4th Carlo Alberto Stochastics Workshop

FUNCTIONAL METHODS IN INFORMATION GEOMETRY

Geometries of the Probability Simplex

Giovanni Pistone

www.giannidiorestino.it

DE CASTRO STATISTICS

Collegio Carlo Alberto

April 18, 2019

PROGRAM

TIME TABLE

18/04 14:30	Giovanni Pistone (de Castro Statistics and Collegio Carlo Alberto) Information geometry of the probability simplex
18/04 16:00	Break
18/04 16:00	Giuseppe Savaré (University of Pavia) Entropic optimal transport and Hellinger-Kantorovich distance
19/04 10:00	Jan Naudts (University of Antwerp) An alternative approach to Quantum Information Geometry
19/04 11:30	Contributed papers and discussion

keywords

Information Geometry, Exponential Manifold, Entropy, Optimal Transport, Hellinger distance, Kantorovich distance, Deformed exponential, Non-parametric, Quantum Information geometry.

Books on Information geometry

- M. K. Murray and J. W. Rice. *Differential Geometry and Statistics*. Number 48 in Monographs on Statistics and Applied Probability. Chapman & Hall, 1993
- R. E. Kass and P. W. Vos. *Geometrical foundations of asymptotic inference*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons, Inc., New York, 1997. A Wiley-Interscience Publication
- S. Amari and H. Nagaoka. *Methods of information geometry*. American Mathematical Society, 2000. Translated from the 1993 Japanese original by Daishi Harada
- S. Amari. *Information geometry and its applications*, volume 194 of *Applied Mathematical Sciences*. Springer, [Tokyo], 2016
- N. Ay, J. Jost, H. V. Lê, and L. 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

Plan

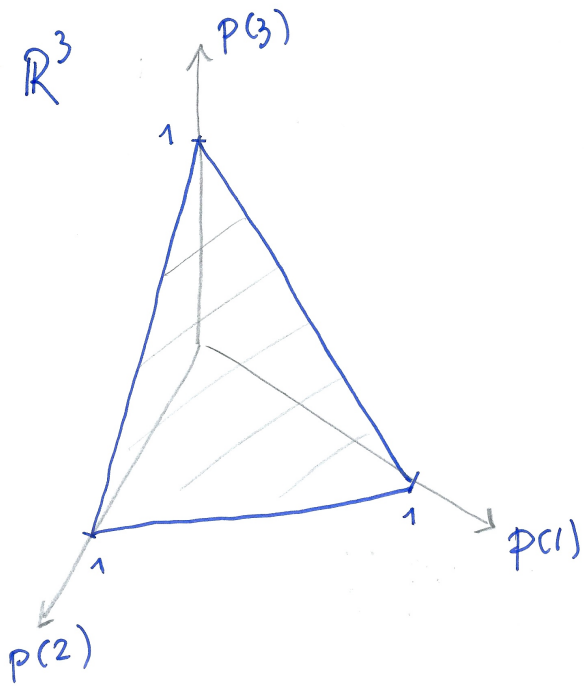
PART I Information Geometry: Statistical bundle, exponential manifold, deformed exponential manifold, second order geometry

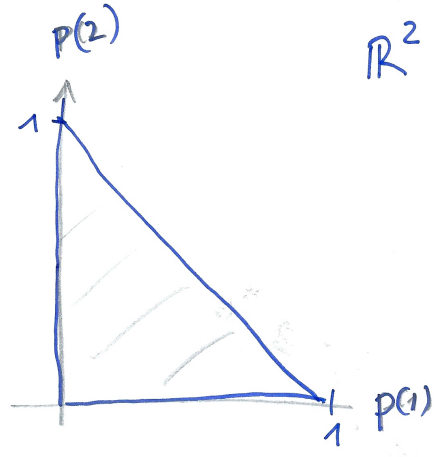
PART II Kantorovich distance: distance defined by a graph

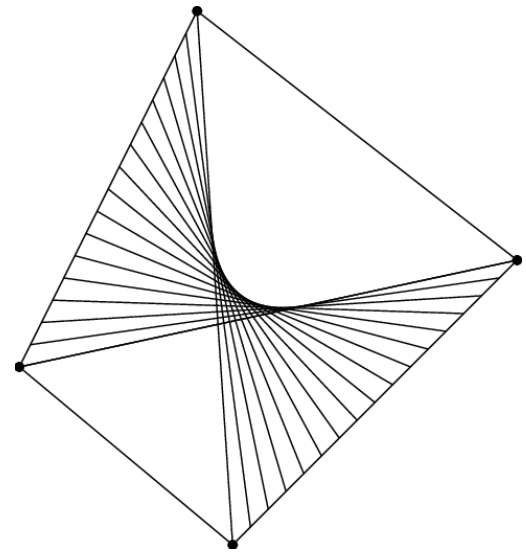
- JW Gibbs
- R Fisher
- CR Rao
- NN Čentsov
- B Efron
- S Amari

PART I: Information Geometry

Ω is a finite set (sample space, configuration space, state space). The set of probability functions is the probability simplex $\Delta(\Omega)$. The set of strictly positive probability functions is the interior $\Delta^\circ(\Omega)$ of the probability simplex.

