Variational Optimization on Lie Groups, with Examples of Leading (Generalized) Eigenvalue Problems

Molei Tao
Georgia Institute of Technology

Tomoki Ohsawa
University of Texas at Dallas

Abstract

The article considers smooth optimization of functions on Lie groups. By generalizing NAG variational principle in vector space (Wibisono et al., 2016) to Lie groups, continuous Lie-NAG dynamics which are guaranteed to converge to local optimum are obtained. They correspond to momentum versions of gradient flow on Lie groups. A particular case of $\mathsf{SO}(n)$ is then studied in details, with objective functions corresponding to leading Generalized EigenValue problems: the Lie-NAG dynamics are first made explicit in coordinates, and then discretized in structure preserving fashions, resulting in optimization algorithms with faithful energy behavior (due to conformal symplecticity) and exactly remaining on the Lie group. Stochastic gradient versions are also investigated. Numerical experiments on both synthetic data and practical problem (LDA for MNIST) demonstrate the effectiveness of the proposed methods as optimization algorithms (*not* as a classification method).

1 Introduction

The algorithmic task of optimization is important in data sciences and other fields. For differentiable objective functions, 1st-order optimization algorithms have been popular choices especially for high dimensional problems, largely due to their scalability, generality, and robustness. A celebrated class of them is based on Nesterov Accelerated Gradient descent (NAG; see e.g., (Nesterov, 1983, 2018)), also known as a major way to add momentum to Gradient Descent (GD). NAGs enjoy great properties such as quadratic decay of error (instead of GD's linear decay) for convex but not strongly convex objective functions. In addition, the introduction of momentum in NAG softens the dependence of convergence rate on the condition number of the problem. Since high dimensional problems often correspond to larger condition numbers, it is conventional wisdom that adding momentum to gradient descent makes it scale better with high dimensional problems (e.g., Ruder (2016), and Cheng et al. (2018) for rigorous results on related problems).

In particular, at least two versions of NAG have been widely used, referred to as NAG-SC and NAG-C for instance in Shi et al. (2018). While their original versions are iterative methods in discrete time, their continuum limits (as the step size goes to zero) have also been studied: for example, Su et al. (2014) thoroughly investigates these limits as ODEs, and Wibisono et al. (2016) establishes a corresponding variational principle (along with other generalizations). Further developments exist; for instance, Shi et al. (2018) discusses how to better approximate the original NAGs by high-resolution NAG-ODEs when step size is small but not infinitesimal, and was followed up by Wang and Tao (2020). Note, however, that no variational principle has been provided yet for the high-resolution NAG-ODEs, to the best of our knowledge.

Although the aforementioned discussions on NAG are in the context of finite dimensional vector space, a variational principle can allow it to be intrinsically generalized to manifolds. Such generalizations are meaningful, because objective functions may not always be a function on vector space, and abundant applications require optimization with respect to parameters in curved spaces. The first part of this article generalizes continuous NAG dynamics to Lie groups, which are differentiable manifolds that are also groups. Special orthogonal group $\mathsf{SO}(n)$, which contains n-by-n real orthogonal matrices with determinant 1, is a classical Lie group, and its optimization is not only relevant to data sciences (see e.g., Sec.3 and Appendix) but also to physical sciences. Some more examples include sym-

Proceedings of the 23[rd]International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

plectic groups, spin groups, and unitary groups, all of which play important roles in contemporary physics (e.g., Sattinger and Weaver (2013)); for instance, optimization on unitary groups found applications in quantum control (e.g., Glaser et al. (1998)), quantum information (e.g., Kitaev and Watrous (2000)), MIMO communication systems (e.g., Abrudan et al. (2009)), and NMR spectroscopy (e.g., Sorensen (1989)).

Variational principles on Lie groups (or more precisely, on the tangent bundle of Lie groups, for introducing velocity) provide a Lagrangian point of view for mechanical systems on Lie groups, and have been extensively studied in geometric mechanics (e.g., Marsden and Ratiu (2013); Holm et al. (2009)). Nevertheless, the application of geometric mechanics to NAG-type optimization in this article is new. The second part of this article will discretize the resulting NAG-dynamics on Lie groups, which lead to actual optimization algorithms. These algorithms are also new, although they can certainly be embedded as part of the profound existing field of geometric numerical integration (e.g., the classic monograph of Hairer et al. (2006)).

It is also important to mention that optimization on manifolds is already a field so rich that only an incomplete list of references can be provided, e.g., Gabay (1982); Smith (1994); Edelman et al. (1998); Absil et al. (2009); Patterson and Teh (2013); Zhang and Sra (2016); Zhang et al. (2016); Liu et al. (2017); Boumal et al. (2018); Ma et al. (2019); Zhang and Sra (2018); Liu et al. (2018). However, a specialization in Lie group will still be helpful, because the additional group structure (joined efforts with NAG) improves the optimization; for instance, a well known reduction is to, under symmetry, pull the velocity at any location on the Lie group back the tangent space at the identity (known as the Lie algebra).

We also note that NAG (either in vector space or on Lie group) is not restricted to convex optimization. In fact, the proposed methods will be demonstrated on an example of (leading) (Generalized) EigenValues (GEV) problems, which is known to be nonconvex (e.g., Chi et al. (2019) and its references therein).

GEV is a classical linear algebra problem behind tasks including Linear Discriminant Analysis (see Sec.4.3 and Appendix) and Canonical Correlation Analysis (e.g., Barnett and Preisendorfer (1987)). Due to its importance, numerous GEV algorithms exist (see e.g., Saad (2011)), some iterative (e.g., variants of power method) and some direct (e.g., Lanczos-based methods). And we choose GEV as an example to demonstrate our method applied to Lie group $\mathsf{SO}(n)$.

Meanwhile, another line of approaches has also been popular, especially for data sciences problems, often referred to as Oja flow (Oja, 1982), Sanger's rule (Sanger, 1989), and Generalized Hebbian Algorithm (Gorrell, 2006). While initially proposed for the leading eigenvalue problem, they extend to the leading GEV problem (e.g., Chen et al. (2019)). For a simple notation, we follow Chen et al. (2019) and denote them by 'GHA'. GHA is based on a matrix-valued ODE, whose long time solution converges to a solution of GEV; more details are reviewed in Appendix. Since the GHA ODE has to be discretized and numerically solved, GHA in practice is still an iterative method, but it is a special one: because of its ODE nature, GHA adapts well to a stochastic generalization of GEV, in which one only has access to noisy/incomplete realizations of the actual matrix (see Sec.3.3 for more details), and hence remains popular in machine learning. The proposed methods will also be based on ODEs and suitable to stochastic problems, and thus they will be compared with GHA (Sec.4.2). Worth mentioning is, GEV is still being actively investigated; besides Chen et al. (2019), recent progress include, for instance, Ge et al. (2016); Allen-Zhu and Li (2017); Arora et al. (2017). While the main contribution of this article is the momentum-based general Lie group optimization methodology (**not** GEV algorithms), the derived GEV algorithms are complementary to states-of-arts, because the proposed methods are indifferent to eigengap unlike Ge et al. (2016), and no direct access or inversion of the constraining matrix as different from Allen-Zhu and Li (2017); Arora et al. (2017); however, our method can be made stochastic but not 'doubly-stochastic'.

This article is organized as follows. Sec.2 derives the continuous Lie-group optimization dynamics based on the NAG variational principle. Sec.3.1 describes, at the continuous level, the case when the Lie group is $\mathsf{SO}(n)$, including the (full) eigenvalue problem and the leading GEV problem; both NAG dynamics and GD (no momentum) are discussed. Sec.3.2 then describes discretized algorithms, and Sec.3.3 extends them to stochastic problems. Sec.4 provides numerical evidence of the efficacy of our methods, with demonstrations on both synthetic and real data.

Quick user guide: For GEV, a family of NAG dynamics were obtained. The simplest ones are

$$\underline{\text{Lie-GD}}: \qquad \dot{R} = R([R^T A R, \mathcal{E}]) \tag{1}$$

Initial condition has to satisfy: $R(0)^T B R(0) = I$.

$$\underline{\text{Lie-NAG}}: \qquad \dot{R} = R\xi,\ \dot{\xi} = -\gamma(t)\xi + [R^T A R, \mathcal{E}] \tag{2}$$

where $\mathcal{E} := \left[\begin{smallmatrix} I_l & 0 \\ 0 & 0 \end{smallmatrix}\right]_{n\times n}$. Initial conditions have to satisfy: $R(0)^T B R(0) = I$ and $\xi(0)^T = -\xi(0)$.

Constant γ and $\gamma(t) = 3/t$ respectively correspond to Lie-NAG-SC and Lie-NAG-C. If it is affordable to tune

the constant γ value, our general recommendation is Lie-NAG-SC. Its associated optimization algorithm is Algm.2, and Algm.1 is also provided for Lie-GD.

2 Variational Optimization on Lie Group: the General Theory

2.1 Gradient Flow

Our focus is optimization problems on Lie groups: Let G be a compact Lie group, $f\colon \mathsf{G} \to \mathbb{R}$ be a smooth function, and consider the optimization problem

$$\min_{g\in\mathsf{G}} f(g).$$

We may define the gradient flow for this problem as follows: Let $T\mathsf{G}$ and $T^*\mathsf{G}$ be the tangent and cotangent bundles of G, $e \in \mathsf{G}$ be the identity, and $\mathfrak{g} := T_e\mathsf{G}$ be the Lie algebra of G. Suppose that $\mathfrak{g}$ is equipped with an inner product $\langle\!\langle \xi, \eta \rangle\!\rangle := \langle \mathbb{I}\xi, \eta\rangle$ with an isomorphism $\mathbb{I}\colon \mathfrak{g} \to \mathfrak{g}^*; \xi \mapsto \mathbb{I}(\xi)$ where $\mathfrak{g}^*$ is the dual of the Lie algebra $\mathfrak{g}$, and $\langle\cdot,\cdot\rangle$ stands for the natural dual pairing. One can naturally extend this metric to a left-invariant metric on G by defining, $\forall g \in \mathsf{G}$ and $\forall v, w \in T_g\mathsf{G}$, $\langle\!\langle v, w\rangle\!\rangle := \big\langle\!\big\langle T_g\mathsf{L}_{g^{-1}}(v), T_g\mathsf{L}_{g^{-1}}(w)\big\rangle\!\big\rangle$, where $\mathsf{L}_g\colon \mathsf{G} \to \mathsf{G}; h \mapsto gh$ is the left translation by $g \in \mathsf{G}$ and $T_h\mathsf{L}_g\colon T_h\mathsf{G} \to T_{gh}\mathsf{G}$ is its tangent map.

