New Results on the Polyak Stepsize: Tight Convergence Analysis and Universal Function Classes

Chang He * Wenzhi Gao † Bo Jiang ‡ Madeleine Udell § Shuzhong Zhang ¶

December 10, 2025

Abstract

In this paper, we revisit a classical adaptive stepsize strategy for gradient descent: the Polyak stepsize (`PolyakGD`), originally proposed in Polyak (1969). We study the convergence behavior of `PolyakGD` from two perspectives: tight worst-case analysis and universality across function classes. As our first main result, we establish the tightness of the known convergence rates of `PolyakGD` by explicitly constructing worst-case functions. In particular, we show that the $\mathcal{O}((1-\frac{1}{\kappa})^K)$ rate for smooth strongly convex functions and the $\mathcal{O}(1/K)$ rate for smooth convex functions are both tight. Moreover, we theoretically show that `PolyakGD` automatically exploits floating-point errors to escape the worst-case behavior. Our second main result provides new convergence guarantees for `PolyakGD` under both Hölder smoothness and Hölder growth conditions. These findings show that the Polyak stepsize is universal, automatically adapting to various function classes without requiring prior knowledge of problem parameters.

1 Introduction

The Polyak stepsize is a stepsize schedule for (sub)gradient descent in convex optimization, originally proposed by Boris T. Polyak (Polyak, 1969, 1987). It was initially designed for the subgradient method in nonsmooth optimization, and has recently been extended to more general settings (Hazan and Kakade, 2019; Loizou et al., 2021; Wang et al., 2023; Deng et al., 2024; Devanathan and Boyd, 2024). The Polyak stepsize requires knowledge of the optimal function value $f^\star$ and is defined as

$$\alpha_k = \frac{f(x^k) - f^\star}{\|\nabla f(x^k)\|^2}, \ k = 0, 1, \ldots, K-1. \tag{1}$$

`PolyakGD` automatically adjusts the stepsize according to the position of the current iterate x^k on the optimization landscape. The resulting stepsizes are typically much larger than standard constant or diminishing stepsizes, making `PolyakGD` more aggressive and often yielding superior practical performance (Zamani and Glineur, 2024). Consequently, the Polyak stepsize is widely used in settings where the optimal function value $f^\star$ is known, such as convex feasibility problems

*The first two authors contribute equally. School of Information Management and Engineering, Shanghai University of Finance and Economics; Department of Industrial and Systems Engineering, University of Minnesota. ischanghe@gmail.com

†ICME, Stanford University. gwz@stanford.edu

‡School of Information Management and Engineering, Shanghai University of Finance and Economics. isyebojiang@gmail.com

§ICME, Stanford University. udell@stanford.edu

¶Department of Industrial and Systems Engineering, University of Minnesota. zhangs@umn.edu

and over-parameterized machine learning models (Boyd, 2013; Loizou et al., 2021). Even when $f^\star$ is not known a priori, `PolyakGD` remains highly effective, provided that a reasonable estimate of the optimal value is available (Polyak, 1969; Hazan and Kakade, 2019; Zheng et al., 2024).

From a theoretical perspective, the convergence of the Polyak stepsize has been extensively studied in nonsmooth convex optimization. Polyak (1987) proved that the subgradient method with the Polyak stepsize satisfies $\lim_{k\to\infty} \sqrt{k}\,(f(x^k) - f^\star) = 0$. Moreover, the convergence rate of the best-iterate[1] is $\mathcal{O}(1/\sqrt{K})$ (see, e.g., Boyd (2013)). Beyond best-iterate guarantees, Zamani and Glineur (2024) analyzed the last-iterate behavior of the Polyak stepsize, establishing an $\mathcal{O}(1/\sqrt[4]{K})$ convergence rate and providing a worst-case function to match this rate. Their analysis uses the methodology based on performance estimation problem (PEP) (Drori and Teboulle, 2014; Taylor et al., 2017), building on earlier work (Zamani and Glineur, 2025)[2]. Additionally, Criscitiello and Boumal (2023) provided a worst-case function for subgradient method with the Polyak stepsize on Riemannian manifolds. More recently, Orabona and D'Orazio (2025) showed that the Polyak stepsize can be interpreted as applying the subgradient method to a specific surrogate function, providing an alternative justification for the necessity of acquiring $f^\star$. In contrast to the well-studied nonsmooth setting, the smooth setting remains largely unexplored. The primary reference is the note by Hazan and Kakade (2019), who established convergence guarantees for `PolyakGD` on both smooth convex and smooth strongly convex objectives. In the smooth convex setting, they proved an $\mathcal{O}(1/K)$ convergence rate for the best-iterate, matching the rate of constant-stepsize gradient descent.

Although `PolyakGD` often exhibits superior practical performance, the existing $\mathcal{O}(1/K)$ convergence rate matches that of gradient descent with a constant stepsize. This agreement was unsurprising prior to recent PEP-based convergence analysis (Grimmer, 2024; Altschuler and Parrilo, 2025). For example, Grimmer (2024) showed that gradient descent with frequent long steps achieves provably faster convergence. Unfortunately, the PEP methodology-based does not readily extend to adaptive stepsizes for general convex functions (Goujaud et al., 2023, 2024), even applying current computer-assisted analysis techniques remains challenging. To our best knowledge, the only two existing papers studying the tight convergence of the Polyak stepsize focus on the strongly convex case and only deal with variants of the original Polyak stepsize (Barré et al., 2020; Huang and Qi, 2024). In the strongly convex setting, PEP remains applicable, as one can analyze the contraction ratio between consecutive iterations. This line of work naturally raises the following question:

Is the existing upper bound for the Polyak stepsize tight?

In this paper, we show that the known convergence rates for `PolyakGD` are indeed tight by constructing worst-case functions. On the other hand, we demonstrate that under floating-point arithmetic, `PolyakGD` can automatically leverage numerical error to escape this worst-case behavior. This interesting phenomenon explains the superior empirical performance of `PolyakGD` in practice. Theoretically, the strong performance of `PolyakGD` is attributed to its inherent adaptivity, a crucial property for modern optimization methods (Malitsky and Mishchenko, 2020; Gao et al., 2024; Li and Lan, 2025; He et al., 2025). Hazan and Kakade (2019) showed that the Polyak stepsize is simultaneously adaptive to both smooth and nonsmooth regimes, encompassing general convex as well as strongly convex objectives. Mishkin et al. (2024) further established that it adapts to directional smoothness. More recently, Orabona and D'Orazio (2025) demonstrated that the Polyak

[1] By *best-iterate*, we refer to the quantity $\min_{1\le k\le K}(f(x^k) - f^\star)$; by *last-iterate*, we refer to $f(x^K) - f^\star$.

[2] Zamani and Glineur (2024, 2025) employ PEP to obtain numerical dual multipliers that guide the derivation of analytical expressions for their proofs.

stepsize also adapts to the curvature of a surrogate function induced by the original objective. Inspired by these results, we further advance this line of research and provide a comprehensive answer to the following question:

Can the adaptivity of the Polyak stepsize be further extended?

Contributions. This work answers the above two questions. To address the first question, we construct a simple two-dimensional quadratic function. Our proof strategy relies on choosing a specific initial point such that the Polyak stepsize reduces to a constant stepsize along the trajectory in the strongly convex case. Subsequently, we derive the worst-case functions for the general convex and Hölder smooth cases by adapting this quadratic function. Based on this construction, we establish the tightness of the linear convergence rate for strongly convex functions, the $\mathcal{O}(1/K)$ rate for general convex functions, and the $\mathcal{O}(1/K^{(\nu+1)/2})$ rate for ν-Hölder smooth objectives. We then show that under floating-point arithmetic, γ-`PolyakGD` escapes this worst-case behavior.

Our second result demonstrates that the Polyak stepsize is universal across various function classes. We establish that `PolyakGD` simultaneously adapts to both Hölder smoothness and Hölder growth conditions (Theorem 4.1). In particular, when only the Hölder growth condition holds, γ-`PolyakGD` achieves the optimal convergence rate; when the function is Hölder smooth, it matches the rate of the universal gradient method of Nesterov (2015). Furthermore, we show that the Polyak stepsize adapts to the recently proposed intrinsic global curvature bound (Nesterov, 2025), and that our universal analysis extends to the stochastic setting under the interpolation condition. A summary of these results is provided in Table 1.

Table 1: Upper/lower bounds of `PolyakGD`.

Function class	Upper bound (rate)	Lower bound (rate)
L-smooth	$\mathcal{O}(1/K)$ Hazan and Kakade (2019)	$\Omega(1/K)$ Theorem 3.2
L-smooth μ-strongly convex	$\mathcal{O}((1-\frac{1}{\kappa})^K)$ Hazan and Kakade (2019)	$\Omega((1-\frac{1}{\kappa})^K)$ Theorem 3.1
ν-Hölder smooth	$\mathcal{O}(K^{-(\nu+1)/2})$ Theorem 4.1	$\Omega(K^{-(\nu+1)/2})$ Theorem 3.3
r-Hölder growth	$\mathcal{O}(K^{-r/(2(r-1))})$ Theorem 4.1	$\Omega(K^{-r/(2(r-1))})$ Nemirovskii and Nesterov (1985)
ν-Hölder smooth + r-Hölder growth	$\mathcal{O}(K^{-r(\nu+1)/(2(r-\nu-1))})$ Theorem 4.1	—

2 Preliminaries

We consider the following (un)constrained optimization problem

$$\min_{x\in\mathbb{R}^n} \; f(x), \tag{2}$$

where the objective function $f:\mathbb{R}^n\to\mathbb{R}$ is convex and differentiable[3]. We assume that the optimal solution set $\mathcal{X}^\star$ is nonempty and the optimal function value $f^\star = f(x^\star)$ is finite. Besides, we

[3]Our analysis can be extended to the projected (sub)gradient method for the constrained setting $\min_{x\in\mathcal{X}} f(x)$ with a closed convex constraint set $\mathcal{X}$ and f is nonsmooth.

assume that $f^\star$ is known. Addressing the scenario where $f^\star$ is unknown is beyond the scope of this paper, and the techniques developed in Polyak (1969); Hazan and Kakade (2019) can be applied to estimate $f^\star$ at the cost of an additional $\log(1/\varepsilon)$ factor in the overall complexity.

Algorithm 1: γ-`PolyakGD`

input initial point x^0, optimal value $f^\star$, scaling parameter $\gamma \in (0, 2]$

for $k = 0, 1, \ldots$ **do**

$\alpha_k = \frac{f(x^k) - f^\star}{\|\nabla f(x^k)\|^2}$

$x^{k+1} = x^k - \gamma \cdot \alpha_k \nabla f(x^k)$

output

Throughout the paper, we study the convergence behavior of gradient descent with a scaled Polyak stepsize (Algorithm 1), where $\gamma \in (0, 2]$ is a scaling parameter; we refer to this update as γ-`PolyakGD`. Clearly, setting $\gamma = 1$ recovers the standard Polyak stepsize gradient descent. For the upper curvature bound, we adopt the Hölder smoothness condition:

Definition 2.1. *A differentiable function f is (ν, L_ν)-Hölder smooth if it satisfies*

$$\|\nabla f(x) - \nabla f(y)\| \leq L_\nu \|x - y\|^\nu, \ \nu \in (0, 1], \ L_\nu > 0, \ \textit{for all } x, y \in \mathbb{R}^n. \tag{3}$$

Specifically, the case $\nu = 1$ corresponds to the standard L-smoothness condition. The (ν, L_ν)-Hölder smoothness condition implies the following two inequalities:

$$\begin{aligned} f(y) &\leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L_\nu}{\nu + 1} \|y - x\|^{\nu+1}, \quad \text{for all } x, y \in \mathbb{R}^n, \\ \|\nabla f(x)\|^{\frac{\nu+1}{\nu}} &\leq \frac{\nu + 1}{\nu} L_\nu^{\frac{1}{\nu}} (f(x) - f^\star), \quad \text{for all } x \in \mathbb{R}^n, \end{aligned} \tag{4}$$

with proofs provided in Orabona (2023). For the lower curvature bound, we adopt the following growth condition:

Definition 2.2. *A continuous function f satisfies the (r, ρ_r)-Hölder growth condition on a nonempty set $\mathcal{K} \subseteq \mathbb{R}^n$ if*

$$f(x) - f^\star \geq \rho_r \operatorname{dist}(x, \mathcal{X}^\star)^r, \ r \geq 1, \ \rho_r > 0, \quad \textit{for all } x \in \mathcal{K}, \tag{5}$$

where $\operatorname{dist}(x, \mathcal{X}^\star) := \inf_{x^\star \in \mathcal{X}^\star} \|x - x^\star\|$.

