Stochastic Weakly Convex Optimization beyond Lipschitz Continuity

Wenzhi Gao [1] **Qi Deng** [2]

Abstract

This paper considers stochastic weakly convex optimization without the standard Lipschitz continuity assumption. Based on new robust regularization (stepsize) strategies, we show that a wide class of stochastic algorithms, including the stochastic subgradient method, preserve the $\mathcal{O}(1/\sqrt{K})$ convergence rate with constant failure rate. Our analyses rest on rather weak assumptions: the Lipschitz parameter can be either bounded by a general growth function of $\|x\|$ or locally estimated through independent random samples. Numerical experiments demonstrate the efficiency and robustness of our proposed stepsize policies.

1. Introduction

This paper studies the stochastic optimization problem

$$\min_{x\in\mathbb{R}^n} \psi(x) \coloneqq f(x) + \omega(x), \tag{1}$$

where $f(x) \coloneqq \mathbb{E}_{\xi\sim\Xi}[f(x,\xi)]$. Here, $f(x,\xi)$ is a continuous function in x, with ξ being a random sample drawn from a particular distribution Ξ. The function $\omega(x)$ is lower-semicontinuous, and its proximal mapping is easy to evaluate. We assume both $f(x,\xi)$ and $\omega(x)$ are weakly convex functions. A function g is defined as λ-weakly convex if $g + \frac{\lambda}{2}\|\cdot\|^2$ is convex, for some $\lambda \geq 0$. When λ is unspecified, g is called weakly convex. Weak convexity has found many important applications, including phase retrieval, robust PCA, reinforcement learning, and many others (Duchi and Ruan, 2019; Charisopoulos et al., 2021; Wang et al., 2023). And recent years witnessed a surge in interest regarding weakly convex optimization, leading to a substantial body of work on efficient algorithms with finite time complexity guarantees (Davis and Drusvyatskiy, 2019; Davis et al., 2018b; Deng and Gao, 2021; Davis et al., 2019; Mai and Johansson, 2020). In particular, under the global Lipschitz continuity assumption, Davis and Drusvyatskiy (2019) develop a model-based approach and analyze the convergence of several stochastic algorithms under a unified framework.

While this global Lipschitz assumption is valid for many problems, such as piece-wise linear functions, it can be overly restrictive. To illustrate, consider the weakly convex function $\psi(x) = |e^x + e^{-x} - 3|$ whose subgradient $\psi'(x)$ explodes exponentially as $\|x\|$ grows (**Figure 1**). Hence, treating the Lipschitz constant as any fixed constant in algorithm design can lead to highly unstable iterations and, potentially, to the algorithm divergence.

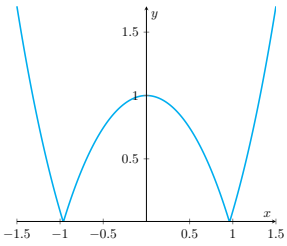

Figure 1: Exponential growth of $\psi(x) = |e^x + e^{-x} - 3|$

To address this issue, a straightforward strategy is to impose an explicit convex compact set constraint, such as $\{x : \|x\| \leq B\}$ to address this issue. However, this introduces extra parameter tuning and may lead to a significantly overestimated Lipschitz constant. The latter phenomenon is evident in the toy example, where the Lipschitz constant grows exponentially with the domain set's diameter. One research direction to deal with non-globally Lipschitz settings is shifting from standard Euclidean geometry to Bregman divergence. Lu (2019) shows that when convex non-Lipschitz functions exhibit "relative" Lipschitz continuity under a carefully chosen divergence kernel, mirror descent still obtains the desired sublinear rate of convergence to optimality defined in the sense of Bregman divergence. However, there are trade-offs to consider. Compared to `SGD`, a mirror descent update is more expensive, often involving a nontrivial root-finding procedure. Additionally, choosing the right kernel is a nuanced and critical task, heavily reliant on an in-depth understanding of the

[1]Institute for Computational and Mathematical Engineering, Stanford University [2]Antai College of Economics and Management, Shanghai Jiao Tong University. Correspondence to: Wenzhi Gao <gwz@stanford.edu>, Qi Deng <qdeng24@sjtu.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

subgradient's growth dynamics (Davis et al., 2018a; Zhang and He, 2018).

Alternatively, recent works aim to develop new algorithms/analyses under relaxed Lipschitz assumptions. For example, Asi and Duchi (2019) show that for the stochastic proximal point method, algorithmic dependency on the global Lipschitz constant can be relaxed to $\mathbb{E}[\|f'(x^\star,\xi)\|^2]$, which only relies on the optimal solution $x^\star$. Mai and Johansson (2021) show that, for stochastic convex optimization with quadratic growth, subgradient methods incorporating a clipping stepsize still ensure convergence, even if the Lipschitz constant exhibits arbitrary growth. In weakly convex optimization, Li et al. (2023b) shows that when the Moreau envelope of objective has a bounded level-set, local Lipschitz continuity alone is sufficient to ensure convergence of the subgradient method. Nevertheless, extending their analysis to stochastic optimization remains challenging. Grimmer (2019) establish the convergence of normalized subgradient in convex optimization without Lipschitz continuity by considering an upper bound of the form

$$f(x) - f(x^\star) \leq \mathcal{G}(\|x - x^\star\|) \tag{2}$$

where $\mathcal{G}$ is a growth function that allows fast growth of f. In (Zhu et al., 2023), the authors propose a relaxed subgradient bound for weakly convex optimization:

$$\mathbb{E}[\|f'(x,\xi)\|^2] \leq c_0 + c_1\|x\|,\ c_0, c_1 \geq 0, \tag{3}$$

which naturally induces a bound on the local Lipschitzness as a function of $\|x\|$. Whether SGD still converges in case of arbitrary non-Lipschitzness, especially those not conforming to the bounded assumption in (3), remains an open area of investigation. The primary difficulty in analyzing stochastic optimization without the standard Lipschitz assumption stems from stability issues. We consider a stochastic algorithm stable if it produces iterations in a bounded set with a probability greater than 0. Unlike the deterministic case, establishing stability in the face of randomness is not straightforward, especially when dealing with non-convex functions. This challenge motivates exploring an appropriate definition of non-Lipschitzness and the development of efficient algorithms for stochastic weakly convex optimization in this non-standard setting.

1.1. Contributions

This paper provides an affirmative answer to the question

Can we optimize stochastic weakly convex problems without assuming global Lipschitz continuity?

We show that carefully chosen robust stepsizes can effectively adapt to arbitrary non-Lipchitzness in stochastic model-based weakly convex optimization. Our contributions are as follows:

1) When the Lipschitz constant is not uniformly bounded above but instead depends on a general growth function $\mathcal{G}(\cdot)$, we design a novel robust adaptive stepsize strategy such that stochastic weakly convex optimization achieves the $\mathcal{O}(1/\sqrt{K})$ convergence rate with a constant failure probability. Our analysis does not assume any specific form of $\mathcal{G}$, such as those implied by (3). To our knowledge, this is the first result of stochastic weakly convex optimization for arbitrary non-Lipschitz objectives. Our analysis applies to a broad class of model-based algorithms (Davis and Drusvyatskiy, 2019; Deng and Gao, 2021), including SGD as a special case. Compared to Davis and Drusvyatskiy (2019), our analysis relaxes the global Lipschitz assumption and makes the model-based framework applicable to a broader range of settings.
2) Even if the growth function $\mathcal{G}$ is unknown, we show that achieving the same convergence guarantee is still possible. To this end, we introduce a new robust stepsize based on the concept of "reference Lipschitz continuity", which allows us to estimate the Lipschitz parameter of a stochastic model function using local samples. Our algorithm is highly flexible and can be applied to most weakly convex problems of interest. Moreover, our analyses can be extended to solving convex stochastic optimization without Lipschitz continuity. A more detailed discussion is left to **Section E**.

Model-based Optimization Model-based optimization, as proposed by (Davis and Drusvyatskiy, 2019), serves as a general framework for analyzing stochastic weakly convex optimization. This framework has been leveraged by several papers (Davis et al., 2018a; Chadha et al., 2022; Deng and Gao, 2021; Gao and Deng, 2024) to obtain convergence rates for a broad class of algorithms. Our analysis also builds on this framework and extends it to several algorithms.

Other Related Works Adaptive stepsize and gradient clipping are two essential tools adopted in our algorithm framework. On the one hand, stepsize selection has been an important topic in stochastic optimization, and it has been justified that adaptive stepsize benefits stochastic first-order methods both in theory and in practice (Duchi et al., 2011; Kingma and Ba, 2014; Li and Orabona, 2019; Hinton et al., 2012; Defazio and Mishchenko, 2023; Ivgi et al., 2023; Malitsky and Mishchenko, 2023). On the other hand, gradient clipping (Zhang et al., 2019) will be employed as a technique in the paper. In theory, gradient clipping was initially identified as a tool to solve problems with generalized Lipschitz smoothness condition (Li et al., 2023a; Xie et al., 2023; Zhang et al., 2019). Recent works (Gorbunov et al., 2020; Koloskova et al., 2023) show that gradient clipping can effectively deal with problems with heavy-tail noise.

It is also observed that gradient clipping improves the robustness and stability of SGD (Mai and Johansson, 2020) in stochastic convex optimization. In our analysis, a generalized version of gradient clipping is developed to alleviate the instability arising from stochastic noise.

2. Preliminaries

Notations Throughout the paper $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ denote the Euclidean inner product and norm. Subdifferential of f is given by $\partial f(x) \coloneqq \{v : f(y) \geq f(x) + \langle v, y-x\rangle + o(\|x-y\|), y \to x\}$ and $f'(x) \in \partial f(x)$ is called a subgradient. A growth function $\mathcal{G}(\cdot)$ is a continuous non-decreasing function mapping from $\mathbb{R}_+$ to $\mathbb{R}_+$.

Model Function and Model-based Optimization Our main algorithm will be presented in a "model-based" fashion, which encompasses several first-order methods, including the most widely used (proximal) subgradient method. Model-based optimization (Davis and Drusvyatskiy, 2019) contains two components: a stochastic model function and a stepsize (regularization) parameter. In each iteration, we can construct a local approximation of $f(x)$ based on random sample ξ^k and the current iterate x^k. The stochastic function, denoted by $f_{x^k}(\cdot,\xi^k)+\omega(x)$, is called a model function. Then we take parameter γ_k and minimize this local approximation under quadratic regularization $\frac{\gamma_k}{2}\|x-x^k\|^2$ to obtain the next iterate x^{k+1}. Typical models include

- (Sub)gradient. Given $\mathbb{E}[f'(x,\xi)] = f'(x) \in \partial f(x)$, $f_x(y,\xi) = f(x,\xi) + \langle f'(x,\xi), y-x\rangle$.
- Prox-linear. Given $f(x,\xi) = h(c(x,\xi))$, $f_x(y,\xi) = h(c(x,\xi) + \langle \nabla c(x,\xi), y-x\rangle)$.
- Truncated. Given a known lower-bound of model ℓ, $f_x(y,\xi) = \max\{f(x,\xi) + \langle \nabla f(x,\xi), y-x\rangle, \ell\}$.

Algorithm 1 summarizes model-based optimization.

Algorithm 1 Stochastic model-based optimization

Input x^1

for k = 1, 2,... **do**

Sample data ξ^k and choose regularization $\gamma_k > 0$

$$x^{k+1} = \arg\min_x \left\{ f_{x^k}(x,\xi^k) + \omega(x) + \tfrac{\gamma_k}{2}\|x-x^k\|^2 \right\}. \tag{4}$$

end

We see that **1)** model function $f_x(\cdot,\xi)$; **2)** regularization parameter γ_k are two core components for our algorithm design. Throughout this paper, we show how properly chosen γ_k improves convergence beyond Lipschitz continuity. We start by making assumptions.

Envelope Smoothing Our analysis adopts the Moreau envelope as the potential function for weakly convex optimization. Let f be a λ-weakly convex function. Given $\rho > \lambda$, the Moreau envelope and the associated proximal mapping of f are given by

$$f_{1/\rho}(x) \coloneqq \min_y \left\{ f(y) + \tfrac{\rho}{2}\|x-y\|^2 \right\}$$
$$\operatorname{prox}_{f/\rho}(x) \coloneqq \arg\min_y \left\{ f(y) + \tfrac{\rho}{2}\|x-y\|^2 \right\}.$$

Moreau envelope can be interpreted as a smooth approximation of the original function. $f_{1/\rho}(x)$ is differentiable with gradient

$$\nabla f_{1/\rho}(x) = \rho(x - \operatorname{prox}_{f/\rho}(x))$$

If $\|\nabla f_{1/\rho}(x)\| \leq \varepsilon$, then x is in the proximity of a near stationary point $\operatorname{prox}_{f/\rho}(x)$. An important observation is that the existence of $f_{1/\rho}(x)$ relies on weak convexity instead of Lipschitz continuity of f.

Assumptions. We make the following assumptions.

A1: It is possible to generate i.i.d. samples $\{\xi^k\}$.

A2: $\omega(x)$ is κ-weakly convex and L_ω-Lipschitz continuous for all $x \in \operatorname{dom}\omega$.

A3: $\mathbb{E}_\xi[f_x(x,\xi)] = f(x)$ for all $x \in \operatorname{dom}\omega$ and

$$\mathbb{E}_\xi[f_x(y,\xi) - f(y)] \leq \tfrac{\tau}{2}\|x-y\|^2$$

for all $x, y \in \operatorname{dom}\omega$. Moreover, $f_x(y,\xi)$ is convex for all $x, y \in \operatorname{dom}\omega$ and $\xi \sim \Xi$.

A4: $\psi_{1/\rho}(x)$ is lower bounded by $-\Lambda \leq 0$.

A5: The v-level-set of $\psi_{1/\rho}(x)$:

$$\mathcal{L}_v = \{x : \psi_{1/\rho}(x) \leq v\}$$

has a bounded diameter $\operatorname{diam}(\mathcal{L}_v) \leq \mathsf{B}_v < \infty$.

Remark 1. Given **A1** to **A3**, it follows that $f(x)$ is τ-weakly convex and $\psi(x)$ is $(\tau+\kappa)$ weakly convex.

Remark 2. In **A3**, we adopt a convex model function since all the models considered in this paper are convex. But the result directly generalizes to weakly convex model functions: for example, if $f_x(x,\xi)$ is λ-weakly convex, then

$$f_x(x,\xi) + \omega(x) = \left[f_x(x,\xi) + \tfrac{\lambda}{2}\|x\|^2\right] + \left[\omega(x) - \tfrac{\lambda}{2}\|x\|^2\right]$$

and we can redefine ω and $f_x(x,\xi)$ to push weak convexity to the proximal term.

Remark 3. **A5** is a key assumption in dealing with non-Lipschitzness. It implies that by bounding the Moreau envelope as the potential function, we can control the Lipschitzness as a function of x. Typically, **A5** is satisfied when ψ is coercive (Li et al., 2023b), which is natural in our non-Lipschitz (e.g., a high-order polynomial) context, as the function growth is faster than linear. It's important to note that non-Lipschitzness can also arise in other cases, such as in the interpolation setting, which may require different analyses (Li et al., 2023b).

Structure of the Paper The paper is organized as follows. **Section 3** discusses the convergence of weakly convex optimization with the standard Lipschitz condition, which serves as a benchmark to provide sufficient intuition for the algorithm design in more challenging scenarios. **Section 4** and **5** discuss two cases where the standard Lipschitz condition fails to hold. **Section 6** conducts numerical experiments to verify our results.

3. Optimization of Standard Lipschitzness

The results in this section are already available in the literature (Davis and Drusvyatskiy, 2019), and the goal is to provide a benchmark result and establish some basic intuitions. We assume that

B1: For any x, y and $\xi \sim \Xi$,

$$f_x(x,\xi) - f_x(y,\xi) \le L_f(\xi)\|x-y\|$$

and $\mathbb{E}[L_f(\xi)^2] \le L_f^2$.

B1 is common in nonsmooth optimization, and the following descent property is available.

Lemma 3.1. *Let $\hat{x}^k = \mathrm{prox}_{\psi/\rho}(x^k)$. Suppose that **A1** to **A3** as well as **B1** holds, then given $\rho > \kappa + \tau, \gamma_k > \rho$,*

$$\begin{aligned}&\mathbb{E}_k[\psi_{1/\rho}(x^{k+1})] \\ &\le \psi_{1/\rho}(x^k) - \tfrac{\rho(\rho-\tau-\kappa)}{2(\gamma_k-\kappa)}\|\hat{x}^k - x^k\|^2 + \tfrac{2\rho L_f^2}{(\gamma_k-\rho)(\gamma_k-\kappa)}\end{aligned} \tag{5}$$

where $\mathbb{E}_k[\cdot] := \mathbb{E}[\cdot|\xi^1,\dots,\xi^k]$ denotes the conditional expectation taken over $\xi^1,\dots,\xi^k$.

γ_k is generally much larger than the other constants in the algorithm, and Lemma 3.1 reveals the following relation

$$\begin{aligned}&\mathbb{E}_k[\psi_{1/\rho}(x^{k+1})] \\ &\le \psi_{1/\rho}(x^k) - \mathcal{O}(\gamma_k^{-1})\|\nabla\psi_{1/\rho}(x^k)\|^2 + \mathcal{O}(L_f^2\gamma_k^{-2}),\end{aligned} \tag{6}$$

where $\mathcal{O}(L_f^2\gamma_k^{-2})$ characterizes the error from both stochastic noise and nonsmoothness. Taking $\gamma_k \equiv \mathcal{O}(\sqrt{K})$ and telescoping Lemma 3.1, we get the convergence result.

Theorem 3.1. *Under the same conditions as Lemma 3.1, if we take $\gamma_k \equiv \rho + \kappa + \alpha\sqrt{K}$, then we have*

$$\begin{aligned}&\min_{1\le k\le K}\mathbb{E}[\|\nabla\psi_{1/\rho}(x^k)\|^2] \\ &\qquad\le \tfrac{2\rho}{\rho-\tau-\kappa}\Big[\tfrac{\rho D}{K} + \tfrac{1}{\sqrt{K}}\big(\alpha D + \tfrac{2\rho}{\alpha}L_f^2\big)\Big],\end{aligned}$$

where $D = \psi(x^1) - \inf_x \psi(x)$.

Theorem 3.1 is standard in the literature (Davis and Drusvyatskiy, 2019). One important intuition we want to establish is that the choice of $\gamma_k \equiv \mathcal{O}(\sqrt{K})$ is a consequence of the following trade-off: suppose we telescope over (6) directly, then

$$\begin{aligned}&\tfrac{1}{K}\textstyle\sum_{k=1}^K \mathcal{O}(\gamma_k^{-1})\mathbb{E}[\|\nabla\psi_{1/\rho}(x^k)\|^2] \\ &\qquad\le \mathcal{O}(\tfrac{1}{K}) + \tfrac{1}{K}\textstyle\sum_{k=1}^K \mathcal{O}(L_f^2\gamma_k^{-2}).\end{aligned}$$

First, we need large γ_k, in other words, a conservative stepsize (large γ), such that the error of potential reduction $\mathcal{O}(\sum_k L_f^1\gamma_k^{-2})$ is properly bounded. Meanwhile, large γ also leads to the amount of potential reduction $\mathcal{O}(\gamma_k^{-1})\|\nabla\psi_{1/\rho}(x^k)\|^2$ being small. Finally, the optimal trade-off is $\gamma_k \equiv \mathcal{O}(\sqrt{K})$ and an $\mathcal{O}(1/\sqrt{K})$ rate of convergence. As we will show in the following sections, with non-Lipschitzness, the error $\mathcal{O}(L_f^2\gamma_k^{-2})$ cannot be bounded by choosing some constant γ_k. What we do is adaptively find suitable and robust γ_k, to reduce the error. Using **A5** and a probabilistic analysis, we achieve this goal without compromising the convergence rate.

