A Differential Equation for Modeling Nesterovs Accelerated Gradient Method

Presenter: Yineng Chen

Department of Mathematics

December 2nd, 2024

Abstract

This paper establishes a connection between Nesterovs Accelerated Gradient Method (NAGM) and a second-order ODE. By deriving this ODE as the continuous-time limit of NAGM, the authors provide deeper insights into the algorithm's dynamics, including its accelerated convergence and oscillatory behavior.

Key contributions include:

1. A rigorous ODE framework for analyzing NAGM.
2. A generalized damping model that extends NAGM to a family of methods.
3. A restarting technique that enhances performance, especially for strongly convex functions.

Su, W., Boyd, S., & Candès, E. J. (2015). A Differential Equation for Modeling Nesterov's Accelerated Gradient Method: Theory and Insights. *arXiv preprint arXiv:1503.01243*.

Nesterov's Accelerated Gradient Method (NAGM)

NAGM Algorithm:

$$\begin{cases} x_k = y_{k-1} - s\nabla f(y_{k-1}) \\ y_k = x_k + \frac{k-1}{k+2}(x_k - x_{k-1}) \end{cases}$$

where $y_0 = x_0$, step size $s \leq \frac{1}{L}$, and L is the Lipschitz constant of ∇f.

inverse quadratic convergence rate:

$$f(x_k) - f^* = \mathcal{O}\left(\frac{\|x_0 - x^*\|^2}{sk^2}\right)$$

Using ODE to model Nesterov's scheme

By taking small step size in NAGM, one can derive an ODE that is the exact limit of Nesterov's scheme:

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

As step size goes to 0, we have $x_k \approx X(k\sqrt{s})$
The initial condition is:

$$X(0) = 0, \ \dot{X}(0) = 0$$

Theorem

For any $f \in \cup_{L>0}\mathcal{F}_L$ ($\mathcal{F}_L$ denotes the class of convex functions f with LLipschitz continuous gradients), as step size $s \to 0$, Nesterov's scheme converges to the ODE above in the sense that for all fixed $T > 0$:

$$\lim_{s \to 0} \max_{0 \leq k \leq T/\sqrt{s}} \|x_k - X(k\sqrt{s})\| = 0$$

Exploring the Link Between Nesterovs Scheme and ODE

- **Objective**: Analyze the approximate equivalence between Nesterovs scheme and its ODE representation.
- **Key Topics**:
  - Convergence equivalence between Nesterovs scheme and ODE.
  - Oscillatory behavior in quadratic and strongly convex functions.
  - Comparison of Nesterovs scheme and gradient descent.

ODE and Nesterovs Scheme: Similar Convergence Rates

Nesterovs Convergence (Discrete):

$$f(x_k) - f^\star \leq \frac{2\|x_0 - x^\star\|^2}{s(k+1)^2}.$$

ODE Convergence:

$$f(X(t)) - f^\star \leq \frac{2\|x_0 - x^\star\|^2}{t^2}.$$

Proven using an energy functional:

$$\mathcal{E}(t) = t^2(f(X(t)) - f^\star) + 2\|X + t\dot{X}/2 - x^\star\|^2.$$

Key Insight: The ODE convergence rate matches Nesterovs scheme for $t \approx k\sqrt{s}$.

Oscillations Explained with Bessel Functions

ODE Solution for Quadratic $f = \frac{1}{2}\langle x, Ax\rangle + \langle b, x\rangle$:

$$\ddot{X}_i + \frac{3}{t}\dot{X}_i + \lambda_i X_i = 0.$$

Solution involves the Bessel function $J_1(t)$:

$$X_i(t) = \frac{2x_{0,i}}{t\sqrt{\lambda_i}} J_1(t\sqrt{\lambda_i}).$$

Asymptotic Form for Large t:

$$J_1(t) \sim \sqrt{\frac{2}{\pi t}} \cos(t - 3\pi/4).$$

Oscillations and decay are explained by this solution.

Oscillation Frequencies for Strongly Convex Functions

Key Insight: Oscillation frequency depends on eigenvalues μ and L:

$$O(\sqrt{\mu}) \leq \text{frequency} \leq O(\sqrt{L}).$$

Root Spacing for Oscillations:

$$t_{i+1} - t_i \sim \frac{\pi}{\sqrt{L}}.$$

This result highlights how strongly convex functions influence the oscillation behavior of the ODE solution.

Why Nesterovs Scheme Moves Faster

Square-Root Scaling:

$$t \approx k\sqrt{s} \quad \text{(Nesterov)} \quad \text{vs.} \quad t \propto ks \quad \text{(Gradient Descent)}.$$

Numerical Stability:

- ODE stable step size: $\Delta t \leq 2/\sqrt{L}$.
- Nesterovs scheme: $s = 1/L$.
- Gradient descent requires $s = 2/L$, slower in practice.

Empirical Comparison: Simulations show that Nesterovs scheme traverses the solution space faster per iteration.

Exploring the Const 3

Overview:

- The constant $r = 3$ in Nesterov's ODE and discrete schemes:

$$\ddot{X} + \frac{r}{t}\dot{X} + \nabla f(X) = 0.$$

- This constant governs the convergence behavior:
  - $r > 3$: High friction, reduced oscillations, maintains $O(1/t^2)$.
  - $r < 3$: Low friction, instability, or slower convergence.
- Goals of this section:
  - Analyze $r > 3$ (high friction).
  - Examine $r < 3$ (low friction).
  - Extend results to strongly convex functions and discrete schemes.

