

Article

A Chebyshev–Halley Method with Gradient Regularization and an Improved Convergence Rate

Jianyu Xiao ¹, Haibin Zhang ¹  and Huan Gao ^{2,*}¹ Beijing Institute for Scientific and Engineering Computing, Beijing University of Technology, Beijing 100124, China; xiaoxiao@emails.bjut.edu.cn (J.X.); zhanghaibin@bjut.edu.cn (H.Z.)² College of Mathematics and Computational Science, Hunan First Normal University, Changsha 410205, China

* Correspondence: sjyhgaohnfnu.edu.cn

Abstract: High-order methods are particularly crucial for achieving highly accurate solutions or satisfying high-order optimality conditions. However, most existing high-order methods require solving complex high-order Taylor polynomial models, which pose significant computational challenges. In this paper, we propose a Chebyshev–Halley method with gradient regularization, which retains the convergence advantages of high-order methods while effectively addressing computational challenges in polynomial model solving. The proposed method incorporates a quadratic regularization term with an adaptive parameter proportional to a certain power of the gradient norm, thereby ensuring a closed-form solution at each iteration. In theory, the method achieves a global convergence rate of $O(k^{-3})$ or even $O(k^{-5})$, attaining the optimal rate of third-order methods without requiring additional acceleration techniques. Moreover, it maintains local superlinear convergence for strongly convex functions. Numerical experiments demonstrate that the proposed method compares favorably with similar methods in terms of efficiency and applicability.

Keywords: Chebyshev–Halley method; quadratic regularization; gradient norm; closed-form solution; global convergence; local superlinear convergence

MSC: 49M37; 90C25; 90C30; 68Q25



Received: 12 March 2025

Revised: 12 April 2025

Accepted: 15 April 2025

Published: 17 April 2025

Citation: Xiao, J.; Zhang, H.; Gao, H. A Chebyshev–Halley Method with Gradient Regularization and an Improved Convergence Rate. *Mathematics* **2025**, *13*, 1319. <https://doi.org/10.3390/math13081319>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In this paper, we consider the following unconstrained convex optimization problem:

$$\min_{x \in \mathbb{R}^n} f(x) \quad (1)$$

where $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is a continuously differentiable function with a globally Lipschitz continuous third-order derivative.

From a theoretical standpoint, high-order optimization methods exploit p ($p \geq 3$)th-order derivative information to attain enhanced convergence rates compared to first- and second-order methods. Furthermore, in the context of non-convex optimization, these methods exhibit a distinct advantage of circumventing degenerate solutions induced by singular Hessian matrices, thereby yielding solutions with higher precision and superior quality [1,2]. The modern development of high-order methods can be traced back to the cubic regularization of Newton’s method (CNM) [3], which laid the foundation for subsequent extensions to p th-order tensor methods [4–8], as well as adaptive [9–12] and

stochastic [13–16] variants. In the aforementioned studies, a p th-order Taylor polynomial model is employed to approximate the objective function $f(x)$ at x_k :

$$\Omega_{x_k, p, \lambda_k}(s) = f(x_k) + \sum_{i=1}^p \frac{1}{i!} \nabla^i f(x_k)[s]^i + \frac{\lambda_k}{(p+1)!} \|s\|^{p+1},$$

where $\nabla^i f(x_k) \in \mathbb{R}^{n^i}$ is the i th-order derivative and $\nabla^i f(x_k)[s]^i \in \mathbb{R}$ is the i th-order derivative along $s \in \mathbb{R}^n$. The next iterate is updated as $x_{k+1} = x_k + s_k$, where $s_k = \arg \min_{s \in \mathbb{R}^n} \Omega_{x_k, p, \lambda_k}(s)$, and it continues until a satisfactory solution is found.

The non-convex nature of the polynomial model $\Omega_{x_k, p, \lambda_k}(s)$ presents a significant challenge for these methods in achieving minimization. Nesterov [17] demonstrated that a sufficiently large regularization parameter λ_k can ensure the convexity of the model $\Omega_{x_k, p, \lambda_k}(s)$ while preserving the theoretical advantages of local high-order approximations, particularly in settings where the i th-order derivative is Lipschitz continuous. Furthermore, Nesterov developed an efficient minimization scheme for a convex polynomial model with $p = 3$, incorporating a linearly convergent inner iteration at each step. The scheme leverages automatic differentiation techniques [18,19] to compute third-order derivatives along specific directions, significantly improving computational efficiency. Doikov et al. [20] proposed an inexact strategy for tensor methods, which initially reduced the accuracy of the model's solution and gradually improved it in later iterations, thus ensuring global convergence. Cartis et al. [21] utilized the cubic-quartic regularization (CQR) framework to successfully address the minimization of a non-convex cubic model with quartic regularization. Ahmadi et al. [22] introduced sum-of-squares (SoS) convex polynomials as an alternative to address the limitations of the Taylor model in intuitive representation, and they solved the resulting model via n -dimensional semidefinite programming. Subsequently, ref. [23] integrated SoS into an adaptive framework, extending its applicability to general non-convex problems. Xiao et al. [24] proposed a regularized Chebyshev–Halley method (RCH) for solving a third-order polynomial model, achieving a global convergence rate of $O(k^{-3})$, which can be further accelerated to $O(k^{-5})$. The method transforms the solution of a polynomial model into the simultaneous resolution of a regularized Newton equation and a Halley equation, while maintaining computational complexity comparable to that of second-order methods.

Another active line of research aims to improve convergence rates through acceleration techniques. In machine learning and deep learning applications, for example, efficient methods not only accelerate model convergence but also significantly reduce computational resource requirements. In this context, two primary directions have been explored. The first is Nesterov-type acceleration [17,25–27], which leverages the estimate sequence technique. However, this technique relies on intricate algebraic constructions, leading to a lack of a unified mathematical explanation for its acceleration mechanism. Furthermore, experimental studies have indicated that accelerated tensor methods may underperform in practice [28]. The second direction is based on the accelerated hybrid proximal extragradient (A-HPE) framework [29–32], which achieves an optimal convergence rate of $O(k^{-\frac{3p+1}{2}})$ for p th-order optimization methods. Despite its theoretical advantages, this framework requires multiple iterative sequences, additional hyperparameters, and a time-consuming bisection line search process that is sensitive to parameter choices, making it less suitable for high-dimensional problems.

In practical scenarios, researchers tend to prefer fundamental methods due to their simplicity, ease of implementation, and generalizability to various optimization problems.

Polyak [33] introduced a quadratic regularization technique in his work and proposed the following iterative scheme:

$$x_{k+1} = x_k - (\nabla^2 f(x_k) + \lambda_k I)^{-1} \nabla f(x_k),$$

where $\lambda_k \propto \|\nabla f(x_k)\|$. This iteration is concise and admits a closed-form solution at each step, circumventing the need to solve the polynomial model $\Omega_{x_k, 2, \lambda_k}(s)$, but it only achieves the $O(k^{-\frac{1}{4}})$ global convergence rate. Ueda et al. [34] employed a similar technique and reconstructed the quadratic Taylor polynomial model as follows:

$$\Omega_{x_k, 2, \lambda_k}^2(s) = f(x_k) + \sum_{i=1}^p \frac{1}{i!} \nabla^i f(x_k)[s]^i + \frac{\lambda_k}{2} \|s\|^2.$$

Although its minimization leads to the same iterative scheme, the global convergence rate remains suboptimal. The aforementioned studies suggest that the quadratic regularization strategy may establish a balance between computational efficiency and global convergence, offering a new pathway for developing fundamental methods.

Recently, Doikov et al. [35] combined quadratic regularization with the Bregman distance, designing a regularization parameter $\lambda_k \propto \sqrt{\|\nabla f(x_k)\|}$. The resulting regularized Newton method can be viewed as a generalized relaxation of CNM, which theoretically not only preserves the global convergence rate of $O(k^{-2})$ but also significantly reduces the computational complexity through closed-form iterations. Mishchenko [36] explored the relationship between the Levenberg–Marquardt penalty and CNM, constructing a quadratic regularization model. Despite differences in theoretical analysis, the proposed method achieves the same $O(k^{-2})$ convergence rate and also admits a closed-form solution. Building on this, ref. [37] further unified quadratic and cubic regularization forms while introducing novel regularization terms such as the elastic net. Refs. [38–41] have also explored simple methods for more general or non-convex problems. Meanwhile, significant progress has been made in fundamental methods based on stepsize adjustment. Hanzely et al. [42,43] designed a novel damping parameter update scheme for the damped Newton method by revealing the implicit duality between the Newton stepsize and Levenberg–Marquardt regularization. This scheme updates the stepsize using the product of the gradient and the Hessian matrix, achieving a global convergence rate of $O(k^{-2})$ or $O(k^{-3})$.

