A block symmetric Gauss-Seidel decomposition theorem for convex composite quadratic programming and its applications

Xudong Li*, Defeng Sun†, Kim-Chuan Toh‡

March 20, 2017

Abstract

For a symmetric positive semidefinite linear system of equations $\mathcal{Q}\boldsymbol{x} = \boldsymbol{b}$, where $\boldsymbol{x} = (x_1, \ldots, x_s)$ is partitioned into s blocks, with $s \geq 2$, we show that each cycle of the classical block symmetric Gauss-Seidel (block sGS) method exactly solves the associated quadratic programming (QP) problem but added with an extra proximal term of the form $\frac{1}{2}\|\boldsymbol{x}-\boldsymbol{x}^k\|_{\mathcal{T}}^2$, where $\mathcal{T}$ is a symmetric positive semidefinite matrix related to the sGS decomposition and $\boldsymbol{x}^k$ is the previous iterate. By leveraging on such a connection to optimization, we are able to extend the result (which we name as the block sGS decomposition theorem) for solving a convex composite QP (CCQP) with an additional possibly nonsmooth term in x_1, i.e., $\min\{p(x_1) + \frac{1}{2}\langle \boldsymbol{x}, \mathcal{Q}\boldsymbol{x}\rangle - \langle \boldsymbol{b}, \boldsymbol{x}\rangle\}$, where $p(\cdot)$ is a proper closed convex function. Based on the block sGS decomposition theorem, we are able to extend the classical block sGS method to solve a CCQP. In addition, our extended block sGS method has the flexibility of allowing for inexact computation in each step of the block sGS cycle. At the same time, we can also accelerate the inexact block sGS method to achieve an iteration complexity of $O(1/k^2)$ after performing k block sGS cycles. As a fundamental building block, the block sGS decomposition theorem has played a key role in various recently developed algorithms such as the inexact semiproximal ALM/ADMM for linearly constrained multi-block convex composite conic programming (CCCP), and the accelerated block coordinate descent method for multi-block CCCP.

Keywords: Convex composite quadratic programming, block symmetric Gauss-Seidel, Schur complement

AMS subject classifications: 90C06, 90C20, 90C25, 65F10

1 Introduction

It is well known that the classical block symmetric Gauss-Seidel (block sGS) method [1, 5, 10, 20] can be used to solve a symmetric positive semidefinite linear system of equations $\mathcal{Q}\boldsymbol{x} = \boldsymbol{b}$ where $\boldsymbol{x} = (x_1; \ldots; x_s)$ is partitioned into s blocks with $s \geq 2$. We are particularly interested in the case when $s > 2$. In this paper, we show that each cycle of the classical block sGS method

*Department of Mathematics, National University of Singapore, 10 Lower Kent Ridge Road, Singapore 119076 (matlixu@nus.edu.sg).

†Department of Mathematics and Risk Management Institute, National University of Singapore, 10 Lower Kent Ridge Road, Singapore 119076 (matsundf@nus.edu.sg).

‡Department of Mathematics, National University of Singapore, 10 Lower Kent Ridge Road, Singapore 119076 (mattohkc@nus.edu.sg).

exactly solves the corresponding convex quadratic programming (QP) problem but added with an extra proximal term depending on the previous iterate (say $\boldsymbol{x}^k$). Through such a connection to optimization, we are able to extend the result (which we name as the block sGS decomposition theorem) to a convex composite QP (CCQP) with an additional possibly nonsmooth term in x_1, and subsequently extend the classical block sGS method to solve a CCQP. We can also extend the classical block sGS method to the inexact setting, where the underlying linear system for each block of the new iterate $\boldsymbol{x}^{k+1}$ need not be solved exactly. Moreover, by borrowing ideas in the optimization literature, we are able to accelerate the classical block sGS method and provide new convergence results. More details will be given later.

Assume that $\mathcal{X}_i = \mathbb{R}^{n_i}$ for $i = 1, \ldots, s$, and $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_s$. Consider the following symmetric positive semidefinite block linear system of equations:

$$\mathcal{Q}\boldsymbol{x} = \boldsymbol{b}, \tag{1}$$

where for some $s \geq 2$, $\boldsymbol{x} = [x_1;\ \ldots;\ x_s] \in \mathcal{X}$, $\boldsymbol{b} = [b_1;\ \ldots;\ b_s] \in \mathcal{X}$, and

$$\mathcal{Q} \;=\; \begin{bmatrix} Q_{1,1} & \cdots & Q_{1,s} \\ \vdots & \vdots & \vdots \\ Q_{1,s}^* & \cdots & Q_{s,s} \end{bmatrix} \tag{2}$$

with $Q_{i,j} \in \mathbb{R}^{n_i \times n_j}$ for $1 \leq i, j \leq s$. It is well known that (1) is the optimality condition for the following unconstrained QP:

$$\text{(QP)} \quad \min\Big\{ q(\boldsymbol{x}) := \frac{1}{2}\langle \boldsymbol{x},\, \mathcal{Q}\boldsymbol{x}\rangle - \langle \boldsymbol{b},\, \boldsymbol{x}\rangle \mid \boldsymbol{x} \in \mathcal{X} \Big\}. \tag{3}$$

Note that even though our problem is phrased in the matrix-vector setting for convenience, one can consider the setting where each $\mathcal{X}_i$ is a real n_i-dimensional inner product space and $Q_{i,j}$ is a linear map from $\mathcal{X}_i$ to $\mathcal{X}_j$. Throughout the paper, we make the following assumption:

Assumption 1. *$\mathcal{Q}$ is symmetric positive semidefinite and each diagonal block $Q_{i,i}$ is symmetric positive definite for $i = 1, \ldots, s$.*

From the following decomposition of $\mathcal{Q}$:

$$\mathcal{Q} \;=\; \mathcal{U} + \mathcal{D} + \mathcal{U}^*, \tag{4}$$

where

$$\mathcal{U} = \begin{bmatrix} \mathbf{0} & Q_{1,2} & \cdots & Q_{1,s} \\ & \ddots & & \vdots \\ & & \ddots & Q_{s-1,s} \\ & & & \mathbf{0} \end{bmatrix}, \quad \mathcal{D} = \begin{bmatrix} Q_{1,1} & & & \\ & Q_{2,2} & & \\ & & \ddots & \\ & & & Q_{s,s} \end{bmatrix}, \tag{5}$$

the classical block sGS iteration in numerical analysis is usually derived as a natural generalization of the classical pointwise sGS for solving a symmetric *positive definite* linear system of equations, and the latter is typically derived as a fixed-point iteration for the sGS matrix splitting based on (4). Specifically, the block sGS fixed-point iteration in the third normal form (in the terminology used in [10]) reads as follows:

$$\widehat{\mathcal{Q}}(\boldsymbol{x}^{k+1} - \boldsymbol{x}^k) = \boldsymbol{b} - \mathcal{Q}\boldsymbol{x}^k, \tag{6}$$

where $\widehat{\mathcal{Q}} = (\mathcal{D} + \mathcal{U})\mathcal{D}^{-1}(\mathcal{D} + \mathcal{U}^*)$.

In this paper, we give a derivation of the classical block sGS method (6) from the optimization perspective. By doing so, we are able to extend the classical block sGS method to solve a structured CCQP problem of the form:

$$\text{(CCQP)} \quad \min\Big\{F(\boldsymbol{x}) \;:=\; p(x_1) + \frac{1}{2}\langle \boldsymbol{x},\, \mathcal{Q}\boldsymbol{x}\rangle - \langle \boldsymbol{b},\, \boldsymbol{x}\rangle \mid \boldsymbol{x} = [x_1; \ldots; x_s] \in \mathcal{X}\Big\}, \tag{7}$$

where $p : \mathcal{X}_1 \to (-\infty, \infty]$ is a proper closed convex function such as $p(x_1) = \|x_1\|_1$ or $p(x_1) = \delta_{\mathbb{R}^{n_1}_+}(x_1)$ (i.e., the indicator function of $\mathbb{R}^{n_1}_+$ which is defined as $\delta_{\mathbb{R}^{n_1}_+}(x_1) = 0$ if $x_1 \geq 0$ and $\delta_{\mathbb{R}^{n_1}_+}(x_1) = \infty$ otherwise). Our specific contributions are described in the next few paragraphs. We note that the main results presented here are parts of the thesis of the first author [15].

First, we establish the key result of the paper, the *block sGS decomposition theorem*, which states that each cycle of the block sGS method, say at the kth iteration, corresponds exactly to solving (7) with an additional proximal term $\frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}^k\|_{\mathcal{T}_{\mathcal{Q}}}$ added to its objective function, i.e.,

$$\min\Big\{p(x_1) + \frac{1}{2}\langle \boldsymbol{x},\, \mathcal{Q}\boldsymbol{x}\rangle - \langle \boldsymbol{b},\, \boldsymbol{x}\rangle + \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}^k\|^2_{\mathcal{T}_{\mathcal{Q}}} \mid \boldsymbol{x} \in \mathcal{X}\Big\}, \tag{8}$$

where $\mathcal{T}_{\mathcal{Q}} = \mathcal{U}\mathcal{D}^{-1}\mathcal{U}^*$, and $\|\boldsymbol{x}\|^2_{\mathcal{T}_{\mathcal{Q}}} = \langle \boldsymbol{x},\, \mathcal{T}_{\mathcal{Q}}\boldsymbol{x}\rangle$. It is clear that when $p(\cdot) \equiv 0$, the problem (7) is exactly the QP (3) associated with the linear system (1). Therefore, we can interpret the classical block sGS method as a proximal-point minimization method for solving the QP (3), and each cycle of the classical block sGS method solves exactly the proximal subproblem (8) associated with the QP (3). As far as we are aware of, this is the first time in which the classical block sGS method (6) (and also the pointwise sGS method) is derived from an optimization perspective.

Second, we also establish a factorization view of the block sGS decomposition theorem and show its equivalence to the Schur complement based (SCB) reduction procedure proposed in [14] for solving a recursively defined variant of the proximal subproblem (8). The SCB reduction procedure in [14] is derived by inductively finding an appropriate proximal term to be added to the objective function of (7) so that the block variables $x_s, x_{s-1}, \ldots, x_2$ can be eliminated in a sequential manner and thus ending with a minimization problem involving only the variable x_1. In a nutshell, we show that the SCB reduction procedure sequentially eliminates the blocks (in the reversed order starting from x_s) in the variable $\boldsymbol{x}$ of the proximal subproblem (8) by decomposing the proximal term $\frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}^k\|_{\mathcal{T}_{\mathcal{Q}}}$ also in a sequential manner. In turn, each of the reduction step corresponds exactly to one step in a cycle of the block sGS method.

