Entropy **2014**, *16*, 4260-4289; doi:10.3390/e16084260

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

Article

Combinatorial Optimization with Information Geometry: The Newton Method

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

[1] Dipartimento di Informatica, Università degli Studi di Milano, Via Comelico, 39/41, 20135 Milano, Italy; E-Mail: malago@di.unimi.it

[2] de Castro Statistics, Collegio Carlo Alberto, Via Real Collegio 30, 10024 Moncalieri, Italy

* Author to whom correspondence should be addressed; E-Mail: giovanni.pistone@carloalberto.org; Tel.: +39-011-670-5033; Fax: +39-011-670-5082.

Received: 31 March 2014; in revised form: 10 July 2014 / Accepted: 11 July 2014 / Published: 28 July 2014

Abstract: We discuss the use of the Newton method in the computation of $\max(p \mapsto \mathbb{E}_p[f])$, where p belongs to a statistical exponential family on a finite state space. In a number of papers, the authors have applied first order search methods based on information geometry. Second order methods have been widely used in optimization on manifolds, e.g., matrix manifolds, but appear to be new in statistical manifolds. These methods require the computation of the Riemannian Hessian in a statistical manifold. We use a non-parametric formulation of information geometry in view of further applications in the continuous state space cases, where the construction of a proper Riemannian structure is still an open problem.

Keywords: statistical manifold; Riemannian Hessian; combinatorial optimization; Newton method

1. Introduction

In this paper, statistical exponential families [1] are thought of as differentiable manifolds along the approach called information geometry [2] or the exponential statistical manifold [3]. Specifically, our aim is to discuss optimization on statistical manifolds using the Newton method, as is suggested in ([4] (Ch. 5 and 6)); see also the monograph [5]. This method is based on classical Riemannian geometry [6], but here, we put our emphasis on coordinate-free differential geometry; see [7,8].

We mainly refer to the above-mentioned references [2,4], with one notable exception in the description of the tangent space. Our manifold will be an exponential family $\mathcal{E}_V$ of positive densities, V being a vector space of sufficient statistics. Given a one-dimensional statistical model $p(t) \in \mathcal{E}_V$, $t \in I$, we define its velocity at time t to be its Fisher score $s(t) = \frac{d}{dt} \ln p(t)$ [9]. The Fisher score $s(t)$ is a random variable with zero expectation with respect to $p(t)$, $\mathbb{E}_{p(t)}[s(t)] = 0$. Because of that, the tangent space at $p \in \mathcal{E}_V$ is a vector space of random variables with zero expectation at p. A vector field is a mapping from p to a random variable $V(p)$, such that for all $p \in \mathcal{E}$, the random variable $V(p)$ is centered at p, $\mathbb{E}_p[V(p)] = 0$. In other words, each point of the manifold has a different tangent space, and this tangent space can be used as a non-parametric model space of the manifold. In this formalism, a vector field is a mapping from densities to centered random variables, that is, it is what in statistics is called a pivot of the statistical model. To avoid confusion with the product of random variables, we do not use the standard notation for the action of a vector field on a real function. This approach is possibly unusual in differential geometry, but it is fully natural from the statistical point of view, where the Fisher score has a central place. Moreover, this approach scales nicely from the finite state space to the general state space; see the discussion in [9] and the review in [3].

A complete construction of the geometric framework based on the idea of using the Fisher scores as elements of the tangent bundle has been actually worked out. In this paper, we go on by considering a second order geometry based on the non-parametric settings.

Our main motivation for such a geometrical construction is its application to combinatorial optimization using exponential families, whose first order version was developed in [10–14]. We give here an illustration of the methods in the following toy example.

Consider the function $f(x_1, x_2) = a_0 + a_1 x_1 + a_2 x_2 + a_{12} x_1 x_2$, with $x_1, x_2 = \pm 1$, $a_0, a_1, a_2, a_{12} \in \mathbb{R}$. The function f is a real random variable on the sample space $\Omega = \{+1, -1\}^2$ with the uniform probability λ. Note that the coordinate mappings X_1, X_2 of Ω generate an orthonormal basis $1, X_1, X_2, X_1 X_2$ of $L^2(\Omega, \lambda)$ and that f is the general form of a real random variable on such a space. Let $\mathcal{P}_>$ be the open simplex of positive densities on (Ω, λ), and let $\mathcal{E}_V$ be a statistical model, *i.e.*, a subset of $\mathcal{P}_>$. The relaxed mapping $F \colon \mathcal{E}_V \to \mathbb{R}$,

$$F(p) = \mathbb{E}_p[f] = a_0 + a_1 \mathbb{E}_p[X_1] + a_2 \mathbb{E}_p[X_2] + a_{12} \mathbb{E}_p[X_1 X_2], \tag{1}$$

is strictly bounded by the maximum of f, $F(p) = \mathbb{E}_p[f] < \max_{x \in \Omega} f(x)$, unless f is constant. We are looking for a sequence p_n, $n \in \mathbb{N}$, such that $\mathbb{E}_{p_n}[f] \to \max_{x \in \Omega} f(x)$ as $n \to \infty$. The existence of such a sequence is a nontrivial condition for the model $\mathcal{E}$. Precisely, the closure of $\mathcal{E}_V$ must contain a density, whose support is contained in the set of maxima $\{x \in \Omega | f(x) = \max f\}$. This condition is satisfied by the independence model, $V = \mathrm{Span}\{X_1, X_2\}$, where we can write:

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

See Figure 1.

The gradient of Equation (2) has components $\partial_1 F = a_1 + a_{12} \eta^2$, $\partial_2 F = a_2 + a_{12} \eta^1$, and the flow along the gradient produces increasing values for F; however, the gradient flow does not converge to the maximum of F; see the dotted line in Figure 2. However, one can follow the suggestion by [15] and use a modified gradient (the "natural" gradient) flow that produces better results in our problem; see Figure 3. Full details on this example are given in Section 2.5.2.

Figure 1. Relaxation of the Function (2) on the independence model. $a_1 = 1$, $a_2 = 2$, $a_{12} = 3$.

