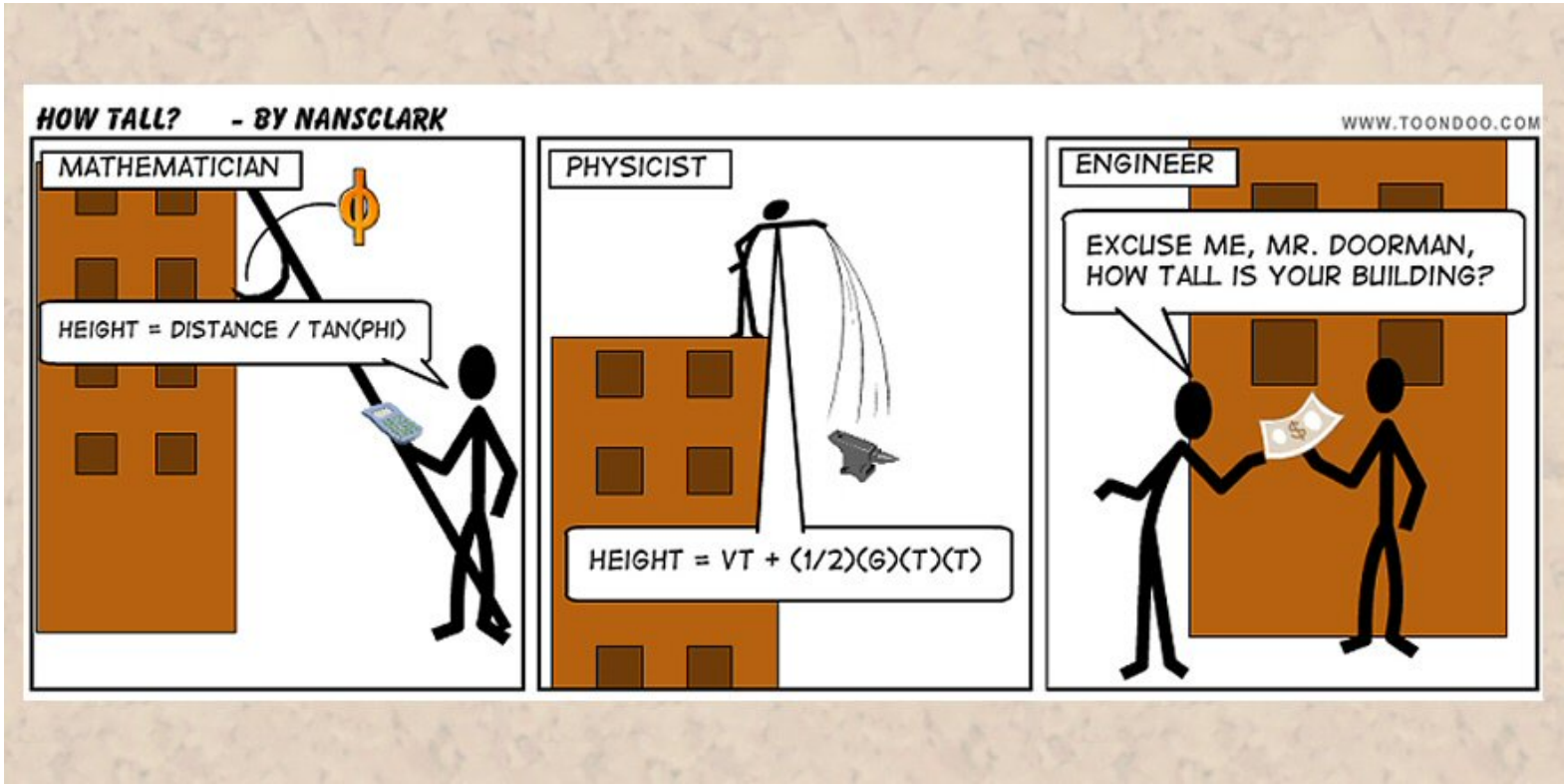
"Metode de optimizare Riemanniene pentru învățare profundă"
Proiect cofinanțat din Fondul European de Dezvoltare Regională prin
Programul Operațional Competitivitate 2014-2020

Mechanics over the Probability Simplex

Goffredo Chirco[1], Luigi Malagò[1], Giovanni Pistone[2]

[1] Romanian Institute of Science and Technology
[2] Collegio Carlo Alberto

SPIG'20 July 28, 2020


HOW TALL? - BY NANSCLARK
WWW.TOONDOO.COM
MATHEMATICIAN
HEIGHT = DISTANCE / TAN(PHI)
PHYSICIST
HEIGHT = VT + (1/2)(G)(T)(T)
ENGINEER
EXCUSE ME, MR. DOORMAN, HOW TALL IS YOUR BUILDING?

Outline of the Talk

Non-parametric Information Geometry

Mechanics of the Simplex: Lagrangians and Hamiltonians

Kullback-Leibler Lagrangian

Some examples

Presentation based on a work-in-progress paper

G. Chirco, L. Malagò, and G. Pistone.
Lagrangian and Hamiltonian Mechanics for Probabilities on the Statistical Manifold, 2020. To appear on the arXiv (soon).

Introduction 1/2

The framework of classical Mechanics is a finite-dimensional Riemannian manifold, e.g. L.D. Landau and E.M. Lifshits (1976), V.I. Arnold (1989), R. Abraham and J.E. Mardsen (1978) and by J.E. Marsden and T.S. Ratiu (1999).

Information Geometry (IG), as firstly defined by S.-I Amari (1982, 2000, 2016), views parametric statistical models as Riemannian manifolds, at the same time described as affine manifolds endowed with a dually-flat connection. See also the monograph by N. Ay, J. Jost, H.V. Lê, and L. Schwachhöfer (2017).

Recently, some authors have started to inquire about the relation between the geometry of classical Mechanics and Information Geometry (Leok and Zhang, 2017, Lods and Pistone, 2015, and Pistone, 2018).

Introduction 2/2

We use an approach based on non-parametric Information Geometry developed by Pistone and collaborators (P. and Sempi, 1995, Gibilisco and P., 1998, P. and Rogantin 1999, P. 2013, 2020)

We restrict our analysis to a finite dimensional state space, in order to avoid technical issues related to the infinite dimensional space modelization

We aim to study the relation between the geometry of Classical Mechanics and Information Geometry, with an emphasis on providing a statistical intuition of the geometric quantities involved.

The continuous evolution of probability functions has been growing a great interest in several areas, such as Population Dynamics, Differential Games, Optimization Methods and Machine Learning and more recently.

Maximal Exponential Family

We consider a finite sample space Ω with cardinality N.

Let $\Delta(\Omega)$ be the probability simplex, and $\Delta^\circ(\Omega)$ its interior.

We denote with μ the uniform probability function $1/N$.

The maximal exponential family $\mathcal{E}(\mu)$ is the set of densities which can be written as $p \propto e^f$, where f is defined up to a constant

Given a reference density $p \in \mathcal{E}(\mu)$, we have

$$q(x) = \exp(v(x) + H(v)) \cdot p(x), \quad \mathbb{E}[v(x)] = 0,$$
$$H(v) = -\log \mathbb{E}_q[e^v] = \mathrm{D}\left(p \,\|\, q\right)$$

with

$$v = \log \frac{q}{p} - \mathbb{E}_q\left[\log \frac{q}{p}\right] .$$

Statistical Bundle

The exponential statistical bundle with base Ω is defined as

$$S\mathcal{E}(\mu) = \{(q, v) \,|\, q \in \mathcal{E}(\mu), \mathbb{E}_q[v] = 0\} \ ,$$

we denote with ${}^*S_q\mathcal{E}(\mu)$ the dual statistical bundle. For finite Ω, $S_q\mathcal{E}(\mu)$ and ${}^*S_q\mathcal{E}(\mu)$ coincide.

