

REVISITING STOCHASTIC PROXIMAL POINT METHODS: GENERALIZED SMOOTHNESS AND SIMILARITY

ZHIRAYR TOVMASYAN, GRIGORY MALINOVSKY, LAURENT CONDAT*, PETER RICHTÁRIK

Computer Science Program, CEMSE Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Kingdom of Saudi Arabia

Abstract. The growing prevalence of nonsmooth optimization problems in machine learning has spurred significant interest in generalized smoothness assumptions. Among these, the (L_0, L_1) -smoothness assumption has emerged as one of the most prominent. While proximal methods are well-suited and effective for nonsmooth problems in deterministic settings, their stochastic counterparts remain under-explored. This work focuses on the stochastic proximal point method (SPPM), valued for its stability and minimal hyperparameter tuning—advantages often missing in stochastic gradient descent (SGD). We propose a novel ϕ -smoothness framework and provide a comprehensive analysis of SPPM without relying on traditional smoothness assumptions. Our results are highly general, encompassing existing findings as special cases. Furthermore, we examine SPPM under the widely adopted expected similarity assumption, thereby extending its applicability to a broader range of scenarios. Our theoretical contributions are illustrated and validated by practical experiments.

Keywords. Fixed-time convergence; Gradient flow; Lyapunov stability; Second-order dynamics; Time-varying systems.

1. INTRODUCTION

In this paper, we address the **stochastic optimization** problem of minimizing the expected function

$$\min_{x \in \mathbb{R}^d} \{ f(x) := \mathbb{E}_{\xi \sim \mathcal{D}} [f_\xi(x)] \}, \quad (1.1)$$

where $\xi \sim \mathcal{D}$ is a random variable drawn from the distribution \mathcal{D} , and $\mathbb{E}[\cdot]$ represents the mathematical expectation. Here, x represents a machine learning (ML) model with d parameters, \mathcal{D} denotes the distribution of labeled examples, $\xi \sim \mathcal{D}$ are the samples, f_ξ represents the loss associated with a single sample ξ , and f corresponds to the generalization error. In this setting, while an unbiased estimator of the gradient $\nabla f_\xi(x)$ can be computed, the gradient $\nabla f(x)$ itself is not directly accessible. Such problems form the backbone of supervised learning theory [8, 76, 77].

*Corresponding author.

E-mail address: laurent.condat@kaust.edu.sa (L. Condat).

Received 4 October 2025; Accepted 2 December 2025; Published online 1 April 2026.

A particular case of interest is the finite-sum optimization problem, where f is the average of a large number of functions [1, 74]:

$$\min_{x \in \mathbb{R}^d} \left\{ f(x) := \frac{1}{n} \sum_{i=1}^n f_i(x) \right\}. \quad (1.2)$$

This problem frequently arises when training supervised ML models using empirical risk minimization [27, 28]. It is an instance of (1.1) with \mathcal{D} the uniform distribution over the finite set $\{1, \dots, n\}$.

Optimization problems in ML and Deep Learning (DL) are frequently **nonsmooth**, meaning that the gradient of the objective function does not necessarily satisfy the Lipschitz continuity condition, or is even not well defined [32, 41]. For instance, even in relatively simple neural networks, the gradient of the standard ℓ_2 -regression loss fails to satisfy Lipschitz continuity [92]. Moreover, the convex but nonsmooth setting often provides an effective framework for capturing the complexities of DL problems [49]. Many widely used and empirically successful adaptive optimization algorithms, such as AdaGrad [25], Adam [51], and Prodigy [63] have been specifically designed for this setting, demonstrating their practical effectiveness across various DL applications.

In contrast, optimization methods that rely on smoothness assumptions and offer strong theoretical guarantees frequently fall short in practical DL tasks [17]. For example, while variance-reduced methods [34, 68] achieve superior convergence rates in theory, they are often outperformed by simpler methods in practice due to the challenges posed by the complex and highly nonconvex landscapes of DL [21]. These challenges have motivated researchers to introduce more realistic smoothness assumptions and develop corresponding theoretical guarantees within these refined frameworks [32, 85].

One of the earliest extensions beyond the standard Lipschitz smoothness assumption is the (L_0, L_1) -**smoothness** condition, which was initially proposed for twice-differentiable functions [92]. This assumption posits that the norm of the Hessian can be bounded linearly by the norm of the gradient. Later, this assumption was generalized to encompass a broader class of differentiable functions [11, 91].

Stochastic Gradient Descent (SGD) methods have been extensively analyzed in both convex and nonconvex settings, with significant attention also given to adaptive variants and other modifications [26, 40, 87, 93]. A natural extension of the (L_0, L_1) -smoothness assumption involves bounding the Hessian norm with a polynomial dependence on the gradient norm, offering a more flexible and generalized formulation. A further and even more general approach to smoothness involves the use of an arbitrary nondecreasing continuous function to bound the Hessian norm [57]. This generalized setting not only encompasses the previously discussed assumptions but also provides a broader and more adaptable framework applicable to a wide range of functions [84].

While SGD methods have been extensively studied in the context of generalized smoothness, **stochastic proximal point methods (SPPM)** remain relatively underexplored. SPPM can serve as an effective alternative when stochastic proximity operators are computationally feasible [2, 3, 47]. We recall that the proximity operator (**prox**) of a function $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ is

$$\text{prox}_f(x) \stackrel{\text{def}}{=} \underset{z \in \mathbb{R}^d}{\text{argmin}} \left\{ f(z) + \frac{1}{2} \|z - x\|^2 \right\} \quad (1.3)$$

[4]. Proximal methods, which leverage prox of functions [16, 67], are known for their robustness and resilience to the choice of stepsize, often allowing for the use of larger stepsizes compared to standard gradient-based methods. [71] provided convergence rate guarantees for SPPM and emphasize its stability with respect to inaccuracies in learning rate selection; a property not typically observed in SGD. [2] investigated a more general framework, AProx, which encompasses SPPM as a special case. They establish both stability and convergence rate results for AProx under convexity assumptions. Moreover, SPPM has been shown to achieve convergence rates comparable to those of SGD across a variety of algorithmic settings [69].

Alternatively, instead of relying on the smoothness assumption and its generalizations, we can consider the **similarity** assumption [75]. It reflects the idea that there is a certain level of similarity or homogeneity among the gradients, which is particularly relevant in ML, where these gradients capture the characteristics of the underlying data [31, 78]. Recognizing this natural property, several recent works have explored various formulations and generalizations of the similarity assumption [37, 73, 79]. In particular, multiple studies have analyzed stochastic proximal methods under different similarity conditions, demonstrating their practical relevance and theoretical significance [29, 73]. Moreover, the concept of similarity offers a more refined perspective on the behavior of optimization algorithms in distributed and federated learning settings [44, 46, 54].

To make the considered methods practical, the computation of a prox often involves an **inexact** approach, where it is approximated by several iterations of a subroutine designed to solve the corresponding auxiliary problem. This technique has been extensively studied in the literature, with various criteria established to ensure the effectiveness of the approximation [7, 35, 54]. These criteria typically include conditions such as relative and absolute error thresholds [47, 55], as well as guarantees on the reduction of the gradient norm [72]. Meeting these criteria is crucial to maintaining the overall convergence properties and efficiency of the optimization algorithm while balancing computational cost and accuracy.

Our **contributions** are the following.

- **New generalized smoothness assumption, called ϕ -smoothness.** We investigate more general conditions under which SPPM converges and introduce the novel notion of ϕ -smoothness (Section 4). Under this assumption, we establish rigorous convergence guarantees and explore various special cases, highlighting the specific effects and implications of the proposed framework.
- **Convergence under ϕ -smoothness.** We conduct a comprehensive analysis of SPPM when the prox is computed inexactly under the newly introduced ϕ -smoothness assumption (Section 5). Specifically, we derive conditions on the number of subroutine steps required to solve the auxiliary problem, ensuring that the overall iteration complexity remains the same as in the case of exact proximal evaluations. Our results provide practical guidelines for balancing computational efficiency and theoretical guarantees. Our convergence analysis covers both strongly convex and general convex settings, and we precisely characterize the convergence rates, offering insights into the tradeoffs between problem complexity and algorithmic performance.
- **Convergence under expected similarity.** We further extend our theoretical contributions to settings with the expected similarity assumption (Section 6), which captures practical scenarios where the different functions share a certain degree of similarity in

expectation. Under this assumption, we derive specific convergence results, offering valuable theoretical insights.

- **Experiments.** To support our theoretical findings, we conduct a series of carefully designed experiments that empirically validate our predictions and provide deeper insights into the practical performance of the proposed methods (Section 7).

2. PRELIMINARIES

Before presenting our convergence results, we define the key concepts and outline the assumptions used throughout the work. We start with the standard assumptions that apply to all our results, primarily that each stochastic function f_ξ is convex and differentiable (for simplicity, to avoid complicated notations associated to set-valued subdifferentials, see their use in [73])

Assumption 2.1 (Differentiability). The function $f_\xi : \mathbb{R}^d \rightarrow \mathbb{R}$ is differentiable for \mathcal{D} -almost every sample ξ .

We implicitly assume that differentiation and expectation can be interchanged, which leads to $\nabla f(x) = \nabla \mathbb{E}_{\xi \sim \mathcal{D}} [f_\xi(x)] = \mathbb{E}_{\xi \sim \mathcal{D}} [\nabla f_\xi(x)]$. Consequently, f is differentiable.

Assumption 2.2 (Convexity). The function $f_\xi : \mathbb{R}^d \rightarrow \mathbb{R}$ is convex for \mathcal{D} -almost every sample ξ ; that is,

$$f_\xi(x) \geq f_\xi(y) + \langle \nabla f_\xi(y), x - y \rangle \quad \forall x, y \in \mathbb{R}^d.$$

Next, we formulate the strong convexity assumption, which is applied to the expected function f in several settings:

Assumption 2.3 (Strong convexity of f). The function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is μ -strongly convex for some $\mu > 0$; that is,

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2} \|x - y\|^2 \quad \forall x, y \in \mathbb{R}^d.$$

A strongly convex function is better behaved and easier to minimize than a merely convex one.

Next, we consider the interpolation condition, where all stochastic functions share a common minimizer. This regime is commonly observed in overparameterized models, particularly large-scale ML models, where the number of parameters exceeds the number of training examples, enabling perfect fit to the data [60, 86].

Assumption 2.4 (Interpolation regime). There exists $x_\star \in \mathbb{R}^d$ such that $\nabla f_\xi(x_\star) = 0$ for \mathcal{D} -almost every sample ξ .

Now we need to formulate several smoothness assumptions which are used in the analysis. We start with the well-known standard assumption of L -smoothness:

Assumption 2.5 (L -smoothness). The function $f_\xi : \mathbb{R}^d \rightarrow \mathbb{R}$ is L -smooth for \mathcal{D} -almost every sample ξ ; that is, it is differentiable and its gradient is L -Lipschitz continuous:

$$\|\nabla f_\xi(x) - \nabla f_\xi(y)\| \leq L \|x - y\| \quad \forall x, y \in \mathbb{R}^d.$$

Next, we introduce the (L_0, L_1) -smoothness assumption, initially formulated by [91].

Assumption 2.6 (Symmetric (L_0, L_1) -smoothness). The function $f_\xi: \mathbb{R}^d \rightarrow \mathbb{R}$ is symmetrically (L_0, L_1) -smooth for \mathcal{D} -almost every sample ξ ; that is,

$$\|\nabla f_\xi(x) - \nabla f_\xi(y)\| \leq \left(L_0 + L_1 \sup_{u \in [x, y]} \|\nabla f_\xi(u)\| \right) \|x - y\| \quad \forall x, y \in \mathbb{R}^d,$$

where $[x, y]$ denotes the line segment connecting points x and y .

A more generalized form of symmetric (L_0, L_1) -smoothness is the so-called α -symmetric generalized smoothness, introduced by [11]. This extension provides a more flexible framework for analyzing smoothness properties by incorporating an additional parameter α , which allows for a broader class of functions to be captured within the smoothness definition.

Assumption 2.7 (α -symmetric generalized smoothness). The function $f_\xi: \mathbb{R}^d \rightarrow \mathbb{R}$ is α -symmetrically generalized-smooth for \mathcal{D} -almost every sample ξ ; that is,

$$\|\nabla f_\xi(x) - \nabla f_\xi(y)\| \leq \left(L_0 + L_1 \sup_{u \in [x, y]} \|\nabla f_\xi(u)\|^\alpha \right) \|x - y\| \quad \forall x, y \in \mathbb{R}^d,$$

where $\alpha \in [0, 1]$ and $[x, y]$ is the interval connecting the points x and y .

It is important to note that when $\alpha = 0$, the definition reduces to the standard L -smoothness, and when $\alpha = 1$, it corresponds to the symmetric (L_0, L_1) -smoothness.

Then we consider the **expected similarity** assumption. Initially, the concept of similarity was introduced for twice-differentiable functions and is commonly referred to as Hessian similarity [82]. For functions that are only once differentiable, analogous assumptions have been employed in the literature. In this work, we consider an even more general assumption introduced by [73]. We call it *Star Similarity* since it only has to hold with respect to a solution x_* .

Assumption 2.8 (Star Similarity). The function $f_\xi: \mathbb{R}^d \rightarrow \mathbb{R}$ has Star Similarity; that is, there exist a solution x_* of (1.1) and a constant $\delta_* > 0$ such that

$$\mathbb{E}_{\xi \sim \mathcal{D}} \left[\|\nabla f_\xi(x) - \nabla f(x) - \nabla f_\xi(x_*)\|^2 \right] \leq \delta_*^2 \|x - x_*\|^2 \quad \forall x \in \mathbb{R}^d.$$

Finally, we consider the inexact computation of the prox by employing a subroutine. To ensure convergence guarantees, we impose an assumption on the convergence rate of the subroutine, which has been previously introduced in [30, 72].

Assumption 2.9 (Inexact Proximal Condition). At the k -th iteration of the algorithm, the subroutine \mathcal{M} , after T iterations, produces an approximate solution \hat{x}_k that satisfies the inexactness condition

$$\|\nabla \Psi_k(\hat{x}_k)\|^2 \leq \frac{\eta \|x_k - x_k^\Psi\|^2}{T^\alpha},$$

where $\eta > 0$ is a scaling factor, $\alpha > 0$ denotes the decay rate, and x_k^Ψ denotes the exact minimizer of the proximal subproblem objective Ψ_k ; that is,

$$x_k^\Psi := \arg \min_{x \in \mathbb{R}^d} \left\{ \Psi_k(x) := f_{\xi_k}(x) + \frac{1}{2\gamma} \|x - x_k\|^2 \right\}. \quad (2.1)$$

Algorithm 1 Stochastic Proximal Point Method (SPPM)

-
- 1: **Parameters:** stepsize $\gamma > 0$, starting point $x_0 \in \mathbb{R}^d$
 - 2: **for** $t = 0, 1, \dots, T - 1$ **do**
 - 3: Sample $\xi_k \sim \mathcal{D}$
 - 4: $x_{k+1} := \operatorname{argmin}_{z \in \mathbb{R}^d} \left\{ f_{\xi_k}(z) + \frac{1}{2\gamma} \|z - x_k\|^2 \right\}$
 - 5: **end for**
-

TABLE 1. Summary of the main notations used in the paper.

Symbol	Description
x_*	optimal solution that minimizes $f(x)$
x_0	initial point
x_k	k -th iterate of SPPM
γ	stepsize of SPPM
$\phi(\cdot, \cdot)$	smoothness function.
$D_{f_\xi}(x, y)$	Bregman divergence of f_ξ between x and y , see (A.1)
\mathcal{F}_k	σ -algebra generated by x_0, \dots, x_k
\hat{x}_k	uniformly chosen iterate from the set $\{x_0, x_1, \dots, x_{k-1}\}$
x_k^Ψ	exact solution to the proximal subproblem, see (2.1)
T	number of inner iterations for inexact computation of the prox
σ_*^2	upper bound on the variance of the stochastic gradients
μ	strong convexity parameter of the function f
δ_*	Star Similarity constant

2.1. The Stochastic Proximal Point Method. The prox of a function f , defined in (1.3), is a well-defined operator when f is proper, closed and convex. Moreover, when f is differentiable, it satisfies

$$y = \operatorname{prox}_{\gamma f}(x) \Leftrightarrow y + \gamma \nabla f(y) = x. \quad (2.2)$$

This relationship provides an insightful connection between proximal operators and gradient-based methods, further motivating their use in optimization frameworks.

In real world applications it is often not possible or feasible to compute the prox of the total objective function f , and a stochastic approach comes in handy. The stochastic proximal point method (SPPM), shown as Algorithm 1, consists in applying at each iteration the prox of a randomly chosen f_ξ : $x_{k+1} := \operatorname{prox}_{\gamma f_\xi}(x_k)$, so that, for differentiable functions,

$$x_{k+1} = x_k - \gamma \nabla f_\xi(x_{k+1}). \quad (2.3)$$

Compared to the standard SGD update, SPPM appears very similar, with one small but conceptually major distinction: the gradient is evaluated at the updated point x_{k+1} . Proximal algorithms, such as SPPM, are thus implicit methods, as their updates require solving an equation in which the new iterate appears on both sides.

The main notations used throughout the paper are summarized in Table 1.

