A Stochastic Primal-Dual Splitting Algorithm with Variance Reduction for Composite Optimization Problems

Van Dung Nguyen[1], Bằng Công Vũ[2], and Dimitri Papadimitriou[2]

[1] Department of Mathematics, University of Transport and Communications,
3 Cau Giay Street, Hanoi, Vietnam
[2] Belgium Research Center (BeRC) - Huawei, Leuven, Belgium
dungnv@utc.edu.vn; bangcvvn@gmail.com; dpapadimitriou@3nlab.org

October 16, 2024

Abstract

This paper revisits the generic structured primal-dual problem involving the infimal convolution in real Hilbert spaces. For this purpose, we develop a stochastic primal-dual splitting with variance reduction for solving this generic problem. Weak almost sure convergence of the iterates is proved. The linear convergence rate of the primal-dual gap is obtained under an additional condition like the strong convexity.

Keywords: Stochastic optimization, Variance reduction, Duality, Saddle point problem, Sublinear convergence, Linear convergence.

Mathematics Subject Classifications (2010): 49M29, 65K10, 65Y20, 90C25.

1 Introduction

In this paper, we revisit the following saddle point problem in real Hilbert spaces.

Problem 1.1 Let $(\mathcal{H}, \langle\cdot \mid \cdot\rangle)$, $(\mathcal{G}, \langle\cdot \mid \cdot\rangle)$ be separable real Hilbert spaces. Let $f\colon \mathcal{H} \to]-\infty, +\infty]$ and $g\colon \mathcal{G} \to]-\infty, +\infty]$ be proper lower semicontinuous convex functions. Let $h\colon \mathcal{H} \to \mathbb{R}$ and $\ell\colon \mathcal{G} \to \mathbb{R}$ be convex differentiable functions. Let $K\colon \mathcal{H} \to \mathcal{G}$ be a non zero bounded linear operator. The problem is to

$$\min_{x\in\mathcal{H}} \max_{v\in\mathcal{G}} G(x, v),$$

where the Lagrangian function G is defined by

$$\begin{aligned} G : \mathcal{H} \times \mathcal{G} &\to \mathbb{R} \cup \{-\infty, +\infty\} \\ (x, v) &\mapsto h(x) + f(x) + \langle Kx \mid v\rangle - g^\star(v) - \ell(v), \end{aligned} \tag{1.1}$$

where $g^\star$ denotes the Fenchel conjugate of the function g, see Section 2 for further details.

Various special cases of this problem with $\ell \equiv 0$ can be found in [6, 8, 9, 11, 12, 15, 17, 19, 21]. These special cases often arise in machine learning, image processing, statistics, game theory, portfolio optimization [1, 12, 14, 16, 18, 19, 20, 27, 28, 33, 35, 36]. We first state additional assumptions related to Problem 1.1 which will be used throughout this paper before motivating its investigation by a couple of examples.

Assumption 1.2 The following assumptions will be used in this paper.

(i) There exists a point $(x^\star, v^\star) \in \mathcal{H} \times \mathcal{G}$ that verifies the following saddle point conditions for the Lagrangian function G defined by (1.1):

$$(\forall x \in \mathcal{H})(\forall v \in \mathcal{G})\ G(x^\star, v) \leq G(x^\star, v^\star) \leq G(x, v^\star), \tag{1.2}$$

Further, we denote by $\mathcal{S}$ the set of all points $(x^\star, v^\star)$ such that conditions (1.2) are fully satisfied.

(ii) The functions h and ℓ are defined by finite sums, i.e., $h = \frac{1}{n}\sum_{i=1}^{n} h_i$ and $\ell = \frac{1}{n'}\sum_{j=1}^{n'} \ell_j$, where n and n' are positive integers and $h_i \colon \mathcal{H} \to \mathbb{R}$ and $\ell_j \colon \mathcal{G} \to \mathbb{R}$ are differentiable convex functions with μ_i and ν_j-Lipschitz gradient, respectively, $\forall i \in \{1, \ldots, n\}$ and $\forall j \in \{1, \ldots, n'\}$.

When $n = n' = 1$, as mentioned above, various examples can be found in the literature. Below, we present an example for the case when $n' > 1$ and $n > 1$.

Example 1.3 [Regularized Wasserstein barycenter problem [32, Remark 5.8]] Let m, p, p' and $(n_k)_{1\leq k\leq p}$ be strictly positive integer. Let Δ^m be the standard simplex in $\mathcal{H} = \mathbb{R}^m$, $\Delta^m = \big\{x \in \mathbb{R}^m \ \big| \ x \geq 0, \sum_{i=1}^m x_i = 1\big\}$. For every $k \in \{1, \ldots, p\}$, let $F^k \colon \mathbb{R}^m \to \mathbb{R}^{n_k}$ be a linear mapping. Let $(\theta^k)_{1\leq k\leq p}$ be the observation of an unknown vector $x \in \Delta^m$ through F^k, $\theta^k \approx F^k x$. Let $\alpha = (\alpha_k)_{1\leq k\leq p} \in \Delta^p$. The regularized Wasserstein barycenter problem can be formulated as the following saddle point problem in the real Hilbert spaces $\mathcal{H} = \mathbb{R}^m$ and $\mathcal{G} = \mathbb{R}^{n_1} \times \ldots \times \mathbb{R}^{n_p} \times \mathbb{R}^{m_1} \times \ldots \times \mathbb{R}^{m_{p'}}$:

$$\min_{x\in\Delta^m} \max_{\substack{\eta^1\in\mathbb{R}^{n_1},\ldots,\eta^p\in\mathbb{R}^{n_p}\\ \zeta^1\in\mathbb{R}^{m_1},\ldots,\zeta^{p'}\in\mathbb{R}^{m_{p'}}}} \sum_{k=1}^{p}\Bigg[\Big\langle \alpha_k\eta^k \mid F^k x\Big\rangle - \alpha_k\gamma_k \sum_{j=1}^{n_k}\theta_j^k \log\Bigg(\sum_{i=1}^{n_k}\exp\Big(\frac{\eta_i^k - C_{i,j}^k}{\gamma_k}\Big)\Bigg)\Bigg]$$
$$+ \sum_{r=1}^{p'}\Big[\langle \zeta^r \mid B^r x\rangle - (J^r)^\star(\zeta^r)\Big],$$

where $(\gamma_k)_{1\leq k\leq p}$ are strictly positive parameters, $C^k = (C_{i,j}^k)_{1\leq i,j\leq n_k}$ is a given matrix, $B^r \colon \mathbb{R}^m \to \mathbb{R}^{m_r}$ is a linear mapping, and $J^r \colon \mathbb{R}^{m_r} \to]-\infty, +\infty]$ is a proper lower semicontinous convex

function. This problem formulation is an instance of Problem 1.1 with

$$
\begin{aligned}
&n' = p,\\
&h = 0,\ \ f = \iota_{\Delta^m},\\
&v = (\eta^1, \ldots, \eta^p, \zeta^1, \ldots, \zeta^{p'}),\\
&\langle K\cdot \mid \cdot\rangle : (x, v) \mapsto \sum_{k=1}^{p} \left\langle \alpha_k \eta^k \mid F^k x \right\rangle + \sum_{r=1}^{p'} \langle \zeta^r \mid B^r x\rangle,\\
&g^\star \colon v \mapsto \sum_{r=1}^{p'} (J^r)^\star(\zeta^r),\\
&\ell \colon v \mapsto \sum_{k=1}^{p} \alpha_k \gamma_k \sum_{j=1}^{n_k} \theta_j^k \log\left(\sum_{i=1}^{n_k} \exp\left(\frac{\eta_i^k - C_{i,j}^k}{\gamma_k}\right)\right).
\end{aligned}
$$

Indeed, it is proved in [32] that ℓ is a differentiable function with Lipschitz continuous gradient.

Example 1.4 [Entropy regularized LPBoost [33]] Let $U = (u_{i,j})_{1\le i\le m, 1\le j\le p}$ be a given matrix; α, β, γ are positive parameters. The problem is to find

$$
\min_{x\in\Delta^m, x_i\le\alpha} \max_{y\in\Delta^p} x^T U y + \beta\|Vx\|^2 - \gamma\|Wy\|^2,
$$

where V and W are linear mappings on $\mathbb{R}^m$ and $\mathbb{R}^p$, respectively. This problem can be formulated as an instance of Problem 1.1 together with

$$
\begin{aligned}
&\mathcal{H} = \mathbb{R}^m,\ \ \mathcal{G} = \mathbb{R}^p\\
&h = \beta\|Vx\|^2,\ \ f = \iota_{x\in\Delta^m, x_i\le\alpha}\\
&\langle K\cdot \mid \cdot\rangle : (x, y) \mapsto x^T U y\\
&g^\star = 0,\ \ \ell \colon y \mapsto \gamma\|Wy\|^2.
\end{aligned}
$$

Example 1.5 Example for non-trivial, non-smooth $g^\star$ was given in [2, Section 5].

The saddle point problem has been investigated using Primal-dual Splitting (PDS) algorithms in [3, 4, 5, 7, 11, 14, 34] and recently in [25, 26, 32]. PDS methods for solving monotone inclusion problems are motivated by the fact that i) a wide variety of convex optimization problems such as location problems, support vector machine problems for classication and regression, problems in clustering and portfolio optimization as well as signal and image processing problems, all of them potentially possessing nonsmooth terms in their objectives, can be reduced to the solving of monotone inclusion problems blending linearly composed maximally monotone operators, parallel sums of maximally monotone operators and/or single-valued Lipschitzian or cocoercive monotone operators, and ii) classical splitting algorithms such as forward-backward algorithm [Bauschke, H.H., Combettes, P.L], Tseng's forward-backward-forward algorithm and Douglas-Rachford algorithm and variants yield considerable limitations when employed on monotone inclusion problems as they require computation of the resolvent(s) of linearly composed maximally monotone operators or of parallel sums of maximally monotone operators, for which exact formulae are available only in very exceptional situations; thus, simply inapplicable in practice. PDS methods overcome this

shortcoming by solving the primal-dual pair formed by the monotone inclusion and its dual (in the sense of Attouch-Thera or the classical Fenchel–Rockafellar duality framework) reformulated as a monotone inclusion problem in a corresponding product space. The PDS algorithmic scheme follows by applying standard splitting algorithms in an appropriate way. Subsequently, primal-dual splitting methods have been extensively investigated and have found many applications in applied mathematics; see [3, 6, 7, 8, 11, 12, 15, 17, 34] for instances. The first PDS algorithmic framework for solving structured composite problems involving infimal convolutions was proposed in [14]. This prototypical problem was then further investigated in [11]. Further developments and convergence analysis of the algorithmic framework developed in [14] can be found in [4].

Deterministic primal-dual splitting methods often evaluate the full gradient of h and ℓ. When n and n' remain relative small, Problem 1.1 can be solved efficiently by various deterministic primal-dual algorithms; see [7, 14, 15, 34] for examples. However, when n and n' are (very) large, the evaluation of the full gradient of h and ℓ becomes prohibitive since the computational cost increases with n and n'.

In turn, stochastic primal-dual splitting methods are often used alternatively. Recently, stochastic methods have found large interest in solving various problems see [1, 20, 21, 26, 30, 31, 33] for instances. Stochastic primal-dual splitting methods with variance reduction have been served as a standard approach to improve their convergence profiles. The reason being that computing the iterates of stochastic gradient does not ensure convergence to the solution without either ensuring the sequence of stepsizes is decreasing or involving variance reduction techniques. Variance reduction methods use $\nabla h_i(x_k)$ to update an estimate t_k of the gradient so that $t_k \approx \nabla h(x_k)$ as opposed to classical methods which use one or more $\nabla h_i(x_k)$ directly as an approximation of $\nabla h(x_k)$. With this gradient estimate, one then takes approximate gradient steps of the form $x_{k+1} = x_k - \gamma t_k$, where $\gamma > 0$ is the stepsize. To ensure its convergence with a constant stepsize, one verifies that the variance of the gradient estimate t_k converges to zero, that is $\mathsf{E}[\|t_k - \nabla h(x_k)\|^2] \xrightarrow{k\to\infty} 0$, where the expectation is taken with respect to all the random variables in the algorithm up to iteration k. This property is responsible for the faster convergence of VR methods and ensures that the VR method will stop when reaching the optimal point.

Challenges: As mentioned there are several primal-dual splitting (PDS) methods that can be used for solving Problem 1.1 in both deterministic and stochastic setting. The main challenge addressed in this paper resides in the use of variance reduction with both primal function h and dual function ℓ to improve the convergence profile. This issue was partly addressed in [25] albeit without strong guarantees for the convergence of the iteration sequence. We highlight below the existing framework proposed to solve the saddle-point Problem 1.1 in the stochastic setting for nontrivial functions ℓ.

(i) The work in [25] structured a stochastic Bregman PDS method for solving Problem 1.1. This work exploited the variance reduction technique and obtain the convergence of the primal-dual function only.

(ii) The work in [32] developed a stochastic Bregman PDS method for solving Problem 1.1. This work can be viewed as a stochastic extension of the work in [15, 34] with Bregman distance. However, this work does not exploit the variance reduction technique; hence, it imposes a quite strong condition on the variance.

(iii) The work in [26, Section 4] is an alternative method for solving Problem 1.1. However, here

too, this work does not exploit the variance reduction technique and imposes consequently a quite strong condition on the variance.

Main objective Recently, several stochastic variance reduction algorithms [1, 10, 24, 25, 33] have appeared that can be used to solve Problem 1.1. The objective of this paper is to develop a novel stochastic primal-dual splitting method with variance reduction for solving Problem 1.1 when n and n' are very large and investigate its convergence properties.

Contribution In this paper, we develop new stochastic primal-dual splitting methods for solving Problem 1.1, which incorporate the following features: (i) Using the acceleration technique in terms of variance reduction to obtain a faster convergence rate; (ii) The proposed algorithm is full splitting. Further, we prove the almost sure convergence of the iteration sequence for the general case. In the strongly convex case, we obtain the linear convergence in expectation of the primal-dual sequences.

Structure The remainder of this paper is organized as follows. In Section 2, we recall the base notions in convex analysis that will be used in the proof of the convergence of the proposed stochastic primal-dual splitting algorithm with variance reduction. Subsequently, in Section 3, we detail the proposed algorithm and the obtained convergence results.