Third, based on the block sGS decomposition theorem, we are able to extend the classical block sGS method for solving the QP (3) to solve the CCQP (7), and each cycle of the extended block sGS method corresponds precisely to solving the proximal subproblem (8). Our extension of the block sGS method has thus overcome the limitation of the classical method by allowing us to solve the nonsmooth CCQP which often arises in practice, for example, in semidefinite programming. Moreover, our extension also allows the updates of the blocks to be inexact. Our derivation even allows us to design an inexact version of the classical block sGS method, where the iterate $\boldsymbol{x}^{k+1}$ need not be computed exactly from (6). Such an inexact extension of the classical block sGS method for (6) appears to be new. We should emphasize that the inexact block sGS method is potentially very useful when a diagonal block, say $Q_{i,i}$, in (2) is large and the computation of $\boldsymbol{x}_i^{k+1}$ must be done via an iterative solver rather than a direct solver.

Fourth, armed with the optimization interpretation of each cycle of the block sGS method, it becomes easy for us to adapt ideas from the optimization literature to establish the iteration complexity of $O(\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2_{\widehat{\mathcal{Q}}}/k)$ for the extended block sGS method as well as to accelerate it to obtain the complexity of $O(\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2_{\widehat{\mathcal{Q}}}/(k+1)^2)$, after running for k cycles, where $\boldsymbol{x}^*$ is an optimal solution for (7). Just as in the classical block sGS method, we can obtain a linear rate of convergence for our extended inexact block sGS method under the assumption that $\mathcal{Q} \succ 0$. With the help of an extensive optimization literature on the linear convergences of proximal gradient methods, we are further able to relax the positive definiteness assumption on $\mathcal{Q}$ to some mild error bound assumptions on F (which hold automatically for many interesting applications) and derive at least R-linear convergence results for our extended block sGS method.

Recent research works in [12, 13, 14, 22] have shown that our block sGS decomposition theorem for the CCQP (7) can play an essential role in the design of efficient algorithms for solving various convex optimization problems such as convex composite quadratic semidefinite programming problems. Our experiences have shown that the inexact block sGS cycle can provide the much needed flexibility when one is designing an algorithm based on the framework of the proximal augmented Lagrangian (ALM) or proximal alternating direction method of multipliers (ADMM) for solving important large scale convex composite optimization problems. As a concrete illustration of the application of our block sGS decomposition theorem, we will briefly describe in section 5 on how to utilize the theorem in the design of the proximal augmented Lagrangian method for solving a linearly constrained convex composite quadratic programming problem.

The idea of sequentially updating the blocks of a multi-block variable, either in the Gauss-Seidel fashion or the successive over-relaxation (SOR) fashion, has been incorporated into quite a number of optimization algorithms [3] and in solving nonlinear equations [19]. Indeed the Gauss-Seidel (also known as the block coordinate descent) approach for solving optimization problems has been considered extensively; we refer the readers to [2, 9] for the literature review on the recent developments, especially for the case where $s > 2$. Here we would like to emphasize that even for the case of an unconstrained *smooth* convex minimization problem $\min\{f(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathcal{X}\}$, whose objective function $f(\boldsymbol{x})$ (not necessarily strongly convex) has a Lipschitz continuous gradient of modulus L, it is only proven recently in [2] that the block coordinate (gradient) descent method is globally convergent with the iteration complexity of $O(Ls/k)$ after k cycles, where s is the number of blocks. When $f(\boldsymbol{x})$ is the quadratic function in (3), the block coordinate descent method is precisely the classical block Gauss-Seidel (GS) method. In contrast to the block sGS method, each iteration of the block GS method does not appear to have an optimization equivalence. Despite the extensive work on the Gauss-Seidel approach for solving convex optimization problems, surprisingly, little is known about the symmetric Gauss-Seidel approach for solving the same problems except for the recent paper [22] which utilized our block sGS decomposition theorem to design an inexact accelerated block coordinate descent method to solve a problem of the form $\min\{p(x_1) + f(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathcal{X}\}$.

The remaining parts of the paper are organized as follows. The next section is devoted to the block sGS decomposition theorem for the CCQP (7). In section 3, we present a factorization view of the block sGS theorem and prove its equivalence to the SCB reduction procedure proposed in [14, 15]. In the following section, we derive the block sGS method from an optimization perspective and extend it to solve the CCQP (7). The convergence results for our extended block sGS method are also presented in this section. In section 5, the application of our block sGS decomposition theorem is demonstrated in the design of a proximal augmented Lagrangian method for solving a linearly constrained convex composite quadratic programming problem. The extension of the classical block symmetric SOR method for solving (7) is presented in section 6. We conclude our paper in the final section.

We end the section by giving some notation. For a symmetric matrix $\mathcal{Q}$, the notation $\mathcal{Q} \succeq 0$ ($\mathcal{Q} \succ 0$) means that the matrix $\mathcal{Q}$ is symmetric positive semidefinite (definite).

2 Derivation of the block sGS decomposition theorem for (7)

In this section, we will present the derivation of one cycle of the block sGS method for (7) from the optimization perspective as mentioned in the introduction.

Recall the decomposition of $\mathcal{Q}$ in (4), $\mathcal{U}$,$\mathcal{D}$ in (5) and the *sGS linear operator* defined by

$$\mathcal{T}_{\mathcal{Q}} = \mathcal{U}\mathcal{D}^{-1}\mathcal{U}^*. \tag{9}$$

Given $\bar{\boldsymbol{x}} \in \mathcal{X}$, corresponding to problem (7), we consider solving the following subproblem

$$\boldsymbol{x}^+ := \mathrm{argmin}_{\boldsymbol{x}\in\mathcal{X}} \Big\{ p(x_1) + q(\boldsymbol{x}) + \frac{1}{2}\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|^2_{\mathcal{T}_{\mathcal{Q}}} - \langle \boldsymbol{x},\, \Delta(\boldsymbol{\delta'}, \boldsymbol{\delta})\rangle \Big\}, \tag{10}$$

where $\boldsymbol{\delta'}, \boldsymbol{\delta} \in \mathcal{X}$ are two given error vectors with $\delta'_1 = \delta_1$, and

$$\Delta(\boldsymbol{\delta'}, \boldsymbol{\delta}) := \boldsymbol{\delta} + \mathcal{U}\mathcal{D}^{-1}(\boldsymbol{\delta} - \boldsymbol{\delta'}). \tag{11}$$

We note that the vectors $\boldsymbol{\delta'}, \boldsymbol{\delta}$ need not be known a priori. We should view $\boldsymbol{x}^+$ as an approximate solution to (10) without the perturbation term $\langle \boldsymbol{x},\, \Delta(\boldsymbol{\delta'}, \boldsymbol{\delta})\rangle$. Once $\boldsymbol{x}^+$ has been computed, the associated error vectors can then be obtained, and $\boldsymbol{x}^+$ is then the exact solution to the perturbed problem (10).

The following theorem shows that $\boldsymbol{x}^+$ can be computed by performing exactly one cycle of the block sGS method for (7). In particular, if $p(x_1) \equiv 0$ and $\boldsymbol{\delta'} = 0 = \boldsymbol{\delta}$, then the computation of $\boldsymbol{x}^+$ corresponds to exactly one cycle of the classical block sGS method. For the proof, we need to define the following notation for a given $\boldsymbol{x} = (x_1; \ldots; x_s)$,

$$x_{\geq i} = (x_i; \ldots; x_s), \quad x_{\leq i} = (x_1; \ldots; x_i), \quad i = 1, \ldots, s.$$

We also define $x_{\geq s+1} = \emptyset$.

Theorem 1 (sGS Decomposition)**.** *Assume that $\mathcal{Q} \succeq 0$ and the self-adjoint linear operators $\mathcal{Q}_{ii}$ are positive definite for all $i = 1, \ldots, s$. Then, it holds that*

$$\widehat{\mathcal{Q}} := \mathcal{Q} + \mathcal{T}_{\mathcal{Q}} = (\mathcal{D} + \mathcal{U})\mathcal{D}^{-1}(\mathcal{D} + \mathcal{U}^*) \succ 0. \tag{12}$$

For $i = s, \ldots, 2$, suppose that we have computed $x'_i \in \mathcal{X}_i$ defined as follows:

$$\begin{aligned} x'_i &:= \operatorname*{argmin}_{x_i \in \mathcal{X}_i} \; p(\bar{x}_1) + q(\bar{x}_{\leq i-1}; x_i; x'_{\geq i+1}) - \langle \delta'_i,\, x_i\rangle \\ &= Q_{ii}^{-1}\big(b_i + \delta'_i - \textstyle\sum_{j=1}^{i-1} Q^*_{ji}\bar{x}_j - \sum_{j=i+1}^{s} Q_{ij}x'_j\big). \end{aligned} \tag{13}$$

Then the optimal solution $\boldsymbol{x}^+$ *for* (10) *can be computed exactly via the following steps:*

$$\begin{cases} x_1^+ = \underset{x_1 \in \mathcal{X}_1}{\operatorname{argmin}} \; p(x_1) + q(x_1; x'_{\geq 2}) - \langle \delta_1, \, x_1 \rangle, \\ x_i^+ = \underset{x_i \in \mathcal{X}_i}{\operatorname{argmin}} \; p(x_1^+) + q(x_{\leq i-1}^+; x_i; x'_{\geq i+1}) - \langle \delta_i, \, x_i \rangle \\ \quad\; = Q_{ii}^{-1}\big(b_i + \delta_i - \sum_{j=1}^{i-1} Q_{ji}^* x_j^+ - \sum_{j=i+1}^{s} Q_{ij} x'_j\big), \quad i = 2, \ldots, s. \end{cases} \tag{14}$$

Proof. Since $\mathcal{D} \succ 0$, we know that $\mathcal{D}$, $\mathcal{D} + \mathcal{U}$ and $\mathcal{D} + \mathcal{U}^*$ are all nonsingular. Then, (12) can easily be obtained from the following observation

$$\mathcal{Q} + \mathcal{T}_{\mathcal{Q}} = \mathcal{D} + \mathcal{U} + \mathcal{U}^* + \mathcal{U}\mathcal{D}^{-1}\mathcal{U}^* = (\mathcal{D} + \mathcal{U})\mathcal{D}^{-1}(\mathcal{D} + \mathcal{U}^*). \tag{15}$$

Next we show the equivalence between (10) and (14). By noting that $\delta_1 = \delta'_1$ and $\mathcal{Q}_{11} \succ 0$, we can define x'_1 as follows:

$$x'_1 = \underset{x_1 \in \mathcal{X}_1}{\operatorname{argmin}} \; p(x_1) + q(x_1; x'_{\geq 2}) - \langle \delta'_1, \, x_1 \rangle = \underset{x_1 \in \mathcal{X}_1}{\operatorname{argmin}} \; p(x_1) + q(x_1; x'_{\geq 2}) - \langle \delta_1, \, x_1 \rangle = x_1^+. \tag{16}$$

