RESEARCH PAPER

Bregman dynamics, contact transformations and convex optimization

Alessandro Bravetti[1] · Maria L. Daza-Torres[2] · Hugo Flores-Arguedas[3] · Michael Betancourt[4]

Abstract

Recent research on accelerated gradient methods of use in optimization has demonstrated that these methods can be derived as discretizations of dynamical systems. This, in turn, has provided a basis for more systematic investigations, especially into the geometric 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. As a consequence, all the deterministic flows used in optimization share an extremely interesting geometric property: they are invariant under contact transformations. In our main result, we exploit this observation to show that the celebrated Bregman Hamiltonian system can always be transformed into an equivalent but separable Hamiltonian by means

Maria L. Daza-Torres, Hugo Flores-Arguedas and Michael Betancourt contributed equally to this work.

✉ Alessandro Bravetti
alessandro.bravetti@iimas.unam.mx

Maria L. Daza-Torres
mdazatorres@ucdavis.edu

Hugo Flores-Arguedas
hfloresarguedas@astate.edu

Michael Betancourt
betan@symplectomorphic.com

[1] Departamento de Matemáticas y Mecánica, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, A. P. 70543, Mexico City 04510, Mexico

[2] Department of Public Health Sciences, University of California Davis, Davis, CA 95616, USA

[3] Arkansas State University, Campus Querétaro (ASUCQ), Querétaro, Mexico

[4] Symplectomorphic LLC, New York, USA

🖄 Springer

of a contact transformation. This in turn enables the development of fast and robust discretizations through geometric *contact splitting integrators*. As an illustration, we propose the Relativistic Bregman algorithm, and show in some paradigmatic examples that it compares favorably with respect to standard optimization algorithms such as classical momentum and Nesterov's accelerated gradient.

Keywords Convex optimization · Bregman Hamiltonian · 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 turn hinders the development of new and more principled schemes. Recently a progression of work has studied the continuum limit of accelerated gradient methods, demonstrating that these methods can be derived as discretizations of continuous dynamical systems. Shifting the 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 [1], 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 similar accelerated gradient methods that achieved comparable convergence rates.

Later, in [2] the authors found what is arguably still the most important development in this direction, by showing that accelerated methods can also be derived as discretizations of a more structured family of *variational* dynamical systems, specified with a time-dependent Lagrangian function, or equivalent Hamiltonian function. Consider an objective function $f : \mathcal{X} \to \mathbb{R}$, which is continuously differentiable, convex, and has a unique minimizer $X^* \in \mathcal{X}$. Moreover assume that $\mathcal{X}$ is a convex set endowed with a distance-generating function $h : \mathcal{X} \to \mathbb{R}$ that is also convex and essentially smooth. 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*

$$L_{\mathrm{Br}}(X, V, t) = \mathrm{e}^{\alpha+\gamma} \left(D_h(X + \mathrm{e}^{-\alpha}V, X) - \mathrm{e}^{\beta} f(X) \right), \tag{5}$$

where α, β, and γ are continuously differentiable functions of time. They then proved that under the *ideal scaling conditions*

$$\dot{\beta} \leq \mathrm{e}^{\alpha} \tag{6}$$

$$\dot{\gamma} = \mathrm{e}^{\alpha}, \tag{7}$$

the solutions of the resulting Euler–Lagrange equations

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

satisfy [2, Theorem 1.1]

$$f(X) - f(X^*) \leq \mathcal{O}(\mathrm{e}^{-\beta}). \tag{9}$$

From a physical perspective the two terms in Eq. (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 $\dot{\beta} = \mathrm{e}^{\alpha}$, i.e. $\beta = \int_{t_0}^{t} \mathrm{e}^{\alpha(s)} ds$, 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

$$\ddot{X} + \left(\mathrm{e}^{\alpha} - \dot{\alpha} \right) \dot{X} + \mathrm{e}^{2\alpha+\beta} \nabla f(X) = 0. \tag{10}$$

Finally the authors developed a heuristic discretization of (8) that yielded optimization algorithms matching the continuous convergence rates. However, these algorithms are of no practical use, due to the extremely high cost of their implementation.

In [3] the authors 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 [4].

In particular they considered the associated *Bregman Hamiltonian* [2],

$$H_{\mathrm{Br}}(X, P, t) = \mathrm{e}^{\alpha+\gamma} \left(D_{h^*}(\mathrm{e}^{-\gamma} P + \nabla h(X), \nabla h(X)) + \mathrm{e}^{\beta} f(X) \right), \tag{11}$$

where $h^*(P) := \sup_V \langle P, V \rangle - h(X)$ 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

$$\dot{X} = \nabla_P H_{\mathrm{Br}}(X, P, t) \tag{12}$$

$$\dot{P} = -\nabla_X H_{\mathrm{Br}}(X, P, t), \tag{13}$$

is with *pre-symplectic integrators* in an extended phase space where t and its associated momentum E are added as dynamical variables. They numerically demonstrated in the Euclidean (i.e., separable) 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 then proposed in [5], 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. The crucial significance of this result lies in the fact that it implies that splitting integrators automatically yield "rate-matching" algorithms, without the need for a discrete convergence analysis. Ideally, this one would like to be able to directly apply splitting integrators to the general Bregman Hamiltonian. Unfortunately, however, when applied to the general (non-separable) Bregman dynamics, these splitting methods yield implicit updates that increase the computational burden and affect the numerical stability. As a first attempt to remedy this problem, in [5] the authors used a strategy that involves doubling the phase space dimension. As we will comment in the following, this strategy is not completely satisfying (see also the discussion in [6]). Other attempts have been proposed later in [6, 7], based respectively on the so-called *Poincaré transform* and the *adaptive Hamiltonian variational integrators* in the first case and on the *cosymplectic geometry* and the associated discrete variational methods for time-dependent Lagrangians in the second. Although both perspectives seem to provide a robust solution to the problem, they are rather sophisticated compared to the simplicity of the splitting mechanism, and in particular, it is not clear how to adapt the fundamental results from [5] to this setting.

Therefore, despite much investigation, there is still a crucial question about the Bregman dynamics that is left open:

> is it possible to find an explicit splitting integrator for the general (non-separable) Bregman Hamiltonian?

