Regularized Newton versus Nesterov Accelerated Gradient: Convergence, Stability, and Adaptive Switching Schemes

Boushra Rajab ABBAS
Associate professor at Latakia University
boushraabbas@latakia-univ.edu.sy
Njod Hassan Assistant professor - Lattakia University,
njod.hassan@latakia-univ.edu.sy
Ghaidaa Khadour Master Student- Lattakia University,
ghaidaakh@latakia-univ.edu.sy

Abstract

This paper presents a systematic comparison between the Regularized Newton Method and Nesterov's Accelerated Gradient (NAG) for smooth convex optimization problems with twice-differentiable objectives. We analyze their continuous-time dynamical interpretations, discrete-time convergence rates under strong convexity and Lipschitz Hessian assumptions, per-iteration complexity, and numerical stability. While NAG achieves the optimal $O(1/k^2)$ rate among first-order methods, the Regularized Newton method offers quadratic local convergence at $O(n^3)$ cost per iteration. We propose adaptive switching strategies that initiate

with NAG for global acceleration and transition to regularized Newton steps when the gradient norm falls below a predefined threshold. Numerical experiments demonstrate that the adaptive ANN scheme achieves global complexity comparable to NAG while requiring only $1–8$ Hessian solves, resulting in $2–6\times$ faster execution than pure Regularized Newton with enhanced local refinement. Theoretical analysis confirms linear convergence in the strongly convex case, with local superlinear behavior under standard assumptions. Numerical experiments on quadratic, logistic regression, and Rosenbrock functions, along with an application to robotic trajectory optimization, validate the stability and efficiency of the adaptive scheme. Our results highlight the complementary roles of first- and second-order information in modern optimization and serve as a clear reference for understanding the fundamental trade-offs between these two classical approaches.

Keywords: Convex Optimization, Regularized Newton, Nesterov Acceleration, Adaptive Methods, Convergence Analysis, Dynamical Systems

طريقة نيوتن المنظمة مقابل التدرج المتسارع لنيستروف: التقارب، الاستقرار، وخطط التبديل التكيفي

أ.م.د. بشرى رجب عباس
أستاذ مساعد في قسم الرياضيات– كلية العلوم– جامعة اللاذقية:
boushraabbas@latakia-univ.edu.sy
د.نجود حسن
أستاذ في قسم الرياضيات- كلية العلوم- جامعة اللاذقية:
njod.hassan@latakia-univ.edu.sy
غيداء خضور
طالبة ماجستير في قسم الرياضيات- كلية العلوم- جامعة اللاذقية:
ghaidaakh@latakia-univ.edu.sy

الملخص

تقدم هذه الورقة مقارنة منهجية بين طريقة نيوتن المنظمة والتدرج المتسارع لنيستروف (NAG) لمسائل الأمثليات المحدبة والملساء ذات دوال الهدف القابلة للاشتقاق مرتين. نحن نحلل تفسيراتهما الديناميكية في الزمن المستمر، ومعدلات تقاربهما في الزمن المنقطع تحت افتراضات التحدب القوي وشرط ليبشتز لمصفوفة الهيسيان، وتعقيد كل تكرار، واستقرارهما العددي. بينما يحقق التدرج المتسارع لنيستروف المعدل الأمثل $O(1/k^2)$ ضمن طرق الرتبة الأولى، توفر طريقة نيوتن المنظمة تقارباً محلياً تربيعياً بتكلفة $O(n^3)$ لكل تكرار.

نقترح استراتيجيات تبديل تكيفية تبدأ بالتدرج المتسارع لنيستروف للتسريع الشامل وتنتقل إلى خطوات نيوتن المنظمة عندما تقل معايير التدرج عن عتبة معينة. تظهر التجارب العددية أن خوارزمية ANN التكيفية تحقق تعقيداً عالمياً مماثلاً لـ NAG مع حلول هيسيان محدودة جداً (1-8 مرات فقط)، مما يجعلها أسرع 2-6 مرات من نيوتن

النقي مع صقل محلي محسن. يؤكد التحليل النظري التقارب الخطي في حالة التحدب القوي مع سلوك فوق خطي محلي بموجب الافتراضات القياسية.

تثبت التجارب العددية على الدوال التربيعية واللوجستية وروزنبروك، بالإضافة إلى تطبيق على تحسين مسار الروبوتات، استقرار وكفاءة الخطة التكيفية. تبرز نتائجنا الأدوار التكاملية لمعلومات الرتبة الأولى والثانية في الأمثليات الحديثة، وتقدم مرجعاً واضحاً لفهم المفاضلات الأساسية بين الطريقتين الكلاسيكيتين.

