Gradient Flow of the Stochastic Relaxation on a Generic Exponential Family

Luigi Malagò*,† and Giovanni Pistone**,‡

*Collegio Carlo Alberto, Via Real Collegio 30, 10024 Moncalieri, Italy.
Current affiliation: Shinshu University, 4-17-1 Wakasato, Nagano 380-8553 , Japan
†malago@shinshu-u.ac.jp
**Collegio Carlo Alberto, Via Real Collegio 30, 10024 Moncalieri, Italy
‡giovanni.pistone@carloalberto.org

Abstract. We study the natural gradient flow of the expected value $\mathrm{E}_p[f]$ of an objective function f for p in an exponential family. We parameterize the exponential family with the expectation parameters and we show that the dynamical system associated to the natural gradient flow can be extended outside the marginal polytope.

Keywords: Information Geometry, Stochastic Relaxation, Natural Gradient Flow.
PACS: 89.20.Ff

1. GRADIENT FLOW OF RELAXED OPTIMIZATION

Let $(\Omega, \mathscr{A}, \mu)$ be a measured space of samples $\boldsymbol{x} \in \Omega$, $\mathscr{P}_\geq$ the simplex of (probability) densities, $\mathscr{P}_> \subset \mathscr{P}_\geq$ the open simplex of strictly positive densities. For a bounded objective function $f\colon \Omega \to \mathbb{R}$ and a statistical model $\mathscr{M} \subset \mathscr{P}_>$, the (stochastic) *relaxation* of f to $\mathscr{M}$ is the function $F(p) = \mathrm{E}_p[f] \in \mathbb{R}$, $p \in \mathscr{M}$, cf. [1]. The minimization of the stochastic relaxation has been studied by many authors [2, 3, 4, 5, 6]. If we have a parameterization $\boldsymbol{\xi} \mapsto p_{\boldsymbol{\xi}}$ of $\mathscr{M}$, 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 $\partial_j \hat{F}(\boldsymbol{\xi}) = \mathrm{E}_{p_{\boldsymbol{\xi}}}\left[\partial_j \log\left(p_{\boldsymbol{\xi}}\right) f\right]$ and $\mathrm{E}_{p_{\boldsymbol{\xi}}}\left[\partial_j \log\left(p_{\boldsymbol{\xi}}\right)\right] = 0$, see [7, 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* $\partial_j \log\left(p_{\boldsymbol{\xi}}\right)$ 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 set-up, in the non finite state space some care has to be taken. Here we follow the set-up discussed in [7] and, in particular, exponential families. Full details are given only in the simplest cases but we claim general applicability of our methods.

We discuss in this Section the finite state space case, while the next Section is devoted to the binary case. Let Ω be a finite set of points $\boldsymbol{x} = (x_1, \dots, x_n)$ and μ the counting measure. In this case a density is a probability function. Given a set $\mathscr{B} = \{T_1, \dots, T_d\}$ of *affinely independent* random variables, we consider the statistical model $\mathscr{E}$ whose elements are uniquely identified by the natural parameters $\boldsymbol{\theta}$ in the exponential family with sufficient statistics $\mathscr{B}$, namely $p_{\boldsymbol{\theta}} \in \mathscr{E}$ if $\log p_{\boldsymbol{\theta}} = \sum_{i=1}^{d} \theta_i T_i - \psi(\boldsymbol{\theta})$, $\theta \in \mathbb{R}^d$, see [12].

The convex function $\boldsymbol{\theta} \mapsto \psi(\boldsymbol{\theta}) = \log \sum_{\boldsymbol{x} \in \Omega} \mathrm{e}^{\boldsymbol{\theta} \cdot \boldsymbol{T}} = \boldsymbol{\theta} \cdot \mathrm{E}_{p_{\boldsymbol{\theta}}}[\boldsymbol{T}] - \mathrm{E}_{p_\theta}[\log(p_\theta)]$ is the cumulant generating function of the sufficient statistics, in particular, $\nabla \psi(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}}[\boldsymbol{T}]$ and $\operatorname{Hess} \psi(\boldsymbol{\theta}) = \operatorname{Cov}_{\boldsymbol{\theta}}(\boldsymbol{T}, \boldsymbol{T})$. It follows that the entropy of p_θ is $H(p_{\boldsymbol{\theta}}) = -\mathrm{E}_{p_\theta}[\log(p_{\boldsymbol{\theta}})] = \psi(\boldsymbol{\theta}) - \boldsymbol{\theta} \cdot \nabla \psi(\boldsymbol{\theta})$.

