Variational Accelerated Optimization on Riemannian Manifolds

Melvin Leok

Mathematics, University of California, San Diego

Joint work with Valentin Duruisseaux, Jeremy Schmitt

Symmetry, Invariants, and their Applications
(A Celebration of Peter Olver's 70th Birthday)
Dalhousie University, Halifax, Nova Scotia, Canada, August 2022

mleok@ucsd.edu
http://www.math.ucsd.edu/~mleok/

NSF DMS-1813635 • CCF-2112665 • DMS-2217293
AFOSR FA9550-18-1-0288 • DoD Newton Award
Simons Fellowship in Mathematics

Congratulations to Peter Olver

On the occasion of his 70th birthday
and his contributions to useful and beautiful mathematics

Accelerated Optimization

Accelerated Optimization

- Consider the **optimization problem** $\min_{x \in X} f(x)$, where $X \subset \mathbb{R}^n$ is a convex domain and $f : X \to \mathbb{R}$ is a continuously differentiable convex function, with a unique minimizer $x^* \in X$.
- Nesterov[1] introduced the **accelerated gradient method**,

$$x_k = y_{k-1} - s\nabla f(y_{k-1}), \qquad y_k = x_k + \frac{k-1}{k+2}(x_k - x_{k-1}),$$

which for any fixed step size $s \leq 1/L$, where L is the Lipschitz constant of ∇f, converges at

$$f(x_k) - f(x^*) \leq O\left(\frac{\|x_0 - x^*\|^2}{sk^2}\right).$$

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

Accelerated Optimization

Accelerated Optimization

- Nesterov[2] proved that $O(1/k^2)$ is optimal for methods that only use information about the gradient of f at consecutive iterates, and vanilla gradient descent methods only achieve $O(1/k)$.
- Su, Boyd, and Candès[3] proved that the Nesterov scheme has a continuous limit given by,

$$X'' + \frac{3}{t}X' + \nabla f(X) = 0,$$

with initial conditions $X(0) = x_0$, $X'(0) = 0$. The time parameter in the ODE is related to the step size s via $t \approx k\sqrt{s}$.

[2] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87 of Applied Optimization. Kluwer Academic Publishers, Boston, MA, 2004.

[3] W. Su, S. Boyd, E. J. Candès, *A Differential Equation for modeling Nesterov's Accelerated Gradient Method: Theory and Insights*, Journal of Machine Learning Research, 17(153), 1–43, 2016.

Variational Accelerated Optimization

Bregman Lagrangian[4]

- The **Bregman Lagrangian** is given by,

$$L(x,v,t) = e^{\alpha(t)+\gamma(t)}(\mathcal{D}_h(x + e^{-\alpha(t)}v, x) - e^{\beta(t)} f(x)),$$

and the **Bregman Hamiltonian** is given by

$$H(x,p,t) = e^{\alpha(t)+\gamma(t)}(\mathcal{D}_{h^*}(\nabla h(x)+e^{-\gamma(t)}p, \nabla h(x))+e^{\beta(t)} f(x)),$$

where $h : X \to \mathbb{R}$ is convex, h^* is its convex dual, and D_h is the **Bregman divergence**,

$$\mathcal{D}_h(y,x) = h(y) - h(x) - \langle \nabla h(x), y - x\rangle.$$

[4] A. Wibisono, A. Wilson, M. I. Jordan, *A variational perspective on accelerated methods in optimization*, Proceedings of the National Academy of Sciences, 133, E7351–E7358, 2016.

Variational Accelerated Optimization

Convergence rates of Euler–Lagrange flow

- Under the growth conditions $\dot{\beta} \leq e^{\alpha}$, $\dot{\gamma} = e^{\alpha}$, the solutions of the associated Euler–Lagrange equations exhibit the following convergence property,

$$f(x(t)) - f(x^*) \leq \mathcal{O}(e^{-\beta(t)}),$$

  which was shown using a Lyapunov function approach.

- In particular, for $p > 0$, if we choose $\alpha(t) = \log p - \log t$, $\beta(t) = p \log t + \log C$, $\gamma(t) = p \log t$, where $C > 0$, then the growth condition above is satisfied, and the Euler–Lagrange flow converges to the optimal value in $\mathcal{O}(1/t^p)$.

