Doubly stochastic primal dual splitting algorithm with variance reduction for saddle point problems

Bằng Công Vũ[1], and Dimitri Papadimitriou[1,2]

[1] Belgium Research Center (BeRC) - Huawei, Leuven, Belgium
bangcvvn@gmail.com
[2] 3nLab@MCO Institute & ULB
dpapadimitriou@mco-inst.be

Abstract

The (structured) saddle-point problem involving the infimal convolution in real Hilbert spaces finds applicability in many applied mathematics disciplines. For this purpose, we develop a stochastic primal-dual splitting (PDS) algorithm with loopless variance-reduction (VR) for solving this generic problem. A PDS algorithm aims to overcome the well-known shortcomings of common splitting methods by solving the primal-dual pair formed by the monotone inclusion and its dual (in the sense of Fenchel–Rockafellar) reformulated as a monotone inclusion problem in a corresponding product space. The stochastic nature of the algorithm prevents from requiring the evaluation of the full gradient at each iteration, operation which can be computationally intensive when realized over the full dataset. The motivation behind variance reduction techniques finds its root in the convergence profile of stochastic first-order methods with fixed step size. From the perspective of the memory vs. computation time tradeoff, loopless VR offers several advantages, in particular in terms of computational time, compared to alternatives such as double-loop VR. In this respect, we first prove the weak almost sure convergence of the iterates; then, demonstrate that our algorithm achieves linear convergence in expectation of its iterates as well as convergence of the (smoothed and duality) gap function value.

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 structured saddle point problem in real Hilbert spaces.

Problem 1.1 Let $\mathcal{H}$, $\mathcal{G}$ be separable real Hilbert spaces. Let $L\colon \mathcal{H} \to \mathcal{G}$ be a bounded linear operator. Let $f\colon \mathcal{H} \to \left]-\infty, +\infty\right]$ and $g\colon \mathcal{G} \to \left]-\infty, +\infty\right]$ be proper lower semicontinuous convex functions. Let n_p and n_d be strictly positive integers. Let $(\mu_i)_{1\leq i\leq n_p}$ and $(\nu_i)_{1\leq i\leq n_d}$ be non-negative sequences. Let $(h_i)_{1\leq i\leq n_p}$ be a sequence of convex differentiable functions from $\mathcal{H}$ to $\mathbb{R}$ such that ∇h_i is μ_i-Lipschitz continuous. Let $(\ell_j)_{1\leq j\leq n_d}$ be a sequence of convex functions from $\mathcal{H}$ to $\mathbb{R}$ such that ℓ_j is $1/\nu_j$-strongly convex. Let $h\colon \mathcal{H}\to\mathbb{R}$ and $\ell\colon \mathcal{G}\to\mathbb{R}$ be convex differentiable functions defined respectively by

$$h := \frac{1}{n_p}\sum_{i=1}^{n_p} h_i \quad\text{and}\quad \ell^\star := \frac{1}{n_d}\sum_{i=1}^{n_d}\ell_i^\star \tag{1.1}$$

The primal problem is to

$$\underset{x\in\mathcal{H}}{\text{minimize}}\; h(x) + (\ell\,\square\, g)(Lx) + f(x), \tag{1.2}$$

where $\ell\,\square\, g$ denotes the infimal convolution of the functions ℓ and g. The dual problem (in the sense of Fenchel-Rockafellar) is to

$$\underset{v\in\mathcal{G}}{\text{minimize}}\; (h+f)^\star(L^\star v) + g^\star(-v) + \ell^\star(-v), \tag{1.3}$$

where $f^\star$ and $L^\star$ denote the Fenchel conjugate of the function f and operator L, respectively.

Stochastic numerical methods for solving saddle points problems have been extensively investigated in the literature, see [2, 3, 8, 9, 18, 20, 21] and [12, 14, 15, 28, 29, 34] for more recent developments. In these papers, the proposed methods find applicability to various problems arising from machine learning, statistical learning, transport optimization, portfolio optimization, eigenvalue optimization as well as many another problems in applied mathematics. Over the last decade, many of these stochastic methods have also exploited the variance reduction (class of) techniques in order to increase the precision of the gradient estimates while decreasing the computation time to obtain them; see for instances [2, 3, 8, 12, 14, 15, 28, 29, 34] and references therein. In this context, Problem 1.1 was first investigated in [9] and then in [30, 4, 5, 13] for the case where $n_p = n_d = 1$. In the case where $n_p+n_d > 2$, the problem has been recently resolved in [28, 21, 22] by means of stochastic variants of primal-dual splitting methods. Let us emphasize that when n_p and n_d are (very) large, the evaluation of the full gradient of h and ℓ becomes prohibitive. In turn, stochastic primal-dual splitting methods are often used as alternative to their deterministic counterpart. Comparatively,

(i) The algorithm in [28] can be viewed as a stochastic extension of [10] by using the Bregman distance. The main advantage of this work is that Hilbert spaces are relaxed to reflexive Banach spaces. Although enabling interesting applications such as the linear inverse problems on the simplex, the condition on the variables is much stronger than expected; moreover, the method does not exploit any variance reduction technique.

(ii) A stochastic method is developed in [21] for solving the Problem 1.1 with Bregman distance. The method exploits the variance reduction technique of [32] in finite dimensional Banach space. However, it reaches only sublinear convergence in expectation of the primal-dual gap (under mild conditions) whereas linear convergence rate is obtained under constraining conditions as the strong convexity relative to Bregman functions.

(iii) The method in [22] continues on the one developed in [21] by partially relaxing the fixed setting of the extrapolation parameters, and exploiting the double-loop variance reduction technique of [32] but still restricted to the usual duality gap function.

This work is motivated by the recent development in [16] of the loopless variance reduction method as well as [1]. In contrast, the methods developed in [21, 22] rely on double-loop variance reduction algorithms following the technique proposed in [32]. In the latter, at the beginning of the outer loop, the full gradient of the smooth functions needs to be computed and then used to build the stochastic gradient. Instead, this work aims at developing a stochastic primal-dual splitting algorithm for Problem 1.1 that relies on loopless variance reduction. In this paper, we also quantify the convergence speed for this class of saddle point problems by means of a generalization of the duality gap (to unbounded domains) based on the smoothing of nonsmooth functions [11] [13]. This gap referred to as smoothed gap takes finite values even for constrained problems, unlike the duality gap. Moreover, if the smoothness parameter is small and the smoothed gap is small, this implies that both the optimality gap and feasibility error are small too.

2 Preliminaries

Notations. The inner product and norm of all Hilbert spaces are denoted by $\langle \cdot \mid \cdot \rangle$ and $\| \cdot \|$. The conjugate of the linear operator L is denoted by $L^\star$. The effective domain of a function $f\colon \mathcal{H} \to \left]-\infty, +\infty\right]$ is $\operatorname{dom}(f) = \big\{x \in \mathcal{H} \mid f(x) < +\infty\big\}$. This function is proper if $\operatorname{dom}(f) \neq \varnothing$. We denote by $\Gamma_0(\mathcal{H})$ the class of all proper lower semicontinuous convex functions f from $\mathcal{H}$ to $\left]-\infty, +\infty\right]$. For $f \in \Gamma_0(\mathcal{H})$, the conjugate (or Fenchel conjugate) of the function f is denoted by $f^\star$. We also use ∂f to refer to the subdifferential of f. Finally, the infimal convolution of two functions f and g from $\mathcal{H}$ to $\left]-\infty, +\infty\right]$ writes as $f \,\square\, g$.

Assumptions. As in [9], throughout this paper, we assume that the set S is defined by

$$S = \big\{x \in \mathcal{H} \mid 0 \in \partial f(x) + \nabla h(x) + \big(L^\star \circ (\partial \ell \,\square\, \partial g) \circ L\big)(x)\big\} \neq \emptyset, \tag{2.1}$$

where

$$\partial f\colon \mathcal{H} \to 2^{\mathcal{H}}\colon x \mapsto \big\{u \in \mathcal{H} \mid (\forall y \in \mathcal{H}) \;\; \langle y - x \mid u\rangle + f(x) \leq f(y)\big\}, \tag{2.2}$$

and[1]

$$\partial \ell \,\square\, \partial g = (\partial \ell^\star + \partial g^\star)^{-1}. \tag{2.3}$$

As demonstrated in [9], under some qualification conditions, the primal problem can be reduced to find a point in S. We denote by

$$\boldsymbol{M}\colon (x, v) \mapsto \partial f(x) \times \partial g^\star(v) \text{ and } \boldsymbol{C}\colon (x, v) \mapsto (\nabla h(x) + L^\star v) \times (\nabla \ell^\star(v) - Lx), \tag{2.4}$$

where the (Fenchel) conjugate of the function g is defined by $g^\star\colon a \mapsto \sup\limits_{x \in \mathcal{H}} \big(\langle a \mid x\rangle - g(x)\big)$. Then, under the condition (2.1), the problem is equivalent to

$$\mathcal{S} = \big\{(x, v) \in \mathcal{H} \times \mathcal{G} \mid 0 \in (\boldsymbol{M} + \boldsymbol{C})(x, v)\big\} \neq \emptyset. \tag{2.5}$$

Definitions. We recall the definition and properties of the smoothed gap as introduced in [13].

[1] The infimal convolution of ℓ and g from $\mathcal{H}$ to $\left]-\infty, +\infty\right]$ is defined by $\ell \,\square\, g\colon x \mapsto \inf_{y \in \mathcal{H}}(\ell(y) + g(x - y))$

Definition 2.1 *Let $\beta \in [0,+\infty[$ and $(\tau,\sigma) \in]0,+\infty[^2$ define the smoothness parameters β/τ and β/σ, respectively*[2]*. Let $\mathsf{x} = (x,v)$ and $\dot{\mathsf{x}} = (\dot{x},\dot{v})$ be in $\mathcal{H} \times \mathcal{G}$ (where $\times$ denotes the Cartesian product). The smoothed gap $G_\beta(\mathsf{x};\dot{\mathsf{x}})$ centered at $\dot{\mathsf{x}}$ is defined by*

$$G_\beta(\mathsf{x};\dot{\mathsf{x}}) := \sup_{x' \in \mathcal{H}, v' \in \mathcal{G}} \left(K(x,v') - K(x',v) - \frac{\beta}{2\tau}\|x' - \dot{x}\|^2 - \frac{\beta}{2\sigma}\|v' - \dot{v}\|^2 \right), \tag{2.6}$$

where the Lagrangian function $K(x,v)$ is given by

$$K(x,v) = h(x) + f(x) + \langle Lx \mid v \rangle - g^\star(v) - \ell^\star(v). \tag{2.7}$$

In [13, Proposition 8], authors demonstrate that if $\mathsf{x} = \mathsf{x}^\dagger$ belongs to the primal-dual space $\mathcal{H}^\dagger \times \mathcal{G}^\dagger$ such that $0 \in \partial K(\mathsf{x}^\dagger)$, i.e., $\mathsf{x}^\dagger$ is a saddle point of the Lagrangian function K; then, the smoothed gap $G_\beta(\mathsf{x};\mathsf{x}^\dagger)$ is a measure of optimality, i.e., $G_\beta(\mathsf{x};\mathsf{x}^\dagger) = 0$. Observe that setting $\beta = 0$ yields the usual duality gap $G_{\beta=0}(\mathsf{x};\mathsf{x}^\dagger)$. The following Lemma recalls this result to our setting.

Lemma 2.2 *Let $\beta \in [0,+\infty[$ and $(\tau,\sigma) \in]0,+\infty[^2$. Let $\mathsf{x}^\dagger = (x^\dagger, v^\dagger) \in \mathcal{S}$ and $\mathsf{x} = (x,v) \in \mathcal{H} \times \mathcal{G}$. Then,*

$$G_\beta(\mathsf{x};\mathsf{x}^\dagger) = 0 \ \text{ if and only if } \ \mathsf{x} \in \mathcal{S}. \tag{2.8}$$

Moreover, define

$$x_\beta(x) := \operatorname{prox}_{\tau(f+h)/\beta}(x^\dagger - \tau L^\star v/\beta) \ \text{ and } \ v_\beta(x) := \operatorname{prox}_{\sigma(g^\star+\ell^\star)/\beta}(v^\dagger + \sigma Lx/\beta), \tag{2.9}$$

where the proximity operator of f is defined by $\operatorname{prox}_f : \mathcal{H} \to \mathcal{H} : x \mapsto \underset{y \in \mathcal{H}}{\operatorname{argmin}}\big(f(y) + \frac{1}{2}\|x - y\|^2\big)$. Then, the following holds,

$$G_\beta(\mathsf{x};\mathsf{x}^\dagger) \geq K(x,v^\dagger) - K(x^\dagger,v) + \frac{\beta}{\sigma}\|v_\beta(x) - v^\dagger\|^2 + \frac{\beta}{\tau}\|x_\beta(v) - x^\dagger\|^2. \tag{2.10}$$

Definition 2.3 [14, Section D.] *Variance reduction (VR): method used to increase the precision of the (gradient) estimates and the speed to obtain them. Formally, assume $\hat{h}_k$ is an estimate of the gradient $\nabla h(x_k)$. A method which verifies the property $\mathsf{E}[\|\hat{h}_k - \nabla h(x_k)\|^2] \xrightarrow[k\to\infty]{} 0$ is referred to as a VR method.*

Note that although stricto sensu a VR method does not require $\hat{h}_k$ being an unbiased estimate of the gradient $\nabla h(x_k)$, the proposed algorithm relies on this property (see Lemma 3.3).