Note that the first inequality in Equation (4) implies

$$f(x) - f^\star \leq \frac{L_\nu}{\nu + 1} \operatorname{dist}(x, \mathcal{X}^\star)^{\nu+1}, \quad \text{for all } x \in \mathbb{R}^n.$$

Combining this with the (r, ρ_r)-Hölder growth condition on the set $\mathcal{K}$, we have

$$\rho_r \operatorname{dist}(x, \mathcal{X}^\star)^r \leq \frac{L_\nu}{\nu + 1} \operatorname{dist}(x, \mathcal{X}^\star)^{\nu+1}, \quad \text{for all } x \in \mathcal{K}.$$

This implies that

$$\frac{\rho_r(\nu + 1)}{L_\nu} \leq \operatorname{dist}(x, \mathcal{X}^\star)^{(\nu+1)-r}, \quad \text{for all } x \in \mathcal{K}. \tag{6}$$

Consequently, several subtle points deserve attention. First, we require that $\mathcal{X}^\star \subseteq \mathcal{K}$, otherwise the Hölder growth condition does not seem intuitive. As a result, we exclude the case $r < \nu +$

1, as choosing x sufficiently close to $\mathcal{X}^\star$ would lead to a contradiction with the inequality (6). Moreover, we require that $\sup_{x\in\mathcal{K}} \operatorname{dist}(x, \mathcal{X}^\star) < \infty$. Otherwise, the case $r > \nu + 1$ would also yield a contradiction by taking x such that $\operatorname{dist}(x, \mathcal{X}^\star) \to \infty$ (see Section 1.2 of Roulet and d'Aspremont (2017) for a similar argument). Fortunately, as discussed at the beginning of Section 4, γ-`PolyakGD` (Algorithm 1) naturally ensures the existence of such a proper set $\mathcal{K}$ due to the Fejér monotonicity.

3 Tight Convergence of γ-PolyakGD

In this section, we study the tightness of the convergence rates of γ-`PolyakGD` for various function classes. Taking the class of convex L-smooth functions as an example, the formal problem is: given a total iteration count K, can we find a convex and L-smooth function f and an initial point x^0 such that the trajectory generated by γ-`PolyakGD` satisfies

$$\min_{1\le k\le K} \frac{f(x^k) - f^\star}{L\|x^0 - x^\star\|^2} \ge \Omega\Big(\frac{1}{K}\Big)?$$

Consequently, the main task is to construct a worst-case function to answer the above question. For instance, Drori and Teboulle (2014); Teboulle and Vaisbourd (2023) utilized the Huber loss to establish the tightness of the $\mathcal{O}(1/K)$ rate for constant stepsize gradient descent. The intuition is that the small gradient magnitude of the Huber loss forces a constant stepsize gradient descent to take conservatively small steps, thereby justifying the tightness. However, this strategy fails due to the adaptivity of γ-`PolyakGD`, as demonstrated below.

Proposition 3.1. *Let $K \ge 1$ be the total iteration count. Consider the Huber loss defined by*

$$\mathcal{H}_K(x) := \begin{cases} \frac{1}{2K+1}\|x\| - \frac{1}{2(2K+1)^2}, & \|x\| > \frac{1}{2K+1}, \\ \frac{1}{2}\|x\|^2, & \|x\| \le \frac{1}{2K+1}. \end{cases}$$

Then, for any initial point $x^0 \in \mathbb{R}^n$, 1-`PolyakGD` converges to $x^\star = 0$ in two steps.

Therefore, our strategy is to construct a function such that the Polyak stepsize reduces to a constant stepsize along the trajectory. We remark that the first two theorems (Theorems 3.1 and 3.2) allow for a wider range of $\gamma \in (0, 4)$, which facilitates the analysis for the worst-case function under Hölder smoothness (Theorem 3.3).

3.1 Worst-case quadratic functions for γ-PolyakGD

We start with the case where the objective function f is μ-strongly convex ($\mu > 0$) and L-smooth.

Theorem 3.1. *Given the parameter $\gamma \in (0, 4)$, let $\kappa > \max\{4/\gamma - 1, \gamma/(4-\gamma)\}$ be the condition number. Consider the following two-dimensional quadratic optimization problem*

$$\min_{x\in\mathbb{R}^2} q(x) := \frac{\kappa}{2}x_{(1)}^2 + \frac{1}{2}x_{(2)}^2.$$

Suppose the initial point $x^0 = (x_{(1)}^0, x_{(2)}^0)$ satisfies

$$(x_{(2)}^0)^2 = \frac{4\kappa^2 - \gamma\kappa(\kappa+1)}{\gamma(\kappa+1) - 4}(x_{(1)}^0)^2. \tag{7}$$

Then γ-`PolyakGD` generates a trajectory of $\{x^k\}$ such that

$$\left|\frac{x_{(1)}^{k+1}}{x_{(1)}^k}\right| = \left|\frac{x_{(2)}^{k+1}}{x_{(2)}^k}\right| = \frac{\kappa-1}{\kappa+1}, \quad \text{and} \quad q(x^k) - q^\star = \Big(\frac{\kappa-1}{\kappa+1}\Big)^{2k}(q(x^0) - q^\star).$$

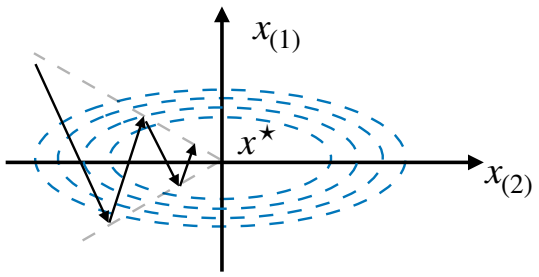

Figure 1: Behavior of γ-`PolyakGD` on the worst-case function

Proof. First, note that the initial point (7) is well-defined provided that $\kappa > \max\{4/\gamma - 1, \gamma/(4-\gamma)\}$ for any $\gamma \in (0, 4)$. We now prove by mathematical induction that the trajectory $\{x^k\}$ generated by γ-`PolyakGD` satisfies the following relationship:

$$(x_{(2)}^k)^2 = \frac{4\kappa^2 - \gamma\kappa(\kappa+1)}{\gamma(\kappa+1) - 4}(x_{(1)}^k)^2, \quad \text{for all } k \geq 0. \tag{8}$$

Base case. The basic case $k = 0$ holds trivially due to the condition (7).

Inductive step. Assuming the case holds for k, we now consider the case $k+1$. Since the optimal function value $q^\star = 0$, the Polyak stepsize at iteration k is given by

$$\alpha_k = \frac{q(x^k)}{\|\nabla q(x^k)\|^2} = \frac{\frac{\kappa}{2}(x_{(1)}^k)^2 + \frac{1}{2}(x_{(2)}^k)^2}{\kappa^2(x_{(1)}^k)^2 + (x_{(2)}^k)^2} = \frac{\frac{\kappa}{2}(x_{(1)}^k)^2 + \frac{1}{2}\frac{4\kappa^2 - \gamma\kappa(\kappa+1)}{\gamma(\kappa+1)-4}(x_{(1)}^k)^2}{\kappa^2(x_{(1)}^k)^2 + \frac{4\kappa^2 - \gamma\kappa(\kappa+1)}{\gamma(\kappa+1)-4}(x_{(1)}^k)^2} = \frac{1}{\gamma}\frac{2}{\kappa+1}.$$

Consequently, the next iterate x^{k+1} is

$$\begin{aligned} x_{(1)}^{k+1} &= x_{(1)}^k - \gamma \cdot \frac{1}{\gamma}\frac{2}{\kappa+1}\kappa x_{(1)}^k = -\frac{\kappa-1}{\kappa+1}x_{(1)}^k, \\ x_{(2)}^{k+1} &= x_{(2)}^k - \gamma \cdot \frac{1}{\gamma}\frac{2}{\kappa+1}x_{(2)}^k = \frac{\kappa-1}{\kappa+1}x_{(2)}^k, \end{aligned}$$

which implies that

$$\left|\frac{x_{(1)}^{k+1}}{x_{(1)}^k}\right| = \left|\frac{x_{(2)}^{k+1}}{x_{(2)}^k}\right| = \frac{\kappa-1}{\kappa+1}. \tag{9}$$

The iterate x^{k+1} thus satisfies the relationship (8), which follows from the inductive hypothesis at iteration k. Therefore, by the principle of mathematical induction, the relationship (9) also holds for each iteration k, and the iterates $\{x^k\}$ converge to $x^\star$ along a fixed trajectory (Figure 1).

Finally, by calculating the function value at iteration $k+1$ with condition (9), we arrive at the following recurrence relation:

$$q(x^k) = \frac{\kappa}{2}(x_{(1)}^{k+1})^2 + \frac{1}{2}(x_{(2)}^{k+1})^2 = \left(\frac{\kappa-1}{\kappa+1}\right)^2\left(\frac{\kappa}{2}(x_{(1)}^k)^2 + \frac{1}{2}(x_{(2)}^k)^2\right) = \left(\frac{\kappa-1}{\kappa+1}\right)^2 q(x^{k-1}).$$

This completes the proof by substituting $q^\star = 0$. □

We now turn to constructing the worst-case function for the class of L-smooth convex functions. Our strategy is to adapt the two-dimensional quadratic function q in Theorem 3.1 via a strongly-convex to convex reduction. Specifically, by setting the condition number κ proportional to K, the

linear convergence (in terms of κ) translates to sublinear convergence (in terms of K). However, this approach makes the gradient Lipschitz constant of q dependent on K, implying that q is not L-smooth. Crucially, we observe that although the Polyak stepsize itself scales inversely with the objective function, the resulting trajectory of γ-`PolyakGD` remains invariant to such scaling. Define a scaled quadratic function

$$\tilde{q}(x) \coloneqq \frac{L}{\kappa} q(x), \tag{10}$$

where $L > 0$ is a constant. Then $\tilde{q}$ is L-smooth by definition, and thus $\tilde{q}$ is an L-smooth convex function. These two components lead to the following theorem.

Theorem 3.2. *Given the parameter $\gamma \in (0, 4)$, let $K \geq 1$ be the total iteration count and $L > 0$ be a constant. Then there exists an L-smooth convex function $\tilde{q}$ and an initial point x^0 such that the trajectory generated by γ-`PolyakGD` satisfies*

$$\min_{1 \leq k \leq K} \frac{\tilde{q}(x^k) - \tilde{q}^\star}{L\|x^0 - x^\star\|^2} \geq \frac{\gamma}{2\mathrm{e}^{2\gamma}((4-\gamma)K + \gamma)},$$

where e *denotes Euler's number.*

Proof. We begin with the same setting as in Theorem 3.1. Suppose the condition number κ satisfies $\kappa > \max\{4/\gamma - 1, \gamma/(4-\gamma)\}$, a requirement that will be justified by our subsequent choice of κ. Without loss of generality, we choose the initial point x^0 on the unit sphere, from which it holds that:

$$1 = (x^0_{(1)})^2 + (x^0_{(2)})^2 = (x^0_{(1)})^2 + \frac{4 - \gamma(\kappa+1)}{\gamma\kappa(\kappa+1) - 4\kappa^2}(x^0_{(1)})^2.$$

Then the initial point x^0 follows as

$$(x^0_{(1)})^2 = \frac{\gamma(\kappa+1) - 4}{(4-\gamma)(\kappa^2-1)} = -\frac{1}{\kappa^2 - 1} + \frac{\gamma\kappa}{(4-\gamma)(\kappa^2-1)}, \tag{11}$$

$$(x^0_{(2)})^2 = 1 - (x^0_{(1)})^2 = \frac{\kappa^2}{\kappa^2 - 1} - \frac{\gamma\kappa}{(4-\gamma)(\kappa^2-1)}, \tag{12}$$

giving

$$q(x^0) = \frac{\kappa}{2}(x^0_{(1)})^2 + \frac{1}{2}(x^0_{(2)})^2 = \frac{2\kappa}{(4-\gamma)(\kappa+1)}.$$

Using the results in Theorem 3.1 yields

$$q(x^K) - q^\star = \left(\frac{\kappa-1}{\kappa+1}\right)^{2K} q(x^0) = \left(\frac{\kappa-1}{\kappa+1}\right)^{2K} \frac{2\kappa}{(4-\gamma)(\kappa+1)}.$$

Consequently, plugging in the definition of $\tilde{q}$, we have

$$\tilde{q}(x^K) - \tilde{q}^\star = \left(\frac{\kappa-1}{\kappa+1}\right)^{2K} \tilde{q}(x^0) = \left(\frac{\kappa-1}{\kappa+1}\right)^{2K} \frac{L}{\kappa} q(x^0) = \left(\frac{\kappa-1}{\kappa+1}\right)^{2K} \frac{2L}{(4-\gamma)(\kappa+1)}.$$

The above equality implies that the function value gap is monotone in this case, and it follows that

$$\tilde{q}(x^K) - \tilde{q}^\star = \min_{1 \leq k \leq K} \tilde{q}(x^k) - \tilde{q}^\star = \left(\frac{\kappa-1}{\kappa+1}\right)^{2K} \frac{2L}{(4-\gamma)(\kappa+1)}.$$

Now we choose $\kappa = 4K/\gamma + \gamma/(4-\gamma) > \max\{4/\gamma - 1, \gamma/(4-\gamma)\}$. Note that $x^\star$ is the origin and $\|x^0\| = 1$. We have

$$\min_{1\le k\le K} \frac{\tilde{q}(x^k) - \tilde{q}^\star}{L\|x^0 - x^\star\|^2} = \left(\frac{8K - 2(1+K)\gamma + \gamma^2}{2K(4-\gamma) + 2\gamma}\right)^{2K} \frac{\gamma}{2((4-\gamma)K + \gamma)}.$$

Therefore, the proof is completed by using technical lemma A.1. □

Remark 3.1. *Our reduction argument is also applicable to the BB stepsize on quadratic functions due to its scale invariance. By leveraging the tight convergence result and the worst-case quadratic function from Li and Sun (2021), a similar $\Omega(1/K)$ result can be established for the BB stepsize.*

We extend the above analysis to the Hölder smooth setting in the following theorem. The proof is deferred to the appendix. Given $\nu \in (0, 1]$, we construct a worst-case function $q_\nu(x) := (\tilde{q}(x))^{\frac{\nu+1}{2}}$, which can be shown to be (ν, L_ν)-Hölder smooth for some constant $L_\nu > 0$.

