha_2}{2\sqrt{4c^2+(\alpha_1-\alpha_2)^2}}}$$

and

$$\inf_{\|z\|=1} z^T\Phi z \geq \frac{\alpha_1+\alpha_2}{2} - \frac{1}{2}\sqrt{4c^2+(\alpha_1-\alpha_2)^2}. \tag{6.4.35}$$

Hence, a sufficient condition for T_1 monotone (strongly monotone) is

$$\frac{\alpha_1+\alpha_2}{2} - \frac{1}{2}\sqrt{4c^2+(\alpha_1-\alpha_2)^2} \geq (>) 0.$$

That is,

$$(\alpha_1+\alpha_2)^2 - (\alpha_1-\alpha_2)^2 \geq (>) 4c^2 = 4\|A\|^2. \tag{6.4.36}$$

(6.4.36) is just the same as (6.4.33). Part (b) is proved. However, (6.4.35) implies that a strongly monotone modulus is given by (6.4.34) under the conditions of Part (c). (6.4.34) is the best estimate for the α in (6.4.10) with general matrices A, Φ_1 and Φ_2 because the α' in (6.4.34) equals α in (6.4.10) for some special Φ such as the special case with $\Phi_1 = I$ and $\Phi_2 = I$. Hence, Part (c) is also proved. □

6.5. Choices of the Splitting Strategies for $T = T_1 + T_2$

In the previous discussion, we always suppose that the T_1 and T_2 are known and monotone with $T = T_1 + T_2$. However, there are questions need to answer. For example, for a given T, how do we split it into $T_1 + T_2$? and what is the influence of different splitting strategies of $T = T_1 + T_2$? In other words, how do we choose the splitting strategy $T = T_1 + T_2$? Hence, let us discuss the splitting strategies of $T = T_1 + T_2$ at this section.

Let

$$T = T_0 + T_f + \mathcal{N}_W. \tag{6.5.1}$$

Where T_0 and T_f are monotone and single-valued on W, T_f is restricted to be in the forward step for whatever reasons and $\mathcal{N}_W$ is the normal cone operator with a nonempty, closed and convex set W. Since $\mathcal{N}_W$ is hard to implement by forward step and easy to implement by backward step, $\mathcal{N}_W$ is always a part of T_2. Hence, splitting T is equivalent to splitting T_0.

In this section, we assume that the operator T of (6.5.1) is strongly monotone over set W with the biggest modulus α.

Definition 6.5.1. *Suppose $P : \mathcal{H} \rightrightarrows \mathcal{H}$ is an operator. a constant η is called a monotone indicator of P on W if*

$$\eta = \inf_{z' \neq z''} \frac{\langle z' - z'', P(z') - P(z'') \rangle}{\|z' - z''\|^2} \qquad z', z'' \in W. \tag{6.5.2}$$

η is simply called M-indicator of P.

Hence, P is monotone (strongly monotone) on W if and only if its M-indicator η on W is non-negative (positive).

Now let us form a mathematical problem about the choice of $T = T_1 + T_2$. Our purpose here is to find suitable single-valued and monotone operators P_1 and P_2 such that $T_0 = P_1 + P_2$. Suppose that η_f, η_0, η_1, η_2 and η' are the M-indicators of T_f, T_0, T_1, P_2 and P_1 on W respectively. Let M_f, M and M_1 be the Lipschitz constants of T_f, T_1 and P_1 on W respectively. If $\eta_1 + \eta_2 > 0$ and $\{z^k\}$ is from the BFBS scheme (2.10.1) with $z^0 \in W$, then similar to Proposition 6.2.3,

$$\|z^{k+1} - z^*\|^2 \leq \frac{M^2 - \eta_1^2}{M^2 - \eta_1^2 + (\eta_1 + \eta_2)^2}\|z^k - z^*\|^2, \quad \text{if } \lambda_k \equiv \frac{\eta_1 + \eta_2}{M^2 + \eta_1\eta_2}, \tag{6.5.3}$$

and $z^* \in T(z^*)$. It is easy to see that the λ_k in (6.5.3) is suboptimally chosen for every k in the concept of Section 6.2. Our problem is to find monotone operators

P_1 and P_2 with $T_0 = P_1 + P_2$ such that the corresponding coefficient in the inequality of (6.5.3) is as small as possible. That is,

$$\min_{P_1,P_2} \frac{M^2-\eta_1^2}{M^2-\eta_1^2+(\eta_1+\eta_2)^2} \quad \text{with} \quad T_1 = T_f + P_1, \quad T_2 = P_2 + \mathcal{N}_W. \tag{6.5.4}$$

In general, it is hard to solve the problem (6.5.4) for general operators P_1 and P_2. However, some special cases of (6.5.4) can be solved easily.

