Accelerated Optimization on Riemannian Manifolds via Projected Variational Integrators

Valentin Duruisseaux and Melvin Leok

Abstract

A variational formulation of accelerated optimization on normed spaces was recently introduced by considering a specific family of time-dependent Bregman Lagrangian and Hamiltonian systems whose corresponding trajectories converge to the minimizer of the given convex function at an arbitrary accelerated rate of $\mathcal{O}(1/t^p)$. This framework has been exploited using time-adaptive geometric integrators to design efficient explicit algorithms for symplectic accelerated optimization. It was observed that geometric discretizations were substantially less prone to stability issues, and were therefore more robust, reliable, and computationally efficient. More recently, this variational framework has been extended to the Riemannian manifold setting by considering a more general family of time-dependent Bregman Lagrangian and Hamiltonian systems on Riemannian manifolds. It is thus natural to develop time-adaptive Hamiltonian variational integrators for accelerated optimization on Riemannian manifolds. In the past, Hamiltonian variational integrators have been constructed with holonomic constraints, but the resulting algorithms were implicit in nature, which significantly increased their cost per iteration. In this paper, we will test the performance of explicit methods based on Hamiltonian variational integrators combined with projections that constrain the numerical solution to remain on the constraint manifold.

1 Introduction

Many data analysis and machine learning algorithms are designed around the minimization of a loss function or the maximization of a likelihood function. Due to the ever-growing size of data sets, there has been a lot of focus on first-order optimization algorithms because of their low cost per iteration. Nesterov's accelerated gradient method [Nesterov, 1983] was shown to converge in $\mathcal{O}(1/k^2)$ to the minimum of the convex objective function f at hand, improving on the $\mathcal{O}(1/k)$ convergence rate exhibited by standard gradient descent methods. This phenomenon in which an algorithm displays this improved rate of convergence is referred to as acceleration. Nesterov's algorithm was shown in [Su *et al.*, 2016] to limit to a second-order ODE as the time-step goes to 0, and $f(x(t))$ converges to its optimal value at a rate of $\mathcal{O}(1/t^2)$ along the trajectories of this ODE. It was later shown that in continuous time, an arbitrary convergence rate $\mathcal{O}(1/t^p)$ can be achieved in normed spaces [Wibisono *et al.*, 2016] and on Riemannian manifolds [Duruisseaux and Leok, 2021b], 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-rescaling property of this family was then exploited in [Duruisseaux *et al.*, 2021] using time-adaptive geometric integrators to design efficient explicit algorithms for accelerated optimization on normed vector spaces. It was observed that a careful use of adaptivity and symplecticity could result in a significant gain in computational efficiency. More generally, when applied to Hamiltonian systems, symplectic integrators yield discrete approximations of the flow that preserve the symplectic two-form [Hairer *et al.*, 2006], and results in the preservation of many qualitative aspects of the underlying dynamical system and, in particular, exhibit excellent long-time near-energy preservation [Reich, 1999; Benettin and Giorgilli, 1994]. Variational integrators provide a systematic method for constructing symplectic integrators of arbitrarily high-order based on the discretization of Hamilton's principle [Marsden and West, 2001].

Recently, there has been some effort to derive accelerated optimization algorithms in the Riemannian manifold setting [Duruisseaux and Leok, 2021a; Duruisseaux and Leok, 2021b; Alimisis *et al.*, 2020; Zhang and Sra, 2016; Zhang and Sra, 2018; Ahn and Sra, 2020; Liu *et al.*, 2017]. The Whitney Embedding Theorems [Whitney, 1944a; Whitney, 1944b] state that any smooth manifold of dimension $n \geq 2$ can be embedded in $\mathbb{R}^{2n}$ and immersed in $\mathbb{R}^{2n-1}$, and is thus diffeomorphic to a submanifold of $\mathbb{R}^{2n}$. Furthermore, the Nash Embedding Theorem [Nash, 1956] states that any Riemannian manifold can be globally isometrically embedded into some Euclidean space. As a consequence, the study of Riemannian manifolds can in principle be reduced to the study of submanifolds of Euclidean spaces. Altogether, this motivates the introduction of time-adaptive variational integrators on Riemannian manifolds that exploit the structure of the embedding Euclidean space. The time-adaptive approach relying on a Poincaré transformation from [Duruisseaux *et al.*, 2021] was extended to the Riemannian manifold setting in [Duruisseaux and Leok, 2021b], and [Duruisseaux and Leok, 2021a] studied how holonomic constraints can be incorporated into variational integrators to constrain the numerical solution to the Riemannian manifold. Although these integrators were carefully justified geometrically as coming from discrete action principles, they were implicit in nature, which significantly increases their cost per iteration as the dimension of the problem becomes large.

In this paper, we present new algorithms based on explicit variational integrators in the embedding space where the manifold constraints are enforced via projections. The resulting explicit algorithms are then used to numerically solve generalized eigenvalue and Procrustes problems on the unit sphere and Stiefel manifold. We believe that these algorithms are the most efficient methods to date which exploit the variational framework from [Duruisseaux and Leok, 2021b].