Discrete Mechanics and Accelerated Optimization

■ Variational Discretization[5]

- Due to the scaling terms $\alpha(t)$, $\beta(t)$, $\gamma(t)$ in the Bregman Lagrangian and Hamiltonian, they are time-dependent.
- Naïvely applying symplectic and variational integrators to such time-dependent systems often yield poor numerical results.
- Given a Hamiltonian, $H(q,p)$, and a desired transformation of time, $t \mapsto \tau$, given by $\frac{dt}{d\tau} = g(q,p)$, a new Hamiltonian system is given by the **Poincaré transformation**[6],

$$\bar{H}(\bar{q},\bar{p}) = g(q,p)\left(H(q,p) + p^t\right),$$

where $(\bar{q},\bar{p}) = \left(\begin{bmatrix} q \\ q^t \end{bmatrix}, \begin{bmatrix} p \\ p^t \end{bmatrix}\right)$.

[5] V. Duruisseaux, J. Schmitt, M. Leok, *Adaptive Hamiltonian Variational Integrators and Symplectic Accelerated Optimization*, SIAM Journal of Scientific Computing, 43(4), A2949-A2980 (32 pages), 2021.

[6] E. Hairer, *Variable time step integration with symplectic methods*, Applied Numerical Mathematics, 25, 219–227, 1997.

Discrete Mechanics and Accelerated Optimization

Variational Discretization

- We let $q^t = t$ and $p^t = -H(q(0), p(0))$. Then, $\bar{H}(\bar{q}, \bar{p}) = 0$ along all integral curves through $(q(0), p(0))$.
- The corresponding Hamilton's equations are given by,

$$\dot{\bar{q}} = \begin{bmatrix} \nabla_p g(q,p) \\ 0 \end{bmatrix} (H(q,p) + p^t) + \begin{bmatrix} \frac{\partial H}{\partial p} \\ 1 \end{bmatrix} g(q,p),$$

$$\dot{\bar{p}} = -\begin{bmatrix} \nabla_q g(q,p) \\ 0 \end{bmatrix} (H(q,p) + p^t) - \begin{bmatrix} \frac{\partial H}{\partial q} \\ 0 \end{bmatrix} g(q,p).$$

- With initial conditions $(q(0), p(0))$, $H(q,p) + p^t = 0$ and

$$\dot{\bar{q}} = \begin{bmatrix} g(q,p)\frac{\partial H}{\partial p} \\ g(q,p) \end{bmatrix}, \qquad \dot{\bar{p}} = \begin{bmatrix} -g(q,p)\frac{\partial H}{\partial q} \\ 0 \end{bmatrix}.$$

Discrete Mechanics and Accelerated Optimization

Variational Discretization

- In general, we obtain a degenerate Hamiltonian, since
$$\frac{\partial^2 \bar{H}}{\partial \bar{p}^2} = \begin{bmatrix} \frac{\partial H}{\partial p}\nabla_p g(q,p)^T + g(q,p)\frac{\partial^2 H}{\partial p^2} + \nabla_p g(q,p)\frac{\partial H}{\partial p}^T & \nabla_p g(q,p) \\ \nabla_p g(q,p)^T & 0 \end{bmatrix}.$$
- Going to extended phase space, corresponds to $g \equiv 1$, which yields a degenerate Hamiltonian for which no Lagrangian analogue exists.
- This necessitates the use of Hamiltonian variational integrators. The variational error analysis result holds so long as
$$\det\left(\frac{\partial H}{\partial p}\nabla_p g(q,p)^T + g(q,p)\frac{\partial^2 H}{\partial p^2} + \nabla_p g(q,p)\frac{\partial H}{\partial p}^T\right) \neq 0.$$
In particular, this holds for non-degenerate Hamiltonians H, and p-independent monitor functions.

Exact Discrete Hamiltonian

Sketch of Approach[7]

- The exact discrete Lagrangian is a Type I generating function,

$$L_d^{\text{exact}}(q_0, q_1) \equiv \operatorname*{ext}_{\substack{q \in C^2([0,h],Q) \\ q(0)=q_0, q(h)=q_1}} \int_0^h L(q(t), \dot{q}(t))dt,$$

  expressed in terms of a continuous Lagrangian.

