Entropy **2015**, *17*, 4215-4254; doi:10.3390/e17064215

OPEN ACCESS
entropy
ISSN 1099-4300
www.mdpi.com/journal/entropy

Article

Natural Gradient Flow in the Mixture Geometry of a Discrete Exponential Family †

Luigi Malagò [1,2,*] and Giovanni Pistone [3]

[1] Department of Electrical and Electronic Engineering, Shinshu University, Nagano, Japan
[2] Inria Saclay, Île-de-France, Orsay Cedex, France
[3] De Castro Statistics, Collegio Carlo Alberto, Moncalieri, Italy;
E-Mail: giovanni.pistone@carloalberto.org

† This paper is an extended version of our paper published in the Proceedings of MaxEnt 2014 Conference on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Amboise, France, 21–26 September 2014.

* Author to whom correspondence should be addressed; E-Mail: malago@shinshu-u.ac.jp;
Tel./Fax: +81-26-269-5235.

Academic Editors: Frédéric Barbaresco and Ali Mohammad-Djafari

Received: 31 January 2015 / Accepted: 2 June 2015 / Published: 18 June 2015

Abstract: In this paper, we study Amari's natural gradient flows of real functions defined on the densities belonging to an exponential family on a finite sample space. Our main example is the minimization of the expected value of a real function defined on the sample space. In such a case, the natural gradient flow converges to densities with reduced support that belong to the border of the exponential family. We have suggested in previous works to use the natural gradient evaluated in the mixture geometry. Here, we show that in some cases, the differential equation can be extended to a bigger domain in such a way that the densities at the border of the exponential family are actually internal points in the extended problem. The extension is based on the algebraic concept of an exponential variety. We study in full detail a toy example and obtain positive partial results in the important case of a binary sample space.

Keywords: information geometry; stochastic relaxation; natural gradient flow; expectation parameters; toric models

1. Introduction

For the purpose of obtaining a clear presentation of our approach to the geometry of statistical models, we start with a recap of nonparametric statistical manifold; see, e.g., the review paper [1]. However, we will shortly move to the actual setup of the present paper, *i.e.*, the finite state space case.

Let $(\Omega, \mathcal{A}, \mu)$ be a measured space of sample points $\boldsymbol{x} \in \Omega$. We denote by $\mathcal{P}_{\geq} \subset L^1(\mu)$ the simplex of (probability) densities and by $\mathcal{P}_{>} \subset \mathcal{P}_{\geq}$ the convex set of strictly positive densities. If Ω is finite, then $\mathcal{P}_{>}$ is the topological interior of $\mathcal{P}_{\geq}$. We denote by $\mathcal{P}^1$ the affine space generated by $\mathcal{P}_{\geq}$.

The set $\mathcal{P}_{>}$ holds the exponential geometry, which is an affine geometry, whose geodesics are curves of the form $t \mapsto p_t \propto p_0^{1-t} p_1^t$. The set $\mathcal{P}^1$ holds the mixture geometry, whose geodesics are of the form $t \mapsto p_t = (1-t)p_0 + tp_1$. A proper definition of the exponential and mixture geometry, where probability densities are considered points, requires the definition of the proper tangent space to hold the vectors representing the velocity of a curve. In both cases, the tangent space T_p at a point p is a space of random variables V with zero expected value, $\mathrm{E}_p[V] = 0$. On the tangent space T_p, a natural scalar product is defined, $\langle U, V \rangle_p = \mathrm{E}_p[UV]$, so that a pseudo-Riemannian structure is available. Note that the Riemannian structure is a third geometry, different from both the exponential and the mixture geometries. Note also that both the expected value and the covariance can be naturally extended to be defined on $\mathcal{P}^1$.

For each lower bounded objective function $f\colon \Omega \to \mathbb{R}$ and each statistical model $\mathcal{M} \subset \mathcal{P}_{>}$, the (stochastic) relaxation of f to $\mathcal{M}$ is the function $F(p) = \mathrm{E}_p[f] \in \mathbb{R}$, $p \in \mathcal{M}$; *cf.* [2]. The minimization of the stochastic relaxation as a tool to minimize the objective function has been studied by many authors [3–7].

If we have a parameterization $\boldsymbol{\xi} \mapsto p_{\boldsymbol{\xi}}$ of $\mathcal{M}$, the parametric expression of the relaxed function is $\hat{F}(\boldsymbol{\xi}) = \mathrm{E}_{p_{\boldsymbol{\xi}}}[f]$. Under integrability and differentiability conditions on both $\boldsymbol{\xi} \mapsto p_{\boldsymbol{\xi}}$ and $\boldsymbol{x} \mapsto f(\boldsymbol{x})$, $\hat{F}$ is differentiable, with $\partial_j \hat{F}(\boldsymbol{\xi}) = \mathrm{E}_{p_{\boldsymbol{\xi}}}[\partial_j \log(p_{\boldsymbol{\xi}})\, f]$ and $\mathrm{E}_{p_{\boldsymbol{\xi}}}[\partial_j \log(p_{\boldsymbol{\xi}})] = 0$; see [1,8]. In order to properly describe the gradient flow of a relaxed random variable, these classical computations are better cast into the formal language of information geometry (see [9]) and, even better, in the language of non-parametric differential geometry [10] that was used in [11]. The previous computations suggest to take the Fisher score $\partial_j \log(p_{\boldsymbol{\xi}})$ as the definition of a tangent vector at the j-th coordinate curve. While the development of this analogy in the finite state space case does not require a special setup, in the non-finite state space, some care has to be taken.

In this paper, we follow the non-parametric setup discussed in [1] and, in particular, the notion of an exponential family $\mathcal{E}$ and the identification of the tangent space at each $p \in \mathcal{E}$ with a space of p-centered random variables.

The paper is organized as follows. We discuss in Section 2 the generalities of the finite state space case; in particular, we carefully define the various notions of the Fisher information matrix and natural gradient that arise from a given parameterization. In Section 3, we discuss a toy example in order to introduce the construction of an algebraic variety extending the exponential family from positive probabilities $\mathcal{P}_{>}$ to signed probabilities $\mathcal{P}^1$; this construction is applied to the natural gradient flow in the expectation parameters; moreover, it is shown that this model has a variety that is ruled. The last Section 4 is devoted to the treatment of the special important case when the sample space is binary.

The present paper is a development of the paper [12], which was presented as a poster at the MaxEnt Conference 2014. While the topic is the same, the actual overlapping between the two papers is minimal and concerns mainly the generalities that are repeated for the convenience of the reader.

2. Gradient Flow of Relaxed Optimization

Let Ω be a finite set of points $\boldsymbol{x} = (x_1, \dots, x_n)$ and μ the counting measure of Ω. In this case, a density $p \in \mathcal{P}_{\geq}$ is a probability function, *i.e.*, $p \colon \Omega \to \mathbb{R}_+$, such that $\sum_{\boldsymbol{x}\in\Omega} p(\boldsymbol{x}) = 1$.

Let $\mathcal{B} = \{T_1, \dots, T_d\}$ be a set of random variables, such that, if $\sum_{j=1}^{d} c_j T_j$ is constant, then $c_1 = \cdots = c_d = 0$; for instance consider $\mathcal{B}$ such that $\sum_{\boldsymbol{x}\in\Omega} T_j(\boldsymbol{x}) = 0$, $j = 0, \dots, d$, and $\mathcal{B}$ is a linear basis. We say that $\mathcal{B}$ is a set of affinely independent random variables. If $\mathcal{B}$ is a linear basis, it is affinely independent if and only if $\{1, T_1, \dots, T_d\}$ is a linear basis.

We consider the statistical model $\mathcal{E}$ whose elements are uniquely identified by the natural parameters $\boldsymbol{\theta}$ in the exponential family with sufficient statistics $\mathcal{B}$, namely:

$$p_{\boldsymbol{\theta}} \in \mathcal{E} \quad \Leftrightarrow \quad \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{i=1}^{d} \theta_i T_i(\boldsymbol{x}) - \psi(\boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \mathbb{R}^d \,,$$

see [13].

The proper convex function $\psi \colon \mathbb{R}^d$,

$$\boldsymbol{\theta} \mapsto \psi(\boldsymbol{\theta}) = \log \sum_{\boldsymbol{x}\in\Omega} \mathrm{e}^{\boldsymbol{\theta}\cdot\boldsymbol{T}(\boldsymbol{x})} = \boldsymbol{\theta} \cdot \mathrm{E}_{p_{\boldsymbol{\theta}}} [\boldsymbol{T}] - \mathrm{E}_{p_{\boldsymbol{\theta}}} [\log (p_{\boldsymbol{\theta}})]$$

is the cumulant generating function of the sufficient statistics $\boldsymbol{T}$, in particular,

$$\nabla\psi(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}} [\boldsymbol{T}] \,, \quad \operatorname{Hess} \psi(\boldsymbol{\theta}) = \operatorname{Cov}_{\boldsymbol{\theta}} (\boldsymbol{T}, \boldsymbol{T}) \,.$$

Moreover, the entropy of $p_{\boldsymbol{\theta}}$ is:

$$H(p_{\boldsymbol{\theta}}) = -\mathrm{E}_{p_{\boldsymbol{\theta}}} [\log (p_{\boldsymbol{\theta}})] = \psi(\boldsymbol{\theta}) - \boldsymbol{\theta} \cdot \nabla\psi(\boldsymbol{\theta}) \,.$$

The mapping $\nabla\psi$ is one-to-one onto the interior M° of the marginal polytope, that is the convex span of the values of the sufficient statistics $M = \{\boldsymbol{T}(\boldsymbol{x}) | \boldsymbol{x} \in \Omega\}$. Note that no extra condition is required, because on a finite state space, all random variables are bounded. Nonetheless, even in this case, the proof is not trivial; see [13].

Convex conjugation applies [14] (Section 25) with the definition:

$$\psi_*(\boldsymbol{\eta}) = \sup \left\{ \boldsymbol{\theta} \in \mathbb{R}^d \middle| \boldsymbol{\theta} \cdot \boldsymbol{\eta} - \psi(\boldsymbol{\theta}) \right\} , \quad \boldsymbol{\eta} \in \mathbb{R}^d \,.$$

The concave function $\boldsymbol{\theta} \mapsto \boldsymbol{\eta} \cdot \boldsymbol{\theta} - \psi(\boldsymbol{\theta})$ has divergence mapping $\boldsymbol{\theta} \mapsto \boldsymbol{\eta} - \nabla\psi(\boldsymbol{\theta})$, and the equation $\boldsymbol{\eta} = \nabla\psi(\boldsymbol{\theta})$ has a solution if and only if $\boldsymbol{\eta}$ belongs to the interior M° of the marginal polytope. The restriction $\phi = \psi_*|_{M^\circ}$ is the Legendre conjugate of ψ, and it is computed by:

$$\phi \colon M^\circ \ni \boldsymbol{\eta} \mapsto \in (\nabla\psi)^{-1}(\boldsymbol{\eta}) \cdot \boldsymbol{\eta} - \psi \circ (\nabla\psi)^{-1}(\boldsymbol{\eta}) \in \mathbb{R} \,.$$

The Legendre conjugate ϕ is such that $\nabla\phi = (\nabla\psi)^{-1}$, and it provides an alternative parameterization of $\mathcal{E}$ with the so-called expectation or mixture parameter $\boldsymbol{\eta} = \nabla\psi(\boldsymbol{\theta})$,

$$p_{\boldsymbol{\eta}} = \exp\left((\boldsymbol{T} - \boldsymbol{\eta}) \cdot \nabla\phi(\boldsymbol{\eta}) + \phi(\boldsymbol{\eta})\right). \tag{1}$$

While in the $\boldsymbol{\theta}$ parameters, the entropy is $H(p_{\boldsymbol{\theta}}) = \psi(\boldsymbol{\theta}) - \boldsymbol{\theta} \cdot \nabla\psi(\boldsymbol{\theta})$, in the $\boldsymbol{\eta}$ parameters, the ϕ function gives the negative entropy: $-H(p_{\boldsymbol{\eta}}) = \mathrm{E}_{p_{\boldsymbol{\eta}}}\left[\log p_{\boldsymbol{\eta}}\right] = \phi(\boldsymbol{\eta})$.

Proposition 1.

1. $\mathrm{Hess}\,\phi(\boldsymbol{\eta}) = (\mathrm{Hess}\,\psi(\boldsymbol{\theta}))^{-1}$ *when* $\boldsymbol{\eta} = \nabla\psi(\boldsymbol{\theta})$.
2. *The Fisher information matrix of the statistical model given by the exponential family in the* $\boldsymbol{\theta}$ *parameters is* $I_e(\boldsymbol{\theta}) = \mathrm{Cov}_{p_{\boldsymbol{\theta}}}\left(\nabla \log p_{\boldsymbol{\theta}}, \nabla \log p_{\boldsymbol{\theta}}\right) = \mathrm{Hess}\,\psi(\boldsymbol{\theta})$.
3. *The Fisher information matrix of the statistical model given by the exponential family in the* $\boldsymbol{\eta}$ *parameters is* $I_m(\boldsymbol{\theta}) = \mathrm{Cov}_{p_{\boldsymbol{\eta}}}\left(\nabla \log p_{\boldsymbol{\eta}}, \nabla \log p_{\boldsymbol{\eta}}\right) = \mathrm{Hess}\,\phi(\boldsymbol{\eta})$.

Proof. Derivation of the equality $\nabla\phi = (\nabla\psi)^{-1}$ gives the first item. The second item is a property of the cumulant generating function ψ. The third item follows from Equation (1). □

2.1. Statistical Manifold

The exponential family $\mathcal{E}$ is an elementary manifold in either the $\boldsymbol{\theta}$ or the $\boldsymbol{\eta}$ parameterization, named respectively exponential or mixture parameterization. We discuss now the proper definition of the tangent bundle $T\mathcal{E}$.

Definition 1 (Velocity)**.** *If* $I \ni t \mapsto p_t$, I *open interval, is a differentiable curve in* $\mathcal{E}$*, then its velocity vector is identified with its Fisher score:*

$$\frac{D}{dt}p(t) = \frac{d}{dt}\log\left(p_t\right) .$$

The capital D notation is taken from differential geometry; see the classical monograph [15].

Definition 2 (Tangent space)**.** *In the expression of the curve by the exponential parameters, the velocity is:*

$$\frac{D}{dt}p(t) = \frac{d}{dt}\log\left(p_t\right) = \frac{d}{dt}\left(\boldsymbol{\theta}(t) \cdot \boldsymbol{T} - \psi(\boldsymbol{\theta}(t))\right) = \dot{\boldsymbol{\theta}}(t) \cdot \left(\boldsymbol{T} - \mathrm{E}_{\boldsymbol{\theta}(t)}\left[\boldsymbol{T}\right]\right), \tag{2}$$

that is it equals the statistics whose coordinates are $\dot{\boldsymbol{\theta}}(t)$ *in the basis of the sufficient statistics centered at* p_t*. As a consequence, we identify the tangent space at each* $p \in \mathcal{E}$ *with the vector space of centered sufficient statistics, that is:*

$$T_p\mathcal{E} = \mathrm{Span}\left(T_j - \mathrm{E}_p\left[T_j\right] | j = 1, \ldots, d\right) .$$

In the mixture parameterization of Equation (1), the computation of the velocity is:

$$\frac{D}{dt}p(t) = \frac{d}{dt}\log(p_t) = \frac{d}{dt}\left(\nabla\phi(\boldsymbol{\eta}(t))\cdot(\boldsymbol{T}-\boldsymbol{\eta}(t)) + \phi(\boldsymbol{\eta}(t))\right) = \\ \left(\mathrm{Hess}\,\phi(\boldsymbol{\eta}(t))\dot{\boldsymbol{\eta}}(t)\right)\cdot(\boldsymbol{T}-\boldsymbol{\eta}(t)) = \dot{\boldsymbol{\eta}}(t)\cdot\left[\mathrm{Hess}\,\phi(\boldsymbol{\eta}(t))\,(\boldsymbol{T}-\boldsymbol{\eta}(t))\right]. \quad (3)$$

The last equality provides the interpretation of $\dot{\boldsymbol{\eta}}(t)$ as the coordinate of the velocity in the conjugate vector basis $\mathrm{Hess}\,\phi(\boldsymbol{\eta}(t))\,(\boldsymbol{T}-\boldsymbol{\eta}(t))$, that is the basis of velocities along the $\boldsymbol{\eta}$ coordinates.

In conclusion, the first order geometry is characterized as follows.

Definition 3 (Tangent bundle $T\mathcal{E}$)**.** *The tangent space at each $p \in \mathcal{E}$ is a vector space of random variables $T_p\mathcal{E} = \mathrm{Span}\,(T_j - \mathrm{E}_p[T_j] | j = 1,\ldots,d)$, and the tangent bundle $T\mathcal{E} = \{(p,V) | p \in \mathcal{E}, V \in T_p\mathcal{E}\}$, as a manifold, is defined by the chart:*

$$T\mathcal{E} \ni (\mathrm{e}^{\boldsymbol{\theta}\cdot\boldsymbol{T}-\psi(\boldsymbol{\theta})}, \boldsymbol{v}\cdot(\boldsymbol{T}-\mathrm{E}_{\boldsymbol{\theta}}[\boldsymbol{T}])) \mapsto (\boldsymbol{\theta},\boldsymbol{v}). \quad (4)$$

Proposition 2.

1. *If $V = \boldsymbol{v}\cdot(\boldsymbol{T}-\boldsymbol{\eta}) \in T_{p_{\boldsymbol{\eta}}}\mathcal{E}$, then V is represented in the conjugate basis as:*

$$V = \boldsymbol{v}\cdot(\boldsymbol{T}-\boldsymbol{\eta}) = \boldsymbol{v}\cdot(\mathrm{Hess}\,\phi(\boldsymbol{\eta}))^{-1}\,\mathrm{Hess}\,\phi(\boldsymbol{\eta})\,(\boldsymbol{T}-\boldsymbol{\eta}) = \\ \left((\mathrm{Hess}\,\phi(\boldsymbol{\eta}))^{-1}\boldsymbol{v}\right)\cdot\mathrm{Hess}\,\phi(\boldsymbol{\eta})\,(\boldsymbol{T}-\boldsymbol{\eta}). \quad (5)$$

2. *The mapping $(\mathrm{Hess}\,\phi(\boldsymbol{\eta}))^{-1}$ maps the coordinates $\boldsymbol{v}$ of a tangent vector $V \in T_{p_{\boldsymbol{\eta}}}\mathcal{E}$ with respect to the basis of centered sufficient statistics to the coordinates $\boldsymbol{v}^*$ with respect to the conjugate basis.*

3. *In the $\boldsymbol{\theta}$ parameters, the transformation is $\boldsymbol{v} \mapsto \boldsymbol{v}^* = \mathrm{Hess}\,\psi(\boldsymbol{\theta})\boldsymbol{v}$.*

Remark 1. *In the finite state space case, it is not necessary to go on to the formal construction of a dual tangent bundle, because all finite dimensional vector spaces are isomorphic. However, this step is compulsory in the infinite state space case, as was done in* [1]*. Moreover, the explicit construction of natural connections and natural parallel transports of the tangent and dual tangent bundle is unavoidable when considering the second-order calculus, as was done in* [1,8]*, in order to compute Hessians and implement Newton methods of optimization. However, the scope of the present paper is restricted to a basic study of gradient flows; hence, from now on, we focus on the Riemannian structure and disregard all second-order topics.*

Proposition 3 (Riemannian metric)**.** *The tangent bundle has a Riemannian structure with the natural scalar product of each $T_p\mathcal{E}$, $\langle V,W\rangle_p = \mathrm{E}_p[VW]$. In the basis of sufficient statistics, the metric is expressed by the Fisher information matrix $I(p) = \mathrm{Cov}_p(\boldsymbol{T},\boldsymbol{T})$, while in the conjugate basis, it is expressed by the inverse Fisher matrix $I^{-1}(p)$.*

Proof. In the basis of the sufficient statistics, $V = \boldsymbol{v}\cdot(\boldsymbol{T}-\mathrm{E}_p[\boldsymbol{T}])$, $W = \boldsymbol{w}\cdot(\boldsymbol{T}-\mathrm{E}_p[\boldsymbol{T}])$, so that:

$$\langle V,W\rangle_p = \boldsymbol{v}'\mathrm{E}_p\left[(\boldsymbol{T}-\mathrm{E}_p[\boldsymbol{T}])\,(\boldsymbol{T}-\mathrm{E}_p[\boldsymbol{T}])'\right]\boldsymbol{w} = \boldsymbol{v}'\,\mathrm{Cov}_p(\boldsymbol{T},\boldsymbol{T})\,\boldsymbol{w} = \boldsymbol{v}'I(p)\boldsymbol{w}\,, \quad (6)$$

where $I(p) = \mathrm{Cov}_p(\boldsymbol{T},\boldsymbol{T})$ is the Fisher information matrix.