Theorem 6.5.1. *If $P_1 = \tau I$ and $\eta_0 > 0$, a solution of the following special case of the problem (6.5.4)*

$$\min_{0\le\tau\le\eta_0} \frac{M^2-\eta_1^2}{M^2-\eta_1^2+(\eta_1+\eta_2)^2} \quad \textit{with} \quad T_1 = T_f + \tau I, \quad P_2 = T_0 - \tau I \tag{6.5.5}$$

is $\tau = 0$ or $T_1 = T_f$.

Proof. At this case,

$$\begin{aligned} M^2 &= \max_{z'\neq z''} \frac{\|T_1(z')-T_1(z'')\|^2}{\|z'-z''\|^2} = \\ &\max_{z'\neq z''} \frac{\|T_f(z')-T_f(z'')\|^2 + \tau^2\|z'-z''\|^2 + 2\tau\langle z'-z'', T_f(z')-T_f(z'')\rangle}{\|z'-z''\|^2} \\ &\ge M_f^2 + \tau^2 + 2\tau\eta_f \end{aligned} \tag{6.5.6}$$

and

$$\eta_1 = \inf_{z'\neq z''} \frac{\langle z'-z'', (T_f+\tau I)(z') - (T_f+\tau I)(z'')\rangle}{\|z'-z''\|^2} = \eta_f + \tau. \tag{6.5.7}$$

Hence,

$$M^2 - \eta_1^2 \ge M_f^2 + \tau^2 + 2\tau\eta_f - (\eta_f+\tau)^2 = M_f^2 - \eta_f^2. \tag{6.5.8}$$

Similar to (6.5.7), we have $\eta_2 = \eta_0 - \tau$ for $0 \le \tau \le \eta_0$. Thus,

$$\eta_1 + \eta_2 \equiv \eta_f + \eta_0 > 0 \qquad \text{for} \qquad 0 \le \tau \le \eta_0. \tag{6.5.9}$$

Consider the following one dimensional function

$$f(\theta) = \frac{\theta}{\theta + (\eta_1+\eta_2)^2}, \qquad \theta \in \mathbb{R}. \tag{6.5.10}$$

Then

$$f'(\theta) = \frac{(\eta_1+\eta_2)^2}{[(\eta_1+\eta_2)^2+\theta]^2} > 0 \qquad \text{for} \qquad \theta \ge 0. \tag{6.5.11}$$

That is, $f(\theta)$ of (6.5.10) is strictly increasing for $\theta \ge 0$. Therefore, $M_f \ge \eta_f$ and (6.5.8), (6.5.11) imply that $\tau = 0$ is a solution of (6.5.5). □

Corollary 6.5.2. *Under the conditions of Theorem 6.5.1, if $T_f(z)$ equals an affine operator expression $\Phi z + \varphi$ and Φ is an anti-symmetric linear operator, then all τ with $0 \le \tau \le \eta_0$ are solutions of (6.5.5).*

Proof. At this situation, $\eta_f = 0$ and (6.5.6), (6.5.7) become respectively

$$M^2 = M_f^2 + \tau^2, \qquad \eta_1 = \tau.$$

And hence

$$M^2 - \eta_1^2 \equiv M_f^2 \qquad \text{for all} \qquad \tau \in [0, \eta_0]. \tag{6.5.12}$$

Therefore, this corollary is true. $\square$

Corollary 6.5.2 tells us that all of the splitting strategies $T_1 = T_f + \tau I$ for $0 \le \tau \le \eta_0$ have same suboptimal convergence behavior when T_f is an affine operator with an anti-symmetric linear operator. Hence, the choice of τ is not important at all.