- Use the continuous Legendre transformation to obtain,

$$L(q, \dot{q}) = p\dot{q} - H(q, p).$$

[7] ML, J. Zhang, *Discrete Hamiltonian Variational Integrators*, IMA Journal of Numerical Analysis, 31(4), 1497–1532, 2011.

Exact Discrete Hamiltonian

Sketch of Approach

- Use the discrete Legendre transformation,

$$\begin{array}{ccc} L_d(q_k, q_{k+1}) & \longrightarrow & H_d^+(q_k, p_{k+1}) \\ \downarrow & & \downarrow \\ H_d^-(p_k, q_{k+1}) & \longrightarrow & R_d(p_k, p_{k+1}) \end{array}$$

  to obtain a Type II generating function,

$$H^+_{d,\text{exact}}(q_k, p_{k+1}) = \underset{\substack{(q,p)\in C^2([t_k,t_{k+1}],T^*Q) \\ q(t_k)=q_k, p(t_{k+1})=p_{k+1}}}{\text{ext}} p(t_{k+1})q(t_{k+1}) - \int_{t_k}^{t_{k+1}} [p\dot{q} - H(q,p)]\, dt.$$

- Discretize using an approximation space for Q (not T^*Q) and a quadrature rule. Equivalent to Lagrangian variational integrator.

Accelerated Optimization using Adaptive Variational Integrators

Transformations of the Bregman Hamiltonian

- Use adaptivity to transform the time-dependent Bregman Hamiltonian corresponding to $p > 0$ into an autonomous Hamiltonian corresponding to a smaller $\mathring{p} < p$ in extended phase-space.
- Integrate higher-order p-Bregman dynamics with the computational efficiency of integrating lower-order $\mathring{p}$-Bregman dynamics.
- The desired monitor function is given by

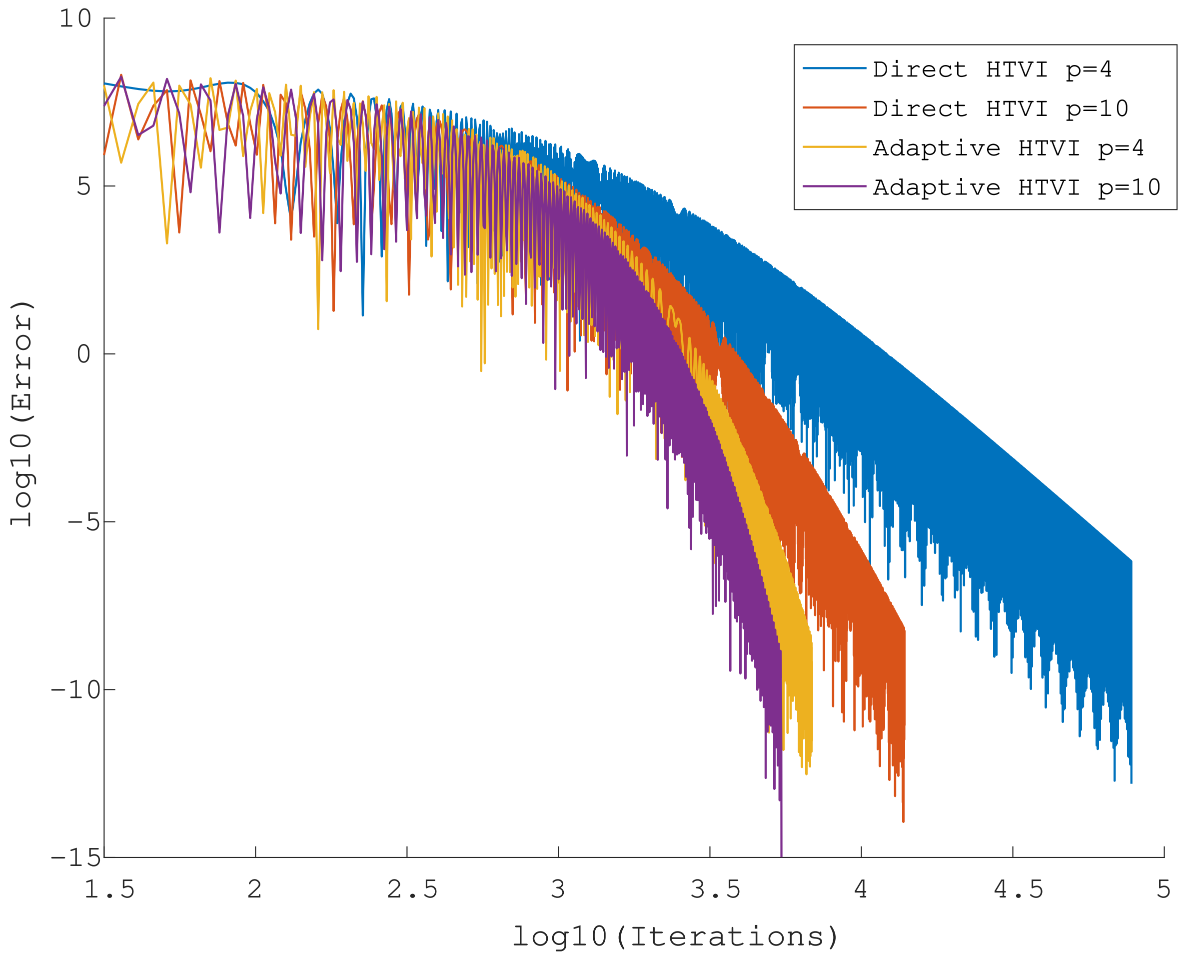
$$\frac{dt}{d\tau} = g(q,t,r) = \frac{p}{\mathring{p}} t^{1-\mathring{p}/p},$$