If $p = p_{\boldsymbol{\theta}} = p_{\boldsymbol{\eta}}$, the conjugate basis at p is:

$$\operatorname{Hess}\phi(\boldsymbol{\eta})(\boldsymbol{T}-\boldsymbol{\eta}) = \operatorname{Hess}\psi(\boldsymbol{\theta})^{-1}(\boldsymbol{T}-\nabla\phi(\boldsymbol{\theta})) = I^{-1}(p)(\boldsymbol{T}-\mathrm{E}_p[\boldsymbol{T}]), \tag{7}$$

so that for elements of the tangent space expressed in the conjugate basis, we have $V = \boldsymbol{v}^* \cdot I^{-1}(p)\,(\boldsymbol{T}-\mathrm{E}_p[\boldsymbol{T}])$, $W = \boldsymbol{w}^* \cdot I^{-1}(p)\,(\boldsymbol{T}-\mathrm{E}_p[\boldsymbol{T}])$; thus:

$$\langle V, W\rangle_p = \boldsymbol{v}^{*\prime}\mathrm{E}_p\left[I^{-1}(p)\cdot(\boldsymbol{T}-\mathrm{E}_p[\boldsymbol{T}])\,(\boldsymbol{T}-\mathrm{E}_p[\boldsymbol{T}])'\,I^{-1}(p)\right]\boldsymbol{w}^* = \boldsymbol{v}^{*\prime}I^{-1}(p)\boldsymbol{w}^*. \tag{8}$$

□

2.2. Gradient

For each C^1 real function $F\colon \mathcal{E}\to\mathbb{R}$, its gradient is defined by taking the derivative along a C^1 curve $I\mapsto p(t)$, $p = p(0)$, and writing it with the Riemannian metrics,

$$\frac{d}{dt}\hat{F}(\boldsymbol{\theta}(t))\Big|_{t=0} = \left\langle \nabla F(p), \frac{D}{dt}p(t)\Big|_{t=0}\right\rangle_p, \quad \nabla F(p)\in T_p\mathcal{E}\,. \tag{9}$$

If $\boldsymbol{\theta}\mapsto\hat{F}(\boldsymbol{\theta})$ is the expression of F in the parameter $\boldsymbol{\theta}$ and $t\mapsto\boldsymbol{\theta}(t)$ is the expression of the curve, then $\frac{d}{dt}\hat{F}(\boldsymbol{\theta}(t)) = \nabla\hat{F}(\boldsymbol{\theta}(t))\cdot\dot{\boldsymbol{\theta}}(t)$, so that at $p = p_{\boldsymbol{\theta}(0)}$, with velocity $V = \frac{D}{dt}p(t)\big|_{t=0} = \dot{\boldsymbol{\theta}}(0)\cdot(\boldsymbol{T}-\nabla\psi(\boldsymbol{\theta}(0))$, so that we obtain the celebrated Amari's natural gradient of [16]:

$$\langle\nabla F(p), V\rangle_p = \left(\operatorname{Hess}\psi(\boldsymbol{\theta}(0))^{-1}\nabla\hat{F}(\boldsymbol{\theta}(0)\right)'\operatorname{Hess}\psi(\boldsymbol{\theta}(0))\dot{\boldsymbol{\theta}}(0)\,. \tag{10}$$

If $\boldsymbol{\eta}\mapsto\check{F}(\boldsymbol{\eta})$ is the expression of F in the parameter $\boldsymbol{\eta}$ and $t\mapsto\boldsymbol{\eta}(t)$ is the expression of the curve, then $\frac{d}{dt}\check{F}(\boldsymbol{\eta}(t)) = \nabla\check{F}(\boldsymbol{\eta}(t))\cdot\dot{\boldsymbol{\eta}}(t)$ so that at $p = p_{\boldsymbol{\eta}(0)}$, with velocity $V = \frac{d}{dt}\log(p(t))\big|_{t=0} = \dot{\boldsymbol{\eta}}(0)\cdot\operatorname{Hess}\phi(\boldsymbol{\eta}(0))(\boldsymbol{T}-\boldsymbol{\eta}(0))$,

$$\langle\nabla F(p), V\rangle_p = (\operatorname{Hess}\phi(\boldsymbol{\eta}(0))^{-1}\nabla\hat{F}(\boldsymbol{\eta}(0))'\operatorname{Hess}\phi(\boldsymbol{\eta}(0))\dot{\boldsymbol{\eta}}(0). \tag{11}$$

We summarize all notions of gradient in the following definition.

Definition 4 (Gradients)**.**

1. *The random variable $\nabla F(p)$ uniquely defined by Equation (9) is called the (geometric) gradient of F at p. The mapping $\nabla F\colon\mathcal{E}\ni p\mapsto\nabla F(p)$ is a vector field of $T\mathcal{E}$.*
2. *The vector $\widetilde{\nabla}\,\hat{F}(\boldsymbol{\theta}) = \operatorname{Hess}\psi(\boldsymbol{\theta})^{-1}\nabla\hat{F}(\boldsymbol{\theta})$ of Equation (10) is the expression of the geometric gradient in the $\boldsymbol{\theta}$ in the basis of sufficient statistics, and it is called the natural gradient, while $\nabla\hat{F}(\boldsymbol{\theta})$, which is the expression in the conjugate basis of the sufficient statistics, is called the vanilla gradient.*
3. *The vector $\widetilde{\nabla}\,\check{F}(\boldsymbol{\eta}) = \operatorname{Hess}\phi(\boldsymbol{\eta})^{-1}\nabla\check{F}(\boldsymbol{\eta})$ of Equation (10) is the expression of the geometric gradient in the $\boldsymbol{\eta}$ parameter and in the conjugate basis of sufficient statistics, and it is called the natural gradient, while $\nabla\check{F}(\boldsymbol{\eta})$, which is the expression in the basis of sufficient statistics, is called the vanilla gradient.*

Given a vector field of $\mathcal{E}$, *i.e.*, a mapping G defined on $\mathcal{E}$, such that $G(p) \in T_p\mathcal{E}$, which is called a section of the tangent bundle in the standard differential geometric language, an integral curve from p is a curve $I \ni t \mapsto p(t)$, such that $p(0) = p$ and $\frac{D}{dt}p(t) = G(p(t))$. In the $\boldsymbol{\theta}$ parameters, $G(p_{\boldsymbol{\theta}}) = \hat{\boldsymbol{G}}(\boldsymbol{\theta}) \cdot (\boldsymbol{T} - \nabla\psi(\boldsymbol{\theta}))$, so that the differential equation is expressed by $\dot{\boldsymbol{\theta}}(t) = \hat{\boldsymbol{G}}(\boldsymbol{\theta}(t))$. In the $\boldsymbol{\eta}$ parameters, $G(p_{\boldsymbol{\eta}}) = \check{\boldsymbol{G}}(\boldsymbol{\eta}) \cdot \operatorname{Hess}\phi(\boldsymbol{\eta})(\boldsymbol{T} - \boldsymbol{\eta})$, and the differential equation is $\dot{\boldsymbol{\eta}}(t) = \check{\boldsymbol{G}}(\boldsymbol{\eta}(t))$.

Definition 5 (Gradient flow)**.** *The gradient flow of the real function $F\colon \mathcal{E}$ is the flow of the differential equation $\frac{D}{dt}p(t) = \nabla F(p(t))$,* i.e., *$\frac{d}{dt}p(t) = p(t)\nabla F(p(t))$. The expression in the $\boldsymbol{\theta}$ parameters is $\dot{\boldsymbol{\theta}}(t) = \widetilde{\nabla}\,\hat{F}(\boldsymbol{\theta}(t))$, and the expression in the $\boldsymbol{\eta}$ parameters is $\dot{\boldsymbol{\eta}}(t) = \widetilde{\nabla}\,\check{F}(\boldsymbol{\eta}(t))$.*

The cases of gradient computation we have discussed above are just a special case of a generic argument. Let us briefly study the gradient flow in a general chart $f\colon \boldsymbol{\zeta} \mapsto p_{\boldsymbol{\zeta}}$. Consider the change of parametrization from $\boldsymbol{\zeta}$ to $\boldsymbol{\theta}$,

$$\boldsymbol{\zeta} \mapsto p_{\boldsymbol{\zeta}} \mapsto \boldsymbol{\theta}(p_{\boldsymbol{\zeta}}) = I(p_{\boldsymbol{\zeta}})^{-1}\operatorname{Cov}_{p_{\boldsymbol{\zeta}}}(\boldsymbol{T}, \log p_{\boldsymbol{\zeta}}) \ ,$$

and denote the Jacobian matrix of the parameters' change by $J(\boldsymbol{\zeta})$. We have:

$$\begin{aligned}\log p_{\boldsymbol{\zeta}} &= \boldsymbol{T}\cdot\boldsymbol{\theta}(\boldsymbol{\zeta}) - \psi(\boldsymbol{\theta}(\boldsymbol{\zeta})) \\ &= \boldsymbol{T}\cdot I(p_{\boldsymbol{\zeta}})^{-1}\operatorname{Cov}_{p_{\boldsymbol{\zeta}}}(\boldsymbol{T}, \log p_{\boldsymbol{\zeta}}) - \psi\left(I(p_{\boldsymbol{\zeta}})^{-1}\operatorname{Cov}_{p_{\boldsymbol{\zeta}}}(\boldsymbol{T}, \log p_{\boldsymbol{\zeta}})\right) \ ,\end{aligned}$$

and the $\boldsymbol{\zeta}$ coordinate basis of the tangent space $T_{p_{\boldsymbol{\zeta}}}\mathcal{E}$ consists of the components of the gradient with respect to $\boldsymbol{\zeta}$,

$$\nabla(\boldsymbol{\zeta} \mapsto \log p_{\boldsymbol{\zeta}}) = J^{-1}(\boldsymbol{\zeta})\left(\boldsymbol{T} - \mathrm{E}_{p_{\boldsymbol{\zeta}}}[\boldsymbol{T}]\right)$$

It should be noted that in this case, the expression of the Fisher information matrix does not have the form of a Hessian of a potential function. In fact, the case of the exponential and the mixture parameters point to a special structure, which is called the Hessian manifold; see [17].

2.3. Gradient Flow in the Mixture Geometry

From now on, we are going to focus on the expression of the gradient flow in the $\boldsymbol{\eta}$ parameters. From Definition 4, we have:

$$\widetilde{\nabla}\,\check{F}(\boldsymbol{\eta}) = \operatorname{Hess}\phi(\boldsymbol{\eta})^{-1}\nabla\check{F}(\boldsymbol{\eta}) = \operatorname{Hess}\psi(\nabla\phi(\boldsymbol{\eta}))\nabla\check{F}(\boldsymbol{\eta}) = I(p_{\boldsymbol{\eta}})\nabla\check{F}(\boldsymbol{\eta}) \ ,$$

where $I(p) = \operatorname{Cov}_p(\boldsymbol{T}, \boldsymbol{T})$. As $p \mapsto \operatorname{Cov}_p(\boldsymbol{T}, \boldsymbol{T})$ is the restriction to the simplex of a quadratic function, while $p \mapsto \boldsymbol{\eta}$ is the restriction to the exponential family $\mathcal{E}$ of a linear function, in some cases, we can naturally consider the extension of the gradient flow equation outside M°. One notable case is when the function F is a relaxation of a non-constant state space function $f\colon \Omega \to \mathbb{R}$, as it is defined in, e.g., [3].

Proposition 4. *Let $f\colon \Omega \to \mathbb{R}$, and let $F(p) = \mathrm{E}_p[f]$ be its relaxation on $p \in \mathcal{E}$. It follows:*

1. *$\nabla F(p)$ is the least square projection of f onto $T_p\mathcal{E}$, that is:*

$$\nabla F(p) = I(p)^{-1}\operatorname{Cov}_p(f, \boldsymbol{T})\cdot(\boldsymbol{T} - \mathrm{E}_p[\boldsymbol{T}]) \ .$$

2. *The expressions in the exponential parameters* $\boldsymbol{\theta}$ *are* $\widetilde{\nabla}\hat{F}(\boldsymbol{\theta}) = (\operatorname{Hess}\psi(\boldsymbol{\theta}))^{-1}\operatorname{Cov}_{\boldsymbol{\theta}}(f,\boldsymbol{T})$, $\nabla\hat{F}(\boldsymbol{\theta}) = \operatorname{Cov}_{\boldsymbol{\theta}}(f,\boldsymbol{T})$, *respectively.*

3. *The expressions in the mixture parameters* $\boldsymbol{\eta}$ *are* $\widetilde{\nabla}\check{F}(\boldsymbol{\eta}) = \operatorname{Cov}_{\boldsymbol{\eta}}(f,\boldsymbol{T})$ *and* $\nabla\check{F}(\boldsymbol{\eta}) = \operatorname{Hess}\phi(\boldsymbol{\eta})\operatorname{Cov}_{\boldsymbol{\eta}}(f,\boldsymbol{T})$, *respectively.*

Proof. On a generic curve through p with velocity V, we have $\frac{d}{dt}\mathrm{E}_{p(t)}[f]\big|_{t=0} = \operatorname{Cov}_p(f,V) = \langle f,V\rangle_p$. If $V \in T_p\mathcal{E}$, we can orthogonally project f to get $\langle\nabla F,V\rangle_p = \langle(I^{-1}(p)\operatorname{Cov}_p(f,\boldsymbol{T}))\cdot(\boldsymbol{T}-\mathrm{E}_p[\boldsymbol{T}]),V\rangle_p$. □

Remark 2. *Let us briefly recall the behavior of the gradient flow in the relaxation case. Let* $\boldsymbol{\theta}_n$, $n = 1,2,\ldots$, *be a minimizing sequence for* $\hat{F}$, *and let* $\bar{p}$ *be a limit point of the sequence* $(p_{\boldsymbol{\theta}_n})_n$. *It follows that* $\bar{p}$ *has a defective support, in particular* $\bar{p}\notin\mathcal{E}$; *see* [18,19]. *For a proof along lines coherent with the present paper, see* [20] *(Theorem 1). It is found that the support* $\underline{F}\subset\Omega$ *is exposed, that is* $\boldsymbol{T}(\underline{F})$ *is a face of the marginal polytope* $M = \operatorname{con}\{\boldsymbol{T}(\boldsymbol{x})|\boldsymbol{x}\in\Omega\}$. *In particular,* $\mathrm{E}_{\bar{p}}[\boldsymbol{T}] = \bar{\boldsymbol{\eta}}$ *belongs to a face of the marginal polytope* M. *If* $\boldsymbol{a}$ *is the (interior) orthogonal of the face, that is* $\boldsymbol{a}\cdot\boldsymbol{T}(\boldsymbol{x})+b\geq 0$ *for all* $\boldsymbol{x}\in\Omega$ *and* $\boldsymbol{a}\cdot\boldsymbol{T}(\boldsymbol{x})+b = 0$ *on the exposed set, then* $\boldsymbol{a}\cdot(\boldsymbol{T}(\boldsymbol{x})-\bar{\boldsymbol{\eta}}) = 0$ *on the face, so that* $\boldsymbol{a}\cdot\operatorname{Cov}_{\bar{p}}(f,\boldsymbol{T}) = 0$. *If we extend the mapping* $\boldsymbol{\eta}\mapsto\operatorname{Cov}_{\boldsymbol{\eta}}(f,\boldsymbol{T})$ *on the closed marginal polytope* M *to be the limit of the vector field of the gradient on the faces of the marginal polytope, we expect to see that such a vector field is tangent to the faces. This remark is further elaborated below in the binary case.*

2.4. The Saturated Model

A case of special tutorial interest is obtained when the exponential family contains all probability densities, that is when $\mathcal{E} = \mathcal{P}_>$. This case has been treated by many authors; here, we use the presentation of [21].

It is convenient to recode the sample space as $\Omega = \{0,\ldots,d\}$, where $\boldsymbol{x} = 0$ is a distinguished point. If X is the identity on Ω, we define the sufficient statistics to be the indicator functions of points $T_j = (X = j)$, $j = 1,\ldots,d$. The saturated exponential family consists of all of the positive densities written as:

$$p(\boldsymbol{x};\boldsymbol{\theta}) = \exp\left(\sum_{j=1}^{d}\theta_j(X=j)-\psi(\boldsymbol{\theta})\right),$$

where:

$$\psi(\boldsymbol{\theta}) = \log\left(1+\sum_{j=1}^{d}\mathrm{e}^{\theta_j}\right).$$

Note that, in this case, the expectation parameter $\eta_j = \mathrm{E}((X=j))$ is the probability of case $\boldsymbol{x} = j$ and the marginal polytope is the probability simplex Δ_d.

The gradient mapping is:

$$\boldsymbol{\eta} = \nabla\psi(\boldsymbol{\theta}) = \left(\frac{\mathrm{e}^{\theta_j}}{1+\sum_{i=1}^{d}\mathrm{e}^{\theta_i}}\middle| j=1,\ldots,d\right),$$

the inverse gradient mapping is defined for $\boldsymbol{\eta}\in]0,1[^d$ by:

$$\boldsymbol{\theta} = (\nabla\psi)^{-1}(\boldsymbol{\eta}) = \nabla\phi(\boldsymbol{\eta}) = \left(\log\left(\frac{\eta_j}{1-\sum_{i=1}^{d}\eta_i}\right)\middle| j=1,\ldots,d\right),$$

the negative entropy (Legendre conjugate) is:

$$\phi(\boldsymbol{\eta}) = \boldsymbol{\eta} \cdot \nabla\phi(\boldsymbol{\eta}) - \psi \circ \nabla\phi(\boldsymbol{\eta}) = \sum_{j=1}^{d} \eta_j \log\left(\frac{\eta_j}{1-\sum_{i=1}^{d}\eta_i}\right) + \log\left(1-\sum_{i=1}^{d}\eta_i\right) ,$$

the $\boldsymbol{\eta}$ parameterization (1) of the probability is:

$$\begin{aligned} p_{\boldsymbol{\eta}} &= \exp\left((\boldsymbol{T}-\boldsymbol{\eta})\cdot\nabla\phi(\boldsymbol{\eta}) + \phi(\boldsymbol{\eta})\right) = \\ &\exp\left(\sum_{j=1}^{d}((X=j)-\eta_j)\log\left(\frac{\eta_j}{1-\sum_{i=1}^{d}\eta_i}\right) + \sum_{j=1}^{d}\eta_j\log\left(\frac{\eta_j}{1-\sum_{i=1}^{d}\eta_i}\right) + \log\left(1-\sum_{i=1}^{d}\eta_i\right)\right) = \\ &\exp\left(\sum_{j=1}^{d}(X=j)\log\left(\frac{\eta_j}{1-\sum_{i=1}^{d}\eta_i}\right) + \log\left(1-\sum_{i=1}^{d}\eta_i\right)\right) = \\ &\prod_{j=1}^{d}\left(\frac{\eta_j}{1-\sum_{i=1}^{d}\eta_i}\right)^{(X=j)}\left(1-\sum_{i=1}^{d}\eta_i\right) = \left(1-\sum_{i=1}^{d}\eta_i\right)^{(X=0)}\prod_{j=1}^{d}\eta_j^{(X=j)} . \end{aligned}$$

Remark 3. *The previous equation prompts three crucial remarks:*

1. *The expression of the probability in the $\boldsymbol{\eta}$ parameters is a normalized monomial in the parameters.*
2. *The expression continuously extends the exponential family to the probabilities in $\mathcal{P}_{\geq}$.*
3. *The expression actually is a polynomial parameterization of the signed densities $\mathcal{P}^1$.*

We proceed to approach the three issues above. The Hessian functions are:

$$\operatorname{Hess}\psi(\boldsymbol{\theta}) = \operatorname{diag}(\boldsymbol{p}) - \boldsymbol{p}\otimes\boldsymbol{p}, \quad \boldsymbol{p} = (1-\sum_{j=1}^{d}\mathrm{e}^{\theta_j})^{-1}\mathrm{e}^{\boldsymbol{p}} ,$$

$$\operatorname{Hess}\phi(\boldsymbol{\eta}) = \operatorname{diag}(\boldsymbol{\eta})^{-1} - \eta_0^{-1}[1]_{i,j=1}^{d}, \quad \eta_0 = 1-\sum_{j=1}^{d}\eta_j .$$

The matrix $\operatorname{Hess}\psi(\boldsymbol{\theta})$ is the Fisher information matrix $I(p)$ of the exponential family at $\boldsymbol{p} = p_{\boldsymbol{\theta}}$, and the matrix $\operatorname{Hess}\phi(\boldsymbol{\eta})$ is the inverse Fisher information matrix $I^{-1}(p)$ at $\boldsymbol{p} = p_{\boldsymbol{\eta}}$. It follows that the natural gradient of a function $\boldsymbol{\eta} \mapsto h(\boldsymbol{\eta})$ will be:

$$\widetilde{\nabla}\, h(\boldsymbol{\eta}) = \operatorname{Hess}\phi(\boldsymbol{\eta})\nabla h(\boldsymbol{\eta}) ,$$

whose behavior depends on the following theorem; see [21] (Proposition 3).