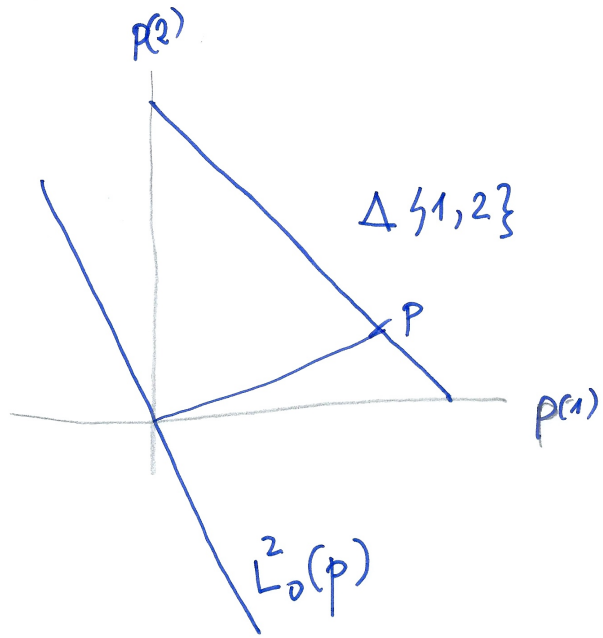






Probabilities and random variables

- $\Delta(\Omega) = \left\{p \in \mathbb{R}^\Omega \middle| \sum_{x\in\Omega} p(x) = 1, p(x) \geq 0\right\}$
- $\Delta^\circ(\Omega) = \left\{p \in \mathbb{R}^\Omega \middle| \sum_{x\in\Omega} p(x) = 1, p(x) > 0\right\}$
- $A_1(\Omega) = \left\{q \in \mathbb{R}^\Omega \middle| \sum_{x\in\Omega} q(x) = 1\right\}$
- $L^2(p) = \left\{U \in \mathbb{R}^\Omega\right\}, \quad \|U\|_p^2 = \sum_{x\in\Omega} U(x)^2 p(x) \quad , p \in \Delta(\Omega)$
- $\mathbb{E}_p[U] = \sum_{x\in\Omega} U(x)p(x), \quad U \in L^2(p), p \in \Delta(\Omega)$
- $L_0^2(p) = \left\{U \in \mathbb{R}^\Omega \middle| \mathbb{E}_p[U] = 0\right\} \qquad L^2(p) = \mathbb{R} \oplus L_0^2(p)$



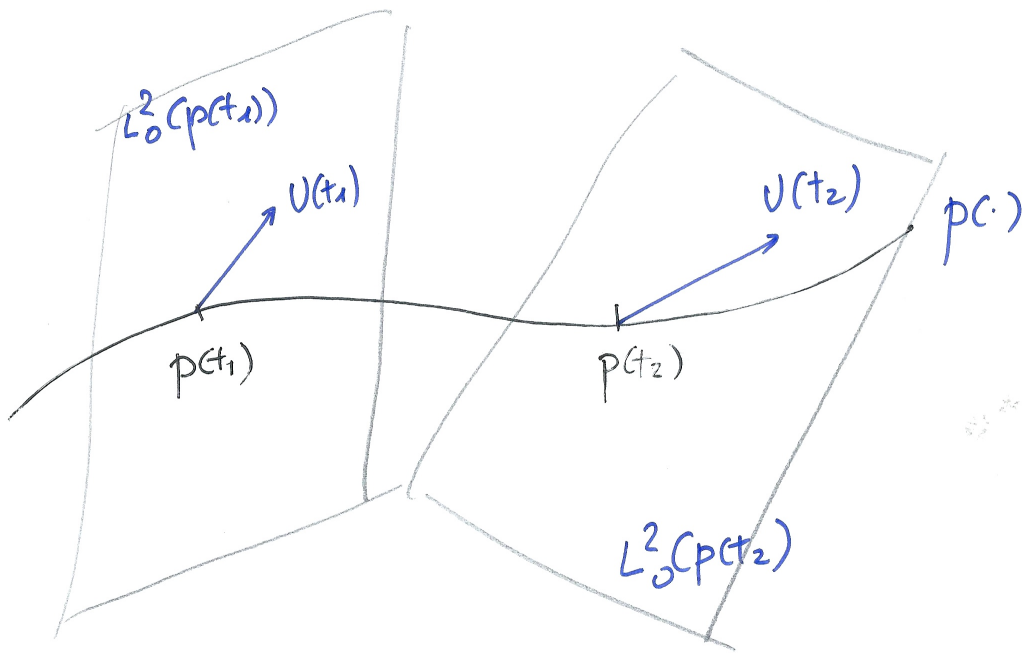
Statistical bundle

- The statistical bundle with base $\Delta(\Omega)$ is

$$S\Delta(\Omega) = \left\{(p, U)\middle| p \in \Delta(\Omega), U \in L_0^2(p)\right\}$$

- The fiber at p is $S_p\Delta(\Omega) = L_0^2(p)$
- The base of the bundle is $\Delta(\Omega)$ but could be smaller i.e. $S\Delta^\circ(\Omega)$
- A curve in the statistical bundle is a mapping

$$I \ni t \mapsto (p(t), U(t)) \quad p(t) \in \Delta(\Omega) \quad U \in S_{p(t)}\Delta(\Omega) = L_0^2(p(t))$$



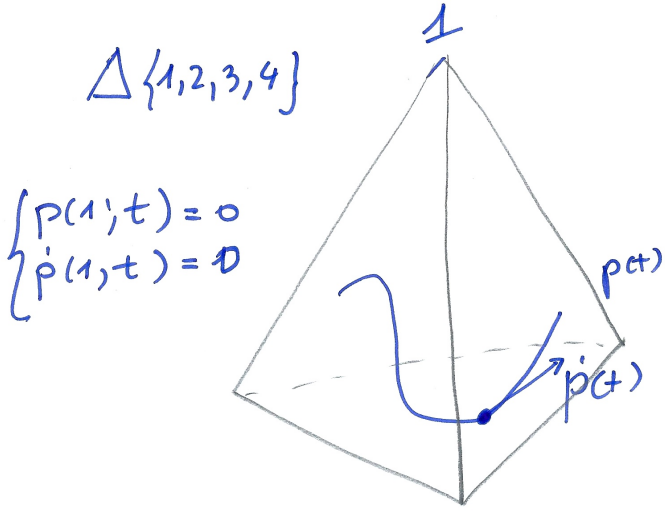
Fisher's score I

Given an $\mathbb{R}^\Omega$-smooth curve $t \mapsto p(t) \in \Delta(\Omega)$, the support of $\dot{p}(t)$ is contained in the support of $p(t)$ for all t. The Fisher's score $Dp(t)$ is such that $\dot{p}(t) = Dp(t) \cdot p(t)$,

$$Dp(x;t) = \begin{cases} 0 & \text{if } p(x;t) = 0, \\ \dfrac{\dot{p}(x;t)}{p(x;t)} = \frac{d}{dt} \log p(x;t) & \text{if } p(x;t) > 0. \end{cases}$$

$t \mapsto (p(t), Sp(t)) \in S\Delta(\Omega)$ is a curve in the statistical bundle.

Proof: For each t and $x \in \Omega$ the condition $p(x;t) = 0$ implies t is a point of minimum, hence $\dot{p}(x;t) = 0$. It follows that $Dp(t)$ is well defined, and $\dot{p}(t) = Dp(t) \cdot p(t)$. Moreover, $\mathbb{E}_{p(t)}[Dp(t)] = \sum_{y \in \Omega} \dot{p}(y;t) = 0$.



