New Perspectives on the Polyak Stepsize: Surrogate Functions and Negative Results

Francesco Orabona
King Abdullah University of Science and Technology (KAUST)
Thuwal, 23955-6900, Kingdom of Saudi Arabia
francesco@orabona.com

Ryan D'Orazio
Mila Québec AI Institute, Université de Montréal
Montréal, QC, Canada
ryan.dorazio@mila.quebec

Abstract

The Polyak stepsize has been proven to be a fundamental stepsize in convex optimization, giving near optimal gradient descent rates across a wide range of assumptions. The universality of the Polyak stepsize has also inspired many stochastic variants, with theoretical guarantees and strong empirical performance. Despite the many theoretical results, our understanding of the convergence properties and shortcomings of the Polyak stepsize or its variants is both incomplete and fractured across different analyses. We propose a new, unified, and simple perspective for the Polyak stepsize and its variants as gradient descent on a surrogate loss. We show that *each variant is equivalent to minimize a surrogate function with stepsizes that adapt to a guaranteed local curvature*. Our general surrogate loss perspective is then used to provide a unified analysis of existing variants across different assumptions. Moreover, we show a number of negative results proving that the non-convergence results in some of the upper bounds is indeed real.

1 Introduction

The iterative optimization of complex functions forms a cornerstone of modern machine learning, scientific computing, and engineering. Among the most foundational first-order methods is gradient descent, which iteratively refines a solution by moving in the direction opposite to the function's gradient. A critical aspect of gradient descent is the selection of an appropriate stepsize (or learning rate), as it dictates both the speed of convergence and the stability of the algorithm. The wrong choice of the stepsizes can lead to slow convergence or, conversely, to divergence, making the tuning process a significant practical hurdle.

In this landscape of stepsize selection strategies, the Polyak stepsize, proposed by Polyak [37] stands out for its theoretical elegance and convergence properties. Starting from an arbitrary $\boldsymbol{x}_1 \in \mathbb{R}^d$, the Polyak stepsize is defined as

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \frac{f(\boldsymbol{x}_t) - f^\star}{\|\boldsymbol{g}_t\|^2} \boldsymbol{g}_t, \tag{1}$$

where $\mathbf{0} \neq \boldsymbol{g}_t \in \partial f(\boldsymbol{x}_t)$ and $f^\star = \min_{\boldsymbol{x}} f(\boldsymbol{x})$. If $\boldsymbol{g}_t = \mathbf{0}$, then $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t$. This update rule can achieve linear convergence for strongly convex and smooth functions, $\mathcal{O}(1/T)$ rate for convex smooth functions, and $\mathcal{O}(1/\sqrt{T})$ rate for non-smooth convex ones. This is particularly interesting because all of these rates are achieved with a unique stepsize and without knowledge of smoothness

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

or curvature constants. In other words, this update rule is *adaptive* to the geometry of the functions to optimize.

Recently, the Polyak stepsize has seen a resurgence in the machine learning literature, with a plethora of variants. However, despite the big number of papers on this topic, one essential research question seems still to be missing: *What makes the Polyak stepsize adaptive and when can it fail?*

Contributions. This paper aims to provide a novel framework to understand the Polyak stepsize, providing a clear geometric explanation of its adaptivity. In particular, we show that the adaptivity is due to a simple but powerful observation: *The Polyak stepsize minimizes a surrogate objective function that is always locally smooth.* As for standard smooth functions, we will show that the knowledge of the local smoothness constant is enough to obtain the correct rates. In addition, the local smoothness will depend only on the gradient itself, removing the need to estimate it. Furthermore, we show that minimal curvature of the surrogate is inherited from the original function as well. We also use this framework to extend its core idea to a family of algorithms. Then, we will show a number of negative results when $f(\boldsymbol{x}^\star)$ is not known and for its use in the stochastic case. These negative results complete our understanding by showing that some non-vanishing terms in existing upper bounds are necessary.

2 Related Work

In the pionnering work of Ermol'ev [14] stepsizes of the form $\eta_t \propto 1/\|\boldsymbol{g}_t\|^2$ were proposed for non-smooth optimization. Despite the many convergence guarantees enabled by Ermol'ev [14]'s framework, in Polyak [37] it is noted that linear convergence is not possible with such stepsizes. As an alternative, Polyak suggests the stepsize (1), which is shown to converge at favourable rates with non-smooth convex functions, and strongly convex and smooth functions. In fact, contrary to common belief, Polyak [37] was the first to show linear convergence with a rate comparable to gradient descent in the smooth and strongly convex case. Furthermore, the case where $f^\star$ is estimated was also studied, showing convergence to a level set if $f^\star$ is overestimated, and best-iterate convergence to a neighboorhood if it is underestimated. The Polyak stepsize has since been extended and studied across several applications and domains.

Non-smooth convex. In non-smooth convex optimization, several schemes have been developed to estimate $f^\star$ on the fly [7, 6, 16, 41, 31, 27]. In the finite-sum case with interpolation, (1) and variants have been studied as an incremental subgradient method [31, 27].

Non-expansive operators. In the context of non-expansive operators, the update (1) has also been studied as a special case of the subgradient projector [2, 8, 10]; where it can be shown that subgradient descent with $\eta_t = (f(x)-c)_+/\|\boldsymbol{g}_t\|^2$ is a quasi-firmly non-expansive operator[1] if f is convex and $c \geq f^\star$, and $\boldsymbol{x} \to \boldsymbol{x}^\star$ if f is continuous [2]. Moreover, in the finite-sum setting, interpolation can also be viewed as iterating different quasi-firmly non-expansive operators with a common fixed point. For example, applying a subgradient projector sequentially (i.e., cycling through the different component functions) in a way such that each function eventually gets visited guarantees that $\boldsymbol{x}_t \to \{f(\boldsymbol{y}) \leq c\}$ [8, Example 5.9.7]. For a survey on this topic see Censor [9].

Deterministic. In modern optimization, (1) has been shown to achieve similar rates to gradient descent in various common assumptions (e.g., Lipschitz, smoothness and strong convexity) [22], and more recently with other assumptions such as weakly convex functions [12], (L_0, L_1)-smooth functions [42, 17], and directional smoothness [30].

Stochastic. The Polyak stepsize has also been extended to the stochastic case with emphasis on applications to machine learning [39, 5, 29, 38]. The ALI-G method [5] and $\texttt{SPS}_{\max}$ [29] use stochastic estimates via the sampled function $f(\boldsymbol{x}, \boldsymbol{\xi})$ and its gradient $\nabla f(\boldsymbol{x}, \boldsymbol{\xi})$ to perform a Polyak-like stepsize. $\texttt{SPS}_{\max}$ in addition uses $\inf_x f(x, \boldsymbol{\xi})$ as an estimate to $f^\star$, and is shown to converge at fast rates without a neighbourhood under interpolation. Folowing $\texttt{SPS}_{\max}$ many variants have been proposed for SGD: StoPS [24], DECSPS and $\texttt{SPS}^{\ell}_{\max}$ [35], $\texttt{SPS}_+$ [15]. Beyond SGD, other extensions include: mirror descent [13], with preconditioning [1], with line-search [25], and with momentum [43, 40, 33].

[1] An operator T is quasi-firmly non-expansive if for all fixed points $\boldsymbol{x}^\star$, $\|T(\boldsymbol{x}) - \boldsymbol{x}^\star\|^2 \leq \|\boldsymbol{x} - \boldsymbol{x}^\star\|^2 - \|T(\boldsymbol{x}) - \boldsymbol{x}\|^2$. Note that a quasi-firmly non-expansive operators are also referred to as *cutters*.

Neighbourhood of Convergence. For $\texttt{SPS}_{\max}$, Loizou et al. [29] proved that the suboptimality gap is only guaranteed to shrink up to a factor that depends on the loss themselves. As we will explain in Section 5.1, this is equivalent to the guarantees in online learning where the regret is proportional to the cumulative loss of the competitor, typically denoted by $L^\star$. Hence, these kind of guarantees are usually called in online learning $L^\star$ bounds [see, e.g., 34, Section 4.2].

Surrogates. Gower et al. [20] propose a fixed stepsize online sgd surrogate perspective to the Polyak stepsize and the TAPS variant. The methods are shown to be equivalent to sgd on a self-bounded surrogate. In comparison, our surrogate approach considers a fixed surrogate loss with local smoothness where the Polyak stepsize is chosen to be the inverse of the local smoothness.

Surprisingly enough, despite the adaptivity of the Polyak stepsize across various assumptions without modification, in previous literature there is no clear explanation why this is the case.

3 Definitions and Notation

We will use the following notation and definitions. All the norms in this paper are L2 norms and will be denoted by $\|\cdot\|$. For a function $f : \mathbb{R}^d \to \mathbb{R}$, we define a *subgradient* of f in $\boldsymbol{x} \in \mathbb{R}^d$ as a vector $\boldsymbol{g} \in \mathbb{R}^d$ that satisfies $f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \boldsymbol{g}, \boldsymbol{y} - \boldsymbol{x} \rangle,\ \forall \boldsymbol{y} \in \mathbb{R}^d$. We denote the set of subgradients of f in $\boldsymbol{x}$ by $\partial f(\boldsymbol{x})$. For a differentiable function we have that $\partial f(\boldsymbol{x}) = \{\nabla f(\boldsymbol{x})\}$. A function $f : V \to \mathbb{R}$, differentiable in an open set containing V, is L-*smooth* w.r.t. $\|\cdot\|$ if $f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2$ for all $\boldsymbol{x}, \boldsymbol{y} \in V$.

Definition 1. *We say that a function f has a s-sharp minimum in $\boldsymbol{x}^\star$ if*

$$f(\boldsymbol{x}) - f(\boldsymbol{x}^\star) \geq s\|\boldsymbol{x} - \boldsymbol{x}^\star\| \ .$$

Note that if f has a sharp minimum then it is not differentiable at $\boldsymbol{x}^*$ [36] and if the function is also convex and G-Lipschitz we immediately have $G \geq s$.

Definition 2. *We say that a function $f : \mathbb{R}^d \to \mathbb{R}$ is L-self-bounded if*

$$\|\boldsymbol{g}\|^2 \leq 2L(f(\boldsymbol{x}) - \inf_{\boldsymbol{x}}\ f(\boldsymbol{x})),\ \forall \boldsymbol{g} \in \partial f(\boldsymbol{x}) \ .$$

It is known that L-smooth functions are L-self-bounded [see, e.g., Lemma 4 in 28], but this definition is strictly weaker because it does not assume differentiability.

4 Polyak Stepsize is Gradient Descent on a Surrogate Function

Let f be convex and $\boldsymbol{x}^\star \in \operatorname{argmin}_{\boldsymbol{x}}\ f(\boldsymbol{x})$. Consider the following function:

$$\phi(\boldsymbol{x}) = \frac{1}{2}\left(f(\boldsymbol{x}) - f(\boldsymbol{x}^\star)\right)^2 \ . \tag{2}$$

Instead of viewing the Polyak stepsize (1) with respect to f we propose to view it equivalently as a subgradient method with respect to ϕ. By the chain rule of subgradients [2][Corollary 16.72], subgradient descent with the Polyak stepsize (1) is equivalent to subgradient descent on ϕ with stepsize $\eta_t = \frac{1}{\|\boldsymbol{g}_t\|^2}$. This perspective may seem superfluous, howerever, we will show that $\frac{1}{\|\boldsymbol{g}_t\|^2}$ is strongly related to a certain notion of local curvature of ϕ, local star upper curvature.

Definition 3 (Local star upper curvature (LSUC))**.** *We say that a function f with minimizer $\boldsymbol{x}^\star$ has $\lambda_{\boldsymbol{y}}$-local star upper curvature (LSUC) around $\boldsymbol{y}$ if there exists $\lambda_{\boldsymbol{y}} > 0$ such that*

$$f(\boldsymbol{x}^\star) - f(\boldsymbol{y}) - \langle \boldsymbol{g}_{\boldsymbol{y}}, \boldsymbol{x}^\star - \boldsymbol{y} \rangle \geq \frac{1}{2\lambda_{\boldsymbol{y}}}\|\boldsymbol{g}_{\boldsymbol{y}}\|_2^2,\ \forall \boldsymbol{g}_{\boldsymbol{y}} \in \partial f(\boldsymbol{y}) \ .$$

Note that if the function is LSUC everywhere, then it must be convex since we assume the existence of a subgradient.[2] It is also immediate to show that convex L-smooth functions are also L-LSUC. Indeed, for convex L-smooth functions we have that [32, Theorem 2.1.5]

$$f(\boldsymbol{x}) - f(\boldsymbol{y}) - \langle \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \geq \frac{1}{2L}\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|^2,\ \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d \ .$$

[2] If $\boldsymbol{g}_t$ is not a subgradient but a directional derivative then f would be guaranteed to be star-convex [26].

