A VARIATIONAL FORMULATION OF ACCELERATED OPTIMIZATION ON RIEMANNIAN MANIFOLDS

VALENTIN DURUISSEAUX AND MELVIN LEOK

ABSTRACT. It was shown recently by [23] that Nesterov's accelerated gradient method for minimizing a smooth convex function f can be thought of as the time discretization of a second-order ODE, and that $f(x(t))$ converges to its optimal value at a rate of $\mathcal{O}(1/t^2)$ along any trajectory $x(t)$ of this ODE. A variational formulation was introduced in [25] which allowed for accelerated convergence at a rate of $\mathcal{O}(1/t^p)$, for arbitrary $p > 0$, in normed vector spaces. This framework was exploited in [8] using time-adaptive geometric integrators to design efficient explicit algorithms for symplectic accelerated optimization. In [3], a second-order ODE was proposed as the continuous-time limit of a Riemannian accelerated algorithm, and it was shown that the objective function $f(x(t))$ converges to its optimal value at a rate of $\mathcal{O}(1/t^2)$ along solutions of this ODE, thereby generalizing the earlier Euclidean result to the Riemannian manifold setting. In this paper, we show that on Riemannian manifolds, the convergence rate of $f(x(t))$ to its optimal value can also be accelerated to an arbitrary convergence rate $\mathcal{O}(1/t^p)$, by considering a family of time-dependent Bregman Lagrangian and Hamiltonian systems on Riemannian manifolds. This generalizes the results of [25] to Riemannian manifolds and also provides a variational framework for accelerated optimization on Riemannian manifolds. In particular, we will establish results for objective functions on Riemannian manifolds that are geodesically convex, weakly-quasi-convex, and strongly convex. An approach based on the time-invariance property of the family of Bregman Lagrangians and Hamiltonians was used to construct very efficient optimization algorithms in [8], and we establish a similar time-invariance property in the Riemannian setting. This lays the foundation for constructing similarly efficient optimization algorithms on Riemannian manifolds, once the Riemannian analogue of time-adaptive Hamiltonian variational integrators has been developed. The experience with the numerical discretization of variational accelerated optimization flows on vector spaces suggests that the combination of time-adaptivity and symplecticity is important for the efficient, robust, and stable discretization of these variational flows describing accelerated optimization. One expects that a geometric numerical integrator that is time-adaptive, symplectic, and Riemannian manifold preserving will yield a class of similarly promising optimization algorithms on manifolds.

1. INTRODUCTION

Efficient optimization has become one of the major concerns in data analysis. Many machine learning algorithms are designed around the minimization of a loss function or the maximization of a likelihood function. Due to the ever-growing scale of the data sets and size of the problems, there has been a lot of focus on first-order optimization algorithms because of their low cost per iteration. The first gradient descent algorithm was proposed in [5] by Cauchy to deal with the very large systems of equations he was facing when trying to simulate orbits of celestial bodies, and many gradient-based optimization methods have been proposed since Cauchy's work in 1847.

In 1983, Nesterov's accelerated gradient method was introduced in [19], and was shown to converge in $\mathcal{O}(1/k^2)$ to the minimum of the convex objective function f, improving on the $\mathcal{O}(1/k)$ convergence rate exhibited by the standard gradient descent methods. This $\mathcal{O}(1/k^2)$ convergence rate was shown in [20] to be optimal among first-order methods using only information about ∇f at consecutive iterates. This phenomenon in which an algorithm displays this improved rate of convergence is referred to as acceleration, and other accelerated algorithms have been derived since Nesterov's algorithm, such as accelerated mirror descent [18] and accelerated cubic-regularized Newton's method [21]. More recently, it was shown in [23] that Nesterov's accelerated gradient method limits to a second-order ODE, as the timestep goes to 0, and that the objective function

$f(x(t))$ converges to its optimal value at a rate of $\mathcal{O}(1/t^2)$ along the trajectories of this ODE. It was then shown in [25] that in continuous time, the convergence rate of $f(x(t))$ can be accelerated to an arbitrary convergence rate $\mathcal{O}(1/t^p)$ in normed spaces, by considering flow maps generated by a family of time-dependent Bregman Lagrangian and Hamiltonian systems which is closed under time rescaling. This variational framework and the time-invariance property of the family of Bregman Lagrangians was then exploited in [8] using time-adaptive geometric integrators to design efficient explicit algorithms for symplectic accelerated optimization. It was observed that a careful use of adaptivity and symplecticity could result in a significant gain in computational efficiency.

In the past few years, there has been some effort to derive accelerated optimization algorithms in the Riemannian manifold setting [2–4; 15; 26; 27]. In [3], a second-order ODE was proposed as the continuous-time limit of a Riemannian accelerated algorithm, and it was shown that the objective function $f(x(t))$ converges to its optimal value at a rate of $\mathcal{O}(1/t^2)$ along solutions of this ODE, generalizing the Euclidean result obtained in [23] to the Riemannian manifold setting.

In this paper, we show that in continuous time, the convergence rate of $f(x(t))$ to its optimal value can be accelerated to an arbitrary convergence rate $\mathcal{O}(1/t^p)$ on Riemannian manifolds, thereby generalizing the results of [25] to the Riemannian setting. This is achieved by considering a family of time-dependent Bregman Lagrangian and Hamiltonian systems on Riemannian manifolds. This also provides a variational framework for accelerated optimization on Riemannian manifolds, generalizing the normed vector space variational formulation of accelerated optimization introduced in [25]. We will then illustrate the derived theoretical convergence rates by integrating the Bregman Euler–Lagrange equations using a simple numerical scheme to solve eigenvalue and distance minimization problems on Riemannian manifolds. Finally, we will show that the family of Bregman dynamics on Riemannian manifolds is closed under time rescaling, and we will draw inspiration from the approach introduced in [8] to take advantage of this invariance property via a carefully chosen Poincaré transformation that will allow for the integration of higher-order Bregman dynamics while benefiting from the computational efficiency of integrating lower-order Bregman dynamics on Riemannian manifolds.

2. Definitions and Preliminaries

We first introduce the main notions from Riemannian geometry and Lagrangian and Hamiltonian mechanics that will be used throughout this paper (see [3; 9; 10; 12; 13; 16] for more details).

2.1. Riemannian Geometry.

Definition 2.1. *Given a manifold $\mathcal{Q}$, the **tangent bundle** $T\mathcal{Q}$ and **cotangent bundle** $T^*\mathcal{Q}$ are defined by*

$$T\mathcal{Q} = \{(q,v) | q \in \mathcal{Q}, v \in T_q\mathcal{Q}\} \qquad \text{and} \qquad T^*\mathcal{Q} = \{(q,p) | q \in \mathcal{Q}, p \in T_q^*\mathcal{Q}\}.$$

Definition 2.2. *Suppose we have a Riemannian manifold $\mathcal{Q}$ with Riemannian metric $g(\cdot,\cdot) = \langle \cdot,\cdot \rangle$, represented by the positive-definite symmetric matrix (g_{ij}) in local coordinates. Then, we define the **musical isomorphism** $g^\flat : T\mathcal{Q} \to T^*\mathcal{Q}$ by*

$$g^\flat(u)(v) = g_p(u,v) \quad \forall p \in \mathcal{Q} \text{ and } \forall u,v \in T_p\mathcal{Q},$$

*and its **inverse musical isomorphism** $g^\sharp : T^*\mathcal{Q} \to T\mathcal{Q}$. The Riemannian metric $g(\cdot,\cdot) = \langle \cdot,\cdot \rangle$ induces a **fiber metric** $g^*(\cdot,\cdot) = \langle\!\langle \cdot,\cdot \rangle\!\rangle$ on $T^*\mathcal{Q}$ by*

$$\langle\!\langle u,v \rangle\!\rangle = \langle g^\sharp(u), g^\sharp(v) \rangle \quad \forall u,v \in T^*\mathcal{Q},$$

represented by the positive definite symmetric matrix (g^{ij}) in local coordinates, which is the inverse of the Riemannian metric matrix (g_{ij}).

Definition 2.3. *The **Riemannian gradient** $\mathrm{gradf}(q) \in T_q\mathcal{Q}$ at a point $q \in \mathcal{Q}$ of a smooth function $f : \mathcal{Q} \to \mathbb{R}$ is the tangent vector at q such that*

$$\langle \mathrm{gradf}(q), u\rangle = df(q)u \qquad \forall u \in T_q\mathcal{Q},$$

where df is the differential of f.

Definition 2.4. *A **vector field** on a Riemannian manifold $\mathcal{Q}$ is a map $X : \mathcal{Q} \to T\mathcal{Q}$ such that $X(q) \in T_q\mathcal{Q}$ for all $q \in \mathcal{Q}$. The set of all vector fields on $\mathcal{Q}$ is denoted $\mathcal{X}(\mathcal{Q})$. The **integral curve** at q of $X \in \mathcal{X}(\mathcal{Q})$ is the smooth curve c on $\mathcal{Q}$ such that $c(0) = q$ and $c'(t) = X(c(t))$.*

Definition 2.5. *A **geodesic** in a Riemannian manifold $\mathcal{Q}$ is a parametrized curve $\gamma : [0,1] \to \mathcal{Q}$ which is of minimal local length. It can be thought of as a curve having zero "acceleration" or constant "speed", that is as a generalization of the notion of straight line from Euclidean spaces to Riemannian manifolds. Given two points $q, \tilde{q} \in \mathcal{Q}$, a vector in $T_q\mathcal{Q}$ can be transported to $T_{\tilde{q}}\mathcal{Q}$ along a geodesic γ by an operation $\Gamma(\gamma)_q^{\tilde{q}} : T_q\mathcal{Q} \to T_{\tilde{q}}\mathcal{Q}$ called **parallel transport along** γ. We will simply write $\Gamma_q^{\tilde{q}}$ to denote the parallel transport along some geodesic connecting the two points $q, \tilde{q} \in \mathcal{Q}$, and given $A \in \mathcal{X}(\mathcal{Q})$, we will denote by $\Gamma(A)$ the parallel transport along integral curves of A. Note that parallel transport preserves inner products: given a geodesic γ from $q \in \mathcal{Q}$ to $\tilde{q} \in \mathcal{Q}$,*

$$g_q(u,v) = g_{\tilde{q}}\left(\Gamma(\gamma)_q^{\tilde{q}}u, \Gamma(\gamma)_q^{\tilde{q}}v\right) \qquad \forall u, v \in T_q\mathcal{Q}.$$

Definition 2.6. *Given $X, Y \in \mathcal{X}(\mathcal{Q})$, the **covariant derivative** $\nabla_X Y \in \mathcal{X}(\mathcal{Q})$ of Y along X is*

$$\nabla_X Y(q) = \lim_{h\to 0} \frac{\Gamma(\gamma)_{\gamma(h)}^{q} Y(\gamma(h)) - Y(q)}{h},$$

where γ is the unique integral curve of X such that $\gamma(0) = q$, for any $q \in \mathcal{Q}$.

Definition 2.7. *A function $f : \mathcal{Q} \to \mathbb{R}$ is called **L-smooth** if for any two points $q, \tilde{q} \in \mathcal{Q}$ and geodesic γ connecting them,*

$$\left\|\mathrm{gradf}(q) - \Gamma(\gamma)_{\tilde{q}}^{q}\mathrm{gradf}(\tilde{q})\right\| \leq L \ \mathrm{length}(\gamma).$$

Definition 2.8. *The **Riemannian Exponential map** $\mathrm{Exp}_q : T_q\mathcal{Q} \to \mathcal{Q}$ at $q \in \mathcal{Q}$ is defined by*

$$\mathrm{Exp}_q(v) = \gamma_v(1),$$

where γ_v is the unique geodesic in $\mathcal{Q}$ such that $\gamma_v(0) = q$ and $\gamma_v'(0) = v$, for any $v \in T_q\mathcal{Q}$. Exp_q *is a diffeomorphism in some neighborhood $U \subset T_q\mathcal{Q}$ containing 0, so we can define its inverse map, the **Riemannian Logarithm map** $\mathrm{Log}_p : \mathrm{Exp}_q(U) \to T_q\mathcal{Q}$.*

Definition 2.9. *Given a Riemannian manifold $\mathcal{Q}$ with sectional curvature bounded below by $K_{\min}$, and an upper bound D for the diameter of the considered domain, define*

$$\zeta = \begin{cases} \sqrt{-K_{\min}}D\coth\left(\sqrt{-K_{\min}}D\right) & \text{if } K_{\min} < 0 \\ 1 & \text{if } K_{\min} \geq 0 \end{cases}. \tag{2.1}$$

Note that $\zeta \geq 1$ since $x\coth x \geq 1$ for all real values of x.