4. Optimization of Generalized Lipschitzness

Before achieving our goal of solving non-Lipschitz weakly convex optimization problems, we start from a less challenging case characterized as follows.

C1: For all x, y, z and $\xi \sim \Xi$,

$$f_x(y,\xi) - f_x(z,\xi) \le L_f(\xi)\mathcal{G}(\|x\|)\|y-z\|$$

and $\mathbb{E}[L_f(\xi)^2] \le L_f^2$; Recall that growth function $\mathcal{G}$ is monotonically increasing.

This assumption implies that our model function is globally Lipschitz, but the Lipschitz constant has some known dependency on the norm of the model's expansion point x. Our analysis also applies if we can estimate a non-trivial upper bound of $\mathcal{G}$. But for the brevity of analysis, we take this upper bound to be $\mathcal{G}$ itself. Many real-life applications have this structure, especially if the source of non-Lipschitzness is a high-order polynomial.

Example 4.1 (Phase retrieval)**.** *Consider*

$$f(x,\xi) = |\langle a, x\rangle^2 - b|, a \in \mathbb{R}^n, b \in \mathbb{R}_+.$$

The subgradient model

$$f_x(y,\xi) = f(x,\xi) + \langle f'(x,\xi), y-x\rangle \\ = f(x,\xi) + \langle 2\cdot \mathrm{sign}(\langle a,x\rangle - b)\langle a,x\rangle a, y-x\rangle$$

satisfies $f_x(y,\xi) - f_x(z,\xi) \le 2\|a\|^2\|x\|\cdot\|y-z\|$.

Example 4.2 (Subgradient with known growth)**.** *SGD corresponds to the model function*

$$f_x(y,\xi) = f(x,\xi) + \langle f'(x,\xi), y-x\rangle.$$

Then, **C1** *is satisfied if* $\|f'(x,\xi)\| \le L_f(\xi)\mathcal{G}(\|x\|)$*. It follows that* $\mathbb{E}[\|f'(x,\xi)\|^2] \le L_f^2\mathcal{G}^2(\|x\|)$*. Particularly,* (3) *corresponds to the case of* $\mathcal{G}^2(\cdot)$ *being a linear function.*

One direct consequence of **C1** is that the Lipschitz constant of $f_x(\cdot,\xi)$ can go to ∞ when $\|x\|\to\infty$. Moreover, we cannot directly rely on Lipschitzness of $f(x)$. Taking subgradient update as an example, this implies $f'(x,\xi)$ can have a large norm, leading to a higher chance of divergence. From the perspective of convergence analysis, the error term $\mathcal{O}(L_f^2\gamma_k^{-2})$ in (6) becomes hard to bound, and Lemma 4.1 quantifies this hardness.

Lemma 4.1. *Suppose* **A1** *to* **A3** *as well as* **C1** *holds, then given* $\rho > \kappa+\tau, \gamma_k > \rho$*,*

$$\mathbb{E}_k[\psi_{1/\rho}(x^{k+1})] \le \psi_{1/\rho}(x^k) - \tfrac{\rho(\rho-\tau-\kappa)}{2(\gamma_k-\kappa)}\|\hat{x}^k - x^k\|^2 \\ + \tfrac{\rho}{2\gamma_k(\gamma_k-\kappa)}(\mathcal{G}(\|x^k\|)L_f + L_\omega)^2$$

where γ_k *is independent of* ξ^k*.*

According to Lemma 4.1, the error of potential reduction involves the norm of x^k. Since $\mathcal{G}(\|x^k\|)$ is not necessarily bounded above, the previous constant stepsize analysis is not applicable. As a natural fix, we can take

$$\gamma_k = \mathcal{O}\big(\mathcal{G}(\|x^k\|)\sqrt{K}\big)$$

to reduce the error. However, according to the trade-off we previously mentioned, unless $\mathcal{G}(\|x^k\|)$ is bounded by some constant independent of K, the reduction in the potential function can be arbitrarily small, and we still cannot obtain a convergence rate. To resolve this issue, we essentially need to show the boundedness of $\{\|x^k\|\}$, and our solution is to associate the boundedness of $\{\|x^k\|\}$ with another bounded quantity during the algorithm: $\mathbb{E}[\psi_{1/\rho}(x^k)]$. Intuitively, since $\mathbb{E}[\psi_{1/\rho}(x^k)]$ is reduced by the algorithm, it remains bounded on expectation, and from **A5** we know that boundedness of $\psi_{1/\rho}(x^k)$ implies boundedness of $\|x^k\|$, giving bounded $\mathcal{G}(\|x^k\|)$ and the $\mathcal{O}(1/\sqrt{K})$ rate we want. The following asymptotic result confirms our intuition.

Theorem 4.1. *Under the same conditions as Lemma 4.1 and* **A4**, **A5**, *if* $\gamma_k = \rho+\kappa+(\mathcal{G}(\|x^k\|)+1)k^\zeta, \zeta\in(\frac{1}{2},1)$*, then as* $k\to\infty$*,* $\{\|x^k\|\}$ *is bounded with probability* 1*; Moreover, the sequence* $\{\inf_{j\le k}\|\nabla\psi_{1/\rho}(x^j)\|\}$ *converges to* 0 *almost surely.*

While it is relatively easy to show asymptotic convergence, we need a more careful analysis of the algorithm behavior to obtain a finite-time convergence rate. One significant difficulty is that the boundedness of $\mathbb{E}[\psi_{1/\rho}(x^k)]$ is may not directly provide information of $\|x^k\|$, since this relation holds only on expectation. To deal with this issue, we resort to probabilistic tools and establish a new probabilistic argument in the following subsection.

4.1. Stability of the Iterations

In this subsection, we aim to analyze the stability of the iterates of a stochastic algorithm on a non-Lipschitz function. The intuition is straightforward: if a stochastic algorithm reduces some potential function that has a bounded level-set, then the iterates will stay in a bounded region with high probability. We provide the basic proof sketch and leave a more rigorous argument in the appendix.

Our analysis relies on two simple facts that we gain from the robust stepsize.

Lemma 4.2 (Informal)**.** *Under the same conditions as Lemma 4.3, if* $\gamma_k = \mathcal{O}\big((\mathcal{G}(\|x^k\|)+1)\sqrt{K}\big)$*, then we have, for all* $k=2,\dots,K$*, that*

$$\mathbb{E}[\|x^{k+1}-x^k\|] \le \mathcal{O}(\tfrac{1}{\sqrt{K}}), \tag{7}$$

$$\mathbb{E}[\psi_{1/\rho}(x^k)] \le \mathcal{O}(1). \tag{8}$$

Relation (7) says that with our robust stepsize strategy, we cautiously explore the feasible region, and at each iteration, we only take a small step of $\mathcal{O}(1/\sqrt{K})$. The second relation (8) comes directly from Lemma 4.1. Indeed, with

$$\gamma_k = \mathcal{O}\big((\mathcal{G}(\|x^k\|)+1)\sqrt{K}\big)$$

we have $\frac{\rho(\mathcal{G}(\|x^k\|)L_f+L_\omega)^2}{2\gamma_k(\gamma_k-\kappa)} = \mathcal{O}(1/K)$. And if we telescope Lemma 4.1 and take expectation over all the randomness for $k=2,\dots K$, we have

$$\mathbb{E}[\psi_{1/\rho}(x^k)] \le \textstyle\sum_{j=1}^k \mathcal{O}(1/K) = \mathcal{O}(1), k=2,\dots,K.$$

Each of the two relations alone may not offer us helpful information, as they both hold on expectation. However, when they are combined, a more useful result is available. Our argument is as follows:

Consider the event "$\|x^k\|$ is large" and we wish to upper-bound its probability. We have the following facts:

1. If $\|x^k\|$ is large, $\|x^{k+1}\| \ge \|x^k\| - \mathcal{O}(1/\sqrt{K})$ by triangle inequality and $\|x^{k+1}\|$ is also large.
2. If $\|x^{k+1}\|$ is large, then $\psi_{1/\rho}(x^{k+1})$ is large by **A5**.
3. $\mathbb{E}[\psi_{1/\rho}(x^{k+1})]$ is bounded by some constant.

In other words, conditioned on the event "$\|x^k\|$ is large", to ensure that $\mathbb{E}[\psi_{1/\rho}(x^{k+1})]$ is still bounded, either **Case**

1): the event happens with low probability, or **Case 2):** x^{k+1} has to immediately jump back to a bounded region of smaller radius. However, since our robust stepsize restricts the "jump" between two consecutive iterations, **Case 2)** cannot happen. Therefore, it is unlikely "$\|x^k\|$ is large".

This argument brings us the following tail-bound which characterizes the behavior of $\|x^k\|$ as a random variable.

Lemma 4.3. *Under the same conditions as Lemma 4.1 as well as* **A4**, **A5**, *if we take* $\gamma_k = \rho + \kappa + \tau + \alpha(\mathcal{G}(\|x^k\|) + 1)\sqrt{K}$, *then the tail bound*

$$\mathbb{P}\{\|x^k\| \geq \mathsf{B}_{a\Delta} + \tfrac{4(L_f+L_\omega)}{\alpha\sqrt{K}}\} \leq \tfrac{2\Delta}{a\Delta+\Lambda},$$

holds for all $2 \leq k \leq K$, *where*

$$\Delta = \psi_{1/\rho}(x^1) + \Lambda + \tfrac{\rho(L_f+L_\omega)^2}{\alpha^2} > 0$$

and recall that $\operatorname{diam}(\mathcal{L}_{a\Delta}) \leq \mathsf{B}_{a\Delta}$.

Lemma 4.3 provides a useful characterization of the tail probability on the norm of the iterations. Now that the bound holds for all x^k, we can immediately condition on the event that $\Theta(K)$ iterations lie in the bounded set to retrieve an $\mathcal{O}(1/\sqrt{K})$ convergence rate.

Theorem 4.2. *Under the same conditions as Lemma 4.3, given* $\delta \in (0, 1/4)$, *at least with probability* $1-p, p \in (2\delta, 1)$, $(1-2p^{-1}\delta)K$ *iterations will be bounded by*

$$\mathsf{R}(\delta) = \mathsf{B}_{\delta^{-1}\Delta} + \tfrac{4(L_f+L_\omega)}{\alpha\sqrt{K}},$$

and conditioned on these iterations,

$$\min_{1\leq k\leq K} \mathbb{E}[\|\nabla\psi_{1/\rho}(x^k)\|^2] \leq \tfrac{pM}{p-2\delta}\left(\tfrac{\rho+\tau+\kappa}{K} + \tfrac{\alpha(\mathsf{G}_\delta+1)}{\sqrt{K}}\right),$$

where $M = \frac{2\rho}{\rho-\tau-\kappa}\left[D + \frac{\rho}{2\alpha^2}(L_f+L_\omega)^2\right]$ *and* $\mathsf{G}_\delta := \max_{z\leq\mathsf{R}(\delta)} \mathcal{G}(z)$.

Theorem 4.2 shows that, with constant probability, we retrieve $\mathcal{O}(1/\sqrt{K})$ convergence rate after K iterations. This probability argument can be further improved, for example, by running the algorithm independently multiple times (Davis and Grimmer, 2019). The analysis in this section serves as a step-stone for the next section, where we deal with non-Lipschitz optimization without knowing $\mathcal{G}(\cdot)$.

5. Optimization of unknown Lipschitzness

While the analysis in **Section 4** extends the solvability of weakly convex optimization to non-Lipschitz functions, it relies on the knowledge of an explicit growth function $\mathcal{G}(\|x\|)$ to bound local Lipschitzness. However, it is possible that either access to $\mathcal{G}(\|x\|)$ is not viable, or the bound lacks a predefined functional form. In these cases, we assume that the growth function is unknown a priori.

D1: For all fixed $x \in \operatorname{dom}\omega$,

$$f_x(z,\xi) - f_x(y,\xi) \leq \mathsf{Lip}(x,\xi)\|z-y\|$$

for all y, z; $\xi \sim \Xi$.

Although **D1** and **C1** look similar, they are fundamentally different. The most direct consequence is that $\mathsf{Lip}(x,\xi)$ is sample-dependent, which means any stepsize strategy based on it will introduce bias in the stochastic algorithm. To resolve this issue, our new stepsize policy relies on constructing an estimator of $\mathsf{Lip}(x,\xi)$. We introduce the property of "admitting a reliable estimation of Lipschitz constant", which we call the *reference Lipschitz continuity*.

5.1. Reference Lipschitz Continuity

Definition 1 (Reference Lipschitz continuity)**.** *Stochastic model* $f_x(y,\xi)$ *satisfies reference Lipschitz continuity if*

- *given an* x, $f_x(\cdot,\xi)$ *is globally Lipschitz with a Lipschitz constant* $\mathsf{Lip}(x,\xi)$*;*
- *given* $\xi, \xi' \sim \Xi$,

$$\mathbb{E}_{\xi,\xi'}[|\mathsf{Lip}(x,\xi) - \mathsf{Lip}(x,\xi')|^2] \leq \sigma^2 < \infty.$$

The first property, entirely determined by the stochastic model function, is typical, as most model functions are compositions of Lipschitz functions and linear expansions. The second property indicates that the expected difference between models' Lipschitzness is bounded by noise parameter σ. This property is also assumed in literature (Mai and Johansson, 2021) when dealing with stochastic subgradient of a function with arbitrary growth. As we will demonstrate in the examples, for most functions it can be deduced from bounded variance assumption.

One direct outcome of reference Lipschitz continuity is that we can cheaply estimate $\mathsf{Lip}(x,\xi)$ based on $\mathsf{Lip}(x,\xi')$, where ξ and ξ' are two independent samples drawn from Ξ.

Example 5.1 ((Sub)gradient)**.**

$$f_x(y,\xi) = f(x,\xi) + \langle\nabla f(x,\xi), y-x\rangle$$

The model is Lipschitz with $\mathsf{Lip}(x,\xi) = \|\nabla f(x,\xi)\|$. If $\mathbb{E}[\|\nabla f(x) - \nabla f(x,\xi)\|^2] \leq \sigma^2$, then

$$\begin{aligned}&\mathbb{E}[|\mathsf{Lip}(x,\xi) - \mathsf{Lip}(x,\xi')|]\\ &\quad\leq \mathbb{E}[\|\nabla f(x,\xi) - \nabla f(x,\xi')\|] \leq 2\sigma\end{aligned}$$

Even if f is nonsmooth, the property may still hold. One example is the composition problem $f(x,\xi) = h(c(x,\xi))$, where h is L_h-Lipschitz continuous and c is differentiable. Then $\partial f(x,\xi) = \nabla c(x,\xi)\partial h(c(x,\xi))$. If we estimate the Lipschitz constant with $L_h\|\nabla c(x,\xi)\|$, then $\mathbb{E}[|\mathsf{Lip}(f_x(\cdot,\xi)) - \mathsf{Lip}(x,\xi')|] \leq 2L_h\sigma$.

Example 5.2 (Proximal linear)**.**

$$f_x(y,\xi) = h(c(x,\xi) + \langle \nabla c(x,\xi), y - x\rangle)$$

When h is L_h Lipschitz continuous, the model is globally Lipschitz with $\mathsf{Lip}(x,\xi) = L_h\|\nabla c(x,\xi)\|$. If $\mathbb{E}[\|\nabla c(x) - \nabla c(x,\xi)\|^2] \le \sigma^2$, then

$$\begin{aligned}&\mathbb{E}[|\mathsf{Lip}(x,\xi) - \mathsf{Lip}(x,\xi')|]\\ &\quad\le L_h\mathbb{E}[\|\nabla c(x,\xi) - \nabla c(x,\xi')\|] \le 2L_h\sigma\end{aligned}$$

Example 5.3 (Truncated model)**.**

$$f_x(y,\xi) = \max\{f(x,\xi) + \langle \nabla f(x,\xi), y - x\rangle, \ell\}.$$

The model is $\|\nabla f(x,\xi)\|$-Lipschitz and the reasoning of reference Lipschitz continuity is the same as in **Example 5.1**. Note that the truncated model encompasses stochastic Polyak stepsize as a special case (Schaipp et al., 2023).

In this section, we would assume that our model satisfies the reference Lipschitz continuity.

D2: The stochastic model $f_x(\cdot,\xi)$ satisfies the reference Lipschitz continuity with noise parameter σ.

5.2. Algorithm Design and Analysis

As we did in **Section 4**, before getting down to the algorithm design, we first need to see what happens to our potential reduction in this new setting. Firstly, Lemma 5.1 characterizes the descent property of our potential function under the assumption that γ_k is independent of ξ^k.

Lemma 5.1. *Suppose* **A1** *to* **A3** *as well as* **D1**, **D2** *hold, then given $\rho > \kappa + \tau$,*

$$\begin{aligned}\mathbb{E}_k[\psi_{1/\rho}(x^{k+1})] \le \psi_{1/\rho}(x^k) - \tfrac{\rho(\rho-\tau-\kappa)}{2(\gamma_k-\kappa)}\|\hat{x}^k - x^k\|^2\\ + \mathbb{E}_k\big[\tfrac{\rho}{2\gamma_k(\gamma_k-\kappa)}(\mathsf{Lip}(x^k,\xi^k) + L_\omega)^2\big]\end{aligned}$$

where γ_k is chosen to be independent of ξ^k and is considered deterministic here.

Compared to Lemma 4.1, bounding the error term $\frac{\rho(\mathsf{Lip}(x^k,\xi^k))+L_\omega)^2}{2\gamma_k(\gamma_k-\kappa)}$ becomes more challenging due to the lack of information about its growth. However, thanks to the reference Lipschitz continuity, by sampling ξ', an independent copy of ξ^k, we can utilize $\mathsf{Lip}(x^k,\xi')$ as a surrogate for $\mathsf{Lip}(x^k,\xi^k)$. The preference of $\mathsf{Lip}(x^k,\xi')$ over $\mathsf{Lip}(x^k,\xi^k)$ is driven by the fact that $\mathsf{Lip}(x^k,\xi^k)$ is correlated with x^{k+1}, which significantly complicates the analysis (Andradóttir, 1996). To mitigate the impact of large noise σ on the accuracy of our estimation, we clip the estimator by a threshold $\alpha > 0$: $\max\{\mathsf{Lip}(x^k,\xi'),\alpha\}$. Consequently, we set

$$\gamma_k = \mathcal{O}(\max\{\mathsf{Lip}(x^k,\xi'),\alpha\}\cdot\sqrt{K}).$$

Remark 4. Our stepsize policy can be seen as a generalization of gradient clipping stepsize to the model-based optimization setting. In particular, when $f_x(y,\xi) = f(x,\xi) + \langle\nabla f(x,\xi), y - x\rangle$ and $\xi = \xi'$, we retrieve the clipping subgradient method.

