Tutorial on Information Geometry and Algebraic Statistics

Luigi Malagò[1], Giovanni Pistone[2]

[1] Shinshu University & INRIA Saclay [2] Collegio Carlo Alberto

Algebraic Statistics Genova, June 11, 2015

Gradient Descent in Learning and Optimization

The use of natural gradient descent in statistics and machine learning was first proposed by Amari in 1998

- Policy Learning in Reinforment Learning
- Neural Networks Training
- Bayesian Variational Inference
- Stochastic Relaxation

Black-box Optimization

Suppose we want to optimized a function $f : \Omega \to \mathbb{R}$, however:

- you don't have direct access to an explicit formula for f
- given $\boldsymbol{x} \in \Omega$, you can evaluate $f(\boldsymbol{x}) \in \mathbb{R}$

Black-box Optimization

Suppose we want to optimized a function $f : \Omega \to \mathbb{R}$, however:

- you don't have direct access to an explicit formula for f
- given $\boldsymbol{x} \in \Omega$, you can evaluate $f(\boldsymbol{x}) \in \mathbb{R}$

One näive approach is a local search:

0. define a neighborhood function $\mathcal{V}(\boldsymbol{x}) \subset \Omega$
   $t = 0$
   $\boldsymbol{x}_0$ chosen randomly
1. $\boldsymbol{x}_{t+1} = \arg\max_{\boldsymbol{x} \in \mathcal{V}(\boldsymbol{x}_t)} f(\boldsymbol{x})$
2. $t = t + 1$
3. repeat 1-2 until convergence

Local Search Has Some Drawbacks

- For $\Omega = \mathbb{R}^n$, gradient cannot be evaluate directly, since f is unknown
- The choice of the $\mathcal{V}(\boldsymbol{x})$ may determine premature convergence to local minima

Local Search Has Some Drawbacks

- For $\Omega = \mathbb{R}^n$, gradient cannot be evaluate directly, since f is unknown
- The choice of the $\mathcal{V}(\boldsymbol{x})$ may determine premature convergence to local minima
- Some randomness may be useful in the exploration of the search space
- For large $\mathcal{V}(\boldsymbol{x})$, the search space can be sampled: random search

Local Search Has Some Drawbacks

- For $\Omega = \mathbb{R}^n$, gradient cannot be evaluate directly, since f is unknown
- The choice of the $\mathcal{V}(\boldsymbol{x})$ may determine premature convergence to local minima
- Some randomness may be useful in the exploration of the search space
- For large $\mathcal{V}(\boldsymbol{x})$, the search space can be sampled: random search

As an alternative approach, we can introduce a statistical model to guide the search for the optimum

A probability density function over Ω can be used to concentrate probability mass around certain regions of the search space

Some Notation: Finite Case

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

Some Notation: Finite Case

- $f(\boldsymbol{x}) : \Omega \to \mathbb{R}$ 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_n the n-dimensional probability simplex

Some Notation: Finite Case

- $f(\boldsymbol{x}) : \Omega \to \mathbb{R}$ 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_n the n-dimensional probability simplex
- $\mathcal{M} = \{p(\boldsymbol{x}; \boldsymbol{\xi}) : \boldsymbol{\xi} \in \Xi\} \subset \Delta$ a parametrized statistical model
- $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_d)$ a parameter vector for p

Stochastic Relaxation

Consider the minimization problem

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

Stochastic Relaxation

Consider the minimization problem

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

We define the Stochastic Relaxation (SR) of f as

$$F(p) = \mathbb{E}_p[f]$$

Stochastic Relaxation

Consider the minimization problem

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

We define the Stochastic Relaxation (SR) of f as

$$F(p) = \mathbb{E}_p[f]$$

We look for the minimum of f by optimizing its SR

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

[Remark 1] We take $\inf$, since in general $\mathcal{M}$ may not be closed

Stochastic Relaxation

Consider the minimization problem

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

We define the Stochastic Relaxation (SR) of f as

$$F(p) = \mathbb{E}_p[f]$$

We look for the minimum of f by optimizing its SR

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

[Remark 1] We take $\inf$, since in general $\mathcal{M}$ may not be closed

[Remark 2] Candidate solutions for P can be obtained by sampling

Optimization over a Statistical Manifold

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 onto a statistical model, from a discrete optimization problem over Ω to a continuous 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 for the SR

A Few Remarks

We move the search onto a statistical model, from a discrete optimization problem over Ω to a continuous 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 for the SR

[Remark 3] 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

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

Equivalence of P and SR

Let us denote the optimal solutions with:

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

Equivalence of P and SR

Let us denote the optimal solutions with:

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

The SR is equivalent to P if $p^*(\boldsymbol{x}^*) = 1$, i.e., we can sample optimal solutions of P from optimal solutions of SR with probability one

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

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

Equivalence of P and SR

Let us denote the optimal solutions with:

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

The SR is equivalent to P if $p^*(\boldsymbol{x}^*) = 1$, i.e., we can sample optimal solutions of P from optimal solutions of SR with probability one

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 $\overline{\mathcal{M}}$

How to solve the SR?

The SR in an optimization problem defined over a statistical model

It can be solved in many different ways, here we focus on natural gradient descent

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

How to solve the SR?

The SR in an optimization problem defined over a statistical model

It can be solved in many different ways, here we focus on natural gradient descent

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

Some references:

- 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

Which Model to Choose in the SR?

Let n be the cardinality of Ω, to parametrize Δ we need $n-1$ parameters

Minimizing $F(p)$ with $p \in \Delta$ is equivalent to an exhaustive search!

Which Model to Choose in the SR?

Let n be the cardinality of Ω, to parametrize Δ we need $n-1$ parameters

Minimizing $F(p)$ with $p \in \Delta$ is equivalent to an exhaustive search!

We need a low-dimensional models in the SR

- The equivalence of P and SR can be easily guaranteed
- The landscape (number of local minima) of $F(p)$ depends on the choice of $\mathcal{M}$
- In practice we need to learn $\mathcal{M}$: model selection
- Often it is conveniente to employ graphical models

Pseudo-Boolean Optimization

Consider the case where $\Omega = \{+1, -1\}^n$, where 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

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

Pseudo-Boolean Optimization

Consider the case where $\Omega = \{+1, -1\}^n$, where 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 $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n) \in L$ uniquely identifies the monomial $\boldsymbol{x}^{\boldsymbol{\alpha}}$ by

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

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

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

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

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

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

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

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

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

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

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

The Independence Model

Let $\mathcal{I}$ be the independence model for $\boldsymbol{X} = (X_1, \ldots, X_n)$

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

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

The Independence Model

Let $\mathcal{I}$ be the independence model for $\boldsymbol{X} = (X_1, \ldots, X_n)$

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

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

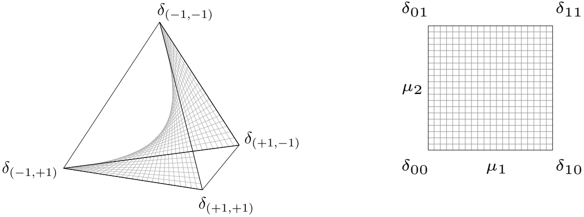
We parametrize $\mathcal{I}$ using $\{\pm 1\}$ Bernoulli distributions for X_i