2.2. Convexity in Riemannian Manifolds.

Definition 2.10. *A subset A of a Riemannian manifold $\mathcal{Q}$ is called **geodesically uniquely convex** if every two points of A are connected by a unique geodesic in A. A function $f : \mathcal{Q} \to \mathbb{R}$ is called **geodesically convex** if for any two points $q, \tilde{q} \in \mathcal{Q}$ and geodesic γ connecting them,*

$$f(\gamma(t)) \leq (1-t)f(q) + tf(\tilde{q}) \qquad \forall t \in [0,1].$$

Note that if f is a smooth geodesically convex function on a geodesically uniquely convex subset A of a Riemannian manifold, then

$$f(q) - f(\tilde{q}) \geq \langle \mathrm{gradf}(\tilde{q}), \mathrm{Log}_{\tilde{q}}(q)\rangle \qquad \forall q, \tilde{q} \in A.$$

A function $f : A \to \mathbb{R}$ is called ***geodesically λ-weakly-quasi-convex*** *with respect to $q \in \mathcal{Q}$ for some $\lambda \in (0,1]$ if*

$$\lambda\left(f(q) - f(\tilde{q})\right) \geq \langle \mathrm{gradf}(\tilde{q}), \mathrm{Log}_{\tilde{q}}(q)\rangle \qquad \forall \tilde{q} \in A.$$

A function $f : A \to \mathbb{R}$ is called ***geodesically μ-strongly-convex*** *for some $\mu > 0$ if*

$$f(q) - f(\tilde{q}) \geq \langle \mathrm{gradf}(\tilde{q}), \mathrm{Log}_{\tilde{q}}(q)\rangle + \frac{\mu}{2}\|\mathrm{Log}_{\tilde{q}}(q)\|^2 \qquad \forall q, \tilde{q} \in A.$$

A local minimum of a geodesically convex or λ-weakly-quasi-convex function is also a global minimum, and a geodesically strongly convex function either has no minimum or a unique global minimum. Also note that a geodesically convex function is λ-weakly-quasi-convex with $\lambda = 1$.

2.3. Lagrangian and Hamiltonian Mechanics.

Given a n-dimensional Riemannian manifold $\mathcal{Q}$ with local coordinates $(q^1, \dots, q^n)$, a **Lagrangian** is a function $L : T\mathcal{Q} \times \mathbb{R} \to \mathbb{R}$. The corresponding **action integral** $\mathcal{S}$ is defined to be the functional

$$\mathcal{S}(q) = \int_0^T L(q, \dot{q}, t)dt, \tag{2.2}$$

over the space of smooth curves $q : [0,T] \to \mathcal{Q}$. **Hamilton's Variational Principle** states that $\delta S = 0$ where the variation δS is induced by an infinitesimal variation δq of the trajectory q that vanishes at the endpoints. Hamilton's Variational Principle can be shown to be equivalent to the **Euler–Lagrange equations**

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}^k}\right) = \frac{\partial L}{\partial q^k} \qquad \text{for } k = 1, \dots, n. \tag{2.3}$$

The **Legendre transform** $\mathbb{F}L : T\mathcal{Q} \to T^*\mathcal{Q}$ of L is defined fiberwise by $\mathbb{F}L : (q^i, \dot{q}^i) \mapsto (q^i, p_i)$ where $p_i = \frac{\partial L}{\partial \dot{q}^i} \in T^*\mathcal{Q}$ is the **conjugate momentum** of q^i. We can then define the associated **Hamiltonian** $H : T^*\mathcal{Q} \to \mathbb{R}$ by

$$H(q,p,t) = \sum_{j=1}^n p_j \dot{q}^j - L(q, \dot{q}, t)\Bigg|_{p_i = \frac{\partial L}{\partial \dot{q}^i}}. \tag{2.4}$$

We can also define a Hamiltonian Variational Principle on the Hamiltonian side in momentum phase space

$$\delta \int_0^T \sum_{j=1}^n \left[p_j \dot{q}^j - H(q,p,t)\right] dt = 0, \tag{2.5}$$

where the variation is induced by an infinitesimal variation δq of the trajectory q that vanishes at the endpoints. This is equivalent to **Hamilton's equations**, given by

$$\dot{p}_k = -\frac{\partial H}{\partial q^k}(p,q), \qquad \dot{q}^k = \frac{\partial H}{\partial p_k}(p,q) \qquad \text{for } k = 1, \dots, n, \tag{2.6}$$

which can also be shown to be equivalent to the Euler–Lagrange equations (2.3).

3. Variational Formulation and Convergence Rates

3.1. Inspiration.

A variational framework was introduced in [25] for accelerated optimization on normed vector spaces. Given a convex, continuously differentiable function $h : \mathcal{X} \to \mathbb{R}$ on a normed vector space $\mathcal{X}$ such that $\|\nabla h(x)\| \to \infty$ as $\|x\| \to \infty$, its corresponding Bregman divergence is defined by

$$D_h(x,y) = h(y) - h(x) - \langle \nabla h(x), y - x\rangle. \tag{3.1}$$

The Bregman Lagrangian and Hamiltonian are then defined to be

$$\begin{aligned} \mathcal{L}_{\alpha,\beta,\gamma}(x,v,t) &= e^{\alpha_t + \gamma_t}\left[D_h\left(x + e^{-\alpha_t}v, x\right) - e^{\beta_t} f(x)\right], \\ \mathcal{H}_{\alpha,\beta,\gamma}(x,r,t) &= e^{\alpha_t + \gamma_t}\left[D_{h^*}\left(\nabla h(x) + e^{-\gamma_t} r, \nabla h(x)\right) + e^{\beta_t} f(x)\right], \end{aligned} \tag{3.2}$$

which are scalar-valued functions of position $x \in \mathcal{X}$, velocity $v \in \mathbb{R}^d$ or momentum $r \in \mathbb{R}^d$, and of time t. Here, $h^* : \mathcal{X}^* \to \mathbb{R}$ denotes the Legendre transform (or convex dual function) of h, defined by $h^*(w) = \sup_{z\in\mathcal{X}} [\langle w, z\rangle - h(z)]$. The Bregman Lagrangian and Hamiltonian family is parametrized by smooth functions of time, $\alpha_t = \alpha(t), \beta_t = \beta(t), \gamma_t = \gamma(t)$, which are said to satisfy the ideal scaling conditions if

$$\dot\beta_t \leq e^{\alpha_t} \qquad \text{and} \qquad \dot\gamma_t = e^{\alpha_t}. \tag{3.3}$$

If the ideal scaling conditions are satisfied, then by Theorem 1.1 in [25],

$$f(x(t)) - f(x^*) \leq \mathcal{O}(e^{-\beta_t}). \tag{3.4}$$

Another very important property of this family of Bregman Lagrangians is its closure under time dilation, proven in Theorem 1.2 of [25]:

Theorem 3.1. *If $x(t)$ satisfies the Euler-Lagrange equations corresponding to the Bregman Lagrangian $\mathcal{L}_{\alpha,\beta,\gamma}$, then the reparametrized curve $y(t) = x(\tau(t))$ satisfies the Euler-Lagrange equations corresponding to the modified Bregman Lagrangian $\mathcal{L}_{\tilde\alpha,\tilde\beta,\tilde\gamma}$ where $\tilde\alpha_t = \alpha_{\tau(t)} + \log\dot\tau(t)$, $\tilde\beta_t = \beta_{\tau(t)}$, and $\tilde\gamma_t = \gamma_{\tau(t)}$. Furthermore α,β,γ satisfy the ideal scaling conditions (3.3) if and only if $\tilde\alpha,\tilde\beta,\tilde\gamma$ do.*

We will now extend these results to the Riemannian manifold setting. Throughout this paper, we will make the following assumptions on the function $f : \mathcal{Q} \to \mathbb{R}$ to be minimized and on the ambient Riemannian manifold $\mathcal{Q}$, which are standard assumptions in Riemannian optimization [3; 4; 26; 27]:

Assumption 1. *Solutions of the differential equations derived in this paper remain inside a geodesically uniquely convex subset A of a complete Riemannian manifold $\mathcal{Q}$ (i.e. any two points in $\mathcal{Q}$ can be connected by a geodesic), such that* $\operatorname{diam}(A)$ *is bounded above by some constant D, that the sectional curvature is bounded from below by $K_{\min}$ on A, and that* Exp_q *is well-defined for any $q \in A$, and its inverse* Log_q *is well-defined and differentiable on A for any $q \in A$. Furthermore, f is bounded below, geodesically L-smooth and all its minima are inside A.*

3.2. **Convex and Weakly-Quasi-Convex Cases.** Suppose that $f : \mathcal{Q} \to \mathbb{R}$ is a given geodesically λ-weakly-quasi-convex function, and that Assumption 1 holds true. Since a geodesically convex function is λ-weakly-quasi-convex with $\lambda = 1$, the following treatment also applies to the case where f is geodesically convex. We define a family of Bregman Lagrangians $\mathcal{L}_{\alpha,\beta,\gamma} : T\mathcal{Q} \times \mathbb{R} \to \mathbb{R}$ parametrized by smooth functions of time α, β, γ by

$$\boxed{\mathcal{L}_{\alpha,\beta,\gamma}(X, V, t) = \frac{1}{2} e^{\lambda^{-1}\zeta\gamma_t - \alpha_t} \langle V, V\rangle - e^{\alpha_t + \beta_t + \lambda^{-1}\zeta\gamma_t} f(X),} \tag{3.5}$$

and the corresponding Bregman Hamiltonians $\mathcal{H}_{\alpha,\beta,\gamma} : T^*\mathcal{Q} \times \mathbb{R} \to \mathbb{R}$ are given by

$$\boxed{\mathcal{H}_{\alpha,\beta,\gamma}(X, R, t) = \frac{1}{2} e^{\alpha_t - \lambda^{-1}\zeta\gamma_t} \langle\!\langle R, R\rangle\!\rangle + e^{\alpha_t + \beta_t + \lambda^{-1}\zeta\gamma_t} f(X),} \tag{3.6}$$

where $X \in \mathcal{Q}$ denotes position on the manifold $\mathcal{Q}$, V is the velocity vector field, R is the momentum covector field, t is the time variable, and ζ is given by equation (2.1). This family of functions is a generalization of the Bregman Lagrangians and Hamiltonians introduced in [25] for the convex continuously differentiable function $h(x) = \frac{1}{2}\langle x, x\rangle$. Throughout this paper, we will assume that the parameter functions α, β, γ satisfy the ideal scaling conditions (3.3).

Theorem 3.2. *The Bregman Euler–Lagrange equation corresponding to the Bregman Lagrangian $\mathcal{L}_{\alpha,\beta,\gamma}$ is given by*

$$\boxed{\nabla_{\dot X}\dot X + \left(\lambda^{-1}\zeta e^{\alpha_t} - \dot\alpha_t\right)\dot X + e^{2\alpha_t + \beta_t}\operatorname{gradf}(X) = 0.} \tag{3.7}$$

Proof. See Appendix A.1.

Theorem 3.3. *Suppose that $f : \mathcal{Q} \to \mathbb{R}$ is a geodesically λ-weakly-quasi-convex function, and that Assumption 1 is satisfied. Then, any solution $X(t)$ to the Bregman Euler–Lagrange equation (3.7) converges to a minimizer x^* of f with rate*

$$\boxed{f(X(t)) - f(x^*) \leq \frac{2\lambda^2 e^{\beta_0}\left(f(x_0) - f(x^*)\right) + \zeta \|\mathrm{Log}_{x_0}(x^*)\|^2}{2\lambda^2 e^{\beta_t}} = \mathcal{O}(e^{-\beta_t}).} \tag{3.8}$$

Proof. See Appendix B.

A $p > 0$ parametrized subfamily of Bregman Lagrangians and Hamiltonians, that is of particular practical interest, is given by the choice of parameter functions

$$\boxed{\alpha_t = \log p - \log t, \qquad \beta_t = p \log t + \log C, \qquad \gamma_t = p \log t,} \tag{3.9}$$

where $C > 0$ is a constant. This yields the p-Bregman Lagrangian and Hamiltonian given by

$$\boxed{\mathcal{L}_p(X, V, t) = \frac{t^{\lambda^{-1}\zeta p + 1}}{2p}\langle V, V\rangle - Cpt^{(\lambda^{-1}\zeta+1)p-1} f(X),} \tag{3.10}$$

$$\boxed{\mathcal{H}_p(X, R, t) = \frac{p}{2t^{\lambda^{-1}\zeta p+1}}\langle\!\langle R, R\rangle\!\rangle + Cpt^{(\lambda^{-1}\zeta+1)p-1} f(X),} \tag{3.11}$$

and the corresponding p-Bregman Euler–Lagrange equations are given by

$$\boxed{\nabla_{\dot{X}}\dot{X} + \frac{\zeta p + \lambda}{\lambda t}\dot{X} + Cp^2 t^{p-2}\mathrm{gradf}(X) = 0.} \tag{3.12}$$

Theorem 3.4. *Suppose that $f : \mathcal{Q} \to \mathbb{R}$ is a geodesically weakly-quasi-convex function, and that Assumption 1 is satisfied. Then, the p-Bregman Euler–Lagrange equation (3.12) has a solution, and any solution $X(t)$ converges to a minimizer x^* of f with rate* $\boxed{f(X(t)) - f(x^*) \leq \mathcal{O}(1/t^p)}$.