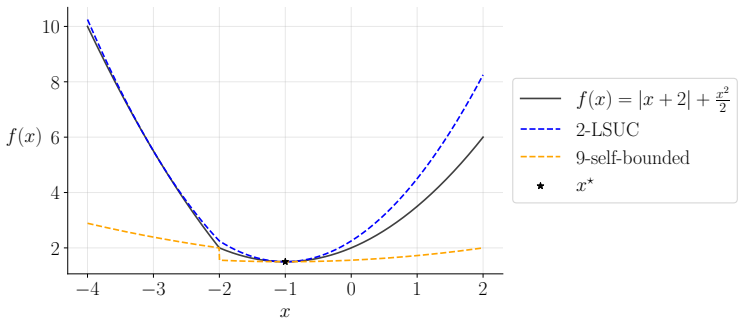

Figure 1: The function $f(x) = |x+2| + \frac{x^2}{2}$ is non-smooth but is 2-LSUC as demonstrated by the blue curve, $f(x^\star) - \langle \boldsymbol{g}, x^\star - x\rangle - 1/4\|\boldsymbol{g}\|^2$, being larger than $f(x)$ for all x and $\boldsymbol{g} \in \partial f(x)$. Similarly, f is self-bounded but with the larger constant $L = 9$.

So, it is enough to set $\boldsymbol{x} = \boldsymbol{x}^\star$ to obtain the above definition. However, the inclusion is strict, because there exist functions that are not smooth and still satisfy the above definition. For example, as shown in Figure 1, one can easily verify that $f(x) = |x+2| + \frac{x^2}{2}$ is 2-LSUC and 9-self-bounded but not differentiable $x = -2$, hence it is not smooth.

Finally, if the star-upper-curvature holds globally, i.e., there exists $0 < \lambda < \lambda_{\boldsymbol{y}}$ for all $\boldsymbol{y}$, then we can show that this condition is equivalent to the upper quadratic growth condition in Guille-Escuret et al. [21].[3] This observation was first made by Goujaud et al. [18, Theorem 2.6], we include the precise statement and proof in the Appendix B.

The key observation in the next Theorem is that ϕ is *always* locally star upper curved, regardless of the curvature (or lack of it) of the function f. Moreover, it will inherit additional curvature from f. The proof can be found in Appendix A.

Theorem 1 (Curvature of the Polyak surrogate)**.** *Let $f(x)$ be convex and define $\boldsymbol{x}^\star \in \operatorname{argmin}_{\boldsymbol{x}} f(\boldsymbol{x})$. Define $\phi(\boldsymbol{x}) = \frac{1}{2}(f(\boldsymbol{x}) - f(\boldsymbol{x}^\star))^2$. Then, we have*

- *ϕ is $\|\boldsymbol{g}_{\boldsymbol{y}}\|^2$-LSUC around any $\boldsymbol{y}$ for any $\boldsymbol{g}_{\boldsymbol{y}} \in \partial f(\boldsymbol{y})$.*
- *If f is s-sharp, then ϕ has s^2-quadratic growth.*
- *If f has μ-quadratic growth and L-self bounded, then ϕ satisfies a local quadratic growth:*

$$\phi(\boldsymbol{x}) \geq \frac{1}{2}\frac{\mu\|\boldsymbol{g}\|^2}{2L}\|\boldsymbol{x} - \boldsymbol{x}^\star\|^2, \ \forall \boldsymbol{g} \in \partial f(\boldsymbol{x}) \ .$$

This theorem tells us that, regardless of the curvature of f, we can always construct the function ϕ that is locally curved. It is well-known that for L-smooth functions one can use the stepsize $\eta = \frac{1}{L}$ and achieve a rate between $\mathcal{O}(1/T)$ and a linear one, depending on the presence of strong convexity. Here, we show a similar result: GD can use stepsizes that depend on the local star upper curvature in *all* cases. Note however, unlike GD with a constant stepsize and smoothness, we do not have a descent lemma with ϕ. Indeed this is not possible as it would guarantee a $\mathcal{O}(1/T)$ rate for the last iterate which was shown to be impossible by Goujaud et al. [18] for $QG + (L)$ functions (i.e., L-LUSC functions).

Lemma 1. *Let ϕ convex and define $\boldsymbol{x}^\star \in \operatorname{argmin}_{\boldsymbol{x}} \phi(\boldsymbol{x})$. Assume ϕ to be $\lambda_{\boldsymbol{x}}$-LSUC around any point $\boldsymbol{x}$. Then, using subgradient descent with stepsizes $\eta_t = \frac{1}{\lambda_{\boldsymbol{x}_t}}$ guarantees*

$$\eta_t \left(\phi(\boldsymbol{x}_t) - \phi(\boldsymbol{x}^\star)\right) \leq \frac{1}{2}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 - \frac{1}{2}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star\|^2 \ . \tag{3}$$

Summing this inequality over time, we also have

$$\phi(\bar{\boldsymbol{x}}_T)\sum_{t=1}^{T}\eta_t \leq \sum_{t=1}^{T}\eta_t\phi(\boldsymbol{x}_t) \leq \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{2}, \tag{4}$$

[3] A function f satisfies the μ-quadratic growth condition if $f(\boldsymbol{x}) - f(\boldsymbol{x}^\star) \geq \mu/2\|\boldsymbol{x} - \boldsymbol{x}^\star\|^2$.

where $\bar{\boldsymbol{x}}_T = \frac{1}{\sum_{t=1}^T \eta_t} \boldsymbol{x}_t$ *or* $\bar{\boldsymbol{x}}_T = \operatorname{argmin}_{\boldsymbol{x} \in \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_T\}} \phi(\boldsymbol{x})$.

Proof. From the classic one-step analysis of GD [see, e.g., 34], we have

$$\eta_t \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}^\star \rangle = \frac{1}{2}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 - \frac{1}{2}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star\|^2 + \frac{\eta_t^2}{2}\|\boldsymbol{g}_t\|^2,$$

for any $\boldsymbol{g}_{\boldsymbol{x}_t} \in \partial\phi(\boldsymbol{x}_t)$. Now, we use the fact that ϕ is LSUC and the definition of η_t we obtain the stated bound. Summing from $t = 1$ to T and discarding the negative term on the right hand side concludes the proof. □

The above discussion can be summarized in the following theorem.

Theorem 2. *The Polyak stepsize in* (1) *is equivalent to subgradient descent on the function* ϕ *in* (2)*, when using stepsizes* η_t *equal to the inverse of the local-star-upper curvature of* ϕ *in* $\boldsymbol{x}_t$*.*

This theorem and Lemma 1 do not give us a rate, however we can immediately observe that if $\sum_t \eta_t = +\infty$ we also have $\phi(\boldsymbol{x}_t) \to 0$, implying convergence of the last iterate (see Remark 6 for more details).

To obtain known rates for the Polyak stepsize, we can use additional assumptions on f. However, we want to stress that, differently from prior results, we explicitly get a convergence rate for the surrogate function ϕ, the function actually minimized by (1). The rates on the original function f are immediate by just taking the square root.

If f is G-Lipschitz, then $\sum_{t=1}^T \eta_t \geq T/G^2$. Hence, the final rate is $\phi(\bar{\boldsymbol{x}}_T) \leq \frac{G^2}{2T}\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2$. So, the surrogate loss converges as $\mathcal{O}(1/T)$, as expected by a loss with upper curvature.

Now, instead let's assume that the function f is L-self-bounded. Using inequalities between harmonic and arithmetic means, we have

$$\frac{1}{\sum_{t=1}^T \frac{1}{\|\boldsymbol{g}_{\boldsymbol{x}_t}\|^2}} \leq \frac{1}{T^2}\sum_{t=1}^T \|\boldsymbol{g}_{\boldsymbol{x}_t}\|^2 = \frac{1}{T^2}\sum_{t=1}^T \frac{\|\boldsymbol{g}_{\boldsymbol{x}_t}\|^4}{\|\boldsymbol{g}_{\boldsymbol{x}_t}\|^2} \leq \frac{1}{T^2}\sum_{t=1}^T \frac{8L^2\phi(\boldsymbol{x}_t)}{\|\boldsymbol{g}_{\boldsymbol{x}_t}\|^2} \leq \frac{8L^2}{T^2}\frac{1}{2}\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2,$$

where in the last inequality we use (4). This implies

$$\phi(\bar{\boldsymbol{x}}_T) \leq \frac{1}{2\sum_{t=1}^T \eta_t}\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2 = \frac{1}{2\sum_{t=1}^T \frac{1}{\|\boldsymbol{g}_{\boldsymbol{x}_t}\|^2}}\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2 \leq \frac{4L^2}{T^2}\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^4 .$$

Similarly, it is equally easy to obtain rates for Hölder-smooth functions, see Theorem 7 for more details.

We can also assume that f is s-sharp and G-Lipschitz. So, from (3) and the fact that ϕ has s^2 quadratic growth from Theorem 1, then by Lemma 1 we have

$$\frac{s^2}{G^2}\frac{1}{2}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 \leq s^2\eta_t\frac{1}{2}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 \leq \eta_t(\phi(\boldsymbol{x}_t) - \phi(\boldsymbol{x}^\star)) \leq \frac{1}{2}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 - \frac{1}{2}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star\|^2 .$$

Using the fact that $\frac{s^2}{G^2} \leq 1$, this immediately gives a linear convergence rate.

5 Generalizing the Polyak Stepsize: More Surrogates and Stochastic Setting

We have shown how the Polyak stepsize is just GD on a particular function with stepsizes adapted to the local curvature of the function. In this section, we show that we can construct an entire family of surrogate losses with similar guarantees, while also preparing ourselves for the stochastic setting.

Instead of the function (2), we consider more generally the surrogate

$$\psi(\boldsymbol{x}) = \frac{1}{2}h^2(\boldsymbol{x}),$$

where $h : \mathbb{R}^d \to \mathbb{R}_{\geq}0$ is convex. As a special case we can recover (2) with $h(\boldsymbol{x}) = f(\boldsymbol{x}) - f^\star$, however, in general we do not need to know $f^\star$. For example we can take $h = (f(\boldsymbol{x}) - a)_+$ for any a. Intuitively, the role of h is to transform f into a positive function. We show that ψ generally has an approximate local-star-upper curvature, where the approximation stems from h potentially being strictly positive in $\boldsymbol{x}^\star$.

Definition 4 (Approximate local-star-upper curvature)**.** *We will say that a function f with minimizer $\boldsymbol{x}^\star$ has ϵ-approximate $\lambda_{\boldsymbol{y}}$-star-upper-curvature around $\boldsymbol{y}$ if there exists ϵ such that*

$$f(\boldsymbol{x}^\star) - f(\boldsymbol{y}) - \langle \boldsymbol{g}_{\boldsymbol{y}}, \boldsymbol{x}^\star - \boldsymbol{y}\rangle \geq \frac{1}{2\lambda_{\boldsymbol{y}}}\|\boldsymbol{g}_{\boldsymbol{y}}\|_2^2 - \epsilon, \ \forall \boldsymbol{g}_{\boldsymbol{y}} \in \partial f(\boldsymbol{y}) \ .$$

Since we do not make explicit assumptions h with respect to f, we can only hope to achieve convergence to the minimum of h or ψ. So, from here onward we denote $\boldsymbol{x}^\star$ as as minizer of ψ.

Lemma 2. *Let $h : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ be convex. Define $\psi = \frac{1}{2}h^2$. Then, ψ is $(2\sqrt{\psi(\boldsymbol{x})\psi(\boldsymbol{x}^\star)} - \psi(\boldsymbol{x}^\star))$-approximate $\|\boldsymbol{g}\|$-LSUC for any $\boldsymbol{g} \in \partial h(\boldsymbol{x})$.*

Proof. Given that the function $\psi(\boldsymbol{x})$ might not be differentiable, we have to be careful in the calculation of its subgradients. We have

$$\begin{aligned}\psi(\boldsymbol{y}) &= \frac{1}{2}h^2(\boldsymbol{y}) \geq \frac{1}{2}h^2(\boldsymbol{x}) + h(\boldsymbol{x})[h(\boldsymbol{y}) - h(\boldsymbol{x})] \\ &\geq \frac{1}{2}h^2(\boldsymbol{x}) + h(\boldsymbol{x})\langle \boldsymbol{g}, \boldsymbol{y} - \boldsymbol{x}\rangle = \psi(\boldsymbol{x}) + \langle h(\boldsymbol{x})\boldsymbol{g}, \boldsymbol{y} - \boldsymbol{x}\rangle,\end{aligned}$$

where $\boldsymbol{g} \in \partial h(\boldsymbol{x})$ and the first inequality is due to the fact that $\frac{1}{2}(\cdot)^2$ is a convex function. Hence, we see that $\tilde{\boldsymbol{g}} := h(\boldsymbol{x})\boldsymbol{g}$ is a subgradient of ψ in $\boldsymbol{x}$. Hence, for any $\boldsymbol{u} \in \mathbb{R}^d$, we have

$$\begin{aligned}\psi(\boldsymbol{x}) - \langle \tilde{\boldsymbol{g}}, \boldsymbol{x} - \boldsymbol{u}\rangle + \frac{1}{2\|\boldsymbol{g}\|^2}\|\tilde{\boldsymbol{g}}\|_2^2 &= \frac{1}{2}h^2(\boldsymbol{x})) - h(\boldsymbol{x})\langle \boldsymbol{g}, \boldsymbol{x} - \boldsymbol{u}\rangle + \frac{1}{2}h(\boldsymbol{x})^2 \\ &= h(\boldsymbol{x})\left(h(\boldsymbol{x}) - \langle \boldsymbol{g}, \boldsymbol{x} - \boldsymbol{u}\rangle\right) \\ &= h(\boldsymbol{x})\left(h(\boldsymbol{x}) - h(\boldsymbol{u}) - \langle \boldsymbol{g}, \boldsymbol{x} - \boldsymbol{u}\rangle + h(\boldsymbol{u})\right) \\ &\leq h(\boldsymbol{x}) \cdot h(\boldsymbol{u}) = 2\sqrt{\psi(\boldsymbol{x})\psi(\boldsymbol{u})},\end{aligned}$$

where the inequality is due to the convexity of h and the fact that $h(\boldsymbol{x}) \geq 0$. Setting $\boldsymbol{u} = \boldsymbol{x}^\star$, we have the stated bound. □

With approximate local curvature, a generalization of Lemma 1 is immediate.

