3RD CARLO ALBERTO STOCHASTICS WORKSHOP

Jan 10–11, 2014

DE CASTRO STATISTICS Collegio Carlo Alberto

Via Real Collegio, 30 10024 Moncalieri (TO), Italy

Optimization via information geometry

Giovanni Pistone

de Castro Statistics Initiative, Collegio Carlo Alberto, Moncalieri, Italy

Jan 11, 2014

Abstract

Combinatorial optimization is the maximization of a real function defined on a finite space $f : \Omega \to \mathbb{R}$. This problem is reduced to a continuous optimization problem by considering the relaxed function $F(p) = \mathrm{E}_p[f]$, where p is a positive density in a statistical model $\mathcal{M}$ on Ω. I will present results from a joint work in progress with Luigi Malagò (Università Statale di Milano). This work expands previous works in Optimization by L. Malagò et al., where the geometry of exponential families is show to provide a suitable setting for model based methods in Combinatorial Optimization under a Black Box assumption on the function f. Some basic tools of Algebraic Design of Experiments are used.

Plan

1. Stochastic Relaxation of an Optimization Problem.
2. Natural Gradient
3. Second Order Geometry

- Luigi Malagò, Matteo Matteucci, and Giovann Pistone. 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, pages 230–242, New York, NY, USA, 2011. ACM

- Luigi Malagò, Matteo Matteucci, and Giovanni Pistone. Natural gradient, fitness modelling and model selection: A unifying perspective.

  In *IEEE Congress on Evolutionary Computation*, pages 486–493. IEEE, 2013

- L. Arnold, A. Auger, N. Hansen, and Y. Ollivier. Information-Geometric Optimization Algorithms: A Unifying Picture via Invariance Principles.

  arXiv:1106.3708, 2011v1; 2013v2

Stochastic relaxation

- $(\Omega, \mathcal{F}, \mu)$ (metric) measure space, $\mathcal{P}_>$ (strictly) positive probability densities.
- An open statistical model $(\mathcal{M}, \theta, B)$ is a parameterized subset of $\mathcal{P}_>$, that is $\mathcal{M} \subset \mathcal{P}_>$ and $s \colon \mathcal{M} \to B$, where s is a one-to-one mapping onto an open subset of a Banach space B.
- If $f \colon \Omega \to \mathbb{R}$ is a bounded continuous function, the mapping $\mathcal{M} \ni p \mapsto \mathrm{E}_p[f]$ is a Stochastic Relaxation SR of f.
- $\mathrm{E}_p[f] < \sup_{\omega \in \Omega} f(\omega)$ for all $p \in \mathcal{M}$ if f is not constant, but $\sup_{p \in \mathcal{M}} \mathrm{E}_p[f] = \sup_{\omega \in \Omega} f(\omega)$ if there exist a probability measure ν in the weak closure of $\mathcal{M} \cdot \mu$ whose support is contained in the set of maximizing points of f, $\{\omega \in \Omega \colon f(\omega) = \sup_{\omega \in \Omega} f(\omega)\}$.
- A SR optimization method is an algorithm to produce a sequence $p_n \in \mathcal{M}$, $n \in \mathbb{N}$, such that $\lim_{n \to \infty} \mathrm{E}_{p_n}[f] = \sup_{\omega \in \Omega} f(\omega)$.
- Such algorithms are best studied in the framework of Information Geometry IG, that is the differential geometry of statistical models.

SR on an exponential family

- The exponential family $q_{\boldsymbol{\theta}} = \exp\left(\sum_{j=1}^{d} \theta_j T_j - \psi(\theta)\right) \cdot p$ is a statistical model $\mathcal{M} = \{q_{\boldsymbol{\theta}}\}$ and parameterization $\sigma\colon q_{\boldsymbol{\theta}} \mapsto \boldsymbol{\theta} \in \mathbb{R}^d$.
- 1. $\psi(\boldsymbol{\theta}) = \log\left(\mathrm{E}_p\left[\mathrm{e}^{\boldsymbol{\theta}\cdot T}\right]\right)$ is convex, lower semi-continuous;
  2. ψ is analytic on the interior $\mathcal{U}$ the proper domain;
  3. $\nabla\psi(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}}[T]$, $\mathrm{Hess}\,\psi(\boldsymbol{\theta}) = \mathrm{Var}_{\boldsymbol{\theta}}(T)$.
  4. $\mathcal{U} \ni \boldsymbol{\theta} \mapsto \nabla\psi(\boldsymbol{\theta}) = \boldsymbol{\eta} \in \mathcal{N}$ is one-to-one, analytic, monotone; $\mathcal{N}$ is the interior of the marginal polytope, i.e. the convex set generated by $\{T(\omega)\colon \omega \in \Omega\}$;
- The SR $\boldsymbol{\theta} \mapsto \mathrm{E}_{\boldsymbol{\theta}}[f]$ is well posed iff the border set $\partial\mathcal{M}$ contains at least one point of $\operatorname{argmax} f$. A sufficient condition is $\delta_{T(\omega)} \in \partial\mathcal{M}$ for all $\omega \in \Omega$
- The gradient of the SR is

