IGADG IV

AFFINE GEOMETRY OF THE STATISTICAL BUNDLE

Giovanni Pistone

DE CASTRO STATISTICS

August 12, 2023

e-mail: giovanni.pistone@carloalberto.org
orcidID: 0000-0003-2841-788X
The author acknowledges the support of de Castro Statistics and Collegio Carlo Alberto. He is a member of INdAM-GNAMPA and of the UMI Group MATHEMATICS FOR ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING.

PART O

Abstract⋆

The Statistical Bundle is the set $S\mathcal{E}$ of couples (p, u) with p strictly positive probability function and u a real random variable such that $E_p(u) = 0$. It is a vector bundle $\pi : S\mathcal{E} \to \mathcal{E}$ where $\mathcal{E}$ is the open probability simplex on a finite set X. For example, if $\theta \mapsto p(\theta) \in \mathcal{E}$ is a smooth one-dimensional probability model, the lift $\theta \mapsto (p(\theta), Dp(\theta))$ is a smooth curve in the Statistical Bundle, where $Dp(\theta)$ is the Fisher's score (logarithmic derivative) of the model.
Given two points (p, u) and (q, v) in $S\mathcal{E}$, one can define affine displacements in the elementary sense of Weyl (1921),

$$((p, u), (q, v)) \mapsto V_{p,u}(q, v) \in S_p\mathcal{E},$$

and correspondingly define an affine geometry on the Statistical Bundle. The further assignment of a duality pairing on the fibres produces by dualization a dually flat geometrical structure. See a tutorial in G Chirco and G Pistone arXiv:2204.00917.
Defining the affine geometry on the Statistical Bundle implicitly defines the connection on the non-parametric affine bundle of the open probability simplex.
The study of Information Geometry of the Statistical Bundle has other distinct advantages—first, a simplified presentation of the transport Problem of the probability simplex. See G. Pistone. Statistical bundle of the transport model. In GSI 5th Proceedings, 752–759. Springer-Verlag, 2021. Second, the vector bundle and its dual provides the proper setting for studying Lagrangian and Hamiltonian mechanics of the probability simplex. See G Chirco, L Malagò, G Pistone. Lagrangian and Hamiltonian dynamics for probabilities on the statistical bundle. *International Journal of Geometric Methods in Modern Physics*, 19(13):2250214.1–46, August 2022.
The talk will mention other relevant references, particularly the generalization to continuous state space.

My presentation will mainly focus on the statistical meaning of the geometric concepts.

Motivation

- If the state space is finite, say $\Omega = \{1, \ldots, n\}$, the set of probability functions $p \in \mathcal{P}$ is the probability simplex, a convex subset of the affine space defined by $A = \left\{ p \in \mathbb{R}^n \,\middle|\, \sum_{j=1}^n p_j = 1 \right\}$. The set of probability functions has naturally an affine structure whose vector space is defined by $B_1 = \sum_{j=1}^n p_j = 0$. If $\theta \mapsto p(\theta)$ is a 1-dimensional statistical model, then the derivative (velocity) is $\frac{d}{d\theta} p(\theta) \in B_1$.
- Ronald Fisher suggests a different affine structure. If the probability functions are strictly positive, $p \in \mathcal{P}_>$, the natural way to compute derivatives is the log-derivative, the Fisher's score, $\frac{d}{d\theta} \log f(\theta) = \dot{f}(\theta)/f(\theta) = \overset{\star}{f}(\theta)$. In this case $\overset{\star}{f}(\theta) \in B_\theta$, where B_θ is the vector space of random variables u such that $\sum_{j=1}^n u_j p_j(\theta) = 0$. Each density p has its own tangent space B_p.
- L. Boltzmann and W. Gibbs suggest considering for each "state" $p \in \mathcal{P}_>$ the fluctuations at p, that is the random variables in B_p. In particular, $p \propto \mathrm{e}^U$, $U - \sum_{j=1}^n U_j \; p_j \in B_p$ has many names in Statistical Physics.

References⋆ I

This talk is about probability measures' affine geometry. I shall not discuss other important related topics, namely, Riemannian geometry and metric geometry.

1921 Hermann Weyl. *Space- time- matter / by Hermann Weyl*. Dover, New York, 1952. translation of the 1921 RAUM ZEIT MATERIE Definition of affine space in physics.

1972 N. N. Čencov. *Statistical decision rules and optimal inference*, volume 53 of *Translations of Mathematical Monographs*. American Mathematical Society, 1982. Translation from the Russian edited by Lev J. Leifman Affine geometry in statistics.

1993 Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*. American Mathematical Society, 2000. Translated from the 1993 Japanese original by Daishi Harada Dually affine geometry in statistics named Information Geometry.

References⋆ II

1995 Giovanni Pistone and Carlo Sempi. An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. *Ann. Statist.*, 23(5):1543–1561, October 1995 Non parametric exponential families with Orlicz spaces.

1998 Paolo Gibilisco and Giovanni Pistone. Connections on non-parametric statistical manifolds by Orlicz space geometry. *IDAQP*, 1(2):325–347, 1998 Affine connections in non parametric Information Geometry.

1999 Giovanni Pistone and Maria Piera Rogantin. The exponential statistical manifold: mean parameters, orthogonality and space transformations. *Bernoulli*, 5(4):721–760, August 1999

2013 Giovanni Pistone. Nonparametric information geometry. In Frank Nielsen and Frédéric Barbaresco, editors, *Geometric science of information*, volume 8085 of *Lecture Notes in Comput. Sci.*, pages 5–36. Springer, Heidelberg, 2013. First International Conference, GSI 2013 Paris, France, August 28-30, 2013 Proceedings Statistical bundle arxiv:1306.0480

References⋆ III

2018 Giovanni Pistone. Information geometry of the Gaussian space. In *Information geometry and its applications*, volume 252 of *Springer Proc. Math. Stat.*, pages 119–155. Springer, Cham, 2018 Information Geometry of the Gaussian space.

2022 Goffredo Chirco, Luigi Malagò, and Giovanni Pistone. Lagrangian and Hamiltonian dynamics for probabilities on the statistical bundle. *International Journal of Geometric Methods in Modern Physics*, 19(13):2250214.1–46, August 2022 Mechanics on the statistical bundle. arxiv 2020

2022 Giovanni Pistone. Affine statistical bundle modeled on a Gaussian Orlicz–Sobolev space. *Information Geometry*, November 2022. Review paper.

2022 Goffredo Chirco and Giovanni Pistone. Dually affine Information Geometry modeled on a Banach space, 2022. arXiv:2204.00917 Handbook chapter.

Plan

- PART 1: Affine Statistical Bundle
- PART 2: Product space
- PART 3: Mechanics

PART I

Finite sample space

- If the sample space Ω is finite, the set of all probability measures is identified with the set $\mathcal{P}$ of all probability functions q,

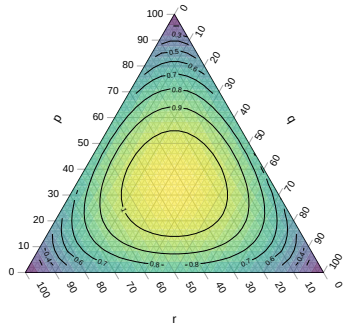
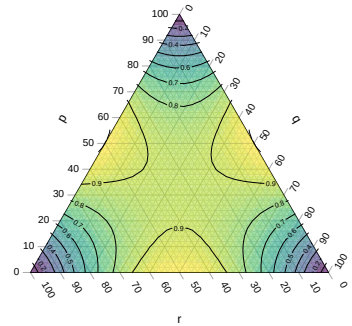
$$q\colon \Omega \ni x \mapsto q(x) \in \mathbb{R}_{\geq} \ , \quad \sum_{x\in\Omega} q(x) = 1 \ .$$

- The set $\mathcal{P}$ is a closed convex set of the vector space $\mathbb{R}^{\Omega}$. It is called standard simplex of $\mathbb{R}^{\Omega}$. A polytope is the convex hull of a finite number of points. In the case of the standard simplex, the generating points are the delta probability functions δ_a, $a \in \Omega$, $\delta_a(x) = (x = a)$. A simplex is a polytope generated by affinely independent points. Simplexes with the same number of vertexes are transformed into one another by an affine function.
- Given the simplex $\mathcal{P}(\Omega)$, its affine space is $\mathcal{A}(\Omega) = \left\{ v \in \mathbb{R}^{\Omega} \,\middle|\, \sum_{x\in\Omega} v(x) = 1 \right\}$, and its tangent space is

$$T\mathcal{P}(\Omega) = \left\{ v \in \mathbb{R}^{\Omega} \,\middle|\, \sum_{x\in\Omega} v(x) = 0 \right\}$$

- Alexander Barvinok. *A course in convexity*, volume 54 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2002

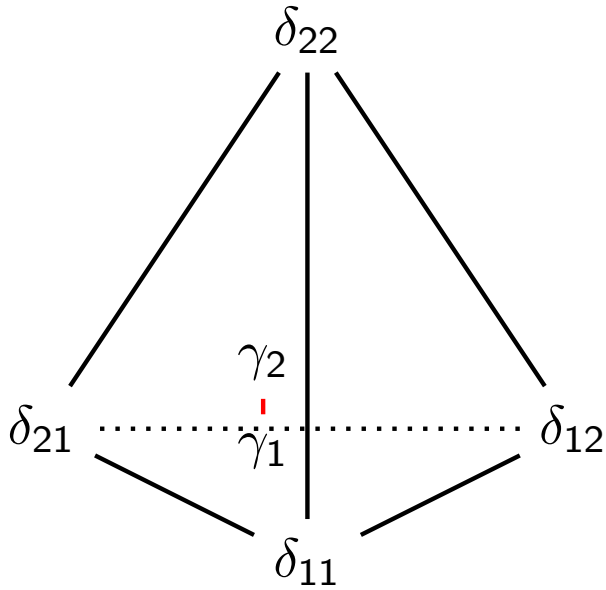
Examples⋆

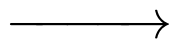
$\mathcal{H}(q) =$  POL$(q) =$ 

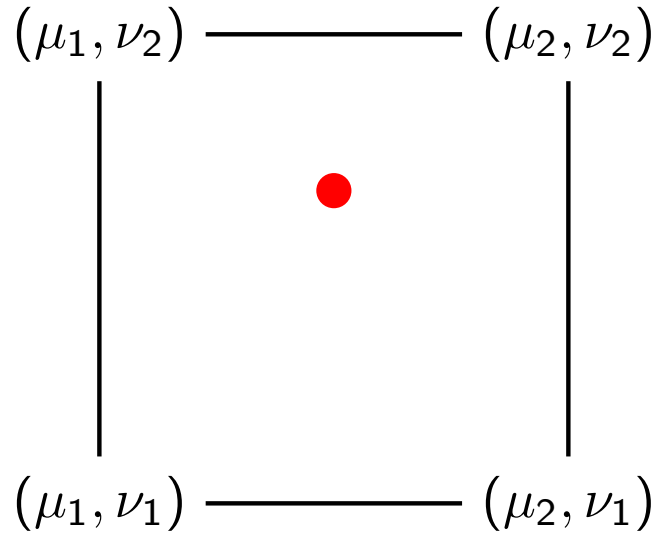
- The entropy $\mathcal{H}(q) = -\sum_{x\in\Omega} \log q(x)\, q(x) = \mathbb{E}_q[-\log q]$ $q(x) > 0$ has a maximum at the uniform probability function and a minimum on δ_a, $a \in \Omega$.
- If X, Y, Z are independent and $\sim q$, the probability of "two equal" is $3\sum_{x\in\Omega} q(x)^2(1-q(x))$. The polarization index is $\mathrm{POL}(q) = 4\sum_{x\in\Omega} q(x)^2(1-q(x)) = 4\,\mathbb{E}_q[q(1-q)]$. It is maximal at the middle point of faces of dimension 1, and minimal on the vertices. The uniform probability function is a critical point.