Lemma 3. *Assume $\psi : \mathbb{R}^d \to \mathbb{R}$ to be ϵ_t-approximately λ_t-star-upper-curve around $\boldsymbol{x}_t$. Then, for any $\eta_t > 0$, any $\boldsymbol{g}_t \in \partial\psi(\boldsymbol{x}_t)$, and $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta_t\boldsymbol{g}_t$ we have*

$$\eta_t(\psi(\boldsymbol{x}_t) - \psi(\boldsymbol{x}^\star)) \leq \frac{\|\boldsymbol{x}^\star - \boldsymbol{x}_t\|_2^2}{2} - \frac{\|\boldsymbol{x}^\star - \boldsymbol{x}_{t+1}\|_2^2}{2} + \frac{\eta_t}{2}\left(\eta_t - \frac{1}{\lambda_t}\right)\|\boldsymbol{g}_t\|^2 + \eta_t\epsilon_t \ .$$

The last two lemmas tell us that the properties of the surrogate functions breaks if $\psi(\boldsymbol{x}^\star) \neq 0$. Hence, we will not be able to prove convergence results, but only that, for example, the suboptimality gap will converge up to a floor that depends on $\psi(\boldsymbol{x}^\star)$. However, in Section 6 we will show that this is not an artifact of the proof. Indeed, we can construct simple one-dimensional functions where the generalized Polyak stepsize does not converge.

5.1 Stochastic Approximation Setting

Consider now the case that we are minimizing $F(\boldsymbol{x}) := \mathbb{E}_{\boldsymbol{\xi}\sim D}[f(\boldsymbol{x}, \boldsymbol{\xi})]$, where $f : \mathbb{R}^d \times \mathcal{S} \to \mathbb{R}$, that covers both the stochastic approximation and finite-sum settings. We do not know the distribution D, but we assume that we can sample $\boldsymbol{\xi}$ i.i.d. from D.

In this setting, we argue that the Polyak stepsize makes sense only in restricted settings. In fact, the interpretation of the Polyak stepsize as minimizing a surrogate function implies that in the stochastic setting we will minimize the function $\mathbb{E}_{\boldsymbol{\xi}\sim D}[\frac{1}{2}h^2(\boldsymbol{x}, \boldsymbol{\xi})]$, where the function $h(\cdot, \boldsymbol{\xi})$ depends on the particular variant of the stochastic Polyak stepsize. It is clear that in general $\operatorname{argmin}_{\boldsymbol{x}} \ \mathbb{E}_{\boldsymbol{\xi}\sim D}[f(\boldsymbol{x}, \boldsymbol{\xi})]$ can be completely different from $\operatorname{argmin}_{\boldsymbol{x}} \ \mathbb{E}_{\boldsymbol{\xi}\sim D}[\frac{1}{2}h^2(\boldsymbol{x}, \boldsymbol{\xi})]$.

Here, starting from ALI-G [5] and $\texttt{SPS}_{\max}$ [29] that use the idea of limiting the stepsizes, we propose a generalized Polyak stepsize algorithm. The proof is in Appendix C.

Algorithm 1 Generalized Polyak Stepsize

Require: $h : \mathbb{R}^d \times \mathcal{S} \to \mathbb{R}$, $\boldsymbol{x}_1 \in \mathbb{R}^d$

for $t = 1, \dots, T$ **do**
  Sample $\boldsymbol{\xi}_t$ from D
  Tranform $f(\boldsymbol{x}, \boldsymbol{\xi}_t)$ into $h(\boldsymbol{x}, \boldsymbol{\xi}_t)$
  Receive $\boldsymbol{g}_t \in \partial h(\boldsymbol{x}, \boldsymbol{\xi}_t)$
  if $\boldsymbol{g}_t \neq \mathbf{0}$ **then**
    $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta_t h(\boldsymbol{x}_t, \boldsymbol{\xi}_t)\boldsymbol{g}_t$ where $\eta_t = \min\left(\frac{1}{\|\boldsymbol{g}_t\|^2}, \frac{\gamma}{h(\boldsymbol{x}_t, \boldsymbol{\xi}_t)}\right)$
  else
    $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t$
  end if
end for

Theorem 3. *Let $h : \mathbb{R}^d \times \mathcal{S} \to \mathbb{R}_{\geq 0}$ be convex in its first argument. Denote by $H(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\xi} \sim D}[h(\boldsymbol{x}, \boldsymbol{\xi})]$. Then, setting $\eta_t = \min\left(\frac{1}{\|\boldsymbol{g}_t\|^2}, \frac{\gamma}{h(\boldsymbol{x}_t, \boldsymbol{\xi}_t)}\right)$ in Algorithm 1, we have*

- *If $h(\cdot, \boldsymbol{\xi}_t)$ is L-self bounded, we have*

$$\frac{1}{T}\sum_{t=1}^{T} \min\left(\frac{1}{2L}, \gamma\right) \mathbb{E}[H(\boldsymbol{x}_t)] \leq \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{T} + 2\gamma H(\boldsymbol{x}^\star)$$

  and

$$\sum_{t=1}^{T} \min\left(\frac{1}{2L}, \gamma\right) \mathbb{E}[H(\boldsymbol{x}_t)] - \gamma \sum_{t=1}^{T} H(\boldsymbol{x}^\star) \leq \frac{1}{2}\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2 + \frac{1}{2}\sum_{t=1}^{T} \gamma^2 \mathbb{E}[\|\boldsymbol{g}_t\|^2] \,.$$

- *If $h(\cdot, \boldsymbol{\xi}_t)$ is G-Lipschitz, then we have*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[H(\boldsymbol{x}_t)] \leq \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{\gamma T} + 2H(\boldsymbol{x}^\star) + \frac{G\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|}{\sqrt{T}} + G\sqrt{2\gamma H(\boldsymbol{x}^\star)} \,.$$

- *If $h(\cdot, \boldsymbol{\xi})$ is L-self-bounded and $H(\boldsymbol{x})$ has μ-quadratic growth, then*

$$\mathbb{E}\left[\|\boldsymbol{x}_{T+1} - \boldsymbol{x}^\star\|^2\right] \leq \mathbb{E}\left[\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2\right] a^{T+1} + b\frac{1 - a^{T+1}}{1 - a} H(\boldsymbol{x}^\star),$$

  where $a = \frac{\mu}{2}\min\left(\frac{1}{2L}, \gamma\right)$ and $b = 2\gamma - \min\left(\frac{1}{2L}, \gamma\right)$.

Choosing the function h, the above theorem covers and extends a number of results in previous papers, for example:

- $\texttt{SPS}_{\max}$ [29]: $h(\boldsymbol{x}, \boldsymbol{\xi}) = f(\boldsymbol{x}, \boldsymbol{\xi}) - \inf_{\boldsymbol{x}} \; f(\boldsymbol{x}, \boldsymbol{\xi})$, so $H(\boldsymbol{x}^\star) = \mathbb{E}[f(\boldsymbol{x}^\star, \boldsymbol{\xi}) - \inf_{\boldsymbol{x}} \; f(\boldsymbol{x}, \boldsymbol{\xi})]$.
- $\texttt{SPS}_{\max}^{\ell}$ [35]: $h(\boldsymbol{x}, \boldsymbol{\xi}) = f(\boldsymbol{x}, \boldsymbol{\xi}) - q(\boldsymbol{\xi})$, where $q(\boldsymbol{\xi})$ is a lower bound to $\inf_{\boldsymbol{x}} \; f(\boldsymbol{x}, \boldsymbol{\xi})$. In this case, $H(\boldsymbol{x}^\star) = \mathbb{E}[f(\boldsymbol{x}^\star, \boldsymbol{\xi}) - q(\boldsymbol{\xi})]$.
- $\texttt{SPS}_{+}$ [15]: $h(\boldsymbol{x}, \boldsymbol{\xi}) = (f(\boldsymbol{x}, \boldsymbol{\xi}) - f(\boldsymbol{x}^\star, \boldsymbol{\xi}))_+$. In this case, $H(\boldsymbol{x}^\star) = 0$ so we can also safely set $\gamma = \infty$. Moreover, $H(\boldsymbol{x}) \geq F(\boldsymbol{x}) - F(\boldsymbol{x}^\star)$, hence any bound on $H(\boldsymbol{x})$ translates to a bound on the suboptimality gap.

Remark 1. *If $h(\boldsymbol{x}^\star, \boldsymbol{\xi}_t) = 0$ for all t, then $\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star\| \leq \|\boldsymbol{x}_t - \boldsymbol{x}^\star\|$. Hence, in this case we only need to consider all the properties of h in the bounded domain $\{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x} - \boldsymbol{x}^\star\| \leq \|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|\}$. This is well-known via the subgradient projector perspective, as $\boldsymbol{x}_t$ is guaranteed to approach the set $\bigcap_{\boldsymbol{\xi}}\{\boldsymbol{x} : h(\boldsymbol{x}, \boldsymbol{\xi}) = 0\}$ at each iteration if it is non-empty. This property was observed in Gower et al. [19] for $\texttt{SPS}_{+}$ [15], but here it holds more generally. For example, $h(\boldsymbol{x}, \boldsymbol{\xi}) = (f(\boldsymbol{x}, \boldsymbol{\xi}) - a)_+$, where $a \geq \sup_{\xi} f(\boldsymbol{x}^\star, \boldsymbol{\xi})$, would also have no neighbourhood of convergence and satisfies the assumption of the theorem.*

Remark 2. *The above theorem also applies to the case where some of the $f(\cdot, \xi)$ are non-convex functions, while still guaranteeing the convergence to the global optimum of F. For example, consider $F(x) = 0.5f_1(x) + 0.5f_2(x)$, where $f_1 = -|x|$ (non-convex) and $f_2 = 2|x|$. We have that $F(x) = \frac{1}{2}|x|$ so $\boldsymbol{x}^\star = 0$. Now, choose $h_1(x) = \max(f_1(x) - f_1(x^\star), 0) = 0$ and $h_2(x) = \max(f_2(x) - f_2(x^\star), 0) = 2|x|$. Hence, the hypotheses of the theorem are verified. Moreover, $H(x) = 0.5h_1(x) + 0.5h_2(x) \geq F(x)$ and $H(x^\star) = 0$, so the theorem implies a convergence rate for the minimization of $F(x) - F(x^\star)$ by using $\texttt{SPS}_+$.*

Remark 3. *In the proof of Theorem 3, if one stops before taking expectations, one obtains a* regret guarantee *on a sequence of arbitrary losses $h(\boldsymbol{x}, \boldsymbol{\xi}_t)$. Such regret scales as the sum of the loss in $\boldsymbol{x}^\star$. This is exactly the $L^\star$ bound that we mentioned in Section 2. Indeed, this kind of updates and guarantees were already obtained in the online learning literature for the special case of linear predictors by the Passive-Aggressive family of algorithms [11].*

Besides covering a number of previous algorithmic variants, we also extend the previous known guarantees. In particular, Loizou et al. [29] only studied SPS in the non-smooth setting but did not include $\texttt{SPS}_{\max}$.[4] Theorem 3 shows for the first time that $\texttt{SPS}_{\max}$ is adaptive to the entire range of upper curvature of the function, from Lipschitz to smooth functions. In Appendix D we also show additional results. Moreover, the second result in the smooth case is new, and it allows to recover the SGD guarantee on H when γ is sufficiently small. For example, we include a precise statement for $\texttt{SPS}_+$ when $f(\boldsymbol{x}, \boldsymbol{\xi})$ is L-self-bounded, that recovers the guarantee in Gower et al. [19, Corollary 2.3].

6 Neighbourhood of Convergence and Instability of the Polyak Stepsize

In Section 5 we demonstrate that a generalized version of the Polyak stepsize and existing variants can be viewed as GD on a function with approximate local curvature, with convergence to a neighbourhood of the optimal solution. This neighbourhood of convergence appears in our analysis just like in all existing variants, therefore *suggesting* it is unavoidable if

$$H(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\xi} \sim D}[h(\boldsymbol{x}, \boldsymbol{\xi})] > 0 \text{ for all } \boldsymbol{x} \,. \tag{5}$$

In this section, we demonstrate that this neighbourhood of convergence is not an artifact of the analysis and indeed cannot be avoided in general. We also show that the positivity condition (5) can fundamentally change the dynamics of Algorithm 1, even in the deterministic setting, thus posing a challenge that is not just associated with interpolation.

