Accelerated Distance-adaptive Methods for Hölder Smooth and Convex Optimization

Anonymous Author(s)
Affiliation
Address
email

Abstract

This paper introduces new parameter-free first-order methods for convex optimization problems in which the objective function exhibits Hölder smoothness. Inspired by the recently proposed distance-over-gradient (DOG) technique, we propose an accelerated distance-adaptive method which achieves optimal anytime convergence rates for Hölder smooth problems without requiring prior knowledge of smoothness parameters or explicit parameter tuning. Importantly, our parameter-free approach removes the necessity of specifying target accuracy in advance, addressing a significant limitation found in the universal fast gradient methods (Nesterov, 2015). We further present a parameter-free accelerated method that eliminates the need for line-search procedures and extend it to convex stochastic optimization. Preliminary experimental results highlight the effectiveness of our approach in convex nonsmooth problems and its advantages over existing parameter-free or accelerated methods.

1 Introduction

In this paper, we consider the following composite function

$$\min_{x\in\mathbb{R}^d} \psi(x) := f(x) + g(x). \tag{1}$$

$f\colon \mathbb{R}^d \to \mathbb{R}$ is a continuous convex function, and $g\colon \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ is a simple proper lower semi-continuous (lsc) convex function of which the proximal operator is easy to evaluate. First-order algorithms—(sub)gradient descent and their accelerated variants—are the work-horses for solving (1), particularly in large-scale machine-learning and AI applications. Their empirical success, however, hinges on judiciously chosen stepsizes (learning rates); manual tuning quickly becomes the dominant practical bottleneck.

In standard analysis, the stepsize policy is often designed based on the smoothness level of the objective function. Specifically, subgradient methods for nonsmooth problems typically employ diminishing stepsizes, whereas gradient descent methods for smooth optimization often utilize a stepsize inversely proportional to the gradient Lipschitz parameter. In a seminar work, Nesterov [24] considered convex and Hölder smooth problem where $f(x)$ satisfies the following condition:

$$\|\nabla f(x) - \nabla f(y)\|_* \le L_\nu \|x-y\|^\nu, \forall x, y \in \mathbb{R}^d. \tag{2}$$

where $\nu \in [0,1]$ continuously interpolates between the nonsmooth ($\nu = 0$) and smooth ($\nu = 1$) settings. Nesterov introduced accelerated gradient methods capable of universally achieving optimal convergence rates without prior knowledge of the smoothness parameters of the objective. Notably,

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

the universal fast gradient method exhibits a convergence rate of

$$\psi(x^k) - \psi(x^*) \leq \mathcal{O}\left(\frac{L_\nu^{\frac{2}{1+\nu}} D_0^2}{\epsilon^{\frac{1-\nu}{1+\nu}} k^{\frac{1+3\nu}{1+\nu}}} + \epsilon\right). \tag{3}$$

where $D_0 := \|x^0 - x^*\|$ denotes the initial distance to the minimizer. An appealing property of this method is its independence from explicit knowledge of the smoothness level ν and the parameter L_ν.

Despite these attractive features, recent studies have highlighted significant limitations of the universal gradient methods. One notable issue is that the universal gradient method relies on a linesearch subroutine to adapt to the unknown smoothness level, which makes it challenging to extend to stochastic optimization. More critically, the performance of universal methods is sensitive to the predetermined target accuracy level ϵ. As ϵ must be set in advance, the method does not guarantee an anytime convergence property, which is crucial for practical implementations where iterations can be halted at an arbitrary stage. In addition, as pointed out by [26], an optimally set ϵ should depend on D_0, which is unfortunately unknown beforehand. Setting the value of ϵ without prior knowledge of the underlying problem structure often results in suboptimal trade-offs between the two error terms in (3). This turns out to be a critical problem in nonsmooth optimization or online learning [20, 2, 5, 4, 36]. Consequently, it remains an open question whether a genuinely parameter-free method can be obtained to achieve optimal convergence rates across different Hölder smoothness regimes.

In this paper, we demonstrate that near-optimal parameter-free convergence rates can be achieved for convex Hölder smooth optimization, up to logarithmic factors. Specifically, instead of fixing the target ϵ, we set variable target levels that are dynamically changing based on the optimization trajectory. As mentioned earlier, an optimal level shall depend on the distance to the minimizer and is difficult to compute. Motivated by the Distance over Gradients (DOG) style stepsize [2, 13], we approximate $\|x^0 - x^*\|$ by the maximum distance to the iterates and use this knowledge to choose stepsize in the accelerated gradient method. Leveraging the technique of distance adaptation, we are able to obtain a parameter-free and anytime convergence rate of

$$\mathcal{O}\left(\frac{D_0^{1+\nu} \log^2 e\frac{D_0}{\bar{r}}}{k^{\frac{1+3\nu}{2}}}\right), \tag{4}$$

without requiring any predefined optimality level, knowing the smoothness level ν or the Hölder smooth parameter.

In addition, we propose a line-search-free accelerated method that achieves optimal convergence rates for both Hölder smooth and stochastic optimization problems. To eliminate the need for line search, we adopt a bounded domain assumption, as originally introduced by Rodomanov et al. [30]. Different from their approach, which explicitly requires the domain diameter D to set stepsizes, our method exploits distance adaptation to approximate D. This can be particularly appealing as computing the diameter of a general convex set can be computationally intractable. Moreover, by estimating D through the observed distance from the initial point, our method naturally adopts more adaptive stepsizes in large domains. Experimental results support our theoretical insights and demonstrate the practical effectiveness of our approach.

1.1 Related works

The increasing computational cost associated with hyperparameter tuning has driven significant research interest in developing adaptive or parameter-free algorithms. The online learning community has extensively studied parameter-free optimization, particularly focusing on achieving nearly-optimal regret bounds without prior knowledge of domain boundedness or the distance to the minimizer. For example, see [20, 21, 25, 4, 36]. A recent breakthrough has been made by Carmon and Hinder [2], which moves beyond regret analysis and focuses directly on stochastic optimization. This algorithm appears to be conceptually simpler and motivates a few more practical SGD algorithms, such as [13, 5, 23]. A notable related study to our work is [23], which applied the distance adaptation technique to Nesterov's dual averaging method. They have offered convergence guarantees across various problem classes, including nonsmooth, smooth, (L_0, L_1)-smooth functions, and many others. However, the convergence rate for the Hölder smooth problem remains suboptimal.

The concept of universal gradient methods adapting to Hölder smoothness was pioneered by Nesterov [24]. Subsequent works extended this approach to nonconvex optimization [10] and stochastic settings [9]. It has been demonstrated that normalized gradient stepsizes [31] can automatically adapt to Hölder smoothness without requiring line search, as shown in [11, 27]. Recently, Rodomanov et al. [30] proposed a line-search-free universally optimal method that is robust to stochastic noise in gradient estimations. However, this method requires the domain to be bounded and the diameter to be known. In parallel to universal methods, bundle-type methods [17, 1] have emerged as an effective approach for nonsmooth optimization, enabling self-adaptation to Hölder smoothness [15]. However, these methods often involve solving a complex cut-constrained subproblem and lack straightforward extensions to stochastic settings. The Polyak stepsize method [28] also exhibits self-adaptation to smoothness [12] and Lipschitz parameters [22]. It can be seen as a special case of the bundle-level method [7]. While imposing Nesterov's acceleration technique [6] in the Polyak stepsize can universally achieve the optimal rates for Hölder smooth problems, it typically requires knowledge of the optimal value.

Finally, it is important to emphasize that the parameter-free algorithms discussed in this paper differ from adaptive gradient methods [8], which have been substantially studied in the literature [14, 29, 32]. While adaptive gradient methods primarily focus on adjusting to Lipschitz constants or constructing a preconditioner to approximate the Hessian inverse [37, 34, 18], parameter-free algorithms in our context do not rely on such adaptation mechanisms.

2 Preliminaries

Let $\mathbb{R}^d$ denote the d-dimensional Euclidean space. Let $\|\cdot\| = \sqrt{\langle\cdot,\cdot\rangle}$ be the norm associated with inner product $\langle\cdot,\cdot\rangle$. Its dual norm is defined by $\|s\|_* = \max_{\|x\|=1}\langle s, x\rangle$, where $s \in \mathbb{R}^d$. For a convex function $f : \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$, we use $\nabla f(x)$ to denote a (sub)gradient of $f(\cdot)$ at the point x. We use $\mathcal{B}_\delta(x) = \{y \in \mathbb{R}^d : \|y - x\| \le \delta\}$ to denote the closed ball of radius δ centered at x. Let $h : \mathbb{R}^d \to \mathbb{R}$ be a kernel function such that it is continuously differentiable and strongly convex with respect to $\|\cdot\|$, the Bregman distance generated by $h(\cdot)$ is given by $B(x, y) := h(y)-h(x)-\langle\nabla h(x), y-x\rangle$. Without loss of generality, we assume $B(x, y) \ge \frac{1}{2}\|x - y\|^2$. In this paper, we choose $B(x, y) = \frac{1}{2}\|x - y\|^2$. We define I_ν as $(\frac{1}{1-\nu})^{\frac{1+\nu}{2}}$ when $\nu < 1$ and 1 when $\nu = 1$ to to simplify the expressions.

A convex function f is said to be *locally Hölder smooth* at z with radius r if there exists a mapping $M_\nu : \mathbb{R}^d \times (0, +\infty) \to (0, +\infty)$ such that, for any $z \in \mathbb{R}^d$, for any $x, y \in \mathcal{B}_r(z)$ $(r > 0)$, we have

$$\|\nabla f(x) - \nabla f(y)\|_* \le M_\nu(z, r)\,\|x - y\|^\nu. \tag{5}$$

Nesterov [24] considered the *global Hölder smooth* functions (2) where the smoothness mapping $M_\nu(z, r)$ is reduced to a constant L_ν. However, we will show that by incorporating both the line search method and the distance adaptation, we can guarantee the boundedness of the iterates. Consequently, the complexity relies on the local Hölder smoothness, which is defined by

$$\hat{M}_\nu := M_\nu(x^*, 3D_0) < +\infty.$$

3 The accelerated distance-adaptive method

Motivation Before describing the main algorithm, we first shed light on the intuition of Nesterov's universal gradient methods [24]. From the definition of global Hölder smoothness (2), we have the error bound condition

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

This bound can be translated into an inexact variant of the usual Lipschitz smooth condition:

$$f(y) \le f(x) + \langle\nabla f(x), y - x\rangle + \frac{\gamma(L_\nu, \delta)}{2}\|x - y\|^2 + \frac{\delta}{2}. \tag{6}$$

where $\gamma(L, \delta) := (\frac{1-\nu}{1+\nu}\frac{1}{\delta})^{\frac{1-\nu}{1+\nu}} L^{\frac{2}{1+\nu}}$, and $\delta \in (0, \infty)$ is the trade-off parameter. Since the Hölder exponent may be unknown a priori, Nesterov proposed to use pre-specify $\delta = \epsilon$, where ϵ is the target

accuracy, and then perform line search over $\gamma(L_\nu, \delta)$ to satisfy the inexact Lipschitz smoothness condition (6).

As described earlier, the potential issue of this approach is in setting the optimal δ. Note that this δ trade-off two terms: 1) the convergence rate of optimizing $\psi(x)$ as a Lipschitz smooth function, and 2) the approximation error of using Lipschitz smooth function. However, choosing the optimal δ necessitates the knowledge of the distance to the minimizer, and can not be done easily when the domain is unconstrained. Moreover, even if the domain is bounded with $D = \max_{x,y\in \mathrm{dom}\, g} \|x - y\| < \infty$, this value can poorly overestimate $x - x^*$.

To address the limitation in the existing universal fast gradient method, we present the accelerated distance-adaptive method (AGDA) in Algorithm 1. Our algorithm can be viewed as a variant of the accelerated regularized dual averaging method [33, 24, 35] which involves the triplets $\{x^k, v^k, y^k\}$. The main difference from the prior work is the new stepsize and linesearch procedure to adapt to the distance to the minimizer.

Specifically, leveraging the distance adaptation technique [23, 13], we approximate D_0 by a sequence of values:

$$\bar{r}_k = \max\{\bar{r}_{k-1}, r_k\}, \text{ where } r_k = \|x^0 - v^k\|, \ k \geq 0, \tag{7}$$

and then set the averaging sequence a_k based on the distance estimation:

$$a_{k+1} = A_{k+1} - A_k, \text{ where } A_{k+1} := \left(\sum_{i=0}^{k} \bar{r}_i^{\frac{1}{2}}\right)^2, \ k = 1, 2, \ldots \tag{8}$$

To deal with the unknown Hölder smooth parameter, we invoke a line search procedure to find the appropriate value of β_{k+1}, which satisfies

$$f(y^{k+1}) \leq f(x^{k+1}) + \langle \nabla f(x^{k+1}), y^{k+1} - x^{k+1} \rangle + \frac{\beta_{k+1}}{64\tau_k^2 A_{k+1}} \|y^{k+1} - x^{k+1}\|^2 + \frac{\tau_k \eta_k}{2}, \tag{9}$$

where η_k measures the inexactness in Lipschitz smooth approximation and τ_k is introduced by Nesterov's momentum ($\tau_k = 1$ in the non-accelerated method). As mentioned earlier, η_k is set to a fixed δ in the universal (fast) gradient method. Different from the earlier approach, we simply take $\eta_k = \frac{\beta_{k+1}\bar{r}_k^2 - \beta_k \bar{r}_{k-1}^2}{8a_{k+1}}$, which dynamically adjusts during the optimization process.

Algorithm 1 Accelerated Gradient Method with Distance Adaption(AGDA)

Input: x^0 and $\bar{r}$;

Initialize $A_0 = 0$, $\bar{r}_{-1} = \bar{r}_0 = \bar{r}$ and β_0 be a small constant, like 10^{-3}.
Set initial solution: $z^0 = y^0 = x^0$
for $k = 0, 1, ..., K-1$ **do**
    Set r_k and $\bar{r}_k$ according to (7);
    Update a_{k+1} and A_{k+1} by (8), and set $\tau_k = \frac{a_{k+1}}{A_{k+1}}$;
    Set $x^{k+1} = \tau_k v^k + (1 - \tau_k) y^k$;
    Apply the line search to find β_{k+1} such that $l_k(\beta_{k+1}) \geq 0$;
    Compute $v^{k+1} = \mathrm{argmin}_y \left\{ \frac{\beta_{k+1}}{2}\|x^0 - y\|^2 + \sum_{i=1}^{k+1} a_i(\langle \nabla f(x^i), y - x^i \rangle + g(y)) \right\}$;
    Set $y^{k+1} = \tau_k v^{k+1} + (1 - \tau_k) y^k$;
end for

Output: $z^K = \arg\min_{y \in \{y^0, y^1, \ldots y^K\}} \psi(y)$

Next, we further provide the details of the line search procedure. We expect to find a suitable β_{k+1} that satisfies the inequality (9). Note that in the inequality (9) y^{k+1} is dependent on β_{k+1}, we can describe the searching process of β_{k+1} by formulating it as finding a root of a continuous function. We first define some notations.

$$\begin{aligned} l_k(\beta) := & - f(y^{k+1}(\beta)) + f(x^{k+1}) + \langle \nabla f(x^{k+1}), y^{k+1}(\beta) - x^{k+1} \rangle \\ & + \frac{\beta}{64\tau_k^2 A_{k+1}} \|y^{k+1}(\beta) - x^{k+1}\|^2 + \frac{\beta \bar{r}_k^2 - \beta_k \bar{r}_{k-1}^2}{16 A_{k+1}}, \end{aligned} \tag{10}$$

where $y^{k+1}(\beta) = \tau_k v_{k+1}(\beta) + (1 - \tau_k) y^k$ is the trial point, and $v_{k+1}(\beta) := \mathrm{argmin}_x \sum_{i=1}^{k+1} a_i(\langle \nabla f(x^i), x \rangle + g(x)) + \frac{\beta}{2}\|x - x^0\|^2$. Our line search consists of two stages,

each involving an iterative procedure. We assume the first stage and the second stage can be terminated in i'_k-th and i^*_k-th iterations, respectively. In the first stage, we find the smallest value $i \in \{0, 1, 2, \dots, \}$ such that $l_k(2^{i-1}\beta_k)$ is nonnegative and set $i'_k = i$. Consequently, we will have two situations:

1. $i'_k = 1$, i.e., $l_k(\beta_k) \geq 0$, then we complete the line search;
2. $i'_k > 1$, then we perform a binary search to find an approximate root of $l_k(\cdot) = 0$ in the interval $[2^{i'_k-2}\beta_k, 2^{i'_k-1}\beta_k]$. The search is terminated when the interval length is no more than the tolerance level $\frac{\beta_0}{2k^2}$. We return the right endpoint of the final interval as $\beta_{k+1} = \beta_{k+1,i^*}$.

The goal of the line search strategy in the k-th iteration is to find the root of $l_k(\cdot)$. Next, we justify the correctness of the line search and efficient implementation.

Proposition 1. *Suppose $f(\cdot)$ is locally Hölder smooth* (5) *in $\mathcal{B}_{3D_0}(x^*)$. In Algorithm 1, for any $k \geq 0$, at least one of the following two conditions must be met:*

1. *$l_k(\beta_k) \geq 0$;*
2. *there exists $\beta^*_{k+1} > \beta_k$ such that $l_k(\beta^*_{k+1}) = 0$ and $\forall \beta > \beta^*_{k+1}, l_k(\beta) > 0$.*

Consequently, we have $\beta_{k+1} \leq \mathcal{O}(k^{\frac{3-3\nu}{2}})$. Moreover, the total number of line searches ($\sum_{k=0}^{K-1}(i'_k + i^*_k)$) required by Algorithm 1 is $\mathcal{O}(K \log K)$.

Remark 1. *Proposition 1 implies that our method requires an additional $\mathcal{O}(\log K)$ function evaluations compared to other algorithms [24]. However, it is worth emphasizing that our line search procedure only requires access to function values, whereas the line search in the universal fast gradient method involves both gradient and function value evaluations.*

Next, we provide an important upper bound on the convergence rate.

Proposition 2. *Suppose $f(\cdot)$ is locally Hölder smooth* (5) *in $\mathcal{B}_{3D_0}(x^*)$. For any $k > 0$, it holds that*

$$\psi(y^k) - \psi(x^*) \leq \frac{\beta_k(D_0^2 - D_k^2)}{2A_k} + \frac{\beta_k \bar{r}_k^2}{8A_k}. \tag{11}$$

One key insight of the distance adaptation is that the inequality (11) implies the boundedness of r_k. To avoid the first step of Algorithm 1 from searching too far and breaking the boundedness of r_k, we should adopt a conservative distance estimation such that $\bar{r} \leq 4D_0$. The smaller $\bar{r}$ is, the more likely it is that $\bar{r} \leq 4D_0$ holds.

Theorem 1. *Suppose $f(\cdot)$ is locally Hölder smooth* (5) *in $\mathcal{B}_{3D_0}(x^*)$ and $\bar{r} \leq 4D_0$ holds. Then it holds that $\|v^k - x^0\| \leq 4D_0$ and $\|v^k - x^*\| \leq 3D_0$, for all $k \geq 0$.*

Since both x^i and y^i are convex combination of v^i, we immediately have that that all the generated points $\{x^i, y^i\}_{i\geq 0}$ are in $\mathcal{B}_{3D_0}(x^*)$.

Next, we further refine the upper bound in (11). Since $D_0^2 - D_k^2 \leq 2D_0 r_k$, $r_k \leq \bar{r}_k$ and $r_k \leq 4D_0$, the upper bound in Proposition 2 can be relaxed to

$$\psi(y^k) - \psi(x^*) \leq \frac{3\beta_k \bar{r}_k D_0}{2A_k}.$$

It remains to control the growth of $\frac{\bar{r}_k}{A_k}$. To this end, we invoke a useful logarithmic bound [13, 19] as follows.

Lemma 1. *Let $(d_i)_{i=0}^{\infty}$ be a positive nondecreasing sequence. Then for any $K \geq 1$,*

$$\min_{1\leq k\leq K} \frac{d_k}{\sum_{i=0}^{k-1} d_i} \leq \frac{\left(\frac{d_K}{d_0}\right)^{\frac{1}{K}} \log \frac{e d_K}{d_0}}{K}. \tag{12}$$

To apply the above result, we simply take $d_k = \sqrt{\bar{r}_k}$. It shows there is always some $k < K$ where the error $\frac{\bar{r}_k}{A_k}$ is bounded by $\mathcal{O}\big(\frac{\log^2 f(\bar{r}_T/\bar{r})}{K^2}\big)$.

Next, we bring all the pieces together. As pointed out earlier, the line search ensures that β_{k+1} is order up to $\mathcal{O}(k^{\frac{3-3\nu}{2}})$. Together with the bound over distance-adaptive term $\frac{\bar{r}_k}{A_k}$, we arrive at our final convergence rate in the following theorem.