2 Preliminaries

In this paper, we use the notations $\langle . \mid . \rangle$ and $\|.\|$ for inner product and norm in the spaces $\mathcal{H}, \mathcal{G}$. The conjugate of the operator K is denoted by $K^\star$. The domain of a function $f\colon \mathcal{H} \to]-\infty, +\infty]$ is $\mathrm{dom}(f) = \{x \in \mathcal{H} \mid f(x) < +\infty\}$. This function is proper if $\mathrm{dom}(f) \neq \varnothing$. We denote $\Gamma_0(\mathcal{H})$ the class of all proper lower semicontinuous convex functions f from $\mathcal{H}$ to $]-\infty, +\infty]$. We define by $\mathcal{H} \times \mathcal{G}$ the standard product space equipped with the norm $(x, v) \mapsto \sqrt{\|x\|^2 + \|v\|^2}$.

Let A be a set-valued operator on $\mathcal{H}$, i.e. $A\colon \mathcal{H} \to 2^{\mathcal{H}}$, the inverse of the operator A is defined as $A^{-1}\colon u \mapsto \{x \in \mathcal{H} \mid u \in Ax\}$. Denote $\overline{1, n} = \{1, 2, 3, \ldots, n\}$.

Definition 2.1 *A mapping $T : \mathcal{H} \to \mathcal{H}$ is $\alpha-$cocoercive ($\alpha \in]0, +\infty[$) if*

$$(\forall x \in \mathcal{H})(\forall y \in \mathcal{H})\ \langle Tx - Ty \mid x - y \rangle \geq \alpha \|Tx - Ty\|^2.$$

If $\alpha = 1$, T is firmly nonexpansive or equivalently

$$(\forall x \in \mathcal{H})(\forall y \in \mathcal{H})\ \|Tx - Ty\|^2 \leq \|x - y\|^2 - \|(\mathrm{Id} - T)x - (\mathrm{Id} - T)y\|^2.$$

Definition 2.2 *For $f \in \Gamma_0(\mathcal{H})$:*

(i) *The conjugate (or Fenchel conjugate) of the function f is*

$$f^\star\colon a \mapsto \sup_{x \in \mathcal{H}} \big(\langle a \mid x \rangle - f(x) \big).$$

(ii) *The subdifferential of f is*

$$\partial f\colon \mathcal{H} \to 2^{\mathcal{H}}\colon x \mapsto \{u \in \mathcal{H} \mid (\forall y \in \mathcal{H})\ \langle y - x \mid u\rangle + f(x) \leq f(y)\}$$

with inverse given by $(\partial f)^{-1} = \partial f^\star$

(iii) *The proximity operator of f is*

$$prox_f : \mathcal{H} \to \mathcal{H} : x \mapsto \underset{y\in\mathcal{H}}{argmin}(f(y) + \frac{1}{2}\|x - y\|^2).$$

Note: $prox_f = \mathrm{Id} - \mathrm{prox}_{f^\star} = (\mathrm{Id} + \partial f)^{-1}$ and $prox_f$ is firmly nonexpansive.

(iv) *The infimal convolution of two functions ℓ and g from $\mathcal{H}$ to $]-\infty, +\infty]$ is*

$$\ell \,\square\, g\colon x \mapsto \inf_{y\in\mathcal{H}}(\ell(y) + g(x - y)).$$

Note: If g and ℓ are in $\Gamma_0(\mathcal{G})$, the conjugate of the function $\ell \,\square\, g$ is $(\ell \,\square\, g)^\star = g^\star + \ell^\star$.

Definition 2.3 *The function $f : \mathcal{H} \to]-\infty, +\infty]$ is said to be strongly convex if there exists $\alpha \in]0, +\infty[$ such that $f - \alpha \|\cdot\|^2/2$ is convex.*

Definition 2.4 *The Lagrangian function G is $\alpha-$strongly convex-concave ($\alpha \in]0, +\infty[$) if*

$$(\forall x \in \mathcal{H})(\forall v \in \mathcal{G})\ G(x, v^\star) - G(x^\star, v) \geq \frac{\alpha}{2}\Big(\|x - x^\star\|^2 + \|v - v^\star\|^2\Big).$$

Let $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ be a probability space where $\Omega_1 = \{1, \ldots, n\}$, $\mathcal{F}_1 = 2^{\Omega_1}$, and $\mathbb{P}_1 = \{q_1, q_2, \ldots, q_n\}$ with $q_i \in]0, 1[$, $\sum_{i=1}^n q_i = 1$. Let $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$ be a probability space where $\Omega_2 = \{1, \ldots, n'\}$, $\mathcal{F}_2 = 2^{\Omega_2}$, and $\mathbb{P}_2 = \{q'_1, q'_2, \ldots, q'_{n'}\}$ with $q'_j \in]0, 1[$, $\sum_{j=1}^{n'} q'_i = 1$. Then $(\Omega, \mathcal{F}, \mathbb{P}) = (\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2, \mathbb{P}_1 \times \mathbb{P}_2)$ defines a probability space. A $\mathcal{H}$-valued random variable is a measurable function $X : \Omega \to \mathcal{H}$, where $\mathcal{H}$ is endowed with the Borel σ-algebra. The $\sigma-$algebra generated by a family Φ of random variables is denoted by $\sigma(\Phi)$. The expectation of a random variable X is denoted by $\mathsf{E}[X]$. The conditional expectation of X given a σ-field $\mathcal{A} \subset \mathcal{F}$ is denoted by $\mathsf{E}[X|\mathcal{A}]$. See [23] for more details on probability Theory in Hilbert spaces. The abbreviation a.s. stands for "almost surely".

Lemma 2.5 ([29, Theorem 1]) *Let $(\mathcal{F}_n)_{n\in\mathbb{N}}$ be an increasing sequence of sub-σ-algebras of $\mathcal{F}$, let $(z_n)_{n\in\mathbb{N}}$, $(\lambda_n)_{n\in\mathbb{N}}$, $(\zeta_n)_{n\in\mathbb{N}}$ and $(t_n)_{n\in\mathbb{N}}$ be $[0, +\infty[$-valued random sequences such that, for every $n \in \mathbb{N}$, z_n, ξ_n, ζ_n and t_n are $\mathcal{F}_n$-measurable. Assume moreover that $\sum_{n\in\mathbb{N}} t_n < +\infty$, $\sum_{n\in\mathbb{N}} \zeta_n < +\infty$ a.s. and*

$$(\forall n \in \mathbb{N})\ \mathsf{E}[z_{n+1}|\mathcal{F}_n] \leq (1 + t_n)z_n + \zeta_n - \lambda_n \text{ a.s.}.$$

Then $(z_n)_{n\in\mathbb{N}}$ converges a.s. to a random variable z_∞ and $(\lambda_n)_{n\in\mathbb{N}}$ is summable a.s..

The following lemma can be viewed as direct consequence of [13, Proposition 2.3].

Lemma 2.6 *Let C be a non-empty closed subset of $\mathcal{H}$ and let $(x_n)_{n\in\mathbb{N}}$ be a $\mathcal{H}$-valued random sequence. Suppose that, for every $x \in C$, $(\|x_{n+1} - x\|)_{n\in\mathbb{N}}$ converges a.s.. Suppose that the set of weak sequentially cluster points of $(x_n)_{n\in\mathbb{N}}$ is a subset of C a.s.. Then $(x_n)_{n\in\mathbb{N}}$ converges weakly a.s. to a C-valued random vector.*

3 Algorithm and convergence properties

Problem 1.1 can be written following the formulation introduced in [14].

Problem 3.1 Let $\mathcal{H}$, $\mathcal{G}$ be separable real Hilbert spaces. Let $f\colon \mathcal{H} \to]-\infty,+\infty]$ and $g\colon \mathcal{G} \to]-\infty,+\infty]$ be proper lower semicontinuous convex functions. Let $h\colon \mathcal{H} \to \mathbb{R}$ and $\ell\colon \mathcal{G} \to \mathbb{R}$ be convex differentiable functions. Let $K\colon \mathcal{H} \to \mathcal{G}$ be a bounded linear operator. The primal problem is to

$$\underset{x\in\mathcal{H}}{\text{minimize}}\ h(x) + (\ell^\star \square g)(Kx) + f(x),$$

and the dual problem (in the sense of Fenchel-Rockafellar) is to

$$\underset{v\in\mathcal{G}}{\text{minimize}}\ (h+f)^\star(-K^\star v) + g^\star(v) + \ell(v).$$

In Problem 3.1, the functions h and ℓ are given by finite sums, i.e., $h = \frac{1}{n}\sum_{i=1}^{n} h_i$ and $\ell = \frac{1}{n'}\sum_{j=1}^{n'} \ell_j$, where n and n' are positive integers and $h_i\colon \mathcal{H} \to \mathbb{R}$ and $\ell_j\colon \mathcal{G} \to \mathbb{R}$ are differentiable convex functions with μ_i and ν_j-Lipschitz gradient, respectively, $\forall i \in [1,n]$ and $j \in [1,n']$.

3.1 Variance reduction step

Before detailing our algorithm, we present the variance reduction step and derive some properties which need to be proven in order to demonstrate the convergence properties of the proposed algorithm.

Algorithm 3.2 Let m be a strictly positive integer, let $(\theta_k)_{0\leq k\leq m-1}$ and $(\gamma_k)_{0\leq k\leq m-1}$ be strictly positive real numbers. Let $(\overline{x}, x_0, x_{-1}) \in \mathcal{H}^3$ and $(\overline{v}, v_0, v_{-1}) \in \mathcal{G}^3$. Let $Q = \{q_1,\dots,q_n\}$ and $Q' = \{q'_1,\dots,q'_{n'}\}$ be the probabilities on $\{1,\dots,n\}$ and $\{1,\dots,n'\}$, respectively and iterate

For $k = 0, 1, \dots, m-1$
Select $i_k \in \{1,\dots,n\}$ randomly according to Q
Select $j_k \in \{1,\dots,n'\}$ randomly according to Q'
Compute

$$\begin{cases} y_k &= x_k + \theta_k(x_k - x_{k-1}) \\ z_k &= \dfrac{\nabla h_{i_k}(y_k) - \nabla h_{i_k}(\overline{x})}{q_{i_k} n} + \nabla h(\overline{x}) \\ u_k &= v_k + \theta_k(v_k - v_{k-1}) \\ t_k &= \dfrac{\nabla \ell_{j_k}(u_k) - \nabla \ell_{j_k}(\overline{v})}{q'_{j_k} n'} + \nabla \ell(\overline{v}) \end{cases}$$

Update

$$\begin{cases} x_{k+1} &= (\mathrm{Id} + \gamma_k \partial f)^{-1}(x_k - \gamma_k z_k - \gamma_k K^\star u_k) \\ v_{k+1} &= (\mathrm{Id} + \gamma_k \partial g^\star)^{-1}(v_k - \gamma_k t_k + \gamma_k K y_k). \end{cases}$$

end

Hereunder, we prove some results for the convergence of the variance of the $\mathcal{H}$-valued random sequences $(z_k)_{k\in\mathbb{N}}$ and $(t_k)_{k\in\mathbb{N}}$, and derive the main estimation of the difference $G(x_{k+1}, v^\star) - G(x^\star, v_{k+1})$ for $(x^\star, v^\star) \in \mathcal{S}$. For this purpose, we need to consider the following results.

Lemma 3.3 [25, Lemma 3.3] *Suppose Assumption 1.2 is satisfied. Set* $L_Q = \max_{i\in\overline{1,n}} \mu_i/(q_i n)$, $L_{Q'} = \max_{j\in\overline{1,n'}} \nu_j/(q'_j n')$. *Then, for* $(x, v) \in \mathrm{dom}(f) \times \mathrm{dom}(g^\star)$ *and* $(x^\star, v^\star) \in \mathcal{S}$, *we have*

$$\frac{1}{n}\sum_{i=1}^{n} \frac{1}{nq_i}\|\nabla h_i(x) - \nabla h_i(x^\star)\|^2 \leq 2L_Q[G(x, v^\star) - G(x^\star, v^\star)],$$

and

$$\frac{1}{n'}\sum_{j=1}^{n'} \frac{1}{n'q'_j}\|\nabla \ell_j(v) - \nabla \ell_j(v^\star)\|^2 \leq 2L_{Q'}[G(x^\star, v^\star) - G(x^\star, v)].$$

Corollary 3.4 *Under the same assumptions as in Lemma 3.3. Let* $(x_k)_{k\in\mathbb{N}}$, $(y_k)_{k\in\mathbb{N}}$, $(u_k)_{k\in\mathbb{N}}$, $(v_k)_{k\in\mathbb{N}}$, $(z_k)_{k\in\mathbb{N}}$ $(t_k)_{k\in\mathbb{N}}$ *be sequences generated by Algorithm 3.2. Let* E_{i_k} *and* E_{j_k} *be the conditional expectations of* i_k *and* j_k *with respect to the history* $\{(i_0, j_0), \ldots, (i_{k-1}, j_{k-1})\}$. *Then,* $(\forall k \in \mathbb{N})$ $\mathsf{E}_{i_k}[z_k]$ *and* $\mathsf{E}_{j_k}[t_k]$ *are unbiased estimators of* $\nabla h(y_k)$ *and* $\nabla \ell(u_k)$, *respectively, i.e., we have*

$$(\forall k \in \mathbb{N})\ \mathsf{E}_{i_k}[z_k] = \nabla h(y_k) \quad \text{and} \quad \mathsf{E}_{j_k}[t_k] = \nabla \ell(u_k). \tag{3.1}$$

Moreover, set $L_1 = \max\{L_Q, L_{Q'}\}$, $L_2 = \max_{i\in\overline{1,n},\ j\in\overline{1,n'}}\{\mu_i^2/(q_i n), \nu_j^2/(q'_j n')\}$. *Then, the following inequalities hold*

$$\begin{cases} \mathsf{E}_{i_k}\|z_k - \nabla h(y_k)\|^2 & \leq 2L_2(\theta_k^2 + \theta_k)\|x_k - x_{k-1}\|^2 + 4(1+\theta_k)L_1[G(x_k, v^\star) - G(x^\star, v^\star)] \\ & \quad +4L_1[G(\bar{x}, v^\star) - G(x^\star, v^\star)) \\ \mathsf{E}_{j_k}\|t_k - \nabla \ell(u_k)\|^2 & \leq 2L_2(\theta_k^2 + \theta_k)\|v_k - v_{k-1}\|^2 + 4(1+\theta_k)L_1[G(x^\star, v^\star) - G(x^\star, v_k)] \\ & \quad +4L_1[G(x^\star, v^\star) - G(x^\star, \bar{v})). \end{cases}$$