Condition (5) occurs with SPS [29] without interpolation, or in the deterministic setting when the optimal value is underestimated, $h(\boldsymbol{x}) = f(\boldsymbol{x}) - c$ where $f^\star > c$. Convergence under this condition was first studied in the deterministic case in Polyak's original paper [37], where it is shown that if $\inf_x h(x) = h^\star > 0$ then $\lim_{t \to \infty} \min_{1 \leq s \leq t} h(\boldsymbol{x}_t) - h^\star \leq h^\star$. That is, the best iterate eventually enters a neighbourhood of the minima where the size of the neighbourhood is dependent on how much $h^\star$ is underestimated by 0. In the stochastic case, convergence of the average iterate to a neighbourhood when understimating the minimum was also studied by Orvieto et al. [35] under $\texttt{SPS}^{\ell}_{\max}$. However, we demonstrate the consequence of condition (5) is far greater than existing results suggest, with instability of fixed points, potential cycles, lower bounds in the sub-optimality gap, and lack of convergence regardless of initialization.

Deterministic Setting. We first demonstrate that in the deterministic setting, for different classes of h, if $h^\star = \min_x h(\boldsymbol{x}) > 0$, then the fixed points of

$$\boldsymbol{x}_{t+1} = T(\boldsymbol{x}_t) := \begin{cases} \boldsymbol{x}_t - \frac{h(\boldsymbol{x}_t)}{\|\boldsymbol{g}_t\|^2} \boldsymbol{g}_t, & \text{if } \boldsymbol{g}_t \in \partial h(\boldsymbol{x}_t),\ 0 \notin \partial h(\boldsymbol{x}_t), \\ \boldsymbol{x}_t, & \text{otherwise} \end{cases} \tag{6}$$

are unstable. Intuitively, this can be explained via our surrogate function view: In fact, denoting the local curvature constant of the surrogate $\frac{1}{2}(h(\boldsymbol{x}) - h^\star)^2$ around $\boldsymbol{x}_t$ as λ_t, we see that the stepsize η_t in update (6) can be equivalently written as

$$\eta_t = \left(\frac{h(\boldsymbol{x}_t)}{h(\boldsymbol{x}_t) - h^\star} \right) \frac{1}{\lambda_t}. \tag{7}$$

[4] Although Loizou et al. [29] state that SPS analysis can be readily extended to $\texttt{SPS}_{\max}$ this does not seem to be the case due to the non-convexity of the min function, as demonstrated by our different proof technique in Appendix C.

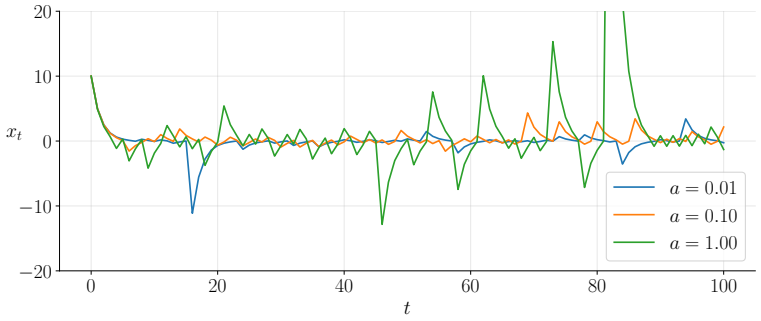

Figure 2: Trajectories under T (6) for $h(x) = \frac{x^2}{2} + a$ with an unstable fixed point at $x^\star = 0$. Lack of convergence is observed for different values of a as predicted by Proposition 3.

If h is Lipschitz or self-bounded then $\eta_t \to +\infty$ as $\boldsymbol{x}_t \to \boldsymbol{x}^\star$. Therefore, if h possesses curvature, then $\boldsymbol{x}_{t+1}$ may move further away from $\boldsymbol{x}^\star$ within a neighborhood of $\boldsymbol{x}^\star$. Indeed, in Proposition 1 we show that for all self-bounded functions with a quadratic growth condition, the fixed point of T in (6) is unstable. A similar result can also be shown if h is L-Lipschitz and has a sharp mininum (see Proposition 6 in the Appendix).

Proposition 1 (Unstable fixed point)**.** *Suppose h is convex, strictly positive, L-self-bounded, and satisfies the quadratic growth condition $h(\boldsymbol{x}) - h^\star \geq \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{x}^\star\|^2$, where $\boldsymbol{x}^\star = \arg\min_{\boldsymbol{x}} h(\boldsymbol{x})$ is the only fixed point of T, defined in* (6)*. Then, for any point $\boldsymbol{x} \in S = \{\boldsymbol{y} : \boldsymbol{y} \neq \boldsymbol{x}^\star, h(\boldsymbol{y}) - h^\star < h^\star \frac{\mu}{8L-\mu}\}$ we have*

$$\|T(\boldsymbol{x}) - \boldsymbol{x}^\star\| > \|\boldsymbol{x} - \boldsymbol{x}^\star\| .$$

Note that this reinforces the need to clip the stepsize as proposed in ALI-G and $\texttt{SPS}_{\max}$. However, clipping will not remove this behaviour unless the maximum value is taken to be small enough. In Proposition 8 we show there is always a subregion where the stepsize is bounded, and this subregion can be made arbitrarily large within the unstable region; therefore, if the clipped value is too large instability is unavoidable.

The importance of $h > 0$ in update (6) has also been studied in Bauschke et al. [3], where they demonstrate with examples that T can fail to be quasi-firmly non-expansive if $h^\star > 0$. Propositions 1, and 6 provide extra insight on this phenomenon as they automatically prove T cannot be quasi-firmly non-expansive and therefore we have the following remark.

Remark 4. *If $h > 0$, and either of the following conditions hold:*

- *h is convex, self-bounded, and satisfies the quadratic growth condition,*
- *h is convex, Lipschitz, and has a sharp minimum,*

then T from (6) *is not quasi-firmly non-expansive.*

While Propositions 1 and 6 establish that minima can be unstable, this property may not fully describe the dynamics of update (6). In fact, instability can admit convergence in the average iterate or last iterate if the local critical neighborhood is skipped. So, in Proposition 2 we provide an example of a function h, which satisfies the assumptions of Proposition 1, where the iterates *cycle and never reach the minimum in best iterate or on average*.

Proposition 2 (Cycling and failure to converge)**.** *There exists a strictly positive smooth and strongly convex function h, and an initial point $\boldsymbol{x}_1$ such that iterates from update* (6) *cycle and satisfy the inequality $h(\frac{1}{t}\sum_{i=1}^t \boldsymbol{x}_i) - h^\star \geq \delta > 0$ for all t.*

Note that since the cycle in Proposition 2 consists of a finite number of points, clipping will not necessarily remove this behaviour (e.g. if the clipped value is taken to be larger than any of the seen stepsizes). In Proposition 2, a specific initialization was chosen to construct a cycle that would not converge. However, in Proposition 3 we show that for 1-d quadratics the lack of convergence is true for all initializations and values of $h^\star$ *up to a set of measure zero*.

Proposition 3 (The set of good initializations can have measure zero)**.** *Let $h : \mathbb{R} \to \mathbb{R}$, defined as $h(x) = \frac{x^2}{2} + a$ for $a > 0$, where $x_{t+1} = x_t - \frac{h(x_t)}{\|\nabla h(x_t)\|^2}\nabla h(x_t)$, and $\boldsymbol{x}_1$ is randomly initialized. Then, $P\{\lim_{t\to\infty} x_t = x^\star\} = 0$. In other words, the set of initializations that can converge to the optimal solution has measure zero.*

Proof. Note h is 1-smooth and 1-strongly convex and therefore satisfies the conditions of Proposition 1 with $\mu = L = 1$ and an unstable unique fixed point. Let T be such that $x_{t+1} = T(x_t)$. $T(x) = x\left(\frac{1}{2} - \frac{a}{x^2}\right)$ if $x \neq 0$ and 0 otherwise. With inverse $T^{-1}(S) = \{x \pm \sqrt{x^2 + 2a} : x \in S\}$. Therefore, $T^{-k}(\{x^\star\})$ has at most 2^k points which has measure zero for all k. By Lemma 5 in Appendix E, the result follows. □

Remark 5. *Note that, by Lemma 5, Proposition 3 can be extended much more generally if T from* (6) *is shown to satisfy the Lusin (N^{-1}) property (see Definition 7) [23, Definition 4.12].*

In the stochastic case with SPS, condition (5) is due to lack of interpolation, and Orvieto et al. [35] show that it can change the expected fixed point. In contrast, in the deterministic setting we have shown lack of convergence despite the fixed point being $\boldsymbol{x}^\star$. Therefore the issue here stems from the instability of the method due to the underestimation of $h^\star$ and not the bias of the expected update.

Stochastic Setting. In the stochastic setting, the positivity condition (5) can occur despite $\min_{\boldsymbol{x}} h(\boldsymbol{x}, \boldsymbol{\xi}) = 0$, such as in SPS ($h(\boldsymbol{x}, \boldsymbol{\xi}) = f(\boldsymbol{x}) - \min_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{\xi})$) without interpolation. Orvieto et al. [35] demonstrate that without interpolation SPS can fail to converge in a 1-d quadratic and has an expected fixed point different than $\min_{\boldsymbol{x}} F(\boldsymbol{x}) = \min_{\boldsymbol{x}} \mathbb{E}_{\boldsymbol{\xi}\sim D}[f(\boldsymbol{x}, \boldsymbol{\xi})]$. Similarly to the deterministic setting, we can show that SPS can have a random walk between a finite number of points.

Proposition 4 (Failure to converge)**.** *There exist f_1 and f_2 quadratic 1-d functions and a starting point x_1 such that SPS on $F(x) = 0.5(f_1(x) + f_2(x))$ satisfies*

$$\mathbb{E}[F(x_t)] - \min_x \; F(x) \geq 2/3, \; \forall t \; .$$

Proof. Let $f_1 = x^2 + 2x + 5$ and $f_2(x) = 2x^2 - 4x + 10$. Let's start from $x_1 = 1$ where $F(x_1) = 8$. If we draw f_1, $x_2 = -1$, while if we draw f_2 then $x_2 = 1$ because $f_2'(1) = 0$. Hence, $\mathbb{E}[F(x_2)] = 0.5\frac{f_1(1)+f_2(1)}{2} + 0.5\frac{f_1(-1)+f_2(-1)}{2} = 9$. Iterating, we have that x_3 has equal probability to be equal to 1 and -1. Hence, again we have $\mathbb{E}[F(x_3)] = 9$. So, we have that this holds for any t. Moreover, we have that $\min_x F(x) = 44/6$. □

7 Discussion and Limitations

We have shown that the design, properties, and failure of the (variants) of the Polyak stepsize can be easily derived through the lens of the minimization of a surrogate objective function. This framework also provides a new and natural explanation on the adaptivity of the stepsize via the local curvature of the surrogate. We believe this framework has the promise to design new variants, by simply designing surrogate functions with the required properties. Furthermore, with our perspective we have provided new insight on the challenge of controlling neighbourhoods of convergence that often appear in variants of the Polyak stepsize. We demonstrate that this neighbourhood is unavoidable and a fundamental issue causing instability. Moreover, we show that this issue is not due to the lack of interpolation, as commonly believed, but instead because the minimum of the surrogate loss is not zero more generally.

The limitations of our framework include the assumption of convex h in the generalized surrogate that must be assumed apriori. It is unclear if this framework can be extended to the more general case of noncovex surrogate functions. The class of such surrogates that admit fast rates and tight neighbourhoods of convergence remains an open question that we leave to future work.

Acknowledgments

We acknowledge the use of Gemini 2.5 in developing the proof of Proposition 2. We also thank Mehdi Inane Ahmed for helpful discussions. Ryan D'Orazio's work is funded by Ioannis Mitliagkas' CIFAR chair.

References

[1] Farshed Abdukhakimov, Chulu Xiang, Dmitry Kamzolov, and Martin Takáč. Stochastic gradient descent with preconditioned Polyak step-size. *Computational Mathematics and Mathematical Physics*, 64(4):621–634, 2024.

[2] Heinz H Bauschke and Patrick L Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2017.

[3] Heinz H Bauschke, Caifang Wang, Xianfu Wang, and Jia Xu. Subgradient projectors: extensions, theory, and characterizations. *Set-Valued and Variational Analysis*, 26:1009–1078, 2018.

[4] A. Beck. *First-order methods in optimization*. SIAM, 2017.

[5] Leonard Berrada, Andrew Zisserman, and M. Pawan Kumar. Training neural networks for and by interpolation. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 799–809. PMLR, 13–18 Jul 2020.

[6] Dimitri Bertsekas. *Nonlinear Programming*, volume 2. Athena Scientific, 1999.

[7] Ulf Brännlund. On relaxation methods for nonsmooth convex optimization, 1993.

[8] Andrzej Cegielski. *Iterative methods for fixed point problems in Hilbert spaces*, volume 2057. Springer, 2012.

[9] Yair Censor. Iterative methods for the convex feasibility problem. In M. Rosenfeld and J. Zaks, editors, *Annals of Discrete Mathematics (20): Convexity and Graph Theory*, volume 87 of *North-Holland Mathematics Studies*, pages 83–91. North-Holland, 1984.

