A note on 'Conformal Symplectic and Relativistic Optimization'

(Spotlight paper at NeurIPS 2020)

December 28, 2020

1 Motivation

- Acceleration is a fundamental phenomenon in optimization, and it is important to understand its nature from a geometrically inclined perspective.
- Is there a principled way to *understand* acceleration and *create* accelerated algorithms in general?
- What does it mean to be able to optimize fast?

Thinking about the acceleration phenomenon *without* delving into the topological aspects of the underlying space where an algorithm is supposed to operate, *does not* (rather cannot) provide any rich answers. The richer understanding therefore has to be built via going into continuous time, via the language of differential equations (differential geometry).

The goal therefore is to

- understand the acceleration phenomena in continuous time first,
- deriving the corresponding continuous time convergence rates (e.g., via a Lyapunov analysis), then
- discretizing (numerically integrate) the differential equations to obtain discrete algorithms which
  - are stable,
  - preserve the continuous time rates, and
  - respect the underlying geometry.

There is a long line of recent work [5, 6, 7, 4, 1, 2] that has been able to formalize these notions to a great extent. This paper considers the popularly known variants of accelerated algorithms, and tries to argue about their *stability* properties. The paper also creates a class of accelerated algorithms that are more stable and fast.

Consider the optimization problem

$$\min_{\mathbf{x}\in M} f(\mathbf{x}),$$

where M is a d dimensional smooth manifold (e.g., $(\mathbb{R}^d, \mathcal{O}_{\text{standard}}, \mathcal{A}_{\text{smooth}})$), and $f\colon M \to \mathbb{R}$ is a continuously differentiable function. The paper considers two popularly used variants of acceleration

1. Classical momentum (CM): Also known as the heavy ball method

$$\mathbf{v}_{k+1} = \mu\mathbf{v}_k - \epsilon\nabla f(\mathbf{x}_k), \qquad \mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{v}_{k+1}. \tag{1.1}$$

2. Nesterov's accelerated gradient (NAG):

$$\mathbf{v}_{k+1} = \mu\mathbf{v}_k - \epsilon\nabla f(\mathbf{x}_k + \mu\mathbf{v}_k), \qquad \mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{v}_{k+1}. \tag{1.2}$$

In both the methods, $\epsilon > 0$ is the step-size, and $\mu \in (0, 1)$ is the momentum factor. Both these methods are *some* naive discretizations of a particular 2^{nd} order differential equation with a friction term. Since there are many ways to discretize a continuous system, the question is which discretization is the most suitable one for optimization in terms of the above stated goals?

2 Background

- A solution to a differential equation provides a vector field $X \in \Gamma(TM)$ whose flow

$$\begin{aligned}\Phi &\colon \mathbb{R} \times M \to M \\ (h, \mathbf{x}(t)) &\mapsto \Phi(h, \mathbf{x}(t)) = \mathbf{x}(t+h)\end{aligned}$$

  is an element of the one parameter group of diffeomorphisms on M.

- In physical systems, a Hamiltonian encodes the laws of motion. These laws of motion typically correspond to conservation of quantities like energy, or angular momentum (through some symmetry via Noether's theorem). The Hamiltonian is a function which given a point $(\mathbf{x}, \mathbf{p}) \in T^*M$ on the phase space, gives us the *energy* of the total system at that state. Here $\mathbf{x} \in M$ is the position and $\mathbf{p} \in T^*_{\mathbf{x}}M$ is the momentum at the point $\mathbf{x}$. With an analogy from classical mechanical systems, we will first consider the standard classical Hamiltonian

$$H(\mathbf{x}, \mathbf{p}) = \underbrace{\frac{\|\mathbf{p}\|_2^2}{2m}}_{\text{Kinetic energy}} + \underbrace{f(\mathbf{x})}_{\text{Potential energy}}, \tag{2.1}$$

  where $m \in \mathbb{R}_{++}$ is the mass of the particle under motion.