- The entropy has many interpretations. In particular, it is a form of potential energy, Giovanni Pistone. Lagrangian function on the finite state space statistical bundle. *Entropy*, 20(2):139, 2018.
- Polarization is discussed in Giovanni Pistone and Maria Piera Rogantin. The gradient flow of the polarization measure. with an appendix. arXiv:1502.06718, 2015.

Example⋆







Probability simplex of $\mathcal{P}(\{1,2\}^2)$ and the marginalization operator

$$\text{margins} : \mathcal{P}(\{1,2\}^2) \ni \gamma \mapsto (\mu, \nu) \in \mathcal{P}(\{1,2\})^2$$

The segment from γ_1 to γ_2 are the vertexes of a coupling polytope with

$$\gamma_1 = \begin{pmatrix} 1/6 & 1/3 \\ 1/2 & 0 \end{pmatrix} , \quad \gamma_2 = \begin{pmatrix} 1/2 & 0 \\ 1/6 & 1/3 \end{pmatrix} .$$

- The algebraic features of Kantorovich distance are discussed in Giovanni Pistone, Fabio Rapallo, and Maria Piera Rogantin. Finite space Kantorovich problem with an MCMC of table moves. *Electron. J. Statist.*, 15(1):880–907, 2021.

Affine space, affine atlas I

Given a set M and a real finite dimensional vector space V, Hermann Weyl[1] considers a displacement mapping

$$M \times M \ni (p, q) \mapsto \overrightarrow{pq} \in V \ ,$$

such that

- for each p the mapping $s_p \colon M \ni q \mapsto s_p(q) = \overrightarrow{pq}$ is 1-to-1, and
- the parallelogram law or Chasles relation,

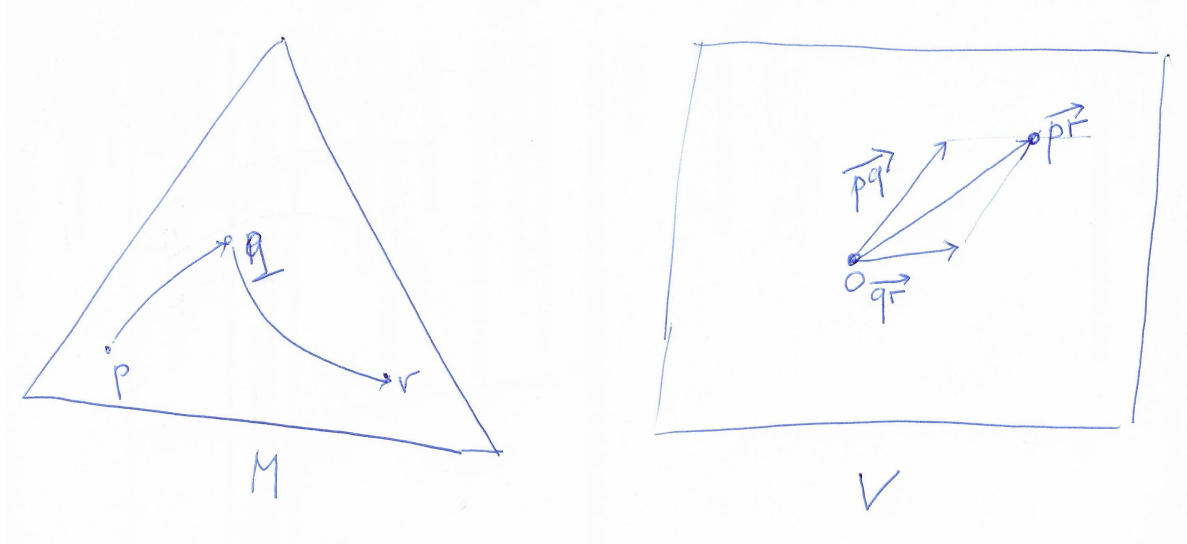
$$\overrightarrow{pq} + \overrightarrow{qr} = \overrightarrow{pr}$$

that is,

$$s_p(q) + s_q(r) = s_p(r)$$

holds.

Affine space, affine atlas II



- In particular, it follows

$$\overrightarrow{pp} = s_p(p) = 0 \quad \text{and} \quad \overrightarrow{pq} + \overrightarrow{qp} = s_p(q) + s_q(p) = 0$$

- $(M, V, \overrightarrow{\cdot\cdot})$ is an affine space. s_p is a chart if the image is open. The atlas of charts $s_p \colon M \to V$, $p \in M$, defines an affine manifold. All change-of-chats mappings of the atlas are vector translations:

$$s_p \circ s_q^{-1} \colon v \mapsto s_p(q) + v$$

[1] Hermann Weyl. *Space- time- matter / by Hermann Weyl*. Dover, New York, 1952. translation of the 1921 RAUM ZEIT MATERIE

Example: the probability simplex I

- The base set M is the probability simplex $\mathcal{P}(\Omega)$ and the vector space V is the tangent space of the simplex,

$$V = T\mathcal{P}(\Omega) = \left\{ v \in \mathbb{R}^{\Omega} \,\middle|\, \sum_{x \in \Omega} v(x) = 0 \right\} .$$

- The displacement is $\overrightarrow{pq} = q - p \in V$ because $\sum_x (q(x) - p(x)) = 1 - 1 = 0$. The mapping $s_p(q) = q - p$ is 1-to-1. The paralellogram law is $(q - p) + (r - q) = (r - p)$.

- A vector base of V produces a faithful parameterization of M. For example, a base of $T\mathcal{P}(\{1, 2, 3\})$ is $v_1 = (1, -1, 0)$, $v_2 = (1, 0, -1)$, and

$$(p, q) \mapsto s_p(q) = -(q(2) - p(2)) \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} - (q(3) - p(3)) \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \mapsto$$

$$(-(q(2) - p(2)), -(q(3) - p(3))) \in$$

$$\text{hull}\,((1, 0), (1, -1), (0, -1), (-1, 0), (-1, 1), (0, 1)) \subset \mathbb{R}^2 .$$

Example: the probability simplex II

- For example, J. Aitchison. *The statistical analysis of compositional data*. Monographs on Statistics and Applied Probability. Chapman& Hall, London, 1986 uses a generating set of contrasts systematically. I am not going to discuss his approach here.
- The use of integer values contrasts produces an algebraic theory as in Giovanni Pistone, Eva Riccomagno, and Henry P. Wynn. *Algebraic statistics: Computational commutative algebra in statistics*, volume 89 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, Boca Raton, FL, 2001. For example, Giovanni Pistone, Fabio Rapallo, and Maria Piera Rogantin. Finite space Kantorovich problem with an MCMC of table moves. *Electron. J. Statist.*, 15(1):880–907, 2021.

Example: the open probability simplex

- In the previous example, the image of each $s_p : q \mapsto q - p$ is a translation by $-p$ of the simplex. Hence it is closed in the tangent space, and the manifold structure is not defined. In fact, we need an open image.[2]
- Let us take as the base set the open simplex

$$\mathcal{P}_{>}(\Omega) = \{q \in \mathcal{P}(\Omega) \,|\, q(x) > 0, x \in \Omega\}$$

and the same vector space as in the previous case, $V = T\mathcal{P}(\Omega) = T\mathcal{P}_{>}(\Omega)$. The displacement $s_p(q) = q - p$ satisfies the parallelogram law and moreover, the image of s_p is $\mathcal{P}_{>}(\Omega) - p$ hence it is open in the tangent space. s_p is a chart. Let the inverse is $s_p^{-1}(v) = v + p$ and the change-of-chart is

$$v \mapsto s_p^{-1}(v) = v + p \mapsto s_q \circ s_p^{-1}(v) = (v + p) - q = v + s_q(p)$$

- However, in the practice of statistical physics and statistics, other affine geometries appear.

[2] We use the nonparametric definition of the manifold as in Serge Lang. *Differential and Riemannian manifolds*, volume 160 of *Graduate Texts in Mathematics*. Springer-Verlag, third edition, 1995

Fisher's score I

- A curve or one-dimensional model is a mapping

$$I \ni \theta \mapsto q(\theta) \in \mathcal{P}(\Omega) \ .$$

- As $\mathcal{P}(\Omega) \subset \mathbb{R}^{\Omega}$, we can assume I open and the mapping differentiable. In such a case,

$$\dot{q}(t) = \lim_{h \to 0} h^{-1}(q(t+h) - q(t)) \in T\mathcal{P}(\Omega) \ .$$

  Precisely because of that, we use the term "tangent space": the definition of tangent space depends on the definition of velocity of variation of the curve.

- Assume now that there exists a couple $(x, \theta_0) \in \Omega \times I$ such that the curve $\theta \mapsto q(\theta)$ hits the facet $\{p \,|\, p(x) = 0\}$ at $\theta = \theta_0$, that is $q(x; \theta_0) = 0$.[3] The differentiable curve $\theta \mapsto q(x, \theta)$ has a minimum at θ_0, hence $\dot{q}(x, \theta_0) = 0$.

