# A stochastic proximal alternating method for non-smooth non-convex optimization

**Citation for published version:**
Driggs, D, Tang, J, Liang, J, Davies, M & Schönlieb, C-B 2021, 'A stochastic proximal alternating method for non-smooth non-convex optimization', *Siam journal on imaging sciences*, vol. 14, no. 4, pp. 1932–1970. https://doi.org/10.1137/20M1387213

**Digital Object Identifier (DOI):**
10.1137/20M1387213



**Published In:**
Siam journal on imaging sciences

# A Stochastic Proximal Alternating Minimization Algorithm for Non-smooth and Non-convex Optimization*

Derek Driggs[†] [‡], Junqi Tang[†] [§], Jingwei Liang[¶], Mike Davies[§], and Carola-Bibiane Schönlieb[‡]

**Abstract.** In this work, we introduce a novel stochastic *proximal alternating linearized minimization* (PALM) algorithm [6] for solving a class of non-smooth and non-convex optimization problems. Large-scale imaging problems are becoming increasingly prevalent due to the advances in data acquisition and computational capabilities. Motivated by the success of stochastic optimization methods, we propose a stochastic variant of proximal alternating linearized minimization. We provide global convergence guarantees, demonstrating that our proposed method with variance-reduced stochastic gradient estimators, such as SAGA [16] and SARAH [27], achieves state-of-the-art oracle complexities. We also demonstrate the efficacy of our algorithm via several numerical examples including sparse non-negative matrix factorization, sparse principal component analysis and blind image deconvolution.

**Key words.** Non-convex and non-smooth optimization, Stochastic optimization, Variance reduction, Alternating minimization, Stochastic PALM, Kurdyka-Łojasiewicz inequality, Sparse principle component analysis

**AMS subject classifications.** 90C26, 90C15, 90C30, 49M27

## 1. Introduction.

With the advent of large-scale machine learning, developing efficient and reliable algorithms for (empirical) risk minimization has become an intense focus of the optimization community. These tasks involve minimizing a loss function measuring the fit between observed data, $x$, and a model's predicted result, $b$: $\min_{x\in\mathbb{R}^{m_1}} \frac{1}{n}\sum_{i=1}^n F(x_i, b_i)$ where $n$ denotes the number of samples and $F$ is the loss function. The two defining qualities of these problems are their large scale (in many applications, $n$ is on the order of billions), and finite-sum structure.

When the value of $n$ is very large, computing the gradient of the loss function is often prohibitively expensive, rendering most traditional deterministic first-order optimization algorithms ineffective. Over the years, randomized optimization algorithms [7, 32] have become increasingly popular due to their efficiency and simplicity. For these algorithms, the full gradient is replaced by a stochastic approximation that is cheap to compute, so that their per-iteration complexity grows slowly with $n$. For objectives with a finite-sum structure, many works have shown that certain randomized algorithms achieve convergence rates similar to those of full-gradient methods, even though their per-iteration complexity is often a factor of $n$ smaller [16, 21, 38].

Outside machine learning, objectives with a finite-sum structure also arise in problems from image processing and computer vision. Recently, randomized optimization algorithms have been explored for image processing tasks including PET reconstruction, deblurring and tomography [12, 36]. As stochastic methods expand into new applications, they move further from smooth, strongly convex finite-sum objectives where they are well-understood theoretically. In this work, we aim to provide a

[†]Contributed equally

*Submitted to the editors DATE.

[‡]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK (d.driggs@maths.cam.ac.uk, cbs31@cam.ac.uk).

[§]School of Engineering, University of Edinburgh, Edinburgh, UK (j.tang@ed.ac.uk, mike.davies@ed.ac.uk).

[¶]Institute of Natural Sciences and School of Mathematical Sciences, Shanghai Jiao Tong University, Shanghai, China (jingwei.liang@sjtu.edu.cn).

better understanding of stochastic algorithms for problems that are neither smooth nor convex.

### 1.1. Non-smooth, non-convex optimization.

Our goal is to minimize composite objectives of the following form:

$$\min_{x\in\mathbb{R}^{m_1}, y\in\mathbb{R}^{m_2}} \left\{ \Phi(x,y) \stackrel{\text{def}}{=} J(x) + F(x,y) + R(y) \right\}, \tag{1.1}$$

where $F(x,y) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n F_i(x,y)$ has a finite-sum structure. In general, functions $J$ and $R$ are non-smooth regularizations that promote structures in the solutions, *e.g.* sparsity or non-negativity. The blocks $x$ and $y$ represent differently structured elements of the solution that are coupled through the loss term, $F(x,y)$. Throughout this work, we impose the following assumptions:

- **(A.1)** $J : \mathbb{R}^{m_1} \to \mathbb{R} \cup \{+\infty\}$ and $R : \mathbb{R}^{m_2} \to \mathbb{R} \cup \{+\infty\}$ are proper lower semi-continuous (lsc) functions that are bounded from below;
- **(A.2)** $F_i : \mathbb{R}^{m_1} \times \mathbb{R}^{m_2} \to \mathbb{R}$ are finite-valued, differentiable, and their gradients $\nabla F_i$ are $M(\mathcal{X},\mathcal{Y})$-Lipschitz continuous on bounded sets $\mathcal{X}\times\mathcal{Y} \subset \mathbb{R}^{m_1}\times\mathbb{R}^{m_2}$ for all $i \in \{1,\cdots,n\}$;[1]
- **(A.3)** The partial gradients $\nabla_x F_i$ are Lipschitz continuous with modulus $L_1(y)$, and $\nabla_y F_i$ are Lipschitz continuous with modulus $L_2(x)$ for all $i \in \{1,\cdots,n\}$;
- **(A.4)** The function $\Phi$ is bounded from below.

No convexity is imposed on any of the functions involved. Problem (1.1) departs from the sum-of-convex-objectives models that populate the majority of the optimization literature. Many models in machine learning, statistics and image processing require the full generality of (1.1). Archetypal examples include non-negative or sparse matrix factorization [20], Sparse PCA [13, 42], Robust PCA [11], trimmed least-squares [1] and blind image deconvolution [10]. Despite the prevalence of these problems, few numerical methods can solve the general problem (1.1), and none that realize match the efficiency that randomized algorithms provide. We outline some existing options below.

*Proximal alternating minimization.* One approach to solve (1.1) is the Proximal Alternating Minimization (PAM) method [3], whose iterations take the following form:

$$\begin{aligned} x_{k+1} &\in \operatorname{Argmin}_{x\in\mathbb{R}^{m_1}} \left\{ \Phi(x,y_k) + \tfrac{1}{2\gamma_{x,k}}\|x-x_k\|^2 \right\}, \\ y_{k+1} &\in \operatorname{Argmin}_{y\in\mathbb{R}^{m_2}} \left\{ \Phi(x_{k+1},y) + \tfrac{1}{2\gamma_{y,k}}\|y-y_k\|^2 \right\}, \end{aligned} \tag{1.2}$$

where $\gamma_{x,k}, \gamma_{y,k} > 0$ are step-sizes. A significant limitation of PAM is that the subproblems in (1.2) do not have closed-form solutions in general. As a consequence, each subproblem requires its own set of inner iterations, which makes PAM inefficient in practice.

*Proximal alternating linearized minimization [6].* To circumvent this limitation of PAM, Proximal Alternating Linearized Minimization (PALM) [6] replaces PAM's two subproblems with their proximal linearizations. PALM's iterations take the form

$$\begin{aligned} x_{k+1} &\in \operatorname{prox}_{\gamma_{x,k}J}\big(x_k - \gamma_{x,k}\nabla_x F(x_k,y_k)\big), \\ y_{k+1} &\in \operatorname{prox}_{\gamma_{y,k}R}\big(y_k - \gamma_{y,k}\nabla_y F(x_{k+1},y_k)\big), \end{aligned} \tag{1.3}$$

[1] Because we consider a particular bounded set in our analysis, we drop the dependence on $\mathcal{X}$ and $\mathcal{Y}$ for the remainder of the paper, writing the Lipschitz constant as $M$.

**Algorithm 1.1** SPRING: **S**tochastic **Pr**oximal Alternat**ing** Linearized Minimization

**Initialize**: $x_0 \in \mathbb{R}^{m_1}, y_0 \in \mathbb{R}^{m_2}$.
**for** $k = 1, 2, \cdots, T-1$ **do**
  $x_{k+1} \in \mathrm{prox}_{\gamma_{x,k} J}\big(x_k - \gamma_{x,k} \widetilde{\nabla}_x(x_k, y_k)\big)$
  $y_{k+1} \in \mathrm{prox}_{\gamma_{y,k} R}\big(y_k - \gamma_{y,k} \widetilde{\nabla}_y(x_{k+1}, y_k)\big)$
**end for**
**return** $(x_T, y_T)$

where $\nabla_x F$ and $\nabla_y F$ are partial derivatives, and $\mathrm{prox}_{\gamma_{x,k} J}$ is called "proximal operator" of $J$ and defined by

$$\mathrm{prox}_{\gamma J}(\cdot) \stackrel{\text{def}}{=} \mathrm{Argmin}_x \gamma J(x) + \tfrac{1}{2}\|x - \cdot\|^2.$$

The proximal mapping is set-valued in general, and becomes single-valued if $J$ is convex.

In contrast to PAM, each subproblem of PALM can be efficiently computed if the proximal maps of $J$ and $R$ are easy to calculate, which is true in many applications. PALM also has the same convergence guarantees as PAM, so linearizing $F$ in each proximal step is a clear improvement over PAM. PALM with momentum is considered in [29], where the authors show that inertia allows PALM to converge to critical points with lower objective values, although accelerated rates might not be obtained.

### 1.2. Stochastic PALM.

In this work, we introduce SPRING, a randomized version of PALM where the partial gradients $\nabla_x F(x_k, y_k)$ and $\nabla_y F(x_{k+1}, y_k)$ in (1.3) are replaced by random estimates, $\widetilde{\nabla}_x(x_k, y_k)$ and $\widetilde{\nabla}_y(x_{k+1}, y_k)$, formed using the gradients of only a few indices $\nabla_x F_j(x_k, y_k)$ and $\nabla_y F_j(x_{k+1}, y_k)$ for $j \in B_k \subset \{1, 2, \cdots, n\}$. The mini-batch $B_k$ is chosen uniformly at random from all subsets of $\{1, 2, \cdots, n\}$ with cardinality $b$. We describe SPRING in Algorithm 1.1.

Many different gradient estimators can be used in SPRING. The simplest one is the stochastic gradient descent (SGD) estimator [33]

$$\widetilde{\nabla}_x^{\text{SGD}}(x_k, y_k) = \tfrac{1}{b} \textstyle\sum_{j \in B_k} \nabla_x F_j(x_k, y_k),$$

which uses the gradient of a randomly sampled batch to represent the full gradient. Another popular choice is SAGA gradient estimator [16], which incorporates the gradient history:

$$\begin{aligned} \widetilde{\nabla}_x^{\text{SAGA}}(x_k, y_k) &= \tfrac{1}{b} \textstyle\sum_{j \in B_k} \big(\nabla_x F_j(x_k, y_k) - g_{k,j}\big) + \tfrac{1}{n} \textstyle\sum_{i=1}^n g_{k,i}, \\ g_{k+1,i} &= \begin{cases} \nabla_x F_i(x_k, y_k) & \text{if } i \in B_k, \\ g_{k,i} & \text{o.w.} \end{cases} \end{aligned}$$

Both SGD and SAGA estimators are *unbiased*. The last gradient estimator we specifically consider in this work is the (loopless) SARAH estimator [24, 27], $\widetilde{\nabla}_x^{\text{SARAH}}(x_k, y_k)$, which is *biased*.

$$\begin{cases} \nabla_x F(x_k, y_k) & \text{w.p. } \frac{1}{p} \\ \frac{1}{b} \sum_{j \in B_k} \big(\nabla_x F_j(x_k, y_k) - \nabla_x F_j(x_{k-1}, y_{k-1})\big) + \widetilde{\nabla}_x^{\text{SARAH}}(x_{k-1}, y_{k-1}) & \text{o.w.} \end{cases}$$

Here, $p$ is a tuning parameter that is generally set to $\mathcal{O}(n)$. Other variance-reduced estimators can be used in SPRING, including the SAG [34] and SVRG [21] estimators, for example, but we consider only the SAGA and SARAH estimators specifically in this work.

Computing the full gradient is generally $n$-times more expensive than computing $\nabla_x F_i$, so when $n$ is large and $b \ll n$, each step of SPRING with any of these estimators is significantly less expensive than that of PALM.

*Remark* 1.1. Although we consider only two variable blocks in (1.1), the results of this paper easily extend to an arbitrary number of blocks to solve problems of the form

$$\min_{x_1,\cdots,x_\ell} \Big\{ \tfrac{1}{n} \textstyle\sum_{i=1}^n F_i(x_1,\cdots,x_\ell) + \sum_{t=1}^\ell R_t(x_t) \Big\},$$

where each $R_t$ is a (possibly non-smooth) regularizer.

## 1.3. Contributions.

By combining PALM with popular stochastic gradient estimators which are variance reduced, we proposed a novel stochastic algorithm for non-convex and non-smooth optimization. Theoretically, we show that the resulted algorithm matches the convergence rates of PALM given that the gradient estimators $\widetilde{\nabla}_x$ and $\widetilde{\nabla}_y$ satisfy a *variance-reduced* property (see Definition 2.1). We prove convergence guarantees of two types.

*Convergence rate of generalized gradient map.* Given a point $z = (x, y)$, the *generalized gradient map* of PALM/SPRING is defined as

$$\mathcal{G}_{\gamma_1,\gamma_2}(z) \stackrel{\text{def}}{=} \begin{pmatrix} 1/\gamma_1\big(x - \mathrm{prox}_{\gamma_1 J}(x - \gamma_1 \nabla_x F(x,y))\big) \\ 1/\gamma_2\big(y - \mathrm{prox}_{\gamma_2 R}(y - \gamma_2 \nabla_y F(x,y))\big) \end{pmatrix}, \tag{1.4}$$

where $\gamma_1, \gamma_2 > 0$ are parameters (not necessarily equal to the step-sizes in Algorithm 1.1). If $\mathrm{dist}(0, \mathcal{G}_{\gamma_1,\gamma_2}(z)) = 0$, then by the definition of the proximal operator, $0 \in (\nabla_x F(x,y) + \partial J(x), \nabla_y F(x,y) + \partial R(y)) = \partial\Phi(z)$, meaning $z$ is a critical point. The point $z$ is an *$\epsilon$-approximate critical point* if it satisfies $\mathrm{dist}(0, \mathcal{G}_{\gamma_1,\gamma_2}(z)) \le \epsilon$ *for some* $\gamma_1, \gamma_2 \in (0, \infty)$.[2] In Section 3, we show that[3]

$$\mathbb{E}[\mathrm{dist}(0, \mathcal{G}_{\frac{\gamma_{x,\alpha}}{2},\frac{\gamma_{y,\alpha}}{2}}(z_\alpha))^2] \le \mathcal{O}\big(\tfrac{1}{k}\big),$$

where $\alpha$ is chosen uniformly at random from the set $\{1, 2, \cdots, k\}$. If $\Phi$ satisfies a certain error bound involving the generalized gradient map (see Eq. (3.1)), then SPRING converges linearly to the global optimum. These results generalize almost all existing results for stochastic gradient methods on non-convex, non-smooth objectives [1, 18, 30, 37, 41].

Specializing these convergence guarantees to specific gradient estimators, the constants appearing in these rates scale with the mean-squared error (MSE, see Definition 2.1) of the gradient estimators.

- For the SAGA estimator with $b \le \mathcal{O}(n^{2/3})$, the iterates of SPRING satisfy

$$\mathbb{E}[\mathrm{dist}(0, \mathcal{G}_{\frac{\gamma_{x,\alpha}}{2},\frac{\gamma_{y,\alpha}}{2}}(z_\alpha))^2] \le \mathcal{O}\big(\tfrac{n^2 L}{b^3 k}\big).$$

- For the SARAH gradient estimator with any batch size, we have

$$\mathbb{E}[\mathrm{dist}(0, \mathcal{G}_{\frac{\gamma_{x,\alpha}}{2},\frac{\gamma_{y,\alpha}}{2}}(z_\alpha))^2] \le \mathcal{O}\big(\tfrac{\sqrt{n}L}{k}\big).$$

[2] The set of $\epsilon$-critical points depends on the parameters $\gamma_1, \gamma_2$, with larger parameter values generally increasing the size of the set for fixed $\epsilon$. For fixed and bounded $\gamma_1$ and $\gamma_2$, the generalized gradient map provides a notion of distance to a critical point. If $\mathcal{S}(\epsilon)$ is the set of $\epsilon$-critical points, then with $\gamma_1$ and $\gamma_2$ fixed and bounded, we have $\mathcal{S}(\epsilon_1) \subset \mathcal{S}(\epsilon_2)$ for $\epsilon_1 \le \epsilon_2$, and as $\epsilon \to 0$, $\mathcal{S}(\epsilon)$ contains only the set of critical points of $\Phi$.

[3] We prove bounds on the expectation of the *squared* norm of the generalized gradient map to facilitate comparisons with existing results [1, 30, 31].

These convergence rates imply complexity bounds with respect to a *stochastic first-order oracle* (SFO) which returns the partial gradient of a single component $F_i$ (for example, $\nabla_x F_i(x_k, y_k)$). To find an $\epsilon$-approximate critical point, SAGA with a mini-batch of size $n^{2/3}$ requires no more than $\mathcal{O}(n^{2/3}L/\epsilon^2)$ SFO calls, and SARAH requires no more than $\mathcal{O}(\sqrt{n}L/\epsilon^2)$. The improved dependence on $n$ when using SARAH gradient estimator exists in all of our convergence rates for SPRING. Because most existing works on stochastic optimization for non-smooth, non-convex problems use models that are special cases of (1.1), our results for SPRING capture most existing work as special cases. In particular, in the case $R \equiv J \equiv 0$, our results recover recent results showing that SARAH achieves the *oracle complexity lower-bound* for non-convex problems with a finite-sum structure [18, 28, 37, 40, 41].

*Convergence under the Kurdyka–Łojasiewicz property.* We also provide convergence guarantees under the Kurdyka–Łojasiewicz property (see Definition 2.4). First, we prove the global convergence of the generated sequence under the assumption that the objective function $\Phi(x, y)$ of (1.1) has the Kurdyka–Łojasiewicz property. Then, under the assumption that $\Phi$ is semi-algebraic with KL-exponent $\theta$ (see Section 2), we show that the sequence $z_k = (x_k, y_k)$ generated by SPRING converges in expectation to a critical point $z^\star$ of problem (1.1) at the following rates:

- If $\theta = 0$, then $\{\mathbb{E}\Phi(z_k)\}_{k\in\mathbb{N}}$ converges to $\mathbb{E}\Phi(z^\star)$ in a finite number of steps.
- If $\theta \in (0, 1/2]$, then $\mathbb{E}\|z_k - z^\star\| \le \mathcal{O}(\tau^k)$ for some $\tau \in (0, 1)$.
- If $\theta \in (1/2, 1)$, then $\mathbb{E}\|z_k - z^\star\| \le \mathcal{O}(k^{-\frac{1-\theta}{2\theta-1}})$.

These rates match the rates of the original PALM algorithm.

### 1.4. Prior Art.

SPRING offers several advantages over existing stochastic algorithms for non-smooth non-convex optimization. Reddi *et al.* investigate proximal SAGA and SVRG for solving problems of the form (1.1) when $y$ is constant and $J$ is convex [30]. Using mini-batches of size $b = n^{2/3}$, SAGA and SVRG require $\mathcal{O}(n^{2/3}L/\epsilon^2)$ stochastic gradient evaluations to converge to an $\epsilon$-approximate critical point. Similarly, Aravkin and Davis introduce TSVRG, a stochastic algorithm based on SVRG gradient estimator, for solving another special case of (1.1) [1]. Our work generalizes their results and improves them in many cases. Most importantly, we show that using SARAH gradient estimator allows SPRING to achieve a complexity of $\mathcal{O}(\sqrt{n}L/\epsilon^2)$ even when the mini-batch size is equal to one. Our results for semi-algebraic objectives offer even sharper convergence rates.

The block stochastic gradient method [39] is closely related to SPRING using the (non-variance-reduced) SGD gradient estimator. In a similar work, Davis *et al.* introduce SAPALM, an asynchronous version of PALM that allows stochastic noise in the gradients [15]. The authors prove convergence rates that scale with the variance of the noise in the gradients, with their best complexity bound for finding an $\epsilon$-approximate critical point equal to $\mathcal{O}(nL/\epsilon^2)$. While significant in their own right, these results are not directly related to ours, as these works require an explicit bound on the variance of the noise in the gradients, and the gradient estimators we consider do not admit such a bound [15].

## 2. Preliminaries.

We use the following definitions and notation throughout the manuscript.

*Variance Reduction.* In our analysis, we mainly focus on stochastic gradient estimators that are variance reduced. We use a general definition of a variance-reduced gradient estimator that includes all existing estimators, for example, SAGA and SARAH, as special cases.

Definition 2.1 (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 1.1 with some gradient estimator $\widetilde{\nabla}$. This gradient estimator is* variance-reduced *with constants $V_1, V_2, V_\Upsilon \ge 0$, and $\rho \in (0, 1]$ if it satisfies the following conditions:*

1. **(MSE Bound)** *There exists a sequence of random variables* $\{\Upsilon_k\}_{k\geq 1}$ *of the form* $\Upsilon_k = \sum_{i=1}^s (v_k^i)^2$ *for some non-negative random variables* $v_k^i \in \mathbb{R}$ *such that*

$$\mathbb{E}_k[\|\widetilde{\nabla}_x(x_k,y_k) - \nabla_x F(x_k,y_k)\|^2 + \|\widetilde{\nabla}_y(x_{k+1},y_k) - \nabla_y F(x_{k+1},y_k)\|^2] \leq \Upsilon_k + V_1(\mathbb{E}_k\|z_{k+1}-z_k\|^2 + \|z_k - z_{k-1}\|^2), \tag{2.1}$$

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

$$\mathbb{E}_k[\|\widetilde{\nabla}_x(x_k,y_k) - \nabla_x F(x_k,y_k)\| + \|\widetilde{\nabla}_y(x_{k+1},y_k) - \nabla_y F(x_{k+1},y_k)\|] \leq \Gamma_k + V_2(\mathbb{E}_k\|z_{k+1}-z_k\| + \|z_k - z_{k-1}\|).$$

2. **(Geometric Decay)** *The sequence* $\{\Upsilon_k\}_{k\geq 1}$ *decays geometrically:*

$$\mathbb{E}_k\Upsilon_{k+1} \leq (1-\rho)\Upsilon_k + V_\Upsilon(\mathbb{E}_k\|z_{k+1}-z_k\|^2 + \|z_k - z_{k-1}\|^2). \tag{2.2}$$

3. **(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* $\mathbb{E}\Gamma_k \to 0$.