The optimality conditions corresponding to x'_1 and x_1^+ in (16) can be written as

$$\begin{cases} Q_{11} x'_1 = b_1 - \gamma_1 + \delta'_1 - \sum_{j=2}^{s} Q_{ij} x'_j, & \text{(17a)} \\ Q_{11} x_1^+ = b_1 - \gamma_1 + \delta_1 - \sum_{j=2}^{s} Q_{ij} x'_j, & \text{(17b)} \end{cases}$$

where $\gamma_1 \in \partial p(x'_1) \equiv \partial p(x_1^+)$. Simple calculations show that (17a) together with (13) can equivalently be rewritten as

$$(\mathcal{D} + \mathcal{U})\boldsymbol{x}' = \boldsymbol{b} - \boldsymbol{\gamma} + \boldsymbol{\delta}' - \mathcal{U}^*\bar{\boldsymbol{x}},$$

where $\boldsymbol{\gamma} = (\gamma_1; 0; \ldots, 0) \in \mathcal{X}$, while (14) can equivalently be recast as

$$(\mathcal{D} + \mathcal{U}^*)\boldsymbol{x}^+ = \boldsymbol{b} - \boldsymbol{\gamma} + \boldsymbol{\delta} - \mathcal{U}\boldsymbol{x}'.$$

By substituting $\boldsymbol{x}' = (\mathcal{D} + \mathcal{U})^{-1}(\boldsymbol{b} - \boldsymbol{\gamma} + \boldsymbol{\delta}' - \mathcal{U}^*\bar{\boldsymbol{x}})$ into the above equation, we obtain that

$$\begin{aligned} (\mathcal{D} + \mathcal{U}^*)\boldsymbol{x}^+ &= \boldsymbol{b} - \boldsymbol{\gamma} + \boldsymbol{\delta} - \mathcal{U}(\mathcal{D} + \mathcal{U})^{-1}(\boldsymbol{b} - \boldsymbol{\gamma} + \boldsymbol{\delta}' - \mathcal{U}^*\bar{\boldsymbol{x}}) \\ &= \mathcal{D}(\mathcal{D} + \mathcal{U})^{-1}(\boldsymbol{b} - \boldsymbol{\gamma}) + \mathcal{U}(\mathcal{D} + \mathcal{U})^{-1}\mathcal{U}^*\bar{\boldsymbol{x}} + \boldsymbol{\delta} - \mathcal{U}(\mathcal{D} + \mathcal{U})^{-1}\boldsymbol{\delta}', \end{aligned}$$

which, together with (15), (11) and the definition of $\mathcal{T}_{\mathcal{Q}}$ in (9), implies that

$$(\mathcal{Q} + \mathcal{T}_{\mathcal{Q}})\boldsymbol{x}^+ = \boldsymbol{b} - \boldsymbol{\gamma} + \mathcal{T}_{\mathcal{Q}}\bar{\boldsymbol{x}} + \Delta(\boldsymbol{\delta}', \boldsymbol{\delta}). \tag{18}$$

In the above, we have used the fact that $(\mathcal{D}+\mathcal{U})\mathcal{D}^{-1}\mathcal{U}(\mathcal{D}+\mathcal{U})^{-1} = \mathcal{U}\mathcal{D}^{-1}$. By noting that (18) is in fact the optimality condition for (10) and $\mathcal{Q} + \mathcal{T}_{\mathcal{Q}} \succ 0$, we have thus obtained the equivalence between (10) and (14). This completes the proof of the theorem. □

We shall explain here the roles of the error vectors $\boldsymbol{\delta}'$ and $\boldsymbol{\delta}$ in the above block sGS decomposition theorem. There is no need to choose these error vectors in advance. We emphasize that x'_i and x_i^+ obtained from (13) and (14) should be viewed as approximate solutions to the minimization problems without the terms involving δ'_i and δ_i. Once these approximate solutions have been computed, they would generate δ_i' and δ_i automatically. With these known error vectors, we know that the computed approximate solutions are the exact solutions to the minimization problems in (13) and (14).

Theorem 1 shows that instead of solving the QP subproblem (10) directly with an N-dimensional variable $\boldsymbol{x}$, where $N = \sum_{i=1}^{s} n_i$, the computation can be decomposed into s pieces of smaller dimensional problems involving only the variable x_i for each $i = 1, \dots, s$. Such a decomposition is obviously highly useful for dealing with a large scale CCQP of the form (7) when N is very large. The benefit is especially important because the computation of x_i for $i = 2, \dots, s$ involves only solving linear systems of equations. Of course, one would still have to solve a potentially difficult subproblem involving the variable x_1 due to the presence of the possibly nonsmooth term $p(x_1)$, i.e.,

$$x_1^+ \;=\; \operatorname{argmin}\Big\{p(x_1) + \frac{1}{2}\langle x_1,\, Q_{11}x_1\rangle - \langle c_1,\, x_1\rangle \mid x_1 \in \mathcal{X}_1\Big\},$$

where c_1 is a known vector depending on the previously computed $x_s', \dots, x_2'$. However, in many applications, $p(x_1)$ is usually a simple nonsmooth function such as $\|x_1\|_1$, $\|x_1\|_\infty$, or $\delta_{\mathbb{R}_+^{n_1}}(x_1)$ for which the corresponding subproblem is not difficult to solve. As a concrete example, suppose that $Q_{11} = I_{n_1}$. Then $x_1^+ = \operatorname{Prox}_p(c_1)$ and the Moreau-Yosida proximal map $\operatorname{Prox}_p(c_1)$ can be computed efficiently for various nonsmooth function $p(\cdot)$ including the examples just mentioned. In fact, one can always make the subproblem easier to solve by (a) adding an additional proximal term $\frac{1}{2}\|x_1 - \bar{x}_1\|_{J_1}$ to (10), where $J_1 = \mu_1 I_{n_1} - Q_{11}$ with $\mu_1 = \|Q_{11}\|_2$; and (b) modifying the sGS operator to $\mathcal{U}\widehat{\mathcal{D}}^{-1}\mathcal{U}^*$, where $\widehat{\mathcal{D}} = \mathcal{D} + \operatorname{diag}(J_1, 0, \dots, 0)$. With the additional proximal term involving J_1, the subproblem corresponding to x_1 then becomes

$$\begin{aligned} x_1^+ &= \operatorname{argmin}\Big\{p(x_1) + \frac{1}{2}\langle x_1,\, Q_{11}x_1\rangle - \langle c_1,\, x_1\rangle + \frac{1}{2}\|x_1 - \bar{x}_1\|_{J_1}^2 \mid x_1 \in \mathbb{R}^{n_1}\Big\} \\ &= \operatorname{argmin}\Big\{p(x_1) + \frac{\mu_1}{2}\langle x_1,\, x_1\rangle - \langle c_1 + J_1\bar{x}_1,\, x_1\rangle \mid x_1 \in \mathbb{R}^{n_1}\Big\} \\ &= \operatorname{Prox}_{p/\mu_1}\big(\mu_1^{-1}(c_1 + J_1\bar{x}_1)\big). \end{aligned}$$

The following proposition is useful in estimating the error term $\Delta(\boldsymbol{\delta}', \boldsymbol{\delta})$ in (10).

Proposition 1. *Denote $\widehat{\mathcal{Q}} := \mathcal{Q} + \mathcal{T}_{\mathcal{Q}}$, which is positive definite. Let $\xi = \|\widehat{\mathcal{Q}}^{-1/2}\Delta(\boldsymbol{\delta}', \boldsymbol{\delta})\|$. It holds that*

$$\xi \le \|\mathcal{D}^{-1/2}(\boldsymbol{\delta} - \boldsymbol{\delta}')\| + \|\widehat{\mathcal{Q}}^{-1/2}\boldsymbol{\delta}'\|.$$

Proof. Recall that $\widehat{\mathcal{Q}} = (\mathcal{D} + \mathcal{U})\mathcal{D}^{-1}(\mathcal{D} + \mathcal{U}^*)$. Thus, we have

$$\widehat{\mathcal{Q}}^{-1} = (\mathcal{D} + \mathcal{U}^*)^{-1}\mathcal{D}(\mathcal{D} + \mathcal{U})^{-1} = (\mathcal{D} + \mathcal{U}^*)^{-1}\mathcal{D}^{1/2}\mathcal{D}^{1/2}(\mathcal{D} + \mathcal{U})^{-1},$$

which, together with the definition of $\Delta(\boldsymbol{\delta}', \boldsymbol{\delta})$ in (11), implies that

$$\xi = \|\mathcal{D}^{1/2}(\mathcal{D} + \mathcal{U})^{-1}\boldsymbol{\delta}' + \mathcal{D}^{-1/2}(\boldsymbol{\delta} - \boldsymbol{\delta}')\| \le \|\mathcal{D}^{1/2}(\mathcal{D} + \mathcal{U})^{-1}\boldsymbol{\delta}'\| + \|\mathcal{D}^{-1/2}(\boldsymbol{\delta} - \boldsymbol{\delta}')\|.$$

The desired result then follows. □

3 A factorization view of the block sGS decomposition theorem and its equivalence to the SCB reduction procedure

In this section, we present a factorization view of the block sGS decomposition theorem and show its equivalence to the Schur complement based (SCB) reduction procedure developed in [14, 15].

Let Θ_1 be the zero matrix in $\mathbb{R}^{n_1\times n_1}$ and $N_1 := n_1$. For $j = 2,\dots,s$, let $N_j := \sum_{i=1}^{j} n_i$ and define $\widehat{\Theta}_j \in \mathbb{R}^{N_{j-1}\times N_{j-1}}$ and $\Theta_j \in \mathbb{R}^{N_j\times N_j}$ as follows:

$$\widehat{\Theta}_j := \begin{bmatrix} Q_{1,j} \\ \vdots \\ Q_{j-1,j} \end{bmatrix} Q_{j,j}^{-1} \; \left[Q_{1,j}^*,\, \dots,\, Q_{j-1,j}^*\right]$$

and

$$\Theta_j := \begin{bmatrix} Q_{1,2} \\ 0 \\ \vdots \\ 0 \end{bmatrix} Q_{2,2}^{-1} \; \left[Q_{1,2}^*,\, 0,\, \dots,\, 0\right] + \cdots + \begin{bmatrix} Q_{1,j} \\ \vdots \\ Q_{j-1,j} \\ 0 \end{bmatrix} Q_{j,j}^{-1} \; \left[Q_{1,j}^*,\, \dots,\, Q_{j-1,j}^*,\, 0\right]. \tag{19}$$

Then, the above definitions indicate that, for $2 \le j \le s$,

$$\Theta_j = \mathrm{diag}(\Theta_{j-1}, 0_{n_j}) + \mathrm{diag}(\widehat{\Theta}_j, 0_{n_j}) \in \mathbb{R}^{N_j\times N_j}. \tag{20}$$