Fisher's score II

It holds $q(x;\theta) = 0 \Rightarrow \dot{q}(x;\theta) = 0$, hence and exists a curve $t \mapsto \overset{\star}{q}(t) \in \mathbb{R}^{\Omega}$, the Fisher's score, such that

$$\dot{q}(x;\theta) = \overset{\star}{q}(x;\theta) q(x;\theta)$$

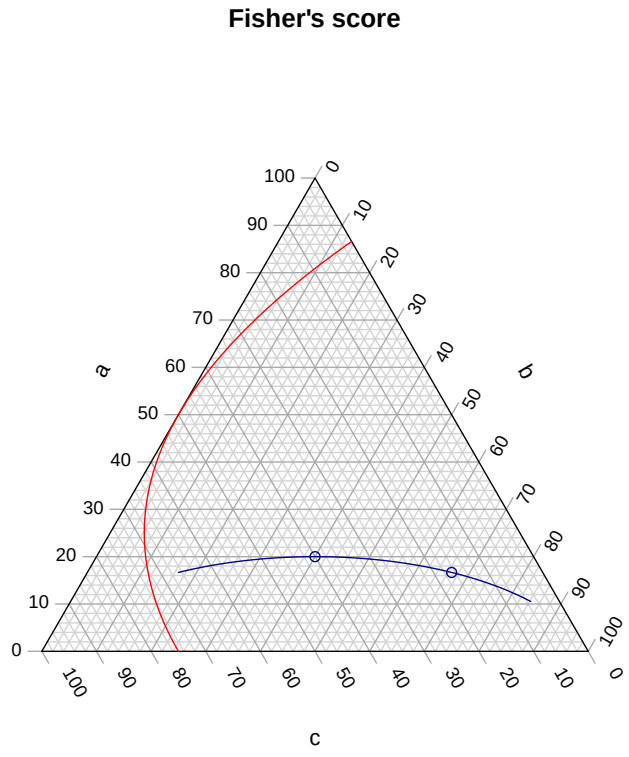
. It is an absolute continuity setting, $\dot{q}(\theta) << q(\theta)$.

- In particular, if $\theta \mapsto q(\theta) \in \mathcal{P}_{>}(\Omega)$, then the Fisher's score is $\overset{\star}{q}(\theta) = \frac{d}{d\theta} \log q(\theta)$. There are many statistical reasons to take the Fisher score as a useful notion of the velocity of variation in the parameters of a statistical model.[4]

[3] For this argument, see Nihat Ay, Jürgen Jost, Hông Vân Lê, and Lorenz Schwachhöfer. *Information geometry*, volume 64 of *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics [Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics]*. Springer, Cham, 2017

[4] See, for example, Bradley Efron and Trevor Hastie. *Computer age statistical inference*, volume 5 of *Institute of Mathematical Statistics (IMS) Monographs*. Cambridge University Press, New York, 2016. Algorithms, evidence, and data science

Examples I



Examples II

$$\theta \mapsto \theta\delta_a + \left(\theta - \frac{1}{2}\right)^2 \delta_b + \left(1 - \theta - \left(\theta - \frac{1}{2}\right)^2\right) \delta_c \qquad \text{(red curve)}$$

$$= \theta^2(\delta_b - \delta_c) + \theta(\delta_a - \delta_b) + \frac{1}{4}(\delta_b - \delta_c)$$

$$\theta \mapsto q(\theta) \propto P^\theta R^{(1-\theta)} = \exp\left(\theta \log \frac{P}{R}\right) \cdot R \qquad \text{(blue curve)}$$

- The Fisher's score is a contrast for the "true" probability:

$$\mathbb{E}_{q(\theta)}\left[\dot{\ell}(\theta)\right] = \sum_{x \in \Omega} \frac{\dot{q}(x;\theta)}{q(x;\theta)} q(x;\theta) = \frac{d}{d\theta} \sum_{x \in \Omega} q(x;\theta) = 0$$

Examples III

- The original example is the Gibbs-Boltzmann model[5]

$$\beta \mapsto q(x;\beta) = \exp\left(-\frac{1}{\beta}H(x) - \psi(\beta)\right)\cdot q(x)\ , \quad \beta > 0\ ,$$

  with $\psi(\beta) = \log \mathbb{E}_q\left[\mathrm{e}^{-\frac{1}{\beta}H}\right]$.

- The Fisher'score is

$$\dot{q}(\beta) = \frac{d}{d\beta}\log q(\beta) = \frac{1}{\beta^2}H - \frac{d}{d\beta}\psi(\beta) = \frac{1}{\beta^2}\left(H - \mathbb{E}_{q(\beta)}\left[H\right]\right)\ .$$

  Fisher's score is proportional to the fluctuation of H in this case.

[5] See sec. 28 of Lev D. Landau and Eugenij M. Lifshits. *Course of Theoretical Physics. Statistical Physics.*, volume V. Butterworth-Heinemann, 3rd edition, 1980

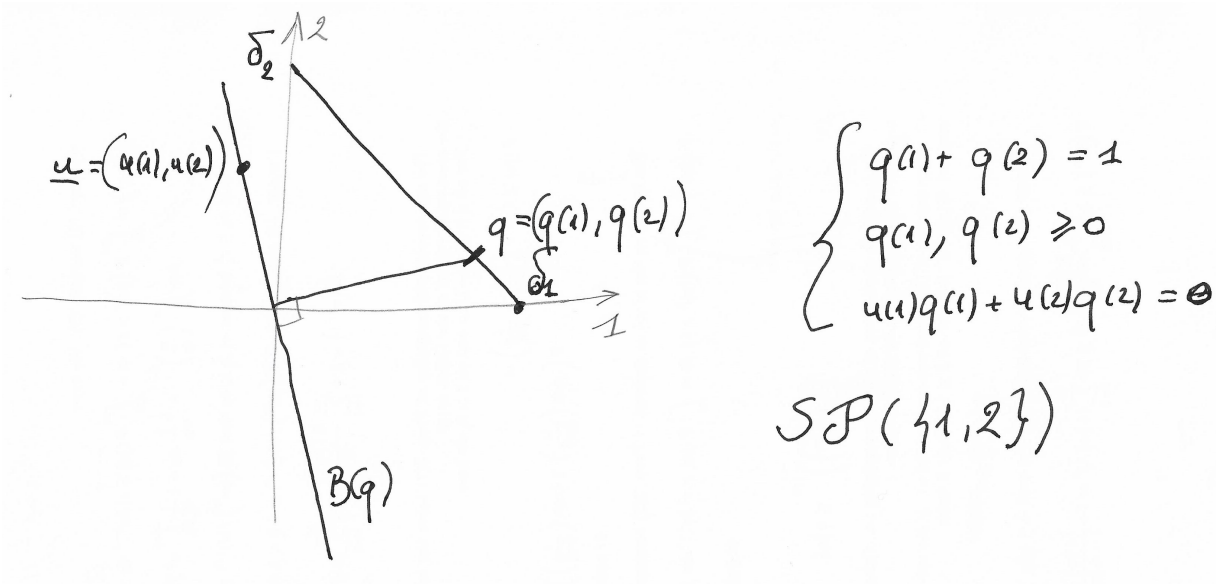
Statistical bundle, contrasts bundle

We are led to consider for each $q \in \mathcal{P}_> (\Omega)$ the vector space of q-contrasts, that is, the random variables which are centered for q,

$$B_q = \left\{ u \in \mathbb{R}^\Omega \,\middle|\, \sum_{x \in \Omega} u(x) q(x) = 0 \right\} .$$

The statistical bundle of $\mathcal{P}_> (\Omega)$ is the bundle $S\mathcal{P}_> (\Omega) \to \mathcal{P}_> (\Omega)$,

$$S\mathcal{P}_> (\Omega) = \left\{ (q, v) \in \mathcal{P}_> (\Omega) \times \mathbb{R}^\Omega \,\middle|\, \mathbb{E}_q [u] = 0 \right\}$$



Parallel transport

- Each fiber B_q represents the tangent vectors (velocities) at q. We need to introduce "connections" between different "expressions" of the tangent space.
- Each fiber B_q has an inner product $\langle u, v \rangle_q = \mathbb{E}_q[uv]$. The mapping

$$(q, u, v) \mapsto \langle u, v \rangle_q$$

  is called the metric of the statistical bundle.

For all $(p, u), (q, v) \in S\mathcal{P}_>(\Omega)$,

- ${}^{\mathrm{e}}\mathbb{U}_p^q \colon B_p \ni u \mapsto u - \mathbb{E}_q[u] \in B_q$ is the exponential transport;
- ${}^{\mathrm{m}}\mathbb{U}_q^p \colon B_p \ni v \mapsto \frac{q}{p} v \in B_p$ is the mixture transport.

For all $(p, u), (q, v) \in S\mathcal{P}_>(\Omega)$,

$$\left\langle u, {}^{\mathrm{m}}\mathbb{U}_q^p v \right\rangle_p = \left\langle {}^{\mathrm{e}}\mathbb{U}_p^q u, v \right\rangle_q$$

Exercise

- model: $\theta \mapsto q(\theta)$, $\theta \in I \subset \mathbb{R}$
- Fisher's score, velocity: $\overset{\star}{q}(\theta) = \frac{d}{d\theta} \log q(\theta) = \frac{\dot{q}(\theta)}{q(\theta)}$
- Fisher's information:

$$\mathbb{E}_{q(\theta)}\left[\left(\overset{\star}{q}(\theta)\right)^2\right] = \langle \overset{\star}{q}(\theta), \overset{\star}{q}(\theta) \rangle_{q(\theta)} =$$

$$\sum_x \left(\frac{\dot{q}(x;\theta)}{q(x;\theta)}\right)^2 q(x;\theta) = \sum_x \frac{\left(\dot{q}(x;\theta)\right)^2}{q(x;\theta)}$$

- Duality with $u \in B_p$ and $v \in B_p$:

$$\left\langle u, {}^{\mathrm{m}}\mathbb{U}_q^p v \right\rangle_p = \sum_x u(x) \frac{q(x)}{p(x)} v(x) p(x) = \sum_x u(x) v(x) q(x) =$$

$$\sum_x \left(u(x) - \mathbb{E}_q\left[u\right]\right) v(x) q(x) = \left\langle {}^{\mathrm{e}}\mathbb{U}_p^q u, v \right\rangle_q$$

Affine space with a parallel transport

Because of what was said about the notion of Fisher's score and the bundle of contrasts, it is useful to rephrase Weyl's definition by moving from the trivial tangent bundle $M \times V$ to the nontrivial bundle of contrasts or fluctuations.[6]

Affine space

Let M be a set and B_p, $p \in M$, a family of topological vector spaces (top-linear spaces). Let $(\mathbb{U}_p^q)$, $p, q \in M$ be a cocycle of top-linear isomorphism $\mathbb{U}_p^q \colon B_p \to B_q$, $\mathbb{U}_q^p \mathbb{U}_p^q = I_{B_p}$. Define a displacement mapping

$$\mathbb{S} \colon M \times M \ni (p, q) \mapsto s_p(q) \in B_p$$

so that:

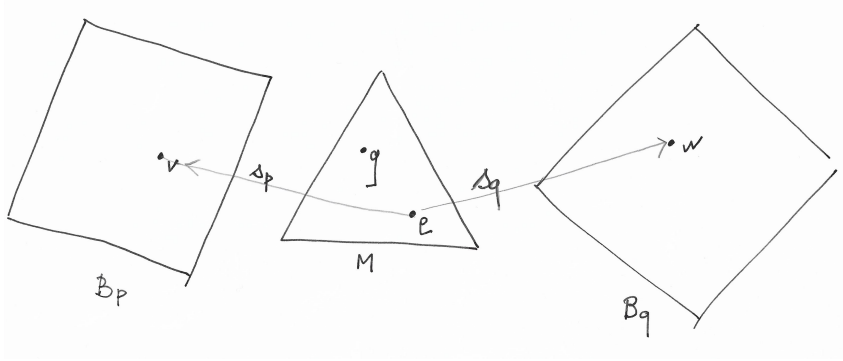
1. For each fixed p the mapping $s_p \colon q \mapsto s_p(q) = \mathbb{S}(s, p)$ is injective
2. (Parallelogram law) $\mathbb{S}(p, q) + \mathbb{U}_q^p \mathbb{S}(q, r) = \mathbb{S}(p, r)$

[6] While still staying in the finite state space case, I start using a more general language taken from infinite-dimensional differential geometry. See Serge Lang. *Differential and Riemannian manifolds*, volume 160 of *Graduate Texts in Mathematics*. Springer-Verlag, third edition, 1995

Affine manifold I

The affine space provides an atlas of charts $s_p \colon M \to B_p$, $p \in M$. The change-of-chart map is $s_p \circ s_q^{-1}$. At $\rho = s_q^{-1}(w)$, $w \in B_q$, it holds

$$s_p \circ s_q^{-1}(w) = s_p(\rho) = s_p(q) + \mathbb{U}_q^p s_q(\rho) = s_p(q) + \mathbb{U}_q^p w \ .$$



The change-of-origin map is the restriction of an affine map whose linear part is the parallel transport.

