SPRING: A fast stochastic proximal alternating method for non-smooth non-convex optimization

Anonymous Authors[1]

Abstract

We propose novel stochastic proximal alternating linearized minimization (PALM) algorithms for solving a class of non-smooth and non-convex optimization problems which arise in many statistical machine learning, computer vision, and imaging applications. We provide a theoretical analysis, showing that our proposed method with variance-reduced stochastic gradient estimators such as SAGA and SARAH achieves state-of-the-art oracle complexities. We also demonstrate the efficiency of our algorithm via numerical experiments including sparse non-negative matrix factorization, sparse principal component analysis, and blind image deconvolution.

todo: add the papers mentioned by Gabriele.

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 challenge the optimizer to minimize the average of a loss function measuring the fit between observed data, x, and a model's predicted result, b: $\min_{x\in\mathbb{R}^{m_1}} 1/n\sum_{i=1}^{n}\mathcal{L}(x_i, b_i)$. The two defining qualities of these problems are their large scale (in many applications, n is of the order of billions), and their finite-sum structure, representing the challenge of these problems and its solution, respectively.

When n is large in ($\mathcal{P}$), computing the gradient of the objective is often prohibitively expensive, rendering most traditional first-order optimization algorithms ineffective. Randomized optimization algorithms (Robbins & Monro, 1951; Bottou, 2010) replace the full gradient with a random estimate that is cheap to compute, so 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 (Defazio et al., 2014; Johnson & Zhang, 2013; Xiao & Zhang, 2014).

Objectives with a finite-sum structure arise in image processing and computer vision applications as well. Recently, randomized optimization algorithms have been explored for image processing tasks including PET reconstruction, deblurring, and tomography (Chambolle et al., 2018; Tang et al., 2019). As randomized optimization expands into new applications, it moves further from the smooth, strongly convex, finite-sum objectives where it is well-understood theoretically. This work offers a better understanding of randomized optimization for objectives that are neither smooth nor convex.

1.1. Non-smooth, non-convex optimization

Our goal is to minimize composite objectives of the form

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

where $F(x,y) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} F_i(x,y)$. The functions J and R are regularizers promoting low-complexity structures such as sparsity or non-negativity in the solution. 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 on J, R, and F:

- **(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), and bounded from below;
- **(A.2)** $F : \mathbb{R}^{m_1}\times\mathbb{R}^{m_2} \to \mathbb{R}$ is finite-valued, differentiable, and its gradient ∇F is M-Lipschitz continuous on bounded sets of $\mathbb{R}^{m_1}\times\mathbb{R}^{m_2}$;
- **(A.3)** The partial gradient $\nabla_x F$ is Lipschitz continuous with modulus $L_1(y)$, and $\nabla_y F$ is Lipschitz continuous with modulus $L_2(x)$;
- **(A.4)** The function Φ is bounded from below.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Throughout, no convexity is imposed on any of the functions involved. The model in ($\mathcal{P}$) departs from the popular 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 ($\mathcal{P}$). Archetypal examples include non-negative or sparse matrix factorization (Hoyer, 2004), sparse PCA (d'Aspremont et al., 2005; Zou et al., 2006), robust PCA (Candès et al., 2011), trimmed least-squares (Aravkin & Davis, 2019), and blind image deconvolution (Campisi & Egiazarian, 2016). Despite the prevalence of these problems, there are only a few methods that can be generically applied to solve ($\mathcal{P}$).

Proximal alternating minimization (Attouch et al., 2010) In (Attouch et al., 2010), the authors propose the Proximal Alternating Minimization (PAM) method for solving ($\mathcal{P}$), which is defined by the following procedure:

$$
\begin{aligned}
&x_{k+1} \in \operatorname{argmin}_{x\in\mathbb{R}^{m_1}} \big\{\Phi(x, y_k) + \frac{1}{2\gamma_{x,k}}\|x - x_k\|^2\big\},\\
&y_{k+1} \in \operatorname{argmin}_{y\in\mathbb{R}^{m_2}} \big\{\Phi(x_{k+1}, y) + \frac{1}{2\gamma_{y,k}}\|y - y_k\|^2\big\},
\end{aligned} \tag{1.1}
$$

where $\gamma_{x,k}, \gamma_{y,k} > 0$ are step-sizes. The inner iterations of PAM require the global minimization of a non-convex objective, which is computationally intractable for large-scale problems. PAM can only produce a high-quality solution in the special case that the two subproblems above are convex, which is not the case in many applications.

Proximal alternating linearized minimization (Bolte et al., 2014) In (Bolte et al., 2014), the authors propose a linearized version of PAM: the Proximal Alternating Linearised Minimization (PALM) algorithm. PALM follows the procedure

$$
\begin{aligned}
x_{k+1} &\in \operatorname{prox}_{\gamma_{x,k} J}\big(x_k - \gamma_{x,k}\nabla_x F(x_k, y_k)\big),\\
y_{k+1} &\in \operatorname{prox}_{\gamma_{y,k} R}\big(y_k - \gamma_{y,k}\nabla_y F(x_{k+1}, y_k)\big),
\end{aligned} \tag{1.2}
$$

where $\nabla_x F$ and $\nabla_y F$ are partial derivatives, and the proximal operator is defined as

$$
\operatorname{prox}_{\eta f}(w) \stackrel{\text{def}}{=} \operatorname{argmin}_x \big\{f(x) + \frac{1}{2\eta}\|x - w\|^2\big\}. \tag{1.3}
$$

In contrast to PAM, each subproblem of PALM is efficiently computable with access to the proximal maps of J and R, and these are accessible in many applications. PALM also has the same convergence rates as PAM, so linearizing F in each proximal step offers significant improvement over PAM. Improving performance further, Pock & Sabach (2016) introduce an inertial variant of PALM with an additional momentum step. Although Pock & Sabach (2016) do not show improved theoretical convergence rates over PALM, they show that inertia often improves PALM's practical performance.

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 (1.2) 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 $\nabla_x F_i(x_k, y_k)$ and $\nabla_y F_i(x_{k+1}, y_k)$ for $i \in B_k \subset \{1, 2, \cdots, n\}$. The mini-batch B_k is chosen uniformly at random from all subsets of $\{1, 2, \cdots, n\}$ with cardinality b. SPRING is outlined in Algorithm 1.

Algorithm 1 SPRING: **S**tochastic **Pr**oximal Alternat**ing** Linearized Minimization

Initialize: $x_0 \in \mathbb{R}^{m_1}, y_0 \in \mathbb{R}^{m_2}$.
repeat

$$
\begin{aligned}
x_{k+1} &\in \operatorname{prox}_{\gamma_{x,k} J}\big(x_k - \gamma_{x,k}\widetilde{\nabla}_x(x_k, y_k)\big).\\
y_{k+1} &\in \operatorname{prox}_{\gamma_{y,k} R}\big(y_k - \gamma_{y,k}\widetilde{\nabla}_y(x_{k+1}, y_k)\big).
\end{aligned} \tag{1.4}
$$

$k = k + 1$
until *convergence*;

Many different gradient estimators can be used in SPRING; the simplest is the stochastic gradient descent (SGD) estimator:

$$
\widetilde{\nabla}_x^{\text{SGD}}(x_k, y_k) = \frac{1}{b}\sum_{j\in B_k} \nabla_x F_j(x_k, y_k). \tag{1.5}
$$

Another popular choice is the SAGA gradient estimator (Defazio et al., 2014), which incorporates the gradient history:

$$
\begin{aligned}
D_k &= \frac{1}{b}\big(\textstyle\sum_{j\in B_k} \nabla_x F_j(x_k, y_k) - g_{k,j}\big),\\
\widetilde{\nabla}_x^{\text{SAGA}}(x_k, y_k) &= D_k + M_k,\\
g_{k+1,i} &= \begin{cases} \nabla_x F_i(x_k, y_k) & \text{if } i \in B_k,\\ g_{k,i} & \text{o.w.}\end{cases}\\
M_{k+1} &= M_k + \frac{b}{n} D_k.
\end{aligned} \tag{1.6}
$$

The last estimator we specifically consider in this work is the SARAH estimator (Nguyen et al., 2017), $\widetilde{\nabla}_x^{\text{SARAH}}(x_k, y_k)$, which is equal to

$$
\begin{cases}
\frac{1}{b}\big(\sum\limits_{j\in B_k} \nabla_x F_j(x_k, y_k) - \nabla_x F_j(x_{k-1}, y_{k-1})\big) & \text{w.p. } \frac{1}{p}\\
\qquad\qquad + \widetilde{\nabla}_x^{\text{SARAH}}(x_{k-1}, y_{k-1}) & \\
\nabla_x F(x_k, y_k) & \text{o.w.}
\end{cases} \tag{1.7}
$$

Here, p is a tuning parameter that is generally set to $\mathcal{O}(n)$. Other popular estimators that we do not specifically consider include the SVRG estimator (Johnson & Zhang, 2013) and the SAG estimator (Schmidt et al., 2017).

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 one step of PALM.

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

$$\min_{\{x_i\}_{i=1}^{\ell}, x_i \in \mathbb{R}^{n_i}} \left\{ \frac{1}{n} \sum_{i=1}^{n} F_i(x_1, \cdots, x_\ell) + \sum_{t=1}^{\ell} R_t(x_t) \right\}, \tag{1.8}$$

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

1.3. Contributions

Our main contribution is to show that if the gradient estimators $\widetilde{\nabla}_x$ and $\widetilde{\nabla}_y$ satisfy a *variance-reduced* property (see Definition 2.1), then the convergence rates of SPRING match the convergence rates of PALM (Bolte et al., 2014). In particular, if Φ is a semialgebraic function with KL-exponent θ (see Section 2), then SPRING produces a sequence of iterates $z_k = (x_k, y_k)$ that converges in expectation to a critical point z^* at the following rates:

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

We also prove convergence with respect to a *generalized gradient map* at a rate that is independent of the KL-exponent. The generalized gradient map is defined as $\mathcal{G}_{\gamma_1,\gamma_2}(x_k, y_k) \stackrel{\text{def}}{=}$

$$\begin{pmatrix} 1/\gamma_1(x_k - \mathrm{prox}_{\gamma_1 J}(x_k - \gamma_1 \nabla_x F(x_k, y_k))) \\ 1/\gamma_2(y_k - \mathrm{prox}_{\gamma_2 R}(y_k - \gamma_2 \nabla_y F(x_{k+1}, y_k))) \end{pmatrix}, \tag{1.9}$$

where $\gamma_1, \gamma_2 > 0$ are step-sizes. We show that

$$\mathbb{E}[\mathrm{dist}(0, \mathcal{G}_{\frac{\gamma_{x,\alpha}}{2}, \frac{\gamma_{y,\alpha}}{2}}(z_\alpha))^2] \le \mathcal{O}\left(\frac{1}{k}\right), \tag{1.10}$$

where α is chosen uniformly at random from the set $\{1, 2, \cdots, k\}$. If Φ satisfies a certain error bound (see (3.1)), then SPRING converges linearly to the global optimum.

The constants appearing in these rates scale with the mean-squared error (MSE) of the gradient estimators. When using the SAGA gradient estimator, the iterates of SPRING satisfy

$$\mathbb{E}[\mathrm{dist}(0, \mathcal{G}_{\frac{\gamma_{x,\alpha}}{2}, \frac{\gamma_{y,\alpha}}{2}}(z_\alpha))^2] \le \mathcal{O}\left(\frac{n^2 L}{b^3 k}\right), \tag{1.11}$$

and for the SARAH gradient estimator, we prove a convergence rate of

$$\mathbb{E}[\mathrm{dist}(0, \mathcal{G}_{\frac{\gamma_{x,\alpha}}{2}, \frac{\gamma_{y,\alpha}}{2}}(z_\alpha))^2] \le \mathcal{O}\left(\frac{\sqrt{n} L}{k}\right). \tag{1.12}$$

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 (e.g. $\nabla_x F_i(x_k, y_k)$). To find an ϵ-approximate critical point (i.e., a point z satisfying $\mathbb{E}\mathrm{dist}(0, \mathcal{G}_{\gamma_1,\gamma_2}(z)) \le \epsilon$ for some γ_1 and γ_2), 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 the 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 utilize models that are special cases of ($\mathcal{P}$), our results for SPRING capture most existing work as special cases. In particular, in the case $R \equiv J \equiv 0$, our result recovers recent results showing that SARAH achieves the *oracle complexity lower-bound* for non-convex problems with a finite-sum structure (Fang et al., 2018; Zhou et al., 2019; Wang et al., 2018; Pham et al., 2019; Zhou & Gu, 2019).

1.4. Prior Art

SPRING offers several advantages over existing stochastic algorithms for non-smooth non-convex optimization. In (Reddi et al., 2016), the authors investigate proximal SAGA and SVRG for solving problems of the form ($\mathcal{P}$) when y is constant and J is convex. 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 ϵ-approximate critical point. Similarly, in (Aravkin & Davis, 2019), the authors introduce TSVRG, a stochastic algorithm based on the SVRG gradient estimator, for solving another special case of ($\mathcal{P}$). This work generalizes their results and improves them in many cases. Most importantly, we show that using the 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 semialgebraic objectives offer even sharper convergence rates.

In (Davis et al., 2016), the authors introduce SAPALM, an asynchronous version of PALM that allows stochastic noise in the computed gradients. The authors prove convergence rates that scale with the variance of the noise in the gradients, with their best complexity bound for finding an ϵ-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 Davis et al. (2016) 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.

2. Preliminaries

We use the following definitions and notation throughout the manuscript.

Variance Reduction For our analysis, we assume that the gradient estimator used in Algorithm 1 is *variance-reduced*, as defined below.

Definition 2.1. A gradient estimator $\widetilde{\nabla}$ is *variance-reduced* with constants $V_1, V_2, V_\Upsilon \geq 0$, 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 random vectors v_k^i such that

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

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

$$\begin{aligned}&\mathbb{E}_k[\|\widetilde{\nabla}_x(x_k,y_k) - \nabla_x F(x_k,y_k)\| \\ &\quad + \|\widetilde{\nabla}_y(x_{k+1},y_k) - \nabla_y F(x_{k+1},y_k)\|] \\ &\leq \Gamma_k + V_2(\mathbb{E}_k\|z_{k+1}-z_k\| + \|z_k - z_{k-1}\|).\end{aligned} \tag{2.2}$$

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

$$\begin{aligned}\mathbb{E}_k\Upsilon_{k+1} \leq &(1-\rho)\Upsilon_k + V_\Upsilon(\mathbb{E}_k\|z_{k+1}-z_k\|^2 \\ &+ \|z_k - z_{k-1}\|^2).\end{aligned} \tag{2.3}$$

3. (Convergence of Estimator): For all sequences $\{z_k\}_{k=0}^\infty$ satisfying $\lim_{k\to\infty}\mathbb{E}\|z_k - z_{k-1}\|^2 \to 0$, it follows that $\mathbb{E}\Upsilon_k \to 0$ and $\mathbb{E}\Gamma_k \to 0$.

Almost all popular stochastic gradient estimators satisfy this property; in this work, we specifically consider the SAGA and SARAH estimators.

Proposition 2.2. *The SAGA gradient estimator is variance-reduced with parameters* $V_1 = 6M^2/b$, $V_2 = \sqrt{6}M/\sqrt{b}$, $V_\Upsilon = \frac{115nL^2}{b^2}$, *and* $\rho = \frac{b}{2n}$. *The 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 slight generalization of existing variance bounds for these estimators. For completeness, we include a proof of Proposition 2.2 in the supplementary material (Section D for the SAGA estimator and Section E for the SARAH estimator).

Remark 2.3. Our convergence results allow Algorithm 1 to use any variance-reduced gradient estimator, and they even allow different estimators to be used to approximate ∇_x and ∇_y. In particular, it is possible to use different mini-batch sizes when approximating the two partial gradients.

Kurdyka–Łojasiewicz property Some of our results assume Φ satisfies the Kurdyka–Łojasiewicz property. Let $H : \mathbb{R}^{m_1} \to \mathbb{R} \cup \{+\infty\}$ be a proper lower semicontinuous function. For ϵ_1, ϵ_2 satisfying $-\infty < \epsilon_1 < \epsilon_2 < +\infty$, define the set $[\epsilon_1 < H < \epsilon_2] \overset{\text{def}}{=} \{x \in \mathbb{R}^{m_1} : \epsilon_1 < H(x) < \epsilon_2\}$.

Definition 2.4 (Kurdyka–Łojasiewicz)**.** 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) $\varphi(0) = 0$, φ 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:

$$\varphi'\big(H(x) - H(\bar{x})\big)\mathrm{dist}\big(0, \partial H(x)\big) \geq 1. \tag{2.4}$$

Proper functions which satisfy the Kurdyka–Łojasiewicz property at each point of $\mathrm{dom}(\partial H)$ are called KL functions.

Roughly speaking, KL functions become sharp up to reparameterization via φ, a *desingularizing function* for H. Typical KL functions include the class of semialgebraic functions, see (Bolte et al., 2007; 2010). For instance, the ℓ_0 pseudo-norm and the rank function are KL. Semialgebraic 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 (Bolte et al., 2007; 2014). For these functions, the KL inequality reads

$$(H(x) - H(\overline{x}))^\theta \leq C\|\zeta\| \quad \forall \zeta \in \partial H(x), \tag{2.5}$$

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

Notation We use $L_x \overset{\text{def}}{=} \max_{k\in\mathbb{N}} L_1(y_k)$ where y_k is an iterate of SPRING, and we define L_y analogously. We set $\bar{L} \overset{\text{def}}{=} \max\{L_x, L_y\}$, $\overline{\gamma}_k \overset{\text{def}}{=} \max\{\gamma_{x,k}, \gamma_{y,k}\}$, $\underline{\gamma}_k \overset{\text{def}}{=} \min\{\gamma_{x,k}, \gamma_{y,k}\}$, and $\underline{\Phi} \overset{\text{def}}{=} \inf_{(x,y)\in\mathrm{dom}(\Phi)} \Phi(x,y)$. We also use L to denote the maximum Lipschitz constant of F, J, and R over the domain of Φ, so that $L_1(y), L_2(x), M \leq L$ for all $(x,y) \in \mathrm{dom}(\Phi)$. We use $\mathbb{E}_k$ to denote the expectation conditional on the first k iterations of SPRING.

3. Main Results

We prove convergence rates of three types. Our first result holds for all functions satisfying assumptions (**A.1**) to (**A.4**) and shows that the norm of the gradient map decays like $\mathcal{O}(1/\sqrt{k})$. If Φ satisfies an additional global error bound

$$\Phi(x,y) - \underline{\Phi} \leq \mu\mathrm{dist}(0, \mathcal{G}_{\gamma_1,\gamma_2}(x,y))^2, \tag{3.1}$$

for all $(x,y) \in \mathrm{dom}(\Phi)$, then the suboptimality decays linearly. These two results generalize many existing convergence guarantees for stochastic gradient methods on non-convex, non-smooth objectives, including those in (Reddi et al., 2016; Fang et al., 2018; Zhou et al., 2019; Wang et al., 2018; Aravkin & Davis, 2019).

Theorem 3.1. *Let $\widetilde{\nabla}$ be a variance-reduced estimator. Suppose $\overline{\gamma}_k$ is non-increasing, and for all k,*

$$\overline{\gamma}_k \le \tfrac{1}{16}\sqrt{\tfrac{\bar{L}^2}{(V_1+V_\Upsilon/\rho)^2}+\tfrac{16}{(V_1+V_\Upsilon/\rho)}}-\tfrac{\bar{L}}{16(V_1+V_\Upsilon/\rho)},$$
$$0<\beta\le\underline{\gamma}_k, \quad \gamma_{x,k}<\tfrac{1}{4L_x}, \quad \text{and} \quad \gamma_{y,k}<\tfrac{1}{4L_y}. \tag{3.2}$$

With α chosen uniformly at random from the set $\{0,1,\cdots,T-1\}$, the generalized gradient at (x_α,y_α) after T iterations satisfies

$$\mathbb{E}[\operatorname{dist}(0,\mathcal{G}_{\frac{\gamma_{x,\alpha}}{2},\frac{\gamma_{y,\alpha}}{2}}(z_\alpha))^2]\le\frac{4(\Phi(x_0,y_0)+\frac{2\overline{\gamma}_0}{\rho}\Upsilon_0)}{T\nu\beta^2}. \tag{3.3}$$

Furthermore, if Φ satisfies the error bound (3.1) and

$$\overline{\gamma}_k\le\tfrac{1}{20}\sqrt{\tfrac{\bar{L}^2}{(V_1+V_\Upsilon/\rho)^2}+\tfrac{20}{(V_1+V_\Upsilon/\rho)}}-\tfrac{\bar{L}}{20(V_1+V_\Upsilon/\rho)}, \tag{3.4}$$

then after T iterations of Algorithm 1,

$$\mathbb{E}[\Phi(x_T,y_T)-\underline{\Phi}]\le(1-\Theta)^T(\Phi(x_0,y_0)-\underline{\Phi}+\tfrac{4\overline{\gamma}_0}{\rho}\Upsilon_0), \tag{3.5}$$

where $\Theta\overset{def}{=}\min\{\mu\nu\beta^2/4,\rho/2\}$ and $\nu\overset{def}{=}\max\{\frac{1}{4\gamma_{x,0}}-L_x,\frac{1}{4\gamma_{y,0}}-L_y\}$.

