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

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

[1] Belgium Research Center (BeRC) - Huawei, Leuven, Belgium
bangcvvn@gmail.com
[2] 3nLab@MCO Institute, Leuven & Université Libre de Bruxelles (ULB), Brussels
bangcvvn@gmail.com; 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 algorithm with loopless variance-reduction for solving this generic problem. We first prove the weak almost sure convergence of the iterates. We then demonstrate that our algorithm achieves linear convergence in expectation of its iterates as well as convergence of the (smoothed primal-dual and duality) gap function value under the assumption of strong convexity. We also derive the total average complexity and compare it to the most recent advances developed in the available literature.

**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 $\nabla h_i$ is $\mu_i$-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 $\ell_j$ 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 $\ell$ and $g$ (see Section 2 for its definition). 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 $\ell^\star$ denote the Fenchel conjugate of the function $f$ and $\ell$, respectively (see Section 2 for the definition) and $L^\star$ is the adjoint of the linear operator $L$.

**Prior and related work**: Stochastic numerical methods for solving saddle points problems have been extensively investigated in the literature, see [3, 5, 11, 12, 21, 23, 24] and [15, 17, 18, 31, 32, 37] 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 [3, 5, 11, 15, 17, 18, 31, 32, 37] and references therein. In this context, Problem 1.1 was first investigated in [12] and then in [33, 6, 7, 16] 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 [31, 24, 25] 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 $\ell$ becomes prohibitive. In turn, stochastic primal-dual splitting methods are often used as alternative to their deterministic counterpart. Comparatively,

(i) The algorithm in [31] can be viewed as a stochastic extension of [13] 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 [24] for solving the Problem 1.1 with Bregman distance. The method exploits the variance reduction technique of [35] in finite dimensional Banach space. It reaches a linear convergence rate in expectation under constraining conditions as the strong convexity relative to Bregman functions.

(iii) The method in [24] was further developed in [25] by partially relaxing the fixed setting of the extrapolation parameters, and exploiting the double-loop variance reduction technique of [35] but still restricted to the usual duality gap function.

The present work is motivated by the recent development in [19] of the loopless variance reduction method which obtains the optimal total average complexity. Their framework is nevertheless less general than the method proposed in this paper because it is concerned by the minimization of the function $h$ only. This limitation has been removed in [1] where authors developed the idea of loopless variance reduction for solving monotone inclusions which can apply to solve primal-dual problems (1.2)-(1.3). The resulting algorithm does not improve the complexity over its deterministic counterpart as confirmed by [2]. The total average complexity of the method proposed in [2] (but also those developed in [8, 18, 5]) remains far from the one obtained in [19]. Instead, this work fills this gap by developing a stochastic primal-dual splitting algorithm for Problem 1.1 that relies on loopless variance reduction and that obtains the optimal total average complexity as in [19].

**Contribution**: The main contributions of this paper are:

(i) The development of a primal-dual full-splitting method with loopless variance reduction as well as the proof of the almost sure weak convergence of the generated sequences and the convergence of the smoothed primal-dual gap function introduced in [14].

(ii) The proof of the linear convergence in expectation of the iterations as well as of the duality gap and the smoothed primal-dual gap function.

(iii) Under the strong convexity assumption, the method obtains the total average complexity as in [19] that focuses on minimizing a single function objective (referring to primal problem (1.2), the function $h$).

**Structure**: Section 2 is devoted to the definition of the notations and the introduction of basic notions This section also includes the basic results used in the next sections of this paper. We present the algorithm and prove its convergence properties in Section 3. The complexity analysis is detailed in Section 4. The last section consists of a brief conclusion.

## 2 Preliminaries

**Notations.** The inner product and norm of all Hilbert spaces are denoted by $\langle \cdot \mid \cdot \rangle$ and $\| \cdot \|$. The adjoint 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 $f$ denoted by $f^\star$ is defined as

$$f^\star(x) = \sup_{y\in\mathcal{H}}(\langle x \mid y\rangle - f(y)). \tag{2.1}$$

We also use $\partial f$ to refer to the subdifferential of $f$. Given the functions $f$ and $g$ from $\mathcal{H}$ to $\left]-\infty, +\infty\right]$, their infimal convolution $f \,\square\, g$,

$$\ell \,\square\, g\colon x \mapsto \inf_{y\in\mathcal{H}}(\ell(y) + g(x-y)). \tag{2.2}$$

The proximity operator of the scaled function $\lambda f$ from $\mathcal{H}$ to $\left]-\infty, +\infty\right]$ with parameter $\lambda > 0$ is defined by

$$\operatorname{prox}_{\lambda f} : \mathcal{H} \to \mathcal{H} : x \mapsto \operatorname*{argmin}_{y\in\mathcal{H}}\big(f(y) + \frac{1}{2\lambda}\|x-y\|^2\big). \tag{2.3}$$

Let $U, V$ be two self-adjoint bounded linear operators from $\mathcal{H}$ to $\mathcal{H}$, we write $U \succeq V$ to indicate that $(\forall x \in \mathcal{H})\ \langle x \mid Ux\rangle \geq \langle x \mid Vx\rangle$. We denote the semi-norm of $x \in \mathcal{H}$ on $U$ by $\|x\|_U = \sqrt{\langle x \mid Ux\rangle}$ where $U$ is semi-definite. We also use the following notation

$$\langle \cdot \mid \cdot \rangle_U : (x, y) \mapsto \langle x \mid Uy\rangle\,,$$

which defines a scalar product on $\mathcal{H}$ if $U$ is self-adjoint, positive definite. The following simple properties of the semi-norm will be used:

$$(\forall x \in \mathcal{H})\ \|x\|_U^2 + \|x\|_V^2 = \|x\|_{U+V}^2 \text{ and } (\forall \alpha \in [0, +\infty[)\ \alpha \|x\|_U^2 = \|x\|_{\alpha U}^2. \tag{2.4}$$

Moreover, if $U \succeq V$ then $(\forall x \in \mathcal{H})\ \|x\|_U^2 \geq \|x\|_V^2$.

**Assumptions.** As in [12], throughout this paper, we assume that the set $S$ is defined by

$$S = \big\{x \in \mathcal{H} \mid 0 \in \partial f(x) + \nabla h(x) + \big(L^\star \circ (\partial \ell \,\square\, \partial g) \circ L\big)(x)\big\} \neq \emptyset, \tag{2.5}$$

where

$$\partial f \colon \mathcal{H} \to 2^{\mathcal{H}} \colon x \mapsto \big\{u \in \mathcal{H} \mid (\forall y \in \mathcal{H})\ \ \langle y - x \mid u\rangle + f(x) \leq f(y)\big\}, \tag{2.6}$$

and

$$\partial \ell \,\square\, \partial g = (\partial \ell^\star + \partial g^\star)^{-1}. \tag{2.7}$$

As demonstrated in [12], under some qualification conditions, the primal problem (1.2) can be reduced to find a point in the set $S$ (2.5). If we denote by

$$\boldsymbol{M} \colon (x, v) \mapsto \partial f(x) \times \partial g^\star(v) \text{ and } \boldsymbol{C} \colon (x, v) \mapsto (\nabla h(x) + L^\star v) \times (\nabla \ell^\star(v) - Lx), \tag{2.8}$$

where the (Fenchel) conjugate $g^\star$ of the function $g$ is defined by (2.1). Then, under the condition (2.5), this problem becomes equivalent to

$$\mathcal{S} = \big\{(x, v) \in \mathcal{H} \times \mathcal{G} \mid 0 \in (\boldsymbol{M} + \boldsymbol{C})(x, v)\big\} \neq \emptyset. \tag{2.9}$$

We recall the definition and properties of the smoothed primal-dual gap function as introduced in [14].

**Definition 2.1** *Let $\beta \in [0, +\infty[$ and $(\tau, \sigma) \in ]0, +\infty[^2$. 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 primal-dual gap function $G_\beta(\mathsf{x}; \dot{\mathsf{x}})$ centered at $\dot{\mathsf{x}}$ is defined by*

$$G_\beta(\mathsf{x}; \dot{\mathsf{x}}) \coloneqq \sup_{x' \in \mathcal{H}, v' \in \mathcal{G}} \left(K(x, v') - K(x', v) - \frac{\beta}{2\tau}\|x' - \dot{x}\|^2 - \frac{\beta}{2\sigma}\|v' - \dot{v}\|^2\right), \tag{2.10}$$

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

$$K(x, v) \coloneqq h(x) + f(x) + \langle Lx \mid v\rangle - g^\star(v) - \ell^\star(v). \tag{2.11}$$

Observe that setting $\beta = 0$ yields the conventional duality gap function defined as

$$G_{\beta=0}(\mathsf{x}) \coloneqq \sup_{x' \in \mathcal{H}, v' \in \mathcal{G}} \big(K(x, v') - K(x', v)\big). \tag{2.12}$$

Moreover, [16] shows that the smoothed primal-dual gap function $G_\beta(\mathsf{x}; \dot{\mathsf{x}})$ as defined by (2.10) can be used as a measure of optimality in the following sense:

**Lemma 2.2** [16, Proposition 8] *Let $\beta \in [0,+\infty[$ and $(\tau,\sigma) \in ]0,+\infty[^2$. Let $\mathsf{x}^\dagger = (x^\dagger, v^\dagger) \in \mathcal{S}$, where $\mathcal{S}$ is defined by (2.9), and $\mathsf{x} = (x,v) \in \mathcal{H} \times \mathcal{G}$. Then,*

$$G_\beta(\mathsf{x};\mathsf{x}^\dagger) = 0 \quad \textit{if and only if} \quad \mathsf{x} \in \mathcal{S}. \tag{2.13}$$

*Moreover, define*

$$\begin{cases} x_\beta(x) := \operatorname{prox}_{\tau(f+h)/\beta}(x^\dagger - \tau L^\star v/\beta) \\ v_\beta(x) := \operatorname{prox}_{\sigma(g^\star+\ell^\star)/\beta}(v^\dagger + \sigma L x/\beta). \end{cases} \tag{2.14}$$

*Then, the following holds,*

$$G_\beta(\mathsf{x};\mathsf{x}^\dagger) \geq K(x,v^\dagger) - K(x^\dagger,v) + \frac{\beta}{\sigma}\|v_\beta(x) - v^\dagger\|^2 + \frac{\beta}{\tau}\|x_\beta(v) - x^\dagger\|^2. \tag{2.15}$$

We provide the main probability theory definitions and notations used throughout this manuscript. We refer the reader to [22] for more details on probability Theory in Hilbert spaces.

**Definition 2.3** *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.*

**Definition 2.4** *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 $\mathsf{E}[X|\mathcal{A}]$. The abbreviation a.s. stands for "almost surely".*

The proposed method developed in Section 3 obeys the general characterization of variance reduction as in the following definition.

**Definition 2.5** [17, Section D.] *Variance reduction (VR): method used to increase the precision of the (gradient) estimates and to improve the speed to obtain them. Formally, assume $\hat{\mathbf{h}}_k$ is an estimate of the gradient $\nabla\mathbf{h}(x_k)$. A method which verifies the property $\mathsf{E}[\|\hat{\mathbf{h}}_k - \nabla\mathbf{h}(x_k)\|^2] \xrightarrow[k\to\infty]{} 0$ is referred to as a VR method.*

In turn, variance reduction implies to specify the rate of improvement of the gradient estimate against the deterministic variant. The total average (computational) complexity is further developed in Section 4.

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

**Lemma 2.6** ([28, Theorem 1]) *Let $(\mathcal{F}_k)_{k\in\mathbb{N}}$ be an increasing sequence of sub-$\boldsymbol{\sigma}$-algebras of the $\boldsymbol{\sigma}$-algebra $\mathcal{F}$. Let $(z_k)_{k\in\mathbb{N}}$, $(\lambda_k)_{k\in\mathbb{N}}$, $(\zeta_k)_{k\in\mathbb{N}}$ and $(t_k)_{k\in\mathbb{N}}$ be sequences of $[0,+\infty[$-valued random variables such that, for every $k \in \mathbb{N}$, $z_k$, $\xi_k$, $\zeta_k$ and $t_k$ are $\mathcal{F}_k$-measurable. Assume moreover that $\sum_{k\in\mathbb{N}} t_k < +\infty$, $\sum_{k\in\mathbb{N}} \zeta_k < +\infty$ a.s. and*

$$(\forall k \in \mathbb{N}) \;\; \mathsf{E}[z_{k+1}|\mathcal{F}_k] \leq (1+t_k)z_k + \zeta_k - \lambda_k \;\; a.s..$$

*Then, the sequence $(z_k)_{k\in\mathbb{N}}$ converges a.s. to a $[0,+\infty[$-valued random variable and the sequence $(\theta_k)_{k\in\mathbb{N}}$ is summable a.s..*

# 3 Algorithm and Convergence properties

## 3.1 Algorithm

In this section, we detail our algorithm to solve the primal-dual problem (1.2)-(1.3) 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 $\nabla \ell^\star(u_k)$. Nonetheless, in our convergence analysis (see Section 3.2), we also use the full gradients at each iteration $k$ through

$$\begin{cases} \hat{x}_{k+1} = \operatorname{prox}_{\tau_k f}(x_k - \tau_k \nabla h(y_k) - \tau_k L^\star u_k) \\ \hat{v}_{k+1} = \operatorname{prox}_{\sigma_k g^\star}(v_k - \sigma_k \nabla \ell^\star(u_k) + \sigma_k L y_k), \end{cases} \tag{3.1}$$

where $\tau_k$ the primal stepsize and $\sigma_k$ the dual stepsize.

**Algorithm 3.1** Let $(\tau_k, \sigma_k)_{k\in\mathbb{N}}$ be (stepsize) sequences in $]0,+\infty[^2$. Let the priority updating probability $(p,q)$ be in $]0,1]^2$. Let $(x_0, x_{-1}) \in \mathcal{H}^2$ and $(v_0, v_{-1}) \in \mathcal{G}^2$.
Set auxiliary variables $w_1 \in \mathcal{H}$ and $w_2 \in \mathcal{G}$ at $k=0$ : $w_{1,0} = w_{1,-1} = x_0$ and $w_{2,0} = w_{2,-1} = v_0$.
Iterate
▷ Step 1. Compute

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

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

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

where

$$w_{1,k} = \begin{cases} y_k \text{ with probability } p \\ w_{1,k-1} \text{ with probability } 1-p \end{cases} \text{ and } w_{2,k} = \begin{cases} u_k \text{ with probability } q \\ w_{2,k-1} \text{ with probability } 1-q \end{cases} \tag{3.4}$$

▷ 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** Here are some remarks concerning this algorithm.

(i) The extrapolation Step 1 of Algorithm 3.1 was introduced in [20] for solving the classical variational inequality problem over a closed convex set in $\mathcal{H}$. Then, it was extended by [9] to solve a monotone inclusion. A stochastic development of [9] has been recently obtained in [26].

(ii) The idea of using the auxiliary variables $w_{1,k}$ and $w_{2,k}$ (as part of Step 2) was presented in [19] with the purpose of finding a minimizer of a single function $h$, without extrapolation Step (i.e., $y_k = x_k, u_k = v_k$). This idea was further developed in [1] for the method introduced in [20]. Algorithm 3.1 can be viewed as combining the auxiliary variables as proposed in [19] with the method developed in [9]. In particular, if $n_P = n_D = 1$, then we obtain the method in [9] for finding a point in $\mathcal{S}$, see (2.9).