Corollary 6.5.3. *In the (5.2.7), when $H_k \equiv H = S^2 \ne I$, if $P_1 = \tau H$ and $T_f(z) = \Phi z + \varphi$ with the linear operater $\Phi^* = -\Phi$, then all τ with $0 \le \tau \le \eta_0$ are solutions of (6.5.5) with T_i being replaced by $S^{-1}T_iS^{-1}$ for $i = 0, 1, 2$, where H is a positive definite matrix.*

Proof. At this situation, if we replace z^* by $w^* = Sz^*$, z^k by $w^k = Sz^k$, T_i by $S^{-1}T_iS^{-1}$ for $i = 0, 1, 2$, and T_f by $S^{-1}T_fS^{-1}$ in (6.5.3), then, we shall get the same problem as (6.5.4) and (6.5.5) under the replacements, similar to the arguments in Section 5.4 and Section 6.3. Now, $P_1 = \tau H$ is exactly $S^{-1}P_1S^{-1} = \tau I$. Since $S^{-1}\Phi S^{-1}$ is still anti-symmetric, Corollary 6.5.2 implies the conclusion. $\square$

Lemma 6.5.4. *Suppose P and Q are two monotone and Lipschitz continuous operators on W with*

$$P = \begin{pmatrix} A & \\ & B \end{pmatrix}, \qquad Q = \begin{pmatrix} & C^T \\ -C & \end{pmatrix}. \tag{6.5.13}$$

Let η_P and η_C be the bigger M-indicator of A^TA, B^TB and the smaller M-indicator of C^TC, CC^T respectively. Then

$$\|P + Q\|^2 \ge \max\left\{\|Q\|^2 + \eta_P, \|P\|^2 + \eta_C\right\}. \tag{6.5.14}$$

Proof. Let $z^T = (u^T, v^T)$. Then

$$\begin{aligned}\|P+Q\|^2 &= \max_{\|z\|=1} \|Au + C^Tv\|^2 + \|Bv - Cu\|^2 \\ &= \max_{\|z\|=1} \|Au\|^2 + \|Bv\|^2 + \|Cu\|^2 + \|C^Tv\|^2 + 2v^T(CA - B^TC)u.\end{aligned} \tag{6.5.15}$$

Now, there exist $\bar{u}$, $\bar{v}$ and $\hat{u}$ such that $\|\bar{u}\| = \|\bar{v}\| = \|\hat{u}\| = 1$ and

$$A^TA\bar{u} = \|A\|^2\bar{u}, \qquad B^TB\bar{v} = \|B\|^2\bar{v}, \qquad C^TC\hat{u} = \|C\|^2\hat{u}. \tag{6.5.16}$$

Let $u = \tau\bar{u}$ and $v = \sqrt{1-\tau^2}\bar{v}$, where $0 \le \tau \le 1$. Then

$$\|P+Q\|^2 \ge \max_{0\le\tau\le1} f(\tau),$$

where

$$\begin{aligned}f(\tau) =&(\|A\|^2 + \|C\bar{u}\|^2)\tau^2 + (\|B\|^2 + \|C^T\bar{v}\|^2)(1-\tau^2) \\ &+ 2(\bar{v}^T(CA - B^TC)\bar{u}\tau\sqrt{1-\tau^2}.\end{aligned}$$

Hence

$$\begin{aligned}\|P+Q\|^2 &\ge \max\left\{f(0), f(1)\right\} \\ &= \max\left\{\|A\|^2 + \|C\bar{u}\|^2, \|B\|^2 + \|C^T\bar{v}\|^2\right\} \\ &\ge \max\left\{\|A\|^2, \|B\|^2\right\} + \min\left\{\|C\bar{u}\|^2, \|C^T\bar{v}\|^2\right\} \\ &\ge \|P\|^2 + \eta_C.\end{aligned}$$

This proves half of (6.5.14).