Theorem 2. *Suppose all the assumptions of Theorem 1 hold. Then, Algorithm 1 exhibits a convergence rate that*

$$\psi(z^K) - \psi(x^*) \in \mathcal{O}\left(\frac{\hat{M}_\nu D_0^{1+\nu} \log^2 e\frac{D_0}{\bar{r}}}{K^{\frac{1+3\nu}{2}}}\right), \tag{13}$$

where $z^K = \arg\min_{y\in\{y^0,y^1,\ldots y^K\}} \psi(y)$.

Remark 2. *To achieve an ϵ-optimality gap, our method attains a near-optimal complexity bound of $\tilde{\mathcal{O}}\big(D_0^{\frac{2(1+\nu)}{1+3\nu}} \epsilon^{-\frac{2}{1+3\nu}}\big)$, where the $\tilde{\mathcal{O}}$ notation hides logarithmic factors arising from line search and distance adaptation, such as $\mathcal{O}(\log\frac{1}{\epsilon})$ and $\mathcal{O}(\log^2\frac{D_0}{\bar{r}})$.*

Remark 3. *Theorems 1 and 2 require the initial guess $\bar{r}$ to lie within a reasonably large neighborhood, specifically $\bar{r} \le 4D_0$. This condition is a key assumption underlying distance-adaptive methods [13]. For theoretical purposes, we provide an automatic initialization strategy for $\bar{r}$ in certain special cases (see Appendix). Empirically, we observe that the performance of the algorithm is largely insensitive to the specific choice of $\bar{r}$.*

4 Stochastic optimization

In this section, we focus on stochastic optimization of Hölder smooth functions, wherein problem (1), $f(x)$ exhibits the expectation form:

$$f(x) = \mathbb{E}_\xi[f(x,\xi)],$$

where ξ is a random sample following from specific distribution. Due to the difficulty in exactly computing the gradient $\nabla f(x)$, it is challenging to perform line search. To bypass this issue, we present a new line-search-free and accelerated distance-adaptive method in Algorithm 2. At the cost of removing linesearch, we require an additional boundedness assumption.

Assumption 1 (Boundedness of domain)**.** *The set* $\operatorname{dom} g$ *is bounded, namely,* $D = \sup_{x,y\in\operatorname{dom} g} \|x - y\| < +\infty$. *We denote* $\tilde{M}_\nu = M_\nu(x^*, D)$ *for simplicity.*

Let us use $\nabla\tilde{f}(x,\xi)$ to represent a stochastic gradient, we further assume the stochastic gradient has a bounded variance: $\sigma^2 := \sup_{x\in\mathbb{R}^d} \mathbb{E}_\xi[\|\nabla\tilde{f}(x,\xi) - \nabla f(x)\|_*^2] < +\infty$. For the sake of notation, we denote $\tilde{\nabla} f(x^k) = \nabla\tilde{f}(x^k,\xi_k^x)$ and $\tilde{\nabla} f(y^k) = \nabla\tilde{f}(y^k,\xi_k^y)$ to present the stochastic gradient in the k-th iteration, where ξ_k^x and ξ_k^y are two i.i.d. samples.

Algorithm 2 AGDA Line Search Free Modification (AGDA LSFM)

Input: x^0, $\bar{r}$;
1: Initialize $A_0 = 0$, $\beta_0 = 0$, $\bar{r}_0 = \bar{r}$;
2: Set initial solution: $z^0 = \hat{x}^0 = y^0 = x^0$;
3: **for** $k = 0, 1, ..., K-1$ **do**
4: Solve $v^k = \arg\min_x \sum_{i=0}^k a_i[f(x^i) + \langle\nabla f(x^i), x - x^i\rangle + g(x)] + \frac{\beta_k}{2}\|x^0 - x\|^2$;
5: Set $d_k = \|x^0 - \hat{x}^k\|$;
6: Update $\bar{r}_k$ and A_{k+1} by (14) and (8)
7: Set $a_{k+1} = A_{k+1} - A_k$, $\tau_k = \frac{a_{k+1}}{A_{k+1}}$;
8: Set $x^{k+1} = \tau_k v^k + (1-\tau_k)y^k$;
9: Compute $\hat{x}^{k+1} = \arg\min_y\{a_{k+1}[\langle\tilde{\nabla} f(x^{k+1}), y - x^{k+1}\rangle + g(y)] + \frac{\beta_k}{2}\|v^k - y\|^2\}$;
10: Set $y^{k+1} = \tau_k\hat{x}^{k+1} + (1-\tau_k)y^k$;
11: Set $\eta_k = \frac{\beta_{k+1}\bar{r}_k^2 - \beta_k\bar{r}_k^2}{8a_{k+1}}$;
12: Solve (15) to obtain the solution β_{k+1};
13: **end for**

Algorithm 2 is equipped with the following rules:

$$\bar{r}_k = \max\{\bar{r}_{k-1}, r_k, d_k\}, k > 0, \tag{14}$$

where $d_k = \|\hat{x}^k - x^0\|$.

In stochastic settings, traditional line search methods cannot be used as they introduce bias. Therefore, it is necessary to develop an approach that does not rely on line search. Rather than performing line search to find the descent direction, Rodomanov et al. [30] proposes a nonlinear balance equation. The core idea is to bound the error term $f(y^{k+1}) - f(x^{k+1}) - \langle \nabla f(x^{k+1}), y^{k+1} - x^{k+1} \rangle - \frac{H_k}{2}\|y^{k+1} - x^{k+1}\|^2$ by constructing a balance equation incorporating D. We demonstrate that the term used to bound the error is effectively equivalent to line search, allowing us to use $\bar{r}_k$ to approximate D, which implies that D is not essential. Subsequently, we will explain how to formulate the balance equation.

As we mentioned in Section 3, line search strategy tries to find β_{k+1} that $l_k(\beta_{k+1}) = 0$. The difficulty is that the y^{k+1} is depend on β_{k+1} and thus $l_k(\beta_{k+1}) = 0$ can not be solved by the closed-form solution. The motivation for applying the balance equation is to decouple the updating rule of y^{k+1} from the β_{k+1}. Once the y^{k+1} has been updated, $l_k(\beta_{k+1}) = 0$ will degenerates to the following balance equation

$$\frac{\beta_{k+1} - \beta_k}{2A_{k+1}}\bar{r}_k^2 = [\langle \tilde{\nabla} f(y^{k+1}) - \tilde{\nabla} f(x^{k+1}), y^{k+1} - x^{k+1} \rangle - \frac{\beta_{k+1}}{64\tau_k^2 A_{k+1}}\|y^{k+1} - x^{k+1}\|^2]_+, \tag{15}$$

where $[\cdot]_+ = \max(0, \cdot)$. We use $\tilde{\nabla} f(y^{k+1}), y^{k+1} - x^{k+1} \rangle$ to replace the $-f(y^{k+1}) + f(x^{k+1})$ since we can not obtain the function value.

Since we decouple y^{k+1} from β_{k+1}, equation (15) has a simple form that is easy to solve. Moreover, it has a unique closed-form solution given by

$$\beta_{k+1} = \beta_k + \frac{[64\tau_k^2 A_{k+1}\langle \tilde{\nabla} f(y^{k+1}) - \tilde{\nabla} f(x^{k+1}), y^{k+1} - x^{k+1} \rangle - \beta_k \|y^{k+1} - x^{k+1}\|^2]_+}{32\tau_k^2 \bar{r}_k^2 + \|y^{k+1} - x^{k+1}\|^2}. \tag{16}$$

We leave the details about conducting the closed-form solution in the appendix.

We next conduct the convergence analysis of Algorithm 2. In order to use the unbiasedness of the inexact oracle, we adopt the balance equation to update the β_{k+1}. Moreover, we use $\bar{r}_k$ is a natural underestimation of D and Lemma 1 resure that the cost of underestimation can be reduced to $\mathcal{O}(\log \frac{D}{\bar{r}})$. We leave the proof in the appendix.

Theorem 3. *Suppose Assumption 1 holds. Algorithm 2 exhibits a convergence rate that*

$$\mathbb{E}[\psi(z^{k^*}) - \psi(x^*)] \in \mathcal{O}\left(\frac{\tilde{M}_\nu D^{1+\nu}}{K^{\frac{1+3\nu}{2}}} + \frac{\sigma D}{\sqrt{K}}\right). \tag{17}$$

where $k^* = \arg\min_{1 \le k \le K}\{\frac{\bar{r}_k}{A_k}\}$.

5 Experiments

We evaluate the performance of our proposed method on a diverse set of convex optimization problems. The goal is to assess its efficiency and robustness across different application scenarios. Additional implementation details and extended results are provided in the appendix.

5.1 Deterministic setting

Softmax The first problem is optimizing the softmax function:

$$\min_{x \in \mathbb{R}^d} \mu \log\left[\sum_{i=1}^n \exp\left(\frac{\langle a_i, x \rangle - b_i}{\mu}\right)\right], \tag{18}$$

where $a_i \in \mathbb{R}^d$, $b_i \in \mathbb{R}$ and μ is a given positive factor. This function can be viewed as a smooth approximation of the maximization function.

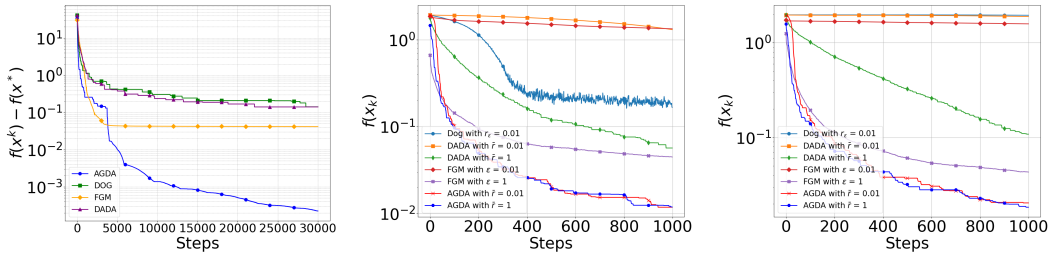

Figure 1: Performance of the compared algorithms. Left: softmax problem. Middle: Matrix game problem of size $(n, m) = (896, 128)$. Right: Matrix game of size and $(n, m) = (448, 64)$.

In our problem setting, we aim to establish a simple baseline solution to easily evaluate the performance of each method. To achieve this, we first generate the i.i.d vectors $\{\hat{a}_i\}$ with their components uniformly distributed within the interval $[-1, 1]$. Similarly, we generate vectors b_i from the same distribution. With these generated vectors, we define a function $\hat{f}(x)$ by the definition 18. Next, let $a_i = \hat{a}_i - \nabla\hat{f}(x)$ and redefine the function $f(x)$ by the definition 18 with a_i and b_i. Under this configuration, the point $x = 0_d$ serves as the optimal solution of $f(x)$.

We employ various methods for comparison, specifically considering DOG [13], DADA [23], and the universal fast gradient method (FGM) [24] as benchmarks. For Dog, we set $r_\epsilon = 0.01$. Both DADA and AGDA are configured with $\bar{r} = 0.01$, while for FGM, we set $\epsilon = 0.01$. We set $n = 1000$, $d = 2000$ and $\mu = 0.005$ as the parameters of the problem. The results of our method are illustrated in the left part of Figure 1. As expected from complexity analysis, FGM, being an accelerated method, outperforms the non-accelerated baselines DOG and DADA. Notably, our proposed algorithm achieves the fastest convergence among all tested methods, which empirically confirms the advantage of our adaptive stepsize selection.

Matrix game The second problem we experimented with is the matrix game problem [24]. We denote Δ_d as the standard simplex with d-dimension, for some $d > 0$. Specifically, consider a payoff matrix $A \in \mathbb{R}^{n\times m}$, where two agents engage in a game by adopting mixed strategies $x \in \Delta_n$ and $y \in \Delta_n$ respectively to play a game without knowledge of each other's strategy. The gain of the first agent is given by $\langle x, Ay\rangle$, which corresponds to the loss of the second agent. The Nash equilibrium of this game can be found by solving the root of the function:

$$\min_{x\in\Delta_n} \max_{y\in\Delta_m} \langle x, Ay\rangle. \tag{19}$$

We generate the payoff matrix A such that each entry is independently and uniformly distributed within the interval $[-1, 1]$. This problem is nonsmooth with Hölder smoothness parameter $\nu = 0$, making it a suitable test case for evaluating the robustness of optimization algorithms under minimal smoothness assumptions. We evaluate all methods on two problem sizes: $(n, m) = (896, 128)$ and $(n, m) = (448, 64)$. The performance of our method, along with the baselines, is shown in the right panels of Figure 1. The results demonstrate that our algorithm remains highly effective even in challenging nonsmooth settings, outperforming the alternatives in both cases.

5.2 Stochastic setting

Least-squares For the stochastic setting, we first consider the following problem:

$$\min_{x\in\mathbb{R}^d} f(x) = \frac{1}{2}\|Ax - b\|_2^2 \quad \text{s.t. } \|x\|_2 \le r. \tag{20}$$

We set the constraint radius $r = 10$ and conduct experiments using real-world datasets from LIBSVM[1]. For the first test, we use the diabetes dataset to examine robustness. For both USFGM and our AGDA-LSFM algorithm, we vary the initialization hyperparameter-D for USFGM and $\bar{r}$ for AGDA-LSFM—to evaluate sensitivity to stepsize-related inputs. As shown in the left panel of Figure 2,

[1] `https://www.csie.ntu.edu.tw/cjlin/libsvm/`

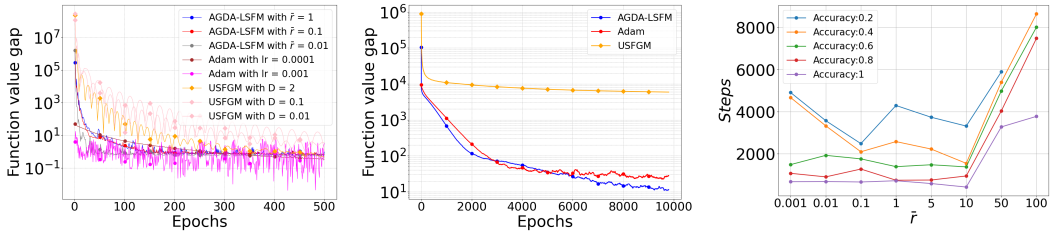

Figure 2: Performance of the compared algorithms Left: robustness test on diabetes dataset. Middle: Long-run test on Boston housing dataset. Right: robustness test on softmax problem

USFGM [30] and Adam exhibit unstable performance when hyperparameters are poorly tuned, while our algorithm maintains strong and consistent convergence across a wide range of settings.

To further validate algorithm efficiency in a practical regime, we repeat the experiment on the Boston housing dataset, tuning all methods with their best-performing hyperparameters. The middle panel of Figure 2 shows that our method achieves competitive long-term performance while preserving its robustness advantage. These results illustrate that our approach is not only stable but also effective in real-world stochastic optimization tasks.

5.3 Robustness

We conduct additional experiments to assess the robustness of our method with respect to the choice of the parameter $\bar{r}$, which reflects an estimate of the initial distance to the optimal solution, D_0. Our goal is to show that the performance of our algorithm remains stable across a wide range of $\bar{r}$ values, thereby reducing the sensitivity to inaccurate user-specified estimates.

To this end, we revisit the softmax minimization problem and vary $\bar{r}$ logarithmically from 10^{-4} to 10^4. For each setting, we fix the target function value tolerance at $\epsilon \in \{0.2, 0.4, 0.6, 0.8, 1\}$ and record the number of iterations required to reach the specified accuracy. The results, shown in the third panel of Figure 2, reveal that our method is highly robust: the number of iterations remains nearly constant across several orders of magnitude of $\bar{r}$. This suggests that our approach can tolerate significant misspecification of D_0 without compromising convergence efficiency. In practice, users may either provide a rough estimate of the initial distance or simply default to a moderate value such as $\bar{r} = 10^{-3}$, which performs consistently well across our tests.

6 Conclusion

This paper introduces a novel parameter-free first-order method for solving composite convex optimization problems without requiring prior knowledge of the initial distance to the optimum (D_0) or the Hölder smoothness parameters. Our method achieves a near-optimal complexity bound for locally Hölder smooth functions in an anytime fashion, making it broadly applicable and practical. In the stochastic setting, we further develop a line-search-free accelerated method that eliminates the need for estimating the problem-dependent diameter D during stepsize selection. This enhances both theoretical generality and practical usability. Preliminary experiments demonstrate that our algorithms are competitive and often outperform existing universal methods for Hölder smooth optimization, particularly in terms of robustness and adaptivity. An important direction for future research is to improve the dependence on the diameter D_0 in the convergence complexity, and to further relax the boundedness assumptions typically required in the stochastic setting.

References

[1] Aharon Ben-Tal and Arkadi Nemirovski. Non-euclidean restricted memory level method for large-scale convex optimization. *Mathematical Programming*, 102:407–456, 2005.

[2] Yair Carmon and Oliver Hinder. Making sgd parameter-free. In *Conference on Learning Theory*, pages 2360–2389. PMLR, 2022.

[3] Gong Chen and Marc Teboulle. Convergence analysis of a proximal-like minimization algorithm using bregman functions. *SIAM Journal on Optimization*, 3(3):538–543, 1993.

[4] Ashok Cutkosky and Kwabena Boahen. Online learning without prior information. In *Conference on learning theory*, pages 643–677. PMLR, 2017.

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

[6] Qi Deng, Guanghui Lan, and Zhenwei Lin. Uniformly optimal and parameter-free first-order methods for convex and function-constrained optimization. *arXiv preprint arXiv:2412.06319*, 2024.

[7] Nikhil Devanathan and Stephen Boyd. Polyak minorant method for convex optimization. *Journal of Optimization Theory and Applications*, pages 1–20, 2024.

[8] John C Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 2011.

[9] Alexander Vladimirovich Gasnikov and Yu E Nesterov. Universal method for stochastic composite optimization problems. *Computational Mathematics and Mathematical Physics*, 58: 48–64, 2018.

[10] Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Generalized uniformly optimal methods for nonlinear programming. *Journal of Scientific Computing*, 79:1854–1881, 2019.

[11] Benjamin Grimmer. Convergence rates for deterministic and stochastic subgradient methods without lipschitz continuity. *SIAM Journal on Optimization*, 29(2):1350–1365, 2019. ISSN 1052-6234.

[12] Elad Hazan and Sham Kakade. Revisiting the polyak step size. *arXiv preprint arXiv:1905.00313*, 2019.

[13] Maor Ivgi, Oliver Hinder, and Yair Carmon. Dog is sgd's best friend: A parameter-free dynamic step size schedule. In *International Conference on Machine Learning*, pages 14465–14499. PMLR, 2023.

[14] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[15] Guanghui Lan. Bundle-level type methods uniformly optimal for smooth and nonsmooth convex optimization. *Mathematical Programming*, 149(1-2):1–45, 2015.

[16] Guanghui Lan, Zhaosong Lu, and Renato DC Monteiro. Primal-dual first-order methods with iteration-complexity for cone programming. *Mathematical Programming*, 126(1):1–29, 2011.

[17] C. Lemaréchal, A. S. Nemirovski, and Y. E. Nesterov. New variants of bundle methods. *Mathematical Programming*, 69:111–148, 1995.

[18] Hong Liu, Zhiyuan Li, David Hall, Percy Liang, and Tengyu Ma. Sophia: A scalable stochastic second-order optimizer for language model pre-training. *arXiv preprint arXiv:2305.14342*, 2023.

[19] Zijian Liu and Zhengyuan Zhou. Stochastic nonsmooth convex optimization with heavy-tailed noises: High-probability bound, in-expectation rate and initial distance adaptation. *arXiv preprint arXiv:2303.12277*, 2023.

[20] Brendan Mcmahan and Matthew Streeter. No-regret algorithms for unconstrained online convex optimization. *Advances in neural information processing systems*, 25, 2012.

[21] H Brendan McMahan and Francesco Orabona. Unconstrained online linear learning in hilbert spaces: Minimax algorithms and normal approximations. In *Conference on Learning Theory*, pages 1020–1039. PMLR, 2014.

[22] Aaron Mishkin, Ahmed Khaled, Yuanhao Wang, Aaron Defazio, and Robert Gower. Directional smoothness and gradient methods: Convergence and adaptivity. *Advances in Neural Information Processing Systems*, 37:14810–14848, 2024.

[23] Mohammad Moshtaghifar, Anton Rodomanov, Daniil Vankov, and Sebastian Stich. Dada: Dual averaging with distance adaptation. *arXiv preprint arXiv:2501.10258*, 2025.

[24] Yu Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152(1):381–404, 2015.

[25] Francesco Orabona. Simultaneous model selection and optimization through parameter-free stochastic learning. *Advances in Neural Information Processing Systems*, 27, 2014.

[26] Francesco Orabona. Is nesterov's universal algorithm really universal?, November 2023. Blog post, Parameter-Free Learning and Optimization Algorithms.

