OPT2025: 17th Annual Workshop on Optimization for Machine Learning

Revisiting Stochastic Proximal Point Methods: Generalized Smoothness and Similarity

Zhirayr Tovmasyan
Grigory Malinovsky GRIGORII.MALINOVSKII@KAUST.EDU.SA
Laurent Condat
Peter Richtárik
KAUST, 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 underexplored. 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.

1. Introduction

We study the **stochastic optimization** problem

$$\min_{x\in\mathbb{R}^d} \left\{ f(x) := \mathbb{E}_{\xi\sim\mathcal{D}} \left[f_\xi(x) \right] \right\}, \tag{1}$$

where $\xi \sim \mathcal{D}$ is a random variable, f_ξ is the loss on sample ξ, and f is the generalization error. While $\nabla f_\xi(x)$ is accessible, $\nabla f(x)$ is not. Such problems underpin supervised learning [8, 75, 76].

A special case is the finite-sum problem

$$\min_{x\in\mathbb{R}^d} \left\{ f(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x) \right\}, \tag{2}$$

arising in empirical risk minimization [1, 27, 28, 73].

Many ML and DL problems are **nonsmooth**, where gradients are not Lipschitz or even undefined [31, 40, 90]. Nonsmooth convex settings often capture DL complexity [48], motivating adaptive algorithms such as AdaGrad, Adam, and Prodigy [25, 50, 62], which remain highly effective in practice. In contrast, smoothness-based methods with strong theory often fail on DL tasks [17, 21, 32, 67], prompting refinements of smoothness assumptions [31, 83].

An early extension is the (L_0, L_1)**-smoothness** condition, bounding the Hessian by the gradient norm [11, 89, 90]. More general approaches bound the Hessian via polynomial or arbitrary functions

of the gradient [56, 82]. While SGD variants are well studied [26, 39, 85, 91], **stochastic proximal point methods (SPPM)** remain less explored. The prox operator of f is

$$\operatorname{prox}_f(x) := \operatorname*{argmin}_{z\in\mathbb{R}^d} \left\{ f(z) + \tfrac{1}{2}\|z-x\|^2 \right\}, \tag{3}$$

[4]. Proximal methods [12, 66] allow larger stepsizes and are robust to stepsize choice. SPPM enjoys stability and convergence guarantees [2, 3, 46, 68, 70].

Another perspective is the **similarity** assumption [74], capturing gradient homogeneity across data [30, 77]. Various formulations have been analyzed for proximal methods, showing both practical and theoretical relevance [29, 36, 72, 78]. Similarity also plays a key role in distributed and federated learning [43, 45, 53].

In practice, prox computations are often **inexact**, approximated by iterative subroutines. Effectiveness is ensured via criteria such as relative/absolute error thresholds [46, 54], gradient-norm reduction [71], or other guarantees [7, 34, 53]. These maintain convergence while balancing cost and accuracy.

2. Preliminaries

We introduce the main assumptions used in our analysis. Throughout, each stochastic function f_ξ is convex and differentiable (avoiding subdifferential notation, see Sadiev et al. [72]).

Assumption 1 (Differentiability) *The function $f_\xi : \mathbb{R}^d \to \mathbb{R}$ is differentiable for $\mathcal{D}$-almost every ξ.*

This implies $\nabla f(x) = \mathbb{E}_\xi[\nabla f_\xi(x)]$, so f is differentiable.

Assumption 2 (Convexity) *Each f_ξ is convex:*

$$f_\xi(x) \geq f_\xi(y) + \langle \nabla f_\xi(y), x-y\rangle \quad \forall x, y.$$

Assumption 3 (Strong convexity of f) *The function f is μ-strongly convex:*

$$f(x) \geq f(y) + \langle \nabla f(y), x-y\rangle + \tfrac{\mu}{2}\|x-y\|^2.$$

Assumption 4 (Interpolation regime) *There exists $x_\star$ such that $\nabla f_\xi(x_\star) = 0$ for $\mathcal{D}$-almost every ξ [59, 84].*

We next state several smoothness assumptions.

Assumption 5 (L-smoothness) *Each f_ξ is L-smooth:*

$$\|\nabla f_\xi(x) - \nabla f_\xi(y)\| \leq L\|x-y\|.$$

Assumption 6 (Symmetric (L_0, L_1)-smoothness) *Each f_ξ satisfies*

$$\|\nabla f_\xi(x) - \nabla f_\xi(y)\| \leq \left(L_0 + L_1 \sup_{u\in[x,y]} \|\nabla f_\xi(u)\|\right)\|x-y\|.$$

Algorithm 1 Stochastic Proximal Point Method (SPPM)

Parameters: stepsize $\gamma > 0$, starting point x_0
for $t = 0, 1, \ldots, T-1$ **do**
  Sample $\xi_k \sim \mathcal{D}$
  $x_{k+1} := \arg\min_z \{f_{\xi_k}(z) + \frac{1}{2\gamma}\|z - x_k\|^2\}$
end for

Assumption 7 (α-symmetric generalized smoothness) *Each f_ξ satisfies*

$$\|\nabla f_\xi(x) - \nabla f_\xi(y)\| \leq \big(L_0 + L_1 \sup_{u\in[x,y]} \|\nabla f_\xi(u)\|^\alpha\big)\|x - y\|,$$

with $\alpha \in [0, 1]$ [11].

When $\alpha = 0$ this reduces to L-smoothness, and $\alpha = 1$ recovers (L_0, L_1)-smoothness.

Assumption 8 (Star Similarity) *There exist $x_\star$ and $\delta_\star > 0$ such that*

$$\mathbb{E}_\xi\big[\|\nabla f_\xi(x) - \nabla f(x) - \nabla f_\xi(x_\star)\|^2\big] \leq \delta_\star^2\|x - x_\star\|^2.$$

Assumption 9 (Inexact Proximal Condition) *At iteration k, the subroutine $\mathcal{M}$ after T steps produces $\hat{x}_k$ with*

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

where $x_k^\Psi := \arg\min_x \{f_{\xi_k}(x) + \frac{1}{2\gamma}\|x - x_k\|^2\}$.

2.1. The Stochastic Proximal Point Method

For differentiable f, the prox operator satisfies

$$y = \mathrm{prox}_{\gamma f}(x) \;\Leftrightarrow\; y + \gamma\nabla f(y) = x.$$

Thus SPPM performs stochastic prox updates, equivalent for differentiable f_ξ to

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

Unlike SGD, the gradient is evaluated at the new iterate x_{k+1}, making SPPM an implicit method.

3. ϕ-Smoothness: A New Generalized Assumption

Analyses of gradient-type methods often assume L-smoothness, but in practice rely on a weaker *path-wise smoothness* condition that holds only along iterates:

$$\|\nabla f_\xi(x_{k+1}) - \nabla f_\xi(x_k)\| \leq L_{\text{path}}\|x_{k+1} - x_k\|. \tag{4}$$

This is harder to verify directly, as it depends on the algorithm's trajectory. To overcome this, we introduce a more general condition, ϕ-smoothness, which recovers path-wise smoothness as a consequence.