3. RELATED WORK

Stochastic Gradient Descent (SGD) [70] is a fundamental and widely used optimization algorithm for training machine learning models. Due to its efficiency and scalability, it has become the backbone of modern deep learning, with state-of-the-art training methods relying on various adaptations of SGD [76, 89, 94]. Over the years, the algorithm has been extensively studied, leading to a deeper understanding of its convergence properties, robustness, and efficiency in different settings [8, 23, 33, 48]. This ongoing theoretical research continues to refine SGD and its variants, ensuring their effectiveness in large-scale and complex learning tasks [10, 65, 90]. Notably, methods designed to leverage the smoothness of the objective often struggle in Deep Learning, where optimization problems are inherently nonsmooth. For instance, while variance-reduced methods [9, 22, 45, 62, 66, 74] theoretically offer faster convergence for finite sums of smooth functions, they are often outperformed in practice by standard, non-variance-reduced methods [21]. These challenges highlight the need to explore alternative assumptions that go beyond the standard smoothness assumption.

One of the most commonly used generalized smoothness assumptions is (L_0, L_1) -smoothness [92]. Several studies have analyzed SGD methods under this condition in the convex setting [32, 52, 57, 59, 81, 85]. The analysis of SGD in the non-convex case was first discussed in [92] and later extended to momentum-based methods in [91].

Similar results have been established for various optimization methods, including Normalized GD [11, 39, 93], SignGD [17], AdaGrad-Norm/AdaGrad [26, 87], Adam [88], and Normalized GD with Momentum [40]. Additionally, methods specifically designed for distributed optimization have been analyzed under generalized smoothness conditions [18, 24, 50].

Stochastic proximal point methods [5] have been extensively studied across different settings due to their versatility and strong theoretical properties. This framework can encompass various optimization algorithms, making it a unifying approach for analyzing and designing new methods. One of its key advantages is enhanced stability, which helps mitigate the challenges of variance in stochastic optimization. Additionally, it is particularly well-suited for nonsmooth problems, where traditional smoothness-based methods may struggle. [6, 19, 83]. These properties make stochastic proximal point methods a valuable tool in both theoretical analysis and practical applications [13, 14, 20].

Stochastic proximal methods have become increasingly important in Federated Learning due to their ability to handle decentralized optimization problems efficiently [53]. Some researchers propose replacing the standard local update steps with the proximity operator, which provides a more robust framework for understanding the behavior of local methods and can lead to faster convergence rates by improving the optimization process [36, 42, 56, 58, 80]. On the other hand, other studies focus on interpreting the aggregation step as a proximity operator, which allows for a more efficient combination of local updates [12, 15, 38, 43, 61, 64].

4. ϕ -SMOOTHNESS: A NEW GENERALIZED ASSUMPTION

The standard analysis of gradient-based methods, including proximal point methods, can be viewed within the framework of gradient-type algorithms due to their update formulation, as given by (2.3). While such analyses typically assume standard L -smoothness, they, in fact, operate under a path-wise smoothness condition, which applies specifically to the sequence of

points generated by the algorithm:

$$\|\nabla f_{\xi}(x_{k+1}) - \nabla f_{\xi}(x_k)\| \leq L_{\text{path}} \|x_{k+1} - x_k\|. \quad (4.1)$$

This condition is less restrictive than the conventional L -smoothness assumption (Assumption 2.5), as it only needs to hold for the iterates produced by the algorithm. However, verifying this condition is considerably more challenging, as it depends on the algorithm's trajectory and specific properties of the method. To circumvent this difficulty, we propose an alternative approach by establishing path-wise smoothness through a different set of assumptions. Specifically, we introduce the generalized ϕ -smoothness assumption, which, when combined with additional conditions, facilitates the derivation of the desired path-wise smoothness property.

Now we are ready to formulate our novel generalized smoothness assumption:

Assumption 4.1 (ϕ -smoothness). The function $f_{\xi}: \mathbb{R}^d \rightarrow \mathbb{R}$ is ϕ -smooth for \mathcal{D} -almost every sample ξ ; that is,

$$\|\nabla f_{\xi}(x) - \nabla f_{\xi}(y)\| \leq \phi(\|x - y\|, \|\nabla f_{\xi}(y)\|) \|x - y\| \quad \forall x, y \in \mathbb{R}^d,$$

where ϕ is a nonnegative and nondecreasing function in both variables.

This assumption is similar in nature to the concept of l -smoothness, as discussed in [57, 84]. However, l -smoothness specifically applies to twice differentiable functions. Next, we present an important lemma for the new class of functions, which plays a crucial role in the analysis.

Lemma 4.1. ϕ -smoothness (Assumption 4.1) implies that, for \mathcal{D} -almost every sample ξ ,

$$f_{\xi}(x) \leq f_{\xi}(y) + \langle \nabla f_{\xi}(y), x - y \rangle + \frac{\phi(\|x - y\|, \|\nabla f_{\xi}(y)\|)}{2} \|x - y\|^2 \quad \forall x, y \in \mathbb{R}^d.$$

Next, we demonstrate that the notion of ϕ -smoothness encompasses (L_0, L_1) -smoothness as a special case.

Lemma 4.2. If the function $f_{\xi}: \mathbb{R}^d \rightarrow \mathbb{R}$ satisfies the (L_0, L_1) -smoothness condition, then it is also ϕ -smooth with the function ϕ defined as

$$\phi(\|x - y\|, \|\nabla f_{\xi}(y)\|) = (L_0 + L_1 \|\nabla f_{\xi}(y)\|) \exp(L_1 \|x - y\|).$$

Another special case of the ϕ -smoothness assumption is

Lemma 4.3. If the function $f_{\xi}: \mathbb{R}^d \rightarrow \mathbb{R}$ satisfies the α -symmetric generalized-smoothness condition, then it is also ϕ -smooth with the function ϕ defined as

$$\phi(\|x - y\|, \|\nabla f_{\xi}(y)\|) = K_0 + K_1 \|\nabla f_{\xi}(y)\|^\alpha + K_2 \|x - y\|^{\frac{\alpha}{1-\alpha}},$$

where $K_0 := L_0 \left(2^{\frac{\alpha^2}{1-\alpha}} + 1\right)$, $K_1 := L_1 \cdot 2^{\frac{\alpha^2}{1-\alpha}} \cdot 3^\alpha$, $K_2 := L_1^{\frac{1}{1-\alpha}} \cdot 2^{\frac{\alpha^2}{1-\alpha}} \cdot 3^\alpha (1 - \alpha)^{\frac{\alpha}{1-\alpha}}$.

5. CONVERGENCE RESULTS UNDER ϕ -SMOOTHNESS

First, we outline some intuitions and key ideas behind our results. Our analysis crucially depends on the fact that the distance between two consecutive iterates x_k and x_{k+1} is bounded. This result is established using the interpolation regime, convexity of the objective function, and nonexpansiveness of the prox.

We begin by proving monotonicity of the distance to the solution, i.e., $\|x_{k+1} - x_*\|^2 \leq \|x_k - x_*\|^2$. Building upon this fundamental property, we derive the following key results:

Algorithm 2 SPPM-inexact

- 1: **Parameters:** stepsize $\gamma > 0$, starting point $x_0 \in \mathbb{R}^d$
 - 2: **for** $t = 0, 1, \dots, T - 1$ **do**
 - 3: Sample $\xi_k \sim \mathcal{D}$
 - 4: $\hat{x}_k := \underset{z \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ f_{\xi_k}(z) + \frac{1}{2\gamma} \|z - x_k\|^2 \right\}$
 - 5: $x_{k+1} := x_k - \gamma \nabla f_{\xi_k}(\hat{x}_k)$
 - 6: **end for**
-

Lemma 5.1. *Let Assumptions 2.2 and 2.4 hold. Then the iterates of SPPM satisfy, for every $k \geq 0$,*

$$\|x_k - x_{k+1}\|^2 \leq \|x_0 - x_*\|^2.$$

Next, we utilize the bound on the distance $\|x_k - x_{k+1}\|^2$ between consecutive iterates in conjunction with the ϕ -smoothness assumption (Assumption 4.1) to establish the concept of pathwise smoothness (4.1). By deriving several intermediate steps, we further obtain a bound on the difference between the function values evaluated at sequential iterates.

Lemma 5.2. *Let Assumptions 2.1, 2.2 and 4.1 hold. Then for any $\gamma > 0$, the iterates of SPPM satisfy, for every $k \geq 0$,*

$$f_{\xi_k}(x_k) - f_{\xi_k}(x_{k+1}) \leq \left(\frac{1}{\gamma} + \frac{\phi\left(\|x_0 - x_*\|, \frac{1}{\gamma}\|x_0 - x_*\|\right)}{2} \right) \|x_k - x_{k+1}\|^2. \quad (5.1)$$

We are now ready to present the final result and the main theorem. We establish convergence results, notably without imposing a bound on the stepsize. This implies that while increasing the stepsize can reduce the number of iterations, it also makes each iteration more computationally complex. We provide convergence results for both the strongly convex case and the general convex case. In the strongly convex setting, we achieve a linear rate of convergence, while in the general convex case, we obtain a sublinear rate of $\mathcal{O}(\frac{1}{k})$.

Theorem 5.1. *Let Assumptions 2.1 (Differentiability), 2.2 (Convexity), 2.4 (Interpolation), and 4.1 (ϕ -smoothness) hold. Then, for any stepsize $\gamma > 0$, the iterates of SPPM satisfy, for every $k \geq 0$,*

$$\mathbb{E}[f(\hat{x}_k)] - f(x_*) \leq \frac{\phi(\|x_0 - x_*\|, \frac{1}{\gamma}\|x_0 - x_*\|) + \frac{2}{\gamma}}{2k} \|x_0 - x_*\|^2,$$

where \hat{x}_k is a vector chosen from the collection of iterates x_0, \dots, x_{k-1} uniformly at random.

If, in addition, Assumption 2.3 holds, then, for any stepsize $\gamma > 0$, the iterates of SPPM satisfy, for every $k \geq 0$,

$$\mathbb{E}[\|x_k - x_*\|^2] \leq \left(1 - \frac{\mu}{\frac{2}{\gamma} + \phi(\|x_0 - x_*\|, \frac{1}{\gamma}\|x_0 - x_*\|)} \right)^k \|x_0 - x_*\|^2.$$

It is worth mentioning that our result guarantees convergence to the exact solution, as the variance in the interpolation regime does not hinder convergence to the optimal point. Additionally, it is important to note that there is an additional dependence on the distance between the starting point and the solution, $\|x_0 - x_*\|^2$, within the function ϕ .

Next, we analyze the inexact computation of the prox within the framework of ϕ -smoothness. In this setting, the prox is computed approximately, followed by an additional gradient step performed from the approximate point:

$$\hat{x}_k := \underset{z \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ f(z) + \frac{1}{2\gamma} \|z - x_k\|^2 \right\},$$

$$x_{k+1} := x_k - \gamma \nabla f(\hat{x}_k).$$

It is worth noting that if the prox is computed exactly, the additional gradient step becomes redundant. In this case, $\nabla f_{\xi}(\hat{x}_k) = \frac{1}{\gamma}(x_k - \hat{x}_k)$, and we have $x_{k+1} = \operatorname{prox}_{\gamma f_{\xi}}(x_k)$. SPPM with inexact prox (SPPM-inexact) is shown as Algorithm 2.

Remarkably, we show that when the number of iterations of the subroutine used to solve the inner subproblem is sufficiently large (satisfying a specific condition), the overall convergence guarantees are preserved up to a constant factor.

Theorem 5.2. *Let Assumptions 2.1 (Differentiability), 2.2 (Convexity), 2.4 (Interpolation) and 4.1 (ϕ -smoothness) hold. Consider SPPM-inexact with every inexact prox satisfying Assumption 2.9. If T is chosen sufficiently large such that $\frac{\eta\gamma^2}{T\alpha} \leq c < 1$, then, for any stepsize $\gamma > 0$ the iterates of SPPM-inexact satisfy, for every $k \geq 0$,*

$$\mathbb{E}[f(\hat{x}_k)] - f(x_*) \leq \frac{\phi(\|x_0 - x_*\|, \frac{1}{\gamma}\|x_0 - x_*\|) + \frac{2}{\gamma}}{2k(1-c)} \|x_0 - x_*\|^2.$$

If, in addition, Assumption 2.3 holds, then, for any $\gamma > 0$, the iterates of SPPM-inexact satisfy, for every $k \geq 0$,

$$\mathbb{E}[\|x_k - x_*\|^2] \leq \left(1 - \frac{(1-c)\mu}{\frac{2}{\gamma} + \phi(\|x_0 - x_*\|, \frac{1}{\gamma}\|x_0 - x_*\|)} \right)^k \|x_0 - x_*\|^2.$$

Remark 5.1 (Realizability of Assumption 2.9 by a parameter-free inner solver). Although ϕ -smoothness does not imply global Lipschitz continuity, the proximal subproblem Ψ_k is $\frac{1}{\gamma}$ -strongly convex and has a *locally* Lipschitz gradient on a bounded set containing the iterates (independent of k). Consequently, gradient descent on Ψ_k with Armijo backtracking (requiring no knowledge of L_0 , L_1 , or explicit values of ϕ) converges linearly. In particular, the inexactness model in Assumption 2.9 is realizable by such a parameter-free inner method. For more details, see Appendix B.2.1.

6. CONVERGENCE RESULTS UNDER EXPECTED SIMILARITY

In this section, we discuss the convergence results under expected similarity (Assumption 2.8) without assuming that all stochastic gradients vanish at the solution. The main concept of the proof is to introduce the average iterate:

$$\bar{x}_{k+1} = \mathbb{E}[x_{k+1} | \mathcal{F}_k],$$

where we denote by \mathcal{F}_k the σ -algebra generated by the randomness (e.g., stochastic gradients or iterates) up to iteration k . This approach enables us to derive the complexity of the algorithm. Since we utilize the expected similarity assumption, we do not require any form of smoothness.

In this setting, convergence is impacted by the variance, defined as follows:

Assumption 6.1 (Bounded Variance at Optimum). Let x_* denote any minimizer of f , supposed to exist. The variance of the stochastic gradients at x_* is bounded as:

$$\mathbb{E}_{\xi \sim \mathcal{D}} [\|\nabla f_{\xi}(x_*)\|^2] \leq \sigma_*^2.$$

Remark 6.1. The interpolation regime, where stochastic gradients vanish at the solution, is a special case of the general setting with $\sigma_*^2 = 0$. Detailed convergence results for this regime are provided in Appendix D.

First, we present the convergence result for the exact formulation of SPPM.

Theorem 6.1. *Let Assumptions 2.1 (Differentiability), 2.2 (Convexity), 2.3 (Strong convexity of f), 2.8 (Star Similarity), and 6.1 hold. If the stepsize satisfies $\gamma \leq \frac{\mu}{4\delta_*^2}$, then SPPM satisfies, for every $k \geq 0$,*

$$\mathbb{E}[\|x_k - x_*\|^2] \leq \left(1 - \frac{1}{2} \min\left(\frac{\gamma\mu}{2}, 1\right)\right)^k \|x_0 - x_*\|^2 + 4 \max\left(\frac{2}{\gamma\mu}, 1\right) \gamma^2 \sigma_*^2.$$

The result shows linear convergence to a neighborhood of the solution, with the neighborhood size proportional to the variance σ_*^2 from Assumption 6.1, and depends on the stepsize γ . Using a decaying stepsize schedule leads to sublinear convergence to the exact solution.

We now present the convergence guarantee for the inexact variant of SPPM.

Theorem 6.2. *Let Assumptions 2.1 (Differentiability), 2.2 (Convexity), 2.8 (Star Similarity), and 6.1 (Bounded Variance) hold. Consider SPPM-inexact with Assumption 2.9 satisfied. If the stepsize satisfies $\gamma \leq \frac{\mu(1-c)}{4\delta_*^2}$ and T is chosen sufficiently large such that $\frac{\eta\gamma^2}{T\alpha} \leq c$, where $0 < c < 1$ is a constant, then SPPM-inexact satisfies, for every $k \geq 0$,*

$$\mathbb{E}[\|x_k - x_*\|^2] \leq \left(1 - \frac{1}{2} \min\left(\frac{\gamma\mu}{2}, 1-c\right)\right)^k \|x_0 - x_*\|^2 + \max\left(\frac{2}{\gamma\mu}, \frac{1}{1-c}\right) \frac{4\gamma^2 \sigma_*^2}{(1-c)}.$$

The result guarantees convergence to a neighborhood of the solution, with a neighborhood size strictly larger than in the exact case due to the approximation error c . Specifically, the radius of this neighborhood increases by a factor of at least $\frac{1}{1-c}$. Additionally, the contraction factor becomes worse (closer to 1), and the allowable stepsize is reduced by a factor of $(1-c)$. Hence, the convergence speed is slower compared to the exact case, but convergence remains guaranteed as long as the approximation error is controlled.

7. EXPERIMENTS

In this section, we present numerical experiments conducted for the optimization problem in the finite-sum form (1.2), where the functions f_i take the specific form $f_i(x) = a_i \|x\|^{2s}$, with $s \in \mathbb{N} \setminus \{1\}$ and $a_i \in \mathbb{R}^+$ for all $i = 1, \dots, n$. Each function f_i is (L_0, L_1) -smooth, with $L_0 = 2s$ and $L_1 = 2s - 1$, and is convex.