In this paper, we propose a Chebyshev–Halley method with gradient regularization, a third-order method that combines rapid convergence with iterative simplicity. The proposed method employs a quadratic regularization strategy, where the regularization parameter is adjusted based on a power of the gradient norm at the current iterate. Therefore, the iterate is updated directly through matrix inversion and vector multiplication, which removes the need for nested inner iterations to solve the polynomial model. Theoretically, the method can achieve a global convergence rate of $O(k^{-3})$ or $O(k^{-5})$ without requiring acceleration techniques, attaining the optimal convergence rate for third-order methods. Numerical experiments demonstrate its effectiveness and applicability across various scenarios.

This paper is organized as follows. Section 2 presents the assumptions and the algorithmic framework of the proposed method. Section 3 focuses on the convergence analysis. Section 4 provides numerical experiments to validate the effectiveness of the method. Section 5 summarizes this paper.

2. Gradient-Regularized Chebyshev–Halley Method

For $x, s \in \mathbb{R}^n$, define $\nabla f(x)[s] = \nabla f(x)^T s$, $\nabla^2 f(x)[s]^2 = s^T \nabla^2 f(x) s$ and $\nabla^3 f(x)[s]^3 = s^T \nabla^3 f(x) s s$, where $\nabla f(x_k)$, $\nabla^2 f(x_k)$ and $\nabla^3 f(x_k)$ represent the gradient, Hessian matrix, and third-order derivative, respectively. The objective function $f(x)$ is assumed to satisfy the following standard conditions.

Assumption 1. The objective function $f(x)$ is convex and differentiable on $\mathbb{R}^n$, with a Lipschitz continuous third-order derivative. Specifically, there exists a constant $0 < L_3 < \infty$ such that for any $x, y \in \mathbb{R}^n$,

$$\|\nabla^3 f(x) - \nabla^3 f(y)\| \leq L_3 \|x - y\|.$$

As a direct consequence, for any $x, s \in \mathbb{R}^n$, the following bounds hold:

$$\left\| \nabla f(x+s) - \nabla f(x_k) - \nabla^2 f(x_k)s - \frac{1}{2} \nabla^3 f(x_k)ss \right\| \leq \frac{L_3}{6} \|s\|^3, \quad (2)$$

$$\left\| \nabla^2 f(x+s) - \nabla^2 f(x_k) - \nabla^3 f(x_k)s \right\| \leq \frac{L_3}{2} \|s\|^2. \quad (3)$$

These bounds in (2) and (3) quantify the approximation errors of the gradient and Hessian, respectively, and play a crucial role in analyzing the convergence behavior of the proposed method.

Assumption 2. The level set $\mathcal{L}(x_0) := \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$ is bounded, satisfying

$$\max_{x, y \in \mathcal{L}(x_0)} \|x - y\| \leq D < +\infty.$$

The classical Chebyshev–Halley method, also known as the improved hyperbolic tangent method, is iteratively formatted as [44] $x_{k+1} = x_k + s_k^1 + s_k^2$, where s_k^1, s_k^2 are the solutions to the Newton equation

$$\nabla^2 f(x_k)s_k^1 = -\nabla f(x_k),$$

and the Halley equation

$$\nabla^2 f(x_k)s_k^2 = -\frac{1}{2} \nabla^3 f(x_k)s_k^1 s_k^1,$$

respectively. To improve the global convergence of the method, ref. [24] approximate $f(x)$ using a third-order Taylor polynomial model with a quartic regularization term and incorporate an acceleration sequence to achieve a global convergence rate of $O(k^{-5})$. However, this method requires nested inner iterations to solve the model, and the introduction of an acceleration sequence further complicates parameter tuning.

In this section, we consider the following third-order Taylor polynomial model with quadratic regularization:

$$s_k = \arg \min_{s \in \mathbb{R}^n} \Omega_{x_k, 3, \lambda_k}^2(s) \quad (4)$$

$$\triangleq \arg \min_{s \in \mathbb{R}^n} \left\{ f(x_k) + \nabla f(x_k)^T s + \frac{1}{2} s^T \nabla^2 f(x_k) s + \frac{1}{6} s^T \nabla^3 f(x_k) s s + \frac{\lambda_k}{2} \|s\|^2 \right\}, \quad (5)$$

where $x_k \in \mathbb{R}^n$, $\lambda_k \geq L_3$ is a regularization parameter proportional to a certain power of the gradient norm at x_k . The basic iterative framework of the algorithm is presented below.

Algorithm 1 incorporates a quadratic regularization term and third-order derivatives to enhance convergence properties. The quadratic regularization term ensures global convergence while also providing a closed-form solution, thereby avoiding the need for

nested inner iterations to solve the high-order Taylor polynomial model. Simultaneously, the utilization of the third-order derivative accelerates convergence, and the computational cost of third-order derivatives is comparable to that of Newton's method [45].

Algorithm 1 Chebyshev–Halley method with gradient regularization (CHGR).

Require: Initial point $x_0 \in \mathbb{R}^n$, $\alpha \in [\frac{2}{3}, 1]$, $\epsilon > 0$ and $H > 0$.

- 1: **for** $k = 0, 1, 2, \dots$ **do**
- 2: Set $\lambda_k \geq H \|\nabla f(x_k)\|^\alpha$.
- 3: Obtain the direction s_k through the following two steps:

$$(\nabla^2 f(x_k) + \lambda_k I) s_k^1 = -\nabla f(x_k), \quad (6)$$

$$(\nabla^2 f(x_k) + \lambda_k I) s_k = -\nabla f(x_k) - \frac{1}{2} \nabla^3 f(x_k) s_k^1 s_k^1. \quad (7)$$

- 4: $x_{k+1} = x_k + s_k$.

- 5: **end for**

We now derive an upper bound for the regularized Newton direction.

Corollary 1. Suppose that Assumption 1 holds. Thus,

$$\|s_k^1\| = \left\| \frac{\nabla f(x_k)}{\nabla^2 f(x_k) + \lambda_k I} \right\| \leq \frac{\|\nabla f(x_k)\|}{\lambda_k} \leq \frac{\|\nabla f(x_k)\|^{1-\alpha}}{H}. \quad (8)$$

In particular, if $H \geq \|\nabla f(x_k)\|^{1-\alpha}$, then

$$\|s_k^1\| \leq 1. \quad (9)$$

Define $s_k^2 = s_k - s_k^1$. From (6) and (7), it follows that

$$\|s_k^2\| = \frac{1}{2} \|\nabla^3 f(x_k) s_k^1 s_k^1\| \leq \frac{L_3}{2} \|s_k^1\|^2.$$

Furthermore, assuming that there exists a constant $C > 0$ such that $C \|s_k^1\| \leq \|s_k\|$, then, combining (9) and the norm triangle inequality, we obtain

$$C \|s_k^1\| \leq \|s_k\| \leq \|s_k^1\| + \|s_k^2\| \leq \|s_k^1\| + \frac{L_3}{2} \|s_k^1\|^2 \leq \left(1 + \frac{L_3}{2}\right) \|s_k^1\|. \quad (10)$$

Thus, $\|s_k\|$ and $\|s_k^1\|$ are of the same order of magnitude, meaning that $\|s_k\|$ is primarily determined by $\|s_k^1\|$, while $\|s_k^2\|$ serves only as a secondary correction term, proportional to the square of $\|s_k^1\|$. As $\|s_k^1\|$ approaches zero, the effect of $\|s_k^2\|$ diminishes, making its contribution to the overall stepsize negligible.

3. Convergence Analysis

This section is dedicated to analyzing the convergence of the proposed method. We begin by estimating the bound for the third-order derivative $\nabla^3 f(x)$ along the direction $h \in \mathbb{R}^n$.