The following theorem establishes an asymptotic result and confirms our intuition.

Theorem 5.1. *Under the same conditions as Lemma 5.1,* **A4**, **A5**, *if $\gamma_k = \rho + \kappa + \tau + \max\{\mathsf{Lip}(x^k,\xi'),\alpha\}k^\zeta$, $\zeta \in (\frac{1}{2},1)$, as $k\to\infty$, $\{\|x^k\|\}$ is bounded with probability 1; $\{\inf_{j\le k}\|\nabla\psi_{1/\rho}(x^j)\|\}$ converges to 0 almost surely.*

To obtain a non-asymptotic result, we again apply the probabilistic analysis to derive the tail bound.

Lemma 5.2. *Under the same conditions of Lemma 5.1 as well as* **A4**, **A5**, *if we take $\gamma_k = \rho + \tau + \kappa + \max\{\mathsf{Lip}(x^k,\xi'),\alpha\}\sqrt{K}$, then the tail bound*

$$\mathbb{P}\big\{\|x^k\| \ge \mathsf{B}_{a\Delta} + \tfrac{4(\alpha+\sigma+L_\omega)}{\alpha\sqrt{K}}\big\} \le \tfrac{2\Delta}{a\Delta+\Lambda},$$

holds for all $2\le k\le K$, where

$$\Delta = \psi_{1/\rho}(x^1) + \Lambda + \tfrac{\rho}{\alpha^2}(\alpha+\sigma+L_\omega)^2 > 0.$$

Theorem 5.2. *Assuming the conditions of Lemma 5.2 hold, then given $\delta\in(0,1/4)$, with probability at least $1-p, p\in(2\delta,1)$, $(1-2p^{-1}\delta)K$ iterations will lie in the ball with radius*

$$\mathsf{R}(\delta) = \mathsf{B}_{\delta^{-1}\Delta} + \tfrac{4(\alpha+\sigma+L_\omega)}{\alpha\sqrt{K}},$$

and conditioned on these iterations,

$$\min_{1\le k\le K}\mathbb{E}[\|\nabla\psi_{1/\rho}(x^k)\|^2] \le \tfrac{pM}{p-2\delta}\big(\tfrac{\rho+\tau+\kappa}{K} + \tfrac{\alpha+\mathsf{G}_\delta}{\sqrt{K}}\big),$$

where $M = \frac{2\rho}{\rho-\tau-\kappa}\big[D + \frac{\rho}{\alpha^2}(\alpha+\sigma+L_\omega)^2\big]$ and $\mathsf{G}_\delta := \max_{\|x\|\le\mathsf{R}(\delta)}\sup_{\xi\sim\Xi}\mathsf{Lip}(x,\xi)$.

Remark 5. We need to assume $\mathsf{G}_\delta < \infty$ in the analysis, which can be satisfied for finite sum optimization or when the support of data distribution Ξ is bounded.

6. Experiments

In this section, we perform experiments to demonstrate the effectiveness of our proposed methods. We consider the following robust nonlinear regression problem:

$$\min_x \quad \frac{1}{m}\sum_{i=1}^m |r(x,a_i) - b_i| =: \frac{1}{m}\sum_{i=1}^m f(x,\xi_i), \qquad (9)$$

where, given observations $\{a_i\}$ from $A\in\mathbb{R}^{m\times n}$, regression model $r(x,a)$ and target label b_i, we aim to fit the model coefficient x given problem data. **Table 1** summarizes the regression models and their Lipschitz properties.

Table 1: Nonlinear regression models. $r(a,x) = \langle a,x\rangle^2$ represents the standard robust phase retrieval problem; $r(a,x) = \langle a,x\rangle^5 + \langle a,x\rangle^3 + 1$ is a high-order polynomial of $\langle a,x\rangle$; while $e^{\langle a,x\rangle} + 10$ exhibits exponential growth.

Loss	$r(x,a)$	$\partial f(x,\xi)$	$\mathcal{G}(\|x\|)$	$\mathsf{Lip}(x,\xi)$
r_1	$\langle a,x\rangle^2$	$\mathrm{sign}(\langle a,x\rangle^2 - b)\cdot 2\langle a,x\rangle a$	$\|x\|$	$2\|\langle a,x\rangle\|\cdot\|a\|$
r_2	$\langle a,x\rangle^5 + \langle a,x\rangle^3 + 1$	$\mathrm{sign}(r_2 - b)\cdot(5\langle a,x\rangle^4 + 3\langle a,x\rangle^2)a$	$5(\|x\|^4 + \|x\|^2)$	$5(\|a\|^4\|x\|^4 + \|a\|^2\|x\|^2)$
r_3	$e^{\langle a,x\rangle} + 10$	$\mathrm{sign}(e^{\langle a,x\rangle} - b)\cdot e^{\langle a,x\rangle}\cdot a$	$e^{A\|x\|}$ $(A=3)$	$e^{\langle a,x\rangle}\cdot\|a\|$

6.1. Experiment Setup

Dataset. We let $m = 300, n = 100$. Data generation is consistent with (Deng and Gao, 2021), where, given condition number parameter $\kappa \geq 1$, we compute $A = QD, Q \in \mathbb{R}^{m\times n}$. Here each element of Q is drawn from $\mathcal{N}(0,1)$; $D = \mathrm{diag}(d), d \in \mathbb{R}^n, d_i \in [1/\kappa, 1]$ for all i. Then a true signal $\hat{x} \sim \mathcal{N}(0, I)$ is generated, giving the measurements b by formula $b_i = r(x, a_i)$. We randomly perturb p_{fail}-fraction of the measurements with $\mathcal{N}(0, 25)$ noise added to them to simulate data corruption.

1) **Dataset**. We follow (Deng and Gao, 2021) and set set $\kappa \in \{1, 10\}$ and $p_{\text{fail}} \in \{0.2, 0.3\}$.
2) **Initial point**. We generate $x' \sim \mathcal{N}(0, I_n)$ and start from $x^1 = \frac{10x'}{\|x'\|}$ for r_1, $x^1 = \frac{x'}{\|x'\|}$ for r_2, r_3.
3) **Stopping criterion**. We run algorithms for 400 epochs ($K = 400m$). Algorithms stop if $f \leq 1.2f(\hat{x})$.
4) **Stepsize**. We let $\gamma_k = \theta\cdot\sqrt{K}$ for vanilla algorithms; $\gamma_k = \theta\cdot\mathcal{G}(\|x^k\|)\sqrt{K}$ for robust stepsize with known growth condition; $\gamma_k = \theta\cdot\max\{\mathsf{Lip}(x^k,\xi'),\alpha\}\sqrt{K}$ for robust stepsize with unknown growth condition. $\theta \in [10^{-2}, 10^1]$ serves as a hyper-parameter.
5) **Clipping**. Clipping parameter α is set to 1.0.
6) **Mirror descent**. For experiments on mirror descent, we use kernels for r_1, r_2 from (Davis et al., 2018a). We leave the detailed kernel choices to **Appendix A**.

6.2. Comparing Different Stepsizes

Figure 2, 3 and **4** investigate the number of iterations for each stepsize to converge under different choices of θ. As the experiments suggest, our robust choices tend to be conservative when the function exhibits low-order growth. However, when the function exhibits high-order growth, our robust stepsize tends to converge within a reasonable range of stepsizes. It is worth noticing that for problem r_2, `SGD` diverges for $\theta \sim 10^8$, while our proposed approaches work robustly. Moreover, we notice that our robust stepsize based on reference Lipschitz property never diverges in practice, although it is sometimes conservative on problems where function growth is mild (such as r_1).

6.3. Comparison with Mirror Descent

Last, we compare our proposed method with the commonly adopted mirror descent approach for non-Lipschitz problems. We test both mirror descent and our proposed `SGD`-based approaches. As **Figure 5** shows, mirror descent indeed often exhibits more stable performance compared to vanilla `SGD`. However, we see that our approaches still exhibit superior convergence performance. It is important to note that the comparison in terms of iteration counts does not fully capture the efficiency of our approach. Specifically, our method tackles a much simpler proximal subproblem compared to the more complex root-finding subproblem in mirror descent, which further demonstrates the advantage of our robust stepsize strategy.

7. Conclusions

We develop novel robust stepsize (regularization) strategies and show that for weakly convex objectives without Lipschitz continuity, stochastic model-based methods can still converge at the desirable $\mathcal{O}(1/\sqrt{K})$ rate with constant failure probability. To our knowledge, this is achieved under the least restrictive assumptions known to date. A promising direction for future research is the adaptation of our analyses to more sophisticated methods, such as momentum-based or adaptive gradient methods.

Acknowledgement and Disclosure of Funding

The authors are grateful to the Area Chairs and the anonymous reviewers for their constructive comments. This research is partially supported by the Major Program of National Natural Science Foundation of China (Grant 72394360, 72394364).

Impact Statement

This paper discusses stochastic weakly convex optimization without standard Lipschitz continuity assumptions. We develop methods that can benefit the optimization community and have no negative social impacts.

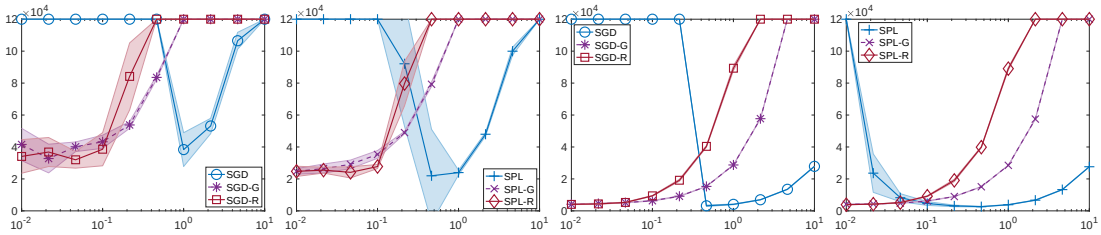

Figure 2: Problem r_1. Left two: $(\kappa, p_{\text{fail}}) = (10, 0.2)$; Right two: $(\kappa, p_{\text{fail}}) = (10, 0.3)$. x-axis: parameter θ; y-axis: number of iterations. `SGD` denotes vanilla SGD; `SGD-G` denotes SGD robust to known Lipschitzness; `SGD-R` denotes SGD robust to unknown Lipschitzness. The same applies to `SPL`.

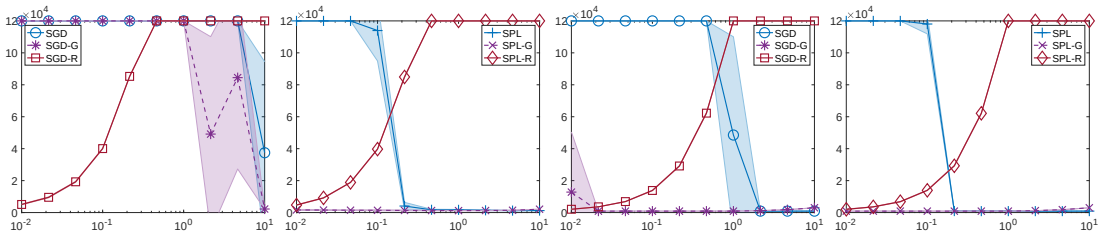

Figure 3: Problem r_2. Left two: $(\kappa, p_{\text{fail}}) = (1, 0.2)$; Right two: $(\kappa, p_{\text{fail}}) = (10, 0.3)$. x-axis: parameter θ; y-axis: number of iterations.

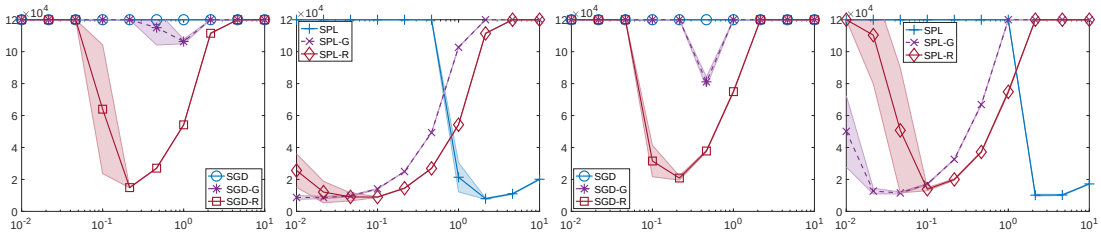

Figure 4: Problem r_3. Left two: $(\kappa, p_{\text{fail}}) = (1, 0.2)$; Right two: $(\kappa, p_{\text{fail}}) = (1, 0.3)$. x-axis: parameter θ; y-axis: number of iterations.

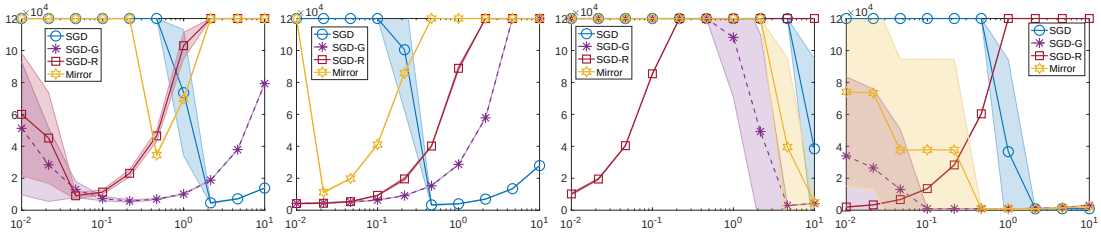

Figure 5: Left two: Problem r_1, $(\kappa, p_{\text{fail}}) = (1, 0.3)$; Right two: Problem r_2, $(\kappa, p_{\text{fail}}) = (1, 0.3)$. x-axis: parameter θ; y-axis: number of iterations.

References

Sigrún Andradóttir. A scaled stochastic approximation algorithm. *Management Science*, 42(4):475–498, 1996.

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

Karan Chadha, Gary Cheng, and John Duchi. Accelerated, optimal and parallel: Some results on model-based stochastic optimization. In *International Conference on Machine Learning*, pages 2811–2827. PMLR, 2022.

Vasileios Charisopoulos, Yudong Chen, Damek Davis, Mateo Díaz, Lijun Ding, and Dmitriy Drusvyatskiy. Low-rank matrix recovery with composite optimization: good

conditioning and rapid convergence. *Foundations of Computational Mathematics*, 21(6):1505–1593, 2021.

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

Damek Davis and Benjamin Grimmer. Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems. *SIAM Journal on Optimization*, 29(3): 1908–1930, 2019.

Damek Davis, Dmitriy Drusvyatskiy, and Kellie J MacPhee. Stochastic model-based minimization under high-order growth. *arXiv preprint arXiv:1807.00255*, 2018a.

Damek Davis, Dmitriy Drusvyatskiy, Kellie J MacPhee, and Courtney Paquette. Subgradient methods for sharp weakly convex functions. *Journal of Optimization Theory and Applications*, 179:962–982, 2018b.

Damek Davis, Dmitriy Drusvyatskiy, and Vasileios Charisopoulos. Stochastic algorithms with geometric step decay converge linearly on sharp functions. *arXiv preprint arXiv:1907.09547*, 2019.

Aaron Defazio and Konstantin Mishchenko. Learning-rate-free learning by d-adaptation. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 7449–7479. PMLR, 23–29 Jul 2023.

Qi Deng and Wenzhi Gao. Minibatch and momentum model-based methods for stochastic weakly convex optimization. *Advances in Neural Information Processing Systems*, 34:23115–23127, 2021.

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

John C Duchi and Feng Ruan. Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval. *Information and Inference: A Journal of the IMA*, 8(3):471–529, 2019.

Wenzhi Gao and Qi Deng. Delayed algorithms for distributed stochastic weakly convex optimization. *Advances in Neural Information Processing Systems*, 36, 2024.

Eduard Gorbunov, Marina Danilova, and Alexander Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. *Advances in Neural Information Processing Systems*, 33:15042–15053, 2020.

Benjamin Grimmer. Convergence rates for deterministic and stochastic subgradient methods without lipschitz continuity. *SIAM Journal on Optimization*, 29(2):1350–1365, 2019.

Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. *Cited on*, 14 (8):2, 2012.

Maor Ivgi, Oliver Hinder, and Yair Carmon. Dog is sgd's best friend: A parameter-free dynamic step size schedule. *arXiv preprint arXiv:2302.12022*, 2023.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Anastasia Koloskova, Hadrien Hendrikx, and Sebastian U Stich. Revisiting gradient clipping: Stochastic bias and tight convergence guarantees. *arXiv preprint arXiv:2305.01588*, 2023.

Haochuan Li, Jian Qian, Yi Tian, Alexander Rakhlin, and Ali Jadbabaie. Convex and non-convex optimization under generalized smoothness. *arXiv preprint arXiv:2306.01264*, 2023a.

Xiao Li, Lei Zhao, Daoli Zhu, and Anthony Man-Cho So. Revisiting subgradient method: Complexity and convergence beyond lipschitz continuity. *arXiv preprint arXiv:2305.14161*, 2023b.

Xiaoyu Li and Francesco Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. In *The 22nd international conference on artificial intelligence and statistics*, pages 983–992. PMLR, 2019.

Haihao Lu. "relative continuity" for non-lipschitz nonsmooth convex optimization using stochastic (or deterministic) mirror descent. *INFORMS Journal on Optimization*, 1(4):288–303, 2019.

Vien Mai and Mikael Johansson. Convergence of a stochastic gradient method with momentum for non-smooth non-convex optimization. In *International conference on machine learning*, pages 6630–6639. PMLR, 2020.

Vien V Mai and Mikael Johansson. Stability and convergence of stochastic gradient clipping: Beyond lipschitz continuity and smoothness. In *International Conference on Machine Learning*, pages 7325–7335. PMLR, 2021.

Yura Malitsky and Konstantin Mishchenko. Adaptive proximal gradient method for convex optimization. *arXiv preprint arXiv:2308.02261*, 2023.

Herbert Robbins and David Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing methods in statistics*, pages 233–257. Elsevier, 1971.

Fabian Schaipp, Robert M. Gower, and Michael Ulbrich. A stochastic proximal polyak step size. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=jWr41htaB3. Reproducibility Certification.

Qiuhao Wang, Chin Pang Ho, and Marek Petrik. Policy gradient in robust mdps with global convergence guarantee. In *International Conference on Machine Learning*, pages 35763–35797. PMLR, 2023.

Chenghan Xie, Chenxi Li, Chuwen Zhang, Qi Deng, Dongdong Ge, and Yinyu Ye. Trust region methods for nonconvex stochastic optimization beyond lipschitz smoothness. *arXiv preprint arXiv:2310.17319*, 2023.

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.

Siqi Zhang and Niao He. On the convergence rate of stochastic mirror descent for nonsmooth nonconvex optimization. *arXiv preprint arXiv:1806.04781*, 2018.

Daoli Zhu, Lei Zhao, and Shuzhong Zhang. A unified analysis for the subgradient methods minimizing composite nonconvex, nonsmooth and non-lipschitz functions. *arXiv preprint arXiv:2308.16362*, 2023.

Appendix

Table of Contents

Structure of the appendix The appendix is organized as follows. In **Section B**, **Section C** and **Section D**, we respectively prove the results for weakly convex optimization under different Lipschitz continuity assumptions. In **Section E** and its subsections, we extend our results to convex optimization.