In [14, 15], the SCB reduction procedure corresponding to problem (7) is derived through the construction of the above self-adjoint linear operator Θ_s on $\mathcal{X}$. Now we recall the key steps in the SCB reduction procedure derived in the previous work. For $j = 1,\dots,s$, define

$$\mathcal{Q}_j := \begin{bmatrix} Q_{1,1} & \cdots & Q_{1,j-1} & Q_{1,j} \\ \vdots & \ddots & \vdots & \vdots \\ Q_{1,j-1}^* & \cdots & Q_{j-1,j-1} & Q_{j-1,j} \\ Q_{1,j}^* & \cdots & Q_{j-1,j}^* & Q_{j,j} \end{bmatrix}, \quad R_j := \begin{bmatrix} Q_{1,j} \\ \vdots \\ Q_{j-1,j} \end{bmatrix}.$$

It is easy to show that

$$\min\left\{p(x_1) + q(x_{\le s-1}; x_s) + \frac{1}{2}\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|_{\Theta_s}^2 \mid \boldsymbol{x} \in \mathcal{X}\right\} \tag{21}$$

$$= \min_{x_{\le s-1}\in\mathcal{X}_{\le s-1}} \left\{ \begin{array}{l} p(x_1) + \frac{1}{2}\langle x_{\le s-1},\, \mathcal{Q}_{s-1}x_{\le s-1}\rangle - \langle b_{\le s-1},\, x_{\le s-1}\rangle + \frac{1}{2}\|x_{\le s-1} - \bar{x}_{\le s-1}\|_{\Theta_{s-1}}^2 \\ +\min_{x_s\in\mathcal{X}_s}\left\{\frac{1}{2}\langle x_s,\, Q_{s,s}x_s\rangle - \langle b_s - R_s^* x_{\le s-1},\, x_s\rangle + \frac{1}{2}\|x_{\le s-1} - \bar{x}_{\le s-1}\|_{\widehat{\Theta}_s}^2\right\} \end{array} \right\}.$$

By first solving the inner minimization problem with respect to x_s, we get the solution as a function of $x_1,\dots,x_{s-1}$ as follows:

$$x_s = Q_{s,s}^{-1}\big(b_s - R_s^* x_{\le s-1}\big). \tag{22}$$

And the minimum value is given by

$$\begin{aligned} & -\frac{1}{2}\langle b_s - R_s^* x_{\le s-1},\, Q_{s,s}^{-1}(b_s - R_s^* x_{\le s-1})\rangle + \frac{1}{2}\|x_{\le s-1} - \bar{x}_{\le s-1}\|_{\widehat{\Theta}_s}^2 \\ = \; & -\frac{1}{2}\langle b_s,\, Q_{s,s}b_s\rangle + \frac{1}{2}\langle \bar{x}_{\le s-1},\, \widehat{\Theta}_s\bar{x}_{\le s-1}\rangle + \langle R_sQ_{s,s}^{-1}(b_s - R_s^*\bar{x}_{\le s-1}),\, x_{\le s-1}\rangle. \end{aligned}$$

Thus (21) reduces to a problem involving only the variables $x_1, \ldots, x_{s-1}$, which, up to a constant, is given by

$$
\begin{aligned}
&\min_{x_{\le s-1} \in \mathcal{X}_{\le s-1}} \left\{ \begin{array}{l} p(x_1) + \frac{1}{2}\langle x_{\le s-1},\, \mathcal{Q}_{s-1} x_{\le s-1}\rangle - \langle b_{\le s-1} - R_s x_s',\, x_{\le s-1}\rangle \\ + \frac{1}{2}\|x_{\le s-1} - \bar{x}_{\le s-1}\|^2_{\Theta_{s-1}} \end{array} \right\} \\
= &\min_{x_{\le s-1} \in \mathcal{X}_{\le s-1}} \left\{ p(x_1) + q(x_{\le s-1}; x_s') + \frac{1}{2}\|x_{\le s-1} - \bar{x}_{\le s-1}\|^2_{\Theta_{s-1}} \right\},
\end{aligned}
\tag{23}
$$

where $x_s' = Q_{s,s}^{-1}(b_s - R_s^* \bar{x}_{\le s-1})$. Observe that (23) has exactly the same form as (21). By repeating the above procedure to sequentially eliminate the variables $x_{s-1}, \ldots, x_1$, we will finally arrive at a minimization problem involving only the variable x_1. Once that minimization problem is solved, we can recover the solutions for $x_2, \ldots, x_s$ in a sequential manner.

Now we will prove the equivalence between the block sGS decomposition theorem and the SCB reduction procedure in the subsequent analysis by proving that $\Theta_s = \mathcal{T}_{\mathcal{Q}}$, where $\mathcal{T}_{\mathcal{Q}}$ is given in (9). For $j = 2, \ldots, s$, define the block matrices $\widehat{\mathcal{V}}_j \in \mathbb{R}^{N_j \times N_j}$ and $\mathcal{V}_j \in \mathbb{R}^{N \times N}$ by

$$
\widehat{\mathcal{V}}_j := \begin{bmatrix} I_{n_1} & & & Q_{1,j}Q_{j,j}^{-1} \\ & \ddots & & \vdots \\ & & I_{n_{j-1}} & Q_{j-1,j}Q_{j,j}^{-1} \\ & & & I_{n_j} \end{bmatrix} \in \mathbb{R}^{N_j \times N_j}, \quad \mathcal{V}_j := \operatorname{diag}(\widehat{\mathcal{V}}_j, I_{N-N_j}) \in \mathbb{R}^{N \times N}, \tag{24}
$$

where I_{n_j} is the $n_j \times n_j$ identity matrix. Note that $\widehat{\mathcal{V}}_s = \mathcal{V}_s$. Given $j \ge 2$, we have, by simple calculations, that for any $k < j$,

$$
\widehat{\mathcal{V}}_j^{-1} \begin{bmatrix} Q_{1,k} \\ \vdots \\ Q_{k-1,k} \\ 0 \end{bmatrix} = \begin{bmatrix} Q_{1,k} \\ \vdots \\ Q_{k-1,k} \\ 0 \end{bmatrix} \quad \text{and} \quad \widehat{\mathcal{V}}_j^{-1} \begin{bmatrix} Q_{1,j} \\ \vdots \\ Q_{j-1,j} \\ Q_{j,j} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ Q_{j,j} \end{bmatrix}. \tag{25}
$$

From (19) and (25), we have that

$$
\widehat{\mathcal{V}}_j^{-1} \Theta_j (\widehat{\mathcal{V}}_j^{-1})^* = \Theta_j, \quad j = 2, \ldots, s. \tag{26}
$$

Lemma 1. *Let $\mathcal{U}$ and $\mathcal{D}$ be given in* (5)*. It holds that*

$$
\mathcal{V}_2^* \cdots \mathcal{V}_s^* = \mathcal{D}^{-1}(\mathcal{D} + \mathcal{U}^*), \quad \mathcal{V}_s \cdots \mathcal{V}_2 = (\mathcal{D} + \mathcal{U})\mathcal{D}^{-1}.
$$

Proof. It can be verified directly that

$$
\mathcal{V}_2^* \cdots \mathcal{V}_s^* = \begin{bmatrix} I & & & \\ Q_{2,2}^{-1}Q_{1,2}^* & I & & \\ \vdots & \ddots & \ddots & \\ Q_{s,s}^{-1}Q_{1,s}^* & \cdots & Q_{s,s}^{-1}Q_{s-1,s}^* & I \end{bmatrix} = \mathcal{D}^{-1}(\mathcal{D} + \mathcal{U}^*).
$$

The second equality follows readily from the first. □

In the proof of the next lemma, we will make use of the well known fact that for given symmetric matrices A, C such that $C \succ 0$ and $M := A - BC^{-1}B^* \succ 0$, we have that

$$\begin{bmatrix} A & B \\ B^* & C \end{bmatrix} = \begin{bmatrix} I & BC^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} M & 0 \\ 0 & C \end{bmatrix} \begin{bmatrix} I & 0 \\ C^{-1}B^* & I \end{bmatrix}. \tag{27}$$

Theorem 2. *It holds that*

$$\mathcal{Q}_s + \Theta_s = \mathcal{V}_s \cdots \mathcal{V}_2 \, \mathcal{D} \, \mathcal{V}_2^* \cdots \mathcal{V}_s^* \quad \text{and} \quad \Theta_s = \mathcal{T}_{\mathcal{Q}}.$$

Proof. By using (27), for $j = 2, \dots, s$, we have that

$$\mathcal{Q}_j = \widehat{\mathcal{V}}_j \operatorname{diag}(\mathcal{M}_{j-1}, Q_{j,j}) \, \widehat{\mathcal{V}}_j^*,$$

where

$$\mathcal{M}_{j-1} = \begin{bmatrix} Q_{1,1} & \dots & Q_{1,j-1} \\ \vdots & \ddots & \vdots \\ Q_{1,j-1}^* & \dots & Q_{j-1,j-1} \end{bmatrix} - \begin{bmatrix} Q_{1,j} \\ \vdots \\ Q_{j-1,j} \end{bmatrix} Q_{j,j}^{-1} \begin{bmatrix} Q_{1,j}^*, & \dots, & Q_{j-1,j}^* \end{bmatrix} \; = \; \mathcal{Q}_{j-1} - \widehat{\Theta}_j.$$

Thus, from (26), we know that for $2 \leq j \leq s$,

$$\mathcal{Q}_j + \Theta_j = \widehat{\mathcal{V}}_j \Big(\operatorname{diag}(\mathcal{M}_{j-1}, Q_{j,j}) + \widehat{\mathcal{V}}_j^{-1} \Theta_j (\widehat{\mathcal{V}}_j^{-1})^* \Big) \widehat{\mathcal{V}}_j^* \; = \; \widehat{\mathcal{V}}_j \Big(\operatorname{diag}(\mathcal{M}_{j-1}, Q_{j,j}) + \Theta_j \Big) \widehat{\mathcal{V}}_j^*.$$

For $2 \leq j \leq s$, by (20), we have that

$$\operatorname{diag}(\mathcal{M}_{j-1}, Q_{j,j}) + \Theta_j = \operatorname{diag}(\mathcal{M}_{j-1} + \Theta_{j-1} + \widehat{\Theta}_j, Q_{j,j}) = \operatorname{diag}(\mathcal{Q}_{j-1} + \Theta_{j-1}, Q_{j,j})$$

and consequently,

$$\mathcal{Q}_j + \Theta_j = \widehat{\mathcal{V}}_j \operatorname{diag}(\mathcal{Q}_{j-1} + \Theta_{j-1}, \mathcal{Q}_{jj}) \, \widehat{\mathcal{V}}_j^*. \tag{28}$$

Thus, by recalling the definitions of $\widehat{\mathcal{V}}_j$ and $\mathcal{V}_j$ in (24) and using (28), we obtain through simple calculations that

$$\begin{aligned} \mathcal{Q}_s + \Theta_s &= \mathcal{V}_s \operatorname{diag}(\mathcal{Q}_{s-1} + \Theta_{s-1}, Q_{s,s}) \; \mathcal{V}_s^* \\ &= \vdots \\ &= \mathcal{V}_s \cdots \mathcal{V}_2 \operatorname{diag}(\mathcal{Q}_1 + \Theta_1, Q_{2,2}, \dots, Q_{s,s}) \; \mathcal{V}_2^* \cdots \mathcal{V}_s^*. \end{aligned}$$