High Friction vs. Low Friction

High Friction ($r > 3$):

- Generalized energy functional:

$$\mathcal{E}(t) = \frac{2t^2}{r-1}(f(X(t)) - f^\star) + (r-1)\|X + \frac{t}{r-1}\dot{X} - x^\star\|^2.$$

- Maintains $O(1/t^2)$ convergence, with a larger constant:

$$f(X(t)) - f^\star \leq \frac{(r-1)^2\|x_0 - x^\star\|^2}{2t^2}.$$

Low Friction ($r < 3$):

- Instability observed with $O(1/t^r)$ convergence for $r < 2$.
- Additional structural assumptions needed for $O(1/t^2)$ convergence:

$$(f - f^\star)^{\frac{r-1}{2}} \text{ must be convex.}$$

Strong Convexity and Improved Convergence

Enhanced Rates for Strongly Convex Functions ($f \in \mathcal{S}_{\mu,L}$):

- New energy functional:

$$\mathcal{E}(t;\alpha) = t^{\alpha}(f(X(t)) - f^{\star}) + \frac{(2r-\alpha)^2 t^{\alpha-2}}{8}\|X + \frac{2t}{2r-\alpha}\dot{X} - x^{\star}\|^2.$$

- For $\alpha = 2r/3$, achieves $O(1/t^{2r/3})$:

$$f(X(t)) - f^{\star} \leq \frac{C\|x_0 - x^{\star}\|^2}{\mu^{\frac{\alpha-2}{2}} t^{\alpha}}.$$

Insights:

- Strong convexity allows faster convergence.
- Highlights the role of $r > 3$ in improving rates for specific problems.

Extending to Discrete Schemes

Generalized Nesterovs Scheme:

- Updates for $r > 3$:

$$x_k = y_{k-1} - sG_s(y_{k-1}), \quad y_k = x_k + \frac{k-1}{k+r-1}(x_k - x_{k-1}).$$

- Key Results:
  - $O(1/k^2)$ for any $r > 3$:

$$f(x_k) - f^\star \leq \frac{(r-1)^2\|x_0 - x^\star\|^2}{2s(k+r-2)^2}.$$

  - $O(1/k^3)$ for $r \geq 9/2$:

$$f(x_k) - f^\star \leq \frac{CL\|x_0 - x^\star\|^2}{k^3}.$$

Numerical Insights:

- Smaller r: Faster initial progress, higher overshoot.
- Larger r: Slower but stable convergence near the solution.

Why Restarting is Necessary?

Challenges with Momentum in Strong Convexity:

- Nesterov's scheme performs worse then vanilla gradient descent in strongly convex function.
- Momentum introduces overshooting, slowing convergence:

$$O(1/\texttt{poly}(k)) \quad \text{vs.} \quad \text{Gradient Method: } O((1-\mu/L)^k).$$

- NAGM can also achieve linear convergence for strongly convex functions but requires knowledge of μ/L, difficult to estimate.

Existing Restarting Approaches:

- **Gradient Restarting:** Restarts when $f(x_{k+1}) > f(x_k)$.
- Effective but lacks theoretical guarantees.

New Proposal: Speed Restarting Scheme

- Maintains high velocity by resetting the trajectory when velocity decreases.
- Provably achieves linear convergence for strongly convex functions.

How Speed Restarting Works

Key Concepts:

- **Speed Restarting Time:** First instance when velocity decreases:

$$T = \sup\{t > 0 : \forall u \in (0, t), \frac{d\|\dot{X}(u)\|^2}{du} > 0\}.$$

- Restart resets $3/t$ in the ODE:

$$\ddot{X}(t) + \frac{3}{t_{sr}}\dot{X}(t) + \nabla f(X(t)) = 0.$$

Linear Convergence Result:

- For $f \in \mathcal{S}_{\mu,L}$, speed restarting achieves:

$$f(X^{sr}(t)) - f^\star \leq \frac{c_1 L\|x_0 - x^\star\|^2}{2} e^{-c_2 t\sqrt{L}}.$$

- Error reduces by a constant factor with each restart.

Numerical Examples: Speed Restarting in Action

Examples:

- **Quadratic:** $f(x) = \frac{1}{2}x^T Ax + b^T x$, A is positive definite.
- **Matrix Completion:** Combines Frobenius norm and nuclear norm regularization.
- **Logistic Regression:** Smooth convex objective with and without ℓ_1-regularization.

Comparison with Other Methods:

- Methods: Speed Restarting (srN), Gradient Restarting (grN), Original Nesterovs Scheme (oN), Proximal Gradient (PG).
- Observations:
  - Speed restarting reduces oscillations and improves stability.
  - Achieves linear convergence empirically, even in non-strongly convex settings.

Discussion and Future Directions

Key Contributions:

- Proposed a second-order ODE framework for analyzing Nesterovs accelerated method.
- Explained oscillations and generalized $O(1/k^2)$ schemes.
- Introduced a speed restarting scheme with linear convergence for strongly convex f.

Future Work:

- Develop a theory linking ODEs to discrete updates to simplify analysis.
- Explore alternative velocity coefficients for new accelerated methods.
- Leverage ODE trajectories (e.g., curvature) for better stopping criteria and adaptive step sizes.