OPT2025: 17th Annual Workshop on Optimization for Machine Learning

Revisiting the Geometrically Decaying Step Size: Linear Convergence for Smooth or Non-Smooth Functions

Jihun Kim JIHUN.KIM@BERKELEY.EDU
University of California, Berkeley, CA 94720, United States

Abstract

We revisit the geometrically decaying step size given a positive inverse condition number, under which a locally Lipschitz function shows linear convergence. The positivity does not require the function to satisfy convexity, weak convexity, quasar convexity, or sharpness, but instead amounts to a property strictly weaker than the assumptions used in existing works (*e.g.*, weak convexity + sharpness). We propose a clean and simple subgradient descent algorithm that requires minimal knowledge of problem constants, applicable to either smooth or non-smooth functions.

1. Introduction

The goal of optimization is to find a true minimizer x^* in the set of minimizers $\mathcal{X}^* \subseteq \mathbb{R}^n$ of a function $f(x)$, whether smooth or non-smooth. Goffin [8] and Shor [19] independently studied the geometrically decaying step sizes in the subgradient descent algorithm, presenting various scenarios of the decaying rate, showing that linear convergence is achieved under convexity and sharpness-type assumptions. Polyak [18] also proved linear convergence under similar conditions; however, the Polyak step size requires the information of the minimum function value $f(x^*)$. Recent studies have proposed alternative methods to achieve linear convergence. For example, [20] suggested averaging past consecutive iterates to obtain a robust estimate, while [12] proposed the descending-stairs step size scheme, which reduces the step size only occasionally. However, these approaches generally require complicated constants or involve redundant subroutines.

In this work, we revisit the classical geometrically decaying step size and propose a simple subgradient descent algorithm that achieves linear convergence for *a locally Lipschitz—smooth or non-smooth—function* $f(x)$ under a general condition and requires minimal knowledge of constants. A similar line of work [7] established linear convergence under weak convexity and sharpness; however, these conditions are somewhat restrictive. In contrast, we show that the "positive inverse condition number" stated in the next section turns out to be a property strictly weaker than the aforementioned weak convexity and sharpness. Meanwhile, the work [11] examines quasar-convex functions—a generalization of star-convexity [16] under which every line segment from a minimizer x^* to any other point preserves convexity. We also establish that the positive inverse condition number holds for all quasar-convex functions and extends to functions beyond this class.

Our focus is on this generalized *positive inverse condition number* assumption, which holds not only for arbitrary norm functions but also for a class of nonconvex functions that may not satisfy sharpness, weak convexity, or quasar convexity. From a geometric perspective, a positive inverse condition number implies that the negative subgradient direction forms an acute angle with the direction toward a minimizer. This property can be interpreted in terms of cosine similarity, which is experimentally investigated in [9]. Such a property frequently arises in applications involving

nonconvex functions without spurious local minima, including phase retrieval [4, 17], matrix recovery [2, 5], image alignment [21], and dynamical systems [10], even under adversarial environments [13]. As one application, in dynamic programming (DP), it has been shown that if each DP subproblem has no spurious local minima, then one can obtain a global solution to the entire (one-shot) optimization problem [1, 14], in which case the one-shot problem can be solved by repeatedly leveraging the property that will be discussed throughout the paper.

Notation: We denote by $\|\cdot\|$ the ℓ_2-norm of a vector, and by $\langle\cdot,\cdot\rangle$ the inner product of two vectors. For a point x and a set X, $\mathrm{dist}(x;X)$ denotes $\inf_{y\in X}\|x-y\|$. For two sets X and Y, $X\setminus Y$ denotes the set of elements in X not in Y.

2. Assumption for Linear Convergence

In this section, we begin by presenting the following assumption for the geometrically decaying step size scheme to achieve linear convergence. In this regard, we define the set of interest $S=\{x\in\mathbb{R}^n : \mathrm{dist}(x;\mathcal{X}^*)\le \mathrm{dist}(x_0;\mathcal{X}^*)\}$, given an initial point $x_0\in\mathbb{R}^n$.