Proposition 5.

1. *The inverse Fisher information matrix $I(p)^{-1}$ is zero on the vertexes of the simplex, only.*
2. *The determinant of the inverse Fisher information matrix $I(p)^{-1}$ is:*

$$\det\left(I(p)^{-1}\right) = \left(1-\sum_{i=1}^{n}p_i\right)\prod_{i=1}^{n}p_i .$$

3. *The determinant of the inverse Fisher information matrix $I(p)^{-1}$ is zero on the borders of the simplex, only.*

4. *On the interior of each facet, the rank of the inverse Fisher information matrix $I(p)^{-1}$ is $(n-1)$, and the $(n-1)$ linear independent column vectors generate the subspace parallel to the facet itself.*

A generic statistical model can be seen as a submanifold of the saturated model, so that the form of the gradient in the submanifold is derived according to the general results in differential geometry. We do not do that here, and we switch to some very specific examples.

3. Toric Models: A Tutorial Example

Exponential families whose sample space is an integer lattice, such as finite subsets of $\mathbb{Z}^2$ or $\{+1,-1\}^d$, have special algebro-combinatorial features that fall under the name of algebraic statistics. Seminal papers have been [22,23]. Monographs on the topic are [24–26]. The book [27] covers both information geometry and algebraic statistics.

We do not assume the reader has detailed information about algebraic statistics. In this section, we work on a toy example intended to show both the basic mechanism of algebraic statistics and how the algebraic concepts are applied to the gradient flow problem as it was described in the previous section.

First, we give a general definition of the object on which we focus. A toric model is an exponential family, such that the orthogonal space of the space generated by the sufficient statistics and the constant has a vector basis of integer-valued random variables. We consider this example:

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

, (12)

which corresponds to a variation of the classical independence model, where the design corresponds to the vertices of a square. It this example we moved the point $\{4\}$ from $(1,1)$ to $(2,1)$.

In Equation (12), T_1 and T_2 are the sufficient statistics of the exponential family:

$$p_{\boldsymbol{\theta}} = \exp\left(\theta_1 T_1 + \theta_2 T_2 - \psi(\boldsymbol{\theta})\right), \quad \psi(\boldsymbol{\theta}) = \log\left(1 + \mathrm{e}^{\theta_2} + \mathrm{e}^{\theta_1} + \mathrm{e}^{2\theta_1+\theta_2}\right), \tag{13}$$

T_3 is an integer-valued vector basis of the orthogonal space $\mathrm{Span}\,(\mathbf{1}, T_1, T_2)^{\perp}$.

For the purpose of the generalization to less trivial examples, it should be noted that $T_3 = T_3^+ - T_3^-$, that is $(-2,1,2,-1) = (0,1,2,0) - (2,0,0,1)$. The couple (T_3^+, T_3^-) connects the lattice defined by:

$$\mathcal{L} = \left\{(Y,Z) \in \mathbb{Z}^4_{\geq} \times \mathbb{Z}^4_{\geq} \middle| B^T y = B^T Z\right\}, \quad B = \begin{bmatrix} \mathbf{1} & T_1 & T_2 \end{bmatrix}.$$

Such a set of generators is called a Markov basis of the lattice; see [22]. Algorithms are available to compute such a set of generators and are implemented, for instance, in the software suite 4ti2; see [28].

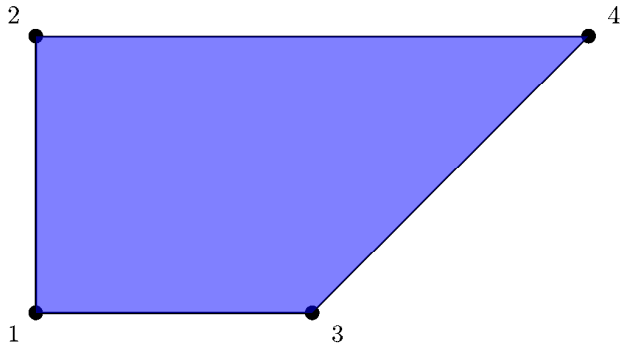


Figure 1. Marginal polytope of the exponential family in Equations (12) and (13). The coordinates of the vertices are given by (T_1, T_2).

The sample space can be identified with the value of the sufficient statistics, hence with a finite subset of $\mathbb{Q}^2 \supset \Omega$, $\Omega = \{(0,0),(0,1),(1,0),(2,1)\}$; see Figure 1. Given a finite subset of $\mathbb{R}^d$, it is a general algebraic fact that there exists a filtering set of monomial functions that is a vector basis of all real functions on the subset itself; see an exposition and the applications to statistics in [24] or [27]. In our case, the monomial basis is $1, T_1, T_2, T_1T_2$, and we define the matrix of the saturated model to be:

$$A = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} \mathbf{1} & T_1 & T_2 & T_1T_2 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 2 & 1 & 2 \end{bmatrix} \end{array}, \quad A^{-1} = \frac{1}{2}\begin{bmatrix} 2 & 0 & 0 & 0 \\ -2 & 0 & 2 & 0 \\ -2 & 2 & 0 & 0 \\ 2 & -1 & -2 & 1 \end{bmatrix}. \tag{14}$$

The matrix A one-to-one maps probabilities into expected values,

$$\begin{bmatrix} 1 & \eta_1 & \eta_2 & \eta_{12} \end{bmatrix} = \begin{bmatrix} 1 & \mathbb{E}[T_1] & \mathbb{E}[T_2] & \mathbb{E}[T_1T_2] \end{bmatrix} = \begin{bmatrix} p_1 & p_2 & p_3 & p_4 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 2 & 1 & 2 \end{bmatrix}, \tag{15}$$

and *vice versa*,

$$\begin{bmatrix} p_1 & p_2 & p_3 & p_4 \end{bmatrix} = \begin{bmatrix} 1 & \eta_1 & \eta_2 & \eta_{12} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 1 & 0 & 0 \\ 1 & -\frac{1}{2} & -1 & \frac{1}{2} \end{bmatrix}. \tag{16}$$

On Model (13), the (positive) probabilities are constrained by the model:

Ω	$p_{\boldsymbol{\theta}}$	$\exp\left(\theta_1T_1 + \theta_2T_2 - \log\left(1 + \mathrm{e}^{\theta_2} + \mathrm{e}^{\theta_1} + \mathrm{e}^{2\theta_1+\theta_2}\right)\right)$
1	$p(1;\boldsymbol{\theta})$	$\exp\left(-\log\left(1 + \mathrm{e}^{\theta_2} + \mathrm{e}^{\theta_1} + \mathrm{e}^{2\theta_1+\theta_2}\right)\right)$
2	$p(2;\boldsymbol{\theta})$	$\exp\left(\theta_2 - \log\left(1 + \mathrm{e}^{\theta_2} + \mathrm{e}^{\theta_1} + \mathrm{e}^{2\theta_1+\theta_2}\right)\right)$
3	$p(3;\boldsymbol{\theta})$	$\exp\left(\theta_1 - \log\left(1 + \mathrm{e}^{\theta_2} + \mathrm{e}^{\theta_1} + \mathrm{e}^{2\theta_1+\theta_2}\right)\right)$
4	$p(4;\boldsymbol{\theta})$	$\exp\left(2\theta_1 + \theta_2 - \log\left(1 + \mathrm{e}^{\theta_2} + \mathrm{e}^{\theta_1} + \mathrm{e}^{2\theta_1+\theta_2}\right)\right)$

. (17)

If we introduce the parameters $\zeta_1 = \exp(\theta_1)$, $\zeta_2 = \exp(\theta_2)$, the model is shown to be a (piece of an) algebraic variety, that is a set described by the rational parametric equations:

Ω	$p_{\boldsymbol{\zeta}}$	$\zeta^{T_1}\zeta^{T_2}/(1+\zeta_2+\zeta_1+\zeta_1^2\zeta_2)$
1	$p(1;\boldsymbol{\zeta})$	$1/(1+\zeta_2+\zeta_1+\zeta_1^2\zeta_2)$
2	$p(2;\boldsymbol{\zeta})$	$\zeta_2/(1+\zeta_2+\zeta_1+\zeta_1^2\zeta_2)$
3	$p(3;\boldsymbol{\zeta})$	$\zeta_1/(1+\zeta_2+\zeta_1+\zeta_1^2\zeta_2)$
4	$p(4;\boldsymbol{\zeta})$	$\zeta_1^2\zeta_2/(1+\zeta_2+\zeta_1+\zeta_1^2\zeta_2)$

. (18)

The peculiar structure of the toric model is best seen by considering the unnormalized probabilities:

Ω	$q_{\boldsymbol{\zeta}}$	$\zeta^{T_1}\zeta^{T_2}$
1	$q(1;\boldsymbol{\zeta})$	1
2	$q(2;\boldsymbol{\zeta})$	ζ_2
3	$q(3;\boldsymbol{\zeta})$	ζ_1
4	$q(4;\boldsymbol{\zeta})$	$\zeta_1^2\zeta_2$

, $\quad p(\boldsymbol{x};\boldsymbol{\zeta}) = \dfrac{q(\boldsymbol{x};\boldsymbol{\zeta})}{1+\zeta_2+\zeta_1+\zeta_1^2\zeta_2}$. (19)

In algebraic terms, the homogeneous coordinates $[q_1 : q_2 : q_3 : q_4]$ belong to the projective space $\boldsymbol{P}^3$. Precisely, the (real) projective space $\boldsymbol{P}^3$ is the set of all non-zero points of $\mathbb{R}^4$ together with the equivalence relation $[q_1 : q_2 : q_3 : q_4] = [\bar{q}_1 : \bar{q}_2 : \bar{q}_3 : \bar{q}_4]$ if, and only if, $[q_1, q_2, q_3, q_4] = k[\bar{q}_1, \bar{q}_2, \bar{q}_3, \bar{q}_4]$, $k \neq 0$. The domain of unnormalized signed probabilities as projective points is the open subset $\mathbb{P}^3_*$ of $\mathbb{P}^3$ where $q_1 + q_2 + q_3 + q_4 \neq 0$. On this set, we can compute the normalization:

$$\mathbb{P}^3_* \ni [q_1 : q_2 : q_3 : q_4] \mapsto [q_1, q_2, q_3, q_4]/(q_1 + q_2 + q_3 + q_4) \in {}^*\mathcal{E}\ ,$$

where ${}^*\mathcal{E}$ is the affine space generated by the simplex Δ_3. Notice that this embedding produces a number of natural geometrical structures on ${}^*\mathcal{E}$.

Because of the form of (13), a positive density p belongs to that family if, and only if, $\log p \in \mathrm{Span}(1, T_1, T_2)$, which, in turn, is equivalent to $\log p \perp T_3$. We can rewrite the orthogonality as:

$$\begin{aligned} 0 = \sum_{\boldsymbol{x}\in\Omega} \log p(\boldsymbol{x}) T_3(\boldsymbol{x}) &= \sum_{\boldsymbol{x}:\, T_3(\boldsymbol{x})>0} \log p(\boldsymbol{x}) T_3^+(\boldsymbol{x}) - \sum_{\boldsymbol{x}:\, T_3(\boldsymbol{x})<0} \log p(\boldsymbol{x}) T_3^-(\boldsymbol{x}) \\ &= \log\left(\prod_{\boldsymbol{x}:\, T_3(\boldsymbol{x})>0} p(\boldsymbol{x})^{T_3^+(\boldsymbol{x})}\right) - \log\left(\prod_{\boldsymbol{x}:\, T_3(\boldsymbol{x})<0} p(\boldsymbol{x})^{T_3^-(\boldsymbol{x})}\right)\ . \end{aligned}$$

Dropping the $\log$ function in the last expression, we observe that the positive probabilities described by either Equation (17) with $\theta_1, \theta_2 \in \mathbb{R}$ or Equation (18) with $\zeta_1, \zeta_2 \in \mathbb{R}_>$ are equivalently described by the equations:

$$p_1 + p_2 + p_3 + p_4 - 1 = 0\ , \tag{20}$$

$$p_1^2 p_4 - p_2 p_3^2 = 0\ . \tag{21}$$

Equation (21) identifies a surface within the probability simplex Δ_3, which is represented in Figure 2 by the triangularization of a grid of points that satisfy the invariant.

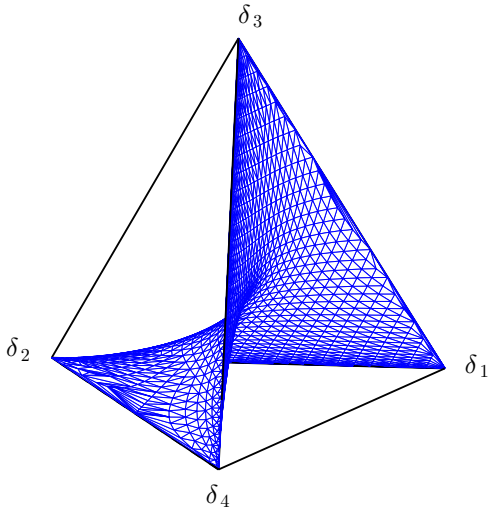


Figure 2. Representation of the exponential family in Equations (12) and (13) as a surface that intersects the probability simplex Δ_3. The surface is obtained by the triangularization of a grid of points that satisfy the invariant in Equation (21).

By choosing a basis for the space orthogonal to $\mathrm{Span}\,(\mathbf{1}, T_1, T_2)^{\perp}$, we can embed the marginal polytope of Figure 1 into the associated full marginal polytope. By expressing probabilities as a function of the expectation parameters, Equation (21) identifies a relationship between η_1, η_2 and the expected values of the chosen basis for the orthogonal space. This corresponds to an equivalent invariant in the expectation parameters, which, in turn, identifies a surface in the full marginal polytope.

For instance, consider the full marginal polytope parametrized by $\boldsymbol{\eta} = (\eta_1, \eta_2, \eta_3)$, with $\eta_3 = \mathbb{E}[T_3]$, which corresponds to the choice of T_3 as a basis for the space orthogonal to the span of the sufficient statistics of the model, together with the constant $\mathbf{1}$, as in Equation (12). We introduce the following matrix:

$$B = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} \mathbf{1} & T_1 & T_2 & T_3 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & -2 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 2 \\ 1 & 2 & 1 & -1 \end{bmatrix} \end{array}, \tag{22}$$

and similarly to Equation (15), we use the B matrix to one-to-one map probabilities into expected values, that is:

$$\begin{bmatrix} 1 \\ \eta_1 \\ \eta_2 \\ \eta_3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 2 \\ 0 & 1 & 0 & 1 \\ -2 & 1 & 2 & -1 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{bmatrix}, \tag{23}$$

and:

$$\begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{bmatrix} = \begin{bmatrix} \frac{3}{5} & -\frac{1}{5} & -\frac{2}{5} & -\frac{1}{5} \\ \frac{1}{5} & -\frac{2}{5} & \frac{7}{10} & \frac{1}{10} \\ \frac{2}{5} & \frac{1}{5} & -\frac{3}{5} & \frac{1}{5} \\ -\frac{1}{5} & \frac{2}{5} & \frac{3}{10} & -\frac{1}{10} \end{bmatrix} \begin{bmatrix} 1 \\ \eta_1 \\ \eta_2 \\ \eta_3 \end{bmatrix}. \tag{24}$$

Then, by expressing probabilities as a function of the expectation parameters in Equation (21), we obtain the following invariant in $\boldsymbol{\eta}$ associated with the model:

$$(4\eta_1 + 3\eta_2 - \eta_3 - 2)(\eta_1 + 2\eta_2 + \eta_3 - 3)^2 + (4\eta_1 - 7\eta_2 - \eta_3 - 2)(\eta_1 - 3\eta_2 + \eta_3 + 2)^2 = 0 \,. \quad (25)$$

From the linear relationship between probabilities and expectation probabilities, we know that on the interior of the full marginal polytope, there exists a unique η_3 which can be computed as a function of the other expectation parameters. Solving Equation (25) for η_3 allows one to express explicitly the value of η_3 given (η_1, η_2) and represent the surface associated with the invariant in the full marginal polytope. However, the cubic polynomial in Equation (25) in general admits three roots. The unique value of η_3 can be obtained from the roots of the cubic polynomial, by imposing that η_3 must be real and belong to the full marginal polytope given by $\mathrm{Conv}\,\{(T_1(\boldsymbol{x}), T_2(\boldsymbol{x}), T_3(\boldsymbol{x}))|\boldsymbol{x} \in \Omega\}$.