- Conservative Hamiltonian systems have *constants of motion* which gives us the system of *classical Hamiltonian differential equations*

$$\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} = \frac{\partial}{\partial \mathbf{p}} H(\mathbf{x}(t), \mathbf{p}(t)) = \frac{\mathbf{p}(t)}{m} \tag{2.2a}$$

$$\frac{\mathrm{d}\mathbf{p}(t)}{\mathrm{d}t} = -\frac{\partial}{\partial \mathbf{x}} H(\mathbf{x}(t), \mathbf{p}(t)) = -\nabla f(\mathbf{x}(t)) \tag{2.2b}$$

- Generically, dynamical systems on a d dimensional smooth manifold M corresponds to preservation of the non-degenerate symplectic structure on T^*M

$$\omega \in \Gamma(T^*M \otimes T^*M),$$

  if and only if it is a conservative Hamiltonian system, i.e., there exists a Hamiltonian function H on T^*M such that the corresponding Hamiltonian vector field X and flow Φ satisfy

$$\mathcal{L}_X \omega(t) = 0 \iff \frac{\mathrm{d}}{\mathrm{d}t} H(\mathbf{x}(t), \mathbf{p}(t)) = 0 \tag{2.3a}$$

$$\equiv \Phi_t^* \omega = \omega, \qquad (\text{i.e., } \Phi_t \text{ is a symplectomorphism } \forall\, t) \tag{2.3b}$$

$$\equiv \left(\frac{\partial \Phi(t, \mathbf{z}_0)}{\partial \mathbf{z}_0}\right)^\top \Omega \left(\frac{\partial \Phi(t, \mathbf{z}_0)}{\partial \mathbf{z}_0}\right) = \Omega, \tag{2.3c}$$

  where $\Omega \coloneqq \begin{bmatrix} \mathbf{0}_d & \mathbf{I}_d \\ -\mathbf{I}_d & \mathbf{0}_d \end{bmatrix}$ contains the coordinates functions of ω under the Darboux chart.

- Naive discretizations (like explicit or implicit Euler methods) of general Hamiltonian differential equations lead to bleeding off of the energy with time and therefore are not the right way to discretize Hamiltonian systems. Symplectic integrators developed in mid 20^{th} century allow us to discretize these differential equation without losing much energy, and therefore allows us to take much larger steps as they are stable and have better error control properties.

- We however, want to decrease both potential energy $f(\mathbf{x})$ and the kinetic energy $\frac{\|\mathbf{p}\|_2^2}{2m}$ at the same time at a controlled pace for desired convergence rates in continuous time. Symplectic integrators on conserved Hamiltonian systems tend to oscillate around the minimum. Therefore we want to dissipate energy so that both $\nabla f(\mathbf{x}(t))$ and $\frac{\mathbf{p}(t)}{m}$ suitably go to $\mathbf{0}$ with t.

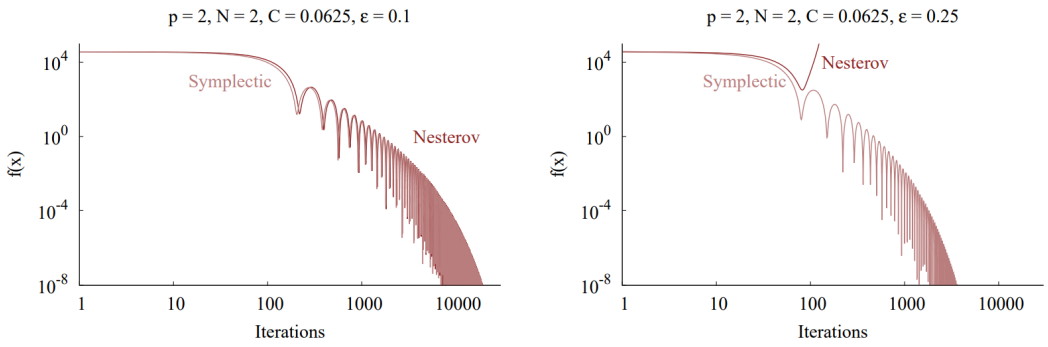

Figure 1: Stability of a symplectic integrator on a quadratic optimization problem (Taken from [1])