Theorem 3.3. *Given the parameter $\gamma \in (0, 2)$ and $\nu \in (0, 1]$, let $K \ge 1$ be the total iteration count and $L_\nu > 0$ be a constant. Then there exists a (ν, L_ν)-Hölder smooth convex function q_ν and an initial point x^0 such that the trajectory generated by γ-`PolyakGD` satisfies*

$$\min_{1\le k\le K} \frac{q_\nu(x^k) - q_\nu(x^\star)}{L_\nu\|x^0 - x^\star\|^{\nu+1}} \ge \frac{2^{\nu-1}\gamma^{\frac{\nu+1}{2}}}{\mathrm{e}^{2\gamma}(\nu+1)((2(\nu+1) - \gamma)K + \gamma)^{\frac{\nu+1}{2}}},$$

where e *denotes Euler's number.*

Remark 3.2. *Setting $\nu \to 0$, the above result reduces to the lower bound $\Omega(1/\sqrt{K})$. We remark that this result does not contradict the $\Omega(1/\sqrt[4]{K})$ bound established in Zamani and Glineur (2024). Our tightness result is established for the best iterate, while theirs applies to the last iterate, which is generally a more challenging criterion. Consequently, the worst-case function constructed in Zamani and Glineur (2024) is more complicated. This distinction reveals the significant gap between the best iterate and the last iterate under bounded gradient condition, highlighting the necessity of analyzing both cases.*

Finally, we conclude this subsection by addressing the tightness of the gradient norm convergence. Since our worst-case function $\tilde{q}$ is quadratic, it naturally allows us to establish this result as a byproduct.

Theorem 3.4. *Given the parameter $\gamma \in (0, 2)$ and $\nu \in (0, 1]$, let $K \ge 1$ be the total iteration count and $L_\nu > 0$ be a constant. Then there exists a (ν, L_ν)-Hölder smooth convex function q_ν and an initial point x^0 such that the trajectory generated by γ-`PolyakGD` satisfies*

$$\min_{1\le k\le K} \frac{\|\nabla q_\nu(x^k)\|}{L_\nu\|x^0 - x^\star\|^{\frac{\nu+1}{2}}} \ge \frac{2^{\frac{\nu-1}{2}}\gamma^{\frac{\nu+1}{2}}(\nu+1)^{\frac{\nu-1}{2}}\sqrt{4-\gamma}}{\mathrm{e}^{\gamma\nu}\sqrt{4(4-\gamma)+\gamma^2}(2(\nu+1)-\gamma)^{\frac{\nu}{2}}K^{\frac{\nu}{2}}},$$

where e *denotes Euler's number.*

Remark 3.3. *Establishing the tight gradient norm convergence of constant stepsize gradient descent involves subtle distinctions. Setting $\nu = 1$, the above result reduces to the lower bound $\Omega(1/\sqrt{K})$. However, Nesterov et al. (2018); Diakonikolas and Wang (2021); Teboulle and Vaisbourd (2023) established that small constant stepsizes achieve an upper bound of $\mathcal{O}(1/K)$, a result shown to be tight by Kim and Fessler (2018) via PEP. Using a slightly large constant stepsize slows the convergence to $\mathcal{O}(1/\sqrt{K})$, as shown in Proposition 4.4 in Rotaru et al. (2022) and Theorem 2.1 in Rotaru et al. (2024), a bound that is also tight according to PEP. Therefore, our result suggests that the Polyak stepsize behaves like a "large constant stepsize" in some sense.*

3.2 Polyak stepsize escapes the worst case under floating-point arithmetic

Our tightness convergence results show that `PolyakGD` reduces to gradient descent with a constant stepsize on the worst-case quadratic function. While this conclusion holds under exact arithmetic, it may not persist in practical implementations due to floating-point errors. Indeed, when implementing the Polyak stepsize on the constructed worst-case quadratic function, the numerical behavior differs from the theoretical results. Figure 2 plots the stepsize sequence $\{\alpha_k\}$ and the corresponding function value gaps $\{q(x^k) - q^\star\}$ generated by 1-`PolyakGD`. As predicted by our theory, 1-`PolyakGD` initially behaves like constant stepsize gradient descent. Yet after a few iterations, it departs from the theoretical trajectory, producing larger stepsizes and converging significantly faster. This suggests that under inexact arithmetic, γ-`PolyakGD` (with $\gamma \in (0, 2)$) can automatically exploit floating-point errors to escape the worst-case behavior, which may explain the superior practical performance of the Polyak stepsize. This section formally quantifies this intuition. The techniques developed in this section have also been applied to analyze the hypergradient descent heuristic recently (Chu et al., 2025).

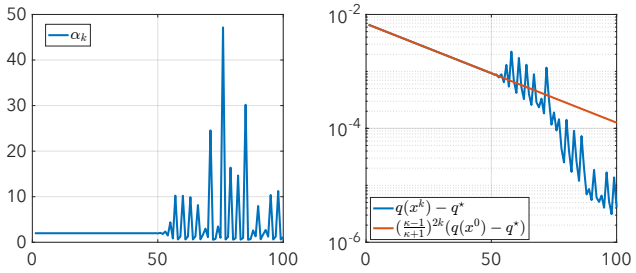

Figure 2: Instability in the presence of floating-point error allows `PolyakGD` to escape the worst-case

Dynamical system formulation of γ-`PolyakGD`. To simplify notation, for this subsection only, we will overload the notation by $z := (x, y) = (x_{(1)}, x_{(2)})$. In this new notation, the worst-case function is $q(z) := \frac{\kappa}{2}x^2 + y^2$. Since q is homogeneous with optimum $z^\star = (0, 0)$, it suffices to consider $\frac{z}{\|z\|}$ to analyze the algorithm trajectory. Our analysis starts by rewriting γ-`PolyakGD` as a nonlinear dynamical system: define the state variable $s = (\frac{x}{\|z\|}, \frac{y}{\|z\|}, \alpha)$. Letting $z^+ = z - \alpha\nabla f(z)$, we have

$$s^+ = \begin{pmatrix} \frac{x^+}{\|z^+\|} \\ \frac{y^+}{\|z^+\|} \\ \alpha^+ \end{pmatrix} = \mathcal{F}_\gamma(s) := \begin{pmatrix} \frac{(1-\alpha\kappa)x}{\sqrt{(1-\alpha\kappa)^2x^2+(1-\alpha)^2y^2}} \\ \frac{(1-\alpha)y}{\sqrt{(1-\alpha\kappa)^2x^2+(1-\alpha)^2y^2}} \\ \frac{\gamma}{2}\frac{\kappa x^2+y^2}{(1-\alpha\kappa)^2\kappa^2x^2+(1-\alpha)^2y^2} \end{pmatrix}. \tag{13}$$

According to Theorem 3.1, when the initial point of γ-`PolyakGD` satisfies Equation (7), the algorithmic trajectory oscillates between two rays. In the language of dynamical systems, there exist two states s_1, s_2 such that

$$s_1 = \mathcal{F}_\gamma(s_2), \qquad s_2 = \mathcal{F}_\gamma(s_1) = \mathcal{F}_\gamma(\mathcal{F}_\gamma(s_2)),$$

and s_1 (or equivalently s_2) is a period-2 orbit of $\mathcal{F}_\gamma$ (Alligood et al., 1997). To understand the local behavior of γ-`PolyakGD` in the presence of floating-point errors, it is natural to examine the stability of this nonlinear map around its orbit. This reduces to analyzing the spectrum of the Jacobian of $\mathcal{F}_\gamma$. The proof of Proposition 3.2 is assisted by `Mathematica` and the source code is available at this link.

Proposition 3.2. *Suppose* $\gamma \in (0,2]$ *and* $\kappa > \max\{4\gamma^{-1}-1, 3+2\sqrt{2}\}$. *Then the nonlinear dynamical system* (13) *has period-two orbits given by*

$$\mathcal{S} = \left\{(s_1,s_2) : s_1 = \begin{pmatrix} x \\ y \\ \frac{2}{\kappa+1} \end{pmatrix}, s_2 = \begin{pmatrix} -x \\ y \\ \frac{2}{\kappa+1} \end{pmatrix}, x^2 = \frac{\gamma(\kappa+1)-4}{(\gamma-4)(\kappa^2-1)}, y^2 = \frac{\kappa(\gamma-4\kappa+\gamma\kappa)}{(\gamma-4)(\kappa^2-1)}\right\}.$$

Moreover, the Jacobian of the dynamical system is given by

$$\mathcal{J}_{\mathcal{F}_\gamma}(s) = \begin{pmatrix} -\frac{(\alpha-1)^2y^2(\alpha\kappa-1)}{((x-\alpha\kappa x)^2+(y-\alpha y)^2)^{3/2}} & \frac{(\alpha-1)^2xy(\alpha\kappa-1)}{((x-\alpha\kappa x)^2+(y-\alpha y)^2)^{3/2}} & \frac{(\alpha-1)(\kappa-1)xy^2}{((x-\alpha\kappa x)^2+(y-\alpha y)^2)^{3/2}} \\ \frac{(\alpha-1)xy(\alpha\kappa-1)^2}{((x-\alpha\kappa x)^2+(y-\alpha y)^2)^{3/2}} & -\frac{(\alpha-1)x^2(\alpha\kappa-1)^2}{((x-\alpha\kappa x)^2+(y-\alpha y)^2)^{3/2}} & -\frac{(\kappa-1)x^2y(\alpha\kappa-1)}{((x-\alpha\kappa x)^2+(y-\alpha y)^2)^{3/2}} \\ -\frac{(\alpha-1)^2\gamma(\kappa-1)\kappa xy^2(\alpha\kappa-1)^2}{(\kappa^2x^2(\alpha\kappa-1)^2+(\alpha-1)^2y^2)^2} & \frac{(\alpha-1)^2\gamma(\kappa-1)\kappa x^2y(\alpha\kappa-1)^2}{(\kappa^2x^2(\alpha\kappa-1)^2+(\alpha-1)^2y^2)^2} & \frac{(\alpha-1)\gamma(\kappa-1)^2\kappa x^2y^2(\alpha\kappa-1)}{(\kappa^2x^2(\alpha\kappa-1)^2+(\alpha-1)^2y^2)^2} \end{pmatrix}.$$

In particular, for any $(s_1,s_2) \in \mathcal{S}$, *the spectral radius of the product Jacobians satisfy*

$$\rho(\mathcal{J}_{\mathcal{F}_1}(s_1)\mathcal{J}_{\mathcal{F}_1}(s_2)) = \frac{4(\kappa^2-4\kappa+1)^2}{(\kappa-1)^4} > 1 \quad \textit{and} \quad \rho(\mathcal{J}_{\mathcal{F}_2}(s_1)\mathcal{J}_{\mathcal{F}_2}(s_2)) = 1.$$

Proposition 3.2 shows that when $\gamma = 1$ and $\kappa > 6$, the spectral radius of the product Jacobian is strictly greater than 1. Hence, the nonlinear dynamical system (13) is unstable around the trajectory induced by the worst-case function. Consequently, once floating-point errors enter the iterations, γ-`PolyakGD` deviates from this trajectory and exhibits accelerated convergence. Similar observation holds for all $\gamma \in (0,2)$. In contrast, when $\gamma = 2$, the spectral radius is exactly 1, implying that 2-`PolyakGD` may become trapped on the worst-case trajectory. It is consistent with the zig-zag behavior observed in Huang and Qi (2024). In summary, the worst-case trajectory is only realized under exact arithmetic, and `PolyakGD` in practice is largely immune to it.

4 Universal Convergence of γ-`PolyakGD` Across Function Classes

In this section, we demonstrate that γ-`PolyakGD` is universal across various function classes, following the terminology "universal" introduced in Nesterov (2015). We first establish convergence under Hölder smoothness and growth conditions, and then provide several extensions, such as relaxations of convexity and the stochastic Polyak stepsize.