We remind that the determinant Δ associated with the cubic function in Equation (25) in the η_3 variable:

$$a\eta_3^3 + b\eta_3^2 + c\eta_3 + d = 0 \,, \quad (26)$$

with:

$$a = 1 \quad (27)$$

$$b = -2\eta_1 + \eta_2 + 1 \quad (28)$$

$$c = -(4\eta_1 + 3\eta_2 - 2)(\eta_1 + 2\eta_2 - 3) + \frac{1}{2}(\eta_1 + 2\eta_2 - 3)^2 - (4\eta_1 - 7\eta_2 - 2)(\eta_1 - 3\eta_2 + 2) + \\ + \frac{1}{2}(\eta_1 - 3\eta_2 + 2)^2 \quad (29)$$

$$d = -\frac{1}{2}(4\eta_1 + 3\eta_2 - 2)(\eta_1 + 2\eta_2 - 3)^2 - \frac{1}{2}(4\eta_1 - 7\eta_2 - 2)(\eta_1 - 3\eta_2 + 2)^2 \quad (30)$$

is given by:

$$\Delta = 18abcd - 4b^3d + b^2c^2 - 4ac^3 - 27a^2d^2 \,. \quad (31)$$

For $\Delta = 0$, the polynomial has a real root with multiplicity equal to three; for $\Delta < 0$, we have one real root and two complex conjugates roots, while for $\Delta > 0$, there exist three real roots. The three roots of the polynomial as a function of the coefficients are given by:

$$\eta_{3,k} = -\frac{1}{3}\left(b + u_k C + \frac{\Delta_0}{u_k C}\right) \,, \quad (32)$$

for $k \in \{1, 2, 3\}$, with:

$$u_1 = 1 \,, \quad (33)$$

$$u_2 = \frac{-1 + i\sqrt{3}}{2} \,, \quad (34)$$

$$u_3 = \frac{-1 - i\sqrt{3}}{2} \,, \quad (35)$$

and:

$$C = \sqrt[3]{\frac{\Delta_1 + \sqrt{(}\Delta_1^2 - 4\Delta_0^3)}{2}} \,, \quad (36)$$

$$\Delta_0 = b^2 - 3ac \,, \quad (37)$$

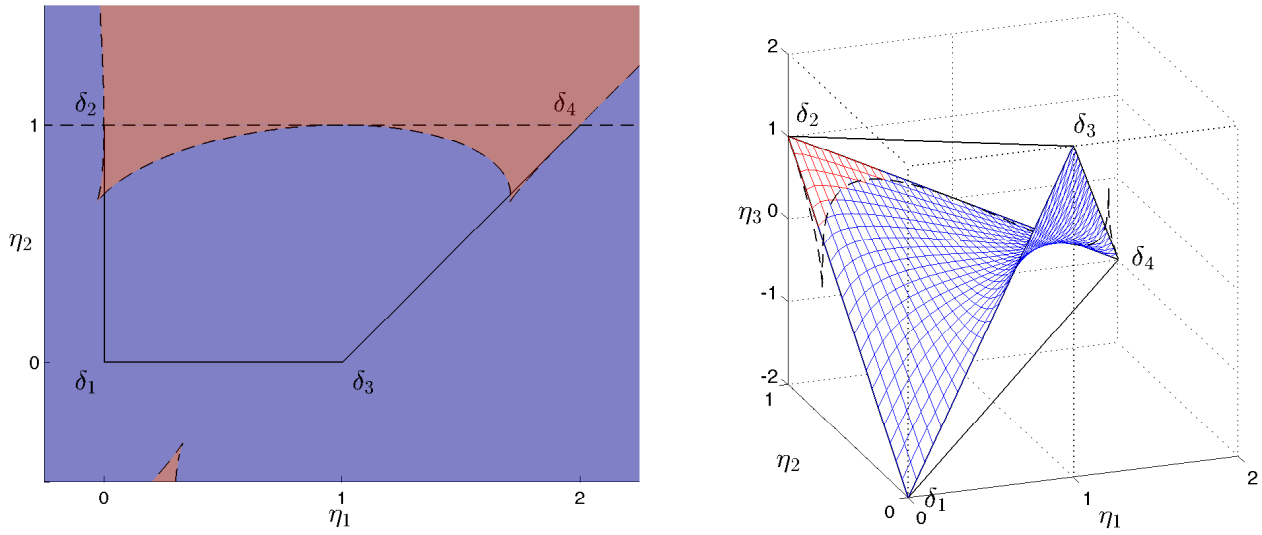
$$\Delta_1 = 2b^3 + 9abc + 27a^2d \,. \quad (38)$$

For the cubic polynomial in η_3 of Equation (25), $\Delta < 0$ for $\eta_2 - 1 \neq 0$ and for:

$$4\eta_1^4 - 8\eta_1^3\eta_2 + 24\eta_1^2\eta_2^2 - 20\eta_1\eta_2^3 - 2\eta_2^4 - 8\eta_1^3 - 12\eta_1^2\eta_2 + 4\eta_2^3 + 8\eta_1^2 + 16\eta_1\eta_2 - \eta_2^2 - 4\eta_1 - 2\eta_2 + 1 > 0\,. \quad (39)$$

In Figure 3(a), we represent in blue the region of the space (η_1, η_2) where $\Delta < 0$, in red where $\Delta > 0$, and the points where $\Delta = 0$ with a dashed line. For $\Delta < 0$, the only real root is $\eta_{3,1}$, which identifies the blue surface in the full marginal polytope in Figure 3(b). For $\Delta > 0$, it is easy to verify that only $\eta_{3,2}$ belongs to the interior of the full marginal polytope parametrized by (η_1, η_2, η_3), since it satisfies the inequalities given by the facets of the marginal polytope, and is represented in Figure 3(b) by the red surface. Finally, the three real roots coincide for $\Delta = 0$, that is, for $\eta_2 = 1$, and where:

$$4\eta_1^4 - 8\eta_1^3\eta_2 + 24\eta_1^2\eta_2^2 - 20\eta_1\eta_2^3 - 2\eta_2^4 - 8\eta_1^3 - 12\eta_1^2\eta_2 + 4\eta_2^3 + 8\eta_1^2 + 16\eta_1\eta_2 - \eta_2^2 - 4\eta_1 - 2\eta_2 + 1 = 0\,. \quad (40)$$



(a) (b)

Figure 3. Marginal polytope of the exponential family in Equations (12) and (13) (**a**). The dashed lines correspond to the points where $\Delta = 0$, where Δ is the discriminant in Equation (31); over the red regions $\Delta > 0$ and over the blue regions $\Delta < 0$. Representation of the exponential family as a surface in the full marginal polytope parametrized by (η_1, η_2, η_3) (**b**). The blue surface is given by the unique real root $\eta_{3,1}$ in Equation (32); the red surface corresponds to the unique real root $\eta_{3,2}$, which belongs to the full marginal polytope; over the dashed lines, which have been computed solving Equation (40) numerically, Equation (26) admits a real root with multiplicity equal to three.

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

$$\mathcal{I} = \left\langle p_1 + p_2 + p_3 + p_4 - 1, p_1^2 p_4 - p_2 p_3^2 \right\rangle \quad (41)$$

consists of all the polynomials of the form:

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

The algebraic variety of $\mathcal{I}$ uniquely extends the exponential family outside the positive octant. In the language of commutative algebra, it is the real Zariski closure of the exponential family model, *cf.* [29]. It is a notable example of toric variety. The general theory is in the monograph [30], and the applications to statistical models were first discussed in [31,32].

Let us discuss in some detail the parameterization of the toric variety as the submanifold of $\mathbb{R}^4$ defined by Equations (20) and (21). 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,

$$2p_1p_4 = -p_3^2 = -2p_2p_3 = p_1^2 .$$

This is equivalent to $p_1^2 = p_3^2 = 0$, which is a subspace of dimension two, whose intersection with Equation (20), is a line $\mathcal{C}$ in the affine space ${}^*\mathcal{E} = \{\boldsymbol{p} \in \mathbb{R}^4 | p_1 + p_2 + p_3 + p_4 = 1\}$. This (double) critical line intersects the simplex along the edge $\delta_2 \leftrightarrow \delta_4$. Outside $\mathcal{C}$, that is in the open complement set, the equations of the toric variety are locally solvable in two among the p_i's under the condition that the corresponding minor is not zero. To have a picture of what this critical set looks like, let us intersect our surface with the plane $p_3 = 0$. On the affine space $p_1 + p_2 + p_4 = 1$, we have $p_1^2 p_4 = 0$, that is the union of the double line $p_1^2 = 0$ with the line $p_4 = 0$.

In the following, we derive a parameterization based on an algebraic argument, the Bézout theorem. In fact, it is remarkable that the cubic surface defined by Equations (20) and (21) is a well known example of ruled surface, see Exercise 5.8.15 in [33]. In fact, the singular line is a double line, so that the intersection of the cubic surface with any plane through the singular line is of degree $1 = 3 - 2$, by the Bézout theorem, and thus, it is a line.

The line $\mathcal{C}$ is said to be double because the polynomial $p_1^2 p_4 - p_2 p_3^2$ belongs to the ideal generated by p_1^2 and p_3^2. Let us consider the sheaf of planes through the singular line defined for each $[\alpha : \beta] \in \boldsymbol{P}^1$ by the equations:

$$\mathcal{P}[\alpha : \beta] = \{p_1 + p_2 + p_3 + p_4 - 1 = 0, \alpha p_1 + \beta p_3 = 0\} .$$

Let us intersect each plane $\mathcal{P}[\alpha : \beta]$ of the sheaf with the model variety $\mathcal{M}$ by solving the system of equations:

$$\begin{cases} p_1 + p_2 + p_3 + p_4 & = 1 \\ p_1^2 p_4 - p_2 p_3^2 & = 0 \\ \alpha p_1 + \beta p_3 & = 0 \end{cases} . \tag{42}$$

On the critical line $\mathcal{C}$, a generic point is parameterized as $\boldsymbol{p}(\tau, 0) = (0, \tau, 0, 1 - \tau)$, which satisfies Equation (42) for $\tau \in \mathbb{R}$. If $0 \leq \tau \leq 1$, then $\boldsymbol{p}(\tau, 0)$ belongs to the edge $\delta_2 \leftrightarrow \delta_4$.

As the critical line is double and the intersection of the model variety with the plane of the sheaf is a cubic curve, we expect the remaining part to be of degree $3 - 2 = 1$, that is to be a line. Assume first $\alpha, \beta \neq 0$. Outside the critical line, as p_1, p_3 are not both zero and $\alpha p_1 + \beta p_3 = 0$, then $\alpha p_1 = -\beta p_3 \neq 0$. It follows $(\alpha p_1)^2 = (\beta p_3)^2 \neq 0$; hence:

$$p_1^2 p_4 - p_2 p_3^2 = 0 \Rightarrow \beta^2 (\alpha p_1)^2 p_4 - \alpha^2 p_2 (\beta p_3)^2 = 0 \Rightarrow \beta^2 p_4 - \alpha^2 p_2 = 0 .$$

We have found that for $\alpha, \beta \neq 0$, the intersection between the plane $\mathcal{P}[\alpha : \beta]$ and the model variety $\mathcal{M}$ is the union of the critical line $\mathcal{C}$ and the line of equations:

$$\begin{cases} p_1 + p_2 + p_3 + p_4 & = 1 \\ \alpha p_1 + \beta p_3 & = 0 \\ -\alpha^2 p_2 + \beta^2 p_4 & = 0 \end{cases} . \tag{43}$$

This line intersects the critical line where:

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

that is in the point:

$$\boldsymbol{p}([\alpha : \beta], 0)) = \left(0, \frac{\beta^2}{\alpha^2 + \beta^2}, 0, \frac{\alpha^2}{\alpha^2 + \beta^2}\right) .$$

In parametric form, the line in Equations (43) is:

$$\boldsymbol{p}([\alpha : \beta], t) = \boldsymbol{p}([\alpha : \beta], 0) + \boldsymbol{u}t \, ,$$

with $\boldsymbol{u} = \left(\beta, \frac{\beta^2(\alpha-\beta)}{\alpha^2+\beta^2}, -\alpha, \frac{\alpha^2(\alpha-\beta)}{\alpha^2+\beta^2}\right)$,

$$\begin{aligned} p_1([\alpha : \beta], t) &= \beta t \\ p_2([\alpha : \beta], t) &= \frac{\beta^2}{\alpha^2 + \beta^2} + \frac{\beta^2(\alpha - \beta)}{\alpha^2 + \beta^2} t \\ p_3([\alpha : \beta], t) &= -\alpha t \\ p_4([\alpha : \beta], t) &= \frac{\alpha^2}{\alpha^2 + \beta^2} + \frac{\alpha^2(\alpha - \beta)}{\alpha^2 + \beta^2} t \end{aligned} . \tag{44}$$

The same equations hold in the previously excluded case $\alpha\beta = 0$.

Positive values of components 1 and 3 of the probability are obtained in Equation (44) for $\alpha\beta < 0$ and $\beta t > 0$, say $\alpha < 0$, $\beta > 0$, $t > 0$. In this case, we have for component 2:

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

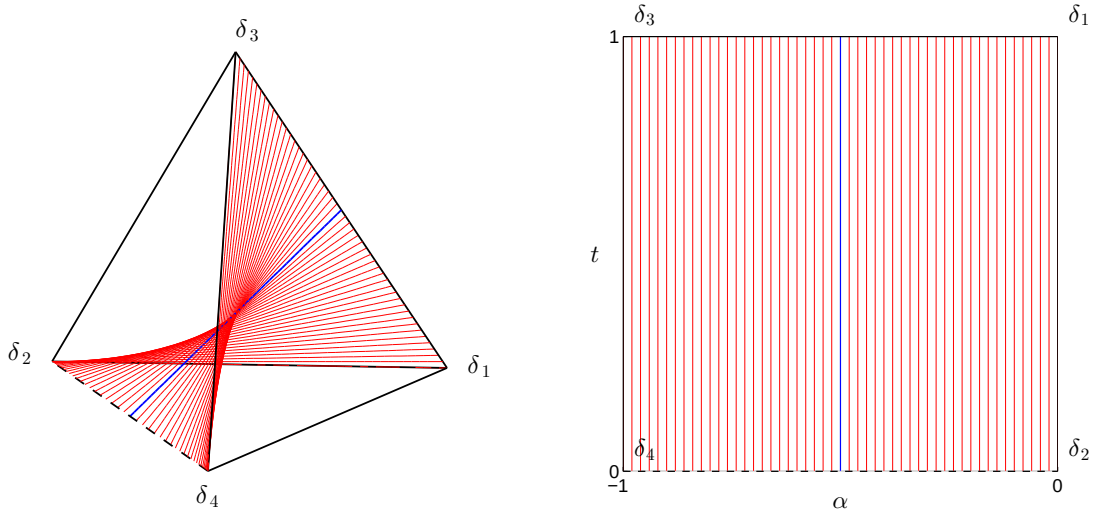
which is positive if $t < (\beta - \alpha)^{-1}$. The same condition applies to component 4. As $[\alpha : \beta] = \left[\frac{\alpha}{\beta-\alpha} : \frac{\beta}{\beta-\alpha}\right]$, we can always assume $\beta > 0$ and $\beta - \alpha = 1$ that is, $\alpha = \beta - 1$; hence $\beta < 1$. The parameterization of the positive probabilities in the model becomes:

$$\begin{aligned} p_1(\alpha, t) &= (\alpha + 1)t \\ p_2(\alpha, t) &= \frac{\alpha^2 - (\alpha^2 + 2\alpha + 1)t + 2\alpha + 1}{2\alpha^2 + 2\alpha + 1} \\ p_3(\alpha, t) &= -\alpha t \\ p_4(\alpha, t) &= -\frac{\alpha^2 t - \alpha^2}{2\alpha^2 + 2\alpha + 1} \end{aligned} , \qquad 0 < t < 1, -1 < \alpha < 0 \, . \tag{45}$$

For example, with $\alpha = -\frac{1}{2}$, we have:

$$\begin{aligned} p_1(\alpha, t) &= \frac{1}{2}t \\ p_2(\alpha, t) &= \frac{1}{2}(1-t) \\ p_3(\alpha, t) &= \frac{1}{2}t \\ p_4(\alpha, t) &= \frac{1}{2}(1-t) \end{aligned} \quad , \qquad 0 < t < 1 \; .$$

In Figure 4(a), we represented the surface associated with the invariant of Equation (21) as a ruled surface in the probability simplex, according to Equations (45), where the blue line corresponds to the case $\alpha = -\frac{1}{2}$. The ruled surface corresponds to the surface in Figure 2 that was approximated by the triangularization of a grid of points satisfying the invariant. In Figure 4(b), we represent the same lines of Figure 4(a) in the chart (α, t).



(a) (b)

Figure 4. Representation of the exponential family in Equations (12) and (13) as a ruled surface in the probability simplex (**a**) and in the parameter space (α, t) (**b**). The dashed line corresponds to the critical edge $\delta_2 \leftrightarrow \delta_4$ and the blue line to the case $\alpha = -\frac{1}{2}$.

From Equation (45), we can express the expectation parameters $\boldsymbol{\eta}$ as a function of (α, t), *i.e.*,

$$\eta_1 = \frac{2\alpha^2 - (2\alpha^3 + 4\alpha^2 + \alpha)t}{2\alpha^2 + 2\alpha + 1} \; , \tag{46}$$

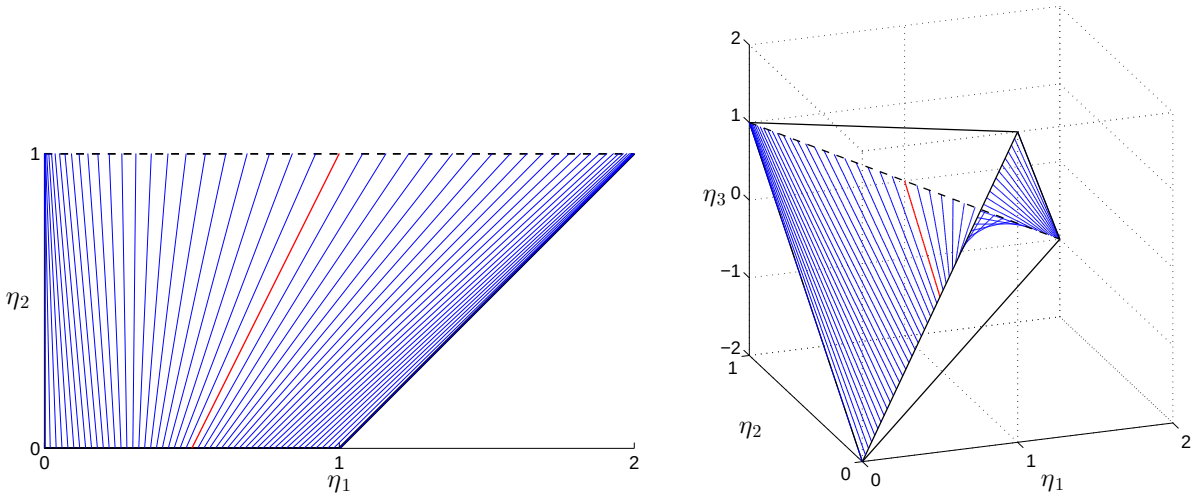
$$\eta_2 = -t + 1 \; , \tag{47}$$

$$\eta_3 = -\frac{(8\alpha^3 + 12\alpha^2 + 10\alpha + 3)t - 2\alpha - 1}{2\alpha^2 + 2\alpha + 1} \; . \tag{48}$$

Notice that the dependence on (α, t) is rational. In Figure 5(a), the ruled surface has been represented in the full marginal polytope, while in Figure 5(a), the lines have been projected over the marginal polytope.

Let us invert Equation (45) to obtain the corresponding chart $\boldsymbol{p} \mapsto (\beta, t)$. From p_1 and p_3, we obtain $\beta = p_1/(p_1 + p_3)$. As $p_2 + p_4 = 1 - t$, we have the chart:

$$\beta = \frac{p_1}{p_1 + p_3} ,$$
$$t = 1 - p_2 - p_4 = p_1 + p_3 .$$



(a) (b)

Figure 5. Representation of the exponential family in Equations (12) and (13) as a ruled surface in the marginal polytope (η_1, η_2) (**a**) and in the full marginal polytope parametrized by (η_1, η_2, η_3) (**b**). The dashed line corresponds to the critical line $\delta_2 \leftrightarrow \delta_4$ and the red line to the case $\alpha = -\frac{1}{2}$.

It is remarkable that the model depends on the probability restricted to $\{1, 3\}$; similarly, the expectation parameters depend on p_1 and p_3 only.

