Rethinking the Variational Interpretation of Accelerated Optimization Methods

Peiyuan Zhang[1], Antonio Orvieto[1], Hadi Daneshmand[1,2]

1: ETH Zurich, Switzerland 2: INRIA Paris, France

Acceleration in the Continuous Domain

For the smooth minimization of f, the Nestrov path (a.k.a. AGD-ODE) is the continuous limit of his famous accelerated gradient method:

$$\ddot{X} + a(t)\dot{X} + \nabla f(X) = 0 \qquad \text{(AGD-ODE)}$$

Assume f is L-Lipschitz smooth, AGD-ODE is referred as an accelerated flow under different damping choice and geometry:

- $f(X(t)) - f(x^*) \leq \mathcal{O}(1/t^2)$ for convex f and $\alpha(t) = 3/t$;
- $f(X(t)) - f(x^*) \leq \mathcal{O}(e^{-\sqrt{\mu}t})$ for μ-strongly convex f and $\alpha = 2\sqrt{\mu}$.

Variational Formulation of Nestrov's Path

The AGD-ODE can be seen as solution to the famous Euler-Lagrange equation in *variational calculus*

$$\frac{d}{dt}\left(\frac{\partial}{\partial \dot{X}} L(X, \dot{X}, t)\right) = \frac{\partial}{\partial X} L(X, \dot{X}, t) \qquad \text{(E-L)}$$

with respect to the time-dependent Lagrangian

$$L(X, \dot{X}, t) = m(t)\left(\|\dot{X}\|/2 - f(X)\right)$$

where $a(t) = m'(t)/m(t)$.

Therefore, the variational formulation in [Wibisono et al., 2016] conjectures AGD-ODE to be a solution to variational problem

$$\min_{y \in \mathcal{C}^1([t_1,t_2],\mathbb{R}^d)} J[Y] := \int_{t_1}^{t_2} L(Y, \dot{Y}, t)dt \qquad \text{(VarP)}$$

by the least action principle.

First and Second Order Variation: Necessary and Sufficient Condition for Optimality

From a perspective of mathematical rigorousness:

- Solving E-L equation does **not** guarantee minimality to (VarP);
- First-order variation condition like E-L is only the **necessary** condi-

Jacobi Condition and Conjugate Points

Jacobi condition provides better and sufficient criteria for the minimality of (VarP).

Proposition 2 [Jacobi condition]. Sufficient conditions for Y to be a minimum for J are: (1) Y satisfy the E-L; (2) $P \succ 0$; (3) (t_1, t_2) contains no points conjugate to t_1.

A point $t \in (t_1, t_2)$ is said to be **conjugate** to t_1 w.r.t J if Jacobi equation

$$d(Ph')/dt - Qh = 0 \qquad \text{(Jacobi equation)}$$

admits a non-trivial solution $h(t) \in \mathcal{C}^1([t_1, t_2], \mathbb{R}^d)$ that vanishes at both t and t_1, where $P = L_{\dot{Y}\dot{Y}}$ and $Q = L_{YY} - dL_{\dot{Y}Y}/dt$.

Nesterov's Path is Saddle with Decreasing Damping

To study the optimality of AGD-ODE with decreasing damping $a(t) = 3/t$, we consider one-dimensional quadratic $f(x) = \beta x^2/2$.

- Applying second order variation, Jacobi's equation reduces to

$$h''(t) + \frac{3}{t}h'(t) + \beta h(t) = 0, \quad h(t_1) = 0.$$

- Points conjugate to t_1 satisfies $h(t) = 0$, which results in identity

$$\mathcal{Y}_1(\sqrt{\beta}\, t) = K_{\beta,t_1}\mathcal{J}_1(\sqrt{\beta}\, t)$$

where K_{β,t_1} is a constant parameterized by curvature β and t_1, $\mathcal{J}$ and $\mathcal{Y}$ are first and second kind Bessel functions.

- Oscillating nature of Bessel function guarantees the existence of conjugate points.

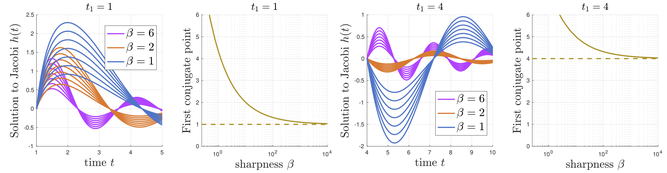

Figure 1: Smallest conjugate pts to $t_1 = 1, 4$ under $f(x) = \beta x^2/2$.

Path is also Saddle for Constant Damping that Accelerates

To study the optimality of AGD-ODE with constant damping α, we consider multi-dimensional quadratic $f(x) = x'Hx/2$ with $\lambda_{\max}(H) = \beta$, $\lambda_{\min}(H) = \mu$.

- Underdamping ($\alpha < 2\sqrt{\mu}$): Jacobi equation admits conjugate point for interval $[t_1, t_2]$ with $|t_2 - t_2| > 2\pi/\sqrt{4\beta - \alpha^2}$. Therefore, AGD-ODE is **saddle** and also **accelerated**.
- Over- and critical damping ($\alpha \geq 2\sqrt{\beta}$): Jacobi equation admits no conjugate points, the **minimality** of Nesterov's path is indeed guaranteed. But the ODE with such high damping is **not** accelerated.

Oscillation, Damping and Acceleration

- The suboptimality is *due precisely to the oscillations*. In contrast, if each coordinate decreases monotonically, Nesterov's path is optimal.
- Very high damping indeed guarantees mininality, but also avoids oscillation, therefore does not lead to acceleration.

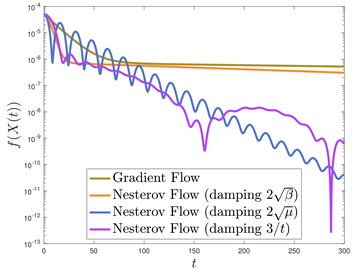

Figure 2 : Non-monotonic trajectories (i.e. the accelerated curves) minimize the action only for short time intervals.

Conclusion: No Optimal Path through the Variational Problem

- Locally Nesterov's path might optimize (VarP). But the property does not hold for complicated geometry and becomes saddle.