A duality mapping between the statistical bundle and its dual the can be defined at the fiber at q by

$${}^*S_q\mathcal{E}(\mu) \times S_q\mathcal{E}(\mu) \ni (\eta, v) \mapsto \langle \eta, v \rangle_q = \mathbb{E}_q[\eta v] \ .$$

Two different affine geometries can be define for $S_q\mathcal{E}(\mu)$ and ${}^*S_q\mathcal{E}(\mu)$., by defining two different transports for each $p, q \in \mathcal{E}(\mu)$, i.e.,

Exponential transport: ${}^{\mathrm{e}}\mathbb{U}_p^q : S_p\mathcal{E}(\mu) \to S_q\mathcal{E}(\mu)$, ${}^{\mathrm{e}}\mathbb{U}_p^q v = v - \mathbb{E}_q[v]$

Mixture transport: ${}^{\mathrm{m}}\mathbb{U}_p^q : {}^*S_p\mathcal{E}(\mu) \to {}^*S_q\mathcal{E}(\mu)$, ${}^{\mathrm{m}}\mathbb{U}_p^q \eta = \frac{p}{q}\eta$.

Duality Between Transports

The two transports defined above are conjugate with respect to the duality pairing,

$$\left\langle {}^{\mathsf{m}}\mathbb{U}_p^q \eta, v \right\rangle_q = \left\langle \eta, {}^{\mathsf{e}}\mathbb{U}_q^p v \right\rangle_p \ , \quad \eta \in {}^*S_p\,\mathcal{E}\,(\mu)\,, v \in S_q\,\mathcal{E}\,(\mu) \ .$$

Moreover, it holds

$$\left\langle {}^{\mathsf{m}}\mathbb{U}_p^q \eta, {}^{\mathsf{e}}\mathbb{U}_p^q v \right\rangle_q = \langle \eta, v \rangle_p \ , \quad \eta \in {}^*S_p\,\mathcal{E}\,(\mu)\,, v \in S_p\,\mathcal{E}\,(\mu) \ .$$

Exponential Atlas

The *exponential atlas* of the exponential statistical bundle $S\mathcal{E}(\mu)$ is the collection of charts given for each $p \in \mathcal{E}(\mu)$ by

$$s_p \colon S\mathcal{E}(\mu) \ni (q, v) \mapsto (s_p(q), {}^{\mathrm{e}}\mathbb{U}_q^p v) \in S_p\mathcal{E}(\mu) \times S_p\mathcal{E}(\mu) \ ,$$

where

$$s_p(q) = \log \frac{q}{p} - \mathbb{E}_p\left[\log \frac{q}{p}\right] \ .$$

As $s_p(p, v) = (0, v)$, we say that s_p is the chart *centered at* p. The *cumulant function* K_p is defined on $S_p\mathcal{E}(\mu)$ by

$$K_p(u) = \log \mathbb{E}_p\left[\mathrm{e}^u\right] = \mathbb{E}_p\left[\log \frac{p}{q}\right] = \mathrm{D}(p \,\|\, q) \ ,$$

that is, $K_p(u)$ is the expression in the chart at p of Kullback-Leibler divergence of $q \mapsto \mathrm{D}(p \,\|\, q)$, and we can write

$$q = \mathrm{e}^{u - K_p(u)} \cdot p = \mathrm{e}_p(u) \ .$$

The *patch centered at* p is

$$s_p^{-1} = e_p \colon (S_p\mathcal{E}(\mu))^2 \ni (u, v) \mapsto (\mathrm{e}_p(u), {}^{\mathrm{e}}\mathbb{U}_p^{\mathrm{e}_p(u)} v) \in S\mathcal{E}(\mu) \ .$$

Dual Atlas

The *dual atlas* of the mixture statistical bundle $S\mathcal{E}(\mu)$ is the collection of charts given for each $p \in \mathcal{E}(\mu)$ by

$$\eta_p \colon {}^*S\mathcal{E}(\mu) \ni (q, w) \mapsto \left(s_p(q), {}^{\mathsf{m}}\mathbb{U}_q^p w\right) \in S_p\mathcal{E}(\mu) \times {}^*S_p\mathcal{E}(\mu) \ .$$

We say that η_p is the chart *centered at* p. The *patch centered at* p is

$$\eta_p^{-1} \colon S_p\mathcal{E}(\mu) \times \times {}^*S_p\mathcal{E}(\mu) \ni (u, v) \mapsto \left(e_p(u), {}^{\mathsf{m}}\mathbb{U}_p^{e_p(u)} v\right) \in {}^*S\mathcal{E}(\mu) \ .$$

Statistical Manifolds are Hessian Manifolds

The base manifold $\mathcal{E}(\mu)$ is actually an Hessian manifold with respect to any of the convex functions $K_p(u) = \log \mathbb{E}_p[\mathrm{e}^u]$, $u \in S_p\mathcal{E}(\mu)$, see H. Shima's (2007) monograph.

Some properties which can be easily checked

$$\mathbb{E}_{\mathrm{e}_p(u)}[h] = dK_p(u)[h]$$

$${}^{\mathsf{e}}\mathbb{U}_p^{\mathrm{e}_p(u)} h = h - dK_p(u)[h]$$

$$d^2K_p(u)[h_1, h_2] = \left\langle {}^{\mathsf{e}}\mathbb{U}_p^{\mathrm{e}_p(u)} h_1, {}^{\mathsf{e}}\mathbb{U}_p^{\mathrm{e}_p(u)} h_2 \right\rangle_{\mathrm{e}_p(u)}$$

$$d^3K_p(u)[h_1, h_2, h_3] = \mathbb{E}_{\mathrm{e}_p(u)}\left[({}^{\mathsf{e}}\mathbb{U}_p^{\mathrm{e}_p(u)} h_1)({}^{\mathsf{e}}\mathbb{U}_p^{\mathrm{e}_p(u)} h_2)({}^{\mathsf{e}}\mathbb{U}_p^{\mathrm{e}_p(u)} h_3) \right]$$

Velocities to a Curve 1/2

Let us compute the expression of the velocity at time t of a smooth curve

$$t \mapsto \gamma(t) = (q(t), w(t)) \in S\mathcal{E}(\mu)$$

in the exponential chart centered at p. The expression of the curve is

$$\gamma_p(t) = \left(s_p(q(t)), {}^{\mathsf{e}}\mathbb{U}^p_{q(t)} w(t)\right) ,$$

and hence we have, by denoting the ordinary derivative of a curve in $\mathbb{R}^N$ by the dot,

$$\frac{d}{dt} s_p(q(t)) = \frac{d}{dt}\left(\log \frac{q(t)}{p} - \mathbb{E}_p\left[\log \frac{q(t)}{p}\right]\right) = \frac{\dot{q}(t)}{q(t)} - \mathbb{E}_p\left[\frac{\dot{q}(t)}{q(t)}\right] =$$

$${}^{\mathsf{e}}\mathbb{U}^p_{q(t)} \frac{\dot{q}(t)}{q(t)} = {}^{\mathsf{e}}\mathbb{U}^p_{q(t)} \frac{d}{dt} \log q(t) ,$$

and

$$\frac{d}{dt} {}^{\mathsf{e}}\mathbb{U}^p_{q(t)} w(t) = \frac{d}{dt}\left(w(t) - \mathbb{E}_p[w(t)]\right) = \dot{w}(t) - \mathbb{E}_p[\dot{w}(t)] .$$

Velocities to a Curve 2/2

We express the tangent at each time t in the moving frame centered at the position $q(t)$ of the curve itself. Because of that, we define the *velocity* of the curve

$$t \mapsto q(t) = \mathrm{e}^{u(t) - K_p(u(t))} \cdot p \ , \quad u(t) = s_p(q(t)) \ ,$$

to be

$$\dot{q}^\star(t) = {}^{\mathsf{e}}\mathbb{U}_p^{q(t)} \frac{d}{dt} s_p(q(t)) = \dot{u}(t) - \mathbb{E}_{q(t)}\left[\dot{u}(t)\right] = \frac{d}{dt} \log q(t) = \frac{\dot{q}(t)}{q(t)} \ .$$

It follows that $t \mapsto (q(t), \dot{q}^\star(t))$ is a curve in the statistical bundle whose expression in the chart centered at p is $t \mapsto (u(t), \dot{u}(t))$. In fact,

$${}^{\mathsf{e}}\mathbb{U}_{q(t)}^{p} \left(\dot{u}(t) - dK_p(u(t))[\dot{u}(t)]\right) = \dot{u}(t) \ .$$

The mapping $q \mapsto (q, \dot{q}^\star)$ is a lift of the curve to the statistical bundle.