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

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



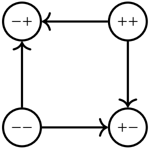
$\mathcal{I}$ is a n-dimensional manifold embedded in the $2^n - 1$ dimensional probability simplex Δ

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
$+1$	$+1$	6
$+1$	-1	-4
-1	$+1$	-2
-1	-1	0



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
$+1$	$+1$	6
$+1$	-1	-4
-1	$+1$	-2
-1	-1	0

$$\begin{tikzcd}
-+ & ++ \arrow[l] \arrow[d] \\
-- \arrow[u] \arrow[r] & +-
\end{tikzcd}$$

The gradient flow is the solution of the differential equation

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

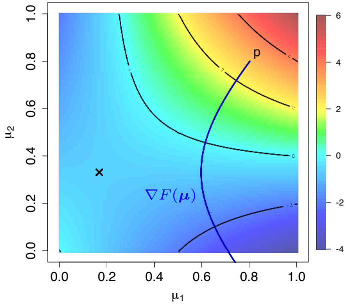
We are interested in studying gradient flows for different parameterization and different statistical models

Gradient Flows on the Independence Model

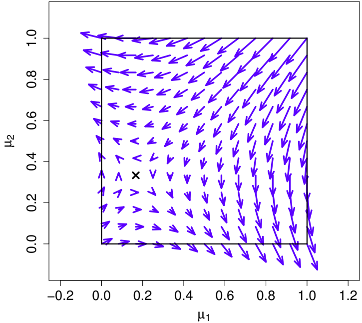
$$F(\boldsymbol{\mu}) = \sum_{\boldsymbol{x}\in\Omega} f(\boldsymbol{x})p_1(x_1)p_2(x_2) = -4\mu_1 - 2\mu_2 + 12\mu_1\mu_2$$

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

Gradient flow in $\boldsymbol{\mu}$



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

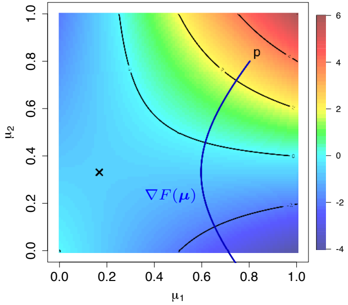


Gradient Flows on the Independence Model

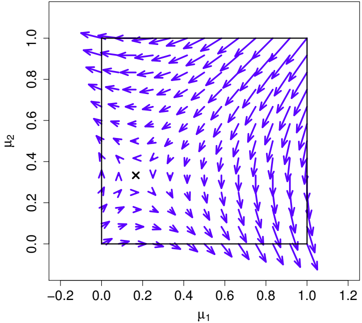
$$F(\boldsymbol{\mu}) = \sum_{\boldsymbol{x}\in\Omega} f(\boldsymbol{x})p_1(x_1)p_2(x_2) = -4\mu_1 - 2\mu_2 + 12\mu_1\mu_2$$

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

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 $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n) \in \mathbb{R}^n$, with

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

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 $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n) \in \mathbb{R}^n$, with

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

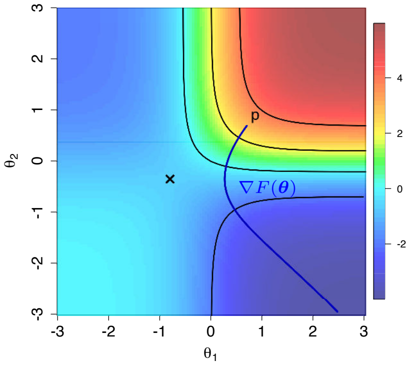
The mapping between marginal probabilities and natural parameters is one-to-one for $p > 0$

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

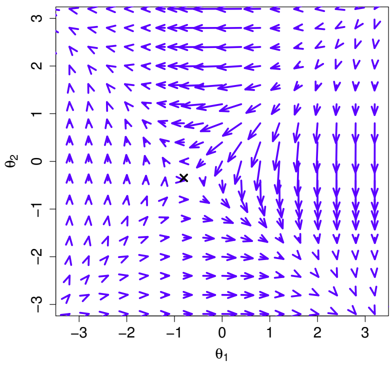
Gradient Flows on the Independence Model

$$F(\boldsymbol{\theta}) = (-4e^{\theta_1-\theta_2} - 2e^{-\theta_1+\theta_2} + 6e^{\theta_1+\theta_2})/Z(\boldsymbol{\theta})$$
$$\nabla F(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}[f(\boldsymbol{X} - \mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{X}])] = \mathrm{Cov}_{\boldsymbol{\theta}}(f, \boldsymbol{X})$$

Gradient flow in $\boldsymbol{\theta}$



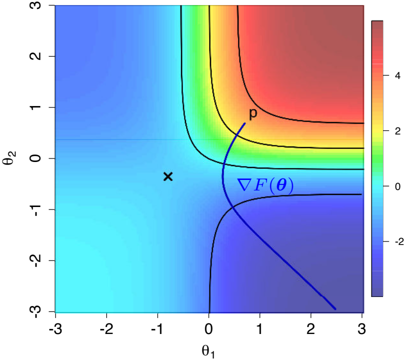
Gradient vectors in $\boldsymbol{\theta}$, $\lambda = 0.15$



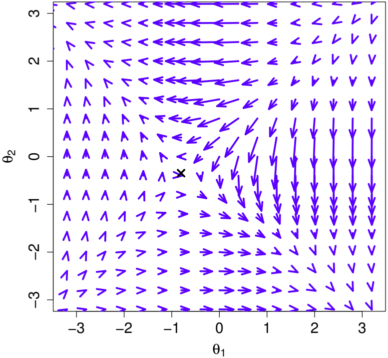
Gradient Flows on the Independence Model

$$F(\boldsymbol{\theta}) = (-4e^{\theta_1-\theta_2} - 2e^{-\theta_1+\theta_2} + 6e^{\theta_1+\theta_2})/Z(\boldsymbol{\theta})$$
$$\nabla F(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}[f(\boldsymbol{X} - \mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{X}])] = \mathrm{Cov}_{\boldsymbol{\theta}}(f, \boldsymbol{X})$$

Gradient flow in $\boldsymbol{\theta}$



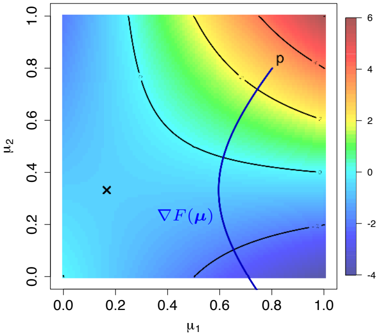
Gradient vectors in $\boldsymbol{\theta}$, $\lambda = 0.15$



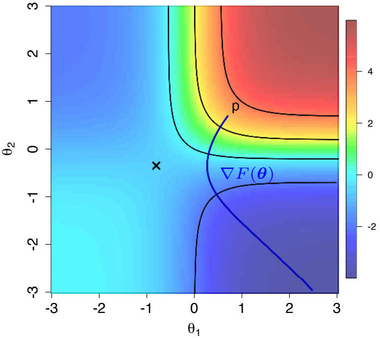
In the $\boldsymbol{\theta}$ parameters, $\nabla F(\boldsymbol{\theta})$ vanishes over the plateaux