2 Preliminaries

2.1 Variational Integration

Variational integrators are derived by discretizing Hamilton's principle, instead of discretizing Hamilton's equations. As a result, variational integrators are symplectic, preserve many invariants and momentum maps, and have excellent long-time near-energy preservation [Marsden and West, 2001]. Traditionally, variational integrators have been designed based on the Type I generating function known as the discrete Lagrangian, but more recently, variational integrators have been extended to the framework of Type II/III generating functions, commonly referred to as discrete Hamiltonians [Lall and West, 2006; Leok and Zhang, 2011; Schmitt and Leok, 2017]. The boundary-value formulation of the exact Type II generating function of the time-h flow of Hamilton's equations is given by the exact discrete right Hamiltonian,

$$H_d^{+,E}(q_0,p_h) = p_h q_h - \int_0^h \left[p(t)\dot{q}(t) - H(q(t),p(t))\right] dt,$$

where (q,p) satisfies Hamilton's equations with boundary conditions $q(0) = q_0$ and $p(h) = p_h$. A Type II Hamiltonian variational integrator is constructed by using an approximate discrete Hamiltonian H_d^+, and applying the discrete right Hamilton's equations,

$$p_0 = D_1 H_d^+(q_0,p_1), \qquad q_1 = D_2 H_d^+(q_0,p_1), \qquad (1)$$

which implicitly defines the discrete right Hamiltonian map $\tilde{F}_{H_d^+} : (q_0,p_0) \mapsto (q_1,p_1)$. Theorem 2.2 from [Schmitt and Leok, 2017] states that if a discrete right Hamiltonian H_d^+ approximates the exact discrete right Hamiltonian $H_d^{+,E}$ to order r, then the discrete right Hamiltonian map $\tilde{F}_{H_d^+}$, viewed as a 1-step method, is order r accurate.

2.2 Riemannian Geometry

We first introduce a few main notions from Riemannian geometry. See [Absil *et al.*, 2008; Boumal, 2020; Duruisseaux and Leok, 2021b; Lee, 2018] for more details on Riemannian manifolds and optimization on manifolds.

Definition 1. *Let $\mathcal{Q}$ be a Riemannian manifold with Riemannian metric $g(\cdot,\cdot) = \langle\cdot,\cdot\rangle$. We define the* ***musical isomorphism*** $g^\flat : T\mathcal{Q} \to T^*\mathcal{Q}$ *via* $g^\flat(u)(v) = g_q(u,v)$ *for all* $q \in \mathcal{Q}$ *and* $u,v \in T_q\mathcal{Q}$*, 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* $T^*\mathcal{Q}$ *via*

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

Definition 2. *Denoting the exterior derivative of f by df, the* ***Riemannian gradient*** $\mathrm{grad}f(q) \in T_q\mathcal{Q}$ *at* $q \in \mathcal{Q}$ *of a smooth function* $f : \mathcal{Q} \to \mathbb{R}$ *is the tangent vector at q such that*

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

Definition 3. *A* ***geodesic*** *in a Riemannian manifold $\mathcal{Q}$ is a parametrized curve $\gamma : [0,1] \to \mathcal{Q}$ which is of minimal local length, and is a generalization of the notion of a straight line from Euclidean spaces to Riemannian manifolds.*

Definition 4. *The* ***Riemannian Exponential*** $\mathrm{Exp}_q : T_q\mathcal{Q} \to \mathcal{Q}$ *at* $q \in \mathcal{Q}$ *is defined via* $\mathrm{Exp}_q(v) = \gamma_v(1)$*, 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 its inverse, the* ***Riemannian Logarithm*** $\mathrm{Log}_p : \mathrm{Exp}_q(U) \to T_q\mathcal{Q}$*, is well-defined.*

Definition 5. *A* ***retraction*** *on a manifold $\mathcal{Q}$ is a smooth mapping* $\mathcal{R} : T\mathcal{Q} \to \mathcal{Q}$*, such that for any* $q \in \mathcal{Q}$*, the restriction* $\mathcal{R}_q : T_q\mathcal{Q} \to \mathcal{Q}$ *of* $\mathcal{R}$ *to* $T_q\mathcal{Q}$ *satisfies*

- $\mathcal{R}_q(0_q) = q$*, where* 0_q *denotes the zero element of* $T_q\mathcal{Q}$*,*
- $T_{0_q}\mathcal{R}_q = \mathbb{I}_{T_q\mathcal{Q}}$ *with the canonical identification* $T_{0_q}T_q\mathcal{Q} \simeq T_q\mathcal{Q}$*, where* $T_{0_q}\mathcal{R}_q$ *is the tangent map of* $\mathcal{R}$ *at* $0_q \in T_q\mathcal{Q}$ *and* $\mathbb{I}_{T_q\mathcal{Q}}$ *is the identity map on* $T_q\mathcal{Q}$*.*

The Riemannian Exponential map is a natural example of a retraction on a Riemannian manifold.

Definition 6. *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 a 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,

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

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*

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

Since a geodesically convex function is λ-weakly-quasi-convex with $\lambda = 1$, the algorithms introduced in this paper can also be used in the geodesically convex case. Note that a local minimum of a geodesically convex or λ-weakly-quasi-convex function is also a global minimum.