Let $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ be a probability space where $\Omega_1 = \{1,\ldots,n_P\}$, $\mathcal{F}_1 = 2^{\Omega_1}$, and $\mathbb{P}_1 = \{p_1, p_2, \ldots, p_{n_p}\}$ with uniformly selected random index $p_i = 1/n_P \in]0,1]$. Let $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$ be a probability space where $\Omega_2 = \{1,\ldots,n_D\}$, $\mathcal{F}_2 = 2^{\Omega_2}$, and $\mathbb{P}_2 = \{q_1, q_2, \ldots, q_{n_d}\}$ with $q_j = 1/n_D \in]0,1]$. 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 $\boldsymbol{\sigma}$-algebra. The expectation of a random variable X is denoted by $\mathsf{E}[X]$. The conditional expectation of X given a $\boldsymbol{\sigma}$-field $\mathcal{A} \subset \mathcal{F}$ is denoted by $\mathsf{E}[X|\mathcal{A}]$. See [19] for more details on probability Theory in Hilbert spaces. The abbreviation a.s. stands for ”almost surely”.

[2] Compared to [13], we set $\beta_x = \beta = \beta_y$ with $\beta \in [0,+\infty[$ instead of $\beta \in [0,+\infty]$

Lemma 2.4 ([25, Theorem 1]) *Let $(\mathcal{F}_n)_{n\in\mathbb{N}}$ be an increasing sequence of sub-$\boldsymbol{\sigma}$-algebras of the $\boldsymbol{\sigma}$-algebra $\mathcal{F}$. 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..

3 Algorithm and Convergence properties

3.1 Algorithm

In this section, we detail our algorithm to solve problem (1.2) where we use the stochastic estimation of the full-gradient incorporating auxiliary variables with priority updating probabilities. Hence, the Algorithm 3.1 does not involve the full gradients $\nabla h(y_k)$ and $\nabla \ell(u_k)$. Nonetheless, in our convergence analysis (see Section 3.2), we also use the full gradients defined by

$$\hat{x}_{k+1} = \operatorname{prox}_{\tau_k f}(x_k - \tau_k \nabla h(y_k) - \tau_k L^\star u_k) \text{ and } \hat{v}_{k+1} = \operatorname{prox}_{\sigma_k g^\star}(v_k - \sigma_k \nabla \ell(u_k) + \sigma_k L y_k). \quad (3.1)$$

Algorithm 3.1 Let $(\tau_k, \sigma_k)_{k\in\mathbb{N}}$ be sequences in $]0,+\infty[^2$. Let $(x_0, x_{-1}) \in \mathcal{H}^2$ and $(v_0, v_{-1}) \in \mathcal{G}^2$. Let $w_{1,0} = w_{1,-1} = x_0$ and $w_{2,0} = w_{2,-1} = v_0$. Let (p,q) be in $]0,1]^2$.
Step 1. Compute

$$\begin{cases} y_k &= 2x_k - x_{k-1} \\ u_k &= 2v_k - v_{k-1} \end{cases} \quad (3.2)$$

Step 2. Pick $i_k \in \{1,\ldots,n_P\}$ and $j_k \in \{1,\ldots,n_D\}$ uniformly at random, and compute

$$\begin{cases} z_k &= -\nabla h_{i_k}(w_{1,k}) + \nabla h_{i_k}(y_k) + \nabla h(w_{1,k}) \\ d_k &= -\nabla \ell^\star_{j_k}(w_{2,k}) + \nabla \ell^\star_{j_k}(u_k) + \nabla \ell^\star(w_{2,k}) \end{cases} \quad (3.3)$$

where

$$w_{1,k+1} = \begin{cases} y_{k+1} \text{ with probability } p \\ w_{1,k} \text{ with probability } 1-p, \end{cases} \text{ and } w_{2,k+1} = \begin{cases} u_{k+1} \text{ with probability } q \\ w_{2,k} \text{ with probability } 1-q, \end{cases} \quad (3.4)$$

Step 3. Update

$$\begin{cases} x_{k+1} &= \operatorname{prox}_{\tau_k f}(x_k - \tau_k z_k - \tau_k L^\star u_k) \\ v_{k+1} &= \operatorname{prox}_{\sigma_k g^\star}(v_k - \sigma_k d_k + \sigma_k L y_k). \end{cases}$$

Remark 3.2 Comparison against related algorithms.

(i) The extrapolation Step 1 of Algorithm 3.1 was introduced in [17] for solving the classical variational inequality problem over a closed convex set in $\mathcal{H}$. Then, it was extended by [6] to solve a monotone inclusion. A stochastic development of [6] has been recently obtained in [23].

(ii) The idea of using the auxiliary variables $w_{1,k}$ and $w_{2,k}$ (as part of Step 2) was presented in [16] with the purpose of finding a minimizer of a single function h, where the extrapolation Step is not used (i.e. $y_k = x_k$). This idea was further developed in [1] for the method introduced in [17]. Algorithm 3.1 can be viewed as combining the auxiliary variables as proposed in [16] with the method developed in [6]. In particular, if $n_P = n_D = 1$, then we obtain the method in [6] for finding a point in $\mathcal{S}$, see (2.5).

(iii) The main differences of Algorithm 3.1 compared to recently published works [21, 23] consists of i) the appearance of auxiliary variables with priority updating probabilities (p, q) and ii) the loopless variance reduction step compared to double-loop variance reduction structure where the outer loop is replaced by a probabilistic switch between two types of updates: with probability (p, q) a full/stochastic gradient computation is performed on the primal/dual, while with probability $1 - p/1 - q$ the previous gradient is reused with an adjustment.

We first demonstrate that, for all $k \in \mathbb{N}$, the random variables z_k and d_k as defined by this algorithm are unbiased estimators of $\nabla h(y_k)$ and $\nabla \ell(u_k)$, and their variances are reduced progressively along with the convergence of the full-gradient. More precisely, we have the following.

Lemma 3.3 *Let E_k be the conditional expectations with respect to the history $\{y_k, w_{1,k-1}, u_k, w_{2,k-1}\}$. Then, $(\forall k \in \mathbb{N})$ z_k and d_k are unbiased estimators of $\nabla h(y_k)$ and $\nabla \ell^\star(u_k)$, respectively, i.e., we have*

$$(\forall k \in \mathbb{N})\ \mathsf{E}_k\left[z_k\right] = \nabla h(y_k) \quad \textit{and} \quad \mathsf{E}_k\left[d_k\right] = \nabla \ell^\star(u_k). \tag{3.5}$$

Moreover, let $\mathsf{x}^\dagger = (x^\dagger, v^\dagger) \in \mathcal{S}$ and define

$$\Xi_h(w_{1,k}, x^\dagger) := \frac{1}{n_p} \sum_{i=1}^{n_p} \|\nabla h_i(w_{1,k}) - \nabla h_i(x^\dagger)\|^2, \tag{3.6}$$

$$\Xi_{\ell^\star}(w_{2,k}, v^\dagger) := \frac{1}{n_q} \sum_{j=1}^{n_q} \|\nabla \ell_j^\star(w_{2,k}) - \nabla \ell_j^\star(v^\dagger)\|^2. \tag{3.7}$$

Then, we have

$$\begin{cases} \mathsf{E}_k[\|z_k - \nabla h(y_k)\|^2] & \leq 2(1-p)\big(\Xi_h(w_{1,k-1}, x^\star) + \Xi_h(y_k, x^\star)\big) \\ \mathsf{E}_k[\|d_k - \nabla \ell^\star(u_k)\|^2] & \leq 2(1-q)\big(\Xi_{\ell^\star}(w_{2,k-1}, v^\star) + \Xi_{\ell^\star}(u_k, v^\star)\big). \end{cases} \tag{3.8}$$

Proof. The unbiased estimation in (3.5) follows directly from the fact that $(\forall x \in \mathcal{H})$ $\mathsf{E}_k[\nabla h_{i_k}(x)] = \nabla h(x)$ and $(\forall v \in \mathcal{G})$ $\mathsf{E}_k[\nabla \ell_{j_k}^\star(v)] = \nabla \ell^\star(v)$. Let us prove (3.8). From (3.2), by substracting $\nabla h(y_k)$ on both left- and right-hand sides, we have

$$(\forall k \in \mathbb{N})\ \|z_k - \nabla h(y_k)\|^2 = \|\nabla h(w_{1,k}) - \nabla h_{i_k}(w_{1,k}) + \nabla h_{i_k}(y_k) - \nabla h(y_k)\|^2. \tag{3.9}$$

This equality implies that the variance of the variable z_k computed over i_k samples is bounded by

$$\mathsf{E}_{i_k}\left[\|z_k - \nabla h(y_k)\|^2\right] \leq \mathsf{E}_{i_k}\left[\|\nabla h_{i_k}(w_{1,k}) - \nabla h_{i_k}(y_k)\|^2\right]. \tag{3.10}$$

Hence, since for any $k \in \mathbb{N}$ for which $w_{1,k} = y_k$ with probability p, $w_{1,k} = w_{1,k-1}$ with probability $(1-p)$ and $\|x-y\|^2 \leq 2\,\|x-z\|^2 + 2\,\|z-y\|^2$, the left-hand side of this inequality verifies

$$\begin{aligned}
\mathsf{E}_{i_k}\left[\|z_k - \nabla h(y_k)\|^2\right] &= (1-p)\mathsf{E}_{i_k}\left[\|\nabla h_{i_k}(w_{1,k-1}) - \nabla h_{i_k}(y_k)\|^2\right] \\
&\leq 2(1-p)\bigg(\mathsf{E}_{i_k}\left[\|\nabla h_{i_k}(w_{1,k-1}) - \nabla h_{i_k}(x^\dagger)\|^2\right] + \mathsf{E}_{i_k}\left[\|\nabla h_{i_k}(y_k) - \nabla h_{i_k}(x^\dagger)\|^2\right]\bigg) \\
&\leq 2\frac{(1-p)}{n_p}\left(\sum_{i=1}^{n_p} \|\nabla h_i(w_{1,k-1}) - \nabla h_i(x^\dagger)\|^2 + \|\nabla h_i(y_k) - \nabla h_i(x^\dagger)\|^2\right) \\
&= 2(1-p)\big(\Xi_h(w_{1,k-1}, x^\dagger) + \Xi_h(y_k, x^\dagger)\big). \qquad (3.11)
\end{aligned}$$

From (3.2), by substracting $\nabla\ell^\star(u_k)$ on both left- and right-hand sides, we also have,

$$(\forall k \in \mathbb{N})\ \|d_k - \nabla\ell^\star(u_k)\|^2 = \left\|\nabla\ell^\star(w_{2,k}) - \nabla\ell^\star_{j_k}(w_{2,k}) + \nabla\ell^\star_{j_k}(u_k) - \nabla\ell^\star(u_k)\right\|^2 \qquad (3.12)$$

This equality implies that the variance of the variable d_k computed over j_k samples is bounded by

$$\mathsf{E}_{j_k}\left[\|d_k - \nabla\ell^\star(u_k)\|^2\right] \leq \mathsf{E}_{j_k}\left[\|\nabla\ell^\star_{j_k}(w_{2,k}) - \nabla\ell^\star_{j_k}(u_k)\|^2\right] \qquad (3.13)$$

Hence, drawing a similar reasoning as for the variance of the variable z_k, we obtain

$$\begin{aligned}
\mathsf{E}_{j_k}[\|d_k - \nabla\ell^\star(u_k)\|^2] &\leq 2\frac{(1-q)}{n_q}\left(\sum_{j=1}^{n_q}\left\|\nabla\ell^\star_j(w_{2,k-1}) - \nabla\ell^\star_j(v^\dagger)\right\|^2 + \left\|\nabla\ell^\star_j(u_k) - \nabla\ell^\star_j(v^\dagger)\right\|^2\right) \\
&= 2(1-q)\big(\Xi_{\ell^\star}(w_{2,k-1}, v^\dagger) + \Xi_{\ell^\star}(u_k, v^\dagger)\big), \qquad (3.14)
\end{aligned}$$

which completes the proof. □

The next Lemma provides an upper bound on the values of the gap function.