Proof. See Appendix C.1 for the existence of a solution to the p-Bregman Euler–Lagrange equations. The $\mathcal{O}(1/t^p)$ convergence rate follows directly from Theorem 3.3.

Note that this theorem reduces to Theorem 5 from [3] when $p = 2$ and $C = 1/4$.

Remark. *To construct this variational framework for accelerated optimization, we first constructed candidate p-equations with the desired $\mathcal{O}(1/t^p)$ convergence rates, and then designed Lagrangians whose p-Bregman Euler–Lagrange equations matched the candidate p-equations, by inspection. We then used a similar approach to extend these results to the general α, β, γ case presented here.*

Remark. *In our generalization of the Bregman Lagrangian and Hamiltonian to Riemannian manifolds, we have specialized to the case where $h(x) = \frac{1}{2}\|x\|^2$, because its Hessian $\nabla^2 h(x)$ is the identity matrix, which significantly simplifies the Euler–Lagrange equations and the analysis. In addition, it avoids the complication of making intrinsic sense of terms like $X + e^{-\alpha}V$ in the vector space Bregman Lagrangians and Hamiltonians, which require the use of Riemannian geodesics and exponentials since $X \in \mathcal{Q}$ while $V \in T_X\mathcal{Q}$.*

3.3. **Strongly Convex Case.** Suppose $f : \mathcal{Q} \to \mathbb{R}$ is a geodesically μ-strongly-convex function, and that Assumption 1 is satisfied. With ζ given by equation (2.1), let

$$\eta = \left(\frac{1}{\sqrt{\zeta}} + \sqrt{\zeta}\right)\sqrt{\mu}. \tag{3.13}$$

We define the corresponding Lagrangian $\mathcal{L}^{SC} : T\mathcal{Q} \times \mathbb{R} \to \mathbb{R}$ by

$$\boxed{\mathcal{L}^{SC}(X, V, t) = \frac{e^{\eta t}}{2}\langle V, V\rangle - e^{\eta t} f(X),} \tag{3.14}$$

and the corresponding Hamiltonian $\mathcal{H}^{SC}: T^*\mathcal{Q} \times \mathbb{R} \to \mathbb{R}$ is given by

$$\boxed{\mathcal{H}^{SC}(X,R,t) = \frac{e^{-\eta t}}{2}\langle\langle R,R\rangle\rangle + e^{\eta t} f(X).} \tag{3.15}$$

Theorem 3.5. *The Euler–Lagrange equation corresponding to the Lagrangian $\mathcal{L}^{SC}$ is given by*

$$\boxed{\nabla_{\dot{X}}\dot{X} + \eta\dot{X} + \mathrm{gradf}(X) = 0.} \tag{3.16}$$

Proof. The derivation of the Euler–Lagrange equation is presented in Appendix A.2.

Theorem 3.6. *Suppose $f : \mathcal{Q} \to \mathbb{R}$ is a geodesically μ-strongly-convex function, and suppose that Assumption 1 is satisfied. Then, the Euler–Lagrange equation (3.16) has a solution, and any solution $X(t)$ converges to a minimizer x^* of f with rate*

$$\boxed{f(X(t)) - f(x^*) \le \frac{\mu\|\mathrm{Log}_{x_0}(x^*)\|^2 + 2\left(f(x_0) - f(x^*)\right)}{2e^{\sqrt{\frac{\mu}{\zeta}}t}}.} \tag{3.17}$$

Proof. See Appendix C.2 for the existence of a solution to the Euler–Lagrange equation (3.16), and Theorem 7 from [3] for the convergence rate.

4. Numerical Experiments

The p-Bregman Euler–Lagrange equation (3.12) can be rewritten as the first-order system

$$\dot{X} = V, \qquad \nabla_V V = -\frac{\zeta p + \lambda}{\lambda t}V - Cp^2 t^{p-2}\mathrm{gradf}(X), \tag{4.1}$$

for the geodesically λ-weakly-quasi-convex case, and the Euler–Lagrange equation (3.16) corresponding to the Lagrangian $\mathcal{L}^{SC}$ can be rewritten as the first-order system

$$\dot{X} = V, \qquad \nabla_V V = -\left(\frac{1}{\sqrt{\zeta}} + \sqrt{\zeta}\right)\sqrt{\mu}V - \mathrm{gradf}(X), \tag{4.2}$$

for the μ-strongly convex case. As in [3], we can adapt a semi-implicit Euler scheme (explicit Euler update for the velocity V followed by an update for position X based on the updated value of V) to the Riemannian setting to obtain the following algorithm:

Algorithm 1: Semi-Implicit Euler Integration of the p-Bregman Euler–Lagrange Equations

```
Input: A function $f : \mathcal{Q} \to \mathbb{R}$. Constants $C, h, p > 0$. $X_0 \in \mathcal{Q}$. $V_0 \in T_{X_0}\mathcal{Q}$.
while convergence criterion is not met do
    if $f$ is $\mu$-geodesically strongly convex then
        $b_k \leftarrow 1 - h\left(\frac{1}{\sqrt{\zeta}} + \sqrt{\zeta}\right)\sqrt{\mu}$,    $c_k \leftarrow 1$
    else if $f$ is $\lambda$-weakly-quasi-convex then
        $b_k \leftarrow 1 - \frac{\zeta p + \lambda}{\lambda k}$,    $c_k \leftarrow Cp^2(kh)^{p-2}$
    Version I: $a_k \leftarrow b_k V_k - hc_k \mathrm{gradf}(X_k)$
    Version II: $a_k \leftarrow b_k V_k - hc_k \mathrm{gradf}\left(\mathrm{Exp}_{X_k}(hb_k V_k)\right)$
    $X_{k+1} \leftarrow \mathrm{Exp}_{X_k}(ha_k)$,    $V_{k+1} \leftarrow \Gamma_{X_k}^{X_{k+1}} a_k$
```

Version I of Algorithm 1 corresponds to the usual update for the Semi-Implicit Euler scheme, while Version II is inspired by the reformulation of Nesterov's method from [24] that uses a corrected gradient $\nabla f(X_k + hb_k V_k)$ instead of the traditional gradient $\nabla f(X_k)$. Note that the SIRNAG algorithm presented in [3] corresponds to the special case where $p = 2$ and $C = 1/4$.

The first problem we have investigated is the problem presented in [3] of minimizing the (strongly convex) distance function $f(x) = \frac{1}{2}d(x,q)^2$ for a given point q, on a subset of chosen finite diameter of the hyperbolic plane $\mathbb{H}^2$, which is a manifold with constant negative curvature $K = -1$.

The second problem we have investigated is Rayleigh quotient optimization. Eigenvectors corresponding to the largest eigenvalue of a symmetric $n \times n$ matrix A maximize the Rayleigh quotient $\frac{v^\top A v}{v^\top v}$ over $\mathbb{R}^n$. Thus, a unit eigenvector v^* corresponding to the largest eigenvalue of the matrix A is a minimizer of the function $f(v) = -v^\top A v$, over the unit sphere $\mathcal{Q} = \mathbb{S}^{n-1}$, which can be thought of as a Riemannian submanifold with constant positive curvature $K = 1$ of $\mathbb{R}^n$ endowed with the Riemannian metric inherited from the Euclidean inner product $g_v(u,w) = u^\top w$. More information concerning the geometry of $\mathbb{S}^{n-1}$, such as its tangent bundle, its orthogonal projection and exponential map can be found in [1]. Solving the Rayleigh quotient optimization problem efficiently is challenging when the given symmetric matrix A is ill-conditioned and high-dimensional. Note that an efficient algorithm that solves the above minimization problem can also be used to find eigenvectors corresponding to the smallest eigenvalue of A by using the fact that the eigenvalues of A are the negative of the eigenvalues of $-A$.

Experiments carried out in [3] showed that SIRNAG (the convex $p = 2$ Algorithm 1) and the strongly convex Algorithm 1 were of comparable efficiency or more efficient than the standard Riemannian Gradient Descent (RGD) method, depending on the properties of the objective function and on the geometry of the Riemannian manifold. We have conducted further numerical experiments to investigate how the simple discretization of higher-order $p = 6$ Bregman dynamics compared to its $p = 2$ counterpart, and to see whether it matches the theoretical $\mathcal{O}(t^{-p})$ convergence rate. The numerical results obtained for the distance minimization and Rayleigh minimization problems are illustrated in Figure 1, where all the algorithms were implemented with the same fixed timestep. We can see that the $p = 6$ algorithms outperform their $p = 2$ counterparts, and that the efficiency improvement is very important. Furthermore, both versions of the $p = 6$ Algorithm 1 exhibit a faster convergence rate than the theoretical $\mathcal{O}(t^{-6})$ rate. While Version I of Algorithm 1 exhibits polynomial rates of $\mathcal{O}(t^{-10.8})$ and $\mathcal{O}(t^{-9})$ on the objective functions considered, Version II of Algorithm 1 exhibits a much faster exponential rate of convergence on both examples.

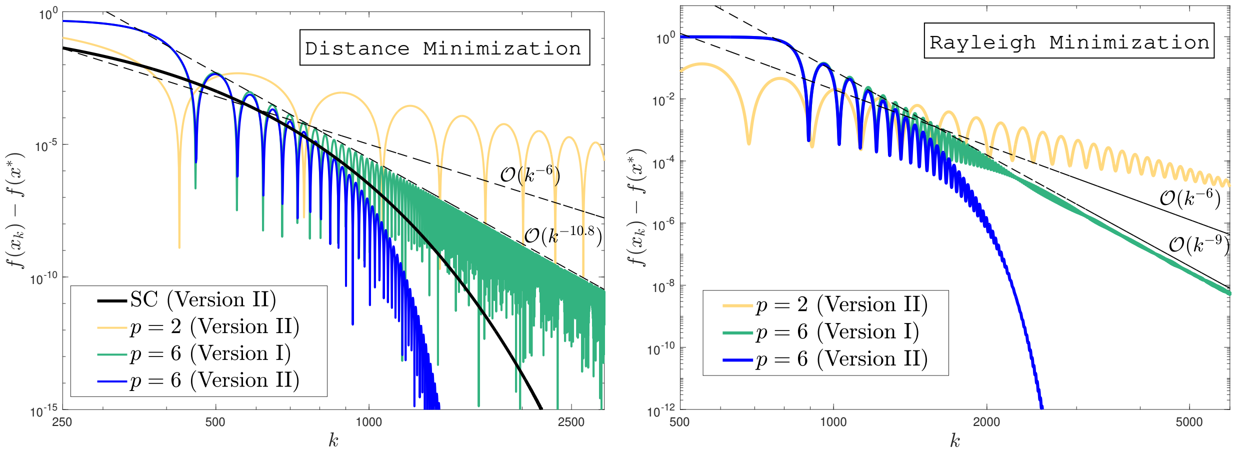

FIGURE 1. Comparison of the rates of convergence of the μ-strongly convex (SC) Algorithm 1 and convex Algorithms 1 with different values of p and with the two versions of the update corresponding to the traditional and corrected gradients. Note that all the algorithms were implemented with the same timestep h.

Figure 2 displays the evolution of the rates of convergence of Version 1 of the convex Algorithm 1 as the value of the parameter p is increased from $p = 4$ to $p = 16$ for the distance minimization and Rayleigh minimization problems. We can clearly see an improvement in the convergence rates as the value of p increases, and for each value of p the algorithm achieves a faster rate of convergence than the theoretical $\mathcal{O}(t^{-p})$ rates.