[27] Francesco Orabona. Normalized gradients for all. *arXiv preprint arXiv:2308.05621*, 2023.

[28] Boris T Polyak. Introduction to optimization. 1987.

[29] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018.

[30] Anton Rodomanov, Ali Kavis, Yongtao Wu, Kimon Antonakopoulos, and Volkan Cevher. Universal gradient methods for stochastic convex optimization. *arXiv preprint arXiv:2402.03210*, 2024.

[31] Naum Zuselevich Shor. *Minimization methods for non-differentiable functions*, volume 3. Springer Science & Business Media, 2012.

[32] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26, 2012.

[33] Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization. *submitted to SIAM Journal on Optimization*, 2(3), 2008.

[34] Nikhil Vyas, Depen Morwani, Rosie Zhao, Mujin Kwun, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham M. Kakade. SOAP: Improving and stabilizing shampoo using adam. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*, 2025.

[35] Lin Xiao. Dual averaging method for regularized stochastic learning and online optimization. *Advances in Neural Information Processing Systems*, 22, 2009.

[36] Jiujia Zhang and Ashok Cutkosky. Parameter-free regret in high probability with heavy tails. *Advances in Neural Information Processing Systems*, 35:8000–8012, 2022.

[37] Yushun Zhang, Congliang Chen, Ziniu Li, Tian Ding, Chenwei Wu, Diederik P Kingma, Yinyu Ye, Zhi-Quan Luo, and Ruoyu Sun. Adam-mini: Use fewer learning rates to gain more. *arXiv preprint arXiv:2406.16793*, 2024.

Appendix

Structure of the Appendix In Section A, we discuss the limitations of our algorithms. Section B presents the proofs of the lemmas for completeness. In Sections C and D, we provide detailed proofs of the main results discussed in Sections 3 and 4. In Section E, we introduce two methods for automatically setting the hyperparameters. Finally, Section F offers additional details regarding the experiments.

A Limitation

Algorithm 1 significantly reduces the multiplicative overhead of choosing a sufficiently small parameter $\bar{r}$ from a polynomial to a logarithmic factor, and lowers the average number of gradient evaluations to one per iteration—compared to four per iteration in the Universal Fast Gradient Method (FGM) [24]. However, this improvement comes at the cost of increased computational burden during the line search procedure. Specifically, to accurately adapt to the local Hölder smoothness, our method requires a more precise selection of the parameter β_k, leading to a total of $\mathcal{O}(k \log k)$ line search operations after k iterations, whereas ,in contrast, FGM only requires $\mathcal{O}(k)$.

B Auxiliary lemmas

B.1 Proof of Lemma 1

This result was first established in [13, Lemma 3] and [[19, Lemma 30]. We give a proof for completeness.

Proof. Let $R_k = \frac{r_k}{\sum_{i=0}^{k-1} r_i}$ and $R_0^{-1} = 0$, then for any $k \geq 0$

$$r_{k+1} R_{k+1}^{-1} = r_k R_k^{-1} + r_k \frac{r_k}{r_{k+1}} = R_{k+1}^{-1} - \frac{r_k}{r_{k+1}} R_k^{-1}.$$

Then

$$\begin{aligned}\sum_{i=0}^{k-1} \frac{r_i}{r_{i+1}} &= \sum_{i=0}^{k-1} R_{i+1}^{-1} - \frac{r_i}{r_{i+1}} R_i^{-1} = \sum_{i=0}^{k-1} R_{i+1}^{-1} - R_i^{-1} + (1 - \frac{r_i}{r_{i+1}}) R_i^{-1} \\ &= R_k^{-1} + \sum_{i=0}^{k-1} (1 - \frac{r_i}{r_{i+1}}) R_i^{-1} \leq R_{k^*}^{-1} (1 + k - \sum_{i=0}^{k-1} \frac{r_i}{r_{i+1}}),\end{aligned}$$

where $k^* = \arg\min_{0 \leq i \leq k} R_i$. It then follows that

$$\begin{aligned}R_{k^*} \leq & \frac{1 + k - \sum_{i=0}^{k-1} \frac{r_i}{r_{i+1}}}{\sum_{i=0}^{k-1} \frac{r_i}{r_{i+1}}} \leq \frac{1 + k - k(\prod_{i=0}^{k-1} \frac{r_i}{r_{i+1}})^{\frac{1}{k}}}{k(\prod_{i=0}^{k-1} \frac{r_i}{r_{i+1}})^{\frac{1}{k}}} \\ = & \frac{1 + k - k(\frac{r_0}{r_k})^{\frac{1}{k}}}{k(\frac{r_0}{r_k})^{\frac{1}{k}}} \leq \frac{1 - k \log(\frac{r_0}{r_k})^{\frac{1}{k}}}{k(\frac{r_0}{r_k})^{\frac{1}{k}}} \\ = & \frac{1 - \log(\frac{r_0}{r_k})}{k(\frac{r_0}{r_k})^{\frac{1}{k}}} = (\frac{r_k}{r_0})^{\frac{1}{k}} \frac{\log(e \frac{r_k}{r_0})}{k}.\end{aligned}$$

□

B.2 Proof of auxiliary Lemmas

The following three-point Lemma is a well-known result. See also [3, Lemma 3.2] and [16, Lemma 6]. We give a proof for the sake of completeness.

Lemma 2. *For any proper lsc convex function $\phi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, any $z \in \operatorname{dom} \phi$ and $\beta > 0$. Let $z_+ = \arg\min_{x \in \operatorname{dom} \phi} \{\phi(x) + \frac{\beta}{2}\|z - x\|^2\}$. Then, we have*

$$\phi(x) + \frac{\beta}{2}\|z - x\|^2 \geq \phi(z_+) + \frac{\beta}{2}\|z - z_+\|^2 + \frac{\beta}{2}\|z_+ - x\|^2, \forall x \in \mathbb{R}^d. \tag{21}$$

Proof.

$$\begin{aligned}\frac{1}{2}\|z_+ - x\|^2 + \frac{1}{2}\|z - z_+\|^2 - \frac{1}{2}\|z - x\|^2 =& \frac{1}{2}\|x\|^2 - \langle z_+, x\rangle + \frac{1}{2}\|z_+\|^2 \\ &+ \frac{1}{2}\|z\|^2 - \langle z, z_+\rangle + \frac{1}{2}\|z_+\|^2 \\ &- \frac{1}{2}\|z\|^2 + \langle z, x\rangle - \frac{1}{2}\|x\|^2 \\ =& \langle z - z_+, x - z_+\rangle.\end{aligned}$$

In view of the first-order optimal condition at z_+, we have

$$\langle \nabla\phi(z_+) + \beta(z_+ - z), x - z_+\rangle \geq 0.$$

Combining the two inequalities above, we have

$$\begin{aligned}\frac{\beta}{2}\|z_+ - x\|^2 + \frac{\beta}{2}\|z - z_+\|^2 - \frac{\beta}{2}\|z - x\|^2 &= \beta\langle z - z_+, x - z_+\rangle \\ &\leq \langle \nabla\phi(z_+), x - z_+\rangle \\ &\leq \phi(x) - \phi(z_+),\end{aligned}$$

where the last inequality uses the convexity of $\phi(\cdot)$. □

Lemma 3. *For any $u \geq 0$, $k \geq 0$, there exits a positive constant c_u such that:*

$$(k+1)^u - k^u \geq c_u(k+1)^{u-1}, \tag{22}$$

where c_u only depends on u.

Proof. When $k = 0$, we have $(1)^u - 0^u = 1^u = 1^{u-1} = 1$. Now consider $k > 0$. We distinguish between two cases.

Case 1: If $u \geq 1$, we have

$$(k+1)^u - k^u = u\int_k^{k+1} x^{u-1}dx \geq uk^{u-1}$$

and hence

$$\left(\frac{k}{k+1}\right)^{u-1} \geq \left(\frac{1}{2}\right)^{u-1}.$$

Therefore,

$$(k+1)^u - k^u \geq u(\frac{1}{2})^{u-1}(k+1)^{u-1}.$$

Case 2: $0 \leq u < 1$,

$$(k+1)^u - k^u = u\int_k^{k+1} x^{u-1}dx \geq u(k+1)^{u-1}.$$

Therefore, we can set $c_u = u(\frac{1}{2})^{u-1}$. □

C Proof details in Section 3

In this section, we provide a detailed convergence analysis of Algorithm 1. For the sake of simplicity, we define the following notations.

$$\begin{aligned}\phi_{k+1}(x) &= \sum_{i=1}^{k+1} a_i[f(x^i) + \langle \nabla f(x^i), x - x^i\rangle + g(x)] + \frac{\beta_{k+1}}{2}\|x^0 - x\|^2, \\ \eta_k &= \frac{\beta_{k+1}\bar{r}_k^2 - \beta_k\bar{r}_{k-1}^2}{8a_{k+1}},\end{aligned}$$

Using this definition, it follows that $\phi_0(x) = \frac{\beta_0}{2}\|x^0 - x\|^2$.

C.1 Proof of important lemmas

To begin our analysis, we first prove the key bound (6) used in universal gradient methods. The bound (6) ensures that these methods can be accelerated by line search without prior knowledge about ν and L_ν.

The following result is from [[24], Lemma 2]. We give a proof for completeness.

Lemma 4. *Define* $\gamma(\hat{M}_\nu, \delta) = \gamma_\nu(\hat{M}_\nu, \delta) = (\frac{1-\nu}{1+\nu}\frac{1}{\delta})^{\frac{1-\nu}{1+\nu}}\hat{M}_\nu^{\frac{2}{1+\nu}}$*, here we regard* $(\frac{1-1}{1+1}\frac{1}{\delta})^{\frac{1-1}{1+1}} = 1$ *since* $\lim_{\nu\to 1}(\frac{1-\nu}{1+\nu}\frac{1}{\delta})^{\frac{1-\nu}{1+\nu}} = 1$.

Suppose $f(\cdot)$ *is locally Hölder smooth, i.e., for any* $x, y \in \mathcal{B}_{3D_0}(x^*)$*, we have*

$$\|\nabla f(x) - \nabla f(y)\|_* \le \hat{M}_\nu \|x - y\|^\nu. \tag{23}$$

Then we have

$$f(y) \le f(x) + \langle \nabla f(x), y - x\rangle + \frac{\gamma(\hat{M}_\nu, \delta)}{2}\|x - y\|^2 + \frac{\delta}{2}. \tag{24}$$

Proof. Note that the condition (23) immediately implies

$$f(y) \le f(x) + \langle \nabla f(x), y - x\rangle + \frac{\hat{M}_\nu}{1+\nu}\|x - y\|^{1+\nu}$$

from basic convex analysis.

For any $a, b \in \mathbb{R}_+$ and $p, q \ge 1$, $\frac{1}{p} + \frac{1}{q} = 1$, applying Young's inequality we obtain

$$\frac{a^p}{p} + \frac{b^q}{q} \ge ab. \tag{25}$$

We choose $p = \frac{2}{1+\nu}$, $q = \frac{2}{1-\nu}$, $a = t^{1+\nu}$ and $b = (\frac{1+\nu}{1-\nu}\frac{\delta}{\hat{M}_\nu})^{\frac{1-\nu}{1+\nu}}$ and have

$$\begin{aligned}
\frac{(1+\nu)t^2}{2} + \frac{(1-\nu)(\frac{1+\nu}{1-\nu}\frac{\delta}{\hat{M}_\nu})^{\frac{2}{1+\nu}}}{2} &\ge t^{1+\nu}(\frac{1+\nu}{1-\nu}\frac{\delta}{\hat{M}_\nu})^{\frac{1-\nu}{1+\nu}} \\
\frac{(1+\nu)t^2}{2}(\frac{1-\nu}{1+\nu}\frac{\hat{M}_\nu}{\delta})^{\frac{1-\nu}{1+\nu}} + \frac{(1+\nu)\delta}{2\hat{M}_\nu} &\ge t^{1+\nu} \\
\frac{t^2}{2}(\frac{1-\nu}{1+\nu}\frac{1}{\delta})^{\frac{1-\nu}{1+\nu}}\hat{M}_\nu^{\frac{2}{1+\nu}} + \frac{\delta}{2} &\ge \frac{t^{1+\nu}}{1+\nu}\hat{M}_\nu \\
\frac{t^2}{2}\gamma(\hat{M}_\nu, \delta) + \frac{\delta}{2} &\ge \frac{t^{1+\nu}}{1+\nu}\hat{M}_\nu.
\end{aligned}$$

We set $t = \|x - y\|$ and obtain (24) directly. □

To demonstrate the primary convergence results from Section 3, we first establish some useful lemmas regarding the well-definedness of line search and the boundedness of the iterates.

Lemma 5. *Given* $c \in \mathbb{R}^d$*, a proper lsc convex function* $g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$*. Let* $h \in \mathbb{R}_+$ *be a variable.*

Define $z(h) := \arg\min_{x\in\mathbb{R}^d}\{\langle c, x\rangle + g(x) + h\|x^0 - x\|^2\}, h \in \mathbb{R}_+$*. Then, the function* $\|x^0 - z(h)\|$ *is monotonically decreasing in* h *and converges to* 0 *as* $h \to +\infty$*.*

Proof. First, we prove $\|x^0 - z(h)\|$ is monotonically decreasing in h. For any h_1, h_2 such that $h_2 > h_1 > 0$, in view of the optimality of $z(h_1)$ and $z(h_2)$, we have

$$\langle c, z(h_1)\rangle + g(z(h_1)) + h_1\|x^0 - z(h_1)\|^2 \le \langle c, z(h_2)\rangle + g(z(h_2)) + h_1\|x^0 - z(h_2)\|^2 \tag{26}$$

and

$$\langle c, z(h_1)\rangle + g(z(h_1)) + h_2\|x^0 - z(h_1)\|^2 \ge \langle c, z(h_2)\rangle + g(z(h_2)) + h_2\|x^0 - z(h_2)\|^2. \tag{27}$$

Combining (27) and (26) and noticing that $h_2 - h_1 > 0$, we have

$$(h_2 - h_1)\|x^0 - z(h_1)\|^2 \geq (h_2 - h_1)\|x^0 - z(h_2)\|^2,$$

which implies

$$\|x^0 - z(h_1)\| \geq \|x^0 - z(h_2)\|.$$

Next, we prove $\lim_{h\to+\infty} \|x^0 - z(h)\| = 0$ by contradiction. If there exists $\delta > 0$, for any $h \in \mathbb{R}_+, \|x^0 - z(h)\| \geq \delta$, then we have

$$\langle c, z(h)\rangle + g(z(h)) + h\|x^0 - z(h)\|^2 \geq \langle c, z(h)\rangle + g(z(h)) + h\delta^2.$$

Let us consider $h > h_0 > 0$. Uusing the optimality of $z(h_0)$, we obtain

$$\langle c, z(h)\rangle + g(z(h)) + h_0\|x^0 - z(h)\|^2 \geq \langle c, z(h_0)\rangle + g(z(h_0)) + h_0\|x^0 - z(h_0)\|^2,$$

Note that monotonicity proved above implies $\|x^0 - z(h)\| \leq \|x^0 - z(h_0)\|$, together with the above inequality, we have

$$\langle c, z(h)\rangle + g(z(h)) \geq \langle c, z(h_0)\rangle + g(z(h_0)).$$

Moreover, using the optimality at $z(h)$ and the lower boundedness $\|x^0 - z(h)\| \geq \delta$, we have

$$\langle c, x^0\rangle + g(x^0) \geq \langle c, z(h)\rangle + g(z(h)) + h\|x^0 - z(h)\|^2 \geq \langle c, z(h_0)\rangle + g(z(h_0)) + h\delta^2. \quad (28)$$

Since h can be arbitrarily large, this result is impossible unless $\delta = 0$. □

Next, we establish an important property about the convergence error.

Lemma 6. *In Algorithm 1, suppose $\bar{r} \leq 4D_0$. If for $i = 0, 1, \ldots, k$, $l_i(\beta_{i+1}) \geq 0$, then we have the error bound*

$$\psi(y^{k+1}) - \psi(x^*) \leq \frac{\beta_{k+1}(D_0^2 - D_{k+1}^2)}{2A_{k+1}} + \frac{\beta_{k+1}\bar{r}_{k+1}^2}{8A_{k+1}}. \quad (29)$$

Proof. Since $l_k(\beta_{k+1}) \geq 0$, we have

$$\begin{aligned} l_k(\beta_{k+1}) =& - f(\tau_k v_{k+1}(\beta_{k+1}) + (1-\tau_k)y^k) + f(x^{k+1}) + \langle \nabla f(x^{k+1}), \tau_k v_{k+1}(\beta_{k+1}) \\ &+(1-\tau_k)y^k - x^{k+1}\rangle + \frac{\beta_{k+1}}{64\tau_k^2 A_{k+1}}\|\tau_k v_{k+1}(\beta) + (1-\tau_k)y^k - x^{k+1}\|^2 + \frac{\beta\bar{r}_k^2 - \beta_k\bar{r}_{k-1}^2}{16A_{k+1}} \geq 0, \end{aligned} \quad (30)$$

i.e.,

$$f(y^{k+1}) \leq f(x^{k+1}) + \langle \nabla f(x^{k+1}), y^{k+1} - x^{k+1}\rangle + \frac{\beta_{k+1}}{64\tau_k^2 A_{k+1}}\|y^{k+1} - x^{k+1}\|^2 + \frac{\tau_k\eta_k}{2}.$$

Because $x^{k+1} = \tau_k v^k + (1-\tau_k)y^k$, $y^{k+1} = \tau_k v^{k+1} + (1-\tau_k)y^k$ and $\eta_k = \frac{\beta_{k+1}\bar{r}_k^2 - \beta_k\bar{r}_{k-1}^2}{8a_{k+1}}$, it holds

$$\begin{aligned} f(y^{k+1}) \leq& (1-\tau_k)(f(x^{k+1}) + \langle \nabla f(x^{k+1}), y^k - x^{k+1}\rangle) + \tau_k(f(x^{k+1}) + \langle \nabla f(x^{k+1}), v^{k+1} - x^{k+1}\rangle) \\ &+ \frac{\beta_{k+1}}{64\tau_k^2 A_{k+1}}\tau_k^2\|v^{k+1} - v^k\|^2 + \frac{\beta_{k+1}\bar{r}_k^2 - \beta_k\bar{r}_{k-1}^2}{16A_{k+1}} \\ \leq& (1-\tau_k)f(y^k) + \tau_k(f(x^{k+1}) + \langle \nabla f(x^{k+1}), v^{k+1} - x^{k+1}\rangle) \\ &+ \frac{\beta_{k+1}}{64A_{k+1}}\|v^{k+1} - v^k\|^2 + \frac{\beta_{k+1}\bar{r}_k^2 - \beta_k\bar{r}_{k-1}^2}{16A_{k+1}}. \end{aligned}$$

Multiplying both sides by A_{k+1}, we have

$$\begin{aligned} A_{k+1}f(y^{k+1}) \leq& A_k f(y^k) + a_{k+1}(f(x^{k+1}) + \langle \nabla f(x^{k+1}), v^{k+1} - x^{k+1}\rangle) \\ &+ \frac{\beta_{k+1}}{64}\|v^{k+1} - v^k\|^2 + \frac{\beta_{k+1}\bar{r}_k^2 - \beta_k\bar{r}_{k-1}^2}{16}. \end{aligned}$$

Note that

$$\|v^{k+1} - v^k\|^2 \leq (\|v^{k+1} - x^0\| + \|v^k - x^0\|)^2 \leq (2\max\{\|v^{k+1} - x^0\|, \|v^k - x^0\|\})^2 \leq 4\bar{r}_{k+1}^2.$$

It follows that

$$\begin{aligned}
A_{k+1}f(y^{k+1}) \leq & A_k f(y^k) + a_{k+1}(f(x^{k+1}) + \langle \nabla f(x^{k+1}), v^{k+1} - x^{k+1}\rangle) \\
& + \frac{\beta_k}{64}\|v^{k+1} - v^k\|^2 + \frac{\beta_{k+1} - \beta_k}{64} 4\bar{r}_{k+1}^2 + \frac{\beta_{k+1}\bar{r}_k^2 - \beta_k \bar{r}_{k-1}^2}{16} \\
\leq & A_k f(y^k) + a_{k+1}(f(x^{k+1}) + \langle \nabla f(x^{k+1}), v^{k+1} - x^{k+1}\rangle) \\
& + \frac{\beta_k}{2}\|v^{k+1} - v^k\|^2 + \frac{\beta_{k+1} - \beta_k}{16} \bar{r}_{k+1}^2 + \frac{\beta_{k+1}\bar{r}_k^2 - \beta_k \bar{r}_{k-1}^2}{16} \\
\leq & A_k f(y^k) + a_{k+1}(f(x^{k+1}) + \langle \nabla f(x^{k+1}), v^{k+1} - x^{k+1}\rangle) \\
& + \frac{\beta_k}{2}\|v^{k+1} - v^k\|^2 + \frac{\beta_{k+1}\bar{r}_{k+1}^2 - \beta_k \bar{r}_k^2}{16} + \frac{\beta_{k+1}\bar{r}_k^2 - \beta_k \bar{r}_{k-1}^2}{16},
\end{aligned} \tag{31}$$

where the last inequality uses $\bar{r}_{k+1} \geq \bar{r}_k$.