(iii) The main differences of Algorithm 3.1 compared to recently published works [24, 26] consists of i) the involvement 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 $\nabla \ell^\star(u_k)$, and their variances are reduced progressively along with the convergence of the full-gradient. More precisely, we have the following.

**Lemma 3.3** *Let $\mathsf{E}_k$ 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*

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

*Moreover, let $\mathsf{x}^\dagger = (x^\dagger, v^\dagger) \in \mathcal{S}$ (2.9) and define the mean square error (MSE) over $n_P$ and $n_D$, i.e., the average squared difference between the point-wise estimation of the gradient $\nabla h_i(w_{1,k})$ and $\nabla \ell_j^\star(w_{2,k})$ (computed at the values the auxiliary variables $w_{1,k}$ and $w_{2,k}$) and the actual gradient value $\nabla h_i(x^\dagger)$ and $\nabla \ell_j^\star(v^\dagger)$, respectively*

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

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

*Then, we have*

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

*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 (3.3), 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

$$\begin{aligned}\mathsf{E}_{i_k}\left[\|z_k-\nabla h(y_k)\|^2\right] &= \mathsf{E}_{i_k}\left[\|\nabla h(w_{1,k})-\nabla h(y_k)-(\nabla h_{i_k}(w_{1,k})-\nabla h_{i_k}(y_k))\|^2\right]\\ &= \mathsf{E}_{i_k}\left[\|\nabla h(w_{1,k})-\nabla h(y_k)\|^2+\|\nabla h_{i_k}(w_{1,k})-\nabla h_{i_k}(y_k)\|^2\right]\\ &\quad -2\mathsf{E}_{i_k}\left[\langle\nabla h(w_{1,k})-\nabla h(y_k)\mid\nabla h_{i_k}(w_{1,k})-\nabla h_{i_k}(y_k)\rangle\right]\\ &= \mathsf{E}_{i_k}\left[\|\nabla h_{i_k}(w_{1,k})-\nabla h_{i_k}(y_k)\|^2\right]-\mathsf{E}_{i_k}\left[\|\nabla h(w_{1,k})-\nabla h(y_k)\|^2\right]\\ &\le \mathsf{E}_{i_k}\left[\|\nabla h_{i_k}(w_{1,k})-\nabla h_{i_k}(y_k)\|^2\right]. \end{aligned}\tag{3.10}$$

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 $\|x-y\|^2\le 2\|x-z\|^2+2\|z-y\|^2$, the left-hand side of this inequality verifies

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

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

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

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

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

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

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

which completes the proof. □

Hence, this Lemma enables to state that the variance of the estimators $z_k$ and $d_k$ is bounded and (being unbiased) is equal to their mean square error. The next Lemma provides an upper bound on the values of the difference of the Lagrangian function. To simplify notations, we also introduce the following definition

**Definition 3.4** *Set* $\mathsf{x}=(x,v)\in\operatorname{dom}(f)\times\operatorname{dom}(g^\star)$. *Define*

$$(\forall k\in\mathbb{N})\begin{cases}\mathsf{x}_k &:= (x_k,v_k),\ \hat{\mathsf{x}}_k:=(\hat{x}_k,\hat{v}_k),\ \mathsf{y}_k:=(y_k,u_k),\\ \mathbf{r}_k &:= (z_k,d_k),\\ \mathbf{R}_k &:= (\nabla h(y_k),\nabla\ell^\star(u_k)),\end{cases}\tag{3.15}$$

*and denote the infimal convolution and addition by*

$$\begin{cases}\mathbf{g} &:= g\,\square\,\ell\\ \mathbf{f} &:= f+h\end{cases}\tag{3.16}$$

**Lemma 3.5** *Set $\mu = \max_{1\le i\le n_p}\mu_i$ and $\nu = \max_{1\le j\le n_d}\nu_j$. Define*

$$\boldsymbol{L} = \begin{pmatrix} 0 & -L^\star \\ L & 0 \end{pmatrix},\ \boldsymbol{U}_k = \begin{pmatrix} \mathrm{Id}/\tau_k & 0 \\ 0 & \mathrm{Id}/\sigma_k \end{pmatrix},\ \textit{and}\ \boldsymbol{D} = \begin{pmatrix} \mu\mathrm{Id} & 0 \\ 0 & \nu\mathrm{Id} \end{pmatrix}. \tag{3.17}$$

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

$$(\forall k\in\mathbb{N})\ \boldsymbol{b}_k(\mathsf{x}) = \langle \mathsf{x}_k - \mathsf{x}_{k-1} \mid \boldsymbol{L}(\mathsf{x}_k - \mathsf{x})\rangle \tag{3.18}$$

*where $\hat{\mathsf{x}}_k$ is defined in (3.1). Then,*

$$\begin{aligned} K(x_{k+1}, v) - K(x, v_{k+1}) \le{}& \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{2\boldsymbol{D}} - \boldsymbol{b}_k(\mathsf{x}) \\ &- \Big(\frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{2\boldsymbol{D}} - \boldsymbol{b}_{k+1}(\mathsf{x})\Big) \\ &+ \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{4\boldsymbol{D}+\boldsymbol{L}^\star\boldsymbol{D}^{-1}\boldsymbol{L}} - \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\boldsymbol{U}_k} \\ &+ \|\mathbf{r}_k - \mathbf{R}_k\|^2_{\boldsymbol{U}_k^{-1}} + \langle \hat{\mathsf{x}}_{k+1} - \mathsf{x} \mid \mathbf{R}_k - \mathbf{r}_k\rangle\,. \end{aligned} \tag{3.19}$$

*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

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

which implies that

$$\begin{aligned} g^\star(v_{k+1}) - g^\star(v) &\le \langle d_k - Ly_k \mid v - v_{k+1}\rangle + \frac{1}{\sigma_k}\langle v_k - v_{k+1} \mid v_{k+1} - v\rangle \\ &= \langle d_k - Ly_k \mid v - v_{k+1}\rangle + \frac{1}{2\sigma_k}\big(\|v - v_k\|^2 - \|v_{k+1} - v_k\|^2 - \|v - v_{k+1}\|^2\big), \end{aligned} \tag{3.20}$$

where the last equality follows from the base identity in [4, Lemma 2.12(i)]. Since $\ell^\star$ is convex and continuously differentiable with $\nu$-Lipschitz gradient, it follows from the descent lemma [4, Lemma 2.64] that

$$\begin{cases} \ell^\star(u_k) - \ell^\star(v) & \le \langle u_k - v \mid \nabla\ell^\star(u_k)\rangle \\ \ell^\star(v_{k+1}) - \ell^\star(u_k) & \le \langle v_{k+1} - u_k \mid \nabla\ell^\star(u_k)\rangle + \frac{\nu}{2}\|v_{k+1} - u_k\|^2. \end{cases}$$

Adding these two inequalities, we obtain

$$\ell^\star(v_{k+1}) - \ell^\star(v) \le \langle v_{k+1} - v \mid \nabla\ell^\star(u_k)\rangle + \frac{\nu}{2}\|v_{k+1} - u_k\|^2. \tag{3.21}$$

We derive from (2.11), (3.20), and (3.21) that, for every $v\in\mathcal{G}$,

$$\begin{aligned} K(x_{k+1}, v) - K(x_{k+1}, v_{k+1}) &= \langle Lx_{k+1} \mid v - v_{k+1}\rangle + \mathbf{g}^\star(v_{k+1}) - \mathbf{g}^\star(v) \\ &\le \langle L(x_{k+1} - y_k) \mid v - v_{k+1}\rangle + \frac{1}{2\sigma_k}\big(\|v - v_k\|^2 - \|v_{k+1} - v_k\|^2 - \|v - v_{k+1}\|^2\big) \\ &\quad + \frac{\nu}{2}\|v_{k+1} - u_k\|^2 + \langle \nabla\ell^\star(u_k) - d_k \mid v_{k+1} - v\rangle\,. \end{aligned} \tag{3.22}$$

Similar to (3.22), we have, for every $x \in \mathcal{H}$,

$$\begin{aligned} K(x_{k+1}, v_{k+1}) - K(x, v_{k+1}) &= \langle L(x_{k+1} - x) \mid v_{k+1} \rangle + \mathbf{f}(x_{k+1}) - \mathbf{f}(x) \\ &\le \langle L(x_{k+1} - x) \mid v_{k+1} - u_k \rangle + \frac{1}{2\tau_k}\left(\|x - x_k\|^2 - \|x_{k+1} - x_k\|^2 - \|x - x_{k+1}\|^2\right) \\ &\quad + \frac{\mu}{2}\|x_{k+1} - y_k\|^2 + \langle x_{k+1} - x \mid \nabla h(y_k) - z_k \rangle \,. \end{aligned} \tag{3.23}$$

Adding (3.22) and (3.23), we obtain for every $(x, v) \in \mathcal{H} \times \mathcal{G}$

$$\begin{aligned} K(x_{k+1}, v) - K(x, v_{k+1}) \le \Big( &\overbrace{\langle L(x_{k+1} - x) \mid v_{k+1} - u_k \rangle}^{\alpha_{1,k}} + \overbrace{\langle L(x_{k+1} - y_k) \mid v - v_{k+1} \rangle}^{\alpha_{2,k}} \Big) \\ + &\underbrace{\frac{1}{2\tau_k}\left(\|x - x_k\|^2 - \|x_{k+1} - x_k\|^2 - \|x - x_{k+1}\|^2\right)}_{\alpha_{5,k}} + \underbrace{\frac{1}{2\sigma_k}\left(\|v - v_k\|^2 - \|v_{k+1} - v_k\|^2 - \|v - v_{k+1}\|^2\right)}_{\alpha_{6,k}} \\ + &\underbrace{\frac{\mu}{2}\|x_{k+1} - y_k\|^2 + \frac{\nu}{2}\|v_{k+1} - u_k\|^2}_{\alpha_{0,k}} + \underbrace{\langle x_{k+1} - x \mid \nabla h(y_k) - z_k \rangle}_{\alpha_{3,k}} + \underbrace{\langle \nabla \ell^\star(u_k) - d_k \mid v_{k+1} - v \rangle}_{\alpha_{4,k}} . \end{aligned} \tag{3.24}$$

Using (3.2), i.e., $u_k = v_k + v_k - v_{k-1}$, the first term in the right hand side of (3.24) can be expressed as

$$\begin{aligned} \alpha_{1,k} &= \langle L(x_{k+1} - x) \mid v_{k+1} - v_k - v_k + v_{k-1} \rangle \\ &= \langle L(x_{k+1} - x) \mid v_{k+1} - v_k \rangle - \langle L(x_{k+1} - x) \mid v_k - v_{k-1} \rangle \\ &= \langle L(x_{k+1} - x) \mid v_{k+1} - v_k \rangle - \langle L(x_{k+1} - x_k) \mid v_k - v_{k-1} \rangle - \langle L(x_k - x) \mid v_k - v_{k-1} \rangle \,. \end{aligned} \tag{3.25}$$

Similar to (3.25), for the second term of (3.24), by expanding the expression of $y_k$ (see (3.2)), we also have

$$\alpha_{2,k} = \langle L(x_{k+1} - x_k) \mid v - v_{k+1} \rangle - \langle L(x_k - x_{k-1}) \mid v - v_k \rangle - \langle L(x_k - x_{k-1}) \mid v_k - v_{k+1} \rangle \,. \tag{3.26}$$

Observe that

$$\begin{cases} \langle L(x_{k+1} - x_k) \mid v_k - v_{k-1} \rangle + \langle L(x_k - x_{k-1}) \mid v_k - v_{k+1} \rangle = \langle \mathsf{x}_k - \mathsf{x}_{k+1} \mid \boldsymbol{L}(\mathsf{x}_k - \mathsf{x}_{k-1}) \rangle \\ \langle L(x_{k+1} - x) \mid v_{k+1} - v_k \rangle + \langle L(x_{k+1} - x_k) \mid v - v_{k+1} \rangle = \langle \mathsf{x}_{k+1} - \mathsf{x}_k \mid \boldsymbol{L}(\mathsf{x}_{k+1} - \mathsf{x}) \rangle \\ \langle L(x_k - x) \mid v_k - v_{k-1} \rangle + \langle L(x_k - x_{k-1}) \mid v - v_k \rangle = \langle \mathsf{x}_k - \mathsf{x}_{k-1} \mid \boldsymbol{L}(\mathsf{x}_k - \mathsf{x}) \rangle \,. \end{cases} \tag{3.27}$$

Hence, we can derive from (3.27), (3.26) and (3.25) that

$$\alpha_{1,k} + \alpha_{2,k} = \boldsymbol{b}_{k+1}(\mathsf{x}) - \boldsymbol{b}_k(\mathsf{x}) - \langle \mathsf{x}_k - \mathsf{x}_{k+1} \mid \boldsymbol{L}(\mathsf{x}_k - \mathsf{x}_{k-1}) \rangle \,. \tag{3.28}$$

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

$$\begin{aligned} \|\hat{x}_{k+1} - x_{k+1}\| &= \left\|\operatorname{prox}_{\tau_k f}\left(x_k - \tau_k \nabla h(y_k) - \tau_k L^\star u_k\right) - \operatorname{prox}_{\tau_k f}\left(x_k - \tau_k z_k - \tau_k L^\star u_k\right)\right\| \\ &\le \tau_k \|z_k - \nabla h(y_k)\| \,. \end{aligned} \tag{3.29}$$

In turn,

$$\begin{aligned}\alpha_{3,k} &= \langle x_{k+1} - \hat{x}_{k+1} \mid \nabla h(y_k) - z_k\rangle + \langle \hat{x}_{k+1} - x \mid \nabla h(y_k) - z_k\rangle \\ &\leq \|z_k - \nabla h(y_k)\| \|x_{k+1} - \hat{x}_{k+1}\| + \langle \hat{x}_{k+1} - x \mid \nabla h(y_k) - z_k\rangle \\ &\leq \tau_k \|z_k - \nabla h(y_k)\|^2 + \langle \hat{x}_{k+1} - x \mid \nabla h(y_k) - z_k\rangle \,. \end{aligned} \tag{3.30}$$

In the same way, we also have

$$\alpha_{4,k} \leq \sigma_k \|d_k - \nabla \ell^\star(u_k)\|^2 + \langle \nabla \ell^\star(u_k) - d_k \mid \hat{v}_{k+1} - v\rangle \,. \tag{3.31}$$

Adding (3.31) and (3.30), we obtain following definition (3.15)

$$\alpha_{3,k} + \alpha_{4,k} \leq \|\mathbf{r}_k - \mathbf{R}_k\|^2_{\boldsymbol{U}_k^{-1}} + \langle \hat{\mathsf{x}}_{k+1} - \mathsf{x} \mid \mathbf{R}_k - \mathbf{r}_k\rangle \,. \tag{3.32}$$

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

$$\begin{aligned}\frac{\mu}{2}\|x_{k+1} - y_k\|^2 &= \frac{\mu}{2}\|x_{k+1} - x_k - (x_k - x_{k-1})\|^2 \\ &= \frac{\mu}{2}\|x_{k+1} - x_k\|^2 + \frac{\mu}{2}\|x_k - x_{k-1}\|^2 - \mu \langle x_{k+1} - x_k \mid x_k - x_{k-1}\rangle \,, \end{aligned} \tag{3.33}$$

and

$$\begin{aligned}\frac{\nu}{2}\|v_{k+1} - u_k\|^2 &= \frac{\nu}{2}\|v_{k+1} - v_k - (v_k - v_{k-1})\|^2 \\ &= \frac{\nu}{2}\|v_{k+1} - v_k\|^2 + \frac{\nu}{2}\|v_k - v_{k-1}\|^2 - \nu \langle v_{k+1} - v_k \mid v_k - v_{k-1}\rangle \,. \end{aligned} \tag{3.34}$$