Lemma 3.4 *Define* $\mathbf{g} := g \,\square\, \ell$ *and* $\mathbf{f} := f + h$. *Set*

$$\boldsymbol{L} = \begin{pmatrix} 0 & -L^\star \\ L & 0 \end{pmatrix}, \quad \boldsymbol{U}_k = \begin{pmatrix} 1/\tau_k & 0 \\ 0 & 1/\sigma_k \end{pmatrix}, \quad \textit{and } \boldsymbol{D} = \operatorname{diag}(\mu, \nu). \qquad (3.15)$$

Set $\mathsf{x} = (x, v) \in \operatorname{dom}(f) \times \operatorname{dom}(g^\star)$. *Define*

$$(\forall k \in \mathbb{N}) \begin{cases} \mathsf{x}_k &= (x_k, v_k),\ \hat{\mathsf{x}}_k = (\hat{x}_k, \hat{v}_k),\ \mathsf{y}_k = (y_k, u_k), \\ \mathbf{r}_k &= (z_k, t_k), \\ \mathsf{R}_k &= (\nabla h(y_k), \nabla\ell(u_k)), \\ \boldsymbol{b}_k(\mathsf{x}) &= \langle \boldsymbol{L}(\mathsf{x}_k - \mathsf{x}_{k-1}) \mid \mathsf{x}_k - \mathsf{x} \rangle\,. \end{cases} \qquad (3.16)$$

Then,

$$\begin{aligned}
K(x_{k+1}, v) - K(x, v_{k+1}) \leq{}& \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{2\boldsymbol{D}} - \boldsymbol{b}_k(\mathsf{x}) \\
&- \Big(\frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{2\boldsymbol{D}} - \boldsymbol{b}_{k+1}(\mathsf{x})\Big) \\
&+ \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{4\boldsymbol{D} + \boldsymbol{L}^\intercal\boldsymbol{D}^{-1}\boldsymbol{L}} - \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\boldsymbol{U}_k} \\
&+ \|\mathbf{r}_k - \mathbf{R}_k\|^2_{\boldsymbol{U}_k^{-1}} + \langle \hat{\mathsf{x}}_{k+1} - \mathsf{x} \mid \mathbf{R}_k - \mathbf{r}_k \rangle\,, \qquad (3.17)
\end{aligned}$$

where $\boldsymbol{L}^\intercal$ *denotes the (conjugate) transpose of* $\boldsymbol{L}$, *i.e.*, $\boldsymbol{L}^\intercal = -\boldsymbol{L}$.

Proof. Let $k \in \mathbb{N}$. We have $v_{k+1} = (\mathrm{Id} + \sigma_k \partial g^\star)^{-1}(v_k - \sigma_k d_k + \sigma_k L y_k)$, which is equivalent to

$$Ly_k - d_k + \frac{1}{\sigma_k}(v_k - v_{k+1}) \in \partial g^\star(v_{k+1}).$$

Since $g^\star$ is a convex function, it follows that,

$$(\forall v \in \mathcal{G})\ g^\star(v) \geq g^\star(v_{k+1}) + \left\langle Ly_k - d_k + \frac{1}{\sigma_k}(v_k - v_{k+1}) \mid v - v_{k+1} \right\rangle,$$

which implies that

$$\begin{aligned} g^\star(v_{k+1}) - g^\star(v) &\leq \langle d_k - Ly_k \mid v - v_{k+1}\rangle + \frac{1}{\sigma_k}\langle v_k - v_{k+1} \mid v_{k+1} - v\rangle \\ &= \langle d_k - Ly_k \mid v - v_{k+1}\rangle + \frac{1}{2\sigma_k}\big(\|v - v_k\|^2 - \|v_{k+1} - v_k\|^2 - \|v - v_{k+1}\|^2\big). \end{aligned} \tag{3.18}$$

Since $\ell^\star$ is convex and continuously differentiable with ν-Lipschitz gradient, we have

$$\ell^\star(v_{k+1}) - \ell^\star(v) \leq \langle v_{k+1} - v \mid \nabla \ell^\star(u_k)\rangle + \frac{\nu}{2}\|v_{k+1} - u_k\|^2. \tag{3.19}$$

We derive from (3.18) and (3.19) that for every $x \in \mathcal{H}$,

$$\begin{aligned} K(x_{k+1}, v) - K(x_{k+1}, v_{k+1}) &= \langle Lx_{k+1} \mid v - v_{k+1}\rangle + \mathbf{g}^\star(v_{k+1}) - \mathbf{g}^\star(v) \\ &\leq \langle L(x_{k+1} - y_k) \mid v - v_{k+1}\rangle + \frac{1}{2\sigma_k}\big(\|v - v_k\|^2 - \|v_{k+1} - v_k\|^2 - \|v - v_{k+1}\|^2\big) \\ &\quad + \frac{\nu}{2}\|v_{k+1} - u_k\|^2 + \langle \nabla\ell^\star(u_k) - d_k \mid v_{k+1} - v\rangle. \end{aligned} \tag{3.20}$$

Similar to (3.20), we have, for every $x \in \mathcal{H}$,

$$\begin{aligned} K(x_{k+1}, v_{k+1}) - K(x, v_{k+1}) &= \langle L(x_{k+1} - x) \mid v_{k+1}\rangle + \mathbf{f}(x_{k+1}) - \mathbf{f}(x) \\ &\leq \langle L(x_{k+1} - x) \mid v_{k+1} - u_k\rangle + \frac{1}{2\tau_k}\big(\|x - x_k\|^2 - \|x_{k+1} - x_k\|^2 - \|x - x_{k+1}\|^2\big) \\ &\quad + \frac{\mu}{2}\|x_{k+1} - y_k\|^2 + \langle x_{k+1} - x \mid \nabla h(y_k) - z_k\rangle. \end{aligned} \tag{3.21}$$

Adding (3.20) and (3.21), we obtain

$$\begin{aligned} K(x_{k+1}, v) - K(x, v_{k+1}) &\leq \Big(\overbrace{\langle L(x_{k+1} - x) \mid v_{k+1} - u_k\rangle}^{\alpha_{1,k}} + \overbrace{\langle L(x_{k+1} - y_k) \mid v - v_{k+1}\rangle}^{\alpha_{2,k}}\Big) \\ &+ \underbrace{\frac{1}{2\tau_k}\big(\|x - x_k\|^2 - \|x_{k+1} - x_k\|^2 - \|x - x_{k+1}\|^2\big)}_{\alpha_{5,k}} + \underbrace{\frac{1}{2\sigma_k}\big(\|v - v_k\|^2 - \|v_{k+1} - v_k\|^2 - \|v - v_{k+1}\|^2\big)}_{\alpha_{6,k}} \\ &+ \underbrace{\frac{\mu}{2}\|x_{k+1} - y_k\|^2 + \frac{\nu}{2}\|v_{k+1} - u_k\|^2}_{\alpha_{0,k}} + \underbrace{\langle x_{k+1} - x \mid \nabla h(y_k) - z_k\rangle}_{\alpha_{3,k}} + \underbrace{\langle \nabla\ell^\star(u_k) - d_k \mid v_{k+1} - v\rangle}_{\alpha_{4,k}}. \end{aligned} \tag{3.22}$$

Using (3.1), i.e., $u_k = v_k + v_k - v_{k-1}$, the first term in the right hand side of (3.22) can be expressed as

$$\begin{aligned}\alpha_{1,k} &= \langle L(x_{k+1}-x) \mid v_{k+1} - v_k - v_k + v_{k-1}\rangle\\ &= \langle L(x_{k+1}-x) \mid v_{k+1} - v_k\rangle - \langle L(x_{k+1}-x) \mid v_k - v_{k-1}\rangle\\ &= \langle L(x_{k+1}-x) \mid v_{k+1} - v_k\rangle - \langle L(x_{k+1}-x_k) \mid v_k - v_{k-1}\rangle - \langle L(x_k-x) \mid v_k - v_{k-1}\rangle . \qquad (3.23)\end{aligned}$$

Similar to (3.23), for the second term of (3.22), by expanding the expression of y_k (see (3.1)), we also have

$$\alpha_{2,k} = \langle L(x_{k+1}-x_k) \mid v - v_{k+1}\rangle - \langle L(x_k - x_{k-1}) \mid v - v_k\rangle - \langle L(x_k-x_{k-1}) \mid v_k - v_{k+1}\rangle . \qquad (3.24)$$

Observe that

$$\begin{cases}\langle L(x_{k+1}-x_k) \mid v_k - v_{k-1}\rangle + \langle L(x_k - x_{k-1}) \mid v_k - v_{k+1}\rangle = \langle \mathsf{x}_k - \mathsf{x}_{k+1} \mid \boldsymbol{L}(\mathsf{x}_k - \mathsf{x}_{k-1})\rangle\\ \langle L(x_{k+1}-x) \mid v_{k+1} - v_k\rangle + \langle L(x_{k+1} - x_k) \mid v - v_{k+1}\rangle = \langle \mathsf{x}_{k+1} - \mathsf{x}_k \mid \boldsymbol{L}(\mathsf{x}_{k+1} - \mathsf{x})\rangle\\ \langle L(x_k-x) \mid v_k - v_{k-1}\rangle + \langle L(x_k - x_{k-1}) \mid v - v_k\rangle = \langle \mathsf{x}_k - \mathsf{x}_{k-1} \mid \boldsymbol{L}(\mathsf{x}_k - \mathsf{x})\rangle .\end{cases} \qquad (3.25)$$

Hence, we can derive from (3.25), (3.24) and (3.23) that

$$\alpha_{1,k} + \alpha_{2,k} = \boldsymbol{b}_{k+1}(\mathsf{x}) - \boldsymbol{b}_k(\mathsf{x}) - \langle \mathsf{x}_k - \mathsf{x}_{k+1} \mid \boldsymbol{L}(\mathsf{x}_k - \mathsf{x}_{k-1})\rangle . \qquad (3.26)$$

Next we estimate $\alpha_{3,k}$ and $\alpha_{4,k}$. Using the non-expansiveness property of $\operatorname{prox}_{\tau_k f}$, we have

$$\begin{aligned}\|\hat{x}_{k+1} - x_{k+1}\| &= \left\|\operatorname{prox}_{\tau_k f}(x_k - \tau_k \nabla h(y_k) - \tau_k L^\star u_k) - \operatorname{prox}_{\tau_k f}(x_k - \tau_k z_k - \tau_k L^\star u_k)\right\|\\ &\le \tau_k \left\|z_k - \nabla h(y_k)\right\| . \qquad (3.27)\end{aligned}$$

In turn,

$$\begin{aligned}\alpha_{3,k} &= \langle x_{k+1} - \hat{x}_{k+1} \mid \nabla h(y_k) - z_k\rangle + \langle \hat{x}_{k+1} - x \mid \nabla h(y_k) - z_k\rangle\\ &\le \|z_k - \nabla h(y_k)\|\|x_{k+1} - \hat{x}_{k+1}\| + \langle \hat{x}_{k+1} - x \mid \nabla h(y_k) - z_k\rangle\\ &\le \tau_k \|z_k - \nabla h(y_k)\|^2 + \langle \hat{x}_{k+1} - x \mid \nabla h(y_k) - z_k\rangle . \qquad (3.28)\end{aligned}$$

In the same way, we also have

$$\alpha_{4,k} \le \sigma_k \|d_k - \nabla \ell(u_k)\|^2 + \langle \nabla \ell(u_k) - d_k \mid \hat{v}_{k+1} - v\rangle . \qquad (3.29)$$

Adding (3.29) and (3.28), we obtain

$$\alpha_{3,k} + \alpha_{4,k} \le \|\mathbf{r}_k - \mathbf{R}_k\|^2_{\boldsymbol{U}_k^{-1}} + \langle \hat{\mathsf{x}}_{k+1} - \mathsf{x} \mid \mathbf{R}_k - \mathbf{r}_k\rangle . \qquad (3.30)$$

