Nesterov's Accelerated Gradient Descent: A Systems Perspective

May 7, 2019

Motivation

Background:

- The convergence of numerical optimization algorithms with discrete steps is often difficult to evaluate directly.
- However if step size $\rightarrow 0$, these problems can often be approximated as continuous dynamics.

Main Idea:

- Leverage well-established, physically intuitive results from systems theory to analyze the performance of optimization algorithms.
- Convergence of these algorithms can be directly interpreted in terms of the stability of a nonlinear system's equilibrium point.

Problem Formulation

Goal: Minimize a convex function $f : \mathbb{R}^n \to \mathbb{R}$ with L-Lipschitz continuous gradients, i.e. find $\min_{x \in \mathbb{R}^n} f(x)$, where $\exists\, L > 0$ such that:

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \ \forall x, y \in \mathbb{R}^n.$$

Nesterov's Gradient Descent Algorithm [1]

- With initial condition $y_0 = x_0$ and step size s, recursively define:

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

- If step size $s \leq 1/L$, this algorithm enjoys the convergence rate:

$$f(x_k) - f^\star \leq O\left(\frac{|x_0 - x^\star|^2}{sk^2}\right)$$

whereas vanilla gradient descent only converges as $O(1/k)$.

Problem Formulation

- **Question:** Can we make sense of this algorithm's convergence by interpreting it as the 'stability' of a dynamical system?

- **Answer:** The exact limit of Nesterov's scheme as the step size $s \to 0$ is given by the second order ODE:

$$\ddot{X}(t) + \frac{3}{t}\dot{X} + \nabla f(X) = 0 \tag{2}$$

  for $t > 0$, and initialized by $X(0) = x_0$ and $\dot{X}(0) = 0$.

- This can be obtained from Nesterov's scheme using the approximate relation $t \approx k\sqrt{s}$ and using a Taylor series expansion (see Theorem 2 below).

Theorems 1 and 2

Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and differentiable, with Lipschitz continuous gradients, and unique global minimum $f(x^\star) \equiv f^\star$.

Theorem 1 (Existence/Uniqueness of Solution to (2) [2])

For any $x_0 \in \mathbb{R}^n$, the ODE (2) with $X(0) = x_0$, $\dot{X}(0) = 0$, has a unique, global C^1 solution $X(t)$.

Theorem 2 (Nesterov's Algorithm Converges to the ODE (2))

As the step size $s \to 0$, Nesterov's scheme (1) converges to the ODE (2), in the sense that for any fixed $T > 0$:

$$\lim_{s \to 0} \max_{0 \le k \le \frac{T}{\sqrt{s}}} \|x_k - X(k\sqrt{s})\| = 0.$$

Sketch of Proof of Theorem 2

Nesterov's Algorithm (1) states that:

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

Rearranging terms and dividing by the discretization step $\sqrt{s}$, we have:

$$\begin{aligned} x_{k+1} &= x_k + \frac{k-1}{k+2}(x_k - x_{k-1}) - s\nabla f(y_k) \\ \Rightarrow \underbrace{\frac{x_{k+1} - x_k}{\sqrt{s}}}_{①} &= \underbrace{\frac{k-1}{k+2}}_{③}\underbrace{\left(\frac{x_k - x_{k-1}}{\sqrt{s}}\right)}_{②} - \sqrt{s}\nabla f(y_k) \end{aligned} \tag{3}$$

Key Idea: Define a smooth $X(t)$ such that $X(n\sqrt{s}) = x_n$ for each $n \in \mathbb{N}$. Then, associate ①, ②, ③ to the ODE (2) with $X(t)$.

Sketch of Proof of Theorem 2, contd.