A. Choice of Mirror Descent Kernel

For $r(x,a) = \langle a, x\rangle^2$, we have

$$\|f'(x,a)\| \le \|2\langle a,x\rangle a\| \le 2\|a\|^2\|x\|$$

and the Bregman divergence kernel is taken to be $d(x) = \frac{1}{2}\|x\|^2 + \frac{1}{4}\|x\|^4$.

For $r(x,a) = \langle a,x\rangle^5 + \langle a,x\rangle^3 + 1$, we have

$$|f'(x,a)| \le \|(5\langle a,x\rangle^4 + 3\langle a,x\rangle^2)a\| \le 5(\|a\|^5\|x\|^4 + \|a\|^3\|x\|^2)$$

and the divergence kernel is taken to be $d(x) = \frac{1}{2}\|x\|^2 + \frac{1}{6}\|x\|^6 + \frac{1}{8}\|x\|^8 + \frac{1}{10}\|x\|^{10}$. Note that to ensure strong convexity we always keep $\frac{1}{2}\|x\|^2$ in the kernel.

B. Proof of Results in Section 3

B.1. Proof of Lemma 3.1

By the optimality conditions of the proximal subproblems (4), we have, for any $\xi^k \sim \Xi$, that

$$f_{x^k}(x^{k+1},\xi^k) + \omega(x^{k+1}) + \frac{\gamma_k}{2}\|x^{k+1}-x^k\|^2 \le f_{x^k}(\hat{x}^k,\xi^k) + \omega(\hat{x}^k) + \frac{\gamma_k}{2}\|\hat{x}^k - x^k\|^2 - \frac{\gamma_k-\kappa}{2}\|x^{k+1}-\hat{x}^k\|^2$$

$$f(\hat{x}^k) + \omega(\hat{x}^k) + \frac{\rho}{2}\|\hat{x}^k - x^k\|^2 \le f(x^{k+1}) + \omega(x^{k+1}) + \frac{\rho}{2}\|x^{k+1}-x^k\|^2$$

Summing over the above two relations, we deduce that

$$\begin{aligned}&\frac{\gamma_k-\rho}{2}\|x^{k+1}-x^k\|^2 - \frac{\gamma_k-\rho}{2}\|\hat{x}^k-x^k\|^2 + \frac{\gamma_k-\kappa}{2}\|x^{k+1}-\hat{x}^k\|^2\\ &\le f(x^{k+1}) - f_{x^k}(x^{k+1},\xi^k) + f_{x^k}(\hat{x}^k,\xi^k) - f(\hat{x}^k)\end{aligned}$$

Conditioned on $\xi^1,\ldots,\xi^{k-1}$ and taking expectation with respect to ξ^k, we have

$$\begin{aligned}&\frac{\gamma_k-\rho}{2}\mathbb{E}_k[\|x^{k+1}-x^k\|^2] - \frac{\gamma_k-\rho}{2}\|\hat{x}^k-x^k\|^2 + \frac{\gamma_k-\kappa}{2}\mathbb{E}_k[\|x^{k+1}-\hat{x}^k\|^2]\\ &\le \mathbb{E}_k[f(x^{k+1}) - f_{x^k}(x^k,\xi^k)] + \mathbb{E}_k[L(\xi^k)\|x^{k+1}-x^k\|] + \frac{\tau}{2}\|\hat{x}^k-x^k\|^2 \qquad (10)\\ &= \mathbb{E}_k[f(x^{k+1})] - f(x^k) + \mathbb{E}_k[L(\xi^k)\|x^{k+1}-x^k\|] + \frac{\tau}{2}\|\hat{x}^k-x^k\|^2 \qquad (11)\\ &\le L_f\mathbb{E}_k[\|x^{k+1}-x^k\|] + \mathbb{E}_k[L(\xi^k)\|x^{k+1}-x^k\|] + \frac{\tau}{2}\|\hat{x}^k-x^k\|^2 \qquad (12)\end{aligned}$$

where (10) uses $L_f(\xi)$-Lipschitzness of $f_{x^k}(x,\xi)$; (11) uses quadratic bound from **A3** and (12) applies L_f-Lipschitzness of $f(x)$ (Davis and Drusvyatskiy, 2019). Re-arranging the terms, we have

$$\begin{aligned}&\frac{\gamma_k-\kappa}{2}\mathbb{E}_k[\|x^{k+1}-\hat{x}^k\|^2]\\ &\le \frac{\gamma_k-\rho+\tau}{2}\|\hat{x}^k-x^k\|^2 - \frac{\gamma_k-\rho}{2}\mathbb{E}_k[\|x^{k+1}-x^k\|^2] + \mathbb{E}_k[(L(\xi^k)+L_f)\|x^{k+1}-x^k\|]\\ &\le \frac{\gamma_k-\rho+\tau}{2}\|\hat{x}^k-x^k\|^2 + \frac{2L_f^2}{\gamma_k-\rho} \qquad (13)\\ &= \frac{\gamma_k-\kappa}{2}\|\hat{x}^k-x^k\|^2 - \frac{\rho-\tau-\kappa}{2}\|\hat{x}^k-x^k\|^2 + \frac{2L_f^2}{\gamma_k-\rho}\end{aligned}$$

where (13) uses the relation $-\frac{a}{2}x^2 + bx \le \frac{b^2}{2a}$ and $\mathbb{E}_\xi[(L(\xi)+L_f)^2] \le 4L_f^2$. Dividing both sides by $\frac{\gamma_k-\kappa}{2}$,

$$\mathbb{E}_k[\|x^{k+1}-\hat{x}^k\|^2] \le \|\hat{x}^k-x^k\|^2 - \frac{\rho-\tau-\kappa}{\gamma_k-\kappa}\|\hat{x}^k-x^k\|^2 + \frac{4L_f^2}{(\gamma_k-\rho)(\gamma_k-\kappa)}$$

and the potential function is reduced by

$$\begin{aligned}\mathbb{E}_k[\psi_{1/\rho}(x^{k+1})] &= \min_x\{f(x)+\omega(x)+\frac{\rho}{2}\|x-x^{k+1}\|^2\}\\ &\le f(\hat{x}^k)+\omega(\hat{x}^k)+\frac{\rho}{2}\|\hat{x}^k-x^{k+1}\|^2\\ &\le f(\hat{x}^k)+\omega(\hat{x}^k)+\frac{\rho}{2}\|\hat{x}^k-x^k\|^2-\frac{\rho(\rho-\tau-\kappa)}{2(\gamma_k-\kappa)}\|\hat{x}^k-x^k\|^2+\frac{2\rho L_f^2}{(\gamma_k-\rho)(\gamma_k-\kappa)}\\ &= \psi_{1/\rho}(x^k)-\frac{\rho(\rho-\tau-\kappa)}{2(\gamma_k-\kappa)}\|\hat{x}^k-x^k\|^2+\frac{2\rho L_f^2}{(\gamma_k-\rho)(\gamma_k-\kappa)},\end{aligned}$$

which completes the proof.

B.2. Proof of Theorem 3.1

Given fixed stepsize $\gamma_k \equiv \gamma = \rho+\kappa+\alpha\sqrt{K}$, where we have, after telescoping, that

$$\frac{\rho(\rho-\tau-\kappa)}{2(\gamma-\kappa)}\sum_{k=1}^K\mathbb{E}[\|\hat{x}^k-x^k\|^2] \le \psi_{1/\rho}(x^1)-\mathbb{E}[\psi_{1/\rho}(x^{K+1})]+\frac{2\rho L_f^2 K}{(\gamma-\rho)(\gamma-\kappa)}.$$

Re-arranging the terms and summing over $k=1,\ldots,K$, we have

$$\begin{aligned}\min_{1\le k\le K}\mathbb{E}[\|\nabla\psi_{1/\rho}(x^k)\|^2] &\le \frac{2\rho}{\rho-\tau-\kappa}\left[\frac{(\gamma-\kappa)D}{K}+\frac{2\rho L_f^2}{\gamma-\rho}\right]\\ &\le \frac{2\rho}{\rho-\tau-\kappa}\left[\frac{\rho D}{K}+\frac{\alpha D}{\sqrt{K}}+\frac{2\rho L_f^2}{\alpha\sqrt{K}}\right],\end{aligned}$$

where $D=\psi_{1/\rho}(x^1)-\inf_x\psi(x)\ge\psi_{1/\rho}(x^1)-\mathbb{E}[\psi_{1/\rho}(x^K)]$ and this completes the proof.

C. Proof of Results in Section 4

For brevity of notation, in the proof we define

$$\mathsf{G}_k := \mathcal{G}(\|x^k\|) \tag{14}$$

and use them interchangeably in this section. We also note that G_k is a random variable whose randomness comes from samples from previous iterations $\xi^1,\ldots,\xi^{k-1}$.

C.1. Proof of Lemma 4.1

Firstly, we still use optimality condition to get, for a given ξ^k, that

$$f_{x^k}(x^{k+1},\xi^k)+\omega(x^{k+1})+\frac{\gamma_k}{2}\|x^{k+1}-x^k\|^2 \le f_{x^k}(\hat{x}^k,\xi^k)+\omega(\hat{x}^k)+\frac{\gamma_k}{2}\|\hat{x}^k-x^k\|^2-\frac{\gamma_k-\kappa}{2}\|x^{k+1}-\hat{x}^k\|^2$$

$$f(\hat{x}^k)+\omega(\hat{x}^k)+\frac{\rho}{2}\|\hat{x}^k-x^k\|^2 \le f(x^k)+\omega(x^k)$$

Summing over the above two relations, we deduce that

$$\begin{aligned}&\frac{\gamma_k}{2}\|x^{k+1}-x^k\|^2-\frac{\gamma_k-\rho}{2}\|\hat{x}^k-x^k\|^2+\frac{\gamma_k-\kappa}{2}\|x^{k+1}-\hat{x}^k\|^2\\ &\le f(x^k)-f_{x^k}(x^{k+1},\xi^k)+f_{x^k}(\hat{x}^k,\xi^k)-f(\hat{x}^k)+L_\omega\|x^{k+1}-x^k\| \qquad (15)\\ &\le f(x^k)-f(x^k,\xi^k)+f_{x^k}(\hat{x}^k,\xi^k)-f(\hat{x}^k)+(\mathsf{G}_kL_f(\xi^k)+L_\omega)\|x^{k+1}-x^k\|, \qquad (16)\end{aligned}$$

where (15) applies L_ω-Lipschitz continuity of $\omega(x)$; (16) applies **C1**. Dividing both sides by $\frac{\gamma_k-\kappa}{2}$ and re-arranging the terms, we have

$$\begin{aligned}\|x^{k+1}-\hat{x}^k\|^2 \le{}& \frac{\gamma_k-\rho}{\gamma_k-\kappa}\|\hat{x}^k-x^k\|^2-\frac{\gamma_k}{\gamma_k-\kappa}\|x^{k+1}-x^k\|^2\\ &+\frac{2}{\gamma_k-\kappa}[f(x^k)-f(x^k,\xi^k)+f_{x^k}(\hat{x}^k,\xi^k)-f(\hat{x}^k)+(\mathsf{G}_kL_f(\xi^k)+L_\omega)\|x^{k+1}-x^k\|]\end{aligned}$$

Next conditioned on $\xi^1, \ldots, \xi^{k-1}$, taking expectation with respect to ξ^k, and recalling that $\mathcal{G}(\|x^k\|)$, and therefore γ_k is fixed given $\xi^1, \ldots, \xi^{k-1}$, we have

$$\begin{aligned}
&\mathbb{E}_k[\|x^{k+1} - \hat{x}^k\|^2] \\
&\leq \frac{\gamma_k - \rho + \tau}{\gamma_k - \kappa}\|\hat{x}^k - x^k\|^2 + \frac{2}{\gamma_k - \kappa}\mathbb{E}_k[(\mathsf{G}_k L_f(\xi^k) + L_\omega)\|x^{k+1} - x^k\| - \frac{\gamma_k}{2}\|x^{k+1} - x^k\|^2] \qquad (17)\\
&\leq \frac{\gamma_k - \rho + \tau}{\gamma_k - \kappa}\|\hat{x}^k - x^k\|^2 + \frac{(\mathsf{G}_k L_f + L_\omega)^2}{\gamma_k(\gamma_k - \kappa)} \qquad (18)\\
&= \|\hat{x}^k - x^k\|^2 - \frac{\rho - \tau - \kappa}{\gamma_k - \kappa}\|\hat{x}^k - x^k\|^2 + \frac{(\mathsf{G}_k L_f + L_\omega)^2}{\gamma_k(\gamma_k - \kappa)},
\end{aligned}$$

where (17) applies $f(x^k) - \mathbb{E}_k[f(x^k, \xi^k)] = 0, \mathbb{E}_k[f_{x^k}(\hat{x}^k, \xi^k)] - f(\hat{x}^k) \leq \frac{\tau}{2}\|x^k - \hat{x}^k\|^2$; (18) applies the relation $-\frac{a}{2}x^2 + bx \leq \frac{b^2}{2a}$ and $\mathbb{E}[L_f(\xi)]^2 \leq \mathbb{E}[L_f(\xi)^2] \leq L_f^2$. Now we can deduce reduction of the potential function by

$$\begin{aligned}
\mathbb{E}_k[\psi_{1/\rho}(x^{k+1})] &= \min_x\{f(x) + \omega(x) + \frac{\rho}{2}\|x - x^{k+1}\|^2\} \\
&\leq f(\hat{x}^k) + \omega(\hat{x}^k) + \frac{\rho}{2}\|\hat{x}^k - x^{k+1}\|^2 \\
&\leq f(\hat{x}^k) + \omega(\hat{x}^k) + \frac{\rho}{2}\|\hat{x}^k - x^k\|^2 - \frac{\rho(\rho - \tau - \kappa)}{2(\gamma_k - \kappa)}\|\hat{x}^k - x^k\|^2 + \frac{\rho(\mathsf{G}_k L_f + L_\omega)^2}{2\gamma_k(\gamma_k - \kappa)} \\
&= \psi_{1/\rho}(x^k) - \frac{\rho(\rho - \tau - \kappa)}{2(\gamma_k - \kappa)}\|\hat{x}^k - x^k\|^2 + \frac{\rho(\mathsf{G}_k L_f + L_\omega)^2}{2\gamma_k(\gamma_k - \kappa)}
\end{aligned}$$

and this completes the proof.

C.2. Proof of Theorem 4.1

First we introduce the following lemma.

Lemma C.1 (Robbins-Siegmund (Robbins and Siegmund, 1971)). *Let A_k, B_k, C_k and V_k be nonnegative random variables adapted to the filtration $\mathcal{F}_k$ and satisfying $\mathbb{E}[V_{k+1}|\mathcal{F}_k] \leq (1 + A_k)V_k + B_k - C_k$. Then on the event $\{\sum_{k=1}^\infty A_k < \infty, \sum_{k=1}^\infty B_k < \infty\}$, there is a random variable V_∞ such that $V_k \xrightarrow{a.s.} V_\infty$ and $\sum_{k=0}^\infty C_k < \infty$ almost surely.*

Now we get down to the proof. Recall that in Lemma 4.1 we have shown that

$$\mathbb{E}_k[\psi_{1/\rho}(x^{k+1})] \leq \psi_{1/\rho}(x^k) - \frac{\rho(\rho - \tau - \kappa)}{2(\gamma_k - \kappa)}\|\hat{x}^k - x^k\|^2 + \frac{\rho(\mathsf{G}_k L_f + L_\omega)^2}{2\gamma_k(\gamma_k - \kappa)}$$

and we can bound

$$\frac{\mathsf{G}_k L_f + L_\omega}{\gamma_k - \kappa} = \frac{\mathsf{G}_k L_f + L_\omega}{\rho + \tau + k^\zeta(\mathsf{G}_k + 1)} \leq \frac{\mathsf{G}_k L_f + L_\omega}{k^\zeta(\mathsf{G}_k + 1)} \leq \frac{L_f + L_\omega}{k^\zeta} \qquad (19)$$

to get

$$\mathbb{E}_k[\psi_{1/\rho}(x^{k+1}) + \Lambda] \leq [\psi_{1/\rho}(x^k) + \Lambda] - \frac{\rho(\rho - \kappa - \tau)}{2(\gamma_k - \kappa)}\|\hat{x}^k - x^k\|^2 + \frac{\rho}{2k^{2\zeta}}(L_f + L_\omega)^2$$

Then we invoke Lemma C.1, plugging in the relation

$$A_k = 0, \quad B_k = \frac{\rho(L_f + L_\omega)^2}{2k^{2\zeta}}, \quad C_k = \frac{\rho(\rho - \kappa - \tau)}{2(\gamma_k - \kappa)}\|\hat{x}^k - x^k\|^2, \quad V_k = \psi_{1/\rho}(x^k) + \Lambda \geq 0 \qquad (20)$$

Then with $\zeta \in (\frac{1}{2}, 1)$ $\sum_{k=1}^\infty B_k = \frac{\rho(L_f + L_\omega)^2}{2k^{2\zeta}} < \infty$ we know that $\{\psi_{1/\rho}(x^k) + \Lambda\} \to \psi_{1/\rho}(x^\infty) + \Lambda < \infty$ and that $\sum_{k=1}^\infty \frac{\rho(\rho - \kappa - \tau)}{2(\gamma_k - \kappa)}\|\hat{x}^k - x^k\|^2 < \infty$. By **A5**, $\|x^k\|$ is bounded with probability 1 and G_k is bounded almost surely. Finally $\sum_{k=1}^\infty \frac{1}{\gamma_k - \kappa} = \infty \Rightarrow \inf_{j \leq k}\|\hat{x}^j - x^j\| \to 0$ almost surely and this completes the proof since $\|\hat{x}^k - x^k\| = \rho^{-1}\|\nabla\psi_{1/\rho}(x^k)\|$.

C.3. Proof of Lemma 4.2 and 4.3

Following (19), we first we bound the error of potential reduction by $\frac{\rho(\mathsf{G}_k L_f + L_\omega)^2}{2\gamma_k(\gamma_k - \kappa)} \leq \frac{\rho(L_f + L_\omega)^2}{\alpha^2 K}$.