The mapping $\nabla \psi$ is 1-to-1 onto the interior M° of the *marginal polytope*, i.e. the convex set generated by the values $\boldsymbol{T}(\boldsymbol{x}) \in \mathbb{R}^d$, $\boldsymbol{x} \in \Omega$, see [12]. Convex conjugation applies, see [13, §25]. The Legendre conjugate $\phi\colon M^\circ$ of ψ is such that $\nabla \phi = (\nabla \psi)^{-1}$ and it provides an alternative parameterization of $\mathscr{E}$ with $\boldsymbol{\eta} = \nabla \psi(\boldsymbol{\theta})$,

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

In the $\boldsymbol{\theta}$-parameters the entropy is $H(p_{\boldsymbol{\theta}}) = -\mathrm{E}_{p_{\boldsymbol{\theta}}}[\log p_{\boldsymbol{\theta}}] = \psi(\boldsymbol{\theta}) - \boldsymbol{\theta} \cdot \nabla \psi(\boldsymbol{\theta})$; in the $\boldsymbol{\eta}$-parameters the entropy is $H(p_{\boldsymbol{\eta}}) = -\mathrm{E}_{p_{\boldsymbol{\eta}}}[\log p_{\boldsymbol{\eta}}] = -\phi(\boldsymbol{\eta})$.

Derivation of the equality $\nabla \phi = (\nabla \psi)^{-1}$ gives $\operatorname{Hess} \phi(\boldsymbol{\eta}) = \operatorname{Hess} \psi(\boldsymbol{\theta})^{-1}$ when $\boldsymbol{\eta} = \nabla \psi(\boldsymbol{\theta})$.

While $\mathscr{E}$ is an elementary manifold in either the $\boldsymbol{\theta}$- or the $\boldsymbol{\eta}$-parameterization, the definition of the tangent bundle $T\mathscr{E}$ requires some care. If $I \ni t \mapsto p_t$ is a curve in $\mathscr{E}$, then we identify the velocity vector with the Fisher score $\frac{d}{dt} \log(p_t)$. In the expression of the curve by the $\boldsymbol{\theta}$ parameters the velocity is

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

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 $p \in \mathscr{E}$ with the vector space of centered sufficient statistics, that is $T_p\mathscr{E} = \operatorname{Span}\left(T_j - \mathrm{E}_p[T_j] \middle| j = 1, \dots, d\right)$.

In the $\boldsymbol{\eta}$-parameterization of (1) the computation of the velocity is

$$\begin{aligned} \frac{d}{dt} \log(p_t) &= \frac{d}{dt}(\nabla \phi(\boldsymbol{\eta}(t)) \cdot (\boldsymbol{T} - \boldsymbol{\eta}(t)) + \phi(\boldsymbol{\eta}(t))) = \\ &(\operatorname{Hess} \phi(\boldsymbol{\eta}(t)) \dot{\boldsymbol{\eta}}(t)) \cdot (\boldsymbol{T} - \boldsymbol{\eta}(t)) = \dot{\boldsymbol{\eta}}(t) \cdot [\operatorname{Hess} \phi(\boldsymbol{\eta}(t))(\boldsymbol{T} - \boldsymbol{\eta}(t))]. \end{aligned} \tag{3}$$

The last equality provides the interpretation of $\dot{\boldsymbol{\eta}}(t)$ as the coordinate of the velocity in the *conjugate* vector basis $\operatorname{Hess} \phi(\boldsymbol{\eta}(t))(\boldsymbol{T} - \boldsymbol{\eta}(t))$, that is the basis of derivatives along the $\boldsymbol{\eta}$ coordinates. In conclusion, the first order geometry is characterized as follows.