Note that, from all the above discussion, it would seem that the answer to this question is negative. However, as we will see, a proper geometric approach will reveal otherwise. To understand this result and to provide the complete general picture, it is convenient to first conclude our brief account of the geometric approaches to the construction of dynamical systems that can be used in optimization: it turns out that we can replace variational dynamical systems that exploit heuristic time-dependencies to achieve dissipation with explicitly *dissipative* dynamical systems. Muehlebach and Jordan [8] and Diakonikolas and Jordan [9] considered a dynamical systems perspective on these systems, showing how relatively simple dissipations can achieve state-of-the-art convergence. França et al. [10] 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

$$\dot{X} = \nabla_P H(X, P) \tag{14}$$

$$\dot{P} = -\nabla_X H(X, P) - c\,P, \tag{15}$$

with $c \in \mathbb{R}$ a constant. Being a geometric dynamical system this approach also admits effective geometric integrators similar to [3]. 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 [10].

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 [11–13]. In particular, first Maddison et al. [14] and then França et al. [10] 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 Maddison et al. [14] introduced another way of incorporating dissipative terms into Hamilton's equations (12)–(13). 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 $(X^*, P^*) = \operatorname{argmin} H(X, P)$. 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 [15]. 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* [16, 17] 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. More importantly, the geometric character of these dynamics directly implies that they are invariant under *contact transformations*. This indeed will be the key property to prove that the Bregman dynamics can always be re-written in new coordinates where the Hamiltonian is *separable*, thus establishing a definitive positive answer to our guiding question (see Theorem 1). This is the main result of our work. We argue that this property is of fundamental interest for practitioners, as it enables to directly construct simple explicit structure-preserving discretizations in the original phase space of the system. Finally, equipped with all these results, we introduce the *Relativistic Bregman Dynamics*, and provide an

explicit contact integrator that accurately follows the continuous flow in the original phase space of the system, thus obtaining the associated *Relativistic Bregman algorithm*. We also include numerical experiments showing that this simple construction is comparable to state-of-the-art approaches to discretize the (non-separable) Relativistic Bregman dynamics in terms of stability and rates of convergence (see [18]).

The structure of this work is as follows: in Sect. 2 we introduce contact Hamiltonian systems and show that all systems corresponding to Eqs. (10), (14)–(15), and (16)–(17) can be easily recovered as particular examples. Based on this result, we stress the importance of (time-dependent) contact transformations and then we prove that they can be used to make the Bregman Hamiltonian separable. To provide an explicit example of this construction, we then introduce the Relativistic Bregman Dynamics. Indeed, after applying Theorem 1, we obtain an equivalent but separable Hamiltonian system that can be integrated by splitting, thus obtaining what we call the related Relativistic Bregman algorithm (RB). Then in Sect. 3 we illustrate numerically that the RB can perform as well as other state-of-the-art algorithms in standard optimization tasks (see also [18] for further numerical tests and for all the codes). Finally, in Sect. 4 we summarize the results and discuss future directions.

2 Contact geometry in optimization

2.1 Definitions and examples

Contact geometry is a rich subject at the intersection among 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 [17, 19–25].

Definition 1 A *contact manifold* is a pair $(C, \mathcal{D})$, where C is a $(2n + 1)$-dimensional manifold and $\mathcal{D}$ is a maximally non-integrable distribution of hyperplanes on C, that is, a smooth assignment at each point p of C of a hyperplane in the tangent space $T_p C$.

One can prove that $\mathcal{D}$ can always be given locally as the kernel of a 1-form η 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, η), where C is a $(2n + 1)$-dimensional manifold and η 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

$$F^* \eta_2 = \alpha_F \eta_1, \tag{18}$$

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

$$\eta_{\text{std1}} := dS - P dX. \tag{19}$$

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

$$\eta_{\text{std2}} := d\tilde{S} - \frac{1}{2} \tilde{P} d\tilde{X} + \frac{1}{2} \tilde{X} d\tilde{P}. \tag{20}$$

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 η_{std1} and η_{std2} above. The map

$$F : (X, P, S) \mapsto \left(\tilde{X} = X + P, \tilde{P} = \frac{P - X}{2}, \tilde{S} = S - \frac{XP}{2} \right) \tag{21}$$

satisfies $F^* \eta_{\text{std2}} = \eta_{\text{std1}}$. 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, Ω) where M is a $2n$-dimensional manifold and Ω a 2-form on M that is closed, $d\Omega = 0$, and non-degenerate, $\Omega^n \neq 0$.

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

$$\Omega(X_H, \cdot) = -dH \qquad (22)$$

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 [26]. 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* (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 on $\theta = P dX - H(X, P, t)dt$. This is called the *Poincaré–Cartan* 1-form. Finally, define the dynamics $\tilde{X}_H$ by the two conditions

$$d\theta(\tilde{X}_H, \cdot) = 0, \qquad dt(\tilde{X}_H) = 1. \qquad (23)$$

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

Remark 3 Geometrically, the dynamics in (23) is precisely the *cosymplectic dynamics* used in [7]. From the point of view of contact geometry, it can be viewed as a very special type of *non-autonomous contact Hamiltonian system* (cf. item i) in Proposition 1 below). This is the point of view of the present work. However, this point of view can be further geometrized by considering $\tilde{X}_H$ as a *cocontact Hamiltonian system*, which is the analogue in contact geometry of the cosymplectic construction (see [27–29]). Similarly, the *pre-symplectic dynamics* used in [5] is exactly the dynamics of $\tilde{X}_H$. 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). A different although related approach to optimization based on the symplectification of contact systems is considered in [30].

We can now define dynamical systems that generalize the Hamiltonian systems arising in symplectic geometries.

Definition 6 (*Contact Hamiltonian systems*) Given a possibly time-dependent differentiable function $\mathcal{H}$ on a contact manifold (C, η), we define the *contact Hamiltonian vector field associated to $\mathcal{H}$* as the vector field $X_{\mathcal{H}}$ satisfying

$$\pounds_{X_{\mathcal{H}}} \eta = -R(\mathcal{H})\,\eta \qquad \eta(X_{\mathcal{H}}) = -\mathcal{H}, \tag{24}$$

where $\pounds_{X_{\mathcal{H}}} \eta$ denotes the Lie derivative of η with respect to $X_{\mathcal{H}}$ and R is the *Reeb vector field*, that is, the vector field satisfying $d\eta(R, \cdot) = 0$ and $\eta(R) = 1$. We denote the flow of $X_{\mathcal{H}}$ the *contact Hamiltonian system associated to $\mathcal{H}$*.