- (Poincaré 1899, Cartan 1913) Any smooth manifold has 6 fundamental classes of diffeomorphisms, one of which is the set of flows called the *conformal symplectic* flows, Diff^c_ω, which contains the diffeomorphisms that preserve the symplectic form ω up to constants. Such systems dissipate energy (symplectic area), i.e., if Φ is a flow of a conformal symplectic vector field X, then

$$\begin{aligned} \mathcal{L}_X\omega &= -\gamma\omega, && (2.4a)\\ \equiv \Phi_t^*\omega &= e^{-\gamma t}\omega, && (2.4b)\\ \equiv \left(\frac{\partial\Phi(t,\mathbf{z}_0)}{\partial\mathbf{z}_0}\right)^\top \Omega\left(\frac{\partial\Phi(t,\mathbf{z}_0)}{\partial\mathbf{z}_0}\right) &= e^{-\gamma t}\Omega, && (2.4c)\\ \equiv \omega_t &= e^{-\gamma t}\omega_0, && (2.4d) \end{aligned}$$

  for some $\gamma > 0$. Conformal symplectic transformations can be composed and form the *conformal group*.

- If X is the vector field solution of the classical Hamiltonian, then the differential equations corresponding to the *classical conformal Hamiltonian system* is

$$\begin{aligned} \frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} &= \nabla_{\mathbf{p}} H(\mathbf{x}(t),\mathbf{p}(t))\,, && (2.5a)\\ \frac{\mathrm{d}\mathbf{p}(t)}{\mathrm{d}t} &= -\nabla_{\mathbf{x}} H(\mathbf{x}(t),\mathbf{p}(t)) - \gamma\mathbf{p}(t)\,, && (2.5b) \end{aligned}$$

$$\implies \frac{\mathrm{d}}{\mathrm{d}t}H(\mathbf{x}(t),\mathbf{p}(t)) = -\frac{\gamma}{m}\|\mathbf{p}(t)\|_2^2 \le 0,$$

  or concisely, for $\mathbf{z}(t) \coloneqq \begin{bmatrix}\mathbf{x}(t)\\ \mathbf{p}(t)\end{bmatrix}$, $\Omega = \begin{bmatrix}\mathbf{0}_d & \mathbf{I}_d\\ -\mathbf{I}_d & \mathbf{0}_d\end{bmatrix}$, and $\mathbf{D} \coloneqq \begin{bmatrix}\mathbf{0}_d & \mathbf{0}_d\\ \mathbf{0}_d & \mathbf{I}_d\end{bmatrix}$, the classical conformal Hamiltonian differential equations (2.5) can be written as

$$\dot{\mathbf{z}} = \underbrace{\Omega\nabla H(\mathbf{z})}_{=:C(\mathbf{z})} + \underbrace{(-\gamma\mathbf{D}\mathbf{z})}_{=:D(\mathbf{z})} \qquad (2.6)$$

- Recently [2], it has been shown that *conformal symplectic integrators* are able to preserve the continuous time rates of convergence up to a controlled error, i.e., they are 'rate matching' up to negligible error. The continuous time rate can be obtained by a Lyapunov analysis for a given Hamiltonian function (time varying). In other words, symplectic integrators provide stability of critical points and preserves continuous time convergence rates, therefore it is possible to take much *larger steps*.

3 Conformal Symplectic Integration

- Associating flows Φ_t^C and Φ_t^D to respective vector fields of the terms $C(\mathbf{z})$ and $D(\mathbf{z})$ in Equation (2.6), we can approximately construct the total flow Φ_t of Equation (2.6) by composing the individual flows. To construct

the numerical map Ψ_h of the continuous flow Φ_t, we first approximate the conservative part Ψ_h^C of the system $\dot{\mathbf{z}} = \Omega \nabla H(\mathbf{z})$, and then the dissipative part Φ_h^D of the system $\dot{\mathbf{z}} = -\gamma \mathbf{D}\mathbf{z}$, which can be integrated exactly as

$$\Psi_h^D \colon (\mathbf{x}, \mathbf{p}) \mapsto (\mathbf{x}, e^{-\gamma h}\mathbf{p}). \tag{3.1}$$