We include the proof of Theorem 3.1 in Section B of the supplementary material. Because the SAGA and SARAH gradient estimators are variance-reduced, this theorem implies specific convergence rates for Algorithm 1 when using these estimators.

Corollary 3.2. *To compute an ϵ-approximate critical point in expectation, Algorithm 1 using*

- *the SARAH gradient estimator with $p=n$ and $\overline{\gamma}_k\le\frac{1}{2L\sqrt{30n}}$ requires no more than $\mathcal{O}\left(L\sqrt{n}/\epsilon^2\right)$ SFO calls;*
- *the SAGA gradient estimator with $b=n^{2/3}$ and $\overline{\gamma}_k\le\frac{1}{2\sqrt{2330}L}$, requires no more than $\mathcal{O}(Ln^{2/3}/\epsilon^2)$ SFO calls.*[1]

If Φ satisfies the error bound (3.1), then to compute an ϵ-suboptimal point in expectation, Algorithm 1 using

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

Our third set of convergence guarantees provide tighter results for semialgebraic Φ. These convergence rates depend on its KL exponent, showing that the full convergence theory of PALM extends to SPRING.

Theorem 3.3. *Suppose Φ is a semialgebraic 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 $\gamma_{x,k},\gamma_{y,k}\in[\beta,\frac{\sqrt{2}}{5(\sqrt{V_1+V_\Upsilon/\rho+\bar{L}})})$, and $\overline{\gamma}_k$ is non-increasing.*

1. *If $\theta=0$, then there exists an $m\in\mathbb{N}$ such that $\mathbb{E}\Phi(z_k)=\Phi(z^*)$ for all $k\ge m$.*
2. *If $\theta\in(0,1/2]$, then there exists $d_1>0$ and $\tau\in[1-\rho,1)$ such that $\mathbb{E}\|z_k-z^*\|\le d_1\tau^k$.*
3. *If $\theta\in(1/2,1)$, then there exists a constant $d_2>0$ such that $\mathbb{E}\|z_k-z^*\|\le d_2k^{-\frac{1-\theta}{2\theta-1}}$.*

We include the proof of Theorem 3.3 in Section C of the supplementary material. The main difference between these convergence rates and the convergence rates of PALM is 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.

4. Numerical Experiments

In this section, we present our numerical study on the practical performance of the proposed SPRING with SAGA and SARAH gradient estimators in comparison to PALM (Bolte et al., 2014) and inertial PALM (Pock & Sabach, 2016). We also present results for SPRING using the stochastic gradient estimator (SGD), although we provide no convergence guarantees in this case. We refer to SPRING using the SGD, SAGA, and SARAH gradient estimators as SPRING-SGD, SPRING-SAGA and SPRING-SARAH, respectively.

We consider three applications of proximal alternating optimization methods in machine learning and computer vision: Sparse Non-negative Matrix Factorization (Sparse-NMF), Sparse Principal Component Analysis (Sparse-PCA), and Blind Image-Deblurring (BID). The algorithms proposed in (Aravkin & Davis, 2019) do not apply to Sparse-NMF, Sparse-PCA, or BID, so these experiments highlight SPRING's broad applicability.

Sparse-NMF: Given a data-matrix A, we seek a factorization $A\approx XY$ where $X\in\mathbb{R}^{n\times r},Y\in\mathbb{R}^{r\times d}$ have non-negative entries, $r\le d$, and X is sparse. We formulate Sparse-NMF as the following problem:

$$\begin{aligned}&\min_{X,Y}\|A-XY\|_F^2,\\&\text{s.t. } X,Y\ge0,\ \|X_i\|_0\le s,\ i=1,...,r.\end{aligned} \tag{4.1}$$

where X_i denotes the i'th column of X. In dictionary learn-

[1]For ease of exposition, we do not optimize over constants, so these step-sizes (particularly for the SAGA estimator) are not optimal.

ing and sparse coding, $X^\star$ is referred to as the learned dictionary with coefficients $Y^\star$. In this formulation, the sparsity on X is strictly enforced using the non-convex ℓ_0 constraint, but one can also use ℓ_1 regularization to preserve convexity.

Sparse-PCA: The problem of Sparse-PCA with r principal components can be written as:

$$\min_{X,Y} \|A - XY\|_F^2 + \lambda_1 \|X\|_1 + \lambda_2 \|Y\|_1, \tag{4.2}$$

where $X \in \mathbb{R}^{n \times r}, Y \in \mathbb{R}^{r \times d}$. We use ℓ_1 regularization on both X and Y to promote sparsity.

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

$$\begin{aligned} &\min_{X,Y} \|Z - X * Y\|_F^2 + \lambda \|X\|_{TV}, \\ &\text{s.t.} 0 \leq X \leq 1,\ 0 \leq Y \leq 1,\ \|Y\|_1 \leq 1, \end{aligned} \tag{4.3}$$

where $*$ is the 2D convolution operator, X is the recovered image, and Y is the estimated blur-kernel. We choose the classic TV semi-norm (Chambolle et al., 2010) as the regularizer in the image domain.

4.1. 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 the local Lipschitz constants are needed for PALM and inertial PALM (Pock & Sabach, 2016). 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.

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 to be inverse proportional to the estimated Lipschitz constants as follows:

- **PALM:** $\gamma_x = \frac{1}{\hat{L}_x(y_k)}$ and $\gamma_y = \frac{1}{\hat{L}_y(x_k)}$ which is standard for PALM (Bolte et al., 2014).
- **Inertial PALM:** $\gamma_x = \frac{0.9}{\hat{L}_x(y_k)}$, $\gamma_y = \frac{0.9}{\hat{L}_y(x_k)}$, and we set the momentum parameter to $\frac{k-1}{k+2}$, where k denotes the number of iterations. Pock & Sabach (2016) assert that this dynamic momentum parameter achieves the best practical performance.[2]
- **SPRING-SGD:** $\gamma_x = \frac{1}{\sqrt{\lceil kb/n \rceil}\tilde{L}_x(y_k)}$ and $\gamma_y = \frac{1}{\sqrt{\lceil kb/n \rceil}\tilde{L}_y(x_k)}$. It is well-known in the literature that a shrinking step-size is necessary for SGD to converge to a critical point (Bottou, 2010; Moulines & Bach, 2011; Konečný et al., 2015).
- **SPRING-SAGA:** $\gamma_x = \frac{1}{3\tilde{L}_x(y_k)}$ and $\gamma_y = \frac{1}{3\tilde{L}_y(x_k)}$.
- **SPRING-SARAH:** $\gamma_x = \frac{1}{2\tilde{L}_x(y_k)}$ and $\gamma_y = \frac{1}{2\tilde{L}_y(x_k)}$.

Remark 4.1. (Practical step-sizes for SPRING-SAGA and SPRING-SARAH.) While the step-sizes suggested in Section 3 lead to state-of-the-art theoretical convergence rate guarantees for ($\mathcal{P}$), we numerically 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 $\frac{1}{3}$ for SAGA (Defazio et al., 2014, Section 2) and $\frac{1}{2}$ for SARAH (Nguyen et al., 2017, Cororllary 3). We find that these choices are near-optimal for our methods in practice.

The same random initialization is used for all of the compared algorithms in our Sparse-NMF and Sparse-PCA experiments. We numerically 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. This simple trick was first reported in (Konečný & Richtárik, 2017) for warm-starting SVRG-type variance-reduced gradient methods.

4.2. Numerical results

We first consider Sparse-NMF on the extended Yale-B dataset and the ORL dataset. These datasets are standard facial recognition benchmarks consisting of human face images.[3] The ORL datasets contain 400 images of size 64×64, and the extended Yale-B dataset contains 2414 cropped images of size 32×32. In this experiment, we extract 49 sparse basis-images for both datasets. In each iteration of

[2] The dynamic choice of momentum parameter is not theoretically analysed in (Pock & Sabach, 2016), but it appears to be superior to the constant inertial parameter choice. Pock & Sabach (2016) suggest the aggressive step-sizes $\gamma_x = \frac{1}{\hat{L}_x(y_k)}$ and $\gamma_y = \frac{1}{\hat{L}_y(x_k)}$ 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 $\gamma_x = \frac{0.9}{\hat{L}_x(y_k)}$ and $\gamma_y = \frac{0.9}{\hat{L}_y(x_k)}$ instead. These choices ensure the algorithm is stable, and we observe that they do not compromise the convergence rate in practice.

[3] Preprocessed versions (Cai et al., 2007a;b) can be found in: http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html

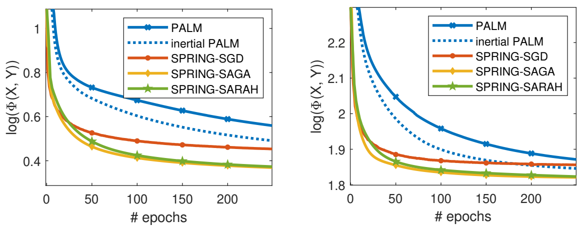

Figure 1. Sparse-NMF on (a) ORL and (b) Yale dataset.

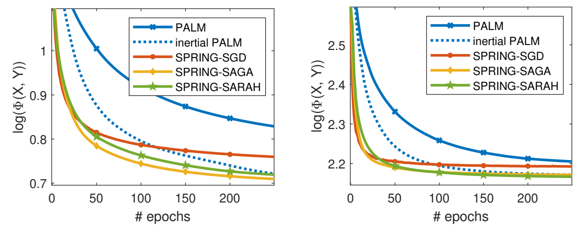

Figure 3. Sparse-PCA on (a) ORL and (b) Yale dataset.

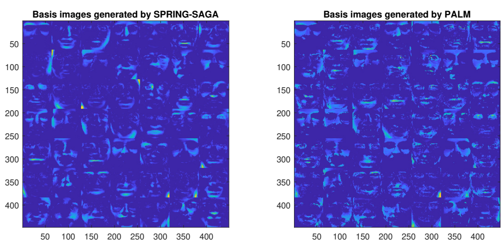

Figure 2. Sparse-NMF experiment: basis images generated by SPRING-SAGA and PALM at 250th epoch for ORL dataset.

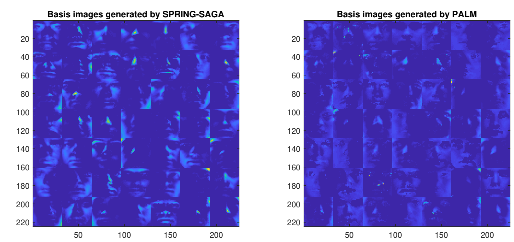

Figure 4. Sparse-NMF experiment: basis images generated by SPRING-SAGA and PALM at 10th epoch for Yale dataset.

the stochastic algorithms we randomly sub-sample 2.5% of the full batch as a mini-batch. From our numerical results shown in Figure 1, we observe that the proposed SPRING using the SAGA and SARAH stochastic variance-reduced gradient estimators achieve superior performance compared to PALM, inertial PALM, and SPRING using the vanilla SGD gradient estimator (which is not variance-reduced). PALM has the worst convergence rate in the Sparse-NMF tasks we considered, but incorporating inertia can offer considerable practical acceleration for PALM. 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.

For further demonstration, in Figure 2 we present the basis images generated by SPRING-SAGA and PALM for the ORL dataset at the 250th epoch. It is clear that the basis images generated by SPRING-SAGA appear natural and smooth, while PALM's results at the 250th iteration still seem noisy and distorted. In Figure 4, we also present the results on the Yale dataset, where we see that at the 10th epoch SPRING-SAGA is already able to provide reasonable basis images that appear natural and smooth, while the results provided by PALM at the 10th epoch contain clearly observable artefacts.

In our Sparse-PCA experiments, we compare SPRING-SAGA, SPRING-SARAH, SPRING-SGD, and PALM on the same datasets. Similar to what we observe in the Sparse-NMF experiments, our results in Figure 3 show that SPRING with stochastic variance-reduced gradient estimators achieves the best convergence rates. We also observe that the inertial scheme is able to provide significant acceleration for PALM in both the Sparse-NMF and Sparse-PCA tasks. We believe that such inertial schemes can also be extended to accelerate SPRING and leave it as an important direction of future research.

For our final experiments, we compare SPRING-SARAH, PALM, and inertial PALM for blind image-deconvolution tasks. In these experiments, we perform blind-deconvolution on blurred versions of 128×128 *Kodim08* and *Kodim15* images with additional Gaussian noise. In each iteration, we randomly sub-sample 4% of the full batch as a mini-batch for SPRING-SARAH. We run 200 epochs for each of the algorithms, and demonstrate the convergence results in Figures 5 and 6 for motion-blur and out-of-focus blur cases respectively. We present the deblurred images given by SPRING-SARAH and PALM at the 200th epoch in Figures 7 and 8. We also present the dynamics of the estimated kernels by SPRING-SARAH and PALM every 20 epochs (from the left to right) in the bottom row of Figures 7 and 8. We observe that with the same amount of computation, the proposed SPRING-SARAH algorithm provides significantly better image recovery and improved blur-kernel estimation quality than PALM and inertial 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 (Tang et al., 2019), SPRING

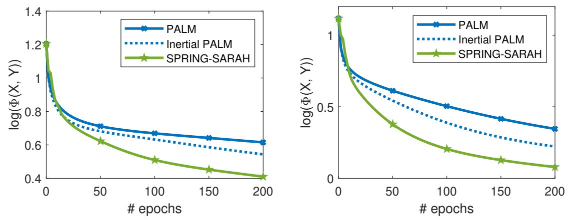

Figure 5. Blind image-deconvolution experiment on (a) Kodim08, (b) Kodim15 images, with motion-blur kernel.

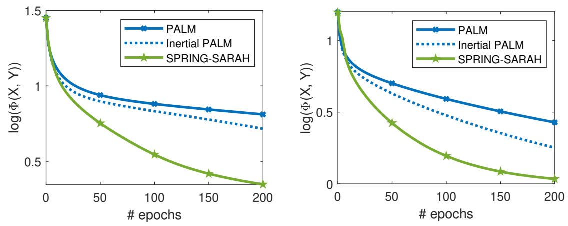

Figure 6. Blind image-deconvolution experiment on (a) Kodim08, (b) Kodim15 images, with out-of-focus blur kernel.

still offers significant acceleration over PALM in terms of epoch counts in certain blind-deblurring tasks. We observe that SPRING estimates the blur kernel much faster than PALM in our examples, hence achieving superior performance. We include additional numerical results in Section F of the supplementary material.

5. Conclusion

We propose stochastic extensions of the well-known and widely-applied PALM algorithm of Bolte et al. (2014) for solving a class of structured non-smooth and non-convex optimization problems occurring in many machine learning and computer vision applications. We analyse the convergence properties of our stochastic PALM with two typical variance-reduced stochastic gradient estimators, SAGA and SARAH. For generic optimization problems of the form ($\mathcal{P}$), we show that SPRING-SAGA and SPRING-SARAH return an ϵ-approximate critical point in expectation in no more than $O(\frac{n^2L}{b^3\epsilon^2})$ 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. These results generalize or improve on almost all existing results for stochastic non-convex optimization.

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. Our proposed methods come not only with provably superior convergence guarantees in theory,

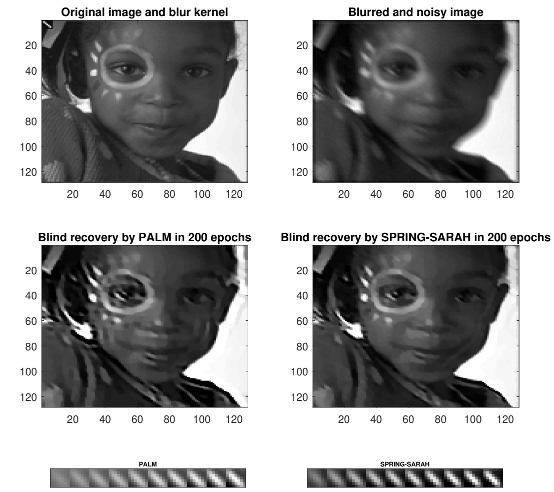

Figure 7. Blind Image-Deconvolution (7 $\times$ 7 motion blur).

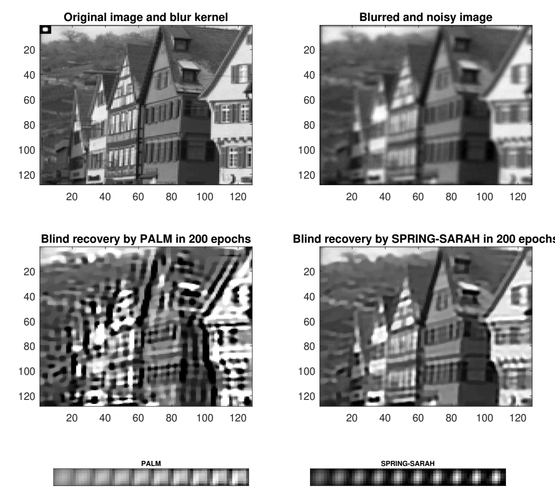

Figure 8. Blind Image-Deconvolution (7 $\times$ 7 out-of-focus blur).

but also improved practical performance, as demonstrated by our experiments.

This work suggests several prospective research directions. For further algorithmic improvements, it would be fruitful to design and analyse inertial variants of SPRING. There are also several applications of SPRING that warrant further investigation. It would be interesting to explore SPRING's performance on imaging and computer vision tasks involving advanced image priors based on deep CNN's and GAN's via the Plug-and-Play (Venkatakrishnan et al., 2013), Regularization-by-Denoising (Romano et al., 2017; Reehorst & Schniter, 2018), and adversarial regularization (Lunz et al., 2018) frameworks.

References

Aravkin, A. and Davis, D. Trimmed statistical estimation via variance reduction. *Mathematics of Operations Research*, 2019.

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

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):438–457, 2010.

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):1205–1223, 2007.

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):3319–3363, 2010.

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

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

Cai, D., He, X., and Han, J. Spectral regression for efficient regularized subspace learning. In *2007 IEEE 11th international conference on computer vision*, pp. 1–8. IEEE, 2007a.

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

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

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

Chambolle, A., Caselles, V., Cremers, D., Novaga, M., and Pock, T. An introduction to total variation for image analysis. *Theoretical foundations and numerical methods for sparse recovery*, 9(263-340):227, 2010.

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):2783–2808, 2018.

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*, pp. 41–48, 2005.

Davis, D. The asynchronous palm algorithm for nonsmooth nonconvex problems. *arXiv:1604.00526*, 2016.

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*, pp. 226–234, 2016.

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*, pp. 1646–1654, 2014.

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.

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

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

Konečnỳ, J. and Richtárik, P. Semi-stochastic gradient descent methods. *Frontiers in Applied Mathematics and Statistics*, 3:9, 2017.

Konečnỳ, 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):242–255, 2015.

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*, 18:1199–1232, 2018.

Lunz, S., Öktem, O., and Schönlieb, C.-B. Adversarial regularizers in inverse problems. In *Advances in Neural Information Processing Systems*, pp. 8507–8516, 2018.

Moulines, E. and Bach, F. R. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, pp. 451–459, 2011.

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*, volume 70, pp. 2613–2621, 2017.

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.