Then a telescopic sum gives, for all $2 \leq k \leq K$ that

$$\mathbb{E}[\psi_{1/\rho}(x^k) + \Lambda] \leq \psi_{1/\rho}(x^1) + \Lambda + \frac{\rho(L_f + L_\omega)^2}{\alpha^2} =: \Delta.$$

Here Δ is a constant that only depends on the initialization of the algorithm. Next we consider one step of the algorithm

$$f_{x^k}(x^{k+1}, \xi^k) + \omega(x^{k+1}) + \frac{\gamma_k}{2}\|x^{k+1} - x^k\|^2 \leq f_{x^k}(x^k, \xi^k) + \omega(x^k)$$

and a re-arrangement gives

$$\begin{aligned}\frac{\gamma_k}{2}\|x^{k+1} - x^k\|^2 &\leq f_{x^k}(x^k, \xi^k) - f_{x^k}(x^{k+1}, \xi^k) + \omega(x^k) - \omega(x^{k+1}) \\ &\leq (L_f(\xi^k)\mathsf{G}_k + L_\omega)\|x^{k+1} - x^k\|,\end{aligned}$$

where we use **C1** and Lipschitz continuity of $\omega(x)$. Dividing both sides by $\|x^{k+1} - x^k\|$, we have

$$\|x^{k+1} - x^k\| \leq \frac{2(L_f(\xi^k)\mathsf{G}_k + L_\omega)}{\gamma_k} \leq \frac{2(L_f(\xi^k)\mathsf{G}_k + L_\omega)}{\alpha(\mathsf{G}_k + 1)\sqrt{K}}.$$

Conditioned on $\xi^1, \ldots, \xi^{k-1}$ and taking expectation with respect to ξ^k, we get

$$\mathbb{E}_k[\|x^{k+1} - x^k\|] \leq \frac{2\mathbb{E}_{\xi^k}[L_f(\xi^k)]\mathsf{G}_k + 2L_\omega}{\alpha(\mathsf{G}_k + 1)\sqrt{K}} \leq \frac{2L_f\mathsf{G}_k + 2L_\omega}{\alpha(\mathsf{G}_k + 1)\sqrt{K}} \leq \frac{2(L_f + L_\omega)}{\alpha\sqrt{K}}.$$

This completes the proof of Lemma 4.2. By Markov's inequality, we know that, for any $2 \leq k \leq K$, the following bound holds

$$\mathbb{P}_{\xi^k \sim \Xi}\left\{\|x^{k+1} - x^k\| \leq \frac{4(L_f + L_\omega)}{\alpha\sqrt{K}} | \xi^1, \ldots, \xi^{k-1}\right\} \geq \frac{1}{2}$$

and without loss of generality we let $\mathsf{Z} = \frac{4(L_f + L_\omega)}{\alpha\sqrt{K}}$, and clearly $\mathsf{Z} = \mathcal{O}(1/\sqrt{K}) = \mathcal{O}(1)$.

This relation says "it's likely that x^k and x^{k+1} are close", and we leverage this intuition to derive a tail-bound on $\|x^k\|$. So far, we have the following properties in hand:

1. $\mathbb{E}[\psi_{1/\rho}(x^k)]$ is bounded by a constant $\Delta - \Lambda$ for all k
2. If $\|x^k\| \geq \mathsf{B}_v$, then $\psi_{1/\rho}(x^k) \geq v$
3. If $\|x^k\| \geq \mathsf{B}_v$, it's likely that $\|x^{k+1}\| \geq \mathsf{B}_v - \mathcal{O}(1/\sqrt{K})$.

Our reasoning is as follows: given large $a > 1$, conditioned on $\|x^k\| \geq \mathsf{B}_{a\Delta}$, then it is likely that $\|x^{k+1}\| \approx \mathsf{B}_{a\Delta}$ since $\|x^{k+1} - x^k\|$ is likely to be small. And it implies $\mathbb{E}[\psi_{1/\rho}(x^{k+1})] \geq a\Delta > \Delta$. However, we know that $\mathbb{E}[\psi_{1/\rho}(x^{k+1})] \leq \Delta$, and this will therefore reversely bound the probability that $\|x^k\| \geq \mathsf{B}_{a\Delta} + \mathcal{O}(1/\sqrt{K})$. We formalize the proof as follows.

First recall that by **A5**, $\|x^k\| \geq \mathsf{B}_v$ implies $\psi_{1/\rho}(x^k) \geq v$. Taking $v = a\Delta, a > 1$, we have $\|x^k\| \geq \mathsf{B}_{a\Delta} \Rightarrow \psi_{1/\rho}(x^k) \geq av$. Now we consider the event $\|x^k\| \geq \mathsf{B}_{a\Delta} + \mathsf{Z}$ and apply law of total expectation to get

$$\begin{aligned}\Delta &\geq \mathbb{E}[\psi_{1/\rho}(x^{k+1}) + \Lambda] \\ &= \mathbb{E}[\psi_{1/\rho}(x^{k+1}) + \Lambda | \|x^k\| \geq \mathsf{B}_{a\Delta} + \mathsf{Z}] \cdot \mathbb{P}\{\|x^k\| \geq \mathsf{B}_{a\Delta} + \mathsf{Z}\} \\ &\quad + \mathbb{E}[\psi_{1/\rho}(x^{k+1}) + \Lambda | \|x^k\| \leq \mathsf{B}_{a\Delta} + \mathsf{Z}] \cdot \mathbb{P}\{\|x^k\| \leq \mathsf{B}_{a\Delta} + \mathsf{Z}\} \\ &\geq \mathbb{E}[\psi_{1/\rho}(x^{k+1}) + \Lambda | \|x^k\| \geq \mathsf{B}_{a\Delta} + \mathsf{Z}] \cdot \mathbb{P}\{\|x^k\| \geq \mathsf{B}_{a\Delta} + \mathsf{Z}\}, \end{aligned} \tag{21}$$

where (21) uses $\psi_{1/\rho}(x) + \Lambda \geq 0$ for all x. Next we consider the expectation

$$\mathbb{E}\big[\psi_{1/\rho}(x^{k+1}) + \Lambda | \|x^k\| \geq \mathsf{B}_{a\Delta} + \mathsf{Z}\big], \tag{22}$$

and successively deduce that

$$\begin{aligned}
&\mathbb{E}[\psi_{1/\rho}(x^{k+1}) + \Lambda | \|x^k\| \geq \mathsf{B}_{a\Delta} + \mathsf{Z}] \\
&= \mathbb{E}[\psi_{1/\rho}(x^{k+1}) + \Lambda | \|x^k\| \geq \mathsf{B}_{a\Delta} + \mathsf{Z}, \|x^{k+1} - x^k\| \leq \mathsf{Z}] \cdot \mathbb{P}\{\|x^{k+1} - x^k\| \leq \mathsf{Z}\} \\
&\quad + \mathbb{E}[\psi_{1/\rho}(x^{k+1}) + \Lambda | \|x^k\| \geq \mathsf{B}_{a\Delta} + \mathsf{Z}, \|x^{k+1} - x^k\| \geq \mathsf{Z}] \cdot \mathbb{P}\{\|x^{k+1} - x^k\| \geq \mathsf{Z}\} \\
&\geq \frac{1}{2}\mathbb{E}[\psi_{1/\rho}(x^{k+1}) + \Lambda | \|x^k\| \geq \mathsf{B}_{a\Delta} + \mathsf{Z}, \|x^{k+1} - x^k\| \leq \mathsf{Z}] \qquad (23) \\
&\geq \frac{1}{2}(a\Delta + \Lambda), \qquad (24)
\end{aligned}$$

where (23) is by Markov's inequality $\mathbb{P}\{\|x^{k+1} - x^k\| \leq \mathsf{Z}\} \geq 0.5$ and (24) uses the triangle inequality

$$\|x^{k+1}\| = \|x^{k+1} - x^k + x^k\| \geq \|x^k\| - \|x^{k+1} - x^k\| \geq \|x^k\| - \mathsf{Z} \geq \mathsf{B}_{a\Delta}$$

and we recall that $\|x^{k+1}\| \geq \mathsf{B}_{a\Delta}$ implies $\psi_{1/\rho}(x^{k+1}) \geq a\Delta$. Chaining the above inequalities, we arrive at

$$\Delta \geq \left(\tfrac{a\Delta+\Lambda}{2}\right) \cdot \mathbb{P}\{\|x^k\| \geq \mathsf{B}_{a\Delta} + \mathsf{Z}\}$$

Dividing both sides by $\frac{a\Delta+\Lambda}{2}$ gives the desired tail bound

$$\mathbb{P}\Big\{\|x^k\| \geq \mathsf{B}_{a\Delta} + \mathsf{Z}\Big\} \leq \frac{2\Delta}{a\Delta + \Lambda}.$$

Since $\Lambda \geq 0$, the bound is nontrivial when $a > 2$, and this completes the proof.

C.4. Proof of Theorem 4.2

Given $\delta \in (0, 1/4)$, we know, from Lemma 4.3 that for $2 \leq k \leq K$,

$$\mathbb{P}\left\{\|x^k\| \geq \mathsf{B}_{\delta^{-1}\Delta} + \mathsf{Z}\right\} \leq \frac{2\Delta}{\delta^{-1}\Delta + \Lambda} \leq 2\delta.$$

Then denote $I_k = \mathbb{I}\big\{\|x^k\| \leq \mathsf{B}_{\delta^{-1}\Delta} + \mathsf{Z}\big\}$, and we have $\sum_{k=1}^K \mathbb{E}[1 - I_k] \leq 2\delta K$. By Markov's inequality, given $p \in (2\delta, 1)$,

$$\mathbb{P}\left\{\sum_{k=1}^K (1 - I_k) \geq \frac{2\delta K}{p}\right\} \leq \frac{2\delta K}{2p^{-1}\delta K} = p$$

and a re-arrangement gives $\mathbb{P}\big\{\sum_{k=1}^K I_k \geq K(1 - 2p^{-1}\delta)\big\} \geq 1 - p$. Now we telescope Lemma 4.1 and get

$$\begin{aligned}
\sum_{k=1}^K \mathbb{E}\left[\frac{\rho(\rho - \kappa - \tau)}{2\gamma_k}\|\hat{x}^k - x^k\|^2\right] &\leq \psi_{1/\rho}(x^1) - \mathbb{E}[\psi_{1/\rho}(x^{K+1})] + \frac{\rho}{2\alpha^2}(L_f + L_\omega)^2 \qquad (25) \\
&\leq \psi_{1/\rho}(x^1) - \inf_x \psi(x) + \frac{\rho}{2\alpha^2}(L_f + L_\omega)^2, \qquad (26) \\
&= D + \frac{\rho}{2\alpha^2}(L_f + L_\omega)^2
\end{aligned}$$

where (25) uses the previously established bound $\mathbb{E}_k\left[\frac{\rho(\mathsf{G}_k L_f + L_\omega)^2}{2\gamma_k(\gamma_k - \kappa)}\right] \leq \frac{\rho(L_f + L_\omega)^2}{\alpha^2 K}$. (26) uses the fact that $\psi_{1/\rho}(x^{K+1}) \geq \inf_x \psi(x)$. Next we notice, conditioned on the event $\sum_{k=1}^K I_k \geq K(1 - 2p^{-1}\delta)$ and these iterations (the set $\{j : I_j = 1\}$),

defining $\mathsf{G}_\delta := \max_z \mathcal{G}(z), z \le \mathsf{B}_{\delta^{-1}\Delta} + \mathsf{Z}$, that

$$\begin{aligned}
&\sum_{k=1}^{K} \mathbb{E}\left[\frac{\rho(\rho-\kappa-\tau)}{2\gamma_k}\|\hat{x}^k - x^k\|^2\right] \\
&= \sum_{k\in\{j:I_j=0\}} \mathbb{E}\left[\frac{\rho(\rho-\kappa-\tau)}{2\gamma_k}\|\hat{x}^k - x^k\|^2\right] + \sum_{k\in\{j:I_j=1\}} \mathbb{E}\left[\frac{\rho(\rho-\kappa-\tau)}{2\gamma_k}\|\hat{x}^k - x^k\|^2\right] \\
&\ge \sum_{k\in\{j:I_j=1\}} \mathbb{E}\left[\frac{\rho(\rho-\kappa-\tau)}{2\gamma_k}\|\hat{x}^k - x^k\|^2\right] \\
&\ge \sum_{k\in\{k:I_j=1\}} \frac{\rho(\rho-\kappa-\tau)}{2(\rho+\kappa+\tau+\alpha(\mathsf{G}_\delta+1)\sqrt{K})}\mathbb{E}[\|\hat{x}^k - x^k\|^2] \qquad (27)\\
&\ge \frac{\rho(\rho-\kappa-\tau)(1-2p^{-1}\delta)K}{2(\rho+\kappa+\tau+\alpha(\mathsf{G}_\delta+1)\sqrt{K})} \min_{k\in\{j:I_j=1\}} \mathbb{E}[\|\hat{x}^k - x^k\|^2], \qquad (28)
\end{aligned}$$

where (27) applies the fact that if $I_j = 1$, $\|x^j\| \le \mathsf{G}_\delta$; (28) utilizes the fact that $|\{k : I_j = 1\}| \ge K(1 - 2p^{-1}\delta)$. Putting the inequalities together, we have

$$\begin{aligned}
\min_{k\in\{j:I_j=1\}} \mathbb{E}[\|\hat{x}^k - x^k\|^2] &\le \frac{2(\rho+\kappa+\tau+\alpha(\mathsf{G}_\delta+1)\sqrt{K})}{\rho(\rho-\kappa-\tau)(1-2p^{-1}\delta)K}\left[D + \frac{\rho(L_f+L_\omega)^2}{2\alpha^2}\right] \\
&= \frac{2(D + \frac{\rho(L_f+L_\omega)^2}{2\alpha^2})}{\rho(\rho-\kappa-\tau)(1-2p^{-1}\delta)}\left[\frac{\rho+\kappa+\tau}{K} + \frac{\alpha(\mathsf{G}_\delta+1)}{\sqrt{K}}\right],
\end{aligned}$$

Recalling that $\|\hat{x}^k - x^k\| = \rho^{-1}\|\nabla\psi_{1/\rho}(x^k)\|$, at least with probability $1-p$,

$$\begin{aligned}
\min_{1\le k\le K} \mathbb{E}[\|\nabla\psi_{1/\rho}(x^k)\|^2] &\le \min_{k\in\{k:I_k=1\}} \mathbb{E}[\|\nabla\psi_{1/\rho}(x^k)\|^2] \\
&\le \frac{p}{p-2\delta}\cdot\frac{2\rho}{\rho-\tau-\kappa}\left[D + \frac{\rho(L_f+L_\omega)^2}{2\alpha^2}\right]\left(\frac{\rho+\tau+\kappa}{K} + \frac{\alpha(\mathsf{G}_\delta+1)}{\sqrt{K}}\right)
\end{aligned}$$

and this completes the proof.

D. Proof of Results in Section 5

For brevity of expression, we define $L_f^k := \mathsf{Lip}(x^k, \xi^k)$, $L_f' := \mathsf{Lip}(x^k, \xi')$ (k is hidden when clear from the context) and use them interchangeably.

D.1. Auxiliary Lemmas

Lemma D.1. *Given independent nonnegative random variables X and Y. If $\mathbb{E}_{X,Y}[|X-Y|^2] \le \sigma^2$, then*

$$\mathbb{E}_{X,Y}\left[\frac{X^2}{\max\{Y^2,\alpha^2\}}\right] \le \left(\tfrac{\sigma+\alpha}{\alpha}\right)^2, \quad \mathbb{E}_{X,Y}\left[\frac{X}{\max\{Y^2,\alpha^2\}}\right] \le \tfrac{\sigma}{\alpha^2} + \tfrac{1}{\alpha}, \quad \mathbb{E}_{X,Y}\left[\frac{X}{\max\{Y,\alpha\}}\right] \le \tfrac{\sigma}{\alpha} + 1$$

where $\alpha > 0$.

Proof. For the first relation, we successively deduce that

$$\begin{aligned}
\mathbb{E}_{X,Y}\left[\frac{X^2}{\max\{Y^2,\alpha^2\}}\right] &= \mathbb{E}_{X,Y}\left[\frac{(X-Y+Y)^2}{\max\{Y^2,\alpha^2\}}\right] \\
&= \mathbb{E}_{X,Y}\left[\frac{(X-Y)^2+Y^2+2Y(X-Y)}{\max\{Y^2,\alpha^2\}}\right] \\
&= \mathbb{E}_{X,Y}\left[\frac{(X-Y)^2}{\max\{Y^2,\alpha^2\}}\right] + \mathbb{E}_{Y}\left[\frac{Y^2}{\max\{Y^2,\alpha^2\}}\right] + \mathbb{E}_{X,Y}\left[\frac{2Y(X-Y)}{\max\{Y^2,\alpha^2\}}\right] \\
&\le \tfrac{\sigma^2}{\alpha^2} + 1 + \mathbb{E}_{X,Y}\left[\frac{2Y|X-Y|}{\max\{Y,\alpha\}\cdot\max\{Y,\alpha\}}\right] \qquad (29)\\
&\le \tfrac{\sigma^2}{\alpha^2} + \tfrac{2\sigma}{\alpha} + 1 = \left(\tfrac{\sigma+\alpha}{\alpha}\right)^2,
\end{aligned}$$

where (29) uses $\frac{a}{\max\{b,c\}} \leq \frac{a}{c}$ and the last inequality applies

$$\frac{2Y|X-Y|}{\max\{Y,\alpha\}\cdot\max\{Y,\alpha\}} = \frac{2Y}{\max\{Y,\alpha\}} \cdot \frac{|X-Y|}{\max\{Y,\alpha\}} \leq 2 \cdot \frac{|X-Y|}{\alpha}. \tag{30}$$

Similarly we can deduce that

$$\mathbb{E}_{X,Y}\left[\frac{X}{\max\{Y^2,\alpha^2\}}\right] \leq \mathbb{E}_{X,Y}\left[\frac{|X-Y|}{\max\{Y^2,\alpha^2\}}\right] + \mathbb{E}_Y\left[\frac{Y}{\max\{Y^2,\alpha^2\}}\right] \leq \frac{\sigma}{\alpha^2} + \frac{1}{\alpha},$$

$$\mathbb{E}_{X,Y}\left[\frac{X}{\max\{Y,\alpha\}}\right] \leq \mathbb{E}_{X,Y}\left[\frac{|X-Y|}{\max\{Y,\alpha\}}\right] + \mathbb{E}_{X,Y}\left[\frac{Y}{\max\{Y,\alpha\}}\right] \leq \frac{\sigma}{\alpha} + 1,$$

which completes the proof. □

D.2. Proof of Lemma 5.1

By the optimality condition, we have

$$f_{x^k}(x^{k+1},\xi^k) + \omega(x^{k+1}) + \frac{\gamma_k}{2}\|x^{k+1}-x^k\|^2 \leq f_{x^k}(\hat{x}^k,\xi^k) + \omega(\hat{x}^k) + \frac{\gamma_k}{2}\|\hat{x}^k-x^k\|^2 - \frac{\gamma_k-\kappa}{2}\|x^{k+1}-\hat{x}^k\|^2$$

$$f(\hat{x}^k) + \omega(\hat{x}^k) + \frac{\rho}{2}\|\hat{x}^k-x^k\|^2 \leq f(x^k) + \omega(x^k)$$

and summation over the two relations gives

$$\begin{aligned}
&\frac{\gamma_k}{2}\|x^{k+1}-x^k\|^2 - \frac{\gamma_k-\rho}{2}\|\hat{x}^k-x^k\|^2 + \frac{\gamma_k-\kappa}{2}\|x^{k+1}-\hat{x}^k\|^2 \\
&\leq f(x^k) - f_{x^k}(x^{k+1},\xi^k) + f_{x^k}(\hat{x}^k,\xi^k) - f(\hat{x}^k) + L_\omega\|x^{k+1}-x^k\| \\
&\leq f(x^k) - f(x^k,\xi^k) + f_{x^k}(\hat{x}^k,\xi^k) - f(\hat{x}^k) + (L_f^k + L_\omega)\|x^{k+1}-x^k\|, \qquad (31)
\end{aligned}$$

where (31) applies **D1**. Fixing ξ', we divide both sides by $\frac{\gamma_k-\kappa}{2}$ and get

$$\begin{aligned}
\|x^{k+1}-\hat{x}^k\|^2 \leq{}& \frac{\gamma_k-\rho}{\gamma_k-\kappa}\|\hat{x}^k-x^k\|^2 - \frac{\gamma_k}{\gamma_k-\kappa}\|x^{k+1}-x^k\|^2 \\
&+ \frac{2}{\gamma_k-\kappa}[f(x^k) - f(x^k,\xi^k) + f_{x^k}(\hat{x}^k,\xi^k) - f(\hat{x}^k) + (L_f^k+L_\omega)\|x^{k+1}-x^k\|]
\end{aligned}$$