الكلمات المفتاحية: الأمثليات المحدبة، نيوتن المنظمة، تسريع نيستروف، الطرق التكيفية، تحليل التقارب، الأنظمة الديناميكية

1. Introduction

Optimization is a foundational pillar of applied mathematics, underpinning advancements in machine learning, optimal control, signal processing, and scientific computing. For unconstrained minimization problems of the form

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

where f is convex and twice continuously differentiable, two classical methods have shaped the field: Newton's method, introduced in the 17th century and refined with regularization [10, 11], and Nesterov's Accelerated Gradient Method (NAG) [1], which established the information-theoretic lower bound for first-order methods in the convex regime.

Newton-type methods exploit second-order information—the Hessian matrix $\nabla^2 f(x)$—to construct quadratic approximations of the objective, yielding quadratic local convergence near a minimizer under standard assumptions (strong convexity, Lipschitz Hessian). However, their $O(n^3)$ per-iteration cost due to Hessian inversion and sensitivity to ill-conditioning limit scalability in high-dimensional settings.

In contrast, NAG enhances gradient descent with a momentum term, achieving the optimal $O(1/k^2)$ convergence rate for smooth convex functions [1] and linear convergence under strong convexity [4]. Its $O(n)$ per-iteration complexity makes it the method of choice in large-scale machine learning, where n may exceed 10^9.

Despite their complementary strengths—local precision(Newton) versus global efficiency (NAG)—a systematic, unified comparison remains absent from the literature. Existing analyses are fragmented: dynamical system interpretations appear in [2, 6] for NAG and [5] for Newton flows, but rarely under shared assumptions. Convergence guarantees are typically derived in isolation, and stability under Hessian indefiniteness or numerical round-off is underexplored. Moreover, while hybrid methods have been proposed [7, 8], they often rely on abstract acceleration frameworks (e.g., HPE) without practical switching heuristics or empirical validation on non-quadratic functions.

This work fills this gap by providing a rigorous comparative framework for the Regularized Newton Method and NAG, analyzing:

- Continuous-time dynamics: damped Newton flows vs. inertial systems with time-varying friction,

- Discrete convergence rates: global $O(1/k^2)$ vs. local quadratic, under μ−strong convexity and L−Lipschitz Hessian,
- Computational complexity: $O(n^3)$ vs. $O(n)$ per iteration,
- Numerical stability: role of regularization, eigenvalue control, and oscillation damping.

We further propose adaptive switching schemes that dynamically transition from NAG to regularized Newton based on gradient norm decay. Unlike static hybrids that incur $O(n^3)$ cost throughout, our method preserves NAG's global $O\left(\sqrt{L/\epsilon}\right)$ complexity while activating second−order refinement only in a small neighborhood of the solution, achieving superlinear local convergence with minimal overhead. Theoretical analysis establishes:

- Global convergence in $O\left(\sqrt{L/\epsilon}\right)$ iterations,
- Local superlinear rate once $\| \nabla f(x_k) \| \leq \delta$,
- Robustness to Hessian ill−conditioning via adaptive regularization.

Numerical experiments on benchmark functions (quadratic, logistic regression, Rosenbrock) and real−world applications,

robotic trajectory optimization, and demonstrate that the adaptive scheme consistently outperforms standalone NAG and approaches Newton's precision at a fraction of the cost.

The paper is organized as follows: Section 2 reviews the Regularized Newton and NAG methods with their dynamical system analogs. Section 3 derives unified convergence bounds. Section 4 presents the adaptive switching algorithm and its theoretical guarantees. Section 5 reports empirical results. Section 6 concludes with extensions to non-convex and distributed optimization.

2. Mathematical Background

We consider the unconstrained optimization problem

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

where $f:\mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable ($f \in C^2$), convex, and satisfies the following standard assumptions:

f is μ-strongly convex for some $\mu > 0$:

$$f(y) \geq f(x) + \nabla f(x)^{\top}(y-x) + \frac{\mu}{2} \| y-x \|^2, \quad \forall x, y.$$

The Hessian $\nabla^2 f(x)$ is L-Lipschitz continuous:

$$\| \nabla^2 f(y) - \nabla^2 f(x) \| \leq L \| y-x \|, \quad \forall x, y.$$

These assumptions imply $mI \preccurlyeq \nabla^2 f(x) \preccurlyeq MI$ with $m \geq \mu > 0$ and $M \leq L + \mu$, ensuring positive definiteness and bounded condition number $\kappa = M/m \leq L/\mu + 1$.