Proposition 2.2. *SAGA gradient estimator is variance-reduced with parameters* $V_1 = 6M^2/b$, $V_2 = \sqrt{6}M/\sqrt{b}$, $V_\Upsilon = \frac{134nL^2}{b^2}$, *and* $\rho = \frac{b}{2n}$. *SARAH estimator is variance-reduced with parameters* $V_1 = V_\Upsilon = 2L^2$, $V_2 = 2L$, *and* $\rho = 1/p$.

Proposition 2.2 is a generalization of existing variance bounds for these estimators. For a derivation of the constants appearing in Proposition 2.2, we refer the reader to our full work [17].

*Remark* 2.3. Our results allow Algorithm 1.1 to use any variance-reduced gradient estimator, even different estimators for $\nabla_x$ and $\nabla_y$. In particular, it is possible to use different mini-batch sizes when approximating the two partial gradients.

*Kurdyka–Łojasiewicz property.* Let $H : \mathbb{R}^{m_1} \to \mathbb{R}\cup\{+\infty\}$ be a proper lower semicontinuous function. For $\epsilon_1, \epsilon_2$ satisfying $-\infty < \epsilon_1 < \epsilon_2 < +\infty$, define the set $[\epsilon_1 < H < \epsilon_2] \stackrel{\text{def}}{=} \{x \in \mathbb{R}^{m_1} : \epsilon_1 < H(x) < \epsilon_2\}$.

Definition 2.4 (Kurdyka–Łojasiewicz). *A function* $H$ *is said to have the Kurdyka-Łojasiewicz property at* $\bar{x} \in \mathrm{dom}(H)$ *if there exists* $\epsilon \in (0,+\infty]$, *a neighborhood* $U$ *of* $\bar{x}$ *and a continuous concave function* $\varphi : [0,\epsilon) \to \mathbb{R}_+$ *such that*
(i) $\varphi(0) = 0$, $\varphi$ *is* $C^1$ *on* $(0,\epsilon)$, *and for all* $r \in (0,\epsilon)$, $\varphi'(r) > 0$;
(ii) *for all* $x \in U \cap [H(\bar{x}) < H < H(\bar{x}) + \epsilon]$, *the Kurdyka–Łojasiewicz inequality holds:*

$$\varphi'\big(H(x) - H(\bar{x})\big)\mathrm{dist}\big(0, \partial H(x)\big) \geq 1. \tag{2.3}$$

*If* $H$ *satisfies the KL property at each point of* $\mathrm{dom}(\partial H)$, *then it is called KL functions.*

Roughly speaking, KL functions become sharp up to reparameterization via $\varphi$, a *desingularizing function* for $H$. Typical KL functions include the class of semi-algebraic functions [4, 5]. For instance, the $\ell_0$ pseudo-norm and the rank function are KL. Semi-algebraic functions admit desingularizing functions of the form $\varphi(r) = ar^{1-\theta}$ for $a > 0$, and $\theta \in [0,1)$ is known as the *KL exponent* of the function [4, 6]. For these functions, the KL inequality reads

$$\big(H(x) - H(\overline{x})\big)^\theta \leq C\|\zeta\| \qquad \forall \zeta \in \partial H(x), \tag{2.4}$$

for some $C > 0$. In the case $H(x) = H(\overline{x})$, we use the convention $0^0 \stackrel{\text{def}}{=} 0$.

*Bounded Iterates.* Many of our results require the assumption that the iterates generated by SPRING are bounded, in addition to assumptions (**A.1**)-(**A.4**). Because assumption (**A.2**) only requires $\nabla F_i$ to be Lipschitz on bounded sets, assuming the iterates are bounded allows us to say $\nabla F_i$ is $M$-Lipschitz continuous. We also require boundedness of the iterates to ensure that a limit point of this sequence exists during the proof of Lemma 4.3. This assumption is required for the same reasons in the analysis of PALM. It is satisfied, for example, if $J$ and $R$ have bounded domains.

*Notation.* We use $\{(x_k, y_k)\}_{k\in\mathbb{N}}$ to denote the sequence generated by SPRING. We use $L_x \stackrel{\text{def}}{=} \max_{k\in\mathbb{N}} L_1\,(y_k)$, and define $L_y$ analogously. We set $\bar{L} \stackrel{\text{def}}{=} \max\{L_x, L_y\}$, $\overline{\gamma}_k \stackrel{\text{def}}{=} \max\{\gamma_{x,k}, \gamma_{y,k}\}$, $\underline{\gamma}_k \stackrel{\text{def}}{=} \min\{\gamma_{x,k}, \gamma_{y,k}\}$, and $\underline{\Phi} \stackrel{\text{def}}{=} \inf_{(x,y)\in\operatorname{dom}(\Phi)} \Phi(x,y)$. We also use $L$ to denote the maximum of $L_x$, $L_y$, and $M$ over the iterates generated by SPRING, so that $\bar{L}, M \le L$. We use $\mathbb{E}_k$ to denote the expectation conditional on the first $k$ iterations of SPRING. Specifically, $\mathbb{E}_k \equiv \mathbb{E}[\cdot|\mathcal{F}_k]$ where $\mathcal{F}_k$ is the $\sigma$-algebra generated by $B_0, \cdots, B_{k-1}$. We require a notion of the expectation of the subdifferential of $\Phi(z_k)$. To define this, let $\overline{n} = \binom{n}{b}$ be the number of possible gradient estimates in one iteration of Algorithm 1.1, and let $\{z_k^i\}_{i=1}^{\overline{n}^k}$ be the set of possible values for $z_k$.[4] We use the notation $\mathbb{E}\partial\Phi(z_k) = \partial\mathbb{E}\Phi(z_k) = \{\frac{1}{\overline{n}^k}\sum_{i=1}^{\overline{n}^k}\xi_i | \xi_i \in \partial\Phi(z_k^i)\}$. Every subgradient $\xi \in \partial\Phi(z_k)$ is of the form $\frac{1}{\overline{n}^k}\sum_{i=1}^{\overline{n}^k}\xi_i$ for $\xi_i \in \partial\Phi(z_k^i)$, and we denote this vector as $\mathbb{E}\xi \in \mathbb{E}\partial\Phi(z_k)$.

## 2.1. Elementary Lemmas.

The following lemmas generalize the sufficient decrease property of proximal gradient descent to the stochastic-gradient setting. They allow us to show that, if the MSE of the stochastic gradient estimator is small enough, then iteratively applying the proximal gradient operator decreases the suboptimality of each iterate in expectation.

Lemma 2.5. *Let $F : \mathbb{R}^m \to \mathbb{R}$ be a function with $L$-Lipschitz continuous gradient, $R : \mathbb{R}^m \to \mathbb{R}$ a proper lower semicontinuous function that is bounded from below, and $z \in \operatorname{prox}_{\eta R}(x - \eta d)$ for some $\eta > 0$ and $x, d \in \mathbb{R}^m$. Then, for all $y \in \mathbb{R}^m$,*

$$0 \le F(y)+R(y)-F(z)-R(z)+\langle\nabla F(x)-d, z-y\rangle+(\tfrac{L}{2}-\tfrac{1}{2\eta})\|x-z\|^2+(\tfrac{L}{2}+\tfrac{1}{2\eta})\|x-y\|^2. \tag{2.5}$$

*Proof.* By the Lipschitz continuity of $\nabla F$, we have the inequalities

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

Furthermore, by the definition of $z$,

$$z \in \operatorname{Argmin}_{v\in\mathbb{R}^m}\left\{\langle d, v-x\rangle + \tfrac{1}{2\eta}\|v-x\|^2 + R(v)\right\}.$$

Taking $v = y$, we obtain

$$0 \le R(y) - R(z) + \langle d, y-z\rangle + \tfrac{1}{2\eta}\left(\|x-y\|^2 - \|x-z\|^2\right).$$

Adding these three inequalities completes the proof. ∎

[4] When the proximal operator is multi-valued, Algorithm 1.1 requires one element to be chosen for each iterate, so we are not counting "possible" values for $z_k$ that arise from choosing other elements of the proximal operator.

If the full gradient estimator is used, Lemma 2.5 implies the well-known sufficient decrease property of proximal gradient descent. Using a gradient estimator, this decrease is offset by the estimator's MSE. The following lemma quantifies this relationship.

Lemma 2.6 (Sufficient Decrease Property). *Let $F$, $R$, and $z$ be defined as in Lemma 2.5. The following inequality holds for any $\lambda > 0$:*

$$0 \le F(x) + R(x) - F(z) - R(z) + \tfrac{1}{2L\lambda}\|d - \nabla F(x)\|^2 + \left(\tfrac{L(\lambda+1)}{2} - \tfrac{1}{2\eta}\right)\|x - z\|^2. \tag{2.6}$$

*Proof.* From Lemma 2.5 with $y = x$, we have

$$0 \le F(x) + R(x) - F(z) - R(z) + \langle \nabla F(x) - d, z - x\rangle + (\tfrac{L}{2} - \tfrac{1}{2\eta})\|x - z\|^2.$$

Using Young's inequality $\langle \nabla F(x) - d, z - x\rangle \le \frac{1}{2L\lambda}\|d - \nabla F(x)\|^2 + \frac{L\lambda}{2}\|x - z\|^2$ we obtain the desired result. ∎

As in a related work [14], we use the *supermartingale convergence theorem* to obtain almost sure convergence of sequences generated by SPRING. Below, we present an implication of this result adapted to our context. We refer to [14, Theorem 4.2] and [33, Theorem 1] for more general presentations.

Lemma 2.7 (Supermartingale Convergence). *Let $\{X_k\}_{k=0}^\infty$ and $\{Y_k\}_{k=0}^\infty$ be sequences of bounded non-negative random variables such that $X_k, Y_k$ are functions of only the first $k$ iterations of SPRING. If*

$$\mathbb{E}_k X_{k+1} + Y_k \le X_k, \tag{2.7}$$

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

## 3. Convergence rates of the generalized gradient map.

To begin, we present our analysis of the convergence rate of the generalized gradient map defined in (1.4). The following results of Theorem 3.1 generalize many existing convergence guarantees for stochastic gradient methods on non-convex, non-smooth objectives [1, 18, 30, 37, 41]. Recall that $\bar{L} \stackrel{\text{def}}{=} \max\{L_x, L_y\}$, $\overline{\gamma}_k \stackrel{\text{def}}{=} \max\{\gamma_{x,k}, \gamma_{y,k}\}$, $\underline{\gamma}_k \stackrel{\text{def}}{=} \min\{\gamma_{x,k}, \gamma_{y,k}\}$, and $\underline{\Phi} \stackrel{\text{def}}{=} \inf_{(x,y)\in\text{dom}(\Phi)} \Phi(x,y)$.

Theorem 3.1. *Suppose that assumptions (**A.1**)-(**A.4**) hold and that the sequence $\{(x_k, y_k)\}_{k\in\mathbb{N}}$ is bounded. Let $\widetilde{\nabla}_x$ and $\widetilde{\nabla}_y$ be variance-reduced gradient estimators following Definition 2.1.*

- *Suppose $\overline{\gamma}_k$ is non-increasing, and for all $k$, $\gamma_{y,k} < \frac{1}{4L_y+2M}$ and*

$$\overline{\gamma}_k \le \frac{1}{16}\sqrt{\frac{(\bar{L}+M)^2}{(V_1+V_\Upsilon/\rho)^2} + \frac{16}{(V_1+V_\Upsilon/\rho)}} - \frac{\bar{L}+M}{16(V_1+V_\Upsilon/\rho)},\ 0 < \beta \le \underline{\gamma}_k,\ \gamma_{x,k} < \frac{1}{4L_x},$$

  *With $\alpha$ chosen uniformly at random from $\{0, 1, \cdots, T-1\}$,*

$$\mathbb{E}[\text{dist}(0, \mathcal{G}_{\frac{\gamma_{x,\alpha}}{2},\frac{\gamma_{y,\alpha}}{2}}(z_\alpha))^2] \le \frac{4(\Phi(x_0,y_0) + \frac{2\overline{\gamma}_0}{\rho}\Upsilon_0)}{T\nu\beta^2},$$

  *where $\nu \stackrel{\text{def}}{=} \min\{\frac{1}{4\gamma_{x,0}} - L_x, \frac{1}{4\gamma_{y,0}} - \frac{M}{2} - L_y\}$.*
- *If, moreover, $\Phi$ satisfies the error bound*

$$\Phi(x_k, y_k) - \underline{\Phi} \le \mu\,\text{dist}\big(0, \mathcal{G}_{\frac{\gamma_{x,k}}{2},\frac{\gamma_{y,k}}{2}}(x_k, y_k)\big)^2, \tag{3.1}$$

*for all $k \in \mathbb{N}$, and $\overline{\gamma}_k$ satisfies*

$$\overline{\gamma}_k \leq \frac{1}{20}\sqrt{\frac{(\bar{L}+M)^2}{(V_1+V_\Upsilon/\rho)^2} + \frac{20}{(V_1+V_\Upsilon/\rho)}} - \frac{\bar{L}+M}{20(V_1+V_\Upsilon/\rho)},$$

*then the iterates of SPRING converge to the set of global minimizers of $\Phi$, and after $T$ iterations of Algorithm 1.1,*

$$\mathbb{E}[\Phi(x_T, y_T) - \underline{\Phi}] \leq (1-\Theta)^T(\Phi(x_0, y_0) - \underline{\Phi} + \tfrac{4\overline{\gamma}_0}{\rho}\Upsilon_0),$$

*where $\Theta \overset{def}{=} \min\{\frac{\nu\beta^2}{4\mu}, \frac{\rho}{2}\}$.*

*Remark* 3.2. We include convergence guarantees under the error bound (3.1) to compare with related works [1]. This error bound is similar to the Kurdyka–Łojasiewicz property for functions with a KL exponent of $1/2$, as can be seen comparing equation (3.1) to equation (2.4) with $\theta = 1/2$ and $H(\overline{x}) = \underline{\Phi}$. Although objectives satisfying this error bound could be non-convex, this condition ensures that convergence to the global minimum is guaranteed.

*Proof of Theorem 3.1, Part 1.* Let $\hat{x}_{k+1} \in \mathrm{prox}_{\frac{\gamma_{x,k}}{2}J}(x_k - \frac{\gamma_{x,k}}{2}\nabla_x F(x_k, y_k))$, and let $\hat{y}_{k+1} \in \mathrm{prox}_{\frac{\gamma_{y,k}}{2}R}(y_k - \frac{\gamma_{y,k}}{2}\nabla_y F(x_k, y_k))$. Applying Lemma 2.5 with $z = \hat{x}_{k+1}$, $y = x = x_k$ and $d = \nabla_x F(x_k, y_k)$, we have

$$F(\hat{x}_{k+1}, y_k) + J(\hat{x}_{k+1}) \leq F(x_k, y_k) + J(x_k) + (\tfrac{L_x}{2} - \tfrac{1}{\gamma_{x,k}})\|\hat{x}_{k+1} - x_k\|^2.$$

Again, applying Lemma 2.5 with $z = x_{k+1}$, $y = \hat{x}_{k+1}$, $x = x_k$, and $d = \widetilde{\nabla}_x(x_k, y_k)$, we obtain

$$\begin{aligned} F(x_{k+1}, y_k) + J(x_{k+1}) \leq F(\hat{x}_{k+1}, y_k) + J(\hat{x}_{k+1}) + \langle \nabla_x F(x_k, y_k) - \widetilde{\nabla}_x(x_k, y_k), x_{k+1} - \hat{x}_{k+1}\rangle \\ + (\tfrac{L_x}{2} - \tfrac{1}{2\gamma_{x,k}})\|x_{k+1} - x_k\|^2 + (\tfrac{L_x}{2} + \tfrac{1}{2\gamma_{x,k}})\|\hat{x}_{k+1} - x_k\|^2. \end{aligned}$$

Adding these two inequalities gives

$$\begin{aligned} & F(x_{k+1}, y_k) + J(x_{k+1}) \\ & \leq F(x_k, y_k) + J(x_k) + (L_x - \tfrac{1}{2\gamma_{x,k}})\|\hat{x}_{k+1} - x_k\|^2 + (\tfrac{L_x}{2} - \tfrac{1}{2\gamma_{x,k}})\|x_{k+1} - x_k\|^2 \\ & \quad + \langle \nabla_x F(x_k, y_k) - \widetilde{\nabla}_x(x_k, y_k), x_{k+1} - \hat{x}_{k+1}\rangle \\ & \overset{①}{\leq} F(x_k, y_k) + J(x_k) + (L_x - \tfrac{1}{2\gamma_{x,k}})\|\hat{x}_{k+1} - x_k\|^2 + (\tfrac{L_x}{2} - \tfrac{1}{2\gamma_{x,k}})\|x_{k+1} - x_k\|^2 \\ & \quad + 2\gamma_{x,k}\|\nabla_x F(x_k, y_k) - \widetilde{\nabla}_x(x_k, y_k)\|^2 + \tfrac{1}{8\gamma_{x,k}}\|\hat{x}_{k+1} - x_{k+1}\|^2 \\ & \overset{②}{\leq} F(x_k, y_k) + J(x_k) + (L_x - \tfrac{1}{4\gamma_{x,k}})\|\hat{x}_{k+1} - x_k\|^2 + (\tfrac{L_x}{2} - \tfrac{1}{4\gamma_{x,k}})\|x_{k+1} - x_k\|^2 \\ & \quad + 2\gamma_{x,k}\|\nabla_x F(x_k, y_k) - \widetilde{\nabla}_x(x_k, y_k)\|^2. \end{aligned} \tag{3.2}$$

Inequality ① is Young's, and ② is the standard inequality $\|a - c\|^2 \leq 2\|a - b\|^2 + 2\|b - c\|^2$. For the updates in $y_k$, we use Lemma 2.5 with $z = \hat{y}_{k+1}, y = x = y_k$, and $d = \nabla_y F(x_k, y_k)$, which gives

$$\begin{aligned} 0 \le\ & F(x_{k+1}, y_k) + R(y_k) - F(x_{k+1}, \hat{y}_{k+1}) - R(\hat{y}_{k+1}) \\ & + \langle \nabla_y F(x_{k+1}, y_k) - \nabla_y F(x_k, y_k), \hat{y}_{k+1} - y_k \rangle + (\tfrac{L_y}{2} - \tfrac{1}{\gamma_{y,k}})\|y_k - \hat{y}_{k+1}\|^2. \end{aligned} \tag{3.3}$$

Finally, we apply Lemma 2.5 with $z = y_{k+1}, y = \hat{y}_{k+1}, x = y_k$, and $d = \widetilde{\nabla}_y(x_{k+1}, y_k)$

$$\begin{aligned} 0 \le\ & F(x_{k+1}, \hat{y}_{k+1}) + R(\hat{y}_{k+1}) - F(x_{k+1}, y_{k+1}) - R(y_{k+1}) \\ & + \langle \nabla_y F(x_{k+1}, y_k) - \widetilde{\nabla}_y(x_{k+1}, y_k), y_{k+1} - \hat{y}_{k+1} \rangle + (\tfrac{L_y}{2} - \tfrac{1}{2\gamma_{y,k}})\|y_k - y_{k+1}\|^2 \\ & + (\tfrac{L_y}{2} + \tfrac{1}{2\gamma_{y,k}})\|y_k - \hat{y}_{k+1}\|^2. \end{aligned} \tag{3.4}$$

Adding these two inequalities and bounding the result as in (3.2), we obtain

$$\begin{aligned} & F(x_{k+1}, y_{k+1}) + R(y_{k+1}) \\ & \le F(x_{k+1}, y_k) + R(y_k) + (L_y - \tfrac{1}{2\gamma_{y,k}})\|\hat{y}_{k+1} - y_k\|^2 + (\tfrac{L_y}{2} - \tfrac{1}{2\gamma_{y,k}})\|y_{k+1} - y_k\|^2 \\ & \quad + \langle \nabla_y F(x_{k+1}, y_k) - \nabla_y F(x_k, y_k), \hat{y}_{k+1} - y_k \rangle + \langle \nabla_y F(x_{k+1}, y_k) - \widetilde{\nabla}_y(x_{k+1}, y_k), y_{k+1} - \hat{y}_{k+1} \rangle \\ & \overset{①}{\le} F(x_{k+1}, y_k) + R(y_k) + (L_y - \tfrac{1}{4\gamma_{y,k}})\|\hat{y}_{k+1} - y_k\|^2 + (\tfrac{L_y}{2} - \tfrac{1}{4\gamma_{y,k}})\|y_{k+1} - y_k\|^2 \\ & \quad + \langle \nabla_y F(x_{k+1}, y_k) - \nabla_y F(x_k, y_k), \hat{y}_{k+1} - y_k \rangle + 2\gamma_{y,k}\|\nabla_y F(x_{k+1}, y_k) - \widetilde{\nabla}_y(x_{k+1}, y_k)\|^2 \\ & \quad + \tfrac{1}{8\gamma_{y,k}}\|y_{k+1} - \hat{y}_{k+1}\|^2 \\ & \overset{②}{\le} F(x_{k+1}, y_k) + R(y_k) + (L_y - \tfrac{1}{4\gamma_{y,k}})\|\hat{y}_{k+1} - y_k\|^2 + (\tfrac{L_y}{2} - \tfrac{1}{4\gamma_{y,k}})\|y_{k+1} - y_k\|^2 \\ & \quad + \langle \nabla_y F(x_{k+1}, y_k) - \nabla_y F(x_k, y_k), \hat{y}_{k+1} - y_k \rangle + 2\gamma_{y,k}\|\nabla_y F(x_{k+1}, y_k) - \widetilde{\nabla}_y(x_{k+1}, y_k)\|^2 \\ & \overset{③}{\le} F(x_{k+1}, y_k) + R(y_k) + (L_y - \tfrac{1}{4\gamma_{y,k}})\|\hat{y}_{k+1} - y_k\|^2 + (\tfrac{L_y}{2} - \tfrac{1}{4\gamma_{y,k}})\|y_{k+1} - y_k\|^2 \\ & \quad + \tfrac{1}{2M}\|\nabla_y F(x_{k+1}, y_k) - \nabla_y F(x_k, y_k)\|^2 + \tfrac{M}{2}\|\hat{y}_{k+1} - y_k\|^2 \\ & \quad + 2\gamma_{y,k}\|\nabla_y F(x_{k+1}, y_k) - \widetilde{\nabla}_y(x_{k+1}, y_k)\|^2 \\ & \overset{④}{\le} F(x_{k+1}, y_k) + R(y_k) + (L_y + \tfrac{M}{2} - \tfrac{1}{4\gamma_{y,k}})\|\hat{y}_{k+1} - y_k\|^2 + (\tfrac{L_y}{2} - \tfrac{1}{4\gamma_{y,k}})\|y_{k+1} - y_k\|^2 \\ & \qquad + 2\gamma_{y,k}\|\nabla_y F(x_{k+1}, y_k) - \widetilde{\nabla}_x(x_{k+1}, y_k)\|^2 + \tfrac{M}{2}\|x_{k+1} - x_k\|^2. \end{aligned} \tag{3.5}$$

Inequalities ① and ③ are Young's, inequality ② follows from the fact that $\|a-c\|^2 \le 2\|a-b\|^2 + 2\|b-c\|^2$, and ④ uses the assumptions that the sequence $\{(x_k, y_k)\}_{k \in \mathbb{N}}$ is bounded and $\nabla F$ is $M$-Lipschitz continuous on this bounded set.