In order to estimate $\alpha_{0,k}$, we deduce by expanding the expression of y_k

$$\begin{aligned}\frac{\mu}{2}\|x_{k+1} - y_k\|^2 &= \frac{\mu}{2}\|x_{k+1} - x_k - (x_k - x_{k-1})\|^2\\ &\le \frac{\mu}{2}\|x_{k+1} - x_k\|^2 + \frac{\mu}{2}\|x_k - x_{k-1}\|^2 - \mu \langle x_{k+1} - x_k \mid x_k - x_{k-1}\rangle , \qquad (3.31)\end{aligned}$$

and

$$\begin{aligned}\frac{\nu}{2}\|v_{k+1}-u_k\|^2 &= \frac{\nu}{2}\|v_{k+1}-v_k-(v_k-v_{k-1})\|^2\\ &\le \frac{\nu}{2}\|v_{k+1}-v_k\|^2+\frac{\nu}{2}\|v_k-v_{k-1}\|^2-\nu\left\langle v_{k+1}-v_k \mid v_k-v_{k-1}\right\rangle .\end{aligned} \tag{3.32}$$

Adding (3.31) and (3.32), we obtain, since $\boldsymbol{D}=\mathrm{diag}(\mu,\nu)$, the following expression for $\alpha_{0,k}$

$$\alpha_{0,k}=\frac{1}{2}\|\mathsf{x}_{k+1}-\mathsf{x}_k\|^2_{\boldsymbol{D}}+\frac{1}{2}\|\mathsf{x}_k-\mathsf{x}_{k-1}\|^2_{\boldsymbol{D}}-\left\langle \mathsf{x}_{k+1}-\mathsf{x}_k \mid \boldsymbol{D}(\mathsf{x}_k-\mathsf{x}_{k-1})\right\rangle . \tag{3.33}$$

Therefore, adding (3.33) and (3.26), we get

$$\begin{aligned}\alpha_{0,k}+\alpha_{1,k}+\alpha_{2,k} &= \frac{1}{2}\|\mathsf{x}_{k+1}-\mathsf{x}_k\|^2_{\boldsymbol{D}}+\frac{1}{2}\|\mathsf{x}_k-\mathsf{x}_{k-1}\|^2_{\boldsymbol{D}}-\left\langle \mathsf{x}_{k+1}-\mathsf{x}_k \mid (\boldsymbol{D}-\boldsymbol{L})(\mathsf{x}_k-\mathsf{x}_{k-1})\right\rangle\\ &\quad+\boldsymbol{b}_{k+1}(\mathsf{x})-\boldsymbol{b}_k(\mathsf{x}).\end{aligned} \tag{3.34}$$

We have

$$\begin{aligned}\left\langle \mathsf{x}_{k+1}-\mathsf{x}_k \mid (\boldsymbol{D}-\boldsymbol{L})(\mathsf{x}_k-\mathsf{x}_{k-1})\right\rangle &= \left\langle \boldsymbol{D}^{-1}(\boldsymbol{D}-\boldsymbol{L})^{\intercal}(\mathsf{x}_{k+1}-\mathsf{x}_k) \mid \mathsf{x}_k-\mathsf{x}_{k-1}\right\rangle_{\boldsymbol{D}}\\ &\le \frac{1}{2}\|\mathsf{x}_k-\mathsf{x}_{k-1}\|^2_{\boldsymbol{D}}+\frac{1}{2}\|\boldsymbol{D}^{-1}(\boldsymbol{D}-\boldsymbol{L})^{\intercal}(\mathsf{x}_{k+1}-\mathsf{x}_k)\|^2_{\boldsymbol{D}}.\end{aligned} \tag{3.35}$$

Hence, (3.34) becomes

$$\alpha_{0,k}+\alpha_{1,k}+\alpha_{2,k}=\frac{1}{2}\|\mathsf{x}_{k+1}-\mathsf{x}_k\|^2_{2\boldsymbol{D}+\boldsymbol{L}^{\intercal}\boldsymbol{D}^{-1}\boldsymbol{L}}+\frac{1}{2}\|\mathsf{x}_k-\mathsf{x}_{k-1}\|^2_{2\boldsymbol{D}}+\boldsymbol{b}_{k+1}(\mathsf{x})-\boldsymbol{b}_k(\mathsf{x}). \tag{3.36}$$

We have next, by using the definition of $\boldsymbol{U}_k$, we can rewrite the sum $\alpha_{5,k}$ and $\alpha_{6,k}$ as

$$\alpha_{5,k}+\alpha_{6,k}=\frac{1}{2}\|\mathsf{x}_k-\mathsf{x}\|^2_{\boldsymbol{U}_k}-\frac{1}{2}\|\mathsf{x}_{k+1}-\mathsf{x}\|^2_{\boldsymbol{U}_k}-\frac{1}{2}\|\mathsf{x}_k-\mathsf{x}_{k+1}\|^2_{\boldsymbol{U}_k}. \tag{3.37}$$

Therefore, by combining (3.36), (3.37), (3.30) into (3.22), we obtain

$$\begin{aligned}K(x_{k+1},v)-K(x,v_{k+1}) &\le \sum_{i=0}^{6}\alpha_{i,k}\\ &\le \frac{1}{2}\|\mathsf{x}_k-\mathsf{x}\|^2_{\boldsymbol{U}_k}+\frac{1}{2}\|\mathsf{x}_k-\mathsf{x}_{k-1}\|^2_{2\boldsymbol{D}}-\boldsymbol{b}_k(\mathsf{x})\\ &\quad-\left(\frac{1}{2}\|\mathsf{x}_{k+1}-\mathsf{x}\|^2_{\boldsymbol{U}_k}+\frac{1}{2}\|\mathsf{x}_{k+1}-\mathsf{x}_k\|^2_{2\boldsymbol{D}}-\boldsymbol{b}_{k+1}(\mathsf{x})\right)\\ &\quad+\frac{1}{2}\|\mathsf{x}_{k+1}-\mathsf{x}_k\|^2_{4\boldsymbol{D}+\boldsymbol{L}^{\intercal}\boldsymbol{D}^{-1}\boldsymbol{L}}-\frac{1}{2}\|\mathsf{x}_{k+1}-\mathsf{x}_k\|^2_{\boldsymbol{U}_k}\\ &\quad+\|\mathbf{r}_k-\mathbf{R}_k\|^2_{\boldsymbol{U}_k^{-1}}+\left\langle \hat{\mathsf{x}}_{k+1}-\mathsf{x} \mid \mathbf{R}_k-\mathbf{r}_k\right\rangle .\end{aligned} \tag{3.38}$$

Hence, the proof is completed. □

Remark 3.5 Suppose that f and $g^\star$ are strongly convex functions with constants θ_1 and θ_2, respectively. Then, under the same notations as Lemma 3.4, we have

$$
\begin{aligned}
K(x_{k+1}, v) - K(x, v_{k+1}) + &\frac{\min\{\theta_1\tau_k, \theta_2\sigma_k\}}{2}\|\mathsf{x}_{k+1} - \mathsf{x}\|^2_{\boldsymbol{U}_k} \\
&\leq \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{2\boldsymbol{D}} - \boldsymbol{b}_k(\mathsf{x}) \\
&\quad - \left(\frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{2\boldsymbol{D}} - \boldsymbol{b}_{k+1}(\mathsf{x})\right) \\
&\quad + \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{4\boldsymbol{D}+\boldsymbol{L}^\intercal\boldsymbol{D}^{-1}\boldsymbol{L}} - \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\boldsymbol{U}_k} \\
&\quad + \|\mathbf{r}_k - \mathbf{R}_k\|^2_{\boldsymbol{U}_k^{-1}} + \langle \hat{\mathsf{x}}_{k+1} - \mathsf{x} \mid \mathbf{R}_k - \mathbf{r}_k \rangle\,.
\end{aligned}
\tag{3.39}
$$

Moreover, if instead of (3.40), we use, with $\mu_0 = \|\boldsymbol{D} - \boldsymbol{L}\|$,

$$
\langle \mathsf{x}_{k+1} - \mathsf{x}_k \mid (\boldsymbol{D} - \boldsymbol{L})(\mathsf{x}_k - \mathsf{x}_{k-1}) \rangle \leq \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{\mu_0\,\mathrm{Id}} + \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\mu_0\,\mathrm{Id}}, \tag{3.40}
$$

then, instead of (3.39), we also have

$$
\begin{aligned}
K(x_{k+1}, v) - K(x, v_{k+1}) + &\frac{\min\{\theta_1\tau_k, \theta_2\sigma_k\}}{2}\|\mathsf{x}_{k+1} - \mathsf{x}\|^2_{\boldsymbol{U}_k} \\
&\leq \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}\|^2_{\boldsymbol{U}_k} - \boldsymbol{b}_k(\mathsf{x}) + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{\boldsymbol{D}+\mu_0\,\mathrm{Id}} \\
&\quad - \left(\frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\boldsymbol{D}+\mu_0\,\mathrm{Id}} - \boldsymbol{b}_{k+1}(\mathsf{x})\right) \\
&\quad + \|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\boldsymbol{D}+\mu_0\,\mathrm{Id}} - \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\boldsymbol{U}_k} \\
&\quad + \|\mathbf{r}_k - \mathbf{R}_k\|^2_{\boldsymbol{U}_k^{-1}} + \langle \hat{\mathsf{x}}_{k+1} - \mathsf{x} \mid \mathbf{R}_k - \mathbf{r}_k \rangle\,.
\end{aligned}
\tag{3.41}
$$

Remark 3.6 In the case where $p = 1 = q$, Lemma 3.3 recovers the one provided in [32], and Lemma 3.4 is similar to [21, Lemma 3.5] where μ_0 is replaced by $\|\boldsymbol{D}\| + \|\boldsymbol{L}\|$.

3.2 Convergence properties

In this section, we characterize the convergence properties of Algorithm 3.1. We start by studying its (weak) convergence profile in Section 3.2.1. Then, in Section 3.2.2, we develop the conditions and assumptions under which this algorithm converges linearly.

3.2.1 Weak convergence

The weak convergence of the iterate as well as the convergence of the gap function value to 0 rely on the following Theorem that establishes the descent property of a suitable Lyapunov function.

Theorem 3.7 *Let $\mathsf{x}^\dagger \in \mathcal{S}$ and define $P = \operatorname{diag}(1-p, 1-q)$. For every $k \in \mathbb{N}$, define*

$$\begin{cases} \mathsf{x}_k = (x_k, v_k),\ \mathsf{w}_k = (w_{1,k}, w_{2,k}), \\ \Theta(\mathsf{x}_k) := \Theta(x_k, v_k) = K(x_k, v^\dagger) - K(x^\dagger, v_k), \\ Q(\mathsf{w}_k) := Q(w_{k,1}, w_{k,2}) = \Xi_h(w_{1,k}, x^\star) + \Xi_{\ell^\star}(w_{2,k}, v^\star), \\ \gamma_k = \max\{\sigma_k, \tau_k\}. \end{cases} \tag{3.42}$$

We also use following operators:

$$\begin{cases} \boldsymbol{V}_k = 2\boldsymbol{D} + 4\boldsymbol{D}(\mathrm{Id} - P) + 8\boldsymbol{D}P\boldsymbol{U}_k^{-1}, \\ \Lambda_{k+1} = \boldsymbol{U}_k - 2\boldsymbol{D} - \boldsymbol{V}_{k+1}. \end{cases} \tag{3.43}$$

Moreover, define the following Lyapunov function

$$\mathcal{L}_k(\mathsf{x}^\dagger) = \Theta(\mathsf{x}_k) + Q(\mathsf{w}_{k-1}) + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_{k-1}} - \boldsymbol{b}_k(\mathsf{x}^\dagger) + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{\boldsymbol{V}_k}, \tag{3.44}$$

where $\boldsymbol{b}_k(\mathsf{x}^\dagger)$ is defined by (3.16). Set $\overline{\mu} = \max\{\mu, \nu\}$, $\underline{p} = \min\{p, q\}$ and $\epsilon \in \left]0, \underline{p}\right[$. Let $(\eta_k)_{k\in\mathbb{N}}$ be a sequence in $\ell^1_+(\mathbb{N})$. Suppose that the following conditions are verified.

$$\begin{cases} 4\overline{\mu}\big(2\gamma_k(1-\underline{p}) + q\big) + \epsilon \le 1 + \eta_k,\ (2\gamma_k + 1)(1-\underline{p}) + \epsilon \le 1 + \eta_k, \\ \boldsymbol{U}_{k-1} \succeq \boldsymbol{U}_k \succeq (\epsilon + \|L\|)\,\mathrm{Id}, \\ \boldsymbol{V}_k \succeq \|L\|\,\mathrm{Id}, \\ \Lambda_k \succeq \epsilon\,\mathrm{Id}\,. \end{cases} \tag{3.45}$$

Then, the following descent property is verified for all k:

$$\mathsf{E}_k\left[\mathcal{L}_{k+1}(\mathsf{x}^\dagger)\right] + \mathsf{E}_k\left[\frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\Lambda_{k+1}}\right] \le (1+\eta_k)\mathcal{L}_k(\mathsf{x}^\dagger) - \epsilon\big(\Theta(\mathsf{x}_k) + Q(\mathsf{w}_{k-1})\big). \tag{3.46}$$