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

  which suggests to take the θ-MS approximation of f on $\mathrm{Span}(T_1, \ldots, T_d)$ as direction of steepest ascent.
- This ideas prompt for a systematic development of the geometric picture of statistical models.

Information Geometry

- Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry.* American Mathematical Society, Providence, RI, 2000. Translated from the 1993 Japanese original by Daishi Harada
- 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
- Paolo Gibilisco and Giovanni Pistone. Connections on non-parametric statistical manifolds by Orlicz space geometry. *IDAQP*, 1(2):325–347, 1998
- Giovanni Pistone and Maria Piera Rogantin. The exponential statistical manifold: mean parameters, orthogonality and space transformations. *Bernoulli*, 5(4):721–760, August 1999
- Alberto Cena. *Geometric structures on the non-parametric statistical manifold.* PhD thesis, Dottorato in Matematica, Università di Milano, 2002
- Alberto Cena and Giovanni Pistone. Exponential statistical manifold. *Ann. Inst. Statist. Math.*, 59(1):27–56, 2007
- Luigi Malagò. *On the geometry of optimization based on the exponential family relaxation.* PhD thesis, Politecnico di Milano, 2012
- Giovanni Pistone. Nonparametric information geometry. In Frank Nielsen and Freèdeèric Barbaresco, editors, *Geometric Science of Information*, number 8085 in LNCS, pages 5–36, Berlin Heidelberg, 2013. Springer-Verlag. First International Conference, GSI 2013 Paris, France, August 28-30, 2013 Proceedings

EDA

- Estimation of Distribution is a model-based optimization algorithm.
- **Input:** N, M ▷ population size, selected population size
  Input: $\mathcal{M} = \{p(x;\xi)\}$ ▷ parametric model
  $t \leftarrow 0$
  $\mathcal{P}^t = \textsc{InitRandom}()$ ▷ random initial population
  repeat
    $\mathcal{P}_s^t = \textsc{Selection}(\mathcal{P}^t, M)$ ▷ select M samples
    $\xi^{t+1} = \textsc{Estimation}(\mathcal{P}_s^t, \mathcal{M})$ ▷ opt. model selection
    $\mathcal{P}^{t+1} = \textsc{Sampler}(\xi^{t+1}, N)$ ▷ N samples
    $t \leftarrow t + 1$
  until $\textsc{StoppingCriteria}()$

SNGD

- Stochastic Natural Gradient Descent is a ST algorithm that requires the estimation of a gradient.
- **Input:** N, λ ▷ population size, learning rate
  Optional: M ▷ selected population size (default $M = N$)
  $t \leftarrow 0$
  $\theta^t \leftarrow (0, \ldots, 0)$ ▷ uniform distribution
  $\mathcal{P}^t \leftarrow \textsc{InitRandom}()$ ▷ random initial population
  repeat
    $\mathcal{P}_s^t = \textsc{Selection}(\mathcal{P}^t, M)$ ▷ opt. select M samples
    $\widehat{\nabla}\mathbb{E}[f] \leftarrow \widehat{\mathrm{Cov}}(f, T_i)_{i=1}^m$ ▷ empirical covariances
    $\widehat{I} \leftarrow [\widehat{\mathrm{Cov}}(T_i, T_j)]_{i,j=1}^m$ ▷ $\{T_i(x)\}$ may be learned
    $\theta^{t+1} \leftarrow \theta^t - \lambda \widehat{I}^{-1} \widehat{\nabla}\mathbb{E}[f]$
    $\mathcal{P}^{t+1} \leftarrow \textsc{GibbsSampler}(\theta^{t+1}, N)$ ▷ N samples
    $t \leftarrow t + 1$
  until $\textsc{StoppingCriteria}()$

A toy example I

- $f(x_1, x_2) = a_0 + a_1 x_1 + a_2 x_2 + a_{12} x_1 x_2$, $x_1, x_2 = \pm 1$, $a_0, a_1, a_2, a_{12} \in \mathbb{R}$.
- f is a real random variable on the sample space $\Omega = \{+1, -1\}^2$ with the uniform probability λ.
- $X_1, X_2 \colon \Omega \to \pm 1$ generate an orthonormal basis $1, X_1, X_2, X_1 X_2$ of $L^2(\Omega, \lambda)$ and f is the general form of a real random variable on such a space.
- $\mathcal{P}_>$ is the open simplex of positive densities on (Ω, λ), $\mathcal{E}$ is a statistical model.