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

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

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

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

In other words, the mapping $(\operatorname{Hess}\phi(\boldsymbol{\eta}))^{-1}$ maps the coordinates $\boldsymbol{v}$ of a tangent vector $V \in T_{p_{\boldsymbol{\eta}}}\mathscr{E}$ with respect to the basis of centered sufficient statistics to the coordinates $\boldsymbol{v}^*$ with respect to the conjugate basis. In the $\boldsymbol{\theta}$-parameters the transformation is $\boldsymbol{v} \mapsto \boldsymbol{v}^* = \operatorname{Hess}\psi(\boldsymbol{\theta})\boldsymbol{v}$.

The explicit construction of the tangent bundle together with its parallel transports is unavoidable when considering the second order calculus as it was done in [7, 8]. 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.

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

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

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

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

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

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

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

□

For each C^1 real function $F\colon \mathscr{E} \to \mathbb{R}$, the derivative along a C^1 curve $I \mapsto p(t)$, $p = p(0)$, is of the form

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

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 $V = \frac{d}{dt}\log(p(t))\big|_{t=0} = \dot{\boldsymbol{\theta}}(0) \cdot (\boldsymbol{T} - \nabla\psi(\boldsymbol{\theta}(0))$,

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

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 $V = \frac{d}{dt}\log(p(t))\big|_{t=0} = \dot{\boldsymbol{\eta}}(0) \cdot \operatorname{Hess}\phi(\boldsymbol{\eta}(0))(\boldsymbol{T} - \boldsymbol{\eta}(0))$,

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

Definition 2 (Gradients)**.**

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

Given a vector field of $\mathscr{E}$, i.e. a mapping $G \colon \mathscr{E}$ such that $G(p) \in T_p\mathscr{E}$, an integral curve from p is a curve $I \ni t \mapsto p(t)$ such that $p(0) = p$ and $\frac{d}{dt}\log(p(t)) = G(p(t))$. 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 3 (Gradient flow)**.** *The* gradient flow *of the real function $F \colon \mathscr{E}$ is the flow of the differential equation $\frac{d}{dt}\log(p(t)) = \nabla F(p(t))$. 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 $\dot{\boldsymbol{\eta}}(t) = \widetilde{\nabla}\check{F}(\boldsymbol{\eta}(t))$.*

We are going to focus on the expression of the gradient flow in the $\boldsymbol{\eta}$ parameters. As $\widetilde{\nabla}\check{F}(\boldsymbol{\eta}) = \operatorname{Hess}\phi(\boldsymbol{\eta})^{-1}\nabla\check{F}(\boldsymbol{\eta}) = \operatorname{Hess}\psi(\nabla\phi(\boldsymbol{\eta}))\nabla\check{F}(\boldsymbol{\eta}) = I(p_{\boldsymbol{\eta}})\nabla\check{F}(\boldsymbol{\eta})$, in some cases we can naturally consider the extension of the equation outside M°. One notable case is when the function F is a relaxation of a non constant state space function f.

Proposition 2. *If $f \colon \Omega \to \mathbb{R}$ and $F(p) = \mathrm{E}_p[f]$ is its relaxation on $\mathscr{E}$, then $\nabla F(p)$ is the least square projection of f onto $T_p\mathscr{E}$, that is $I(p)^{-1}\operatorname{Cov}_p(f, \boldsymbol{T}) \cdot (\boldsymbol{T} - \mathrm{E}_p[\boldsymbol{T}])$. The expression in $\boldsymbol{\theta}$ are $\widetilde{\nabla}\hat{F}(\boldsymbol{\theta}) = (\operatorname{Hess}\psi(\boldsymbol{\theta}))^{-1}\operatorname{Cov}_{\boldsymbol{\theta}}(f, \boldsymbol{T})$, $\nabla\hat{F}(\boldsymbol{\theta}) = \operatorname{Cov}_{\boldsymbol{\theta}}(f, \boldsymbol{T})$. The expressions in $\boldsymbol{\eta}$ are $\widetilde{\nabla}\check{F}(\boldsymbol{\eta}) = \operatorname{Cov}_{\boldsymbol{\eta}}(f, \boldsymbol{T})$ and $\nabla\check{F}(\boldsymbol{\eta}) = \operatorname{Hess}\phi(\boldsymbol{\eta})\operatorname{Cov}_{\boldsymbol{\eta}}(f, \boldsymbol{T})$.*

