Numerical Algorithms (2024) 97:51–100
https://doi.org/10.1007/s11075-023-01693-9

ORIGINAL PAPER

A stochastic two-step inertial Bregman proximal alternating linearized minimization algorithm for nonconvex and nonsmooth problems

Chenzheng Guo[1] · **Jing Zhao**[1] · **Qiao-Li Dong**[1]

Received: 10 July 2023 / Accepted: 19 October 2023 / Published online: 9 November 2023

Abstract
In this paper, for solving a broad class of large-scale nonconvex and nonsmooth optimization problems, we propose a stochastic two-step inertial Bregman proximal alternating linearized minimization (STiBPALM) algorithm with variance-reduced stochastic gradient estimators. And we show that SAGA and SARAH are variance-reduced gradient estimators. Under expectation conditions with the Kurdyka–Łojasiewicz property and some suitable conditions on the parameters, we obtain that the sequence generated by the proposed algorithm converges to a critical point. And the general convergence rate is also provided. Numerical experiments on sparse nonnegative matrix factorization and blind image-deblurring are presented to demonstrate the performance of the proposed algorithm.

Keywords Nonconvex and nonsmooth optimization · Stochastic · Bregman · Variance-reduced · Kurdyka–Łojasiewicz property

Mathematics Subject Classification (2010) 47J06 · 49J52 · 65K10 · 90C26 · 90C30

1 Introduction

In this paper, we are interested in solving the following composite optimization problem:

$$\min_{x\in\mathbb{R}^l,\, y\in\mathbb{R}^m} \Phi(x, y) = f(x) + H(x, y) + g(y), \tag{1.1}$$

✉ Jing Zhao
zhaojing200103@163.com

Chenzheng Guo
g13526199036@163.com

Qiao-Li Dong
dongql@lsec.cc.ac.cn

[1] College of Science, Civil Aviation University of China, Tianjin 300300, China

where $f : \mathbb{R}^l \to (-\infty, +\infty]$ and $g : \mathbb{R}^m \to (-\infty, +\infty]$ are proper lower semicontinuous. $H(x, y) = \frac{1}{n}\sum_{i=1}^{n} H_i(x, y)$ has a finite-sum structure, $H_i : \mathbb{R}^l \times \mathbb{R}^m \to \mathbb{R}$ is continuously differentiable, and ∇H_i is Lipschitz continuous on bounded subsets. Note that here and throughout the paper, no convexity is imposed on Φ. In practical application, numerous problems can be formulated into the form of (1.1), such as signal and image processing [1, 2], nonnegative matrix factorization [3–5], blind image-deblurring [5, 6], sparse principal component analysis [7, 8], and compressed sensing [9, 10]. Here, we list two applications of (1.1), which will also be used in the numerical experiments.

(1) Sparse nonnegative matrix factorization (S-NMF). The S-NMF has important applications in image processing (face recognition) and bioinformatics (clustering of gene expressions) (see [4] for details). Given a matrix $A \in \mathbb{R}^{l\times m}$ and an integer $r > 0$, we want to seek a factorization $A \approx XY$, where $X \in \mathbb{R}^{l\times r}$ and $Y \in \mathbb{R}^{r\times m}$ are nonnegative with $r \leq \min\{l, m\}$ and X is sparse. One way to solve this problem is by finding a solution for the nonnegative least squares model given by

$$\min_{X,Y}\left\{\frac{\eta}{2}\|A - XY\|_F^2 : \; X, Y \geq 0, \; \|X_i\|_0 \leq s, \; i = 1, 2, \dots, r\right\}, \tag{1.2}$$

where $\eta > 0$, X_i denotes the ith column of X, and $\|X_i\|_0$ denotes the number of nonzero elements of the ith column of X. In this formulation, the sparsity on X is strictly enforced using the nonconvex l_0 constraint. Let $H(X, Y) = \frac{\eta}{2}\|A - XY\|_F^2 = \sum_{i=1}^{l}\frac{\eta}{2}\|A_i - X_iY\|_F^2$, $f(X) = \iota_{X\geq 0}(X) + \iota_{\|X_1\|_0\geq s}(X) + \cdots + \iota_{\|X_r\|_0\geq s}(X)$, $g(Y) = \iota_{Y\geq 0}(Y)$, where A_i denotes the ith low of A, and ι_C is the indicator function on C. Then, this model (1.2) can be converted to (1.1).

(2) Blind image deconvolution (BID). Let A be the observed blurred image, and let X be the unknown sharp image of the same size. Furthermore, let Y denote a small unknown blur kernel, and a typical variational formulation of the blind deconvolution problem is given by the following:

$$\min_{X,Y}\left\{\frac{1}{2}\|A - X \odot Y\|_F^2 + \eta\sum_{r=1}^{2d} R([D(X)]_r) : \; 0 \leq X \leq 1, \; 0 \leq Y \leq 1, \; \|Y\|_1 \leq 1\right\}, \tag{1.3}$$

where $\eta > 0$, $\odot$ is the two-dimensional convolution operator, X is the image to recover, and Y is the blur kernel to estimate. Here, $R(\cdot)$ is an image regularization term, that imposes sparsity on the image gradient and hence favors sharp images. $D(\cdot)$ is the differential operator, computing the horizontal and vertical gradients for each pixel. This model (1.3) can be converted to (1.1), where $H(X, Y) = \frac{1}{2}\|A - X \odot Y\|_F^2 + \eta\sum_{r=1}^{2d} R([D(X)]_r)$, $f(X) = \iota_{0\leq X\leq 1}(X)$, $g(Y) = \iota_{\|Y\|_1\leq 1}(Y) + \iota_{0\leq Y\leq 1}(Y)$. See [6] for details.

For solving problem (1.1), a frequently applied algorithm is the following proximal alternating linearized minimization algorithm (PALM) by Bolte et al. [11] based on results in [12, 13]:

$$\begin{cases} x_{k+1} \in \arg\min_{x\in\mathbb{R}^l}\{f(x) + \langle x, \nabla_x H(x_k, y_k)\rangle + \frac{1}{2\lambda_k}\|x - x_k\|_2^2\}, \\ y_{k+1} \in \arg\min_{y\in\mathbb{R}^m}\{g(y) + \langle y, \nabla_y H(x_{k+1}, y_k)\rangle + \frac{1}{2\mu_k}\|y - y_k\|_2^2\}, \end{cases} \tag{1.4}$$

where $\{\lambda_k\}_{k\in\mathbb{N}}$ and $\{\mu_k\}_{k\in\mathbb{N}}$ are positive sequences. To further improve the performance of PALM, Pock and Sabach [6] introduced an inertial step to PALM and proposed the following inertial proximal alternating linearized minimization (iPALM) algorithm:

$$\begin{cases} u_{1k} = x_k + \alpha_{1k}(x_k - x_{k-1}),\ v_{1k} = x_k + \beta_{1k}(x_k - x_{k-1}),\\ x_{k+1} \in \arg\min_{x\in\mathbb{R}^l}\{f(x) + \langle x, \nabla_x H(v_{1k}, y_k)\rangle + \frac{1}{2\lambda_k}\|x - u_{1k}\|_2^2\},\\ u_{2k} = y_k + \alpha_{2k}(y_k - y_{k-1}),\ v_{2k} = y_k + \beta_{2k}(y_k - y_{k-1}),\\ y_{k+1} \in \arg\min_{y\in\mathbb{R}^m}\{g(y) + \langle y, \nabla_y H(x_{k+1}, v_{2k})\rangle + \frac{1}{2\mu_k}\|y - u_{2k}\|_2^2\}, \end{cases} \tag{1.5}$$

where $\alpha_{1k}, \alpha_{2k}, \beta_{1k}, \beta_{2k} \in [0, 1]$. Then, Gao et al. [14] presented a Gauss–Seidel type inertial proximal alternating linearized minimization (GiPALM) algorithm, in which the inertial step is performed whenever the x or y-subproblem is updated. In order to use the existing information as much as possible to further improve the numerical performance, Wang et al. [15] proposed a new inertial version of proximal alternating linearized minimization (NiPALM) algorithm, which inherits both advantages of iPALM and GiPALM.

The Bregman distance regularization is an effective way to improve the numerical results of the algorithm. In [16], the authors constructed the following two-step inertial Bregman alternating minimization (TiBAM) algorithm using the information of the previous three iterates:

$$\begin{cases} x_{k+1} \in \arg\min_{x\in\mathbb{R}^l}\{\Phi(x, y_k) + D_{\phi_1}(x, x_k) + \alpha_{1k}\langle x, x_{k-1} - x_k\rangle + \alpha_{2k}\langle x, x_{k-2} - x_{k-1}\rangle\},\\ y_{k+1} \in \arg\min_{y\in\mathbb{R}^m}\{\Phi(x_{k+1}, y) + D_{\phi_2}(y, y_k) + \beta_{1k}\langle y, y_{k-1} - y_k\rangle + \beta_{2k}\langle y, y_{k-2} - y_{k-1}\},\end{cases} \tag{1.6}$$

where $D_{\phi_i}(i = 1, 2)$ denotes the Bregman distance with respect to $\phi_i(i = 1, 2)$. By linearizing $H(x, y)$ in TiBAM algorithm, the authors [17] proposed the following two-step inertial Bregman proximal alternating linearized minimization (TiBPALM) algorithm:

$$\begin{cases} x_{k+1} \in \arg\min_{x\in\mathbb{R}^l}\{f(x) + \langle x, \nabla_x H(x_k, y_k)\rangle + D_{\phi_1}(x, x_k) + \alpha_{1k}\langle x, x_{k-1} - x_k\rangle\\ \qquad\qquad + \alpha_{2k}\langle x, x_{k-2} - x_{k-1}\rangle\},\\ y_{k+1} \in \arg\min_{y\in\mathbb{R}^m}\{g(y) + \langle y, \nabla_y H(x_{k+1}, y_k)\rangle + D_{\phi_2}(y, y_k) + \beta_{1k}\langle y, y_{k-1} - y_k\rangle\\ \qquad\qquad + \beta_{2k}\langle y, y_{k-2} - y_{k-1}\rangle\}. \end{cases} \tag{1.7}$$

If we take $\phi_1(x) = \frac{1}{2\lambda}\|x\|_2^2$ and $\phi_2(y) = \frac{1}{2\mu}\|y\|_2^2$ for all $x \in \mathbb{R}^l$ and $y \in \mathbb{R}^m$, then (1.7) becomes two-step inertial proximal alternating linearized minimization (TiPALM) algorithm. Then, based on alternating minimization algorithm, Chao et al. [18] proposed inertial alternating minimization with the Bregman distance (BIAM) algorithm. Other related work can be found in [19, 20] and their references.

It should be noted that all these works are obtained for deterministic methods, i.e., no randomness involved. But when the dimension of data is very large, the computing cost of the full gradient of the function $H(x, y)$ is often prohibitively expensive. In order to overcome this difficulty, stochastic gradient approximations were applied (see, e.g., [21] and the references therein). A block stochastic gradient iteration combining a simple stochastic gradient descent (SGD) estimator with PALM was first proposed by Xu and Yin [22]. To weaken the assumptions on the objective function in [22] and improve the estimates on the convergence rate of a stochastic PALM algorithm, Driggs et al. [23] used more sophisticated so-called variance-reduced gradient estimators instead of the simple stochastic gradient descent estimators and proposed the following stochastic proximal alternating linearized minimization (SPRING) algorithm:

$$\begin{cases} x_{k+1} \in \arg\min_{x\in\mathbb{R}^l}\{f(x) + \langle x, \widetilde{\nabla}_x(x_k, y_k)\rangle + \frac{1}{2\lambda_k}\|x - x_k\|_2^2\}, \\ y_{k+1} \in \arg\min_{y\in\mathbb{R}^m}\{g(y) + \langle y, \widetilde{\nabla}_y(x_{k+1}, y_k)\rangle + \frac{1}{2\mu_k}\|y - y_k\|_2^2\}. \end{cases} \tag{1.8}$$

The key of SPRING algorithm is replacing the full gradient computations $\nabla_x H(x_k, y_k)$ and $\nabla_y H(x_{k+1}, y_k)$ with stochastic estimations $\widetilde{\nabla}_x(x_k, y_k)$ and $\widetilde{\nabla}_y(x_{k+1}, y_k)$, respectively. Then, Hertrich et al. [24] introduced the following inertial variant of a stochastic PALM algorithm with a variance-reduced gradient estimator, called SiPALM:

$$\begin{cases} u_{1k} = x_k + \alpha_{1k}(x_k - x_{k-1}),\ v_{1k} = x_k + \beta_{1k}(x_k - x_{k-1}), \\ x_{k+1} \in \arg\min_{x\in\mathbb{R}^l}\{f(x) + \langle x, \widetilde{\nabla}_x(v_{1k}, y_k)\rangle + \frac{1}{2\lambda_k}\|x - u_{1k}\|_2^2\}, \\ u_{2k} = y_k + \alpha_{2k}(y_k - y_{k-1}),\ v_{2k} = y_k + \beta_{2k}(y_k - y_{k-1}), \\ y_{k+1} \in \arg\min_{y\in\mathbb{R}^m}\{g(y) + \langle y, \widetilde{\nabla}_y(x_{k+1}, v_{2k})\rangle + \frac{1}{2\mu_k}\|y - u_{2k}\|_2^2\}, \end{cases} \tag{1.9}$$

where $\alpha_{1k}, \alpha_{2k}, \beta_{1k}, \beta_{2k} \in [0, 1]$. Also, some variance-reduced gradient estimators are proposed to solve the nonconvex optimization problem. The classical stochastic gradient direction is modified in various ways so as to drive the variance of the gradient estimator towards zero, such as SAG [25], SVRG [26, 27], SAGA [28], and SARAH [29, 30].

In this paper, we combine the inertial technique, Bregman distance, and stochastic gradient estimators to develop a stochastic two-step inertial Bregman proximal alternating linearized minimization (STiBPALM) algorithm to solve the nonconvex optimization problem (1.1). Our contributions are listed as follows:

(1) We propose the STiBPALM algorithm with variance-reduced stochastic gradient estimators to solve the nonconvex optimization problem (1.1). And we show that SAGA and SARAH are variance-reduced gradient estimators (Definition 3.4) in the appendix.
(2) We provide theoretical analysis to show that the proposed algorithm with the variance-reduced stochastic gradient estimator has global convergence under expectation conditions. Under the expectation version of Kurdyka–Łojasiewicz (KŁ) property, the sequence generated by the proposed algorithm converges to a critical point and the general convergence rate is also obtained.
(3) We use several well-studied stochastic gradient estimators (e.g., SGD, SAGA, and SARAH) to test the performance of STiBPALM for sparse nonnegative matrix factorization and blind image-deblurring problems. And compared with some existing algorithms (e.g., PALM, iPALM, SPRING, and SiPALM) in the literature,

we report some preliminary numerical results to demonstrate the effectiveness of the proposed algorithm.

This paper is organized as follows. In Sect. 2, we recall some concepts and important lemmas which will be used in the proof of main results. Section 3 introduces our STiBPALM algorithm in detail. We discuss the convergence behavior of STiBPALM in Sect. 4. In Sect. 5, we perform some numerical experiments and compare the results with other algorithms. We give the specific theoretical analysis to show that SAGA and SARAH have variance-reduced stochastic gradient estimators in the appendix.

2 Preliminaries

In this section, we summarize some useful definitions and lemmas.

Definition 2.1 Let $F:\mathbb{R}^d\to(-\infty,+\infty]$ be a proper and lower semicontinuous function. For $x\in \mathrm{dom}F$, the Fréchet subdifferential of F at x, written $\hat{\partial}F(x)$, is the set of vectors $v\in\mathbb{R}^d$ which satisfy

$$\liminf_{y\to x}\frac{1}{\|x-y\|_2}[F(y)-F(x)-\langle v,y-x\rangle]\geq 0.$$

If $x\notin \mathrm{dom}F$, then $\hat{\partial}F(x)=\emptyset$. The limiting-subdifferential, or simply the subdifferential for short, of F at $x\in\mathrm{dom}F$, written $\partial F(x)$, is defined as follows:

$$\partial F(x):=\{v\in\mathbb{R}^d:\exists x_k\to x,\ F(x_k)\to F(x),\ v_k\in\hat{\partial}F(x_k),\ v_k\to v\}.$$

Remark 2.1 (a) The above definition implies that $\hat{\partial}F(x)\subseteq\partial F(x)$ for each $x\in\mathbb{R}^d$, where the first set is convex and closed while the second one is closed. (see [31]).

(b) (Closedness of ∂F) Let $\{x_k\}_{k\in\mathbb{N}}$ and $\{v_k\}_{k\in\mathbb{N}}$ be sequences in $\mathbb{R}^d$ such that $v_k\in\partial F(x_k)$ for all $k\in\mathbb{N}$. If $(x_k,v_k)\to(x,v)$ and $F(x_k)\to F(x)$ as $k\to\infty$, then $v\in\partial F(x)$.

(c) If $F:\mathbb{R}^d\to(-\infty,+\infty]$ be a proper and lower semicontinuous and $H:\mathbb{R}^d\to\mathbb{R}$ is a continuously differentiable function, then $\partial(F+H)(x)=\partial F(x)+\nabla H(x)$ for all $x\in\mathbb{R}^d$.

(d) A necessary (but not sufficient) condition for $x\in\mathbb{R}^d$ to be a minimizer of F is

$$0\in\partial F(x).$$

A point satisfying $0\in\partial F(x)$ is called limiting-critical or simply critical. The set of critical points of F is denoted by $\mathrm{crit}F$.

Definition 2.2 (Kurdyka–Łojasiewicz property [12]) Let $F:\mathbb{R}^d\to(-\infty,+\infty]$ be a proper and lower semicontinuous function.

(i) The function $F:\mathbb{R}^d\to(-\infty,+\infty]$ is said to have the Kurdyka–Łojasiewicz (KŁ) property at $x^*\in\mathrm{dom}F$ if there exist $\eta\in(0,+\infty]$, a neighborhood U of x^* and a continuous concave function $\varphi:[0,\eta)\to\mathbb{R}_+$ such that $\varphi(0)=0$, φ is C^1

on $(0, \eta)$, for all $s \in (0, \eta)$, it is $\varphi'(s) > 0$, and for all x in $U \cap [F(x^*) < F < F(x^*) + \eta]$, the Kurdyka–Łojasiewicz inequality holds

$$\varphi'(F(x) - F(x^*))\mathrm{dist}(0, \partial F(x)) \geq 1.$$

(ii) Proper lower semicontinuous functions which satisfy the Kurdyka–Łojasiewicz inequality at each point of the domain of its subdifferential are called Kurdyka–Łojasiewicz (KŁ) functions.

Roughly speaking, KŁ functions become sharp up to reparameterization via φ, a desingularizing function for F. Typical KŁ functions include the class of semialgebraic functions [32, 33]. For instance, the l_0 pseudonorm and the rank function are KŁ. Semialgebraic functions admit desingularizing functions of the form $\varphi(r) = ar^{1-\vartheta}$ for $a > 0$, and $\vartheta \in [0, 1)$ is known as the KŁ exponent of the function [11, 32]. For these functions, the KŁ inequality reads

$$(F(x) - F(x^*))^{\vartheta} \leq C \|\xi\|, \quad \forall \xi \in \partial F(x) \tag{2.1}$$

for some $C > 0$.

Definition 2.3 A function F is said convex if dom$F$ is a convex set and if, for all x, y ∈dom$F$, $\alpha \in [0, 1]$,

$$F(\alpha x + (1 - \alpha)y) \leq \alpha F(x) + (1 - \alpha)F(y).$$

F is said θ-strongly convex with $\theta > 0$ if $F - \frac{\theta}{2}\| \cdot \|^2$ is convex, i.e.,

$$F(\alpha x + (1 - \alpha)y) \leq \alpha F(x) + (1 - \alpha)F(y) - \frac{1}{2}\theta\alpha(1 - \alpha)\|x - y\|^2$$

for all x, y ∈dom$F$ and $\alpha \in [0, 1]$.

Suppose that the function F is differentiable. Then, F is convex if and only if dom$F$ is a convex set and

$$F(x) \geq F(y) + \langle \nabla F(y), x - y \rangle$$

holds for all x, y ∈dom$F$. Moreover, F is θ-strongly convex with $\theta > 0$ if and only if

$$F(x) \geq F(y) + \langle \nabla F(y), x - y \rangle + \frac{\theta}{2}\|x - y\|^2$$

for all x, y ∈dom$F$.

Definition 2.4 Let $\phi : \mathbb{R}^d \to (-\infty, +\infty]$ be a convex and Gâteaux differentiable function. The function D_ϕ : domϕ × intdom$\phi \to [0, +\infty)$, defined by

$$D_\phi(x, y) = \phi(x) - \phi(y) - \langle \nabla\phi(y), x - y \rangle,$$

is called the Bregman distance with respect to ϕ.

From the above definition, it follows that

$$D_\phi(x, y) \geq \frac{\theta}{2}\|x - y\|^2, \tag{2.2}$$

if ϕ is θ-strongly convex.

Lemma 2.1 (Descent lemma[34]) *Let $F : \mathbb{R}^d \to \mathbb{R}$ be a continuously differentiable function with gradient ∇F assumed L-Lipschitz continuous. Then,*

$$|F(y) - F(x) - \langle y - x, \nabla F(x)\rangle| \leq \frac{L}{2}\|x - y\|^2, \ \forall x, y \in \mathbb{R}^d. \tag{2.3}$$

Lemma 2.2 *Let $F : \mathbb{R}^d \to \mathbb{R}$ be a function with L-Lipschitz continuous gradient, $G : \mathbb{R}^d \to \mathbb{R}$ a proper lower semicontinuous function, and $z \in \arg\min_{v\in\mathbb{R}^d}\{G(v) + \langle d, v - x\rangle + D_\phi(v, x) + \gamma\langle v, u\rangle + \mu\langle v, w\rangle\}$, where D_ϕ denotes the Bregman distance with respect to ϕ, and x, d, u, $w \in \mathbb{R}^d$. Then, for all $y \in \mathbb{R}^d$,*

$$\begin{aligned} F(z) + G(z) \leq & F(y) + G(y) + \langle \nabla F(x) - d, z - y\rangle + \frac{L}{2}\|x - y\|^2 + D_\phi(y, x) \\ & + \frac{L}{2}\|z - x\|^2 - D_\phi(z, x) + \gamma\langle y - z, u\rangle + \mu\langle y - z, w\rangle. \end{aligned} \tag{2.4}$$

Proof By Lemma 2.1, we have the inequalities

$$\begin{aligned} F(x) - F(y) &\leq \langle \nabla F(x), x - y\rangle + \frac{L}{2}\|x - y\|^2, \\ F(z) - F(x) &\leq \langle \nabla F(x), z - x\rangle + \frac{L}{2}\|z - x\|^2, \end{aligned}$$

which implies that

$$F(z) \leq F(y) + \langle \nabla F(x), z - y\rangle + \frac{L}{2}\|x - y\|^2 + \frac{L}{2}\|z - x\|^2. \tag{2.5}$$

Furthermore, by the definition of z, taking $v = y$, we obtain

$$\begin{aligned} & G(z) + \langle d, z - x\rangle + D_\phi(z, x) + \gamma\langle z, u\rangle + \mu\langle z, w\rangle \\ \leq & G(y) + \langle d, y - x\rangle + D_\phi(y, x) + \gamma\langle y, u\rangle + \mu\langle y, w\rangle, \end{aligned}$$

which implies that

$$G(z) \leq G(y) + \langle d, y - z\rangle + D_\phi(y, x) - D_\phi(z, x) + \gamma\langle y - z, u\rangle + \mu\langle y - z, w\rangle. \tag{2.6}$$

Adding (2.5) and (2.6) completes the proof. $\square$

Lemma 2.3 (sufficient decrease property) *Let F, G, and z be defined as in Lemma 2.2, where $x,\ d,\ u,\ w\ \in\ \mathbb{R}^d$. Assume that ϕ is θ-strongly convex. Then, the following inequality holds, for any $\lambda > 0$,*

$$F(z) + G(z) \leq F(x) + G(x) + \frac{1}{2L\lambda}\|d - \nabla F(x)\|^2 + \frac{L(\lambda+1)-\theta}{2}\|x - z\|^2 + \gamma\langle x - z, u\rangle + \mu\langle x - z, w\rangle. \tag{2.7}$$

Proof From Lemma 2.2 with $y = x$, we have

$$F(z) + G(z) \leq F(x) + G(x) + \langle \nabla F(x) - d, z - x\rangle + \frac{L}{2}\|x - z\|^2 - D_\phi(z, x) + \gamma\langle x - z, u\rangle + \mu\langle x - z, w\rangle.$$

Using Young's inequality $\langle \nabla F(x) - d, z - x\rangle \leq \frac{1}{2L\lambda}\|d - \nabla F(x)\|^2 + \frac{L\lambda}{2}\|x - z\|^2$ and (2.2), we can obtain

$$F(z) + G(z) \leq F(x) + G(x) + \frac{1}{2L\lambda}\|d - \nabla F(x)\|^2 + \frac{L\lambda}{2}\|x - z\|^2 + \frac{L}{2}\|x - z\|^2 - \frac{\theta}{2}\|z - x\|^2 + \gamma\langle x - z, u\rangle + \mu\langle x - z, w\rangle,$$

which can be abbreviated as the desired result. □

3 Stochastic two-step inertial Bregman proximal alternating linearized minimization algorithm

Throughout this paper, we impose the following assumptions.

Assumption 3.1 (i) The function Φ is bounded from below, i.e., $\Phi(x, y) \geq \underline{\Phi}$.

(ii) For any fixed y, the partial gradient $\nabla_x H_i(\cdot, y)$ is globally Lipschitz with module L_y for all $i \in \{1, \ldots, n\}$, that is,

$$\|\nabla_x H_i(x_1, y) - \nabla_x H_i(x_2, y)\| \leq L_y\|x_1 - x_2\|,\ \forall x_1, x_2 \in \mathbb{R}^l.$$

Likewise, for any fixed x, the partial gradient $\nabla_y H_i(x, \cdot)$ is globally Lipschitz with module L_x,

$$\left\|\nabla_y H_i(x, y_1) - \nabla_y H_i(x, y_2)\right\| \leq L_x\|y_1 - y_2\|,\ \forall y_1, y_2 \in \mathbb{R}^m.$$

(iii) ∇H is Lipschitz continuous on bounded subsets of $\mathbb{R}^l \times \mathbb{R}^m$. In other words, for each bounded subset $B_1 \times B_2$ of $\mathbb{R}^l \times \mathbb{R}^m$, there exists $M_{B_1\times B_2} > 0$ such that

$$\left\|\left(\nabla_x H(x_1, y_1) - \nabla_x H(x_2, y_2), \nabla_y H(x_1, y_1) - \nabla_y H(x_2, y_2)\right)\right\| \leq M_{B_1\times B_2}\|(x_1 - x_2, y_1 - y_2)\|.$$

for all $(x_1, y_1), (x_2, y_2) \in B_1 \times B_2$.