The relaxed mapping $F \colon \mathcal{E} \to \mathbb{R}$,

$$F(p) = \mathrm{E}_p[f] = a_0 + a_1 \mathrm{E}_p[X_1] + a_2 \mathrm{E}_p[X_2] + a_{12} \mathrm{E}_p[X_1 X_2]$$

is strictly bounded by the maximum of f, $\mathrm{E}_p[f] < \max_{x \in \Omega}$ if f is not constant.

A toy example II

- We are looking for a sequence p_n, $n \in \mathbb{N}$, such that $\mathrm{E}_{p_n}[f] \to \max_{x \in \Omega} f(x)$ as $n \to \infty$.
- The existence of such a sequence in a nontrivial condition for the model $\mathcal{E}$.

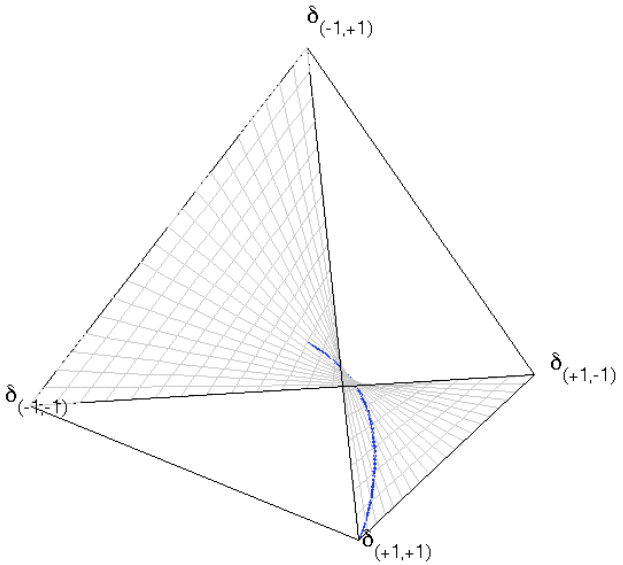
This condition is satisfied by the independence model, when we can write

$$F(\eta^1, \eta^2) = a_0 + a_1 \eta^1 + a_2 \eta^2 + a_{12} \eta^1 \eta^2, \quad \eta^i = \mathrm{E}_p[X_i],$$

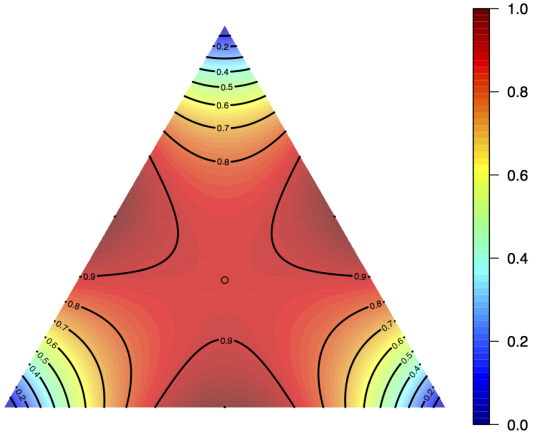
- Not all functions on $\mathcal{P}_>$ are extremised on the vertices. For example the entropy, or the polarization measure