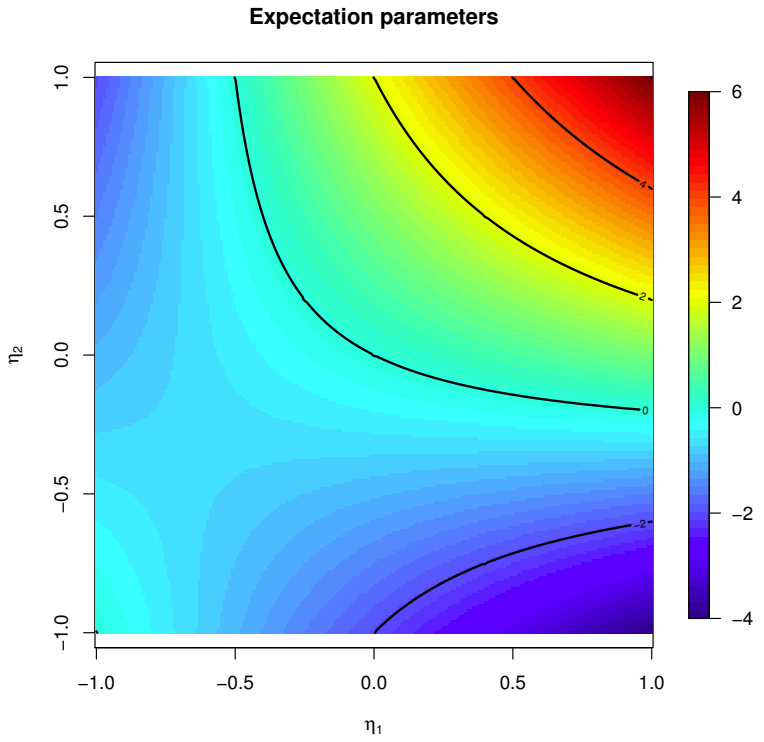


Figure 2. Gradient flow of the Function (2). The domain has been increased to include values outside the square $[-1, +1]^2$.

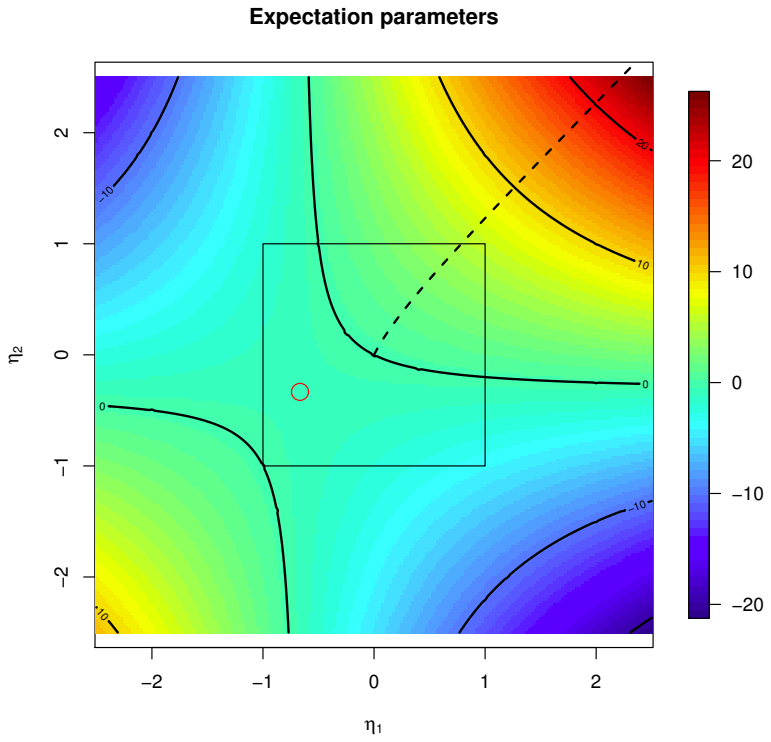
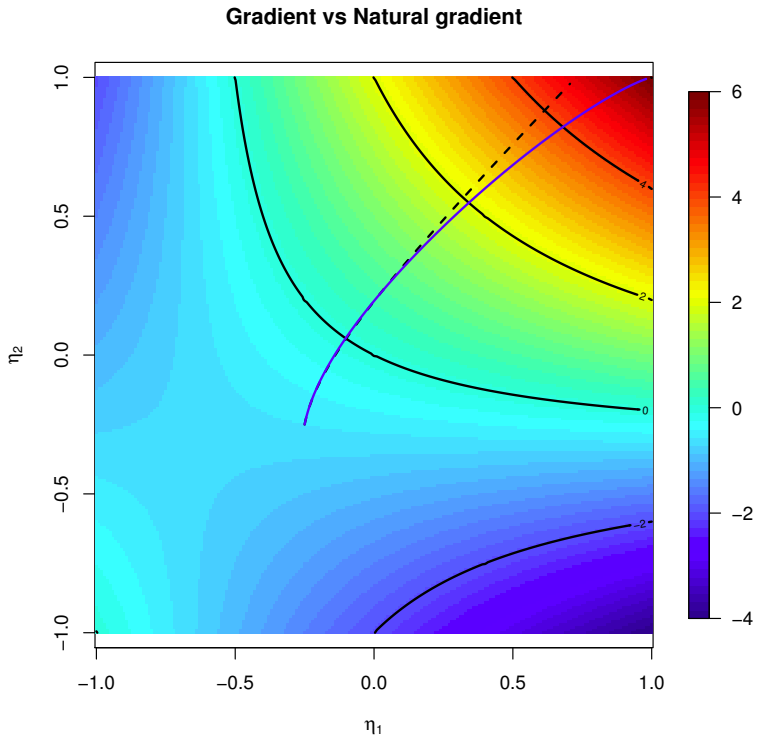


Figure 3. Gradient flow (blue line) and natural gradient flow (black line) for the Function (2), starting at $(-1/4, -1/4)$.



In combinatorial optimization, the values of the function f are assumed to be available at each point, and the curve of steepest ascent of the relaxed function is learned through a simulation procedure based on exponential statistical models.

In this paper, we introduce, in Section 2, the geometry of exponential families and its first order calculus. The second order calculus and the Hessian are discussed in Section 3. Finally, in Section 4, we apply the formalism to the discussion of the Newton method in the context of the maximization of the relaxed function.

2. Models on a Finite State Space

We consider here the exponential statistical manifold on the set of positive densities on a measure space (Ω, μ) with Ω finite and counting measure μ. The setup we describe below is not strictly required in the finite case, because in such a case, other approaches are possible, but it provides a mathematical formalism that has its own pros and that scales naturally to the infinite case.

We provide below a schematic presentation of our formalism as an introduction to this section.

- Two different exponential families can actually be the same statistical model, as the set of densities in the two exponential families are actually equal. This fact is due to both the arbitrariness of the reference density and the fact that sufficient statistics are actually a vector basis of the vector space generated by the sufficient statistics. In a non-parametric approach, we can refer directly to the vector space of centered log-densities, while the change of reference density is geometrically interpreted as a change of chart. The set of all possible such charts defines a manifold.

- We make a specific interpretation of the tangent bundle as the vector space of Fisher's scores at each density and use such tangent spaces as the space of coordinates. This produces a different tangent space/space of coordinates at each density, and different tangent spaces are mapped one onto another by a proper parallel transport, which is nothing else than the re-centering of random variables.
- If a basis is chosen, a parametrization is given, and such a parametrization is, in fact, a new chart, whose values are real vectors. In the real parametrization, the natural scalar product in each scores space is given by Fisher's information matrix.
- Riemannian gradients are defined in the usual way. It is customary in information geometry to call "natural gradient" the real coordinate presentation of the Riemannian gradient. The natural gradient is computed by applying the inverse of the Fisher information matrix to the Euclidean gradient. It seems that there are tree gradients involved, but they all represent the same object when correctly understood.
- The classical notion of expectation parameters for exponential families carries on as another chart on the statistical manifold, which gives rise to a further presentation of a geometrical object.
- While the statistical manifold is unique, there are at least three relevant connections as structures on the vector bundles of the manifold: one relating to the exponential charts, one relating to the expectation charts and one depending on the Riemannian structure.