Assumption 1 (Positive Inverse Condition Number) *Consider a locally Lipschitz function $f(x)$ over the set S. We define the inverse condition number $\bar{\mu}$ as*

$$\bar{\mu}=\inf_{x^*\in\mathcal{X}^*,x\in S\setminus\mathcal{X}^*}\ \inf_{u\in\partial^\circ f(x)\setminus\{0\}}\frac{\langle u,x-x^*\rangle}{\|u\|\|x-x^*\|},\tag{1}$$

where $\partial^\circ f(x)$ is a Clarke differential [6], whose elements are called generalized subgradients. We assume that $\bar{\mu}>0$.

Remark 1 *Note that $\partial^\circ f(x)$ is nonempty, closed, and convex for all $x\in S$, since $f(x)$ is assumed to be locally Lipschitz. We also note that $\bar{\mu}\le 1$ holds due to the Cauchy-Schwarz inequality.*

Under Assumption 1, $0\in\partial^\circ f(x)$ is an equivalent condition to optimality, and thus serves as a stopping criterion for the generalized subgradient descent algorithm.

Lemma 2 *Suppose that $\langle u,x-x^*\rangle>0$ holds for all $x^*\in\mathcal{X}^*, x\in S\setminus\mathcal{X}^*$, and any $u\in\partial^\circ f(x)$. Then, a point $s\in S$ is a minimizer of f if and only if $0\in\partial^\circ f(s)$.*

Proof If $s\in S$ is a minimizer, it is evident that $0\in\partial^\circ f(s)$ (see Proposition 2.3.2 in [6]). To prove the converse, suppose that $0\in\partial^\circ f(s)$ holds but s is not a minimizer of f. Then, due to the precondition of the lemma, we arrive at $\langle 0,s-x^*\rangle>0$ for all $x^*\in\mathcal{X}^*$, which yields the contradiction. Thus, s is a minimizer of f. ■

Remark 3 *Lemma 2 implies that if $x\in S\setminus\mathcal{X}^*$ (i.e., not a minimizer), then $0\notin\partial^\circ f(x)$. Hence, in the definition of the inverse condition number in (1), one can replace $u\in\partial^\circ f(x)\setminus\{0\}$ with $u\in\partial^\circ f(x)$, while keeping the quantity well-defined.*

In the following two lemmas, we now provide some natural sufficient conditions for $\bar{\mu}>0$ to hold.

Lemma 4 *Suppose $f(x)$ is locally Lipschitz over the set S and ρ-weakly convex with respect to x^*; i.e., there exists $M > 0,\ \rho \geq 0$ such that*

$$\|u\| \leq M \quad \textit{and} \quad f(x^*) \geq f(x) + \langle u, x^* - x\rangle - \frac{\rho}{2}\|x^* - x\|^2, \quad \forall u \in \partial^\circ f(x)$$

for all $x^* \in \mathcal{X}^*$ and $x \in S \setminus \mathcal{X}^*$. Suppose also that the function has sharpness; i.e., there exists $m > 0$ such that $f(x) - f(x^*) \geq m\|x - x^*\|$ for all $x^* \in \mathcal{X}^*$ and $x \in S \setminus \mathcal{X}^*$. When dist$(x_0; \mathcal{X}^*) \leq \frac{m}{\rho}$, we have $\bar{\mu} \geq \frac{m}{2M} > 0$.

