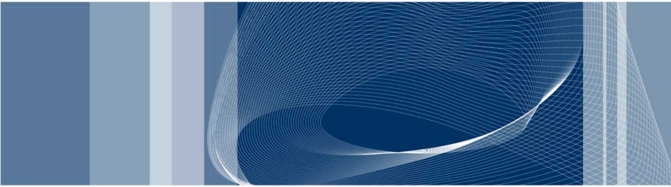
PhD Course on Information Geometry in Learning and Optimization
http://image.diku.dk/MLLab/IG.php

University of Copenhagen

September 22-26, 2014



Applications of Information Geometry to Combinatorial Optimization

Luigi Malagò

Shinshu University, Nagano, Japan – malago@shinshu-u.ac.jp
(09/14 → 09/15 @ INRIA TAO, Paris)

26 Sept 2014

The following slide contains spoilers!

In this lecture we discuss first order optimization over manifolds of distributions

We optimize functionals defined over statistical manifolds by means of (stochastic) natural gradient descent

We focus on discrete statistical models in the exponential family, even if most of the discussion applies also to the continuous case

Motivation

The main motivation is stochastic optimization, i.e., randomized search methods which introduce randomness into the search process

In particular we are interested in model-based optimization, i.e., a broad family of algorithms which employ statistical models to guide the search for the optimum of a function

Model-based algorithms often generate minimizing sequences of probability distributions in a statistical model

Some Notation

- $f(\boldsymbol{x}) : \Omega \to \mathbb{R}$ the objective function
- Ω a finite search search space

- $p(\boldsymbol{x})$ a probability distribution over the sample space Ω
- p_0 the uniform distribution over Ω
- Δ the probability simplex
- $\mathcal{M} = \{p(\boldsymbol{x};\boldsymbol{\xi}) : \boldsymbol{\xi} \in \Xi\} \subset \Delta$ a parametrized statistical model
- $\boldsymbol{\xi}$ a parameter vector for p

Stochastic Relaxation

Consider the minimization problem

$$\text{(P)} \qquad \min_{x \in \Omega} f(\boldsymbol{x})$$

Let $F(p) = \mathbb{E}_p[f]$, we look for the minimum of f by optimizing the Stochastic Relaxation (SR) of f

$$\text{(SR)} \qquad \inf_{p \in \mathcal{M}} F(p)$$

[Remark] We take $\inf$, since in general $\mathcal{M}$ may not be closed

We get candidate solutions for P by sampling optimal solutions of the SR

We introduce a chart $\boldsymbol{\xi}$ over $\mathcal{M} = \{p(\boldsymbol{x};\boldsymbol{\xi}) : \boldsymbol{\xi} \in \Xi\}$, let $F(\boldsymbol{\xi}) = \mathbb{E}_{\boldsymbol{\xi}}[f]$, we have a parametric representation (in coordinates) of the SR

$$\text{(SR)} \qquad \inf_{\boldsymbol{\xi} \in \Xi} F(\boldsymbol{\xi})$$

A Few Remarks

We move the search to the space of probability distributions, from a discrete optimization problem over Ω to a continuous optimization problem over $\mathcal{M}$

In the parametric representation of F, the parameters $\boldsymbol{\xi}$ become the new variables of the SR

Since $\boldsymbol{\xi} \in \Xi$, we may have a constrained formulation of the SR

[Disclaimer] The SR does not provide lower bounds for P, indeed

$$\min_{\boldsymbol{x}\in\Omega} f(\boldsymbol{x}) \leq F(p) \leq \max_{\boldsymbol{x}\in\Omega} f(\boldsymbol{x})$$

Let $\mathcal{M} = \Delta$, $\min_{\boldsymbol{x}\in\Omega} f(\boldsymbol{x}) = \min_{p\in\Delta} F(p)$

More in general, for $\mathcal{M} \subset \Delta$, $\min_{\boldsymbol{x}\in\Omega} f(\boldsymbol{x}) \leq \inf_{p\in\mathcal{M}} F(p)$

Closure of $\mathcal{M}$

We denote with $\overline{\mathcal{M}}$ the topological closure of $\mathcal{M}$, i.e., $\mathcal{M}$ together with the limits of (weakly convergent) sequences of distributions in $\mathcal{M}$

Moreover, we suppose $\overline{\mathcal{M}}$ is compact so that by the extreme value theorem $F(p)$ attains its infimum over $\overline{\mathcal{M}}$

Equivalence of P and SR

Let us denote the optimal solutions

- $\boldsymbol{x}^* \in \Omega^* = \arg\min_{x \in \Omega} f(\boldsymbol{x})$
- $p^* \in P^* = \arg\min_{p \in \overline{\mathcal{M}}} F(p)$

The SR is equivalent to P if $p^*(\boldsymbol{x}^*) = 1$, i.e., the probability of sampling optimal solutions of P from optimal solutions of SR is equal to 1

In other words, there must exists a sequence $\{p_t\}$ in $\mathcal{M}$ such that

$$\lim_{t \to \infty} p_t(\boldsymbol{x}^*) = 1$$

A sufficient condition for the equivalence of SR and P is that all Dirac distribution δ_x are included in $\overline{\mathcal{M}}$

"Why are we doing this?"

Let us clarify some points

- Let k be the cardinality of Ω, to parametrize Δ we need $k-1$ parameters
- Minimizing $F(p)$ with $p \in \Delta$ is as computationally expensive as an exhaustive search

And so, why going from Ω to $\mathcal{M}$?

- In practice we restrict the search to some lower-dimensional model
- The equivalence of P and SR can be easily guaranteed by low dimensional models
- We can develop efficient blackbox optimization algorithms
- It is possible to identify (e.g., learn) submodels for which "nice" properties hold (e.g., gradient descent converges to global optimum)

Which Model to Choose?

[Remark] The landscape of $F(p)$ depends on the choice of $\mathcal{M}$, i.e., the number of local minima of $F(p)$ depends on $\mathcal{M}$

Often it is convenient to work with graphical models

- There exists an interpretation of the model in terms of conditional independence assumptions among variables
- We can define hierarchical structures and control model dimension
- There is a strong literature about estimation, sampling, learning, ...

In the following we work (mainly) with exponential families

The Gibbs or Boltzmann Distribution

- The Gibbs distribution is the one-dimensional exponential family

$$p(\boldsymbol{x};\beta) = \frac{e^{-\beta f(\boldsymbol{x})}}{\mathbb{E}_q[e^{-\beta f(\boldsymbol{x})}]} q(\boldsymbol{x}), \quad \beta > 0$$

- $f(\boldsymbol{x})$ the energy function
- β the inverse temperature
- $q(\boldsymbol{x})$ the reference measure

- The set of distributions is not weakly closed
- Suppose $f(\boldsymbol{x}) \geq 0$ and $f(\boldsymbol{x}) = 0$ for some $\boldsymbol{x} \in \Omega$, but not everywhere

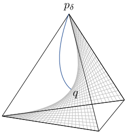
$$\lim_{\beta \to 0} e^{-\beta f(\boldsymbol{x})} = q(\boldsymbol{x}) \qquad \lim_{\beta \to 0} \mathbb{E}_q[e^{-\beta f(\boldsymbol{x})}] = 1$$

$$\lim_{\beta \to \infty} e^{-\beta f(\boldsymbol{x})} = \begin{cases} q(\boldsymbol{x}) & \text{if } f(\boldsymbol{x}) = 0, \\ 0 & \text{otherwise} \end{cases} \qquad \lim_{\beta \to \infty} \mathbb{E}_q[e^{-\beta f(\boldsymbol{x})}] = \sum_{\Omega^*} q(\boldsymbol{x})$$

Closure of the Gibbs Distribution

$$\lim_{\beta\to 0} p(\boldsymbol{x};\beta) = q(\boldsymbol{x})$$

$$\lim_{\beta\to\infty} p(\boldsymbol{x};\beta) = \frac{q(\boldsymbol{x})}{\sum_{\Omega^*} q(\boldsymbol{x})}$$



The Gibbs distribution is in principle an optimal choice for the SR, indeed

- The limit for $\beta \to \infty$ has support over the minima of f, and in particular for $q(x) = p_0$, it is the uniform distribution δ_{Ω^*}
- Since
  $$\nabla F(\beta) = -\operatorname{Var}_\beta(f) < 0$$
  $F(\beta)$ decreases monotonically as $\beta \to \infty$
- The Gibbs distribution satisfies the principle of maximum entropy