Gradient Flows on the Independence Model

Marginal probabilities $\boldsymbol{\mu}$

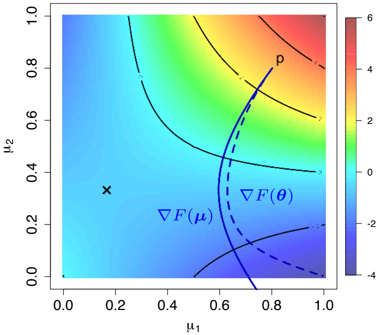


Natural parameters $\boldsymbol{\theta}$

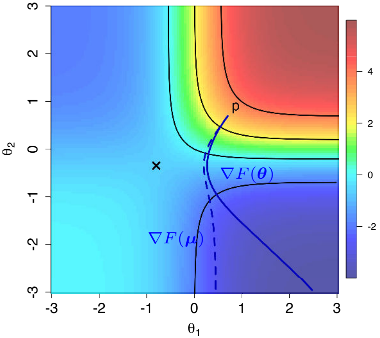


Gradient Flows on the Independence Model

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

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 $\psi(\boldsymbol{\theta})$

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
- 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

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

We have an 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

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 $\boldsymbol{\eta} = \mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{T}]$

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

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

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

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

Variable Transformations

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

Variable Transformations

[Recall] Let $A^n = \underbrace{A^1 \otimes \ldots \otimes A^1}_{\text{n times}}$, $\otimes$ 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 $\sum_{\boldsymbol{x} \in \Omega} \rho_{\boldsymbol{x}} = 1$

Variable Transformations

[Recall] Let $A^n = \underbrace{A^1 \otimes \ldots \otimes A^1}_{\text{n times}}$, $\otimes$ 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 $0 \le \rho_{\boldsymbol{x}} \le 1$ 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 sum to one, give us $\eta_0 = 1$ and $A^n \boldsymbol{\eta} \ge 0$.

Variable Transformations

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

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

Positivity constraints and 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 $\theta_{\mathbf{0}} = -\log \mathbb{E}_{\boldsymbol{\theta}}[\exp \sum_{\boldsymbol{\alpha} \in L \setminus \{\mathbf{0}\}} \theta_{\boldsymbol{\alpha}} \boldsymbol{x}^{\boldsymbol{\alpha}}]$

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

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 $\boldsymbol{\theta} = ((\boldsymbol{\theta})_{\boldsymbol{\alpha} \in M}; \mathbf{0})$

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 $\boldsymbol{\eta}^*_{\boldsymbol{\alpha} \notin M} \neq \mathbf{0}$

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 $\boldsymbol{\eta}_{\boldsymbol{\alpha} \in M}$

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 $(\boldsymbol{\eta}_{\boldsymbol{\alpha} \in M}; \mathbf{0})$

Algebraic Statistics: Invariants in $\boldsymbol{\rho}$ and $\boldsymbol{\eta}$

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

Algebraic Statistics: Invariants in $\boldsymbol{\rho}$ and $\boldsymbol{\eta}$

[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

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

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

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

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

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

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

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 $n = 2$, $\boldsymbol{T} = (x_1, x_1 x_2)$

$$A = \begin{array}{c} \\ ++ \\ +- \\ -+ \\ -- \end{array} \begin{array}{c} \begin{array}{cc} x_1 & x_1x_2 \end{array} \\ \left[\begin{array}{cc} -1 & +1 \\ +1 & -1 \\ +1 & +1 \\ -1 & -1 \end{array}\right] \end{array}$$

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

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

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

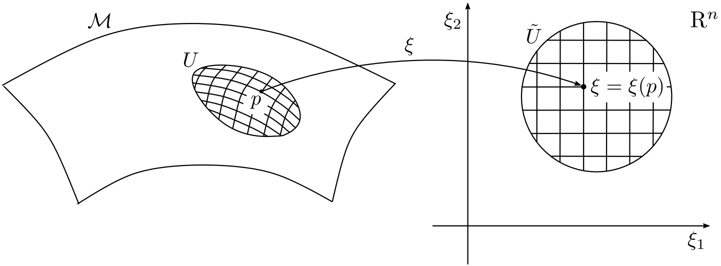
Information Geometry (IG) consists of the study of statistical models as manifolds of distributions endowed with the Fisher information metric (Amari 1982, 2001)

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)



Amari's Natural Gradient

Why $\widetilde{\nabla}_{\boldsymbol{\xi}} F(\boldsymbol{\xi})$ and not just $\nabla_{\boldsymbol{\xi}} F(\boldsymbol{\xi})$?

Amari's Natural Gradient

Why $\widetilde{\nabla}_{\boldsymbol{\xi}} F(\boldsymbol{\xi})$ and not just $\nabla_{\boldsymbol{\xi}} F(\boldsymbol{\xi})$?

[Short answer]

The geometry of $\mathcal{M}$ is not Euclidean

$\widetilde{\nabla}_{\boldsymbol{\xi}} F(\boldsymbol{\xi})$ is the natural gradient, i.e., the direction of steepest descent evaluated over a statistical model

In general $\widetilde{\nabla}_{\boldsymbol{\xi}} F(\boldsymbol{\xi})$ does not coincide with the vector of partial derivatives with respect to $\boldsymbol{\xi}$ denoted by $\nabla_{\boldsymbol{\xi}} F(\boldsymbol{\xi})$

Amari's Natural Gradient

[Longer answer]

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

Amari's Natural Gradient

[Longer answer]

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

For each vector field U over $\mathcal{M}$, the natural gradient $\widetilde{\nabla} F$, is the unique vector that satisfies

$$\langle \widetilde{\nabla} F, U \rangle_g = \mathrm{D}_U\, F,$$

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

Amari's Natural Gradient

[Longer answer]

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

For each vector field U over $\mathcal{M}$, the natural gradient $\widetilde{\nabla} F$, is the unique vector that satisfies

$$\langle \widetilde{\nabla} F, U \rangle_g = \mathrm{D}_U\, F,$$

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

Given a coordinate chart (a parameterization) $\boldsymbol{\xi}$ for $\mathcal{M}$, the representation in coordinates of $\widetilde{\nabla}_{\boldsymbol{\xi}} F(\boldsymbol{\xi})$ reads

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

Amari's Natural Gradient

[Longer answer]

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

For each vector field U over $\mathcal{M}$, the natural gradient $\widetilde{\nabla} F$, is the unique vector that satisfies

$$\langle \widetilde{\nabla} F, U \rangle_g = \mathrm{D}_U \, F,$$

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

Given a coordinate chart (a parameterization) $\boldsymbol{\xi}$ for $\mathcal{M}$, the representation in coordinates of $\widetilde{\nabla}_{\boldsymbol{\xi}} F(\boldsymbol{\xi})$ reads