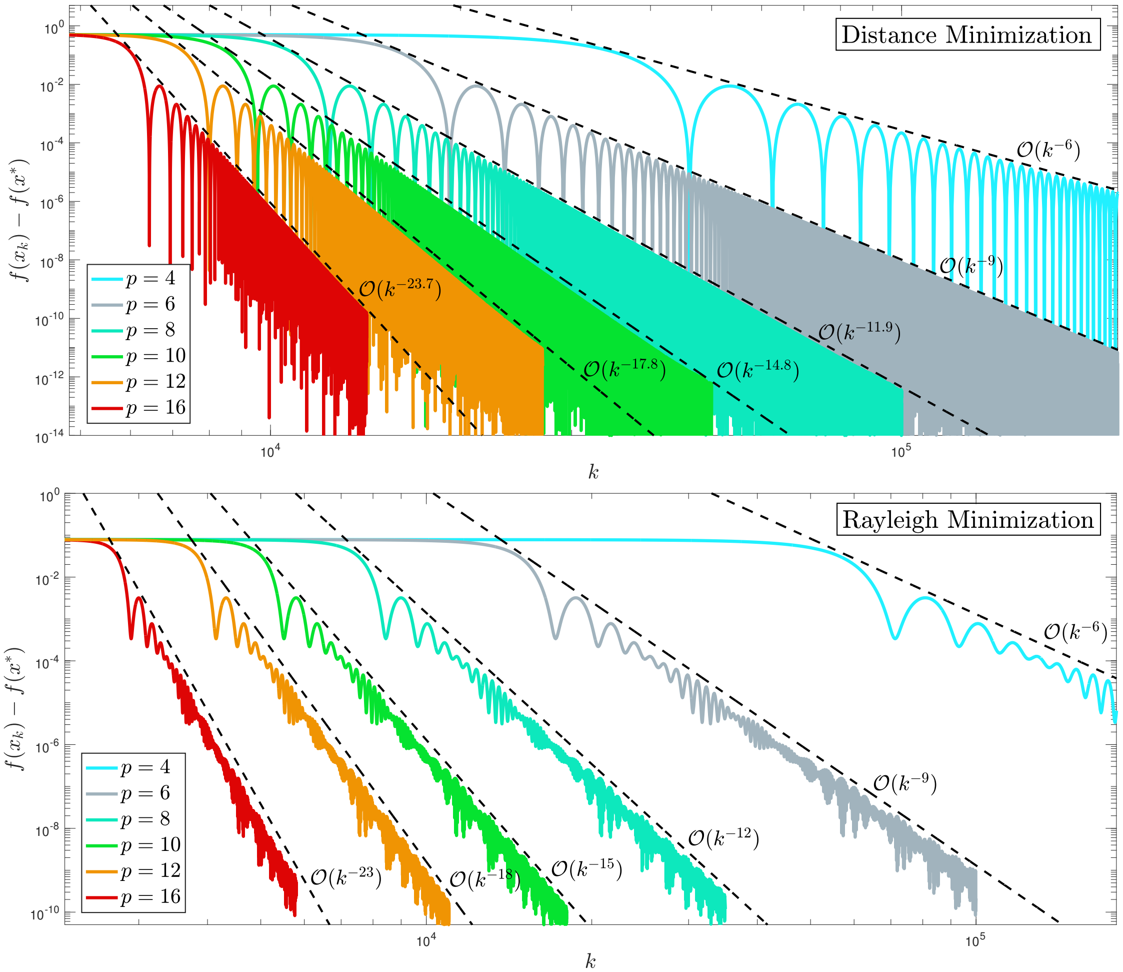

FIGURE 2. Evolution of the rates of convergence of Version 1 of the convex Algorithm 1 with different values of p. Note that all the algorithms were implemented with the same timestep h.

Note however that an increase in the value of p in Algorithm 1, which corresponds to an increase in the order of the Bregman dynamics integrated, requires a decrease in the timestep, in agreement with intuitive expectations. This timestep decrease requirement is especially important due to the polynomially growing $h(kh)^{p-2}$ coefficient multiplying the gradient of f in the updates of the algorithm. Such a decrease in the timestep does not really affect the convergence rate, but the transition between the initialization and convergence phases takes longer. As a consequence, by using larger timesteps, the algorithm corresponding to a smaller value of p might achieve a desired convergence criterion with fewer iterations than the algorithm corresponding to a larger value of p, despite having a slower convergence rate. Similar issues arise when discretizing the continuous Euler–Lagrange flow associated with accelerated optimization on vector spaces, and in that situation, it was observed that time-adaptive symplectic integrators based on Hamiltonian variational integrators resulted in dramatically improved robustness and stability. As such, it will be natural to explore generalizations of time-adaptive symplectic integrators based on Hamiltonian variational integrators applied to Poincaré transformed Hamiltonians, that respect the Riemannian manifold structure in order to yield more robust and stable numerical discretizations of the flows we have studied in this paper in order to construct accelerated optimization algorithms on Riemannian manifolds. We will lay the foundation for such time-adaptive symplectic integrators in Section 5.

Finally, Figure 3 shows that the discretization empirically converges to the solution of the ODE as the timestep h goes to 0. Note that although all the discretizations follow the ODE trajectory closely, smaller timesteps result in a larger number of iterations, especially to transition from the initialization plateau to the convergence phase (around time $t = 4$ in the example presented in Figure 3). A theoretical shadowing result bounding the error between the discrete-time RGD and its continuous-time limiting ODE was obtained in [3]. It would be desirable to obtain similar shadowing results in the future for discretizations of the class of ODEs considered here, perhaps drawing inspiration from [28]. However, such a result might be very difficult to obtain because momentum methods lack contraction, are nondescending, and are highly oscillatory [3; 22]. While it is hoped that the continuous analysis in this paper will eventually guide the convergence analysis of discrete-time algorithms, this does not appear to be a straightforward exercise, as one would first need to reconcile the arbitrarily fast $\mathcal{O}(1/t^p)$ rate of convergence of the continuous-time trajectories with Nesterov's barrier theorem of $\mathcal{O}(1/k^2)$ for discrete-time algorithms.

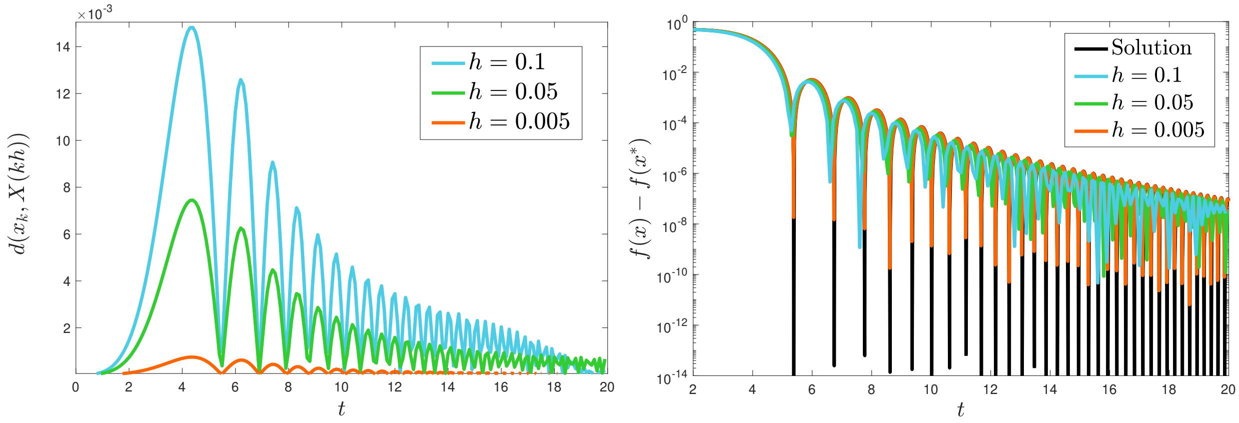

FIGURE 3. Discretization errors (top graph) and convergence rates (bottom graphs) of Version I of the $p = 5$ convex Algorithm 1 with different values of h for the distance minimization problem. The true solution of the differential equation was approximated by the same algorithm with a very small timestep $h = 10^{-5}$.

5. Time Invariance and Poincaré Transformation

Let $f : \mathcal{Q} \to \mathbb{R}$ be a given λ-weakly-quasi-convex function, and suppose Assumption 1 is satisfied. In Section 3, we formulated a variational framework for the minimization of f, via Bregman Lagrangians and Hamiltonians. We now extend Theorem 3.1 to Riemannian manifolds.

Theorem 5.1. *Suppose that Assumption 1 is satisfied and that the curve $X(t)$ satisfies the Riemannian Bregman Euler–Lagrange equation (3.7) corresponding to $\mathcal{L}_{\alpha,\beta,\gamma}$. Then the reparametrized curve $X(\tau(t))$ satisfies the Bregman Euler–Lagrange equation (3.7) corresponding to the modified Riemannian Bregman Lagrangian $\mathcal{L}_{\tilde{\alpha},\tilde{\beta},\tilde{\gamma}}$ where $\tilde{\alpha}_t = \alpha_{\tau(t)} + \log \dot{\tau}(t)$, $\tilde{\beta}_t = \beta_{\tau(t)}$, and $\tilde{\gamma}_t = \gamma_{\tau(t)}$. Furthermore α, β, γ satisfy the ideal scaling conditions (3.3) if and only if $\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma}$ do.*

Proof. See Appendix D.

As a special case, we have the following theorem:

Theorem 5.2. *Suppose that $f : \mathcal{Q} \to \mathbb{R}$ is a geodesically λ-weakly-quasi-convex function, and that Assumption 1 is satisfied. Suppose $X(t)$ satisfies the p-Bregman Euler–Lagrange equation (3.12). Then, the reparametrized curve $X(t^{\mathring{p}/p})$ satisfies the $\mathring{p}$-Bregman Euler–Lagrange equation (3.12).*

Thus, the entire subfamily of Bregman trajectories indexed by the parameter p can be obtained by speeding up or slowing down along the Bregman curve in spacetime corresponding to any specific value of p. Inspired by the computational efficiency of the approach introduced in [8], it is natural to attempt to exploit the time-rescaling property of the Bregman dynamics together with a carefully chosen Poincaré transformation to transform the p-Bregman Hamiltonian into an autonomous version of the $\mathring{p}$-Bregman Hamiltonian in extended phase-space, where $\mathring{p} < p$. This would allow us to integrate the higher-order p-Bregman dynamics while benefiting from the computational efficiency of integrating the lower-order $\mathring{p}$-Bregman dynamics. Explicitly, the time rescaling $\tau(t) = t^{\mathring{p}/p}$ is associated to the monitor function

$$\frac{dt}{d\tau} = g_{p\to\mathring{p}}(t) = \frac{p}{\mathring{p}} t^{1-\mathring{p}/p}, \tag{5.1}$$

and generates a Poincaré transformed Hamiltonian

$$\bar{\mathcal{H}}_{p\to\mathring{p}}(\bar{X},\bar{R}) = g_{p\to\mathring{p}}(X^t)\left(\mathcal{H}_p\left(\bar{X},R\right) + R^t\right), \tag{5.2}$$

in the extended space $\bar{\mathcal{Q}} = \mathcal{Q} \times \mathbb{R}$ where $\bar{X} = \begin{bmatrix} X \\ X^t \end{bmatrix}$ and $\bar{R} = \begin{bmatrix} R \\ R^t \end{bmatrix}$. We will make the conventional choice $X^t = t$, with conjugate momentum R^t, and $R^t(0) = -\mathcal{H}_p(X(0), R(0), 0) = -H_0$, which is chosen so that $\bar{\mathcal{H}}_{p\to\mathring{p}}(\bar{X},\bar{R}) = 0$ along all integral curves through $(\bar{X}(0), \bar{R}(0))$. The time t shall be referred to as the physical time, while τ will be referred to as the fictive time. The corresponding Hamiltonian equations of motion in the extended phase space are then given by

$$\dot{\bar{X}} = \frac{\partial \bar{\mathcal{H}}_{p\to\mathring{p}}}{\partial \bar{R}}, \qquad \dot{\bar{R}} = -\frac{\partial \bar{\mathcal{H}}_{p\to\mathring{p}}}{\partial \bar{X}}. \tag{5.3}$$

Now, suppose $(\bar{X}(\tau), \bar{R}(\tau))$ are solutions to these extended equations of motion, and let $(x(t), r(t))$ solve Hamilton's equations for the original Hamiltonian $\mathcal{H}_p$. Then

$$\bar{\mathcal{H}}_{p\to\mathring{p}}(\bar{X}(\tau), \bar{R}(\tau)) = \bar{\mathcal{H}}_{p\to\mathring{p}}(\bar{X}(0), \bar{R}(0)) = 0.$$

Thus, the components $(X(\tau), R(\tau))$ in the original phase space of $(\bar{X}(\tau), \bar{R}(\tau))$ satisfy

$$\mathcal{H}_p(X(\tau), R(\tau), \tau) = -R^t(\tau), \qquad \mathcal{H}_p(X(0), R(0), 0) = -R^t(0) = \mathcal{H}_p(x(0), r(0), 0).$$

Therefore, $(X(\tau), R(\tau))$ and $(x(t), r(t))$ both satisfy Hamilton's equations for the original Hamiltonian $\mathcal{H}_p$ with the same initial values, so they must be the same.