Definition 7. *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*

$$\zeta = \begin{cases} \sqrt{-K_{\min}}D\coth\left(\sqrt{-K_{\min}}D\right) & \text{if } K_{\min} < 0 \\ 1 & \text{if } K_{\min} \geq 0 \end{cases}. \qquad (2)$$

3 Variational Accelerated Optimization

3.1 Riemannian Bregman Hamiltonian Approach

[Duruisseaux and Leok, 2021b] formulated a variational framework for the minimization of any λ-weakly-quasi-convex function $f : \mathcal{Q} \to \mathbb{R}$, via a p-Bregman Lagrangian $\mathcal{L}_p : T\mathcal{Q} \times \mathbb{R} \to \mathbb{R}$ and a corresponding p-Bregman Hamiltonian $\mathcal{H}_p : T^*\mathcal{Q} \times \mathbb{R} \to \mathbb{R}$ for $p > 0$ of the form

$$\mathcal{L}_p(X, V, t) = \frac{t^{\frac{\zeta}{\lambda}p+1}}{2p}\langle V, V\rangle - Cpt^{\left(\frac{\zeta}{\lambda}+1\right)p-1} f(X), \qquad (3)$$

$$\mathcal{H}_p(X, R, t) = \frac{p}{2t^{\frac{\zeta}{\lambda}p+1}}\langle\langle R, R\rangle\rangle + Cpt^{\left(\frac{\zeta}{\lambda}+1\right)p-1} f(X), \qquad (4)$$

where ζ is given by equation (2). [Duruisseaux and Leok, 2021b] showed that solutions to the p-Bregman Euler–Lagrange equations converge to a minimizer of f at a convergence rate of $\mathcal{O}(1/t^p)$, under suitable assumptions.

Furthermore, [Duruisseaux and Leok, 2021b] proved that time-rescaling the p-Bregman dynamics via $\tau(t) = t^{\mathring{p}/p}$ yields the $\mathring{p}$-Bregman dynamics. 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 corresponding to any value of p. Inspired by the computational efficiency of the approach introduced in [Duruisseaux *et al.*, 2021] on vector spaces, we can 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 allows one to integrate the higher-order p-Bregman dynamics while benefiting from the computational efficiency of integrating the lower-order $\mathring{p}$-Bregman dynamics. Explicitly, it was shown in [Duruisseaux and Leok, 2021b] that the use of the time rescaling $\tau(t) = t^{\mathring{p}/p}$ within the Poincaré transformation framework yields the Direct approach Riemannian p-Bregman Hamiltonian

$$\bar{\mathcal{H}}_p(\bar{Q}, \bar{R}) = \frac{p\langle\langle R, R\rangle\rangle}{2\mathfrak{Q}^{\frac{\zeta}{\lambda}p+1}} + \mathfrak{R} + Cp\mathfrak{Q}^{\left(\frac{\zeta}{\lambda}+1\right)p-1} f(Q),$$

and the Adaptive Riemannian $p \to \mathring{p}$ Bregman Hamiltonian

$$\bar{\mathcal{H}}_{p\to\mathring{p}}(\bar{Q}, \bar{R}) = \frac{p^2}{2\mathring{p}\mathfrak{Q}^{\frac{\zeta}{\lambda}p+\frac{\mathring{p}}{p}}}\langle\langle R, R\rangle\rangle + \frac{p}{\mathring{p}}\mathfrak{Q}^{1-\frac{\mathring{p}}{p}}\mathfrak{R} + \frac{Cp^2}{\mathring{p}}\mathfrak{Q}^{\left(\frac{\zeta}{\lambda}+1\right)p-\frac{\mathring{p}}{p}} f(Q),$$

in the extended phase space defined by $\bar{Q} = \left[\begin{smallmatrix} Q \\ \mathfrak{Q} \end{smallmatrix}\right]$ and $\bar{R} = \left[\begin{smallmatrix} R \\ \mathfrak{R} \end{smallmatrix}\right]$ where $\mathfrak{Q} = t$ and $\mathfrak{R}$ is its conjugate momentum.

On normed vector spaces, these Riemannian Bregman Hamiltonians reduce to the Bregman Hamiltonians derived in [Duruisseaux *et al.*, 2021]. The careful computational study from [Duruisseaux *et al.*, 2021] showed that time-adaptive Hamiltonian variational discretizations, which are automatically symplectic, with adaptive time-steps informed by the time invariance of the family of p-Bregman Hamiltonians yielded the most robust and computationally efficient 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.

[Duruisseaux and Leok, 2021a] incorporated holonomic constraints into variational integrators to constrain the numerical solution to the Riemannian manifold, but the resulting integrators were implicit, which significantly increases their cost. Here, we take a different approach using the fact that the Bregman Hamiltonian in the embedding space restricts to the Riemannian Bregman Hamiltonian on the Riemannian submanifold $\mathcal{Q}$, and the projection of the Bregman Hamiltonian vector field in the embedding space onto the tangent bundle $T\mathcal{Q}$ of the Riemannian submanifold recovers the Hamiltonian vector field of the Riemannian Bregman Hamiltonian. As such, we will numerically integrate the Bregman dynamics in the embedding space and use projections to force the numerical solution to lie on $\mathcal{Q}$. If projections onto the constraint manifold $\mathcal{Q}$ can be computed exactly or approximately very efficiently, we can simply project the updated position onto $\mathcal{Q}$ after every iteration. Furthermore, if projections onto tangent spaces $T_q\mathcal{Q}$ for any point $q \in \mathcal{Q}$ are also available at a low computational cost, it might sometimes be helpful to project the update vector onto $T_q\mathcal{Q}$. Projections have been found for most Riemannian manifolds of practical interest (see [Absil *et al.*, 2008; Boumal, 2020]). These typically involve standard matrix factorizations which can be efficiently computed using iterative methods, and if they are expensive to compute, there are usually ways to accelerate the computations via approximations.