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

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.

Lemma 1 (Contact Hamiltonian systems: std1) *Given a (possibly time-dependent) differentiable function $\mathcal{H}(X, P, S, t)$ on the contact state space $(\mathbb{R}^{2n+1}, \eta_{\mathrm{std1}})$, 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 $(\mathbb{R}^{2n+1}, \eta_{\mathrm{std2}})$, 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 η_{std1} and η_{std2} respectively, using Cartan's identity for the Lie derivative of a 1-form, and from the fact that $R = \partial/\partial S$ in both cases, as can be seen by writing its definition in the corresponding coordinates.

Remark 5 (Variational formulation) Contact systems can alternatively be introduced starting from Herglotz' variational principle [31–34], with the Lagrangian function $\mathcal{L}(X, V, S, t)$ and its corresponding generalized Euler–Lagrange equations

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

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).

2.2 Main results

We arrive at our first main result with a direct calculation using Eqs. (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 $\mathcal{H} = H(X, P, t)$, that is, if $\mathcal{H}$ does not depend explicitly on S, then from Eqs. (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 $\mathcal{H} = H(X, P) + c\, S$, then from Eqs. (25)–(26) we obtain the standard equations for conformally symplectic systems (14)–(15), with (27) once again completely decoupled from the system.*
- *(iii) If $\mathcal{H} = H(X, P) + \langle X^*, P \rangle - \langle P^*, X \rangle + 2\, S$, then from Eqs. (28)–(29) we obtain the Hamiltonian descent equations (16)–(17), with (30) decoupled from the system.*
- *(iv) If $\mathcal{H} = \frac{1}{2} \|P\|^2 + f(X) + \frac{3}{t} S$, then from Eqs. (25)–(26) we obtain the continuous limit of NAG (3), with (27) decoupled from the system.*

Three immediate but powerful consequences of Proposition 1 are given in the following corollaries.

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

In particular, by choosing the contact Hamiltonian $\mathcal{H} = H_{\mathrm{Br}}(X, P, t)$, and for the appropriate choice of the free functions α, β and γ we can obtain contact systems with polynomial convergence rates of any order and even exponential convergence rates (see Eq. 9). So we see that in 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. As a second consequence of Proposition 1, we have

Corollary 2 *All the dynamics in Proposition 1 have a variational formulation.*

This important observation outlines the fact that all such dynamics may be studied by variational methods, either in the continuous case (see e.g. the interesting analysis in [35] and compare with [36]) or in the discrete one (cf. [7] and [32, 34]). All these aspects can now be analyzed through the lens of the corresponding contact tools, cf. Remark 5. In this work we focus on the following (third) key aspect of the geometric approach:

Corollary 3 *All the systems in Proposition 1 share invariance under a very large group of transformations, the group of (time-dependent) contact transformations.*

Corollary 3 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, are much less sensitive to the conditioning of the problem rather than other algorithms that do not possess a geometric structure (we refer to Maddison et al. [14] for an illuminating discussion on this, and we remark that contact transformations are more general than the symplectic transformations considered therein in Sect. 2.1). On the other hand, one can exploit contact transformations to re-write the dynamics in particular sets of coordinates where the system takes a simpler form, with great benefit for the ensuing discretization.

As a first "warm-up" example of the utility of contact transformations in optimization, we now prove that NAG is a composition of a contact transformation followed by a simple gradient descent step. This result is inspired by the conjecture put forward in [3], 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 instead we provide an actual proof that NAG is based on the composition of a *contact* map and a gradient step.

Proposition 2 (NAG is contact + gradient) *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

$$
dS_{k+1} - \frac{1}{2} P_{k+1} dX_{k+1} + \frac{1}{2} X_{k+1} dP_{k+1}
$$
$$
= f(X_k, P_k, S_k) \left(dS_k - \frac{1}{2} P_k dX_k + \frac{1}{2} X_k dP_k \right), \tag{33}
$$

for some function $f(X_k, P_k, S_k)$ that is nowhere 0. Then one can directly verify 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

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

followed by a standard gradient descent map,

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

$$P_{k+1} = P_k$$
(39)

$$S_{k+1} = S_k.$$
(40)

$\square$

Now we come to the main point of our work: in order to further illustrate the utility of contact transformations in a case of great current interest in the literature (see e.g. [37–41]), in the remainder of this section we focus on the Bregman dynamics and show how to use time-dependent contact transformation in order to re-write the Bregman dynamics in its most general form in such a way that it is clear that it is *always* derived from a separable Hamiltonian, and is thus amenable of simple, geometric and explicit discretizations by direct splitting.

As a first step, we need to introduce time-dependent contact transformations, which we do now following closely [17], and proceeding analogously to the case of time-dependent canonical (symplectic) transformations usually encountered in classical mechanics [19]. First, we extend the contact manifold C to $C \times \mathbb{R}$ by including the time variable as a coordinate on $\mathbb{R}$; then we also extend the contact form η to

$$\eta^E = \eta + \mathcal{H}(X, P, S, t)dt,$$
(41)

where $\mathcal{H}(X, P, S, t)$ is the contact Hamiltonian of the system. Recall as usual that locally we can think $C \simeq \mathbb{R}^{2n+1}$ and $\eta^E = dS - PdX + \mathcal{H}(X, P, S, t)dt$. Then, we define a time-dependent contact transformation as follows.

Definition 7 (*Time-dependent contact transformation*) A *time-dependent contact transformation* for a system with contact Hamiltonian $\mathcal{H}$ is a map $F : C \times \mathbb{R} \to C \times \mathbb{R}$ such that

$$F^*\eta^E = \sigma\, \eta^E \quad \text{and} \quad F^*t = t,$$
(42)

where $\sigma : C \times \mathbb{R} \to \mathbb{R}$ is a nowhere-vanishing function.

Remark 6 In local coordinates (X, P, S, t) on $C \times \mathbb{R}$, this is equivalent to saying that a contact transformation maps to new coordinates

$$\tilde{X} = \tilde{X}(X, P, S, t), \quad \tilde{P} = \tilde{P}(X, P, S, t), \quad \tilde{S} = \tilde{S}(X, P, S, t), \quad \tilde{t} = t,$$

such that

$$d\tilde{S} - \tilde{P}d\tilde{X} + K(\tilde{X}, \tilde{P}, \tilde{S}, t)dt = \sigma(X, P, S, t)(dS - PdX + \mathcal{H}(X, P, S, t)dt),$$
(43)

where $K(\tilde{X}, \tilde{P}, \tilde{S}, t)$ is the new contact Hamiltonian in the new coordinates.

