STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **8**, September 2020, pp 656–667.
Published online in International Academic Press (www.IAPress.org)

Convergence Analysis of a Stochastic Progressive Hedging Algorithm for Stochastic Programming

Zhenguo Mu [1,*], Junfeng Yang[2,*]

[1] *College of Science, Nanjing University of Posts and Telecommunications, China*
[2] *Department of Mathematics, Nanjing University, China*

Abstract Stochastic programming is an approach for solving optimization problems with uncertain data whose probability distribution is assumed to be known, and *progressive hedging algorithm* (PHA) is a well-known decomposition method for solving the underlying model. However, the per iteration computation of PHA could be very costly since it solves a large number of subproblems corresponding to all the scenarios. In this paper, a stochastic variant of PHA is studied. At each iteration, only a small fraction of the scenarios are selected uniformly at random and the corresponding variable components are updated accordingly, while the variable components corresponding to those not selected scenarios are kept untouch. Therefore, the per iteration cost can be controlled freely to achieve very fast iterations. We show that, though the per iteration cost is reduced significantly, the proposed stochastic PHA converges in an ergodic sense at the same sublinear rate as the original PHA.

Keywords Progressive hedging algorithm, stochastic programming, nonanticipativity, alternating direction method of multipliers.

AMS 2010 subject classifications 65K05, 65K10, 65J22, 90C25

DOI:10.19139/soic-2310-5070-964

1. Introduction

Stochastic programming (SP) and robust optimization are two main approaches to deal with optimization problems with uncertain data. Unlike robust optimization, which usually assumes the unknown data lies in certain region, SP takes advantage of the fact that the probability distribution of the uncertain data are known or can be estimated. Therefore, SP is frequently applied to settings in which decisions are made repeatedly in essentially the same circumstances, and the goal is to make decisions that perform well on average. It has many applications in finance, portfolio investment, energy optimization, wireless communication, machine learning, etc., see the monographs [1, 17].

We now review very briefly the multistage convex SP model introduced by Rockafellar and Wets [15]. In an N-stage decision problem, decisions are made and uncertain information is revealed alternatingly and sequentially as follows

$$x_1 \to \xi_1 \to x_2 \to \xi_2 \to \cdots \to \xi_{N-1} \to x_N \to \xi_N.$$

Here, for each j, x_j denotes the decision to be made at stage j, and ξ_j represents the uncertain information revealed after decision x_j and before x_{j+1}. SP aims at determining an optimal response to the information available at each

*Correspondence to: Zhenguo Mu (Email: mzg@njupt.edu.cn), College of Science, Nanjing University of Posts and Telecommunications, Nanjing, 210023, China; Junfeng Yang (Email: jfyang@nju.edu.cn), Department of Mathematics, Nanjing University, 22 Hankou Road, Nanjing, 210093, China

ISSN 2310-5070 (online) ISSN 2311-004X (print)

stage, resulting a decision sequence essentially of the form

$$x_1 \to x_2(\xi_1) \to x_3(\xi_1, \xi_2) \to \cdots \to x_N(\xi_1, \ldots, \xi_{N-1}).$$

Note that each x_j is dependent on $\{\xi_i : i < j\}$ but not on $\{\xi_i : i \geq j\}$ since the later is not yet available at the time the decision x_j is to be made. This restriction on the decision is called nonanticipativity constraint, which plays an important role in SP. For each j, we denote the set of all possibilities of ξ_j by Ξ_j. The set of all scenarios is then given by

$$\Xi = \otimes_{j=1}^{N} \Xi_j = \left\{ \xi = \left(\xi_1, \xi_2, \ldots, \xi_N\right) \mid \xi_j \in \Xi_j, j = 1, 2, \ldots, N \right\}.$$

Throughout this paper, we assume that Ξ has a finite cardinality and the probability distribution of Ξ is given, i.e., each $\xi \in \Xi$ has a known probability $p(\xi)$ such that $\sum_{\xi \in \Xi} p(\xi) = 1$.

Assume that $x_j \in \Re^{n_j}$ and let $n := \sum_{j=1}^{N} n_j$. A response mapping $x(\cdot) : \Xi \to \Re^n$ is given by

$$x(\cdot): \quad \xi \longrightarrow x(\xi) = (x_1(\xi), x_2(\xi), \ldots, x_N(\xi)) \in \Re^{n_1} \times \Re^{n_2} \times \cdots \times \Re^{n_N} = \Re^n, \ \forall \xi \in \Xi.$$

The set of all response mappings from Ξ to $\Re^n$, which forms a linear space, is denoted by $\mathcal{L}$. For convenience, we endow $\mathcal{L}$ with the following weighted inner product

$$\langle x(\cdot), y(\cdot) \rangle = \sum_{\xi \in \Xi} p(\xi) \sum_{j=1}^{N} \langle x_j(\xi), y_j(\xi) \rangle, \quad \forall\, x(\cdot), y(\cdot) \in \mathcal{L}, \tag{1.1}$$

where $\langle x_j(\xi), y_j(\xi) \rangle = x_j(\xi)^T y_j(\xi)$ denotes the usual inner product in $\Re^{n_j}$. The nonanticipativity constraint on the decision mapping is denoted by

$$\mathcal{N} = \{x(\cdot) \in \mathcal{L} \mid x_j(\xi) \text{ does not depend on } \xi_j, \ldots, \xi_N, j = 1, 2, \ldots, N\}.$$

It is easy to show that $\mathcal{N}$ forms a subspace of $\mathcal{L}$ and its orthogonal complement space is given by

$$\mathcal{N}^{\perp} = \{y(\cdot) = (y_1(\cdot), \ldots, y_N(\cdot)) \in \mathcal{L} \mid \mathbb{E}_{\xi|\xi_1,\ldots,\xi_{j-1}} y_j(\xi) = 0 \text{ for all } j\},$$

where $\mathbb{E}_{\xi|\xi_1,\ldots,\xi_{j-1}} y_j(\xi)$ denotes the conditional expectation of $y_j(\xi)$ given $\xi_1, \ldots, \xi_{j-1}$.

For each $\xi \in \Xi$, let $C(\xi)$ be a nonempty closed convex subset of $\Re^n$ and define

$$\mathcal{C} = \{x(\cdot) \in \mathcal{L} \mid x(\xi) \in C(\xi), \forall\, \xi \in \Xi\}.$$

Besides the nonanticipativity constraint, we assume that the decision mapping has the abstract constraint $x(\cdot) \in \mathcal{C}$. This constraint on response mapping can be specified in particular applications. In this paper, we are interested in the *progressive hedging algorithm* (PHA) and the convergence analysis of a stochastic variant of it, and it is not necessary to go beyond an abstract representation of the constraint on the response mapping. We assume that $\mathcal{C} \cap \mathcal{N} \neq \emptyset$.

For each $\xi \in \Xi$, we let $g(\cdot, \xi) : C(\xi) \to \Re$ be a lower semicontinuous convex function. An alternating sequence of decisions and observations $(x_1, \xi_1, x_2, \xi_2, \ldots, x_N, \xi_N)$ will result in a cost $g(x, \xi)$, which depends on both x and ξ, and each choice of $x(\cdot) \in \mathcal{L}$ yields a function from Ξ to $\Re$, i.e.,

$$g(x(\cdot), \cdot): \quad \xi \to g(x(\xi), \xi), \quad \xi \in \Xi.$$

Clearly, $g(x(\cdot), \cdot)$ can be regarded as a random variable. A commonly used risk measure in SP is the expected cost $\mathcal{G} : \mathcal{L} \to \Re$ given by

$$\mathcal{G}(x(\cdot)) = \mathbb{E}_{\xi}[g(x(\xi), \xi)] = \sum\nolimits_{\xi \in \Xi} p(\xi) g(x(\xi), \xi).$$

Here $\mathbb{E}_{\xi}$ denotes the expectation taken with respect to ξ. The classical SP model with nonanticipativity constraint originally proposed by Rockafellar and Wets in [15] aims to find $x(\cdot) \in \mathcal{C} \cap \mathcal{N}$ such that the expected cost $\mathcal{G}(x(\cdot))$

is minimized, namely

$$\min_{x(\cdot)\in\mathcal{L}} \{\mathcal{G}(x(\cdot)) \mid x(\cdot) \in \mathcal{C} \cap \mathcal{N}\}. \tag{1.2}$$

An important property of the expected cost is that it is completely separable with respect to scenarios, which often fulfills an important requirement in designing decomposition algorithms, e.g., PHA. By using this risk measure, it is also implied that the decision maker is risk neutral since a response mapping that performs well on average is pursued. Risk averse measures are also popular in SP, e.g., the conditional value-at-risk measure. However, the separability feature is lost in that case and some additional efforts are required to reactivate PHA, see the recent work [12].