Algorithm 2 SPPM-inexact

Input: stepsize $\gamma > 0$, start x_0
for $t = 0, 1, \ldots, T-1$ **do**
  Sample $\xi_k \sim \mathcal{D}$
  $\hat{x}_k \approx \arg\min_z \{f_{\xi_k}(z) + \frac{1}{2\gamma}\|z - x_k\|^2\}$
  $x_{k+1} = x_k - \gamma \nabla f_{\xi_k}(\hat{x}_k)$
end for

Assumption 10 (ϕ-smoothness) *For $\mathcal{D}$-almost every ξ, $f_\xi : \mathbb{R}^d \to \mathbb{R}$ satisfies*

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

where ϕ is nonnegative and nondecreasing in both arguments.

This extends l-smoothness [56, 82], which applies to twice differentiable functions.

Lemma 11 *Under Assumption 10, for $\mathcal{D}$-a.e. ξ,*

$$f_\xi(x) \leq f_\xi(y) + \langle \nabla f_\xi(y), x - y\rangle + \frac{\phi(\|x-y\|, \|\nabla f_\xi(y)\|)}{2}\|x - y\|^2.$$

Lemma 12 *If f_ξ is (L_0, L_1)-smooth, then it is ϕ-smooth with*

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

Lemma 13 *If f_ξ is α-symmetric smooth, then it is ϕ-smooth with*

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

where K_0, K_1, K_2 depend on (L_0, L_1, α) as in the original definition.

4. Convergence Results under ϕ-Smoothness

Our analysis relies on bounding $\|x_{k+1} - x_k\|$, using convexity, interpolation, and prox nonexpansiveness.

Lemma 14 *Under Assumptions 2, 4, $\|x_{k+1} - x_k\|^2 \leq \|x_0 - x_*\|^2$ for all $k \geq 0$.*

Lemma 15 *Under Assumptions 1, 2, 10, SPPM iterates satisfy*

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

Theorem 16 *Under Assumptions 1, 2, 4, 10, for any $\gamma > 0$,*

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

If additionally f is μ-strongly convex, then

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

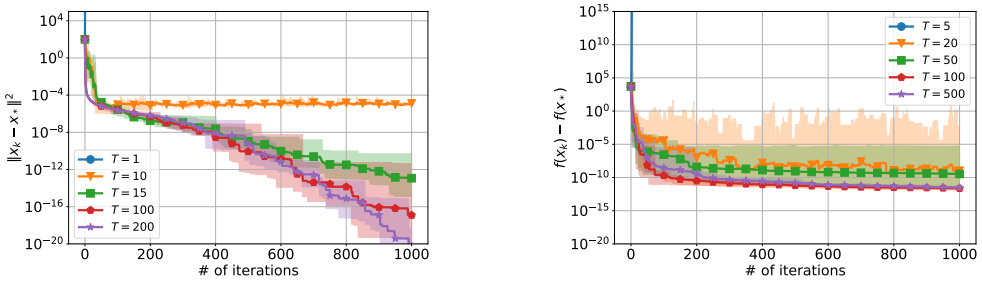

Figure 1: Convergence behavior of SPPM-inexact with different inner iterations in strongly convex and convex settings.

Finally, for inexact prox computations:

Theorem 17 *Under the above assumptions and Assumption 9, if* $\frac{\eta\gamma^2}{T^\alpha} \leq c < 1$*, then SPPM-inexact satisfies*

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

If f *is* μ*-strongly convex, then*

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

5. Experiments

In this section, we present numerical experiments conducted for the optimization problem in the finite-sum form (6), 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 $i = 1, \ldots, n$. Each function f_i is (L_0, L_1)-smooth, with $L_0 = 2s$ and $L_1 = 2s - 1$, and is convex.

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.

Additional experiments are provided in Appendix E.

6. Conclusion

We analyze SPPM methods under assumptions beyond Lipschitz smoothness. A generalized ϕ-assumption unifies many cases, yielding convergence guarantees for strongly convex and general interpolation regimes. We also study convergence under expected similarity in strongly convex settings. Our results advance understanding of stochastic methods beyond smoothness, and extending SPPM to general convex regimes with expected similarity remains future work.

Acknowledgements

The research reported in this publication 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, under award number 5940, iii) SDAIA-KAUST Center of Excellence in Artificial Intelligence and Data Science.

References

[1] Alekh Agarwal and Leon Bottou. A lower bound for the optimization of finite sums. In *International conference on machine learning*, pages 78–86. PMLR, 2015.

[2] Hilal Asi and John C Duchi. Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity. *SIAM Journal on Optimization*, 29(3):2257–2290, 2019.

[3] Hilal Asi, Karan Chadha, Gary Cheng, and John C Duchi. Minibatch stochastic approximate proximal point methods. *Advances in neural information processing systems*, 33:21958–21968, 2020.

[4] Heinz H Bauschke, Patrick L Combettes, Heinz H Bauschke, and Patrick L Combettes. *Correction to: convex analysis and monotone operator theory in Hilbert spaces*. Springer, 2017.

[5] Dimitri P Bertsekas. Incremental proximal methods for large scale convex optimization. *Mathematical programming*, 129(2):163–195, 2011.

[6] Pascal Bianchi. Ergodic convergence of a stochastic proximal point algorithm. *SIAM Journal on Optimization*, 26(4):2235–2260, 2016.

[7] Ekaterina Borodich, Georgiy Kormakov, Dmitry Kovalev, Aleksandr Beznosikov, and Alexander Gasnikov. Optimal algorithm with complexity separation for strongly convex-strongly concave composite saddle point problems. *arXiv preprint arXiv:2307.12946*, 2023.

[8] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311, 2018.

[9] Xufeng Cai, Chaobing Song, Stephen Wright, and Jelena Diakonikolas. Cyclic block coordinate descent with variance reduction for composite nonconvex optimization. In *International Conference on Machine Learning*, pages 3469–3494. PMLR, 2023.

[10] Luigi Carratino, Alessandro Rudi, and Lorenzo Rosasco. Learning with sgd and random features. *Advances in neural information processing systems*, 31, 2018.

[11] Ziyi Chen, Yi Zhou, Yingbin Liang, and Zhaosong Lu. Generalized-smooth nonconvex optimization is as efficient as smooth nonconvex optimization. In *International Conference on Machine Learning*, pages 5396–5427. PMLR, 2023.

[12] L. Condat, D. Kitahara, A. Contreras, and A. Hirabayashi. Proximal splitting algorithms for convex optimization: A tour of recent advances, with new twists. *SIAM Review*, 65(2):375–435, 2023.

[13] Laurent Condat and Peter Richtárik. Randprox: Primal-dual optimization algorithms with randomized proximal updates. *arXiv preprint arXiv:2207.12891*, 2022.

[14] Laurent Condat and Peter Richtárik. A simple linear convergence analysis of the point-saga algorithm. *arXiv preprint arXiv:2405.19951*, 2024.

[15] Laurent Condat, Grigory Malinovsky, and Peter Richtárik. Distributed proximal splitting algorithms with rates and acceleration. *Frontiers in Signal Processing*, 1:776825, 2022.