Adding (3.33) and (3.34), we obtain, since per (3.17), $\boldsymbol{D} = \operatorname{diag}(\mu \operatorname{Id}, \nu \operatorname{Id})$, the following expression for $\alpha_{0,k}$

$$\alpha_{0,k} = \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\boldsymbol{D}} + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{\boldsymbol{D}} - \langle \mathsf{x}_{k+1} - \mathsf{x}_k \mid \boldsymbol{D}(\mathsf{x}_k - \mathsf{x}_{k-1})\rangle \,. \tag{3.35}$$

Therefore, adding (3.35) and (3.28), we get

$$\begin{aligned}\alpha_{0,k} + \alpha_{1,k} + \alpha_{2,k} = {}& \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\boldsymbol{D}} + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{\boldsymbol{D}} - \langle \mathsf{x}_{k+1} - \mathsf{x}_k \mid (\boldsymbol{D} - \boldsymbol{L})(\mathsf{x}_k - \mathsf{x}_{k-1})\rangle \\ &+ \boldsymbol{b}_{k+1}(\mathsf{x}) - \boldsymbol{b}_k(\mathsf{x}). \end{aligned} \tag{3.36}$$

We have

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

Following (3.17), since $\boldsymbol{D}^\star = \boldsymbol{D}$ and $\boldsymbol{L}^\star = -\boldsymbol{L}$, we have

$$\begin{aligned}\frac{1}{2}\|\boldsymbol{D}^{-1}(\boldsymbol{D} - \boldsymbol{L})^\star(\mathsf{x}_{k+1} - \mathsf{x}_k)\|^2_{\boldsymbol{D}} &= \frac{1}{2}\left\langle \boldsymbol{D}^{-1}(\boldsymbol{D} - \boldsymbol{L})^\star(\mathsf{x}_{k+1} - \mathsf{x}_k) \mid (\boldsymbol{D} - \boldsymbol{L})^\star(\mathsf{x}_{k+1} - \mathsf{x}_k)\right\rangle \\ &= \frac{1}{2}\left\langle (\boldsymbol{D} + \boldsymbol{L}^\star)(\operatorname{Id} + \boldsymbol{D}^{-1}\boldsymbol{L})(\mathsf{x}_{k+1} - \mathsf{x}_k) \mid \mathsf{x}_{k+1} - \mathsf{x}_k\right\rangle \\ &= \frac{1}{2}\left\langle (\boldsymbol{D} + \boldsymbol{L}^\star\boldsymbol{D}^{-1}\boldsymbol{L})(\mathsf{x}_{k+1} - \mathsf{x}_k) \mid \mathsf{x}_{k+1} - \mathsf{x}_k\right\rangle \\ &= \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\boldsymbol{D} + \boldsymbol{L}^\star\boldsymbol{D}^{-1}\boldsymbol{L}}, \end{aligned}$$

which implies that

$$\langle \mathsf{x}_{k+1}-\mathsf{x}_k \mid (\boldsymbol{D}-\boldsymbol{L})(\mathsf{x}_k-\mathsf{x}_{k-1})\rangle \le \frac{1}{2}\|\mathsf{x}_k-\mathsf{x}_{k-1}\|^2_{\boldsymbol{D}} + \frac{1}{2}\|\mathsf{x}_{k+1}-\mathsf{x}_k\|^2_{\boldsymbol{D}+\boldsymbol{L}^\star\boldsymbol{D}^{-1}\boldsymbol{L}}. \tag{3.37}$$

Hence, by using (2.4), the expression (3.36) becomes

$$\alpha_{0,k}+\alpha_{1,k}+\alpha_{2,k} = \frac{1}{2}\|\mathsf{x}_{k+1}-\mathsf{x}_k\|^2_{2\boldsymbol{D}+\boldsymbol{L}^\star\boldsymbol{D}^{-1}\boldsymbol{L}} + \frac{1}{2}\|\mathsf{x}_k-\mathsf{x}_{k-1}\|^2_{2\boldsymbol{D}} + \boldsymbol{b}_{k+1}(\mathsf{x}) - \boldsymbol{b}_k(\mathsf{x}). \tag{3.38}$$

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

$$\alpha_{5,k}+\alpha_{6,k} = \frac{1}{2}\|\mathsf{x}_k-\mathsf{x}\|^2_{\boldsymbol{U}_k} - \frac{1}{2}\|\mathsf{x}_{k+1}-\mathsf{x}\|^2_{\boldsymbol{U}_k} - \frac{1}{2}\|\mathsf{x}_k-\mathsf{x}_{k+1}\|^2_{\boldsymbol{U}_k}. \tag{3.39}$$

Therefore, by combining (3.38), (3.39), (3.32) into (3.24), we obtain

$$\begin{aligned} K(x_{k+1},v)-K(x,v_{k+1}) &\le \sum_{i=0}^{6}\alpha_{i,k} \\ &\le \frac{1}{2}\|\mathsf{x}_k-\mathsf{x}\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_k-\mathsf{x}_{k-1}\|^2_{2\boldsymbol{D}} - \boldsymbol{b}_k(\mathsf{x}) \\ &\quad - \Big(\frac{1}{2}\|\mathsf{x}_{k+1}-\mathsf{x}\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_{k+1}-\mathsf{x}_k\|^2_{2\boldsymbol{D}} - \boldsymbol{b}_{k+1}(\mathsf{x})\Big) \\ &\quad + \frac{1}{2}\|\mathsf{x}_{k+1}-\mathsf{x}_k\|^2_{4\boldsymbol{D}+\boldsymbol{L}^\star\boldsymbol{D}^{-1}\boldsymbol{L}} - \frac{1}{2}\|\mathsf{x}_{k+1}-\mathsf{x}_k\|^2_{\boldsymbol{U}_k} \\ &\quad + \|\mathbf{r}_k-\mathbf{R}_k\|^2_{\boldsymbol{U}_k^{-1}} + \langle \hat{\mathsf{x}}_{k+1}-\mathsf{x} \mid \mathbf{R}_k-\mathbf{r}_k\rangle\,. \end{aligned} \tag{3.40}$$

Hence, the proof is completed. □

**Remark 3.6** Suppose that $f$ and $g^\star$ are strongly convex functions with constants $\theta_1$ and $\theta_2$, respectively. Then, using the same notations as Lemma 3.5, we have

$$\begin{aligned} K(x_{k+1},v)-K(x,v_{k+1}) + \frac{\min\{\theta_1\tau_k,\theta_2\sigma_k\}}{2}\|\mathsf{x}_{k+1}-\mathsf{x}\|^2_{\boldsymbol{U}_k} &\\ \le \frac{1}{2}\|\mathsf{x}_k-\mathsf{x}\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_k-\mathsf{x}_{k-1}\|^2_{2\boldsymbol{D}} - \boldsymbol{b}_k(\mathsf{x}) &\\ - \Big(\frac{1}{2}\|\mathsf{x}_{k+1}-\mathsf{x}\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_{k+1}-\mathsf{x}_k\|^2_{2\boldsymbol{D}} - \boldsymbol{b}_{k+1}(\mathsf{x})\Big) &\\ + \frac{1}{2}\|\mathsf{x}_{k+1}-\mathsf{x}_k\|^2_{4\boldsymbol{D}+\boldsymbol{L}^\star\boldsymbol{D}^{-1}\boldsymbol{L}} - \frac{1}{2}\|\mathsf{x}_{k+1}-\mathsf{x}_k\|^2_{\boldsymbol{U}_k} &\\ + \|\mathbf{r}_k-\mathbf{R}_k\|^2_{\boldsymbol{U}_k^{-1}} + \langle \hat{\mathsf{x}}_{k+1}-\mathsf{x} \mid \mathbf{R}_k-\mathbf{r}_k\rangle\,. & \end{aligned} \tag{3.41}$$

Moreover, by replacing $\mu_0 = \|\boldsymbol{D}-\boldsymbol{L}\|$ in (3.37), it follows

$$\langle \mathsf{x}_{k+1}-\mathsf{x}_k \mid (\boldsymbol{D}-\boldsymbol{L})(\mathsf{x}_k-\mathsf{x}_{k-1})\rangle \le \frac{1}{2}\|\mathsf{x}_k-\mathsf{x}_{k-1}\|^2_{\mu_0\,\mathrm{Id}} + \frac{1}{2}\|\mathsf{x}_{k+1}-\mathsf{x}_k\|^2_{\mu_0\,\mathrm{Id}}. \tag{3.42}$$

Then, using the latter inequality, the expression (3.41) becomes

$$\begin{aligned} K(x_{k+1}, v) - K(x, v_{k+1}) + \frac{\min\{\theta_1\tau_k, \theta_2\sigma_k\}}{2}\|\mathsf{x}_{k+1} - \mathsf{x}\|^2_{\boldsymbol{U}_k} \\ \leq \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}\|^2_{\boldsymbol{U}_k} - \boldsymbol{b}_k(\mathsf{x}) + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{\boldsymbol{D}+\mu_0 \,\mathrm{Id}} \\ - \left(\frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\boldsymbol{D}+\mu_0 \,\mathrm{Id}} - \boldsymbol{b}_{k+1}(\mathsf{x})\right) \\ + \|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\boldsymbol{D}+\mu_0 \,\mathrm{Id}} - \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\boldsymbol{U}_k} \\ + \|\mathbf{r}_k - \mathbf{R}_k\|^2_{\boldsymbol{U}_k^{-1}} + \langle \hat{\mathsf{x}}_{k+1} - \mathsf{x} \mid \mathbf{R}_k - \mathbf{r}_k \rangle . \end{aligned} \tag{3.43}$$

**Remark 3.7** When $p = 1 = q$, Lemma 3.3 recovers the one provided in [35], and Lemma 3.5 is similar to [24, Lemma 3.5] where $\mu_0$ is replaced by $\|\boldsymbol{D}\| + \|\boldsymbol{L}\|$.

Next, we extend Lemma 3.5 to upper bound the smoothed gap $G_{\beta_k}$ (2.10) as detailed in Definition 2.1.

**Lemma 3.8** *Let $(\beta_k)_{k\in\mathbb{N}}$ be a sequence in $]0, +\infty[$. Under the same setting as of Lemma 3.5, define*

$$\mathbf{S}_k = \begin{pmatrix} 2\mu + \frac{4\tau_k}{\beta_k}L^\star L & 0 \\ 0 & 2\nu + \frac{4\sigma_k}{\beta_k}LL^\star \end{pmatrix} \quad \textit{and}$$

$$\mathbf{T}_{k+1} = \begin{pmatrix} 2\mu + \frac{4}{\tau_k\beta_k} + 4(\frac{\tau_k}{\beta_k} + \frac{1}{\nu})L^\star L & 0 \\ 0 & 2\nu + \frac{4}{\beta_k\sigma_k} + 4(\frac{\sigma_k}{\beta_k} + \frac{1}{\mu})LL^\star \end{pmatrix}. \tag{3.44}$$

*Then, for every $k \in \mathbb{N}$, the smoothed primal-dual gap $G_{\beta_k}$ centered at $\mathsf{x}^\dagger \in \mathcal{S}$ is bounded by*

$$\begin{aligned} G_{\beta_k}(\mathsf{x}_{k+1}; \mathsf{x}^\dagger) \leq \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{\mathbf{S}_k} - \boldsymbol{b}_k(\mathsf{x}^\dagger) \\ - \left(\frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\mathbf{S}_{k+1}} - \boldsymbol{b}_{k+1}(\mathsf{x}^\dagger)\right) \\ + \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\mathbf{S}_{k+1}+\mathbf{T}_{k+1}} - \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\boldsymbol{U}_k} \\ + (1 + 2/\beta_k)\|\mathbf{r}_k - \mathbf{R}_k\|^2_{\boldsymbol{U}_k^{-1}} + \left\langle \hat{\mathsf{x}}_{k+1} - \mathsf{x}^\dagger \mid \mathbf{R}_k - \mathbf{r}_k \right\rangle . \end{aligned} \tag{3.45}$$

*Proof.* We fist observe that we can rewrite (3.44) as

$$\begin{cases} \mathbf{S}_k = 2\boldsymbol{D} + 4\beta_k^{-1}\boldsymbol{L}^\star\boldsymbol{U}_k^{-1}\boldsymbol{L} \\ \mathbf{T}_{k+1} = \mathbf{S}_k + 4\beta_k^{-1}\boldsymbol{U}_k + 4\boldsymbol{L}^\star\boldsymbol{D}^{-1}\boldsymbol{L}. \end{cases} \tag{3.46}$$

We have the following estimations

$$\begin{aligned} \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}\|^2_{\boldsymbol{U}_k} - \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}\|^2_{\boldsymbol{U}_k} = \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} - \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \left\langle \boldsymbol{U}_k(\mathsf{x}_k - \mathsf{x}_{k+1}) \mid \mathsf{x}^\dagger - \mathsf{x} \right\rangle \\ \leq \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} - \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{2}{\beta_k}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\boldsymbol{U}_k} + \frac{\beta_k}{8}\|\mathsf{x} - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k}, \end{aligned} \tag{3.47}$$

and

$$\begin{aligned}\boldsymbol{b}_k(\mathsf{x}) - \boldsymbol{b}_{k+1}(\mathsf{x}) &= \boldsymbol{b}_k(\mathsf{x}^\dagger) - \boldsymbol{b}_{k+1}(\mathsf{x}^\dagger) + \left\langle \mathbf{L}^\star(\mathsf{x}_k - \mathsf{x}_{k-1}) \mid \mathsf{x}^\dagger - \mathsf{x} \right\rangle - \left\langle \mathbf{L}^\star(\mathsf{x}_{k+1} - \mathsf{x}_k) \mid \mathsf{x}^\dagger - \mathsf{x} \right\rangle \\ &= \boldsymbol{b}_k(\mathsf{x}^\dagger) - \boldsymbol{b}_{k+1}(\mathsf{x}^\dagger) + \frac{2}{\beta_k}\|\boldsymbol{U}_k^{-1}\boldsymbol{L}^\star(\mathsf{x}_k - \mathsf{x}_{k-1})\|^2_{\boldsymbol{U}_k} + \frac{2}{\beta_k}\|\boldsymbol{U}_k^{-1}\boldsymbol{L}^\star(\mathsf{x}_{k+1} - \mathsf{x}_k)\|^2 \\ &\quad + \frac{\beta_k}{4}\|\mathsf{x} - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} \\ &= \boldsymbol{b}_k(\mathsf{x}^\dagger) - \boldsymbol{b}_{k+1}(\mathsf{x}^\dagger) + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{4\beta_k^{-1}\boldsymbol{L}^\star\boldsymbol{U}_k^{-1}\boldsymbol{L}} + \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{4\beta_k^{-1}\boldsymbol{L}^\star\boldsymbol{U}_k^{-1}\boldsymbol{L}} \\ &\quad + \frac{\beta_k}{4}\|\mathsf{x} - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k}, \end{aligned} \tag{3.48}$$

and

$$\begin{aligned}\langle \hat{\mathsf{x}}_{k+1} - \mathsf{x} \mid \mathbf{R}_k - \mathbf{r}_k\rangle &= \left\langle \hat{\mathsf{x}}_{k+1} - \mathsf{x}^\dagger \mid \mathbf{R}_k - \mathbf{r}_k \right\rangle + \left\langle \mathsf{x}^\dagger - \mathsf{x} \mid \mathbf{R}_k - \mathbf{r}_k \right\rangle \\ &\le \left\langle \hat{\mathsf{x}}_{k+1} - \mathsf{x}^\dagger \mid \mathbf{R}_k - \mathbf{r}_k \right\rangle + \frac{2}{\beta_k}\|\boldsymbol{U}_k^{-1}(\mathbf{r}_k - \mathbf{R}_k)\|^2_{\boldsymbol{U}_k} + \frac{\beta_k}{8}\|\mathsf{x} - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} \\ &\le \left\langle \hat{\mathsf{x}}_{k+1} - \mathsf{x}^\dagger \mid \mathbf{R}_k - \mathbf{r}_k \right\rangle + \frac{2}{\beta_k}\|\mathbf{r}_k - \mathbf{R}_k\|^2_{\boldsymbol{U}_k^{-1}} + \frac{\beta_k}{8}\|\mathsf{x} - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k}. \end{aligned} \tag{3.49}$$