On the other hand, since $g(\cdot)$ is convex, we have

$$g(y^{k+1}) \leq (1 - \tau_k) g(y^k) + \tau_k g(v^{k+1}). \tag{32}$$

Combining (31) and (32), we obtain

$$\begin{aligned}
A_{k+1}\psi(y^{k+1}) \leq & A_k \psi(y^k) + a_{k+1}(f(x^{k+1}) + \langle \nabla f(x^{k+1}), v^{k+1} - x^{k+1}\rangle + g(v^{k+1})) \\
& + \frac{\beta_k}{2}\|v^{k+1} - v^k\|^2 + \frac{\beta_{k+1}\bar{r}_{k+1}^2 - \beta_k \bar{r}_k^2}{16} + \frac{\beta_{k+1}\bar{r}_k^2 - \beta_k \bar{r}_{k-1}^2}{16},
\end{aligned}$$

For $\frac{\beta_k}{2}\|v^{k+1} - v^k\|^2$, we use Lemma 2, then

$$\begin{aligned}
A_{k+1}\psi(y^{k+1}) \leq & A_k \psi(y^k) + a_{k+1}(f(x^{k+1}) + \langle \nabla f(x^{k+1}), v^{k+1} - x^{k+1}\rangle + g(v^{k+1})) \\
& + \sum_{i=1}^{k} a_i (f(x^i) + \langle \nabla f(x^i), v^{k+1} - x^i\rangle + g(v^{k+1})) + \frac{\beta_k}{2}\|x^0 - v^{k+1}\|^2 \\
& - \sum_{i=1}^{k} a_i (f(x^i) + \langle \nabla f(x^i), v^{k} - x^i\rangle + g(v^{k})) - \frac{\beta_k}{2}\|x^0 - v^{k}\|^2 \\
& + \frac{\beta_{k+1}\bar{r}_{k+1}^2 - \beta_k \bar{r}_k^2}{16} + \frac{\beta_{k+1}\bar{r}_k^2 - \beta_k \bar{r}_{k-1}^2}{16} \\
\leq & A_k \psi(y^k) + \sum_{i=1}^{k+1} a_i (f(x^i) + \langle \nabla f(x^i), v^{k+1} - x^i\rangle + g(v^{k+1})) + \frac{\beta_{k+1}}{2}\|x^0 - v^{k+1}\|^2 \\
& - \sum_{i=1}^{k} a_i (f(x^i) + \langle \nabla f(x^i), v^{k} - x^i\rangle + g(v^{k})) - \frac{\beta_k}{2}\|x^0 - v^{k}\|^2 \\
& + \frac{\beta_{k+1}\bar{r}_{k+1}^2 - \beta_k \bar{r}_k^2}{16} + \frac{\beta_{k+1}\bar{r}_k^2 - \beta_k \bar{r}_{k-1}^2}{16}.
\end{aligned} \tag{33}$$

We can shorten the inequality (33) by using the definition of $\phi_k(\cdot)$:

$$A_{k+1}\psi(y^{k+1}) \leq A_k \psi(y^k) + \phi_{k+1}(v^{k+1}) - \phi_k(v^k) + \frac{\beta_{k+1}\bar{r}_{k+1}^2 - \beta_k \bar{r}_k^2}{16} + \frac{\beta_{k+1}\bar{r}_k^2 - \beta_k \bar{r}_{k-1}^2}{16}.$$

Applying the upper inequality recursively, it holds

$$\begin{aligned}
A_{k+1}\psi(y^{k+1}) \leq & \phi_{k+1}(v^{k+1}) - \phi_0(v^0) + \sum_{i=0}^{k} \frac{\beta_{i+1}\bar{r}_{i+1}^2 - \beta_i \bar{r}_i^2}{16} + \sum_{i=0}^{k} \frac{\beta_{i+1}\bar{r}_i^2 - \beta_i \bar{r}_{i-1}^2}{16} \\
\leq & \phi_{k+1}(v^{k+1}) + \frac{\beta_{k+1}}{16}\bar{r}_{k+1}^2 - \frac{\beta_0}{16}\bar{r}_0^2 + \frac{\beta_{k+1}}{16}\bar{r}_k^2 - \frac{\beta_0}{16}\bar{r}_{-1}^2 \\
\leq & \phi_{k+1}(v^{k+1}) + \frac{\beta_{k+1}}{16}\bar{r}_{k+1}^2 + \frac{\beta_{k+1}}{16}\bar{r}_{k+1}^2 \\
\leq & \phi_{k+1}(v^{k+1}) + \frac{\beta_{k+1}}{8}\bar{r}_{k+1}^2.
\end{aligned}$$

where $\phi_0(v^0) = 0$ and $\beta_0 > 0$.

Since $v^{k+1} = \arg\min_x \phi_{k+1}(x)$, we use Lemma 2 again and obtain that:

$$\begin{aligned}
A_{k+1}\psi(y^{k+1}) \leq & \phi_{k+1}(v^{k+1}) + \frac{\beta_{k+1}}{8}\bar{r}_{k+1}^2 \\
= & \sum_{i=1}^{k+1} a_i(f(x^i) + \langle \nabla f(x^i), v^k - x^i\rangle + g(v^k)) + \frac{\beta_{k+1}}{2}\|x^0 - v^{k+1}\|^2 + \frac{\beta_{k+1}}{8}\bar{r}_{k+1}^2 \\
\leq & \sum_{i=1}^{k+1} a_i(f(x^i) + \langle \nabla f(x^i), x^* - x^i\rangle + g(x^*)) + \frac{\beta_{k+1}}{2}\|x^0 - x^*\|^2 - \frac{\beta_{k+1}}{2}\|v^{k+1} - x^*\|^2 \\
& + \frac{\beta_{k+1}}{8}\bar{r}_{k+1}^2 \\
\leq & A_{k+1}\psi(x^*) + \frac{\beta_{k+1}}{2}\|x^0 - x^*\|^2 - \frac{\beta_{k+1}}{2}\|v^{k+1} - x^*\|^2 + \frac{\beta_{k+1}}{8}\bar{r}_{k+1}^2.
\end{aligned}$$

Finally, we use D_0 and D_{k+1} to replace $\|x^0 - x^*\|$ and $\|v^{k+1} - x^*\|$ and have

$$\begin{aligned}
A_{k+1}\psi(y^{k+1}) &\leq A_{k+1}\psi(x^*) + \frac{\beta_{k+1}}{2}D_0^2 - \frac{\beta_{k+1}}{2}D_{k+1}^2 + \frac{\beta_{k+1}}{8}\bar{r}_{k+1}^2 \\
\psi(y^{k+1}) - \psi(x^*) &\leq \frac{\beta_k(D_0^2 - D_{k+1}^2)}{2A_{k+1}} + \frac{\beta_k\bar{r}_{k+1}^2}{8A_{k+1}}.
\end{aligned}$$

□

Note that the convergence result above is conditioned on the success of the line search, which further requires the boundedness of the iterates. We prove these important properties in the following lemma.

Lemma 7. *Consider Algorithm 1. Suppose that for all $i = 0, 1, \ldots, k$, the iterates x^i, y^i and v^i lie within the set $\mathcal{B}_{3D_0}(x^*)$. Then, the line search in the k-th iteration terminates in a finite number of steps, and $x^{k+1}, y^{k+1}, v^{k+1}$ remain within $\mathcal{B}_{3D_0}(x^*)$.*

Proof. For clarity, we divide the proof into the following parts.

Part 1: Finite termination of the line search.

Given the value of x^k, y^k, A_{k+1}, τ_k , $\bar{r}_k$, $\bar{r}_{k-1}$, and β_k, $l_k(\beta)$ is defined by

$$\begin{aligned}
l_k(\beta) := & - f(\tau_k v_{k+1}(\beta) + (1-\tau_k)y^k) + f(x^{k+1}) + \langle \nabla f(x^{k+1}), \tau_k v_{k+1}(\beta) \\
& + (1-\tau_k)y^k - x^{k+1}\rangle + \frac{\beta}{64\tau_k^2 A_{k+1}}\|\tau_k v_{k+1}(\beta) + (1-\tau_k)y^k - x^{k+1}\|^2 + \frac{\beta\bar{r}_k^2 - \beta_k\bar{r}_{k-1}^2}{16A_{k+1}},
\end{aligned} \tag{34}$$

where $v_{k+1}(\beta) := \arg\min_{x\in\mathbb{R}^d} \sum_{i=1}^{k+1} a_i(\langle \nabla f(x^i), x\rangle + g(x)) + \frac{\beta}{2}\|x - x^0\|^2, \beta \in \mathbb{R}_+$.

We analyze the function $v_{k+1}(\beta)$ and $l_k(\beta)$ first. $v_{k+1}(\beta) \in \operatorname{dom} g$ is well-defined and unique since $\sum_{i=1}^k a_i(\langle \nabla f(x^i), x\rangle + g(x)) + \frac{\beta}{2}\|x - x^0\|, \beta \in \mathbb{R}_+$ is strong convex and has a unique optimum. We claim that $g(x)$ restricted to $\operatorname{dom} g$ is continuous since it is convex and lsc. The convexity guarantees $g(x)$ is continuous on the interior point of $\operatorname{dom} g$ and lower semicontinuity guarantees that it maintains the continuity on the remaining points of $\operatorname{dom} g$. Thus $v_k(\beta)$ is continuous. Since $l_k(\beta)$ is the composition of some continuous function, it is continuous as well.

Next, we discuss the behavior of $l_k(\beta)$ when $\beta \to +\infty$. Recall that we assume $x^k, v^k, y^k \in \mathcal{B}_{3D_0}(x^*)$, we shall first prove that the line search for y^{k+1} must be finitely terminated. Specifically, applying Lemma 5 with $c = \sum_{i=1}^{k+1} \nabla f(x^i)$ and $h = \frac{\beta}{2}$, we have that for a sufficiently large value $\hat{\beta}$, when $\beta \geq \hat{\beta}$, $\|x^0 - v_{k+1}(\beta)\| \leq 2D_0$, which further implies $\|x^* - v_{k+1}(\beta)\| \leq 3D_0$.

Let us consider $\beta \geq \beta_{k+1}^{\text{TH}}$, $\delta > 0$, where

$$\beta_{k+1}^{\text{TH}} := \max\left\{\hat{\beta}, \frac{8A_{k+1}\delta + \beta_k\bar{r}_{k-1}^2}{\bar{r}_k^2}, 32\tau_k^2 A_{k+1}\gamma(\hat{M}_\nu, \delta), \beta_k\right\}$$

we must have $v_{k+1}(\beta) \in \mathcal{B}_{3D_0}(x^*)$. Since $y^{k+1}(\beta)$ is a convex combination of $v_{k+1}(\beta)$ and y^k, we have $y^{k+1}(\beta) \in \mathcal{B}_{3D_0}(x^*)$. Similarly, we have $x^{k+1} \in \mathcal{B}_{3D_0}(x^*)$. Lemma 4 implies

$$f(y^{k+1}(\beta)) \leq f(x^{k+1}) + \langle \nabla f(x), y - x \rangle + \frac{\gamma(\hat{M}_\nu, \delta)}{2}\|x - y\|^2 + \frac{\delta}{2}, \forall x, y \in \mathcal{B}_{3D_0}(x^*). \quad (35)$$

Moreover, due to the definition of β_{k+1}^{TH}, we have

$$\frac{\gamma(\hat{M}_\nu, \delta)}{2} \leq \frac{\beta}{64\tau_k^2 A_{k+1}}, \quad \text{and} \quad \frac{\delta}{2} \leq \frac{\beta \bar{r}_k^2 - \beta_k \bar{r}_{k-1}^2}{16 A_{k+1}}.$$

Combining the above two results, we conclude that $l_k(\beta) \geq 0$. That is, $l_k(\beta)$ remains nonnegative when β exceeds a certain threshold, and hence the search will terminate in finitely many steps.

Furthermore, we would like to point out that $2\beta_{k+1}^{\text{TH}}$ is another threshold. For any $\beta \geq 2\beta_{k+1}^{\text{TH}}$, we have $l_k(\beta) > 0$. The reason is that the following inequalities hold:

$$\frac{\gamma(\hat{M}_\nu, \delta)}{2} \leq \frac{\beta}{64\tau_k^2 A_{k+1}} < \frac{2\beta}{64\tau_k^2 A_{k+1}}, \quad \text{and} \quad \frac{\delta}{2} \leq \frac{\beta \bar{r}_k^2 - \beta_k \bar{r}_{k-1}^2}{16 A_{k+1}} < \frac{2\beta \bar{r}_k^2 - \beta_k \bar{r}_{k-1}^2}{16 A_{k+1}}.$$

The second stage of the line search procedure also ends in finite steps as it employs a simple bisection method.

Part 2: Boundedness of the $(k+1)$-th iterates.

First, we immediately have $x^{k+1} = \tau_k v^k + (1 - \tau_k) y^k \in \mathcal{B}_{3D_0}(x^*)$ by the assumption. Next, we prove $y^{k+1}, v^{k+1} \in \mathcal{B}_{3D_0}(x^*)$. Part 1 implies $l_i(\beta_{i+1}) \geq 0$ $(i = 0, 1, \ldots, k)$. Applying Lemma 6, we have

$$\psi(y^{k+1}) - \psi(x^*) \leq \frac{\beta_{k+1}(D_0^2 - D_{k+1}^2)}{2A_{k+1}} + \frac{\beta_{k+1}\bar{r}_{k+1}^2}{8A_{k+1}}. \quad (36)$$

We shall consider two cases.

Case 1: $\bar{r}_{k+1} = \bar{r}_k$, then $r_{k+1} \leq \bar{r}_k \leq 4D_0$;

Case 2: $\bar{r}_{k+1} = r_{k+1}$.

Due to the non-negativity of the function value gap and (36), we have

$$0 \leq \frac{\beta_{k+1}[D_0^2 - D_{k+1}^2]}{2A_{k+1}} + \frac{\beta_{k+1} r_{k+1}^2}{8A_{k+1}}.$$

By dividing both sides by $\frac{\beta_{k+1}}{2A_{k+1}}$, we obtain

$$0 \leq D_0^2 - D_{k+1}^2 + \frac{r_{k+1}^2}{4},$$

which implies:

$$D_{k+1} \leq \sqrt{D_0^2 + \frac{r_{k+1}^2}{4}} \leq D_0 + \frac{r_{k+1}}{2}.$$

By the triangle inequality, we have:

$$r_{k+1} \leq D_0 + D_{k+1} \leq D_0 + D_0 + \frac{r_{k+1}}{2}$$
$$r_{k+1} \leq 4D_0.$$

By repeatedly using the $D_k \leq D_0 + \frac{\bar{r}_k}{2}$, we have:

$$D_k \leq D_0 + \frac{r_k}{2} \leq D_0 + 2D_0 \leq 3D_0.$$

That is, $\|v^{k+1} - x^*\| \leq 3D_0$ and thus $v^{k+1} \in \mathcal{B}_{3D_0}(x^*)$. Thus, we have $y^{k+1} \in \mathcal{B}_{3D_0}(x^*)$ as well since y^{k+1} is convex combination of v^{k+1} and y^k.

□

The following lemma develops an upper bound of β_k.

Lemma 8. *Suppose $f(\cdot)$ is locally Hölder smooth in $\mathcal{B}_{3D_0}(x^*)$ and $\beta_0 \leq 2^7 I_\nu \hat{M}_\nu \bar{r}^\nu$. In Algorithm 1, for any $k \geq 0$, given $\epsilon_{k+1}^l > 0$, if at least one of the following two propositions holds:*

1. *there exists β_{k+1}^*, such that $\beta_{k+1} \leq \beta_{k+1}^* + \epsilon_{k+1}^l$ and $l_k(\beta_{k+1}^*) = 0$;*
2. *$\beta_{k+1} = \beta_k$, where $l_k(\beta_k) \geq 0$*

then we have

$$\beta_k \leq 2^7 I_\nu \hat{M}_\nu \bar{r}_{k-1}^\nu k^{\frac{3-3v}{2}} + \sum_{i=1}^{k} \epsilon_i^l. \tag{37}$$

Proof. First, we estimate the growth of a_{k+1}. We have

$$\begin{aligned} a_{k+1} &= A_{k+1} - A_k = (A_{k+1}^{\frac{1}{2}})^2 - (A_k^{\frac{1}{2}})^2 \\ &= (A_{k+1}^{\frac{1}{2}} - A_k^{\frac{1}{2}})(A_{k+1}^{\frac{1}{2}} + A_k^{\frac{1}{2}}) \\ &\leq 2\bar{r}_k^{\frac{1}{2}} A_{k+1}^{\frac{1}{2}}, \end{aligned}$$

which gives

$$a_{k+1}^2 \leq 4\bar{r}_k A_{k+1}.$$

Next, for β_{k+1}^* that satisfies $l_k(\beta_{k+1}^*) = 0$, applying Lemma 6, we have

$$y^{k+1}(\beta_{k+1}^*) \in \mathcal{B}_{3D_0}(x^*), \tag{38}$$

where $y^{k+1}(\beta)$ is defined in the main text of the paper.

$l_k(\beta_{k+1}^*) = 0$ implies that

$$\begin{aligned} f(y^{k+1}(\beta_{k+1}^*)) = f(x^{k+1}) &+ \langle \nabla f(x^{k+1}), y^{k+1}(\beta_{k+1}^*) - x^{k+1} \rangle \\ &+ \frac{\beta_{k+1}^*}{64\tau_k^2 A_{k+1}} \|y^{k+1}(\beta_{k+1}^*) - x^{k+1}\|^2 + \frac{\beta_{k+1}^* \bar{r}_k^2 - \beta_k \bar{r}_{k-1}^2}{16 A_{k+1}}. \end{aligned} \tag{39}$$

Applying Lemma 4, we have

$$\begin{aligned} f(y^{k+1}(\beta_{k+1}^*)) = f(x^{k+1}) &+ \langle \nabla f(x^{k+1}), y^{k+1}(\beta_{k+1}^*) - x^{k+1} \rangle \\ &+ \frac{\gamma(\hat{M}_\nu, \delta)}{2} \|y^{k+1}(\beta_{k+1}^*) - x^{k+1}\|^2 + \frac{\delta}{2}. \end{aligned} \tag{40}$$

We take $\delta = \frac{\beta_{k+1}^* \bar{r}_k^2 - \beta_k \bar{r}_{k-1}^2}{8A_{k+1}}$, then combine (39) and (40), we have

$$\frac{\beta_{k+1}^*}{32\tau_k^2 A_{k+1}} \leq \gamma \left(\hat{M}_\nu, \frac{\beta_{k+1}^* \bar{r}_k^2 - \beta_k \bar{r}_{k-1}^2}{8A_{k+1}} \right); \tag{41}$$

We prove this lemma by induction. By the assumption on β_0, it holds for $k = 0$. Next, we assume it is valid for some certain k.

Case 1: $\beta_{k+1} = \beta_k$. Then we prove it directly since

$$\beta_{k+1} = \beta_k \leq 2^7 I_\nu \hat{M}_\nu \bar{r}_{k-1}^\nu k^{\frac{3-3v}{2}} + \sum_{i=1}^{k} \epsilon_i^l \leq 2^7 I_\nu \hat{M}_\nu \bar{r}_k^\nu (k+1)^{\frac{3-3v}{2}} + \sum_{i=1}^{k+1} \epsilon_i^l.$$

Case 2.a: $\beta_{k+1} \leq \beta_{k+1}^* + \epsilon_{k+1}^l$ and $\nu = 1$.

$$\begin{aligned}\frac{\beta^*_{k+1}}{32\tau_k^2 A_{k+1}} \leq & (\frac{8A_{k+1}}{\beta^*_{k+1}\bar{r}_k^2 - \beta_k \bar{r}_{k-1}^2})^0 \hat{M}_\nu = \hat{M}_\nu \\ \beta^*_{k+1} \leq & 2^7 \hat{M}_\nu \bar{r}_k = 2^7 I_\nu \hat{M}_\nu \bar{r}_k \\ \beta_{k+1} \leq & 2^7 I_\nu \hat{M}_\nu \bar{r}_k + \epsilon_i^l \leq 2^7 I_\nu \hat{M}_\nu \bar{r}_k + \sum_{i=1}^{k+1} \epsilon_i^l.\end{aligned}$$

Case 2.b: $\beta_{k+1} \leq \beta^*_{k+1} + \epsilon^l_{k+1}$ and $\nu \neq 1$. First we analyze β^*_{k+1}. Since β^*_{k+1} is the maximal zero point of $l_k(\cdot)$, applying Lemma 4, we have

$$\frac{\beta^*_{k+1}}{32\tau_k^2 A_{k+1}} \leq \gamma(\hat{M}_\nu, \frac{\beta^*_{k+1}\bar{r}_k^2 - \beta_k \bar{r}_{k-1}^2}{8A_{k+1}}),$$

i.e.

$$\frac{\beta^*_{k+1}}{32\tau_k^2 A_{k+1}} \leq \left(\frac{1-\nu}{1+\nu}\frac{8A_{k+1}}{\beta^*_{k+1}\bar{r}_k^2 - \beta_k \bar{r}_{k-1}^2}\right)^{\frac{1-\nu}{1+\nu}} \hat{M}_\nu{}^{\frac{2}{1+\nu}} \leq \left(\frac{8A_{k+1}}{(\beta^*_{k+1} - \beta_k)\bar{r}_k^2}\right)^{\frac{1-\nu}{1+\nu}} \hat{M}_\nu{}^{\frac{2}{1+\nu}}.$$

It can be rewritten in the following form:

$$\beta^*_{k+1}(\beta^*_{k+1} - \beta_k)^{\frac{1-\nu}{1+\nu}} \leq 2^{\frac{10+4\nu}{1+\nu}} \bar{r}_k^{\frac{2\nu}{1+\nu}} (k+1)^{2\frac{1-\nu}{1+\nu}} \hat{M}_\nu{}^{\frac{2}{1+\nu}}.$$

As β_{k+1} increases with $\beta_{k+1} \geq \beta_k$, the left-hand side also increases. Thus, by identifying a value where the left-hand side is at most equal to the right-hand side, we can determine an upper bound for β_{k+1}.