It follows directly from (43), that a necessary condition for a time-dependent transformation to be contact is

$$\frac{\partial \tilde{S}}{\partial t} - \tilde{P}\frac{\partial \tilde{X}}{\partial t} + K = \sigma \, \mathcal{H},$$
(44)

that is, the new contact Hamiltonian is found from the original one according to

$$K = \sigma \, \mathcal{H} - \frac{\partial \tilde{S}}{\partial t} + \tilde{P}\frac{\partial \tilde{X}}{\partial t}.$$
(45)

Finally, it can be proved that the time-dependent contact Hamiltonian equations are preserved under the above transformations if and only if

$$-\sigma\frac{\partial K}{\partial \tilde{S}} = -\sigma\frac{\partial \mathcal{H}}{\partial S} - d\sigma(X_{\mathcal{H}}^E),$$
(46)

where $X_{\mathcal{H}}^E = X_{\mathcal{H}} + \partial_t$ (see [17] for more details).

Now we can use the above results to simplify the Bregman Hamiltonian. In particular, we construct a time-dependent contact transformation that brings H_{Br} to a Hamiltonian K_{Br} which is equivalent to the original one but separable, and hence directly amenable to an explicit discretization by splitting. Let us consider the transformation

$$\tilde{X} = X, \quad \tilde{P} = e^{-\gamma}p + \nabla h(X), \quad \tilde{S} = e^{-\gamma}S + h(X), \quad \tilde{t} = t,$$
(47)

with $h(X)$ the same convex function as in (11). We have the following

Lemma 3 *For any contact Hamiltonian system with Hamiltonian $\mathcal{H}$, the transformation (47) is not canonical (symplectic) but it is contact, with*

$$\sigma(X, P, S, t) = e^{-\gamma}, \quad K = e^{-\gamma}(\mathcal{H} + \dot{\gamma}S).$$
(48)

Proof The proof proceeds by directly calculating the differential of the transformation (47). In this way we obtain that

$$d\tilde{P} \wedge d\tilde{X} = e^{-\gamma}dP \wedge dX,$$

showing that it is not a canonical (symplectic) transformation. Moreover, one can also verify that

$$d\tilde{S} - \tilde{P}d\tilde{X} + K(\tilde{X}, \tilde{P}, \tilde{S}, t)dt = e^{-\gamma}\left(dS - PdX + \mathcal{H}(X, P, S, t)dt\right),$$

and that (46) holds for K as in (48), showing that indeed it is a time-dependent contact transformation that preserves the dynamics.   □

Now we are finally ready to prove the second main result of our work.

Theorem 1 (The Bregman Hamiltonian is always separable) *Under the ideal scaling condition $\dot{\gamma} = e^{\alpha}$, the generally non-separable Bregman Hamiltonian*

$$H_{Br}(X, P, t) = e^{\alpha+\gamma}\left(D_{h^*}(e^{-\gamma}P + \nabla h(X), \nabla h(X)) + e^{\beta}f(X)\right),\qquad(49)$$

is equivalent to the separable contact Hamiltonian

$$K_{Br}(\tilde{X}, \tilde{P}, \tilde{S}, t) = e^{\alpha}\left(h^*(\tilde{P}) - \langle\tilde{P}, \tilde{X}\rangle + e^{\beta}f(\tilde{X}) + \tilde{S}\right),\qquad(50)$$

for any choice of the convex function $h(X)$.

Proof We start by applying the above contact transformation (47) to the Bregman Hamiltonian (49) and use that $\dot{\gamma} = e^{\alpha}$ (ideal scaling condition from [2]) to obtain the new Bregman Hamiltonian in the new coordinates, which, using (48), reads

$$K_{Br}(\tilde{X}, \tilde{P}, \tilde{S}, t) = e^{\alpha}\left(D_{h^*}(\tilde{P}, \nabla h(\tilde{X})) + e^{\beta}f(\tilde{X}) + \tilde{S} - h(\tilde{X})\right).\qquad(51)$$

Interestingly, by the very definition of the Bregman divergence, Eq. (4), we can rewrite the above expression as

$$K_{Br}(\tilde{X}, \tilde{P}, \tilde{S}, t) = e^{\alpha}\left(h^*(\tilde{P}) - h^*(\nabla h(\tilde{X})) - \langle\tilde{P}, \tilde{X}\rangle + \langle\nabla h(\tilde{X}), \tilde{X}\rangle + e^{\beta}f(\tilde{X}) + \tilde{S} - h(\tilde{X})\right).$$
$$(52)$$

Moreover, by definition of the Legendre transform,

$$h^*(\nabla h(\tilde{X})) = \langle\nabla h(\tilde{X}), \tilde{X}\rangle - h(\tilde{X}),$$

and thus some of the terms in (52) cancel, to give the sought-for final result (50).   □

We remark at this point that non-separable Hamiltonians systems are notoriously hard to discretize, since a direct application of standard symplectic (or contact) integrators leads to implicit algorithms. To bypass this problem for the Bregman dynamics, in [5] the authors suggested to use a technique first proposed in [42] that consists in doubling the phase space and then propose an "augmented Hamiltonian" on such space that is separable and thus can be integrated using standard explicit symplectic

integrators. However, it must be remarked that the geometric (pre-symplectic) character of such algorithms is only guaranteed in the doubled phase space, not in the original one, and this signals that such procedure has to be handled with care. Instead, using Theorem 1, we will be able to perform a direct splitting that on the one side simplifies the algorithm and requires less computational burden, and on the other side guarantees structure-preservation in the original phase space of the system.

2.3 Relativistic Bregman algorithm

To illustrate the benefits of our approach and motivated by the arguments in [10] on the advantage of choosing a "relativistic" kinetic term in the Hamiltonian approach to optimization, we first define the *Relativistic Bregman dynamics* as a motivating example of a non-separable Bregman Hamiltonian that can be turned into an equivalent (but separable) one by means of Theorem 1. Afterwards, by a direct splitting we obtain the corresponding *Relativistic Bregman algorithm* (RB).