Pock, T. and Sabach, S. Inertial proximal alternating linearized minimization (ipalm) for nonconvex and nonsmooth problems. *SIAM Journal on Imaging Sciences*, 9 (4):1756–1787, 2016.

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.

Reehorst, E. T. and Schniter, P. Regularization by denoising: Clarifications and new interpretations. *IEEE transactions on computational imaging*, 5(1):52–67, 2018.

Robbins, H. and Monro, S. A stochastic approximation method. *Annals of Mathematical Statistics*, 22(3):400–407, 1951.

Robbins, H. and Siegmund, D. A convergence theorem for non-negative almost supermartingales and some applications. *Optimizing Methods in Statistics*, pp. 233–257, 1971.

Romano, Y., Elad, M., and Milanfar, P. The little engine that could: Regularization by denoising (red). *SIAM Journal on Imaging Sciences*, 10(4):1804–1844, 2017.

Schmidt, M., Roux, N. L., and Bach, F. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162:83–112, 2017.

Tang, J., Egiazarian, K., Golbabaee, M., and Davies, M. The practicality of stochastic optimization in imaging inverse problems. *arXiv:1910.10100*, 2019.

Venkatakrishnan, S. V., Bouman, C. A., and Wohlberg, B. Plug-and-play priors for model based reconstruction. In *2013 IEEE Global Conference on Signal and Information Processing*, pp. 945–948. IEEE, 2013.

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.

Xiao, L. and Zhang, T. A proximal stochastic gradient method with progressive variance reduction. *Technical report, Microsoft Research*, 2014.

Zhou, D. and Gu, Q. Lower bounds for smooth nonconvex finite-sum optimization. *arXiv preprint arXiv:1901.11224*, 2019.

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.

Zou, H., Hastie, T., and Tibshirani, R. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.

A. Elementary Lemmas

Lemma A.1. *Let $f : \mathbb{R}^m \to \mathbb{R}$ be a function with L-Lipschitz continuous gradient, define $\sigma : \mathbb{R}^m \to \mathbb{R}$, and let $z \in \mathrm{prox}_{\eta\sigma}(x - \eta d)$. Then*

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

Proof. By the Lipschitz continuity of ∇f, we have the inequalities

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

Furthermore, by the definition of z,

$$z \in \mathrm{argmin}_{v \in \mathbb{R}^m} \left\{ \langle d, v - x\rangle + \frac{1}{2\eta}\|v - x\|^2 + \sigma(v) \right\}. \tag{A.3}$$

Taking $v = y$, we obtain

$$0 \le \sigma(y) - \sigma(z) + \langle d, y - z\rangle + \tfrac{1}{2\eta}\left(\|x - y\|^2 - \|x - z\|^2\right). \tag{A.4}$$

Adding these three inequalities completes the proof. □

If the full gradient estimator is used, Lemma A.1 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 A.2 (Sufficient Decrease Property)**.** *Let $f : \mathbb{R}^m \to \mathbb{R}$ be a function with L-Lipschitz continuous gradient, define $\sigma : \mathbb{R}^m \to \mathbb{R}$, and let $z \in \mathrm{prox}_{\eta\sigma}(x - \eta d)$. The following inequality holds:*

$$0 \le f(x) + \sigma(x) - f(z) - \sigma(z) + \frac{1}{2L\lambda}\|d - \nabla f(x)\|^2 + \left(\frac{L(\lambda + 1)}{2} - \frac{1}{2\eta}\right)\|x - z\|^2, \tag{A.5}$$

for any $\lambda > 0$.

Proof. From Lemma A.1 with $x = y$, we have

$$0 \le f(x) + \sigma(x) - f(z) - \sigma(z) + \langle \nabla f(x) - d, z - x\rangle + (\tfrac{L}{2} - \tfrac{1}{2\eta})\|x - z\|^2. \tag{A.6}$$

Using Young's inequality to say

$$\langle \nabla f(x) - d, z - x\rangle \le \frac{1}{2L\lambda}\|d - \nabla f(x)\|^2 + \frac{L\lambda}{2}\|x - z\|^2, \tag{A.7}$$

we achieve the desired result. □

As in (Davis, 2016), we use the *supermartingale convergence theorem* to obtain almost sure convergence of certain sequences generated by SPRING. Below, we present a version of this result adapted to our context. We refer to (Davis, 2016, Thm. 4.2) and (Robbins & Siegmund, 1971, Thm. 1) for more general presentations.

Lemma A.3 (Supermartingale Convergence Theorem)**.** *Let $\mathbb{E}_k$ denote the expectation conditional on the first k iterations of SPRING. Let $\{X_k\}_{k=0}^\infty$ and $\{Y_k\}_{k=0}^\infty$ be sequences of bounded non-negative random variables such that X_k and Y_k depend only on the first k iterations of SPRING. If*

$$\mathbb{E}_k X_{k+1} + Y_k \le X_k, \tag{A.8}$$

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

To prove the convergence rates of Theorem 3.3, we use the Uniformized KL Property, which is a simple extension of Definition 2.4. We include a definition of the Uniformized KL Property from (Bolte et al., 2014) for completeness.

Lemma A.4 (Bolte et al., 2014, Lem. 6). *Let Ω be a compact set and let $\sigma : \mathbb{R}^m \to (-\infty, \infty]$ be a proper lower semicontinuous function that is constant on Ω and satisfies the KL property at each point in Ω. Then there exist $\epsilon_1, \epsilon_2 > 0$ and a continuous concave function $\varphi : [0, \epsilon_2) \to \mathbb{R}_+$ such that for all $\bar{x} \in \Omega$ and all x in the set*

$$\{x \in \mathbb{R}^m : \operatorname{dist}(x, \Omega) < \epsilon_1\} \cap [x \in \mathbb{R}^m : \sigma(\bar{x}) < \sigma(x) < \sigma(\bar{x}) + \epsilon_2], \tag{A.9}$$

one has

$$\varphi'\big(\sigma(x) - \sigma(\bar{x})\big)\operatorname{dist}\big(0, \partial\sigma(x)\big) \geq 1. \tag{A.10}$$

B. Proof of Theorem 3.1

Theorem 3.1 follows immediately from the descent property of Lemma A.1 and the fact that $\widetilde{\nabla}$ is a variance reduced estimator.

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

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

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

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

Adding these two inequalities gives

$$\begin{aligned} F(x_{k+1}, y_k) + J(x_{k+1}) &\leq F(x_k, y_k) + J(x_k) + (L_x - \tfrac{1}{2\gamma_{x,k}})\|\hat{x}_{k+1} - x_k\|^2 + (\tfrac{L_x}{2} - \tfrac{1}{2\gamma_{x,k}})\|x_{k+1} - x_k\|^2 \\ &\quad + \langle \nabla_x F(x_k, y_k) - \widetilde{\nabla}_x(x_k, y_k), x_{k+1} - \hat{x}_{k+1}\rangle \\ &\overset{①}{\leq} F(x_k, y_k) + J(x_k) + (L_x - \tfrac{1}{2\gamma_{x,k}})\|\hat{x}_{k+1} - x_k\|^2 + (\tfrac{L_x}{2} - \tfrac{1}{2\gamma_{x,k}})\|x_{k+1} - x_k\|^2 \\ &\quad + 2\gamma_{x,k}\|\nabla_x F(x_k, y_k) - \widetilde{\nabla}_x(x_k, y_k)\|^2 + \tfrac{1}{8\gamma_{x,k}}\|\hat{x}_{k+1} - x_{k+1}\|^2 \\ &\overset{②}{\leq} F(x_k, y_k) + J(x_k) + (L_x - \tfrac{1}{4\gamma_{x,k}})\|\hat{x}_{k+1} - x_k\|^2 + (\tfrac{L_x}{2} - \tfrac{1}{4\gamma_{x,k}})\|x_{k+1} - x_k\|^2 \\ &\quad + 2\gamma_{x,k}\|\nabla_x F(x_k, y_k) - \widetilde{\nabla}_x(x_k, y_k)\|^2. \end{aligned} \tag{B.3}$$

Inequality ① is Young's, and ② is the standard inequality $\|a - c\|^2 \leq 2\|a - b\|^2 + 2\|b - c\|^2$. Performing the same procedure for the updates in y_k gives

$$\begin{aligned} F(x_{k+1}, y_{k+1}) + R(y_{k+1}) \leq\ & F(x_{k+1}, y_k) + R(y_k) + (L_x - \tfrac{1}{4\gamma_{y,k}})\|\hat{y}_{k+1} - y_k\|^2 + (\tfrac{L_y}{2} - \tfrac{1}{4\gamma_{y,k}})\|y_{k+1} - y_k\|^2 \\ & + 2\gamma_{y,k}\|\nabla_y F(x_{k+1}, y_k) - \widetilde{\nabla}_y(x_{k+1}, y_k)\|^2. \end{aligned} \tag{B.4}$$

Adding inequality (B.3) and inequality (B.4), we have

$$\begin{aligned} \Phi(x_{k+1}, y_{k+1}) \leq\ & \Phi(x_k, y_k) + (L_x - \tfrac{1}{4\gamma_{x,k}})\|\hat{x}_{k+1} - x_k\|^2 + (L_y - \tfrac{1}{4\gamma_{y,k}})\|\hat{y}_{k+1} - y_k\|^2 + (\tfrac{L_x}{2} - \tfrac{1}{4\gamma_{x,k}})\|x_{k+1} - x_k\|^2 \\ & + (\tfrac{L_y}{2} - \tfrac{1}{4\gamma_{y,k}})\|y_{k+1} - y_k\|^2 \\ & + 2\overline{\gamma}_k\left(\|\nabla_x F(x_k, y_k) - \widetilde{\nabla}_x(x_k, y_k)\|^2 + \|\nabla_y F(x_{k+1}, y_k) - \widetilde{\nabla}_y(x_{k+1}, y_k)\|^2\right), \end{aligned} \tag{B.5}$$

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

$$\begin{aligned} &\mathbb{E}_k[\Phi(x_{k+1}, y_{k+1}) + (-\tfrac{L_x}{2} - 2V_1\gamma_{x,k} + \tfrac{1}{4\gamma_{x,k}})\|x_{k+1} - x_k\|^2 + (-\tfrac{L_y}{2} - 2V_1\gamma_{y,k} + \tfrac{1}{4\gamma_{y,k}})\|y_{k+1} - y_k\|^2] \\ \leq\ &\Phi(x_k, y_k) + (L_x - \tfrac{1}{4\gamma_{x,k}})\|\hat{x}_{k+1} - x_k\|^2 + (L_y - \tfrac{1}{4\gamma_{y,k}})\|\hat{y}_{k+1} - y_k\|^2 + 2\overline{\gamma}_k\Upsilon_k + 2V_1\overline{\gamma}_k\|z_k - z_{k-1}\|^2. \end{aligned} \tag{B.6}$$

Next, we use (2.3) to say

$$2\overline{\gamma}_k \Upsilon_k \leq \frac{2\overline{\gamma}_k}{\rho} \left(-\mathbb{E}_k \Upsilon_{k+1} + \Upsilon_k + V_\Upsilon (\mathbb{E}_k \|z_{k+1} - z_k\|^2 + \|z_k - z_{k-1}\|^2) \right). \tag{B.7}$$

Adding the previous two inequalities, we have

$$\begin{aligned} &\mathbb{E}_k[\Phi(x_{k+1}, y_{k+1}) + (-\tfrac{L_x}{2} - 2V_1\gamma_{x,k} - \tfrac{2V_\Upsilon \overline{\gamma}_k}{\rho} + \tfrac{1}{4\gamma_{x,k}})\|x_{k+1} - x_k\|^2 \\ &\quad + (-\tfrac{L_y}{2} - 2V_1\gamma_{y,k} - \tfrac{2V_\Upsilon \overline{\gamma}_k}{\rho} + \tfrac{1}{4\gamma_{y,k}})\|y_{k+1} - y_k\|^2 + \tfrac{2\overline{\gamma}_k}{\rho}\Upsilon_{k+1}] \\ \leq &\Phi(x_k, y_k) + (L_x - \tfrac{1}{4\gamma_{x,k}})\|\hat{x}_{k+1} - x_k\|^2 + (L_y - \tfrac{1}{4\gamma_{y,k}})\|\hat{y}_{k+1} - y_k\|^2 + \tfrac{2\overline{\gamma}_k}{\rho}\Upsilon_k \\ &\quad + 2\overline{\gamma}_k(V_1 + \tfrac{V_\Upsilon}{\rho})\|z_k - z_{k-1}\|^2. \end{aligned} \tag{B.8}$$

Let $\bar{L} = \max\{L_x, L_y\}$. Setting the step-sizes so that, for all $k \geq 0$,

$$\overline{\gamma}_k \leq \frac{1}{16}\sqrt{\frac{\bar{L}^2}{(V_1 + V_\Upsilon/\rho)^2} + \frac{16}{(V_1 + V_\Upsilon/\rho)}} - \frac{\bar{L}}{16(V_1 + V_\Upsilon/\rho)}, \quad \gamma_{x,k} < \frac{1}{4L_x}, \quad \gamma_{y,k} < \frac{1}{4L_y}, \tag{B.9}$$

we have

$$\begin{aligned} &\mathbb{E}_k[\Phi(x_{k+1}, y_{k+1}) + 2\overline{\gamma}_k(V_1 + V_\Upsilon/\rho)\|z_{k+1} - z_k\|^2 + \tfrac{2\overline{\gamma}_k}{\rho}\Upsilon_{k+1}] \\ \leq &\Phi(x_k, y_k) + (L_x - \tfrac{1}{4\gamma_{x,k}})\|\hat{x}_{k+1} - x_k\|^2 + (L_y - \tfrac{1}{4\gamma_{y,k}})\|\hat{y}_{k+1} - y_k\|^2 + 2\overline{\gamma}_k(V_1 + V_\Upsilon/\rho)\|z_k - z_{k-1}\|^2 + \tfrac{2\overline{\gamma}_k}{\rho}\Upsilon_k. \end{aligned} \tag{B.10}$$

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

$$\begin{aligned} &\mathbb{E}_k[\Phi(x_{k+1}, y_{k+1}) + 2\overline{\gamma}_{k+1}(V_1 + V_\Upsilon/\rho)\|z_{k+1} - z_k\|^2 + \tfrac{2\overline{\gamma}_{k+1}}{\rho}\Upsilon_{k+1}] \\ \leq &\Phi(x_k, y_k) - \nu\|\hat{z}_{k+1} - z_k\|^2 + 2\overline{\gamma}_k(V_1 + V_\Upsilon/\rho)\|z_k - z_{k-1}\|^2 + \tfrac{2\overline{\gamma}_k}{\rho}\Upsilon_k, \end{aligned} \tag{B.11}$$

where $\nu = \max\{\frac{1}{4\gamma_{x,0}} - L_x, \frac{1}{4\gamma_{y,0}} - L_y\}$ Applying the full expectation operator and summing from $k = 0$ to $k = T - 1$ gives

$$\frac{2\overline{\gamma}_T}{\rho}\Upsilon_T + 2\overline{\gamma}_T(V_1 + V_\Upsilon/\rho)\|z_T - z_{T-1}\|^2 + \sum_{k=0}^{T-1}\mathbb{E}\|\hat{z}_{k+1} - z_k\|^2 \leq \frac{\Phi(x_0, y_0) + \frac{2\overline{\gamma}_0}{\rho}\Upsilon_0}{\nu}. \tag{B.12}$$

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

$$\mathbb{E}\text{dist}(0, \mathcal{G}_{\frac{\gamma_{x,\alpha}}{2}, \frac{\gamma_{y,\alpha}}{2}}(z_\alpha))^2 \leq \frac{4(\Phi(x_0, y_0) + \frac{2\overline{\gamma}_0}{\rho}\Upsilon_0)}{T\nu\beta^2}. \tag{B.13}$$

□

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 (B.6):

$$\begin{aligned} &\mathbb{E}_k[\Phi(x_{k+1}, y_{k+1}) + (-\tfrac{L_x}{2} - 2V_1\gamma_{x,k} + \tfrac{1}{4\gamma_{x,k}})\|x_{k+1} - x_k\|^2 + (-\tfrac{L_y}{2} - 2V_1\gamma_{y,k} + \tfrac{1}{4\gamma_{y,k}})\|y_{k+1} - y_k\|^2] \\ \leq &\Phi(x_k, y_k) - \nu\|\hat{z}_{k+1} - z_k\|^2 + 2\overline{\gamma}_k\Upsilon_k + 2V_1\overline{\gamma}_k\|z_k - z_{k-1}\|^2. \end{aligned} \tag{B.14}$$

Using (2.3), we can say for any $c > 0$,

$$0 \leq \frac{2c\overline{\gamma}_k}{\rho}\left(-\mathbb{E}_k\Upsilon_{k+1} + (1-\rho)\Upsilon_k + V_\Upsilon(\|z_{k+1} - z_k\|^2 + \|z_k - z_{k-1}\|^2) \right). \tag{B.15}$$

Adding the previous two inequalities, we have

$$\begin{aligned} &\mathbb{E}_k[\Phi(x_{k+1}, y_{k+1}) + (-\tfrac{L_x}{2} - 2V_1\gamma_{x,k} - \tfrac{2cV_\Upsilon\overline{\gamma}_k}{\rho} + \tfrac{1}{4\gamma_{x,k}})\|x_{k+1} - x_k\|^2 \\ &\quad + (-\tfrac{L_y}{2} - 2V_1\gamma_{y,k} - \tfrac{2cV_\Upsilon\overline{\gamma}_k}{\rho} + \tfrac{1}{4\gamma_{y,k}})\|y_{k+1} - y_k\|^2 + \tfrac{2c\overline{\gamma}_k}{\rho}\Upsilon_{k+1}] \\ \leq &\Phi(x_k, y_k) - \nu\|\hat{z}_{k+1} - z_k\|^2 + 2\overline{\gamma}_k(V_1 + \tfrac{cV_\Upsilon}{\rho}\|z_k - z_{k-1}\|^2 + \tfrac{2c\overline{\gamma}_k}{\rho}(1 + \tfrac{\rho}{c} - \rho)\Upsilon_k. \end{aligned} \tag{B.16}$$

Because $\gamma_{x,k} < \frac{1}{4L_x}$ and $\gamma_{y,k} < \frac{1}{4L_y}$, we can apply the error bound assumption (3.1) to say

$$-\nu\|\hat{z}_{k+1} - z_k\|^2 \leq -\frac{\nu\underline{\gamma}_k^2}{4}\mathrm{dist}(0, \mathcal{G}_{\frac{\gamma_{x,k}}{2},\frac{\gamma_{y,k}}{2}}(z_k))^2 \leq -\frac{\mu\nu\underline{\gamma}_k^2}{4}(\Phi(x_k, y_k) - \underline{\Phi}). \tag{B.17}$$

In total, we have

$$\begin{aligned}
&\mathbb{E}_k[\Phi(x_{k+1}, y_{k+1}) - \underline{\Phi} + (-\tfrac{L_x}{2} - 2V_1\gamma_{x,k} - \tfrac{2cV_\Upsilon\overline{\gamma}_k}{\rho} + \tfrac{1}{4\gamma_{x,k}})\|x_{k+1} - x_k\|^2 \\
&\quad + (-\tfrac{L_y}{2} - 2V_1\gamma_{y,k} - \tfrac{2cV_\Upsilon\overline{\gamma}_k}{\rho} + \tfrac{1}{4\gamma_{y,k}})\|y_{k+1} - y_k\|^2 + \tfrac{2c\overline{\gamma}_k}{\rho}\Upsilon_{k+1}] \\
&\leq (1 - \tfrac{\mu\nu\underline{\gamma}_k^2}{4})(\Phi(x_k, y_k) - \underline{\Phi}) + 2\overline{\gamma}_k(V_1 + \tfrac{cV_\Upsilon}{\rho}\|z_k - z_{k-1}\|^2 + \tfrac{2c\overline{\gamma}_k}{\rho}(1 + \tfrac{\rho}{c} - \rho)\Upsilon_k.
\end{aligned} \tag{B.18}$$

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

$$\overline{\gamma}_k \leq \frac{1}{20}\sqrt{\frac{\bar{L}^2}{(V_1+2V_\Upsilon/\rho)^2} + \frac{20}{(V_1+2V_\Upsilon/\rho)}} - \frac{\bar{L}}{20(V_1+2V_\Upsilon/\rho)}, \quad \gamma_{x,k} < \frac{1}{4L_x}, \quad \gamma_{y,k} < \frac{1}{4L_y}, \quad 0 < \beta \leq \underline{\gamma}_k \quad \forall k, \tag{B.19}$$

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

$$\begin{aligned}
&\mathbb{E}_k[\Phi(x_{k+1}, y_{k+1}) - \underline{\Phi} + 2\overline{\gamma}_k(V_1 + \tfrac{2V_\Upsilon}{\rho}\|z_{k+1} - z_k\|^2 + \tfrac{4\overline{\gamma}_k}{\rho}\Upsilon_{k+1}] \\
&\leq (1 - \Theta)(\Phi(x_k, y_k) - \underline{\Phi} + 2\overline{\gamma}_k(V_1 + \tfrac{2V_\Upsilon}{\rho}\|z_k - z_{k-1}\|^2 + \tfrac{4\overline{\gamma}_k}{\rho}\Upsilon_k).
\end{aligned} \tag{B.20}$$

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

$$\begin{aligned}
&\mathbb{E}_k[\Phi(x_{k+1}, y_{k+1}) - \underline{\Phi} + 2\overline{\gamma}_{k+1}(V_1 + \tfrac{2V_\Upsilon}{\rho}\|z_{k+1} - z_k\|^2 + \tfrac{4\overline{\gamma}_{k+1}}{\rho}\Upsilon_{k+1}] \\
&\leq (1 - \Theta)(\Phi(x_k, y_k) - \underline{\Phi} + 2\overline{\gamma}_k(V_1 + \tfrac{2V_\Upsilon}{\rho}\|z_k - z_{k-1}\|^2 + \tfrac{4\overline{\gamma}_k}{\rho}\Upsilon_k).
\end{aligned} \tag{B.21}$$

Applying the full expectation operator and chaining this inequality over the iterations $k = 0$ to $k = T - 1$, we have

$$\mathbb{E}[\Phi(x_T, y_T) - \underline{\Phi}] \leq (1 - \Theta)^T\left(\Phi(x_0, y_0) - \underline{\Phi} + \frac{4\overline{\gamma}_0}{\rho}\Upsilon_0\right). \tag{B.22}$$

This completes the proof. $\square$

C. Proof of Theorem 3.3

To prove convergence rates that depend on the KL exponent of Φ, we use a procedure similar to the general approach of (Bolte et al., 2014): first, we prove the monotonic decrease (in expectation) of a non-negative Lyapunov functional, then we bound the quantity $\operatorname{dist}(0, \partial\Phi(z_k))$. Combining these results, we prove that the sequence of iterates that SPRING generates is Cauchy and converges to a critical point of Φ.