Now from (6.5.16), let $u = \tau\hat{u}$ and $v = \sqrt{1-\tau^2}\hat{v}$, where $\hat{v} = \|C\|^{-1}C\hat{u}$. Then $\|\hat{v}\| = 1$ and $C^T\hat{v} = \|C\|\hat{u}$. And hence (6.5.15) implies

$$\|P+Q\|^2 \ge \max_{0\le\tau\le1} g(\tau),$$

here

$$g(\tau) = \|C\|^2 + \tau^2\|A\hat{u}\|^2 + (1-\tau^2)\|B\hat{v}\|^2 + 2\tau\sqrt{1-\tau^2}\|C\|(\hat{u}A\hat{u} - \hat{v}B\hat{v}).$$

Thus, from $\|Q\| = \|C\|$,

$$\begin{aligned}\|P+Q\|^2 &\ge \max\left\{g(0), g(1)\right\} \\ &= \max\left\{\|A\hat{u}\|^2 + \|C\|^2, \|B\hat{v}\|^2 + \|C\|^2\right\} \\ &\ge \max\left\{\|A\hat{u}\|^2, \|B\hat{v}\|^2\right\} + \|Q\|^2 \\ &\ge \|Q\|^2 + \eta_P.\end{aligned}$$

This is another half of (6.5.14). ◻

Theorem 6.5.5. *Suppose $\eta_0 > 0$. If $T_f(z) = Qz + q_f$ with Q defined in (6.5.13) and $P_1 = P$ of (6.5.13), then $\eta_f = 0$, $\eta_1 = \eta'$ and $T_1 = T_f$ is a global solution of the following special case of the problem (6.5.4)*

$$\min_{A,B} \frac{M^2 - \eta_1^2}{M^2 - \eta_1^2 + (\eta_1 + \eta_2)^2} \quad \text{with} \quad T_1 = T_f + P, \quad P_2 = T_0 - P. \tag{6.5.17}$$

Proof. Lemma 6.5.4 gives that

$$M^2 \geq M_f^2 + \eta_P \geq M_f^2 + \eta'^2 \tag{6.5.18}$$

and at this case, $\eta_f = 0$ obviously and

$$\eta_1 = \inf_{z' \neq z''} \frac{\langle z' - z'', (T_f + P)(z') - (T_f + P)(z'')\rangle}{\|z' - z''\|^2} = \eta'. \tag{6.5.19}$$

Hence

$$M^2 - \eta_1^2 \geq M_f^2 + \eta_1^2 - \eta_1^2 = M_f^2.$$

From $\eta_0 \geq \eta_1 + \eta_2$ and $\eta_f = 0$, we have

$$\frac{M^2 - \eta_1^2}{M^2 - \eta_1^2 + (\eta_1 + \eta_2)^2} \geq \frac{M_f^2}{M_f^2 + (\eta_1 + \eta_2)^2} \geq \frac{M_f^2}{M_f^2 + \eta_0^2}. \tag{6.5.20}$$

(6.5.20) implies that $T_1 = T_f$ or P_1 being zero operator is a solution of (6.5.17).
□

6.6. The Choice of the Metric H in the EFBS Methods

Definition 6.6.1. *Suppose T is a bounded linear operator on $\mathcal{H}$. We call T as symmetric if $T^* = T$, call T as positive definite if there exists a $\alpha > 0$ such that*

$$\langle x, Tx\rangle \geq \alpha\langle x, x\rangle, \forall x \in \mathcal{H} \tag{6.6.1}$$

and call T as positive semidefinite if $\langle x, Tx\rangle \geq 0$ for any $x \in \mathcal{H}$.

Proposition 6.6.1. *If T is a bounded linear operator on $\mathcal{H}$, then T is strongly monotone if and only if T is positive definite and T is monotone if and only if T is positive semidefinite.*

Proof. These are obvious from the linearity of T. □

Definition 6.6.2. *Suppose H is a symmetric positive-definite bounded linear operator. Then*

(a) *define* $\| * \|_H$ as

$$\|z\|_H = \sqrt{\langle z, Hz\rangle} = \|H^{\frac{1}{2}} z\|. \tag{6.6.2}$$

(b) *an operator T will be called (H, H^{-1})-Lipschitz continuous with constant κ if*

$$\|T(z') - T(z)\|_{H^{-1}} \leq \kappa \|z' - z\|_H \quad \text{for all} \quad z,\ z'. \tag{6.6.3}$$