As a consequence, instead of integrating the p-Bregman Hamiltonian system (3.11), we can focus on the Poincaré transformed Hamiltonian $\bar{\mathcal{H}}_{p\to\mathring{p}}$ in extended phase-space given by equation (5.2), with $\mathcal{H}_p$ and $g_{p\to\mathring{p}}$ given by equations (3.11) and (5.1), that is

$$\boxed{\bar{\mathcal{H}}_{p\to\mathring{p}}(\bar{X},\bar{R}) = \frac{p^2}{2\mathring{p}(X^t)^{\lambda^{-1}\zeta p + \mathring{p}/p}} \langle\!\langle R, R \rangle\!\rangle + \frac{Cp^2}{\mathring{p}} (X^t)^{(\lambda^{-1}\zeta+1)p - \mathring{p}/p} f(X) + \frac{p}{\mathring{p}} (X^t)^{1-\mathring{p}/p} R^t,} \tag{5.4}$$

The resulting integrator has constant timestep in fictive time τ but variable timestep in physical time t. In our prior work on discretizations of variational formulations of accelerated optimization on normed spaces [8], we performed a very careful computational study of how time-adaptivity and symplecticity of the numerical scheme improve the performance of the resulting numerical optimization algorithm. In particular, we observed that time-adaptive Hamiltonian variational discretizations, which are automatically symplectic, with adaptive timesteps informed by the time invariance of the family of p-Bregman Lagrangians and Hamiltonians yielded the most robust and computationally efficient numerical optimization algorithms, outperforming fixed-timestep symplectic discretizations, adaptive-timestep non-symplectic discretizations, and Nesterov's accelerated gradient algorithm which is neither time-adaptive nor symplectic. As such, it would be desirable to generalize the time-adaptive Hamiltonian variational integrator framework to Riemannian manifolds, and apply it to the variational formulation of accelerated optimization on Riemannian manifolds.

6. Conclusion

We have shown that on Riemannian manifolds, the convergence rate in continuous time of a geodesically convex or weakly-quasi-convex function $f(x(t))$ to its optimal value can be accelerated to an arbitrary convergence rate, which extended the results of [25] from normed vector spaces to Riemannian manifolds. This rate of convergence is achieved along solutions of the Euler–Lagrange and Hamilton's equations corresponding to a family of time-dependent Bregman Lagrangian and Hamiltonian systems on Riemannian manifolds. As was demonstrated in the normed vector space setting, such families of Bregman Lagrangians and Hamiltonians can be used to construct practical, robust, and computationally efficient numerical optimization algorithms that outperform Nesterov's accelerated gradient method by considering geometric structure-preserving discretizations of the continuous-time flows.

Numerical experiments implementing a simple discretization of the p-Bregman Euler–Lagrange equations applied to a distance minimization and Rayleigh minimization problems confirmed that the higher-order algorithms outperform significantly their lower-order counterparts and their theoretical $\mathcal{O}(1/t^p)$ convergence rates. Numerical results also showed that using a corrected gradient in the update instead of the traditional gradient, as was done in [24], improved the theoretically predicted polynomial convergence rate to an exponential rate of convergence in practice. While higher values of p result in faster rates of convergence, they usually require smaller timesteps and also appear to be more prone to stability issues under numerical discretization, which can cause the numerical optimization algorithm to diverge, but we anticipate that symplectic discretizations will address these stability issues.

Finally, in analogy to what was done in [25] for normed vector spaces, we proved that the family of time-dependent Bregman Lagrangian and Hamiltonians on Riemannian manifolds is closed under time rescaling. Inspired by the computational efficiency of the approach introduced in [8], we can then exploit this invariance property via a carefully chosen Poincaré transformation that will allow us to integrate higher-order p-Bregman dynamics while benefiting from the computational efficiency of integrating a lower-order $\mathring{p}$-Bregman Hamiltonian system.

It was observed in our prior computational experiments in the normed vector space case [8] that geometric discretizations which respect the time-rescaling invariance and symplecticity of the Bregman Lagrangian and Hamiltonian flows were substantially less prone to stability issues, and were therefore more robust, reliable, and computationally efficient. As such, it is natural to develop time-adaptive Hamiltonian variational integrators for the Bregman Hamiltonian introduced in this paper describing accelerated optimization on Riemannian manifolds.

Developing an intrinsic extension of Hamiltonian variational integrators to manifolds will require some additional work, since the current approach involves Type II/Type III generating functions $H_d^+(q_k.p_{k+1})$, $H_d^-(p_k, q_{k+1})$, which depend on the position at one boundary point, and the momentum at the other boundary point. However, this does not make intrinsic sense on a manifold, since one needs the base point in order to specify the corresponding cotangent space, and one should ideally consider a Hamiltonian variational integrator construction based on discrete Dirac mechanics [14], which would yield a generating function $E_d^+(q_k, q_{k+1}, p_{k+1})$, $E_d^-(q_k, p_k, q_{k+1})$, that depends on the position at both boundary points and the momentum at one of the boundary points. This approach can be viewed as a discretization of the generalized energy $E(q,v,p) = \langle p,v\rangle - L(q,v)$, in contrast to the Hamiltonian $H(q,p) = \mathrm{ext}_v \langle p,v\rangle - L(q,v) = \langle p,v\rangle - L(q,v)\big|_{p=\frac{\partial L}{\partial v}}$.

However, a more practical method relies on the fact that we have a Riemannian manifold, which is endowed with a Riemannian exponential and Riemannian logarithm that can be used to construct an extension of Hamiltonian variational integrators using geodesic normal coordinates. For many important matrix manifolds, one can replace the Riemannian exponential in the geodesic normal coordinates by a retraction [1], which is often constructed using matrix factorizations.

Another important case involves Riemannian submanifolds that are embedded in a Riemannian linear manifold and are realized as the level set of a submersion. The characterization of the submanifold as the level set of a submersion, together with the linear space structure of the embedding space, and the variational characterization of the dynamics naturally lends itself to the use of the Lagrange multiplier theorem, which allows one to use Hamiltonian variational integrators defined on the embedding space by including a Lagrange multiplier term involving the submersion in the Lagrangian or Hamiltonian [6]. This is analogous to the derivation of the SHAKE and RATTLE methods as a variational integrator for constrained systems (see, for example, §3.5 of [17]). Another practical method can be obtained by projecting the updates of Hamiltonian variational integrators defined on the embedding space onto the constraint manifold [7].

We anticipate that applying an appropriate generalization of Hamiltonian variational integrators to the Bregman Hamiltonians introduced in this paper will yield a novel class of robust and efficient accelerated optimization algorithms on Riemannian manifolds. It would also be desirable to analyze the resulting discrete-time algorithms and rigorously establish their rates of convergence. In addition, we would like to better understand how to reconcile the arbitrarily high rate of convergence one expects from the continuous-time analysis, with Nesterov's barrier theorem on the rate of convergence of discrete-time algorithms.

Acknowledgments

The authors were supported in part by NSF under grants DMS-1411792, DMS-1345013, DMS-1813635, by AFOSR under grant FA9550-18-1-0288, and by the DoD under grant 13106725 (Newton Award for Transformative Ideas during the COVID-19 Pandemic).

Appendix A. Derivation of the Euler–Lagrange Equations

A.1. **Convex and Weakly-Quasi-Convex Cases.**

Theorem A.1. *The Euler–Lagrange equation corresponding to the Lagrangian*

$$\mathcal{L}_{\alpha,\beta,\gamma}(X,V,t) = \frac{1}{2}e^{\lambda^{-1}\zeta\gamma_t-\alpha_t}\langle V,V\rangle - e^{\alpha_t+\beta_t+\lambda^{-1}\zeta\gamma_t}f(X),$$

is given by

$$\nabla_{\dot{X}}\dot{X} + \left(\lambda^{-1}\zeta e^{\alpha_t} - \dot{\alpha}_t\right)\dot{X} + e^{2\alpha_t+\beta_t}\mathrm{gradf}(X) = 0,$$

Proof. Consider a path on the manifold $\mathcal{Q}$ described in coordinates by

$$(x(t),\dot{x}(t)) = \left(q^1(t),\dots,q^n(t),v^1(t),\dots,v^n(t)\right).$$

Then, with $\langle\cdot,\cdot\rangle = \sum_{i,j=1}^n g_{ij}dx^i dx^j$, the Bregman Lagrangian $\mathcal{L}_{\alpha,\beta,\gamma}$ can be written as

$$\mathcal{L}_{\alpha,\beta,\gamma}\left(x(t),\dot{x}(t),t\right) = \frac{1}{2}e^{\lambda^{-1}\zeta\gamma_t-\alpha_t}\sum_{i,j=1}^n g_{ij}(x(t))v^i(t)v^j(t) - e^{\alpha_t+\beta_t+\lambda^{-1}\zeta\gamma_t}f(x(t)).$$

For $k = 1, \ldots n$,

$$\frac{d}{dt}\left(\frac{\partial \mathcal{L}_{\alpha,\beta,\gamma}}{\partial v^k}\left(x(t),\dot{x}(t),t\right)\right) = e^{\lambda^{-1}\zeta\gamma_t-\alpha_t}\sum_{i=1}^{n} g_{ik}(x(t))\frac{dv^i}{dt}(t) + e^{\lambda^{-1}\zeta\gamma_t-\alpha_t}\sum_{i,j=1}^{n}\frac{\partial g_{kj}}{\partial q^i}(x(t))v^i(t)v^j(t)$$
$$+ (\lambda^{-1}\zeta\dot{\gamma}_t - \dot{\alpha}_t)e^{\lambda^{-1}\zeta\gamma_t-\alpha_t}\sum_{i=1}^{n} g_{ik}(x(t))v^i(t),$$

$$\frac{\partial \mathcal{L}_{\alpha,\beta,\gamma}}{\partial q^k}\left(x(t),\dot{x}(t),t\right) = \frac{1}{2}e^{\lambda^{-1}\zeta\gamma_t-\alpha_t}\sum_{i,j=1}^{n}\frac{\partial g_{ij}}{\partial q^k}(x(t))v^i(t)v^j(t) - e^{\alpha_t+\beta_t+\lambda^{-1}\zeta\gamma_t}\frac{\partial f}{\partial q^k}(x(t)).$$

Multiplying both terms by $e^{\alpha_t-\lambda^{-1}\zeta\gamma_t}$, the Euler–Lagrange equations (2.3) for the Bregman Lagrangian $\mathcal{L}_{\alpha,\beta,\gamma}$ are given, for $k = 1, \ldots, n$, by

$$0 = \sum_{i=1}^{n} g_{ik}(x(t))\frac{dv^i}{dt}(t) + \sum_{i,j=1}^{n}\frac{\partial g_{kj}}{\partial q^i}(x(t))v^i(t)v^j(t) + (\lambda^{-1}\zeta\dot{\gamma}_t - \dot{\alpha}_t)\sum_{i=1}^{n} g_{ik}(x(t))v^i(t)$$
$$- \frac{1}{2}\sum_{i,j=1}^{n}\frac{\partial g_{ij}}{\partial q^k}(x(t))v^i(t)v^j(t) + e^{2\alpha_t+\beta_t}\frac{\partial f}{\partial q^k}(x(t)).$$

Rearranging terms, and multiplying by the matrix (g^{ij}) which is the inverse of (g_{ij}), we get, for $k = 1, \ldots n$, the equation

$$\left(\frac{dv^k}{dt}(t) + \sum_{i,j=1}^{n}\Gamma^k_{ij}(x(t))v^i(t)v^j(t)\right) + \left(\lambda^{-1}\zeta\dot{\gamma}_t - \dot{\alpha}_t\right)v^k(t) + e^{2\alpha_t+\beta_t}\left(\mathrm{gradf}(x(t))\right)^k = 0,$$

where Γ^k_{ij} are the Christoffel symbols given by $\Gamma^k_{ij} = \frac{1}{2}\sum_{l=1}^{n} g^{kl}\left[\frac{\partial g_{jl}}{\partial x^i} + \frac{\partial g_{li}}{\partial x^j} - \frac{\partial g_{ij}}{\partial x^l}\right]$, which gives the desired Euler–Lagrange equation once we use the ideal scaling equation $\dot{\gamma}_t = e^{\alpha_t}$. □