Proof. Using the same notations as defined for Lemma 3.4 and the expression (3.17) with $\mathsf{x} = \mathsf{x}^\dagger$, we obtain

$$\begin{aligned} \Theta(\mathsf{x}_{k+1}) \le\ & \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{2\boldsymbol{D}} - \boldsymbol{b}_k(\mathsf{x}^\dagger) \\ & - \Big(\frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{2\boldsymbol{D}} - \boldsymbol{b}_{k+1}(\mathsf{x}^\dagger)\Big) \\ & + \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{4\boldsymbol{D} + \boldsymbol{L}^\intercal \boldsymbol{D}^{-1}\boldsymbol{L}} - \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\boldsymbol{U}_k} \\ & + \|\mathbf{r}_k - \mathbf{R}_k\|^2_{\boldsymbol{U}_k^{-1}} + \Big\langle \hat{\mathsf{x}}_{k+1} - \mathsf{x}^\dagger \mid \mathbf{R}_k - \mathbf{r}_k \Big\rangle\,. \end{aligned} \tag{3.47}$$

From Lemma 3.3, we deduce

$$\mathsf{E}_k\left[\Big\langle \hat{\mathsf{x}}_{k+1} - \mathsf{x}^\dagger \mid \mathbf{R}_k - \mathbf{r}_k \Big\rangle\right] = 0. \tag{3.48}$$

Using (3.8), we also have

$$\begin{aligned} \mathsf{E}_k\left[\|\mathbf{r}_k - \mathbf{R}_k\|^2_{\boldsymbol{U}_k^{-1}}\right] \le\ & 2\tau_k(1-p)\Big(\Xi_h(w_{1,k-1}, x^\star) + \Xi_h(y_k, x^\star)\Big) \\ & + 2\sigma_k(1-q)\Big(\Xi_{\ell^\star}(w_{2,k-1}, v^\star) + \Xi_{\ell^\star}(u_k, v^\star)\Big). \end{aligned} \tag{3.49}$$

By definition of Ξ_h and $\Xi_{\ell^\star}$, we derive the following inequalities

$$\begin{cases} \Xi_h(y_k, x^\star) & \leq 2\big(\Xi_h(y_k, x_k) + \Xi_h(x_k, x^\star)\big) \leq 2\big(\mu\|x_k - x_{k-1}\|^2 + \Xi_h(x_k, x^\star)\big) \\ \Xi_{\ell^\star}(u_k, v^\star) & \leq 2\big(\Xi_{\ell^\star}(u_k, v_k) + \Xi_{\ell^\star}(v_k, v^\star)\big) \leq 2\big(\nu\|v_k - v_{k-1}\|^2 + \Xi_{\ell^\star}(v_k, v^\star)\big). \end{cases} \tag{3.50}$$

We can further estimate (3.49) as

$$\begin{aligned} \mathsf{E}_k\left[\|\mathbf{r}_k - \mathbf{R}_k\|^2_{\boldsymbol{U}_k^{-1}}\right] &\leq 2\tau_k(1-p)\Big(\Xi_h(w_{1,k-1}, x^\star) + 2\mu\|x_k - x_{k-1}\|^2 + 2\Xi_h(x_k, x^\star)\Big) \\ &\quad + 2\sigma_k(1-q)\Big(\Xi_{\ell^\star}(w_{2,k-1}, v^\star) + 2\nu\|v_k - v_{k-1}\|^2 + 2\Xi_{\ell^*}(v_k, v^\star)\Big) \\ &\leq 2\gamma_k(1-\underline{p})\Big(\big(\Xi_h(w_{1,k-1}, x^\star) + \Xi_{\ell^\star}(w_{2,k-1}, v^\star)\big) + 2\big(\Xi_h(x_k, x^\star) + \Xi_{\ell^\star}(v_k, v^\star)\big)\Big) \\ &\quad + \|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{4\boldsymbol{D}P\boldsymbol{U}_k^{-1}} \\ &= 2\gamma_k(1-\underline{p})\Big(Q(\mathsf{w}_{k-1}) + 2\big(\Xi_h(x_k, x^\star) + \Xi_{\ell^\star}(v_k, v^\star)\big)\Big) \\ &\quad + \|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{4\boldsymbol{D}P\boldsymbol{U}_k^{-1}}. \end{aligned} \tag{3.51}$$

The second term in (3.51) are bound by the gap as indicated by Lemma 3.3 in [21],

$$\Xi_h(x_k, x^\star) + \Xi_{\ell^\star}(v_k, v^\star) \leq 2\overline{\mu}\Theta(\mathsf{x}_k). \tag{3.52}$$

Therefore,

$$\mathsf{E}_k\left[\|\mathbf{r}_k - \mathbf{R}_k\|^2_{\boldsymbol{U}_k^{-1}}\right] \leq 2\gamma_k(1-\underline{p})\big(Q(\mathsf{w}_{k-1}) + 4\overline{\mu}\Theta(\mathsf{x}_k)\big) + \|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{4\boldsymbol{D}P\boldsymbol{U}_k^{-1}}. \tag{3.53}$$

Now, by taking the expectation E_k on both sides of (3.47) and invoking (3.53), we obtain

$$\begin{aligned} \mathsf{E}_k\left[\Theta(\mathsf{x}_{k+1})\right] \leq\ & 8\gamma_k\overline{\mu}(1-\underline{p})\Theta(\mathsf{x}_k) + 2\gamma_k(1-\underline{p})Q(\mathsf{w}_{k-1}) \\ & + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{2\boldsymbol{D}} + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{8\boldsymbol{D}P\boldsymbol{U}_k^{-1}} - \boldsymbol{b}_k(\mathsf{x}^\dagger) \\ & - \mathsf{E}_k\left[\frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{2\boldsymbol{D}} + \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{8\boldsymbol{D}P\boldsymbol{U}_{k+1}^{-1}} - \boldsymbol{b}_{k+1}(\mathsf{x}^\dagger)\right] \\ & + \mathsf{E}_k\left[\frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{4\boldsymbol{D}+\boldsymbol{L}^\intercal\boldsymbol{D}^{-1}\boldsymbol{L}+8\boldsymbol{D}P\boldsymbol{U}_{k+1}^{-1}} - \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\boldsymbol{U}_k}\right]. \end{aligned} \tag{3.54}$$

Since following (3.42), $Q(\mathsf{w}_k) = \Xi_h(w_{1,k}, x^\star) + \Xi_{\ell^\star}(w_{2,k}, v^\star)$, its expectation $\mathsf{E}_k[Q(\mathsf{w}_k)]$ can be upper bounded by using inequalities (3.50) and (3.52)

$$\begin{aligned} \mathsf{E}_k\left[Q(\mathsf{w}_k)\right] &= \mathsf{E}_k\left[\Xi_h(w_{1,k}, x^\star) + \Xi_{\ell^\star}(w_{2,k}, v^\star)\right] \\ &= (1-p)\Xi_h(w_{1,k-1}, x^\star) + (1-q)\Xi_{\ell^\star}(w_{2,k-1}, v^\star) + p\Xi_h(y_k, x^\star) + q\Xi_{\ell^\star}(u_k, v^\star) \\ &\leq (1-\underline{p})Q(\mathsf{w}_{k-1}) + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{4\boldsymbol{D}(\mathrm{Id}-P)} + 4\overline{\mu}q\Theta(\mathsf{x}_k). \end{aligned} \tag{3.55}$$

Adding (3.54) to (3.55), and using the definition of $\boldsymbol{V}_k$ in (3.42), we obtain[3]

$$\begin{aligned}\mathsf{E}_k\left[\Theta(\mathsf{x}_{k+1})\right]+\mathsf{E}_k\left[Q(\mathsf{w}_k)\right]\leq{}&4\overline{\mu}\big(2\gamma_k(1-\underline{p})+q\big)\Theta(\mathsf{x}_k)+(2\gamma_k+1)(1-\underline{p})Q(\mathsf{w}_{k-1})\\&+\frac{1}{2}\|\mathsf{x}_k-\mathsf{x}\|^2_{\boldsymbol{U}_k}+\frac{1}{2}\|\mathsf{x}_k-\mathsf{x}_{k-1}\|^2_{\boldsymbol{V}_k}-\boldsymbol{b}_k(\mathsf{x}^\dagger)\\&-\mathsf{E}_k\left[\frac{1}{2}\|\mathsf{x}_{k+1}-\mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k}+\frac{1}{2}\|\mathsf{x}_{k+1}-\mathsf{x}_k\|^2_{\boldsymbol{V}_{k+1}}-\boldsymbol{b}_{k+1}(\mathsf{x}^\dagger)\right]\\&-\mathsf{E}_k\left[\frac{1}{2}\|\mathsf{x}_{k+1}-\mathsf{x}_k\|^2_{\Lambda_{k+1}}\right].\end{aligned}\tag{3.56}$$

Now using the definition of $\boldsymbol{b}_k(\mathsf{x})$, we have

$$\begin{aligned}\boldsymbol{b}_k(\mathsf{x}^\dagger)&=\Big\langle\boldsymbol{L}(\mathsf{x}_k-\mathsf{x}_{k-1})\mid\mathsf{x}_k-\mathsf{x}^\dagger\Big\rangle\\&\leq\frac{\|L\|}{2}\Big(\|\mathsf{x}_k-\mathsf{x}^\dagger\|^2+\|\mathsf{x}_k-\mathsf{x}_{k-1}\|^2\Big).\end{aligned}\tag{3.57}$$

The inequality (3.57) implies that

$$\frac{1}{2}\|\mathsf{x}_k-\mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k}+\frac{1}{2}\|\mathsf{x}_k-\mathsf{x}_{k-1}\|^2_{\boldsymbol{V}_k}-\boldsymbol{b}_k(\mathsf{x}^\dagger)\geq\frac{1}{2}\|\mathsf{x}_k-\mathsf{x}_{k-1}\|^2_{\boldsymbol{V}_k-\|L\|\,\mathrm{Id}}+\frac{1}{2}\|\mathsf{x}_k-\mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k-\|L\|\,\mathrm{Id}},\tag{3.58}$$

where the last inequality follows from $\boldsymbol{V}_k-\|L\|\,\mathrm{Id}\succeq 0$ in (3.45). Hence,

$$\mathcal{L}_{k+1}(\mathsf{x}^\dagger)\geq\frac{1}{2}\|\mathsf{x}_{k+1}-\mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k-\|L\|\,\mathrm{Id}}\geq\frac{\epsilon}{2}\|\mathsf{x}_{k+1}-\mathsf{x}^\dagger\|^2.\tag{3.59}$$

Moreover, in terms of the Lyapunov function defined by (3.81), we can rewrite (3.56) as

$$\mathsf{E}_k\left[\mathcal{L}_{k+1}(\mathsf{x}^\dagger)\right]+\mathsf{E}_k\left[\frac{1}{2}\|\mathsf{x}_{k+1}-\mathsf{x}_k\|^2_{\Lambda_{k+1}}\right]\leq(1+\eta_k)\mathcal{L}_k(\mathsf{x}^\dagger)-\epsilon\big(\Theta(\mathsf{x}_k)+Q(\mathsf{w}_{k-1})\big),\tag{3.60}$$

which proves (3.46). □

Example 3.8 Assume $\mu=\nu=\overline{\mu}$. Then the conditions (3.43) are satisfied when the strictly positive sequence $(\tau_n,\sigma_n)_{n\in\mathbb{N}}$ verifies the following

(i) $\gamma_n\leq\max\{\tau_n,\sigma_n\}\leq\min\left\{\dfrac{\underline{p}-\epsilon}{2(1-\underline{p})},\dfrac{1-\epsilon-4\overline{\mu}q}{8\overline{\mu}(1-\underline{p})}\right\}$.

(ii) $\tau_n\leq(\|L\|+\epsilon)^{-1}$ and $\sigma_n\leq(\|L\|+\epsilon)^{-1}$.

(iii) $\tau_n\big(\|L\|-2\overline{\mu}(1+2p)\big)\leq 4(1-p)$ and $\sigma_n\big(\|L\|-2\overline{\mu}(1+2q)\big)\leq 4(1-q)$.

(iv) $1/\tau_n\geq 4\overline{\mu}(2-p)+\chi_0$ and $1/\sigma_n\geq 4\overline{\mu}(2-q)+\chi_0$, where $\chi_0=\mu+\mu^{-1}\|L\|^2$.