Therefore, (3.19) becomes

$$\begin{aligned} &K(x_{k+1}, v) - K(x, v_{k+1}) - \frac{\beta_k}{2}\|\mathsf{x} - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} \\ &\quad \le \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{2\boldsymbol{D}+4\beta_k^{-1}\boldsymbol{L}^\star\boldsymbol{U}_k^{-1}\boldsymbol{L}} - \boldsymbol{b}_k(\mathsf{x}^\dagger) \\ &\quad\quad - \left(\frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{2\boldsymbol{D}} - \boldsymbol{b}_{k+1}(\mathsf{x}^\dagger)\right) \\ &\quad\quad + \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{4\boldsymbol{D}+\boldsymbol{L}^\star\boldsymbol{D}^{-1}\boldsymbol{L}+4\beta_k^{-1}\boldsymbol{U}_k+4\beta_k^{-1}\boldsymbol{L}^\star\boldsymbol{U}_k^{-1}\boldsymbol{L}} - \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\boldsymbol{U}_k} \\ &\quad\quad + (1 + 2/\beta_k)\|\mathbf{r}_k - \mathbf{R}_k\|^2_{\boldsymbol{U}_k^{-1}} + \left\langle \hat{\mathsf{x}}_{k+1} - \mathsf{x}^\dagger \mid \mathbf{R}_k - \mathbf{r}_k \right\rangle. \end{aligned} \tag{3.50}$$

Taking the supremun over $\mathsf{x} \in \mathrm{dom}(f) \times \mathrm{dom}(g^\star)$ and using (3.46), we obtain (3.45). □

## 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 smoothed primal-dual gap function value to 0 rely on the following results that establish the bound of the variance $\mathsf{E}_k[\|\mathbf{r}_k - \mathbf{R}_k\|^2_{\boldsymbol{U}_k^{-1}}]$ as well as the descent property of a suitable Lyapunov function.

In view of (3.19) in Lemma 3.5, we need to estimate the variance $\mathsf{E}_k[\|\mathbf{r}_k - \mathbf{R}_k\|^2_{\boldsymbol{U}_k^{-1}}]$ by means of the difference $K(x_k, v^\dagger) - K(x^\dagger, v_k)$ for some $\mathsf{x}^\dagger = (x^\dagger, v^\dagger) \in \mathcal{S}$, where $\mathcal{S}$ is defined by (2.9). For this purpose, we introduce the following definition.

**Definition 3.9** *For every $k \in \mathbb{N}$, let $\mathsf{x}_k = (x_k, v_k)$ and $\mathsf{w}_k = (w_{1,k}, w_{2,k})$. For $\mathsf{x}^\dagger = (x^\dagger, v^\dagger) \in \mathcal{S}$, the function $\Theta(\mathsf{x}_k)$ is defined by the difference*

$$\begin{aligned}
&\Theta(\mathsf{x}_k) := \Theta(x_k, v_k) = K(x_k, v^\dagger) - K(x^\dagger, v_k),\\
&\qquad \textit{where the Lagrangian function } K \textit{ is defined by } (2.11);\\
&\textit{and the (total) MSE associated to the auxiliary variables } (w_{1,k}, w_{2,k}) \textit{ against } (x^\dagger, v^\dagger)\\
&Q(\mathsf{w}_k) := Q(w_{1,k}, w_{2,k}) = \Xi_h(w_{1,k}, x^\dagger) + \Xi_{\ell^\star}(w_{2,k}, v^\dagger),\\
&\qquad \textit{where } \Xi_h \textit{ and } \Xi_{\ell^\star} \textit{ are defined by } (3.6) \textit{ and } (3.7)\textit{, respectively.}
\end{aligned} \tag{3.51}$$

Using this definition, we can bound the variance $\mathsf{E}_k[\|\mathbf{r}_k - \mathbf{R}_k\|^2_{\boldsymbol{U}_k^{-1}}]$ wherewith the following Lemma.

**Lemma 3.10** *Let $\mathsf{x}^\dagger \in \mathcal{S}$ and define*

$$\boldsymbol{P} = \operatorname{diag}((1-p)\operatorname{Id}, (1-q)\operatorname{Id}) := \begin{pmatrix} (1-p)\operatorname{Id} & 0 \\ 0 & (1-q)\operatorname{Id} \end{pmatrix}.$$

*Then, following Definition 3.9,*

$$\mathsf{E}_k\left[\|\mathbf{r}_k - \mathbf{R}_k\|^2_{\boldsymbol{U}_k^{-1}}\right] \le 2\gamma_k(1-\underline{p})\big(Q(\mathsf{w}_{k-1}) + 4\overline{\mu}\Theta(\mathsf{x}_k)\big) + \|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{4\boldsymbol{D}^2\boldsymbol{P}\boldsymbol{U}_k^{-1}}, \tag{3.52}$$

*where we set $\underline{p} = \min\{p, q\}$ and $\overline{\mu} = \max\{\mu, \nu\}$.*

*Proof.* Using Lemma 3.3, we first have

$$\begin{aligned}
\mathsf{E}_k\left[\|\mathbf{r}_k - \mathbf{R}_k\|^2_{\boldsymbol{U}_k^{-1}}\right] &= \tau_k \mathsf{E}_k[\|z_k - \nabla h(y_k)\|^2] + \sigma_k \mathsf{E}_k[\|d_k - \nabla \ell^\star(u_k)\|^2]\\
&\le 2\tau_k(1-p)\Big(\Xi_h(w_{1,k-1}, x^\dagger) + \Xi_h(y_k, x^\dagger)\Big)\\
&\quad + 2\sigma_k(1-q)\Big(\Xi_{\ell^\star}(w_{2,k-1}, v^\dagger) + \Xi_{\ell^\star}(u_k, v^\dagger)\Big).
\end{aligned} \tag{3.53}$$

By definition of $\Xi_h$ and $\Xi_{\ell^\star}$, using the triangle inequality, the Lipschitz continuity of $\nabla h_i$ and $\nabla \ell_j^\star$ with respect to the constant $\mu$ and $\nu$, respectively, as well as (3.2), we derive the following inequalities

$$\begin{aligned}
\Xi_h(y_k, x^\dagger) &= \frac{1}{n_p}\sum_{i=1}^{n_p} \|\nabla h_i(y_k) - \nabla h_i(x^\dagger)\|^2\\
&\le \frac{2}{n_p}\sum_{i=1}^{n_p} \|\nabla h_i(y_k) - \nabla h_i(x_k)\|^2 + \frac{2}{n_p}\sum_{i=1}^{n_p} \|\nabla h_i(x_k) - \nabla h_i(x^\dagger)\|^2\\
&\le \frac{2}{n_p}\mu^2 \sum_{i=1}^{n_p} \|y_k - x_k\|^2 + 2\Xi_h(x_k, x^\dagger))\\
&\le 2\big(\mu^2\|x_k - x_{k-1}\|^2 + \Xi_h(x_k, x^\dagger)\big),
\end{aligned} \tag{3.54}$$

and similarly

$$\Xi_{\ell^\star}(u_k, v^\dagger) \leq 2\big(\nu^2\|v_k - v_{k-1}\|^2 + \Xi_{\ell^\star}(v_k, v^\dagger)\big). \tag{3.55}$$

Inserting (3.54) and (3.55) to (3.53), we can further estimate (3.53) as

$$\begin{aligned}
\mathsf{E}_k\left[\|\mathbf{r}_k - \mathbf{R}_k\|^2_{\boldsymbol{U}_k^{-1}}\right] &\leq 2\tau_k(1-p)\Big(\Xi_h(w_{1,k-1}, x^\dagger) + 2\mu^2\|x_k - x_{k-1}\|^2 + 2\Xi_h(x_k, x^\dagger)\Big) \\
&\quad + 2\sigma_k(1-q)\Big(\Xi_{\ell^\star}(w_{2,k-1}, v^\dagger) + 2\nu^2\|v_k - v_{k-1}\|^2 + 2\Xi_{\ell^*}(v_k, v^\dagger)\Big) \\
&\leq 2\gamma_k(1-\underline{p})\Big(\big(\Xi_h(w_{1,k-1}, x^\dagger) + \Xi_{\ell^\star}(w_{2,k-1}, v^\dagger)\big) + 2\big(\Xi_h(x_k, x^\dagger) + \Xi_{\ell^\star}(v_k, v^\dagger)\big)\Big) \\
&\quad + \|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{4\boldsymbol{D}^2\boldsymbol{P}\boldsymbol{U}_k^{-1}} \\
&= 2\gamma_k(1-\underline{p})\Big(Q(\mathsf{w}_{k-1}) + 2\big(\Xi_h(x_k, x^\dagger) + \Xi_{\ell^\star}(v_k, v^\dagger)\big)\Big) \\
&\quad + \|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{4\boldsymbol{D}^2\boldsymbol{P}\boldsymbol{U}_k^{-1}}.
\end{aligned} \tag{3.56}$$

The second term in (3.56) is bounded by $\Theta(\mathsf{x}_k)$, as indicated by Lemma 3.3 in [24],

$$\Xi_h(x_k, x^\dagger) + \Xi_{\ell^\star}(v_k, v^\dagger) \leq 2\overline{\mu}\Theta(\mathsf{x}_k). \tag{3.57}$$

Therefore, the variance is bounded by

$$\mathsf{E}_k\left[\|\mathbf{r}_k - \mathbf{R}_k\|^2_{\boldsymbol{U}_k^{-1}}\right] \leq 2\gamma_k(1-\underline{p})\big(Q(\mathsf{w}_{k-1}) + 4\overline{\mu}\Theta(\mathsf{x}_k)\big) + \|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{4\boldsymbol{D}^2\boldsymbol{P}\boldsymbol{U}_k^{-1}}, \tag{3.58}$$

which proves (3.52). □

Next, we introduce the following Lyapunov function

**Definition 3.11** *For every $k \in \mathbb{N}$, the Lyapunov function $\mathcal{L}_k(\mathsf{x}^\dagger)$ is defined by*

$$\mathcal{L}_k(\mathsf{x}^\dagger) = \Theta(\mathsf{x}_k) + Q(\mathsf{w}_{k-1}) + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} - \boldsymbol{b}_k(\mathsf{x}^\dagger) + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{\boldsymbol{V}_k}, \tag{3.59}$$

*where $\Theta(\mathsf{x}_k)$ and $Q(\mathsf{w}_{k-1})$ are defined following (3.51) and $\boldsymbol{b}_k(\mathsf{x}^\dagger)$ by (3.18).*

The following Theorem proves that the Lyapunov function verifies a descent property based on the bound of the variance $\mathsf{E}_k[\|\mathbf{r}_k - \mathbf{R}_k\|^2_{\boldsymbol{U}_k^{-1}}]$ as obtained in Lemma 3.10.

**Theorem 3.12** *Let $\mathsf{x}^\dagger \in \mathcal{S}$, $\overline{p} = \max\{p, q\}$, $\hat{\boldsymbol{V}} = \operatorname{diag}\big((2\mu + 4p\mu^2)\operatorname{Id}, (2\nu + 4q\nu^2)\operatorname{Id}\big)$ and $k \in \mathbb{N}$. Define $\gamma_k = \max\{\sigma_k, \tau_k\}$ and*

$$\begin{cases}
\Lambda_{k+1} = \begin{pmatrix} (\frac{1}{\tau_k} - 2\mu - 8(1-p)\tau_{k+1}\mu^2)\operatorname{Id} - \frac{L^\star L}{\nu} & 0 \\ 0 & (\frac{1}{\sigma_k} - 2\nu - 8(1-q)\sigma_{k+1}\nu^2)\operatorname{Id} - \frac{LL^\star}{\mu} \end{pmatrix} - \hat{\boldsymbol{V}} \\
\\
\boldsymbol{V}_k = \hat{\boldsymbol{V}} + \begin{pmatrix} 8(1-p)\mu^2\tau_k\operatorname{Id} & 0 \\ 0 & 8(1-q)\nu^2\sigma_k\operatorname{Id} \end{pmatrix}.
\end{cases} \tag{3.60}$$

*Set* $\epsilon \in \left]0, \underline{p}\right[$, *where* $\underline{p} = \min\{p, q\}$. *Let* $(\eta_k)_{k\in\mathbb{N}}$ *be a sequence in* $\ell^1_+(\mathbb{N})$. *Suppose that the following conditions are verified.*

$$\begin{cases} 4\overline{\mu}\big(2\gamma_k(1-\underline{p}) + \overline{p}\big) + \epsilon \leq 1 + \eta_k, \ \textit{with } \overline{\mu} = \max\{\mu, \nu\} \\ (2\gamma_k + 1)(1-\underline{p}) + \epsilon \leq 1 + \eta_k, \\ \boldsymbol{U}_{k-1} \succeq \boldsymbol{U}_k \succeq \epsilon \operatorname{Id} + \|L\|^2 \hat{\boldsymbol{V}}^{-1}, \\ \Lambda_k \succeq \epsilon \operatorname{Id}. \end{cases} \tag{3.61}$$

*Then, for all* $k$*, the following descent property is verified by the Lyapunov function* $\mathcal{L}_k(\mathsf{x}^\dagger)$ (3.59)*:*

$$\mathsf{E}_k\left[\mathcal{L}_{k+1}(\mathsf{x}^\dagger)\right] + \frac{1}{2}\mathsf{E}_k\left[\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\Lambda_{k+1}}\right] \leq (1+\eta_k)\mathcal{L}_k(\mathsf{x}^\dagger) - \epsilon\big(\Theta(\mathsf{x}_k) + Q(\mathsf{w}_{k-1})\big). \tag{3.62}$$

*Proof.* Using the same notations as defined for Lemma 3.5 and Lemma 3.10, and the expression (3.19) with $\mathsf{x} = \mathsf{x}^\dagger$, we obtain

$$\begin{aligned} \Theta(\mathsf{x}_{k+1}) \leq\ & \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{2\boldsymbol{D}} - \boldsymbol{b}_k(\mathsf{x}^\dagger) \\ & - \left(\frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{2\boldsymbol{D}} - \boldsymbol{b}_{k+1}(\mathsf{x}^\dagger)\right) \\ & + \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{4\boldsymbol{D}+\boldsymbol{L}^\star\boldsymbol{D}^{-1}\boldsymbol{L}} - \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\boldsymbol{U}_k} \\ & + \|\mathbf{r}_k - \mathbf{R}_k\|^2_{\boldsymbol{U}_k^{-1}} + \left\langle \hat{\mathsf{x}}_{k+1} - \mathsf{x}^\dagger \mid \mathbf{R}_k - \mathbf{r}_k \right\rangle. \end{aligned} \tag{3.63}$$

