Chapter 1
Optimization via Information Geometry

Luigi Malagò and Giovanni Pistone

Abstract Information Geometry has been used to inspire efficient algorithms for stochastic optimization, both in the combinatorial and the continuous case. We give an overview of the authors' research program and some specific contributions to the underlying theory.

1.1 Introduction

The present paper is based on the talk given by the second author on May 21, 2013, to the Seventh International Workshop on Simulation in Rimini. Some pieces of research that were announced in that talk have been subsequently published [17, 21, 22]. Here we give a general overview, references to latest published results, and a number of specific topics that have not been published elsewhere.

Let $(\Omega, \mathscr{F}, \mu)$ be a measure space, whose strictly positive probability densities form the algebraically open convex set $\mathscr{P}_>$. An *open statistical model* $(\mathscr{M}, \theta, B)$ is a parametrized subset of $\mathscr{P}_>$, that is, $\mathscr{M} \subset \mathscr{P}_>$ and $\theta \colon \mathscr{M} \to B$, where θ is a one-to-one mapping onto an open subset of a Banach space B. We assume in the following that Ω is endowed with a distance and $\mathscr{F}$ is its Borel σ-algebra.

If $f \colon \Omega \to \mathbb{R}$ is a bounded continuous function, the mapping $\mathscr{M} \ni p \mapsto \mathbb{E}_p[f]$ is a *Stochastic Relaxation* (SR) of f. The strict inequality $\mathbb{E}_p[f] < \sup_{\omega \in \Omega} f(\omega)$ holds for all $p \in \mathscr{M}$, unless f is constant. However, $\sup_{p \in \mathscr{M}} \mathbb{E}_p[f] = \sup_{\omega \in \Omega} f(\omega)$ if there exist a probability measure ν in the weak closure of $\mathscr{M} \cdot \mu$ whose support is contained in the set of maximizing points of f, that is to say

Luigi Malagò
Dipartimento di Informatica, Università degli Studi di Milano, Via Comelico, 39/41, 20135 Milano, Italy, e-mail: malago@di.unimi.it

Giovanni Pistone
de Castro Statistics, Collegio Carlo Alberto, Via Real Collegio 30, 10024 Moncalieri, Italy, e-mail: giovanni.pistone@carloalberto.org

$$\nu\left\{\omega\in\Omega\colon f(\omega)=\sup_{\omega\in\Omega}f(\omega)\right\}=1,\quad\text{or}\quad\int f\,d\nu=\sup_{\omega\in\Omega}f(\omega).$$

Such a ν belongs to the border of $\mathscr{M}\cdot\mu$. For a discussion of the border issue for finite Ω, see [18]. Other relaxation methods have been considered, e.g., [4, 25].

A *SR optimization method* is an algorithm producing a sequence $p_n\in\mathscr{M}$, $n\in\mathbb{N}$, which is expected to converge to the probability measure ν, so that $\lim_{n\to\infty}\mathbb{E}_{p_n}[f]=\sup_{\omega\in\Omega}f(\omega)$. Such algorithms are best studied in the framework of *Information Geometry* (IG), that is, the differential geometry of statistical models. See [2] for a general treatment of IG and [4, 6, 13, 14, 15, 16, 17] for applications to SR. All the quoted literature refers to the case where the model Banach space of the statistical manifold, i.e., the parameter space, is finite dimensional, $B=\mathbb{R}^d$. An infinite dimensional version of IG has been developed, see [22] for a recent presentation together with new results, and references therein for a detailed bibliography. The nonparametric version is unavoidable in applications to evolution equations in Physics [21], and it is useful even when the sample space is finite [19].

1.2 Stochastic relaxation on an exponential family

We recall some basic facts on exponential families, see [8].