[16] Laurent Condat, Ivan Agarskỳ, Grigory Malinovsky, and Peter Richtárik. Tamuna: Doubly accelerated federated learning with local training, compression, and partial participation. In *International Workshop on Federated Learning in the Age of Foundation Models in Conjunction with NeurIPS 2023*, 2023.

[17] Michael Crawshaw, Mingrui Liu, Francesco Orabona, Wei Zhang, and Zhenxun Zhuang. Robustness to unbounded smoothness of generalized signsgd. *Advances in neural information processing systems*, 35:9955–9968, 2022.

[18] Michael Crawshaw, Yajie Bao, and Mingrui Liu. Federated learning with client subsampling, data heterogeneity, and unbounded smoothness: A new algorithm and lower bounds. *Advances in Neural Information Processing Systems*, 36, 2024.

[19] Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

[20] Damek Davis and Wotao Yin. A three-operator splitting scheme and its optimization applications. *Set-valued and variational analysis*, 25:829–858, 2017.

[21] Aaron Defazio and Léon Bottou. On the ineffectiveness of variance reduced optimization for deep learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[22] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 27, 2014.

[23] Yury Demidovich, Grigory Malinovsky, Igor Sokolov, and Peter Richtárik. A guide through the zoo of biased sgd. *Advances in Neural Information Processing Systems*, 36:23158–23171, 2023.

[24] Yury Demidovich, Petr Ostroukhov, Grigory Malinovsky, Samuel Horváth, Martin Takáč, Peter Richtárik, and Eduard Gorbunov. Methods with local steps and random reshuffling for generally smooth non-convex federated optimization. *arXiv preprint arXiv:2412.02781*, 2024.

[25] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

[26] Matthew Faw, Litu Rout, Constantine Caramanis, and Sanjay Shakkottai. Beyond uniform smoothness: A stopped analysis of adaptive sgd. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 89–160. PMLR, 2023.

[27] Vitaly Feldman. Generalization of erm in stochastic convex optimization: The dimension strikes back. *Advances in Neural Information Processing Systems*, 29, 2016.

[28] Claudio Gambella, Bissan Ghaddar, and Joe Naoum-Sawaya. Optimization problems for machine learning: A survey. *European Journal of Operational Research*, 290(3):807–828, 2021.

[29] Elnur Gasanov and Peter Richtarik. Speeding up stochastic proximal optimization in the high hessian dissimilarity setting. *arXiv preprint arXiv:2412.13619*, 2024.

[30] Ian Goodfellow. Deep learning, 2016.

[31] Eduard Gorbunov, Nazarii Tupitsa, Sayantan Choudhury, Alen Aliev, Peter Richtárik, Samuel Horváth, and Martin Takáč. Methods for convex (l_0, l_1)-smooth optimization: Clipping, acceleration, and adaptivity. *arXiv preprint arXiv:2409.14989*, 2024.

[32] Robert M Gower, Mark Schmidt, Francis Bach, and Peter Richtárik. Variance-reduced methods for machine learning. *Proceedings of the IEEE*, 108(11):1968–1983, 2020.

[33] Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. Sgd: General analysis and improved rates. In *International conference on machine learning*, pages 5200–5209. PMLR, 2019.

[34] Michał Grudzień, Grigory Malinovsky, and Peter Richtárik. Can 5th generation local training methods support client sampling? yes! In *International Conference on Artificial Intelligence and Statistics*, pages 1055–1092. PMLR, 2023.

[35] Michał Grudzień, Grigory Malinovsky, and Peter Richtárik. Improving accelerated federated learning with compression and importance sampling. *arXiv preprint arXiv:2306.03240*, 2023.

[36] Hadrien Hendrikx, Lin Xiao, Sebastien Bubeck, Francis Bach, and Laurent Massoulie. Statistically preconditioned accelerated gradient method for distributed optimization. In *International conference on machine learning*, pages 4203–4227. PMLR, 2020.

[37] Zhengmian Hu and Heng Huang. Tighter analysis for proxskip. In *International Conference on Machine Learning*, pages 13469–13496. PMLR, 2023.

[38] Florian Hübler, Ilyas Fatkhullin, and Niao He. From gradient clipping to normalization for heavy tailed sgd. *arXiv preprint arXiv:2410.13849*, 2024.

[39] Florian Hübler, Junchi Yang, Xiang Li, and Niao He. Parameter-agnostic optimization under relaxed smoothness. In *International Conference on Artificial Intelligence and Statistics*, pages 4861–4869. PMLR, 2024.

[40] Franck Iutzeler and Jérôme Malick. Nonsmoothness in machine learning: specific structure, proximal identification, and applications. *Set-Valued and Variational Analysis*, 28(4):661–678, 2020.

[41] Divyansh Jhunjhunwala, Shiqiang Wang, and Gauri Joshi. Fedexp: Speeding up federated averaging via extrapolation. *arXiv preprint arXiv:2301.09604*, 2023.

[42] Xiaowen Jiang, Anton Rodomanov, and Sebastian U Stich. Federated optimization with doubly regularized drift correction. *arXiv preprint arXiv:2404.08447*, 2024.

[43] Xiaowen Jiang, Anton Rodomanov, and Sebastian U Stich. Stabilized proximal-point methods for federated optimization. *arXiv preprint arXiv:2407.07084*, 2024.

[44] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26, 2013.

[45] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pages 5132–5143. PMLR, 2020.

[46] Ahmed Khaled and Chi Jin. Faster federated optimization under second-order similarity. *arXiv preprint arXiv:2209.02257*, 2022.

[47] Ahmed Khaled and Peter Richtárik. Better theory for sgd in the nonconvex world. *arXiv preprint arXiv:2002.03329*, 2020.

[48] Ahmed Khaled, Konstantin Mishchenko, and Chi Jin. Dowg unleashed: An efficient universal parameter-free gradient descent method. *Advances in Neural Information Processing Systems*, 36:6748–6769, 2023.

[49] Sarit Khirirat, Abdurakhmon Sadiev, Artem Riabinin, Eduard Gorbunov, and Peter Richtárik. Error feedback under (l_0, l_1)-smoothness: Normalization and momentum. *arXiv preprint arXiv:2410.16871*, 2024.

[50] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[51] Anastasia Koloskova, Hadrien Hendrikx, and Sebastian U Stich. Revisiting gradient clipping: Stochastic bias and tight convergence guarantees. In *International Conference on Machine Learning*, pages 17343–17363. PMLR, 2023.

[52] Jakub Konečnỳ. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

[53] Dmitry Kovalev, Aleksandr Beznosikov, Ekaterina Borodich, Alexander Gasnikov, and Gesualdo Scutari. Optimal gradient sliding and its application to optimal distributed optimization under similarity. *Advances in Neural Information Processing Systems*, 35:33494–33507, 2022.

[54] Hanmin Li and Peter Richtárik. On the convergence of fedprox with extrapolation and inexact prox. *arXiv preprint arXiv:2410.01410*, 2024.