and generates the Poincaré transformed Hamiltonian

$$\bar{H}(\bar{q},\bar{r}) = \frac{1}{\mathring{p}} \left[\frac{p^2}{2(q^t)^{p+\frac{\mathring{p}}{p}}} \langle r, r \rangle + Cp^2 (q^t)^{2p-\frac{\mathring{p}}{p}} f(q) + p r^t (q^t)^{1-\frac{\mathring{p}}{p}} \right].$$

Accelerated Optimization using Adaptive Variational Integrators

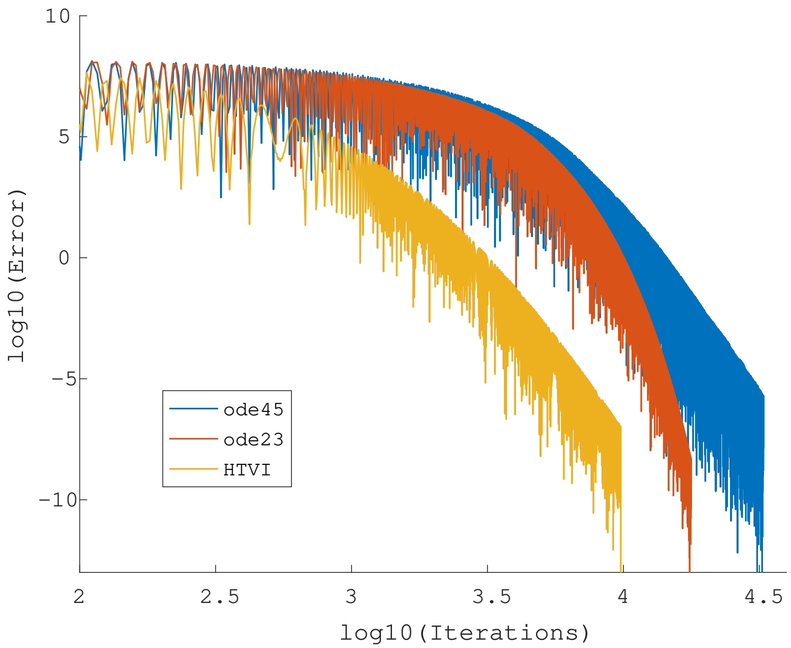
■ Adaptive versus Direct approach

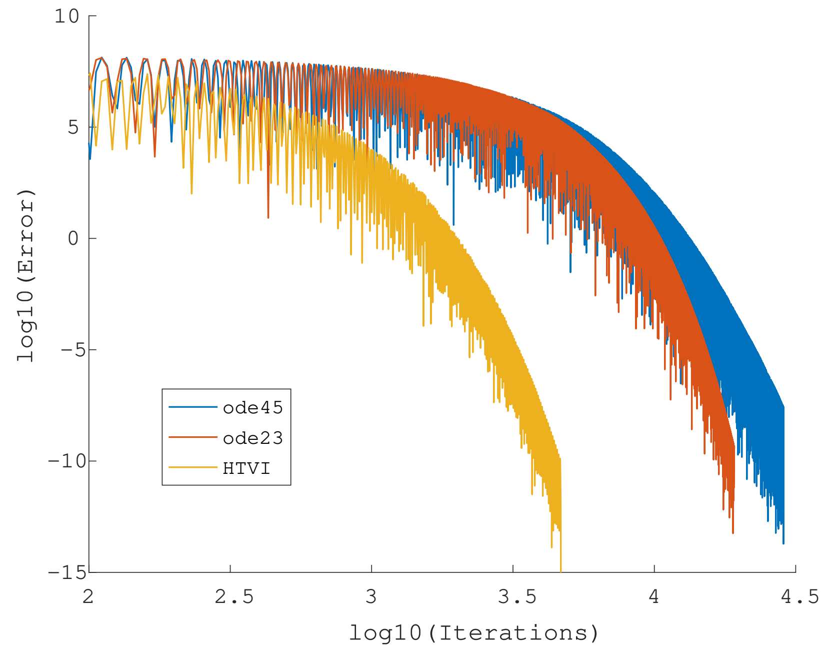


.