Now, we define the gradient vector field $\operatorname{grad} f$ on G as follows: For any $g \in \mathsf{G}$ and any $\dot g \in T_g\mathsf{G}$,

$$\langle\!\langle (\operatorname{grad} f)(g), \dot g\rangle\!\rangle = \langle \mathbf{d}f(g), \dot g\rangle \quad \forall g \in \mathsf{G} \quad \forall \dot g \in T_g\mathsf{G},$$

where $\mathbf{d}$ stands for the exterior differential. This gives

$$(\operatorname{grad} f)(g) = T_e\mathsf{L}_g \circ \mathbb{I}^{-1} \circ T_e^*\mathsf{L}_g(\mathbf{d}f(g)),$$

where $T_e^*\mathsf{L}_g$ is the dual of $T_e\mathsf{L}_g$, i.e., $\forall \alpha_g \in T_g^*\mathsf{G}$ and $\forall \xi \in \mathfrak{g}$, $\langle T_e^*\mathsf{L}_g(\alpha_g), \xi\rangle = \langle \alpha_g, T_e\mathsf{L}_g(\xi)\rangle$. Hence the gradient descent equation is given by

$$\dot g = -(\operatorname{grad} f)(g) = -T_e\mathsf{L}_g \circ \mathbb{I}^{-1} \circ T_e^*\mathsf{L}_g(\mathbf{d}f(g)). \quad (3)$$

2.2 Adding Momentum: the Variational Optimization

Our work provides a natural extension of variational optimization of Wibisono et al. (2016) to Lie groups making use of the geometric formulation of the Euler–Lagrange equation on Lie groups. Specifically, let us define the Lagrangian $L\colon T\mathsf{G} \times \mathbb{R} \to \mathbb{R}$ as follows:

$$L(g, \dot g, t) := r(t)\left(\frac{1}{2}\langle\!\langle \dot g, \dot g\rangle\!\rangle - f(g)\right), \quad (4)$$

where $r\colon \mathbb{R} \to \mathbb{R}_{>0}$ is a smooth positive-valued function. Instead of working with the tangent bundle $T\mathsf{G}$ directly, it is more convenient to use the left-trivialization of $T\mathsf{G}$, i.e., we may identify $T\mathsf{G}$ with $\mathsf{G} \times \mathfrak{g}$ via the map $\mathsf{G} \times \mathfrak{g} \to T\mathsf{G}; (g, \xi) \mapsto (g, T_e\mathsf{L}_g(\xi))$. Under this identification, we have the Lagrangian $L\colon \mathsf{G} \times \mathfrak{g} \times \mathbb{R} \to \mathbb{R}$ defined as

$$L(g, \xi, t) := r(t)\left(\frac{1}{2}\langle \mathbb{I}(\xi), \xi\rangle - f(g)\right). \quad (5)$$

The Euler–Lagrange equation for this Lagrangian is (see, e.g., Holm et al. (2009, Section 7.3) and also Marsden and Ratiu (2013))

$$\frac{d}{dt}\left(\frac{\delta L}{\delta \xi}\right) = \operatorname{ad}_\xi^* \frac{\delta L}{\delta \xi} + T_e^*\mathsf{L}_g(\mathbf{d}_g L),$$

along with

$$\dot g = T_e\mathsf{L}_g(\xi) =: g\xi, \quad (6)$$

where ad^* is the coadjoint operator; $\delta L/\delta\xi \in \mathfrak{g}^*$ is defined so that, for any $\delta\xi \in \mathfrak{g}$,

$$\left\langle \frac{\delta L}{\delta \xi}, \delta\xi \right\rangle = \frac{d}{ds} L(g, \xi + s\delta\xi, t)\bigg|_{s=0};$$

also note that $\mathbf{d}_g L$ stands for the exterior differential of $g \mapsto L(g, \xi, t)$. Using the above expression (5) of the Lagrangian, we obtain

$$\frac{d}{dt}\mathbb{I}(\xi) = -\gamma(t)\mathbb{I}(\xi) + \operatorname{ad}_\xi^* \mathbb{I}(\xi) - T_e^*\mathsf{L}_g(\mathbf{d}f(g)), \quad (7)$$

where we defined $\gamma(t) := r'(t)/r(t)$.

Choices of γ. We will mainly consider $\gamma(t) = \gamma$ (constant) and $\gamma(t) = 3/t$, derived from $r = \exp(\gamma t)$ and $r = t^3$. In vector space, these two choices respectively correspond to, as termed for instance in Shi et al. (2018), NAG-SC and NAG-C, which are the continuum limits of two classical versions of Nesterov's Accelerated Gradient methods (Nesterov, 1983, 2018).

Lyapunov function. Let $t \mapsto (g(t), \xi(t))$ be a solution of eq. (7). Assuming that g_0 is an isolated local minimum of f, we can show that the dynamics starting in a neighborhood of g_0 converges to g_0 as follows. Define the "energy" function $E\colon \mathsf{G} \times \mathfrak{g} \to \mathbb{R}$ as

$$E(g, \xi) := \frac{1}{2}\langle\!\langle \xi, \xi\rangle\!\rangle + f(g) = \frac{1}{2}\langle \mathbb{I}(\xi), \xi\rangle + f(g). \quad (8)$$

This gives a Lyapunov function. In fact, there exists a neighborhood U of $(g_0, 0)$ such that $E(g, \xi) \geq f(g) > f(g_0)$ for any $(g, \xi) \in U\backslash\{(g_0, 0)\}$. Moreover, we have $\frac{d}{dt}E(g(t), \xi(t)) = -\gamma\langle\!\langle \xi, \xi\rangle\!\rangle \leq 0$, where the equality implies $\xi = 0$, for which (7) gives $\mathbf{d}f(g) = 0$, which locally gives $g = g_0$.

3 The Example of $\mathsf{SO}(n)$ and Its Application to Leading GEV

3.1 The Continuous Formulations

3.1.1 The Symmetric Eigenvalue Problem

Let A be a real symmetric $n \times n$ matrix, and define, as in Brockett (1989); Mahony and Manton (2002),

$$f\colon \mathsf{SO}(n) \to \mathbb{R}; \qquad f(R) := \operatorname{tr}(R^T A R \mathcal{N}),$$

where $\mathcal{N} := \operatorname{diag}(1, 2, \dots, n)$. We equip the Lie algebra $\mathfrak{so}(n)$ with the inner product $\langle\!\langle \xi, \eta \rangle\!\rangle := \operatorname{tr}(\xi^T \eta)$. Then we may identify $\mathfrak{so}(n)^*$ with $\mathfrak{so}(n)$ via this inner product. Then the "force" term in (7) is given by $T_I^* \mathsf{L}_R(\mathbf{d}f(R)) = [R^T A R, \mathcal{N}]$. Since $\operatorname{ad}_\xi^* \mu = [\mu, \xi]$ for any $\xi \in \mathfrak{so}(n)$ and $\mu \in \mathfrak{so}(n)^* \cong \mathfrak{so}(n)$, (7) becomes

$$\dot{R} = R\xi, \ \dot{\xi} = -\gamma\xi + \mathbb{I}^{-1}\big([\mathbb{I}(\xi), \xi] - [R^T A R, \mathcal{N}]\big), \quad (9)$$

whereas the gradient descent equation (3) gives

$$\dot{R} = -R\mathbb{I}^{-1}([R^T A R, \mathcal{N}]). \quad (10)$$

Remark 3.1 (Rigorous results v.s. intuitive addition of momentum)**.** The above dynamics work for any positive definite isomorphism $\mathbb{I}\colon \mathfrak{g} \to \mathfrak{g}^*$. For simplicity, we will use $\mathbb{I} = \mathrm{id}$ (where $\mathfrak{g}^*$ is identified with $\mathfrak{g}$) in implementations in this article. In this case, the $[\mathbb{I}(\xi), \xi]$ term and the $\mathbb{I}^{-1}$ operation vanish, and the momentum version (9) is heuristically obtainable from (10) just like how momentum was added to gradient flow in vector spaces. Otherwise, they create additional nontrivial nonlinearities that account for the curved space.

Remark 3.2 (Relation to double-bracket)**.** When $\mathbb{I} = \mathrm{id}$, the gradient flow (10) becomes $\dot{R} = -R([R^T A R, \mathcal{N}])$. By setting $M(t) := R(t)^T A R(t)$, we recover the double-bracket equation $\dot{M} = -[M, [M, \mathcal{N}]]$ of Brockett (1991) (see also Bloch et al. (1992)). Note that there is a sign difference from Brockett (1991) because Brockett's is gradient *ascent*.

Remark 3.3 (Generality)**.** The proposed methods, Lie-NAG (9) and Lie-GD (10), are indifferent to the absolute location of A's eigenvalues, because they are invariant to the shift $A \mapsto A + \lambda I$. To see this, note $[R^T A R, \mathcal{N}] \mapsto [R^T(A + \lambda I)R, \mathcal{N}] = [R^T A R, \mathcal{N}] + \lambda[R^T R, \mathcal{N}] = [R^T A R, \mathcal{N}] + \lambda[I, \mathcal{N}] = [R^T A R, \mathcal{N}]$. Therefore, the proposed methods work the same no matter whether A is positive/negative-definite. In the generalized eigenvalue setting (see future Sec.3.1.3), the same reasoning and invariance hold for $L^{-T} A L^{-1} \mapsto L^{-T} A L^{-1} + \lambda I$ where $L^T L = B$.

3.1.2 The Leading l Eigenvalue Problem

Let A be a real symmetric $n \times n$ matrix. Since finding the smallest l eigenvalues of A is the same as finding the largest l eigenvalues of $-A$, define

$$f\colon \mathsf{SO}(n) \to \mathbb{R}; \qquad f(R) := -\operatorname{tr}(E^T R^T A R E), \quad (11)$$

where $E := \left[\begin{smallmatrix} I_l \\ 0 \end{smallmatrix}\right]$ is $n \times l$ where I_l is the $l \times l$ identity matrix and 0 is the $(n-l) \times l$ zero matrix.

The cost function is almost the same as the previous case except that $\mathcal{N}$ is now replaced by

$$\mathcal{E} := EE^T = \left[\begin{smallmatrix} I_l & 0 \\ 0 & 0 \end{smallmatrix}\right].$$

So we have $T_I^* \mathsf{L}_R(\mathbf{d}f(R)) = -[R^T A R, \mathcal{E}]$.

3.1.3 The Leading l Generalized Eigenvalues

Consider the leading l Generalized EigenValues problem (GEV): given n-by-n symmetric A and n-by-n positive definite B, we seek an optimizer of

$$\max_{V \in \mathbb{R}^{n \times l}} \operatorname{tr}(V^T A V) \quad \text{s.t.} \quad V^T B V = I_{l \times l}. \quad (12)$$

It can be seen, by Cholesky decomposition $B = L^T L$ and a Lie group isomorphism $X \mapsto LX$, that

Proposition 3.1. *$G = \{X | X \in \mathbb{R}^{n \times n}, X^T B X = I\}$ is a Lie group. Its identity is L^{-1}, and its multiplication is not the usual matrix multiplication but $X_1 \cdot X_2 = X_1 L X_2$.*

Therefore, in theory, GEV can be solved by padding V into X and then following our general approach (7).