2.1. Exponential Families As Manifolds

On the finite sample space Ω, $\#\Omega = n$, let a set of random variables $\mathcal{B} = \{X_1, \dots, X_m\}$ be given, such that $\sum_J \alpha_j X_j$ is constant if, and only if, the α_j's are zero, or, equivalently, such that $X_0 = 1, X_1, \dots, X_m$ are affinely independent. The condition implies, necessarily, the linear independence of $\mathcal{B}$. A common choice is to take a set of linearly independent and μ-centered random variables.

We write $\mathcal{V} = \operatorname{Span}\{X_1, \dots, X_m\}$ and define the following exponential family of positive densities $p \in \mathcal{P}_>$:

$$\mathcal{E}_\mathcal{V} = \left\{ q \in \mathcal{P}_> \middle| q \propto \mathrm{e}^V p, V \in \mathcal{V} \right\}. \tag{3}$$

Given any couple $p, q \in \mathcal{E}_\mathcal{V}$, then there exist a unique set of parameters $\boldsymbol{\theta} = \boldsymbol{\theta}_p(q)$, such that:

$$q = \exp\left(\sum_j \theta^j \, {}^e\mathbb{U}^p X_j - \psi_p(\boldsymbol{\theta}) \right) \cdot p \tag{4}$$

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

$${}^e\mathbb{U}^p \colon \mathcal{V} \ni U \mapsto U - \mathbb{E}_p[U] \in {}^e\mathbb{U}^p\mathcal{V}. \tag{5}$$

The linear mapping ${}^e\mathbb{U}^p$ is one-to-one on $\mathcal{V}$ and ${}^e\mathbb{U}^p X_j$, $j = 1, \dots, m$, and is a basis of ${}^e\mathbb{U}^p\mathcal{V}$. We view each choice of a specific reference p as providing a chart centered at p on the exponential family $\mathcal{E}_\mathcal{V}$, namely:

$$\sigma_p \colon \exp\left(\sum_j \theta^j \, {}^e\mathbb{U}^p X_j - \psi_p(\boldsymbol{\theta}) \right) \cdot p \mapsto \boldsymbol{\theta}, \tag{6}$$

If:

$$U = {}^{e}\mathbb{U}^{p}U + \mathbb{E}_{p}\left[U\right] = \sum_{j=1}^{m} \theta^{j}\, {}^{e}\mathbb{U}^{p}X_{j} + \mathbb{E}_{p}\left[U\right], \tag{7}$$

then:

$$\mathbb{E}_{p}\left[U\, {}^{e}\mathbb{U}^{p}X_{i}\right] = \sum_{j=1}^{m} \theta^{j}\, \mathbb{E}_{p}\left[{}^{e}\mathbb{U}^{p}X_{i}\, {}^{e}\mathbb{U}^{p}X_{j}\right], \tag{8}$$

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

$$I_{\mathcal{B}}(p) = \left[\mathrm{Cov}_{p}\left(X_{i}, X_{j}\right)\right]_{ij} = \mathbb{E}_{p}\left[\boldsymbol{X}\boldsymbol{X}'\right] - \mathbb{E}_{p}\left[\boldsymbol{X}\right]\mathbb{E}_{p}\left[\boldsymbol{X}'\right] \tag{9}$$

is the Fisher information matrix of the basis $\mathcal{B} = \{X_{1}, \ldots, X_{m}\}$.

The mappings:

$$\sigma_{p}\colon \mathcal{E}_{\mathcal{V}} \ni q \mapsto U \mapsto \boldsymbol{\theta} \in \mathbb{R}^{m} \tag{10}$$

where:

$$s_{p}\colon q \mapsto U = \log\left(\frac{q}{p}\right) - \mathbb{E}_{p}\left[\log\left(\frac{q}{p}\right)\right], \tag{11}$$

$$\sigma_{p}\colon q \mapsto \boldsymbol{\theta} = I_{\mathcal{B}}^{-1}(p)\, \mathbb{E}_{p}\left[U\, {}^{e}\mathbb{U}^{p}\boldsymbol{X}\right] = I_{\mathcal{B}}^{-1}(p)\, \mathbb{E}_{p}\left[\log\left(\frac{q}{p}\right) {}^{e}\mathbb{U}^{p}\boldsymbol{X}\right], \tag{12}$$

are global charts in the non-parametric and parametric coordinates, respectively. Notice that Equation (12) provides the regression coefficients of the least squares estimate on ${}^{e}\mathbb{U}^{p}\mathcal{V}$ of the log-likelihood.

We denote by $e_{p}\colon \mathbb{R}^{m} \to \mathcal{E}_{\mathcal{V}}$ the inverse of σ_{p}, *i.e.*,

$$e_{p}(\boldsymbol{\theta}) = \exp\left(\sum_{j=1}^{m} \theta^{j}\, {}^{e}\mathbb{U}^{p}X_{j} - \psi_{p}(\boldsymbol{\theta})\right) \cdot p, \tag{13}$$

so that the representation of the divergence $q \mapsto \mathrm{D}\left(p \,\|q\right)$ in the chart σ_{p} is ψ_{p}:

$$\psi_{p}(\boldsymbol{\theta}) = \log\left(\mathbb{E}_{p}\left[\mathrm{e}^{\sum_{j=1}^{m} \theta^{j}\, {}^{e}\mathbb{U}^{p}X_{j}}\right]\right) = \mathbb{E}_{\boldsymbol{\theta}}\left[\log\left(\frac{p}{e_{p}(\boldsymbol{\theta})}\right)\right] = \mathrm{D}\left(p \,\|e_{p}(\boldsymbol{\theta})\right). \tag{14}$$

The mapping $I_{\mathcal{B}}\colon p \mapsto \mathrm{Cov}_{p}\left(\boldsymbol{X}, \boldsymbol{X}\right) \in \mathbb{R}^{m\times m}$ is represented in the chart centered at p by:

$$I_{\mathcal{B},p}(\boldsymbol{\theta}) = I_{\mathcal{B}}(e_{p}(\boldsymbol{\theta})) = \left[\mathrm{Cov}_{e_{p}(\boldsymbol{\theta})}\left(X_{i}, X_{j}\right)\right]_{i,j} = \mathrm{Hess}\, \psi_{p}(\boldsymbol{\theta}), \tag{15}$$

See [1].