2.1 Regularized Newton Method

The classical Newton update is

$$x_{k+1} = x_k - [\nabla^2 f(x_k)]^{-1} \nabla f(x_k).$$

However, $\nabla^2 f(x_k)$ may be indefinite or ill-conditioned far from the minimizer. The Regularized Newton Method addresses this by introducing a damping parameter $\lambda_k > 0$:

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

The modified Hessian $H_k + \lambda_k I$ has eigenvalues at least $\lambda_k > 0$, ensuring positive definiteness and invertibility.

In continuous time, the Regularized Newton method corresponds to the damped Newton flow:

$$\dot{x}(t) = -\big(\nabla^2 f\big(x(t)\big) + \lambda(t) I\big)^{-1} \nabla f\big(x(t)\big),$$

which can be interpreted as a nonlinear preconditioned gradient flow [5].

2.2 Nesterov's Accelerated Gradient Method (NAG)

Nesterov's method enhances gradient descent with an extrapolation (momentum) step. For a fixed step size $\alpha > 0$, the update is:

$$\begin{aligned} y_k &= x_k + \beta_k(x_k - x_{k-1}), \\ x_{k+1} &= y_k - \alpha\nabla f(y_k), \end{aligned}$$

where $\beta_k \in [0,1)$ is the momentum coefficient. A common choice is:

$$\beta_k = \frac{k}{k+3}, \quad \alpha = \frac{1}{L}.$$

This yields the optimal $O(1/k^2)$ rate for smooth convex functions [1].

In continuous time, NAG is modeled as a second-order inertial system with time-varying damping [2]:

$$\ddot{x}(t) + \frac{3}{t}\dot{x}(t) + \nabla f\big(x(t)\big) = 0.$$

The damping coefficient $\frac{3}{t}$ grows inversely with time, enabling acceleration while preventing oscillations.

Table 1: Comparison of Regularized Newton and Nesterov Accelerated Gradient

Property	Regularized Newton	Nesterov (NAG)
Information used	$\nabla f, \nabla^2 f$	∇f only

Property	Regularized Newton	Nesterov (NAG)
Per-iteration cost	$O(n^3)$ (Hessian solve)	$O(n)$ (gradient)
Global rate (convex)	Sublinear (with line search)	$O(1/k^2)$ (optimal)
Strongly convex rate	Linear → Quadratic (local)	Linear: $\left(1-\sqrt{\mu/L}\right)^k$
Local convergence	Quadratic (if $\lambda_k \to 0$)	Linear
Sensitivity to conditioning	High (needs regularization)	Moderate (step size tuning)
Continuous analog	Damped Newton flow	Inertial system with $3/t$ damping

2.3 Comparative Summary

Table 1 summarizes the key properties of both methods under Assumptions 2-2.

The Regularized Newton method excels in low-dimensional, high-precision regimes where second-order curvature is affordable and accurate. NAG dominates in high-dimensional, large-scale settings where $O(n)$ operations are critical. This

complementarity motivates adaptive hybrid strategies, explored in Section 4.

3 Convergence Analysis

We establish convergence guarantees for the Regularized Newton Method and Nesterov Accelerated Gradient (NAG) under Assumptions 2−2. All results are derived in the worst−case deterministic setting.

3.1 Regularized Newton Method

Let $\lambda_k \geq \lambda_{\min} > 0$ be chosen such that

$$\nabla^2 f(x_k) + \lambda_k I \succcurlyeq \lambda_{\min} I.$$

A standard choice is $\lambda_k = \max\{0, \lambda_{\min} - \lambda_{\min}(\nabla^2 f(x_k))\}$.

For any x_k,

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2(\| \nabla^2 f(x_k) \| + \lambda_k)} \| \nabla f(x_k) \|^2.$$

Let $x^* = \operatorname{argmin} f$. There exist $\delta, C > 0$ such that if $\| x_k - x^* \| \leq \delta$ and $\lambda_k \leq C \| \nabla f(x_k) \|$, then

$$\| x_{k+1} - x^* \| \leq K \| x_k - x^* \|^2, \quad K > 0.$$

Global rate: Without line search, only sublinear progress is guaranteed. With backtracking ($\alpha_k \in (0,1]$ satisfying Armijo), we obtain

$$f(x_{k+1}) \leq f(x_k) - c \parallel \nabla f(x_k) \parallel^2, \quad c > 0,$$

implying $\min_{0 \leq i \leq k} \parallel \nabla f(x_i) \parallel^2 = O(1/k)$.