However, evaluating the partition function is computationally infeasible

Pseudo-Boolean Optimization

In the following we restrict to $\Omega = \{+1,-1\}^n$, and we use the harmonic encoding $\{+1,-1\}$ for a binary variable

$$-1^0 = +1 \qquad\qquad -1^1 = -1$$

A pseudo-Boolean function f is a real-valued mapping

$$f(\boldsymbol{x}) : \Omega = \{+1,-1\}^n \to \mathbb{R}$$

Any f can be expanded uniquely as a square free polynomial

$$f(x) = \sum_{\boldsymbol{\alpha} \in L} c_{\boldsymbol{\alpha}} \boldsymbol{x}^{\boldsymbol{\alpha}},$$

by employing a multi-index notation. Let $L = \{0,1\}^n$, then $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_n) \in L$ uniquely identifies the monomial $\boldsymbol{x}^{\boldsymbol{\alpha}}$ by

$$\boldsymbol{\alpha} \mapsto \prod_{i=1}^{n} x_i^{\alpha_i}$$

The degree of the monomials represents the order of the interactions in f

Monomial Representation of PS Functions

Let $A^n = \underbrace{A^1 \otimes \ldots \otimes A^1}_{\text{n times}}$, where $\otimes$ denotes the Kronecker product and

$$A^1 = \begin{array}{c} \\ + \\ - \end{array} \begin{array}{cc} 0 & 1 \\ \hline \end{array} \hspace{-3em}$$

$$A^1 = \begin{array}{c|cc} & 0 & 1 \\ \hline + & +1 & +1 \\ - & +1 & -1 \end{array}$$

let $\boldsymbol{a} = (f(\boldsymbol{x}))_{\boldsymbol{x} \in \Omega}$, we have $A^n \boldsymbol{c} = \boldsymbol{a}$, with $\boldsymbol{c} = (c_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha} \in L}$ and $\boldsymbol{c} = 2^{-n} A^n \boldsymbol{a}$

[Example] In case of two variables $\boldsymbol{x} = (x_1, x_2)$, we have

$$f(\boldsymbol{x}) = c_0 + c_1 x_1 + c_2 x_2 + c_{12} x_1 x_2$$

x_1	x_2	$f(\boldsymbol{x})$
$+1$	$+1$	a_{++}
$+1$	-1	a_{+-}
-1	$+1$	a_{-+}
-1	-1	a_{--}

$$\begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_{12} \end{bmatrix} = \frac{1}{4} \times \begin{array}{c|cccc} & 00 & 10 & 01 & 11 \\ \hline ++ & +1 & +1 & +1 & +1 \\ +- & +1 & +1 & -1 & -1 \\ -+ & +1 & -1 & +1 & -1 \\ -- & +1 & -1 & -1 & +1 \end{array} \begin{bmatrix} a_{++} \\ a_{+-} \\ a_{-+} \\ a_{--} \end{bmatrix}$$

The Independence Model

Let $\mathcal{M}_1$ be the independence model for $\boldsymbol{X} = (X_1, \dots, X_n)$

$$\mathcal{M}_1 = \{p : p(\boldsymbol{x}) = \prod_{i=1}^{n} p_i(x_i)\}$$

with marginal probabilities $p_i(x_i) = \mathbb{P}(X_i = x_i)$

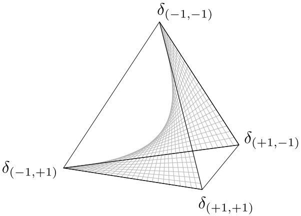
We parametrize $\mathcal{M}_1$ using ±1 Bernoulli distributions for X_i

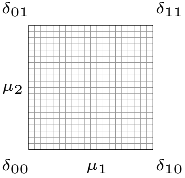
$$\begin{aligned} p(\boldsymbol{x}; \boldsymbol{\mu}) &= \prod_{i=1}^{n} \mu_i^{(1+x)/2} (1 - \mu_i)^{(1-x)/2} \\ &= \prod_{i=1}^{n} (2\mu_i x_i - x_i + 1)/2 \end{aligned}$$

with $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n) \in [0,1]^n$ and

$$\begin{aligned} \mu_i &= \mathbb{P}(X_i = +1) \\ 1 - \mu_i &= \mathbb{P}(X_i = -1) \end{aligned}$$

Marginal Parameters for the Independence Model





$\mathcal{M}_1$ is a n-dimensional manifold embedded in the $2^n - 1$ dimensional probability simplex Δ

Stochastic Relaxation: Who?

The SR can be solved in different ways

- Sampling, Selection and Estimation paradigm (EDAs, Larrañaga and Lozano, 2002, CE method, Rubinstein, 1997)
- Fitness Modeling (DEUM, Shakya et al., 2005)
- Covariance Matrix Adaptation (CMA-ES, Hansen et al., 2001)
- GAs (Holland, 1975), ACO (Dorigo, 1992), ESs (Rechenberg, 1960)
- Boltzmann distribution and Gibbs sampler (Geman and Geman, 1984)
- Simulated Annealing and Boltzmann Machines (Aarts and Korst, 1989)
- Method of the Moments (SDP, Meziat et al., 2001)
- LP relaxation in pseudo-Boolean optimization (Boros and Hammer, 2001)
- ...and many others

Gradient Descent

In the following we will focus on gradient descent techniques of the form

$$\boldsymbol{\xi}^{t+1} = \boldsymbol{\xi}^{t} - \lambda \nabla F(\boldsymbol{\xi}), \qquad \lambda > 0$$

In particular we refer to gradient-based algorithms such as

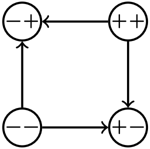
- Covariance Matrix Adaptation Evolutionary Strategy (CMA-ES), Hansen et al., 2001
- Natural Evolutionary Strategies (NES), Wierstra et al., 2008
- Stochastic Natural Gradient Descent (SNGD), M. et al., 2011
- Information Geometry Optimization (IGO), Arnold et al., 2011

A Toy Example

Let $n = 2$, $\Omega = \{-1, +1\}^2$, we want to minimize

$$f(\boldsymbol{x}) = x_1 + 2x_2 + 3x_1x_2$$

x_1	x_2	$f(\boldsymbol{x})$
$+1$	$+1$	6
$+1$	-1	-4
-1	$+1$	-2
-1	-1	0



The gradient flow is the solution of the following differential equation

$$\dot{\boldsymbol{\xi}} = \nabla F(\boldsymbol{\xi}),$$

where we set the step size $\lambda \to \epsilon$

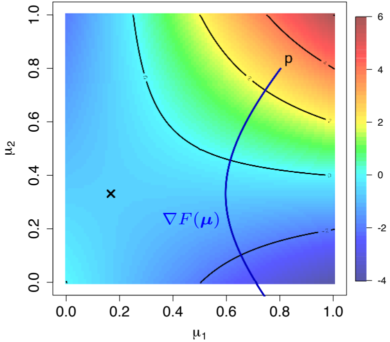
We are interested in studying gradient flows for different parameterization and different statistical models

Gradient Flows on the Independence Model

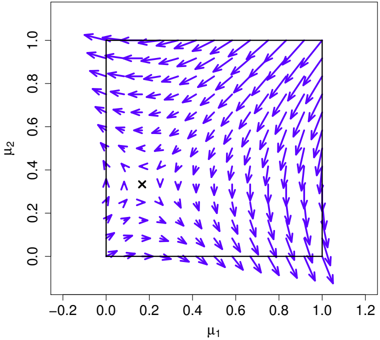
$$F(\boldsymbol{\mu}) = \sum_{\boldsymbol{x}\in\Omega} f(\boldsymbol{x})p_1(x_1)p_2(x_2) = -4\mu_1 - 2\mu_2 + 12\mu_1\mu_2$$

$$\nabla F(\boldsymbol{\mu}) = (-4 + 12\mu_2, -2 + 12\mu_1)^{\mathrm{T}}$$

Gradient flow in $\boldsymbol{\mu}$



Gradient vector in $\boldsymbol{\mu}$, $\lambda = 0.025$



$\nabla F(\boldsymbol{\eta})$ does not convergence to (local) optima, a projection over the hyperplanes given by the constraints is required

Natural Parameters for the Independence Model

If we restrict to positive probabilities $p > 0$, we can represent the interior of the independence model as the exponential family

$$p(\boldsymbol{x};\boldsymbol{\theta}) = \exp\left\{\sum_{i=1}^{n} \theta_i x_i - \psi(\boldsymbol{\theta})\right\}$$

where $\psi(\boldsymbol{\theta}) = \ln Z(\boldsymbol{\theta})$ is the log partition function

The natural parameters of the independence model $\mathcal{M}_1$ represented by an exponential family are $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n) \in \mathbb{R}^n$, with

$$p_i(x_i) = \frac{e^{\theta_i x_i}}{e^{\theta_i} + e^{-\theta_i}}$$

The mapping between marginal probabilities and natural parameters is one-to-one for $p > 0$

