Continuous-Time Analysis of AGM via Conservation Laws in Dilated Coordinate Systems

Jaewook J. Suh[1], Gyumin Roh[1], **Ernest K. Ryu**[1]

Long Presentation
International Conference on Machine Learning, 2022

[1] Department of Mathematical Sciences, Seoul National University

Acceleration of first-order convex minimization

Consider

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad f(x)$$

where f is L-smooth convex. Gradient descent

$$x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$$

converge with the rate $f(x_k) - f_\star \le \mathcal{O}(1/k)$.
Nesterov's celebrated accelerated gradient method (AGM)

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

converges with the accelerated rate $f(x_k) - f_\star \le \mathcal{O}(1/k^2)$.

Question) How does acceleration work?

Continuous-time model of AGM

To gain insight into Nesterov's acceleration, Su et al. analyzed a continuous time model of AGM in the limit of small stepsizes:

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

with initial condition $X(0) = X_0$, $\dot{X}(0) = 0$.

They defined the Lyapunov function

$$\Phi(t) = t^2 \left(f(X) - f_\star\right) + \frac{1}{2}\left\|t\dot{X} + 2(X - X_\star)\right\|^2$$

and performed direct calculations to show that $\frac{d}{dt}\Phi(t) \leq 0$. This leads to

$$t^2 \left(f(X) - f_\star\right) \leq \Phi(t) \leq \Phi(0) = 2\left\|X_0 - X_\star\right\|^2,$$

and dividing both sides by t^2 yields an $\mathcal{O}(1/t^2)$ rate.

Su, Boyd, and Candes, A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights, *NeurIPS*, 2014.

Continuous-time model of AGM

However, the element of mystery remains. Where does

$$\Phi(t) = t^2 \left(f(X) - f_\star \right) + \frac{1}{2} \left\| t\dot{X} + 2(X - X_\star) \right\|^2$$

come from? What do the terms mean?

This work addresses this question.

Outline

Conservation laws in dilated coordinates

Semi-second-order symplectic Euler discretization in dilated coordinates

Conservation laws of AGM ODE

Interestingly, the ODE

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

exhibits many conservation laws such as

$$E \equiv f(X) - f_\star + \frac{1}{2}\left\|\dot{X}\right\|^2 + \int_0^t \frac{3}{s}\|\dot{X}\|\, ds$$

and

$$\begin{aligned} E \equiv\, & t^2\left(f(X) - f_\star\right) + \frac{1}{2}\left\|t\dot{X} + 2(X - X_\star)\right\|^2 \\ & + \int_0^t 2s \underbrace{\left(f_\star - f\left(X\right) - \langle \nabla f(X), X_\star - X\rangle\right)}_{\geq 0 \text{ by convexity of } f} ds \end{aligned}$$

(E is independent of time.) Where do these come from?

Energy and conservation law

Let $A\colon (0,\infty) \to \mathbb{R}$ be differentiable and $B\colon (0,\infty) \to \mathbb{R}$ be integrable. Suppose

$$0 = \dot{A}(t) + B(t).$$

Integrating from 0 to t gives us the *conservation law*

$$E \equiv A(0) = A(t) + \int_0^t B(s)\,ds,$$

where the *energy* E is independent of time.

First conservation law

Multiply $\dot{X}$ to both sides of

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

and integrate from 0 to t to get

$$E \equiv f(X) - f_\star + \frac{1}{2}\left\|\dot{X}\right\|^2 + \int_0^t \frac{3}{s}\|\dot{X}\|\,ds$$

This was our first conservation law.

First conservation law: Calculation steps

Starting from

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

multiply $\dot{X}$ to both sides

$$0 = \langle \dot{X}, \ddot{X} \rangle + \frac{3}{t}\|\dot{X}\|^2 + \langle \nabla f(X), \dot{X} \rangle$$

and integrate from 0 to t to get

$$0 = \frac{1}{2}\|\dot{X}\|^2\Big|_0^t + \int_0^t \frac{3}{s}\|\dot{X}\|^2 \, ds + (f(X) - f_\star)\Big|_0^t$$

Reorganizing, we get

$$E \equiv f(X) - f_\star + \frac{1}{2}\left\|\dot{X}\right\|^2 + \int_0^t \frac{3}{s}\|\dot{X}\| \, ds.$$

Second conservation law via dilated coordinates

How about other conservation laws? Use a change of variables!