Covariant Derivatives

Given the exponential parallel transport, we define a *covariant derivative* by setting

$$\begin{aligned}\frac{D}{dt}w(t) &= {}^{\mathsf{e}}\mathbb{U}_{p}^{q(t)}\frac{d}{dt}\,{}^{\mathsf{e}}\mathbb{U}_{q(t)}^{p}w(t) = {}^{\mathsf{e}}\mathbb{U}_{p}^{q(t)}\Big(\dot{w}(t) - \mathbb{E}_{p}\left[\dot{w}(t)\right]\Big) \\ &= \dot{w}(t) - \mathbb{E}_{q(t)}\left[\dot{w}(t)\right] \ .\end{aligned}$$

The notation $\frac{D}{dt}$ denotes the covariant time derivative

In the *dual bundle*, the curve is $\zeta(t) = (q(t), \eta(t))$ and the expression of the second component is ${}^{\mathsf{m}}\mathbb{U}_{q(t)}^{p}\eta(t) = \frac{q(t)}{p}\eta(t)$. Then

$$\frac{d}{dt}\,{}^{\mathsf{m}}\mathbb{U}_{q(t)}^{p}\eta(t) = \frac{d}{dt}\frac{q(t)}{p}\eta(t) = \frac{1}{p}\left(\dot{q}(t)\eta(t) + q(t)\dot{\eta}(t)\right) \ ,$$

which, in turn, gives the dual covariant derivative

$$\frac{D}{dt}\eta(t) = {}^{\mathsf{m}}\mathbb{U}_{p}^{q(t)}\frac{d}{dt}\,{}^{\mathsf{m}}\mathbb{U}_{q(t)}^{p}\eta(t) = \overset{\star}{q}(t)\eta(t) + \dot{\eta}(t) \ .$$

Second Statistical Bundle

We define the second statistical bundle to be

$$S^2\,\mathcal{E}\,(\mu) = \{(q, w_1, w_2, w_3)\,|\,(q \in \mathcal{E}\,(\mu)\,, w_1, w_2, w_3 \in S_q\,\mathcal{E}\,(\mu)\}\ ,$$

with charts centered at each $p \in \mathcal{E}\,(\mu)$ defined by

$$s_p(q, w_1, w_2, w_3) = \left(s_p(q), {}^{\mathsf{e}}\mathbb{U}_q^p w_1, {}^{\mathsf{e}}\mathbb{U}_q^p w_2, {}^{\mathsf{e}}\mathbb{U}_q^p w_3\right)\ .$$

The second bundle is an expression of the tangent bundle of the exponential bundle. For each curve $t \mapsto \gamma(t) = (q(t), w(t))$ in the statistical bundle, we define its *velocity at* t to be

$$\dot{\gamma}(t) = \left(q(t), w(t), \dot{q}(t), \frac{D}{dt} w(t)\right)$$

Accelerations

In particular, for each smooth curve $t \mapsto q(t)$, the velocity of the lift $t \mapsto \gamma(t) = (q(t), \overset{\star}{q}(t))$ is

$$\overset{\star}{\chi}(t) = \left(q(t), \overset{\star}{q}(t), \overset{\star}{q}(t), \overset{\star\star}{q}(t)\right) \ ,$$

where the *acceleration at* t $\overset{\star\star}{q}(t)$ is

$$\overset{\star\star}{q}(t) = \frac{D}{dt}\overset{\star}{q}(t) = \frac{d}{dt}\frac{\dot{q}(t)}{q(t)} - \mathbb{E}_{q(t)}\left[\frac{d}{dt}\frac{\dot{q}(t)}{q(t)}\right] = \frac{\ddot{q}(t)}{q(t)} - \left(\overset{\star}{q}(t)^2 - \mathbb{E}_{q(t)}\left[\overset{\star}{q}(t)^2\right]\right).$$

We have three different interpretation of the lifted curve, namely, we can consider $t \mapsto (q(t), \overset{\star}{q}(t))$ as a curve in the statistical bundle $S\mathcal{E}(\mu)$, or, a curve in the dual bundle ${}^{*}S\mathcal{E}(\mu)$. Each of these frameworks provides a different derivation, hence, a different acceleration.

We have the already defined *exponential acceleration* ${}^{e}\mathrm{D}^2 q(t) = \overset{\star\star}{q}(t)$, and we can define, the *mixture acceleration* as

$${}^{m}\mathrm{D}^2 q(t) = \frac{D^{\mathsf{m}}}{dt}\overset{\star}{q}(t) = {}^{\mathsf{m}}\mathbb{U}_p^{q(t)}\frac{d}{dt}\,{}^{\mathsf{m}}\mathbb{U}_{q(t)}^{p}\overset{\star}{q}(t) = \ddot{q}(t)/q(t)$$

Total Derivative

Let be given a scalar field $F\colon {}^1S^1\,\mathcal{E}\,(\mu)\times\mathcal{D}\to\mathbb{R}$, $\mathcal{D}$ a domain of $\mathbb{R}^k$, and a generic smooth curve

$$t\mapsto (q(t),\eta(t),w(t),c(t))\in {}^1S^1\,\mathcal{E}\,(\mu)\times\mathcal{D}\ .$$

The total derivative can be computed by

$$\begin{aligned}\frac{d}{dt}F\big(q(t),&\eta(t),w(t),c(t)\big)=\\ &\Big\langle \operatorname{grad}F\big(q(t),\eta(t),w(t),c(t)\big),\dot{q}^{\star}(t)\Big\rangle_{q(t)}+\\ &\left\langle \frac{D}{dt}\eta(t),\operatorname{grad}_{\mathsf{m}}F\big(q(t),\eta(t),w(t),c(t)\big)\right\rangle_{q(t)}+\\ &\left\langle \operatorname{grad}_{\mathsf{e}}F\big(q(t),\eta(t),w(t),c(t)\big),\frac{D}{dt}w(t)\right\rangle_{q(t)}+\\ &\nabla F\big(q(t),\eta(t),w(t),c(t)\big)\cdot\dot{c}(t)\end{aligned}$$

Action Integral

If $q\colon [0,1] \ni t \mapsto q(t)$ is a smooth curve in the exponential manifold $\mathcal{E}(\mu)$ and $t \mapsto (q(t), \overset{\star}{q}(t))$, $\overset{\star}{q}(t) = \frac{d}{dt} \log q(t)$, is its lift to the statistical bundle $S\mathcal{E}(\mu)$, an *action integral* is

$$q \mapsto A(q) = \int_0^1 L(q(t), \overset{\star}{q}(t), t)\ dt\ ,$$

where $L\colon S\mathcal{E}(\mu) \times [0,1] \to \mathbb{R}$ is a smooth Lagrangian function.