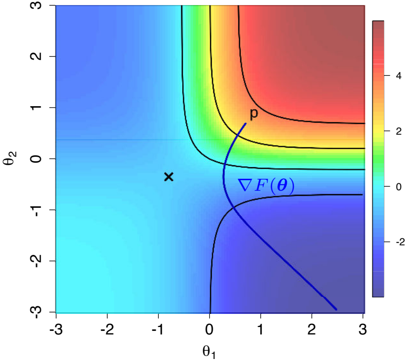
$$\theta_i = \left(\ln(\mu_i) - \ln(1 - \mu_i)\right)/2 \qquad\qquad \mu_i = \frac{e^{\theta_i}}{e^{\theta_i} + e^{-\theta_i}}$$

Gradient Flows on the Independence Model

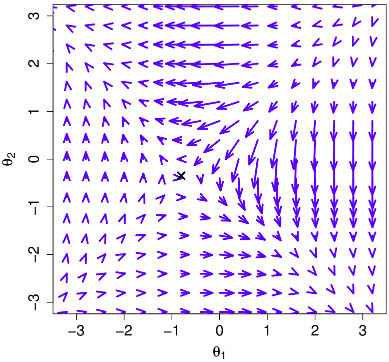
$$F(\boldsymbol{\theta}) = (-4e^{\theta_1-\theta_2} - 2e^{-\theta_1+\theta_2} + 6e^{\theta_1+\theta_2})/Z(\boldsymbol{\theta})$$

$$\nabla F(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}[f(\boldsymbol{X} - \mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{X}])] = \mathrm{Cov}_{\boldsymbol{\theta}}(f, \boldsymbol{X})$$

Gradient flow in $\boldsymbol{\theta}$



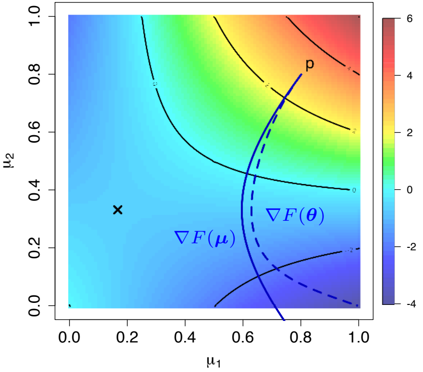
Gradient vectors in $\boldsymbol{\theta}$, $\lambda = 0.15$



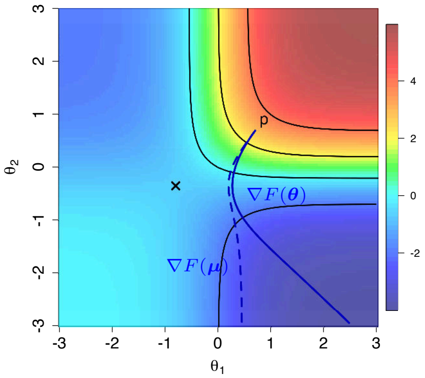
In the $\boldsymbol{\theta}$ parameters, $\nabla F(\boldsymbol{\theta})$ vanishes over the plateaux

Gradient Flows on the Independence Model

Marginal probabilities $\boldsymbol{\mu}$



Natural parameters $\boldsymbol{\theta}$



Gradient flows $\nabla F(\boldsymbol{\xi})$ depend on the parameterization

In the $\boldsymbol{\eta}$ parameters, $\nabla F(\boldsymbol{\eta})$ does not convergence to the expected distribution $\delta_{\boldsymbol{x}^*}$ over an optimum

Summary of the Intro

- Iterative algorithm generate sequences of distributions which can be compared to the gradient flow of $F(\boldsymbol{\xi})$
- The choice of the model and of the parameterization play a role in terms of appearance of local minima
- Euclidean geometry does not appear to be the proper geometry for statistical models

We need a more general mathematical framework able to deal with non-Euclidean geometries

The Exponential Family

In the following, we consider models in the exponential family $\mathcal{E}$

$$p(\boldsymbol{x}, \boldsymbol{\theta}) = \exp\left(\sum_{i=1}^{m} \theta_i T_i(\boldsymbol{x}) - \psi(\boldsymbol{\theta})\right)$$

- sufficient statistics $\boldsymbol{T} = (T_1(\boldsymbol{x}), \ldots, T_m(\boldsymbol{x}))$
- natural parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m) \in \Theta \subset \mathbb{R}^m$
- log-partition function $\psi(\boldsymbol{\theta})$

Several statistical models belong to the exponential family, both in the continuous and discrete case, among them

- the independence model
- the Gibbs distribution
- Markov random fields
- multivariate Gaussians

Markov Random Fields

[Recall] The monomials $\{\boldsymbol{x}^{\boldsymbol{\alpha}}\}, \boldsymbol{\alpha} \in L$, define a basis for f

By choosing a subset of $\{\boldsymbol{x}^{\boldsymbol{\alpha}}\}$ as sufficient statistics, we can identify a low-dimensional exponential family parametrized by $\boldsymbol{\theta}$

$$p(\boldsymbol{x};\boldsymbol{\theta}) = \exp\left(\sum_{\boldsymbol{\alpha} \in M \subset L_0} \theta_{\boldsymbol{\alpha}} \boldsymbol{x}^{\boldsymbol{\alpha}} - \psi(\boldsymbol{\theta})\right), \qquad L_0 = L \setminus \{0\}$$

Such models are known as

- log-liner models
- Markov random fields
- Boltzmann machines

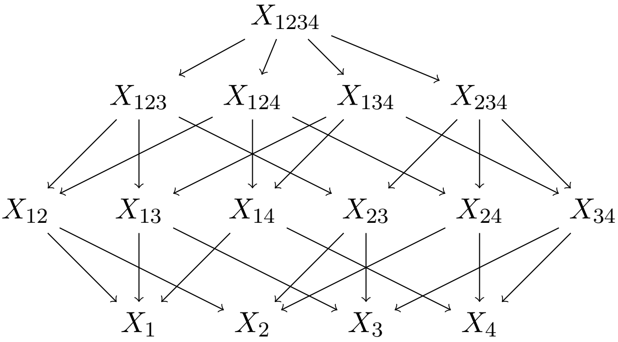
There exists a statistical interpretation for the topology of the model

- The absence of edges in an undirected graphical model implies conditional independence among variables
- Joint probability distributions factorize over the cliques of the graph

Hierarchical Markov Random Fields, $n = 4$

In some cases it may be convenient to work with hierarchical models

A hierarchical model has the property that if $\boldsymbol{x}^{\boldsymbol{\alpha}}$ is a sufficient statistic of the exponential family then all $\{\boldsymbol{x}^{\boldsymbol{\beta}} : \boldsymbol{\beta} \in L \setminus \{\mathbf{0}\} \wedge \boldsymbol{\beta} \to \boldsymbol{\alpha}\}$, must be sufficient statistics as well, where → is the bitwise implication operator



Dual Parameterization for the Exponential Family

$$p(\boldsymbol{x};\boldsymbol{\theta}) = \exp\left(\sum_{i=1}^{m} \boldsymbol{\theta}_i T_i(\boldsymbol{x}) - \psi(\boldsymbol{\theta})\right)$$

- Exponential families admit a dual parametrization to the natural parameters, given by the expectation parameters with $\boldsymbol{\eta} = \mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{T}]$
- Let $\varphi(\boldsymbol{\eta})$ be the negative entropy of p, then $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are connected by the Legendre transform

$$\psi(\boldsymbol{\theta}) + \varphi(\boldsymbol{\eta}) - \langle \boldsymbol{\theta}, \boldsymbol{\eta} \rangle = 0$$

- Variable transformations are given by

$$\begin{aligned} \boldsymbol{\eta} &= \nabla_{\boldsymbol{\theta}}\psi(\boldsymbol{\theta}) = (\nabla_{\boldsymbol{\eta}}\varphi)^{-1}(\boldsymbol{\theta}) \\ \boldsymbol{\theta} &= \nabla_{\boldsymbol{\eta}}\varphi(\boldsymbol{\eta}) = (\nabla_{\boldsymbol{\theta}}\psi)^{-1}(\boldsymbol{\eta}) \end{aligned}$$

Variable Transformations

[Recall] Let $A^n = \underbrace{A^1 \otimes \ldots \otimes A^1}_{\text{n times}}$, where $\otimes$ denotes the Kronecker product

A probability distribution $p \in \Delta$ requires 2^n parameters $\boldsymbol{\rho} = (p(\boldsymbol{x}))_{\boldsymbol{x} \in \Omega}$ to be uniquely identified, with constraints $0 \leq \rho_{\boldsymbol{x}} \leq 1$ and $\sum_{\boldsymbol{x} \in \Omega} \rho_{\boldsymbol{x}} = 1$

The expectation parameters $\boldsymbol{\eta} = (\eta_{\boldsymbol{\alpha}}), \boldsymbol{\alpha} \in L$, provide an equivalent parameterization for p, and since $p(\boldsymbol{x})$ is a pseudo-Boolean function itself, we have

$$\boldsymbol{\rho} = 2^{-n} A^n \boldsymbol{\eta} \qquad\qquad \boldsymbol{\eta} = A^n \boldsymbol{\rho}$$

Positivity constraints and the fact that probabilities sum to one, give us $\eta_0 = 1$ and $A^n \boldsymbol{\eta} \geq 0$.