3.2 Riemannian Optimization Problems

Rayleigh Quotient Minimization on the Unit Sphere

Eigenvectors corresponding to the largest eigenvalue of a symmetric $n \times n$ matrix A maximize the Rayleigh quotient $\frac{v^\top Av}{v^\top v}$ over $\mathbb{R}^n$. Thus, a unit eigenvector corresponding to the largest eigenvalue of the matrix A is a minimizer of the function $f(v) = -v^\top Av$, 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 $g_v(u, w) = u^\top w$. A choice of projection from $\mathbb{R}^n$ to $\mathbb{S}^{n-1}$ is the rescaling $v \mapsto \frac{v}{\|v\|_2}$. 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, since the eigenvalues of A are the negative of the eigenvalues of $-A$.

Eigenvalue and Procrustes Problems on St(m, n)

When endowed with the Riemannian metric $g_X(A, B) = \text{Trace}(A^\top B)$, the Stiefel manifold

$$\text{St}(m, n) = \{X \in \mathbb{R}^{n\times m} | X^\top X = I_m\}, \qquad (5)$$

is a Riemannian submanifold of $\mathbb{R}^{n\times m}$. The tangent space at any $X \in \text{St}(m, n)$ is given by

$$T_X\text{St}(m, n) = \{Z \in \mathbb{R}^{n\times m} | X^\top Z + Z^\top X = 0\}, \qquad (6)$$

and the orthogonal projection P_X onto $T_X\text{St}(m,n)$ is given by $P_X Z = Z - \frac{1}{2}X(X^\top Z + Z^\top X)$. We can define a projection of any matrix $\tilde{X} \in \mathbb{R}^{n\times m}$ onto $\text{St}(m,n)$ as the solution of

$$\underset{X\in\text{St}(m,n)}{\operatorname{argmin}} \|X - \tilde{X}\|_F.$$

From [Hairer *et al.*, 2006], the solution of this problem is given by $X = UV^\top$ where $\tilde{X} = U\Sigma V^\top$ is the Singular Value Decomposition of $\tilde{X}$ where Σ is a square diagonal $m\times m$ matrix. The solution X of this problem can also be thought of as the first component of the polar decomposition $\tilde{X} = XS^{1/2}$ where $X \in \text{St}(m,n)$ and S is a $m \times m$ symmetric positive-definite matrix. This solution can be written in closed form as $X = \tilde{X}(\tilde{X}^\top\tilde{X})^{-1/2}$ (and $S = \tilde{X}^\top\tilde{X}$). Thus, a first projection of any given matrix $Q \in \mathbb{R}^{n\times m}$ with Singular Value Decomposition $Q = U\Sigma V^\top$ onto $\text{St}(m,n)$ is given by

$$Q \mapsto Q(Q^\top Q)^{-1/2} \quad \text{or equivalently} \quad Q \mapsto UV^\top.$$

Another method to project a matrix $Y \in \mathbb{R}^{n\times m}$ onto $\text{St}(m,n)$ is obtained via the matrix orthogonalization $Y \mapsto \text{qf}(Y)$, which maps the matrix Y to the Q factor of its QR factorization $Y = QR$ where $Q \in \text{St}(m,n)$ and R is an upper triangular $n \times m$ matrix with strictly positive diagonal elements [Absil *et al.*, 2008].

These polar decomposition and matrix orthogonalization can also be used to construct retractions on $\text{St}(m,n)$:

$$\mathscr{R}_X(\xi) = (X + \xi)(I_m + \xi^\top\xi)^{-1/2}, \quad \mathscr{R}_X(\xi) = \text{qf}(X + \xi).$$

A generalized eigenvector problem consists of finding the m smallest eigenvalues of a $n \times n$ symmetric matrix A and corresponding eigenvectors. This problem can be formulated as a Riemannian optimization problem on the Stiefel manifold $\text{St}(m,n)$ via the Brockett cost function

$$f : \text{St}(m,n) \to \mathbb{R}, \quad f(X) = \text{Trace}(X^\top AXN), \tag{7}$$

where $N = \text{diag}(\mu_1,\ldots,\mu_m)$ for arbitrary $0 \le \mu_1 \le \ldots \le \mu_m$. The columns of a global minimizer of f are eigenvectors corresponding to the m smallest eigenvalues of A [Absil *et al.*, 2008]. If we define $\bar{f} : \mathbb{R}^{n\times m} \to \mathbb{R}$ via $X \mapsto \bar{f}(X) = \text{Trace}(X^\top AXN)$, then $f = \bar{f}|_{\text{St}(m,n)}$ so

$$\text{grad}f(X) = P_X\text{grad}\bar{f}(X) = P_X(2AXN).$$

The unbalanced orthogonal Procrustes problem consists of minimizing the function

$$f : \text{St}(m,n) \to \mathbb{R}, \quad f(X) = \|AX - B\|_F^2, \tag{8}$$

on $\text{St}(m,n)$, for given matrices $A \in \mathbb{R}^{l\times n}$ and $B \in \mathbb{R}^{l\times m}$ with $l \ge n$ and $l > m$, where $\|\cdot\|_F$ is the Frobenius norm $\|X\|_F^2 = \text{Trace}(X^\top X) = \sum_{ij} X_{ij}^2$. If we define $\bar{f} : \mathbb{R}^{n\times m} \to \mathbb{R}$ via $X \mapsto \bar{f}(X) = \|AX - B\|_F^2$, then $f = \bar{f}|_{\text{St}(m,n)}$ so

