CrossMark

Randomized Primal–Dual Proximal Block Coordinate Updates

Xiang Gao[1] · **Yang-Yang Xu**[2] · **Shu-Zhong Zhang**[1]

Abstract

In this paper, we propose a randomized primal–dual proximal block coordinate updating framework for a general multi-block convex optimization model with coupled objective function and linear constraints. Assuming mere convexity, we establish its $O(1/t)$ convergence rate in terms of the objective value and feasibility measure. The framework includes several existing algorithms as special cases such as a primal–dual method for bilinear saddle-point problems (PD-S), the proximal Jacobian alternating direction method of multipliers (Prox-JADMM) and a randomized variant of the ADMM for multi-block convex optimization. Our analysis recovers and/or strengthens the convergence properties of several existing algorithms. For example, for PD-S our result leads to the same order of convergence rate without the previously assumed boundedness condition on the constraint sets, and for Prox-JADMM the new result provides convergence rate in terms of the objective value and the feasibility violation. It is well known that the original ADMM may fail to converge when the number of blocks exceeds two. Our result shows that if an appropriate randomization procedure is invoked to select the updating blocks, then a sublinear rate of convergence in expectation can be guaranteed for multi-block ADMM, without assuming any strong convexity. The new approach is also extended to solve problems where only a stochastic approximation of the subgradient of the objective is available, and we establish an

✉ Yang-Yang Xu
xuy21@rpi.edu

Xiang Gao
gaoxx460@umn.edu

Shu-Zhong Zhang
zhangs@umn.edu

1   Department of Industrial and Systems Engineering, University of Minnesota, Minneapolis, USA

2   Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, USA

$O(1/\sqrt{t})$ convergence rate of the extended approach for solving stochastic programming.

1 Introduction

In this paper, we consider the following multi-block structured convex optimization model

$$
\min_{x,y} \quad f(x_1, \cdots, x_N) + \sum_{i=1}^{N} u_i(x_i) + g(y_1, \cdots, y_M) + \sum_{j=1}^{M} v_j(y_j)
$$

$$
\text{s.t.} \sum_{i=1}^{N} A_i x_i + \sum_{j=1}^{M} B_j y_j = b, \tag{1.1}
$$

$$
x_i \in \mathcal{X}_i, \ i = 1, \cdots, N; \ y_j \in \mathcal{Y}_j, \ j = 1, \cdots, M,
$$

where the variables $x = (x_1; \cdots; x_N)$ and $y = (y_1; \cdots; y_M)$ are naturally partitioned into N and M blocks, respectively, $A = (A_1, \cdots, A_N)$ and $B = (B_1, \cdots, B_M)$ are block matrices, $\mathcal{X}_i$ and $\mathcal{Y}_j$ are some closed convex sets, f and g are smooth convex functions, and u_i and v_j are proper closed convex (possibly nonsmooth) functions.

1.1 Motivating Examples

Optimization problems in the form of (1.1) have many emerging applications from various fields. For example, the constrained lasso (classo) problem that was first studied by James et al. [1] as a generalization of the lasso problem can be formulated as

$$
\min_{x} \frac{1}{2} \| Ax - b \|_2^2 + \tau \| x \|_1
$$
$$
\text{s.t.} \ Cx \leqslant d, \tag{1.2}
$$

where $A \in \mathbb{R}^{m \times p}$, $b \in \mathbb{R}^m$ are the observed data, and $C \in \mathbb{R}^{n \times p}$, $d \in \mathbb{R}^n$ are the predefined data matrix and vector. Many widely used statistical models can be viewed as special cases of (1.2), including the monotone curve estimation, fused lasso, generalized lasso, and so on [1]. By partitioning the variable x into blocks as $x = (x_1; \cdots; x_K)$ where $x_i \in \mathbb{R}^{p_i}$ as well as other matrices and vectors in (1.2) correspondingly, and introducing another slack variable y, the classo problem can be

transformed to

$$\min_{x,y} \frac{1}{2} \left\| \sum_{i=1}^{K} A_i x_i - b \right\|_2^2 + \tau \sum_{i=1}^{K} \|x_i\|_1$$
$$\text{s.t.} \sum_{i=1}^{K} C_i x_i + y = d, \ y \geqslant 0, \tag{1.3}$$

which is in the form of (1.1).

Another interesting example is the extended linear-quadratic programming [2] that can be formulated as

$$\min_{x} \frac{1}{2} x^{\top} P x + a^{\top} x + \max_{s \in \mathcal{S}} \left\{ (d - Cx)^{\top} s - \frac{1}{2} s^{\top} Q s \right\}$$
$$\text{s.t.} \ Ax \leqslant b \tag{1.4}$$

where P and Q are symmetric positive semidefinite matrices, and $\mathcal{S}$ is a polyhedral set. Apparently, (1.4) includes quadratic programming as a special case. In general, its objective is a piece-wise linear-quadratic convex function. Let $g(s) = \frac{1}{2} s^{\top} Q s + \iota_{\mathcal{S}}(s)$, where $\iota_{\mathcal{S}}$ denotes the indicator function of $\mathcal{S}$. Then,

$$\max_{s \in \mathcal{S}} \left\{ (d - Cx)^{\top} s - \frac{1}{2} s^{\top} Q s \right\} = g^*(d - Cx),$$

where g^* denotes the convex conjugate of g. Replacing $d - Cx$ by y and introducing slack variable z, we can equivalently write (1.4) into the form of (1.1):

$$\min_{x,y,z} \frac{1}{2} x^{\top} P x + a^{\top} x + g^*(y)$$
$$\text{s.t.} \ Ax + z = b, \ z \geqslant 0, \ Cx + y = d, \tag{1.5}$$

for which one can further partition the x-variable into a number of disjoint blocks.

Many other interesting applications in various areas can be formulated as optimization problems in the form of (1.1), including those arising from signal processing, image processing, machine learning, and statistical learning; see [3–6] and the references therein.

Finally, we mention that computing a point on the central path for a generic convex programming in block variables $(x_1; \cdots ; x_N)$:

$$\min_{x} \ f(x_1, \cdots, x_N)$$
$$\text{s.t.} \ \sum_{i=1}^{N} A_i x_i \leqslant b, \ x_i \geqslant 0, \ i = 1, 2, \cdots, N$$

boils down to

$$\min_{x,y} \ f(x_1, \cdots, x_N) - \mu e^\top \ln x - \mu e^\top \ln y$$
$$\text{s.t.} \ \sum_{i=1}^{N} A_i x_i + y = b,$$

where $\mu > 0$ and $e^\top \ln v$ indicates the sum of the logarithm of all the components of v. This model is again in the form of (1.1).

1.2 Related Works in the Literature

Our work relates to two recently very popular topics: the alternating direction method of multipliers (ADMM) for multi-block structured problems and the first-order primal–dual method for bilinear saddle-point problems. Below, we review the two methods and their convergence results. More complete discussion on their connections to our method will be provided after presenting our algorithm.

Multi-block ADMM and its Variants

One well-known approach for solving a linear constrained problem in the form of (1.1) is the augmented Lagrangian method, which iteratively updates the primal variable (x, y) by minimizing the augmented Lagrangian function in (1.7) and then the multiplier λ through dual gradient ascent. However, since the linear constraint couples $x_1, \cdots, x_N$ and $y_1, \cdots, y_M$ all together, it can be very expensive to minimize the augmented Lagrangian function simultaneously with respect to all block variables. Utilizing the multi-block structure of the problem, the multi-block ADMM updates the block variables sequentially, one at a time with the others fixed to their most recent values, followed by the update of multiplier. Specifically, it performs the following updates iteratively (by assuming the absence of the coupled functions f and g):

$$\begin{cases} x_1^{k+1} = \arg\min_{x_1 \in \mathcal{X}_1} \mathcal{L}_\rho(x_1, x_2^k, \cdots, x_N^k, y^k, \lambda^k), \\ \quad \vdots \\ x_N^{k+1} = \arg\min_{x_N \in \mathcal{X}_N} \mathcal{L}_\rho(x_1^{k+1}, \cdots, x_{N-1}^{k+1}, x_N, y^k, \lambda^k), \\ y_1^{k+1} = \arg\min_{y_1 \in \mathcal{Y}_1} \mathcal{L}_\rho(x^{k+1}, y_1, y_2^k, \cdots, y_M^k, \lambda^k), \\ \quad \vdots \\ y_M^{k+1} = \arg\min_{y_M \in \mathcal{Y}_M} \mathcal{L}_\rho(x^{k+1}, y_1^{k+1}, \cdots, y_{M-1}^{k+1}, y_M, \lambda^k), \\ \lambda^{k+1} = \lambda^k - \rho(Ax^{k+1} + By^{k+1} - b). \end{cases} \quad (1.6)$$

In the above updates, $\mathcal{L}_\rho$ is the augmented Lagrangian function, defined as:

$$\mathcal{L}_\rho(x, y, \lambda) = \sum_{i=1}^{N} u_i(x_i) + \sum_{j=1}^{M} v_j(y_j) - \lambda^\top (Ax + By - b) + \frac{\rho}{2} \|Ax + By - b\|^2, \quad (1.7)$$

where ρ is the penalty parameter and λ is the Lagrangian multiplier.

When there are only two blocks, i.e., $N = M = 1$, the update scheme in (1.6) reduces to the classic 2-block ADMM [7,8]. The convergence properties of the ADMM for solving 2-block separable convex problems have been studied extensively. Since the 2-block ADMM can be viewed as a manifestation of some kinds of operator splitting, its convergence follows from that of the so-called Douglas–Rachford operator splitting method; see [9,10]. Moreover, the convergence rate of the 2-block ADMM has been established recently by many authors; see, e.g., [11–16].

Although the multi-block ADMM scheme in (1.6) performs very well for many instances encountered in practice (e.g., [17,18]), it may fail to converge for some instances if there are more than 2 block variables, i.e., $N + M \geqslant 3$. In particular, an example was presented in [19] to show that the ADMM may even diverge with 3 blocks of variables, when solving a linear system of equations. Thus, some additional assumptions or modifications will have to be in place to ensure convergence of the multi-block ADMM. In fact, by incorporating some extra correction steps or changing the Gauss–Seidel updating rule, [20–24] showed that the convergence can still be achieved for the multi-block ADMM. Moreover, if some parts of the objective function are strongly convex or the objective has certain regularity property, then it can be shown that the convergence holds under various conditions; see [14,25–30]. Using some other conditions including the error bound condition and taking small dual stepsizes, or by adding some perturbations to the original problem, authors of [15,31] established the rate of convergence results even without strong convexity. Not only for the problem with linear constraint, in [32–34] multi-block ADMM were extended to solve convex linear/quadratic conic programming problems. In a very recent work [35], Sun, Luo and Ye proposed a randomly permuted ADMM (RP-ADMM) that basically chose a random permutation of the block indices and performs the ADMM update according to the order of indices in that permutation, and they showed that the RP-ADMM converges in expectation for solving non-singular square linear system of equations.

In [3], the authors proposed a block successive upper-bound minimization method of multipliers (BSUMM) to solve problem (1.1) without y-variable. Essentially, at every iteration, the BSUMM replaced the nonseparable part $f(x)$ by an upper-bound function and works on that modified function in an ADMM manner. Under some error bound conditions and a diminishing dual stepsize assumption, the authors are able to show that the iterates produced by the BSUMM algorithm converge to the set of primal–dual optimal solutions. Along a similar direction, Cui et al. [4] introduced a quadratic upper-bound function for the nonseparable function f to solve 2-block problems; they showed that their algorithm has an $O(1/t)$ convergence rate, where t is the number of total iterations. Very recently, [6] proposed a set of variants of the ADMM by adding some proximal terms into the algorithm; the authors have managed to prove $O(1/t)$ convergence rate for the 2-block case, and the same results applied for general multi-block case under some strong convexity assumptions. Moreover, [5] showed the convergence of the ADMM for 2-block problems by imposing quadratic structure on the coupled function $f(x)$ and also the convergence of RP-ADMM for multi-block case where all separable functions vanish (i.e., $u_i(x_i) = 0$, $\forall i$).

Primal–Dual Method for Bilinear Saddle-Point Problems

Recently, the work [36] generalized the first-order primal–dual method in [37] to a randomized method for solving a class of saddle-point problems in the following form:

$$\min_{z \in \mathcal{Z}} \left\{ h(z) + \max_{x \in \mathcal{X}} \left\langle z, \sum_{i=1}^{N} A_i x_i \right\rangle - \sum_{i=1}^{N} u_i(x_i) \right\}, \tag{1.8}$$

where $x = (x_1; \cdots; x_N)$ and $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_N$. Let $\mathcal{Z} = \mathbb{R}^p$ and $h(z) = -b^\top z$. Then, it is easy to see that (1.8) is a saddle-point reformulation of the multi-block structured optimization problem

$$\min_{x \in \mathcal{X}} \sum_{i=1}^{N} u_i(x_i), \text{ s.t. } \sum_{i=1}^{N} A_i x_i = b,$$

which is a special case of (1.1) without y-variable or the coupled function f.

At each iteration, the algorithm in [36] chose one block of x-variable uniformly at random and performs a proximal update to it, followed by another proximal update to the z-variable. More precisely, it iteratively performs the updates:

$$x_i^{k+1} = \begin{cases} \arg\min_{x_i \in \mathcal{X}_i} \langle -\bar{z}^k, A_i x_i \rangle + u_i(x_i) + \frac{\tau}{2} \|x_i - x_i^k\|_2^2, & \text{if } i = i_k, \\ x_i^k, & \text{if } i \neq i_k, \end{cases} \tag{1.9a}$$

$$z^{k+1} = \arg\min_{z \in \mathcal{Z}} h(z) + \langle z, A x^{k+1} \rangle + \frac{\eta}{2} \|z - z^k\|_2^2, \tag{1.9b}$$

$$\bar{z}^{k+1} = q(z^{k+1} - z^k) + z^{k+1}, \tag{1.9c}$$

where i_k is a randomly selected block, and τ, η and q are certain parameters.[1] When there is only one block of x-variable, i.e., $N = 1$, the scheme in (1.9) becomes exactly the primal–dual method in [37]. Assuming the boundedness of the constraint sets $\mathcal{X}$ and $\mathcal{Z}$, [36] showed that under convexity, $O(1/t)$ convergence rate result of the scheme can be established by choosing appropriate parameters, and if u_i are all strongly convex, the scheme can be accelerated to have $O(1/t^2)$ convergence rate by adapting the parameters.

1.3 Contributions and Organization of this Paper

- We propose a randomized primal–dual coordinate update algorithm to solve problems in the form of (1.1). The key feature is to introduce randomization as done in (1.9) to the multi-block ADMM framework (1.6). Unlike the random permutation scheme as previously investigated in [5,35], we simply choose a subset of blocks of variables based on the uniform distribution. In addition, we perform a proximal

---

[1] Actually, [36] presented its algorithm in a more general way with the parameters adaptive to the iteration. However, its convergence result assumes constant values of these parameters for the convex case.

update to that selected subset of variables. With appropriate proximal terms [e.g., the setting in (2.15)], the selected block variables can be decoupled, and thus the updates can be done in parallel.

- More general than (1.6), we can accommodate coupled terms in the objective function in our algorithm by linearizing such terms. By imposing Lipschitz continuity condition on the partial gradient of the coupled functions f and g and using proximal terms, we show that our method has an expected $O(1/t)$ convergence rate for solving problem (1.1) under mere convexity assumption.

- We show that our algorithm includes several existing methods as special cases such as the scheme in (1.9) and the proximal Jacobian ADMM in [20]. Our result indicates that the $O(1/t)$ convergence rate of the scheme in (1.9) can be shown without assuming boundedness of the constraint sets. In addition, the same order of convergence rate of the proximal Jacobian ADMM can be established in terms of a better measure.

- Furthermore, the linearization scheme allows us to deal with stochastic objective function, for instance, when the function f is given in a form of expectation $f = E_\xi[f_\xi(x)]$ where ξ is a random vector. As long as an unbiased estimator of the subgradient of f is available, we can extend our method to the stochastic problem and an expected $O(1/\sqrt{t})$ convergence rate is achievable.

The rest of the paper is organized as follows. In Sect. 2, we introduce our algorithm and present some preliminary results. In Sect. 3, we present the sublinear convergence rate results of the proposed algorithm. Depending on the multi-block structure of y, different conditions and parameter settings are presented in Sects. 3.1, 3.2 and 3.3, respectively. In Sect. 4, we present an extension of our algorithm where we assume the availability of only first-order stochastic approximation of the objective function. The convergence analysis is extended to such settings accordingly. Numerical results are shown in Sect. 5. In Sect. 6, we discuss the connections of our algorithm to other well-known methods in the literature. Finally, we conclude the paper in Sect. 7. The proofs for the technical lemmas are presented in Appendix A, and the proofs for the main theorems are in Appendix B.

2 Randomized Primal–Dual Block Coordinate Update Algorithm

In this section, we first present some notations and then introduce our algorithm as well as some preliminary lemmas.

2.1 Notations

We denote $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_N$ and $\mathcal{Y} = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_M$. For any symmetric positive semidefinite matrix W, we define $\|z\|_W = \sqrt{z^\top W z}$. Given an integer $\ell > 0$, $[\ell]$ denotes the set $\{1, 2, \cdots, \ell\}$. We use I and J as index sets, while I is also used to denote the identity matrix; we believe that the intention is evident in the context. Given $I = \{i_1, i_2, \cdots, i_n\}$, we denote:

- Block-indexed variable: $x_I = (x_{i_1}; x_{i_2}; \cdots; x_{i_n})$;
- Block-indexed set: $\mathcal{X}_I = \mathcal{X}_{i_1} \times \cdots \times \mathcal{X}_{i_n}$;
- Block-indexed function: $u_I(x_I) = u_{i_1}(x_{i_1}) + u_{i_2}(x_{i_2}) + \cdots + u_{i_n}(x_{i_n})$;
- Block-indexed gradient: $\nabla_I f(x) = (\nabla_{i_1} f(x); \nabla_{i_2} f(x); \cdots; \nabla_{i_n} f(x))$;
- Block-indexed matrix: $A_I = [A_{i_1}, A_{i_2}, \cdots, A_{i_n}]$.