Proof. On a generic curve thought 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 $V \in T_p\mathscr{E}$ we can orthogonally project f to get $\langle \nabla F, V \rangle_p = \left\langle (I^{-1}(p)\operatorname{Cov}_p(f, \boldsymbol{T})) \cdot (\boldsymbol{T} - \mathrm{E}_p[\boldsymbol{T}], V \right\rangle_p$. □

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 $\bar{p} \notin \mathscr{E}$, and it is proved in [14, Th. 1] that its support $F \subset \Omega$ is *exposed*, that is $\boldsymbol{T}(F)$ 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 $\boldsymbol{a} \cdot \mathrm{Cov}_{\bar{p}}(f, \boldsymbol{T}) = 0$. If we take the mapping $\boldsymbol{\eta} \mapsto \mathrm{Cov}_{\boldsymbol{\eta}}(f, \boldsymbol{T})$ to be the limit of the vector field of the gradient on the faces of the marginal polytope, we see that such a vector field is tangent to the faces. This remark is further elaborated below in the binary case.

2. PSEUDO-BOOLEAN OBJECTIVE FUNCTIONS

We turn to the case of binary variables, $\boldsymbol{x} = (x_1, \ldots, x_n) \in \{+1, -1\}^n = \Omega$. 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 $M \subset L^* = L \setminus \{\boldsymbol{0}\}$, the model where $p \in \mathscr{E}$ 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 [2, 3, 4]. The following are results in Algebraic Statistics, cf. [15, 14]. Let $\mathscr{P}^1 = \left\{f \in \mathbb{R}^\Omega \middle| \sum_{\boldsymbol{x} \in \Omega} p(\boldsymbol{x}) = 1\right\}$.

Proposition 3. *Given $p \in \mathbb{R}^\Omega$, then $p \in \mathscr{E}$ holds if, and only if,*

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* $\boldsymbol{\beta} \in L^* \setminus M$.

The following proposition is given here without proof. It is intended to motivate the example of Fig. 1.

Proposition 4. *1. The closure $\overline{\mathscr{E}}$ of $\mathscr{E}$ in $\mathscr{P}_{\geq}$ is characterized by $p(\boldsymbol{x}) \geq 0$, $\boldsymbol{x} \in \Omega$, together with items 2 and 3 of Prop. 3.*

2. *The algebraic variety of the ring $\mathbb{R}[p(\boldsymbol{x}) \colon \boldsymbol{x} \in \Omega]$ 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 $\mathscr{E}^1$ of $\mathscr{E}$ to $\mathscr{P}^1$.*
3. *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 $\mathscr{E} \ni p \mapsto \boldsymbol{\eta}(p) \in M^\circ$ to a mapping defined on $\mathscr{E}^1$.*

Example. If $T_i(\boldsymbol{x}) = x_i$, for $i = 1, \ldots, n$, the model $\mathscr{E}$ consists of the interior of the independence model, that is, all positive probability distributions $p(\boldsymbol{x})$ over Ω that factorize as the product of marginal probabilities, i.e., $p(x; \boldsymbol{\theta}) = \prod_{i=1}^n p_i(x_i; \theta_i)$. In terms of the moments defined in Prop. 4.3 we have $F(2^{-n} \sum_{\boldsymbol{\alpha} \in L} \boldsymbol{x}^{\boldsymbol{\alpha}} \eta_{\boldsymbol{\alpha}}) = \sum_{\boldsymbol{\alpha} \in L} a_{\boldsymbol{\alpha}} \eta_{\boldsymbol{\alpha}}$. Not that this is *not* the $\boldsymbol{\eta}$ parameterization of the model, but it is a parameterization on the full $\mathscr{P}_{>}$, constrained by Prop. 4.3. In this case we obtain easily the $\boldsymbol{\eta}$-parameterization from the independence, $\check{F}(\boldsymbol{\eta}) = \sum_{\boldsymbol{\alpha} \in L} a_{\boldsymbol{\alpha}} \boldsymbol{\eta}^{\boldsymbol{\alpha}}$, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)$, $\boldsymbol{\eta}^{\boldsymbol{\alpha}} = \prod_{i=1}^n \eta_i^{\alpha_i}$.

