Optimization algorithms inspired by the geometry of dissipative systems

Alessandro Bravetti¹, Maria L. Daza-Torres², Hugo Flores-Arguedas, and Michael Betancourt³

¹Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas (IIMAS), Universidad Nacional Autónoma de México, A.P. 70-543, 04510 Ciudad de México, México
alessandro.bravetti@iimas.unam.mx
²University of California, Davis, United States, mdazatorres@ucdavis.edu
³Symplectomorphic LLC, New York, USA, betan@symplectomorphic.com

Abstract
Accelerated gradient methods are a powerful optimization tool in machine learning and statistics but their development has traditionally been driven by heuristic motivations. Recent research, however, has demonstrated that these methods can be derived as discretizations of dynamical systems, which in turn has provided a basis for more systematic investigations, especially into the structure of those dynamical systems and their structure-preserving discretizations. In this work we introduce dynamical systems defined through a contact geometry which are not only naturally suited to the optimization goal but also subsume all previous methods based on geometric dynamical systems. These contact dynamical systems also admit a natural, robust discretization through geometric contact integrators. We illustrate these features in paradigmatic examples which show that we can indeed obtain optimization algorithms that achieve oracle lower bounds on convergence rates while also improving on previous proposals in terms of stability.

Keywords: Optimization, Accelerated Gradient, Dynamical Systems, Geometric Integrators, Contact Geometry

1 Introduction
Despite their practical utility and explicit convergence bounds, accelerated gradient methods have long been difficult to motivate from a fundamental theory. This lack of understanding limits the theoretical foundations of the methods, which in turns hinders the development of new and more principled schemes. Recently a progression of work has studied some continuum limits of accelerated gradient methods,
demonstrating that these methods can be derived as discretizations of continuous dynamical systems. Shifting focus to the structure and discretization of these latent dynamical systems provides a foundation for the systematic development and implementation of new accelerated gradient methods.

This recent direction of research began with [26] which found a continuum limit of Nesterov’s accelerated gradient method (NAG)

\[
X_k = P_{k-1} - \tau \nabla f(P_{k-1}) \tag{1}
\]

\[
P_k = X_k + \frac{k-1}{k+2}(X_k - X_{k-1}) \tag{2}
\]

by discretizing the ordinary differential equation

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

for \( t > 0 \) with the initial conditions \( X(0) = X_0 \) and \( \dot{X}(0) = 0 \). By generalizing the ordinary differential equation (3) they were then able to derive new accelerated gradient methods that achieved comparable convergence rates.

[29] continued in this direction by showing that accelerated methods can also be derived as a discretization of a more structured variational dynamical system specified with a time–dependent Lagrangian, or equivalent Hamiltonian. Consider an objective function \( f : X \to \mathbb{R} \) which is continuously differentiable, convex, and has a unique minimizer \( X^* \in X \). Moreover assume that \( X \) is a convex set endowed with a distance–generating function \( h : X \to \mathbb{R} \) that is also convex and essentially smooth, and indicate by \( \langle P, X \rangle \) the pairing between \( X \in X \) and \( P \in X^* \). From the Bregman divergence induced by \( h \),

\[
D_h(Y, X) = h(Y) - h(X) - \langle \nabla h(X), Y - X \rangle \tag{4}
\]

they derived the Bregman Lagrangian

\[
L_{BH}(X, V, t) = e^{\alpha t} \left( D_h(X + e^{-\alpha}V, X) - e^{\beta t} f(X) \right) \tag{5}
\]

where \( \alpha, \beta, \) and \( \gamma \) are continuously differentiable functions of time. They then proved that under the ideal scaling conditions

\[
\dot{\beta} \leq e^\alpha \tag{6}
\]

\[
\dot{\gamma} = e^\alpha, \tag{7}
\]

the solutions of the resulting Euler–Lagrange equations

\[
\ddot{X} + (e^\alpha - \dot{\alpha}) \dot{X} + e^{2\alpha + \beta} \left[ \nabla^2 h(X + e^{-\alpha} \dot{X}) \right]^{-1} \nabla f(X) = 0 \tag{8}
\]

satisfy [29, Theorem 1.1]

\[
f(X) - f(X^*) \leq O(e^{-\beta t}). \tag{9}
\]

From a physical perspective the two terms in equation (5) play the role of a kinetic and a potential energies, respectively. At the same time the explicit time–dependence of the Lagrangian (5) is a necessary ingredient in order for the dynamical system to dissipate energy and relax to a minimum of the potential, and hence to
a minimum of the objective function. Moreover, by (6), the optimal convergence rate is achieved by choosing $\beta = e^{\alpha}$, i.e. $\beta = \int_t^s e^{\alpha(s)} ds$, and we observe that in the Euclidean case, $h(X) = \frac{1}{2} \|X\|^2$, the Hessian is the identity matrix and thus (8) simplifies to
\[
\ddot{X} + (e^\alpha - \dot{\alpha}) \dot{X} + e^{2\alpha + \beta} \nabla f(X) = 0.
\] (10)
Finally the authors developed a heuristic discretization of (8) that yielded optimization algorithms matching the continuous convergence rates.

[5] considered more systematic discretizations of these variational dynamical systems that exploited the fact that they are well suited for numerical discretizations that preserve their geometric structure [21]. In particular they Legendre transformed the Bregman Lagrangian (5) to derive the Bregman Hamiltonian,
\[
H_{Br}(X, P, t) = e^{\alpha + \gamma} \left( D_{h^*} (e^{-\gamma} P + \nabla h(X), \nabla h(X)) + e^\beta f(X) \right),
\] (11)
where $h^*(P) = \sup_X \{\langle P, X \rangle - h(X) \}$ is the Legendre transform of $h(X)$, and then argued that a principled way to obtain reliable and rate–matching discretizations of the resulting dynamical system
\[
\dot{X} = \nabla_P H_{Br}(X, P, t)
\] (12)
\[
\dot{P} = -\nabla_X H_{Br}(X, P, t),
\] (13)
is with symplectic integrators. They numerically demonstrated in the Euclidean case that a standard leapfrog integrator yields an optimization algorithm that achieves polynomial convergence rates and showed how the introduction of a gradient flow could achieve late–time exponential convergence rates matching those seen empirically in other accelerated gradient methods. A more theoretical approach to rate–matching geometric discretizations for the Bregman dynamics has been recently proposed in [18], where the authors prove that pre–symplectic integrators provide a principled way to discretize Bregman dynamics while preserving the continuous rates of convergence up to a negligible error, provided some assumptions are met.

Perhaps most importantly these variational methods allowed the focus to shift from existing accelerated gradient methods to the structure of the latent dynamical systems. In particular the existence of variational dynamical systems whose geometric discretization gives rise to accelerated gradient methods, suggests the introduction of new, principled dynamical systems that promise entirely new methods. For example we can replace variational dynamical systems that exploit heuristic time–dependencies to achieve dissipation with explicitly dissipative dynamical systems. [24] and [16] considered a dynamical systems perspective on these systems, showing how relatively simple dissipations can achieve state-of-the-art convergence. [17] took a more geometric perspective, replacing the time–dependent Hamiltonian geometry of the variational systems with a conformally symplectic geometry that generates dynamical systems of the form
\[
\dot{X} = \nabla_P H(X, P)
\] (14)
\[
\dot{P} = -\nabla_X H(X, P) - c P,
\] (15)
with $c \in \mathbb{R}$ a constant. Being a geometric dynamical system this approach also admits effective geometric integrators similar to [5]. These conformally symplectic dynamical systems, however, are less general than the time–dependent variational
dynamical systems; in particular NAG cannot be exactly recovered in this framework [17].