2.2 Algorithm

Our algorithm is rather general. Its major ingredients are randomization in selecting block variables, linearization of the coupled functions f and g, and adding proximal terms. Specifically, at each iteration k, it first randomly samples a subset I_k of blocks of x, and then a subset J_k of blocks of y according to the uniform distribution over the indices. The randomized sampling rule is as follows:

Randomization Rule (U): For the given integers $n \leqslant N$ and $m \leqslant M$, it randomly chooses index sets $I_k \subset [N]$ with $|I_k| = n$ and $J_k \subset [M]$ with $|J_k| = m$ uniformly; i.e., for any subsets $\{i_1, i_2, \cdots, i_n\} \subset [N]$ and $\{j_1, j_2, \cdots, j_m\} \subset [M]$, the following holds

$$\text{Prob}[I_k = \{i_1, i_2, \cdots, i_n\}] = 1 / \binom{N}{n},$$

$$\text{Prob}[J_k = \{j_1, j_2, \cdots, j_m\}] = 1 / \binom{M}{m}.$$

After those subsets have been selected, it performs a prox-linear update to those selected blocks based on the augmented Lagrangian function, followed by an update of the Lagrangian multiplier. The details of the method are given in Algorithm 1.

In Algorithm 1, P^k and Q^k are predetermined positive semidefinite matrices with appropriate dimensions. For the selected blocks in I_k and J_k, instead of implementing the exact minimization of the augmented Lagrangian function, we perform a block proximal gradient update. In particular, before minimization, we first linearize the coupled functions f, g and add some proximal terms to it. Note that one can always select all blocks, i.e., $I_k = [N]$ and $J_k = [M]$. Empirically, however, the block coordinate update method usually outperforms the full coordinate update method if the problem possesses certain structures; see [38] for an example. To decouple the selected x blocks and also y blocks, we choose the matrices P^k and Q^k in Algorithm 1 as follows:

$$P^k = \hat{P}_{I_k} - \rho_x A_{I_k}^\top A_{I_k}, \qquad Q^k = \hat{Q}_{J_k} - \rho_y B_{J_k}^\top B_{J_k}, \tag{2.15}$$

where $\hat{P}$ and $\hat{Q}$ are symmetric positive semidefinite and block diagonal matrices, $\hat{P}_{I_k}$ denotes the diagonal blocks of $\hat{P}$ indexed by I_k. With such setting of P^k and Q^k, (2.10) and (2.12), respectively, become

$$x_I^{k+1} = \underset{x_I \in \mathcal{X}_I}{\arg\min} \langle \nabla_I f(x^k) - A_I^\top(\lambda^k - \rho_x r^k), x_I \rangle + u_I(x_I) + \frac{1}{2}\|x_I - x_I^k\|_{\hat{P}_I}^2, \tag{2.16}$$

Algorithm 1: Randomized **P**rimal-**D**ual **B**lock Coordinate Update (RPDBU) Method

Step 1 Initialization: choose x^0, y^0 and $\lambda^0 = 0$; let $r^0 = Ax^0 + By^0 - b$;
        choose ρ, ρ_x, ρ_y

Step 2 for $k = 0, 1, \cdots$ **do**

Step 3   Randomly select $I_k \subset [N]$ and $J_k \subset [M]$ with $|I_k| = n$ and
           $|J_k| = m$ according to (**U**).

Step 4   Let $x_i^{k+1} = x_i^k$, $\forall i \notin I_k$ and $y_j^{k+1} = y_j^k$, $\forall j \notin J_k$.

Step 5   For $I = I_k$, perform the update

$$x_I^{k+1} = \arg\min_{x_I \in \mathcal{X}_I} \langle \nabla_I f(x^k) - A_I^\top \lambda^k, x_I \rangle + u_I(x_I) \qquad (2.10)$$

$$+ \frac{\rho_x}{2} \|A_I(x_I - x_I^k) + r^k\|^2 + \frac{1}{2}\|x_I - x_I^k\|_{P_k}^2,$$

$$r^{k+\frac{1}{2}} = r^k + A_I(x_I^{k+1} - x_I^k). \qquad (2.11)$$

    For $J = J_k$, perform the update

$$y_J^{k+1} = \arg\min_{y_J \in \mathcal{Y}_J} \langle \nabla_J g(y^k) - B_J^\top \lambda^k, y_J \rangle + v_J(y_J) \qquad (2.12)$$

$$+ \frac{\rho_y}{2} \|B_J(y_J - y_J^k) + r^{k+\frac{1}{2}}\|^2 + \frac{1}{2}\|y_J - y_J^k\|_{Q_k}^2,$$

$$r^{k+1} = r^{k+\frac{1}{2}} + B_J(y_J^{k+1} - y_J^k). \qquad (2.13)$$

    Update the multiplier by

$$\lambda^{k+1} = \lambda^k - \rho r^{k+1}. \qquad (2.14)$$

Step 6 end

$$y_J^{k+1} = \arg\min_{y_J \in \mathcal{Y}_J} \langle \nabla_J g(y^k) - B_J^\top(\lambda^k - \rho_y r^{k+\frac{1}{2}}), y_J \rangle + v_J(y_J) + \frac{1}{2}\|y_J - y_J^k\|_{\hat{Q}_J}^2. \qquad (2.17)$$

Due to the block diagonal structure of $\hat{P}$ and $\hat{Q}$, both x-update and y-update can be computed in parallel.

2.3 Preliminaries

Let w be the aggregated primal–dual variables and $H(w)$ the primal–dual linear mapping; namely

$$w = \begin{pmatrix} x \\ y \\ \lambda \end{pmatrix}, \quad H(w) = \begin{pmatrix} -A^\top \lambda \\ -B^\top \lambda \\ Ax + By - b \end{pmatrix}, \qquad (2.18)$$

and also let

$$u(x) = \sum_{i=1}^{N} u_i(x_i), \quad v(y) = \sum_{j=1}^{M} v_j(y_j),$$

$$F(x) = f(x) + u(x), \quad G(y) = g(y) + v(y), \quad \Phi(x, y) = F(x) + G(y).$$

The point (x^*, y^*) is a solution to (1.1) if and only if there exists λ^* such that

$$\Phi(x, y) - \Phi(x^*, y^*) + (w - w^*)^\top H(w^*) \geqslant 0, \ \forall (x, y) \in \mathcal{X} \times \mathcal{Y}, \ \forall \lambda, \quad (2.19a)$$

$$Ax^* + By^* = b, \quad (2.19b)$$

$$x^* \in \mathcal{X}, \quad y^* \in \mathcal{Y}. \quad (2.19c)$$

The following lemmas will be used in our subsequent analysis, whose proofs are elementary and thus are omitted here.

Lemma 2.1 *For any two vectors w and $\tilde{w}$, it holds*

$$(w - \tilde{w})^\top H(w) = (w - \tilde{w})^\top H(\tilde{w}). \quad (2.20)$$

Lemma 2.2 *For any two vectors u, v and a positive semidefinite matrix W,*

$$u^\top W v = \frac{1}{2}\big(\|u\|_W^2 + \|v\|_W^2 - \|u - v\|_W^2\big). \quad (2.21)$$

Lemma 2.3 *For any nonzero positive semidefinite matrix W, it holds for any z and $\hat{z}$ of appropriate size that*

$$\|z - \hat{z}\|^2 \geqslant \frac{1}{\|W\|_2}\|z - \hat{z}\|_W^2, \quad (2.22)$$

where $\|W\|_2$ denotes the matrix operator norm of W.

The following lemma presents a useful property of $H(w)$, which essentially follows from (2.20).

Lemma 2.4 *For any vectors $w^0, w^1, \cdots, w^t$, and sequence of positive numbers $\beta^0, \beta^1, \cdots, \beta^t$, it holds that*

$$\left(\frac{\sum_{k=0}^{t} \beta^k w^k}{\sum_{k=0}^{t} \beta^k} - w\right)^\top H\left(\frac{\sum_{k=0}^{t} \beta^k w^k}{\sum_{k=0}^{t} \beta^k}\right) = \frac{1}{\sum_{k=0}^{t} \beta^k} \sum_{k=0}^{t} \beta^k (w^k - w)^\top H(w^k). \quad (2.23)$$

3 Convergence Rate Results

In this section, we establish sublinear convergence rate results of Algorithm 1 for three different cases. We differentiate those cases based on whether or not y in problem (1.1) also has the multi-block structure. The first case assumes no y-variable at all.

It falls into the second and third cases, and we discuss this case separately since it requires weaker conditions to guarantee the same convergence rate. The second case is that y is treated as a single-block variable, and this can be reflected in our algorithm by simply selecting all y-blocks every time, i.e., $m = M$. In the third case where y is a multi-block variable, it requires $\frac{n}{N} = \frac{m}{M}$ where n and m are the cardinalities of the subsets of x and y selected in our algorithm, respectively. Since the analysis only requires convexity, we can ensure the condition to hold by adding zero component functions if necessary, in such a way that $N = M$ and then choosing $n = m$. In particular, we make the following assumptions:

Assumption 3.1 (*Convexity*) For (1.1), $\mathcal{X}_i$ and $\mathcal{Y}_j$ are some closed convex sets, f and g are smooth convex functions, and u_i and v_j are proper closed convex function.

Assumption 3.2 (*Existence of a solution*) There is at least one point $w^* = (x^*, y^*, \lambda^*)$ satisfying the conditions in (2.19).

Assumption 3.3 (*Lipschitz continuous partial gradient*) There exist constants L_f and L_g such that for any subset I of $[N]$ with $|I| = n$ and any subset J of $[M]$ with $|J| = m$, it holds that

$$\|\nabla_I f(x + U_I \tilde{x}) - \nabla_I f(x)\| \leqslant L_f \|\tilde{x}_I\|, \quad \forall x, \tilde{x}, \tag{3.1a}$$

$$\|\nabla_J g(y + U_J \tilde{y}) - \nabla_J g(y)\| \leqslant L_g \|\tilde{y}_J\|, \quad \forall y, \tilde{y}, \tag{3.1b}$$

where $U_I \tilde{x}$ keeps the blocks of $\tilde{x}$ that are indexed by I and zero elsewhere.

Before presenting the main convergence rate result, we first establish a few key lemmas.

Lemma 3.1 (One-step analysis) *Let $\{(x^k, y^k, r^k, \lambda^k)\}$ be the sequence generated from Algorithm 1 with matrices P^k and Q^k defined as in (2.15). Then, the following inequalities hold*

$$\begin{aligned}
E_{I_k} &\Big[F(x^{k+1}) - F(x) + (x^{k+1} - x)^\top (-A^\top \lambda^{k+1}) \\
&+ (\rho_x - \rho)(x^{k+1} - x)^\top A^\top r^{k+1} - \rho_x (x^{k+1} - x)^\top A^\top B(y^{k+1} - y^k) \Big] \\
&+ E_{I_k} (x^{k+1} - x)^\top (\hat{P} - \rho_x A^\top A)(x^{k+1} - x^k) - \frac{L_f}{2} E_{I_k} \|x^k - x^{k+1}\|^2 \\
\leqslant &\left(1 - \frac{n}{N}\right) \big[F(x^k) - F(x) + (x^k - x)^\top (-A^\top \lambda^k) + \rho_x (x^k - x)^\top A^\top r^k \big],
\end{aligned} \tag{3.2}$$

and

$$\begin{aligned}
E_{J_k} &\big[G(y^{k+1}) - G(y) + (y^{k+1} - y)^\top (-B^\top \lambda^{k+1}) + (\rho_y - \rho)(y^{k+1} - y)^\top B^\top r^{k+1} \big] \\
&+ E_{J_k} (y^{k+1} - y)^\top (\hat{Q} - \rho_y B^\top B)(y^{k+1} - y^k) - \frac{L_g}{2} E_{J_k} \|y^k - y^{k+1}\|^2 \\
&- \left(1 - \frac{m}{M}\right) \rho_y (y^k - y)^\top B^\top A(x^{k+1} - x^k)
\end{aligned}$$

$$\leqslant \left(1 - \frac{m}{M}\right)\left[G(y^k) - G(y) + (y^k - y)^\top(-B^\top\lambda^k) + \rho_y(y^k - y)^\top B^\top r^k\right], \quad (3.3)$$

where E_{I_k} denotes expectation over I_k and conditional on all previous history.

Note that for any feasible point (x, y) (namely, $x \in \mathcal{X}$, $y \in \mathcal{Y}$ and $Ax + By = b$),

$$Ax^{k+1} - Ax = \frac{1}{\rho}(\lambda^k - \lambda^{k+1}) - (By^{k+1} - b) + (By - b)$$

$$= \frac{1}{\rho}(\lambda^k - \lambda^{k+1}) - B(y^{k+1} - y) \quad (3.4)$$

and

$$By^k - By = \frac{1}{\rho}(\lambda^{k-1} - \lambda^k) - (Ax^k - b) + (Ax - b)$$

$$= \frac{1}{\rho}(\lambda^{k-1} - \lambda^k) - A(x^k - x). \quad (3.5)$$

Hence,

$$(x^{k+1} - x)^\top A^\top B(y^{k+1} - y^k) = \frac{1}{\rho}(\lambda^k - \lambda^{k+1})^\top B(y^{k+1} - y^k)$$

$$-(y^{k+1} - y)^\top B^\top B(y^{k+1} - y^k),$$

and

$$(y^k - y)^\top B^\top A(x^{k+1} - x^k) = \frac{1}{\rho}(\lambda^{k-1} - \lambda^k)^\top A(x^{k+1} - x^k) - (x^k - x)^\top A^\top A(x^{k+1} - x^k).$$

Then, using (2.21) we have the following result.

Lemma 3.2 *For any feasible point (x, y) and integer t, it holds*

$$\sum_{k=0}^{t}(x^{k+1} - x)^\top A^\top B(y^{k+1} - y^k)$$

$$= \frac{1}{\rho}\sum_{k=0}^{t}(\lambda^k - \lambda^{k+1})^\top B(y^{k+1} - y^k)$$

$$-\frac{1}{2}\left(\|y^{t+1} - y\|_{B^\top B}^2 - \|y^0 - y\|_{B^\top B}^2 + \sum_{k=0}^{t}\|y^{k+1} - y^k\|_{B^\top B}^2\right) \quad (3.6)$$

and

$$\sum_{k=0}^{t}(y^k - y)^\top B^\top A(x^{k+1} - x^k)$$

$$= \frac{1}{\rho} \sum_{k=0}^{t} (\lambda^{k-1} - \lambda^k)^\top A(x^{k+1} - x^k)$$

$$+ \frac{1}{2} \left(\|x^0 - x\|_{A^\top A}^2 - \|x^{t+1} - x\|_{A^\top A}^2 + \sum_{k=0}^{t} \|x^{k+1} - x^k\|_{A^\top A}^2 \right). \quad (3.7)$$

Lemma 3.3 *Given a continuous function h, for a random vector $\hat{w} = (\hat{x}, \hat{y}, \hat{\lambda})$, if for any feasible point $w = (x, y, \lambda)$ that may depend on $\hat{w}$, we have*

$$E\big[\Phi(\hat{x}, \hat{y}) - \Phi(x, y) + (\hat{w} - w)^\top H(w)\big] \leqslant E[h(w)], \quad (3.8)$$

then for any $\gamma > 0$ and any optimal solution (x^*, y^*) to (1.1) we also have

$$E\big[\Phi(\hat{x}, \hat{y}) - \Phi(x^*, y^*) + \gamma \|A\hat{x} + B\hat{y} - b\|\big] \leqslant \sup_{\|\lambda\| \leqslant \gamma} h(x^*, y^*, \lambda).$$

Noting

$$\Phi(x, y) - \Phi(x^*, y^*) + (w - w^*)^\top H(w^*)$$
$$= \Phi(x, y) - \Phi(x^*, y^*) - (\lambda^*)^\top (Ax + By - b),$$

we can easily show the following lemma by the optimality of (x^*, y^*, λ^*) and the Cauchy-Schwarz inequality.

Lemma 3.4 *Assume (x^*, y^*, λ^*) satisfies (2.19). Then, for any point $(\hat{x}, \hat{y}) \in \mathcal{X} \times \mathcal{Y}$, we have*

$$\Phi(\hat{x}, \hat{y}) - \Phi(x^*, y^*) \geqslant -\|\lambda^*\| \cdot \|A\hat{x} + B\hat{y} - b\|. \quad (3.9)$$

The following lemma shows a connection between different convergence measures, and it can be simply proved by using (3.9). If both w and $\hat{w}$ are deterministic, it implies Lemma 2.4 in [39].

Lemma 3.5 *Assume that (x^*, y^*, λ^*) satisfies the optimality conditions in (2.19). Let $\gamma > \|\lambda^*\|$ and $\varepsilon \geqslant 0$. If a random vector $(\hat{x}, \hat{y})$ satisfies*

$$E\big[\Phi(\hat{x}, \hat{y}) - \Phi(x^*, y^*) + \gamma \|A\hat{x} + B\hat{y} - b\|\big] \leqslant \varepsilon,$$

then

$$E\|A\hat{x} + B\hat{y} - b\| \leqslant \frac{\varepsilon}{\gamma - \|\lambda^*\|},$$

$$E\big|\Phi(\hat{x}, \hat{y}) - \Phi(x^*, y^*)\big| \leqslant \left(\frac{2\|\lambda^*\|}{\gamma - \|\lambda^*\|} + 1 \right) \varepsilon.$$

The convergence analysis for Algorithm 1 requires slightly different parameter settings under different structures. In fact, the underlying analysis and results also differ. To account for the differences, we present in the next three subsections the corresponding convergence results. The first one assumes there is no y part at all; the second case assumes a single block on the y side; the last one deals with the general case where the ratios n/N are assumed to be equal to m/M.

3.1 Multiple *x* Blocks and No *y*-Variable

We first consider a special case with no y-variable, namely $g = v = 0$ and $B = 0$ in (1.1). This case has its own importance. It is a parallel block coordinate update version of the linearized augmented Lagrangian method (ALM).