Assume that the vector fibers of the affine space $(M, (B_p), \mathbb{S})$ are Banach spaces and assume that for each p, $s_p M$ is a neighborhood of 0 in B_μ. Define $U_p = s_p^{-1}\left(s_p(M)^\circ\right)$. Then $(s_p \colon U_p)$ is a chart on M. The charts are compatible, and the resulting manifold is the affine manifold of the affine space.

Affine bundle

The specific form of the atlas defining the affine manifold allows the extension of the same atlas to define an affine bundle.

Given the affine manifold $\mathcal{M}$, consider the set

$$\{(p, v) \,|\, p \in M, v \in B_q\}$$

and, for each $p \in M$ define the chart

$$s_p(q, v) = (s_p(q), \mathbb{U}_p^q v) \in B_p \times B_p$$

to define the manifold $S\,\mathcal{M}$.

The statistical bundle of the affine manifold is where we define the velocity of a curve. The affine bundle has a tangent bundle where we define the second order geometry that is, accellerations.

Exponential affine space I

- We define the exponential displacement on $\mathcal{P}_{>}(\Omega)$ by

$$\mathcal{P}_{>}(\Omega) \times \mathcal{P}_{>}(\Omega) \ni (p, q) \mapsto s_p(q) = \log \frac{q}{p} - \mathbb{E}_p\left[\log \frac{q}{p}\right] \in B_p \ ,$$

  and the exponential transport by

$$^{\mathrm{e}}\mathbb{U}_p^q \colon B_p \ni v \mapsto v - \mathbb{E}_q[v] \in B_q \ .$$

- The (generalized) parallelogram law is

$$\left(\log \frac{q}{p} - \mathbb{E}_p\left[\log \frac{q}{p}\right]\right) + {}^{\mathrm{e}}\mathbb{U}_q^p\left(\log \frac{r}{q} - \mathbb{E}_q\left[\log \frac{r}{q}\right]\right) =$$
$$\left(\log \frac{q}{p} - \mathbb{E}_p\left[\log \frac{q}{p}\right]\right) + \left(\log \frac{r}{q} - \mathbb{E}_p\left[\log \frac{q}{p}\right]\right) =$$
$$\log \frac{q}{p} - \mathbb{E}_p\left[\log \frac{r}{p}\right] \ .$$

Exponential affine space II

- The inverse chart is defined on all of B_p by

$$s_p^{-1}(v) = \exp\left(v - K_p(v)\right) \cdot p = e_p(v) \ , \quad K_p(v) = \log \mathbb{E}_p\left[\mathrm{e}^v\right] \ .$$

- The cumulant functional K_p has several important well-known properties.

1. $\mathrm{D}\left(p \,\|q\right) = \mathbb{E}_p\left[\log \frac{p}{q}\right] = \mathbb{E}_p\left[\log \frac{p}{\exp(v - K_p(v)) \cdot p}\right] = K_p(v)$, that is,
$$K_p(s_p(q)) = \mathrm{D}\left(p \,\|q\right) \ .$$
2. $dK_p(v)[h] = \mathbb{E}_{e_p(v)}\left[h\right] = \left\langle \frac{e_p(v)}{p} - 1, h \right\rangle_{e_p(v)}$
3. $d^2K_p(v)[h,k] = \mathrm{Cov}_{e_p(v)}\left(h,k\right) = \left\langle {}^{\mathrm{e}}\mathbb{U}_p^{e_p(v)} h, {}^{\mathrm{e}}\mathbb{U}_p^{e_p(v)} k \right\rangle_{e_p(v)} = \left\langle h, \left({}^{\mathrm{m}}\mathbb{U}_{e_p(v)}^{p} \, {}^{\mathrm{e}}\mathbb{U}_p^{e_p(v)}\right) k \right\rangle_p .$

Mixture affine space

- We define the mixture displacement on $\mathcal{P}_>(\Omega)$ by

$$\mathcal{P}_>(\Omega) \times \mathcal{P}_>(\Omega) \ni (p, q) \mapsto \eta_p(q) = \frac{q}{p} - 1 \in B_p \ ,$$

  and the mixture transport by

$${}^{\mathrm{m}}\mathbb{U}_p^q \colon B_p \ni v \mapsto \frac{p}{q}\, v \in B_q \ .$$

- The (generalized) parallelogram law is

$$\left(\frac{q}{p} - 1\right) + \frac{q}{p}\left(\frac{r}{q} - 1\right) = \left(\frac{r}{p} - 1\right) \ .$$

- The inverse chart is defined for all $v > 1$, $w \in B_p$ by

$$\eta_p^{-1}(v) = (1 + v) \cdot p \ .$$

Kinematics I

Definition

The velocity of the smooth curve $t \mapsto \gamma(t)$ of the affine manifold is the curve $t \mapsto (\gamma(t), \dot{\gamma}(t))$ of the affine bundle whose second component is

$$\dot{\gamma}(t) = \lim_{h \to 0} h^{-1}(s_{\gamma(t)}(\gamma(t+h)) = \left. \frac{d}{dh} s_{\gamma(t)}(\gamma(t+h)) \right|_{h=0} .$$

Definition

Let F be a section of the affine bundle (vector field), that is, $(p, F(p)) \in \mathcal{S}\mathcal{M}$, $p \in M$. An integral curve of the section F is a curve $t \mapsto \gamma(t)$ such that $\dot{\gamma}(t) = F(\gamma(t))$. A flow of the section F is a mapping

$$M \times I \ni (\nu, t) \mapsto \Gamma_t(\nu)$$

such that for each ν the curve $t \mapsto \Gamma_t(\nu)$ is an integral curve and $\gamma(0, \nu) = \nu$.

Kinematics II

Definition

A curve $I : t \mapsto \gamma(t)$ is auto-parallel in the affine bundle if

$$\overset{\star}{\gamma}(t) = \mathbb{U}^{\gamma(t)}_{\gamma(s)} \overset{\star}{\gamma}(s) \quad s, t \in I \ .$$

Proposition

The following conditions are equivalent.

1. The curve γ is autoparallel.
2. The expression of the cuve in each chart is affine.
3. For all s, t, $\gamma(t) = s^{-1}_{\gamma(s)} \left((t - s) \overset{\star}{\gamma}(s) \right)$.

Consider the curve $t \mapsto \gamma(t)$ with velocity $t \mapsto \overset{\star}{\gamma}(t)$. The acceleration $t \mapsto \overset{\star\star}{\gamma}(t)$ is the velocity $t \mapsto (\mu(t), \overset{\star}{\gamma}(t))$.

Examples of Kinematics I

- Velocity in the mixture affine manifold

$$\overset{\star}{p}(t) = \lim_{h \to 0} h^{-1} \, \mathbb{S}(p(t), p(t+h)) =$$

$$\lim_{h \to 0} h^{-1} \left(\frac{p(t+h)}{p(t)} - 1 \right) = \frac{\dot{p}(t)}{p(t)}$$

- Velocity in the exponential affine manifold

$$\overset{\star}{p}(t) = \lim_{h \to 0} h^{-1} \, \mathbb{S}(p(t), p(t+h)) =$$

$$\lim_{h \to 0} h^{-1} \left(\log \frac{p(t+h)}{p(t)} - \int \log \frac{p(t+h)}{p(t)} \; p(t) \; dm \right) = \frac{\dot{p}(t)}{p(t)}$$

Examples of Kinematics II

- It is remarkable that the expression of the velocity is the same in both cases.The exponential velocity for a curve of the form of a Gibbs model $p(t) \propto \mathrm{e}^{\alpha(t)U} \cdot p$, that is $p(t) = \mathrm{e}^{\alpha(t)U - \psi(t)} \cdot p$, is

$$\overset{\star}{p}(t) = \frac{d}{dt}\left(\alpha(t)U - \psi(t)\right) =$$

$$\dot{\alpha}(t)U - \dot{\psi}(t) = \dot{\alpha}(t)\left(U - \int U\ p(t)\ dm\right)\ .$$

  In this case, the quantity $\overset{\star}{p}(t)$ is seen as $\dot{\alpha}(t)$ times the fluctuation $\left(U - \int U\ p(t)\ dm\right)$.