[55] Hanmin Li, Kirill Acharya, and Peter Richtarik. The power of extrapolation in federated learning. *arXiv preprint arXiv:2405.13766*, 2024.

[56] Haochuan Li, Jian Qian, Yi Tian, Alexander Rakhlin, and Ali Jadbabaie. Convex and non-convex optimization under generalized smoothness. *Advances in Neural Information Processing Systems*, 36, 2024.

[57] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

[58] Aleksandr Lobanov, Alexander Gasnikov, Eduard Gorbunov, and Martin 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.

[59] Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning. In *International Conference on Machine Learning*, pages 3325–3334. PMLR, 2018.

[60] Grigory Malinovsky, Kai Yi, and Peter Richtárik. Variance reduced proxskip: Algorithm, theory and application to federated learning. *Advances in Neural Information Processing Systems*, 35: 15176–15189, 2022.

[61] Grigory Malinovsky, Alibek Sailanbayev, and Peter Richtárik. Random reshuffling with variance reduction: New analysis and better rates. In *Uncertainty in Artificial Intelligence*, pages 1347–1357. PMLR, 2023.

[62] Konstantin Mishchenko and Aaron Defazio. Prodigy: An expeditiously adaptive parameter-free learner. *arXiv preprint arXiv:2306.06101*, 2023.

[63] Konstantin Mishchenko, Grigory Malinovsky, Sebastian Stich, and Peter Richtárik. Proxskip: Yes! local gradient steps provably lead to communication acceleration! finally! In *International Conference on Machine Learning*, pages 15750–15769. PMLR, 2022.

[64] Lam Nguyen, Phuong Ha Nguyen, Marten Dijk, Peter Richtárik, Katya Scheinberg, and Martin Takác. Sgd and hogwild! convergence without the bounded gradients assumption. In *International Conference on Machine Learning*, pages 3750–3758. PMLR, 2018.

[65] Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *International conference on machine learning*, pages 2613–2621. PMLR, 2017.

[66] N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 3(1): 127–239, 2014.

[67] Sashank J Reddi, Ahmed Hefny, Suvrit Sra, Barnabas Poczos, and Alex Smola. Stochastic variance reduction for nonconvex optimization. In *International conference on machine learning*, pages 314–323. PMLR, 2016.

[68] Peter Richtárik, Abdurakhmon Sadiev, and Yury Demidovich. A unified theory of stochastic proximal point methods without smoothness. *arXiv preprint arXiv:2405.15941*, 2024.

[69] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[70] Ernest K Ryu and Stephen Boyd. Stochastic proximal iteration: a non-asymptotic improvement upon stochastic gradient descent. *Author website, early draft*, 2014.

[71] Abdurakhmon Sadiev, Dmitry Kovalev, and Peter Richtárik. Communication acceleration of local gradient methods via an accelerated primal-dual algorithm with an inexact prox. *Advances in Neural Information Processing Systems*, 35:21777–21791, 2022.

[72] Abdurakhmon Sadiev, Laurent Condat, and Peter Richtárik. Stochastic proximal point methods for monotone inclusions under expected similarity. *arXiv preprint arXiv:2405.14255*, 2024.

[73] Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162:83–112, 2017.

[74] Ohad Shamir, Nati Srebro, and Tong Zhang. Communication-efficient distributed optimization using an approximate newton-type method. In *International conference on machine learning*, pages 1000–1008. PMLR, 2014.

[75] Ruo-Yu Sun. Optimization for deep learning: An overview. *Journal of the Operations Research Society of China*, 8(2):249–294, 2020.

[76] Shiliang Sun, Zehui Cao, Han Zhu, and Jing Zhao. A survey of optimization methods from a machine learning perspective. *IEEE transactions on cybernetics*, 50(8):3668–3681, 2019.

[77] Ying Sun, Gesualdo Scutari, and Amir Daneshmand. Distributed optimization based on gradient tracking revisited: Enhancing convergence rate via surrogation. *SIAM Journal on Optimization*, 32(2):354–385, 2022.

[78] Rafał Szlendak, Alexander Tyurin, and Peter Richtárik. Permutation compressors for provably faster distributed nonconvex optimization. *arXiv preprint arXiv:2110.03300*, 2021.

[79] Canh T Dinh, Nguyen Tran, and Josh Nguyen. Personalized federated learning with moreau envelopes. *Advances in neural information processing systems*, 33:21394–21405, 2020.

[80] Yuki Takezawa, Han Bao, Ryoma Sato, Kenta Niwa, and Makoto Yamada. Polyak meets parameter-free clipped gradient descent. *arXiv preprint arXiv:2405.15010*, 2024.

[81] Panos Toulis, Dustin Tran, and Edo Airoldi. Towards stability and optimality in stochastic gradient descent. In *Artificial Intelligence and Statistics*, pages 1290–1298. PMLR, 2016.

[82] Alexander Tyurin. Toward a unified theory of gradient descent under generalized smoothness. *arXiv preprint arXiv:2412.11773*, 2024.

[83] Daniil Vankov, Anton Rodomanov, Angelia Nedich, Lalitha Sankar, and Sebastian U Stich. Optimizing (l_0, l_1)-smooth functions by gradient methods. *arXiv preprint arXiv:2410.10800*, 2024.

[84] Aditya Vardhan Varre, Loucas Pillaud-Vivien, and Nicolas Flammarion. Last iterate convergence of sgd for least-squares in the interpolation regime. *Advances in Neural Information Processing Systems*, 34:21581–21591, 2021.

[85] Bohan Wang, Huishuai Zhang, Zhiming Ma, and Wei Chen. Convergence of adagrad for non-convex objectives: Simple proofs and relaxed assumptions. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 161–190. PMLR, 2023.

[86] Bohan Wang, Yushun Zhang, Huishuai Zhang, Qi Meng, Ruoyu Sun, Zhi-Ming Ma, Tie-Yan Liu, Zhi-Quan Luo, and Wei Chen. Provable adaptivity of adam under non-uniform smoothness. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2960–2969, 2024.

[87] Jingfeng Wu, Wenqing Hu, Haoyi Xiong, Jun Huan, Vladimir Braverman, and Zhanxing Zhu. On the noisy gradient descent that generalizes as sgd. In *International Conference on Machine Learning*, pages 10367–10376. PMLR, 2020.

[88] Junchi Yang, Xiang Li, Ilyas Fatkhullin, and Niao He. Two sides of one coin: the limits of untuned sgd and the power of adaptive methods. *Advances in Neural Information Processing Systems*, 36, 2024.

[89] Bohang Zhang, Jikai Jin, Cong Fang, and Liwei Wang. Improved analysis of clipping algorithms for non-convex optimization. *Advances in Neural Information Processing Systems*, 33:15511–15521, 2020.

[90] Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. *arXiv preprint arXiv:1905.11881*, 2019.

[91] Shen-Yi Zhao, Yin-Peng Xie, and Wu-Jun Li. On the convergence and improvement of stochastic normalized gradient descent. *Science China Information Sciences*, 64:1–13, 2021.