Fisher's score II

- From now on, we restrict the discussion to the open simplex

$$S\Delta^{\circ}(\Omega) = \left\{(p, U)\middle| p \in \Delta^{\circ}(\Omega), U \in L_0^2(p)\right\},$$

with fiber $S_p\Delta(\Omega) = L_0^2(p)$. In this case,

$$Dp(x;t) = \frac{\dot{p}(x;t)}{p(x;t)} = \frac{d}{dt}\log p(x;t)$$

- Later on, the statistical bundle will be identified with the tangent bundle of the mapping $p \mapsto \log p$.
- Example: Gibbs probability function

$$p(x,t) = \frac{\exp\left(-\frac{1}{t}V(x)\right)}{\sum_y \exp\left(-\frac{1}{t}V(y)\right)} = \exp\left(-\frac{1}{t}V(x) - \log\sum_y \exp\left(-\frac{1}{t}V(y)\right)\right)$$

$$Dp(t) = \frac{d}{dt}\left(-\frac{1}{t}V(x) - \log\sum_y \exp\left(-\frac{1}{t}V(y)\right)\right) =$$

$$\frac{1}{t^2}\left(V - \frac{\sum_y V(y)\exp\left(-\frac{1}{t}V(y)\right)}{\sum_y \exp\left(-\frac{1}{t}V(y)\right)}\right) = \frac{1}{t^2}\left(V - \mathbb{E}_{p(t)}[V]\right)$$

Fisher's score is a derivation in the statistical bundle

Let X be real random variable on Ω.

$$\frac{d}{dt}\mathbb{E}_{p(t)}[X] = \frac{d}{dt}\sum_y X(y)p(y;t) = \sum_y X(y)\dot{p}(y;t) =$$

$$\sum_y X(y)\frac{\dot{p}(y;t)}{p(y;t)}p(y;t) = \mathbb{E}_{p(t)}[XDp(t)] = \mathbb{E}_{p(t)}\left[(X - \mathbb{E}_{p(t)}[X])Dp(t)\right] ,$$

where we have used the fact that $\mathbb{E}_{p(t)}[Dp(t)] = 0$.

Each fiber $S_p\Delta^\circ(\Omega)$ is an Hilbert space for the inner product $\langle U, V\rangle_p = \mathbb{E}_p[UV]$, and $X - \mathbb{E}_p[X] \in S_p\Delta^\circ(\Omega)$, so

$$\frac{d}{dt}\mathbb{E}_{p(t)}[X] = \left\langle X - \mathbb{E}_{p(t)}[X], Dp(t)\right\rangle_{p(t)}$$

The random variable $Dp(t) \in S_{p(t)}\Delta^\circ(\Omega)$ represents the velocity of variation of the point information $\log p(x;t)$.

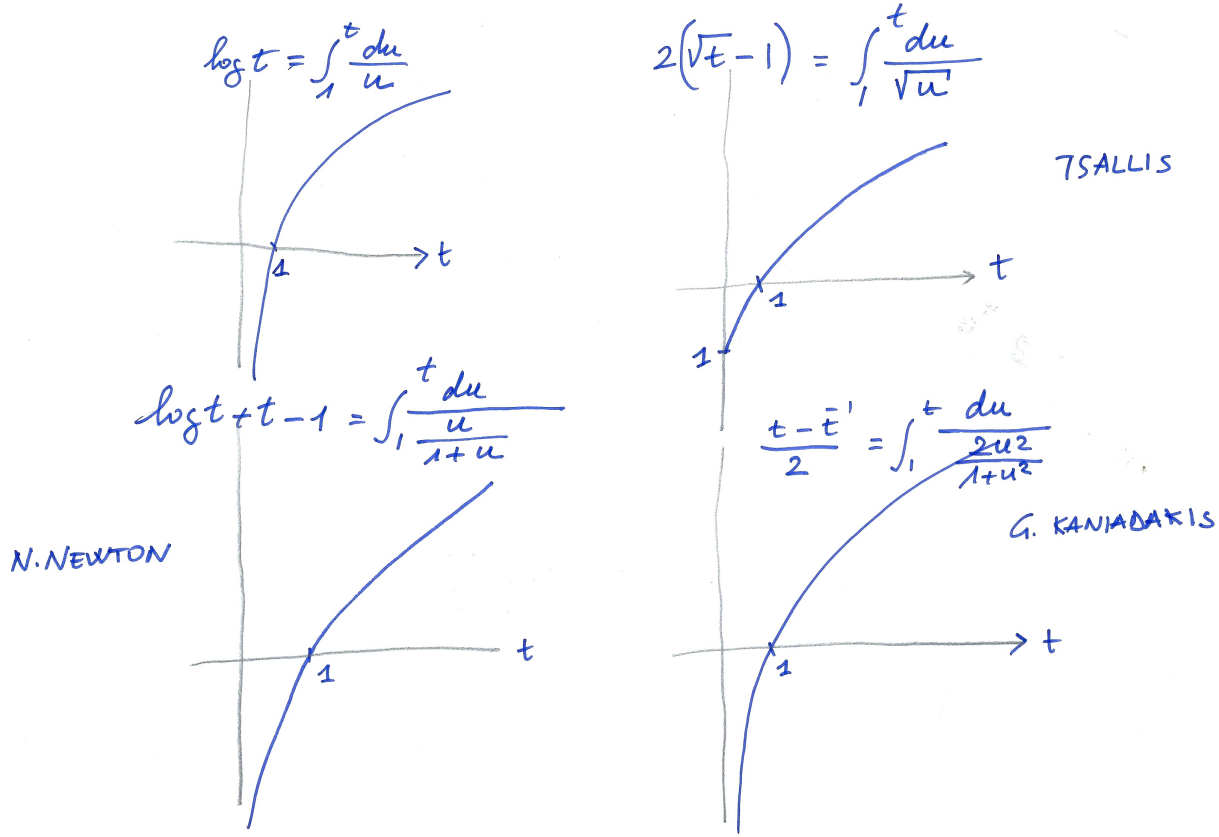
"Deformed" statistical bundle

If $\log_A(t) = \int_1^t \frac{du}{A(u)}$, then

$$\frac{d}{dt}\mathbb{E}_{p(t)}[X] = \sum_y X(y)\dot{p}(y;t) = \sum_y X(y)\frac{\dot{p}(y;t)}{A(p(y;t))}A(p(y;t)) =$$

$$\mathbb{E}_{A(p(t))}[XD_Ap(t)] = \mathbb{E}_{A(p(t))}\left[(X - \mathbb{E}_{A(p(t))}[X])D_Ap(t)\right] =$$

$$\left\langle X - \mathbb{E}_{A(p(t))}[X], D_Ap(t)\right\rangle_{A(p(t))}$$



Exponential expression I

- If $p \in \Delta^\circ(\Omega)$, then $V = \log p$ gives

$$p = \mathrm{e}^V = \mathrm{e}^{(V - \mathbb{E}_p[V]) + \mathbb{E}_p[V]} = \mathrm{e}^{U - \mathcal{H}(p)}$$

  with $U = \log p - \mathbb{E}_p[\log p] \in S_p\Delta^\circ(\Omega)$ and $\mathcal{H}(p) = -\mathbb{E}_p[\log p]$ is the entropy of p.