Adding inequality (3.2) and inequality (3.5), we have

$$\begin{aligned} \Phi(x_{k+1}, y_{k+1}) \le\ & \Phi(x_k, y_k) + (L_x - \tfrac{1}{4\gamma_{x,k}})\|\hat{x}_{k+1} - x_k\|^2 + (L_y + \tfrac{M}{2} - \tfrac{1}{4\gamma_{y,k}})\|\hat{y}_{k+1} - y_k\|^2 \\ & + (\tfrac{L_x}{2} + \tfrac{M}{2} - \tfrac{1}{4\gamma_{x,k}})\|x_{k+1} - x_k\|^2 + (\tfrac{L_y}{2} - \tfrac{1}{4\gamma_{y,k}})\|y_{k+1} - y_k\|^2 \\ & + 2\overline{\gamma}_k\big(\|\nabla_x F(x_k, y_k) - \widetilde{\nabla}_x(x_k, y_k)\|^2 + \|\nabla_y F(x_{k+1}, y_k) - \widetilde{\nabla}_y(x_{k+1}, y_k)\|^2\big), \end{aligned} \tag{3.6}$$

where $\overline{\gamma}_k = \max\{\gamma_{x,k}, \gamma_{y,k}\}$. We apply the conditional expectation operator $\mathbb{E}_k$ and bound the MSE terms using (2.1). This gives

$$
\begin{aligned}
&\mathbb{E}_k[\Phi(x_{k+1}, y_{k+1}) + (-\tfrac{L_x}{2} - \tfrac{M}{2} - 2V_1\overline{\gamma}_k + \tfrac{1}{4\gamma_{x,k}})\|x_{k+1} - x_k\|^2 \\
&\qquad + (-\tfrac{L_y}{2} - 2V_1\overline{\gamma}_k + \tfrac{1}{4\gamma_{y,k}})\|y_{k+1} - y_k\|^2] \\
&\leq \Phi(x_k, y_k) + (L_x - \tfrac{1}{4\gamma_{x,k}})\|\hat{x}_{k+1} - x_k\|^2 + (L_y + \tfrac{M}{2} - \tfrac{1}{4\gamma_{y,k}})\|\hat{y}_{k+1} - y_k\|^2 + 2\overline{\gamma}_k\Upsilon_k \\
&\qquad + 2V_1\overline{\gamma}_k\|z_k - z_{k-1}\|^2.
\end{aligned} \tag{3.7}
$$

Next, we use (2.2) to say

$$
2\overline{\gamma}_k\Upsilon_k \leq \tfrac{2\overline{\gamma}_k}{\rho}\big(-\mathbb{E}_k\Upsilon_{k+1} + \Upsilon_k + V_\Upsilon(\mathbb{E}_k\|z_{k+1} - z_k\|^2 + \|z_k - z_{k-1}\|^2)\big).
$$

Adding the previous two inequalities, we have

$$
\begin{aligned}
&\mathbb{E}_k[\Phi(x_{k+1}, y_{k+1}) + (-\tfrac{L_x}{2} - \tfrac{M}{2} - 2V_1\overline{\gamma}_k - \tfrac{2V_\Upsilon\overline{\gamma}_k}{\rho} + \tfrac{1}{4\gamma_{x,k}})\|x_{k+1} - x_k\|^2 \\
&\qquad + (-\tfrac{L_y}{2} - 2V_1\overline{\gamma}_k - \tfrac{2V_\Upsilon\overline{\gamma}_k}{\rho} + \tfrac{1}{4\gamma_{y,k}})\|y_{k+1} - y_k\|^2 + \tfrac{2\overline{\gamma}_k}{\rho}\Upsilon_{k+1}] \\
&\leq \Phi(x_k, y_k) + (L_x - \tfrac{1}{4\gamma_{x,k}})\|\hat{x}_{k+1} - x_k\|^2 + (L_y + \tfrac{M}{2} - \tfrac{1}{4\gamma_{y,k}})\|\hat{y}_{k+1} - y_k\|^2 + \tfrac{2\overline{\gamma}_k}{\rho}\Upsilon_k \\
&\qquad + 2\overline{\gamma}_k(V_1 + \tfrac{V_\Upsilon}{\rho})\|z_k - z_{k-1}\|^2.
\end{aligned}
$$

Let $\bar{L} = \max\{L_x, L_y\}$. To ensure that the coefficients of $\|x_{k+1} - x_k\|^2$ and $\|y_{k+1} - y_k\|^2$ are non-negative, we set

$$
\overline{\gamma}_k \leq \tfrac{1}{16}\sqrt{\tfrac{(\bar{L}+M)^2}{(V_1+V_\Upsilon/\rho)^2} + \tfrac{16}{(V_1+V_\Upsilon/\rho)}} - \tfrac{\bar{L}+M}{16(V_1+V_\Upsilon/\rho)}, \tag{3.8}
$$

for all $k \in \mathbb{N}$. With this choice,

$$
\begin{aligned}
&(-\tfrac{L_x+M}{2} - 2V_1\overline{\gamma}_k - \tfrac{2V_\Upsilon\overline{\gamma}_k}{\rho} + \tfrac{1}{4\gamma_{x,k}})\|x_{k+1} - x_k\|^2 + (-\tfrac{L_y}{2} - 2V_1\overline{\gamma}_k - \tfrac{2V_\Upsilon\overline{\gamma}_k}{\rho} \\
&\qquad + \tfrac{1}{4\gamma_{y,k}})\|y_{k+1} - y_k\|^2 \\
&\geq (-\tfrac{\bar{L}+M}{2} - 2V_1\overline{\gamma}_k - \tfrac{2V_\Upsilon\overline{\gamma}_k}{\rho} + \tfrac{1}{4\overline{\gamma}_k})\|z_{k+1} - z_k\|^2 \\
&\geq 2\overline{\gamma}_k(V_1 + V_\Upsilon/\rho)\|z_{k+1} - z_k\|^2.
\end{aligned} \tag{3.9}
$$

The final inequality is due to the bound in (3.8). To ensure that the coefficients of $\|\hat{x}_{k+1} - x_k\|^2$ and $\|\hat{y}_{k+1} - y_k\|^2$ are non-positive, we set $\gamma_{x,k} < \frac{1}{4L_x}$ and $\gamma_{y,k} < \frac{1}{4L_y+2M}$, which yields

$$
\begin{aligned}
&\mathbb{E}_k[\Phi(x_{k+1}, y_{k+1}) + 2\overline{\gamma}_k(V_1 + V_\Upsilon/\rho)\|z_{k+1} - z_k\|^2 + \tfrac{2\overline{\gamma}_k}{\rho}\Upsilon_{k+1}] \\
&\leq \Phi(x_k, y_k) + (L_x - \tfrac{1}{4\gamma_{x,k}})\|\hat{x}_{k+1} - x_k\|^2 + (L_y - \tfrac{1}{4\gamma_{y,k}})\|\hat{y}_{k+1} - y_k\|^2 \\
&\qquad + 2\overline{\gamma}_k(V_1 + V_\Upsilon/\rho)\|z_k - z_{k-1}\|^2 + \tfrac{2\overline{\gamma}_k}{\rho}\Upsilon_k.
\end{aligned}
$$

Because $\overline{\gamma}_k$ is non-increasing,

$$
\begin{aligned}
&\mathbb{E}_k[\Phi(x_{k+1}, y_{k+1}) + 2\overline{\gamma}_{k+1}(V_1 + V_\Upsilon/\rho)\|z_{k+1} - z_k\|^2 + \tfrac{2\overline{\gamma}_{k+1}}{\rho}\Upsilon_{k+1}] \\
&\leq \Phi(x_k, y_k) - \nu\|\hat{z}_{k+1} - z_k\|^2 + 2\overline{\gamma}_k(V_1 + V_\Upsilon/\rho)\|z_k - z_{k-1}\|^2 + \tfrac{2\overline{\gamma}_k}{\rho}\Upsilon_k,
\end{aligned}
$$

where $\nu = \min\{\frac{1}{4\gamma_{x,0}} - L_x, \frac{1}{4\gamma_{y,0}} - \frac{M}{2} - L_y\}$ Applying the full expectation operator, summing from $k = 0$ to $k = T-1$, and using the convention $z_{-1} = z_0$ gives

$$\tfrac{2\overline{\gamma}_T}{\rho}\Upsilon_T + 2\overline{\gamma}_T(V_1 + V_\Upsilon/\rho)\|z_T - z_{T-1}\|^2 + \nu\sum_{k=0}^{T-1}\mathbb{E}\|\hat{z}_{k+1} - z_k\|^2 \leq \Phi(x_0, y_0) + \tfrac{2\overline{\gamma}_0}{\rho}\Upsilon_0.$$

We drop the first two terms on the left from the inequality as they are non-negative. Let $\alpha$ be drawn uniformly at random from the set $\{0, 1, \cdots, T-1\}$, and recall $\underline{\gamma}_k \geq \beta$. Using the fact that $\|\hat{z}_{k+1} - z_k\|^2 \geq \frac{\beta^2}{4}\mathrm{dist}(0, \mathcal{G}_{\frac{\gamma_{x,k}}{2}, \frac{\gamma_{y,k}}{2}}(z_k))^2$,

$$\mathbb{E}\mathrm{dist}\Big(0, \mathcal{G}_{\frac{\gamma_{x,\alpha}}{2}, \frac{\gamma_{y,\alpha}}{2}}(z_\alpha)\Big)^2 \leq \frac{4(\Phi(x_0,y_0) + \frac{2\overline{\gamma}_0}{\rho}\Upsilon_0)}{T\nu\beta^2},$$

which completes the proof of the first claim. ∎

Combining the same argument with the error bound (3.1), we obtain a linear convergence rate to the global optimum.

*Proof of Theorem 3.1, Part 2.* We begin with equation (3.7):

$$\begin{aligned}
&\mathbb{E}_k[\Phi(x_{k+1}, y_{k+1}) + (-\tfrac{L_x}{2} - \tfrac{M}{2} - 2V_1\gamma_{x,k} + \tfrac{1}{4\gamma_{x,k}})\|x_{k+1} - x_k\|^2 \\
&\quad + (-\tfrac{L_y}{2} - 2V_1\gamma_{y,k} + \tfrac{1}{4\gamma_{y,k}})\|y_{k+1} - y_k\|^2] \\
&\leq \Phi(x_k, y_k) - \nu\|\hat{z}_{k+1} - z_k\|^2 + 2\overline{\gamma}_k\Upsilon_k + 2V_1\overline{\gamma}_k\|z_k - z_{k-1}\|^2.
\end{aligned}$$

Using (2.2), we can say for any $c > 0$,

$$0 \leq \tfrac{2c\overline{\gamma}_k}{\rho}\big(-\mathbb{E}_k\Upsilon_{k+1} + (1-\rho)\Upsilon_k + V_\Upsilon(\|z_{k+1} - z_k\|^2 + \|z_k - z_{k-1}\|^2)\big).$$

Adding the previous two inequalities, we have

$$\begin{aligned}
&\mathbb{E}_k[\Phi(x_{k+1}, y_{k+1}) + (-\tfrac{L_x}{2} - \tfrac{M}{2} - 2V_1\gamma_{x,k} - \tfrac{2cV_\Upsilon\overline{\gamma}_k}{\rho} + \tfrac{1}{4\gamma_{x,k}})\|x_{k+1} - x_k\|^2 \\
&\quad + (-\tfrac{L_y}{2} - 2V_1\gamma_{y,k} - \tfrac{2cV_\Upsilon\overline{\gamma}_k}{\rho} + \tfrac{1}{4\gamma_{y,k}})\|y_{k+1} - y_k\|^2 + \tfrac{2c\overline{\gamma}_k}{\rho}\Upsilon_{k+1}] \\
&\leq \Phi(x_k, y_k) - \nu\|\hat{z}_{k+1} - z_k\|^2 + 2\overline{\gamma}_k(V_1 + \tfrac{cV_\Upsilon}{\rho})\|z_k - z_{k-1}\|^2 + \tfrac{2c\overline{\gamma}_k}{\rho}(1 + \tfrac{\rho}{c} - \rho)\Upsilon_k.
\end{aligned}$$

We apply the error bound assumption (3.1) to say

$$-\nu\|\hat{z}_{k+1} - z_k\|^2 \leq -\tfrac{\nu\underline{\gamma}_k^2}{4}\mathrm{dist}(0, \mathcal{G}_{\frac{\gamma_{x,k}}{2}, \frac{\gamma_{y,k}}{2}}(z_k))^2 \leq -\tfrac{\nu\underline{\gamma}_k^2}{4\mu}(\Phi(x_k, y_k) - \underline{\Phi}).$$

In total, we have

$$\begin{aligned}
&\mathbb{E}_k[\Phi(x_{k+1}, y_{k+1}) - \underline{\Phi} + (-\tfrac{L_x}{2} - \tfrac{M}{2} - 2V_1\gamma_{x,k} - \tfrac{2cV_\Upsilon\overline{\gamma}_k}{\rho} + \tfrac{1}{4\gamma_{x,k}})\|x_{k+1} - x_k\|^2 \\
&\quad + (-\tfrac{L_y}{2} - 2V_1\gamma_{y,k} - \tfrac{2cV_\Upsilon\overline{\gamma}_k}{\rho} + \tfrac{1}{4\gamma_{y,k}})\|y_{k+1} - y_k\|^2 + \tfrac{2c\overline{\gamma}_k}{\rho}\Upsilon_{k+1}] \\
&\leq (1 - \tfrac{\nu\underline{\gamma}_k^2}{4\mu})(\Phi(x_k, y_k) - \underline{\Phi}) + 2\overline{\gamma}_k(V_1 + \tfrac{cV_\Upsilon}{\rho})\|z_k - z_{k-1}\|^2 + \tfrac{2c\overline{\gamma}_k}{\rho}(1 + \tfrac{\rho}{c} - \rho)\Upsilon_k.
\end{aligned}$$

Choosing $c = 2$, setting the step-sizes so that they satisfy, for all $k$,

$$\overline{\gamma}_k \leq \frac{1}{20}\sqrt{\frac{(\bar{L}+M)^2}{(V_1+2V_\Upsilon/\rho)^2} + \frac{20}{(V_1+2V_\Upsilon/\rho)}} - \frac{\bar{L}+M}{20(V_1+2V_\Upsilon/\rho)},\ \gamma_{x,k} < \frac{1}{4L_x},\ \gamma_{y,k} < \frac{1}{4L_y+2M},\ 0 < \beta \leq \underline{\gamma}_k,$$

and letting $\Theta = \min\{\frac{\nu\beta^2}{4\mu}, \frac{\rho}{2}\}$, we have

$$\begin{aligned}&\mathbb{E}_k[\Phi(x_{k+1}, y_{k+1}) - \underline{\Phi} + 2\overline{\gamma}_k(V_1 + \tfrac{2V_\Upsilon}{\rho})\|z_{k+1} - z_k\|^2 + \tfrac{4\overline{\gamma}_k}{\rho}\Upsilon_{k+1}]\\ &\leq (1-\Theta)(\Phi(x_k, y_k) - \underline{\Phi} + 2\overline{\gamma}_k(V_1 + \tfrac{2V_\Upsilon}{\rho})\|z_k - z_{k-1}\|^2 + \tfrac{4\overline{\gamma}_k}{\rho}\Upsilon_k).\end{aligned}$$

Because $\overline{\gamma}_k$ is non-increasing,

$$\begin{aligned}&\mathbb{E}_k[\Phi(x_{k+1}, y_{k+1}) - \underline{\Phi} + 2\overline{\gamma}_{k+1}(V_1 + \tfrac{2V_\Upsilon}{\rho})\|z_{k+1} - z_k\|^2 + \tfrac{4\overline{\gamma}_{k+1}}{\rho}\Upsilon_{k+1}]\\ &\leq (1-\Theta)(\Phi(x_k, y_k) - \underline{\Phi} + 2\overline{\gamma}_k(V_1 + \tfrac{2V_\Upsilon}{\rho})\|z_k - z_{k-1}\|^2 + \tfrac{4\overline{\gamma}_k}{\rho}\Upsilon_k).\end{aligned}$$

Applying the full expectation operator, chaining this inequality over the iterations $k = 0$ to $k = T - 1$, and using the convention $z_{-1} = z_0$,

$$\mathbb{E}[\Phi(x_T, y_T) - \underline{\Phi}] \leq (1-\Theta)^T \big(\Phi(x_0, y_0) - \underline{\Phi} + \tfrac{4\overline{\gamma}_0}{\rho}\Upsilon_0\big),$$

which completes the proof. ■

Because SAGA and SARAH gradient estimators are variance-reduced, Theorem 3.1 implies specific convergence rates for Algorithm 1.1 when using these estimators.

Corollary 3.3. *To compute an $\epsilon$-approximate critical point in expectation, Algorithm 1.1 using*

- *SARAH gradient estimator with $p = n$, $\overline{\gamma}_k \leq \frac{1}{2L\sqrt{30n}}$ and any batch size requires no more than $\mathcal{O}\big(L\sqrt{n}/\epsilon^2\big)$ SFO calls;*
- *SAGA gradient estimator with $b = n^{2/3}$ and $\overline{\gamma}_k \leq \frac{1}{2\sqrt{2710}L}$ requires no more than $\mathcal{O}(Ln^{2/3}/\epsilon^2)$ SFO calls.*[5]

*If $\Phi$ satisfies the error bound condition (3.1), then to compute an $\epsilon$-suboptimal point in expectation, Algorithm 1.1 using*

- *the SARAH gradient estimator requires no more than $\mathcal{O}((n + L\sqrt{n}/\mu)\log\big(1/\epsilon\big))$ SFO calls;*
- *the SAGA gradient estimator requires no more than $\mathcal{O}((n + Ln^{2/3}/\mu)\log(1/\epsilon))$ SFO calls.*

*Remark* 3.4. The improved dependence on $n$ when using SARAH gradient estimator exists in all of our convergence rates for SPRING. Because most existing works on stochastic optimization for non-smooth, non-convex problems use models that are special cases of (1.1), our results for SPRING capture most existing work as special cases. In particular, in the case $R \equiv J \equiv 0$, our results recover recent results showing that SARAH achieves the *oracle complexity lower-bound* for non-convex problems with a finite-sum structure [18, 28, 37, 40, 41].

[5] For ease of exposition, we do not optimize over constants, so these step-sizes (particularly for the SAGA estimator) are not optimal. In general, we find the step-sizes suggested by theory to be conservative in practice (see Section 5 for details regarding practical step-sizes).

**4. Convergence Rate under the KL Property.** The results from previous section require only assumptions (**A.1**) to (**A.4**). To prove convergence of the sequence of the algorithm, and to obtain convergence rates depending on the KL exponent of the objective, we further require that $\Phi$ is semi-algebraic. In this section, under these assumptions, we prove convergence of the sequence and extend the convergence rates of PALM to SPRING. To derive these results, we first derive some preparatory results which generalize claims of PALM [6] to the stochastic setting. Given $k \in \mathbb{N}$, define the quantity

$$\Psi_k \overset{\text{def}}{=} \Phi(z_k) + \frac{1}{2\rho\sqrt{2(V_1+V_\Upsilon/\rho)}}\Upsilon_k + \frac{\sqrt{V_1+V_\Upsilon/\rho}}{\sqrt{2}}\|z_k - z_{k-1}\|^2. \tag{4.1}$$

Our first result guarantees that $\Psi_k$ is decreasing in expectation.