Let $\boldsymbol{\beta}_i \in \{0,1\}^n$ be the vector such that $\beta_j = 1$ for $j = i$, and 0 otherwise. We denote with $\otimes$ the bitwise XOR. Let i denote the i-component of the gradient vector, and i, j the indices of I, follows that

$$\nabla \check{F}(\boldsymbol{\eta}) = \sum_{\boldsymbol{\alpha}\in L:\alpha_i=1} a_{\boldsymbol{\alpha}} \boldsymbol{\eta}^{\boldsymbol{\alpha}\otimes\boldsymbol{\beta}_i} \;, \tag{12}$$

$$I(\boldsymbol{\eta})^{-1} = \mathrm{Cov}_{\boldsymbol{\eta}}(X_i,X_j) = \mathbb{E}_{\boldsymbol{\eta}}[X_iX_j] - \eta_i\eta_j = \mathrm{diag}(1-\eta_i^2) \;, \tag{13}$$

$$\widetilde{\nabla}\check{F}(\boldsymbol{\eta}) = (1-\eta_i^2) \sum_{\boldsymbol{\alpha}\in L:\alpha_i=1} a_{\boldsymbol{\alpha}} \boldsymbol{\eta}^{\boldsymbol{\alpha}\otimes\boldsymbol{\beta}_i} \;. \tag{14}$$

Similarly we have

$$\begin{aligned}\widetilde{\nabla}\check{F}(\boldsymbol{\eta}) &= \mathrm{Cov}_{\boldsymbol{\eta}}(f,X_i) = \sum_{\boldsymbol{\alpha}\in L} a_{\boldsymbol{\alpha}} \mathrm{Cov}_{\boldsymbol{\eta}}(\boldsymbol{X}^{\boldsymbol{\alpha}},X_i) = \sum_{\boldsymbol{\alpha}\in L} a_{\boldsymbol{\alpha}}(\mathbb{E}_{\boldsymbol{\eta}}[\boldsymbol{X}^{\boldsymbol{\alpha}}X_i] - \mathbb{E}_{\boldsymbol{\eta}}[\boldsymbol{X}^{\boldsymbol{\alpha}}]\eta_i) \\ &= \sum_{\boldsymbol{\alpha}\in L:\alpha_i=1} a_{\boldsymbol{\alpha}}(\boldsymbol{\eta}^{\boldsymbol{\alpha}\otimes\boldsymbol{\beta}_i} - \boldsymbol{\eta}^{\boldsymbol{\alpha}}\eta_i) = (1-\eta_i^2) \sum_{\boldsymbol{\alpha}\in L:\alpha_i=1} a_{\boldsymbol{\alpha}} \boldsymbol{\eta}^{\boldsymbol{\alpha}\otimes\boldsymbol{\beta}_i}\end{aligned} \tag{15}$$

It is easy to show that the natural gradient vanish over the vertexes of the hypercube $[-1,+1]^n$, and that it is orthogonal to its exposed facets, i.e., trajectories with initial condition in M remain in M.

Example. We now study a toy example with $n=2$, which allows us to represent natural gradient flows in the (η_1,η_2) plane. Consider a vector of two binary variables $\boldsymbol{x}=(x_1,x_2)$, and let $f = a_0 + a_1x_1 + a_2x_2 + a_{12}x_1x_2$, where $\boldsymbol{a}$ is a vector or real numbers. For a given initial state, gradient flows are given by the solutions of the following differential equations

$$\dot{\eta}_1 = (1-\eta_1^2)(a_1 + a_{12}\eta_2) \;, \qquad \dot{\eta}_2 = (1-\eta_2^2)(a_2 + a_{12}\eta_1) \;. \tag{16}$$