Let $c_\nu = 2^7(\frac{1}{1-\nu})^{\frac{1+\nu}{2}}\hat{M}_\nu$. For $c_\nu \bar{r}_k^\nu (k+1)^{\frac{3-3\nu}{2}} + \sum_{i=1}^k \epsilon_i^l$, we have

$$\begin{aligned}&(c_\nu \bar{r}_k^\nu (k+1)^{\frac{3-3\nu}{2}} + \sum_{i=1}^k \epsilon_i^l)(c_\nu \bar{r}_k^\nu (k+1)^{\frac{3-3\nu}{2}} + \sum_{i=1}^k \epsilon_i^l - \beta_k)^{\frac{1-\nu}{1+\nu}} \\ \geq & (c_\nu \bar{r}_k^\nu (k+1)^{\frac{3-3\nu}{2}} + \sum_{i=1}^k \epsilon_i^l)(c_\nu \bar{r}_k^\nu (k+1)^{\frac{3-3\nu}{2}} + \sum_{i=1}^k \epsilon_i^l - c_\nu \bar{r}_{k-1}^\nu k^{\frac{3-3\nu}{2}} - \sum_{i=1}^k \epsilon_i^l)^{\frac{1-\nu}{1+\nu}} \\ \geq & (c_\nu \bar{r}_k^\nu (k+1)^{\frac{3-3\nu}{2}} + \sum_{i=1}^k \epsilon_i^l)(c_\nu \bar{r}_k^\nu (k+1)^{\frac{3-3\nu}{2}} - c_\nu \bar{r}_{k-1}^\nu k^{\frac{3-3\nu}{2}})^{\frac{1-\nu}{1+\nu}} \\ \geq & c_\nu \bar{r}_k^\nu (k+1)^{\frac{3-3\nu}{2}} (c_\nu \bar{r}_k^\nu (k+1)^{\frac{3-3\nu}{2}} - c_\nu \bar{r}_k^\nu k^{\frac{3-3\nu}{2}})^{\frac{1-\nu}{1+\nu}} \\ \geq & c_\nu^{\frac{2}{1+\nu}} \bar{r}_k^{\frac{2\nu}{1+\nu}} (k+1)^{\frac{3-3\nu}{2}} ((k+1)^{\frac{3-3\nu}{2}} - k^{\frac{3-3\nu}{2}})^{\frac{1-\nu}{1+\nu}}\end{aligned} \tag{42}$$

Since $\frac{3-3v}{2} \in [0, \frac{3}{2}]$, and $\min_{1 \leq u \leq \frac{3}{2}} (\frac{1}{2})^{u-1} = 2^{-\frac{1}{2}} > \frac{1}{2}$, $\frac{3-3v}{2}(\frac{1}{2})^{\frac{3-3v}{2}-1} \geq \frac{3-3v}{4}$. Applying lemma 3, it holds that

$$\begin{aligned}&c_\nu^{\frac{2}{1+\nu}} \bar{r}_k^{\frac{2\nu}{1+\nu}} (k+1)^{\frac{3-3\nu}{2}} ((k+1)^{\frac{3-3\nu}{2}} - k^{\frac{3-3\nu}{2}})^{\frac{1-\nu}{1+\nu}} \\ \geq & \frac{3-3\nu}{4} c_\nu^{\frac{2}{1+\nu}} \bar{r}_k^{\frac{2\nu}{1+\nu}} (k+1)^{\frac{3-3\nu}{2}} ((k+1)^{\frac{1-3\nu}{2}})^{\frac{1-\nu}{1+\nu}} \\ \geq & \frac{3-3\nu}{4} (\frac{1}{1-\nu}) 2^{\frac{14}{1+\nu}} \hat{M}_\nu^{\frac{2}{1+\nu}} \bar{r}_k^{\frac{2\nu}{1+\nu}} (k+1)^{\frac{3-3\nu}{2}} ((k+1)^{\frac{1-3\nu}{2}})^{\frac{1-\nu}{1+\nu}} \\ \geq & 2^{\frac{10+4\nu}{1+\nu}} \hat{M}_\nu^{\frac{2\nu}{1+\nu}} \bar{r}_k^{\frac{2\nu}{1+\nu}} (k+1)^{2\frac{1-\nu}{1+\nu}}\end{aligned} \tag{43}$$

This inequality implies that $\beta^*_{k+1} \leq 2^7 I_\nu \hat{M}_\nu \bar{r}_k^\nu (k+1)^{\frac{3-3\nu}{2}} + \sum_{i=1}^k \epsilon_i^l$ and $\beta_{k+1} \leq \beta^*_{k+1} + \epsilon^l_{k+1} \leq 2^7 I_\nu \hat{M}_\nu \bar{r}_k^\nu (k+1)^{\frac{3-3\nu}{2}} + \sum_{i=1}^{k+1} \epsilon_i^l$. □

Moreover, since we set $\epsilon_k^l = \frac{\beta_0}{2k^2}$, we can obtain an upper bound of β_k as follows

$$\beta_{k+1} \leq 2^7 I_\nu \hat{M}_\nu \bar{r}_k^\nu (k+1)^{\frac{3-3\nu}{2}} + \sum_{i=1}^{k+1} \frac{\beta_0}{2i^2} \leq 2^7 I_\nu \hat{M}_\nu \bar{r}_k^\nu (k+1)^{\frac{3-3\nu}{2}} + \beta_0 \leq 2^8 I_\nu \hat{M}_\nu \bar{r}_k^\nu (k+1)^{\frac{3-3\nu}{2}}.$$

C.2 Proof of Proposition 1

Proof. In the proof of Lemma 7, we have proved that in the first stage of the line search terminates in a finite number of steps, and thus $l_k(2^{i'_k-1}\beta_k) \geq 0$. Moreover, we also proved that $l_k(\cdot)$ is continuous. Next, we show that at least one of the two propositions mentioned in Proposition 1 is correct.

When $i'_k = 1$, we have $l_k(2^{i'_k-1}\beta_k) = l_k(\beta_k) \geq 0$. Thus, the first proposition holds in this case.

When $i'_k > 1$, we have $l_k(2^{i'_k-1}\beta_k) \geq 0$ and $l_k(2^{i'_k-2}\beta_k) < 0$. Since $l_k(\cdot)$ is continuous, based on the intermediate value theorem, there exists at least one root in the interval $[2^{i'_k-2}\beta_k, 2^{i'_k-1}\beta_k]$. Moreover, as previously discussed in Lemma 7, there exists a threshold $2\beta^{\text{TH}}$ such that for any $\beta \geq 2\beta^{\text{TH}}_{k+1}$, $l_k(\beta) > 0$. Now, since $l_k(\beta)$ has at least one root and the set of roots has an upper bound, there exists a maximal root β^*_{k+1} of the continuous function $l_k(\cdot)$, and therefore the second proposition holds.

It remains to estimate the upper bound of the amount of searching. Without loss of generality, we assume that $\beta_0 \leq 2^7 I_\nu \hat{M}_\nu \bar{r}^v$ in the Algorithm 1. This is a common assumption in the previous work, for example, see [24], and it is reasonable since the upper bound of the searched value increases polynomial in k. The initial value of the searched value is not very sensitive. Thus all the conditions of Lemma 8 are satisfied, we apply it to obtain that $\beta_{k+1} \leq \mathcal{O}(k^{\frac{3-3\nu}{2}})$.

The first stage in the line search in the k-th iteration starts from β_k to at most $2\beta_{k+1}$. So the length of the interval is at most $2\beta_{k+1} - \beta_k$ and this stage in line search procedure requires at most $i'_k \leq 1 + \log(\frac{2\beta_{k+1}}{\beta_k})$ times in the k-th iteration. We sum them up and obtain that up to the k-th iteration, the total amount of line search operations in the first stage is at most

$$\sum_{j=1}^{k} i'_j \leq (1 + \log 2)k + \log(\frac{\beta_{k+1}}{\beta_0}) \leq \mathcal{O}(k + \log k).$$

The second step in line search process in the k-th iteration start from $2^{i_k-1}\beta_k$ to at most $2^{i_k}\beta_k$. Hence, the interval length is at most $2^{i_k}\beta_k - 2^{i_k-1}\beta_k = 2^{i_k-1}\beta_k \leq \beta_{k+1}$ and this step of process requires at most $i^*_k - i'_k \leq 1 + \log(\frac{\beta_{k+1}}{\epsilon^l_{k+1}})$ times in the k-th iteration. We sum them up and obtain that up to the k-th iteration, the total amount of line search operations in the first part is at most

$$\sum_{j=1}^{k} i^*_j \leq k + \sum_{i=1}^{k} \log\left(\frac{\beta_i}{\epsilon^l_i}\right) = k + \log\left(\prod_{i=1}^{k} \beta_i\right) - \log\left(\prod_{i=1}^{k} \epsilon^l_i\right) \leq \mathcal{O}(k \log k),$$

where we apply Lemma!8 to estimate β_k.

To summarize, the total amount of line search operations is $\mathcal{O}(k \log k)$. □

C.3 Proof of Proposition 2

Proof. We apply Lemma 6 and 7 to prove it by induction. First, $x^0 = v^0 = y^0 \in \mathcal{B}_{3D_0}(x^*)$. Next, we assume it holds for some $k \geq 0$.

Since $x^k, v^k, y^k \in \mathcal{B}_{3D_0}(x^*)$, we have $l_k(\beta_{k+1}) \geq 0$ and $x^{k+1}, v^{k+1}, y^{k+1} \in \mathcal{B}_{3D_0}(x^*)$ by Lemma 7. Thus, apply Lemma 6, it holds that

$$\psi(y^{k+1}) - \psi(x^*) \leq \frac{\beta_{k+1}(D_0^2 - D_{k+1}^2)}{2A_{k+1}} + \frac{\beta_{k+1}\bar{r}_{k+1}^2}{8A_{k+1}}. \tag{44}$$

Therefore, for any $k > 0$, we have

$$\psi(y^k) - \psi(x^*) \leq \frac{\beta_k(D_0^2 - D_k^2)}{2A_k} + \frac{\beta_k\bar{r}_k^2}{8A_k}. \tag{45}$$

□

C.4 Proof of Theorem 1

Theorem 1 is almost a direct corollary of Proposition 2.

Proof. The proof is similar to the proof of Lemma 7. For completeness, we give a proof that directly uses Proposition 2 and induction.

Since we assume that $\bar{r}$ is small enough such that $\bar{r} \leq 4D_0$, then $\bar{r}_0 = \bar{r} \leq 4D_0$. Next, we assume it holds for some certain $k \geq 0$, then $x^{k+1} = \tau_k v^k + (1-\tau_k)y^k$ lies in $\mathcal{B}_{3D_0}(x^*)$.

Case 1: $\bar{r}_{k+1} = \bar{r}_k$, then $r_{k+1} \leq \bar{r}_k \leq 4D_0$;

Case 2: $\bar{r}_{k+1} = r_{k+1}$. Applying Proposition 2, we have

$$0 \leq \psi(y^{k+1}) - \psi(x^*) \leq \frac{\beta_{k+1}(D_0^2 - D_{k+1}^2)}{2A_{k+1}} + \frac{\beta_{k+1}\bar{r}_{k+1}^2}{8A_{k+1}}. \tag{46}$$

Then it holds that

$$\begin{aligned}
0 &\leq D_0^2 - D_{k+1}^2 + \frac{r_{k+1}^2}{4} \\
D_{k+1}^2 &\leq D_0^2 + \frac{r_{k+1}^2}{4} \\
D_{k+1} &\leq \sqrt{D_0^2 + \frac{r_{k+1}^2}{4}} \leq D_0 + \frac{r_{k+1}}{2}.
\end{aligned}$$

By the triangle inequality, we have:

$$\begin{aligned}
r_{k+1} &\leq D_0 + D_{k+1} \leq D_0 + D_0 + \frac{r_{k+1}}{2} \\
r_{k+1} &\leq 4D_0.
\end{aligned}$$

Repeat using the $D_k \leq D_0 + \frac{\bar{r}_k}{2}$, we have:

$$D_k \leq D_0 + \frac{r_k}{2} \leq D_0 + 2D_0 \leq 3D_0.$$

That is, $\|v^{k+1} - x^*\| \leq 3D_0$ and thus $v^{k+1} \in \mathcal{B}_{3D_0}(x^*)$. $y^{k+1} \in \mathcal{B}_{3D_0}(x^*)$ as well since y^{k+1} is convex combination of v^{k+1} and y^k.

In conclusion, we prove that for any $i \in \mathbb{N}$, $\|v^i - x^0\| \leq 4D_0$ and $\|v^i - x^*\| \leq 3D_0$. □

The boundedness property is important for us to remove the standard global Hölder smoothness assumption on the full domain. Instead, we can safely use the local Hölder smoothness assumption as all the iterates remain in the ball $\mathcal{B}_{3D_0}(x^*)$.

Finally, all the preparatory work for Theorem 2 is now complete. Proposition 1 ensures the practical implementability of the algorithm, while Proposition 2 provides the tool needed to analyze the convergence rate. Furthermore, Theorem 1, Lemma 1 and Lemma 8 ensure that the convergence rate achieves the best-known rate.

C.5 Proof of Theorem 2

Proof. Using the triangle inequality, it holds that

$$\begin{aligned}
D_k &\geq |D_0 - r_k| \\
D_k^2 &\geq D_0^2 - 2D_0 r_k + r_k^2 \\
2D_0 r_k &\geq D_0^2 - D_k^2.
\end{aligned} \tag{47}$$

In view of Proposition 2, we have

$$\psi(y^k) - \psi(x^*) \leq \frac{\beta_k[D_0^2 - D_k^2]}{2A_k} + \frac{\beta_k \bar{r}_k^2}{8A_k} \leq \frac{2\beta_k D_0 r_k}{2A_k} + \frac{\beta_k \bar{r}_k^2}{8A_k} \leq \frac{\beta_k D_0 \bar{r}_k}{A_k} + \frac{\beta_k \bar{r}_k^2}{8A_k}, \tag{48}$$

where we use (47).

Applying Theorem 1, we can match the right hand side of inequality (48):

$$\psi(y^k) - \psi(x^*) \leq \frac{\beta_k D_0 \bar{r}_k}{A_k} + \frac{4\beta_k D_0 \bar{r}_k}{8A_k} \leq \frac{3\beta_k D_0 \bar{r}_k}{2A_k}, \tag{49}$$

Denote $k^* = \arg\min_{0\leq i\leq k} \frac{\bar{r}_k^{\frac{1}{2}}}{\sum_{i=0}^{k-1} \bar{r}_i^{\frac{1}{2}}}$. Applying Lemma 1 gives

$$\frac{\bar{r}_{k^*}^{\frac{1}{2}}}{\sum_{i=0}^{k^*-1} \bar{r}_i^{\frac{1}{2}}} \leq \frac{(\frac{\bar{r}_k}{\bar{r}_0})^{\frac{1}{2}\times\frac{1}{k}} \log e(\frac{\bar{r}_k}{\bar{r}_0})^{\frac{1}{2}}}{k} \leq \frac{(\frac{4D_0}{\bar{r}})^{\frac{1}{2k}} \log e\frac{4D_0}{\bar{r}}}{2k}. \tag{50}$$

Without loss of generality, we assume that $\beta_0 \leq 2^7 I_\nu \hat{M}_\nu \bar{r}^v$ in Algorithm 1. The reason is as same as that we mentioned in the proof of Proposition 1.

Thus, for k^*, combining the inequality (50) and Lemma 8, it holds that

$$\begin{aligned}
\min_{y\in\{y^0,y^1\dots y^k\}} \psi(y) - \psi(x^*) \leq & \psi(y^{k^*}) - \psi(x^*) \\
\leq & \frac{3\beta_{k^*} D_0 \bar{r}_{k^*}}{2A_{k^*}} \\
\leq & \frac{3}{2}\beta_{k^*} D_0 \left(\frac{\bar{r}_{k^*}^{\frac{1}{2}}}{\sum_{i=0}^{k^*-1} \bar{r}_i^{\frac{1}{2}}}\right)^2 \\
\leq & \frac{3}{2} \times 2^8 I_\nu \hat{M}_\nu D_0 \bar{r}_{k^*}^\nu {k^*}^{\frac{3-3v}{2}} \left(\frac{(\frac{4D_0}{\bar{r}})^{\frac{1}{2k}} \log e\frac{4D_0}{\bar{r}}}{2k}\right)^2 \\
\leq & 384 I_\nu (\frac{4D_0}{\bar{r}})^{\frac{1}{k}} \log^2 e\frac{4D_0}{\bar{r}} \frac{\hat{M}_\nu D_0 \bar{r}_{k^*}^\nu {k^*}^{\frac{3-3v}{2}}}{k^2} \\
\leq & 384 I_\nu (\frac{4D_0}{\bar{r}})^{\frac{1}{k}} \log^2 e\frac{4D_0}{\bar{r}} \frac{\hat{M}_\nu D_0^{1+\nu} k^{\frac{3-3v}{2}}}{k^2} \\
\leq & 384 I_\nu (\frac{4D_0}{\bar{r}})^{\frac{1}{k}} \log^2 e\frac{4D_0}{\bar{r}} \frac{\hat{M}_\nu D_0^{1+\nu}}{k^{\frac{1+3v}{2}}},
\end{aligned} \tag{51}$$

where we use the fact $(k^*)^{\frac{3-3v}{2}} \leq k^{\frac{3-3v}{2}}$.

It remains to use that $\psi(z^k) = \min_{y\in\{y^0,y^1\dots y^k\}} \psi(y)$ by definition. Since $(\frac{4D_0}{\bar{r}})^{\frac{1}{k}} \leq 2$ when $k \geq \log(\frac{4D_0}{\bar{r}})/\log 2$, the convergence rate of Algorithm 1 is

$$\mathcal{O}\left(\frac{\hat{M}_\nu D_0^{1+\nu} \log^2 e\frac{D_0}{\bar{r}}}{k^{\frac{1+3\nu}{2}}}\right). \tag{52}$$

□

Remark 4. *In the proof of Theorem 2, we show that the convergence rate of Algorithm 1 has a multiplicative factor I_ν. In fact, we can set $A_{k+1} = (\sum_{i=0}^k \bar{r}_i^{\frac{1}{n}})^n$, where $n \geq 2, n \in \mathbb{Z}_+$. By doing this, we can achieve a result similar to that of Theorem 2. If we choose $n \geq 3$, then the multiplicative factor I_ν will be replaced by another constant, which does not depend on ν, as I_ν is generated by Lemma 3.*

D Proof details in Section 4

In this section, we provide the detailed proof of the results in Section 4 and conduct the convergence rate of Algorithm 2. For the stochastic case, we use the notation $-\xi_k$ to present the other random variables except ξ_k. The proofs are similar to the deterministic case, however, we do not need to prove the boundedness since we have assumed it under Assumption 1.

Lemma 9 and 10 are provided to analyze the balance equation to estimate of β_k.

Lemma 9. *$\forall \alpha \geq 0, \beta > 0, \nu \in [0, 1)$, we have*

$$\alpha r^{1+\nu} - \beta r^2 \leq \frac{1+\nu}{2}(\frac{1-\nu}{2})^{\frac{1+\nu}{1-\nu}}(\alpha^{\frac{2}{1-\nu}}/\beta^{\frac{1+\nu}{1-\nu}}), r \geq 0. \tag{53}$$

This auxiliary result has been used in [[30], Lemma E.3]. We give proof for completeness.