We begin with the decreasing Lyapunov functional.

Lemma C.1. *Let $\{z_k\}_{k=0}^{\infty}$ be a sequence of iterates generated by SPRING with step-sizes satisfying*

$$\overline{\gamma}_k < \frac{\sqrt{2}}{5(\sqrt{V_1 + V_\Upsilon/\rho} + \bar{L})} \quad \forall k, \tag{C.1}$$

and $\overline{\gamma}_k$ is non-increasing. The Lyapunov functional

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

satisfies

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

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

$$\sum_{k=0}^{\infty}\mathbb{E}\left[\|x_{k+1} - x_k\|^2 + \|y_{k+1} - y_k\|^2\right] = \sum_{k=0}^{\infty}\mathbb{E}\|z_{k+1} - z_k\|^2 < \infty. \tag{C.4}$$

Proof. Applying Lemma A.2 twice, once for the update in x_k and once for the update in y_k, we have

$$\begin{aligned} F(x_{k+1}, y_k) + J(x_{k+1}) \le\, & F(x_k, y_k) + J(x_k) + \frac{1}{2\bar{L}\lambda}\|\widetilde{\nabla}_x(x_k, y_k) - \nabla_x F(x_k, y_k)\|^2 \\ & + \left(\frac{\bar{L}(\lambda+1)}{2} - \frac{1}{2\gamma_{x,k}}\right)\|x_{k+1} - x_k\|^2, \end{aligned} \tag{C.5}$$

as well as

$$\begin{aligned} F(x_{k+1}, y_{k+1}) + R(y_{k+1}) \le\, & F(x_{k+1}, y_k) + R(y_k) + \frac{1}{2\bar{L}\lambda}\|\widetilde{\nabla}_y(x_{k+1}, y_k) - \nabla_y F(x_{k+1}, y_k)\|^2 \\ & + \left(\frac{\bar{L}(\lambda+1)}{2} - \frac{1}{2\gamma_{y,k}}\right)\|y_{k+1} - y_k\|^2. \end{aligned} \tag{C.6}$$

Adding these inequalities together,

$$\begin{aligned} \Phi(x_{k+1}, y_{k+1}) \le\, & \Phi(x_k, y_k) + \frac{1}{2\bar{L}\lambda}\|\widetilde{\nabla}_x(x_k, y_k) - \nabla_x F(x_k, y_k)\|^2 + \frac{1}{2\bar{L}\lambda}\|\widetilde{\nabla}_y(x_{k+1}, y_k) - \nabla_y F(x_{k+1}, y_k)\|^2 \\ & + \left(\frac{\bar{L}(\lambda+1)}{2} - \frac{1}{2\overline{\gamma}_k}\right)\|z_{k+1} - z_k\|^2. \end{aligned} \tag{C.7}$$

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

$$\mathbb{E}_k\left[\Phi(z_{k+1}) + \left(-\frac{\bar{L}(\lambda+1)}{2} - \frac{V_1}{2\bar{L}\lambda} + \frac{1}{2\overline{\gamma}_k}\right)\|z_{k+1} - z_k\|^2\right] \le \Phi(z_k) + \frac{1}{2\bar{L}\lambda}\Upsilon_k + \frac{V_1}{2\bar{L}\lambda}\|z_k - z_{k-1}\|^2. \tag{C.8}$$

Next, we use (2.3) to say

$$\frac{1}{2\bar{L}\lambda}\Upsilon_k \le \frac{1}{2\bar{L}\lambda\rho}\left(-\mathbb{E}_k\Upsilon_{k+1} + \Upsilon_k + V_\Upsilon(\mathbb{E}_k\|z_{k+1} - z_k\|^2 + \|z_k - z_{k-1}\|^2)\right). \tag{C.9}$$

Combining these inequalities, we have

$$\begin{aligned}&\mathbb{E}_k\left[\Phi(z_{k+1})+\frac{1}{2\bar{L}\lambda\rho}\Upsilon_{k+1}+\left(-\frac{\bar{L}(\lambda+1)}{2}-\frac{V_1+V_\Upsilon/\rho}{2\bar{L}\lambda}+\frac{1}{2\overline{\gamma}_k}\right)\|z_{k+1}-z_k\|^2\right]\\&\leq\Phi(z_k)+\frac{1}{2\bar{L}\lambda\rho}\Upsilon_k+\frac{V_1+V_\Upsilon/\rho}{2\bar{L}\lambda}\|z_k-z_{k-1}\|^2.\end{aligned}\tag{C.10}$$

This is equivalent to

$$\begin{aligned}&\mathbb{E}_k\left[\Phi(z_{k+1})+\frac{1}{2\bar{L}\lambda\rho}\Upsilon_{k+1}+\left(\frac{V_1+V_\Upsilon/\rho}{2\bar{L}\lambda}+Z\right)\|z_{k+1}-z_k\|^2\left(-\frac{\bar{L}(\lambda+1)}{2}-\frac{V_1+V_\Upsilon/\rho}{\bar{L}\lambda}-Z+\frac{1}{2\overline{\gamma}_k}\right)\|z_{k+1}-z_k\|^2\right]\\&\leq\Phi(z_k)+\frac{1}{2\bar{L}\lambda\rho}\Upsilon_k+\left(\frac{V_1+V_\Upsilon/\rho}{2\bar{L}\lambda}+Z\right)\|z_k-z_{k-1}\|^2-Z\|z_k-z_{k-1}\|^2,\end{aligned}\tag{C.11}$$

for some constant $Z\geq 0$. Setting $\overline{\gamma}_k\leq(2(\frac{\bar{L}(\lambda+1)}{2}+\frac{V_1+V_\Upsilon/\rho}{\bar{L}\lambda}+Z))^{-1}$ and using the fact that $\overline{\gamma}_k$ is non-increasing, we have

$$\mathbb{E}_k\Psi_{k+1}\leq\Psi_k+\left(\frac{\bar{L}(\lambda+1)}{2}+\frac{V_1+V_\Upsilon/\rho}{\bar{L}\lambda}+Z-\frac{1}{2\overline{\gamma}_k}\right)\mathbb{E}_k\|z_{k+1}-z_k\|^2-Z\|z_k-z_{k-1}\|^2.\tag{C.12}$$

proving the first claim that Ψ_k is decreasing in expectation. To approximately maximize our bound on $\overline{\gamma}_k$, we set $\lambda=\frac{\sqrt{2(V_1+V_\Upsilon/\rho)}}{\bar{L}}$.

To prove the second claim, we apply the full expectation operator to (C.12) and sum the resulting inequality from $k=0$ to $k=T-1$,

$$\mathbb{E}\Psi_T\leq\Psi_0+\frac{1}{2\bar{L}\lambda\rho}\|\Upsilon_0\|^2+\left(\frac{\bar{L}(\lambda+1)}{2}+\frac{V_1+V_\Upsilon/\rho}{\bar{L}\lambda}+Z-\frac{1}{2\overline{\gamma}_k}\right)\sum_{k=0}^{T-1}\mathbb{E}\|z_{k+1}-z_k\|^2-Z\mathbb{E}\|z_k-z_{k-1}\|^2.\tag{C.13}$$

Rearranging and using the fact that $\underline{\Phi}\leq\Psi_T$,

$$\left(\frac{1}{2\overline{\gamma}_k}-\frac{\bar{L}(\lambda+1)}{2}-\frac{V_1+V_\Upsilon/\rho}{\bar{L}\lambda}-Z\right)\sum_{k=0}^{T-1}\mathbb{E}\|z_{k+1}-z_k\|^2+Z\mathbb{E}\|z_k-z_{k-1}\|^2\leq\Psi_0-\underline{\Phi}+\frac{1}{2\bar{L}\lambda\rho}\Upsilon_0.\tag{C.14}$$

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

Inequalities (C.12) and (C.14) hold for any choice of $Z\geq 0$; we set $Z=\frac{\sqrt{V_1+V_\Upsilon/\rho}}{2\sqrt{2}}$ to simplify later arguments. □

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

Lemma C.2 (Subgradient Bound). *Let $\{z_k\}_{k\in\mathbb{N}}$ be a bounded sequence generated by SPRING with step-sizes satisfying $0<\beta\leq\underline{\gamma}_k$. Define*

$$A_x^k\overset{\text{def}}{=}1/\gamma_{x,k}(x_{k-1}-x_k)+\nabla_x F(x_k,y_k)-\widetilde{\nabla}_x(x_{k-1},y_{k-1}).\tag{C.15}$$

and

$$A_y^k\overset{\text{def}}{=}1/\gamma_{y,k}(y_{k-1}-y_k)+\nabla_y F(x_k,y_k)-\widetilde{\nabla}_y(x_k,y_{k-1}).\tag{C.16}$$

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

$$\mathbb{E}_{k-1}\|(A_x^k,A_y^k)\|\leq p(\mathbb{E}_{k-1}\|z_k-z_{k-1}\|+\|z_{k-1}-z_{k-2}\|)+\Gamma_{k-1},\quad\forall k\geq 1,\tag{C.17}$$

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

$$\frac{1}{\gamma_{x,k}}(x_{k-1}-x_k)-\widetilde{\nabla}_x(x_{k-1},y_{k-1})\in\partial J(x_k),\quad\text{and}\quad\frac{1}{\gamma_{y,k}}(y_{k-1}-y_k)-\widetilde{\nabla}_y(x_k,y_{k-1})\in\partial R(y_k).\tag{C.18}$$

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 ∇F is M-Lipschitz continuous on bounded sets and we assume that the sequence $\{z_k\}_{k=0}^\infty$ is bounded, we can say

$$\begin{aligned}
&\mathbb{E}_{k-1}\|A_x^k\| \\
&\leq \frac{1}{\gamma_{x,k}}\mathbb{E}_{k-1}\|x_{k-1} - x_k\| + \mathbb{E}_{k-1}\|\nabla_x F(x_k, y_k) - \widetilde{\nabla}_x(x_{k-1}, y_{k-1})\| \\
&\leq \frac{1}{\gamma_{x,k}}\mathbb{E}_{k-1}\|x_{k-1} - x_k\| + \mathbb{E}_{k-1}\|\nabla_x F(x_k, y_k) - \nabla_x F(x_{k-1}, y_{k-1})\| + \mathbb{E}_{k-1}\|\nabla_x F(x_{k-1}, y_{k-1}) - \widetilde{\nabla}_x(x_{k-1}, y_{k-1})\| \\
&\leq \left(\frac{1}{\gamma_{x,k}} + M\right)\mathbb{E}_{k-1}\|x_{k-1} - x_k\| + M\mathbb{E}_{k-1}\|y_k - y_{k-1}\| + \mathbb{E}_{k-1}\|\nabla_x F(x_{k-1}, y_{k-1}) - \widetilde{\nabla}_x(x_{k-1}, y_{k-1})\|.
\end{aligned} \tag{C.19}$$

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

$$\begin{aligned}
&\mathbb{E}_{k-1}\|A_y^k\| \\
&\leq \frac{1}{\gamma_{y,k}}\mathbb{E}_{k-1}\|y_{k-1} - y_k\| + \mathbb{E}_{k-1}\|\nabla_y F(x_k, y_k) - \widetilde{\nabla}_y(x_k, y_{k-1})\| \\
&\leq \frac{1}{\gamma_{y,k}}\mathbb{E}_{k-1}\|y_{k-1} - y_k\| + \mathbb{E}_{k-1}\|\nabla_y F(x_k, y_k) - \nabla_y F(x_k, y_{k-1})\| + \mathbb{E}_{k-1}\|\nabla_y F(x_k, y_{k-1}) - \widetilde{\nabla}_y(x_k, y_{k-1})\| \\
&\leq \left(\frac{1}{\gamma_{y,k}} + L_y\right)\mathbb{E}_{k-1}\|y_{k-1} - y_k\| + \mathbb{E}_{k-1}\|\nabla_y F(x_k, y_{k-1}) - \widetilde{\nabla}_y(x_k, y_{k-1})\|.
\end{aligned} \tag{C.20}$$

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

$$\mathbb{E}_{k-1}\|(A_x^k, A_y^k)\| \leq \mathbb{E}_{k-1}\left[\|A_x^k\| + \|A_y^k\|\right] \leq p(\mathbb{E}_{k-1}\|z_k - z_{k-1}\| + \|z_{k-1} - z_{k-2}\|) + \Gamma_{k-1}. \tag{C.21}$$

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

Lemma C.3. *Let $\{z_k\}_{k=0}^\infty$ be a bounded sequence of iterates of SPRING using a variance-reduced gradient estimator and step-sizes satisfying*

$$\gamma_{x,k}, \gamma_{y,k} \in \left[\beta, \frac{\sqrt{2}}{5(\sqrt{V_1 + V_\Upsilon/\rho} + \bar{L})}\right) \quad \forall k, \tag{C.22}$$

and $\overline{\gamma}_k$ is non-increasing. Define the set of limit points of $\{z_k\}_{k=0}^\infty$ as

$$\omega(z_0) \stackrel{\text{def}}{=} \{z : \exists \text{ an increasing sequence of integers } \{k_\ell\}_{\ell\in\mathbb{N}} \text{ such that } z_{k_\ell} \to z \text{ as } \ell \to \infty\}. \tag{C.23}$$

Then

1. $\sum_{k=1}^\infty \|z_k - z_{k-1}\|^2 < \infty$ *a.s., and* $\|z_k - z_{k-1}\| \to 0$ *a.s.;*
2. $\mathbb{E}\Phi(z_k) \to \Phi^*$*, where* $\Phi^* \in [\underline{\Phi}, \infty)$*;*
3. $\mathbb{E}\mathrm{dist}(0, \partial\Phi(z_k)) \to 0$*;*
4. *The set $\omega(z_0)$ a.s. is non-empty and contains only critical points of Φ;*
5. $\mathrm{dist}(z_k, \omega(z_0)) \to 0$ *a.s.;*
6. *$\omega(z_0)$ is a.s. compact and connected;*
7. $\mathbb{E}\Phi(z^*) = \Phi^*$ *for all* $z^* \in \omega(z_0)$.

Proof. By Lemma C.1, we have

$$\mathbb{E}_k\Psi_{k+1} + \mathcal{O}\left(\|z_k - z_{k-1}\|^2\right) \leq \Psi_k. \tag{C.24}$$

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 Ψ_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 $\mathbb{E}\Psi_k$ converges to a finite value bounded below by $\underline{\Phi}$ for all k, implying Claim 2.

Claim 3 holds because, by Lemma C.2,

$$\mathbb{E}\|(A_x^k, A_y^k)\| \leq p\mathbb{E}[\|z_k - z_{k-1}\| + \|z_{k-1} - z_{k-2}\|] + \mathbb{E}\Gamma_{k-1}. \tag{C.25}$$

We have that $\|z_k - z_{k-1}\| \to 0$ a.s. 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 $z^* = (x^*, y^*)$ 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 $\lim_{q\to\infty} z_{k_q} \to z^*$. Because R and J are lower semicontinuous,

$$\liminf_{q\to\infty} R(x_{k_q}) \geq R(x^*), \quad \text{and} \quad \liminf_{q\to\infty} J(x_{k_q}) \geq J(x^*). \tag{C.26}$$

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

$$x_{k+1} \in \operatorname{argmin}_x \left\{ \langle x - x_k, \widetilde{\nabla}_x(x_k, y_k)\rangle + \frac{1}{2\gamma_{x,k}}\|x - x_k\|^2 + R(x) \right\}. \tag{C.27}$$

Letting $x = x^*$,

$$\begin{aligned} &\langle x_{k+1} - x_k, \widetilde{\nabla}_x(x_k, y_k)\rangle + \frac{1}{2\gamma_{x,k}}\|x_{k+1} - x_k\|^2 + R(x_{k+1}) \\ \leq &\langle x^* - x_k, \nabla_x F(x_k, y_k)\rangle + \langle x^* - x_k, \widetilde{\nabla}_x(x_k, y_k) - \nabla_x F(x_k, y_k)\rangle + \frac{1}{2\gamma_{x,k}}\|x^* - x_k\|^2 + R(x^*). \end{aligned} \tag{C.28}$$

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

$$\begin{aligned} &\limsup_{q\to\infty} R(x_{k_q+1}) \\ \leq &\limsup_{q\to\infty} \langle x^* - x_{k_q}, \nabla_x F(x_{k_q}, y_{k_q})\rangle + \langle x^* - x_{k_q}, \widetilde{\nabla}_x(x_{k_q}, y_{k_q}) - \nabla_x F(x_{k_q}, y_{k_q})\rangle + \frac{1}{2\gamma_{x,k}}\|x^* - x_{k_q}\|^2 + R(x^*). \end{aligned} \tag{C.29}$$

Because $x_{k_q} \to x^*$, we can say $\limsup_{q\to\infty} R(x_{k_q+1}) \leq R(x^*)$, which, together with equation (C.26), implies $R(x_{k_q+1}) \to R(x^*)$. The same argument holds for J and y_k, and it follows that

$$\lim_{q\to\infty} \Phi(x_{k_q}, y_{k_q}) = \Phi(x^*, y^*). \tag{C.30}$$

Lemma C.2 ensures that (x^*, y^*) is a critical point of Φ because $\mathbb{E}\|(A_x^k, A_y^k)\| \to (0,0)$ as $k \to \infty$ and $\partial\Phi(x^*, y^*)$ is closed.

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 (Bolte et al., 2014, Remark 5) and (Davis, 2016, Remark 4.1)).

Finally, we must show that Φ has constant expectation over $\omega(z_0)$. From Claim 2, we have that $\mathbb{E}\Phi(z_k) \to \Phi^*$, which implies that $\mathbb{E}\Phi(z_{k_q}) \to \Phi^*$ for every subsequence $\{z_{k_q}\}_{q=0}^\infty$ converging to some $z^* \in \omega(z_0)$. In the proof of Claim 4, we show that $\Phi(z_{k_q}) \to \Phi(z^*)$, so $\mathbb{E}\Phi(z^*) = \Phi^*$ for all $z^* \in \omega(z_0)$.