Proof. We take the conditional expectation with respect to i_k to obtain

$$\mathsf{E}_{i_k}[\frac{1}{nq_{i_k}}\nabla h_{i_k}(y_k)] = \sum_{i=1}^{n}\frac{q_i}{nq_i}\nabla h_i(y_k) = \sum_{i=1}^{n}\frac{1}{n}\nabla h_i(y_k) = \nabla h(y_k).$$

Similarly, we have $\mathsf{E}_{i_k}[(1/(nq_{i_k}))\nabla h_{i_k}(\bar{x})] = \nabla h(\bar{x})$. Therefore,

$$\mathsf{E}_{i_k}[z_k] = \mathsf{E}_{i_k}[\frac{\nabla h_{i_k}(y_k) - \nabla h_{i_k}(\bar{x})}{nq_{i_k}} + \nabla h(\bar{x})] = \nabla h(y_k).$$

Using the same argument, we also obtain $\mathsf{E}_{j_k}[t_k] = \nabla \ell(u_k)$. Hence, (3.1) is proved. Next, we bound the variance.

$$\begin{aligned}
\mathsf{E}_{i_k}\left[\|z_k - \nabla h(y_k)\|^2\right] &= \mathsf{E}_{i_k}\left[\|\frac{1}{nq_{i_k}}(\nabla h_{i_k}(y_k) - \nabla h_{i_k}(\bar{x})) + \nabla h(\bar{x}) - \nabla h(y_k)\|^2\right] \\
&= \mathsf{E}_{i_k}\left[\frac{1}{(nq_{i_k})^2}\|\nabla h_{i_k}(y_k) - \nabla h_{i_k}(\bar{x})\|^2 - \|\nabla h(y_k) - \nabla h(\bar{x})\|^2\right] \\
&\leq \mathsf{E}_{i_k}\left[\frac{1}{(nq_{i_k})^2}\|\nabla h_{i_k}(y_k) - \nabla h_{i_k}(\bar{x})\|^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{nq_i}\|\nabla h_i(y_k) - \nabla h_i(\bar{x})\|^2 \\
&\leq \frac{2}{n}\sum_{i=1}^{n}\frac{1}{nq_i}(\|\nabla h_i(y_k) - \nabla h_i(x^\star)\|^2 + \|\nabla h_i(\bar{x}) - \nabla h_i(x^\star)\|^2). \qquad (3.2)
\end{aligned}$$

We have

$$\begin{aligned}
\|\nabla h_i(y_k) - \nabla h_i(x^\star)\|^2 &= \|\nabla h_i(y_k) - \nabla h_i(x_k)\|^2 + \|\nabla h_i(x_k) - \nabla h_i(x^\star)\|^2 \\
&\quad + 2\langle \nabla h_i(y_k) - \nabla h_i(x_k) \mid \nabla h_i(x_k) - \nabla h_i(x^\star)\rangle \\
&\leq \|\nabla h_i(y_k) - \nabla h_i(x_k)\|^2 + \|\nabla h_i(x_k) - \nabla h_i(x^\star)\|^2 \\
&\quad + 2\|\nabla h_i(y_k) - \nabla h_i(x_k)\|\|\nabla h_i(x_k) - \nabla h_i(x^\star)\| \\
&\leq \|\nabla h_i(y_k) - \nabla h_i(x_k)\|^2 + \|\nabla h_i(x_k) - \nabla h_i(x^\star)\|^2 \\
&\quad + 2\|\nabla h_i(y_k) - \nabla h_i(x_k)\|\|\nabla h_i(x_k) - \nabla h_i(x^\star)\|.
\end{aligned}$$

Using the Lipschitzianity of ∇h_i, and $y_k - x_k = \theta_k(x_k - x_{k-1})$, we derive

$$\begin{aligned}
\|\nabla h_i(y_k) - \nabla h_i(x^\star)\|^2 &\leq \mu_i^2\theta_k^2\|x_k - x_{k-1}\|^2 + \|\nabla h_i(x_k) - \nabla h_i(x^\star)\|^2 \\
&\quad + 2\mu_i\theta_k\|x_k - x_{k-1}\|\|\nabla h_i(x_k) - \nabla h_i(x^\star)\|. \\
&\leq \mu_i^2\theta_k^2\|x_k - x_{k-1}\|^2 + \|\nabla h_i(x_k) - \nabla h_i(x^\star)\|^2 \\
&\quad + \theta_k(\mu_i^2\|x_k - x_{k-1}\|^2 + \|\nabla h_i(x_k) - \nabla h_i(x^\star)\|^2) \\
&= \mu_i^2(\theta_k^2 + \theta_k)\|x_k - x_{k-1}\|^2 + (1 + \theta_k)\|\nabla h_i(x_k) - \nabla h_i(x^\star)\|^2. \qquad (3.3)
\end{aligned}$$

Relations (3.2) and (3.3) together with Lemma 3.3 imply that

$$\begin{aligned}
\mathsf{E}_{i_k}\left[\|z_k - \nabla h(y_k)\|^2\right] &\leq \frac{2}{n}\sum_{i=1}^{n}\frac{1}{nq_i}(\mu_i^2(\theta_k^2 + \theta_k)\|x_k - x_{k-1}\|^2 + (1 + \theta_k)\|\nabla h_i(x_k) - \nabla h_i(x^\star)\|^2) \\
&\quad + \frac{2}{n}\sum_{i=1}^{n}\frac{1}{nq_i}\|\nabla h_i(\bar{x}) - \nabla h_i(x^\star)\|^2 \\
&\leq \frac{2(\theta_k^2 + \theta_k)}{n}\sum_{i=1}^{n}\frac{\mu_i^2}{nq_i}\|x_k - x_{k-1}\|^2 + 4(1 + \theta_k)L_Q[G(x_k, v^\star) - G(x^\star, v^\star)] \\
&\quad + 4L_Q[G(\bar{x}, v^\star) - G(x^\star, v^\star)] \\
&\leq 2L_2(\theta_k^2 + \theta_k)\|x_k - x_{k-1}\|^2 + 4(1 + \theta_k)L_1[G(x_k, v^\star) - G(x^\star, v^\star)] \\
&\quad + 4L_1[G(\bar{x}, v^\star) - G(x^\star, v^\star)].
\end{aligned}$$

Here, the second inequality is obtained by using Lemma 3.3.
Similarly, we also have

$$\mathsf{E}_{j_k}\left[\|t_k - \nabla\ell(u_k)\|^2\right] \leq 2L_2(\theta_k^2+\theta_k)\|v_k - v_{k-1}\|^2 + 4(1+\theta_k)L_1[G(x^\star, v^\star) - G(x^\star, v_k)] + 4L_1[G(x^\star, v^\star) - G(x^\star, \bar{v})).$$

Hence, the proof is completed. □

Remark 3.5 In Corollary 3.4, when $\theta_k \equiv 0$ or $\theta_k \equiv 1$, it is shown in [25, Corollary 3.4].

The gradient of the functions h and ℓ are, respectively, μ and ν-Lipschitz continuous, where $\mu = \frac{1}{n}\sum_{i=1}^{n}\mu_i$, and $\nu = \frac{1}{n'}\sum_{j=1}^{n'}\nu_j$.

Lemma 3.6 *Suppose that Assumption 1.2 is satisfied. Let $(x_k)_{k\in\mathbb{N}}, (y_k)_{k\in\mathbb{N}}$, $(u_k)_{k\in\mathbb{N}}, (v_k)_{k\in\mathbb{N}}$, $(z_k)_{k\in\mathbb{N}}$, $(t_k)_{k\in\mathbb{N}}$ be sequences generated by Algorithm 3.2. Let $\mathsf{x} = (x, v) \in \mathcal{H}\times\mathcal{G}$ and set*

$$\begin{cases} \hat{x}_{k+1} &= (\mathrm{Id} + \gamma_k\partial f)^{-1}(x_k - \gamma_k\nabla h(y_k) - \gamma_k K^* u_k), \\ \hat{v}_{k+1} &= (\mathrm{Id} + \gamma_k\partial g^*)^{-1}(v_k - \gamma_k\nabla\ell(u_k) + \gamma_k K y_k). \end{cases}$$

Define

$$(\forall k\in\mathbb{N})\begin{cases} \mathsf{x}_k &= (x_k, v_k),\ \mathsf{y}_k = (y_k, u_k),\ \hat{\mathsf{x}}_k = (\hat{x}_k, \hat{v}_k), \\ r_k &= (z_k, t_k), \\ \mathsf{R}_k &= (\nabla h(y_k), \nabla\ell(u_k)), \\ \mathbf{L}\colon &\mathcal{H}\times\mathcal{G}\to\mathcal{H}\times\mathcal{G}\colon (x,v)\mapsto(K^*v, -Kx), \\ b_k(\mathsf{x}) &= \langle \mathbf{L}(\mathsf{x}_k - \mathsf{x}_{k-1}), \mathsf{x}_k - \mathsf{x}\rangle. \end{cases} \tag{3.4}$$

Set $\mu_0 = \max\{\mu, \nu\}$. Then, the following inequality holds $(\forall k\in\{0,\ldots,m-1\})$,

$$\begin{aligned} 2\gamma_k[G(x_{k+1}, v) - G(x, v_{k+1})] \leq{}& 2\gamma_k b_{k+1}(\mathsf{x}) - 2\gamma_k\theta_k b_k(\mathsf{x}) + \|\mathsf{x}_k - \mathsf{x}\|^2 - \|\mathsf{x}_{k+1} - \mathsf{x}\|^2 \\ &- (1 - \gamma_k\theta_k\|K\| - \gamma_k\mu_0(1+\theta_k))\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2 \\ &+ (\gamma_k\theta_k\|K\| + \gamma_k\mu_0(\theta_k^2+\theta_k))\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2 \\ &+ 2\gamma_k^2\|r_k - \mathsf{R}_k\|^2 + 2\gamma_k\langle\hat{\mathsf{x}}_{k+1} - \mathsf{x}, \mathsf{R}_k - r_k\rangle. \end{aligned} \tag{3.5}$$

Proof. Let $k\in\overline{0, m-1}$. We have $v_{k+1} = (\mathrm{Id} + \gamma_k\partial g^\star)^{-1}(v_k - \gamma_k t_k + \gamma_k K y_k)$, which is equivalent to

$$Ky_k - t_k + \frac{1}{\gamma_k}(v_k - v_{k+1}) \in \partial g^\star(v_{k+1}).$$

Since $g^\star$ is convex, which implies that, for every $v\in\mathcal{G}$,

$$g^\star(v) \geq g^\star(v_{k+1}) + \left\langle Ky_k - t_k + \frac{1}{\gamma_k}(v_k - v_{k+1}) \mid v - v_{k+1}\right\rangle.$$

Therefore,

$$\begin{aligned} g^\star(v_{k+1}) - g^\star(v) &\leq \langle t_k - Ky_k \mid v - v_{k+1}\rangle + \frac{1}{\gamma_k}\langle v_k - v_{k+1} \mid v_{k+1} - v\rangle \\ &= \langle t_k - Ky_k \mid v - v_{k+1}\rangle + \frac{1}{2\gamma_k}(\|v - v_k\|^2 - \|v_{k+1} - v_k\|^2 - \|v - v_{k+1}\|^2). \end{aligned} \tag{3.6}$$

Since ℓ is convex and continuously differentiable with ν-Lipschitz gradient, we have

$$\ell(v_{k+1}) - \ell(v) \leq \langle v_{k+1} - v, \nabla\ell(u_k)\rangle + \frac{\nu}{2}\|v_{k+1} - u_k\|^2. \tag{3.7}$$

We derive from (3.6) and (3.7) that

$$\begin{aligned} G(x_{k+1}, v) - G(x_{k+1}, v_{k+1}) &= \langle Kx_{k+1} \mid v - v_{k+1}\rangle - g^\star(v) + g^\star(v_{k+1}) - \ell(v) + \ell(v_{k+1}) \\ &\leq \langle K(x_{k+1} - y_k) \mid v - v_{k+1}\rangle + \frac{1}{2\gamma_k}(\|v - v_k\|^2 - \|v_{k+1} - v_k\|^2 - \|v - v_{k+1}\|^2) \\ &\quad + \frac{\nu}{2}\|v_{k+1} - u_k\|^2 + \langle \nabla\ell(u_k) - t_k \mid v_{k+1} - v\rangle\,. \end{aligned} \tag{3.8}$$

Similar to (3.8), we have, for every $x \in \mathcal{H}$,

$$\begin{aligned} G(x_{k+1}, v_{k+1}) - G(x, v_{k+1}) &= h(x_{k+1}) - h(x) + \langle K(x_{k+1} - x) \mid v_{k+1}\rangle + f(x_{k+1}) - f(x) \\ &\leq \langle K(x_{k+1} - x) \mid v_{k+1} - u_k\rangle + \frac{1}{2\gamma_k}(\|x - x_k\|^2 - \|x_{k+1} - x_k\|^2 - \|x - x_{k+1}\|^2) \\ &\quad + \frac{\mu}{2}\|x_{k+1} - y_k\|^2 + \langle x_{k+1} - x \mid \nabla h(y_k) - z_k\rangle\,. \end{aligned} \tag{3.9}$$