The natural parameters $\boldsymbol{\theta} = (\theta_{\boldsymbol{\alpha}})$, $\boldsymbol{\alpha} \in L$, can be obtained from raw probabilities, with the constraint $\theta_{\mathbf{0}} = -\log \mathbb{E}_{\boldsymbol{\theta}}[\exp \sum_{\boldsymbol{\alpha} \in L \setminus \{\mathbf{0}\}} \theta_{\boldsymbol{\alpha}} \boldsymbol{x}^{\boldsymbol{\alpha}}]$

$$\ln \boldsymbol{\rho} = 2^{-n} A^n \boldsymbol{\theta} \qquad\qquad \boldsymbol{\theta} = A^n \ln \boldsymbol{\rho}$$

Mixed Parametrization for Markov Random Fields

An exponential family $\mathcal{M}$ given by the sufficient statistics $\{\boldsymbol{x}^{\boldsymbol{\alpha}}\}, \boldsymbol{\alpha} \in M$, identifies a submanifold in Δ, parametrized by $\boldsymbol{\theta} = ((\boldsymbol{\theta})_{\boldsymbol{\alpha} \in M}; \mathbf{0})$

By the one-to-one correspondence between $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$, $\mathcal{M}$ can be parametrized by $\boldsymbol{\eta} = (\boldsymbol{\eta}_{\boldsymbol{\alpha} \in M}; \boldsymbol{\eta}^*_{\boldsymbol{\alpha} \notin M})$, where in general $\boldsymbol{\eta}^*_{\boldsymbol{\alpha} \notin M} \neq \mathbf{0}$

However, the $\boldsymbol{\eta}^*_{\boldsymbol{\alpha} \notin M}$ parameters are not free and it can be proved they are given by implicit polynomial algebraic equations in $\boldsymbol{\eta}_{\boldsymbol{\alpha} \in M}$

Due to the duality between $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$, we can employ a mixed parametrization for $\mathcal{M}$ and parametrize the model as $(\boldsymbol{\eta}_{\boldsymbol{\alpha} \in M}; \mathbf{0})$

[Remark] The study of statistical models using tools from computational algebra is called Algebraic Statistics (Pistone et al., 2000)

Algebraic Statistics: Invariants in ρ and η

[Example] Let $n = 2$, we consider the independence model parametrized by $(\theta_1, \theta_2; 0)$, with $\theta_{12} = 0$

The same model can be parametrized by $(\eta_1, \eta_2; 0)$, we show $\eta_{12} = \eta_1\eta_2$

Since $\boldsymbol{\theta} = A^n \ln \boldsymbol{\rho}$, by imposing $\theta_{12} = 0$ we have

$$\ln \rho_{++} + \ln \rho_{--} = \ln \rho_{+-} + \ln \rho_{-+}$$

$$\rho_{++}\rho_{--} = \rho_{+-}\rho_{-+}$$

$$\begin{bmatrix} \rho_{++} \\ \rho_{+-} \\ \rho_{-+} \\ \rho_{--} \end{bmatrix} = \frac{1}{4} \times \begin{array}{cc} & \begin{array}{cccc} 00 & 10 & 01 & 11 \end{array} \\ \begin{array}{c} ++ \\ +- \\ -+ \\ -- \end{array} & \begin{bmatrix} +1 & +1 & +1 & +1 \\ +1 & +1 & -1 & -1 \\ +1 & -1 & +1 & -1 \\ +1 & -1 & -1 & +1 \end{bmatrix} \end{array} \begin{bmatrix} 1 \\ \eta_1 \\ \eta_2 \\ \eta_{12} \end{bmatrix}$$

$$(1 + \eta_1 + \eta_2 + \eta_{12})(1 - \eta_1 - \eta_2 + \eta_{12}) = (1 + \eta_1 - \eta_2 - \eta_{12})(1 - \eta_1 + \eta_2 - \eta_{12})$$

$$\eta_{12} = \eta_1\eta_2$$

Marginal Polytope

The range of the expectation parameters $\boldsymbol{\eta} = \mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{T}]$ identifies a polytope M in $\mathbb{R}^m$ called the marginal polytope

The marginal polytope can be obtained as the convex hull of $\boldsymbol{T}(\Omega)$, there $\boldsymbol{T}$ is the vector of sufficient statistics of the model

[Example] Let $n = 2$, $\boldsymbol{T} = (x_1, x_1 x_2)$

$$A = \begin{array}{c} \\ {\scriptstyle ++} \\ {\scriptstyle +-} \\ {\scriptstyle -+} \\ {\scriptstyle --} \end{array} \begin{array}{c} \begin{array}{cc} x_1 & x_1 x_2 \end{array} \\ \left[\begin{array}{cc} -1 & +1 \\ +1 & -1 \\ +1 & +1 \\ -1 & -1 \end{array}\right] \end{array}$$

Convex hull of

$$(+1,+1)$$
$$(+1,-1)$$
$$(-1,-1)$$
$$(-1,+1)$$

Marginal Polytope

The marginal polytope corresponds to the domain for the $\boldsymbol{\eta}$ parameters in the SR

- For the independence model $\mathrm{M} = [-1, 1]^n$
- For the saturated model $\mathrm{M} = \Delta$
- In the other cases, things can get very "nasty", indeed the number of its faces can grow more than exponentially in n

[Example] Let $n = 3$, consider the exponential model with sufficient statistics given by

$$\{x_1, x_2, x_3, x_{12}, x_{23}, x_{13}\}$$

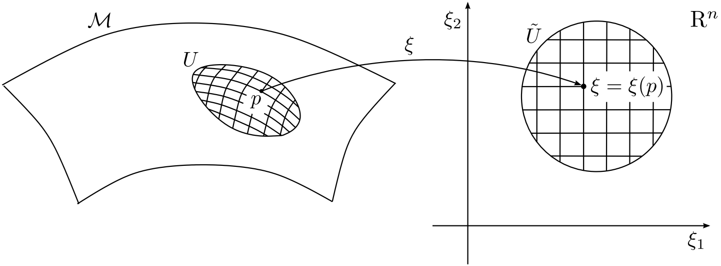
then the number of hyperplanes of M is 16

Information Geometry

The geometry of statistical models is not Euclidean

We need tools from differential geometry to define notions such as tangent vectors, shortest paths and distances between distributions

Information Geometry (IG) consists of the study of statistical models as manifolds of distributions endowed with the Fisher information metric (Amari 1982, 2001)



Characterization of the Tangent Space of $\mathcal{E}$

Over the manifold of distributions we can introduce an affine chart in p such that any density q is locally represented w.r.t. to the reference measure p by $\frac{q}{p} - 1$

The tangent space at each point p is defined by

$$\mathsf{T}_p = \{v : \mathbb{E}_p[v] = 0\}$$

Consider a curve $p(\theta)$ such that $p(0) = p$, then $\frac{\dot{p}}{p} \in \mathsf{T}_p$

In a moving coordinate system, tangent (velocity) vectors in $\mathsf{T}_{p(\theta)}$ to the curve are given by logarithmic derivative

$$\frac{\dot{p}(\theta)}{p(\theta)} = \frac{d}{d\theta} \log p(\theta)$$

Characterization of the Tangent Space of $\mathcal{E}$

The one dimensional model

$$p(\theta) = \exp\{\theta T - \psi(\theta)\}$$

is a curve in the manifold, with tangent (velocity) vector

$$\frac{\dot{p}(\theta)}{p(\theta)} = T - \frac{d}{d\theta}\psi(\theta)$$

On the other side, given a vector field, at each p we have a vector $U(p)$ tangent to some curve, then we obtain the differential equation

$$\frac{d}{d\theta}\log p(\theta) = U(p),$$

whose solutions are one-dimensional models in Δ

Natural Gradient

Let $(\mathcal{M}, I)$ be a statistical manifold endowed with a metric $I = [g_{ij}]$, and let $F(p) : \mathcal{M} \mapsto \mathbb{R}$ be a smooth function

For each vector field U over $\mathcal{M}$, the natural gradient of $\widetilde{\nabla} F(p)$, i.e., the direction of steepest descent of $F(p)$ is the unique vector that satisfies

$$g(\widetilde{\nabla} F, U) = \mathrm{D}_U F,$$

where $\mathrm{D}_U F$ is the directional derivative of F in the direction U

Given a coordinate system $\boldsymbol{\xi}$ for $\mathcal{M}$ we have

$$\widetilde{\nabla} F(\boldsymbol{\xi}) = \sum_{i=1}^{n} \sum_{j=1}^{n} g^{ij} \frac{\partial F}{\partial \xi_i} \frac{\partial}{\partial \xi_j} = I(\boldsymbol{\xi})^{-1} \nabla F(\boldsymbol{\xi})$$