Lemma 4.1 ($\ell_2$ summability). *Let $\{z_k\}_{k=0}^{\infty}$ be the sequence generated by SPRING with $\overline{\gamma}_k$ non-increasing and satisfying $\overline{\gamma}_k < \frac{\sqrt{2}}{5(\sqrt{V_1+V_\Upsilon/\rho}+\bar{L})}$, $\forall k$, then $\Psi_k$ satisfies*

$$\mathbb{E}_k\Psi_{k+1} \le \Psi_k + \left(\tfrac{\bar{L}}{2} + \tfrac{3}{2}\sqrt{2(V_1 + V_\Upsilon/\rho)} - \tfrac{1}{2\overline{\gamma}_k}\right)\mathbb{E}_k\|z_{k+1} - z_k\|^2 - \frac{\sqrt{V_1+V_\Upsilon/\rho}}{2\sqrt{2}}\|z_k - z_{k-1}\|^2, \tag{4.2}$$

*and the expectation of the squared distance between the iterates is summable:*

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

*Proof.* Applying Lemma 2.6 twice, once for the update in $x_k$ and once for the update in $y_k$, we have

$$\begin{aligned} F(x_{k+1}, y_k) + J(x_{k+1}) \le F(x_k, y_k) + J(x_k) + \tfrac{1}{2\bar{L}\lambda}\|\widetilde{\nabla}_x(x_k, y_k) - \nabla_x F(x_k, y_k)\|^2 \\ + \left(\tfrac{\bar{L}(\lambda+1)}{2} - \tfrac{1}{2\gamma_{x,k}}\right)\|x_{k+1} - x_k\|^2, \end{aligned}$$

as well as

$$\begin{aligned} F(x_{k+1}, y_{k+1}) + R(y_{k+1}) \le F(x_{k+1}, y_k) + R(y_k) + \left(\tfrac{\bar{L}(\lambda+1)}{2} - \tfrac{1}{2\gamma_{y,k}}\right)\|y_{k+1} - y_k\|^2 \\ + \tfrac{1}{2\bar{L}\lambda}\|\widetilde{\nabla}_y(x_{k+1}, y_k) - \nabla_y F(x_{k+1}, y_k)\|^2. \end{aligned}$$

Adding these inequalities together,

$$\begin{aligned} \Phi(x_{k+1}, y_{k+1}) \le \Phi(x_k, y_k) + \tfrac{1}{2\bar{L}\lambda}\|\widetilde{\nabla}_x(x_k, y_k) - \nabla_x F(x_k, y_k)\|^2 \\ + \tfrac{1}{2\bar{L}\lambda}\|\widetilde{\nabla}_y(x_{k+1}, y_k) - \nabla_y F(x_{k+1}, y_k)\|^2 + \left(\tfrac{\bar{L}(\lambda+1)}{2} - \tfrac{1}{2\overline{\gamma}_k}\right)\|z_{k+1} - z_k\|^2. \end{aligned}$$

Applying the conditional expectation operator $\mathbb{E}_k$, we can bound the MSE terms using (2.1). This gives

$$\mathbb{E}_k\left[\Phi(z_{k+1}) + \left(-\tfrac{\bar{L}(\lambda+1)}{2} - \tfrac{V_1}{2\bar{L}\lambda} + \tfrac{1}{2\overline{\gamma}_k}\right)\|z_{k+1} - z_k\|^2\right] \le \Phi(z_k) + \tfrac{1}{2\bar{L}\lambda}\Upsilon_k + \tfrac{V_1}{2\bar{L}\lambda}\|z_k - z_{k-1}\|^2. \tag{4.3}$$

Next, we use (2.2) to say that

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

Combining these inequalities, we have

$$\begin{aligned}&\mathbb{E}_k\big[\Phi(z_{k+1})+\tfrac{1}{2\bar{L}\lambda\rho}\Upsilon_{k+1}+\big(-\tfrac{\bar{L}(\lambda+1)}{2}-\tfrac{V_1+V_\Upsilon/\rho}{2\bar{L}\lambda}+\tfrac{1}{2\overline{\gamma}_k}\big)\|z_{k+1}-z_k\|^2\big]\\&\le\Phi(z_k)+\tfrac{1}{2\bar{L}\lambda\rho}\Upsilon_k+\tfrac{V_1+V_\Upsilon/\rho}{2\bar{L}\lambda}\|z_k-z_{k-1}\|^2.\end{aligned}$$

This is equivalent to

$$\begin{aligned}&\mathbb{E}_k\big[\Phi(z_{k+1})+\tfrac{1}{2\bar{L}\lambda\rho}\Upsilon_{k+1}+\big(\tfrac{V_1+V_\Upsilon/\rho}{2\bar{L}\lambda}+Z\big)\|z_{k+1}-z_k\|^2\\&\qquad+\big(-\tfrac{\bar{L}(\lambda+1)}{2}-\tfrac{V_1+V_\Upsilon/\rho}{\bar{L}\lambda}-Z+\tfrac{1}{2\overline{\gamma}_k}\big)\|z_{k+1}-z_k\|^2\big]\\&\le\Phi(z_k)+\tfrac{1}{2\bar{L}\lambda\rho}\Upsilon_k+\big(\tfrac{V_1+V_\Upsilon/\rho}{2\bar{L}\lambda}+Z\big)\|z_k-z_{k-1}\|^2-Z\|z_k-z_{k-1}\|^2,\end{aligned}$$

for any constant $Z\ge 0$. We use the choice $Z=\frac{\sqrt{V_1+V_\Upsilon/\rho}}{2\sqrt{2}}$ to simplify later arguments. Setting $\overline{\gamma}_k\le(2(\frac{\bar{L}(\lambda+1)}{2}+\frac{V_1+V_\Upsilon/\rho}{\bar{L}\lambda}+Z))^{-1}$, setting $\lambda=\frac{\sqrt{2(V_1+V_\Upsilon/\rho)}}{\bar{L}}$ to approximately maximize this bound on $\overline{\gamma}_k$, and using the fact that $\overline{\gamma}_k$ is non-increasing, we have

$$\mathbb{E}_k\Psi_{k+1}\le\Psi_k+\big(\tfrac{\bar{L}(\lambda+1)}{2}+\tfrac{V_1+V_\Upsilon/\rho}{\bar{L}\lambda}+Z-\tfrac{1}{2\overline{\gamma}_k}\big)\mathbb{E}_k\|z_{k+1}-z_k\|^2-Z\|z_k-z_{k-1}\|^2, \tag{4.4}$$

proving the first claim that $\Psi_k$ is decreasing in expectation.

To prove the second claim, we apply the full expectation operator to (4.4) and sum the resulting inequality from $k=0$ to $k=T-1$,

$$\mathbb{E}\Psi_T\le\Psi_0+\sum_{k=0}^{T-1}\big(\tfrac{\bar{L}(\lambda+1)}{2}+\tfrac{V_1+V_\Upsilon/\rho}{\bar{L}\lambda}+Z-\tfrac{1}{2\overline{\gamma}_k}\big)\mathbb{E}\|z_{k+1}-z_k\|^2-Z\mathbb{E}\|z_k-z_{k-1}\|^2.$$

Rearranging and using the facts that $\underline{\Phi}\le\Psi_T$ and $\overline{\gamma}_k$ is non-increasing,

$$\sum_{k=0}^{T-1}\big(\tfrac{1}{2\overline{\gamma}_k}-\tfrac{\bar{L}(\lambda+1)}{2}-\tfrac{V_1+V_\Upsilon/\rho}{\bar{L}\lambda}-Z\big)\mathbb{E}\|z_{k+1}-z_k\|^2+Z\mathbb{E}\|z_k-z_{k-1}\|^2\le\Psi_0-\underline{\Phi}. \tag{4.5}$$

Taking the limit $T\to+\infty$ proves that the sequence $\mathbb{E}\|z_{k+1}-z_k\|^2$ is summable. ∎

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

Lemma 4.2 (Subgradient Bound). *Let $\{z_k\}_{k\in\mathbb{N}}$ be the sequence generated by SPRING with step-sizes satisfying $0<\beta\le\underline{\gamma}_k$. Define*

$$\begin{aligned}A_x^k&\stackrel{\text{def}}{=}1/\gamma_{x,k}(x_{k-1}-x_k)+\nabla_x F(x_k,y_k)-\widetilde{\nabla}_x(x_{k-1},y_{k-1})\quad\textit{and}\\A_y^k&\stackrel{\text{def}}{=}1/\gamma_{y,k}(y_{k-1}-y_k)+\nabla_y F(x_k,y_k)-\widetilde{\nabla}_y(x_k,y_{k-1}).\end{aligned}$$

*Then $(A_x^k,A_y^k)\in\partial\Phi(x_k,y_k)$ and, with $p=1/\beta+M+L_y+V_2$,*

$$\mathbb{E}_{k-1}\|(A_x^k,A_y^k)\|\le p(\mathbb{E}_{k-1}\|z_k-z_{k-1}\|+\|z_{k-1}-z_{k-2}\|)+\Gamma_{k-1}. \tag{4.6}$$

*Proof.* The fact that $(A_x^k, A_y^k) \in \partial\Phi(x_k, y_k)$ is clear from the definition of the proximal operator:

$$\begin{aligned} \tfrac{1}{\gamma_{x,k}}(x_{k-1} - x_k) - \widetilde{\nabla}_x(x_{k-1}, y_{k-1}) &\in \partial J(x_k), \\ \tfrac{1}{\gamma_{y,k}}(y_{k-1} - y_k) - \widetilde{\nabla}_y(x_k, y_{k-1}) &\in \partial R(y_k). \end{aligned}$$

Combining this with the fact that $\partial\Phi(x_k, y_k) = (\nabla_x F(x_k, y_k) + \partial J(x_k), \nabla_y F(x_k, y_k) + \partial R(y_k))$ makes it clear that $(A_x^k, A_y^k) \in \partial\Phi(x_k, y_k)$. All that remains is to bound the norms of $A_x^k$ and $A_y^k$. Because $\nabla F$ is $M$-Lipschitz continuous on bounded sets,

$$\begin{aligned} \mathbb{E}_{k-1}\|A_x^k\| &\le \tfrac{1}{\gamma_{x,k}}\mathbb{E}_{k-1}\|x_{k-1} - x_k\| + \mathbb{E}_{k-1}\|\nabla_x F(x_k, y_k) - \widetilde{\nabla}_x(x_{k-1}, y_{k-1})\| \\ &\le \tfrac{1}{\gamma_{x,k}}\mathbb{E}_{k-1}\|x_{k-1} - x_k\| + \mathbb{E}_{k-1}\|\nabla_x F(x_k, y_k) - \nabla_x F(x_{k-1}, y_{k-1})\| \\ &\quad + \mathbb{E}_{k-1}\|\nabla_x F(x_{k-1}, y_{k-1}) - \widetilde{\nabla}_x(x_{k-1}, y_{k-1})\| \\ &\le \left(\tfrac{1}{\gamma_{x,k}} + M\right)\mathbb{E}_{k-1}\|x_{k-1} - x_k\| + M\mathbb{E}_{k-1}\|y_k - y_{k-1}\| \\ &\quad + \mathbb{E}_{k-1}\|\nabla_x F(x_{k-1}, y_{k-1}) - \widetilde{\nabla}_x(x_{k-1}, y_{k-1})\|. \end{aligned} \tag{4.7}$$

A similar argument holds for $\|A_y^k\|$.

$$\begin{aligned} \mathbb{E}_{k-1}\|A_y^k\| &\le \tfrac{1}{\gamma_{y,k}}\mathbb{E}_{k-1}\|y_{k-1} - y_k\| + \mathbb{E}_{k-1}\|\nabla_y F(x_k, y_k) - \widetilde{\nabla}_y(x_k, y_{k-1})\| \\ &\le \tfrac{1}{\gamma_{y,k}}\mathbb{E}_{k-1}\|y_{k-1} - y_k\| + \mathbb{E}_{k-1}\|\nabla_y F(x_k, y_k) - \nabla_y F(x_k, y_{k-1})\| \\ &\quad + \mathbb{E}_{k-1}\|\nabla_y F(x_k, y_{k-1}) - \widetilde{\nabla}_y(x_k, y_{k-1})\| \\ &\le \left(\tfrac{1}{\gamma_{y,k}} + L_y\right)\mathbb{E}_{k-1}\|y_{k-1} - y_k\| + \mathbb{E}_{k-1}\|\nabla_y F(x_k, y_{k-1}) - \widetilde{\nabla}_y(x_k, y_{k-1})\|. \end{aligned}$$

Adding these two inequalities together and using equation (2.1) to bound the MSE terms, we get

$$\mathbb{E}_{k-1}\|(A_x^k, A_y^k)\| \le \mathbb{E}_{k-1}\left[\|A_x^k\| + \|A_y^k\|\right] \le p(\mathbb{E}_{k-1}\|z_k - z_{k-1}\| + \|z_{k-1} - z_{k-2}\|) + \Gamma_{k-1},$$

where $p = 1/\beta + M + L_y + V_2$. ∎

Define the set of limit points of $\{z_k\}_{k=0}^\infty$ as

$$\omega \stackrel{\text{def}}{=} \{z : \exists \text{ an increasing sequence of integers } \{k_\ell\}_{\ell\in\mathbb{N}} \text{ such that } z_{k_\ell} \to z \text{ as } \ell \to +\infty\}.$$

The following lemma describes properties of $\omega$.

**Lemma 4.3** (Limit points of $\{z_k\}_{k=0}^\infty$). *Suppose assumptions* (**A.1**)-(**A.4**) *hold, that the sequence* $z_k = (x_k, y_k)$ *is* bounded*, and the step-sizes of Algorithm 1.1 satisfy the following conditions:*

$$\gamma_{x,k}, \gamma_{y,k} \in \left[\beta, \tfrac{\sqrt{2}}{5(\sqrt{V_1 + V_\Upsilon/\rho} + \bar{L})}\right) \qquad \forall k,$$

*and* $\overline{\gamma}_k$ *is non-increasing. 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). $\mathbb{E}\Phi(z_k) \to \Phi^\star$*, where* $\Phi^\star \in [\underline{\Phi}, \infty)$*;*
(3). $\mathbb{E}\mathrm{dist}(0, \partial\Phi(z_k)) \to 0$*;*

(4). *The set* $\omega$ *is non-empty, and for all* $z^\star \in \omega$, $\mathbb{E}\mathrm{dist}(0, \partial\Phi(z^\star)) = 0$*;*
(5). $\mathrm{dist}(z_k, \omega) \to 0$ *a.s.;*
(6). $\omega$ *is a.s. compact and connected;*
(7). $\mathbb{E}\Phi(z^\star) = \Phi^\star$ *for all* $z^\star \in \omega$.

*Remark* 4.4. The boundedness of $z_k$ is also imposed in the original PALM [6] and asynchronous PALM [14], it can be satisfied automatically if, for instance, each regularizer has bounded domain.

*Proof.* By Lemma 4.1, we have

$$\mathbb{E}_k\Psi_{k+1} + \mathcal{O}\big(\|z_k - z_{k-1}\|^2\big) \le \Psi_k.$$

The supermartingale convergence theorem implies that $\sum_{k=1}^{\infty} \|z_k - z_{k-1}\|^2 < +\infty$ a.s., and it follows that $\|z_k - z_{k-1}\| \to 0$ a.s. This proves Claim 1.

The supermartingale convergence theorem also ensures $\Psi_k$ converges a.s. to a finite, positive random variable. Because $\|z_k - z_{k-1}\| \to 0$ a.s. and $\widetilde{\nabla}$ is variance-reduced so $\mathbb{E}\Upsilon_k \to 0$, we can say $\lim_{k\to\infty} \mathbb{E}\Psi_k = \lim_{k\to\infty} \mathbb{E}\Phi(z_k) \in [\underline{\Phi}, \infty)$, implying Claim 2.

Claim 3 holds because, by Lemma 4.2,

$$\mathbb{E}\|(A_x^k, A_y^k)\| \le p\mathbb{E}[\|z_k - z_{k-1}\| + \|z_{k-1} - z_{k-2}\|] + \mathbb{E}\Gamma_{k-1}.$$

We have that $\mathbb{E}\|z_k - z_{k-1}\| \to 0$ and $\mathbb{E}\Gamma_k \to 0$. This ensures that $\mathbb{E}\|(A_x^k, A_y^k)\| \to 0$.

To prove Claim 4, suppose $z^\star = (x^\star, y^\star)$ is a limit point of the sequence $\{z_k\}_{k=0}^{\infty}$ (a limit point must exist because we suppose the sequence $\{z_k\}_{k=0}^{\infty}$ is bounded). This means there exists a subsequence $z_{k_q}$ satisfying $\lim_{q\to\infty} z_{k_q} \to z^\star$. Furthermore, by the variance-reduced property of $\widetilde{\nabla}_x(x_{k_q}, y_{k_q}$, we have $\mathbb{E}\|\widetilde{\nabla}_x(x_{k_q}, y_{k_q}) - \nabla_x F(x_{k_q}, y_{k_q})\|^2 \to 0$, which implies that there exists a subsequence of $\{z_{k_q}\}_{q\in\mathbb{N}}$ (call it $\{z_{k_q}\}_{q\in\mathcal{I}}$ for some index set $\mathcal{I} \subset \mathbb{N}$) such that $\widetilde{\nabla}_x(x_{k_q}, y_{k_q}) - \nabla_x F(x_{k_q}, y_{k_q}) \to 0$ a.s. Because $R$ and $J$ are lower semicontinuous,

$$\liminf_{q\to\infty} R(x_{k_q}) \ge R(x^\star) \qquad \text{and} \qquad \liminf_{q\to\infty} J(x_{k_q}) \ge J(x^\star). \tag{4.8}$$

By the update rule for $x_{k+1}$,

$$x_{k+1} \in \mathrm{argmin}_x\big\{\langle x - x_k, \widetilde{\nabla}_x(x_k, y_k)\rangle + \tfrac{1}{2\gamma_{x,k}}\|x - x_k\|^2 + R(x)\big\}.$$

Letting $x = x^\star$,

$$\begin{aligned}&\langle x_{k+1} - x_k, \widetilde{\nabla}_x(x_k, y_k)\rangle + \tfrac{1}{2\gamma_{x,k}}\|x_{k+1} - x_k\|^2 + R(x_{k+1})\\ &\le \langle x^\star - x_k, \nabla_x F(x_k, y_k)\rangle + \langle x^\star - x_k, \widetilde{\nabla}_x(x_k, y_k) - \nabla_x F(x_k, y_k)\rangle + \tfrac{1}{2\gamma_{x,k}}\|x^\star - x_k\|^2 + R(x^\star).\end{aligned}$$

Setting $k = k_q$, taking the expectation, and taking the limit $q \to \infty$,

$$\begin{aligned}\limsup_{q\to\infty} R(x_{k_q+1}) \le{}& \limsup_{q\to\infty} \langle x^\star - x_{k_q}, \nabla_x F(x_{k_q}, y_{k_q})\rangle\\ &+ \langle x^\star - x_{k_q}, \widetilde{\nabla}_x(x_{k_q}, y_{k_q}) - \nabla_x F(x_{k_q}, y_{k_q})\rangle + \tfrac{1}{2\gamma_{x,k}}\|x^\star - x_{k_q}\|^2 + R(x^\star).\end{aligned}$$

The first term on the right goes to zero because $x_{k_q} \to x^\star$ and $\nabla_x F(x_{k_q}, y_{k_q})$ is bounded. The second term is zero almost surely because it is bounded above by $\|x_{k_q} - x^*\|^2 + \|\widetilde{\nabla}_x(x_{k_q}, y_{k_q}) - \nabla_x F(x_{k_q}, y_{k_q})\|^2$, and we have $\widetilde{\nabla}_x(x_{k_q}, y_{k_q}) - \nabla_x F(x_{k_q}, y_{k_q}) \to 0$ a.s. Therefore, $\limsup_{q\to\infty} R(x_{k_q+1}) \leq R(x^\star)$ a.s., which, together with equation (4.8), implies $R(x_{k_q+1}) \to R(x^\star)$ a.s. The same argument holds for $J$ and $y_k$, and it follows that

$$\lim_{q\to\infty} \Phi(x_{k_q}, y_{k_q}) = \Phi(x^\star, y^\star) \qquad \text{a.s.} \tag{4.9}$$

Claim 3 ensures that $\mathbb{E}\|(A_x^k, A_y^k)\| \to 0$. Combining Claim 3 with (4.9) and the fact that the subdifferential of $\Phi$ is closed, we have $\mathbb{E}\text{dist}(0, \partial\Phi(z^\star)) = 0$.

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 [6, Remark 5] and [14, Remark 4.1]).

Finally, we must show that $\Phi$ has constant expectation over $\omega$. From Claim 2, we have $\mathbb{E}\Phi(z_k) \to \Phi^\star$ which implies $\mathbb{E}\Phi(z_{k_q}) \to \Phi^\star$ for every subsequence $\{z_{k_q}\}_{q=0}^\infty$ converging to some $z^\star \in \omega$. In the proof of Claim 4, we show that $\Phi(z_{k_q}) \to \Phi(z^\star)$, so $\mathbb{E}\Phi(z^\star) = \Phi^\star$ for all $z^\star \in \omega$. ∎

The following lemma is analogous to the Uniformized Kurdyka–Łojasiewicz Property [6]. It is a slight generalization of the Kurdyka–Łojasiewicz property showing that $z_k$ eventually enters a region of $\overline{z}$ for some $\overline{z}$ satisfying $\Phi(\overline{z}) = \Phi(z^*)$, and in this region, the Kurdyka–Łojasiewicz inequality holds.

Lemma 4.5. *Assume the conditions of Lemma 4.3 hold and that $z_k$ is not a critical point of $\Phi$ after a finite number of iterations. Let $\Phi$ be a semi-algebraic function with KL exponent $\theta$. Then there exists an index $m$ and a desingularizing function $\phi$ so that the following bound holds:*

$$\phi'(\mathbb{E}[\Phi(z_k) - \Phi_k^\star])\mathbb{E}\text{dist}\big(0, \partial\Phi(z_k)\big) \geq 1 \qquad \forall k > m,$$

*where $\Phi_k^\star$ is a non-decreasing sequence converging to $\mathbb{E}\Phi(z^\star)$ for some $z^\star \in \omega$.*

*Proof.* First, we show that $\mathbb{E}\Phi(z_k)$ satisfies the KL property. Recall that $b$ is the mini-batch size. Let $\overline{n} = \binom{n}{b}$ be the number of possible gradient estimates in one iteration, and let $\{z_k^i\}_{i=1}^{\overline{n}^k}$ be the set of possible values for $z_k$. Considering $\mathbb{E}\Phi$ as a function of $\{z_k^i\}_{i=1}^{\overline{n}^k}$, we have

$$\mathbb{E}\Phi(z_k) = \tfrac{1}{\overline{n}^k} \sum\nolimits_{i=1}^{\overline{n}^k} \Phi(z_k^i).$$

Because $\mathbb{E}\Phi(z_k)$ can be written as $\sum_i f_i(x_i)$ where $f_i$ are KL functions with exponent $\theta$, $\mathbb{E}\Phi(z_k)$ (as a function of $\{z_k^i\}_{i=1}^{\overline{n}^k}$) is also KL with exponent $\theta$ [25, Theorem 3.3]. Hence, $\mathbb{E}\Phi$ satisfies the KL inequality at every point in its domain. Therefore, for every point $(z_k^1, \cdots, z_k^{\overline{n}^k})$ in a neighborhood $U_k$ of $(\overline{z}_k^1, \overline{z}_k^2, \cdots, \overline{z}_k^{\overline{n}^k})$ and satisfying

$$\tfrac{1}{\overline{n}^k} \sum\nolimits_{i=1}^{\overline{n}^k} \Phi(\overline{z}_k^i) < \tfrac{1}{\overline{n}^k} \sum\nolimits_{i=1}^{\overline{n}^k} \Phi(z_k^i) < \tfrac{1}{\overline{n}^k} \sum\nolimits_{i=1}^{\overline{n}^k} \Phi(\overline{z}_k^i) + \epsilon_k \tag{4.10}$$

for some $\epsilon_k > 0$, the Kurdyka–Łojasiewicz inequality holds with the desingularizing function $\phi_k$:

$$\phi_k'\Big(\tfrac{1}{\overline{n}^k} \sum\nolimits_{i=1}^{\overline{n}^k} \Phi(z_k^i) - \tfrac{1}{\overline{n}^k} \sum\nolimits_{i=1}^{\overline{n}^k} \Phi(\overline{z}_k^i)\Big)\text{dist}\Big(0, \tfrac{1}{\overline{n}^k} \sum\nolimits_{i=1}^{\overline{n}^k} \partial\Phi(z_k^i)\Big) \geq 1.\text{[6]} \tag{4.11}$$

[6]For the subdifferential terms we are taking the Minkowski sum: $\frac{1}{\overline{n}^k} \sum_{i=1}^{\overline{n}^k} \partial\Phi(z_k^i) = \{\frac{1}{\overline{n}^k} \sum_{i=1}^{\overline{n}^k} \xi_i | \xi_i \in \partial\Phi(z_k^i)\}$.