Thus, by using the fact that $\mathcal{Q}_1 + \Theta_1 = Q_{1,1}$, we get

$$\mathcal{Q}_s + \Theta_s = \mathcal{V}_s \cdots \mathcal{V}_2 \operatorname{diag}(Q_{1,1}, Q_{2,2}, \dots, Q_{s,s}) \; \mathcal{V}_2^* \cdots \mathcal{V}_s^*.$$

By Lemma 1, it follows that

$$\mathcal{Q}_s + \Theta_s = (\mathcal{D} + \mathcal{U}) \mathcal{D}^{-1} \mathcal{D} \mathcal{D}^{-1} (\mathcal{D} + \mathcal{U}^*) \; = \; (\mathcal{D} + \mathcal{U}) \mathcal{D}^{-1} (\mathcal{D} + \mathcal{U}^*) \; = \; \mathcal{Q} + \mathcal{T}_{\mathcal{Q}},$$

where the last equation follows from (12) in Theorem 1. Since $\mathcal{Q}_s = \mathcal{Q}$, we know that

$$\Theta_s = \mathcal{T}_{\mathcal{Q}}.$$

This completes the proof of the theorem. □

4 An extended block sGS method for solving the CCQP (7)

With the block sGS decomposition theorem (Theorem 1) and Proposition 1, we can now extend the classic block sGS method to solve the CCQP (7). The detail steps of the algorithm for solving (7) are given as follows.

Algorithm 1: An sGS based inexact proximal gradient method for (7).

Input $\widetilde{\boldsymbol{x}}^1 = \boldsymbol{x}^0 \in \mathrm{dom}(p) \times \mathbb{R}^{n_2} \times \ldots \times \mathbb{R}^{n_s}$, $t_1 = 1$ and a summable sequence of nonnegative numbers $\{\epsilon_k\}$. For $k = 1, 2, \ldots$, perform the following steps in each iteration.

Step 1. Compute

$$\boldsymbol{x}^k = \mathrm{argmin}_{\boldsymbol{x}\in\mathcal{X}} \Big\{ p(x_1) + q(\boldsymbol{x}) + \frac{1}{2}\|\boldsymbol{x} - \widetilde{\boldsymbol{x}}^k\|^2_{\mathcal{T}_{\mathcal{Q}}} - \langle \boldsymbol{x},\, \Delta(\widetilde{\boldsymbol{\delta}}^k, \boldsymbol{\delta}^k)\rangle \Big\}, \tag{29}$$

via the sGS decomposition procedure described in Theorem 1, where $\widetilde{\boldsymbol{\delta}}^k$, $\boldsymbol{\delta}^k \in \mathcal{X}$ are error vectors such that

$$\max\{\|\widetilde{\boldsymbol{\delta}}^k\|, \|\boldsymbol{\delta}^k\|\} \le \frac{\epsilon_k}{t_k}. \tag{30}$$

Step 2. Choose t_{k+1} such that $t_{k+1}^2 - t_{k+1} \le t_k^2$ and set $\beta_k = \frac{t_k - 1}{t_{k+1}}$. Compute

$$\widetilde{\boldsymbol{x}}^{k+1} = \boldsymbol{x}^k + \beta_k(\boldsymbol{x}^k - \boldsymbol{x}^{k-1}).$$

We have the following iteration complexity convergence results for Algorithm 1.

Proposition 2. *Suppose that $\boldsymbol{x}^*$ is an optimal solution of problem* (7). *Let $\{\boldsymbol{x}^k\}$ be the sequence generated by Algorithm 1. Define $M = 2\|\mathcal{D}^{-1/2}\|_2 + \|\widehat{\mathcal{Q}}^{-1/2}\|_2$.*

(a) If $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$ for all $k \ge 1$, it holds that

$$0 \le F(\boldsymbol{x}^k) - F(\boldsymbol{x}^*) \le \frac{2}{(k+1)^2}\Big(\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_{\widehat{\mathcal{Q}}} + \bar{\epsilon}_k\Big)^2,$$

where $\widehat{\mathcal{Q}} = \mathcal{Q} + \mathcal{T}_{\mathcal{Q}}$ and $\bar{\epsilon}_k = 2M\sum_{i=1}^k \epsilon_i$.

(b) If $t_k = 1$ for all $k \ge 1$, it holds that

$$0 \le F(\boldsymbol{x}^k) - F(\boldsymbol{x}^*) \le \frac{1}{2k}\Big(\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_{\widehat{\mathcal{Q}}} + \tilde{\epsilon}_k\Big)^2,$$

where $\tilde{\epsilon}_k = 4M\sum_{i=1}^k i\epsilon_i$.

Proof. (a) The result can be proved by applying Theorem 2.1 in [11]. In order to apply the theorem, we need to verify that the error $\boldsymbol{e} := \boldsymbol{\gamma} + \mathcal{Q}\boldsymbol{x}^k - \boldsymbol{b} + \mathcal{T}_{\mathcal{Q}}(\boldsymbol{x}^k - \widetilde{\boldsymbol{x}}^k)$, where $\boldsymbol{\gamma} = (\gamma_1; 0; \ldots; 0)$ and $\gamma_1 \in \partial p(x_1^k)$, incurred for solving the subproblem (without the perturbation term $\Delta(\widetilde{\boldsymbol{\delta}}^k, \boldsymbol{\delta}^k)$) in Step 1 inexactly is sufficiently small. From Theorem 1, we know that

$$\boldsymbol{e} := \boldsymbol{\gamma} + \mathcal{Q}\boldsymbol{x}^k - \boldsymbol{b} + \mathcal{T}_{\mathcal{Q}}(\boldsymbol{x}^k - \widetilde{\boldsymbol{x}}^k) = \Delta(\widetilde{\boldsymbol{\delta}}^k, \boldsymbol{\delta}^k).$$

The theorem is proved via Theorem 2.1 in [11] if we can show that $\|\widehat{\mathcal{Q}}^{-1/2}\Delta(\widetilde{\boldsymbol{\delta}}^k,\boldsymbol{\delta}^k)\| \leq M\frac{\epsilon_k}{t_k}$. But from (30) and Proposition 1, we have that

$$\|\widehat{\mathcal{Q}}^{-1/2}\Delta(\widetilde{\boldsymbol{\delta}}^k,\boldsymbol{\delta}^k)\| \leq \|\mathcal{D}^{-1/2}\boldsymbol{\delta}^k\| + \|\mathcal{D}^{-1/2}\widetilde{\boldsymbol{\delta}}^k\| + \|\widehat{\mathcal{Q}}^{-1/2}\widetilde{\boldsymbol{\delta}}^k\| \leq M\epsilon_k/t_k,$$

thus the required inequality indeed holds true, and the proof is completed.

(b) There is no straightforward theorem for which we can apply to prove the result, we will provide the proof in the Appendix. □

Remark 1. *It is not difficult to show that if* $p(\cdot) \equiv 0$, $t_k = 1$, *and* $\boldsymbol{\delta}^k = \widetilde{\boldsymbol{\delta}}^k = 0$ *for all* $k \geq 1$, *then Algorithm 1 exactly coincides with the classical block sGS method* (6). *Proposition 2 shows that the classical block sGS method for solving* (1) *can be extended to solve the convex composite QP* (7). *It also demonstrates the advantage of interpreting the block sGS method from the optimization respective. For example, one can obtain the* $O(1/k)$ *iteration complexity result for the classical block sGS method without assuming that* $\mathcal{Q}$ *is positive definite. To the best of our knowledge, such a complexity result for the classical block sGS is new. More importantly, inexact and accelerated versions of the block sGS method can also be derived for* (1).

Remark 2. *In solving* (29) *via the sGS decomposition procedure to satisfy the error condition* (30), *let* $\boldsymbol{x}' = [x_1'; \ldots; x_s']$ *be the intermediate solution computed during the backward GS sweep and the associated error vector be* $\widetilde{\boldsymbol{\delta}}^k = [\widetilde{\delta}_1^k; \ldots; \widetilde{\delta}_s^k]$. *In the forward GS sweep, one can often save computations by using the computed* x_i' *to estimate* x_i^{k+1} *for* $i \geq 2$, *and the resulting error vector will be given by* $\delta_i^k = \widetilde{\delta}_i^k + \sum_{j=1}^{i-1} Q_{ji}^*(x_j^{k+1} - \widetilde{x}_j^k)$. *If we have that*

$$\|\textstyle\sum_{j=1}^{i-1} Q_{ji}^*(x_j^{k+1} - \widetilde{x}_j^k)\| \;\leq\; \rho, \tag{31}$$

where $\rho = \frac{c}{\sqrt{s}}\|\widetilde{\boldsymbol{\delta}}^k\|$ *and* $c > 0$ *is some given constant, then clearly* $\|\delta_i^k\|^2 \leq 2\|\widetilde{\delta}_i^k\|^2 + 2\rho^2$. *When all the error components* $\|\delta_i^k\|^2$ *satisfy the previous bound for* $i = 1, \ldots, s$, *regardless of whether* x_i^{k+1} *is estimated from* x_i' *or computed afresh, we get* $\|\boldsymbol{\delta}^k\| \leq \sqrt{2(1+c^2)}\|\widetilde{\boldsymbol{\delta}}^k\|$. *Consequently the error condition* (30) *can be satisfied with a slightly larger error tolerance* $\sqrt{2(1+c^2)}\,\epsilon_k/t_k$. *It is easy to see that one can use the condition in* (31) *to decide whether* x_i^{k+1} *can be estimated from* x_i' *without contributing a large error to* $\|\boldsymbol{\delta}^k\|$ *for each* $i = 2, \ldots, s$.

Besides the above iteration complexity results, one can also study the linear convergence rate of Algorithm 1. Indeed, just as in the case of the classical block sGS method, the convergence rate of our extended inexact block sGS method for solving (7) can also be established when $\mathcal{Q} \succ 0$. The precise result is given in the next theorem.

Theorem 3. *Suppose that* $\mathrm{ri}(\mathrm{dom}(p)) \neq \emptyset$, $\mathcal{Q} \succ 0$ *and* $t_k = 1$ *for all* $k \geq 1$. *Then*

$$\|\widehat{\mathcal{Q}}^{-1/2}(\boldsymbol{x}^k - \boldsymbol{x}^*)\| \leq \|\mathcal{B}\|_2^k \|\widehat{\mathcal{Q}}^{-1/2}(\boldsymbol{x}^0 - \boldsymbol{x}^*)\| + M\|\mathcal{B}\|_2^k \sum_{j=1}^{k} \|\mathcal{B}\|_2^{-j}\epsilon_j, \tag{32}$$

where $\mathcal{B} = I - \widehat{\mathcal{Q}}^{-1/2}\mathcal{Q}\widehat{\mathcal{Q}}^{-1/2}$. *Note that* $0 \preceq \mathcal{B} \prec I$.