□

Lemma C.4. *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 C.3, and suppose that z_k is not a critical point after a finite number of iterations. Let Φ be a semialgebraic function satisfying the Kurdyka–Łojasiewicz property with exponent θ. Then there exists an index m and a desingularizing function $\phi = ar^{1-\theta}$ and the following bound holds almost surely:*

$$\phi'(\mathbb{E}[\Phi(z_k) - \Phi_k^*])\mathbb{E}\operatorname{dist}(0, \partial\Phi(z_k)) \geq 1 \quad \forall k > m, \tag{C.31}$$

where Φ_k^* is a non-decreasing sequence converging to $\mathbb{E}\Phi(z^*)$ for some $z^* \in \omega(z_0)$.

Proof. First, we show that $\mathbb{E}\Phi(z_k)$ satisfies the KL property. 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. It is clear that $\mathbb{E}\Phi$ is a function of $\{z_k^i\}_{i=1}^{\overline{n}^k}$:

$$\mathbb{E}\Phi(z_k) = \frac{1}{\overline{n}^k}\sum_{i=1}^{\overline{n}^k} \Phi(z_k^i). \tag{C.32}$$

Because $\mathbb{E}\Phi(z_k)$ can be written as $\sum_i f_i(x_i)$ where f_i are KL functions with exponent θ, $\mathbb{E}\Phi(z_k)$ (as a function of $\{z_k^i\}_{i=0}^{\overline{n}^k}$) is also KL with exponent θ (Li & Pong, 2018, Thm. 3.3). Hence, $\mathbb{E}\Phi$ satisfies the KL inequality at every point in its domain. Therefore, for every point $(z_k^1, \cdots, z_k^{n_k})$ in a neighborhood U_k of $(\overline{z}_k^1, \overline{z}_k^2, \cdots, \overline{z}_k^{n_k})$ and satisfying

$$\frac{1}{\overline{n}^k}\sum_{i=1}^{\overline{n}^k}\Phi(\overline{z}_k^i) < \frac{1}{\overline{n}^k}\sum_{i=1}^{\overline{n}^k}\Phi(z_k^i) < \frac{1}{\overline{n}^k}\sum_{i=1}^{\overline{n}^k}\Phi(\overline{z}_k^i) + \epsilon_k \tag{C.33}$$

for some $\epsilon_k > 0$, the Kurdyka–Łojasiewicz inequality holds:

$$\phi'\left(\frac{1}{\overline{n}^k}\sum_{i=1}^{\overline{n}^k}\Phi(z_k^i) - \frac{1}{\overline{n}^k}\sum_{i=1}^{\overline{n}^k}\Phi(\overline{z}_k^i)\right)\operatorname{dist}\left(0, \frac{1}{\overline{n}^k}\sum_{i=1}^{\overline{n}^k}\partial\Phi(z_k^i)\right) \geq 1. \tag{C.34}$$

There always exists a choice of $(\overline{z}_k^1, \overline{z}_k^2, \cdots, \overline{z}_k^{n_k})$ satisfying (C.33) unless $\mathbb{E}\Phi(z_k)$ is a local minimum.

Let $\Phi_k^* \stackrel{\text{def}}{=} \frac{1}{\overline{n}^k}\sum_{i=1}^{\overline{n}^k}\Phi(\overline{z}_k^i)$. By Lemma C.3, Claim 5 implies $\operatorname{dist}(z_k, \omega(z_0)) \to 0$ for some $z^* \in \omega(z_0)$ a.s., and Claims 2 and 7 imply $\Phi_k^* \to \mathbb{E}\Phi(z^*)$. These results show a.s. that there exists an index m such that for all $k \geq m$, we can choose $\overline{z}_k^i$ so that Φ_k^* is non-decreasing and bounded above by $\Phi(z^*)$. Hence, we have shown

$$\phi'(\mathbb{E}[\Phi(z_k) - \Phi_k^*])\operatorname{dist}(0, \mathbb{E}\partial\Phi(z_k)) \geq 1 \quad \forall k > m, \tag{C.35}$$

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

Lemma C.5 (Finite Length). *Suppose Φ is a semialgebraic 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 C.3. Then $\{z_k\}_{k=0}^\infty$ almost surely satisfies the finite length property in expectation:*

$$\sum_{k=0}^{\infty}\mathbb{E}\|z_{k+1} - z_k\| < \infty. \tag{C.36}$$

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

$$\begin{aligned}\sum_{k=m}^{i}\mathbb{E}\|z_{k+1} - z_k\| + \mathbb{E}\|z_k - z_{k-1}\| \leq& \sqrt{\mathbb{E}\|z_m - z_{m-1}\|^2} + \sqrt{\mathbb{E}\|z_{m-1} - z_{m-2}\|^2} + \frac{2\sqrt{s}}{K_1\rho}\sqrt{\mathbb{E}\Upsilon_{m-1}} \\ &+ K_3\Delta_{m,i+1},\end{aligned} \tag{C.37}$$

where

$$K_1 \stackrel{\text{def}}{=} p + 2\sqrt{sV_\Upsilon}/\rho, \quad K_2 \stackrel{\text{def}}{=} \frac{1}{2\overline{\gamma}_0} - \frac{\bar{L}}{2} - \frac{3\sqrt{2}}{4}\sqrt{V_1 + V_\Upsilon/\rho}, \quad K_3 \stackrel{\text{def}}{=} \frac{2K_1(K_2 + Z)}{K_2 Z}, \tag{C.38}$$

p is as in Lemma C.2, and $\Delta_{p,q} \stackrel{\text{def}}{=} \phi(\mathbb{E}[\Psi_p - \Psi^*]) - \phi(\mathbb{E}[\Psi_q - \Psi^*])$.

Proof. If $\theta \in (0, 1/2)$, then Φ satisfies the KL property with exponent $1/2$, so we consider only the case $\theta \in [1/2, 1)$. By Lemma C.4, there exists a function $\phi_0(r) = ar^{1-\theta}$ such that, almost surely,

$$\phi_0'(\mathbb{E}[\Phi(z_k) - \Phi(z^*)])\mathbb{E}\operatorname{dist}(0, \partial\Phi(z_k)) \geq 1 \quad \forall k > m. \tag{C.39}$$

Lemma C.2 provides a bound on $\mathbb{E}\operatorname{dist}(0, \partial\Phi(z_k))$.

$$\begin{aligned}\mathbb{E}\operatorname{dist}(0, \partial\Phi(z_k)) &\leq \mathbb{E}\|(A_x^k, A_y^k)\| \leq p\mathbb{E}[\|z_k - z_{k-1}\| + \|z_{k-1} - z_{k-2}\|] + \mathbb{E}\Gamma_{k-1} \\ &\leq p(\sqrt{\mathbb{E}\|z_k - z_{k-1}\|^2} + \sqrt{\mathbb{E}\|z_{k-1} - z_{k-2}\|^2}) + \sqrt{s\mathbb{E}\Upsilon_{k-1}}.\end{aligned} \tag{C.40}$$

The final inequality is Jensen's. Because $\Gamma_k = \sum_{i=1}^s \|v_k^i\|$ for some vectors v_k^i, we can say $\mathbb{E}\Gamma_k = \mathbb{E}\sum_{i=1}^s \|v_k^i\| \leq \mathbb{E}\sqrt{s\sum_{i=1}^s \|v_k^i\|^2} \leq \sqrt{s\mathbb{E}\Upsilon_k}$. We can bound the term $\sqrt{\mathbb{E}\Upsilon_k}$ using (2.3):

$$\begin{aligned}\sqrt{\mathbb{E}\Upsilon_k} &\leq \sqrt{(1-\rho)\mathbb{E}\Upsilon_{k-1} + V_\Upsilon\mathbb{E}[\|z_k - z_{k-1}\|^2 + \|z_{k-1} - z_{k-2}\|^2]} \\ &\leq \sqrt{(1-\rho)}\sqrt{\mathbb{E}\Upsilon_{k-1}} + \sqrt{V_\Upsilon}(\sqrt{\mathbb{E}\|z_k - z_{k-1}\|^2} + \sqrt{\mathbb{E}\|z_{k-1} - z_{k-2}\|^2}) \\ &\leq \left(1 - \frac{\rho}{2}\right)\sqrt{\mathbb{E}\Upsilon_{k-1}} + \sqrt{V_\Upsilon}(\sqrt{\mathbb{E}\|z_k - z_{k-1}\|^2} + \sqrt{\mathbb{E}\|z_{k-1} - z_{k-2}\|^2}).\end{aligned} \tag{C.41}$$

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

$$\mathbb{E}\mathrm{dist}\,(0, \partial\Phi(z_k)) \leq K_1\sqrt{\mathbb{E}\|z_k - z_{k-1}\|^2} + K_1\sqrt{\mathbb{E}\|z_{k-1} - z_{k-2}\|^2} + \frac{2\sqrt{s}}{\rho}(\sqrt{\mathbb{E}\Upsilon_{k-1}} - \sqrt{\mathbb{E}\Upsilon_k}), \tag{C.42}$$

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

$$C_k \stackrel{\text{def}}{=} K_1\sqrt{\mathbb{E}\|z_k - z_{k-1}\|^2} + K_1\sqrt{\mathbb{E}\|z_{k-1} - z_{k-2}\|^2} + \frac{2\sqrt{s}}{\rho}(\sqrt{\mathbb{E}\Upsilon_{k-1}} - \sqrt{\mathbb{E}\Upsilon_k}). \tag{C.43}$$

We then have

$$\phi_0'(\mathbb{E}[\Phi(z_k) - \Phi(z^*)])C_k \geq 1 \quad \forall k > m. \tag{C.44}$$

By the definition of ϕ_0, this is equivalent to

$$\frac{a(1-\theta)C_k}{(\mathbb{E}[\Phi(z_k) - \Phi(z^*)])^\theta} \geq 1 \quad \forall k > m. \tag{C.45}$$

We would like the inequality above to hold for Ψ_k rather than $\Phi(z_k)$. Let $\Psi^* = \Phi(z^*)$. 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 (C.45) still holds after this adjustment because these terms are small compared to C_k.

The quantity $C_k \geq \mathcal{O}(\sqrt{\mathbb{E}\|z_k - z_{k-1}\|^2} + \sqrt{\mathbb{E}\|z_{k-1} - z_{k-2}\|^2} + \sqrt{\mathbb{E}\Upsilon_{k-1}})$, 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 constant $c > 0$ such that

$$\begin{aligned}\left(\mathbb{E}\left[\frac{1}{2\rho\sqrt{2(V_1 + V_\Upsilon/\rho)}}\Upsilon_k + \frac{\sqrt{V_1 + V_\Upsilon/\rho}}{\sqrt{2}}\|z_k - z_{k-1}\|^2\right]\right)^\theta &\leq \mathcal{O}\left(\left(\mathbb{E}\left[\Upsilon_{k-1} + \|z_k - z_{k-1}\|^2 + \|z_{k-1} - z_{k-2}\|^2\right]\right)^\theta\right) \\ &\leq cC_k \quad \forall k > m.\end{aligned} \tag{C.46}$$

The first inequality uses (2.3). Because the terms above are small compared to C_k, there exists a constant $d > c$ such that

$$\frac{ad(1-\theta)C_k}{(\mathbb{E}[\Phi(z_k) - \Phi(z^*)])^\theta + \left(\mathbb{E}\left[\frac{1}{2\rho\sqrt{2(V_1 + V_\Upsilon/\rho)}}\Upsilon_k + \frac{\sqrt{V_1 + V_\Upsilon/\rho}}{\sqrt{2}}\|z_k - z_{k-1}\|^2\right]\right)^\theta} \geq 1. \tag{C.47}$$

Using the fact that $(a+b)^\theta \leq a^\theta + b^\theta$ because $\theta \in [1/2, 1)$, we finally have

$$\begin{aligned}\frac{ad(1-\theta)C_k}{(\mathbb{E}[\Psi_k - \Psi^*])^\theta} &= \frac{ad(1-\theta)C_k}{\left(\mathbb{E}\left[\Phi(z_k) - \Phi(z^*) + \frac{1}{2\rho\sqrt{2(V_1+V_\Upsilon/\rho)}}\Upsilon_k + \frac{\sqrt{V_1+V_\Upsilon/\rho}}{\sqrt{2}}\|z_k - z_{k-1}\|^2\right]\right)^\theta} \\ &\geq \frac{ad(1-\theta)C_k}{(\mathbb{E}\left[\Phi(z_k) - \Phi(z^*)\right])^\theta + \left(\mathbb{E}\left[\frac{1}{2\rho\sqrt{2(V_1+V_\Upsilon/\rho)}}\Upsilon_k + \frac{\sqrt{V_1+V_\Upsilon/\rho}}{\sqrt{2}}\|z_k - z_{k-1}\|^2\right]\right)^\theta} \\ &\geq 1 \quad \forall k > m.\end{aligned} \tag{C.48}$$

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

$$\phi'(\mathbb{E}[\Psi_k - \Psi^*])C_k \geq 1 \quad \forall k > m. \tag{C.49}$$

By the concavity of ϕ,

$$\phi(\mathbb{E}[\Psi_k - \Psi^*]) - \phi(\mathbb{E}[\Psi_{k+1} - \Psi^*]) \geq \phi'(\mathbb{E}[\Psi_k - \Psi^*])(\mathbb{E}[\Psi_k - \Psi_{k+1}]). \tag{C.50}$$

With $\Delta_{p,q} \stackrel{\text{def}}{=} \phi(\mathbb{E}[\Psi_p - \Psi^*]) - \phi(\mathbb{E}[\Psi_q - \Psi^*])]$, we have shown

$$\Delta_{k,k+1}C_k \geq \mathbb{E}[\Psi_k - \Psi_{k+1}]. \tag{C.51}$$

Using Lemma C.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,

$$\Delta_{k,k+1} C_k \geq Z\mathbb{E}[\|z_k - z_{k-1}\|^2] \tag{C.52}$$

as well as

$$\Delta_{k,k+1} C_k \geq K_2 \mathbb{E}[\|z_{k+1} - z_k\|^2], \tag{C.53}$$

where

$$K_2 \stackrel{\text{def}}{=} -\left(\frac{\bar{L}(\lambda+1)}{2} + \frac{V_1 + V_\Upsilon/\rho}{\bar{L}\lambda} + Z - \frac{1}{2\overline{\gamma}_0}\right), \tag{C.54}$$

and λ and Z are set as in Lemma C.1. Let us use the first of these inequalities to begin. Applying Young's inequality to (C.52) yields

$$2\sqrt{\mathbb{E}\|z_k - z_{k-1}\|^2} \leq 2\sqrt{C_k \Delta_{k,k+1} Z^{-1}} \leq \frac{C_k}{2K_1} + \frac{2K_1 \Delta_{k,k+1}}{Z} \tag{C.55}$$

Summing inequality (C.55) from $k = m$ to $k = i$,

$$\begin{aligned} 2\sum_{k=m}^{i} \sqrt{\mathbb{E}\|z_k - z_{k-1}\|^2} &\leq \sum_{k=m}^{i} \frac{C_k}{2K_1} + \frac{2K_1 \Delta_{m,i+1}}{Z} \\ &\leq \sum_{k=m}^{i} \frac{1}{2}\sqrt{\mathbb{E}\|z_k - z_{k-1}\|^2} + \frac{1}{2}\sqrt{\mathbb{E}\|z_{k-1} - z_{k-2}\|^2} \\ &\quad - \frac{\sqrt{s}}{K_1\rho}\left(\sqrt{\mathbb{E}\Upsilon_i} - \sqrt{\mathbb{E}\Upsilon_{m-1}}\right) + \frac{2K_1 \Delta_{m,i+1}}{Z}, \end{aligned} \tag{C.56}$$

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

$$\frac{3}{2}\sum_{k=m}^{i} \sqrt{\mathbb{E}\|z_k - z_{k-1}\|^2} \leq \frac{1}{2}\sqrt{\mathbb{E}\|z_{m-1} - z_{m-2}\|^2} + \frac{\sqrt{s}}{K_1\rho}\sqrt{\mathbb{E}\Upsilon_{m-1}} + \frac{2K_1 \Delta_{m,i+1}}{Z}. \tag{C.57}$$

Applying the same argument using inequality (C.53) instead of (C.52), we obtain

$$\frac{3}{2}\sum_{k=m}^{i} \sqrt{\mathbb{E}\|z_{k+1} - z_k\|^2} \leq \frac{1}{2}\sqrt{\mathbb{E}\|z_m - z_{m-1}\|^2} + \frac{1}{2}\sqrt{\mathbb{E}\|z_{m-1} - z_{m-2}\|^2} + \frac{\sqrt{s}}{K_1\rho}\sqrt{\mathbb{E}\Upsilon_{m-1}} + \frac{2K_1 \Delta_{m,i+1}}{K_2}. \tag{C.58}$$

Adding these inequalities together, we have

$$\begin{aligned} \frac{3}{2}\left(\sum_{k=m}^{i} \sqrt{\mathbb{E}\|z_{k+1} - z_k\|^2} + \sqrt{\mathbb{E}\|z_k - z_{k-1}\|^2}\right) \leq& \frac{1}{2}\sqrt{\mathbb{E}\|z_m - z_{m-1}\|^2} + \sqrt{\mathbb{E}\|z_{m-1} - z_{m-2}\|^2} \\ &+ \frac{2\sqrt{s}}{K_1\rho}\sqrt{\mathbb{E}\Upsilon_{m-1}} + \frac{2K_1(K_2+Z)\Delta_{m,i+1}}{K_2 Z}. \end{aligned} \tag{C.59}$$

This implies that

$$\begin{aligned} \sum_{k=m}^{i} \sqrt{\mathbb{E}\|z_{k+1} - z_k\|^2} + \sqrt{\mathbb{E}\|z_k - z_{k-1}\|^2} \leq& \sqrt{\mathbb{E}\|z_m - z_{m-1}\|^2} + \sqrt{\mathbb{E}\|z_{m-1} - z_{m-2}\|^2} + \frac{2\sqrt{s}}{K_1\rho}\sqrt{\mathbb{E}\Upsilon_{m-1}} \\ &+ \frac{2K_1(K_2+Z)\Delta_{m,i+1}}{K_2 Z}. \end{aligned} \tag{C.60}$$

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

$$\begin{aligned} \sum_{k=m}^{i} \mathbb{E}\|z_{k+1} - z_k\| + \mathbb{E}\|z_k - z_{k-1}\| \leq& \sqrt{\mathbb{E}\|z_m - z_{m-1}\|^2} + \sqrt{\mathbb{E}\|z_{m-1} - z_{m-2}\|^2} + \frac{2\sqrt{s}}{K_1\rho}\sqrt{\mathbb{E}\Upsilon_{m-1}} \\ &+ \frac{2K_1(K_2+Z)\Delta_{m,i+1}}{K_2 Z}, \end{aligned} \tag{C.61}$$

and letting $i \to \infty$ proves the assertion. □

Theorem C.6 (Convergence Rates)**.** *Suppose Φ is a semialgebraic 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 C.3, and let K_1, K_2, and K_3 be as in Lemma C.5. The following convergence rates hold almost surely:*

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

Proof. As in the proof of the previous lemma, if $\theta \in (0, 1/2)$, then Φ 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 (C.60),

$$\begin{aligned}\sum_{k=m}^{\infty} \sqrt{\mathbb{E}\|z_{k+1} - z_k\|^2} + \sqrt{\mathbb{E}\|z_k - z_{k-1}\|^2} \leq& \sqrt{\mathbb{E}\|z_m - z_{m-1}\|^2} + \sqrt{\mathbb{E}\|z_{m-1} - z_{m-2}\|^2} + \frac{2\sqrt{s}}{K_1\rho}\sqrt{\mathbb{E}\Upsilon_{m-1}} \\ &+ aK_3(\mathbb{E}[\Psi_m - \Psi^*])^{1-\theta}.\end{aligned} \tag{C.62}$$