There always exists a choice of $(\overline{z}_k^1, \overline{z}_k^2, \cdots, \overline{z}_k^{\overline{n}^k})$ satisfying (4.10) unless $\mathbb{E}\Phi(z_k)$ is a local minimum. Lemma 4.3 Claim 5 implies $\operatorname{dist}(z_k, \omega) \to 0$ a.s., and Claims 2 and 7 imply $\mathbb{E}\Phi(z_k) \to \mathbb{E}\Phi(z^\star)$, so we can choose $\overline{z}_k$ such that $\frac{1}{\overline{n}^k}\sum_{i=1}^{\overline{n}^k}\Phi(\overline{z}_k^i) \to \mathbb{E}\Phi(z^\star)$ as well. To summarize, we have shown that there exists a sequence $(\overline{z}_k^1, \cdots, \overline{z}_k^{\overline{n}^k})$ such that

1. The point $(z_k^1, \cdots, z_k^{\overline{n}^k})$ lies in a neighborhood $U_k$ of $(\overline{z}_k^1, \cdots, \overline{z}_k^{\overline{n}^k})$,
2. The inequality (4.10) is satisfied, and
3. We have $\frac{1}{\overline{n}^k}\sum_{i=1}^{\overline{n}^k}\Phi(\overline{z}_k^i) \to \mathbb{E}\Phi(z^\star)$.

Points 1.) and 2.) imply the Kurdyka–Łojasiewicz inequality (4.11). This ensures that the Kurdyka–Łojasiewicz inequality holds at every iteration, but the desingularizing function $\phi_k$ changes every iteration. We now show that the Kurdyka–Łojasiewicz inequality holds using a single function $\phi$.

Because $\Phi$ is semi-algebraic with KL exponent $\theta$, each desingularizing function is of the form $\phi_k(s) = a_k s^{1-\theta}$. Each $a_k$ is bounded, so $a_{\max} \stackrel{\text{def}}{=} \max\{a_k\}_{k\geq 1}$ is bounded, and inequality (4.11) holds with the desingularizing function $\phi_{\max}(s) = a_{\max}s^{1-\theta}$.

Let $\Phi_k^\star \stackrel{\text{def}}{=} \min_{j\geq k}\frac{1}{\overline{n}^j}\sum_{i=1}^{\overline{n}^j}\Phi(\overline{z}_j^i)$. It is clear that $\Phi_k^*$ is non-decreasing and $\Phi_k^\star \to \mathbb{E}\Phi(z^*)$. From point 3, we can say there exists an index $m$ and a constant $a$ such that for all $k \geq m$,

$$a\Big(\frac{1}{\overline{n}^k}\sum_{i=1}^{\overline{n}^k}\Phi(z_k^i) - \Phi_k^\star\Big)^{-\theta} \geq a_{\max}\Big(\frac{1}{\overline{n}^k}\sum_{i=1}^{\overline{n}^k}\Phi(z_k^i) - \frac{1}{\overline{n}^k}\sum_{i=1}^{\overline{n}^k}\Phi(\overline{z}_k^i)\Big)^{-\theta}. \tag{4.12}$$

The constant $a$ exists; we can take $a$ to be

$$\max_{k\geq 1} \quad \left\{\left(\frac{\frac{1}{\overline{n}^k}\sum_{i=1}^{\overline{n}^k}\Phi(z_k^i) - \Phi_k^\star}{\frac{1}{\overline{n}^k}\sum_{i=1}^{\overline{n}^k}\Phi(z_k^i) - \frac{1}{\overline{n}^k}\sum_{i=1}^{\overline{n}^k}\Phi(\overline{z}_k^i)}\right)^{\theta}\right\}_{k\geq 1}, \tag{4.13}$$

which is bounded. To see this, we acknowledge that this ratio is bounded for every $k$, and

$$\lim_{k\to\infty}\left(\frac{\frac{1}{\overline{n}^k}\sum_{i=1}^{\overline{n}^k}\Phi(z_k^i) - \Phi_k^\star}{\frac{1}{\overline{n}^k}\sum_{i=1}^{\overline{n}^k}\Phi(z_k^i) - \frac{1}{\overline{n}^k}\sum_{i=1}^{\overline{n}^k}\Phi(\overline{z}_k^i)}\right) = \lim_{k\to\infty}\left(\frac{\frac{1}{\overline{n}^k}\sum_{i=1}^{\overline{n}^k}\Phi(z_k^i) - \mathbb{E}\Phi(z^*)}{\frac{1}{\overline{n}^k}\sum_{i=1}^{\overline{n}^k}\Phi(z_k^i) - \mathbb{E}\Phi(z^*)}\right) = 1. \tag{4.14}$$

Therefore, with $\phi(s) = as^{1-\theta}$, we have

$$\phi'(\mathbb{E}[\Phi(z_k) - \Phi_k^\star])\operatorname{dist}\big(0, \mathbb{E}\partial\Phi(z_k)\big) \geq \phi'_{\max}(\mathbb{E}[\Phi(z_k) - \Phi_k^\star])\operatorname{dist}\big(0, \mathbb{E}\partial\Phi(z_k)\big) \geq 1, \ \forall k > m,$$

The desired inequality follows from Jensen's inequality and the convexity of $x \mapsto \operatorname{dist}(0, x)$. ∎

We now show that the iterates of SPRING have finite length in expectation.

Lemma 4.6 (Finite Length). *Suppose $\Phi$ is a semi-algebraic function with KL exponent $\theta \in [0, 1)$. Let $\{z_k\}_{k=0}^\infty$ be a bounded sequence of iterates of SPRING using a variance-reduced gradient estimator and step-sizes satisfying the hypotheses of Lemma 4.3.*

(1). *Either $z_k$ is a critical point after a finite number of iterations, or $\{z_k\}_{k=0}^\infty$ satisfies the finite length property in expectation:*

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

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

$$\sum\nolimits_{k=m}^{i} \mathbb{E}\|z_{k+1} - z_k\| + \mathbb{E}\|z_k - z_{k-1}\| \le \sqrt{\mathbb{E}\|z_m - z_{m-1}\|^2} + \sqrt{\mathbb{E}\|z_{m-1} - z_{m-2}\|^2} + \tfrac{2\sqrt{s}}{K_1\rho}\sqrt{\mathbb{E}\Upsilon_{m-1}} + K_3\Delta_{m,i+1},$$

*where*

$$K_1 \overset{\text{def}}{=} p + 2\sqrt{sV_\Upsilon}/\rho, \qquad K_2 \overset{\text{def}}{=} \tfrac{1}{2\overline{\gamma}_0} - \tfrac{\bar{L}}{2} - \tfrac{3\sqrt{2}}{4}\sqrt{V_1 + V_\Upsilon/\rho}, \qquad K_3 \overset{\text{def}}{=} \tfrac{2K_1(K_2+Z)}{K_2 Z},$$

*$p$ is as in Lemma 4.2, and $\Delta_{p,q} \overset{\text{def}}{=} \phi(\mathbb{E}[\Psi_p - \Phi_p^\star]) - \phi(\mathbb{E}[\Psi_q - \Phi_q^\star])$.*

(2). *The iterates of SPRING $\{z_k\}_{k=0}^\infty$ converge to a critical point of $\Phi$ in expectation.*

*Remark* 4.7. Our analysis for SPRING requires $\Phi$ to be semi-algebraic for the finite-length property to hold, but in the analysis of PALM, the finite-length property requires only that $\Phi$ is KL (and not necessarily semi-algebraic) [6, Thm. 1]. This difference arises because SPRING does not necessarily decrease the objective every iteration (even in expectation), but PALM does [6, Lem. 3]. Instead, we prove that the iterates of SPRING decrease $\Psi_k$ in expectation. Related works [14] solve this problem by requiring an analog of $\Psi_k$ to be KL, but this is not a straightforward approach for SPRING because of the complex variance bounds required to analyze variance-reduced gradient estimators.

*Proof.* We begin with a proof of Claim 1. If $\theta \in (0, 1/2)$, then $\Phi$ satisfies the KL property with exponent $1/2$, so we consider only the case $\theta \in [1/2, 1)$. By Lemma 4.5, there exists a function $\phi_0(r) = ar^{1-\theta}$ such that

$$\phi_0'(\mathbb{E}[\Phi(z_k) - \Phi_k^\star])\mathbb{E}\mathrm{dist}\big(0, \partial\Phi(z_k)\big) \ge 1 \qquad \forall k > m.$$

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

$$\begin{aligned} \mathbb{E}\mathrm{dist}\big(0, \partial\Phi(z_k)\big) &\le \mathbb{E}\|(A_x^k, A_y^k)\| \le p\mathbb{E}[\|z_k - z_{k-1}\| + \|z_{k-1} - z_{k-2}\|] + \mathbb{E}\Gamma_{k-1} \\ &\le p\big(\sqrt{\mathbb{E}\|z_k - z_{k-1}\|^2} + \sqrt{\mathbb{E}\|z_{k-1} - z_{k-2}\|^2}\big) + \sqrt{s\mathbb{E}\Upsilon_{k-1}}. \end{aligned} \tag{4.15}$$

The final inequality is Jensen's. 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 (2.2):

$$\begin{aligned} \sqrt{\mathbb{E}\Upsilon_k} &\le \sqrt{(1-\rho)\mathbb{E}\Upsilon_{k-1} + V_\Upsilon\mathbb{E}[\|z_k - z_{k-1}\|^2 + \|z_{k-1} - z_{k-2}\|^2]} \\ &\le \sqrt{(1-\rho)}\sqrt{\mathbb{E}\Upsilon_{k-1}} + \sqrt{V_\Upsilon}\big(\sqrt{\mathbb{E}\|z_k - z_{k-1}\|^2} + \sqrt{\mathbb{E}\|z_{k-1} - z_{k-2}\|^2}\big) \\ &\le \big(1 - \tfrac{\rho}{2}\big)\sqrt{\mathbb{E}\Upsilon_{k-1}} + \sqrt{V_\Upsilon}\big(\sqrt{\mathbb{E}\|z_k - z_{k-1}\|^2} + \sqrt{\mathbb{E}\|z_{k-1} - z_{k-2}\|^2}\big). \end{aligned} \tag{4.16}$$

The final inequality uses the fact that $\sqrt{1-\rho} = 1 - \rho/2 - \rho^2/8 - \cdots$. This allows us to say

$$\mathbb{E}\mathrm{dist}\big(0, \partial\Phi(z_k)\big) \le K_1\sqrt{\mathbb{E}\|z_k - z_{k-1}\|^2} + K_1\sqrt{\mathbb{E}\|z_{k-1} - z_{k-2}\|^2} + \tfrac{2\sqrt{s}}{\rho}\big(\sqrt{\mathbb{E}\Upsilon_{k-1}} - \sqrt{\mathbb{E}\Upsilon_k}\big), \tag{4.17}$$

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

$$C_k \overset{\text{def}}{=} K_1\sqrt{\mathbb{E}\|z_k - z_{k-1}\|^2} + K_1\sqrt{\mathbb{E}\|z_{k-1} - z_{k-2}\|^2} + \tfrac{2\sqrt{s}}{\rho}\big(\sqrt{\mathbb{E}\Upsilon_{k-1}} - \sqrt{\mathbb{E}\Upsilon_k}\big).$$

We then have

$$\phi_0'(\mathbb{E}[\Phi(z_k) - \Phi_k^\star])C_k \geq 1 \qquad \forall k > m. \tag{4.18}$$

By the definition of $\phi_0$, this is equivalent to

$$\frac{a(1-\theta)C_k}{(\mathbb{E}[\Phi(z_k) - \Phi_k^\star])^\theta} \geq 1 \qquad \forall k > m. \tag{4.19}$$

We would like the inequality above to hold for $\Psi_k$ rather than $\Phi(z_k)$. Replacing $\mathbb{E}\Phi(z_k)$ with $\mathbb{E}\Psi_k$ introduces a term of $\mathcal{O}((\mathbb{E}[\|z_k - z_{k-1}\|^2 + \Upsilon_k])^\theta)$ in the denominator. We show that inequality (4.19) still holds after this adjustment because these terms are small compared to $C_k$.

The quantity $C_k \geq c_1(\sqrt{\mathbb{E}\|z_k - z_{k-1}\|^2} + \sqrt{\mathbb{E}\|z_{k-1} - z_{k-2}\|^2} + \sqrt{\mathbb{E}\Upsilon_{k-1}})$ for some constant $c_1 > 0$, and because $\mathbb{E}\|z_k - z_{k-1}\|^2, \mathbb{E}\Upsilon_k \to 0$, and $\theta \geq 1/2$, there exists an index $m$ and a constants $c_2, c_3 > 0$ such that

$$\begin{aligned}&\left(\mathbb{E}\left[\frac{1}{2\rho\sqrt{2(V_1+V_\Upsilon/\rho)}}\Upsilon_k + \frac{\sqrt{V_1+V_\Upsilon/\rho}}{\sqrt{2}}\|z_k - z_{k-1}\|^2\right]\right)^\theta \\ &\leq c_2\Big(\left(\mathbb{E}\left[\Upsilon_{k-1} + \|z_k - z_{k-1}\|^2 + \|z_{k-1} - z_{k-2}\|^2\right]\right)^\theta\Big) \leq c_3 C_k \qquad \forall k > m.\end{aligned}$$

The first inequality uses (2.2). Because the terms above are small compared to $C_k$, there exists a constant $+\infty > d > c_3$ such that

$$\frac{ad(1-\theta)C_k}{(\mathbb{E}[\Phi(z_k)-\Phi_k^\star])^\theta + \left(\mathbb{E}[\frac{1}{2\rho\sqrt{2(V_1+V_\Upsilon/\rho)}}\Upsilon_k + \frac{\sqrt{V_1+V_\Upsilon/\rho}}{\sqrt{2}}\|z_k-z_{k-1}\|^2]\right)^\theta} \geq 1,$$

for all $k > m$. Using the fact that $(a+b)^\theta \leq a^\theta + b^\theta$ for all $a, b \geq 0$ because $\theta \in [1/2, 1)$, we have

$$\begin{aligned}\frac{ad(1-\theta)C_k}{(\mathbb{E}[\Psi_k - \Psi^\star])^\theta} &= \frac{ad(1-\theta)C_k}{\left(\mathbb{E}\left[\Phi(z_k) - \Phi_k^\star + \frac{1}{2\rho\sqrt{2(V_1+V_\Upsilon/\rho)}}\Upsilon_k + \frac{\sqrt{V_1+V_\Upsilon/\rho}}{\sqrt{2}}\|z_k - z_{k-1}\|^2\right]\right)^\theta} \\ &\geq \frac{ad(1-\theta)C_k}{\left(\mathbb{E}\left[\Phi(z_k)-\Phi_k^\star\right]\right)^\theta + \left(\mathbb{E}\left[\frac{1}{2\rho\sqrt{2(V_1+V_\Upsilon/\rho)}}\Upsilon_k + \frac{\sqrt{V_1+V_\Upsilon/\rho}}{\sqrt{2}}\|z_k-z_{k-1}\|^2\right]\right)^\theta} \geq 1 \qquad \forall k > m.\end{aligned}$$

Therefore, with $\phi(r) = adr^{1-\theta}$,

$$\phi'(\mathbb{E}[\Psi_k - \Phi_k^\star])C_k \geq 1 \qquad \forall k > m.$$

By the concavity of $\phi$,

$$\begin{aligned}\phi(\mathbb{E}[\Psi_k - \Phi_k^\star]) - \phi(\mathbb{E}[\Psi_{k+1} - \Phi_{k+1}^\star]) &\geq \phi'(\mathbb{E}[\Psi_k - \Phi_k^\star])(\mathbb{E}[\Psi_k - \Phi_k^\star + \Phi_{k+1}^\star - \Psi_{k+1}]) \\ &\geq \phi'(\mathbb{E}[\Psi_k - \Phi_k^\star])(\mathbb{E}[\Psi_k - \Psi_{k+1}]),\end{aligned} \tag{4.20}$$

where the last inequality follows from the fact that $\Phi_k^\star$ is non-decreasing. With $\Delta_{p,q} \stackrel{\text{def}}{=} \phi(\mathbb{E}[\Psi_p - \Phi_p^\star]) - \phi(\mathbb{E}[\Psi_q - \Phi_q^\star])$, we have shown

$$\Delta_{k,k+1}C_k \geq \mathbb{E}[\Psi_k - \Psi_{k+1}].$$

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

$$\Delta_{k,k+1} C_k \geq Z \mathbb{E}[\|z_k - z_{k-1}\|^2], \tag{4.21}$$

as well as

$$\Delta_{k,k+1} C_k \geq K_2 \mathbb{E}[\|z_{k+1} - z_k\|^2], \tag{4.22}$$

where $K_2 \overset{\text{def}}{=} -\left(\frac{\bar{L}(\lambda+1)}{2} + \frac{V_1 + V_\Upsilon/\rho}{\bar{L}\lambda} + Z - \frac{1}{2\bar{\gamma}_0}\right)$ and $\lambda$ and $Z$ are set as in Lemma 4.1. Let us use the first of these inequalities to begin. Applying Young's inequality to (4.21) yields

$$2\sqrt{\mathbb{E}\|z_k - z_{k-1}\|^2} \leq 2\sqrt{C_k \Delta_{k,k+1} Z^{-1}} \leq \frac{C_k}{2K_1} + \frac{2K_1 \Delta_{k,k+1}}{Z} \tag{4.23}$$

Summing inequality (4.23) from $k = m$ to $k = i$,

$$\begin{aligned} 2\sum_{k=m}^{i} \sqrt{\mathbb{E}\|z_k - z_{k-1}\|^2} &\leq \sum_{k=m}^{i} \frac{C_k}{2K_1} + \frac{2K_1 \Delta_{m,i+1}}{Z} \\ &\leq \sum_{k=m}^{i} \frac{1}{2}\sqrt{\mathbb{E}\|z_k - z_{k-1}\|^2} + \frac{1}{2}\sqrt{\mathbb{E}\|z_{k-1} - z_{k-2}\|^2} \\ &\quad - \frac{\sqrt{s}}{K_1 \rho}\left(\sqrt{\mathbb{E}\Upsilon_i} - \sqrt{\mathbb{E}\Upsilon_{m-1}}\right) + \frac{2K_1 \Delta_{m,i+1}}{Z}, \end{aligned} \tag{4.24}$$

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

$$\sum_{k=m}^{i} \sqrt{\mathbb{E}\|z_k - z_{k-1}\|^2} \leq \frac{1}{2}\sqrt{\mathbb{E}\|z_{m-1} - z_{m-2}\|^2} + \frac{\sqrt{s}}{K_1 \rho}\sqrt{\mathbb{E}\Upsilon_{m-1}} + \frac{2K_1 \Delta_{m,i+1}}{Z}.$$

Applying the same argument using inequality (4.22) instead of (4.21), we obtain

$$\begin{aligned} &\sum_{k=m}^{i} \sqrt{\mathbb{E}\|z_{k+1} - z_k\|^2} \\ &\leq \frac{1}{2}\sqrt{\mathbb{E}\|z_m - z_{m-1}\|^2} + \frac{1}{2}\sqrt{\mathbb{E}\|z_{m-1} - z_{m-2}\|^2} + \frac{\sqrt{s}}{K_1 \rho}\sqrt{\mathbb{E}\Upsilon_{m-1}} + \frac{2K_1 \Delta_{m,i+1}}{K_2}. \end{aligned}$$

Adding these inequalities together, we have

$$\begin{aligned} \sum_{k=m}^{i} \sqrt{\mathbb{E}\|z_{k+1} - z_k\|^2} + \sqrt{\mathbb{E}\|z_k - z_{k-1}\|^2} &\leq \frac{1}{2}\sqrt{\mathbb{E}\|z_m - z_{m-1}\|^2} + \sqrt{\mathbb{E}\|z_{m-1} - z_{m-2}\|^2} \\ &\quad + \frac{2\sqrt{s}}{K_1 \rho}\sqrt{\mathbb{E}\Upsilon_{m-1}} + \frac{2K_1(K_2 + Z)\Delta_{m,i+1}}{K_2 Z}. \end{aligned}$$

For easier analysis, we add $\frac{1}{2}\sqrt{\mathbb{E}\|z_m - z_{m-1}\|^2}$ to the right side:

$$\begin{aligned} &\sum_{k=m}^{i} \sqrt{\mathbb{E}\|z_{k+1} - z_k\|^2} + \sqrt{\mathbb{E}\|z_k - z_{k-1}\|^2} \\ &\leq \sqrt{\mathbb{E}\|z_m - z_{m-1}\|^2} + \sqrt{\mathbb{E}\|z_{m-1} - z_{m-2}\|^2} + \frac{2\sqrt{s}}{K_1 \rho}\sqrt{\mathbb{E}\Upsilon_{m-1}} + \frac{2K_1(K_2 + Z)\Delta_{m,i+1}}{K_2 Z}. \end{aligned} \tag{4.25}$$

Applying Jensen's inequality to the terms on the left gives

$$\begin{aligned} &\sum_{k=m}^{i} \mathbb{E}\|z_{k+1} - z_k\| + \mathbb{E}\|z_k - z_{k-1}\| \\ &\leq \sqrt{\mathbb{E}\|z_m - z_{m-1}\|^2} + \sqrt{\mathbb{E}\|z_{m-1} - z_{m-2}\|^2} + \frac{2\sqrt{s}}{K_1 \rho}\sqrt{\mathbb{E}\Upsilon_{m-1}} + \frac{2K_1(K_2 + Z)\Delta_{m,i+1}}{K_2 Z}. \end{aligned}$$

The term $\lim_{i\to\infty} \Delta_{m,i+1}$ is bounded because $\mathbb{E}\Psi_k$ is bounded due to Lemma 4.1, so letting $i \to \infty$ proves the assertion.

An immediate consequence of Claim 1 is that the sequence $\mathbb{E}\|z_{k+1} - z_k\|$ is Cauchy, so the sequence $\{z_k\}_{k=0}^\infty$ converges in expectation to a critical point. This is because, for any $p, q \in \mathbb{N}$ with $p \geq q$, $\mathbb{E}\|z_p - z_q\| = \mathbb{E}\| \sum_{k=q}^{p-1} z_{k+1} - z_k\| \leq \sum_{k=q}^{p-1} \mathbb{E}\|z_{k+1} - z_k\|$, and the finite length property implies this final sum converges to zero. This proves Claim 2. ■

Finally, we prove convergence rates for SPRING depending on the KL exponent of the objective function, demonstrating that the full convergence theory of PALM extends to SPRING.