Adding (3.8) and (3.9), we obtain

$$\begin{aligned} G(x_{k+1}, v) - G(x, v_{k+1}) &\leq \Big(\langle K(x_{k+1} - x) \mid v_{k+1} - u_k\rangle + \langle K(x_{k+1} - y_k) \mid v - v_{k+1}\rangle\Big) \\ &+ \frac{1}{2\gamma_k}\Big(\|x - x_k\|^2 - \|x_{k+1} - x_k\|^2 - \|x - x_{k+1}\|^2 + \|v - v_k\|^2 - \|v_{k+1} - v_k\|^2 - \|v - v_{k+1}\|^2\Big) \\ &+ \frac{\mu}{2}\|x_{k+1} - y_k\|^2 + \frac{\nu}{2}\|v_{k+1} - u_k\|^2 + \langle x_{k+1} - x \mid \nabla h(y_k) - z_k\rangle + \langle \nabla\ell(u_k) - t_k \mid v_{k+1} - v\rangle\,. \end{aligned} \tag{3.10}$$

The first term in the right-hand side of (3.10) can be expressed as

$$\begin{aligned} \langle K(x_{k+1} - x) \mid v_{k+1} - u_k\rangle &= \langle K(x_{k+1} - x) \mid v_{k+1} - v_k - \theta_k(v_k - v_{k-1})\rangle \\ &= \langle K(x_{k+1} - x) \mid v_{k+1} - v_k\rangle - \theta_k \langle K(x_{k+1} - x) \mid v_k - v_{k-1}\rangle \\ &= \langle K(x_{k+1} - x) \mid v_{k+1} - v_k\rangle - \theta_k \langle K(x_k - x) \mid v_k - v_{k-1}\rangle \\ &\quad - \theta_k \langle K(x_{k+1} - x_k) \mid v_k - v_{k-1}\rangle\,. \end{aligned} \tag{3.11}$$

Similar to (3.11), for the second term of (3.10), we also have

$$\begin{aligned} \langle K(x_{k+1} - y_k) \mid v - v_{k+1}\rangle &= \langle K(x_{k+1} - x_k) \mid v - v_{k+1}\rangle - \theta_k \langle K(x_k - x_{k-1}) \mid v - v_k\rangle \\ &\quad - \theta_k \langle K(x_k - x_{k-1}) \mid v_k - v_{k+1}\rangle\,. \end{aligned} \tag{3.12}$$

For the next to the last term in (3.10), we rewrite the formulas of $\hat{x}_{k+1}$ and x_{k+1} as

$$\begin{cases} \hat{x}_{k+1} &= (\mathrm{Id} + \gamma_k \partial f)^{-1}(x_k - \gamma_k \nabla h(y_k) - \gamma_k K^\star u_k), \\ x_{k+1} &= (\mathrm{Id} + \gamma_k \partial f)^{-1}(x_k - \gamma_k z_k - \gamma_k K^\star u_k). \end{cases}$$

Using the non-expansiveness property of prox_f, we have

$$\|\hat{x}_{k+1} - x_{k+1}\| \leq \gamma_k \|z_k - \nabla h(y_k)\|. \tag{3.13}$$

In turn,

$$\begin{aligned}
&\langle x_{k+1} - x \mid \nabla h(y_k) - z_k \rangle \\
&\quad = \langle x_{k+1} - \hat{x}_{k+1} \mid \nabla h(y_k) - z_k \rangle + \langle \hat{x}_{k+1} - x \mid \nabla h(y_k) - z_k \rangle \\
&\quad \le \|z_k - \nabla h(y_k)\| \|x_{k+1} - \hat{x}_{k+1}\| + \langle \hat{x}_{k+1} - x \mid \nabla h(y_k) - z_k \rangle \\
&\quad \le \gamma_k \|z_k - \nabla h(y_k)\|^2 + \langle \hat{x}_{k+1} - x \mid \nabla h(y_k) - z_k \rangle \,. \qquad (3.14)
\end{aligned}$$

By the same way,

$$\langle \nabla \ell(u_k) - t_k \mid v_{k+1} - v \rangle \le \gamma_k \|t_k - \nabla \ell(u_k)\|^2 + \langle \nabla \ell(u_k) - t_k \mid \hat{v}_{k+1} - v \rangle \,. \qquad (3.15)$$

From (3.11), (3.12) and the definitions provided in (3.4), we derive the following identity

$$\langle K(x_{k+1} - x) \mid v_{k+1} - u_k \rangle + \langle K(x_{k+1} - y_k) \mid v - v_{k+1} \rangle = b_{k+1} - \theta_k b_k + \theta_k \langle \mathbf{L}(\mathsf{x}_{k+1} - \mathsf{x}_k), \mathsf{x}_k - \mathsf{x}_{k-1} \rangle \qquad (3.16)$$

Therefore, from (3.14), (3.15), (3.16), using $\mu_0 = \max\{\mu, \nu\}$, (3.10) implies

$$\begin{aligned}
2\gamma_k [G(x_{k+1}, v) - G(x, v_{k+1})] \le\ & 2\gamma_k b_{k+1} - 2\gamma_k \theta_k b_k + 2\gamma_k \theta_k \langle \mathbf{L}(\mathsf{x}_{k+1} - \mathsf{x}_k), \mathsf{x}_k - \mathsf{x}_{k-1} \rangle \\
& + \|\mathsf{x}_k - \mathsf{x}\|^2 - \|\mathsf{x}_{k+1} - \mathsf{x}\|^2 - \|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2 \\
& + \gamma_k \mu_0 \|\mathsf{x}_{k+1} - \mathsf{y}_k\|^2 + 2\gamma_k^2 \|r_k - \mathsf{R}_k\|^2 + 2\gamma_k \langle \hat{\mathsf{x}}_{k+1} - \mathsf{x}, \mathsf{R}_k - r_k \rangle \qquad (3.17)
\end{aligned}$$

Using the Cauchy-Schwartz inequality and the identity $\|\mathbf{L}\| = \|K\|$, we have

$$\begin{aligned}
\langle \mathbf{L}(\mathsf{x}_{k+1} - \mathsf{x}_k), \mathsf{x}_k - \mathsf{x}_{k-1} \rangle &\le \|\mathbf{L}\| \|\mathsf{x}_{k+1} - \mathsf{x}_k\| \|\mathsf{x}_k - \mathsf{x}_{k-1}\| \\
&\le \frac{\|K\|}{2} (\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2 + \|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2)
\end{aligned}$$

and

$$\begin{aligned}
\|\mathsf{x}_{k+1} - \mathsf{y}_k\|^2 &= \|\mathsf{x}_{k+1} - \mathsf{x}_k - \theta_k(\mathsf{x}_k - \mathsf{x}_{k-1})\|^2 \\
&\le \|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2 + \theta_k^2 \|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2 + 2\theta_k \|\mathsf{x}_{k+1} - \mathsf{x}_k\| \|\mathsf{x}_k - \mathsf{x}_{k-1}\| \\
&\le (1 + \theta_k) \|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2 + (\theta_k^2 + \theta_k) \|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2.
\end{aligned}$$

Hence, we derive from (3.17)

$$\begin{aligned}
2\gamma_k [G(x_{k+1}, v) - G(x, v_{k+1})] \le\ & 2\gamma_k b_{k+1} - 2\gamma_k \theta_k b_k + \gamma_k \theta_k \|K\| (\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2 + \|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2) \\
& + \|\mathsf{x}_k - \mathsf{x}\|^2 - \|\mathsf{x}_{k+1} - \mathsf{x}\|^2 - \|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2 \\
& + \gamma_k \mu_0 ((1 + \theta_k) \|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2 + (\theta_k^2 + \theta_k) \|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2) \\
& + 2\gamma_k^2 \|r_k - \mathsf{R}_k\|^2 + 2\gamma_k \langle \hat{\mathsf{x}}_{k+1} - \mathsf{x}, \mathsf{R}_k - r_k \rangle,
\end{aligned}$$

which implies the desired result. The proof is completed. □

3.2 Proposed algorithm

We propose the following stochastic primal-dual splitting algorithm, where the main step was already presented in Algorithm 3.2, for solving Problem 3.1.

Algorithm 3.7 Let m be a strictly positive integer, $\forall k \in \{-1, 0, 1, \dots, m\}$; let $(\gamma_k^s)_{s\in\mathbb{N}}$, and $(\theta_k^s)_{s\in\mathbb{N}}$ be bounded strictly positive sequences. Let $(\bar{x}_0, \bar{v}_0) \in \mathcal{H}\times\mathcal{G}$, $(x_0^s, x_{-1}^s)_{s\in\mathbb{N}}$, $(v_0^s, v_{-1}^s)_{s\in\mathbb{N}}$ be sequences in $\mathcal{H}^2$ and $\mathcal{G}^2$ with $x_0^0 = x_{-1}^0 = \bar{x}_0,\ v_0^0 = v_{-1}^0 = \bar{v}_0$. Let $Q = \{q_1, \dots, q_n\}$ and $Q' = \{q'_1, \dots, q'_{n'}\}$ be the probabilities on $\{1, \dots, n\}$ and $\{1, \dots, n'\}$, respectively.

For $s = 0, 1, 2, \dots$
$\bar{x} := \bar{x_s},\ x_0 := x_0^s,\ x_{-1} := x_{-1}^s$
$\bar{v} := \bar{v_s},\ v_0 := v_0^s,\ v_{-1} := v_{-1}^s$
For $k = 0, 1, \dots, m-1$
Select $i_k \in \{1, \dots, n\}$ randomly according to Q
Select $j_k \in \{1, \dots, n'\}$ randomly according to Q'

Compute

$$\begin{cases} y_k &= x_k + \theta_k(x_k - x_{k-1}) \\ z_k &= \dfrac{\nabla h_{i_k}(y_k) - \nabla h_{i_k}(\bar{x})}{q_{i_k} n} + \nabla h(\bar{x}) \\ u_k &= v_k + \theta_k(v_k - v_{k-1}) \\ t_k &= \dfrac{\nabla \ell_{j_k}(u_k) - \nabla \ell_{j_k}(\bar{v})}{q'_{j_k} n'} + \nabla \ell(\bar{v}) \end{cases}$$

where (γ_k, θ_k) stands for (γ_k^s, θ_k^s)

Update

$$\begin{cases} x_{k+1} &= (\mathrm{Id} + \gamma_k \partial f)^{-1}(x_k - \gamma_k z_k - \gamma_k K^\star u_k) \\ v_{k+1} &= (\mathrm{Id} + \gamma_k \partial g^\star)^{-1}(v_k - \gamma_k t_k + \gamma_k K y_k) \end{cases}$$

end
where for any $k \in \{0, 1, 2, \dots, m-1\}$, (y_k, z_k, u_k, t_k) stands for $(y_k^s, z_k^s,\ u_k^s, t_k^s)$
and, for any $k \in \{-1, 0, 1, 2, \dots, m\}, (x_k, v_k)$ stands for (x_k^s, v_k^s)
Update

$$\begin{cases} \bar{x}_{s+1} &= \Big(\sum_{k=0}^{m-1} \gamma_k x_{k+1}\Big) \Big/ \Big(\sum_{k=0}^{m-1} \gamma_k\Big) \\ \bar{v}_{s+1} &= \Big(\sum_{k=0}^{m-1} \gamma_k v_{k+1}\Big) \Big/ \Big(\sum_{k=0}^{m-1} \gamma_k\Big). \end{cases}$$

end

Note that in Algorithm 3.7, the full gradients $\nabla h(\bar{x})$ and $\nabla \ell(\bar{v})$ have to be computed only once per iteration s.

Related work. Recently, a series of primal-dual stochastic methods to solve the convex-concave saddle point problems have been proposed; see [1, 10, 20, 22, 24, 26, 30, 31, 33, 37, 38] for instances and the references therein. These methods are different from our proposed algorithm. We highlight here the comparisons to the one in [25] which is the closest to us. Basically, Algorithm 3.7 shares the same structure as the one proposed in [25]. The main differences are listed below.

(i) The work in [25] considered only the case when $\theta_k \equiv 0$ or $\theta_k \equiv 1$ and $\gamma_k \equiv \gamma$ in non-Euclidean spaces with Bregman distances.

(ii) The updating rule of $(\overline{x}_{s+1}, \overline{v}_{s+1})$ in the strongly convex case is different from the work in [25].

3.3 Convergence results

The convergence of Algorithm 3.7 depends on the choices of $(\theta_k^s)_{s\in\mathbb{N}}$, and $(\gamma_k^s)_{s\in\mathbb{N}}$ $(\forall k \in \{0,\ldots,m-1\}, m \in \mathbb{N})$ as well as the choices of $(x_0^s, x_{-1}^s)_{s\in\mathbb{N}}$, and $(v_0^s, v_{-1}^s)_{s\in\mathbb{N}}$. Here below, we prove the almost sure weak convergence of the sequence $(\bar{x}_s, \bar{v}_s)_{s\in\mathbb{N}}$ in the general case, i.e., the Lagrangian function G is convex-concave.

Our proof technique of the almost sure weak convergence relies on the reduction of the variance with respect to the Lagrangian function G. A main advantage of this approach is that one can remove the condition imposed on the summability of the variance in [26].