We organize this paper as follows. In Section 2, we state the connections between the PHA and the classical *alternating direction method of multipliers* (ADMM). A stochastic variant of PHA is proposed in Section 3 and the main convergence results are given in Section 4. A summary is given in Section 5.

2. PHA and its connection with ADMM

PHA was originally introduced in [15] for multistage convex SP and has recently been extended in [13] to multistage stochastic monotone linear complementarity and variational inequality problems. It was shown in [15, Theorem 5.1] and [13, Theorem 1] that PHA is a customized application of the proximal point algorithm [7, 9]. In this section, we first review very briefly the PHA for SP. In particular, we show that PHA for SP problem (1.2) is an application of the classical ADMM for linearly constrained separable convex optimization. This explanation serves as a basis of the stochastic PHA which can be viewed essentially as a randomized variant of ADMM.

Let $\mathcal{P}_{\mathcal{N}}$ and $\mathcal{P}_{\mathcal{N}^\perp}$ be the projection operators onto $\mathcal{N}$ and $\mathcal{N}^\perp$, respectively, under the norm induced by the weighted inner product (1.1). The PHA for convex SP (1.2) is summarized below.

Algorithm 1 (PHA for SP (1.2), [15, page 8], [13, Eq. (2.3)]). *Given $\beta > 0$. Initialize $x^0(\cdot) \in \mathcal{N}$, $w^k(\cdot) \in \mathcal{N}^\perp$, and set $k = 0$.*

1. *Solve for each $\xi \in \Xi$ the following convex optimization problem to obtain the unique solution $\hat{x}^k(\xi)$:*

$$\min_{x(\xi)\in C(\xi)} g(x(\xi),\xi) + w^k(\xi)^T x(\xi) + \frac{\beta}{2}\|x(\xi) - x^k(\xi)\|^2.$$

2. *Compute $x^{k+1}(\cdot) = \mathcal{P}_{\mathcal{N}}(\hat{x}^k(\cdot))$ and $w^{k+1}(\cdot) = w^k(\cdot) + \beta\mathcal{P}_{\mathcal{N}^\perp}(\hat{x}^k(\cdot))$.*
3. *Set $k = k + 1$ and repeat until convergence.*

Note that the norm in the minimand of Step 1 is the common 2-norm. By categorizing PHA as a proximal point algorithm, it is explained in [15, 13] that the sequence generated by PHA is contractive with respect to the solution set. Linear convergence rate results are also given in the linear quadratic case [15, Theorem 5.2], i.e., $C(\xi)$'s are polyhedron and $g(\cdot,\xi)$'s are convex quadratic functions.

To apply ADMM, we introduce an auxiliary variable $y(\cdot)$ and rewrite (1.2) as

$$\min_{x(\cdot),y(\cdot)} \mathcal{G}(y(\cdot)) \quad \text{s.t.} \quad x(\cdot) = y(\cdot),\ x(\cdot) \in \mathcal{N},\ y(\cdot) \in \mathcal{C}. \tag{2.1}$$

Although the variables are duplicated, the constraint $x(\cdot) \in \mathcal{C} \cap \mathcal{N}$ is now separated into $x(\cdot) \in \mathcal{N}$ and $y(\cdot) \in \mathcal{C}$. Moreover, the objective function, as well as the constraints $x(\cdot) = y(\cdot)$ and $y(\cdot) \in \mathcal{C}$ are now separable with respect to each scenario. This makes the classical ADMM [4, 2] for linearly constrained separable convex optimization problem readily applicable.

Let $\beta > 0$ be a penalty parameter and $w(\cdot) \in \mathcal{L}$ be a Lagrange multiplier. The augmented Lagrangian function associated with (2.1) is given by

$$\mathcal{L}_\beta(x(\cdot), y(\cdot), w(\cdot)) = \mathcal{G}(y(\cdot)) + \langle w(\cdot), y(\cdot) - x(\cdot)\rangle + \frac{\beta}{2}\|y(\cdot) - x(\cdot)\|^2.$$

Note that here the norm in the last term is the one induced by the weighted inner product (1.1). Given $x^0(\cdot) \in \mathcal{L}$ and $w^0(\cdot) \in \mathcal{L}$, the ADMM iterates, for $k = 0, 1, \ldots$, as

$$\hat{x}^k(\cdot) = \arg\min_{x(\cdot)\in\mathcal{C}} \mathcal{L}_\beta(x^k(\cdot), x(\cdot), w^k(\cdot)), \tag{2.2a}$$

$$x^{k+1}(\cdot) = \arg\min_{x(\cdot)\in\mathcal{N}} \mathcal{L}_\beta(x(\cdot), \hat{x}^k(\cdot), w^k(\cdot)), \tag{2.2b}$$

$$w^{k+1}(\cdot) = w^k(\cdot) + \beta(\hat{x}^k(\cdot) - x^{k+1}(\cdot)). \tag{2.2c}$$

The following theorem summarizes the well known fact that PHA is an application of ADMM. For completeness, we give a proof.

Theorem 2. *Assume that $w^0(\cdot) \in \mathcal{N}^\perp$. Then, the ADMM (2.2) generates exactly the same sequence of points as the PHA given in Algorithm 1, i.e., PHA is an application of ADMM.*

Proof
It is easy to observe that (2.2a) is separable with respect to each scenario $\xi \in \Xi$, and (2.2b) is no more than a projection onto $\mathcal{N}$. Thus, (2.2) is equivalent to

$$\hat{x}^k(\xi) = \arg\min_{x(\xi)\in C(\xi)} g(x(\xi), \xi) + w^k(\xi)^T x(\xi) + \frac{\beta}{2}\|x(\xi) - x^k(\xi)\|^2, \forall\, \xi \in \Xi, \tag{2.3a}$$

$$x^{k+1}(\cdot) = \mathcal{P}_{\mathcal{N}}(\hat{x}^k(\cdot) + \frac{1}{\beta} w^k(\cdot)), \tag{2.3b}$$

$$w^{k+1}(\cdot) = w^k(\cdot) + \beta(\hat{x}^k(\cdot) - x^{k+1}(\cdot)). \tag{2.3c}$$

Note that the norm in the minimand of (2.3a) is the common 2-norm. Since $w^0(\cdot) \in \mathcal{N}^\perp$, it follows from (2.3b) and (2.3c) that

$$x^{k+1}(\cdot) = \mathcal{P}_{\mathcal{N}}(\hat{x}^k(\cdot)) \text{ and } w^{k+1}(\cdot) = w^k(\cdot) + \beta\mathcal{P}_{\mathcal{N}^\perp}(\hat{x}^k(\cdot))$$

for all $k \geq 0$. Thus, the ADMM (2.3) generates exactly the same sequence of points as the PHA given in Algorithm 1, i.e., the PHA for multistage convex SP is an application of the ADMM. □

Compared with the classical augmented Lagrangian method [6, 8], which minimizes $\mathcal{L}_\beta(x(\cdot), y(\cdot), w(\cdot))$, for fixed $w(\cdot)$, jointly with $x(\cdot) \in \mathcal{L}$ and $y(\cdot) \in \mathcal{C}$, an important advantage of PHA/ADMM is that it decomposes this large and coupled subproblem into many smaller pieces. In particular, PHA needs to solve at each iteration the following subproblems

$$\min_{x(\xi)\in C(\xi)} g(x(\xi), \xi) + w^k(\xi)^T x(\xi) + \frac{\beta}{2}\|x(\xi) - x^k(\xi)\|^2, \quad \xi \in \Xi.$$

Although this step is fully separable, in cases when the cardinality of Ξ is large, even a single iteration of PHA could be very costly. Note that $|\Xi| = \Pi_{j=1}^N |\Xi_j|$, which can be very large when either N or some $|\Xi_j|$'s are large. On the other hand, the projection onto $\mathcal{N}$ is much cheaper. Given the above connection between PHA and ADMM, we propose in the next section a stochastic variant of PHA, which at each iteration selects a small fraction of ξ's in Ξ and updates the corresponding variables only. Since the number of selected scenarios can be controlled, the per iteration cost of this stochastic PHA can be significantly lower than that of the original algorithm.