Theorem 4.8 (Convergence Rates). *Suppose $\Phi$ is a semi-algebraic function with KL exponent $\theta \in [0, 1)$. Let $\{z_k\}_{k=0}^\infty$ be a bounded sequence of iterates of SPRING using a variance-reduced gradient estimator and step-sizes satisfying the hypotheses of Lemma 4.3. The following convergence rates hold:*

(1). *If $\theta \in (0, 1/2]$, then there exists $d_1 > 0$ and $\tau \in [1-\rho, 1)$ such that $\mathbb{E}\|z_k - z^\star\| \leq d_1 \tau^k$.*

(2). *If $\theta \in (1/2, 1)$, then there exists a constant $d_2 > 0$ such that $\mathbb{E}\|z_k - z^\star\| \leq d_2 k^{-\frac{1-\theta}{2\theta-1}}$.*

(3). *If $\theta = 0$, then there exists an $m \in \mathbb{N}$ such that $\mathbb{E}\Phi(z_k) = \mathbb{E}\Phi(z^\star)$ for all $k \geq m$.*

*Proof.* As in the proof of the previous lemma, if $\theta \in (0, 1/2)$, then $\Phi$ satisfies the KL property with exponent $1/2$, so we consider only the case $\theta \in [1/2, 1)$.

Substituting the desingularizing function $\phi(r) = ar^{1-\theta}$ into (4.25),

$$\begin{aligned} &\sum_{k=m}^\infty \sqrt{\mathbb{E}\|z_{k+1} - z_k\|^2} + \sqrt{\mathbb{E}\|z_k - z_{k-1}\|^2} \\ &\leq \sqrt{\mathbb{E}\|z_m - z_{m-1}\|^2} + \sqrt{\mathbb{E}\|z_{m-1} - z_{m-2}\|^2} + \tfrac{2\sqrt{s}}{K_1\rho}\sqrt{\mathbb{E}\Upsilon_{m-1}} + aK_3(\mathbb{E}[\Psi_m - \Phi_m^\star])^{1-\theta}. \end{aligned} \tag{4.26}$$

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

$$\begin{aligned} (\mathbb{E}[\Psi_m - \Phi_m^\star])^{1-\theta} &= (\mathbb{E}[\Phi(z_m) - \Phi_m^\star + \tfrac{1}{2L\lambda\rho}\Upsilon_m + \tfrac{V_1 + V_\Upsilon/\rho}{2L\lambda}\|z_m - z_{m-1}\|^2])^{1-\theta} \\ &\overset{①}{\leq} (\mathbb{E}[\Phi(z_m) - \Phi_m^\star])^{1-\theta} + \big(\tfrac{1}{2L\lambda\rho}\mathbb{E}\Upsilon_m\big)^{1-\theta} + \big(\tfrac{V_1 + V_\Upsilon/\rho}{2L\lambda}\mathbb{E}\|z_m - z_{m-1}\|^2\big)^{1-\theta}. \end{aligned}$$

Inequality ① is due to the fact that $(a+b)^{1-\theta} \leq a^{1-\theta} + b^{1-\theta}$. This yields the inequality

$$\begin{aligned} &\sum_{k=m}^\infty \sqrt{\mathbb{E}\|z_{k+1} - z_k\|^2} + \sqrt{\mathbb{E}\|z_k - z_{k-1}\|^2} \\ &\leq \sqrt{\mathbb{E}\|z_m - z_{m-1}\|^2} + \sqrt{\mathbb{E}\|z_{m-1} - z_{m-2}\|^2} + \tfrac{2\sqrt{s}}{K_1\rho}\sqrt{\mathbb{E}\Upsilon_{m-1}} + aK_3(\mathbb{E}[\Phi(z_m) - \Phi_m^\star])^{1-\theta} \\ &\quad + aK_3\big(\tfrac{1}{2L\lambda\rho}\mathbb{E}\Upsilon_m\big)^{1-\theta} + aK_3\big(\tfrac{V_1 + V_\Upsilon/\rho}{2L\lambda}\mathbb{E}\|z_m - z_{m-1}\|^2\big)^{1-\theta}. \end{aligned}$$

Applying the Kurdyka–Łojasiewicz inequality (2.4),

$$aK_3(\mathbb{E}\,[\Phi(z_m) - \Phi_m^\star])^{1-\theta} \leq aK_3(\mathbb{E}\|\zeta_m\|)^{\frac{1-\theta}{\theta}}, \tag{4.27}$$

for all $\zeta_m \in \partial\Phi(z_m)$ and we have absorbed the constant $C$ into $a$. Equation (4.15) provides a bound on the norm of the subgradient:

$$(\mathbb{E}\|\zeta_m\|)^{\frac{1-\theta}{\theta}} \leq \big(p(\sqrt{\mathbb{E}\|z_m - z_{m-1}\|^2} + \sqrt{\mathbb{E}\|z_{m-1} - z_{m-2}\|^2}) + \sqrt{s\mathbb{E}\Upsilon_{m-1}}\big)^{\frac{1-\theta}{\theta}}.$$

Denote the right side of this inequality $\Theta_m^{\frac{1-\theta}{\theta}}$. Therefore,

$$\begin{aligned}
&\textstyle\sum_{k=m}^{\infty} \sqrt{\mathbb{E}\|z_{k+1}-z_k\|^2} + \sqrt{\mathbb{E}\|z_k - z_{k-1}\|^2} \\
&\le \sqrt{\mathbb{E}\|z_m - z_{m-1}\|^2} + \sqrt{\mathbb{E}\|z_{m-1}-z_{m-2}\|^2} + \tfrac{2\sqrt{s}}{K_1\rho}\sqrt{\mathbb{E}\Upsilon_{m-1}} + aK_3\Theta_m^{\frac{1-\theta}{\theta}} \\
&\quad + aK_3\big(\tfrac{1}{2L\lambda\rho}\mathbb{E}\Upsilon_m\big)^{1-\theta} + aK_3\big(\tfrac{V_1+V_\Upsilon/\rho}{2L\lambda}\mathbb{E}\|z_m - z_{m-1}\|^2\big)^{1-\theta}.
\end{aligned} \tag{4.28}$$

Suppose $\theta \in (1/2, 1)$. Each of the terms on the right side of this inequality are converging to zero, but at different rates. Because $\Theta_m = \mathcal{O}(\sqrt{\mathbb{E}\|z_m - z_{m-1}\|^2} + \sqrt{\mathbb{E}\|z_{m-1}-z_{m-2}\|^2} + \sqrt{\mathbb{E}\Upsilon_{m-1}})$, and $\theta$ satisfies $\frac{1-\theta}{\theta} < 1$, the term $\Theta_m^{\frac{1-\theta}{\theta}}$ dominates the first three terms on the right side of this inequality for large $m$. Also, because $\frac{1-\theta}{2\theta} \le 1-\theta$, $\Theta_m^{\frac{1-\theta}{\theta}}$ dominates the final two terms as well. Combining these facts, there exists a natural number $M_1$ such that for all $m \ge M_1$,

$$\Big(\textstyle\sum_{k=m}^{\infty} \sqrt{\mathbb{E}\|z_{k+1}-z_k\|^2} + \sqrt{\mathbb{E}\|z_k - z_{k-1}\|^2}\Big)^{\frac{\theta}{1-\theta}} \le P\Theta_m, \tag{4.29}$$

for some constant $P > (aK_3)^{\frac{\theta}{1-\theta}}$. The bound of (4.16) implies

$$2\sqrt{s\mathbb{E}\Upsilon_{m-1}} \le \tfrac{4\sqrt{s}}{\rho}\big(\sqrt{\mathbb{E}\Upsilon_{m-1}} - \sqrt{\mathbb{E}\Upsilon_m} + \sqrt{V_\Upsilon}(\sqrt{\mathbb{E}\|z_m - z_{m-1}\|^2} + \sqrt{\mathbb{E}\|z_{m-1}-z_{m-2}\|^2})\big).$$

Therefore,

$$\begin{aligned}
\Theta_m &= p(\sqrt{\mathbb{E}\|z_m - z_{m-1}\|^2} + \sqrt{\mathbb{E}\|z_{m-1}-z_{m-2}\|^2}) + (2\sqrt{s\mathbb{E}\Upsilon_{m-1}} - \sqrt{s\mathbb{E}\Upsilon_{m-1}}) \\
&\le \big(p + \tfrac{4\sqrt{sV_\Upsilon}}{\rho}\big)(\sqrt{\mathbb{E}\|z_m - z_{m-1}\|^2} + \sqrt{\mathbb{E}\|z_{m-1}-z_{m-2}\|^2}) \\
&\quad + \tfrac{4\sqrt{s}}{\rho}(\sqrt{\mathbb{E}\Upsilon_{m-1}} - \sqrt{\mathbb{E}\Upsilon_m}) - \sqrt{s\mathbb{E}\Upsilon_{m-1}}.
\end{aligned} \tag{4.30}$$

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

$$\big(\tfrac{4\sqrt{s}(1-\rho/4)}{\rho(p+4\sqrt{sV_\Upsilon}/\rho)}\sqrt{\mathbb{E}\Upsilon_m}\big)^{\frac{\theta}{1-\theta}} \le P\sqrt{s\mathbb{E}\Upsilon_m}. \tag{4.31}$$

The constant appearing on the left was chosen to simplify later arguments. Therefore, (4.29) implies

$$\begin{aligned}
&\big(\textstyle\sum_{k=m}^{\infty} \sqrt{\mathbb{E}\|z_{k+1}-z_k\|^2} + \sqrt{\mathbb{E}\|z_k - z_{k-1}\|^2} + \tfrac{4\sqrt{s}(1-\rho/4)}{\rho(p+4\sqrt{sV_\Upsilon}/\rho)}\sqrt{\mathbb{E}\Upsilon_m}\big)^{\frac{\theta}{1-\theta}} \\
&\overset{①}{\le} \tfrac{2^{\frac{\theta}{1-\theta}}}{2}\big(\textstyle\sum_{k=m}^{\infty} \sqrt{\mathbb{E}\|z_{k+1}-z_k\|^2} + \sqrt{\mathbb{E}\|z_k - z_{k-1}\|^2}\big)^{\frac{\theta}{1-\theta}} \\
&\quad + \tfrac{2^{\frac{\theta}{1-\theta}}}{2}\big(\tfrac{4\sqrt{s}(1-\rho/4)}{\rho(p+4\sqrt{sV_\Upsilon}/\rho)}\sqrt{\mathbb{E}\Upsilon_m}\big)^{\frac{\theta}{1-\theta}} \\
&\overset{②}{\le} \tfrac{2^{\frac{\theta}{1-\theta}}}{2}\big(\textstyle\sum_{k=m}^{\infty} \sqrt{\mathbb{E}\|z_{k+1}-z_k\|^2} + \sqrt{\mathbb{E}\|z_k - z_{k-1}\|^2}\big)^{\frac{\theta}{1-\theta}} + \tfrac{2^{\frac{\theta}{1-\theta}}}{2}\big(P\sqrt{s\mathbb{E}\Upsilon_m}\big) \\
&\overset{③}{\le} \tfrac{2^{\frac{\theta}{1-\theta}}}{2}\Big(P(p + 4\sqrt{sV_\Upsilon}/\rho)\big(\sqrt{\mathbb{E}\,\|z_m - z_{m-1}\|^2} + \sqrt{\|z_{m-1}-z_{m-2}\|^2}\big) \\
&\quad + \tfrac{4\sqrt{s}P(1-\rho/4)}{\rho}\big(\sqrt{\mathbb{E}\Upsilon_{m-1}} - \sqrt{\mathbb{E}\Upsilon_m}\big)\Big).
\end{aligned}$$

Here, ① follows by convexity of the function $x^{\frac{\theta}{1-\theta}}$ for $\theta \in [1/2, 1)$ and $x \geq 0$, ② is (4.31), and ③ is (4.29) combined with (4.30). We absorb the constant $\frac{2^{\frac{\theta}{1-\theta}}}{2}$ into $P$. Define

$$S_m \overset{\text{def}}{=} \sum_{k=m}^{\infty} \sqrt{\mathbb{E}\|z_{k+1} - z_k\|^2} + \sqrt{\mathbb{E}\|z_k - z_{k-1}\|^2} + \tfrac{4\sqrt{s}P(1-\rho/4)}{\rho(p+4\sqrt{sV_\Upsilon}/\rho)} \sqrt{\mathbb{E}\Upsilon_m}.$$

$S_m$ is bounded for all $m$ because $\sum_{k=m}^{\infty} \sqrt{\mathbb{E}\|z_{k+1} - z_k\|^2}$ is bounded by equation (4.26). Hence, we have shown

$$S_m^{\frac{\theta}{1-\theta}} \leq P(p + 4\sqrt{sV_\Upsilon}/\rho)(S_{m-1} - S_m). \tag{4.32}$$

The rest of the proof follows the proof of [2, Theorem 5]. Let $h(r) \overset{\text{def}}{=} r^{-\frac{\theta}{1-\theta}}$. First, suppose that $h(S_m) \leq R h(S_{m-1})$ for some $R \in (1, \infty)$. Then (4.32) ensures that

$$\begin{aligned} 1 \leq P(p + 4\sqrt{sV_\Upsilon}/\rho)(S_{m-1} - S_m)h(S_m) &\leq RP(p + 4\sqrt{sV_\Upsilon}/\rho)(S_{m-1} - S_m)h(S_{m-1}) \\ &\leq RP(p + 4\sqrt{sV_\Upsilon}/\rho) \int_{S_m}^{S_{m-1}} h(r)dr \\ &= \tfrac{RP(p+4\sqrt{sV_\Upsilon}/\rho)(1-\theta)}{1-2\theta} \left[ S_{m-1}^{\frac{1-2\theta}{1-\theta}} - S_m^{\frac{1-2\theta}{1-\theta}} \right]. \end{aligned}$$

Hence,

$$0 < -\tfrac{1-2\theta}{RP(p+4\sqrt{sV_\Upsilon}/\rho)(1-\theta)} \leq S_m^{\frac{1-2\theta}{1-\theta}} - S_{m-1}^{\frac{1-2\theta}{1-\theta}}.$$

Now suppose $h(S_m) > Rh(S_{m-1})$, so that $S_m < R^{-\frac{1-\theta}{\theta}} S_{m-1}$ and $S_m^{\frac{1-2\theta}{1-\theta}} > q^{\frac{1-2\theta}{1-\theta}} S_{m-1}^{\frac{1-2\theta}{1-\theta}}$ where $q = R^{-\frac{1-\theta}{\theta}}$. This implies that

$$\big(q^{\frac{1-2\theta}{1-\theta}} - 1\big) S_{m-1}^{\frac{1-2\theta}{1-\theta}} \leq S_m^{\frac{1-2\theta}{1-\theta}} - S_{m-1}^{\frac{1-2\theta}{1-\theta}},$$

and the quantity on the left is clearly bounded away from zero because $q < 1$, $\frac{1-2\theta}{1-\theta} < 0$, and $S_{m-1} \to 0$. This shows that in either case, there exists a $\mu' > 0$ such that

$$\mu' \leq S_m^{\frac{1-2\theta}{1-\theta}} - S_{m-1}^{\frac{1-2\theta}{1-\theta}}.$$

Summing this inequality from $m = M_2$ to $m = M_3$, we obtain $(M_3 - M_2)\mu' \leq S_{M_3}^{\frac{1-2\theta}{1-\theta}} - S_{M_2-1}^{\frac{1-2\theta}{1-\theta}}$, and because the function $x \mapsto x^{\frac{1-\theta}{1-2\theta}}$ is decreasing, this implies

$$S_{M_3} \leq \big(S_{M_2-1}^{\frac{1-2\theta}{1-\theta}} + (M_3 - M_2)\mu'\big)^{\frac{1-\theta}{1-2\theta}} \leq d M_3^{\frac{1-\theta}{1-2\theta}},$$

for some constant $d$. By Jensen's inequality, we can say $\sum_{k=M_3}^{\infty} \mathbb{E}\|z_k - z_{k-1}\| \leq S_{M_3} \leq d M_3^{-\frac{1-\theta}{2\theta-1}}$. Using the fact that $\mathbb{E}\|z_m - z^\star\| = \mathbb{E}\|\sum_{k=m+1}^{\infty} z_k - z_{k-1}\| \leq \mathbb{E}\sum_{k=m}^{\infty} \|z_k - z_{k-1}\|$ proves Claim 1.

If $\theta = 1/2$, then $(\mathbb{E}\|\zeta_m\|)^{\frac{1-\theta}{\theta}} = \mathbb{E}\|\zeta_m\|$. Equation (4.28) then gives

$$\begin{aligned}
&\sum\nolimits_{i=m}^{\infty} \sqrt{\mathbb{E}\|z_{k+1}-z_k\|^2} + \sqrt{\mathbb{E}\|z_k - z_{k-1}\|^2} \\
&\le \left(1 + aK_3\big(p + \sqrt{\tfrac{V_1+V_\Upsilon/\rho}{2\bar{L}\lambda}}\big)\right)\big(\sqrt{\mathbb{E}\|z_m - z_{m-1}\|^2} + \sqrt{\mathbb{E}\|z_{m-1}-z_{m-2}\|^2}\big) \\
&\quad + \left(\tfrac{2\sqrt{s}}{K_1\rho} + aK_3\sqrt{s}\right)\sqrt{\mathbb{E}\Upsilon_{m-1}} + aK_3\sqrt{\tfrac{1}{2\bar{L}\lambda\rho}}\sqrt{\mathbb{E}\Upsilon_m},
\end{aligned} \tag{4.33}$$

where we have added the non-negative term $aK_3\sqrt{\frac{V_1+V_\Upsilon/\rho}{2\bar{L}\lambda}}\sqrt{\mathbb{E}\|z_{m-1}-z_{m-2}\|^2}$ to the right to simplify the presentation. Using equation (4.16), we have that, for any constant $c > 0$,

$$0 \le -c\sqrt{\mathbb{E}\Upsilon_m} + c\big(1-\tfrac{\rho}{2}\big)\sqrt{\mathbb{E}\Upsilon_{m-1}} + c\sqrt{V_\Upsilon}\big(\sqrt{\mathbb{E}\|z_m - z_{m-1}\|^2} + \sqrt{\mathbb{E}\|z_{m-1}-z_{m-2}\|^2}\big).$$

Combining this inequality with (4.33),

$$\begin{aligned}
&\sum\nolimits_{i=m}^{\infty} \sqrt{\mathbb{E}\|z_{k+1}-z_k\|^2} + \sqrt{\mathbb{E}\|z_k - z_{k-1}\|^2} \\
&\le \big(1 + aK_3\big(p + \sqrt{\tfrac{V_1+V_\Upsilon/\rho}{2\bar{L}\lambda}}\big) + c\sqrt{V_\Upsilon}\big)\big(\sqrt{\mathbb{E}\|z_m - z_{m-1}\|^2} + \sqrt{\mathbb{E}\|z_{m-1}-z_{m-2}\|^2}\big) \\
&\quad + c\big(1 - \tfrac{\rho}{2} + \tfrac{2\sqrt{s}}{cK_1\rho} + \tfrac{aK_3\sqrt{s}}{c}\big)\sqrt{\mathbb{E}\Upsilon_{m-1}} - c\big(1 - aK_3c^{-1}\sqrt{\tfrac{1}{2\bar{L}\lambda\rho}}\big)\sqrt{\mathbb{E}\Upsilon_m}.
\end{aligned}$$

Defining

$$T_m \stackrel{\text{def}}{=} \sum\nolimits_{i=m}^{\infty} \sqrt{\mathbb{E}\|z_{i+1}-z_i\|^2} + \sqrt{\mathbb{E}\|z_i - z_{i-1}\|^2},$$

and $P_2 = 1 + aK_3\big(p + 4\sqrt{sV_\Upsilon}/\rho + \sqrt{\frac{V_1+V_\Upsilon/\rho}{2\bar{L}\lambda}}\big) + c\sqrt{V_\Upsilon}$, we have shown

$$\begin{aligned}
&T_m + c\big(1 - aK_3c^{-1}\sqrt{\tfrac{1}{2\bar{L}\lambda\rho}}\big)\sqrt{\mathbb{E}\Upsilon_m} \\
&\le P_2(T_{m-1} - T_m) + c\big(1 - \tfrac{\rho}{2} + \tfrac{2\sqrt{s}}{cK_1\rho} + \tfrac{aK_3\sqrt{s}}{c}\big)\sqrt{\mathbb{E}\Upsilon_{m-1}}.
\end{aligned}$$

Rearranging,

$$(1+P_2)T_m + c\big(1 - aK_3c^{-1}\sqrt{\tfrac{1}{2\bar{L}\lambda\rho}}\big)\sqrt{\mathbb{E}\Upsilon_m} \le P_2 T_{m-1} + c\big(1 - \tfrac{\rho}{2} + \tfrac{2\sqrt{s}}{cK_1\rho} + \tfrac{aK_3\sqrt{s}}{c}\big)\sqrt{\mathbb{E}\Upsilon_{m-1}}.$$

This implies

$$\begin{aligned}
&T_m + \sqrt{\mathbb{E}\Upsilon_m} \\
&\le \max\left\{\frac{P_2}{1+P_2}, \big(1 - \tfrac{\rho}{2} + \tfrac{2\sqrt{s}}{cK_1\rho} + \tfrac{aK_3\sqrt{s}}{c}\big)\big(1 - aK_3c^{-1}\sqrt{\tfrac{1}{2\bar{L}\lambda\rho}}\big)^{-1}\right\}\big(T_{m-1} + \sqrt{\mathbb{E}\Upsilon_{m-1}}\big).
\end{aligned}$$

For large $c$, the second coefficient in the above expression approaches $1 - \rho/2$. This proves the linear rate of Claim 2.