Conditioned on $\xi^1,\ldots,\xi^{k-1}$ and taking expectation with respect to ξ^k, we have

$$\begin{aligned}
&\mathbb{E}_k[\|x^{k+1}-\hat{x}^k\|^2] \\
&\leq \frac{\gamma_k-\rho+\tau}{\gamma_k-\kappa}\|\hat{x}^k-x^k\|^2 + \frac{2}{\gamma_k-\kappa}\mathbb{E}_k[(L_f^k+L_\omega)\|x^{k+1}-x^k\| - \frac{\gamma_k}{2}\|x^{k+1}-x^k\|^2] \qquad (32) \\
&\leq \frac{\gamma_k-\rho+\tau}{\gamma_k-\kappa}\|\hat{x}^k-x^k\|^2 + \mathbb{E}_k\Big[\frac{1}{\gamma_k(\gamma_k-\kappa)}(L_f^k+L_\omega)^2\Big] \qquad (33)
\end{aligned}$$

where (32) again uses $f(x^k) - \mathbb{E}_k[f(x^k,\xi^k)] = 0, \mathbb{E}_k[f_{x^k}(\hat{x}^k,\xi^k)] - f(\hat{x}^k) \leq \frac{\tau}{2}\|x^k-\hat{x}^k\|^2$; (33) uses the relation $-\frac{a}{2}x^2 + bx \leq \frac{b^2}{2a}$. Now we have

$$\mathbb{E}_k[\|x^{k+1}-\hat{x}^k\|^2] \leq \|\hat{x}^k-x^k\|^2 - \frac{\rho-\tau-\kappa}{\gamma_k-\kappa}\|\hat{x}^k-x^k\|^2 + \mathbb{E}_k\Big[\frac{1}{\gamma_k(\gamma_k-\kappa)}(L_f^k+L_\omega)^2\Big].$$

In view of our potential function, we have

$$\begin{aligned}
\mathbb{E}_k[\psi_{1/\rho}(x^{k+1})] &= \min_x\{f(x) + \omega(x) + \frac{\rho}{2}\|x-x^{k+1}\|^2\} \\
&\leq f(\hat{x}^k) + \omega(\hat{x}^k) + \frac{\rho}{2}\|\hat{x}^k-x^{k+1}\|^2 \\
&\leq f(\hat{x}^k) + \omega(\hat{x}^k) + \frac{\rho}{2}\|\hat{x}^k-x^k\|^2 - \frac{\rho(\rho-\tau-\kappa)}{2(\gamma_k-\kappa)}\|\hat{x}^k-x^k\|^2 + \mathbb{E}_k\Big[\frac{\rho}{2\gamma_k(\gamma_k-\kappa)}(L_f^k+L_\omega)^2\Big] \\
&= \psi_{1/\rho}(x^k) - \frac{\rho(\rho-\tau-\kappa)}{2(\gamma_k-\kappa)}\|\hat{x}^k-x^k\|^2 + \mathbb{E}_k\Big[\frac{\rho}{2\gamma_k(\gamma_k-\kappa)}(L_f^k+L_\omega)^2\Big]
\end{aligned}$$

and this completes the proof.

D.3. Proof of Theorem 5.1

Our reasoning is the same as in Theorem 4.1, and we start by bounding $\mathbb{E}_{\xi'}\mathbb{E}_k\left[\frac{(L_f^k+L_\omega)^2}{\gamma_k(\gamma_k-\kappa)}\right]$.

For brevity we omit $\mathbb{E}_k[\cdot]$ and notice that, for $\gamma_k \geq 2\kappa$, that,

$$\frac{(L_f^k+L_\omega)^2}{\gamma_k(\gamma_k-\kappa)} \leq \frac{2(L_f^k+L_\omega)^2}{\gamma_k^2} = \frac{(L_f^k)^2}{\gamma_k^2} + \frac{2L_f^k L_\omega}{\gamma_k^2} + \frac{L_\omega^2}{\gamma_k^2}$$

so that we can bound the three terms respectively using

$$\mathbb{E}_{\xi'}\left[\frac{(L_f^k)^2}{\gamma_k^2}\right] \leq \mathbb{E}_{\xi'}\left[\frac{(L_f^k)^2}{\max\{(L_f')^2,\alpha^2\}k^{2\zeta}}\right] \leq \left(\frac{\alpha+\sigma}{\alpha}\right)^2\frac{1}{k^{2\zeta}}$$

where we invoked Lemma D.1 and take $X = L_f^k, Y = L_f'$ and we recall that by reference Lipschitz property **D2**: $\mathbb{E}[|L_f^k - L_f'|^2] \leq \sigma^2$. Similarly, we can deduce that

$$\mathbb{E}_{\xi'}\left[\frac{2L_f^k L_\omega}{\gamma_k^2}\right] \leq \mathbb{E}_{\xi'}\left[\frac{2L_f^k L_\omega}{\max\{\alpha L_f',\alpha^2\}k^{2\zeta}}\right] \leq \left(\frac{\alpha+\sigma}{\alpha^2}\right)\frac{2L_\omega}{k^{2\zeta}}$$

and $\frac{L_\omega^2}{\gamma_k^2} \leq \frac{L_\omega^2}{\alpha^2 k^{2\zeta}}$ since $\gamma_k \geq \alpha k^\zeta$. Putting the things together, we have

$$\mathbb{E}_{\xi'}\left[\frac{(L_f^k+L_\omega)^2}{\gamma_k^2}\right] \leq \frac{(\alpha+\sigma)^2 + 2L_\omega(\alpha+\sigma) + L_\omega^2}{\alpha^2 k^{2\zeta}} = \frac{(\alpha+\sigma+L_\omega)^2}{\alpha^2 k^{2\zeta}}.$$

and

$$\mathbb{E}_k[\psi_{1/\rho}(x^{k+1})] \leq \psi_{1/\rho}(x^k) - \mathbb{E}_{\xi'}\left[\frac{\rho(\rho-\tau-\kappa)}{2(\gamma_k-\kappa)}\right]\|\hat{x}^k - x^k\|^2 + \frac{\rho(\alpha+\sigma+L_\omega)^2}{\alpha^2 k^{2\zeta}},$$

or

$$\mathbb{E}_k[\psi_{1/\rho}(x^{k+1}) + \Lambda] \leq [\psi_{1/\rho}(x^k) + \Lambda] - \mathbb{E}_{\xi'}\left[\frac{\rho(\rho-\tau-\kappa)}{2(\gamma_k-\kappa)}\right]\|\hat{x}^k - x^k\|^2 + \frac{\rho(\alpha+\sigma+L_\omega)^2}{\alpha^2 k^{2\zeta}}.$$

Invoking Lemma C.1, plugging in the relation $A_k = 0, V_k = \psi_{1/\rho}(x^k) + \Lambda \geq 0$, $\mathsf{B}_k = \frac{\rho(\alpha+\sigma+L_\omega)^2}{\alpha^2 k^{2\zeta}}$ and $C_k = \mathbb{E}_{\xi'}\left[\frac{\rho(\rho-\tau-\kappa)}{\gamma_k-\kappa}\right]\|\hat{x}^k - x^k\|^2$. Note that since $\mathbb{E}_{\xi'}\left[\frac{\rho(\rho-\tau-\kappa)}{\gamma_k-\kappa}\right]$ is determined by x^k, we can view $C_k = g(\|x^k\|)\|\hat{x}^k - x^k\|^2$ for some function g and the rest of reasoning is the same as in Lemma 4.1.

D.4. Proof of Lemma 5.2

Similar to Lemma 4.3 we first bound the error of potential reduction $\mathbb{E}_k\left[\frac{\rho(L_f^k+L_\omega)^2}{2\gamma_k(\gamma_k-\kappa)}\right]$, and according to the proof of Lemma 5.1,

$$\mathbb{E}_{\xi'}\mathbb{E}_k\left[\frac{\rho(L_f^k+L_\omega)^2}{2\gamma_k(\gamma_k-\kappa)}\right] \leq \frac{\rho(\alpha+\sigma+L_\omega)^2}{\alpha^2 K}.$$

Telescoping the relation $\mathbb{E}[\psi_{1/\rho}(x^{k+1})] \leq \psi_{1/\rho}(x^k) + \frac{\rho(\alpha+\sigma+L_\omega)^2}{2\alpha^2 K}$ gives

$$\mathbb{E}[\psi_{1/\rho}(x^k) + \Lambda] \leq \psi_{1/\rho}(x^1) + \Lambda + \frac{\rho(\alpha+\sigma+L_\omega)^2}{\alpha^2} =: \Delta.$$

Next we show that $\mathbb{E}_k[\|x^{k+1} - x^k\|]$ is bounded. By the optimality condition we have

$$f_{x^k}(x^{k+1},\xi^k) + \omega(x^{k+1}) + \frac{\gamma_k}{2}\|x^{k+1} - x^k\|^2 \leq f_{x^k}(x^k,\xi^k) + \omega(x^k) - \frac{\gamma_k-\kappa}{2}\|x^{k+1} - x^k\|^2$$

and

$$\begin{aligned}\frac{\gamma_k}{2}\|x^{k+1} - x^k\|^2 &\leq f_{x^k}(x^k,\xi^k) - f_{x^k}(x^{k+1},\xi^k) + \omega(x^k) - \omega(x^{k+1}) \\ &\leq (L_f^k + L_\omega)\|x^{k+1} - x^k\|.\end{aligned}$$

Dividing both sides by $\|x^{k+1}-x^k\|$, we get $\|x^{k+1}-x^k\| \le \frac{2(L_f^k+L_\omega)}{\gamma_k} \le \frac{2L_f^k}{\gamma_k} + \frac{2L_\omega}{\alpha\sqrt{K}}$. Taking expectation on both sides, we have $\mathbb{E}_{\xi'}\mathbb{E}_k[L_f^k/\gamma_k] \le \frac{\alpha+\sigma}{\alpha\sqrt{K}}$, where we invoke Lemma D.1 with $X = L_f^k, Y = L_f'$ again. Now

$$\mathbb{E}_k[\|x^{k+1}-x^k\|] \le \frac{2(\alpha+\sigma+L_\omega)}{\alpha\sqrt{K}}.$$

The rest of the reasoning is consistent with Lemma 4.3 up to difference in constants. And we still present them for completeness. Applying Markov's inequality, we know that

$$\mathbb{P}_{\xi^k,\xi'\sim\Xi}\{\|x^{k+1}-x^k\| \le \tfrac{4(\alpha+\sigma+L_\omega)}{\alpha\sqrt{K}} | \xi_1,\dots,\xi_{k-1}\} \ge \frac{1}{2}.$$

By A5, $\|x^k\| \ge \mathsf{B}_v$ implies $\psi_{1/\rho}(x^k) \ge v$. Taking $v = a\Delta$, we have $\|x^k\| \ge \mathsf{B}_{a\Delta} \Rightarrow \psi_{1/\rho}(x^k) \ge av$. Without loss of generality, let $\mathsf{Z} = \frac{4(\alpha+\sigma+L_\omega)}{\alpha\sqrt{K}} > 0$, and we condition on the event $\|x^k\| \ge \mathsf{B}_{a\Delta}+\mathsf{Z}$ to deduce that

$$\begin{aligned}
\Delta &\ge \mathbb{E}[\psi_{1/\rho}(x^{k+1}) + \Lambda] \\
&= \mathbb{E}[\psi_{1/\rho}(x^{k+1}) + \Lambda | \|x^k\| \ge \mathsf{B}_{a\Delta}+\mathsf{Z}] \cdot \mathbb{P}\{\|x^k\| \ge \mathsf{B}_{a\Delta}+\mathsf{Z}\} \\
&\quad + \mathbb{E}[\psi_{1/\rho}(x^{k+1}) + \Lambda | \|x^k\| \le \mathsf{B}_{a\Delta}+\mathsf{Z}] \cdot \mathbb{P}\{\|x^k\| \le \mathsf{B}_{a\Delta}+\mathsf{Z}\} \\
&\ge \mathbb{E}[\psi_{1/\rho}(x^{k+1}) + \Lambda | \|x^k\| \ge \mathsf{B}_{a\Delta}+\mathsf{Z}] \cdot \mathbb{P}\{\|x^k\| \ge \mathsf{B}_{a\Delta}+\mathsf{Z}\}, \qquad (34)
\end{aligned}$$

where (34) uses $\psi_{1/\rho}(x)+\Lambda \ge 0$ for all x. Next we consider the expectation $\mathbb{E}[\psi_{1/\rho}(x^{k+1}) + \Lambda | \|x^k\| \ge \mathsf{B}_{a\Delta}+\mathsf{Z}]$ and we successively deduce that

$$\begin{aligned}
&\mathbb{E}[\psi_{1/\rho}(x^{k+1}) + \Lambda | \|x^k\| \ge \mathsf{B}_{a\Delta}+\mathsf{Z}] \\
&= \mathbb{E}[\psi_{1/\rho}(x^{k+1}) + \Lambda | \|x^k\| \ge \mathsf{B}_{a\Delta}+\mathsf{Z}, \|x^{k+1}-x^k\| \le \mathsf{Z}] \cdot \mathbb{P}\{\|x^{k+1}-x^k\| \le \mathsf{Z}\} \\
&\quad + \mathbb{E}[\psi_{1/\rho}(x^{k+1}) + \Lambda | \|x^k\| \ge \mathsf{B}_{a\Delta}+\mathsf{Z}, \|x^{k+1}-x^k\| \ge \mathsf{Z}] \cdot \mathbb{P}\{\|x^{k+1}-x^k\| \ge \mathsf{Z}\} \\
&\ge \Big(\frac{a\Delta+\Lambda}{2}\Big) \cdot \mathbb{P}\{\|x^{k+1}-x^k\| \le \mathsf{Z}\}, \qquad (35)
\end{aligned}$$

where (35) is by $\mathbb{P}\{\|x^{k+1}-x^k\| \le \mathsf{Z}\} \ge 0.5$ and that, conditioned on $\|x^{k+1}-x^k\| \le \mathsf{Z}$,

$$\|x^{k+1}\| = \|x^{k+1}-x^k+x^k\| \ge \|x^k\| - \|x^{k+1}-x^k\| \ge \|x^k\| - \mathsf{Z} \ge \mathsf{B}_{a\Delta}.$$

Chaining the above inequalities, we arrive at

$$\Delta \ge \Big(\frac{a\Delta+\Lambda}{2}\Big) \cdot \mathbb{P}\{\|x^k\| \ge \mathsf{B}_{a\Delta}+\mathsf{Z}\}$$

Dividing both sides by $\frac{a\Delta+\Lambda}{2}$ gives the desired tail bound.

D.5. Proof of Theorem 5.2

The proof again follows the clue of Theorem 4.2. Recall that $\mathsf{Z} = \frac{4(\alpha+\sigma+L_\omega)}{\alpha\sqrt{K}}$ and given $\delta \in (0, 1/4)$,

$$\mathbb{P}\{\|x^k\| \ge \mathsf{B}_{\delta^{-1}\Delta}+\mathsf{Z}\} \le \frac{2\Delta}{\delta^{-1}\Delta+\Lambda} \le 2\delta.$$

Denoting $I_k = \mathbb{I}\{\|x^k\| \le \mathsf{B}_{\delta^{-1}\Delta}+\mathsf{Z}\}$, we have $\sum_{k=1}^K \mathbb{E}[1-I_k] \le 2\delta K$ and by Markov's inequality, given $p \in (2\delta, 1)$, we have

$$\mathbb{P}\Big\{\sum_{k=1}^K (1-I_k) \ge \frac{2\delta K}{p}\Big\} \le \frac{2\delta K}{2p^{-1}\delta K} = p$$

and $\mathbb{P}\{\sum_{k=1}^K I_k \ge K(1-2p^{-1}\delta)\} \ge 1-p$. Now we telescope over Lemma 5.1 and get

$$\sum_{k=1}^{K}\mathbb{E}\left[\frac{\rho(\rho-\tau-\kappa)}{2\gamma_k}\|\hat{x}^k-x^k\|^2\right]\leq\psi_{1/\rho}(x^1)-\mathbb{E}[\psi_{1/\rho}(x^{K+1})]+\frac{\rho(\alpha+\sigma+L_\omega)^2}{2\alpha^2}$$
$$\leq\psi_{1/\rho}(x^1)-\inf_x\psi(x)+\frac{\rho}{2\alpha^2}(\alpha+\sigma+L_\omega)^2$$
$$=D+\frac{\rho}{2\alpha^2}(\alpha+\sigma+L_\omega)^2.$$

Next define

$$\mathsf{G}_\delta\coloneqq\max_x\sup_{\xi\sim\Xi}\mathsf{Lip}(x,\xi)\quad\text{subject to}\quad\|x\|\leq\mathsf{B}_{\delta^{-1}\Delta}+\mathsf{Z},$$

and we have, conditioned on the event $\sum_{k=1}^{K}I_k\geq K(1-2p^{-1}\delta)$,

$$\begin{aligned}
&\sum_{k=1}^{K}\mathbb{E}\left[\frac{\rho(\rho-\tau-\kappa)}{2\gamma_k}\|\hat{x}^k-x^k\|^2\right]\\
=&\sum_{k\in\{j:I_j=0\}}\mathbb{E}\left[\frac{\rho(\rho-\tau-\kappa)}{2\gamma_k}\|\hat{x}^k-x^k\|^2\right]+\sum_{k\in\{j:I_j=1\}}\mathbb{E}\left[\frac{\rho(\rho-\tau-\kappa)}{2\gamma_k}\|\hat{x}^k-x^k\|^2\right]\\
\geq&\sum_{k\in\{j:I_j=1\}}\mathbb{E}\left[\frac{\rho(\rho-\tau-\kappa)}{2\gamma_k}\|\hat{x}^k-x^k\|^2\right]\\
\geq&\sum_{k\in\{j:I_j=1\}}\mathbb{E}\left[\frac{\rho(\rho-\tau-\kappa)}{2(\rho+\kappa+\tau+(\alpha+\mathsf{G}_\delta)\sqrt{K})}\|\hat{x}^k-x^k\|^2\right]\\
=&\frac{\rho(\rho-\tau-\kappa)}{2(\rho+\kappa+\tau+(\alpha+\mathsf{G}_\delta)\sqrt{K})}\sum_{k\in\{j:I_j=1\}}\mathbb{E}[\|\hat{x}^k-x^k\|^2]\\
\geq&\frac{\rho(\rho-\tau-\kappa)(1-2p^{-1}\delta)K}{2(\rho+\kappa+\tau+(\alpha+\mathsf{G}_\delta)\sqrt{K})}\min_{k\in\{j:I_j=1\}}\mathbb{E}[\|\hat{x}^k-x^k\|^2]
\end{aligned}$$

Re-arranging the terms, we have, at least with probability $1-p$, that

$$\begin{aligned}
\min_{1\leq k\leq K}\mathbb{E}[\|\nabla\psi_{1/\rho}(x^k)\|^2]\leq&\min_{k\in\{k:I_k=1\}}\mathbb{E}[\|\nabla\psi_{1/\rho}(x^k)\|^2]\\
\leq&\frac{p}{p-2\delta}\cdot\frac{2\rho}{\rho-\tau-\kappa}\left[D+\frac{\rho}{\alpha^2}(\alpha+\sigma+L_\omega)^2\right]\left(\frac{\rho+\lambda}{K}+\frac{\alpha+\mathsf{G}_\delta}{\sqrt{K}}\right)
\end{aligned}$$

and this completes the proof after re-arrangement.