(iv) $\phi_i (i = 1, 2)$ is θ_i-strongly convex differentiable function. And the gradient $\nabla\phi_i$ is η_i-Lipschitz continuous, i.e.,

$$\|\nabla\phi_1(x_1) - \nabla\phi_1(x_2)\| \leq \eta_1\|x_1 - x_2\|, \ \forall x_1, x_2 \in \mathbb{R}^l,$$
$$\|\nabla\phi_2(y_1) - \nabla\phi_2(y_2)\| \leq \eta_2\|y_1 - y_2\|, \ \ \forall y_1, y_2 \in \mathbb{R}^m.$$

We now introduce a stochastic version of the two-step inertial Bregman proximal alternating linearized minimization algorithm. The key of our algorithm is replacing the full gradient computations $\nabla_x H(u_k, y_k)$ and $\nabla_y(x_{k+1}, v_k)$ with stochastic estimations $\widetilde{\nabla}_x(u_k, y_k)$ and $\widetilde{\nabla}_y(x_{k+1}, v_k)$, respectively. We describe the resulting algorithm as follows.

Algorithm 3.1 Choose $(x_0, y_0) \in$domΦ and set $(x_{-i}, y_{-i}) = (x_0, y_0)$, $i = 1, 2$. Take the sequences $\{\gamma_{1k}\}$, $\{\mu_{1k}\} \subseteq [0, \gamma_1]$, $\{\gamma_{2k}\}$, $\{\mu_{2k}\} \subseteq [0, \gamma_2]$, $\{\alpha_{1k}\}$, $\{\beta_{1k}\} \subseteq [0, \alpha_1]$ and $\{\alpha_{2k}\}$, $\{\beta_{2k}\} \subseteq [0, \alpha_2]$, where $\gamma_1 \geq 0$, $\gamma_2 \geq 0$, $\alpha_1 \geq 0$ and $\alpha_2 \geq 0$. For $k \geq 0$, let

$$\begin{cases} u_k = x_k + \gamma_{1k}(x_k - x_{k-1}) + \gamma_{2k}(x_{k-1} - x_{k-2}), \\ x_{k+1} \in \arg\min_{x\in\mathbb{R}^l}\{f(x) + \langle x, \widetilde{\nabla}_x(u_k, y_k)\rangle + D_{\phi_1}(x, x_k) + \alpha_{1k}\langle x, x_{k-1} - x_k\rangle \\ \qquad\qquad + \alpha_{2k}\langle x, x_{k-2} - x_{k-1}\rangle\}, \\ v_k = y_k + \mu_{1k}(y_k - y_{k-1}) + \mu_{2k}(y_{k-1} - y_{k-2}), \\ y_{k+1} \in \arg\min_{y\in\mathbb{R}^m}\{g(y) + \langle y, \widetilde{\nabla}_y(x_{k+1}, v_k)\rangle + D_{\phi_2}(y, y_k) + \beta_{1k}\langle y, y_{k-1} - y_k\rangle \\ \qquad\qquad + \beta_{2k}\langle y, y_{k-2} - y_{k-1}\rangle\}, \end{cases} \tag{3.1}$$

where D_{ϕ_1} and D_{ϕ_2} denote the Bregman distance with respect to ϕ_1 and ϕ_2, respectively.

Stochastic gradients $\widetilde{\nabla}_x(u_k, y_k)$ and $\widetilde{\nabla}_y(x_{k+1}, v_k)$ use the gradients of only a few indices $\nabla_x H_i(u_k, y_k)$ and $\nabla_y H_i(x_{k+1}, v_k)$ for $i \in B_k \subset \{1, 2, \ldots, n\}$. The minibatch B_k is chosen uniformly at random from all subsets of $\{1, 2, \ldots, n\}$ with cardinality b. The simplest one is the stochastic gradient descent (SGD) estimator [35]. While the SGD estimator is not variance-reduced, many popular gradient estimators as the SAGA [28] and SARAH [29, 30] estimators have this property. In this paper, we mainly consider SAGA (Appendix A) and SARAH (Appendix B) gradient estimators.

Definition 3.1 (SGD [35]) The SGD gradient estimator $\widetilde{\nabla}_x^{SGD}(x_k, y_k)$ is defined as follows:

$$\widetilde{\nabla}_x^{SGD}(x_k, y_k) = \frac{1}{b}\sum_{i\in B_k} \nabla_x H_i(x_k, y_k),$$

where B_k are mini-batches containing b indices.

The SGD gradient estimator uses the gradient of a randomly sampled batch to represent the full gradient.

Definition 3.2 (SAGA [28]) The SAGA gradient estimator $\widetilde{\nabla}_x^{SAGA}(x_k, y_k)$ is defined as follows:

$$\widetilde{\nabla}_x^{SAGA}(x_k, y_k) = \frac{1}{b} \sum_{i \in B_k} \left(\nabla_x H_i(x_k, y_k) - \nabla_x H_i(\varphi_k^i, y_k) \right) + \frac{1}{n} \sum_{j=1}^{n} \nabla_x H_j(\varphi_k^j, y_k),$$

where B_k are mini-batches containing b indices. The variables φ_k^i follow the update rules $\varphi_{k+1}^i = x_k$ if $i \in B_k$ and $\varphi_{k+1}^i = \varphi_k^i$ otherwise.

Definition 3.3 (SARAH [29, 30]) The SARAH gradient estimator reads for $k = 0$ as

$$\widetilde{\nabla}_x^{SARAH}(x_0, y_0) = \nabla_x H(x_0, y_0).$$

For $k = 1, 2, \ldots,$ we define random variables $p_k \in \{0, 1\}$ with $P(p_k = 0) = \frac{1}{p}$ and $P(p_k = 1) = 1 - \frac{1}{p}$, where $p \in (1, \infty)$ is a fixed chosen parameter. Let B_k be a random subset uniformly drawn from $\{1, \ldots, n\}$ of fixed batch size b. Then, for $k = 1, 2, \ldots,$ the SARAH gradient estimator reads as

$$\begin{aligned} &\widetilde{\nabla}_x^{SARAH}(x_k, y_k) \\ &= \begin{cases} \nabla_x H(x_k, y_k), & \text{if } p_k = 0, \\ \frac{1}{b} \sum_{i \in B_k} (\nabla_x H_i(x_k, y_k) - \nabla_x H_i(x_{k-1}, y_{k-1})) + \widetilde{\nabla}_x^{SARAH}(x_{k-1}, y_{k-1}), & \text{if } p_k = 1. \end{cases} \end{aligned}$$

In our analysis, we assume that stochastic gradient estimator used in Algorithm 3.1 is variance-reduced, which is a quite general assumption in stochastic gradient algorithms [23, 24]. The following definition is analogous to Definition 2.1 in [23].

Definition 3.4 (Variance-reduced gradient estimator) Let $\{z_k\}_{k \in \mathbb{N}} = \{(x_k, y_k)\}_{k \in \mathbb{N}}$ be the sequence generated by Algorithm 3.1 with some gradient estimator $\widetilde{\nabla}$. This gradient estimator is called variance-reduced with constants $V_1, V_2, V_\Upsilon \geq 0$, and $\rho \in (0, 1]$ if it satisfies the following conditions:

(i) (MSE bound) There exists a sequence of random variables $\{\Upsilon_k\}_{k \in \mathbb{N}}$ of the form $\Upsilon_k = \sum_{i=1}^{s} (v_k^i)^2$ for some nonnegative random variables $v_k^i \in \mathbb{R}$ such that

$$\begin{aligned} &\mathbb{E}_k \left[\left\| \widetilde{\nabla}_x(u_k, y_k) - \nabla_x H(u_k, y_k) \right\|^2 + \left\| \widetilde{\nabla}_y(x_{k+1}, v_k) - \nabla_y H(x_{k+1}, v_k) \right\|^2 \right] \\ &\leq \Upsilon_k + V_1 \left(\mathbb{E}_k \|z_{k+1} - z_k\|^2 + \|z_k - z_{k-1}\|^2 + \|z_{k-1} - z_{k-2}\|^2 + \|z_{k-2} - z_{k-3}\|^2 \right), \end{aligned} \tag{3.2}$$

and, with $\Gamma_k = \sum_{i=1}^{s} v_k^i$

$$\begin{aligned} &\mathbb{E}_k \left[\left\| \widetilde{\nabla}_x(u_k, y_k) - \nabla_x H(u_k, y_k) \right\| + \left\| \widetilde{\nabla}_y(x_{k+1}, v_k) - \nabla_y H(x_{k+1}, v_k) \right\| \right] \\ &\leq \Gamma_k + V_2 \left(\mathbb{E}_k \|z_{k+1} - z_k\| + \|z_k - z_{k-1}\| + \|z_{k-1} - z_{k-2}\| + \|z_{k-2} - z_{k-3}\| \right). \end{aligned} \tag{3.3}$$

(ii) (Geometric decay) The sequence $\{\Upsilon_k\}_{k\in\mathbb{N}}$ decays geometrically:

$$\mathbb{E}_k\Upsilon_{k+1} \leq (1-\rho)\Upsilon_k + V_\Upsilon\left(\mathbb{E}_k\|z_{k+1}-z_k\|^2 + \|z_k - z_{k-1}\|^2 + \|z_{k-1}-z_{k-2}\|^2 + \|z_{k-2}-z_{k-3}\|^2\right). \tag{3.4}$$

(iii) (Convergence of estimator) If $\{z_k\}_{k\in\mathbb{N}}$ satisfies $\lim_{k\to\infty}\mathbb{E}\|z_k - z_{k-1}\|^2 = 0$, then $\mathbb{E}\Upsilon_k \to 0$ and $\mathbb{E}\Gamma_k \to 0$.

In the following, if $\{z_k\}_{k\in\mathbb{N}} = \{(x_k, y_k)\}_{k\in\mathbb{N}}$ is the bounded sequence generated by Algorithm 3.1, we assume ∇H is M-Lipschitz continuous on $\{(x_k, y_k)\}_{k\in\mathbb{N}}$.

Assumption 3.2 For the sequences $\{x_k\}_{k\in\mathbb{N}}$ and $\{y_k\}_{k\in\mathbb{N}}$ generated by Algorithm 3.1, there exists $L > 0$ such that

$$\sup\left\{L_{y_k} : k \in \mathbb{N}\right\} \leq L \ \text{ and } \ \sup\left\{L_{x_k} : k \in \mathbb{N}\right\} \leq L,$$

where L_{y_k} and L_{x_k} are the Lipschitz constants for $\nabla_x H_i(\cdot, y_k)$ and $\nabla_y H_i(x_k, \cdot)$, respectively.

Proposition 3.1 *Let $\{z_k\}_{k\in\mathbb{N}} = \{(x_k, y_k)\}_{k\in\mathbb{N}}$ be the bounded sequence generated by Algorithm 3.1. Then, the SAGA gradient estimator is variance-reduced with parameters $V_1 = \frac{16N^2\gamma^2}{b}$, $V_2 = \frac{4N\gamma}{\sqrt{b}}$, $V_\Upsilon = \frac{408nN^2(1+2\gamma_1^2+\gamma_2^2)}{b^2}$ and $\rho = \frac{b}{2n}$, where $N = \max\{M, L\}$, $\gamma = \max\{\gamma_1, \gamma_2\}$. The SARAH estimator is variance-reduced with parameters $V_1 = 6\left(1-\frac{1}{p}\right)M^2(1+2\gamma_1^2+\gamma_2^2)$, $V_2 = M\sqrt{6(1-\frac{1}{p})(1+2\gamma_1^2+\gamma_2^2)}$, $V_\Upsilon = 6\left(1-\frac{1}{p}\right)M^2(1+2\gamma_1^2+\gamma_2^2)$ and $\rho = \frac{1}{p}$.*

See the detailed proof of Proposition 3.1 in Appendix A and B. And the conclusion that SVRG gradient estimator is variance-reduced can be obtained similarly.

Below, we give the supermartingale convergence theorem that will be applied to obtain almost sure convergence of sequences generated by STiBPALM (Algorithm 3.1).

Lemma 3.1 (Supermartingale convergence) *Let $\{X_k\}_{k\in\mathbb{N}}$ and $\{Y_k\}_{k\in\mathbb{N}}$ be sequences of bounded nonnegative random variables such that X_k and Y_k depend only on the first k iterations of Algorithm 3.1. If*

$$\mathbb{E}_k X_{k+1} + Y_k \leq X_k \tag{3.5}$$

for all k, then $\sum_{k=0}^{\infty} Y_k < +\infty$ a.s. and $\{X_k\}$ converges a.s.

4 Convergence analysis under the KŁ property

In this section, under Assumptions 3.1 and 3.2, we prove convergence of the sequence and extend the convergence rates of SPRING to Algorithm 3.1, for semialgebraic function Φ. Given $k \in \mathbb{N}$, define the quantity

$$\begin{aligned}\Psi_k =&\Phi(z_k) + \frac{1}{L\lambda\rho}\Upsilon_k + \left(\frac{V_1 + V_\Upsilon/\rho}{L\lambda} + \frac{\alpha_1+\alpha_2}{2} + \frac{2L(\gamma_1^2+\gamma_2^2)}{\lambda} + 3Z\right)\|z_k - z_{k-1}\|^2 \\ &+\left(\frac{V_1+V_\Upsilon/\rho}{L\lambda} + \frac{\alpha_2}{2} + \frac{2L\gamma_2^2}{\lambda} + 2Z\right)\|z_{k-1}-z_{k-2}\|^2 + \left(\frac{V_1+V_\Upsilon/\rho}{L\lambda} + Z\right)\|z_{k-2}-z_{k-3}\|^2,\end{aligned} \tag{4.1}$$

where $\lambda = \sqrt{\frac{10(V_1+V_\Upsilon/\rho)+4L^2(\gamma_1^2+\gamma_2^2)}{L^2}}$, $Z = \frac{V_1+V_\Upsilon/\rho}{\sqrt{10(V_1+V_\Upsilon/\rho)+4L^2(\gamma_1^2+\gamma_2^2)}} + \epsilon > 0, \epsilon > 0$ is small enough. Our first result guarantees that Ψ_k is decreasing in expectation.

Lemma 4.1 (l_2 summability) *Suppose Assumptions 3.1 and 3.2 hold. Let $\{z_k\}_{k\in\mathbb{N}}$ be the sequence generated by Algorithm 3.1 with variance-reduced gradient estimator, and let*

$$\theta \triangleq \min\{\theta_1, \theta_2\} > L + 2\alpha_1 + 2\alpha_2 + 2\sqrt{10(V_1 + V_\Upsilon/\rho) + 4L^2(\gamma_1^2+\gamma_2^2)} + 6\epsilon,$$

then the following conclusions hold.

(i) Ψ_k *satisfies*

$$\mathbb{E}_k\left[\Psi_{k+1} + \kappa\|z_{k+1}-z_k\|^2 + \epsilon\|z_k - z_{k-1}\|^2 + \epsilon\|z_{k-1}-z_{k-2}\|^2 + Z\|z_{k-2}-z_{k-3}\|^2\right] \le \Psi_k, \tag{4.2}$$

where $\kappa = -\frac{L-\theta}{2} - \alpha_1 - \alpha_2 - \sqrt{10(V_1+V_\Upsilon/\rho)+4L^2(\gamma_1^2+\gamma_2^2)} - 3\epsilon > 0.$

(ii) *The expectation of the squared distance between the iterates is summable:*

$$\sum_{k=0}^{\infty}\mathbb{E}[\|x_{k+1}-x_k\|^2 + \|y_{k+1}-y_k\|^2] = \sum_{k=0}^{\infty}\mathbb{E}\|z_{k+1}-z_k\|^2 < \infty.$$

Proof (i) Applying Lemma 2.3 with $F(\cdot) = H(\cdot, y_k), G(\cdot) = f(\cdot), z = x_{k+1}, x = x_k$, $d = \widetilde{\nabla}_x(u_k, y_k)$, $u = x_{k-1} - x_k$ and $w = x_{k-2} - x_{k-1}$, for any $\lambda > 0$, we have

$$\begin{aligned}&H(x_{k+1}, y_k) + f(x_{k+1}) \\ \le& H(x_k, y_k) + f(x_k) + \frac{1}{2L\lambda}\left\|\widetilde{\nabla}_x(u_k, y_k) - \nabla_x H(x_k, y_k)\right\|^2 + \frac{L(\lambda+1)-\theta_1}{2}\|x_{k+1}-x_k\|^2 \\ &+ \alpha_{1k}\langle x_{k+1}-x_k, x_k - x_{k-1}\rangle + \alpha_{2k}\langle x_{k+1}-x_k, x_{k-1}-x_{k-2}\rangle \\ \overset{(1)}{\le}& H(x_k, y_k) + f(x_k) + \frac{1}{L\lambda}\left\|\widetilde{\nabla}_x(u_k, y_k) - \nabla_x H(u_k, y_k)\right\|^2 + \frac{1}{L\lambda}\|\nabla_x H(u_k, y_k) - \nabla_x H(x_k, y_k)\|^2 \\ &+ \frac{L(\lambda+1)-\theta_1}{2}\|x_{k+1}-x_k\|^2 + \frac{\alpha_{1k}}{2}(\|x_{k+1}-x_k\|^2 + \|x_k - x_{k-1}\|^2) \\ &+ \frac{\alpha_{2k}}{2}(\|x_{k+1}-x_k\|^2 + \|x_{k-1}-x_{k-2}\|^2)\end{aligned}$$

$$\stackrel{(2)}{\leq} H(x_k, y_k) + f(x_k) + \frac{1}{L\lambda} \left\| \widetilde{\nabla}_x(u_k, y_k) - \nabla_x H(u_k, y_k) \right\|^2 + \frac{L}{\lambda} \|u_k - x_k\|^2$$
$$+ \left(\frac{L(\lambda+1) - \theta_1}{2} + \frac{\alpha_1 + \alpha_2}{2} \right) \|x_{k+1} - x_k\|^2 + \frac{\alpha_1}{2} \|x_k - x_{k-1}\|^2 + \frac{\alpha_2}{2} \|x_{k-1} - x_{k-2}\|^2$$
$$\leq H(x_k, y_k) + f(x_k) + \frac{1}{L\lambda} \left\| \widetilde{\nabla}_x(u_k, y_k) - \nabla_x H(u_k, y_k) \right\|^2 + \left(\frac{2L\gamma_{1k}^2}{\lambda} + \frac{\alpha_1}{2} \right) \|x_k - x_{k-1}\|^2$$
$$+ \left(\frac{2L\gamma_{2k}^2}{\lambda} + \frac{\alpha_2}{2} \right) \|x_{k-1} - x_{k-2}\|^2 + \left(\frac{L(\lambda+1) - \theta_1}{2} + \frac{\alpha_1 + \alpha_2}{2} \right) \|x_{k+1} - x_k\|^2. \quad (4.3)$$

Inequality (1) is the standard inequality $\|a - c\|^2 \leq 2\|a - b\|^2 + 2\|b - c\|^2$, and (2) uses Assumption 3.1 (ii) and Assumption 3.2. Analogously, for the updates in y_k, we use Lemma 2.3 with $F(\cdot) = H(x_{k+1}, \cdot)$, $G(\cdot) = g(\cdot)$, $z = y_{k+1}$, $x = y_k$, $d = \widetilde{\nabla}_y(x_{k+1}, v_k)$, $u = y_{k-1} - y_k$ and $w = y_{k-2} - y_{k-1}$, we have

$$H(x_{k+1}, y_{k+1}) + g(y_{k+1})$$
$$\leq H(x_{k+1}, y_k) + g(y_k) + \frac{1}{L\lambda} \left\| \widetilde{\nabla}_y(x_{k+1}, v_k) - \nabla_y H(x_{k+1}, v_k) \right\|^2 + \left(\frac{2L\mu_{1k}^2}{\lambda} + \frac{\alpha_1}{2} \right) \|y_k - y_{k-1}\|^2$$
$$+ \left(\frac{2L\mu_{2k}^2}{\lambda} + \frac{\alpha_2}{2} \right) \|y_{k-1} - y_{k-2}\|^2 + \left(\frac{L(\lambda+1) - \theta_2}{2} + \frac{\alpha_1 + \alpha_2}{2} \right) \|y_{k+1} - y_k\|^2. \quad (4.4)$$

Adding (4.3) and (4.4), we have

$$\Phi(x_{k+1}, y_{k+1})$$
$$\leq \Phi(x_k, y_k) + \frac{1}{L\lambda} \left(\left\| \widetilde{\nabla}_x(u_k, y_k) - \nabla_x H(u_k, y_k) \right\|^2 + \left\| \widetilde{\nabla}_y(x_{k+1}, v_k) - \nabla_y H(x_{k+1}, v_k) \right\|^2 \right)$$
$$+ \left(\frac{L(\lambda+1) - \theta}{2} + \frac{\alpha_1 + \alpha_2}{2} \right) \|z_{k+1} - z_k\|^2 + \left(\frac{2L\gamma_1^2}{\lambda} + \frac{\alpha_1}{2} \right) \|z_k - z_{k-1}\|^2$$
$$+ \left(\frac{2L\gamma_2^2}{\lambda} + \frac{\alpha_2}{2} \right) \|z_{k-1} - z_{k-2}\|^2,$$

where $\theta = \min\{\theta_1, \theta_2\}$. Applying the conditional expectation operator $\mathbb{E}_k$, we can bound the MSE terms using (3.2). This gives

$$\mathbb{E}_k \left[\Phi(z_{k+1}) + \left(-\frac{L(\lambda+1) - \theta}{2} - \frac{\alpha_1 + \alpha_2}{2} - \frac{V_1}{L\lambda} \right) \|z_{k+1} - z_k\|^2 \right]$$
$$\leq \Phi(z_k) + \frac{1}{L\lambda} \Upsilon_k + \left(\frac{V_1}{L\lambda} + \frac{2L\gamma_1^2}{\lambda} + \frac{\alpha_1}{2} \right) \|z_k - z_{k-1}\|^2 + \left(\frac{V_1}{L\lambda} + \frac{2L\gamma_2^2}{\lambda} + \frac{\alpha_2}{2} \right) \|z_{k-1} - z_{k-2}\|^2$$
$$+ \frac{V_1}{L\lambda} \|z_{k-2} - z_{k-3}\|^2. \quad (4.5)$$

Next, we use (3.4) to say that

$$\frac{1}{L\lambda} \Upsilon_k \leq \frac{1}{L\lambda\rho} \left(-\mathbb{E}_k \Upsilon_{k+1} + \Upsilon_k + V_\Upsilon \left(\mathbb{E}_k \|z_{k+1} - z_k\|^2 + \|z_k - z_{k-1}\|^2 \right.\right.$$

$$+\|z_{k-1}-z_{k-2}\|^2+\|z_{k-2}-z_{k-3}\|^2\Big)\Big).$$

Combining these inequalities, we have

$$\begin{aligned}
&\mathbb{E}_k\left[\Phi(z_{k+1})+\frac{1}{L\lambda\rho}\Upsilon_{k+1}+\left(-\frac{L(\lambda+1)-\theta}{2}-\frac{\alpha_1+\alpha_2}{2}-\frac{V_1+V_\Upsilon/\rho}{L\lambda}\right)\|z_{k+1}-z_k\|^2\right]\\
\leq&\Phi(z_k)+\frac{1}{L\lambda\rho}\Upsilon_k+\left(\frac{V_1+V_\Upsilon/\rho}{L\lambda}+\frac{2L\gamma_1^2}{\lambda}+\frac{\alpha_1}{2}\right)\|z_k-z_{k-1}\|^2\\
&+\left(\frac{V_1+V_\Upsilon/\rho}{L\lambda}+\frac{2L\gamma_2^2}{\lambda}+\frac{\alpha_2}{2}\right)\|z_{k-1}-z_{k-2}\|^2+\frac{V_1+V_\Upsilon/\rho}{L\lambda}\|z_{k-2}-z_{k-3}\|^2.
\end{aligned}$$

This is equivalent to

$$\begin{aligned}
&\mathbb{E}_k\left[\Phi(z_{k+1})+\frac{1}{L\lambda\rho}\Upsilon_{k+1}+\left(\frac{V_1+V_\Upsilon/\rho}{L\lambda}+\frac{\alpha_1+\alpha_2}{2}+\frac{2L(\gamma_1^2+\gamma_2^2)}{\lambda}+3Z\right)\|z_{k+1}-z_k\|^2\right.\\
&+\left(\frac{V_1+V_\Upsilon/\rho}{L\lambda}+\frac{\alpha_2}{2}+\frac{2L\gamma_2^2}{\lambda}+2Z\right)\|z_k-z_{k-1}\|^2+\left(\frac{V_1+V_\Upsilon/\rho}{L\lambda}+Z\right)\|z_{k-1}-z_{k-2}\|^2\\
&\left.+\left(-\frac{L(\lambda+1)-\theta}{2}-\frac{2(V_1+V_\Upsilon/\rho)}{L\lambda}-\alpha_1-\alpha_2-\frac{2L(\gamma_1^2+\gamma_2^2)}{\lambda}-3Z\right)\|z_{k+1}-z_k\|^2\right]\\
\leq&\Phi(z_k)+\frac{1}{L\lambda\rho}\Upsilon_k+\left(\frac{V_1+V_\Upsilon/\rho}{L\lambda}+\frac{\alpha_1+\alpha_2}{2}+\frac{2L(\gamma_1^2+\gamma_2^2)}{\lambda}+3Z\right)\|z_k-z_{k-1}\|^2\\
&+\left(\frac{V_1+V_\Upsilon/\rho}{L\lambda}+\frac{\alpha_2}{2}+\frac{2L\gamma_2^2}{\lambda}+2Z\right)\|z_{k-1}-z_{k-2}\|^2+\left(\frac{V_1+V_\Upsilon/\rho}{L\lambda}+Z\right)\|z_{k-2}-z_{k-3}\|^2\\
&-\left(Z-\frac{V_1+V_\Upsilon/\rho}{L\lambda}\right)\|z_k-z_{k-1}\|^2-\left(Z-\frac{V_1+V_\Upsilon/\rho}{L\lambda}\right)\|z_{k-1}-z_{k-2}\|^2-Z\|z_{k-2}-z_{k-3}\|^2.
\end{aligned}\tag{4.6}$$

We have

$$\begin{aligned}
&\mathbb{E}_k\left[\Psi_{k+1}+\left(-\frac{L(\lambda+1)-\theta}{2}-\frac{2(V_1+V_\Upsilon/\rho)}{L\lambda}-\alpha_1-\alpha_2-\frac{2L(\gamma_1^2+\gamma_2^2)}{\lambda}-3Z\right)\|z_{k+1}-z_k\|^2\right]\\
\leq&\Psi_k-\left(Z-\frac{V_1+V_\Upsilon/\rho}{L\lambda}\right)\|z_k-z_{k-1}\|^2-\left(Z-\frac{V_1+V_\Upsilon/\rho}{L\lambda}\right)\|z_{k-1}-z_{k-2}\|^2-Z\|z_{k-2}-z_{k-3}\|^2.
\end{aligned}\tag{4.7}$$

By $\lambda=\sqrt{\frac{10(V_1+V_\Upsilon/\rho)+4L^2(\gamma_1^2+\gamma_2^2)}{L^2}}$, we have $-\frac{L(\lambda+1)-\theta}{2}-\frac{2(V_1+V_\Upsilon/\rho)}{L\lambda}-\alpha_1-\alpha_2-\frac{2L(\gamma_1^2+\gamma_2^2)}{\lambda}-3Z=-\frac{L-\theta}{2}-\alpha_1-\alpha_2-\sqrt{10(V_1+V_\Upsilon/\rho)+4L^2(\gamma_1^2+\gamma_2^2)}-3\epsilon=\kappa$. Hence, (4.7) becomes

$$\mathbb{E}_k\left[\Psi_{k+1}+\kappa\|z_{k+1}-z_k\|^2+\epsilon\|z_k-z_{k-1}\|^2+\epsilon\|z_{k-1}-z_{k-2}\|^2+Z\|z_{k-2}-z_{k-3}\|^2\right]\leq\Psi_k.\tag{4.8}$$

According to $\theta > L + 2\alpha_1 + 2\alpha_2 + 2\sqrt{10(V_1 + V_\Upsilon/\rho) + 4L^2(\gamma_1^2 + \gamma_2^2)} + 6\epsilon$, we have $\kappa > 0$. So we prove the first claim.