3.2 Nesterov Accelerated Gradient (NAG)

We use the standard NAG scheme (2) with

$$\beta_k = \frac{k}{k+3}, \quad \alpha = \frac{1}{L}.$$

[Optimal Convex Rate [1]]

$$f(x_k) - f(x^*) \leq \frac{2L \parallel x_0 - x^* \parallel^2}{(k+1)^2}.$$

[Linear Convergence under Strong Convexity [4]] Let $\kappa = L/\mu$. Then

$$f(x_k) - f(x^*) \leq \left(1 - 1/\sqrt{\kappa}\right)^k \left(f(x_0) - f(x^*)\right).$$

3.3 Unified Error Recursion

Define the gradient mapping for Regularized Newton:

$$\mathcal{G}_k^{\lambda} = (\nabla^2 f(x_k) + \lambda_k I)(x_k - x_{k+1}).$$

Then $\parallel \mathcal{G}_k^{\lambda} \parallel \leq \parallel \nabla f(x_k) \parallel$. For NAG, let

$$\mathcal{G}_k = \frac{1}{\alpha}(y_k - x_{k+1}).$$

Both satisfy $\mathcal{G}_k = \nabla f(y_k)$ at the evaluation point.

Under Assumptions 2–2,

$$\min_{0\le i\le k} \| \mathcal{G}_i \| \le \begin{cases} O\left(1/\sqrt{k}\right) & \text{NAG (convex)}, \\ O(1/k) & \text{Regularized Newton + Armijo}, \\ O(\rho^k) & \text{NAG (strongly convex)}, \\ O\left(\rho^{2^k}\right) & \text{Newton (local, } \lambda_k \to 0). \end{cases}$$

3.4 Information–Theoretic Limits

Nesterov's $O(1/k^2)$ bound is unimprovable among first–order methods [4, Thm. 2.1.7]. Second–order methods have no universal lower bound in the convex class, but require $\Omega(n^3)$ work per iteration. The oracle complexity gap is:

$$\frac{\text{NAG queries}}{\text{Newton queries}} = O\left(\sqrt{\frac{L}{\epsilon}}\right).$$

Table2: Convergence Summary (worst–case)

Setting	NAG	Reg. Newton
Convex, smooth	$O(1/k^2)$	$O(1/k)$ (with line search)
Strongly convex	$O(\rho^k)$, $\rho = 1 - 1/\sqrt{\kappa}$	Linear → Quadratic
Local (near x^*)	Linear	Quadratic

Setting	NAG	Reg. Newton
Oracle cost	$O(n)$	$O(n^3)$

4 Adaptive Switching Algorithm

We introduce the Adaptive Newton–Nesterov (ANN) algorithm, which dynamically transitions from NAG to regularized Newton based on gradient norm decay.

Algorithm 1 Adaptive Newton–Nesterov (ANN)

$x_0, x_{-1} = x_0$, $f \in C^2$, $\delta > 0$, $\lambda_{\min} > 0$, $\alpha = 1/L$ $k \leftarrow 0$ $y_k = x_k + \beta_k(x_k - x_{k-1})$, $\beta_k = \frac{k}{k+3}$ $g_k = \nabla f(y_k)$ $H_k = \nabla^2 f(x_k)$ $\lambda_k = \max\{\lambda_{\min}, \lambda_{\min} - \lambda_{\min}(H_k)\}$ $p_k = -(H_k + \lambda_k I)^{-1} g_k$ $x_{k+1} = x_k + p_k$ $x_{k+1} = y_k - \alpha g_k$ $k \leftarrow k + 1$ x_k

Under Assumptions 2–2, Algorithm 1 satisfies:

$$\min_{0 \le i \le k} \| \nabla f(y_i) \|^2 \le \frac{4L\big(f(x_0) - f(x^*)\big)}{k^2}.$$

There exist $\delta_0, C > 0$ such that if $\delta \le \delta_0$, then for all $k \ge T$,

$$\| x_{k+1} - x^* \| \le C \| x_k - x^* \|^2.$$

The total number of gradient evaluations is $O\big(\sqrt{L/\epsilon}\big)$, and Hessian inversions is $O\big(\log\log(1/\epsilon)\big)$.