7.1. Main Experiments. We begin with two main experiments. The first experiment investigates the effect of varying the maximum number of iterations for the inner solver, without enforcing a stopping condition based on the gradient norm. The purpose of this analysis is to demonstrate that if the inner solver fails to achieve sufficient accuracy in solving the proximal step, the method may either diverge or exhibit slower convergence. We conduct two variations of this experiment.

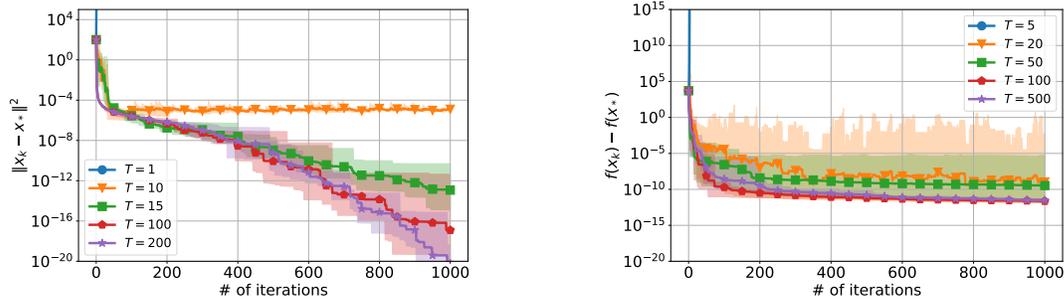


FIGURE 1. Convergence behavior of SPPM-inexact with different inner iterations in strongly convex and convex settings.

In the first variation, we consider functions with different values of s , ranging from 2 to 20, and run the algorithm with varying maximum iteration limits for the inner solver: $T = 5, 20, 50, 100, 500$. The number of functions is set to $n = 100$, and the dimension is $d = 5$. For the second variation, we modify the original problem to ensure strong convexity of f , enabling us to verify that the convergence rate becomes linear in this setting. The modified problem (1.2) is defined as follows: $f_i(x) = a_i \|x\|^{2s} + \lambda \langle e_i, x \rangle^2$, where $a_i \in \mathbb{R}^+$ for all $i = 1, \dots, n$, $\lambda \in \mathbb{R}^+$, and e_i is a unit vector with its i -th coordinate equal to one. In this case, each function f_i is (L_0, L_1) -smooth, with $L_0 = 2s + 2\lambda$ and $L_1 = 2s - 1$, and convex. Furthermore, f is $\frac{\lambda}{n}$ -strongly convex. We set $n = d = 100$, $\lambda = 2$, and $s = 2$, and conduct experiments with inner solver iteration limits of 1, 10, 15, 100, 200.

As shown in Figure 1, if the number of iterations of the subroutine is sufficiently small, the method may either diverge or exhibit significantly slower convergence. Once the number of inner iterations reaches a sufficiently large value, the convergence improves. It is worth noting that increasing the number of iterations beyond this point does not result in a significant further improvement in convergence. These observations confirm our theoretical findings.

In the second experiment, we analyze the performance of SPPM-inexact and compare it with stochastic gradient descent (SGD) using constant stepsizes. Each function in this experiment has the form $f_i(x) = a_i |x|^{2-\alpha_i}$, each of which is individually α_i -smooth, as shown by [11]. It is straightforward to verify that each f_i is also $1 > \beta > \alpha_i > 0$ -symmetric smooth, implying all functions are $\alpha = \max_{i=1 \dots N} \alpha_i$ -symmetric smooth. We set $N = 1000$, with coefficients a_i uniformly distributed in the range $[10, 1000)$.

We run both SPPM-inexact and SGD with three constant stepsizes: $\gamma \in \{10^{-3}, 5 \times 10^{-7}, 10^{-8}\}$. Figure 2 illustrates that for smaller stepsizes ($\gamma = 5 \times 10^{-7}, 10^{-8}$), both methods exhibit slow convergence, stagnating at suboptimal solutions within the number of iterations considered. However, for the larger stepsize ($\gamma = 10^{-3}$), SPPM-inexact converges consistently to the optimal solution, whereas SGD diverges.

This experiment highlights an advantage of SPPM-inexact: its capability to utilize relatively larger constant stepsizes effectively, resulting in faster and stable convergence.

On the other hand, constant stepsize SGD may require smaller stepsizes or careful tuning to achieve convergence. Hence, while the considered optimization scenario is not inherently challenging, it clearly demonstrates the practical benefits of the flexibility in step-size selection provided by SPPM-inexact.

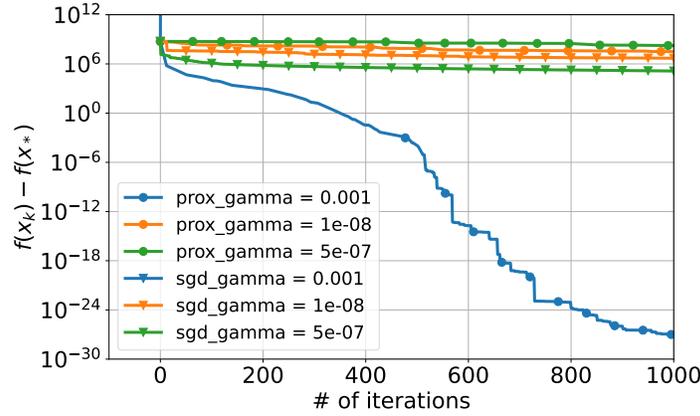


FIGURE 2. Comparison between SPPM-inexact and SGD with constant stepsizes γ .

7.2. Further Experiments. In addition to the main comparisons, we further examine the robustness of SPPM-inexact with respect to step-size selection and initialization. These experiments complement the previous results by providing deeper insight into the practical behavior of the method.

7.2.1. Impact of Different Stepsizes. We next investigate the impact of different stepsizes on the convergence of the proposed method. Specifically, we aim to demonstrate that the method converges for any positive stepsize, provided an appropriately chosen tolerance for the inner solver is used, and that larger stepsizes lead to faster convergence rates.

To validate this, we analyze three different values of s , namely $s = 2, 3, 4$. For each case, five different stepsizes are tested: $\gamma = 0.1, 1, 10, 100, 1000$. The number of functions is set to $n = 1000$, the dimension to $d = 100$, and an inexact solver is employed with a stopping criterion based on the squared norm of the gradient, with an accuracy threshold of 10^{-12} : $\|\nabla\Psi_k(\hat{x}_k)\|^2 \leq 10^{-12}$. As shown in Figure 3, the method converges for all chosen stepsizes. While larger stepsizes accelerate convergence in terms of the number of iterations, they also increase the computational cost per iteration. These observations confirm our theoretical findings.

7.2.2. Dependence on the Initial Point. Finally, we study the dependence of convergence on the initial point. The objective is to verify that the convergence of SPPM-inexact is independent of the initial point and to assess the tightness of the theoretical analysis, i.e., whether the convergence rate depends on the distance between the initial and optimal points. For this experiment, we analyze three different values of s ($s = 2, 3, 4$) and, for each case, randomly select five different initial points with norms $\|x_0\| = 0.1, 1, 10, 50, 100$, while keeping all other parameters unchanged. As shown in Figure 4, the convergence rates are nearly identical across all cases, suggesting that the upper bound $\|x_k - x_{k+1}\|^2 \leq \|x_0 - x_*\|^2$ provided in Lemma 5.1 may be overly restrictive for certain problems. Further investigation of this effect is a promising direction for future research.

It is worth noting that increasing the parameter s , which influences the problem formulation, makes the problem more challenging to solve. This is because the parameters L_0 and L_1 increase as s grows. This observation aligns with our theoretical understanding.

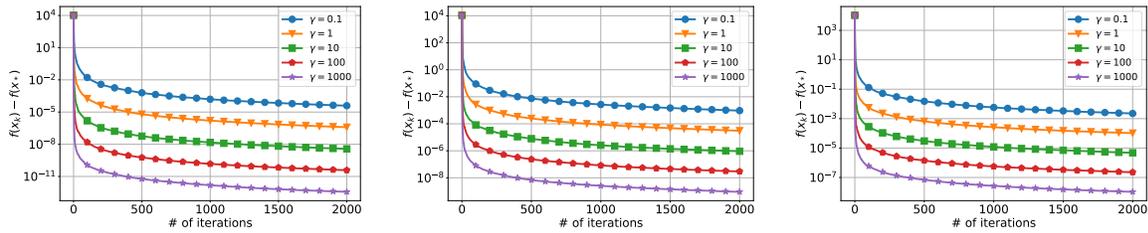


FIGURE 3. Convergence behavior of SPPM-inexact with different stepsizes.

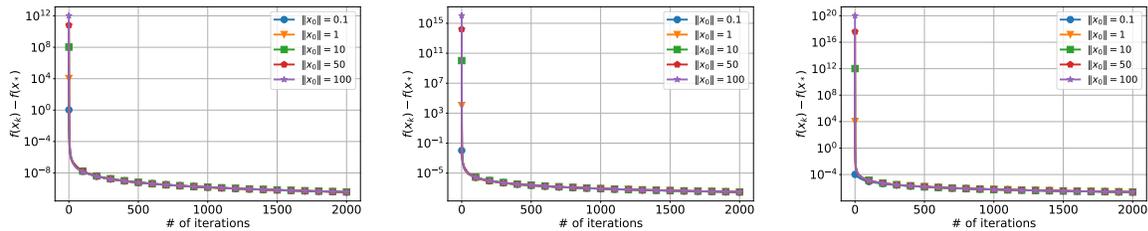


FIGURE 4. Convergence behavior of SPPM-inexact with different starting points.

8. CONCLUSION

In this work, we analyzed SPPM methods under assumptions beyond standard Lipschitz smoothness. We introduced a generalized ϕ -assumption that covers many special cases and provide convergence guarantees for both strongly convex and general settings in the interpolation regime. Additionally, we studied convergence under the expected similarity assumption in strongly convex cases for both interpolation and non-interpolation regimes. Our analysis of stochastic methods beyond smoothness is a step forward in the understanding of their practical performance in ML. Further exploring SPPM in general convex settings with expected similarity is an important direction for future research.

Acknowledgments

This work was supported by funding from King Abdullah University of Science and Technology (KAUST):

- i) KAUST Baseline Research Scheme,
- ii) Center of Excellence for Generative AI (award no. 5940),
- iii) Competitive Research Grant (CRG) Program (award no. 6460),
- iv) SDAIA-KAUST Center of Excellence in Data Science and Artificial Intelligence (SDAIA-KAUST AI).

REFERENCES

- [1] A. Agarwal, L. Bottou, lower bound for the optimization of finite sums, In: International Conference on Machine Learning, pp. 78–86. PMLR, 2015.
- [2] H. Asi, J.C. Duchi, Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity, *SIAM J. Optim.* 29 (2019), 2257–2290.
- [3] H. Asi, K. Chadha, G. Cheng, J.C. Duchi, Minibatch stochastic approximate proximal point methods, *Advances in Neural Information Processing Systems*, 33 (2020), 21958–21968.

- [4] H.H. Bauschke, P.L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. 2nd edition, CMS Books in Mathematics, Springer, Cham, 2017.
- [5] D.P. Bertsekas, Incremental proximal methods for large scale convex optimization, *Math. Program.* 129 (2011), 163–195.
- [6] P. Bianchi, Ergodic convergence of a stochastic proximal point algorithm, *SIAM J. Optim.* 26 (2016), 2235–2260.
- [7] E. Borodich, G. Kormakov, D. Kovalev, A. Beznosikov, A. Gasnikov, Optimal algorithm with complexity separation for strongly convex-strongly concave composite saddle point problems, *arXiv preprint arXiv:2307.12946*, 2023.
- [8] L. Bottou, F.E. Curtis, J. Nocedal, Optimization methods for large-scale machine learning, *SIAM Rev.* 60 (2010), 223–311.
- [9] X. Cai, C. Song, S. Wright, J. Diakonikolas, Cyclic block coordinate descent with variance reduction for composite nonconvex optimization, In: *International Conference on Machine Learning*, pp. 3469–3494, PMLR, 2023.
- [10] L. Carratino, A. Rudi, L. Rosasco, Learning with SGD and random features, In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 10213–10224, 2018.
- [11] Z. Chen, Y. Zhou, Y. Liang, Z. Lu, Generalized-smooth nonconvex optimization is as efficient as smooth nonconvex optimization, In: *International Conference on Machine Learning*, pp. 5396–5427, PMLR, 2023.
- [12] L. Condat, P. Richtárik, Randprox: Primal-dual optimization algorithms with randomized proximal updates, In: *International Conference on Learning Representations*, 2023.
- [13] L. Condat, P. Richtárik, A simple linear convergence analysis of the Point-SAGA algorithm, *arXiv preprint arXiv:2405.19951*, 2024.
- [14] L. Condat, G. Malinovsky, P. Richtárik, Distributed proximal splitting algorithms with rates and acceleration *Frontiers in Signal Processing*, 1 (2022), 776825.
- [15] L. Condat, I. Agarský, G. Malinovsky, P. Richtárik, TAMUNA: Doubly accelerated federated learning with local training, compression, and partial participation, *arXiv preprint arXiv:2302.09832*, 2023.
- [16] L. Condat, D. Kitahara, A. Contreras, A. Hirabayashi, Proximal splitting algorithms for convex optimization: A tour of recent advances, with new twists, *SIAM Rev.* 65 (2023), 375–435.
- [17] M. Crawshaw, M. Liu, F. Orabona, W. Zhang, Z. Zhuang, Robustness to unbounded smoothness of generalized signsgd, *Advances in Neural Information Processing Systems*, 35 (2022), 9955–9968.
- [18] M. Crawshaw, Y. Bao, M. Liu, Federated learning with client subsampling, data heterogeneity, and unbounded smoothness: A new algorithm and lower bounds, *Advances in Neural Information Processing Systems* 36 (NeurIPS 2023), 2024.
- [19] D. Davis, D. Drusvyatskiy, Stochastic model-based minimization of weakly convex functions, *SIAM J. Optim.* 29 (2019), 207–239.
- [20] D. Davis, W. Yin, A three-operator splitting scheme and its optimization applications, *Set-valued Var. Anal.* 25 (2017), 829–858.
- [21] A. Defazio, L. Bottou, On the ineffectiveness of variance reduced optimization for deep learning, In: *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, Vancouver, 2019.
- [22] A. Defazio, F. Bach, S. Lacoste-Julien, SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives, *Advances in Neural Information Processing Systems*, 27, pp. 1646–1654, 2014.
- [23] Y. Demidovich, G. Malinovsky, I. Sokolov, P. Richtárik, A guide through the zoo of biased SGD, *Advances in Neural Information Processing Systems*, pp. 23158–23171, 2023.
- [24] Y. Demidovich, P. Ostroukhov, G. Malinovsky, S. Horváth, M. Takáč, P. Richtárik, E. Gorbunov, Methods with local steps and random reshuffling for generally smooth non-convex federated optimization, *arXiv preprint arXiv:2412.02781*, 2024.
- [25] J. Duchi, E. Hazan, Y. Singer, Adaptive subgradient methods for online learning and stochastic optimization, *J. Mach. Learn. Res.* 12 (2011), 2121–2159.
- [26] M. Faw, L. Rout, C. Caramanis, S. Shakkottai, Beyond uniform smoothness: A stopped analysis of adaptive sgd, In: *The Thirty Sixth Annual Conference on Learning Theory*, pp. 89–160, PMLR, 2023.