The point of this section is to make this solution explicit, and more importantly, to show L is never explicitly needed, which leads to computational efficiency. In fact, the same NAG dynamics

$$\dot{R} = R\xi, \qquad \dot{\xi} = -\gamma(t)\xi + [R^T A R, \mathcal{E}] \quad (13)$$

with initial conditions satisfying

$$R(0)^T B R(0) = I, \quad \xi(0)^T = -\xi(0)$$

will solve (12) upon projecting the first l columns of R into V.

Note the only difference from the previous two sections is the initial condition on R. In addition, although positive definite B is needed for the group isomorphism, it is only a sufficient (not necessary) condition for NAG (13) to work.

A rigorous justification of why (13) works for not only EV but also GEV can be found in Appendix, where one will also find the proof of a quick sanity check:

Theorem 3.1. *Under* (13) *and consistent initial condition, $R(t)^T B R(t) = I$ and $\xi(t)^T = -\xi(t)$ for all t.*

The objective function itself does not decrease monotonically in NAG, because it acts as potential energy, which exchanges with kinetic energy, but the total energy decreases (eq.8).

On the other hand, if one considers Lie-GD, which can be shown to generalize to GEV also by only modifying the initial condition (given by (1)), then not only does $R(t)$ stay on the Lie group G (see Appendix), but also is the objective function $\mathrm{tr}[-(R^T(t)AR(t)\mathcal{E})]$ monotone (by construction).

3.2 The Discrete Algorithms

Define Cayley transformation[1] as $\mathrm{Cayley}(\xi) := (I - \xi/2)^{-1}(I + \xi/2)$. It will be useful as a 2nd-order structure-preserving approximation of matrix exp, the latter of which is computationally too expensive. More precisely, $\exp(h\xi) = \mathrm{Cayley}(h\xi) + \mathcal{O}(h^3)$.

Lie-GD. We adopt a 1st-order (in h) explicit discretization of the dynamics $\dot{R} = R([R^T AR, \mathcal{E}])$:

Algorithm 1 A 1st-order Lie-GD for leading GEV

1: Initialize with some R_0 satisfying $R_0^T B R_0 = I$.
2: **for** $i = 0, \cdots$,TotalSteps-1 **do**
3: $\quad f_i \leftarrow R_i^T AR_i\mathcal{E} - \mathcal{E}R_i^T AR_i$.
4: $\quad R_{i+1} \leftarrow R_i\mathrm{Cayley}(hf_i)$
5: **end for**
6: Output $R_{\mathrm{TotalSteps}}$ as $\mathrm{argmin}\, f$ in (11).

Note Algm.1 is more accurate than forward Euler discretization despite that both are 1st-order. This is because all R_i's it produces will remain on the Lie group (i.e., $R_i^T B R_i = I$; see Thm.4.2 in Appendix).

Lie-NAG. We present a 2nd-order (in h) explicit discretization of the dynamics $\dot{R} = R\xi$, $\dot{\xi} = -\gamma(t)\xi + [R^T AR, \mathcal{E}]$. Unlike the Lie-GD case, the discretization was achieved by the powerful machinery of operator splitting, and can be easily generalized to arbitrarily high-order (e.g., McLachlan and Quispel (2002); Tao (2016)), provided that Cayley transformation was replaced by a higher-order Lie-group-preserving approximation of matrix exponential.

More precisely, denote by ϕ^h the exact h-time flow of the NAG dynamics, and by ϕ_1^h and ϕ_2^h some p-th order approximations of the h-time flows of $\dot{R} = R\xi$, $\dot{\xi} = 0$ and $\dot{R} = 0$, $\dot{\xi} = -\gamma(t)\xi + [R^T AR, \mathcal{E}]$. Note even though ϕ is unavailable, the latter systems are analytically solvable, so if $\exp(\xi h)$ is exactly computed, ϕ_1 and ϕ_2 can be made exact. Even if they are just p-th order approximations ($p \geq 2$), operator splitting yields $\phi^h = \phi_2^{h/2} \circ \phi_1^h \circ \phi_2^{h/2} + \mathcal{O}(h^3)$. Other ways of composing ϕ_1,ϕ_2 can lead to higher order methods (Appendix describes some 4th-order options), with maximum order capped by p. For simpler coding, $\dot{\xi} = -\gamma(t)\xi + [R^T AR, \mathcal{E}]$ can be further split into $\dot{\xi} = -\gamma(t)\xi$ and $\dot{\xi} = [R^T AR, \mathcal{E}]$, and Algm.2 is based on $\phi_3^{h/2} \circ \phi_2^{h/2} \circ \phi_1^h \circ \phi_2^{h/2} \circ \phi_3^{h/2}$:

Algorithm 2 A 2nd-order Lie-NAG for leading GEV

1: Initialize with some R_0 and ξ_0 satisfying $R_0^T B R_0 = I$ and $\xi_0^T = -\xi_0$.
2: **for** $i = 0, \cdots$,TotalSteps-1 **do**
3: $\quad \xi_{i'} \leftarrow \xi_i + h/2(R_i^T AR_i\mathcal{E} - \mathcal{E}R_i^T AR_i)$.
4: $\quad \xi_{i'} \leftarrow \begin{cases} \exp(-\gamma h/2)\xi_{i'}, & \text{for NAG-SC} \\ ((ih)^3/((i+1/2)h)^3)\xi_{i'}, & \text{for NAG-C} \end{cases}$.
5: $\quad R_{i+1} \leftarrow R_i\mathrm{Cayley}(h\xi_{i'})$.
6: $\quad \xi_{i'} \leftarrow \begin{cases} \exp(-\gamma h/2)\xi_{i'}, & \text{NAG-SC} \\ (((i+1/2)h)^3/((i+1)h)^3)\xi_{i'}, & \text{NAG-C} \end{cases}$.
7: $\quad \xi_{i+1} \leftarrow \xi_{i'} + h/2(R_{i+1}^T AR_{i+1}\mathcal{E} - \mathcal{E}R_{i+1}^T AR_{i+1})$.
8: **end for**
9: Output $R_{\mathrm{TotalSteps}}$ as $\mathrm{argmin}\, f$ in (11).

Also by Thm.4.2, all R_i's remain on the Lie group if arithmetics have infinite machine precision.

In addition, Algm.2 is conformal symplectic (see Appendix), which is indicative of favorable accuracy in long time energy behavior. To prove so, note both ϕ_1 and ϕ_3 as exact Hamiltonian flows preserve the canonical symplectic form, and two substeps of ϕ_2 as linear maps discount it by a multiplicative factor of $r(t_i)/r(t_{i+1})$. This exactly agrees with the continuous theory in Appendix.

3.3 Generalization to Stochastic Problems

Setup: now let us consider a Stochastic Gradient (SG) setup, where one may not have full access to A but only a finite collection of its noisy realizations. More precisely, given one realization of i.i.d. random matrices $A_1, \cdots, A_K$, the goal is to compute the leading (generalized) eigenvalues of $A = \frac{1}{K}\sum_{k=1}^K A_k$ based on A_k's without explicitly using A.

Implementation: following the classical stochastic gradient approach, we simply replace A in each algorithm by A_κ, where κ is a uniform random variable on $[K]$, independently drawn at each timestep.

Remark 3.4. Like Ge et al. (2016) and unlike Chen et al. (2019), the proposed methods do not allow B to be a stochastic approximation. Only A can be stochastic. On the other hand, unlike both Ge et al. (2016) and Chen et al. (2019), we do not require a direct access to B, and all information about B is reflected in the initial condition $R(0)$.

Intuition: we now make heuristic arguments to gain insights about the performance of the method.

[1] It is the same as Pade(1,1) approximation.

First, based on the common approximation of stochastic gradient as batch gradient plus Gaussian noise (see e.g., Li et al. (2019) for some state-of-art quantifications of the accuracy of this approximation), assume $A_\kappa = A+\sigma H$ where H is a symmetric Gaussian matrix (assumed as $H = \Xi+\Xi^T$ where Ξ is an n-by-n matrix with i.i.d. standard normal elements), i.i.d. at each step. Then the gradient $[R^T A_\kappa R, \mathcal{E}]$ is, in distribution and conditioned on R, equal to $[R^T AR, \mathcal{E}] + 2\sigma\Xi$. This is because $[R^T A_\kappa R, \mathcal{E}]$ is Gaussian and its mean is $[R^T AR, \mathcal{E}]$ and covariance is $\sigma^2 \text{covar}[[R^T HR, \mathcal{E}]|R]$, which can be computed to be $4\sigma^2 I$, independent of R as long as $R^T R = I$ and $\mathcal{E}$ is a degenerate identity. Therefore, at least in the case of $\mathbb{I} = \text{id}$, the Lie-NAG SG dynamics can be understood through

$$\dot{R} = R\xi, \qquad \dot{\xi} = -\gamma(t)\xi + [R^T AR, \mathcal{E}] + 2\hat{\sigma}E, \quad (14)$$

where E is a skew-symmetric white-noise, i.e., E_{ij} with $i < j$ being i.i.d. white noise, $E_{ji} = -E_{ij}$, and $E_{ii} = 0$, and $\hat{\sigma} = \sigma$ in this continuous setting.

Worth mentioning is, once one uses a numerical discretization, namely

$$\begin{aligned}\xi_{i+1} &= \xi_i - h\gamma(t_i)\xi_i + h[R_i^T A_{\kappa_i} R_i, \mathcal{E}] + o(h),\\ &\overset{D}{=} \xi_i - h\gamma(t_i)\xi_i + h[R_i^T AR_i, \mathcal{E}] + h2\sigma E_i + o(h)\end{aligned}$$

then since κ does not randomize infinitely frequently, the effective noise amplitude $\hat{\sigma}$ gets scaled as

$$\hat{\sigma} = \sqrt{h}\sigma + o(\sqrt{h}), \quad (15)$$

because a 1st-order discretization of (14) should have its ξ component being

$$\xi_{i+1} = \xi_i - h\gamma(t_i)\xi_i + h[R_i^T AR_i, \mathcal{E}] + \sqrt{h}2\hat{\sigma}E_i + o(h)$$

due to stochastic calculus. This leads to $h2\sigma E_i = \sqrt{h}2\hat{\sigma}E_i + o(h)$, and hence (15).

Secondly, recall an analogous vector space setting, in which one considers

$$\dot{q} = p, \qquad \dot{p} = -\gamma p - \nabla V(q) + \hat{\sigma}e$$

where e is standard vectorial white-noise. It is well known that under reasonable assumptions (e.g., Pavliotis (2014)) this diffusion process admits, and converges weakly to an invariant distribution of $Z^{-1}\exp(-H(q,p)/kT)dqdp$, where $H = \|p\|^2/2+V(q)$ is the Hamiltonian, Z is some normalization constant, and $kT = \hat{\sigma}^2/(2\gamma)$ is the temperature (with unit).

It is easy to see that for the purpose of optimization, the temperature should be small. If one uses vanishing stepsizes, since $\hat{\sigma}^2 = h\sigma^2$, $kT \to 0$, and stochastic optimization can be guaranteed to work (more details in Robbins and Monro (1951)). If h is small but not infinitesimal, q (or R) is still concentrated near the optimum value(s) with high probability.