1. The exponential family $q_\theta=\exp\left(\sum_{j=1}^d\theta_jT_j-\psi(\theta)\right)\cdot p$, $\mathbb{E}_p[T_j]=0$, is a statistical model $\mathscr{M}=\{q_\theta\}$ with parametrization $q_\theta\mapsto\theta\in\mathbb{R}^d$.
2. $\psi(\theta)=\log\left(\mathbb{E}_p\left[\mathrm{e}^{\theta\cdot T}\right]\right)$, $\theta\in\mathbb{R}^d$, is convex and lower semi-continuous.
3. ψ is analytic on the (non empty) interior $\mathscr{U}$ of its proper domain.
4. $\nabla\psi(\theta)=\mathbb{E}_\theta[T]$, $T=(T_1,\dots,T_d)$.
5. $\mathrm{Hess}\,\psi(\theta)=\mathrm{Var}_\theta(T)$.
6. $\mathscr{U}\ni\theta\mapsto\nabla\psi(\theta)=\eta\in\mathscr{N}$ is one-to-one, analytic, and monotone; $\mathscr{N}$ is the interior of the *marginal polytope*, i.e., the convex set generated by $\{T(\omega)\colon\omega\in\Omega\}$.
7. The gradient of the SR of f is

$$\nabla(\theta\mapsto\mathbb{E}_\theta[f])=(\mathrm{Cov}_\theta(f,T_1),\dots,\mathrm{Cov}_\theta(f,T_d)),$$

   which suggests to take the least squares approximation of f on $\mathrm{Span}(T_1,\dots,T_d)$ as direction of *steepest ascent*, see [16].
8. The representation of the gradient in the scalar product with respect to θ is called *natural gradient*, see [2, 3, 19].

Different methods can be employed to generate a maximizing sequence of densities p_n is a statistical model $\mathscr{M}$. A first example is given by Estimation of Distribution Algorithms (EDAs) [12], a large family of iterative algorithms where the parameters of a density are estimated after sampling and selection, in order to favor samples with larger values for f, see Example 1. Another approach is to evaluate

the gradient of $\mathbb{E}_p[f]$ and follow the direction of the natural gradient over $\mathscr{M}$, as illustrated in Example 2.

Example 1 (EDA from [17]). An *Estimation of Distribution Algorithm* is a SR optimization algorithm based on sampling, selection and estimation, see [12].

Input: N, M ▷ population size, selected population size
Input: $\mathscr{M} = \{p(x;\xi)\}$ ▷ parametric model
  $t \leftarrow 0$
  $\mathscr{P}^t = \text{INITRANDOM}()$ ▷ random initial population
  repeat
    $\mathscr{P}^t_s = \text{SELECTION}(\mathscr{P}^t, M)$ ▷ select M samples
    $\xi^{t+1} = \text{ESTIMATION}(\mathscr{P}^t_s, \mathscr{M})$ ▷ opt. model selection
    $\mathscr{P}^{t+1} = \text{SAMPLER}(\xi^{t+1}, N)$ ▷ N samples
    $t \leftarrow t+1$
  until STOPPINGCRITERIA()

Example 2 (SNGD from [17]). Stochastic Natural Gradient Descent [16] is a SR algorithm that requires the estimation of the gradient.

Input: N, λ ▷ population size, learning rate
Optional: M ▷ selected population size (default $M = N$)
  $t \leftarrow 0$
  $\theta^t \leftarrow (0, \dots, 0)$ ▷ uniform distribution
  $\mathscr{P}^t \leftarrow \text{INITRANDOM}()$ ▷ random initial population
  repeat
    $\mathscr{P}^t_s = \text{SELECTION}(\mathscr{P}^t, M)$ ▷ opt. select M samples
    $\widehat{\nabla}\mathbb{E}[f] \leftarrow \widehat{\text{Cov}}(f, T_i)_{i=1}^d$ ▷ empirical covariances
    $\widehat{I} \leftarrow [\widehat{\text{Cov}}(T_i, T_j)]_{i,j=1}^d$ ▷ $\{T_i(x)\}$ may be learned
    $\theta^{t+1} \leftarrow \theta^t - \lambda \widehat{I}^{-1} \widehat{\nabla}\mathbb{E}[f]$
    $\mathscr{P}^{t+1} \leftarrow \text{GIBBSSAMPLER}(\theta^{t+1}, N)$ ▷ N samples
    $t \leftarrow t+1$
  until STOPPINGCRITERIA()

Finally, other algorithms are based on Bregman divergence. Example 3 illustrates the connection with the exponential family.