Theorem 3.8 *Suppose Assumption 1.2 is satisfied. Let $(\bar{x}_s)_{s\in\mathbb{N}}, (\bar{v}_s)_{s\in\mathbb{N}}$ be sequences generated by Algorithm 3.7 with $x_{-1}^{s+1} = x_{m-1}^s$, $x_0^{s+1} = x_m^s$. Assume that*

(i) $(\forall s \in \mathbb{N})$,

$$\begin{cases} \gamma_{k+1}^s = \gamma_k^s/\theta_{k+1}^s \ \forall k \in \{0,\ldots,m-1\}, \\ \gamma_m^s = \gamma_{m-1}^s = \gamma_0^{s+1}, \\ \theta_m^s = \theta_0^{s+1} = 1. \end{cases} \tag{3.18}$$

and

$$\sum_{k=0}^{m-1} \gamma_k^s \geq \sum_{k=0}^{m-1} \gamma_k^{s+1}. \tag{3.19}$$

(ii) *There exist positive constants $c, \alpha, \gamma, \theta$ such that*

$$(\forall s \in \mathbb{N},\ \forall k \in \{0,\ldots,m\}) \begin{cases} \alpha \leq \gamma_k^s \leq \gamma \\ \theta_k^s \leq \theta \\ (\gamma_k^s)^2 \leq c\gamma\gamma_{k-1}^s \end{cases} \tag{3.20}$$

and

$$\begin{cases} 2\gamma\theta\|K\| + \gamma\mu_0(\theta+1)^2 + 4(\theta^2+\theta)L_2\gamma^2 < 1 \\ 4L_1\gamma((1+\theta)c+1) < 1. \end{cases} \tag{3.21}$$

Then $(\bar{x}_s, \bar{v}_s)_{s\in\mathbb{N}}$ *converges weakly to a point in* $\mathcal{S}$ *a.s..*

Proof. At stage s, for $k \in \{0, 1, 2, \ldots, m-1\}$, we rewrite (3.5) with $x = x^\star$, $v = v^\star$, using $\gamma_k = \theta_{k+1}\gamma_{k+1}$, we get

$$\begin{aligned} 2\gamma_k[G(x_{k+1}, v^\star) - G(x^\star, v_{k+1})] \leq\ & 2\gamma_{k+1}\theta_{k+1}b_{k+1} - 2\gamma_k\theta_k b_k + \|\mathsf{x}_k - \mathsf{x}^\star\|^2 - \|\mathsf{x}_{k+1} - \mathsf{x}^\star\|^2 \\ & - (1 - \gamma_k\theta_k\|K\| - \gamma_k\mu_0(1+\theta_k))\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2 \\ & + (\gamma_k\theta_k\|K\| + \gamma_k\mu_0(\theta_k^2+\theta_k))\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2 \\ & + 2\gamma_k^2\|r_k - \mathsf{R}_k\|^2 + 2\gamma_k\langle \hat{\mathsf{x}}_{k+1} - \mathsf{x}^\star, \mathsf{R}_k - r_k\rangle. \end{aligned} \tag{3.22}$$

Denote $\xi_k = (i_k, j_k)$. Let $\mathsf{E}_{\xi_k} := \mathsf{E}_{\xi_k^s}$ be the conditional expectation with respect to the history $\{(i_0, j_0), \ldots, (i_{k-1}, j_{k-1})\}$. Using Corollary 3.4 and the fact $\hat{\mathsf{x}}_{k+1}$ is ξ_k-measurable, we derive

$$\mathsf{E}_{\xi_k}[\|r_k - \mathsf{R}_k\|^2] \leq 2L_2(\theta_k^2+\theta_k)\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2 + 4L_1(1+\theta_k)[G(x_k, v^\star) - G(x^\star, v_k)] + 4L_1[G(\bar{x}, v^\star) - G(x^\star, \bar{v})].$$

Therefore, inequality (3.22) implies

$$\begin{aligned} & 2\gamma_k\mathsf{E}_{\xi_k}[G(x_{k+1}, v^\star) - G(x^\star, v_{k+1})] \\ & \quad \leq \|\mathsf{x}_k - \mathsf{x}^\star\|^2 - 2\gamma_k\theta_k b_k - \mathsf{E}_{\xi_k}[\|\mathsf{x}_{k+1} - \mathsf{x}^\star\|^2] + 2\gamma_{k+1}\theta_{k+1}\mathsf{E}_{\xi_k}[b_{k+1}] \\ & \qquad - (1 - \gamma\theta\|K\| - \gamma\mu_0(1+\theta))\mathsf{E}_{\xi_k}[\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2] \\ & \qquad + (\gamma\theta\|K\| + \gamma\mu_0(\theta^2+\theta) + 4L_2(\theta^2+\theta)\gamma^2)\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2 \\ & \qquad + 8(1+\theta)L_1\gamma_k^2[G(x_k, v^\star) - G(x^\star, v_k)] + 8L_1\gamma_k^2[G(\bar{x}, v^\star) - G(x^\star, \bar{v})] \\ & \quad \leq e_k - \mathsf{E}_{\xi_k}[e_{k+1}] + 8(1+\theta)L_1\gamma_k^2[G(x_k, v^\star) - G(x^\star, v_k)] \\ & \qquad + (2\gamma\theta\|K\| + \gamma\mu_0(\theta+1)^2 + 4(\theta^2+\theta)L_2\gamma^2 - 1)\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2 \\ & \qquad + 8L_1\gamma_k^2[G(\bar{x}, v^\star) - G(x^\star, \bar{v})], \end{aligned} \tag{3.23}$$

where

$$\begin{aligned} e_k := e_k^s &= \|\mathsf{x}_k^s - \mathsf{x}^\star\|^2 - 2\gamma_k^s\theta_k^s b_k^s + (1 - \gamma\theta\|K\| - \gamma\mu_0(1+\theta))\|\mathsf{x}_k^s - \mathsf{x}_{k-1}^s\|^2 \\ &= \|\mathsf{x}_k - \mathsf{x}^\star\|^2 - 2\gamma_k\theta_k b_k + (1 - \gamma\theta\|K\| - \gamma\mu_0(1+\theta))\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2. \end{aligned}$$

Using the Cauchy-Schwarz inequality, we have that

$$\begin{aligned} 2\gamma_k\theta_k b_k &= 2\gamma_k\theta_k\langle \mathbf{L}(\mathsf{x}_k - \mathsf{x}_{k-1}) \mid \mathsf{x}_k - \mathsf{x}^\star\rangle \\ &\leq 2\gamma\theta\|K\|\|\mathsf{x}_k - \mathsf{x}_{k-1}\|\|\mathsf{x}_k - \mathsf{x}^\star\| \\ &\leq \gamma\theta\|K\|(\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2 + \|\mathsf{x}_k - \mathsf{x}^\star\|^2). \end{aligned}$$

Hence, from the first condition in (3.21), we obtain

$$e_k \geq (1 - \gamma\theta\|K\|)\|\mathsf{x}_k - \mathsf{x}^\star\|^2 + (1 - 2\gamma\theta\|K\| - \gamma\mu_0(1+\theta))\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2 \geq \frac{\|\mathsf{x}_k - \mathsf{x}^\star\|^2}{2} \geq 0. \tag{3.24}$$

Set $S_k := S_k^s = G(x_k^s, v^\star) - G(x^\star, v_k^s) = G(x_k, v^\star) - G(x^\star, v_k)$. Taking the expectation with respect to all the history in the stage s (denote the resulting expectation by E_s), summing the inequality (3.23) from $k = 0$ to $m-1$, and using the condition (3.21), i.e. $2\gamma\theta\|K\| + \gamma\mu_0(\theta+1)^2 + 4(\theta^2+\theta)L_2\gamma^2 - 1 < 0$, we obtain

$$\begin{aligned}
2\sum_{k=0}^{m-1}\gamma_k\mathsf{E}_s[S_{k+1}] &\le e_0 - \mathsf{E}_s[e_m] + 8(1+\theta)L_1\sum_{k=0}^{m-1}\gamma_k^2\mathsf{E}_s[S_k] + 8L_1\sum_{k=0}^{m-1}\gamma_k^2\big[G(\bar{x}_s, v^\star) - G(x^\star, \bar{v}_s)\big] \\
&\le (e_0 + 8(1+\theta)L_1\gamma_0^2 S_0) - (\mathsf{E}_s[e_m] + 8(1+\theta)L_1\gamma_m^2\mathsf{E}_s[S_m]) + 8(1+\theta)L_1\sum_{k=1}^{m}\gamma_k^2\mathsf{E}_s[S_k] \\
&\quad + 8L_1\sum_{k=0}^{m-1}\gamma_k^2\big[G(\bar{x}_s, v^\star) - G(x^\star, \bar{v}_s)\big]. \qquad (3.25)
\end{aligned}$$

Using the condition (3.20), we derive from (3.25)

$$\begin{aligned}
2(1 - 4(1+\theta)c\gamma.L_1)\sum_{k=0}^{m-1}\gamma_k\mathsf{E}_s[S_{k+1}] &\le (e_0 + 8(1+\theta)L_1\gamma_0^2 S_0) - (\mathsf{E}_s[e_m] + 8(1+\theta)L_1\gamma_m^2\mathsf{E}_s[S_m]) \\
&\quad + 8L_1\gamma\sum_{k=0}^{m-1}\gamma_k\big[G(\bar{x}_s, v^\star) - G(x^\star, \bar{v}_s)\big].
\end{aligned}$$

Using the convex-concave property of the Lagrangian function G, we obtain

$$\begin{aligned}
&2\sum_{k=0}^{m-1}\gamma_k(1 - 4(1+\theta)c\gamma L_1)\mathsf{E}_s[G(\bar{x}_{s+1}, v^\star) - G(x^\star, \bar{v}_{s+1})] \\
&\le (e_0 + 8(1+\theta)L_1\gamma_0^2 S_0) - (\mathsf{E}_s[e_m] + 8(1+\theta)L_1\gamma_m^2\mathsf{E}_s[S_m]) + 8L_1\gamma\sum_{k=0}^{m-1}\gamma_k\big[G(\bar{x}_s, v^\star) - G(x^\star, \bar{v}_s)\big]. \qquad (3.26)
\end{aligned}$$

Set $T^s = \sum_{k=0}^{m-1}\gamma_k^s$. It follows from condition (3.19) that $T^s \ge T^{s+1}$ $(\forall s \in \mathbb{N})$. Hence, we obtain from (3.26)

$$\begin{aligned}
&2T^{s+1}((1 - 4(1+\theta)c\gamma L_1)\mathsf{E}_s[G(\bar{x}_{s+1}, v^\star) - G(x^\star, \bar{v}_{s+1})] \\
&\le (e_0 + 8(1+\theta)L_1\gamma_0^2 S_0) - (\mathsf{E}_s[e_m] + 8(1+\theta)L_1\gamma_m^2\mathsf{E}_s[S_m]) + 8L_1\gamma T^s\big[G(\bar{x}_s, v^\star) - G(x^\star, \bar{v}_s)\big]. \qquad (3.27)
\end{aligned}$$

Note that by the choices of $(x_0^s)_{s\in\mathbb{N}}$ and $(x_{-1}^s)_{s\in\mathbb{N}}$, we have

$$\mathsf{E}_s[e_m^s] + 8(1+\theta)L_1(\gamma_m^s)^2\mathsf{E}_s[S_m^s] = \mathsf{E}_s[e_0^{s+1}] + 8(1+\theta)L_1(\gamma_0^{s+1})^2\mathsf{E}_s[S_0^{s+1}].$$

Hence, we can rewrite (3.27) as

$$\begin{aligned}
&\mathsf{E}_s\big[e_0^{s+1} + 8(1+\theta)L_1(\gamma_0^{s+1})^2 S_0^{s+1} + 2T^{s+1}((1 - 4(1+\theta)c\gamma L_1)(G(\bar{x}_{s+1}, v^\star) - G(x^\star, \bar{v}_{s+1}))\big] \\
&\quad \le e_0^s + 8(1+\theta)L_1(\gamma_0^s)^2 S_0^s + 2T^s(1 - 4(1+\theta)c\gamma L_1)(G(\bar{x}_s, v^\star) - G(x^\star, \bar{v}_s)) \\
&\quad\quad + 2T^s(4L_1(1+\theta)c\gamma + 4L_1\gamma - 1)(G(\bar{x}_s, v^\star) - G(x^\star, \bar{v}_s)). \qquad (3.28)
\end{aligned}$$

Using condition (3.21) and Lemma 2.5, we derive from the above inequality that

$$e_0^s + 8(1+\theta)L_1(\gamma_0^s)^2 S_0^s + 2T^s(1 - 4(1+\theta)c\gamma L_1)(G(\bar{x}_s, v^\star) - G(x^\star, \bar{v}_s)) \text{ converges } a.s.$$

and

$$\sum_{s\in\mathbb{N}}(G(\bar{x}_s, v^\star) - G(x^\star, \bar{v}_s)) < +\infty \ a.s..$$

Let us consider the stage $s+1$, $\forall k \in \{1, 2, \ldots, m\}$, it follows from (3.23) and condition (3.21) that

$$\begin{aligned}\mathsf{E}_{\xi_{k-1}^{s+1}}[2\gamma_{k-1}(G(x_k, v^\star) - G(x^\star, v_k)) + e_k] \leq\ & 2\gamma_{k-2}(G(x_{k-1}, v^\star) - G(x^\star, v_{k-1})) + e_{k-1} \\ & + 8\gamma^2 L_1[G(\bar{x}_{s+1}, v^\star) - G(x^\star, \bar{v}_{s+1})].\end{aligned}$$

Let $\mathcal{F}_k^s$ be $\sigma-$ algebra generated by (x_i^j, v_i^j) for all $i \in \{-1, 0, \ldots, m\}$, $j \in \{0, 1, \ldots, s-1\}$ and (x_i^s, v_i^s) for all $i = \{-1, 0, 1, \ldots, k\}$. Taking the conditional expectation on both sides of the above inequality sequentially k times, we deduce

$$\begin{aligned}&\mathsf{E}[2\gamma_{k-1}(G(x_k^{s+1}, v^\star) - G(x^\star, v_k^{s+1})) + e_k^{s+1}|\mathcal{F}_0^{s+1}] \\ &\leq 2\gamma_{-1}(G(x_0, v^\star) - G(x^\star, v_0)) + e_0 + 8k\gamma^2 L_1[G(\bar{x}_{s+1}, v^\star) - G(x^\star, \bar{v}_{s+1})] \\ &= 2\gamma_m^s(G(x_m^s, v^\star) - G(x^\star, v_m^s)) + e_m^s + 8k\gamma^2 L_1[G(\bar{x}_{s+1}, v^\star) - G(x^\star, \bar{v}_{s+1})]. \qquad (3.29)\end{aligned}$$

In stage s, using (3.23) again, we have

$$\begin{aligned}&\mathsf{E}[2\gamma_{m-1}(G(x_m^s, v^\star) - G(x^\star, v_m^s)) + e_m^s|\mathcal{F}_k^s] \\ &\leq 2\gamma_{k-1}(G(x_k^s, v^\star) - G(x^\star, v_k^s)) + e_k^s + (8(1+\theta)L_1\gamma_k^2 - 2\gamma_{k-1})(G(x_k^s, v^\star) - G(x^\star, v_k^s)) \\ &+ (2\gamma\theta\|K\| + \gamma\mu_0(\theta+1)^2 + 4(\theta^2+\theta)L_2\gamma^2 - 1)\|\mathsf{x}_k^s - \mathsf{x}_{k-1}^s\|^2 + 8\gamma^2 L_1(m-k)[G(\bar{x}_s, v^\star) - G(x^\star, \bar{v}_s)]. \qquad (3.30)\end{aligned}$$