Now recall Lie-NAG-SC uses constant γ; Lie-NAG-C, on the contrary, uses $\gamma(t) = 3/t$. This means Lie-NAG-SC equipped with SG converges to some invariant distribution at temperature $h\sigma^2/(2\gamma)$, but Lie-NAG-C-SG's 'temperature' $kT = h\sigma^2/(6/t)$ grows unbounded with t for constant h; i.e., constant stepsize Lie-NAG-C-SG doesn't converge even in a weak sense.

This is another reason that our general recommendation is Lie-NAG-SC over Lie-NAG-C. On the other hand, there are multiple possibilities to correct the non-convergence of Lie-NAG-C: (i) appropriately vanishing h can lead to recovery of an invariant distribution, but to obtain a fixed accuracy one would need more steps; (ii) one can add a correction to the dynamics (Wang and Tao, 2020); (iii) modify $\gamma(t)$.

Corrected dissipation coefficient: this article experimented with option (iii) with

$$\gamma = 3/t + ct, \quad \text{where } c \text{ is a small constant;} \quad (16)$$

see Sec.4.2. This choice corresponds to $r(t) = \exp(ct^2/2)t^3$ in the variational formulation. Formally, it leads to 0 temperature, but in practice early stopping is needed because any finite h cannot properly numerical-integrate the dynamics when γ becomes sufficiently large.

The reason for choosing the specific linear form of the correction $+ct$ is in Appendix.

4 Experiments

4.1 Leading Eigenvalue Problems

4.1.1 Bounded Spectrum

We first test the proposed methods on a synthetic problem: finding the l largest eigenvalues of $A = (\Xi+\Xi^T)/2/\sqrt{n}$, where Ξ is a sample of an n-by-n matrix with i.i.d. standard normal elements. The scaling of $1/\sqrt{n}$ ensures[2] the leading eigenvalues are bounded by a constant independent of n; for an unbounded case, see the next example.

Fig. 1 shows results for a generic sample of 500-dimensional A. The proposed Lie-NAG's, i.e. variational methods with momentum, converge significantly faster than the popular GHA. This advantage is even more significant in higher dimensions (see Fig. 6 in Appendix). Note Fig. 1 plots accuracy as a function of

[2]For more precise statement and justification, see random matrix theory for Gaussian Orthogonal Ensemble (GOE), or more generally Wigner matrix Wigner (1958)

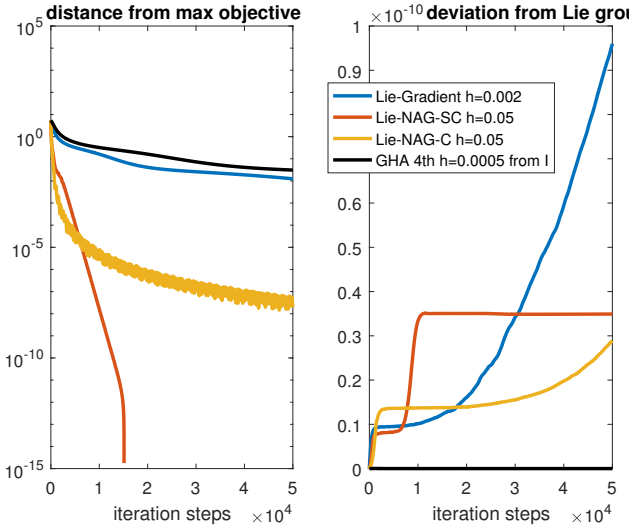

Figure 1: Performances of proposed Lie-GD, Lie-NAG-C and Lie-NAG-SC, compared with GHA, for computing the leading $l = 2$ eigenvalues of scaled GOE. All algorithms use step sizes tuned to minimize error in 5×10^4 iterations (although the proposed methods do not need much tuning), and identity initial condition. GHA was based on Runge-Kutta-4 integration of $\dot{Q} = (I - QQ^T)AQ$ for accuracy, and an Euler integration did not result in any notable error reduction. NAG-SC uses friction coefficient untuned $\gamma = 1$. The deviations of Lie-NAGs and Lie-GD from the Lie group are machine/platform (MATLAB) precision artifacts.

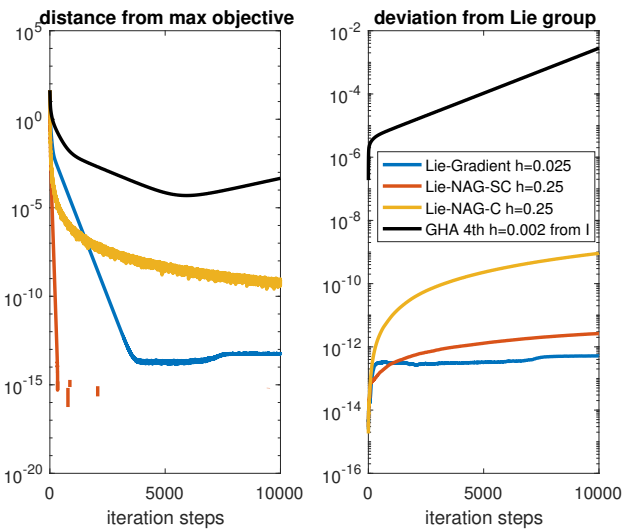

Figure 2: Proposed Lie-GD, Lie-NAG-C and Lie-NAG-SC, compared with GHA, for computing the leading $l = 2$ eigenvalues of $A = -\Xi\Xi^T/2$. Ξ is 25-dimensional. Other descriptions are same as in Fig.1.

the number of iterations, and readers interested in accuracy as a function of wallclock are referred to Fig. 7 (note wallclock count is platform dependent and therefore the latter illustration is only qualitative but not quantitative, thus placed in the Appendix). In any case, for this problem at least, if low-moderate accuracy is desired, Lie-NAG-C is the most efficient among tested methods; if high accuracy is desired instead, Lie-NAG-SC is the optimal choice.

Note the fact that A has both positive and negative eigenvalues should not impair the credibility of this demonstration. This is because one can shift A to make it positive definite or negative definite, and the convergences will be precisely the same. See Rmk.3.3.

4.1.2 Unbounded Spectrum

Now consider computing the leading eigenvalues of $A = -\Xi\Xi^T/2$ (Ξ similarly defined as in Sec.4.1.1). This is equivalent to finding the l smallest eigenvalues of $\Xi\Xi^T/2$. Doing so is relevant, for instance, in graph theory, where the 2nd smallest eigenvalue of graph Laplacian is the algebraic connectivity of the graph (Fiedler, 1973; Von Luxburg, 2007).

Fig.2 shows the advantage of variational methods (i.e., with momentum), even when the dimension is relatively low $n = 25$. A is defined such that its spectrum grows linearly with n, and GHA thus needs to use tiny timesteps. Although the proposed methods also need to use reduced step sizes for bigger n, the rate of reduction is much slower than that for GHA (results omitted).

4.2 Stochastic Leading Eigenvalue Problems

To investigate the efficacy of the proposed methods in the stochastic setup (Sec.3.3), we take the same A from Sec.4.1.1, and add $K = 100$ random perturbations to it to form a batch $A_1, \cdots, A_K$. Each random perturbation is $(\Xi + \Xi^T)/4/\sqrt{n}$ for i.i.d. Ξ; note these are large fluctuations when compared to A. Then A is refreshed to be the mean of A_k's, whose leading $l = 2$ eigenvalues are accurately computed as the ground truth.

Fig.3 shows the advantage of variational methods, even though their larger step sizes lead to much higher variances of the stochastic gradient approximation. The corrected dissipation (16) enabled the convergence of NAG-C. The same correction slows down the convergence of NAG-SC in the beginning, but significantly improves its long time performance, which otherwise stagnates at small but not infinitesimal error.

The reason NAG-SC-original stagnates is, over long time, it samples from an invariant distribution at a nonzero temperature. This invariant distribution, however, is not the exact one of the continuous limit; the latter of which would concentrate around the minimizer with 0 error. Instead, the numerical method's invariant distribution, if existent, is $\mathcal{O}(h^p)$ away from the exact one (Bou-Rabee and Owhadi, 2010; Abdulle

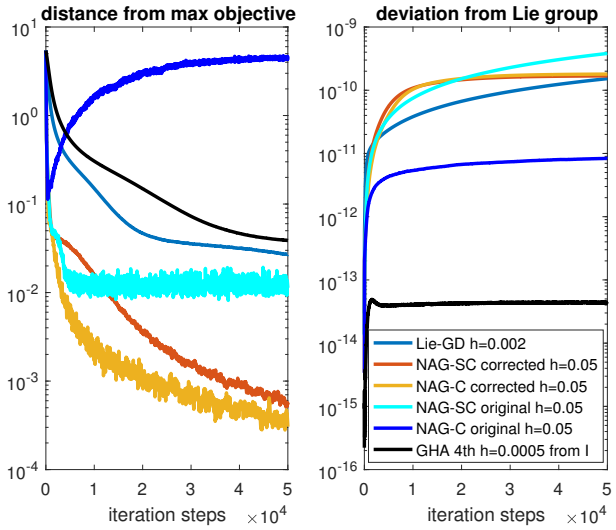

Figure 3: The computation of leading $l = 2$ eigenvalues of $A = \frac{1}{K}\sum_{k=1}^{K} A_k$ based on stochastic gradients from batch $A_1, \cdots, A_K$ without A. NAG-SC corrected, NAG-C corrected, NAG-SC and NAG-C use, respectively, $\gamma = 1 + 0.01t$, $3/t + 0.01t$, 1, and $3/t$. Other descriptions are same as in Fig.1.

et al., 2014) under suitable assumptions, which means as the numerical method converges, it gives R's that are $\mathcal{O}(h^p)$ away from the exact minimizer with high probability. NAG-SC-corrected alleviated this issue.

4.3 Leading Generalized Eigenvalue: a Demonstration Based on LDA

We report numerical experiments on multiclass Fisher Linear Discriminant Analysis (LDA) of the handwritten-digits database MNIST (LeCun et al., 1998). Since it is known that LDA can be formulated as a leading generalized eigenvalue problem (e.g., reviewed in Li et al. (2006); Welling (2005); see appendix for a summary), we use it as an example to test our leading GEV algorithm. Important to note is, our purpose is NOT to construct an algorithm for MNIST classification, as it is known that LDA does not achieve state-of-art performance in that regard (test error based on exact leading GEV solution was 10% in our experiment). Instead, we simply would like to quantify the efficacy of our algorithm applied to a leading generalized eigenvalue problem based on real life data.

The 60000 training data of MNIST were employed to compute the 'inter-class scatter matrix' A and the 'intra-class scatter matrix' B (see appendix for more details). Each 28×28 image had its white margins cropped, resulting in a 400-dimensional vector, and thus A and B are both 400-by-400, respectively positive semi-definite and positive definite. Furthermore, to avoid laborious tuning of timestep sizes, both A and B are normalized by their respective 2-norm; this is without loss of generality, because $\arg\min_Q \frac{\det(Q^T A Q)}{\det(Q^T B Q)}$ is invariant to scaling of A and/or B. Since there are 10 classes, $l = 9$ is chosen.