To approximate Φ_t^C, we can use any standard symplectic integrator. Considering the symplectic Euler method, we have

$$\begin{aligned}
\Psi_h^C \colon &\quad (\mathbf{x}_k, \mathbf{p}_k) \mapsto (\mathbf{x}_{k+1}, \mathbf{p}_{k+1}) \ , &&\text{where it can be either} \\
\Psi_h^{C_1} \colon &\quad \mathbf{p}_{k+1} = \mathbf{p}_k - h\nabla_{\mathbf{x}} H(\mathbf{x}_k, \mathbf{p}_{k+1}) \ , &&\mathbf{x}_{k+1} = \mathbf{x}_k + h\nabla_{\mathbf{p}} H(\mathbf{x}_k, \mathbf{p}_{k+1}) \ , \qquad (3.2) \\
\text{or,} \quad \Psi_h^{C_2} \colon &\quad \mathbf{x}_{k+1} = \mathbf{x}_k + h\nabla_{\mathbf{p}} H(\mathbf{x}_{k+1}, \mathbf{p}_k) \ , &&\mathbf{p}_{k+1} = \mathbf{p}_k - h\nabla_{\mathbf{x}} H(\mathbf{x}_{k+1}, \mathbf{p}_k) \ . \qquad (3.3)
\end{aligned}$$

Note that they are implicit, but if H is separable in $\mathbf{x}$ and $\mathbf{p}$ as in our case, it turns into an explicit update rule.

- We can create the total numerical flow Ψ_h as
  1. (Dissipative Symplectic Euler method): $\Psi_h^{C_1} \circ \Psi_h^D$, or $\Psi_h^{C_2} \circ \Psi_h^D$,
  2. (Dissipative Leapfrog method): $\Psi_{h/2}^D \circ \Psi_{h/2}^{C_1} \circ \Psi_{h/2}^{C_2} \circ \Psi_{h/2}^D$.

Definition 3.1 (Order of a symplectic integrator)**.** A numerical map Ψ of a continuous flow Φ is of order $r \geq 1$ if $\|\Psi_h(\mathbf{z}) - \Phi_h(\mathbf{z})\|_2 = \mathcal{O}\big(h^{r+1}\big)$ for all $\mathbf{z} \in T^*M$.

Theorem 3.2. *Both integrators 1 and Integrator 2 are conformal symplectic, i.e., they both satisfy Equation (2.4c) exactly. Integrator 1 is of order* 1*, and Integrator 2 is of order* 2*.*

Remark 3.3. The numerical map $\Psi_h^{C_1} \circ \Psi_h^D$ is nothing but the Classical Momentum (CM) method (1.1) for the classical Hamiltonian (2.1), which from Theorem 3.2 is a **conformal symplectic** integrator of order 1.

Remark 3.4. The numerical map $\Psi_{h/2}^D \circ \Psi_{h/2}^{C_1} \circ \Psi_{h/2}^{C_2} \circ \Psi_{h/2}^D$, for the classical Hamiltonian (2.1) can be written as

$$\mathbf{x}_{k+1/2} = \mathbf{x}_k + \mu \mathbf{v}_k, \qquad \mathbf{v}_{k+1} = \mu \mathbf{v}_k - \epsilon \nabla f(\mathbf{x}_{k+1/2}), \qquad \mathbf{x}_{k+1} = \mathbf{x}_{k+1/2} + \mathbf{v}_{k+1}, \tag{3.4}$$

which will be the Nesterov's Accelerated Gradient (NAG) method if we replace $\mathbf{x}_{k+1/2} \to \mathbf{x}_k$ in the third update above. NAG is an integrator of order 1, but is **not conformal symplectic**, rather contracts the symplectic form as

$$\omega_{k+1} = e^{-\gamma h}\left[\mathbf{I} - \frac{h^2}{m}\nabla^2 f(\mathbf{x}_k)\right]\omega_k + \mathcal{O}\big(h^3\big). \tag{3.5}$$

The spurious dissipation in NAG may improve the convergence rate sightly, but at the cost of making the method less stable.