3. A stochastic PHA

For clearness of presentation and analysis, we denote the decision variables by vectors instead of mappings. For this purpose, we now redefine some notation. For any positive integer q, we let $[q] = \{1, 2, \ldots, q\}$. Let $m = \prod_{j=1}^N |\Xi_j|$ be the cardinality of $|\Xi|$. For convenience, we rewrite Ξ as $\Xi = \{\xi^i = (\xi_1^i, \ldots, \xi_N^i) \mid i \in [m]\}$. Let

$x_{ij} = x_j(\xi^i) \in \Re^{n_j}$ be the decision made at the jth stage corresponding to the ith scenario ξ^i, $i \in [m]$ and $j \in [N]$. Define $\mathbf{x}_i = (x_{i1}, \ldots, x_{iN}) \in \Re^n$, $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_m) \in \Re^{mn}$ and $p_i = p(\xi^i)$ for $i \in [m]$. Let $g_i(\mathbf{y}_i) = g(y(\xi^i), \xi^i)$, $i \in [m]$, and $\tilde{g}(\mathbf{y}) = \sum_{i=1}^m \tilde{g}_i(\mathbf{y}_i)$ with $\tilde{g}_i(\mathbf{y}_i) = p_i g_i(\mathbf{y}_i)$ for all $i \in [m]$. Throughout this paper, we endow $\Re^{qn}$ with the usually inner product for $q \leq m-1$, while for $q = m$, i.e., $\Re^{mn}$, we endow it with the weighted inner product

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^m p_i \langle \mathbf{x}_i, \mathbf{y}_i \rangle \text{ for all } \mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_m) \in \Re^{mn} \text{ and } \mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_m) \in \Re^{mn}. \tag{3.1}$$

The norm $\|\cdot\|$ always denotes the induced norm by the corresponding endowed inner product. With all the above notation, (2.1) can be simplified to

$$\begin{aligned} \min_{\mathbf{x},\mathbf{y} \in \Re^{mn}} \ & \tilde{g}(\mathbf{y}), \\ \text{s.t.} \ & \mathbf{x} = \mathbf{y}, \\ & \mathbf{x} \in \mathcal{N}, \\ & \mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_m) \in \mathcal{C} = C_1 \times \cdots \times C_m, \end{aligned} \tag{3.2}$$

where $\mathcal{N}$ is a subspace of $\Re^{mn}$ corresponding to the nonanticipativity constraint and each $C_i = C(\xi^i)$ is a closed convex set in $\Re^n$. Let $\beta > 0$ and $\mathbf{w} \in \Re^{mn}$. The augmented Lagrange function of (3.2) is

$$\mathcal{L}_\beta(\mathbf{y}, \mathbf{x}, \mathbf{w}) = \tilde{g}(\mathbf{y}) + \langle \mathbf{w}, \mathbf{y} - \mathbf{x} \rangle + \frac{\beta}{2} \|\mathbf{y} - \mathbf{x}\|^2.$$

Note that here the inner product in the middle term is that defined in (3.1) and the norm in the last term is the corresponding induced norm. Based on the equivalence of PHA and ADMM established in Section 2, we propose to solve (3.2) by the following stochastic variant of PHA, which can essentially be viewed as a randomized variant of ADMM.

Algorithm 3 (Stochastic PHA for SP (1.2))**.** *Let $\beta > 0$ and $0 < \theta < 1$ be such that θm is an integer. Define $\rho = \theta\beta$. Initialize $\mathbf{x}^0 \in \Re^{mn}$, $\mathbf{y}^0 \in \Re^{mn}$, $\mathbf{w}^0 \in \mathcal{N}^\perp$, and set $k = 0$.*

1. *Generate $S_k \subseteq [m]$ uniformly at random with a fixed cardinality $|S_k| = \theta m$.*
2. *Perform the following updates:*

$$\mathbf{y}_i^{k+1} = \begin{cases} \arg\min_{\mathbf{y}_i \in C_i} g_i(\mathbf{y}_i) + \langle \mathbf{w}_i^k, \mathbf{y}_i \rangle + \frac{\beta}{2} \|\mathbf{y}_i - \mathbf{x}_i^k\|^2, & \forall i \in S_k, \\ \mathbf{y}_i^k, & \forall i \notin S_k \end{cases} \tag{3.3a}$$

$$\mathbf{x}^{k+1} = \mathcal{P}_{\mathcal{N}}(\mathbf{y}^{k+1}), \tag{3.3b}$$

$$\mathbf{w}^{k+1} = \mathbf{w}^k + \rho(\mathbf{y}^{k+1} - \mathbf{x}^{k+1}). \tag{3.3c}$$

3. *Set $k = k + 1$ and repeat until convergence.*

Here, for simplicity of analysis we assumed that S_k has a fixed cardinality θm. Since $\mathbf{w}^0 \in \mathcal{N}^\perp$, it follows from (3.3b)-(3.3c) that $\mathbf{w}^k \in \mathcal{N}^\perp$ for all $k \geq 1$. From (3.3a), only the variable components $\{\mathbf{y}_i : i \in S_k\}$ are updated at the kth iteration and $\{\mathbf{y}_i : i \notin S_k\}$ are kept untouch. Since the cardinality of S_k can be controlled freely, the per iteration cost of this stochastic variant of PHA can be significantly smaller than than of the original PHA. Our main contribution is to show in the next section that this stochastic PHA shares in an ergodic sense the same sublinear rate of convergence as the original algorithm.

4. Convergence analysis

Below we analyze the convergence property of the stochastic PHA given in Algorithm 3. Throughout the convergence analysis, we make the following assumption.

Assumption 4. *Assume that the set of KKT points of (3.2) is nonempty, i.e., there exists* $(\mathbf{x}^*, \mathbf{y}^*, \mathbf{w}^*)$ *such that* $\mathbf{x}^* = \mathbf{y}^* \in \mathcal{C} \cap \mathcal{N}$, $\mathbf{w}^* \in \mathcal{N}^\perp$, *and* $\tilde{g}(\mathbf{y}) - \tilde{g}(\mathbf{y}^*) + \langle \mathbf{w}^*, \mathbf{y} - \mathbf{y}^* \rangle \geq 0$ *for all* $\mathbf{y} \in \mathcal{C}$.

Let $S = S_k$, $\mathbf{y}_S = (\mathbf{y}_i)_{i \in S}$, $p_S = (p_i)_{i \in S}$ and $\tilde{g}_S(\mathbf{y}_S) = \sum_{i \in S} \tilde{g}_i(\mathbf{y}_i)$. Let $\mathbb{E}_S$ be the expectation about S conditional on all previous history, while $\mathbb{E}$ denote the overall expectation. For convenient of analysis, we further introduce a sequence of dummy vectors

$$\tilde{\mathbf{w}}^{k+1} = \mathbf{w}^k + \beta(\mathbf{y}^{k+1} - \mathbf{x}^{k+1}), \ \forall\, k \geq 0. \tag{4.1}$$

We first establish some lemmas, which characterize the one-step convergence behavior of the stochastic PHA.

Lemma 5. *For all* $\boldsymbol{y} \in \mathcal{C}$ *and* $k \geq 0$, *it holds that*

$$\begin{aligned} &\mathbb{E}_S\big[\tilde{g}(\mathbf{y}^{k+1}) - \tilde{g}(\mathbf{y}) + \big\langle \mathbf{y}^{k+1} - \mathbf{y}, \mathbf{w}^k + \beta(\mathbf{y}^{k+1} - \mathbf{x}^k)\big\rangle\big] \\ &\leq (1-\theta)\big[\tilde{g}(\mathbf{y}^k) - \tilde{g}(\mathbf{y}) + \langle \mathbf{y}^k - \mathbf{y}, \mathbf{w}^k + \beta(\mathbf{y}^k - \mathbf{x}^k)\rangle\big]. \end{aligned} \tag{4.2}$$

Proof
Let $\mathbf{y} \in \mathcal{C}$ be arbitrarily fixed. Then, the optimality condition of (3.3a) reads

$$\big\langle \mathbf{y}_i^{k+1} - \mathbf{y}_i, \nabla g_i(\mathbf{y}_i^{k+1}) + \mathbf{w}_i^k + \beta(\mathbf{y}_i^{k+1} - \mathbf{x}_i^k)\big\rangle \leq 0, \ \forall i \in S.$$

Multiplying both sides by p_i and taking a sum over $i \in S$, we obtain

$$\big\langle \mathbf{y}_S^{k+1} - \mathbf{y}_S, \nabla \tilde{g}_S(\mathbf{y}_S^{k+1}) + p_S \otimes \big[\mathbf{w}_S^k + \beta(\mathbf{y}_S^{k+1} - \mathbf{x}_S^k)\big]\big\rangle \leq 0. \tag{4.3}$$

Here and hereafter, we let "$\otimes$" be the Kronecker product. Note that since we assume $\theta < 1$ and thus $|S| = \theta m < m$. Therefore, the inner product in (4.3) is the standard one. It follows from the convexity of $\tilde{g}_i$, $i \in S$, and the fact $\mathbf{y}_i^{k+1} = \mathbf{y}_i^k$, $i \notin S$, that

$$\big\langle \mathbf{y}_S^{k+1} - \mathbf{y}_S, \nabla \tilde{g}_S(\mathbf{y}_S^{k+1})\big\rangle \geq \tilde{g}_S(\mathbf{y}_S^{k+1}) - \tilde{g}_S(\mathbf{y}_S) = \tilde{g}(\mathbf{y}^{k+1}) - \tilde{g}(\mathbf{y}^k) + \tilde{g}_S(\mathbf{y}_S^k) - \tilde{g}_S(\mathbf{y}_S).$$