Accelerated Optimization using Adaptive Variational Integrators

■ Comparison to non-symplectic adaptive RK method

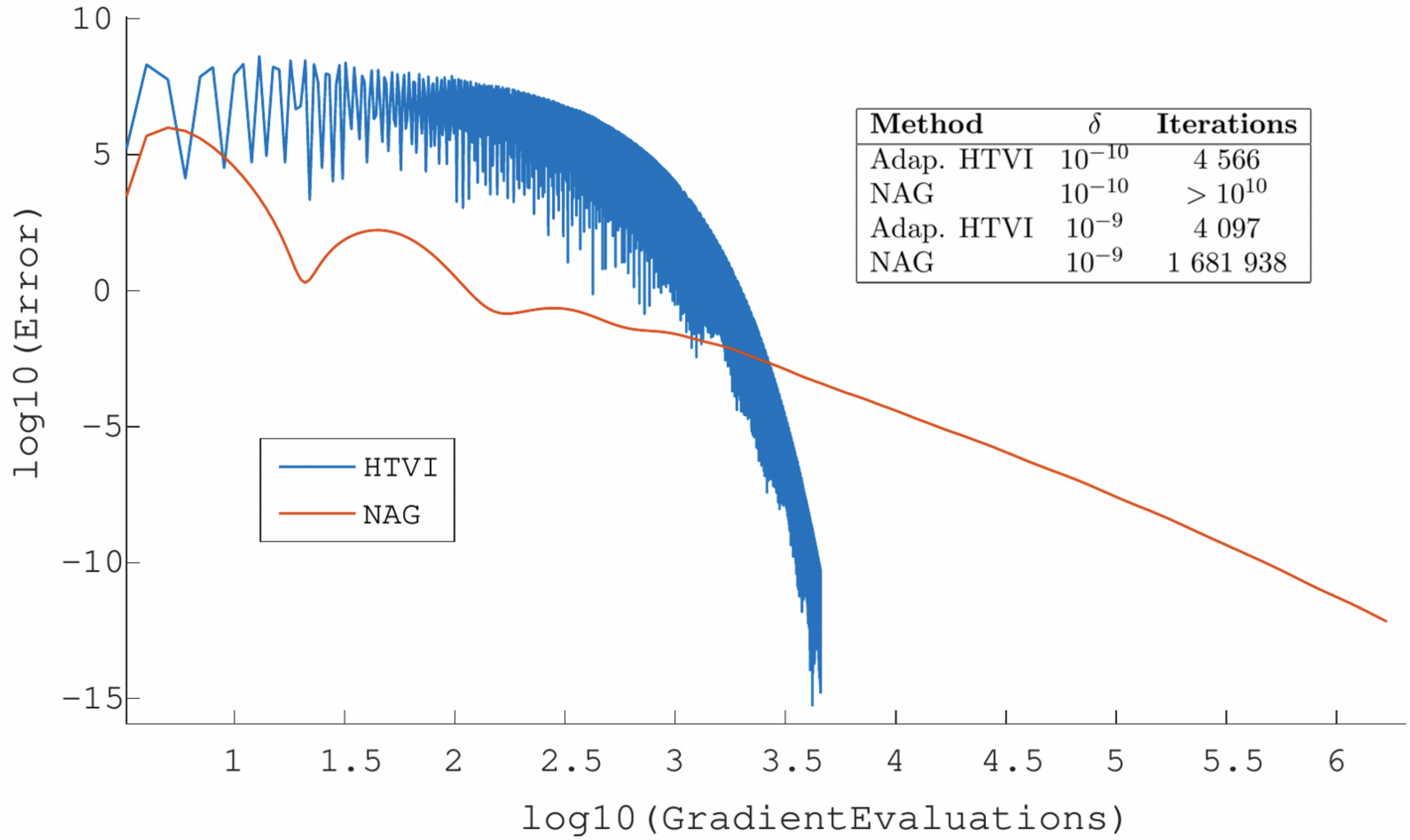




Direct approach (left) and Adaptive approach (right)

Accelerated Optimization using Adaptive Variational Integrators

■ Comparison to Nesterov's Accelerated Gradient (NAG)



Method	δ	Iterations
Adap. HTVI	10^{-10}	4 566
NAG	10^{-10}	$> 10^{10}$
Adap. HTVI	10^{-9}	4 097
NAG	10^{-9}	1 681 938

Accelerated Optimization on Riemannian Manifolds[8]

Convex and Weakly-Quasi-Convex Case

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

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