4.1 Convergence results under Hölder smoothness and growth

Before presenting the main convergence result, we justify the choice of the set $\mathcal{K}$ for the Hölder growth condition in (5). Given the parameter $\gamma \in (0,2]$, first note that for any iteration $k \geq 0$ and any $x^\star \in \mathcal{X}^\star$, the distance to $x^\star$ satisfies

$$\begin{aligned} \|x^{k+1}-x^\star\|^2 &= \|x^k - \gamma\cdot\alpha_k\nabla f(x^k) - x^\star\|^2 \\ &= \|x^k-x^\star\|^2 - 2\gamma\cdot\alpha_k\langle\nabla f(x^k), x^k-x^\star\rangle + \gamma^2\cdot\alpha_k^2\|\nabla f(x^k)\|^2 \\ &\leq \|x^k-x^\star\|^2 - 2\gamma\cdot\alpha_k(f(x^k)-f^\star) + \gamma^2\cdot\alpha_k^2\|\nabla f(x^k)\|^2 \\ &= \|x^k-x^\star\|^2 - \gamma(2-\gamma)\frac{(f(x^k)-f^\star)^2}{\|\nabla f(x^k)\|^2}. \end{aligned} \tag{14}$$

This inequality implies that the iterate sequence generated by γ-`PolyakGD` is Fejér monotone. Accordingly, we define

$$\mathcal{K}(x^\star) \coloneqq \mathbb{B}(x^\star, \|x^0-x^\star\|), \text{ and } \mathcal{K} \coloneqq \bigcup_{x^\star\in\mathcal{X}^\star}\mathcal{K}(x^\star), \tag{15}$$

where $\mathbb{B}(x^\star, \|x^0 - x^\star\|)$ denotes the closed ball centered at $x^\star$ with radius $\|x^0 - x^\star\|$. Clearly, we have $\mathcal{X}^\star \subseteq \bigcup_{x^\star \in \mathcal{X}^\star} \mathcal{K}(x^\star)$ and $\sup_{x \in \mathcal{K}} \operatorname{dist}(x, \mathcal{X}^\star) < \infty$, satisfying the requirement discussed in Section 2.

Theorem 4.1. *Let $\gamma \in (0, 2)$ and $x^0 \in \mathbb{R}^n$. Suppose that the objective function f satisfies the (ν, L_ν)-Hölder smoothness condition (Definition 2.1) and the (r, ρ_r)-Hölder growth condition (Definition 2.2) on $\mathcal{K}$ defined in (15). Then, for any (even) iteration count $K \geq 1$, the iterates generated by γ-`PolyakGD` satisfy:*

- *When $r = \nu + 1$,*

$$\min_{1 \leq k \leq K} f(x^k) - f^\star \leq \left(1 - \frac{\gamma(2-\gamma)\nu^{\frac{2\nu}{\nu+1}} \rho_r^{\frac{2}{\nu+1}}}{(\nu+1)^{\frac{2\nu}{\nu+1}} L_\nu^{\frac{2}{\nu+1}}}\right)^{\frac{K}{2}} \sup_{x \in \mathcal{K}} \|\nabla f(x)\| \cdot \operatorname{dist}(x^0, \mathcal{X}^\star),$$

- *When $r > \nu + 1$,*

$$\min_{1 \leq k \leq K} f(x^k) - f^\star \leq \frac{(r-\nu-1)^{\frac{(\nu+1)^2}{2(\nu+1-r)}} (\gamma(2-\gamma))^{\frac{r(\nu+1)}{2(\nu+1-r)}} \nu^{\frac{\nu r}{\nu+1-r}}}{2^{\frac{r(\nu+1)}{2(\nu+1-r)}} (\nu+1)^{\frac{(\nu+1)^2+2\nu r}{2(\nu+1-r)}}} \cdot \frac{L_\nu^{\frac{r}{r-\nu-1}}}{\rho_r^{\frac{\nu+1}{r-\nu-1}} K^{\frac{r(\nu+1)}{2(r-\nu-1)}}}.$$

Proof. Let $\mathrm{D}_k \coloneqq \operatorname{dist}(x^k, \mathcal{X}^\star)$ and $\Delta_k \coloneqq f(x^k) - f^\star$ for short. From Equation (14), we obtain

$$\mathrm{D}_{k+1}^2 \leq \|x^{k+1} - x^\star\|^2 \leq \|x^k - x^\star\|^2 - \gamma(2-\gamma)\frac{\Delta_k^2}{\|\nabla f(x^k)\|^2}.$$

Since the right-hand side in the above inequality holds for all $x^\star \in \mathcal{X}^\star$, we derive

$$\mathrm{D}_{k+1}^2 \leq \mathrm{D}_k^2 - \gamma(2-\gamma)\frac{\Delta_k^2}{\|\nabla f(x^k)\|^2}. \tag{16}$$

Using the second inequality in Equation (4) yields

$$\mathrm{D}_{k+1}^2 \leq \mathrm{D}_k^2 - \frac{\gamma(2-\gamma)\nu^{\frac{2\nu}{\nu+1}}}{(\nu+1)^{\frac{2\nu}{\nu+1}} L_\nu^{\frac{2}{\nu+1}}} \Delta_k^{\frac{2}{\nu+1}}.$$

Now we substitute the Hölder growth condition to get

$$\mathrm{D}_{k+1}^2 \leq \mathrm{D}_k^2 - \frac{\gamma(2-\gamma)\nu^{\frac{2\nu}{\nu+1}} \rho_r^{\frac{2}{\nu+1}}}{(\nu+1)^{\frac{2\nu}{\nu+1}} L_\nu^{\frac{2}{\nu+1}}} \mathrm{D}_k^{\frac{2r}{\nu+1}}. \tag{17}$$

Case 1. $r = \nu + 1$. the above equation reduces to

$$\mathrm{D}_{k+1}^2 \leq \mathrm{D}_k^2 - \frac{\gamma(2-\gamma)\nu^{\frac{2\nu}{\nu+1}} \rho_r^{\frac{2}{\nu+1}}}{(\nu+1)^{\frac{2\nu}{\nu+1}} L_\nu^{\frac{2}{\nu+1}}} \mathrm{D}_k^2.$$

Note that in this case, Equation (6) implies $\frac{\rho_r}{L_\nu} \leq \frac{1}{\nu+1} \leq 1$. Hence, the linear convergence follows

$$\mathrm{D}_{k+1}^2 \leq \left(1 - \frac{\gamma(2-\gamma)\nu^{\frac{2\nu}{\nu+1}} \rho_r^{\frac{2}{\nu+1}}}{(\nu+1)^{\frac{2\nu}{\nu+1}} L_\nu^{\frac{2}{\nu+1}}}\right) \mathrm{D}_k^2,$$

as these constants satisfy $\gamma(2-\gamma) < 1$, $\nu/(\nu+1) < 1$ and $\rho_r/L_v \leq 1$. By convexity, it holds that $\Delta_k \leq \langle \nabla f(x^k), x^k - x^\star \rangle \leq G\,\mathrm{D}_k$, for all $k \geq 0$, where $G \coloneqq \sup_{x \in \mathcal{K}} \|\nabla f(x)\|$. Therefore, we conclude

$$\min_{1 \leq k \leq K} f(x^k) - f^\star \leq \left(1 - \frac{\gamma(2-\gamma)\nu^{\frac{2\nu}{\nu+1}}\rho_r^{\frac{2}{\nu+1}}}{(\nu+1)^{\frac{2\nu}{\nu+1}}L_\nu^{\frac{2}{\nu+1}}}\right)^{\frac{K}{2}} G\,\mathrm{D}_0, \quad \text{for all } K \geq 1.$$

Case 2. $r > \nu + 1$. From Equation (17), we use the technical lemma A.2 in the Appendix (by setting $a_k = \mathrm{D}_k^2$, $\tau = \frac{r}{\nu+1} > 1$, $c = \frac{\gamma(2-\gamma)\nu^{\frac{2\nu}{\nu+1}}\rho_r^{\frac{2}{\nu+1}}}{(\nu+1)^{\frac{2\nu}{\nu+1}}L_\nu^{\frac{2}{\nu+1}}}$.) It immediately follows

$$\begin{aligned}
\mathrm{D}_k^2 &\leq \left(\mathrm{D}_0^{\frac{2(\nu+1-r)}{\nu+1}} + \frac{r-\nu-1}{\nu+1} \cdot \frac{\gamma(2-\gamma)\nu^{\frac{2\nu}{\nu+1}}\rho_r^{\frac{2}{\nu+1}}}{(\nu+1)^{\frac{2\nu}{\nu+1}}L_\nu^{\frac{2}{\nu+1}}} k\right)^{\frac{\nu+1}{\nu+1-r}} \\
&\leq \left(\frac{r-\nu-1}{\nu+1} \cdot \frac{\gamma(2-\gamma)\nu^{\frac{2\nu}{\nu+1}}\rho_r^{\frac{2}{\nu+1}}}{(\nu+1)^{\frac{2\nu}{\nu+1}}L_\nu^{\frac{2}{\nu+1}}} k\right)^{\frac{\nu+1}{\nu+1-r}}.
\end{aligned}$$

Now we combine Equation (16) with the second inequality in Equation (4), and it follows

$$\gamma(2-\gamma)\Delta_k^{\frac{2}{\nu+1}} \leq \frac{(\nu+1)^{\frac{2\nu}{\nu+1}}L_\nu^{\frac{2}{\nu+1}}}{\nu^{\frac{2\nu}{\nu+1}}}(\mathrm{D}_k^2 - \mathrm{D}_{k+1}^2).$$

Without loss of generality, we choose K as an even number such that $K \geq 2$. Then we get

$$\begin{aligned}
\frac{1}{K/2}\sum_{k=K/2}^{K-1}\gamma(2-\gamma)\Delta_k^{\frac{2}{\nu+1}} &\leq \frac{2(\nu+1)^{\frac{2\nu}{\nu+1}}L_\nu^{\frac{2}{\nu+1}}}{\nu^{\frac{2\nu}{\nu+1}}K}\sum_{k=K/2}^{K-1}(\mathrm{D}_k^2 - \mathrm{D}_{k+1}^2) \\
&\leq \frac{2(\nu+1)^{\frac{2\nu}{\nu+1}}L_\nu^{\frac{2}{\nu+1}}}{\nu^{\frac{2\nu}{\nu+1}}K}\mathrm{D}_{K/2}^2 \\
&\leq \frac{2(\nu+1)^{\frac{2\nu}{\nu+1}}L_\nu^{\frac{2}{\nu+1}}}{\nu^{\frac{2\nu}{\nu+1}}K}\left(\frac{r-\nu-1}{\nu+1} \cdot \frac{\gamma(2-\gamma)\nu^{\frac{2\nu}{\nu+1}}\rho_r^{\frac{2}{\nu+1}}}{2(\nu+1)^{\frac{2\nu}{\nu+1}}L_\nu^{\frac{2}{\nu+1}}}K\right)^{\frac{\nu+1}{\nu+1-r}}.
\end{aligned}$$

This implies

$$\begin{aligned}
\min_{1 \leq k \leq K}(f(x^k) - f^\star)^{\frac{2}{\nu+1}} &\leq \frac{2(\nu+1)^{\frac{2\nu}{\nu+1}}L_\nu^{\frac{2}{\nu+1}}}{\gamma(2-\gamma)\nu^{\frac{2\nu}{\nu+1}}K}\left(\frac{r-\nu-1}{\nu+1} \cdot \frac{\gamma(2-\gamma)\nu^{\frac{2\nu}{\nu+1}}\rho_r^{\frac{2}{\nu+1}}}{2(\nu+1)^{\frac{2\nu}{\nu+1}}L_\nu^{\frac{2}{\nu+1}}}K\right)^{\frac{\nu+1}{\nu+1-r}} \\
&= \frac{(r-\nu-1)^{\frac{\nu+1}{\nu+1-r}}(\gamma(2-\gamma))^{\frac{r}{\nu+1-r}}\nu^{\frac{2\nu r}{(\nu+1)(\nu+1-r)}}}{2^{\frac{r}{\nu+1-r}}(\nu+1)^{\frac{(\nu+1)^2+2\nu r}{(\nu+1)(\nu+1-r)}}} \cdot \frac{L_\nu^{\frac{2r}{(\nu+1)(r-\nu-1)}}}{\rho_r^{\frac{2}{r-\nu-1}}K^{\frac{r}{r-\nu-1}}}.
\end{aligned}$$

Therefore, we conclude

$$\min_{1 \leq k \leq K} f(x^k) - f^\star \leq \frac{(r-\nu-1)^{\frac{(\nu+1)^2}{2(\nu+1-r)}}(\gamma(2-\gamma))^{\frac{r(\nu+1)}{2(\nu+1-r)}}\nu^{\frac{\nu r}{\nu+1-r}}}{2^{\frac{r(\nu+1)}{2(\nu+1-r)}}(\nu+1)^{\frac{(\nu+1)^2+2\nu r}{2(\nu+1-r)}}} \cdot \frac{L_\nu^{\frac{r}{r-\nu-1}}}{\rho_r^{\frac{\nu+1}{r-\nu-1}}K^{\frac{r(\nu+1)}{2(r-\nu-1)}}}.$$

The proof is completed. □

Remark 4.1. *The above theorem demonstrates that γ-PolyakGD adapts to both Hölder smoothness and Hölder growth conditions. Specifically, letting $r \to \infty$, the above result reduces to $\mathcal{O}(1/K^{(\nu+1)/2})$, which matches the rate achieved by the universal gradient method in Nesterov (2015). This result is shown to be tight in Theorem 3.3. On the other hand, setting $\nu = 0$, the above result reduces to $\mathcal{O}(1/K^{r/(2(r-1))})$, which matches the lower bound (Nemirovskii and Nesterov, 1985). This rate is also achieved by restart-based optimization methods (Roulet and d'Aspremont, 2017; Yang and Lin, 2018; Xu, 2019; d'Aspremont et al., 2021; Renegar and Grimmer, 2022; Grimmer, 2023).*

4.2 Extensions

In this subsection, we extend the previous analysis to more general settings.