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

The metric for $\mathcal{M}$ is the Fisher information matrix

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

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

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

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

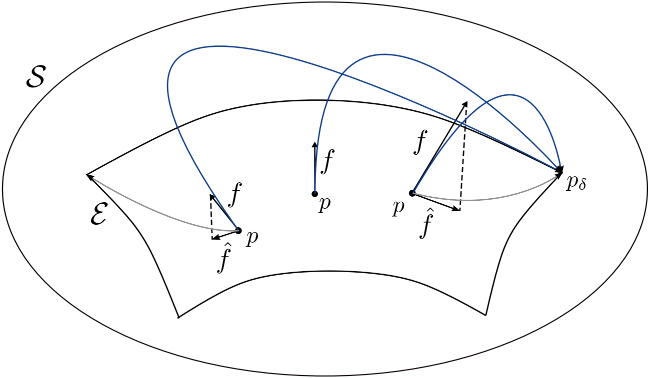
Since $\nabla F(\boldsymbol{\theta}) = \mathrm{Cov}_{\boldsymbol{\theta}}(f, T)$, if $f \in \mathsf{T}_p$, the steepest direction is given by $f - \mathbb{E}_{\boldsymbol{\theta}}[f]$, otherwise we take the projection $\widehat{f}$ of f onto T_p

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

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

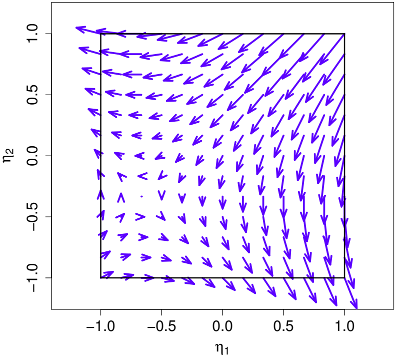
The Big Picture

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



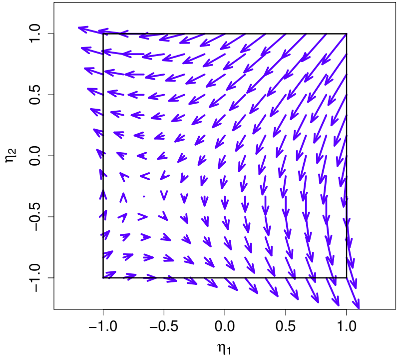
Vanilla vs Natural Gradient: $\boldsymbol{\eta}, \lambda = 0.05$

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

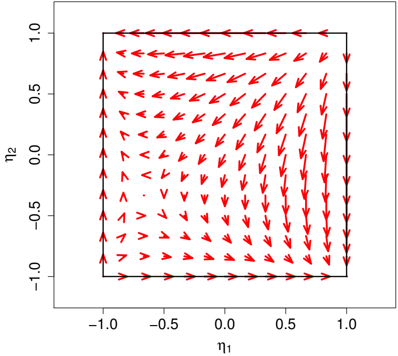


Vanilla vs Natural Gradient: $\boldsymbol{\eta}, \lambda = 0.05$

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



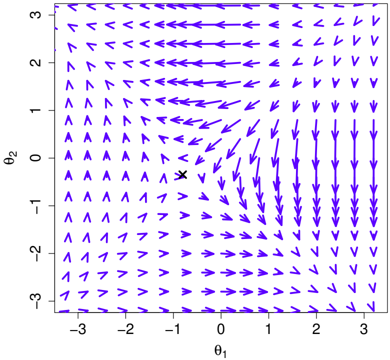
Natural gradient $\widetilde{\nabla} F(\boldsymbol{\eta})$



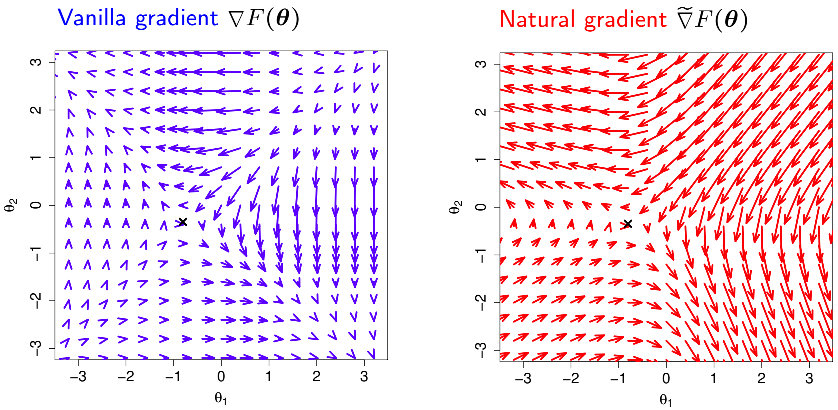
In both cases there exist two basins of attraction, however $\widetilde{\nabla} F(\boldsymbol{\eta})$ convergences to $\delta_{\boldsymbol{x}}$ distributions, which are local optima for $F(\boldsymbol{\eta})$ and where $\widetilde{\nabla} F(\delta_{\boldsymbol{x}}) = 0$

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

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

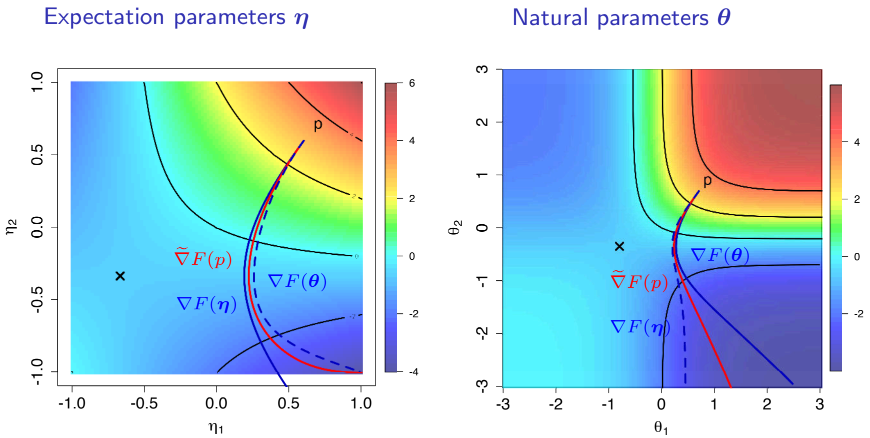


Vanilla vs Natural Gradient: $\boldsymbol{\theta}, \lambda = 0.15$



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



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

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

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

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

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

Back to the Toy Example

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

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

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

Back to the Toy Example

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

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

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

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

Back to the Toy Example

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

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

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

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

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

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)

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

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 a 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

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., $\boldsymbol{c} \notin \mathrm{M}$

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

Back to the Toy Example: Bifurcation Diagram

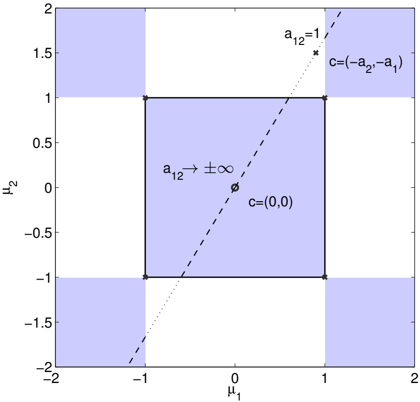
We can interpret $|a_{12}|$ as the strength of the interaction among x_1 and x_2