Equivalently, for each $t \in \mathbb{R}$ that is an integer multiple of $\sqrt{s}$, associate $k \leftrightarrow t/\sqrt{s}$. Then, for ①, ②, ③, respectively, we have:

$$\frac{x_{k+1} - x_k}{\sqrt{s}} = \frac{X(t+\sqrt{s}) - X(t)}{\sqrt{s}} = \dot{X}(t) + \frac{1}{2}\ddot{X}(t)\sqrt{s} + o(\sqrt{s}),$$

$$\frac{x_k - x_{k-1}}{\sqrt{s}} = \frac{X(t) - X(t-\sqrt{s})}{\sqrt{s}} = \dot{X}(t) - \frac{1}{2}\ddot{X}(t)\sqrt{s} + o(\sqrt{s}),$$

$$\frac{k-1}{k+2} = 1 - \frac{3}{k+2} = 1 - \frac{3\sqrt{s}}{t + 2\sqrt{s}}.$$

Sketch of Proof of Theorem 2, contd.

Substituting back into (3), we have:

$$\begin{aligned}
&\dot{X}(t) + \frac{1}{2}\ddot{X}(t)\cdot\sqrt{s} + o(\sqrt{s}) \\
= &\left(1 - \frac{3\sqrt{s}}{t}\right)\left(\dot{X}(t) - \frac{1}{2}\ddot{X}(t)\cdot\sqrt{s}\right) - \sqrt{s}\cdot\nabla f(X(t)) + o(\sqrt{s})
\end{aligned}$$

Finally, we compare coefficients in $\sqrt{s}$ to recover (2):

$$\begin{aligned}
&\frac{1}{2}\ddot{X}(t) = -\frac{3}{t}\dot{X}(t) - \frac{1}{2}\ddot{X}(t) - \nabla f(X(t)), \\
\Rightarrow &\ddot{X}(t) + \frac{3}{t}\dot{X}(t) + \nabla f(X(t)) = 0.
\end{aligned}$$

Theorem 3

Theorem 3 (Convergence rate of the ODE (2))

If $X(t)$ is the unique global solution to ODE (2) with $X(0) = x_0$ and $\dot{X}(0) = 0$, we have that for all $t > 0$:

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

Proof of Theorem 3

From the ODE (2), we have:

$$\ddot{X}(t) + \frac{3}{t}\dot{X} + \nabla f(X) = 0,$$
$$\Rightarrow \frac{t}{2}\ddot{X} + \frac{3}{2}\dot{X} = -\frac{t}{2}\nabla f(X).$$

Consider the energy functional:

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

$$\dot{\mathcal{E}}(t) = 2t(f(X(t)) - f^\star) + t^2\langle \nabla f, \dot{X} \rangle + 4\left\langle X + \frac{t}{2}\dot{X} - x^\star, \frac{3}{2}\dot{X} + \frac{t}{2}\ddot{X} \right\rangle$$

$$= 2t(f(X(t)) - f^\star) + 4\left\langle X - x^\star, -\frac{t}{2}\nabla f(X) \right\rangle$$

$$= 2t\Big[f(X(t)) - f^\star - \langle X - x^\star, \nabla f(X) \rangle \Big]$$

$\dot{\mathcal{E}}(t) \leq 0$ using the convexity of f

Proof of Theorem 3, contd.

From the definition of the energy functional:

$$\begin{aligned}\mathcal{E}(t) &= t^2(f(X(t)) - f^\star) + 2\left\| X + \frac{t}{2}\dot{X} - x^\star \right\|^2 \\ &\geq t^2(f(X(t)) - f^\star)\end{aligned}$$

Since the second term above is non-negative, we have:

$$f(X(t)) - f^\star \leq \frac{\mathcal{E}(t)}{t^2} \leq \frac{\mathcal{E}(0)}{t^2} = \frac{2\|x_0 - x^\star\|^2}{t^2}$$

Nesterov's Algorithm as a Damped Oscillator

Nesterov's Algorithm (1) can be slightly generalized using a constant $r > 0$ in the momentum coefficient as follows:

$$
\begin{aligned}
x_k &= y_{k-1} - s\nabla f(y_{k-1}) \\
y_k &= x_k + \frac{k-1}{k+r-1}(x_k - x_{k-1}).
\end{aligned}
\tag{4}
$$

The corresponding ODE is given as:

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

Note that $r = 3$ in the original Nesterov's scheme.