Star-convexity. The proof of Theorem 4.1 relies on the convexity of f to obtain

$$\langle \nabla f(x^k), x^k - x^\star \rangle \geq f(x^k) - f^\star$$

in Equation (14). This inequality corresponds to a weaker condition known as star-convexity (see Definition 4.1 in Nesterov et al. (2018), for example). Therefore, the convexity assumption can naturally be relaxed to star-convexity. For example, when $f(x) := h(g(x))$ is the composition between an outer scalar function $h : \mathbb{R} \to \mathbb{R}$ and a function $g : \mathbb{R}^n \to \mathbb{R}$, f can be shown to be star-convex under mild conditions of h and g.

Lemma 4.1. *Suppose the scalar function h is nondecreasing and convex, and the function g is differentiable and (star-)convex, i.e., $\langle \nabla g(x), x - x_g^\star \rangle \geq g(x) - g(x_g^\star)$, where $x_g^\star \in \mathcal{X}_g^\star := \operatorname{argmin}_{x \in \mathbb{R}^n} g(x)$. For any $x_g^\star \in \mathcal{X}_g^\star$, the objective function f satisfies $\langle \xi(x) \nabla g(x), x - x_g^\star \rangle \geq f(x) - f(x_g^\star)$, for all $x \in \mathbb{R}^n$, and $\xi(x) \in \partial h(g(x))$.*

This structure is common in machine learning optimization tasks (Chen et al., 2021; Jiang et al., 2022), where the function f itself may be nonconvex and nonsmooth. Star-convexity suffices to ensure that Theorem 4.1 remains valid. The only modification is that we need to use the subgradient of f in γ-PolyakGD (Algorithm 1).

Convergence of 2-PolyakGD. Note that Theorem 4.1 is not applicable to the case $\gamma = 2$. To establish a convergence guarantee for this case, we introduce the following inequality. This inequality is an extension of inequality (2.1.10)[4] in Nesterov et al. (2018) to Hölder smoothness.

Lemma 4.2. *Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is (L_ν, ν)-Hölder smooth with $\nu \in (0, 1]$ and $L_\nu > 0$, then it holds that*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\nu}{\nu + 1} \frac{1}{L_\nu^{\frac{1}{\nu}}} \|\nabla f(x) - \nabla f(y)\|^{\frac{\nu+1}{\nu}}, \quad \text{for all } x, y \in \mathbb{R}^n.$$

Theorem 4.2. *Suppose the objective function f satisfies (ν, L_ν)-Hölder smoothness. For any total iteration count $K \geq 1$, the iterates generated by γ-PolyakGD with $\gamma = 2$ satisfy*

$$\min_{1 \leq k \leq K} f(x^k) - f^\star \leq \left(\frac{\nu + 1}{4\nu}\right)^\nu \frac{L_\nu \operatorname{dist}(x^0, \mathcal{X}^\star)^{\nu+1}}{K^\nu}.$$

[4] This inequality is also known as cocoercivity in the literature (see, e.g., Malitsky and Mishchenko (2024); Latafat et al. (2024)).

Remark 4.2. *When $\gamma = 2$, the upper bound $\mathcal{O}(1/K^\nu)$ is slightly weaker than the result $\mathcal{O}(1/K^{(\nu+1)/2})$ in Theorem 4.1 for $\nu \in (0, 1]$. In the case $\nu = 1$ (i.e., L-smoothness), the two bounds coincide. We argue that the rate $\mathcal{O}(1/K^\nu)$ is reasonable because when $\nu \to 0$ (the nonsmooth case), 2-PolyakGD may not converge. For example, consider the absolute value function $|x|$ with initial point $x^0 = 1$; 2-PolyakGD oscillates between the two points -1 and 1.*

As a byproduct, the tighter inequality in Lemma 4.2 can also be used to establish an $\mathcal{O}(1/K^{\nu/2})$ convergence rate for the gradient norm. According to Theorem 3.4, this upper bound is tight.

Theorem 4.3. *Suppose the objective function f satisfies (ν, L_ν)-Hölder smoothness. For any total iteration count $K \geq 1$, the iterates generated by γ-PolyakGD with $\gamma \in (0, 2]$ satisfy*

$$\min_{1 \leq k \leq K} \|\nabla f(x^k)\| \leq \frac{(\nu+1)^\nu L_\nu \operatorname{dist}(x^0, \mathcal{X}^\star)^\nu}{2^{\frac{\nu}{2}} \gamma^{\frac{\nu}{2}} \nu^\nu K^{\frac{\nu}{2}}}.$$

Global curvature bound. We now consider a more general upper curvature bound than Hölder smoothness, known as the global curvature bound (Nesterov, 2025; Doikov and Nesterov, 2025). This bound is intrinsic as it does not require the objective function f to belong to any specific function class; it is defined by the following quantity:

$$\hat{\mu}_f(t) \coloneqq \sup_{\substack{\|x-y\| \leq t, \\ \alpha \in (0,1)}} \frac{|\alpha f(x) + (1-\alpha) f(y) - f(\alpha x + (1-\alpha) y)|}{\alpha(1-\alpha)}. \tag{18}$$

Given $\hat{\mu}_f$, its inverse function s_f, known as the complexity gauge, corresponds to a stationary criterion for f (see Definition 1 in Nesterov (2025)). Specifically, it satisfies $\hat{\mu}_f(s_f(t)) = t$ for all $t \geq 0$. The following theorem shows that γ-PolyakGD also adapts to this global curvature bound, matching the result of the universal primal gradient method in Nesterov (2025).

Theorem 4.4. *Let $\gamma \in (0, 2)$ and $x^0 \in \mathbb{R}^n$. Suppose $\hat{\mu}_f$ is well-defined for any $t \in [0, \infty)$ and satisfies $\lim_{t \to 0} \hat{\mu}(t)/t = 0$. For any total iteration count $K \geq 1$, the iterates generated by γ-PolyakGD satisfy*

$$\min_{1 \leq k \leq K} f(x^k) - f^\star \leq \hat{\mu}_f\left(\frac{3 \operatorname{dist}(x^0, \mathcal{X}^\star)^2}{\sqrt{K}}\right).$$

Consequently, given an accuracy $\varepsilon > 0$, γ-PolyakGD guarantees $\min_{1 \leq k \leq K} f(x^k) - f^\star \leq \varepsilon$ after at most $K = \mathcal{O}(\operatorname{dist}(x^0, \mathcal{X}^\star)^2 / s_f^2(\varepsilon))$ iterations.

Stochastic Polyak stepsize. To conclude this section, we extend γ-PolyakGD to stochastic Polyak stepsize (Loizou et al., 2021). Consider the following stochastic optimization problem

$$\min_{x \in \mathbb{R}^n} f(x) \coloneqq \mathbb{E}_\xi[f(x, \xi)],$$

where f is assumed to satisfy the interpolation condition: $\min_x \mathbb{E}[f(x, \xi)] - \mathbb{E}[\min_x f(x, \xi)] = 0$. The stochastic γ-PolyakGD update is given by

$$x^{k+1} = x^k - \gamma \cdot \frac{f(x^k, \xi^k) - \min_x f(x, \xi^k)}{\|\nabla f(x^k, \xi^k)\|^2} \nabla f(x^k, \xi^k),$$

where $\mathbb{E}[\nabla f(x^k, \xi^k) | x^k] = \nabla f(x^k)$. The convergence guarantees of stochastic γ-PolyakGD resemble those of the deterministic case.

Theorem 4.5. *Suppose the interpolation condition holds and that $\mathcal{X}^\star = \{x^\star\}$ is a singleton. Then stochastic γ-PolyakGD has the same rates as Theorem 4.1 in terms of $\mathbb{E}[f(x^k) - f^\star]$.*

5 Conclusions

This work revisits the classical Polyak stepsize and establishes several new results. We establish the tightness of the current convergence results and provide new convergence rates for new function classes. An interesting future direction is how to construct a universal worst-case function for general adaptive stepsizes, beyond just the Polyak stepsize.

References

K. T. Alligood, T. D. Sauer, and J. A. Yorke. *Chaos: an introduction to dynamical systems*. Springer, 1997. 9

J. M. Altschuler and P. A. Parrilo. Acceleration by stepsize hedging: Multi-step descent and the silver stepsize schedule. *Journal of the ACM*, 72(2):1–38, 2025. 2

M. Barré, A. Taylor, and A. d'Aspremont. Complexity guarantees for Polyak steps with momentum. In *Conference on Learning Theory*, pages 452–478. PMLR, 2020. 2

S. Boyd. Subgradient methods. *Notes for EE364b, Stanford University, Spring 2013–14*, 2013. 2

T. Chen, Y. Sun, and W. Yin. Solving stochastic compositional optimization is nearly as easy as solving stochastic optimization. *IEEE Transactions on Signal Processing*, 69:4937–4948, 2021. 13

Y.-C. Chu, W. Gao, Y. Ye, and M. Udell. Gradient methods with online scaling. Part II: Practical aspects. *arXiv preprint arXiv:2509.11007*, 2025. 9

C. Criscitiello and N. Boumal. Curvature and complexity: Better lower bounds for geodesically convex optimization. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2969–3013. PMLR, 2023. 2

Q. Deng, G. Lan, and Z. Lin. Uniformly optimal and parameter-free first-order methods for convex and function-constrained optimization. *arXiv preprint arXiv:2412.06319*, 2024. 1

N. Devanathan and S. Boyd. Polyak minorant method for convex optimization. *Journal of Optimization Theory and Applications*, 203(3):2263–2282, 2024. 1

J. Diakonikolas and P. Wang. Potential function-based framework for making the gradients small in convex and min-max optimization. *arXiv preprint arXiv:2101.12101*, 2021. 8

N. Doikov and Y. Nesterov. Universal reduced-operator method and high-order global curvature bounds. *arXiv preprint arXiv:2511.07341*, 2025. 14

Y. Drori and M. Teboulle. Performance of first-order methods for smooth convex minimization: a novel approach. *Mathematical Programming*, 145(1):451–482, 2014. 2, 5

D. Drusvyatskiy. Convex analysis and nonsmooth optimization. *University Lecture*, page 40, 2020. 24

A. d'Aspremont, D. Scieur, A. Taylor, et al. Acceleration methods. *Foundations and Trends® in Optimization*, 5(1-2):1–245, 2021. 13

W. Gao, Y.-C. Chu, Y. Ye, and M. Udell. Gradient methods with online scaling. *arXiv preprint arXiv:2411.01803*, 2024. 2

B. Goujaud, A. Dieuleveut, and A. Taylor. On fundamental proof structures in first-order optimization. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pages 3023–3030. IEEE, 2023. 2

B. Goujaud, C. Moucer, F. Glineur, J. M. Hendrickx, A. B. Taylor, and A. Dieuleveut. PEPit: computer-assisted worst-case analyses of first-order optimization methods in python. *Mathematical Programming Computation*, 16(3):337–367, 2024. 2

B. Grimmer. General Hölder smooth convergence rates follow from specialized rates assuming growth bounds. *Journal of Optimization Theory and Applications*, 197(1):51–70, 2023. 13

B. Grimmer. Provably faster gradient descent via long steps. *SIAM Journal on Optimization*, 34 (3):2588–2608, 2024. 2

E. Hazan and S. Kakade. Revisiting the Polyak stepsize. *arXiv preprint arXiv:1905.00313*, 2019. 1, 2, 3, 4, 19

C. He, B. Jiang, Y. Jiang, C. Zhang, and S. Zhang. History-aware adaptive high-order tensor regularization. *arXiv preprint arXiv:2511.05788*, 2025. 2

Y.-K. Huang and H.-D. Qi. Analytic analysis of the worst-case complexity of the gradient method with exact line search and the Polyak stepsize. *arXiv preprint arXiv:2407.04914*, 2024. 2, 10

W. Jiang, B. Wang, Y. Wang, L. Zhang, and T. Yang. Optimal algorithms for stochastic multi-level compositional optimization. In *International Conference on Machine Learning*, pages 10195–10216. PMLR, 2022. 13

D. Kim and J. A. Fessler. Generalizing the optimized gradient method for smooth convex minimization. *SIAM Journal on Optimization*, 28(2):1920–1950, 2018. 8

P. Latafat, A. Themelis, L. Stella, and P. Patrinos. Adaptive proximal algorithms for convex optimization under local lipschitz continuity of the gradient. *Mathematical Programming*, pages 1–39, 2024. 13

D. Li and R. Sun. On a faster R-linear convergence rate of the Barzilai-Borwein method. *arXiv preprint arXiv:2101.00205*, 2021. 8

T. Li and G. Lan. A simple uniformly optimal method without line search for convex optimization. *Mathematical Programming*, pages 1–38, 2025. 2

N. Loizou, S. Vaswani, I. H. Laradji, and S. Lacoste-Julien. Stochastic Polyak stepsize for sgd: An adaptive learning rate for fast convergence. In *International Conference on Artificial Intelligence and Statistics*, pages 1306–1314. PMLR, 2021. 1, 2, 14