Another relevant aspect that has been uncovered by studying optimization algorithms from a variational or Hamiltonian analysis is the focus on a very important degree of freedom, the choice of the kinetic energy, that plays a fundamental role in the construction of fast and stable algorithms that can possibly escape local minima, in direct analogy with what happens in Hamiltonian Monte Carlo methods [4, 3, 22]. In particular, first [23] and then [17] have motivated that a careful choice of the kinetic energy term can stabilize the dynamical systems when the objective function is rapidly changing, similar to the regularization induced by trust regions. Indeed, like the popular Adam, AdaGrad and RMSprop algorithms, the resulting Relativistic Gradient Descent (RGD) algorithm regularizes the dynamical velocities to achieve a faster convergence and improved stability.

Finally [23] introduced another way of incorporating dissipative terms into Hamilton’s equations (12)–(13) (see also [25]). Their Hamiltonian descent algorithm is derived from the equations of motion

\[
\dot{X} = \nabla_P H(X, P) + X^* - X \tag{16}
\]
\[
\dot{P} = -\nabla_X H(X, P) + P^* - P, \tag{17}
\]

where \((X^*, P^*) = \arg\min H(X, P)\). Because the dynamics are defined using terms only linear in \(X\) and \(P\) they converge to the optimal solution exponentially quickly under mild conditions on \(H\) [25]. That said, this exponential convergence requires already knowing the optimum \((X^*, P^*)\) in order to generate the dynamics. Additionally this particular dynamical system lies outside of the variational and conformal symplectic families of dynamical systems and so can not take advantage of the geometric integrators.

In this work we show that all of the above–mentioned dynamical systems can be incorporated into a single family of contact Hamiltonian systems [6, 8] endowed with a contact geometry. The geometric foundation provides a powerful set of tools for both studying the convergence of the continuous dynamics as well as generating structure–preserving discretizations. Moreover, it automatically implies the invariance of the derived algorithms with respect to the group of contactomorphisms. We will comment on this property and its implications, although more technical results obtained from this fact regarding the Bregman dynamics are deferred to a companion paper [12]. Then we illustrate some of the possible advantages of this enriched perspective by producing a new version of RGD, named Contact Relativistic Gradient Descent (CRGD), and provide numerical experiments as a way to illustrate the potential advantages of using CRGD over RGD.

The structure of this work is as follows: in Section 2 we introduce contact Hamiltonian systems and show that all systems corresponding to Equations (10), (14)–(15), and (16)–(17) can be easily recovered as particular examples. In Section 3 we provide the basics of the geometric theory of time–discretization of contact systems by means of splitting, thus deriving optimization algorithms similar in spirit to, but more general than, those introduced in [5, 17, 18]. Then in Section 4 and Section 5 we show numerically that our algorithms can compete in terms of speed of convergence and stability with state-of-the-art algorithms such as CM, NAG and RGD. Finally, in Section 6 we summarize the results and discuss future directions.
2 Continuous–time contact optimization

Contact geometry is a rich subject at the intersection between differential geometry, topology and dynamical systems. Here in order to ease the exposition, we will present some of the basic facts needed to compare with previous works using symplectic and pre–symplectic structures in optimization in full generality, but we will soon specialize them to the cases of interest. For a treatment of the more general theory see [2, 20, 7, 8, 13, 15, 19, 14].

**Definition 1.** A contact manifold is a pair \((C, D)\), where \(C\) is a \((2n+1)\)-dimensional manifold and \(D\) is a maximally non–integrable distribution of hyperplanes on \(C\), that is, a smooth assignment at each point \(p\) of \(C\) of a subspace of dimension \(2n\) of the tangent space \(T_p C\), such that the maximal dimension of the integral submanifolds of \(D\) is \(n\).

One can prove that locally \(D\) can always be given as the kernel of a 1–form \(\eta\) satisfying the condition \(\eta \wedge (d\eta)^n \neq 0\), where \(\wedge\) is the wedge product and \((d\eta)^n\) means \(n\) times the wedge product of the 2–form \(d\eta\). This characterization will be enough for our purposes, and therefore we can introduce the following less general but more direct definition.

**Definition 2.** An exact contact manifold is a pair \((C, \eta)\), where \(C\) is a \((2n+1)\)-dimensional manifold and \(\eta\) is a 1–form satisfying \(\eta \wedge (d\eta)^n \neq 0\).

In what follows we will always restrict to the case of exact contact manifolds. As it is standard in geometry, transformations that preserve the contact structure, and hence the contact geometry, play a special role on these spaces. By noticing that the geometric object of interest is the kernel of a 1–form, one then defines isomorphisms in the contact setting in the following way.

**Definition 3.** A contact transformation or contactomorphism \(F : (C_1, \eta_1) \to (C_2, \eta_2)\) is a map that preserves the contact structure

\[
F^* \eta_2 = \alpha_F \eta_1,
\]

where \(F^*\) is the pullback induced by \(F\), and \(\alpha_F : C_1 \to \mathbb{R}\) is a nowhere–vanishing function.

**Remark 1.** In words, Definition 3 states that a contact map re–scales the contact 1–form by multiplying it by a nowhere–vanishing function. Indeed, such multiplication preserves the kernel of the 1–form, and hence the resulting geometry.

Let us present a simple but important example of contact manifold: we take \(C = \mathbb{R}^{2n+1}\) and specify the same contact structure in 2 different ways, corresponding to different choices of coordinates. Afterwards we prove that it amounts to the same structure by providing an explicit contactomorphism between the two.

**Example 1** (The standard contact structure in canonical coordinates). The standard structure is defined as the kernel of the 1–form

\[
\eta_{std1} := dS - PdX,
\]

where \(X \in \mathbb{R}^n\), \(P \in \mathbb{R}^n\) and \(S \in \mathbb{R}\) are Cartesian coordinates in \(\mathbb{R}^{2n+1}\).

We use “standard” because one can show that a contact structure on any manifold looks like this one locally [20].
Example 2 (The standard contact structure in non–canonical coordinates). This structure is defined as the kernel of the 1–form

$$\eta_{\text{std}2} := d\tilde{S} - \frac{1}{2} \tilde{P}d\tilde{X} + \frac{1}{2} \tilde{X}d\tilde{P},$$

(20)

where ($\tilde{X}, \tilde{P}, \tilde{S}$) are related to the Cartesian coordinates ($X, P, S$) of the previous example by the transformation (21) defined below.

Although this appears different from the structure in Example 1 they define equivalent geometries, as we now show.

Remark 2. We can explicitly construct a contact transformation between $\eta_{\text{std}1}$ and $\eta_{\text{std}2}$ above. The map

$$F : (X, P, S) \mapsto \left(\tilde{X} = X + P, \tilde{P} = P - X, \tilde{S} = S - \frac{XP}{2}\right)$$

(21)

satisfies $F^*\eta_{\text{std}2} = \eta_{\text{std}1}$. Consequently the two structures defined in Examples 1 and 2 are equivalent. The superficial difference arises only because they are written in different coordinates.

Historically, one of the main motivations to introduce contact geometry is the study of time–dependent Hamiltonian systems on a symplectic manifold. We briefly sketch these ideas because they will be relevant for our discussion.

Definition 4. A symplectic manifold is a pair $(M, \Omega)$ where $M$ is a 2n–dimensional manifold and $\Omega$ a 2–form on $M$ that is closed, $d\Omega = 0$, and non–degenerate, $(d\Omega)^n \neq 0$.

Definition 5. A Hamiltonian system on a symplectic manifold is a vector field $X_H$ which is defined by the condition

$$\Omega(X_H, \cdot) = -dH$$

(22)

where $H : M \to \mathbb{R}$ is a function called the Hamiltonian. The equations for the integral curves of $X_H$ are usually called Hamilton’s equations, and they are the foundations of the geometric treatment of mechanical systems.