- For each given p, define for all $q \in \Delta^\circ(\Omega)$ the chart

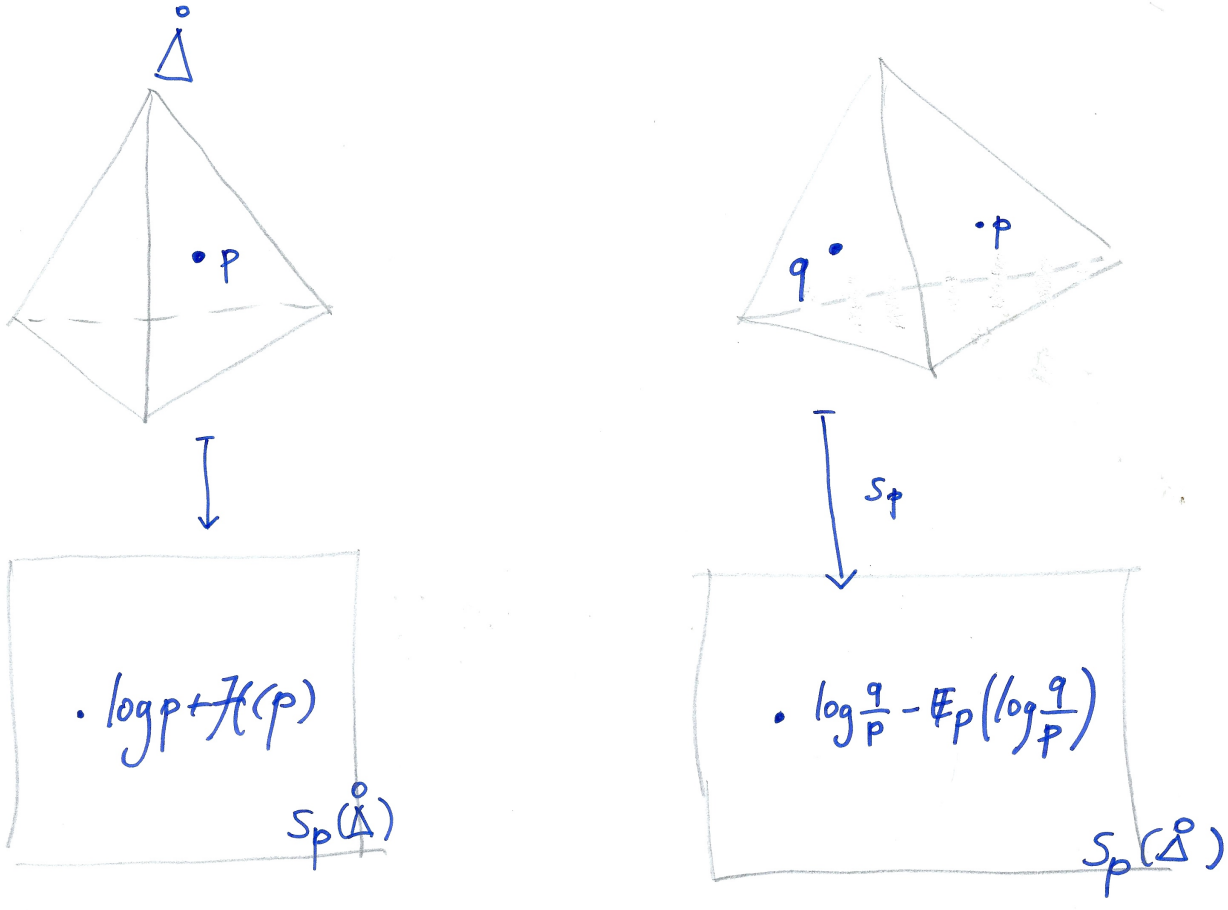
$$s_p(q) = \log \frac{q}{p} - \mathbb{E}_p\left[\log \frac{q}{p}\right] \in S_p\Delta^\circ(\Omega) \ .$$

- Conversely, given any $U \in S_p\Delta^\circ(\Omega)$ the equation $q = \mathrm{e}^{U - K(U)}p$, with $K(U) = \log \mathbb{E}_p\left[\mathrm{e}^U\right]$, defines a probability function such that

$$s_p(q) = s_p(\mathrm{e}^{U-K(U)}p) = \log \frac{\mathrm{e}^{U-K(U)}p}{p} - \mathbb{E}_p\left[\log \frac{\mathrm{e}^{U-K(U)}p}{p}\right] =$$

$$U - K(U) - \mathbb{E}_p[U] + K(U) = U$$

Exponential expression II



Natural gradient I

Definition

Given a function $f\colon \Delta^\circ(\Omega) \to \mathbb{R}$, its natural gradient is a section $\Delta^\circ(\Omega) \ni p \mapsto (p, \operatorname{grad} f(p)) \in S\Delta^\circ(\Omega)$ such that for each smooth curve $I \ni t \mapsto p(t)$ it holds

$$\frac{d}{dt} f(p(t)) = \langle \operatorname{grad} f(p(t)), Dp(t) \rangle_{p(t)}\,, \quad t \in I$$

Computing grad

Let f be a C^1 real function on the open simplex $\Delta^\circ(\Omega)$, $f\colon p \mapsto f(p(x)\colon x \in \Omega)$. For each $p \in \Delta^\circ(\Omega)$, define the random variable $\nabla f(p)$ that takes value $\frac{\partial}{\partial p(x)} f(p)$ at $x \in \Omega$. The natural gradient is

$$\operatorname{grad} f(p) = \nabla f(p) - \mathbb{E}_p\left[\nabla f(p)\right]\ .$$