Proof Using the definition of the inverse condition number, we have

$$\bar{\mu} \geq \frac{f(x) - f(x^*) - \frac{\rho}{2}\|x^* - x\|^2}{\|u\|\|x - x^*\|} \geq \frac{1}{M}\left(\frac{f(x) - f(x^*)}{\|x - x^*\|} - \frac{\rho}{2}\|x^* - x\|\right) \geq \frac{1}{M}(m - \frac{\rho}{2} \cdot \frac{m}{\rho}) = \frac{m}{2M}$$

for all $x^* \in \mathcal{X}^*, x \in S \setminus \mathcal{X}^*$. This completes the proof. ■

Lemma 5 *Suppose $f(x)$ is locally Lipschitz over the set S and γ-quasar convex with respect to x^*; i.e., there exists $M > 0$, $\gamma \in (0, 1]$ such that*

$$\|u\| \leq M \quad \textit{and} \quad f(x^*) \geq f(x) + \frac{1}{\gamma}\langle u, x^* - x\rangle, \quad \forall u \in \partial^\circ f(x)$$

for all $x^* \in \mathcal{X}^*$ and $x \in S \setminus \mathcal{X}^*$. Suppose also that the function has sharpness; i.e., there exists $m > 0$ such that $f(x) - f(x^*) \geq m\|x - x^*\|$ for all $x^* \in \mathcal{X}^*$ and $x \in S \setminus \mathcal{X}^*$. Then, we have $\bar{\mu} \geq \frac{\gamma m}{M} > 0$.

Proof One can observe that $\bar{\mu} \geq \frac{\gamma(f(x) - f(x^*))}{\|u\|\|x - x^*\|} \geq \frac{\gamma m}{M}$ for all $x^* \in \mathcal{X}^*$ and $x \in S \setminus \mathcal{X}^*$. ■

Lemmas 4 and 5 respectively provide sufficient conditions for Assumption 1 to hold, but it requires the function to have sharpness and be weakly-convex or quasar-convex. To illustrate that our Assumption 1 can be regarded as a more general notion, we present examples where it holds even though the function lacks convexity, weak convexity, quasar convexity, or sharpness. For a pictorial illustration, we present graphs of these examples in Figure 1.

Example 1 (Assumption 1 does not imply convexity) *Consider a function $f(x) = |3x| + \sin(|x|)$, which attains its unique minimum at 0. Although the function is nonconvex, we have $\bar{\mu} > 0$ since*

$$\langle f'(x), x\rangle = |x| \cdot (3 + \cos(x)), \quad x \neq 0,$$

which is exactly $|x||f'(x)|$. This implies that the inverse condition number is 1.

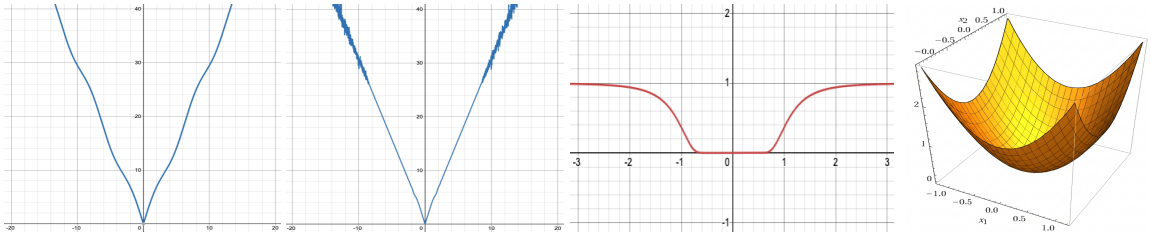
Example 2 (Assumption 1 does not imply weak convexity) *Consider a function $f(x) = \int_0^{|x|}(3 + \cos(s^3))ds$, which attains its minimum at 0. Since the function is twice-differentiable, the weak convexity is equivalent to finding $\rho \geq 0$ such that $f''(x) \geq -\rho$ for all x. For $x > 0$, we have*

$$f''(x) = \frac{d}{dx}(3 + \cos(x^3)) = -3x^2 \sin(x^3),$$

which is unbounded below. However, we have $\bar{\mu} > 0$ since

$$\langle f'(x), x\rangle = |x| \cdot (3 + \cos(x^3)), \quad x \neq 0,$$

which is exactly $|x||f'(x)|$. This implies that the inverse condition number is 1.