Example 3 (Binomial $B(n,p)$). On the finite sample space $\Omega = \{0, \dots, n\}$ with $\mu(x) = \binom{n}{x}$, consider the exponential family $p(x;\theta) = \exp\left(\theta x - n \log\left(1 + e^\theta\right)\right)$. With respect to the expectation parameter $\eta = n e^\theta / (1 + e^\theta) \in]0, n[$ we have $p(x;\eta) = (\eta/n)^x (1 - \eta/n)^{n-x}$, which is the standard presentation of the binomial density.

The standard presentation is defined for $\eta = 0, n$, where the exponential formula is not. In fact, the conjugate $\psi_*(\eta)$ of $\psi(\theta) = n \log\left(1 + e^\theta\right)$ is

$$\psi_*(\eta) = \begin{cases} +\infty & \text{if } \eta < 0 \text{ or } \eta > n, \\ 0 & \text{if } \eta = 0, n, \\ \eta \log\left(\frac{\eta}{n-\eta}\right) - n \log\left(\frac{n}{n-\eta}\right) & \text{if } 0 < \eta < n. \end{cases}$$

We have

$$\begin{aligned}\log p(x;\eta) &= \log\left(\frac{\eta}{n-\eta}\right)(x-\eta)+\psi_*(\eta), \quad \eta\in]0,n[\\ &= \psi_*'(\eta)(x-\eta)+\psi_*(\eta)\le\psi_*(x).\end{aligned}$$

For $x\neq 0,n$, the sign of $\psi_*'(\eta)(x-\eta)$ is eventually negative as $\eta\to 0,n$, hence

$$\lim_{\eta\to 0,n}\log p(x;\eta)=\lim_{\eta\to 0,n}\psi_*'(\eta)(x-\eta)+\psi_*(\eta)=-\infty.$$

If $x=0,n$, the sign of both $\psi_*'(\eta)(0-\eta)$ and $\psi_*'(\eta)(n-\eta)$ is eventually positive as $\eta\to 0$ and $\eta\to n$, respectively. The limit is bounded by $0=\psi_*(x)$, for $x=0,n$.

The argument above is actually general. It has been observed by [5] that the Bregman divergence $D_{\psi_*}(x\|\eta)=\psi_*(x)-\psi_*(\eta)-\psi_*'(\eta)(x-\eta)\ge 0$ provides an interesting form of the density as $p(x;\eta)=\mathrm{e}^{-D_{\psi_*}(x\|\eta)}\mathrm{e}^{\psi_*(x)}\propto\mathrm{e}^{-D_{\psi_*}(x\|\eta)}$.

1.3 Exponential manifold

The set of positive probability densities $\mathscr{P}_>$ is a convex subset of $L^1(\mu)$. Given a $p\in\mathscr{P}_>$, every $q\in\mathscr{P}_>$ can be written as $q=\mathrm{e}^v\cdot p$ where $v=\log\left(\frac{q}{p}\right)$. Below we summarize, together with a few new details, results from [21, 22] and references therein, and the unpublished [24].

Definition 1 (Orlicz Φ-space [11], [20, Chapter II], [23]). Define $\phi(y)=\cosh y-1$. The Orlicz Φ-space $L^\Phi(p)$ is the vector space of all random variables such that $\mathbb{E}_p[\Phi(\alpha u)]$ is finite for some $\alpha>0$. Equivalently, it is the set of all random variables u whose Laplace transform under $p\cdot\mu$, $t\mapsto\hat{u}_p(t)=\mathbb{E}_p[\mathrm{e}^{tu}]$ is finite in a neighborhood of 0. We denote by $M^\Phi(p)\subset L^\Phi(p)$ the vector space of random variables whose Laplace transform is always finite.

Proposition 1 (Properties of the Φ-space).