From the theory of exponential families, we know that the gradient mapping:

$$(\theta_1, \theta_2) \mapsto \nabla\psi(\theta_1, \theta_2) = \left[\begin{array}{cc} \frac{2\,e^{(2\,\theta_1+\theta_2)}+e^{\theta_1}}{e^{(2\,\theta_1+\theta_2)}+e^{\theta_1}+e^{\theta_2}+1} & \frac{e^{(2\,\theta_1+\theta_2)}+e^{\theta_2}}{e^{(2\,\theta_1+\theta_2)}+e^{\theta_1}+e^{\theta_2}+1} \end{array} \right]$$

is one-to-one from $\mathbb{R}^2$ onto the interior of the marginal polytope M; see Figure 3(b). The equations:

$$\eta_1 = \frac{\zeta_1 + 2\zeta_1^2\zeta_2}{1 + \zeta_2 + \zeta_1 + \zeta_1^2\zeta_2} ,$$
$$\eta_2 = \frac{\zeta_2 + \zeta_1^2\zeta_2}{1 + \zeta_2 + \zeta_1 + \zeta_1^2\zeta_2} ,$$

are uniquely solvable for $(\eta_1, \eta_2) \in M^\circ$. We study the local solvability in ζ_1, ζ_2 of:

$$(1 + \zeta_2 + \zeta_1 + \zeta_1^2\zeta_2)\eta_1 = \zeta_1 + 2\zeta_1^2\zeta_2 ,$$
$$(1 + \zeta_2 + \zeta_1 + \zeta_1^2\zeta_2)\eta_2 = \zeta_2 + \zeta_1^2\zeta_2 ,$$

that is,

$$0 = \eta_1 + (\eta_1 - 1)\zeta_1 + \eta_1\zeta_2 + (\eta_1 - 2)\zeta_1^2\zeta_2 ,$$
$$0 = \eta_2 + \eta_2\zeta_1 + (\eta_2 - 1)\zeta_2 + (\eta_2 - 1)\zeta_1^2\zeta_2 .$$

The Jacobian is:

$$\begin{bmatrix} (\eta_1 - 1) + 2(\eta_1 - 2)\zeta_1\zeta_2 & \eta_1 + (\eta_1 - 2)\zeta_1^2 \\ \eta_2 + 2(\eta_2 - 1)\zeta_1\zeta_2 & (\eta_2 - 1) + (\eta_2 - 1)\zeta_1^2 \end{bmatrix} .$$

If we introduce the extra variable η_{12}, from Equations (15) and (18) we have the system:

$$\begin{aligned} (1 + \zeta_2 + \zeta_1 + \zeta_1^2\zeta_2)\eta_1 &= \zeta_1 + 2\zeta_1^2\zeta_2 , \\ (1 + \zeta_2 + \zeta_1 + \zeta_1^2\zeta_2)\eta_2 &= \zeta_2 + \zeta_1^2\zeta_2 , \\ (1 + \zeta_2 + \zeta_1 + \zeta_1^2\zeta_2)\eta_{12} &= 2\zeta_1^2\zeta_2 , \end{aligned}$$

Instead, if we use the variable η_3, from Equations (16) and (41), it is possible to derive the equation of the model variety in the η_1, η_2, η_3 parameters. From Equation (18), we have:

$$\begin{aligned} \eta_1 = \mathrm{E}_{\boldsymbol{\zeta}}[T_1] &= \frac{\zeta_1 + 2\zeta_1^2\zeta_2}{1 + \zeta_2 + \zeta_1 + \zeta_1^2\zeta_2} , \\ \eta_2 = \mathrm{E}_{\boldsymbol{\zeta}}[T_2] &= \frac{\zeta_2 + \zeta_1^2\zeta_2}{1 + \zeta_2 + \zeta_1 + \zeta_1^2\zeta_2} , \\ \eta_3 = \mathrm{E}_{\boldsymbol{\zeta}}[T_3] &= \frac{-2 + \zeta_2 + 2\zeta_1 - \zeta_1^2\zeta_2}{1 + \zeta_2 + \zeta_1 + \zeta_1^2\zeta_2} . \end{aligned}$$

Let us solve for the $\boldsymbol{\zeta}$, that is:

$$\begin{aligned} (1 + \zeta_2 + \zeta_1 + \zeta_1^2\zeta_2)\eta_1 &= \zeta_1 + 2\zeta_1^2\zeta_2 , \\ (1 + \zeta_2 + \zeta_1 + \zeta_1^2\zeta_2)\eta_2 &= \zeta_2 + \zeta_1^2\zeta_2 , \\ (1 + \zeta_2 + \zeta_1 + \zeta_1^2\zeta_2)\eta_3 &= -2 + \zeta_2 + 2\zeta_1 - \zeta_1^2\zeta_2 . \end{aligned}$$

There is another way to derive the model constraint in the $\boldsymbol{\eta}$. In the example, the sample space has four points; the monomials $1, T_1, T_2, T_1T_2$ are a vector basis of the linear space of the columns of the matrix A, in particular T_3 is a linear combination:

Ω	$\mathbf{1}$	T_1	T_2	T_1T_2	T_3
1	1	0	0	0	-2
2	1	0	1	0	1
3	1	1	0	0	2
4	1	2	1	2	-1
	-2	4	3	-5	$=$

.

It follows that:

$$\begin{aligned} \eta_3 = \mathrm{E}_{\boldsymbol{\theta}}[T_3] &= \mathrm{E}_{\boldsymbol{\theta}}[-2 + 4T_1 + 3T_2 - 5T_1T_2] \\ &= -2 + 4\mathrm{E}_{\boldsymbol{\theta}}[T_1] + 3\mathrm{E}_{\boldsymbol{\theta}}[T_2] + 3\,\mathrm{Cov}_{\boldsymbol{\theta}}(T_1, T_2) + 3\mathrm{E}_{\boldsymbol{\theta}}[T_1]\,\mathrm{E}_{\boldsymbol{\theta}}[T_2] \\ &= -2 + 4\partial_1\psi(\boldsymbol{\theta}) + 3\partial_2\psi(\boldsymbol{\theta}) - 5\partial_1\partial_2\psi(\boldsymbol{\theta}) - 5\partial_1\psi(\boldsymbol{\theta})\partial_2\psi(\boldsymbol{\theta}) \\ &= -2 + 4\eta_1 + 3\eta_2 - 5\partial_1\partial_2\psi(\boldsymbol{\theta}) - 5\eta_1\eta_2 . \end{aligned}$$

3.1. Border

Let us consider the points in the model variety that are probabilities, that is,

$$p_1 + p_2 + p_3 + p_4 = 1, \quad p_1^2 p_4 = p_2 p_3^2, \quad p_1, p_2, p_3, p_4 \geq 0\ . \tag{49}$$

From the equation above, we see that single zeros are not allowed, that is to say there are no intersections between the model in Equation (49) and the open facets of the probability simplex. We now consider the full marginal polytope obtained by adding the sufficient statistics T_1T_2, and parametrized by $(\eta_1, \eta_2, \eta_{12})$. By Equation (16), the marginal polytope is represented by the inequalities:

$$\begin{aligned} p_1 &= 1 - \eta_1 - \eta_2 + \eta_{12} \geq 0\ , \\ p_2 &= \eta_2 - \frac{1}{2}\eta_{12} \geq 0\ , \\ p_3 &= \eta_1 - \eta_{12} \geq 0\ , \\ p_4 &= \frac{1}{2}\eta_3 \geq 0\ , \end{aligned}$$

which is a convex set with vertexes $(0,0,0)$, $(0,1,0)$, $(1,0,0)$, $(2,1,2)$, which corresponds to the full marginal polytope associated to the sufficient statistics $\{T_1, T_2, T_1T_2\}$. As the critical set is the edge $\delta_2 \leftrightarrow \delta_4$ in the $\boldsymbol{p}$ space, it is the edge $(0,1,0) \leftrightarrow (2,1,2)$ in the $\boldsymbol{\eta}$ space.

We have the following possible models on the border of the probability simplex and on the border of the full marginal polytope, where the values for η_1 and η_2 are obtained from Equation (15).

p_1	p_2	p_3	p_4	η_1	η_2
0	0	+	+	$p_3 + 2p_4$	p_4
0	+	0	+	$2p_4$	$p_2 + p_4$
+	0	+	0	p_3	0
+	+	0	0	0	p_2

p_1	p_2	p_3	p_4	η_1	η_2
+	0	0	0	0	0
0	+	0	0	0	1
0	0	+	0	1	0
0	0	0	+	2	1

That is, the domains that can be support of probabilities in the algebraic model are the faces of the marginal polytope. This is general; see [20,34].

3.2. Fisher Information

Let us consider the covariance matrix of the sufficient statistics. Let us denote by $A_{|12}$ the block of the two central columns in A in Equation (14) and by $\boldsymbol{p}$ the row vector of probabilities. Then, the variance matrix is:

$$A_{|12}^T \operatorname{diag}(\boldsymbol{p})\, A_{|12} - (\boldsymbol{p}A_{|12})^T \boldsymbol{p}A_{|12} = A_{|12}^T \operatorname{diag}(\boldsymbol{p})\, A_{|12} - A_{|12}^T \boldsymbol{p}^T \boldsymbol{p} A_{|12} = A_{|12}^T \left(\operatorname{diag}(\boldsymbol{p}) - \boldsymbol{p}^T\boldsymbol{p}\right) A_{|12}\ .$$

On each of the cases of probabilities supported by a single point, the matrix $\boldsymbol{p} - \boldsymbol{p}^T\boldsymbol{p}$ is zero; hence, the covariance matrix is zero. In each of the cases where the probability is supported by a facet, say $\{1, 2\}$, the matrix $\boldsymbol{p} - \boldsymbol{p}^T\boldsymbol{p}$ reduces to the corresponding block, and the covariance matrix is:

$$\begin{aligned} & \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} p_1 - p_1^2 & -p_1p_2 & 0 & 0 \\ -p_1p_2 & p_2 - p_2^2 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 2 & 1 \end{bmatrix} \\ = & \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} p_1 - p_1^2 & -p_1p_2 \\ -p_1p_2 & p_2 - p_2^2 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \\ = & \begin{bmatrix} 0 & 0 \\ 0 & p_2 - p_2^2 \end{bmatrix} . \end{aligned}$$

The space generated by the covariance matrix is $\mathbb{Q}(0, 1)$, that is the affine space that contains the facets itself. Analogous results hold for each facet, and this result is general.

We note that the determinant of the covariance matrix is a polynomial of degree six in the indeterminates p_1, p_2, p_3. This polynomial is zero on each facet.

The $\boldsymbol{\eta}$ parameters can be given as a function of either $\boldsymbol{\theta}$ or $\boldsymbol{\zeta}$. We have:

$\boldsymbol{\eta}$	$A^T[p_{\boldsymbol{\zeta}}]$
η_1	$(\zeta_1 + 2\zeta_1^2\zeta_2)/(1 + \zeta_2 + \zeta_1 + \zeta_1^2\zeta_2)$
η_2	$(\zeta_2 + \zeta_1^2\zeta_2)/(1 + \zeta_2 + \zeta_1 + \zeta_1^2\zeta_2)$
η_3	$(-2 + \zeta_2 + 2\zeta_1 - \zeta_1^2\zeta_2)/(1 + \zeta_2 + \zeta_1 + \zeta_1^2\zeta_2)$

(50)

We know from the theory of exponential families that the mapping:

$$]0, \infty[\times]0, \infty[\ni (\zeta_1, \zeta_2) \mapsto (\eta_1, \eta_2) \in \mathrm{Conv}\,\{(T_1(x), T_2(x)) | x \in \Omega\}^\circ$$

is one-to-one. We look for an algebraic inversion of the equations:

$$\begin{aligned} (1 + \zeta_2 + \zeta_1 + \zeta_1^2\zeta_2)\eta_1 &= \zeta_1 + 2\zeta_1^2\zeta_2 , \\ (1 + \zeta_2 + \zeta_1 + \zeta_1^2\zeta_2)\eta_2 &= \zeta_2 + \zeta_1^2\zeta_2 . \end{aligned}$$

If we rewrite Equations (50) as polynomials in ζ_1, ζ_2, we obtain:

$$\eta_1 + (\eta_1 - 1)\zeta_1 + \eta_1\zeta_2 + (\eta_1 - 2)\zeta_1^2\zeta_2 = 0 , \tag{51}$$

$$\eta_2 + \eta_2\zeta_1 + (\eta_2 - 1)\zeta_2 + (\eta_2 - 1)\zeta_1^2\zeta_2 = 0 , \tag{52}$$

$$-\eta_3 + (\eta_3 - 2)\zeta_1 + (\eta_3 - 1)\zeta_2 + (\eta_3 + 1)\zeta_1^2\zeta_2 = 0 . \tag{53}$$

Gauss elimination produces a linear system in ζ_1, ζ_2 with coefficients that are polynomials in η_1, η_2, η_3 to be considered with the implicit equation derived from $p_1^2p_4 - p_2p_3^2 = 0$. The system is:

$$\begin{aligned} -2\eta_2\eta_3 - 2\eta_1 + 2\eta_2 &= (-2\eta_2\eta_3 - 2\eta_1 + 2)\zeta_1 + (-2\eta_2\eta_3 + 2\eta_2 + 2\eta_3 - 2)\zeta_2 , \\ \eta_2 &= \eta_2\zeta 1 + (\eta_2 - 1)\zeta_2 . \end{aligned}$$

3.3. Extension of the Model

In this subsection, we study an extension to signed probabilities of the exponential family in Equations (12) and (13) based on the representation of the statistical model as a ruled surface in the probability simplex. Our motivation for such an analysis is the study of the stability of the critical points of a gradient field in the $\boldsymbol{\eta}$ parameters, in particular when the critical points belong to the boundary of the model. Indeed, by extending the gradient field outside the marginal polytope, we can identify open neighborhoods for critical points on the boundary of the polytope, which allow one to study the convergence of the differential equations associated with the gradient flows, for instance by means of Lyapunov stability.

In the following, we describe more in detail how the extension can be obtained. Let $\boldsymbol{a}$ be a point along the edge $\delta_2 \leftrightarrow \delta_4$ of the full marginal polytope parametrized by (η_1, η_2, η_3) and $\boldsymbol{b}$ the coordinates of the corresponding point over $\delta_1 \leftrightarrow \delta_3$ obtained by intersecting the line of the ruled surface through $\boldsymbol{a}$ with the edge $\delta_1 \leftrightarrow \delta_3$. The values of the η_2 coordinate for $\boldsymbol{a}$ and $\boldsymbol{b}$ are one and zero, respectively. The other coordinates of $\boldsymbol{b}$ depend on those of $\boldsymbol{a}$ though α. First, we obtain the values of the η_3 coordinates as a function of the η_1 coordinate. For $\boldsymbol{a}$, we find the equation of the line to which $\delta_2 \leftrightarrow \delta_4$ belongs, given by:

$$\begin{pmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} + u \begin{pmatrix} 2 \\ 0 \\ -2 \end{pmatrix} = \begin{pmatrix} 2u \\ 1 \\ 1 - 2u \end{pmatrix} , \tag{54}$$

from which we obtain $\eta_3 = 1 - \eta_1$. Similarly, for the η_3 coordinate of $\boldsymbol{b}$, we consider the line through $\delta_1 \leftrightarrow \delta_3$, that is:

$$\begin{pmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ -2 \end{pmatrix} + u \begin{pmatrix} 1 \\ 0 \\ 4 \end{pmatrix} = \begin{pmatrix} u \\ 0 \\ 4u - 2 \end{pmatrix} , \tag{55}$$

which gives us $\eta_3 = 4\eta_1 - 2$. Finally, for the η_1 coordinate, we use Equations (44). In $\boldsymbol{a}$, since $t = 0$ and $p_1 = p_3 = 0$, then $p_2 = \frac{\beta^2}{\alpha^2+\beta^2}$ and $p_4 = \frac{\alpha^2}{\alpha^2+\beta^2}$. From Equation (24), it follows that:

$$\eta_1 = \frac{2\alpha^2}{2\alpha^2 + 2\alpha + 1} . \tag{56}$$

Similarly, for $\boldsymbol{b}$, we have $p_2 = p_4 = 0$ and $t = 1$, so that $p_1 = \alpha + 1$ and $p_3 = -\alpha$. From Equation (24), it follows that:

$$\eta_1 = -\alpha . \tag{57}$$

As a result, the coordinates of $\boldsymbol{a}$ and $\boldsymbol{b}$ both depend on α as follows,

$$\boldsymbol{a} = \left(\frac{2\alpha^2}{2\alpha^2 + 2\alpha + 1}, 1, \frac{2\alpha + 1}{2\alpha^2 + 2\alpha + 1} \right) \tag{58}$$

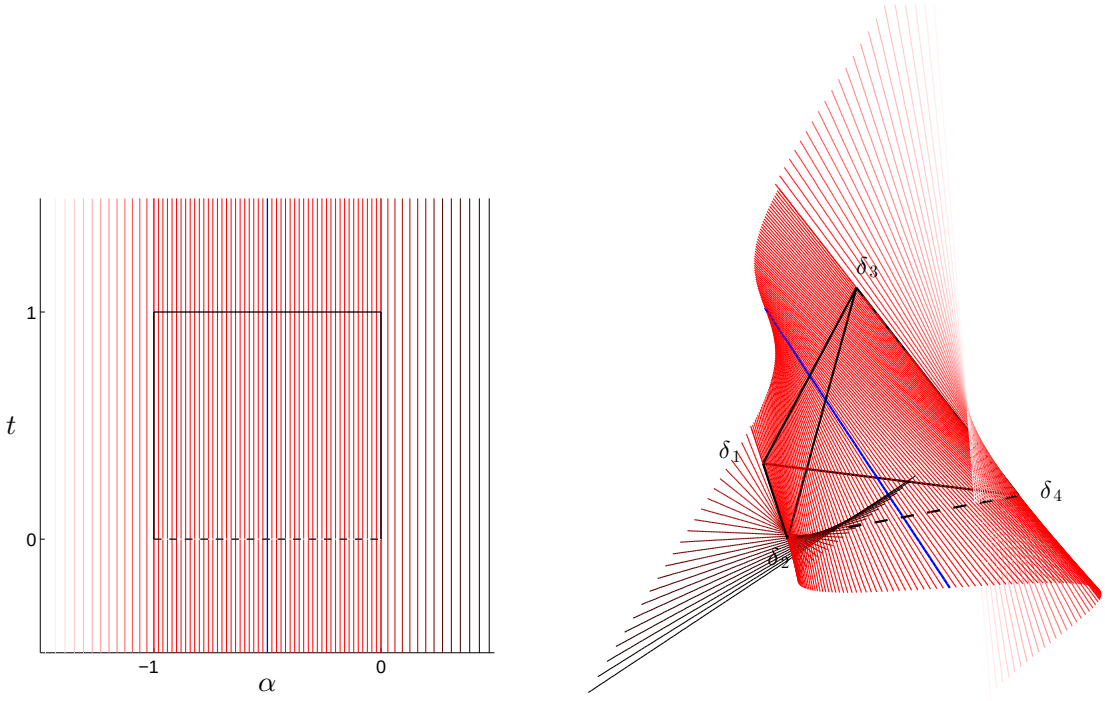
$$\boldsymbol{b} = (-\alpha, 0, -4\alpha - 2) \tag{59}$$

The ruled surface in the full marginal polytope is given by the lines through $\boldsymbol{a}$ and $\boldsymbol{b}$ described by the following implicit representation, for $-1 < \alpha < 1$ and $0 < t < 1$,

$$\begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{bmatrix} = \begin{bmatrix} -\alpha \\ 0 \\ -4\alpha - 2 \end{bmatrix} + t \begin{bmatrix} \frac{2\alpha^3+4\alpha^2+\alpha}{2\alpha^2+2\alpha+1} \\ 1 \\ \frac{8\alpha^3+12\alpha^2+10\alpha+3}{2\alpha^2+2\alpha+1} \end{bmatrix}. \tag{60}$$

The ruled surface can be extended outside the marginal polytope by taking values of $\alpha, t \in \mathbb{R}$ and considering the set of lines through $\boldsymbol{a}$ and $\boldsymbol{b}$ for different values of α. For $\alpha \to \pm\infty$, the η_1 coordinate of $\boldsymbol{b}$ tends to $\mp\infty$, while the η_1 of $\boldsymbol{a}$ tends to one. For $\alpha \to \pm\infty$, the ruled surface admits the same limit given by the line parallel to $\delta_1 \leftrightarrow \delta_3$ passing through $(1, 1, 0)$. The surface intersects the interior of the marginal polytope for $t \in (0, 1)$ and $\alpha \in (-1, 0)$. Moreover, the surface intersects the critical line twice, for $t = 0, \alpha \in [-1, 0]$ and for $t = 0, \alpha \notin [-1, 0]$.

In Figures 6 and 7, we represent the extension of the ruled surface outside the probability simplex and in the (α, t) chart, while in Figures 8 and 9, the extended surface has been represented in the full marginal polytope parametrized by (η_1, η_2, η_3) and in the marginal polytope parametrized by (η_1, η_2).



(a) (b)

Figure 6. The segments that form the ruled surface in Figure 4 have been extended, for $-0.5 < t < 1.5$. New lines described by Equations (60) have been represented for $0 < \alpha < \exp(0.7)$ (shading from red to black for increasing values of α) and for $\exp(0.7) - 1 < \alpha < -1$ (shading from red to white for decreasing values of α). The simplex in (**b**) has been rotated with respect to Figure 4(a) to better visualize the intersection of the lines with the critical edge $\delta_2 \leftrightarrow \delta_4$.