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

$$\begin{aligned}(\mathbb{E}[\Psi_m - \Psi^*])^{1-\theta} &= (\mathbb{E}[\Phi(z_m) - \Phi(z^*) + \frac{1}{2\bar{L}\lambda\rho}\Upsilon_m + \frac{V_1 + V_\Upsilon/\rho}{2\bar{L}\lambda}\|z_m - z_{m-1}\|^2])^{1-\theta} \\ &\overset{①}{\leq} (\mathbb{E}[\Phi(z_m) - \Phi(z^*)])^{1-\theta} + \left(\frac{1}{2\bar{L}\lambda\rho}\mathbb{E}\Upsilon_m\right)^{1-\theta} + \left(\frac{V_1 + V_\Upsilon/\rho}{2\bar{L}\lambda}\mathbb{E}\|z_m - z_{m-1}\|^2\right)^{1-\theta}.\end{aligned} \tag{C.63}$$

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

$$\begin{aligned}&\sum_{k=m}^{\infty} \sqrt{\mathbb{E}\|z_{k+1} - z_k\|^2} + \sqrt{\mathbb{E}\|z_k - z_{k-1}\|^2} \\ &\leq \sqrt{\mathbb{E}\|z_m - z_{m-1}\|^2} + \sqrt{\mathbb{E}\|z_{m-1} - z_{m-2}\|^2} + \frac{2\sqrt{s}}{K_1\rho}\sqrt{\mathbb{E}\Upsilon_{m-1}} + aK_3(\mathbb{E}[\Phi(z_m) - \Phi(z^*)])^{1-\theta} \\ &\quad + \left(\frac{1}{2\bar{L}\lambda\rho}\mathbb{E}\Upsilon_m\right)^{1-\theta} + \left(\frac{V_1 + V_\Upsilon/\rho}{2\bar{L}\lambda}\mathbb{E}\|z_m - z_{m-1}\|^2\right)^{1-\theta}.\end{aligned} \tag{C.64}$$

Applying Łojasiewicz's inequality (2.5),

$$aK_3(\mathbb{E}\left[\Phi(z_m) - \Phi(z^*)\right])^{1-\theta} \leq aK_3(\mathbb{E}\|\zeta_m\|)^{\frac{1-\theta}{\theta}}, \tag{C.65}$$

where $\zeta_m \in \partial\Phi(z_m)$ and we have absorbed the constant C into a. Equation C.40 provides a bound on the norm of the subgradient:

$$(\mathbb{E}\|\zeta_m\|)^{\frac{1-\theta}{\theta}} \leq \left(p(\sqrt{\mathbb{E}\|z_m - z_{m-1}\|^2} + \sqrt{\mathbb{E}\|z_{m-1} - z_{m-2}\|^2}) + \sqrt{s\mathbb{E}\Upsilon_{m-1}}\right)^{\frac{1-\theta}{\theta}}. \tag{C.66}$$

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

$$\begin{aligned}&\sum_{k=m}^{\infty} \sqrt{\mathbb{E}\|z_{k+1} - z_k\|^2} + \sqrt{\mathbb{E}\|z_k - z_{k-1}\|^2} \\ &\leq \sqrt{\mathbb{E}\|z_m - z_{m-1}\|^2} + \sqrt{\mathbb{E}\|z_{m-1} - z_{m-2}\|^2} + \frac{2\sqrt{s}}{K_1\rho}\sqrt{\mathbb{E}\Upsilon_{m-1}} + aK_3\Theta_m^{\frac{1-\theta}{\theta}} + \left(\frac{1}{2\bar{L}\lambda\rho}\mathbb{E}\Upsilon_m\right)^{1-\theta} \\ &\quad + \left(\frac{V_1 + V_\Upsilon/\rho}{2\bar{L}\lambda}\mathbb{E}\|z_m - z_{m-1}\|^2\right)^{1-\theta}.\end{aligned} \tag{C.67}$$

Suppose $\theta \in (1/2, 1)$. 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 θ satisfies $\frac{1-\theta}{2\theta} \leq 1-\theta$ and $\frac{1-\theta}{\theta} < 1$, the term $\Theta_m^{\frac{1-\theta}{\theta}}$ is dominant for large m. Precisely, there exists a natural number M_1 such that for all $m \geq M_1$,

$$\left(\sum_{k=m}^{\infty} \sqrt{\mathbb{E}\|z_{k+1} - z_k\|^2} + \sqrt{\mathbb{E}\|z_k - z_{k-1}\|^2}\right)^{\frac{\theta}{1-\theta}} \leq P\Theta_m, \tag{C.68}$$

for some constant $P > (aK_3)^{\frac{\theta}{1-\theta}}$. The bound of (C.41) implies

$$2\sqrt{s\mathbb{E}\Upsilon_{m-1}} \leq \frac{4\sqrt{s}}{\rho}\big(\sqrt{\mathbb{E}\Upsilon_{m-1}} - \sqrt{\mathbb{E}\Upsilon_m} + \sqrt{V_\Upsilon}(\sqrt{\mathbb{E}\|z_m - z_{m-1}\|^2} + \sqrt{\mathbb{E}\|z_{m-1} - z_{m-2}\|^2})\big). \tag{C.69}$$

Therefore,

$$\begin{aligned}
\Theta_m &= p(\sqrt{\mathbb{E}\|z_m - z_{m-1}\|^2} + \sqrt{\mathbb{E}\|z_{m-1} - z_{m-2}\|^2}) + \sqrt{s\mathbb{E}\Upsilon_{m-1}} \\
&\leq \left(p + \frac{4\sqrt{sV_\Upsilon}}{\rho}\right)(\sqrt{\mathbb{E}\|z_m - z_{m-1}\|^2} + \sqrt{\mathbb{E}\|z_{m-1} - z_{m-2}\|^2}) + \frac{4\sqrt{s}}{\rho}(\sqrt{\mathbb{E}\Upsilon_{m-1}} - \sqrt{\mathbb{E}\Upsilon_m}) - \sqrt{s\mathbb{E}\Upsilon_{m-1}}.
\end{aligned} \tag{C.70}$$

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 $m \geq M_2$,

$$\left(\frac{4\sqrt{s}(1-\rho/4)}{\rho(p + 4\sqrt{sV_\Upsilon}/\rho)}\sqrt{\mathbb{E}\Upsilon_m}\right)^{\frac{\theta}{1-\theta}} \leq P\sqrt{s\mathbb{E}\Upsilon_m}. \tag{C.71}$$

Therefore, (C.68) implies

$$\begin{aligned}
&\left(\sum_{k=m}^{\infty} \sqrt{\mathbb{E}\|z_{k+1} - z_k\|^2} + \sqrt{\mathbb{E}\|z_k - z_{k-1}\|^2} + \frac{4\sqrt{s}(1-\rho/4)}{\rho(p + 4\sqrt{sV_\Upsilon}/\rho)}\sqrt{\mathbb{E}\Upsilon_m}\right)^{\frac{\theta}{1-\theta}} \\
&\overset{①}{\leq} \frac{2^{\frac{\theta}{1-\theta}}}{2}\left(\sum_{k=m}^{\infty} \sqrt{\mathbb{E}\|z_{k+1} - z_k\|^2} + \sqrt{\mathbb{E}\|z_k - z_{k-1}\|^2}\right)^{\frac{\theta}{1-\theta}} + \frac{2^{\frac{\theta}{1-\theta}}}{2}\left(\frac{4\sqrt{s}(1-\rho/4)}{\rho(p + 4\sqrt{sV_\Upsilon}/\rho)}\sqrt{\mathbb{E}\Upsilon_m}\right)^{\frac{\theta}{1-\theta}} \\
&\overset{②}{\leq} \frac{2^{\frac{\theta}{1-\theta}}}{2}\left(\sum_{k=m}^{\infty} \sqrt{\mathbb{E}\|z_{k+1} - z_k\|^2} + \sqrt{\mathbb{E}\|z_k - z_{k-1}\|^2}\right)^{\frac{\theta}{1-\theta}} + \frac{2^{\frac{\theta}{1-\theta}}}{2}\left(P\sqrt{s\mathbb{E}\Upsilon_m}\right) \\
&\overset{③}{\leq} \frac{2^{\frac{\theta}{1-\theta}}}{2}\left(P(p + 4\sqrt{sV_\Upsilon}/\rho)\left(\sqrt{\mathbb{E}\left\|z_m - z_{m-1}\right\|^2} + \sqrt{\left\|z_{m-1} - z_{m-2}\right\|^2}\right) + \frac{4\sqrt{s}P(1-\rho/4)}{\rho}\left(\sqrt{\mathbb{E}\Upsilon_{m-1}} - \sqrt{\mathbb{E}\Upsilon_m}\right)\right).
\end{aligned} \tag{C.72}$$

Here, ① follows by convexity of the function $x^{\frac{\theta}{1-\theta}}$ for $\theta \in [1/2, 1)$, ② is (C.71), and ③ is (C.68) combined with (C.70). We absorb the constant $\frac{2^{\frac{\theta}{1-\theta}}}{2}$ into P. With

$$S_m \overset{\text{def}}{=} \sum_{k=m}^{\infty} \sqrt{\mathbb{E}\|z_{k+1} - z_k\|^2} + \sqrt{\mathbb{E}\|z_k - z_{k-1}\|^2} + \frac{4\sqrt{s}P(1-\rho/4)}{\rho(p + 4\sqrt{sV_\Upsilon}/\rho)}\sqrt{\mathbb{E}\Upsilon_m}, \tag{C.73}$$

we have shown

$$S_m^{\frac{\theta}{1-\theta}} \leq P(p + 4\sqrt{sV_\Upsilon}/\rho)(S_{m-1} - S_m), \tag{C.74}$$

The rest of the proof follows the proof of (Attouch & Bolte, 2007, Thm. 5). Let $h(r) \overset{\text{def}}{=} r^{-\frac{\theta}{1-\theta}}$. First, suppose that $h(S_m) \leq Rh(S_{m-1})$ for some $R \in (1, \infty)$. Then (C.74) ensures that

$$\begin{aligned}
1 &\leq P(p + 4\sqrt{sV_\Upsilon}/\rho)(S_{m-1} - S_m)h(S_m) \\
&\leq RP(p + 4\sqrt{sV_\Upsilon}/\rho)(S_{m-1} - S_m)h(S_{m-1}) \\
&\leq RP(p + 4\sqrt{sV_\Upsilon}/\rho)\int_{S_m}^{S_{m-1}} h(r)dr \\
&= \frac{RP(p + 4\sqrt{sV_\Upsilon}/\rho)(1-\theta)}{1-2\theta}\left[S_{m-1}^{\frac{1-2\theta}{1-\theta}} - S_m^{\frac{1-2\theta}{1-\theta}}\right].
\end{aligned} \tag{C.75}$$

Hence,

$$0 < -\frac{1-2\theta}{RP(p+4\sqrt{sV_\Upsilon}/\rho)(1-\theta)} \le S_m^{\frac{1-2\theta}{1-\theta}} - S_{m-1}^{\frac{1-2\theta}{1-\theta}}. \tag{C.76}$$

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

$$\left(q^{\frac{1-2\theta}{1-\theta}} - 1\right) S_{m-1}^{\frac{1-2\theta}{1-\theta}} \le S_m^{\frac{1-2\theta}{1-\theta}} - S_{m-1}^{\frac{1-2\theta}{1-\theta}}, \tag{C.77}$$

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

$$\mu \le S_m^{\frac{1-2\theta}{1-\theta}} - S_{m-1}^{\frac{1-2\theta}{1-\theta}}. \tag{C.78}$$

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

$$S_M \le \left(S_{M_2-1}^{\frac{1-2\theta}{1-\theta}} + (M - M_2)\mu\right)^{\frac{1-\theta}{1-2\theta}} \le dM^{\frac{1-\theta}{1-2\theta}}, \tag{C.79}$$

for some constant d. By Jensen's inequality and the fact that z_k converges to z^*, we can say $\mathbb{E}\|z_k - z^*\| \le \sum_{k=m}^{\infty} \mathbb{E}\|z_k - z_{k-1}\| \le S_M \le dM_2^{-\frac{1-\theta}{2\theta-1}}$, proving Claim 1.

If $\theta = 1/2$, then $\|\zeta_m\|^{\frac{1-\theta}{\theta}} = \|\zeta_m\|$. Equation (C.67) then reads

$$\begin{aligned}
&\sum_{i=m}^{\infty} \sqrt{\mathbb{E}\|z_{k+1} - z_k\|^2} + \sqrt{\mathbb{E}\|z_k - z_{k-1}\|^2} \\
&\le \left(1 + aK_3(p + \frac{4\sqrt{sV_\Upsilon}}{\rho}) + \sqrt{\frac{V_1 + V_\Upsilon/\rho}{2\bar{L}\lambda}}\right)\left(\sqrt{\mathbb{E}\|z_m - z_{m-1}\|^2} + \sqrt{\mathbb{E}\|z_{m-1} - z_{m-2}\|^2}\right) \\
&\quad + \frac{2\sqrt{s}}{K_1\rho}\sqrt{\mathbb{E}\Upsilon_{m-1}} + \left(aK_3\sqrt{s} + \sqrt{\frac{1}{2\bar{L}\lambda\rho}}\right)\sqrt{\mathbb{E}\Upsilon_m}.
\end{aligned} \tag{C.80}$$

Using equation (C.41), we have that, for any constant $c > 0$,

$$0 \le -c\sqrt{\mathbb{E}\Upsilon_m} + c\left(1 - \frac{\rho}{2}\right)\sqrt{\mathbb{E}\Upsilon_{m-1}} + c\sqrt{V_\Upsilon}(\sqrt{\mathbb{E}\|z_m - z_{m-1}\|^2} + \sqrt{\mathbb{E}\|z_{m-1} - z_{m-2}\|^2}). \tag{C.81}$$

Combining this inequality with (C.80),

$$\begin{aligned}
&\sum_{i=m}^{\infty} \sqrt{\mathbb{E}\|z_{k+1} - z_k\|^2} + \sqrt{\mathbb{E}\|z_k - z_{k-1}\|^2} \\
&\le \left(1 + aK_3(p + \frac{4\sqrt{sV_\Upsilon}}{\rho}) + \sqrt{\frac{V_1 + V_\Upsilon/\rho}{2\bar{L}\lambda}} + c\sqrt{V_\Upsilon}\right)\left(\sqrt{\mathbb{E}\|z_m - z_{m-1}\|^2} + \sqrt{\mathbb{E}\|z_{m-1} - z_{m-2}\|^2}\right) \\
&\quad + c\left(1 - \frac{\rho}{2} + \frac{2\sqrt{s}}{cK_1\rho}\right)\sqrt{\mathbb{E}\Upsilon_{m-1}} - c\left(1 - c^{-1}\left(aK_3\sqrt{s} + \sqrt{\frac{1}{2\bar{L}\lambda\rho}}\right)\right)\sqrt{\mathbb{E}\Upsilon_m}.
\end{aligned} \tag{C.82}$$

Defining

$$T_m \stackrel{\text{def}}{=} \sum_{i=m}^{\infty} \sqrt{\mathbb{E}\|z_{i+1} - z_i\|^2} + \sqrt{\mathbb{E}\|z_i - z_{i-1}\|^2}, \tag{C.83}$$

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

$$T_m + c\left(1 - c^{-1}\left(aK_3\sqrt{s} + \sqrt{\frac{1}{2\bar{L}\lambda\rho}}\right)\right)\sqrt{\mathbb{E}\Upsilon_m} \le P_2(T_{m-1} - T_m) + c\left(1 - \frac{\rho}{2} + \frac{2\sqrt{s}}{cK_1\rho}\right)\sqrt{\mathbb{E}\Upsilon_{m-1}}. \tag{C.84}$$

Rearranging,

$$(1+P_2)T_m + c\left(1 - c^{-1}\left(aK_3\sqrt{s} + \sqrt{\frac{1}{2\bar{L}\lambda\rho}}\right)\right)\sqrt{\mathbb{E}\Upsilon_m} \le P_2 T_{m-1} + c\left(1 - \frac{\rho}{2} + \frac{2\sqrt{s}}{cK_1\rho}\right)\sqrt{\mathbb{E}\Upsilon_{m-1}}. \tag{C.85}$$

This implies

$$T_m + \sqrt{\mathbb{E}\Upsilon_m} \le \max\left\{\frac{P_2}{1+P_2}, \left(1 - \frac{\rho}{2} + \frac{2\sqrt{s}}{cK_1\rho}\right)\left(1 - c^{-1}\left(aK_3\sqrt{s} + \sqrt{\frac{1}{2\bar{L}\lambda\rho}}\right)\right)^{-1}\right\}(T_{m-1} + \sqrt{\mathbb{E}\Upsilon_{m-1}}). \tag{C.86}$$

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 (2.5) implies that exactly one of the following two scenarios holds: either $\mathbb{E}\Phi(z_k) \neq \Phi(z^*)$ and

$$0 < C_0 \le \mathbb{E}\|\zeta_k\| \quad \forall \zeta_k \in \mathbb{E}\partial\Phi(z_k), \tag{C.87}$$

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

Using the subgradient bound, the first scenario implies

$$\begin{aligned} C_0^2 &\le (\mathbb{E}\|\zeta_k\|)^2 \\ &\le (p\mathbb{E}\|z_k - z_{k-1}\| + p\mathbb{E}\|z_{k-1} - z_{k-2}\| + \mathbb{E}\Gamma_{k-1})^2, \\ &\le 3p^2(\mathbb{E}\|z_k - z_{k-1}\|)^2 + 3p^2(\mathbb{E}\|z_{k-1} - z_{k-2}\|)^2 + 3(\mathbb{E}\Gamma_{k-1})^2, \\ &\le 3p^2\mathbb{E}\|z_k - z_{k-1}\|^2 + 3p^2\mathbb{E}\|z_{k-1} - z_{k-2}\|^2 + 3s\mathbb{E}\Upsilon_{k-1}. \end{aligned} \tag{C.88}$$

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 Ψ_k (C.3), we obtain

$$\begin{aligned} \mathbb{E}\Psi_k &\le \mathbb{E}\Psi_{k-1} + \left(\frac{\bar{L}(\lambda+1)}{2} + \frac{V_1 + V_\Gamma/\rho}{2\bar{L}\lambda} + Z - \frac{1}{2\eta}\right)\mathbb{E}\|z_k - z_{k-1}\|^2 - Z\mathbb{E}\|z_{k-1} - z_{k-2}\|^2 \\ &\le \mathbb{E}\Psi_{k-1} - C_0^2 + \mathcal{O}(\mathbb{E}\|z_k - z_{k-1}\|^2) + \mathcal{O}(\mathbb{E}\|z_{k-1} - z_{k-2}\|^2) + \mathcal{O}(\mathbb{E}\Upsilon_{k-1}), \end{aligned} \tag{C.89}$$

for some constant C_0^2.[4] Because the final three terms go to zero as $k \to \infty$, there exists an index M_3 so that the sum of these three terms is bounded above by $C_0^2/2$ for all $k \ge M_3$. Therefore,

$$\mathbb{E}\Psi_k \le \mathbb{E}\Psi_{k-1} - \frac{C_0^2}{2}, \quad \forall k \ge M_3. \tag{C.90}$$

Because Ψ_k is bounded below for all k, this inequality can only hold for $N < \infty$ steps. After N steps, it is no longer possible for the bound (C.87) to hold, so it must be that $\Phi(z_k) = \Phi(z^*)$. □

D. SAGA Variance Bound

We define the SAGA gradient estimators $\widetilde{\nabla}_x^{\text{SAGA}}$ and $\widetilde{\nabla}_y^{\text{SAGA}}$ as follows:

$$\begin{aligned} \widetilde{\nabla}_x^{\text{SAGA}}(x_k, y_k) &= \frac{1}{b}\left(\sum_{j \in J_k^x} \nabla_x F_j(x_k, y_k) - \nabla_x F_j(\varphi_k^j, y_k)\right) + \frac{1}{n}\sum_{i=1}^n \nabla_x F_i(\varphi_k^i, y_k) \\ \widetilde{\nabla}_y^{\text{SAGA}}(x_{k+1}, y_k) &= \frac{1}{b}\left(\sum_{j \in J_k^y} \nabla_y F_j(x_{k+1}, y_k) - \nabla_x F_j(x_{k+1}, \xi_k^j)\right) + \frac{1}{n}\sum_{i=1}^n \nabla_x F_i(x_{k+1}, \xi_k^i), \end{aligned} \tag{D.1}$$

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

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

[4] We have ignored extraneous constants in the final three terms for clarity.