From Lemma 3.3, we deduce

$$\mathsf{E}_k\left[\left\langle \hat{\mathsf{x}}_{k+1} - \mathsf{x}^\dagger \mid \mathbf{R}_k - \mathbf{r}_k \right\rangle\right] = 0. \tag{3.64}$$

Now, by taking the expectation $\mathsf{E}_k$ on both sides of (3.63) and invoking (3.52) in Lemma 3.10 as well as (2.4), we obtain

$$\begin{aligned} \mathsf{E}_k\left[\Theta(\mathsf{x}_{k+1})\right] \leq\ & 8\gamma_k\overline{\mu}(1-\underline{p})\Theta(\mathsf{x}_k) + 2\gamma_k(1-\underline{p})Q(\mathsf{w}_{k-1}) \\ & + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{2\boldsymbol{D}} + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{8\boldsymbol{D}^2P\boldsymbol{U}_k^{-1}} - \boldsymbol{b}_k(\mathsf{x}^\dagger) \\ & - \mathsf{E}_k\left[\frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{2\boldsymbol{D}} + \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{8\boldsymbol{D}^2P\boldsymbol{U}_{k+1}^{-1}} - \boldsymbol{b}_{k+1}(\mathsf{x}^\dagger)\right] \\ & + \mathsf{E}_k\left[\frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{4\boldsymbol{D}+\boldsymbol{L}^\star\boldsymbol{D}^{-1}\boldsymbol{L}+8\boldsymbol{D}^2P\boldsymbol{U}_{k+1}^{-1}} - \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\boldsymbol{U}_k}\right]. \end{aligned} \tag{3.65}$$

Since following (3.51), $Q(\mathsf{w}_k) = \Xi_h(w_{1,k}, x^\dagger) + \Xi_{\ell^\star}(w_{2,k}, v^\dagger)$, its expectation $\mathsf{E}_k[Q(\mathsf{w}_k)]$ can be upper bounded by using inequalities (3.54), (3.55) and (3.57),

$$\begin{aligned} \mathsf{E}_k\left[Q(\mathsf{w}_k)\right] &= \mathsf{E}_k\left[\Xi_h(w_{1,k}, x^\dagger) + \Xi_{\ell^\star}(w_{2,k}, v^\dagger)\right] \\ &= (1-p)\Xi_h(w_{1,k-1}, x^\dagger) + (1-q)\Xi_{\ell^\star}(w_{2,k-1}, v^\dagger) + p\Xi_h(y_k, x^\dagger) + q\Xi_{\ell^\star}(u_k, v^\dagger) \\ &\leq (1-\underline{p})Q(\mathsf{w}_{k-1}) + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{4\boldsymbol{D}^2(\operatorname{Id}-P)} + 4\overline{\mu}\,\overline{p}\,\Theta(\mathsf{x}_k). \end{aligned} \tag{3.66}$$

We note that the the definition of $\boldsymbol{V}_k$ and $\Lambda_{k+1}$ in (3.60) can be rewritten as

$$\begin{cases} \boldsymbol{V}_k = 2\boldsymbol{D} + 4\boldsymbol{D}^2(\mathrm{Id} - P) + 8\boldsymbol{D}^2 P\boldsymbol{U}_k^{-1} \\ \Lambda_{k+1} = \boldsymbol{U}_k - 2\boldsymbol{D} - \boldsymbol{V}_{k+1} - \boldsymbol{L}^*\boldsymbol{D}^{-1}\boldsymbol{L}. \end{cases} \tag{3.67}$$

Therefore, adding (3.65) to (3.66), we obtain[1] by using the first three conditions in (3.61)

$$\begin{aligned} \mathsf{E}_k\left[\Theta(\mathsf{x}_{k+1})\right] + \mathsf{E}_k\left[Q(\mathsf{w}_k)\right] &\leq 4\overline{\mu}\big(2\gamma_k(1-\underline{p}) + \overline{p}\big)\Theta(\mathsf{x}_k) + (2\gamma_k + 1)(1-\underline{p})Q(\mathsf{w}_{k-1}) \\ &\quad + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{\boldsymbol{V}_k} - \boldsymbol{b}_k(\mathsf{x}^\dagger) \\ &\quad - \mathsf{E}_k\left[\frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\boldsymbol{V}_{k+1}} - \boldsymbol{b}_{k+1}(\mathsf{x}^\dagger)\right] \\ &\quad - \mathsf{E}_k\left[\frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\Lambda_{k+1}}\right] \\ &\leq (1 + \eta_k - \epsilon)\big(\Theta(\mathsf{x}_k) + Q(\mathsf{w}_{k-1})\big) \\ &\quad + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{\boldsymbol{V}_k} - \boldsymbol{b}_k(\mathsf{x}^\dagger) \\ &\quad - \mathsf{E}_k\left[\frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_{k+1}} + \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\boldsymbol{V}_{k+1}} - \boldsymbol{b}_{k+1}(\mathsf{x}^\dagger)\right] \\ &\quad - \mathsf{E}_k\left[\frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\Lambda_{k+1}}\right]. \end{aligned} \tag{3.68}$$

Now using the definition of $\boldsymbol{b}_k(\mathsf{x})$ (see (3.18)), we have

$$\begin{aligned} \boldsymbol{b}_k(\mathsf{x}^\dagger) &= \left\langle \mathsf{x}_k - \mathsf{x}_{k-1} \mid \boldsymbol{L}(\mathsf{x}_k - \mathsf{x}^\dagger)\right\rangle = \langle L(x_k - x) \mid v_k - v_{k-1}\rangle + \langle L(x_k - x_{k-1}) \mid v - v_k\rangle \\ &\leq \frac{1}{2}\left(\frac{\|L\|^2}{2\mu + 4p\mu^2}\|x_k - x^\dagger\|^2 + (2\mu + 4p\mu^2)\|x_k - x_{k-1}\|^2\right) \\ &\qquad + \frac{1}{2}\left(\frac{\|L\|^2}{2\nu + 4q\nu^2}\|v_k - v^\dagger\|^2 + (2\nu + 4q\nu^2)\|v_k - v_{k-1}\|^2\right). \\ &= \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}^\dagger\|^2_{\|L\|^2\hat{\boldsymbol{V}}^{-1}} + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{\hat{\boldsymbol{V}}}. \end{aligned} \tag{3.69}$$

The inequality (3.69) implies that

$$\begin{aligned} \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{\boldsymbol{V}_k} - \boldsymbol{b}_k(\mathsf{x}^\dagger) &\geq \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{\boldsymbol{V}_k - \hat{\boldsymbol{V}}} + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k - \|L\|^2\hat{\boldsymbol{V}}^{-1}} \\ &\geq 0, \end{aligned} \tag{3.70}$$

where the last inequality follows from $\boldsymbol{V}_k - \hat{\boldsymbol{V}} = 8P\boldsymbol{U}_k^{-1} \succeq 0$ and $\boldsymbol{U}_k - \|L\|^2\hat{\boldsymbol{V}}^{-1} \succeq \epsilon\,\mathrm{Id}$ in (3.61). Hence,

$$\begin{cases} \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{\boldsymbol{V}_k} - \boldsymbol{b}_k(\mathsf{x}^\dagger) \leq (1+\eta_k)\left(\frac{1}{2}\|\mathsf{x}_k - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{\boldsymbol{V}_k} - \boldsymbol{b}_k(\mathsf{x}^\dagger)\right) \\ \mathcal{L}_k(\mathsf{x}^\dagger) \geq \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k - \|L\|^2\,\mathrm{Id}/\underline{\mu}} \geq \frac{\epsilon}{2}\|\mathsf{x}_k - \mathsf{x}^\dagger\|^2. \end{cases} \tag{3.71}$$

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

Moreover, in terms of the Lyapunov function defined by (3.59), we can rewrite (3.68) as

$$\mathsf{E}_k\left[\mathcal{L}_{k+1}(\mathsf{x}^\dagger)\right] + \mathsf{E}_k\left[\frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\Lambda_{k+1}}\right] \leq (1+\eta_k)\mathcal{L}_k(\mathsf{x}^\dagger) - \epsilon\big(\Theta(\mathsf{x}_k) + Q(\mathsf{w}_{k-1})\big), \tag{3.72}$$

which proves (3.62). □

We specify the condition (3.61) in the following examples.

**Example 3.13** Assume the Lipschitz constants verify $\mu = \nu$; thus, $\overline{\mu} = \mu$. Set $\eta_k \equiv 0$. Then the conditions (3.61) are satisfied when the strictly positive stepsize sequence $(\tau_k, \sigma_k)_{k\in\mathbb{N}}$ verifies the following conditions

(i) $\gamma_k = \max\{\tau_k, \sigma_k\} \leq \min\left\{\dfrac{\underline{p} - \epsilon}{2(1-\underline{p})}, \dfrac{1 - \epsilon - 4\mu\,\overline{p}}{8\mu(1-\underline{p})}\right\}.$

(ii) $\tau_{k-1}^{-1} \geq \tau_k^{-1} \geq \epsilon + \dfrac{\|L\|^2}{2\mu(1+2p\mu)}$ and $\sigma_{k-1}^{-1} \geq \sigma_k^{-1} \geq \epsilon + \dfrac{\|L\|^2}{2\nu(1+2q\nu)}$

(iii) $\tau_k^{-1} \geq 4\mu + 4p\mu^2 + 4\mu^2(\underline{p} - \epsilon) + \dfrac{1}{\nu}\|L\|^2 + \epsilon$ and $\sigma_k^{-1} \geq 4\nu + 4q\nu^2 + 4\nu^2(\underline{p} - \epsilon) + \dfrac{1}{\mu}\|L\|^2 + \epsilon.$

**Example 3.14** Assume the stepsize $\tau_k = \sigma_k \equiv \gamma$. Set $\eta_k \equiv 0$ and $s = 2\mu + 4p\mu^2$. For simplicity, further assume $\mu = \nu$ and $p = q$, thus $\underline{p} = \overline{p} = p$. Then, $\boldsymbol{U}_k = \gamma^{-1}\,\mathrm{Id}$ and $\boldsymbol{D} = \mu\,\mathrm{Id}$. Then we can simplify the conditions in Example 3.13 as

$$0 < \gamma \leq \min\left\{\frac{p-\epsilon}{2(1-p)}, \frac{1-\epsilon-4p\mu}{8\mu(1-p)}, \frac{s}{\|L\|^2 + s\epsilon}, \frac{1}{[4\mu(1+4p\mu+\mu(p-\epsilon)) + \mu^{-1}\|L\|^2] + \epsilon}\right\}. \tag{3.73}$$

**Remark 3.15** Note that the first term $\frac{p-\epsilon}{2(1-p)}$ appearing in (3.73) depends only on $p$. Let $N$ be the epoch and $q = p = 1/N$. We can take $N$ large enough and $\epsilon = \frac{1}{2N}$ such that

$$\gamma = \frac{p-\epsilon}{2(1-p)} = \frac{1}{4(N-1)}, \tag{3.74}$$

which is much better than $\gamma = \dfrac{1}{4N(\overline{\mu} + \|L\|)}$ used in [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 $\mathsf{x}^\dagger \in \mathcal{S}$ (2.9) and the convergence of the sequence of $(\Theta(\mathsf{x}_k))_{k\in\mathbb{N}}$ to 0.

**Theorem 3.16** *Under the same setting as Theorem 3.12, the following hold for* $\mathsf{x}^\dagger = (x^\dagger, v^\dagger) \in \mathcal{S}$.

$$\begin{cases} \Theta(\mathsf{x}_k) = \Theta(x_k, v_k) = K(x_k, v^\dagger) - K(x^\dagger, v_k) \to 0 \textit{ almost surely;} \\ Q(\mathsf{w}_k) = Q(w_{k,1}, w_{k,2}) = \Xi_h(w_{1,k}, x^\dagger) + \Xi_{\ell^\star}(w_{2,k}, v^\dagger) \to 0 \textit{ almost surely.} \end{cases} \tag{3.75}$$

*Moreover, if the following conditions (the lower boundedness of the primal stepsize* $\tau_k$ *and dual stepsize* $\sigma_k$*) are verified*

$$\inf_{k\in\mathbb{N}} \tau_k \geq \epsilon \textit{ and } \inf_{k\in\mathbb{N}} \sigma_k \geq \epsilon, \tag{3.76}$$

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

*Proof.* Under the setting of Theorem 3.12, all the conditions stated in Lemma 2.6 are satisfied. Consequently, there exists a random variable defined as $\mathcal{L}_\infty(\mathsf{x}^\dagger)$ such that

$$\mathcal{L}_k(\mathsf{x}^\dagger) \to \mathcal{L}_\infty(\mathsf{x}^\dagger) \text{ a.s. as } k \to \infty, \tag{3.77}$$

and

$$\epsilon \sum_{k\in\mathbb{N}} \mathsf{E}_k[\frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2] \leq \sum_{k\in\mathbb{N}} \mathsf{E}_k[\frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\Lambda_k}] < +\infty \text{ a.s.}, \tag{3.78}$$

and

$$\sum_{k\in\mathbb{N}} \big(\Theta(\mathsf{x}_k) + Q(\mathsf{w}_{k-1})\big) < +\infty. \tag{3.79}$$

Hence, by [26, Corollary 2.6], we also obtain

$$\sum_{k\in\mathbb{N}} \|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2 < +\infty \tag{3.80}$$

$$\text{as well as } \mathsf{x}_{k+1} - \mathsf{x}_k \to 0, \text{ and } \Theta(\mathsf{x}_k) + Q(\mathsf{w}_{k-1}) \to 0 \text{ a.s.} \tag{3.81}$$

Therefore, (3.75) is proved.