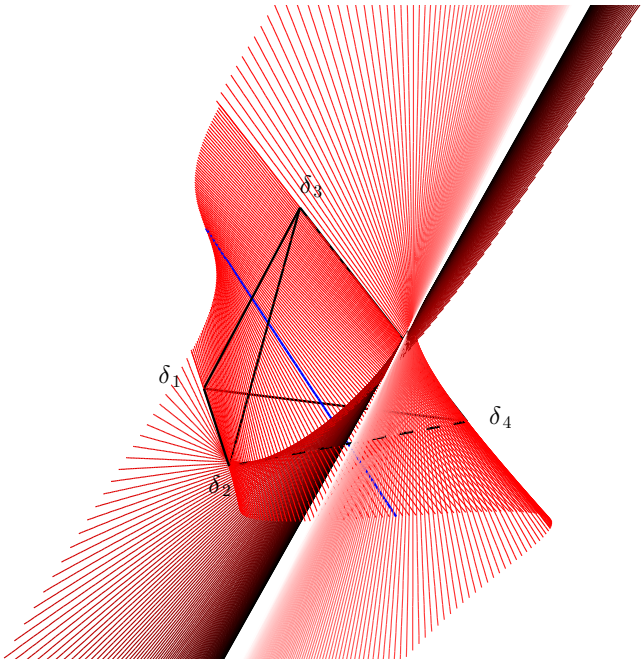
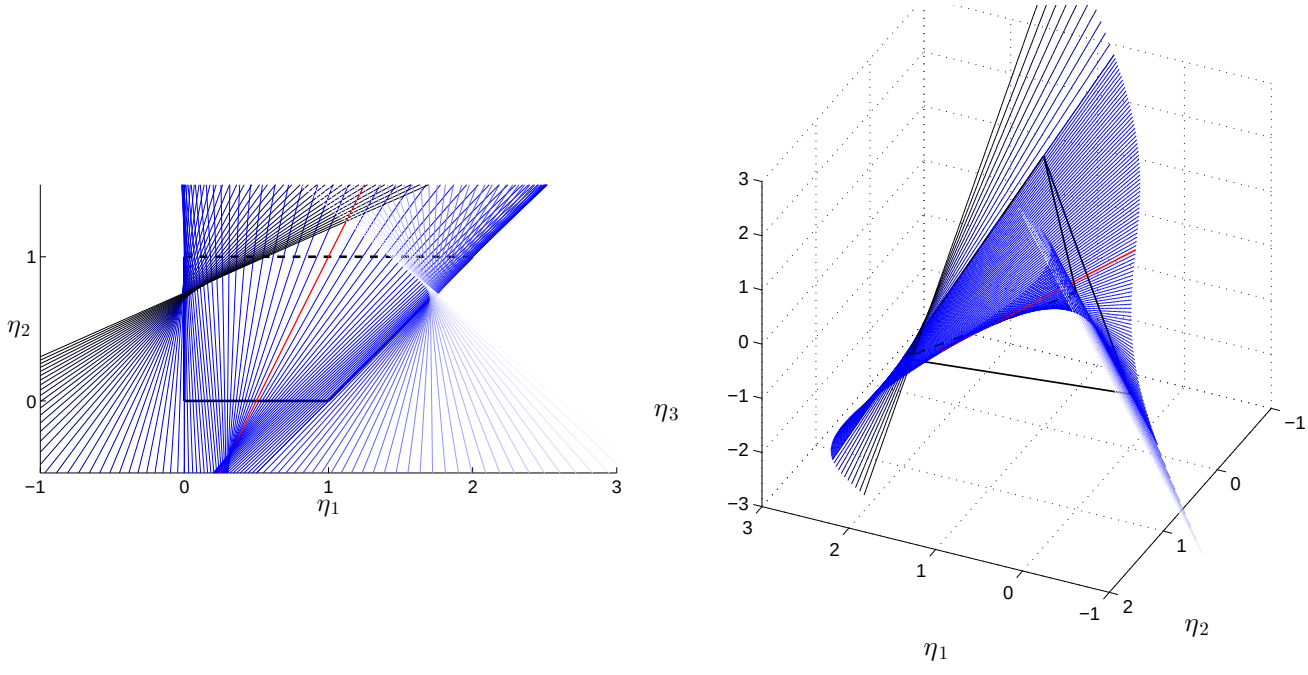


Figure 7. Extension of the ruled surface associated with the exponential family in Equations (12) and (13) as in Figure 6(b), for $\exp(3.5) - 1 < \alpha < \exp(3.5)$ and $-0.5 < t < 1.5$; for $\alpha \to \pm\infty$, the lines of the extended surface admit the same limit.



(a) (b)

Figure 8. The segments that form the ruled surface in Figure 5 have been extended, for $-0.5 < t < 1.5$. New lines described by Equations (60) have been represented for $0 < \alpha < \exp(1)$ (shading from blue to black for increasing values of α) and $\exp(1) - 1 < \alpha < -1$ (shading from blue to white for decreasing values of α). The full marginal polytope in (**b**) has been rotated with respect to Figure 5(b) to better visualize the intersection of the lines with the critical edge $\delta_2 \leftrightarrow \delta_4$.

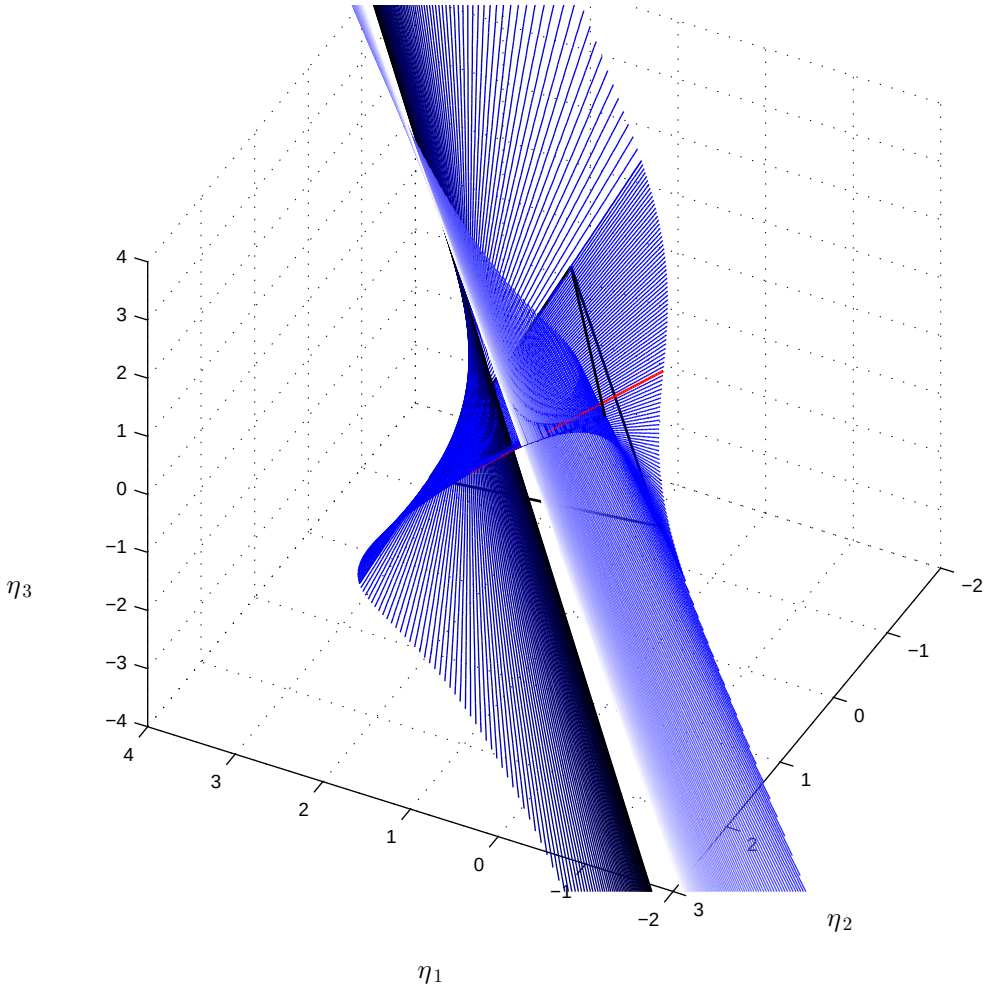


Figure 9. Extension of the ruled surface associated with the exponential family in Equations (12) and (13) as in Figure 8(b), for $\exp(3) - 1 < \alpha < \exp(3)$ and $-0.5 < t < 1.5$; notice that for $\alpha \to \pm\infty$, the lines of the extended surface admit the same limit.

3.4. Optimization and Natural Gradient Flows

We are interested in the study of natural gradient flows of functions defined over statistical models. Our motivation is the study of the optimization of the stochastic relaxation of a function, *i.e.*, the optimization of the expected value of the function itself with respect to a distribution p in a statistical model. Natural gradient flows associated with the stochastic relaxation converge to the boundary of the model, where the probability mass is concentrated on some instances of the search space. To study the convergence over the boundary, we proposed to extend the natural gradient field outside the marginal polytope and the probability simplex, by employing a parameterization that describes the model as a ruled surface, as we described in the tutorial example of this section.

In the following, we focus on the optimization of a function $f : \Omega \to \mathbb{R}$, and we consider its stochastic relaxation with respect to a probability distribution in the exponential family in Equations (12) and (13). First, we compute a basis for all real-valued functions defined over Ω using algebraic arguments. Consider the zero-dimensional ideal I associated with the set of points in Ω, and let R be the polynomial

ring with the field of real coefficients; a vector space basis for the quotient ring R/I defines a basis for all functions defined over Ω. In CoCoA [36], this can be computed with the command `QuotientBasis`.

Coming back to our example, with $\Omega = \{1, 2, 3, 4\}$, by fixing the graded reverse lexicographical monomial order, which is the default one in CoCoA [36], we obtain a basis given by $\{1, x_1, x_2, x_1x_2\}$, so that any $f : \Omega \to \mathbb{R}$ can be written as:

$$f = c_0 + c_1x_1 + c_2x_2 + c_{12}x_1x_2 \, . \tag{61}$$

We are interested in the study of the natural gradient field of $F(p) = \mathbb{E}_p[f]$. Recall that $T_3 = 4x_1 + 3x_2 - 5x_1x_2 - 2$ and $\eta_3 = \mathbb{E}[T_3]$, so that:

$$\mathbb{E}[x_1x_2] = \frac{1}{5}(4\eta_1 + 3\eta_2 - \eta_3 - 2) \, , \tag{62}$$

which implies:

$$F_\eta(\boldsymbol{\eta}) = c_0 - \frac{2}{5}c_{12} + \left(c_1 + \frac{4}{5}c_{12}\right)\eta_1 + \left(c_2 + \frac{3}{5}c_{12}\right)\eta_2 - \frac{1}{5}c_{12}\eta_3 \, . \tag{63}$$

In order to study the gradient field of $F_\eta(\boldsymbol{\eta})$ over the marginal polytope parameterized by (η_1, η_2), we need to express η_3 as a function of η_1 and η_2. In order to do that, we parametrize the exponential family as a ruled surface by means of the (α, t) parameters. Moreover, this parametrization has a natural extension outside the marginal polytope, which allows one to study the stability of the critical points on the boundary of the marginal polytope. We start by evaluating the gradient field of $F_{\alpha,t}(\alpha, t)$ in the (α, t) parametrization, then we map it to the marginal polytope in the $\boldsymbol{\eta}$ parameterization.

By expressing (η_1, η_2) as a function of (α, t), we obtain:

$$F_{\alpha,t}(\alpha, t) = \frac{2\alpha^2(c_1 + c_{12}) + (2\alpha^2 + 2\alpha + 1)(c_0 + c_2) - (2\alpha^2(c_1 + c_{12}) + (2\alpha^2 + 2\alpha + 1)(c_1\alpha + c_2))t}{2\alpha^2 + 2\alpha + 1} \, . \tag{64}$$

If we take partial derivatives of Equation (64) with respect to α and t, we have:

$$\partial_\alpha F_{\alpha,t}(\alpha, t) = \frac{4(\alpha^2 + \alpha)(c_1 + c_{12}) - ((4\alpha^4 + 8\alpha^3 + 12\alpha^2 + 8\alpha + 1)c_1 + 4(\alpha^2 + \alpha)c_{12})t}{4\alpha^4 + 8\alpha^3 + 8\alpha^2 + 4\alpha + 1} \, , \tag{65}$$

$$\partial_t F_{\alpha,t}(\alpha, t) = -\frac{2\alpha^2 c_{12} + (2\alpha^3 + 4\alpha^2 + \alpha)c_1 + (2\alpha^2 + 2\alpha + 1)c_2}{2\alpha^2 + 2\alpha + 1} \, . \tag{66}$$

In the (α, t) parameterization, the Fisher information matrix reads:

$$I_{\alpha,t}(\alpha, t) = \mathbb{E}_{\alpha,t}[-\partial^2 \log p(x; \alpha, t)] = \begin{bmatrix} \frac{4\alpha^2 - (4\alpha^4 + 8\alpha^3 + 12\alpha^2 + 8\alpha + 1)t + 4\alpha}{4\alpha^6 + 12\alpha^5 + 16\alpha^4 + 12\alpha^3 + 5\alpha^2 + \alpha} & 0 \\ 0 & -(t^2 - t)^{-1} \end{bmatrix} \, . \tag{67}$$

Finally, the natural gradient becomes:

$$\widetilde{\nabla} F_{\alpha,t}(\alpha, t) = I_{\alpha,t}(\alpha, t)^{-1} \nabla F_{\alpha,t}(\alpha, t) \tag{68}$$

$$= \begin{bmatrix} \frac{(4\alpha^6 + 12\alpha^5 + 16\alpha^4 + 12\alpha^3 + 5\alpha^2 + \alpha)(4(\alpha^2 + \alpha)c_1 + 4(\alpha^2 + \alpha)c_{12} - ((4\alpha^4 + 8\alpha^3 + 12\alpha^2 + 8\alpha + 1)c_1 + 4(\alpha^2 + \alpha)c_{12})t)}{(4\alpha^4 + 8\alpha^3 + 8\alpha^2 + 4\alpha + 1)(4\alpha^2 - (4\alpha^4 + 8\alpha^3 + 12\alpha^2 + 8\alpha + 1)t + 4\alpha)} \\ \frac{(2\alpha^2 c_{12} + (2\alpha^3 + 4\alpha^2 + \alpha)c_1 + (2\alpha^2 + 2\alpha + 1)c_2)(t^2 - t)}{2\alpha^2 + 2\alpha + 1} \end{bmatrix}$$

We obtained a rational formula for the natural gradient in the (α, t) parameterization, which can be easily extended outside the marginal polytope. However, notice that the inverse Fisher information matrix and the natural gradient are not defined for:

$$t = \frac{4(\alpha^2+\alpha)}{4\alpha^4+8\alpha^3+12\alpha^2+8\alpha+1} \,. \tag{69}$$

We also remark that over the boundary of the model, for $t \in \{0,1\}$ and $\alpha \in \{-1,0\}$, the determinant of the inverse Fisher information vanishes, so that the matrix is not full rank. It follows that the trajectories associated with natural gradient flows with initial conditions in the interior of the marginal polytope remain in the marginal polytope.

In order to study the natural gradient field over the marginal polytope, we apply a reparameterization of a tangent vector from the (α, t) parameterization to the (η_1, η_2) parameterization. Indeed, by the chain rule and the inverse function theorem, we have:

$$\nabla F_\eta(\alpha, t) = \nabla F_{\alpha,t}(\alpha, t)^T J(\alpha, t)^{-1} \tag{70}$$

The Jacobian of the map $(\alpha, t) \mapsto (\eta_1, \eta_2)$ is:

$$J(\alpha,t) = \begin{bmatrix} -\frac{(6\alpha^2+8\alpha+1)t-4\alpha}{2\alpha^2+2\alpha+1} & -\frac{2(2\alpha^2-(2\alpha^3+4\alpha^2+\alpha)t)(2\alpha+1)}{(2\alpha^2+2\alpha+1)^2} & -\frac{2\alpha^3+4\alpha^2+\alpha}{2\alpha^2+2\alpha+1} \\ 0 & & -1 \end{bmatrix} , \tag{71}$$

with inverse:

$$J(\alpha,t)^{-1} = \begin{bmatrix} \frac{4\alpha^4+8\alpha^3+8\alpha^2+4\alpha+1}{4\alpha^2-(4\alpha^4+8\alpha^3+12\alpha^2+8\alpha+1)t+4\alpha} & -\frac{4\alpha^5+12\alpha^4+12\alpha^3+6\alpha^2+\alpha}{4\alpha^2-(4\alpha^4+8\alpha^3+12\alpha^2+8\alpha+1)t+4\alpha} \\ 0 & -1 \end{bmatrix} . \tag{72}$$

It follows that:

$$\nabla F_\eta(\alpha,t) = \begin{bmatrix} \frac{4(\alpha^2+\alpha)c1+4(\alpha^2+\alpha)c12-((4\alpha^4+8\alpha^3+12\alpha^2+8\alpha+1)c1+4(\alpha^2+\alpha)c12)t}{4\alpha^2-(4\alpha^4+8\alpha^3+12\alpha^2+8\alpha+1)t+4\alpha} \\ -\frac{4(\alpha^3+\alpha^2)c12-4(\alpha^2+\alpha)c2+(2(2\alpha^4-\alpha^2)c12+(4\alpha^4+8\alpha^3+12\alpha^2+8\alpha+1)c2)t}{4\alpha^2-(4\alpha^4+8\alpha^3+12\alpha^2+8\alpha+1)t+4\alpha} \end{bmatrix} . \tag{73}$$

Notice that, as for the inverse Fisher information matrix, the inverse Jacobian $J(\alpha, t)^{-1}$ is not defined for t which satisfies Equation (69).