- The paper also asks for which dynamical systems, CM and NAG turn out to be 2^{nd} order integrators of.
  - Both their Shadow Hamiltonian dynamical systems depend on the step-size h (reflecting upon the intrinsic behavior of the discrete time algorithm). If we set $h \to 0$ in both their dynamical systems, we obtain a 2^{nd} order differential equation
  $$\ddot{\mathbf{x}} + \gamma \dot{\mathbf{x}} = -\frac{1}{m}\nabla f(\mathbf{x}),$$
  which is of the same form obtained in [SBC16] 'A differential equation for modeling Nesterov's accelerated gradient descent method: Theory and insights', up to γ being depending on t.
  - The perturbed system for CM turns out to be conformal Hamiltonian, contrary to that of NAG. Being a conformal Hamiltonian system allows structure preservation and their conformal symplectic discretizations tend to be more stable since perturbed trajectories are always close to that of the original Hamiltonian dynamics.

4 A dissipative Relativistic system

- We, in optimization (unlike in physics) can choose our own Hamiltonian system, and define a geometry with properties we are interested in. The meaning of acceleration will not change, but its characterization will, depending on the Hamiltonian function.

- For the classical Hamiltonian i.e., Equation (2.1), since time is independent of space (like in Newtonian spacetime, that is parabolic), large magnitudes of gradient can be lead to large momentum. This leads to large spatial speed, and cause position updates to diverge.

- In relativistic spacetime physics, space and time carry no sharp distinctions. The base space is a $(d+1)$ dimensional space called *Minkwoski spacetime* (hyperbolic). Distances on spacetime arise through a *Lorentzian metric* under which an infinitesimal distance $\mathrm{d}\mathbf{s}$ satisfies $(\mathrm{d}\mathbf{s})^2 = -(c\,\mathrm{d}t)^2 + \|\mathrm{d}\mathbf{x}\|_2^2$, i.e., no particle can travel faster than c (a universal constant in the physical world). This imposes constraint on the geometry (hyperbolic) where trajectories take place. Therefore, by discretizing a relativistic system we can hope to create optimization algorithms that are more stable.

- The relativistic Hamiltonian (in contrast to the classical Hamiltonian (2.1)) is now defined as

$$H(\mathbf{x},\mathbf{p}) = c\sqrt{\|\mathbf{p}\|_2^2 + m^2c^2} + f(\mathbf{x}), \tag{4.1}$$

  and its corresponding *relativistic conformal Hamiltonian dynamics* is gives by

$$\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} = \frac{c\,\mathbf{p}(t)}{\sqrt{\|\mathbf{p}(t)\|_2^2 + m^2c^2}} \tag{4.2a}$$

$$\frac{\mathrm{d}\mathbf{p}(t)}{\mathrm{d}t} = -\nabla f(\mathbf{x}(t)) - \gamma\mathbf{p}(t) \tag{4.2b}$$

  Note from Equation (4.2a) how $\|\dot{\mathbf{x}}\|_2$ is bounded even for large $\|\mathbf{p}(t)\|_2$.

- The corresponding dissipative symplectic Euler numerical map $\Psi_h^{C_1} \circ \Psi_h^D$ is an order 1 conformal symplectic integrator, and is exactly CM if $c \to \infty$.

- The corresponding dissipative leapfrog numerical map $\Psi_{h/2}^D \circ \Psi_{h/2}^{C_1} \circ \Psi_{h/2}^{C_2} \circ \Psi_{h/2}^D$ is an order 2 conformal symplectic integrator. Just like NAG is a deviation from this classical version of this integrator (Equation (3.4)), we can obtain a family of numerical maps by introducing a convex combination $\alpha\mathbf{x}_{k+1/2} + (1-\alpha)\mathbf{x}_k$ for $\alpha \in [0,1]$, with $\alpha = 1$ corresponding to the conformal leapfrog version (order 2), and $\alpha = 0$ to the Hessian damping version.