$$\text{grad}f(X) = P_X\text{grad}\bar{f}(X) = P_X(2A^\top(AX - B)).$$

The special case where $n = m$ is the balanced orthogonal Procrustes problem. In this case, $\text{St}(m,n) = O(n)$ so $\|AX\|_F^2 = \|A\|_F^2$ and thus the minimization problem becomes the problem of maximizing $\text{Trace}(X^\top A^\top B)$ over $X \in O(n)$. In this special case, a solution is given by $X^* = UV^\top$ where $B^\top A = U\Sigma V^\top$ where U and V are orthogonal square matrices is the Singular Value Decomposition of $B^\top A$, and the solution is unique provided that the matrix $B^\top A$ is nonsingular [Eldén and Park, 1999; Golub and Van Loan, 2013].

3.3 Numerical Methods

Euler–Lagrange Simple Discretization

In [Duruisseaux and Leok, 2021b], the p-Bregman Euler–Lagrange equations were rewritten as a first order system of differential equations, for which a Riemannian version of a semi-implicit Euler scheme was applied to obtain the iterates presented in Algorithm 1, when given a λ-weakly-quasi-convex function $f : \mathcal{Q} \to \mathbb{R}$, a retraction $\mathscr{R}$ from $T\mathcal{Q}$ to $\mathcal{Q}$, constants $C, h, p > 0$, and $X_0 \in \mathcal{Q}$, $V_0 \in T_{X_0}\mathcal{Q}$.

Algorithm 1 Euler–Lagrange Discretization

```
$b_k \leftarrow 1 - \frac{\zeta p+\lambda}{\lambda k}, \quad c_k \leftarrow Cp^2(kh)^{p-2}$
Version I: $a_k \leftarrow b_k V_k - hc_k \,\text{grad}f(X_k)$
Version II: $a_k \leftarrow b_k V_k - hc_k \,\text{grad}f(\mathscr{R}_{X_k}(hb_kV_k))$
$X_{k+1} \leftarrow \mathscr{R}_{X_k}(ha_k)$   and   $V_{k+1} \leftarrow \Gamma_{X_k}^{X_{k+1}} a_k$
```

Version I of Algorithm 1 corresponds to the update with the traditional gradient $\nabla f(X_k)$ for the semi-implicit Euler scheme, while Version II is inspired by the reformulation of Nesterov's method from [Sutskever *et al.*, 2013] that uses a corrected gradient $\nabla f(X_k + hb_kV_k)$ instead of $\nabla f(X_k)$.

Hamiltonian Taylor Variational Integrators

Hamiltonian Taylor variational integrators (HTVIs) form a class of variational integrators described in [Schmitt *et al.*, 2018]. A discrete approximate Hamiltonian is constructed by approximating the flow map and the trajectory associated with the boundary values using a Taylor method, and approximating the integral by a quadrature rule. The variational integrator is then generated by the discrete Hamilton's equations. We will use projected versions of the HTVIs constructed in [Duruisseaux *et al.*, 2021]. Given a function $f : \mathcal{Q} \to \mathbb{R}$, a projection $\mathscr{P}_{\mathcal{Q}}$ onto $\mathcal{Q}$, $(q_0, r_0) \in T^*_{q_0}\mathcal{Q}$, and constants $C, h, p, \mathring{p}, \mathfrak{q}_0 > 0$, the Direct and Adaptive HTVIs are obtained by iterating the updates given in Algorithm 2.

Algorithm 2 Direct and Adaptive HTVIs

Adaptive HTVI

$$r_{k+1} = r_k - \frac{p^2}{\mathring{p}}hC\mathfrak{q}_k^{2p-\mathring{p}/p}\text{grad}f(q_k),$$

$$q_{k+1} = \mathscr{P}_{\mathcal{Q}}\left(q_k + \frac{p^2}{\mathring{p}}h\mathfrak{q}_k^{-p-\mathring{p}/p}r_{k+1}\right),$$

$$\mathfrak{q}_{k+1} = \mathfrak{q}_k + \frac{p}{\mathring{p}}h\mathfrak{q}_k^{1-\mathring{p}/p}.$$

Direct HTVI is the special case where $\mathring{p} = p$.

Riemannian Gradient Descent

Riemannian Gradient Descent is a generalization of Gradient Descent to Riemannian manifolds which involves the Riemannian gradient and a retraction. Given a function $f : \mathcal{Q} \to \mathbb{R}$ with a retraction $\mathscr{R}$ from $T\mathcal{Q}$ to $\mathcal{Q}$, $h > 0$, and $X_0 \in \mathcal{Q}$, the Riemannian Gradient Descent iterations are given by

$$X_{k+1} = \mathscr{R}_{X_k}(-h \,\text{grad}f(X_k)). \tag{9}$$

3.4 Numerical Results

Comparison of the Adaptive and Direct approaches

Numerical experiments were conducted for the Rayleigh quotient minimization problem on $\mathbb{S}^{n-1}$ with the projection based method. As was observed in [Duruisseaux *et al.*, 2021] on vector spaces, Figure 1 shows that the Adaptive approach can be significantly more efficient than the Direct approach, and that both methods enjoy faster convergence as p increases.