$$f(X(t)) - f(x^*) \leq \frac{2\lambda^2 e^{\beta_0}\left(f(x_0)-f(x^*)\right) + \zeta\|Log_{x_0}(x^*)\|^2}{2\lambda^2 e^{\beta_t}} = \mathcal{O}(e^{-\beta_t})$$

Strongly Convex Case

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

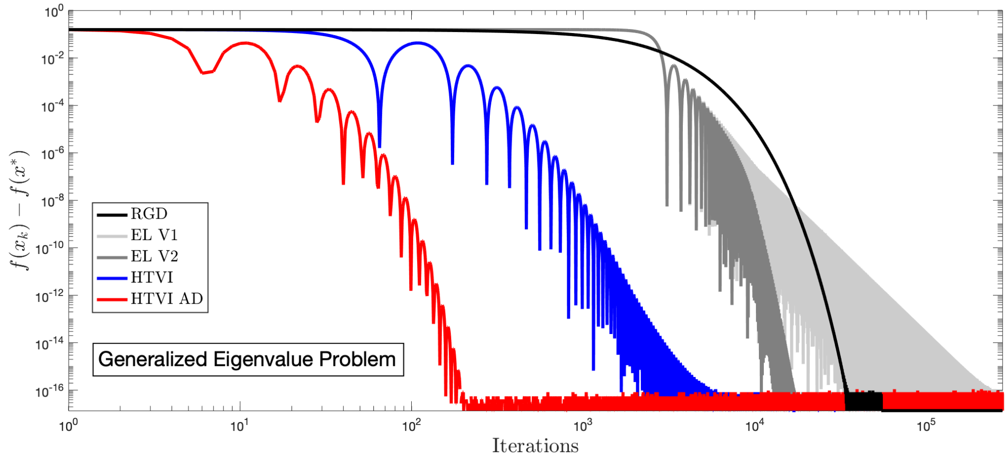
$$\mathcal{H}^{SC}(X,R,t) = \frac{e^{-\eta t}}{2}\langle\langle R,R\rangle\rangle + e^{\eta t}f(X)$$

$$f(X(t)) - f(x^*) \leq \frac{\mu\|Log_{x_0}(x^*)\|^2 + 2\left(f(x_0)-f(x^*)\right)}{2e^{\sqrt{\frac{\mu}{\zeta}}t}}$$

[8] V. Duruisseaux, M. Leok, *A Variational Formulation of Accelerated Optimization on Riemannian Manifolds*, SIAM Journal on Mathematics of Data Science, 4(2), 649-674, 2022.