Note this is a positive semi-definite problem by construction. Some generalized eigenvalue methods require or prefer such a property (e.g., Oja flow (Yan et al., 1994)), but the proposed algorithms are indifferent to the definiteness (see Rmk.3.3).

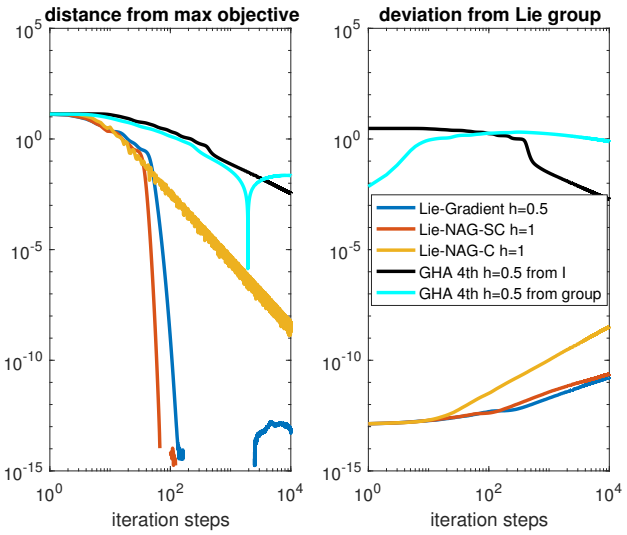

Figure 4: Lie-GD, Lie-NAG-C, Lie-NAG-SC, and GHA, for computing the leading l generalized eigenvalues associated with LDA for the MNIST dataset. All algorithms use step sizes tuned to minimize error in 10^4 iterations (although the proposed methods do not need much tuning). Two GHA runs use two initial conditions, $Q(0) = I$ which is not on the Lie group $Q^T B Q = I$, and $Q(0)$ being the first l columns of L^{-1} which is on the Lie group; all others use initial condition L^{-1}. GHA was based on Runge-Kutta-4 integration of $\dot{Q} = (I - BQQ^T)AQ$ for accuracy, and an Euler integration did not result in any notable error reduction. NAG-SC uses friction coefficient untuned $\gamma = 1$. The pollution of NAG simulations near the end is a machine precision artifact, and so are the deviations of Lie-NAGs and Lie-GD from the Lie group.

Fig.4 shows that all proposed methods converge significantly faster than GHA. Interestingly, although Lie-NAG-SC still converges faster than Lie-GD, the acceleration due to momentum is not as drastic as before.

In addition, Fig.5 in Appendix shows that our methods do not require an eigengap, and thus are widely applicable. Great methods have been continuously proposed for GEV; for instance, a globally linear convergent algorithm was recently proposed based on power method (Ge et al., 2016), but its convergence is affected by eigengap. The proposed methods do not have this restriction.

Acknowledgements

The authors thank Tuo Zhao and Justin Romberg for insightful discussions. Generous support from NSF DMS-1521667, DMS-1847802 and ECCS-1936776 (MT) and CMMI-1824798 (TO) are acknowledged.

References

Abdulle, A., Vilmart, G., and Zygalakis, K. C. (2014). High order numerical approximation of the invariant measure of ergodic SDEs. *SIAM Journal on Numerical Analysis*, 52(4):1600–1622.

Abraham, R. and Marsden, J. E. (1978). *Foundations of Mechanics*. Addison–Wesley, 2nd edition.

Abrudan, T., Eriksson, J., and Koivunen, V. (2009). Conjugate gradient algorithm for optimization under unitary matrix constraint. *Signal Processing*, 89(9):1704–1714.

Absil, P.-A., Mahony, R., and Sepulchre, R. (2009). *Optimization algorithms on matrix manifolds*. Princeton University Press.

Allen-Zhu, Z. and Li, Y. (2017). Doubly accelerated methods for faster CCA and generalized eigendecomposition. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 98–106. JMLR. org.

Arora, R., Marinov, T. V., Mianjy, P., and Srebro, N. (2017). Stochastic approximation for canonical correlation analysis. In *Advances in Neural Information Processing Systems*, pages 4775–4784.

Artstein, Z. and Infante, E. (1976). On the asymptotic stability of oscillators with unbounded damping. *Quarterly of Applied Mathematics*, 34(2):195–199.

Barnett, T. and Preisendorfer, R. (1987). Origins and levels of monthly and seasonal forecast skill for united states surface air temperatures determined by canonical correlation analysis. *Monthly Weather Review*, 115(9):1825–1850.

Bloch, A. M., Brockett, R. W., and Ratiu, T. S. (1992). Completely integrable gradient flows. *Communications in Mathematical Physics*, 147(1):57–74.

Bou-Rabee, N. and Owhadi, H. (2010). Long-run accuracy of variational integrators in the stochastic context. *SIAM Journal on Numerical Analysis*, 48(1):278–297.

Boumal, N., Absil, P.-A., and Cartis, C. (2018). Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, 39(1):1–33.

Brockett, R. W. (1989). Least squares matching problems. *Linear Algebra and its applications*, 122:761–777.

Brockett, R. W. (1991). Dynamical systems that sort lists, diagonalize matrices, and solve linear programming problems. *Linear Algebra and its Applications*, 146(0):79–91.

Chen, Z., Li, X., Yang, L., Haupt, J., and Zhao, T. (2019). On constrained nonconvex stochastic optimization: A case study for generalized eigenvalue decomposition. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 916–925.

Cheng, X., Chatterji, N. S., Bartlett, P. L., and Jordan, M. I. (2018). Underdamped Langevin MCMC: A non-asymptotic analysis. In *Conference On Learning Theory*, pages 300–323.

Chi, Y., Lu, Y. M., and Chen, Y. (2019). Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269.

Edelman, A., Arias, T. A., and Smith, S. T. (1998). The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353.

Fiedler, M. (1973). Algebraic connectivity of graphs. *Czechoslovak mathematical journal*, 23(2):298–305.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188.

Gabay, D. (1982). Minimizing a differentiable function over a differential manifold. *Journal of Optimization Theory and Applications*, 37(2):177–219.

Ge, R., Jin, C., Netrapalli, P., Sidford, A., et al. (2016). Efficient algorithms for large-scale generalized eigenvector computation and canonical correlation analysis. In *International Conference on Machine Learning*, pages 2741–2750.

Glaser, S. J., Schulte-Herbrüggen, T., Sieveking, M., Schedletzky, O., Nielsen, N. C., Sørensen, O. W., and Griesinger, C. (1998). Unitary control in quantum ensembles: Maximizing signal intensity in coherent spectroscopy. *Science*, 280(5362):421–424.

Gorrell, G. (2006). Generalized Hebbian algorithm for incremental singular value decomposition in natural language processing. *11th Conference of the European Chapter of the Association for Computational Linguistics*, page 8.

Hairer, E., Lubich, C., and Wanner, G. (2006). *Geometric numerical integration: structure-preserving algorithms for ordinary differential equations*, volume 31. Springer Science & Business Media.

Holm, D., Schmah, T., and Stoica, C. (2009). *Geometric mechanics and symmetry: from finite to infinite dimensions*. Oxford texts in applied and engineering mathematics. Oxford University Press.

Johnson, R. A., Wichern, D. W., et al. (2002). *Applied multivariate statistical analysis*, volume 5. Prentice hall Upper Saddle River, NJ.

Kitaev, A. and Watrous, J. (2000). Parallelization, amplification, and exponential time simulation of quantum interactive proof systems. In *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, pages 608–617.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., et al. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Lee, J. M. (2013). *Introduction to Smooth Manifolds*, volume 218 of *Graduate Studies in Mathematics*. Springer, 2nd edition.

Li, Q., Tai, C., and Weinan, E. (2019). Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations. *Journal of Machine Learning Research*, 20(40):1–40.

Li, T., Zhu, S., and Ogihara, M. (2006). Using discriminant analysis for multi-class classification: an experimental investigation. *Knowledge and information systems*, 10(4):453–472.

Liu, C., Zhuo, J., Cheng, P., Zhang, R., Zhu, J., and Carin, L. (2018). Accelerated first-order methods on the Wasserstein space for Bayesian inference. *arXiv preprint arXiv:1807.01750*.

Liu, Y., Shang, F., Cheng, J., Cheng, H., and Jiao, L. (2017). Accelerated first-order methods for geodesically convex optimization on Riemannian manifolds. In *Advances in Neural Information Processing Systems*, pages 4868–4877.

Ma, Y.-A., Chatterji, N., Cheng, X., Flammarion, N., Bartlett, P., and Jordan, M. I. (2019). Is there an analog of Nesterov acceleration for MCMC? *arXiv preprint arXiv:1902.00996*.

Mahony, R. and Manton, J. H. (2002). The geometry of the Newton method on non-compact Lie groups. *Journal of Global Optimization*, 23(3):309–327.

Marsden, J. E. and Ratiu, T. S. (2013). *Introduction to mechanics and symmetry: a basic exposition of classical mechanical systems*, volume 17. Springer Science & Business Media.

McLachlan, R. and Perlmutter, M. (2001). Conformal Hamiltonian systems. *Journal of Geometry and Physics*, 39(4):276–300.

McLachlan, R. I. and Quispel, G. R. W. (2002). Splitting methods. *Acta Numerica*, 11:341–434.

Nesterov, Y. (2018). *Lectures on convex optimization*, volume 137. Springer.

Nesterov, Y. E. (1983). A method for solving the convex programming problem with convergence rate O $(1/k^2)$. In *Dokl. akad. nauk Sssr*, volume 269, pages 543–547.

Oja, E. (1982). Simplified neuron model as a principal component analyzer. 15(3):267–273.

Patterson, S. and Teh, Y. W. (2013). Stochastic gradient riemannian langevin dynamics on the probability simplex. In *Advances in neural information processing systems*, pages 3102–3110.

Pavliotis, G. A. (2014). *Stochastic processes and applications: diffusion processes, the Fokker-Planck and Langevin equations*, volume 60. Springer.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407.

Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

Saad, Y. (2011). *Numerical methods for large eigenvalue problems: revised edition*, volume 66. SIAM.

Sanger, T. D. (1989). Optimal unsupervised learning in a single-layer linear feedforward neural network. 2(6):459–473.

Sattinger, D. H. and Weaver, O. L. (2013). *Lie groups and algebras with applications to physics, geometry, and mechanics*, volume 61. Springer Science & Business Media.

Shi, B., Du, S. S., Jordan, M. I., and Su, W. J. (2018). Understanding the acceleration phenomenon via high-resolution differential equations. *arXiv preprint arXiv:1810.08907*.

Smith, S. T. (1994). Optimization techniques on riemannian manifolds. *Fields institute communications*, 3(3):113–135.

Sorensen, O. W. (1989). Polarization transfer experiments in high-resolution nmr spectroscopy. *Progress in nuclear magnetic resonance spectroscopy*, 21.