Let us consider the case in which H_{Br} in (11) is generated by the convex function

$$h(X) = -mv^2 \sqrt{1 - \frac{\|X\|^2}{v^2}}, \tag{53}$$

where this function is inspired in the relativistic Lagrangian for a particle of mass m and v is the analogue of the speed of light. Moreover, let us fix the functions

$$\alpha = \log c - \log t, \qquad \beta = c \log t + \log C, \qquad \gamma = c \log t, \qquad c, C > 0,$$

which satisfy the ideal scaling conditions (6) and (7). With this choice, the Bregman dynamics is guaranteed to have a polynomial convergence rate of order c [2].

From (53) and (11) we obtain the *Relativistic Bregman Hamiltonian*

$$
H_{\mathrm{Br}}^R = e^{\alpha+\gamma} \left[v \sqrt{\left\| e^{-\gamma} P + \frac{mX}{\sqrt{1 - \frac{\|X\|^2}{v^2}}} \right\|^2 + m^2 v^2} \right.
$$

$$
\left. - \frac{mv^3}{\sqrt{v^2 - \|X\|^2}} - e^{-\gamma}\langle P, X \rangle + e^{\beta} f(X) \right], \tag{54}
$$

which is clearly non-separable. However, using (50) we get the much simpler but equivalent

$$K_{\mathrm{Br}}^R = e^{\alpha} \left(v\sqrt{\|P\|^2 + m^2 v^2} - \langle P, X \rangle + e^{\beta} f(X) + S \right), \tag{55}$$

where, from now on, we drop the notation with a tilde above the new coordinates in (55), as there will be no problem of confusion. We emphasize that the Relativistic Bregman Hamiltonian (54) is different from the Relativistic Gradient Descent proposed in [10]. Indeed, although both are defined using a relativistic kinetic term, the former belongs to the Bregman family of Hamiltonian functions, while the latter does not.

In [5] in order to integrate Hamilton's equation stemming from a non-separable Hamiltonian like (54), the authors suggest to use the technique first proposed in [42], according to which one first doubles the phase space dimensions to a space with coordinates $(X, P, t, \bar{X}, \bar{P}, \bar{t})$, and then defines the *augmented Hamiltonian*

$$
H(X, P, t, \bar{X}, \bar{P}, \bar{t}) = H_{\mathrm{Br}}^R(X, \bar{P}, t) + H_{\mathrm{Br}}^R(\bar{X}, P, \bar{t})
$$
$$
+ \frac{\xi}{2} \left(\|X - \bar{X}\|^2 + \|P - \bar{P}\|^2 \right), \tag{56}
$$

where $\xi > 0$ is a free parameter whose value controls the strength of the last term, that is included in order to bias the dynamics towards $X = \bar{X}$ and $P = \bar{P}$, and whose value has to be tuned in practice. The equations of motion for the augmented Hamiltonian are equivalent to those of the original system when $X = \bar{X}$, $P = \bar{P}$, and $t = \bar{t}$ and the Hamiltonian is now separable and thus one can integrate the dynamics using a splitting scheme.

On the other hand, the Hamiltonian K_{Br}^R in (55) can be splitted directly in the original contact phase space, with coordinates (X, P, S) (and time). For instance, using the splitting

$$
K_A := e^\alpha v \sqrt{\|P\|^2 + v^2 m^2}, \quad K_B = -e^\alpha \langle X, P \rangle, \quad K_C = e^{\alpha+\beta} f(X), \quad K_D = e^\alpha S \tag{57}
$$

we get the maps

$$
\varphi_\tau^A \begin{pmatrix} X \\ P \\ S \end{pmatrix} = \begin{pmatrix} X + \frac{e^\alpha v P}{\sqrt{\|P\|^2 + v^2 m^2}} \tau \\ P \\ S + e^\alpha v \left(\frac{-v^2 m^2}{\sqrt{\|P\|^2 + v^2 m^2}} \right) \tau \end{pmatrix}
$$

$$
\varphi_\tau^B \begin{pmatrix} X \\ P \\ S \end{pmatrix} = \begin{pmatrix} X \exp\left(-e^\alpha \tau\right) \\ P \exp\left(e^\alpha \tau\right) \\ S \end{pmatrix}
$$

$$
\varphi_\tau^C \begin{pmatrix} X \\ P \\ S \end{pmatrix} = \begin{pmatrix} X \\ -e^{\alpha+\beta} \nabla f(X)\tau + P \\ -e^{\alpha+\beta} f(X)\tau + S \end{pmatrix}
$$

$$
\varphi_\tau^D \begin{pmatrix} X \\ P \\ S \end{pmatrix} = \begin{pmatrix} X \\ P \exp\left(-e^\alpha \tau\right) \\ S \exp\left(-e^\alpha \tau\right) \end{pmatrix}
$$

and, using the composition

$$\varphi^D_{\tau/2} \circ \varphi^B_{\tau/2} \circ \varphi^C_{\tau/2} \circ \varphi^A_{\tau} \circ \varphi^C_{\tau/2} \circ \varphi^B_{\tau/2} \circ \varphi^D_{\tau/2},$$

we obtain a second-order contact integrator (we refer to e.g. [43] for the study of contact integrators derived by splitting. See also [30] for a different but related approach to optimization using these types of algorithms). Clearly this integrator yields an optimization algorithm that we shall call the *Relativistic Bregman algorithm* (RB).

We remark on the most important difference between the strategy put forward in [5] as compared to ours: while the integrators proposed in [5] for the "non-separable" case are geometric (pre-symplectic) only in the enlarged phase space, thus not guaranteeing that such property holds in the original phase space of the system, the RB that we have just described, as well as any other algorithm based on the use of contact splitting integrators after Theorem 1, are structure-preserving in the original (contact) phase space. Therefore we expect the RB to be both more stable and efficient, with direct benefits for the optimization task, especially as the dimension of the problem increases. The numerical simulations reported in [18] provide evidence for these conclusions.

3 Numerical experiments

The purpose of this section is just illustrative: we aim to show that the Relativistic Bregman algorithm (RB) proposed above can effectively be used in benchmark optimization and machine learning tasks. To show this point, we report here the performance of RB on some reference examples and present a comparison with its Euclidean counterpart, namely the *Euclidean Bregman algorithm* (EB), and two standard optimization algorithms such as NAG and CM. We refer to [18] for the details of the EB and for a much larger comparison with further test functions and algorithms.