- [27] V. Feldman, Generalization of erm in stochastic convex optimization: The dimension strikes back, *Advances in Neural Information Processing Systems* 29, 2016.
- [28] C. Gambella, B. Ghaddar, J. Naoum-Sawaya, Optimization problems for machine learning: A survey, *Eur. J. Oper. Res.* 290 (2021), 807–828.
- [29] E. Gasanov, P. Richtárik, Speeding up stochastic proximal optimization in the high Hessian dissimilarity setting, *arXiv preprint arXiv:2412.13619*, 2024.
- [30] A. Gasnikov, D. Kovalev, G. and Malinovsky, An optimal algorithm for strongly convex min-min optimization, *arXiv preprint arXiv:2212.14439*, 2022.
- [31] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*, MIT Press, 2016.
- [32] E. Gorbunov, N. Tupitsa, S. Choudhury, A. Aliev, P. Richtárik, S. Horváth, M. Takáč, Methods for convex (L_0, L_1) -smooth optimization: Clipping, acceleration, and adaptivity, *arXiv preprint arXiv:2409.14989*, 2024.
- [33] R.M. Gower, N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin, P. Richtárik, Sgd: General analysis and improved rates, In: *International Conference on Machine Learning*, pp. 5200–5209, PMLR, 2019.
- [34] R.M. Gower, M. Schmidt, F. Bach, P. Richtárik, Variance-reduced methods for machine learning, *Proce. IEEE* 108 (2020), 1968–1983.
- [35] M. Grudzień, G. Malinovsky, P. Richtárik, Can 5th generation local training methods support client sampling? yes!, In: *International Conference on Artificial Intelligence and Statistics*, pp. 1055–1092, PMLR, 2023.
- [36] M. Grudzień, G. Malinovsky, P. Richtárik, Improving accelerated federated learning with compression and importance sampling, *arXiv preprint arXiv:2306.03240*, 2023.
- [37] H. Hendriks, L. Xiao, S. Bubeck, F. Bach, L. Massoulié, Statistically preconditioned accelerated gradient method for distributed optimization, In: *International Conference on Machine Learning*, pp. 4203–4227, PMLR, 2020.
- [38] Z. Hu, H. Huang, Tighter analysis for ProxSkip, In: *International Conference on Machine Learning*, pp. 13469–13496, PMLR, 2023.
- [39] F. Hübler, I. Fatkhullin, N. He, From gradient clipping to normalization for heavy tailed SGD, *arXiv preprint arXiv:2410.13849*, 2024.
- [40] F. Hübler, J. Yang, X. Li, N. He, Parameter-agnostic optimization under relaxed smoothness, In: *International Conference on Artificial Intelligence and Statistics*, pp. 4861–4869, PMLR, 2024.
- [41] F. Iutzeler, J. Malick, Nonsmoothness in machine learning: Specific structure, proximal identification, and applications, *Set-Valued Var. Anal.* 28 (2020), 661–678.
- [42] D. Jhunjhunwala, S. Wang, G. Joshi, FedExP: Speeding up federated averaging via extrapolation, *arXiv preprint arXiv:2301.09604*, 2023.
- [43] X. Jiang, A. Rodomanov, S.U. Stich, Federated optimization with doubly regularized drift correction, *arXiv preprint arXiv:2404.08447*, 2024.
- [44] X. Jiang, A. Rodomanov, S.U. Stich, Stabilized proximal-point methods for federated optimization, *arXiv preprint arXiv:2407.07084*, 2024.
- [45] R. Johnson, T. Zhang, Accelerating stochastic gradient descent using predictive variance reduction, *Proceedings of the 27th International Conference on Neural Information Processing Systems*, 1 (2013), 315–323.
- [46] S.P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, A.T. Suresh, SCAFFOLD: Stochastic controlled averaging for federated learning, In: *International Conference on Machine Learning*, pp. 5132–5143, PMLR, 2020.
- [47] A. Khaled, C. Jin, Faster federated optimization under second-order similarity, *arXiv preprint arXiv:2209.02257*, 2022.
- [48] A. Khaled, P. Richtárik, Better theory for SGD in the nonconvex world, *arXiv preprint arXiv:2002.03329*, 2020.
- [49] A. Khaled, K. Mishchenko, C. Jin, DoWG unleashed: An efficient universal parameter-free gradient descent method, *Advances in Neural Information Processing Systems*, 36 (2023), 6748–6769.
- [50] S. Khirirat, A. Sadiev, A. Riabinin, E. Gorbunov, P. Richtárik, Error feedback under (L_0, L_1) -smoothness: Normalization and momentum, *arXiv preprint arXiv:2410.16871*, 2024.
- [51] D.P. Kingma, Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980*, 2014.
- [52] A. Koloskova, H. Hendriks, S.U. Stich, Revisiting gradient clipping: Stochastic bias and tight convergence guarantees, In: *International Conference on Machine Learning*, pp. 17343–17363, PMLR, 2023.

- [53] J. Konečný, Federated learning: Strategies for improving communication efficiency, arXiv preprint arXiv:1610.05492, 2016.
- [54] D. Kovalev, A. Beznosikov, E. Borodich, A. Gasnikov, G. Scutari, Optimal gradient sliding and its application to optimal distributed optimization under similarity, Advances in Neural Information Processing Systems, 35 (2022), 33494–33507.
- [55] H. Li, P. Richtárik, On the convergence of FedProx with extrapolation and inexact prox, arXiv preprint arXiv:2410.01410, 2024.
- [56] H. Li, K. Acharya, P. Richtarik, The power of extrapolation in federated learning, arXiv preprint arXiv:2405.13766, 2024a.
- [57] H. Li, J. Qian, Y. Tian, A. Rakhlin, A. Jadbabaie, Convex and non-convex optimization under generalized smoothness, In: Proceedings of the 37th International Conference on Neural Information Processing Systems, pp. 40238–4027, 2023.
- [58] T. Li, A.K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, V. Smith, Federated optimization in heterogeneous networks, Proceedings of Machine Learning and Systems, 2 (2020), 429–450.
- [59] A. Lobanov, A. Gasnikov, E. Gorbunov, M. Takác, Linear convergence rate in convex setup is possible! Gradient descent method variants under (L_0, L_1) -smoothness, arXiv preprint arXiv:2412.17050, 2024.
- [60] S. Ma, R. Bassily, M. Belkin, The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning, In: International Conference on Machine Learning, pp. 3325–3334, PMLR, 2018.
- [61] G. Malinovsky, K. Yi, P. Richtárik, Variance reduced ProxSkip: Algorithm, theory and application to federated learning, Advances in Neural Information Processing Systems, 35 (2022), 15176–15189.
- [62] G. Malinovsky, A. Sailanbayev, P. Richtárik, Random reshuffling with variance reduction: New analysis and better rates, In: Uncertainty in Artificial Intelligence, pp. 1347–1357. PMLR, 2023.
- [63] K. Mishchenko, A. Defazio, Prodigy: An expeditiously adaptive parameter-free learner, arXiv preprint arXiv:2306.06101, 2023.
- [64] K. Mishchenko, G. Malinovsky, S. Stich, P. Richtárik, ProxSkip: Yes! Local gradient steps provably lead to communication acceleration! finally!, In: International Conference on Machine Learning, pp. 15750–15769. PMLR, 2022.
- [65] L. Nguyen, P.H. Nguyen, M. Dijk, P. Richtárik, K. Scheinberg, M. Takác, SGD and Hogwild! convergence without the bounded gradients assumption, In: International Conference on Machine Learning, pp. 3750–3758. PMLR, 2018.
- [66] L.M. Nguyen, J. Liu, K. Scheinberg, M. Takáč, Sarah: A novel method for machine learning problems using stochastic recursive gradient, In: International Conference on Machine Learning, pp. 2613–2621. PMLR, 2017.
- [67] N. Parikh, S. Boyd, Proximal algorithms, Foundations and Trends in Optimization, 3 (2014), 127–239.
- [68] S.J. Reddi, A. Hefny, S. Sra, B. Póczos, A. Smola, Stochastic variance reduction for nonconvex optimization, In: International Conference on Machine Learning, pp. 314–323. PMLR, 2016.
- [69] P. Richtárik, A. Sadiev, Y. Demidovich, A unified theory of stochastic proximal point methods without smoothness, arXiv preprint arXiv:2405.15941, 2024.
- [70] H. Robbins, S. Monro, A stochastic approximation method, Ann. Math. Statist. 22 (1951), 400–407.
- [71] E.K. Ryu, S. Boyd, Stochastic proximal iteration: A non-asymptotic improvement upon stochastic gradient descent, <https://web.stanford.edu/~boyd/papers/spi.html>
- [72] A. Sadiev, D. Kovalev, P. Richtárik, Communication acceleration of local gradient methods via an accelerated primal-dual algorithm with an inexact prox, Advances in Neural Information Processing Systems, 35 (2022), 21777–21791.
- [73] A. Sadiev, L. Condat, P. Richtárik, Stochastic proximal point methods for monotone inclusions under expected similarity, arXiv preprint arXiv:2405.14255, 2024.
- [74] M. Schmidt, N. Le Roux, F. Bach, Minimizing finite sums with the stochastic average gradient, Math. Program. 162 (2017), 83–112.
- [75] O. Shamir, N. Srebro, T. Zhang, Communication-efficient distributed optimization using an approximate newton-type method, In: International Conference on Machine Learning, pp. 1000–1008. PMLR, 2014.
- [76] R.Y. Sun, Optimization for deep learning: An overview, J. Oper. Res. Chin. 8 (2020), 249–294.

- [77] S. Sun, Z. Cao, H. Zhu, J. Zhao, A survey of optimization methods from a machine learning perspective, *IEEE Trans. Cybernetics*, 50 (2019), 3668–3681.
- [78] Y. Sun, G. Scutari, A. Daneshmand, Distributed optimization based on gradient tracking revisited: Enhancing convergence rate via surrogation, *SIAM J. Optim.* 32 (2022), 354–385.
- [79] R. Szlendak, A. Tyurin, P. Richtárik, Permutation compressors for provably faster distributed nonconvex optimization, *arXiv preprint arXiv:2110.03300*, 2021.
- [80] C.T. Dinh, N. Tran, J. Nguyen, Personalized federated learning with Moreau envelopes, *Advances in Neural Information Processing Systems*, 33 (2020), 21394–21405.
- [81] Y. Takezawa, H. Bao, R. Sato, K. Niwa, M. Yamada, Polyak meets parameter-free clipped gradient descent, *arXiv preprint arXiv:2405.15010*, 2024.
- [82] Y. Tian, G. Scutari, T. Cao, A. Gasnikov, Acceleration in distributed optimization under similarity, In: *International Conference on Artificial Intelligence and Statistics*, pp. 5721–5756. PMLR, 2022.
- [83] P. Toulis, D. Tran, E. Airoldi, Towards stability and optimality in stochastic gradient descent, In: *Artificial Intelligence and Statistics*, pp. 1290–1298. PMLR, 2016.
- [84] A. Tyurin, Toward a unified theory of gradient descent under generalized smoothness, *arXiv preprint arXiv:2412.11773*, 2024.
- [85] D. Vankov, A. Rodomanov, A. Nedich, L. Sankar, S.U. Stich, Optimizing (L_0, L_1) -smooth functions by gradient methods, *arXiv preprint arXiv:2410.10800*, 2024.
- [86] A.V. Varre, L. Pillaud-Vivien, N. Flammarion, Last iterate convergence of SGD for least-squares in the interpolation regime, *Advances in Neural Information Processing Systems*, 34 (2021), 21581–21591.
- [87] B. Wang, H. Zhang, Z. Ma, W. Chen, Convergence of adagrad for non-convex objectives: Simple proofs and relaxed assumptions, In: *The Thirty Sixth Annual Conference on Learning Theory*, pp. 161–190. PMLR, 2023.
- [88] B. Wang, Y. Zhang, H. Zhang, Q. Meng, R. Sun, Z.M. Ma, T.Y. Liu, Z.Q. Luo, W. Chen, Provable adaptivity of adam under non-uniform smoothness, In: *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2960–2969, 2024.
- [89] J. Wu, W. Hu, H. Xiong, J. Huan, V. Braverman, Z. Zhu, On the noisy gradient descent that generalizes as SGD, In: *International Conference on Machine Learning*, pp. 10367–10376. PMLR, 2020.
- [90] J. Yang, X. Li, I. Fatkhullin, N. He, Two sides of one coin: the limits of untuned SGD and the power of adaptive methods, In: *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pp. 74257–74288, 2023.
- [91] B. Zhang, J. Jin, C. Fang, L. Wang, Improved analysis of clipping algorithms for non-convex optimization, *Advances in Neural Information Processing Systems*, 33 (2020), 15511–15521.
- [92] J. Zhang, T. He, S. Sra, A. Jadbabaie, Why gradient clipping accelerates training: A theoretical justification for adaptivity, *arXiv preprint arXiv:1905.11881*, 2019.
- [93] S.Y. Zhao, Y.P. Xie, W.J. Li, On the convergence and improvement of stochastic normalized gradient descent, *Sci. Chin. Info. Sci.* 64 (2021), 1–13.
- [94] P. Zhou, J. Feng, C. Ma, C. Xiong, S. Hoi, E. Weinan, Towards theoretically understanding why SGD generalizes better than Adam in deep learning, *Advances in Neural Information Processing Systems*, 33 (2020), 21285–21296.

APPENDIX A. FUNDAMENTAL LEMMAS

We define the Bregman divergence of a function g as

$$D_g(x, y) := g(x) - g(y) - \langle \nabla g(y), x - y \rangle \quad \forall x, y \in \mathbb{R}^d. \quad (\text{A.1})$$

Lemma A.1. *Let Assumptions 2.1, 2.2 and 2.4 hold. Then, for any stepsize $\gamma > 0$, the iterates of SPPM satisfy, for every $k \geq 0$,*

$$\langle x_{k+1} - x_*, x_k - x_{k+1} \rangle \geq \gamma(f_{\xi_k}(x_{k+1}) - f_{\xi_k}(x_*)) \geq 0. \quad (\text{A.2})$$

Proof. Since

$$x_{k+1} = \arg \min_{y \in \mathbb{R}^d} \left(f_{\xi_k}(y) + \frac{1}{2\gamma} \|y - x_k\|^2 \right),$$

the first-order optimality condition implies that:

$$\nabla f_{\xi_k}(x_{k+1}) + \frac{1}{\gamma}(x_{k+1} - x_k) = 0. \quad (\text{A.3})$$

Since f_{ξ_k} is differentiable and convex, we obtain the following inequality:

$$\begin{aligned} f_{\xi_k}(x_*) &\geq f_{\xi_k}(x_{k+1}) + \langle \nabla f_{\xi_k}(x_{k+1}), x_* - x_{k+1} \rangle \\ &\stackrel{(\text{A.3})}{=} f_{\xi_k}(x_{k+1}) + \frac{1}{\gamma} \langle x_k - x_{k+1}, x_* - x_{k+1} \rangle. \end{aligned}$$

Rearranging this expression yields the first inequality in (A.2). Furthermore, we also have

$$f_{\xi_k}(x_{k+1}) \geq \inf_{x \in \mathbb{R}^d} f_{\xi_k}(x) = f_{\xi_k}(x_*),$$

which follows from Assumption 2.4. □

A.1. Proof of Lemma 4.1.

$$\begin{aligned} f_{\xi}(x) - f_{\xi}(y) &= \int_0^1 \langle \nabla f_{\xi}(y + t(x - y)), x - y \rangle dt \\ &= \int_0^1 \langle \nabla f_{\xi}(y + t(x - y)) - \nabla f_{\xi}(y), x - y \rangle dt + \langle \nabla f_{\xi}(y), x - y \rangle. \end{aligned}$$

Moving $\langle \nabla f_{\xi}(y), x - y \rangle$ to the left-hand side and using Cauchy–Schwarz inequality, we get

$$\begin{aligned} D_{f_{\xi}}(x, y) &= \int_0^1 \langle \nabla f_{\xi}(y + t(x - y)) - \nabla f_{\xi}(y), x - y \rangle dt \\ &\leq \int_0^1 \|\nabla f_{\xi}(y + t(x - y)) - \nabla f_{\xi}(y)\| \|x - y\| dt. \end{aligned}$$

Using ϕ -smoothness of f_{ξ} and that ϕ is nondecreasing on both variables, we get

$$\begin{aligned} D_{f_{\xi}}(x, y) &\leq \int_0^1 (\phi(t\|x - y\|, \|\nabla f_{\xi}(y)\|)) t \|x - y\|^2 dt \\ &\leq \int_0^1 (\phi(\|x - y\|, \|\nabla f_{\xi}(y)\|)) t \|x - y\|^2 dt \\ &= \frac{\phi(\|x - y\|, \|\nabla f_{\xi}(y)\|)}{2} \|x - y\|^2. \end{aligned} \quad (\text{A.4})$$

A.2. Proof of Lemma 4.2. From Proposition 3.2 (Point 2) in [11], the function $f_\xi : \mathbb{R}^d \rightarrow \mathbb{R}$ is symmetrically (L_0, L_1) -smooth if and only if for all $x, y \in \mathbb{R}^d$,

$$\|\nabla f_\xi(x) - \nabla f_\xi(y)\| \leq (L_0 + L_1 \|\nabla f_\xi(y)\|) \exp(L_1 \|x - y\|) \|x - y\|.$$

Hence, defining

$$\phi(\|x - y\|, \|\nabla f_\xi(y)\|) := (L_0 + L_1 \|\nabla f_\xi(y)\|) \exp(L_1 \|x - y\|),$$

we obtain

$$\|\nabla f_\xi(x) - \nabla f_\xi(y)\| \leq \phi(\|x - y\|, \|\nabla f_\xi(y)\|) \|x - y\|,$$

which confirms that f_ξ is ϕ -smooth according to Assumption 4.1.

A.3. Proof of Lemma 4.3. From Proposition 3.2 (Point 1) in [11], the function $f_\xi : \mathbb{R}^d \rightarrow \mathbb{R}$ is said to satisfy the α -symmetric generalized-smoothness condition if and only if for all $x, y \in \mathbb{R}^d$,

$$\|\nabla f_\xi(x) - \nabla f_\xi(y)\| \leq \|x - y\| \left(K_0 + K_1 \|\nabla f_\xi(y)\|^\alpha + K_2 \|x - y\|^{\frac{\alpha}{1-\alpha}} \right),$$

where the constants K_0, K_1, K_2 are defined as:

$$K_0 := L_0 \left(2^{\frac{\alpha^2}{1-\alpha}} + 1 \right),$$

$$K_1 := L_1 \cdot 2^{\frac{\alpha^2}{1-\alpha}} \cdot 3^\alpha,$$

$$K_2 := L_1^{\frac{1}{1-\alpha}} \cdot 2^{\frac{\alpha^2}{1-\alpha}} \cdot 3^\alpha (1 - \alpha)^{\frac{\alpha}{1-\alpha}}.$$

Hence, defining

$$\phi(\|x - y\|, \|\nabla f_\xi(y)\|) := K_0 + K_1 \|\nabla f_\xi(y)\|^\alpha + K_2 \|x - y\|^{\frac{\alpha}{1-\alpha}},$$

we obtain

$$\|\nabla f_\xi(x) - \nabla f_\xi(y)\| \leq \phi(\|x - y\|, \|\nabla f_\xi(y)\|) \|x - y\|,$$

which confirms that f_ξ is ϕ -smooth in the sense of Assumption 4.1.