(a) Example 1 (b) Example 2 (c) Example 3 (d) Example 4

Figure 1: Examples of functions with positive inverse condition numbers but without (a) convexity, (b) weak convexity, (c) quasar convexity, or (d) sharpness

Example 3 (Assumption 1 does not imply quasar convexity) *Consider a function* $f(x) = e^{-1/x^4}$ *for* $x \neq 0$*, and* $f(0) = 0$*, which attains its minimum at* 0*. It fails to satisfy quasar convexity since there does not exist* $\gamma > 0$ *such that*

$$\gamma \leq \frac{\langle f'(x), x\rangle}{f(x)} = \frac{\frac{4}{x^5}e^{-1/x^4} \cdot x}{e^{-1/x^4}} = \frac{4}{x^4}, \quad \forall x \neq 0,$$

since the right term converges to 0 *as* $x \to \pm\infty$*. However, we have* $\bar{\mu} > 0$ *since* $\langle f'(x), x\rangle = \frac{4}{x^4}e^{-1/x^4} = |x||f'(x)|$*, implying that the inverse condition number is 1.*

Example 4 (Assumption 1 does not imply sharpness) *Consider a function* $f(x_1, x_2) = x_1^2 + x_2^2$*, which attains its unique minimum at* 0*. This function does not have sharpness in the sense that for any* $m > 0$*, one can always find* (x_1, x_2) *such that*

$$f(x_1, x_2) = x_1^2 + x_2^2 < m\sqrt{x_1^2 + x_2^2},$$

especially when $0 < x_1^2 + x_2^2 < m^2$*. However, we have* $\langle \nabla f(x_1, x_2), (x_1, x_2)\rangle = 2\sqrt{x_1^2 + x_2^2} \cdot \sqrt{x_1^2 + x_2^2}$*, which implies that the inverse condition number is 1.*

For non-smooth functions, existing works [7, 11] relied on weak convexity or quasar convexity, together with sharpness. On the other hand, for smooth functions, the quadratic growth condition $f(x) - f(x^*) \geq m\|x - x^*\|^2$ derived from strong convexity is a standard assumption to ensure linear convergence [3, 15]. However, $x_1^2 + x_2^2$ has quadratic growth with $m = 1$, yet not sharpness (see Example 4), indicating that sharpness is an overly restrictive requirement applicable only to functions that are non-smooth at the minimum. In the next section, we will show that a *positive inverse condition number* suffices to guarantee linear convergence for either smooth or non-smooth functions.

3. Geometrically Decaying Step Size Design

In this section, we provide the generalized subgradient descent algorithm and provide our convergence analysis under the *positive inverse condition number*. We note that [7] and [11] introduced a linear convergent algorithm only under the assumptions of Lemma 4 and Lemma 5, respectively, whereas our analysis shows that linear convergence still holds under a strictly weaker condition.

Algorithm 1 Generalized Subgradient Descent Algorithm

Input: Inverse Condition number $\bar{\mu}$. Initial point x_0.

Let $0 < r \leq \min\{\bar{\mu}, \frac{1}{\sqrt{2}}\}$.
for time $t = 0, 1, \ldots$ **do**
  if $0 \in \partial^\circ f(x_t)$ **then**
    Break. x_t is a minimizer.
  else
    Pick $g_t \in \partial^\circ f(x_t)$.
    Set the step size $\eta_t = \frac{r\left(1-r^2\right)^{t/2}\|x_0 - x^*\|}{\|g_t\|}$.
    Let $x_{t+1} = x_t - \eta_t g_t$.
  end if
end for