By taking expectation $\mathbb{E}_S$ on both sides of the above inequality, we obtain

$$\begin{aligned} \mathbb{E}_S\big\langle \mathbf{y}_S^{k+1} - \mathbf{y}_S, \nabla \tilde{g}_S(\mathbf{y}_S^{k+1})\big\rangle &\geq \mathbb{E}_S\big[\tilde{g}(\mathbf{y}^{k+1}) - \tilde{g}(\mathbf{y}^k)\big] + \theta\big[\tilde{g}(\mathbf{y}^k) - \tilde{g}(\mathbf{y})\big] \\ &= \mathbb{E}_S\big[\tilde{g}(\mathbf{y}^{k+1}) - \tilde{g}(\mathbf{y})\big] - (1-\theta)\big[\tilde{g}(\mathbf{y}^k) - \tilde{g}(\mathbf{y})\big]. \end{aligned} \tag{4.4}$$

We use the fact $\mathbf{y}_i^{k+1} = \mathbf{y}_i^k$, $\forall i \notin S$, and write

$$\begin{aligned} \big\langle \mathbf{y}_S^{k+1} - \mathbf{y}_S, p_S \otimes \mathbf{w}_S^k\big\rangle &= \big\langle \mathbf{y}_S^{k+1} - \mathbf{y}_S^k, p_S \otimes \mathbf{w}_S^k\big\rangle + \big\langle \mathbf{y}_S^k - \mathbf{y}_S, p_S \otimes \mathbf{w}_S^k\big\rangle \\ &= \big\langle \mathbf{y}^{k+1} - \mathbf{y}^k, \mathbf{w}^k\big\rangle + \big\langle \mathbf{y}_S^k - \mathbf{y}_S, p_S \otimes \mathbf{w}_S^k\big\rangle. \end{aligned} \tag{4.5}$$

Note that the inner product in the term $\big\langle \mathbf{y}^{k+1} - \mathbf{y}^k, \mathbf{w}^k\big\rangle$ in (4.5) is defined by (3.1) since the corresponding space is $\Re^{mn}$, while inner products in all other terms in this equation are standard. Take expectation $\mathbb{E}_S$ on both sides and obtain

$$\begin{aligned} \mathbb{E}_S\big\langle \mathbf{y}_S^{k+1} - \mathbf{y}_S, p_S \otimes \mathbf{w}_S^k\big\rangle &= \mathbb{E}_S\big\langle \mathbf{y}_S^{k+1} - \mathbf{y}_S^k, p_S \otimes \mathbf{w}_S^k\big\rangle + \mathbb{E}_S\big\langle \mathbf{y}_S^k - \mathbf{y}_S, p_S \otimes \mathbf{w}_S^k\big\rangle \\ &= \mathbb{E}_S\big\langle \mathbf{y}^{k+1} - \mathbf{y}^k, \mathbf{w}^k\big\rangle + \theta\langle \mathbf{y}^k - \mathbf{y}, \mathbf{w}^k\rangle \\ &= \mathbb{E}_S\big\langle \mathbf{y}^{k+1} - \mathbf{y}, \mathbf{w}^k\big\rangle - (1-\theta)\langle \mathbf{y}^k - \mathbf{y}, \mathbf{w}^k\rangle. \end{aligned} \tag{4.6}$$

By taking expectation $\mathbb{E}_S$ on (4.3), plugging (4.4) and (4.6) and rearranging terms, we obtain

$$\begin{aligned} &\mathbb{E}_S\big[\tilde{g}(\mathbf{y}^{k+1}) - \tilde{g}(\mathbf{y}) + \big\langle \mathbf{y}^{k+1} - \mathbf{y}, \mathbf{w}^k\big\rangle\big] + \mathbb{E}_S\big\langle \mathbf{y}_S^{k+1} - \mathbf{y}_S, p_S \otimes \beta(\mathbf{y}_S^{k+1} - \mathbf{x}_S^k)\big\rangle \\ &\leq (1-\theta)\big[\tilde{g}(\mathbf{y}^k) - \tilde{g}(\mathbf{y}) + \langle \mathbf{y}^k - \mathbf{y}, \mathbf{w}^k\rangle\big]. \end{aligned} \tag{4.7}$$

Again, by using $\mathbf{y}_i^{k+1} = \mathbf{y}_i^k$ for all $i \notin S$, we have

$$\begin{aligned}\left\langle \mathbf{y}_S^{k+1} - \mathbf{y}_S, p_S \otimes \beta(\mathbf{y}_S^{k+1} - \mathbf{x}_S^k)\right\rangle &= \left\langle \mathbf{y}_S^{k+1} - \mathbf{y}_S, p_S \otimes \beta(\mathbf{y}_S^{k+1} - \mathbf{y}_S^k)\right\rangle + \left\langle \mathbf{y}_S^{k+1} - \mathbf{y}_S, p_S \otimes \beta(\mathbf{y}_S^k - \mathbf{x}_S^k)\right\rangle \\ &= \left\langle \mathbf{y}^{k+1} - \mathbf{y}, \beta(\mathbf{y}^{k+1} - \mathbf{y}^k)\right\rangle + \left\langle \mathbf{y}_S^{k+1} - \mathbf{y}_S, p_S \otimes \beta(\mathbf{y}_S^k - \mathbf{x}_S^k)\right\rangle.\end{aligned}$$

Similar to (4.6), we have

$$\begin{aligned}\mathbb{E}_S\left\langle \mathbf{y}_S^{k+1} - \mathbf{y}_S, p_S \otimes \beta(\mathbf{y}_S^k - \mathbf{x}_S^k)\right\rangle &= \mathbb{E}_S\left\langle \mathbf{y}_S^{k+1} - \mathbf{y}_S^k, p_S \otimes \beta(\mathbf{y}_S^k - \mathbf{x}_S^k)\right\rangle + \mathbb{E}_S\left\langle \mathbf{y}_S^k - \mathbf{y}_S, p_S \otimes \beta(\mathbf{y}_S^k - \mathbf{x}_S^k)\right\rangle \\ &= \mathbb{E}_S\left\langle \mathbf{y}^{k+1} - \mathbf{y}^k, \beta(\mathbf{y}^k - \mathbf{x}^k)\right\rangle + \theta\left\langle \mathbf{y}^k - \mathbf{y}, \beta(\mathbf{y}^k - \mathbf{x}^k)\right\rangle \\ &= \mathbb{E}_S\left\langle \mathbf{y}^{k+1} - \mathbf{y}, \beta(\mathbf{y}^k - \mathbf{x}^k)\right\rangle - (1-\theta)\left\langle \mathbf{y}^k - \mathbf{y}, \beta(\mathbf{y}^k - \mathbf{x}^k)\right\rangle.\end{aligned}$$

Therefore,

$$\mathbb{E}_S\left\langle \mathbf{y}_S^{k+1} - \mathbf{y}_S, p_S \otimes \beta(\mathbf{y}_S^{k+1} - \mathbf{x}_S^k)\right\rangle = \mathbb{E}_S\left\langle \mathbf{y}^{k+1} - \mathbf{y}, \beta(\mathbf{y}^{k+1} - \mathbf{x}^k)\right\rangle - (1-\theta)\left\langle \mathbf{y}^k - \mathbf{y}, \beta(\mathbf{y}^k - \mathbf{x}^k)\right\rangle. \quad (4.8)$$

Plugging (4.8) into (4.7) and rearranging terms, we obtain (4.2). □

Lemma 6. *Let $T > 0$ be an integer. For any $\mathbf{w}$ and $(\mathbf{x}, \mathbf{y})$ such that $\mathbf{x} = \mathbf{y} \in \mathcal{N} \cap \mathcal{C}$, it holds that*

$$\begin{aligned}&\mathbb{E}\left[\tilde{g}(\mathbf{y}^{T+1}) - \tilde{g}(\mathbf{y}) + \left\langle \mathbf{y}^{T+1} - \mathbf{x}^{T+1}, \tilde{\mathbf{w}}^{T+1}\right\rangle\right] + \theta\sum_{k=0}^{T-1}\mathbb{E}\left[\tilde{g}(\mathbf{y}^{k+1}) - \tilde{g}(\mathbf{y}) + \left\langle \mathbf{y}^{k+1} - \mathbf{x}^{k+1}, \mathbf{w}\right\rangle\right] \\ &+ \frac{1}{2\beta}\left[\mathbb{E}\|\mathbf{w}^T - \mathbf{w}\|^2 - \|\mathbf{w}^0 - \mathbf{w}\|^2 + \sum_{k=0}^{T-1}\mathbb{E}\|\mathbf{w}^{k+1} - \mathbf{w}^k\|^2\right] + \mathbb{E}\langle \tilde{\mathbf{w}}^{T+1} - \mathbf{w}^T, \mathbf{x}^{T+1} - \mathbf{x}^T\rangle \\ &+ \frac{\beta}{2}\left[\mathbb{E}\|\mathbf{x}^{T+1} - \mathbf{x}\|^2 - \|\mathbf{x}^0 - \mathbf{x}\|^2 + \sum_{k=0}^{T}\mathbb{E}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2\right] + \frac{1}{\theta}\sum_{k=0}^{T-1}\mathbb{E}\langle \mathbf{w}^{k+1} - \mathbf{w}^k, \mathbf{x}^{k+1} - \mathbf{x}^k\rangle \\ &\leq (1-\theta)\left[\tilde{g}(\mathbf{y}^0) - \tilde{g}(\mathbf{y}) + \langle \mathbf{y}^0 - \mathbf{y}, \mathbf{w}^0\rangle\right] + (1-\theta)\left\langle \mathbf{y}^0 - \mathbf{y}, \beta(\mathbf{y}^0 - \mathbf{x}^0)\right\rangle.\end{aligned} \quad (4.9)$$