Let us express the action integral in the exponential chart s_p centered at p. If $q(t) = \mathrm{e}^{u(t) - K_p(u(t))} \cdot p$, with $t \mapsto u(t) \in S_p\mathcal{E}(\mu)$, we have

$$s_p(q(t), \overset{\star}{q}(t)) = (u(t), \dot{u}(t))\ \ ,$$

hence,

$$L(q(t), \overset{\star}{q}(t), t) = L\left(e_p(u(t)), {}^{\mathrm{e}}\mathbb{U}_p^{e_p(u(t))} \dot{u}(t), t\right) = L_p(u(t), \dot{u}(t), t)\ ,$$

so that the expression of the action integral is

$$u \mapsto A_p(u) = \int_0^1 L_p(u(t), \dot{u}(t), t)\ dt\ .$$

Euler Lagrange Equation

The Euler-Lagrange equation, written with partial derivatives, that is, without the gradients to be computed below, is

$$d_1 L_p(u(t), \dot{u}(t), t)[h] = \frac{d}{dt} d_2 L_p(u(t), \dot{u}(t), t)[h]$$

with $t \in [0,1]$, $h \in S_p \mathcal{E}(\mu)$

If q is an extremal of the action integral, then

$$\frac{D}{dt} \operatorname{grad}_{\mathsf{e}} L(q(t), \dot{q}(t), t) = \operatorname{grad} L(q(t), \dot{q}(t), t) \ .$$

Legendre Transform 1/2

At each fixed density $q \in \mathcal{E}(\mu)$, and each time t, the mapping

$$S_q\mathcal{E}(\mu) \ni w \mapsto L_{q,t}(w) = L(q,w,t)$$

is defined on the vector space $S_q\mathcal{E}(\mu)$, and its gradient mapping in the duality of ${}^*S_q\mathcal{E}(\mu) \times S_q\mathcal{E}(\mu)$ is the mapping $w \mapsto \mathrm{grad}_{\mathsf{e}} L(q,w,t)$.

The Legendre transform $H_{q,t}$ of $L_{q,t}$ is defined for each $\eta \in {}^*S_q\mathcal{E}(\mu)$ of the image of $\mathrm{grad}_{\mathsf{e}} L(q,\cdot,t)$ by

$$H_{q,t}(\eta) = \left\langle \eta, (\mathrm{grad}_{\mathsf{e}} L_{q,t})^{-1}(\eta) \right\rangle_q - L_q((\mathrm{grad}_{\mathsf{e}} L_{q,t})^{-1}(\eta)) \ ,$$

which, in turn, defines the Hamiltonian

$$H(q,\eta,t) = \left\langle \eta, (\mathrm{grad}_{\mathsf{e}} L_{q,t})^{-1}(\eta) \right\rangle_q - L(q,(\mathrm{grad}_{\mathsf{e}} L_{q,t})^{-1}(\eta)) \ .$$

Legendre Transform 2/2

It is a general property of the Legendre transform that

$$\operatorname{grad}_{\mathsf{m}} H_{q,t}(\eta) = (\operatorname{grad}_{\mathsf{e}} L_{q,t})^{-1}(\eta) \ ,$$

which, in turn, implies the Young equality

$$H(q,\eta,t) + L(q,w,t) = \langle \eta, w \rangle_q$$

if $\eta = \operatorname{grad}_{\mathsf{e}} L(q,w,t)$ or $\operatorname{grad}_{\mathsf{m}} H(q,\eta,t) = w$

Hamiltonian Equations

For partial mappings $w \mapsto L_q(w)$ are strictly convex for each q, $w \mapsto \eta = \operatorname{grad}_{\mathsf{e}} L_q(w)$ is a 1-to-1 mapping from $S_q \mathcal{E}(\mu)$ to ${}^*S_q \mathcal{E}(\mu)$ and thus the Euler-Lagrange becomes

$$\frac{D}{dt}\eta(t) = \frac{D}{dt} \operatorname{grad}_{\mathsf{e}} L(q(t), \dot{q}^\star(t))$$

and the *Hamilton equations* hold, namely,

$$\begin{cases} \dfrac{D}{dt}\eta(t) = -\operatorname{grad} H(q(t), \eta(t), t) \\ \dot{q}^\star(t) = \operatorname{grad}_{\mathsf{m}} H(q(t), \eta(t), t). \end{cases}$$

For each solution of the Hamilton equations, it holds

$$\frac{d}{dt} H(q(t), \eta(t), t) = \frac{\partial}{\partial t} H(q(t), \eta(t), t) \ .$$

Free Particle Lagrangian 1/2

Let m be the inertial mass

$$L(q,w) = \frac{m}{2}\,\mathbb{E}_q\left[w^2\right] = \frac{m}{2}\,\langle w,w\rangle_q \;\;, \quad m \geq 0,\ (q,w) \in S\mathcal{E}\,(\mu)\;\;.$$

From

$$d^2K_p(u)[h_1,h_2] = \left\langle {}^{\mathsf{e}}\mathbb{U}_p^{\mathrm{e}_p(u)} h_1, {}^{\mathsf{e}}\mathbb{U}_p^{\mathrm{e}_p(u)} h_2 \right\rangle_{\mathrm{e}_p(u)}$$

we can obtain an expression in the chart centered in p for the Lagrangian

$$L_p(u,v) = \frac{m}{2}\left\langle {}^{\mathsf{e}}\mathbb{U}_p^{\mathrm{e}_p(u)} v, {}^{\mathsf{e}}\mathbb{U}_p^{\mathrm{e}_p(u)} v \right\rangle_{\mathrm{e}_p(u)} = \frac{m}{2} d^2K_p(u)[v,v]\;,$$

where $q = e_p(u)$ and $w = {}^{\mathsf{e}}\mathbb{U}_p^q v$

Free Particle Lagrangian 2/2

By computing the total derivative in the chart of L,

$$dL_p(u,v)[h,k] = \frac{m}{2}\left\langle w^2 - \mathbb{E}_q\left[w^2\right], {}^{\mathsf{e}}\mathbb{U}_p^q u\right\rangle_q + m\left\langle w, {}^{\mathsf{e}}\mathbb{U}_p^q k\right\rangle_q \ .$$

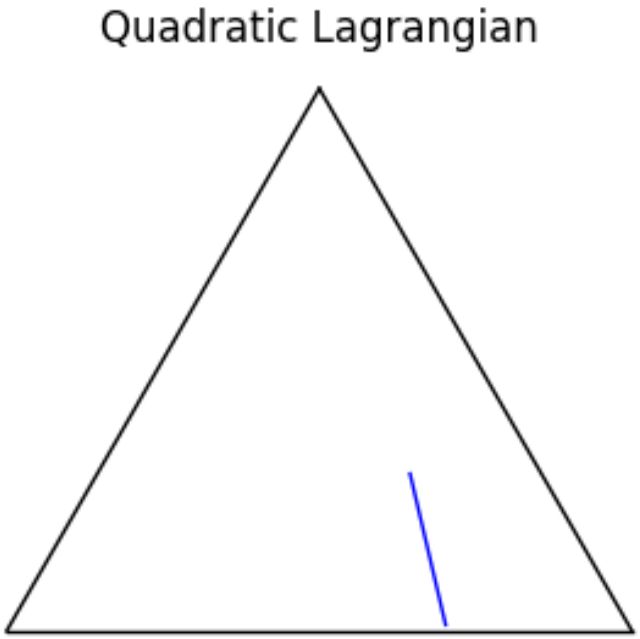
we can obtain the Euler-Lagrange equation