Su, W., Boyd, S., and Candes, E. (2014). A differential equation for modeling Nesterovs accelerated gradient method: Theory and insights. In *Advances in Neural Information Processing Systems*, pages 2510–2518.

Tao, M. (2016). Explicit symplectic approximation of nonseparable Hamiltonians: Algorithm and long time performance. *Physical Review E*, 94(4):043303.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416.

Wang, Y. and Tao, M. (2020). Hessian-Free High-Resolution ODE for Nesterov Accelerated Gradient method. *preprint*.

Wei-Yong Yan, Helmke, U., and Moore, J. B. (1994). Global analysis of Oja's flow for neural networks. 5(5):674–683.

Welling, M. (2005). Fisher linear discriminant analysis. *Department of Computer Science, University of Toronto*, 3(1).

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

Wigner, E. P. (1958). On the distribution of the roots of certain symmetric matrices. *Ann. Math*, 67(2):325–327.

Yan, W.-Y., Helmke, U., and Moore, J. B. (1994). Global analysis of Oja's flow for neural networks. *IEEE Transactions on Neural Networks*, 5(5):674–683.

Zhang, H., Reddi, S. J., and Sra, S. (2016). Riemannian SVRG: Fast stochastic optimization on Riemannian manifolds. In *Advances in Neural Information Processing Systems*, pages 4592–4600.

Zhang, H. and Sra, S. (2016). First-order methods for geodesically convex optimization. In *Conference on Learning Theory*, pages 1617–1638.

Zhang, H. and Sra, S. (2018). Towards Riemannian accelerated gradient methods. *arXiv preprint arXiv:1806.02812*.

Appendix

Justification of NAG dynamics for GEV

This section justifies why one can simply use the same NAG flow for eigenvalue problem and only modify R's initial condition. It is rigorous when B is positive definite, since its Cholesky decomposition will be used; otherwise, the justification is formal, and the same NAG dynamics is still well defined.

First, rewrite (12) as

$$\max_{R\in\mathbb{R}^{n\times n}} \operatorname{tr}(E^T R^T A R E) \quad \text{s.t.} \quad R^T B R = I_{n\times n}.$$

Cholesky decompose B as $B = L^T L$, let $Q = LR$ and $\hat{A} = L^{-T} A L$, then the GEV is equivalently

$$\max_{Q\in\mathbb{R}^{n\times n}} \operatorname{tr}(Q^T \hat{A} Q \mathcal{E}) \quad \text{s.t.} \quad Q^T Q = I_{n\times n}.$$

One can write down the NAG dynamics for variationally optimizing this problem:

$$\dot{Q} = Q\xi, \qquad \dot{\xi} = -\gamma(t)\xi + [Q^T \hat{A} Q, \mathcal{E}]$$

Note this is

$$L\dot{R} = LR\xi, \qquad \dot{\xi} = -\gamma\xi + [R^T L^T L^{-T} A L^{-1} L R, \mathcal{E}],$$

and all L's can be canceled, leading to (2).

In terms of initial condition, since $Q(0)^T Q(0) = I$, $R(0)^T L^T L R(0) = R(0)^T B R(0) = I$. $\xi(0)$ needs to be skew-symmetric throughout.

Preservation of Lie group structure

(This section explicitly demonstrates several facts of geometric mechanics; for more information about geometric mechanics less in coordinates, see e.g., Marsden and Ratiu (2013); Holm et al. (2009).)

For continuous dynamics, we have

Theorem 4.1. *Consider $\dot{R}(t) = R(t)F(t)$ where R and F are n-by-n matrices. If $R(t_0)^T B R(t_0) = I$ and $F(t)$ is skew-symmetric for all $t \geq t_0$, then $R(t)^T B R(t) = I$, $\forall t \geq t_0$.*

Proof.

$$\frac{d}{dt}(R^T B R) = \dot{R}^T B R + R^T B \dot{R}$$
$$= F^T R^T B R + R^T B R F = F^T + F = 0. \qquad \square$$

Corollary 4.1. *We thus have Theorem 3.1.*

Proof. We only need to show $F := \xi(t)$ remains skew-symmetric. This is true because

$$\xi(t) = e^{-\Gamma(t)} \left(\xi(0) + \int_0^t e^{\Gamma(s)} [R(s)^T A R(s), \mathcal{E}] ds \right),$$

where $\Gamma(t) := \int_0^t \gamma(s) ds$ is a scalar. However, $\xi(0)$ is skew-symmetric by assumption, and so is the integrand because

$$[R(s)^T A R(s), \mathcal{E}]^T = [\mathcal{E}^T, (R(s)^T A R(s))^T]$$
$$= [\mathcal{E}, R(s)^T A R(s)] = -[R(s)^T A R(s), \mathcal{E}]. \qquad \square$$

Corollary 4.2. *Lie-GD $\dot{R} = R[R^T A R, \mathcal{E}]$ also maintains $R^T B R = I$.*

For discrete timesteppings, we have

Theorem 4.2. *Define Cayley transformation as Cayley$(\xi) := (I - \xi/2)^{-1}(I + \xi/2)$. Consider $\dot{R}(t) = R(t)F(t)$ where R and F are n-by-n matrices. If $R(t_0)^T B R(t_0) = I$ and $F(t_0)$ is skew-symmetric, then the discrete updates given by $\hat{R} = R(t_0)\exp(F(t_0)h)$ and $\hat{R} = R(t_0)\text{Cayley}(F(t_0)h)$ both satisfy $\hat{R}^T B \hat{R} = I$.*

Proof. Consider $\hat{R} = RQ$. If $Q^T Q = I$, then

$$\hat{R}^T B \hat{R} = Q^T R^T B R Q = Q^T Q = I.$$

$Q = \exp(Fh)$ for skew-symmetric F satisfies this condition because

$$Q^T Q = \exp(F^T h)\exp(Fh) = \exp(-Fh)\exp(Fh) = I.$$

$Q = \text{Cayley}(Fh)$ for skew-symmetric F satisfies this condition because

$$Q^T Q = (I + Fh/2)^T (I - Fh/2)^{-T} (I - Fh/2)^{-1} (I + Fh/2)$$
$$= (I - Fh/2)(I + Fh/2)^{-1} (I - Fh/2)^{-1} (I + Fh/2) = I$$

the last equality because $I - Fh/2$ and $I + Fh/2$ commute. $\square$

A brief recap of GHA

(This subsection is not new research but for the self-containment of the article.)

Oja flow / Sanger's rule / Generalized Hebbian Algorithm (e.g., Oja (1982); Sanger (1989); Gorrell (2006); Wei-Yong Yan et al. (1994)) is a celebrated type of methods based on continuous dynamics for finding leading eigenvalues of a symmetric matrix. Only for the reason of a concise presentation, we refer to them as GHA in this article.

GHA works as follows: given n-by-n symmetric A, to find the eigenspace associated with its largest l eigenvalues, one denotes by $V(t)$ an n-by-l matrix and uses the long time limit of dynamics

$$\dot{V} = (I - VV^T)AV$$

as a span of the corresponding orthonormal eigenvectors.

This approach can be extended to GEV (12) by using GHA dynamics

$$\dot{V} = (I - BVV^T)AV; \tag{17}$$

see e.g., Chen et al. (2019) and references therein.

To implement GHA in practice, the continuous dynamics need to be numerically discretized. A 1st-order discretization is based on Euler scheme, namely

$$V_{i+1} = V_i + h(I - BV_iV_i^T)AV_i,$$

and it is most commonly used. However, if a smaller deviation from the continuous dynamics is desired, a higher-order discretization can also be used, e.g., a 4th-order Runge-Kutta given by

$$\begin{aligned}
k_1 &= \left(I - BV_iV_i^T\right)AV_i \\
k_2 &= \left(I - B\left(V_i + \frac{h}{2}k_1\right)\left(V_i + \frac{h}{2}k_1\right)^T\right)A\left(V_i + \frac{h}{2}k_1\right) \\
k_3 &= \left(I - B\left(V_i + \frac{h}{2}k_2\right)\left(V_i + \frac{h}{2}k_2\right)^T\right)A\left(V_i + \frac{h}{2}k_2\right) \\
k_4 &= \left(I - B\left(V_i + hk_3\right)\left(V_i + hk_3\right)^T\right)A\left(V_i + hk_3\right) \\
V_{i+1} &= V_i + \frac{h}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right).
\end{aligned}$$

Roughly 4 times the flops of Euler are needed per step, but the deviation from (17) is $\mathcal{O}(h^4)$ instead of $\mathcal{O}(h)$ for Euler.

A brief recap of multiclass Fisher Linear Discriminant Analysis (LDA)

(This subsection is not new research but, for the self-containment of the article, a quick excerpt of the existing methods of Fisher Linear Discriminant Analysis Fisher (1936) and Multiple Discriminant Analysis (e.g., Johnson et al. (2002)), mainly based on Li et al. (2006)).

Given d-by-1 vectorial data x_i, $i = 1, \cdots, N$ labeled into M-classes, define 'inter-class scatter matrix' A and 'intra-class class scatter matrix' B by

$$\begin{aligned}
\mu_m &= \frac{1}{|\mathcal{C}_m|}\sum_{i\in\mathcal{C}_m} x_i, \\
\bar{x} &= \frac{1}{N}\sum_{i=1}^{N} x_i, \\
A &= \sum_{m=1}^{M}(\mu_m - \bar{x})(\mu_m - \bar{x})^T, \\
B &= \sum_{m=1}^{M}\sum_{i\in\mathcal{C}_m}(x_i - \mu_m)(x_i - \mu_m)^T,
\end{aligned}$$

where $\mathcal{C}_m$ is the set of indices corresponding to class-m. FDA seeks a projection represented by a d-by-l matrix Q that maximizes the Rayleigh quotient:

$$\max_Q \frac{\det(Q^TAQ)}{\det(Q^TBQ)},$$

where a standard choice of l is $l = M - 1$. This problem can be reformulated as the generalized eigenvalue problem $Aw = \lambda Bw$ (e.g., Li et al. (2006); Welling (2005)), and thus equivalent to

$$\begin{aligned}
&\max && \text{tr}\,(Q^TAQ) \\
&\text{s.t.} && Q^TBQ = I.
\end{aligned}$$

Additional LDA experimental results

To demonstrate that the proposed methods still work when there is no eigengap (i.e., two largest eigenvalues being identical), we take A and B from LDA for MNIST, Cholesky decompose B as $B = L^TL$, let $\hat{A} = L^{-T}AL^{-1}$, diagonalize $\hat{A} = VDV^{-1}$, and then replace D's largest diagonal element by the value of the 2nd largest. Denoting the result by $\tilde{D}$, we replace A by $\tilde{A} = L^TV\tilde{D}V^{-1}L$. The generalized eigenvalue problem associated with $\{\tilde{A}, B\}$ now has a zero eigengap, which prevents, for example, power-method based approaches from working. However, Fig. 5 shows that the proposed methods perform almost identically to the original $\{A, B\}$ case (c.f., Fig. 4).