For the second part, we derive from (3.52) and (3.81) that

$$\sum_{k\in\mathbb{N}} \mathsf{E}_k\left[\|\mathbf{r}_k - \mathbf{R}_k\|^2\right] < +\infty \text{ a.s..} \tag{3.82}$$

Using [26, Corollary 2.6] again, we get by expanding $\mathbf{r}_k$ and $\mathbf{R}_k$ following definition (3.15)

$$\|\mathbf{r}_k - \mathbf{R}_k\|^2 = \|z_k - \nabla h(y_k)\|^2 + \|d_k - \nabla \ell^\star(u_k)\|^2 \to 0 \text{ a.s.}, \tag{3.83}$$

and thus, by (3.29), that

$$\begin{cases} \|\hat{\mathsf{x}}_{k+1} - \mathsf{x}_{k+1}\|^2 \leq \tau_k^2\|z_k - \nabla h(y_k)\|^2 + \sigma_k^2\|t_k - \nabla\ell^\star(u_k)\|^2 \to 0 \text{ a.s} \\ \|\hat{\mathsf{x}}_{k+1} - \mathsf{x}_k\| \leq \|\hat{\mathsf{x}}_{k+1} - \mathsf{x}_{k+1}\| + \|\mathsf{x}_{k+1} - \mathsf{x}_k\| \to 0 \text{ a.s.} \end{cases} \tag{3.84}$$

Moreover, (3.77) implies that $(\mathcal{L}_k(\mathsf{x}^\dagger))_{k\in\mathbb{N}}$ is bounded a.s. Hence, by (3.71), $(\|\mathsf{x}_{k+1} - \mathsf{x}^\dagger\|_{\boldsymbol{U}_k})_{k\in\mathbb{N}}$ is also bounded a.s. In turn, using the definition of $\boldsymbol{b}_k(\mathsf{x}^\dagger)$ (3.18),

$$|\boldsymbol{b}_{k+1}(\mathsf{x}^\dagger)| \leq \|\boldsymbol{L}\|\|\mathsf{x}_{k+1} - \mathsf{x}_k\|\|\mathsf{x}_{k+1} - \mathsf{x}^\dagger\|_{\boldsymbol{U}_k} = \|L\|\|\mathsf{x}_{k+1} - \mathsf{x}_k\|\|\mathsf{x}_{k+1} - \mathsf{x}^\dagger\|_{\boldsymbol{U}_k} \to 0. \tag{3.85}$$

Next, we derive from (3.81), (3.85) and (3.77) that

$$\lim \mathcal{L}_{k+1}(\mathsf{x}^\dagger) = \lim \|\mathsf{x}_{k+1} - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} = \mathcal{L}_\infty(\mathsf{x}^\dagger) \text{ a.s.}, \tag{3.86}$$

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

$$\boldsymbol{U}_k^{-1}(\mathsf{x}_k - \hat{\mathsf{x}}_{k+1} - \boldsymbol{C}\mathsf{y}_k) \in \boldsymbol{M}\hat{\mathsf{x}}_{k+1}, \tag{3.87}$$

and (3.76), we obtain $\bar{\mathsf{x}} \in \operatorname{zer}(\boldsymbol{M} + \boldsymbol{C}) = \mathcal{S}$ a.s. Therefore, by [34, Proposition 2.5], $(\mathsf{x}_k)_{k\in\mathbb{N}}$ converges weakly a.s. to a point in $\mathcal{S}$. □

We next show that the convergence of the function $\Theta(\mathsf{x}_k)$ (3.51) to 0 and the strong convergence of $(\mathsf{x}_k)_{k\in\mathbb{N}}$ imply the convergence of the partial duality gap function defined by

$$G_{\mathcal{Z}}(x_k, v_k) = \sup_{\mathsf{x}\in\mathcal{Z}}(K(x_k, v) - K(x, v_k)), \tag{3.88}$$

where $\mathcal{Z}$ is a bounded set of $\mathcal{H}\times\mathcal{G}$.

**Corollary 3.17** *Suppose that the conditions of Theorem 3.16 are satisfied and* $\dim(\mathcal{H}\times\mathcal{G}) < +\infty$*. Then, for any bounded set* $\mathcal{Z}$ *of* $\mathcal{H}\times\mathcal{G}$*, which have nonempty intersection with the set of solutions* $\mathcal{S}$*, the partial duality gap converges to* 0*, i.e.,*

$$\sup_{\mathsf{x}\in\mathcal{Z}}(K(x_k, v) - K(x, v_k)) \to 0 \ a.s.. \tag{3.89}$$

*Proof.* Since $\mathcal{Z}\cap\mathcal{S}\neq\emptyset$, the partial gap $\sup_{\mathsf{x}\in\mathcal{Z}}(K(x_k, v) - K(x, v_k))$ is nonnegative. By Theorem 3.16 and $\dim(\mathcal{H}\times\mathcal{G}) < +\infty$, $\mathsf{x}_k \to \mathsf{x}^\dagger$ a.s. and $\lambda_0 = \sup_{\mathsf{x}\in\mathcal{Z}}\|x\| < +\infty$. Simple calculations show that

$$\begin{aligned}
&\Big(K(x_k, v) - G(x_k, v^\dagger)\Big) - \Big(K(x, v_k) - G(x^\dagger, v_{s+1})\Big)\\
&= G(x^\dagger, v) - G(x, v^\dagger) + \Big\langle L(x_k - x^\dagger) \mid v - v^\dagger\Big\rangle - \Big\langle L(x - x^\dagger) \mid v_k - v^\dagger\Big\rangle.
\end{aligned} \tag{3.90}$$

Since $\mathsf{x}^\dagger\in\mathcal{S}$, the convexity of $f, h, g, \ell$ and the linearity of $L$ imply that $G(x^\dagger, v) - G(x, v^\dagger) \leq 0$. Therefore, it follows from (3.90) that

$$\begin{aligned}
&\Big(K(x_k, v) - G(x_k, v^\dagger)\Big) - \Big(K(x, v_k) - G(x^\dagger, v_k)\Big)\\
&\leq \Big\langle L(x_k - x^\dagger) \mid v - v^\dagger\Big\rangle - \Big\langle L(x - x^\dagger) \mid v_k - v^\dagger\Big\rangle\\
&\leq \|L\|(\lambda_0 + \|x^\dagger\|)\|x_k - x^\dagger\| + \|L\|(\lambda_0 + \|v^\dagger\|)\|v_k - v^\dagger\|,
\end{aligned} \tag{3.91}$$

which implies that

$$\sup_{\mathsf{x}\in\mathcal{Z}}(K(x_k, v) - K(x, v_k)) \leq \Theta(\mathsf{x}_k) + \|L\|(\lambda_0 + \|x^\dagger\|)\|x_k - x^\dagger\| + \|L\|(\lambda_0 + \|v^\dagger\|)\|v_k - v^\dagger\|. \tag{3.92}$$

Since $\Theta(x_k) \to 0$ and $\mathsf{x}_k \to \mathsf{x}^\dagger$ a.s., (3.89) follows from (3.92). □

We next show that the proposed method is a variance reduction method.

**Corollary 3.18** *Under the same setting as Theorem 3.12, Algorithm 3.1 is indeed a variance reduction method in the sense of Definition 2.5.*

*Proof.* The conclusion follows directly from (3.82). □

The following theorem proves the almost sure convergence of $G_{\beta_k}(\mathsf{x}_k; \mathsf{x}^\dagger)$ to 0 and the almost sure weak convergence of the sequence $(\mathsf{x}_k)_{k\in\mathbb{N}}$ to a point $\mathsf{x}^\dagger\in\mathcal{S}$.

**Theorem 3.19** *Let $k \in \mathbb{N}$ and define*

$$\begin{cases} \boldsymbol{Z}_k & = \begin{pmatrix} 8\mu^2\tau_k(1+\frac{2}{\beta_k})(1-p) + \frac{4\tau_k}{\beta_k}L^\star L & 0 \\ 0 & 8\nu^2\sigma_k(1+\frac{2}{\beta_k})(1-q) + \frac{4\sigma_k}{\beta_k}LL^\star \end{pmatrix} + \hat{\boldsymbol{V}} \\ \overline{\Lambda}_{k+1} & = \boldsymbol{U}_k - \mathbf{T}_{k+1} - \boldsymbol{Z}_{k+1}. \end{cases} \tag{3.93}$$

*Define the following Lyapunov function*

$$\mathcal{L}_{\beta_k}(\mathsf{x}^\dagger) = G_{\beta_k}(\mathsf{x}_k;\mathsf{x}^\dagger) + Q(\mathsf{w}_{k-1}) + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} - \boldsymbol{b}_k(\mathsf{x}^\dagger) + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{\boldsymbol{Z}_k}, \tag{3.94}$$

*where $G_{\beta_k}(\mathsf{x}_k;\mathsf{x}^\dagger)$ is defined at each iterate $k \in \mathbb{N}$ following (2.10) and $Q(\mathsf{w}_{k-1})$ by (3.51). Let $\overline{p} = \max\{p,q\}$ and $\underline{p} = \min\{p,q\}$. 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.*

$$\begin{cases} \beta_k \geq \beta_{k-1}, \\ 4\overline{\mu}\big(2\overline{\gamma}_k(1-\underline{p}) + \overline{p}\big) + \epsilon \leq 1 + \eta_k \;\textit{with}\; \overline{\gamma}_k = \gamma_k(1+2/\beta_k), \\ (2\overline{\gamma}_k + 1)(1-\underline{p}) + \epsilon \leq 1 + \eta_k \;\textit{with}\; \overline{\gamma}_k = \gamma_k(1+2/\beta_k), \\ \boldsymbol{U}_{k-1} \succeq \boldsymbol{U}_k \succeq \epsilon\,\mathrm{Id} + \|L\|^2\hat{\boldsymbol{V}}^{-1}, \\ \overline{\Lambda}_k \succeq \epsilon\,\mathrm{Id}\,. \end{cases} \tag{3.95}$$

*Then, for all $k \in \mathbb{N}$, the following descent property is verified by the Lyapunov function (3.94)*

$$\mathsf{E}_k\left[\mathcal{L}_{\beta_{k+1}}(\mathsf{x}^\dagger)\right] + \mathsf{E}_k\left[\frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\overline{\Lambda}_{k+1}}\right] \leq (1+\eta_k)\mathcal{L}_{\beta_k}(\mathsf{x}^\dagger) - \epsilon\big(G_{\beta_k}(\mathsf{x}_k;\mathsf{x}^\dagger) + Q(\mathsf{w}_{k-1})\big). \tag{3.96}$$

*Consequently,*

$$G_{\beta_k}(\mathsf{x}_k;\mathsf{x}^\dagger) \to 0 \textit{ and } Q(\mathsf{w}_k) \to 0 \textit{ a.s.}. \tag{3.97}$$

*Moreoover, if (3.76) is satisfied, then $(\mathsf{x}_k)_{k\in\mathbb{N}}$ converges weakly to some random variable $\overline{\mathsf{x}} \in \mathcal{S}$ almost surely.*

*Proof.* Obseve that we can rewrite (3.93) as

$$\begin{cases} \boldsymbol{Z}_k & = \mathbf{S}_k + 8\boldsymbol{D}^2 P_k \boldsymbol{U}_k^{-1} + 4\boldsymbol{D}^2(\mathrm{Id} - \boldsymbol{P}), \\ \overline{\Lambda}_{k+1} & = \boldsymbol{U}_k - \mathbf{T}_{k+1} - \boldsymbol{Z}_{k+1} \end{cases} \tag{3.98}$$

where $P_k = (1+2/\beta_k)\boldsymbol{P}$. It follows from Lemma 3.10 that

$$(1+2/\beta_k)\mathsf{E}_k\left[\|\mathbf{r}_k - \mathbf{R}_k\|^2_{\boldsymbol{U}_k^{-1}}\right] \leq 2\overline{\gamma}_k(1-\underline{p})\big(Q(\mathsf{w}_{k-1}) + 4\overline{\mu}\Theta(\mathsf{x}_k)\big) + \|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{4\boldsymbol{D}^2 P_k \boldsymbol{U}_k^{-1}} \tag{3.99}$$

Hence, by taking conditional expectation on both sides of (3.45) in Lemma 3.8, we obtain

$$\begin{aligned} \mathsf{E}_k\left[G_{\beta_k}(\mathsf{x}_{k+1};\mathsf{x}^\dagger)\right] \leq{}& 2\overline{\gamma}_k(1-\underline{p})\big(Q(\mathsf{w}_{k-1}) + 4\overline{\mu}\Theta(\mathsf{x}_k)\big) + \|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{4\boldsymbol{D}^2 P_k \boldsymbol{U}_k^{-1}} \\ &+ \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{\mathbf{S}_k} - \boldsymbol{b}_k(\mathsf{x}^\dagger) \\ &- \left(\frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\mathbf{S}_{k+1}} - \boldsymbol{b}_{k+1}(\mathsf{x}^\dagger)\right) \\ &+ \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\mathbf{S}_{k+1}+\mathbf{T}_{k+1}} - \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\boldsymbol{U}_k}. \end{aligned} \tag{3.100}$$

Adding (3.66) to (3.100), we obtain

$$\begin{aligned}
\mathsf{E}_k\left[G_{\beta_k}(\mathsf{x}_{k+1};\mathsf{x}^\dagger)\right] + \mathsf{E}_k\left[Q(\mathsf{w}_k)\right] \le{}& 4\overline{\mu}\big(2\overline{\gamma}_k(1-\underline{p})+\overline{p}\big)\Theta(\mathsf{x}_k) + (2\overline{\gamma}_k+1)(1-\underline{p})Q(\mathsf{w}_{k-1})\\
&+\frac{1}{2}\|\mathsf{x}_k-\mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_k-\mathsf{x}_{k-1}\|^2_{\mathbf{S}_k+8\boldsymbol{D}^2P_k\boldsymbol{U}_k^{-1}+4\boldsymbol{D}^2(\mathrm{Id}-\boldsymbol{P})} - \boldsymbol{b}_k(\mathsf{x}^\dagger)\\
&-\left(\frac{1}{2}\|\mathsf{x}_{k+1}-\mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_{k+1}-\mathsf{x}_k\|^2_{\mathbf{S}_{k+1}} - \boldsymbol{b}_{k+1}(\mathsf{x}^\dagger)\right)\\
&+\frac{1}{2}\|\mathsf{x}_{k+1}-\mathsf{x}_k\|^2_{\mathbf{S}_{k+1}+\mathbf{T}_{k+1}} - \frac{1}{2}\|\mathsf{x}_{k+1}-\mathsf{x}_k\|^2_{\boldsymbol{U}_k}.
\end{aligned}\tag{3.101}$$

In view of the notations defined in (3.98), we can rewrite (3.101) as

$$\begin{aligned}
\mathsf{E}_k\left[G_{\beta_k}(\mathsf{x}_{k+1};\mathsf{x}^\dagger)\right] + \mathsf{E}_k\left[Q(\mathsf{w}_k)\right] \le{}& 4\overline{\mu}\big(2\overline{\gamma}_k(1-\underline{p})+\overline{p}\big)\Theta(\mathsf{x}_k) + (2\overline{\gamma}_k+1)(1-\underline{p})Q(\mathsf{w}_{k-1})\\
&+\frac{1}{2}\|\mathsf{x}_k-\mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_k-\mathsf{x}_{k-1}\|^2_{\boldsymbol{Z}_k} - \boldsymbol{b}_k(\mathsf{x}^\dagger)\\
&-\left(\frac{1}{2}\|\mathsf{x}_{k+1}-\mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_{k+1}-\mathsf{x}_k\|^2_{\boldsymbol{Z}_{k+1}} - \boldsymbol{b}_{k+1}(\mathsf{x}^\dagger)\right)\\
&-\frac{1}{2}\|\mathsf{x}_{k+1}-\mathsf{x}_k\|^2_{\Lambda_{k+1}}.
\end{aligned}\tag{3.102}$$

Since $(\beta_k)_{k\in\mathbb{N}}$ is assumed increasing, $G_{\beta_k}(\mathsf{x}_{k+1};\mathsf{x}^\dagger) \ge G_{\beta_{k+1}}(\mathsf{x}_{k+1};\mathsf{x}^\dagger)$. Moreover, by Lemma 2.2, $\Theta(\mathsf{x}_k) \le G_{\beta_k}(\mathsf{x}_k;\mathsf{x}^\dagger)$. Therefore, (3.102) can be further estimated as follows

$$\begin{aligned}
\mathsf{E}_k\left[\mathcal{L}_{\beta_{k+1}}(\mathsf{x}^\dagger)\right] + \mathsf{E}_k\left[\frac{1}{2}\|\mathsf{x}_{k+1}-\mathsf{x}_k\|^2_{\Lambda_{k+1}}\right] \le{}& 4\overline{\mu}\big(2\overline{\gamma}_k(1-\underline{p})+\overline{p}\big)G_{\beta_k}(\mathsf{x}_k;\mathsf{x}^\dagger) + (2\overline{\gamma}_k+1)(1-\underline{p})Q(\mathsf{w}_{k-1})\\
&+\frac{1}{2}\|\mathsf{x}_k-\mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_k-\mathsf{x}_{k-1}\|^2_{\boldsymbol{Z}_k} - \boldsymbol{b}_k(\mathsf{x}^\dagger)\\
\le{}& (1+\eta_k)\big(G_{\beta_k}(\mathsf{x}_k;\mathsf{x}^\dagger) + Q(\mathsf{w}_{k-1})\big)\\
&+\frac{1}{2}\|\mathsf{x}_k-\mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_k-\mathsf{x}_{k-1}\|^2_{\boldsymbol{Z}_k} - \boldsymbol{b}_k(\mathsf{x}^\dagger)\\
&-\epsilon\big(G_{\beta_k}(\mathsf{x}_k;\mathsf{x}^\dagger) + Q(\mathsf{w}_{k-1})\big).
\end{aligned}\tag{3.103}$$