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

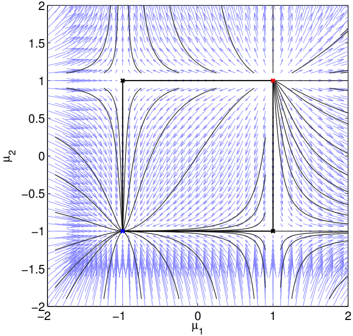
Projection of the bifurcation diagram (η_1, η_2, a_{12}) over (η_1, η_2) for arbitrary a_1, a_2 and $0 \leq a_{12} < \infty$



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}

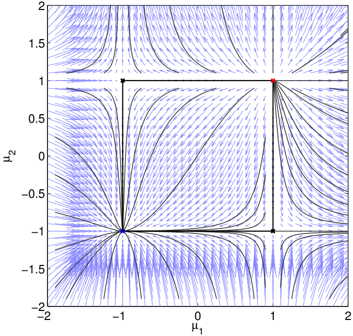


$(a_{12} = 0)$ 1 SN, 1 UN, 2 SPs

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

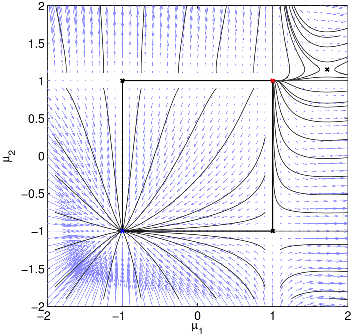
Back to the Toy Example (M. et al., 2014)

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



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

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

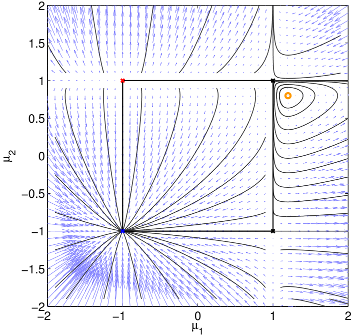


$(a_{12} = 0.85)$ 1 SN, 1 UN, 3 SPs

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}

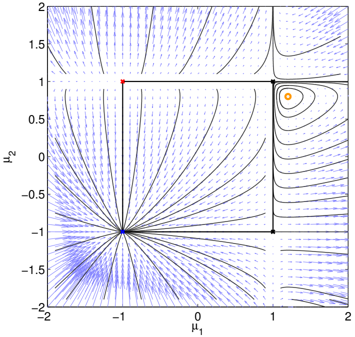


$(a_{12} = 1.25)$ 1 SN, 1 UN, SPs,1 C

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

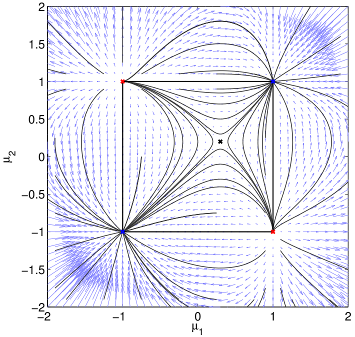
Back to the Toy Example (M. et al., 2014)

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



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

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



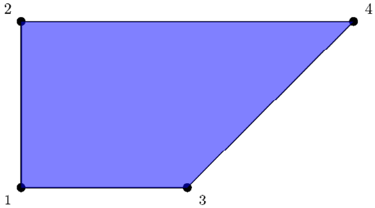
$(a_{12} = 5)$ 2 SNs, 2 UNs, 1 SP

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

A Second Toy Example

Consider the exponential family over $\Omega = \{1, 2, 3, 4\}$ given by the sufficient statistics T_1, T_2:

Ω	T_1	T_2
1	0	0
2	0	1
3	1	0
4	2	1



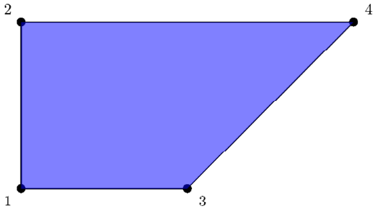
Marginal Polytope

$$p_{\boldsymbol{\theta}} = \exp\left(\theta_1 T_1 + \theta_2 T_2 - \psi(\boldsymbol{\theta})\right), \quad \psi(\boldsymbol{\theta}) = \log\left(1 + e^{\theta_2} + e^{\theta_1} + e^{2\theta_1 + \theta_2}\right)$$

A Second Toy Example

Consider the exponential family over $\Omega = \{1,2,3,4\}$ given by the sufficient statistics T_1, T_2:

Ω	T_1	T_2
1	0	0
2	0	1
3	1	0
4	2	1



Marginal Polytope

$$p_{\boldsymbol{\theta}} = \exp\left(\theta_1 T_1 + \theta_2 T_2 - \psi(\boldsymbol{\theta})\right), \quad \psi(\boldsymbol{\theta}) = \log\left(1 + e^{\theta_2} + e^{\theta_1} + e^{2\theta_1 + \theta_2}\right)$$

We are interested in natural gradient flows in the mixture geometry

Stochastic Relaxation

We generate a basis for all $f : \Omega \to \mathbb{R}$

$$\{1, x_1, x_2, x_{12}\}$$

Any f can be written as

$$f = c_0 + c_1 x_1 + c_2 x_2 + c_{12} x_1 x_2$$

Stochastic Relaxation

We generate a basis for all $f : \Omega \to \mathbb{R}$

$$\{1, x_1, x_2, x_{12}\}$$

Any f can be written as

$$f = c_0 + c_1 x_1 + c_2 x_2 + c_{12} x_1 x_2$$

We move to the SR with respect to the model identified by T_1, T_2

$$F(\boldsymbol{\eta}) = \mathbb{E}_{\boldsymbol{\eta}}[f] = c_0 + c_1 \eta_1 + c_2 \eta_2 + c_{12} \mathbb{E}_{\boldsymbol{\eta}}[x_1 x_2]$$

How do express $\mathbb{E}_{\boldsymbol{\eta}}[x_1 x_2]$ as a function of η_1, η_2?

Orthogonal Space and Markov Basis

Notice that the exponential family is a toric model

We can derive a Markov basis $\{T_3\}$ the orthogonal space of the space spanned by $\{1, T_1, T_2\}$

Ω	1	T_1	T_2	T_3
1	1	0	0	−2
2	1	0	1	1
3	1	1	0	2
4	1	2	1	−1

Derivation of the Invariant

From the exponential family $\log p_{\boldsymbol{\theta}} \in \operatorname{Span}(\{1, T_1, T_2\})$

$$\log p_{\boldsymbol{\theta}} = \theta_1 T_1 + \theta_2 T_2 - \psi(\boldsymbol{\theta}) \ ,$$

and thus $\log p_{\boldsymbol{\theta}} \perp T_3$

Let $T_3 = T_3^+ - T_3^- = (0,1,2,0) - (2,0,0,1)$, orthogonality can be rewritten as

$$\begin{aligned} 0 &= \sum_{x=1}^{4} \log p(x) T_3(x) \\ &= \sum_{x:T_3(x)>0} \log p(x) T_3^+(x) - \sum_{x:T_3(x)<0} \log p(x) T_3^-(x) \\ &= \log \left(\prod_{x:T_3(x)>0} p(x)^{T_3^+(x)} \right) - \log \left(\prod_{x:T_3(x)<0} p(x)^{T_3^-(x)} \right) \end{aligned}$$

Derivation of the Invariant (cont.)