2.2. Change of Chart

Fix $p, \bar{p} \in \mathcal{E}_{\mathcal{V}}$; then, we can express p in the chart centered at $\bar{p}$,

$$p = \exp\left(\bar{U} - k_{p}(\bar{U})\right) \cdot \bar{p}, \quad \bar{U} \in {}^{e}\mathbb{U}^{\bar{p}}\mathcal{V}, \quad k_{\bar{p}}(\bar{U}) = \log\left(\mathbb{E}_{\bar{p}}\left[\mathrm{e}^{\bar{U}}\right]\right). \tag{16}$$

In coordinates $\bar{U} = \sum_{j=1^{\cdot}}^{m} \bar{\theta}^{j}\, {}^{e}\mathbb{U}^{\bar{p}}X_{j}$.

For all $q \in \mathcal{E}_{\mathcal{V}}$, $q = \exp(U - k_p(U))\, p$, $U \in {}^e\mathbb{U}^p\mathcal{V}$, $k_p(U) = \log\left(\mathbb{E}_p\left[\mathrm{e}^U\right]\right)$, in coordinates $U = \sum_{j=1^,}^{m} \theta^j\, {}^e\mathbb{U}^p X_j$, we can write:

$$\begin{aligned} q &= \exp(U - k_p(U)) \cdot p \\ &= \exp(U - k_p(U)) \exp\left(\bar{U} - k_{\bar{p}}(\bar{U})\right) \cdot \bar{p} \\ &= \exp\left(U - k_p(U) + \bar{U} - k_{\bar{p}}(\bar{U})\right) \cdot \bar{p} \\ &= \exp\left(\left((U + \bar{U}) - \mathbb{E}_{\bar{p}}[U]\right) - \left(k_p(U) - k_{\bar{p}}(\bar{U}) + \mathbb{E}_{\bar{p}}[U]\right)\right) \cdot \bar{p}, \end{aligned} \tag{17}$$

hence, the non-parametric coordinate of q in the chart centered at $\bar{p}$ is $U + \bar{U} - \mathbb{E}_{\bar{p}}[U] = {}^e\mathbb{U}^{\bar{p}}(U) + \bar{U}$.

From Equation (12):

$$\begin{aligned} \sigma_{\bar{p}}(q) &= I_{\mathcal{V}}^{-1}(\bar{p})\, \mathbb{E}_{\bar{p}}\left[\left({}^e\mathbb{U}^{\bar{p}} U + \bar{U}\right) {}^e\mathbb{U}^{\bar{p}} \boldsymbol{X}\right] \\ &= \boldsymbol{\theta} + \bar{\boldsymbol{\theta}} \end{aligned} \tag{18}$$

This provides the change of charts $\sigma_{\bar{p}} \circ \sigma_p^{-1} \colon \boldsymbol{\theta} \mapsto \boldsymbol{\theta} + \bar{\boldsymbol{\theta}}$. This atlas of charts defines the affine manifold $(\mathcal{E}_{\mathcal{V}}, (\sigma_p))$. This fact has deep consequences that we do not discuss here, e.g., our manifold is an instance of a Hessian manifold [16].

2.3. Tangent Bundle

The space of Fisher scores at p is ${}^e\mathbb{U}^p\mathcal{V}$, and it is identified with the tangent space of the manifold at p, $T_q\mathcal{E}_{\mathcal{V}}$; see the discussion in [3,9]. Let us check the consistency of this statement with our θ-parametrization.

Let:

$$q(\tau) = \exp\left(\sum_{j=1}^{m} \theta^j(\tau)\, {}^e\mathbb{U}^{q(0)} \boldsymbol{X} - \psi_{q(0)}(\tau)\right) \cdot q(0), \tag{19}$$

$\tau \in I$, I an open interval containing zero, a curve in $\mathcal{E}_{\mathcal{V}}$. In the chart centered at $q(0)$, we have from Equation (12):

$$\begin{aligned} \sigma_{q(0)}(q(\tau)) &= I_{\mathcal{B}}^{-1}(q(0))\, \mathbb{E}_{q(0)}\left[\log\left(\frac{q(\tau)}{q(0)}\right) {}^e\mathbb{U}^{q(0)} \boldsymbol{X}\right] \\ &= I_{\mathcal{B}}^{-1}(q(0))\, \mathbb{E}_{q(0)}\left[\left(\sum_{j=1}^{m} \theta^j(\tau)\, {}^e\mathbb{U}^{q(0)} X_j - \psi_{q(0)}(\boldsymbol{\theta}(\tau))\right) {}^e\mathbb{U}^{q(0)} \boldsymbol{X}\right] \\ &= I_{\mathcal{B}}^{-1}(q(0)) \sum_{j=1}^{m} \theta^j(\tau)\, \mathbb{E}_{q(0)}\left[{}^e\mathbb{U}^{q(0)} X_j\, {}^e\mathbb{U}^{q(0)} \boldsymbol{X}\right] \\ &= I_{\mathcal{B}}^{-1}(q(0))\, \mathbb{E}_{q(0)}\left[{}^e\mathbb{U}^{q(0)} \boldsymbol{X}\, {}^e\mathbb{U}^{q(0)} \boldsymbol{X}\right] \boldsymbol{\theta} \\ &= \boldsymbol{\theta}(\tau). \end{aligned} \tag{20}$$

The vector space ${}^e\mathbb{U}^p\mathcal{V}$ is represented by the coordinates in the base ${}^e\mathbb{U}^p\mathcal{B}$. The tangent bundle $T\mathcal{E}_{\mathcal{V}}$ as a manifold is defined by the charts $(\sigma_p, \dot{\sigma}_p)$ on the domain:

$$T\mathcal{E}_{\mathcal{V}} = \{(p, v) | p \in \mathcal{E}_{\mathcal{V}}, v \in T_p\mathcal{E}_{\mathcal{V}}\} \tag{21}$$

with:

$$(\sigma_p, \dot{\sigma}_p)\colon (q, V) \mapsto \left(I_{\mathcal{B}}^{-1}(p)\, \mathbb{E}_p \left[\log\left(\frac{q}{p}\right) {}^{e}\mathbb{U}^p \boldsymbol{X} \right], I_{\mathcal{B}}^{-1}(p)\, \mathbb{E}_p \left[V\, {}^{e}\mathbb{U}^p \boldsymbol{X} \right] \right). \tag{22}$$

The dot notation $\dot{\sigma}_p$ for the charts on the tangent spaces is justified by the computation in Equation (23) below:

$$\frac{d}{dt}\sigma_{q(0)}(q(\tau))\Big|_{\tau=0} = I_{\mathcal{B}}^{-1}(q(0))\, \mathbb{E}_{q(0)} \left[\frac{d}{d\tau} \log\left(q(\tau)\right)\Big|_{\tau=0} {}^{e}\mathbb{U}^{q(0)} \boldsymbol{X} \right] = I_{\mathcal{B}}^{-1}(q(0))\, \mathbb{E}_{q(0)} \left[\delta q(0)\, {}^{e}\mathbb{U}^{q(0)} \boldsymbol{X} \right] = \dot{\sigma}_{q(0)}(\delta q(0)). \tag{23}$$

The velocity at $\tau = 0$ is $\delta q(0) = \frac{d}{d\tau} \log\left(q(\tau)\right)\big|_{\tau=0} \in T_{q(0)}\mathcal{E}_{\mathcal{V}}$ and:

$$\begin{aligned} \frac{d}{d\tau}\boldsymbol{\theta}(\tau)\Big|_{\tau=0} &= I_{\mathcal{B}}^{-1}(q(0))\, \mathbb{E}_{q(0)} \left[\frac{d}{d\tau} \log\left(q(\tau)\right)\Big|_{\tau=0} {}^{e}\mathbb{U}^{q(0)} \boldsymbol{X} \right] \\ &= I_{\mathcal{B}}^{-1}(q(0))\, \mathbb{E}_{q(0)} \left[\delta q(0)\, {}^{e}\mathbb{U}^{q(0)} \boldsymbol{X} \right], \end{aligned} \tag{24}$$

which is consistent with both the definition of tangent space as set of Fisher scores and with the chart of the tangent bundle as defined in Equation (22).