Every vertex of the marginal polytope $[-1,+1]^2$ is a critical point. In order to evaluate the nature of each critical point, we look at the eigenvalues of the Jacobian given by the partial derivatives of (16) evaluated at the vertices, cf. [16]. Let $\boldsymbol{v}\in\{-1,+1\}^2$ be a vertex of $[-1,+1]^2$. The eigenvalues of the Jacobian in $\boldsymbol{v}=(v_1,v_2)$ are given by

$$\lambda_1 = -2v_1(a_1 + a_{12}v_2) \;, \qquad \lambda_2 = -2v_2(a_2 + a_{12}v_1) \;. \tag{17}$$

In the following we suppose $a_1,a_2 \neq 0$. If $a_{12}=0$, or if $a_{12}\neq 0$ and $|a_{12}|<|a_1| \vee |a_{12}| < |a_2|$, then there exists exactly one vertex which is a stable node where $\lambda_1,\lambda_2<0$, and another vertex which is an unstable node, where $\lambda_1,\lambda_2>0$. The two remaining vertices are saddle points. Similarly, it is easy to verify that when $|a_{12}|>|a_1| \wedge |a_{12}|>|a_2|$ two vertices are stable nodes, one local optimum and one global optimum for $\check{F}(\boldsymbol{\eta})$, and the remaining two are saddle points.

Moreover, for $a_{12}\neq 0$ there exists a critical point at $\boldsymbol{c}=(c_1,c_2)=(-a_2/a_{12},-a_1/a_{12})$. The Jacobian matrix evaluated at $\boldsymbol{c}$ has trace equal to zero, and eigenvalues given by

$$\lambda_{1,2} = \pm\sqrt{(a_{12}^2-a_1^2)(a_{12}^2-a_2^2)/a_{12}^2} \;. \tag{18}$$

Follows that for $|a_{12}|<|a_1| \veebar |a_{12}|<|a_2|$, i.e., $|c_1|>1 \veebar |c_2|>1$, we have complex eigenvalues, $\boldsymbol{c}$ is a center and the flows correspond to periodic trajectories. For $|c_1| =$