Next, we use (3.69) to obtain

$$\begin{aligned}
\frac{1}{2}\|\mathsf{x}_k-\mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_k-\mathsf{x}_{k-1}\|^2_{\boldsymbol{Z}_k} - \boldsymbol{b}_k(\mathsf{x}^\dagger) \ge{}& \frac{1}{2}\|\mathsf{x}_k-\mathsf{x}_{k-1}\|^2_{\boldsymbol{Z}_k-\hat{\boldsymbol{V}}} + \frac{1}{2}\|\mathsf{x}_k-\mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k-\|L\|^2\hat{\boldsymbol{V}}^{-1}}\\
\ge{}& 0,
\end{aligned}\tag{3.104}$$

where the last inequality follows from $\boldsymbol{Z}_k - \hat{\boldsymbol{V}} \succeq 0$ and $\boldsymbol{U}_k - \|L\|^2\hat{\boldsymbol{V}}^{-1} \succeq \epsilon\,\mathrm{Id}$ in (3.95). Therefore, we can further estimate (3.103) as

$$\begin{aligned}
\mathsf{E}_k\left[\mathcal{L}_{\beta_{k+1}}(\mathsf{x}^\dagger)\right] + \mathsf{E}_k\left[\frac{1}{2}\|\mathsf{x}_{k+1}-\mathsf{x}_k\|^2_{\Lambda_{k+1}}\right] \le{}& (1+\eta_k)\bigg(G_{\beta_k}(\mathsf{x}_k;\mathsf{x}^\dagger) + Q(\mathsf{w}_{k-1})\\
&+\frac{1}{2}\|\mathsf{x}_k-\mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_k-\mathsf{x}_{k-1}\|^2_{\boldsymbol{Z}_k} - \boldsymbol{b}_k(\mathsf{x}^\dagger)\bigg)\\
&-\epsilon\big(G_{\beta_k}(\mathsf{x}_k;\mathsf{x}^\dagger) + Q(\mathsf{w}_{k-1})\big)\\
={}& (1+\eta_k)\mathcal{L}_{\beta_k}(\mathsf{x}^\dagger) - \epsilon\big(G_{\beta_k}(\mathsf{x}_k;\mathsf{x}^\dagger) + Q(\mathsf{w}_{k-1})\big).
\end{aligned}\tag{3.105}$$

Hence, (3.96) is proved. The remainder of the proof is similar to the proof of Theorem 3.16, and we omit it here. □

**Remark 3.20** Here are some comments.

(i) The weak convergence of the iterate as well as the convergence of the smoothed primal-dual gap function appear to be new in the context of loopless variance reduction method for solving primal-dual problems. In the case of non-loopless variance reduction method, this kind of result has also been obtained in [25]. While the proof of the almost sure convergence of the iterations based on the gap function is not new approach even in the stochastic; see [31, 25] for instance.

(ii) To the best of our knowledge, our results appear to be the first establishing the convergence of the smoothed primal-dual gap introduced by [16] 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 duality and the smoothed primal-dual gap as well as the iteration.

**Theorem 3.21** *Suppose that $f$ and $g^\star$ are strongly convex functions with strictly positive constants $\theta_1$ and $\theta_2$, respectively. Let $\overline{\mu} = \max\{\mu, \nu\}$, $\overline{p} = \max\{p, q\}$ and $\underline{p} = \min\{p, q\}$. For every $k \in \mathbb{N}$, set $\underline{\epsilon} = \inf_{k\in\mathbb{N}} \min\{\theta_1\tau_k, \theta_2\sigma_k\}$ and $\gamma_k = \max\{\tau_k, \sigma_k\}$. Suppose that $\rho_0$ verifies*

$$(\forall k \in \mathbb{N}) \begin{cases} (2+\underline{\epsilon})(1-\rho_0) \leq \underline{\epsilon} \\ 4\overline{\mu}\big(2\gamma_k(1-\underline{p}) + \overline{p}\big) \leq \rho_0 < 1 \\ (2\gamma_k + 1)(1-\underline{p}) \leq \rho_0 < 1. \end{cases} \tag{3.106}$$

*and that*

$$(\forall k \in \mathbb{N}) \begin{cases} \boldsymbol{U}_{k-1} \succeq \boldsymbol{U}_k \\ \Lambda_{k+1} \succeq \dfrac{1-\rho_0}{\rho_0}(\boldsymbol{V}_{k+1} + \boldsymbol{L}^\star \boldsymbol{U}_{k+1}^{-1}\boldsymbol{L}). \end{cases} \tag{3.107}$$

*Then, the following hold:*

$$\mathsf{E}_k[\Theta(\mathsf{x}_{k+1})] = \mathcal{O}(\rho_0^k),\ \mathsf{E}_k[Q(\mathsf{w}_k)] = \mathcal{O}(\rho_0^k) \textit{ and } \mathsf{E}_k\left[[\epsilon\|\mathsf{x}_{k+1} - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_{k+1}}\right] \leq \mathcal{O}(\rho_0^k). \tag{3.108}$$

*Proof.* By using (3.41), instead of (3.68), we obtain

$$\begin{aligned} \mathsf{E}_k\left[\Theta(\mathsf{x}_{k+1}) + Q(\mathsf{w}_k) + \frac{\epsilon}{2}\|\mathsf{x}_{k+1} - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k}\right] &\leq 4\overline{\mu}\big(2\gamma_k(1-\underline{p}) + \overline{p}\big)\Theta(\mathsf{x}_k) + (2\gamma_k + 1)(1-\underline{p})Q(\mathsf{w}_{k-1}) \\ &\quad + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{\boldsymbol{V}_k} - \boldsymbol{b}_k(\mathsf{x}^\dagger) \\ &\quad - \mathsf{E}_k\left[\frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\boldsymbol{V}_{k+1}} - \boldsymbol{b}_{k+1}(\mathsf{x}^\dagger)\right] \\ &\quad - \mathsf{E}_k\left[\frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\Lambda_{k+1}}\right]. \end{aligned} \tag{3.109}$$

This inequality together with the condition (3.106) gives

$$\begin{aligned}
\mathsf{E}_k\left[\Theta(\mathsf{x}_{k+1})\right] + \mathsf{E}_k\left[Q(\mathsf{w}_k)\right] \leq{}& \rho_0\big(\Theta(\mathsf{x}_k) + Q(\mathsf{w}_{k-1})\big)\\
&\quad + \frac{1+\underline{\epsilon}}{2}\|\mathsf{x}_k - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{\boldsymbol{V}_k} - \boldsymbol{b}_k(\mathsf{x}^\dagger) - \frac{\epsilon}{2}\|\mathsf{x}_k - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k}\\
&\quad - \mathsf{E}_k\left[\frac{1+\underline{\epsilon}}{2}\|\mathsf{x}_{k+1} - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\boldsymbol{V}_{k+1}} - \boldsymbol{b}_{k+1}(\mathsf{x}^\dagger)\right]\\
&\quad - \mathsf{E}_k\left[\frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\Lambda_{k+1}}\right]\\
\leq{}& \rho_0\big(\Theta(\mathsf{x}_k) + Q(\mathsf{w}_{k-1})\big)\\
&\quad + \frac{1+\underline{\epsilon}}{2}\|\mathsf{x}_k - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{\boldsymbol{V}_k} - \boldsymbol{b}_k(\mathsf{x}^\dagger) - \frac{\epsilon}{2}\|\mathsf{x}_k - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k}\\
&\quad - \mathsf{E}_k\left[\frac{1+\underline{\epsilon}}{2}\|\mathsf{x}_{k+1} - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_{k+1}} + \frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\boldsymbol{V}_{k+1}} - \boldsymbol{b}_{k+1}(\mathsf{x}^\dagger)\right]\\
&\quad - \mathsf{E}_k\left[\frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\Lambda_{k+1}}\right]\\
={}& \rho_0\big(\Theta(\mathsf{x}_k) + Q(\mathsf{w}_{k-1})\big) + \boldsymbol{a}_k - \frac{\epsilon}{2}\|\mathsf{x}_k - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} - \boldsymbol{a}_{k+1}\\
&\quad - \mathsf{E}_k\left[\frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\Lambda_{k+1}}\right],
\end{aligned}\tag{3.110}$$

where the last inequality follows from the first condition in (3.107) and

$$\boldsymbol{a}_k = \frac{1+\underline{\epsilon}}{2}\|\mathsf{x}_k - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{\boldsymbol{V}_k} - \boldsymbol{b}_k(\mathsf{x}^\dagger).\tag{3.111}$$

Then,

$$\begin{aligned}
(1-\rho_0)\boldsymbol{b}_k(\mathsf{x}^\dagger) &= (1-\rho_0)\left\langle \boldsymbol{U}_k^{-1}\mathbf{L}^\star(\mathsf{x}_k - \mathsf{x}_{k-1}) \mid \mathsf{x}_k - \mathsf{x}^\dagger\right\rangle_{\boldsymbol{U}_k}\\
&\geq -\frac{(1-\rho_0)}{2}\|\mathsf{x}_k - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} - \frac{(1-\rho_0)}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{\boldsymbol{L}^\star\boldsymbol{U}_k^{-1}\boldsymbol{L}}
\end{aligned}\tag{3.112}$$

Therefore,

$$\begin{aligned}
\boldsymbol{a}_k &= \rho_0\boldsymbol{a}_k + (1-\rho_0)\boldsymbol{a}_k\\
&= \rho_0\boldsymbol{a}_k + \frac{(1+\underline{\epsilon})(1-\rho_0)}{2}\|\mathsf{x}_k - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1-\rho_0}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{\boldsymbol{V}_k} - (1-\rho_0)\boldsymbol{b}_k(\mathsf{x}^\dagger)\\
&\leq \rho_0\boldsymbol{a}_k + \frac{(2+\underline{\epsilon})(1-\rho_0)}{2}\|\mathsf{x}_k - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} + \frac{1-\rho_0}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{\boldsymbol{V}_k + \boldsymbol{L}^\star\boldsymbol{U}_k^{-1}\boldsymbol{L}}.
\end{aligned}\tag{3.113}$$

Now, using the second condition in (3.106), i.e., $(2+\underline{\epsilon})(1-\rho_0) \leq \underline{\epsilon}$, we obtain

$$\begin{aligned}
\boldsymbol{a}_k - \frac{\epsilon}{2}\|\mathsf{x}_k - \mathsf{x}^\dagger\|^2_{\boldsymbol{U}_k} &\leq \rho_0\boldsymbol{a}_k + \frac{1-\rho_0}{2}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{\boldsymbol{V}_k + \boldsymbol{L}^\star\boldsymbol{U}_k^{-1}\boldsymbol{L}}\\
&= \rho_0\Big(\boldsymbol{a}_k + \frac{1-\rho_0}{2\rho_0}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{\boldsymbol{V}_k + \boldsymbol{L}^\star\boldsymbol{U}_k^{-1}\boldsymbol{L}}\Big).
\end{aligned}\tag{3.114}$$

Therefore, (3.110) can be further estimated as

$$
\begin{aligned}
\mathsf{E}_k\left[\Theta(\mathsf{x}_{k+1})\right] + \mathsf{E}_k\left[Q(\mathsf{w}_k)\right] \le{}& \rho_0\Big(\Theta(\mathsf{x}_k) + Q(\mathsf{w}_{k-1}) + \boldsymbol{a}_k + \frac{1-\rho_0}{2\rho_0}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{\boldsymbol{V}_k + \boldsymbol{L}^\star \boldsymbol{U}_k^{-1}\boldsymbol{L}}\Big) \\
&- \mathsf{E}_k\left[\boldsymbol{a}_{k+1}\right] - \mathsf{E}_k\left[\frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\Lambda_k}\right] \\
={}& \rho_0\Big(\Theta(\mathsf{x}_k) + Q(\mathsf{w}_{k-1}) + \boldsymbol{a}_k + \frac{1-\rho_0}{2\rho_0}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{\boldsymbol{V}_k + \boldsymbol{L}^\star \boldsymbol{U}_k^{-1}\boldsymbol{L}}\Big) \\
&- \mathsf{E}_k\left[\boldsymbol{a}_{k+1} + \frac{1-\rho_0}{2\rho_0}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\boldsymbol{V}_{k+1} + \boldsymbol{L}^\star \boldsymbol{U}_{k+1}^{-1}\boldsymbol{L}}\right] \\
&+ \mathsf{E}_k\left[\frac{1-\rho_0}{2\rho_0}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\boldsymbol{V}_{k+1} + \boldsymbol{L}^\star \boldsymbol{U}_{k+1}^{-1}\boldsymbol{L}}\right] - \mathsf{E}_k\left[\frac{1}{2}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\Lambda_{k+1}}\right]
\end{aligned}
\tag{3.115}
$$

The difference between the last two terms in (3.115) is negative due to the condition (3.107). Therefore,

$$
\begin{aligned}
\mathsf{E}_k\left[\Theta(\mathsf{x}_{k+1})\right] + &\mathsf{E}_k\left[Q(\mathsf{w}_k) + \boldsymbol{a}_{k+1} + \frac{1-\rho_0}{2\rho_0}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\boldsymbol{V}_{k+1} + \boldsymbol{L}^\star \boldsymbol{U}_{k+1}^{-1}\boldsymbol{L}}\right] \\
&\le \rho_0\Big(\Theta(\mathsf{x}_k) + Q(\mathsf{w}_{k-1}) + \boldsymbol{a}_k + \frac{1-\rho_0}{2\rho_0}\|\mathsf{x}_k - \mathsf{x}_{k-1}\|^2_{\boldsymbol{V}_k + \boldsymbol{L}^\star \boldsymbol{U}_k^{-1}\boldsymbol{L}}\Big)
\end{aligned}
\tag{3.116}
$$

Using this expression recursively, we obtain

$$
\mathsf{E}_k\left[\Theta(\mathsf{x}_{k+1})\right] + \mathsf{E}_k\left[Q(\mathsf{w}_k) + \boldsymbol{a}_{k+1} + \frac{1-\rho_0}{2\rho_0}\|\mathsf{x}_{k+1} - \mathsf{x}_k\|^2_{\boldsymbol{V}_{k+1} + \boldsymbol{L}^\star \boldsymbol{U}_{k+1}^{-1}\boldsymbol{L}}\right] \le \mathcal{O}(\rho_0^k), \tag{3.117}
$$

which proves the desired results. □

**Remark 3.22** Under the strong convexity of $f$ and $g^\star$, the linear convergence of $\Theta(\mathsf{x}_k)$ implies the linear convergence of the duality gap defined by (2.12), i.e., $\sup_{\mathsf{x}\in\mathcal{H}\times\mathcal{G}} K(x_k, v) - K(x, v_k) = \mathcal{O}(\rho_0^k)$ [25]. Moreover, by using the same technique, the linear convergence of the smoothed primal-dual gap function can be obtained. Hence, we omit it here.