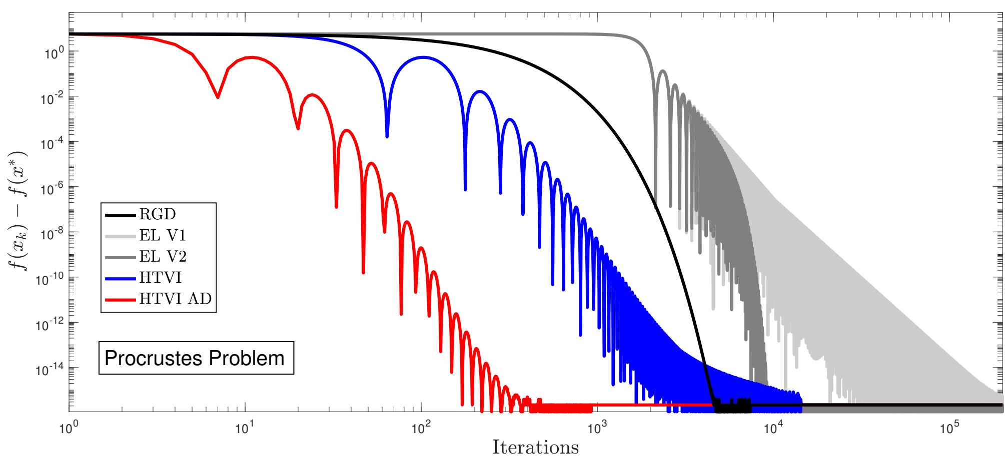
Accelerated Optimization on Riemannian Manifolds

Generalized Eigenvector Problem



$$f : \mathrm{St}(m,n) \to \mathbb{R}$$
$$X \mapsto f(X) = \mathrm{Tr}(X^\top A X N)$$

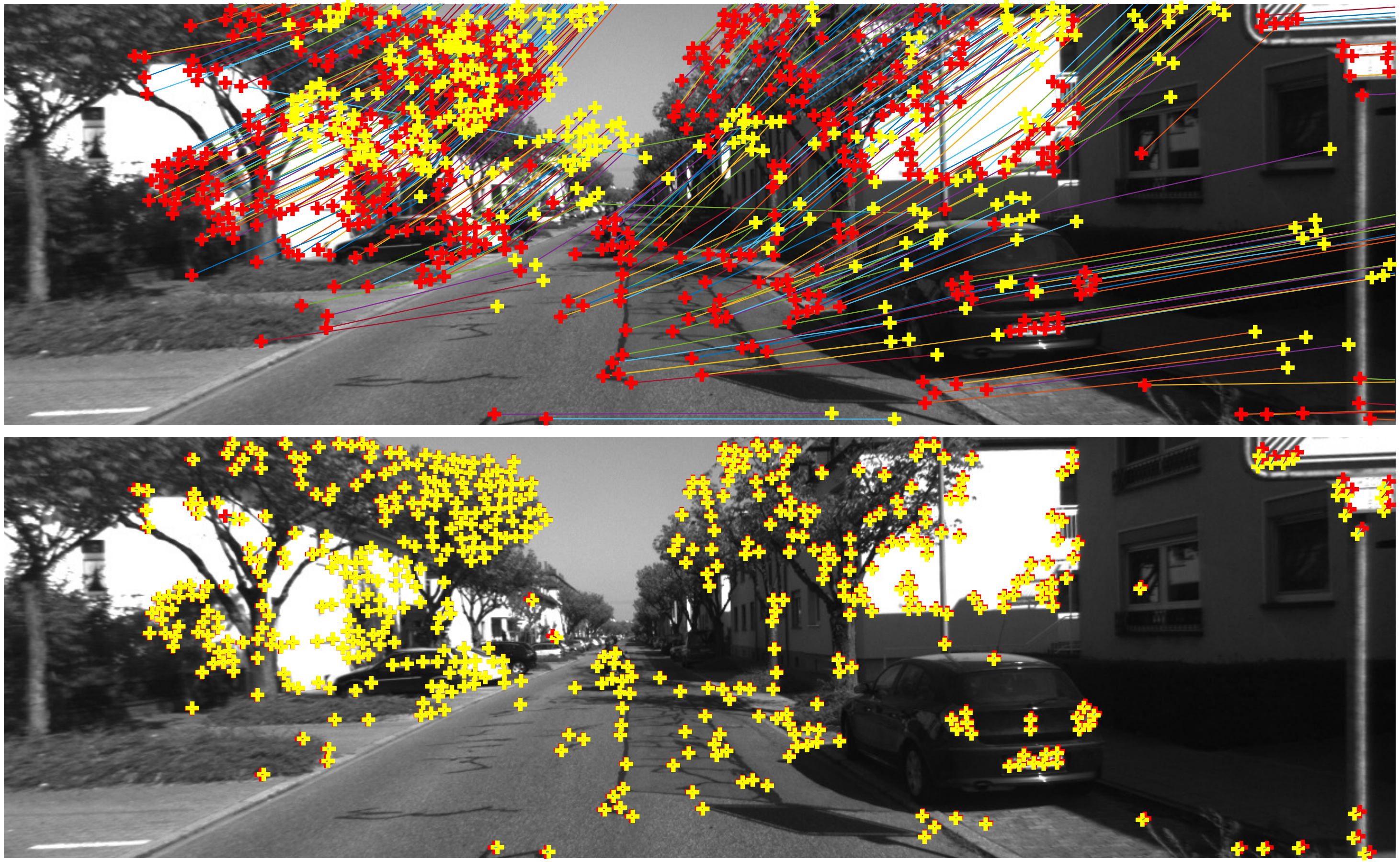
Unbalanced Orthogonal Procrustes Problem