Theorem 6 *Consider a minimizer $x^* \in \mathcal{X}^*$. Under Assumption 1, Algorithm 1 achieves*

$$\|x_t - x^*\| \leq \left(1 - r^2\right)^{t/2} \|x_0 - x^*\|. \tag{2}$$

Proof We prove this by induction. The base case $t = 0$ is trivial. For the induction step, suppose that

$$\|x_s - x^*\| \leq \left(1 - r^2\right)^{s/2} \|x_0 - x^*\|$$

holds. Let α be the constant that satisfies

$$\alpha\|x_s - x^*\| = \left(1 - r^2\right)^{s/2} \|x_0 - x^*\|. \tag{3}$$

By the induction hypothesis, we clearly have $\alpha \geq 1$. Then, from the definition of the step size, we arrive at

$$\eta_s\|g_s\| = r\alpha\|x_s - x^*\|. \tag{4}$$

Then, we have

$$\begin{aligned}
\|x_{s+1} - x^*\|^2 &= \|x_s - \eta_s g_s - x^*\|^2 = \|x_s - x^*\|^2 - 2\eta_s\langle g_s, x_s - x^*\rangle + \eta_s^2\|g_s\|^2 \\
&\underset{\text{(a)}}{\leq} \|x_s - x^*\|^2 - 2\eta_s\|g_s\|\|x_s - x^*\| \cdot r + \eta_s^2\|g_s\|^2 \\
&\underset{\text{(b)}}{=} \|x_s - x^*\|^2 - 2r\alpha\|x_s - x^*\| \cdot \|x_s - x^*\|r + (r\alpha\|x_s - x^*\|)^2 \\
&= \left(1 - 2r^2\alpha + r^2\alpha^2\right)\|x_s - x^*\|^2 \\
&\underset{\text{(c)}}{\leq} (1 - r^2)\alpha^2\|x_s - x^*\|^2 \\
&\underset{\text{(d)}}{=} (1 - r^2)^{s+1}\|x_0 - x^*\|^2
\end{aligned}$$

where (a) is due to the definition of the inverse condition number and $\bar{\mu} \geq r$, (b) is due to (4), and (d) is from (3). The inequality (c) is derived from the fact that $r \leq \frac{1}{\sqrt{2}}$ and $\alpha \geq 1$ yields

$$(1 - 2r^2)\alpha^2 + 2r^2\alpha - 1 = (\alpha - 1)\left((1 - 2r^2)\alpha + 1\right) \geq 0.$$

Thus, we achieve $\|x_{s+1} - x^*\| \leq (1 - r^2)^{\frac{s+1}{2}} \|x_0 - x^*\|$, which completes the proof. ■

Algorithm 2 Generalized Subgradient Descent Algorithm without the exact estimation of $\|x_0 - x^*\|$

Input: Inverse Condition number $\bar{\mu}$. Initial point x_0. A constant R that satisfies

$$\beta\|x_0 - x^*\| \leq R \leq (1-\beta)\|x_0 - x^*\| \tag{5}$$

for a pre-determined $0 < \beta \leq 0.5$.

Let $0 < r \leq \bar{\mu}$.
for time $t = 0, 1, \ldots$ **do**
  if $0 \in \partial^\circ f(x_t)$ **then**
    Break. x_t is a minimizer.
  else
    Pick $g_t \in \partial^\circ f(x_t)$.
    Set the step size $\eta_t = \frac{r\left(1+(\beta^2-2\beta)r^2\right)^{t/2}R}{\|g_t\|}$.
    Let $x_{t+1} = x_t - \eta_t g_t$.
  end if
end for

In Algorithm 1, one can choose r from the range $(0, \min\{\bar{\mu}, \frac{1}{\sqrt{2}}\}]$. A natural choice to maximize the linear convergence rate $\sqrt{1-r^2}$ would be to replace $\bar{\mu}$ with $\frac{m}{2M}$ or $\frac{\gamma m}{M}$, given that we can estimate m and M (see Lemmas 4 and 5). Then, the only term that we are unaware of in the step size η_t is $\|x_0 - x^*\|$; *i.e.* the distance from the initial point to any true minimizer $x^* \in \mathcal{X}^*$. In the next theorem, we propose that the unknown $\|x_0 - x^*\|$ need not be exactly estimated, but can instead be replaced by a sufficiently small constant and still achieve linear convergence.