[Remark] There is only one (natural) gradient given by the geometry of $\mathcal{M}$

We use $\widetilde{\nabla} F(\boldsymbol{\xi})$ to distinguish the natural gradient from the vanilla gradient $\nabla F(\boldsymbol{\xi})$, i.e., the vector of partial derivatives of F w. r. t. $\boldsymbol{\xi}$

Geometry of the Exponential Family

In case of a finite sample space Ω, we have

$$p(\boldsymbol{x};\boldsymbol{\theta}) = \exp\left(\sum_{i=1}^{m} \theta_i T_i(\boldsymbol{x}) - \psi(\boldsymbol{\theta})\right) \quad \boldsymbol{\theta} \in \mathbb{R}^m$$

and

$$\mathsf{T}_{\boldsymbol{\theta}} = \left\{ v : v = \sum_{i=1}^{k} a_i (T_i(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{\theta}}[T_i]), a_i \in \mathbb{R} \right\}$$

Since $\nabla F(\boldsymbol{\theta}) = \mathrm{Cov}_{\boldsymbol{\theta}}(f, T)$, if $f \in \mathsf{T}_p$, the steepest direction is given by $f - \mathbb{E}_{\boldsymbol{\theta}}[f]$, otherwise we take the projection $\widehat{f}$ of f onto T_p

$$\widehat{f} = \sum_{i=1}^{m} \widehat{a}_i (T_i(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{\theta}}[T_i]),$$

and obtain $\widehat{f}$ by solving a system of linear equations

Geometry of Statistical Models

Since $f - \widehat{f}$ is orthogonal to T_p

$$\mathbb{E}_{\boldsymbol{\theta}}[(f - \widehat{f}_{\boldsymbol{\theta}})(T - \mathbb{E}_{\boldsymbol{\theta}}[T])] = \operatorname{Cov}_{\boldsymbol{\theta}}(f - \widehat{f}_{\boldsymbol{\theta}}, T) = 0,$$

from which we obtain, for $i = 1, \dots, m$

$$\operatorname{Cov}_{\boldsymbol{\theta}}(f, T_i) = \operatorname{Cov}_{\boldsymbol{\theta}}(\widehat{f}_{\boldsymbol{\theta}}, T_i) = \sum_{j=1}^{m} \widehat{a}_j \operatorname{Cov}_{\boldsymbol{\theta}}(T_j, T_i)$$

As the Hessian matrix of $\psi(\boldsymbol{\theta})$ is invertible, we have

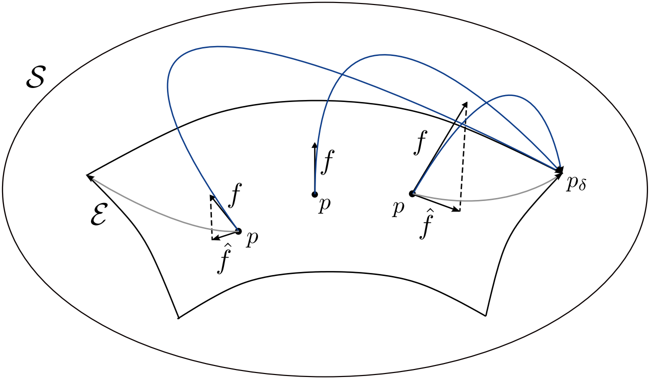
$$\widehat{\boldsymbol{a}} = [\operatorname{Cov}_{\boldsymbol{\theta}}(T_i, T_j)]^{-1} \operatorname{Cov}_{\boldsymbol{\theta}}(f, T) = I(\boldsymbol{\theta})^{-1} \nabla F(\boldsymbol{\theta})$$

In case $f \in \operatorname{Span}\{T_1, \dots, T_m\}$, then $\widehat{f}_{\boldsymbol{\theta}} = f$

By taking the projection of f onto T_p, we obtained the natural gradient $\widetilde{\nabla} F$, i.e., the gradient evaluated w.r.t. the Fisher information metric I

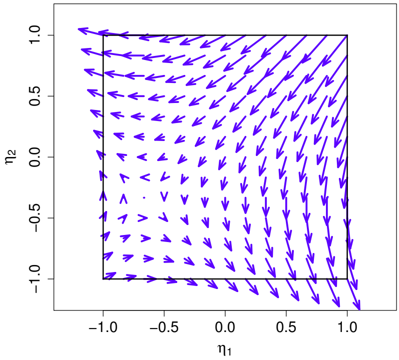
The Big Picture

If $f \notin \mathsf{T}_p$, the projection $\widehat{f}$ may vanish, and local minima may appear

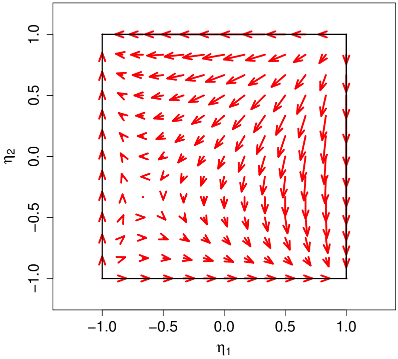


Vanilla vs Natural Gradient: $\eta, \lambda = 0.05$

Vanilla gradient $\nabla F(\boldsymbol{\eta})$

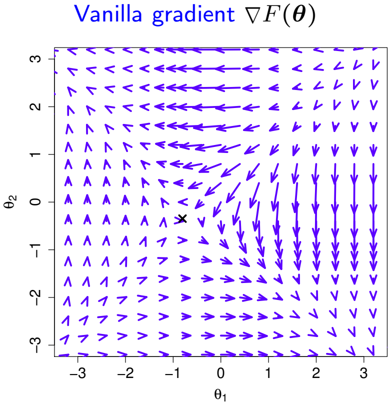


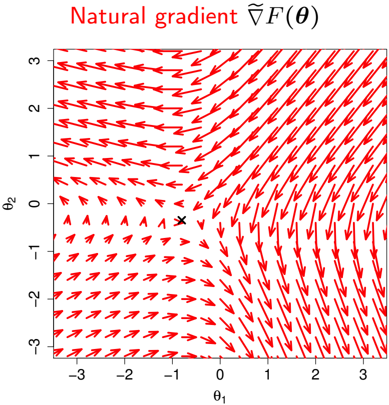
Natural gradient $\widetilde{\nabla} F(\boldsymbol{\eta})$



In both cases there exist two basins of attraction, however $\widetilde{\nabla} F(\boldsymbol{\eta})$ convergences to $\delta_{\boldsymbol{x}}$ distributions, which are local optima for $F(\boldsymbol{\eta})$ and where $\widetilde{\nabla} F(\delta_{\boldsymbol{x}}) = 0$

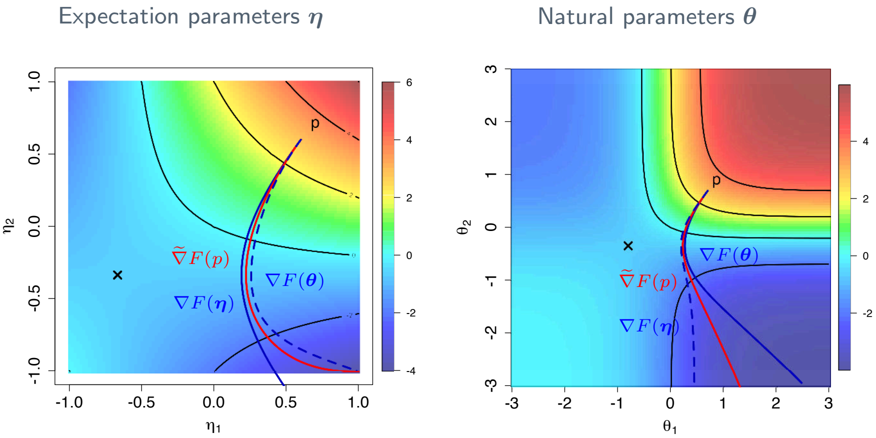
Vanilla vs Natural Gradient: $\boldsymbol{\theta}, \lambda = 0.15$


Vanilla gradient $\nabla F(\boldsymbol{\theta})$


Natural gradient $\widetilde{\nabla} F(\boldsymbol{\theta})$

In both cases there exist two basins of attraction, however in the natural parameters $\widetilde{\nabla} F(\boldsymbol{\theta})$ “speeds up” over the plateaux

Vanilla vs Natural Gradient


Expectation parameters $\boldsymbol{\eta}$

Natural parameters $\boldsymbol{\theta}$

Vanilla gradient ∇F vs Natural gradient $\widetilde{\nabla} F$

The natural gradient flow is invariant to parameterization

Stochastic Natural Gradient Descent

In the exponential family, the natural gradient descent updating rule reads

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^{t} - \lambda I(\boldsymbol{\theta})^{-1} \nabla F(\boldsymbol{\theta}), \qquad \lambda > 0$$