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

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

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

$$\mathbb{E}_k \|X_1 + \cdots + X_t\|^2 = \sum_{i,j=1}^{t} \mathbb{E}_k\langle X_i, X_j\rangle = \mathbb{E}_k[\|X_1\|^2 + \cdots + \|X_t\|^2]. \tag{D.3}$$

□

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

Lemma D.2. *The SAGA gradient estimator satisfies*

$$\begin{aligned} \mathbb{E}_k \|\widetilde{\nabla}_x^{\text{SAGA}}(x_k, y_k) - \nabla_x F(x_k, y_k)\|^2 &\le \frac{1}{bn} \sum_{i=1}^{n} \left\|\nabla_x F_i(x_k, y_k) - \nabla_x F_i(\varphi_k^i, y_k)\right\|^2, \\ \mathbb{E}_k \|\widetilde{\nabla}_y^{\text{SAGA}}(x_{k+1}, y_k) - \nabla_y F(x_{k+1}, y_k)\|^2 &\le \frac{4}{bn} \sum_{i=1}^{n} \left\|\nabla_y F_i(x_k, y_k) - \nabla_y F_i(x_k, \xi_k^i)\right\|^2 \\ &\quad + \frac{6M^2}{b} \mathbb{E}_k \|x_{k+1} - x_k\|^2, \end{aligned} \tag{D.4}$$

as well as

$$\begin{aligned} \mathbb{E}_k \|\widetilde{\nabla}_x^{\text{SAGA}}(x_k, y_k) - \nabla_x F(x_k, y_k)\| &\le \frac{1}{\sqrt{bn}} \sum_{i=1}^{n} \left\|\nabla_x F_i(x_k, y_k) - \nabla_x F_i(\varphi_k^i, y_k)\right\|, \\ \mathbb{E}_k \|\widetilde{\nabla}_y^{\text{SAGA}}(x_{k+1}, y_k) - \nabla_y F(x_{k+1}, y_k)\| &\le \frac{2}{\sqrt{bn}} \sum_{i=1}^{n} \left\|\nabla_y F_i(x_k, y_k) - \nabla_y F_i(x_k, \xi_k^i)\right\| \\ &\quad + \frac{\sqrt{6}M}{\sqrt{b}} \mathbb{E}_k \|x_{k+1} - x_k\|. \end{aligned} \tag{D.5}$$

Proof. The proof amounts to computing expectations and applying the Lipschitz continuity of $\nabla_x F_i$.

$$\begin{aligned} &\mathbb{E}_k \|\widetilde{\nabla}_x^{\text{SAGA}}(x_k, y_k) - \nabla_x F(x_k, y_k)\|^2 \\ =&\mathbb{E}_k \left\| \frac{1}{b} \sum_{j \in J_k^x} \Big(\nabla_x F_j(x_k, y_k) - \nabla_x F_j(\varphi_k^j, y_k)\Big) - \nabla_x F(x_k, y_k) + \frac{1}{n}\sum_{i=1}^{n} \nabla_x F_i(\varphi_k^i, y_k) \right\|^2 \\ \overset{①}{\le}& \frac{1}{b^2} \mathbb{E}_k \sum_{j \in J_k^x} \left\|\nabla_x F_j(x_k, y_k) - \nabla_x F_j(\varphi_k^j, y_k)\right\|^2 \\ =&\frac{1}{bn} \sum_{i=1}^{n} \left\|\nabla_x F_i(x_k, y_k) - \nabla_x F_i(\varphi_k^i, y_k)\right\|^2. \end{aligned} \tag{D.6}$$

Inequality ① follows from Lemma D.1. We can also say that

$$\begin{aligned} \mathbb{E}_k \|\widetilde{\nabla}_x^{\text{SAGA}}(x_k, y_k) - \nabla_x F(x_k, y_k)\| &\overset{①}{\le} \sqrt{\mathbb{E}_k \|\widetilde{\nabla}_x^{\text{SAGA}}(x_k, y_k) - \nabla_x F(x_k, y_k)\|^2} \\ &\le \frac{1}{\sqrt{bn}} \sqrt{\sum_{i=1}^{n} \left\|\nabla_x F_i(x_k, y_k) - \nabla_x F_i(\varphi_k^i, y_k)\right\|^2} \\ &\le \frac{1}{\sqrt{bn}} \sum_{i=1}^{n} \left\|\nabla_x F_i(x_k, y_k) - \nabla_x F_i(\varphi_k^i, y_k)\right\|. \end{aligned} \tag{D.7}$$

Inequality ① is Jensen's.

We use an analogous argument for $\widetilde{\nabla}_y^{\text{SAGA}}$. Let $\mathbb{E}_{k,x}$ denote the expectation conditional on the first k iterations and J_k^x. By the same reasoning as in (D.6),

$$\mathbb{E}_{k,x}\|\widetilde{\nabla}_y^{\text{SAGA}}(x_{k+1},y_k)-\nabla_y F(x_{k+1},y_k)\|^2 \leq \frac{1}{bn}\sum_{i=1}^n \left\|\nabla_y F_i(x_{k+1},y_k)-\nabla_y F_i(x_{k+1},\xi_k^i)\right\|^2. \tag{D.8}$$

Applying the Lipschitz continuity of $\nabla_y F_i$,

$$\begin{aligned}
&\frac{1}{bn}\sum_{i=1}^n \left\|\nabla_y F_i(x_{k+1},y_k)-\nabla_y F_i(x_{k+1},\xi_k^i)\right\|^2\\
&\leq \frac{2}{bn}\sum_{i=1}^n \|\nabla_y F_i(x_{k+1},y_k)-\nabla_y F_i(x_k,y_k)\|^2+\frac{2}{bn}\sum_{i=1}^n \left\|\nabla_y F_i(x_k,y_k)-\nabla_y F_i(x_{k+1},\xi_k^i)\right\|^2\\
&\leq \frac{2M^2}{b}\|x_{k+1}-x_k\|^2+\frac{4}{bn}\sum_{i=1}^n \left\|\nabla_y F_i(x_k,\xi_k^i)-\nabla_y F_i(x_{k+1},\xi_k^i)\right\|^2+\frac{4}{bn}\sum_{i=1}^n \left\|\nabla_y F_i(x_k,y_k)-\nabla_y F_i(x_k,\xi_k^i)\right\|^2\\
&\leq \frac{2M^2}{b}\|x_{k+1}-x_k\|^2+\frac{4M^2}{b}\|x_k-x_{k+1}\|^2+\frac{4}{bn}\sum_{i=1}^n \left\|\nabla_y F_i(x_k,y_k)-\nabla_y F_i(x_k,\xi_k^i)\right\|^2.
\end{aligned} \tag{D.9}$$

Also, by the same reasoning as in (D.7),

$$\begin{aligned}
\mathbb{E}_{k,x}\|\widetilde{\nabla}_y^{\text{SAGA}}(x_{k+1},y_k)-\nabla_y F(x_{k+1},y_k)\| &\overset{①}{\leq}\sqrt{\mathbb{E}_{k,x}\|\widetilde{\nabla}_y^{\text{SAGA}}(x_{k+1},y_k)-\nabla_x F(x_{k+1},y_k)\|^2}\\
&\leq\sqrt{\frac{4}{bn}\sum_{i=1}^n \left\|\nabla_y F_i(x_k,y_k)-\nabla_y F_i(x_k,\xi_k^i)\right\|^2+\frac{6M^2}{b}\|x_{k+1}-x_k\|^2}\\
&\leq\frac{2}{\sqrt{bn}}\sum_{i=1}^n \left\|\nabla_y F_i(x_k,y_k)-\nabla_y F_i(x_k,\xi_k^i)\right\|+\frac{\sqrt{6}M}{\sqrt{b}}\|x_{k+1}-x_k\|.
\end{aligned} \tag{D.10}$$

Applying the operator $\mathbb{E}_k$ to these two inequalities gives the desired result. □

Lemma D.3. *The SAGA gradient estimator is variance-reduced with*

$$\begin{aligned}
\Upsilon_{k+1}&=\frac{1}{bn}\left(\sum_{i=1}^n \|\nabla_x F_i(x_{k+1},y_{k+1})-\nabla_x F_i(\varphi_{k+1}^i,y_{k+1})\|^2+4\|\nabla_y F_i(x_{k+1},y_{k+1})-\nabla_y F_i(x_{k+1},\xi_{k+1}^i)\|^2\right),\\
\Gamma_{k+1}&=\frac{1}{\sqrt{bn}}\left(\sum_{i=1}^n \|\nabla_x F_i(x_{k+1},y_{k+1})-\nabla_x F_i(\varphi_{k+1}^i,y_{k+1})\|+2\|\nabla_y F_i(x_{k+1},y_{k+1})-\nabla_y F_i(x_{k+1},\xi_{k+1}^i)\|\right),
\end{aligned} \tag{D.11}$$

and constants $V_1=6M^2/b$, $V_2=\sqrt{6}M/\sqrt{b}$, $V_\Upsilon=\frac{115nL^2}{b^2}$, *and* $\rho=\frac{b}{2n}$.

Proof. We must show that $\mathbb{E}_k\Upsilon_{k+1}$ decreases at a geometric rate. We first bound the MSE of the estimator $\widetilde{\nabla}_x^{\text{SAGA}}$. Applying the inequality $\|a-c\|^2 \leq (1+\delta)\|a-b\|^2 + (1+\delta^{-1})\|b-c\|^2$ twice,

$$\begin{aligned}
&\frac{1}{bn}\sum_{i=1}^{n}\mathbb{E}_k\left\|\nabla_x F_i(x_{k+1},y_{k+1}) - \nabla_x F_i(\varphi_{k+1}^i, y_{k+1})\right\|^2 \\
&\leq \frac{1+\delta}{bn}\mathbb{E}_k\sum_{i=1}^{n}\left\|\nabla_x F_i(x_k,y_k) - \nabla_x F_i(\varphi_{k+1}^i, y_{k+1})\right\|^2 + \frac{\delta^{-1}-1}{bn}\sum_{i=1}^{n}\|\nabla_x F_i(x_{k+1},y_{k+1}) - \nabla_x F_i(x_k,y_k)\|^2 \\
&\leq \frac{(1+\delta)^2}{bn}\mathbb{E}_k\sum_{i=1}^{n}\left\|\nabla_x F_i(x_k,y_k) - \nabla_x F_i(\varphi_{k+1}^i, y_k)\right\|^2 + \frac{(\delta^{-1}-1)(1+\delta)}{bn}\mathbb{E}_k\sum_{i=1}^{n}\left\|\nabla_x F_i(\varphi_{k+1},y_{k+1}) - \nabla_x F_i(\varphi_{k+1}^i, y_k)\right\|^2 \\
&\quad + \frac{\delta^{-1}-1}{bn}\sum_{i=1}^{n}\|\nabla_x F_i(x_{k+1},y_{k+1}) - \nabla_x F_i(x_k,y_k)\|^2.
\end{aligned} \tag{D.12}$$

Next, we compute the expectation of the first term.

$$\begin{aligned}
&\leq \tfrac{(1+\delta)^2(1-b/n)}{bn}\sum_{i=1}^{n}\left\|\nabla_x F_i(x_k,y_k) - \nabla_x F_i(\varphi_k^i, y_k)\right\|^2 + \tfrac{(\delta^{-1}-1)(1+\delta)}{bn}\mathbb{E}_k\sum_{i=1}^{n}\left\|\nabla_x F_i(\varphi_{k+1}^i,y_{k+1}) - \nabla_x F_i(\varphi_{k+1}^i, y_k)\right\|^2 \\
&\quad + \tfrac{\delta^{-1}-1}{bn}\sum_{i=1}^{n}\|\nabla_x F_i(x_{k+1},y_{k+1}) - \nabla_x F_i(x_k,y_k)\|^2 \\
&\leq \tfrac{(1+\delta)^2(1-b/n)}{bn}\sum_{i=1}^{n}\left\|\nabla_x F_i(x_k,y_k) - \nabla_x F_i(\varphi_k^i, y_k)\right\|^2 + \tfrac{(\delta^{-1}-1)(1+\delta)M^2}{b}\mathbb{E}_k\left\|y_{k+1}-y_k\right\|^2 \\
&\quad + \tfrac{(\delta^{-1}-1)M^2}{b}\mathbb{E}_k\|z_{k+1}-z_k\|^2.
\end{aligned} \tag{D.13}$$

We bound the MSE of the estimator $\widetilde{\nabla}_y^{\text{SAGA}}$ similarly.

$$\begin{aligned}
&\tfrac{1}{bn}\sum_{i=1}^{n}\mathbb{E}_k\left\|\nabla_y F_i(x_{k+1},y_{k+1}) - \nabla_y F_i(x_{k+1}, \xi_{k+1}^i)\right\|^2 \\
&\leq \tfrac{1+\delta}{bn}\mathbb{E}_k\sum_{i=1}^{n}\left\|\nabla_y F_i(x_{k+1},y_k) - \nabla_y F_i(x_{k+1}, \xi_{k+1}^i)\right\|^2 + \tfrac{\delta^{-1}-1}{bn}\mathbb{E}_k\sum_{i=1}^{n}\|\nabla_y F_i(x_{k+1},y_{k+1}) - \nabla_y F_i(x_{k+1},y_k)\|^2 \\
&= \tfrac{(1+\delta)(1-b/n)}{bn}\mathbb{E}_k\sum_{i=1}^{n}\left\|\nabla_y F_i(x_{k+1},y_k) - \nabla_y F_i(x_{k+1}, \xi_k^i)\right\|^2 + \tfrac{\delta^{-1}-1}{bn}\mathbb{E}_k\sum_{i=1}^{n}\|\nabla_y F_i(x_{k+1},y_{k+1}) - \nabla_y F_i(x_{k+1},y_k)\|^2 \\
&\leq \tfrac{(1+\delta)^2(1-b/n)}{bn}\mathbb{E}_k\sum_{i=1}^{n}\left\|\nabla_y F_i(x_k,y_k) - \nabla_y F_i(x_{k+1}, \xi_k^i)\right\|^2 + \tfrac{\delta^{-1}-1}{bn}\mathbb{E}_k\sum_{i=1}^{n}\|\nabla_y F_i(x_{k+1},y_{k+1}) - \nabla_y F_i(x_{k+1},y_k)\|^2 \\
&\quad + \tfrac{(1+\delta)(\delta^{-1}-1)(1-b/n)}{bn}\mathbb{E}_k\sum_{i=1}^{n}\|\nabla_y F_i(x_{k+1},y_k) - \nabla_y F_i(x_k,y_k)\|^2 \\
&\leq \tfrac{(1+\delta)^3(1-b/n)}{bn}\mathbb{E}_k\sum_{i=1}^{n}\left\|\nabla_y F_i(x_k,y_k) - \nabla_y F_i(x_k, \xi_k^i)\right\|^2 + \tfrac{\delta^{-1}-1}{bn}\mathbb{E}_k\sum_{i=1}^{n}\|\nabla_y F_i(x_{k+1},y_{k+1}) - \nabla_y F_i(x_{k+1},y_k)\|^2 \\
&\quad + \tfrac{(1+\delta)(\delta^{-1}-1)(1-b/n)}{bn}\mathbb{E}_k\sum_{i=1}^{n}\|\nabla_y F_i(x_{k+1},y_k) - \nabla_y F_i(x_k,y_k)\|^2 \\
&\quad + \tfrac{(1+\delta)^2(\delta^{-1}-1)(1-b/n)}{bn}\mathbb{E}_k\sum_{i=1}^{n}\|\nabla_y F_i(x_{k+1},\xi_k^i) - \nabla_y F_i(x_k,\xi_k^i)\|^2,
\end{aligned} \tag{D.14}$$

and, by the Lipschitz continuity of $\nabla_y F_i$,

$$\begin{aligned}
&\leq \frac{(1+\delta)^3(1-b/n)}{bn}\mathbb{E}_k \sum_{i=1}^n \left\|\nabla_y F_i(x_k,y_k) - \nabla_y F_i(x_k,\xi_k^i)\right\|^2 + \frac{(\delta^{-1}-1)L_y^2}{b}\mathbb{E}_k\|y_{k+1}-y_k\|^2 \\
&\quad + \frac{(1+\delta)(\delta^{-1}-1)(1-b/n)M^2}{b}\mathbb{E}_k\|x_{k+1}-x_k\|^2 \\
&\quad + \frac{(1+\delta)^2(\delta^{-1}-1)(1-b/n)M^2}{b}\mathbb{E}_k\|x_{k+1}-x_k\|^2.
\end{aligned} \tag{D.15}$$

With

$$\Upsilon_{k+1} = \frac{1}{bn}\left(\sum_{i=1}^n \|\nabla_x F_i(x_{k+1},y_{k+1}) - \nabla_x F_i(\varphi_{k+1}^i, y_{k+1})\|^2 + 4\|\nabla_y F_i(x_{k+1},y_{k+1}) - \nabla_y F_i(x_{k+1},\xi_{k+1}^i)\|^2\right), \tag{D.16}$$

we can now say

$$\begin{aligned}
\mathbb{E}_k\Upsilon_{k+1} \leq &(1+\delta)^3(1-b/n)\Upsilon_k + \frac{4(\delta^{-1}-1)L_y^2}{b}\mathbb{E}_k\|y_{k+1}-y_k\|^2 \\
&+ \frac{8(1+\delta)^2(\delta^{-1}-1)(1-b/n)M^2}{b}\mathbb{E}_k\|x_{k+1}-x_k\|^2 \\
&+ \frac{(1+\delta)(\delta^{-1}-1)M^2}{b}\mathbb{E}_k\|y_{k+1}-y_k\|^2 + \frac{(\delta^{-1}-1)M^2}{b}\mathbb{E}_k\|z_{k+1}-z_k\|^2 \\
\leq &(1+\delta)^3(1-b/n)\Upsilon_k + \frac{14(1+\delta)^2\delta^{-1}L^2}{b}\mathbb{E}_k[\|z_{k+1}-z_k\|^2],
\end{aligned} \tag{D.17}$$

where $L \stackrel{\text{def}}{=} \max\{L_x, L_y, M\}$. Choosing $\delta = \frac{b}{6n}$, we are ensured that $(1+\delta)^3(1-b/n) \leq 1 - \frac{b}{2n}$, producing the inequality

$$\begin{aligned}
\mathbb{E}_k\Upsilon_{k+1} &\leq (1-\frac{b}{2n})\Upsilon_k + \frac{84(1+\frac{b}{6n})^2 nL^2}{b^2}\mathbb{E}_k[\|z_{k+1}-z_k\|^2] \\
&\leq (1-\frac{b}{2n})\Upsilon_k + \frac{115nL^2}{b^2}\mathbb{E}_k[\|z_{k+1}-z_k\|^2].
\end{aligned} \tag{D.18}$$

This proves the geometric decay of Υ_k in expectation.