Proof. It is easy to see that $\alpha r^{1+\nu} - \beta r^2$ increases first and then decreases later as $r \geq 0$ increases. It achieves maximum on $[0, +\infty)$ iff its gradient equals to zero, i.e $r = (\frac{(1-\nu)\alpha}{2\beta})^{\frac{1}{1-\nu}}$.

Thus, for $r \geq 0$,

$$\begin{aligned}
\alpha r^{1+\nu} - \beta r^2 \leq & \alpha((\frac{(1-\nu)\alpha}{2\beta})^{\frac{1}{1-\nu}})^{1+\nu} - \beta((\frac{(1-\nu)\alpha}{2\beta})^{\frac{1}{1-\nu}})^2 \\
\leq & \alpha(\frac{(1-\nu)\alpha}{2\beta})^{\frac{1+\nu}{1-\nu}} - \beta(\frac{(1-\nu)\alpha}{2\beta})^{\frac{2}{1-\nu}} \\
\leq & (\frac{1-\nu}{2})^{\frac{1+\nu}{1-\nu}}\frac{\alpha^{\frac{2}{1-\nu}}}{\beta^{\frac{1+\nu}{1-\nu}}} - (\frac{1-\nu}{2})^{\frac{2}{1-\nu}}\frac{\alpha^{\frac{2}{1-\nu}}}{\beta^{\frac{1+\nu}{1-\nu}}} \\
\leq & (\frac{1+\nu}{2})(\frac{1-\nu}{2})^{\frac{1+\nu}{1-\nu}}\frac{\alpha^{\frac{2}{1-\nu}}}{\beta^{\frac{1+\nu}{1-\nu}}}.
\end{aligned}$$

□

Lemma 10. *For nonnegative sequences $\{\alpha_i\}_{i\in\mathbb{N}}$ and $\{\gamma_i\}_{i\in\mathbb{N}}$, the sequence $\{h_i\}_{i\in\mathbb{N}}$ satisfies that*

$$h_{k+1} - h_k \leq \frac{(1-\nu)\alpha_{k+1}}{h_{k+1}^{\frac{\nu}{1-\nu}}} + \gamma_{k+1}, \tag{54}$$

with $h_0 = 0$. Then for $k \geq 1$, we have

$$h_k \leq (\sum_{i=1}^{k}\alpha_i)^{1-\nu} + \sum_{i=1}^{k}\gamma_i. \tag{55}$$

This auxiliary result has been used in [[30], Lemma E.9]. We give proof for completeness.

Proof. We prove it by induction.

Since $h_0 = 0 \leq 0$, we assume it is valid for some $k \geq 0$, then

$$\begin{aligned}
h_{k+1} - \frac{(1-\nu)\alpha_{k+1}}{h_{k+1}^{\frac{\nu}{1-\nu}}} \leq & \gamma_{k+1} + h_k \\
\leq & \gamma_{k+1} + (\sum_{i=1}^{k}\alpha_i)^{1-\nu} + \sum_{i=1}^{k}\gamma_i \\
\leq & (\sum_{i=1}^{k}\alpha_i)^{1-\nu} + \sum_{i=1}^{k+1}\gamma_i.
\end{aligned}$$

Define $\Gamma_{k+1}(x) := x - \frac{(1-\nu)\alpha_{k+1}}{x^{\frac{\nu}{1-\nu}}}$. It is easy to verify that $\Gamma_{k+1}(x)$ is increasing in x. Hence, it suffices to show

$$(\sum_{i=1}^{k}\alpha_i)^{1-\nu} + \sum_{i=1}^{k+1}\gamma_i \leq \Gamma_{k+1}((\sum_{i=1}^{k+1}\alpha_i)^{1-\nu} + \sum_{i=1}^{k+1}\gamma_i),$$

which means

$$(\sum_{i=1}^{k}\alpha_i)^{1-\nu} + \sum_{i=1}^{k+1}\gamma_i \leq (\sum_{i=1}^{k+1}\alpha_i)^{1-\nu} + \sum_{i=1}^{k+1}\gamma_i - (1-\nu)\alpha_{k+1}((\sum_{i=1}^{k+1}\alpha_i)^{1-\nu} + \sum_{i=1}^{k+1}\gamma_i)^{\frac{-\nu}{1-\nu}}. \tag{56}$$

Rearranging (56) gives us

$$(1-\nu)\alpha_{k+1}((\sum_{i=1}^{k+1}\alpha_i)^{1-\nu}+\sum_{i=1}^{k+1}\gamma_i)^{\frac{-\nu}{1-\nu}} \leq (\sum_{i=1}^{k+1}\alpha_i)^{1-\nu}-(\sum_{i=1}^{k}\alpha_i)^{1-\nu},$$

which is implied by

$$(1-\nu)\alpha_{k+1} \leq ((\sum_{i=1}^{k+1}\alpha_i)^{1-\nu}-(\sum_{i=1}^{k}\alpha_i)^{1-\nu})(\sum_{i=1}^{k+1}\alpha_i)^{\nu}. \tag{57}$$

The inequality (57) is valid since

$$\begin{aligned}
&((\sum_{i=1}^{k+1}\alpha_i)^{1-\nu}-(\sum_{i=1}^{k}\alpha_i)^{1-\nu})(\sum_{i=1}^{k+1}\alpha_i)^{\nu}\\
\geq&((\sum_{i=1}^{k+1}\alpha_i)^{1-\nu}-(\sum_{i=1}^{k+1}\alpha_i)^{1-\nu}+(1-\nu)(\sum_{i=1}^{k+1}\alpha_i)^{-\nu}a_{k+1})(\sum_{i=1}^{k+1}\alpha_i)^{\nu}\\
=&((1-\nu)(\sum_{i=1}^{k+1}\alpha_i)^{-\nu}a_{k+1})(\sum_{i=1}^{k+1}\alpha_i)^{\nu}\\
=&(1-\nu)a_{k+1},
\end{aligned}$$

where the first inequality is due to the first order condition of the concave function $(\cdot)^{1-\nu}, \nu \in [0,1]$. □

Next, we provide the upper bound of the expectation of β_k.

Lemma 11. *Suppose the Assumption 1 holds and $f(\cdot)$ is locally Hölder smooth in $\mathcal{B}_{3D_0}(x^*)$. In Algorithm 2, it holds that*

$$\mathbb{E}[\beta_k] \leq 2^{\frac{9+9\nu}{2}}\tilde{M}_\nu D^\nu k^{\frac{3-3\nu}{2}}+2^5k^{\frac{3}{2}}\sigma. \tag{58}$$

Proof. We denote $\|\nabla f(x^{k+1})-\tilde{\nabla} f(x^{k+1})\|_* = \Delta_{k+1}^x$ and $\|\nabla f(y^{k+1})-\tilde{\nabla} f(y^{k+1})\|_* = \Delta_{k+1}^y$.

The expectation of $(\Delta_k^x)^2$ and $(\Delta_k^y)^2$ satisfies

$$\begin{aligned}
\mathbb{E}[(\Delta_k^x)^2] = \mathbb{E}[\|\nabla f(x^{k+1})-\tilde{\nabla} f(x^{k+1})\|_*^2] \leq \sigma^2,\\
\mathbb{E}[(\Delta_k^y)^2] = \mathbb{E}[\|\nabla f(y^{k+1})-\tilde{\nabla} f(y^{k+1})\|_*^2] \leq \sigma^2.
\end{aligned} \tag{59}$$

From the balance equation, we have

$$\begin{aligned}
\frac{\tau_k\eta_k}{2} =&[\langle\tilde{\nabla} f(y^{k+1})-\tilde{\nabla} f(x^{k+1}), y^{k+1}-x^{k+1}\rangle-\frac{\beta_{k+1}}{64\tau_k^2A_{k+1}}\|y^{k+1}-x^{k+1}\|^2]_+\\
\leq&[\langle\nabla f(y^{k+1})-\nabla f(x^{k+1}), y^{k+1}-x^{k+1}\rangle+\langle\nabla f(x^{k+1})-\tilde{\nabla} f(x^{k+1}), y^{k+1}-x^{k+1}\rangle\\
&+\langle\tilde{\nabla} f(y^{k+1})-\nabla f(y^{k+1}), y^{k+1}-x^{k+1}\rangle-\frac{\beta_{k+1}}{64\tau_k^2A_{k+1}}\|y^{k+1}-x^{k+1}\|^2]_+,\\
\leq&[\tilde{M}_\nu\|y^{k+1}-x^{k+1}\|^{1+\nu}+(\Delta_{k+1}^y+\Delta_{k+1}^x)\|y^{k+1}-x^{k+1}\|-\frac{\beta_{k+1}}{64\tau_k^2A_{k+1}}\|y^{k+1}-x^{k+1}\|^2]_+.
\end{aligned}$$

Note that $[\cdot]_+$ is a monotonically increasing function. The first inequality uses the convexity of $f(\cdot)$ and the second inequality applies the locally Hölder smoothness and Cauchy-Schwarz inequality.

Case 1: $\nu = 1$

$$
\begin{aligned}
\frac{\beta_{k+1}\bar{r}_k^2-\beta_k\bar{r}_{k-1}^2}{2A_{k+1}} \leq & [(\tilde{M}_\nu-\frac{\beta_{k+1}}{64\tau_k^2A_{k+1}})\|y^{k+1}-x^{k+1}\|^2+(\Delta_{k+1}^y+\Delta_{k+1}^x)\|y^{k+1}-x^{k+1}\|]_+\\
\beta_{k+1}\bar{r}_k^2-\beta_k\bar{r}_{k-1}^2 \leq & 2A_{k+1}[(\tilde{M}_\nu-\frac{\beta_{k+1}}{64\tau_k^2A_{k+1}})\|y^{k+1}-x^{k+1}\|^2+(\Delta_{k+1}^y+\Delta_{k+1}^x)\|y^{k+1}-x^{k+1}\|]_+\\
\beta_{k+1}\bar{r}_k^2-\beta_k\bar{r}_{k-1}^2 \leq & 2A_{k+1}[(\tilde{M}_\nu-\frac{\beta_{k+1}}{2^8\bar{r}_k})\|y^{k+1}-x^{k+1}\|^2+(\Delta_{k+1}^y+\Delta_{k+1}^x)\|y^{k+1}-x^{k+1}\|]_+\\
\beta_{k+1}\bar{r}_k^2-\beta_k\bar{r}_{k-1}^2 \leq & 2(k+1)^2[\bar{r}_k(\tilde{M}_\nu-\frac{\beta_{k+1}}{2^8\bar{r}_k})\|y^{k+1}-x^{k+1}\|^2+\bar{r}_k(\Delta_{k+1}^y+\Delta_{k+1}^x)\|y^{k+1}-x^{k+1}\|]_+\\
\beta_{k+1}-\beta_k \leq & \frac{2(k+1)^2}{\bar{r}_k}[(\tilde{M}_\nu-\frac{\beta_{k+1}}{2^8\bar{r}_k})\|y^{k+1}-x^{k+1}\|^2+(\Delta_{k+1}^y+\Delta_{k+1}^x)\|y^{k+1}-x^{k+1}\|]_+,
\end{aligned}
$$

where we use $A_{k+1}\leq(k+1)^2\bar{r}_k$ and $D\geq\bar{r}_{k+1}\geq\bar{r}_k$.

We prove that

$$
\beta_k^2\leq\sum_{i=1}^k 2^{10}i^2((\Delta_i^x)^2+(\Delta_i^y)^2)+2^{18}\tilde{M}_\nu^2D^2. \tag{60}
$$

Since $\beta_0^2=0<2^{18}\tilde{M}_\nu^2D^2$, we prove it by induction. Define $k^*=\max\{i|\beta_i\leq 2^9\tilde{M}_\nu D\}\geq 0$, so $\forall k\leq k^*$ satisfies the inequality 60. We assume it is valid for certain $k\geq k^*$, then

$$
\begin{aligned}
\beta_{k+1}-\beta_k \leq & \frac{1}{\bar{r}_k}[2(k+1)^2(\frac{\beta_{k+1}}{2^9D}-\frac{\beta_{k+1}}{2^8\bar{r}_k})\|y^{k+1}-x^{k+1}\|^2+2(k+1)^2(\Delta_{k+1}^x+\Delta_{k+1}^y)\|y^{k+1}-x^{k+1}\|]_+\\
\leq & \frac{1}{\bar{r}_k}[-2(k+1)^2\frac{\beta_{k+1}}{2^9\bar{r}_k}\|y^{k+1}-x^{k+1}\|^2+2(k+1)^2(\Delta_{k+1}^x+\Delta_{k+1}^y)\|y^{k+1}-x^{k+1}\|]_+\\
\leq & \frac{1}{\bar{r}_k}2(k+1)^2\frac{2^9\bar{r}_k(\Delta_{k+1}^x)^2}{2\beta_{k+1}}+\frac{1}{\bar{r}_k}2(k+1)^2\frac{2^9\bar{r}_k(\Delta_{k+1}^y)^2}{2\beta_{k+1}}\\
= & 2^9(k+1)^2\frac{(\Delta_{k+1}^x)^2+(\Delta_{k+1}^y)^2}{\beta_{k+1}}.
\end{aligned}
$$

Then

$$
\begin{aligned}
\frac{1}{2}(\beta_{k+1}^2-\beta_k^2)\leq\beta_{k+1}(\beta_{k+1}-\beta_k)\leq & 2^9(k+1)^2(\Delta_{k+1}^y+\Delta_{k+1}^x)^2\\
\beta_{k+1}^2\leq & 2^{10}(k+1)^2(\Delta_{k+1}^y+\Delta_{k+1}^x)^2+\beta_k^2\\
\beta_{k+1}^2\leq & 2^{10}(k+1)^2(\Delta_{k+1}^y+\Delta_{k+1}^x)^2+\beta_k^2\\
\beta_{k+1}^2\leq & 2^{10}(k+1)^2(\Delta_{k+1}^y+\Delta_{k+1}^x)^2+\sum_{i=1}^k 2^{10}i^2(\Delta_i^y+\Delta_i^x)^2+2^{18}\tilde{M}_\nu^2D^2\\
\beta_{k+1}^2\leq & \sum_{i=1}^{k+1}2^{10}i^2(\Delta_i^y+\Delta_i^x)^2+2^{18}\tilde{M}_\nu^2D^2,
\end{aligned}
$$

where we use $\beta_k^2\leq\sum_{i=1}^k 2^{10}i^2\Delta_i^2+2^{18}\tilde{M}_\nu^2D^2$ by induction. Applying the inequality $a^2+b^2\leq(a+b)^2$ where $a,b\geq 0$, we obtain

$$
\beta_{k+1}\leq 2^5(\sum_{i=1}^{k+1}i^2(\Delta_i^x+\Delta_i^y)^2)^{\frac{1}{2}}+2^9\tilde{M}_\nu D.
$$

We take the expectation of β_k:

$$\begin{aligned}\mathbb{E}[\beta_k] \leq& \mathbb{E}[2^5(\sum_{i=1}^{k} i^2(\Delta_i^x+\Delta_i^y)^2))^{\frac{1}{2}} + 2^9\tilde{M}_\nu D]\\ \leq& 2^5(\sum_{i=1}^{k} i^2(\mathbb{E}[(\Delta_i^x)^2]+\mathbb{E}[(\Delta_i^y)^2])^{\frac{1}{2}} + 2^9\tilde{M}_\nu D\\ \leq& 2^5\sum_{i=1}^{k}\sqrt{2}i^2\sigma + 2^9\tilde{M}_\nu D\\ \leq& 2^{\frac{11}{2}}k^{\frac{3}{2}}\sigma + 2^9\tilde{M}_\nu D,\end{aligned}$$

where we use the Jensen's inequality that $\mathbb{E}[X^{\frac{1}{2}}] \leq (\mathbb{E}[X])^{\frac{1}{2}}$ and estimate $\sum_{i=1}^{k} i^2 \leq k^3$ roughly.

Case 2: $\nu \neq 1$ Applying Lemma 9 for three times with $r = \|y^{k+1}-x^{k+1}\|$, $\alpha = \tilde{M}_\nu$, $\beta = \frac{\beta_{k+1}}{128\tau_k^2 A_{k+1}}$; $r' = \|y^{k+1}-x^{k+1}\|$, $\alpha' = \Delta_{k+1}^x$, $\beta' = \frac{\beta_{k+1}}{256\tau_k^2 A_{k+1}}$; $r'' = \|y^{k+1}-x^{k+1}\|$, $\alpha'' = \Delta_{k+1}^y$, $\beta'' = \frac{\beta_{k+1}}{256\tau_k^2 A_{k+1}}$, it holds that

$$\begin{aligned}\frac{\tau_k\eta_k}{2} \leq& [\frac{1+\nu}{2}(\frac{1-\nu}{2})^{\frac{1+\nu}{1-\nu}}\tilde{M}_\nu^{\frac{2}{1-\nu}}(\frac{128\tau_k^2A_{k+1}}{\beta_{k+1}})^{\frac{1+\nu}{1-\nu}} + \frac{1}{2}(\Delta_{k+1}^y+\Delta_{k+1}^x)^2\frac{256\tau_k^2A_{k+1}}{\beta_{k+1}}]_+\\ =& \frac{1+\nu}{2}(\frac{1-\nu}{2})^{\frac{1+\nu}{1-\nu}}\tilde{M}_\nu^{\frac{2}{1-\nu}}(\frac{128\tau_k^2A_{k+1}}{\beta_{k+1}})^{\frac{1+\nu}{1-\nu}} + (\Delta_{k+1}^y+\Delta_{k+1}^x)^2\frac{128\tau_k^2A_{k+1}}{\beta_{k+1}},\end{aligned}$$

where the last equality is obviously nonnegative.

Similar to the proof of Lemma 8, we have

$$a_{k+1}^2 \leq 4\bar{r}_k A_{k+1},$$

which implies $\tau_k^2 A_{k+1} \leq 4\bar{r}_k$. Consequently,

$$\begin{aligned}\frac{\tau_k\eta_k}{2} \leq& \frac{1+\nu}{2}\tilde{M}_\nu^{\frac{2}{1-\nu}}(\frac{(1-\nu)2^8\bar{r}_k}{\beta_{k+1}})^{\frac{1+\nu}{1-\nu}} + (\Delta_{k+1}^y+\Delta_{k+1}^x)^2\frac{2^9\bar{r}_k}{\beta_{k+1}}\\ \beta_{k+1}\bar{r}_k^2 - \beta_k\bar{r}_{k-1}^2 \leq& A_{k+1}\tilde{M}_\nu^{\frac{2}{1-\nu}}(\frac{(1-\nu)2^8\bar{r}_k}{\beta_{k+1}})^{\frac{1+\nu}{1-\nu}} + A_{k+1}(\Delta_{k+1}^y+\Delta_{k+1}^x)^2\frac{2^{10}\bar{r}_k}{\beta_{k+1}}\\ \beta_{k+1}\bar{r}_k^2 - \beta_k\bar{r}_k^2 \leq& (k+1)^2\bar{r}_k\tilde{M}_\nu^{\frac{2}{1-\nu}}(\frac{(1-\nu)2^8\bar{r}_k}{\beta_{k+1}})^{\frac{1+\nu}{1-\nu}} + (k+1)^2\bar{r}_k(\Delta_{k+1}^y+\Delta_{k+1}^x)^2\frac{2^{10}\bar{r}_k}{\beta_{k+1}}\\ \beta_{k+1}-\beta_k \leq& (k+1)^2\tilde{M}_\nu^{\frac{2}{1-\nu}}(\frac{(1-\nu)2^8}{\beta_{k+1}})^{\frac{1+\nu}{1-\nu}}\bar{r}_k^{\frac{2\nu}{1-\nu}} + (k+1)^2(\Delta_{k+1}^y+\Delta_{k+1}^x)^2\frac{2^{10}}{\beta_{k+1}},\end{aligned}$$

where we use $A_{k+1} \leq (k+1)^2\bar{r}_k$ and $\bar{r}_k \geq \bar{r}_{k-1}$. It then follows that

$$\begin{aligned}\beta_{k+1}(\beta_{k+1}-\beta_k) \leq& (k+1)^2\tilde{M}_\nu^{\frac{2}{1-\nu}}\frac{((1-\nu)2^8)^{\frac{1+\nu}{1-\nu}}}{\beta_{k+1}^{\frac{2\nu}{1-\nu}}}\bar{r}_k^{\frac{2\nu}{1-\nu}} + 2^{10}(k+1)^2\Delta_{k+1}^2\\ \beta_{k+1}^2-\beta_k^2 \leq& 2(k+1)^2\tilde{M}_\nu^{\frac{2}{1-\nu}}\frac{((1-\nu)2^8)^{\frac{1+\nu}{1-\nu}}}{\beta_{k+1}^{\frac{2\nu}{1-\nu}}}\bar{r}_k^{\frac{2\nu}{1-\nu}} + 2^{11}(k+1)^2\Delta_{k+1}^2.\end{aligned}$$