Theorem 3.6 (Sublinear ergodic convergence I) *Assume* $g(y) = 0$, $v_j(y_j) = 0$, $\forall j$ *and* $B = 0$ *in* (1.1). *Let* $\{(x^k, y^k, \lambda^k)\}$ *be the sequence generated from Algorithm* 1 *with* $y^k \equiv y^0$. *Assume* $\frac{n}{N} = \theta$, $\rho = \theta \rho_x$, *and*

$$\hat{P}_{I_k} \succcurlyeq L_f I + \rho_x A_{I_k}^\top A_{I_k}, \ \forall k. \tag{3.10}$$

Let

$$\hat{x}^t = \frac{x^{t+1} + \theta \sum_{k=1}^{t} x^k}{1 + \theta t}. \tag{3.11}$$

Then, under Assumptions 3.1, 3.2, *and* 3.3, *we have*

$$\max \left\{ E \left| F(\hat{x}^t) - F(x^*) \right|, \ E \| A\hat{x}^t - b \| \right\}$$
$$\leqslant \frac{2}{1 + \theta t} \left[(1 - \theta) \left(F(x^0) - F(x^*) + \frac{\rho_x}{2} \| r^0 \|^2 \right) \right.$$
$$\left. + \frac{1}{2} \| x^0 - x^* \|_{\hat{P}}^2 + \frac{\max\{(0.5 + \| \lambda^* \|)^2, 9 \| \lambda^* \|^2\}}{2\rho_x} \right], \tag{3.12}$$

where (x^*, λ^*) *is an arbitrary primal–dual solution.*

Our result recovers the convergence of the proximal Jacobian ADMM introduced in [20]. In fact, the above theorem strengthens the convergence result in [20] by establishing an $O(1/t)$ rate of convergence in terms of the feasibility measure and the objective value. If strong convexity is assumed on the objective function, the algorithm can be accelerated to have the rate $O(1/t^2)$ as shown in [40].

3.2 Multiple *x* Blocks and a Single *y* Block

When the y-variable is simple to update, it could be beneficial to renew the whole of it at every iteration, such as the problem (1.3). In this subsection, we consider the case that there are multiple x-blocks but a single y-block (or equivalently, $m = M$), and

we establish a sublinear convergence rate result with a different technique of dealing with the y-variable.

Theorem 3.7 (Sublinear ergodic convergence II) *Let $\{(x^k, y^k, \lambda^k)\}$ be the sequence generated from Algorithm 1 with $m = M$ and $\rho = \rho_y = \theta\rho_x$, where $\theta = \frac{n}{N}$. Assume*

$$\hat{P} \succcurlyeq L_f I + \rho_x A^\top A, \qquad \hat{Q} \succcurlyeq \frac{L_g}{\theta} I + \left(\frac{\rho}{\theta^4} - \frac{\rho}{\theta^2} + \rho_y\right) B^\top B. \qquad (3.13)$$

Let

$$\hat{x}^t = \frac{x^{t+1} + \theta \sum_{k=1}^{t} x^k}{1 + \theta t}, \quad \hat{y}^t = \frac{\tilde{y}^{t+1} + \theta \sum_{k=1}^{t} y^k}{1 + \theta t}, \qquad (3.14)$$

where

$$\tilde{y}^{t+1} = \underset{y \in \mathcal{Y}}{\arg\min} \langle \nabla g(y^t) - B^\top \lambda^t, y \rangle + v(y) + \frac{\rho_x}{2} \|Ax^{t+1} + By - b\|^2$$
$$+ \frac{\theta}{2} \|y - y^t\|_{\hat{Q} - \rho_y B^\top B}^2. \qquad (3.15)$$

Then, under Assumptions 3.1, 3.2, and 3.3, we have

$$\max\left\{ E\big|\Phi(\hat{x}^t, \hat{y}^t) - \Phi(x^*, y^*)\big|, \ E\|A\hat{x}^t + B\hat{y}^t - b\| \right\}$$
$$\leqslant \frac{2}{1 + \theta t}\left[(1 - \theta)\left(\Phi(x^0, y^0) - \Phi(x^*, y^*) + \frac{\rho_x}{2}\|r^0\|^2\right) + \frac{1}{2}\|x^0 - x^*\|_{\hat{P}}^2 \right.$$
$$\left. + \frac{1}{2}\|y^0 - y^*\|_{\theta\hat{Q} + (\rho_x - \theta\rho_y)B^\top B}^2 + \frac{\max\{(0.5 + \|\lambda^*\|)^2, 9\|\lambda^*\|^2\}}{2\rho_x} \right], \qquad (3.16)$$

where (x^*, y^*, λ^*) is an arbitrary primal–dual solution.

Remark 3.1 It is easy to see that if $\theta = 1$, the result in Theorem 3.7 becomes exactly the same as that in Theorem 3.8. In general, they are different because the conditions in (3.13) on $\hat{P}$ and $\hat{Q}$ are different from those in (3.18a).

3.3 Multiple *x* and *y* Blocks

In this subsection, we consider the most general case where both x and y have multi-block structure. Assuming $\frac{n}{N} = \frac{m}{M}$, we can still have the $O(1/t)$ convergence rate. The assumption can be made without losing generality, e.g., by adding zero components if necessary (which is essentially equivalent to varying the probabilities of the variable selection).

Theorem 3.8 (Sublinear ergodic convergence III) *Let $\{(x^k, y^k, \lambda^k)\}$ be the sequence generated from Algorithm 1 with the parameters satisfying*

$$\rho = \frac{n\rho_x}{N} = \frac{m\rho_y}{M} > 0. \qquad (3.17)$$

Assume $\frac{n}{N} = \frac{m}{M} = \theta$, and $\hat{P}$, $\hat{Q}$ satisfy one of the following conditions:

$$\hat{P} \succcurlyeq (2 - \theta) \left(\frac{1 - \theta}{\theta^2} + 1 \right) \rho_x A^\top A + L_f I, \qquad \hat{Q} \succcurlyeq \frac{(2 - \theta)}{\theta^2} \rho_y B^\top B + L_g I, \quad (3.18a)$$

$$\hat{P}_i \succcurlyeq (2 - \theta) \left(\frac{1 - \theta}{\theta^2} + 1 \right) n \rho_x A_i^\top A_i + L_f I, \, \forall i,$$

$$\hat{Q}_j \succcurlyeq \frac{(2 - \theta)}{\theta^2} m \rho_y B_j^\top B_j + L_g I, \, \forall j. \qquad (3.18b)$$

Let

$$\hat{x}^t = \frac{x^{t+1} + \theta \sum_{k=1}^{t} x^k}{1 + \theta t}, \quad \hat{y}^t = \frac{y^{t+1} + \theta \sum_{k=1}^{t} y^k}{1 + \theta t}. \qquad (3.19)$$

Then, under Assumptions 3.1, 3.2, and 3.3, we have

$$
\begin{aligned}
\max &\left\{ E \left| \Phi(\hat{x}^t, \hat{y}^t) - \Phi(x^*, y^*) \right|, \, E \| A\hat{x}^t + B\hat{y}^t - b \| \right\} \\
&\leqslant \frac{2}{1 + \theta t} \left[(1 - \theta) \left(\Phi(x^0, y^0) - \Phi(x^*, y^*) + \rho_x \| r^0 \|^2 \right) \right. \\
&\quad + \frac{1}{2} \| x^0 - x^* \|_{\hat{P} - \theta \rho_x A^\top A}^2 \\
&\quad \left. + \frac{1}{2} \| y^0 - y^* \|_{\hat{Q}}^2 + \frac{\max\{ (0.5 + \| \lambda^* \|)^2, 9 \| \lambda^* \|^2 \}}{2 \rho_x} \right],
\end{aligned} \qquad (3.20)
$$

where (x^*, y^*, λ^*) is an arbitrary primal–dual solution.

Remark 3.2 When $N = M = 1$, the two conditions in (3.18) become the same. However, in general, neither of the two conditions in (3.18) implies the other one. Roughly speaking, for the case of $n \approx N$ and $m \approx M$, the one in (3.18a) can be weaker, and for the case of $n \ll N$ and $m \ll M$, the one in (3.18b) is more likely weaker. In addition, (3.18b) provides an explicit way to choose block diagonal $\hat{P}$ and $\hat{Q}$ by simply setting $\hat{P}_i$ and $\hat{Q}_j$ to the lower bounds there.

4 Randomized Primal–Dual Coordinate Approach for Stochastic Programming

In this section, we extend our method to solve a stochastic optimization problem where the objective function involves an expectation. Specifically, we assume the coupled function to be in the form of $f(x) = E_\xi f_\xi(x)$ where ξ is a random vector.

For simplicity, we assume $g = v = 0$, namely, we consider the following problem:

$$\min_x \ E_\xi f_\xi(x) + \sum_{i=1}^N u_i(x_i)$$

$$\text{s.t.} \ \sum_{i=1}^N A_i x_i = b, \ x_i \in \mathcal{X}_i, \ i = 1, 2, \cdots, N. \tag{4.1}$$

One can easily extend our analysis to the case where $g \neq 0, v \neq 0$ and g is also stochastic. An example of (4.1) is the penalized and constrained regression problem that includes (1.2) as a special case.

Due to the expectation form of f, it is natural that the exact gradient of f is not available or very expensive to compute. Instead, we assume that its stochastic gradient is readily accessible. By some slight abuse of the notation, we denote

$$w = \begin{bmatrix} x \\ \lambda \end{bmatrix}, \quad H(w) = \begin{bmatrix} -A^\top \lambda \\ Ax - b \end{bmatrix}, \quad F(x) := E_\xi f_\xi(x) + \sum_{i=1}^N u_i(x_i). \tag{4.2}$$

A point x^* is a solution to (4.1) if and only if there exists λ^* such that

$$F(x) - F(x^*) + (w - w^*)^\top H(w^*) \geqslant 0, \ \forall w, \tag{4.3a}$$

$$Ax^* = b, \ x^* \in \mathcal{X}. \tag{4.3b}$$

Modifying Algorithm 1 to (4.1), we present the stochastic primal–dual coordinate update method of multipliers, summarized in Algorithm 2, where G^k is a stochastic approximation of $\nabla f(x^k)$. The strategy of block coordinate update with stochastic gradient information was first proposed in [41,42], which considered problems without linear constraint. We make the following assumption on the stochastic gradient G^k.

Assumption 4.1 Let $\delta^k = G^k - \nabla f(x^k)$. There exists a constant σ such that for all k,

$$E[\delta^k \mid x^k] = 0, \tag{4.6a}$$

$$E\|\delta^k\|^2 \leqslant \sigma^2. \tag{4.6b}$$

Following the proof of Lemma 3.1 and also noting

$$E_{I_k}\big[(x_{I_k} - x_{I_k}^{k+1})^\top \delta_{I_k}^k \mid x^k\big] = E_{I_k}(x^k - x^{k+1})^\top \delta^k, \tag{4.7}$$

we immediately have the following result.

Lemma 4.1 (One-step analysis) *Let* $\{(x^k, r^k, \lambda^k)\}$ *be the sequence generated from Algorithm* 2 *where* P^k *is given in* (2.15) *with* $\rho_x = \rho$. *Then,*

$$E_{I_k}\big[F(x^{k+1}) - F(x) + (x^{k+1} - x)^\top(-A^\top \lambda^k) + \rho(x^{k+1} - x)^\top A^\top r^{k+1}\big]$$

Algorithm 2: Randomized **P**rimal-**D**ual **B**lock Coordinate **U**pdate Method for **S**tochastic (RPDBUS) Programming

Step 1 Initialization: choose x^0, λ^0 and set parameters ρ, α_k's
Step 2 for $k = 0, 1, \cdots$ **do**
Step 3   Randomly select $I_k \subset [N]$ with $|I_k| = n$ according to **(U)**.
Step 4   Let $x_i^{k+1} = x_i^k$, $\forall i \notin I_k$, and for $I = I_k$, do the update

$$x_I^{k+1} = \arg\min_{x_I \in \mathcal{X}_I} \langle G_I^k - A_I^\top \lambda^k, x_I \rangle + u_I(x_I) \tag{4.4}$$

$$+ \frac{\rho}{2} \|A_I(x_I - x_I^k) + r^k\|^2 + \frac{1}{2}\|x_I - x_I^k\|_{P^k + \frac{I}{\alpha_k}}^2.$$

Update the residual $r^{k+1} = r^k + A_I(x_I^{k+1} - x_I^k)$.
Step 5   Update the multiplier by

$$\lambda^{k+1} = \lambda^k - \left(1 - \frac{(N - n)\alpha_{k+1}}{N\alpha_k}\right)\rho r^{k+1}. \tag{4.5}$$

Step 6 end

$$+ E_{I_k}(x^{k+1} - x)^\top \left(\hat{P} - \rho A^\top A + \frac{I}{\alpha_k}\right)(x^{k+1} - x^k)$$

$$- \frac{L_f}{2} E_{I_k}\|x^k - x^{k+1}\|^2 + E_{I_k}(x^{k+1} - x^k)^\top \delta^k$$

$$\leqslant \left(1 - \frac{n}{N}\right)\left[F(x^k) - F(x) + (x^k - x)^\top(-A^\top \lambda^k) + \rho(x^k - x)^\top A^\top r^k\right]. \tag{4.8}$$

The following theorem is a key result, from which we can choose appropriate α_k to obtain the $O(1/\sqrt{t})$ convergence rate.

Theorem 4.2 *Let $\{(x^k, \lambda^k)\}$ be the sequence generated from Algorithm* 2. *Let $\theta = \frac{n}{N}$ and denote*

$$\beta_k = \frac{\alpha_k}{\left(1 - \frac{\alpha_k(1-\theta)}{\alpha_{k-1}}\right)\rho}, \forall k.$$

Assume $\alpha_k > 0$ is nonincreasing, and

$$Ax^0 = b, \quad \lambda^0 = 0, \tag{4.9a}$$

$$\hat{P} \succcurlyeq L_f I + \rho A^\top A, \tag{4.9b}$$

$$\frac{\alpha_{k-1}\beta_k}{2\alpha_k} + \frac{(1-\theta)\beta_{k+1}}{2} - \frac{\alpha_k\beta_{k+1}}{2\alpha_{k+1}} - \frac{(1-\theta)\beta_k}{2} \geqslant 0, \forall k, \tag{4.9c}$$

$$\frac{\alpha_t}{2\rho} \geqslant \left| \frac{\alpha_{t-1}\beta_t}{\alpha_t} - (1-\theta)\beta_t - \frac{\alpha_t}{\rho} \right|, \text{ for some } t. \tag{4.9d}$$

Let

$$\hat{x}^t = \frac{\alpha_{t+1}x^{t+1} + \theta \sum_{k=1}^{t} \alpha_k x^k}{\alpha_{t+1} + \theta \sum_{k=1}^{t} \alpha_k}. \tag{4.10}$$

Then, under Assumptions 3.1, 3.2, 3.3, and 4.1, we have

$$(\alpha_{t+1} + \theta \sum_{k=1}^{t} \alpha_k) E \left[F(\hat{x}^t) - F(x^*) + \gamma \|A\hat{x}^t - b\| \right]$$

$$\leqslant (1-\theta)\alpha_0 \left[F(x^0) - F(x^*) \right] + \frac{\alpha_0}{2}\|x^0 - x^*\|^2_{\hat{P}-\rho A^\top A} + \frac{1}{2}\|x^0 - x^*\|^2$$

$$+ \left| \frac{\alpha_0\beta_1}{2\alpha_1} - \frac{(1-\theta)\beta_1}{2} \right| \gamma^2 + \sum_{k=0}^{t} \frac{\alpha_k^2}{2} E\|\delta^k\|^2. \tag{4.11}$$

The following proposition gives sublinear convergence rate of Algorithm 2 by specifying the values of its parameters. The choice of α_k depends on whether we fix the total number of iterations.

Proposition 4.3 *Let $\{(x^k, \lambda^k)\}$ be the sequence generated from Algorithm 2 with P^k given in (2.15), $\hat{P}$ satisfying (4.9b), and the initial point satisfying $Ax^0 = b$ and $\lambda^0 = 0$. Let C_0 be*

$$C_0 = (1-\theta)\alpha_0 \left[F(x^0) - F(x^*) \right] + \frac{1}{2}\|x^0 - x^*\|^2_{D_x}$$

$$+ \frac{\alpha_0}{2\rho} \max\{(0.5 + \|\lambda^*\|)^2, 9\|\lambda^*\|^2\}, \tag{4.12}$$

where (x^*, λ^*) is a primal–dual solution, and $D_x := \alpha_0(\hat{P} - \rho A^\top A) + I$.

1. *If $\alpha_k = \frac{\alpha_0}{\sqrt{k}}, \forall k \geqslant 1$ for a certain $\alpha_0 > 0$, then for $t \geqslant 2$,*

$$\max \left\{ E \left| F(\hat{x}^t) - F(x^*) \right|, E\|A\hat{x}^t - b\| \right\} \leqslant \frac{2C_0}{\theta\alpha_0\sqrt{t}} + \frac{\alpha_0(\log t + 2)\sigma^2}{\theta\sqrt{t}}. \tag{4.13}$$

2. *If the number of maximum number of iterations is fixed a priori, then by choosing $\alpha_k = \frac{\alpha_0}{\sqrt{t}}, \forall k \geqslant 1$ with any given $\alpha_0 > 0$, we have*

$$\max \left\{ E \left| F(\hat{x}^t) - F(x^*) \right|, E\|A\hat{x}^t - b\| \right\} \leqslant \frac{2C_0}{\theta\alpha_0\sqrt{t}} + \frac{2\alpha_0\sigma^2}{\theta\sqrt{t}}. \tag{4.14}$$

Proof When $\alpha_k = \frac{\alpha_0}{\sqrt{k}}$, we can show that (4.9c) and (4.9d) hold for $t \geqslant 2$; see Appendix A.4. Hence, the result in (4.13) follows from (4.11), the convexity of F, Lemma 3.5 with $\gamma = \max\{1 + \|\lambda^*\|, 2\|\lambda^*\|\}$, and the inequalities

$$\sum_{k=1}^{t} \frac{1}{\sqrt{k}} \geqslant \sqrt{t}, \quad \sum_{k=1}^{t} \frac{1}{k} \leqslant \log t + 1.$$

When α_k is a constant, the terms on the left-hand side of (4.9c) and on the right-hand side of (4.9d) are both zero, so they are satisfied. Hence, the result in (4.14) immediately follows by noting $\sum_{k=1}^{t} \alpha_k = \alpha_0 \sqrt{t}$ and $\sum_{k=0}^{t} \alpha_k^2 \leqslant 2\alpha_0^2$.