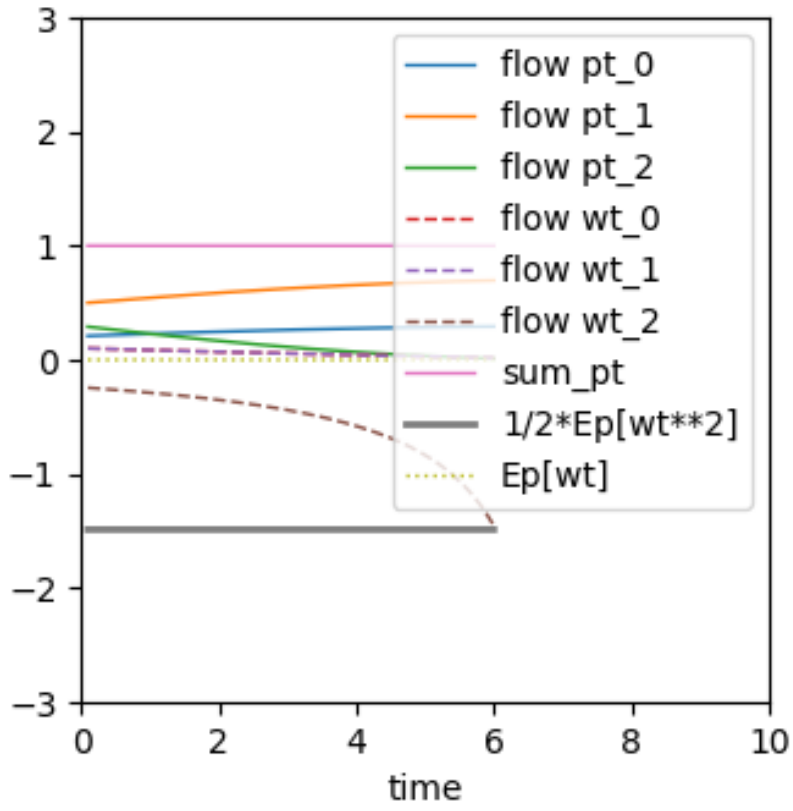
$$\frac{D}{dt}\dot{\overset{\star}{q}}(t) = \frac{1}{2}\left(\dot{\overset{\star}{q}}(t)^2 - \mathbb{E}_{q(t)}\left[\dot{\overset{\star}{q}}(t)^2\right]\right) \ ,$$

which can be expressed as a system of N second-order ODEs

$$\ddot{q}_j(t) = \frac{\dot{q}_j(t)^2}{2q_j(t)} - \frac{q_j(t)}{2N}\sum_{i=1}^{N}\frac{\dot{q}_i(t)^2}{q_i(t)^2} \ , \quad j = 1,\ldots,N \ .$$

Examples of Trajectories: Free Particle





Motion in Entropic Potential 1/2

Consider the case of a Lagrangian function given by the difference of the quadratic form and a potential on the bundle,

$$L(q,w) = \frac{m}{2}\langle w,w\rangle_q - \kappa\,\mathbb{E}_q\left[\log q\right] \ ,$$

with the negative entropy $f(q) = -\mathcal{H}(q)$ playing the role of the convex potential well, see Pistone (2018)

The Euler-Lagrange equation can be derived as

$$m\frac{D}{dt}\dot{q} = \frac{m}{2}\left(\dot{q}(t)^2 - \mathbb{E}_{q(t)}\left[\dot{q}(t)^2\right]\right) + \kappa\,\mathrm{grad}\,\mathcal{H}(q) \ .$$

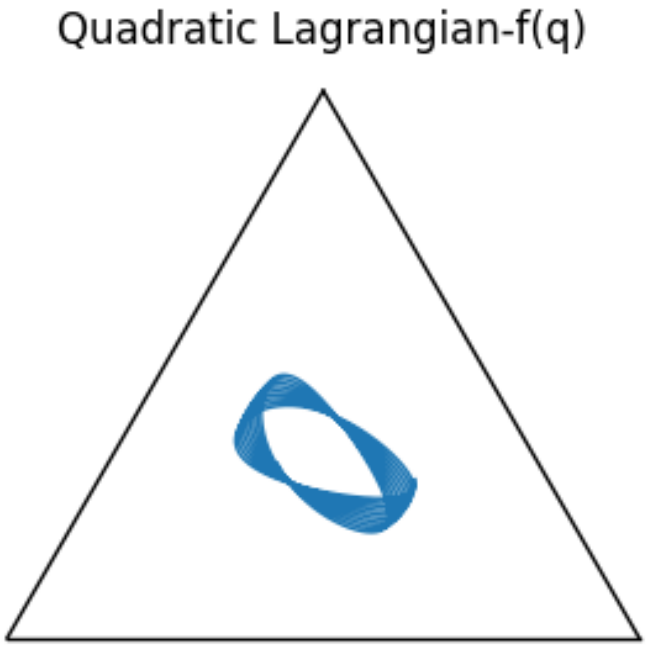
Motion in Entropic Potential 2/2

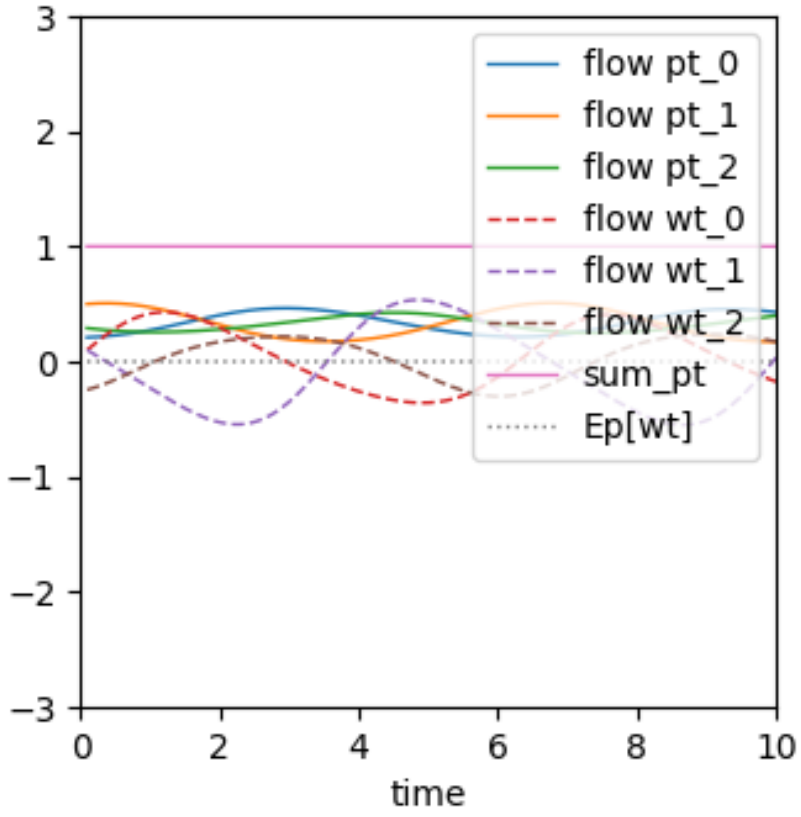
Let $A(q,v) = v^2/2 + \frac{\kappa}{m}\log(q)$ and $B(q,v) = v^2/2 - \frac{\kappa}{m}\log(q)$, the associated system of first-order ODEs is

$$\begin{cases} \dfrac{d}{dt}q(x;t) = q(x;t)v(x;t) \\ \dfrac{d}{dt}v(x;t) = -A(q(x;t),v(x;t)) - \dfrac{1}{N}\displaystyle\sum_y q(y;t)B(q(y;t),v(y;t)) \end{cases},$$

for $x \in \Omega$.

Examples of Trajectories: Motion in Potential





Divergence Lagrangian

A divergence is a smooth mapping $D\colon \mathcal{E}(\mu) \times \mathcal{E}(\mu) \to \mathbb{R}$, such that for all $p, q \in \mathcal{E}(\mu)$ it holds $D(p,q) \geq 0$ and $D(p,q) = 0$ if, and only if, $p = q$.

Every divergence can be associated to a Lagrangian by the canonical mapping

$$\mathcal{E}(\mu)^2 \ni (q,r) \mapsto (q, s_q(r)) = (q,w) \in S\mathcal{E}(\mu) \ ,$$

with $q = \mathrm{e}^{v - K_p(v)} \cdot p$, that is, $v = s_p(q)$.