1. *The set $S_{\le 1}=\left\{u\in L^\Phi(p)\colon \mathbb{E}_p[\Phi(u)]\le 1\right\}$ is the closed unit ball of the complete norm*
$$\|u\|_p=\inf\left\{\rho>0\colon \mathbb{E}_p\left[\Phi\left(\frac{u}{\rho}\right)\right]\le 1\right\}$$
on the Φ-space. For all $a\ge 1$ the continuous injections $L^\infty(\mu)\hookrightarrow L^\Phi(p)\hookrightarrow L^a(p)$ hold.
2. *$\|u\|_p=1$ if either $\mathbb{E}_p[\Phi(u)]=1$ or $\mathbb{E}_p[\Phi(u)]<1$ and $\mathbb{E}_p\left[\Phi\left(\frac{u}{\rho}\right)\right]=\infty$ for $\rho>1$. If $\|u\|_p>1$ then $\|u\|_p\le\mathbb{E}_p[\Phi(u)]$. In particular, $\lim_{\|u\|_p\to\infty}\mathbb{E}_p[\Phi(u)]=\infty$.*
3. *$M^\Phi(p)$ is a closed and separable subspace of $L^\Phi(p)$.*
4. *$L^\Phi(p)=L^\Phi(q)$ as Banach spaces if, and only if, $\int p^{1-\theta}q^\theta\,d\mu$ is finite on a neighborhood of $[0,1]$.*

Proof.

1. See [11], [20, Chapter II], [23].
2. The function $\mathbb{R}_{\geq} \ni \alpha \mapsto \hat{u}(t) = \mathbb{E}_p[\Phi(\alpha u)]$ is increasing, convex, lower semi-continuous. If for some $t_+ > 1$ the value $\hat{u}(t_+)$ is finite, we are in the first case and $\hat{u}(1) = 1$. Otherwise, we have $\hat{u}(1) \leq 1$. If $\|u\|_p > a > 1$, so that $\left\| \frac{a}{\|u\|_p} u \right\|_p > 1$, hence
$$1 < \mathbb{E}_p \left[\Phi \left(\frac{a}{\|u\|_p} u \right) \right] \leq \frac{a}{\|u\|_p} \mathbb{E}_p [\Phi(u)],$$
and $\|u\|_p < a\mathbb{E}_p[\Phi(u)]$, for all $a > 1$.
3. See [11], [20, Chapter II], [23].
4. See [9, 24].

Example 4 (Boolean state space). In the case of a finite state space, the moment generating function is finite everywhere, but its computation can be challenging. We discuss in particular the Boolean case $\Omega = \{+1, -1\}^n$ with counting reference measure μ and uniform density $p(x) = 2^{-n}$, $x \in \Omega$. In this case there is a huge literature from statistical physics, e.g., [10, Ch. VII]. A generic real function on Ω—called pseudo-Boolean [7] in the combinatorial optimization literature—has the form $u(x) = \sum_{\alpha \in L} \hat{u}(\alpha) x^\alpha$, with $L = \{0,1\}^n$, $x^\alpha = \prod_{i=1}^n x_i^{\alpha_i}$, $\hat{u}(\alpha) = 2^{-n} \sum_{x \in \Omega} u(x) x^\alpha$.

As $e^{ax} = \cosh(a) + \sinh(a)x$ if $x^2 = 1$ i.e., $x = \pm 1$, we have

$$\begin{aligned} e^{tu(x)} &= \exp\left(\sum_{\alpha \in \operatorname{Supp} \hat{u}} t\hat{u}(\alpha) x^\alpha \right) = \prod_{\alpha \in \operatorname{Supp} \hat{u}} e^{t\hat{u}(\alpha) x^\alpha} \\ &= \prod_{\alpha \in \operatorname{Supp} \hat{u}} (\cosh(t\hat{u}(\alpha)) + \sinh(t\hat{u}(\alpha)) x^\alpha) \\ &= \sum_{B \subset \operatorname{Supp} \hat{u}} \prod_{\alpha \in B^c} \cosh(t\hat{u}(\alpha)) \prod_{\alpha \in B} \sinh(t\hat{u}(\alpha)) x^{\sum_{\alpha \in B} \alpha}. \end{aligned}$$