The sublinear convergence result of Algorithm 2 can also be shown if f is non-differentiable convex and Lipschitz continuous. Indeed, if f is Lipschitz continuous with constant L_c, i.e.,

$$\|f(x) - f(y)\| \leqslant L_c \|x - y\|, \ \forall x, y,$$

then $\|\tilde{\nabla} f(x)\| \leqslant L_c, \ \forall x$, where $\tilde{\nabla} f(x)$ is a subgradient of f at x. Hence,

$$
\begin{aligned}
& E_{I_k}(x_{I_k} - x_{I_k}^{k+1})^\top \tilde{\nabla}_{I_k} f(x^k) \\
&= E_{I_k}(x_{I_k} - x_{I_k}^k)^\top \tilde{\nabla}_{I_k} f(x^k) + E_{I_k}(x_{I_k}^k - x_{I_k}^{k+1})^\top \tilde{\nabla}_{I_k} f(x^k) \\
&= \frac{n}{N}(x - x^k)^\top \tilde{\nabla} f(x^k) + E_{I_k}(x^k - x^{k+1})^\top \tilde{\nabla} f(x^{k+1}) \\
& \quad + E_{I_k}(x^k - x^{k+1})^\top \big(\tilde{\nabla} f(x^k) - \tilde{\nabla} f(x^{k+1})\big) \\
&\leqslant \frac{n}{N}(f(x) - f(x^k)) + E_{I_k}[f(x^k) - f(x^{k+1})] \\
& \quad + E_{I_k}(x^k - x^{k+1})^\top \big(\tilde{\nabla} f(x^k) - \tilde{\nabla} f(x^{k+1})\big) \\
&= \frac{n - N}{N}(f(x) - f(x^k)) + E_{I_k}[f(x) - f(x^{k+1})] \\
& \quad + E_{I_k}(x^k - x^{k+1})^\top \big(\tilde{\nabla} f(x^k) - \tilde{\nabla} f(x^{k+1})\big).
\end{aligned}
$$

Now following the proof of Lemma 3.1, we can have a result similar to (4.8), and then through the same arguments as those in the proof of Theorem 4.2, we can establish sublinear convergence rate of $O(1/\sqrt{t})$.

5 Numerical Experiments

In this section, we test the proposed randomized primal–dual method on solving three problems. The first one is the nonnegativity constrained quadratic programming (NCQP):

$$\min_{x \in \mathbb{R}^n} F(x) \equiv \frac{1}{2} x^\top Q x + c^\top x, \text{ s.t. } Ax = b, x_i \geqslant 0, i = 1, \cdots, n, \quad (5.1)$$

where $A \in \mathbb{R}^{m \times n}$, and $Q \in \mathbb{R}^{n \times n}$ is a symmetric positive semidefinite (PSD) matrix. The second problem is the constrained Lasso:

$$\min_{x \in \mathbb{R}^p} F(x) \equiv \frac{1}{2} \|Ax - b\|^2 + \tau \|x\|_1 \text{ s.t. } Cx = d, \tag{5.2}$$

where $A \in \mathbb{R}^{m \times p}$ and $C \in \mathbb{R}^{n \times p}$, and the third one is the log-barrier penalized problem of a linear program:

$$\min_{x, y} F(x, y) \equiv c^\top x - \mu e^\top \log x - \mu e^\top \log y \text{ s.t. } Ax + y$$
$$= b, \ x_i \leqslant u_i, \ i = 1, \cdots, n, \tag{5.3}$$

where $A \in \mathbb{R}^{m \times n}$. The first two problems fall into the case in Theorem 3.6, i.e., multiple x blocks and no y, and the third one falls into Theorem 3.7, i.e., multiple x blocks and single y block. We perform all experiments on a Macbook Pro with 4 cores. The first experiment demonstrates the parallelization performance of the proposed method, and the rest ones compare it to other methods.

5.1 Parallelization

This test is to illustrate the power unleashed in our new method, which is flexible in terms of parallel and distributive computing. We test the algorithm on NCQP (5.1), and we set $m = 200, n = 2\,000$ and generate $Q = HH^\top$, where the components of $H \in \mathbb{R}^{n \times n}$ follow the standard Gaussian distribution. The matrix A and vectors b, c are also randomly generated. We treat every component of x as one block, and at every iteration we select and update p blocks, where p is the number of used cores. Figure 1 shows the running time by using 1, 2, and 4 cores, where the optimal value $F(x^*)$ is obtained by calling Matlab function quadprog with tolerance 10^{-16}. From the figure, we see that our proposed method achieves nearly linear speedup.
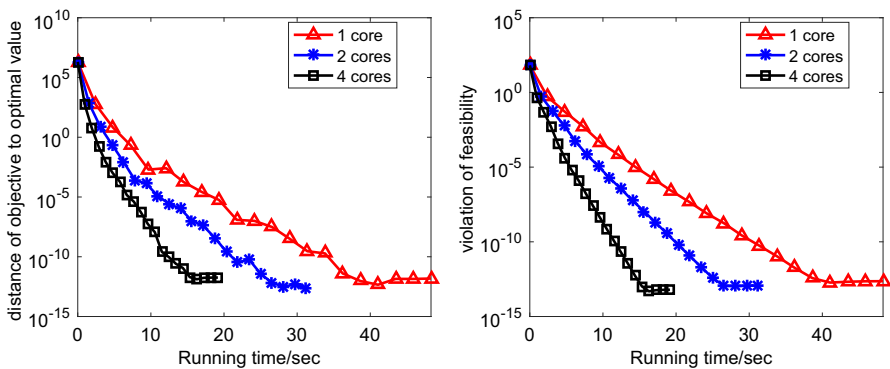

Fig. 1 Nearly linear speedup performance of the proposed primal–dual method for solving (5.1) on a 4-core machine. Left: distance of objective to optimal value $|F(x^k) - F(x^*)|$; Right: violation of feasibility $\|Ax^k - b\|$

5.2 Comparison to Other Methods

In this set of experiments, we compare the proposed method to the linearized ALM and the cyclic linearized ADMM. Note that the linearized ALM is a fully Jacobian update method, while the cyclic linearized ADMM is fully Gauss–Seidel. Without linearization, the former method reduces to the proximal Jacobi ADMM (see section 6.4) and the latter one to the direct multi-block ADMM (see section 6.3). All three methods are tested on NCQP (5.1), the constrained Lasso (5.2), and also the log-barrier penalized problem (5.3).

For the NCQP (5.1), we set $m = 1\,000$, $n = 5\,000$ and generate $Q = HH^\top$, where the components of $H \in \mathbb{R}^{n \times (n-50)}$ follow standard Gaussian distribution. Note that Q is singular, and thus (5.1) is not strongly convex. We evenly partition the variable into 100 blocks. At each iteration of our method, we uniformly at random select one block variable to update. Figure 2 shows the performance by the three compared methods, where one epoch is equivalent to updating all blocks once. The optimal value is computed via Matlab built-in function quadprog. From the figure, we see that our proposed method is comparable to the cyclic linearized ADMM (L-ADMM) and significantly better than the linearized ALM (L-ALM). Although the cyclic ADMM performs well on this instance, in general it can diverge if the problem has more than two blocks; see [19].

For the constrained Lasso (5.2), we set $m = n = 200$ and $p = 2\,000$. A and C are generated according to the standard Gaussian distribution. A sparse vector x^o is generated with 20 nonzero entries. The locations of these nonzeros are chosen uniformly at random, and their values follow the standard Gaussian distribution. Then, we let $b = Ax^o + 0.01\xi$ and $d = Cx^o$, where ξ is generated by the standard Gaussian distribution. We evenly partition the variable into 400 blocks. The same as the previous test, at each iteration of our method, we uniformly at random select one block variable to update. Figure 3 plots the results given by the three compared methods. The optimal value is computed via CVX [43] with high-precision tolerance. From the figure, we see that our proposed method is significantly better than the linearized ALM but slightly
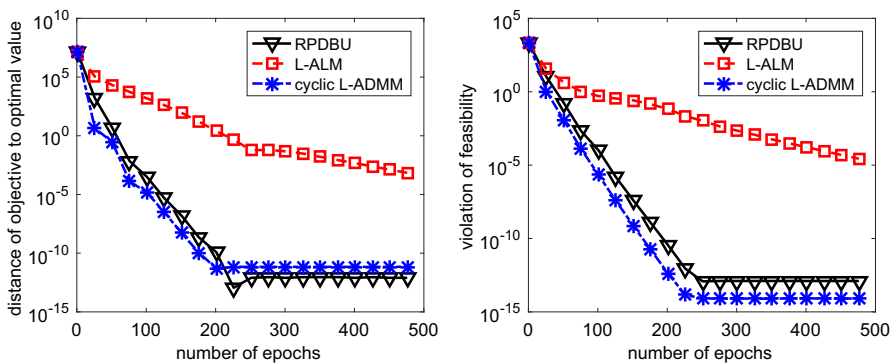

Fig. 2 Comparison of the proposed RPDBU to the L-ALM and the cyclic L-ADMM on solving the nonnegativity constrained quadratic programming (5.1). Left: distance of objective to optimal value $|F(x^k) - F(x^*)|$; Right: violation of feasibility $\|Ax^k - b\|$
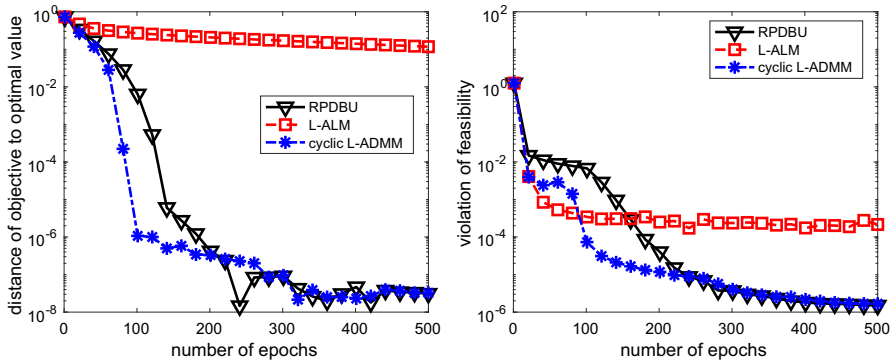
Fig. 3 Comparison of the proposed RPDBU to the L-ALM and the cyclic L-ADMM on solving the constrained Lasso (5.2). Left: distance of objective to optimal value $|F(x^k) - F(x^*)|$; right: violation of feasibility $\|Cx^k - d\|$

worse than the cyclic linearized ADMM. However, the convergence of the cyclic ADMM is not guaranteed. In addition, our method can be parallelized to have nearly linear speedup, while the cyclic ADMM is intrinsically serial.

For the log-barrier penalized problem (5.3), we set $m = 500, n = 2\,000$. μ is set to one, and $u_i = 10$ for all i. The entries of A follow standard Gaussian distribution, and each entry of b is set as a positive number to ensure the existence of a feasible solution. We evenly partition the x-variable into 400 blocks and treat y as a single block. At each iteration, we uniformly at random choose one x block to update, then perform an update to y, and finally renew the multipliers. The results by all three methods are shown in Fig. 4, where the optimal solution is computed via CVX with high-precision tolerance. The figure again demonstrates the significant better performance of our method over the linearized ALM. The cyclic ADMM is slightly better. However, note again that our proposed method has the capability of nearly linear speedup by parallelization, while the cyclic ADMM can only be implemented in a serial way.
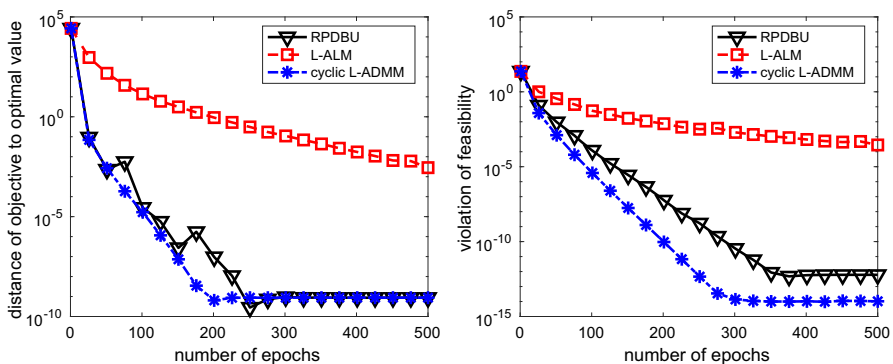

Fig. 4 Comparison of the proposed RPDBU to the L-ALM and the cyclic L-ADMM on solving the log-barrier penalized problem (5.3). Left: distance of objective to optimal value $|F(x^k, y^k) - F(x^*, y^*)|$; right: violation of feasibility $\|Ax^k + y^k - b\|$

6 Connections to Existing Methods

In this section, we discuss how Algorithms 1 and 2 are related to several existing methods in the literature, and we also compare their convergence results. It turns out that the proposed algorithms specialize to several known methods or their variants in the literature under various specific conditions. Therefore, our convergence analysis recovers some existing results as special cases, as well as provides new convergence results for certain existing algorithms such as the Jacobian proximal parallel ADMM and the primal–dual scheme in (1.9).

6.1 Randomized Proximal Coordinate Descent

The randomized proximal coordinate descent (RPCD) was proposed in [44], where smooth convex optimization problems are considered. It was then extended in [45,46] to deal with nonsmooth problems that can be formulated as

$$\min_x f(x_1, \cdots, x_N) + \sum_{i=1}^{N} u_i(x_i), \tag{6.1}$$

where $x = (x_1; \cdots; x_N)$. Toward solving (6.1), at each iteration k, the RPCD method first randomly selects one block i_k and then performs the update:

$$x_i^{k+1} = \begin{cases} \arg\min_{x_i} \langle \nabla_i f(x^k), x_i \rangle + \frac{L_i}{2} \|x_i - x_i^k\|_2^2 + u_i(x_i), & \text{if } i = i_k, \\ x_i^k, & \text{if } i \neq i_k, \end{cases} \tag{6.2}$$

where L_i is the Lipschitz continuity constant of the partial gradient $\nabla_i f(x)$. With more than one block selected every time, (6.2) has been further extended into parallel coordinate descent in [47].

When there is no linear constraint and no y-variable in (1.1), then Algorithm 1 reduces to the scheme in (6.2) if $I_k = \{i_k\}$, i.e., only one block is chosen, and $P^k = L_{i_k} I$, $\lambda^k = 0$, $\forall k$, and to the parallel coordinate descent in [47] if $I_k = \{i_k^1, \cdots, i_k^n\}$ and $P^k = \text{blkdiag}(L_{i_k^1} I, \cdots, L_{i_k^n} I)$, $\lambda^k = 0$, $\forall k$. Although the convergence rate results in [45–47] are non-ergodic, we can easily strengthen our result to a non-ergodic one by noticing that (3.2) implies nonincreasing monotonicity of the objective if Algorithm 1 is applied to (6.1).

6.2 Stochastic Block Proximal Gradient

For solving the problem (6.1) with a stochastic f, [41] proposed a stochastic block proximal gradient (SBPG) method, which iteratively performs the update in (6.2) with $\nabla_i f(x^k)$ replaced by a stochastic approximation. If f is Lipschitz differentiable, then an ergodic $O(1/\sqrt{t})$ convergence rate was shown. Setting $I_k = \{i_k\}$, $\forall k$, we reduce Algorithm 2 to the SBPG method, and thus our convergence results in Proposition 4.3 recover that in [41].

6.3 Multi-block ADMM

Without coupled functions or proximal terms, Algorithm 1 can be regarded as a randomized variant of the multi-block ADMM scheme in (1.6). While multi-block ADMM can diverge if the problem has three or more blocks, our result in Theorem 3.6 shows that $O(1/t)$ convergence rate is guaranteed if at each iteration, one randomly selected block is updated, followed by an update to the multiplier. Note that in the case of no coupled function and $n = 1$, (3.10) indicates that we can choose $P^k = 0$, i.e., without proximal term. Hence, randomization is a key to convergence.

When there are only two blocks, ADMM has been shown (e.g., [14]) to have an ergodic $O(1/t)$ convergence rate. If there are no coupled functions, (3.13) and (3.18a) both indicate that we can choose $\hat{P} = \rho_x A^\top A$, $\hat{Q} = \rho_y B^\top B$ if $\theta = 1$, i.e., all x and y blocks are selected. Thus, according to (2.15), we can set $P^k = 0$, $Q^k = 0$, $\forall k$, in which case Algorithm 1 reduces to the classic 2-block ADMM. Hence, our results in Theorems 3.7 and 3.8 both recover the ergodic $O(1/t)$ convergence rate of ADMM for two-block convex optimization problems.

6.4 Proximal Jacobian Parallel ADMM

In [20], the proximal Jacobian parallel ADMM (Prox-JADMM) was proposed to solve the linearly constrained multi-block separable convex optimization model

$$\min_x \sum_{i=1}^N u_i(x_i), \text{ s.t. } \sum_{i=1}^N A_i x_i = b. \tag{6.3}$$

At each iteration, the Prox-JADMM performs the updates for $i = 1, \cdots, n$ in parallel:

$$x_i^{k+1} = \arg\min_{x_i} u_i(x_i) - \langle \lambda^k, A_i x_i \rangle + \frac{\rho}{2} \Big\| A_i x_i + \sum_{j \neq i} A_j x_j^k - b \Big\|_2^2 + \frac{1}{2} \|x_i - x_i^k\|_{P_i}^2, \tag{6.4}$$

and then updates the multiplier by

$$\lambda^{k+1} = \lambda^k - \gamma\rho \left(\sum_{i=1}^N A_i x_i^{k+1} - b \right), \tag{6.5}$$

where $P_i \succ 0, \forall i$ and $\gamma > 0$ is a damping parameter. By choosing appropriate parameters, [20] established convergence rate of order $1/t$ based on norm square of the difference of two consecutive iterates.

If there is no y-variable or the coupled function f in (1.1), setting $I_k = [N]$, $P^k =$ blkdiag$(\rho_x A_1^\top A_1 + P_1, \cdots, \rho_x A_N^\top A_N + P_N) - \rho_x A^\top A \succcurlyeq 0$, $\forall k$, where P_i are the same as those in (6.4), then Algorithm 1 reduces to the Prox-JADMM with $\gamma = 1$, and Theorem 3.6 provides a new convergence result in terms of the objective value and the feasibility measure.