We have an equivalence of a couple of a point and a vector and a couple of points. Every divergence D is mapped into a *divergence Lagrangian*, and conversely,

$$L(q,w) = D(q, e_q(w)) \ , \quad D(q,r) = L(q, s_q(r)) \ .$$

Kullback-Leibler Lagrangian

We focus on the case of the Kullback-Leibler divergence (KL), which lies at the intersection of the family of Csiszár's f-divergences and Bregman divergences (Amari, 2016).

Up to second-order approximation the KL provides a *locally quadratic measure*, motivating its interpretation as a local, non-symmetric generalization of the kinetic energy of classical mechanics.

The Lagrangian

$$D(q, r) = \mathrm{D}\left(q \,\|\, r\right) = \mathbb{E}_q\left[\log \frac{q}{r}\right] \ ,$$

can be written in chart at q as

$$\mathrm{D}\left(q \,\|\, e_q(w)\right) = \mathbb{E}_q\left[\log \frac{q}{e_q(w)}\right] = \mathbb{E}_q\left[-w + K_q(w)\right] = K_q(w) \ .$$

Kullback-Leibler Lagrangian: Euler-Lagrange

The expression of the divergence Lagrangian in chart at p is

$$L_p(u,v) = L(e_p(u), {}^{\mathsf{e}}\mathbb{U}_p^{e_p(u)} v) = D(e_p(u), e_{e_p(u)}({}^{\mathsf{e}}\mathbb{U}_p^{e_p(u)} v) \\ = D(e_p(u), e_p(u+v)) \ .$$

The Euler-Lagrange equation is obtained by plugging in $w(t) = \dot{\overset{\star}{q}}(t)$,

$$\frac{D}{dt}\left(e^{\overset{\star}{q}(t) - K_{q(t)}(\overset{\star}{q}(t))} - 1\right) = e^{\overset{\star}{q}(t) - K_{q(t)}(\overset{\star}{q}(t))} - 1 - \overset{\star}{q}(t) \ ,$$

which takes the form of a second-order equation

$$\left(\mathrm{e}^{\overset{\star}{q}(t) - K_{q(t)}(\overset{\star}{q}(t))}\right)\left(\overset{\star}{q}(t) + \ddot{\overset{\star}{q}}(t) - \mathbb{E}_{e_{q(t)}(\overset{\star}{q}(t))}\left[\overset{\star}{q}(t) + \ddot{\overset{\star}{q}}(t)\right]\right) = \\ = \mathrm{e}^{\overset{\star}{q}(t) - K_{q(t)}(\overset{\star}{q}(t))} - 1 \ .$$

Kullback-Leibler Lagrangian: ODE

By using $\dot{q}(t) = v(t)$ we have

$$\frac{d}{dt}v(t) = \ddot{q}(t) - \mathbb{E}_{q(t)}\left[v(t)^2\right] =$$

$$-v(t) + \frac{e^{\dot{q}(t) - K_{q(t)}(\dot{q}(t))} - 1}{e^{\dot{q}(t) - K_{q(t)}(\dot{q}(t))}} - \mathbb{E}_{q(t)}\left[\frac{e^{\dot{q}(t) - K_{q(t)}(\dot{q}(t))} - 1}{e^{\dot{q}(t) - K_{q(t)}(\dot{q}(t))}}\right] - \mathbb{E}_{q(t)}\left[v(t)^2\right]$$

which leads to

$$\begin{cases} \dfrac{d}{dt}q(x;t) = q(x;t)v(x;t) \\ \dfrac{d}{dt}v(x;t) = -v(x;t) + \dfrac{e^{v(x;t) - K_{q(t)}(v(x;t))} - 1}{e^{v(x;t) - K_{q(t)}(v(x;t))}} - \dfrac{1}{N}\displaystyle\sum_x q(x;t)\, v(x;t)^2 \\ \qquad\qquad - \dfrac{1}{N}\displaystyle\sum_x q(x;t)\, \dfrac{e^{v(x;t) - K_{q(t)}(v(x;t))} - 1}{e^{v(x;t) - K_{q(t)}(v(x;t))}} \end{cases},$$

for $x \in \Omega$.

Kullback-Leibler Lagrangian: Hamiltonian 1/2

The strong convexity of the KL generating function ensures the existence of an invertible Legendre transform, naturally allowing for a Hamiltonian formulation.

Using the equation for $\mathrm{grad}_{\mathsf{e}} K_q(w)$ and its inverse the Legendre transform of $w \mapsto K_q(w)$ is

$$
\begin{aligned}
H_q(\eta) = \langle \eta, \log(1+\eta) - \mathbb{E}_q\left[\log(1+\eta)\right] \rangle_q + \\
- K_q\Big(\log(1+\eta) - \mathbb{E}_q\left[\log(1+\eta)\right] \Big) \\
= \mathbb{E}_q\left[\eta \log(1+\eta)\right] - \mathbb{E}_q\left[\log\left(1+\eta\right)\right] = \mathbb{E}_q\left[(1+\eta)\log(1+\eta)\right] \ .
\end{aligned}
$$

In the chart at p, $q = e_p(u) = \mathrm{e}^{u - K_p(u)} \cdot p$,
$\eta = {}^{\mathsf{m}}\mathbb{U}_p^{e_p(u)} \zeta = \mathrm{e}^{-u + K_p(u)} \zeta$, so that

$$
\begin{aligned}
H_p(u,\zeta) = \mathbb{E}_{e_p(u)}\left[(1 + {}^{\mathsf{m}}\mathbb{U}_p^{e_p(u)} \zeta) \log(1 + {}^{\mathsf{m}}\mathbb{U}_p^{e_p(u)} \zeta) \right] = \\
\mathbb{E}_p\left[(\mathrm{e}^{u - K_p(u)} + \zeta) \log\left(1 + \mathrm{e}^{-u + K_p(u)} \zeta\right) \right] \ .
\end{aligned}
$$

Kullback-Leibler Lagrangian: Hamiltonian 2/2

By taking the derivative wrt u, and going back to the original variables, the Hamilton equations are

$$\begin{cases} \dfrac{D}{dt}\eta(t) = \eta(t) - \log(1+\eta(t)) + \mathbb{E}_{q(t)}\left[\log(1+\eta(t))\right] \\ \quad \overset{\star}{q}(t) = \log(1+\eta(t)) - \mathbb{E}_{q(t)}\left[\log(1+\eta(t))\right] \end{cases}$$

The solution curve and its derivatives can be expressed in the global space in which the dual bundle is embedded by

$$\frac{D}{dt}\eta(t) = \frac{\dot{q}(t)}{q(t)}\eta(t) + \dot{\eta}(t), \qquad \overset{\star}{q}(t) = \frac{\dot{q}(t)}{q(t)} \ ,$$

so that the resulting system of ODEs becomes

$$\begin{cases} \dot{\eta}(x;t) = \eta(x;t) - (1+\eta(x;t))\left(\log(1+\eta(x;t)) \right. \\ \qquad\qquad \left. -\frac{1}{N}\sum_y q(y;t)\log(1+\eta(y;t))\right) \ , \\ \dot{q}(x;t) = q(x;t)\left(\log(1+\eta(x;t)) - \frac{1}{N}\sum_y q(y;t)\log(1+\eta(y;t))\right) \ . \end{cases}$$