(ii) We apply the full expectation operator to (4.8) and sum the resulting inequality from $k = 0$ to $k = T - 1$,

$$\begin{aligned}
&\mathbb{E}\Psi_T + \kappa \sum_{k=0}^{T-1} \mathbb{E}\,\|z_{k+1} - z_k\|^2 + \epsilon \sum_{k=0}^{T-1} \mathbb{E}\,\|z_k - z_{k-1}\|^2 + \epsilon \sum_{k=0}^{T-1} \mathbb{E}\,\|z_{k-1} - z_{k-2}\|^2 \\
&+ Z \sum_{k=0}^{T-1} \mathbb{E}\,\|z_{k-2} - z_{k-3}\|^2 \\
\leq &\Psi_0,
\end{aligned}$$

Using the fact that $\underline{\Phi} \leq \Psi_T$,

$$\begin{aligned}
&\kappa \sum_{k=0}^{T-1} \mathbb{E}\,\|z_{k+1} - z_k\|^2 + \epsilon \sum_{k=0}^{T-1} \mathbb{E}\,\|z_k - z_{k-1}\|^2 + \epsilon \sum_{k=0}^{T-1} \mathbb{E}\,\|z_{k-1} - z_{k-2}\|^2 \\
&+ Z \sum_{k=0}^{T-1} \mathbb{E}\,\|z_{k-2} - z_{k-3}\|^2 \\
\leq &\Psi_0 - \underline{\Phi}.
\end{aligned} \tag{4.9}$$

Taking the limit $T \to +\infty$, we have the sequence $\left\{\mathbb{E}\,\|z_{k+1} - z_k\|^2\right\}$ is summable. □

The next lemma establishes a bound on the norm of the subgradients of $\Phi(z_k)$.

Lemma 4.2 (Subgradient bound) *Suppose Assumptions 3.1 and 3.2 hold. Let* $\{z_k\}_{k\in\mathbb{N}}$ *be a bounded sequence, which is generated by Algorithm 3.1 with variance-reduced gradient estimator. For* $k \geq 0$*, define*

$$\begin{aligned}
A_x^k =&\nabla_x H(x_k, y_k) - \widetilde{\nabla}_x(u_{k-1}, y_{k-1}) + \nabla\phi_1(x_{k-1}) - \nabla\phi_1(x_k) + \alpha_{1,k-1}(x_{k-1} - x_{k-2}) \\
&+ \alpha_{2,k-1}(x_{k-2} - x_{k-3}), \\
A_y^k =&\nabla_y H(x_k, y_k) - \widetilde{\nabla}_y(x_k, v_{k-1}) + \nabla\phi_2(y_{k-1}) - \nabla\phi_2(y_k) + \beta_{1,k-1}(y_{k-1} - y_{k-2}) \\
&+ \beta_{2,k-1}(y_{k-2} - y_{k-3}).
\end{aligned}$$

Then, $(A_x^k, A_y^k) \in \partial\Phi(x_k, y_k)$ *and*

$$\begin{aligned}
&\mathbb{E}_{k-1}\left\|(A_x^k, A_y^k)\right\| \\
\leq\; &p\,(\mathbb{E}_{k-1}\,\|z_k - z_{k-1}\| + \|z_{k-1} - z_{k-2}\| + \|z_{k-2} - z_{k-3}\| + \|z_{k-3} - z_{k-4}\|) + \Gamma_{k-1},
\end{aligned} \tag{4.10}$$

where $p = 2(2N + \eta + N\gamma_1 + N\gamma_2 + \alpha_1 + \alpha_2) + V_2$*,* $N = \max\{M, L\}$*,* $\eta = \max\{\eta_1, \eta_2\}$.

Proof By the definition of x_k, we have that 0 must lie in the subdifferential at point x_k of the function

$$x \longmapsto f(x)+\langle x, \widetilde{\nabla}_x(u_{k-1}, y_{k-1})\rangle+D_{\phi_1}(x, x_{k-1})+\alpha_{1,k-1}\langle x, x_{k-2}-x_{k-1}\rangle+\alpha_{2,k-1}\langle x, x_{k-3}-x_{k-2}\rangle.$$

Since ϕ are differential, we have

$$\begin{aligned} 0 \in\ & \partial f(x_k) + \widetilde{\nabla}_x(u_{k-1}, y_{k-1}) + \nabla\phi_1(x_k) - \nabla\phi_1(x_{k-1}) + \alpha_{1,k-1}(x_{k-2} - x_{k-1}) \\ & +\alpha_{2,k-1}(x_{k-3} - x_{k-2}), \end{aligned}$$

which implies that

$$\begin{aligned} & \nabla_x H(x_k, y_k) - \widetilde{\nabla}_x(u_{k-1}, y_{k-1}) + \nabla\phi_1(x_{k-1}) - \nabla\phi_1(x_k) \\ & + \alpha_{1,k-1}(x_{k-1} - x_{k-2}) + \alpha_{2,k-1}(x_{k-2} - x_{k-3}) \\ & \in \nabla_x H(x_k, y_k) + \partial f(x_k). \end{aligned} \tag{4.11}$$

Similarly, we have

$$\begin{aligned} & \nabla_y H(x_k, y_k) - \widetilde{\nabla}_y(x_k, v_{k-1}) + \nabla\phi_2(y_{k-1}) - \nabla\phi_2(y_k) \\ & + \beta_{1,k-1}(y_{k-1} - y_{k-2}) + \beta_{2,k-1}(y_{k-2} - y_{k-3}) \\ & \in \nabla_y H(x_k, y_k) + \partial g(y_k). \end{aligned} \tag{4.12}$$

Because of the structure of Φ, from (4.11) and (4.12), we have $(A_x^k, A_y^k) \in \partial\Phi(x_k, y_k)$. All that remains is to bound the norms of A_x^k and A_y^k. Because ∇H is M-Lipschitz continuous on bounded sets, then from Assumption 3.1 (iii) and (iv), we have

$$\begin{aligned} & \left\| A_x^k \right\| \\ \le & \left\| \nabla_x H(x_k, y_k) - \widetilde{\nabla}_x(u_{k-1}, y_{k-1}) \right\| + \|\nabla\phi_1(x_{k-1}) - \nabla\phi_1(x_k)\| \\ & + \alpha_{1,k-1}\|x_{k-1} - x_{k-2}\| + \alpha_{2,k-1}\|x_{k-2} - x_{k-3}\| \\ \le & \|\nabla_x H(x_k, y_k) - \nabla_x H(u_{k-1}, y_{k-1})\| + \left\|\nabla_x H(u_{k-1}, y_{k-1}) - \widetilde{\nabla}_x(u_{k-1}, y_{k-1})\right\| \\ & + \eta_1\|x_{k-1} - x_k\| + \alpha_{1,k-1}\|x_{k-1} - x_{k-2}\| + \alpha_{2,k-1}\|x_{k-2} - x_{k-3}\| \\ \le & \left\|\nabla_x H(u_{k-1}, y_{k-1}) - \widetilde{\nabla}_x(u_{k-1}, y_{k-1})\right\| + M\|x_k - u_{k-1}\| + M\|y_k - y_{k-1}\| \\ & + \eta_1\|x_{k-1} - x_k\| + \alpha_{1,k-1}\|x_{k-1} - x_{k-2}\| + \alpha_{2,k-1}\|x_{k-2} - x_{k-3}\| \\ \le & \left\|\nabla_x H(u_{k-1}, y_{k-1}) - \widetilde{\nabla}_x(u_{k-1}, y_{k-1})\right\| + (M+\eta_1)\|x_k - x_{k-1}\| + M\|y_k - y_{k-1}\| \\ & (M\gamma_1 + \alpha_1)\|x_{k-1} - x_{k-2}\| + (M\gamma_2 + \alpha_2)\|x_{k-2} - x_{k-3}\|. \end{aligned} \tag{4.13}$$

A similar argument holds for A_y^k:

$$\begin{aligned} & \left\| A_y^k \right\| \\ \le & \left\|\nabla_y H(x_k, y_k) - \nabla_y H(x_k, v_{k-1})\right\| + \left\|\nabla_y H(x_k, v_{k-1}) - \widetilde{\nabla}_y(x_k, v_{k-1})\right\| \\ & + \eta_2\|y_{k-1} - y_k\| + \beta_{1,k-1}\|y_{k-1} - y_{k-2}\| + \beta_{2,k-1}\|y_{k-2} - y_{k-3}\| \end{aligned}$$

$$\le \left\| \nabla_y H(x_k, v_{k-1}) - \widetilde{\nabla}_y(x_k, v_{k-1}) \right\| + (L + \eta_2) \|y_k - y_{k-1}\|$$
$$(L\gamma_1 + \alpha_1) \|y_{k-1} - y_{k-2}\| + (L\gamma_2 + \alpha_2) \|y_{k-2} - y_{k-3}\| . \tag{4.14}$$

Adding (4.13) and (4.14), we get

$$\left\| A_x^k \right\| + \left\| A_y^k \right\|$$
$$\le \left\| \nabla_x H(u_{k-1}, y_{k-1}) - \widetilde{\nabla}_x(u_{k-1}, y_{k-1}) \right\| + \left\| \nabla_y H(x_k, v_{k-1}) - \widetilde{\nabla}_y(x_k, v_{k-1}) \right\|$$
$$+2(2N+\eta) \|z_k - z_{k-1}\| + 2(N\gamma_1+\alpha_1) \|z_{k-1} - z_{k-2}\| + 2(N\gamma_2+\alpha_2) \|z_{k-2} - z_{k-3}\| ,$$

where $N = \max\{M, L\}$, $\eta = \max\{\eta_1, \eta_2\}$. Applying the conditional expectation operator and using (3.3) to bound the MSE terms, we can obtain

$$\mathbb{E}_{k-1} \left\| (A_x^k, A_y^k) \right\| \le \mathbb{E}_{k-1} \left[\left\| A_x^k \right\| + \left\| A_y^k \right\| \right]$$
$$\le (4N + 2\eta + V_2)\mathbb{E}_{k-1} \|z_k - z_{k-1}\| + (2N\gamma_1 + 2\alpha_1 + V_2) \|z_{k-1} - z_{k-2}\|$$
$$+ (2N\gamma_2 + 2\alpha_2 + V_2) \|z_{k-2} - z_{k-3}\| + V_2 \|z_{k-3} - z_{k-4}\| + \Gamma_{k-1}$$
$$\le p \left(\mathbb{E}_{k-1} \|z_k - z_{k-1}\| + \|z_{k-1} - z_{k-2}\| + \|z_{k-2} - z_{k-3}\| + \|z_{k-3} - z_{k-4}\|\right) + \Gamma_{k-1},$$

where $p = 2(2N + \eta + N\gamma_1 + N\gamma_2 + \alpha_1 + \alpha_2) + V_2$. □

Define the set of limit points of $\{z_k\}_{k\in\mathbb{N}}$ as

$$\Omega := \{\hat{z} : \text{ there exists a subsequence } \left\{z_{k_l}\right\} \text{ of } \{z_k\} \text{ such that } z_{k_l} \to \hat{z} \text{ as } l \to \infty\}.$$

The following lemma describes properties of Ω.

Lemma 4.3 (Limit points of $\{z_k\}_{k\in\mathbb{N}}$) *Suppose Assumptions 3.1 and 3.2 hold. Let $\{z_k\}_{k\in\mathbb{N}}$ be a bounded sequence, which is generated by Algorithm 3.1 with variance-reduced gradient estimator, and let*

$$\theta > L + 2\alpha_1 + 2\alpha_2 + 2\sqrt{10(V_1 + V_\Upsilon/\rho) + 4L^2(\gamma_1^2 + \gamma_2^2)} + 6\epsilon.$$

where $\epsilon > 0$ is small enough. Then,

(1) $\sum_{k=1}^{\infty} \|z_k - z_{k-1}\|^2 < \infty$ *a.s., and* $\|z_k - z_{k-1}\| \to 0$ *a.s.;*
(2) $\mathbb{E}\Phi(z_k) \to \Phi^*$, *where* $\Phi^* \in [\underline{\Phi}, \infty)$*;*
(3) $\mathbb{E}\mathrm{dist}(0, \partial\Phi(z_k)) \to 0$*;*
(4) *the set Ω is nonempty, and for all $z^* \in \Omega$,* $\mathbb{E}\mathrm{dist}(0, \partial\Phi(z^*)) = 0$*;*
(5) $\mathrm{dist}(z_k, \Omega) \to 0$ *a.s.;*
(6) Ω *is a.s. compact and connected;*
(7) $\mathbb{E}\Phi(z^*) = \Phi^*$ *for all $z^* \in \Omega$.*

Proof By Lemma 4.1, we have claim (1) holds.

According to (4.2), the supermartingale convergence theorem ensures $\{\Psi_k\}$ converges to a finite, positive random variable. Because $\|z_k - z_{k-1}\| \to 0$ a.s.,

$\|z_{k-1} - z_{k-2}\| \to 0$ a.s., $\|z_{k-2} - z_{k-3}\| \to 0$ a.s. and $\widetilde{\nabla}$ is variance-reduced so $\mathbb{E}\Upsilon_k \to 0$, we can say

$$\lim_{k\to\infty} \mathbb{E}\Psi_k = \lim_{k\to\infty} \mathbb{E}\Phi(z_k) \in [\underline{\Phi}, \infty),$$

which implys claim (2).

Claim (3) holds because, by Lemma 4.2,

$$\begin{aligned}&\mathbb{E}\left\|(A_x^k, A_y^k)\right\| \\ \leq &p\mathbb{E}\left(\|z_k - z_{k-1}\| + \|z_{k-1} - z_{k-2}\| + \|z_{k-2} - z_{k-3}\| + \|z_{k-3} - z_{k-4}\|\right) + \mathbb{E}\Gamma_{k-1}.\end{aligned}$$

We have that $\mathbb{E}\,\|z_k - z_{k-1}\| \to 0$ and $\mathbb{E}\Gamma_{k-1} \to 0$. This ensures that $\mathbb{E}\left\|(A_x^k, A_y^k)\right\| \to 0$. Since (A_x^k, A_y^k) is one element of $\partial\Phi(z_k)$, we obtain $\mathbb{E}\text{dist}(0, \partial\Phi(z_k)) \leq \mathbb{E}\left\|(A_x^k, A_y^k)\right\| \to 0$.

To prove claim (4), suppose $z^* = (x^*, y^*)$ is a limit point of the sequence $\{z_k\}_{k\in\mathbb{N}}$ (a limit point must exist because we suppose the sequence $\{z_k\}_{k\in\mathbb{N}}$ is bounded). This means there exists a subsequence $\left\{z_{k_j}\right\}$ satisfying $\lim_{j\to\infty} z_{k_j} = z^*$. Furthermore, by the variance-reduced property of $\widetilde{\nabla}(u_{k_j-1}, y_{k_j-1})$, we have $\mathbb{E}\left\|\widetilde{\nabla}_x(u_{k_j-1}, y_{k_j-1}) - \nabla_x H(u_{k_j-1}, y_{k_j-1})\right\|^2 \to 0$.

Because f and g are lower semicontinuous, we have

$$\begin{aligned}&\liminf_{j\to\infty} f(x_{k_j}) \geq f(x^*),\\ &\liminf_{j\to\infty} g(y_{k_j}) \geq g(y^*).\end{aligned} \tag{4.15}$$

By the update rule for x_{k_j}, letting $x = x^*$, we have

$$\begin{aligned}&f(x_{k_j}) + \langle x_{k_j}, \widetilde{\nabla}_x(u_{k_j-1}, y_{k_j-1})\rangle + D_{\phi_1}(x_{k_j}, x_{k_j-1}) + \alpha_{1,k_j-1}\langle x_{k_j}, x_{k_j-2} - x_{k_j-1}\rangle \\ &+ \alpha_{2,k_j-1}\langle x_{k_j}, x_{k_j-3} - x_{k_j-2}\rangle \\ \leq &f(x^*) + \langle x^*, \widetilde{\nabla}_x(u_{k_j-1}, y_{k_j-1})\rangle + D_{\phi_1}(x^*, x_{k_j-1}) + \alpha_{1,k_j-1}\langle x^*, x_{k_j-2} - x_{k_j-1}\rangle \\ &+ \alpha_{2,k_j-1}\langle x^*, x_{k_j-3} - x_{k_j-2}\rangle.\end{aligned}$$

Taking the expectation and taking the limit $j \to \infty$,

$$\begin{aligned}&\limsup_{j\to\infty} f(x_{k_j}) \\ \leq &\limsup_{j\to\infty} f(x^*) + \langle x^* - x_{k_j}, \nabla_x H(u_{k_j-1}, y_{k_j-1})\rangle + \langle x^* - x_{k_j}, \widetilde{\nabla}_x(u_{k_j-1}, y_{k_j-1}) \\ &- \nabla_x H(u_{k_j-1}, y_{k_j-1})\rangle + \phi_1(x^*) - \phi_1(x_{k_j}) + \left\langle \nabla\phi_1(x_{k_j-1}), x^* - x_{k_j-1}\right\rangle\end{aligned}$$

$$+\alpha_{1,k_j-1}\langle x^* - x_{k_j}, x_{k_j-2} - x_{k_j-1}\rangle + \alpha_{2,k_j-1}\langle x^* - x_{k_j}, x_{k_j-3} - x_{k_j-2}\rangle.$$

The second term on the right goes to zero because $x_{k_j} \to x^*$ and $\{\nabla_x H(u_{k_j-1}, y_{k_j-1})\}$ is bounded. The thrid term is zero almost surely because it is bounded above by $\|x^* - x_{k_j}\|^2$, and $\widetilde{\nabla}_x(u_{k_j-1}, y_{k_j-1}) - \nabla_x H(u_{k_j-1}, y_{k_j-1}) \to 0$ a.s. Noting that ϕ_1 is differentiable, so $\limsup_{j\to\infty} f(x_{k_j}) \le f(x^*)$ a.s., which, together with (4.15), implies that $\lim_{j\to\infty} f(x_{k_j}) = f(x^*)$ a.s. Similarly, we have $\lim_{j\to\infty} g(y_{k_j}) = g(y^*)$ a.s., and hence

$$\lim_{j\to\infty} \Phi(x_{k_j}, y_{k_j}) = \Phi(x^*, y^*) \text{ a.s.} \tag{4.16}$$

Claim (3) ensures that $\mathbb{E}\text{dist}(0, \partial\Phi(z_k)) \to 0$. Combining (4.16) and the fact that the subdifferential of Φ is closed, we have $\mathbb{E}\text{dist}(0, \partial\Phi(z^*)) = 0$.

Claims (5) and (6) hold for any sequence satisfying $\|z_k - z_{k-1}\| \to 0$ a.s. (this fact is used in the same context in [11, 36]).

Finally, we must show that Φ has constant expectation over Ω. From claim (2), we have $\mathbb{E}\Phi(z_k) \to \Phi^*$, which implies $\mathbb{E}\Phi(z_{k_j}) \to \Phi^*$ for every subsequence $\{z_{k_j}\}_{j\in\mathbb{N}}$ converging to some $z^* \in \Omega$. In the proof of claim (4), we show that $\Phi(z_{k_j}) \to \Phi(z^*)$ a.s., so $\mathbb{E}\Phi(z^*) = \Phi^*$ for all $z^* \in \Omega$. □

The following lemma is analogous to the uniformized Kurdyka–Łojasiewicz property [11]. It is a slight generalization of the KŁ property showing that z_k eventually enters a region of $\tilde{z}$ for some $\tilde{z}$ satisfying $\Phi(\tilde{z}) = \Phi(z^*)$, and in this region, the KŁ inequality holds.

Lemma 4.4 *Assume that the conditions of Lemma 4.3 hold and that z_k is not a critical point of Φ after a finite number of iterations. Let Φ be a semialgebraic function with KŁ exponent ϑ. Then, there exists an index m and a desingularizing function φ so that the following bound holds:*

$$\varphi'(\mathbb{E}[\Phi(z_k) - \Phi_k^*])\mathbb{E}\text{dist}(0, \partial\Phi(z_k)) \ge 1, \quad \forall k > m,$$

where Φ_k^* is a nondecreasing sequence converging to $\mathbb{E}\Phi(z^*)$ for all $z^* \in \Omega$.

The proof is almost the same as that of Lemma 4.5 in [23]. We omit the proof here. We now show that the iterates of Algorithm 3.1 have finite length in expectation.