We apply Lemma 10 with $h_k = \beta_k^2$, $\alpha_k = 2\times(2^8)^{\frac{1+\nu}{1-\nu}}k^2\tilde{M}_\nu^{\frac{2}{1-\nu}}(1-\nu)^{\frac{2\nu}{1-\nu}}\bar{r}_{k-1}^{\frac{2\nu}{1-\nu}}$ and $\gamma_k = 2^{11}(k+1)^2\Delta_k^2$, it holds that

$$\begin{aligned}\beta_k^2 \leq& (\sum_{i=1}^{k} 2\times(2^8)^{\frac{1+\nu}{1-\nu}}i^2\tilde{M}_\nu^{\frac{2}{1-\nu}}(1-\nu)^{\frac{2\nu}{1-\nu}}\bar{r}_{i-1}^{\frac{2\nu}{1-\nu}})^{1-\nu} + \sum_{i=1}^{k}2^{11}i^2\Delta_k^2\\ \leq& (2^{9+7\nu})\tilde{M}_\nu^2(1-\nu)^{2\nu}\bar{r}_{k-1}^{2\nu}(\sum_{i=1}^{k}i^2)^{1-\nu} + 2^{11}\sum_{i=1}^{k}i^2\Delta_i^2\\ \leq& (2^{9+7\nu})\tilde{M}_\nu^2(1-\nu)^{2\nu}D^{2\nu}k^{3-3\nu} + 2^{11}\sum_{i=1}^{k}i^2\Delta_i^2.\end{aligned}$$

Here we estimate $\sum_{i=1}^{k} i^2 \leq k^3$ roughly. Applying the inequality $a^2+b^2 \leq (a+b)^2$ where $a, b \geq 0$, we obtain

$$\beta_k \leq (2^{\frac{9+7\nu}{2}})\tilde{M}_\nu (1-\nu)^\nu D^\nu k^{\frac{3-3\nu}{2}} + 2^{\frac{11}{2}} (\sum_{i=1}^{k} i^2 \Delta_i^2)^{\frac{1}{2}}.$$

Finally, we take the expectation of β_k:

$$\begin{aligned}
\mathbb{E}[\beta_k] \leq & \mathbb{E}[(2^{\frac{9+7\nu}{2}})\tilde{M}_\nu (1-\nu)^\nu D^\nu k^{\frac{3-3\nu}{2}} + 2^{\frac{11}{2}} (\sum_{i=1}^{k} i^2 \Delta_i^2)^{\frac{1}{2}}] \\
\leq & (2^{\frac{9+7\nu}{2}})\tilde{M}_\nu (1-\nu)^\nu D^\nu k^{\frac{3-3\nu}{2}} + 2^{\frac{11}{2}} (\sum_{i=1}^{k} i^2 \mathbb{E}[\Delta_i^2])^{\frac{1}{2}} \\
\leq & (2^{\frac{9+7\nu}{2}})\tilde{M}_\nu (1-\nu)^\nu D^\nu k^{\frac{3-3\nu}{2}} + 2^{\frac{11}{2}} \sigma (\sum_{i=1}^{k} i^2)^{\frac{1}{2}} \\
\leq & (2^{\frac{9+7\nu}{2}})\tilde{M}_\nu (1-\nu)^\nu D^\nu k^{\frac{3-3\nu}{2}} + 2^{\frac{11}{2}} k^{\frac{3}{2}} \sigma,
\end{aligned}$$

where we use Jensen's inequality again.

In conclusion, we have $\mathbb{E}[\beta_k] \leq 2^{\frac{9+9\nu}{2}} \tilde{M}_\nu D^\nu k^{\frac{3-3\nu}{2}} + 2^{\frac{11}{2}} k^{\frac{3}{2}} \sigma$. □

We are now ready to derive the convergence rate of Algorithm 2.

D.1 Proof of Theorem 3

Proof. In view of the balance equation (15) and the fact $[x]_+ \geq x$, it holds that

$$\begin{aligned}
0 \leq & \langle \tilde{\nabla} f(x^{k+1}) - \tilde{\nabla} f(y^{k+1}), y^{k+1} - x^{k+1} \rangle + \frac{\beta_{k+1}}{64\tau_k^2 A_{k+1}} \|y^{k+1} - x^{k+1}\|^2 + \frac{\tau_k \eta_k}{2} \\
\langle \nabla f(y^{k+1}), y^{k+1} - x^{k+1} \rangle \leq & \langle \tilde{\nabla} f(x^{k+1}), y^{k+1} - x^{k+1} \rangle + \frac{\beta_{k+1}}{64\tau_k^2 A_{k+1}} \|y^{k+1} - x^{k+1}\|^2 + \frac{\tau_k \eta_k}{2} \\
& + \langle \nabla f(y^{k+1}) - \tilde{\nabla} f(y^{k+1}), y^{k+1} - x^{k+1} \rangle
\end{aligned}$$

The first order condition of convex function $f(\cdot)$ implies $f(y^{k+1}) - f(x^{k+1}) \leq \langle \nabla f(y^{k+1}), y^{k+1} - x^{k+1} \rangle$, thus

$$\begin{aligned}
f(y^{k+1}) \leq & f(x^{k+1}) + \langle \tilde{\nabla} f(x^{k+1}), y^{k+1} - x^{k+1} \rangle + \frac{\beta_{k+1}}{64\tau_k^2 A_{k+1}} \|y^{k+1} - x^{k+1}\|^2 + \frac{\tau_k \eta_k}{2} \\
& + \langle \nabla f(y^{k+1}) - \tilde{\nabla} f(y^{k+1}), y^{k+1} - x^{k+1} \rangle.
\end{aligned}$$

Because $x^{k+1} = \tau_k v^k + (1-\tau_k) y^k$, $y^{k+1} = \tau_k \hat{x}^{k+1} + (1-\tau_k) y^k$ and $\eta_k = \frac{\beta_{k+1} \bar{r}_k^2 - \beta_k \bar{r}_{k-1}^2}{8 a_{k+1}}$, it holds that

$$\begin{aligned}
f(y^{k+1}) \leq & (1-\tau_k)(f(x^{k+1}) + \langle \tilde{\nabla} f(x^{k+1}), y^k - x^{k+1} \rangle) + \tau_k (f(x^{k+1}) + \langle \tilde{\nabla} f(x^{k+1}), \hat{x}^{k+1} - x^{k+1} \rangle) \\
& + \frac{\beta_{k+1}}{64\tau_k^2 A_{k+1}} \tau_k^2 \|\hat{x}^{k+1} - v^k\|^2 + \frac{\beta_{k+1} - \beta_k}{16 A_{k+1}} \bar{r}_k^2 \\
& + \langle \nabla f(y^{k+1}) - \tilde{\nabla} f(y^{k+1}), y^{k+1} - x^{k+1} \rangle \\
\leq & (1-\tau_k) f(y^k) + \tau_k (f(x^{k+1}) + \langle \tilde{\nabla} f(x^{k+1}), \hat{x}^{k+1} - x^{k+1} \rangle) \\
& + \frac{\beta_k}{64 A_{k+1}} \|\hat{x}^{k+1} - v^k\|^2 + \frac{\beta_{k+1} - \beta_k}{16 A_{k+1}} \bar{r}_k^2 \\
& + (1-\tau_k) \langle \tilde{\nabla} f(x^{k+1}) - \nabla f(x^{k+1}), y^k - x^{k+1} \rangle + \frac{\beta_{k+1} - \beta_k}{64 A_{k+1}} \|\hat{x}^{k+1} - v^k\|^2 \\
& + \langle \nabla f(y^{k+1}) - \tilde{\nabla} f(y^{k+1}), y^{k+1} - x^{k+1} \rangle.
\end{aligned}$$

Since $a_{k+1}\langle\tilde{\nabla}f(x^{k+1}),\hat{x}^{k+1}\rangle+\frac{\beta_k}{2}\|\hat{x}^{k+1}-v^k\|^2\le a_{k+1}\langle\tilde{\nabla}f(x^{k+1}),v^{k+1}\rangle+\frac{\beta_k}{2}\|v^{k+1}-v^k\|^2$, we have

$$\begin{aligned}
f(y^{k+1})\le&(1-\tau_k)f(y^k)+\tau_k(f(x^{k+1})+\langle\tilde{\nabla}f(x^{k+1}),v^{k+1}-x^{k+1}\rangle)\\
&+\frac{\beta_k}{64A_{k+1}}\|v^{k+1}-v^k\|^2+\frac{\beta_{k+1}-\beta_k}{16A_{k+1}}\bar{r}_k^2\\
&+(1-\tau_k)\langle\tilde{\nabla}f(x^{k+1})-\nabla f(x^{k+1}),y^k-x^{k+1}\rangle+\frac{\beta_{k+1}-\beta_k}{64A_{k+1}}\|\hat{x}^{k+1}-v^k\|^2\\
&+\langle\nabla f(y^{k+1})-\tilde{\nabla}f(y^{k+1}),y^{k+1}-x^{k+1}\rangle\\
\le&(1-\tau_k)f(y^k)+\tau_k(f(x^{k+1})+\langle\tilde{\nabla}f(x^{k+1}),v^{k+1}-x^{k+1}\rangle)\\
&+\frac{\beta_k}{64A_{k+1}}\|v^{k+1}-v^k\|^2+\frac{\beta_{k+1}-\beta_k}{16A_{k+1}}\bar{r}_k^2\\
&+(1-\tau_k)\langle\tilde{\nabla}f(x^{k+1})-\nabla f(x^{k+1}),y^k-x^{k+1}\rangle+\frac{\beta_{k+1}-\beta_k}{16A_{k+1}}\bar{r}_{k+1}^2\\
&+\langle\nabla f(y^{k+1})-\tilde{\nabla}f(y^{k+1}),y^{k+1}-x^{k+1}\rangle.
\end{aligned}$$

Here, the last inequality is due to $\|\hat{x}^{k+1}-v^k\|^2\le(d_{k+1}+r_k)^2\le 4\bar{r}_{k+1}^2$.

We match the two error terms by $\frac{\beta_{k+1}\bar{r}_k^2-\beta_k\bar{r}_k^2}{16A_{k+1}}+\frac{\beta_{k+1}-\beta_k}{16A_{k+1}}\bar{r}_{k+1}^2\le\frac{\beta_{k+1}-\beta_k}{8A_{k+1}}\bar{r}_{k+1}^2$. Multiplying both sides by A_{k+1}, we have

$$\begin{aligned}
A_{k+1}f(y^{k+1})\le&A_kf(y^k)+a_{k+1}(f(x^{k+1})+\langle\tilde{\nabla}f(x^{k+1}),v^{k+1}-x^{k+1}\rangle)\\
&+\frac{\beta_k}{64}\|v^{k+1}-v^k\|^2+\frac{\beta_{k+1}-\beta_k}{8A_{k+1}}\bar{r}_{k+1}^2\\
&+A_k\langle\tilde{\nabla}f(x^{k+1})-\nabla f(x^{k+1}),y^k-x^{k+1}\rangle\\
&+A_{k+1}\langle\nabla f(y^{k+1})-\tilde{\nabla}f(y^{k+1}),y^{k+1}-x^{k+1}\rangle.
\end{aligned}$$

On the other hand, it holds that

$$g(y^{k+1})\le(1-\tau_k)g(y^k)+\tau_kg(v^{k+1}).\tag{61}$$

Combining (31) and (61), we obtain

$$\begin{aligned}
A_{k+1}\psi(y^{k+1})\le&A_k\psi(y^k)+a_{k+1}(f(x^{k+1})+\langle\tilde{\nabla}f(x^{k+1}),v^{k+1}-x^{k+1}\rangle+g(v^{k+1}))\\
&+\frac{\beta_k}{2}\|v^{k+1}-v^k\|^2+\frac{\beta_{k+1}-\beta_k}{8A_{k+1}}\bar{r}_{k+1}^2\\
&+A_k\langle\tilde{\nabla}f(x^{k+1})-\nabla f(x^{k+1}),y^k-x^{k+1}\rangle\\
&+A_{k+1}\langle\nabla f(y^{k+1})-\tilde{\nabla}f(y^{k+1}),y^{k+1}-x^{k+1}\rangle.
\end{aligned}$$

For $\frac{\beta_k}{2}\|v^{k+1}-v^k\|^2$, we use Lemma 2, then

$$
\begin{aligned}
A_{k+1}\psi(y^{k+1}) \leq & A_k\psi(y^k) + a_{k+1}(f(x^{k+1}) + \langle\tilde{\nabla}f(x^{k+1}), v^{k+1}-x^{k+1}\rangle + g(v^{k+1})) \\
& + \sum_{i=1}^{k} a_i(f(x^i) + \langle\tilde{\nabla}f(x^i), v^{k+1}-x^i\rangle + g(v^{k+1})) + \frac{\beta_k}{2}\|x^0-v^{k+1}\|^2 \\
& - \sum_{i=1}^{k} a_i(f(x^i) + \langle\tilde{\nabla}f(x^i), v^{k}-x^i\rangle + g(v^{k})) - \frac{\beta_k}{2}\|x^0-v^{k}\|^2 \\
& + \frac{\beta_{k+1}-\beta_k}{8A_{k+1}}\bar{r}_{k+1}^2 + A_k\langle\tilde{\nabla}f(x^{k+1})-\nabla f(x^{k+1}), y^k-x^{k+1}\rangle \\
& + A_{k+1}\langle\nabla f(y^{k+1})-\tilde{\nabla}f(y^{k+1}), y^{k+1}-x^{k+1}\rangle \\
\leq & A_k\psi(y^k) + \sum_{i=1}^{k+1} a_i(f(x^i) + \langle\tilde{\nabla}f(x^i), v^{k+1}-x^i\rangle + g(v^{k+1})) + \frac{\beta_{k+1}}{2}\|x^0-v^{k+1}\|^2 \\
& - \sum_{i=1}^{k} a_i(f(x^i) + \langle\tilde{\nabla}f(x^i), v^{k}-x^i\rangle + g(v^{k}))\frac{\beta_k}{2}\|x^0-v^{k}\|^2 \\
& + \frac{\beta_{k+1}-\beta_k}{8A_{k+1}}\bar{r}_{k+1}^2 + A_k\langle\tilde{\nabla}f(x^{k+1})-\nabla f(x^{k+1}), y^k-x^{k+1}\rangle \\
& + A_{k+1}\langle\nabla f(y^{k+1})-\tilde{\nabla}f(y^{k+1}), y^{k+1}-x^{k+1}\rangle.
\end{aligned} \tag{62}
$$

We can simplify (33) by using the definition of $\phi_k(\cdot)$:

$$
\begin{aligned}
A_{k+1}\psi(y^{k+1}) \leq & A_k\psi(y^k) + \phi_{k+1}(v^{k+1}) - \phi_k(v^k) \\
& + \frac{\beta_{k+1}-\beta_k}{8A_{k+1}}\bar{r}_{k+1}^2 + A_k\langle\tilde{\nabla}f(x^{k+1})-\nabla f(x^{k+1}), y^k-x^{k+1}\rangle \\
& + A_{k+1}\langle\nabla f(y^{k+1})-\tilde{\nabla}f(y^{k+1}), y^{k+1}-x^{k+1}\rangle.
\end{aligned} \tag{63}
$$

Applying the upper inequality recursively, it holds that

$$
\begin{aligned}
A_k\psi(y^k) \leq & \phi_k(v^k) - \phi_0(v^0) \\
& + \sum_{i=0}^{k-1} \frac{\beta_{i+1}-\beta_i}{8}\bar{r}_{i+1}^2 + A_i\langle\tilde{\nabla}f(x^{i+1})-\nabla f(x^{i+1}), y^i-x^{i+1}\rangle \\
& + A_{i+1}\langle\nabla f(y^{i+1})-\tilde{\nabla}f(y^{i+1}), y^{i+1}-x^{i+1}\rangle \\
\leq & \phi_k(v^k) + \frac{\beta_k}{8}\bar{r}_k^2 - \frac{\beta_0}{8}\bar{r}_1^2 + \sum_{i=0}^{k-1} A_i\langle\tilde{\nabla}f(x^{i+1})-\nabla f(x^{i+1}), y^i-x^{i+1}\rangle \\
& + A_{i+1}\langle\nabla f(y^{i+1})-\tilde{\nabla}f(y^{i+1}), y^{i+1}-x^{i+1}\rangle \\
\leq & \phi_k(v^k) + \frac{\beta_k}{8}\bar{r}_k^2 + \sum_{i=0}^{k-1} A_i\langle\tilde{\nabla}f(x^{i+1})-\nabla f(x^{i+1}), y^i-x^{i+1}\rangle \\
& + A_{i+1}\langle\nabla f(y^{i+1})-\tilde{\nabla}f(y^{i+1}), y^{i+1}-x^{i+1}\rangle,
\end{aligned}
$$

where $\phi_0(v^0)=0$ and $\beta_0=0$.

Since $v^k = \arg\min_x \phi_k(x)$, we apply Lemma 2 again and obtain that:

$$
\begin{aligned}
A_k\psi(y^k) \leq & \phi_k(v^k) + \frac{\beta_k}{8}\bar{r}_k^2 + \sum_{i=0}^{k-1} A_i\langle\tilde{\nabla}f(x^{i+1}) - \nabla f(x^{i+1}), y^i - x^{i+1}\rangle \\
& + A_{i+1}\langle\nabla f(y^{i+1}) - \tilde{\nabla}f(y^{i+1}), y^{i+1} - x^{i+1}\rangle \\
= & \sum_{i=1}^{k} a_i(f(x^i) + \langle\tilde{\nabla}f(x^i), v^k - x^i\rangle + g(v^k)) + \frac{\beta_k}{2}\|x^0 - v^k\|^2 + \frac{\beta_k}{8}\bar{r}_k^2 \\
& + \sum_{i=0}^{k-1} A_i\langle\tilde{\nabla}f(x^{i+1}) - \nabla f(x^{i+1}), y^i - x^{i+1}\rangle + A_{i+1}\langle\nabla f(y^{i+1}) - \tilde{\nabla}f(y^{i+1}), y^{i+1} - x^{i+1}\rangle \\
\leq & \sum_{i=1}^{k} a_i(f(x^i) + \langle\tilde{\nabla}f(x^i), x^* - x^i\rangle + g(x^*)) + \frac{\beta_k}{2}\|x^0 - x^*\|^2 - \frac{\beta_k}{2}\|v^{k+1} - x^*\|^2 \\
& + \frac{\beta_k}{8}\bar{r}_k^2 + \sum_{i=0}^{k-1} A_i\langle\tilde{\nabla}f(x^{i+1}) - \nabla f(x^{i+1}), y^i - x^{i+1}\rangle \\
& + A_{i+1}\langle\nabla f(y^{i+1}) - \tilde{\nabla}f(y^{i+1}), y^{i+1} - x^{i+1}\rangle \\
\leq & \sum_{i=1}^{k} a_i(f(x^i) + \langle\nabla f(x^i), x^* - x^i\rangle + g(x^*)) + \frac{\beta_k}{2}\|x^0 - x^*\|^2 - \frac{\beta_k}{2}\|v^{k+1} - x^*\|^2 \\
& + \frac{\beta_k}{8}\bar{r}_k^2 + \sum_{i=0}^{k-1} a_{i+1}\langle\tilde{\nabla}f(x^{i+1}) - \nabla f(x^{i+1}), v^i - x^*\rangle \\
& + A_{i+1}\langle\nabla f(y^{i+1}) - \tilde{\nabla}f(y^{i+1}), y^{i+1} - x^{i+1}\rangle \\
\leq & A_k\psi(x^*) + \frac{\beta_k}{2}\|x^0 - x^*\|^2 - \frac{\beta_k}{2}\|v^{k+1} - x^*\|^2 + \frac{\beta_k}{8}\bar{r}_k^2 \\
& + \sum_{i=0}^{k-1} a_{i+1}\langle\tilde{\nabla}f(x^{i+1}) - \nabla f(x^{i+1}), v^i - x^*\rangle \\
& + A_{i+1}\langle\nabla f(y^{i+1}) - \tilde{\nabla}f(y^{i+1}), y^{i+1} - x^{i+1}\rangle.
\end{aligned}
$$