Lemma 1. Suppose that Assumption 1 holds. For any $x, h \in \mathbb{R}^n$ and $\tau > 0$, we have

$$\|\nabla^3 f(x) h h\| \leq \sqrt{2L_3} \|h\| \sqrt{h^T \nabla^2 f(x) h}. \quad (11)$$

Proof. For any $x, y, h \in \mathbb{R}^n$, it follows from (3) that

$$0 \leq h^T \nabla^2 f(y) h \leq h^T \nabla^2 f(x) h + (y - x)^T \nabla^3 f(x) h h + \frac{L_3}{2} \|y - x\|^2 \|h\|^2.$$

Let $y = x - \tau u$, where $\tau > 0$ and $u \in \mathbb{R}^n$ with $\|u\| = 1$. Then,

$$(\tau u)^T \nabla^3 f(x) h h \leq h^T \nabla^2 f(x) h + \frac{L_3}{2} \|\tau u\|^2 \|h\|^2.$$

By dividing both sides of the inequality by τ and taking the norm, we obtain

$$\left\| \nabla^3 f(x) h h \right\| \leq \frac{1}{\tau} h^T \nabla^2 f(x) h + \tau \frac{L_3}{2} \|h\|^2.$$

By minimizing the right-hand side of the inequality with respect to τ , we obtain the optimal parameter $\tau_* = \sqrt{\frac{2h^T \nabla^2 f(x) h}{L_3 \|h\|^2}}$. Substituting τ_* into the inequality above yields the upper bound (11). $\square$

Lemma 2. Suppose that Assumption 1 holds. For any $s_k \in \mathbb{R}^n$, we have

$$s_k^T \nabla^2 f(x_k) s_k \leq \frac{3 \|\nabla f(x_k)\|^2}{16 \lambda_k}. \quad (12)$$

Proof. From (7), we have

$$\begin{aligned} s_k^T \nabla^2 f(x_k) s_k &= s_k^T \left(-\nabla f(x_k) - \lambda_k s_k - \frac{1}{2} \nabla^3 f(x_k) s_k^1 s_k^1 \right) \\ &\leq \|s_k\| \|\nabla f(x_k)\| + \frac{1}{2} \|s_k\| \|\nabla^3 f(x_k)\| \|s_k^1\|^2 - \lambda_k \|s_k\|^2 \\ &\leq \|s_k\| \|\nabla f(x_k)\| + \frac{L_3 \|\nabla f(x_k)\|}{2 \lambda_k} \|s_k^1\| \|s_k\| - \lambda_k \|s_k\|^2 \\ &\leq \frac{3}{2} \|s_k\| \|\nabla f(x_k)\| - \lambda_k \|s_k\|^2, \end{aligned}$$

where the first inequality follows from the norm triangle inequality, the second inequality follows from Assumption 1 and (8), and the last inequality follows from $\lambda_k \geq L_3$ and (9). Maximizing the right-hand side of the above inequality with respect to $\|s_k\|$ yields (12). $\square$

The following theorem establishes the descent property of the objective function, ensuring that each iteration sufficiently reduces the gradient norm.

Theorem 1. Suppose that Assumption 1 holds and that the regularization parameter λ_k satisfies

$$\lambda_k \geq \max \left\{ c_1 \|\nabla f(x_k)\|^{\frac{2}{3}}, c_2 \|\nabla f(x_k)\|^{\frac{4}{5}}, 1 \right\}, \quad (13)$$

where $c_1 = \max \left\{ \frac{75L_3}{8}, \frac{5(2+L_3)^2 L_3^2}{16C^2}, \frac{5(2+L_3)^2 L_3^2}{24C} \right\}^{\frac{1}{3}}$, $c_2 = \left(\frac{5(2+L_3)^4 L_3^2}{576} \right)^{\frac{1}{5}}$. Then, we have

$$\langle \nabla f(x_{k+1}), -s_k \rangle \geq \frac{1}{4 \lambda_k} \|\nabla f(x_{k+1})\|^2. \quad (14)$$

Proof. From (2) and (7), we obtain

$$\begin{aligned}\frac{L_3}{6}\|s_k\|^3 &\geq \left\| \nabla f(x_{k+1}) - \nabla f(x_k) - \nabla^2 f(x_k)s_k - \frac{1}{2}\nabla^3 f(x_k)s_k s_k \right\| \\ &= \left\| \nabla f(x_{k+1}) + \lambda_k s_k + \frac{1}{2}\nabla^3 f(x_k)s_k^1 s_k^1 - \frac{1}{2}\nabla^3 f(x_k)s_k s_k \right\| \\ &\geq \left\| \nabla f(x_{k+1}) + \lambda_k s_k - \frac{1}{2}\nabla^3 f(x_k)s_k s_k \right\| - \left\| \frac{1}{2}\nabla^3 f(x_k)s_k^1 s_k^1 \right\|.\end{aligned}$$

Squaring both sides of the above inequality yields

$$\begin{aligned}&\left(\frac{1}{2}\|\nabla^3 f(x_k)s_k^1 s_k^1\| + \frac{L_3}{6}\|s_k\|^3 \right)^2 \\ &\geq \|\nabla f(x_{k+1})\|^2 + \lambda_k^2\|s_k\|^2 + \frac{1}{4}\|\nabla^3 f(x_k)s_k s_k\|^2 + 2\lambda_k\langle \nabla f(x_{k+1}), s_k \rangle \\ &\quad - \langle \nabla f(x_{k+1}) + \lambda_k s_k, \nabla^3 f(x_k)s_k s_k \rangle.\end{aligned}\tag{15}$$

By using the Cauchy–Schwarz inequality and the norm triangle inequality, we obtain

$$\begin{aligned}\langle \nabla f(x_{k+1}), \nabla^3 f(x_k)s_k s_k \rangle &\leq \|\nabla f(x_{k+1})\| \|\nabla^3 f(x_k)s_k s_k\| \\ &\leq \frac{\|\nabla f(x_{k+1})\|^2 + \|\nabla^3 f(x_k)s_k s_k\|^2}{2},\end{aligned}$$

and

$$\lambda_k \langle s_k, \nabla^3 f(x_k)s_k s_k \rangle \leq \lambda_k \|s_k\| \|\nabla^3 f(x_k)s_k s_k\|.$$

By substituting these inequalities into (15) and using $\lambda_k \geq 1$, we obtain

$$\begin{aligned}&\langle \nabla f(x_{k+1}), -s_k \rangle \\ &\geq \frac{1}{4\lambda_k}\|\nabla f(x_{k+1})\|^2 + \frac{\lambda_k\|s_k\|^2}{2} - \frac{1}{8\lambda_k}\|\nabla^3 f(x_k)s_k s_k\|^2 \\ &\quad - \frac{\|s_k\|}{2}\|\nabla^3 f(x_k)s_k s_k\| - \frac{1}{2}\left(\frac{1}{2}\|\nabla^3 f(x_k)s_k^1 s_k^1\| + \frac{L_3}{6}\|s_k\|^3 \right)^2 \\ &= \frac{1}{4\lambda_k}\|\nabla f(x_{k+1})\|^2 + \frac{\lambda_k\|s_k\|^2}{2} - \frac{1}{8\lambda_k}\|\nabla^3 f(x_k)s_k s_k\|^2 \\ &\quad - \frac{\|s_k\|}{2}\|\nabla^3 f(x_k)s_k s_k\| - \frac{1}{8}\|\nabla^3 f(x_k)s_k^1 s_k^1\|^2 \\ &\quad - \frac{L_3^2}{72}\|s_k\|^6 - \frac{L_3}{12}\|\nabla^3 f(x_k)s_k^1 s_k^1\|\|s_k\|^3.\end{aligned}$$

By combining (8), (10)–(12), we obtain

$$\|s_k\| \leq \left(1 + \frac{L_3}{2} \right) \|s_k^1\| \leq \left(1 + \frac{L_3}{2} \right) \frac{\|\nabla f(x_k)\|}{\lambda_k},$$

and

$$\|\nabla^3 f(x_k)s_k s_k\| \leq \frac{3L_3}{8\lambda_k}\|\nabla f(x_k)\|^2\|s_k\|^2.$$

Thus, we obtain

$$\begin{aligned} & \langle \nabla f(x_{k+1}), -s_k \rangle \\ & \geq \frac{1}{4\lambda_k} \|\nabla f(x_{k+1})\|^2 + \frac{\lambda_k \|s_k\|^2}{2} - \frac{3L_3}{64\lambda_k^2} \|\nabla f(x_k)\|^2 \|s_k\|^2 \\ & \quad - \sqrt{\frac{3L_3}{32\lambda_k}} \|\nabla f(x_k)\| \|s_k\|^2 - \frac{\left(1 + \frac{L_3}{2}\right)^2 L_3^2}{8C^2 \lambda_k^2} \|\nabla f(x_k)\|^2 \|s_k\|^2 \\ & \quad - \frac{\left(1 + \frac{L_3}{2}\right)^4 L_3^2}{72\lambda_k^4} \|\nabla f(x_k)\|^4 \|s_k\|^2 - \frac{\left(1 + \frac{L_3}{2}\right)^2 L_3^2}{12C\lambda_k^2} \|\nabla f(x_k)\|^2 \|s_k\|^2. \end{aligned}$$

To ensure that (14) holds, we divide $\frac{\lambda_k \|s_k\|^2}{2}$ into five parts, each satisfying the following conditions:

- (i). $\frac{\lambda_k \|s_k\|^2}{10} \geq \frac{3L_3}{64\lambda_k^2} \|\nabla f(x_k)\|^2 \|s_k\|^2$, which leads to $\lambda_k \geq \left(\frac{15L_3}{32}\right)^{\frac{1}{3}} \|\nabla f(x_k)\|^{\frac{2}{3}}$.
- (ii). $\frac{\lambda_k \|s_k\|^2}{10} \geq \sqrt{\frac{3L_3}{32\lambda_k}} \|\nabla f(x_k)\| \|s_k\|^2$, which leads to $\lambda_k \geq \left(\frac{75L_3}{8}\right)^{\frac{1}{3}} \|\nabla f(x_k)\|^{\frac{2}{3}}$.
- (iii). $\frac{\lambda_k \|s_k\|^2}{10} \geq \frac{\left(1 + \frac{L_3}{2}\right)^2 L_3^2}{8C^2 \lambda_k^2} \|\nabla f(x_k)\|^2 \|s_k\|^2$, which leads to $\lambda_k \geq \left(\frac{5(2+L_3)^2 L_3^2}{16C^2}\right)^{\frac{1}{3}} \|\nabla f(x_k)\|^{\frac{2}{3}}$.
- (iv). $\frac{\lambda_k \|s_k\|^2}{10} \geq \frac{\left(1 + \frac{L_3}{2}\right)^4 L_3^2}{72\lambda_k^4} \|\nabla f(x_k)\|^4 \|s_k\|^2$, which leads to $\lambda_k \geq \left(\frac{5(2+L_3)^4 L_3^2}{576}\right)^{\frac{1}{5}} \|\nabla f(x_k)\|^{\frac{4}{5}}$.
- (v). $\frac{\lambda_k \|s_k\|^2}{10} \geq \frac{\left(1 + \frac{L_3}{2}\right)^2 L_3^2}{12C\lambda_k^2} \|\nabla f(x_k)\|^2 \|s_k\|^2$, which leads to $\lambda_k \geq \left(\frac{5(2+L_3)^2 L_3^2}{24C}\right)^{\frac{1}{3}} \|\nabla f(x_k)\|^{\frac{2}{3}}$.

By combining these conditions, we obtain the requirement (13). Thus, the theorem is proven. $\square$

Based on the previous results, we now establish the global convergence of the proposed method.

Theorem 2. Suppose that Assumptions 1–2 hold. Thus, the sequence $\{x_k\}$ generated by the Algorithm 1 satisfies

$$f(x_k) - f^* \leq \begin{cases} \left(\frac{12c_1 D^{\frac{4}{3}}}{k}\right)^3, & \text{if } \lambda_k = c_1 \|\nabla f(x_k)\|^{\frac{2}{3}}, \\ \left(\frac{20c_2 D^{\frac{6}{5}}}{k}\right)^5, & \text{if } \lambda_k = c_2 \|\nabla f(x_k)\|^{\frac{4}{5}}. \end{cases} \quad (16)$$

Proof. From Assumption 2, for any $x \in \mathcal{L}(x_0)$, we have

$$\langle \nabla f(x_k), x_k - x^* \rangle \geq f(x_k) - f^*,$$

which implies

$$\|\nabla f(x_k)\| \geq \frac{f(x_k) - f^*}{D}, \quad (17)$$

where D is the bound of the level set $\mathcal{L}(x_0)$.

For $0 \leq \beta \leq 1$ and $x \geq 0$, the function $y(x) = x^\beta$ is concave. Consequently, we have

$$a^\beta - b^\beta \geq \frac{\beta}{a^{1-\beta}}(a - b), \quad \forall a > b \geq 0. \quad (18)$$

Let $\lambda_k = c_1 \|\nabla f(x_k)\|^{\frac{2}{3}}$. By combining this with (14), we obtain

$$f(x_k) - f(x_{k+1}) \geq \langle \nabla f(x_{k+1}), x_k - x_{k+1} \rangle \quad (19)$$

$$\geq \frac{1}{4\lambda_k} \|\nabla f(x_{k+1})\|^2 = \frac{\|\nabla f(x_{k+1})\|^2}{4c_1 \|\nabla f(x_k)\|^{\frac{2}{3}}}. \quad (20)$$

Let $\delta_k = f(x_k) - f^*$. Thus,

$$\begin{aligned} \frac{1}{\delta_{k+1}^\beta} - \frac{1}{\delta_k^\beta} &= \frac{\delta_k^\beta - \delta_{k+1}^\beta}{\delta_{k+1}^\beta \delta_k^\beta} \\ &\geq \frac{\beta(\delta_k - \delta_{k+1})}{\delta_k \delta_{k+1}^\beta} \\ &= \frac{\beta(f(x_k) - f(x_{k+1}))}{\delta_k \delta_{k+1}^\beta} \\ &\geq \frac{\beta \|\nabla f(x_{k+1})\|^2 \|\nabla f(x_k)\|^{\frac{4}{3}}}{4c_1 \|\nabla f(x_k)\|^2 \delta_k \delta_{k+1}^\beta} \\ &\geq \frac{\beta \|\nabla f(x_{k+1})\|^2 \delta_k^{\frac{4}{3}}}{4c_1 D^{\frac{4}{3}} \|\nabla f(x_k)\|^2 \delta_k \delta_{k+1}^\beta}, \end{aligned}$$

where the first inequality follows from (18), the second inequality follows from (19), and the third inequality follows from (17). By setting $\beta = \frac{1}{3}$ and using the fact that $\delta_{k+1} < \delta_k$, the above inequality is simplified to

$$\frac{1}{\delta_{k+1}^\beta} - \frac{1}{\delta_k^\beta} \geq \frac{1}{12c_1 D^{\frac{4}{3}}} \frac{\|\nabla f(x_{k+1})\|^2}{\|\nabla f(x_k)\|^2}.$$

By summing these inequalities from 1 to k , we obtain

$$\begin{aligned} \frac{1}{\delta_k^\beta} - \frac{1}{\delta_0^\beta} &\geq \frac{1}{12c_1 D^{\frac{4}{3}}} \sum_{i=0}^{k-1} \frac{\|\nabla f(x_{i+1})\|^2}{\|\nabla f(x_i)\|^2} \\ &\geq \frac{1}{12c_1 D^{\frac{4}{3}}} \left(\prod_{i=0}^{k-1} \frac{\|\nabla f(x_{i+1})\|}{\|\nabla f(x_i)\|} \right)^{\frac{2}{k}} \\ &\geq \frac{1}{12c_1 D^{\frac{4}{3}}} \left(\frac{\delta_k}{D \|\nabla f(x_0)\|} \right)^{\frac{2}{k}} \\ &= \frac{1}{12c_1 D^{\frac{4}{3}}} \exp \left(-\frac{2}{k} \ln \frac{D \|\nabla f(x_0)\|}{\delta_k} \right) \\ &\geq \frac{1}{12c_1 D^{\frac{4}{3}}} \left(1 - \frac{2}{k} \ln \frac{D \|\nabla f(x_0)\|}{\delta_k} \right), \end{aligned}$$

where the second inequality follows from the geometric–arithmetic mean inequality, and the third inequality follows from (17).

As $k \rightarrow \infty$, the inequality $1 < \frac{2}{k} \ln \frac{D \|\nabla f(x_0)\|}{\delta_k}$ does not hold. Therefore, we have

$$1 \geq \frac{2}{k} \ln \frac{D \|\nabla f(x_0)\|}{\delta_k},$$

which implies

$$\frac{1}{\delta_k^\beta} \geq \frac{1}{12c_1 D^{\frac{4}{3}}}.$$

Thus, the first case in (16) is proven.

Let $\lambda_k = c_2 \|\nabla f(x_k)\|^{\frac{4}{5}}$ and $\beta = \frac{1}{5}$. Following a similar derivation process, the second case in (16) can be proven. $\square$

In Algorithm 1, we set $\lambda_k = H \|\nabla f(x_k)\|^\alpha$. Based on the convergence analysis, we know that α can be chosen as $\frac{2}{3}$ or $\frac{4}{5}$. Correspondingly, the value of H is given by $H = \max \left\{ \frac{75L_3}{8}, \frac{5(2+L_3)^2 L_3^2}{16C^2}, \frac{5(2+L_3)^2 L_3^2}{24C} \right\}^{\frac{1}{3}}$, or $\left(\frac{5(2+L_3)^4 L_3^2}{576} \right)^{\frac{1}{5}}$.

Next, we investigate the local convergence of the proposed method.