[92] Pan Zhou, Jiashi Feng, Chao Ma, Caiming Xiong, Steven Chu Hong Hoi, et al. Towards theoretically understanding why sgd generalizes better than adam in deep learning. *Advances in Neural Information Processing Systems*, 33:21285–21296, 2020.

Appendix A. Extended Introduction

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

$$\min_{x\in\mathbb{R}^d} \left\{ f(x) := \mathbb{E}_{\xi\sim\mathcal{D}}\left[f_\xi(x)\right] \right\}, \tag{5}$$

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, 75, 76].

A particular case of interest is the finite-sum optimization problem, where f is the average of a large number of functions [1, 73]:

$$\min_{x\in\mathbb{R}^d} \left\{ f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x) \right\}. \tag{6}$$

This problem frequently arises when training supervised ML models using empirical risk minimization [27, 28]. It is an instance of (5) 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 [31, 40]. For instance, even in relatively simple neural networks, the gradient of the standard ℓ_2-regression loss fails to satisfy Lipschitz continuity [90]. Moreover, the convex but nonsmooth setting often provides an effective framework for capturing the complexities of DL problems [48]. Many widely used and empirically successful adaptive optimization algorithms, such as AdaGrad [25], Adam [50], and Prodigy [62] 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 [32, 67] 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 [31, 83].

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 [90]. 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, 89].

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, 39, 85, 91]. 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 [56]. 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 [82].

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, 46]. We recall that the proximity operator (**prox**) of a function $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is

$$\operatorname{prox}_f(x) \stackrel{\text{def}}{=} \operatorname*{argmin}_{z \in \mathbb{R}^d} \left\{ f(z) + \frac{1}{2}\|z - x\|^2 \right\} \tag{7}$$

[4]. Proximal methods, which leverage proxs of functions [12, 66], 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. Ryu and Boyd [70] provide convergence rate guarantees for SPPM and emphasize its stability with respect to inaccuracies in learning rate selection; a property not typically observed in SGD. Asi and Duchi [2] investigate 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 [68].

Alternatively, instead of relying on the smoothness assumption and its generalizations, we can consider the **similarity** assumption [74]. 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 [30, 77]. Recognizing this natural property, several recent works have explored various formulations and generalizations of the similarity assumption [36, 72, 78]. In particular, multiple studies have analyzed stochastic proximal methods under different similarity conditions, demonstrating their practical relevance and theoretical significance [29, 72]. Moreover, the concept of similarity offers a more refined perspective on the behavior of optimization algorithms in distributed and federated learning settings [43, 45, 53].

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, 34, 53]. These criteria typically include conditions such as relative and absolute error thresholds [46, 54], as well as guarantees on the reduction of the gradient norm [71]. 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 the most general conditions required for the convergence of SPPM and introduce the novel notion of ϕ-smoothness (Section 3). 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 4). 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 (Appendix C), 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 (Appendix D).

Appendix B. Related work

Stochastic Gradient Descent (SGD) [69] 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 [75, 87, 92]. 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, 47]. This ongoing theoretical research continues to refine SGD and its variants, ensuring their effectiveness in large-scale and complex learning tasks [10, 64, 88]. Notably, methods designed to leverage the smoothness of the objective often struggle in Deep Learning, where optimization problems are inherently non-smooth. For instance, while variance-reduced methods [9, 22, 44, 61, 65, 73] 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 [90]. Several studies have analyzed SGD methods under this condition in the convex setting [31, 51, 56, 58, 80, 83]. The analysis of SGD in the non-convex case was first discussed in [90] and later extended to momentum-based methods in [89].

Similar results have been established for various optimization methods, including Normalized GD [11, 38, 91], SignGD [17], AdaGrad-Norm/AdaGrad [26, 85], Adam [86], and Normalized GD with Momentum [39]. Additionally, methods specifically designed for distributed optimization have been analyzed under generalized smoothness conditions [18, 24, 49].

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 non-smooth problems, where traditional smoothness-based methods may struggle. [6, 19, 81] These properties make stochastic proximal point methods a valuable tool in both theoretical analysis and practical applications [14, 15, 20].

Stochastic proximal methods have become increasingly important in Federated Learning due to their ability to handle decentralized optimization problems efficiently [52]. 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 [35, 41, 55, 57, 79]. 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 [13, 16, 37, 42, 60, 63].

Appendix C. Convergence Results under Expected Similarity

In this section, we discuss the convergence results under expected similarity (Assumption 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 18 (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:*

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

Remark 19 *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 J.*

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

Theorem 20 *Let Assumptions 1 (Differentiability), 2 (Convexity), 3 (Strong convexity of f), 8 (Star Similarity), and 18 hold. If the stepsize satisfies $\gamma \leq \frac{\mu}{4\delta_*^2}$, then SPPM satisfies, for every $k \geq 0$,*

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

The result shows linear convergence to a neighborhood of the solution, with the neighborhood size proportional to the variance σ_*^2 from Assumption 18, 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 21 *Let Assumptions 1 (Differentiability), 2 (Convexity), 8 (Star Similarity), and 18 (Bounded Variance) hold. Consider SPPM-inexact with Assumption 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 $k \geq 0$,*

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

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.

Appendix D. Experiments

In this section, we present numerical experiments conducted for the optimization problem in the finite-sum form (6), 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 $i = 1, \ldots, n$. Each function f_i is (L_0, L_1)-smooth, with $L_0 = 2s$ and $L_1 = 2s - 1$, and is convex.

Additional experiments exploring the impact of stepsizes and dependence on the initial point are provided in Appendix E.

Here, we present two main experiments. The first experiment (originally Experiment 3) 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.

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 (6) 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 $i = 1, \ldots, n$, $\lambda \in \mathbb{R}^+$, 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 $f_i(x) = a_i|x|^{\frac{2-\alpha_i}{1-\alpha_i}}$, each of which is individually α_i-smooth, as shown by Chen et al. [11]. It is straightforward to verify that each f_i is also $1 > \beta > \alpha_i > 0$-symmetric smooth, implying all functions are $\alpha = \max_{i=1\ldots N} \alpha_i$-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.

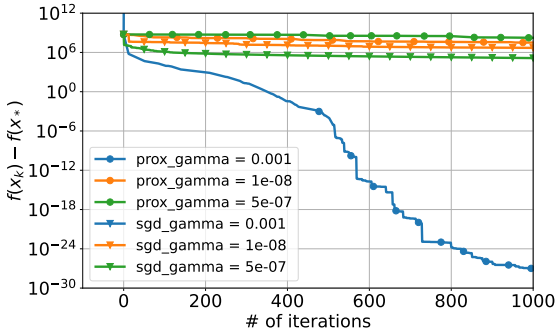

Figure 2: Comparison between SPPM-inexact and SGD with constant stepsizes γ.