Algorithm 1 Relativistic Gradient Descent

Input: $\mathbf{x}_0, \mathbf{v}_0 \in \mathbb{R}^d$, $\epsilon, \delta \geq 0$, $\mu \in (0,1]$, $\alpha \in [0,1]$
for $k \in \mathbb{Z}_+$ **do**
  $\mathbf{x}_{k+1} \leftarrow \mathbf{x}_k + \frac{\sqrt{\mu}}{\mu\delta\|\mathbf{v}_k\|_2^2+1}\mathbf{v}_k$
  $\mathbf{v}_{k+1/2} \leftarrow \sqrt{\mu}\mathbf{v}_k - \epsilon\nabla f(\mathbf{x}_{k+1/2})$
  $\mathbf{x}_{k+1} \leftarrow \big(\alpha\mathbf{x}_{k+1/2} + (1-\alpha)\mathbf{x}_k\big) + \frac{1}{\delta\|\mathbf{v}_{k+1/2}\|_2^2+1}\mathbf{v}_{k+1/2}$
  $\mathbf{v}_{k+1} \leftarrow \sqrt{\mu}\mathbf{v}_{k+1/2}$
end for

  In Algorithm 1, $\delta = 0$ corresponds to the the leapfrog integrator for classical (non-relativistic) Hamiltonian dynamics.

- It can be shown that when $\delta > 0$ (relativistic), then

$$\big\|\mathbf{x}_{k+1} - \big(\alpha\mathbf{x}_{k+1/2} + (1-\alpha)\mathbf{x}_k\big)\big\|_2 \leq 1/\delta,$$

  showing that the spatial speed is always bounded irrespective of large magnitudes of gradient.

5 Trade-off between stability and convergence rate

- Higher contraction, i.e., by introducing spurious dissipation in numerical method (like in NAG, Equation (3.5)) in the symplectic form leads to improved convergence rates, but at the cost of stability.
- A geometry/structure preserving integrator (like CM) may have lesser contraction (obeys conformal symplecticity), but will have higher stability. Such integrators also better preserve the continuous time convergence rates.
- The paper looks at a simple function $f(\mathbf{x}) = \lambda\|\mathbf{x}\|_2^2/2$ for $d = 1$ and computes the eigenvalues of the update rules $\begin{bmatrix}\mathbf{x}_{k+1}\\ \mathbf{p}_{k+1}\end{bmatrix} = \mathbf{T}_{\text{alg}}(h, \lambda, m, \gamma)\begin{bmatrix}\mathbf{x}_k\\ \mathbf{p}_k\end{bmatrix}$ for alg $\in \{\text{CM}, \text{NAG}, \text{RGD}\}$.

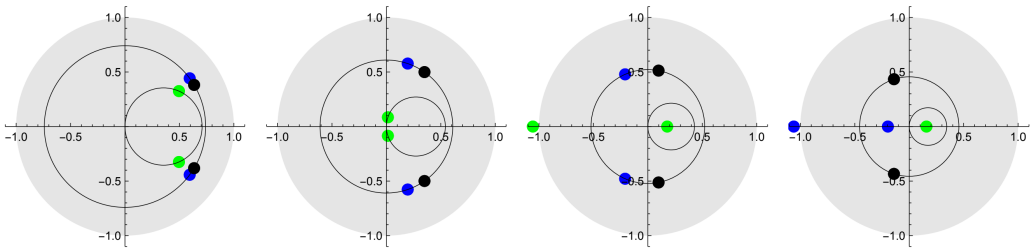

Figure 2: Region of eigenvalues in $\mathbb{C}$ of the update operators to understand stability of CM (blue), NAG (green), and RGD (black) ($c \to \infty$, $\alpha = 1$, $m = 1$, $\gamma = 1$) with increasing step size h for $f(\mathbf{x}) = \|\mathbf{x}\|_2^2/2$ for $d = 1$.