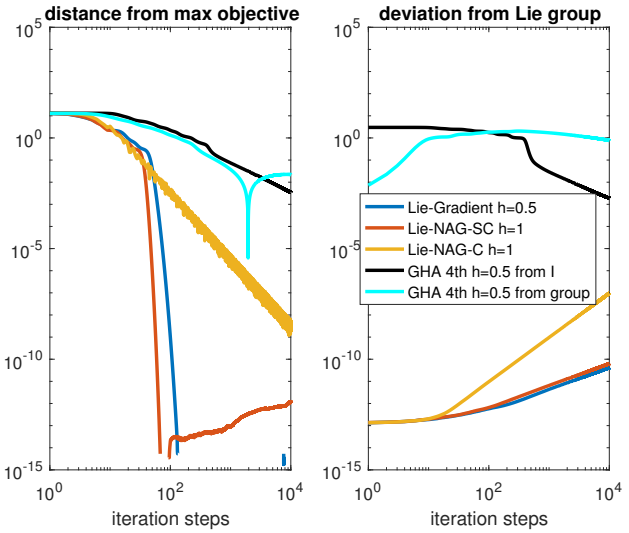

Figure 5: Same experiment as in Fig.4 for modified MNIST with 0 eigengap.

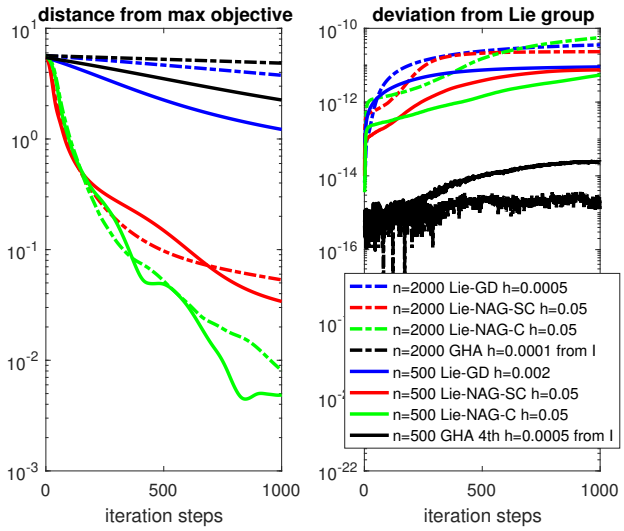

Figure 6: The computation of leading $l = 2$ eigenvalues of 2000-dimensional scaled GOE, compared with that for 500-dimension. Other descriptions are same as in Fig.1.

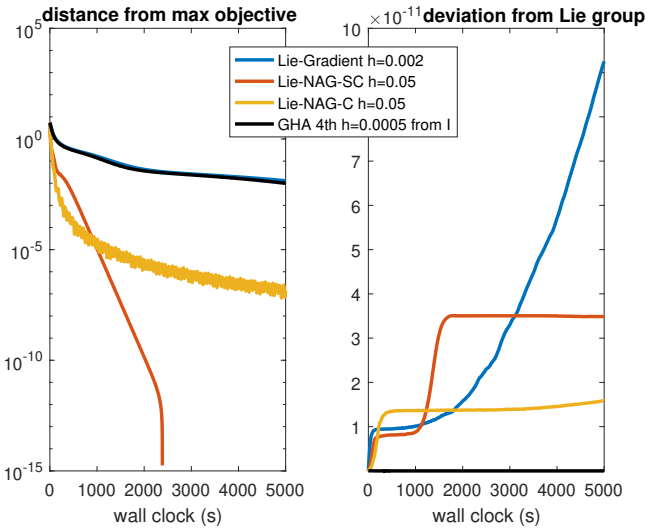

Figure 7: The computation of leading $l = 2$ eigenvalues of 500-dimensional scaled GOE. All descriptions are same as in Fig.1, except that x-axis is no longer in iteration steps but in wallclock.

l largest eigenvalues of $A = (\Xi + \Xi^T)/2/\sqrt{n}$: $n = 2000$ result

Fig.6 describes the same experiment as in Sec.4.1.1 when the dimension is $n = 2000$ instead of 500. When compared with the $n = 500$ case, one sees Lie-GD and GHA converge much slower, but Lie-NAG's converge only marginally slower. This suggests that the advantage of variational methods increases in higher dimension, at least in this experiment.

l largest eigenvalues of $A = (\Xi + \Xi^T)/2/\sqrt{n}$: $n = 500$ result in wallclock count

Fig.7 illustrates the actual computational costs of methods used in this paper by reproducing Fig.1 with x-axis replaced by the time it took for each method to run. All qualitative conclusions remain unchanged. Experiments were conducted on a 4th-gen Intel Core laptop with integrated graphics unit running 64-bit Windows 7 and MATLAB R2016b.

Two 4th-order versions of Lie-NAG algorithms

Version 1: more accurate but more computation

$$\phi^h = \phi_2^{a_1 h} \circ \phi_1^{b_1 h} \circ \phi_2^{a_2 h} \circ \phi_1^{b_2 h} \circ \phi_2^{a_3 h} \circ \phi_1^{b_3 h} \circ \phi_2^{a_4 h} \circ \phi_1^{b_3 h} \circ \phi_2^{a_3 h} \circ \phi_1^{b_2 h} \circ \phi_2^{a_2 h} \circ \phi_1^{b_1 h} \circ \phi_2^{a_1 h} + \mathcal{O}(h^5)$$

where

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} = \begin{bmatrix} 0.079203696431196 \\ 0.353172906049774 \\ -0.042065080357719 \\ 0.219376955753500 \end{bmatrix},$$

$$\begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 0.209515106613362 \\ -0.143851773179818 \\ 0.434336666566456 \end{bmatrix}.$$

Version 2: less accurate but less computation

$$\phi_2^{a_1 h} \circ \phi_1^{b_1 h} \circ \phi_2^{a_2 h} \circ \phi_1^{b_2 h} \circ \phi_2^{a_2 h} \circ \phi_1^{b_1 h} \circ \phi_2^{a_1 h}$$

where

$$\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \gamma_4/2 \\ (1-\gamma_4)/2 \end{bmatrix}, \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} \gamma_4 \\ 1-2\gamma_4 \end{bmatrix}, \gamma_4 = \frac{1}{2 - 2^{1/3}}.$$

Details can be found, e.g., in McLachlan and Quispel (2002). Swapping ϕ_1 and ϕ_2 will yield additional methods at the same order of accuracy. We present the above because ϕ_1 is computationally more costly due to Cayley transform.

Some heuristic insights on the correction of the NAG dissipation coefficient in SG context

Based on the discussion in the main text, heuristically, large γ values correspond to lower 'temperatures' and reduced variances accumulated from stochastic gradients. However, they also slow down the convergences of the stochastic processes, and yet we'd like to

take advantage of the fast convergence of deterministic NAG dynamics. Therefore, we consider an additive correction that is small for small t and increasing to infinity.

For simplicity, restrict the correction to be a monomial of t, i.e., $\delta\gamma = ct^p$. Then we select the value of p by resorting to intuitions first gained from a linear deterministic case, for which our choice of p has to lead to convergence because the deterministic solution is the mean of the stochastic solution. It is proved in Artstein and Infante (1976) that a sufficient condition for asymptotic stability of $\ddot{q} + \gamma(t)\dot{q} + q = 0$ is

$$\limsup_{T\to\infty}\left(\frac{1}{T^2}\int_0^T \gamma(t)dt\right) < \infty \quad \text{and} \quad \gamma(t) \geq \gamma_0$$

for some constant $\gamma_0 > 0$. It is easy to check that $\gamma(t) = \gamma_0 + ct^p$ or $3/t + ct^p$ satisfies this condition if $p \leq 1$, but not when $p > 1$. We thus inspect the boundary case of $p = 1$ for a fast decay of variance at large t, now in a stochastic setup:

$$\begin{cases} dq &= pdt \\ dp &= (-(\gamma_0(t) + ct)p - q)dt + \sigma dW \end{cases}, \tag{18}$$

where γ_0 is either a constant or $3/t$. Since this is a linear SDE whose solution is Gaussian, it suffices to show the convergences of the (deterministic) mean and covariance evolutions in order to establish the SDE's convergence.

It is standard to show the mean $x(t) := \mathbb{E}[q(t), p(t)]$ satisfies a closed non-autonomous ODE system, and the covariance $V(t) := \mathbb{E}\big[[q(t) - \mathbb{E}[q(t)],\ p(t) - \mathbb{E}[p(t)]]^T[q(t)-\mathbb{E}[q(t)],\ p(t)-\mathbb{E}[p(t)]]\big]$ satisfies another. These systems are not analytically solvable, but we can analyze their long time behavior by asymptotic analysis.

More precisely, under the ansatz of $\mathbb{E}[q] = bt^a + o(t^a)$, matching leading order terms in the mean ODE leads to

$$\mathbb{E}[q(t)] \sim t^{-1/c}, \qquad \mathbb{E}[p(t)] \sim t^{-1/c-1}$$

for both constant γ_0 and $\gamma_0(t) = 3/t$ in (18).

Under the ansatz of $\mathrm{Var}[q] = b_1 t^{a_1} + o(t^{a_1})$, $\mathrm{Var}[p] = b_2 t^{a_2} + o(t^{a_2})$, $\mathbb{E}[(q - \mathbb{E}q)(p - \mathbb{E}p)] = b_3 t^{a_3} + o(t^{a_3})$, matching leading order terms in the covariance ODE leads to

$$\mathrm{Var}[q] = \frac{1}{c(2-c)}t^{-1}, \quad \mathrm{Var}[p] = \frac{1}{2c}t^{-1},$$
$$\mathbb{E}[(q - \mathbb{E}q)(p - \mathbb{E}p)] = \frac{1}{2c(c-2)}t^{-2}.$$

Note this means, for small but positive c, convergence is guaranteed, and covariance converges slower than mean, at the rate independent of c.

Therefore, adding ct to γ in the original NAG's works in the linear case, and thus it has a potential to work for nonlinear cases (e.g., Lie group versions). And it does in experiments (Sec.4.2).

Hamiltonian Formulation

In this section, we give a Hamiltonian formulation of the variational optimization equation (7) and prove the conformal symplecticity of its flow.