Natural gradient II

- Proof:

$$\frac{d}{dt} f(p(t)) = \sum_{x \in \Omega} f_x(p(t)) \frac{d}{dt} p(x;t) =$$

$$\sum_{x \in \Omega} f_x(p(t)) \frac{d}{dt} \log p(x;t) p(x;t) = \mathbb{E}_{p(t)} \left[\nabla f(p(t)) D p(x;t) \right] =$$

$$\mathbb{E}_{p(t)} \left[\left(\nabla f(p(t)) - \mathbb{E}_{p(t)} \left[\nabla f(p(t)) \right] \right) D p(x;t) \right] \ .$$

- Natural gradient of the entropy functional:
  $\mathcal{H}(p) = -\sum_{x \in \Omega} p(x) \log p(x)$. The partial derivatives are $f_x(p) = -\log p(x) - 1$, so that $\nabla f(p) = -\log p - 1$ and

$$\operatorname{grad} \mathcal{H}(p) = -\log p - 1 - \mathbb{E}_p \left[-\log p - 1 \right] = -\log p - \mathcal{H}(p) \ .$$

- Here $\operatorname{grad} f(p)$ is the projection of $\nabla f(p)$ onto $S_p \Delta^\circ(\Omega)$ with respect to the inner product $\langle ., . \rangle_p$.

Flows

Given a section $F\colon \Delta^\circ(\Omega)$, $F(p) \in S_p\Delta^\circ(\Omega)$, the trajectories along the section are the solutions of the differential equation $Dp(t) = F(p(t))$.

- The differential equation is equivalent to the system of ordinary differential equations

$$\frac{d}{dt}p(x;t) = p(x;t)F(p(t)) \qquad x \in \Omega$$

- A flow of the section F is the collection of all trajectories along the section.
- The gradient flow is the flow of the section $\pm \operatorname{grad} f$.
- Let $f\colon \Omega \to \mathbb{R}$ a real function to maximize. Relax to $f(p) = \mathbb{E}_p[f]$, $p \in \Delta^\circ(\Omega)$. We have $\operatorname{grad} F(p) = f - \mathbb{E}_p[f]$ and consider the gradient flow $Dp(t) = \operatorname{grad} F(p(t)) = f - \mathbb{E}_{p(t)}[f]$. The solution is the exponential family $p(t) = \mathrm{e}^{tf-\psi(t)}p(0)$. As $t \to \infty$, the solution goes to the probability function uniform on $\{x \in \Omega | f(x) = \max_y f(y)\}$ and zero elsewhere.

Gradient flow of the entropy I

The model example of gradient flow is the gradient flow of the entropy $Dp(t) = -\operatorname{grad} \mathcal{H}(p(t))$. The equation is

$$\frac{d}{dt} \log p(t) = \operatorname{grad} \mathcal{H}(p(t)) = \log p(t) + \mathcal{H}(p(t)) \ ,$$

which is a system of ordinary differential equation,

$$\frac{d}{dt} p(x;t) = p(x;t) \log p(x;t) - \sum_{y \in \Omega} p(y;t) \log p(y;t) \ .$$

Gradient flow of the entropy II

Let us show that the solution starting at $p = p(0)$ is $p(t) \propto p^{\mathrm{e}^t}$. At each $x \in \Omega$,

$$p(x;t) = \frac{p(x)^{\mathrm{e}^t}}{\sum_{y\in\Omega} p(y)^{\mathrm{e}^t}}$$

$$\log p(x;t) = \mathrm{e}^t \log p(x) - \log \sum_{y\in\Omega} p(y)^{\mathrm{e}^t}$$

$$\mathcal{H}\left(p(t)\right) = -\mathbb{E}_{p(t)}\left[\log p(t)\right] = -\mathrm{e}^t \mathbb{E}_{p(t)}\left[\log p\right] + \log \sum_{y\in\Omega} p(y)^{\mathrm{e}^t}$$

$$\log p(x;t) + \mathcal{H}\left(p(t)\right) = \mathrm{e}^t \left(p(x) - \mathbb{E}_{p(t)}\left[p\right]\right)$$

$$Dp(t) = \frac{d}{dt} \log p(x;t) = \mathrm{e}^t p(x) - \frac{\sum_{y\in\Omega} p(y)^{\mathrm{e}^t} \mathrm{e}^t \log p(y)}{\sum_{y\in\Omega} p(y)^{\mathrm{e}^t}} =$$

$$\mathrm{e}^t \left(p(x) - \mathbb{E}_{p(t)}\left[p\right]\right)$$

Exponential families are the orthogonal trajectories to the level sets of the entropy. Here, orthogonality at p is with respect to the inner product of $S_p\Delta^\circ(\Omega)$.

Transports between fibers I

- If $U \in S_p\Delta^\circ(\Omega)$ and $q \in \Delta^\circ(\Omega)$,

$$\mathbb{E}_q\left[U - \mathbb{E}_q\left[U\right]\right] = \mathbb{E}_q\left[U\right] - \mathbb{E}_q\left[U\right] = 0$$

$$\mathbb{E}_q\left[\frac{p}{q}U\right] = \sum_{y\in\Omega} q(y)\frac{p(y)}{q(y)}U(y) = \mathbb{E}_p\left[U\right] = 0$$

- The exponential parallel transport, or e-transport, is the family of linear mappings

$${}^e\mathbb{U}_p^q \colon S_p\Delta^\circ(\Omega) \ni U \mapsto U - \mathbb{E}_q\left[U\right] \in S_q\Delta^\circ(\Omega)\ .$$

- The mixture parallel transport, or m-transport, is the family of linear mappings

$${}^m\mathbb{U}_p^q \colon S_p\Delta^\circ(\Omega) \ni U \mapsto \frac{p}{q}U \in S_q\Delta^\circ(\Omega)\ .$$