E. Stochastic Convex Optimization beyond Lipschitz Continuity

In this section, we consider applying the above mentioned ideas to convex optimization. When dealing with convex optimization problems, instead of relying on Moreau envelope smoothing, we have a better potential function $\|x-x^\star\|$ directly relevant to distance to optimal set $\mathcal{X}^\star$. This turns out greatly simplifies our assumptions.

E1: $f(x,\xi)$ is convex for all $\xi\sim\Xi$. $\lambda=\kappa=0$ and $\tau=0$.

It is rather straight-forward to extend our results to convex optimization. And we remark that our analysis is different from (Mai and Johansson, 2021), where the authors focus on subgradient method and assume quadratic growth condition.

E.1. Convex Optimization under Standard Lipschitzness

Lemma E.1. *Suppose that* **A1** *to* **A3**, **E1** *as well as* **B1** *holds, then given* $\gamma_k>0$

$$\mathbb{E}_k[\|x^{k+1}-x^\star\|^2]\leq\|x^k-x^\star\|^2-\frac{2}{\gamma_k}[\psi(x^k)-\psi(x^\star)]+\frac{(L_f+L_\omega)^2}{\gamma_k^2},\tag{36}$$

where $x^\star \in \mathcal{X}^\star$ is any optimal solution.

Theorem E.1. *Under the same assumptions as Lemma E.1, if we take $\gamma_k \equiv \gamma = \alpha\sqrt{K}$, then*

$$\min_{1\le k\le K} \mathbb{E}[\psi(x^k) - \psi(x^\star)] \le \frac{1}{2\sqrt{K}}\left[\|x^1 - x^\star\|^2\alpha + \frac{(L_f + L_\omega)^2}{\alpha}\right],$$

where $x^\star \in \mathcal{X}^\star$ is an optimal solution.

Remark 6. We observe the same trade-off as in weakly convex optimization, where we have, given telescopic sum of (36), that

$$\frac{1}{K}\sum_{k=1}^K \mathcal{O}(\gamma_k^{-1})\mathbb{E}[\psi(x^k) - \psi(x^\star)] \le \mathcal{O}\Big(\frac{1}{K}\Big) + \frac{1}{K}\sum_{k=1}^K \mathcal{O}(L_f^2\gamma_k^{-2}).$$

Compared with weakly convex case

$$\frac{1}{K}\sum_{k=1}^K \mathcal{O}(\gamma_k^{-1})\mathbb{E}[\|\nabla\psi_{1/\rho}(x^k)\|^2] \le \mathcal{O}\Big(\frac{1}{K}\Big) + \frac{1}{K}\sum_{k=1}^K \mathcal{O}(L_f^2\gamma_k^{-2}),$$

this resemblance implies our previous analysis for weakly convex optimization is immediately applicable.

E.2. Convex Optimization under Generalized Lipschitzness

Lemma E.2. *Suppose A1 to A3, E1 as well as C1 holds, then given $\gamma_k > 0$,*

$$\mathbb{E}_k[\|x^{k+1} - x^\star\|^2] \le \|x^k - x^\star\|^2 - \frac{2}{\gamma_k}[\psi(x^k) - \psi(x^\star)] + \frac{(\mathcal{G}(\|x^k\|)L_f + L_\omega)^2}{\gamma_k^2},$$

where $x^\star \in \mathcal{X}^\star$ is an optimal solution.

Theorem E.2. *With the same conditions as Lemma E.2, if $\gamma_k = (\mathcal{G}(\|x^k\|) + 1)k^\zeta, \zeta \in \big(\frac{1}{2}, 1\big)$, then as $k \to \infty$, $\{\|x^k\|\}$ is bounded with probability 1 and $\{\inf_{j\le k} f(x^j) - f(x^\star)\}$ converges to 0 almost surely.*

Lemma E.3. *Under the same conditions as Lemma E.2, if we take $\gamma_k = \alpha(\mathcal{G}(\|x^k\|) + 1)\sqrt{K}$, then the tail bound*

$$\mathbb{P}\left\{\|x^k - x^\star\| \ge \frac{2(L_f + L_\omega)}{\alpha\sqrt{K}} + a\right\} \le \frac{2\Delta}{a},$$

holds for all $2 \le k \le K$, where $\Delta = \|x^1 - x^\star\| + \frac{L_f + L_\omega}{\alpha}$.

Theorem E.3. *Under the same conditions as Lemma E.2, given $\delta \in (0, 1/4), p \in (2\delta, 1)$, $(1 - 2p^{-1}\delta)K$ iterations will lie in the ball centered around $x^\star$ with radius $\mathsf{R}(\delta) = \delta^{-1}\Delta + \frac{2(L_f+L_\omega)}{\alpha\sqrt{K}}$ and*

$$\min_{1\le k\le K} \mathbb{E}[\psi(x^k) - \psi(x^\star)] \le \frac{p}{p - 2\delta}\cdot\frac{\mathsf{G}_\delta + 1}{2\sqrt{K}}\left[\|x^1 - x^\star\|^2\alpha + \frac{(L_f + L_\omega)^2}{\alpha},\right] \tag{37}$$

where $\mathsf{G}_\delta := \max_z \mathcal{G}(z), \|z - x^\star\| \le \delta^{-1}\Delta + \frac{2(L_f+L_\omega)}{\alpha\sqrt{K}}$.

Remark 7. We note that $x^\star$ is actually arbitrary. Therefore we can take it to be a minimum norm optimal solution to get a tighter bound.

E.3. Convex Optimization under Unknown Lipschitzness

Lemma E.4. *Suppose that A1 to A3, E1 as well as D1, D2 hold, then given $\gamma > 0$,*

$$\mathbb{E}_k[\|x^{k+1} - x^\star\|^2] \le \|x^k - x^\star\|^2 - \frac{2}{\gamma_k}[\psi(x^k) - \psi(x^\star)] + \mathbb{E}_k\left[\frac{(\mathsf{Lip}(x^k, \xi^k)) + L_\omega)^2}{\gamma_k^2}\right], \tag{38}$$

where γ_k is chosen to be independent of ξ^k and is considered deterministic here.

Theorem E.4. *With the same conditions as Lemma E.4, if* $\gamma_k = \max\{\mathsf{Lip}(x^k, \xi^k)), \alpha\} k^{\zeta}, \zeta \in (\frac{1}{2}, 1)$*, then as* $k \to \infty$*,* $\{\|x^k\|\}$ *is bounded with probability 1 and* $\{\inf_{j \le k} f(x^j) - f(x^\star)\}$ *converges to 0 almost surely.*

Lemma E.5. *Under the same conditions as Lemma E.4, if we take* $\gamma_k = \max\{\mathsf{Lip}(f(x^k, \xi^k), \alpha\}\sqrt{K}$*, then the tail bound*

$$\mathbb{P}\left\{\|x^k - x^\star\| \ge \frac{2(\alpha + \sigma + L_\omega)}{\alpha\sqrt{K}} + a\right\} \le \frac{2\Delta}{a},$$

holds for all $2 \le k \le K$*, where* $\Delta = \|x^1 - x^\star\| + \frac{\alpha+\sigma+L_\omega}{\alpha}$.

Remark 8. Now that in convex optimization our potential function $\|x - x^\star\|^2$ itself already defines a bounded set. The proof of **E.5** can also be done based on a conditional probability argument.

Theorem E.5. *Under the same conditions as Lemma E.4, given* $\delta \in (0, 1/4), p \in (2\delta, 1)$*,* $(1 - 2p^{-1}\delta)K$ *iterations will lie in the ball centered around* $x^\star$ *with radius* $\mathsf{R}(\delta) = \delta^{-1}\Delta + \frac{2(\alpha+\sigma+L_\omega)}{\alpha\sqrt{K}}$ *and*

$$\min_{1\le k\le K} \mathbb{E}[\psi(x^k) - \psi(x^\star)] \le \frac{p}{p - 2\delta} \cdot \frac{\mathsf{G}_\delta + \alpha}{2\sqrt{K}} \left[\|x^1 - x^\star\|^2\alpha + \frac{(\alpha + \sigma + L_\omega)^2}{\alpha}, \right] \tag{39}$$

where $\mathsf{G}_\delta := \max_x \sup_{\xi\sim\Xi} \mathsf{Lip}(x, \xi), \|x - x^\star\| \le \delta^{-1}\Delta + \frac{2(\alpha+\sigma+L_\omega)}{\alpha\sqrt{K}}$.

E.4. Proof of Results in Subsection E.1

E.4.1. Proof of Lemma E.1

Let $x^\star \in \mathcal{X}^\star$ be an optimal solution to the problem. Then by three-point lemma, we have

$$\begin{aligned}
&f_{x^k}(x^{k+1}, \xi^k) + \omega(x^{k+1}) + \frac{\gamma_k}{2}\|x^{k+1} - x^k\|^2 \\
&\le f_{x^k}(x^\star, \xi^k) + \omega(x^\star) + \frac{\gamma_k}{2}\|x^k - x^\star\|^2 - \frac{\gamma_k}{2}\|x^{k+1} - x^\star\|^2.
\end{aligned}$$

Re-arranging the terms, we deduce that

$$\begin{aligned}
&\frac{\gamma_k}{2}\|x^{k+1} - x^\star\|^2 \\
&\le \frac{\gamma_k}{2}\|x^k - x^\star\|^2 - \frac{\gamma_k}{2}\|x^{k+1} - x^k\|^2 + f_{x^k}(x^\star, \xi^k) + \omega(x^\star) - f_{x^k}(x^{k+1}, \xi^k) - \omega(x^{k+1}) \\
&\le \frac{\gamma_k}{2}\|x^k - x^\star\|^2 - \frac{\gamma_k}{2}\|x^{k+1} - x^k\|^2 + (L_f(\xi^k) + L_\omega)\|x^{k+1} - x^k\| \qquad (40) \\
&\quad + f_{x^k}(x^\star, \xi^k) + \omega(x^\star) - f(x^k) - \omega(x^k),
\end{aligned}$$

where (40) applies **B1** to get $f_{x^k}(x^{k+1}, \xi^k) - f_{x^k}(x^k, \xi^k) \le L_f(\xi^k)\|x^{k+1} - x^k\|$. Dividing both sides by $\frac{\gamma_k}{2}$,

$$\begin{aligned}
\|x^{k+1} - x^\star\|^2 \le \|x^k - x^\star\|^2 - \|x^{k+1} - x^k\|^2 + \frac{2(L_f(\xi^k) + L_\omega)}{\gamma_k}\|x^{k+1} - x^k\| \\
+ \frac{2}{\gamma_k}[f(x^\star, \xi^k) + \omega(x^\star) - f(x^k) - \omega(x^k)] + \frac{2}{\gamma_k}[f_{x^k}(x^\star, \xi) - f(x^\star, \xi)].
\end{aligned}$$

Conditioned on $x^1, \ldots, x^k$ and taking expectation with respect to ξ^k, we successively deduce that

$$\begin{aligned}
&\mathbb{E}_k[\|x^{k+1} - x^\star\|^2] \\
&\le \|x^k - x^\star\|^2 - \mathbb{E}_k[\|x^{k+1} - x^k\|^2] + \mathbb{E}_k[2\gamma_k^{-1}(L_f(\xi^k) + L_\omega)\|x^{k+1} - x^k\|] \\
&\quad + \frac{2}{\gamma_k}[f(x^\star) + \omega(x^\star) - f(x^k) - \omega(x^k)] \qquad (41) \\
&\le \|x^k - x^\star\|^2 + \frac{(L_f + L_\omega)^2}{\gamma_k^2} + \frac{2}{\gamma_k}[f(x^\star) + \omega(x^\star) - f(x^k) - \omega(x^k)] \qquad (42) \\
&= \|x^k - x^\star\|^2 + \frac{(L_f + L_\omega)^2}{\gamma_k^2} + \frac{2}{\gamma_k}[\psi(x^\star) - \psi(x^k)],
\end{aligned}$$

where (41) applies **E1** to get $\mathbb{E}_{\xi^k}[f_{x^k}(x^\star,\xi)-f(x^\star,\xi)]\leq 0$; (42) uses $-\frac{a}{2}x^2+bx\leq\frac{b^2}{2a}$ and that $\mathbb{E}[L_f(\xi)^2]\leq L_f^2$. Re-arranging the terms, we arrive at

$$\mathbb{E}_k[\|x^{k+1}-x^\star\|^2]\leq\|x^k-x^\star\|^2-\frac{2}{\gamma_k}[\psi(x^k)-\psi(x^\star)]+\frac{(L_f+L_\omega)^2}{\gamma_k^2}$$

and this completes the proof.

E.4.2. Proof of Theorem E.1

Taking $\gamma_k\equiv\gamma=\alpha\sqrt{K}$ and telescoping from $1,\ldots,K$, we have

$$\begin{aligned}\min_{1\leq k\leq K}\mathbb{E}[\psi(x^k)-\psi(x^\star)]&\leq\frac{1}{K}\sum_{k=1}^K\mathbb{E}[\psi(x^k)-\psi(x^\star)]\\&\leq\frac{1}{2\sqrt{K}}\Big[\|x^1-x^\star\|^2\alpha+\frac{(L_f+L_\omega)^2}{\alpha}\Big]\end{aligned}$$

and this completes the proof.

E.5. Proof of Results in Subsection E.2

E.5.1. Proof of Lemma E.2

Define $\mathsf{G}_k:=\mathcal{G}(\|x^k\|)$. Let $x^\star\in\mathcal{X}^\star$ be an optimal solution to the problem. Similarly, we have

$$\begin{aligned}&f_{x^k}(x^{k+1},\xi^k)+\omega(x^{k+1})+\frac{\gamma_k}{2}\|x^{k+1}-x^k\|^2\\&\leq f_{x^k}(x^\star,\xi^k)+\omega(x^\star)+\frac{\gamma_k}{2}\|x^k-x^\star\|^2-\frac{\gamma_k}{2}\|x^{k+1}-x^\star\|^2.\end{aligned}$$

Re-arranging the terms, for $\xi^k\sim\Xi$, we deduce that

$$\begin{aligned}&\frac{\gamma_k}{2}\|x^{k+1}-x^\star\|^2\\&\leq\frac{\gamma_k}{2}\|x^k-x^\star\|^2-\frac{\gamma_k}{2}\|x^{k+1}-x^k\|^2+f_{x^k}(x^\star,\xi^k)+\omega(x^\star)-f_{x^k}(x^{k+1},\xi^k)-\omega(x^{k+1})\\&\leq\frac{\gamma_k}{2}\|x^k-x^\star\|^2-\frac{\gamma_k}{2}\|x^{k+1}-x^k\|^2+(\mathsf{G}_kL_f(\xi^k)+L_\omega)\|x^{k+1}-x^k\|\\&\quad+f(x^\star,\xi^k)+\omega(x^\star)-f(x^k)-\omega(x^k),\end{aligned}\tag{43}$$

where (43) applies convexity. Dividing both sides by $\frac{\gamma_k}{2}$, we have

$$\begin{aligned}\|x^{k+1}-x^\star\|^2\leq\|x^k-x^\star\|^2-\|x^{k+1}-x^k\|^2+\frac{2(\mathsf{G}_kL_f(\xi^k)+L_\omega)}{\gamma_k}\|x^{k+1}-x^k\|\\+\frac{2}{\gamma_k}[f(x^\star,\xi^k)+\omega(x^\star)-f(x^k)-\omega(x^k)]\end{aligned}$$

Next, conditioned on $x^1,\ldots,x^k$ and taking expectation with respect to ξ^k, we successively deduce that

$$\begin{aligned}&\mathbb{E}_k[\|x^{k+1}-x^\star\|^2]\\&\leq\|x^k-x^\star\|^2-\mathbb{E}_k[\|x^{k+1}-x^k\|^2]+\mathbb{E}_k[2\gamma_k^{-1}(\mathsf{G}_kL_f(\xi^k)+L_\omega)\|x^{k+1}-x^k\|]\\&\quad+\frac{2}{\gamma_k}[f(x^\star)+\omega(x^\star)-f(x^k)-\omega(x^k)]\\&\leq\|x^k-x^\star\|^2+\frac{(\mathsf{G}_kL_f+L_\omega)^2}{\gamma_k^2}+\frac{2}{\gamma_k}[f(x^\star)+\omega(x^\star)-f(x^k)-\omega(x^k)]\\&=\|x^k-x^\star\|^2+\frac{(\mathsf{G}_kL_f+L_\omega)^2}{\gamma_k^2}+\frac{2}{\gamma_k}[\psi(x^\star)-\psi(x^k)],\end{aligned}\tag{44}$$

where (44) again uses $-\frac{a}{2}x^2 + bx \leq \frac{b^2}{2a}$ and the assumption $\mathbb{E}_\xi[L_f(\xi)^2] \leq L_f^2$. Re-arranging the terms, we get

$$\mathbb{E}_k[\|x^{k+1} - x^\star\|^2] \leq \|x^k - x^\star\|^2 - \frac{2}{\gamma_k}[\psi(x^k) - \psi(x^\star)] + \frac{(\mathsf{G}_k L_f + L_\omega)^2}{\gamma_k^2} \tag{45}$$

and this completes the proof.

E.5.2. Proof of Theorem E.2

Now that our recursive potential reduction is changed into

$$\mathbb{E}_k[\|x^{k+1} - x^\star\|^2] \leq \|x^k - x^\star\|^2 - \frac{2}{\gamma_k}[\psi(x^k) - \psi(x^\star)] + \frac{(\mathsf{G}_k L_f + L_\omega)^2}{\gamma_k^2}$$

and we bound

$$\frac{\mathsf{G}_k L_f + L_\omega}{\gamma_k} = \frac{\mathsf{G}_k L_f + L_\omega}{(\mathsf{G}_k + 1)k^\zeta} \leq \frac{L_f + L_\omega}{k^\zeta},$$

giving

$$\mathbb{E}_k[\|x^{k+1} - x^\star\|^2] \leq \|x^k - x^\star\|^2 - \frac{2}{\gamma_k}[\psi(x^k) - \psi(x^\star)] + \frac{(L_f + L_\omega)^2}{k^{2\zeta}}.$$

Invoking Lemma C.1 with $A_k = 0, V_k = \|x^k - x^\star\|^2 \geq 0, B_k = \frac{(L_f+L_\omega)^2}{k^{2\zeta}}$ and $C_k = \frac{2}{\gamma_k}[\psi(x^k) - \psi(x^\star)]$, we complete the proof with the same argument as in Theorem 4.1.