Analogously to what happens in the contact case, it can be proved that all symplectic manifolds locally look the same; this result is known as Darboux’s Theorem ([1]). In particular, they all look like the Euclidean space $\mathbb{R}^{2n}$ with coordinates $(X, P)$, where $X \in \mathbb{R}^n$ play the role in physics of the generalized coordinates, and $P \in \mathbb{R}^n$ of the corresponding momenta. In such coordinates $\Omega = dP \wedge dQ$, and thus Hamilton’s equations (22) read like (12) and (13), but with the important difference that $H : M \to \mathbb{R}$ does not depend on time, thus giving rise to the so–called conservative mechanical systems. Here the name is due to the fact that $H(X, P)$ is usually the total mechanical energy of the system and it is easy to show that it is conserved along the dynamics.

To allow for $H$ to depend also on time and thus to describe dissipative systems in the geometric setting, one usually performs the following extension: first extend the manifold $M$ to $M \times \mathbb{R}$, where time $t$ is defined to be the coordinate on $\mathbb{R}$. At this point $H(X, P, t)$ is a well defined function on $M \times \mathbb{R}$. Now, define (locally at least)
the 1–form \( \theta = P \, dX - H(X,P,t) \, dt \). This is called the Poincaré–Cartan 1–form.

Finally, define the dynamics \( \dot{X}_H \) by the two conditions

\[
d\theta(\dot{X}_H, \cdot) = 0, \quad \theta(\dot{X}_H) = 1.
\]

(23)

It follows that the resulting equations for the integral curves of \( \dot{X}_H \) in the coordinates \((X,P,t)\) are just (12) and (13) for a generic \( H \) (now with the explicit time dependence in the function \( H \)), plus a trivial equation for \( t \), namely \( \dot{t} = 1 \).

**Remark 3.** For us there are 4 important points that are worth being remarked about this construction:

1. \( \theta \) is a contact form on \( M \times \mathbb{R} \) (unless \( H \) is a homogeneous function of degree 1 in \( P \), a case which is not interesting here).

2. From the point of view of the contact geometry, the dynamics in (23) is a very special type of Hamiltonian system on a contact manifold (to be defined below), corresponding to a constant Hamiltonian function with value -1 (see Remark 5 below). Indeed it may be shown that any contact flow can be split into a part of the type in (23) and another vector field of special geometric character, called a Liouville vector field [10].

3. Notice that, although Hamilton’s equations (12) and (13) dissipate the energy function \( H(X,P,t) \), from the geometric point of view the system of Hamilton’s equations (12) and (13), together with \( \dot{t} = 1 \), is conservative. This is because \( \mathcal{L}_{\dot{X}_H} \theta = 0 \), with \( \mathcal{L} \) the Lie derivative, i.e. the dynamics preserves \( \theta \), and hence the volume of the space given by \( \theta \wedge (d\theta)^n \).

4. The pre–symplectic dynamics used in [18] is exactly the dynamics of \( \dot{X}_H \). Indeed one can check that their final manifold of states and dynamical equations, after fixing the appropriate gauge, coincide with \( M \times \mathbb{R} \) and (23) respectively (indeed, it suffices to specialize our discussion to the case \( M = T^*\mathcal{M} \), and note that \( d\theta = \Omega \) in their notation).

We can now define dynamical systems that generalize the Hamiltonian systems arising in symplectic geometries (we refer again to [2, 20, 7, 8, 13, 15, 19, 14] for further details).

**Definition 6** (Contact Hamiltonian systems). Given a possibly time–dependent differentiable function \( H \) on a contact manifold \((C, \eta)\), we define the contact Hamiltonian vector field associated to \( H \) as the vector field \( X_H \) satisfying

\[
\mathcal{L}_{X_H} \eta = -R(\mathcal{H}) \eta \quad \eta(X_H) = -\mathcal{H},
\]

(24)

where \( \mathcal{L}_{X_H} \eta \) denotes the Lie derivative of \( \eta \) with respect to \( X_H \) and \( R \) is the Reeb vector field defined by the contact form, \( d\eta(R, \cdot) = 0 \) with \( \eta(R) = 1 \). We denote the flow of \( X_H \) the contact Hamiltonian system associated to \( \mathcal{H} \).

**Remark 4.** The first condition in (24) simply ensures that the flow of \( X_H \) generates contact transformations, while the second condition requires the vector field to be generated by a Hamiltonian function.

**Remark 5.** It follows directly from Definition 6 and from identifying \( C = M \times \mathbb{R} \) and \( \eta = \theta \) that the dynamics \( \dot{X}_H \) describing time–dependent symplectic Hamiltonian systems is the contact Hamiltonian system corresponding to the contact Hamiltonian \( \mathcal{H} = -1 \), and as such it is a very particular instance of a contact Hamiltonian system.
In our discussion we will consider more general instances, in the sense that the contact manifold $C$ is not restricted to be of the form $M \times \mathbb{R}$ and the Hamiltonian can be any function on $C$. Notice that for a general contact Hamiltonian system it follows directly from (24) that the derivative along the flow of the Hamiltonian is $\mathcal{L}_X \mathcal{H} = - R(\mathcal{H}) \mathcal{H} + \frac{\partial \mathcal{H}}{\partial t}$, meaning that $\mathcal{H}$ is not conserved. For instance, when $R(\mathcal{H})$ is a constant energy is constantly dissipated; however, more general dissipative terms may be allowed. A similarly direct calculation shows that $\mathcal{L}_X \mathcal{H} \eta \wedge (d\eta)^n = -(n+1) R(\mathcal{H}) \eta \wedge (d\eta)^n$, meaning that also the volume is contracted, to be compared with $\mathcal{L}_\tilde{X} \theta \wedge (d\theta)^n = 0$, as it happens for the geometric description of time–dependent symplectic Hamiltonian systems in the Poincaré–Cartan (or pre–symplectic) setting. Therefore we see that in the general contact case we have also a geometric description of dissipation, defined as a contraction of the relevant volume form.

As we did above where we introduced the standard model of contact manifolds in two different useful coordinate systems, we write here the corresponding models for contact Hamiltonian systems. Note that in each case we use the corresponding coordinates as in Examples 1 and 2, but in the second one we drop the tilde in the notation. In this way we can write the Reeb vector field as $R = \partial / \partial S$ in both cases, as can be seen by writing its definition in the corresponding coordinates.

**Lemma 1** (Contact Hamiltonian systems: std1). Given a (possibly time–dependent) differentiable function $\mathcal{H}(X, P, S, t)$ on the contact state space $(\mathbb{R}^{2n+1}, \eta_{std1})$, the associated contact Hamiltonian system is the following dynamical system

\[
\dot{X} = \nabla_P \mathcal{H} \tag{25}
\]
\[
\dot{P} = - \nabla_X \mathcal{H} - P \frac{\partial \mathcal{H}}{\partial S} \tag{26}
\]
\[
\dot{S} = \langle \nabla_P \mathcal{H}, P \rangle - \mathcal{H} \tag{27}
\]

**Lemma 2** (Contact Hamiltonian system: std2). Given a (possibly time–dependent) differentiable function $\mathcal{H}(X, P, S, t)$ on the contact state space $(\mathbb{R}^{2n+1}, \eta_{std2})$, the associated contact Hamiltonian system is the following dynamical system

\[
\dot{X} = \nabla_P \mathcal{H} - \frac{1}{2} X \frac{\partial \mathcal{H}}{\partial S} \tag{28}
\]
\[
\dot{P} = - \nabla_X \mathcal{H} - \frac{1}{2} P \frac{\partial \mathcal{H}}{\partial S} \tag{29}
\]
\[
\dot{S} = \frac{1}{2} \left( \langle X, \nabla_X \mathcal{H} \rangle + \langle P, \nabla_P \mathcal{H} \rangle \right) - \mathcal{H} \tag{30}
\]

The proofs of the above lemmas follow from writing explicitly the conditions in (24) for $\eta_{std1}$ and $\eta_{std2}$ respectively and using Cartan’s identity for the Lie derivative of a 1–form.

