Conformal Symplectic and Relativistic Optimization

Guilherme França

Johns Hopkins University
University of California, Berkeley

NeurIPS 2020

Accelerated Optimization Methods

Acceleration has a fundamental importance in optimization and machine learning. It is at the core of modern applications (e.g. deep learning). However, "acceleration phenomena" are not well-understood, e.g. the reason why such methods have a faster convergence remains unknown. An underlying principle to construct accelerated algorithms is also unknown.

Accelerated Optimization Methods

Acceleration has a fundamental importance in optimization and machine learning. It is at the core of modern applications (e.g. deep learning). However, "acceleration phenomena" are not well-understood, e.g. the reason why such methods have a faster convergence remains unknown. An underlying principle to construct accelerated algorithms is also unknown.

Consider a smooth optimization problem:

$$\min_{x \in \mathbb{R}^n} f(x). \tag{1}$$

The two most important algorithms are the Heavy Ball or classical momentum (CM) method,

$$v_{k+1} = \mu v_k - \epsilon \nabla f(x_k), \qquad x_{k+1} = x_k + v_{k+1}, \tag{2}$$

and Nesterov's accelerated gradient (NAG) method,

$$v_{k+1} = \mu v_k - \epsilon \nabla f(x_k + \mu v_k), \qquad x_{k+1} = x_k + v_{k+1}. \tag{3}$$

The only difference between them is the point inside the gradient.

Connection with Continuum Systems

Remarkably, these two optimization algorithms are 1st order integrators to the following ODE:

$$m\ddot{x}(t) + m\gamma\dot{x}(t) = -\nabla f(x(t)). \tag{4}$$

r-th order means $\|x_k - x(t)\| = O(h^{r+1})$.

This a special kind of system, i.e. it is a *Conformal Hamiltonian System* of the general form

$$\dot{x} = -\nabla_p H(x,p), \qquad \dot{p} = -\nabla_x H(x,p) - \gamma p. \tag{5}$$

For the above, $H = \frac{1}{2m}\|p\|^2 + f(x)$. (Conformal) Hamiltonian systems are ubiquitous in physics and have a special mathematical structure.

Conformal Hamiltonian Systems

The most fundamental property of conformal Hamiltonian systems is the contraction of the symplectic form:

$$\omega(t) \equiv dx(t) \wedge dp(t) \quad \Longrightarrow \quad \boxed{\omega(t) = e^{-\gamma t}\omega(0).} \tag{6}$$

- The phase space is a conformal symplectic manifold.
- The flow composition has a (conformal) group structure.
- This property is related to stability and convergence rate.
- Conformal Symplectic Integrator: it is a structure-perserving discretization, namely one such that

$$\omega_{k+1} = e^{-\gamma h}\omega_k \qquad (h > 0). \tag{7}$$

  This ensures that numerical trajectories lies on the same (conformal symplectic) manifold as the continuum system. As a consequence, the *phase portrait, stability, and convergence rates* are all preserved.

CM and NAG from a Symplectic Perspective

In the paper we constructed two general symplectic integrators (1st and 2nd order) whose details are not important here. We then show that:

Theorem (CM is symplectic)

CM is a 1st order integrator. Moreover, it turns out to be conformal symplectic:

$$\omega_{k+1} = e^{-\gamma h}\omega_k \qquad (\mu \equiv e^{-\gamma h}). \tag{8}$$

CM and NAG from a Symplectic Perspective

In the paper we constructed two general symplectic integrators (1st and 2nd order) whose details are not important here. We then show that:

Theorem (CM is symplectic)

CM is a 1st order integrator. Moreover, it turns out to be conformal symplectic:

$$\omega_{k+1} = e^{-\gamma h}\omega_k \qquad (\mu \equiv e^{-\gamma h}). \tag{8}$$

Theorem (NAG is not symplectic)

NAG is also a 1st order integrator. However it is not conformal symplectic:

$$\omega_{k+1} = e^{-\gamma h}\left[I - \frac{h^2}{m}\nabla^2 f(x_k)\right]\omega_k + O(h^3). \tag{9}$$

Shadow Dynamical Systems

We thus see that NAG introduces some spurious dissipation. To describe this more precisely, we ask: *for which dynamical system CM or NAG turns out to be a 2nd order integrator?* The answer is as follows.

Theorem

CM is a 2nd order integrator to the perturbed system

$$m\ddot{x} + m\gamma\dot{x} = -\left[I + \tfrac{h\gamma}{2} I - \tfrac{h^2\gamma^2}{4} I - \tfrac{h^2}{4m}\nabla^2 f(x)\right]\nabla f(x). \tag{10}$$

Shadow Dynamical Systems

We thus see that NAG introduces some spurious dissipation. To describe this more precisely, we ask: *for which dynamical system CM or NAG turns out to be a 2nd order integrator?* The answer is as follows.

Theorem

CM is a 2nd order integrator to the perturbed system

$$m\ddot{x} + m\gamma\dot{x} = -\left[I + \tfrac{h\gamma}{2}I - \tfrac{h^2\gamma^2}{4}I - \tfrac{h^2}{4m}\nabla^2 f(x)\right]\nabla f(x). \tag{10}$$

Theorem

NAG is a 2nd order integrator to the perturbed system

$$m\ddot{x} + \left[m\gamma + h\nabla^2 f(x)\right]\dot{x} = -\left[I + \tfrac{h\gamma}{2}I - \tfrac{h^2\gamma^2}{4}I + \tfrac{h^2}{4m}\nabla^2 f(x)\right]\nabla f(x). \tag{11}$$

This captures dependence on the step size and the "geometry" of $f(x)$.