Combining (3.29) and (3.30), we derive, for $\forall k \in \{1, 2, \ldots, m\}$

$$\begin{aligned}&\mathsf{E}[2\gamma_{k-1}^{s+1}(G(x_k^{s+1}, v^\star) - G(x^\star, v_k^{s+1})) + e_k^{s+1}|\mathcal{F}_k^s] \\ &\leq 2\gamma_{k-1}^s(G(x_k^s, v^\star) - G(x^\star, v_k^s)) + e_k^s + (8(1+\theta)L_1(\gamma_k^s)^2 - 2\gamma_{k-1}^s)(G(x_k^s, v^\star) - G(x^\star, v_k^s)) \\ &+ (2\gamma\theta\|K\| + \gamma\mu_0(\theta+1)^2 + 4(\theta^2+\theta)L_2\gamma^2 - 1)\|\mathsf{x}_k^s - \mathsf{x}_{k-1}^s\|^2 + 8\gamma^2 L_1(m-k)[G(\bar{x}_s, v^\star) - G(x^\star, \bar{v}_s)] \\ &+ 8k\gamma^2 L_1\mathsf{E}[G(\bar{x}_{s+1}, v^\star) - G(x^\star, \bar{v}_{s+1})|\mathcal{F}_k^s]. \qquad (3.31)\end{aligned}$$

Taking the expectation on both sides of (3.28), we obtain

$$\sum_{s\in\mathbb{N}}\mathsf{E}[G(\bar{x}_s, v^\star) - G(x^\star, \bar{v}_s)] < +\infty, \qquad (3.32)$$

which implies

$$\sum_{s\in\mathbb{N}}\mathsf{E}[G(\bar{x}_{s+1}, v^\star) - G(x^\star, \bar{v}_{s+1})|\mathcal{F}_k^s] < +\infty \quad a.s..$$

From (3.31), condition (3.21), and using Lemma 2.5, the following limits exist

$$\lim_{s\to+\infty} 2\gamma_{k-1}^s(G(x_k^s, v^\star) - G(x^\star, v_k^s)) + e_k^s \ \ a.s. \qquad (3.33)$$

and

$$\lim_{s\to\infty}[G(x_k^s, v^\star) - G(x^\star, v_k^s)] = \lim_{s\to\infty} \|\mathsf{x}_k^s - \mathsf{x}_{k-1}^s\| = 0 \ a.s.. \tag{3.34}$$

Then, from (3.33) and (3.34), $\forall k \in \{1, 2, \ldots, m\}$, there exists

$$\lim_{s\to\infty} e_k^s \ \ a.s..$$

We recall the definition of e_k^s, i.e., $e_k^s = \|\mathsf{x}_k^s - \mathsf{x}^\star\|^2 - 2\gamma_k^s\theta_k^s b_k^s + (1-\gamma\theta\|K\| - \gamma\mu_0(1+\theta))\|\mathsf{x}_k^s - \mathsf{x}_{k-1}^s\|^2$. It follows from (3.24) and (3.33) that $(\|\mathsf{x}_k^s - \mathsf{x}^\star\|)_{s\in\mathbb{N}}$ is bounded, which implies $\lim_{s\to\infty} b_k^s = 0$ a.s.. Therefore, $\forall k \in \{1, 2, \ldots, m\}$, there exists

$$\lim_{s\to\infty} \|\mathsf{x}_k^s - \mathsf{x}^\star\|^2 \quad a.s..$$

Suppose that $\hat{\mathsf{x}} = (\hat{x}, \hat{v})$ is a weak sequential cluster point of the sequence $(\mathsf{x}_k^s)_{s\in\mathbb{N}}$. We rewrite (3.13) and the same inequality

$$\begin{aligned}\|\hat{x}_{k+1} - x_{k+1}\| &\le \gamma_k\|z_k - \nabla h(y_k)\|,\\ \|\hat{v}_{k+1} - v_{k+1}\| &\le \gamma_k\|t_k - \nabla\ell(u_k)\|,\end{aligned}$$

which implies

$$\|\mathsf{x}_{k+1} - \hat{\mathsf{x}}_{k+1}\|^2 \le \gamma_k^2\|r_k - \mathsf{R}_k\|^2.$$

Taking the conditional expectation E_{ξ_k} and using Corollary 3.4, we have

$$\begin{aligned}\mathsf{E}_{\xi_k}[\|\mathsf{x}_{k+1} - \hat{\mathsf{x}}_{k+1}\|^2] \le{}& 2L_2(\theta^2+\theta)\gamma_k^2\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2 + 4L_1(1+\theta)\gamma_k^2[G(x_k^s, v^\star) - G(x^\star, v_k^s)]\\ &+ 4L_1\gamma_k^2[G(\bar{x}, v^\star) - G(x^\star, \bar{v})].\end{aligned}$$

Hence

$$\begin{aligned}\mathsf{E}[\|\mathsf{x}_{k+1} - \hat{\mathsf{x}}_{k+1}\|^2] \le{}& 2L_2(\theta^2+\theta)\gamma^2\mathsf{E}[\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2] + 4L_1(1+\theta)\gamma^2\mathsf{E}[G(x_k^s, v^\star) - G(x^\star, v_k^s)]\\ &+ 4L_1\gamma^2\mathsf{E}[G(\bar{x}, v^\star) - G(x^\star, \bar{v})].\end{aligned} \tag{3.35}$$

Taking the expectation on both sides of (3.31), we derive

$$\begin{aligned}&\mathsf{E}[2\gamma_{k-1}^{s+1}(G(x_k^{s+1}, v^\star) - G(x^\star, v_k^{s+1})) + e_k^{s+1}]\\ &\le \mathsf{E}[2\gamma_{k-1}^s(G(x_k^s, v^\star) - G(x^\star, v_k^s)) + e_k^s] + (8(1+\theta)L_1(\gamma_k^s)^2 - 2\gamma_{k-1}^s)\mathsf{E}[G(x_k^s, v^\star) - G(x^\star, v_k^s)]\\ &+ (2\gamma\theta\|K\| + \gamma\mu_0(\theta+1)^2 + 4(\theta^2+\theta)L_2\gamma^2 - 1)\mathsf{E}[\|\mathsf{x}_k^s - \mathsf{x}_{k-1}^s\|^2] + 8\gamma^2L_1(m-k)\mathsf{E}[G(\bar{x}_s, v^\star) - G(x^\star, \bar{v}_s)]\\ &+ 8k\gamma^2L_1\mathsf{E}[G(\bar{x}_{s+1}, v^\star) - G(x^\star, \bar{v}_{s+1})].\end{aligned} \tag{3.36}$$

From (3.20), we have

$$8(1+\theta)L_1(\gamma_k^s)^2 - 2\gamma_{k-1}^s \le 2\gamma_{k-1}^s(4(1+\theta)c\gamma - 1) < 0.$$

Hence, using (3.32), i.e. $\sum_{s\in\mathbb{N}} \mathsf{E}[G(\bar{x}_s, v^\star) - G(x^\star, \bar{v}_s)] < +\infty$, we can deduce from (3.36) that

$$\begin{cases}\sum_{s\in\mathbb{N}} \mathsf{E}[G(x_k^s, v^\star) - G(x^\star, v_k^s)] < +\infty\\ \sum_{s\in\mathbb{N}} \mathsf{E}[\|\mathsf{x}_k^s - \mathsf{x}_{k-1}^s\|^2] < +\infty.\end{cases} \tag{3.37}$$

Therefore, (3.35) implies

$$\sum_{s\in\mathbb{N}} \mathsf{E}[\|\mathsf{x}_{k+1}^s - \hat{\mathsf{x}}_{k+1}^s\|^2] < +\infty$$

which, in turn, implies

$$\lim_{s\to\infty} \|\mathsf{x}_{k+1}^s - \hat{\mathsf{x}}_{k+1}^s\| = 0 \quad a.s..$$

Using the triangle inequality $\|\hat{\mathsf{x}}_{k+1}^s - y_k^s\| \leq \|\mathsf{x}_{k+1}^s - \hat{\mathsf{x}}_{k+1}^s\| + \|\mathsf{x}_{k+1}^s - \mathsf{x}_k^s\| + \|\mathsf{x}_k^s - \mathsf{x}_{k-1}^s\|$, we also have $\lim_{s\to\infty} \|\hat{\mathsf{x}}_{k+1}^s - \mathsf{y}_k^s\| = 0$. a.s.. From the definition of $\hat{x}_{k+1}^s$, we obtain

$$\frac{x_k^s - \hat{x}_{k+1}^s}{\gamma} - \nabla h(y_k^s) + \nabla h(\hat{x}_{k+1}^s) \in \partial f(\hat{x}_{k+1}^s) + \nabla h(\hat{x}_{k+1}^s) + K^* u_k^s$$

which implies $0 \in \partial f(\hat{x}) + \nabla h(\hat{x}) + K^\star \hat{v}$. We also have $0 \in \partial g^\star(\hat{v}) + \nabla \ell(\hat{v}) - K\hat{x}$. Therefore, $(\hat{x}, \hat{v}) \in \mathcal{S}$.

Using Lemma 2.6, the sequence $(x_k^s, v_k^s)_{s\in\mathbb{N}}$ converges weakly to a point in $\mathcal{S}$. From (3.37)

$$\sum_{s\in\mathbb{N}} \mathsf{E}[\|\mathsf{x}_k^s - \mathsf{x}_{k-1}^s\|^2] < +\infty,$$

we have that: for all $k \in \{1, 2, \ldots, m\}$, the limit of the sequence (x_k^s, v_k^s) when $s \to +\infty$ is the same a.s.. Hence, $(\bar{x}_s, \bar{v}_s)$ converges weakly to a point in $\mathcal{S}$ a.s..
The proof is completed. □

Remark 3.9 We show some cases of sequences $(\theta_k^s)_{s\in\mathbb{N}}$ and $(\gamma_k^s)_{s\in\mathbb{N}}$ that satisfy the conditions of Theorem 3.8 .

(i) In the constant-case, $\gamma_k^s = \gamma, \theta_k^s = 1$, $(\forall k, \forall s)$, the conditions of Theorem 3.8 become

$$\begin{cases} 4\gamma\mu_0 + 2\gamma\|K\| + 8L_2\gamma^2 < 1, \\ 12\gamma L_1 < 1. \end{cases} \tag{3.38}$$

In [25], with condition (3.38), the authors proved the convergence of the primal-dual function only. Here, we show the convergence of the iterative sequence.

(ii) Let $(\beta_k)_{k\in\mathbb{N}}$ be a non-increasing positive sequence. Assume there exist positive constants α, γ, θ such that

$$(\forall k \in \mathbb{N}) \quad \begin{cases} \alpha \leq \beta_k \leq \gamma \\ \frac{\gamma_k}{\gamma_{k+1}} \leq \theta. \end{cases}$$

At stage $s \in \mathbb{N}$, we choose $\gamma_k = \beta_{s(m-1)+k}$, $\theta_{k+1} = \frac{\gamma_k}{\gamma_{k+1}}$ $(\forall k \in \{0, \ldots, m-1\})$. Then the conditions (3.18), (3.19), (3.20) of Theorem 3.8 are satisfied for $c = 1$ and the condition (3.21) becomes

$$\begin{cases} 2\gamma\theta\|K\| + \gamma\mu_0(\theta+1)^2 + 4\theta\gamma^2 L_2(\theta+1) < 1 \\ 4L_1\gamma(\theta+2) < 1. \end{cases}$$

(iii) Let (ζ_k) be a positive sequence ($k \in \{0, \dots, m-1\}$). Set $\theta = \max_{i \neq j} \frac{\zeta_i}{\zeta_j}$. For $s \in \mathbb{N}$, let (γ_k^s) be a permutation of (ζ_k) such that $\gamma_{m-1}^s = \gamma_0^{s+1}$. Set $\theta_{k+1}^s = \frac{\gamma_k^s}{\gamma_{k+1}^s}$. Then the conditions (3.18), (3.19), (3.20) of Theorem 3.8 are satisfied for $c = \theta$ and the condition (3.21) becomes

$$\begin{cases} 2\gamma\theta\|K\| + \gamma\mu_0(\theta+1)^2 + 4\theta\gamma^2 L_2(\theta+1) < 1 \\ 4L_1\gamma(\theta(1+\theta)+1) < 1. \end{cases} \tag{3.39}$$

We can choose (ζ_k) that satisfies (3.39) by scaling this sequence with an arbitrarily large constant.

3.3.1 Particular case: G is $\alpha-$strongly convex-concave

For the particular case where G is $\alpha-$strongly convex-concave ($\alpha > 0$) and ($\forall s \in \mathbb{N}$), $x_0^s = x_{-1}^s = \bar{x}_s, v_0^s = v_{-1}^s = \bar{v}_s, \gamma_k^s = \gamma_k^{s+1} = \gamma_k$, ($\forall k \in \{0, \dots, m-1\}$. Here below we prove the linear convergence rate in expectation of the difference of the Lagragian function.