Remember that $T_3 = T_3^+ - T_3^- = (0,1,2,0) - (2,0,0,1)$, by dropping the log in

$$0 = \log\left(\prod_{x:T_3(x)>0} p(x)^{T_3^+(x)}\right) - \log\left(\prod_{x:T_3(x)<0} p(x)^{T_3^-(x)}\right),$$

we obtain the polynomial invariant

$$p_1^2 p_4 - p_2 p_3^2 = 0$$

Derivation of the Invariant (cont.)

Remember that $T_3 = T_3^+ - T_3^- = (0,1,2,0) - (2,0,0,1)$, by dropping the log in

$$0 = \log\left(\prod_{x:T_3(x)>0} p(x)^{T_3^+(x)}\right) - \log\left(\prod_{x:T_3(x)<0} p(x)^{T_3^-(x)}\right),$$

we obtain the polynomial invariant

$$p_1^2 p_4 - p_2 p_3^2 = 0$$

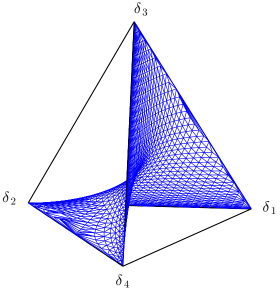
Our exponential family for positive probabilities is equivalently described by

$$\begin{aligned} p_1 + p_2 + p_3 + p_4 - 1 &= 0 \\ p_1^2 p_4 - p_2 p_3^2 &= 0 \end{aligned}$$

A Surface in the Probability Simplex

The model identifies a surface in the probability simplex

$$p_1 + p_2 + p_3 + p_4 - 1 = 0$$
$$p_1^2 p_4 - p_2 p_3^2 = 0$$



Probability Simplex Δ_3

Notice, the surface is not the independence model as in the previous example

Expectation Parameters

We introduce the following matrix:

$$
B = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array}
\begin{array}{c}
\begin{array}{cccc} \mathbf{1} & T_1 & T_2 & T_3 \end{array} \\
\begin{bmatrix}
1 & 0 & 0 & -2 \\
1 & 0 & 1 & 1 \\
1 & 1 & 0 & 2 \\
1 & 2 & 1 & -1
\end{bmatrix}
\end{array}
$$

In the simplex, probabilities maps into expected values one-to-one

$$
\begin{bmatrix} 1 \\ \eta_1 \\ \eta_2 \\ \eta_3 \end{bmatrix} =
\begin{bmatrix}
1 & 1 & 1 & 1 \\
0 & 0 & 1 & 2 \\
0 & 1 & 0 & 1 \\
-2 & 1 & 2 & -1
\end{bmatrix}
\begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{bmatrix}
$$

Expectation Parameters

We introduce the following matrix:

$$B = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} \mathbf{1} & T_1 & T_2 & T_3 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & -2 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 2 \\ 1 & 2 & 1 & -1 \end{bmatrix} \end{array}$$

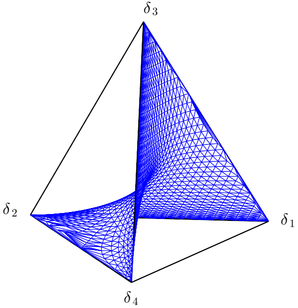
In the simplex, probabilities maps into expected values one-to-one

$$\begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{bmatrix} = \begin{bmatrix} \frac{3}{5} & -\frac{1}{5} & -\frac{2}{5} & -\frac{1}{5} \\ \frac{1}{5} & -\frac{2}{5} & \frac{7}{10} & \frac{1}{10} \\ \frac{2}{5} & \frac{1}{5} & -\frac{3}{5} & \frac{1}{5} \\ -\frac{1}{5} & \frac{2}{5} & \frac{3}{10} & -\frac{1}{10} \end{bmatrix} \begin{bmatrix} 1 \\ \eta_1 \\ \eta_2 \\ \eta_3 \end{bmatrix}$$

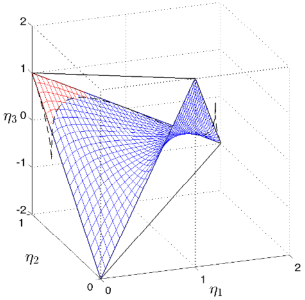
A Surface in the Full Marginal Polytope

Then, in the $\boldsymbol{\eta}$ parameters $p_1^2 p_4 - p_2 p_3^2 = 0$ becomes

$$(4\eta_1+3\eta_2-\eta_3-2)(\eta_1+2\eta_2+\eta_3-3)^2+(4\eta_1-7\eta_2-\eta_3-2)(\eta_1-3\eta_2+\eta_3+2)^2 = 0$$



Probability simplex Δ_3



Full marginal polytope

The surface on the right has been plotted by evaluating the unique real root in the interior of the marginal polytope

Back to the Stochastic Relaxation

We stopped at

$$F(\boldsymbol{\eta}) = c_0 + c_1\eta_1 + c_2\eta_2 + c_{12}\mathbb{E}_{\boldsymbol{\eta}}[x_1x_2]$$

Back to the Stochastic Relaxation

We stopped at

$$F(\boldsymbol{\eta}) = c_0 + c_1\eta_1 + c_2\eta_2 + c_{12}\mathbb{E}_{\boldsymbol{\eta}}[x_1x_2]$$

We have $T_3 = 4x_1 + 3x_2 - 5x_1x_2 - 2$ and $\eta_3 = \mathbb{E}[T_3]$, so that

$$\mathbb{E}[x_1x_2] = \frac{1}{5}(4\eta_1 + 3\eta_2 - \eta_3 - 2) \ ,$$

Back to the Stochastic Relaxation

We stopped at

$$F(\boldsymbol{\eta}) = c_0 + c_1\eta_1 + c_2\eta_2 + c_{12}\mathbb{E}_{\boldsymbol{\eta}}[x_1x_2]$$

We have $T_3 = 4x_1 + 3x_2 - 5x_1x_2 - 2$ and $\eta_3 = \mathbb{E}[T_3]$, so that

$$\mathbb{E}[x_1x_2] = \frac{1}{5}(4\eta_1 + 3\eta_2 - \eta_3 - 2) \ ,$$

which implies

$$F_\eta(\eta) = c_0 - \frac{2}{5}c_{12} + \left(c_1 + \frac{4}{5}c_{12}\right)\eta_1 + \left(c_2 + \frac{3}{5}c_{12}\right)\eta_2 - \frac{1}{5}c_{12}\eta_3 \ ,$$

where η_3 is the unique real root as a function of η_1, η_2

[Remark] The solution of the problem relies on being able to find real root of the invariant

Ruled Surfaces

Mauro C. Beltrametti
Ettore Carletti
Dionisio Gallarati
Giacomo Monti Bragadin

Lectures on Curves, Surfaces and Projective Varieties

A Classical View of Algebraic Geometry

Translated from the Italian
by Francis Sullivan

Ruled Surfaces

Mauro C. Beltrametti
Ettore Carletti
Dionisio Gallarati
Giacomo Monti Bragadin

Lectures on Curves, Surfaces and Projective Varieties

A Classical View of Algebraic Geometry

Translated from the Italian
by Francis Sullivan

5.8.15. Let $\mathcal{F}$ be the surface with equation $x_0^2x_1 - x_2^2x_3 = 0$. Noting that $\mathcal{F}$ has a double line and then observing that it is a ruled surface, find the singular generators and the pinch-points on the double line.