Theorem 4.1 (Finite length) *Assume that the conditions of Lemma 4.3 hold and Φ is a semialgebraic function with KŁ exponent $\vartheta \in [0, 1)$. Let $\{z_k\}_{k\in\mathbb{N}}$ be a bounded sequence, which is generated by Algorithm 3.1 with variance-reduced gradient estimator.*

(i) *Either z_k is a critical point after a finite number of iterations or $\{z_k\}_{k\in\mathbb{N}}$ satisfies the finite length property in expectation:*

$$\sum_{k=0}^{\infty} \mathbb{E}\|z_{k+1} - z_k\| < \infty,$$

and there exists an integer m so that, for all $i > m$,

$$\begin{aligned}&\sum_{k=m}^{i}\mathbb{E}\,\|z_{k+1}-z_k\|+\sum_{k=m}^{i}\mathbb{E}\,\|z_k-z_{k-1}\|+\sum_{k=m}^{i}\mathbb{E}\,\|z_{k-1}-z_{k-2}\|+\sum_{k=m}^{i}\mathbb{E}\,\|z_{k-2}-z_{k-3}\|\\&\le\sqrt{\mathbb{E}\,\|z_m-z_{m-1}\|^2}+\sqrt{\mathbb{E}\,\|z_{m-1}-z_{m-2}\|^2}+\sqrt{\mathbb{E}\,\|z_{m-2}-z_{m-3}\|^2}\\&\quad+\sqrt{\mathbb{E}\,\|z_{m-3}-z_{m-4}\|^2}+\frac{2\sqrt{s}}{K_1\rho}\sqrt{\mathbb{E}\Upsilon_{m-1}}+K_3\triangle_{m,i+1},\end{aligned}\tag{4.17}$$

where

$$K_1=p+\frac{2\sqrt{sV_\Upsilon}}{\rho},\ K_3=\frac{4K_1}{K_2},\ K_2=\min\{\kappa,\epsilon,Z\},$$

p is as in Lemma 4.2, and $\triangle_{p,q}=(\mathbb{E}[\Psi_p-\Phi_p^*]-\mathbb{E}[\Psi_q-\Phi_q^*])$.
(ii) *$\{z_k\}_{k\in\mathbb{N}}$ generated by Algorithm 3.1 converge to a critical point of Φ in expectation.*

Proof (i) If $\vartheta\in(0,\frac{1}{2})$, then Φ satisfies the KŁ property with exponent $\frac{1}{2}$, so we consider only the case $\vartheta\in[\frac{1}{2},1)$. By Lemma 4.4, there exists a function $\varphi_0(r)=ar^{1-\vartheta}$ such that

$$\varphi_0'(\mathbb{E}[\Phi(z_k)-\Phi_k^*])\mathbb{E}\mathrm{dist}(0,\partial\Phi(z_k))\ge 1,\ \ \forall k>m.$$

Lemma 4.2 provides a bound on $\mathbb{E}\mathrm{dist}(0,\partial\Phi(z_k))$.

$$\begin{aligned}&\mathbb{E}\mathrm{dist}(0,\partial\Phi(z_k))\le\mathbb{E}\left\|(A_x^k,A_y^k)\right\|\\&\le p\mathbb{E}\left(\|z_k-z_{k-1}\|+\|z_{k-1}-z_{k-2}\|+\|z_{k-2}-z_{k-3}\|+\|z_{k-3}-z_{k-4}\|\right)+\mathbb{E}\Gamma_{k-1}\\&\le p\Big(\sqrt{\mathbb{E}\,\|z_k-z_{k-1}\|^2}+\sqrt{\mathbb{E}\,\|z_{k-1}-z_{k-2}\|^2}+\sqrt{\mathbb{E}\,\|z_{k-2}-z_{k-3}\|^2}\\&\quad+\sqrt{\mathbb{E}\,\|z_{k-3}-z_{k-4}\|^2}\Big)+\sqrt{s\mathbb{E}\Upsilon_{k-1}}.\end{aligned}\tag{4.18}$$

The final inequality is Jensen's inequality. Because $\Gamma_k=\sum_{i=1}^s v_k^i$ for some non-negative random variables v_k^i, we can say $\mathbb{E}\Gamma_k=\mathbb{E}\sum_{i=1}^s v_k^i\le\mathbb{E}\sqrt{s\sum_{i=1}^s(v_k^i)^2}\le\sqrt{s\mathbb{E}\Upsilon_k}$. We can bound the term $\sqrt{\mathbb{E}\Upsilon_k}$ using (3.4):

$$\begin{aligned}&\sqrt{\mathbb{E}\Upsilon_k}\\&\le\sqrt{(1-\rho)\mathbb{E}\Upsilon_{k-1}+V_\Upsilon\mathbb{E}\left(\|z_k-z_{k-1}\|^2+\|z_{k-1}-z_{k-2}\|^2+\|z_{k-2}-z_{k-3}\|^2+\|z_{k-3}-z_{k-4}\|^2\right)}\\&\le\sqrt{(1-\rho)}\sqrt{\mathbb{E}\Upsilon_{k-1}}+\sqrt{V_\Upsilon}\Big(\sqrt{\mathbb{E}\,\|z_k-z_{k-1}\|^2}+\sqrt{\mathbb{E}\,\|z_{k-1}-z_{k-2}\|^2}+\sqrt{\mathbb{E}\,\|z_{k-2}-z_{k-3}\|^2}\\&\quad+\sqrt{\mathbb{E}\,\|z_{k-3}-z_{k-4}\|^2}\Big)\\&\le(1-\frac{\rho}{2})\sqrt{\mathbb{E}\Upsilon_{k-1}}+\sqrt{V_\Upsilon}\Big(\sqrt{\mathbb{E}\,\|z_k-z_{k-1}\|^2}+\sqrt{\mathbb{E}\,\|z_{k-1}-z_{k-2}\|^2}+\sqrt{\mathbb{E}\,\|z_{k-2}-z_{k-3}\|^2}\end{aligned}$$

$$+\sqrt{\mathbb{E}\left\|z_{k-3}-z_{k-4}\right\|^2}\Big). \tag{4.19}$$

The final inequality uses the fact that $\sqrt{1-\rho}=1-\frac{\rho}{2}-\frac{\rho^2}{8}-\cdots$. This implies that

$$\begin{aligned}&\sqrt{s\mathbb{E}\Upsilon_{k-1}}\\ \leq&\frac{2\sqrt{s}}{\rho}\left(\sqrt{\mathbb{E}\Upsilon_{k-1}}-\sqrt{\mathbb{E}\Upsilon_k}\right)+\frac{2\sqrt{sV_\Upsilon}}{\rho}\Big(\sqrt{\mathbb{E}\left\|z_k-z_{k-1}\right\|^2}+\sqrt{\mathbb{E}\left\|z_{k-1}-z_{k-2}\right\|^2}\\ &+\sqrt{\mathbb{E}\left\|z_{k-2}-z_{k-3}\right\|^2}+\sqrt{\mathbb{E}\left\|z_{k-3}-z_{k-4}\right\|^2}\Big).\end{aligned} \tag{4.20}$$

Then, from (4.18) and (4.20), we have

$$\begin{aligned}&\mathbb{E}\mathrm{dist}(0,\partial\Phi(z_k))\\ \leq&\left(p+\frac{2\sqrt{sV_\Upsilon}}{\rho}\right)\Big(\sqrt{\mathbb{E}\left\|z_k-z_{k-1}\right\|^2}+\sqrt{\mathbb{E}\left\|z_{k-1}-z_{k-2}\right\|^2}+\sqrt{\mathbb{E}\left\|z_{k-2}-z_{k-3}\right\|^2}\\ &+\sqrt{\mathbb{E}\left\|z_{k-3}-z_{k-4}\right\|^2}\Big)+\frac{2\sqrt{s}}{\rho}\left(\sqrt{\mathbb{E}\Upsilon_{k-1}}-\sqrt{\mathbb{E}\Upsilon_k}\right)\\ =&K_1\Big(\sqrt{\mathbb{E}\left\|z_k-z_{k-1}\right\|^2}+\sqrt{\mathbb{E}\left\|z_{k-1}-z_{k-2}\right\|^2}+\sqrt{\mathbb{E}\left\|z_{k-2}-z_{k-3}\right\|^2}\\ &+\sqrt{\mathbb{E}\left\|z_{k-3}-z_{k-4}\right\|^2}\Big)+\frac{2\sqrt{s}}{\rho}\left(\sqrt{\mathbb{E}\Upsilon_{k-1}}-\sqrt{\mathbb{E}\Upsilon_k}\right),\end{aligned}$$

where $K_1=p+\frac{2\sqrt{sV_\Upsilon}}{\rho}$. Define C_k to be the right side of this inequality:

$$\begin{aligned}C_k=&K_1\sqrt{\mathbb{E}\left\|z_k-z_{k-1}\right\|^2}+K_1\sqrt{\mathbb{E}\left\|z_{k-1}-z_{k-2}\right\|^2}+K_1\sqrt{\mathbb{E}\left\|z_{k-2}-z_{k-3}\right\|^2}\\ &+K_1\sqrt{\mathbb{E}\left\|z_{k-3}-z_{k-4}\right\|^2}+\frac{2\sqrt{s}}{\rho}\left(\sqrt{\mathbb{E}\Upsilon_{k-1}}-\sqrt{\mathbb{E}\Upsilon_k}\right).\end{aligned}$$

We then have

$$\varphi_0'(\mathbb{E}[\Phi(z_k)-\Phi_k^*])C_k\geq 1,\quad\forall k>m. \tag{4.21}$$

By the definition of φ_0, this is equivalent to

$$\frac{a(1-\vartheta)C_k}{(\mathbb{E}[\Phi(z_k)-\Phi_k^*])^{\vartheta}}\geq 1,\quad\forall k>m. \tag{4.22}$$

We would like to hold the inequality above for Ψ_k rather than $\Phi(z_k)$. Replace $\mathbb{E}\Phi(z_k)$ with $\mathbb{E}\Psi_k$ by introducing a term of $\mathcal{O}\left(\left(\mathbb{E}\left[\|z_k-z_{k-1}\|^2+\|z_{k-1}-z_{k-2}\|^2+\|z_{k-2}-z_{k-3}\|^2+\Upsilon_k\right]\right)^{\vartheta}\right)$ in the denominator. We show that inequality (4.22) still holds after this adjustment because these terms are small compared to C_k. Indeed, the quantity

$$C_k \geq c_1 \left(\sqrt{\mathbb{E}\,\|z_k - z_{k-1}\|^2} + \sqrt{\mathbb{E}\,\|z_{k-1} - z_{k-2}\|^2} + \sqrt{\mathbb{E}\,\|z_{k-2} - z_{k-3}\|^2} \right.$$
$$\left. + \sqrt{\mathbb{E}\,\|z_{k-3} - z_{k-4}\|^2} + \sqrt{\mathbb{E}\Upsilon_{k-1}} \right)$$

for some constant $c_1 > 0$. And because $\mathbb{E}\,\|z_k - z_{k-1}\|^2 \to 0$, $\mathbb{E}\Upsilon_k \to 0$, and $\vartheta > \frac{1}{2}$, there exists an index m and constants $c_2, c_3 > 0$ such that

$$\begin{aligned}
&(\mathbb{E}[\Psi_k - \Phi(z_k)])^{\vartheta} \\
&= \left(\mathbb{E}\left[\frac{1}{L\lambda\rho}\Upsilon_k + \left(\frac{V_1 + V_\Upsilon/\rho}{L\lambda} + \frac{\alpha_1 + \alpha_2}{2} + \frac{2L(\gamma_1^2 + \gamma_2^2)}{\lambda} + 3Z \right) \|z_k - z_{k-1}\|^2 + \left(\frac{V_1 + V_\Upsilon/\rho}{L\lambda} \right.\right.\right. \\
&\quad \left.\left.\left. + \frac{\alpha_2}{2} + \frac{2L\gamma_2^2}{\lambda} + 2Z \right) \|z_{k-1} - z_{k-2}\|^2 + \left(\frac{V_1 + V_\Upsilon/\rho}{L\lambda} + Z \right) \|z_{k-2} - z_{k-3}\|^2 \right] \right)^{\vartheta} \\
&\leq c_2 \left(\left(\mathbb{E}\left[\Upsilon_{k-1} + \|z_k - z_{k-1}\|^2 + \|z_{k-1} - z_{k-2}\|^2 + \|z_{k-2} - z_{k-3}\|^2 + \|z_{k-3} - z_{k-4}\|^2 \right] \right)^{\vartheta} \right) \\
&\leq c_3 C_k, \quad \forall k > m.
\end{aligned}$$

The first inequality uses (3.4). Because the terms above are small compared to C_k, there exists a constant d such that $c_3 < d < +\infty$ and

$$\frac{ad(1-\vartheta)C_k}{(\mathbb{E}[\Phi(z_k) - \Phi_k^*])^{\vartheta} + (\mathbb{E}[\Psi_k - \Phi(z_k)])^{\vartheta}} \geq 1, \quad \forall k > m.$$

For $\vartheta \in [\frac{1}{2}, 1)$, using the fact that $(a+b)^{\vartheta} \leq a^{\vartheta} + b^{\vartheta}$ for all $a, b \geq 0$, we have

$$\begin{aligned}
\frac{ad(1-\vartheta)C_k}{\left(\mathbb{E}[\Psi_k - \Phi_k^*]\right)^{\vartheta}} &= \frac{ad(1-\vartheta)C_k}{\left(\mathbb{E}[\Phi(z_k) - \Phi_k^* + \Psi_k - \Phi(z_k)]\right)^{\vartheta}} \\
&\geq \frac{ad(1-\vartheta)C_k}{\left(\mathbb{E}[\Phi(z_k) - \Phi_k^*]\right)^{\vartheta} + (\mathbb{E}[\Psi_k - \Phi(z_k)])^{\vartheta}} \\
&\geq 1, \quad \forall k > m.
\end{aligned}$$

Therefore, with $\varphi(r) = adr^{1-\vartheta}$,

$$\varphi'(\mathbb{E}[\Psi_k - \Phi_k^*])C_k \geq 1, \quad \forall k > m. \tag{4.23}$$

By the concavity of φ,

$$\begin{aligned}
\varphi(\mathbb{E}[\Psi_k - \Phi_k^*]) - \varphi(\mathbb{E}[\Psi_{k+1} - \Phi_{k+1}^*]) &\geq \varphi'(\mathbb{E}[\Psi_k - \Phi_k^*])(\mathbb{E}[\Psi_k - \Phi_k^* + \Phi_{k+1}^* - \Psi_{k+1}]) \\
&\geq \varphi'(\mathbb{E}[\Psi_k - \Phi_k^*])(\mathbb{E}[\Psi_k - \Psi_{k+1}]),
\end{aligned}$$

where the last inequality follows from the fact that Φ_k^* is nondecreasing. With $\triangle_{p,q} = \varphi(\mathbb{E}[\Psi_p - \Phi_p^*]) - \varphi(\mathbb{E}[\Psi_q - \Phi_q^*])$, we have shown

$$\triangle_{k,k+1} C_k \geq \mathbb{E}[\Psi_k - \Psi_{k+1}], \ \forall k > m.$$

Using Lemma 4.1, we can bound $\mathbb{E}[\Psi_k - \Psi_{k+1}]$ below by both $\mathbb{E}\,\|z_{k+1} - z_k\|^2$, $\mathbb{E}\,\|z_k - z_{k-1}\|^2$, $\mathbb{E}\,\|z_{k-1} - z_{k-2}\|^2$ and $\mathbb{E}\,\|z_{k-2} - z_{k-3}\|^2$. Specifically,

$$\begin{aligned}\triangle_{k,k+1} C_k &\geq \kappa \mathbb{E}\,\|z_{k+1} - z_k\|^2 + \epsilon \mathbb{E}\,\|z_k - z_{k-1}\|^2 + \epsilon \mathbb{E}\,\|z_{k-1} - z_{k-2}\|^2 + Z \mathbb{E}\,\|z_{k-2} - z_{k-3}\|^2 \\ &\geq K_2 \mathbb{E}\,\|z_{k+1} - z_k\|^2 + K_2 \mathbb{E}\,\|z_k - z_{k-1}\|^2 + K_2 \mathbb{E}\,\|z_{k-1} - z_{k-2}\|^2 + K_2 \mathbb{E}\,\|z_{k-2} - z_{k-3}\|^2,\end{aligned} \tag{4.24}$$

where $K_2 = \min\{\kappa, \epsilon, Z\} > 0$, κ, λ, ϵ and Z are set as in Lemma 4.1. Let us use the first of these inequalities to begin. Applying Young's inequality to (4.24) yields

$$\begin{aligned}&\sqrt{\mathbb{E}\,\|z_{k+1} - z_k\|^2} + \sqrt{\mathbb{E}\,\|z_k - z_{k-1}\|^2} + \sqrt{\mathbb{E}\,\|z_{k-1} - z_{k-2}\|^2} + \sqrt{\mathbb{E}\,\|z_{k-2} - z_{k-3}\|^2} \\ &\leq 2\sqrt{\mathbb{E}\,\|z_{k+1} - z_k\|^2 + \mathbb{E}\,\|z_k - z_{k-1}\|^2 + \mathbb{E}\,\|z_{k-1} - z_{k-2}\|^2 + \mathbb{E}\,\|z_{k-2} - z_{k-3}\|^2} \\ &\leq 2\sqrt{K_2^{-1} C_k \triangle_{k,k+1}} \leq \frac{C_k}{2K_1} + \frac{2K_1 \triangle_{k,k+1}}{K_2} \\ &\leq \frac{1}{2}\sqrt{\mathbb{E}\,\|z_k - z_{k-1}\|^2} + \frac{1}{2}\sqrt{\mathbb{E}\,\|z_{k-1} - z_{k-2}\|^2} + \frac{1}{2}\sqrt{\mathbb{E}\,\|z_{k-2} - z_{k-3}\|^2} \\ &\quad + \frac{1}{2}\sqrt{\mathbb{E}\,\|z_{k-3} - z_{k-4}\|^2} + \frac{\sqrt{s}}{K_1 \rho}\left(\sqrt{\mathbb{E}\Upsilon_{k-1}} - \sqrt{\mathbb{E}\Upsilon_k}\right) + \frac{2K_1 \triangle_{k,k+1}}{K_2}.\end{aligned} \tag{4.25}$$

Summing inequality (4.25) from $k = m$ to $k = i$, set

$$\begin{aligned}T_m^i = &\sum_{k=m}^{i} \sqrt{\mathbb{E}\,\|z_{k+1} - z_k\|^2} + \sum_{k=m}^{i} \sqrt{\mathbb{E}\,\|z_k - z_{k-1}\|^2} + \sum_{k=m}^{i} \sqrt{\mathbb{E}\,\|z_{k-1} - z_{k-2}\|^2} \\ &+ \sum_{k=m}^{i} \sqrt{\mathbb{E}\,\|z_{k-2} - z_{k-3}\|^2}.\end{aligned} \tag{4.26}$$

Then,

$$T_m^i \leq \frac{1}{2} T_{m-1}^{i-1} + \frac{\sqrt{s}}{K_1 \rho}\left(\sqrt{\mathbb{E}\Upsilon_{m-1}} - \sqrt{\mathbb{E}\Upsilon_i}\right) + \frac{2K_1}{K_2} \triangle_{m,i+1},$$

which implies that

$$\begin{aligned}\frac{1}{2} T_m^i \leq &\frac{1}{2}\sqrt{\mathbb{E}\,\|z_m - z_{m-1}\|^2} + \frac{1}{2}\sqrt{\mathbb{E}\,\|z_{m-1} - z_{m-2}\|^2} + \frac{1}{2}\sqrt{\mathbb{E}\,\|z_{m-2} - z_{m-3}\|^2} \\ &+ \frac{1}{2}\sqrt{\mathbb{E}\,\|z_{m-3} - z_{m-4}\|^2} + \frac{\sqrt{s}}{K_1 \rho}\left(\sqrt{\mathbb{E}\Upsilon_{m-1}} - \sqrt{\mathbb{E}\Upsilon_i}\right) + \frac{2K_1}{K_2} \triangle_{m,i+1}.\end{aligned}$$

Dropping the nonpositive term $-\sqrt{\mathbb{E}\Upsilon_i}$, this shows that

$$T_m^i \leq \sqrt{\mathbb{E}\left\|z_m - z_{m-1}\right\|^2} + \sqrt{\mathbb{E}\left\|z_{m-1} - z_{m-2}\right\|^2} + \sqrt{\mathbb{E}\left\|z_{m-2} - z_{m-3}\right\|^2} + \sqrt{\mathbb{E}\left\|z_{m-3} - z_{m-4}\right\|^2} + \frac{2\sqrt{s}}{K_1\rho}\sqrt{\mathbb{E}\Upsilon_{m-1}} + K_3\triangle_{m,i+1}. \tag{4.27}$$

where $K_3 = \frac{4K_1}{K_2}$. Applying Jensen's inequality to the terms on the left gives

$$\sum_{k=m}^{i}\mathbb{E}\left\|z_{k+1}-z_k\right\| + \sum_{k=m}^{i}\mathbb{E}\left\|z_k - z_{k-1}\right\| + \sum_{k=m}^{i}\mathbb{E}\left\|z_{k-1} - z_{k-2}\right\| + \sum_{k=m}^{i}\mathbb{E}\left\|z_{k-2} - z_{k-3}\right\| \leq T_m^i$$
$$\leq \sqrt{\mathbb{E}\left\|z_m - z_{m-1}\right\|^2} + \sqrt{\mathbb{E}\left\|z_{m-1} - z_{m-2}\right\|^2} + \sqrt{\mathbb{E}\left\|z_{m-2} - z_{m-3}\right\|^2} + \sqrt{\mathbb{E}\left\|z_{m-3} - z_{m-4}\right\|^2} + \frac{2\sqrt{s}}{K_1\rho}\sqrt{\mathbb{E}\Upsilon_{m-1}} + K_3\triangle_{m,i+1}.$$

The term $\lim_{i\to\infty}\triangle_{m,i+1}$ is bounded because $\mathbb{E}\Psi_k$ is bounded due to Lemma 4.1. Letting $i \to \infty$, we prove the assertion.

(ii) An immediate consequence of claim (i) is that the sequence $\{z_k\}_{k\in\mathbb{N}}$ converges in expectation to a critical point. This is because, for any $p, q \in \mathbb{N}$ with $p \geq q$, $\mathbb{E}\left\|z_p - z_q\right\| = \mathbb{E}\left\|\sum_{k=q}^{p-1}(z_{k+1} - z_k)\right\| \leq \sum_{k=q}^{p-1}\mathbb{E}\left\|z_{k+1} - z_k\right\|$, and the finite length property implies this final sum converges to zero. This proves claim (ii). □

Theorem 4.2 *Assume that the conditions of Lemma 4.3 hold and Φ is a semialgebraic function with KŁ exponent $\vartheta \in [0, 1)$. Let $\{z_k\}_{k\in\mathbb{N}}$ be a bounded sequence, which is generated by Algorithm 3.1 with variance-reduced gradient estimator. The following convergence rates hold:*

(i) *If $\vartheta \in (0, \frac{1}{2}]$, then there exist $d_1 > 0$ and $\tau \in [1-\rho, 1)$ such that $\mathbb{E}\left\|z_k - z^*\right\| \leq d_1\tau^k$.*
(ii) *If $\vartheta \in (\frac{1}{2}, 1)$, then there exists a constant $d_2 > 0$ such that $\mathbb{E}\left\|z_k - z^*\right\| \leq d_2 k^{-\frac{1-\vartheta}{2\vartheta-1}}$.*
(iii) *If $\vartheta = 0$, then there exists an $m \in \mathbb{N}$ such that $\mathbb{E}\Phi(z_k) = \mathbb{E}\Phi(z^*)$ for all $k \geq m$.*

Proof As in the proof of Theorem 4.1, if $\vartheta \in (0, \frac{1}{2})$, then Φ satisfies the KŁ property with exponent $\frac{1}{2}$, so we consider only the case $\vartheta \in [\frac{1}{2}, 1)$.
Let

$$T_m = \sum_{k=m}^{\infty}\sqrt{\mathbb{E}\left\|z_{k+1} - z_k\right\|^2} + \sum_{k=m}^{\infty}\sqrt{\mathbb{E}\left\|z_k - z_{k-1}\right\|^2} + \sum_{k=m}^{\infty}\sqrt{\mathbb{E}\left\|z_{k-1} - z_{k-2}\right\|^2} + \sum_{k=m}^{\infty}\sqrt{\mathbb{E}\left\|z_{k-2} - z_{k-3}\right\|^2}.$$

Substituting the desingularizing function $\varphi(r) = ar^{1-\vartheta}$ into (4.27), let $i \to \infty$, then we have

$$T_m \leq \sqrt{\mathbb{E}\,\|z_m - z_{m-1}\|^2} + \sqrt{\mathbb{E}\,\|z_{m-1} - z_{m-2}\|^2} + \sqrt{\mathbb{E}\,\|z_{m-2} - z_{m-3}\|^2} + \sqrt{\mathbb{E}\,\|z_{m-3} - z_{m-4}\|^2} + \frac{2\sqrt{s}}{K_1\rho}\sqrt{\mathbb{E}\Upsilon_{m-1}} + aK_3(\mathbb{E}[\Psi_m - \Phi_m^*])^{1-\vartheta}. \quad (4.28)$$

Because $\Psi_m = \Phi(z_m) + \mathcal{O}(\|z_m - z_{m-1}\|^2 + \|z_{m-1} - z_{m-2}\|^2 + \|z_{m-2} - z_{m-3}\|^2 + \Upsilon_m)$, we can rewrite the final term as $\Phi(z_m) - \Phi_m^*$.

$$\begin{aligned}
&(\mathbb{E}[\Psi_m - \Phi_m^*])^{1-\vartheta} \\
&= \Bigg(\mathbb{E}\Bigg[\Phi(z_m) - \Phi_m^* + \frac{1}{L\lambda\rho}\Upsilon_k + \left(\frac{V_1 + V_\Upsilon/\rho}{L\lambda} + \frac{\alpha_1+\alpha_2}{2} + \frac{2L(\gamma_1^2+\gamma_2^2)}{\lambda} + 3Z\right)\|z_m - z_{m-1}\|^2 \\
&\quad + \left(\frac{V_1 + V_\Upsilon/\rho}{L\lambda} + \frac{\alpha_2}{2} + \frac{2L\gamma_2^2}{\lambda} + 2Z\right)\|z_{m-1} - z_{m-2}\|^2 + \left(\frac{V_1 + V_\Upsilon/\rho}{L\lambda} + Z\right) \\
&\quad \|z_{m-2} - z_{m-3}\|^2\Bigg]\Bigg)^{1-\vartheta} \\
&\overset{(1)}{\leq} \left(\mathbb{E}[\Phi(z_m) - \Phi_m^*]\right)^{1-\vartheta} + \left(\frac{1}{L\lambda\rho}\mathbb{E}\Upsilon_m\right)^{1-\vartheta} + \Bigg(\left(\frac{V_1+V_\Upsilon/\rho}{L\lambda} + \frac{\alpha_1+\alpha_2}{2} + \frac{2L(\gamma_1^2+\gamma_2^2)}{\lambda} + 3Z\right) \\
&\quad \mathbb{E}\|z_m - z_{m-1}\|^2\Bigg)^{1-\vartheta} + \left(\left(\frac{V_1 + V_\Upsilon/\rho}{L\lambda} + \frac{\alpha_2}{2} + \frac{2L\gamma_2^2}{\lambda} + 2Z\right)\mathbb{E}\|z_{m-1} - z_{m-2}\|^2\right)^{1-\vartheta} \\
&\quad + \left(\left(\frac{V_1 + V_\Upsilon/\rho}{L\lambda} + Z\right)\mathbb{E}\|z_{m-2} - z_{m-3}\|^2\right)^{1-\vartheta}. \qquad (4.29)
\end{aligned}$$

Inequality (1) is due to the fact that $(a + b)^{1-\vartheta} \leq a^{1-\vartheta} + b^{1-\vartheta}$. Applying the KŁ inequality (2.1),

$$aK_3\left(\mathbb{E}[\Phi(z_m) - \Phi_m^*]\right)^{1-\vartheta} \leq aK_4\left(\mathbb{E}\,\|\xi_m\|\right)^{\frac{1-\vartheta}{\vartheta}} \qquad (4.30)$$

for all $\xi_m \in \partial\Phi(z_m)$ and we have absorbed the constant C into K_4. Inequality (4.18) provides a bound on the norm of the subgradient:

$$(\mathbb{E}\,\|\xi_m\|)^{\frac{1-\vartheta}{\vartheta}} \leq \left(p\left(\sqrt{\mathbb{E}\,\|z_m - z_{m-1}\|^2} + \sqrt{\mathbb{E}\,\|z_{m-1} - z_{m-2}\|^2} + \sqrt{\mathbb{E}\,\|z_{m-2} - z_{m-3}\|^2} + \sqrt{\mathbb{E}\,\|z_{m-3} - z_{m-4}\|^2}\right) + \sqrt{s\mathbb{E}\Upsilon_{m-1}}\right)^{\frac{1-\vartheta}{\vartheta}}.$$

Let

$$\Theta_m = p\left(\sqrt{\mathbb{E}\,\|z_m - z_{m-1}\|^2} + \sqrt{\mathbb{E}\,\|z_{m-1} - z_{m-2}\|^2} + \sqrt{\mathbb{E}\,\|z_{m-2} - z_{m-3}\|^2} + \sqrt{\mathbb{E}\,\|z_{m-3} - z_{m-4}\|^2}\right) + \sqrt{s\mathbb{E}\Upsilon_{m-1}}.$$