The moment generating function of u under the uniform density p is

$$t \mapsto \sum_{B \in \mathscr{B}(\hat{u})} \prod_{\alpha \in B^c} \cosh(t\hat{u}(\alpha)) \prod_{\alpha \in B} \sinh(t\hat{u}(\alpha)),$$

where $\mathscr{B}(\hat{u})$ are those $B \subset \operatorname{Supp} \hat{u}$ such that $\sum_{\alpha \in B} \alpha = 0 \mod 2$. We have

$$\mathbb{E}_p[\Phi](tu) = \sum_{B \in \mathscr{B}_0(\hat{u})} \prod_{\alpha \in B^c} \cosh(t\hat{u}(\alpha)) \prod_{\alpha \in B} \sinh(t\hat{u}(\alpha)) - 1,$$

where $\mathscr{B}_0(\hat{u})$ are those $B \subset \operatorname{Supp} \hat{u}$ such that $\sum_{\alpha \in B} \alpha = 0 \mod 2$ and $\sum_{\alpha \in \operatorname{Supp} \hat{u}} \alpha = 0$.

If S is the $\{1, \ldots, n\} \times \operatorname{Supp} \hat{u}$ matrix with elements α_i we want to solve the system $Sb = 0 \mod 2$ to find all elements of $\mathscr{B}$; we add the equation $\sum b = 0 \mod 2$ to find $\mathscr{B}_0$. The simplest example is $u(x) = \sum_{i=1}^n c_i x_i$,

Example 5 (The sphere is not smooth in general). We look for the moment generating function of the density

$$p(x) \propto (a+x)^{-\frac{3}{2}} \mathrm{e}^{-x}, \quad x > 0,$$

where a is a positive constant. From the incomplete gamma integral

$$\Gamma\left(-\frac{1}{2}, x\right) = \int_x^\infty s^{-\frac{1}{2}-1} \mathrm{e}^{-s}\, ds, \quad x > 0,$$

we have for $\theta, a > 0$,

$$\frac{d}{dx}\Gamma\left(-\frac{1}{2}, \theta(a+x)\right) = -\theta^{-\frac{1}{2}} \mathrm{e}^{-\theta a}(a+x)^{-\frac{3}{2}} \mathrm{e}^{-\theta x}.$$

We have, for $\theta \in \mathbb{R}$,

$$C(\theta, a) = \int_0^\infty (a+x)^{-\frac{3}{2}} \mathrm{e}^{-\theta x}\, dx = \begin{cases} \sqrt{\theta}\mathrm{e}^{\theta a}\Gamma\left(-\frac{1}{2}, \theta a\right) & \text{if } \theta > 0. \\ \frac{1}{2\sqrt{a}} & \text{if } \theta = 0, \\ +\infty & \text{if } \theta < 0. \end{cases}$$

or, $C(\theta, a) = \frac{1}{2}a^{-\frac{1}{2}} - \frac{\sqrt{\pi}\theta}{2}\mathrm{e}^{\theta a} R_{1/2,1}(\theta a)$ if $\theta \leq 1$, $+\infty$ otherwise, where $R_{1/2,1}$ is the survival function of the Gamma distribution with shape $1/2$ and scale 1.

The density p is obtained with $\theta = 1$,

$$p(x) = C(1,a)^{-1}(a+x)^{-\frac{3}{2}} \mathrm{e}^{-x} = \frac{(a+x)^{-\frac{3}{2}} \mathrm{e}^{-x}}{\mathrm{e}^a \Gamma\left(-\frac{1}{2}, a\right)}, \quad x > o,$$

and, for the random variable $u(x) = x$, the function

$$\begin{aligned} \alpha \mapsto \mathbb{E}_p\left[\Phi(\alpha u)\right] &= \frac{1}{\mathrm{e}^a \Gamma\left(-\frac{1}{2}, a\right)} \int_0^\infty (a+x)^{-\frac{3}{2}} \frac{\mathrm{e}^{-(1-\alpha)x} + \mathrm{e}^{-(1+\alpha)x}}{2}\, dx - 1 \\ &= \frac{C(1-\alpha, a) + C(1+\alpha, a)}{2C(1,a)} - 1 \end{aligned}$$