The velocity at a generic τ is $\delta q(\tau) = \frac{d}{d\tau} \log\left(q(\tau)\right) \in T_{q(\tau)}\mathcal{E}_{\mathcal{V}}$ and has coordinates at p:

$$\begin{aligned} \frac{d}{d\tau}\boldsymbol{\theta}(\tau) &= I_{\mathcal{B}}^{-1}(q(0))\, \mathbb{E}_{q(0)} \left[\frac{d}{d\tau} \log\left(q(\tau)\right) {}^{e}\mathbb{U}^{q(0)} \boldsymbol{X} \right] \\ &= I_{\mathcal{B}}^{-1}(q(0))\, \mathbb{E}_{q(0)} \left[\delta q(\tau)\, {}^{e}\mathbb{U}^{q(0)} \boldsymbol{X} \right]. \end{aligned} \tag{25}$$

If V, W are vector fields on $T\mathcal{E}_{\mathcal{V}}$, *i.e.*, $V(p), W(p) \in T_p\mathcal{E}_{\mathcal{V}} = {}^{e}\mathbb{U}^p\mathcal{V}$, $p \in \mathcal{E}_{\mathcal{V}}$, we define a Riemannian metric $g(V, W))$ by:

$$g(V, W)(p) = g_p(V(p), W(p)) = \mathbb{E}_p\left[V(p)W(p)\right] \tag{26}$$

In coordinates at p, $V(p) = \sum_j \dot{\sigma}_p^j(V)\, {}^{e}\mathbb{U}^p X_j$, $W(p) = \sum_j \dot{\sigma}_p^j(W)\, {}^{e}\mathbb{U}^p X_j$, so that:

$$g_p(V(p), W(p)) = \dot{\sigma}_p(V)' I_{\mathcal{B}}(p) \dot{\sigma}_p(W). \tag{27}$$

2.4. Gradients

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

$$\begin{array}{ccc} \mathcal{E}_{\mathcal{V}} & \xrightarrow{\phi} & \mathbb{R} \\ {\scriptstyle e_p}\uparrow & \nearrow_{\phi_p} & \\ \mathbb{R}^m & & \end{array} \tag{28}$$

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

$$\nabla\phi_p(\mathbf{0})\boldsymbol{\alpha} = \nabla\phi_p(\mathbf{0}) I_{\mathcal{B}}^{-1}(p) I_{\mathcal{B}}(p)\boldsymbol{\alpha} = \left(I_{\mathcal{B}}^{-1}(p)\nabla\phi_p(\mathbf{0})'\right)' I_{\mathcal{B}}(p)\boldsymbol{\alpha} = g_p(I_{\mathcal{B}}^{-1}(p)\nabla\phi_p(\mathbf{0})', \boldsymbol{\alpha}). \tag{29}$$

The mapping $\widetilde{\nabla}\phi\colon p \mapsto I_{\mathcal{B}}^{-1}(p)(\nabla\phi_p(\mathbf{0}))' \in \mathbb{R}^m$ that appears in Equation (29) is Amari's natural gradient of $\phi\colon \mathcal{E}_{\mathcal{V}}$; see [15]. It is a standard notion in Riemannian geometry; *cf.* [4] (p. 46).

More generally, the derivative of $\boldsymbol{\theta} \mapsto \phi_p(\boldsymbol{\theta})$ at $\boldsymbol{\theta}$ along $\boldsymbol{\alpha} \in \mathbb{R}^m$ is:

$$\nabla\phi_p(\boldsymbol{\theta})\boldsymbol{\alpha} = \nabla\phi_p(\boldsymbol{\theta})I_{\mathcal{B}}^{-1}(e_p(\boldsymbol{\theta}))I_{\mathcal{B}}(e_p(\boldsymbol{\theta}))\boldsymbol{\alpha} = \left(I_{\mathcal{B}}^{-1}(e_p(\boldsymbol{\theta}))\nabla\phi_p(\boldsymbol{\theta})'\right)' I_{\mathcal{B}}(e_p(\boldsymbol{\theta}))\boldsymbol{\alpha} = g_{e_p(\boldsymbol{\theta})}(I_{\mathcal{B}}^{-1}(e_p(\boldsymbol{\theta}))\nabla\phi_p(\boldsymbol{\theta})', \boldsymbol{\alpha}). \quad (30)$$

Let us compare $\nabla\phi_q(\mathbf{0})$ and $\nabla\phi_p(\boldsymbol{\theta})$ when $q = e_p(\boldsymbol{\theta})$. As $\phi_p = \phi \circ e_p$ and $\phi_q = \phi \circ e_q$, we have the change of charts:

$$\phi_q = \phi \circ e_q = \phi \circ e_p \circ \sigma_p \circ e_q = \phi_p \circ \sigma_p \circ e_q, \quad (31)$$

hence $\nabla\phi_q(\mathbf{0}) = \nabla\phi_p(\sigma_p(q))J(\sigma_p \circ e_q)(\mathbf{0})$, where $J(\sigma_p \circ e_q)$ is the Jacobian of $\sigma_p \circ e_q$. As $\sigma_p \circ e_q(\theta) = \theta + \sigma_p(q)$, we have $J(\sigma_p \circ e_q) = \mathrm{Id}$, and in conclusion, $\nabla\phi_{e_p(\theta)}(\mathbf{0}) = \nabla\phi_p(\boldsymbol{\theta})$. For all $p \in \mathcal{E}_{\mathcal{V}}$ and $\boldsymbol{\theta} \in \mathbb{R}^m$,

$$\widetilde{\nabla}\phi(e_p(\theta)) = I_{\mathcal{B}}^{-1}(e_p(\theta))\nabla\phi_p(\theta). \quad (32)$$

Alternatively, for all $q, p \in \mathcal{E}_{\mathcal{V}}$, $\widetilde{\nabla}\phi\colon \mathcal{E}_{\mathcal{V}} \to \mathbb{R}^m$ is defined by:

$$\widetilde{\nabla}\phi(q) = I_{\mathcal{B}}^{-1}(q)\nabla\phi_p(\sigma_p(q)). \quad (33)$$