[10] Patrick L. Combettes. *Fejér monotonicity in convex optimization*, pages 1016–1024. Springer US, Boston, MA, 2009.

[11] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7:551–585, 2006.

[12] Damek Davis, Dmitriy Drusvyatskiy, Kellie J MacPhee, and Courtney Paquette. Subgradient methods for sharp weakly convex functions. *Journal of Optimization Theory and Applications*, 179(3):962–982, 2018.

[13] Ryan D'Orazio, Nicolas Loizou, Issam H. Laradji, and Ioannis Mitliagkas. Stochastic mirror descent: Convergence analysis and adaptive variants via the mirror stochastic polyak stepsize. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

[14] Yu. M. Ermol'ev. Methods of solution of nonlinear extremal problems. *Cybernetics*, 2(4):1–14, July 1966. ISSN 1573-8337. doi: 10.1007/BF01071403.

[15] G. Garrigos, R. M. Gower, and F. Schaipp. Function value learning: Adaptive learning rates based on the polyak stepsize and function splitting in ERM. *arXiv preprint arXiv:2307.14528*, 2023.

[16] Jean-Louis Goffin and Krzysztof C. Kiwiel. Convergence of a simple subgradient level method. *Mathematical Programming*, 85(1):207–211, May 1999. ISSN 1436-4646. doi: 10.1007/s101070050053.

[17] Eduard Gorbunov, Nazarii Tupitsa, Sayantan Choudhury, Alen Aliev, Peter Richtárik, Samuel Horváth, and Martin Takáč. Methods for convex (L_0, L_1)-smooth optimization: Clipping, acceleration, and adaptivity. In *The Thirteenth International Conference on Learning Representations*, 2025.

[18] B. Goujaud, A. Taylor, and A. Dieuleveut. Optimal first-order methods for convex functions with a quadratic upper bound. *arXiv preprint arXiv:2205.15033*, 2022.

[19] R. M. Gower, G. Garrigos, N. Loizou, D. Oikonomou, K. Mishchenko, and F. Schaipp. Analysis of an idealized stochastic Polyak method and its application to black-box model distillation. *arXiv preprint arXiv:2504.01898*, 2025.

[20] Robert M Gower, Aaron Defazio, and Michael Rabbat. Stochastic polyak stepsize with a moving target. *arXiv preprint arXiv:2106.11851*, 2021.

[21] C. Guille-Escuret, M. Girotti, B. Goujaud, and I. Mitliagkas. A study of condition numbers for first-order optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 1261–1269. PMLR, 2021.

[22] E. Hazan and S. Kakade. Revisiting the Polyak step size. *arXiv preprint arXiv:1905.00313*, 2019.

[23] Stanislav Hencl and Pekka Koskela. *Lectures on mappings of finite distortion*, volume 2096. Springer, 2014.

[24] Samuel Horváth, Konstantin Mishchenko, and Peter Richtárik. Adaptive learning rates for faster stochastic gradient methods. *arXiv preprint arXiv:2208.05287*, 2022.

[25] Xiaowen Jiang and Sebastian U Stich. Adaptive SGD with Polyak stepsize and line-search: Robust convergence and variance reduction. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[26] Pooria Joulani, András György, and Csaba Szepesvári. A modular analysis of adaptive (non-)convex optimization: Optimism, composite objectives, variance reduction, and variational bounds. *Theoretical Computer Science*, 808:108–138, 2020. ISSN 0304-3975. Special Issue on Algorithmic Learning Theory.

[27] Krzysztof C. Kiwiel. Convergence of approximate and incremental subgradient methods for convex optimization. *SIAM Journal on Optimization*, 14(3):807–840, 2004.

[28] X. Li and F. Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. In *Proc. of the 22nd International Conference on Artificial Intelligence and Statistics, AISTATS*, 2019.

[29] N. Loizou, S. Vaswani, I. H. Laradji, and S. Lacoste-Julien. Stochastic Polyak step-size for SGD: An adaptive learning rate for fast convergence. In *International Conference on Artificial Intelligence and Statistics*, pages 1306–1314. PMLR, 2021.

[30] Aaron Mishkin, Ahmed Khaled, Yuanhao Wang, Aaron Defazio, and Robert M. Gower. Directional smoothness and gradient methods: Convergence and adaptivity. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[31] Angelia Nedic and Dimitri P. Bertsekas. Incremental subgradient methods for nondifferentiable optimization. *SIAM Journal on Optimization*, 12(1):109–138, 2001.

[32] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer, 2004.

[33] Dimitris Oikonomou and Nicolas Loizou. Stochastic Polyak step-sizes and momentum: Convergence guarantees and practical performance. In *The Thirteenth International Conference on Learning Representations*, 2025.

[34] F. Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019. Version 7.

[35] A. Orvieto, S. Lacoste-Julien, and N. Loizou. Dynamics of SGD with stochastic Polyak stepsizes: Truly adaptive variants and convergence to exact solution. In *Advances in Neural Information Processing Systems*, volume 35, pages 26943–26954, 2022.

[36] B. T. Polyak. *Introduction to Optimization*. Translations Series in Mathematics and Engineering. Optimization Software, Inc., 1987.

[37] Boris Teodorovich Polyak. Minimization of unsmooth functionals. *USSR Computational Mathematics and Mathematical Physics*, 9(3):14–29, 1969.

[38] Mariana Prazeres and Adam M. Oberman. Stochastic gradient descent with Polyak's learning rate. *Journal of Scientific Computing*, 89(1):25, Sep 2021.

[39] Michal Rolinek and Georg Martius. L4: Practical loss-based stepsize adaptation for deep learning. *Advances in neural information processing systems*, 31, 2018.

[40] Fabian Schaipp, Ruben Ohana, Michael Eickenberg, Aaron Defazio, and Robert M Gower. MoMo: Momentum models for adaptive learning rates. In *International Conference on Machine Learning*, pages 43542–43570. PMLR, 2024.

[41] Hanif D Sherali, Gyunghyun Choi, and Cihan H Tuncbilek. A variable target value method for nondifferentiable optimization. *Operations Research Letters*, 26(1):1–8, 2000.

[42] Yuki Takezawa, Han Bao, Ryoma Sato, Kenta Niwa, and Makoto Yamada. Parameter-free clipped gradient descent meets Polyak. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[43] Xiaoyu Wang, Mikael Johansson, and Tong Zhang. Generalized Polyak step size for first order optimization with momentum. In *International Conference on Machine Learning*, pages 35836–35863. PMLR, 2023.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: To the best of our knowledge our framework in Section 5 that both generalizes and analyzes the Polyak stepsize is novel. Additionally, we have included novel negative results in Section 6.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our framework and assumptions in Section 7.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All theoretical statements outline assumptions used in their respective proof. Complete proofs are either presented directly in the main body or in the appendix. We do not provide proof sketches for proofs not in the main body but do provide intuition and discussion in such cases.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [NA]

Justification: We do not include any experiments in our paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [NA]

   Justification: We do not include any experiments in our paper.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: We do not include any experiments in our paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: We do not include any experiments in our paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: We do not include any experiments in our paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: As our paper is theoretical we do not have any concerns regarding: research with human subjects, and data concerns. Nor do we foresee any societal impact.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: Due to the fundamental research of our paper we do not foresee broader societal impacts as per the guidelines below.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not use nor provide any models or data.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: No assets of the kind are used.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We do not use or release any such assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No human subjects are used.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No human subjects are used.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: An LLM was used to help derive result but not our core results or analyses.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Proofs for the Surrogate ϕ

Theorem 1 (Curvature of the Polyak surrogate)**.** *Let $f(x)$ be convex and define $\boldsymbol{x}^\star \in \operatorname{argmin}_{\boldsymbol{x}}\ f(\boldsymbol{x})$. Define $\phi(\boldsymbol{x}) = \frac{1}{2}(f(\boldsymbol{x}) - f(\boldsymbol{x}^\star))^2$. Then, we have*

- *ϕ is $\|\boldsymbol{g}_{\boldsymbol{y}}\|^2$-LSUC around any $\boldsymbol{y}$ for any $\boldsymbol{g}_{\boldsymbol{y}} \in \partial f(\boldsymbol{y})$.*
- *If f is s-sharp, then ϕ has s^2-quadratic growth.*
- *If f has μ-quadratic growth and L-self bounded, then ϕ satisfies a local quadratic growth:*

$$\phi(\boldsymbol{x}) \geq \frac{1}{2}\frac{\mu\|\boldsymbol{g}\|^2}{2L}\|\boldsymbol{x} - \boldsymbol{x}^\star\|^2,\ \forall \boldsymbol{g} \in \partial f(\boldsymbol{x})\ .$$

Proof.

$$\begin{aligned}
&\phi(\boldsymbol{y}) - \phi(\boldsymbol{x}^\star) - \langle \nabla\phi(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{x}^\star\rangle + \frac{1}{2\|\boldsymbol{g}_{\boldsymbol{y}}\|^2}\|\nabla\phi(\boldsymbol{y})\|_2^2 \\
&= \frac{1}{2}(f(\boldsymbol{y}) - f(\boldsymbol{x}^\star))^2 - (f(\boldsymbol{y}) - f(\boldsymbol{x}^\star))\langle \boldsymbol{g}_{\boldsymbol{y}}, \boldsymbol{y} - \boldsymbol{x}^\star\rangle + \frac{1}{2\|\boldsymbol{g}_{\boldsymbol{y}}\|^2}\|\nabla\phi(\boldsymbol{y})\|^2 \\
&= \frac{1}{2}(f(\boldsymbol{y}) - f(\boldsymbol{x}^\star))^2 - (f(\boldsymbol{y}) - f(\boldsymbol{x}^\star))\langle \boldsymbol{g}_{\boldsymbol{y}}, \boldsymbol{y} - \boldsymbol{x}^\star\rangle + \frac{(f(\boldsymbol{y}) - f(\boldsymbol{x}^\star))^2}{2} \\
&= (f(\boldsymbol{y}) - f(\boldsymbol{x}^\star))\left(f(\boldsymbol{y}) - f(\boldsymbol{x}^\star) - \langle \boldsymbol{g}_{\boldsymbol{y}}, \boldsymbol{y} - \boldsymbol{x}^\star\rangle\right) \\
&\leq 0,
\end{aligned}$$

where the inequality is due to the convexity of f and the fact that $f(\boldsymbol{y}) - f(\boldsymbol{x}^\star) \geq 0$.

For the second property, we have

$$f(\boldsymbol{x}) - f(\boldsymbol{x}^\star) \geq \alpha\|\boldsymbol{x} - \boldsymbol{x}^\star\| \Rightarrow \phi(\boldsymbol{x}) - \phi(\boldsymbol{x}^\star) \geq \frac{\alpha^2}{2}\|\boldsymbol{x} - \boldsymbol{x}^\star\|^2\ .$$

For the third property we have

$$\phi(\boldsymbol{x}) = \frac{1}{2}(f(\boldsymbol{x}) - f(\boldsymbol{x}^\star))^2 \geq \frac{\|\boldsymbol{g}\|^2}{2L}(f(\boldsymbol{x}) - f(\boldsymbol{x}^\star)) \geq \frac{1}{2}\frac{\mu\|\boldsymbol{g}\|^2}{2L}\|\boldsymbol{x} - \boldsymbol{x}^\star\|^2.$$

□

Remark 6. *Convergence of the last iterate follows from a classic argument with Fejér monotone sequences. From Lemma 1 we have that the distance to any solution is decreasing $\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star\| \leq \|\boldsymbol{x}_t - \boldsymbol{x}^\star\|$ for any minimizer $\boldsymbol{x}^\star$ of ϕ, that is, $\{\boldsymbol{x}_t\}_{t\geq 0}$ is a Fejér monotone sequence with respect to the solution set. Since we have $\phi(\boldsymbol{x}_t) \to 0$ and ϕ is continuous then for every limit point $\boldsymbol{x}'$ of the sequence it also holds that $\phi(\boldsymbol{x}') = 0$ implying $\boldsymbol{x}'$ is also a minimizer of ϕ. Therefore we can use the fact that if $\{\boldsymbol{x}_t\}_{t\geq 0}$ is Fejér monotone with respect to the solution set and the set contains all the limit points of the sequence then the sequence must converge to a point in the solution set (see Theorem 8.16 in Beck [4]).*

B Relationship between Star Upper Curvature and Upper Quadratic Growth

For a function f, denote by $\mathcal{X}^\star := \{\boldsymbol{x} : f(\boldsymbol{x}) = \min_{\boldsymbol{x}}\ f(\boldsymbol{x})\}$. In Guille-Escuret et al. [21], they define the following function class.

Definition 5. *A function f is L-quadratically upper bounded (denoted L-QG$^+$) if for all $\boldsymbol{x} \in \mathbb{R}^d$:*

$$f(\boldsymbol{x}) - f^\star \leq \frac{L}{2}\min_{\boldsymbol{x}^\star \in \mathcal{X}^\star}\|\boldsymbol{x} - \boldsymbol{x}^\star\|_2^2\ .$$

We now show that convex L-QG$^+$ are globally L-star upper curved, while the other direction is true for the local version of the two definitions.