- In the mixture case, the acceleration is

$$\overset{\star\star}{p}(t) = {}^{\mathrm{m}}\mathbb{U}_{p}^{p(t)} \frac{d}{dt} {}^{\mathrm{m}}\mathbb{U}_{p(t)}^{p} \overset{\star}{p}(t) = \frac{p}{p(t)} \frac{d}{dt} \frac{p(t)}{p} \frac{\dot{p}(t)}{p(t)} = \frac{\ddot{p}(t)}{p(t)}$$

Examples of Kinematics III

- In the exponential case, the acceleration is

$$\overset{**}{p}(t) = {}^{e}\mathbb{U}_{p}^{p(t)} \frac{d}{dt} {}^{e}\mathbb{U}_{p(t)}^{p} \overset{\star}{p}(t) = {}^{e}\mathbb{U}_{p}^{p(t)} \frac{d}{dt} \left(\frac{\dot{p}(t)}{p(t)} - \int \frac{\dot{p}(t)}{p(t)} \; p \; dm \right) =$$

$$\frac{\ddot{p}(t)}{p(t)} - \left(\frac{\dot{p}(t)}{p(t)} \right)^2 - \int \left(\frac{\ddot{p}(t)}{p(t)} - \left(\frac{\dot{p}(t)}{p(t)} \right)^2 \right) \; p(t) \; dm =$$

$$\frac{\ddot{p}(t)}{p(t)} - \left[\left(\frac{\dot{p}(t)}{p(t)} \right)^2 - \int \left(\frac{\dot{p}(t)}{p(t)} \right)^2 \; p(t) \; dm \right]$$

- For the Gibbs model above, the exponential acceleration is proportional to the velocity, namely

$$\overset{**}{p}(t) = \ddot{\alpha}(t) \left(U - \int U \; p(t) \; dm \right) = \frac{\ddot{\alpha}(t)}{\dot{\alpha}(t)} \overset{\star}{p}(t) \; .$$

Examples of Kinematics IV

- The auto-parallel curves (geodesics) in the mixture geometry are of the form

$$\gamma(t) = \gamma(0) + \dot{\gamma}(0)t = (1 + \overset{\star}{\gamma}(0))\gamma(0) = (1-t)\gamma(0) + t\gamma(1)$$

  The last expression explains the name.

- In the exponential geometry, the form of the auto-parallel curve (geodesic) is

$$\gamma(t) = s^{-1}_{\gamma(0)}(t(\overset{\star}{\gamma}(0))) = \mathrm{e}^{t\overset{\star}{\gamma}(0) - K_{\gamma(0)}(\overset{\star}{\gamma}(0))} \cdot \gamma(0)$$

  that is, it is an exponential family.

- The auto-parallel (geodesic) interval is the Hellinger arc

$$\gamma(t) \propto \gamma(0)^{1-t}\gamma(1)^t$$

Example: gradient of the entropy I

The entropy of $q \in \mathcal{P}_{>}(X)$ is

$$\mathcal{H}(q) = \mathbb{E}_q[-\log q] = \sum_{x \in X} -\log q(x)\, q(x) = \sum_{x \in X} L(q(x))$$

By definition, the gradient $\operatorname{grad} \mathcal{H}$ of a scalar field

$$\mathcal{H} \colon \mathcal{P}_{>}(X) \to \mathbb{R}$$

is defined by

Example: gradient of the entropy II

$$
\begin{aligned}
\left\langle \operatorname{grad} \mathcal{H}\left(q(t)\right), \frac{\mathrm{D}}{dt} q(t) \right\rangle_{q(t)} &= \frac{d}{dt} \mathcal{H}\left(q(t)\right) \\
&= \frac{d}{dt} \sum_{x \in X} L(q(x;t)) \\
&= \sum_{x \in X} L'(q(x;t)) \frac{d}{dt} q(x;t) \\
&= \sum_{x \in X} (-1 - \log\left(q(x;t)\right)) \frac{\mathrm{D}}{dt} q(x;t)\, q(x;t) \\
&= \mathbb{E}_{q(t)} \left[-(1 + \log\left(q(t)\right)) \frac{\mathrm{D}}{dt} q(t) \right]
\end{aligned}
$$

where $t \mapsto q(t)$ is a generic smooth curve.
The gradient must belong to B_q so that

$$
\operatorname{grad} \mathcal{H}\left(q\right) = -(1 + \log\left(q\right)) - \mathbb{E}_q\left[-(1 + \log\left(q\right))\right] = -\log\left(q\right) - \mathcal{H}\left(q\right)
$$

Example: gradient of the entropy III

Notice that $\operatorname{grad} \mathcal{H}(m) = 0$ provided m is the uniform probability function.
The gradient flow equation of the entropy is

$$\boxed{\frac{\mathrm{D}}{dt} q(t) = -\operatorname{grad} \mathcal{H}(q(t))}$$

which is a replicator equation[7]

$$\frac{d}{dt} q(x;t) = q(x;t) \left(-\log(q(x;t)) - \sum_{y \in X} -\log(q(y;t))\ q(y;t) \right) .$$

If $q(0) = q_0$, then

$$\overset{\star}{q}(0) = \left. \frac{\mathrm{D}}{dt} q(t) \right|_{t=0} = -\log(q_0) - \mathcal{H}(q_0)$$

Example: gradient of the entropy IV

A computation shows that the solution of the gradient flow equation is a (non-canonical) exponential family of the form

$$q(t) = \exp\left(\theta(t)\overset{\star}{q}(0) - K_{q_0}(\theta(t)\overset{\star}{q}(0))\right) \cdot q_0 \ .$$

with $\theta(0) = 0$

- The first step is to compute the velocity. From

$$\log q(t) = \theta(t)\overset{\star}{q}(0) - K_{q_0}(\theta(t)\overset{\star}{q}(0)) + \log q_0$$

and the equation for the derivative of the cumulant function, the velocity is

$$\overset{\star}{q}(t) = \dot{\theta}(t)\overset{\star}{q}(0) - \mathbb{E}_{q(t)}\left[\dot{\theta}(t)\overset{\star}{q}(0)\right] = \dot{\theta}(t)\left(-\log q_0 + \mathbb{E}_{q(t)}\left[\log q_0\right]\right)$$

- The entropy is

$$\mathcal{H}\left(q(t)\right) = \mathbb{E}_{q(t)}\left[-\log q(t)\right] = \\ \mathbb{E}_{q(t)}\left[-\theta(t)\overset{\star}{q}(0) - \log q_0\right] + K_{q_0}(\theta(t)\overset{\star}{q}(0))$$

Example: gradient of the entropy V

- The gradient of the entropy is

$$
\begin{aligned}
&-\log q(t) - \mathcal{H}\left(q(t)\right) = \\
&\qquad \theta(t)(-\overset{\star}{q}(0) + \mathbb{E}_{q(t)}\left[\overset{\star}{q}(0)\right]) - \log q_0 + \mathbb{E}_{q(t)}\left[\log q_0\right] = \\
&\qquad\qquad (\theta(t) - 1)\left(\log q(0) - \mathbb{E}_{q(t)}\left[\log q(0)\right]\right)
\end{aligned}
$$

- The gradient flow equation is satisfied by

$$\dot{\theta}(t) = \theta(t) - 1 \Rightarrow \theta(t) = \mathrm{e}^t - 1$$

This in turn, gives

$$q(t) \propto q_0^{\mathrm{e}^t} \ .$$

Computing the limits $t \to \pm\infty$ is also interesting. Precisely, $\lim_{t\to-\infty} q(t) = m$, and the other case follows from

$$\left(\frac{q_0(x)}{\max_x q_0(x)}\right)^{\mathrm{e}^t} \to \begin{cases} 1 & \text{if } q_0(x) = \max_x q_0(x) \\ 0 & \text{if } q_0(x) < \max_x q_0(x) \end{cases}$$

[7] See, for example, §6.2 of Nihat Ay, Jürgen Jost, Hông Vân Lê, and Lorenz Schwachhöfer. *Information geometry*, volume 64 of *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics [Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics]*. Springer, Cham, 2017

PART 2

ANOVA with two non-independent factors I

- Consider a finite product sample space $\Omega = \Omega_1 \times \Omega_2$ with a joint probability function

$$q \colon \Omega_1 \times \Omega_2 \ni (x_1, x_2) \mapsto q(x_1, x_2) \ .$$

  We denote the two margins by

$$q_1(x_1) = \sum_y f(x_1, y) \ , \quad q_2(x_2) = \sum_x q(x, x_2) \ .$$

- We focus on the case $q \neq q_1 \otimes q_2$.
- For each random variable $f \in L^2(q)$ we look for an orthogonal decomposition of the form

$$f(x_1, x_2) = f_0 \oplus (f_1(x) + f_2(x_2)) \oplus f_{12}(x_1, x_2)$$

- Notice that we do not require $f_1 \perp f_2$ as it is done in case of independence, cf Hajek:1968 and Sobol':2001.

ANOVA with two non-independent factors II

- We call factors the two projections

$$X_1(x_1, x_2) = x_1 \ , \quad X_2(x_1, x_2) = x_2 \ .$$

- Consider the subsets of $\{1, 2\}$, that is $\emptyset, \{1\}, \{2\}, \{1, 2\}$, partially ordered by inclusion. Let X_I be the components projection on I, $X_I = (X_j : j \in I)$. Each I is an interaction.
- An effect of the interaction I is a random variable of the form $f \circ X_I$ which is q-orthogonal to all $g \circ X_J$ for all $J \prec I$, that is, $J \subset I$ and $J \neq I$.
- The order of the interaction I is $\#I$. Let H_k be the vector space generated by the I-interactions of order k.
  - An effect of order 0 is a grand mean;
  - an effect of order 1 is a simple effect;
  - an effect of order 2 is an interaction.
- H_0 contains random variables which do not depend on any X_j $j = 1, 2$. that is, $H_0 = \mathbb{R}$.

ANOVA with two non-independent factors III

- The space H_1 is generated by the random variables of the form $f_1 \circ X_1$ and $f_2 \circ X_2$ with

$$\mathbb{E}_q [f_1 \circ X_1] = \mathbb{E}_{q_1} [f_1] = 0$$
$$\mathbb{E}_q [f_2 \circ X_2] = \mathbb{E}_{q_2} [f_2] = 0$$

  In fact, q-orthogonal to $H_0 = \mathbb{R}$ is the same as zero q-expectation.

- An element of H_1 is of the form

$$f_1 \circ X_1 + f_2 \circ X_2 \ , \quad f_1 \in L_0^2(q_1) \ , f_2 \in L_0^2(q_2)$$

- The representation above is unique. In fact, if

$$f_1(x_1) + f_2(x_2) = 0 \ , x_1 \in \Omega_1, x_2 \in \Omega_2 \ ,$$

  then both f_1 and f_2 must be constant. As the q-expectation is zero, $f_1 = f_2 = 0$.

ANOVA with two non-independent factors IV

- An element of H_2 is of the form $f_{12} \circ (X_1, X_2)$ with

$$f_{12} \circ (X_1, X_2) \perp H_\emptyset, H_{\{1\}}, H_{\{2\}}$$

The orthogonality with respect to $H_\emptyset$ implies zero q-expectation $\mathbb{E}_q[f_{12}] = 0$. The orthogonality with respect to $H_\emptyset + H_{\{1\}}$ and $H_\emptyset + H_{\{2\}}$ implies zero conditional expectation with respect to each factor:

$$\mathrm{E}_q(f_{12} \circ (X_1, X_2)|X_1) = 0 \ , \quad \mathrm{E}_q(f_{12} \circ (X_1, X_2)|X_2) = 0$$

- We look for a decomposition of $f \in L^2(q)$ of the form

$$0 = f_0 + (f_1 \circ X_1 + f_2 \circ X_2) + f_{12} \circ (X_1, X_2)$$

with $f_0 \in H_0$, $(f_1 \circ X_1 + f_2 \circ X_2) \in H_1$, and $f_{12} \circ (X_1, X_2) \in H_2$.