Relativistic Gradient Descent (RGD)

Can we derive new optimization algorithms inspired by physical systems?

Consider a dissipative relativistic system $\left(H = c\sqrt{\|p\|^2 + m^2c^2} + f(x)\right)$:

$$\dot{x} = \frac{cp}{\sqrt{\|p|^2 + m^2c^2}}, \qquad \dot{p} = -\nabla f(x) - \gamma p. \tag{12}$$

After discretizing with our general 2nd order method—and a change of variables—we get the following solver:

$$x_{k+1/2} = x_k + \sqrt{\mu} v_k \Big/ \sqrt{\mu\delta\|v_k\|^2 + 1}, \tag{13a}$$

$$v_{k+1/2} = \sqrt{\mu} v_k - \epsilon \nabla f(x_{k+1/2}), \tag{13b}$$

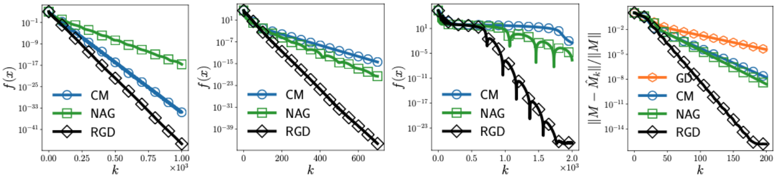
$$x_{k+1} = \alpha x_{k+1/2} + (1-\alpha) x_k + v_{k+1/2} \Big/ \sqrt{\delta\|v_{k+1/2}\|^2 + 1}, \tag{13c}$$

$$v_{k+1} = \sqrt{\mu} v_{k+1/2}. \tag{13d}$$

Some interesting properties of RGD:

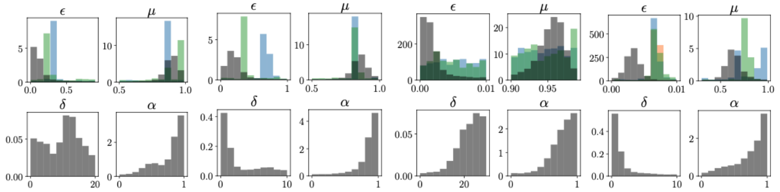
$$
\begin{aligned}
x_{k+1/2} &= x_k + \sqrt{\mu} v_k \Big/ \sqrt{\mu\delta\|v_k\|^2 + 1}, \\
v_{k+1/2} &= \sqrt{\mu} v_k - \epsilon \nabla f(x_{k+1/2}), \\
x_{k+1} &= \alpha x_{k+1/2} + (1-\alpha) x_k + v_{k+1/2} \Big/ \sqrt{\delta\|v_{k+1/2}\|^2 + 1}, \\
v_{k+1} &= \sqrt{\mu} v_{k+1/2}.
\end{aligned}
$$

- RGD recover the behaviour of CM when $\delta = 0$ and $\alpha = 1$.
- RGD can also recover NAG when $\delta = 0$ and $\alpha = 0$.
- In general, it can considerably improve the convergence and stability of CM and NAG since it can control the kinetic energy of the system.
- We expect that RGD stands out on setting with large gradients.
- Let us show a few examples ...



(a) Correlated quadratic (b) Random quadratic (c) Rosenbrock (d) Matrix completion

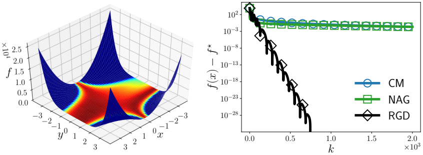
Figure 2: Convergence rate showing improved performance of RGD (Algorithm 1); see text.



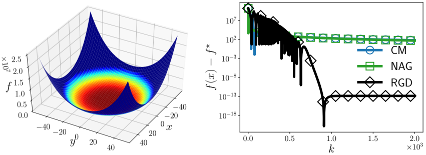
(a) Correlated quadratic (b) Random quadratic (c) Rosenbrock (d) Matrix completion

Figure 3: Histograms of hyperparameter tuning by Bayesian optimization. Tendency towards $\alpha \approx 1$ indicates benefits of being symplectic, while $\alpha \approx 0$ of being extra damped as in NAG. Tendency towards $\delta > 0$ indicates benefits of relativistic normalization. (Colors follow Fig. 2.)

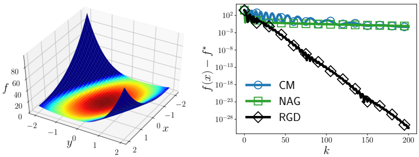
More Examples



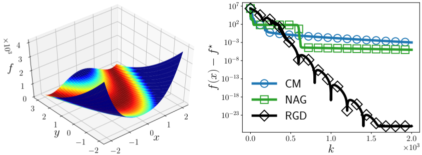
Beale function, 2 dimensions.



Chung-Reynolds, 50 dimensions.



Zakharov function, 5 dimensions.



Rosenbrock function, 1000 dimensions.

- Check the paper for many other examples and further insights.

Summary

- We considered accelerated methods (CM and NAG) from a "symplectic" or structure-preserving perspective.
- We elucidated how CM and NAG preserves, or not, the underlying symplectic structure of the continuum system.
- We proposed a perturbed dynamical systems that describe CM and NAG to a higher degree of resolution.
- We introduced a new method (RGD) that generalizes CM and NAG, and may have better stability and improved convergence.
- More importantly, this paper brings a first-principles approach to understand/construct accelerated methods.
- A complete theoretical justification is provided in our more recent paper: GF, MI Jordan, R Vidal, "On Dissipative Symplectic Integration with Applications to Gradient-Based Optimization" (2020).

Acknowledgements: ARO MURI W911NF-17-1-0304 and NSF 1447822.