All that is left is to show that if $\mathbb{E}\|z_k - z_{k-1}\|^2 \to 0$, then so do Υ_k and Γ_k. We begin by showing that $\sum_{i=1}^n \mathbb{E}\|\nabla_x F_i(x_k,y_k) - \nabla_x F_i(\varphi_k^i, y_k)\|^2 \to 0$.

$$\begin{aligned}
\sum_{i=1}^n \mathbb{E}\|\nabla_x F_i(x_k,y_k) - \nabla_x F_i(\varphi_k^i,y_k)\|^2 &\leq L_x^2 \sum_{i=1}^n \mathbb{E}\|x_k - \varphi_k^i\|^2 \\
&\leq L_x^2 n\left(1+\frac{2n}{b}\right)\mathbb{E}\|x_k - x_{k-1}\|^2 + \left(1+\frac{b}{2n}\right)\sum_{i=1}^n \mathbb{E}\|x_{k-1}-\varphi_k^i\|^2 \\
&\leq L_x^2 n\left(1+\frac{2n}{b}\right)\mathbb{E}\|x_k - x_{k-1}\|^2 + \left(1+\frac{b}{2n}\right)\left(1-\frac{b}{n}\right)\sum_{i=1}^n \mathbb{E}\|x_{k-1}-\varphi_{k-1}^i\|^2 \\
&\leq L_x^2 n\left(1+\frac{2n}{b}\right)\mathbb{E}\|x_k - x_{k-1}\|^2 + \left(1-\frac{b}{2n}\right)\sum_{i=1}^n \mathbb{E}\|x_{k-1}-\varphi_{k-1}^i\|^2 \\
&\leq L_x^2 n\left(1+\frac{2n}{b}\right)\sum_{\ell=1}^k \left(1-\frac{b}{2n}\right)^{k-\ell}\mathbb{E}\|x_\ell - x_{\ell-1}\|^2.
\end{aligned} \tag{D.19}$$

Because $\|x_k - x_{k-1}\|^2 \to 0$, it is clear that the bound on the right goes to zero as $k \to \infty$. An analogous argument shows that $\sum_{i=1}^n \mathbb{E}\|\nabla_x F_i(x_k, y_k) - \nabla_x F_i(x_k, \xi_k^i)\|^2 \to 0$ as well. The fact that $\Gamma_k \to 0$ follows similarly:

$$\begin{aligned}\sum_{i=1}^n \mathbb{E}\|\nabla_x F_i(x_k, y_k) - \nabla_x F_i(\varphi_k^i, y_k)\| &\le L_x \sum_{i=1}^n \mathbb{E}\|x_k - \varphi_k^i\| \\ &\le nL_x\|x_k - x_{k-1}\| + \sum_{i=1}^n \mathbb{E}\|x_{k-1} - \varphi_k^i\| \\ &\le nL_x\|x_k - x_{k-1}\| + \left(1 - \frac{b}{n}\right)\sum_{i=1}^n \mathbb{E}\|x_{k-1} - \varphi_{k-1}^i\| \\ &\le nL_x \sum_{\ell=1}^k \left(1 - \frac{b}{n}\right)^{k-\ell} \mathbb{E}\|x_\ell - x_{\ell-1}\|.\end{aligned} \tag{D.20}$$

As $\|x_k - x_{k-1}\|^2 \to 0$, it follows that $\|x_k - x_{k-1}\| \to 0$ (because Jensen's inequality implies $\mathbb{E}\|x_k - x_{k-1}\| \le \sqrt{\mathbb{E}\|x_k - x_{k-1}\|^2} \to 0$), so the bound above implies $\Gamma_k \to 0$ as well. $\square$

E. SARAH Variance Bound

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

Lemma E.1. *The SARAH gradient estimator is variance reduced with*

$$\begin{aligned}\Upsilon_{k+1} &= \|\widetilde{\nabla}_x^{\text{SARAH}}(x_k, y_k) - \nabla_x F(x_k, y_k)\|^2 + \|\widetilde{\nabla}_y^{\text{SARAH}}(x_{k+1}, y_k) - \nabla_y F(x_{k+1}, y_k)\|^2, \\ \Gamma_{k+1} &= \|\widetilde{\nabla}_x^{\text{SARAH}}(x_k, y_k) - \nabla_x F(x_k, y_k)\| + \|\widetilde{\nabla}_y^{\text{SARAH}}(x_{k+1}, y_k) - \nabla_y F(x_{k+1}, y_k)\|,\end{aligned} \tag{E.1}$$

and constants $\rho = 1/p$, $V_1 = V_\Upsilon = 2L^2$, and $V_2 = 2L$.

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

$$\begin{aligned}\mathbb{E}_{k,p}\widetilde{\nabla}_x^{\text{SARAH}}(x_k, y_k) &= \frac{1}{b}\mathbb{E}_{k,p}\left(\sum_{j \in J_k^x} \nabla_x F_j(x_k, y_k) - F_j(x_{k-1}, y_{k-1})\right) + \widetilde{\nabla}_x^{\text{SARAH}}(x_{k-1}, y_{k-1}) \\ &= \nabla_x F(x_k, y_k) - \nabla_x F(x_{k-1}, y_{k-1}) + \widetilde{\nabla}_x^{\text{SARAH}}(x_{k-1}, y_{k-1}).\end{aligned} \tag{E.2}$$

We begin with a bound on $\mathbb{E}_{k,p}\|\widetilde{\nabla}_x^{\text{SARAH}}(x_k, y_k) - \nabla_x F(x_k, y_k)\|^2$.

$$\begin{aligned}&\mathbb{E}_{k,p}\|\widetilde{\nabla}_x^{\text{SARAH}}(x_k, y_k) - \nabla_x F(x_k, y_k)\|^2 \\ =&\mathbb{E}_{k,p}\|\widetilde{\nabla}_x^{\text{SARAH}}(x_{k-1}, y_{k-1}) - \nabla_x F(x_{k-1}, y_{k-1}) + \nabla_x F(x_{k-1}, y_{k-1}) - \nabla_x F(x_k, y_k) \\ &\qquad + \widetilde{\nabla}_x^{\text{SARAH}}(x_k, y_k) - \widetilde{\nabla}_x^{\text{SARAH}}(x_{k-1}, y_{k-1})\|^2 \\ =& \left\|\widetilde{\nabla}_x^{\text{SARAH}}(x_{k-1}, y_{k-1}) - \nabla_x F(x_{k-1}, y_{k-1})\right\|^2 + \|\nabla_x F(x_{k-1}, y_{k-1}) - \nabla_x F(x_k, y_k)\|^2 \\ &+ \mathbb{E}_{k,p}\left\|\widetilde{\nabla}_x^{\text{SARAH}}(x_k, y_k) - \widetilde{\nabla}_x^{\text{SARAH}}(x_{k-1}, y_{k-1})\right\|^2 \\ &+ 2\langle \nabla_x F(x_{k-1}, y_{k-1}) - \widetilde{\nabla}_x^{\text{SARAH}}(x_{k-1}, y_{k-1}), \nabla_x F(x_k, y_k) - \nabla_x F(x_{k-1}, y_{k-1})\rangle \\ &- 2\left\langle \nabla_x F(x_{k-1}, y_{k-1}) - \widetilde{\nabla}_x^{\text{SARAH}}(x_{k-1}, y_{k-1}), \mathbb{E}_{k,p}\left[\widetilde{\nabla}_x^{\text{SARAH}}(x_k, y_k) - \widetilde{\nabla}_x^{\text{SARAH}}(x_{k-1}, y_{k-1})\right]\right\rangle \\ &- 2\left\langle \nabla_x F(x_k, y_k) - \nabla_x F(x_{k-1}, y_{k-1}), \mathbb{E}_{k,p}\left[\widetilde{\nabla}_x^{\text{SARAH}}(x_k, y_k) - \widetilde{\nabla}_x^{\text{SARAH}}(x_{k-1}, y_{k-1})\right]\right\rangle.\end{aligned}$$

To simplify the inner-product terms, we use the fact that

$$\mathbb{E}_{k,p}[\widetilde{\nabla}_x^{\text{SARAH}}(x_k, y_k) - \widetilde{\nabla}_x^{\text{SARAH}}(x_{k-1}, y_{k-1})] = \nabla_x F(x_k, y_k) - \nabla_x F(x_{k-1}, y_{k-1}). \tag{E.3}$$

With this equality established, we see that the second inner product is equal to

$$\begin{aligned}
&-2\left\langle \nabla_x F(x_{k-1},y_{k-1}) - \widetilde{\nabla}_x^{\text{SARAH}}(x_{k-1},y_{k-1}), \mathbb{E}_{k,p}\left[\widetilde{\nabla}_x^{\text{SARAH}}(x_k,y_k) - \widetilde{\nabla}_x^{\text{SARAH}}(x_{k-1},y_{k-1})\right]\right\rangle \\
=&-2\langle \nabla_x F(x_{k-1},y_{k-1}) - \widetilde{\nabla}_x^{\text{SARAH}}(x_{k-1},y_{k-1}), \nabla_x F(x_k,y_k) - \nabla_x F(x_{k-1},y_{k-1})\rangle,
\end{aligned}$$

so the first two inner-products sum to zero. The third inner product is equal to

$$\begin{aligned}
&-2\left\langle \nabla_x F(x_k,y_k) - \nabla_x F(x_{k-1},y_{k-1}), \mathbb{E}_{k,p}\left[\widetilde{\nabla}_x^{\text{SARAH}}(x_k,y_k) - \widetilde{\nabla}_x^{\text{SARAH}}(x_{k-1},y_{k-1})\right]\right\rangle \\
=&-2\langle \nabla_x F(x_k,y_k) - \nabla_x F(x_{k-1},y_{k-1}), \nabla_x F(x_k,y_k) - \nabla_x F(x_{k-1},y_{k-1})\rangle \\
=&-2\|\nabla_x F(x_k,y_k) - \nabla_x F(x_{k-1},y_{k-1})\|^2.
\end{aligned}$$

Altogether, we have

$$\begin{aligned}
&\mathbb{E}_{k,p}\|\widetilde{\nabla}_x^{\text{SARAH}}(x_k,y_k) - \nabla_x F(x_k,y_k)\|^2 \\
\leq&\left\|\widetilde{\nabla}_x^{\text{SARAH}}(x_{k-1},y_{k-1}) - \nabla_x F(x_{k-1},y_{k-1})\right\|^2 - \|\nabla_x F(x_k,y_k) - \nabla_x F(x_{k-1},y_{k-1})\|^2 \\
&+\mathbb{E}_{k,p}\|\widetilde{\nabla}_x^{\text{SARAH}}(x_k,y_k) - \widetilde{\nabla}_x^{\text{SARAH}}(x_{k-1},y_{k-1})\|^2 \\
\leq&\left\|\widetilde{\nabla}_x^{\text{SARAH}}(x_{k-1},y_{k-1}) - \nabla_x F(x_{k-1},y_{k-1})\right\|^2 + \mathbb{E}_{k,p}\|\widetilde{\nabla}_x^{\text{SARAH}}(x_k,y_k) - \widetilde{\nabla}_x^{\text{SARAH}}(x_{k-1},y_{k-1})\|^2.
\end{aligned}$$

We can bound the second term by computing the expectation.

$$\begin{aligned}
\mathbb{E}_{k,p}\|\widetilde{\nabla}_x^{\text{SARAH}}(x_k,y_k) - \widetilde{\nabla}_x^{\text{SARAH}}(x_{k-1},y_{k-1})\|^2 =&\mathbb{E}_{k,p}\left\|\frac{1}{b}\left(\sum_{j\in J_k^x}\nabla_x F_j(x_k,y_k) - \nabla_x F_j(x_{k-1},y_{k-1})\right)\right\|^2 \\
\leq&\frac{1}{b}\mathbb{E}_{k,p}\left[\sum_{j\in J_k^x}\|\nabla_x F_j(x_k,y_k) - \nabla_x F_j(x_{k-1},y_{k-1})\|^2\right] \\
=&\frac{1}{n}\sum_{i=1}^n\|\nabla_x F_i(x_k,y_k) - \nabla_x F_i(x_{k-1},y_{k-1})\|^2.
\end{aligned} \tag{E.4}$$

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

$$\begin{aligned}
&\mathbb{E}_{k,p}\|\widetilde{\nabla}_x^{\text{SARAH}}(x_k,y_k) - \nabla_x F(x_k,y_k)\|^2 \\
\leq&\left\|\widetilde{\nabla}_x^{\text{SARAH}}(x_{k-1},y_{k-1}) - \nabla_x F(x_{k-1},y_{k-1})\right\|^2 + \frac{1}{n}\sum_{i=1}^n\|\nabla_x F_i(x_k,y_k) - \nabla_x F_i(x_{k-1},y_{k-1})\|^2.
\end{aligned}$$

This bounds the MSE under the condition that the full gradient is not computed. When the full gradient is computed, the MSE is equal to zero, so

$$\begin{aligned}
&\mathbb{E}_k\|\widetilde{\nabla}_x^{\text{SARAH}}(x_k,y_k) - \nabla_x F(x_k,y_k)\|^2 \\
\leq&\left(1-\frac{1}{p}\right)\left(\left\|\widetilde{\nabla}_x^{\text{SARAH}}(x_{k-1},y_{k-1}) - \nabla_x F(x_{k-1},y_{k-1})\right\|^2 + \frac{1}{n}\sum_{i=1}^n\|\nabla_x F_i(x_k,y_k) - \nabla_x F_i(x_{k-1},y_{k-1})\|^2\right) \\
\leq&\left(1-\frac{1}{p}\right)\left\|\widetilde{\nabla}_x^{\text{SARAH}}(x_{k-1},y_{k-1}) - \nabla_x F(x_{k-1},y_{k-1})\right\|^2 + M^2\|z_k - z_{k-1}\|^2.
\end{aligned}$$

By symmetric arguments, analogous results hold for $\mathbb{E}_k\|\widetilde{\nabla}_y^{\text{SARAH}}(x_{k+1},y_k) - \nabla_y F(x_{k+1},y_k)\|^2$:

$$\begin{aligned}
&\mathbb{E}_k\|\widetilde{\nabla}_y^{\text{SARAH}}(x_{k+1},y_k) - \nabla_y F(x_{k+1},y_k)\|^2 \\
\leq&\left(1-\frac{1}{p}\right)\left(\left\|\widetilde{\nabla}_y^{\text{SARAH}}(x_k,y_{k-1}) - \nabla_y F(x_k,y_{k-1})\right\|^2 + \frac{1}{n}\sum_{i=1}^n\|\nabla_y F_i(x_{k+1},y_k) - \nabla_y F_i(x_k,y_{k-1})\|^2\right) \\
\leq&\left(1-\frac{1}{p}\right)\left\|\widetilde{\nabla}_y^{\text{SARAH}}(x_k,y_{k-1}) - \nabla_y F(x_k,y_{k-1})\right\|^2 + M^2(\|x_{k+1} - x_k\|^2 + \|y_k - y_{k-1}\|^2).
\end{aligned}$$

Combining the two inequalities above, we have shown

$$\begin{aligned}
&\mathbb{E}_k[\|\widetilde{\nabla}_x^{\text{SARAH}}(x_k, y_k) - \nabla_x F(x_k, y_k)\|^2 + \|\widetilde{\nabla}_y^{\text{SARAH}}(x_{k+1}, y_k) - \nabla_y F(x_{k+1}, y_k)\|^2] \\
\leq &\left(1 - \frac{1}{p}\right) \left(\|\widetilde{\nabla}_x^{\text{SARAH}}(x_{k-1}, y_{k-1}) - \nabla_x F(x_{k-1}, y_{k-1})\|^2 + \|\widetilde{\nabla}_y^{\text{SARAH}}(x_k, y_{k-1}) - \nabla_y F(x_k, y_{k-1})\|^2\right) \\
&+ 2L^2 \mathbb{E}_k[\|z_{k+1} - z_k\|^2 + \|z_k - z_{k-1}\|^2]
\end{aligned} \tag{E.5}$$

We have also established the geometric decay property:

$$\mathbb{E}_k \Upsilon_{k+1} \leq \left(1 - \frac{1}{p}\right) \Upsilon_k + 2L^2 \mathbb{E}_k[\|z_{k+1} - z_k\|^2 + \|z_k - z_{k-1}\|^2], \tag{E.6}$$

justifying the choice of constants $\rho = 1/p$ and $V_1 = V_\Upsilon = 2L^2$. Similar bounds hold for Γ_k due to Jensen's inequality:

$$\begin{aligned}
&\mathbb{E}_k \|\widetilde{\nabla}_x^{\text{SARAH}}(x_k, y_k) - \nabla_x F(x_k, y_k)\| \\
\leq &\sqrt{\mathbb{E}_k \|\widetilde{\nabla}_x^{\text{SARAH}}(x_k, y_k) - \nabla_x F(x_k, y_k)\|^2} \\
\leq &\sqrt{\left(1 - \frac{1}{p}\right) \left\|\widetilde{\nabla}_x^{\text{SARAH}}(x_{k-1}, y_{k-1}) - \nabla_x F(x_{k-1}, y_{k-1})\right\|^2 + M^2 \|z_k - z_{k-1}\|^2} \\
\leq &\sqrt{\left(1 - \frac{1}{p}\right)} \left\|\widetilde{\nabla}_x^{\text{SARAH}}(x_{k-1}, y_{k-1}) - \nabla_x F(x_{k-1}, y_{k-1})\right\| + M \|z_k - z_{k-1}\|.
\end{aligned}$$

Applying an analogous result for $\widetilde{\nabla}_y$ gives the desired bound on Γ_k.

It is also easy to see that $\mathbb{E}\|z_k - z_{k-1}\|^2 \to 0$ implies $\mathbb{E}\Upsilon_k \to 0$:

$$\begin{aligned}
\mathbb{E}\Upsilon_k &\leq \left(1 - \frac{1}{p}\right) \mathbb{E}\Upsilon_{k-1} + 2L^2 \mathbb{E}[\|z_{k+1} - z_k\|^2 + \|z_k - z_{k-1}\|^2] \\
&\leq 2L^2 \sum_{\ell=1}^{k} \left(1 - \frac{1}{p}\right)^{k-\ell} \mathbb{E}[\|z_{\ell+1} - z_\ell\|^2 + \|z_\ell - z_{\ell-1}\|^2.
\end{aligned} \tag{E.7}$$

As $\mathbb{E}\|\widetilde{\nabla}_x^{\text{SARAH}}(x_k, y_k) - \nabla_x F(x_k, y_k)\|^2 \to 0$, so does $\mathbb{E}\|\widetilde{\nabla}_x^{\text{SARAH}}(x_k, y_k) - \nabla_x F(x_k, y_k)\| \to 0$ by Jensen's inequality, so it is clear that $\Gamma_k \to 0$ as well. $\square$

F. Additional Experiments

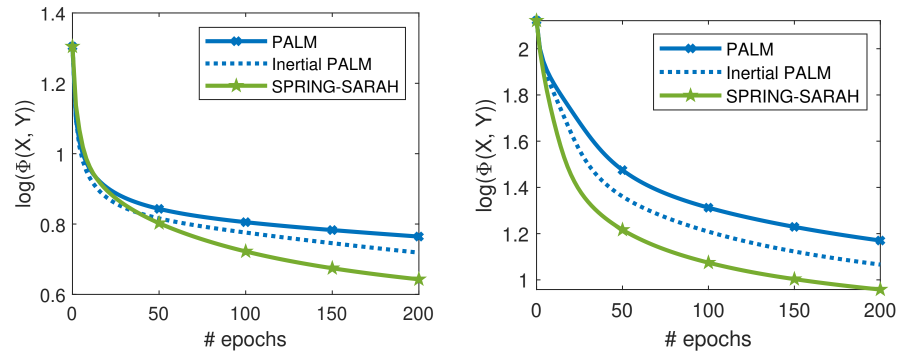

Figure 9. Blind image-deconvolution experiment on (a) Kodim04, (b) Kodim05 images, with motion-blur kernel.

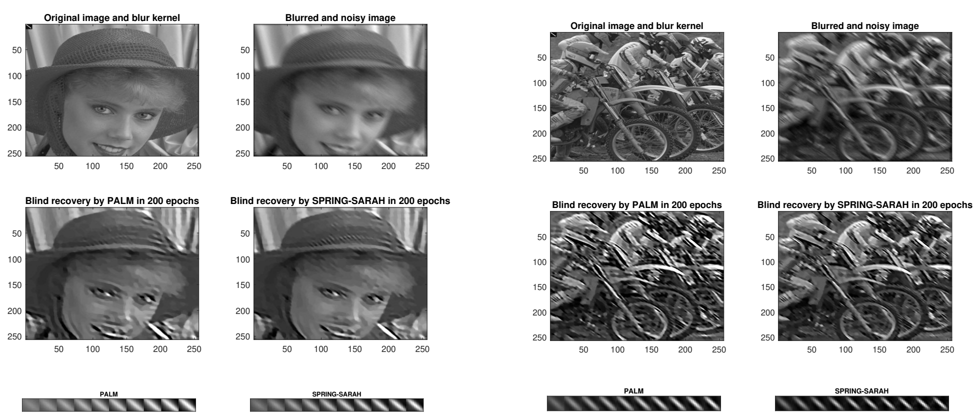

Figure 10. Blind Image-Deconvolution (11×11 motion blur).

For additional experiments in blind image deconvolution, we consider a 256×256 version of *Kodim04* and *Kodim05* images, blurred with an 11×11 motion blur kernel. The images are further degraded by additional Gaussian noise. From these results we observe that with the same number of epochs, SPRING-SARAH provides better recovery than both PALM and inertial PALM.