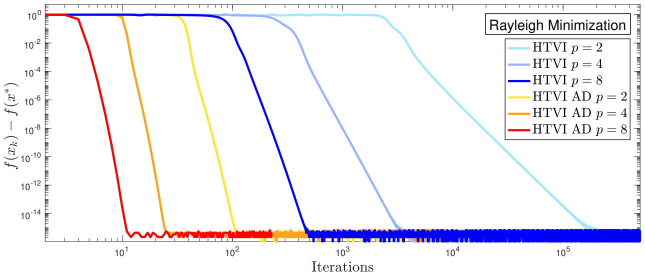

Figure 1: Comparison of the Direct and Adaptive (AD) projection based HTVIs with different values of the parameter p and the same time-step $h = 0.01$, for the Rayleigh minimization problem on $\mathbb{S}^{n-1}$.

Rayleigh minimization problem on the unit sphere $\mathbb{S}^{n-1}$

It was noted in [Duruisseaux and Leok, 2021b] that although higher values of p in Algorithm 1 result in provably faster rates of convergence, they also appear to be more prone to stability issues under numerical discretization, which can cause the numerical optimization algorithm to diverge. Numerical experiments in [Duruisseaux *et al.*, 2021] showed that on normed vector spaces, geometric discretizations which respect the time-rescaling invariance and symplecticity of the Bregman Hamiltonian flows were substantially less prone to these stability issues, and were therefore more robust, reliable, and computationally efficient. This was one of the motivations to develop time-adaptive variational integrators for the Bregman Hamiltonians. Numerical experiments were conducted for the Rayleigh quotient minimization problem on $\mathbb{S}^{n-1}$, and the results are presented in Figure 2.

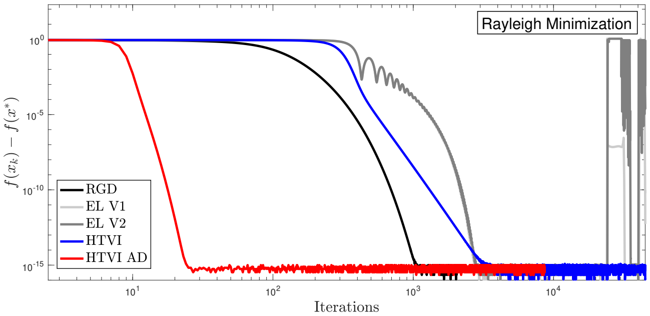

Figure 2: Comparison of the Direct and Adaptive (AD) projection based HTVIs with the Riemannian Gradient Descent (RGD) method and the Euler–Lagrange discretizations (EL V1 and EL V2), with $p = 4$ and the same time-step h.

The Adaptive HTVI clearly outperforms the other algorithms for this problem. Note that the Euler–Lagrange discretizations suffer from stability issues leading to a loss of convergence (after $\approx 10^4$ iterations), due to the polynomially growing unbounded term $Cp^2(kh)^{p-2}$ paired with the grad$f$ term to 0 which can only achieve a finite order of accuracy due to numerical roundoff error. This issue can be fixed by adding a suitable upper bound to the coefficient $Cp^2(kh)^{p-2}$ in the updates, or by stopping the iterating process once a desired convergence criterion is achieved.

Optimization Problems on the Stiefel manifold St(m, n)

Numerical experiments were conducted for the generalized eigenvalue and Procustes problems on St(m, n) to observe how the projection based HTVIs compare to the Euler–Lagrange discretizations from [Duruisseaux and Leok, 2021b] and the standard Riemannian gradient descent. The results are presented in Figure 3. Note that for certain instances of the Procrustes problem with certain initial values $X_0 \in$ St(m, n), all the algorithms converged to a local minimizer which was not a global minimizer.

The projection based Adaptive HTVIs clearly outperform their Direct approach counterparts, Riemannian gradient descent and both versions of the Euler–Lagrange discretization in terms of number of iterations required, when all the algorithms are implemented with the same time step (see the two bottom plots in Figure 3). As can be seen from the top two plots in Figure 3, the Adaptive HTVIs are still the best performing algorithms, even when larger time steps are taken for the other algorithms and in particular even when the Riemannian gradient descent algorithm has been tuned optimally. Note that both the Euler–Lagrange discretizations and Hamiltonian variational integrators suffered from the numerical roundoff issue described in the previous subsection, but this issue was resolved by adding a suitable upper bound to the ever-growing problematic coefficient in the updates.