A.2. **Strongly Convex Case.**

Theorem A.2. *The Euler–Lagrange equation corresponding to the Lagrangian $\mathcal{L}^{SC}$ is given by*

$$\nabla_{\dot{X}}\dot{X} + \eta\dot{X} + \mathrm{gradf}(X) = 0.$$

Proof. Consider a path on the manifold $\mathcal{Q}$ described in coordinates by

$$(x(t),\dot{x}(t)) = \left(q^1(t),\ldots,q^n(t),v^1(t),\ldots,v^n(t)\right).$$

Then, with $\langle\cdot,\cdot\rangle = \sum_{i,j=1}^{n} g_{ij}dx^i dx^j$, the Lagrangian $\mathcal{L}^{SC}$ can be written as

$$\mathcal{L}^{SC}\left(x(t),\dot{x}(t),t\right) = \frac{e^{\eta t}}{2}\sum_{i,j=1}^{n} g_{ij}(x(t))v^i(t)v^j(t) - e^{\eta t}f(x(t)).$$

For $k = 1, \ldots n$,

$$\frac{d}{dt}\left(\frac{\partial \mathcal{L}^{SC}}{\partial v^k}\left(x(t),\dot{x}(t),t\right)\right) = e^{\eta t}\sum_{i=1}^{n} g_{ik}(x(t))\frac{dv^i}{dt}(t) + e^{\eta t}\sum_{i,j=1}^{n}\frac{\partial g_{kj}}{\partial q^i}(x(t))v^i(t)v^j(t)$$
$$+ \eta e^{\eta t}\sum_{i=1}^{n} g_{ik}(x(t))v^i(t),$$

$$\frac{\partial \mathcal{L}^{SC}}{\partial q^k}\left(x(t),\dot{x}(t),t\right) = e^{\eta t}\sum_{i,j=1}^{n}\frac{\partial g_{ij}}{\partial q^k}(x(t))v^i(t)v^j(t) - e^{\eta t}\frac{\partial f}{\partial q^k}(x(t)).$$

If we multiply both terms by $e^{-\eta t}$, the Euler–Lagrange equations (2.3) for the Lagrangian $\mathcal{L}^{SC}$ are given, for $k = 1, \dots, n$, by

$$\begin{aligned} 0 = \sum_{i=1}^{n} g_{ik}(x(t))\frac{dv^i}{dt}(t) + \sum_{i,j=1}^{n} \frac{\partial g_{kj}}{\partial q^i}(x(t))v^i(t)v^j(t) + \eta \sum_{i=1}^{n} g_{ik}(x(t))v^i(t) \\ - \frac{1}{2}\sum_{i,j=1}^{n} \frac{\partial g_{ij}}{\partial q^k}(x(t))v^i(t)v^j(t) + \frac{\partial f}{\partial q^k}(x(t)). \end{aligned}$$

Rearranging terms, and multiplying by the matrix (g^{ij}) which is the inverse of (g_{ij}), we get, for $k = 1, \dots n$, the equation

$$\left(\frac{dv^k}{dt}(t) + \sum_{i,j=1}^{n} \Gamma_{ij}^k(x(t))v^i(t)v^j(t)\right) + \eta v^k(t) + (\text{gradf}(x(t)))^k = 0,$$

where Γ_{ij}^k are the Christoffel symbols given by $\Gamma_{ij}^k = \frac{1}{2}\sum_{l=1}^{n} g^{kl}\left[\frac{\partial g_{jl}}{\partial x^i} + \frac{\partial g_{li}}{\partial x^j} - \frac{\partial g_{ij}}{\partial x^l}\right]$, which gives the desired Euler–Lagrange equation. □

Appendix B. Proof of the Convergence Rates

The proofs of the convergence rates of solutions to the Bregman Euler–Lagrange equations are inspired by those of Theorems 5 and 6 from [3], and make use of Lemmas 2 and 12 therein:

Lemma B.1. *Given a Riemannian manifold $\mathcal{Q}$ with sectional curvature bounded above by $K_{\max}$ and below by $K_{\min}$, with ζ given by equation (2.1), and such that*

$$\text{diam}(\mathcal{Q}) < \begin{cases} \frac{\pi}{\sqrt{K_{\max}}} & \text{if } K_{\max} > 0 \\ \infty & \text{if } K_{\max} \leq 0 \end{cases},$$

we have that

$$\langle \nabla_{\dot{X}} \text{Log}_X(p), -\dot{X} \rangle \leq \zeta \|\dot{X}\|^2.$$

Lemma B.2. *Given a point q and a smooth curve $X(t)$ on a Riemannian manifold $\mathcal{Q}$,*

$$\frac{d}{dt}\|\text{Log}_{X(t)}(q)\|^2 = 2\langle \text{Log}_{X(t)}(q), \nabla_{\dot{X}} \text{Log}_{X(t)}(q) \rangle = 2\langle \text{Log}_{X(t)}(q), -\dot{X}(t) \rangle.$$

Theorem B.1. *Suppose $f : \mathcal{Q} \to \mathbb{R}$ is a λ-weakly-quasi-convex function, and suppose that Assumption 1 is satisfied. Then, any solution $X(t)$ of the Bregman Euler–Lagrange equation*

$$\nabla_{\dot{X}}\dot{X} + \left(\lambda^{-1}\zeta e^{\alpha_t} - \dot{\alpha}_t\right)\dot{X} + e^{2\alpha_t + \beta_t}\text{gradf}(X) = 0,$$

with $X(0) = x_0$ and $\dot{X}(0) = 0$, converges to a minimizer x^* of f with rate

$$f(X(t)) - f(x^*) \leq \frac{2\lambda^2 e^{\beta_0}\left(f(x_0) - f(x^*)\right) + \zeta\|\text{Log}_{x_0}(x^*)\|^2}{2\lambda^2 e^{\beta_t}}.$$

Proof. Let

$$\mathcal{E}(t) = \lambda^2 e^{\beta_t}\left(f(X) - f(x^*)\right) + \frac{1}{2}(\zeta - 1)\|\text{Log}_X(x^*)\|^2 + \frac{1}{2}\left\|\lambda e^{-\alpha_t}\dot{X} - \text{Log}_X(x^*)\right\|^2.$$

Then, using Lemma B.2,

$$\begin{aligned} \dot{\mathcal{E}}(t) &= \lambda^2\dot{\beta}_t e^{\beta_t}\left(f(X) - f(x^*)\right) + \lambda^2 e^{\beta_t}\langle \text{gradf}(X), \dot{X}\rangle + (\zeta - 1)\langle \text{Log}_X(x^*), -\dot{X}\rangle \\ &\quad + \langle \lambda e^{-\alpha_t}\dot{X} - \text{Log}_X(x^*), -\dot{\alpha}_t\lambda e^{-\alpha}\dot{X} + \lambda e^{-\alpha_t}\nabla_{\dot{X}}\dot{X} - \nabla_{\dot{X}}\text{Log}_X(x^*)\rangle \\ &= \lambda^2\dot{\beta}_t e^{\beta_t}\left(f(X) - f(x^*)\right) + \lambda^2 e^{\beta_t}\langle \text{gradf}(X), \dot{X}\rangle + (\zeta - 1)\langle \text{Log}_X(x^*), -\dot{X}\rangle \end{aligned}$$

$$+\left\langle \lambda e^{-\alpha_t}\dot{X} - \mathrm{Log}_X(x^*), \lambda e^{-\alpha_t}\left(-\dot{\alpha}_t\dot{X} + \nabla_{\dot{X}}\dot{X}\right) - \nabla_{\dot{X}}\mathrm{Log}_X(x^*)\right\rangle.$$

Now, from the Bregman Euler–Lagrange equation,

$$-\dot{\alpha}_t\dot{X} + \nabla_{\dot{X}}\dot{X} = -\lambda^{-1}\zeta e^{\alpha_t}\dot{X} - e^{2\alpha_t+\beta_t}\mathrm{gradf}(X).$$

Thus,

$$\begin{aligned}\dot{\mathcal{E}}(t) &= \lambda^2\dot{\beta}_t e^{\beta_t}\left(f(X) - f(x^*)\right) + \lambda^2 e^{\beta_t}\langle \mathrm{gradf}(X), \dot{X}\rangle + (\zeta-1)\langle \mathrm{Log}_X(x^*), -\dot{X}\rangle \\ &\quad + \left\langle \lambda e^{-\alpha_t}\dot{X} - \mathrm{Log}_X(x^*), -\zeta\dot{X} - \lambda e^{\alpha_t+\beta_t}\mathrm{gradf}(X) - \nabla_{\dot{X}}\mathrm{Log}_X(x^*)\right\rangle \\ &= \lambda^2\dot{\beta}_t e^{\beta_t}\left(f(X) - f(x^*)\right) + \lambda^2 e^{\beta_t}\langle \mathrm{gradf}(X), \dot{X}\rangle + (\zeta-1)\langle \mathrm{Log}_X(x^*), -\dot{X}\rangle - \lambda\zeta e^{-\alpha_t}\langle \dot{X}, \dot{X}\rangle \\ &\quad - \lambda^2 e^{\beta_t}\langle \dot{X}, \mathrm{gradf}(X)\rangle - \lambda e^{-\alpha_t}\langle \dot{X}, \nabla_{\dot{X}}\mathrm{Log}_X(x^*)\rangle + \zeta\langle \mathrm{Log}_X(x^*), \dot{X}\rangle \\ &\quad + \lambda e^{\alpha_t+\beta_t}\langle \mathrm{Log}_X(x^*), \mathrm{gradf}(X)\rangle + \langle \mathrm{Log}_X(x^*), \nabla_{\dot{X}}\mathrm{Log}_X(x^*)\rangle.\end{aligned}$$

Canceling the $\langle \mathrm{gradf}(X), \dot{X}\rangle$ and $\langle \mathrm{Log}_X(x^*), -\dot{X}\rangle$ terms out using Lemma B.2, we get

$$\begin{aligned}\dot{\mathcal{E}}(t) &= \lambda^2\dot{\beta}_t e^{\beta_t}\left(f(X) - f(x^*)\right) + \lambda e^{\alpha_t+\beta_t}\langle \mathrm{Log}_X(x^*), \mathrm{gradf}(X)\rangle \\ &\quad - \lambda\zeta e^{-\alpha_t}\langle \dot{X}, \dot{X}\rangle - \lambda e^{-\alpha_t}\langle \dot{X}, \nabla_{\dot{X}}\mathrm{Log}_X(x^*)\rangle \\ &= \lambda e^{\beta_t}\left[\dot{\beta}_t\lambda\left(f(X) - f(x^*)\right) + e^{\alpha_t}\langle \mathrm{Log}_X(x^*), \mathrm{gradf}(X)\rangle\right] \\ &\quad - \lambda e^{-\alpha_t}\left[\zeta\langle \dot{X}, \dot{X}\rangle + \langle \dot{X}, \nabla_{\dot{X}}\mathrm{Log}_X(x^*)\rangle\right].\end{aligned}$$

Now, since f is geodesically λ-weakly-quasi-convex, we have that

$$\lambda\left(f(X) - f(x^*)\right) + \langle \mathrm{Log}_X(x^*), \mathrm{gradf}(X)\rangle \leq 0,$$

so the ideal scaling equation $\dot{\beta}_t \leq e^{\alpha_t}$ implies that

$$\lambda e^{\beta_t}\left[\dot{\beta}_t\lambda\left(f(X) - f(x^*)\right) + e^{\alpha_t}\langle \mathrm{Log}_X(x^*), \mathrm{gradf}(X)\rangle\right] \leq 0.$$

Moreover, Lemma B.1 yields $\left[\zeta\langle \dot{X}, \dot{X}\rangle + \langle \dot{X}, \nabla_{\dot{X}}\mathrm{Log}_X(x^*)\rangle\right] \geq 0$, so

$$-\lambda e^{-\alpha_t}\left[\zeta\langle \dot{X}, \dot{X}\rangle + \langle \dot{X}, \nabla_{\dot{X}}\mathrm{Log}_X(x^*)\rangle\right] \leq 0.$$

Therefore, $\dot{\mathcal{E}}(t) \leq 0$, and so

$$\begin{aligned}\lambda^2 e^{\beta_t}\left(f(X) - f(x^*)\right) &\leq \lambda^2 e^{\beta_t}\left(f(X) - f(x^*)\right) + \frac{1}{2}(\zeta-1)\|\mathrm{Log}_X(x^*)\|^2 + \frac{1}{2}\left\|\lambda e^{-\alpha_t}\dot{X} - \mathrm{Log}_X(x^*)\right\|^2 \\ &= \mathcal{E}(t) \leq \mathcal{E}(0) = \lambda^2 e^{\beta_0}\left(f(x_0) - f(x^*)\right) + \frac{1}{2}\zeta\|\mathrm{Log}_{x_0}(x^*)\|^2,\end{aligned}$$

which gives the desired rate of convergence

$$f(X(t)) - f(x^*) \leq \frac{2\lambda^2 e^{\beta_0}\left(f(x_0) - f(x^*)\right) + \zeta\|\mathrm{Log}_{x_0}(x^*)\|^2}{2\lambda^2 e^{\beta_t}}.$$

□

Appendix C. Proof of Existence Theorems

C.1. Convex and Weakly-Quasi-Convex Cases.