Symplectic Structure on $\mathsf{G} \times \mathfrak{g}^*$
Let λ be the left trivialization of $T^*\mathsf{G}$, i.e.,

$$\lambda\colon T^*\mathsf{G} \to \mathsf{G} \times \mathfrak{g}^*; \qquad p_g \mapsto (g, T_e^* L_g(p_g))\,.$$

Then its inverse is given by

$$\lambda^{-1}\colon \mathsf{G} \times \mathfrak{g}^* \to T^*\mathsf{G}; \qquad (g,\mu) \mapsto T_g^* L_{g^{-1}}(\mu).$$

Let Θ and $\Omega := -\mathbf{d}\Theta$ be the canonical one-form and the symplectic structure on $T^*\mathsf{G}$, and θ and ω be their pull-backs via the left trivialization, i.e.,

$$\theta := (\lambda^{-1})^*\Theta, \qquad \omega := (\lambda^{-1})^*\Omega.$$

According to Abraham and Marsden (1978, Proposition 4.4.1 on p. 315) (see also the reference therein), for any $(g,\mu) \in \mathsf{G} \times \mathfrak{g}^*$ and any $(v,\alpha),(w,\beta) \in T_{(g,\mu)}(\mathsf{G} \times \mathfrak{g}^*)$,

$$\theta_{(g,\mu)}(w,\beta) = \big\langle \mu, T_g L_{g^{-1}}(w)\big\rangle \tag{19}$$

and

$$\begin{aligned}\omega_{(g,\mu)}&((v,\alpha),(w,\beta)) \\ &= \big\langle \beta, T_g L_{g^{-1}}(v)\big\rangle - \big\langle \alpha, T_g L_{g^{-1}}(w)\big\rangle \\ &\quad + \big\langle \mu, [T_g L_{g^{-1}}(v), T_g L_{g^{-1}}(w)]\big\rangle.\end{aligned} \tag{20}$$

Given a function $h\colon \mathsf{G} \times \mathfrak{g}^* \to \mathbb{R}$, the corresponding Hamiltonian vector field $X_h \in \mathfrak{X}(\mathsf{G} \times \mathfrak{g}^*)$ defined by $\mathbf{i}_{X_h}\omega = \mathbf{d}h$ is given by

$$X_h(g,\mu) = \left(T_e L_g\left(\frac{\delta h}{\delta \mu}\right),\ \mathrm{ad}^*_{\frac{\delta h}{\delta \mu}}\,\mu - T_e^* L_g(\mathbf{d}_g h)\right),$$

where $\mathbf{d}_g$ stands for the exterior differential with respect to g.

Legendre Transform and Hamiltonian Formulation
We may apply a time-independent Legendre transform using the initial Lagrangian as follows: Let us define the initial Lagrangian $L_0\colon \mathsf{G} \times \mathfrak{g} \to \mathbb{R}$ by setting $L_0(g,\xi) := L(g,\xi,0)$, and the time-independent Legendre transform

$$\mathbb{F}L_0\colon \mathfrak{g} \to \mathfrak{g}^*; \qquad \xi \mapsto \frac{\delta L_0}{\delta \xi}(g,\xi,t) = r(0)\,\mathbb{I}(\xi),$$

whose inverse is given by

$$(\mathbb{F}L_0)^{-1}\colon \mathfrak{g}^* \to \mathfrak{g}; \qquad \mu \mapsto \frac{1}{r(0)}\mathbb{I}^{-1}(\mu).$$

We define the initial Hamiltonian $H\colon \mathsf{G} \times \mathfrak{g}^* \to \mathbb{R}$ as follows:

$$\begin{aligned} H(g,\mu) &:= \big\langle \mu, (\mathbb{F}L_0)^{-1}(\mu)\big\rangle - L_0\big(g, (\mathbb{F}L_0)^{-1}(\mu)\big) \\ &= \frac{1}{2r(0)}\big\langle \mu, \mathbb{I}^{-1}(\mu)\big\rangle + r(0)f(g). \end{aligned}$$

Its associated Hamiltonian vector field X_H on $\mathfrak{g}^*$ is defined as $\mathbf{i}_{X_H}\omega = \mathbf{d}H$ using the symplectic form ω on $\mathsf{G} \times \mathfrak{g}^*$ (see (20)):

$$X_H(\mu) = \mathrm{ad}^*_{\frac{\delta H}{\delta \mu}}\,\mu - T_e^*\mathsf{L}_g(\mathbf{d}_g H).$$

Then we may rewrite (7) as follows:

$$\begin{aligned} \dot\mu &= -\gamma(t)\mu + \mathrm{ad}^*_{\frac{\delta H}{\delta \mu}}\,\mu - T_e^*\mathsf{L}_g(\mathbf{d}_g H) \\ &= X_H(\mu) - \gamma(t)\mu, \end{aligned} \tag{21}$$

where we set $\gamma(t) := r'(t)/r(t)$.

Conformal Symplecticity

Given the Lagrangian of the form $r(t)L_0(q,\dot q)$, the Euler–Lagrange equation is

$$\frac{d}{dt}\left(r(t)\frac{\partial L_0}{\partial \dot q}\right) - r(t)\frac{\partial L_0}{\partial q} = 0. \tag{22}$$

We would like to show that the two-form $r(t)\mathbf{d}p \wedge \mathbf{d}q$ with $p := \partial L_0/\partial \dot q$ is preserved in time in two different ways. The first is based on the variational principle: Consider

$$\mathbf{dd}\int_{t_0}^{t_1} r(t)L_0(q,\dot q)dt,$$

which is obviously 0 because any exact form is closed. On the other hand, it is the same as (due to integration by parts)

$$\mathbf{d}\left(\int_{t_0}^{t_1}\left(r\frac{\partial L_0}{\partial q}\mathbf{d}q - \frac{d}{dt}\left(r\frac{\partial L_0}{\partial \dot q}\right)\mathbf{d}q\right)dt + r\frac{\partial L_0}{\partial \dot q}\mathbf{d}q\Big|_{t_0}^{t_1}\right)$$

The first term is zero because of (22). Therefore,

$$0 = \mathbf{d}\left(r\frac{\partial L_0}{\partial \dot q}\mathbf{d}q\Big|_{t_0}^{t_1}\right) = \mathbf{d}(rp\mathbf{d}q)|_{t_0}^{t_1} = r\mathbf{d}p \wedge \mathbf{d}q|_{t_0}^{t_1}$$

The second proof uses the Hamiltonian formulation. We may write the Hamiltonian system corresponding to the Euler–Lagrange equation for the Lagrangian of the form $r(t)L_0(q,\dot q)$ as follows:

$$\dot q = \frac{\partial H}{\partial p}, \qquad \dot p = -\frac{\partial H}{\partial q} - \gamma(t)p, \tag{23}$$

where the Hamiltonian H is obtained via the Legendre transform of $L_0(q,\dot q)$ not $r(t)L_0(q,\dot q)$.

In what follows, we would like to generalize the work of McLachlan and Perlmutter (2001)—in which γ is set to be constant—to derive the conformal symplecticity of dissipative Hamiltonian systems of the above type. Let P be an (exact) symplectic manifold with symplectic form $\Omega = -\mathbf{d}\Theta$ and $H\colon P \to \mathbb{R}$ be a (time-independent) Hamiltonian. Let us define a time-dependent vector field $X_{H,(\cdot)}\colon \mathbb{R} \times P \to TP$ by defining, for any $t \in \mathbb{R}$, a vector field $X_{H,t}$ on P by setting

$$X_{H,t} := X_H - Z_t,$$

where X_H is the Hamiltonian vector field on P defined by

$$\mathbf{i}_{X_H}\Omega = \mathbf{d}H,$$

and the time-dependent vector field $Z_{(\cdot)}\colon \mathbb{R}\times P \to TP$ is defined as follows: Let $\Omega_{(\cdot)}$ be the time-dependent symplectic form on P defined as, for any $t \in \mathbb{R}$,

$$\Omega_t := r(t)\Omega.$$

We define Z_t by setting

$$\mathbf{i}_{Z_t}\Omega_t = -r'(t)\Theta.$$

In terms of the canonical coordinates (q,p) for P, we have

$$Z_t = p_i\frac{\partial}{\partial p_i},$$

and hence we have

$$X_{H,t}(q,p) = \frac{\partial H}{\partial p_i}\frac{\partial}{\partial q^i} + \left(\frac{\partial H}{\partial q^i} + \gamma(t)p_i\right)\frac{\partial}{\partial p_i}.$$

Therefore, $X_{H,t}$ yields the dissipative Hamiltonian system (23).

Let $\Phi\colon \mathbb{R}\times\mathbb{R}\times P \to P$ be the time-dependent flow of $X_{H,(\cdot)}$ (assuming for simplicity that the solutions exist for any time $t \in \mathbb{R}$ with any initial time $t_0 \in \mathbb{R}$). Then, for any $t_0, t_1 \in \mathbb{R}$ (see, e.g., Lee (2013, Proposition 22.15)),

$$\begin{aligned} &\frac{d}{dt}\Phi_{t,t_0}^*\Omega_t\Big|_{t=t_1} \\ &= \Phi_{t_1,t_0}^*\left(\frac{\partial}{\partial t}\Omega_t\Big|_{t=t_1} + \mathcal{L}_{X_{H,t_1}}\Omega_{t_1}\right) \\ &= \Phi_{t_1,t_0}^*(r'(t_1)\Omega + \mathcal{L}_{X_H}\Omega_{t_1} + \mathcal{L}_{Z^{t_1}}\Omega_{t_1}) \\ &= \Phi_{t_1,t_0}^*\big(r'(t_1)\Omega + r(t_1)\mathcal{L}_{X_H}\Omega + r(t_1)\mathcal{L}_{Z_{t_1}}\Omega\big) \\ &= \Phi_{t_1,t_0}^*\big(r'(t_1)\Omega - r(t_1)\big(\mathbf{d}\mathbf{i}_{Z_{t_1}}\Omega + \mathbf{i}_{Z_{t_1}}\mathbf{d}\Omega\big)\big) \\ &= \Phi_{t_1,t_0}^*\big(r'(t_1)\Omega - \mathbf{d}\mathbf{i}_{Z_{t_1}}\Omega_{t_1}\big) \\ &= \Phi_{t_1,t_0}^*(r'(t_1)\Omega - \mathbf{d}(-r'(t_1)\Theta)) \\ &= \Phi_{t_1,t_0}^*(r'(t_1)\Omega + r'(t_1)\mathbf{d}\Theta) \\ &= 0. \end{aligned}$$

Therefore, we have

$$\Phi^*_{t_1,t_0}\Omega_{t_1} = \Omega_{t_0}. \tag{24}$$

Now, (21) is a special case of the above setting. Specifically, we may define a time-dependent vector field $Z_{(\cdot)}\colon \mathbb{R} \times (\mathsf{G} \times \mathfrak{g}^*) \to T(\mathsf{G} \times \mathfrak{g}^*)$ by setting, for any $t \in \mathbb{R}$,

$$\mathbf{i}_{Z_t}\omega_t = -r'(t)\theta,$$

where $\omega_t := r(t)\omega$. This yields $Z_t(\mu) = \gamma(t)\mu$. Then we may write (21) as

$$\dot{\mu}(t) = (X_H - Z_t)(\mu(t)).$$

Let $\varphi\colon \mathbb{R}\times\mathbb{R}\times(\mathsf{G}\times\mathfrak{g}^*) \to \mathsf{G}\times\mathfrak{g}^*$ be the time-dependent flow of this system. Then, the conformal symplecticity (24) implies that, for any $t_0, t_1 \in \mathbb{R}$,

$$\varphi^*_{t,t_0}\omega_t = \omega_{t_0}.$$