A.4. Proof of Lemma 5.1. Let $k \geq 0$. Since $x_{k+1} = \arg \min_{y \in \mathbb{R}^d} \left(f_{\xi_k}(y) + \frac{1}{2\gamma} \|y - x_k\|^2 \right)$, we have

$$f_{\xi_k}(x_{k+1}) + \frac{1}{2\gamma} \|x_k - x_{k+1}\|^2 \leq f_{\xi_k}(x_*) + \frac{1}{2\gamma} \|x_k - x_*\|^2,$$

Rearranging the terms, we obtain

$$\|x_k - x_{k+1}\|^2 \leq \|x_k - x_*\|^2 - 2\gamma (f_{\xi_k}(x_{k+1}) - f_{\xi_k}(x_*)).$$

Since $f_{\xi_k}(x_{k+1}) - f_{\xi_k}(x_*) \geq 0$, it follows that

$$\|x_k - x_{k+1}\|^2 \leq \|x_k - x_*\|^2.$$

On the other hand, due to the nonexpansiveness of the proximal operator, we know that

$$\|x_k - x_*\|^2 \leq \|x_{k-1} - x_*\|^2 \quad \text{for any } k \geq 1.$$

By applying this recursively, we obtain

$$\|x_k - x_*\|^2 \leq \|x_0 - x_*\|^2.$$

Combining the above inequalities, we arrive at the desired result.

A.5. Proof of Lemma 5.2. Let $k \geq 0$. Applying (4.1) to the iterates, we obtain

$$f_{\xi_k}(x_k) \leq f_{\xi_k}(x_{k+1}) + \langle \nabla f_{\xi_k}(x_{k+1}), x_k - x_{k+1} \rangle + \frac{\phi(\|x_k - x_{k+1}\|, \|\nabla f(x_{k+1})\|)}{2} \|x_k - x_{k+1}\|^2.$$

Using (2.2) we obtain

$$f_{\xi_k}(x_k) \leq f_{\xi_k}(x_{k+1}) + \frac{1}{\gamma} \|x_k - x_{k+1}\|^2 + \frac{\phi\left(\|x_k - x_{k+1}\|, \frac{1}{\gamma} \|x_k - x_{k+1}\|\right)}{2} \|x_k - x_{k+1}\|^2.$$

Finally, applying (5.1) and rearranging terms, we obtain

$$f_{\xi_k}(x_k) - f_{\xi_k}(x_{k+1}) \leq \left(\frac{1}{\gamma} + \frac{\phi\left(\|x_0 - x_*\|, \frac{1}{\gamma} \|x_0 - x_*\|\right)}{2} \right) \|x_k - x_{k+1}\|^2.$$

APPENDIX B. PROOF OF THEOREMS 5.1 AND 5.2

B.1. Proof of Theorem 5.1. For convenience, we restate Theorem 5.1 here:

Theorem B.1. *Let Assumptions 2.1 (Differentiability), 2.2 (Convexity), 2.4 (Interpolation) and 4.1 (ϕ -smoothness) hold. Then for any stepsize $\gamma > 0$ we have, for every $k \geq 0$,*

$$\mathbb{E}[f(\hat{x}_k)] - f(x_*) \leq \frac{\phi(\|x_0 - x_*\|, \frac{1}{\gamma} \|x_0 - x_*\|) + \frac{2}{\gamma}}{2k} \|x_0 - x_*\|^2, \quad (\text{B.1})$$

where \hat{x}^k is a vector chosen from the set of iterates x_0, \dots, x_{k-1} uniformly at random.

Additionally, if Assumption 2.3 holds, then, for any stepsize $\gamma > 0$, for every $k \geq 0$,

$$\mathbb{E}[\|x_k - x_*\|^2] \leq \left(1 - \frac{\mu}{\frac{2}{\gamma} + \phi(\|x_0 - x_*\|, \frac{1}{\gamma} \|x_0 - x_*\|)} \right)^k \|x_0 - x_*\|^2. \quad (\text{B.2})$$

Proof. We have

$$\begin{aligned} \|x_{k+1} - x_*\|^2 &= \|x_k - x_* - (x_k - x_{k+1})\|^2 \\ &= \|x_k - x_*\|^2 - 2\langle x_k - x_*, x_k - x_{k+1} \rangle + \|x_k - x_{k+1}\|^2 \\ &= \|x_k - x_*\|^2 - 2\langle x_k - x_{k+1} + x_{k+1} - x_*, x_k - x_{k+1} \rangle + \|x_k - x_{k+1}\|^2 \\ &= \|x_k - x_*\|^2 - 2\langle x_{k+1} - x_*, x_k - x_{k+1} \rangle - \|x_k - x_{k+1}\|^2. \end{aligned} \quad (\text{B.3})$$

Using (A.2) and (5.1), we obtain

$$\begin{aligned} &\|x_{k+1} - x_*\|^2 \\ &\leq \|x_k - x_*\|^2 - 2\gamma(f_{\xi_k}(x_{k+1}) - f_{\xi_k}(x_*)) - \frac{1}{\frac{1}{\gamma} + \frac{\phi(\|x_0 - x_*\|, \frac{1}{\gamma} \|x_0 - x_*\|)}{2}} (f_{\xi_k}(x_k) - f_{\xi_k}(x_{k+1})). \end{aligned}$$

Since $2\gamma > \frac{1}{\frac{1}{\gamma} + \frac{\phi(\|x_0 - x_*\|, \frac{1}{\gamma} \|x_0 - x_*\|)}{2}}$, we obtain

$$\|x_{k+1} - x_*\|^2 \leq \|x_k - x_*\|^2 - \frac{1}{\frac{1}{\gamma} + \frac{\phi(\|x_0 - x_*\|, \frac{1}{\gamma} \|x_0 - x_*\|)}{2}} (f_{\xi_k}(x_k) - f_{\xi_k}(x_*)).$$

Let \mathcal{F}_k denote the σ -algebra generated by the collection of random variables (x_0, \dots, x_k) . Taking the expectation conditioned on \mathcal{F}_k , we have

$$\mathbb{E}[\|x_{k+1} - x_*\|^2 \mid \mathcal{F}_k] \leq \|x_k - x_*\|^2 - \frac{1}{\frac{1}{\gamma} + \frac{\phi\left(\|x_0 - x_*\|, \frac{1}{\gamma}\|x_0 - x_*\|\right)}{2}} (f(x_k) - f(x_*)). \quad (\text{B.4})$$

Taking full expectation, we see that

$$\mathbb{E}[\|x_{k+1} - x_*\|^2] \leq \mathbb{E}[\|x_k - x_*\|^2] - \frac{1}{\frac{1}{\gamma} + \frac{\phi\left(\|x_0 - x_*\|, \frac{1}{\gamma}\|x_0 - x_*\|\right)}{2}} (\mathbb{E}[f(x_k)] - f(x_*)).$$

By summing up the inequalities telescopically for $t = 0, \dots, k$, we obtain

$$\begin{aligned} \sum_{t=0}^k \mathbb{E}[f(x_t)] - f(x_*) &\leq \left(\frac{1}{\gamma} + \frac{\phi\left(\|x_0 - x_*\|, \frac{1}{\gamma}\|x_0 - x_*\|\right)}{2} \right) (\|x_0 - x_*\|^2 - \mathbb{E}[\|x_{k+1} - x_*\|^2]) \\ &\leq \left(\frac{1}{\gamma} + \frac{\phi\left(\|x_0 - x_*\|, \frac{1}{\gamma}\|x_0 - x_*\|\right)}{2} \right) \|x_0 - x_*\|^2. \end{aligned}$$

Notice that:

$$\mathbb{E}[f(\hat{x}_{k+1})] = \mathbb{E}[\mathbb{E}[f(\hat{x}_{k+1}) \mid \mathcal{F}_k]] = \mathbb{E}\left[\frac{1}{k+1} \sum_{t=0}^k f(x_t)\right] = \frac{1}{k+1} \sum_{t=0}^k \mathbb{E}[f(x_t)].$$

Thus, we have

$$\mathbb{E}[f(\hat{x}_{k+1})] - f(x_*) \leq \frac{\frac{2}{\gamma} + \phi\left(\|x_0 - x_*\|, \frac{1}{\gamma}\|x_0 - x_*\|\right)}{2(k+1)} \|x_0 - x_*\|^2.$$

If we assume 2.3, then in step (B.4), applying the strong convexity of f , we get

$$\begin{aligned} \mathbb{E}[\|x_{k+1} - x_*\|^2 \mid \mathcal{F}_k] &\leq \|x_k - x_*\|^2 - \frac{\mu}{\frac{2}{\gamma} + \phi\left(\|x_0 - x_*\|, \frac{1}{\gamma}\|x_0 - x_*\|\right)} \|x_k - x_*\|^2 \\ &\leq \left(1 - \frac{\mu}{\frac{2}{\gamma} + \phi\left(\|x_0 - x_*\|, \frac{1}{\gamma}\|x_0 - x_*\|\right)} \right) \|x_k - x_*\|^2. \end{aligned}$$

Taking full expectation, we obtain

$$\mathbb{E}[\|x_{k+1} - x_*\|^2] \leq \left(1 - \frac{\mu}{\frac{2}{\gamma} + \phi\left(\|x_0 - x_*\|, \frac{1}{\gamma}\|x_0 - x_*\|\right)} \right) \mathbb{E}[\|x_k - x_*\|^2].$$

Applying this recursively, we get

$$\mathbb{E}[\|x_{k+1} - x_*\|^2] \leq \left(1 - \frac{\mu}{\frac{2}{\gamma} + \phi\left(\|x_0 - x_*\|, \frac{1}{\gamma}\|x_0 - x_*\|\right)} \right)^{k+1} \|x_0 - x_*\|^2.$$

□

B.2. Proof of Theorem 5.2. For convenience, we restate Theorem 5.2 as follows.

Theorem B.2. *Let Assumptions 2.1 (Differentiability), 2.2 (Convexity), 2.4 (Interpolation) and 4.1 (ϕ -smoothness) hold. Consider SPPM-inexact with every inexact prox satisfying Assumption 2.9. If T is chosen sufficiently large such that $\frac{\eta\gamma^2}{T\alpha} \leq c < 1$, then, for any stepsize $\gamma > 0$ the iterates of SPPM-inexact satisfy, for every $k \geq 0$,*

$$\mathbb{E}[f(\hat{x}_k)] - f(x_*) \leq \frac{\phi(\|x_0 - x_*\|, \frac{1}{\gamma}\|x_0 - x_*\|) + \frac{2}{\gamma}}{2k(1-c)} \|x_0 - x_*\|^2.$$

If, in addition, Assumption 2.3 holds, then, for any $\gamma > 0$, the iterates of SPPM-inexact satisfy, for every $k \geq 0$,

$$\mathbb{E}[\|x_k - x_*\|^2] \leq \left(1 - \frac{(1-c)\mu}{\frac{2}{\gamma} + \phi(\|x_0 - x_*\|, \frac{1}{\gamma}\|x_0 - x_*\|)}\right)^k \|x_0 - x_*\|^2.$$

Proof. We have

$$\|x_{k+1} - x_*\| = \|x_k - x_*\|^2 - 2\langle x_{k+1} - x_*, x_k - x_{k+1} \rangle - \|x_k - x_{k+1}\|^2.$$

Substituting $x_{k+1} = x_k - \gamma \nabla f_{\xi_k}(\hat{x}_k)$, we obtain

$$\begin{aligned} \|x_{k+1} - x_*\|^2 &= \|x_k - x_*\|^2 - \|x_{k+1} - x_k\|^2 - 2\gamma \langle \nabla f_{\xi_k}(\hat{x}_k), x_{k+1} - x_* \rangle \\ &= \|x_k - x_*\|^2 - \|x_{k+1} - x_k\|^2 - 2\gamma \langle \nabla f_{\xi_k}(\hat{x}_k), x_{k+1} - \hat{x}_k + \hat{x}_k - x_* \rangle \\ &= \|x_k - x_*\|^2 - \|x_{k+1} - x_k\|^2 - 2\gamma \langle \nabla f_{\xi_k}(\hat{x}_k), \hat{x}_k - x_* \rangle - 2\gamma \langle \nabla f_{\xi_k}(\hat{x}_k), x_{k+1} - \hat{x}_k \rangle. \end{aligned}$$

Using the identity $-2\langle a, b \rangle = -\|a + b\|^2 + \|a\|^2 + \|b\|^2$, we rewrite the expression as

$$\begin{aligned} \|x_{k+1} - x_*\|^2 &= \|x_k - x_*\|^2 - \|x_{k+1} - x_k\|^2 - 2\gamma \langle \nabla f_{\xi_k}(\hat{x}_k), \hat{x}_k - x_* \rangle \\ &\quad - \|x_{k+1} + \gamma \nabla f_{\xi_k}(\hat{x}_k) - \hat{x}_k\|^2 + \|\gamma \nabla f_{\xi_k}(\hat{x}_k)\|^2 + \|x_{k+1} - \hat{x}_k\|^2. \end{aligned}$$

Noting that $\|x_{k+1} + \gamma \nabla f_{\xi_k}(\hat{x}_k) - x_k\|^2 = \|x_k - \hat{x}_k\|^2$ and $\|\gamma \nabla f_{\xi_k}(\hat{x}_k)\|^2 = \|x_{k+1} - x_k\|^2$, we simplify as

$$\|x_{k+1} - x_*\|^2 = \|x_k - x_*\|^2 - 2\gamma \langle \nabla f_{\xi_k}(\hat{x}_k), \hat{x}_k - x_* \rangle - \|x_k - \hat{x}_k\|^2 + \|x_{k+1} - \hat{x}_k\|^2.$$

Using $\|x_{k+1} - \hat{x}_k\|^2 = \|x_k - \gamma \nabla f_{\xi_k}(\hat{x}_k) - \hat{x}_k\|^2 = \|-\gamma \nabla \Psi_k(\hat{x}_k)\|^2$, we derive

$$\|x_{k+1} - x_*\|^2 = \|x_k - x_*\|^2 - 2\gamma \langle \nabla f_{\xi_k}(\hat{x}_k), \hat{x}_k - x_* \rangle - \|x_k - \hat{x}_k\|^2 + \gamma^2 \|\nabla \Psi_k(\hat{x}_k)\|^2. \quad (\text{B.5})$$

By convexity of f_{ξ_k} , we have

$$f_{\xi_k}(x_*) \geq f_{\xi_k}(\hat{x}_k) + \langle \nabla f_{\xi_k}(\hat{x}_k), x_* - \hat{x}_k \rangle,$$

which implies

$$\langle \nabla f_{\xi_k}(\hat{x}_k), \hat{x}_k - x_* \rangle \geq f_{\xi_k}(\hat{x}_k) - f_{\xi_k}(x_*). \quad (\text{B.6})$$

Using (B.6) in (B.5), we get

$$\|x_{k+1} - x_*\|^2 \leq \|x_k - x_*\|^2 - 2\gamma(f_{\xi_k}(\hat{x}_k) - f_{\xi_k}(x_*)) - \|x_k - \hat{x}_k\|^2 + \gamma^2 \|\nabla \Psi_k(\hat{x}_k)\|^2. \quad (\text{B.7})$$

Since x_k^Ψ is the minimizer of $\Psi_k(x)$, we have

$$f_{\xi_k}(x_k^\Psi) + \frac{1}{2\gamma} \|x_k^\Psi - x_k\|^2 \leq f_{\xi_k}(\hat{x}_k) + \frac{1}{2\gamma} \|\hat{x}_k - x_k\|^2,$$

from which it follows that

$$f_{\xi_k}(\hat{x}_k) \geq f_{\xi_k}(x_k^\Psi) + \frac{1}{2\gamma} \|x_k^\Psi - x_k\|^2 - \frac{1}{2\gamma} \|\hat{x}_k - x_k\|^2.$$

Replacing it in (B.7), we obtain

$$\begin{aligned} \|x_{k+1} - x_*\|^2 &\leq \|x_k - x_*\|^2 - 2\gamma(f_{\xi_k}(x_k^\Psi) - f_{\xi_k}(x_*)) - \|x_k^\Psi - x_k\|^2 + \|\hat{x}_k - x_k\|^2 - \|x_k - \hat{x}_k\|^2 \\ &\quad + \gamma^2 \|\nabla \Psi_k(\hat{x}_k)\|^2 \\ &= \|x_k - x_*\|^2 - 2\gamma(f_{\xi_k}(x_k^\Psi) - f_{\xi_k}(x_*)) - \|x_k^\Psi - x_k\|^2 + \gamma^2 \|\nabla \Psi_k(\hat{x}_k)\|^2. \end{aligned}$$

Using (2.9), we get

$$\begin{aligned} \|x_{k+1} - x_*\|^2 &\leq \|x_k - x_*\|^2 - 2\gamma(f_{\xi_k}(x_k^\Psi) - f_{\xi_k}(x_*)) - \|x_k^\Psi - x_k\|^2 + \frac{\eta\gamma^2 \|x_k - x_k^\Psi\|^2}{T\alpha} \\ &= \|x_k - x_*\|^2 - 2\gamma(f_{\xi_k}(x_k^\Psi) - f_{\xi_k}(x_*)) - \left(1 - \frac{\eta\gamma^2}{T\alpha}\right) \|x_k - x_k^\Psi\|^2. \end{aligned} \quad (\text{B.8})$$