When $\theta = 0$, the KL property (2.4) implies that exactly one of the following two scenarios holds: either $\mathbb{E}\Phi(z_k) \neq \Phi_k^\star$ and

$$0 < C \le \mathbb{E}\|\zeta_k\| \qquad \forall \zeta_k \in \partial\Phi(z_k), \tag{4.34}$$

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

Using the subgradient bound, the first scenario implies

$$
\begin{aligned}
C^2 \le (\mathbb{E}\|\zeta_k\|)^2 &\le (p\mathbb{E}\|z_k - z_{k-1}\| + p\mathbb{E}\|z_{k-1} - z_{k-2}\| + \mathbb{E}\Gamma_{k-1})^2, \\
&\le 3p^2(\mathbb{E}\|z_k - z_{k-1}\|)^2 + 3p^2(\mathbb{E}\|z_{k-1} - z_{k-2}\|)^2 + 3(\mathbb{E}\Gamma_{k-1})^2, \\
&\le 3p^2\mathbb{E}\|z_k - z_{k-1}\|^2 + 3p^2\mathbb{E}\|z_{k-1} - z_{k-2}\|^2 + 3s\mathbb{E}\Upsilon_{k-1}.
\end{aligned}
$$

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

$$
\begin{aligned}
\mathbb{E}\Psi_k &\le \mathbb{E}\Psi_{k-1} + \left(\tfrac{\bar{L}(\lambda+1)}{2} + \tfrac{V_1 + V_\Gamma/\rho}{2\bar{L}\lambda} + Z - \tfrac{1}{2\eta}\right)\mathbb{E}\|z_k - z_{k-1}\|^2 - Z\mathbb{E}\|z_{k-1} - z_{k-2}\|^2 \\
&\le \mathbb{E}\Psi_{k-1} - C^2 + \mathcal{O}(\mathbb{E}\|z_k - z_{k-1}\|^2) + \mathcal{O}(\mathbb{E}\|z_{k-1} - z_{k-2}\|^2) + \mathcal{O}(\mathbb{E}\Upsilon_{k-1}),
\end{aligned}
$$

for some constant $C^2$.[7] Because the final three terms go to zero as $k \to \infty$, there exists an index $M_4$ so that the sum of these three terms is bounded above by $C^2/2$ for all $k \ge M_4$. Therefore,

$$
\mathbb{E}\Psi_k \le \mathbb{E}\Psi_{k-1} - \tfrac{C^2}{2}, \qquad \forall k \ge M_4.
$$

Because $\Psi_k$ 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 (4.34) to hold, so it must be that $\Phi(z_k) = \Phi_k^\star$. Because $\Phi_k^\star < \Phi(z^\star)$, $\Phi_k^\star < \mathbb{E}\Phi(z_k)$, and both $\mathbb{E}\Phi(z_k)$, $\Phi_k^\star$ converge to $\mathbb{E}\Phi(z^\star)$, we must have $\Phi_k^\star = \mathbb{E}\Phi(z_k) = \mathbb{E}\Phi(z^\star)$. ■

The main difference between these convergence rates and those of PALM occurs when $\theta \in (0, 1/2]$. In this case, the linear convergence rate cannot be faster than the geometric decay of the MSE of the gradient estimator, which is of order $(1-\rho)^k$ after $k$ iterations. Without mini-batching (i.e. $b = 1$), this rate is approximately $(1 - 1/n)^k$ for the SAGA estimator and $(1 - 1/p)^k$ for the SARAH estimator.

## 5. Numerical Experiments.

To demonstrate the advantages of SPRING, we compare SPRING using the SAGA and SARAH gradient estimators to PALM [6] and inertial PALM [29]. We also present results for SPRING using the (non-variance-reduced) SGD estimator (a case studied by Xu and Yin [39]). We refer to SPRING using the SGD, SAGA, and SARAH gradient estimators as SPRING-SGD, SPRING-SAGA, and SPRING-SARAH, respectively. Two applications are considered here for comparison: sparse non-negative matrix factorization (Sparse-NMF) and blind image-deblurring (BID)[8]. We also provide in the appendix additional results on sparse principal component analysis (Sparse-PCA).

**Sparse-NMF:** Given a data-matrix $A$, we seek a factorization $A \approx XY$ where $X \in \mathbb{R}^{n\times r}, Y \in \mathbb{R}^{r\times d}$ are non-negative with $r \le d$ and $X$ sparse. Sparse-NMF has the following formulation:

$$
(5.1) \qquad \min_{X,Y} \|A - XY\|_F^2, \qquad \text{s.t. } X, Y \ge 0, \;\; \|X_i\|_0 \le s, \; i = 1, ..., r.
$$

Here, $X_i$ denotes the $i^{th}$ column of $X$. In dictionary learning and sparse coding, $X$ is called the learned dictionary with coefficients $Y$. In this formulation, the sparsity on $X$ is strictly enforced using the non-convex $\ell_0$ constraint, restricting 75% of the entries to be 0.

[7] We have ignored extraneous constants in the final three terms for clarity.
[8] The implementations are available at https://junqitang.com/

**Blind Image-Deblurring:** Let $Z$ be a blurred image. The problem of blind deconvolution reads:

$$\min_{X,Y} \|Z - X \odot Y\|_F^2 + \lambda \sum\nolimits_{r=1}^{2d} \Phi([D(X)]_r) \quad \text{s.t.} \quad 0 \le X \le 1,\ 0 \le Y \le 1,\ \|Y\|_1 \le 1, \tag{5.2}$$

where $\odot$ is the 2D convolution operator, $X$ is the image to recover, and $Y$ is the blur-kernel to estimate. We choose a classic smooth edge-preserving regularizer in the image domain, with $D(\cdot)$ being the 2D differential operator computing the horizontal and vertical gradients for each pixel. For the potential function $\Phi(\cdot)$, we choose $\Phi(v) := \log(1+\theta v^2)$ as in [29]. This potential function promotes sparsity in image gradients, hence yielding sharp images. We choose $\theta = 10^3$ and $\lambda = 5 \times 10^{-5}$ in our experiments

One of the benefits of SPRING and PALM is that the two step-sizes, $\gamma_{X,k}$ and $\gamma_{Y,k}$, depend separately on the Lipschitz constants $\hat{L}_X(Y_k)$ and $\hat{L}_Y(X_k)$. The practical performance of these algorithms depends significantly on the step-size choices. The following section describes how we use adaptive step-sizes in our experiments.

## 5.1. Parameter choices and on-the-fly estimation of Lipschitz constants.

The global Lipschitz constants of the partial gradients of $F$ are usually unknown and difficult to estimate. In practice, adaptive step-size choices based on estimating local Lipschitz constants are needed for PALM and inertial PALM [29]. In our experiments, we use the power method to estimate the Lipschitz constants on-the-fly in every iteration of the compared algorithms. For SPRING-SGD, SPRING-SAGA, and SPRING-SARAH, we find that it is sufficient to randomly sub-sample a mini-batch and run 5 iterations of the power method to get an estimate of the Lipschitz constants of the stochastic gradients. For PALM, we run 5 iterations of the power method in each iteration on the full batch to get an estimate of the Lipschitz constants of the full partial gradients.

For example, consider estimating the Lipschitz constants of the gradients corresponding to the objective function of Sparse-NMF (5.1). Let $X_k$ and $Y_k$ be the updates of $k$-th iteration, then $L_Y(X_k) = \|X_k\|^2$, which is the largest squared singular value of $X_k$, and can be computed via power iteration:

$$v_i = \frac{X_k^T(X_k v_{i-1})}{\|X_k^T(X_k v_{i-1})\|_2},$$

with a random initialization satisfying $\|v_0\|_2 = 1$. We find that using 5 iterations is sufficient to provide good estimates, so we approximate $L_Y(X_k)$ by $\|X_k^T(X_k v_5)\|_2$. We use the same strategy for $L_X(Y_k)$.

Denote the estimated Lipschitz constants of the full gradients as $\hat{L}_X(Y_k)$ and $\hat{L}_Y(X_k)$, and denote the estimated Lipschitz constants of the stochastic estimates as $\tilde{L}_X(Y_k)$ and $\tilde{L}_Y(X_k)$. We set the step-sizes of the compared algorithms as follows:

- **PALM:** $\gamma_{X,k} = \frac{1}{\hat{L}_X(Y_k)}$ and $\gamma_{Y,k} = \frac{1}{\hat{L}_Y(X_k)}$ (these are the standard step-sizes [6]).
- **Inertial PALM:** $\gamma_{X,k} = \frac{0.9}{\hat{L}_X(Y_k)}$, $\gamma_{Y,k} = \frac{0.9}{\hat{L}_Y(X_k)}$, and we set the momentum parameter to $\frac{k-1}{k+2}$, where $k$ denotes the number of iterations. Pock and Sabach [29] assert that this dynamic momentum parameter achieves the best practical performance.[9]

[9]The dynamic choice of momentum parameter is not theoretically analyzed by Pock and Sabach [29], but it appears to be superior to the constant inertial parameter choice. Pock and Sabach suggest the aggressive step-sizes $\gamma_{X,k} = \frac{1}{\hat{L}_X(Y_k)}$ and $\gamma_{Y,k} = \frac{1}{\hat{L}_Y(X_k)}$ for the dynamic scheme, but we find these choices sometimes lead to unstable/divergent behavior in the late iterations. Hence, we use the slightly smaller step-sizes $\gamma_{X,k} = \frac{0.9}{\hat{L}_X(Y_k)}$ and $\gamma_{Y,k} = \frac{0.9}{\hat{L}_Y(X_k)}$ instead. These choices ensure the algorithm is stable, and we observe that they do not compromise the convergence rate in practice.

- **SPRING-SGD:** $\gamma_{X,k} = \frac{1}{\sqrt{\lceil kb/n \rceil}\tilde{L}_X(Y_k)}$ and $\gamma_{Y,k} = \frac{1}{\sqrt{\lceil kb/n \rceil}\tilde{L}_Y(X_k)}$. It is well-known in the literature that a shrinking step-size is necessary for SGD to converge to a critical point [7, 22, 26, 39].
- **SPRING-SAGA:** $\gamma_{X,k} = \frac{1}{3\tilde{L}_X(Y_k)}$ and $\gamma_{Y,k} = \frac{1}{3\tilde{L}_Y(X_k)}$.
- **SPRING-SARAH:** $\gamma_{X,k} = \frac{1}{2\tilde{L}_X(Y_k)}$ and $\gamma_{Y,k} = \frac{1}{2\tilde{L}_Y(X_k)}$.

*Remark* 5.1 (Practical step-sizes for SPRING-SAGA and SPRING-SARAH). While the step-sizes suggested in Sections 3 and 4 lead to state-of-the-art convergence rates for (1.1), we observe that those step-size choices are conservative for SPRING-SAGA and SPRING-SARAH in practice. Hence, we adopt the suggested step-size choices in the original works with scale factors $1/3$ for SAGA [16, Section 2] and $1/2$ for SARAH [27, Corollary 3]. For all tested methods, the step-sizes we use are optimal in practice while ensuring convergence in all experiments with extensive tests.

The same random initialization is used for all of the compared algorithms in our Sparse-NMF experiments, while for BID we initialize the image estimate with the blurred image and the kernel estimate with all ones. We observe that SPRING with variance-reduced gradients can be sensitive to poor initialization, and this may initially compromise convergence. However, this initialization issue can be effectively resolved if we use plain stochastic gradient without variance-reduction in the first epoch of SPRING-SARAH/SPRING-SAGA as a warm-start, which is suggested in [23].

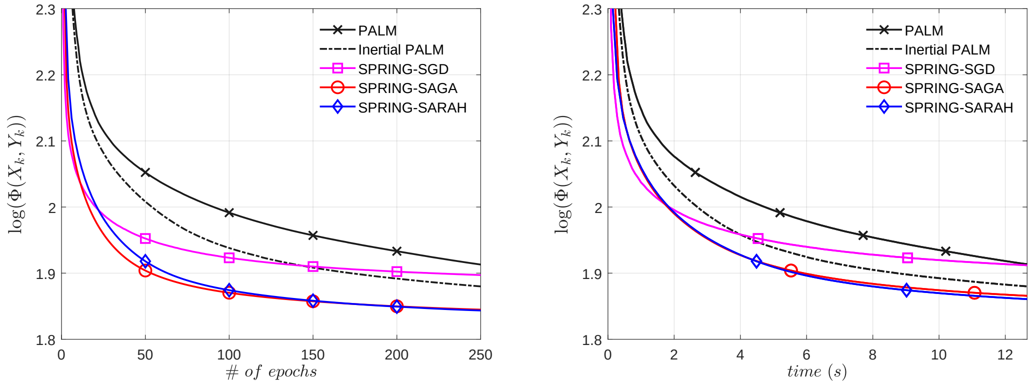
In all the convergence plots for our experiments, we report the average results for stochastic methods with 10 independent runs. We comment here that from our numerical observations, the final objective values achieved by the stochastic algorithms vary very little from the average.

### 5.2. Sparse-NMF.

We consider the extended Yale-B dataset and ORL dataset, which are standard facial recognition benchmarks consisting of human face images.[10] The extended Yale-B dataset contains 2414 cropped images of size $32 \times 32$, while the ORL dataset contains 400 images sized $64 \times 64$. In the experiment for Yale dataset, we extract 49 sparse basis-images for the dataset. For ORL dataset we extract 25 sparse basis-images. In each iteration of the stochastic algorithms, we randomly sub-sample $5\%$ of the full batch as a mini-batch. Here for SPRING-SARAH we set $p = \frac{1}{20}$. To reflect the effect of the algorithmic randomness within our methods, we report the average results (over 10 independent runs) of objective values in Figure 1. Meanwhile we also report the variance of the objective value at termination in Table 1. The obtained results are shown in Figures 1 and 2, from which we observe:

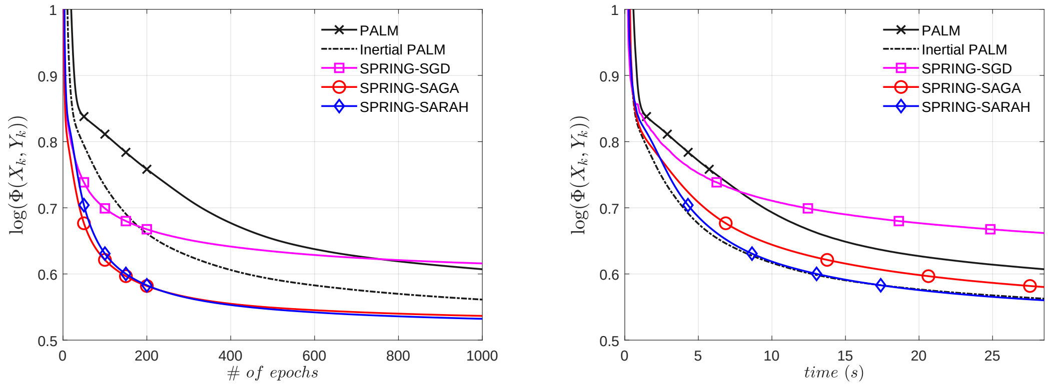
- Overall, SPRING using SAGA and SARAH estimators achieves superior performance compared to PALM, inertial PALM, and SPRING using the vanilla SGD gradient estimator.
- PALM has the worst performance in the considered Sparse-NMF tasks, which is not surprising since PALM is the baseline method in this comparison. Incorporating inertia can offer considerable acceleration for PALM. We believe that such inertial schemes can also be extended to accelerate SPRING and leave it as an important direction of future research (see [19] for some work in this direction).
- SPRING using the vanilla SGD gradient estimator achieves fast convergence initially, but gradually slows its convergence due to the shrinking step-size that is necessary to combat the non-reducing variance. However, using variance-reduced gradient estimators SAGA and SARAH, SPRING is able to overcome this issue and achieve the best overall convergence rates.

[10] Preprocessed versions [8, 9] can be found in: http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html




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

Figure 1: Objective decrease comparison of Sparse-NMF on Yale dataset.




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

Figure 2: Objective decrease comparison of Sparse-NMF on ORL dataset.

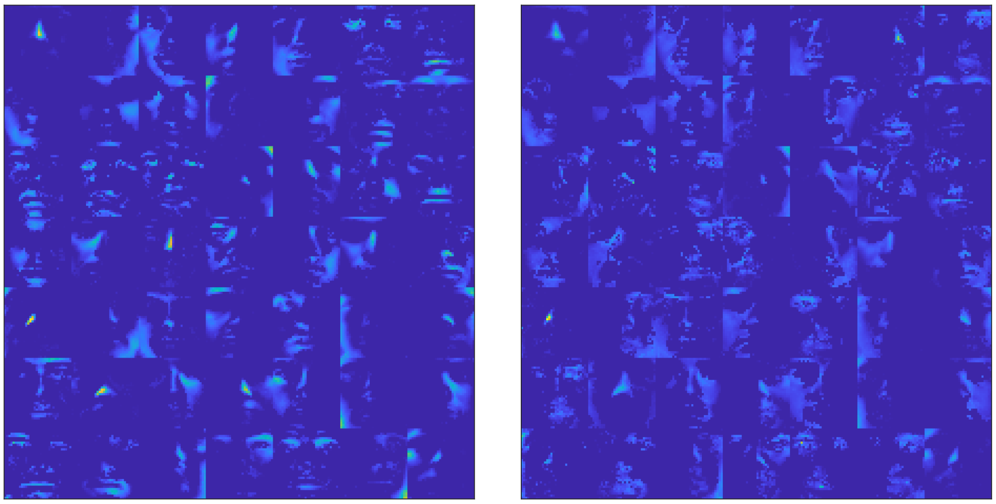
*Remark* 5.2 (Computational overheads for stochastic gradient methods). While the epoch count metric measures the gradient complexities of the algorithms, it does not reflect the computation overheads of the stochastic algorithms. The most important overhead for stochastic gradient methods in our setting would be the multiple calls to the proximal operator [35, 36]. Even though the proximal operators in our settings are not computationally expensive, computing such an operation many times still accumulates to a non-negligible overhead. Although our epoch count results confirm the complexity advantage predicted by theory, we can only observe compromised benefits from the wall-clock time comparison.

*Remark* 5.3 (The effect of algorithmic randomness). In order to reflect the algorithmic randomness of our stochastic methods, we present in log-scale the averaged convergence curves over 10 independent runs (in Figure 1 and 2). We also report that the variation of these results are virtually negligible, as we show in Table 1. The variances of the objective values at termination ($250^{th}$ epoch for Yale dataset, and $1000^{th}$ epoch for ORL dataset) in the same log-scale are very small.

Table 1: The variation of the objective value (log-scale) at termination for randomized methods

| Dataset/Algorithm | SPRING-SGD | SPRING-SAGA | SPRING-SARAH |
|---|---|---|---|
| Yale | $1.8711 \times 10^{-5}$ | $7.5532 \times 10^{-6}$ | $8.3064 \times 10^{-6}$ |
| ORL | $9.9082 \times 10^{-5}$ | $1.6723 \times 10^{-5}$ | $1.2961 \times 10^{-5}$ |

As a visual illustration we also present in Figure 3 the basis images generated by SPRING-SAGA and PALM for the Yale dataset at the $50^{th}$ epoch. It is clear that the basis images generated by SPRING-SAGA appear natural and smooth quickly at an early epoch, while PALM's results at the same epoch appear noisy and distorted.

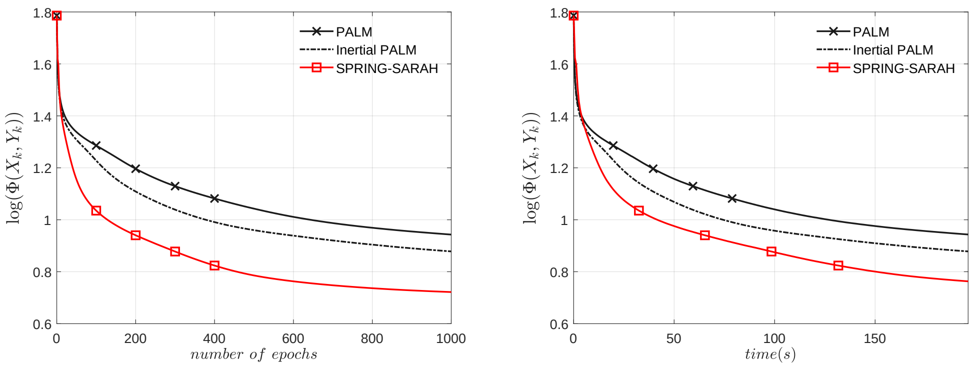


(a) SPRING-SAGA (b) PALM

Figure 3: Basis images from the Sparse-NMF experiment generated by SPRING-SAGA and PALM on the $50^{th}$ epoch for the Yale dataset.

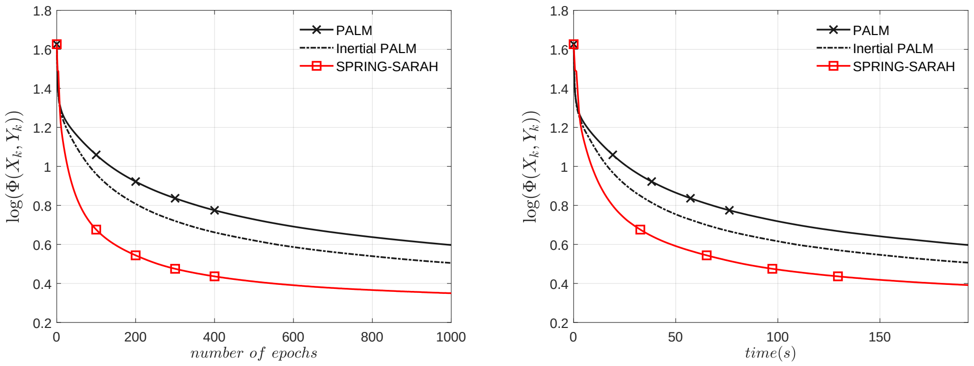
**5.3. Blind Image-Deblurring.** For blind image-deconvolution, we choose to compare SPRING-SARAH, PALM and inertial PALM. We use two images, *Kodim08* and *Kodim15*, of size $256 \times 256$ for testing. For each image, two blur kernels—linear motion blur and out-of-focus blur—are considered with additional additive Gaussian noise. For SPRING-SARAH, the mini-batch size is $1/64$ of the full batch (and also we set $p = \frac{1}{64}$). For this mini-batch size, we choose smaller step sizes $\gamma_{X,k} = \frac{1}{8\tilde{L}_X(Y_k)}$, $\gamma_{Y,k} = \frac{1}{3\tilde{L}_Y(X_k)}$ than the default choices to encourage stability. As above, we present results of SPRING in terms of an average of 10 independent runs in Figures 6 and 7, and we report that the variance due to the algorithmic randomness evaluated at termination is also negligible (on the order of $10^{-6}$).

For both images with motion blur, the convergence comparisons of the algorithms are provided in Figures 4 and 5, from which we observe SPRING-SARAH is faster than the other two methods in both cases. Figures 6 and 7 provide comparisons of the recovered image and blur kernel. We observe superior performance of SPRING-SARAH over PALM in these figures as well. In particular, when comparing




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

Figure 4: Objective decrease comparison (epoch counts) of blind image-deconvolution experiment on Kodim08 image using an $11 \times 11$ motion-blur kernel.