Theorem 7 *Consider a minimizer $x^* \in \mathcal{X}^*$. Under Assumption 1, Algorithm 2 achieves*

$$\|x_t - x^*\| \leq \left(1 + (\beta^2 - 2\beta)r^2\right)^{t/2} \|x_0 - x^*\|. \tag{6}$$

Proof We use a similar induction technique. The proof details can be found in Appendix A. ■

Remark 8 *Since $0 < \beta \leq 0.5$ in Algorithm 2, we have $(\beta^2 - 2\beta)r^2 < 0$, which again ensures linear convergence of the algorithm. The additional constants required to determine the step size are β and R. If β is chosen sufficiently small, one can increase confidence to obtain a correspondingly small value of R that satisfies (5). However, one should account for a natural trade-off: setting β too small may lead to a worse linear convergence ratio.*

4. Conclusion

In this work, we propose a geometically decaying step size scheme that achieves linear convergence to a minimizer under the assumption of a positive inverse condition number. We identify that existing assumptions—such as weak convexity or quasar-convexity, combined with sharpness—are common in applications yet impose strictly stronger requirements than a positive inverse condition number. We first develop a generalized subgradient descent algorithm that requires knowledge of the distance from the initial point to a minimizer, and subsequently show that it can be replaced by a user-defined constant, potentially altering the convergence rate while preserving linear convergence. Our results provide a foundation for a more general assumption and convergence analysis to incorporate the possibility of extending the framework to a broader class of functions in a wider range of optimization areas, including *stochastic, robust, nonconvex, and distributed* optimization.

References

[1] Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. *Operations Research*, 72(5):1906–1927, 2024.

[2] Yingjie Bi, Haixiang Zhang, and Javad Lavaei. Local and global linear convergence of general low-rank matrix recovery problems. In *AAAI Conference on Artificial Intelligence*, volume 36, 2022.

[3] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[4] Emmanuel J. Candès, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

[5] Laming Chen and Yuantao Gu. The convergence guarantees of a non-convex approach for sparse recovery. *IEEE Transactions on Signal Processing*, 62(15):3754–3767, 2014.

[6] Frank H. Clarke. *Optimization and Nonsmooth Analysis*. Society for Industrial and Applied Mathematics, 1990.

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

[8] J. L. Goffin. On convergence rates of subgradient optimization methods. *Mathematical Programming*, 13:329–347, 1977.

[9] Charles Guille-Escuret, Hiroki Naganuma, Kilian Fatras, and Ioannis Mitliagkas. No wrong turns: The simple geometry of neural networks optimization paths. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 16751–16772. PMLR, 2024.

[10] Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient descent learns linear dynamical systems. *Journal of Machine Learning Research*, 19(29):1–44, 2018.

[11] Oliver Hinder, Aaron Sidford, and Nimit Sohoni. Near-optimal methods for minimizing star-convex functions and beyond. In *Conference on Learning Theory*, volume 125, pages 1894–1938. PMLR, 2020.

[12] Patrick R. Johnstone and Pierre Moulin. Faster subgradient methods for functions with hölderian growth. *Mathematical Programming*, 180:417–450, 2020.

[13] Jihun Kim and Javad Lavaei. Prevailing against adversarial noncentral disturbances: Exact recovery of linear systems with the ℓ_1-norm estimator. In *American Control Conference (ACC)*, pages 1161–1168. IEEE, 2025.

[14] Jihun Kim, Yuhao Ding, Yingjie Bi, and Javad Lavaei. The landscape of deterministic and stochastic optimal control problems: One-shot optimization versus dynamic programming. *IEEE Transactions on Automatic Control*, 69(12):8587–8602, 2024.