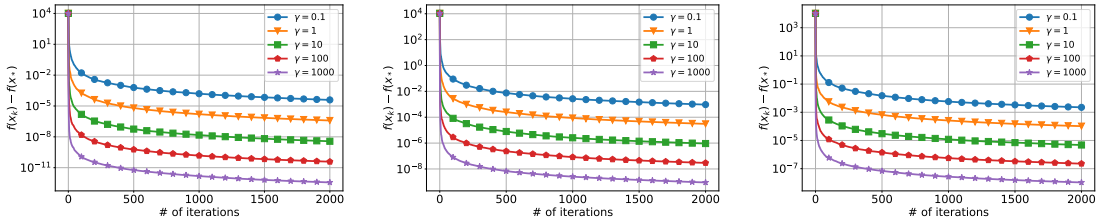

Figure 3: Convergence behavior of SPPM-inexact with different stepsizes.

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.

Appendix E. Additional Experiments

In this appendix, we present additional experiments originally conducted to analyze the behavior of SPPM-inexact under various conditions.

E.1. Impact of Different Stepsizes

This experiment investigates 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

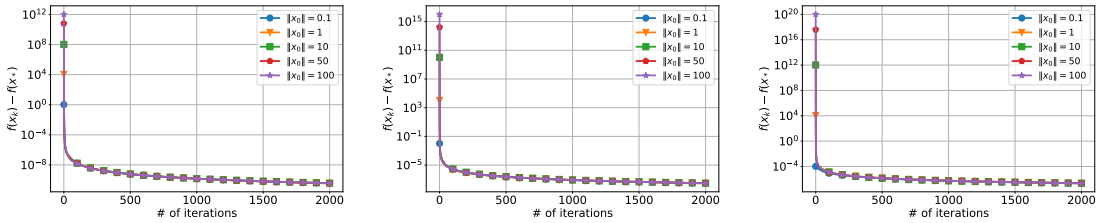

Figure 4: Convergence behavior of SPPM-inexact with different starting points.

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.

E.2. Dependence on the Initial Point

This experiment investigates 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 14 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.

Appendix F. Notations

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 (8)
$\mathcal{F}_k$	σ-algebra generated by $x_0,\ldots,x_k$
$\hat{x}_k$	uniformly chosen iterate from the set $\{x_0,x_1,\ldots,x_{k-1}\}$
x_k^Ψ	exact solution to the proximal subproblem, see (**??**)
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

Appendix G. Fundamental Lemmas

We define the Bregman divergence of a function g as

$$D_g(x,y) := g(x) - g(y) - \langle \nabla g(y), x - y\rangle \quad \forall x, y \in \mathbb{R}^d. \tag{8}$$

Lemma 22 *Let Assumptions 1, 2 and 4 hold. Then for any stepsize $\gamma > 0$, the iterates of SPPM satisfy, for every $k \geq 0$,*

$$\langle x_{k+1} - x_*, x_k - x_{k+1}\rangle \geq \gamma(f_{\xi_k}(x_{k+1}) - f_{\xi_k}(x_*)) \geq 0. \tag{9}$$

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:

$$\nabla f_{\xi_k}(x_{k+1}) + \frac{1}{\gamma}(x_{k+1} - x_k) = 0. \tag{10}$$

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

$$\begin{aligned} f_{\xi_k}(x_*) &\geq f_{\xi_k}(x_{k+1}) + \langle \nabla f_{\xi_k}(x_{k+1}), x_* - x_{k+1}\rangle \\ &\overset{(10)}{=} f_{\xi_k}(x_{k+1}) + \frac{1}{\gamma}\langle x_k - x_{k+1}, x_* - x_{k+1}\rangle. \end{aligned}$$

Rearranging this expression yields the first inequality in (9). 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 4. ■

G.1. Proof of Lemma 11

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

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:

$$\begin{aligned} D_{f_\xi}(x,y) &\leq \int_0^1 (\phi(t\|x-y\|, \|\nabla f_\xi(y)\|) t\|x-y\|^2 \, dt \\ &\leq \int_0^1 (\phi(\|x-y\|, \|\nabla f_\xi(y)\|) t\|x-y\|^2 \, dt \\ &= \frac{\phi(\|x-y\|, \|\nabla f_\xi(y)\|)}{2}\|x-y\|^2. \end{aligned} \tag{11}$$

G.2. Proof of Lemma 12

From Proposition 3.2 (Point 2) in Chen et al. [11], the function $f_\xi \colon \mathbb{R}^d \to \mathbb{R}$ is symmetrically (L_0, L_1)-smooth if and only if for all $x, y \in \mathbb{R}^d$,

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

Hence, defining

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

we obtain

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

which confirms that f_ξ is ϕ-smooth according to Assumption 10.

G.3. Proof of Lemma 13

From Proposition 3.2 (Point 1) in Chen et al. [11], the function $f_\xi \colon \mathbb{R}^d \to \mathbb{R}$ 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:

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

Hence, defining

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

we obtain

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

which confirms that f_ξ is ϕ-smooth in the sense of Assumption 10.

G.4. Proof of Lemma 14

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:

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

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.

G.5. Proof of Lemma 15

Let $k \geq 0$. Applying (11) to the iterates, we obtain:

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

Using (**??**) 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 (14) 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 H. Proof of Theorems 16 and 17

H.1. Proof of Theorem 16

For convenience, we restate Theorem 16 here:

Theorem 23 *Let Assumptions 1 (Differentiability), 2 (Convexity), 4 (Interpolation) and 10 (ϕ-smoothness) hold. Then for any stepsize $\gamma > 0$ we have, for every $k \geq 0$,*

$$\mathbb{E}[f(\hat{x}_k)] - f(x_*) \leq \frac{\phi(\|x_0 - x_*\|, \frac{1}{\gamma}\|x_0 - x_*\|) + \frac{2}{\gamma}}{2k}\|x_0 - x_*\|^2, \tag{12}$$

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

If additionally Assumption 3 holds, then for any stepsize $\gamma > 0$ and $k \in \mathbb{N}^+$, we have:

$$\mathbb{E}\left[\|x_k - x_*\|^2\right] \leq \left(1 - \frac{\mu}{\frac{2}{\gamma} + \phi(\|x_0 - x_*\|, \frac{1}{\gamma}\|x_0 - x_*\|)}\right)^k \|x_0 - x_*\|^2, \tag{13}$$

Proof We have:

$$\begin{aligned}
\|x_{k+1} - x_*\|^2 &= \|x_k - x_* - (x_k - x_{k+1})\|^2 \\
&= \|x_k - x_*\|^2 - 2\langle x_k - x_*, x_k - x_{k+1}\rangle + \|x_k - x_{k+1}\|^2 \\
&= \|x_k - x_*\|^2 - 2\langle x_k - x_{k+1} + x_{k+1} - x_*, x_k - x_{k+1}\rangle + \|x_k - x_{k+1}\|^2 \\
&= \|x_k - x_*\|^2 - 2\langle x_{k+1} - x_*, x_k - x_{k+1}\rangle - \|x_k - x_{k+1}\|^2.
\end{aligned} \tag{14}$$

Using (9) and Lemma 15, we obtain:

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

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

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

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

$$\mathbb{E}[\|x_{k+1} - x_*\|^2 \mid \mathcal{F}_k] \leq \|x_k - x_*\|^2 - \frac{1}{\frac{1}{\gamma} + \frac{\phi\left(\|x_0 - x_*\|, \frac{1}{\gamma}\|x_0 - x_*\|\right)}{2}}(f(x_k) - f(x_*)). \tag{15}$$

Taking full expectation, we get:

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

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

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

Notice that:

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

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 3, then in step (15), applying the strong convexity of f, we get:

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

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:

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

■

H.2. Proof of Theorem 17

For convenience, we restate Theorem 17 here:

Theorem 24 *Let Assumptions 1 (Differentiability), 2 (Convexity), 4 (Interpolation) and 10 (ϕ-smoothness) hold. Consider SPPM-inexact with every inexact prox satisfying Assumption 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 3 holds, then for any $\gamma > 0$, the iterates of SPPM-inexact satisfy, for every $k \geq 0$,

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

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:

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

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:

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

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:

$$\|x_{k+1} - x_*\|^2 = \|x_k - x_*\|^2 - 2\gamma\langle\nabla f_{\xi_k}(\hat{x}_k), \hat{x}_k - x_*\rangle - \|x_k - \hat{x}_k\|^2 + \gamma^2\|\nabla\Psi_k(\hat{x}_k)\|^2. \quad (16)$$

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:

$$\langle\nabla f_{\xi_k}(\hat{x}_k), \hat{x}_k - x_*\rangle \geq f_{\xi_k}(\hat{x}_k) - f_{\xi_k}(x_*). \quad (17)$$

Using (17) in (16), we get:

$$\|x_{k+1} - x_*\|^2 \leq \|x_k - x_*\|^2 - 2\gamma(f_{\xi_k}(\hat{x}_k) - f_{\xi_k}(x_*)) - \|x_k - \hat{x}_k\|^2 + \gamma^2\|\nabla\Psi_k(\hat{x}_k)\|^2. \quad (18)$$

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 (18), we obtain:

$$\begin{aligned}\|x_{k+1} - x_*\|^2 &\leq \|x_k - x_*\|^2 - 2\gamma(f_{\xi_k}(x_k^\Psi) - f_{\xi_k}(x_*)) - \|x_k^\Psi - x_k\|^2 + \|\hat{x}_k - x_k\|^2 - \|x_k - \hat{x}_k\|^2 + \gamma^2\|\nabla\Psi_k(\hat{x}_k)\|^2 \\ &= \|x_k - x_*\|^2 - 2\gamma(f_{\xi_k}(x_k^\Psi) - f_{\xi_k}(x_*)) - \|x_k^\Psi - x_k\|^2 + \gamma^2\|\nabla\Psi_k(\hat{x}_k)\|^2 \qquad (19)\end{aligned}$$

Using (**??**), we get:

$$\begin{aligned}\|x_{k+1} - x_*\|^2 &\leq \|x_k - x_*\|^2 - 2\gamma(f_{\xi_k}(x_k^\Psi) - f_{\xi_k}(x_*)) - \|x_k^\Psi - x_k\|^2 + \frac{\eta\gamma^2\|x_k - x_k^\Psi\|^2}{T^\alpha} \\ &= \|x_k - x_*\|^2 - 2\gamma(f_{\xi_k}(x_k^\Psi) - f_{\xi_k}(x_*)) - (1 - \frac{\eta\gamma^2}{T^\alpha})\|x_k - x_k^\Psi\|^2 \qquad (20)\end{aligned}$$

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 (20), 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 (**??**) in (20):

$$\|x_{k+1} - x_*\|^2 \leq \|x_k - x_*\|^2 - 2\gamma(f_{\xi_k}(x_k^\Psi) - f_{\xi_k}(x_*)) - \frac{1 - \frac{\eta\gamma^2}{T^\alpha}}{\frac{1}{\gamma} + \frac{\phi\left(\|x_0 - x_*\|, \frac{1}{\gamma}\|x_0 - x_*\|\right)}{2}}\left(f_{\xi_k}(x_k) - f_{\xi_k}(x_k^\Psi)\right) \qquad (21)$$

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

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

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

$$\mathbb{E}[\|x_{k+1}-x_*\|^2 \mid \mathcal{F}_k] \leq \|x_k-x_*\|^2 - \frac{1-\frac{\eta\gamma^2}{T^\alpha}}{\frac{1}{\gamma}+\frac{\phi\left(\|x_0-x_*\|,\frac{1}{\gamma}\|x_0-x_*\|\right)}{2}}\left(f(x_k)-f(x_*)\right) \tag{22}$$

Now, by taking the full expectation, we get:

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

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

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

Notice that:

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

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 3, then in step (22), 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:

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

■

Appendix I. Proof of Theorems 20 and 21

For convenience, we restate Theorem 20 here:

Theorem 25 *Let Assumptions 1 (Differentiability), 2 (Convexity), 3(Strong convexity of f), 8 (Star Similarity) and 18 (Bounded Variance at Optimum) hold. If the stepsize satisfies $\gamma \leq \frac{\mu}{4\delta_*^2}$, we have, for every $k \geq 0$,*

$$\mathbb{E}\left[\|x_k - x_*\|^2\right] \leq \left(1 - \min\left(\frac{\gamma\mu}{4}, \frac{1}{2}\right)\right)^k \|x_0 - x_*\|^2 + \max\left(\frac{4}{\gamma\mu}, 2\right) 2\gamma^2\sigma_*^2. \tag{23}$$

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 get:

$$\begin{aligned}
\mathbb{E}\left[\|x_{k+1} - x_*\|^2|\mathcal{F}_k\right] &\leq \|x_k - x_*\|^2 - 2\gamma\mathbb{E}\left[f_{\xi_k}(\bar{x}_{k+1}) + \langle \nabla f_{\xi_k}(\bar{x}_{k+1}), x_{k+1} - \bar{x}_{k+1}\rangle - f_{\xi_k}(x_*)|\mathcal{F}_k\right] \\
&\quad - \mathbb{E}\left[\|x_k - x_{k+1}\|^2|\mathcal{F}_k\right] \\
&= \|x_k - x_*\|^2 - 2\gamma\left(f(\bar{x}_{k+1}) - f(x_*) + \mathbb{E}\left[\langle \nabla f_{\xi_k}(\bar{x}_{k+1}), x_{k+1} - \bar{x}_{k+1}\rangle|\mathcal{F}_k\right]\right) \\
&\quad - \mathbb{E}\left[\|x_{k+1} - \bar{x}_{k+1}\|^2|x_k\right] - \|x_k - \bar{x}_{k+1}\|^2.
\end{aligned} \tag{24}$$

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

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

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

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

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

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

Using Assumption 8 and Young's inequality, we get:

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

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

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

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

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

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:

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

■

For convenience, we restate Theorem 21 here:

Theorem 26 *Let Assumptions 1 (Differentiability), 2 (Convexity), 8 (Star Similarity) and 18 (Bounded Variance) hold. Consider SPPM-inexact with Assumption 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 $k \geq 0$,*

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

Proof To avoid repetition, we start from (28):

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

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

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

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

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

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

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

Using Assumption 8 and Young's inequality, we get:

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

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

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

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

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

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:

$$\begin{aligned}\mathbb{E}\left[\|x_{k+1} - x_*\|^2 \mid x_k\right] &\leq \|x_k - x_*\|^2 - \frac{\gamma\mu}{2}\|\bar{x}_{k+1} - x_*\|^2 - (1-c)\|x_k - \bar{x}_{k+1}\|^2 + \frac{2\gamma^2}{1-c}\mathbb{E}\left[\|\nabla f_{\xi_k}(x_*)\|^2 \mid x_k\right] \\ &\leq \|x_k - x_*\|^2 - \frac{1}{2}\min\left(\frac{\gamma\mu}{2}, 1-c\right)\|x_k - \bar{x}_*\|^2 + \frac{2\gamma^2}{1-c}\mathbb{E}\left[\|\nabla f_{\xi_k}(x_*)\|^2 \mid x_k\right] \\ &\overset{18}{\leq} \left(1 - \frac{1}{2}\min\left(\frac{\gamma\mu}{2}, 1-c\right)\right)\|x_k - x_*\|^2 + \frac{2\gamma^2\sigma_*^2}{1-c}.\end{aligned}$$

Taking the full expectation, we obtain:

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

By applying the inequality recursively, we derive:

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

■

Appendix J. Convergence under Interpolation Regime and Expected Similarity

Theorem 27 *Let Assumptions 1 (Differentiability), 2 (Convexity), 3(Strong convexity of f), 4 (Interpolation), and 8 (Star Similarity) hold. If the stepsize satisfies $\gamma \leq \frac{\mu}{2\delta_*^2}$, then we have, for every $k \geq 0$,*

$$\mathbb{E}\left[\|x_k - x_*\|^2\right] \leq \left(1 - \min\left(\frac{\gamma\mu}{4}, \frac{1}{2}\right)\right)^k \|x_0 - x_*\|^2. \tag{25}$$

Proof *Define*

$$\bar{x}_{k+1} = \mathbb{E}[x_{k+1}|\mathcal{F}_k]. \tag{26}$$

Then

$$\begin{aligned}
\|x_{k+1} - x_*\|^2 &= \|x_k - x_*\|^2 - 2\langle x_{k+1} - x_*, x_k - x_{k+1}\rangle - \|x_k - x_{k+1}\|^2 \\
&\overset{9}{\leq} \|x_k - x_*\|^2 - 2\gamma(f_{\xi_k}(x_{k+1}) - f_{\xi_k}(x_*)) - \|x_k - x_{k+1}\|^2 \\
&\overset{2}{\leq} \|x_k - x_*\|^2 - 2\gamma(f_{\xi_k}(\bar{x}_{k+1}) + \langle \nabla f_{\xi_k}(\bar{x}_{k+1}), x_{k+1} - \bar{x}_{k+1}\rangle - f_{\xi_k}(x_*)) - \|x_k - x_{k+1}\|^2.
\end{aligned}$$

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:

$$\begin{aligned}
\mathbb{E}\left[\|x_{k+1} - x_*\|^2|\mathcal{F}_k\right] &\leq \|x_k - x_*\|^2 - 2\gamma\mathbb{E}\left[f_{\xi_k}(\bar{x}_{k+1}) + \langle \nabla f_{\xi_k}(\bar{x}_{k+1}), x_{k+1} - \bar{x}_{k+1}\rangle - f_{\xi_k}(x_*)|\mathcal{F}_k\right] \\
&\quad - \mathbb{E}\left[\|x_k - x_{k+1}\|^2|\mathcal{F}_k\right] \\
&= \|x_k - x_*\|^2 - 2\gamma\left(f(\bar{x}_{k+1}) - f(x_*) + \mathbb{E}\left[\langle \nabla f_{\xi_k}(\bar{x}_{k+1}), x_{k+1} - \bar{x}_{k+1}\rangle|\mathcal{F}_k\right]\right) \\
&\quad - \mathbb{E}\left[\|x_{k+1} - \bar{x}_{k+1}\|^2|\mathcal{F}_k\right] - \|x_k - \bar{x}_{k+1}\|^2.
\end{aligned} \tag{27}$$

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

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

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

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

Finally, using the star similarity condition, we obtain:

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

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

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

Taking the full expectation, we obtain:

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

Applying this inequality recursively, we get:

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

■

Theorem 28 *Let Assumptions 1 (Differentiability), 2 (Convexity), 4 (Interpolation) and 8 (Star Similarity) hold. Consider SPPM-inexact with Assumption 9 satsified. 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 $k \geq 0$,*

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

Proof To avoid repetitions, we start from (20):

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

Instead of (26), we define $\bar{x}_{k+1} = \mathbb{E}[x_k^\Psi|x_k]$.

Using convexity of f_{ξ_k}, we get:

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

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:

$$\begin{aligned}\mathbb{E}\left[\|x_{k+1}-x_*\|^2|x_k\right] &\leq \|x_k-x_*\|^2 - (1-\frac{\eta\gamma^2}{T^\alpha})E\left[\|x_k-x_k^\Psi\|^2|x_k\right] \\ &\quad - 2\gamma\mathbb{E}\left[f_{\xi_k}(\bar{x}_{k+1}) + \langle\nabla f_{\xi_k}(\bar{x}_{k+1}), x_k^\Psi-\bar{x}_{k+1}\rangle - f_{\xi_k}(x_*)|x_k\right] \\ &= \|x_k-x_*\|^2 - 2\gamma\left(f(\bar{x}_{k+1}) - f(x_*) + \mathbb{E}\left[\langle\nabla f_{\xi_k}(\bar{x}_{k+1}), x_k^\Psi-\bar{x}_{k+1}\rangle|x_k\right]\right) \\ &\quad - (1-\frac{\eta\gamma^2}{T^\alpha})\mathbb{E}\left[\|x_k^\Psi-\bar{x}_{k+1}\|^2|x_k\right] - (1-\frac{\eta\gamma^2}{T^\alpha})\|x_k-\bar{x}^{k+1}\|^2. \qquad (28)\end{aligned}$$

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

$$\mathbb{E}\left[\|x_{k+1} - x_*\|^2 | x_k\right] \leq \|x_k - x_*\|^2 - 2\gamma\left(f(\bar{x}_{k+1}) - f(x_*) + \mathbb{E}\left[\langle \nabla f_{\xi_k}(\bar{x}_{k+1}) - f(\bar{x}_{k+1}), x_k^{\Psi} - \bar{x}_{k+1}\rangle | x_k\right]\right)$$
$$- (1 - \frac{\eta\gamma^2}{T^\alpha})\mathbb{E}\left[\|x_k^{\Psi} - \bar{x}_{k+1}\|^2 | x_k\right] - (1 - \frac{\eta\gamma^2}{T^\alpha})\|x_k - \bar{x}^{k+1}\|^2. \tag{29}$$

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

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

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

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

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

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

Now, using the star similarity assumption (8), we get:

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

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

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

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:

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

Taking the full expectation, we obtain:

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

Iterating over k, the result follows:

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

■