Theorem 3.10 *Let $(\bar{x}_s)_{s\in\mathbb{N}}$, and $(\bar{v}_s)_{s\in\mathbb{N}}$ be sequences generated by Algorithm 3.7 with $x_0^s = x_{-1}^s = \bar{x}_s$, $v_0^s = v_{-1}^s = \bar{v}_s$. Suppose that $(\gamma_k)_{k\in\mathbb{N}}$ is a non-increasing sequence. Set $\theta = \sup_{k\in\mathbb{N}} |\theta_k| < +\infty$, and $c = \sup_{k\in\mathbb{N}} |\gamma_k - \gamma_{k+1}\theta_{k+1}|/\gamma_k$. Let M be such that $M > \frac{\|K\|c}{\alpha}$. Assume that*

$$\begin{cases} \gamma_0\mu_0(\theta+1)^2 + 2\gamma_0\theta\|K\| + \gamma_0\|K\|cM + 4L_2(\theta^2+\theta)\gamma_0^2 \leq 1, \\ 1 - \dfrac{\|K\|c}{M\alpha} - 4L_1(\theta+1)\gamma_0 > 0, \end{cases} \tag{3.40}$$

and

$$\rho = \frac{1}{\alpha(1 - \frac{\|K\|c}{M\alpha} - 4L_1(\theta+1)\gamma_0)(\sum_{k=0}^{m-1}\gamma_k)} + \frac{4L_1(\sum_{k=1}^{m-1}\gamma_k^2 + (\theta+2)\gamma_0^2)}{(1 - \frac{\|K\|c}{M\alpha} - 4L_1(\theta+1)\gamma_0)(\sum_{k=0}^{m-1}\gamma_k)} < 1, \tag{3.41}$$

then

$$\mathsf{E}[G(\bar{x}_{s+1}, v^\star) - G(x^\star, \bar{v}_{s+1})] \leq \rho^{s+1}[G(\bar{x}_0, v^\star) - G(x^\star, \bar{v}_0)]. \tag{3.42}$$

Proof. First, we rewrite (3.5) with $x = x^\star$, $v = v^\star$, we get

$$\begin{aligned} 2\gamma_k[G(x_{k+1}, v^\star) - G(x^\star, v_{k+1})] \leq\ & 2\gamma_{k+1}\theta_{k+1}b_{k+1} - 2\gamma_k\theta_k b_k + \|\mathsf{x}_k - \mathsf{x}^\star\|^2 - \|\mathsf{x}_{k+1} - \mathsf{x}^\star\|^2 \\ & + 2(\gamma_k - \gamma_{k+1}\theta_{k+1})b_{k+1} \\ & - (1 - \gamma_k\theta_k\|K\| - \gamma_k\mu_0(1+\theta_k))\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2 \\ & + (\gamma_k\theta_k\|K\| + \gamma_k\mu_0(\theta_k^2+\theta_k))\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2 \\ & + 2\gamma_k^2\|r_k - \mathsf{R}_k\|^2 + 2\gamma_k\langle\hat{\mathsf{x}}_{k+1} - \mathsf{x}^\star, \mathsf{R}_k - r_k\rangle. \end{aligned} \tag{3.43}$$

We have

$$\begin{aligned}2|b_{k+1}| &= 2|\langle \mathbf{L}(\mathsf{x}_{k+1}-\mathsf{x}_k) \mid \mathsf{x}_{k+1}-\mathsf{x}^\star\rangle| \\ &\le 2\|K\|\|\mathsf{x}_{k+1}-\mathsf{x}^\star\|\|\mathsf{x}_{k+1}-\mathsf{x}_k\| \\ &\le \|K\|(M\|\mathsf{x}_{k+1}-\mathsf{x}_k\|^2+\frac{\|\mathsf{x}_{k+1}-\mathsf{x}^\star\|^2}{M}) \\ &\le \|K\|M\|\mathsf{x}_{k+1}-\mathsf{x}_k\|^2+\frac{2\|K\|}{M\alpha}[G(x_{k+1},v^\star)-G(x^\star,v_{k+1})],\end{aligned}$$

where the last inequality is derived from the $\alpha-$strongly convex-concave property of G. Then (3.43) implies that

$$\begin{aligned}2\gamma_k\Big(1-\frac{\|K\|c}{M\alpha}\Big)&[G(x_{k+1},v^\star)-G(x^\star,v_{k+1})] \\ &\le 2\gamma_{k+1}\theta_{k+1}b_{k+1}-2\gamma_k\theta_k b_k+\|\mathsf{x}_k-\mathsf{x}^\star\|^2-\|\mathsf{x}_{k+1}-\mathsf{x}^\star\|^2 \\ &\quad-(1-\gamma_k\mu_0(1+\theta_k)-\gamma_k\theta_k\|K\|-\gamma_k\|K\|cM)\|\mathsf{x}_{k+1}-\mathsf{x}_k\|^2 \\ &\quad+(\gamma_k\mu_0(\theta_k^2+\theta_k)+\gamma_k\theta_k\|K\|)\|\mathsf{x}_k-\mathsf{x}_{k-1}\|^2 \\ &\quad+2\gamma_k^2\|r_k-\mathsf{R}_k\|^2+2\gamma_k\langle\hat{\mathsf{x}}_{k+1}-\mathsf{x}^\star,\mathsf{R}_k-r_k\rangle. \qquad (3.44)\end{aligned}$$

Using Corollary 2.5, we have

$$\begin{aligned}\mathsf{E}_{\xi_k}[\|r_k-\mathsf{R}_k\|^2] &\le 2(\theta_k^2+\theta_k)L_2\|\mathsf{x}_k-\mathsf{x}_{k-1}\|^2+4(\theta_k+1)L_1[G(x_k,v^\star)-G(x^\star,v_k)] \\ &\quad+4L_1[G(\bar{x},v^\star)-G(x^\star,\bar{v})].\end{aligned}$$

Therefore, by taking the conditional expectation on both sides of (3.44) and using the condition (3.40), we get

$$\begin{aligned}2\gamma_k\Big(1-\frac{\|K\|c}{M\alpha}\Big)&\mathsf{E}_{\xi_k}[G(x_{k+1},v^\star)-G(x^\star,v_{k+1})] \\ &\le 2\gamma_{k+1}\theta_{k+1}\mathsf{E}_{\xi_k}[b_{k+1}]-2\gamma_k\theta_k b_k+\|\mathsf{x}_k-\mathsf{x}^\star\|^2-\mathsf{E}_{\xi_k}[\|\mathsf{x}_{k+1}-\mathsf{x}^\star\|^2] \\ &\quad-(1-\gamma_k\mu_0(1+\theta_k)-\gamma_k\theta_k\|K\|-\gamma_k\|K\|cM)\mathsf{E}_{\xi_k}[\|\mathsf{x}_{k+1}-\mathsf{x}_k\|^2] \\ &\quad+(\gamma_k\mu_0(\theta_k^2+\theta_k)+\gamma_k\theta_k\|K\|+4L_2\gamma_k^2(\theta_k^2+\theta_k))\|\mathsf{x}_k-\mathsf{x}_{k-1}\|^2 \\ &\quad+8L_1\gamma_k^2(\theta_k+1)[G(x_k,v^\star)-G(x^\star,v_k)]+8L_1\gamma_k^2(G(\bar{x},v^\star)-G(x^\star,\bar{v})) \\ &\le c_k-\mathsf{E}_{\xi_k}[c_{k+1}]+8L_1\gamma_k^2(\theta_k+1)[G(x_k,v^\star)-G(x^\star,v_k)]+8L_1\gamma_k^2(G(\bar{x},v^\star)-G(x^\star,\bar{v})), \qquad (3.45)\end{aligned}$$

where we set

$$c_k=\|\mathsf{x}_k-\mathsf{x}^\star\|^2-2\gamma_k\theta_k b_k+(\gamma_k\mu_0(\theta_k^2+\theta_k)+\gamma_k\theta_k\|K\|+4L_2\gamma_k^2(\theta_k^2+\theta_k))\|\mathsf{x}_k-\mathsf{x}_{k-1}\|^2.$$

We have

$$\begin{aligned}|b_k| &\le \|L\|\|\mathsf{x}_k-\mathsf{x}_{k-1}\|\|\mathsf{x}_k-\mathsf{x}^\star\| \\ &\le \frac{\|K\|}{2}(\|\mathsf{x}_k-\mathsf{x}_{k-1}\|^2+\|\mathsf{x}_k-\mathsf{x}^\star\|^2).\end{aligned}$$

So the condition (3.40) implies that $c_k\ge 0\ \ \forall k\in\mathbb{N}$.

By summing both sides of (3.45) over $k = 0, \ldots, m-1$, using the decrease of $(\gamma_k)_{k\in\mathbb{N}}$, we obtain

$$
\begin{aligned}
&2\Big(1 - \frac{\|K\|c}{M\alpha}\Big)\sum_{k=0}^{m-1}\gamma_k \mathsf{E}_s[G(x_{k+1}, v^\star) - G(x^\star, v_{k+1})] \\
&\le c_0 - \mathsf{E}_s[c_m] + 8L_1\sum_{k=0}^{m-1}(\theta_k+1)\gamma_k^2\mathsf{E}_s[G(x_k, v^\star) - G(x^\star, v_k)] + 8L_1\sum_{k=0}^{m-1}\gamma_k^2[G(\bar{x}, v^\star) - G(x^\star, \bar{v})] \\
&\le c_0 - \mathsf{E}_s[c_m] + 8L_1(\theta+1)\sum_{k=0}^{m-1}\gamma_k^2\mathsf{E}_s[G(x_{k+1}, v^\star) - G(x^\star, v_{k+1})] \\
&+ 8L_1\Big(\sum_{k=0}^{m-1}\gamma_k^2 + (\theta+1)\gamma_0^2\Big)[G(\bar{x}, v^\star) - G(x^\star, \bar{v})].
\end{aligned}
$$

For the choice $\mathsf{x}_0 = \mathsf{x}_{-1} = \bar{\mathsf{x}}$, we get $c_0 = \|\bar{\mathsf{x}} - \mathsf{x}^\star\|^2$. Thus, using the strongly convex-concave property of G, we derive

$$
\begin{aligned}
&2\Big(1 - \frac{\|K\|c}{M\alpha} - 4L_1(\theta+1)\gamma_0\Big)\sum_{k=0}^{m-1}\gamma_k\mathsf{E}_s[G(x_{k+1}, v^\star) - G(x^\star, v_{k+1})] \\
&\qquad\le \|\bar{\mathsf{x}} - \mathsf{x}^\star\|^2 + 8L_1\Big(\sum_{k=1}^{m-1}\gamma_k^2 + (\theta+2)\gamma_0^2\Big)[G(\bar{x}, v^\star) - G(x^\star, \bar{v})] \\
&\qquad\le \Big(\frac{2}{\alpha} + 8L_1\Big(\sum_{k=1}^{m-1}\gamma_k^2 + (\theta+2)\gamma_0^2\Big)\Big)[G(\bar{x}_s, v^\star) - G(x^\star, \bar{v}_s)],
\end{aligned}
$$

which is equivalent to

$$
\begin{aligned}
&\mathsf{E}_s[G(\bar{x}_{s+1}, v^\star) - G(x^\star, \bar{v}_{s+1})] \\
&\le \Bigg(\frac{1}{\alpha(1 - \frac{\|K\|c}{M\alpha} - 4L_1(\theta+1)\gamma_0)\sum_{k=0}^{m-1}\gamma_k} + \frac{4L_1\Big(\sum_{k=1}^{m-1}\gamma_k^2 + (\theta+2)\gamma_0^2\Big)}{(1 - \frac{\|K\|c}{M\alpha} - 4L_1(\theta+1)\gamma_0)\sum_{k=0}^{m-1}\gamma_k}\Bigg)[G(\bar{x}_s, v^\star) - G(x^\star, \bar{v}_s)].
\end{aligned}
$$

Using this inequality recursively, we obtain (3.42). □

Remark 3.11 Here are some remarks.

(i) For the strongly convex-concave case, we also obtain the linear convergence rate in expectation of the primal-dual function as in [25]. However, our algorithm is completely different from the one in [25].

(ii) Here are some examples where we provide some cases of the stepsizes and m ensuring $\rho < 1$.

- In case $\theta_k \equiv 0$, by choosing $0 < \gamma_0 < \frac{1}{8L_1}(1 - \frac{\|K\|}{M\alpha})$ and $\sum_{k\in\mathbb{N}}\gamma_k = +\infty$. Then, when m is

large enough, we have

$$\rho = \frac{1}{\alpha(1-\frac{\|K\|}{M\alpha}-4\gamma_0 L_1)\sum_{k=0}^{m-1}\gamma_k} + \frac{4\gamma_0^2 L_1}{(1-\frac{\|K\|}{M\alpha}-4\gamma_0 L_1)\sum_{k=0}^{m-1}\gamma_k} + \frac{4L_1\sum_{k=0}^{m-1}\gamma_k^2}{(1-\frac{\|K\|}{M\alpha}-4\gamma_0 L_1)\sum_{k=0}^{m-1}\gamma_k}$$
$$\leq \frac{1}{\alpha(1-\frac{\|K\|}{M\alpha}-4\gamma_0 L_1)\sum_{k=0}^{m-1}\gamma_k} + \frac{4\gamma_0^2 L_1}{(1-\frac{\|K\|}{M\alpha}-4\gamma_0 L_1)\sum_{k=0}^{m-1}\gamma_k} + \frac{4\gamma_0 L_1}{(1-\frac{\|K\|}{M\alpha}-4\gamma_0 L_1)} < 1.$$

- In case $\theta_k \equiv 1$ and $\gamma_k \equiv \gamma$. Then, we have $c = 0$. Hence, the conditions (3.40) and (3.41), respectively, become

$$\begin{cases} 4\gamma\mu_0 + 2\gamma\|K\| + 8L_2\gamma^2 < 1, \\ 1 - 8L_1\gamma > 0, \end{cases}$$

and

$$\rho = \frac{1}{\alpha(1-8L_1\gamma)m\gamma} + \frac{4L_1\gamma(m+2)}{(1-8L_1\gamma)m} < 1.$$

Therefore, when $0 < \gamma < \frac{1}{12L_1}$ and m is large enough, we obtain $\rho < 1$.

(iii) Note that in every stage, the choice of two initial values is different from that of Theorem 3.8.

Theorem 3.12 *Under the same conditions as Theorem 3.10. Then the gap*

$$\sup_{x\in\mathcal{H}, v\in\mathcal{G}} \Big(G(\bar{x}_s, v) - G(x, \bar{v}_s) \Big)$$

converges linearly to 0 *in expectation.*

The sequence produced by the algorithm converges (in expectation) to a point $(\bar{x}_s, \bar{v}_s)$ such that $\sup_v \inf_x G(x,v) \leq G(\bar{x}_s, \bar{v}_s) \leq \inf_x \sup_v G(x,v)$. Hence, the Saddle point Theorem applies.