Transports between fibers II

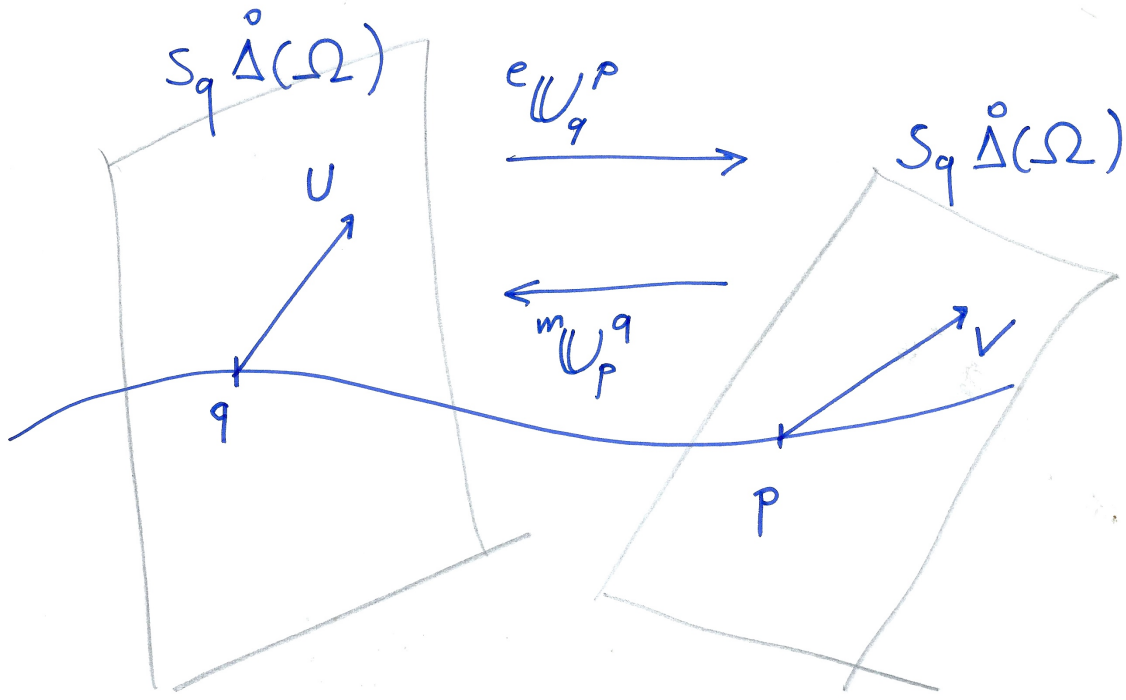
The following properties hold for the e-transport and the m-transport

- semi-group property: ${}^{e}\mathbb{U}_{q}^{r}\,{}^{e}\mathbb{U}_{p}^{q} = {}^{e}\mathbb{U}_{p}^{r}$
- semi-group property: ${}^{m}\mathbb{U}_{q}^{r}\,{}^{m}\mathbb{U}_{p}^{q} = {}^{m}\mathbb{U}_{p}^{r}$
- duality: $\left\langle {}^{e}\mathbb{U}_{p}^{q} U, V \right\rangle_{q} = \left\langle U, {}^{m}\mathbb{U}_{q}^{p} V \right\rangle_{p}$, $U \in S_q\Delta^{\circ}(\Omega)$ and $V \in S_p\Delta^{\circ}(\Omega)$
- transport of the inner product: $\left\langle {}^{e}\mathbb{U}_{p}^{q} U, {}^{m}\mathbb{U}_{p}^{q} V \right\rangle_{q} = \langle U, V \rangle_{p}$, $U, V \in S_p\Delta^{\circ}(\Omega)$

$$\left\langle {}^{e}\mathbb{U}_{p}^{q} U, V \right\rangle_{q} = \mathbb{E}_q\left[(U - \mathbb{E}_q[U])V\right] = \mathbb{E}_q[U]\,V - \mathbb{E}_q[U]\,\mathbb{E}_q[V]$$

$$\mathbb{E}_q[UV] = \mathbb{E}_p\left[\frac{q}{p}UV\right] = \left\langle U, {}^{m}\mathbb{U}_{q}^{p} V \right\rangle_{p}$$

Transports between fibers III



- Given $U \in S_p\Delta^\circ(\Omega)$, each transport defines a section of the statistical bundle:

$$q \mapsto {}^{e}\mathbb{U}_p^q U, \quad q \mapsto {}^{m}\mathbb{U}_p^q U \ ,$$

and we can compute their respective flows.

Transports between fibers IV

- The flow of the vector field $p \mapsto {}^{e}\mathbb{U}_q^p U$, $U \in B_q$ i.e., the solution of

$$Dp(t) = {}^{e}\mathbb{U}_q^{p(t)} U, \quad p(0) = p,$$

is

$$\Delta^\circ(\Omega) \times \mathbb{R} \ni (p, t) \mapsto \mathrm{e}^{t({}^{e}\mathbb{U}_q^p U) - \psi(t)} \cdot p$$

- The flow of the vector field $p \mapsto {}^{m}\mathbb{U}_q^p U$, $U \in B_q$ i.e., the solution of

$$Dp(t) = {}^{m}\mathbb{U}_q^{p(t)} U, \quad p(0) = p$$

is

$$\Delta^\circ(\Omega) \times I \ni (p, t) \mapsto (1 + t\, {}^{m}\mathbb{U}_q^p U)p$$

There are two types of "straight lines", the exponential families and the mixture families

Accelerations I

- Second order differential geometry is usually based on a notion of covariant derivative or affine connection. To each covariant derivative there is associated a parallel transport. Here we procede the other way round because we have a simple definition of parallel transport from which to derive a definition of acceleration.
- A typical application is in optimization where the Hessian provides an estimate of convergence of first order methods and can be used to develop second order methods.
- Let us compute the acceleration of a curve $I \mapsto p(t)$. The velocity is is replaced by Fisher's score
  $t \mapsto (p(t), Dp(t)) = \left(p(t), \frac{d}{dt} \log(p(t))\right) \in S\Delta^\circ(\Omega)$.
- Each random variable $Dp(t)$ has to be checked against an element of $S_{p(t)}\Delta^\circ(\Omega)$, say ${}^m\mathbb{U}_p^{p(t)} V$, $V \in S_p\Delta^\circ(\Omega)$.