**Remark 6** (Lagrangian formulation). Contact systems can alternatively be introduced starting from the Lagrangian function $\mathcal{L}(X, V, S, t)$ and its corresponding generalized Euler–Lagrange equations

\[
\frac{d}{dt} \left( \frac{\partial \mathcal{L}}{\partial V} \right) - \frac{\partial \mathcal{L}}{\partial X} - \frac{\partial \mathcal{L}}{\partial S} \frac{\partial \mathcal{L}}{\partial V} \frac{\partial \mathcal{L}}{\partial S} = 0 \tag{31}
\]

together with the action equation

\[
\dot{S} = \mathcal{L}(X, V, S, t) \tag{32}
\]
Indeed, for regular Lagrangians it can be shown that (25)–(27) are equivalent to the system (31)–(32).

We arrive at our first main result with a direct calculation using equations (25)–(27) and (28)–(30).

**Proposition 1** (Recovering previous frameworks). All the previously-mentioned frameworks for describing continuous-time optimization methods can be recovered as follows:

i) If $H = H(X,P,t)$, that is, if $H$ does not depend explicitly on $S$, then from equations (25)–(26) we obtain the standard Hamiltonian equations (12)–(13), with (27) completely decoupled from the system. In particular, this includes the Bregman dynamics (12) and (13) as a special case of contact Hamiltonian systems.

ii) If $H = H(X,P) + cS$, then from equations (25)–(26) we obtain the standard equations for conformally symplectic systems (14)–(15), with (27) once again completely decoupled from the system.

iii) If $H = H(X,P) + \langle X^*, P \rangle - \langle P^*, X \rangle + 2S$, then from equations (28)–(29) we obtain the Hamiltonian descent equations (16)–(17), with (30) decoupled from the system.

iv) If $H = \frac{1}{2} \|P\|^2 + f(X) + \frac{3}{2}S$, then from equations (14)–(15) we obtain the continuous limit of NAG (3), with (30) decoupled from the system.

Two immediate and powerful consequences of Proposition 1 are the following.

**Corollary 1.** All the convergence analyses of the above dynamics provided in the corresponding literature hold.

In particular, by choosing the contact Hamiltonian $H = H_{Br}(X,P,t)$, and for the appropriate choice of the free functions $\alpha, \beta$ and $\gamma$ we can obtain contact systems with polynomial convergence rates of any order and even exponential convergence rates (see [29]). Thus we see that within the class of contact Hamiltonian systems we can at least reproduce all the conventional approaches to convergence rates from the Hamiltonian perspective. Moreover, contact systems provide the opportunity to generalize all these dynamics, for instance by fixing nonlinear dependences on the additional variable $S$, a strategy that will be considered in future works.

**Corollary 2.** All the systems above (Bregman dynamics, conformally symplectic systems, Hamiltonian descent equations and the continuous limit of NAG) share invariance under a very large group of transformations, the group of (time-dependent) contact transformations.

Notice that the property stated in Corollary 2 was unknown (and hard to prove) by previous approaches. Furthermore, it has interesting practical ramifications: on the one hand, invariance under all contact transformations guarantees that all such dynamics, and the corresponding algorithms obtained through geometric discretizations as discussed in the next section, are much less sensitive to the conditioning of the problem than other algorithms that do not possess a geometric structure. To see an illuminating discussion on this, we refer to [23] and we remark that contact transformations are more general than their symplectic counterparts ($MJMT = J$) considered therein in Section 2.1. On the other hand, one can exploit contact transformations to re-write the dynamics in coordinates where the system takes a simpler
form, with great benefit for the ensuing discretization. Due to the current central importance of the Bregman dynamics in the system dynamical approach to optimization (see also the more recent [27]), we will dedicate a companion paper to discuss this aspect and its utility in optimization [12]. We anticipate that the use of contact transformations leads to a version of the Bregman Hamiltonian that is separable and therefore much more amenable to explicit discretizations, thus reducing the computational burden and the instabilities involved in its discretization.

3 Discrete–time contact optimization

Before considering new optimization algorithms that stem from the discretization of contact Hamiltonian systems with geometric integrators, we will first prove the discrete–time analogue of item iv) in Proposition 1 and show that discrete–time NAG is not given by a dynamical system alone but rather a composition of a contact map and a gradient descent. This result is inspired by the conjecture put forward in [5], who argued that symplectic maps followed by gradient descent steps can generate the exponential convergence near convex optima empirically observed in discrete–time NAG. Here we provide an actual proof that NAG is based on the composition of a contact map and a gradient step.

Proposition 2 (Recovering NAG). Discrete–time NAG, (1)–(2), is given by the composition of a contact map and a gradient descent step.

Proof. First we recall from Definition 3 that a contact transformation for the contact structure given by (20) is a map that satisfies

\[
dS_{k+1} - \frac{1}{2} P_{k+1} dX_{k+1} + \frac{1}{2} X_{k+1} dP_{k+1} = f(X_k, P_k, S_k) \left( dS_k - \frac{1}{2} P_k dX_k + \frac{1}{2} X_k dP_k \right),
\]

for some function \(f(X_k, P_k, S_k)\) that is nowhere 0. Then we claim that NAG can be exactly decomposed in the contact state space as the composition of the map

\[
X_{k+1} = P_k \tag{34}
\]

\[
P_{k+1} = X_{k+1} + \frac{k - 1}{k + 2} (X_{k+1} - X_k), \tag{35}
\]

\[
S_{k+1} = \frac{k - 1}{k + 2} S_k, \tag{36}
\]

which is readily seen to be a contact transformation satisfying

\[
dS_{k+1} - \frac{1}{2} P_{k+1} dX_{k+1} + \frac{1}{2} X_{k+1} dP_{k+1} = \frac{k - 1}{k + 2} \left[ dS_k - \frac{1}{2} P_k dX_k + \frac{1}{2} X_k dP_k \right], \tag{37}
\]

followed by a standard gradient descent map,

\[
X_{k+1} = X_k - \tau \nabla f(X_k) \tag{38}
\]

\[
P_{k+1} = P_k \tag{39}
\]

\[
S_{k+1} = S_k. \tag{40}
\]
Remark 7. The fact that NAG has a latent geometric nature is already a step forward towards understanding its effectiveness. It is also of interest that we have been able to prove exactly the conjecture in [5] that discrete–time NAG can be obtained by composing structure–preserving maps, in this case a contact transformation, with gradient descent steps. In light of our result, it seems that this can indeed be the intrinsic mechanism responsible for the late–stage exponential convergence so often seen in NAG. We will not pursue this direction here, leaving it to future work.

We now review a more systematic procedure to discretize contact Hamiltonian systems, which, when applied e.g. to any of the Hamiltonians in Proposition 1 leads to new optimization algorithms. First we introduce the following lemmas from [11] which show how and when we can construct contact integrators of any even order.

Lemma 3 (Second–order contact integrator). Let the possibly time–dependent contact Hamiltonian be separable into the sum of functions

\[ H(X, P, S, t) = \sum_{j=1}^{n} \phi_j(X, P, S, t), \]

such that each of the associated contact Hamiltonian vector fields \( X_{\phi_j} \) are exactly integrable. Then the integrator

\[ S_2(\tau) = e^{\tau \partial_t} e^{\tau X_{\phi_1}} e^{\tau X_{\phi_2}} \cdots e^{\tau X_{\phi_n}} e^{\tau \partial_t} \]

is a second–order contact integrator.