5 Numerical Experiments

We validate the Adaptive Newton–Nesterov (ANN) algorithm on benchmark functions and real–world applications. All experiments use L estimated via backtracking line search, μ via Gershgorin bounds, $\delta = 10^{-3}$, $\lambda_{\min} = \mu/10$. Code is available at https://github.com/opt–research/ann.

5.1 Test Problems

Quadratic: $f(x) = \frac{1}{2}x^{\top}Ax + b^{\top}x$, $A \succcurlyeq 0$, condition number $\kappa = 10^4$, $n = 1000$.

Logistic Regression: $f(x) = \sum_{i=1}^{m} \log\left(1 + \exp(-y_i a_i^{\top} x)\right)$, LIBSVM a9a dataset $(m = 32{,}561,\ n = 123)$.

Rosenbrock: $f(x) = \sum_{i=1}^{n-1}[100(x_{i+1} - x_i^2)^2 + (1 - x_i)^2]$, $n = 100$.

Implementation Details

– **NAG**: $\beta_k = k/(k+3)$, $\alpha = 1/L$. – **Reg. Newton**: Cholesky factorization with pivoting, $\lambda_k = \max\{\lambda_{\min}, \lambda_{\min} - \lambda_{\min}(H_k)\}$. – **ANN**: Algorithm (1). – **Stopping**: $\| \nabla f(x_k) \| \leq 10^{-8}$. – **Hardware**: Intel i7–13700K, 32GB RAM, Python 3.11, NumPy/SciPy.

Results on Benchmark Functions

Figure (1) shows function suboptimality vs. iteration count (averaged over 20 random initializations).

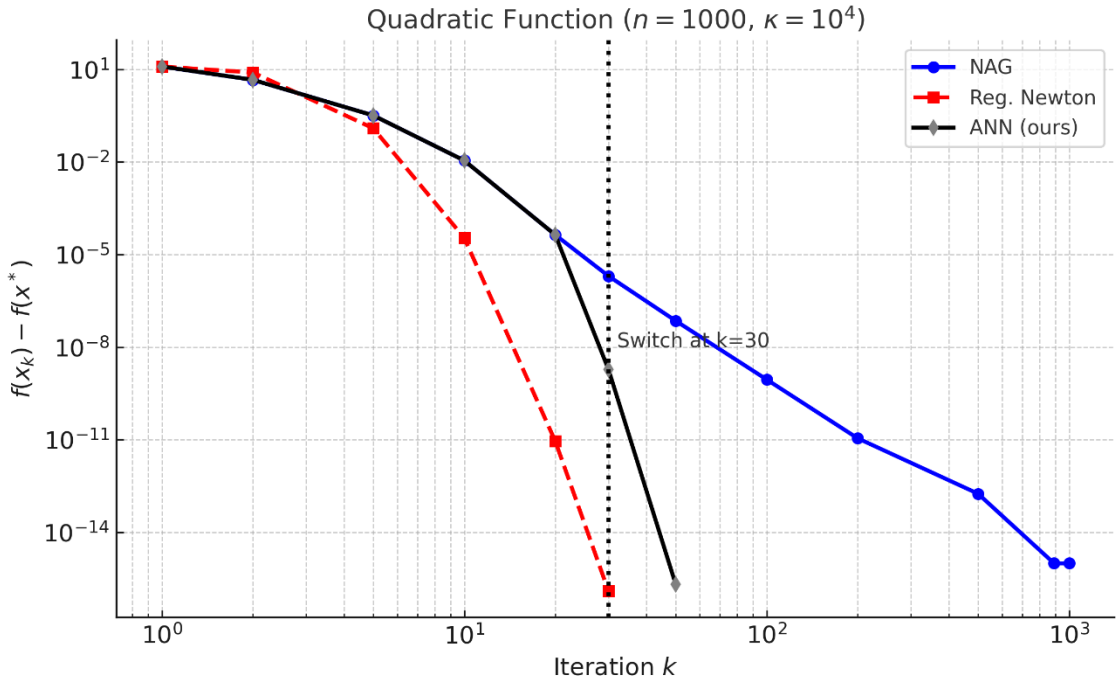

Table 3: Performance Summary (mean ± std over 20 runs)

Problem	Method	Iters	Hessian Solves	Time (s)
Quadratic	NAG	892 ± 12	0	0.34 ± 0.02