The properties of the algorithm like stability, being conformal, convergence speeds, can be visualized from the locus of the eigenvalues of the matrix $\mathbf{T}_{\text{alg}}(h, \lambda, m, \gamma)$ as we increase the step-size h. For the case when $c \to \infty$, and $\alpha = 1$, the paper analytically finds that

$$h^{\max}_{\text{NAG}} < h^{\max}_{\text{CM}} < h^{\max}_{\text{RGD}}.$$

6 Numerical Experiments

- For several popularly known hard (to optimize) functions, the paper shows that when hyper-parameters are tuned using Bayesian optimization, there is a preference towards selecting $\alpha \approx 1$ (indicating benefits being conformal), and $\delta > 0$ (indicating benefits of being relativistic).
- In all cases, RGD is shown to be empirically much faster and stable than CM and NAG, even all the hyper-parameters are tuned to be the best. The most interesting (and hard) ones are below:

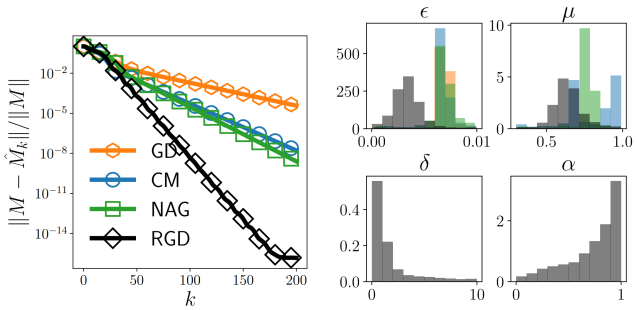

Figure 3: Matrix completion of $\mathbb{R}^{100\times100} \ni M = RS^\top$ for $R, S \in \mathbb{R}^{100\times5}$ with entries i.i.d. $\mathcal{N}(1, 2)$. Sampling ratio 0.3. Solve $\min_{U,V\in\mathbb{R}^{100\times5}} \left\|P_{\text{obs}}(M - UV^\top)\right\|_{\text{F}}$, using alternating minimization with initializations standard normal. Number of observed entries $0.3 \times 100 \times 5$.

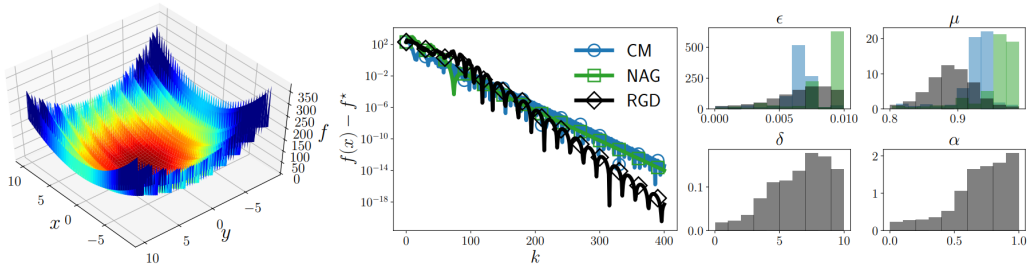

Figure 4: Lévi function: $f(x,y) = \sin^2(3\pi x) + (x-1)^2\big(1+\sin^2(3\pi y)\big) + (y-1)^2\big(1+\sin^2(2\pi y)\big)$. The function is highly non-convex, with global minimum at (1,1). On multiple runs, CM and NAG get stuck in local minima more often then RGD.

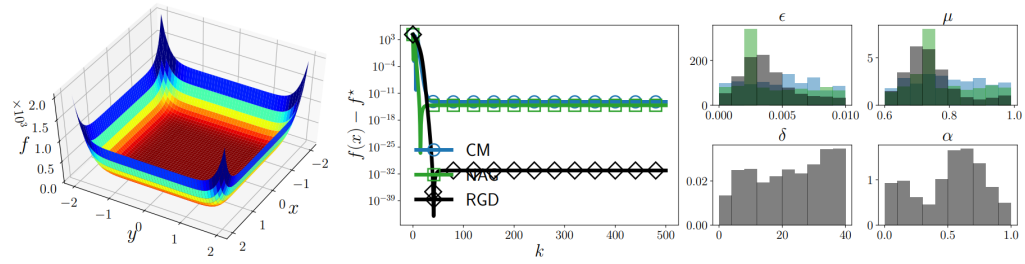

Figure 5: Schwefel function: $f(\mathbf{x}) = \|\mathbf{x}\|_{10}^{10}$, $\mathbf{x} \in \mathbb{R}^{20}$.