Proof
Taking expectation on both sides of (4.2) and summing it from $k = 0$ through T, we have

$$\begin{aligned}&\mathbb{E}\left[\tilde{g}(\mathbf{y}^{T+1}) - \tilde{g}(\mathbf{y})\right] + \theta\sum_{k=0}^{T-1}\mathbb{E}\left[\tilde{g}(\mathbf{y}^{k+1}) - \tilde{g}(\mathbf{y})\right] + \sum_{k=0}^{T}\mathbb{E}\left\langle \mathbf{y}^{k+1} - \mathbf{y}, \mathbf{w}^k + \beta(\mathbf{y}^{k+1} - \mathbf{x}^k)\right\rangle \\ &\leq (1-\theta)\left[\tilde{g}(\mathbf{y}^0) - \tilde{g}(\mathbf{y}) + \sum_{k=0}^{T}\mathbb{E}\langle \mathbf{y}^k - \mathbf{y}, \mathbf{w}^k + \beta(\mathbf{y}^k - \mathbf{x}^k)\rangle\right].\end{aligned} \quad (4.10)$$

Recall that $\tilde{\mathbf{w}}^{T+1}$ is defined in (4.1). Plug the identity

$$\begin{aligned}\sum_{k=0}^{T}\mathbb{E}\left\langle \mathbf{y}^{k+1} - \mathbf{y}, \mathbf{w}^k\right\rangle &= \mathbb{E}\left\langle \mathbf{y}^{T+1} - \mathbf{y}, \tilde{\mathbf{w}}^{T+1}\right\rangle - \mathbb{E}\left\langle \mathbf{y}^{T+1} - \mathbf{y}, \tilde{\mathbf{w}}^{T+1} - \mathbf{w}^T\right\rangle \\ &\quad + \sum_{k=0}^{T-1}\mathbb{E}\left\langle \mathbf{y}^{k+1} - \mathbf{y}, \mathbf{w}^{k+1}\right\rangle - \sum_{k=0}^{T-1}\mathbb{E}\left\langle \mathbf{y}^{k+1} - \mathbf{y}, \mathbf{w}^{k+1} - \mathbf{w}^k\right\rangle\end{aligned}$$

into (4.10) and rearrange terms, we obtain

$$\begin{aligned}
&\mathbb{E}\big[\tilde{g}(\mathbf{y}^{T+1}) - \tilde{g}(\mathbf{y}) + \langle \mathbf{y}^{T+1} - \mathbf{y}, \tilde{\mathbf{w}}^{T+1}\rangle\big] + \theta \sum_{k=0}^{T-1} \mathbb{E}\big[\tilde{g}(\mathbf{y}^{k+1}) - \tilde{g}(\mathbf{y}) + \langle \mathbf{y}^{k+1} - \mathbf{y}, \mathbf{w}^{k+1}\rangle\big] \\
&+ \sum_{k=0}^{T} \mathbb{E}\left\langle \mathbf{y}^{k+1} - \mathbf{y}, \beta(\mathbf{y}^{k+1} - \mathbf{x}^k)\right\rangle - \mathbb{E}\left\langle \mathbf{y}^{T+1} - \mathbf{y}, \tilde{\mathbf{w}}^{T+1} - \mathbf{w}^T\right\rangle - \sum_{k=0}^{T-1} \mathbb{E}\left\langle \mathbf{y}^{k+1} - \mathbf{y}, \mathbf{w}^{k+1} - \mathbf{w}^k\right\rangle \\
&\le (1-\theta)\big[\tilde{g}(\mathbf{y}^0) - \tilde{g}(\mathbf{y}) + \langle \mathbf{y}^0 - \mathbf{y}, \mathbf{w}^0\rangle\big] + (1-\theta)\sum_{k=0}^{T} \mathbb{E}\left\langle \mathbf{y}^k - \mathbf{y}, \beta(\mathbf{y}^k - \mathbf{x}^k)\right\rangle. \qquad (4.11)
\end{aligned}$$

Since $\mathbf{w}^k, \tilde{\mathbf{w}}^k \in \mathcal{N}^\perp$ and $\mathbf{x}^k \in \mathcal{N}$ for all $k \ge 1$ and $\mathbf{x} = \mathbf{y} \in \mathcal{N} \cap \mathcal{C}$, it follows from (3.3b) that

$$\begin{array}{rcccl}
\left\langle \mathbf{y}^{T+1} - \mathbf{y}, \tilde{\mathbf{w}}^{T+1}\right\rangle & = & \left\langle \mathbf{y}^{T+1} - \mathbf{x}^{T+1} + \mathbf{x}^{T+1} - \mathbf{x}, \tilde{\mathbf{w}}^{T+1}\right\rangle & = & \left\langle \mathbf{y}^{T+1} - \mathbf{x}^{T+1}, \tilde{\mathbf{w}}^{T+1}\right\rangle, \\
\left\langle \mathbf{y}^{k+1} - \mathbf{y}, \mathbf{w}^{k+1}\right\rangle & = & \left\langle \mathbf{y}^{k+1} - \mathbf{x}^{k+1} + \mathbf{x}^{k+1} - \mathbf{x}, \mathbf{w}^{k+1}\right\rangle & = & \left\langle \mathbf{y}^{k+1} - \mathbf{x}^{k+1}, \mathbf{w}^{k+1}\right\rangle,
\end{array}$$

for all $k = 0, 1, \ldots, T-1$. Plug the above identities into (4.11), we obtain

$$\begin{aligned}
&\mathbb{E}\big[\tilde{g}(\mathbf{y}^{T+1}) - \tilde{g}(\mathbf{y}) + \langle \mathbf{y}^{T+1} - \mathbf{x}^{T+1}, \tilde{\mathbf{w}}^{T+1}\rangle\big] + \theta \sum_{k=0}^{T-1} \mathbb{E}\left[\tilde{g}(\mathbf{y}^{k+1}) - \tilde{g}(\mathbf{y}) + \langle \mathbf{y}^{k+1} - \mathbf{x}^{k+1}, \mathbf{w}^{k+1}\rangle\right] \\
&+ \sum_{k=0}^{T} \mathbb{E}\left\langle \mathbf{y}^{k+1} - \mathbf{y}, \beta(\mathbf{y}^{k+1} - \mathbf{x}^k)\right\rangle - \mathbb{E}\left\langle \mathbf{y}^{T+1} - \mathbf{y}, \tilde{\mathbf{w}}^{T+1} - \mathbf{w}^T\right\rangle - \sum_{k=0}^{T-1} \mathbb{E}\left\langle \mathbf{y}^{k+1} - \mathbf{y}, \mathbf{w}^{k+1} - \mathbf{w}^k\right\rangle \\
&\le (1-\theta)\big[\tilde{g}(\mathbf{y}^0) - \tilde{g}(\mathbf{y}) + \langle \mathbf{y}^0 - \mathbf{y}, \mathbf{w}^0\rangle\big] + (1-\theta)\sum_{k=0}^{T} \mathbb{E}\left\langle \mathbf{y}^k - \mathbf{y}, \beta(\mathbf{y}^k - \mathbf{x}^k)\right\rangle. \qquad (4.12)
\end{aligned}$$

Since $\mathbf{y}^{k+1} - \mathbf{x}^{k+1} = \frac{1}{\rho}(\mathbf{w}^{k+1} - \mathbf{w}^k)$ and $\rho = \theta\beta$, it holds that

$$\begin{aligned}
\theta \sum_{k=0}^{T-1} \left\langle \mathbf{y}^{k+1} - \mathbf{x}^{k+1}, \mathbf{w}^{k+1} - \mathbf{w}\right\rangle &= \frac{1}{\beta} \sum_{k=0}^{T-1} \left\langle \mathbf{w}^{k+1} - \mathbf{w}^k, \mathbf{w}^{k+1} - \mathbf{w}\right\rangle \\
&= \frac{1}{2\beta}\big[\|\mathbf{w}^T - \mathbf{w}\|^2 - \|\mathbf{w}^0 - \mathbf{w}\|^2 + \sum_{k=0}^{T-1} \|\mathbf{w}^{k+1} - \mathbf{w}^k\|^2\big]. \qquad (4.13)
\end{aligned}$$