Therefore, it follows from (4.28) to (4.30) that

$$
\begin{aligned}
T_m \leq & \sqrt{\mathbb{E}\,\|z_m - z_{m-1}\|^2} + \sqrt{\mathbb{E}\,\|z_{m-1} - z_{m-2}\|^2} + \sqrt{\mathbb{E}\,\|z_{m-2} - z_{m-3}\|^2} \\
& + \sqrt{\mathbb{E}\,\|z_{m-3} - z_{m-4}\|^2} + \frac{2\sqrt{s}}{K_1\rho}\sqrt{\mathbb{E}\Upsilon_{m-1}} + aK_4\Theta_m^{\frac{1-\vartheta}{\vartheta}} + aK_3\left(\frac{1}{L\lambda\rho}\mathbb{E}\Upsilon_m\right)^{1-\vartheta} \\
& + aK_3\left(\left(\frac{V_1 + V_\Upsilon/\rho}{L\lambda} + \frac{\alpha_1+\alpha_2}{2} + \frac{2L(\gamma_1^2+\gamma_2^2)}{\lambda} + 3Z\right)\mathbb{E}\|z_m - z_{m-1}\|^2\right)^{1-\vartheta} \\
& + aK_3\left(\left(\frac{V_1 + V_\Upsilon/\rho}{L\lambda} + \frac{\alpha_2}{2} + \frac{2L\gamma_2^2}{\lambda} + 2Z\right)\mathbb{E}\|z_{m-1} - z_{m-2}\|^2\right)^{1-\vartheta} \\
& + aK_3\left(\left(\frac{V_1 + V_\Upsilon/\rho}{L\lambda} + Z\right)\mathbb{E}\|z_{m-2} - z_{m-3}\|^2\right)^{1-\vartheta}. \qquad (4.31)
\end{aligned}
$$

(i) If $\vartheta = \frac{1}{2}$, then $(\mathbb{E}\,\|\xi_m\|)^{\frac{1-\vartheta}{\vartheta}} = \mathbb{E}\,\|\xi_m\|$. Equation (4.31) then gives

$$
\begin{aligned}
T_m \leq & \sqrt{\mathbb{E}\,\|z_m - z_{m-1}\|^2} + \sqrt{\mathbb{E}\,\|z_{m-1} - z_{m-2}\|^2} + \sqrt{\mathbb{E}\,\|z_{m-2} - z_{m-3}\|^2} + \sqrt{\mathbb{E}\,\|z_{m-3} - z_{m-4}\|^2} \\
& + \frac{2\sqrt{s}}{K_1\rho}\sqrt{\mathbb{E}\Upsilon_{m-1}} + aK_4\left(p\left(\sqrt{\mathbb{E}\,\|z_m - z_{m-1}\|^2} + \sqrt{\mathbb{E}\,\|z_{m-1} - z_{m-2}\|^2}\right.\right. \\
& \left.\left. + \sqrt{\mathbb{E}\,\|z_{m-2} - z_{m-3}\|^2} + \sqrt{\mathbb{E}\,\|z_{m-3} - z_{m-4}\|^2}\right) + \sqrt{s\mathbb{E}\Upsilon_{m-1}}\right) + aK_3\sqrt{\frac{1}{L\lambda\rho}}\sqrt{\mathbb{E}\Upsilon_m} \\
& + \left(aK_3\sqrt{\frac{V_1 + V_\Upsilon/\rho}{L\lambda} + \frac{\alpha_1+\alpha_2}{2} + \frac{2L(\gamma_1^2+\gamma_2^2)}{\lambda} + 3Z}\right)\sqrt{\mathbb{E}\|z_m - z_{m-1}\|^2} \\
& + \left(aK_3\sqrt{\frac{V_1 + V_\Upsilon/\rho}{L\lambda} + \frac{\alpha_2}{2} + \frac{2L\gamma_2^2}{\lambda} + 2Z}\right)\sqrt{\mathbb{E}\|z_{m-1} - z_{m-2}\|^2} \\
& + \left(aK_3\sqrt{\frac{V_1 + V_\Upsilon/\rho}{L\lambda} + Z}\right)\sqrt{\mathbb{E}\|z_{m-2} - z_{m-3}\|^2} \\
\leq & \left(1 + aK_5\left(p + \sqrt{\frac{V_1 + V_\Upsilon/\rho}{L\lambda} + \frac{\alpha_1+\alpha_2}{2} + \frac{2L(\gamma_1^2+\gamma_2^2)}{\lambda} + 3Z}\right)\right)\left(\sqrt{\mathbb{E}\,\|z_m - z_{m-1}\|^2}\right. \\
& \left. + \sqrt{\mathbb{E}\,\|z_{m-1} - z_{m-2}\|^2} + \sqrt{\mathbb{E}\,\|z_{m-2} - z_{m-3}\|^2} + \sqrt{\mathbb{E}\,\|z_{m-3} - z_{m-4}\|^2}\right) \\
& + \left(\frac{2\sqrt{s}}{K_1\rho} + aK_5\sqrt{s}\right)\sqrt{\mathbb{E}\Upsilon_{m-1}} + aK_5\sqrt{\frac{1}{L\lambda\rho}}\sqrt{\mathbb{E}\Upsilon_m}, \qquad (4.32)
\end{aligned}
$$

where $K_5 = \max\{K_3, K_4\}$. Using (4.19), we have that, for any constant $c > 0$,

$$
0 \leq -c\sqrt{\mathbb{E}\Upsilon_k} + c(1 - \frac{\rho}{2})\sqrt{\mathbb{E}\Upsilon_{k-1}} + c\sqrt{V_\Upsilon}\left(\sqrt{\mathbb{E}\,\|z_k - z_{k-1}\|^2} + \sqrt{\mathbb{E}\,\|z_{k-1} - z_{k-2}\|^2}\right.
$$

$$+\sqrt{\mathbb{E}\left\|z_{k-2}-z_{k-3}\right\|^{2}}+\sqrt{\mathbb{E}\left\|z_{k-3}-z_{k-4}\right\|^{2}}\bigg).$$

Combining this inequality with (4.32),

$$\begin{aligned}T_m \leq & \left(1+aK_5\left(p+\sqrt{\frac{V_1+V_\Upsilon/\rho}{L\lambda}+\frac{\alpha_1+\alpha_2}{2}+\frac{2L(\gamma_1^2+\gamma_2^2)}{\lambda}+3Z+c\sqrt{V_\Upsilon}}\right)\right)\left(\sqrt{\mathbb{E}\left\|z_m-z_{m-1}\right\|^2}\right.\\&\left.+\sqrt{\mathbb{E}\left\|z_{m-1}-z_{m-2}\right\|^2}+\sqrt{\mathbb{E}\left\|z_{m-2}-z_{m-3}\right\|^2}+\sqrt{\mathbb{E}\left\|z_{m-3}-z_{m-4}\right\|^2}\right)\\&+c\left(1-\frac{\rho}{2}+\frac{2\sqrt{s}}{K_1\rho c}+\frac{aK_5\sqrt{s}}{c}\right)\sqrt{\mathbb{E}\Upsilon_{m-1}}-c\left(1-\frac{aK_5}{c}\sqrt{\frac{1}{L\lambda\rho}}\right)\sqrt{\mathbb{E}\Upsilon_m}.\end{aligned}$$

Defining $A=1+aK_5\left(p+\sqrt{\frac{V_1+V_\Upsilon/\rho}{L\lambda}+\frac{\alpha_1+\alpha_2}{2}+\frac{2L(\gamma_1^2+\gamma_2^2)}{\lambda}+3Z+c\sqrt{V_\Upsilon}}\right)$, we have shown

$$\begin{aligned}&T_m+c\left(1-\frac{aK_5}{c}\sqrt{\frac{1}{L\lambda\rho}}\right)\sqrt{\mathbb{E}\Upsilon_m}\\&\leq A\left(T_{m-1}-T_m\right)+c\left(1-\frac{\rho}{2}+\frac{2\sqrt{s}}{K_1\rho c}+\frac{aK_5\sqrt{s}}{c}\right)\sqrt{\mathbb{E}\Upsilon_{m-1}}.\end{aligned}$$

Then, we get

$$\begin{aligned}&(1+A)T_m+c\left(1-\frac{aK_5}{c}\sqrt{\frac{1}{L\lambda\rho}}\right)\sqrt{\mathbb{E}\Upsilon_m}\\&\leq AT_{m-1}+c\left(1-\frac{\rho}{2}+\frac{2\sqrt{s}}{K_1\rho c}+\frac{aK_5\sqrt{s}}{c}\right)\sqrt{\mathbb{E}\Upsilon_{m-1}}.\end{aligned}$$

This implies

$$\begin{aligned}&T_m+\sqrt{\mathbb{E}\Upsilon_m}\\&\leq\max\left\{\frac{A}{1+A},\left(1-\frac{\rho}{2}+\frac{2\sqrt{s}}{K_1\rho c}+\frac{aK_5\sqrt{s}}{c}\right)\left(1-\frac{aK_5}{c}\sqrt{\frac{1}{L\lambda\rho}}\right)^{-1}\right\}\left(T_{m-1}+\sqrt{\mathbb{E}\Upsilon_{m-1}}\right).\end{aligned}$$

For large c, the second coefficient in the above expression approaches $1-\frac{\rho}{2}$. So there exist $\tau\in[1-\rho,1)$ such that

$$\sum_{k=m}^{\infty}\sqrt{\mathbb{E}\left\|z_k-z_{k-1}\right\|^2}\leq\tau^k\left(T_0+\sqrt{\mathbb{E}\Upsilon_0}\right)\leq d_1\tau^k$$

for some constant d_1. Then, using the fact that $\mathbb{E}\,\|z_m - z^*\| = \mathbb{E}\,\left\|\sum_{k=m+1}^{\infty}(z_k - z_{k-1})\right\| \le \sum_{k=m}^{\infty}\mathbb{E}\,\|z_k - z_{k-1}\|$, we prove claim (i).

(ii) Suppose $\vartheta \in (\frac{1}{2}, 1)$. Each term on the right side of (4.31) converges to zero, but at different rates. Because

$$\Theta_m = \mathcal{O}\left(\sqrt{\mathbb{E}\,\|z_m - z_{m-1}\|^2} + \sqrt{\mathbb{E}\,\|z_{m-1} - z_{m-2}\|^2} + \sqrt{\mathbb{E}\,\|z_{m-2} - z_{m-3}\|^2} + \sqrt{\mathbb{E}\,\|z_{m-3} - z_{m-4}\|^2} + \sqrt{s\mathbb{E}\Upsilon_{m-1}}\right),$$

and ϑ satisfies $\frac{1-\vartheta}{\vartheta} < 1$, the term $\Theta_m^{\frac{1-\vartheta}{\vartheta}}$ dominates the first five terms on the right side of (4.31) for large m. Also, because $\frac{1-\vartheta}{2\vartheta} < 1-\vartheta$, $\Theta_m^{\frac{1-\vartheta}{\vartheta}}$ dominates the final four terms as well. Combining these facts, there exists a natural number M_1 such that for all $m \ge M_1$,

$$T_m \le P\Theta_m \tag{4.33}$$

for some constant $P > (aK_3)^{\frac{\vartheta}{1-\vartheta}}$. The bound of (4.20) implies

$$\begin{aligned} &2\sqrt{s\mathbb{E}\Upsilon_{m-1}} \\ &\le \frac{4\sqrt{s}}{\rho}\left(\sqrt{\mathbb{E}\Upsilon_{m-1}} - \sqrt{\mathbb{E}\Upsilon_m} + \sqrt{V_\Upsilon}\left(\sqrt{\mathbb{E}\,\|z_m - z_{m-1}\|^2} + \sqrt{\mathbb{E}\,\|z_{m-1} - z_{m-2}\|^2}\right.\right. \\ &\left.\left.+\sqrt{\mathbb{E}\,\|z_{m-2} - z_{m-3}\|^2} + \sqrt{\mathbb{E}\,\|z_{m-3} - z_{m-4}\|^2}\right)\right). \end{aligned}$$

Therefore,

$$\begin{aligned} \Theta_m =& p\left(\sqrt{\mathbb{E}\,\|z_m - z_{m-1}\|^2} + \sqrt{\mathbb{E}\,\|z_{m-1} - z_{m-2}\|^2} + \sqrt{\mathbb{E}\,\|z_{m-2} - z_{m-3}\|^2}\right. \\ &\left.+\sqrt{\mathbb{E}\,\|z_{m-3} - z_{m-4}\|^2}\right) + \left(2\sqrt{s\mathbb{E}\Upsilon_{m-1}} - \sqrt{s\mathbb{E}\Upsilon_{m-1}}\right) \\ \le& \left(p + \frac{4\sqrt{sV_\Upsilon}}{\rho}\right)\left(\sqrt{\mathbb{E}\,\|z_m - z_{m-1}\|^2} + \sqrt{\mathbb{E}\,\|z_{m-1} - z_{m-2}\|^2} + \sqrt{\mathbb{E}\,\|z_{m-2} - z_{m-3}\|^2}\right. \\ &\left.+\sqrt{\mathbb{E}\,\|z_{m-3} - z_{m-4}\|^2}\right) + \frac{4\sqrt{s}}{\rho}\left(\sqrt{\mathbb{E}\Upsilon_{m-1}} - \sqrt{\mathbb{E}\Upsilon_m}\right) - \sqrt{s\mathbb{E}\Upsilon_{m-1}}. \end{aligned} \tag{4.34}$$

Furthermore, because $\frac{\vartheta}{1-\vartheta} > 1$ and $\mathbb{E}\Upsilon_m \to 0$, for large enough m, we have $\left(\sqrt{\mathbb{E}\Upsilon_m}\right)^{\frac{\vartheta}{1-\vartheta}} \ll \sqrt{\mathbb{E}\Upsilon_m}$. This ensures that there exists a natural number M_2 such that for every $m \ge M_2$,

$$\left(\frac{4\sqrt{s}(1-\rho/4)}{\rho(p + 4\sqrt{sV_\Upsilon}/\rho)}\sqrt{\mathbb{E}\Upsilon_m}\right)^{\frac{\vartheta}{1-\vartheta}} \le P\sqrt{s\mathbb{E}\Upsilon_m}. \tag{4.35}$$

The constant appearing on the left was chosen to simplify later arguments. Therefore, (4.33) implies

$$\begin{aligned}
&\left(T_m + \frac{4\sqrt{s}(1-\rho/4)}{\rho(p+4\sqrt{sV_\Upsilon}/\rho)}\sqrt{\mathbb{E}\Upsilon_m}\right)^{\frac{\vartheta}{1-\vartheta}}\\
&\overset{(1)}{\le}\frac{2^{\frac{\vartheta}{1-\vartheta}}}{2}(T_m)^{\frac{\vartheta}{1-\vartheta}}+\frac{2^{\frac{\vartheta}{1-\vartheta}}}{2}\left(\frac{4\sqrt{s}(1-\rho/4)}{\rho(p+4\sqrt{sV_\Upsilon}/\rho)}\sqrt{\mathbb{E}\Upsilon_m}\right)^{\frac{\vartheta}{1-\vartheta}}\overset{(2)}{\le}\frac{2^{\frac{\vartheta}{1-\vartheta}}}{2}(T_m)^{\frac{\vartheta}{1-\vartheta}}+\frac{2^{\frac{\vartheta}{1-\vartheta}}}{2}\left(P\sqrt{s\mathbb{E}\Upsilon_m}\right)\\
&\overset{(3)}{\le}\frac{2^{\frac{\vartheta}{1-\vartheta}}}{2}\left(P\left(p+\frac{4\sqrt{sV_\Upsilon}}{\rho}\right)\left(\sqrt{\mathbb{E}\|z_m-z_{m-1}\|^2}+\sqrt{\mathbb{E}\|z_{m-1}-z_{m-2}\|^2}+\sqrt{\mathbb{E}\|z_{m-2}-z_{m-3}\|^2}\right.\right.\\
&\quad\left.\left.+\sqrt{\mathbb{E}\|z_{m-3}-z_{m-4}\|^2}\right)+\frac{4\sqrt{s}P}{\rho}\left(\sqrt{\mathbb{E}\Upsilon_{m-1}}-\sqrt{\mathbb{E}\Upsilon_m}\right)-P\sqrt{s\mathbb{E}\Upsilon_{m-1}}\right)+\frac{2^{\frac{\vartheta}{1-\vartheta}}}{2}\left(P\sqrt{s\mathbb{E}\Upsilon_m}\right)\\
&\le\frac{2^{\frac{\vartheta}{1-\vartheta}}}{2}\left(P\left(p+\frac{4\sqrt{sV_\Upsilon}}{\rho}\right)\left(\sqrt{\mathbb{E}\|z_m-z_{m-1}\|^2}+\sqrt{\mathbb{E}\|z_{m-1}-z_{m-2}\|^2}+\sqrt{\mathbb{E}\|z_{m-2}-z_{m-3}\|^2}\right.\right.\\
&\quad\left.\left.+\sqrt{\mathbb{E}\|z_{m-3}-z_{m-4}\|^2}\right)+\frac{4\sqrt{s}P(1-\rho/4)}{\rho}\left(\sqrt{\mathbb{E}\Upsilon_{m-1}}-\sqrt{\mathbb{E}\Upsilon_m}\right)\right).
\end{aligned}$$

Here, (1) follows by convexity of the function $x^{\frac{\vartheta}{1-\vartheta}}$ for $\vartheta\in[1/2,1)$ and $x\ge 0$, (2) is (4.35), and (3) is (4.33) combined with (4.34). We absorb the constant $\frac{2^{\frac{\vartheta}{1-\vartheta}}}{2}$ into P. Define

$$S_m = T_m + \frac{4\sqrt{s}(1-\rho/4)}{\rho(p+4\sqrt{sV_\Upsilon}/\rho)}\sqrt{\mathbb{E}\Upsilon_m}.$$

S_m is bounded for all m because $\sum_{k=m}^{\infty}\sqrt{\mathbb{E}\|z_{k+1}-z_k\|^2}$ is bounded by (4.28). Hence, we have shown

$$S_m^{\frac{\vartheta}{1-\vartheta}} \le P\left(p+\frac{4\sqrt{sV_\Upsilon}}{\rho}\right)(S_{m-1}-S_m). \tag{4.36}$$

The rest of the proof is almost the same as what was mentioned in [23, 37]. We omit the proof here. (iii) When $\vartheta=0$, the KŁ property (2.1) implies that exactly one of the following two scenarios holds: either $\mathbb{E}\Phi(z_k)\neq\Phi_k^*$ and

$$0 < C \le \mathbb{E}\|\xi_k\|, \quad \forall \xi_k\in\partial\Phi(z_k) \tag{4.37}$$

or $\mathbb{E}\Phi(z_k)=\Phi_k^*$. We show that the above inequality can hold only for a finite number of iterations.

Using the subgradient bound (4.10), the first scenario implies

$$\begin{aligned}
C^2 &\le (\mathbb{E}\|\xi_k\|)^2\\
&\le\left(p\left(\mathbb{E}\|z_k-z_{k-1}\|+\mathbb{E}\|z_{k-1}-z_{k-2}\|+\mathbb{E}\|z_{k-2}-z_{k-3}\|+\mathbb{E}\|z_{k-3}-z_{k-4}\|\right)+\Gamma_{k-1}\right)^2\\
&\le 5p^2\left(\mathbb{E}\|z_k-z_{k-1}\|\right)^2+5p^2\left(\mathbb{E}\|z_{k-1}-z_{k-2}\|\right)^2+5p^2\left(\mathbb{E}\|z_{k-2}-z_{k-3}\|\right)^2\\
&\quad+5p^2\left(\mathbb{E}\|z_{k-3}-z_{k-4}\|\right)^2+5(\mathbb{E}\Gamma_{k-1})^2\\
&\le 5p^2\left(\mathbb{E}\|z_k-z_{k-1}\|\right)^2+5p^2\left(\mathbb{E}\|z_{k-1}-z_{k-2}\|\right)^2+5p^2\left(\mathbb{E}\|z_{k-2}-z_{k-3}\|\right)^2\\
&\quad+5p^2\left(\mathbb{E}\|z_{k-3}-z_{k-4}\|\right)^2+5s\mathbb{E}\Upsilon_{k-1},
\end{aligned}$$

where we have used the inequality $(a_1 + a_2 + \cdots + a_s)^2 \leq s(a_1^2 + a_2^2 + \cdots + a_s^2)$ and Jensen's inequality. Applying this inequality to the decrease of Ψ_k (4.2), we obtain

$$\begin{aligned}
&\mathbb{E}_k \Psi_k \\
\leq &\mathbb{E}_k \Psi_{k-1} - \kappa \|z_{k+1} - z_k\|^2 - \epsilon \|z_k - z_{k-1}\|^2 - \epsilon \|z_{k-1} - z_{k-2}\|^2 - Z \|z_{k-2} - z_{k-3}\|^2 \\
\leq &\mathbb{E}_k \Psi_{k-1} - C^2 + \mathcal{O}\left(\|z_{k+1} - z_k\|^2\right) + \mathcal{O}\left(\|z_k - z_{k-1}\|^2\right) + \mathcal{O}\left(\|z_{k-1} - z_{k-2}\|^2\right) \\
&+ \mathcal{O}\left(\|z_{k-2} - z_{k-3}\|^2\right) + \mathcal{O}\left(\mathbb{E}\Upsilon_{k-1}\right)
\end{aligned}$$

for some constant C^2. Because the final five terms go to zero as $k \to \infty$, there exists an index M_4 so that the sum of these five terms is bounded above by $\frac{C^2}{2}$ for all $k \geq M_4$. Therefore,

$$\mathbb{E}_k \Psi_k \leq \mathbb{E}_k \Psi - \frac{C^2}{2}, \quad \forall k \geq M_4.$$

Because Ψ_k is bounded below for all k, this inequality can only hold for $N < \infty$ steps. After N steps, it is no longer possible for the bound (4.37) to hold, so it must be that $\mathbb{E}\Phi(z_k) = \Phi_k^*$. Because $\Phi_k^* < \Phi(z^*)$, $\Phi_k^* < \mathbb{E}\Phi(z_k)$, and both $\mathbb{E}\Phi(z_k)$, Φ_k^* converge to $\mathbb{E}\Phi(z^*)$, we must have $\Phi_k^* = \mathbb{E}\Phi(z_k) = \mathbb{E}\Phi(z^*)$. □

5 Numerical experiments

In this section, to demonstrate the advantages of STiBPALM (Algorithm 3.1), we present our numerical study on the practical performance of the proposed STiBPALM with three different stochastic gradient estimators, i.e., SGD estimator [35] (STiBPALM-SGD), SAGA gradient [28] estimator (STiBPALM-SAGA), and SARAH gradient [29] estimator (STiBPALM-SARAH), compared with PALM [11], iPALM [6], TiPALM [17], SPRING [23], and SiPALM [24] algorithms. We refer to SPRING with SGD, SAGA, and SARAH gradient estimators as SPRING-SGD, SPRING-SAGA, and SPRING-SARAH; and SiPALM using the SGD, SAGA, and SARAH gradient estimators as SiPALM-SGD, SiPALM-SAGA, and SiPALM-SARAH, respectively. Two applications are considered here for comparison: sparse nonnegative matrix factorization (S-NMF) and blind image-deblurring (BID).

Since the proposed algorithm is based on the stochastic gradient estimator, we report the average results (over 10 independent runs) of objective values for all algorithms. The initial point is also the same for all algorithms. In addition, we choose step size which is suggested in [11] for PALM and in [6] for iPALM, respectively, and the same step size based on [23] for all stochastic algorithms for simplicity.

Fig. 1 ORL face database which includes 400 normalized cropped frontal faces which we used in our S-NMF example

5.1 Sparse nonnegative matrix factorization

Given a matrix A, sparse nonnegative matrix factorization (S-NMF) [38–40] problem can be formulated as the following model:

$$\min_{X,Y} \left\{ \frac{\eta}{2} \|A - XY\|_F^2 : \; X, Y \geq 0, \; \|X_i\|_0 \leq s, \; i = 1, 2, \ldots, r \right\}. \tag{5.1}$$

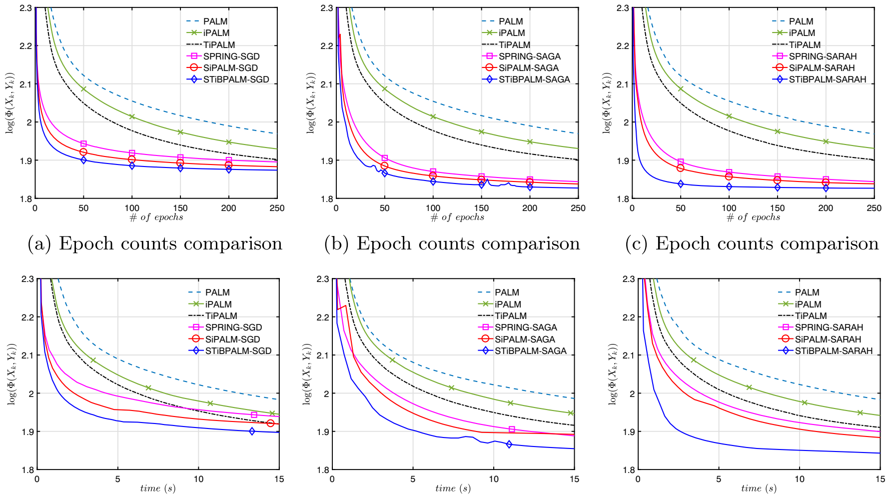
In dictionary learning and sparse coding, X is called the learned dictionary with coefficients Y. In this formulation, the sparsity on X is restricted 75% of the entries to be 0.

We use the extended Yale-B dataset and the ORL dataset, which are standard facial recognition benchmarks consisting of human face images.[1] For solving this S-NMF problem (5.1), [6, 14] gave the details on how to solve the X-subproblems and Y-subproblems. The extended Yale-B dataset contains 2414 cropped images of size 32×32, while the ORL dataset contains 400 images sized 64×64 (see Fig. 1). In the experiment for the Yale dataset, we extract 49 sparse basis images for the dataset. For the ORL dataset, we extract 25 sparse basis images. In each iteration of the stochastic algorithms, we randomly subsample 5% of the full batch as a minibatch. Here, for SARAH gradient estimator, we set $p = \frac{1}{20}$.