$$p \mapsto \sum_{\omega \in \Omega} p(\omega)^2 (1 - p(\omega))$$

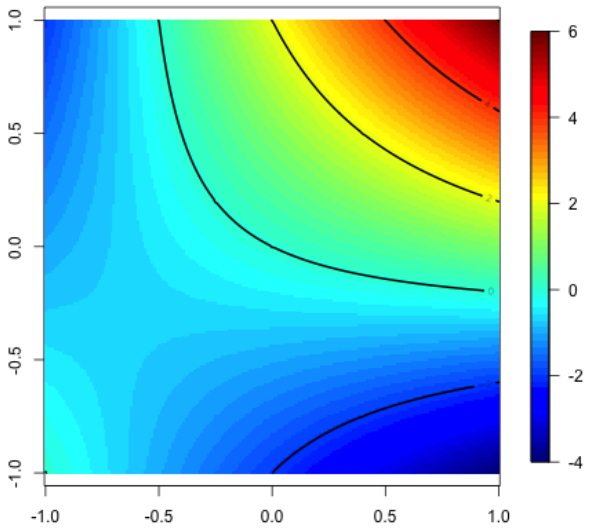
Model, border



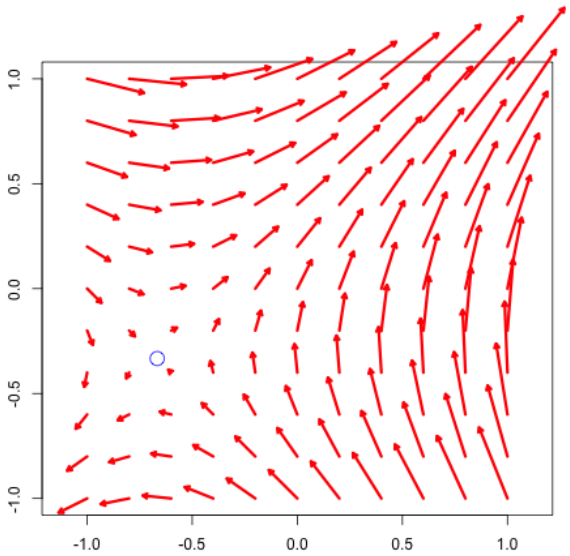
Border: polarization



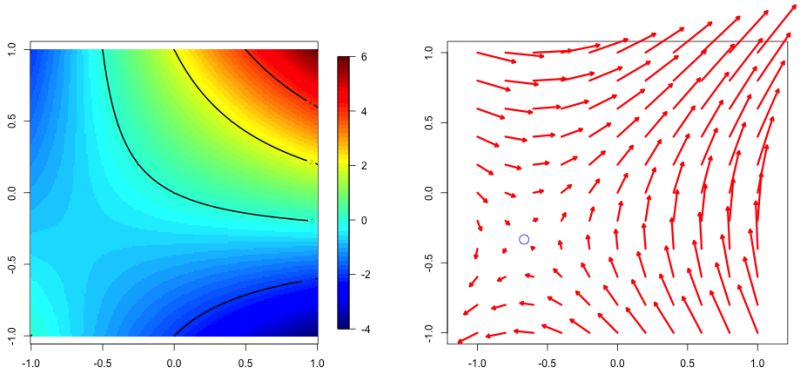
$F(\eta_1, \eta_2) = \eta_1 + 2\eta_2 + 3\eta_1\eta_2$



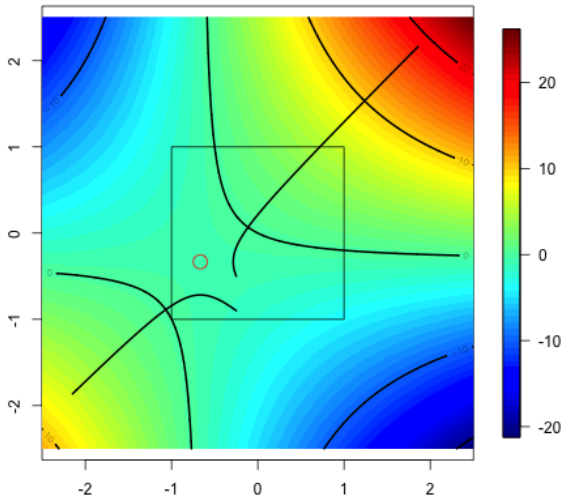
$\nabla F(\eta_1, \eta_2) = (1 + 3\eta_2, 2 + 3\eta_1)$



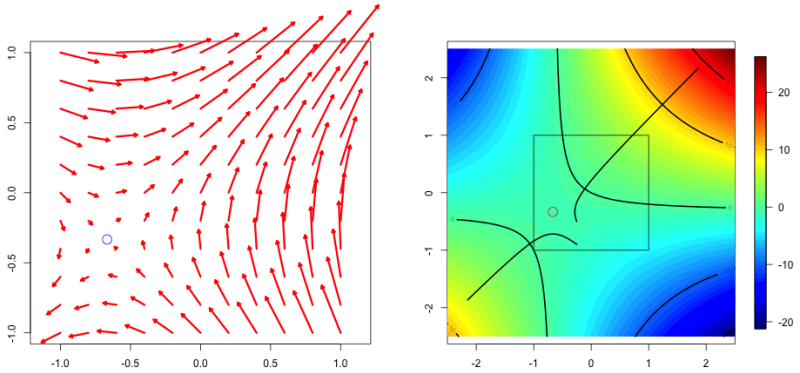
F and ∇F



Integral curves



∇F and integral curves



Issues/Actions

Issues

issue 1 There are critical points in the interior of $[-1,+1]^2$.

issue 2 The gradient points in the right direction outside $[-1,+1]^2$.