Theorem 3. Suppose that Assumption 1 holds and that $f(x)$ is a σ_2 -strongly convex function on $\mathbb{R}^n$. If there exists a constant $k^* > 0$ such that for all $k \geq k^*$, $H \|s_k^1\| \geq c_0$, where c_0 is a sufficiently small positive constant, then

$$\|\nabla f(x_{k+1})\| \leq \left(\frac{L_3 H}{6\sigma_2^3} + \frac{H}{\sigma_2 c_0^{\frac{1}{1-\alpha}}} \right) \|\nabla f(x_k)\|^{2+\alpha}. \quad (21)$$

Proof. From the strong convexity of $f(x)$, it follows that

$$\langle \nabla f(x_k), x_k - x_{k+1} \rangle \geq \langle \nabla f(x_{k+1}), x_k - x_{k+1} \rangle + \sigma_2 \|x_k - x_{k+1}\|^2.$$

By combining (14), we obtain

$$\|\nabla f(x_k)\| \geq \sigma_2 \|s_k\|. \quad (22)$$

Further, by combining (2) and the first-order optimality condition of Model (4), we have

$$\begin{aligned} \|\nabla f(x_{k+1})\| &\leq \left\| \nabla f(x_{k+1}) - \left(\nabla f(x_k) + \nabla^2 f(x_k) s + \frac{1}{2} \nabla^3 f(x_k) s s \right) \right\| + \lambda_k \|s_k\| \\ &\leq \frac{L_3}{6} \|s_k\|^3 + \lambda_k \|s_k\| \\ &\leq \frac{L_3}{6\sigma_2^3} \|\nabla f(x_k)\|^3 + \frac{H}{\sigma_2} \|\nabla f(x_k)\|^{1+\alpha} \\ &\leq \left(\frac{L_3 \|\nabla f(x_k)\|^{1-\alpha}}{6\sigma_2^3} + \frac{H}{\sigma_2 (H \|s_k^1\|)^{\frac{1}{1-\alpha}}} \right) \|\nabla f(x_k)\|^{2+\alpha}, \\ &\leq \left(\frac{L_3 H}{6\sigma_2^3} + \frac{H}{\sigma_2 c_0^{\frac{1}{1-\alpha}}} \right) \|\nabla f(x_k)\|^{2+\alpha}, \end{aligned}$$

where the third inequality follows from (22), the fourth inequality follows from (8), and the last inequality follows from Corollary 1. $\square$

Under the strong convexity condition, most existing works (e.g., refs. [34–38]) have established the local quadratic convergence of various Newton-type methods. However, the third-order method proposed in this paper exhibits a faster convergence rate, achieving an order of $2 + \alpha$, where $\alpha \in [\frac{2}{3}, 1]$.

4. Numerical Experiments

In this section, we evaluate the performance of CHGR (Algorithm 1) on two classical test problems: l_2 -regularized logistic regression and log-sum-exp. The experiments

compare CHGR with various gradient-regularized methods, including the regularized Newton method (RegN [35,36]), the super-universal Newton method (SUN [39]), CNM [3] and RCH [24]. Additionally, we focus on several damped Newton methods with simple iterations and reliable global convergence properties, such as the affine-invariant cubic Newton method (AICN [42]) and the root Newton stepsize schedule (RN [43]). All numerical experiments were conducted in the MATLAB R2019b environment on a laptop equipped with an Intel Core i7 processor and 8GB of memory. Table 1 summarizes the iteration schemes and global convergence properties of the aforementioned methods.

Table 1. Summary of iteration schemes and convergence properties for third-order methods ($p = 3$).

Algorithm	Expression	Global Convergence Rate
CNM	$x_{k+1} = x_k - (\nabla^2 f(x_k) + H\ x_{k+1} - x_k\ I)^{-1} \nabla f(x_k)$	$O(k^{-2})$
SUN	$x_{k+1} = x_k - (\nabla^2 f(x_k) + H\ \nabla f(x_k)\ ^a I)^{-1} \nabla f(x_k)$	$O(k^{-3})$
RegN	$x_{k+1} = x_k - \left(\nabla^2 f(x_k) + \sqrt{H\ \nabla f(x_k)\ } I \right)^{-1} \nabla f(x_k)$	$O(k^{-2})$
AICN	$x_{k+1} = x_k - \eta_k (\nabla^2 f(x_k))^{-1} \nabla f(x_k)$ $\eta_k = \frac{-1 + \sqrt{1 + 2H \sqrt{\nabla f(x_k)^T (\nabla^2 f(x_k))^{-1} \nabla f(x_k)}}}{\sqrt{\nabla f(x_k)^T (\nabla^2 f(x_k))^{-1} \nabla f(x_k)}}$	$O(k^{-2})$
RN	$x_{k+1} = x_k - \eta_k (\nabla^2 f(x_k))^{-1} \nabla f(x_k)$ $\eta_k = \frac{1}{(9H)^{\frac{1}{3}} \left(\nabla f(x_k)^T (\nabla^2 f(x_k))^{-1} \nabla f(x_k) \right)^{\frac{1}{3}}}$	$O(k^{-3})$
RCH	$\left(\nabla^2 f(x_k) + \frac{H}{2} \ s_k\ ^2 I \right) s_k^1 = -\nabla f(x_k)$ $\left(\nabla^2 f(x_k) + \frac{H}{2} \ s_k\ ^2 I \right) s_k = -\nabla f(x_k) - \frac{1}{2} \nabla^3 f(x_k) s_k^1 s_k^1$	$O(k^{-3})$
CHGR	$x_{k+1} = x_k + s_k$ $(6), (7)$	$O(k^{-3})$ or $O(k^{-5})$

The following two problems are widely used in the fields of optimization and machine learning, commonly employed to evaluate the effectiveness and robustness of optimization algorithms in practical tasks.

Problem 1. l_2 -regularized logistic regression:

$$\min_{x \in \mathbb{R}^n} f(x) = \frac{1}{d} \sum_{i=1}^d \log \left(1 + \exp \left(-y_i \left(\omega_i^T x \right) \right) \right) + \frac{l_2}{2} \|x\|^2,$$

where $(y_i, \omega_i)_{i=1}^d$ are the given data samples, with $\omega_i \in \mathbb{R}^n$ representing the feature vectors and $y_i \in \{-1, 1\}$ representing the class label. l_2 is the regularization coefficient that controls model complexity and prevents overfitting. If l_2 is too small, it may result in ill-conditioned problems, negatively affecting the stability and convergence of the methods.

The experiments are conducted on datasets provided by LIBSVM (<https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/>, accessed on 14 April 2025), including the ionosphere, mushroom, a4a, and a9a datasets. All linear systems are solved using the MATLAB backslash operator “\”. The initial point for AICN and RN is chosen as $x_0 = 10 * (1, \dots, 1)^T \in \mathbb{R}^n$ [42,43], while the initial points for the remaining methods are randomly generated from a standard normal distribution. The termination criterion for all methods is set as $\|\nabla f(x_k)\| \leq 1 \times 10^{-10}$, and H is selected to be either 0.1 or 0.5. To evaluate the adaptability of the methods under different regularization strengths, l_2 is set to 1×10^{-2} or 1×10^{-6} .

Problem 2. *log-sum-exp:*

$$\min_{x \in \mathbb{R}^n} f(x) = \rho \log \left(\sum_{i=1}^d \exp \left(\frac{a_i^T x - b_i}{\rho} \right) \right),$$

where the components of vectors $a_1, \dots, a_d \in \mathbb{R}^n$ and $b \in \mathbb{R}^d$ are independently sampled from a uniform distribution over the interval $[-1, 1]$. The parameter $\rho > 0$ is a smoothing factor that controls the approximation accuracy. This function serves as a smooth approximation of $\max\{a_1^T x - b_1, \dots, a_n^T x - b_n\}$. When ρ is small, the function becomes poorly conditioned.

In the experiments, the initial points x_0 for all methods are generated from a standard normal distribution, with the initial parameter set as $H = 0.1$. The termination criterion is set as $\|\nabla f(x_k)\| \leq 1 \times 10^{-10}$. To comprehensively evaluate the performance of the methods, two problem sizes are tested: $n = 300$, $d = 100$ and $n = 700$, $d = 300$. Additionally, different smoothing parameters ρ are selected to assess the convergence stability of the methods.

Figures 1–4 illustrate the performance comparison of the methods on Problem 1. The detailed analysis is as follows:

- **Efficient Convergence**
 - In most cases, CHGR performs best with $\alpha = \frac{4}{5}$, which is consistent with theoretical results. In terms of both iteration count and computational efficiency, CHGR significantly outperforms second-order methods such as AICN, RN, RegN, and CNM. This is due to its two-step iterative framework, which accelerates convergence by incorporating third-order derivative information while maintaining computational complexity comparable to that of second-order methods.
 - For $l_2 = 10^{-2}$, AICN has a slightly less or comparable computational time to CHGR, due to fewer matrix and vector multiplication operations. However, for $l_2 = 10^{-6}$, AICN fails after only two steps of a normal iteration due to its high sensitivity to parameters. A similar issue also occurs in RN.
 - For $\alpha = \frac{2}{3}$, CHGR outperforms SUN in terms of the computational time on a9a ($l_2 = 10^{-2}$) but performs slightly worse on a4a ($l_2 = 10^{-2}$). SUN is a single-step Newton iteration with an adaptive update strategy for L_3 , and it can accelerate convergence in some cases. However, the need to verify the descent condition for the adaptive strategy introduces additional computational costs.
- **Numerical Stability**
 - CHGR maintains stable convergence across different regularization strengths, while CNM and RCH fail to achieve satisfactory solutions within 100 steps for $l_2 = 10^{-6}$. This may be due to the more flexible regularization strategy of CHGR, which dynamically adjusts the regularization parameter based on the gradient norm, thereby improving convergence.
 - For $l_2 = 10^{-2}$, AICN and CHGR exhibit comparable performance. However, for $l_2 = 10^{-6}$, the weaker regularization causes the problem to become ill-conditioned, leading to the failure of AICN and RN to converge. Additionally, these two methods are more sensitive to the initial point, often failing due to numerical instability when initialized with $x_0 = (1, 1, \dots, 1)^T$ or a random initialization.