Proof. For notational convenience, we let $\Delta^j = \Delta(\widetilde{\boldsymbol{\delta}}^j, \boldsymbol{\delta}^j)$ in this proof.

Define $E_1 : \mathcal{X} \to \mathcal{X}_1$ by $E_1(\boldsymbol{x}) = x_1$ and $\widehat{p} : \mathcal{X} \to (-\infty, \infty]$ by $\widehat{p}(\boldsymbol{x}) = p(E_1 \widehat{\mathcal{Q}}^{-1/2}\boldsymbol{x})$. Since $\widehat{\mathcal{Q}} \succ 0$, it is clear that $\text{range}(E_1\widehat{\mathcal{Q}}^{-1/2}) = \mathcal{X}_1$ and hence $\text{ri}(\text{dom}(p)) \cap \text{range}(E_1\widehat{\mathcal{Q}}^{-1/2}) \neq \emptyset$. By [18, Theorem 23.9], we have that

$$\partial \widehat{p}(\boldsymbol{x}) = \widehat{\mathcal{Q}}^{-1/2} E_1^* \partial p(E_1 \widehat{\mathcal{Q}}^{-1/2}\boldsymbol{x}) \quad \forall\, \boldsymbol{x} \in \mathcal{X}. \tag{33}$$

From the optimality condition of $\boldsymbol{x}^j$, we have that

$$\begin{aligned} & 0 = \boldsymbol{\gamma}^j + \widehat{\mathcal{Q}}(\boldsymbol{x}^j - \boldsymbol{x}^{j-1}) - \boldsymbol{b} + \mathcal{Q}\boldsymbol{x}^{j-1} - \Delta^j \\ \Leftrightarrow \quad & \widehat{\mathcal{Q}}^{1/2}\boldsymbol{x}^{j-1} + \widehat{\mathcal{Q}}^{-1/2}(\boldsymbol{b} - \mathcal{Q}\boldsymbol{x}^{j-1}) + \widehat{\mathcal{Q}}^{-1/2}\Delta^j = \widehat{\mathcal{Q}}^{-1/2}\boldsymbol{\gamma}^j + \widehat{\mathcal{Q}}^{1/2}\boldsymbol{x}^j, \end{aligned}$$

where $\boldsymbol{\gamma}^j = (\gamma_1^j; 0; \ldots; 0)$ with $\gamma_1^j \in \partial p(x_1^j)$. Let $\hat{\boldsymbol{x}}^j = \widehat{\mathcal{Q}}^{1/2}\boldsymbol{x}^j$ and $\hat{\boldsymbol{x}}^{j-1} = \widehat{\mathcal{Q}}^{1/2}\boldsymbol{x}^{j-1}$. Then we have that

$$\begin{aligned} & \hat{\boldsymbol{x}}^{j-1} + \widehat{\mathcal{Q}}^{-1/2}(\boldsymbol{b} - \mathcal{Q}\boldsymbol{x}^{j-1}) + \widehat{\mathcal{Q}}^{-1/2}\Delta^j \in (I + \partial\widehat{p})(\hat{\boldsymbol{x}}^j) \\ \Leftrightarrow \quad & \hat{\boldsymbol{x}}^j = \text{Prox}_{\widehat{p}}\big(\mathcal{B}\hat{\boldsymbol{x}}^{j-1} + \widehat{\mathcal{Q}}^{-1/2}\boldsymbol{b} + \widehat{\mathcal{Q}}^{-1/2}\Delta^j\big). \end{aligned}$$

Similarly if $\boldsymbol{x}^*$ is an optimal solution of (7), then we have that

$$\hat{\boldsymbol{x}}^* = \text{Prox}_{\widehat{p}}(\mathcal{B}\hat{\boldsymbol{x}}^* + \widehat{\mathcal{Q}}^{-1/2}\boldsymbol{b}).$$

By using the nonexpansive property of $\text{Prox}_{\widehat{p}}$, we have that

$$\begin{aligned} \|\hat{\boldsymbol{x}}^j - \hat{\boldsymbol{x}}^*\| &= \|\text{Prox}_{\widehat{p}}\big(\mathcal{B}\hat{\boldsymbol{x}}^{j-1} + \widehat{\mathcal{Q}}^{-1/2}\boldsymbol{b} + \widehat{\mathcal{Q}}^{-1/2}\Delta^j\big) - \text{Prox}_{\widehat{p}}(\mathcal{B}\hat{\boldsymbol{x}}^* + \widehat{\mathcal{Q}}^{-1/2}\boldsymbol{b})\| \\ &\leq \|(\mathcal{B}\hat{\boldsymbol{x}}^{j-1} + \widehat{\mathcal{Q}}^{-1/2}\boldsymbol{b} + \widehat{\mathcal{Q}}^{-1/2}\Delta^j) - (\mathcal{B}\hat{\boldsymbol{x}}^* + \widehat{\mathcal{Q}}^{-1/2}\boldsymbol{b})\| \\ &\leq \|\mathcal{B}\|_2\|\hat{\boldsymbol{x}}^{j-1} - \hat{\boldsymbol{x}}^*\| + M\epsilon_j. \end{aligned}$$

By applying the above inequality sequentially for $j = k, k-1, \ldots, 1$, we get the required result in (32). □

Remark 3. *In fact, one can weaken the positive definiteness assumption of $\mathcal{Q}$ in the above theorem and still expect a linear rate of convergence. As a simple illustration, we only discuss here the exact version of Algorithm 1, i.e., $\widetilde{\boldsymbol{\delta}}^k = \boldsymbol{\delta}^k = 0$, under the error bound condition [16, 17] on F which holds automatically if p is a convex piecewise quadratic/linear function such as $p(x_1) = \|x_1\|_1$, $p(x_1) = 0$ for $x_1 \in \mathbb{R}^{n_1}$ or if $\mathcal{Q} \succ 0$. When $t_k = 1$ for all $k \geq 1$, one can prove that $\{F(\boldsymbol{x}^k)\}$ converges at least Q-linearly and $\{\boldsymbol{x}^k\}$ converges at least R-linearly to an optimal solution of problem* (7) *by using the techniques developed in [16, 17]. Interested readers may refer to [24, 26] for more details. For the accelerated case, with the additional fixed restarting scheme incorporated in Algorithm 1, both the R-linear convergences of $\{F(\boldsymbol{x}^k)\}$ and $\{\boldsymbol{x}^k\}$ can be obtained from [25, Corollary 3.8].*

5 An illustration on the application of the block sGS decomposition theorem in designing an efficient proximal ALM

In this section, we demonstrate the usefulness of our block sGS decomposition theorem as a building block for designing an efficient proximal ALM for solving a linearly constrained convex composite QP problem given by

$$\min \Big\{ p(x_1) + \frac{1}{2}\langle \boldsymbol{x},\, \mathcal{P}\boldsymbol{x}\rangle - \langle \boldsymbol{g},\, \boldsymbol{x}\rangle \mid \mathcal{A}\boldsymbol{x} = d \Big\}, \tag{34}$$

where $\mathcal{P}$ is a positive semidefinite linear operator on $\mathcal{X}$, $\mathcal{A} : \mathcal{X} \to \mathcal{Y}$ is a given linear map, and $\boldsymbol{g} \in \mathcal{X}$, $d \in \mathcal{Y}$ are given data. Here $\mathcal{X}$ and $\mathcal{Y}$ are two finite dimensional inner product spaces. Specifically, we show how the block sGS decomposition theorem given in Theorem 1 can be applied within the proximal ALM. We must emphasize that our main purpose here is to briefly illustrate the usefulness of the block sGS decomposition theorem but not to focus on the proximal ALM itself. Indeed, simply being capable of handling the nonsmooth function $p(\cdot)$ has already distinguished our approach from other approaches of using the sGS technique in optimization algorithms, e.g, [6, 7], where the authors incorporated the pointwise sGS splitting as a preconditioner within the Douglas–Rachford splitting method for a convex-concave saddle point problem.

In depth analysis of various recently developed ADMM-type algorithms and accelerated block coordinate descent algorithms employing the block sGS decomposition theorem as a building block can be found in [4, 12, 13, 14, 22]. Thus we shall not elaborate here again on the essential role played by the block sGS decomposition theorem in the design of those algorithms.

Although the problem (34) looks deceivingly simple, in fact it is a powerful model which includes the important class of standard convex quadratic semidefinite programming (QSDP) in the dual form given by

$$\min\Big\{\frac{1}{2}\langle W,\, \mathcal{H}W\rangle - \langle h,\, \xi\rangle \mid Z + \mathcal{B}^*\xi + \mathcal{H}W = C,\ \xi \in \mathbb{R}^p,\ Z \in \mathbb{S}^n_+,\ W \in \mathcal{W}\Big\}, \tag{35}$$

where $h \in \mathbb{R}^p$, $C \in \mathbb{S}^n$ are given data, $\mathcal{B} : \mathbb{S}^n \to \mathbb{R}^p$ is a given linear map that is assumed to be surjective, $\mathcal{H} : \mathbb{S}^n \to \mathbb{S}^n$ is a self-adjoint positive semidefinite linear operator, and $\mathcal{W} \subseteq \mathbb{S}^n$ is any subspace containing $\mathrm{Ran}(\mathcal{H})$, the range space of $\mathcal{H}$. Here $\mathbb{S}^n$ denotes the space of $n \times n$ symmetric matrices and $\mathbb{S}^n_+$ denotes the cone of symmetric positive semidefinite matrices in $\mathbb{S}^n$. One can obviously express the QSDP problem (35) in the form of (34) by defining $\boldsymbol{x} = (Z; \xi; W)$, $p(Z) = \delta_{\mathbb{S}^n_+}(Z)$, $\mathcal{P} = \mathrm{diag}(0, 0, \mathcal{H})$, and $\mathcal{A} = (\mathcal{I}, \mathcal{B}^*, \mathcal{H})$.

We begin with the augmented Lagrangian function associated with (34):

$$L_\sigma(\boldsymbol{x}; y) = p(x_1) + \frac{1}{2}\langle \boldsymbol{x},\, \mathcal{P}\boldsymbol{x}\rangle - \langle \boldsymbol{g},\, \boldsymbol{x}\rangle + \frac{\sigma}{2}\|\mathcal{A}\boldsymbol{x} - d + \sigma^{-1}y\|^2 - \frac{1}{2\sigma}\|y\|^2, \tag{36}$$

where $\sigma > 0$ is a given penalty parameter and $y \in \mathcal{Y}$ is the multiplier associated with the equality constraint. The template for a proximal ALM is given as follows. Given $\mathcal{T} \succeq 0$, $\boldsymbol{x}^0 \in \mathcal{X}$ and $y^0 \in \mathcal{Y}$. Perform the following steps in each iteration.