Our numerical experiments do not suggest that there is a clear benefit in using the polar decomposition based projection over the matrix orthogonalization, or vice versa. Both projection strategies led to very efficient algorithms for Riemannian accelerated optimization with seemingly similar performance and stability properties. Computing the QR decomposition of a $n \times m$ matrix via the standard Householder QR algorithm requires approximately $2m^2(n - m/3)$ floating point operations, while computing the Singular Value Decomposition of a $n \times m$ is more expensive and often relies on intermediate QR decompositions [Trefethen and Bau, 1997]. Thus, these operations can become very costly as the dimension of the problem becomes large, in which case it might be beneficial to use approximate QR decompositions and Singular Value Decompositions. For instance, the projection based on the polar decomposition $Q \mapsto Q(Q^\top Q)^{-1/2}$ can be rewritten as $Q \mapsto Q(I_m + (Q^\top Q - I_m))^{-1/2}$, and provided the distance away from the Stiefel manifold is sufficiently small, the norm of $E = (Q^\top Q - I_m)$ is small and we can approximate the projection by truncating its series expansion

$$Q(I_m + D)^{-1/2} = Q\left(I_m - \frac{1}{2}D + \frac{3}{8}D^2 - \frac{5}{16}D^3 + \dots\right).$$

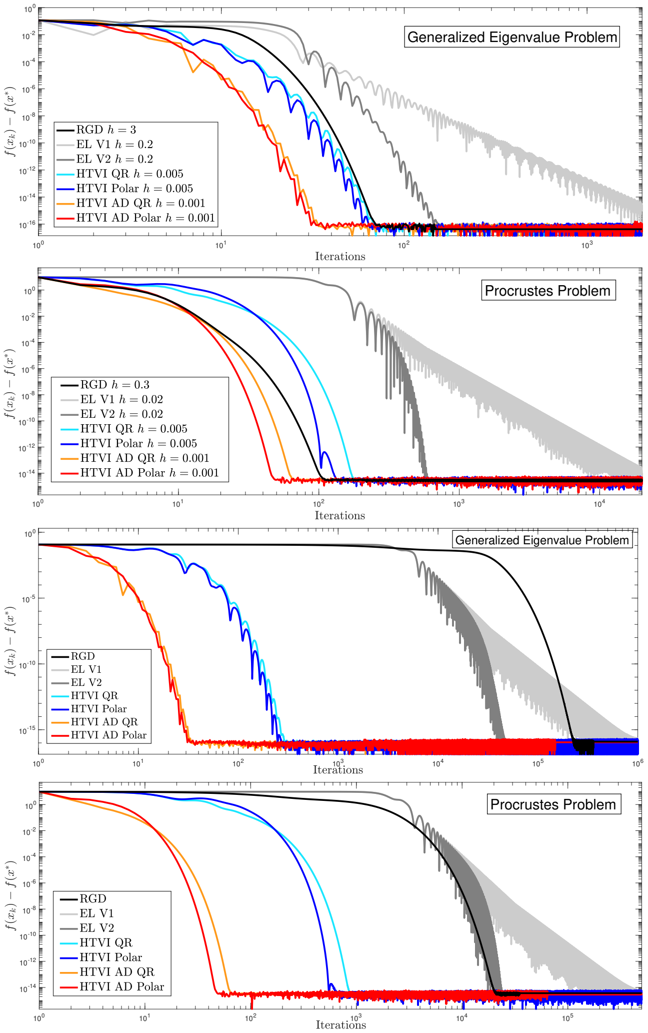

Figure 3: Comparison of the Direct and Adaptive (AD) Type II HTVIs with the Riemannian Gradient Descent (RGD) method and the Euler–Lagrange discretizations (EL V1 and EL V2) with $p = 5$ with different time steps (top two plots) and with the same time-step $h = 0.001$ (bottom two plots), for the generalized eigenvalue and Procrustes problems on $\mathrm{St}(m, n)$.

We also tested the projection algorithm against the implicit algorithm from [Duruisseaux and Leok, 2021a] on the same optimization problems on $\mathbb{S}^{n-1}$ and $\mathrm{St}(m, n)$ Although both algorithms produced very similar graphs for the error as a function of the iteration number, the explicit nature of our projection algorithm made every iteration significantly faster and overall the running time was reduced by several orders of magnitude, even on low-dimensional problems (for instance, 3 orders of magnitude on $\mathbb{S}^{5-1}$ and $\mathrm{St}(3, 2)$, and 4 orders of magnitude on $\mathbb{S}^{100-1}$). Note that the projection algorithm was also easier to implement and tune than the implicit algorithm.

4 Conclusion

Motivated by the observation made in the normed space setting in [Duruisseaux *et al.*, 2021] that a careful use of adaptivity and symplecticity within the variational formulation of accelerated optimization could result in a significant gain in computational efficiency, discrete variational integrators incorporating holonomic constraints were constructed in [Duruisseaux and Leok, 2021a] within the variational framework for Riemannian accelerated optimization of [Duruisseaux and Leok, 2021b]. The resulting algorithms performed well in terms of number of iterations required to achieve convergence but were implicit which can lead to high computational costs. In this paper, we saw that the gain in computational efficiency is preserved when the constraints are enforced via projections instead of being incorporated directly into the variational principles, and that the explicit nature of the resulting algorithms makes every iteration significantly faster and easier to tune than for the implicit algorithms from [Duruisseaux and Leok, 2021a]. As a consequence, if projections onto the constraint manifold can be computed efficiently, these projection based variational integrators form a class of efficient explicit algorithms for Riemannian accelerated optimization, and we believe that these algorithms are the most efficient methods to date which exploit the variational framework from [Duruisseaux and Leok, 2021b].