In all the following examples we set $P_0 = 0$ and $S_0 = 0$ whenever such initial conditions are required.

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

$$f(X) = \frac{1}{2} X^T A X, \quad X \in \mathbb{R}^{500}, \quad \lambda(A) \sim \mathcal{U}(10^{-3}, 1), \tag{58}$$

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 Fig. 1a, we compare the performance of EB and RB on this problem using initial condition $X_0 = (1, 1, ..., 1)$, step size $\tau = 10^{-4}$, speed of light $v = 1000$, mass $m = 10^{-3}$, and varying $c \in \{2, 4, 8\}$. For these parameters, both RB and EB exhibit similar rates of convergence, which are close to the theoretical ones, see Table 1.

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

$$f(X) = [(X - 1)^T \Sigma (X - 1)]^2, \quad X \in \mathbb{R}^{50}, \quad \Sigma_{ij} = 0.9^{|i-j|}. \tag{59}$$
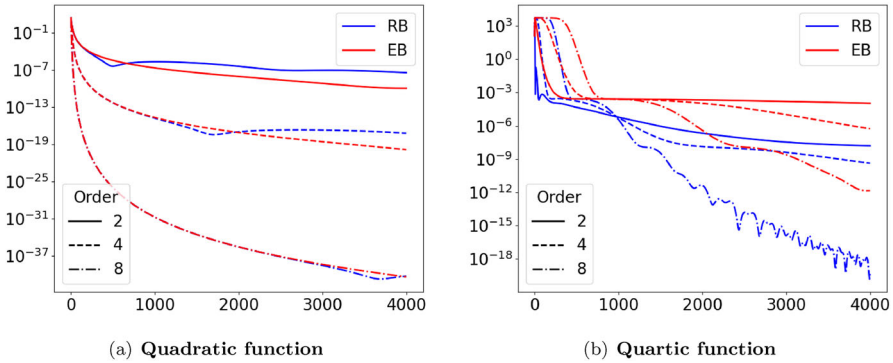
(a) **Quadratic function**   (b) **Quartic function**

Fig. 1 Comparison of the performance of the RB and EB methods, both with the same parameters, and for varying $c \in \{2, 4, 8\}$. Left panel: Quadratic function. Right panel: Quartic function

Table 1 Numerical convergence rates of the RB and EB algorithms for the Quadratic function (58) achieved at 1000, 2000, and 4000 iterations respectively

Order	$c = 2$			$c = 4$			$c = 8$		
Iterations	10^3	2×10^3	4×10^3	10^3	2×10^3	4×10^3	10^3	2×10^3	4×10^3
RB	$t^{-1.75}$	$t^{-1.91}$	$t^{-2.03}$	$t^{-4.99}$	$t^{-4.80}$	$t^{-4.84}$	$t^{-10.25}$	$t^{-10.77}$	$t^{-10.84}$
EB	$t^{-2.26}$	$t^{-2.59}$	$t^{-2.37}$	$t^{-4.89}$	$t^{-5.19}$	$t^{-5.67}$	$t^{-10.25}$	$t^{-10.76}$	$t^{-11.21}$

Table 2 Numerical convergence rates of the RB and EB algorithms for the Quartic function (59), achieved at 1000, 2000, and 4000 iterations respectively

Order	$c = 2$			$c = 4$			$c = 8$		
Iterations	10^3	2×10^3	4×10^3	10^3	2×10^3	4×10^3	10^3	2×10^3	4×10^3
RB	$t^{-2.07}$	$t^{-2.29}$	$t^{-2.28}$	$t^{-2.17}$	$t^{-2.47}$	$t^{-3.01}$	$t^{-2.34}$	$t^{-3.86}$	$t^{-5.64}$
EB	$t^{-1.20}$	$t^{-1.16}$	$t^{-1.24}$	$t^{-1.21}$	$t^{-1.35}$	$t^{-2.03}$	$t^{-1.22}$	$t^{-2.30}$	$t^{-3.69}$

This convex function achieves its minimum value 0 at $X^* = 1$. In Fig. 1b, we compare the performance of EB and RB on this problem using initial condition $X_0 \sim \mathcal{U}(0, 1)$, step size $\tau = 10^{-3}$, and the rest of the parameters as in the previous example. In this case, RB shows better convergence rates than those of EB, being also closer to the theoretical ones, see Table 2.

Example 5 (Machine learning examples) Now we test the RB in two classical machine learning tasks, and compare its performance with those of EB, CM and NAG. We use two popular datasets for classification from the UCI machine learning repository and try to fit a Two-regularized logistic regression model. The profiles of these datasets are summarized in Table 3. We set the regularization parameter for all methods to be $\lambda_{reg} = 10^{-2}$; for the RB and EB algorithms, we set $v = 1000$, $m = 10^{-3}$, $C = 1$, and $c = 2$. The rest of the parameters were tuned in the validation dataset.

Table 3 Profiles of the datasets

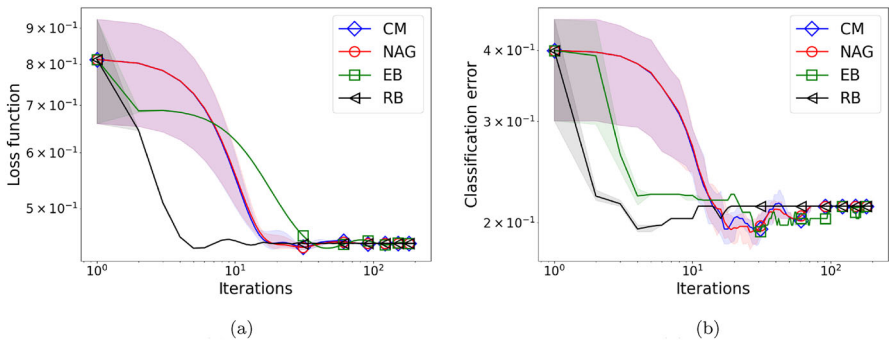| Dataset	Diabetes	Breast cancer
Train	460	340
Test	192	143
Validation	115	86
Feat	8	6
Class	2	2

Fig. 2 Two-regularized logistic regression model trained with RB, EB, CM, and NAG on the Pima Indians Diabetes dataset. Left panel: Objective function values. Right panel: Classification error. Both panels compare with respect to the iteration of optimization on log–log scale. The initial conditions are taken uniformly on $(0, 1)$

Diabetes

In the first example, we use the Pima Indians Diabetes dataset. The objective is to predict, based on diagnostic measurements, whether a patient has diabetes. The dataset was separated into training 460, test 192, and validation 115 sets. Figure 2 shows the mean over 50 random initial conditions, the 0.025 and 0.975 quantiles of the loss function, and the classification error evaluated in the test set. The RB algorithm converges faster to the optimum, which is reached by the other algorithms a few iterations later. The RB and EB algorithms showed less sensitivity to the initial condition compared to CM and NAG.