Accelerations II

- We can compute an acceleration as

$$\begin{aligned}\frac{d}{dt}\left\langle Dp(t), {}^{m}\mathbb{U}_{p}^{p(t)} V\right\rangle_{p(t)} &= \frac{d}{dt}\left\langle {}^{e}\mathbb{U}_{p(t)}^{p} Dp(t), V\right\rangle_{p} \\ &= \left\langle \frac{d}{dt}\, {}^{e}\mathbb{U}_{p(t)}^{p} Dp(t), V\right\rangle_{p} \\ &= \left\langle {}^{e}\mathbb{U}_{p}^{p(t)} \frac{d}{dt}\, {}^{e}\mathbb{U}_{p(t)}^{p} Dp(t), {}^{m}\mathbb{U}_{p}^{p(t)} V\right\rangle_{p(t)}\end{aligned}$$

- In the computation of ${}^{e}\mathbb{U}_{p}^{p(t)} \frac{d}{dt}\, {}^{e}\mathbb{U}_{p(t)}^{p} Dp(t)$ we first move back $Dp(t)$ to the fixed space $S_p\Delta^{\circ}(\Omega)$, then take the derivative. Finally, move back the derivative to the original space $S_{p(t)}\Delta^{\circ}(\Omega)$.

Accelerations III

The exponential acceleration is

$$
\begin{aligned}
{}^{e}\mathrm{D}^2 p(t) &= {}^{e}\mathbb{U}_{p}^{p(t)} \frac{d}{dt} {}^{e}\mathbb{U}_{p(t)}^{p} Dp(t) \\
&= {}^{e}\mathbb{U}_{p}^{p(t)} \frac{d}{dt} {}^{e}\mathbb{U}_{p(t)}^{p} \frac{d}{dt} \log p(t) \\
&= {}^{e}\mathbb{U}_{p}^{p(t)} \frac{d}{dt} \left(\frac{\dot{p}(t)}{p(t)} - \mathbb{E}_p \left[\frac{\dot{p}(t)}{p(t)} \right] \right) \\
&= {}^{e}\mathbb{U}_{p}^{p(t)} \left(\frac{\ddot{p}(t)p(t) - \dot{p}(t)^2}{p(t)^2} - \mathbb{E}_p \left[\frac{\ddot{p}(t)p(t) - \dot{p}(t)^2}{p(t)^2} \right] \right) \\
&= \frac{\ddot{p}(t)p(t) - \dot{p}(t)^2}{p(t)^2} - \mathbb{E}_{p(t)} \left[\frac{\ddot{p}(t)p(t) - \dot{p}(t)^2}{p(t)^2} \right] \\
&= \frac{\ddot{p}(t)}{p(t)} - (Dp(t))^2 + \mathbb{E}_{p(t)} \left[(Dp(t))^2 \right]
\end{aligned}
$$

Accelerations IV

Exponential families have null exponential acceleration. In fact for $p(t) = \exp(tU - \psi(t))\, p$, we have ${}^{e}\mathbb{U}^{p}_{p(t)} Dp(t) = U - \mathbb{E}_p[U]$, so that $\frac{d}{dt}\, {}^{e}\mathbb{U}^{p}_{p(t)} Dp(t) = 0$. Moreover, note that

$$(Dp(t))^2 = (u - \dot{\psi}(t))^2 \ ,$$

$$\ddot{p}(t) = \frac{d}{dt}[p(t)(u - \dot{\psi}(t))] = p(t)(u - \dot{\psi}(t))^2 - p(t)\ddot{\psi}(t) \ .$$

- A second option is to compute the acceleration as

$$\begin{aligned}
\frac{d}{dt}\left\langle Dp(t), {}^{e}\mathbb{U}^{p(t)}_{p} V \right\rangle_{p(t)} &= \frac{d}{dt}\left\langle {}^{m}\mathbb{U}^{p}_{p(t)} Dp(t), V \right\rangle_{p} \\
&= \left\langle \frac{d}{dt}\, {}^{m}\mathbb{U}^{p}_{p(t)} Dp(t), V \right\rangle_{p} \\
&= \left\langle {}^{m}\mathbb{U}^{p(t)}_{p} \frac{d}{dt}\, {}^{m}\mathbb{U}^{p}_{p(t)} Dp(t), {}^{e}\mathbb{U}^{p(t)}_{p} V \right\rangle_{p(t)}
\end{aligned}$$

Accelerations V

The mixture acceleration is

$$
\begin{aligned}
{}^{m}\mathrm{D}^2 p(t) = & \\
& = {}^{m}\mathbb{U}_{p}^{p(t)} \frac{d}{dt} {}^{m}\mathbb{U}_{p(t)}^{p} Dp(t) \\
& = \frac{p}{p(t)} \frac{d}{dt} \left(\frac{p(t)}{p} \frac{\dot{p}(t)}{p(t)} \right) \\
& = \frac{\ddot{p}(t)}{p(t)}
\end{aligned}
$$

- In follows that mixture models $t \mapsto (1 + tU)p$ have null mixture acceleration.

Taylor formula, Hessian I

- Given $q, p \in \Delta^\circ(\Omega)$, the exponential model

$$p(t) = \frac{p^{1-t}q^t}{\sum_{y\in\Omega} p(y)^{1-t}q(y)^t} = \mathbb{E}_p\left[\left(\frac{q}{p}\right)^t\right]^{-1}\left(\frac{q}{p}\right)^t = \mathrm{e}^{tU-\psi(t)} ,$$

were

$$U = \log\frac{q}{p} - \mathbb{E}_p\left[\log\right]\frac{q}{p} \in S_p\Delta^\circ(\Omega) ,$$

$$\psi(t) = t\mathbb{E}_p\left[\log\right]\frac{q}{p} - \log\mathbb{E}_p\left[\left(\frac{q}{p}\right)^t\right] ,$$

connects $p(0) = p$ and $p(1) = q$.

- We know that $Dp(t) = U - \dot{\psi}(t)$ and ${}^e\mathrm{D}^2 p(t) = 0$.
- Define the mixture Hessian (m-Hessian) of f to be

$${}^m\mathrm{Hess}_U f(p) = {}^m\mathbb{U}_p^{p(t)}\frac{d}{dt}\,{}^m\mathbb{U}_{p(t)}^{p}\,\mathrm{grad}\, f(p(t))\Big|_{t=0} \in S_p\Delta^\circ(\Omega) .$$

Taylor formula, Hessian II

- In such a case the first and the second derivative of $t \mapsto f(p(t))$ reduce to

$$\frac{d}{dt} f(p(t)) = \left\langle \operatorname{grad} f(p(t)), U - \dot{\psi}(t) \right\rangle_{p(t)} ,$$

$$\frac{d^2}{dt^2} f(p(t)) = \left\langle {}^m\operatorname{Hess}_{Dp(t)} f(p(t)), Dp(t) \right\rangle_{p(t)} =$$

$$\left\langle {}^m\operatorname{Hess}_{U-\dot{\psi}(t)} f(p(t)), U - \dot{\psi}(t) \right\rangle_{p(t)} .$$

- We can write the Taylor formula as

$$f(q) - f(p) = \left. \frac{d}{dt} f(p(t)) \right|_{t=0} + \frac{1}{2} \left. \frac{d^2}{dt^2} f(p(t)) \right|_{t=0} + R(p,q)$$

$$= \langle \operatorname{grad} f(p), U \rangle_p + \frac{1}{2} \langle {}^m\operatorname{Hess}_U f(p), U \rangle_p + R(p,U) .$$

- The evaluation of the remainder can be done in the L^2 norm, or, better, in stronger norm.