6.5 Randomized Primal–Dual Scheme in (1.9)

In this subsection, we show that the scheme in (1.9) is a special case of Algorithm 1. Let g be the convex conjugate of $g^* := h + \iota_{\mathcal{Z}}$, namely, $g(y) = \sup_z \langle y, z \rangle - h(z) - \iota_{\mathcal{Z}}(z)$. Then, (1.8) is equivalent to the optimization problem:

$$\min_{x \in \mathcal{X}} \sum_{i=1}^{N} u_i(x_i) + g(-Ax),$$

which can be further written as

$$\min_{x \in \mathcal{X}, y} \sum_{i=1}^{N} u_i(x_i) + g(y), \text{ s.t. } Ax + y = 0. \tag{6.6}$$

Proposition 6.1 *The scheme in* (1.9) *is equivalent to the following updates*:

$$x_i^{k+1} = \begin{cases} \arg\min_{x_i \in \mathcal{X}_i} \langle -z^k, A_i x_i \rangle + u_i(x_i) + \frac{q}{2\eta} \|A_i(x_i - x_i^k) + r^k\|^2 + \frac{1}{2} \|x_i - x_i^k\|_{\tau I - \frac{q}{\eta} A_i^\top A_i}^2, & i = i_k, \\ x_i^k, \quad i \neq i_k, \end{cases} \tag{6.7a}$$

$$y^{k+1} = \arg\min_y g(y) - \langle y, z^k \rangle + \frac{1}{2\eta} \|y + Ax^{k+1}\|^2, \tag{6.7b}$$

$$z^{k+1} = z^k - \frac{1}{\eta}(Ax^{k+1} + y^{k+1}), \tag{6.7c}$$

where $r^k = Ax^k + y^k$. *Therefore, it is a special case of Algorithm* 1 *applied to* (6.6) *with the setting of* $I_k = \{i_k\}$, $\rho_x = \frac{q}{\eta}$, $\rho_y = \rho = \frac{1}{\eta}$ *and* $P^k = \tau I - \frac{q}{\eta} A_{i_k}^\top A_{i_k}$, $Q^k = 0, \forall k$.

While the sublinear convergence rate result in [36] requires the boundedness of $\mathcal{X}$ and $\mathcal{Z}$, the result in Theorem 3.7 indicates that the boundedness assumption can be removed if we add one proximal term to the y-update in (6.7b).

7 Concluding Remarks

We have proposed a randomized primal–dual coordinate update algorithm, called RPDBU, for solving linearly constrained convex optimization with multi-block decision variables and coupled terms in the objective. By using a randomization scheme and the proximal gradient mappings, we show a sublinear convergence rate of the RPDBU method. In particular, without any assumptions other than convexity on the objective function and without imposing any restrictions on the constraint matrices, an $O(1/t)$ convergence rate is established. We have also extended RPDBU to solve the problem where the objective is stochastic. If a stochastic subgradient estimator is available, then we show that by adaptively choosing the parameter α_k in the added proximal term, an $O(1/\sqrt{t})$ convergence rate can be established. Furthermore, if there is no coupled function f, then we can remove the proximal term, and the algorithm reduces to a randomized multi-block ADMM. Hence, the convergence of the original randomized

multi-block ADMM follows as a consequence of our analysis. Remark also that by taking the sampling set I_k as the whole set and P^k as some special matrices, our algorithm specializes to the proximal Jacobian ADMM. Finally, we pose as an open problem to decide whether or not a deterministic counterpart of the RPDBU exists, retaining similar convergence properties for solving problem (1.1). For instance, it would be interesting to know whether the algorithm would still be convergent if a deterministic cyclic update rule is applied while a proper proximal term is incorporated.

A Proofs of Lemmas

We give proofs of several lemmas that are used to show our main results.

A.1 Proof of Lemma 3.1

We prove (3.2), and (3.3) can be shown by the same arguments. By the optimality of $x_{I_k}^{k+1}$ from (2.10) or equivalently (2.16), we have for any $x_{I_k} \in \mathcal{X}_{I_k}$,

$$(x_{I_k} - x_{I_k}^{k+1})^\top \left(\nabla_{I_k} f(x^k) - A_{I_k}^\top \lambda^k + \rho_x A_{I_k}^\top r^k + \tilde{\nabla} u_{I_k}(x_{I_k}^{k+1}) + \hat{P}_{I_k}(x_{I_k}^{k+1} - x_{I_k}^k) \right)$$
$$\geq 0, \tag{A.1}$$

where $\tilde{\nabla} u_{I_k}(x_{I_k}^{k+1})$ is a subgradient of u_{I_k} at $x_{I_k}^{k+1}$, and we have used the formula of $r^{k+\frac{1}{2}}$ given in (2.11). We compute the expectation of each term in (A.1) as follows. First, we have

$$E_{I_k}(x_{I_k} - x_{I_k}^{k+1})^\top \nabla_{I_k} f(x^k)$$
$$= E_{I_k} \left(x_{I_k} - x_{I_k}^k \right)^\top \nabla_{I_k} f(x^k) + E_{I_k}(x_{I_k}^k - x_{I_k}^{k+1})^\top \nabla_{I_k} f(x^k)$$
$$= \frac{n}{N} \left(x - x^k \right)^\top \nabla f(x^k) + E_{I_k}(x^k - x^{k+1})^\top \nabla f(x^k) \tag{A.2}$$
$$\leq \frac{n}{N} \left(f(x) - f(x^k) \right) + E_{I_k} \left[f(x^k) - f(x^{k+1}) + \frac{L_f}{2} \|x^k - x^{k+1}\|^2 \right]$$
$$= \frac{n - N}{N} \left(f(x) - f(x^k) \right)$$
$$+ E_{I_k} \left[f(x) - f(x^{k+1}) + \frac{L_f}{2} \|x^k - x^{k+1}\|^2 \right], \tag{A.3}$$

where the last inequality is from the convexity of f and the Lipschitz continuity of $\nabla_{I_k} f(x)$. Secondly,

$$E_{I_k}(x_{I_k} - x_{I_k}^{k+1})^\top (-A_{I_k}^\top \lambda^k)$$
$$= E_{I_k}(x_{I_k} - x_{I_k}^k)^\top (-A_{I_k}^\top \lambda^k) + E_{I_k}(x_{I_k}^k - x_{I_k}^{k+1})^\top (-A_{I_k}^\top \lambda^k)$$

$$= \frac{n}{N}\left(x - x^k\right)^\top (-A^\top \lambda^k) + E_{I_k}(x^k - x^{k+1})^\top (-A^\top \lambda^k)$$

$$= \frac{n-N}{N}\left(x - x^k\right)^\top (-A^\top \lambda^k) + E_{I_k}(x - x^{k+1})^\top (-A^\top \lambda^k). \quad \text{(A.4)}$$

Similarly,

$$\rho_x E_{I_k}(x_{I_k} - x_{I_k}^{k+1})^\top A_{I_k}^\top r^k = \frac{n-N}{N}\rho_x(x - x^k)^\top A^\top r^k$$
$$+ \rho_x E_{I_k}(x - x^{k+1})^\top A^\top r^k. \quad \text{(A.5)}$$

For the fourth term of (A.1), we have

$$E_{I_k}(x_{I_k} - x_{I_k}^{k+1})^\top \tilde{\nabla} u_{I_k}(x_{I_k}^{k+1})$$

$$\leqslant E_{I_k}\left[u_{I_k}(x_{I_k}) - u_{I_k}(x_{I_k}^{k+1})\right]$$

$$= \frac{n}{N} u(x) - E_{I_k}[u(x^{k+1}) - u(x^k) + u_{I_k}(x_{I_k}^k)]$$

$$= \frac{n}{N}\left[u(x) - u(x^k)\right] + E_{I_k}[u(x^k) - u(x^{k+1})]$$

$$= \frac{n-N}{N}\left[u(x) - u(x^k)\right] + E_{I_k}[u(x) - u(x^{k+1})], \quad \text{(A.6)}$$

where the inequality is from the convexity of u_{I_k}. Finally, we have

$$E_{I_k}(x_{I_k} - x_{I_k}^{k+1})^\top \hat{P}_{I_k}(x_{I_k}^{k+1} - x_{I_k}^k) = E_{I_k}(x - x^{k+1})^\top \hat{P}(x^{k+1} - x^k). \quad \text{(A.7)}$$

Plugging (A.3) through (A.7) into (A.1) and recalling $F(x) = f(x) + u(x)$, by rearranging terms we have

$$E_{I_k}\left[F(x^{k+1}) - F(x) + (x^{k+1} - x)^\top (-A^\top \lambda^k) + \rho_x(x^{k+1} - x)^\top A^\top r^k\right]$$

$$+ E_{I_k}\left(x^{k+1} - x\right)^\top \hat{P}(x^{k+1} - x^k) - \frac{L_f}{2}E_{I_k}\|x^k - x^{k+1}\|^2$$

$$\leqslant \frac{N-n}{N}\left[F(x^k) - F(x) + (x^k - x)^\top (-A^\top \lambda^k) + \rho_x(x^k - x)^\top A^\top r^k\right]. \quad \text{(A.8)}$$

Note

$$(x^{k+1} - x)^\top (-A^\top \lambda^k) + \rho_x(x^{k+1} - x)^\top A^\top r^k$$
$$= (x^{k+1} - x)^\top (-A^\top \lambda^k) + \rho_x(x^{k+1} - x)^\top A^\top r^{k+1}$$
$$- \rho_x(x^{k+1} - x)^\top A^\top A(x^{k+1} - x^k) - \rho_x(x^{k+1} - x)^\top A^\top B(y^{k+1} - y^k)$$
$$\overset{(2.14)}{=} (x^{k+1} - x)^\top (-A^\top \lambda^{k+1}) + (\rho_x - \rho)(x^{k+1} - x)^\top A^\top r^{k+1}$$
$$- \rho_x(x^{k+1} - x)^\top A^\top A(x^{k+1} - x^k) - \rho_x(x^{k+1} - x)^\top A^\top B(y^{k+1} - y^k).$$

Hence, we can rewrite (A.8) equivalently into (3.2). Through the same arguments, one can show (3.3), thus completing the proof.

A.2 Proof of Lemma 3.3

Letting $x = x^*$, $y = y^*$ in (3.8), we have for any λ that

$$
\begin{aligned}
&E[h(x^*, y^*, \lambda)] \\
&\geqslant E\Big[\Phi(\hat{x}, \hat{y}) - \Phi(x^*, y^*) + (\hat{x} - x^*)^\top(-A^\top\hat{\lambda}) + (\hat{y} - y^*)^\top(-B^\top\hat{\lambda}) \\
&\quad + (\hat{\lambda} - \lambda)^\top(A\hat{x} + B\hat{y} - b)\Big] \\
&= E\Big[\Phi(\hat{x}, \hat{y}) - \Phi(x^*, y^*) + \langle\hat{\lambda}, Ax^* + By^* - b\rangle - \langle\lambda, A\hat{x} + B\hat{y} - b\rangle\Big] \\
&= E\Big[\Phi(\hat{x}, \hat{y}) - \Phi(x^*, y^*) - \langle\lambda, A\hat{x} + B\hat{y} - b\rangle\Big], \qquad (A.9)
\end{aligned}
$$

where the last equality follows from the feasibility of (x^*, y^*). For any $\gamma > 0$, restricting λ in $\mathcal{B}_\gamma$, we have

$$
E[h(x^*, y^*, \lambda)] \leqslant \sup_{\lambda \in \mathcal{B}_\gamma} h(x^*, y^*, \lambda).
$$

Hence, letting $\lambda = -\frac{\gamma(A\hat{x} + B\hat{y} - b)}{\|A\hat{x} + B\hat{y} - b\|} \in \mathcal{B}_\gamma$ in (A.9) gives the desired result.

A.3 Proof of Lemma 3.5

In view of (3.9), we have

$$
\begin{aligned}
&E\big[(\gamma - \|\lambda^*\|)\|A\hat{x} + B\hat{y} - b\|\big] \\
&\leqslant E\big[\Phi(\hat{x}, \hat{y}) - \Phi(x^*, y^*) + \gamma\|A\hat{x} + B\hat{y} - b\|\big] \leqslant \varepsilon,
\end{aligned}
$$

which implies

$$
E\|A\hat{x} + B\hat{y} - b\| \leqslant \frac{\varepsilon}{\gamma - \|\lambda^*\|}, \text{ and } E\big[\Phi(\hat{x}, \hat{y}) - \Phi(x^*, y^*)\big] \leqslant \varepsilon. \quad (A.10)
$$

Denote $a^- = \max(0, -a)$ for a real number a. Then, from (3.9) and (A.10), it follows that

$$
E\big(\Phi(\hat{x}, \hat{y}) - \Phi(x^*, y^*)\big)^- \leqslant \|\lambda^*\| \cdot E\|A\hat{x} + B\hat{y} - b\| \leqslant \frac{\|\lambda^*\|\varepsilon}{\gamma - \|\lambda^*\|}.
$$

Noting $|a| = a + 2a^-$ for any real number a, we have

$$
E\big|\Phi(\hat{x}, \hat{y}) - \Phi(x^*, y^*)\big|
$$

$$= E\big[\Phi(\hat{x}, \hat{y}) - \Phi(x^*, y^*)\big] + 2E\big(\Phi(\hat{x}, \hat{y}) - \Phi(x^*, y^*)\big)^-$$

$$\leqslant \left(\frac{2\|\lambda^*\|}{\gamma - \|\lambda^*\|} + 1\right)\varepsilon. \tag{A.11}$$

A.4 Proof of Inequalities (4.9c) and (4.9d) with $\alpha_k = \frac{\alpha_0}{\sqrt{k}}$

We have $\beta_k = \frac{\alpha_0}{\rho\left(\sqrt{k} - (1-\theta)\sqrt{k-1}\right)}$, and

$$\frac{\alpha_{k-1}}{\alpha_k}\beta_k + (1-\theta)\beta_{k+1} - \frac{\alpha_k}{\alpha_{k+1}}\beta_{k+1} - (1-\theta)\beta_k$$

$$= \frac{\alpha_0}{\rho}\left[\left(\frac{\sqrt{k}}{\sqrt{k-1}} - (1-\theta)\right)\frac{1}{\left(\sqrt{k} - (1-\theta)\sqrt{k-1}\right)}\right.$$

$$\left. -\left(\frac{\sqrt{k+1}}{\sqrt{k}} - (1-\theta)\right)\frac{1}{\left(\sqrt{k+1} - (1-\theta)\sqrt{k}\right)}\right]$$

$$=: \frac{\alpha_0}{\rho}[\psi(k) - \psi(k+1)].$$

By elementary calculus, we have

$$\psi'(k) = \frac{\frac{\sqrt{k-1}}{\sqrt{k}} - \frac{\sqrt{k}}{\sqrt{k-1}}}{2(k-1)}\frac{1}{\left(\sqrt{k} - (1-\theta)\sqrt{k-1}\right)}$$

$$+ \left(\frac{\sqrt{k}}{\sqrt{k-1}} - (1-\theta)\right)\frac{-1}{2\left(\sqrt{k} - (1-\theta)\sqrt{k-1}\right)^2}\left(\frac{1}{\sqrt{k}} - \frac{1-\theta}{\sqrt{k-1}}\right)$$

$$= \frac{1}{2(k-1)(\sqrt{k} - (1-\theta)\sqrt{k-1})}$$

$$\times\left[\frac{\sqrt{k-1}}{\sqrt{k}} - \frac{\sqrt{k}}{\sqrt{k-1}} - \sqrt{k-1}\left(\frac{1}{\sqrt{k}} - \frac{1-\theta}{\sqrt{k-1}}\right)\right]$$

$$= \frac{1}{2(k-1)(\sqrt{k} - (1-\theta)\sqrt{k-1})}\left((1-\theta) - \frac{\sqrt{k}}{\sqrt{k-1}}\right) < 0.$$

Hence, $\psi(k)$ is decreasing with respect to k, and thus (4.9c) holds.

When $\alpha_k = \frac{\alpha_0}{\sqrt{k}}$, (4.9d) becomes

$$\frac{\alpha_0}{2\rho\sqrt{t}} \geqslant \left|\left(\frac{\sqrt{t}}{\sqrt{t-1}} - (1-\theta)\right)\frac{\alpha_0}{\rho(\sqrt{t} - (1-\theta)\sqrt{t-1})} - \frac{\alpha_0}{\rho\sqrt{t}}\right|,$$

which is equivalent to

$$\frac{1}{2} \geqslant \frac{\sqrt{t}}{\sqrt{t-1}} - 1 \Longleftrightarrow t \geqslant \frac{9}{5}.$$

This completes the proof.