Remark 8. It should be noticed that the assumption that the vector fields \( X_{\phi_j} \) are exactly integrable is not necessary and can be relaxed by using any second–order contact integrator instead of the exact flow for each field. In the literature one usually refers to composition methods in the former case and to splitting methods in the latter [21]. This extremely enlarges the potential applications of the methods presented here (see e.g. the discussion in [5]). However, for the purposes of this work we will only need cases in which each of the vector fields is exactly integrable.

Moreover, there is a standard procedure to combine such integrators that yields higher–order integrators, that is, numerical integrators that follow the original dynamics with higher accuracy, at the price of requiring more gradient evaluations. This is the content of the next two lemmas.

Lemma 4 (Higher–order integrator with exact coefficients). If \( S_{2n}(\tau) \) is an integrator of order \( 2n \) then the map

\[ S_{2n+2}(\tau) = S_{2n}(z_1 \tau) S_{2n}(z_0 \tau) S_{2n}(z_1 \tau), \]

with \( z_0 \) and \( z_1 \) given by

\[ z_0(n) = -\frac{2^{1-n}}{2 - 2^{1-n}}, \quad z_1(n) = \frac{1}{2 - 2^{1-n}}, \]

is an integrator of order \( 2n + 2 \).

Lemma 5 (Higher–order integrator with approximated coefficients). There exist \( m \in \mathbb{N} \) and a set of real coefficients \( \{w_j\}_{j=0}^{m} \) such that the map

\[ S^{(m)}(\tau) = S_2(w_m \tau) S_2(w_{m-1} \tau) \cdots S_2(w_0 \tau) S_2(w_{m-1} \tau) S_2(w_m \tau), \]

11
is an integrator of order $2n$. The coefficients $w_1, \ldots, w_m$ are obtained as approximated solutions to an algebraic equation derived from the Baker–Campbell–Hausdorff formula.

We refer to [11] for the proofs of these lemmas and to [11, 28] for the analysis of the corresponding geometric integrators.

Consequently all we need to find in order to obtain contact integrators for contact Hamiltonian systems is a splitting of the corresponding contact Hamiltonian into a sum of contact Hamiltonians whose vector fields are exactly integrable.

We end this section with the following two remarks, that highlight some of the ongoing work in the context of a deeper understanding of the relationship between geometric integration and acceleration in optimization.

**Remark 9** (Matching rates). Given that the thus–obtained optimization algorithms are based on contact integrators of any even order, in principle one can use backward error analysis and show that the convergence rates of the corresponding discrete maps match those of the continuous differential equation up to the order of the integrator. See, for example, the discussion in [17] and the main result in [18]. This goes beyond the scope of the present work and will be presented elsewhere.

**Remark 10** (Increased stability). Since contact integrators are very stable under the increase of the step size $\tau$ (see [11]), we can use larger steps than standard optimization algorithms and achieve an effective increase in the rate of convergence. In particular, the higher the order of the integrator, the larger we can choose $\tau$. See, for example, Example 3 below (although, see also [30] for some very recent results that show that the geometric character of the integrator is not always essential for acceleration).

### 3.1 Contact relativistic gradient descent

As an example of how the contact geometry framework can guide the generalization of optimization algorithms, let us consider generalizing the Relativistic Gradient Descent proposed in [17].

We start with a brief introduction of the main features of RGD, referring to [17] for more details. At the continuous level, the dynamical system corresponding to RGD is obtained by taking a Hamiltonian energy of the type describing a relativistic particle subject to a potential $f(X)$, that is $H(X, P) = c\sqrt{|P|^2 + (mc)^2 + f(X)}$, and then plugging it into the Hamiltonian equations for conformally symplectic systems (14)–(15). In this case the energy $H(X, P)$ is an obvious Lyapunov function and therefore convergence to the desired minima can be guaranteed. To discretize this system, in [17] the authors provided an ingenious discretization that has several desirable properties, such as the fact that it can be viewed as an interpolation between a relativistic generalization of classical momentum (CM) and NAG, depending on the value of a hyperparameter $\alpha$, whose value ranges between 0 (generalizing NAG) and 1 (generalizing CM). Moreover, by construction when $\alpha = 1$ (and only at such value), such discretization is geometric, meaning conformally symplectic, providing a second order (relativistic) version of CM. Finally, a remarkable property of RGD is that the implementation of the relativistic kinetic energy yields a stabilization mechanism that is seen to be very effective both by linear stability analysis and especially by means of extensive numerical experiments.
In the following we will propose a generalization of RGD based on the following observations: on the one hand, since in this work we focus only on geometric algorithms, we will extend only the geometric discretization given in [17], that is, the one corresponding to the value $\alpha = 1$. Indeed, as the authors find in [17], most of the experiments have shown a preference for such value of the hyperparameter.

We shall also illustrate this point in the numerical experiments, by comparing the results of our geometric integrator with those obtained by optimizing the parameter $\alpha$ in RGD. On the other hand, contrary to [17], we consider a more general geometric setting, the contact geometry, which allows us to consider e.g. an explicit time dependence in the Hamiltonian function. Our purpose in this section is to show that even such an apparently immediate generalization can lead to interesting results, for a right choice of the free function $h(t)$ and when combined with the appropriate geometric integrators.

For all the above reasons, we consider the following contact extension of the Relativistic Gradient Descent (RGD) introduced in [17].

**Proposition 3** (Contact Relativistic Gradient Descent). We start with the contact Hamiltonian

$$H(X, P, S, t) = c\sqrt{||P||^2 + (mc)^2} + f(X) + h(t)S,$$

with a time–dependent dissipative term $h(t) = \gamma(1 + \frac{1}{t})$ instead of the constant factor $\gamma$ which according to Proposition 1.i) would give the corresponding conformally symplectic system considered in [17]. Splitting the contact Hamiltonian into the sum

$$H_{\phi_1} = h(t)S,$$

$$H_{\phi_2} = f(X),$$

$$H_{\phi_3} = c\sqrt{||P||^2 + (mc)^2},$$

yields the following second–order contact integrator,

$$S_2(\tau) = e^{\tau} e^{\tau h(t)} e^{\tau} X_{\phi_1} e^{\tau} X_{\phi_2} e^{\tau} X_{\phi_3} e^{\tau} X_{\phi_1} e^{\tau} X_{\phi_2} e^{\tau} X_{\phi_3} e^{\tau} X_{\phi_1} e^{\tau} X_{\phi_2} e^{\tau} X_{\phi_3},$$

where each map is given explicitly by

$$e^{\tau X_{\phi_1}} \begin{bmatrix} X \\ P \\ S \\ t \end{bmatrix} = \begin{bmatrix} X \\ P \\ S \\ t + \frac{c}{2} \end{bmatrix}$$

$$e^{\tau X_{\phi_2}} \begin{bmatrix} X \\ P \\ S \\ t \end{bmatrix} = \begin{bmatrix} X - \nabla f(X) \tau \\ P - \nabla f(X) \tau \\ S - f(X) \tau \\ t \end{bmatrix}$$

$$e^{\tau X_{\phi_3}} \begin{bmatrix} X \\ P \\ S \\ t \end{bmatrix} = \begin{bmatrix} X \frac{cP}{\sqrt{||P||^2 + (mc)^2}} \tau \\ P \frac{cP}{\sqrt{||P||^2 + (mc)^2}} \tau \\ S \frac{cP}{\sqrt{||P||^2 + (mc)^2}} \tau \\ t \end{bmatrix}.$$