is convex lower semi-continuous on $\alpha \in \mathbb{R}$, finite for $\alpha \in [-1, 1]$, infinite otherwise, hence not steep. Its value at $\alpha = 1$ is

$$\begin{aligned} \mathbb{E}_p\left[\Phi(u)\right] &= \frac{1}{\mathrm{e}^a \Gamma\left(-\frac{1}{2}, a\right)} \int_0^\infty (a+x)^{-\frac{3}{2}} \frac{1 + \mathrm{e}^{-2x}}{2}\, dx - 1 \\ &= \frac{C(0,a) + C(2,a)}{2C(1,a)} - 1 \end{aligned}$$

Example 6 (Normal density). Let $p(x) = (2\pi)^{-1/2}\mathrm{e}^{-(1/2)x^2}$. Consider a generic quadratic polynomial $u(x) = a + bx + \frac{1}{2}cx^2$. We have for $tc \neq 1$

$$t(a+bx+\frac{1}{2}cx^2)-\frac{1}{2}x^2 == -\frac{1}{2(1-tc)^{-1}}\left(x-\frac{tb}{1-tc}\right)^2+\frac{1}{2}\frac{t^2b^2-2ta(1-tc)}{(1-tc)},$$

hence

$$\mathbb{E}_p\left[\mathrm{e}^{tu}\right]=\begin{cases}+\infty & \text{if } tc\leq 1,\\ \sqrt{1-tc}\exp\left(\frac{1}{2}\frac{t^2b^2-2ta(1-tc)}{(1-tc)}\right) & \text{if } tc<1.\end{cases}$$

If, and only if, $-1<c<1$, we have

$$\mathbb{E}_p\left[\Phi(u)\right]=\\ \frac{1}{2}\sqrt{1-c}\exp\left(\frac{1}{2}\frac{b^2-a(1-c)}{(1-c)}\right)+\frac{1}{2}\sqrt{1+c}\exp\left(\frac{1}{2}\frac{b^2-a(1+c)}{(1+c)}\right)-1.$$

1.4 Vector bundles

Vector bundles are constructed as sets of couples (p,v) with $p\in\mathscr{P}_>$ and v is some space of random variables such that $\mathbb{E}_p[v]=0$. The tangent bundle is obtained when the vector space is $L_0^\Phi(p)$. The Hilbert bundle is defined as $H\mathscr{P}_>=\left\{(p,v)\colon p\in\mathscr{P}_>, v\in L_0^2(p)\right\}$. We refer to [21] and [19] were charts and affine connections on the Hilbert bundle are derived from the isometric transport

$$L_0^2(p)\ni u\mapsto\sqrt{\frac{p}{q}}u-\left(1+\mathbb{E}_q\left[\sqrt{\frac{p}{q}}\right]\right)^{-1}\left(1+\sqrt{\frac{p}{q}}\right)\mathbb{E}_q\left[\sqrt{\frac{p}{q}}u\right]\in L_0^2(q).$$

In turn, an isometric trasport $U_p^q\colon L_0^2(p)\to L_0^2(q)$ can be used to compute the derivative of a vector field in the Hilbert bundle, for example the derivative of the gradient of a relaxed function.

The resulting second order structure is instrumental in computing the Hessian of the natural gradient of the SR function. This allows the design a second order approximation method, as it is suggested in [1] for general Riemannian manifolds, and applied to SR in [19]. A second order structure is also used to define the curvature of a statistical manifold and, possibly, to compute its geodesics, see [6] for applications to optimization.