Y. Malitsky and K. Mishchenko. Adaptive gradient descent without descent. In *International Conference on Machine Learning*, pages 6702–6712. PMLR, 2020. 2

Y. Malitsky and K. Mishchenko. Adaptive proximal gradient method for convex optimization. *Advances in Neural Information Processing Systems*, 37:100670–100697, 2024. 13

A. Mishkin, A. Khaled, Y. Wang, A. Defazio, and R. Gower. Directional smoothness and gradient methods: Convergence and adaptivity. *Advances in Neural Information Processing Systems*, 37: 14810–14848, 2024. 2

A. S. Nemirovskii and Y. E. Nesterov. Optimal methods of smooth convex minimization. *USSR Computational Mathematics and Mathematical Physics*, 25(2):21–30, 1985. 3, 13

Y. Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152(1):381–404, 2015. 3, 10, 13

Y. Nesterov. Universal complexity bounds for universal gradient methods in nonlinear optimization. *arXiv preprint arXiv:2509.20902*, 2025. 3, 14, 24

Y. Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018. 8, 13

F. Orabona. Normalized gradients for all. *arXiv preprint arXiv:2308.05621*, 2023. 4

F. Orabona and R. D'Orazio. New perspectives on the Polyak stepsize: Surrogate functions and negative results. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=7GwcxPIkip. 2

B. T. Polyak. Minimization of unsmooth functionals. *USSR Computational Mathematics and Mathematical Physics*, 9(3):14–29, 1969. 1, 2, 4

B. T. Polyak. Introduction to optimization. 1987. 1, 2

J. Renegar and B. Grimmer. A simple nearly optimal restart scheme for speeding up first-order methods. *Foundations of computational mathematics*, 22(1):211–256, 2022. 13

A. Rodomanov and Y. Nesterov. Smoothness parameter of power of euclidean norm. *Journal of Optimization Theory and Applications*, 185(2):303–326, 2020. 19

T. Rotaru, F. Glineur, and P. Panagiotis. Tight convergence rates of the gradient method on hypoconvex functions. *arXiv preprint arXiv:2203.00775*, 2022. 8

T. Rotaru, F. Glineur, and P. Patrinos. Exact worst-case convergence rates of gradient descent: a complete analysis for all constant stepsizes over nonconvex and convex functions. *arXiv preprint arXiv:2406.17506*, 2024. 8

V. Roulet and A. d'Aspremont. Sharpness, restart and acceleration. *Advances in Neural Information Processing Systems*, 30, 2017. 5, 13

A. B. Taylor, J. M. Hendrickx, and F. Glineur. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 161(1):307–345, 2017. 2

M. Teboulle and Y. Vaisbourd. An elementary approach to tight worst-case complexity analysis of gradient based methods. *Mathematical Programming*, 201(1):63–96, 2023. 5, 8

X. Wang, M. Johansson, and T. Zhang. Generalized Polyak stepsize for first-order optimization with momentum. In *International Conference on Machine Learning*, pages 35836–35863. PMLR, 2023. 1

Y. Xu. *Accelerating convex optimization in machine learning by leveraging functional growth conditions*. The University of Iowa, 2019. 13

T. Yang and Q. Lin. RSG: Beating subgradient method without smoothness and strong convexity. *Journal of Machine Learning Research*, 19(6):1–33, 2018. 13

M. Zamani and F. Glineur. Exact convergence rate of the subgradient method by using Polyak stepsize. *arXiv preprint arXiv:2407.15195*, 2024. 1, 2, 8

M. Zamani and F. Glineur. Exact convergence rate of the last iterate in subgradient methods. *SIAM Journal on Optimization*, 35(3):2182–2201, 2025. 2

Z. Zheng, N. S. Aybat, S. Ma, and L. Xue. Adaptive algorithms for robust phase retrieval. *arXiv preprint arXiv:2409.19162*, 2024. 2

Appendix

A Technical Lemmas

Lemma A.1. *For any $\gamma \in (0,4)$, the following relation holds for all $K \geq 1$:*

$$\left(\frac{8K - 2(1+K)\gamma + \gamma^2}{2K(4-\gamma)+2\gamma}\right)^{2K} \geq \left(\frac{K}{\gamma+K}\right)^{2K} \geq \lim_{K\to\infty}\left(\frac{K}{\gamma+K}\right)^{2K} = \mathrm{e}^{-2\gamma}.$$

Proof. The proof is available at this link. □

Lemma A.2. *Let $\{a_k\}_{k\geq 0}$ be a nonnegative sequence satisfying $a_{k+1} \leq a_k - c \cdot a_k^\tau$ for some $c > 0$ and $\tau > 1$. Then it holds that $a_k \leq \left(a_0^{1-\tau} + (\tau-1)ck\right)^{-\frac{1}{\tau-1}}$.*

Proof. Let $\phi(t) \coloneqq t^{1-\tau}$, $t > 0$. Note that ϕ is strictly decreasing over $(0,\infty)$ when $\tau > 1$ since $\phi'(t) = (1-\tau)t^{-\tau} < 0$. Fix $k \geq 0$. Recall the sequence satisfies $a_k - a_{k+1} \geq c \cdot a_k^\tau > 0$. Then by Lagrange's mean value theorem, there exists a value $\xi_k \in [a_{k+1}, a_k]$ such that

$$a_{k+1}^{1-\tau} - a_k^{1-\tau} = \phi(a_{k+1}) - \phi(a_k) = \phi'(\xi_k)(a_{k+1} - a_k) = (\tau-1)\xi_k^{-\tau}(a_k - a_{k+1}).$$

Since ϕ is decreasing, then $\phi(\xi_k) \geq \phi(a_k)$, which yields that

$$a_{k+1}^{1-\tau} - a_k^{1-\tau} \geq (\tau-1)a_k^{-\tau}(a_k - a_{k+1}) \geq (\tau-1)a_k^{-\tau}c \cdot a_k^\tau = (\tau-1)c.$$

Therefore, we directly obtain $a_k^{1-\tau} \geq a_0^{1-\tau} + (\tau-1)ck$, which implies that

$$a_k \leq \left(a_0^{1-\tau} + (\tau-1)ck\right)^{-\frac{1}{\tau-1}}.$$

The proof is completed. □

Lemma A.3. *For any symmetric matrix $A \in \mathbb{R}^{n\times n}$ and $\nu \in (0,1]$, the function $p_A(x) = \|Ax\|^{1+\nu}$ is (ν, L_ν)-Hölder smooth, where $L_\nu = 2^{1-\nu}(\nu+1)\|A\|^{\nu+1}$.*

Proof. According to Theorem 6.3 in Rodomanov and Nesterov (2020), the function $p(x) = \|x\|^{1+\nu}$ is $(\nu, 2^{1-\nu}(\nu+1))$-Hölder smooth. Note that $\nabla p_A(x) = \nabla(p(Ax)) = A\nabla p(Ax)$. Hence, for any $x, y \in \mathbb{R}^n$, we have

$$\begin{aligned}\|\nabla p_A(x) - \nabla p_A(y)\| &= \|A(\nabla p(Ax) - \nabla p(Ay))\| \\ &\leq \|A\| \cdot 2^{1-\nu}(\nu+1)\|Ax - Ay\|^\nu \\ &\leq 2^{1-\nu}(\nu+1)\|A\|^{\nu+1}\|x-y\|^\nu.\end{aligned}$$

Setting $L_\nu = 2^{1-\nu}(\nu+1)\|A\|^{\nu+1}$ completes the proof. □

B Missing Proofs in Section 3

B.1 Proof of Proposition 3.1

First, observe that the Huber loss is a quadratic function with condition number 1 when $\|x\| \leq 1/(2K+1)$. By Theorem 1 in Hazan and Kakade (2019), 1-`PolyakGD` converges to $x^\star$ in one step once the iterate enters this region. Hence, it suffices to consider an initial point x^0 satisfying

$\|x^0\| > 1/(2K+1)$. We now show that after a single iteration, the iterate x^1 generated by 1-`PolyakGD` satisfies $\|x^1\| \leq 1/(2K+1)$. Specifically, the stepsize is given by

$$\alpha_0 = \frac{\mathcal{H}_K(x^0) - \mathcal{H}_K^\star}{\|\nabla \mathcal{H}_K(x^0)\|^2} = \frac{\frac{1}{2K+1}\|x^0\| - \frac{1}{2(2K+1)^2}}{\frac{1}{(2K+1)^2}} = (2K+1)\|x^0\| - \frac{1}{2}.$$

Substituting the stepsize into the update rule yields

$$x^1 = x^0 - \alpha_0 \nabla \mathcal{H}_K(x^0) = x^0 - (2K+1)\|x^0\| \cdot \frac{x^0}{(2K+1)\|x^0\|} + \frac{1}{2}\frac{x^0}{(2K+1)\|x^0\|} = \frac{x^0}{2(2K+1)\|x^0\|}.$$

Clearly, we have $\|x^1\| = 1/(2(2K+1)) < 1/(2K+1)$, which completes the proof.

B.2 Proof of Theorem 3.3

Let $\tilde{q}$ be the quadratic function defined in Equation (10) and Theorem 3.2. Define $q_\nu(x) \coloneqq \tilde{q}(x)^{\frac{\nu+1}{2}}$. Then, we have

$$\nabla q_\nu(x) = \frac{\nu+1}{2}\tilde{q}(x)^{\frac{\nu-1}{2}}\nabla \tilde{q}(x).$$

Consequently, the γ-`PolyakGD` update for q_ν is given by:

$$\begin{aligned}
x^{k+1} &= x^k - \gamma \cdot \frac{q_\nu(x^k)}{\|\nabla q_\nu(x^k)\|^2}\nabla q_\nu(x^k) \\
&= x^k - \gamma \cdot \frac{\tilde{q}(x^k)^{\frac{\nu+1}{2}}}{\left(\frac{\nu+1}{2}\right)^2 \tilde{q}(x^k)^{\nu-1}} \frac{\nu+1}{2}\tilde{q}(x^k)^{\frac{\nu-1}{2}} \frac{1}{\|\nabla \tilde{q}(x^k)\|^2}\nabla \tilde{q}(x^k) \\
&= x^k - \frac{2\gamma}{\nu+1}\frac{\tilde{q}(x^k)}{\|\nabla \tilde{q}(x^k)\|^2}\nabla \tilde{q}(x^k).
\end{aligned}$$

Observe that the trajectory coincides with that of $\frac{2\gamma}{\nu+1}$-`PolyakGD` applied to $\tilde{q}$. Given that $\frac{2\gamma}{\nu+1} \in (0,4)$ for all $\gamma \in (0,2)$ and $\nu \in (0,1]$, a direct application of Theorem 3.2 establishes shows, for a specific initial point x^0, that

$$\min_{1\leq k\leq K} \frac{\tilde{q}(x^k) - \tilde{q}^\star}{L\|x^0 - x^\star\|^2} \geq \frac{\gamma}{2\mathrm{e}^{\frac{4\gamma}{\nu+1}}((2(\nu+1)-\gamma)K + \gamma)}.$$

Using the fact that $q_\nu^\star = \tilde{q}^\star = 0$, we obtain

$$\min_{1\leq k\leq K} q_\nu(x^k) = \min_{1\leq k\leq K} \tilde{q}(x^k)^{\frac{\nu+1}{2}} \geq \left(\frac{\gamma L\|x^0 - x^\star\|^2}{2\mathrm{e}^{\frac{4\gamma}{\nu+1}}((2(\nu+1)-\gamma)K + \gamma)}\right)^{\frac{\nu+1}{2}}. \tag{19}$$

Recall that

$$\tilde{q}(x) = \frac{L}{\kappa}q(x) = \frac{1}{2}\langle x, Ax\rangle = \frac{1}{2}\|A^{1/2}x\|^2, \; A = \begin{pmatrix} L & 0 \\ 0 & L/\kappa \end{pmatrix}.$$

Hence, we have $q_\nu(x) = 2^{-\frac{\nu+1}{2}}\|A^{1/2}x\|^{\nu+1}$. By technical lemma A.3, q_ν is (ν, L_ν)-Hölder smooth with

$$L_\nu = 2^{-\frac{\nu+1}{2}}2^{1-\nu}(\nu+1)\|A^{1/2}\|^{\nu+1} = 2^{\frac{-3\nu+1}{2}}(\nu+1)L^{\frac{\nu+1}{2}}.$$

Finally, combining this with Equation (19) yields

$$\min_{1\leq k\leq K} \frac{q_\nu(x^k) - q_\nu^\star}{L_\nu\|x^0 - x^\star\|^{\nu+1}} \geq \frac{2^{\nu-1}\gamma^{\frac{\nu+1}{2}}}{\mathrm{e}^{2\gamma}(\nu+1)((2(\nu+1)-\gamma)K+\gamma)^{\frac{\nu+1}{2}}}.$$

This completes the proof.