These discretized dynamics define a new accelerated optimization algorithm that we call Contact Relativistic Gradient Descent (CRGD). Following [17], we define the new parameters

$$V_k \equiv \frac{\tau}{2m} P_k, \quad \epsilon \equiv \frac{\tau^2}{2m}, \quad \mu \equiv e^{-\gamma \tau}, \quad \delta \equiv \frac{4}{(c\tau)^2},$$

13
and we further introduce
\[ \mu_t \equiv \mu^{1+\frac{1}{t}} \] (54)
so that we can finally write the algorithm in the form
\[
X_{k+1/2} = X_k + \sqrt{\mu_{t+1/2}} \frac{V_k}{\sqrt{\mu_{t+1/2} \delta \|V_k\|^2 + 1}}
\] (55)
\[
V_{k+1/2} = \sqrt{\mu_{t+1/2}} V_k - \epsilon \nabla f(X_{k+1/2})
\] (56)
\[
X_{k+1} = X_{k+1/2} + \frac{V_{k+1/2}}{\sqrt{\delta \|V_{k+1/2}\|^2 + 1}}
\] (57)
\[
V_{k+1} = \sqrt{\mu_{t+1/2}} V_{k+1/2}
\] (58)
\[
S_{k+1} = \mu_{t+1/2} S_k - \frac{2\sqrt{\mu_{t+1/2}}}{\sqrt{\delta c}} \left[ \frac{1}{e^\delta} \left( \frac{1}{\sqrt{\delta \mu_{t+1/2} \|V_k\|^2 + 1}} + \frac{1}{\sqrt{\delta \|V_{k+1/2}\|^2 + 1}} \right) - f(X_{k+1/2}) \right]
\] (59)

**Remark 11.** Note that, as equation (59) is decoupled from (55)–(58), one can equivalently think of CRGD as being the algorithm (55)–(58).

**Remark 12.** If we take \( h(t) = \gamma \) applied to only the variables \( (X, P) \) then we re-obtain the RGD algorithm proposed in [17] (for the case \( \alpha = 1 \)), which, as is the case for any other algorithm based on conformally symplectic systems, can be considered as a particular contact optimization algorithm.

**Remark 13.** Let us explain now the particular choice of the function \( h(t) \) used in CRGD. If we fix \( \mu = e^{-\gamma t} \), then we can rewrite the map for \( X_\phi \) as
\[
\begin{bmatrix}
X \\
P \\
S \\
t
\end{bmatrix} =
\begin{bmatrix}
X \\
P \mu^{\frac{1}{2}(1+\frac{1}{t})} \\
S \mu^{\frac{1}{2}(1+\frac{1}{t})} \\
t/n\mu
\end{bmatrix}.
\] (60)

The dissipation parameter \( \mu \in (0, 1) \) in RGD is constant and therefore it has to be carefully tuned; this is because in regions where \( f(X) \) has a high curvature one would prefer \( \mu \) to be small (in order to suppress the momentum as much as possible) while in flat regions one would like it to be as close to 1 as possible (in order to suppress the momentum as little as possible). Our dynamically tuned \( \mu_t \) does exactly so. Indeed, if we assume convex functions that are relatively flat around the minimum, then the dynamics are likely to start in a region of high curvature before converging to the low curvature near the optimum. In other words we would want \( \mu_t \) to transition from a small value to a larger value as the system evolves, which it does, as it interpolates between 0 and \( \mu < 1 \) with increasing time. Consequently this choice should improve convergence for objective functions that exhibit nearly flat regions. This explains how we decided to fix the functional form for \( h(t) \). We explore this possibility in the numerical experiments below.

This is just one example of the type of reasoning that can guide the generalization from standard symplectic and conformally symplectic to contact optimization algorithms.
4 Numerical experiments

In this section we compare CM, NAG, RGD and CRGD on some benchmark examples. Importantly, for RGD we use the same range for the parameters as in [17], which means that in particular in this section we also vary the $\alpha$ parameter for what regards RGD, in order to understand if there is some advantage in these cases to use a non-geometric and NAG-like integrator over the second order contact splitting integrator that corresponds to CRGD (see Table 1). This will give an excellent heuristic of the fact that indeed the two schemes behave similarly at best (that is, when the NAG-like version of RGD has its best value), and therefore a geometric approach with less hyperparameters may be preferable. This, together with the fact that CRGD has a less stringent dependence on the specific value of the parameter $\mu$ due to the careful choice of the function $h(t)$, hints to the study of CRGD as a viable alternative for large-scale optimization problems.

| Example | $\alpha$ |
|---------|----------|
| 3       | 0.74     |
| 4       | 0.2      |
| 5 (b)   | 0.89     |
| 5 (c)   | 0.67     |
| 6 (b)   | 0.33     |
| 6 (c)   | 0.84     |

Table 1: Mean for the estimated value of $\alpha$ for RGB in each example. Recall that $\alpha \in [0, 1]$ and that a value $\alpha = 0$ corresponds to a “NAG-like” integrator for the conformally symplectic relativistic system, introducing a spurious dissipative term, while a value $\alpha = 1$ corresponds to a geometric integrator. See Section 5 for details on the tuning process.

Example 3 (Quadratic function). Let us start with a simple quadratic function

$$f(X) = \frac{1}{2} X^T A X, \quad X \in \mathbb{R}^{500}, \quad \lambda(A) \sim U(10^{-3}, 1), \quad (61)$$

where $A \in \mathbb{R}^{500 \times 500}$ is a positive–definite random matrix with eigenvalues uniformly distributed over the range $[10^{-3}, 1]$. In Figure 1 we show the convergence rate of each algorithm when minimizing $f(X)$ in each of 50 Monte Carlo simulations. We see that all the algorithms behave similarly in such a simple case. Note also that for RGD we obtain $\alpha \sim 0.74$ (see Table 1), meaning that an extra “NAG-like” boost is needed for such algorithm to perform at best. However this implies a decrease in stability as e.g. the range of allowed time steps is shrunk (cf. the discussion on such “trade-off” in [17]).
Figure 1: All the algorithms display a similar convergence for the random quadratic function (61). The initial state for each run is always $X_0 = (1, \ldots, 1)^T$, $P_0 = 0$ (and $S_0 = 0$ for CRGD). Solid lines are the median, and shaded areas are the quantiles 0.025 and 0.975.

**Example 4** (Quartic function). Next let us consider the quartic function

$$f(X) = \sum_{i=1}^{n} iX_i^4, \quad n = 50. \quad (62)$$

This function is an example of a convex function that has a global minimum at $f(0) = 0$, surrounded by a very flat region, as seen in Fig. 2(a). For this reason in this case we expect CRGD to outperform RGD. In Fig. 2(b) we set the initial state $X_0 = (2, \ldots, 2)^T$, $P_0 = 0$ (and $S_0 = 0$ for CRGD). Indeed, we observe that CRGD converges the fastest. Such good performance of CRGD is due to the time-dependent momentum term $\mu_t$, allowing it to transition from a small value to a larger value as the system goes from an area of high curvature to an area of low curvature near the minimum. Note also that for RGD we obtain $\alpha \sim 0.2$ (see Table 1), meaning that an extra “NAG-like” boost is needed for such algorithm to perform at best. However this implies a decrease in stability as e.g. the range of allowed time steps is shrunk (cf. the discussion on such “trade-off” in [17]).