E.5.3. Proof of Lemma E.3

Our proof is a duplicate of Lemma 4.3 using a different potential function. We start by bounding the error of potential reduction by $\frac{(\mathsf{G}_k L_f+L_\omega)^2}{\gamma_k^2} \leq \frac{(L_f+L_\omega)^2}{\alpha^2 K}$. Then telescoping of (45) gives us, for all $2 \leq k \leq K$, that

$$\mathbb{E}[\|x^k - x^\star\|]^2 \leq \mathbb{E}[\|x^k - x^\star\|^2] \leq \|x^1 - x^\star\|^2 + \frac{(L_f + L_\omega)^2}{\alpha^2} \leq \left(\|x^1 - x^\star\| + \frac{L_f + L_\omega}{\alpha}\right)^2 =: \Delta^2$$

and $\mathbb{E}[\|x^k - x^\star\|] \leq \Delta$. Next consider, by optimality condition, that

$$f_{x^k}(x^{k+1}, \xi^k) + \omega(x^{k+1}) + \frac{\gamma_k}{2}\|x^{k+1} - x^k\|^2 \leq f_{x^k}(x^k, \xi^k) + \omega(x^k) - \frac{\gamma_k}{2}\|x^{k+1} - x^k\|^2.$$

A re-arrangement gives

$$\gamma_k\|x^{k+1} - x^k\|^2 \leq (L_f(\xi^k)\mathsf{G}_k + L_\omega)\|x^{k+1} - x^k\|.$$

Dividing both sides by $\|x^{k+1} - x^k\|$,

$$\|x^{k+1} - x^k\| \leq \frac{L_f(\xi^k)\mathsf{G}_k + L_\omega}{\gamma_k} = \frac{L_f(\xi^k)\mathsf{G}_k + L_\omega}{\alpha(\mathsf{G}_k + 1)\sqrt{K}}.$$

Taking expectation, we get $\mathbb{E}_k[\|x^{k+1} - x^k\|] \leq \frac{L_f+L_\omega}{\alpha\sqrt{K}}$. Then by Markov's inequality,

$$\mathbb{P}_{\xi^k \sim \Xi}\left\{\|x^{k+1} - x^k\| \leq \frac{2(L_f + L_\omega)}{\alpha\sqrt{K}} | \xi^1, \ldots, \xi^{k-1}\right\} \geq \frac{1}{2}$$

and without loss of generality, we let $\mathsf{Z} = \frac{2(L_f+L_\omega)}{\alpha\sqrt{K}} > 0$. Then we successively deduce that

$$\begin{aligned}
\Delta &\geq \mathbb{E}[\|x^{k+1} - x^\star\|] \\
&= \mathbb{E}[\|x^{k+1} - x^\star\| | \|x^k - x^\star\| \leq \mathsf{Z} + z] \cdot \mathbb{P}\{\|x^k - x^\star\| \leq \mathsf{Z} + z\} \\
&\quad + \mathbb{E}[\|x^{k+1} - x^\star\| | \|x^k - x^\star\| \geq \mathsf{Z} + z] \cdot \mathbb{P}\{\|x^k - x^\star\| \geq \mathsf{Z} + z\} \\
&\geq \mathbb{E}[\|x^{k+1} - x^\star\| | \|x^k - x^\star\| \geq \mathsf{Z} + z].
\end{aligned}$$

Next, consider the expectation $\mathbb{E}[\|x^{k+1}-x^\star\| \,|\, \|x^k-x^\star\| \geq \mathsf{Z}+z]$ and we successively deduce that

$$\begin{aligned}
&\mathbb{E}[\|x^{k+1}-x^\star\| \,|\, \|x^k-x^\star\| \geq \mathsf{Z}+z] \\
=\;&\mathbb{E}[\|x^{k+1}-x^\star\| \,|\, \|x^k-x^\star\| \geq \mathsf{Z}+z, \|x^{k+1}-x^k\| \leq \mathsf{Z}] \cdot \mathbb{P}\{\|x^{k+1}-x^k\| \leq \mathsf{Z}\} \\
&+\mathbb{E}[\|x^{k+1}-x^\star\| \,|\, \|x^k-x^\star\| \geq \mathsf{Z}+z, \|x^{k+1}-x^k\| \geq \mathsf{Z}] \cdot \mathbb{P}\{\|x^{k+1}-x^k\| \geq \mathsf{Z}\} \\
\geq\;&\frac{z}{2} \cdot \mathbb{P}\{\|x^{k+1}-x^k\| \geq \mathsf{Z}\} \qquad (46)
\end{aligned}$$

where (46) is by $\mathbb{P}\{\|x^{k+1}-x^k\| \leq \mathsf{Z}\} \geq 0.5$ and that, conditioned on $\|x^{k+1}-x^k\| \leq \mathsf{Z}$,

$$\|x^{k+1}-x^\star\| = \|x^{k+1}-x^k+x^k-x^\star\| \geq \|x^k-x^\star\| - \|x^{k+1}-x^k\| \geq z.$$

Chaining the above inequalities, we arrive at

$$\Delta \geq \frac{z}{2} \cdot \mathbb{P}\{\|x^k-x^\star\| \geq \mathsf{Z}+z\}.$$

Dividing both sides by $\frac{z}{2}$ and taking $z=a$ gives the desired tail bound.

E.5.4. Proof of Theorem E.3

Given $\delta \in (0, 1/4)$, we have

$$\mathbb{P}\{\|x^k-x^\star\| \geq \mathsf{Z}+\delta^{-1}\Delta\} \leq \frac{2\Delta}{\delta^{-1}\Delta} \leq 2\delta.$$

Denote $I_k = \mathbb{I}\{\|x^k-x^\star\| \leq \delta^{-1}\Delta+\mathsf{Z}\}$. We have $\sum_{k=1}^K \mathbb{E}[1-I_k] \leq 2\delta K$ and by Markov's inequality, given $p \in (2\delta, 1)$,

$$\mathbb{P}\left\{\sum_{k=1}^K I_k \geq K(1-2p^{-1}\delta)\right\} \geq 1-p.$$

Now we telescope over Lemma E.3 and deduce that

$$\begin{aligned}
\sum_{k=1}^K \mathbb{E}\left[\frac{2}{\gamma_k}(\psi(x^k)-\psi(x^\star))\right] &\leq \|x^1-x^\star\|^2 - \mathbb{E}[\|x^{K+1}-x^\star\|^2] + \frac{(L_f+L_\omega)^2}{\alpha^2} \\
&\leq \|x^1-x^\star\|^2 + \frac{(L_f+L_\omega)^2}{\alpha^2}.
\end{aligned}$$

Next we condition on the event $\sum_{k=1}^K I_k \geq K(1-2p^{-1}\delta)$, define $\mathsf{G}_\delta := \max_z \mathcal{G}(z), \|z-x^\star\| \leq \delta^{-1}\Delta+\mathsf{Z}$, and successively deduce that

$$\begin{aligned}
&\sum_{k=1}^K \mathbb{E}\left[\frac{2}{\gamma_k}(\psi(x^k)-\psi(x^\star))\right] \\
=\;&\sum_{k\in\{j:I_j=0\}} \mathbb{E}\left[\frac{2}{\gamma_k}(\psi(x^k)-\psi(x^\star))\right] + \sum_{k\in\{j:I_j=1\}} \mathbb{E}\left[\frac{2}{\gamma_k}(\psi(x^k)-\psi(x^\star))\right] \\
\geq\;&\sum_{k\in\{j:I_j=1\}} \mathbb{E}\left[\frac{2}{\gamma_k}(\psi(x^k)-\psi(x^\star))\right] \\
\geq\;&\sum_{k\in\{j:I_j=1\}} \mathbb{E}\left[\frac{2}{\alpha(\mathsf{G}_\delta+1)\sqrt{K}}(\psi(x^k)-\psi(x^\star))\right] \\
\geq\;&\frac{2(1-2p^{-1}\delta)\sqrt{K}}{\alpha(\mathsf{G}_\delta+1)} \min_{k\in\{j:I_j=1\}} \mathbb{E}[(\psi(x^k)-\psi(x^\star))], \qquad (47)
\end{aligned}$$

where (47) follows from the event $\sum_{k=1}^{K} I_k \geq K(1-2p^{-1}\delta)$. Putting the results together, we have

$$\min_{1\leq k\leq K} \mathbb{E}[\psi(x^k)-\psi(x^\star)] \leq \min_{k\in\{j:I_j=0\}} \mathbb{E}[\psi(x^k)-\psi(x^\star)] \tag{48}$$

$$\leq \frac{p}{p-2\delta}\cdot\frac{\mathsf{G}_\delta+1}{2\sqrt{K}}\left[\|x^1-x^\star\|^2\alpha+\frac{(L_f+L_\omega)^2}{\alpha}\right] \tag{49}$$

and this completes the proof.

E.6. Proof of Results in Subsection E.3

In this section, we again define $L_f^k := \mathsf{Lip}(x^k,\xi^k), L_f' := \mathsf{Lip}(x^k,\xi')$ to simplify notation.

E.6.1. Proof of Lemma E.3

By the optimality condition we have

$$f_{x^k}(x^{k+1},\xi^k)+\omega(x^{k+1})+\frac{\gamma_k}{2}\|x^{k+1}-x^k\|^2 \leq f_{x^k}(x^\star,\xi^k)+\omega(x^\star)+\frac{\gamma_k}{2}\|x^k-x^\star\|^2-\frac{\gamma_k}{2}\|x^{k+1}-x^\star\|^2.$$

Re-arranging the terms, we get

$$\begin{aligned}
&\frac{\gamma_k}{2}\|x^{k+1}-x^\star\|^2\\
\leq\ &\frac{\gamma_k}{2}\|x^k-x^\star\|^2-\frac{\gamma_k}{2}\|x^{k+1}-x^k\|^2+f_{x^k}(x^\star,\xi^k)+\omega(x^\star)-f_{x^k}(x^{k+1},\xi^k)-\omega(x^{k+1})\\
\leq\ &\frac{\gamma_k}{2}\|x^k-x^\star\|^2-\frac{\gamma_k}{2}\|x^{k+1}-x^k\|^2+(L_f^k+L_\omega)\|x^{k+1}-x^k\|\\
&+f(x^\star,\xi^k)+\omega(x^\star)-f(x^k)-\omega(x^k).
\end{aligned}$$

Dividing both sides by $\frac{\gamma_k}{2}$,

$$\begin{aligned}
\|x^{k+1}-x^\star\|^2 \leq \|x^k-x^\star\|^2-\|x^{k+1}-x^k\|^2+\frac{2(L_f^k+L_\omega)}{\gamma_k}\|x^{k+1}-x^k\|\\
+\frac{2}{\gamma_k}[f(x^\star,\xi^k)+\omega(x^\star)-f(x^k)-\omega(x^k)]
\end{aligned}$$

Conditioned on $x^1,\ldots,x^k$, taking expectation with respect to ξ^k, we have

$$\begin{aligned}
&\mathbb{E}_k[\|x^{k+1}-x^\star\|^2]\\
\leq\ &\|x^k-x^\star\|^2-\mathbb{E}_k[\|x^{k+1}-x^k\|^2]+\mathbb{E}_k[2\gamma_k^{-1}(L_f^k+L_\omega)\|x^{k+1}-x^k\|]\\
&+\frac{2}{\gamma_k}[f(x^\star)+\omega(x^\star)-f(x^k)-\omega(x^k)] \qquad (50)\\
\leq\ &\|x^k-x^\star\|^2+\frac{(L_f^k+L_\omega)^2}{\gamma_k^2}+\frac{2}{\gamma_k}[f(x^\star)+\omega(x^\star)-f(x^k)-\omega(x^k)]\\
=\ &\|x^k-x^\star\|^2+\frac{(L_f^k+L_\omega)^2}{\gamma_k^2}+\frac{2}{\gamma_k}[\psi(x^\star)-\psi(x^k)],
\end{aligned}$$

where (50) use the fact that γ_k does not inherit randomness from ξ^k. Re-arranging the terms, we get

$$\mathbb{E}_k[\|x^{k+1}-x^\star\|^2] \leq \|x^k-x^\star\|^2-\frac{2}{\gamma_k}[\psi(x^k)-\psi(x^\star)]+\mathbb{E}_k\left[\tfrac{(L_f^k+L_\omega)^2}{\gamma_k^2}\right]$$

and this completes the proof.

E.6.2. Proof of Theorem E.4

We start by bounding $\mathbb{E}_k\left[\frac{(L_f^k+L_\omega)^2}{\gamma_k^2}\right]$ and notice that

$$\frac{(L_f^k+L_\omega)^2}{\gamma_k^2} = \frac{L_f^k}{\gamma_k^2} + \frac{2L_f^k L_\omega}{\gamma_k^2} + \frac{L_\omega^2}{\gamma_k^2}$$

so that we can bound

$$\mathbb{E}_{\xi'}\left[\frac{L_f^k}{\gamma_k^2}\right] = \mathbb{E}_{\xi'}\left[\frac{(L_f^k)^2}{\max\{(L_f')^2,\alpha\}k^{2\zeta}}\right] \le \left(\frac{\alpha+\sigma}{\alpha}\right)^2 \frac{1}{k^{2\zeta}},$$

where we invoked Lemma D.1 and take $X = L_f^k, Y = L_f'$, and we recall that by **D2**, $\mathbb{E}[|L_f^k - L_f'|^2] \le \sigma^2$. Similarly, we can deduce that

$$\mathbb{E}_{\xi'}\left[\frac{2L_f^k L_\omega}{\gamma_k^2}\right] \le \mathbb{E}_{\xi'}\left[\frac{2L_f^k L_\omega}{\max\{L_f',\alpha\}^2}\right] \le \mathbb{E}_{\xi'}\left[\frac{2L_f^k}{\alpha L_f'}\right] \le \left(\frac{\alpha+\sigma}{\alpha^2}\right)\frac{2L_\omega}{k^{2\zeta}}$$

$$\mathbb{E}_{\xi'}\left[\frac{L_\omega^2}{\gamma_k^2}\right] \le \frac{L_\omega^2}{\alpha^2 k^{2\zeta}}.$$

Putting the bounds together,

$$\mathbb{E}_{\xi'}\left[\frac{(L_f^k+L_\omega)^2}{\gamma_k^2}\right] \le \frac{(\alpha+\sigma+L_\omega)^2}{\alpha^2 k^{2\zeta}}.$$

Then we have

$$\mathbb{E}_k[\|x^{k+1}-x^\star\|^2] \le \|x^k - x^\star\|^2 - \frac{2}{\gamma_k}[\psi(x^k)-\psi(x^\star)] + \frac{(\alpha+\sigma+L_\omega)^2}{\alpha^2 k^{2\zeta}}$$

and we complete the proof by invoking Lemma C.1.

E.6.3. Proof of Lemma E.5

We start by bounding $\mathbb{E}_k\left[\frac{(L_f^k+L_\omega)^2}{\gamma_k^2}\right] \le \frac{(\alpha+\sigma+L_\omega)^2}{\alpha^2 K}$. Telescoping the relation

$$\mathbb{E}_k[\|x^{k+1}-x^\star\|^2] \le \|x^k - x^\star\|^2 - \frac{2}{\gamma_k}[\psi(x^k)-\psi(x^\star)] + \frac{(\alpha+\sigma+L_\omega)^2}{\alpha^2 K}$$

gives, for all $2 \le k \le K$, that

$$\mathbb{E}[\|x^k - x^\star\|^2] \le \|x^1 - x^\star\|^2 + \frac{(\alpha+\sigma+L_\omega)^2}{\alpha^2} \le \left(\|x^1-x^\star\| + \frac{\alpha+\sigma+L_\omega}{\alpha}\right)^2 =: \Delta^2.$$

and $\mathbb{E}[\|x^k - x^\star\|] \le \Delta$. Next consider

$$f_{x^k}(x^{k+1},\xi^k) + \omega(x^{k+1}) + \frac{\gamma_k}{2}\|x^{k+1}-x^k\|^2 \le f_{x^k}(x^k,\xi^k) + \omega(x^k) - \frac{\gamma_k}{2}\|x^{k+1}-x^k\|^2$$

and re-arrangement gives $\gamma_k\|x^{k+1}-x^k\|^2 \le (L_f^k + L_\omega)\|x^{k+1}-x^k\|$. Dividing both sides by $\|x^{k+1}-x^k\|$, we get $\|x^{k+1}-x^k\| \le \frac{L_f^k+L_\omega}{\gamma_k}$. Taking expectation, $\mathbb{E}_k[\|x^{k+1}-x^k\|] \le \frac{\alpha+\sigma+L_\omega}{\alpha\sqrt{K}}$. Then by Markov's inequality, $\mathbb{P}_{\xi^k,\xi'\sim\Xi}\left\{\|x^{k+1}-x^k\| \le \frac{2(\alpha+\sigma+L_\omega)}{\alpha\sqrt{K}} | \xi^1,\ldots,\xi^{k-1}\right\} \ge \frac{1}{2}$. and without loss of generality, let $\mathsf{Z} = \frac{2(\alpha+\sigma+L_\omega)}{\alpha\sqrt{K}} > 0$. By the same reasoning, we arrive at $\Delta \ge \frac{z}{2}\cdot\mathbb{P}\{\|x^k-x^\star\| \ge \mathsf{Z}+z\}$ and dividing both sides by $\frac{z}{2}$ gives the desired tail bound.

E.6.4. Proof of Theorem E.5

Following the same argument as Theorem E.4, we have, for $\delta \in (0,1/4)$, that $\mathbb{P}\{\|x^k-x^\star\| \ge \mathsf{Z}+\delta^{-1}\Delta\} \le 2\delta$. Define $I_k = \mathbb{I}\{\|x^k-x^\star\| \le \delta^{-1}\Delta + \mathsf{Z}\}$. We have, by Markov's inequality, that $\mathbb{P}\{\sum_{k=1}^K I_k \ge K(1-2p^{-1}\delta)\} \ge 1-p$. Then telescoping over Lemma E.5 gives

$$\sum_{k=1}^K \mathbb{E}\left[\frac{2}{\gamma_k}(\psi(x^k)-\psi(x^\star))\right] \le \|x^1-x^\star\|^2 - \mathbb{E}[\|x^{K+1}-x^\star\|^2] + \frac{(\alpha+\sigma+L_\omega)^2}{\alpha^2}$$

$$\le \|x^1-x^\star\|^2 + \left(\frac{\alpha+\sigma+L_\omega}{\alpha}\right)^2.$$

Conditioned on $\sum_{k=1}^{K} I_k \geq K(1-2p^{-1}\delta)$, we get

$$\sum_{k=1}^{K} \mathbb{E}\left[\frac{2}{\gamma_k}(\psi(x^k)-\psi(x^\star))\right] \geq \frac{2(1-2p^{-1}\delta)\sqrt{K}}{\alpha+\mathsf{G}_\delta} \min_{k\in\{j:I_j=1\}} \mathbb{E}[\psi(x^k)-\psi(x^\star)]$$

where $\mathsf{G}_\delta \coloneqq \max_x \sup_{\xi\sim\Xi} \mathsf{Lip}(x,\xi), \|x-x^\star\| \leq \delta^{-1}\Delta + \mathsf{Z}$. Combining two inequalities completes the proof.