We compute the inverse Fisher information matrix by evaluating the covariance between the sufficient statistics of the exponential family. Since over Ω, we have $x_1^2 = x_1 + x_1x_2$ and $x_1^2 = x_1$, it follows that:

$$I_\eta(\eta)^{-1} = \begin{bmatrix} \frac{1}{5}(9\eta_1+3\eta_2-\eta_3-2)-\eta_1^2 & \frac{1}{5}(4\eta_1+3\eta_2-\eta_3-2)-\eta_1\eta_2 \\ \frac{1}{5}(4\eta_1+3\eta_2-\eta_3-2)-\eta_1\eta_2 & \eta_2-\eta_2^2 \end{bmatrix} . \tag{74}$$

By parameterizing I_η^{-1} with (α, t), we have:

$$I_\eta(\alpha,t)^{-1} = \begin{bmatrix} \frac{4\alpha^4+8\alpha^3-(4\alpha^6+16\alpha^5+20\alpha^4+8\alpha^3+\alpha^2)t^2+4\alpha^2+(4\alpha^5-12\alpha^3-8\alpha^2-\alpha)t}{4\alpha^4+8\alpha^3+8\alpha^2+4\alpha+1} & -\frac{(2\alpha^3+4\alpha^2+\alpha)t^2-(2\alpha^3+4\alpha^2+\alpha)t}{2\alpha^2+2\alpha+1} \\ -\frac{(2\alpha^3+4\alpha^2+\alpha)t^2-(2\alpha^3+4\alpha^2+\alpha)t}{2\alpha^2+2\alpha+1} & -t^2+t \end{bmatrix}. \tag{75}$$

Finally, we derive the following rational formula for the natural gradient over the marginal polytope parametrized as a ruled surface by (α, t):

$$\widetilde{\nabla} F_\eta(\alpha,t) = I_\eta(\alpha,t)^{-1}\nabla F_\eta(\alpha,t) = \begin{bmatrix} \frac{((4\alpha^6+16\alpha^5+20\alpha^4+8\alpha^3+\alpha^2)c_1+2(2\alpha^5+4\alpha^4+\alpha^3)c_{12}+(4\alpha^5+12\alpha^4+12\alpha^3+6\alpha^2+\alpha)c_2)t^2-4(\alpha^4+2\alpha^3+\alpha^2)c_1+}{-4(\alpha^4+2\alpha^3+\alpha^2)c_{12}-((4\alpha^5-12\alpha^3-8\alpha^2-\alpha)c_1+2(2\alpha^5+2\alpha^4-3\alpha^3-2\alpha^2)c_{12}+(4\alpha^5+12\alpha^4+12\alpha^3+6\alpha^2+\alpha)c_2)t} \\ \hline 4\alpha^4+8\alpha^3+8\alpha^2+4\alpha+1 \\ -\frac{(2\alpha^2c_{12}+(2\alpha^3+4\alpha^2+\alpha)c_1+(2\alpha^2+2\alpha+1)c_2)t^2-(2\alpha^2c_{12}+(2\alpha^3+4\alpha^2+\alpha)c_1+(2\alpha^2+2\alpha+1)c_2)t}{2\alpha^2+2\alpha+1} \end{bmatrix}. \tag{76}$$

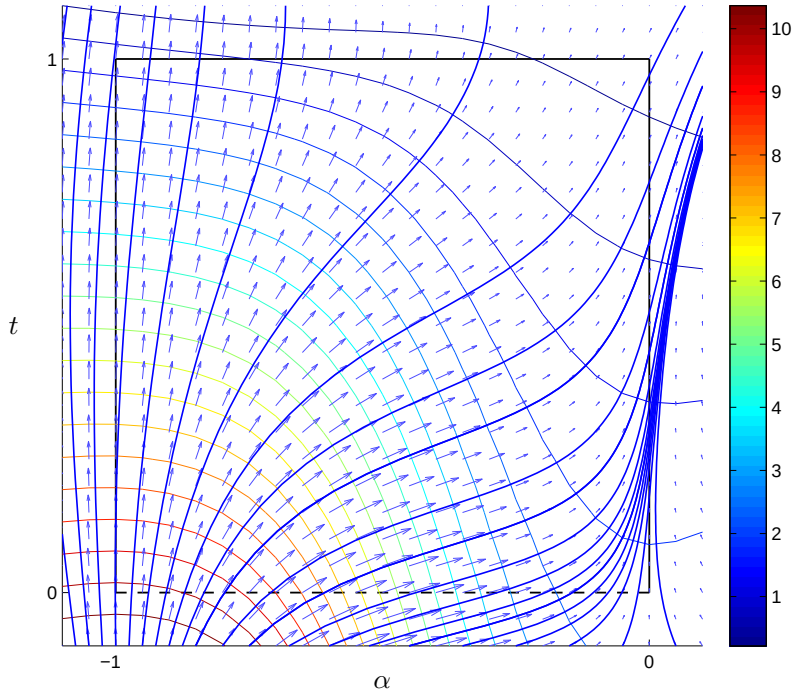
3.5. Examples with Global and Local Optima

We conclude this section with two examples of natural gradient flows associated with two different f functions. First, consider the case where $c_0 = 0, c_1 = 1, c_2 = 2, c_3 = 3$, so that:

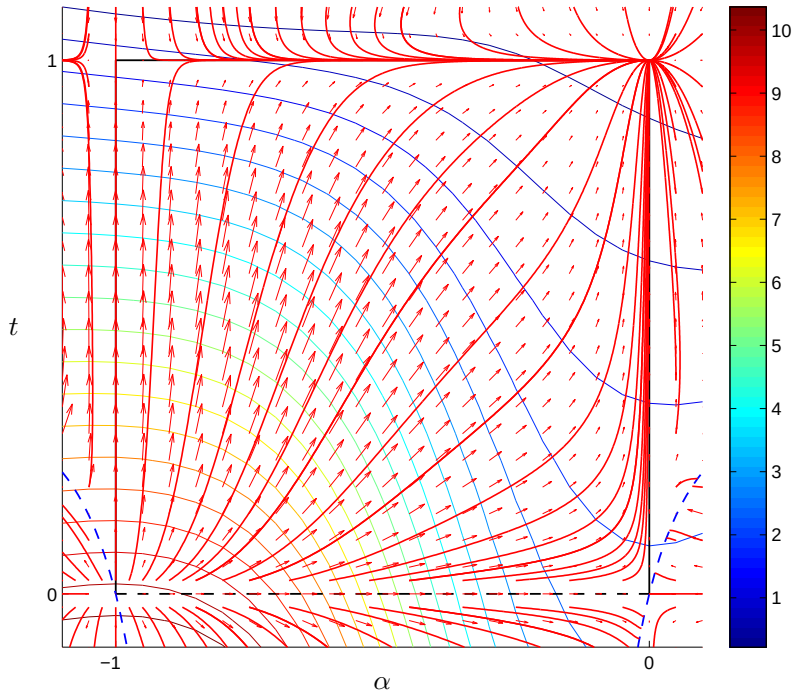
Ω	x_1	x_2	f_1
1	0	0	0
2	0	1	2
3	1	0	1
4	2	1	10

. (77)

The function admits a minimum on $\{1\}$. In Figure 10, we plotted the vector fields associated with the vanilla and natural gradient, together with some gradient flows for different initial conditions, in the (α, t) parameterization. In Figure 11, we represent the vanilla and natural gradient field over the marginal polytope in the (η_1, η_2) parameterization. Notice that, as expected, differently from the vanilla gradient, the natural gradient flows converge to the unique global optima, which corresponds to the vertex where all of the probability is concentrated over $\{1\}$. In the (α, t) parameterization, the flows have been extended outside the statistical model by prolonging the lines of the ruled surface, and as we can see, they remain compatible with the flows on the interior of the model, in the sense that the nature of the critical point is the same for trajectories with initial conditions on the interior and on the exterior of the model. In other words, the global optima is an attractor from both the interior and the exterior of the model and similarly for the other critical points on the vertices, both for saddle points and the unstable points, where the natural gradient vanishes.

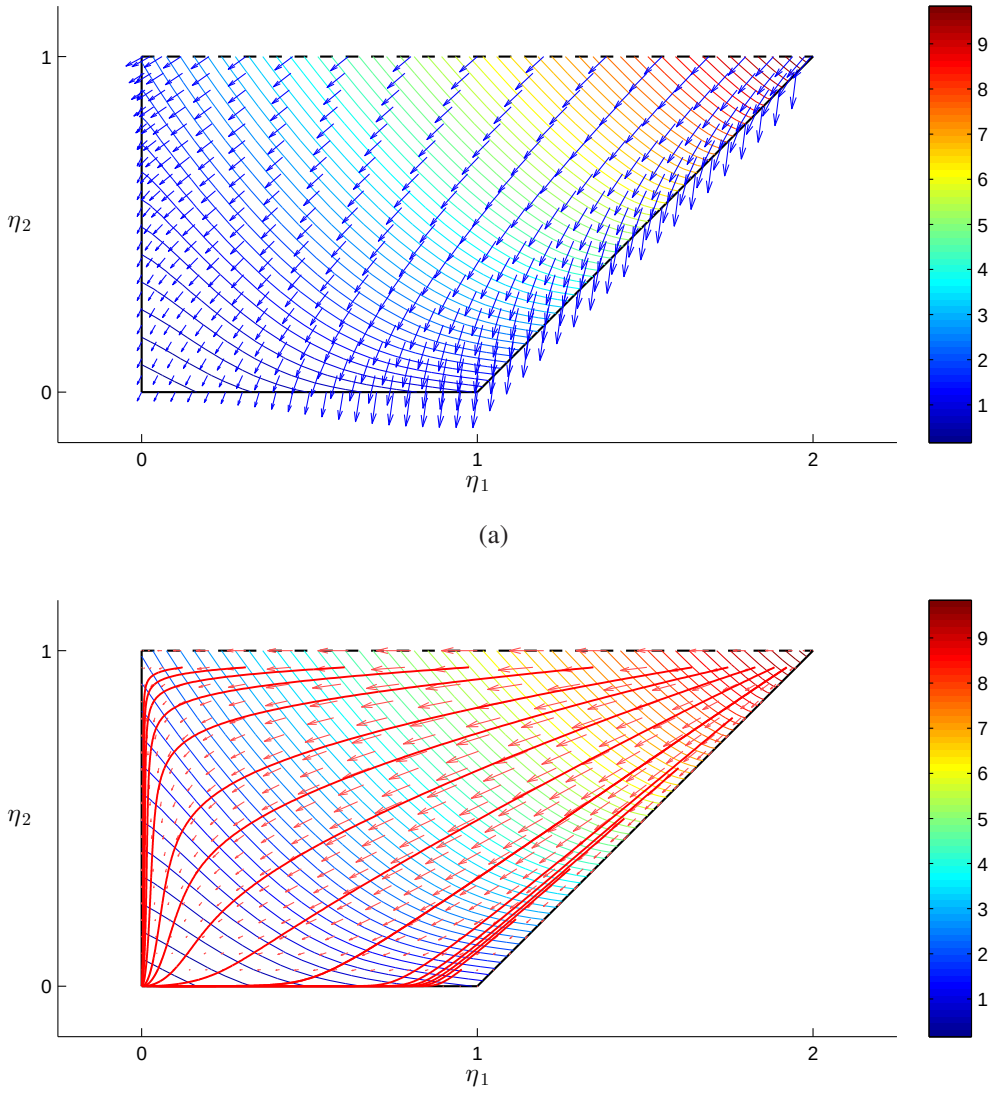


(a)



(b)

Figure 10. Vanilla gradient field and flows in blue (**a**) and natural gradient field and flows in red (**b**), together with level lines associated with $F_{\alpha,t}(\alpha, t)$ in the (α, t) parameterization, for $c_0 = 0, c_1 = 1, c_2 = 2$ and $c_3 = 3$; the dashed blue lines in (**b**) represent the points where $\widetilde{\nabla} F_{\alpha,t}(\alpha, t)$ is not defined; see Equation (68).



(a)

(b)

Figure 11. Vanilla gradient field in blue (**a**) and natural gradient field and flows in red (**b**), together with level lines associated with $F_\eta(\alpha, t)$ over the marginal polytope, for $c_0 = 0$, $c_1 = 1, c_2 = 2$ and $c_3 = 3$.

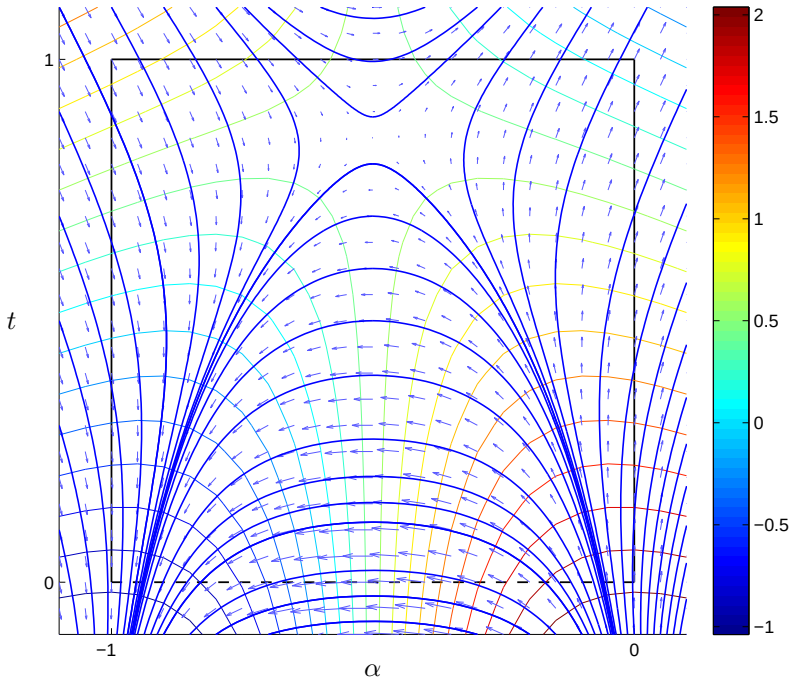
In the second example, we set $c_0 = 0, c_1 = 1, c_2 = 2, c_3 = -5/2$, and we have:

Ω	x_1	x_2	f_2
1	0	0	0
2	0	1	2
3	1	0	1
4	2	1	-1

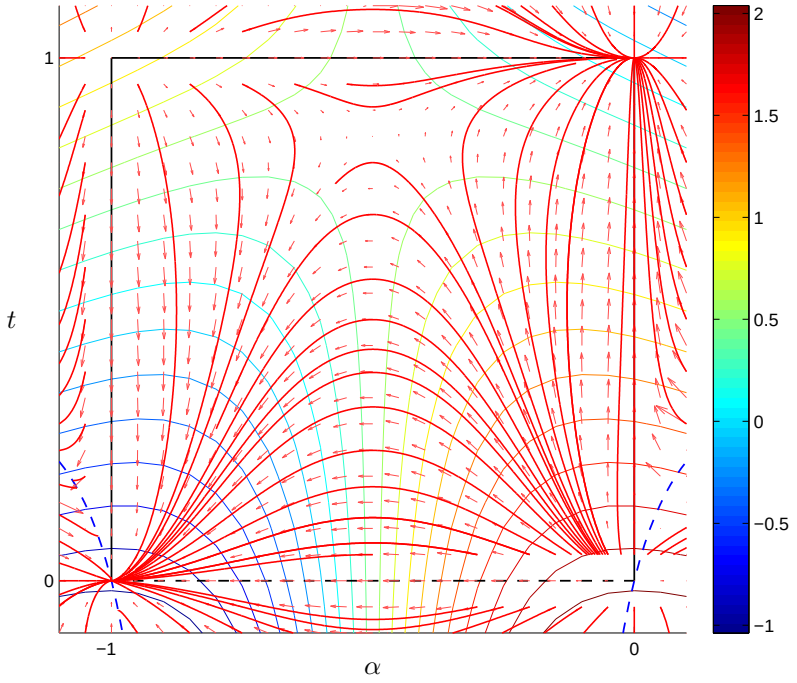
(78)

so that f_2 admits a minimum on $\{4\}$. In Figures 12 and 13, we plotted the vector fields associated with the vanilla and natural gradient, together with some gradient flows for different initial conditions, in the (α, t) and (η_1, η_2) parameterization, respectively. As in the previous example, natural gradient flows converge to the vertices of the model; however, in this case, we have one local optima in $\{1\}$ and one

global optima in $\{4\}$, together with a saddle point in the interior of the model. Similarly to the previous example, in the (α, t) parameterization, the flows have been extended outside the statistical model, and the nature of the critical points is the same for trajectories with initial conditions in the statistical model and in the extension of the statistical model.

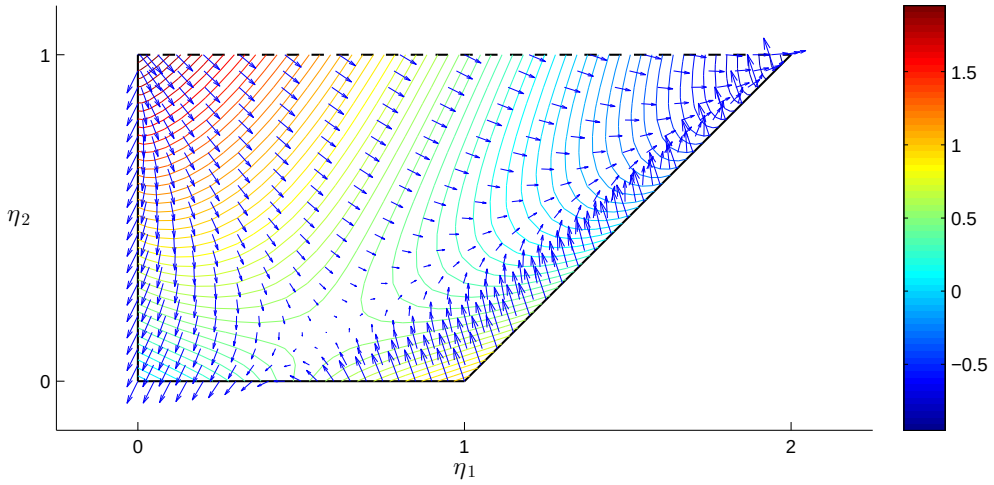


(a)

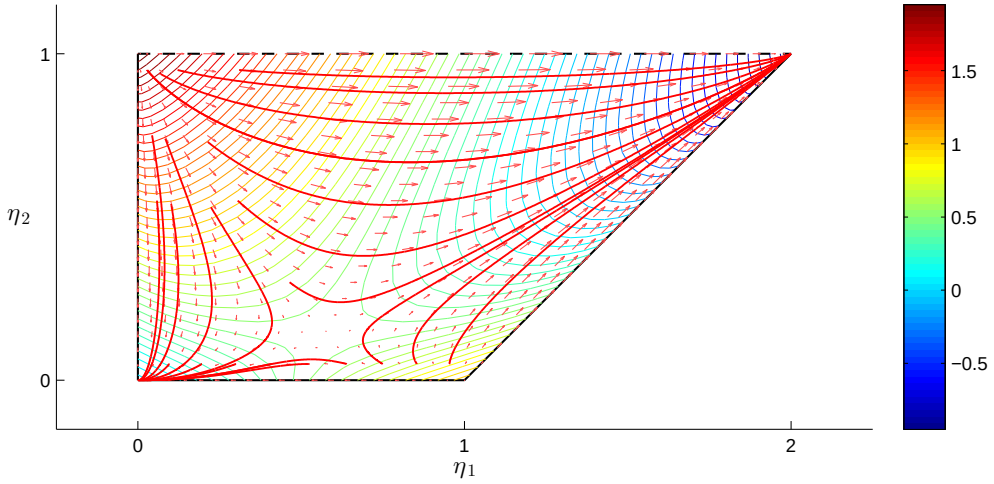


(b)

Figure 12. Vanilla gradient field and flows in blue (**a**) and natural gradient field and flows in red (**b**) as in Figure 10, for $c_0 = 0, c_1 = 1, c_2 = 2$ and $c_3 = -\frac{5}{2}$.



(a)



(b)

Figure 13. Vanilla gradient field in blue (**a**) and natural gradient field and flows in red (**b**) as in Figure 11, for $c_0 = 0, c_1 = 1, c_2 = 2$ and $c_3 = -\frac{5}{2}$.

We conclude the section by noticing that in both examples, for certain values of t in Equation (69), the natural gradient flows are not defined on the extension of the statistical model. As represented in the figures, once a trajectory encounters the dashed blue line in the (α, t) parameterization, the flow stops at that point.

4. Pseudo-Boolean Functions

We turn to discuss a case of considerable practical interest to see which of the results obtained in the example of the previous section we are able to extend.

For binary variables, we use the coding ± 1, that is $\boldsymbol{x} = (x_1, \ldots, x_n) \in \{+1, -1\}^n = \Omega$. For any function $f \colon \Omega \mapsto \mathbb{R}$, with multi-index notation, $f(\boldsymbol{x}) = \sum_{\boldsymbol{\alpha} \in L} a_{\boldsymbol{\alpha}} \boldsymbol{x}^{\boldsymbol{\alpha}}$, with $L = \{0, 1\}^n$ and $\boldsymbol{x}^{\boldsymbol{\alpha}} = \prod_{i=1}^n x_i^{\alpha_i}$, $0^0 = 1$. If $M \subset L^* = L \setminus \{\mathbf{0}\}$, the model where $p \in \mathcal{E}$ if:

$$p \propto \exp\left(\sum_{\boldsymbol{\alpha}\in M} \theta_{\boldsymbol{\alpha}} \boldsymbol{X}^{\boldsymbol{\alpha}}\right) = \prod_{\boldsymbol{\alpha}\in M} \left(\mathrm{e}^{\theta_{\boldsymbol{\alpha}}}\right)^{\boldsymbol{X}^{\boldsymbol{\alpha}}}$$

has been considered in a number of papers on combinatorial optimization; see [3–5]. The following statements are results in algebraic statistics; *cf.* [20,35]. Let $\mathcal{P}^1 = \left\{ f \in \mathbb{R}^{\Omega} \middle| \sum_{\boldsymbol{x}\in\Omega} p(\boldsymbol{x}) = 1 \right\}$.

Proposition 6 (Implicitization of the exponential family)**.** *Given a function* $p\colon \Omega \to \mathbb{R}$*, then* $p \in \mathcal{E}$ *if, and only if, the following conditions all hold:*

1. $p(\boldsymbol{x}) > 0$, $\boldsymbol{x} \in \Omega$*;*
2. $\sum_{\boldsymbol{x}\in\Omega} p(\boldsymbol{x}) = 1$*;*
3. $\prod_{\boldsymbol{x}\colon \boldsymbol{x}^{\boldsymbol{\beta}}=1} p(\boldsymbol{x}) = \prod_{\boldsymbol{x}\colon \boldsymbol{x}^{\boldsymbol{\beta}}=-1} p(\boldsymbol{x})$ *for all* $\boldsymbol{\beta} \in L^* \setminus M$.