Breast cancer

The objective is to classify breast cancer from some features. The dataset was separated into training 460, test 192 and validation 115 sets. The RB algorithm exhibits similar behavior as in the previous example, converging faster than the other algorithms in the first few iterations, see Fig. 3.
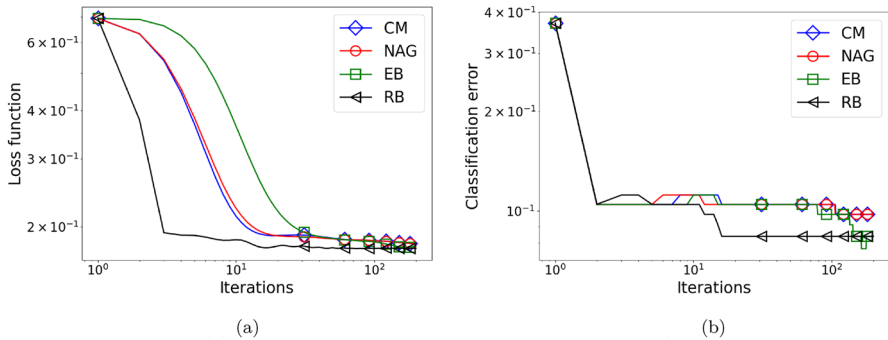
Fig. 3 Two-regularized logistic regression model trained with RB, EB, CM, and NAG on the Breast Cancer dataset. Left panel: Objective function values. Right panel: Classification error. Both panels compare with respect to the iteration of optimization on a log–log scale. All the methods are initiated at the initial condition $X_0 = 0$

4 Conclusions

Geometry is a powerful tool in pure and applied mathematics. Among other things, it enables to describe phenomena independently of the particular choice of coordinates. This is because geometric objects are invariant under some group of transformations. In this work, we have shown that contact geometry and the related contact transformations can be extremely useful in optimization. Indeed, we have shown that invariance under contact transformations is a common feature of all the recently-proposed dynamics in the context of the dynamical systems approach to convex optimization (Proposition 1). More importantly, we have proved that the Bregman dynamics, which is perhaps the single most important recent finding in this context but whose implementation is seriously hindered by the fact that the Hamiltonian is non-separable, is actually always separable up to a contact transformation (Theorem 1). This opens the way to applying explicit and fast numerical integration methods for simulating the Bregman dynamics, which in turn provides new efficient optimization algorithms, even when the original Hamiltonian looks non-separable. Finally, in order to illustrate the relevance of considering algorithms that stem from non-separable Bregman Hamiltonians, we have shown in some benchmark examples from the optimization and machine learning literature that the thus-proposed Relativistic Bregman algorithm compares favorably with respect to the standard Classical Momentum and Nesterov's Accelerated Gradient algorithms, and also to the Euclidean Bregman algorithm.

In future work, we plan to perform a more in-depth analysis of the properties of the Relativistic Bregman algorithm. Moreover, as commented in the text, it will be interesting to study several optimization dynamics through the lens of Herglotz's variational principle. Finally, all the results presented here have a natural generalization to the case of general differentiable manifolds, and therefore we can extend our analysis with methods similar to those employed in [39–41].

Data availibility The datasets analysed and all the codes used for the study are available at https://github.com/mdazatorres/Breg_dynamic_contact_algorithm.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Su, W., Boyd, S., Candès, E.J.: A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. J. Mach. Learn. Res. **17**(153), 1–43 (2016)
2. Wibisono, A., Wilson, A.C., Jordan, M.I.: A variational perspective on accelerated methods in optimization. Proc. Natl. Acad. Sci. **113**(47), 7351–7358 (2016). https://doi.org/10.1073/pnas.1614734113
3. Betancourt, M., Jordan, M.I., Wilson, A.C.: On symplectic optimization (2018). arXiv preprint arXiv:1802.03653
4. Hairer, E., Lubich, C., Wanner, G.: Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations. Springer Series in Computational Mathematics, vol. 31. Springer, Berlin (2006). https://doi.org/10.1007/3-540-30666-8
5. França, G., Jordan, M.I., Vidal, R.: On dissipative symplectic integration with applications to gradient-based optimization. J. Stat. Mech. Theory Exp. **2021**(4), 043402 (2021). https://doi.org/10.1088/1742-5468/abf5d4
6. Duruisseaux, V., Schmitt, J., Leok, M.: Adaptive Hamiltonian variational integrators and applications to symplectic accelerated optimization. SIAM J. Sci. Comput. **43**(4), 2949–2980 (2021). https://doi.org/10.1137/20M1383835
7. Campos, C.M., Mahillo, A., Diego, D.M.: A discrete variational derivation of accelerated methods in optimization (2021). arXiv preprint arXiv:2106.02700
8. Muehlebach, M., Jordan, M.: A dynamical systems perspective on Nesterov acceleration. In: International Conference on Machine Learning, pp. 4656–4662. PMLR (2019)
9. Diakonikolas, J., Jordan, M.I.: Generalized momentum-based methods: a Hamiltonian perspective. SIAM J. Optim. **31**(1), 915–944 (2021). https://doi.org/10.1137/20M1322716
10. França, G., Sulam, J., Robinson, D.P., Vidal, R.: Conformal symplectic and relativistic optimization. J. Stat. Mech. Theory Exp. **2020**(12), 124008 (2020). https://doi.org/10.1088/1742-5468/abcaee
11. Betancourt, M., Byrne, S., Livingstone, S., Girolami, M.: The geometric foundations of Hamiltonian Monte Carlo. Bernoulli **23**(4A), 2257–2298 (2017). https://doi.org/10.3150/16-BEJ810
12. Betancourt, M.: A conceptual introduction to Hamiltonian Monte Carlo (2017). arXiv preprint arXiv:1701.02434
13. Livingstone, S., Faulkner, M.F., Roberts, G.O.: Kinetic energy choice in Hamiltonian/hybrid Monte Carlo. Biometrika **106**(2), 303–319 (2019). https://doi.org/10.1093/biomet/asz013
14. Maddison, C.J., Paulin, D., Teh, Y.W., O'Donoghue, B., Doucet, A.: Hamiltonian descent methods (2018). arXiv preprint arXiv:1809.05042
15. O'Donoghue, B., Maddison, C.J.: Hamiltonian descent for composite objectives. Adv. Neural Inf. Process. Syst. **32** (2019)
16. Bravetti, A.: Contact Hamiltonian dynamics: the concept and its use. Entropy **19**(10) (2017). https://doi.org/10.3390/e19100535