Consider the *dilated coordinate* $W = t^2(X - X_\star)$. ODE becomes

$$0 = \frac{1}{t^2}\ddot{W} - \frac{1}{t^3}\dot{W} + \nabla_W U(W,t)$$

with time-dependent potential

$$U(W,t) = t^2\left(f\left(X(W,t)\right) - f_\star\right).$$

Multiply $\dot{W}$ to both sides and integrate from 0 to t to get

$$\begin{aligned} E \equiv t^2\left(f(X) - f_\star\right) + \frac{1}{2}\left\|t\dot{X} + 2(X - X_\star)\right\|^2 \\ + \int_0^t 2s\Big(f_\star - f\left(X\right) - \langle \nabla f(X), X_\star - X\rangle\Big)\, ds. \end{aligned}$$

Second conservation law: Calculation step

Starting from

$$0 = \frac{1}{t^2}\ddot{W} - \frac{1}{t^3}\dot{W} + \nabla_W U(W,t)$$

multiply $\dot{W}$ to both sides

$$\begin{aligned} 0 &= \frac{1}{t^2}\langle \dot{W}, \ddot{W}\rangle - \frac{1}{t^3}\|\dot{W}\|^2 + \langle \nabla_W U(W,t), \dot{W}\rangle \\ &= \frac{d}{dt}\left(\frac{1}{2t^2}\|\dot{W}\|^2 + U(W,t)\right) - \frac{\partial}{\partial t}U(W,t) \end{aligned}$$

and integrate from 0 to t to get

$$\begin{aligned} 0 &= \left(\frac{1}{2t^2}\|\dot{W}\|^2 + U(W,t)\right)\Big|_0^t - \int_0^t \frac{\partial}{\partial s}U(W(s),s)ds \\ &= \left(\frac{1}{2}\left\|t\dot{X} + 2(X - X_\star)\right\|^2 + t^2\left(f(X) - f_\star\right)\right)\Big|_0^t + \int_0^t 2s\big(f_\star - f\left(X\right) - \langle \nabla f(X), X_\star - X\rangle\big)\, ds \end{aligned}$$

Reorganizing, we get the stated conservation law.

Analysis recipe

Central thesis: Continuous-time analyses of accelerated gradient methods significantly simplify under an alternate dilated coordinate system.

Proposed recipe for continuous-time analysis:

(i) Change the ODE into dilated coordinates.

(ii) Obtain conservation law in the dilated coordinates.

Connection with Lyapunov analyses

Our analyses based on conservation laws are not fundamentally different from the Lyapunov analyses of prior work. The first two terms of the second conservation law for the AGM ODE

$$\Phi(t) = t^2 \left(f(X) - f_\star \right) + \frac{1}{2} \left\| t\dot{X} + 2(X - X_\star) \right\|^2 ,$$

is the Lyapunov function of prior work.

Once $\Phi(t)$ is stated, verifying $\dot{\Phi}(t) \leq 0$ is relatively straightforward. A core motivation of our work is to provide a systematic methodology for obtaining such Lyapunov functions.

AGM ODE $r > 3$

Consider

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

with $r > 3$. With $W = t^2(X - X_\star)$, we get

$$\begin{aligned} E \equiv & -2(r-3)\left\|X_0 - X_\star\right\|^2 + t^2\left(f(X) - f_\star\right) + \frac{1}{2}\left\|t\dot{X} + 2(X - X_\star)\right\|^2 \\ & + (r-3)\left\|X - X_\star\right\|^2 + \int_0^t \frac{r-3}{s}\left\|s\dot{X}\right\|^2 ds \\ & + \int_0^t 2s\left(f_\star - f(X) - \langle \nabla f(X), X_\star - X\rangle\right) ds. \end{aligned}$$

Reorganizing terms, we conclude

$$f(X) - f_\star \leq \frac{(r-1)\left\|X_0 - X_\star\right\|^2}{t^2}.$$

Rate improves upon (Su et al. 2014) and matches (Attouch et al. 2018).

Su, Boyd, and Candes, A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights, *NeurIPS*, 2014.

Attouch, Chbani, Peypouquet, and Redont, Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity. *MPA*, 2018.

