Metode de optimizare riemanniene pentru învăţare profundă
Proiect cofinanţat din Fondul European de Dezvoltare Regională prin Programul Operaţional Competitivitate 2014-2020

Lagrangian and Hamiltonian Dynamics on the Simplex

Goffredo Chirco, Luigi Malagò, Giovanni Pistone
{chirco,malago}@rist.ro, giovanni.pistone@carloalberto.org
Romanian Institute of Science and Technology
Collegio Carlo Alberto

Abstract

The statistical bundle is the set of couples (q, w) of a positive probability density q and a random variable w such that $\mathbb{E}_q[w] = 0$. On a finite state space, we assume q to be a probability density with respect to the uniform probability and express it in the affine atlases of exponential charts. Velocity and acceleration of a one-dimensional statistical model are computed using the canonical dual pair of parallel transports. We define Lagrangian and Hamiltonian mechanics on the bundle and we provide explicit examples of a time-independent and time-dependent Lagrangian functions, leading to diverse accelerated natural gradient dynamics. Within this formalism we can reproduce Nesterov's flow for convex constrained optimization problems on the statistical bundle.

Statistical Bundle

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

$$q(x) = \exp(v(x) + H(v)) \cdot p(x),$$

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

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

The **exponential statistical bundle** with base Ω is defined as

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

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

Affine Geometries

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

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

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

(a) **Exponential transport** ${}^{\mathrm{e}}\mathbb{U}_p^q\colon S_p\mathcal{E}\left(\mu\right) \to S_q\mathcal{E}\left(\mu\right), {}^{\mathrm{e}}\mathbb{U}_p^q v = v - \mathbb{E}_q\left[v\right]$,
(b) **Mixture transport** ${}^{\mathrm{m}}\mathbb{U}_p^q\colon {}^*S_p\mathcal{E}\left(\mu\right) \to {}^*S_q\mathcal{E}\left(\mu\right), {}^{\mathrm{m}}\mathbb{U}_p^q\eta = \frac{p}{q}\eta$.

Quadratic Lagrangian

Let m be the inertial mass

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

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

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

where $q = e_p(u)$ and $w = {}^{\mathrm{e}}\mathbb{U}_p^q v$
By computing the total derivative in the chart of L,

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

we can obtain the Euler-Lagrange equation

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

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

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

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 $f(q) = -\mathcal{H}\left(q\right)$ playing the role of the convex potential well.

The Euler-Lagrange equation can be derived as

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

Let $A(q, v) = v^2/2 + \frac{\kappa}{m}\log\left(q\right)$ and $B(q, v) = v^2/2 - \frac{\kappa}{m}\log\left(q\right)$, 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 $x \in \Omega$.

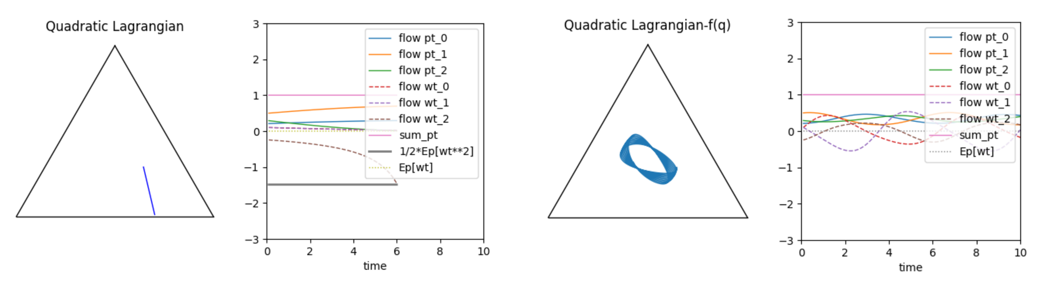

Figure 1: (Left) Free particle quadratic potential; (Right) Motion in Potential.