The Riemannian gradient of $\phi\colon \mathcal{E}_{\mathcal{V}}$ is the vector field $\nabla\phi$, such that $D_Y\phi = g(\nabla\phi, Y)$. Note that the Riemannian gradient takes values in the tangent bundle, while the natural gradient takes values in $\mathbb{R}^m$. We compute the Riemannian gradient at p as follows. If $\boldsymbol{y} = \dot{\sigma}_p(Y(p))$,

$$D_Y\phi(p) = d\phi_p(\mathbf{0})\boldsymbol{y} = g_p(\widetilde{\nabla}\phi(p), \boldsymbol{y}) = \mathbb{E}_p\left[\nabla\phi(p)Y(p)\right], \quad (34)$$

hence $\widetilde{\nabla}\phi(p) = I_{\mathcal{B}}^{-1}(p)\nabla\phi_p(\mathbf{0})'$ is the representation in the chart centered at p of the vector field $\nabla\phi\colon \mathcal{E}_{\mathcal{V}}$. Explicitly, we have (see Equation (22)),

$$\widetilde{\nabla}\phi(p) = I_{\mathcal{B}}^{-1}(p)(\nabla\phi_p(\mathbf{0}))' = I_{\mathcal{B}}^{-1}(p)\,\mathbb{E}_p\left[\nabla\phi(p)\,{}^e\mathbb{U}^p\boldsymbol{X}\right], \quad (35)$$

$$\nabla\phi(p) = \sum_j (\widetilde{\nabla}\phi(p))^j\,{}^e\mathbb{U}^p X_j \quad (36)$$

The Euclidean gradient $\nabla\phi_p(\boldsymbol{\theta})$ is sometimes called the "vanilla gradient." It is equal to the covariance between the Riemannian gradient $\nabla\phi(p)$ and the basis $\boldsymbol{X}$, $(\nabla\phi_p(\mathbf{0}))' = \mathbb{E}_p\left[\nabla\phi(p)\,{}^e\mathbb{U}^p\boldsymbol{X}\right]$.

We summarize in a display the relations between our three gradients: Euclidean $\nabla\phi_p(\mathbf{0})$, natural $\widetilde{\nabla}\phi(p)$ and Riemannian $\nabla\phi(p)$.

$$\begin{array}{ccc} T\mathcal{E}_{\mathcal{V}} & \xrightarrow{(\sigma_p, \dot{\sigma}_p)} & \mathbb{R}^{2m} \\ {\scriptstyle\pi}\downarrow & & \downarrow{\scriptstyle\pi_1} \\ \mathcal{E}_{\mathcal{V}} & \xrightarrow[\sigma_p]{} & \mathbb{R}^m \end{array} \qquad \begin{array}{ccc} T_p\mathcal{E}_{\mathcal{V}} & \xrightarrow{\dot{\sigma}_p} & \mathbb{R}^m \\ {\scriptstyle\nabla\phi(p)}\uparrow & & \downarrow{\scriptstyle I_{\mathcal{B}}(p)} \\ \mathcal{E}_{\mathcal{V}} & \xrightarrow[\nabla\phi_p(\mathbf{0})]{} & \mathbb{R}^m \end{array} \qquad \dot{\sigma}_p \circ \nabla\phi(p) = I_{\mathcal{B}}^{-1}\nabla\phi_p(\mathbf{0}) = \widetilde{\nabla}\phi(p) \quad (37)$$

In the following, we shall frequently use the fact that the representation of the gradient vector field $\nabla\phi$ in a generic chart centered at p is:

$$(\nabla\phi)_p(\boldsymbol{\theta}) = \dot{\sigma}_p(\nabla\phi(e_p(\boldsymbol{\theta}))) = (\widetilde{\nabla}\phi)(e_p(\boldsymbol{\theta})) = I_{\mathcal{B},p}^{-1}(\boldsymbol{\theta})\nabla\phi_p(\boldsymbol{\theta}). \quad (38)$$

It should be noted that the leftmost term $(\nabla\phi)_p(\boldsymbol{\theta})$ is the presentation of the gradient in the charts of the tangent bundle, while in the rightmost term, $\nabla\phi_p(\boldsymbol{\theta})$ denotes the Euclidean gradient of the presentation of the function ϕ in the charts of the manifold.

2.4.1. Expectation Parameters

As ψ_p is strictly convex, the gradient mapping $\boldsymbol{\theta} \mapsto (\nabla\psi_p(\boldsymbol{\theta}))'$ is a homeomorphism from the space of parameters $\mathbb{R}^m$ to the interior of the convex set generated by the image of ${}^e\mathbb{U}^p\boldsymbol{X}$; see [1]. The function $\mu_p\colon \mathcal{E}_\mathcal{V}$ defined by:

$$\mu_p(q) = \mathbb{E}_q\left[{}^e\mathbb{U}^p\boldsymbol{X}\right] = \mathbb{E}_q\left[\boldsymbol{X}\right] - \mathbb{E}_p\left[\boldsymbol{X}\right] = (\nabla\psi_p(\boldsymbol{\theta}))', \quad \boldsymbol{\theta} = \sigma_p(q) \tag{39}$$

is a chart for all $p \in \mathcal{E}_\mathcal{V}$. The value of the inverse $q = L_p(\boldsymbol{\mu})$ is characterized as the unique $q \in \mathcal{E}_\mathcal{V}$, such that $\boldsymbol{\mu} = \mathbb{E}_q\left[{}^e\mathbb{U}^p\boldsymbol{X}\right]$, *i.e.*, the maximum likelihood estimator.

Let us compute the change of chart from p to $\bar{p}$:

$$\mu_{\bar{p}} \circ \mu_p^{-1}(\boldsymbol{\eta}) = \bar{\boldsymbol{\eta}} = \boldsymbol{\eta} + \mathbb{E}_p\left[\boldsymbol{X}\right] - \mathbb{E}_{\bar{p}}\left[\boldsymbol{X}\right]. \tag{40}$$

In fact, $\boldsymbol{\mu} = \mathbb{E}_{L_p(\boldsymbol{\mu})}\left[{}^e\mathbb{U}^p\boldsymbol{X}\right]$ and $\bar{\boldsymbol{\mu}} = \mu_{\bar{p}}(L_p(\boldsymbol{\mu})) = \mathbb{E}_{L_p(\boldsymbol{\mu})}\left[{}^e\mathbb{U}^{\bar{p}}\boldsymbol{X}\right]$.