B Proofs of Theorems

In this section, we give the technical details for showing all theorems. For simplicity of notation, throughout the proofs of this section, we define $\tilde{P}$ and $\tilde{Q}$ as follows:

$$\tilde{P} = \hat{P} - \rho_x A^\top A, \qquad \tilde{Q} = \hat{Q} - \rho_y B^\top B. \tag{B.1}$$

B.1 Proof of Theorem 3.6

Taking expectation over both sides of (3.2) and summing it over $k = 0$ through t, we have

$$E\big[F(x^{t+1}) - F(x) + (x^{t+1}-x)^\top(-A^\top \lambda^{t+1})\big] + (1-\theta)\rho_x E(x^{t+1} - x)^\top A^\top r^{t+1}$$

$$+\theta \sum_{k=0}^{t-1} E\big[F(x^{k+1}) - F(x) + (x^{k+1} - x)^\top(-A^\top \lambda^{k+1})\big]$$

$$-\sum_{k=0}^{t} \rho_x E(x^{k+1} - x)^\top A^\top B(y^{k+1} - y^k)$$

$$+\sum_{k=0}^{t} E(x^{k+1} - x)^\top \tilde{P}(x^{k+1} - x^k) - \frac{L_f}{2} \sum_{k=0}^{t} E\|x^k - x^{k+1}\|^2$$

$$\leqslant (1 - \theta)\Big[F(x^0) - F(x) + (x^0 - x)^\top(-A^\top \lambda^0) + \rho_x(x^0 - x)^\top A^\top r^0\Big], \tag{B.2}$$

where we have used $\frac{n}{N} = \theta$, the condition in (3.17) and the definition of $\tilde{P}$ in (B.1). Similarly, taking expectation over both sides of (3.3), summing it over $k = 0$ through t, we have

$$E\big[G(y^{t+1}) - G(y) + (y^{t+1} - y)^\top(-B^\top \lambda^{t+1})\big]$$

$$+(1 - \theta)\rho_y E(y^{t+1} - y)^\top B^\top r^{t+1}$$

$$+\theta \sum_{k=0}^{t-1} E\big[G(y^{k+1}) - G(y) + (y^{k+1} - y)^\top(-B^\top \lambda^{k+1})\big]$$

$$+\sum_{k=0}^{t} E(y^{k+1} - y)^\top \tilde{Q}(y^{k+1} - y^k) - \frac{L_g}{2} \sum_{k=0}^{t} E\|y^k - y^{k+1}\|^2$$

$$\leqslant (1 - \theta) \left[G(y^0) - G(y) + (y^0 - y)^\top (-B^\top \lambda^0) + \rho_y (y^0 - y)^\top B^\top r^0 \right]$$

$$+ (1 - \theta) \sum_{k=0}^{t} E \rho_y (y^k - y)^\top B^\top A(x^{k+1} - x^k). \tag{B.3}$$

Recall $\lambda^{k+1} = \lambda^k - \rho r^{k+1}$, thus

$$(\lambda^{k+1} - \lambda)^\top r^{k+1} = -\frac{1}{\rho} (\lambda^{k+1} - \lambda)^\top (\lambda^{k+1} - \lambda^k), \tag{B.4}$$

where λ is an arbitrary vector and possibly random. Denote $\tilde{\lambda}^{t+1} = \lambda^t - \rho_x r^{t+1}$. Then, similar to (B.4), we have

$$(\tilde{\lambda}^{t+1} - \lambda)^\top r^{t+1} = -\frac{1}{\rho_x} (\tilde{\lambda}^{t+1} - \lambda)^\top (\tilde{\lambda}^{t+1} - \lambda^t). \tag{B.5}$$

Summing (B.2) and (B.3) together and using (B.4) and (B.5), we have

$$E \left[\Phi(x^{t+1}, y^{t+1}) - \Phi(x, y) + (\tilde{w}^{t+1} - w)^\top H(\tilde{w}^{t+1}) + \frac{1}{\rho_x} (\tilde{\lambda}^{t+1} - \lambda)^\top (\tilde{\lambda}^{t+1} - \lambda^t) \right]$$

$$+ \theta \sum_{k=0}^{t-1} E \left[\Phi(x^{k+1}, y^{k+1}) - \Phi(x, y) + (w^{k+1} - w)^\top H(w^{k+1}) \right.$$

$$\left. + \frac{1}{\rho} (\lambda^{k+1} - \lambda)^\top (\lambda^{k+1} - \lambda^k) \right]$$

$$\leqslant (1 - \theta) \left[F(x^0) - F(x) + (x^0 - x)^\top (-A^\top \lambda^0) + \rho_x (x^0 - x)^\top A^\top r^0 \right]$$

$$+ (1 - \theta) \left[G(y^0) - G(y) + (y^0 - y)^\top (-B^\top \lambda^0) + \rho_y (y^0 - y)^\top B^\top r^0 \right]$$

$$+ \sum_{k=0}^{t} \rho_x E(x^{k+1} - x)^\top A^\top B(y^{k+1} - y^k) + (1 - \theta) \sum_{k=0}^{t} \rho_y E(y^k - y)^\top B^\top A(x^{k+1} - x^k)$$

$$- \sum_{k=0}^{t} E(x^{k+1} - x)^\top \tilde{P}(x^{k+1} - x^k) + \frac{L_f}{2} \sum_{k=0}^{t} E \|x^k - x^{k+1}\|^2$$

$$- \sum_{k=0}^{t} E(y^{k+1} - y)^\top \tilde{Q}(y^{k+1} - y^k) + \frac{L_g}{2} \sum_{k=0}^{t} E \|y^k - y^{k+1}\|^2, \tag{B.6}$$

where we have used $\Phi(x, y) = F(x) + G(y)$ and the definition of H given in (2.18). When $B = 0$ and $y^k \equiv y^0$, (B.6) reduces to

$$E \left[F(x^{t+1}) - F(x) + (\tilde{w}^{t+1} - w)^\top H(\tilde{w}^{t+1}) + \frac{1}{\rho_x} (\tilde{\lambda}^{t+1} - \lambda)^\top (\tilde{\lambda}^{t+1} - \lambda^t) \right]$$

$$+ \theta \sum_{k=0}^{t-1} E \left[F(x^{k+1}) - F(x) + (w^{k+1} - w)^\top H(w^{k+1}) \right.$$

$$+ \frac{1}{\rho} (\lambda^{k+1} - \lambda)^\top (\lambda^{k+1} - \lambda^k) \Big]$$

$$\leqslant (1 - \theta) \Big[F(x^0) - F(x) + (x^0 - x)^\top (-A^\top \lambda^0) + \rho_x (x^0 - x)^\top A^\top r^0 \Big]$$

$$- \sum_{k=0}^{t} E(x^{k+1} - x)^\top \tilde{P}(x^{k+1} - x^k) + \frac{L_f}{2} \sum_{k=0}^{t} E \|x^k - x^{k+1}\|^2.$$

Using (2.21) and noting $\theta = \frac{\rho}{\rho_x}$, from the above inequality after canceling terms we have

$$E \Big[F(x^{t+1}) - F(x) + (\tilde{w}^{t+1} - w)^\top H(\tilde{w}^{t+1}) \Big]$$

$$+ \theta \sum_{k=0}^{t-1} E \Big[F(x^{k+1}) - F(x) + (w^{k+1} - w)^\top H(w^{k+1}) \Big]$$

$$+ \frac{1}{2\rho_x} E \Big[\|\tilde{\lambda}^{t+1} - \lambda\|^2 - \|\lambda^0 - \lambda\|^2 + \|\tilde{\lambda}^{t+1} - \lambda^t\|^2 + \sum_{k=0}^{t-1} \|\lambda^{k+1} - \lambda^k\|^2 \Big]$$

$$\leqslant (1 - \theta) \Big[F(x^0) - F(x) + (x^0 - x)^\top (-A^\top \lambda^0) + \rho_x (x^0 - x)^\top A^\top r^0 \Big]$$

$$- \frac{1}{2} E \Big[\|x^{t+1} - x\|_{\tilde{P}}^2 - \|x^0 - x\|_{\tilde{P}}^2 + \sum_{k=0}^{t} \|x^{k+1} - x^k\|_{\tilde{P}}^2 \Big]$$

$$+ \frac{L_f}{2} \sum_{k=0}^{t} E \|x^k - x^{k+1}\|^2. \tag{B.7}$$

For any feasible x, we note $\tilde{\lambda}^{t+1} - \lambda^t = \rho_x A(x^{t+1} - x)$ and thus

$$\frac{1}{\rho_x} \|\tilde{\lambda}^{t+1} - \lambda^t\|^2 = \rho_x \|x^{t+1} - x\|_{A^\top A}^2. \tag{B.8}$$

In addition, since x^{k+1} and x^k differ only on the index set I_k, we have by recalling $\tilde{P} = \hat{P} - \rho_x A^\top A$ that

$$\|x^{k+1} - x^k\|_{\tilde{P}}^2 - L_f \|x^{k+1} - x^k\|^2 = \|x_{I_k}^{k+1} - x_{I_k}^k\|_{\hat{P}_{I_k}}^2 - \|x_{I_k}^{k+1} - x_{I_k}^k\|_{\rho_x A_{I_k}^\top A_{I_k}}^2$$

$$- L_f \|x_{I_k}^{k+1} - x_{I_k}^k\|^2. \tag{B.9}$$

Plugging (B.8) and (B.9) into (B.7) and using (3.10) lead to

$$E \Big[F(x^{t+1}) - F(x) + (\tilde{w}^{t+1} - w)^\top H(\tilde{w}^{t+1}) \Big]$$

$$+ \theta \sum_{k=0}^{t-1} E \Big[F(x^{k+1}) - F(x) + (w^{k+1} - w)^\top H(w^{k+1}) \Big]$$

$$\leqslant (1 - \theta) \left[F(x^0) - F(x) + (x^0 - x)^\top (-A^\top \lambda^0) + \rho_x (x^0 - x)^\top A^\top r^0 \right]$$
$$+ \frac{1}{2\rho_x} E \|\lambda^0 - \lambda\|^2 + \frac{1}{2} \|x^0 - x\|_{\tilde{P}}^2.$$

The desired result follows from $\lambda^0 = 0$, and Lemmas 3.3 and 3.5 with $\gamma = \max\{0.5 + \|\lambda^*\|, 3\|\lambda^*\|\}$.

B.2 Proof of Theorem 3.7

It follows from (3.3) with $\rho_y = \rho$ and $m = M$ that (recall the definition of $\tilde{Q}$ in (B.1)) for any $y \in \mathcal{Y}$,

$$G(y^{k+1}) - G(y) - \frac{L_g}{2} \|y^k - y^{k+1}\|^2 + (y^{k+1} - y)^\top (-B^\top \lambda^{k+1})$$
$$+ (y^{k+1} - y)^\top \tilde{Q}(y^{k+1} - y^k) \leqslant 0. \tag{B.10}$$

Similar to (B.10), and recalling the definition of $\tilde{y}^{t+1}$, we have for any $y \in \mathcal{Y}$,

$$G(\tilde{y}^{t+1}) - G(y) - \frac{L_g}{2} \|\tilde{y}^{t+1} - y^t\|^2 + (\tilde{y}^{t+1} - y)^\top (-B^\top \tilde{\lambda}^{t+1})$$
$$+ \theta (\tilde{y}^{t+1} - y)^\top \tilde{Q}(y^{t+1} - y^t) \leqslant 0, \tag{B.11}$$

where

$$\tilde{\lambda}^{t+1} = \lambda^t - \rho_x (Ax^{t+1} + B\tilde{y}^{t+1} - b). \tag{B.12}$$

Adding (B.10) and (B.11) to (B.2) and using the formula of λ^k give

$$E \left[F(x^{t+1}) - F(x) + (x^{t+1} - x)^\top (-A^\top \tilde{\lambda}^{t+1}) \right]$$
$$+ E \left(\tilde{\lambda}^{t+1} - \lambda \right)^\top \left(Ax^{t+1} + B\tilde{y}^{t+1} - b + \frac{1}{\rho_x} (\tilde{\lambda}^{t+1} - \lambda^t) \right)$$
$$+ E \left[G(\tilde{y}^{t+1}) - G(y) + (\tilde{y}^{t+1} - y)^\top (-B^\top \tilde{\lambda}^{t+1}) \right.$$
$$\left. + \theta (\tilde{y}^{t+1} - y)^\top \tilde{Q}(\tilde{y}^{t+1} - y^k) \right] - \frac{L_g}{2} E \|\tilde{y}^{t+1} - y^t\|^2$$
$$+ \theta \sum_{k=0}^{t-1} E \left[F(x^{k+1}) - F(x) + (x^{k+1} - x)^\top (-A^\top \lambda^{k+1}) \right]$$
$$- \sum_{k=0}^{t-1} \rho_x E(x^{k+1} - x)^\top A^\top B(y^{k+1} - y^k) - \rho_x E(x^{t+1} - x)^\top A^\top B(\tilde{y}^{t+1} - y^t)$$
$$+ \theta \sum_{k=0}^{t-1} E \left[G(y^{k+1}) - G(y) - \frac{L_g}{2} \|y^k - y^{k+1}\|^2 + (y^{k+1} - y)^\top (-B^\top \lambda^{k+1}) \right.$$

$$+(y^{k+1} - y)^\top \tilde{Q}(y^{k+1} - y^k)\Big]$$

$$+\theta \sum_{k=0}^{t-1} E(\lambda^{k+1} - \lambda)^\top \left(r^{k+1} + \frac{1}{\rho}(\lambda^{k+1} - \lambda^k)\right)$$

$$\leqslant (1 - \theta)\left[F(x^0) - F(x) + (x^0 - x)^\top(-A^\top\lambda^0) + \rho_x(x^0 - x)^\top A^\top r^0\right]$$

$$-\sum_{k=0}^{t} E(x^{k+1} - x)^\top \tilde{P}(x^{k+1} - x^k) + \frac{L_f}{2}\sum_{k=0}^{t} E\|x^k - x^{k+1}\|^2. \qquad (B.13)$$

By the notation in (2.18) and using (3.6), (B.13) can be written into

$$E\left[\Phi(x^{t+1}, \tilde{y}^{t+1}) - \Phi(x, y) + (\tilde{w}^{t+1} - w)^\top H(\tilde{w}^{t+1})\right]$$

$$+\theta \sum_{k=0}^{t-1} E\left[\Phi(x^{k+1}, y^{k+1}) - \Phi(x, y) + (w^{k+1} - w)^\top H(w^{k+1})\right]$$

$$+\theta \sum_{k=0}^{t-1} E(y^{k+1} - y)^\top \tilde{Q}(y^{k+1} - y^k) + \theta E(\tilde{y}^{t+1} - y)^\top \tilde{Q}(\tilde{y}^{t+1} - y^t)$$

$$-\sum_{k=0}^{t-1} \rho_x E\left(\frac{1}{\rho}(\lambda^k - \lambda^{k+1})^\top B(y^{k+1} - y^k) - (y^{k+1} - y)^\top B^\top B(y^{k+1} - y^k)\right)$$

$$-\rho_x E\left(\frac{1}{\rho_x}(\lambda^t - \tilde{\lambda}^{t+1})^\top B(\tilde{y}^{t+1} - y^t) - (\tilde{y}^{t+1} - y)^\top B^\top B(\tilde{y}^{t+1} - y^t)\right)$$

$$+\frac{\theta}{\rho}E(\tilde{\lambda}^{t+1} - \lambda)^\top \left(\tilde{\lambda}^{t+1} - \lambda^t\right) + \frac{\theta}{\rho}\sum_{k=0}^{t-1} E(\lambda^{k+1} - \lambda)^\top \left(\lambda^{k+1} - \lambda^k\right)$$

$$\leqslant (1 - \theta)\left[F(x^0) - F(x) + (x^0 - x)^\top(-A^\top\lambda^0) + \rho_x(x^0 - x)^\top A^\top r^0\right]$$

$$-\sum_{k=0}^{t} E(x^{k+1} - x)^\top \tilde{P}(x^{k+1} - x^k) + \frac{L_f}{2}\sum_{k=0}^{t} E\|x^k - x^{k+1}\|^2$$

$$+\frac{\theta L_g}{2}\sum_{k=0}^{t-1} E\|y^k - y^{k+1}\|^2 + \frac{L_g}{2}E\|\tilde{y}^{t+1} - y^t\|^2.$$

Now use (2.21) to derive from the above inequality that

$$E\left[\Phi(x^{t+1}, \tilde{y}^{t+1}) - \Phi(x, y) + (\tilde{w}^{t+1} - w)^\top H(\tilde{w}^{t+1})\right]$$

$$+\theta \sum_{k=0}^{t-1} E\left[\Phi(x^{k+1}, y^{k+1}) - \Phi(x, y) + (w^{k+1} - w)^\top H(w^{k+1})\right]$$

$$+\frac{\theta}{2}\left(E\|\tilde{y}^{t+1} - y\|_{\tilde{Q}}^2 - \|y^0 - y\|_{\tilde{Q}}^2\right)$$

$$+\frac{\theta}{2}\sum_{k=0}^{t-1}E\|y^{k+1}-y^k\|_{\tilde{Q}}^2 + \frac{\theta}{2}E\|\tilde{y}^{t+1}-y^t\|_{\tilde{Q}}^2$$

$$+\frac{\rho_x}{2}\left(E\|\tilde{y}^{t+1}-y\|_{B^\top B}^2 - \|y^0-y\|_{B^\top B}^2\right)$$

$$+\frac{\rho_x}{2}\sum_{k=0}^{t-1}E\|y^{k+1}-y^k\|_{B^\top B}^2 + \frac{\rho_x}{2}E\|\tilde{y}^{t+1}-y^t\|_{B^\top B}^2$$

$$-\sum_{k=0}^{t-1}E\frac{\rho_x}{\rho}\left(\lambda^k-\lambda^{k+1}\right)^\top B(y^{k+1}-y^k) - E(\lambda^t-\tilde{\lambda}^{t+1})^\top B(\tilde{y}^{t+1}-y^t)$$

$$+\frac{\theta}{2\rho}\left(E\|\tilde{\lambda}^{t+1}-\lambda\|^2 - \|\lambda^0-\lambda\|^2\right)$$

$$+\frac{\theta}{2\rho}\sum_{k=0}^{t-1}E\|\lambda^{k+1}-\lambda^k\|^2 + \frac{\theta}{2\rho}E\|\tilde{\lambda}^{t+1}-\lambda^t\|^2$$

$$\leqslant (1-\theta)\left[F(x^0)-F(x)+(x^0-x)^\top(-A^\top\lambda^0)+\rho_x(x^0-x)^\top A^\top r^0\right]$$

$$-\frac{1}{2}\left[E\|x^{t+1}-x\|_{\tilde{P}}^2 - \|x^0-x\|_{\tilde{P}}^2\right.$$

$$\left.+\sum_{k=0}^{t}E\|x^k-x^{k+1}\|_{\tilde{P}}^2\right] + \frac{L_f}{2}\sum_{k=0}^{t}E\|x^k-x^{k+1}\|^2$$

$$+\frac{\theta L_g}{2}\sum_{k=0}^{t-1}E\|y^k-y^{k+1}\|^2 + \frac{L_g}{2}E\|\tilde{y}^{t+1}-y^t\|^2. \tag{B.14}$$

Note that for $k \leqslant t-1$,

$$-\frac{\rho_x}{\rho}(\lambda^k-\lambda^{k+1})^\top B(y^{k+1}-y^k) + \frac{\theta}{2\rho}\|\lambda^{k+1}-\lambda^k\|^2 \geqslant -\frac{\rho}{2\theta^3}\|y^{k+1}-y^k\|_{B^\top B}^2$$

and

$$-(\lambda^t-\tilde{\lambda}^{t+1})^\top B(\tilde{y}^{t+1}-y^t) + \frac{\theta}{2\rho}\|\tilde{\lambda}^{t+1}-\lambda^t\|^2 \geqslant -\frac{\rho}{2\theta}\|\tilde{y}^{t+1}-y^t\|_{B^\top B}^2.$$

Because $\tilde{P}$, $\tilde{Q}$ and ρ satisfy (3.13), we have from (B.14) that

$$E\left[\Phi(x^{t+1},\tilde{y}^{t+1})-\Phi(x,y)+(\tilde{w}^{t+1}-w)^\top H(\tilde{w}^{t+1})\right]$$

$$+\theta\sum_{k=0}^{t-1}E\left[\Phi(x^{k+1},y^{k+1})-\Phi(x,y)+(w^{k+1}-w)^\top H(w^{k+1})\right]$$

$$\leqslant (1 - \theta) \left[F(x^0) - F(x) + (x^0 - x)^\top (-A^\top \lambda^0) + \rho_x (x^0 - x)^\top A^\top r^0 \right]$$
$$+ \frac{1}{2} \|x^0 - x\|_{\tilde{P}}^2 + \frac{\theta}{2} \|y^0 - y\|_{\tilde{Q}}^2 + \frac{\rho}{2\theta} \|y^0 - y\|_{B^\top B}^2 + \frac{\theta}{2\rho} E \|\lambda^0 - \lambda\|^2.$$

Similar to Theorem 3.8, from the convexity of Φ and (2.23), we have

$$(1 + \theta t) E \left[\Phi(\hat{x}^t, \hat{y}^t) - \Phi(x, y) + (\hat{w}^{t+1} - w)^\top H(w) \right]$$
$$\leqslant (1 - \theta) \left[F(x^0) - F(x) + (x^0 - x)^\top (-A^\top \lambda^0) + \rho_x (x^0 - x)^\top A^\top r^0 \right]$$
$$+ \frac{1}{2} \|x^0 - x\|_{\tilde{P}}^2 + \frac{\theta}{2} \|y^0 - y\|_{\tilde{Q}}^2 + \frac{\rho}{2\theta} \|y^0 - y\|_{B^\top B}^2 + \frac{\theta}{2\rho} E \|\lambda^0 - \lambda\|^2. \quad \text{(B.15)}$$

Noting $\lambda^0 = 0$ and $(x^0 - x)^\top A^\top r^0 \leqslant \frac{1}{2} \left[\|x^0 - x\|_{A^\top A} + \|r^0\|^2 \right]$, and using Lemmas 3.3 and 3.5 with $\gamma = \max\{0.5 + \|\lambda^*\|, 3\|\lambda^*\|\}$, we obtain the result (3.16).