Actions

action 1 Order according decreasing effects.

action 2 Use a modified gradient.

$Y_1 = X_1X_2,\ Y_2 = X_2,\ Y_1Y_2 = X_1$

			$\mathbf{1}$	X_1	X_2	X_1X_2	Y_1	Y_2	Y_1Y_2
1	1	1	1	1	1	1	1	1	1
2	−1	1	1	−1	1	−1	−1	1	−1
3	1	−1	1	1	−1	−1	−1	−1	1
4	−1	−1	1	−1	−1	1	1	−1	−1

Same function

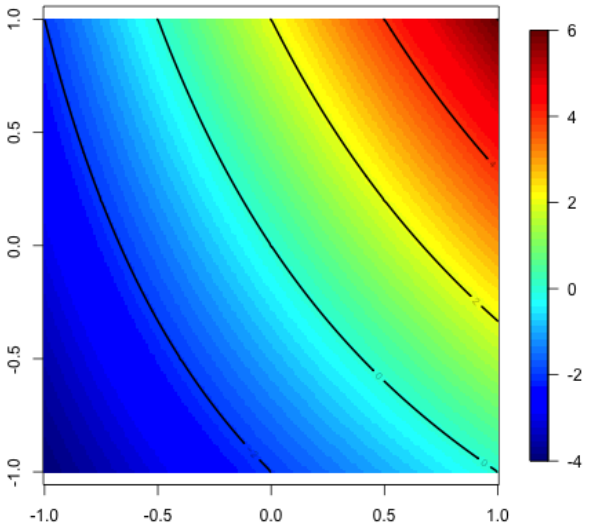
$$X_1 + 2X_2 + 3X_1X_2 = 3Y_1 + 2Y_2 + Y_1Y_2$$

New model

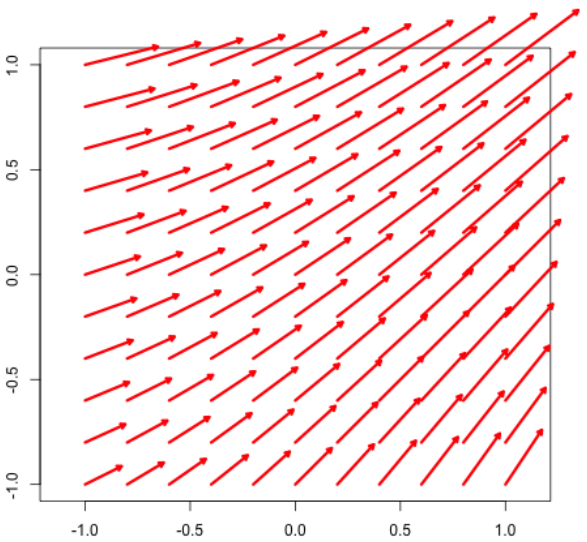
$Y_1 = X_1X_2,\ Y_2 = X_2$ independent

- Emanuele Corsano, Davide Cucci, Luigi Malagò, and Matteo Matteucci. Implicit model selection based on variable transformations in estimation of distribution.
  In *LION*, pages 360–365, 2012
- Davide Cucci, Luigi Malagò, and Matteo Matteucci. Variable transformations in estimation of distribution algorithms.
  In *PPSN (1)*, pages 428–437, 2012

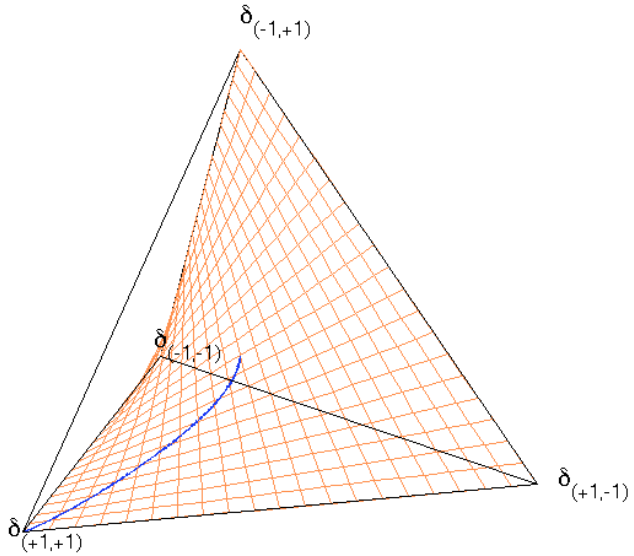
$F(\zeta_1, \zeta_2) = 3\zeta_1 + 2\zeta_2 + \zeta_1\zeta_2$