ANOVA with two non-independent factors V

- Let $f \mapsto \text{Hajek}\,(q)\,f$ be the orthogonal projection of $L^2(q)$ onto H_1, the Hajek projection. It is a general device to approximate a complex random variable with an additive model, cf Hajek:1968 and Efron-Stein:1981.[8]
- Then, the orthogonal decomposition is

$$f = \mathbb{E}_q\, f + \text{Hajek}\,(q)\,f + (\text{I} - \mathbb{E}_q - \text{Hajek}\,(q))f \ .$$

- We want to compute the Hajek projection. It is a least square problem

$$\min \mathbb{E}_q \left[\left| f - f_0 - f_1 \circ X_1 - f_2 \circ X_2 \right|^2 \right]$$
$$\text{s.t.} f_0 \in \mathbb{R}, \mathbb{E}_{q_1}\left[f_1\right] = 0, \mathbb{E}_{q_2}\left[f_2\right] = 0$$

ANOVA with two non-independent factors VI

- The following equations are the gradient equations of the least square problem. They also derive from conditioning the decomposition.

$$\begin{aligned}
\mathbb{E}_q[f] &= f_0 \\
\mathrm{E}_q(f|X_1) &= f_0 + f_1 \circ X_1 + \mathrm{E}_q(f_2 \circ X_2|X_1) \\
\mathrm{E}_q(f|X_2) &= f_0 + \mathrm{E}_q(f_1 \circ X_1|X_2) + f_2 \circ X_2
\end{aligned}$$

- Assume $q = q_1 \otimes q_2$. The system of equations becomes

$$\begin{aligned}
\mathbb{E}_q[f] &= f_0 \\
\mathrm{E}_q(f|X_1) &= f_0 + f_1 \circ X_1 + \cancel{\mathbb{E}_{q_2}[f_2]} \\
\mathrm{E}_q(f|X_2) &= f_0 + \cancel{\mathbb{E}_{q_1}[f_1]} + f_2 \circ X_2 \ .
\end{aligned}$$

ANOVA with two non-independent factors VII

- The decomposition becomes

$$f = \mathbb{E}_q[f] + \\ (\mathsf{E}_q(f - \mathbb{E}_q[f]|X_1) + \mathsf{E}_q(f - \mathbb{E}_q[f]|X_2)) + \\ (f - \mathbb{E}_q[f] - \mathsf{E}_q(f|X_1) + \mathsf{E}_q(f|X_2)) \ .$$

[8] I suspect that it is what some call mean field approximation.

Computing the Hajek projection

- If we assume $\mathbb{E}_q[f] = 0$, the gradient equations for the Hajek projection are[9]

$$\mathrm{E}_q(f|X_1) = f_1 \circ X_1 + \mathrm{E}_q(f_2 \circ X_2|X_1)$$
$$\mathrm{E}_q(f|X_2) = \mathrm{E}_q(f_1 \circ X_1|X_2) + f_2 \circ X_2$$

- With the kernel expression $q = k \cdot q_1 \otimes q_2$

$$\sum_y f(x_1, y)\, k(x_1, y) q_2(y) = f_1(x_1) + \sum_y f_2(y)\, k(x_1, y) q_2(y)\ ,$$
$$\sum_x f(x, x_2)\, k(x, x_2) q_1(x) = \sum_x f_1(x)\, k(x, x_2) q_1(x) + f_2(x_2)\ .$$

[9] Giovanni Pistone. Information geometry of smooth densities on the Gaussian space: Poincaré inequalities. In *Signals and Communication Technology*, pages 1–17. Springer International Publishing, 2021

Aside on Linear Programming[10] I

- The primal problem in canonical form is

$$\text{Find} \quad c = \inf \sum_x c(x) r(x)$$

$$\text{Subject to} \quad \sum_x A(y,x) r(x) = \beta(y) \quad y \in Y$$

$$r(x) \geq 0$$

- r is the primal plan
- a plan is feasible if the constraints hold
- The dual problem in standard form is

$$\text{Find} \quad \beta = \sup \sum_y \beta(y) \lambda(y)$$

$$\text{Subject to} \quad \sum_y A(y,x) \lambda(y) \leq c(x)$$

Aside on Linear Programming[11] II

- λ is the dual plan

Strong duality theorem

If a feasible primal plan exists, then $c = \beta$. If moreover, $c > -\infty$, then primal optimal and dual optimal plans exist.

[10] See §IV.8 in Alexander Barvinok. *A course in convexity*, volume 54 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2002

[11] See §IV.8 in Alexander Barvinok. *A course in convexity*, volume 54 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2002

Fixed margins I

- Assume a product sample space $X = \Omega_1 \times \Omega_2$ and consider the probability simplex $\mathcal{P}(\Omega_1 \times \Omega_2)$. The two marginalisation mappings are

$$X_1 : \mathcal{P}(\Omega_1 \times \Omega_2) \ni q(\cdot,\cdot) \mapsto \sum_{x_2} q(\cdot, x_2) \in \mathcal{P}(X_1)$$

$$X_2 : \mathcal{P}(\Omega_1 \times \Omega_2) \ni q(\cdot,\cdot) \mapsto \sum_{x_1} q(x_1, \cdot) \in \mathcal{P}(X_2)$$

- For each given $q_1 \in \mathcal{P}(X_1)$ and $q_2 \in \mathcal{P}(X_2)$ define

$$\Pi(q_1, q_2) = \{q \in \mathcal{P}(\Omega_1 \times \Omega_2) \,|\, X_1 q = q_1, X_2 q = q_2\} \ .$$

- The set of transport plans $\Pi(q_1, q_2)$ is
  - non-empty,
  - convex,
  - compact.

Fixed margins II

- The marginalization conditions

$$\sum_{x_2 \in X_2} q(y_1, x_2) = q_1(y_1) \quad y_1 \in X_1$$
$$\sum_{x_1 \in X_1} q(x_1, y_2) = q_2(y_2) \quad y_2 \in X_2$$

  can be written as

$$\sum_{(x_1,x_2) \in X_1 \times X_2} (y_1 = x_1) q(x_1, x_2) = q_1(y_1)$$
$$\sum_{(x_1,x_2) \in X_1 \times X_2} (y_2 = x_2) q(x_1, x_2) = q_2(y_2)$$

  thus identifying an operator $A\colon (X_1 \cup X_2) \times (\Omega_1 \times \Omega_2)$ with

$$\sum_{(x_1,x_2) \in \Omega_1 \times \Omega_2} A(y; x_1, x_2) q(x_1, x_2) = (q_1 \cup q_2)(y)$$

Kantorovich optimal transport I

- Given the cost $c\colon \Omega_1 \times \Omega_2 \to \mathbb{R}$, Kantorovich looks for the transport plan with minimal expected cost

$$\inf \left\{ \sum_{x_1,x_2} c(x_1,x_2) q(x_1,x_2) \,\middle|\, q \in \Pi(q_1,q_2) \right\}$$

- It is a primal problem in canonical form:

$$\begin{aligned} \text{Find} \quad & c = \inf \sum_{x_1,x_2} c(x_1,x_2) q(x_1,x_2) \\ \text{Subject to} \quad & \sum_{x} A(y;x_1,x_2) q(x_1,x_2) = (q_1 \cup q_2)(y) \\ & q(x_1,x_2) \geq 0 \ . \end{aligned}$$

Kantorovich optimal transport II

- The dual problem in standard form is

$$\text{Find} \quad \beta = \sup \sum_y (q_1 \cup q_2)(y)\lambda(y)$$

$$\text{Subject to} \quad \sum_y A(y; x_1, x_2)\lambda(y) \leq c(x_1, x_2)$$

- that is, given the form of A,

$$\text{Find} \quad \beta = \sup \left(\sum_{y_1 \in X_1} q_1(y_1)\lambda_1(y_1) + \sum_{y_2 \in X_2} q_2(y_2)\lambda_2(y_2) \right)$$

$$\text{Subject to} \quad \lambda_1(x_1) + \lambda_2(x_2) \leq c(x_1, x_2) \ .$$

- There exists a feasible transport plan, and the set of plans is compact. It follows that the full strong duality theorem holds.

Kantorovich optimal transport III

- Let $\bar{q}$, $\bar{\lambda}$ be the optimal plan and dual plan. The equality of values gives

$$\sum_{x_1,x_2} c(x_1,x_2)\bar{q}(x_1,x_2) =$$

$$\sum_{x_1 \in X_1} q_1(x_1)\bar{\lambda}_1(x_1) + \sum_{x_2 \in X_2} q_2(x_2)\bar{\lambda}_2(x_2) =$$

$$\sum_{x_1,x_2} \left(\bar{\lambda}_1(x_1) + \bar{\lambda}_2(x_2)\right) \bar{q}(x_1,x_2)$$

- Now, the inequality $c \geq \lambda_1 \oplus \lambda_2$ implies

$$c(x_1,x_2) = \bar{\lambda}_1(x_1) + \bar{\lambda}_2(x_2) \quad \text{provided } \bar{q}(x_1,x_2) \neq 0$$

Transport plans in $\mathcal{E}(\Omega_1 \times \Omega_2)$

- Define for all $q_1 \in \mathcal{E}(\Omega_1)$ and $q_2 \in \mathcal{E}(\Omega_2)$

$$\overset{\circ}{\Pi}(q_1, q_2) = \{q \in \mathcal{E}(\Omega_1 \times \Omega_2) \,|\, X_1 q = q_1, X_2 q = q_2\}$$

- $\overset{\circ}{\Pi}$ is the set of positive couplings or positive transport plans of q_1 and q_2.
- $\overset{\circ}{\Pi}$ is the base set of a sub-manifold of the affine statistical manifold on $\mathcal{E}(\Omega_1 \times \Omega_2)$.
- A sub-manifold of the affine manifold $(M, s_p, B_p, \mathbb{U}_p^q \colon p, q \in M)$ is a subset $N \subset M$ such that for each $q \in N$ there exists a smooth splitting of the fibre $B_q = S_q N \oplus R_q N$ and the vector space $S_q N$ is the set of all velocities of curves in N through q.
- Basic examples of sub-manifolds of affine statistical manifolds are mixture models and exponential families, to be discussed elsewhere. Here we discuss the set of couplings or transport plans.

$\overset{\circ}{\Pi}(q_1, q_2)$ as a sub-manifold of $\mathcal{E}(\Omega_1 \times \Omega_2)$

- Finite dimensional exponential families and finite-dimensional mixture models are notable examples of sub-manifolds of the affine manifold structure because they are flat in one of the two dual geometries.
- Notice that the orthogonal projection on the space of velocities provides the required splitting.
- Precisely, the space $S_q N$ of velocities is the affine expression of the tangent space to N.
- The set of couplings $\Pi(q_1, q_2)$ is an affine subset of the probability simplex. Hence it is a polytope, a convex set generated by a finite number of vertices.
- Hence, $\overset{\circ}{\Pi}(q_1, q_2)$ is an open mixture model.