References

1. Absil, P.A., Mahony, R., Sepulchre, R.: Optimization algorithms on matrix manifolds. Princeton University Press, Princeton, NJ (2008). With a foreword by Paul Van Dooren
2. Amari, S., Nagaoka, H.: Methods of information geometry. American Mathematical Society, Providence, RI (2000). Translated from the 1993 Japanese original by Daishi Harada

3. Amari, S.I.: Natural gradient works efficiently in learning. Neural Computation **10**(2), 251–276 (1998)
4. Arnold, L., Auger, A., Hansen, N., Ollivier, Y.: Information-Geometric Optimization Algorithms: A Unifying Picture via Invariance Principles (2011v1; 2013v2). ArXiv:1106.3708
5. Banerjee, A., Merugu, S., Dhillon, I.S., Ghosh, J.: Clustering with bregman divergences. Journal of Machine Learning Research **6**, 1705–1749 (2005)
6. Bensadon, J.: Black-box optimization using geodesics in statistical manifolds. ArXiv:1309.7168
7. Boros, E., Hammer, P.L.: Pseudo-Boolean optimization. Discrete Appl. Math. **123**(1-3), 155–225 (2002). Workshop on Discrete Optimization, DO'99 (Piscataway, NJ)
8. Brown, L.D.: Fundamentals of statistical exponential families with applications in statistical decision theory. No. 9 in IMS Lecture Notes. Monograph Series. Institute of Mathematical Statistics (1986)
9. Cena, A., Pistone, G.: Exponential statistical manifold. Ann. Inst. Statist. Math. **59**(1), 27–56 (2007)
10. Gallavotti, G.: Statistical mechanics: A short treatise. Texts and Monographs in Physics. Springer-Verlag, Berlin (1999)
11. Krasnosel'skii, M.A., Rutickii, Y.B.: Convex Functions and Orlicz Spaces. Noordhoff, Groningen (1961). Russian original: (1958) Fizmatgiz, Moskva
12. Larrañaga, P., Lozano, J.A. (eds.): Estimation of Distribution Algoritms. A New Tool for evolutionary Computation. No. 2 in Genetic Algorithms and Evolutionary Computation. Springer (2001)
13. Malagò, L.: On the geometry of optimization based on the exponential family relaxation. Ph.D. thesis, Politecnico di Milano (2012)
14. Malagò, L., Matteucci, M., Pistone, G.: Stochastic relaxation as a unifying approach in 0/1 programming (2009). NIPS 2009 Workshop on Discrete Optimization in Machine Learning: Submodularity, Sparsity & Polyhedra (DISCML), Dec 11 2009, Whistler, Canada
15. Malagò, L., Matteucci, M., Pistone, G.: Stochastic natural gradient descent by estimation of empirical covariances. In: Proc. of IEEE CEC, pp. 949–956 (2011)
16. Malagò, L., Matteucci, M., Pistone, G.: Towards the geometry of estimation of distribution algorithms based on the exponential family. In: Proceedings of the 11th workshop on Foundations of genetic algorithms, FOGA '11, pp. 230–242. ACM, New York, NY, USA (2011)
17. Malagò, L., Matteucci, M., Pistone, G.: Natural gradient, fitness modelling and model selection: A unifying perspective. In: Proc. of IEEE CEC, pp. 486–493 (2013)
18. Malagò, L., Pistone, G.: A note on the border of an exponential family (2010). ArXiv:1012.0637v1
19. Malagò, L., Pistone, G.: Combinatorial optimization with information geometry: Newton method (2013). In progress
20. Musielak, J.: Orlicz spaces and modular spaces, *Lecture Notes in Mathematics*, vol. 1034. Springer-Verlag, Berlin (1983)
21. Pistone, G.: Examples of application of nonparametric information geometry to statistical physics. Entropy **15**(10), 4042–4065 (2013)
22. Pistone, G.: Nonparametric information geometry. In: F. Nielsen, F. Barbaresco (eds.) Geometric Science of Information, no. 8085 in LNCS, pp. 5–36. Springer-Verlag, Berlin Heidelberg (2013). GSI 2013 Paris, France, August 28-30, 2013 Proceedings
23. Rao, M.M., Ren, Z.D.: Applications of Orlicz spaces, *Monographs and Textbooks in Pure and Applied Mathematics*, vol. 250. Marcel Dekker Inc., New York (2002)
24. Santacroce, M., Siri, P., Trivellato, B.: New results on mixture and exponential models by Orlicz spaces (2013). In progress
25. Wierstra, D., Schaul, T., Peters, J., Schmidhuber, J.: Natural evolution strategies. In: Proc. of IEEE CEC, pp. 3381–3387 (2008)