Nesterov's Algorithm as a Damped Oscillator

$$\ddot{X}(t) + \frac{r}{t}\dot{X} + \nabla f(X) = 0 \tag{5}$$

By viewing (5) as a damped oscillator with damping ratio $\frac{r}{t}$, we see that

- At the start of the algorithm (small t), we have an over damped system that moves towards the origin without oscillating.
- As time progresses, we have an under-damped system that oscillates with amplitude decreasing to zero.
- This explains oscillations in Nesterov's algorithm in later stages.

Faster Convergence with Larger r

If f satisfies a stronger form of convexity, the convergence of the ODE (4) improves. Suppose f is differentiable with L-Lipschitz gradients and μ-strongly convex, i.e. $\exists\, \mu \in \mathbb{R}^+$ such that $f(x) - \frac{1}{2}\mu\|x\|^2$ is convex.

Theorem 4

$\forall\, r \geq 3$, $\exists\, C_r > 0$ such that the solution $X(t)$ to the ODE (5) satisfies:

$$f(X(t)) - f^\star \leq \frac{C_r\|x_0 - x^\star\|^2}{\mu^{\frac{r-3}{3}}} t^{-\frac{2}{3}r}$$

Theorem 5

For $r \geq \frac{9}{2}$, $\exists\, C_r > 0$ such that the generalized Nesterov's Algorithm (4) converges as:

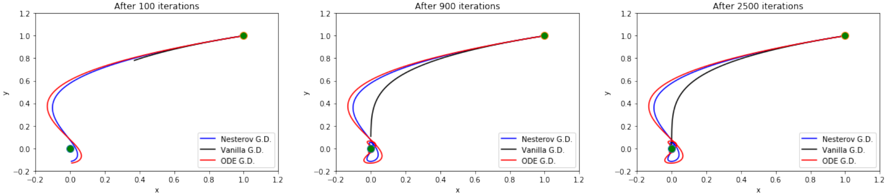
$$f(x_k) - f^\star \leq C_r\sqrt{\frac{L^3}{\mu}}\frac{\|x_0 - x^\star\|^2}{k^3}$$

Experiments: 2D quadratic cost function

Objective function:

$$f(x) = 0.02x_1^2 + 0.005x_2^2$$

Log Error $f - f^{\star}$ during Nesterov's Gradient Descent Algorithm:



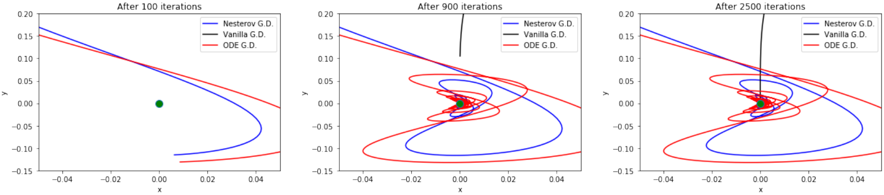
(a) After 100 iterations (b) After 900 iterations (c) After 2500 iterations

Figure 1: Trajectory of (x_1, x_2)

Objective function:

$$f(x) = 0.02x_1^2 + 0.005x_2^2$$

Log Error $f - f^\star$ during Nesterov's Gradient Descent Algorithm:



(a) After 100 iterations (b) After 900 iterations (c) After 2500 iterations

Figure 2: Trajectory of (x_1, x_2), Closeup

Consider the following objective function:

$$f(x) = 0.02x_1^2 + 0.005x_2^2 + 0.0001x_3^2$$

Log Error $f - f^\star$ during Nesterov's Gradient Descent ($r = 3$) Algorithm:

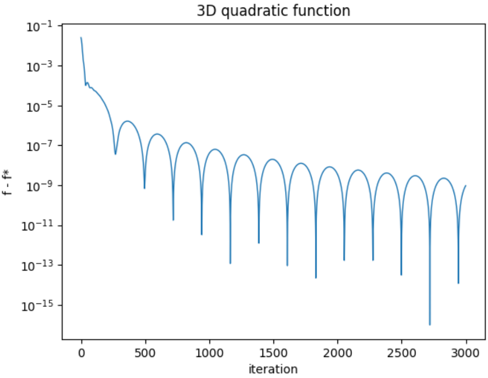
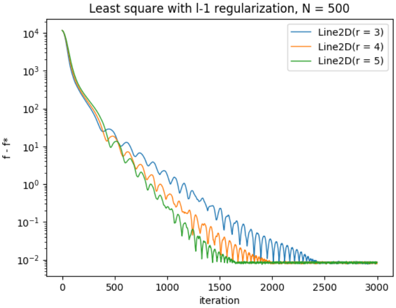


Figure 3: $log(f - f^\star)$ over iteration

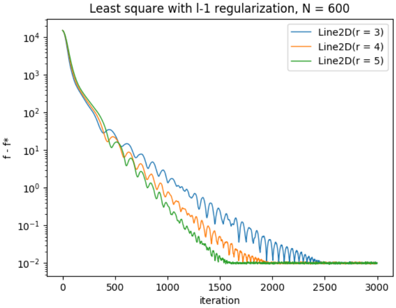
Objective function:

$$f(x) = \frac{1}{2}\|Ax - b\|_2^2 + \gamma\|x\|_1$$

Log Error $f - f^{\star}$ during Nesterov's Gradient Descent Algorithm:



(a) $A \in \mathbf{R}^{500\times 500}$, $b \in \mathbf{R}^{500\times 1}$



(b) $A \in \mathbf{R}^{600\times 600}$, $b \in \mathbf{R}^{600\times 1}$

Figure 4: $log(f - f^{\star})$ over iteration

Conclusion

- Acceleration methods in optimization can help machine learning algorithms converge faster.
- In particular, the Nesterov's gradient descent scheme enjoys a convergence rate of $O(1/k^2)$ for convex functions f as compared to a rate of $O(1/k)$ with vanilla gradient descent.
- Continuous time ODEs and systems theory can be used to understand and justify such behaviour of gradient descent algorithms on convex functions.
- Energy functionals ('Lyapunov functions') were used to deduce these convergence rates.
- We confirm the expected performance of Nesterov's scheme on commonly used loss functions in machine learning such as LASSO, regularized least squares, etc.

[1] Y. E. Nesterov. "A method for solving the convex programming problem with convergence rate $O(1/k^2)$". In: *Dokl. Akad. Nauk SSSR* 269 (1983), pp. 543–547. URL: https://ci.nii.ac.jp/naid/10029946121/en/.

[2] Weijie Su, Stephen Boyd, and Emmanuel J. Candes. "A Differential Equation for Modeling Nesterov's Accelerated Gradient Method: Theory and Insights". In: *Journal of Machine Learning Research* 17.153 (2016), pp. 1–43. URL: http://jmlr.org/papers/v17/15-084.html.

Thank You!

Questions?

Appendix: Proof of Theorem 4

- **Proof:** Analogous to Theorem 3, with modified energy functional:

$$\mathcal{E}(t;r) = t^{\frac{2r}{3}}\left(f(X(t)) - f^\star\right) + \frac{2}{9}r^2 t^{\frac{2r-6}{3}}\left\|X(t) + \frac{3t}{2r}\dot{X}(t) - x^\star\right\|^2$$

- **Key Idea:** Increased damping can lead to higher convergence rate.
- Note that the constant $C_r > 0$ in Theorem 4 grows with r. Therefore, simply increasing r may not guarantee higher convergence rate.
- Nonetheless for $r \geq 9/2$, Theorem 4 guarantees an $O(1/k^3)$ convergence rate for the Nesterov's Generalized Algorithm (4).