Figures 5 and 6 present the comparative results of the methods on Problem 2. The experimental results demonstrate that CHGR consistently achieves fast and stable convergence, showcasing its excellent convergence performance and strong adaptability to various problems. In comparison, AICN and RN succeed only when the problem is well

conditioned, failing to converge for ill-conditioned cases like $d = 300, n = 100, \rho = 0.1$ or $d = 700, n = 300, \rho = 0.5$. Additionally, CNM and RCH also diverge for $d = 300, n = 100, \rho = 0.1$, indicating their limitations in handling ill-conditioned problems.

In summary, CHGR demonstrates excellent convergence and numerical stability across various test problems and different regularization strengths, indicating its strong flexibility and applicability in handling complex problems.

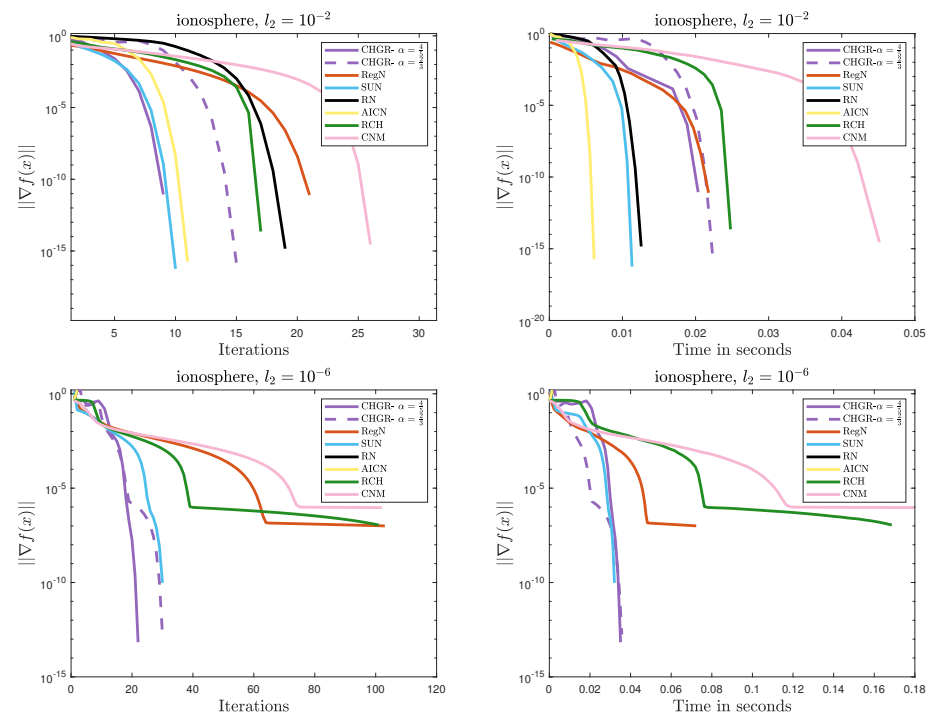


Figure 1. Performance comparison of the methods on Problem 1 (ionosphere: $n = 34, d = 351$).

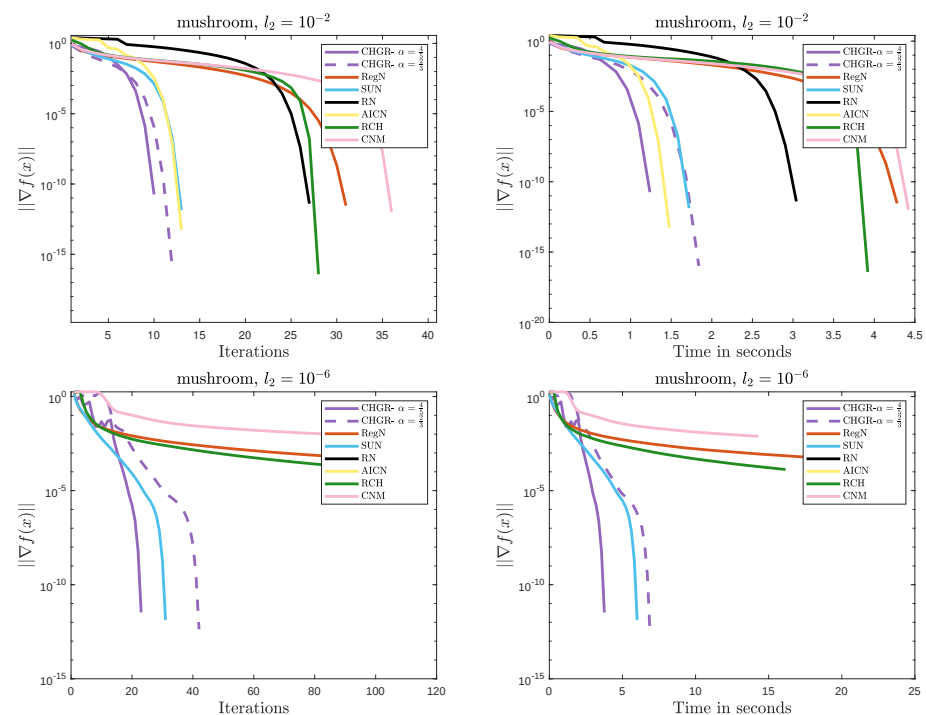


Figure 2. Performance comparison of the methods on Problem 1 (mushroom: $n = 112, d = 8124$).

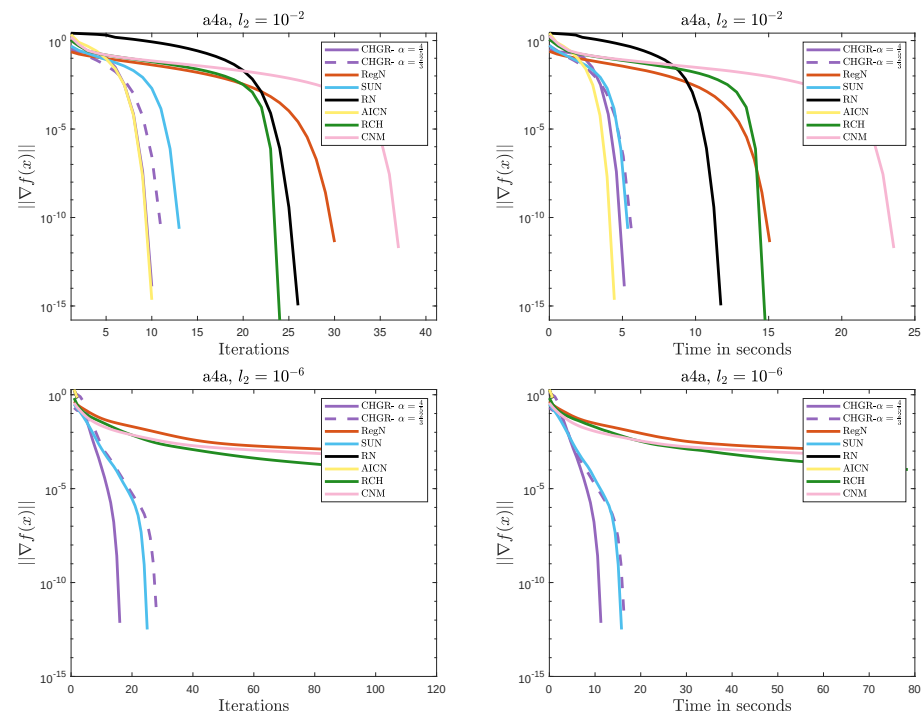


Figure 3. Performance comparison of the methods on Problem 1 (a4a: $n = 123$, $d = 27,780$).