Example 3.9 Set $\eta_k\equiv 0$ and $\tau_k=\sigma_k\equiv\gamma$. For simplicity, assume $\mu=\nu=\overline{\mu}$ and $p=q$, thus $\underline{p}=p$. Then, $\boldsymbol{U}_k=\gamma^{-1}\,\mathrm{Id}$ and $\boldsymbol{D}=\overline{\mu}\,\mathrm{Id}$. Set

$$\chi=\begin{cases}\infty & \text{if } \|L\|>2\mu(1+2p)\\ 4(1-p)/\big(\|L\|-2\mu(1+2p)\big) & \text{otherwise.}\end{cases}$$

[3] Since both expectations exist, the LHS can be equivalently written as $\mathsf{E}_k\left[\Theta(\mathsf{x}_{k+1})+Q(\mathsf{w}_k)\right]$.

It follows that the first and the second condition in (3.45) are satisfied with

$$\gamma \leq \min\left\{\frac{p-\epsilon}{2(1-p)}, \frac{1-\epsilon-4\overline{\mu}p}{8\overline{\mu}(1-p)}, \frac{1}{\|L\|+\epsilon}\right\}. \tag{3.61}$$

The third condition of (3.43) is automatically satisfied when $\gamma \leq \chi$. Moreover, following (3.43),

$$\Lambda_k \equiv \gamma^{-1}\,\mathrm{Id} - \overline{\mu}^{-1}\boldsymbol{L}^{\top}\boldsymbol{L} - 4\overline{\mu}\,\mathrm{Id}\big(1+p+2\gamma(1-p)\big). \tag{3.62}$$

Hence, the last condition of (3.45) is also satisfied when

$$\gamma^{-1} \geq 1 + \overline{\mu}^{-1}\|L\|^2 + 4\overline{\mu} \geq \epsilon + \overline{\mu}^{-1}\|L\|^2 + 4\overline{\mu}\big(1+p+2\gamma(1-p)\big). \tag{3.63}$$

Therefore, all conditions of (3.45) are verified when

$$0 < \epsilon \leq \gamma \leq \min\left\{\frac{p-\epsilon}{2(1-p)}, \frac{1-\epsilon-4\overline{\mu}p}{8\overline{\mu}(1-p)}, \frac{1}{\|L\|+\epsilon}, \frac{1}{1+4\overline{\mu}+\overline{\mu}^{-1}\|L\|^2}, \chi\right\}. \tag{3.64}$$

Example 3.10 Let n be the epoch, and $q = p = 1/n$. Suppose that n is taken large enough such that $n \gg \max\{\|L\|+\epsilon, 1+4\overline{\mu}+\overline{\mu}^{-1}\|L\|^2\}$. Set $\epsilon = 1/(2n)$. Then, by the first element of (3.64), we obtain

$$\gamma = \frac{1}{4(n-1)}, \tag{3.65}$$

which is much better than $\gamma = \dfrac{1}{4n(\overline{\mu}+\|L\|)}$ per [1] for Problem 1.1 whenever $\overline{\mu}+\|L\| > 1$.

The main result of this Subsection can be now stated. The following theorem proves the almost sure weak convergence of the sequence $(\mathsf{x}_k)_{k\in\mathbb{N}}$ to a point $\mathsf{x}^{\dagger} \in \mathcal{S}$ and the convergence of the sequence of the gap function values to 0.

Theorem 3.11 *Under the same setting as Theorem 3.7, the following hold*

$$\Theta(\mathsf{x}_k) \to 0 \text{ and } Q(\mathsf{w}_k) \to 0. \tag{3.66}$$

Moreover, if the following condition is verified

$$\boldsymbol{U}_k^{-1} \succeq \epsilon\,\mathrm{Id}; \tag{3.67}$$

then, $(\mathsf{x}_k)_{k\in\mathbb{N}}$ converges weakly to some random variable $\overline{\mathsf{x}} \in \mathcal{S}$ almost surely (a.s.).

Proof. Under the setting of Theorem 3.7, all the conditions stated in Lemma 2.4 are satisfied. Consequently, there exists a random variable defined as $\mathcal{L}_\infty(\mathsf{x}^{\dagger})$ such that

$$\mathcal{L}_k(\mathsf{x}^{\dagger}) \to \mathcal{L}_\infty(\mathsf{x}^{\dagger}) \text{ a.s. as } k \to \infty, \tag{3.68}$$

and

$$\epsilon\sum_{k\in\mathbb{N}} \mathsf{E}_k[\frac{1}{2}\|\mathsf{x}_{k+1}-\mathsf{x}_k\|^2] \leq \sum_{k\in\mathbb{N}} \mathsf{E}_k[\frac{1}{2}\|\mathsf{x}_{k+1}-\mathsf{x}_k\|^2_{\Lambda_k}] < +\infty \text{ a.s.}, \tag{3.69}$$

and

$$\sum_{k\in\mathbb{N}} \big(\Theta(\mathsf{x}_k) + Q(\mathsf{w}_{k-1})\big) < +\infty. \tag{3.70}$$

Hence, by [31, Corollary 2.6], we also obtain

$$\sum_{k\in\mathbb{N}} \|\mathsf{x}_{k+1}-\mathsf{x}_k\|^2 < +\infty \tag{3.71}$$

$$\text{as well as } \mathsf{x}_{k+1}-\mathsf{x}_k \to 0, \text{ and } \Theta(\mathsf{x}_k)+Q(\mathsf{w}_{k-1}) \to 0 \text{ a.s.} \tag{3.72}$$

Therefore, (3.66) is proved. We also derive from (3.53) that

$$\mathsf{E}_k\left[\|\mathbf{r}_k-\mathbf{R}_k\|^2\right] \to 0, \tag{3.73}$$

and thus, that

$$\|\hat{\mathsf{x}}_{k+1}-\mathsf{x}_{k+1}\| \to 0 \text{ and } \|\hat{\mathsf{x}}_{k+1}-\mathsf{x}_k\| \to 0. \tag{3.74}$$

Moreover, (3.68) implies that $(\mathcal{L}_k(\mathsf{x}^\dagger))_{k\in\mathbb{N}}$ is bounded a.s. Hence, by (3.59), $(\|\mathsf{x}_{k+1}-\mathsf{x}^\dagger\|_{\boldsymbol{U}_k})_{k\in\mathbb{N}}$ is also bounded a.s. In turn,

$$|\boldsymbol{b}_{k+1}(\mathsf{x}^\dagger)| \leq \|L\|\|\mathsf{x}_{k+1}-\mathsf{x}_k\|\|\mathsf{x}_{k+1}-\mathsf{x}^\dagger\|_{\boldsymbol{U}_k} \to 0. \tag{3.75}$$

Now, we derive from (3.72), (3.75) and (3.68) that

$$\lim \mathcal{L}_{k+1}(\mathsf{x}^\dagger) = \lim \|\mathsf{x}_{k+1}-\mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} = \mathcal{L}_\infty(\mathsf{x}^\dagger) \text{ a.s.}, \tag{3.76}$$

which, in particular, implies that $(\mathsf{x}_k)_{k\in\mathbb{N}}$ is bounded almost surely. Let $\bar{\mathsf{x}}$ be a weak cluster point of $(\mathsf{x}_k)_{k\in\mathbb{N}}$, i.e., there exists a subsequence $(\mathsf{x}_{n_k})_{k\in\mathbb{N}}$ that converges weakly a.s to $\bar{\mathsf{x}}$. Note that $(\mathsf{y}_{n_k})_{k\in\mathbb{N}}$ and $(\hat{\mathsf{x}}_{n_k})_{k\in\mathbb{N}}$ also converge weakly a.s to $\bar{\mathsf{x}}$. As $k\to\infty$, from

$$\boldsymbol{U}_k^{-1}(\mathsf{x}_k-\hat{\mathsf{x}}_{k+1}-\boldsymbol{C}\mathsf{y}_k) \in \boldsymbol{M}\hat{\mathsf{x}}_{k+1}, \tag{3.77}$$

and (3.67), we obtain $\bar{\mathsf{x}} \in \operatorname{zer}(\boldsymbol{M}+\boldsymbol{C}) = \mathcal{S}$ a.s. Therefore, by [31, Proposition 2.5], $(\mathsf{x}_k)_{k\in\mathbb{N}}$ converges weakly a.s. to a point in $\mathcal{S}$. □

Theorem 3.12 *Let $(\beta_k)_{k\in\mathbb{N}}$ be a sequence in $]0,+\infty[$. Under the same setting as of Lemma 3.4, define*

$$\begin{cases} \mathbf{S}_k = 2\boldsymbol{D} + 4\beta_k^{-1}\boldsymbol{L}^\intercal\boldsymbol{U}_k^{-1}\boldsymbol{L} \\ \mathbf{T}_{k+1} = \mathbf{S}_k + 4\beta_k^{-1}\boldsymbol{U}_k + 4\boldsymbol{L}^\intercal\boldsymbol{D}^{-1}\boldsymbol{L}. \end{cases} \tag{3.78}$$

Then, for every $k\in\mathbb{N}$, the smoothed gap G_{β_k} is bounded by

$$\begin{aligned} G_{\beta_k}(\mathsf{x}_{k+1};\mathsf{x}^\dagger) \leq{}& \frac{1}{2}\|\mathsf{x}_k-\mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_k-\mathsf{x}_{k-1}\|^2_{\mathbf{S}_k} - \boldsymbol{b}_k(\mathsf{x}^\dagger) \\ &- \left(\frac{1}{2}\|\mathsf{x}_{k+1}-\mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_{k+1}-\mathsf{x}_k\|^2_{\mathbf{S}_{k+1}} - \boldsymbol{b}_{k+1}(\mathsf{x}^\dagger)\right) \\ &+ \frac{1}{2}\|\mathsf{x}_{k+1}-\mathsf{x}_k\|^2_{\mathbf{S}_{k+1}+\mathbf{T}_{k+1}} - \frac{1}{2}\|\mathsf{x}_{k+1}-\mathsf{x}_k\|^2_{\boldsymbol{U}_k} \\ &+ (1+2/\beta_k)\|\mathbf{r}_k-\mathbf{R}_k\|^2_{\boldsymbol{U}_k^{-1}} + \left\langle \hat{\mathsf{x}}_{k+1}-\mathsf{x}^\dagger \mid \mathbf{R}_k-\mathbf{r}_k \right\rangle. \end{aligned} \tag{3.79}$$

Moreover, for all $k\in\mathbb{N}$, set

$$\begin{cases} \boldsymbol{Z}_k &= \mathbf{S}_k + 8\boldsymbol{D}P_k\boldsymbol{U}_k^{-1} + 4\boldsymbol{D}(\operatorname{Id}-P), \\ \Lambda_{k+1} &= \boldsymbol{U}_k - \mathbf{T}_{k+1} - \boldsymbol{Z}_{k+1} \end{cases} \tag{3.80}$$

where $P_k = (1 + 2/\beta_k)P$*, and define the following Lyapunov function*

$$\mathcal{L}_{\beta_k}(\mathsf{x}^\dagger) = G_{\beta_k}(\mathsf{x}_k; \mathsf{x}^\dagger) + Q(\mathsf{w}_{k-1}) + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_{k-1}} - \boldsymbol{b}_k(\mathsf{x}^\dagger) + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{\boldsymbol{Z}_k}. \tag{3.81}$$

Let $(\eta_k)_{k\in\mathbb{N}}$ *be a sequence in* $\ell^1_+(\mathbb{N})$*. Suppose that for all* $k \in \mathbb{N}$*, the following conditions are verified.*

$$\begin{cases} \beta_k \geq \beta_{k-1} \\ 4\overline{\mu}\big(2\overline{\gamma}_k(1-\underline{p}) + q\big) + \epsilon \leq 1 + \eta_k;\ (2\overline{\gamma}_k + 1)(1-\underline{p}) + \epsilon \leq 1 + \eta_k,\ \textit{with}\ \overline{\gamma}_k = \gamma_k(1 + 2/\beta_k) \\ \boldsymbol{U}_{k-1} \succeq \boldsymbol{U}_k \succeq (\epsilon + \|L\|)\,\mathrm{Id} \\ \boldsymbol{Z}_k \succeq \|L\|\,\mathrm{Id} \\ \overline{\Lambda}_k \succeq \epsilon\,\mathrm{Id}\,. \end{cases} \tag{3.82}$$

Then, the following descent property is verified for all $k \in \mathbb{N}$

$$\mathsf{E}_k\left[\mathcal{L}_{\beta_{k+1}}(\mathsf{x}^\dagger)\right] + \mathsf{E}_k\left[\frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\Lambda_{k+1}}\right] \leq (1+\eta_k)\mathcal{L}_{\beta_k}(\mathsf{x}^\dagger) - \epsilon\big(G_{\beta_k}(\mathsf{x}_k; \mathsf{x}^\dagger) + Q(\mathsf{w}_{k-1})\big). \tag{3.83}$$