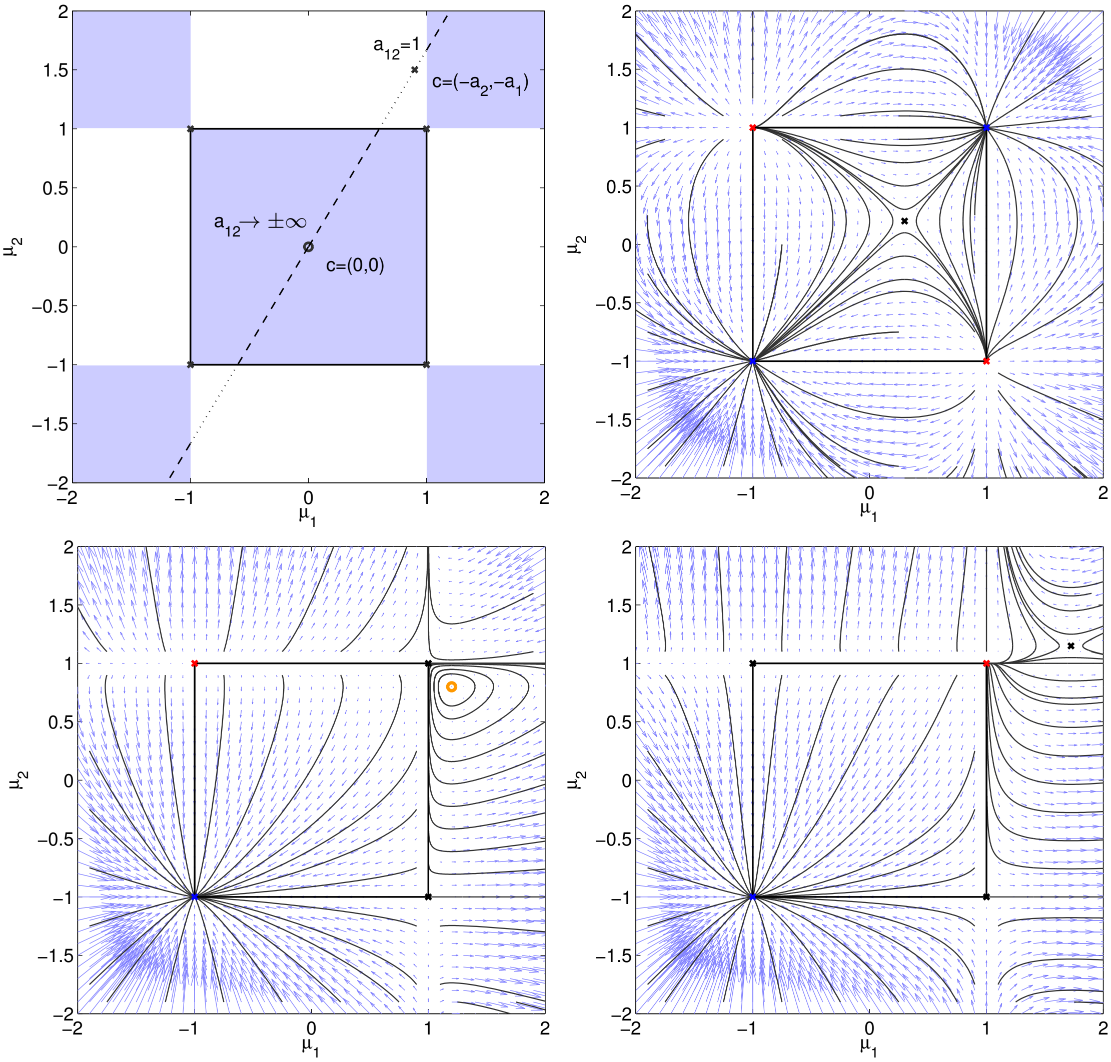

FIGURE 1. (top left) Projection of the bifurcation diagram (η_1, η_2, a_{12}) onto (η_1, η_2) for fixed a_1 and a_2, and free a_{12}. The dashed and dotted lines show the position of the critical point $\boldsymbol{c}$ as a function of a_{12}. The $[-1,+1]^2$ square corresponds to the marginal polytope. In the shaded regions the critical point is a saddle point, in the white regions there are infinite periodic trajectories. In the remaining figures we represent negative natural gradient fields and flows for fixed $a_1 = 1$ and $a_1 = 1.5$, and different values of a_{12}: (top right) $a_{12} = 5$; (bottom left) $a_{12} = 1.25$; (bottom right) $a_{12} = 0.87$. Stable nodes are represented in blue, unstable nodes in red, saddle points in black, and centers in orange.

$1 \veebar |c_2| = 1$, i.e., when $\boldsymbol{c}$ belongs to the boundary of the model, $\boldsymbol{c}$ is unstable. In the remaining cases $\boldsymbol{c}$ is a saddle point. For $a_{12} \to \pm\infty$, $\boldsymbol{c}$ tends to the center of M. In Fig. 1 we represented the projection of the bifurcation diagram (η_1, η_2, a_{12}) onto (η_1, η_2) parameterized by a_{12}, for fixed a_1 and a_2, together with negative gradient flows over (η_1, η_2) for different values of a_{12}. We represent negative gradient flows since we are interested in the minimization of F.

Example. We now consider the case of the full saturated model identified by all $2^n - 1$ the monomials $\{\boldsymbol{x}^{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in L^*\}$ as sufficient statistics. This model consists of all the distributions in the interior of the probability simplex Δ. It follows that

$$\nabla \check{F}(\boldsymbol{\eta}) = \boldsymbol{a} \,, \tag{19}$$

$$I(\boldsymbol{\eta})^{-1} = \mathrm{Cov}_{\boldsymbol{\eta}}(X_{\boldsymbol{\alpha}}, X_{\boldsymbol{\beta}}) = [\mathbb{E}_{\boldsymbol{\eta}}[X_{\boldsymbol{\alpha}} X_{\boldsymbol{\beta}}] - \eta_{\boldsymbol{\alpha}} \eta_{\boldsymbol{\beta}}] = [\eta_{\boldsymbol{\alpha} \otimes \boldsymbol{\beta}} - \eta_{\boldsymbol{\alpha}} \eta_{\boldsymbol{\beta}}] \,, \tag{20}$$

$$\widetilde{\nabla} \check{F}(\boldsymbol{\eta}) = [\eta_{\boldsymbol{\alpha} \otimes \boldsymbol{\beta}} - \eta_{\boldsymbol{\alpha}} \eta_{\boldsymbol{\beta}}] \boldsymbol{a} \,. \tag{21}$$

As in the case of the independence model, it is easy to show that the natural gradient $\widetilde{\nabla} \check{F}(\boldsymbol{\eta})$ vanishes over every vertex of the probability simplex, and that the trajectories associated to the gradient flow in Δ never leave the probability simplex.