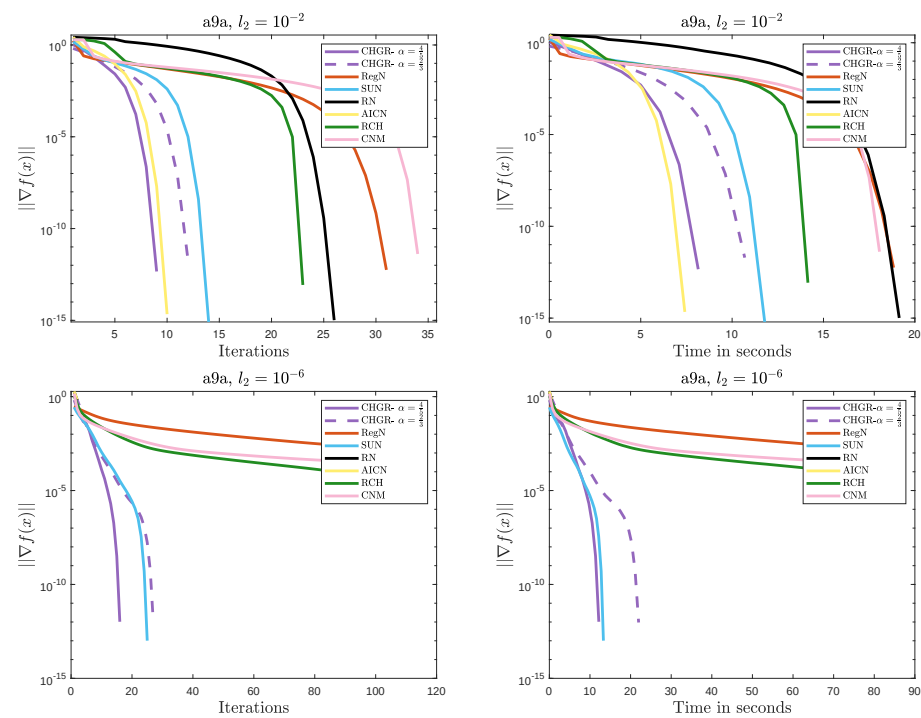


Figure 4. Performance comparison of the methods on Problem 1 (a9a: $n = 123$, $d = 32,561$).

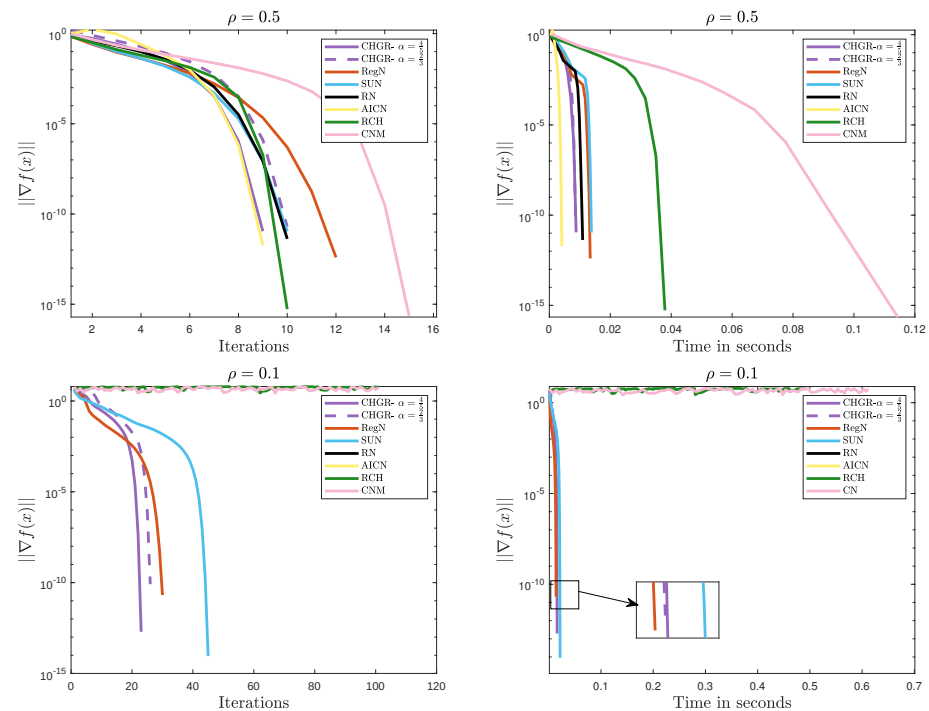


Figure 5. Performance comparison of the methods on Problem 2 ($d = 300$, $n = 100$).

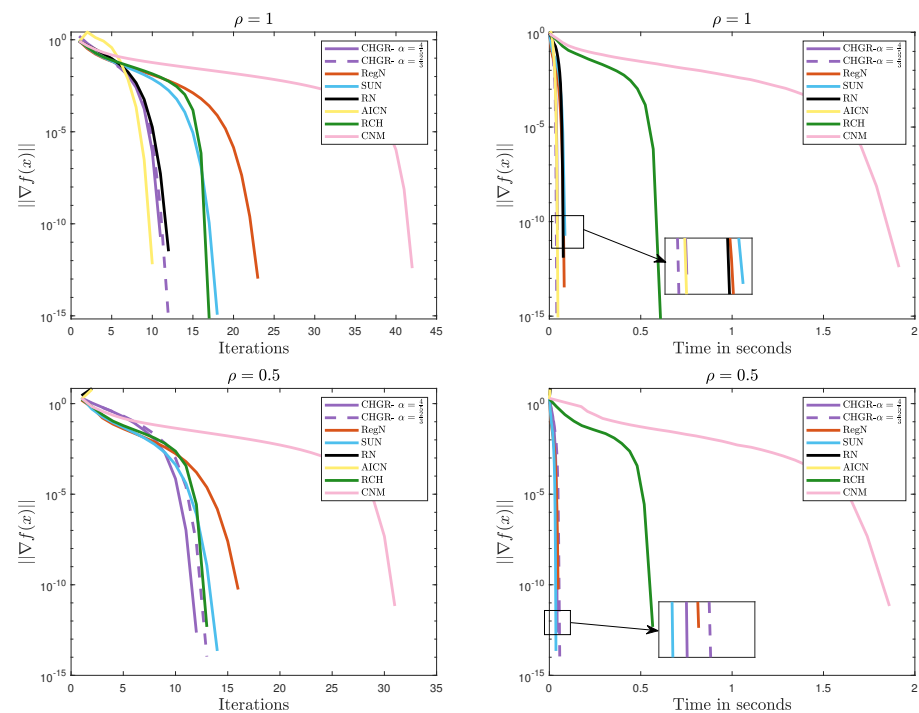


Figure 6. Performance comparison of the methods on Problem 2 ($d = 700$, $n = 300$).

5. Results

This paper proposes a Chebyshev–Halley method with gradient regularization, which demonstrates excellent global convergence properties while maintaining iteration simplicity. The method employs a third-order Taylor polynomial model with a quadratic regularization term and adjusts the regularization parameter based on a power of the gradient norm. In each iteration, the method requires only matrix inversion and vector products to obtain a closed-form solution, avoiding the complexity of nested sub-iterations for solving the polynomial model. Despite not incorporating acceleration techniques, the

method theoretically achieves a global convergence rate of $O(k^{-3})$ or $O(k^{-5})$ and exhibits locally superlinear convergence for strongly convex problems. Numerical experiments show that the proposed method outperforms gradient-regularized Newton-type methods, such as RegN and SUN. Moreover, it exhibits greater adaptability than damped Newton methods like AICN and RN.

Author Contributions: Methodology, software, formal analysis and writing—original draft preparation: J.X.; methodology, funding Acquisition and supervision: H.Z.; formal analysis, writing—review and editing, funding acquisition: H.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Science Foundation of China (Grant No. 12471295, Grant No. 12171145) and the Training Program for Excellent Young Innovators of Changsha (Grant No. kq2209021).