Alternative Parameterization 1/2

By writing $\chi(t) = e_{q(t)}(\overset{\star}{q}(t)) = \mathrm{e}^{\overset{\star}{q}(t) - K_{q(t)}(\overset{\star}{q}(t))} \cdot q(t)$ and $\overset{\star}{\chi}$ in terms of $\overset{\star}{q}$ and $\overset{\star\star}{q}$, the covariant derivative in the lhs of the Euler-Lagrange becomes

$$\frac{D}{dt}\left(\frac{\chi(t) - q(t)}{q(t)}\right) = \frac{1}{q(t)}\frac{d}{dt}\frac{q(t)}{1}\left(\frac{\chi(t) - q(t)}{q(t)}\right) = \frac{\dot{\chi}(t) - \dot{q}(t)}{q(t)} =$$

$$\frac{\overset{\star}{\chi}(t)\chi(t)}{q(t)} - \overset{\star}{q}(t) = \frac{{}^{\mathsf{e}}\mathbb{U}^{\chi(t)}_{q(t)}(\overset{\star\star}{q}(t) + \overset{\star}{q}(t))\chi(t)}{q(t)} - \overset{\star}{q}(t) \ ,$$

while the rhs is

$$\frac{\chi(t) - q(t)}{q(t)} - \overset{\star}{q}(t) \ ,$$

Finally, the Euler-Lagrange equation can be written as

$$\overset{\star}{\chi}(t)\chi(t) = \chi(t) - q(t) \ .$$

Alternative Parameterization 2/2

The introduction of two unknowns q and χ reduces the Euler-Lagrange equation to a system of evolution equations is the statistical bundle,

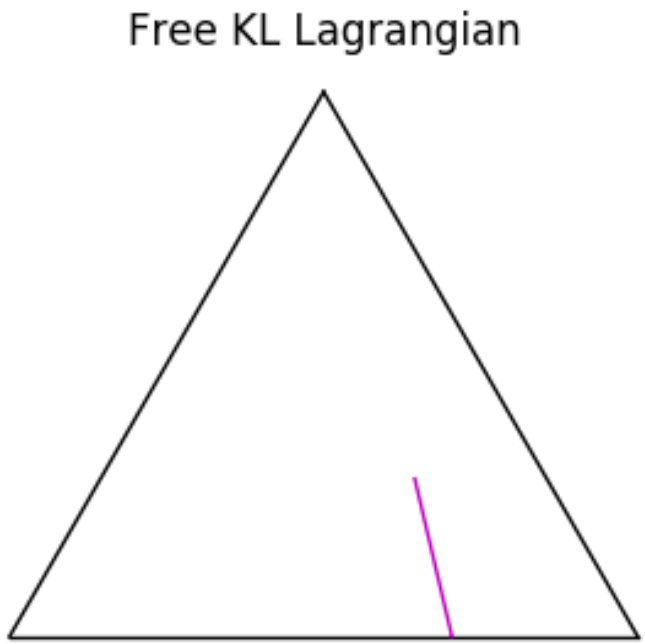
$$\begin{cases} \overset{\star}{\chi} = 1 - q\chi^{-1} \\ \overset{\star}{q} = \log \dfrac{\chi}{q} - \mathbb{E}_q \left[\log \dfrac{\chi}{q} \right] \end{cases},$$

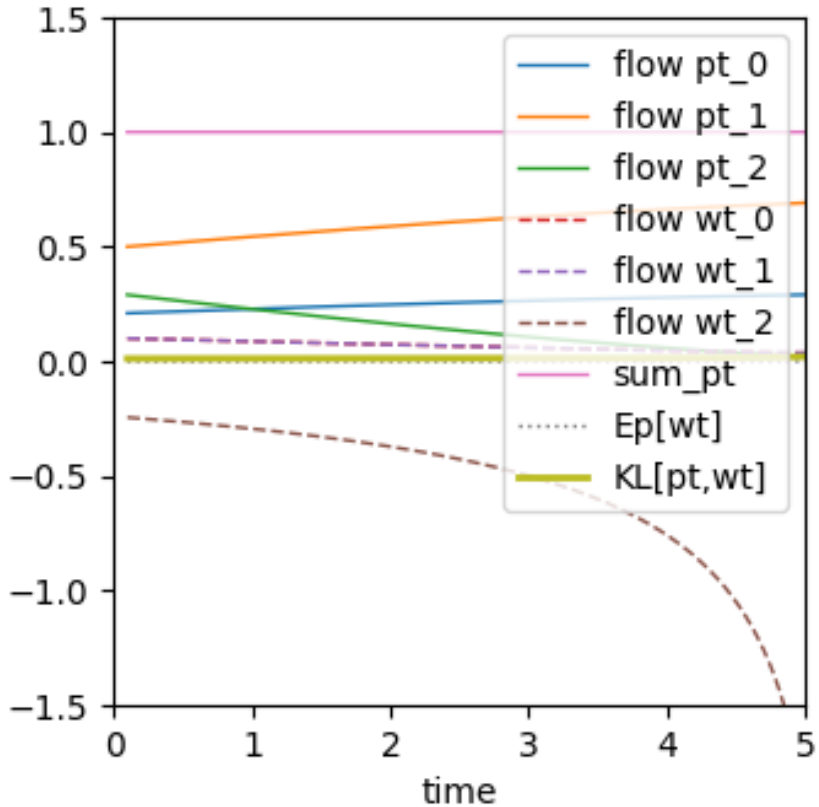
that is, to a system of replicator equations,

$$\begin{cases} \dot{\chi} = \chi - q \\ \dot{q} = q \left(\log \dfrac{\chi}{q} - \mathbb{E}_q \left[\log \dfrac{\chi}{q} \right] \right) \end{cases},$$

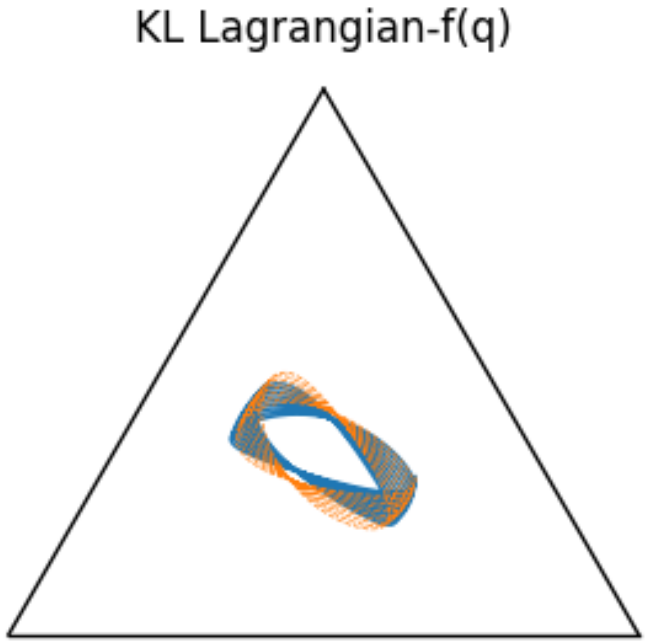
Notice that the vector field is zero if, and only if, $\chi = q$.

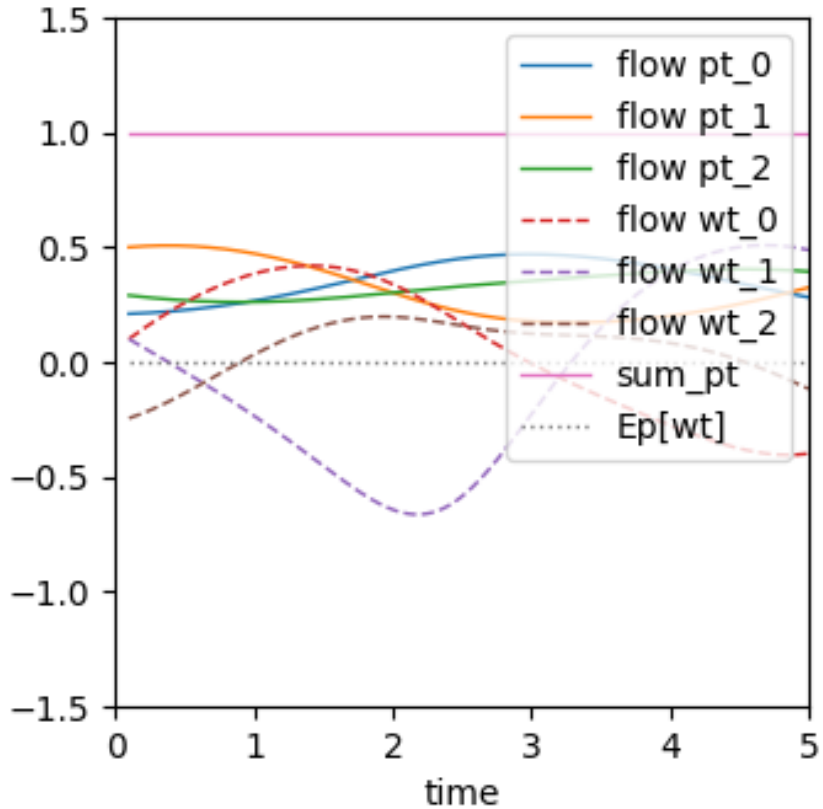
Examples of Trajectories: Free Motion





Examples of Trajectories: Motion in a Potential





Time Dependence

We can introduce an explicit time dependence in the Lagrangian.