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

Accelerated Optimization on Lie Groups[9]

■ Camera Pose Estimation Problem



[9]T. Lee, M. Tao, M. Leok, *Variational Symplectic Accelerated Optimization on Lie Groups*, Proc. IEEE Conf. Decision and Control, 233–240, 2021.

Summary and Future Directions

- Accelerated optimization algorithms can be viewed as discretizations of continuous-time flows that converge to the minimizer.
- One such class of flows are described by Bregman Lagrangian and Hamiltonian dynamics, which are time-dependent symplectic flows.
- Geometric discretization of these flows requires the use of time-adaptive Hamiltonian variational integrators.
- The problem of accelerated optimization on Riemannian manifolds and Lie groups motivates the development of time-adaptive Hamiltonian variational integrators on nonlinear manifolds.
- This requires the development of discrete Hamiltonian mechanics based on a discrete generalized energy, instead of a Type II/Type III generating function.

References

1. A. Wibisono, A. Wilson, M. I. Jordan, *A variational perspective on accelerated methods in optimization*, Proceedings of the National Academy of Sciences, 133, E7351–E7358, 2016. `doi:10.1073/pnas.1614734113`
2. V. Duruisseaux, J. Schmitt, M. Leok, *Adaptive Hamiltonian Variational Integrators and Symplectic Accelerated Optimization*, SIAM Journal of Scientific Computing, 43(4), A2949–A2980, 2021. `doi:10.1137/20M1383835`
3. V. Duruisseaux, M. Leok, *A Variational Formulation of Accelerated Optimization on Riemannian Manifolds*, SIAM Journal on Mathematics of Data Science, 4(2), 649–674, 2022. `doi:10.1137/21M1395648`
4. V. Duruisseaux, M. Leok, *Accelerated Optimization on Riemannian Manifolds via Discrete Constrained Variational Integrators*, Journal of Nonlinear Science, 32:42 (34 pages), 2022. `doi:10.1007/s00332-022-09795-9`
5. T. Lee, M. Tao, M. Leok, *Variational Symplectic Accelerated Optimization on Lie Groups*, Proc. IEEE Conf. Decision and Control, 233-240, 2021. `doi:10.1109/CDC45484.2021.9683657`
6. M. Leok, J. Zhang, *Discrete Hamiltonian Variational Integrators*, IMA Journal of Numerical Analysis, 31(4), 1497–1532, 2011. `doi:10.1093/imanum/drq027`
7. J. Schmitt, T. Shingel, M. Leok, *Lagrangian and Hamiltonian Taylor variational integrators*, BIT Numer. Math., 58, 457–488, 2018. `doi:10.1007/s10543-017-0690-9`

`github.com/vduruiss/AccOpt_via_GNI`