Theorem C.1. *Suppose Assumption 1 is satisfied, and let $C, p > 0$ and $v > 1$ be given constants. Then the differential equation*

$$\nabla_{\dot{X}}\dot{X} + \frac{v}{t}\dot{X} + Ct^{p-2}\mathrm{gradf}(X) = 0,$$

has a global solution $X : [0,\infty) \to \mathcal{Q}$ under the initial conditions $X(0) = x_0 \in \mathcal{Q}$ and $\dot{X}(0) = 0$.

Proof. The proof is similar to that of Lemma 3 in [3], which extended Theorem 1 in [23] to the Riemannian setting. We first define a family of smoothed equations for which we then show existence of a solution for all time. After choosing an equicontinuous and uniformly bounded subfamily of smoothed solutions, we use the Arzela–Ascoli Theorem on the complete Riemannian manifold $\mathcal{Q}$ to obtain a subsequence converging uniformly, and argue that the limit of this subsequence solves the original problem. When $p = 2$, we recover the simpler case considered in Lemma 3 of [3], so we assume $p \neq 2$ in this proof. Consider the following families of smoothed equations for $\delta > 0$:

$$\nabla_{\dot X}\dot X + \frac{v}{\max(\delta,t)}\dot X + C(\max(\delta,t))^{p-2}\mathrm{gradf}(X) = 0 \qquad \text{if } p < 2,$$

$$\nabla_{\dot X}\dot X + \frac{v}{\max(\delta,t)}\dot X + Ct^{p-2}\mathrm{gradf}(X) = 0 \qquad \text{if } p > 2.$$

Exp and Log are defined globally on $\mathcal{Q}$ by Assumption 1, so we can choose geodesically normal coordinates $\phi = \psi^{-1}$ around x_0 defined globally on $\mathcal{Q}$ and put $c = \phi \circ X$. Using the smoothness of f and letting $u = \dot c$ gives a system of first-order ODEs defining a local representation for a vector field in $T\mathcal{Q}$, and Section IV.3 of [12] guarantees that the smoothed ODE has a unique solution X_δ locally around 0. Actually, X_δ exists on $[0,\infty)$. Indeed, by contradiction, let $[0,T)$ be the maximal interval of existence of X_δ, for some finite $T > 0$. Using $\frac{d}{dt}f(X_\delta(t)) = \langle \mathrm{gradf}(X_\delta), \dot X_\delta\rangle$ gives

$$\frac{d}{dt}f(X_\delta) = -\frac{\delta^{2-p}}{C}\langle\nabla_{\dot X_\delta}\dot X_\delta, \dot X_\delta\rangle - \frac{v\delta^{1-p}}{C}\langle\dot X_\delta, \dot X_\delta\rangle = -\frac{\delta^{2-p}}{2C}\frac{d}{dt}\|\dot X_\delta\|^2 - \frac{v\delta^{1-p}}{C}\|\dot X_\delta\|^2 \quad \text{if } \delta > t,\ p < 2,$$

$$\frac{d}{dt}f(X_\delta) = -\frac{t^{2-p}}{C}\langle\nabla_{\dot X_\delta}\dot X_\delta, \dot X_\delta\rangle - \frac{vt^{2-p}}{C\delta}\langle\dot X_\delta, \dot X_\delta\rangle = -\frac{t^{2-p}}{2C}\frac{d}{dt}\|\dot X_\delta\|^2 - \frac{vt^{2-p}}{C\delta}\|\dot X_\delta\|^2 \quad \text{if } \delta > t,\ p > 2,$$

$$\frac{d}{dt}f(X_\delta) = -\frac{t^{2-p}}{C}\langle\nabla_{\dot X_\delta}\dot X_\delta, \dot X_\delta\rangle - \frac{vt^{1-p}}{C}\langle\dot X_\delta, \dot X_\delta\rangle = -\frac{1}{2C}\frac{d}{dt}\left(t^{2-p}\|\dot X_\delta\|^2\right) - \frac{2v(2-p)-1}{2C(2-p)}t^{1-p}\|\dot X_\delta\|^2 \quad \text{if } \delta < t.$$

Let $\theta = \frac{2v(2-p)-1}{2C(2-p)}$. Integrating and using the Cauchy-Schwarz inequality for the $p < 2$ case gives

$$\begin{aligned}
\int_0^T \sqrt{(\max(\delta,t))^{1-p}}\|\dot X_\delta\|dt &= \int_0^\delta \sqrt{\delta^{1-p}}\|\dot X_\delta\|dt + \int_\delta^T \sqrt{t^{1-p}}\|\dot X_\delta\|dt \\
&\leq \sqrt{\frac{C\delta}{v}(f(x_0) - \inf_u f(u)) + \frac{\delta^{2-p}}{2v}\left(\|\dot X_\delta(0)\|^2 - \inf_{t\in[0,T)}\|\dot X_\delta(t)\|^2\right)} \\
&\quad + \sqrt{\frac{T-\delta}{\theta}(f(X_\delta(\delta)) - \inf_u f(u)) + \frac{T-\delta}{2C\theta}\left(\delta^{2-p}\|\dot X_\delta(\delta)\|^2 - \inf_{t\in[0,T)} t^{2-p}\|\dot X_\delta(t)\|^2\right)} < \infty,
\end{aligned}$$

since f is bounded below by Assumption 1. If $\delta \geq T$, then $\sqrt{\delta^{1-p}}\dot X_\delta$ is integrable on $[0,T)$. If $\delta < T$, then the integrals on $[0,T)$ and $[0,\delta)$ are finite, so the integral on $[\delta,T)$ must also be finite, and thus $\sqrt{t^{1-p}}\dot X_\delta$ is integrable on $[\delta,T)$. Now, $\|\int_a^T \dot X_\delta dt\| \leq \int_a^T \|\dot X_\delta\|dt < \infty$ for $a = 0, \delta$ implies that $\lim_{t\to T} X_\delta(t)$ exists. Since $\mathcal{Q}$ is complete by Assumption 1, the limit is in $\mathcal{Q}$, contradicting the maximality of $[0,T)$. The $p > 2$ case is similar: the integrand is replaced by $\sqrt{t^{2-p}(\max(\delta,t))^{-1}}\|\dot X_\delta\|$, and the integral on $[\delta,T)$ remains unchanged while the integral on $[0,\delta)$ can be bounded by the same expression using $t < \delta$. Thus, in both cases, we can find a solution $X_\delta : [0,\infty) \to \mathcal{Q}$ to the smooth initial-value ODE, and its corresponding solution $X_\delta : [0,\infty) \to \mathbb{R}^n$ in local coordinates.

Now let

$$M_\delta(t) = \sup_{u\in(0,t]}\frac{\|\dot X_\delta(u)\|}{u}.$$

When $0 < t \leq \delta$, the smoothed ODE can be written as

$$\nabla_{\dot X_\delta}\left(\dot X_\delta e^{\frac{v}{\delta}}\right) = -C\delta^{p-2}\mathrm{gradf}(X_\delta)e^{\frac{v}{\delta}} \text{ if } p < 2, \quad \nabla_{\dot X_\delta}\left(\dot X_\delta e^{\frac{v}{\delta}}\right) = -Ct^{p-2}\mathrm{gradf}(X_\delta)e^{\frac{v}{\delta}} \text{ if } p > 2.$$

Thus, we can use Lemma 4 in [3] to get for $p > 2$ that

$$\begin{aligned}\Gamma_{X_\delta(t)}^{x_0}\dot{X}_\delta(t) = -e^{-\frac{v}{\delta}t}\int_0^t \left(\Gamma_{X_\delta(u)}^{x_0}\text{gradf}(X_\delta(u)) - \Gamma_{X_\delta(u)}^{x_0}\Gamma(X_\delta)_{x_0}^{X_\delta(u)}\text{gradf}(x_0)\right)Cu^{p-2}e^{\frac{v}{\delta}u}du \\ - e^{-\frac{v}{\delta}t}\int_0^t Cu^{p-2}\Gamma_{X_\delta(u)}^{x_0}\Gamma(X_\delta)_{x_0}^{X_\delta(u)}\text{gradf}(x_0)e^{\frac{v}{\delta}u}du.\end{aligned}$$

From the Lipschitz assumption on f, we have that

$$\left\|\text{gradf}(X_\delta(u)) - \Gamma_{x_0}^{X_\delta(u)}\text{gradf}(x_0)\right\| \leq L\int_0^u \|\dot{X}_\delta(s)\|ds = L\int_0^u s\frac{\|\dot{X}_\delta(s)\|}{s}ds \leq \frac{1}{2}LM_\delta(u)u^2.$$

Thus, since parallel transport preserves inner products,

$$\begin{aligned}\frac{\|\dot{X}_\delta(t)\|}{t} &\leq \left(\frac{1}{2}CLM_\delta(\delta)\delta^p + C\delta^p\|\text{gradf}(x_0)\|\right)\frac{e^{-\frac{v}{\delta}t}}{t}\int_0^t e^{\frac{v}{\delta}u}du \\ &\leq \left(\frac{1}{2}CLM_\delta(\delta)\delta^p + C\delta^p\|\text{gradf}(x_0)\|\right)\frac{\delta}{vt}(1-e^{-\frac{v}{\delta}t}) \leq \frac{1}{2}CLM_\delta(\delta)\delta^p + C\delta^p\|\text{gradf}(x_0)\|.\end{aligned}$$

Taking the supremum over $0 < t \leq \delta$ and rearranging gives for $\delta < \delta_M = \left(\frac{2}{CL}\right)^{\frac{1}{p}}$ that

$$M_\delta(\delta) \leq \frac{2C\delta^p\|\text{gradf}(x_0)\|}{2 - CL\delta^p}.$$

The case $p < 2$ is done exactly in the same way except that we do not need to bound u^{p-2} by δ^{p-2} in the integrals since the t^{p-2} term in the differential equation is already replaced by δ^{p-2}. Note that when $\delta < \delta_M$ and $\delta < t < t_M = \left(\frac{2(v+p+1)}{CL}\right)^{\frac{1}{p}}$, the smoothed ODE can be rewritten as

$$\frac{d}{dt}\left(t^v\dot{X}_\delta(t)\right) = -Ct^{v+p-2}\text{gradf}(X_\delta).$$

Therefore, we can use Lemma 4 in [3] once again to obtain

$$\begin{aligned}\Gamma_{X_\delta(t)}^{X_\delta(\delta)}t^v\dot{X}_\delta(t) - \delta^v\dot{X}_\delta(\delta) = \int_0^t\left(\Gamma_{X_\delta(u)}^{X_\delta(\delta)}\text{gradf}(X_\delta(u)) - \Gamma_{X_\delta(u)}^{X_\delta(\delta)}\Gamma(X_\delta)_{x_0}^{X_\delta(u)}\text{gradf}(x_0)\right)Cu^{v+p-2}du \\ - \int_0^t Cu^{v+p-2}\Gamma_{X_\delta(u)}^{X_\delta(\delta)}\Gamma(X_\delta)_{x_0}^{X_\delta(u)}\text{gradf}(x_0)du.\end{aligned}$$

Using the fact that parallel transport preserves inner products, and dividing by t^{v+1} gives

$$\begin{aligned}\frac{\|\dot{X}_\delta(t)\|}{t} &\leq \frac{\delta^{v+1}}{t^{v+1}}\frac{\|\dot{X}_\delta(\delta)\|}{\delta} + \frac{CL}{2t^{v+1}}\int_\delta^t M_\delta(u)u^{v+p}du + \frac{C}{t^{v+1}}\|\text{gradf}(x_0)\|\int_\delta^t u^{v+p-2}du \\ &\leq \frac{\delta^{v+1}}{t^{v+1}}\frac{2C\delta^p\|\text{gradf}(x_0)\|}{2-CL\delta^p} + \frac{CL}{2(v+p+1)}M_\delta(t)t^p + \frac{C(t^{v+p-1} - \delta^{v+p-1})}{(v+p-1)t^{v+1}}\|\text{gradf}(x_0)\|,\end{aligned}$$

and since this upper bound is an increasing function of t, we have for any $t' \in (\delta, t)$ that

$$\frac{\|\dot{X}_\delta(t')\|}{t'} \leq \frac{2C\delta^p\|\text{gradf}(x_0)\|}{2-CL\delta^p} + \frac{CL}{2(v+p+1)}M_\delta(t)t^p + \frac{Ct^{p-2}}{v+p-1}\|\text{gradf}(x_0)\|.$$