It is a well-known result that $p_1^2p_4 - p_2p_3^2 = 0$ is a ruled surface in Δ_3

Algebraic Varieties

In the polynomial ring $\mathbb{Q}[p_1, p_2, p_3, p_4]$, the model ideal

$$I = \left\langle p_1 + p_2 + p_3 + p_4 - 1, p_1^2 p_4 - p_2 p_3^2 \right\rangle$$

consists of all the polynomials of the form

$$A\,(p_1 + p_2 + p_3 + p_4 - 1) + B\left(p_1^2 p_4 - p_2 p_3^2\right), \quad \forall A, B \in \mathbb{Q}[p_1, p_2, p_3, p_4]$$

The algebraic variety I uniquely extends the exponential family outside of Δ_3, by means of the Zarinski closure

Exploiting Ruled Surfaces

Let us discuss in more detail the ruled parameterization of the toric variety

$$p_1 + p_2 + p_3 + p_4 - 1 = 0$$
$$p_1^2 p_4 - p_2 p_3^2 = 0$$

Exploiting Ruled Surfaces

Let us discuss in more detail the ruled parameterization of the toric variety

$$
\begin{aligned}
p_1 + p_2 + p_3 + p_4 - 1 &= 0 \\
p_1^2 p_4 - p_2 p_3^2 &= 0
\end{aligned}
$$

The Jacobian matrix is

$$
J = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2p_1p_4 & -p_3^2 & -2p_2p_3 & p_1^2 \end{bmatrix}
$$

It has rank one, that is, there is a singularity, if, and only if

$$
2p_1p_4 = -p_3^2 = -2p_2p_3 = p_1^2 \ ,
$$

which is equivalent to $p_1^2 = p_3^2 = 0$

Exploiting Ruled Surfaces

The subspace $p_1^2 = p_3^2 = 0$ intersects the model along a double critical line $\mathcal{C}$, and the simplex along the edge $\delta_2 \leftrightarrow \delta_4$

If we take a sheaf of planes through $\mathcal{C}$, by the Bezuot theorem, it intersects the cubic surface along $\mathcal{C}$ and on a a space of degree $3 - 2 = 1$

Exploiting Ruled Surfaces

The subspace $p_1^2 = p_3^2 = 0$ intersects the model along a double critical line $\mathcal{C}$, and the simplex along the edge $\delta_2 \leftrightarrow \delta_4$

If we take a sheaf of planes through $\mathcal{C}$, by the Bezuot theorem, it intersects the cubic surface along $\mathcal{C}$ and on a a space of degree $3 - 2 = 1$

That is, the system of equations

$$p_1 + p_2 + p_3 + p_4 - 1 = 0$$
$$p_1^2 p_4 - p_2 p_3^2 = 0$$
$$\alpha p_1 + \beta p_3 = 0$$

admits as a solution $\mathcal{C}$ and a line

Lowering the Degree on the Invariant

The system of equations

$$
\begin{aligned}
p_1 + p_2 + p_3 + p_4 - 1 &= 0\\
p_1^2 p_4 - p_2 p_3^2 &= 0\\
\alpha p_1 + \beta p_3 &= 0
\end{aligned}
$$

can be reduced to

$$
\begin{aligned}
p_1 + p_2 + p_3 + p_4 - 1 &= 0\\
\alpha p_1 + \beta p_3 &= 0\\
-\alpha^2 p_2 + \beta^2 p_4 &= 0
\end{aligned}
$$

A New Parametrization for the Model

In parametric form, the line in becomes

$$p_1([\alpha:\beta],t) = \beta t$$

$$p_2([\alpha:\beta],t) = \frac{\beta^2}{\alpha^2+\beta^2} + \frac{\beta^2(\alpha-\beta)}{\alpha^2+\beta^2}t$$

$$p_3([\alpha:\beta],t) = -\alpha t$$

$$p_4([\alpha:\beta],t) = \frac{\alpha^2}{\alpha^2+\beta^2} + \frac{\alpha^2(\alpha-\beta)}{\alpha^2+\beta^2}t$$

By setting $\alpha = \beta - 1$, $0 < t < 1, -1 < \alpha < 0$, we get:

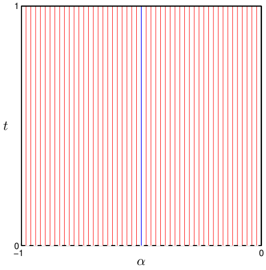
$$p_1(\alpha,t) = (\alpha+1)t$$

$$p_2(\alpha,t) = \frac{\alpha^2 - (\alpha^2+2\alpha+1)t + 2\alpha + 1}{2\alpha^2+2\alpha+1}$$

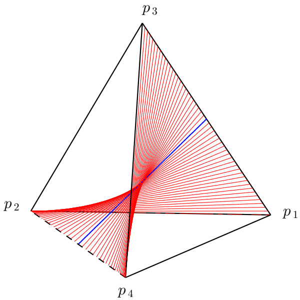
$$p_3(\alpha,t) = -\alpha t$$

$$p_4(\alpha,t) = -\frac{\alpha^2 t - \alpha^2}{2\alpha^2+2\alpha+1}$$

The Ruled Surface in Δ_3



(α, t) parameterization

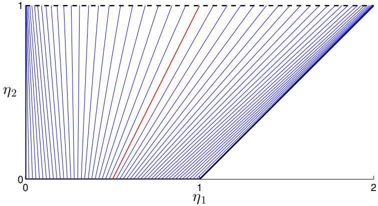


Probability simplex Δ_3

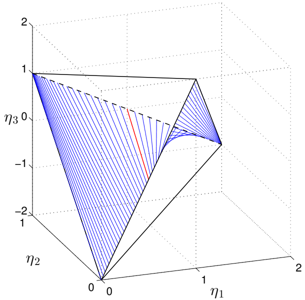
The critical line $\mathcal{C}$ is the dashed line

The Ruled Surface in the (Full) Marginal Polytope

By the linear mapping between $\boldsymbol{p}$ and $\boldsymbol{\eta}$, lines map to lines



Marginal polytope



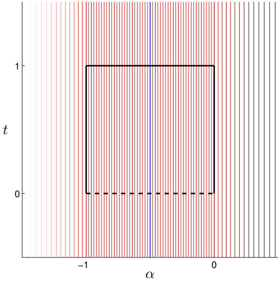
Full marginal polytope

Each line intersects $\delta_2 \leftrightarrow \delta_4$ and $\delta_1 \leftrightarrow \delta_3$ in