Proof. From the definition of the Lagrangian function G, we have

$$\begin{aligned} &\Big(G(\bar{x}_{s+1}, v) - G(\bar{x}_{s+1}, v^\star)\Big) - \Big(G(x^\star, v) - G(x^\star, v^\star)\Big) = \langle K(\bar{x}_{s+1} - x^\star) \mid v - v^\star \rangle, \\ &\Big(G(x, \bar{v}_{s+1}) - G(x^\star, \bar{v}_{s+1})\Big) - \Big(G(x, v^\star) - G(x^\star, v^\star)\Big) = \langle K(x - x^\star) \mid \bar{v}_{s+1} - v^\star \rangle. \end{aligned} \tag{3.46}$$

Hence

$$\begin{aligned} &\Big(G(\bar{x}_{s+1}, v) - G(\bar{x}_{s+1}, v^\star)\Big) - \Big(G(x, \bar{v}_{s+1}) - G(x^\star, \bar{v}_{s+1})\Big) \\ &= G(x^\star, v) - G(x, v^\star) + \langle K(\bar{x}_{s+1} - x^\star) \mid v - v^\star \rangle - \langle K(x - x^\star) \mid \bar{v}_{s+1} - v^\star \rangle. \end{aligned} \tag{3.47}$$

Using the Cauchy-Schwarz inequality, we get

$$\begin{aligned}
&\langle K(\bar{x}_{s+1} - x^\star) \mid v - v^\star \rangle - \langle K(x - x^\star) \mid \bar{v}_{s+1} - v^\star \rangle \\
&\leq \|K\|\|\bar{x}_{s+1} - x^\star\|\|v - v^\star\| + \|K\|\|x - x^\star\|\|\bar{v}_{s+1} - v^\star\| \\
&\leq \Big(\frac{\|K\|^2}{2\alpha}\|\bar{x}_{s+1} - x^\star\|^2 + \frac{\alpha}{2}\|v - v^\star\|^2\Big) + \Big(\frac{\|K\|^2}{2\alpha}\|\bar{v}_{s+1} - v^\star\|^2 + \frac{\alpha}{2}\|x - x^\star\|^2\Big) \\
&= \frac{\|K\|^2}{2\alpha}(\|\bar{x}_{s+1} - x^\star\|^2 + \|\bar{v}_{s+1} - v^\star\|^2) + \frac{\alpha}{2}(\|x - x^\star\|^2 + \|v - v^\star\|^2). \qquad (3.48)
\end{aligned}$$

The strong convexity-concavity of the function G imply that

$$G(x, v^\star) - G(x^\star, v) \geq \frac{\alpha}{2}(\|x - x^\star\|^2 + \|v - v^\star\|^2), \qquad (3.49)$$

and

$$G(\bar{x}_{s+1}, v^\star) - G(x^\star, \bar{v}_{s+1}) \geq \frac{\alpha}{2}(\|\bar{x}_{s+1} - x^\star\|^2 + \|\bar{v}_{s+1} - v^\star\|^2). \qquad (3.50)$$

We derive from (3.47), (3.48), (3.49), and (3.50) that

$$\begin{aligned}
&\Big(G(\bar{x}_{s+1}, v) - G(x, \bar{v}_{s+1})\Big) - \Big(G(\bar{x}_{s+1}, v^\star) - G(x^\star, \bar{v}_{s+1})\Big) \\
&\qquad \leq \frac{\|K\|^2}{\alpha^2}\Big(G(\bar{x}_{s+1}, v^\star) - G(x^\star, \bar{v}_{s+1})\Big),
\end{aligned}$$

which implies

$$G(\bar{x}_{s+1}, v) - G(x, \bar{v}_{s+1}) \leq \Big(1 + \frac{\|K\|^2}{\alpha^2}\Big)\Big(G(\bar{x}_{s+1}, v^\star) - G(x^\star, \bar{v}_{s+1})\Big).$$

From the linear convergence of $G(\bar{x}_{s+1}, v^\star) - G(x^\star, \bar{v}_{s+1})$ in Theorem 3.10, we deduce that $G(\bar{x}_{s+1}, v) - G(x, \bar{v}_{s+1})$ converges linearly to 0. The proof is completed.
□

Corollary 3.13 *Under the same conditions as Theorem 3.10. The sequences $(\bar{x}_s)_{s\in\mathbb{N}}$ and $(\bar{v}_s)_{s\in\mathbb{N}}$ converges linearly in expectation to $x^\star$ and $v^\star$, respectively.*

Proof. We have

$$\|\bar{x}_s - x^\star\|^2 + \|\bar{v}_s - v^\star\|^2 \leq \frac{2}{\alpha}\Big(G(\bar{x}_s, v^\star) - G(x^\star, \bar{v}_s)\Big). \qquad (3.51)$$

Since the difference $G(\bar{x}_s, v^\star) - G(x^\star, \bar{v}_s)$ converges linearly in expectation to 0 by Theorem 3.10. We derive from (3.51) that both $(\bar{x}_s)_{s\in\mathbb{N}}$ and $(\bar{v}_s)_{s\in\mathbb{N}}$ converge linearly in expectation to $x^\star$ and $v^\star$, respectively. □

3.3.2 Related methods

Linear convergence in expectation of the primal-dual gap was established in [25, 33] for Bregman distance and for a different stochastic variance reduction algorithm. In [25, 33], the authors require

the stepsize γ_k is constant and $\theta_k = 0$ for $\forall k \in \mathbb{N}$. Our Algorithm is more general, i.e. γ_k is not constant and $\theta_k \neq 0$. We also update the value $\bar{x}_{s+1}$ which is different that in [25, 33], our update is more simple and more natural than the update of $\bar{x}_{s+1}$ in [25, 33].

The authors in [1, 16] also proposed a stochastic variance reduction algorithm for saddle point problems with linear convergence in expectation of the iterates. Here, we also obtain the linear convergence of the iteration sequence as formalized in Corollary 3.13.

For a special case of Problem 3.1 where $f = 0$, $g^\star = 0$ and $\mathcal{H} = \mathbb{R}^{d_1}$, $\mathcal{G} = \mathbb{R}^{d_2}$; under the additional assumption that the operator K is full rank, i.e. $\text{rank}(K) = d_1$, the method proposed in [18] also achieves convergence rate that is linear but only when the function ℓ is strongly convex.

In the particular case where $\theta_k \equiv 0$, $K = 0$, $\gamma_k = \gamma$, $g^\star = 0$, $\ell = 0$, our results recover [35, Theorem 1] as a special case. Indeed, the condition (3.40) becomes

$$\begin{cases} \gamma\mu \leq 1, \\ 1 - 4L_1\gamma > 0, \end{cases} \tag{3.52}$$

Following the fact that $\mu \leq L_1$, (3.52) is equivalent to the condition $\gamma < \frac{1}{4L_1}$ as in [35, Theorem1].

References

[1] P. Balamurugan and F. Bach, Stochastic Variance Reduction Methods for Saddle-Point Problems, *Advances in Neural Information Processing Systems*, pp. 1416–1424, 2016.

[2] S. R. Becker and P. L. Combettes, An algorithm for splitting parallel sums of linearly composed monotone operators, with applications to signal recovery, *J. Nonlinear Convex Anal.* , vol. 15, no. 1, pp. 137–159,

[3] R. I. Boţ and C. Hendrich, A Douglas–Rachford type primal-dual method for solving inclusions with mixtures of composite and parallel-sum type monotone operators, *SIAM J. Optim.*, Vol. 23, pp. 2541–2565, 2013.

[4] R. I. Boţ and C. Hendrich, Convergence Analysis for a Primal-Dual Monotone + Skew Splitting Algorithm with Applications to Total Variation Minimization, *J. Math. Imaging Vis.*, Vol. 49, pp. 551–568, 2014.

[5] R. I. Boţ, E.R. Csetnek and C. Hendrich, On the convergence rate improvement of a primal-dual splitting algorithm for solving monotone inclusion problems, *Math. Program.*, Vol. 150, pp. 251-279, 2015.

[6] L. M. Briceño-Arias and P. L. Combettes, A monotone+skew splitting model for composite monotone inclusions in duality, *SIAM J. Optim.*, Vol. 21, pp. 1230–1250, 2011.

[7] M. N. Bùi and P. L . Combettes, Multivariate monotone inclusions in saddle form, *Mathematics of Operations Research*, 2021.

[8] V. Cevher and B. C. Vũ, A reflected forward-backward splitting method for monotone inclusions involving Lipschitzian operators, *Set-Valued Var. Anal.,* Vol. 29, pp. 163-174, 2021.

[9] A. Chambolle, T. Pock, On the ergodic convergence rates of a first-order primal–dual algorithm, *Math. Program.*, Vol. 159, pp. 253-287, 2016.

[10] Y. Chen, G. Lan, Y. Ouyang, Optimal primal–dual methods for a class of saddle point problems, *SIAM J. Optim.*, Vol 24, pp. 1779-1814, 2014.

[11] P. L. Combettes, Systems of structured monotone inclusions: duality, algorithms, and applications, *SIAM J. Optim.*, Vol. 23, pp. 2420–2447, 2013.

[12] P. L. Combettes, L. Condat, J.-C. Pesquet, and B. C. Vũ, A forward-backward view of some primal-dual optimization methods in image recovery, *Proceedings of the IEEE International Conference on Image Processing*. Paris, France, October 27-30, 2014.

[13] P. L. Combettes and J. -C. Pesquet, Stochastic quasi-Fejér block-coordinate fixed point iterations with random sweeping, *SIAM J. Optim.*, Vol. 25, pp. 1221-1248, 2015.

[14] P. L. Combettes and J.-C. Pesquet, Primal-dual splitting algorithm for solving inclusions with mixtures of composite, Lipschitzian, and parallel-sum type monotone operators, *Set-Valued Var. Anal.*, Vol. 20, pp. 307-330, 2012.

[15] L. Condat, A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms. *J. Optim. Theory Appl.*, Vol. 158, pp. 460–479, 2013.

[16] A. M. Devraj, J. Chen, Stochastic variance reduced primal dual algorithms for empirical composition optimization, *Advances in Neural Information Processing Systems*, pp. 9882-9892, 2019.

[17] Y. Drori, S. Sabach, and M. Teboulle, A simple algorithm for a class of nonsmooth convex-concave saddle-point problems, *Oper. Res. Lett.*, Vol. 43, pp. 209-214, 2015.

[18] S. S. Du, W. Hu, Linear convergence of the primal-dual gradient method for convex-concave saddle point problems without strong convexity, *Proc. International Conference on Artificial Intelligence and Statistics*, pp. 196-205, 2019.

[19] R. M. Gower, M. Schmidt, F. Bach, P. Richtarik, Variance-Reduced Methods for Machine Learning, *Proceedings of the IEEE*, Vol. 108, 2020.

[20] E. Y. Hamedani, A. Jalilzadeh, A stochastic variance-reduced accelerated primal-dual method for finite-sum saddle-point problems, *Computational Optimization and Applications*, Vol. 85, pp. 653–679, 2023.

[21] E. Y. Hamedani, N. S. Aybat, A primal-dual algorithm with line search for general convex-concave saddle point problems, *SIAM J. Optim.*, Vol. 31(2), pp. 1299-1329, 2021.

[22] A. Juditsky, A. S. Nemirovski,C. Tauvel, Solving variational inequalities with stochastic mirror-prox algorithm, *Stochastic Systems*, Vol. 1 pp. 17-58, 2011.

[23] M. Ledoux and M. Talagrand, Probability in Banach spaces: isoperimetry and processes, Springer, New York, 1991.

[24] A. Nemirovski, A. Juditsky, G. Lan and A. Shapiro, Robust stochastic approximation approach to stochastic programming, *SIAM Journal on Optimization*, Vol. 19, pp. 1574–1609, 2009.

[25] V. D. Nguyen, B. C. Vũ, A Stochastic Variance Reduction Algorithm with Bregman Distances for Structured Composite Problems, *Optimization*, Vol. 72(6), pp. 1463–1484, 2023.

[26] V. D. Nguyen and B. C. Vũ , Convergence analysis of the stochastic reflected forward-backward splitting algorithm, *Optimization letter*, Vol. 16, pp. 2649–2679, 2022.

[27] A. Nitanda, Stochastic proximal gradient descent with acceleration techniques, *Advances in Neural Information Processing Systems*, pp. 1574–1582, 2014.

[28] T. Pethick, O. Fercoq, P. Latafat, P. Patrinos, and V. Cevher, Solving stochastic weak Minty variational inequalities without increasing batch size, *The 11th International Conference on Learning Representations* , 2023.

[29] H. Robbins and D. Siegmund, A convergence theorem for non negative almost supermartingales and some applications. In: Rustagi JS, editor. *Optimizing methods in statistic*, New York (NY): Academic Press, pp. 233-257, 1971.

[30] L. Rosasco, S. Villa, B. C. Vũ, A stochastic inertial forward-backward splitting algorithm for multivariate monotone inclusions, *Optimization*, Vol. 65, pp. 1293-1314, 2016.

[31] L. Rosasco, S. Villa, B. C. Vũ, A First-order stochastic primal-dual algorithm with correction step, *Numer. Funct. Anal. Optim.*, Vol. 38, pp. 602-626, 2017.

[32] A. Silveti-Falls, C. Molinari, and J. Fadili, A Stochastic Bregman Primal-Dual Splitting Algorithm for Composite Optimization, 2021. arXiv preprint arXiv:2112.11928.

[33] Z. Shi, X. Zhang, and Y. Yu, Bregman divergence for stochastic variance reduction: Saddle-point and adversarial prediction, *In Advances in Neural Information Processing Systems*, 2017.

[34] B. C. Vũ, A splitting algorithm for dual monotone inclusions involving cocoercive operators, *Adv. Comput. Math.*, Vol. 38, pp. 667–681, 2013.

[35] L. Xiao and T. Zhang, A proximal stochastic gradient method with progressive variance reduction, *SIAM J. Optim.*, Vol. 24, pp. 2057-2075, 2014.

[36] Z. Allen-Zhu and E. Hazan, Variance Reduction for Faster Non-Convex Optimization, *Proceedings of The 33rd International Conference on Machine Learning, PMLR*, Vol. 48, pp. 699-707, 2016.

[37] R. Zhao, Accelerated stochastic algorithms for convex-concave saddle-point problems, *Mathematics of Operation Research*, Vol. 47, pp. 1443-1473, 2021.

[38] J. Wang, L. Xiao, Exploiting strong convexity from data with primal-dual first-order algorithms, *Proceedings of the 34th International Conference on Machine Learning*; 2017, Aug 6-11; Sydney, Australia; Vol. 70, pp. 3694-3702. JMLR.org.