17. Bravetti, A., Cruz, H., Tapias, D.: Contact Hamiltonian mechanics. Ann. Phys. **376**, 17–39 (2017). https://doi.org/10.1016/j.aop.2016.11.003

18. Daza-Torres, M.L., Bravetti, A., Flores-Arguedas, H., Betancourt, M.: Bregman dynamics, contact transformations and convex optimization. https://github.com/mdazatorres/Breg_dynamic_contact_algorithm. GitHub repository (2022)

19. Arnold, V.I.: Mathematical Methods of Classical Mechanics. Graduate Texts in Mathematics, vol. 60. Springer, Berlin (1989)

20. Geiges, H.: An Introduction to Contact Topology, vol. 109. Cambridge University Press, Cambridge (2008)

21. Bravetti, A.: Contact geometry and thermodynamics. Int. J. Geom. Methods Mod. Phys., 1940003 (2018). https://doi.org/10.1142/S0219887819400036

22. Ciaglia, F.M., Cruz, H., Marmo, G.: Contact manifolds and dissipation, classical and quantum. Ann. Phys. **398**, 159–179 (2018). https://doi.org/10.1016/j.aop.2018.09.012

23. León, M., Sardón, C.: Cosymplectic and contact structures for time-dependent and dissipative Hamiltonian systems. J. Phys. A Math. Theor. **50**(25), 255205 (2017). https://doi.org/10.1088/1751-8121/aa711d

24. 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. Int. J. Geom. Methods Mod. Phys. **17**(06), 2050090 (2020). https://doi.org/10.1142/S0219887820500905

25. León, M., Lainz Valcázar, M.: Contact Hamiltonian systems. J. Math. Phys. **60**(10), 102902 (2019). https://doi.org/10.1063/1.5096475

26. Abraham, R., Marsden, J.E.: Foundations of Mechanics. Benjamin/Cummings Publishing Company, Reading (1978)

27. León, M., Gaset, J., Gràcia, X., Muñoz-Lecanda, M.C., Rivas, X.: Time-dependent contact mechanics. Monatsh. Math., 1–35 (2022). https://doi.org/10.1007/s00605-022-01767-1

28. Gaset, J., López-Gordón, A., Rivas, X.: Symmetries, conservation and dissipation in time-dependent contact systems (2022). arXiv preprint arXiv:2212.14848

29. Azuaje, R., Escobar-Ruiz, A.: Canonical and canonoid transformations for Hamiltonian systems on (co) symplectic and (co) contact manifolds. J. Math. Phys. **64**(3), 033501 (2023). https://doi.org/10.1063/5.0135045

30. Goto, S.-I., Hino, H.: Fast symplectic integrator for Nesterov-type acceleration method (2021). arXiv preprint arXiv:2106.07620

31. Georgieva, B., Guenther, R., Bodurov, T.: Generalized variational principle of Herglotz for several independent variables. First Noether-type theorem. J. Math. Phys. **44**(9), 3911–3927 (2003). https://doi.org/10.1063/1.1597419

32. Vermeeren, M., Bravetti, A., Seri, M.: Contact variational integrators. J. Phys. A Math. Theor. **52**(44), 445206 (2019). https://doi.org/10.1088/1751-8121/ab4767

33. De Leon, M., Lainz Valcázar, M.: Singular Lagrangians and precontact Hamiltonian systems. Int. J. Geom. Methods Mod. Phys. **16**(10), 1950158 (2019). https://doi.org/10.1142/S0219887819501585

34. Anahory Simoes, A., Diego, D., Lainz Valcázar, M., León, M.: On the geometry of discrete contact mechanics. J. Nonlinear Sci. **31**(3), 1–30 (2021). https://doi.org/10.1007/s00332-021-09708-2

35. Zhang, P., Orvieto, A., Daneshmand, H.: Rethinking the variational interpretation of Nesterov's accelerated method (2021). arXiv preprint arXiv:2107.05040

36. Ryan, J.: When action is not least for systems with action-dependent Lagrangians. J. Math. Phys. **64**(3), 032901 (2023). https://doi.org/10.1063/5.0099612

37. Sun, C., Hu, G.: A continuous-time Nesterov accelerated gradient method for centralized and distributed online convex optimization (2020). arXiv preprint arXiv:2009.12545

38. Wilson, A.C., Recht, B., Jordan, M.I.: A Lyapunov analysis of accelerated methods in optimization. J. Mach. Learn. Res. **22**, 113–1 (2021)

39. França, G., Barp, A., Girolami, M., Jordan, M.I.: Optimization on manifolds: a symplectic approach (2021). arXiv preprint arXiv:2107.11231

40. Duruisseaux, V., Leok, M.: A variational formulation of accelerated optimization on Riemannian manifolds. SIAM J. Math. Data Sci. **4**(2), 649–674 (2022). https://doi.org/10.1137/21M1395648

41. Duruisseaux, V., Leok, M.: Time-adaptive Lagrangian variational integrators for accelerated optimization on manifolds (2022). arXiv preprint arXiv:2201.03774

42. Tao, M.: Explicit symplectic approximation of nonseparable Hamiltonians: algorithm and long time performance. Phys. Rev. E **94**(4), 043303 (2016). https://doi.org/10.1103/PhysRevE.94.043303
43. Bravetti, A., Seri, M., Vermeeren, M., Zadra, F.: Numerical integration in celestial mechanics: a case for contact geometry. Celest. Mech. Dyn. Astron. **132**(1), 1–29 (2020). https://doi.org/10.1007/s10569-019-9946-9

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.