Consequently, if $\boldsymbol{U}_k^{-1} \succeq \epsilon\,\mathrm{Id}$*,* $(\mathsf{x}_k)_{k\in\mathbb{N}}$ *converges weakly to some random variable* $\overline{\mathsf{x}} \in \mathcal{S}$ *almost surely, and*

$$G_{\beta_k}(\mathsf{x}_k; \mathsf{x}^\dagger) \to 0 \textit{ and } Q(\mathsf{w}_k) \to 0 \textit{ a.s.} \tag{3.84}$$

Proof. We have the following estimations

$$\begin{aligned} \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}\|^2_{\boldsymbol{U}_k} - \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}\|^2_{\boldsymbol{U}_k} &= \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} - \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}^\star\|^2_{\boldsymbol{U}_k} + \Big\langle \boldsymbol{U}_k(\mathsf{x}_k - \mathsf{x}_{k+1}) \mid \mathsf{x}^\dagger - \mathsf{x}\Big\rangle \\ &\leq \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} - \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}^\star\|^2_{\boldsymbol{U}_k} + \frac{2}{\beta_k}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\boldsymbol{U}_k} + \frac{\beta_k}{8}\|\mathsf{x} - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k}, \end{aligned} \tag{3.85}$$

and

$$\begin{aligned} \boldsymbol{b}_k(\mathsf{x}) - \boldsymbol{b}_{k+1}(\mathsf{x}) &= \boldsymbol{b}_k(\mathsf{x}^\dagger) - \boldsymbol{b}_{k+1}(\mathsf{x}^\dagger) + \Big\langle \mathbf{L}(\mathsf{x}_k - \mathsf{x}_{k-1}) \mid \mathsf{x}^\dagger - \mathsf{x}\Big\rangle - \Big\langle \mathbf{L}(\mathsf{x}_{k+1} - \mathsf{x}_k) \mid \mathsf{x}^\dagger - \mathsf{x}\Big\rangle \\ &= \boldsymbol{b}_k(\mathsf{x}^\dagger) - \boldsymbol{b}_{k+1}(\mathsf{x}^\dagger) + \frac{2}{\beta_k}\|\boldsymbol{U}_k^{-1}\boldsymbol{L}(\mathsf{x}_k - \mathsf{x}_{k-1})\|^2_{\boldsymbol{U}_k} + \frac{2}{\beta_k}\|\boldsymbol{U}_k^{-1}\boldsymbol{L}(\mathsf{x}_{k+1} - \mathsf{x}_k)\|^2 \\ &\quad + \frac{\beta_k}{4}\|\mathsf{x} - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} \\ &= \boldsymbol{b}_k(\mathsf{x}^\dagger) - \boldsymbol{b}_{k+1}(\mathsf{x}^\dagger) + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{4\beta_k^{-1}\boldsymbol{L}^\intercal\boldsymbol{U}_k^{-1}\boldsymbol{L}} + \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{4\beta_k^{-1}\boldsymbol{L}^\intercal\boldsymbol{U}_k^{-1}\boldsymbol{L}} \\ &\quad + \frac{\beta_k}{4}\|\mathsf{x} - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k}, \end{aligned} \tag{3.86}$$

and

$$\begin{aligned} \langle \hat{\mathsf{x}}_{k+1} - \mathsf{x} \mid \mathbf{R}_k - \mathbf{r}_k\rangle &= \Big\langle \hat{\mathsf{x}}_{k+1} - \mathsf{x}^\dagger \mid \mathbf{R}_k - \mathbf{r}_k\Big\rangle + \Big\langle \mathsf{x}^\dagger - \mathsf{x} \mid \mathbf{R}_k - \mathbf{r}_k\Big\rangle \\ &\leq \Big\langle \hat{\mathsf{x}}_{k+1} - \mathsf{x}^\dagger \mid \mathbf{R}_k - \mathbf{r}_k\Big\rangle + \frac{2}{\beta_k}\|\boldsymbol{U}_k^{-1}(\mathbf{r}_k - \mathbf{R}_k)\|^2_{\boldsymbol{U}_k} + \frac{\beta_k}{8}\|\mathsf{x} - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} \\ &\leq \Big\langle \hat{\mathsf{x}}_{k+1} - \mathsf{x}^\dagger \mid \mathbf{R}_k - \mathbf{r}_k\Big\rangle + \frac{2}{\beta_k}\|\mathbf{r}_k - \mathbf{R}_k\|^2_{\boldsymbol{U}_k^{-1}} + \frac{\beta_k}{8}\|\mathsf{x} - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k}. \end{aligned} \tag{3.87}$$

Therefore, (3.17) becomes

$$\begin{aligned}
K(x_{k+1}, v) - K(x, v_{k+1}) - \frac{\beta_k}{2}\|\mathsf{x} - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} \\
\leq \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{2\boldsymbol{D}+4\beta_k^{-1}\boldsymbol{L}^\intercal \boldsymbol{U}_k^{-1}\boldsymbol{L}} - \boldsymbol{b}_k(\mathsf{x}^\dagger) \\
- \left(\frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{2\boldsymbol{D}} - \boldsymbol{b}_{k+1}(\mathsf{x}^\dagger)\right) \\
+ \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{4\boldsymbol{D}+\boldsymbol{L}^\intercal\boldsymbol{D}^{-1}\boldsymbol{L}+4\beta_k^{-1}\boldsymbol{U}_k+4\beta_k^{-1}\boldsymbol{L}^\intercal\boldsymbol{U}_k^{-1}\boldsymbol{L}} - \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\boldsymbol{U}_k} \\
+ (1 + 2/\beta_k)\|\mathbf{r}_k - \mathbf{R}_k\|^2_{\boldsymbol{U}_k^{-1}} + \left\langle \hat{\mathsf{x}}_{k+1} - \mathsf{x}^\dagger \mid \mathbf{R}_k - \mathbf{r}_k \right\rangle. \qquad (3.88)
\end{aligned}$$

Taking the supremun over $\mathsf{x} \in \mathrm{dom}(f) \times \mathrm{dom}(g^\star)$ and using (3.78), we obtain (3.79). Moreover, it follows from (3.53) that

$$(1 + 2/\beta_k)\mathsf{E}_k\left[\|\mathbf{r}_k - \mathbf{R}_k\|^2_{\boldsymbol{U}_k^{-1}}\right] \leq 2\overline{\gamma}_k(1 - \underline{p})\big(Q(\mathsf{w}_{k-1}) + 4\overline{\mu}\Theta(\mathsf{x}_k)\big) + \|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{4\boldsymbol{D}P_k\boldsymbol{U}_k^{-1}} \qquad (3.89)$$

Hence, by taking conditional expectation on both sides of (3.79), we obtain

$$\begin{aligned}
\mathsf{E}_k\left[G_{\beta_k}(\mathsf{x}_{k+1}; \mathsf{x}^\dagger)\right] \leq 2\overline{\gamma}_k(1 - \underline{p})\big(Q(\mathsf{w}_{k-1}) + 4\overline{\mu}\Theta(\mathsf{x}_k)\big) + \|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{4\boldsymbol{D}P_k\boldsymbol{U}_k^{-1}} \\
+ \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{\mathbf{S}_k} - \boldsymbol{b}_k(\mathsf{x}^\dagger) \\
- \left(\frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\mathbf{S}_{k+1}} - \boldsymbol{b}_{k+1}(\mathsf{x}^\dagger)\right) \\
+ \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\mathbf{S}_{k+1}+\mathbf{T}_{k+1}} - \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\boldsymbol{U}_k}. \qquad (3.90)
\end{aligned}$$

Adding (3.55) to (3.90), we obtain

$$\begin{aligned}
\mathsf{E}_k\left[G_{\beta_k}(\mathsf{x}_{k+1}; \mathsf{x}^\dagger)\right] + \mathsf{E}_k\left[Q(\mathsf{w}_k)\right] \leq 4\overline{\mu}\big(2\overline{\gamma}_k(1 - \underline{p}) + q\big)\Theta(\mathsf{x}_k) + (2\overline{\gamma}_k + 1)(1 - \underline{p})Q(\mathsf{w}_{k-1}) \\
+ \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{\mathbf{S}_k+8\boldsymbol{D}P_k\boldsymbol{U}_k^{-1}+4\boldsymbol{D}(\mathrm{Id}-P)} - \boldsymbol{b}_k(\mathsf{x}^\dagger) \\
- \left(\frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\mathbf{S}_{k+1}} - \boldsymbol{b}_{k+1}(\mathsf{x}^\dagger)\right) \\
+ \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\mathbf{S}_{k+1}+\mathbf{T}_{k+1}} - \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\boldsymbol{U}_k}. \qquad (3.91)
\end{aligned}$$

In view of the notations defined in (3.80), we can rewrite (3.91) as

$$\begin{aligned}
\mathsf{E}_k\left[G_{\beta_k}(\mathsf{x}_{k+1}; \mathsf{x}^\dagger)\right] + \mathsf{E}_k\left[Q(\mathsf{w}_k)\right] \leq 4\overline{\mu}\big(2\overline{\gamma}_k(1 - \underline{p}) + q\big)\Theta(\mathsf{x}_k) + (2\overline{\gamma}_k + 1)(1 - \underline{p})Q(\mathsf{w}_{k-1}) \\
+ \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{\boldsymbol{Z}_k} - \boldsymbol{b}_k(\mathsf{x}^\dagger) \\
- \left(\frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\boldsymbol{Z}_{k+1}} - \boldsymbol{b}_{k+1}(\mathsf{x}^\dagger)\right) \\
- \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\Lambda_{k+1}}. \qquad (3.92)
\end{aligned}$$

Since $(\beta_k)_{k\in\mathbb{N}}$ is assumed increasing, $G_{\beta_k}(\mathsf{x}_{k+1};\mathsf{x}^\dagger) \geq G_{\beta_{k+1}}(\mathsf{x}_{k+1};\mathsf{x}^\dagger)$. Moreover, by Lemma 2.2, $\Theta(\mathsf{x}_k) \leq G_{\beta_k}(\mathsf{x}_k;\mathsf{x}^\dagger)$. Therefore, (3.92) can be further estimated as follows

$$\begin{aligned} \mathsf{E}_k\left[\mathcal{L}_{\beta_{k+1}}(\mathsf{x}^\dagger)\right] + \mathsf{E}_k\left[\frac{1}{2}\|\mathsf{x}_{k+1}-\mathsf{x}_k\|^2_{\Lambda_{k+1}}\right] &\leq \big(2\overline{\gamma}_k(1-\underline{p})+1-\underline{p}\big)\big(4\overline{\mu}G_{\beta_k}(\mathsf{x}_k;\mathsf{x}^\dagger)+Q(\mathsf{w}_{k-1})\big) \\ &\quad + \frac{1}{2}\|\mathsf{x}_k-\mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_k-\mathsf{x}_{k-1}\|^2_{\boldsymbol{Z}_k} - \boldsymbol{b}_k(\mathsf{x}^\dagger) \\ &\leq (1+\eta_k)\mathcal{L}_{\beta_k}(\mathsf{x}^\dagger) - \epsilon\big(G_{\beta_k}(\mathsf{x}_k;\mathsf{x}^\dagger)+Q(\mathsf{w}_{k-1})\big). \end{aligned} \tag{3.93}$$

Hence, (3.83) is proved. The remainder of the proof is similar to the proof of Theorem 3.11, and we omit it here. □

Remark 3.13 Here are some comments.

(i) The weak convergence of the iterate as well as the convergence of the gap appear to be new in the context of loopless variance reduction method for solving primal-dual problem. In the case of non-loopless variance reduction method, this kind of result has also been obtained in [22]. While, the proof of the almost sure convergence of iteration based on the gap function is not new approach even in the stochastic; see [28, 22] for instances.

(ii) To the best of our knowledge, our results appear to be the first establishing the (weak) convergence of the smoothed gap introduced by [13] in the stochastic setting.

3.2.2 Linear convergence

In this section, we study the linear convergence properties of the proposed algorithm. More precisely, we establish the linear convergence in expectation of the gap as well as the iteration.