B.3 Proof of Theorem 3.4

Let us start with the L-smooth convex case. Consider the same setting as in Theorem 3.2. By definition of $\tilde{q}$,

$$\|\nabla \tilde{q}(x^k)\|^2 = \frac{L^2}{\kappa^2}\|\nabla q(x^k)\|^2 = \frac{L^2}{\kappa^2}\left(\kappa^2 (x^k_{(1)})^2 + (x^k_{(2)})^2\right), \quad \text{for all } k \geq 0.$$

Recall that the trajectory satisfies the relationship

$$(x^k_{(2)})^2 = \frac{4\kappa^2 - \gamma\kappa(\kappa+1)}{\gamma(\kappa+1)-4}(x^k_{(1)})^2, \ \left|\frac{x^{k+1}_{(1)}}{x^k_{(1)}}\right| = \frac{\kappa-1}{\kappa+1}, \quad \text{for all } k \geq 0.$$

Consequently, for all $k \geq 1$, the trajectory generated by γ-`PolyakGD` satisfies

$$\|\nabla \tilde{q}(x^k)\|^2 = \left(\frac{\kappa-1}{\kappa+1}\right)^2 \|\nabla \tilde{q}(x^{k-1})\|^2 = \left(\frac{\kappa-1}{\kappa+1}\right)^{2k} \|\nabla \tilde{q}(x^0)\|^2 = \left(\frac{\kappa-1}{\kappa+1}\right)^{2k} \frac{\gamma L^2}{(4-\gamma)\kappa}. \tag{20}$$

According to the proof of Theorem 3.2, for all $k \geq 1$, the function value satisfies

$$\tilde{q}(x^k) = \left(\frac{\kappa-1}{\kappa+1}\right)^{2k} \tilde{q}(x^0) = \left(\frac{\kappa-1}{\kappa+1}\right)^{2k} \frac{2L}{(4-\gamma)(\kappa+1)}. \tag{21}$$

Following a similar argument to that in the proof of Theorem 3.3, let $q_\nu(x) = \tilde{q}(x)^{\frac{\nu+1}{2}}$. Recall from the proof of Theorem 3.3 that the trajectory of γ-`PolyakGD` generated on q_ν coincides with that of $\frac{2\gamma}{\nu+1}$-`PolyakGD` applied to $\tilde{q}$. Since $\nabla q_\nu(x) = \frac{\nu+1}{2}\tilde{q}(x)^{\frac{\nu-1}{2}}\nabla \tilde{q}(x)$, by substituting $\frac{2\gamma}{\nu+1}$ for γ in Equations (20) and (21), we obtain that the trajectory satisfies

$$\begin{aligned}
\|\nabla q_\nu(x^k)\| &= \frac{\sqrt{\gamma}(\nu+1)^{\frac{\nu+1}{2}} L^{\frac{\nu+1}{2}}}{2\sqrt{\kappa}\left(2(\nu+1)-\gamma\right)^{\frac{\nu}{2}}(\kappa+1)^{\frac{\nu-1}{2}}}\left(\frac{\kappa-1}{\kappa+1}\right)^{k\nu} \\
&= \frac{2^{\frac{3(\nu-1)}{2}}\sqrt{\gamma}\,(\nu+1)^{\frac{\nu-1}{2}} L_\nu \|x^0 - x^\star\|^{\frac{\nu+1}{2}}}{\sqrt{\kappa}\left(2(\nu+1)-\gamma\right)^{\frac{\nu}{2}}(\kappa+1)^{\frac{\nu-1}{2}}}\left(\frac{\kappa-1}{\kappa+1}\right)^{k\nu}, \quad \text{for all } k \geq 1,
\end{aligned}$$

where the second equality holds since $L_\nu = 2^{\frac{-3\nu+1}{2}}(\nu+1)L^{\frac{\nu+1}{2}}$, $\|x^0\| = 1$ and $\|x^\star\| = 0$. Note that the gradient norm is monotone in k. Choosing $\kappa = 4K/\gamma + \gamma/(4-\gamma)$, we conclude

$$\begin{aligned}
&\min_{1\leq k\leq K} \frac{\|\nabla q_\nu(x^k)\|}{L_\nu \|x^0 - x^\star\|^{\frac{\nu+1}{2}}} \\
&= \frac{2^{\frac{\nu-1}{2}}\gamma^{\frac{\nu+1}{2}}(4-\gamma)^{\frac{\nu}{2}}(\nu+1)^{\frac{\nu-1}{2}}}{\sqrt{4K(4-\gamma)+\gamma^2}\left((4-\gamma)K+\gamma\right)^{\frac{\nu-1}{2}}\left(2(\nu+1)-\gamma\right)^{\frac{\nu}{2}}}\left(\frac{4K - K\gamma - \gamma + \frac{\gamma^2}{2}}{(4-\gamma)K+\gamma}\right)^{K\nu} \\
&\geq \frac{2^{\frac{\nu-1}{2}}\gamma^{\frac{\nu+1}{2}}(\nu+1)^{\frac{\nu-1}{2}}\sqrt{4-\gamma}}{\mathrm{e}^{\gamma\nu}\sqrt{4(4-\gamma)+\gamma^2}(2(\nu+1)-\gamma)^{\frac{\nu}{2}}K^{\frac{\nu}{2}}}.
\end{aligned}$$

The proof is completed.

C Missing Proofs in Section 4

C.1 Proof of Lemma 4.1

First note that

$$h(g(x_g^\star)) \geq h(g(x)) + \xi(x)(g(x_g^\star) - g(x)), \quad \text{for all } \xi(x) \in \partial h(g(x)),$$

by the convexity of h. Also, since the function h is nondecreasing, we have $\xi(x) \geq 0$, which implies

$$f(x) - f(x_g^\star) \leq \xi(x)(g(x) - g(x_g^\star)) \leq \xi(x)(\langle \nabla g(x), x - x_g^\star \rangle).$$

This completes the proof.

C.2 Proof of Lemma 4.2

Define $x^+ \coloneqq \operatorname{argmin}_x \{f(y) + \langle \nabla f(y), x - y\rangle + \frac{L_\nu}{\nu+1}\|x - y\|^{\nu+1}\}$. By the optimality condition,

$$x^+ = y - \frac{1}{L_\nu^{\frac{1}{\nu}}} \|\nabla f(y)\|^{\frac{1-\nu}{\nu}} \nabla f(y).$$

Hence, it follows that

$$\begin{aligned} f(x^+) &\leq f(y) + \langle \nabla f(y), -\frac{1}{L_\nu^{\frac{1}{\nu}}} \|\nabla f(y)\|^{\frac{1-\nu}{\nu}} \nabla f(y)\rangle + \frac{L_\nu}{\nu+1} \|\frac{1}{L_\nu^{\frac{1}{\nu}}} \|\nabla f(y)\|^{\frac{1-\nu}{\nu}} \nabla f(y)\|^{\nu+1} \\ &= f(y) - \frac{\nu}{\nu+1} \frac{1}{L_\nu^{\frac{1}{\nu}}} \|\nabla f(y)\|^{\frac{\nu+1}{\nu}}. \end{aligned}$$

Next, define $\varphi(y) = f(y) - \langle \nabla f(x), y\rangle$. We have $x = \operatorname{argmin}_y \varphi(y)$ and

$$\begin{aligned} f(x) - \langle \nabla f(x), x\rangle = \varphi(x) &\leq \varphi\big(y - \frac{1}{L_\nu^{\frac{1}{\nu}}} \|\nabla f(y)\|^{\frac{1-\nu}{\nu}}\big) \\ &\leq f(y) - \langle \nabla f(x), y\rangle - \frac{\nu}{\nu+1} \frac{1}{L_\nu^{\frac{1}{\nu}}} \|\nabla f(x) - \nabla f(y)\|^{\frac{1+\nu}{\nu}}. \end{aligned}$$

Rearranging the terms gives

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\nu}{\nu+1} \frac{1}{L_\nu^{\frac{1}{\nu}}} \|\nabla f(x) - \nabla f(y)\|^{\frac{1+\nu}{\nu}}$$

and this completes the proof.

C.3 Proof of Theorem 4.2

Using Lemma 4.2, we deduce that

$$\begin{aligned} \|x^{k+1} - x^\star\|^2 &= \|x^k - 2\alpha_k \nabla f(x^k) - x^\star\|^2 \\ &= \|x^k - x^\star\|^2 - 4\alpha_k \langle \nabla f(x^k), x^k - x^\star\rangle + 4\alpha_k^2 \|\nabla f(x^k)\|^2 \\ &\leq \|x^k - x^\star\|^2 - 4\alpha_k \Big(f(x^k) - f^\star + \frac{\nu}{\nu+1} \frac{1}{L_\nu^{\frac{1}{\nu}}} \|\nabla f(x^k)\|^{\frac{\nu+1}{\nu}}\Big) + 4\alpha_k^2 \|\nabla f(x^k)\|^2 \\ &= \|x^k - x^\star\|^2 - \frac{4\nu}{\nu+1} \frac{1}{L_\nu^{\frac{1}{\nu}}} (f(x^k) - f^\star) \|\nabla f(x^k)\|^{\frac{1-\nu}{\nu}}. \end{aligned}$$

By a similar argument in the proof of Theorem 4.1, we have

$$\mathrm{D}_{k+1}^2 \leq \mathrm{D}_k^2 - \frac{4\nu}{\nu+1}\frac{\Delta_k}{L_\nu^{\frac{1}{\nu}}}\|\nabla f(x^k)\|^{\frac{1-\nu}{\nu}},$$

where $\mathrm{D}_k = \mathrm{dist}(x^k, \mathcal{X}^\star)$ and $\Delta_k = f(x^k) - f^\star$. Note that due to convexity and Fejér monotonicity, we have $\Delta_k \leq \|\nabla f(x^k)\|\|x^k - x^\star\| \leq \|\nabla f(x^k)\|\|x^0 - x^\star\|$ for any $x^\star \in \mathcal{X}^\star$. Since $\nu \in (0, 1]$, substituting this inequality back yields

$$\mathrm{D}_{k+1}^2 \leq \mathrm{D}_k^2 - \frac{4\nu}{\nu+1}\frac{\mathrm{D}_0^{\frac{\nu-1}{\nu}}}{L_\nu^{\frac{1}{\nu}}}\Delta_k^{\frac{1}{\nu}}.$$

By summining over k from $k = 0$ to $k = K - 1$, we obtain

$$\sum_{k=0}^{K-1}\Delta_k^{\frac{1}{\nu}} \leq \frac{(\nu+1)L_\nu^{\frac{1}{\nu}}\,\mathrm{D}_0^{\frac{1-\nu}{\nu}}}{4\nu}D_0^2,$$

which immediately leads to

$$\min_{1\leq k\leq K} f(x^k) - f^\star \leq \left(\frac{\nu+1}{4\nu}\right)^\nu \frac{L_\nu\,\mathrm{D}_0^{\nu+1}}{K^\nu}.$$

This completes the proof.

C.4 Proof of Theorem 4.3

Using Lemma 4.2, we deduce

$$\begin{aligned}
\|x^{k+1} - x^\star\|^2 &= \|x^k - \gamma\cdot\alpha_k\nabla f(x^k) - x^\star\|^2 \\
&= \|x^k - x^\star\|^2 - 2\gamma\cdot\alpha_k\langle\nabla f(x^k), x^k - x^\star\rangle + \gamma^2\cdot\alpha_k^2\|\nabla f(x^k)\|^2 \\
&\leq \|x^k - x^\star\|^2 - 2\gamma\cdot\alpha_k\left(f(x^k) - f^\star + \frac{\nu}{\nu+1}\frac{1}{L_\nu^{\frac{1}{\nu}}}\|\nabla f(x^k)\|^{\frac{\nu+1}{\nu}}\right) + \gamma^2\cdot\alpha_k^2\|\nabla f(x^k)\|^2 \\
&= \|x^k - x^\star\|^2 - \gamma(2-\gamma)\frac{(f(x^k) - f^\star)^2}{\|\nabla f(x^k)\|^2} - 2\gamma\frac{f(x^k) - f^\star}{\|\nabla f(x^k)\|^2}\frac{\nu}{\nu+1}\frac{1}{L_\nu^{\frac{1}{\nu}}}\|\nabla f(x^k)\|^{\frac{\nu+1}{\nu}} \\
&\leq \|x^k - x^\star\|^2 - \frac{2\gamma\nu}{\nu+1}\frac{1}{L_\nu^{\frac{1}{\nu}}}(f(x^k) - f^\star)\|\nabla f(x^k)\|^{\frac{1-\nu}{\nu}}
\end{aligned}$$

Then by substituting the inequality in Lemma 4.2 with $x = x^\star$ and $y = x^k$, it holds that

$$\Delta_k = f(x^k) - f^\star \geq \frac{\nu}{\nu+1}\frac{1}{L_\nu^{\frac{1}{\nu}}}\|\nabla f(x^k)\|^{\frac{1+\nu}{\nu}}.$$

Consequently, we have

$$\frac{2\gamma\nu^2}{(\nu+1)^2}\frac{1}{L_\nu^{\frac{2}{\nu}}}\|\nabla f(x^k)\|^{\frac{2}{\nu}} \leq \|x^k - x^\star\|^2 - \|x^{k+1} - x^\star\|^2 \leq \|x^k - x^\star\|^2 - \mathrm{D}_{k+1}^2.$$

Since $x^\star$ is taken arbitrarily, it implies

$$\frac{2\gamma\nu^2}{(\nu+1)^2}\frac{1}{L_\nu^{\frac{2}{\nu}}}\|\nabla f(x^k)\|^{\frac{2}{\nu}} \le \mathrm{D}_k^2 - \mathrm{D}_{k+1}^2 \,.$$

Summing over k from $k=0$ to $k=K-1$ and dividing K on both sides, we conclude

$$\min_{1\le k\le K} \|\nabla f(x^k)\| \le \frac{(\nu+1)^\nu L_\nu \,\mathrm{D}_0^\nu}{2^{\frac{\nu}{2}}\gamma^{\frac{\nu}{2}}\nu^\nu K^{\frac{\nu}{2}}}.$$

This completes the proof.