Velocities of curves in $\overset{\circ}{\Pi}(q_1, q_2)$ I

- $t \mapsto q(t)$ is a smooth curve of $\mathcal{E}(\Omega_1 \times \Omega_2)$ with values in the set of strictly positive transport plans. We can say

$$t \mapsto q(t) \in \overset{\circ}{\Pi}(q_1, q_2)$$

- Recall Fisher's score properties,

$$\overset{\star}{q}(t) = \frac{d}{dt} \log q(t) = \frac{\dot{q}(t)}{q(t)}$$

$$\frac{d}{dt} \mathbb{E}_{q(t)}[f] = \left\langle f - \mathbb{E}_{q(t)}[f], \overset{\star}{q}(t) \right\rangle_{q(t)} .$$

- For each random variable depending on the first variable only, $f_1 \circ X_1$ it holds

$$0 = \frac{d}{dt} \mathbb{E}_{q_1}[f_1] = \frac{d}{dt} \mathbb{E}_{q(t)}[f_1 \circ X_1] =$$
$$\left\langle f_1 \circ X_1 - \mathbb{E}_{q(t)}[f_1 \circ X_1], \overset{\star}{q}(t) \right\rangle_{q(t)} = \mathbb{E}_{q(t)}[f_1 \circ X_1 \overset{\star}{q}(t)] ,$$

Similarly on the other projection.

Velocities of curves in $\overset{\circ}{\Pi}(q_1, q_2)$ II

- It follows that

$$\boxed{\mathbb{E}_{q(t)}\left[\overset{\star}{q}(t)|X_1\right] = 0 \quad \text{and} \quad \mathbb{E}_{q(t)}\left[\overset{\star}{q}(t)|X_2\right] = 0}$$

  that is, if $t \mapsto q(t) \in \overset{\circ}{\Pi}(q_1, q_2)$, then $\overset{\star}{q}(t)$ is an interaction at $q(t)$, $\overset{\star}{q}(t) \in H_2(q(t))$.

- Conversely, let $q \in \overset{\circ}{\Pi}(q_1, q_2)$ and $c_{12} \in H(q)$. The curve $t \mapsto (1 + tc) \cdot q$ is defined in a neighborhood of 0. The margins are correct,

$$\mathbb{E}_{(1+tc_{12})\cdot q}\left[g \circ X_i\right] = \mathbb{E}_q\left[(1 + tc_{12})g \circ X_i\right] = \mathbb{E}_{q_1}\left[g\right] \ ,$$

  and the velocity at 0 is c_{12},

$$\frac{d}{dt}\log\left((1 + tc_{12}) \cdot q\right)\Big|_{t=0} = \frac{c_{12}q}{(1 + tc_{12})q}\Big|_{t=0} = c_{12} \ .$$

Velocities of curves in $\overset{\circ}{\Pi}(q_1, q_2)$ III

- In conclusion, for all $q \in \overset{\circ}{\Pi}(q_1, q_2)$, the velocities fiber is the vector space of interactions,

$$\boxed{S_q \overset{\circ}{\Pi}(q_1, q_2) = H_2(q)}$$

- The orthogonal splitting of the statistical bundle is

$$S_q \mathcal{E}(\Omega_1 \times \Omega_2) = S_q \overset{\circ}{\Pi}(q_1, q_2) \oplus_q \text{Hajek}(q)\, S_q \mathcal{E}(\Omega_1 \times \Omega_2),$$

where q_1, q_2 are the margins of q.

- Notice that the complement fiber is $H_1(q)$,

$$\text{Hajek}(q)\, S_q \mathcal{E}(\Omega_1 \times \Omega_2) = \{f_1 \circ X_1 + f_2 \circ X_2 \,|\, \mathbb{E}_{q_1}[f_1] = \mathbb{E}_{q_2}[X_2] = 0\},$$

which is in turn related to the exponential families of additive statistics

$$\exp(f_1 \circ X_1 + f_2 \circ X_2 - K_q(f_1 \circ X_1 + f_2 \circ X_2)) \cdot q\ .$$

$\overset{\circ}{\Pi}(q_1, q_2)$ as an affine space I

- If $q, r \in \overset{\circ}{\Pi}(q_1, q_2)$, given $c \in H_2(q) = S_q\overset{\circ}{\Pi}(q_1, q_2)$,

$$\mathbb{E}_r\left[\left(\frac{q}{r}c\right) g_i \circ X_i\right] = \mathbb{E}_q\left[c\, g_i \circ X_i\right] = 0$$

  that is, $\frac{q}{r}c \in H_2(r) = S_r\overset{\circ}{\Pi}(q_1, q_2)$.

- We have defined a co-cycle of parallel transports on the bundle

$$S\overset{\circ}{\Pi}(q_1, q_2) = \left\{(q, c) \,\middle|\, q \in \overset{\circ}{\Pi}(q_1, q_2), c \in H_2(q)\right\}$$

- The dual transport is computed as follows. If

$$q, r \in \overset{\circ}{\Pi}(q_1, q_2)$$

$$c \in S_q\overset{\circ}{\Pi}(q_1, q_2) = H_2(q)$$

$$d \in S_r\overset{\circ}{\Pi}(q_1, q_2) = H_2(r)$$

$\overset{\circ}{\Pi}(q_1, q_2)$ as an affine space II

then

$$\left\langle {}^{\mathrm{m}}\mathbb{U}_q^r c, d \right\rangle_r = \mathbb{E}_q\left[cd\right] = \left\langle c, d - \mathrm{Hajek}\left(q\right) d \right\rangle_q$$

- Let us compute the mixture geodesic. If $(q, c) \in S\overset{\circ}{\Pi}(q_1, q_2)$, an m-geodesic is a curve in $t \mapsto q(t) \in \overset{\circ}{\Pi}(q_1, q_2)$ such that $(q(0), \overset{\star}{q}(0)) = (q, c)$ and $\overset{\star}{q}(t) = {}^{\mathrm{m}}\mathbb{U}_q^{q(t)} c$. It follows

$$\frac{\dot{q}(t)}{q(t)} = \frac{q}{q(t)} c \quad \text{then} \quad q(t) = (1 + tc) \cdot q \ .$$

  The m-geodesic from q in the direction c is $t \mapsto (1 + tc) \cdot q$.

- The the affine displacement is the geodesic at $t = 1$:

$$\overset{\circ}{\Pi}(q_1, q_2) \times \overset{\circ}{\Pi}(q_1, q_2) \ni (q, r) \mapsto \frac{r}{q} - 1$$

$\overset{\circ}{\Pi}(q_1, q_2)$ as an affine space III

- The e-geodesic from q in the direction d is the solution of

$$\overset{\star}{q}(t) = (I - \text{Hajek}\,(q(t)))d \ .$$

- A solution of this equation seems to require a solution of the Hajek projection.

Gradient flow of the OT problem I

- Let us discuss the Optimal Transport OT problem in the framework of the affine statistical bundle.
- $c\colon \Omega_1 \times \Omega_2 \to \mathbb{R}$ be a cost function and the expected cost function as a function of the probability function is

$$C\colon \mathcal{E}(\Omega_1 \times \Omega_2) \ni q \mapsto \mathbb{E}_q[c]$$

- The function $q \mapsto C(q)$ restricted to the open transport model $\overset{\circ}{\Pi}(q_1, q_2)$ has gradient in $S\overset{\circ}{\Pi}(q_1, q_2)$ given by

$$\operatorname{grad} C\colon q \mapsto c_{12}(q) = c - \mathbb{E}_q[c] - \operatorname{Hajek}(q)\, c$$

- In fact for all curve $t \mapsto q(t) \in \overset{\circ}{\Pi}(q_1, q_2)$, $q(0) = q$,

$$\frac{d}{dt} C(q(t)) = \frac{d}{dt} \mathbb{E}_{q(t)}[c] = \left\langle c - \mathbb{E}_{q(t)}[c], \overset{\star}{q}(t) \right\rangle_{q(t)} =$$
$$\left\langle c - \mathbb{E}_{q(t)}[c] - \operatorname{Hajek}(q(t))\, c, \overset{\star}{q}(t) \right\rangle_{q(t)}$$

Gradient flow of the OT problem II

- The equation of the gradient flow of C is

$$\overset{\star}{q}(t) = -c_{12}(q(t)) = -(c - \mathbb{E}_{q(t)}[c] - \text{Hajek}\,(q(t))\,c)$$

- Notice that $c_{12}(q) = c - \mathbb{E}_q[c] - \text{Hajek}\,(q)\,c$ is defined for all $q \in \Pi(q_1, q_2)$. If $\hat{q}$ is a zero of the extended gradient map,

$$\text{grad}\, C(\hat{q}) = c_{12}(\hat{q}) = 0\ ,$$

  then c equals the sum of two functions in one variable on the support of $\hat{q}$.

- We expect any solution $t \mapsto q(t)$ of the gradient flow equation to converge to a plan $\bar{q} = \lim_{t\to\infty} q(t) \in \Pi(q_1, q_2)$ such that $\mathbb{E}_{\bar{q}}[c]$ is the value of the Kantorovich optimal transport problem. The form of the gradient is compatible with the classical result in OT.

PART 3

Mechanics I

Classical mechanics systematically exploits the duality between the so called tangent and cotangent bundle.[12] The mixture bundle and the exponential bundle support the mechanics formalism.