Figure 2: (a) A slice of the quartic function (62) where $X = (X_1, X_2, 0, \ldots, 0)$. (b) CRGD exhibits the best convergence when targeting a convex objective function with a high curvature far from the minimum and a flat region near the minimum.
**Example 5** (Camelback Function). To push the algorithms further we consider the nonconvex Camelback objective function,

\[ f(X_1, X_2) = 2X_1^2 - 1.05X_1^4 + \frac{1}{6}X_1^6 + X_1X_2 + X_2^2. \]  

(63)

The contour plot in Fig. 3(a) demonstrates the multimodality of this objective function, with three locally convex neighborhoods separated by nonconvex valleys. A unique global minimum can be found at \( f(0, 0) = 0 \) while two local minima are located at \((X_1^+, X_2^+) \approx \pm(-1.75, 0.87)\) with \( f(X_1^+, X_2^+) \approx 0.3 \). In Fig. 3(b) we set the initial state \( X_0 = (5, 5)^T \), \( P_0 = 0 \), and \( S_0 = 0 \) for CRGD. RGD and CRGD show a similar convergence, clearly outperforming CM and NAG. For Fig. 3(c) we repeat the same experiment but initialize each algorithm close to one of the local minimizers. While CM and NAG are unable to escape the local minimum, both RGD and CRGD escape to the global minimum, with CRGD converging to the global minimum slightly faster. For a quantitative comparison, we report the numerical estimation for the rate of convergence of RGD and CRGD in Table 2. In order to compare the sensitivity with respect to the tuning of the parameters between RGD and CRGD in such a demanding task, in Figs. 3(d)-(e) we perform 500 experiments where the initial position is sampled uniformly in the range \(-5 \leq X_{0,i} \leq 5\), while the parameters are held fixed to the optimal values used for Figs. 3(b)-(c) respectively. Again, we observe a similar behavior between RGD and CRGD, although CRGD slightly outperforms RGD in speed and guarantees a smaller (almost invisible) dispersion, thus providing a good heuristic of the fact that indeed CRGD is less sensitive to the choice of the parameters. Indeed, in Fig. 3(e) there is a much larger dispersion for RGD than in Fig. 3(d), indicating the fact that when RGD is optimized for an initial condition that happens to be near a local minimum, then the value found for \( \mu \) is not optimal for other initial conditions, a situation which clearly does not happen with CRGD. Note also that for RGD \( \alpha \sim 0.89 \) for Figs. 3(b) and 3(d) while \( \alpha \sim 0.67 \) for Figs. 3(c) and 3(e) (see Table 1).

**Example 6** (Rosenbrock Function). For a higher–dimensional challenge, let us consider the nonconvex Rosenbrock function,

\[ f(X) = \sum_{i=1}^{n-1} \left( 100(X_{i+1}^2 - X_i^2)^2 + (1 - X_i)^2 \right), \]  

(64)
Figure 3: (a) The Camelback function (63) features three modes, one global minimum in the center surrounded by two local minima. (b) When initialized away from all of the minima, $X_0 = (5, 5)^T$, CRGD converges the fastest. (c) When initialized near one of the local minima, $X_0 = (1.8, -0.9)^T$, CRGD continues to dominate. (d)-(e) When we perform 500 experiments with the initialization chosen at random within $-5 \leq X_{0,i} \leq 5$ at fixed values of the hyperparameters, RGD and CRGD perform similarly. Here solid lines are the median and shaded areas are the quantiles 0.025 and 0.975. We observe that CRGD has practically no dispersion, indicating that fixed values of the parameters fit a large region of initial conditions.

with $n = 100$ dimensions. The Rosenbrock landscape is quite complex; for instance there are only two local minimizers, one global at $X^* = (1, 1, \ldots, 1)^T$ where $f(X^*) = 0$, and one local near $X \approx (-1, 1, \ldots, 1)^T$. RGD and CRGD demonstrate faster convergence than CM and NAG when the algorithms are initialized close to the local minimum, $X_{0,2i} = 1$ and $X_{0,2i-1} = -1.2$, $i = 1, \ldots, 50$ (Fig. 4(b)), and when initialized in the tails of the Rosenbrock function, $X_{0,2i} = 5$ and $X_{0,2i-1} = -5$, $i = 1, \ldots, 50$ (Fig. 4(c)). Furthermore, in order to compare the sensitivity with respect to the tuning of the parameters between RGD and CRGD in such a demanding task, in Figs. 4(d)-(e) we perform 500 experiments where the initial position is sampled uniformly in the range $-2.048 \leq X_{0,i} \leq 2.048$, while the parameters are held fixed to the optimal values used for Figs. 4(b)-(c) respectively. All the trajectories for CM and NAG diverge, while we observe a similar behavior between RGD and CRGD, with CRGD again guaranteeing a smaller (almost invisible) dispersion, thus providing a good heuristic of the fact that indeed CRGD is less sensitive to the choice of the parameters. Indeed, in Fig. 4(e) there is a much larger dispersion.
for RGD than in Fig. 4(d), indicating the fact that when RGD is optimized for an initial condition that happens to be in a region far from the global minimum, then the value found for $\mu$ is not optimal for other initial conditions, a situation which clearly does not happen with CRGD. Note also that for RGD $\alpha \sim 0.33$ for Figs. 4(b) and 4(d) while $\alpha \sim 0.84$ for Figs. 4(c) and 4(e) (see Table 1).

![Figure 4](image)

**Figure 4:** (a) A slice of the Rosenbrock function (64) where $X = (X_1, X_2, 1, \ldots, 1)$ demonstrates the complex landscape that can frustrate optimization algorithms. RGD and CRGD converge fast when the optimization algorithms are initialized (b) close to and (c) far from the local minimum. (d)-(e) When we perform 500 experiments with the initialization chosen at random within $-2.048 \leq X_{0,i} \leq 2.048$ at fixed values of the hyperparameters, RGD and CRGD perform similarly. Here solid lines are the median and shaded areas are the quantiles 0.025 and 0.975. We observe that CRGD has practically no dispersion, indicating that fixed values of the parameters fit a large region of initial conditions.

## 5 Numerical Details

In this section we describe the tuning process for the parameters used in the examples of Section 4. In all the experiments we use an exhaustive random search on parameter space and the same number of Monte Carlo runs for each algorithm. Moreover we always set $P_0 = 0$ and $S_0 = 0$. All the numerical implementations have been performed in Python, using the scipy, numpy and matplotlib packages. The codes are available in a Github repository (see [9]).
Quadratic function

In Fig. 1 we performed 50 samples of $A$ and for each sample we ran each algorithm 150 times and for 200 iterations, and chose the parameters that give the lowest objective function value. Search ranges are shown in Table 3.

| Algorithm | $\tau$ | $\epsilon$ | $\mu$ | $\delta$ | $\alpha$ |
|-----------|--------|------------|-------|---------|--------|
| CM        | $[10^{-2}, 8 \cdot 10^{-1}]$ | [0.8, 0.99] |       |         |       |
| NAG       | $[10^{-3}, 5 \cdot 10^{-1}]$ | [0.8, 0.99] |       |         |       |
| RGD       | $[0, 6 \cdot 10^{-1}]$ | $[0.49, 0.95]$ | $[0, 20]$ | $[0, 1]$ |       |
| CRGD      | $[0, 6 \cdot 10^{-1}]$ | $[0.49, 0.95]$ | $[0, 20]$ |       |       |

Table 3: Ranges for parameters search in the experiment of Fig. 1.

Quartic function

For the experiments in Fig. 2, we run each algorithm 1000 times and for 500 iterations, and choose the parameters which give the lowest objective function value. Search ranges are shown in Table 4.

| Algorithm | $\tau$ | $\epsilon$ | $\mu$ | $\delta$ | $\alpha$ |
|-----------|--------|------------|-------|---------|--------|
| CM        | $[10^{-5}, 10^{-1}]$ | [0.8, 0.99] |       |         |       |
| NAG       | $[10^{-5}, 10^{-1}]$ | [0.8, 0.99] |       |         |       |
| RGD       | $[10^{-5}, 10^{-2}]$ | $[0.6, 0.99]$ | $[0, 30]$ | $[0, 1]$ |       |
| CRGD      | $[10^{-5}, 10^{-2}]$ | $[0.6, 0.99]$ | $[0, 30]$ |       |       |

Table 4: Ranges for parameters search in the experiment of Fig. 2.