Theorem 5. *Let f be a convex L-QG$^+$ function, then f is globally L-star upper curved. On the other hand, let f be $L_{\boldsymbol{x}}$-LSUC, then for all $\boldsymbol{x}$ we have*

$$f(\boldsymbol{x}) - f^\star \leq \frac{L_{\boldsymbol{x}}}{2} \min_{\boldsymbol{x}^\star \in \mathcal{X}^\star} \|\boldsymbol{x} - \boldsymbol{x}^\star\|_2^2 \,.$$

Proof. Assume that f is L-QG$^+$. Then, we have

$$f(\boldsymbol{y}) + \langle \boldsymbol{g}, \boldsymbol{x} - \boldsymbol{y} \rangle \leq f(\boldsymbol{x}) \leq f^\star + \frac{L}{2} \min_{\boldsymbol{x}' \in \mathcal{X}^\star} \|\boldsymbol{x} - \boldsymbol{x}'\|_2^2,$$

where the first inequality is due to convexity and $\boldsymbol{g} \in \partial f(\boldsymbol{y})$. Now, set $\boldsymbol{x} = \boldsymbol{x}^\star + \frac{1}{L}\boldsymbol{g}$ for any $\boldsymbol{x}^\star \in \mathcal{X}^\star$, to have

$$\begin{aligned} f(\boldsymbol{y}) - f^\star &\leq -\langle \boldsymbol{g}, \boldsymbol{x}^\star + \frac{1}{L}\boldsymbol{g} - \boldsymbol{y} \rangle + \frac{L}{2} \min_{\boldsymbol{x}' \in \mathcal{X}^\star} \left\| \boldsymbol{x}^\star + \frac{1}{L}\boldsymbol{g} - \boldsymbol{x}' \right\|_2^2 \leq -\langle \boldsymbol{g}, \boldsymbol{x}^\star + \frac{1}{L}\boldsymbol{g} - \boldsymbol{y} \rangle + \frac{\|\boldsymbol{g}\|_2^2}{2L} \\ &= \langle \boldsymbol{g}, \boldsymbol{y} - \boldsymbol{x}^\star \rangle - \frac{1}{2L}\|\boldsymbol{g}\|_2^2 \,. \end{aligned}$$

Now, assume that f is $\lambda_{\boldsymbol{x}}$-LSUC and set $\boldsymbol{g} \in \partial f(\boldsymbol{x})$. For any $\boldsymbol{x}^\star \in \operatorname{argmin}_{\boldsymbol{x}} \; f(\boldsymbol{x})$, using Cauchy-Schwarz's inequality, we have

$$f(\boldsymbol{x}) - f(\boldsymbol{x}^\star) \leq \langle \boldsymbol{g}, \boldsymbol{x} - \boldsymbol{x}^\star \rangle - \frac{1}{2\lambda_{\boldsymbol{x}}}\|\boldsymbol{g}\|_2^2 \leq \|\boldsymbol{g}\|_2 \|\boldsymbol{x} - \boldsymbol{x}^\star\|_2 - \frac{1}{2\lambda_{\boldsymbol{x}}}\|\boldsymbol{g}\|_2^2 \leq \frac{\lambda_{\boldsymbol{x}}}{2}\|\boldsymbol{x} - \boldsymbol{x}^\star\|_2^2 \,.$$

Given that this holds for all $\boldsymbol{x}^\star \in \mathcal{X}^\star$, it implies

$$f(\boldsymbol{x}) - f^\star \leq \frac{\lambda_{\boldsymbol{x}}}{2} \min_{\boldsymbol{x}' \in \mathcal{X}^\star} \|\boldsymbol{x} - \boldsymbol{x}'\|_2^2 \,. \qquad \square$$

C Proofs for the Stochastic Surrogate ψ

Theorem 3. *Let $h : \mathbb{R}^d \times \mathcal{S} \to \mathbb{R}_{\geq 0}$ be convex. Denote by $H(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\xi} \sim D}[h(\boldsymbol{x}, \boldsymbol{\xi})]$. Then, setting $\eta_t = \min\left(\frac{1}{\|\boldsymbol{g}_t\|^2}, \frac{\gamma}{h(\boldsymbol{x}_t, \boldsymbol{\xi}_t)}\right)$ in Algorithm 1, we have*

- *If $h(\cdot, \boldsymbol{\xi}_t)$ is L-self bounded, we have*

$$\frac{1}{T}\sum_{t=1}^{T} \min\left(\frac{1}{2L}, \gamma\right) \mathbb{E}[H(\boldsymbol{x}_t)] \leq \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{T} + 2\gamma H(\boldsymbol{x}^\star)$$

  and

$$\sum_{t=1}^{T} \min\left(\frac{1}{2L}, \gamma\right) \mathbb{E}[H(\boldsymbol{x}_t)] \leq \frac{1}{2}\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2 + \frac{1}{2}\sum_{t=1}^{T} \gamma^2 \mathbb{E}[\|\boldsymbol{g}_t\|^2] + \gamma \sum_{t=1}^{T} H(\boldsymbol{x}^\star) \,.$$

- *If $h(\cdot, \boldsymbol{\xi}_t)$ is G-Lipschitz, then we have*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[H(\boldsymbol{x}_t)] \leq \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{\gamma T} + 2H(\boldsymbol{x}^\star) + \frac{G\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|}{\sqrt{T}} + G\sqrt{2\gamma H(\boldsymbol{x}^\star)} \,.$$

- *If $h(\cdot, \boldsymbol{\xi})$ is L-self-bounded and $H(\boldsymbol{x})$ has μ-quadratic growth, then*

$$\mathbb{E}\left[\|\boldsymbol{x}_{T+1} - \boldsymbol{x}^\star\|^2\right] \leq \mathbb{E}\left[\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2\right] a^{T+1} + b\frac{1 - a^{T+1}}{1 - a} H(\boldsymbol{x}^\star),$$

  where $a = \frac{\mu}{2}\min\left(\frac{1}{2L}, \gamma\right)$ and $b = 2\gamma - \min\left(\frac{1}{2L}, \gamma\right)$.

Proof. For simplicity, denote by $h_t(\boldsymbol{x}) = h(\boldsymbol{x}, \xi_t)$.

From the Lemma 3, we have

$$\sum_{t=1}^{T} \eta_t \frac{1}{2} h_t^2(\boldsymbol{x}_t) \leq \frac{1}{2}\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2 + \sum_{t=1}^{T} \frac{\eta_t}{2}\left(\eta_t - \frac{1}{\|\boldsymbol{g}_t\|^2}\right)\|\boldsymbol{g}_t\|^2 h_t^2(\boldsymbol{x}_t) + \sum_{t=1}^{T} \eta_t h_t(\boldsymbol{x}_t) h_t(\boldsymbol{x}^\star) \,.$$

For the last term in the r.h.s., we have

$$\eta_t h_t(\boldsymbol{x}_t) h_t(\boldsymbol{x}^\star) = \min\left(\frac{1}{\|\boldsymbol{g}_t\|^2}, \frac{\gamma}{h_t(\boldsymbol{x}_t)}\right) h_t(\boldsymbol{x}_t) h_t(\boldsymbol{x}^\star) \leq \gamma h_t(\boldsymbol{x}^\star) \,.$$

Observe that if h_t is L-self bounded then $\|g_t\|^2 \leq 2L(h_t(\boldsymbol{x}) - \inf_{\boldsymbol{x}} h_t(\boldsymbol{x})) \leq 2L h_t(\boldsymbol{x})$. Therefore, we have

$$\min\left(\frac{1}{2L}, \gamma\right) h_t(\boldsymbol{x}_t) \leq \min\left(\frac{h_t(\boldsymbol{x}_t)}{\|\boldsymbol{g}_t\|^2}, \gamma\right) h_t(\boldsymbol{x}_t) = \eta_t h_t^2(\boldsymbol{x}_t) \,.$$

Now, since $\eta_t \leq 1/\|\boldsymbol{g}_t\|^2$ the second term on the r.h.s can be discarded because it's negative. Taking expectations, we have the first stated bound.

For the second result, bring on the l.h.s. the terms $\frac{\eta_t}{2} h_t^2(\boldsymbol{x}_t)$. Taking expectations, we have the stated bound.

For the third result, first of all observe that for any $a, b > 0$ we have

$$\min(a, b) = \frac{1}{\max(1/a, 1/b)} \geq \frac{1}{1/a + 1/b} \,.$$

Hence, we have

$$\eta_t h_t^2(\boldsymbol{x}_t) = \min\left(\frac{h_t(\boldsymbol{x}_t)}{\|\boldsymbol{g}_t\|^2}, \gamma\right) h_t(\boldsymbol{x}_t) \geq \gamma \frac{h_t^2(\boldsymbol{x}_t)}{h_t(\boldsymbol{x}_t) + \gamma\|\boldsymbol{g}_t\|^2} \geq \gamma \frac{h_t^2(\boldsymbol{x}_t)}{h_t(\boldsymbol{x}_t) + \gamma G^2} \,.$$

Now, observe that the function $B(x) = \frac{x^2}{x + \gamma G^2}$ is convex $x \geq 0$, because $B''(x) = \frac{2\gamma^2 G^4}{(\gamma G^2 + x)^3}$. So, summing over time and using Jensen's inequality, we have

$$B\left(\frac{1}{T}\sum_{t=1}^{T} h_t(\boldsymbol{x}_t)\right) \leq \frac{1}{T}\sum_{t=1}^{T} B(h_t(\boldsymbol{x}_t)) = \frac{1}{T}\sum_{t=1}^{T} \frac{h_t^2(\boldsymbol{x}_t)}{h_t(\boldsymbol{x}_t) + \gamma G^2} \leq \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{\gamma T} + \frac{2}{T}\sum_{t=1}^{T} h_t(\boldsymbol{x}^\star) \,.$$

Note that $B^{-1}(x) = \frac{x + \sqrt{x^2 + 4x\gamma G^2}}{2} \leq x + G\sqrt{x\gamma}$[5], that is an increasing concave function for $x \geq 0$. So, inverting B and taking expectation, we have

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[H(\boldsymbol{x}_t)] &\leq \mathbb{E}\left[B^{-1}\left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{\gamma T} + \frac{2}{T}\sum_{t=1}^{T} h_t(\boldsymbol{x}^\star)\right)\right] \\
&\leq \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{\gamma T} + 2H(\boldsymbol{x}^\star) + \frac{G\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|}{\sqrt{T}} + G\sqrt{2\gamma H(\boldsymbol{x}^\star)} \,.
\end{aligned}$$

For the smooth and quadratic growth case, we have

$$\begin{aligned}
\min\left(\frac{1}{2L}, \gamma\right) h_t(\boldsymbol{x}_t) &\leq \eta_t h_t^2(\boldsymbol{x}_t) \\
&\leq \|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 - \|\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star\|^2 + 2\eta_t h_t(\boldsymbol{x}_t) h_t(\boldsymbol{x}^\star) \\
&\leq \|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 - \|\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star\|^2 + 2\gamma h_t(\boldsymbol{x}^\star) \,.
\end{aligned}$$

Taking expectations and using the quadratic growth assumption on H, we have

$$\begin{aligned}
\frac{\mu}{2}\min\left(\frac{1}{2L}, \gamma\right)\mathbb{E}\left[\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2\right] &\leq \min\left(\frac{1}{2L}, \gamma\right)\mathbb{E}[H(\boldsymbol{x}_t) - H(\boldsymbol{x}^\star)] \\
&\leq \mathbb{E}\left[\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2\right] - \mathbb{E}\left[\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star\|^2\right] + \left(2\gamma - \min\left(\frac{1}{2L}, \gamma\right)\right) H(\boldsymbol{x}^\star) \,.
\end{aligned}$$

[5] Using the inequality $\sqrt{z + y} \leq \sqrt{z} + \sqrt{y} \quad \forall z, y \geq 0$.

Hence, we obtain

$$\mathbb{E}\left[\|\boldsymbol{x}_{t+1}-\boldsymbol{x}^\star\|^2\right] \leq (1-a)\mathbb{E}\left[\|\boldsymbol{x}_t-\boldsymbol{x}^\star\|^2\right]+b,$$

where $a=\frac{\mu}{2}\min\left(\frac{1}{2L},\gamma\right)$ and $b=\left(2\gamma-\min\left(\frac{1}{2L},\gamma\right)\right)H(\boldsymbol{x}^\star)$. Note we have $0\leq a\leq \mu/4L\leq 1$. From this inequality, it is immediate to obtain

$$\mathbb{E}\left[\|\boldsymbol{x}_{T+1}-\boldsymbol{x}^\star\|^2\right] \leq \mathbb{E}\left[\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|^2\right]a^{T+1}+b\frac{1-a^{T+1}}{1-a}\ . \qquad \square$$

D Additional Convergence Result for Generalized Polyak Stepsize

Corollary 1. *Let $f:\mathbb{R}^d\times\mathcal{S}$ be convex and L-self-bounded. Define $\boldsymbol{x}^\star=\operatorname{argmin}_{\boldsymbol{x}}\left(\mathbb{E}_{\boldsymbol{\xi}\sim D}[f(\boldsymbol{x},\boldsymbol{\xi})]:=F(\boldsymbol{x})\right)$. Let $h(\boldsymbol{x},\boldsymbol{\xi})=(f(\boldsymbol{x},\boldsymbol{\xi})-f(\boldsymbol{x}^\star,\boldsymbol{\xi}))_+$. Then, running Algorithm 1 with $\gamma=\infty$, we have*

$$\mathbb{E}[F(\bar{\boldsymbol{x}}_T)]-F(\boldsymbol{x}^\star)\leq\frac{2L\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|^2}{T}+\frac{\mathbb{E}[\sqrt{2L(F(\boldsymbol{x}^\star)-\mathbb{E}[\inf_{\boldsymbol{x}}f(\boldsymbol{x},\boldsymbol{\xi})])}\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|}{\sqrt{T}},$$

where $\bar{\boldsymbol{x}}_T=\frac{1}{T}\sum_{t=1}^T\boldsymbol{x}_t$ or $\bar{\boldsymbol{x}}_T=\operatorname{argmin}_{\boldsymbol{x}\in\{\boldsymbol{x}_1,\dots,\boldsymbol{x}_T\}}F(\boldsymbol{x})$.