In STiBPALM, let $\phi_1(X) = \frac{\theta_1}{2}\|X\|^2, \phi_2(Y) = \frac{\theta_2}{2}\|Y\|^2$. In a numerical experiment, we choose $\eta = 3$ and calculate θ_1 and θ_2 by computing the largest eigenvalues of ηYY^T and $\eta X^T X$ at k-th iteration, respectively. We choose $\alpha_{1k} = \beta_{1k} = \gamma_{1k} = \mu_{1k} = \frac{k-1}{k+2}$, $\alpha_{2k} = \beta_{2k} = \gamma_{2k} = \mu_{2k} = \frac{k-1}{k+2}$ in TiPALM and STiBPALM and $\alpha_{1k} = \beta_{1k} = \gamma_{1k} = \mu_{1k} = \frac{k-1}{k+2}$ in iPALM and SiPALM. We use BTiPALM and BSTiPALM to denote TiPALM and STiBPALM with $\phi_1(X) = \frac{\theta_1^2}{4}\|X\|^4$, $\phi_2(Y) = \frac{\theta_2}{2}\|Y\|^2$, respectively. We refer to BSTiPALM using the SGD, SAGA, and SARAH gradient estimators as BSTiPALM-SGD, BSTiPALM-SAGA, and BSTiPALM-SARAH, respectively.

In Figs. 2 and 3, we report the numerical results for Yale-B dataset. A similar result for the ORL dataset is plotted in Figs. 4 and 5. One can observe from these four figures that the STiBPALM can get slightly lower values than the other algorithms within almost the same computation time. In addition, STiBPALM can get better performance than the SPRING and SiPALM stochastic algorithms with epoch changes.

[1] http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html.



(a) Epoch counts comparison (b) Epoch counts comparison (c) Epoch counts comparison

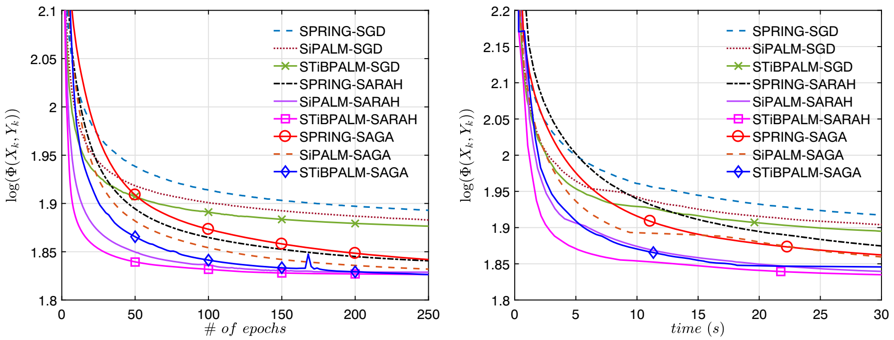
(d) Wall-clock time comparison (e) Wall-clock time comparison (f) Wall-clock time comparison

Fig. 2 Objective decrease comparison of S-NMF with $s = 25\%$ on Yale dataset. From the left column to the right column are the results of SGD, SAGA, and SARAH, respectively

The stochastic algorithms can improve the numerical results compared with the corresponding deterministic method. Furthermore, compared with the stochastic gradient algorithm without variance reduction (SGD), the variance-reduced stochastic gradient (SAGA, SARAH) algorithm can get better numerical results.

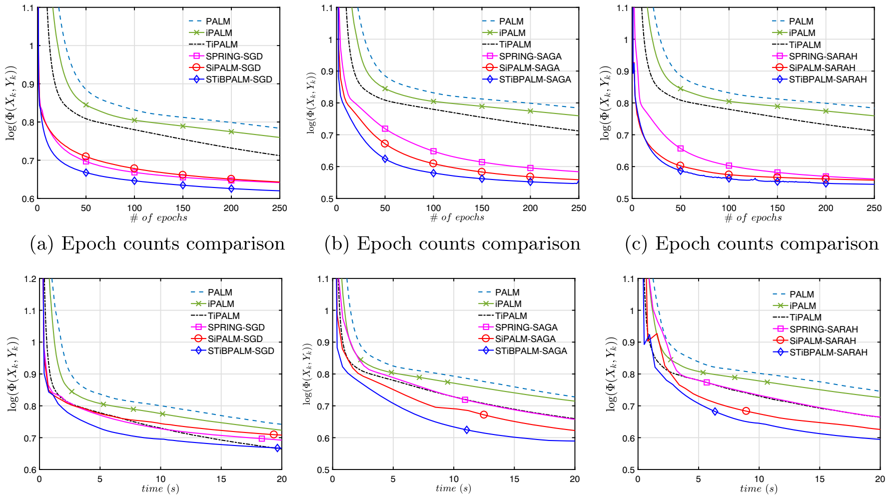
The numerical results applying different Bregman distances under the Yale-B dataset and ORL dataset are reported in Figs. 6 and 7, respectively. We can observe that BSTiPALM algorithm can obtain better numerical results compared to STiBPALM algorithm, where SARAH gradient estimator can get the best performance with epoch changes.

We also compare STiBPALM with SGD, SAGA, and SARAH for different sparsity settings (the value of s). The results of the basis images are shown in Fig. 8. One can



(a) Epoch counts comparison (b) Wall-clock time comparison

Fig. 3 Objective decrease comparison of S-NMF with $s = 25\%$ on Yale dataset



(a) Epoch counts comparison (b) Epoch counts comparison (c) Epoch counts comparison

(d) Wall-clock time comparison (e) Wall-clock time comparison (f) Wall-clock time comparison

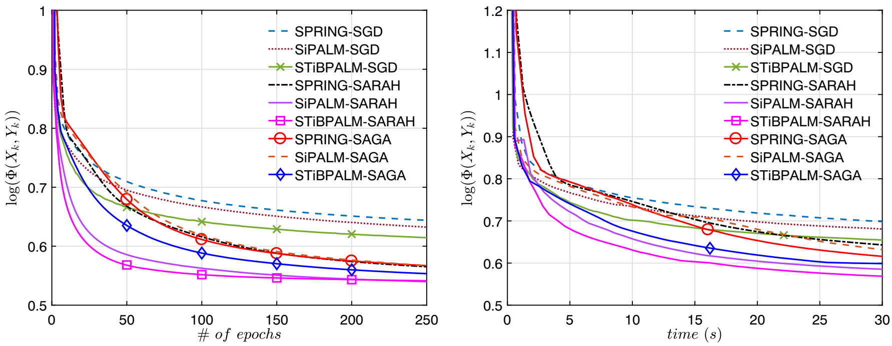
Fig. 4 Objective decrease comparison of S-NMF with $s = 25\%$ on ORL dataset. From the left column to the right column are the results of SGD, SAGA, and SARAH, respectively

observe from Fig. 8 that for smaller values of s, the four algorithms lead to more compact representations. This might improve the generalization capabilities of the representation.

5.2 Blind image-deblurring

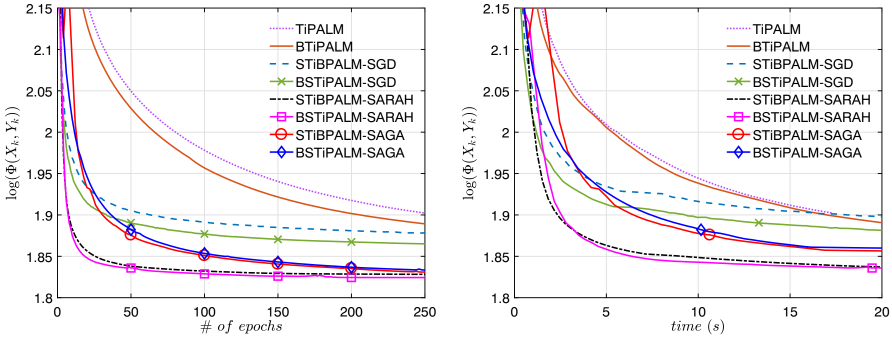
Let A be a blurred image, the problem of blind deconvolution is given by

$$\min_{X,Y}\left\{\frac{1}{2}\|A-X\odot Y\|_F^2+\eta\sum_{r=1}^{2d}R([D(X)]_r):\ 0\leq X\leq 1,\ 0\leq Y\leq 1,\ \|Y\|_1\leq 1\right\}. \tag{5.2}$$



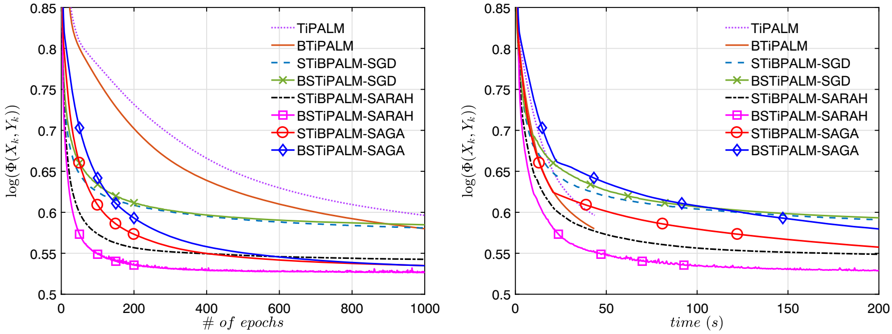
(a) Epoch counts comparison (b) Wall-clock time comparison

Fig. 5 Objective decrease comparison of S-NMF with $s = 25\%$ on ORL dataset



(a) Epoch counts comparison (b) Wall-clock time comparison

Fig. 6 Objective decrease comparison of S-NMF with $s = 25\%$ on Yale dataset with different Brengman distance



(a) Epoch counts comparison (b) Wall-clock time comparison

Fig. 7 Objective decrease comparison of S-NMF with $s = 25\%$ on ORL dataset with different Brengman distance

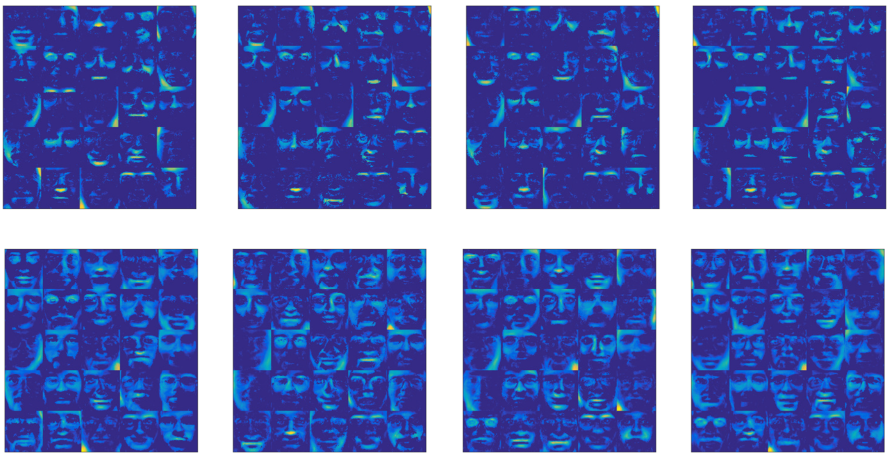
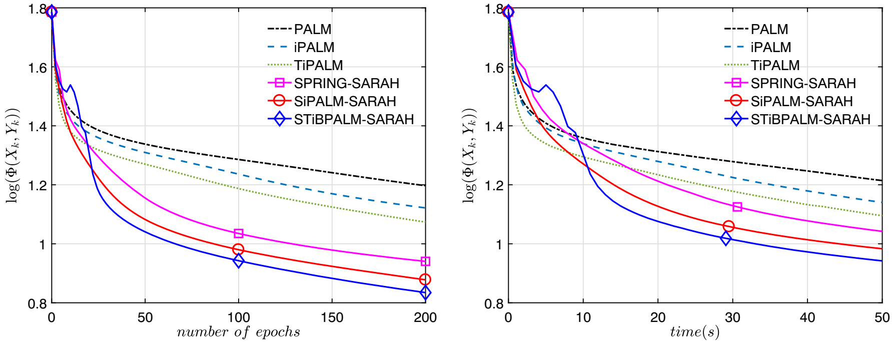


Fig. 8 The results for 25 basis faces using different sparsity settings. From the left column to the right column are the results of TiPALM, STiBPALM-SGD, STiBPALM-SAGA, and STiBPALM-SARAH, respectively. From top row to bottom row are the result of $s = 25\%$ and $s = 50\%$, respectively



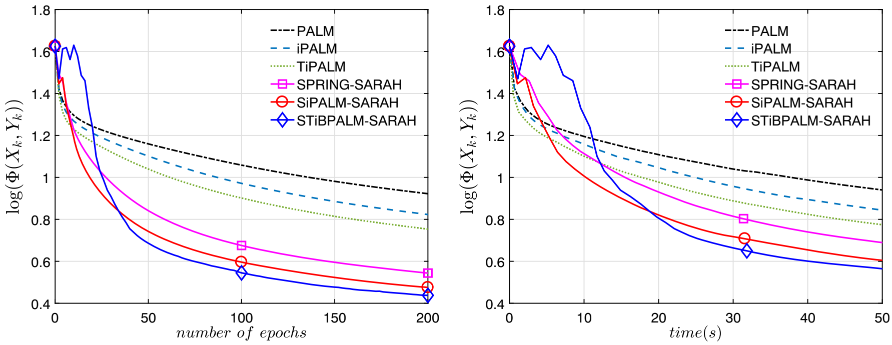
(a) Epoch counts comparison (b) Wall-clock time comparison

Fig. 9 Objective decrease comparison (epoch counts) of blind image-deconvolution experiment on Kodim08 image using an 11×11 motion blur kernel

In numerical experiment, we choose $R(v) = \log(1 + \sigma v^2)$ as in [6], where $\sigma = 10^3$ and $\eta = 5 \times 10^{-5}$.

We consider two images, Kodim08 and Kodim15, of size 256×256 for testing. For each image, two blur kernels—linear motion blur and out-of-focus blur—are considered with additional additive Gaussian noise. In this numerical experiment, we mainly use SARAH gradient estimator and set $p = \frac{1}{64}$. We take $\alpha_{1k} = \beta_{1k} = \gamma_{1k} = \mu_{1k} = \frac{k-1}{k+2}$, $\alpha_{2k} = \beta_{2k} = \gamma_{2k} = \mu_{2k} = \frac{k-1}{k+2}$ in TiPALM and STiBPALM and $\alpha_{1k} = \beta_{1k} = \gamma_{1k} = \mu_{1k} = \frac{k-1}{k+2}$ in iPALM.

The convergence comparisons of the algorithms for both images with motion blur are provided in Figs. 9 and 10, from which we observe STiBPALM-SARAH is faster than the other methods. Figures 11 and 12 provide comparisons of the recovered image and blur kernel. We observe superior performance of stochastic algorithms over deterministic algorithms in these figures as well. In particular, when comparing the estimated blur kernels of the two algorithms every 20 epochs, we clearly see that STiBPALM-SARAH more quickly recovers more accurate solutions than TiPALM.



(a) Epoch counts comparison (b) Wall-clock time comparison

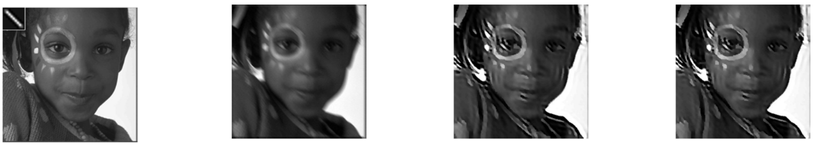
Fig. 10 Objective decrease comparison (epoch counts) of blind image-deconvolution experiment on Kodim15 image using an 11×11 motion blur kernel

(a) Original image and kernel (b) Blurred image (c) Recovered by PALM (d) Recovered by iPALM

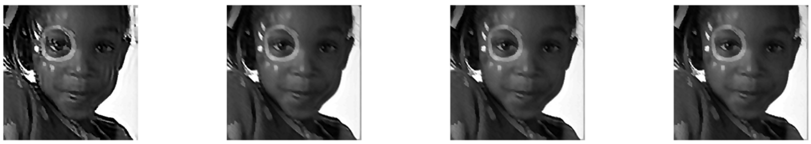
(e) Recovered by TiPALM (f) Recovered by SPRING-SARAH (g) Recovered by SiPALM-SARAH (h) Recovered by STiB PALM-SARAH

(i) Estimated kernel by TiPALM (j) Estimated kernel by STiBPALM-SARAH

Fig. 11 Image and kernel reconstructions from the blind image-deconvolution experiment on the Kodim08 image using an 11×11 motion blur kernel



(a) Original image and kernel (b) Blurred image (c) Recovered by PALM (d) Recovered by iPALM



(e) Recovered by TiPALM (f) Recovered by SPRING-SARAH (g) Recovered by SiPALM-SARAH (h) Recovered by STiB PALM-SARAH

(i) Estimated kernel by TiPALM (j) Estimated kernel by STiBPALM-SARAH

Fig. 12 Image and kernel reconstructions from the blind image-deconvolution experiment on the Kodim08 image using an 11×11 motion blur kernel

6 Conclusion

In this paper, we propose a stochastic two-step inertial Bregman proximal alternating linearized minimization (STiBPALM) algorithm with the variance-reduced gradient estimator to solve a class of nonconvex nonsmooth optimization problems. Under some mild conditions, we analyze the convergence properties of STiBPALM when using a variety of variance-reduced gradient estimators and prove specific convergence rates using the SAGA and SARAH estimators. We also implement the STiBPALM algorithm to sparse nonnegative matrix factorization and blind image-deblurring problems and perform some numerical experiments to demonstrate the effectiveness of the proposed algorithm.

Appendix

A SAGA variance bound

We define the SAGA gradient estimators $\widetilde{\nabla}_x(u_k, y_k)$ and $\widetilde{\nabla}_y(x_{k+1}, v_k)$ as follows:

$$\widetilde{\nabla}_x(u_k, y_k) = \frac{1}{b}\sum_{i\in I_k^x}\Big(\nabla_x H_i(u_k, y_k) - \nabla_x H_i(\varphi_k^i, y_k)\Big) + \frac{1}{n}\sum_{j=1}^{n}\nabla_x H_j(\varphi_k^j, y_k), \tag{A.1}$$

$$\widetilde{\nabla}_y(x_{k+1}, v_k) = \frac{1}{b}\sum_{i\in I_k^y}\Big(\nabla_y H_i(x_{k+1}, v_k) - \nabla_y H_i(x_{k+1}, \xi_k^i)\Big) + \frac{1}{n}\sum_{j=1}^{n}\nabla_y H_j(x_{k+1}, \xi_k^j),$$

where I_k^x and I_k^y are mini-batches containing b indices. The variables φ_k^i and ξ_k^i follow the update rules $\varphi_{k+1}^i = u_k$ if $i \in I_k^x$ and $\varphi_{k+1}^i = \varphi_k^i$ otherwise, and $\xi_{k+1}^i = v_k$ if $i \in I_k^y$ and $\xi_{k+1}^i = \xi_k^i$ otherwise.

To prove our variance bounds, we require the following lemma.

Lemma A.1 *Suppose $X_1, \cdots, X_t$ are independent random variables satisfying $\mathbb{E}_k X_i = 0$ for $1 \le i \le t$. Then*

$$\mathbb{E}_k \|X_1 + \cdots + X_t\|^2 = \mathbb{E}_k\left[\|X_1\|^2 + \cdots + \|X_t\|^2\right]. \tag{A.2}$$

Proof Our hypotheses on these random variables imply $\mathbb{E}_k\langle X_i, X_j\rangle = 0$ for $i \ne j$. Therefore,

$$\mathbb{E}_k \|X_1 + \cdots + X_t\|^2 = \mathbb{E}_k \sum_{i,j=1}^{t}\langle X_i, X_j\rangle = \mathbb{E}_k\left[\|X_1\|^2 + \cdots + \|X_t\|^2\right].$$

□

We are now prepared to prove that the SAGA gradient estimator is variance-reduced.

Lemma A.2 *The SAGA gradient estimator satisfies*

$$\mathbb{E}_k \left\| \widetilde{\nabla}_x(u_k, y_k) - \nabla_x H(u_k, y_k) \right\|^2 \leq \frac{1}{bn} \sum_{j=1}^{n} \left\| \nabla_x H_j(u_k, y_k) - \nabla_x H_j(\varphi_k^j, y_k) \right\|^2,$$

$$\mathbb{E}_k \left\| \widetilde{\nabla}_y(x_{k+1}, v_k) - \nabla_y H(x_{k+1}, v_k) \right\|^2 \leq \frac{4}{bn} \sum_{j=1}^{n} \left\| \nabla_y H_j(x_k, v_k) - \nabla_y H_j(x_k, \xi_k^j) \right\|^2$$

$$+ \frac{16N^2\gamma^2}{b} \left(\mathbb{E}_k \|z_{k+1} - z_k\|^2 + \|z_k - z_{k-1}\|^2 + \|z_{k-1} - z_{k-2}\|^2 \right), \tag{A.3}$$

as well as

$$\mathbb{E}_k \left\| \widetilde{\nabla}_x(u_k, y_k) - \nabla_x H(u_k, y_k) \right\| \leq \frac{1}{\sqrt{bn}} \sum_{j=1}^{n} \left\| \nabla_x H_j(u_k, y_k) - \nabla_x H_j(\varphi_k^j, y_k) \right\|,$$

$$\mathbb{E}_k \left\| \widetilde{\nabla}_y(x_{k+1}, v_k) - \nabla_y H(x_{k+1}, v_k) \right\| \leq \frac{2}{\sqrt{bn}} \sum_{j=1}^{n} \left\| \nabla_y H_j(x_k, v_k) - \nabla_y H_j(x_k, \xi_k^j) \right\|$$

$$+ \frac{4N\gamma}{\sqrt{b}} \left(\mathbb{E}_k \|z_{k+1} - z_k\| + \|z_k - z_{k-1}\| + \|z_{k-1} - z_{k-2}\| \right), \tag{A.4}$$

where $N = \max\{M, L\}$, $\gamma = \max\{\gamma_1, \gamma_2\}$.

Proof According to (A.1), we have

$$\begin{aligned} &\mathbb{E}_k \left\| \widetilde{\nabla}_x(u_k, y_k) - \nabla_x H(u_k, y_k) \right\|^2 \qquad (A.5)\\ =&\mathbb{E}_k \left\| \frac{1}{b} \sum_{i \in I_k^x} \left(\nabla_x H_i(u_k, y_k) - \nabla_x H_i(\varphi_k^i, y_k) \right) - \nabla_x H(u_k, y_k) + \frac{1}{n} \sum_{j=1}^{n} \nabla_x H_j(\varphi_k^j, y_k) \right\|^2 \\ \overset{(1)}{\leq}& \frac{1}{b^2} \mathbb{E}_k \sum_{i \in I_k^x} \left\| \nabla_x H_i(u_k, y_k) - \nabla_x H_i(\varphi_k^i, y_k) \right\|^2 \\ =& \frac{1}{bn} \sum_{j=1}^{n} \left\| \nabla_x H_j(u_k, y_k) - \nabla_x H_j(\varphi_k^j, y_k) \right\|^2. \end{aligned}$$

Inequality (1) follows from Lemma A.1. By the Jensen's inequality, we can say that

$$\begin{aligned} \mathbb{E}_k \left\| \widetilde{\nabla}_x(u_k, y_k) - \nabla_x H(u_k, y_k) \right\| &\leq \sqrt{\mathbb{E}_k \left\| \widetilde{\nabla}_x(u_k, y_k) - \nabla_x H(u_k, y_k) \right\|^2} \qquad (A.6)\\ &\leq \frac{1}{\sqrt{bn}} \sqrt{\sum_{j=1}^{n} \left\| \nabla_x H_j(u_k, y_k) - \nabla_x H_j(\varphi_k^j, y_k) \right\|^2} \end{aligned}$$

$$\leq \frac{1}{\sqrt{bn}} \sum_{j=1}^{n} \left\| \nabla_x H_j(u_k, y_k) - \nabla_x H_j(\varphi_k^j, y_k) \right\| .$$

We use an analogous argument for $\widetilde{\nabla}_y(x_{k+1}, v_k)$. Let $\mathbb{E}_{k,x}$ denote the expectation conditional on the first k iterations and I_k^x. By the same reasoning as in (A.5), applying the Lipschitz continuity of $\nabla_y H_j$, we obtain that

$$\begin{aligned}
&\mathbb{E}_{k,x} \left\| \widetilde{\nabla}_y(x_{k+1}, v_k) - \nabla_y H(x_{k+1}, v_k) \right\|^2 \\
\leq &\frac{1}{bn} \sum_{j=1}^{n} \left\| \nabla_y H_j(x_{k+1}, v_k) - \nabla_y H_j(x_{k+1}, \xi_k^j) \right\|^2 \\
\leq &\frac{4}{bn} \sum_{j=1}^{n} \left\| \nabla_y H_j(x_{k+1}, v_k) - \nabla_y H_j(x_k, y_k) \right\|^2 + \frac{4}{bn} \sum_{j=1}^{n} \left\| \nabla_y H_j(x_k, y_k) - \nabla_y H_j(x_k, v_k) \right\|^2 \\
&+ \frac{4}{bn} \sum_{j=1}^{n} \left\| \nabla_y H_j(x_k, v_k) - \nabla_y H_j(x_k, \xi_k^j) \right\|^2 + \frac{4}{bn} \sum_{j=1}^{n} \left\| \nabla_y H_j(x_k, \xi_k^j) - \nabla_y H_j(x_{k+1}, \xi_k^j) \right\|^2 \\
\leq &\frac{4M^2}{b} \|x_{k+1} - x_k\|^2 + \frac{4M^2}{b} \|v_k - y_k\|^2 + \frac{4L^2}{b} \|y_k - v_k\|^2 \\
&+ \frac{4}{bn} \sum_{j=1}^{n} \left\| \nabla_y H_j(x_k, v_k) - \nabla_y H_j(x_k, \xi_k^j) \right\|^2 + \frac{4M^2}{b} \|x_{k+1} - x_k\|^2 \\
\leq &\frac{4}{bn} \sum_{j=1}^{n} \left\| \nabla_y H_j(x_k, v_k) - \nabla_y H_j(x_k, \xi_k^j) \right\|^2 + \frac{8M^2}{b} \|x_{k+1} - x_k\|^2 \\
&+ \frac{4(M^2 + L^2)}{b} \left(2\gamma_1^2 \|y_k - y_{k-1}\|^2 + 2\gamma_2^2 \|y_{k-1} - y_{k-2}\|^2 \right) \\
\leq &\frac{4}{bn} \sum_{j=1}^{n} \left\| \nabla_y H_j(x_k, v_k) - \nabla_y H_j(x_k, \xi_k^j) \right\|^2 + \frac{16N^2\gamma^2}{b} \Big(\|z_{k+1} - z_k\|^2 + \|z_k - z_{k-1}\|^2 \\
&+ \|z_{k-1} - z_{k-2}\|^2 \Big),
\end{aligned} \tag{A.7}$$

where $N = \max\{M, L\}$, $\gamma = \max\{\gamma_1, \gamma_2\}$. Also, by the same reasoning as in (A.6),

$$\begin{aligned}
&\mathbb{E}_{k,x} \left\| \widetilde{\nabla}_y(x_{k+1}, v_k) - \nabla_y H(x_{k+1}, v_k) \right\| \\
\leq &\sqrt{\mathbb{E}_{k,x} \left\| \widetilde{\nabla}_y(x_{k+1}, v_k) - \nabla_y H(x_{k+1}, v_k) \right\|^2} \\
\leq &\frac{2}{\sqrt{bn}} \sum_{j=1}^{n} \left\| \nabla_y H_j(x_k, v_k) - \nabla_y H_j(x_k, \xi_k^j) \right\| + \frac{4N\gamma}{\sqrt{b}} \left(\|z_{k+1} - z_k\| + \|z_k - z_{k-1}\| \right. \\
&\left. + \|z_{k-1} - z_{k-2}\| \right),
\end{aligned} \tag{A.8}$$