Although the Whitney and Nash Embedding Theorems imply that there is no loss of generality when studying Riemannian manifolds only as submanifolds of Euclidean spaces, designing intrinsic methods that would exploit and preserve the symmetries and geometric properties of the Riemannian manifold and of the problem at hand could have advantages both in terms of computation and in terms of improving our understanding of the acceleration phenomenon on Riemannian manifolds. Developing an intrinsic extension of Hamiltonian variational integrators to manifolds would require some additional work, since the current approach involves Type II/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. This does not make intrinsic sense on a manifold, since one needs the base point in order to specify the corresponding cotangent space, and instead one should ideally consider a construction based on the more general discrete Dirac mechanics [Leok and Ohsawa, 2011].

It would be desirable to have some convergence guarantees, but proving that the discrete time algorithms perform analogously to the continuous dynamics is far from direct, as the $\mathcal{O}(1/t^p)$ convergence rate for the continuous-time dynamics conflicts with the $\mathcal{O}(1/k^2)$ Nesterov barrier theorem for discrete-time algorithms. Some shadowing results can be obtained for certain Riemannian optimization algorithms when the associated dynamical system is uniformly contracting. However, momentum methods such as the ones presented here, are notoriously non-descending and heavily oscillatory, and lack contraction as a result. Even on vector spaces, obtaining theoretical guarantees was a challenging task, achieved in [Zhang *et al.*, 2018] under additional assumptions. Generalizing these results to Riemannian manifolds would be much more challenging than a trivial generalization of these results since vector space operations and objects have to be replaced by their Riemannian generalizations which involve geodesics, parallel transport, covariant derivatives, Riemannian exponentials and logarithms.

References

[Absil *et al.*, 2008] P. A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2008.

[Ahn and Sra, 2020] K. Ahn and S. Sra. From Nesterov's estimate sequence to Riemannian acceleration. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 84–118. PMLR, 09–12 Jul 2020.

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

[Benettin and Giorgilli, 1994] G. Benettin and A. Giorgilli. On the Hamiltonian interpolation of near-to-the identity symplectic mappings with application to symplectic integration algorithms. *Journal of Statistical Physics*, 74:1117–1143, 03 1994.

[Boumal, 2020] N. Boumal. An introduction to optimization on smooth manifolds, 2020. Available online at http://www.nicolasboumal.net/book.

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

[Duruisseaux and Leok, 2021b] V. Duruisseaux and M. Leok. A variational formulation of accelerated optimization on Riemannian manifolds. 2021.

[Duruisseaux *et al.*, 2021] V. Duruisseaux, J. Schmitt, and M. Leok. Adaptive Hamiltonian variational integrators and applications to symplectic accelerated optimization. *SIAM Journal on Scientific Computing*, 43(4):A2949–A2980, 2021.

[Eldén and Park, 1999] L. Eldén and H. Park. A Procrustes problem on the Stiefel manifold. *Numerische Mathematik*, 82(4):599–619, 1999.

[Golub and Van Loan, 2013] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, 2013.

[Hairer *et al.*, 2006] E. Hairer, C. Lubich, and G. Wanner. *Geometric Numerical Integration*, volume 31 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, second edition, 2006.

[Lall and West, 2006] S. Lall and M. West. Discrete variational Hamiltonian mechanics. *J. Phys. A*, 39(19):5509–5519, 2006.

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

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

[Leok and Zhang, 2011] M. Leok and J. Zhang. Discrete Hamiltonian variational integrators. *IMA Journal of Numerical Analysis*, 31(4):1497–1532, 2011.

[Liu *et al.*, 2017] Y. Liu, F. Shang, J. Cheng, H. Cheng, and L. Jiao. Accelerated first-order methods for geodesically convex optimization on Riemannian manifolds. In *NeurIPS*, volume 30, pages 4868–4877, 2017.

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

[Nash, 1956] J. Nash. The imbedding problem for Riemannian manifolds. *Annals of Mathematics*, 63(1):20–63, 1956.

[Nesterov, 1983] 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.

[Reich, 1999] S. Reich. Backward error analysis for numerical integrators. *SIAM J. Numer. Anal.*, 36:1549–1570, 1999.

[Schmitt and Leok, 2017] J. M. Schmitt and M. Leok. Properties of Hamiltonian variational integrators. *IMA Journal of Numerical Analysis*, 38(1):377–398, 03 2017.

[Schmitt *et al.*, 2018] J. M. Schmitt, T. Shingel, and M. Leok. Lagrangian and Hamiltonian Taylor variational integrators. *BIT Numerical Mathematics*, 58:457–488, 2018.

[Su *et al.*, 2016] 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.

[Sutskever *et al.*, 2013] 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, pages 1139–1147, Atlanta, GA, USA, 2013.

[Trefethen and Bau, 1997] L.N. Trefethen and D. Bau. *Numerical Linear Algebra*. Other Titles in Applied Mathematics. SIAM, 1997.

[Whitney, 1944a] H. Whitney. The self-intersections of a smooth n-manifold in $2n$-space. *Annals of Mathematics*, 45(2):220–246, 1944.

[Whitney, 1944b] H. Whitney. The singularities of a smooth n-manifold in $(2n - 1)$-space. *Annals of Mathematics*, 45(2):247–293, 1944.

[Wibisono *et al.*, 2016] 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.

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

[Zhang and Sra, 2018] H. Zhang and S. Sra. An estimate sequence for geodesically convex optimization. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1703–1723, Jul 2018.

[Zhang *et al.*, 2018] J. Zhang, A. Mokhtari, S. Sra, and A. Jadbabaie. Direct Runge-Kutta discretization achieves acceleration. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.