Note that

$$\begin{aligned}
&\sum_{k=0}^{T} \left\langle \mathbf{y}^{k+1} - \mathbf{y}, \beta(\mathbf{y}^{k+1} - \mathbf{x}^{k+1})\right\rangle - (1-\theta)\sum_{k=0}^{T} \left\langle \mathbf{y}^k - \mathbf{y}, \beta(\mathbf{y}^k - \mathbf{x}^k)\right\rangle \\
= \;& \left\langle \mathbf{y}^{T+1} - \mathbf{y}, \beta(\mathbf{y}^{T+1} - \mathbf{x}^{T+1})\right\rangle + \theta \sum_{k=0}^{T-1} \left\langle \mathbf{y}^{k+1} - \mathbf{y}, \beta(\mathbf{y}^{k+1} - \mathbf{x}^{k+1})\right\rangle - (1-\theta)\left\langle \mathbf{y}^0 - \mathbf{y}, \beta(\mathbf{y}^0 - \mathbf{x}^0)\right\rangle \\
= \;& \left\langle \mathbf{y}^{T+1} - \mathbf{y}, \tilde{\mathbf{w}}^{T+1} - \mathbf{w}^T\right\rangle + \sum_{k=0}^{T-1} \left\langle \mathbf{y}^{k+1} - \mathbf{y}, \mathbf{w}^{k+1} - \mathbf{w}^k\right\rangle - (1-\theta)\left\langle \mathbf{y}^0 - \mathbf{y}, \beta(\mathbf{y}^0 - \mathbf{x}^0)\right\rangle. \qquad (4.14)
\end{aligned}$$

Recalling $\mathbf{x} = \mathbf{y}$, we have

$$\left\langle \mathbf{y}^{k+1} - \mathbf{y}, \beta(\mathbf{x}^{k+1} - \mathbf{x}^k)\right\rangle = \frac{1}{\theta}\langle \mathbf{w}^{k+1} - \mathbf{w}^k, \mathbf{x}^{k+1} - \mathbf{x}^k\rangle + \beta\langle \mathbf{x}^{k+1} - \mathbf{x}, \mathbf{x}^{k+1} - \mathbf{x}^k\rangle, \qquad (4.15)$$

$$\left\langle \mathbf{y}^{k+1} - \mathbf{y}, \beta(\mathbf{x}^{k+1} - \mathbf{x}^k)\right\rangle = \langle \tilde{\mathbf{w}}^{k+1} - \mathbf{w}^k, \mathbf{x}^{k+1} - \mathbf{x}^k\rangle + \beta\langle \mathbf{x}^{k+1} - \mathbf{x}, \mathbf{x}^{k+1} - \mathbf{x}^k\rangle. \qquad (4.16)$$

Summing (4.15) from $k=0$ through $T-1$ and adding (4.16) with $k=T$ yield

$$\begin{aligned}
&\sum_{k=0}^{T}\left\langle \mathbf{y}^{k+1}-\mathbf{y},\beta(\mathbf{x}^{k+1}-\mathbf{x}^{k})\right\rangle \\
&=\frac{1}{\theta}\sum_{k=0}^{T-1}\langle \mathbf{w}^{k+1}-\mathbf{w}^{k},\mathbf{x}^{k+1}-\mathbf{x}^{k}\rangle+\langle \tilde{\mathbf{w}}^{T+1}-\mathbf{w}^{T},\mathbf{x}^{T+1}-\mathbf{x}^{T}\rangle+\sum_{k=0}^{T}\beta\langle \mathbf{x}^{k+1}-\mathbf{x},\mathbf{x}^{k+1}-\mathbf{x}^{k}\rangle \\
&=\frac{1}{\theta}\sum_{k=0}^{T-1}\langle \mathbf{w}^{k+1}-\mathbf{w}^{k},\mathbf{x}^{k+1}-\mathbf{x}^{k}\rangle+\langle \tilde{\mathbf{w}}^{T+1}-\mathbf{w}^{T},\mathbf{x}^{T+1}-\mathbf{x}^{T}\rangle \\
&\quad+\frac{\beta}{2}\left[\|\mathbf{x}^{T+1}-\mathbf{x}\|^{2}-\|\mathbf{x}^{0}-\mathbf{x}\|^{2}+\sum_{k=0}^{T}\|\mathbf{x}^{k+1}-\mathbf{x}^{k}\|^{2}\right].
\end{aligned}\tag{4.17}$$

It follows from (4.14) and (4.17) that

$$\begin{aligned}
&\sum_{k=0}^{T}\left\langle \mathbf{y}^{k+1}-\mathbf{y},\beta(\mathbf{y}^{k+1}-\mathbf{x}^{k})\right\rangle-(1-\theta)\sum_{k=0}^{T}\left\langle \mathbf{y}^{k}-\mathbf{y},\beta(\mathbf{y}^{k}-\mathbf{x}^{k})\right\rangle \\
&=\sum_{k=0}^{T}\left\langle \mathbf{y}^{k+1}-\mathbf{y},\beta[(\mathbf{y}^{k+1}-\mathbf{x}^{k+1})+(\mathbf{x}^{k+1}-\mathbf{x}^{k})]\right\rangle-(1-\theta)\sum_{k=0}^{T}\left\langle \mathbf{y}^{k}-\mathbf{y},\beta(\mathbf{y}^{k}-\mathbf{x}^{k})\right\rangle \\
&=\left\langle \mathbf{y}^{T+1}-\mathbf{y},\tilde{\mathbf{w}}^{T+1}-\mathbf{w}^{T}\right\rangle+\sum_{k=0}^{T-1}\left\langle \mathbf{y}^{k+1}-\mathbf{y},\mathbf{w}^{k+1}-\mathbf{w}^{k}\right\rangle-(1-\theta)\left\langle \mathbf{y}^{0}-\mathbf{y},\beta(\mathbf{y}^{0}-\mathbf{x}^{0})\right\rangle \\
&\quad+\frac{1}{\theta}\sum_{k=0}^{T-1}\langle \mathbf{w}^{k+1}-\mathbf{w}^{k},\mathbf{x}^{k+1}-\mathbf{x}^{k}\rangle+\langle \tilde{\mathbf{w}}^{T+1}-\mathbf{w}^{T},\mathbf{x}^{T+1}-\mathbf{x}^{T}\rangle \\
&\quad+\frac{\beta}{2}\left[\|\mathbf{x}^{T+1}-\mathbf{x}\|^{2}-\|\mathbf{x}^{0}-\mathbf{x}\|^{2}+\sum_{k=0}^{T}\|\mathbf{x}^{k+1}-\mathbf{x}^{k}\|^{2}\right].
\end{aligned}\tag{4.18}$$

Taking expectation on (4.13) and (4.18) and adding them into (4.12), we obtain (4.9). □

Now, we are ready to establish a key result for the ergodic convergence.

Theorem 7. *Assume* $\mathbf{x}^{0}=\mathbf{y}^{0}$ *and* $\mathbf{w}^{0}=\mathbf{0}$. *Let* $T>0$ *be any integer and define*

$$\bar{\mathbf{x}}^{T+1}=\frac{\mathbf{x}^{T+1}+\theta\sum_{k=0}^{T-1}\mathbf{x}^{k+1}}{1+\theta T}\quad\text{and}\quad\bar{\mathbf{y}}^{T+1}=\frac{\mathbf{y}^{T+1}+\theta\sum_{k=0}^{T-1}\mathbf{y}^{k+1}}{1+\theta T}.$$

For any $\mathbf{w}$ *and* $\mathbf{x}=\mathbf{y}\in\mathcal{N}\cap\mathcal{C}$, *we have*

$$\mathbb{E}\left[\tilde{g}(\bar{\mathbf{y}}^{T+1})-\tilde{g}(\mathbf{y})+\left\langle \bar{\mathbf{y}}^{T+1}-\bar{\mathbf{x}}^{T+1},\mathbf{w}\right\rangle\right]\leq\frac{(1-\theta)\left[\tilde{g}(\mathbf{y}^{0})-\tilde{g}(\mathbf{y})\right]+\frac{1}{2\beta}\|\mathbf{w}\|^{2}+\frac{\beta}{2}\|\mathbf{x}^{0}-\mathbf{x}\|^{2}}{1+\theta T}.\tag{4.19}$$