Step 1. Compute

$$\begin{aligned} \boldsymbol{x}^{k+1} &= \mathrm{argmin}\Big\{L_\sigma(\boldsymbol{x}; y^k) + \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}^k\|^2_{\mathcal{T}} \mid \boldsymbol{x} \in \mathcal{X}\Big\} \\ &= \mathrm{argmin}\Big\{p(x_1) + \frac{1}{2}\langle \boldsymbol{x},\, (\mathcal{P} + \sigma\mathcal{A}^*\mathcal{A})\boldsymbol{x}\rangle - \langle \boldsymbol{b},\, \boldsymbol{x}\rangle + \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}^k\|^2_{\mathcal{T}} \mid \boldsymbol{x} \in \mathcal{X}\Big\}, \end{aligned} \tag{37}$$

where $\boldsymbol{b} = \boldsymbol{g} + \mathcal{A}^*(\sigma d - y^k)$.

Step 2. Compute $y^{k+1} = y^k + \tau\sigma(\mathcal{A}\boldsymbol{x}^k - d)$, where $\tau \in (0, 2)$ is the dual step-length.

It is clear that the subproblem (37) has the form given in (7). Thus, one can apply the block sGS decomposition theorem to efficiently solve the subproblem if we choose $\mathcal{T} = \mathcal{T}_{\mathcal{P}+\sigma\mathcal{A}^*\mathcal{A}}$, i.e., the sGS operator associated with $\mathcal{Q} := \mathcal{P} + \sigma\mathcal{A}^*\mathcal{A}$. For the QSDP problem (35) with $\mathcal{W} := \mathrm{Ran}\mathcal{H}$, we have that

$$\mathcal{Q} = \sigma \begin{pmatrix} \mathcal{I} & \mathcal{B}^* & \mathcal{H} \\ \mathcal{B} & \mathcal{B}\mathcal{B}^* & \mathcal{B}\mathcal{H} \\ \mathcal{H} & \mathcal{H}\mathcal{B}^* & \sigma^{-1}\mathcal{H} + \mathcal{H}^2 \end{pmatrix}$$

and that the subproblem (37) can be efficiently solved by one cycle of the extended block sGS method explicitly as follows, given the iterate $(Z^k, \xi^k, \mathcal{H}W^k)$ and multiplier y^k.

Step 1a. Compute $\mathcal{H}W'$ as the solution of $(\sigma^{-1}\mathcal{I} + \mathcal{H})\mathcal{H}W' = \sigma^{-1}b_W - \mathcal{H}Z^k - \mathcal{H}\mathcal{B}^*\xi^k$, where $b_W = \mathcal{H}(\sigma C - Y^k)$.

Step 1b. Compute ξ' from $\mathcal{B}\mathcal{B}^*\xi' = \sigma^{-1}b_\xi - \mathcal{B}Z^k - \mathcal{B}\mathcal{H}W'$, where $b_\xi = h + \mathcal{B}(\sigma C - Y^k)$.

Step 1c. Compute $Z^{k+1} = \mathrm{argmin}\Big\{\delta_{\mathbb{S}^n_+}(Z) + \frac{\sigma}{2}\|Z + \mathcal{B}^*\xi' + \mathcal{H}W' - \sigma^{-1}b_Z\|^2\Big\}$, where $b_Z = \sigma C - Y^k$.

Step 1d. Compute ξ^{k+1} from $\mathcal{B}\mathcal{B}^*\xi^{k+1} = \sigma^{-1}b_\xi - \mathcal{B}Z^{k+1} - \mathcal{B}\mathcal{H}W'$.

Step 1e. Compute $\mathcal{H}W^{k+1}$ from $(\sigma^{-1}\mathcal{I} + \mathcal{H})\mathcal{H}W^{k+1} = \sigma^{-1}b_W - \mathcal{H}Z^{k+1} - \mathcal{H}\mathcal{B}^*\xi^{k+1}$.

From the above implementation, one can see how simple it is for one to apply the block sGS decomposition theorem to solve the complicated subproblem (37) arising from QSDP. Note that in Step 1a and Step 1e, we only need to compute $\mathcal{H}W'$ and $\mathcal{H}W^{k+1}$, respectively, and we do not need the values of W' and W^{k+1} explicitly. Here, for simplicity, we only write down the exact version of a proximal ALM by using our exact block sGS decomposition theorem. Without any difficulty, one can also apply our inexact version of the block sGS decomposition theorem to derive a more practical inexact proximal ALM for solving (34), say when the linear systems involved are large scale and have to be solved by a Krylov iterative method.

6 Extension of the block symmetric SOR method for solving (7)

In a way similar to what we have done in section 2, we show in this section that the block symmetric SOR (block sSOR) method can also be interpreted from an optimization perspective.

Given a parameter $\omega \in [1, 2)$, the kth iteration of the block sSOR method in the third normal form is defined by

$$\mathcal{W}(\boldsymbol{x}^{k+1} - \boldsymbol{x}^k) = b - \mathcal{Q}\boldsymbol{x}^k, \tag{38}$$

where

$$\mathcal{W} = (\tau\mathcal{D} + \mathcal{U})^{-1}(\rho\mathcal{D})^{-1}(\tau\mathcal{D} + \mathcal{U}^*),$$

$\tau = \omega^{-1}$, and $\rho = 2\tau - 1$. Note that for $\omega \in [1, 2)$, we have that $\tau \in (1/2, 1]$ and $\rho \in (0, 1]$. We should mention that the block sSOR method is typically not derived in the form given in (38), see for example [10, p.117], but one can show with some algebraic manipulations that (38) is an equivalent reformulation.

Denote

$$\mathcal{T}_{\mathrm{sSOR}} := \big((1-\tau)\mathcal{D} + \mathcal{U}\big)(\rho\mathcal{D})^{-1}\big((1-\tau)\mathcal{D} + \mathcal{U}^*\big).$$

In the next proposition, we show that $\mathcal{W}$ can be decomposed as the sum of $\mathcal{Q}$ and $\mathcal{T}_{\mathrm{sSOR}}$. Similar as the linear operator $\mathcal{T}_{\mathcal{Q}}$ in section 2, $\mathcal{T}_{\mathrm{sSOR}}$ is the key ingredient which enable us to derive the block sSOR method from the optimization perspective.

Proposition 3. *Let $\tau \in (1/2, 1]$ and $\rho = 2\tau - 1$. It holds that*

$$\mathcal{W} = \mathcal{Q} + \mathcal{T}_{\rm sSOR}. \tag{39}$$

Proof. Let $\bar{\tau} := \tau - \frac{1}{2} > 0$ and $\overline{\mathcal{U}} = \mathcal{U} + \frac{1}{2}\mathcal{D}$. Note that $\rho = 2\bar{\tau}$ and

$$\begin{aligned}
\mathcal{W} &= (\bar{\tau}\mathcal{D} + \overline{\mathcal{U}})(2\bar{\tau}\mathcal{D})^{-1}(\bar{\tau}\mathcal{D} + \overline{\mathcal{U}}^*) \\
&= \frac{1}{2}(\bar{\tau}\mathcal{D} + \overline{\mathcal{U}})(I + (\bar{\tau}\mathcal{D})^{-1}\overline{\mathcal{U}}^*) = \frac{1}{2}(\bar{\tau}\mathcal{D} + \overline{\mathcal{U}} + \overline{\mathcal{U}}^* + \overline{\mathcal{U}}(\bar{\tau}\mathcal{D})^{-1}\overline{\mathcal{U}}^*) \\
&= \frac{1}{2}(\mathcal{Q} + \bar{\tau}\mathcal{D} + \overline{\mathcal{U}}(\bar{\tau}\mathcal{D})^{-1}\overline{\mathcal{U}}^*) \\
&= \mathcal{Q} + \frac{1}{2}\big(\bar{\tau}\mathcal{D} + \overline{\mathcal{U}}(\bar{\tau}\mathcal{D})^{-1}\overline{\mathcal{U}}^* - \mathcal{Q}\big).
\end{aligned}$$

Now

$$\begin{aligned}
&\bar{\tau}\mathcal{D} + \overline{\mathcal{U}}(\bar{\tau}\mathcal{D})^{-1}\overline{\mathcal{U}}^* - \mathcal{Q} \;=\; \bar{\tau}\mathcal{D} + \overline{\mathcal{U}}(\bar{\tau}\mathcal{D})^{-1}\overline{\mathcal{U}}^* - \overline{\mathcal{U}} - \overline{\mathcal{U}}^* \\
&= \;(\bar{\tau}\mathcal{D} - \overline{\mathcal{U}})(\bar{\tau}\mathcal{D})^{-1}(\bar{\tau}\mathcal{D} - \overline{\mathcal{U}}^*) \;=\; \big((1-\tau)\mathcal{D} + \mathcal{U}\big)(\bar{\tau}\mathcal{D})^{-1}\big((1-\tau)\mathcal{D} + \mathcal{U}^*\big).
\end{aligned}$$

From here, we get the required expression for $\mathcal{W}$ in (39). □

Given two error tolerance vectors $\boldsymbol{\delta}$ and $\boldsymbol{\delta}'$ with $\delta_1 = \delta_1'$, let

$$\Delta_{\rm sSOR}(\boldsymbol{\delta}', \boldsymbol{\delta}) := \boldsymbol{\delta}' + (\tau\mathcal{D} + \mathcal{U})(\rho\mathcal{D})^{-1}(\boldsymbol{\delta} - \boldsymbol{\delta}').$$

Given $\bar{\boldsymbol{x}} \in \mathcal{X}$, similar to Theorem 1, one can prove without much difficulty that the optimal solution of the following minimization problem

$$\min_{\boldsymbol{x}\in\mathcal{X}} \Big\{ p(x_1) + q(\boldsymbol{x}) + \frac{1}{2}\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|^2_{\mathcal{T}_{\rm sSOR}} - \langle \boldsymbol{x},\, \Delta_{\rm sSOR}(\boldsymbol{\delta}', \boldsymbol{\delta})\rangle \Big\}, \tag{40}$$

can be obtained by performing one cycle of the block sSOR method. In particular, when $p(\cdot) \equiv 0$ and $\boldsymbol{\delta} = \boldsymbol{\delta}' = \boldsymbol{0}$, the optimal solution to (40) can be computed by (38), i.e., set $\bar{\boldsymbol{x}} = \boldsymbol{x}^k$, then $\boldsymbol{x}^{k+1}$ obtained from (38) is the optimal solution to (40). By replacing $\mathcal{T}_{\mathcal{Q}}$ and $\Delta(\cdot, \cdot)$ in Algorithm 1 with $\mathcal{T}_{\rm sSOR}$ and $\Delta_{\rm sSOR}(\cdot, \cdot)$, respectively, one can obtain a block sSOR based inexact proximal gradient method for (7) for which the convergence results presented in Proposition 2 still hold.