Kullback-Leibler Lagrangian

A divergence is a smooth mapping $D\colon \mathcal{E}\left(\mu\right) \times \mathcal{E}\left(\mu\right) \to \mathbb{R}$, such that for all $p, q \in \mathcal{E}\left(\mu\right)$ 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

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

with $q = e^{v - K_p(v)} \cdot p$, 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,

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

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

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

The expression of the divergence Lagrangian in chart at p is

$$\begin{aligned} L_p(u, v) = L(e_p(u), {}^{\mathrm{e}}\mathbb{U}_p^{e_p(u)} v) &= D(e_p(u), e_{e_p(u)}({}^{\mathrm{e}}\mathbb{U}_p^{e_p(u)} v) \\ &= D(e_p(u), e_p(u + v)) . \end{aligned}$$

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

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

which takes the form of a second-order equation

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

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

$$\begin{aligned} \frac{d}{dt}v(t) = \overset{\star\star}{q}(t) - \mathbb{E}_{q(t)}\left[v(t)^2\right] &= -v(t) + \frac{\mathrm{e}^{\overset{\star}{q}(t) - K_{q(t)}(\overset{\star}{q}(t))} - 1}{\mathrm{e}^{\overset{\star}{q}(t) - K_{q(t)}(\overset{\star}{q}(t))}} - \\ &+ \mathbb{E}_{q(t)}\left[\frac{\mathrm{e}^{\overset{\star}{q}(t) - K_{q(t)}(\overset{\star}{q}(t))} - 1}{\mathrm{e}^{\overset{\star}{q}(t) - K_{q(t)}(\overset{\star}{q}(t))}}\right] - \mathbb{E}_{q(t)}\left[v(t)^2\right] \end{aligned}$$

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

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

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

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

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

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

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

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

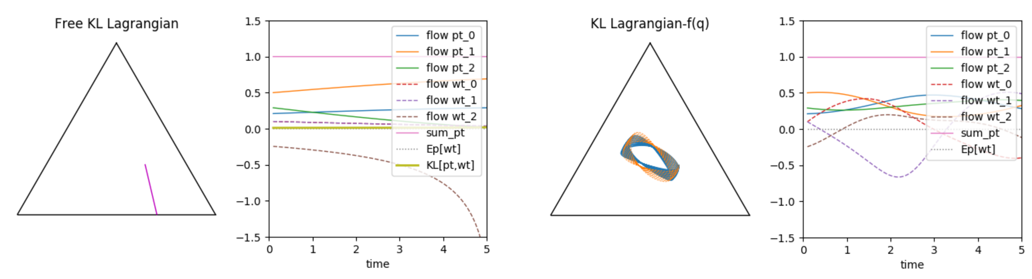

Figure 2: (Left) Free particle Kullback-Leibler divergence potential; (Right) Kullback-Leibler divergence motion in Potential.

Time Dependent KL Lagrangian

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 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 $s_p(\chi) = u + \mathrm{e}^{-\alpha_t} v \in S_p\mathcal{E}\left(\mu\right)$, with $\alpha_t : I \to \mathbb{R}$ smooth, $I \subset \mathbb{R}$ open time interval. With this choice the KL Lagrangian reads

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

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

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

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

$$\begin{aligned} &d^2K_p(u(t) + \mathrm{e}^{-\alpha_t}\dot{u}(t))[(\mathrm{e}^{\alpha_t} - \dot{\alpha}_t)\,\dot{u}(t) + \ddot{u}(t), h] = \\ &\quad = \mathrm{e}^{2\alpha_t}\left(dK_p(u(t) + \mathrm{e}^{-\alpha_t}\dot{u}(t))[h] - dK_p(u(t))[h]\right) , \end{aligned}$$

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

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

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 $\overset{\star}{q}$.

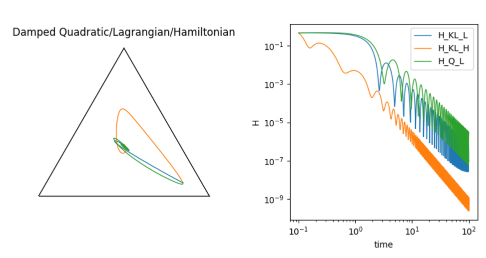

Figure 3: Comparison of different damped systems.