(c) *an operator T will be called H-strongly monotone with modulus β if*

$$\langle w' - w, z' - z\rangle \geq \beta \|z' - z\|_H \qquad \textit{for all} \qquad w \in T(z), \quad w' \in T(z'). \tag{6.6.4}$$

Proposition 6.6.2. *Suppose H is a symmetric positive-definite bounded linear operator. Then*

(a) *an operator T is H-strongly monotone with modulus β, if and only if, $H^{-\frac{1}{2}} T H^{-\frac{1}{2}}$ is strongly monotone with modulus β.*

(b) *an operator T is (H, H^{-1})-Lipschitz continuous with constant κ, if and only if, $H^{-\frac{1}{2}} T H^{-\frac{1}{2}}$ is Lipschitz continuous with constant κ.*

Proof. Let $\bar{T} = H^{-\frac{1}{2}} T H^{-\frac{1}{2}}$. Then from (6.6.2) and

$$\begin{aligned}\langle w' - w, z' - z\rangle &= \langle w' - w, H^{-\frac{1}{2}} H^{\frac{1}{2}}(z' - z)\rangle \\ &= \langle H^{-\frac{1}{2}}(w' - w), H^{\frac{1}{2}}(z' - z)\rangle \\ &= \langle H^{-\frac{1}{2}} w' - H^{-\frac{1}{2}} w, H^{\frac{1}{2}} z' - H^{\frac{1}{2}} z\rangle,\end{aligned}$$

(6.6.4) is equivalent to

$$\langle H^{-\frac{1}{2}} w' - H^{-\frac{1}{2}} w, H^{\frac{1}{2}} z' - H^{\frac{1}{2}} z\rangle \geq \beta \|H^{\frac{1}{2}} z' - H^{\frac{1}{2}} z\|$$

for all $H^{-\frac{1}{2}} w' \in \bar{T}(H^{\frac{1}{2}} z')$ and $H^{-\frac{1}{2}} w \in \bar{T}(H^{\frac{1}{2}} z)$. Hence, Part (a) is true.

Again from (6.6.2) and the linearity of H, (6.6.3) is the same as

$$\|\bar{T}(H^{\frac{1}{2}} z') - \bar{T}(H^{\frac{1}{2}} z)\| \leq \kappa \|H^{\frac{1}{2}} z' - H^{\frac{1}{2}} z\| \qquad \text{for all} \qquad z,\ z'.$$

This proves Part (b). $\square$

Now suppose that $T = T_1 + T_2$ is given and $\mathcal{G}$ be the set of all symmetric positive-definite bounded linear operator over $\mathcal{H}$. Similar to Section 6.5, let η'_H, η''_H be the M-indicators of $H^{-\frac{1}{2}}T_1H^{-\frac{1}{2}}$, $H^{-\frac{1}{2}}T_2H^{-\frac{1}{2}}$ and κ_H be the (H, H^{-1})-Lipschitz constant of T_1. If $\eta'_H + \eta''_H > 0$ and $\{z^k\}$ is from the EFBS scheme (5.2.2) with $z^0 \in \mathrm{dom}(T_1)$ and $H_k \equiv H$, then at the iteration k, Proposition 6.6.2, Theorem 5.4.2 and Proposition 6.3.1 imply that

$$\|z^{k+1} - z^*\|_H^2 \le \frac{\kappa_H^2 - {\eta'_H}^2}{\kappa_H^2 - {\eta'_H}^2 + (\eta'_H + \eta''_H)^2}\|z^k - z^*\|_H^2 = c_H\|z^k - z^*\|_H^2 \qquad (6.6.5)$$

with

$$\lambda_k \equiv \frac{\eta'_H + \eta''_H}{\kappa_H^2 + \eta'_H\eta''_H}, \qquad (6.6.6)$$

where $z^* \in T(z^*)$, $H \in \mathcal{S}$ and $\mathcal{S}$ is a given subset of $\mathcal{G}$. Hence, our problem in this section is to find a $H \in \mathcal{S}$ such that the c_H in (6.6.5) as small as possible. That is,

$$\underset{H\in\mathcal{S}}{\text{minimize}} \quad c_H := \frac{\kappa_H^2 - {\eta'_H}^2}{\kappa_H^2 - {\eta'_H}^2 + (\eta'_H + \eta''_H)^2}. \qquad (6.6.7)$$

However, the above problem is hard to solve for general given set $\mathcal{S}$ and operators T_1, T_2. Thus, we consider two special cases of (6.6.7).