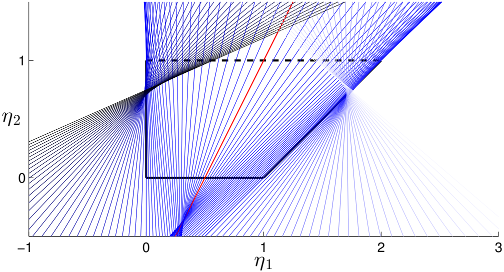
$$\boldsymbol{a} = \left(\frac{2\alpha^2}{2\alpha^2 + 2\alpha + 1}, 1, \frac{2\alpha + 1}{2\alpha^2 + 2\alpha + 1} \right) \qquad \boldsymbol{b} = (-\alpha, 0, -4\alpha - 2)$$

Extension of the Model

Lines can be extended outside of Δ_3 and of the marginal polytope

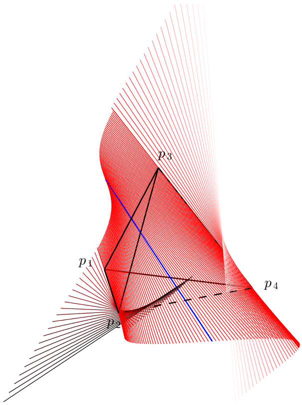


(α, t) parameterization

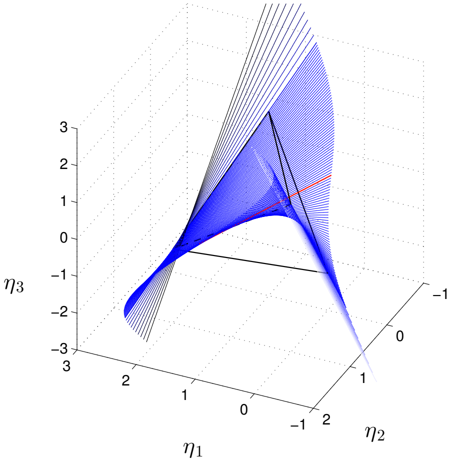


Marginal polytope

Extension of the Model



Probability Simplex Δ_3



Full marginal polytope

Back to the SR

The expectation parameters become rational functions of (α, t)

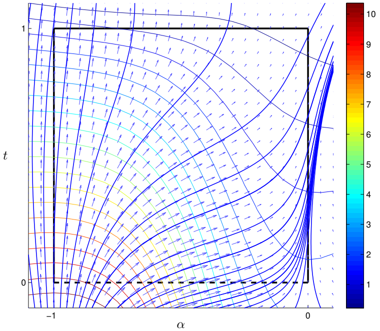
$$\begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{bmatrix} = \begin{bmatrix} -\alpha \\ 0 \\ -4\alpha - 2 \end{bmatrix} + t \begin{bmatrix} \frac{2\alpha^3+4\alpha^2+\alpha}{2\alpha^2+2\alpha+1} \\ 1 \\ \frac{8\alpha^3+12\alpha^2+10\alpha+3}{2\alpha^2+2\alpha+1} \end{bmatrix}$$

The same applies to the (inverse) Fisher Information matrix and the natural gradient, which now can be computed by

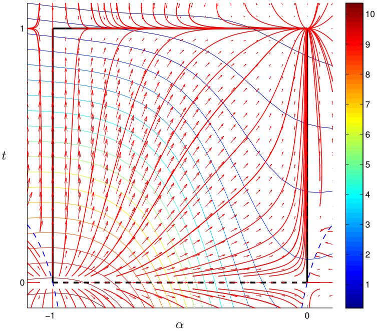
$$\widetilde{\nabla} F_\eta(\alpha, t) = I_\eta(\alpha, t)^{-1} \nabla F_\eta(\alpha, t)$$

Case 1: Gradient Flows in (α, t)

Consider the case with $c_0 = 0, c_1 = 1, c_2 = 2, c_3 = 3$



Vanilla gradient

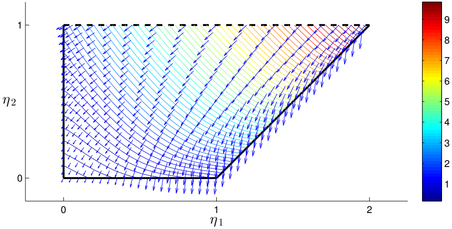


Natural gradient

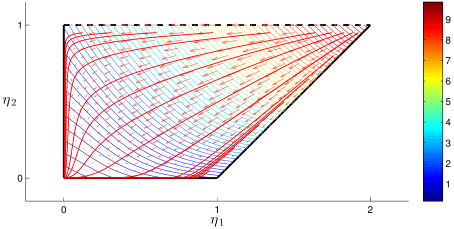
The function f admis one global minima

Case 1: Gradient Flows on the Marginal Polytope

Consider the case with $c_0 = 0, c_1 = 1, c_2 = 2, c_3 = 3$



Vanilla gradient

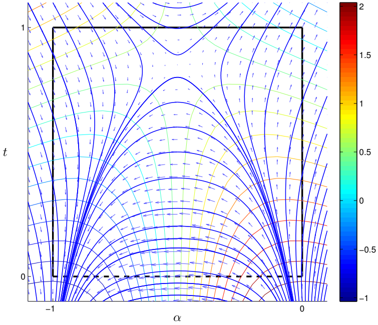


Natural gradient

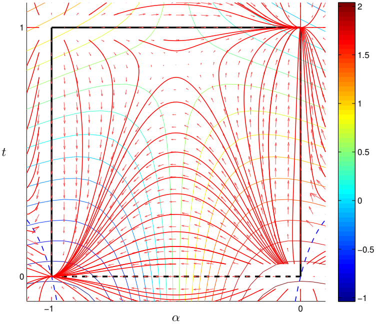
The function f admis one global minima

Case 2: Gradient Flows in (α, t)

Consider the case with $c_0 = 0, c_1 = 1, c_2 = 2, c_3 = -5/2$



Vanilla gradient

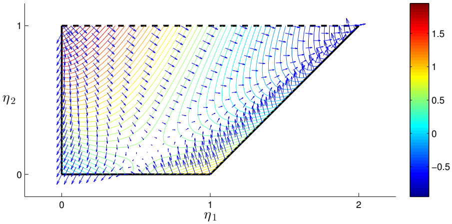


Natural gradient

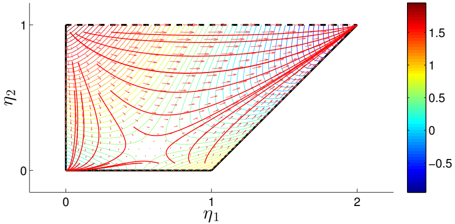
The function f admis two local minima

Case 2: Gradient Flows on the Marginal Polytope

Consider the case with $c_0 = 0, c_1 = 1, c_2 = 2, c_3 = -5/2$



Vanilla gradient



Natural gradient

The function f admis two local minima

Some Remarks

By exploiting the fact that surface in the probability simplex given by the invariant is a ruled surface, we introduced a new parametrization for the model

Some Remarks

By exploiting the fact that surface in the probability simplex given by the invariant is a ruled surface, we introduced a new parametrization for the model

In the new parametrization, the natural gradient is given by a rational formula

The model can be extended, the Fisher information matrix, and the natural gradient can be evaluated also for negative probabilities

The approach is more general than this specific example, and is based on the evaluation of the Markov basis for the orthogonal space and on the intersection of sheaf of planes on exposed faces of the model

Work in progress: the tutorial example is going to appear on Entropy this month, another paper is in preparation