Applying the operator $\mathbb{E}_k$ to (A.7) and (A.8), we get the desired result. □

Now, define

$$\Upsilon_{k+1} = \frac{1}{bn}\sum_{j=1}^{n}\left(\left\|\nabla_x H_j(u_{k+1}, y_{k+1}) - \nabla_x H_j(\varphi_{k+1}^{j}, y_{k+1})\right\|^2 + 4\left\|\nabla_y H_j(x_{k+1}, v_{k+1}) - \nabla_y H_j(x_{k+1}, \xi_{k+1}^{j})\right\|^2\right), \tag{A.9}$$

$$\Gamma_{k+1} = \frac{1}{\sqrt{bn}}\sum_{j=1}^{n}\left(\left\|\nabla_x H_j(u_{k+1}, y_{k+1}) - \nabla_x H_j(\varphi_{k+1}^{j}, y_{k+1})\right\|^2 + 2\left\|\nabla_y H_j(x_{k+1}, v_{k+1}) - \nabla_y H_j(x_{k+1}, \xi_{k+1}^{j})\right\|^2\right).$$

By Lemma A.2, we have

$$\begin{aligned}&\mathbb{E}_k\left[\left\|\widetilde{\nabla}_x(u_k, y_k) - \nabla_x H(u_k, y_k)\right\|^2 + \left\|\widetilde{\nabla}_y(x_{k+1}, v_k) - \nabla_y H(x_{k+1}, v_k)\right\|^2\right]\\ \le&\Upsilon_k + V_1\left(\mathbb{E}_k\|z_{k+1} - z_k\|^2 + \|z_k - z_{k-1}\|^2 + \|z_{k-1} - z_{k-2}\|^2\right),\end{aligned}$$

and

$$\begin{aligned}&\mathbb{E}_k\left[\left\|\widetilde{\nabla}_x(u_k, y_k) - \nabla_x H(u_k, y_k)\right\| + \left\|\widetilde{\nabla}_y(x_{k+1}, v_k) - \nabla_y H(x_{k+1}, v_k)\right\|\right]\\ \le&\Gamma_k + V_2\left(\mathbb{E}_k\|z_{k+1} - z_k\| + \|z_k - z_{k-1}\| + \|z_{k-1} - z_{k-2}\|\right).\end{aligned}$$

This is exactly the MSE bound, where $V_1 = \frac{16N^2\gamma^2}{b}$ and $V_2 = \frac{4N\gamma}{\sqrt{b}}$.

Lemma A.3 (Geometric decay) *Let Υ_k be defined as in (A.9), then we can establish the geometric decay property:*

$$\mathbb{E}_k\Upsilon_{k+1} \le (1-\rho)\,\Upsilon_k + V_\Upsilon\left(\mathbb{E}_k\|z_{k+1} - z_k\|^2 + \|z_k - z_{k-1}\|^2 + \|z_{k-1} - z_{k-2}\|^2\right), \tag{A.10}$$

where $\rho = \frac{b}{2n}$, $V_\Upsilon = \frac{408nN^2(1+2\gamma_1^2+\gamma_2^2)}{b^2}$.

Proof We show that $\mathbb{E}_k\Upsilon_{k+1}$ is decreasing at a geometric rate. By applying the inequality $\|a-c\|^2 \le (1+\varepsilon)\|a-b\|^2 + (1+\varepsilon^{-1})\|b-c\|^2$ twice, it follows that

$$\begin{aligned}&\frac{1}{bn}\sum_{j=1}^{n}\mathbb{E}_k\left\|\nabla_x H_j(u_{k+1}, y_{k+1}) - \nabla_x H_j(\varphi_{k+1}^{j}, y_{k+1})\right\|^2\\ \le&\frac{1+\varepsilon}{bn}\sum_{j=1}^{n}\mathbb{E}_k\left\|\nabla_x H_j(u_k, y_k) - \nabla_x H_j(\varphi_{k+1}^{j}, y_{k+1})\right\|^2 + \frac{1+\varepsilon^{-1}}{bn}\sum_{j=1}^{n}\mathbb{E}_k\left\|\nabla_x H_j(u_{k+1}, y_{k+1})\right.\\ &\left.-\nabla_x H_j(u_k, y_k)\right\|^2\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{(1+\varepsilon)^2}{bn} \sum_{j=1}^{n} \mathbb{E}_k \left\| \nabla_x H_j(u_k, y_k) - \nabla_x H_j(\varphi_{k+1}^j, y_k) \right\|^2 \\
&\quad + \frac{(1+\varepsilon)(1+\varepsilon^{-1})}{bn} \sum_{j=1}^{n} \mathbb{E}_k \left\| \nabla_x H_j(\varphi_{k+1}^j, y_k) - \nabla_x H_j(\varphi_{k+1}^j, y_{k+1}) \right\|^2 \\
&\quad + \frac{1+\varepsilon^{-1}}{bn} \sum_{j=1}^{n} \mathbb{E}_k \left\| \nabla_x H_j(u_{k+1}, y_{k+1}) - \nabla_x H_j(u_k, y_k) \right\|^2 \\
&\leq \frac{(1+\varepsilon)^2(1-b/n)}{bn} \sum_{j=1}^{n} \left\| \nabla_x H_j(u_k, y_k) - \nabla_x H_j(\varphi_k^j, y_k) \right\|^2 + \frac{(1+\varepsilon)(1+\varepsilon^{-1})M^2}{b} \mathbb{E}_k \left\| y_k - y_{k+1} \right\|^2 \\
&\quad + \frac{(1+\varepsilon^{-1})M^2}{b} \mathbb{E}_k \left(\left\| u_{k+1} - u_k \right\|^2 + \left\| y_{k+1} - y_k \right\|^2 \right) \\
&\leq \frac{(1+\varepsilon)^2(1-b/n)}{bn} \sum_{j=1}^{n} \left\| \nabla_x H_j(u_k, y_k) - \nabla_x H_j(\varphi_k^j, y_k) \right\|^2 + \frac{(2+\varepsilon)(1+\varepsilon^{-1})M^2}{b} \mathbb{E}_k \left\| y_{k+1} - y_k \right\|^2 \\
&\quad + \frac{(1+\varepsilon^{-1})M^2}{b} \mathbb{E}_k \left(3\left\| u_{k+1} - x_{k+1} \right\|^2 + 3\left\| x_{k+1} - x_k \right\|^2 + 3\left\| x_k - u_k \right\|^2 \right) \\
&\leq \frac{(1+\varepsilon)^2(1-b/n)}{bn} \sum_{j=1}^{n} \left\| \nabla_x H_j(u_k, y_k) - \nabla_x H_j(\varphi_k^j, y_k) \right\|^2 + \frac{(2+\varepsilon)(1+\varepsilon^{-1})M^2}{b} \mathbb{E}_k \left\| y_{k+1} - y_k \right\|^2 \\
&\quad + \frac{3M^2(1+\varepsilon^{-1})(1+2\gamma_1^2)}{b} \mathbb{E}_k \left\| x_{k+1} - x_k \right\|^2 + \frac{6M^2(1+\varepsilon^{-1})(\gamma_1^2+\gamma_2^2)}{b} \left\| x_k - x_{k-1} \right\|^2 \\
&\quad + \frac{6M^2(1+\varepsilon^{-1})\gamma_2^2}{b} \left\| x_{k-1} - x_{k-2} \right\|^2 . \qquad \text{(A.11)}
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
&\frac{1}{bn} \sum_{j=1}^{n} \mathbb{E}_k \left\| \nabla_y H_j(x_{k+1}, v_{k+1}) - \nabla_y H_j(x_{k+1}, \xi_{k+1}^j) \right\|^2 \\
&\leq \frac{1+\varepsilon}{bn} \sum_{j=1}^{n} \mathbb{E}_k \left\| \nabla_y H_j(x_{k+1}, v_k) - \nabla_y H_j(x_{k+1}, \xi_{k+1}^j) \right\|^2 \\
&\quad + \frac{1+\varepsilon^{-1}}{bn} \sum_{j=1}^{n} \mathbb{E}_k \left\| \nabla_y H_j(x_{k+1}, v_{k+1}) - \nabla_y H_j(x_{k+1}, v_k) \right\|^2 \\
&\leq \frac{(1+\varepsilon)^2(1-b/n)}{bn} \sum_{j=1}^{n} \mathbb{E}_k \left\| \nabla_y H_j(x_k, v_k) - \nabla_y H_j(x_{k+1}, \xi_k^j) \right\|^2 \\
&\quad + \frac{(1+\varepsilon)(1+\varepsilon^{-1})(1-b/n)}{bn} \sum_{j=1}^{n} \mathbb{E}_k \left\| \nabla_y H_j(x_{k+1}, v_k) - \nabla_y H_j(x_k, v_k) \right\|^2 \\
&\quad + \frac{1+\varepsilon^{-1}}{bn} \sum_{j=1}^{n} \mathbb{E}_k \left\| \nabla_y H_j(x_{k+1}, v_{k+1}) - \nabla_y H_j(x_{k+1}, v_k) \right\|^2 \\
&\leq \frac{(1+\varepsilon)^3(1-b/n)}{bn} \sum_{j=1}^{n} \left\| \nabla_y H_j(x_k, v_k) - \nabla_y H_j(x_k, \xi_k^j) \right\|^2
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{(1+\varepsilon)^2(1+\varepsilon^{-1})(1-b/n)}{bn}\sum_{j=1}^{n}\mathbb{E}_k\left\|\nabla_y H_j(x_k,\xi_k^j)-\nabla_y H_j(x_{k+1},\xi_k^j)\right\|^2\\
&+\frac{(1+\varepsilon)(1+\varepsilon^{-1})(1-b/n)}{bn}\sum_{j=1}^{n}\mathbb{E}_k\left\|\nabla_y H_j(x_{k+1},v_k)-\nabla_y H_j(x_k,v_k)\right\|^2\\
&+\frac{1+\varepsilon^{-1}}{bn}\sum_{j=1}^{n}\mathbb{E}_k\left\|\nabla_y H_j(x_{k+1},v_{k+1})-\nabla_y H_j(x_{k+1},v_k)\right\|^2\\
\leq&\frac{(1+\varepsilon)^3(1-b/n)}{bn}\sum_{j=1}^{n}\left\|\nabla_y H_j(x_k,v_k)-\nabla_y H_j(x_k,\xi_k^j)\right\|^2+\frac{(1+\varepsilon)^2(1+\varepsilon^{-1})(1-b/n)M^2}{b}\\
&\mathbb{E}_k\left\|x_{k+1}-x_k\right\|^2+\frac{(1+\varepsilon)(1+\varepsilon^{-1})(1-b/n)M^2}{b}\mathbb{E}_k\left\|x_{k+1}-x_k\right\|^2+\frac{(1+\varepsilon^{-1})L^2}{b}\mathbb{E}_k\left\|v_{k+1}-v_k\right\|^2\\
\leq&\frac{(1+\varepsilon)^3(1-b/n)}{bn}\sum_{j=1}^{n}\left\|\nabla_y H_j(x_k,v_k)-\nabla_y H_j(x_k,\xi_k^j)\right\|^2+\frac{(2+\varepsilon)(1+\varepsilon)(1+\varepsilon^{-1})(1-b/n)M^2}{b}\\
&\mathbb{E}_k\left\|x_{k+1}-x_k\right\|^2+\frac{(1+\varepsilon^{-1})L^2}{b}\mathbb{E}_k\left(3\left\|v_{k+1}-y_{k+1}\right\|^2+3\left\|y_{k+1}-y_k\right\|^2+3\left\|y_k-v_k\right\|^2\right)\\
\leq&\frac{(1+\varepsilon)^3(1-b/n)}{bn}\sum_{j=1}^{n}\left\|\nabla_y H_j(x_k,v_k)-\nabla_y H_j(x_k,\xi_k^j)\right\|^2+\frac{(2+\varepsilon)(1+\varepsilon)(1+\varepsilon^{-1})(1-b/n)M^2}{b}\\
&\mathbb{E}_k\left\|x_{k+1}-x_k\right\|^2+\frac{3L^2(1+\varepsilon^{-1})(1+2\gamma_1^2)}{b}\mathbb{E}_k\left\|y_{k+1}-y_k\right\|^2+\frac{6L^2(1+\varepsilon^{-1})(\gamma_1^2+\gamma_2^2)}{b}\\
&\left\|y_k-y_{k-1}\right\|^2+\frac{6L^2(1+\varepsilon^{-1})\gamma_2^2}{b}\left\|y_{k-1}-y_{k-2}\right\|^2. \qquad \text{(A.12)}
\end{aligned}
$$

With

$$
\begin{aligned}
\Upsilon_{k+1}=&\frac{1}{bn}\sum_{j=1}^{n}\Bigg(\left\|\nabla_x H_j(u_{k+1},y_{k+1})-\nabla_x H_j(\varphi_{k+1}^j,y_{k+1})\right\|^2\\
&+4\left\|\nabla_y H_j(x_{k+1},v_{k+1})-\nabla_y H_j(x_{k+1},\xi_{k+1}^j)\right\|^2\Bigg),
\end{aligned}
$$

adding (A.11) and (A.12), we can obtain

$$
\begin{aligned}
&\mathbb{E}_k\Upsilon_{k+1}\\
\leq&(1+\varepsilon)^3(1-b/n)\Upsilon_k+\frac{(2+\varepsilon)(1+\varepsilon^{-1})M^2}{b}\mathbb{E}_k\left\|y_{k+1}-y_k\right\|^2+\frac{3M^2(1+\varepsilon^{-1})(1+2\gamma_1^2)}{b}\\
&\mathbb{E}_k\left\|x_{k+1}-x_k\right\|^2+\frac{6M^2(1+\varepsilon^{-1})(\gamma_1^2+\gamma_2^2)}{b}\left\|x_k-x_{k-1}\right\|^2+\frac{6M^2(1+\varepsilon^{-1})\gamma_2^2}{b}\\
&\left\|x_{k-1}-x_{k-2}\right\|^2+\frac{4(1+\varepsilon)(1+\varepsilon^{-1})(1-b/n)M^2(2+\varepsilon)}{b}\mathbb{E}_k\left\|x_{k+1}-x_k\right\|^2\\
&+\frac{12L^2(1+\varepsilon^{-1})(1+2\gamma_1^2)}{b}\mathbb{E}_k\left\|y_{k+1}-y_k\right\|^2+\frac{24L^2(1+\varepsilon^{-1})(\gamma_1^2+\gamma_2^2)}{b}\left\|y_k-y_{k-1}\right\|^2
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{24L^2(1+\varepsilon^{-1})\gamma_2^2}{b}\|y_{k-1}-y_{k-2}\|^2\\
\leq&(1+\varepsilon)^3(1-b/n)\Upsilon_k+\frac{13N^2(1+\varepsilon)(2+\varepsilon)(1+\varepsilon^{-1})(1+2\gamma_1^2)}{b}\mathbb{E}_k\|z_{k+1}-z_k\|^2\\
&+\frac{24N^2(1+\varepsilon^{-1})(\gamma_1^2+\gamma_2^2)}{b}\|z_k-z_{k-1}\|^2+\frac{24N^2\gamma_2^2(1+\varepsilon^{-1})}{b}\|z_{k-1}-z_{k-2}\|^2\\
\leq&(1+\varepsilon)^3(1-b/n)\Upsilon_k+\frac{24N^2(1+\varepsilon)(2+\varepsilon)(1+\varepsilon^{-1})(1+2\gamma_1^2+\gamma_2^2)}{b}\Big(\mathbb{E}_k\|z_{k+1}-z_k\|^2\\
&+\|z_k-z_{k-1}\|^2+\|z_{k-1}-z_{k-2}\|^2\Big),
\end{aligned}
$$

where $N=\max\{M,L\}$. Choosing $\varepsilon=\frac{b}{6n}$, we have $(1+\varepsilon)^3(1-\frac{b}{n})\leq 1-\frac{b}{2n}$, producing the inequality

$$
\begin{aligned}
\mathbb{E}_k\Upsilon_{k+1}\leq&(1-\frac{b}{2n})\Upsilon_k+\frac{24N^2(1+\frac{b}{6n})(2+\frac{b}{6n})(1+\frac{6n}{b})(1+2\gamma_1^2+\gamma_2^2)}{b}\big(\mathbb{E}_k\|z_{k+1}-z_k\|^2\\
&+\|z_k-z_{k-1}\|^2+\|z_{k-1}-z_{k-2}\|^2\big)\\
\leq&(1-\frac{b}{2n})\Upsilon_k+\frac{408nN^2(1+2\gamma_1^2+\gamma_2^2)}{b^2}\big(\mathbb{E}_k\|z_{k+1}-z_k\|^2+\|z_k-z_{k-1}\|^2+\|z_{k-1}-z_{k-2}\|^2\big).
\end{aligned}
\tag{A.13}
$$

This completes the proof. □

Lemma A.4 (Convergence of estimator) *If* $\{z_k\}_{k\in\mathbb{N}}$ *satisfies* $\lim_{k\to\infty}\mathbb{E}\|z_k-z_{k-1}\|^2=0$*, then* $\mathbb{E}\Upsilon_k\to 0$ *and* $\mathbb{E}\Gamma_k\to 0$ *as* $k\to\infty$.

Proof We frist show that $\sum_{j=1}^n\mathbb{E}\left\|\nabla_x H_j(u_k,y_k)-\nabla_x H_j(\varphi_k^j,y_k)\right\|^2\to 0$ as $k\to\infty$. Indeed,

$$
\begin{aligned}
&\sum_{j=1}^n\mathbb{E}\left\|\nabla_x H_j(u_k,y_k)-\nabla_x H_j(\varphi_k^j,y_k)\right\|^2\leq L^2\sum_{j=1}^n\mathbb{E}\left\|u_k-\varphi_k^j\right\|^2\\
\leq&nL^2(1+\frac{2n}{b})\mathbb{E}\|u_k-u_{k-1}\|^2+L^2(1+\frac{b}{2n})\sum_{j=1}^n\mathbb{E}\left\|u_{k-1}-\varphi_k^j\right\|^2\\
\leq&nL^2(1+\frac{2n}{b})\mathbb{E}\|u_k-u_{k-1}\|^2+L^2(1+\frac{b}{2n})(1-\frac{b}{n})\sum_{j=1}^n\mathbb{E}\left\|u_{k-1}-\varphi_{k-1}^j\right\|^2\\
\leq&nL^2(1+\frac{2n}{b})\mathbb{E}\|u_k-u_{k-1}\|^2+L^2(1-\frac{b}{2n})\sum_{j=1}^n\mathbb{E}\left\|u_{k-1}-\varphi_{k-1}^j\right\|^2\\
\leq&nL^2(1+\frac{2n}{b})\sum_{l=1}^k(1-\frac{b}{2n})^{k-l}\mathbb{E}\|u_l-u_{l-1}\|^2.
\end{aligned}
\tag{A.14}
$$

As $\mathbb{E}\|z_k - z_{k-1}\|^2 \to 0$, so $\mathbb{E}\|u_k - u_{k-1}\|^2 \to 0$, it is clear that $\sum_{l=1}^{k}(1 - \frac{b}{2n})^{k-l}\mathbb{E}\|u_l - u_{l-1}\|^2 \to 0$, and hence $\sum_{j=1}^{n}\mathbb{E}\left\|\nabla_x H_j(u_k, y_k) - \nabla_x H_j(\varphi_k^j, y_k)\right\|^2 \to 0$ as $k \to \infty$. An analogous argument shows that $\sum_{j=1}^{n}\mathbb{E}\left\|\nabla_y H_j(x_k, v_k) - \nabla_y H_j(x_k, \xi_k^j)\right\|^2 \to 0$ as $k \to \infty$. So $\mathbb{E}\Upsilon_k \to 0$ as $k \to \infty$. Similarly, we can get $\mathbb{E}\Gamma_k \to 0$ as $k \to \infty$. Indeed,

$$\begin{aligned}
&\sum_{j=1}^{n}\mathbb{E}\left\|\nabla_x H_j(u_k, y_k) - \nabla_x H_j(\varphi_k^j, y_k)\right\| \leq L\sum_{j=1}^{n}\mathbb{E}\left\|u_k - \varphi_k^j\right\| \\
\leq & nL\mathbb{E}\|u_k - u_{k-1}\| + L\sum_{j=1}^{n}\mathbb{E}\left\|u_{k-1} - \varphi_k^j\right\| \\
\leq & nL\mathbb{E}\|u_k - u_{k-1}\| + L(1 - \frac{b}{n})\sum_{j=1}^{n}\mathbb{E}\left\|u_{k-1} - \varphi_{k-1}^j\right\| \\
\leq & nL\sum_{l=1}^{k}(1 - \frac{b}{n})^{k-l}\mathbb{E}\|u_l - u_{l-1}\|.
\end{aligned} \tag{A.15}$$

Because $\mathbb{E}\|z_k - z_{k-1}\|^2 \to 0$, it follows that $\mathbb{E}\|z_k - z_{k-1}\| \to 0$ (because Jensen's inequality implies $\mathbb{E}\|z_k - z_{k-1}\| \leq \sqrt{\mathbb{E}\|z_k - z_{k-1}\|^2} \to 0$). So $\mathbb{E}\|u_k - u_{k-1}\| \to 0$, then it follows that the bound on the right goes to zero as $k \to \infty$, hence $\mathbb{E}\Gamma_k \to 0$. □

B SARAH variance bound

As in the previous section, we use I_k^x and I_k^y to denote the mini-batches used to approximate $\nabla_x H(u_k, y_k)$ and $\nabla_y H(x_{k+1}, v_k)$, respectively.

Lemma B.1 *The SARAH gradient estimator satisfies*

$$\begin{aligned}
&\mathbb{E}_k\left(\left\|\widetilde{\nabla}_x(u_k, y_k) - \nabla_x H(u_k, y_k)\right\|^2 + \left\|\widetilde{\nabla}_y(x_{k+1}, v_k) - \nabla_y H(x_{k+1}, v_k)\right\|^2\right) \\
\leq & \left(1 - \frac{1}{p}\right)\left(\left\|\widetilde{\nabla}_x(u_{k-1}, y_{k-1}) - \nabla_x H(u_{k-1}, y_{k-1})\right\|^2 + \left\|\widetilde{\nabla}_y(x_k, v_{k-1}) - \nabla_y H(x_k, v_{k-1})\right\|^2\right) \\
& + V_1\left(\mathbb{E}_k\left\|z_{k+1} - z_k\right\|^2 + \left\|z_k - z_{k-1}\right\|^2 + \left\|z_{k-1} - z_{k-2}\right\|^2 + \left\|z_{k-2} - z_{k-3}\right\|^2\right),
\end{aligned}$$

as well as

$$\begin{aligned}
&\mathbb{E}_k\left(\left\|\widetilde{\nabla}_x(u_k, y_k) - \nabla_x H(u_k, y_k)\right\| + \left\|\widetilde{\nabla}_y(x_{k+1}, v_k) - \nabla_y H(x_{k+1}, v_k)\right\|\right) \\
\leq & \sqrt{1 - \frac{1}{p}}\left(\left\|\widetilde{\nabla}_x(u_k, y_k) - \nabla_x H(u_k, y_k)\right\| + \left\|\widetilde{\nabla}_y(x_{k+1}, v_k) - \nabla_y H(x_{k+1}, v_k)\right\|\right) \\
& + V_2\left(\mathbb{E}_k\|z_{k+1} - z_k\| + \|z_k - z_{k-1}\| + \|z_{k-1} - z_{k-2}\| + \|z_{k-2} - z_{k-3}\|\right),
\end{aligned}$$

where $V_1 = 6\left(1-\frac{1}{p}\right)M^2(1+2\gamma_1^2+\gamma_2^2)$ *and* $V_2 = M\sqrt{6(1-\frac{1}{p})(1+2\gamma_1^2+\gamma_2^2)}$.