Proof. Given the definition of h, we have that $H(\boldsymbol{x}^\star)=0$. Moreover, $h(\boldsymbol{x},\boldsymbol{\xi})\geq f(\boldsymbol{x},\boldsymbol{\xi})-f(\boldsymbol{x}^\star,\boldsymbol{\xi})$, hence $\mathbb{E}[H(\boldsymbol{x}_t)]\geq\mathbb{E}[F(\boldsymbol{x}_t)]-F(\boldsymbol{x}^\star)$ for any t.

We have that

$$\partial h(\boldsymbol{x},\boldsymbol{\xi})=\begin{cases}\partial f(\boldsymbol{x},\boldsymbol{\xi}), & \text{if } f(\boldsymbol{x},\boldsymbol{\xi})>f(\boldsymbol{x}^\star,\boldsymbol{\xi})\\ \{0\}, & \text{if } f(\boldsymbol{x},\boldsymbol{\xi})<f(\boldsymbol{x}^\star,\boldsymbol{\xi})\\ \{\alpha\boldsymbol{g}:\alpha\in[0,1],\boldsymbol{g}\in\partial f(\boldsymbol{x},\boldsymbol{\xi})\}, & \text{if } f(\boldsymbol{x},\boldsymbol{\xi})=f(\boldsymbol{x}^\star,\boldsymbol{\xi})\end{cases}.$$

Hence, for all $\boldsymbol{g}_t\in\partial h(\boldsymbol{x}_t,\boldsymbol{\xi}_t)$ we have

$$\begin{aligned}\|\boldsymbol{g}_t\|^2&\leq 2L(f(\boldsymbol{x},\boldsymbol{\xi}_t)-\inf_{\boldsymbol{x}}\ f(\boldsymbol{x},\boldsymbol{\xi}_t))\\&=2L(f(\boldsymbol{x},\boldsymbol{\xi}_t)-f(\boldsymbol{x}^\star,\boldsymbol{\xi}_t)+f(\boldsymbol{x}^\star,\boldsymbol{\xi}_t)-\inf_{\boldsymbol{x}}f(\boldsymbol{x},\boldsymbol{\xi}_t))\\&\leq 2L(h(\boldsymbol{x},\boldsymbol{\xi}_t)+f(\boldsymbol{x}^\star,\boldsymbol{\xi}_t)-\inf_{\boldsymbol{x}}f(\boldsymbol{x},\boldsymbol{\xi}_t))\ .\end{aligned}$$

So, using this inequality in Lemma 3 gives

$$\sum_{t=1}^T\frac{h_t^2(\boldsymbol{x}_t)}{4L(h_t(\boldsymbol{x}_t)+f_t(\boldsymbol{x}^\star)-f_t^\star)}\leq\sum_{t=1}^T\frac{h_t^2(\boldsymbol{x}_t)}{2\|\boldsymbol{g}_t\|^2}=\sum_{t=1}^T\eta_t\frac{1}{2}h_t^2(\boldsymbol{x}_t)\leq\frac{1}{2}\|\boldsymbol{x}_1-\boldsymbol{x}^\star\|^2\ . \tag{8}$$

Now, from Cauchy–Schwarz inequality, for any non-negative random variable Y and random variable X, we have $\mathbb{E}[X^2/Y]\geq(\mathbb{E}[X])^2/\mathbb{E}[Y]$. Denote by $f_t^\star=\inf_{\boldsymbol{x}}\ f(\boldsymbol{x},\boldsymbol{\xi}_t)$. Given that $f_t(\boldsymbol{x}^\star)-f_t^\star\geq 0$, if $f_t(\boldsymbol{x}^\star)-f_t^\star=0$ with probability 1, i.e., $F(\boldsymbol{x}^\star)-\mathbb{E}[\inf_{\boldsymbol{x}}f(\boldsymbol{x},\boldsymbol{\xi})]=0$, then the expectation of the l.h.s. of the previous inequality is $\mathbb{E}[h_t(\boldsymbol{x}_t)]$. Otherwise, if we assume $F(\boldsymbol{x}^\star)-\mathbb{E}[\inf_{\boldsymbol{x}}f(\boldsymbol{x},\boldsymbol{\xi})]>0$, we have

$$\mathbb{E}\left[\frac{h_t^2(\boldsymbol{x}_t)}{4L(h_t(\boldsymbol{x}_t)+f_t(\boldsymbol{x}^\star)-f_t^\star)}\right]\geq\frac{(\mathbb{E}[h_t(\boldsymbol{x}_t)])^2}{4L(\mathbb{E}[h_t(\boldsymbol{x}_t)]+F(\boldsymbol{x}^\star)-\mathbb{E}[\inf_{\boldsymbol{x}}f(\boldsymbol{x},\boldsymbol{\xi})])}\ .$$

Hence, in all cases we have the last expression is a lower bound to the l.h.s. of (8). We now can proceed as in the proof of the Lipschitz case in Theorem 3, to have the stated bound. $\square$

We now extend Theorem 3 to Hölder-self-bounded functions.

Definition 6. *We say that f is (L_ν,ν) Hölder-self-bounded if there exits $\nu\in[0,1]$ and L_ν such That*

$$\|\boldsymbol{g}\|^2\leq\left(1+\frac{1}{\nu}\right)^{\frac{2\nu}{1+\nu}}L_\nu^{\frac{2}{1+\nu}}(f(\boldsymbol{x})-f(\boldsymbol{x}^\star))^{\frac{2\nu}{1+\nu}},\ \forall\boldsymbol{g}\in\partial f(\boldsymbol{x})\ .$$

This definition is weaker than both Lipschitz and smoothness and it is easy to see that L-smooth functions satisfies this condition with $\nu = 1$ and $L_1 = L$.

The following theorem generalizes both the Lipschitz and the smooth case, recovering both bounds up to constant factors.

Theorem 7. *Let $h : \mathbb{R}^d \times \mathcal{S} \to \mathbb{R}_{\geq 0}$ be convex. Denote by $H(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\xi}\sim D}[h(\boldsymbol{x}, \boldsymbol{\xi})]$. Assume that $h(\cdot, \boldsymbol{\xi}_t)$ is (L_ν, ν)-Hölder-self bounded. Then, setting $\eta_t = \min\left(\frac{1}{\|\boldsymbol{g}_t\|^2}, \frac{\gamma}{h(\boldsymbol{x}_t, \boldsymbol{\xi}_t)}\right)$ in Algorithm 1, we have*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[H(\boldsymbol{x}_t)] \leq Q\left(\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{T\gamma} + 2H(\boldsymbol{x}^\star)\right),$$

where $Q(y) = 2y + L_\nu(2\gamma y)^{\frac{1+\nu}{2}}\left(1 + \frac{1}{\nu}\right)^\nu$.

Proof. For simplicity, denote by $h_t(\boldsymbol{x}) = h(\boldsymbol{x}, \xi_t)$.

From the Lemma 3, we have

$$\sum_{t=1}^{T}\eta_t\frac{1}{2}h_t^2(\boldsymbol{x}_t) \leq \frac{1}{2}\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2 + \sum_{t=1}^{T}\frac{\eta_t}{2}\left(\eta_t - \frac{1}{\|\boldsymbol{g}_t\|^2}\right)\|\boldsymbol{g}_t\|^2 h_t^2(\boldsymbol{x}_t) + \sum_{t=1}^{T}\eta_t h_t(\boldsymbol{x}_t)h_t(\boldsymbol{x}^\star)\,.$$

For the last term in the r.h.s., we have

$$\eta_t h_t(\boldsymbol{x}_t)h_t(\boldsymbol{x}^\star) = \min\left(\frac{1}{\|\boldsymbol{g}_t\|^2}, \frac{\gamma}{h_t(\boldsymbol{x}_t)}\right)h_t(\boldsymbol{x}_t)h_t(\boldsymbol{x}^\star) \leq \gamma h_t(\boldsymbol{x}^\star)\,.$$

Observe that if h_t is (L_ν, ν)-Hölder-self-bounded then

$$\|g_t\|^2 \leq \left(1 + \frac{1}{\nu}\right)^{\frac{2\nu}{1+\nu}} L_\nu^{\frac{2}{1+\nu}}(h_t(\boldsymbol{x}) - \inf_{\boldsymbol{x}} h_t(\boldsymbol{x}))^{\frac{2\nu}{1+\nu}} \leq K_\nu h_t(\boldsymbol{x})^{\frac{2\nu}{1+\nu}},$$

where $K_\nu = \left(1 + \frac{1}{\nu}\right)^{\frac{2\nu}{1+\nu}} L_\nu^{\frac{2}{1+\nu}}$. Therefore, we have

$$\min\left(\frac{h_t^{\frac{1-\nu}{1+\nu}}(\boldsymbol{x}_t)}{K_\nu}, \gamma\right)h_t(\boldsymbol{x}_t) \leq \min\left(\frac{h_t(\boldsymbol{x}_t)}{\|\boldsymbol{g}_t\|^2}, \gamma\right)h_t(\boldsymbol{x}_t) = \eta_t h_t^2(\boldsymbol{x}_t)\,.$$

As before, we lower bound the minimum with the convex function $B(x) = \frac{x^{\frac{2}{1+\nu}}}{x^{\frac{1-\nu}{1+\nu}} + \gamma K_\nu}$:

$$\min\left(\frac{h_t^{\frac{1-\nu}{1+\nu}}(\boldsymbol{x}_t)}{K_\nu}, \gamma\right)h_t(\boldsymbol{x}_t) \geq \frac{\gamma h_t^{\frac{2}{1+\nu}}(\boldsymbol{x}_t)}{h_t^{\frac{1-\nu}{1+\nu}}(\boldsymbol{x}_t) + \gamma K_\nu} = \gamma B(h_t(\boldsymbol{x}_t))\,.$$

As before, this allows us to use Jensen's inequality:

$$B\left(\frac{1}{T}\sum_{t=1}^{T}h_t(\boldsymbol{x}_t)\right) \leq \frac{1}{T}\sum_{t=1}^{T}B(h_t(\boldsymbol{x}_t)) \leq \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^\star\|^2}{\gamma T} + \frac{2}{T}\sum_{t=1}^{T}h_t(\boldsymbol{x}^\star)\,.$$

For simplicity of calculations, we now lower bound $B(x)$,

$$C(x) \coloneqq 0.5\min\left(x, \frac{x^{\frac{2}{1+\nu}}}{\gamma K_\nu}\right) = \frac{x^{\frac{2}{1+\nu}}}{2\max\{x^{\frac{1-\nu}{1+\nu}}, \gamma K_\nu\}} \leq B(x).$$

Note that $C(x)$ is invertible and its inverse is

$$\begin{aligned} C^{-1}(y) &= \begin{cases} 2y, & \text{if } y \geq (\gamma K_\nu)^{\frac{1+\nu}{1-\nu}} \\ (2\gamma K_\nu y)^{\frac{1+\nu}{2}}, & \text{if } y < (\gamma K_\nu)^{\frac{1+\nu}{1-\nu}} \end{cases} \\ &\leq 2y + (2\gamma K_\nu y)^{\frac{1+\nu}{2}} \\ &= 2y + L_\nu(2\gamma y)^{\frac{1+\nu}{2}}\left(1 + \frac{1}{\nu}\right)^\nu\,. \end{aligned}$$

Taking expectations and using Jensen's inequality gives the stated bound. □

E Proofs for Section 6

Lemma 4. *Let $f : \mathbb{R}^n \to \mathbb{R}^+$ where $f^\star = \inf_{\boldsymbol{x}} f(\boldsymbol{x})$. Then for any $c \geq 0$ the following are equivalent:*

- $f(\boldsymbol{x}) - f^\star < cf^\star$,
- $f(\boldsymbol{x}) - f^\star < \frac{c}{c+1} f(\boldsymbol{x})$.