**Remark 3.23** The linear convergence of the duality gap as well as the smoothed primal-dual 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 [5, 16, 24, 32]. If $\ell^\star = 0$ and $f = 0$, under additional assumptions on the linear operator $L$, [15] achieves the linear convergence rate even when the strongly convex-concave condition is not full-filled.

The following proposition provides an explicit expression of the stepsize and will be further used when developing the computational complexity results in Section 4.

**Proposition 3.24** *Under the same conditions stated in Theorem 3.21. Set* $\chi := 10\mu + \frac{\|L\|^2}{\mu}$. *Suppose that* $\mu = \nu$, $p = q < \min\{1/\mu, 1/5\}$ *together with*

$$
\theta_1 = \theta_2 = \theta \le \min\{(2\mu + \chi)/(4\chi), \mu\} \tag{3.118}
$$

*and* $\tau_k \equiv \sigma_k = \gamma \le 0.5/\mu$. *Then the condition* (3.106) *and* (3.107) *are satisfied when*

$$
\gamma = \sqrt{2}\min\left\{\frac{p/(1-p)}{(\theta+4)\theta}; \frac{(1/(4\mu) - p)}{4(1-p) + p\theta}; \frac{1}{2(2\mu+\chi)}\right\}. \tag{3.119}
$$

*Proof.* Under the conditions $p = q$, $\mu = \nu$, $\theta_1 = \theta_2 = \theta$ and $\tau_k \equiv \sigma_k = \gamma_k \equiv \gamma$, we have

$$\underline{\epsilon} = \inf_{k \in \mathbb{N}} \min\{\theta_1 \tau_k, \theta_2 \sigma_k\} = \gamma\theta. \tag{3.120}$$

If we take

$$\rho_0 = \frac{2}{2 + \underline{\epsilon}} = \frac{2}{2 + \gamma\theta} < 1, \tag{3.121}$$

then the first condition of (3.106) is satisfied. Moreover, from (3.106), simple calculations show that the condition $(2\gamma + 1)(1 - \underline{p}) \leq \rho_0$ is satisfied when

$$\begin{aligned}
\gamma \leq \gamma_{00} &:= \frac{-(\theta + 4) + \sqrt{(\theta + 4)^2 + 16p/(1 - p)}}{4\theta} \\
&= \frac{16p/(1 - p)}{4\theta[\theta + 4 + \sqrt{(\theta + 4)^2 + 16p/(1 - p)}]} \\
&\geq \frac{16p/(1 - p)}{8\theta\sqrt{(\theta + 4)^2 + 16p/(1 - p)}} && (3.122) \\
&\geq \frac{\sqrt{2}p}{(\theta + 4)(1 - p)\theta}, && (3.123)
\end{aligned}$$

where the last inequality follows from the condition $p < 1/5$ which implies

$$\frac{16p/(1 - p)}{\sqrt{(\theta + 4)^2 + 16p/(1 - p)}} \geq \frac{16p/(1 - p)}{\sqrt{2}(\theta + 4)}. \tag{3.124}$$

From (3.106), the condition $4\overline{\mu}\big(2\gamma(1 - \underline{p}) + \overline{p}\big) \leq \rho_0$ is satisfied when

$$\begin{aligned}
\gamma \leq \gamma_{01} &:= \frac{-4(1 - p) - p\theta + \sqrt{(4(1 - p) + p\theta)^2 - 16\theta(1 - p)(p - 1/(4\mu))}}{4\theta(1 - p)} \\
&\geq \frac{16\theta(1 - p)(1/(4\mu) - p)}{8\sqrt{2}\theta(1 - p)(4(1 - p) + p\theta)} \\
&= \frac{\sqrt{2}(1/(4\mu) - p)}{4(1 - p) + p\theta}, && (3.125)
\end{aligned}$$

where the last inequality follows from $\theta \leq \mu$. The first condition of (3.107) is trivially satisfied since $\tau_k = \sigma_k \equiv \gamma$. Since (3.67), $\Lambda_{k+1} = \boldsymbol{U}_k - 2\boldsymbol{D} - \boldsymbol{V}_{k+1} - \boldsymbol{L}^*\boldsymbol{D}^{-1}\boldsymbol{L}$, the second condition in (3.107) is equivalent to

$$\begin{aligned}
\boldsymbol{U}_k &\succeq 2\boldsymbol{D} + \frac{1}{\rho_0}\boldsymbol{V}_{k+1} + \boldsymbol{L}^*\boldsymbol{D}^{-1}\boldsymbol{L} + \frac{1 - \rho_0}{\rho_0}\boldsymbol{L}^\star\boldsymbol{U}_{k+1}^{-1}\boldsymbol{L} \\
&= 2\boldsymbol{D} + \frac{1}{\rho_0}\boldsymbol{V}_{k+1} + \mu^{-1}\boldsymbol{L}^\star\boldsymbol{L} + \gamma\frac{1 - \rho_0}{\rho_0}\boldsymbol{L}^\star\boldsymbol{L} \\
&\preceq 2\boldsymbol{D} + \frac{1}{\rho_0}(\boldsymbol{V}_{k+1} + \mu^{-1}\boldsymbol{L}^\star\boldsymbol{L}) \\
&= 2\boldsymbol{D} + \frac{1}{\rho_0}(2\boldsymbol{D} + 4\boldsymbol{D}^2(\mathrm{Id} - P) + 8\boldsymbol{D}^2P\boldsymbol{U}_k^{-1} + \mu^{-1}\boldsymbol{L}^\star\boldsymbol{L}) \\
&\preceq 2\boldsymbol{D} + \frac{1}{\rho_0}(2\boldsymbol{D} + 4\boldsymbol{D} + 4\boldsymbol{D} + \mu^{-1}\boldsymbol{L}^\star\boldsymbol{L}). && (3.126)
\end{aligned}$$

Therefore, the second condition in (3.107) is satisfied when

$$\frac{1}{\gamma} \geq 2\mu + \frac{1}{\rho_0}(10\mu + \frac{\|L\|^2}{\mu}), \tag{3.127}$$

which implies since $\theta \leq (2\mu + \chi)/(4\chi)$ that

$$\begin{aligned} \gamma \leq \gamma_{02} &:= \frac{-(4\mu + 2\chi) + \sqrt{(4\mu + 2\chi)^2 + 8\chi\theta}}{2\chi\theta} \\ &\geq \frac{\sqrt{2}}{2(2\mu + \chi)}. \end{aligned} \tag{3.128}$$

□

# 4 Complexity

The complexity analysis detailed in this section assumes that the functions $f$ and $g^\star$ are strongly convex. From Theorem 3.21 and Proposition 3.24, we derive the following result for the total average complexity. By convention (as usually performed in the literature), we measure the per-iteration complexity in terms of the number of the stochastic gradient calls, i.e., the number of calls to the so-called Stochastic First-order Oracle (SFO)[2].

**Corollary 4.1** *Under the same conditions stated in Theorem 3.21. Set $\chi := 10\mu + \frac{\|L\|^2}{\mu}$. Suppose that $\mu = \nu$, $\theta_1 = \theta_2 = \theta, \sigma_k = \gamma \leq 1/(2\mu), \tau_k = \gamma$, $n_d = n_p = N > \max\{5, 2\mu\}$, and $p = q = 1/N$. Assume that $\theta \leq \min\{(2\mu + \chi)/(4\chi), \mu\}$. Then, to reach an $\epsilon$-accurate point, the total average complexity is $\mathcal{O}\Big((N + \mu/\theta)\log(1/\epsilon)\Big)$.*

*Proof.* First, referring to Theorem 3.21, the number of iterations to reach an $\epsilon$-accurate point is driven by $\rho_0$. If we take $\rho_0$ as in Proposition 3.24 and set $\rho = \frac{1}{2}\theta\gamma$, then

$$\rho_0 = \frac{1}{1 + \frac{1}{2}\gamma\theta} = \frac{1}{1 + \rho}, \text{ where } \rho = \frac{1}{2}\theta\gamma.$$

Hence, since $\log(1/(1+\rho)) = -\log(1+\rho)$, the method reaches an $\epsilon$-accurate point after

$$\mathcal{O}\Big(\frac{\log(1/\epsilon)}{\log(1+\rho)}\Big) \tag{4.1}$$

iterations. Next, at each iteration, the number of calls in expectation to the stochastic first-order oracle is

$$\mathcal{O}(2 + pN). \tag{4.2}$$

By multiplying the term (4.1) with (4.2) and approximating $\log(1+t) \sim t$ $(t \ll 1)$, we have

$$\mathcal{O}\Big((2 + pN)\frac{\log(1/\epsilon)}{\log(1+\rho)}\Big) \sim \mathcal{O}\Big(\frac{2 + pN}{\rho}\log(1/\epsilon)\Big). \tag{4.3}$$

[2] Whence, this complexity measure is often referred to as the oracle complexity.

Since $p = q = 1/N$, it follows that

$$\frac{2 + pN}{\rho} = \frac{3}{\rho} = \frac{6}{\theta\gamma} \tag{4.4}$$

Referring to Proposition 3.24

$$\frac{6}{\theta\gamma} = (3\sqrt{2}) \max \left\{ \frac{(\theta+4)(1-p)}{p}; \frac{4(1-p)+p\theta}{(1/(4\mu)-p)\theta}; \frac{2(2\mu+\chi)}{\theta} \right\};$$

hence,

$$\mathcal{O}\Big((2 + pN)/\rho\Big) \sim \mathcal{O}(N + \frac{\mu}{\theta}). \tag{4.5}$$

It follows that the the total average complexity is $\mathcal{O}\Big((N + \mu/\theta)\log(1/\epsilon)\Big)$. □

**Remark 4.2** In view of Corollary 4.1, the proposed method obtains the optimal total average complexity $\mathcal{O}(N + \mu/\theta)\log(1/\epsilon)$. This recovers the complexity result in [19] obtained when minimizing only one $\theta$–strongly convex function $h$ defined by finite sums. This result significantly improves the total average complexity $\mathcal{O}(N + \sqrt{N}(\mu + \|L\|)/\theta)\log(1/\epsilon)$ obtained in [8] and [2]. Moreover, we can also observe that the total average complexity obtained for proposed method improves the complexity of the deterministic method from $\mathcal{O}(N(\mu + \|L\|)/\theta)\log(1/\epsilon)$ to $\mathcal{O}(N + \mu/\theta)\log(1/\epsilon)$.

**Remark 4.3** In view of (2.9), one can apply directly several existing methods for solving the monotone inclusions $0 \in (\boldsymbol{M} + \boldsymbol{C})\mathsf{x}$ as in [5] and [2] to obtain suboptimal total complexity $\mathcal{O}(N + (\mu + \|L\|)^2/\theta^2)\log(1/\epsilon)$ and $\mathcal{O}(N + \sqrt{N}(\mu + \|L\|)/\theta)\log(1/\epsilon)$, respectively.

## 5 Conclusion

In this paper, we developed a new primal-dual splitting algorithm with loopless variance reduction for solving Problem 1.1. We proved the weak almost sure convergence of the iterations and the convergence of the duality gap and the smoothed primal-dual gap functions 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 $\omega_k$ as those used in [25], all results presented in this paper can be extended to this general case with minor modification of the conditions. In terms of computational complexity, the proposed stochastic primal-dual splitting algorithm reaches the optimal total average complexity as in [19].

## References

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

[2] A. Alacaoglu and Y. Malitsky, Stochastic variance reduction for variational inequality methods, *Proceedings of Thirty Fifth Conference on Learning Theory*, PMLR, Vol. 178, pp. 778-816, 2022.

[3] 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.

[4] H. H. Bauschke and P. L. Combettes, Convex Analysis and Monotone Operator Theory in Hilbert Spaces, Springer, New York, 2nd ed., 2017.

[5] P. Balamurugan and F. Bach, Stochastic Variance Reduction Methods for Saddle-Point Problems, *Adv. Neural Inf. Process. Syst.*, Vol. 29, pp. 1416–1424, 2016.

[6] 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.

[7] M.N. Bùi and P. L . Combettes, Multivariate monotone inclusions in saddle form, *Math. Oper. Res.*, Vol. 47, pp. 1082-1109, 2022.

[8] Y. Carmon, Y. Jin, A. Sidford, and K. Tian, Variance reduction for matrix games, *Adv. Neural Inf. Process. Syst.*, Vol. 32, pp. 11381–11392, 2019.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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*,Vol. 89, pp. 196-205, 2019.

[16] O. Fercoq, Quadratic error bound of the smoothed gap and the restarted averaged primal-dual hybrid gradient, *Open J. Math. Optim.*, Vol. 4, No. 6. 34p., 2023

[17] R. M. Gower, M. Schmidt, F. Bach and P. Richtarik, Variance-Reduced Methods for Machine Learning, *Proceedings of the IEEE*, Vol. 108, pp. 1968-1983 2020.

[18] 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.

[19] 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*, Vol. 117, pp. 451-467, 2020.

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

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

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

[23] A. Nemirovski, A. Juditsky, G. Lan and A. Shapiro, Robust stochastic approximation approach to stochastic programming, *SIAM J. Optim.*, Vol. 19, pp. 1574-1609, 2009.

[24] 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.

[25] V. D. Nguyen, B. C. Vũ and D. Papadimitriou, A Stochastic Primal-Dual Splitting Algorithm with Variance Reduction for Composite Optimization Problems, In press, Available at https://optimization-online.org/?p=28050.

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

[27] A. Nitanda, Stochastic proximal gradient descent with acceleration techniques, *Adv. Neural Inf. Process. Syst.*, Vol.1, pp. 1574-1582, 2014.

[28] 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.

[29] 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.

[30] 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.

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

[32] Z. Shi, X. Zhang and Y. Yu, Bregman divergence for stochastic variance reduction: Saddle-point and adversarial prediction, *Adv. Neural Inf. Process. Syst.*, Vol. 30, pp. 6031-6041, 2017.

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

[34] B. C. Vũ, Almost sure convergence of the stochastic forward-backward-forward splitting algorithm, *Optim. Lett.*, Vol. 10, pp. 781-803, 2016.

[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] 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.

[37] R. Zhao, Accelerated stochastic algorithms for convex-concave saddle-point problems, *Math. Oper. Res.*, Vol. 47, pp. 1443-1473, 2021.

**Appendix A** [26, Corollary 2.6] Let $(\mathcal{F}_n)_{n\in\mathbb{N}}$ be an increasing sequence of sub-$\sigma$-algebras of $\mathcal{F}$, let $(x_n)_{n\in\mathbb{N}}$ be a $[0,+\infty[$-valued random sequence such that, for every $n\in\mathbb{N}$, $x_{n-1}$ is $\mathcal{F}_n$-measurable and

$$\sum_{n\in\mathbb{N}} \mathsf{E}[x_n|\mathcal{F}_n] < +\infty \quad a.s.. \tag{5.1}$$

Then $\sum\limits_{n\in\mathbb{N}} x_n < +\infty$ a.s..

*Proof.* Let us set

$$(\forall n\in\mathbb{N})\ z_n = \sum_{k=1}^{n-1} x_k.$$

Then, $z_n$ is $\mathcal{F}_n$ measurable. Moreover,

$$\mathsf{E}[z_{n+1}|\mathcal{F}_n] = z_n + \mathsf{E}[x_n|\mathcal{F}_n]. \quad a.s..$$

Hence, it follows from Lemma 2.6 and (5.1) that $(z_n)_{n\in\mathbb{N}}$ converges a.s.. □