[15] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer New York, 2004.

[16] Yurii Nesterov and B. T. Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

[17] Praneeth Netrapalli, Prateek Jain, and Sujay Sanghavi. Phase retrieval using alternating minimization. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

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

[19] Naum Zuselevich Shor. *Minimization Methods for Non-Differentiable Functions*. Springer Berlin, Heidelberg, 1985.

[20] Tianbao Yang and Qihang Lin. Rsg: Beating subgradient method without smoothness and strong convexity. *Journal of Machine Learning Research*, 19(6):1–33, 2018.

[21] Yiqing Zhang, Xinming Huang, and Ziming Zhang. Prise: Demystifying deep lucas-kanade with strongly star-convex constraints for multimodel image alignment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Appendix A. Proof for Theorem 7

We leverage a similar induction technique from Theorem 6. The base case $t = 0$ is trivial. For the induction step, suppose that

$$\|x_s - x^*\| \leq \left(1 + (\beta^2 - 2\beta)r^2\right)^{s/2} \|x_0 - x^*\|$$

holds. Let α be a constant satisfying

$$\alpha\|x_s - x^*\| = \left(1 + (\beta^2 - 2\beta)r^2\right)^{s/2} \|x_0 - x^*\|, \tag{7}$$

where $\alpha \geq 1$ naturally holds. Let $q := \frac{R}{\|x_0 - x^*\|}$. Then, we have

$$\eta_s\|g_s\| = r\alpha\|x_s - x^*\| \frac{R}{\|x_0 - x^*\|} = rq\alpha\|x_s - x^*\|. \tag{8}$$

Similar to the proof in Theorem 6, we now arrive at

$$\begin{aligned}
\|x_{s+1} - x^*\|^2 &\leq \|x_s - x^*\|^2 - 2\eta_s\|g_s\|\|x_s - x^*\| \cdot r + \eta_s^2\|g_s\|^2 \\
&\underset{\text{(a)}}{=} \|x_s - x^*\|^2 - 2rq\alpha\|x_s - x^*\| \cdot \|x_s - x^*\|r + (rq\alpha\|x_s - x^*\|)^2 \\
&= \left(1 - 2r^2q\alpha + r^2q^2\alpha^2\right) \|x_s - x^*\|^2 \\
&\underset{\text{(b)}}{\leq} (1 + (\beta^2 - 2\beta)r^2)\alpha^2\|x_s - x^*\|^2 \\
&\underset{\text{(c)}}{=} (1 + (\beta^2 - 2\beta)r^2)^{s+1}\|x_0 - x^*\|^2,
\end{aligned}$$

where (a) and (c) comes from (8) and (7), respectively. For (b), observe that $\beta \leq q \leq 1 - \beta$. It follows that

$$\beta^2 - 2\beta - q^2 \geq \beta^2 - 2\beta - (1 - \beta)^2 = -1, \tag{9a}$$

$$\beta^2 - 2\beta - q(q - 2) \geq \beta^2 - 2\beta - \beta(\beta - 2) = 0, \tag{9b}$$

which leads to

$$\begin{aligned}
&\left[1 + \left(\beta^2 - 2\beta - q^2\right) r^2\right]\alpha^2 + 2r^2q\alpha - 1 \\
&\qquad = (\alpha - 1)\left(\left[1 + \left(\beta^2 - 2\beta - q^2\right) r^2\right] \alpha + 1\right) + (\beta^2 - 2\beta - q^2 + 2q)r^2\alpha \\
&\qquad \geq (\alpha - 1)((1 - r^2)\alpha + 1) + 0 \geq 0,
\end{aligned}$$

where the first inequality follows from $\alpha \geq 1$ and (9), and the second from $\alpha \geq 1$ and $r = \bar{\mu} \leq 1$. Thus, the induction step is complete.