Unfortunately, exact gradients cannot be computed efficiently

- in general the partition function must be evaluated
- or a change of parametrization from $\boldsymbol{\theta}$ to $\boldsymbol{\eta}$ is required

However, due to the properties of the exponential family, natural gradient can be evaluated by means of covariances

$$\nabla F(\boldsymbol{\theta}) = \mathrm{Cov}_{\boldsymbol{\theta}}(f, \boldsymbol{T}) \qquad\qquad I(\boldsymbol{\theta}) = \mathrm{Cov}_{\boldsymbol{\theta}}(\boldsymbol{T}, \boldsymbol{T})$$

As a consequence, stochastic natural gradient can be estimated by replacing exact gradients with empirical estimates, so that

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^{t} - \lambda \widehat{\mathrm{Cov}}_{\boldsymbol{\theta}^t}(\boldsymbol{T}, \boldsymbol{T})^{-1} \widehat{\mathrm{Cov}}_{\boldsymbol{\theta}^t}(f, \boldsymbol{T}), \qquad \lambda > 0$$

Stochastic Natural Gradient Descent

We can now move from gradient flows to iterative algorithms that discretize the flow

A naïve SNGD (M. et al., 2011) implements the following pseudo-code

```
0 $t = 0, \boldsymbol{\theta}_t = \mathbf{0}$
1 Start with an initial random sample $\mathcal{P}_t$
2 Evaluate $f$ for each $\boldsymbol{x} \in \mathcal{P}_t$
3 (Subsample $\mathcal{P}_t$)
4 $\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \lambda \widehat{\mathrm{Cov}}_{\boldsymbol{\theta}}(\boldsymbol{T}, \boldsymbol{T})^{-1} \widehat{\mathrm{Cov}}_{\boldsymbol{\theta}}(f, \boldsymbol{T})$
5 Sample from $p(\boldsymbol{x}; \boldsymbol{\theta}^{t+1})$
6 Repeat 2-5 until convergence
```

Sampling from $p(\boldsymbol{x}; \boldsymbol{\theta}^{t+1})$ can be done using the Gibbs sampler or other sampling algorithms, for instance exploiting special conditional structure and properties of p

Gibbs Sampler with Cooling Scheme

```
Input: $\boldsymbol{\theta}, N$                                   ▷ natural parameters, sample size
Optional: $\mathcal{P}_0, T_0$                                    ▷ pool of samples, initial temperature
function GibbsSampler($\boldsymbol{\theta}, N, P_0, T_0$)
    $\mathcal{P} \leftarrow \varnothing$
    $t \leftarrow 1$
    repeat
        $\boldsymbol{x} \leftarrow$ Random($\mathcal{P}_0$)                        ▷ random point if $\mathcal{P}_0 = \varnothing$
        $T \leftarrow T_0$                                        ▷ initial temp, $T = 1$ default value
        repeat
            $i \leftarrow$ Random($\{1, \dots, n\}$)                     ▷ random variable
            $\boldsymbol{x}_{\setminus i} \leftarrow (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$
            $p_i(x_i|\boldsymbol{x}_{\setminus i}; \boldsymbol{\theta}) = \frac{1}{1+\exp\{2T^{-1} x_i \sum_{\alpha \in M_i} \theta_{\alpha \setminus i} \boldsymbol{x}^{\alpha \setminus i}\}}$
            $x_i \leftarrow \begin{cases} +1, & \text{with } \mathbb{P}_i(X_i = 1|\boldsymbol{X}_{\setminus i} = \boldsymbol{x}_{\setminus i}; \boldsymbol{\theta}) \\ -1, & \text{otherwise} \end{cases}$
            $T \leftarrow$ CoolingScheme($T$)                        ▷ decrease T
        until StoppingCriteria()
        $\mathcal{P} \leftarrow \mathcal{P} \cup \{\boldsymbol{x}\}$                                 ▷ add new point
        $t \leftarrow t + 1$
    until $t = N$
    return $\mathcal{P}$
end function
```

Convergence of the Natural Gradient Flows

Theorem 1 (M. et al., FOGA 2011)

If the sufficient statistics $\{\boldsymbol{x}^{\boldsymbol{\alpha}}\}$ of the exponential family

$$p(x;\boldsymbol{\theta}) = \exp\left(\sum_{\boldsymbol{\alpha}\in M\subset L_0} \theta_{\boldsymbol{\alpha}}\boldsymbol{x}^{\boldsymbol{\alpha}} - \psi(\boldsymbol{\theta})\right)$$

form a basis for

$$f(\boldsymbol{x}) = \sum_{\boldsymbol{\alpha}\in I} c_{\boldsymbol{\alpha}}\boldsymbol{x}^{\boldsymbol{\alpha}}$$

i.e., $I\setminus\{0\}\subset M$, then the natural gradient flows from every distribution p in the exponential family converge to the global optimum of the SR

Natural Gradient and Fitness Modeling

Theorem 2 (M. et al., CEC 2013)

If the sufficient statistics $\{\boldsymbol{x}^{\boldsymbol{\alpha}}\}$ of the exponential family

$$p(x;\boldsymbol{\theta}) = \exp\left(\sum_{\boldsymbol{\alpha}\in M\subset L_0} \theta_{\boldsymbol{\alpha}}\boldsymbol{x}^{\boldsymbol{\alpha}} - \psi(\boldsymbol{\theta})\right)$$

are centered in $\boldsymbol{\theta}$, i.e., $\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{x}^{\boldsymbol{\alpha}}] = 0$ for every $\boldsymbol{\alpha} \in M$, then the least squares estimator $\widehat{a}_N$ of the regression model

$$\widehat{f}(\boldsymbol{x}) = \sum_{\boldsymbol{\alpha}\in M\subset L_0} a_{\boldsymbol{\alpha}}\boldsymbol{x}^{\boldsymbol{\alpha}}$$

estimated from a random sample $\mathcal{P}$ i. i. d. $\sim p(x;\boldsymbol{\theta})$ converges to the natural gradient $\widetilde{\nabla}F(\boldsymbol{\theta})$, as $N \to \infty$

Natural Gradient and Fitness Modeling

Proof. Let A be the design matrix, e.g.,

$$A = \begin{array}{c|cccc} & x_1 & x_2 & x_1x_2 & \dots \\ \hline +--++- & +1 & -1 & -1 & \dots \\ --+++- & -1 & -1 & +1 & \dots \\ ++-+-+ & +1 & +1 & +1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{array}$$

The least squares estimator reads

$$\begin{aligned} \widehat{a}_N &= (A^\top A)^{-1} A^\top \boldsymbol{y} \\ &= \left[\frac{1}{N}\sum_{\boldsymbol{x}\in\mathcal{P}} \boldsymbol{x}^{\boldsymbol{\alpha}}\boldsymbol{x}^{\boldsymbol{\beta}}\right]^{-1} \left(\frac{1}{N}\sum_{\boldsymbol{x}\in\mathcal{P}} f(\boldsymbol{x})\boldsymbol{x}^{\boldsymbol{\alpha}}\right) \\ &= \left[\widehat{\mathrm{Cov}}(\boldsymbol{x}^{\boldsymbol{\alpha}},\boldsymbol{x}^{\boldsymbol{\beta}}) + \widehat{\mathbb{E}}[\boldsymbol{x}^{\boldsymbol{\alpha}}]\widehat{\mathbb{E}}[\boldsymbol{x}^{\boldsymbol{\beta}}]\right]^{-1} \left(\widehat{\mathrm{Cov}}(f,\boldsymbol{x}^{\boldsymbol{\alpha}}) + \widehat{\mathbb{E}}[f]\widehat{\mathbb{E}}[\boldsymbol{x}^{\boldsymbol{\alpha}}]\right) \\ &= \left[\widehat{\mathrm{Cov}}(\boldsymbol{x}^{\boldsymbol{\alpha}},\boldsymbol{x}^{\boldsymbol{\beta}})\right]^{-1} \left(\widehat{\mathrm{Cov}}(f,\boldsymbol{x}^{\boldsymbol{\alpha}})\right) \end{aligned}$$

Model selection and Model Selection

The previous theorem determines a relationship between linear regression in machine learning and stochastic gradient estimation in optimization

- Natural gradient estimation can be solved by linear regression
- Stochastic natural gradient performs fitness modeling
- Model selection can be performed by subset selection during linear regression

Forward subset selection-style algorithms can be employed to simultaneously perform model selection and gradient estimation

1. Start with no sufficient statistics in the model
2. Choose among a set of candidates variables the one that improves the model the most
3. Evaluate the new residual vector $\boldsymbol{f} - \widehat{\boldsymbol{f}}$
4. Repeat 2-3 until RSS is less then a threshold or the maximum number of variables is reached

Choice of $\mathcal{M}$ in the SR

The choice of the statistical model $\mathcal{M}$ determines the landscape of F

To remove the local minima of F over $\mathcal{M}$, we can add higher-order interactions $\boldsymbol{x}^{\boldsymbol{\alpha}}$ to the model

