# 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).

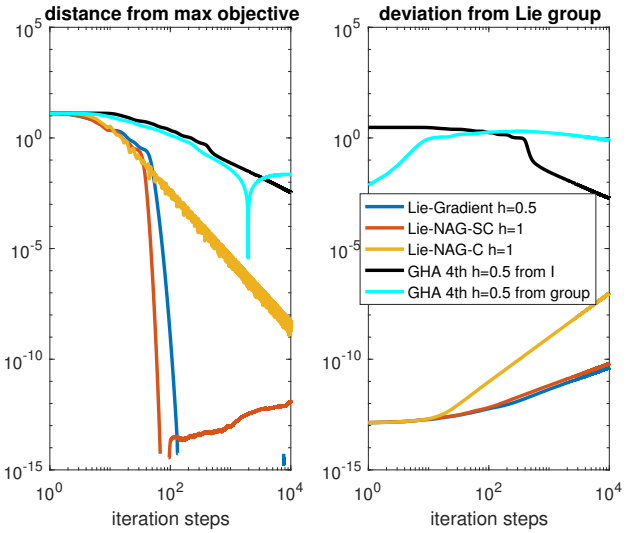


Figure 5: Same experiment as in Fig.4 for modified MNIST with 0 eigengap.

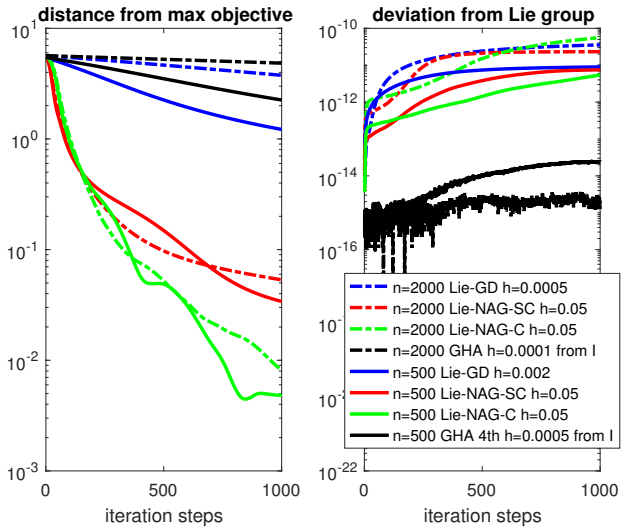


Figure 6: The computation of leading $l = 2$ eigenvalues of 2000-dimensional scaled GOE, compared with that for 500-dimension. Other descriptions are same as in Fig.1.

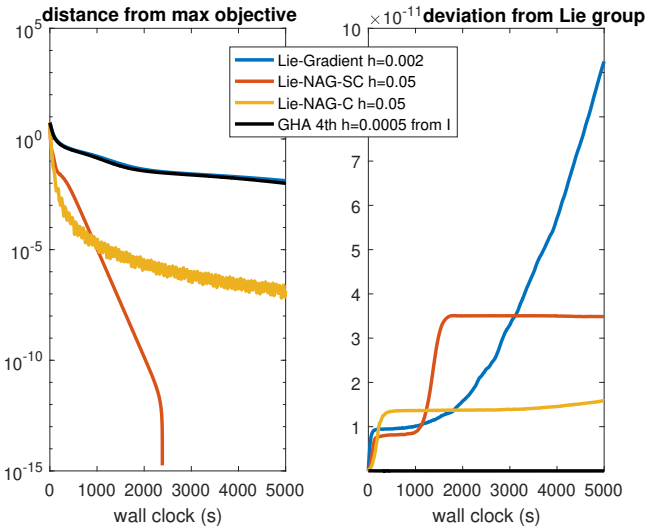


Figure 7: The computation of leading $l = 2$ eigenvalues of 500-dimensional scaled GOE. All descriptions are same as in Fig.1, except that x-axis is no longer in iteration steps but in wallclock.

### $l$ 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 $\phi_1$ and $\phi_2$ will yield additional methods at the same order of accuracy. We present the above because $\phi_1$ 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 $\gamma$ 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 $\gamma_0$ 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 $\gamma_0$ 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 $\gamma$ 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 $\lambda$ 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 $\Theta$ and $\Omega := -\mathbf{d}\Theta$ be the canonical one-form and the symplectic structure on $T^*\mathsf{G}$, and $\theta$ and $\omega$ 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 $\omega$ 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 $\gamma$ 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 $t_0, t_1 \in \mathbb{R}$,

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