This choice is motivated by the role time in generating a *dissipative* accelerated dynamics, which is of central interest in optimization.

In the exponential map, we consider a time-dependent scaling of the shift vector, such that $\chi = e_q(\mathrm{e}^{-\alpha_t} w)$ and $s_p(\chi) = u + \mathrm{e}^{-\alpha_t} v \in S_p\,\mathcal{E}\,(\mu)$, with $\alpha_t : I \to \mathbb{R}$ smooth, $I \subset \mathbb{R}$ open time interval. With this choice the KL Lagrangian reads

$$D\colon I \times S\mathcal{E}\,(\mu) \ni (q, w, t) \mapsto \mathrm{D}\,(q \,\|\, e_q(\mathrm{e}^{-\alpha_t} w)) \in \mathbb{R}\ .$$

In presence of explicit time-dependence, desirable closure under time-dilation can be achieved by an overall scaling of the divergence by a factor e^{α_t}, such that the new Lagrangian

$$L(q, w, t) = \mathrm{e}^{\alpha_t}\ \mathrm{D}\,(q \,\|\, e_q(\mathrm{e}^{-\alpha_t}\, w))\ \ ,$$

leads to fully time-reparametrization invariant action.

Time Dependent KL Lagrangian

We can derive the Euler-Lagrange equation in presence of the time-scaling for $v(t) = \dot{u}(t)$), we get

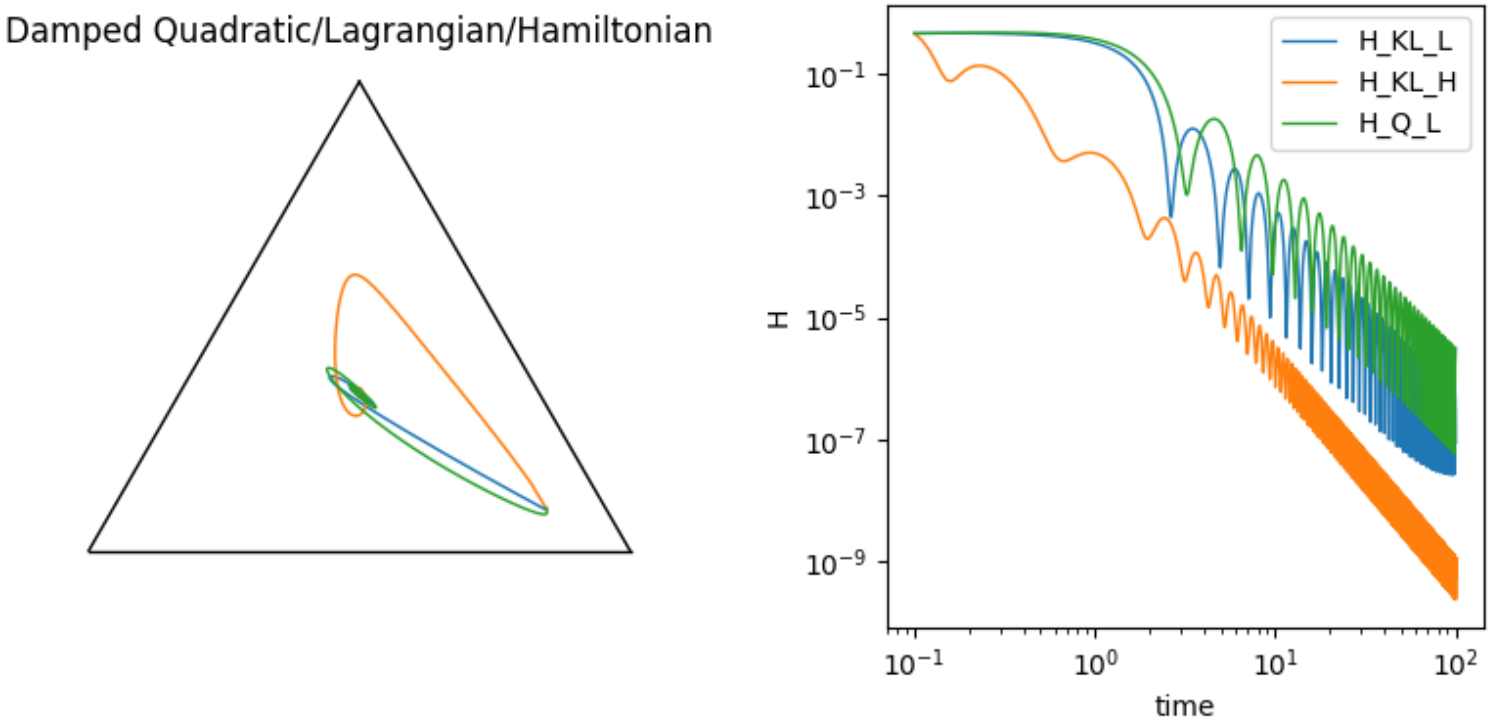
$$d^2 K_p(u(t) + \mathrm{e}^{-\alpha_t}\dot{u}(t))[(\mathrm{e}^{\alpha_t} - \dot{\alpha}_t)\,\dot{u}(t) + \ddot{u}(t), h] = \\ = \mathrm{e}^{2\alpha_t}\ (dK_p(u(t) + \mathrm{e}^{-\alpha_t}\dot{u}(t))[h] - dK_p(u(t))[h])\ ,$$

We can then transport the equation back on the statistical bundle to get

$$\frac{e_q(\mathrm{e}^{-\alpha_t}\,\overset{\star}{q})}{q}\left((\mathrm{e}^{\alpha_t} - \dot{\alpha}_t)\overset{\star}{q}(t) + \overset{\star\star}{q}(t) - \mathbb{E}_{e_p(u+\mathrm{e}^{-\alpha_t}v)}\left[(\mathrm{e}^{\alpha_t} - \dot{\alpha}_t)\overset{\star}{q}(t) + \overset{\star\star}{q}(t)\right]\right) \\ = \mathrm{e}^{2\alpha_t}\left(\frac{e_q(\mathrm{e}^{-\alpha_t}\,\overset{\star}{q})}{q} - 1\right)\ ,$$

with respect to the equation derived for the cumulant Lagrangian, the time-dependent scaling leads to a extra *damping* contribution in the velocity, which redefines the coefficient of $\overset{\star}{q}$.

Examples of Trajectories: Damped Systems



Take Home Message and Future Work

We have proposed a new formalism for the study of the evolution of probability densities on a finite space

A non-parametric presentation of the Lagrangian and Hamiltonian dynamics on the statistical bundle is feasible

The mechanical formalism acts on the statistical bundle, which has a natural interpretation in statistical terms

Future works include

- Implementation of discretization schemes, compatible with the geometry of the ODEs, to obtain efficient optimization algorithms
- Define dynamics on submanifolds of the probability simplex
- Define dynamics over the manifold of Gaussian measures