[Example] In our toy example, with

$$f(\boldsymbol{x}) = x_1 + 2x_2 + 3x_1x_2$$

we can remove the critical point of $\widetilde{\nabla} F$ in $\mathcal{M}_1$ by adding the sufficient statistic X_{12} to the independence model and obtain

$$\{X_1, X_2\} \Rightarrow \{X_1, X_2, X_{12}\}$$

The independence model is a submanifold of the new (saturated) model: a distribution p in the independence model has the parameters associated to higher-order interactions set to 0

$$(\theta_1, \theta_2, 0) \Rightarrow (\theta_1, \theta_2, \theta_{12})$$

Variable transformation (Cucci et al., 2012)

An equivalent approach to change the landscape of F is based on variables transformation: we fix the model over a set of new variables $\boldsymbol{y} = (y_1, \ldots, y_n) \in \Omega$ obtained from $\boldsymbol{x}$

For instance, consider the mapping $\tau_{ij} : \Omega \to \Omega : \boldsymbol{x} \mapsto \boldsymbol{y}$ with

$$y_k = \begin{cases} x_k x_j & \text{if } k = i \\ x_k & \text{otherwise} \end{cases}$$

A large class of transformations can be obtained by concatenating simple τ_{ij} transformations

Mapping $\boldsymbol{y}$ back to $\boldsymbol{x}$ is easy, since $\tau_{ij} = \tau_{ij}^{-1}$

Learning a transformation of $\boldsymbol{x}$ which simplifies the landscape of F for a fixed $\mathcal{M}$ over $\boldsymbol{y}$ is equivalent to finding a good model for $\boldsymbol{x}$

Variables transformation for the independence model

[Example] Let us apply the transformation τ_{12} to $\boldsymbol{x} = (x_1, x_2)$

$$y_1 = x_1 x_2$$

$$y_2 = x_2$$

and consider the independence model over $\boldsymbol{y}$

$$p(\boldsymbol{y}; \boldsymbol{\zeta}) = \exp\{\zeta_1 y_1 + \zeta_2 y_2 - \psi(\boldsymbol{\zeta})\}$$

By representing the independence model for the $\boldsymbol{y}$ variables as a model for $\boldsymbol{x}$ in the probability simple Δ, we obtain the exponential model

$$p(\boldsymbol{x}; \boldsymbol{\zeta}) = \exp\{\zeta_2 x_2 + \zeta_1 x_1 x_2 - \psi(\boldsymbol{\zeta})\}$$

which can be expressed in $\boldsymbol{\theta}$ by renaming the variables

$$p(\boldsymbol{x}; \boldsymbol{\theta}) = \exp\{\theta_2 x_2 + \theta_{12} x_1 x_2 - \psi(\boldsymbol{\theta})\}$$

The independence models for $\boldsymbol{x}$ and $\boldsymbol{y}$ correspond to different models in Δ

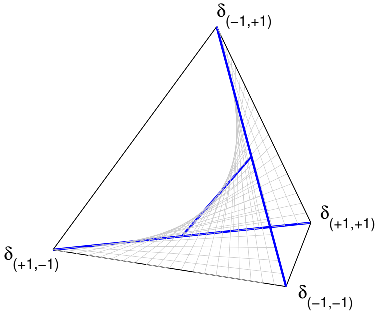
Back to the Toy Example: Submanifolds in Δ

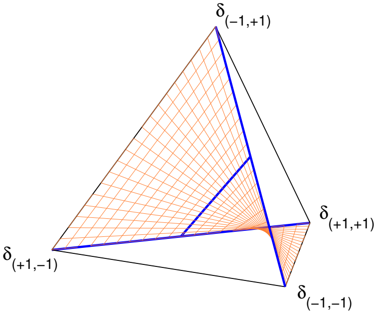
Independence model: $\boldsymbol{\theta} = (\theta_1, \theta_2, 0)$

$$p(\boldsymbol{x}; \boldsymbol{\theta}) = \exp\{\theta_1 x_1 + \theta_2 x_2 - \psi(\boldsymbol{\theta})\}$$

Exponential family: $\boldsymbol{\theta} = (0, \theta_2, \theta_{12})$

$$p(\boldsymbol{x}; \boldsymbol{\theta}) = \exp\{\theta_2 x_2 + \theta_{12} x_1 x_2 - \psi(\boldsymbol{\theta})\}$$





The lines represented in blue, together with all vertices of Δ, correspond to the distributions in the intersection of the closures of the models

Back to the Toy Example: Gradient Fields

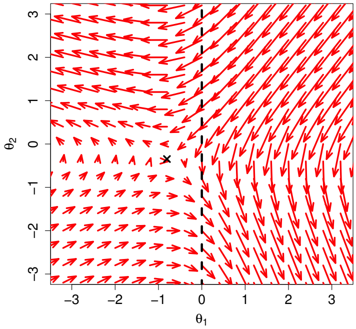
Let $f(\boldsymbol{x}) = x_1 + 2x_2 + 3x_1x_2$

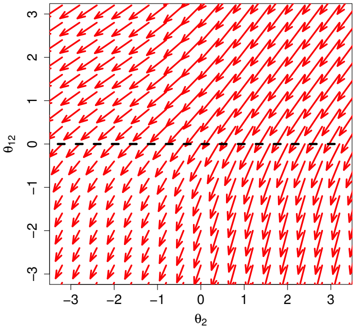
Independence model: $\boldsymbol{\theta} = (\theta_1, \theta_2, 0)$

$$p(\boldsymbol{x};\boldsymbol{\theta}) = \exp\{\theta_1 x_1 + \theta_2 x_2 - \psi(\boldsymbol{\theta})\}$$

Exponential family: $\boldsymbol{\theta} = (0, \theta_2, \theta_{12})$

$$p(\boldsymbol{x};\boldsymbol{\theta}) = \exp\{\theta_2 x_2 + \theta_{12} x_1 x_2 - \psi(\boldsymbol{\theta})\}$$





The dashed lines represent the intersection of the two models in Δ

In the new model there are no critical points for $\widetilde{\nabla} F$

Back to the Toy Example: Gradient Flows

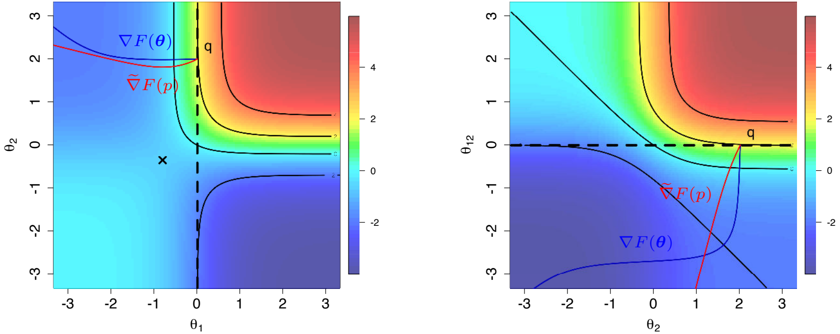
Let $f(\boldsymbol{x}) = x_1 + 2x_2 + 3x_1x_2$

Independence model: $\boldsymbol{\theta} = (\theta_1, \theta_2, 0)$

$$p(\boldsymbol{x};\boldsymbol{\theta}) = \exp\{\theta_1 x_1 + \theta_2 x_2 - \psi(\boldsymbol{\theta})\}$$

Exponential family: $\boldsymbol{\theta} = (0, \theta_2, \theta_{12})$

$$p(\boldsymbol{x};\boldsymbol{\theta}) = \exp\{\theta_2 x_2 + \theta_{12} x_1 x_2 - \psi(\boldsymbol{\theta})\}$$



Vanilla gradient ∇F vs Natural gradient $\widetilde{\nabla} F$

Linear Programming Relaxation

- A standard method to solve pseudo-Boolean optimization is linearization (see Boros and Hammer, 2001 for a survey)
- Every monomial $\boldsymbol{x}^{\boldsymbol{\alpha}}$ in f is replaced by a new binary variable $z_{\boldsymbol{\alpha}}$
- The problem is translated into a continuous linear problem by relaxing integer constraints for z, i.e.,

$$\text{(P)} \quad \arg\min \sum_{\alpha \in I} c_{\boldsymbol{\alpha}} \boldsymbol{x}^{\boldsymbol{\alpha}} \quad \Rightarrow \quad \text{(LP)} \quad \arg\min \sum_{\alpha \in I} c_{\boldsymbol{\alpha}} z_{\boldsymbol{\alpha}}$$

$$\text{s.t. } x_i \in \{+1, -1\} \qquad \qquad \text{s.t. } z_{\boldsymbol{\alpha}} \in [+1, -1]$$