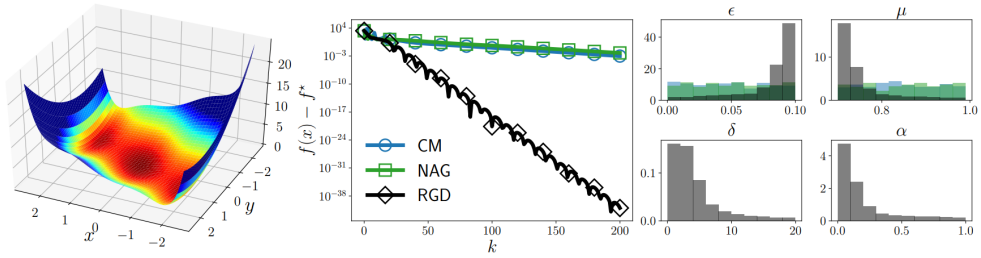

Figure 6: Three-hump camel back function: $f(x,y) = 2x^2 - 1.05x^4 + x^6/6 + xy + y^2$. This function is multi-modal with minima at $(0,0)$. Algorithms were initialized at $(5,5)$.

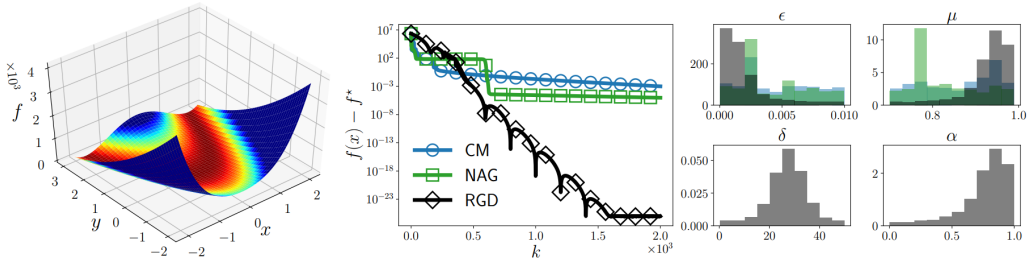

Figure 7: Rosenbrock function: $f(\mathbf{x}) = \sum_{i=1}^{n-1}\big[100(x_{i+1} - x_i^2)^2 + (x_i - 1)^2\big]$, $n = 1000$. One global minima at $\mathbf{1}$. One local minima at $\big[-1, \mathbf{1}_{d-1}\big]$. Has exponentially many saddle points of which 2 are actually hard to escape. Initialized at $x_{0,i} = \pm 2$ for i odd/even.

References

[1] Michael Betancourt, Michael I Jordan, and Ashia C Wilson. On symplectic optimization. *arXiv preprint arXiv:1802.03653*, 2018.

[2] Guilherme França, Michael I Jordan, and René Vidal. On dissipative symplectic integration with applications to gradient-based optimization. *arXiv preprint arXiv:2004.06840*, 2020.

[3] Guilherme França, Jeremias Sulam, Daniel P Robinson, and René Vidal. Conformal symplectic and relativistic optimization. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(12):124008, 2020.

[4] Bin Shi, Simon S Du, Michael I Jordan, and Weijie J Su. Understanding the acceleration phenomenon via high-resolution differential equations. *arXiv preprint arXiv:1810.08907*, 2018.

[5] Weijie Su, Stephen Boyd, and Emmanuel Candes. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. *Advances in neural information processing systems*, 27:2510–2518, 2014.

[6] Andre Wibisono, Ashia C Wilson, and Michael I Jordan. A variational perspective on accelerated methods in optimization. *proceedings of the National Academy of Sciences*, 113(47):E7351–E7358, 2016.

[7] Jingzhao Zhang, Aryan Mokhtari, Suvrit Sra, and Ali Jadbabaie. Direct runge-kutta discretization achieves acceleration. In *Advances in neural information processing systems*, pages 3900–3909, 2018.