Camelback Function

For the experiments in Fig. 3(b)–(c), we run each algorithm 1500 times and for 300 iterations, and choose the parameters which give the lowest objective function value. Search ranges are show in Table 5. For the experiments in Fig. 3(d)-(e) we performed 100 samples of the initial position, where each sample is chosen uniformly in the range $-5 \leq X_0 \leq 5$, using the optimal values of the parameters as for Fig. 3(b)-(c), respectively.

Rosenbrock Function

In the experiments in Fig. 4(b)–(c), we run each algorithm 500 times and for 1200 iterations, and choose the parameters which give the lowest objective function value.
6 Conclusions

In this paper we have demonstrated that contact geometry, and contact Hamiltonian systems, naturally generate the dissipative dynamical systems whose discretizations give rise to accelerated gradients algorithms. Not only do these geometric systems subsume a wide range of dynamical systems previously considered in the literature, their geometric integration provides a principled means of constructing discretizations that preserve the important structure of the latent dynamics. Quite remarkably NAG itself can be decomposed as a contact transformation followed by a standard gradient descent step, demonstrating the fundamental nature of these systems.

We expect that unifying the development of optimization algorithms through this contact geometric lens will allow us to not only better understand these algorithms but also identify how their structure translates to ultimate performance and hence derive improved algorithms more effectively. As a preliminary example we have shown that the RGD algorithm can be immediately generalized to the contact case, resulting in a more stable and generally faster algorithm.

The geometric foundation also brings with it a wealth of mathematical tools for the thorough analysis of the convergence of these algorithms, which we defer to subsequent work.
Acknowledgements

We would like to thank Diego Tapias, Mario Díaz, and Shin-itiro Goto for valuable comments, and Ana Pérez Arteaga and Ramiro Chávez Tovar for their help.

References

[1] Abraham R. & Marsden J. E. *Foundations of mechanics*. Benjamin/Cummings Publishing Company, Massachusetts, 1978.

[2] Arnold V. I. *Mathematical methods of classical mechanics*. Volume 60 of Graduate Texts in Mathematics. Springer-Verlag, New York, 1989.

[3] Betancourt M. *A conceptual introduction to Hamiltonian Monte Carlo*. arXiv:1701.02434, 2017.

[4] Betancourt M., Byrne S., Livingstone S., Girolami M. et al. *The geometric foundations of Hamiltonian Monte Carlo*. Bernoulli, 23:2257–2298, 2017.

[5] Betancourt M., Jordan M. I. & Wilson A. C. *On symplectic optimization*. arXiv:1802.03653, 2018.

[6] Bravetti A. *Contact Hamiltonian dynamics: The concept and its use*. Entropy, 19, 2017.

[7] Bravetti A. *Contact geometry and thermodynamics*. International Journal of Geometric Methods in Modern Physics: 1940003, 2018.

[8] Bravetti A., Cruz H. & Tapias D. *Contact Hamiltonian mechanics*. Annals of Physics, 376:17–39, 2017.

[9] Bravetti A., Daza-Torres M. L., Flores-Arguedas H. & Betancourt M. *Contact optimization algorithms*. https://github.com/mdazatorres/Contact-optimization-algorithms, 2020.

[10] Bravetti A., de León M., Marrero J. C. & Padrón E. *Invariant measures for contact Hamiltonian systems: symplectic sandwiches with contact bread*. Journal of Physics A: Mathematical and Theoretical, 53:455205, 2020.

[11] Bravetti A., Seri M., Vermeeren M. & Zadra F. *Numerical integration in celestial mechanics: a case for contact geometry*. Celestial Mechanics and Dynamical Astronomy, 132:1–29, 2020.

[12] Bravetti A., Daza-Torres M. L. & Michael B. *Bregman dynamics, contact transformations and optimization*. in preparation, 2021.

[13] Ciaglia F. M., Cruz H. & Marmo G. *Contact manifolds and dissipation, classical and quantum*. Annals of Physics, 398:159–179, 2018.

[14] de León M. & Lainz Valcázar M. *Contact Hamiltonian systems*. Journal of Mathematical Physics, 60:102902, 2019.

[15] de León M. & Sardón C. *Cosymplectic and contact structures for time-dependent and dissipative Hamiltonian systems*. Journal of Physics A: Mathematical and Theoretical, 50:255205, 2017.

[16] Diakonikolas J. & Jordan M. I. *Generalized momentum-based methods: a Hamiltonian perspective*. SIAM Journal on Optimization, 31:915–944, 2021.

[17] França G., Sulam J., Robinson D. P. & Vidal R. *Conformal symplectic and rela-
tivistic optimization. Journal of Statistical Mechanics: Theory and Experiment, 2020:124008, 2020.

[18] França G., Jordan M. I. & Vidal R. On dissipative symplectic integration with applications to gradient-based optimization. Journal of Statistical Mechanics: Theory and Experiment, 2021:043402, 2021.

[19] Gaset J., Gràcia X., Muñoz-Lecanda M. C., Rivas X. & Román-Roy N. New contributions to the Hamiltonian and Lagrangian contact formalisms for dissipative mechanical systems and their symmetries. International Journal of Geometric Methods in Modern Physics, 17:2050090, 2020.

[20] Geiges H. An introduction to contact topology. Volume 109. Cambridge University Press, 2008.

[21] Hairer E., Lubich C. & Wanner G. Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations. Volume 31 of Springer Series in Computational Mathematics. Springer-Verlag, Berlin Heidelberg, 2006.

[22] Livingstone S., Faulkner M. F. & Roberts G. O. Kinetic energy choice in Hamiltonian/hybrid Monte Carlo. Biometrika, 106:303–319, 2019.

[23] Maddison C. J., Paulin D., Teh Y. W., O’Donoghue B. & Doucet A. Hamiltonian descent methods. arXiv:1809.05042, 2018.

[24] Muehlebach M. & Jordan M. A dynamical systems perspective on Nesterov acceleration. In International Conference on Machine Learning, pages 4656–4662. 2019.

[25] O’Donoghue B. & Maddison C. J. Hamiltonian descent for composite objectives. In NeurIPS, pages 14443–14453. http://papers.nips.cc/paper/9590-hamiltonian-descent-for-composite-objectives, 2019.

[26] Su W., Boyd S. & Candès E. J. A differential equation for modeling Nesterov’s accelerated gradient method: Theory and insights. Journal of Machine Learning Research, 17:1–43. http://jmlr.org/papers/v17/15-084.html, 2016.

[27] Sun C. & Hu G. A continuous-time Nesterov accelerated gradient method for centralized and distributed online convex optimization. arXiv:2009.12545, 2020.

[28] Vermeeren M., Bravetti A. & Seri M. Contact variational integrators. Journal of Physics A: Mathematical and Theoretical, 52:445206, 2019.

[29] Wibisono A., Wilson A. C. & Jordan M. I. A variational perspective on accelerated methods in optimization. Proceedings of the National Academy of Sciences, 113:E7351–E7358, 2016.

[30] Zhang P., Orvieto A., Daneshmand H., Hofmann T. & Smith R. S. Revisiting the role of Euler numerical integration on acceleration and stability in convex optimization. In International Conference on Artificial Intelligence and Statistics, pages 3979–3987. 2021.