Let the Hilbert space $\mathcal{H} = \mathcal{H}_1 \times \mathcal{H}_2$ and define

$$\mathcal{S} = \{H \,|\, H = \begin{pmatrix} H_1 & \\ & H_2 \end{pmatrix} \in \mathcal{G}\}. \qquad (6.6.8)$$

Now consider

$$T_1 = \begin{pmatrix} & -R \\ R^* & \end{pmatrix}, \qquad T_2 = \begin{pmatrix} P & -D \\ D^* & Q \end{pmatrix} + \mathcal{N}_W, \qquad (6.6.9)$$

where P and H_1 are positive-definite bounded linear operator over $\mathcal{H}_1$, Q and H_2 are positive-definite bounded linear operator over $\mathcal{H}_2$, R and D are bounded linear operators from $\mathcal{H}_2$ to $\mathcal{H}_1$, W represents a subset of $\mathcal{H}$.

Proposition 6.6.3. *For the given $\mathcal{S}$ in (6.6.8) and T_1, T_2 in (6.6.9), we have*

$$\eta'_H = 0, \qquad \kappa_H = \|H_1^{-\frac{1}{2}} R H_2^{-\frac{1}{2}}\|, \qquad \eta''_H = \min\{\eta_P, \eta_Q\} \tag{6.6.10}$$

and if $\sigma_H = \beta_H/\kappa_H$ and $\beta_H = \eta''_H$, then Problem (6.6.7) is equivalent to the following optimization problem:

$$\textit{maximize} \quad \sigma_H \quad \textit{subject to} \quad H \in \mathcal{S}. \tag{6.6.11}$$

Where η_P and η_Q in (6.6.10) are the M-indicators of $H_1^{-\frac{1}{2}} P H_1^{-\frac{1}{2}}$ and $H_2^{-\frac{1}{2}} Q H_2^{-\frac{1}{2}}$ respectively.

Proof. (6.6.10) is obviously true from Proposition 6.6.2, Definition 6.5.1 and Proposition 6.4.4. Hence, c_H of (6.6.7) becomes at this case

$$c_H = \frac{\kappa_H^2}{\kappa_H^2 + \eta''^2_H} = \frac{1}{1 + \sigma_H^2}.$$

This implies that solving (6.6.7) is equivalent to solving (6.6.11). ◻

Theorem 6.6.4. *Under the conditions of Proposition 6.6.3, One global solution of Problem (6.6.11) [and hence a global solution of (6.6.7)] is*

$$\hat{H} = \begin{pmatrix} P & \\ & Q \end{pmatrix}. \tag{6.6.12}$$

Proof. From (6.6.10) and Definition 6.5.1, we know that

$$\kappa_{\hat{H}} = \|P^{-\frac{1}{2}} R Q^{-\frac{1}{2}}\|, \qquad \beta_{\hat{H}} = 1.$$

Hence

$$\sigma_{\hat{H}} = \|P^{-\frac{1}{2}} R Q^{-\frac{1}{2}}\|^{-1}.$$

For general $H \in \mathcal{S}$, (6.6.10) and Theorem 6.4.1 gives that

$$\kappa_H = \|H^{-\frac{1}{2}} T_1 H^{-\frac{1}{2}}\| = \|H_1^{-\frac{1}{2}} R H_2^{-\frac{1}{2}}\|,$$

$$\begin{aligned}\beta_H &= \min\{\eta_P, \eta_Q\}\\ &= \min\{\|H_1^{\frac{1}{2}} P^{-1} H_1^{\frac{1}{2}}\|^{-1}, \|H_2^{\frac{1}{2}} Q^{-1} H_2^{\frac{1}{2}}\|^{-1}\}\\ &= \min\{\|P^{-\frac{1}{2}} H_1^{\frac{1}{2}}\|^{-2}, \|H_2^{\frac{1}{2}} Q^{-\frac{1}{2}}\|^{-2}\}\end{aligned}$$