Proof
Note that

$$\begin{aligned}\frac{1}{\theta}\sum_{k=0}^{T-1}\langle \mathbf{w}^{k+1}-\mathbf{w}^k, \mathbf{x}^{k+1}-\mathbf{x}^k\rangle &= \beta\sum_{k=0}^{T-1}\langle \mathbf{y}^{k+1}-\mathbf{x}^{k+1}, \mathbf{x}^{k+1}-\mathbf{x}^k\rangle\\ &= -\frac{\beta}{2}\sum_{k=0}^{T-1}\left[\|\mathbf{x}^{k+1}-\mathbf{y}^{k+1}\|^2-\|\mathbf{x}^k-\mathbf{y}^{k+1}\|^2+\|\mathbf{x}^{k+1}-\mathbf{x}^k\|^2\right]\\ &\geq -\frac{\beta}{2}\sum_{k=0}^{T-1}\|\mathbf{x}^{k+1}-\mathbf{x}^k\|^2, \end{aligned} \tag{4.20}$$

where the inequality follows from $\mathbf{x}^{k+1}=\mathcal{P}_{\mathcal{N}}(\mathbf{y}^{k+1})$. Apparently, this holds trivially if $\mathbf{x}^0\in\mathcal{N}$ since in that case $\mathbf{x}^k\in\mathcal{N}$ and $\mathbf{w}^k\in\mathcal{N}^\perp$ for all $k\geq 0$. In addition, there hold

$$\begin{aligned}\left\langle \mathbf{y}^{T+1}-\mathbf{x}^{T+1}, \tilde{\mathbf{w}}^{T+1}-\mathbf{w}\right\rangle &= \frac{1}{\beta}\left\langle \tilde{\mathbf{w}}^{T+1}-\mathbf{w}^T, \tilde{\mathbf{w}}^{T+1}-\mathbf{w}\right\rangle\\ &= \frac{1}{2\beta}\left[\|\tilde{\mathbf{w}}^{T+1}-\mathbf{w}^T\|^2-\|\mathbf{w}^T-\mathbf{w}\|^2+\|\tilde{\mathbf{w}}^{T+1}-\mathbf{w}\|^2\right]\\ &= \frac{\beta}{2}\|\mathbf{y}^{T+1}-\mathbf{x}^{T+1}\|^2+\frac{1}{2\beta}\left[\|\tilde{\mathbf{w}}^{T+1}-\mathbf{w}\|^2-\|\mathbf{w}^T-\mathbf{w}\|^2\right] \end{aligned} \tag{4.21}$$

and

$$\begin{aligned}\langle \tilde{\mathbf{w}}^{T+1}-\mathbf{w}^T, \mathbf{x}^{T+1}-\mathbf{x}^T\rangle &= \beta\langle \mathbf{y}^{T+1}-\mathbf{x}^{T+1}, \mathbf{x}^{T+1}-\mathbf{x}^T\rangle\\ &= -\frac{\beta}{2}\left[\|\mathbf{y}^{T+1}-\mathbf{x}^{T+1}\|^2+\|\mathbf{x}^{T+1}-\mathbf{x}^T\|^2-\|\mathbf{x}^T-\mathbf{y}^{T+1}\|^2\right]. \end{aligned} \tag{4.22}$$

Adding (4.20), (4.21) and (4.22) together, we

$$\begin{aligned}&\frac{1}{\theta}\sum_{k=0}^{T-1}\langle \mathbf{w}^{k+1}-\mathbf{w}^k, \mathbf{x}^{k+1}-\mathbf{x}^k\rangle+\left\langle \mathbf{y}^{T+1}-\mathbf{x}^{T+1}, \tilde{\mathbf{w}}^{T+1}-\mathbf{w}\right\rangle+\langle \tilde{\mathbf{w}}^{T+1}-\mathbf{w}^T, \mathbf{x}^{T+1}-\mathbf{x}^T\rangle\\ \geq& -\frac{\beta}{2}\sum_{k=0}^{T}\|\mathbf{x}^{k+1}-\mathbf{x}^k\|^2+\frac{1}{2\beta}\left[\|\tilde{\mathbf{w}}^{T+1}-\mathbf{w}\|^2-\|\mathbf{w}^T-\mathbf{w}\|^2\right]. \end{aligned} \tag{4.23}$$

Taking expectation on both sides of (4.23) and adding it to (4.9), we obtain

$$\begin{aligned}&\mathbb{E}\left[\tilde{g}(\mathbf{y}^{T+1})-\tilde{g}(\mathbf{y})+\left\langle \mathbf{y}^{T+1}-\mathbf{x}^{T+1}, \mathbf{w}\right\rangle\right]+\theta\sum_{k=0}^{T-1}\mathbb{E}\left[\tilde{g}(\mathbf{y}^{k+1})-\tilde{g}(\mathbf{y})+\left\langle \mathbf{y}^{k+1}-\mathbf{x}^{k+1}, \mathbf{w}\right\rangle\right]\\ &+\frac{1}{2\beta}\left[\mathbb{E}\|\tilde{\mathbf{w}}^{T+1}-\mathbf{w}\|^2-\|\mathbf{w}^0-\mathbf{w}\|^2+\sum_{k=0}^{T-1}\mathbb{E}\|\mathbf{w}^{k+1}-\mathbf{w}^k\|^2\right]+\frac{\beta}{2}\left[\mathbb{E}\|\mathbf{x}^{T+1}-\mathbf{x}\|^2-\|\mathbf{x}^0-\mathbf{x}\|^2\right]\\ &\leq(1-\theta)\left[\tilde{g}(\mathbf{y}^0)-\tilde{g}(\mathbf{y})+\langle \mathbf{y}^0-\mathbf{y}, \mathbf{w}^0\rangle\right]+(1-\theta)\left\langle \mathbf{y}^0-\mathbf{y}, \beta(\mathbf{y}^0-\mathbf{x}^0)\right\rangle, \end{aligned}$$

which implies

$$\begin{aligned}&\mathbb{E}\left[\tilde{g}(\mathbf{y}^{T+1})-\tilde{g}(\mathbf{y})+\left\langle \mathbf{y}^{T+1}-\mathbf{x}^{T+1}, \mathbf{w}\right\rangle\right]+\theta\sum_{k=0}^{T-1}\mathbb{E}\left[\tilde{g}(\mathbf{y}^{k+1})-\tilde{g}(\mathbf{y})+\left\langle \mathbf{y}^{k+1}-\mathbf{x}^{k+1}, \mathbf{w}\right\rangle\right]\\ &\leq(1-\theta)\left[\tilde{g}(\mathbf{y}^0)-\tilde{g}(\mathbf{y})+\langle \mathbf{y}^0-\mathbf{y}, \mathbf{w}^0\rangle\right]+(1-\theta)\left\langle \mathbf{y}^0-\mathbf{y}, \beta(\mathbf{y}^0-\mathbf{x}^0)\right\rangle+\frac{1}{2\beta}\|\mathbf{w}^0-\mathbf{w}\|^2+\frac{\beta}{2}\|\mathbf{x}^0-\mathbf{x}\|^2. \end{aligned} \tag{4.24}$$

Since $\mathbf{x}^0 = \mathbf{y}^0$ and $\mathbf{w}^0 = \mathbf{0}$, we have from the convexity of g and (4.24) that

$$\begin{aligned}
&(1+\theta T)\mathbb{E}\big[\tilde{g}(\bar{\mathbf{y}}^{T+1}) - \tilde{g}(\mathbf{y}) + \big\langle \bar{\mathbf{y}}^{T+1} - \bar{\mathbf{x}}^{T+1}, \mathbf{w}\big\rangle\big] \\
&\le \mathbb{E}\big[\tilde{g}(\mathbf{y}^{T+1}) - \tilde{g}(\mathbf{y}) + \big\langle \mathbf{y}^{T+1} - \mathbf{x}^{T+1}, \mathbf{w}\big\rangle\big] + \theta \sum_{k=0}^{T-1} \mathbb{E}\left[\tilde{g}(\mathbf{y}^{k+1}) - \tilde{g}(\mathbf{y}) + \big\langle \mathbf{y}^{k+1} - \mathbf{x}^{k+1}, \mathbf{w}\big\rangle\right] \\
&\le (1-\theta)\big[\tilde{g}(\mathbf{y}^0) - \tilde{g}(\mathbf{y})\big] + \frac{1}{2\beta}\|\mathbf{w}\|^2 + \frac{\beta}{2}\|\mathbf{x}^0 - \mathbf{x}\|^2.
\end{aligned}$$

We thus have proved (4.19). □

To prove the main convergence result, we need the following lemmas, which follow directly from [3, Lemmas 3.3, 3.5].