$\nabla F(\zeta_1, \zeta_2) = (3 + \zeta_2, 2 + \zeta_1)$



Y variables, optimization



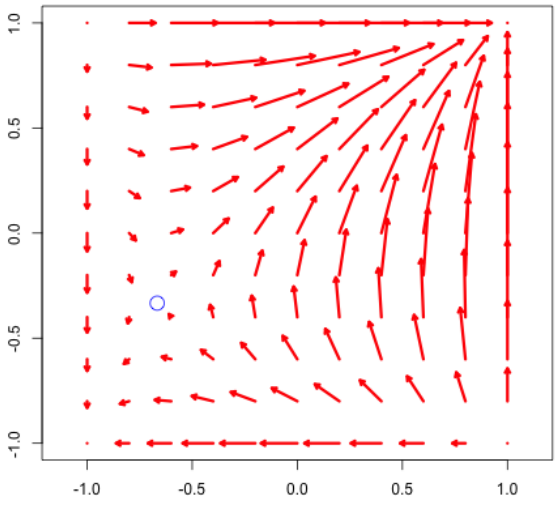
Natural gradient

$$\mathrm{E}_{\boldsymbol{\theta}}\left[(X_1 + 2X_2 + 3X_1X_2)X_1\right] = 1 + 2\eta_1\eta_2 + 3\eta_2$$
$$\mathrm{E}_{\boldsymbol{\theta}}\left[X_1 + 2X_2 + 3 + X_3\right]\mathrm{E}_{\boldsymbol{\theta}}\left[X_1\right] = \eta_1^2 + 2\eta_1\eta_2 + 3\eta_1^2\eta_2$$
$$\mathrm{Cov}_{\boldsymbol{\theta}}\left(X_1 + 2X_2 + 3X_1X_2, X_1\right) = (1 - \eta_1^2)(1 + 3\eta_2)$$

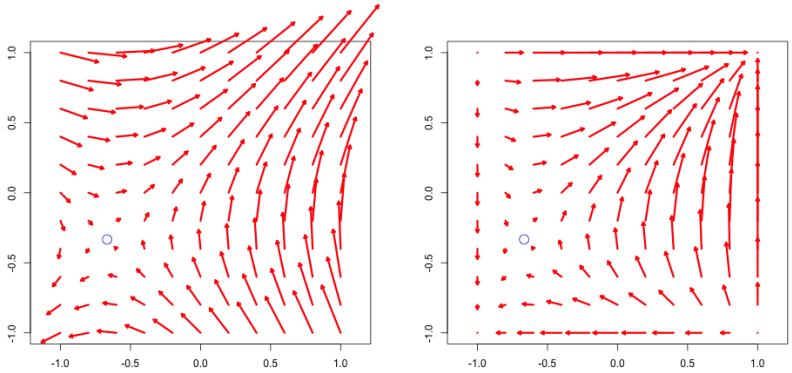
$$\mathrm{E}_{\boldsymbol{\theta}}\left[(X_1 + 2X_2 + 3X_1X_2)X_2\right] = \eta_1\eta_2 + 2 + 3\eta_1$$
$$\mathrm{E}_{\boldsymbol{\theta}}\left[X_1 + 2X_2 + 3 + X_3\right]\mathrm{E}_{\boldsymbol{\theta}}\left[X_2\right] = \eta_1\eta_2 + 2\eta_2^2 + 3\eta_1\eta_2^2$$
$$\mathrm{Cov}_{\boldsymbol{\theta}}\left(X_1 + 2X_2 + 3X_1X_2, X_2\right) = (1 - \eta_2^2)(2 + 3\eta_1)$$

$$\begin{aligned}\widetilde{\nabla} F(\eta_1, \eta_2) &= \left[(1 - \eta_1^2)(1 + 3\eta_2), (1 - \eta_2^2)(2 + 3\eta_1)\right] \\ &= \nabla F(\eta_1, \eta_2)\begin{bmatrix} 1 - \eta_1^2 & 0 \\ 0 & 1 - \eta_2^2 \end{bmatrix}\end{aligned}$$