Taking the supremum over all $t' \in (0,t)$ gives for $\delta < \delta_M$ and $\delta < t < t_M$,

$$M_\delta(t) \leq \frac{1}{1 - \frac{CL}{2(v+p+1)}t^p}\left(\frac{2C\delta^p}{2-CL\delta^p} + \frac{Ct^{p-2}}{v+p-1}\right)\|\text{gradf}(x_0)\|.$$

Now consider the family of functions

$$\mathcal{F} = \left\{X_\delta : [0,T] \to \mathbb{R} \middle| \delta = 2^{-n}\tilde{\delta}, n = 0, 1, \dots\right\},$$

where $T = \left(\frac{v+p+1}{CL}\right)^{\frac{1}{p}}$ and $\tilde{\delta} = \left(\frac{1}{CL}\right)^{\frac{1}{p}}$. By definition of M_δ, we have for $t \in [0,T]$ and $\delta \in (0,\tilde{\delta})$ that

$$\|\dot{X}_\delta\| \leq T M_\delta(T) \leq 2CT\left(\tilde{\delta} + \frac{CT^{p-2}}{v+p-1}\right) \quad \text{and} \quad d(X_\delta(t), X_\delta(0)) \leq \int_0^t \|\dot{X}_\delta(u)\| du \leq t\|\dot{X}_\delta\| \leq T\|\dot{X}_\delta\|.$$

Thus, $\mathcal{F}$ is equicontinuous and uniformly bounded, and the Riemannian manifold $\mathcal{Q}$ is complete by Assumption 1, so by the Arzela–Ascoli Theorem (Theorem 17 in [11]), $\mathcal{F}$ contains a subsequence that converges uniformly on $[0,T]$ to some function X^*. The same argument as in part 5 of the proof of Lemma 3 of [3] shows that X^* is a solution to the original initial-value ODE on $[0,T]$ which can then be extended to get a global solution on $[0,\infty)$. □

C.2. **Strongly Convex Case.**

Theorem C.2. *Suppose that Assumption 1 is satisfied, and that $\eta > 0$ is a given constant. Then, the differential equation*

$$\nabla_{\dot{X}}\dot{X} + \eta\dot{X} + \text{gradf}(X) = 0,$$

has a global solution $X : [0,\infty) \to \mathcal{Q}$ under the initial conditions $X(0) = x_0 \in \mathcal{Q}$ and $\dot{X}(0) = 0$.

Proof. Exp and Log are defined globally on $\mathcal{Q}$ by Assumption 1, so we can choose geodesically normal coordinates $\phi = \psi^{-1}$ around x_0 defined globally on $\mathcal{Q}$ and put $c = \phi \circ X$. As in [3], using the smoothness of f and letting $u = \dot{c}$ gives a system of first-order ODEs which defines a local representation for a vector field in $T\mathcal{Q}$, and results from Section IV.3 of [12] guarantee that the initial-value differential equation has a unique solution locally around 0. It remains to show that this solution actually exists on $[0,\infty)$. Towards contradiction, suppose $[0,T)$ is the maximal interval of existence of the solution X, for some finite $T > 0$. Then,

$$\frac{d}{dt}f(X(t)) = \langle \text{gradf}(X), \dot{X}\rangle = -\langle \nabla_{\dot{X}}\dot{X}, \dot{X}\rangle - C\langle \dot{X}, \dot{X}\rangle = -\frac{1}{2}\frac{d}{dt}\|\dot{X}\|^2 - C\|\dot{X}\|^2.$$

Rearranging, integrating both sides and using the Cauchy-Schwarz inequality gives

$$\int_0^T \|\dot{X}\| dt = \sqrt{T(f(x_0) - \inf_u f(u)) + \frac{T}{2}\left(\|\dot{X}(0)\|^2 - \inf_{t\in[0,T)} \|\dot{X}(t)\|^2\right)} < \infty,$$

since f is bounded from below by Assumption 1. Thus, $\lim_{t\to T} X(t)$ exists, and since $\mathcal{Q}$ is complete, the limit is in $\mathcal{Q}$, contradicting the maximality of $[0,T)$, thereby concluding the proof. □

Appendix D. Proof of Invariance Theorem

Theorem D.1. *Suppose that Assumption 1 is satisfied and that the curve $X(t)$ satisfies the Riemannian Bregman Euler–Lagrange equation (3.7) corresponding to $\mathcal{L}_{\alpha,\beta,\gamma}$. Then the reparametrized curve $X(\tau(t))$ satisfies the Bregman Euler–Lagrange equation (3.7) corresponding to the modified Riemannian Bregman Lagrangian $\mathcal{L}_{\tilde{\alpha},\tilde{\beta},\tilde{\gamma}}$ where $\tilde{\alpha}_t = \alpha_{\tau(t)} + \log\dot{\tau}(t)$, $\tilde{\beta}_t = \beta_{\tau(t)}$, and $\tilde{\gamma}_t = \gamma_{\tau(t)}$. Furthermore α,β,γ satisfy the ideal scaling conditions (3.3) if and only if $\tilde{\alpha},\tilde{\beta},\tilde{\gamma}$ do.*

Proof. Let $Y(t) = X(\tau(t))$. Then

$$\dot{Y}(t) = \dot{\tau}(t)\dot{X}(\tau(t)), \qquad \text{and} \qquad \nabla_{\dot{Y}(t)}\dot{Y}(t) = \ddot{\tau}(t)\dot{X}(\tau(t)) + \dot{\tau}^2(t)\nabla_{\dot{X}(\tau(t))}\dot{X}(\tau(t)).$$

Inverting these relations gives

$$\dot{X}(\tau(t)) = \frac{1}{\dot{\tau}(t)}\dot{Y}(t), \qquad \text{and} \qquad \nabla_{\dot{X}(\tau(t))}\dot{X}(\tau(t)) = \frac{1}{\dot{\tau}^2(t)}\nabla_{\dot{Y}(t)}\dot{Y}(t) - \frac{\ddot{\tau}(t)}{\dot{\tau}^3(t)}\dot{Y}(t).$$

The Bregman Euler–Lagrange equation (3.7) at time $\tau(t)$ is given by

$$\nabla_{\dot{X}(\tau(t))}\dot{X}(\tau(t)) + \left(\lambda^{-1}\zeta e^{\alpha_{\tau(t)}} - \dot{\alpha}_{\tau(t)}\right)\dot{X}(\tau(t)) + e^{2\alpha_{\tau(t)}+\beta_{\tau(t)}}\text{gradf}(X(\tau(t))) = 0.$$

Substituting the expressions for $X(\tau(t)), \dot{X}(\tau(t))$ and $\nabla_{\dot{X}(\tau(t))}\dot{X}(\tau(t))$ in terms of $Y(t)$ and its derivatives, and multiplying by $\dot{\tau}^2(t)$, we get

$$\nabla_{\dot{Y}(t)}\dot{Y}(t) - \frac{\ddot{\tau}(t)}{\dot{\tau}(t)}\dot{Y}(t) + \left(\lambda^{-1}\zeta e^{\alpha_{\tau(t)}} - \dot{\alpha}_{\tau(t)}\right)\dot{\tau}(t)\dot{Y}(t) + \dot{\tau}^2(t)e^{2\alpha_{\tau(t)}+\beta_{\tau(t)}}\text{gradf}(Y(t)) = 0.$$

Substituting the expressions for α, β, γ in terms of $\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma}$ yields

$$\nabla_{\dot{Y}(t)}\dot{Y}(t) - \frac{\ddot{\tau}(t)}{\dot{\tau}(t)}\dot{Y}(t) + \left(\lambda^{-1}\zeta\frac{1}{\dot{\tau}(t)}e^{\tilde{\alpha}_t} - \frac{1}{\dot{\tau}(t)}\left[\dot{\tilde{\alpha}}(t) + \frac{\ddot{\tau}(t)}{\dot{\tau}(t)}\right]\right)\dot{\tau}(t)\dot{Y}(t) + e^{2\tilde{\alpha}_t+\tilde{\beta}_t}\text{gradf}(Y(t)) = 0,$$

which gives the Bregman Euler–Lagrange equation (3.7) corresponding to $\mathcal{L}_{\tilde{\alpha},\tilde{\beta},\tilde{\gamma}}$,

$$\nabla_{\dot{Y}(t)}\dot{Y}(t) + \left(\lambda^{-1}\zeta e^{\tilde{\alpha}_t} - \frac{1}{\dot{\tau}(t)}\dot{\tilde{\alpha}}(t)\right)\dot{Y}(t) + e^{2\tilde{\alpha}_t+\tilde{\beta}_t}\text{gradf}(Y(t)) = 0.$$

The fact that α, β, γ satisfy the ideal scaling conditions (3.3) if and only if $\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma}$ do is established in the proof of Theorem 1.2 of [25]. □

References

[1] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*, volume 78. 12 2008.

[2] K. Ahn and S. Sra. From nesterov's estimate sequence to riemannian acceleration, 2020.

[3] F. Alimisis, A. Orvieto, G. Bécigneul, and A. Lucchi. A continuous-time perspective for modeling acceleration in Riemannian optimization. In *Proceedings of the 23rd International AISTATS Conference*, volume 108 of *PMLR*, pages 1297–1307, 2020.

[4] F. Alimisis, A. Orvieto, G. Bécigneul, and A. Lucchi. Practical accelerated optimization on Riemannian manifolds, 2020.

[5] A.-L. Cauchy. Méthode générale pour la résolution des systèmes déquations simultanées. *Acad. Sci. Paris*, 25:536–538, 1847.

[6] V. Duruisseaux and M. Leok. Accelerated optimization on Riemannian manifolds via discrete constrained variational integrators. 2021.

[7] V. Duruisseaux and M. Leok. Accelerated optimization on Riemannian manifolds via projected variational integrators. 2021.

[8] V. Duruisseaux, J. Schmitt, and M. Leok. Adaptive Hamiltonian variational integrators and applications to symplectic accelerated optimization, 2020.

[9] E. Hairer, C. Lubich, and G. Wanner. *Geometric numerical integration*, volume 31 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, second edition, 2006.

[10] J. Jost. *Riemannian geometry and geometric analysis*. Universitext. Springer, Cham, seventh edition, 2017.

[11] J. L. Kelley. *General Topology*. Graduate Texts in Mathematics. Springer New York, 1975.

[12] S. Lang. *Fundamentals of Differential Geometry*, volume 191 of *Graduate Texts in Mathematics*. Springer -Verlag, New York, 1999.

[13] J. M. Lee. *Introduction to Riemannian Manifolds*, volume 176 of *Graduate Texts in Mathematics*. Springer, Cham, second edition, 2018.

[14] M. Leok and T. Ohsawa. Variational and geometric structures of discrete Dirac mechanics. *Found. Comput. Math.*, 11(5):529–562, 2011.

[15] Y. Liu, F. Shang, J. Cheng, H. Cheng, and L. Jiao. Accelerated first-order methods for geodesically convex optimization on Riemannian manifolds. In *Advances in Neural Information Processing Systems*, volume 30, pages 4868–4877. Curran Associates, Inc., 2017.

[16] J. E. Marsden and T. S. Ratiu. *Introduction to mechanics and symmetry*, volume 17 of *Texts in Applied Mathematics*. Springer-Verlag, New York, second edition, 1999.

[17] J. E. Marsden and M. West. Discrete mechanics and variational integrators. *Acta Numer.*, 10:357–514, 2001.

[18] A. S. Nemirovsky and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization.* Wiley - Interscience series in discrete mathematics. Wiley, 1983.

[19] Y. Nesterov. A method of solving a convex programming problem with convergence rate $\mathcal{O}(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

[20] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87 of *Applied Optimization.* Kluwer Academic Publishers, Boston, MA, 2004.

[21] Y. Nesterov. Accelerating the cubic regularization of Newton's method on convex problems. *Math. Program.*, 112:159–181, 2008.

[22] A. Orvieto and A. Lucchi. Shadowing properties of optimization algorithms. In *Advances in Neural Information Processing Systems*, volume 32, pages 12692–12703, 2019.

[23] W. Su, S. Boyd, and E. Candes. A differential equation for modeling Nesterov's Accelerated Gradient method: theory and insights. *Journal of Machine Learning Research*, 17(153):1–43, 2016.

[24] I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, page III1139III1147, 2013.

[25] A. Wibisono, A. Wilson, and M. Jordan. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, 113(47):E7351–E7358, 2016.

[26] H. Zhang and S. Sra. First-order methods for geodesically convex optimization. In *29th Annual Conference on Learning Theory*, pages 1617–1638, 2016.

[27] H. Zhang and S. Sra. Towards riemannian accelerated gradient methods. 2018.

[28] J. Zhang, A. Mokhtari, S. Sra, and A. Jadbabaie. Direct runge-kutta discretization achieves acceleration. 2018.