We do not discuss here the rich theory started in [2] about the duality between σ_p and μ_p. We limit ourselves to the computation of the Riemannian gradient in the expectation parameters. If $\phi\colon \mathcal{E}_\mathcal{V}$,

$$\phi_p(\boldsymbol{\theta}) = \phi \circ e_p(\boldsymbol{\theta}) = \phi \circ L_p \circ \mu_p \circ e_p(\boldsymbol{\theta}) = (\phi \circ L_p) \circ (\nabla\psi_p)(\boldsymbol{\theta}), \tag{41}$$

because $\mu_p \circ e_p(\boldsymbol{\theta}) = \mathbb{E}_{e_p(\boldsymbol{\theta})}\left[{}^e\mathbb{U}^p\boldsymbol{X}\right] = \nabla\phi_p(\boldsymbol{\theta})$, hence:

$$\nabla\phi_p(\boldsymbol{\theta}) = \nabla(\phi \circ L_p)(\nabla\psi_p(\boldsymbol{\theta}))\,\mathrm{Hess}\,\psi_p(\boldsymbol{\theta}), \tag{42}$$

$$\widetilde{\nabla}\phi(p) = I_\mathcal{V}(p)^{-1}(\nabla(\phi \circ L_p)(\mathbf{0})\,\mathrm{Hess}\,\psi_p(\mathbf{0}))' = (\nabla(\phi \circ L_p)(\mathbf{0}))', \tag{43}$$

$$\nabla\phi(p) = \nabla(\phi \circ L_p)(\mathbf{0})\,{}^e\mathbb{U}^p\boldsymbol{X}, \tag{44}$$

that is, the natural gradient $\widetilde{\nabla}\phi$ at $p = L_p(\boldsymbol{\mu})$ is equal to the Euclidean gradient of $\boldsymbol{\mu} \mapsto \phi \circ L_p(\boldsymbol{\mu})$ at $\boldsymbol{\mu} = \mathbf{0}$.

2.4.2. Vector Fields

If V is a vector field of $T\mathcal{E}_\mathcal{V}$ and $\phi\colon \mathcal{E}_\mathcal{V}$ is a real function, then we define the action of V on ϕ, $\nabla_V\phi$, to be the real function:

$$\nabla_V\phi\colon \mathcal{E}_\mathcal{V} \ni p \mapsto \nabla_V\phi(p) = \nabla\phi_p(\mathbf{0})\dot{\sigma}_p\left(V(p)\right). \tag{45}$$

We prefer to avoid the standard notation $V\phi$, because in our setting, $V(p)$ is a random variable, and the product $V(p)\phi(p)$ is otherwise defined as the ordinary product.

Let us represent $\nabla_V\phi$ in the chart centered at p:

$$(\nabla_V\phi)_p(\boldsymbol{\theta}) = \nabla_V\phi(e_p(\boldsymbol{\theta})) = \nabla\phi_{e_p(\boldsymbol{\theta})}(\mathbf{0})\dot{\sigma}_{e_p(\boldsymbol{\theta})}\left(V(e_p(\boldsymbol{\theta}))\right) = \nabla\phi_p(\boldsymbol{\theta})V_p(\boldsymbol{\theta}), \tag{46}$$

where we have used the equality $\nabla\phi_{e_p(\boldsymbol{\theta})}(\mathbf{0}) = \nabla\phi_p(\boldsymbol{\theta})$ and $V_p(\boldsymbol{\theta}) = \dot{\sigma}_{e_p(\boldsymbol{\theta})}\left(V(e_p(\boldsymbol{\theta}))\right)$.

If W is a vector field, we can compute $\nabla_W\nabla_V\phi$ at p as:

$$\begin{aligned}\nabla_W\nabla_V\phi(p) &= \nabla(\nabla_V\phi)_p(\mathbf{0})\dot{\sigma}_p(W(p)) \\ &= V_p(\mathbf{0})'\,\mathrm{Hess}\,\phi_p(\mathbf{0})W_p(\mathbf{0}) + \nabla\phi_p(\mathbf{0})JV_p(\mathbf{0})W_p(\mathbf{0}),\end{aligned} \tag{47}$$

where J denotes the Jacobian matrix.

The Lie bracket $[W, V]\phi$ (see [7] (§4.2), [8] (V, §1), [4] (Section 5.3.1)) is given by:

$$[W, V]\phi(p) = \nabla_W \nabla_V \phi(p) - \nabla_V \nabla_w \phi(p) = \nabla\phi_p(\mathbf{0}) \left(JV_p(\mathbf{0})W_p(\mathbf{0}) - JW_p(\mathbf{0})V_p(\mathbf{0})\right), \tag{48}$$

because of Equation (47) and the symmetry of the Hessian.

The flow of the smooth vector field $V\colon \mathcal{E}_\mathcal{V}$ is a family of curves $\gamma(t, p)$, $p \in \mathcal{E}_\mathcal{V}$, $t \in J_p$, J_p open real interval containing zero, such that for all $p \in \mathcal{E}_\mathcal{V}$ and $t \in J_p$,

$$\gamma(0, p) = p, \tag{49}$$

$$\delta\gamma(t, p) = V(\gamma(t, p)). \tag{50}$$

As uniqueness holds in Equation (50) (see [8] (VI, §1) or [7] (§4.1)), we have semi-group property $\gamma(s + t, p) = \gamma(s, \gamma(t, p))$, and Equation (50) is equivalent to $\delta\gamma(0, p) = V(\gamma(0, p))$, $p \in \mathcal{E}_\mathcal{V}$.

If a flow of V is available, we have an interpretation of $\nabla_V \phi$ as a derivative of ϕ along $\gamma(t, p)$,

$$\frac{d}{dt}\phi(\gamma(t, p))\bigg|_{t=0} = \nabla\phi_p(\sigma_p(\gamma(t, p))) \left(\frac{d}{dt}\sigma_p(\gamma(t, p))\right)\bigg|_{t=0} = \nabla\phi_p(\mathbf{0})V(p) = \nabla_V\phi(p). \tag{51}$$

2.5. Examples

The following examples are intended to show how the formalism of gradients is usable in performing basic computations.

2.5.1. Expectation

Let f be any random variable, and define $F\colon \mathcal{E}_\mathcal{V}$ by $F(p) = \mathbb{E}_p[f]$. In the chart centered at p, we have:

$$F_p(\boldsymbol{\theta}) = \int f \exp\left(\sum_j \theta^j \, {}^e\mathbb{U}^p X_j - \psi_p(\boldsymbol{\theta})\right) \cdot p \, d\mu \tag{52}$$

and the Euclidean gradient:

$$\nabla F_p(\mathbf{0}) = \mathrm{Cov}_p(f, \boldsymbol{X}) \in (\mathbb{R}^m)'. \tag{53}$$

The natural gradient is:

$$\widetilde{\nabla} F(p) = \mathrm{Cov}_p(\boldsymbol{X}, \boldsymbol{X})^{-1} \mathrm{Cov}_p(\boldsymbol{X}, f) \in \mathbb{R}^m, \tag{54}$$

and the Riemannian gradient is:

$$\nabla F(p) = (\widetilde{\nabla} F(p))' \, {}^e\mathbb{U}^p \boldsymbol{X} = \mathrm{Cov}_p(f, \boldsymbol{X}) \, \mathrm{Cov}_p(\boldsymbol{X}, \boldsymbol{X})^{-1} \, {}^e\mathbb{U}^p \boldsymbol{X} \in T_p\mathcal{E}_\mathcal{V}. \tag{55}$$