From the proximal step, we have

$$f_{\xi_k}(x_k^\Psi) + \frac{1}{2\gamma} \|x_k - x_k^\Psi\|^2 \leq f_{\xi_k}(x_*) + \frac{1}{2\gamma} \|x_* - x_k\|^2.$$

Rearranging the terms, we get

$$\begin{aligned} \|x_k - x_k^\Psi\|^2 &\leq \|x_* - x_k\|^2 - 2\gamma(f_{\xi_k}(x_k^\Psi) - f_{\xi_k}(x_*)) \\ &\leq \|x_* - x_k\|^2, \end{aligned}$$

and from the (B.8), we know that $\|x_{k+1} - x_*\|^2 \leq \|x_k - x_*\|^2$, so using recursion, we get $\|x_k - x_k^\Psi\|^2 \leq \|x_0 - x_*\|^2$, which means we can use (5.1) in (B.8):

$$\begin{aligned} \|x_{k+1} - x_*\|^2 &\leq \|x_k - x_*\|^2 - 2\gamma(f_{\xi_k}(x_k^\Psi) - f_{\xi_k}(x_*)) \\ &\quad - \frac{1 - \frac{\eta\gamma^2}{T\alpha}}{\frac{1}{\gamma} + \frac{\phi(\|x_0 - x_*\|, \frac{1}{\gamma}\|x_0 - x_*\|)}{2}} \left(f_{\xi_k}(x_k) - f_{\xi_k}(x_k^\Psi)\right), \end{aligned} \quad (\text{B.9})$$

and since $2\gamma > \frac{1 - \frac{\eta\gamma^2}{T\alpha}}{\frac{1}{\gamma} + \frac{\phi(\|x_0 - x_*\|, \frac{1}{\gamma}\|x_0 - x_*\|)}{2}}$, we can write

$$\begin{aligned} \|x_{k+1} - x_*\|^2 &\leq \|x_k - x_*\|^2 - \frac{1 - \frac{\eta\gamma^2}{T\alpha}}{\frac{1}{\gamma} + \frac{\phi(\|x_0 - x_*\|, \frac{1}{\gamma}\|x_0 - x_*\|)}{2}} (f_{\xi_k}(x_k^\Psi) - f_{\xi_k}(x_*)) \\ &\quad - \frac{1 - \frac{\eta\gamma^2}{T\alpha}}{\frac{1}{\gamma} + \frac{\phi(\|x_0 - x_*\|, \frac{1}{\gamma}\|x_0 - x_*\|)}{2}} \left(f_{\xi_k}(x_k) - f_{\xi_k}(x_k^\Psi)\right) \\ &= \|x_k - x_*\|^2 - \frac{1 - \frac{\eta\gamma^2}{T\alpha}}{\frac{1}{\gamma} + \frac{\phi(\|x_0 - x_*\|, \frac{1}{\gamma}\|x_0 - x_*\|)}{2}} \left(f_{\xi_k}(x_k) - f_{\xi_k}(x_*)\right). \end{aligned}$$

Taking the expectation conditioned on \mathcal{F}_k , we have

$$\mathbb{E}[\|x_{k+1} - x_*\|^2 \mid \mathcal{F}_k] \leq \|x_k - x_*\|^2 - \frac{1 - \frac{\eta\gamma^2}{T^\alpha}}{\frac{1}{\gamma} + \frac{\phi\left(\|x_0 - x_*\|, \frac{1}{\gamma}\|x_0 - x_*\|\right)}{2}} (f(x_k) - f(x_*)). \quad (\text{B.10})$$

Now, by taking the full expectation, we get

$$\mathbb{E}[\|x_{k+1} - x_*\|^2] \leq \mathbb{E}[\|x_k - x_*\|^2] - \frac{1 - \frac{\eta\gamma^2}{T^\alpha}}{\frac{1}{\gamma} + \frac{\phi\left(\|x_0 - x_*\|, \frac{1}{\gamma}\|x_0 - x_*\|\right)}{2}} (\mathbb{E}[f(x_k)] - f(x_*)).$$

By summing up the inequalities telescopically for $t = 0, \dots, k$, we obtain

$$\begin{aligned} \sum_{t=0}^k \mathbb{E}[f(x_t)] - f(x_*) &\leq \frac{\frac{1}{\gamma} + \frac{\phi\left(\|x_0 - x_*\|, \frac{1}{\gamma}\|x_0 - x_*\|\right)}{2}}{1 - \frac{\eta\gamma^2}{T^\alpha}} (\|x_0 - x_*\|^2 - \mathbb{E}[\|x_{k+1} - x_*\|^2]) \\ &\leq \frac{\frac{1}{\gamma} + \frac{\phi\left(\|x_0 - x_*\|, \frac{1}{\gamma}\|x_0 - x_*\|\right)}{2}}{1 - \frac{\eta\gamma^2}{T^\alpha}} \|x_0 - x_*\|^2. \end{aligned}$$

Notice that

$$\mathbb{E}[f(\hat{x}_{k+1})] = \mathbb{E}[\mathbb{E}[f(\hat{x}_{k+1}) \mid \mathcal{F}_k]] = \mathbb{E}\left[\frac{1}{k+1} \sum_{t=0}^k f(x_t)\right] = \frac{1}{k+1} \sum_{t=0}^k \mathbb{E}[f(x_t)].$$

Thus, we have

$$\begin{aligned} \mathbb{E}[f(\hat{x}_{k+1})] - f(x_*) &\leq \frac{\frac{2}{\gamma} + \phi\left(\|x_0 - x_*\|, \frac{1}{\gamma}\|x_0 - x_*\|\right)}{2(k+1)\left(1 - \frac{\eta\gamma^2}{T^\alpha}\right)} \|x_0 - x_*\|^2 \\ &\leq \frac{\frac{2}{\gamma} + \phi\left(\|x_0 - x_*\|, \frac{1}{\gamma}\|x_0 - x_*\|\right)}{2(k+1)(1-c)} \|x_0 - x_*\|^2. \end{aligned}$$

If we assume 2.3, then in step (B.10), applying the strong convexity of f , we get

$$\begin{aligned} \mathbb{E}[\|x_{k+1} - x_*\|^2 \mid \mathcal{F}_k] &\leq \|x_k - x_*\|^2 - \frac{\mu\left(1 - \frac{\eta\gamma^2}{T^\alpha}\right)}{\frac{2}{\gamma} + \phi\left(\|x_0 - x_*\|, \frac{1}{\gamma}\|x_0 - x_*\|\right)} \|x_k - x_*\|^2 \\ &\leq \left(1 - \frac{\mu\left(1 - \frac{\eta\gamma^2}{T^\alpha}\right)}{\frac{2}{\gamma} + \phi\left(\|x_0 - x_*\|, \frac{1}{\gamma}\|x_0 - x_*\|\right)}\right) \|x_k - x_*\|^2. \end{aligned}$$

Taking the full expectation, we obtain

$$\mathbb{E}[\|x_{k+1} - x_*\|^2] \leq \left(1 - \frac{\mu\left(1 - \frac{\eta\gamma^2}{T^\alpha}\right)}{\frac{2}{\gamma} + \phi\left(\|x_0 - x_*\|, \frac{1}{\gamma}\|x_0 - x_*\|\right)}\right) \mathbb{E}[\|x_k - x_*\|^2].$$

Applying this recursively, we get

$$\begin{aligned} \mathbb{E}[\|x_{k+1} - x_*\|^2] &\leq \left(1 - \frac{\mu \left(1 - \frac{\eta\gamma^2}{T^\alpha}\right)}{\frac{2}{\gamma} + \phi\left(\|x_0 - x_*\|, \frac{1}{\gamma}\|x_0 - x_*\|\right)}\right)^{k+1} \|x_0 - x_*\|^2 \\ &\leq \left(1 - \frac{\mu(1-c)}{\frac{2}{\gamma} + \phi\left(\|x_0 - x_*\|, \frac{1}{\gamma}\|x_0 - x_*\|\right)}\right)^{k+1} \|x_0 - x_*\|^2. \end{aligned}$$

□

B.2.1. Practical details (parameter-free Armijo on Ψ_k). Fix k . Recall Ψ_k and x_k^Ψ from (2.1). By the optimality condition,

$$x_k^\Psi + \gamma \nabla f_{\xi_k}(x_k^\Psi) = x_k \implies \|\nabla f_{\xi_k}(x_k^\Psi)\| = \frac{1}{\gamma} \|x_k - x_k^\Psi\|. \quad (\text{B.11})$$

From the proof of Theorem 5.2,

$$\|x_k - x_k^\Psi\| \leq \|x_0 - x_*\| =: R, \quad (\text{B.12})$$

hence by (B.11),

$$\|\nabla f_{\xi_k}(x_k^\Psi)\| \leq \frac{R}{\gamma}. \quad (\text{B.13})$$

Let $B_k := \{z : \|z - x_k^\Psi\| \leq R\}$. For $y \in B_k$, by ϕ -smoothness and (B.13),

$$\begin{aligned} \|\nabla f_{\xi_k}(y) - \nabla f_{\xi_k}(x_k^\Psi)\| &\leq \phi(\|y - x_k^\Psi\|, \|\nabla f_{\xi_k}(x_k^\Psi)\|) \|y - x_k^\Psi\| \\ &\leq \phi\left(R, \frac{R}{\gamma}\right) R. \end{aligned} \quad (\text{B.14})$$

Thus

$$\|\nabla f_{\xi_k}(y)\| \leq G := \frac{R}{\gamma} + \phi\left(R, \frac{R}{\gamma}\right) R, \quad y \in B_k. \quad (\text{B.15})$$

Now, for any $x, y \in B_k$,

$$\begin{aligned} \|\nabla f_{\xi_k}(x) - \nabla f_{\xi_k}(y)\| &\leq \phi(\|x - y\|, \|\nabla f_{\xi_k}(y)\|) \|x - y\| \\ &\leq \phi(2R, G) \|x - y\|. \end{aligned} \quad (\text{B.16})$$

Since $\nabla \Psi_k(x) = \nabla f_{\xi_k}(x) + \frac{1}{\gamma}(x - x_k)$, we obtain

$$\|\nabla \Psi_k(x) - \nabla \Psi_k(y)\| \leq (\phi(2R, G) + \frac{1}{\gamma}) \|x - y\| =: L_{\text{loc}} \|x - y\|, \quad \forall x, y \in B_k. \quad (\text{B.17})$$

Here L_{loc} is independent of k , and Ψ_k is μ -strongly convex with $\mu = \frac{1}{\gamma}$.

Implication for Assumption 2.9. Since Ψ_k is μ -strongly convex and has a locally Lipschitz gradient on B_k , gradient descent on Ψ_k starting at x_k with *Armijo backtracking* enjoys linear convergence. Hence there exist $\rho \in (0, 1)$ and $C > 0$ (independent of k) such that

$$\|x^t - x_k^\Psi\| \leq \rho^t \|x_k - x_k^\Psi\|, \quad \|\nabla \Psi_k(x^t)\| \leq C \rho^t \|x_k - x_k^\Psi\|.$$

Therefore, for any $\alpha > 0$,

$$\|\nabla \Psi_k(x^T)\|^2 \leq \eta \frac{\|x_k - x_k^\Psi\|^2}{T^\alpha},$$

with $\eta = C^2 C_\alpha$ where C_α satisfies $\rho^{2T} \leq C_\alpha / T^\alpha$ for all $T \geq 1$.

APPENDIX C. PROOF OF THEOREMS 6.1 AND 6.2

For convenience, we restate Theorem 6.1 here.

Theorem C.1. *Let Assumptions 2.1 (Differentiability), 2.2 (Convexity), 2.3 (Strong convexity of f), 2.8 (Star Similarity) and 6.1 (Bounded Variance at Optimum) hold. If the stepsize satisfies $\gamma \leq \frac{\mu}{48\delta_*^2}$, we have, for every $k \geq 0$,*

$$\mathbb{E}[\|x_k - x_*\|^2] \leq \left(1 - \min\left(\frac{\gamma\mu}{4}, \frac{1}{2}\right)\right)^k \|x_0 - x_*\|^2 + \max\left(\frac{4}{\gamma\mu}, 2\right) 2\gamma^2 \sigma_*^2. \quad (\text{C.1})$$

Proof. Define $\bar{x}_{k+1} = \mathbb{E}[x_{k+1} | \mathcal{F}_k]$. Then

$$\begin{aligned} & \|x_{k+1} - x_*\|^2 \\ &= \|x_k - x_*\|^2 - 2\langle x_{k+1} - x_*, x_k - x_{k+1} \rangle - \|x_k - x_{k+1}\|^2 \\ &\leq \|x_k - x_*\|^2 - 2\gamma(f_{\xi_k}(x_{k+1}) - f_{\xi_k}(x_*)) - \|x_k - x_{k+1}\|^2 \\ &\leq \|x_k - x_*\|^2 - 2\gamma(f_{\xi_k}(\bar{x}_{k+1}) + \langle \nabla f_{\xi_k}(\bar{x}_{k+1}), x_{k+1} - \bar{x}_{k+1} \rangle - f_{\xi_k}(x_*)) - \|x_k - x_{k+1}\|^2. \end{aligned}$$

Taking the expectation conditioned on \mathcal{F}_k and using the equality $\mathbb{E}[\|x - c\|^2] = \mathbb{E}[\|x - \mathbb{E}[x]\|^2] + \|\mathbb{E}[x] - c\|^2$, we have

$$\begin{aligned} & \mathbb{E}[\|x_{k+1} - x_*\|^2 | \mathcal{F}_k] \\ &\leq \|x_k - x_*\|^2 - 2\gamma \mathbb{E}[f_{\xi_k}(\bar{x}_{k+1}) + \langle \nabla f_{\xi_k}(\bar{x}_{k+1}), x_{k+1} - \bar{x}_{k+1} \rangle - f_{\xi_k}(x_*) | \mathcal{F}_k] \\ &\quad - \mathbb{E}[\|x_k - x_{k+1}\|^2 | \mathcal{F}_k] \\ &= \|x_k - x_*\|^2 - 2\gamma(f(\bar{x}_{k+1}) - f(x_*) + \mathbb{E}[\langle \nabla f_{\xi_k}(\bar{x}_{k+1}), x_{k+1} - \bar{x}_{k+1} \rangle | \mathcal{F}_k]) \\ &\quad - \mathbb{E}[\|x_{k+1} - \bar{x}_{k+1}\|^2 | x_k] - \|x_k - \bar{x}_{k+1}\|^2. \end{aligned} \quad (\text{C.2})$$

Since $\mathbb{E}[\langle f(\bar{x}_{k+1}), x_{k+1} - \bar{x}_{k+1} \rangle | \mathcal{F}_k] = 0$, we can write

$$\begin{aligned} & \mathbb{E}[\|x_{k+1} - x_*\|^2 | \mathcal{F}_k] \\ &\leq \|x_k - x_*\|^2 - 2\gamma(f(\bar{x}_{k+1}) - f(x_*)) + \mathbb{E}[\langle \nabla f_{\xi_k}(\bar{x}_{k+1}), x_{k+1} - \bar{x}_{k+1} \rangle | \mathcal{F}_k] \\ &\quad - \mathbb{E}[\|x_{k+1} - \bar{x}_{k+1}\|^2 | \mathcal{F}_k] - \|x_k - \bar{x}_{k+1}\|^2 \\ &= \|x_k - x_*\|^2 - 2\gamma(f(\bar{x}_{k+1}) - f(x_*)) \\ &\quad - 2\gamma \mathbb{E}[\langle \nabla f_{\xi_k}(\bar{x}_{k+1}) - \nabla f(\bar{x}_{k+1}) - \nabla f_{\xi_k}(x_*), x_{k+1} - \bar{x}_{k+1} \rangle | \mathcal{F}_k] \\ &\quad + 2\gamma(\mathbb{E}[\langle \nabla f_{\xi_k}(x_*), \bar{x}_{k+1} - x_{k+1} \rangle | \mathcal{F}_k]) - \mathbb{E}[\|x_{k+1} - \bar{x}_{k+1}\|^2 | \mathcal{F}_k] - \|x_k - \bar{x}_{k+1}\|^2. \end{aligned}$$

Let $a \in (0, 1)$. Then

$$\begin{aligned} & \mathbb{E}[\|x_{k+1} - x_*\|^2 | \mathcal{F}_k] \\ &\leq \|x_k - x_*\|^2 - 2\gamma(f(\bar{x}_{k+1}) - f(x_*)) \\ &\quad - \frac{2}{a} \mathbb{E}[\langle \gamma(\nabla f_{\xi_k}(\bar{x}_{k+1}) - \nabla f(\bar{x}_{k+1}) - \nabla f_{\xi_k}(x_*)), a(x_{k+1} - \bar{x}_{k+1}) \rangle | \mathcal{F}_k] \\ &\quad + 2\gamma(\mathbb{E}[\langle \nabla f_{\xi_k}(x_*), \bar{x}_{k+1} - x_{k+1} \rangle | \mathcal{F}_k]) - \mathbb{E}[\|x_{k+1} - \bar{x}_{k+1}\|^2 | \mathcal{F}_k] - \|x_k - \bar{x}_{k+1}\|^2. \end{aligned}$$