- The problem becomes linear, but additional constraints are required to link the $\boldsymbol{z} = (x_{\boldsymbol{\alpha}}), \boldsymbol{\alpha} \in L$ variables to the $\boldsymbol{x} = (x_1, \dots, x_n)$ variables

$$z_{\boldsymbol{\alpha}} = \prod_{i=1}^{n} z_i^{\alpha_i} = \prod_{\{i:\alpha_i=1\}} z_i$$

Linear Programming Relaxation

- The new constraints are nonlinear, but can be modeled by linear inequalities (see Buchheim and Rinaldi, 2007, for a review)
- [Remark] For $x_i \in \{0, 1\}$ a standard linearization is given by

$$z_{\boldsymbol{\alpha}} \leq z_i \qquad \forall i : \alpha_i = 1 \qquad\qquad z_{\boldsymbol{\alpha}} \geq \sum_{i:\alpha_i=1} z_i - \|\boldsymbol{\alpha}\|_1 + 1$$

- As n increases, the LP relaxation requires a good (tight) half-space representation of the marginal polytope $\mathbb{M}$

Theorem 3 (M. et al., DISCML 2009)

The LP relaxation in the new variables $\boldsymbol{z}$ corresponds to the SR in $\boldsymbol{\eta}$

$$F(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{\boldsymbol{\alpha}\in L} c_{\boldsymbol{\alpha}} \boldsymbol{x}^{\boldsymbol{\alpha}}\right] = \sum_{\boldsymbol{\alpha}\in L} c_{\boldsymbol{\alpha}} \mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{x}^{\boldsymbol{\alpha}}] \rightarrow F(\boldsymbol{\eta}) = \sum_{\boldsymbol{\alpha}\in L} c_{\boldsymbol{\alpha}} \eta_{\boldsymbol{\alpha}} = \sum_{\boldsymbol{\alpha}\in L} c_{\boldsymbol{\alpha}} z_{\boldsymbol{\alpha}}$$

Solving the LP relaxation we obtain a lower bound for the SR since it is usually defined over an approximation of the marginal polytope $\mathbb{M}$

Back to the Toy Example

The landscape of $F(\boldsymbol{\eta})$ changes according to f and the choice of $\mathcal{M}$

[Example] Natural gradient flows in the $\boldsymbol{\eta}$ are given by

$$\begin{aligned}\dot{\eta}_1 &= (1-\eta_1^2)(a_1 + a_{12}\eta_2)\\ \dot{\eta}_2 &= (1-\eta_2^2)(a_2 + a_{12}\eta_1)\end{aligned}$$

We fix $\mathcal{M}$ as the independence model and study the flows for different a_{12}

The natural gradient vanishes over

- the vertices of the marginal polytope M
- $\boldsymbol{c} = (-a_2/a_{12}, -a_1/a_{12})^{\mathrm{T}}$

The nature of the critical points can be determined by studying the eigenvalues of the Hessian

$$M = \begin{bmatrix} -2\eta_1(a_1 + a_{12}\eta_2) & a_{12}(1-\eta_1^2) \\ a_{12}(1-\eta_2^2) & -2\eta_2(a_2 + a_{12}\eta_1) \end{bmatrix}$$

Back to the Toy Example: Critical Points

The solutions of the differential equations associated to the flows can be studied for every value of $\boldsymbol{\eta}$, even outside of M, where points do not represent distributions, since positivity constraints are violated

Let $\boldsymbol{v} \in \{-1,+1\}^2$ be a vertex of M, the eigenvalues of H are

$$\lambda_1 = -2v_1(a_{12}v_2 + a_1)$$
$$\lambda_2 = -2v_2(a_{12}v_1 + a_2)$$

According to the signs of λ_1 and λ_2, each vertex can be either a stable node (SN), an unstable node (UN) or a saddle point (SP)

For $\boldsymbol{c} = (-a_2/a_{12}, -a_1/a_{12})^{\mathrm{T}}$

$$\lambda_{1,2} = \pm\sqrt{(a_{12}^2 - a_2^2)(a_{12}^2 - a_1^2)/a_{12}^2}$$

Follows that $\boldsymbol{c}$ is saddle point for $(|a_{12}| \geq |a_1| \wedge |a_{12}| \geq |a_2|) \vee (|a_{12}| \leq |a_1| \wedge |a_{12}| \leq |a_2|)$, in the other cases, it is center (C)

Back to the Toy Example: Bifurcation Diagram

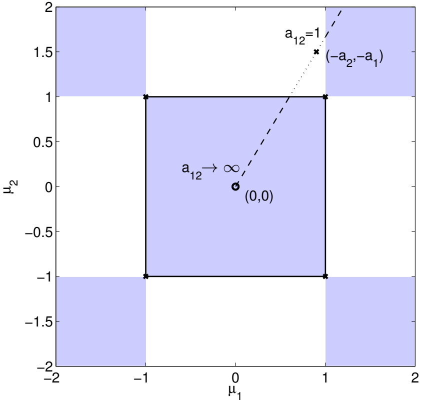
We can interpret $|a_{12}|$ as the strength of the interaction among x_1 and x_2

For $|a_{12}| \neq 0$, $\boldsymbol{c}$ is a saddle point in the shaded regions, where there exist

- strong interactions, $|a_{12}| > |a_1| \wedge |a_{12}| > |a_2|$, i.e. $\boldsymbol{c} \in \mathrm{M}$
- weak interactions, $|a_{12}| < |a_1| \wedge |a_{12}| < |a_2|$, i.e., $\boldsymbol{c} \notin \mathrm{M}$

In the remaining cases $\boldsymbol{c}$ is a center

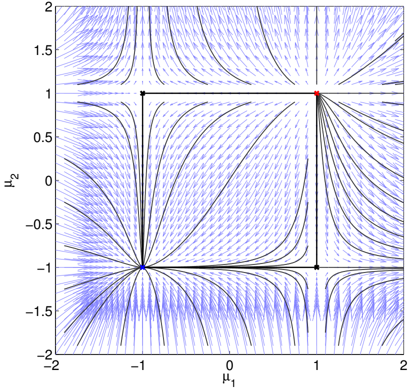
Projection of the bifurcation diagram (η_1, η_2, a_{12}) over (η_1, η_2) for arbitrary a_1, a_2 and $0 \leq a_{12} < \infty$

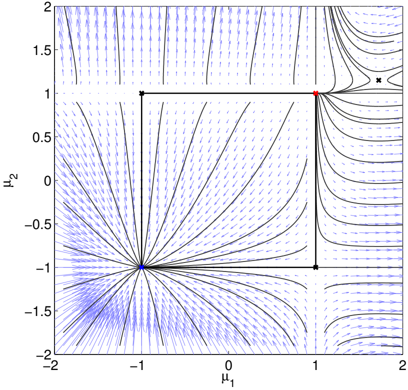


The coordinates of $\boldsymbol{c}$ depends on a_{12}, $\boldsymbol{c}$ is a SP on the dashed lines and a C on the dotted line; for $a_{12} \to \infty$, $\boldsymbol{c}$ converges to the center of M

Back to the Toy Example (M. et al., 2014)

Natural Gradient Flows over (η_1, η_2) for fixed a_{12}





$(a_{12} = 0)$ 1 SN, 1 UN and 2 SPs

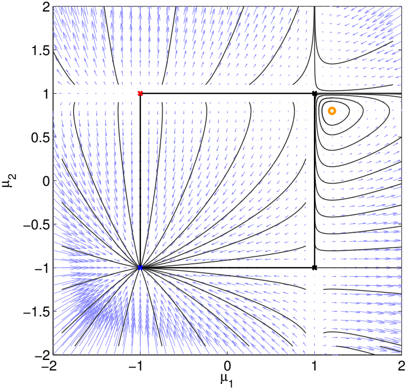
No critical points besides the vertices of M, all trajectories in M converge to the global optimum

$(a_{12} = 0.85)$ 1 SN, 1 UN and 3 SPs

The interaction is weak, $\boldsymbol{c}$ is a SP and is outside of M so that all flows converge to the global optimum

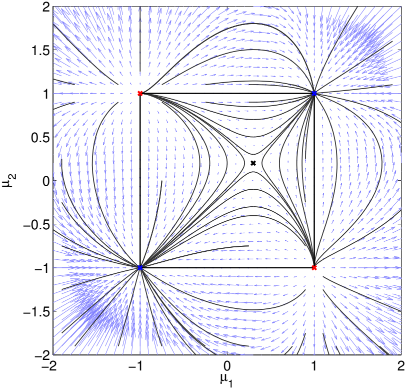
Back to the Toy Example (M. et al., 2014)

Natural Gradient Flows over (η_1, η_2) for fixed a_{12}



$(a_{12} = 1.25)$ 1 SN, 1 UN, SPs and 1 C

The interaction is not strong enough to have $c \in \mathrm{M}$ and to generate local minima, we have period solutions



$(a_{12} = 5)$ 2 SNs, 2 UNs and 1 SP

The interaction is strong, c is a SP and belongs to M, flows converge to either local or global optimum