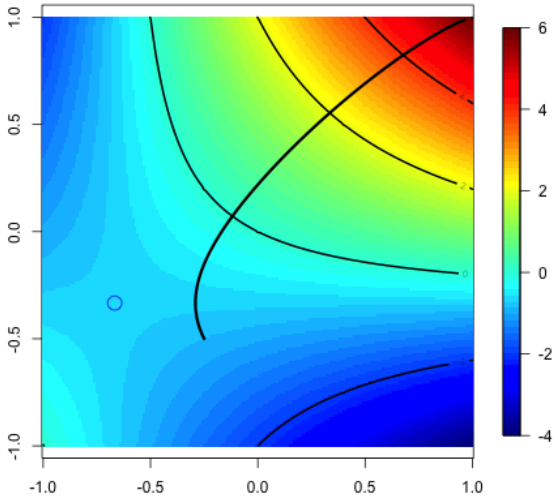
$\widetilde{\nabla} F(\eta_1, \eta_2) = \left((1 - \eta_1^2)(1 + 3\eta_2), (1 - \eta_2^2)(2 + 3\eta_1)\right)$



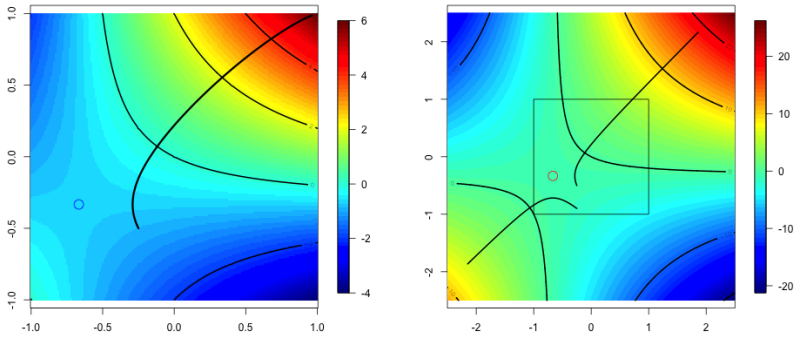
∇F vs $\widetilde{\nabla} F$



Integral curves from $\widetilde{\nabla} F$



$\widetilde{\nabla} F$ flow vs ∇F flow



IGO

- Other relaxation are possible.
- If λ is a positive density, the function

$$\tilde{f} \colon \Omega \ni x \mapsto \mathrm{P}_\lambda \left[f \leq f(x) \right]$$

has the same maximum as f, and

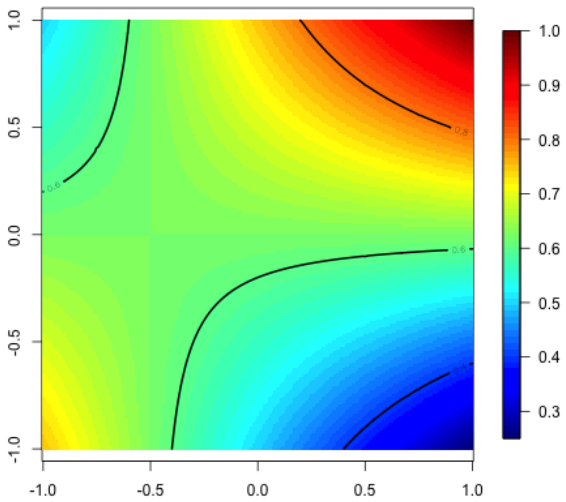
- $$\widehat{F} \colon \mathcal{M} \ni p \mapsto \mathrm{E}_p \left[\hat{f} \right]$$

is a relaxation of $\hat{f}$.

- $\hat{f}$ is invariant under monotone transformations of the values of f and it is a stochastic ordering measure..

- L. Arnold, A. Auger, N. Hansen, and Y. Ollivier. Information-Geometric Optimization Algorithms: A Unifying Picture via Invariance Principles.
arXiv:1106.3708, 2011v1; 2013v2

IGO: $F(\eta_1, \eta_2) = \frac{5}{8} + \frac{1}{8}\eta_2 + \frac{1}{4}\eta_1\eta_2$



Charts I

- On the finite sample space Ω, $\#\Omega = n$, consider a set of random variables $\mathcal{B} = \{X_1, \ldots, X_m\}$ such that $\sum_J \alpha_j X_j$ is constant only if the α_j's are zero, which implies, in turn, the linear independence of $\mathcal{B}$.
- We define $\mathcal{V} = \mathrm{Span}\,(X_1, \ldots, X_m)$ and

$$\mathcal{E}_{\mathcal{V}}(p) = \left\{ q \in \mathcal{P}_> \colon q \propto \mathrm{e}^U p, U \in \mathcal{V} \right\}.$$

- As $\mathcal{E}_{\mathcal{V}}(q) = \mathcal{E}_{\mathcal{V}}(p)$ if, and only if, $q \in \mathcal{E}_{\mathcal{V}}(p) = \mathcal{E}_{\mathcal{V}}$, each choice of a specific reference p is the chart centered at p

$$\sigma_p \colon \exp\left(\sum_j \theta^j {}^e\mathbb{U}^p X_j - \psi_p(\boldsymbol{\theta}) \right) \cdot p \mapsto \boldsymbol{\theta},$$

where ${}^e\mathbb{U}^p$ is the centering at p.