B.3 Proof of Theorem 3.8

Using (3.6) and (3.7), applying (2.21) to the cross terms, and also noting the definition of $\tilde{P}$ and $\tilde{Q}$ in (B.1), we have

$$-\frac{\theta}{\rho} E \left[\sum_{k=0}^{t-1} (\lambda^{k+1} - \lambda)^\top (\lambda^{k+1} - \lambda^k) + (\tilde{\lambda}^{t+1} - \lambda)^\top (\tilde{\lambda}^{t+1} - \lambda^t) \right]$$
$$+ \sum_{k=0}^{t} \rho_x E(x^{k+1} - x)^\top A^\top B(y^{k+1} - y^k) + (1 - \theta) \sum_{k=0}^{t} \rho_y E(y^k - y)^\top B^\top A(x^{k+1} - x^k)$$
$$- \sum_{k=0}^{t} E(x^{k+1} - x)^\top \tilde{P}(x^{k+1} - x^k) + \frac{L_f}{2} \sum_{k=0}^{t} E \|x^k - x^{k+1}\|^2$$
$$- \sum_{k=0}^{t} E(y^{k+1} - y)^\top \tilde{Q}(y^{k+1} - y^k) + \frac{L_g}{2} \sum_{k=0}^{t} E \|y^k - y^{k+1}\|^2$$
$$= -\frac{\theta}{2\rho} E \left[\|\tilde{\lambda}^{t+1} - \lambda\|^2 - \|\lambda^0 - \lambda\|^2 + \sum_{k=0}^{t-1} \|\lambda^{k+1} - \lambda^k\|^2 + \|\tilde{\lambda}^{t+1} - \lambda^t\|^2 \right]$$
$$+ \frac{\rho_x}{\rho} \sum_{k=0}^{t} E(\lambda^k - \lambda^{k+1})^\top B(y^{k+1} - y^k) + \frac{(1 - \theta)\rho_y}{\rho} \sum_{k=0}^{t} E(\lambda^{k-1} - \lambda^k)^\top A(x^{k+1} - x^k)$$
$$- \frac{\theta\rho_y}{2} E \left(\|x^0 - x\|_{A^\top A}^2 - \|x^{t+1} - x\|_{A^\top A}^2 \right) + \frac{(2 - \theta)\rho_y}{2} \sum_{k=0}^{t} E \|x^{k+1} - x^k\|_{A^\top A}^2$$
$$- \frac{1}{2} E \left(\|x^{t+1} - x\|_{\hat{P}}^2 - \|x^0 - x\|_{\hat{P}}^2 + \sum_{k=0}^{t} \|x^{k+1} - x^k\|_{\hat{P}}^2 \right)$$
$$+ \frac{L_f}{2} \sum_{k=0}^{t} E \|x^k - x^{k+1}\|^2$$

$$-\frac{1}{2}E\left(\|y^{t+1} - y\|_{\hat{Q}}^2 - \|y^0 - y\|_{\hat{Q}}^2 + \sum_{k=0}^{t}\|y^{k+1} - y^k\|_{\hat{Q}}^2\right)$$

$$+\frac{L_g}{2}\sum_{k=0}^{t}E\|y^k - y^{k+1}\|^2, \tag{B.16}$$

where we have used the conditions in (3.17). By Young's inequality, we have that for $0 \leqslant k \leqslant t$,

$$\frac{\rho_x}{\rho}(\lambda^k - \lambda^{k+1})^\top B(y^{k+1} - y^k) - \frac{\theta}{2\rho}\frac{1}{2-\theta}\|\lambda^{k+1} - \lambda^k\|^2$$

$$\leqslant \frac{\rho}{\theta}\frac{2-\theta}{2}\frac{\rho_x^2}{\rho^2}\|B(y^{k+1} - y^k)\|^2 \overset{(3.17)}{=} \frac{(2-\theta)\rho_y}{2\theta^2}\|y^{k+1} - y^k\|_{B^\top B}^2, \tag{B.17}$$

and for $1 \leqslant k \leqslant t$,

$$\frac{(1-\theta)\rho_y}{\rho}(\lambda^{k-1} - \lambda^k)^\top A(x^{k+1} - x^k) - \frac{\theta}{2\rho}\frac{1-\theta}{2-\theta}\|\lambda^{k-1} - \lambda^k\|^2$$

$$\leqslant (1-\theta)\frac{\rho}{\theta}\frac{(2-\theta)\rho_y^2}{2\rho^2}\|A(x^{k+1} - x^k)\|^2$$

$$\overset{(3.17)}{=} \frac{(1-\theta)(2-\theta)}{2\theta^2}\rho_x\|x^{k+1} - x^k\|_{A^\top A}^2. \tag{B.18}$$

Plugging (B.17) and (B.18) and also noting $\|\tilde{\lambda}^{t+1} - \lambda^t\|^2 \geqslant \|\lambda^{t+1} - \lambda^t\|^2$, we can upper bound the right-hand side of (B.16) by

$$-\frac{\theta}{2\rho}E\left[\|\tilde{\lambda}^{t+1} - \lambda\|^2 - \|\lambda^0 - \lambda\|^2\right] - \frac{\theta\rho_y}{2}E\left(\|x^0 - x\|_{A^\top A}^2 - \|x^{t+1} - x\|_{A^\top A}^2\right)$$

$$+\left(\frac{(1-\theta)(2-\theta)}{2\theta^2}\rho_x + \frac{(2-\theta)\rho_y}{2}\right)\sum_{k=0}^{t}E\|x^{k+1} - x^k\|_{A^\top A}^2$$

$$+\frac{(2-\theta)\rho_y}{2\theta^2}\sum_{k=0}^{t}E\|y^{k+1} - y^k\|_{B^\top B}^2$$

$$-\frac{1}{2}E\left(\|x^{t+1} - x\|_{\hat{P}}^2 - \|x^0 - x\|_{\hat{P}}^2\right.$$

$$+\sum_{k=0}^{t}\|x^{k+1} - x^k\|_{\hat{P}}^2\right) + \frac{L_f}{2}\sum_{k=0}^{t}E\|x^k - x^{k+1}\|^2$$

$$-\frac{1}{2}E\left(\|y^{t+1} - y\|_{\hat{Q}}^2 - \|y^0 - y\|_{\hat{Q}}^2 + \sum_{k=0}^{t}\|y^{k+1} - y^k\|_{\hat{Q}}^2\right)$$

$$+\frac{L_g}{2}\sum_{k=0}^{t}E\|y^k - y^{k+1}\|^2$$

$$\overset{(3.18a)}{\leqslant} \frac{1}{2}\left(\|x^0 - x\|_{\hat{P}-\theta\rho_x A^\top A}^2 + \|y^0 - y\|_{\hat{Q}}^2\right) + \frac{\theta}{2\rho}E\|\lambda^0 - \lambda\|^2. \tag{B.19}$$

In addition, note that

$$\theta \|x^{t+1} - x\|^2_{A^\top A} = \frac{n}{N} \left\| \sum_{i=1}^{N} A_i (x_i^{t+1} - x_i) \right\|^2 \leqslant n \sum_{i=1}^{N} \|x_i^{t+1} - x_i\|^2_{A_i^\top A_i},$$

$$\|x^k - x^{k+1}\|^2_{A^\top A} = \left\| \sum_{i \in I_k} A_i (x_i^k - x_i^{k+1}) \right\|^2 \leqslant n \sum_{i=1}^{N} \|x_i^k - x_i^{k+1}\|^2_{A_i^\top A_i},$$

$$\|y^k - y^{k+1}\|^2_{B^\top B} = \left\| \sum_{j \in J_k} B_j (y_j^k - y_j^{k+1}) \right\|^2 \leqslant m \sum_{j=1}^{M} \|y_j^k - y_j^{k+1}\|_{B_j^\top B_j}.$$

Hence, if $\hat{P}$ and $\hat{Q}$ satisfy (3.18b), then (B.19) also holds.

Combining (B.6), (B.16), and (B.19) yields

$$E\left[\Phi(x^{t+1}, y^{t+1}) - \Phi(x, y) + (\tilde{w}^{t+1} - w)^\top H(\tilde{w}^{t+1})\right]$$

$$+\theta \sum_{k=0}^{t-1} E\left[\Phi(x^{k+1}, y^{k+1}) - \Phi(x, y) + (w^{k+1} - w)^\top H(w^{k+1})\right]$$

$$\leqslant (1-\theta)\left[\Phi(x^0, y^0) - \Phi(x, y)\right]$$

$$+(1-\theta)\left[(x^0 - x)^\top(-A^\top \lambda^0) + \rho_x(x^0 - x)^\top A^\top r^0 + (y^0 - y)^\top(-B^\top \lambda^0)\right.$$

$$\left.+\rho_y(y^0 - y)^\top B^\top r^0\right]$$

$$+\frac{1}{2}\left(\|x^0 - x\|^2_{\hat{P} - \theta \rho_x A^\top A} + \|y^0 - y\|^2_{\hat{Q}}\right) + \frac{\theta}{2\rho} E\|\lambda^0 - \lambda\|^2. \tag{B.20}$$

Applying the convexity of Φ and the properties (2.23) of H, we have

$$(1 + \theta t)E\left[\Phi(\hat{x}^t, \hat{y}^t) - \Phi(x, y) + (\hat{w}^{t+1} - w)^\top H(w)\right]$$

$$\overset{(2.20)}{=} (1 + \theta t)E\left[\Phi(\hat{x}^t, \hat{y}^t) - \Phi(x, y) + (\hat{w}^{t+1} - w)^\top H(\hat{w}^{t+1})\right]$$

$$\overset{(2.23)}{\leqslant} E\left[\Phi(x^{t+1}, y^{t+1}) - \Phi(x, y) + (\tilde{w}^{t+1} - w)^\top H(\tilde{w}^{t+1})\right]$$

$$+\theta \sum_{k=0}^{t-1} E\left[\Phi(x^{k+1}, y^{k+1}) - \Phi(x, y) + (w^{k+1} - w)^\top H(w^{k+1})\right]. \tag{B.21}$$

Now combining (B.21) and (B.20), we have

$$(1 + \theta t)E\left[\Phi(\hat{x}^t, \hat{y}^t) - \Phi(x, y) + (\hat{w}^{t+1} - w)^\top H(w)\right]$$

$$\leqslant (1-\theta)\left[\Phi(x^0, y^0) - \Phi(x, y)\right]$$

$$+(1-\theta)\left[(x^0-x)^\top(-A^\top\lambda^0)+\rho_x(x^0-x)^\top A^\top r^0+(y^0-y)^\top(-B^\top\lambda^0)\right.$$

$$\left.+\rho_y(y^0-y)^\top B^\top r^0\right]$$

$$+\frac{1}{2}\left(\|x^0-x\|^2_{\hat{P}-\theta\rho_x A^\top A}+\|y^0-y\|^2_{\hat{Q}}\right)+\frac{\theta}{2\rho}E\|\lambda^0-\lambda\|^2. \tag{B.22}$$

By Lemmas 3.3 and 3.5 with $\gamma=\max\{0.5+\|\lambda^*\|,3\|\lambda^*\|\}$, we have the desired result.

B.4 Proof of Theorem 4.2

From the nonincreasing monotonicity of α_k, one can easily show the following result.

Lemma B.1 *Assume* $\lambda^{-1}=\lambda^0$*. It holds that*

$$\sum_{k=0}^{t}\frac{(1-\theta)\beta_k}{2}\left[\|\lambda^k-\lambda\|^2-\|\lambda^{k-1}-\lambda\|^2+\|\lambda^k-\lambda^{k-1}\|^2\right]$$

$$-\sum_{k=0}^{t-1}\frac{\alpha_k\beta_{k+1}}{2\alpha_{k+1}}\left[\|\lambda^{k+1}-\lambda\|^2-\|\lambda^k-\lambda\|^2+\|\lambda^{k+1}-\lambda^k\|^2\right]$$

$$\leqslant-\sum_{k=0}^{t-1}\frac{\beta_{k+1}}{2}\|\lambda^{k+1}-\lambda^k\|^2+\sum_{k=1}^{t}\frac{(1-\theta)\beta_k}{2}\|\lambda^k-\lambda^{k-1}\|^2$$

$$-\sum_{k=0}^{t-1}\frac{\alpha_k\beta_{k+1}}{2\alpha_{k+1}}\|\lambda^{k+1}-\lambda\|^2-\sum_{k=1}^{t}\frac{(1-\theta)\beta_k}{2}\|\lambda^{k-1}-\lambda\|^2$$

$$+\frac{\alpha_0\beta_1}{2\alpha_1}\|\lambda^0-\lambda\|^2+\sum_{k=1}^{t-1}\frac{\alpha_k\beta_{k+1}}{2\alpha_{k+1}}\|\lambda^k-\lambda\|^2+\sum_{k=1}^{t}\frac{(1-\theta)\beta_k}{2}\|\lambda^k-\lambda\|^2$$

$$=-\sum_{k=0}^{t-1}\frac{\theta\beta_{k+1}}{2}\|\lambda^{k+1}-\lambda^k\|^2+\left(\frac{\alpha_0\beta_1}{2\alpha_1}-\frac{(1-\theta)\beta_1}{2}\right)\|\lambda^0-\lambda\|^2$$

$$-\left(\frac{\alpha_{t-1}\beta_t}{2\alpha_t}-\frac{(1-\theta)\beta_t}{2}\right)\|\lambda^t-\lambda\|^2$$

$$-\sum_{k=1}^{t-1}\left(\frac{\alpha_{k-1}\beta_k}{2\alpha_k}+\frac{(1-\theta)\beta_{k+1}}{2}-\frac{\alpha_k\beta_{k+1}}{2\alpha_{k+1}}-\frac{(1-\theta)\beta_k}{2}\right)\|\lambda^k-\lambda\|^2. \tag{B.23}$$

By the update formula of λ in (4.5), we have from (4.8) that

$$E\left[F(x^{k+1})-F(x)+(x^{k+1}-x)^\top(-A^\top\lambda^{k+1})+(\lambda^{k+1}-\lambda)^\top r^{k+1}\right]$$

$$+E\left[\frac{(\lambda^{k+1}-\lambda)^\top(\lambda^{k+1}-\lambda^k)}{\left(1-\frac{(1-\theta)\alpha_{k+1}}{\alpha_k}\right)\rho}+\frac{(1-\theta)\alpha_{k+1}}{\alpha_k}\rho(x^{k+1}-x)^\top A^\top r^{k+1}\right]$$

$$+E(x^{k+1} - x)^\top \left(\tilde{P} + \frac{I}{\alpha_k}\right)(x^{k+1} - x^k) - \frac{L_f}{2}E\|x^k - x^{k+1}\|^2 + E(x^{k+1} - x^k)^\top \delta^k$$

$$\leqslant (1-\theta)E\left[F(x^k) - F(x) + (x^k - x)^\top(-A^\top\lambda^k) + (\lambda^k - \lambda)^\top r^k + \frac{(\lambda^k - \lambda)^\top(\lambda^k - \lambda^{k-1})}{\left(1 - \frac{(1-\theta)\alpha_k}{\alpha_{k-1}}\right)\rho}\right]$$

$$+(1-\theta)\rho E(x^k - x)^\top A^\top r^k, \tag{B.24}$$

where similar to (B.1), we have defined $\tilde{P} = \hat{P} - \rho A^\top A$.