AGM ODE $r < 3$

Consider

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

with $r < 3$. With $W = t^{2r/3}(X - X_\star)$, we get

$$E \equiv t^{\frac{2r}{3}}\left(f\left(X\right) - f_\star\right) + \frac{r(3-r)}{9}t^{\frac{2r}{3}-2}\|X - X_\star\|^2 + \frac{1}{2}t^{\frac{2r}{3}-2}\left\|t\dot{X} + \frac{2r}{3}(X - X_\star)\right\|^2$$

$$+ \int_{t_0}^{t} \frac{2}{27}r(3-r)(3+r)s^{\frac{2r}{3}-3}\left\|X - X_\star\right\|^2\, ds$$

$$+ \int_{t_0}^{t} \frac{2r}{3}s^{\frac{2r}{3}-1}\left(f_\star - f(X) - \langle \nabla f(X), X_\star - X\rangle\right)\, ds.$$

Reorganizing terms, we conclude

$$f(X) - f_\star \leq \frac{E}{t^{\frac{2r}{3}}}.$$

This recovers the rate of (Attouch et al. 2019).

Attouch, Chbani, and Riahi, Rate of convergence of the Nesterov accelerated gradient method in the subcritical case $\alpha \leq 3$, *ESAIM: COCV*, 2019.

Strongly convex AGM

Consider

$$0 = \ddot{X} + 2\sqrt{\mu}\dot{X} + \nabla f(X)$$

with initial condition $X(0) = X_0$, $\dot{X}(0) = 0$, which models Nesterov's AGM in the μ-strongly convex setup. With $W = e^{\sqrt{\mu}t}(X - X_\star)$, we get

$$E \equiv -\frac{\mu}{2}\left\|X_0 - X_\star\right\|^2 + e^{\sqrt{\mu}t}\left(f(X) - f_\star + \frac{1}{2}\left\|\dot{X} + \sqrt{\mu}(X - X_\star)\right\|^2\right)$$
$$+ \int_0^t \frac{\sqrt{\mu}e^{\sqrt{\mu}s}}{2}\left\|\dot{X}\right\|^2 ds + \int_0^t \sqrt{\mu}e^{\sqrt{\mu}s}(\dots)\, ds,$$

where

$$(\dots) = f_\star - f(X) - \langle \nabla f(X), X_\star - X\rangle - \frac{\mu}{2}\left\|X - X_\star\right\|^2 \geq 0.$$

Reorganizing terms, we conclude

$$f(X) - f_\star \leq e^{-\sqrt{\mu}t}\left(f(X_0) - f_\star + \frac{\mu}{2}\left\|X_0 - X_\star\right\|^2\right).$$

This recovers the rate of (Wilson et al. 2021).

Wilson, Recht, and Jordan, A Lyapunov analysis of accelerated methods in optimization, *JMLR*, 2021.

Gradient flow

Consider

$$0 = \dot{X} + \nabla f(X)$$

with $X(0) = X_0$, which models gradient descent. With $W = t(X - X_\star)$, we get

$$E \equiv t\left(f(X) - f_\star\right) + \frac{1}{2}\left\|X - X_\star\right\|^2 - \left\|X_0 - X_\star\right\|^2$$
$$+ \int_0^t s\left\|\dot{X}\right\|^2 ds + \int_0^t \left(f_\star - f(X) - \langle \nabla f(X), X_\star - X\rangle\right)\, ds.$$

Reorganizing terms, we conclude the well known rate

$$f(X) - f_\star \leq \frac{\left\|X_0 - X_\star\right\|^2}{2t}.$$

OGM-G

Consider

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

for $t \in (0,T)$ with initial value $X(0) = X_0$, $\dot{X}(0) = 0$. This models the OGM-G.[2] With $W = t^{-2}(X - X(T))$, we get

$$\begin{aligned} E \equiv\ & \frac{2}{(T-t)^2}\left(f(X) - f(X(T))\right) - \frac{2}{(T-t)^4}\left\|X - X(T)\right\|^2 \\ &+ \frac{1}{2(T-t)^4}\left\|(T-t)\dot{X} + 2(X - X(T))\right\|^2 \\ &+ \int_0^t \frac{4}{(T-s)^3}\left(f(X(T)) - f(X) - \langle \nabla f(X), X(T) - X\rangle\right) ds. \end{aligned}$$

Reorganizing terms and using L'Hôpital's rule, we conclude the new rate

$$\frac{1}{2}\left\|\nabla f(X(T))\right\|^2 \le \frac{2}{T^2}\left(f(X_0) - f(X(T))\right).$$

[2] Kim and Fessler. Optimizing the efficiency of first-order methods for decreasing the gradient of smooth convex functions. *JOTA*, 2021.