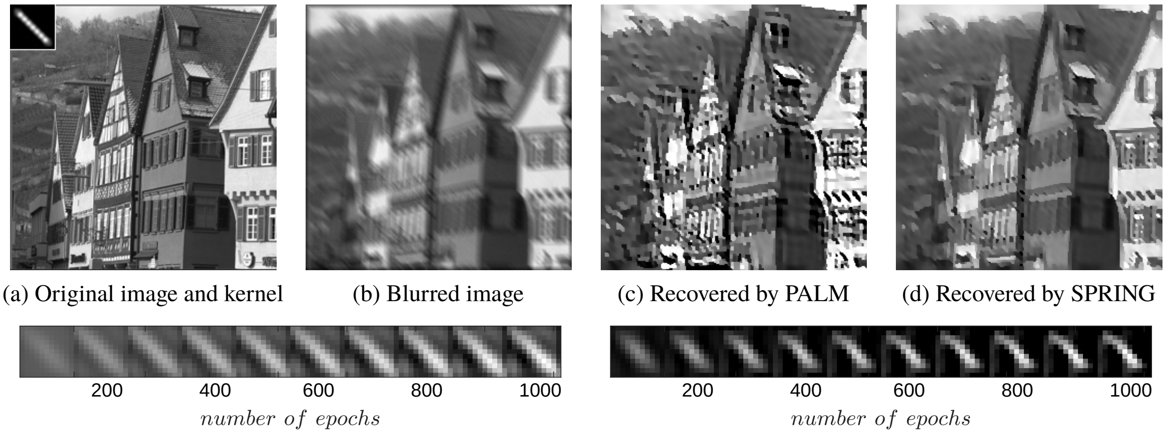

(a) `Epoch counts comparison` (b) `Wall-clock time comparison`

Figure 5: Objective decrease comparison (wall-clock time) of blind image-deconvolution experiment on Kodim15 image using an $11 \times 11$ motion-blur kernel.

the estimated blur kernels of the two algorithms every 100 epochs, we clearly see that SPRING-SARAH more quickly recovers more accurate solutions than PALM. It is worth noting that, although stochastic gradient methods have been shown to be inherently inefficient for non-blind and non-uniform deblurring task where the blur kernels are known or estimated beforehand [36], SPRING still offers significant acceleration over PALM in blind-deblurring tasks. Additional experiments using out-of-focus blur kernels are provided in the appendix.

## 6. Conclusion.

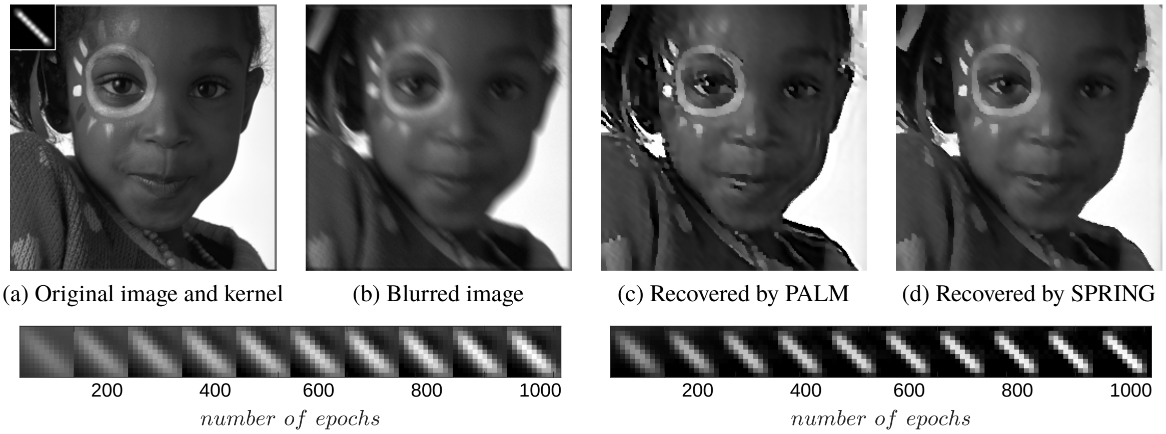
We propose SPRING, a stochastic extension of the PALM algorithm for solving a class of structured non-smooth and non-convex optimization problems. We analyze the convergence properties of SPRING when using a variety of variance-reduced gradient estimators, and we prove specific convergence rates using the SAGA and SARAH estimators. For generic optimization problems of the form (1.1), we show that SPRING-SAGA (with $b \leq \mathcal{O}(n^{2/3})$) and SPRING-SARAH return an




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

(e) Estimated kernel by PALM (f) Estimated kernel by SPRING

Figure 6: Image and kernel reconstructions from the blind image-deconvolution experiment on the Kodim08 image using an $11 \times 11$ motion blur kernel.




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

(e) Estimated kernel by PALM (f) Estimated kernel by SPRING

Figure 7: Image and kernel reconstructions from the blind image-deconvolution experiment on the Kodim15 image using an $11 \times 11$ motion blur kernel.

$\epsilon$-approximate critical point in expectation in no more than $O(\frac{n^2 L}{b^3 \epsilon^2})$ and $O(\frac{\sqrt{n} L}{\epsilon^2})$ SFO calls, respectively, showing that SPRING-SARAH achieves the complexity lower bound for stochastic non-convex optimization. For objectives satisfying an error bound, we further demonstrate that our methods converge linearly to the global optimum. Because of the generality of our results, they contain almost all existing results for stochastic non-convex optimization as special cases, and they improve on them in many settings. Most importantly, we extend the full convergence theory of PALM to the stochastic setting, showing that SPRING achieves the same convergence rates as PALM on semialgebraic objectives.

**Acknowledgements.** JT and MD acknowledge support from the ERC Advanced grant, project 694888, C-SENSE. JL acknowledges support from the Leverhulme Trust. CBS acknowledges support from the Leverhulme Trust project on Breaking the Non-Convexity Barrier, and on Unveiling the Invisible, the Philip Leverhulme Prize, the EPSRC grant No. EP/S026045/1, EPSRC grant No. EP/M00483X/1, and EPSRC Centre No. EP/N014588/1, the European Union Horizon 2020 research and innovation programmes under the Marie Skłodowska-Curie grant agreement No. 691070 CHiPS and the Marie Skłodowska-Curie grant agreement No 777826, the Cantab Capital Institute for the Mathematics of Information, and the Alan Turing Institute.

## REFERENCES

[1] Aravkin, A., and Davis, D. Trimmed statistical estimation via variance reduction. *Mathematics of Operations Research* (2019).

[2] Attouch, H., and Bolte, J. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming B 116*, 1 (2007), 5–16.

[3] Attouch, H., Bolte, J., Redont, P., and Soubeyran, A. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Lojasiewicz inequality. *Mathematics of Operations Research 35*, 2 (2010), 438–457.

[4] Bolte, J., Daniilidis, A., and Lewis, A. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization 17*, 4 (2007), 1205–1223.

[5] Bolte, J., Daniilidis, A., Ley, O., and Mazet, L. Characterizations of Lojasiewicz inequalities: subgradient flows, talweg, convexity. *Transactions of the American Mathematical Society 362*, 6 (2010), 3319–3363.

[6] Bolte, J., Sabach, S., and Teboulle, M. Proximal alternating linearised minimization for nonconvex and nonsmooth problems. *Mathematical Programming 146*, 1-2 (2014), 459–494.

[7] Bottou, L. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT 2010*. Springer, 2010, pp. 177–186.

[8] Cai, D., He, X., and Han, J. Spectral regression for efficient regularized subspace learning. In *IEEE 11th International Conference on Computer Vision* (2007), IEEE, pp. 1–8.

[9] Cai, D., He, X., Hu, Y., Han, J., and Huang, T. Learning a spatially smooth subspace for face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition* (2007), IEEE, pp. 1–7.

[10] Campisi, P., and Egiazarian, K. *Blind image deconvolution: theory and applications*. CRC press, 2016.

[11] Candès, E. J., Li, X., Ma, Y., and Wright, J. Robust principal component analysis? *Journal of the ACM (JACM) 58*, 3 (2011), 11.

[12] Chambolle, A., Ehrhardt, M. J., Richtárik, P., and Schönlieb, C.-B. Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging applications. *SIAM J. Optim. 28*, 4 (2018), 2783–2808.

[13] d'Aspremont, A., Ghaoui, L. E., Jordan, M. I., and Lanckriet, G. R. A direct formulation for sparse pca using semidefinite programming. In *Advances in neural information processing systems* (2005), pp. 41–48.

[14] Davis, D. The asynchronous palm algorithm for nonsmooth nonconvex problems. *arXiv:1604.00526* (2016).

[15] Davis, D., Edmunds, B., and Udell, M. The sound of APALM clapping: Faster nonsmooth nonconvex optimization with stochastic asynchronous palm. In *Advances in Neural Information Processing Systems* (2016), pp. 226–234.

[16] Defazio, A., Bach, F., and Lacoste-Julien, S. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems* (2014), pp. 1646–1654.

[17] Driggs, D., Tang, J., Liang, J., Davies, M., and Schönlieb, C. Spring: A fast stochastic proximal alternating method for non-smooth non-convex optimization. *arXiv preprint arXiv:2002.12266* (2020).

[18] Fang, C., Li, C. J., Lin, Z., and Zhang, T. Spider: Near-optimal non-convex optimization via stochastic path integrated differential estimator. In $32^{nd}$ *Conference on Neural Information Processing Systems* (2018).

[19] Hertrich, J., and Steidl, G. Inertial stochastic palm and its application for learning student-t mixture models. *arXiv:2005.02204* (2020).

[20] Hoyer, P. O. Non-negative matrix factorization with sparseness constraints. *Journal of machine learning research 5*, Nov (2004), 1457–1469.

[21] Johnson, R., and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems* (2013), pp. 315–323.

[22] Konečný, J., Liu, J., Richtárik, P., and Takáč, M. Mini-batch semi-stochastic gradient descent in the proximal setting. *IEEE Journal of Selected Topics in Signal Processing 10*, 2 (2015), 242–255.

[23] Konečnỳ, J., and Richtárik, P. Semi-stochastic gradient descent methods. *Frontiers in Applied Mathematics and Statistics 3* (2017), 9.
[24] Li, B., Ma, M., and Giannakis, G. B. On the convergence of sarah and beyond. *arXiv:1906.02351v2* (2020).
[25] Li, G., and Pong, T. K. Calculus of the exponent of kurdyka–Łojasiewicz inequality and its applications to linear convergence of first-order methods. *Foundations of Computational Mathematics 18* (2018), 1199–1232.
[26] Moulines, E., and Bach, F. R. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems* (2011), pp. 451–459.
[27] Nguyen, L. M., Liu, J., Scheinberg, K., and Takáĉ, M. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *Proceedings of the 34th International Conference on Machine Learning* (2017), vol. 70, pp. 2613–2621.
[28] Pham, N. H., Nguyen, L. M., Phan, D. T., and Tran-Dinh, Q. ProxSARAH: An efficient algorithmic framework for stochastic composite nonconvex optimization. *arXiv:1902.05679* (2019).
[29] Pock, T., and Sabach, S. Inertial proximal alternating linearized minimization (iPALM) for nonconvex and nonsmooth problems. *SIAM Journal on Imaging Sciences 9*, 4 (2016), 1756–1787.
[30] Reddi, S. J., Hefny, A., Sra, S., Póczos, B., and Smola, A. Stochastic variance reduction for nonconvex optimization. In *Proc. 33rd International Conference on Machine Learning* (2016).
[31] Reddi, S. J., Sra, S., Póczos, B., and Smola, A. Fast stochastic methods for nonsmooth nonconvex optimization. In *Proc. 30th Annual Conference on Neural Information Processing Systems* (2016).
[32] Robbins, H., and Monro, S. A stochastic approximation method. *Annals of Mathematical Statistics 22*, 3 (1951), 400–407.
[33] Robbins, H., and Siegmund, D. A convergence theorem for non-negative almost supermartingales and some applications. *Optimizing Methods in Statistics* (1971), 233–257.
[34] Schmidt, M., Roux, N. L., and Bach, F. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming 162* (2017), 83–112.
[35] Tang, J., Egiazarian, K., and Davies, M. The limitation and practical acceleration of stochastic gradient algorithms in inverse problems. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2019), IEEE, pp. 7680–7684.
[36] Tang, J., Egiazarian, K., Golbabaee, M., and Davies, M. The practicality of stochastic optimization in imaging inverse problems. *IEEE Transactions on Computational Imaging 6* (2020), 1471–1485.
[37] Wang, Z., Ji, K., Zhou, Y., Liang, Y., and Tarokh, V. SpiderBoost: A class of faster variance-reduced algorithms for nonconvex optimization. *arXiv:1810.10690* (2018).
[38] Xiao, L., and Zhang, T. A proximal stochastic gradient method with progressive variance reduction. *Technical report, Microsoft Research* (2014).
[39] Xu, Y., and Yin, W. Block stochastic gradient iteration for convex and nonconvex optimization. *SIAM Journal on Optimization 25*, 3 (2015), 1686–1716.
[40] Zhou, D., and Gu, Q. Lower bounds for smooth nonconvex finite-sum optimization. *arXiv preprint arXiv:1901.11224* (2019).
[41] Zhou, Y., Wang, Z., Ji, K., Liang, Y., and Tarokh, V. Momentum schemes with stochastic variance reduction for nonconvex composite optimization. *arXiv:1902.02715* (2019).
[42] Zou, H., Hastie, T., and Tibshirani, R. Sparse principal component analysis. *Journal of computational and graphical statistics 15*, 2 (2006), 265–286.

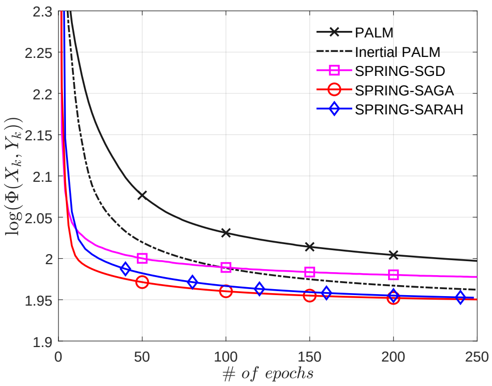
## Appendix A. Additional numerical experiments.

This section contains additional numerical experiments demonstrating the superiority of SPRING over PALM.

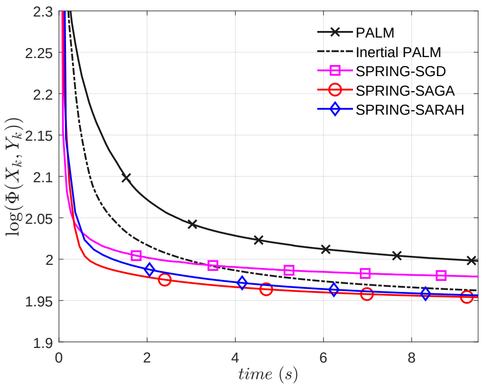
We first present our additional results on the Sparse-PCA example for the Yale and ORL datasets. The problem of Sparse-PCA with $r$ principal components can be written as:

$$\min_{X,Y} \|A - XY\|_F^2 + \lambda_1\|X\|_1 + \lambda_2\|Y\|_1. \tag{A.1}$$

where $X \in \mathbb{R}^{n\times r}, Y \in \mathbb{R}^{r\times d}$. We use $\ell_1$ regularization on both $X$ and $Y$ to promote sparsity with $\lambda_1 = 10^{-3}$ and $\lambda_2 = 5\times 10^{-3}$, and $r = 25$. We compare SPRING-SAGA, SPRING-SARAH, SPRING-SGD and PALM. We choose the mini-batch size to be $\frac{1}{40}$ of the full batch (for SPRING-SARAH we set $p = \frac{1}{40}$). We report the results of 10 independent runs of the stochastic methods in Figure 8 and 9, and we denote that the variance due to the algorithmic randomness evaluated at termination is also negligible (on the order of $10^{-5}$). Similar to what we observe in the Sparse-NMF experiments, our results in Figure 8 and 9 show that SPRING with stochastic variance-reduced gradient estimators achieves the fastest convergence.



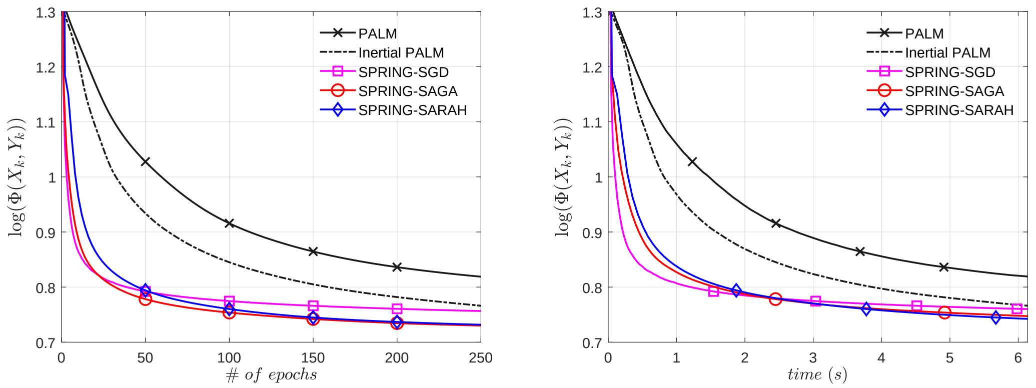

(a) `Epoch counts comparison`




(b) `Wall-clock time comparison`

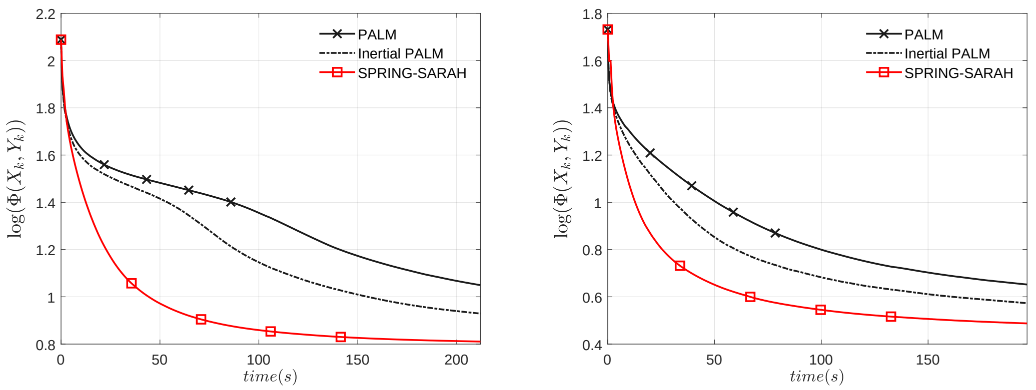
Figure 8: Objective decrease comparison of Sparse-PCA on Yale dataset.

Figures 10 to 12 show additional comparisons for blind image-deblurring where the images are blurred with an out-of-focus kernel. We choose the regularization parameter $\lambda = 1 \times 10^{-4}$ and the other settings are the same for the BID experiments presented in the main text. Again, we observe that our SPRING-SARAH algorithm outperforms PALM and inertial-PALM.



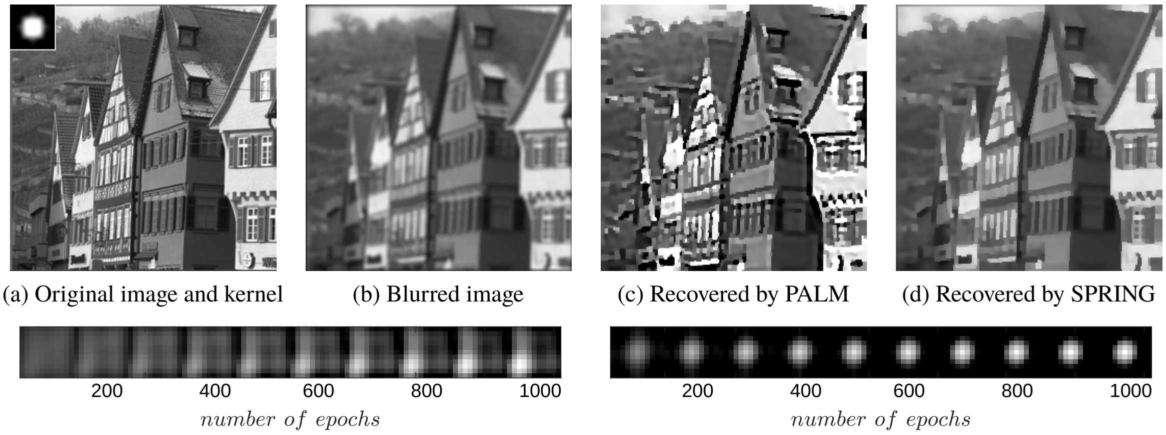

(a) `Epoch counts comparison` (b) `Wall-clock time comparison`

Figure 9: Objective decrease comparison of Sparse-PCA on ORL dataset.




(a) `Kodim 08` (b) `Kodim 15`

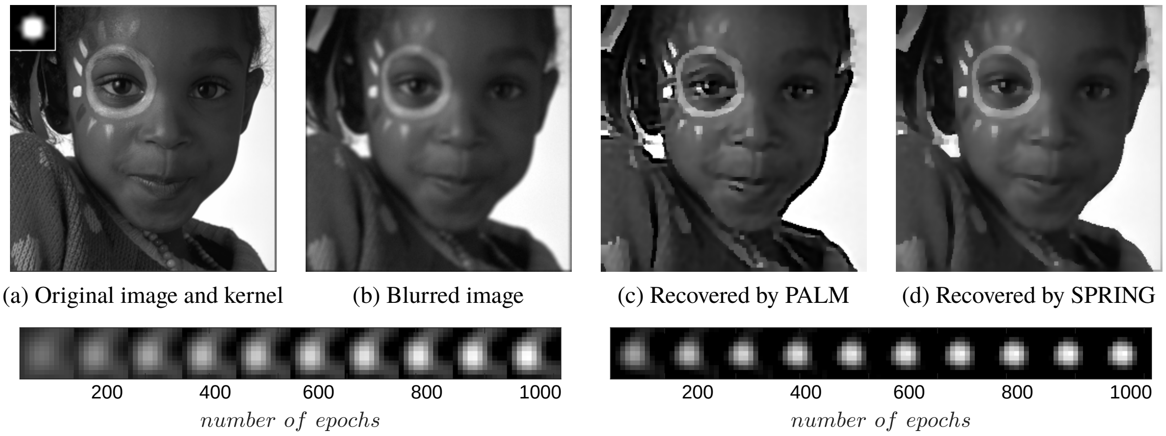
Figure 10: Objective decrease comparison (versus run time) of blind image-deconvolution experiment on Kodim08 and Kodim15 images using an out-of-focus blur kernel.




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

(e) Estimated kernel by PALM (f) Estimated kernel by SPRING

Figure 11: Image and kernel reconstructions from the blind image-deconvolution experiment on the Kodim08 image using an out-of-focus blur kernel.




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

(e) Estimated kernel by PALM (f) Estimated kernel by SPRING

Figure 12: Image and kernel reconstructions from the blind image-deconvolution experiment on the Kodim15 image using an out-of-focus blur kernel.