Using strong convexity of f and the identity $2\langle c, b \rangle = \|c + b\|^2 - \|c\|^2 - \|b\|^2$, we obtain

$$\begin{aligned}
& \mathbb{E} [\|x_{k+1} - x_*\|^2] \\
& \leq \|x_k - x_*\|^2 - \gamma\mu \|\bar{x}_{k+1} - x_*\|^2 \\
& \quad - \frac{1}{a} \mathbb{E} [\|\gamma(\nabla f_{\xi_k}(\bar{x}_{k+1}) - \nabla f(\bar{x}_{k+1}) - \nabla f_{\xi_k}(x_*)) + a(x_{k+1} - \bar{x}_{k+1})\|^2 \mid \mathcal{F}_k] \\
& \quad + \frac{\gamma^2}{a} \mathbb{E} [\|(\nabla f_{\xi_k}(\bar{x}_{k+1}) - \nabla f(\bar{x}_{k+1}) - \nabla f_{\xi_k}(x_*))\|^2 \mid \mathcal{F}_k] - (1-a) \mathbb{E} [\|x_{k+1} - \bar{x}_{k+1}\|^2 \mid \mathcal{F}_k] \\
& \quad + 2\gamma (\mathbb{E} [\langle \nabla f_{\xi_k}(x_*), \bar{x}_{k+1} - x_{k+1} \rangle \mid \mathcal{F}_k]) - \|x_k - \bar{x}_{k+1}\|^2.
\end{aligned}$$

Using Assumption 2.8 and Young's inequality, we see that

$$\begin{aligned}
& \mathbb{E} [\|x_{k+1} - x_*\|^2] \\
& \leq \|x_k - x_*\|^2 - \gamma\mu \|\bar{x}_{k+1} - x_*\|^2 + \frac{\gamma^2 \delta_*^2}{a} \|\bar{x}_{k+1} - x_*\|^2 - (1-a) \mathbb{E} [\|x_{k+1} - \bar{x}_{k+1}\|^2 \mid x_k] \\
& \quad + \frac{\gamma^2}{s} \mathbb{E} [\|\nabla f_{\xi_k}(x_*)\|^2 \mid x_k] + s \mathbb{E} [\|x_{k+1} - \bar{x}_{k+1}\|^2 \mid x_k] - \|x_k - \bar{x}_{k+1}\|^2.
\end{aligned}$$

By substituting $s = 1 - a$ and choosing $a = \frac{1}{2}$, we conclude

$$\begin{aligned}
& \mathbb{E} [\|x_{k+1} - x_*\|^2 \mid \mathcal{F}_k] \\
& \leq \|x_k - x_*\|^2 - \gamma\mu \|\bar{x}_{k+1} - x_*\|^2 + 2\gamma^2 \delta_*^2 \|\bar{x}_{k+1} - x_*\|^2 + 2\gamma^2 \mathbb{E} [\|\nabla f_{\xi_k}(x_*)\|^2 \mid x_k] \\
& \quad - \|x_k - \bar{x}_{k+1}\|^2.
\end{aligned}$$

We require $2\gamma^2 \delta_*^2 \leq \frac{\gamma\mu}{2}$, which implies $\gamma \leq \frac{\mu}{4\delta_*^2}$. Therefore,

$$\begin{aligned}
& \mathbb{E} [\|x_{k+1} - x_*\|^2 \mid \mathcal{F}_k] \\
& \leq \|x_k - x_*\|^2 - \frac{\gamma\mu}{2} \|\bar{x}_{k+1} - x_*\|^2 - \|x_k - \bar{x}_{k+1}\|^2 + 2\gamma^2 \mathbb{E} [\|\nabla f_{\xi_k}(x_*)\|^2 \mid x_k] \\
& \leq \|x_k - x_*\|^2 - \min\left(\frac{\gamma\mu}{2}, 1\right) (\|\bar{x}_{k+1} - x_*\|^2 - \|x_k - \bar{x}_{k+1}\|^2) + 2\gamma^2 \mathbb{E} [\|\nabla f_{\xi_k}(x_*)\|^2 \mid x_k] \\
& \leq \left(1 - \min\left(\frac{\gamma\mu}{4}, \frac{1}{2}\right)\right) \|x_k - x_*\|^2 + 2\gamma^2 \sigma_*^2.
\end{aligned}$$

Taking the full expectation, we obtain

$$\mathbb{E} [\|x_{k+1} - x_*\|^2] \leq \left(1 - \min\left(\frac{\gamma\mu}{4}, \frac{1}{2}\right)\right) \mathbb{E} [\|x_k - x_*\|^2] + 2\gamma^2 \sigma_*^2.$$

By applying the inequality recursively, we derive

$$\begin{aligned}
 & \mathbb{E}[\|x_{k+1} - x_*\|^2] \\
 & \leq \left(1 - \min\left(\frac{\gamma\mu}{4}, \frac{1}{2}\right)\right)^{k+1} \|x_0 - x_*\|^2 \\
 & \quad + \left(\left(1 - \min\left(\frac{\gamma\mu}{4}, \frac{1}{2}\right)\right)^0 + \dots + \left(1 - \min\left(\frac{\gamma\mu}{4}, \frac{1}{2}\right)\right)^k\right) 2\gamma^2\sigma_*^2 \\
 & \leq \left(1 - \min\left(\frac{\gamma\mu}{4}, \frac{1}{2}\right)\right)^{k+1} \|x_0 - x_*\|^2 + \max\left(\frac{4}{\gamma\mu}, 2\right) 2\gamma^2\sigma_*^2.
 \end{aligned}$$

□

For convenience, we restate Theorem 6.2 here:

Theorem C.2. *Let Assumptions 2.1 (Differentiability), 2.2 (Convexity), 2.8 (Star Similarity) and 6.1 (Bounded Variance) hold. Consider SPPM-inexact with Assumption 2.9 satisfied. If the stepsize satisfies $\gamma \leq \frac{\mu(1-c)}{4\delta_*^2}$ and T is chosen sufficiently large such that $\frac{\eta\gamma^2}{T\alpha} \leq c$, where $0 < c < 1$ is a constant, then SPPM-inexact satisfies, for every $k \geq 0$,*

$$\mathbb{E}[\|x_k - x_*\|^2] \leq \left(1 - \frac{1}{2} \min\left(\frac{\gamma\mu}{2}, 1-c\right)\right)^k \|x_0 - x_*\|^2 + \max\left(\frac{2}{\gamma\mu}, \frac{1}{1-c}\right) \frac{4\gamma^2\sigma_*^2}{(1-c)}.$$

Proof. To avoid repetition, we start from (D.4):

$$\begin{aligned}
 & \mathbb{E}[\|x_{k+1} - x_*\|^2 | x_k] \\
 & \leq \|x_k - x_*\|^2 - 2\gamma \left(f(\bar{x}_{k+1}) - f(x_*) + \mathbb{E} \left[\langle \nabla f_{\xi_k}(\bar{x}_{k+1}), x_k^\Psi - \bar{x}_{k+1} \rangle | x_k \right] \right) \\
 & \quad - \left(1 - \frac{\eta\gamma^2}{T\alpha}\right) \mathbb{E}[\|x_k^\Psi - \bar{x}_{k+1}\|^2 | x_k] - \left(1 - \frac{\eta\gamma^2}{T\alpha}\right) \|x_k - \bar{x}^{k+1}\|^2.
 \end{aligned}$$

Now since $\mathbb{E}[\langle f(\bar{x}_{k+1}), x_k^\Psi - \bar{x}_{k+1} \rangle | x_k] = 0$, we can write

$$\begin{aligned}
 & \mathbb{E}[\|x_{k+1} - x_*\|^2 | x_k] \\
 & \leq \|x_k - x_*\|^2 - 2\gamma(f(\bar{x}_{k+1}) - f(x_*)) \\
 & \quad - 2\gamma \mathbb{E} \left[\langle \nabla f_{\xi_k}(\bar{x}_{k+1}) - \nabla f(\bar{x}_{k+1}) - \nabla f_{\xi_k}(x_*), x_k^\Psi - \bar{x}_{k+1} \rangle | x_k \right] \\
 & \quad + 2\gamma \left(\mathbb{E} \left[\langle \nabla f_{\xi_k}(x_*), \bar{x}_{k+1} - x_k^\Psi \rangle | x_k \right] \right) - \left(1 - \frac{\eta\gamma^2}{T\alpha}\right) \mathbb{E}[\|x_k^\Psi - \bar{x}_{k+1}\|^2 | x_k] \\
 & \quad - \left(1 - \frac{\eta\gamma^2}{T\alpha}\right) \|x_k - \bar{x}_{k+1}\|^2.
 \end{aligned}$$

Let $a \in (0, 1)$. Then

$$\begin{aligned}
& \mathbb{E} [\|x_{k+1} - x_*\|^2 \mid x_k] \\
& \leq \|x_k - x_*\|^2 - 2\gamma(f(\bar{x}_{k+1}) - f(x_*)) \\
& \quad - 2\frac{1}{a} \left(\mathbb{E} \left[\langle \gamma(\nabla f_{\xi_k}(\bar{x}_{k+1}) - \nabla f(\bar{x}_{k+1}) - \nabla f_{\xi_k}(x_*)), a(x_k^\Psi - \bar{x}_{k+1}) \rangle \mid x_k \right] \right) \\
& \quad + 2\gamma \left(\mathbb{E} \left[\langle \nabla f_{\xi_k}(x_*), \bar{x}_{k+1} - x_k^\Psi \rangle \mid x_k \right] \right) - \left(1 - \frac{\eta\gamma^2}{T\alpha}\right) \mathbb{E} [\|x_k^\Psi - \bar{x}_{k+1}\|^2 \mid x_k] \\
& \quad - \left(1 - \frac{\eta\gamma^2}{T\alpha}\right) \|x_k - \bar{x}_{k+1}\|^2.
\end{aligned}$$

Using strong convexity of f and the identity $2\langle c, b \rangle = \|c + b\|^2 - \|c\|^2 - \|b\|^2$, we obtain

$$\begin{aligned}
& \mathbb{E} [\|x_{k+1} - x_*\|^2 \mid x_k] \\
& \leq \|x_k - x_*\|^2 - \gamma\mu \|\bar{x}_{k+1} - x_*\|^2 \\
& \quad - \frac{1}{a} \left(\mathbb{E} \left[\|\gamma(\nabla f_{\xi_k}(\bar{x}_{k+1}) - \nabla f(\bar{x}_{k+1}) - \nabla f_{\xi_k}(x_*)) + a(x_k^\Psi - \bar{x}_{k+1})\|^2 \mid x_k \right] \right) \\
& \quad + \frac{\gamma^2}{a} \mathbb{E} [\|(\nabla f_{\xi_k}(\bar{x}_{k+1}) - \nabla f(\bar{x}_{k+1}) - \nabla f_{\xi_k}(x_*))\|^2 \mid x_k] + a \mathbb{E} [\|x_k^\Psi - \bar{x}_{k+1}\|^2 \mid x_k] \\
& \quad + 2\gamma \left(\mathbb{E} \left[\langle \nabla f_{\xi_k}(x_*), \bar{x}_{k+1} - x_k^\Psi \rangle \mid x_k \right] \right) - \left(1 - \frac{\eta\gamma^2}{T\alpha}\right) \mathbb{E} [\|x_k^\Psi - \bar{x}_{k+1}\|^2 \mid x_k] \\
& \quad - \left(1 - \frac{\eta\gamma^2}{T\alpha}\right) \|x_k - \bar{x}_{k+1}\|^2.
\end{aligned}$$

Using Assumption 2.8 and Young's inequality, we get

$$\begin{aligned}
& \mathbb{E} [\|x_{k+1} - x_*\|^2 \mid x_k] \\
& \leq \|x_k - x_*\|^2 - \gamma\mu \|\bar{x}_{k+1} - x_*\|^2 + \frac{\gamma^2 \delta_*^2}{a} \|\bar{x}_{k+1} - x_*\|^2 + a \mathbb{E} [\|x_k^\Psi - \bar{x}_{k+1}\|^2 \mid x_k] \\
& \quad + \frac{\gamma^2}{s} \mathbb{E} [\|\nabla f_{\xi_k}(x_*)\|^2 \mid x_k] + s \mathbb{E} [\|\bar{x}_{k+1} - x_k^\Psi\|^2 \mid x_k] \\
& \quad - \left(1 - \frac{\eta\gamma^2}{T\alpha}\right) \mathbb{E} [\|x_k^\Psi - \bar{x}_{k+1}\|^2 \mid x_k] - \left(1 - \frac{\eta\gamma^2}{T\alpha}\right) \|x_k - \bar{x}_{k+1}\|^2.
\end{aligned}$$

Choosing $s = a = \frac{1 - \frac{\eta\gamma^2}{T\alpha}}{2}$, we derive

$$\begin{aligned}
& \mathbb{E} [\|x_{k+1} - x_*\|^2 \mid x_k] \\
& \leq \|x_k - x_*\|^2 - \gamma\mu \|\bar{x}_{k+1} - x_*\|^2 + \frac{2\gamma^2 \delta_*^2}{1 - \frac{\eta\gamma^2}{T\alpha}} \|\bar{x}_{k+1} - x_*\|^2 \\
& \quad + \frac{2\gamma^2}{1 - \frac{\eta\gamma^2}{T\alpha}} \mathbb{E} [\|\nabla f_{\xi_k}(x_*)\|^2 \mid x_k] - \left(1 - \frac{\eta\gamma^2}{T\alpha}\right) \|x_k - \bar{x}_{k+1}\|^2.
\end{aligned}$$

Let T be large enough such that $\frac{\eta\gamma^2}{T\alpha} \leq c < 1$. Then

$$\begin{aligned} & \mathbb{E} [\|x_{k+1} - x_*\|^2 \mid x_k] \\ & \leq \|x_k - x_*\|^2 - \gamma\mu \|\bar{x}_{k+1} - x_*\|^2 + \frac{2\gamma^2\delta_*^2}{1-c} \|\bar{x}_{k+1} - x_*\|^2 + \frac{2\gamma^2}{1-c} \mathbb{E} [\|\nabla f_{\xi_k}(x_*)\|^2 \mid x_k] \\ & \quad - (1-c) \|x_k - \bar{x}_{k+1}\|^2. \end{aligned}$$

We want $2\gamma^2\delta_*^2 \leq \frac{\gamma\mu(1-c)}{2}$, which implies $\gamma \leq \frac{\mu(1-c)}{4\delta_*^2}$, so we get

$$\begin{aligned} & \mathbb{E} [\|x_{k+1} - x_*\|^2 \mid x_k] \\ & \leq \|x_k - x_*\|^2 - \frac{\gamma\mu}{2} \|\bar{x}_{k+1} - x_*\|^2 - (1-c) \|x_k - \bar{x}_{k+1}\|^2 + \frac{2\gamma^2}{1-c} \mathbb{E} [\|\nabla f_{\xi_k}(x_*)\|^2 \mid x_k] \\ & \leq \|x_k - x_*\|^2 - \frac{1}{2} \min\left(\frac{\gamma\mu}{2}, 1-c\right) \|x_k - \bar{x}_{k+1}\|^2 + \frac{2\gamma^2}{1-c} \mathbb{E} [\|\nabla f_{\xi_k}(x_*)\|^2 \mid x_k] \\ & \stackrel{6.1}{\leq} \left(1 - \frac{1}{2} \min\left(\frac{\gamma\mu}{2}, 1-c\right)\right) \|x_k - x_*\|^2 + \frac{2\gamma^2\sigma_*^2}{1-c}. \end{aligned}$$

Taking the full expectation, we obtain

$$\mathbb{E} [\|x_{k+1} - x_*\|^2] \leq \left(1 - \frac{1}{2} \min\left(\frac{\gamma\mu}{2}, 1-c\right)\right) \mathbb{E} [\|x_k - x_*\|^2] + \frac{2\gamma^2\sigma_*^2}{1-c}.$$

By applying the inequality recursively, we derive

$$\begin{aligned} & \mathbb{E} [\|x_{k+1} - x_*\|^2] \\ & \leq \left(1 - \frac{1}{2} \min\left(\frac{\gamma\mu}{2}, 1-c\right)\right)^{k+1} \|x_0 - x_*\|^2 \\ & \quad + \left(\left(1 - \frac{1}{2} \min\left(\frac{\gamma\mu}{2}, 1-c\right)\right)^0 + \dots + \left(1 - \frac{1}{2} \min\left(\frac{\gamma\mu}{2}, 1-c\right)\right)^k\right) \frac{2\gamma^2\sigma_*^2}{1-c} \\ & \leq \left(1 - \frac{1}{2} \min\left(\frac{\gamma\mu}{2}, 1-c\right)\right)^{k+1} \|x_0 - x_*\|^2 + \max\left(\frac{2}{\gamma\mu}, \frac{1}{1-c}\right) \frac{4\gamma^2\sigma_*^2}{(1-c)}. \end{aligned}$$

□

APPENDIX D. CONVERGENCE UNDER INTERPOLATION REGIME AND EXPECTED SIMILARITY

Theorem D.1. *Let Assumptions 2.1 (Differentiability), 2.2 (Convexity), 2.3 (Strong convexity of f), 2.4 (Interpolation), and 2.8 (Star Similarity) hold. If the stepsize satisfies $\gamma \leq \frac{\mu}{2\delta_*^2}$, then we have, for every $k \geq 0$,*

$$\mathbb{E} [\|x_k - x_*\|^2] \leq \left(1 - \min\left(\frac{\gamma\mu}{4}, \frac{1}{2}\right)\right)^k \|x_0 - x_*\|^2. \quad (\text{D.1})$$