Outline

Conservation laws in dilated coordinates

Semi-second-order symplectic Euler discretization in dilated coordinates

Discrete → continuous → discrete

So far, we studied continuous-time models of discrete-time algorithms. Conversely, can we obtain a discrete-time algorithm from the continuous-time analysis?

This has been surprisingly difficult. Prior continuous → discrete attempts "do not flow natural from the dynamical-systems framework."[3]

Using dilated coordinates, we provide a direct discretization scheme achieving an accelerated $\mathcal{O}(1/k^2)$ rate.

[3] Jordan, Dynamical, symplectic and stochastic perspectives on gradient-based optimization, *In International Congress of Mathematicians*, 2018.

ODE in dilated coordinates and conjugate momentum

Again, consider $0 = \ddot{X} + \frac{3}{t}\dot{X} + \nabla f(X)$. With $W = t^2(X - X_\star)$, the ODE becomes

$$0 = \frac{1}{t^2}\ddot{W} - \frac{1}{t^3}\dot{W} + \nabla_W U(W,t).$$

We can equivalently express this ODE as

$$\begin{aligned}\dot{P} &= -t\nabla f(X(W,t))\\ \dot{W} &= tP\end{aligned}$$

where $P = t\dot{X} + 2(X - X_\star)$ is the conjugate momentum pair of W in the sense of Lagrangian and Hamiltonian mechanics.

Semi-second-order symplectic Euler discretization

Discretize with alternating updates of W and P but use a second-order update for W:

$$
\begin{aligned}
P(t+h) &\approx P(t) - t\nabla f(X)h \\
W(t+h) &\approx W(t) + \dot{W}(t)h + \ddot{W}(t)\frac{h^2}{2} \\
&= W(t) + tP(t)h + \big(P(t) - t^2\nabla f(X(W,t))\big)\frac{h^2}{2}.
\end{aligned}
$$

We refer to this discretization as a semi-second-order symplectic Euler.

Reorganizing, letting $s = h^2$, $\theta_k = \frac{k}{2}$ and $z_k = \frac{p_k}{2} + X_\star$, we get

$$
\begin{aligned}
x_k^+ &= x_k - \frac{s}{2}\nabla f(x_k) \\
z_{k+1} &= z_k - s\theta_k \nabla f(x_k) \\
x_{k+1} &= \frac{\theta_k^2}{\theta_{k+1}^2} x_k^+ + \left(1 - \frac{\theta_k^2}{\theta_{k+1}^2}\right) z_{k+1}
\end{aligned}
$$

Accelerated $\mathcal{O}(1/k^2)$ discrete rate

Discretized method

$$
\begin{aligned}
x_k^+ &= x_k - \frac{s}{2}\nabla f(x_k) \\
z_{k+1} &= z_k - s\theta_k \nabla f(x_k) \\
x_{k+1} &= \frac{\theta_k^2}{\theta_{k+1}^2} x_k^+ + \left(1 - \frac{\theta_k^2}{\theta_{k+1}^2}\right) z_{k+1}
\end{aligned}
$$

with $\theta_k = \frac{k}{2}$, exhibits an accelerated rate.

Theorem

Assume f is convex and L-smooth. Assume f has a minimizer $X_\star$. For $s \in \left(0, \frac{2}{L}\right]$, the discretized method exhibits the rate

$$
f(x_k^+) - f_\star \le \frac{2\left\|X_0 - X_\star\right\|^2}{sk^2}.
$$

Interpreting z_k as conjugate momentum

(Lee et al. 2021) point out that many known accelerated gradient methods have an auxiliary z_k-sequence satisfying a geometric structure.

This mysterious quantity plays a crucial role in the convergence analysis.

We identify that z_k is (up to a factor-2 scaling and translation with $X_\star$) the conjugate momentum $P = \dot{W}/t = t\dot{X} + 2(X - X_\star)$ of the dilated coordinate $W = t^2(X - X_\star)$.

Lee, Park, and Ryu, A geometric structure of acceleration and its role in making gradients small fast. *NeurIPS*, 2021.

Conclusion

We present a new methodology for analyzing continuous-time models of accelerated gradient methods through deriving conservation laws in dilated coordinate systems.