Acknowledgments. L. Malagò was partially supported by de Castro Statistics, Collegio Carlo Alberto, Moncalieri Italy. G. Pistone is supported by de Castro Statistics and is a member of INdAM/GNAMPA.

REFERENCES

1. L. Malagò, M. Matteucci, and G. Pistone, Stochastic relaxation as a unifying approach in 0/1 programming (2009), NIPS 2009 Workshop on Discrete Optimization in Machine Learning: Submodularity, Sparsity & Polyhedra (DISCML), December 11-12, 2009, Whistler Resort & Spa, Canada.
2. L. Malagò, M. Matteucci, and G. Pistone, "Towards the geometry of estimation of distribution algorithms based on the exponential family," in *Proceedings of the 11th workshop on Foundations of genetic algorithms*, FOGA '11, ACM, New York, NY, USA, 2011, pp. 230–242.
3. L. Malagò, M. Matteucci, and G. Pistone, "Stochastic Natural Gradient Descent by estimation of empirical covariances.," in *IEEE Congress on Evolutionary Computation*, IEEE, 2011, pp. 949–956.
4. L. Malagò, M. Matteucci, and G. Pistone, "Natural gradient, fitness modelling and model selection: A unifying perspective," in *IEEE Congress on Evolutionary Computation*, IEEE, 2013, pp. 486–493.
5. D. Wierstra, T. Schaul, J. Peters, and J. Schmidhuber, "Natural evolution strategies," in *IEEE Congress on Evolutionary Computation*, 2008, pp. 3381–3387.
6. Y. Ollivier, L. Arnold, A. Auger, and N. Hansen, Information-Geometric Optimization Algorithms: A Unifying Picture via Invariance Principles (2011v1; 2013v2), arXiv:1106.3708.
7. G. Pistone, "Nonparametric Information Geometry," in *Geometric Science of Information*, edited by F. Nielsen, and F. Barbaresco, LNCS 8085, Springer-Verlag, Berlin Heidelberg, 2013, pp. 5–36, first International Conference, GSI 2013 Paris, France, August 28-30, 2013 Proceedings.
8. L. Malagò, and G. Pistone, *Entropy* **16**, 4260–4289 (2014).
9. S. Amari, and H. Nagaoka, *Methods of information geometry*, American Mathematical Society, Providence, RI, 2000, translated from the 1993 Japanese original by Daishi Harada.
10. N. Bourbaki, *Variétés differentielles et analytiques. Fascicule de résultats / Paragraphes 1 à 7*, Éléments de mathématiques XXXIII, Hermann, Paris, 1971.
11. G. Pistone, and C. Sempi, *Ann. Statist.* **23**, 1543–1561 (1995), ISSN 0090-5364.
12. L. D. Brown, *Fundamentals of statistical exponential families with applications in statistical decision theory*, IMS Lecture Notes. Monograph Series 9, Institute of Mathematical Statistics, 1986.
13. R. T. Rockafellar, *Convex analysis*, Princeton Mathematical Series, No. 28, Princeton University Press, Princeton, N.J., 1970.
14. L. Malagò, and G. Pistone, A note on the border of an exponential family (2010), arXiv:1012.0637v1.
15. G. Pistone, "Algebraic varieties vs. differentiable manifolds in statistical models," in *Algebraic and Geometric Methods in Statistics*, edited by P. Gibilisco, E. Riccomagno, M. Rogantin, and H. P. Wynn, Cambridge University Press, 2009, chap. 21, pp. 339–363.
16. S. H. Strogatz, *Nonlinear Dynamics And Chaos: With Applications To Physics, Biology, Chemistry, And Engineering (Studies in Nonlinearity)*, Studies in nonlinearity, Westview Press, 2001.