Charts II

- The centering is a parallel transport on the tangent spaces of the manifold:

$${}^e\mathbb{U}^p \colon \mathcal{V} \ni U \mapsto U - \mathrm{E}_p[U] \in {}^e\mathbb{U}^p\mathcal{V}.$$

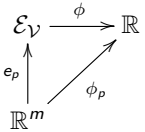
- $\boldsymbol{\theta} = I_{\mathcal{B}}^{-1}(p)\mathrm{E}_p\left[U\, {}^e\mathbb{U}^p X\right]$, where

$$I_{\mathcal{B}}(p) = \left[\mathrm{Cov}_p(X_i, X_j)\right]_{ij} = \mathrm{E}_p\left[XX'\right] - \mathrm{E}_p[X]\,\mathrm{E}_p\left[X'\right] \quad (1)$$

is the Fisher matrix of the basis $\mathcal{B} = \{X_1, \ldots, X_m\}$.

Gradients I

Given a function $\phi \colon \mathcal{E}_{\mathcal{V}} \to \mathbb{R}$ let $\phi_p = \phi \circ e_p$, $e_p = \sigma_p^{-1}$, its representation in the chart centered at p:

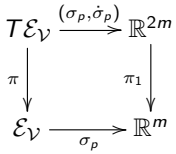


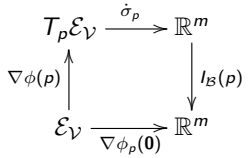
The derivative of $\boldsymbol{\theta} \mapsto \phi_p(\boldsymbol{\theta})$ at $\boldsymbol{\theta} = \mathbf{0}$ along $\boldsymbol{\alpha} \in \mathbb{R}^m$ is

$$\nabla\phi_p(\mathbf{0})\boldsymbol{\alpha} = \nabla\phi_p(\mathbf{0}) I_{\mathcal{B}}^{-1}(p) I_{\mathcal{B}}(p)\boldsymbol{\alpha} = \left(I_{\mathcal{B}}^{-1}(p)\nabla\phi_p(\mathbf{0})'\right)' I_{\mathcal{B}}(p)\boldsymbol{\alpha} = g_p(I_{\mathcal{B}}^{-1}(p)\nabla\phi_p(\mathbf{0})', \boldsymbol{\alpha}).$$

The mapping $\widetilde{\nabla}\phi \colon p \mapsto I_{\mathcal{B}}^{-1}(p)(\nabla\phi_p(\mathbf{0}))' \in \mathbb{R}^m$ is Amari's natural gradient.

Gradients II





$$\dot{\sigma}_p \circ \nabla\phi(p) = I_{\mathcal{B}}^{-1} \nabla\phi_p(\mathbf{0}) = \widetilde{\nabla}\phi(p)$$

Levi-Civita connection

If $D_Y V$ is the vector field on $\mathcal{E}_{\mathcal{V}}$ whose value at p has coordinates

$$\dot{\sigma}_p(D_Y V(p)) = dV_p(\mathbf{0})\boldsymbol{\alpha} + \frac{1}{2} I_{\mathcal{B}}^{-1}(p)\left(dI_{\mathcal{B},p}(\mathbf{0})\boldsymbol{\alpha}\right) V_p(\mathbf{0}), \quad \boldsymbol{\alpha} = \dot{\sigma}_p(Y(p)),$$

then

$$D_Y g(V, W) = g(D_Y V, W) + g(V, D_Y W),$$
$$\dot{\sigma}_p\left(D_W V(p) - D_V W(p)\right) = \dot{\sigma}[V, W](p),$$

i.e. $D_Y V$ is the metric covariant derivative.

- See VIII, §4 of Serge Lang. *Differential and Riemannian manifolds*, volume 160 of *Graduate Texts in Mathematics*.
  Springer-Verlag, New York, third edition, 1995,
- §5.3.2 of P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization algorithms on matrix manifolds*.
  Princeton University Press, Princeton, NJ, 2008.
  With a foreword by Paul Van Dooren.
- L. Malagò, G. Pistone, work in progress on the Newton method.