Data Availability Statement: No new data were created or analyzed in this study.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Anandkumar, A.; Ge, R. Efficient approaches for escaping higher order saddle points in non-convex optimization. In Proceedings of the 29th Annual Conference on Learning Theory, New York, NY, USA, 23–26 June 2016; pp. 81–102.
2. Chen, X.; Toint, P.L. High-order evaluation complexity for convexly-constrained optimization with non-Lipschitzian group sparsity terms. *Math. Program.* **2021**, *187*, 47–78. [\[CrossRef\]](#)
3. Nesterov, Y.; Polyak, B.T. Cubic regularization of Newton method and its global performance. *Math. Program.* **2006**, *108*, 177–205. [\[CrossRef\]](#)
4. Baes, M. *Estimate Sequence Methods: Extensions and Approximations*; Institute for Operations Research, ETH: Zürich, Switzerland, 2009; Volume 2.
5. Gould, N.I.M.; Rees, T.; Scott, J.A. *A Higher Order Method for Solving Nonlinear Least-Squares Problems*; RAL Preprint RAL-P-2017–010; STFC Rutherford Appleton Laboratory: Oxfordshire, UK, 2017.
6. Calandra, H.; Gratton, S.; Riccietti, E.; Vasseur, X. On high-order multilevel optimization strategies. *SIAM J. Optimiz.* **2021**, *31*, 307–330. [\[CrossRef\]](#)
7. Jiang, B.; Wang, H.; Zhang, S. An optimal high-order tensor method for convex optimization. *Math. Oper. Res.* **2021**, *46*, 1390–1412. [\[CrossRef\]](#)
8. Ahookhosh, M.; Nesterov, Y. High-order methods beyond the classical complexity bounds: Inexact high-order proximal-point methods. *Math. Program.* **2024**, *208*, 365–407. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Cartis, C.; Gould, N.I.M.; Toint, P.L. Adaptive cubic regularisation methods for unconstrained optimization. Part I: Motivation, convergence and numerical results. *Math. Program.* **2011**, *127*, 245–295. [\[CrossRef\]](#)
10. Cartis, C.; Gould, N.I.M.; Toint, P.L. Adaptive cubic regularisation methods for unconstrained optimization. Part II: Worst-case function-and derivative-evaluation complexity. *Math. Program.* **2011**, *130*, 295–319. [\[CrossRef\]](#)
11. Jiang, B.; Lin, T.; Zhang, S. A unified adaptive tensor approximation scheme to accelerate composite convex optimization. *SIAM J. Optimiz.* **2020**, *30*, 2897–2926. [\[CrossRef\]](#)
12. Grapiglia, G.N.; Nesterov, Y. Adaptive third-order methods for composite convex optimization. *SIAM J. Optimiz.* **2023**, *33*, 1855–1883. [\[CrossRef\]](#)
13. Chen, X.; Jiang, B.; Lin, T.; Zhang, S. Accelerating adaptive cubic regularization of Newton’s method via random sampling. *J. Mach. Learn. Res.* **2022**, *23*, 1–38.
14. Lucchi, A.; Kohler, J. A sub-sampled tensor method for nonconvex optimization. *IMA J. Numer. Anal.* **2023**, *43*, 2856–2891. [\[CrossRef\]](#)
15. Agafonov, A.; Kamzolov, D.; Dvurechensky, P.; Gasnikov, A.; Takáč, M. Inexact tensor methods and their application to stochastic convex optimization. *Optim. Methods Softw.* **2024**, *39*, 42–83. [\[CrossRef\]](#)
16. Zhao, J.; Doikov, N.; Lucchi, A. Cubic regularized subspace Newton for non-convex optimization. In Proceedings of the 28th International Conference on Artificial Intelligence and Statistics, Phuket, Thailand, 3–5 May 2025.
17. Nesterov, Y. Implementable tensor methods in unconstrained convex optimization. *Math. Program.* **2021**, *186*, 157–183. [\[CrossRef\]](#)
18. Griewank, A. A mathematical view of automatic differentiation. *Acta Numer.* **2003**, *12*, 321–398. [\[CrossRef\]](#)
19. Zhang, H.; Deng, N. An improved inexact Newton method. *J. Glob. Optim.* **2007**, *39*, 221–234. [\[CrossRef\]](#)

20. Doikov, N.; Nesterov, Y.E. Inexact Tensor Methods with Dynamic Accuracies. In Proceedings of the 37th International Conference on Machine Learning, Virtual Event, 13–18 July 2020; pp. 2577–2586.
21. Cartis, C.; Zhu, W. Cubic-quartic regularization models for solving polynomial subproblems in third-order tensor methods. *Math. Program.* **2025**, 1–53. [\[CrossRef\]](#)
22. Ahmadi, A.A.; Chaudhry, A.; Zhang, J. Higher-order newton methods with polynomial work per iteration. *Adv. Math.* **2024**, 452, 109808. [\[CrossRef\]](#)
23. Cartis, C.; Zhu, W. Global Convergence of High-Order Regularization Methods with Sums-of-Squares Taylor Models. *arXiv* **2024**, arXiv:2404.03035.
24. Xiao, J.; Zhang, H.; Gao, H. An Accelerated Regularized Chebyshev–Halley Method for Unconstrained Optimization. *Asia Pac. J. Oper. Res.* **2023**, 40, 2340008. [\[CrossRef\]](#)
25. Nesterov, Y. Accelerating the cubic regularization of Newton’s method on convex problems. *Math. Program.* **2008**, 112, 159–181. [\[CrossRef\]](#)
26. Nesterov, Y. Inexact accelerated high-order proximal-point methods. *Math. Program.* **2023**, 197, 1–26. [\[CrossRef\]](#)
27. Kamzolov, D.; Pasechnyuk, D.; Agafonov, A.; Gasnikov, A.; Takáč, M. OPTAMI: Global Superlinear Convergence of High-order Methods. *arXiv* **2024**, arXiv:2410.04083.
28. Antonakopoulos, K.; Kavis, A.; Cevher, V. Extra-newton: A first approach to noise-adaptive accelerated second-order methods. *NeurIPS* **2022**, 35, 29859–29872.
29. Monteiro, R.D.C.; Svaiter, B.F. An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods. *SIAM J. Optimiz.* **2013**, 23, 1092–1125. [\[CrossRef\]](#)
30. Birgin, E.G.; Gardenghi, J.L.; Martínez, J.M.; Santos, S.A.; Toint, P.L. Worst-case evaluation complexity for unconstrained nonlinear optimization using high-order regularized models. *Math. Program.* **2017**, 163, 359–368. [\[CrossRef\]](#)
31. Gould, N.I.M.; Rees, T.; Scott, J.A. Convergence and evaluation-complexity analysis of a regularized tensor-Newton method for solving nonlinear least-squares problems. *Comput. Optim. Appl.* **2019**, 73, 1–35. [\[CrossRef\]](#)
32. Huang, Z.; Jiang, B.; Jiang, Y. Inexact and Implementable Accelerated Newton Proximal Extragradient Method for Convex Optimization. *arXiv* **2024**, arXiv:2402.11951.
33. Polyak, R.A. Regularized Newton method for unconstrained convex optimization. *Math. Program.* **2009**, 120, 125–145. [\[CrossRef\]](#)
34. Ueda, K.; Yamashita, N. A regularized Newton method without line search for unconstrained optimization. *Comput. Optim. Appl.* **2014**, 59, 321–351. [\[CrossRef\]](#)
35. Doikov, N.; Nesterov, Y. Gradient regularization of Newton method with Bregman distances. *Math. Program.* **2024**, 204, 1–25. [\[CrossRef\]](#)
36. Mishchenko, K. Regularized Newton method with global $O(1/k^2)$ convergence. *SIAM J. Optimiz.* **2023**, 33, 1440–1462. [\[CrossRef\]](#)
37. Yamakawa, Y.; Yamashita, N. Convergence analysis of a regularized Newton method with generalized regularization terms for unconstrained convex optimization problems. *Appl. Math. Comput.* **2025**, 491, 129219. [\[CrossRef\]](#)
38. Jiang, Y.; He, C.; Zhang, C.; Ge, D.; Jiang, B.; Ye, Y. Beyond Nonconvexity: A Universal Trust-Region Method with New Analyses. *arXiv* **2023**, arXiv:2311.11489.
39. Doikov, N.; Mishchenko, K.; Nesterov, Y. Super-universal regularized newton method. *SIAM J. Optimiz.* **2024**, 34, 27–56. [\[CrossRef\]](#)
40. Gratton, S.; Jerad, S.; Toint, P.L. Yet another fast variant of Newton’s method for nonconvex optimization. *IMA J. Numer. Anal.* **2025**, 45, 971–1008. [\[CrossRef\]](#)
41. Zhou, Y.; Xu, J.; Bao, C.; Ding, C.; Zhu, J. A Regularized Newton Method for Nonconvex Optimization with Global and Local Complexity Guarantees. *arXiv* **2025**, arXiv:2502.04799.
42. Hanzely, S.; Kamzolov, D.; Pasechnyuk, D.; Gasnikov, A.; Richtárik, P.; Takáč, M. A Damped Newton Method Achieves Global $O\left(\frac{1}{k^2}\right)$ and Local Quadratic Convergence Rate. *NeurIPS* **2022**, 35, 25320–25334.
43. Hanzely, S.; Abdukhakimov, F.; Takáč, M. Damped Newton Method with Near-Optimal Global $O(k^{-3})$ Convergence Rate. *arXiv* **2024**, arXiv:2405.18926.
44. Jackson, R.H.F.; McCormick, G.P. The polyadic structure of factorable function tensors with applications to high-order minimization techniques. *J. Optimiz. Theory Appl.* **1986**, 51, 63–94. [\[CrossRef\]](#)
45. Zhang, H. On the Halley class of methods for unconstrained optimization problems. *Optim. Methods Softw.* **2010**, 25, 753–762. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.