Proof. ($\Rightarrow$) If $p \in \mathcal{E}$, then $p(\boldsymbol{x}) > 0$, $\boldsymbol{x} \in \Omega$ (Item 1) and $\sum_{\boldsymbol{x}\in\Omega} p(\boldsymbol{x}) = 1$ (Item 2). Moreover, $\log p(\boldsymbol{x}) = \sum_{\boldsymbol{\alpha}\in M} \theta_{\alpha} \boldsymbol{x}^{\boldsymbol{\alpha}} - \psi(\boldsymbol{\theta})$. The function $\log p$ is orthogonal to each $\boldsymbol{X}^{\beta}$, $\beta \in L^* \setminus M$. Hence:

$$0 = \sum_{\boldsymbol{x}\in\Omega} \log p(\boldsymbol{x}) \boldsymbol{x}^{\boldsymbol{\beta}} = \sum_{\boldsymbol{x}\colon \boldsymbol{x}^{\boldsymbol{\beta}}=1} \log p(\boldsymbol{x}) - \sum_{\boldsymbol{x}\colon \boldsymbol{x}^{\boldsymbol{\beta}}=-1} \log p(\boldsymbol{x}) = \log \prod_{\boldsymbol{x}\colon \boldsymbol{x}^{\boldsymbol{\beta}}=1} p(\boldsymbol{x}) - \log \prod_{\boldsymbol{x}\colon \boldsymbol{x}^{\boldsymbol{\beta}}=-1} p(\boldsymbol{x}) \,, \quad (79)$$

which is equivalent to Item 3.

($\Leftarrow$) Oppositely, the computation in Equation (79) implies that $\log p$ is orthogonal to each $\boldsymbol{X}^{\boldsymbol{\beta}}$; hence, there exists $\boldsymbol{\theta}$, such that $\log p = \sum_{\boldsymbol{\alpha}\in M} \theta_{\boldsymbol{\alpha}} \boldsymbol{X}^{\boldsymbol{\alpha}} + C$. Now, Item 2 implies $C = -\psi(\boldsymbol{\theta})$.

□

Let $\mathbb{R}[\Omega]$ denote the ring of polynomials in the indeterminates $\{p(\boldsymbol{x}) | \boldsymbol{x} \in \Omega\}$. Given a binary model M, the set of polynomials:

$$\left\{ \prod_{\boldsymbol{x}\colon \boldsymbol{x}^{\boldsymbol{\beta}}=1} p(\boldsymbol{x}) - \prod_{\boldsymbol{x}\colon \boldsymbol{x}^{\boldsymbol{\beta}}=-1} p(\boldsymbol{x}) \middle| \boldsymbol{\beta} \in L^* \setminus M \right\},$$

generates an ideal $\mathcal{J}(M)$, which is called the toric ideal of the model M. Its variety $\mathcal{V}(M)$ is called the exponential variety of M.

Proposition 7.

1. *The exponential variety of* M *is the Zariski closure of the exponential model* $\mathcal{E}$*.*
2. *The closure* $\overline{\mathcal{E}}$ *of* $\mathcal{E}$ *in* $\mathcal{P}_{\geq}$ *is characterized by* $p(\boldsymbol{x}) \geq 0$, $\boldsymbol{x} \in \Omega$*, together with Items 2 and 3 of Proposition 6.*

3. *The algebraic variety of the ring* $\mathbb{R}[p(\boldsymbol{x}) \colon \boldsymbol{x} \in \Omega]$, *which is generated by the polynomials* $\sum_{\boldsymbol{x}\in\Omega} p(\boldsymbol{x}) - 1$, $\prod_{\boldsymbol{x}\colon \boldsymbol{x}^{\boldsymbol{\beta}}=1} p(\boldsymbol{x}) - \prod_{\boldsymbol{x}\colon \boldsymbol{x}^{\boldsymbol{\beta}}=-1} p(\boldsymbol{x})$, $\boldsymbol{\beta} \in L^* \setminus M$, *is an extension* $\mathcal{E}^1$ *of* $\mathcal{E}$ *to* $\mathcal{P}^1$.

4. *Define the moments* $\eta_{\boldsymbol{\alpha}} = \sum_{\boldsymbol{x}\in\Omega} \boldsymbol{x}^{\boldsymbol{\alpha}} p(\boldsymbol{x})$, $\boldsymbol{\alpha} \in L$, i.e., *the discrete Fourier transform of* p, *with inverse* $p(\boldsymbol{x}) = 2^{-n} \sum_{\boldsymbol{\alpha}\in L} \boldsymbol{x}^{\boldsymbol{\alpha}} \eta_{\boldsymbol{\alpha}}$. *There exists an algebraic extension of the moment function* $\mathcal{E} \ni p \mapsto \boldsymbol{\eta}(p) \in M^\circ$ *to a mapping defined on* $\mathcal{E}^1$.

Proof. 1. According to the implicitization Proposition 6, the exponential family is characterized by the positivity condition together with the algebraic binomial conditions.

2. This follows from the implicit form, and it is proven, for example, in [20].

3. By definition.

4. As the mapping from the probabilities to the moments is affine and one-to-one, such a transformation extends to a one-to-one mapping from the extended model to the affine space of the marginal polytope.

□

We conclude this section by introducing the so-called *no three-way interaction* example. On $\Omega = \{0,1\}^3$, the full model in the statistics $0 \mapsto 1$, $1 \mapsto -1$, that is $t = (-1)^x = 1 - 2x$, is described by the matrix:

$$D_3 = \begin{array}{c} \\ 000 \\ 001 \\ 010 \\ 011 \\ 100 \\ 101 \\ 110 \\ 111 \end{array} \begin{array}{c} \begin{array}{cccccccc} 1 & T_3 & T_2 & T_2T_3 & T_1 & T_1T_3 & T_1T_2 & T_1T_2T_3 \end{array} \\ \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{bmatrix} \end{array}. \tag{80}$$

Note the lexicographic order of both the sample points and the statistics' exponents.

The exponential family without the interaction term $T_1T_2T_3$ is the same model as the toric model without the three-way interaction, which is based on the matrix:

$$B = \begin{array}{c} \\ 000 \\ 001 \\ 010 \\ 011 \\ 100 \\ 101 \\ 110 \\ 111 \end{array} \begin{array}{c} \begin{array}{ccccccc} C & \zeta_1 & \zeta_2 & \zeta_3 & \zeta_4 & \zeta_5 & \zeta_6 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 \end{bmatrix} \end{array}. \tag{81}$$

that is the probabilities as a function of the ζ's are:

$$\begin{cases} p_1 = & c \\ p_2 = & c\zeta_1\zeta_3\zeta_5 \\ p_3 = & c\zeta_2\zeta_3\zeta_6 \\ p_4 = & c\zeta_1\zeta_2\zeta_5\zeta_6 \\ p_5 = & c\zeta_4\zeta_5\zeta_6 \\ p_6 = & c\zeta_1\zeta_3\zeta_4\zeta_6 \\ p_7 = & c\zeta_2\zeta_3\zeta_4\zeta_5 \\ p_8 = & c\zeta_1\zeta_2\zeta_4 \end{cases} . \tag{82}$$

The toric ideal of the toric model in Equation (82) is generated by the polynomial:

$$p_2p_3p_5p_8 - p_1p_4p_6p_7 = 0 \, , \tag{83}$$

this means that the closure of the exponential family is given by the solution of the equations:

$$\begin{cases} p_1 + p_2 + p_3 + p_4 + p_5 + p_6 + p_7 + p_8 = 1 \\ p_2p_3p_5p_8 - p_1p_4p_6p_7 = 0 \end{cases} . \tag{84}$$

The η parameters are the expected values of the sufficient statistics of the full model,

$$\begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \eta_4 \\ \eta_5 \\ \eta_6 \\ \eta_7 \end{bmatrix} = \begin{array}{c} \\ 001 \\ 010 \\ 011 \\ 100 \\ 101 \\ 110 \\ 111 \end{array} \begin{array}{c} \begin{array}{cccccccc} 000 & 001 & 010 & 011 & 100 & 101 & 110 & 111 \end{array} \\ \begin{bmatrix} 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{bmatrix} \end{array} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \\ p_6 \\ p_7 \\ p_8 \end{bmatrix} . \tag{85}$$

In the ring:

$$R = \mathbb{Q}[p_1, p_2, p_3, p_4, p_5, p_6, p_7, p_8, \eta_1, \eta_2, \eta_3, \eta_4, \eta_5, \eta_6, \eta_7] \tag{86}$$

we can consider the ideal $\mathcal{I}$ generated by the Equations (84) together with Equations (85). The elimination ideal:

$$\mathcal{J} = \mathcal{I} \cap \mathbb{Q}[\eta_1, \eta_2, \eta_3, \eta_4, \eta_5, \eta_6, \eta_7] \tag{87}$$

will express the model as a dependence between the η's.

Computation with CoCoA [36] gives the following polynomial:

$$\begin{aligned} &f(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5, \eta_6; \eta_7) = \\ &\eta_1^2\eta_3\eta_4 + \eta_2^2\eta_3\eta_4 - \eta_3^3\eta_4 - \eta_3\eta_4^3 + \eta_1^2\eta_2\eta_5 - \eta_2^3\eta_5 + \eta_2\eta_3^2\eta_5 + \eta_2\eta_4^2\eta_5 + \eta_3\eta_4\eta_5^2 - \eta_2\eta_5^3 - \eta_1^3\eta_6 + \eta_1\eta_2^2\eta_6 + \eta_1\eta_3^2\eta_6 \\ &+ \eta_1\eta_4^2\eta_6 + \eta_1\eta_5^2\eta_6 + \eta_3\eta_4\eta_6^2 + \eta_2\eta_5\eta_6^2 - \eta_1\eta_6^3 - 2\eta_1\eta_2\eta_4 - 2\eta_1\eta_3\eta_5 - 2\eta_2\eta_3\eta_6 - 2\eta_4\eta_5\eta_6 + \eta_3\eta_4 + \eta_2\eta_5 + \eta_1\eta_6 \\ &+ \left(-2\eta_1\eta_2\eta_3 - 2\eta_1\eta_4\eta_5 - 2\eta_2\eta_4\eta_6 - 2\eta_3\eta_5\eta_6 + \eta_1^2 + \eta_2^2 + \eta_3^2 + \eta_4^2 + \eta_5^2 + \eta_6^2 - 1\right)\eta_7 \\ &+ \left(\eta_3\eta_4 + \eta_2\eta_5 + \eta_1\eta_6\right)\eta_7^2 + (-1)\eta_7^3 \, . \end{aligned} \tag{88}$$

The equation:

$$f(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5, \eta_6; \eta_7) = 0 \tag{89}$$

is an expression of the model in the expectation parameters, and this expression is a polynomial equation. We know unique solvability in η_7 if $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5, \eta_6)$ is in the interior of the marginal polytope. As in the example of the previous section, it is possible to intersect the polynomial invariant in Equation (83) with one or more sheaves of hyperplanes around some faces of the simplex, in order to lower the degree of the invariant and thus decompose the model as the convex hull of probabilities on the boundary of the model. We do not describe the details here, and we postpone the discussion of this example to a paper which is in preparation.

5. Conclusions

Geometry and algebra play a fundamental role in the study of statistical models, and in particular in the exponential family. In the fist part of the paper, starting from the definition of the natural gradient over an exponential family, we described the relationship between its expression in the basis of the sufficient statistics and in the conjugate basis. From this perspective, the terms natural gradient and vanilla gradient, to denote gradients evaluated with respect to the Fisher and the Euclidean geometry, together with their duality in the natural and expectation parameters, assume a new meaning, since these definitions depend on the choice of the basis for the tangent space.

In order to study natural gradient flows for a generic discrete exponential model and, in particular, their convergence, it is convenient to move to the mixture geometry of the expectation parameters and to study trajectories over the marginal polytope. However, in order to obtain explicit equations for the flows, it is necessary to determine the dependence between the moments associated with the sufficient statistics of the model, which are constrained to belong to the marginal polytope, and the remaining moments, which on the other side are not free. Such a relationship, which for finite search spaces is given by a system of polynomial invariants, cannot be easily solved explicitly in general. In the second part of the paper, by using algebraic tools, we proposed a novel parameterization based on ruled surfaces for an exponential family, which does not require to solve the polynomial invariant explicitly. We applied our approach to a simple example, and we showed that the surface associated with the model in the full marginal polytope is a ruled surface. We claim that these results are not peculiar to the example we described, and we are working towards an extension of this approach in a more general case.

Acknowledgments

The authors would like to thank Gianfranco Casnati from Politecnico di Torino for the useful discussions on the geometry of ruled surfaces. Giovanni Pistone is supported by de Castro Statistics of Collegio Carlo Alberto at Moncalieri and is a member of INdAM/GNAMPA.

Author Contributions

Both authors contributed to the design of the research. The research was carried out by all of the authors. The manuscript was written by Luigi Malagò and Giovanni Pistone. Both authors read and approved the final manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Pistone, G. In Proceedings of the First International Conference (GSI 2013), Paris, France, 28–30 August 2013; Nonparametric information geometry. In *Geometric Science of Information*; Nielsen, F., Barbaresco, F., Eds.; Lecture Notes in Computer Science, Volume 8085; Springer: Heidelberg, Germany, 2013; pp. 5–36.
2. Malagò, L.; Matteucci, M.; Pistone, G. Stochastic Relaxation as a Unifying Approach in 0/1 Programming, 2009. In Proceedings of the NIPS 2009 Workshop on Discrete Optimization in Machine Learning: Submodularity, Sparsity & Polyhedra (DISCML), Whistler Resort & Spa, BC, Canada, 11–12 December 2009.
3. Malagò, L.; Matteucci, M.; Pistone, G. Towards the geometry of estimation of distribution algorithms based on the exponential family. In Proceedings of the 11th Workshop on Foundations of Genetic Algorithms (FOGA '11), Schwarzenberg, Austria, 5–8 January 2011; ACM: New York, NY, USA, 2011; pp. 230–242.
4. Malagò, L.; Matteucci, M.; Pistone, G. Stochastic Natural Gradient Descent by estimation of empirical covariances. In Proceedings of the 2011 IEEE Congress on Evolutionary Computation (CEC), New Orleans, LA, USA, 5–8 June 2011; pp. 949–956.
5. Malagò, L.; Matteucci, M.; Pistone, G. Natural gradient, fitness modelling and model selection: A unifying perspective. In Proceedings of the 2013 IEEE Congress on Evolutionary Computation (CEC), Cancun, Mexico, 20–23 June 2013; pp. 486–493.
6. Wierstra, D.; Schaul, T.; Peters, J.; Schmidhuber, J. Natural evolution strategies. In Proceedings of the 2008 IEEE Congress on Evolutionary Computation, Hong Kong, China, 1–6 June 2008; pp. 3381–3387.
7. Ollivier, Y.; Arnold, L.; Auger, A.; Hansen, N. Information-Geometric Optimization Algorithms: A Unifying Picture via Invariance Principles. **2011**, arXiv:1106.3708.
8. Malagò, L.; Pistone, G. Combinatorial Optimization with Information Geometry: Newton method. *Entropy* **2014**, *16*, 4260–4289.
9. Amari, S.; Nagaoka, H. *Methods of Information Geometry*; American Mathematical Society: Providence, RI, USA, 2000; Translated from the 1993 Japanese original by Daishi Harada.
10. Bourbaki, N. *Variétés differentielles et analytiques. Fascicule de résultats / Paragraphes 1 à 7*; Number XXXIII in Éléments de mathématiques; Hermann: Paris, France, 1971.

11. Pistone, G.; Sempi, C. An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. *Ann. Stat.* **1995**, *23*, 1543–1561.
12. Malagò, L.; Pistone, G. Gradient Flow of the Stochastic Relaxation on a Generic Exponential Family. In Proceedings of Conference of Bayesian Inference and Maximum Entropy Methods in Science and Engineering (MaxEnt 2014), Clos Lucé, Amboise, France, 21–26 September 2014; Mohammad-Djafari, A., Barbaresco, F., Eds.; pp. 353–360.
13. Brown, L.D. *Fundamentals of Statistical Exponential Families With Applications in Statistical Decision Theory*; Number 9 in IMS Lecture Notes, Monograph Series; Institute of Mathematical Statistics: Hayward, CA, USA, 1986;
14. Rockafellar, R.T. *Convex Analysis*; Princeton Mathematical Series, No. 28; Princeton University Press: Princeton, NJ, USA, 1970.
15. Do Carmo, M.P. *Riemannian Geometry*; Mathematics: Theory & Applications, Birkhäuser Boston Inc.: Boston, MA, USA, 1992; Translated from the second Portuguese edition by Francis Flaherty.
16. Amari, S.I. Natural gradient works efficiently in learning. *Neur. Comput.* **1998**, *10*, 251–276.
17. Shima, H. *The Geometry of Hessian Structures*; World Scientific Publishing Co. Pte. Ltd.: Hackensack, NJ, USA, 2007.
18. Rinaldo, A.; Fienberg, S.E.; Zhou, Y. On the geometry of discrete exponential families with application to exponential random graph models. *Electron. J. Stat.* **2009**, *3*, 446–484.
19. Rauh, J.; Kahle, T.; Ay, N. Support Sets in Exponential Families and Oriented Matroid Theory. *Int. J. Approx. Reas.* **2011**, *52*, 613–626.
20. Malagò, L.; Pistone, G. A note on the border of an exponential family. **2010**, arXiv:1012.0637v1.
21. Pistone, G.; Rogantin, M. The gradient flow of the polarization measure. With an appendix. **2015**, doi:arXiv:1502.06718.
22. Diaconis, P.; Sturmfels, B. Algebraic algorithms for sampling from conditional distributions. *Ann. Stat.* **1998**, *26*, 363–397.
23. Pistone, G.; Wynn, H.P. Generalised confounding with Gröbner bases. *Biometrika* **1996**, *83*, 653–666.
24. Pistone, G.; Riccomagno, E.; Wynn, H.P. *Algebraic Statistics: Computational Commutative Algebra in Statistics*; Volume 89, Monographs on Statistics and Applied Probability, Chapman & Hall/CRC: Boca Raton, FL, USA, 2001.
25. Drton, M.; Sturmfels, B.; Sullivant, S. *Lectures on Algebraic Statistics*; Volume 39, Oberwolfach Seminars; Birkhäuser Verlag: Basel, Germany, 2009.
26. Pachter, L., Sturmfels, B., Eds. *Algebraic Statistics for Computational Biology*; Cambridge University Press: Cambridge, UK, 2005.
27. Gibilisco, P., Riccomagno, E., Rogantin, M.P., Wynn, H.P., Eds. *Algebraic and Geometric Methods in Statistics*; Cambridge University Press: Cambridge, UK, 2010.
28. 4ti2 team. 4ti2—A software package for algebraic, geometric and combinatorial problems on linear spaces. Available online: http://www.4ti2.de (accessed on 2 June 2015).
29. Michałek, M.; Sturmfels, B.; Uhler, C.; Zwiernik, P. Exponential Varieties. **2014**, arXiv:1412.6185.
30. Sturmfels, B. *Gröbner Bases and Convex Polytopes*; American Mathematical Society: Providence, RI, USA, 1996.

31. Geiger, D.; Meek, C.; Sturmfels, B. On the toric algebra of graphical models. *Ann. Stat.* **2006**, *34*, 1463–1492.
32. Rapallo, F. Toric statistical models: Parametric and binomial representations. *Ann. Inst. Stat. Math.* **2007**, *59*, 727–740.
33. Beltrametti, M.; Carletti, E.; Gallarati, D.; Monti Bragadin, G. *Lectures on Curves, Surfaces and Projective Varieties: A Classical View of Algebraic Geometry*; EMS textbooks in mathematics; European Mathematical Society: Zürich, Switzerland, 2009.
34. Rinaldo, A.; Fienberg, S.E.; Zhou, Y. On the geometry of discrete exponential families with application to exponential random graph models. *Electron. J. Stat.* **2009**, *3*, 446–484.
35. Pistone, G. Algebraic varieties *vs.* differentiable manifolds in statistical models. In *Algebraic and Geometric Methods in Statistics*; Gibilisco, P., Riccomagno, E., Rogantin, M., Wynn, H.P., Eds.; Cambridge University Press: Cambridge, UK, 2009; Chapter 21, pp. 339–363.
36. Abbott, J.; Bigatti, A.; Lagorio, G. CoCoA-5: A system for doing Computations in Commutative Algebra. Available online: http://cocoa.dima.unige.it (accessed on 2 June 2015).