7 Conclusion

In this paper, we give an optimization interpretation that each cycle of the classical block sGS method is equivalent to solving the associated convex QP problem with an additional proximal term. This equivalence is fully characterized via our block sGS decomposition theorem. A factorization view of this theorem and its equivalence to the SCB reduction procedure are also established. The classical block sGS method, viewed from the optimization perspective via the block sGS decomposition theorem, is then extended to the inexact setting for solving a class of convex composite QP problems involving nonsmooth functions. Moreover, we are able to derive $O(1/k)$ and $O(1/k^2)$ iteration complexities for our inexact block sGS method and its accelerated version, respectively. These new interpretations and convergence results, together with the incorporation of the (inexact) sGS decomposition techniques in the design of efficient algorithms for core optimization problems, demonstrate the power and usefulness of our simple yet elegant block sGS decomposition theorem. We believe this decomposition theorem will be proven to be even more versatile in solving various optimization problems and beyond.

Appendix: Proof of part (b) of Proposition 2

To begin the proof, we state the following lemma from [21].

Lemma 2. *Suppose that* $\{u_k\}$ *and* $\{\lambda_k\}$ *are two sequences of nonnegative scalars, and* $\{s_k\}$ *is a nondecreasing sequence of scalars such that* $s_0 \geq u_0^2$*. Suppose that for all* $k \geq 1$*, the inequality* $u_k^2 \leq s_k + 2\sum_{i=1}^{k} \lambda_i u_i$ *holds. Then for all* $k \geq 1$*,* $u_k \leq \bar{\lambda}_k + \sqrt{s_k + \bar{\lambda}_k^2}$*, where* $\bar{\lambda}_k = \sum_{i=1}^{k} \lambda_i$*.*

Proof. In this proof, we let $\Delta^j = \Delta(\tilde{\boldsymbol{\delta}}^j, \boldsymbol{\delta}^j)$. Note that under the assumption that $t_j = 1$ for all $j \geq 1$, $\widetilde{\boldsymbol{x}}^j = \boldsymbol{x}^{j-1}$. Note also that from (30), we have that $\|\widehat{\mathcal{Q}}^{-1/2}\Delta^j\| \leq M\epsilon_j$, where M is given as in Proposition 2.

From the optimality of $\boldsymbol{x}^j$ in (29), one can show that

$$F(\boldsymbol{x}) - F(\boldsymbol{x}^j) \geq \frac{1}{2}\|\boldsymbol{x}^j - \boldsymbol{x}^{j-1}\|_{\widehat{\mathcal{Q}}}^2 + \langle \boldsymbol{x}^{j-1} - \boldsymbol{x},\, \widehat{\mathcal{Q}}(\boldsymbol{x}^j - \boldsymbol{x}^{j-1})\rangle + \langle \Delta^j,\, \boldsymbol{x} - \boldsymbol{x}^j\rangle \quad \forall\, \boldsymbol{x}. \tag{41}$$

Let $\boldsymbol{e}^j = \boldsymbol{x}^j - \boldsymbol{x}^*$. By setting $\boldsymbol{x} = \boldsymbol{x}^{j-1}$ and $\boldsymbol{x} = \boldsymbol{x}^*$ in (41), we get

$$F(\boldsymbol{x}^{j-1}) - F(\boldsymbol{x}^j) \;\geq\; \frac{1}{2}\|\boldsymbol{e}^j - \boldsymbol{e}^{j-1}\|_{\widehat{\mathcal{Q}}}^2 + \langle \Delta^j,\, \boldsymbol{e}^{j-1} - \boldsymbol{e}^j\rangle, \tag{42}$$

$$F(\boldsymbol{x}^*) - F(\boldsymbol{x}^j) \;\geq\; \frac{1}{2}\|\boldsymbol{e}^j\|_{\widehat{\mathcal{Q}}}^2 - \frac{1}{2}\|\boldsymbol{e}^{j-1}\|_{\widehat{\mathcal{Q}}}^2 - \langle \Delta^j,\, \boldsymbol{e}^j\rangle. \tag{43}$$

By multiplying $j-1$ to (42) and combining with (43), we get

$$\begin{aligned}
(a_j + b_j^2) &\leq (a_{j-1} + b_{j-1}) - (j-1)\|\boldsymbol{e}^j - \boldsymbol{e}^{j-1}\|_{\widehat{\mathcal{Q}}}^2 + 2\langle \Delta^j,\, j\boldsymbol{e}^j - (j-1)\boldsymbol{e}^{j-1}\rangle \\
&\leq (a_{j-1} + b_{j-1}^2) + 2\|\widehat{\mathcal{Q}}^{-1/2}\Delta^j\|\|j\boldsymbol{e}^j - (j-1)\boldsymbol{e}^{j-1}\|_{\widehat{\mathcal{Q}}} \\
&\leq (a_{j-1} + b_{j-1}^2) + 2\|\widehat{\mathcal{Q}}^{-1/2}\Delta^j\|(jb_j + (j-1)b_{j-1}) \\
&\leq \cdots \\
&\leq a_1 + b_1^2 + 2\sum_{i=2}^{j} M\epsilon_i(ib_i + (i-1)b_{i-1}) \;\leq\; b_0^2 + 2\sum_{i=1}^{j} 2Mi\epsilon_i b_i,
\end{aligned} \tag{44}$$

where $a_j = 2j[F(\boldsymbol{x}^j) - F(\boldsymbol{x}^*)]$ and $b_j = \|\boldsymbol{e}^j\|_{\widehat{\mathcal{Q}}}$. Note that the last inequality follows from (43) with $j = 1$ and some simple manipulations. To summarize, we have $b_j^2 \leq b_0^2 + 2\sum_{i=1}^{j} 2M\epsilon_i i b_i$. By applying Lemma 2, we get

$$b_j \;\leq\; \bar{\lambda}_j + \sqrt{b_0^2 + \bar{\lambda}_j^2} \;\leq\; b_0 + 2\bar{\lambda}_j,$$

where $\bar{\lambda}_j = \sum_{i=1}^{j} \lambda_i$ with $\lambda_i = 2Mi\epsilon_i$. Applying the above result to (44), we get

$$a_j \;\leq\; b_0^2 + 2\sum_{i=1}^{j} \lambda_i(2\bar{\lambda}_i + b_0) \;\leq\; (b_0 + 2\bar{\lambda}_j)^2.$$

From here, the required result in Part (b) of Proposition 2 follows. □

References

[1] O. Axelsson, *Iterative Solution Methods*, Cambridge University Press, 1994.

[2] A. Beck and L. Tetruashvili, *On the convergence of block coordinate descent type methods*, SIAM Journal on Optimization, 23 (2013), pp. 2037–2060.

[3] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed., Athena Scientific, Belmont, Massachusetts, 1995.

[4] L. Chen, D. F. Sun, and K.-C. Toh, *An efficient inexact symmetric Gauss-Seidel based majorized ADMM for high-dimensional convex composite conic programming*, Mathematical Programming, 161 (2017), pp. 237–270.

[5] A. Greenbaum, *Iterative Methods for Solving Linear Systems*, SIAM, Philadelphia, 1997.

[6] B. Kristian and H. P. Sun, *Preconditioned Douglas–Rachford splitting methods for convex-concave saddle-point problems*, SIAM Journal on Numerical Analysis, 53 (2015), pp. 421–444.

[7] B. Kristian and H. P. Sun, *Preconditioned Douglas–Rachford algorithms for TV-and TGV-regularized variational imaging problems*, Journal of Mathematical Imaging and Vision, 52 (2015), pp. 317–344.

[8] R. W. Freund, *Preconditioning of symmetric, but highly indefinite linear systems*, In Proceedings of 15th IMACS World Congress on Scientific Computation Modelling and Applied Mathematics, Berlin, Germany, 1997, pp. 551–556.

[9] L. Grippo and M. Sciandrone, *On the convergence of the block nonlinear Gauss–Seidel method under convex constraints*, Operations Research Letters, 26 (2000), pp. 127–136.

[10] W. Hackbusch, *Iterative Solutions of Large Sparse Systems of Equations*, Springer-Verlag, New York, 1994.

[11] K. F. Jiang, D. F. Sun, and K.-C. Toh, *An inexact accelerated proximal gradient method for large scale linearly constrained convex SDP*, SIAM Journal on Optimization, 22 (2012), pp. 1042–1064.

[12] X. Y. Lam, J. S. Marron, D. F. Sun, and K.-C. Toh, *Fast algorithms for large scale extended distance weighted discrimination*, arXiv:1604.05473, 2016.

[13] X. D. Li, D. F. Sun, and K.-C. Toh, *QSDPNAL: A two-phase augmented Lagrangian method for convex quadratic semidefinite programming*, arXiv:1512.08872, 2015.

[14] X. D. Li, D. F. Sun, and K.-C. Toh, *A Schur complement based semi-proximal ADMM for convex quadratic conic programming and extensions*, Mathematical Programming, 155 (2016), pp. 333–373.

[15] X. D. Li, *A Two-Phase Augmented Lagrangian Method for Convex Composite Quadratic Programming*, PhD thesis, Department of Mathematics, National University of Singapore, 2015.

[16] Z.-Q. Luo and P. Tseng, *On the linear convergence of descent methods for convex essentially smooth minimization*, SIAM Journal on Control and Optimization, 30 (1992), pp. 408–425.

[17] Z.-Q. Luo and P. Tseng, *Error bounds and convergence analysis of feasible descent methods: A general approach*, Annals of Operations Research, 46 (1993), pp. 157–178.

[18] R. T. Rockafellar, Convex Analysis, Princeton University Press, Princeton, N.J., 1970.

[19] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, SIAM, Philadelphia, 2000.

[20] Y. Sadd, *Iterative Methods for Sparse Linear Systems*, SIAM, Philadelphia, 2003.

[21] M. Schmidt, N. Le Roux, and F. Bach, *Convergence rates of inexact proximal-gradient methods for convex optimization*, Advances in Neural Information Processing Systems (NIPS), 2011.

[22] D. F. Sun, K.-C. Toh, and L. Q. Yang, *An efficient inexact ABCD method for least squares semidefinite programming*, SIAM Journal on Optimization, 26 (2016), pp. 1072–1100.

[23] P. Tseng, *Convergence of a block coordinate descent method for nondifferentiable minimization*, Journal on Optimization Theory and Applications, 109 (2001), pp. 475–494.

[24] P. Tseng and S. Yun, *A coordinate gradient descent method for nonsmooth separable minimization*, Mathematical Programming, 125 (2010), pp. 387–423.

[25] B. Wen, X. Chen, and T. K. Pong, *Linear convergence of proximal gradient algorithm with extrapolation for a class of nonconvex nonsmooth minimization problems*, SIAM Journal on Optimization, 27 (2017), pp. 124–145.

[26] Z. Z. Zhou and A. M. C. So, *A unified approach to error bounds for structured convex optimization problems*, arXiv:1512.03518, 2015. To appear in Mathematical Programming.