Proof.

$$\begin{aligned} f(\boldsymbol{x}) - f^\star < cf^\star &\Leftrightarrow f(\boldsymbol{x}) < (c+1)f^\star \\ &\Leftrightarrow \frac{f(\boldsymbol{x})}{(c+1)} < f^\star \\ &\Leftrightarrow f(\boldsymbol{x}) - f^\star < \left(1 - \frac{1}{c+1}\right) f(\boldsymbol{x}) = \frac{c}{c+1} f(\boldsymbol{x}) \,. \end{aligned}$$

□

Proposition 1. *Suppose h is convex, strictly positive, L-self-bounded, and satisfies the quadratic growth condition $h(\boldsymbol{x}) - h^\star \geq \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{x}^\star\|^2$, where $\boldsymbol{x}^\star = \arg\min_{\boldsymbol{x}} h(\boldsymbol{x})$ is the only fixed point of T (6). Then for any point $\boldsymbol{x} \in S = \{\boldsymbol{y} : h(\boldsymbol{y}) - h^\star < h^\star \frac{\mu}{8L-\mu}\}$ we have*

$$\|T(\boldsymbol{x}) - \boldsymbol{x}^\star\| > \|\boldsymbol{x} - \boldsymbol{x}^\star\| \,.$$

Proof. Let $\boldsymbol{x}_t$ be in S then by definition of T we have

$$\begin{aligned} \frac{1}{2}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star\|^2 &= \frac{1}{2}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 - \eta_t \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}^\star \rangle + \frac{\eta_t^2}{2}\|\boldsymbol{g}_t\|^2 \\ &\geq \frac{1}{2}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 - \eta_t \|\boldsymbol{g}_t\| \|\boldsymbol{x}_t - \boldsymbol{x}^\star\| + \frac{\eta_t^2}{2}\|\boldsymbol{g}_t\|^2 \\ &= \frac{1}{2}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 - \frac{h(\boldsymbol{x}_t)}{\|\boldsymbol{g}_t\|}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\| + \frac{h(\boldsymbol{x}_t)^2}{2\|\boldsymbol{g}_t\|^2} \\ &= \frac{1}{2}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 + \frac{h(\boldsymbol{x}_t)}{\|\boldsymbol{g}_t\|}\left[\frac{h(\boldsymbol{x}_t)}{2\|\boldsymbol{g}_t\|} - \|\boldsymbol{x}_t - \boldsymbol{x}^\star\|\right]. \end{aligned}$$

If $h(\boldsymbol{x}_t) - h^\star < h^\star \frac{\mu}{8L-\mu}$ then we have by Lemma 4

$$h(\boldsymbol{x}_t) > \left(\frac{\mu/(8L-\mu) + 1}{\mu/(8L-\mu)}\right)(h(\boldsymbol{x}_t) - h^\star) = \frac{8L}{\mu}(h(\boldsymbol{x}_t) - h^\star) \,.$$

Consequently,

$$\frac{h(\boldsymbol{x}_t)}{2\|\boldsymbol{g}_t\|} > \frac{4L}{\mu}\frac{(h(\boldsymbol{x}_t) - h(\boldsymbol{x}^\star))}{\|\boldsymbol{g}_t\|} \geq 2L\frac{\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2}{\|\boldsymbol{g}_t\|} \geq \|\boldsymbol{x}_t - \boldsymbol{x}^\star\| \,.$$

Where the last inequality follows from h being self-bounded and convex,

$$\frac{1}{2L}\|\boldsymbol{g}_t\|^2 \leq h(\boldsymbol{x}_t) - h^\star \leq \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}^\star \rangle \leq \|\boldsymbol{g}_t\| \|\boldsymbol{x}_t - \boldsymbol{x}^\star\| \,.$$

□

Proposition 6. *Suppose h is convex, strictly positive, L-Lipschitz, and has a μ-sharp minimum $h(\boldsymbol{x}) - h^\star \geq \mu\|\boldsymbol{x} - \boldsymbol{x}^\star\|$, where $\boldsymbol{x}^\star = \arg\min_{\boldsymbol{x}} h(\boldsymbol{x})$ is the only fixed point of T (6). Then for any point $\boldsymbol{x} \in S = \{\boldsymbol{y} : \boldsymbol{y} \neq \boldsymbol{x}^\star, h(\boldsymbol{y}) - h^\star < h^\star \frac{\mu}{2L-\mu}\}$ we have*

$$\|T(\boldsymbol{x}) - \boldsymbol{x}^\star\| > \|\boldsymbol{x} - \boldsymbol{x}^\star\| \,.$$

Proof. Let $\boldsymbol{x}_t \in S$ then by following similar steps to Lemma 1 we have

$$\frac{1}{2}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star\|^2 \geq \frac{1}{2}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|^2 + \frac{h(\boldsymbol{x}_t)}{\|\boldsymbol{g}_t\|}\left[\frac{h(\boldsymbol{x}_t)}{2\|\boldsymbol{g}_t\|} - \|\boldsymbol{x}_t - \boldsymbol{x}^\star\|\right] \,.$$

By Lemma 4, we have

$$h(\boldsymbol{x}_t) - h^\star < h^\star \frac{\mu}{2L-\mu}$$
$$\Leftrightarrow h(\boldsymbol{x}_t) > \left(\frac{\mu/(2L-\mu)+1}{\mu/(2L-\mu)}\right)(h(\boldsymbol{x}_t) - h^\star) = \frac{2L}{\mu}(h(\boldsymbol{x}_t) - h^\star) .$$

Therefore, by sharpness and Lipschitz property of h, we have

$$\frac{h(\boldsymbol{x}_t)}{2\|\boldsymbol{g}_t\|} > \frac{L(h(\boldsymbol{x}_t) - h^\star)}{\mu\|\boldsymbol{g}_t\|} \geq \frac{L\|\boldsymbol{x} - \boldsymbol{x}^\star\|}{\|\boldsymbol{g}_t\|} \geq \|\boldsymbol{x}_t - \boldsymbol{x}^\star\| . \qquad \square$$

Proposition 2 (Cycling and failure to converge)**.** *There exists a strictly positive smooth and strongly convex function h, and initial point $\boldsymbol{x}_1$ such that iterates from update* (6) *cycle and satisfy the inequality $h(\frac{1}{t}\sum_{i=1}^t \boldsymbol{x}_i) - h^\star \geq \delta > 0$ for all t.*

Proof. The proof is constructive: consider $h : \mathbb{R} \to \mathbb{R}, h(x) = x^2 + 1$, so $h^\star = 1$. Observe that the update is

$$x_{t+1} = x_t - \frac{x_t^2 + 1}{2x_t} = \frac{x_t^2 - 1}{2x_t} .$$

Now, we want to choose x_1 so that we oscillate between 3 possible values.

Set $x_1 = \cot\theta$ where θ has to be determined. The update becomes

$$x_2 = \frac{x_1^2 - 1}{2x_1} = \frac{\cot^2\theta - 1}{2\cot\theta} = \cot(2\theta),$$

where in the last equality we used the identity for $\cot$. Hence, we have $x_t = \cot(2^t\theta)$. Given that we want to oscillate between 3 values, we want $x_{t+3} = x_t$, that is, $\cot(2^{t+3}\theta) = \cot(2^t\theta)$. We can achieve it if we select $\theta = \pi/7$. Indeed, we have

$$x_1 = \cot(\pi/7)$$
$$x_2 = \cot(2\pi/7)$$
$$x_3 = \cot(4\pi/7)$$
$$x_5 = \cot(8\pi/7) = \cot(\pi + \pi/7) = \cot(\pi/7) = x_1 .$$

Finally, one can verify numerically that $f(\frac{1}{t}\sum_{i=1}^t x_t) - f^\star > 0.77$. $\square$

Proposition 8. *There exists subregions within the unstable regions in Propositions 1 and 6 where the stepsizes are upper bounded.*

Proof. By convexity of h we have $h - h^\star \leq \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}^\star\rangle \leq \|\boldsymbol{g}_t\|\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|$, so $\|\boldsymbol{g}_t\| \geq \frac{h(\boldsymbol{x})-h^\star}{\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|}$. By assumption in Lemmas 1 and 6 the unstable region is $S = \{\boldsymbol{x} : h(\boldsymbol{x}) - h^\star < ch^\star\}$ where c depends on the properties of h. Therefore, for $\boldsymbol{x} \in S$ and denoting $\boldsymbol{g} \in \partial h(\boldsymbol{x})$ as any subgradient at $\boldsymbol{x}$, we have

$$\frac{h(\boldsymbol{x})}{\|\boldsymbol{g}\|^2} \leq \frac{h(\boldsymbol{x})\|\boldsymbol{x} - \boldsymbol{x}^\star\|^2}{(h(\boldsymbol{x}) - h^\star)^2} < \frac{(c+1)h^\star\|\boldsymbol{x} - \boldsymbol{x}^\star\|^2}{(h(\boldsymbol{x}) - h^\star)^2} .$$

If h has a sharp minimum, $h(\boldsymbol{x}) - h^\star \geq \mu\|\boldsymbol{x} - \boldsymbol{x}^\star\|$, then we have $\frac{h(\boldsymbol{x})}{\|\boldsymbol{g}_t\|^2} < \frac{c+1}{\mu^2}h^\star$. Therefore, the stepsizes are always bounded within S.

Now consider the subregion $S_k = \{\boldsymbol{x} : h(\boldsymbol{x}) - h^\star < \frac{c}{k}h^\star\}$ for some $k > 1$. Consider $\boldsymbol{x} \in S \setminus S_k$, that is $(1 + \frac{c}{k})h^* \leq h(\boldsymbol{x}) \leq (1 + c)h^*$. If h statisfies the quadratic growth condition $h(\boldsymbol{x}) - h^\star \geq \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{x}^\star\|^2$ then

$$\frac{h(\boldsymbol{x})}{\|\boldsymbol{g}\|^2} < \frac{(c+1)h^\star\|\boldsymbol{x} - \boldsymbol{x}^\star\|^2}{(h(\boldsymbol{x}) - h^\star)^2} \leq \frac{2(c+1)h^\star}{\mu(h(\boldsymbol{x}) - h^\star)} \leq \frac{2k(c+1)}{\mu c} . \tag{9}$$

Where the last inequality follows since $\boldsymbol{x} \notin S_k$. Therefore, stepsize is bounded within $S \setminus S_k$, and grows as we increase k. $\square$

Definition 7 (Lusin (N^{-1}) condition)**.** *Let $T : \mathbb{R}^d \to \mathbb{R}^k$. We define T^{-1} over a set $S \subseteq \mathbb{R}^k$ as*

$$T^{-1}(S) = \{\boldsymbol{x} : T(\boldsymbol{x}) \in S\}.$$

We say that T satisfies (N^{-1}) condition if for every set E of measure zero we have that $T^{-1}(E)$ also has measure zero.

Lemma 5. *Let $T : \mathbb{R}^n \to \mathbb{R}^n$ with a unique fixed point $\boldsymbol{x}^\star$. If $\boldsymbol{x}^\star$ is unstable, that is, there exists δ such that $\boldsymbol{x} \neq \boldsymbol{x}^\star$ and $\|\boldsymbol{x} - \boldsymbol{x}^\star\| \leq \delta$, then $\|T(\boldsymbol{x}) - \boldsymbol{x}^\star\| > \|\boldsymbol{x} - \boldsymbol{x}^\star\|$. Define T^{-1} over a set $S \subseteq \mathbb{R}^n$ as $T^{-1}(S) = \{\boldsymbol{x} : T(\boldsymbol{x}) \in S\}$. If $T^{-k}(\{\boldsymbol{x}^\star\})$ is of measure zero for any k, then*

$$P\left(\lim_{t\to\infty} \boldsymbol{x}_t = \boldsymbol{x}^\star\right) = 0\,.$$

In other words, the set of initializations that can converge to the fixed point has measure zero.

Proof. We divide $\mathbb{R}^n$ into two sets, $S = \bigcup_{k=1}^{\infty} T^{-k}(\{\boldsymbol{x}^\star\})$, and its compliment S^c. S represents the points that can exactly reach the unique minimizer $\boldsymbol{x}^\star$. If $T^k(\{\boldsymbol{x}^\star\})$ is a null set for every k then so is S since the countable union of null sets is a null set.

Now we show that for all initializations in $\boldsymbol{x}_1 \in S^c$, $\boldsymbol{x}_1$ cannot converge to $\boldsymbol{x}^\star$. Suppose the contrary, $\lim_{t\to\infty} \boldsymbol{x}_t = \boldsymbol{x}^\star$. Let $B = \{\boldsymbol{y} : \boldsymbol{y} \neq \boldsymbol{x}^\star, \|\boldsymbol{x} - \boldsymbol{x}^\star\| \leq \delta\}$. Since $\boldsymbol{x}_t \to \boldsymbol{x}^\star$ there exists a step n where $\{\boldsymbol{x}_t\}_{t\geq n} \subseteq B$. Similarly, there exists $n' \geq n$ where $\|\boldsymbol{x}_t - \boldsymbol{x}^\star\| \leq \|\boldsymbol{x}_n - \boldsymbol{x}^\star\|$ for all $t \geq n'$. This is a contradiction as we have that $\|\boldsymbol{x}_n - x^\star\| < \|\boldsymbol{x}_{n+1} - \boldsymbol{x}^\star\| < \cdots < \|\boldsymbol{x}_{n'} - \boldsymbol{x}^\star\|$. Therefore, the initializations that allow for $\boldsymbol{x}_t \to \boldsymbol{x}^\star$ coincide exactly with the null set S. □