Multiplying α_k to both sides of (B.24) and using (4.2) and (2.21), we have

$$\alpha_k E\left[F(x^{k+1}) - F(x) + (w^{k+1} - w)^\top H(w^{k+1})\right]$$

$$+\frac{\alpha_k \beta_{k+1}}{2\alpha_{k+1}}E\left[\|\lambda^{k+1} - \lambda\|^2 - \|\lambda^k - \lambda\|^2 + \|\lambda^{k+1} - \lambda^k\|^2\right]$$

$$+E\left[(1-\theta)\alpha_{k+1}\rho(x^{k+1} - x)^\top A^\top r^{k+1}\right]$$

$$+\frac{\alpha_k}{2}E\left[\|x^{k+1} - x\|_{\tilde{P}}^2 - \|x^k - x\|_{\tilde{P}}^2 + \|x^{k+1} - x^k\|_{\tilde{P}}^2\right]$$

$$+\frac{1}{2}E\left[\|x^{k+1} - x\|^2 - \|x^k - x\|^2 + \|x^{k+1} - x^k\|^2\right]$$

$$-\frac{\alpha_k L_f}{2}E\|x^k - x^{k+1}\|^2 + \alpha_k E(x^{k+1} - x^k)^\top \delta^k$$

$$\leqslant (1-\theta)\alpha_k E\left[F(x^k) - F(x) + (w^k - w)^\top H(w^k)\right]$$

$$+\frac{(1-\theta)\beta_k}{2}E\left[\|\lambda^k - \lambda\|^2 - \|\lambda^{k-1} - \lambda\|^2 + \|\lambda^k - \lambda^{k-1}\|^2\right]$$

$$+\alpha_k(1-\theta)\rho E(x^k - x)^\top A^\top r^k. \tag{B.25}$$

Denote $\tilde{\lambda}^{t+1} = \lambda^t - \rho r^{t+1}$. Then, for $k = t$, it is easy to see that (B.25) becomes

$$\alpha_t E\left[F(x^{t+1}) - F(x) + (\tilde{w}^{t+1} - w)^\top H(\tilde{w}^{t+1})\right]$$

$$+\frac{\alpha_t}{2\rho}E\left[\|\tilde{\lambda}^{t+1} - \lambda\|^2 - \|\lambda^t - \lambda\|^2 + \|\tilde{\lambda}^{t+1} - \lambda^t\|^2\right]$$

$$+\frac{\alpha_t}{2}E\left[\|x^{t+1} - x\|_{\tilde{P}}^2 - \|x^t - x\|_{\tilde{P}}^2 + \|x^{t+1} - x^t\|_{\tilde{P}}^2\right]$$

$$+\frac{1}{2}E\left[\|x^{t+1} - x\|^2 - \|x^t - x\|^2 + \|x^{t+1} - x^t\|^2\right]$$

$$-\frac{\alpha_t L_f}{2}E\|x^t - x^{t+1}\|^2 + \alpha_t E(x^{t+1} - x^t)^\top \delta^t$$

$$\leqslant (1-\theta)\alpha_t E\left[F(x^t) - F(x) + (w^t - w)^\top H(w^t)\right]$$

$$+\frac{(1-\theta)\beta_t}{2}E\left[\|\lambda^t - \lambda\|^2 - \|\lambda^{t-1} - \lambda\|^2 + \|\lambda^t - \lambda^{t-1}\|^2\right]$$

$$+\alpha_t(1-\theta)E\rho(x^t - x)^\top A^\top r^t. \tag{B.26}$$

By the nonincreasing monotonicity of α_k, summing (B.25) from $k = 0$ through $t - 1$ and (B.26) and plugging (B.23) give

$$
\alpha_t E\left[F(x^{t+1}) - F(x) + (\tilde{w}^{t+1} - w)^\top H(\tilde{w}^{t+1})\right]
$$

$$
+\theta\alpha_{k+1}\sum_{k=0}^{t-1} E\left[F(x^{k+1}) - F(x) + (w^{k+1} - w)^\top H(w^{k+1})\right]
$$

$$
+\frac{\alpha_t}{2\rho}E\left[\|\tilde{\lambda}^{t+1} - \lambda\|^2 - \|\lambda^t - \lambda\|^2 + \|\tilde{\lambda}^{t+1} - \lambda^t\|^2\right]
$$

$$
+\frac{\alpha_{t+1}}{2}E\|x^{t+1} - x\|_{\tilde{P}}^2 + \sum_{k=0}^{t}\frac{\alpha_k}{2}E\|x^{k+1} - x^k\|_{\tilde{P}}^2
$$

$$
+\frac{1}{2}E\left[\|x^{t+1} - x\|^2 - \|x^0 - x\|^2 + \sum_{k=0}^{t}\|x^{k+1} - x^k\|^2\right]
$$

$$
-\sum_{k=0}^{t}\frac{\alpha_k L_f}{2}E\|x^k - x^{k+1}\|^2 + \sum_{k=0}^{t}\alpha_k E(x^{k+1} - x^k)^\top \delta^k
$$

$$
\leqslant (1 - \theta)\alpha_0 E\left[F(x^0) - F(x) + (w^0 - w)^\top H(w^0)\right]
$$

$$
+\alpha_0(1 - \theta)\rho(x^0 - x)^\top A^\top r^0 + \frac{\alpha_0}{2}\|x^0 - x\|_{\tilde{P}}^2
$$

$$
-\sum_{k=0}^{t-1}\frac{\theta\beta_{k+1}}{2}E\|\lambda^{k+1} - \lambda^k\|^2 + \left(\frac{\alpha_0\beta_1}{2\alpha_1} - \frac{(1 - \theta)\beta_1}{2}\right)E\|\lambda^0 - \lambda\|^2
$$

$$
-\left(\frac{\alpha_{t-1}\beta_t}{2\alpha_t} - \frac{(1 - \theta)\beta_t}{2}\right)E\|\lambda^t - \lambda\|^2
$$

$$
-\sum_{k=1}^{t-1}\left(\frac{\alpha_{k-1}\beta_k}{2\alpha_k} + \frac{(1 - \theta)\beta_{k+1}}{2} - \frac{\alpha_k\beta_{k+1}}{2\alpha_{k+1}} - \frac{(1 - \theta)\beta_k}{2}\right)E\|\lambda^k - \lambda\|^2. \quad \text{(B.27)}
$$

From (4.9d), we have

$$
\frac{\alpha_t}{2\rho}\left[\|\tilde{\lambda}^{t+1} - \lambda\|^2 - \|\lambda^t - \lambda\|^2 + \|\tilde{\lambda}^{t+1} - \lambda^t\|^2\right]
$$

$$
\geqslant -\left(\frac{\alpha_{t-1}\beta_t}{2\alpha_t} - \frac{(1 - \theta)\beta_t}{2}\right)\|\lambda^t - \lambda\|^2.
$$

In addition, from Young's inequality, it holds that

$$
\frac{1}{2}\|x^{k+1} - x^k\|^2 + \alpha_k E(x^{k+1} - x^k)^\top \delta^k \geqslant \frac{\alpha_k^2}{2}\|\delta\|^2.
$$

Hence, dropping negative terms on the right-hand side of (B.27), from the convexity of Φ and (2.23), we have

$$\left(\alpha_{t+1} + \theta \sum_{k=1}^{t} \alpha_k\right) E\left[F(\hat{x}^t) - F(x) + (\hat{w}^t - w)^\top H(\hat{w}^t)\right]$$

$$\alpha_t E\left[F(x^{t+1}) - F(x) + (\tilde{w}^{t+1} - w)^\top H(\tilde{w}^{t+1})\right]$$

$$+\theta\alpha_{k+1} \sum_{k=0}^{t-1} E\left[F(x^{k+1}) - F(x) + (w^{k+1} - w)^\top H(w^{k+1})\right]$$

$$\leqslant (1-\theta)\alpha_0 \left[F(x^0) - F(x) + (w^0 - w)^\top H(w^0)\right]$$

$$+(1-\theta)\alpha_0\rho(x^0 - x)^\top A^\top r^0 + \frac{\alpha_0}{2}\|x^0 - x\|_{\tilde{P}}^2 + \frac{1}{2}\|x^0 - x\|^2$$

$$+\left(\frac{\alpha_0\beta_1}{2\alpha_1} - \frac{(1-\theta)\beta_1}{2}\right) E\|\lambda^0 - \lambda\|^2 + \sum_{k=0}^{t} \frac{\alpha_k^2}{2} E\|\delta^k\|^2. \tag{B.28}$$

Using Lemma 3.3 and the properties of H, we derive the desired result.

B.5 Proof of Proposition 6.1

Let $(I + \partial\phi)^{-1}(x) := \arg\min_z \phi(z) + \frac{1}{2}\|z - x\|_2^2$ denote the proximal mapping of ϕ at x. Then, the update in (1.9b) can be written to

$$z^{k+1} = \left(I + \partial\left(\frac{g^*}{\eta}\right)\right)^{-1} \left(z^k - \frac{1}{\eta}Ax^{k+1}\right).$$

Define y^{k+1} as that in (6.7b). Then,

$$\frac{1}{\eta}y^{k+1} = \frac{1}{\eta}\left\{\arg\min_y g(y) - \langle y, z^k \rangle + \frac{1}{2\eta}\|y + Ax^{k+1}\|^2\right\}$$

$$= \frac{1}{\eta}\left\{\arg\min_y g(y) + \frac{\eta}{2}\|\frac{1}{\eta}y - (z^k - \frac{1}{\eta}Ax^{k+1})\|^2\right\}$$

$$= \arg\min_y g(\eta y) + \frac{\eta}{2}\|y - \left(z^k - \frac{1}{\eta}Ax^{k+1}\right)\|^2$$

$$= \left(I + \partial\left(\frac{1}{\eta}g(\eta\cdot)\right)\right)^{-1} \left(z^k - \frac{1}{\eta}Ax^{k+1}\right).$$

Hence, using the fact that the conjugate of $\frac{1}{\eta} g^*$ is $\frac{1}{\eta} g(\eta \cdot)$ and the Moreau's identity $(I + \partial \phi)^{-1} + (I + \partial \phi^*)^{-1} = I$ for any convex function ϕ, we have

$$\left(I + \partial \left(\frac{g^*}{\eta} \right) \right)^{-1} \left(z^k - \frac{1}{\eta} A x^{k+1} \right)$$
$$+ \left(I + \partial \left(\frac{1}{\eta} g(\eta \cdot) \right) \right)^{-1} \left(z^k - \frac{1}{\eta} A x^{k+1} \right) = z^k - \frac{1}{\eta} A x^{k+1}.$$

Therefore, (6.7c) holds, and thus from (1.9c) it follows

$$\bar{z}^{k+1} = z^{k+1} - \frac{q}{\eta}(A x^{k+1} + y^{k+1}).$$

Substituting the formula of $\bar{z}^k$ into (1.9a), we have for $i = i_k$,

$$\begin{aligned}
x_i^{k+1} &= \arg \min_{x_i \in \mathcal{X}_i} \langle -\bar{z}^k, A_i x_i \rangle + u_i(x_i) + \frac{\tau}{2} \| x_i - x_i^k \|^2 \\
&= \arg \min_{x_i \in \mathcal{X}_i} \langle -z^k, A_i x_i \rangle + \frac{q}{\eta} \langle A x^k + y^k, A_i x_i \rangle + u_i(x_i) + \frac{\tau}{2} \| x_i - x_i^k \|^2 \\
&= \arg \min_{x_i \in \mathcal{X}_i} \langle -z^k, A_i x_i \rangle + u_i(x_i) + \frac{q}{2\eta} \| A_i x_i + A_{\neq i} x_{\neq i}^k + y^k \|^2 \\
&\quad + \frac{1}{2} \| x_i - x_i^k \|_{\tau I - \frac{q}{\eta} A_i^\top A_i},
\end{aligned}$$

which is exactly (6.7a). Hence, we complete the proof.

References

[1] James, G.M., Paulson, C., Rusmevichientong, P.: The constrained lasso. Technical report. University of Southern California (2013)
[2] Rockafellar, R.T.: Large-scale extended linear-quadratic programming and multistage optimization. Advances in Numerical Partial Differential Equations and Optimization: In: Proceedings of the Fifth Mexico-United States Workshop, vol. 2, pp. 247–261 (1991)
[3] Hong, M., Chang, T.-H., Wang, X., Razaviyayn, M., Ma, S., Luo, Z.-Q.: A block successive upper bound minimization method of multipliers for linearly constrained convex optimization (2014). arXiv:1401.7079
[4] Cui, Y., Li, X., Sun, D., Toh, K.-C.: On the convergence properties of a majorized ADMM for linearly constrained convex optimization problems with coupled objective functions. J. Optim. Theory Appl. **169**(3), 1013–1041 (2016)
[5] Chen, C., Li, M., Liu, X., Ye, Y.: On the convergence of multi-block alternating direction method of multipliers and block coordinate descent method (2015). arXiv:1508.00193
[6] Gao, X., Zhang, S.: First-order algorithms for convex optimization with nonseparate objective and coupled constraints. J. Oper. Res. Soc. China **5**(2), 131–159 (2017)
[7] Glowinski, R., Marrocco, A.: Sur l'approximation, par eléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de dirichlet non linéaires. ESAIM Math. Model. Numer. Anal. **9**(R2), 41–76 (1975)
[8] Gabay, D., Mercier, B.: A dual algorithm for the solution of nonlinear variational problems via finite element approximation. Comput. Math. Appl. **2**(1), 17–40 (1976)

[9] Glowinski, R., Le Tallec, P.: Augmented Lagrangian and Operator-Splitting Methods in Nonlinear Mechanics, vol. 9. SIAM, Philadelphia (1989)

[10] Eckstein, J., Bertsekas, D.: On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators. Math. Program. **55**(1–3), 293–318 (1992)

[11] He, B., Yuan, X.: On the $O(1/n)$ convergence rate of the Douglas-Rachford alternating direction method. SIAM J. Numer. Anal. **50**(2), 700–709 (2012)

[12] Monteiro, R.D., Svaiter, B.F.: Iteration-complexity of block-decomposition algorithms and the alternating direction method of multipliers. SIAM J. Optim. **23**(1), 475–507 (2013)

[13] Deng, W., Yin, W.: On the global and linear convergence of the generalized alternating direction method of multipliers. J. Sci. Comput. **66**(3), 889–916 (2015)

[14] Lin, T., Ma, S., Zhang, S.: On the sublinear convergence rate of multi-block admm. J. Oper. Res. Soc. China **3**(3), 251–274 (2015)

[15] Hong, M., Luo, Z.: On the linear convergence of the alternating direction method of multipliers. Math. Program. **162**, 165–199 (2017)

[16] Boley, D.: Local linear convergence of the alternating direction method of multipliers on quadratic or linear programs. SIAM J. Optim. **23**(4), 2183–2207 (2013)

[17] Peng, Y., Ganesh, A., Wright, J., Xu, W., Ma, Y.: Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images. IEEE Trans. Pattern Anal. Mach. Intell. **34**(11), 2233–2246 (2012)

[18] Tao, M., Yuan, X.: Recovering low-rank and sparse components of matrices from incomplete and noisy observations. SIAM J. Optim. **21**(1), 57–81 (2011)

[19] Chen, C., He, B., Ye, Y., Yuan, X.: The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent. Math. Program. **155**(1–2), 57–79 (2016)

[20] Deng, W., Lai, M.-J., Peng, Z., Yin, W.: Parallel multi-block ADMM with $o(1/k)$ convergence. J. Sci. Comput. **71**, 1–25 (2016)

[21] He, B., Tao, M., Yuan, X.: Convergence rate and iteration complexity on the alternating direction method of multipliers with a substitution procedure for separable convex programming. Math. Oper. Res. **42**(3), 662–691 (2017)

[22] He, B., Hou, L., Yuan, X.: On full Jacobian decomposition of the augmented Lagrangian method for separable convex programming. SIAM J. Optim. **25**(4), 2274–2312 (2015)

[23] He, B., Tao, M., Yuan, X.: Alternating direction method with gaussian back substitution for separable convex programming. SIAM J. Optim. **22**(2), 313–340 (2012)

[24] Xu, Y.: Hybrid Jacobian and Gauss–Seidel proximal block coordinate update methods for linearly constrained convex programming. SIAM J. Optim. **28**(1), 646–670 (2018)

[25] Chen, C., Shen, Y., You, Y.: On the convergence analysis of the alternating direction method of multipliers with three blocks. In: Abstract and Applied Analysis, vol. 2013. Hindawi Publishing Corporation (2013)

[26] Cai, X., Han, D., Yuan, X.: On the convergence of the direct extension of ADMM for three-block separable convex minimization models with one strongly convex function. Comput. Optim. Appl. **66**(1), 39–73 (2017)

[27] Lin, T., Ma, S., Zhang, S.: On the global linear convergence of the admm with multiblock variables. SIAM J. Optim. **25**(3), 1478–1497 (2015)

[28] Li, M., Sun, D., Toh, K.-C.: A convergent 3-block semi-proximal ADMM for convex minimization problems with one strongly convex block. Asia-Pac. J. Oper. Res. **32**(04), 1550024 (2015)

[29] Han, D., Yuan, X.: A note on the alternating direction method of multipliers. J. Optim. Theory Appl. **155**(1), 227–238 (2012)

[30] Wang, Y., Yin, W., Zeng, J.: Global convergence of ADMM in nonconvex nonsmooth optimization (2015). arXiv:1511.06324

[31] Lin, T., Ma, S., Zhang, S.: Iteration complexity analysis of multi-block admm for a family of convex minimization without strong convexity. J. Sci. Comput. **69**, 1–30 (2016)

[32] Chen, L., Sun, D., Toh, K.-C.: An efficient inexact symmetric gauss-seidel based majorized ADMM for high-dimensional convex composite conic programming. Math. Program. **161**, 237–270 (2017)

[33] Li, X., Sun, D., Toh, K.-C.: A schur complement based semi-proximal ADMM for convex quadratic conic programming and extensions. Math. Program. **155**(1–2), 333–373 (2016)

[34] Sun, D., Toh, K.-C., Yang, L.: A convergent 3-block semiproximal alternating direction method of multipliers for conic programming with 4-type constraints. SIAM J. Optim. **25**(2), 882–915 (2015)

[35] Sun, R., Luo, Z.-Q., Ye, Y.: On the expected convergence of randomly permuted ADMM (2015). arXiv:1503.06387

[36] Dang, C., Lan, G.: Randomized first-order methods for saddle point optimization (2014). arXiv:1409.8625

[37] Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. J. Math. Imaging Vis. **40**(1), 120–145 (2011)

[38] Peng, Z., Wu, T., Xu, Y., Yan, M., Yin, W.: Coordinate friendly structures, algorithms and applications. Ann. Math. Sci. Appl. **1**(1), 57–119 (2016)

[39] Gao, X., Jiang, B., Zhang, S.: On the information-adaptive variants of the ADMM: an iteration complexity perspective. J. Sci. Comput. **76**(1), 327–363 (2018)

[40] Xu, Y., Zhang, S.: Accelerated primal-dual proximal block coordinate updating methods for constrained convex optimization. Comput. Optim. Appl. **70**(1), 91–128 (2018)

[41] Dang, C.D., Lan, G.: Stochastic block mirror descent methods for nonsmooth and stochastic optimization. SIAM J. Optim. **25**(2), 856–881 (2015)

[42] Xu, Y., Yin, W.: Block stochastic gradient iteration for convex and nonconvex optimization. SIAM J. Optim. **25**(3), 1686–1716 (2015)

[43] Grant, M., Boyd, S., Ye, Y.: CVX: Matlab software for disciplined convex programming, version 2.0 beta. http://cvxr.com/cvx (2013)

[44] Nesterov, Y.: Efficiency of coordinate descent methods on huge-scale optimization problems. SIAM J. Optim. **22**(2), 341–362 (2012)

[45] Richtárik, P., Takáč, M.: Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. Math. Program. **144**(1–2), 1–38 (2014)

[46] Lu, Z., Xiao, L.: On the complexity analysis of randomized block-coordinate descent methods. Math. Program. **152**(1–2), 615–642 (2015)

[47] Richtárik, P., Takáč, M.: Parallel coordinate descent methods for big data optimization. Math. Program. **156**, 1–52 (2015)