C.5 Proof of Theorem 4.4

Let $\hat{\sigma}_f(t) \coloneqq \int_0^t \frac{\hat{\mu}(\tau)}{\tau}\,\mathrm{d}\,\tau$. By Theorem 1 in Nesterov (2025), it holds that

$$\left|f(y) - f(x) - \langle \nabla f(x), y-x\rangle\right| \le 2\hat{\sigma}_f(t) - \hat{\mu}_f(t) + \frac{\hat{\mu}_f(t)}{t^2}\|y-x\|^2 \tag{22}$$

for any $x, y \in \mathbb{R}^n$ and $t \in [0,\infty)$. We first establish the following upper bound for the gradient norm:

$$\|\nabla f(x)\| \le \inf_{t\ge 0} \frac{f(x) - f^\star + 2\hat{\sigma}_f(t)}{t}, \tag{23}$$

which can be viewed as a counterpart to the second inequality in Equation (4). Fix $\zeta \in [0,\infty)$ and $x \in \mathbb{R}^n$. By substituting $y = x - \zeta\cdot\frac{\nabla f(x)}{\|\nabla f(x)\|}$ and $t = \|x-y\| = \zeta$ into inequality (22), it yields

$$f(y) - f(x) + \zeta\|\nabla f(x)\| \le 2\hat{\sigma}_f(\zeta) - \hat{\mu}_f(\zeta) + \hat{\mu}_f(\zeta) = 2\hat{\sigma}_f(\zeta),$$

which further implies $f(x) - f^\star \ge \zeta\|\nabla f(x)\| - 2\hat{\sigma}_f(\zeta)$ due to the fact that $f(y) \ge f^\star$. Since ζ is chosen arbitrarily, we have

$$f(x) - f^\star \ge \sup_{\zeta\ge 0} \zeta\|\nabla f(x)\| - 2\hat{\sigma}_f(\zeta) = (2\hat{\sigma}_f)^*(\|\nabla f(x)\|),$$

where $(2\hat{\sigma}_f)^*$ is the Fenchel conjugate function of $2\hat{\sigma}_f$ (see Definition 3.22 in Drusvyatskiy (2020)). Consequently, by the Fenchel-Young inequality, it follows $\zeta\|\nabla f(x)\| \le 2\hat{\sigma}_f(\zeta) + (2\hat{\sigma}_f)^*(\|\nabla f(x)\|)$. Note that $f(x) - f^\star \ge (2\hat{\sigma}_f)^*(\|\nabla f(x)\|)$. We derive

$$\zeta\|\nabla f(x)\| \le 2\hat{\sigma}_f(\zeta) + f(x) - f^\star.$$

Again, recall that ζ is arbitrary. This proves Equation (23). Since the function f is convex, by Lemma 4 in Nesterov (2025), it holds that $\hat{\sigma}_f(t) \le \hat{\mu}_f(t)$ for any $t \in [0,\infty)$. From Equation (23), we obtain

$$\|\nabla f(x)\| \le \inf_{t\ge 0} \frac{f(x) - f^\star + 2\hat{\mu}_f(t)}{t} \le \frac{f(x) - f^\star + 2\hat{\mu}_f(s_f(f(x)-f^\star))}{s_f(f(x)-f^\star)}.$$

Since the complexity gauge is the inverse of the global curvature bound, it follows $\hat{\mu}_f(s_f(f(x) - f^\star)) = f(x) - f^\star$, which leads to

$$\|\nabla f(x)\| \le \frac{3(f(x)-f^\star)}{s_f(f(x)-f^\star)}.$$

Combining the above inequality with Equation (16), we have

$$\mathrm{D}_{k+1}^2 \leq \mathrm{D}_k^2 - \gamma(2-\gamma)\frac{\Delta_k^2}{\|\nabla f(x^k)\|^2} \leq \mathrm{D}_k^2 - \frac{s_f^2(\Delta_k)}{9},$$

where $\mathrm{D}_k = \mathrm{dist}(x^k, \mathcal{X}^\star)$ and $\Delta_k = f(x^k) - f^\star$. Summing the relation over k from $k=0$ to $k=K-1$ gives

$$\frac{1}{K}\sum_{k=0}^{K-1} s_f^2(\Delta_k) \leq \frac{9\,\mathrm{D}_0^2}{K}.$$

It follows that

$$\min_{1\leq k\leq K} \; s_f(\Delta_k) \leq \frac{3\,\mathrm{D}_0}{\sqrt{K}},$$

which, by taking inverse on both sides, implies

$$\min_{1\leq k\leq K} \; f(x^k) - f^\star \leq \hat{\mu}_f\Big(\frac{3\,\mathrm{D}_0}{\sqrt{K}}\Big).$$

Setting $K \geq (3\,\mathrm{D}_0\,/s_f(\varepsilon))^2$ ensures $\hat{\mu}_f(3\,\mathrm{D}_0\,/\sqrt{K}) \leq \varepsilon$ and completes the proof.

C.6 Proof of Theorem 4.5

Define $f_\xi^\star := \min_x f(x,\xi)$. The interpolation condition implies $\mathbb{E}[f_\xi^\star] = f^\star$. Using the update formula of the Polyak stepsize, we deduce that

$$\begin{aligned}
\|x^{k+1} - x^\star\|^2 &= \|x^k - x^\star\|^2 - 2\gamma\cdot\alpha_k\langle\nabla f(x^k,\xi^k), x^k - x^\star\rangle + \gamma^2\cdot\alpha_k^2\|\nabla f(x^k,\xi^k)\|^2 \\
&\leq \|x^k - x^\star\|^2 - 2\gamma\cdot\alpha_k(f(x^k,\xi^k) - f_{\xi^k}^\star) + \gamma^2\cdot\alpha_k^2\|\nabla f(x^k,\xi^k)\|^2 \\
&= \|x^k - x^\star\|^2 - \gamma(2-\gamma)\frac{(f(x^k,\xi^k) - f_{\xi^k}^\star)^2}{\|\nabla f(x^k,\xi^k)\|^2}.
\end{aligned}$$

With the stochastic version of Equation (4):

$$\|\nabla f(x^k,\xi^k)\|^{\frac{\nu+1}{\nu}} \leq \frac{\nu+1}{\nu}L_\nu^{\frac{1}{\nu}}(f(x^k,\xi^k) - f_{\xi^k}^\star)$$

we deduce that

$$\begin{aligned}
\|x^{k+1} - x^\star\|^2 &\leq \|x^k - x^\star\|^2 - \gamma(2-\gamma)\frac{(f(x^k,\xi^k) - f_{\xi^k}^\star)^2}{\|\nabla f(x^k,\xi^k)\|^2} \\
&\leq \|x^k - x^\star\|^2 - \gamma(2-\gamma)\frac{(f(x^k,\xi^k) - f_{\xi^k}^\star)^2}{\big(\frac{\nu+1}{\nu}\big)^{\frac{2\nu}{\nu+1}}L_\nu^{\frac{2}{\nu+1}}(f(x^k,\xi^k) - f_{\xi^k}^\star)^{\frac{2\nu}{\nu+1}}} \\
&= \|x^k - x^\star\|^2 - \frac{\gamma(2-\gamma)\nu^{\frac{2\nu}{\nu+1}}}{(\nu+1)^{\frac{2\nu}{\nu+1}}L_\nu^{\frac{2}{\nu+1}}}(f(x^k,\xi^k) - f_{\xi^k}^\star)^{\frac{2}{\nu+1}}
\end{aligned}$$

Taking the conditional expectation on both sides, we have

$$\begin{aligned}
\mathbb{E}[\|x^{k+1}-x^\star\|^2|x^k] &\le \|x^k-x^\star\|^2 - \frac{\gamma(2-\gamma)\nu^{\frac{2\nu}{\nu+1}}}{(\nu+1)^{\frac{2\nu}{\nu+1}}L_\nu^{\frac{2}{\nu+1}}}\mathbb{E}\left[(f(x^k,\xi^k)-f_{\xi^k}^\star)^{\frac{2}{\nu+1}}|x^k\right] \\
&\le \|x^k-x^\star\|^2 - \frac{\gamma(2-\gamma)\nu^{\frac{2\nu}{\nu+1}}}{(\nu+1)^{\frac{2\nu}{\nu+1}}L_\nu^{\frac{2}{\nu+1}}}\mathbb{E}[f(x^k,\xi^k)-f_{\xi^k}^\star|x^k]^{\frac{2}{\nu+1}} \qquad (24)\\
&= \|x^k-x^\star\|^2 - \frac{\gamma(2-\gamma)\nu^{\frac{2\nu}{\nu+1}}}{(\nu+1)^{\frac{2\nu}{\nu+1}}L_\nu^{\frac{2}{\nu+1}}}[f(x^k)-f^\star]^{\frac{2}{\nu+1}} \qquad (25)\\
&= \|x^k-x^\star\|^2 - \frac{\gamma(2-\gamma)\nu^{\frac{2\nu}{\nu+1}}}{(\nu+1)^{\frac{2\nu}{\nu+1}}L_\nu^{\frac{2}{\nu+1}}}\Delta_k^{\frac{2}{\nu+1}},
\end{aligned}$$

where (24) uses Jensen's inequality $\mathbb{E}[X^{\frac{2}{\nu+1}}] \ge \mathbb{E}[X]^{\frac{2}{\nu+1}}$ for nonnegative random variable X and $\nu \in (0,1]$; (25) uses the interpolation condition $\mathbb{E}[f(x,\xi)-f_\xi^\star|x] = f(x)-f^\star$.With the Hölder growth condition, we have

$$\mathbb{E}[\|x^{k+1}-x^\star\|^2|x^k] \le \|x^k-x^\star\|^2 - \frac{\gamma(2-\gamma)\nu^{\frac{2\nu}{\nu+1}}\rho_r^{\frac{2}{\nu+1}}}{(\nu+1)^{\frac{2\nu}{\nu+1}}L_\nu^{\frac{2}{\nu+1}}}\|x^k-x^\star\|^{\frac{2r}{\nu+1}}.$$

Taking total expectation and using $r \ge \nu+1$, we have

$$\mathbb{E}\left[\|x^k-x^\star\|^{\frac{2r}{\nu+1}}\right] = \mathbb{E}\left[(\|x^k-x^\star\|^2)^{\frac{r}{\nu+1}}\right] \ge \mathbb{E}[\|x^k-x^\star\|^2]^{\frac{r}{\nu+1}},$$

which implies

$$\begin{aligned}
\mathbb{E}[\|x^{k+1}-x^\star\|^2] &\le \mathbb{E}[\|x^k-x^\star\|^2] - \frac{\gamma(2-\gamma)\nu^{\frac{2\nu}{\nu+1}}\rho_r^{\frac{2}{\nu+1}}}{(\nu+1)^{\frac{2\nu}{\nu+1}}L_\nu^{\frac{2}{\nu+1}}}\mathbb{E}\left[\|x^k-x^\star\|^{\frac{2r}{\nu+1}}\right] \\
&\le \mathbb{E}[\|x^k-x^\star\|^2] - \frac{\gamma(2-\gamma)\nu^{\frac{2\nu}{\nu+1}}\rho_r^{\frac{2}{\nu+1}}}{(\nu+1)^{\frac{2\nu}{\nu+1}}L_\nu^{\frac{2}{\nu+1}}}\mathbb{E}[\|x^k-x^\star\|^2]^{\frac{r}{\nu+1}}.
\end{aligned}$$

Define $\mathbb{D}_k \coloneqq \mathbb{E}[\|x^k-x^\star\|^2]$. We have exactly the same recursive relation as in (17):

$$\mathbb{D}_{k+1} \le \mathbb{D}_k - \frac{\gamma(2-\gamma)\nu^{\frac{2\nu}{\nu+1}}\rho_r^{\frac{2}{\nu+1}}}{(\nu+1)^{\frac{2\nu}{\nu+1}}L_\nu^{\frac{2}{\nu+1}}}\mathbb{D}_k^{\frac{r}{\nu+1}},$$

and the same results from the analysis of the deterministic case. Using

$$\frac{\gamma(2-\gamma)\nu^{\frac{2\nu}{\nu+1}}}{(\nu+1)^{\frac{2\nu}{\nu+1}}L_\nu^{\frac{2}{\nu+1}}}\mathbb{E}[\Delta_k]^{\frac{2}{\nu+1}} \le \frac{\gamma(2-\gamma)\nu^{\frac{2\nu}{\nu+1}}}{(\nu+1)^{\frac{2\nu}{\nu+1}}L_\nu^{\frac{2}{\nu+1}}}\mathbb{E}\left[\Delta_k^{\frac{2}{\nu+1}}\right] \le \mathbb{D}_k - \mathbb{D}_{k+1},$$

following Theorem 4.1 completes the proof.