However,

$$\begin{aligned}\sigma_{\hat{H}} &\geq (\|P^{-\frac{1}{2}} H_1^{\frac{1}{2}}\|\kappa_H\|H_2^{\frac{1}{2}} Q^{-\frac{1}{2}}\|)^{-1}\\ &\geq \kappa_H^{-1}\beta_H = \sigma_H.\end{aligned}$$

That is, $\hat{H}$ is a global solution of Problem (6.6.11). □

Now let us consider another case with $\mathcal{S}$ given by (6.6.8) and

$$T_1 = \begin{pmatrix} P & -R \\ R^* & Q \end{pmatrix}, \qquad T_2 = \begin{pmatrix} & -D \\ D^* & \end{pmatrix} + \mathcal{N}_W, \tag{6.6.13}$$

where P, Q ,R, D and W are the same as those in (6.6.9).

Proposition 6.6.5. *For the given $\mathcal{S}$ in (6.6.8) and T_1, T_2 in (6.6.13), we have*

$$\kappa_H^2 = \|H^{-\frac{1}{2}} T_1 H^{-\frac{1}{2}}\|^2 \geq \|H_1^{-\frac{1}{2}} R H_2^{-\frac{1}{2}}\|^2 + \max\{\eta_P^2, \eta_Q^2\}, \tag{6.6.14}$$

$$\eta''_H = 0, \qquad \eta'_H = \min\{\eta_P, \eta_Q\} \tag{6.6.15}$$

and if $\sigma_H = \beta_H/\kappa_H$ and $\beta_H = \eta'_H$ here, then Problem (6.6.7) is equivalent to the optimization problem (6.6.11). Where η_P and η_Q in (6.6.14) and (6.6.15) are the M-indicators of $H_1^{-\frac{1}{2}} P H_1^{-\frac{1}{2}}$ and $H_2^{-\frac{1}{2}} Q H_2^{-\frac{1}{2}}$ respectively.

Proof. (6.6.15) is obviously true from Proposition 6.6.2 and Definition 6.5.1. On the other hand, (6.6.14) is just a direct corollary of Lemma 6.5.4. Hence, c_H of (6.6.7) becomes at this case

$$c_H = \frac{\kappa_H^2 - \beta_H^2}{\kappa_H^2} = 1 - \sigma_H^2.$$

This implies that solving (6.6.7) is equivalent to solving (6.6.11). □

Lemma 6.6.6. *Suppose T is a monotone and Lipschitz continuous operator on $\mathcal{H}$ with*

$$T = \begin{pmatrix} I & C^T \\ -C & I \end{pmatrix}. \tag{6.6.16}$$

Then

$$\|T\|^2 = 1 + \|C\|^2. \tag{6.6.17}$$

Proof. Let $z^T = (u^T, v^T)$. Then

$$\begin{aligned} \|T\|^2 &= \max_{\|z\|=1} \|u + C^T v\|^2 + \|v - Cu\|^2 \\ &= \max_{\|z\|=1} \|u\|^2 + \|v\|^2 + \|Cu\|^2 + \|C^T v\|^2 \\ &= 1 + \max_{\|z\|=1} \|Cu\|^2 + \|C^T v\|^2. \end{aligned} \tag{6.6.18}$$

Now, there exist $\bar{u}$ and $\bar{v}$ such that $\|\bar{u}\| = \|\bar{v}\| = 1$ and

$$C^T C\bar{u} = \|C\|^2 \bar{u}, \qquad CC^T \bar{v} = \|C\|^2 \bar{v}.$$

Let $u = \tau\bar{u}$ and $v = \sqrt{1-\tau^2}\,\bar{v}$, where $0 \le \tau \le 1$. Then from (6.6.18),

$$\|T\|^2 \ge 1 + \|C\|^2.$$

However, (6.6.18) also implies that

$$\begin{aligned} \|T\|^2 &\le 1 + \|C\|^2 \left(\max_{\|z\|=1} \|u\|^2 + \|v\|^2 \right) \\ &= 1 + \|C\|^2. \end{aligned}$$