We use D_0 and D_k to replace $\|x^0 - x^*\|$ and $\|v^k - x^*\|$. Since $D_0^2 - D_k^2 \leq 2D_0 r_k \leq 2D\bar{r}_k$, we have

$$
\begin{aligned}
A_k\psi(y^k) \leq & A_k\psi(x^*) + \frac{\beta_k}{2}D_0^2 - \frac{\beta_k}{2}D_k^2 + \frac{\beta_k}{8}\bar{r}_k^2 \\
& + \sum_{i=0}^{k-1} a_{i+1}\langle\tilde{\nabla}f(x^{i+1}) - \nabla f(x^{i+1}), v^i - x^*\rangle \\
& + A_{i+1}\langle\nabla f(y^{i+1}) - \tilde{\nabla}f(y^{i+1}), y^{i+1} - x^{i+1}\rangle,
\end{aligned}
$$

which implies

$$
\begin{aligned}
\psi(y^k) - \psi(x^*) \leq & \frac{\beta_k(D_0^2 - D_k^2)}{2A_k} + \frac{\beta_k\bar{r}_k^2}{8A_k} + \sum_{i=0}^{k-1} a_{i+1}\langle\tilde{\nabla}f(x^{i+1}) - \nabla f(x^{i+1}), v^i - x^*\rangle \\
& + A_{i+1}\langle\nabla f(y^{i+1}) - \tilde{\nabla}f(y^{i+1}), y^{i+1} - x^{i+1}\rangle \\
\leq & \frac{9\beta_k D\bar{r}_k}{8A_k} + \sum_{i=0}^{k-1} a_{i+1}\langle\tilde{\nabla}f(x^{i+1}) - \nabla f(x^{i+1}), v^i - x^*\rangle \\
& + A_{i+1}\langle\nabla f(y^{i+1}) - \tilde{\nabla}f(y^{i+1}), y^{i+1} - x^{i+1}\rangle.
\end{aligned} \tag{64}
$$

We take the expectations of both sides and obtain

$$\begin{aligned}\mathbb{E}[\psi(y^k)] - E[\psi(x^*)] \leq & \mathbb{E}[\frac{9\beta_k D\bar{r}_k}{8A_k}] + \mathbb{E}[\sum_{i=0}^{k-1} a_{i+1}\langle\tilde{\nabla}f(x^{i+1}) - \nabla f(x^{i+1}), v^i - x^*\rangle] \\ & + \mathbb{E}[\sum_{i=0}^{k-1} A_{i+1}\langle\nabla f(y^{i+1}) - \tilde{\nabla}f(y^{i+1}), y^{i+1} - x^{i+1}\rangle].\end{aligned}$$

Since $\mathbb{E}[\tilde{\nabla}f(x^{i+1}) - \nabla f(x^{i+1})] = 0$ and v^i, x^*, ξ_{i+1} are independent, we have

$$\begin{aligned}&\mathbb{E}\left[\sum_{i=0}^{k-1} a_{i+1}\langle\tilde{\nabla}f(x^{i+1}) - \nabla f(x^{i+1}), v^i - x^*\rangle\right] \\ =&\mathbb{E}_{-\xi_{i+1}^x}\left[\mathbb{E}_{\xi_{i+1}^x}\left[\sum_{i=0}^{k-1} a_{i+1}\langle\tilde{\nabla}f(x^{i+1}) - \nabla f(x^{i+1}), v^i - x^*\rangle\right]\right] = 0.\end{aligned}$$

For the same reason, we have

$$\begin{aligned}&\mathbb{E}\left[\sum_{i=0}^{k-1} A_{i+1}\langle\nabla f(y^{i+1}) - \tilde{\nabla}f(y^{i+1}), y^{i+1} - x^{i+1}\rangle\right] \\ =&\mathbb{E}_{-\xi_{i+1}^y}\left[\mathbb{E}_{\xi_{i+1}^y}\left[\sum_{i=0}^{k-1} A_{i+1}\langle\nabla f(y^{i+1}) - \tilde{\nabla}f(y^{i+1}), y^{i+1} - x^{i+1}\rangle\right]\right] = 0.\end{aligned}$$

Thus

$$\mathbb{E}[\psi(y^k)] - \psi(x^*) \leq \frac{9}{8}\mathbb{E}\left[\frac{\beta_k D\bar{r}_k}{A_k}\right].$$

We will apply Lemma 11 and 1 to obtain the final complexity. Note that $k^* = \underset{0\leq i\leq k}{\arg\min} \frac{\bar{r}_k}{\sum_{i=0}^{k-1}\bar{r}_i}$.

Applying Lemma 1, we obtain that:

$$\frac{\bar{r}_{k^*}^{\frac{1}{2}}}{\sum_{i=0}^{k^*-1}\bar{r}_i^{\frac{1}{2}}} \leq \frac{(\frac{\bar{r}_k}{\bar{r}_0})^{\frac{1}{2}\times\frac{1}{k}}\log e(\frac{\bar{r}_k}{\bar{r}_0})^{\frac{1}{2}}}{k} \leq \frac{(\frac{4D_0}{\bar{r}})^{\frac{1}{2k}}\log e\frac{4D_0}{\bar{r}}}{2k}. \tag{65}$$

Thus, for k^*, combining the inequality (65) and Lemma 11, it holds that

$$\begin{aligned}&\mathbb{E}\big[\psi(y^{k^*})\big] - \psi(x^*) \\ &\leq \frac{9}{8}\mathbb{E}\left[\frac{\beta_{k^*}D\bar{r}_{k^*}}{A_{k^*}}\right] \\ &\leq \frac{9}{8}\mathbb{E}\left[\beta_k D\left(\frac{\bar{r}_{k^*}^{\frac{1}{2}}}{\sum_{i=0}^{k^*-1}\bar{r}_i^{\frac{1}{2}}}\right)^2\right] \\ &\leq \frac{9}{8}\mathbb{E}[(2^{\frac{9+9\nu}{2}}\tilde{M}_\nu D^{1+\nu}k^{\frac{3-3\nu}{2}} + 2^5 k^{\frac{3}{2}}D\sigma)\left(\frac{(\frac{4D}{\bar{r}})^{\frac{1}{2k}}\log e\frac{4D}{\bar{r}}}{2k}\right)^2] \\ &\leq 36(\frac{4D}{\bar{r}})^{\frac{1}{k}}\log^2 e\frac{4D}{\bar{r}}\frac{\tilde{M}_\nu D^{1+\nu}k^{\frac{3-3\nu}{2}} + k^{\frac{3}{2}}D\sigma}{k^2} \\ &\leq 36(\frac{4D}{\bar{r}})^{\frac{1}{k}}\log^2 e\frac{4D}{\bar{r}}(\frac{\tilde{M}_\nu D^{1+\nu}}{k^{\frac{1+3\nu}{2}}} + \frac{D\sigma}{\sqrt{k}}).\end{aligned} \tag{66}$$

where we use the facts $\{\beta_i\}_{i\in\mathbb{N}}$ is nondecreasing and the random variable B is independent of both k and D.

Finally, it remains to use $z^k = y^{k^*}$ by definition. □

The following lemma is used to show that the balance equation admits a closed-form solution.

Lemma 12. *Let β,l,d ≥ 0 and $r > 0$. Then the equation*

$$(\beta_+ - \beta)r = [l - \beta_+ d]_+ \tag{67}$$

has a unique solution given by

$$\beta_+ = \beta + \frac{[l - \beta d]_+}{r + d}. \tag{68}$$

This auxiliary result has been used in [[30], Lemma E.1]. We give proof for completeness.

Proof. First, the equation has a unique solution since the left-hand side increases from zero to infinity monotonically with respect to β_+ while the right-hand side decreases from a nonnegative number to zero monotonically with respect to β_+.

We show that the β_+ in (68) is the very solution of (67).

When $l - \beta d \leq 0$, $\beta_+ = \beta$. $LHS = (\beta_+ - \beta)r = 0$ and $RHS = [l - \beta_+ d]_+ = [l - \beta d]_+ \leq 0$, which implies $RHS = 0$. Therefore, $LHS = RHS$ and β_+ is the solution.

When $l - \beta d > 0$, $\beta_+ = \beta + \frac{l-\beta d}{r+d}$. then $LHS = (\beta_+ - \beta)r = (\beta + [\frac{l-\beta d}{r+d}]_+ - \beta)r = r\frac{l-\beta d}{r+d}$ and $RHS = [l - \beta_+ d]_+ = [l - \beta d - \frac{l-\beta d}{r+d} d]_+ = [\frac{r}{r+d}(l - \beta d)]_+ = \frac{r}{r+d}(l - \beta d)$. Therefore, $LHS = RHS$ and β_+ is the solution as well. □

E Two approaches for automatic initialization of parameters in Algorithm 1

E.1 Automatic initialization of β_0

In the proof of Proposition 1 and Theorem 2, we assume that $\beta_0 \leq 2^7 I_\nu \hat{M}_\nu \bar{r}^\nu$. This is reasonable since the upper bound of β_k increases polynomial in k, it still holds for enough large k. Previous works often ignore the choice of a legal β_0. Nevertheless, we provide a simple method for choosing an admissible β_0.

Algorithm 3 β_0 Initialization Method

Input: $x^0, \bar{r}$, any other point $x' \in \mathbb{R}^d$ that satisfies $\|x' - x^0\| \leq \bar{r}$ and $f(x') - f(x^0) - \langle \nabla f(x^0), x' - x^0 \rangle > 0$;

Output: $\bar{\beta} \leq 2^7 I_\nu \hat{M}_\nu \bar{r}^\nu$;

1: Set $c = \min\{[f(x') - f(x^0) - \langle \nabla f(x^0), x^0 - x' \rangle]/\bar{r}^2, \frac{1}{2}\}$;
and $M = 2\frac{f(x') - f(x^0) - \langle \nabla f(x^0), x^0 - x' \rangle - c\bar{r}^2/2}{\|x^0 - x'\|^2}$;

2: Let $\bar{\beta} = \bar{r} \max\{8\sqrt{2M}, 128M\} \min\{1, \sqrt{c}\}$;

Proposition 3. *Suppose $f(\cdot)$ is locally Hölder smooth and $f(\cdot)$ is not a linear function in* $\operatorname{dom} g$. *If $\bar{r}$ is small enough such that $\bar{r} \leq 4D_0$, then Algorithm 3 can generate a β_0 that satisfies*

$$\beta_0 \leq 2^7 I_\nu \hat{M}_\nu \bar{r}^v, \tag{69}$$

and this method can be implemented in one operation.

Proof. Since

$$M = 2\frac{f(x') - f(x^0) - \langle \nabla f(x^0), x^0 - x' \rangle - c\frac{\bar{r}^2}{2}}{\|x^0 - x'\|^2} \geq 0,$$

we have

$$f(x') = f(x^0) + \langle \nabla f(x^0), x^0 - x' \rangle + \frac{M}{2}\|x^0 - x'\|^2 + c\frac{\bar{r}^2}{2}. \tag{70}$$

The equation (70) is tight and $x^0, x' \in \mathcal{B}_{3D_0}(x^*)$, so that we have

$$M \leq \gamma(\hat{M}_\nu, c\bar{r}^2) = (\frac{1-\nu}{1+\nu}\frac{1}{c\bar{r}^2})^{\frac{1-\nu}{1+\nu}} \hat{M}_\nu^{\frac{2}{1+\nu}}. \tag{71}$$

$$c^{\frac{1}{2}}\bar{r}\min\{(128\bar{M})^{\frac{1}{2}},128\bar{M}\}\leq c^{\frac{1-\nu}{2}}\bar{r}(128\bar{M})^{\frac{1+\nu}{2}}=128(\frac{1-\nu}{1+\nu})^{\frac{1-\nu}{2}}\hat{M}_\nu\bar{r}^v. \tag{72}$$

Thus

$$c^{\frac{1}{2}}\bar{r}\min\{(128\bar{M})^{\frac{1}{2}},128\bar{M}\}\leq 2^7 I_\nu\hat{M}_\nu\bar{r}^v. \tag{73}$$

So we can initialize $\beta_0=c^{\frac{1}{2}}\bar{r}\min\{(128\bar{M})^{\frac{1}{2}},128\bar{M}\}$. □

E.2 Automatic initialization of $\bar{r}$

As mentioned in the context, Algorithm 1 requires an input $\bar{r}$ as a guess of $4D_0$ that satisfies $\bar{r}\leq 4D_0$. Setting a small enough $\bar{r}$ to meet $\bar{r}\leq 4D_0$ will incur a multiplicative cost of $(\frac{4D_0}{\bar{r}})^{1+\nu}$ in the convergence rate for other algorithms without distance adaptation. In contrast, we reduce the multiplicative cost to $\log^2(\frac{4D_0}{\bar{r}})$. Moreover, we provide a simple method for obtaining an admissible $\bar{r}\leq 4D_0$ in some special cases.

Proposition 4 can handle the cases where we can choose x^0 and make sure $f(\cdot)+g(\cdot)$ is weakly smooth on one of its neighborhoods. Then we can modify the problem by setting $f'(x)=f(x)+g(x)$ and $g'(x)=0$ to get a legal $\bar{r}$.

Algorithm 4 $\bar{r}$ Initialization Method

Input: x^0 and an initial guess r for $4D_0$
Output: $\bar{r}\leq 4D_0$

```
Initialize i ← −1
repeat
    i ← i + 1
    d_i ← 2^{−i} r
    Run Algorithm 1 with parameter d by one iteration and collect the point v_i^1 and the coefficient β_{1,i}
    Denote r_{1,i} = ‖v_i^1 − x^0‖
until v_i^1 is an interior point in dom g and r_{1,i} ≥ d_i
Set r̄ = d_i
```

Proposition 4. *For any $x\in\operatorname{dom} g$, $g(x)=0$, if there exists $\delta>0$, $S=\mathcal{B}_\delta(x^0)\subset\operatorname{dom} g$, and let ν_S be the maximal Hölder exponent of $f(\cdot)$ on S with finite local Hölder continuous constant $\hat{M}_{\nu_S}<+\infty$, $\nu_S>0$, then Algorithm 4 can generate $\bar{r}\leq 4D_0$, and this method can be implemented in a finite number of iterations.*

Proof. Without loss of generality, we assume $\delta\leq 3D_0$, since if $\delta>3D_0$, we can always take a smaller $\delta'\leq 3D_0$ that still satisfies the condition.

Note that Theorem 1 implies that if $\bar{r}_{1,i}=r_{1,i}$, then $\bar{r}\leq r_{1,i}\leq 4D_0$, and this conclusion is independent of the value of β_1. So we only need to prove that this method completes in a finite number of operations. Note that $x^1=\tau_0 v^0+(1-\tau_0)y^0=\tau_0 x^0+(1-\tau_0)x^0=x^0$ does not depend on the value of τ_0. The condition of this method ensures that there exists an interior point in the direction of $-\nabla f(x^0)$.

Applying Lemma 5, it holds that

$$\|v_i'-x^0\|\geq\|v_i^1-x^0\|,$$

where we define

$$\begin{aligned} v_i'&=\arg\min_{x\in\operatorname{dom} g}(d_i\langle\nabla f(x^1),x-x^1\rangle+\frac{\beta_0\|x-x^0\|^2}{2}\\ &=\arg\min_{x\in\operatorname{dom} g}\langle\nabla f(x^0),x-x^0\rangle+\frac{\beta_0\|x-x^0\|^2}{2d_i}. \end{aligned} \tag{74}$$

Applying Lemma 5 again with $h_i = \frac{\beta_0}{2d_i}$, we have $\|v_i' - x^0\| \to 0$ as $i \to +\infty$. Therefore, there exists large enough i^* such that $\forall i \geq i^*$,it satisfies that

$$\delta \geq \|v_i' - x^0\| \geq \|v_i^1 - x^0\|, \tag{75}$$

Case 1: If this method requires at most i^* operations. Then we prove it directly.

Case 2: If this method requires more than i^* operations.

Inequality (75) show that $v_i^1 \in \mathcal{B}_\delta(x^0) \subset \operatorname{dom} g$, which means v_i^1 is an interior point. Then, applying the first order condition to the interior point v_i^1, we have:

$$\begin{aligned} d_i \nabla f(x^0) + \beta_{1,i}(v_i^1 - x^0) &= 0 \\ v_i^1 - x^0 &= d_i \frac{\nabla f(x^0)}{\beta_{1,i}}. \end{aligned} \tag{76}$$

We can adapt the method 3 to obtain a legal β_0 to produce $\beta_1 \leq c_\nu d_i^\nu$, which is guaranteed by Lemma 8, where $c_\nu = 2^7 I_\nu \hat{M}_\nu$.

$$\begin{aligned} r_{1,i} =& \|v_i^1 - x^0\| = d_i \frac{\|\nabla f(x^0)\|}{\beta_1} \\ \geq& d_i^{1-\nu} \frac{\|\nabla f(x^0)\|}{c_\nu}. \end{aligned} \tag{77}$$

Thus $r_{1,i} = d_i^{1-\nu} \frac{\|\nabla f(x^0)\|}{c_\nu} \geq d_i$ when $2^{-i} r = d_i \leq \left(\frac{\|\nabla f(x^0)\|}{c_\nu}\right)^{\frac{1}{v}}$, $v \neq 0$ and this method require at most $-\frac{1}{\nu} \log\left(\frac{\|\nabla f(x^0)\|}{c_\nu}\right) / \log 2$ loops. □

F More experiment details

In this section, we provide more details about the experiments.

Softmax problem We first reexamine the softmax problem with more parameter settings. Specifically, we set $\mu \in \{0.1, 0.01, 0.001\}$. In all the results, we find that our AGDA consistently performs better than the other compared methods.

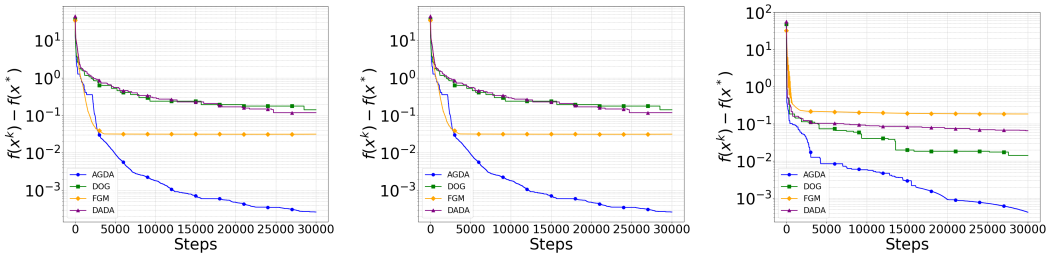

Figure 3: Performance of the compared algorithms on the softmax problem. From left to right: $\mu = 0.1$, $\mu = 0.01$ and $\mu = 0.001$.

L_p norm problem We consider the following problem as an illustrative example, where the smoothness property can be directly adjusted by modifying a parameter $p \in [1, 2]$:

$$\min_{x \in \mathbb{R}^d} f(x) = \|Ax - b\|_p, \tag{78}$$

where $A \in \mathbb{R}^{n \times d}, b \in \mathbb{R}^n$ are taken from real-world datasets in LIBSVM. It is important to note that the smoothness of this problem can be controlled by changing the parameter p; as p increases, the degree of smoothness decreases.

We use the same comparison methods as in the softmax problem, adhering to the same parameter settings. In this problem, our algorithm significantly outperforms the other methods, especially when p is small. This suggests that our algorithm is more adaptive in nonsmooth settings and highlights its greater stability.

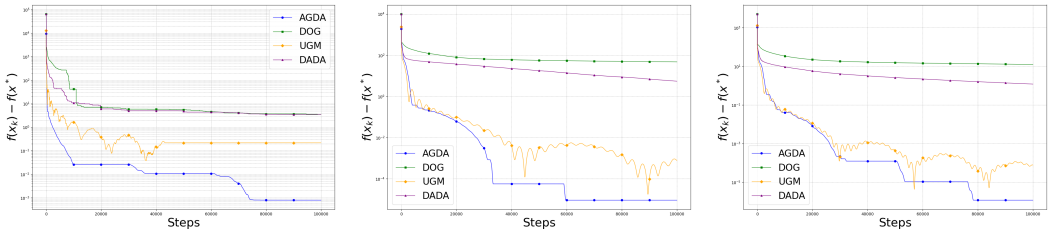

Figure 4: Performance of the compared algorithms on L_p norm problem. Left: $p = 1$ with diabetes. Middle: $p = 1.5$ with `boston`. Right: $p = 2$ with `boston`.

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: We state the complete contributions in Section 1.

   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 work in Appendix A.

   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: We give all assumptions needed in Section 2, 3 and 4. We leave all the complete proofs in Appendix.

   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: [Yes]

   Justification: Our experimental reproduction scripts will been placed in the supplementary material.

   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: [Yes]

   Justification: Our experiments use publicly available datasets or randomly generated data based on specific methodologies. We are committed to making our code completely open source.

   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: [Yes]

   Justification: All details can be found in the paper and supplemental material.

   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: [No]

Justification: Our experimental results do not report standard deviation correlation results in the deterministic case. We focus on the convergence rate of algorithm in the stochastic case and do not report them too.

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: [Yes]

   Justification: All experiments were conducted on personal computers, utilizing CPUs for computations, with 16GB of RAM.

   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: The code in submission is fully compliant with the NeurIPS code of Ethics.

   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: This paper focuses on a theoretical problem without any societal impact.

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: There are no such risks in this paper.

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: [Yes]

Justification: The creators or original owners of assets are properly credited, and the license and terms of use are explicitly mentioned and respected in the paper.

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: [Yes]

Justification: We provide a complete document of our code.

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: This paper does not involve such research.

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: This paper does not involve such research.

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: This paper does not involve any LLMs.

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.