Proof. Define

$$\bar{x}_{k+1} = \mathbb{E}[x_{k+1} \mid \mathcal{F}_k]. \quad (\text{D.2})$$

Then

$$\begin{aligned}
& \|x_{k+1} - x_*\|^2 \\
&= \|x_k - x_*\|^2 - 2\langle x_{k+1} - x_*, x_k - x_{k+1} \rangle - \|x_k - x_{k+1}\|^2 \\
&\stackrel{A.2}{\leq} \|x_k - x_*\|^2 - 2\gamma(f_{\xi_k}(x_{k+1}) - f_{\xi_k}(x_*)) - \|x_k - x_{k+1}\|^2 \\
&\stackrel{2.2}{\leq} \|x_k - x_*\|^2 - 2\gamma(f_{\xi_k}(\bar{x}_{k+1}) + \langle \nabla f_{\xi_k}(\bar{x}_{k+1}), x_{k+1} - \bar{x}_{k+1} \rangle - f_{\xi_k}(x_*)) - \|x_k - x_{k+1}\|^2.
\end{aligned}$$

Taking the expectation conditioned on \mathcal{F}_k and using the identity $\mathbb{E}[\|x - c\|^2] = E[\|x - \mathbb{E}[x]\|^2] + \|\mathbb{E}[x] - c\|^2$, we obtain

$$\begin{aligned}
& \mathbb{E}[\|x_{k+1} - x_*\|^2 | \mathcal{F}_k] \\
&\leq \|x_k - x_*\|^2 - 2\gamma \mathbb{E}[f_{\xi_k}(\bar{x}_{k+1}) + \langle \nabla f_{\xi_k}(\bar{x}_{k+1}), x_{k+1} - \bar{x}_{k+1} \rangle - f_{\xi_k}(x_*) | \mathcal{F}_k] \\
&\quad - \mathbb{E}[\|x_k - x_{k+1}\|^2 | \mathcal{F}_k] \\
&= \|x_k - x_*\|^2 - 2\gamma(f(\bar{x}_{k+1}) - f(x_*) + \mathbb{E}[\langle \nabla f_{\xi_k}(\bar{x}_{k+1}), x_{k+1} - \bar{x}_{k+1} \rangle | \mathcal{F}_k]) \\
&\quad - \mathbb{E}[\|x_{k+1} - \bar{x}_{k+1}\|^2 | \mathcal{F}_k] - \|x_k - \bar{x}_{k+1}\|^2. \tag{D.3}
\end{aligned}$$

Since $\mathbb{E}[\langle f(\bar{x}_{k+1}), x_{k+1} - \bar{x}_{k+1} \rangle | \mathcal{F}_k] = 0$, we can write

$$\begin{aligned}
\mathbb{E}[\|x_{k+1} - x_*\|^2 | x_k] &= \|x_k - x_*\|^2 - 2\gamma(f(\bar{x}_{k+1}) - f(x_*)) \\
&\quad + \mathbb{E}[\langle \nabla f_{\xi_k}(\bar{x}_{k+1}) - \nabla f(\bar{x}_{k+1}), x_{k+1} - \bar{x}_{k+1} \rangle | x_k] \\
&\quad - \mathbb{E}[\|x_{k+1} - \bar{x}_{k+1}\|^2 | x_k] - \|x_k - \bar{x}_{k+1}\|^2.
\end{aligned}$$

Using strong convexity of f and the identity $2\langle a, b \rangle = \|a + b\|^2 - \|a\|^2 - \|b\|^2$, we derive

$$\begin{aligned}
& \mathbb{E}[\|x_{k+1} - x_*\|^2 | \mathcal{F}_k] \\
&\leq \|x_k - x_*\|^2 - \gamma\mu\|\bar{x}_{k+1} - x_*\|^2 - \mathbb{E}[\|x_{k+1} - \bar{x}_{k+1}\|^2 | \mathcal{F}_k] - \|x_k - \bar{x}_{k+1}\|^2 \\
&\quad - \mathbb{E}\left[\|\gamma(\nabla f_{\xi_k}(\bar{x}_{k+1}) - \nabla f(\bar{x}_{k+1})) + (x_{k+1} - \bar{x}_{k+1})\|^2\right. \\
&\quad \left. - \gamma^2\|\nabla f_{\xi_k}(\bar{x}_{k+1}) - \nabla f(\bar{x}_{k+1})\|^2 - \|x_{k+1} - \bar{x}_{k+1}\|^2 | \mathcal{F}_k\right] \\
&\leq \|x_k - x_*\|^2 - \gamma\mu\|\bar{x}_{k+1} - x_*\|^2 - \|x_k - \bar{x}_{k+1}\|^2 \\
&\quad + \gamma^2\mathbb{E}[\|\nabla f_{\xi_k}(\bar{x}_{k+1}) - \nabla f(\bar{x}_{k+1})\|^2 | \mathcal{F}_k].
\end{aligned}$$

Finally, using the star similarity condition, we obtain

$$\mathbb{E}[\|x_{k+1} - x_*\|^2 | \mathcal{F}_k] \leq \|x_k - x_*\|^2 - (\gamma\mu - \gamma^2\delta_*^2)\|\bar{x}_{k+1} - x_*\|^2 - \|x_k - \bar{x}_{k+1}\|^2.$$

If $\gamma^2 \delta_*^2 \leq \frac{1}{2} \gamma \mu$, then $\gamma \leq \frac{\mu}{2\delta_*^2}$. Under this condition, we have

$$\begin{aligned} & \mathbb{E} [\|x_{k+1} - x_*\|^2 | \mathcal{F}_k] \\ & \leq \|x_k - x_*\|^2 - \frac{\gamma\mu}{2} \|\bar{x}_{k+1} - x_*\|^2 - \|x_k - \bar{x}_{k+1}\|^2 \\ & \leq \|x_k - x_*\|^2 - \min\left(\frac{\gamma\mu}{2}, 1\right) (\|\bar{x}_{k+1} - x_*\|^2 - \|x_k - \bar{x}_{k+1}\|^2) \\ & \leq \|x_k - x_*\|^2 - \min\left(\frac{\gamma\mu}{4}, \frac{1}{2}\right) \|x_k - x_*\|^2 = \left(1 - \min\left(\frac{\gamma\mu}{4}, \frac{1}{2}\right)\right) \|x_k - x_*\|^2. \end{aligned}$$

Taking the full expectation, we obtain

$$\mathbb{E} [\|x_{k+1} - x_*\|^2] \leq \left(1 - \min\left(\frac{\gamma\mu}{4}, \frac{1}{2}\right)\right) \mathbb{E} [\|x_k - x_*\|^2].$$

Applying this inequality recursively, we get

$$\mathbb{E} [\|x_{k+1} - x_*\|^2] \leq \left(1 - \min\left(\frac{\gamma\mu}{4}, \frac{1}{2}\right)\right)^{k+1} \|x_0 - x_*\|^2.$$

□

Theorem D.2. *Let Assumptions 2.1 (Differentiability), 2.2 (Convexity), 2.4 (Interpolation) and 2.8 (Star Similarity) hold. Consider SPPM-inexact with Assumption 2.9 satisfied. If the stepsize satisfies $\gamma \leq \frac{\mu(1-c)}{2\delta_*^2}$ and T is chosen sufficiently large such that $\frac{\eta\gamma^2}{T\alpha} \leq c < 1$, then the iterates of SPPM-inexact satisfy, for every $k \geq 0$,*

$$\mathbb{E} [\|x_k - x_*\|^2] \leq \left(1 - \frac{1}{2} \min\left(\frac{\gamma\mu}{2}, 1 - c\right)\right)^k \|x_0 - x_*\|^2.$$

Proof. To avoid repetitions, we start from (B.8):

$$\|x_{k+1} - x_*\| \leq \|x_k - x_*\|^2 - 2\gamma(f_{\xi_k}^{\Psi}(x_k^{\Psi}) - f_{\xi_k}^{\xi}(x_*)) - \left(1 - \frac{\eta\gamma^2}{T\alpha}\right) \|x_k - x_k^{\Psi}\|^2.$$

Instead of (D.2), we define $\bar{x}_{k+1} = \mathbb{E}[x_k^{\Psi} | x_k]$.

Using convexity of $f_{\xi_k}^{\xi}$, we get

$$\begin{aligned} \|x_{k+1} - x_*\| & \leq \|x_k - x_*\|^2 - 2\gamma(f_{\xi_k}^{\xi}(\bar{x}_{k+1}) + \langle \nabla f_{\xi_k}^{\xi}(\bar{x}_{k+1}), x_k^{\Psi} - \bar{x}_{k+1} \rangle - f_{\xi_k}^{\xi}(x_*)) \\ & \quad - \left(1 - \frac{\eta\gamma^2}{T\alpha}\right) \|x_k - x_k^{\Psi}\|^2. \end{aligned}$$

Taking the expectation conditioned on x_k and using the equality $\mathbb{E}[\|x - c\|^2] = \mathbb{E}[\|x - E[x]\|^2] + \|\mathbb{E}[x] - c\|^2$, we obtain

$$\begin{aligned}
& \mathbb{E} [\|x_{k+1} - x_*\|^2 | x_k] \\
& \leq \|x_k - x_*\|^2 - \left(1 - \frac{\eta\gamma^2}{T\alpha}\right) E \left[\|x_k - x_k^\Psi\|^2 | x_k \right] \\
& \quad - 2\gamma \mathbb{E} \left[f_{\xi_k}(\bar{x}_{k+1}) + \langle \nabla f_{\xi_k}(\bar{x}_{k+1}), x_k^\Psi - \bar{x}_{k+1} \rangle - f_{\xi_k}(x_*) | x_k \right] \\
& = \|x_k - x_*\|^2 - 2\gamma \left(f(\bar{x}_{k+1}) - f(x_*) + \mathbb{E} \left[\langle \nabla f_{\xi_k}(\bar{x}_{k+1}), x_k^\Psi - \bar{x}_{k+1} \rangle | x_k \right] \right) \\
& \quad - \left(1 - \frac{\eta\gamma^2}{T\alpha}\right) \mathbb{E} \left[\|x_k^\Psi - \bar{x}_{k+1}\|^2 | x_k \right] - \left(1 - \frac{\eta\gamma^2}{T\alpha}\right) \|x_k - \bar{x}^{k+1}\|^2. \tag{D.4}
\end{aligned}$$

Since $\mathbb{E} [\langle f(\bar{x}_{k+1}), x_k^\Psi - \bar{x}_{k+1} \rangle | x_k] = 0$, we can write

$$\begin{aligned}
& \mathbb{E} [\|x_{k+1} - x_*\|^2 | x_k] \\
& \leq \|x_k - x_*\|^2 - 2\gamma \left(f(\bar{x}_{k+1}) - f(x_*) + \mathbb{E} \left[\langle \nabla f_{\xi_k}(\bar{x}_{k+1}) - f(\bar{x}_{k+1}), x_k^\Psi - \bar{x}_{k+1} \rangle | x_k \right] \right) \\
& \quad - \left(1 - \frac{\eta\gamma^2}{T\alpha}\right) \mathbb{E} \left[\|x_k^\Psi - \bar{x}_{k+1}\|^2 | x_k \right] - \left(1 - \frac{\eta\gamma^2}{T\alpha}\right) \|x_k - \bar{x}^{k+1}\|^2. \tag{D.5}
\end{aligned}$$

Let $a \in (0, 1)$. Then

$$\begin{aligned}
\mathbb{E} [\|x_{k+1} - x_*\|^2 | x_k] & \leq \|x_k - x_*\|^2 - 2\gamma (f(\bar{x}_{k+1}) - f(x_*)) \\
& \quad - 2\frac{1}{a} \mathbb{E} \left[\langle \gamma(\nabla f_{\xi_k}(\bar{x}_{k+1}) - f(\bar{x}_{k+1})), a(x_k^\Psi - \bar{x}_{k+1}) \rangle | x_k \right] \\
& \quad - \left(1 - \frac{\eta\gamma^2}{T\alpha}\right) \mathbb{E} \left[\|x_k^\Psi - \bar{x}_{k+1}\|^2 | x_k \right] - \left(1 - \frac{\eta\gamma^2}{T\alpha}\right) \|x_k - \bar{x}^{k+1}\|^2.
\end{aligned}$$

By applying strong convexity of f and using the identity $2\langle c, b \rangle = \|c + b\|^2 - \|c\|^2 - \|b\|^2$, we have

$$\begin{aligned}
\mathbb{E} [\|x_{k+1} - x_*\|^2 | x_k] & \leq \|x_k - x_*\|^2 - \gamma\mu \|\bar{x}_{k+1} - x_*\|^2 - \left(1 - \frac{\eta\gamma^2}{T\alpha}\right) \mathbb{E} \left[\|x_k^\Psi - \bar{x}_{k+1}\|^2 | x_k \right] \\
& \quad - \left(1 - \frac{\eta\gamma^2}{T\alpha}\right) \|x_k - \bar{x}_{k+1}\|^2 \\
& \quad - \frac{1}{a} \mathbb{E} \left[\|\gamma(\nabla f_{\xi_k}(\bar{x}_{k+1}) - \nabla f(\bar{x}_{k+1})) + (x_k^\Psi - \bar{x}_{k+1})\|^2 | x_k \right] \\
& \quad + \frac{\gamma^2}{a} \mathbb{E} \left[\|\nabla f_{\xi_k}(\bar{x}_{k+1}) - \nabla f(\bar{x}_{k+1})\|^2 | x_k \right] + a \mathbb{E} \left[\|\bar{x}_{k+1} - x_k^\Psi\|^2 | x_k \right].
\end{aligned}$$

By setting $a = \left(1 - \frac{\eta\gamma^2}{T\alpha}\right)$, we obtain

$$\begin{aligned}
\mathbb{E} [\|x_{k+1} - x_*\|^2 | x_k] & \leq \|x_k - x_*\|^2 - \gamma\mu \|\bar{x}_{k+1} - x_*\|^2 - \left(1 - \frac{\eta\gamma^2}{T\alpha}\right) \|x_k - \bar{x}_{k+1}\|^2 \\
& \quad + \frac{\gamma^2}{\left(1 - \frac{\eta\gamma^2}{T\alpha}\right)} \mathbb{E} \left[\|\nabla f_{\xi_k}(\bar{x}_{k+1}) - \nabla f(\bar{x}_{k+1})\|^2 | x_k \right].
\end{aligned}$$

Now, using the star similarity assumption (2.8), we get

$$\begin{aligned} \mathbb{E} [\|x_{k+1} - x_*\|^2 | x_k] &\leq \|x_k - x_*\|^2 - \gamma\mu \|\bar{x}_{k+1} - x_*\|^2 - \left(1 - \frac{\eta\gamma^2}{T\alpha}\right) \|x_k - \bar{x}_{k+1}\|^2 \\ &\quad + \frac{\gamma^2 \delta_*^2}{\left(1 - \frac{\eta\gamma^2}{T\alpha}\right)} \|\bar{x}_{k+1} - x_*\|^2. \end{aligned}$$

Assuming T sufficiently large such that $\frac{\eta\gamma^2}{T\alpha} \leq c < 1$, we obtain

$$\begin{aligned} \mathbb{E} [\|x_{k+1} - x_*\|^2 | x_k] &\leq \|x_k - x_*\|^2 - \gamma\mu \|\bar{x}_{k+1} - x_*\|^2 - (1 - c) \|x_k - \bar{x}_{k+1}\|^2 \\ &\quad + \frac{\gamma^2 \delta_*^2}{1 - c} \|\bar{x}_{k+1} - x_*\|^2. \end{aligned}$$

We want $\frac{\gamma^2 \delta_*^2}{1 - c} \leq \frac{\gamma\mu}{2}$, which means $\gamma \leq \frac{\mu(1 - c)}{2\delta_*^2}$, so we get

$$\begin{aligned} \mathbb{E} [\|x_{k+1} - x_*\|^2 | x_k] &\leq \|x_k - x_*\|^2 - \min\left(\frac{\gamma\mu}{2}, 1 - c\right) (\|\bar{x}_{k+1} - x_*\|^2 - \|x_k - \bar{x}_{k+1}\|^2) \\ &\leq \|x_k - x_*\|^2 - \frac{1}{2} \min\left(\frac{\gamma\mu}{2}, 1 - c\right) \|x_k - x_*\|^2 \\ &= \left(1 - \frac{1}{2} \min\left(\frac{\gamma\mu}{2}, 1 - c\right)\right) \|x_k - x_*\|^2. \end{aligned}$$

Taking the full expectation, we obtain

$$\mathbb{E} [\|x_{k+1} - x_*\|^2] \leq \left(1 - \frac{1}{2} \min\left(\frac{\gamma\mu}{2}, 1 - c\right)\right) \mathbb{E} [\|x_k - x_*\|^2].$$

Iterating over k , the result follows

$$\mathbb{E} [\|x_{k+1} - x_*\|^2] \leq \left(1 - \frac{1}{2} \min\left(\frac{\gamma\mu}{2}, 1 - c\right)\right)^{k+1} \|x_0 - x_*\|^2.$$

□