Lemma 8. *Let h be a continuous function and $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ are random vectors. If for any $(\mathbf{x}, \mathbf{y}, \mathbf{w})$ with $\mathbf{x} = \mathbf{y}$, which may depend on $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$, it holds that*

$$\mathbb{E}\big[\tilde{g}(\bar{\mathbf{y}}) - \tilde{g}(\mathbf{y}) + (\bar{\mathbf{x}} - \bar{\mathbf{y}})^T\mathbf{w}\big] \le \mathbb{E}[h(\mathbf{x}, \mathbf{y}, \mathbf{w})],$$

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

$$\mathbb{E}\big[\tilde{g}(\bar{\mathbf{y}}) - \tilde{g}(\mathbf{y}^*) + \gamma\|\bar{\mathbf{x}} - \bar{\mathbf{y}}\|\big] \le \sup_{\|\mathbf{w}\| \le \gamma} \mathbb{E}[h(\mathbf{x}^*, \mathbf{y}^*, \mathbf{w})].$$

Lemma 9. *Let $(\mathbf{x}^*, \mathbf{y}^*, \mathbf{w}^*)$ be an optimal solution and $\gamma > \|\mathbf{w}^*\|$. If a random vector $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ satisfies*

$$\mathbb{E}\big[\tilde{g}(\bar{\mathbf{y}}) - \tilde{g}(\mathbf{y}^*) + \gamma\|\bar{\mathbf{x}} - \bar{\mathbf{y}}\|\big] \le \varepsilon,$$

then

$$\mathbb{E}\|\bar{\mathbf{x}} - \bar{\mathbf{y}}\| \le \frac{\varepsilon}{\gamma - \|\mathbf{w}^*\|} \quad \text{and} \quad \mathbb{E}|\tilde{g}(\bar{\mathbf{y}}) - \tilde{g}(\mathbf{y}^*)| \le \left[\frac{2\|\mathbf{w}^*\|}{\gamma - \|\mathbf{w}^*\|} + 1\right]\varepsilon.$$

The following theorem summarizes the main convergence result of this paper.

Theorem 10. *Let $T > 0$ be an integer and $(\mathbf{x}^*, \mathbf{y}^*, \mathbf{w}^*)$ be any primal-dual optimal solution. Under the conditions $\mathbf{x}^0 = \mathbf{y}^0$ and $\mathbf{w}^0 = 0$, we have $\bar{\mathbf{x}}^{T+1} \in \mathcal{N}$, $\bar{\mathbf{y}}^{T+1} \in \mathcal{C}$, and*

$$\max\left\{\mathbb{E}\|\bar{\mathbf{x}}^{T+1} - \bar{\mathbf{y}}^{T+1}\|, \mathbb{E}|\tilde{g}(\bar{\mathbf{y}}^{T+1}) - \tilde{g}(\mathbf{y}^*)|\right\} \le \frac{2C}{1+\theta T}, \tag{4.25}$$

where C is dependent on $(\mathbf{x}^*, \mathbf{y}^*, \mathbf{w}^*)$ and $(\mathbf{x}^0, \mathbf{y}^0)$ and is given by

$$C = (1-\theta)|\tilde{g}(\mathbf{y}^0) - \tilde{g}(\mathbf{y}^*)| + \frac{\max\{(\|\mathbf{w}^*\| + 0.5)^2, 9\|\mathbf{w}^*\|^2\}}{2\beta} + \frac{\beta}{2}\|\mathbf{x}^0 - \mathbf{x}^*\|^2.$$

Proof
Set $\gamma = \max(\|\mathbf{w}^*\| + 0.5, 3\|\mathbf{w}^*\|)$. Then

$$\frac{1}{\gamma - \|\mathbf{w}^*\|} \le 2 \quad \text{and} \quad \frac{2\|\mathbf{w}^*\|}{\gamma - \|\mathbf{w}^*\|} + 1 \le 2.$$

The theorem follows directly from (4.19), Lemmas 8 and 9. □

Theorem 10 basically says that the proposed stochastic PHA converges in expectation and in an ergodic sense, because $\bar{\mathbf{x}}^{T+1} \in \mathcal{N}$ and $\bar{\mathbf{y}}^{T+1} \in \mathcal{C}$ satisfying (4.25) can naturally be viewed as an approximate solution of (3.2). Furthermore, the convergence rate is sublinear, i.e., $O(1/(1+\theta T))$, which reduces to $O(1/(1+T))$ for the classical ADMM when $\theta = 1$, see [5].

5. Concluding remarks

Progressive hedging algorithm (PHA) is a classical decomposition approach for solving multistage convex stochastic programming (SP). It decomposes a large and complicated problem into many smaller pieces and can be implemented in parallel. Due to this feature, it is especially advantageous in large scale computing. Thus, PHA has been used in various domains ever since it was invented. Lately, it becomes even more popular due to the series of work by Rockafellar and his collaborators, see [16, 13, 12, 10, 11, 14].

At each iteration, PHA solves for each scenario a convex programming problem. Since the cardinality of the scenario set can be very large [18, 1, 17], the per iteration cost of PHA can be prohibitive in practice. Based on a connection between PHA and the alternating direction method of multipliers, in this paper we proposed a stochastic PHA, which selects a small fraction of the scenarios and updates the corresponding variables only. The number of selected scenarios can be controlled freely. Thus, the per iteration cost of the new algorithm can be significantly lower. Furthermore, our analyses have shown that the new algorithm shares the same sublinear ergodic convergence rate in the sense of expectation. An interesting problem might be extending the present analysis for SP to the case of multistage stochastic variational inequality [16, 13].

Acknowledgement

The work of Z. Mu was supported by Jiangsu Key Laboratory for Numerical Simulation of Large Scale Complex Systems. The work of J. Yang was supported by the Natural Science Foundation of China (NSFC-11771208, 11922111). We thank Prof. Xu Yangyang of Rensselaer Polytechnic Institute for the helpful discussions.

REFERENCES

1. J. Birge and F. Louveaux, *Introduction to stochastic programming*, Springer Science & Business Media, (2011).
2. D. Gabay and B. Mercier, *A dual algorithm for the solution of nonlinear variational problems via finite element approximation*, Computers and Mathematics with Applications, 2 (1976), pp. 17–40.
3. X. Gao, Y. Y. Xu, and S. Z. Zhang, *Randomized primal-dual proximal block coordinate updates*, Journal of the Operations Research Society of China, 7(2), pp. 205–250, 2019.
4. R. Glowinski and A. Marrocco, *Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité, d'une classe de problèmes de Dirichlet non linéaires*, R.A.I.R.O., R2, 9 (1975), pp. 41–76.
5. B. He and X. Yuan, *On the $O(1/n)$ convergence rate of the Douglas-Rachford alternating direction method*, SIAM J. Numer. Anal., 50 (2012), pp. 700–709.
6. M. R. Hestenes, *Multiplier and gradient methods*, J. Optimization Theory Appl., 4 (1969), pp. 303–320.
7. B. Martinet, *Régularisation d'inéquations variationnelles par approximations successives*, Rev. Française Informat. Recherche Opérationnelle, 4 (1970), pp. 154–158.
8. M. J. D. Powell, *A method for nonlinear constraints in minimization problems*, in Optimization (Sympos., Univ. Keele, Keele, 1968), Academic Press, London, 1969, pp. 283–298.
9. R. T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optimization, 14 (1976), pp. 877–898.
10. ———, *Progressive decoupling of linkages in monotone variational inequalities and convex optimization*, in Proceedings of the 10th International Conference on Nonlinear Analysis and Convex Analysis (Chitose, Japanm 2017, Yokohama Publishers, Japan).
11. ———, *Progressive decoupling of linkages in optimization and variational inequalities with elicitable convexity or monotonicity*, Set-valued and Variational Analysis, 27(2019), pp. 863–893.
12. ———, *Solving stochastic programming problems with risk measures by progressive hedging*, Set-Valued and Variational Analysis, 26 (2018), pp. 759–768.
13. R. T. Rockafellar and J. Sun, *Solving monotone stochastic variational inequalities and complementarity problems by progressive headging*, Math. Program., Series B, 174(2019), pp. 453–471.
14. R. T. Rockafellar and S. Uryasev, *Minimizing buffered probability of exceedance by progressive hedging*, Math. Program., Series B, 181(2020), pp. 453–472.
15. R. T. Rockafellar and R. J.-B. Wets, *Scenarios and policy aggregation in optimization under uncertainty*, Math. Oper. Res., 16 (1991), pp. 119–147.
16. ———, *Stochastic variaitional inequalities: single-stage to multistage*, Math. Program., 165 (2017), pp. 331–360.
17. A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on stochastic programming*, vol. 9 of MOS-SIAM Series on Optimization, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA; Mathematical Optimization Society, Philadelphia, PA, second ed., 2014. Modeling and theory.
18. A. Shapiro and A. Philpott, *A tutorial on stochastic programming*, Introductory Lecture Notes, (2007).