My case

- The basic structure of Information Geometry is a exponential-mixture affine manifold. The Riemannian structure is intermediate.
- A systematic parametric presentation could possibly fit the needs of Statistics and Machine Learning, but it is not natural in other applications e.g., Statistical Physics, Evolution Equation.
- It is useful, even in studying statistical models on a finite state space to avoid a premature parametrization.
- The affine structure is feasible with a continuous state space. There are many options in implementing that generalization, the exponential representation of positive densities being one. The use of smooth densities is another possible choice.

PART II: Kantorovich-Rubinstein distance

- Kantorovich distance is a special case of a distance defined on $\Delta(\Omega)$. A more general case is called p-Wassestein distance.
- The formalism of Information geometry does not depend on the specific characteristics of the sample space. In contrast to that "feature" (or "bug") of IG, K-distance is based on the assumption that the sample space is endowed with a distance $d\colon \Omega \times \Omega \to \mathbb{R}$ and the K-distance is an extension of the base distance i.e., $d(\delta_x, \delta_y) = d(x, y)$.
- In turn, this is a part of a larger topic called Optimal Transport.
- The use of distances and divergences between probability distributions is quite common in Statistics. However, it should be observed that a distance does not define a proper geometry, unless the distance admits geodesics i.e. curves of minimal length such that the length of a curve between two distribution equals the distance between the extreme distributions.

Books

- L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, second edition, 2008
- C. Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008
- F. Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Birkhäuser, 2015
- G. Peyré and M. Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019

Couplings I

- Let X be a finite set and $\Delta(X)$ be the simplex of probability functions on X. We assume that the sample space X is provided with a distance d.
- Given probability functions $\mu, \nu \in \Delta(X)$, the joint probability function $\gamma \in \Delta(X \times X)$ is a coupling of μ and ν if μ and ν are the two margins of γ, respectively. For instance, $\mu \otimes \nu$ is a coupling of μ and ν.
- The set of all couplings of μ and ν is the subset of $\Delta(X \times X)$ defined by

$$\mathcal{P}(\mu, \nu) = \left\{ \gamma \in \Delta(X \times X) \middle| \sum_{y \in X} \gamma(x, y) = \mu(x), \sum_{x \in X} \gamma(x, y) = \nu(y) \right\} .$$

- This set is a nonempty bounded polyhedron in $\mathbb{R}^{X \times X}$, hence it is a polytope i.e., the convex combination of its vertexes.

Couplings II

The Kantorovich distance (the K-distance) is defined by

$$d(\mu,\nu) = \inf\left\{ \sum_{x,y\in X} d(x,y)\gamma(x,y) \middle| \gamma \in \mathcal{P}(\mu,\nu) \right\} .$$

- The minimum value is actually a distance on the probability simplex.
- The compactness of $\mathcal{P}(\mu,\nu)$ implies that the minimum is reached at some coupling $\widetilde{\gamma}$ namely, $d(\mu,\nu) = \sum_{x,y} d(x,y)\widetilde{\gamma}(x,y)$. More precisely, as we have a minimum of a linear function on a polytope, the set of optimal couplings is a face of the polytope $\mathcal{P}(\mu,\nu)$.
- Couplings, when they are seen as transport plans, are conveniently represented as special transitions, $\gamma(x,y) = \mu(x)P(x,y) = \nu(y)Q(y,x)$, where P and Q are Markov matrices. The Markov matrix P provides a way to map the initial probability function μ to a final probability function ν.

Optimal coupling

- If $\widetilde{\gamma}$ is an optimal coupling i.e. $d(\mu,\nu) = \sum_{x,y\in X} d(x,y)\widetilde{\gamma}(x,y)$, the support of $\widetilde{\gamma}$ is "small". Let us define the graph on X whose edges are defined by $\widetilde{\gamma}(x,y) > 0$.

The graph of the support of an optimal coupling does not contain any cycle.

Given two probability functions μ and ν, the mixture curve $\mu(t) = (1-t)\mu + t\nu$, $0 \le t \le 1$, is a metric geodesic for the d-distance i.e.,

$$d(\mu(t),\mu(s)) = (t-s)d(\mu,\nu)\ , \quad 0 \le s \le t \le 1\ .$$

Moreover, is $\widetilde{\gamma}$ is optimal for $d(\mu,\nu)$, then the coupling defined by

$$\widetilde{\gamma}(x,y;s,t) = (1-t)\mu(x)(x=y) + s\nu(y) + (t-s)\gamma(x,y)$$

is optimal for $d(\mu(0),\mu(t))$.

Duality I

- The problem is clearly a linear programming problem: optimum of a linear function, plus affine and inequality constrains. As such, we expect it to a have a dual linear programming problem.
- A real function u on X is called 1-Lipschitz for the distance d, if $|u(x) - u(y)| \leq d(x, y)$, for all $x, y \in X$. Equivalently, $u(y) \leq d(x, y) + u(x)$. The set of 1-Lipschitz functions will denoted by $\mathrm{Lip}_1(d)$.

Duality II

Let μ and ν be given probability functions on the finite metric space (X, d) and let $\mathcal{P}(\mu, \nu)$ be the set of couplings. Then

$$\min \left\{ \sum_{x,y \in X} d(x,y)\gamma(x,y) \middle| \gamma \in \mathcal{P}(\mu,\nu) \right\} =$$

$$\max \left\{ \sum_{x \in X} \phi(x)\mu(x) - \sum_{y \in X} \psi(y)\nu(y) \middle| \phi \ominus \psi \leq d \right\} =$$

$$\max \left\{ \sum_{x \in X} u(x)(\mu(x) - \nu(x)) \middle| u \in \mathrm{Lip}_1(d) \right\} \ .$$