Let be given two affine manifolds on the same base set M, $\mathcal{M}_i = (M, (B^i_\mu)_{\mu\in M}, ({}^i\mathbb{U}^\mu_\nu)_{\mu,\nu\in M}, {}^i\mathbb{S})$, $i = 1, 2$, and let be given for each $\mu \in M$ a duality pairing

$$B^1_\mu \times B^2_\mu \ni (u_1, u_2) \mapsto \langle u_1, u_2 \rangle_\mu \quad .$$

The affine manifolds $\mathcal{M}_1$ and M_2 are in duality if for all $\mu, \nu \in M$, $u \in B_\mu$, $v \in B^2_\nu$, it holds

$$\left\langle u, {}^2\mathbb{U}^\mu_\nu v \right\rangle_\mu = \left\langle {}^1\mathbb{U}^\nu_\mu u, v \right\rangle_\nu \quad .$$

Mechanics II

- Here, the mixture and the exponential fibers are equal, ${}^{\mathrm{m}}B_p = {}^{\mathrm{e}}B_p = B_p$, and the separating pairing is $\langle u, v\rangle_p = \int u\ v\ p\ dm$. The mixture affine manifold and the exponential affine manifold are dual. For $u \in B_p$ and $v \in B_q$

$$\left\langle {}^{\mathrm{m}}\mathbb{U}_p^q u, v\right\rangle_q = \int \frac{p}{q} u\ v\ q\ dm = \int u\ v\ p\ dm =$$
$$\int u \left(v - \int vp\, dm\right)\ p\ dm = \left\langle u, {}^{\mathrm{e}}\mathbb{U}_q^p v\right\rangle_p\ .$$

Definition

Consider a M which is base of two dual affine manifolds $\mathcal{M}_1$ and $\mathcal{M}_2$. A real function ϕ on $\mathcal{M}_1$ has a gradient $\operatorname{grad}\phi$ if $\operatorname{grad}\phi$ is a section of the affine bundle $S\,\mathcal{M}_2$ and for all smooth curve $t \mapsto \gamma(t) \in M$ it holds

$$\frac{d}{dt}\phi(\gamma(t)) = \langle \operatorname{grad}\phi(\gamma(t)), \overset{\star}{\gamma}(t)\rangle_\sigma\ .$$

Mechanics III

- The full bundle is

$$ {}^1S^1\,\mathcal{E}\,(\mu) = \{(q,\eta,w)\,|\,q\in\mathcal{E}\,(\mu)\,,\eta\in{}^*S_q\,\mathcal{E}\,(\mu)\,,w\in S_q\,\mathcal{E}\,(\mu)\}\ . $$

- In terms of the exponential parallel transport, we define an exponential covariant derivative by setting

$$ \frac{\mathrm{D}}{dt}w(t) = {}^{\mathrm{e}}\mathbb{U}_p^{q(t)}\frac{d}{dt}\,{}^{\mathrm{e}}\mathbb{U}_{q(t)}^{p}w(t) = $$
$$ {}^{\mathrm{e}}\mathbb{U}_p^{q(t)}\left(\dot{w}(t)-\mathbb{E}_p\left[\dot{w}(t)\right]\right) = \dot{w}(t)-\mathbb{E}_{q(t)}\left[\dot{w}(t)\right]\ . $$

- Let us do now the computation in the mixture bundle. The curve is

$$ t\mapsto\zeta(t) = (q(t),\eta(t))\in{}^*S\,\mathcal{E}\,(\mu) = {}^1S^0\,\mathcal{E}\,(\mu) $$

and the mixture covariant derivative as

$$ \frac{\mathrm{D}}{dt}\eta(t) = {}^{\mathrm{m}}\mathbb{U}_p^{q(t)}\frac{d}{dt}\,{}^{\mathrm{m}}\mathbb{U}_{q(t)}^{p}\eta(t) = $$
$$ \frac{p}{q(t)}\frac{1}{p}\left(\dot{q}(t)\eta(t)+q(t)\dot{\eta}(t)\right) = \overset{\star}{q}(t)\eta(t)+\dot{\eta}(t)\ . $$

Mechanics IV

For each smooth curve in the full statistical bundle,

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

it holds

$$\frac{d}{dt}\,\langle \eta(t), w(t)\rangle_{q(t)} = \left\langle \frac{\mathrm{D_m}}{dt}\eta(t), w(t)\right\rangle_{q(t)} + \left\langle \eta(t), \frac{\mathrm{D_e}}{dt} w(t)\right\rangle_{q(t)} \ .$$

- Given a scalar field $F\colon\ \mathcal{E}\,(\mu) \to \mathbb{R}$ the gradient of F is the section $q \mapsto \operatorname{grad} F(q)$ of the mixture bundle ${}^*S\,\mathcal{E}\,(\mu)$ such that for all smooth curve $t \mapsto q(t) \in \mathcal{E}\,(\mu)$ it holds

$$\frac{d}{dt} F(q(t)) = \langle \operatorname{grad} F(q(t)), \overset{\star}{q}(t)\rangle_{q(t)} \ .$$

Mechanics V

- Let be given a real function $F\colon {}^1S^1\,\mathcal{E}\,(\mu)\times\mathcal{D}\to\mathbb{R}$, where $\mathcal{D}$ a domain of $\mathbb{R}^k$. For a generic smooth curve

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

we want to write

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

Mechanics VI

where the four components of the gradient are

$$
{}^{1}S^{1}\,\mathcal{E}\,(\mu)\times\mathcal{D}\ni(q,\eta,w,c)\mapsto\begin{cases}(q,\operatorname{grad}F\big(q,\eta,w,c\big))\in{}^{*}S_{q}\,\mathcal{E}\,(\mu)\\(q,\operatorname{grad}_{\mathsf{m}}F\big(q,\eta,w,c\big))\in S_{q}\,\mathcal{E}\,(\mu)\\(q,\operatorname{grad}_{\mathsf{e}}F\big(q,\eta,w,c\big))\in{}^{*}S_{q}\,\mathcal{E}\,(\mu)\\(q,\nabla F\big(q,\eta,w,c\big))\in\mathcal{E}\,(\mu)\times\mathbb{R}^{k}\end{cases}
$$

Mechanics VII

In the total derivative,

1. $\operatorname{grad} F\big(q, \eta, w, c\big)$ is the natural gradient of

$$q \mapsto F(q, \mathbb{U}_p^q \zeta, \mathbb{U}_p^q v, c) \ ,$$

that is, with the representation in p-chart

$$F_p(u, \zeta, w, c) = F(\mathrm{e}_p(u), \mathbb{U}_p^{\mathrm{e}_p(u)} \zeta, \mathbb{U}_p^{\mathrm{e}_p(u)} v, c) \ ,$$

it is defined by

$$\langle \operatorname{grad} F(q, \zeta, w, c), \overset{\star}{q} \rangle_q = d_1 F_p(u, \zeta, w, c) \left[\mathbb{U}_q^p \overset{\star}{q}\right] \ , \quad (q, \overset{\star}{q}) \in S\,\mathcal{E}\,(\mu) \ ;$$

2. $\operatorname{grad}_{\mathsf{m}} F\big(q, \eta, w, c\big)$ and $\operatorname{grad}_{\mathsf{e}} F\big(q, \eta, w, c\big)$ are the fiber gradients;
3. $\nabla F\big(q, \eta, w, c\big)$ is the Euclidean gradient w.r.t. the last variable.

Mechanics VIII

- The dually affine geometry of the statistical bundle is naturally well suited for describing the dynamics of probability densities in a Lagrangian and Hamiltonian formalism.
- The Lagrangian formulation of mechanics derives the fundamental laws of force balance from variational principles. In our context, the exponential model $\mathcal{E}(\mu)$ corresponds to the configuration space, while the statistical bundle is associated to the velocity phase space.
- For a given smooth curve $q\colon [0,1] \ni t \mapsto q(t)$ in $\mathcal{E}(\mu)$ and its lift $t \mapsto (q(t), \overset{\star}{q}(t)) \in S\,\mathcal{E}(\mu)$, we introduce a generic Lagrangian function

$$L(q(t), \overset{\star}{q}(t))\colon S\,\mathcal{E}(\mu) \times [0,1] \to \mathbb{R}$$

and define an action as the integral of the Lagrangian along the curve over the fixed time interval $[0,1]$,

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

Mechanics IX

- Hamilton's principle states that this function has a critical point at a solution within the space of curves on $\mathcal{E}(\mu)$. If q is an extremal of the action integral, then

$$\frac{\mathrm{D}}{dt}\,\mathrm{grad}_{\mathrm{e}}\,L(q(t),\overset{\star}{q}(t),t)=\mathrm{grad}\,L(q(t),\overset{\star}{q}(t),t)\ .$$

- At each fixed density $q\in\mathcal{E}(\mu)$, and each time t, the partial 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 $w\mapsto\mathrm{grad}_{\mathrm{e}}\,L(q,w,t)$. The standard argument involving the Legendre transform provides the intrinsic form of the Hamilton equations.

- The Hamiltonian is

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

Mechanics X

- If $t \mapsto q(t)$ a solution of Euler-Lagrange equation, the curve $t \mapsto \zeta(t) = (q(t), \eta(t))$ in ${}^*S\,\mathcal{E}\,(\mu)$, where $\eta(t) = \mathrm{grad}_{\mathsf{e}}\, L(q(t), \overset{\star}{q}(t), t)$ is the momentum. The mixture bundle ${}^*S\,\mathcal{E}\,(\mu)$ then plays the role of the cotangent bundle in mechanics.
- The momentum curve satisfies the Hamilton equations,

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

Moreover,

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

[12] We refer to basic facts as presented in V. I. Arnold. *Mathematical methods of classical mechanics*, volume 60 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1989. Translated from the 1974 Russian original by K. Vogtmann and A. Weinstein, Corrected reprint of the second (1989) edition

Example of Mechanics I

- If $L(q,w) = \frac{1}{2}\langle w, w\rangle_q$ is our Lagrangian, then via Legendre transform, we obtain the Hamiltonian $H(q,\eta) = \frac{1}{2}\langle \eta, \eta\rangle_q$.
- The gradients are

$$\begin{aligned}
\operatorname{grad} H(q,\eta) &= -\frac{1}{2}\left(\eta^2 - \mathbb{E}_q\left[\eta^2\right]\right) \\
\operatorname{grad}_{\mathsf{m}} H(q,\eta) &= \eta \\
\operatorname{grad} L(q,w) &= \frac{1}{2}(w^2 - \mathbb{E}_q\left[w^2\right]) \\
\operatorname{grad}_{\mathsf{e}} L(q,w) &= w
\end{aligned}$$

- For $\overset{\star}{q} = w \in {}^*S\,\mathcal{E}\,(\mu)$, the Euler-Lagrange equation is

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

where the covariant derivative is computed in ${}^*S\,\mathcal{E}\,(\mu)$, that is, $\frac{\mathsf{D}}{dt}\overset{\star}{q}(t) = \ddot{q}(t)/q(t)$.

Example of Mechanics II

- In terms of the exponential acceleration $\overset{**}{q}(t) = \ddot{q}(t)/q(t) - \left(\overset{*}{q}(t)^2 - \mathbb{E}_{q(t)}\left[\overset{*}{q}(t)^2\right]\right)$, the Euler-Lagrange equation reads

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

- The Hamilton equations are

$$\begin{cases} \dfrac{\mathrm{D}}{dt}\eta(t) = \dfrac{1}{2}\left(\eta^2 - \mathbb{E}_q\left[\eta^2\right]\right) \\ \quad \overset{*}{q}(t) = \eta(t) \end{cases} ,$$

  with the covariant derivative again computed in ${}^*S\,\mathcal{E}\,(\mu)$.

- The conserved energy is

$$H(q(t), \eta(t)) = \frac{1}{2}\left\langle \overset{*}{q}(t), \overset{*}{q}(t)\right\rangle_{q(t)} = \frac{1}{2}\,\mathbb{E}_1\left[\frac{\dot{q}(t)^2}{q(t)}\right] \ .$$

  which reflects in the conservation of the Fisher information.