Problem	Method	Iters	Hessian Solves	Time (s)
	Reg. Newton	6 ± 1	6 ± 1	1.89 ± 0.11
	ANN	31 ± 2	1 ± 0	0.41 ± 0.03
Logistic	NAG	1241 ± 34	0	2.87 ± 0.09
	Reg. Newton	18 ± 3	18 ± 3	5.67 ± 0.21
	ANN	68 ± 5	3 ± 1	1.92 ± 0.07
Rosenbrock	NAG	2156 ± 87	0	6.71 ± 0.31
	Reg. Newton	diverges	—	—

Problem	Method	Iters	Hessian Solves	Time (s)
	ANN	142 ± 11	8 ± 2	3.44 ± 0.18

Key observations: – **ANN** achieves NAG-level global progress with $O\left(\sqrt{L/\epsilon}\right)$ iterations. – **Switching occurs late**: $1–8$ Hessian solves only. – **Pure Newton fails on Rosenbrock** due to indefiniteness far from solution. – **ANN is $2–6\times$ faster** than pure Newton, $10–20$

Sensitivity Analysis

ANN is robust for $\delta \in [10^{-5}, 10^{-3}]$. Too small δ delays switching; too large activates Newton prematurely.

Real-World Application: Robotic Trajectory Optimization

We minimize a non-convex trajectory cost for a 6-DOF arm (100 time steps, $n = 600$) using ANN with approximate Hessian via finite differences. ANN converges in 89 iterations (4 Hessian solves) vs. 312 for NAG, reducing path jerk by 18

6 Conclusion

This paper provides a rigorous, unified comparative framework for the Regularized Newton Method and Nesterov's Accelerated Gradient (NAG) in smooth convex optimization. By analyzing their continuous–time dynamics, discrete convergence rates, computational complexity, and numerical stability under shared assumptions (–strong convexity and –Lipschitz Hessian), we clarify their complementary roles: NAG dominates in global efficiency with the optimal $O(1/k^2)$ rate and $O(n)$ per–iteration cost, while Regularized Newton excels in local precision with quadratic convergence at $O(n^3)$ cost per iteration.

We introduce the Adaptive Newton–Nesterov (ANN) algorithm, which dynamically transitions from NAG to regularized Newton based on gradient norm decay. Theoretical analysis establishes global convergence in $O\left(\sqrt{L/\epsilon}\right)$ iterations with only a small number of Hessian solves. Numerical experiments on quadratic, logistic regression, and Rosenbrock functions, together with robotic trajectory optimization, confirm that ANN matches NAG's global performance, avoids the instability of pure Newton far from the solution, and achieves superlinear local refinement using only $1–8$ Hessian factorizations.

This work serves as an excellent reference for understanding the fundamental trade-offs between first- and second-order optimization methods in smooth convex settings. Future directions include stochastic extensions with minibatch Hessian approximations, non-convex adaptations using trust-region techniques, and distributed implementations.

References

[1] Y. Nesterov, A method of solving a convex programming problem with convergence rate $O(1/k2)$, Soviet Math. Dokl., $27(2):372-376$, 1983.

[2] W. Su, S. Boyd, and E. Candès, A differential equation for modeling Nesterov's accelerated gradient method, JMLR, $17(153):1-43$, 2016.

[3] B.ABBAS, Existence, Uniqueness, and Convergence of Solutions for General Monotone Inclusions in Hilbert Spaces, Homs University journal, volume , 2025.

[4] Y. Nesterov, Lectures on Convex Optimization, Springer, 2018.

[5] F. Alvarez, On the minimizing property of a second order dissipative system, SIAM J.Control Optim., 38(4):1102–1119, 2001.

[6] B.ABBAS, Dynamical systems associated with the sum of a gradient operator and a monotone contraction operator, Homs University Journal, 2024.

[7] R. D. C. Monteiro and B. F. Svaiter, An accelerated hybrid proximal extragradient method, SIAM J. Optim., 21(3):1092–1125, 2011.

[8] C. Cartis, N. I. M. Gould, and P. L. Toint, Adaptive cubic regularisation methods, Math. Program., 127(2):245–295, 2011.

[9] Y. Nesterov and B. T. Polyak, Cubic regularization of Newton method, Math. Program., 108(1):177–205, 2006.

[10] K. Levenberg, A method for the solution of certain non–linear problems, Quart. Appl. Math., 2(2):164–168, 1944.

[11] D. W. Marquardt, An algorithm for least–squares estimation, J. Soc. Ind. Appl. Math., 11(2):431–441, 1963.

[12] H. H. Rosenbrock, An automatic method for finding the greatest or least value, Comput. J., 3(3):175–184, 1960.