Theorem 3.14 *Suppose that f and $g^\star$ are strongly convex functions with strictly positive constants θ_1 and θ_2, respectively. For every $k\in\mathbb{N}$, set $\epsilon_1 = \inf_{k\in\mathbb{N}}\min\{\theta_1\tau_k, \theta_2\sigma_k\}$. Suppose that*

$$(\forall k\in\mathbb{N})\max\{4\overline{\mu}\big(2\gamma_k(1-\underline{p})+q\big);(2\gamma_k+1)(1-\underline{p})\} \leq \rho_0 < 1;\ (2+\epsilon_1)(1-\rho_0) \leq \epsilon_1. \tag{3.94}$$

and that

$$\Lambda_k \succeq \frac{1-\rho_0}{\rho_0}(\boldsymbol{V}_{k+1} + \boldsymbol{L}^\intercal \boldsymbol{U}^{-1}_{k+1}\boldsymbol{L}). \tag{3.95}$$

Then, the following holds.

$$\mathsf{E}_k[\Theta(\mathsf{x}_{k+1})] = \mathcal{O}(\rho_0^k),\ \mathsf{E}_k[Q(\mathsf{w}_k)] = \mathcal{O}(\rho_0^k) \text{ and } \mathsf{E}_k\left[[\|\mathsf{x}_{k+1}-\mathsf{x}^\dagger\|^2_{\boldsymbol{U}_{k+1}}\right] \leq \mathcal{O}(\rho_0^k). \tag{3.96}$$

Proof. By using (3.39), instead of (3.56), we obtain

$$\begin{aligned} \mathsf{E}_k\left[\Theta(\mathsf{x}_{k+1}) + Q(\mathsf{w}_k) + \frac{\epsilon_1}{2}\|\mathsf{x}_{k+1}-\mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k}\right] &\leq 4\overline{\mu}\big(2\gamma_k(1-\underline{p})+q\big)\Theta(\mathsf{x}_k) + (2\gamma_k+1)(1-\underline{p})Q(\mathsf{w}_{k-1}) \\ &\quad + \frac{1}{2}\|\mathsf{x}_k-\mathsf{x}\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_k-\mathsf{x}_{k-1}\|^2_{\boldsymbol{V}_k} - \boldsymbol{b}_k(\mathsf{x}^\dagger) \\ &\quad - \mathsf{E}_k\left[\frac{1}{2}\|\mathsf{x}_{k+1}-\mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_{k+1}-\mathsf{x}_k\|^2_{\boldsymbol{V}_{k+1}} - \boldsymbol{b}_{k+1}(\mathsf{x}^\dagger)\right] \\ &\quad - \mathsf{E}_k\left[\frac{1}{2}\|\mathsf{x}_{k+1}-\mathsf{x}_k\|^2_{\Lambda_k}\right]. \end{aligned} \tag{3.97}$$

This inequality together with the condition (3.94) gives

$$
\begin{aligned}
\mathsf{E}_k\left[\Theta(\mathsf{x}_{k+1})\right] + \mathsf{E}_k\left[Q(\mathsf{w}_k)\right] \leq \rho_0\big(\Theta(\mathsf{x}_k) + Q(\mathsf{w}_{k-1})\big)& \\
+ \frac{1+\epsilon_1}{2}\|\mathsf{x}_k - \mathsf{x}\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{\boldsymbol{V}_k} - \boldsymbol{b}_k(\mathsf{x}^\dagger) - \frac{\epsilon_1}{2}\|\mathsf{x}_k - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k}& \\
- \mathsf{E}_k\left[\frac{1+\epsilon_1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\boldsymbol{V}_{k+1}} - \boldsymbol{b}_{k+1}(\mathsf{x}^\dagger)\right]& \\
- \mathsf{E}_k\left[\frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\Lambda_k}\right].& \qquad (3.98)
\end{aligned}
$$

Let us set

$$
\boldsymbol{a}_k = \frac{1+\epsilon_1}{2}\|\mathsf{x}_k - \mathsf{x}\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{\boldsymbol{V}_k} - \boldsymbol{b}_k(\mathsf{x}^\dagger). \qquad (3.99)
$$

Then,

$$
\begin{aligned}
(1-\rho_0)\boldsymbol{b}_k(\mathsf{x}^\star) &= (1-\rho_0)\left\langle \boldsymbol{U}_k^{-1}\mathbf{L}(\mathsf{x}_k - \mathsf{x}_{k-1}) \mid \mathsf{x}_k - \mathsf{x}^\dagger\right\rangle_{\boldsymbol{U}_k} \\
&\leq \frac{(1-\rho_0)}{2}\|\mathsf{x}_k - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{(1-\rho_0)}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{\boldsymbol{L}^\intercal\boldsymbol{U}_k^{-1}\boldsymbol{L}} \qquad (3.100)
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\boldsymbol{a}_k &= \rho_0\boldsymbol{a}_k + (1-\rho_0)\boldsymbol{a}_k \\
&= \rho_0\boldsymbol{a}_k + \frac{(2+\epsilon_1)(1-\rho_0)}{2}\|\mathsf{x}_k - \mathsf{x}\|^2_{\boldsymbol{U}_k} + \frac{1-\rho_0}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{\boldsymbol{V}_k + \boldsymbol{L}^\intercal\boldsymbol{U}_k^{-1}\boldsymbol{L}}. \qquad (3.101)
\end{aligned}
$$

Now, using the second condition in (3.94), i.e., $(2+\epsilon_1)(1-\rho_0) \leq \epsilon_1$, we obtain

$$
\begin{aligned}
\boldsymbol{a}_k - \frac{\epsilon_1}{2}\|\mathsf{x}_k - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} &\leq \rho_0\boldsymbol{a}_k + \frac{1-\rho_0}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{\boldsymbol{V}_k + \boldsymbol{L}^\intercal\boldsymbol{U}_k^{-1}\boldsymbol{L}} \\
&= \rho_0\Big(\boldsymbol{a}_k + \frac{1-\rho_0}{2\rho_0}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{\boldsymbol{V}_k + \boldsymbol{L}^\intercal\boldsymbol{U}_k^{-1}\boldsymbol{L}}\Big). \qquad (3.102)
\end{aligned}
$$

Therefore, (3.98) can be further estimated as

$$
\begin{aligned}
\mathsf{E}_k\left[\Theta(\mathsf{x}_{k+1})\right] + \mathsf{E}_k\left[Q(\mathsf{w}_k)\right] &\leq \rho_0\Big(\Theta(\mathsf{x}_k) + Q(\mathsf{w}_{k-1}) + \boldsymbol{a}_k + \frac{1-\rho_0}{2\rho_0}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{\boldsymbol{V}_k + \boldsymbol{L}^\intercal\boldsymbol{U}_k^{-1}\boldsymbol{L}}\Big) \\
&\quad - \mathsf{E}_k\left[\boldsymbol{a}_{k+1}\right] - \mathsf{E}_k\left[\frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\Lambda_k}\right] \\
&= \rho_0\Big(\Theta(\mathsf{x}_k) + Q(\mathsf{w}_{k-1}) + \boldsymbol{a}_k + \frac{1-\rho_0}{2\rho_0}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{\boldsymbol{V}_k + \boldsymbol{L}^\intercal\boldsymbol{U}_k^{-1}\boldsymbol{L}}\Big) \\
&\quad - \mathsf{E}_k\left[\boldsymbol{a}_{k+1} + \frac{1-\rho_0}{2\rho_0}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\boldsymbol{V}_{k+1} + \boldsymbol{L}^\intercal\boldsymbol{U}_{k+1}^{-1}\boldsymbol{L}}\right] \\
&\quad + \mathsf{E}_k\left[\frac{1-\rho_0}{2\rho_0}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\boldsymbol{V}_{k+1} + \boldsymbol{L}^\intercal\boldsymbol{U}_{k+1}^{-1}\boldsymbol{L}}\right] - \mathsf{E}_k\left[\frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\Lambda_k}\right] \\
&\qquad (3.103)
\end{aligned}
$$

The difference between the last two terms in (3.103) is negative due to the condition (3.95). Therefore,

$$\begin{aligned} \mathsf{E}_k\left[\Theta(\mathsf{x}_{k+1})\right] + &\mathsf{E}_k\left[Q(\mathsf{w}_k) + \boldsymbol{a}_{k+1} + \frac{1-\rho_0}{2\rho_0}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\boldsymbol{V}_{k+1} + \boldsymbol{L}^{\intercal}\boldsymbol{U}^{-1}_{k+1}\boldsymbol{L}}\right] \\ &\leq \rho_0\Big(\Theta(\mathsf{x}_k) + Q(\mathsf{w}_{k-1}) + \boldsymbol{a}_k + \frac{1-\rho_0}{2\rho_0}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{\boldsymbol{V}_k + \boldsymbol{L}^{\intercal}\boldsymbol{U}_k^{-1}\boldsymbol{L}}\Big) \end{aligned} \tag{3.104}$$

Using this expression recursively, we obtain

$$\mathsf{E}_k\left[\Theta(\mathsf{x}_{k+1})\right] + \mathsf{E}_k\left[Q(\mathsf{w}_k) + \boldsymbol{a}_{k+1} + \frac{1-\rho_0}{2\rho_0}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\boldsymbol{V}_{k+1} + \boldsymbol{L}^{\intercal}\boldsymbol{U}^{-1}_{k+1}\boldsymbol{L}}\right] \leq \mathbf{O}(\rho_0^k), \tag{3.105}$$

which proves the desired results. □

Remark 3.15 By using the same technique, the linear convergence of the smoothed gap function value can be obtained. Hence, we omit it here.

Remark 3.16 The linear convergence of the duality gap as well as the smoothed gap function values under an additional condition like the strong convexity-concavity or the quadratic error bound are well-known in both stochastic and deterministic settings; see for examples [3, 13, 21, 29]. If $\ell^\star = 0$ and $f = 0$, under additional assumptions on the linear operator L, [12] achieves the linear convergence rate even when the strongly convex-concave condition is not full-filled.

4 Conclusion

In this paper, we developed a new primal-dual splitting with loopless variance reduction. We proved the weak almost sure convergence of the iterations and the convergence of the gap function as well as of the full gradient. Linear convergence is also obtained under the strong convexity condition. We also note that when Step 1 of Algorithm 3.1 is modified as

$$\begin{cases} y_k &= (1+\omega_k)x_k - \omega_k x_{k-1} \\ u_k &= (1+\omega_k)v_k - \omega_k v_{k-1} \end{cases}$$

where $\omega_k \geq 0$; then, under the same conditions on ω_k as those used in [22], all results presented in this paper can be extend to this general case with minor modification of the conditions.

References

[1] Alacaoglu, A., Malitsky, Y. Cevher, V, Forward-reflected-backward method with variance reduction, *Comput. Optim. Appl.*, Vol. 80, pp.321-346, 2021.

[2] 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.

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

[4] 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.

[5] M.N. Bùi and P. L . Combettes, Multivariate monotone inclusions in saddle form, *Mathematics of Operations Research*, Vol. 47, pp. 1082-1109, 2022.

[6] 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.

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

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

[9] 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.

[10] 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.

[11] Q. Tran-Dinh, O. Fercoq and V. Cevher, A Smooth Primal-Dual Optimization Framework for Nonsmooth Composite Convex Minimization, *SIAM J. Optim.*, Vol. 28, pp. 96-134, 2018.

[12] S. S. Du and 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.

[13] O. Fercoq, Quadratic error bound of the smoothed gap and the restarted averaged primal-dual hybrid gradient, *Open Journal of Mathematical Optimization*, Vol. 4, art. no. 6. 34p., 2023

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

[15] E. Y. Hamedani and A. Jalilzadeh, A stochastic variance-reduced accelerated primal-dual method for finite-sum saddle-point problems, *Comput. Optim. Appl.*, Vol. 85, pp. 653-679, 2023.

[16] D. Kovalev, S. Horvath, and P. Richtarik, Don't jump through hoops and remove those loops: SVRG and Katyusha are better without the outer loop, *In Proceedings of the 31st International Conference on Algorithmic Learning Theory*, pp 451-467, 2020.

[17] Y. Malitsky, Projected reflected gradient methods for monotone variational inequalities, *SIAM J. Control Optim.*, Vol. 25, pp. 502–520, 2015.

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

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

[20] 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.

[21] V. D. Nguyen and B. C. Vũ, A Stochastic Variance Reduction Algorithm with Bregman Distances for Structured Composite Problems, *Optimization*, Vol. 72, pp. 1463-1484, 2023.

[22] V. D. Nguyen, B. C. Vũ and D. Papadimitriou, A Stochastic Primal-Dual Splitting Algorithm with Variance Reduction for Composite Optimization Problems, submitted, 2023.

[23] 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.

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

[25] 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.

[26] 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.

[27] 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.

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

[29] 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.

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

[31] B. C. Vu, Almost sure convergence of the stochastic forward-backward-forward splitting algorithm, *Optimization Letter*, vol. 10, pp. 781-803, 2016.

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

[33] 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.

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