Proof Let $\mathbb{E}_{k,p}$ denote the expectation conditional on the first k iterations and the event that we do not compute the full gradient at iteration k. The conditional expectation of the SARAH gradient estimator in this case is

$$\begin{aligned}\mathbb{E}_{k,p}\widetilde{\nabla}_x(u_k,y_k) =&\frac{1}{b}\mathbb{E}_{k,p}\left(\sum_{i\in I_k^x}\nabla_x H_i(u_k,y_k)-\nabla_x H_i(u_{k-1},y_{k-1})\right)+\widetilde{\nabla}_x(u_{k-1},y_{k-1})\\ =&\nabla_x H(u_k,y_k)-\nabla_x H(u_{k-1},y_{k-1})+\widetilde{\nabla}_x(u_{k-1},y_{k-1}),\end{aligned} \tag{B.1}$$

and further

$$\begin{aligned}&\mathbb{E}_{k,p}\left\|\widetilde{\nabla}_x(u_k,y_k)-\nabla_x H(u_k,y_k)\right\|^2\\ =&\mathbb{E}_{k,p}\left\|\widetilde{\nabla}_x(u_{k-1},y_{k-1})-\nabla_x H(u_{k-1},y_{k-1})+\nabla_x H(u_{k-1},y_{k-1})-\nabla_x H(u_k,y_k)\right.\\ &\left.+\widetilde{\nabla}_x(u_k,y_k)-\widetilde{\nabla}_x(u_{k-1},y_{k-1})\right\|^2\\ =&\left\|\widetilde{\nabla}_x(u_{k-1},y_{k-1})-\nabla_x H(u_{k-1},y_{k-1})\right\|^2+\|\nabla_x H(u_{k-1},y_{k-1})-\nabla_x H(u_k,y_k)\|^2\\ &+\mathbb{E}_{k,p}\left\|\widetilde{\nabla}_x(u_k,y_k)-\widetilde{\nabla}_x(u_{k-1},y_{k-1})\right\|^2\\ &+2\left\langle\widetilde{\nabla}_x(u_{k-1},y_{k-1})-\nabla_x H(u_{k-1},y_{k-1}),\nabla_x H(u_{k-1},y_{k-1})-\nabla_x H(u_k,y_k)\right\rangle\\ &-2\left\langle\nabla_x H(u_{k-1},y_{k-1})-\widetilde{\nabla}_x(u_{k-1},y_{k-1}),\mathbb{E}_{k,p}\left(\widetilde{\nabla}_x(u_k,y_k)-\widetilde{\nabla}_x(u_{k-1},y_{k-1})\right)\right\rangle\\ &-2\left\langle\nabla_x H(u_k,y_k)-\nabla_x H(u_{k-1},y_{k-1}),\mathbb{E}_{k,p}\left(\widetilde{\nabla}_x(u_k,y_k)-\widetilde{\nabla}_x(u_{k-1},y_{k-1})\right)\right\rangle.\end{aligned} \tag{B.2}$$

By (B.1), we see that

$$\mathbb{E}_{k,p}\left(\widetilde{\nabla}_x(u_k,y_k)-\widetilde{\nabla}_x(u_{k-1},y_{k-1})\right)=\nabla_x H(u_k,y_k)-\nabla_x H(u_{k-1},y_{k-1}).$$

Thus, the first two inner products in (B.2) sum to zero and the third one is equal to

$$\begin{aligned}&-2\left\langle\nabla_x H(u_k,y_k)-\nabla_x H(u_{k-1},y_{k-1}),\mathbb{E}_{k,p}\left(\widetilde{\nabla}_x(u_k,y_k)-\widetilde{\nabla}_x(u_{k-1},y_{k-1})\right)\right\rangle\\ =&-2\left\langle\nabla_x H(u_k,y_k)-\nabla_x H(u_{k-1},y_{k-1}),\nabla_x H(u_k,y_k)-\nabla_x H(u_{k-1},y_{k-1})\right\rangle\\ =&-2\left\|\nabla_x H(u_k,y_k)-\nabla_x H(u_{k-1},y_{k-1})\right\|^2.\end{aligned}$$

This yields

$$\begin{aligned}&\mathbb{E}_{k,p}\left\|\widetilde{\nabla}_x(u_k,y_k)-\nabla_x H(u_k,y_k)\right\|^2\\ =&\left\|\widetilde{\nabla}_x(u_{k-1},y_{k-1})-\nabla_x H(u_{k-1},y_{k-1})\right\|^2-\|\nabla_x H(u_{k-1},y_{k-1})-\nabla_x H(u_k,y_k)\|^2\\ &+\mathbb{E}_{k,p}\left\|\widetilde{\nabla}_x(u_k,y_k)-\widetilde{\nabla}_x(u_{k-1},y_{k-1})\right\|^2\\ \le&\left\|\widetilde{\nabla}_x(u_{k-1},y_{k-1})-\nabla_x H(u_{k-1},y_{k-1})\right\|^2+\mathbb{E}_{k,p}\left\|\widetilde{\nabla}_x(u_k,y_k)-\widetilde{\nabla}_x(u_{k-1},y_{k-1})\right\|^2.\end{aligned}$$

We can bound the second term by computing the expectation.

$$
\begin{aligned}
&\mathbb{E}_{k,p} \left\| \widetilde{\nabla}_x(u_k, y_k) - \widetilde{\nabla}_x(u_{k-1}, y_{k-1}) \right\|^2 \\
=&\mathbb{E}_{k,p} \left\| \frac{1}{b} \left(\sum_{i \in I_k^x} \nabla_x H_i(u_k, y_k) - \nabla_x H_i(u_{k-1}, y_{k-1}) \right) \right\|^2 \\
\leq&\frac{1}{b} \mathbb{E}_{k,p} \left[\sum_{i \in I_k^x} \| \nabla_x H_i(u_k, y_k) - \nabla_x H_i(u_{k-1}, y_{k-1}) \|^2 \right] \\
=&\frac{1}{n} \sum_{j=1}^{n} \left\| \nabla_x H_j(u_k, y_k) - \nabla_x H_j(u_{k-1}, y_{k-1}) \right\|^2 .
\end{aligned}
$$

The inequality is due to the convexity of the function $x \mapsto \|x\|^2$. This results in the recursive inequality

$$
\begin{aligned}
&\mathbb{E}_{k,p} \left\| \widetilde{\nabla}_x(u_k, y_k) - \nabla_x H(u_k, y_k) \right\|^2 \\
\leq& \left\| \widetilde{\nabla}_x(u_{k-1}, y_{k-1}) - \nabla_x H(u_{k-1}, y_{k-1}) \right\|^2 + \frac{1}{n} \sum_{j=1}^{n} \left\| \nabla_x H_j(u_k, y_k) - \nabla_x H_j(u_{k-1}, y_{k-1}) \right\|^2 .
\end{aligned}
$$

This bounds the MSE under the condition that the full gradient is not computed. When the full gradient is computed, the MSE is equal to zero, so taking the M-Lipschitz continuity of the gradients of the H_j into account, we get

$$
\begin{aligned}
&\mathbb{E}_k \left\| \widetilde{\nabla}_x(u_k, y_k) - \nabla_x H(u_k, y_k) \right\|^2 \\
\leq& \left(1 - \frac{1}{p}\right) \left(\left\| \widetilde{\nabla}_x(u_{k-1}, y_{k-1}) - \nabla_x H(u_{k-1}, y_{k-1}) \right\|^2 + \frac{1}{n} \sum_{j=1}^{n} \left\| \nabla_x H_j(u_k, y_k) - \nabla_x H_j(u_{k-1}, y_{k-1}) \right\|^2 \right) \\
\leq& \left(1 - \frac{1}{p}\right) \left(\left\| \widetilde{\nabla}_x(u_{k-1}, y_{k-1}) - \nabla_x H(u_{k-1}, y_{k-1}) \right\|^2 + M^2 \left\| (u_k, y_k) - (u_{k-1}, y_{k-1}) \right\|^2 \right).
\end{aligned}
$$

Using $(a + b + c)^2 \leq 3(a^2 + b^2 + c^2)$, we can estimate

$$
\begin{aligned}
&\|(u_k, y_k) - (u_{k-1}, y_{k-1})\|^2 = \|u_k - u_{k-1}\|^2 + \|y_k - y_{k-1}\|^2 \\
\leq& 3 \|u_k - x_k\|^2 + 3 \|x_k - x_{k-1}\|^2 + 3 \|x_{k-1} - u_{k-1}\|^2 + \|y_k - y_{k-1}\|^2 \\
\leq& 3(1 + 2\gamma_1^2) \|x_k - x_{k-1}\|^2 + 6(\gamma_1^2 + \gamma_2^2) \|x_{k-1} - x_{k-2}\|^2 + 6\gamma_2^2 \|x_{k-2} - x_{k-3}\|^2 + \|y_k - y_{k-1}\|^2 .
\end{aligned}
$$

Substituting the above inequality, we can obtain

$$
\begin{aligned}
&\mathbb{E}_k \left\| \widetilde{\nabla}_x(u_k, y_k) - \nabla_x H(u_k, y_k) \right\|^2 \\
\leq& \left(1 - \frac{1}{p}\right) \left(\left\| \widetilde{\nabla}_x(u_{k-1}, y_{k-1}) - \nabla_x H(u_{k-1}, y_{k-1}) \right\|^2 + 3M^2(1 + 2\gamma_1^2) \|x_k - x_{k-1}\|^2 \right.
\end{aligned}
$$

$$+6M^2(\gamma_1^2+\gamma_2^2)\,\|x_{k-1}-x_{k-2}\|^2+6M^2\gamma_2^2\,\|x_{k-2}-x_{k-3}\|^2+M^2\,\|y_k-y_{k-1}\|^2\Big). \tag{B.3}$$

By symmetric arguments, it holds

$$\begin{aligned}
&\mathbb{E}_k \left\|\widetilde{\nabla}_y(x_{k+1}, v_k) - \nabla_y H(x_{k+1}, v_k)\right\|^2 \\
&\leq \left(1-\frac{1}{p}\right)\left(\left\|\widetilde{\nabla}_y(x_k, v_{k-1})-\nabla_y H(x_k, v_{k-1})\right\|^2+M^2\mathbb{E}_k\left\|(x_{k+1}, v_k)-(x_k, v_{k-1})\right\|^2\right) \\
&\leq \left(1-\frac{1}{p}\right)\left(\left\|\widetilde{\nabla}_y(x_k, v_{k-1})-\nabla_y H(x_k, v_{k-1})\right\|^2+M^2\mathbb{E}_k\left\|x_{k+1}-x_k\right\|^2+3M^2(1+2\mu_{1k}^2)\right. \\
&\quad \left.\left\|y_k-y_{k-1}\right\|^2+6M^2(\mu_{1,k-1}^2+\mu_{2k}^2)\left\|y_{k-1}-y_{k-2}\right\|^2+6M^2\mu_{2,k-1}^2\left\|y_{k-2}-y_{k-3}\right\|^2\right) \\
&\leq \left(1-\frac{1}{p}\right)\left(\left\|\widetilde{\nabla}_y(x_k, v_{k-1})-\nabla_y H(x_k, v_{k-1})\right\|^2+M^2\mathbb{E}_k\left\|x_{k+1}-x_k\right\|^2+3M^2(1+2\gamma_1^2)\right. \\
&\quad \left.\left\|y_k - y_{k-1}\right\|^2 + 6M^2(\gamma_1^2 + \gamma_2^2)\left\|y_{k-1} - y_{k-2}\right\|^2 + 6M^2\gamma_2^2\left\|y_{k-2} - y_{k-3}\right\|^2\right).
\end{aligned} \tag{B.4}$$

Combining (B.3) and (B.4), we can obtain

$$\begin{aligned}
&\mathbb{E}_k\left(\left\|\widetilde{\nabla}_x(u_k, y_k) - \nabla_x H(u_k, y_k)\right\|^2 + \left\|\widetilde{\nabla}_y(x_{k+1}, v_k) - \nabla_y H(x_{k+1}, v_k)\right\|^2\right) \\
&\leq \left(1-\frac{1}{p}\right)\left(\left\|\widetilde{\nabla}_x(u_{k-1}, y_{k-1})-\nabla_x H(u_{k-1}, y_{k-1})\right\|^2+\left\|\widetilde{\nabla}_y(x_k, v_{k-1})-\nabla_y H(x_k, v_{k-1})\right\|^2\right. \\
&\quad +M^2\mathbb{E}_k\left\|x_{k+1} - x_k\right\|^2 + M^2\left\|y_k - y_{k-1}\right\|^2 + 3M^2(1 + 2\gamma_1^2)\left\|z_k - z_{k-1}\right\|^2 \\
&\quad \left.+6M^2(\gamma_1^2 + \gamma_2^2)\left\|z_{k-1} - z_{k-2}\right\|^2 + 6M^2\gamma_2^2\left\|z_{k-2} - z_{k-3}\right\|^2\right) \\
&\leq \left(1 - \frac{1}{p}\right)\Upsilon_k + 6\left(1 - \frac{1}{p}\right)M^2(1 + 2\gamma_1^2 + \gamma_2^2)\left(\mathbb{E}_k\left\|z_{k+1} - z_k\right\|^2 + \left\|z_k - z_{k-1}\right\|^2\right. \\
&\quad \left.+\left\|z_{k-1} - z_{k-2}\right\|^2 + \left\|z_{k-2} - z_{k-3}\right\|^2\right).
\end{aligned}$$

Similar bounds hold for Γ_k due to Jensen's inequality:

$$\begin{aligned}
&\mathbb{E}_k\left(\left\|\widetilde{\nabla}_x(u_k, y_k) - \nabla_x H(u_k, y_k)\right\| + \left\|\widetilde{\nabla}_y(x_{k+1}, v_k) - \nabla_y H(x_{k+1}, v_k)\right\|\right) \\
&\leq\sqrt{1 - \frac{1}{p}}\left(\left\|\widetilde{\nabla}_x(u_k, y_k) - \nabla_x H(u_k, y_k)\right\| + \left\|\widetilde{\nabla}_y(x_{k+1}, v_k) - \nabla_y H(x_{k+1}, v_k)\right\|\right) \\
&\quad +M\sqrt{6(1-\frac{1}{p})(1+2\gamma_1^2+\gamma_2^2)}\left(\mathbb{E}_k\left\|z_{k+1}-z_k\right\|+\left\|z_k-z_{k-1}\right\|+\left\|z_{k-1}-z_{k-2}\right\|+\left\|z_{k-2}-z_{k-3}\right\|\right).
\end{aligned}$$

This completes the proof. □

Now, define

$$\begin{aligned}
\Upsilon_{k+1} &= \left\|\widetilde{\nabla}_x(u_k, y_k) - \nabla_x H(u_k, y_k)\right\|^2 + \left\|\widetilde{\nabla}_y(x_{k+1}, v_k) - \nabla_y H(x_{k+1}, v_k)\right\|^2, \\
\Gamma_{k+1} &= \left\|\widetilde{\nabla}_x(u_k, y_k) - \nabla_x H(u_k, y_k)\right\| + \left\|\widetilde{\nabla}_y(x_{k+1}, v_k) - \nabla_y H(x_{k+1}, v_k)\right\|.
\end{aligned} \tag{B.5}$$

By Lemma B.1, we have

$$\begin{aligned}&\mathbb{E}_k\left[\left\|\widetilde{\nabla}_x(u_k, y_k) - \nabla_x H(u_k, y_k)\right\|^2 + \left\|\widetilde{\nabla}_y(x_{k+1}, v_k) - \nabla_y H(x_{k+1}, v_k)\right\|^2\right]\\ \le& \Upsilon_k + V_1\left(\mathbb{E}_k\left\|z_{k+1} - z_k\right\|^2 + \left\|z_k - z_{k-1}\right\|^2 + \left\|z_{k-1} - z_{k-2}\right\|^2 + \left\|z_{k-2} - z_{k-3}\right\|^2\right),\end{aligned}$$

and

$$\begin{aligned}&\mathbb{E}_k\left[\left\|\widetilde{\nabla}_x(u_k, y_k) - \nabla_x H(u_k, y_k)\right\| + \left\|\widetilde{\nabla}_y(x_{k+1}, v_k) - \nabla_y H(x_{k+1}, v_k)\right\|\right]\\ \le& \Gamma_k + V_2\left(\mathbb{E}_k\left\|z_{k+1} - z_k\right\| + \left\|z_k - z_{k-1}\right\| + \left\|z_{k-1} - z_{k-2}\right\| + \left\|z_{k-2} - z_{k-3}\right\|\right).\end{aligned}$$

This is exactly the MSE bound, where $V_1 = 6\left(1 - \frac{1}{p}\right) M^2(1 + 2\gamma_1^2 + \gamma_2^2)$ and $V_2 = M\sqrt{6(1 - \frac{1}{p})(1 + 2\gamma_1^2 + \gamma_2^2)}$.

Lemma B.2 (Geometric decay) *Let Υ_k be defined as in (B.5), then we can establish the geometric decay property:*

$$\mathbb{E}_k \Upsilon_{k+1} \le (1-\rho)\,\Upsilon_k + V_\Upsilon\left(\mathbb{E}_k\left\|z_{k+1} - z_k\right\|^2 + \left\|z_k - z_{k-1}\right\|^2 + \left\|z_{k-1} - z_{k-2}\right\|^2 + \left\|z_{k-2} - z_{k-3}\right\|^2\right), \tag{B.6}$$

where $\rho = \frac{1}{p}$, $V_\Upsilon = 6\left(1 - \frac{1}{p}\right) M^2(1 + 2\gamma_1^2 + \gamma_2^2)$.

Proof This is a direct result of Lemma B.1. □

Lemma B.3 (Convergence of estimator) *If $\{z_k\}_{k\in\mathbb{N}}$ satisfies $\lim_{k\to\infty}\mathbb{E}\left\|z_k - z_{k-1}\right\|^2 = 0$, then $\mathbb{E}\Upsilon_k \to 0$ and $\mathbb{E}\Gamma_k \to 0$ as $k \to \infty$.*

Proof By (B.6), we have

$$\begin{aligned}&\mathbb{E}\Upsilon_k\\ \le& (1-\rho)\,\mathbb{E}\Upsilon_{k-1} + V_\Upsilon\mathbb{E}\left(\left\|z_k - z_{k-1}\right\|^2 + \left\|z_{k-1} - z_{k-2}\right\|^2 + \left\|z_{k-2} - z_{k-3}\right\|^2 + \left\|z_{k-3} - z_{k-4}\right\|^2\right)\\ \le& V_\Upsilon\sum_{l=1}^{k}(1-\rho)^{k-l}\,\mathbb{E}\left(\left\|z_l - z_{l-1}\right\|^2 + \left\|z_{l-1} - z_{l-2}\right\|^2 + \left\|z_{l-2} - z_{l-3}\right\|^2 + \left\|z_{l-3} - z_{l-4}\right\|^2\right),\end{aligned}$$

which implies $\mathbb{E}\Upsilon_k \to 0$ as $k \to \infty$. By Jensen's inequality, we have $\mathbb{E}\Gamma_k \to 0$ as $k \to \infty$. □

Acknowledgements We are grateful to the editor and reviewers for their comments that improved the quality of our paper.

Author Contributions All authors contributed to the manuscript and approved the submitted version.

Funding This work is supported by the Scientific Research Project of Tianjin Municipal Education Commission (2022ZD007).

Data Availability All data generated or analyzed during this study are included in this article.

Declarations

Ethics approval Not applicable

Conflict of interest The authors declare no competing interests.

References

1. Chao, M.T., Han, D.R., Cai, X.J.: Convergence of the Peaceman-Rachford splitting method for a class of nonconvex programs. Numer. Math. Theory Methods Appl. **14**(2), 438–460 (2021)
2. Fu, X., Huang, K., Sidiropoulos, N.D., Ma, W.: Nonnegative matrix factorization for signal and data analytics: identifiability, algorithms, and applications. IEEE Signal Process. Mag. **36**(2), 59–80 (2019)
3. Paatero, P., Tapper, U.: Positive matrix factorization: a nonnegative factor model with optimal utilization of error estimates of data values. Environmetrics **5**, 111–126 (1994)
4. Lee, D.D., Seung, H.S.: Learning the parts of objects by nonnegative matrix factorization. Nat. **401**, 788–791 (1999)
5. Ma, Y., Hu, X., He, T., Jiang, X.: Clustering and integrating of heterogeneous microbiome data by joint symmetric nonnegative matrix factorization with Laplacian regularization. IEEE/ACM Trans. Comput. Biol. Bioinform. **17**(3), 788–795 (2020)
6. Pock, T., Sabach, S.: Inertial proximal alternating linearized minimization (iPALM) for nonconvex and nonsmooth problems. SIAM J. Imaging Sci. **9**, 1756–1787 (2017)
7. Aspremont, A., Ghaoui, L. E., Jordan, M. I., Lanckriet, G. R.: A direct formulation for sparse PCA using semidefinite programming. in Advances in Neural Information Processing Systems 41–48 (2005)
8. Zou, H., Hastie, T., Tibshirani, R.: Sparse principal component analysis. J. Comput. Graph. Statist. **15**, 265–286 (2006)
9. Attouch, H., Bolte, J., Svaiter, B.F.: Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Guass-Seidel methods. Math. Program. **137**, 91–129 (2013)
10. Donoho, D.L.: Compressed sensing. IEEE Trans. Inform. Theory **4**, 1289–1306 (2006)
11. Bolte, J., Sabach, S., Teboulle, M.: Proximal alternating linearised minimization for nonconvex and nonsmooth problems. Math. Program. **146**, 459–494 (2014)
12. Attouch, H., Bolte, J., Redont, P., Soubeyran, A.: Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the Kurdyka-Łojasiewicz inequality. Math. Oper. Res. **35**, 438–457 (2010)
13. Attouch, H., Bolte, J.: On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. Math. Program. Publ. Math. Program. Soc. **116**(1–2), 5–16 (2009)
14. Gao, X., Cai, X.J., Han, D.R.: A Gauss-Seidel type inertial proximal alternating linearized minimization for a class of nonconvex optimization problems. J. Glob. Optim. **76**, 863–887 (2020)
15. Wang, Q.X., Han, D.R.: A generalized inertial proximal alternating linearized minimization method for nonconvex nonsmooth problems. Appl. Numer. Math. **189**, 66–87 (2023)
16. Zhao, J., Dong, Q.L., Michael, Th.R., Wang, F.H.: Two-step inertial Bregman alternating minimization algorithm for nonconvex and nonsmooth problems. J. Glob. Optim. **84**, 941–966 (2022)
17. Guo, C. Z., Zhao, J.: Two-step inertial Bregman proximal alternating linearized minimization algorithm for nonconvex and nonsmooth problems, (2023). arXiv:2306.07614v1
18. Chao, M.T., Nong, F.F., Zhao, M.Y.: An inertial alternating minimization with Bregman distance for a class of nonconvex and nonsmooth problems. J. Appl. Math. Comput. **69**, 1559–1581 (2023)
19. Mukkamala, M.C., Ochs, P., Pock, T., Sabach, S.: Convex-concave backtracking for inertial Bregman proximal gradient algorithms in nonconvex optimization. SIAM J. Math. Data Sci. **2**, 658–682 (2020)
20. Ahookhosh, M., Hien, L.T.K., Gillis, N., Patrinos, P.: A block inertial Bregman proximal algorithm for nonsmooth nonconvex problems with application to symmetric nonnegative matrix tri-factorization. J. Optim. Theory Appl. **190**, 234–258 (2021)
21. Bottou, L.: In: Large-scale machine learning with stochastic gradient descent. In Proceedings of COMP-STAT'2010, **1**, 177–186 (2010)
22. Xu, Y., Yin, W.: Block stochastic gradient iteration for convex and nonconvex optimization. SIAM J. Optim. **25**(3), 1686–1716 (2015)

23. Driggs, D., Tang, J.Q., Liang, J.W., Davies, M., Schonlieb, C.B.: SPRING: a stochastic proximal alternating minimization for nonsmooth and nonconvex optimization. SIAM J. Imaing Sci. **4**, 1932–1970 (2021)
24. Hertrich, J., Steidl, G.: Inertial stochastic PALM and applications in machine learning. Sampl. Theory Sign Process. Data Anal. **20**, (2022). https://doi.org/10.1007/s43670-022-00021-x
25. Schmidt, M., Le Roux, N., Bach, F.: Minimizing finite sums with the stochastic average gradient. Math. Program. **162**, 83–112 (2017)
26. Johnson, R., Zhang, T.: Accelerating stochastic gradient descent using predictive variance reduction. in Advances in Neural Information Processing Systems 315–323 (2013)
27. Konecny, J., Liu, J., Richtarik, P., Takac, M.: Mini-batch semi-stochastic gradient descent in the proximal setting. IEEE J. Sel. Top. Sign Process. **10**, 242–255 (2016)
28. Defazio, A., Bach, F., Lacoste-Julien, S.: SAGA: a fast incremental gradient method with support for non-strongly convex composite objectives. in Advances in Neural Information Processing Systems 1646–1654 (2014)
29. Li, B., Ma, M., Giannakis, G. B.: On the convergence of SARAH and beyond. in International Conference on Artificial Intelligence and Statistics, PMLR 223–233 (2020)
30. Nguyen, L. M., Liu, J., Scheinberg, K., and Tak*áĉ*, M.: SARAH: a novel method for machine learning problems using stochastic recursive gradient. in Proceedings of the 34th International Conference on Machine Learning, 2613–2621 (2017)
31. Rockafellar, R.T., Wets, R.: Variational analysis, Grundlehren der Mathematischen Wissenschaften, vol. 317. Springer, Berlin (1998)
32. Bolte, J., Daniilidis, A., Lewis, A.: The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. SIAM J. Optim. **17**, 1205–1223 (2007)
33. Bolte, J., Daniilidis, A., Ley, O., Mazet, L.: Characterizations of Lojasiewicz inequalities: subgradient flows, talweg, convexity. Trans. Amer. Math. Soc. **362**, 3319–3363 (2010)
34. Bertsekas, D.P., Tsitsiklis, J.N.: Parallel and distributed computation: numerical methods. Prentice hall, Englewood Cliffs, NJ (1989)
35. Robbins, H., Siegmund, D.: A convergence theorem for non-negative almost supermartingales and some applications. Optimizing Methods in Statistics, Academic Press, New York, 233–257 (1971)
36. Damek, D.: The asynchronous PALM algorithm for nonsmooth nonconvex problems (2016). arXiv:1604.00526
37. Attouch, H., Bolte, J.: On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. Math. Program. B **116**, 5–16 (2007)
38. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nat. 788–791 (1999)
39. Pan, J., Gillis, N.: Generalized separable nonnegative matrix factorization. IEEE Trans. Pattern Anal. Mach. Intell. **43**(5), 1546–1561 (2021)
40. Rousset, F., Peyrin, F., Ducros, N.: A semi nonnegative matrix factorization technique for pattern generalization in single-pixel imaging. IEEE Trans. Comput. Imaging **4**(2), 284–294 (2018)
41. Peharz, R., Pernkopf, F.: Sparse nonnegative matrix factorization with l_0-constraints. Neurocomput. **80**, 38–46 (2012)