Therefore, (6.6.17) is true. $\square$

Theorem 6.6.7. *Under the conditions of Proposition 6.6.5, One global solution of Problem (6.6.11) [and hence a global solution of (6.6.7)] is*

$$\hat{H} = \begin{pmatrix} P & \\ & Q \end{pmatrix}. \tag{6.6.19}$$

Proof. From (6.6.13) and (6.6.19), we know that

$$\hat{H}^{-\frac{1}{2}} T_1 \hat{H}^{-\frac{1}{2}} = \begin{pmatrix} I & -P^{-\frac{1}{2}} R Q^{-\frac{1}{2}} \\ Q^{-\frac{1}{2}} R^* P^{-\frac{1}{2}} & I \end{pmatrix}. \tag{6.6.20}$$

Thus, (6.6.14), (6.6.20), Lemma 6.6.6 and Definition 6.5.1 imply that

$$\kappa_{\hat{H}}^2 = 1 + \|P^{-\frac{1}{2}} R Q^{-\frac{1}{2}}\|^2, \qquad \beta_{\hat{H}} = 1.$$

Hence

$$\sigma_{\hat{H}} = \frac{1}{\sqrt{1 + \|P^{-\frac{1}{2}} R Q^{-\frac{1}{2}}\|^2}}.$$

For general $H \in \mathcal{S}$, we have

$$H^{-\frac{1}{2}} T_1 H^{-\frac{1}{2}} = \begin{pmatrix} H_1^{-\frac{1}{2}} P H_1^{-\frac{1}{2}} & -H_1^{-\frac{1}{2}} R H_2^{-\frac{1}{2}} \\ H_2^{-\frac{1}{2}} R^* H_1^{-\frac{1}{2}} & H_2^{-\frac{1}{2}} Q H_2^{-\frac{1}{2}} \end{pmatrix}$$

and (6.6.14), (6.6.15) are true. However, Theorem 6.4.1 implies that

$$\begin{aligned} \beta_H &= \min\{\eta_P, \eta_Q\} \\ &= \min\{\|H_1^{\frac{1}{2}} P^{-1} H_1^{\frac{1}{2}}\|^{-1}, \|H_2^{\frac{1}{2}} Q^{-1} H_2^{\frac{1}{2}}\|^{-1}\} \\ &= \min\{\|P^{-\frac{1}{2}} H_1^{\frac{1}{2}}\|^{-2}, \|H_2^{\frac{1}{2}} Q^{-\frac{1}{2}}\|^{-2}\} \end{aligned} \tag{6.6.21}$$

Hence,

$$\sigma_H \le \frac{\beta_H}{\sqrt{\|H_1^{-\frac{1}{2}} R H_2^{-\frac{1}{2}}\|^2 + \max\{\eta_P^2, \eta_Q^2\}}}. \tag{6.6.22}$$

(6.6.15) and (6.6.21) imply that

$$\frac{\max\{\eta_P^2, \eta_Q^2\}}{\beta_H^2} \ge 1$$

and

$$\begin{aligned} &\beta_H^{-1} \|H_1^{-\frac{1}{2}} R H_2^{-\frac{1}{2}}\| \\ &\ge \|P^{-\frac{1}{2}} H_1^{\frac{1}{2}}\| \|H_1^{-\frac{1}{2}} R H_2^{-\frac{1}{2}}\| \|H_2^{\frac{1}{2}} Q^{-\frac{1}{2}}\| \\ &\ge \|P^{-\frac{1}{2}} R Q^{-\frac{1}{2}}\|. \end{aligned}$$

Therefore, the above inequalities and (6.6.22) give that

$$\sigma_H \le \frac{1}{\sqrt{1 + \|P^{-\frac{1}{2}} R Q^{-\frac{1}{2}}\|^2}} = \sigma_{\hat{H}}.$$

That is, $\hat{H}$ is a global solution of Problem (6.6.7). □

For those $\{z^k\}$ from (5.2.2) with $T^k = T_1^k + T_2^k$, the choice of H_k will be the same as the above at each fixed iteration k. Hence, the process is the same except one needs to solve the problem (6.6.7) for every H_k.