From Equation (55), it follows that $\nabla F(p)$ is the $L^2(p)$-projection f onto ${}^e\mathbb{U}^p\mathcal{V}$, while $\widetilde{\nabla} F(p)$ in Equation (54) are the coordinates of the projection. Let us consider the family of curves:

$$\gamma(t, p) = \exp\left(\sum_{j=1}^m t(\widetilde{\nabla} F(p))^j \, {}^e\mathbb{U}^p X_j - \psi_p(t\widetilde{\nabla} F(p))\right) \cdot p, \quad t \in \mathbb{R}. \tag{56}$$

The velocity is:

$$\delta\gamma(t,p) = \frac{d}{dt}\left(\sum_{j=1}^{m} t(\widetilde{\nabla}F(p))^j\, {}^e\mathbb{U}^p X_j - \psi_p(t\widetilde{\nabla}F(p))\right) = \nabla F(p) - \mathbb{E}_{\gamma(t,p)}\left[\nabla F(p)\right], \tag{57}$$

which is different from $\nabla F(\gamma(t,p))$, unless $f \in \mathcal{V} \oplus \mathbb{R}$. Then, γ is not, in general, the flow of ∇F, but it is a local approximation, as $\delta\gamma(0,p) = \nabla F(p)$.

These computation are the basis of model-based methods in combinatorial optimization; see [10–14].

2.5.2. Binary Independent Variables

Here, we present, in full generality, the toy example of the Introduction; see [17] for more information on the application to combinatorial optimization. Our example is a very special case of Ising exactly solvable models [18], our aim being here to explore the geometric framework.

Let $\Omega = \{+1,-1\}^m$ with counting measure μ, and let the space $\mathcal{V}$ be generated by the coordinate projections $\mathcal{B} = \{X_1,\dots,X_d\}$. Note that we use here the coding $+1,-1$ (from physics) instead of the coding $0,1$, which is more common in combinatorial optimization. The exponential family is $\mathcal{E}_{\mathcal{V}} = \{\exp\left(\sum_{J=1}^m \theta^j X_j - \psi_\lambda(\boldsymbol{\theta})\right)\cdot 2^{-m}\}$, $\lambda(\boldsymbol{x}) = 2^{-m}$ for $x \in \Omega$ being the uniform density. The independence of the sufficient statistics X_j under all distributions in $\mathcal{E}_{\mathcal{V}}$ implies:

$$\psi_\lambda(\boldsymbol{\theta}) = \sum_{j=1}^{m} \psi(\theta^j), \quad \psi(\theta) = \log\left(\cosh(\theta)\right). \tag{58}$$

We have:

$$\begin{aligned}
\nabla\psi_\lambda(\boldsymbol{\theta}) &= [\tanh(\theta^j)\colon j = 1,\dots,d] \\
&= \eta_\lambda(\boldsymbol{\theta}), \qquad (59)\\
\operatorname{Hess}\psi_\lambda(\boldsymbol{\theta}) &= \operatorname{diag}\left(\cosh^{-2}(\theta^j)\colon j = 1,\dots,d\right) \\
&= \operatorname{diag}\left(\mathrm{e}^{-2\psi(\theta^j)}\colon j = 1,\dots,d\right) \\
&= I_{\mathcal{B},\lambda}(\boldsymbol{\theta}), \qquad (60)\\
I_{\mathcal{B},\lambda}(\boldsymbol{\theta})^{-1} &= \operatorname{diag}\left(\cosh^{2}(\theta^j)\colon j = 1,\dots,d\right) \\
&= \operatorname{diag}\left(\mathrm{e}^{2\psi(\theta^j)}\colon j = 1,\dots,d\right). \qquad (61)
\end{aligned}$$

The quadratic function $f(\boldsymbol{X}) = a_0 + \sum_j a_j X_j + \sum_{\{i,j\}} a_{i,j} X_i X_j$ has expected value at $p = e_\lambda(\boldsymbol{\theta})$, *i.e.*, relaxed value, equal to:

$$F(p) = F_\lambda(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}\left[f(\boldsymbol{X})\right] = a_0 + \sum_j a_j \tanh(\theta^j) + \sum_{\{i,j\}} a_{i,j}\tanh(\theta^i)\tanh(\theta^j), \tag{62}$$

and covariance with $X_k \in \mathcal{B}$ equal to:

$$\begin{aligned}
\operatorname{Cov}_{\boldsymbol{\theta}}\left(f(\boldsymbol{X}), X_k\right) &= \sum_j a_j \operatorname{Cov}_{\boldsymbol{\theta}}\left(X_j, X_k\right) + \sum_{\{i,j\}} a_{i,j}\operatorname{Cov}_{\boldsymbol{\theta}}\left(X_i X_j, X_k\right) \\
&= a_k \operatorname{Var}_{\boldsymbol{\theta}}\left(X_k\right) + \sum_{i\neq k} a_{i,k}\,\mathbb{E}_{\boldsymbol{\theta}}\left[X_i\right]\operatorname{Var}_{\boldsymbol{\theta}}\left(X_k\right) \\
&= \cosh^{-2}(\theta^k)\left(a_k + \sum_{i\neq k} a_{i,k}\tanh(\theta^i)\right). \qquad (63)
\end{aligned}$$

In the computation, we have used the independence and the special algebra of ± 1, which implies $X_i^2 = 1$, so that $\mathrm{Cov}_\theta\left(X_i X_j, X_k\right) = 0$ if $i, j \neq k$, otherwise $\mathrm{Cov}_\theta\left(X_i X_k, X_k\right) = \mathbb{E}_\theta\left[X_i\right] - \mathbb{E}_\theta\left[X_i\right]\mathbb{E}_\theta\left[X_k\right]^2$; see [13].

The Euclidean gradient, the natural gradient and the Riemannian gradient are, respectively,

$$\nabla F_\lambda(\boldsymbol{\theta}) = \left[\cosh^{-2}(\theta^j)\left(a_j + \sum_{i\neq j} a_{i,j}\tanh(\theta^i)\right) : j = 1, \ldots, d\right], \tag{64}$$

$$\widetilde{\nabla} F(e_\lambda(\boldsymbol{\theta})) = \left[a_j + \sum_{i\neq j} a_{i,j}\tanh(\theta^i) \colon j = 1, \ldots, d\right], \tag{65}$$

$$\nabla F(e_\lambda(\boldsymbol{\theta})) = \sum_{j=1}^{m}\left(a_j + \sum_{i\neq j} a_{i,j}\,\mathbb{E}_\theta\left[X_i\right]\right)\left(X_j - \mathbb{E}_\theta\left[X_j\right]\right). \tag{66}$$

The (natural) gradient flow equations are:

$$\dot{\theta}^j(t) = a_j + \sum_{i\neq j} a_{i,j}\tanh(\theta^i(t)), \quad j = 1, \ldots, d. \tag{67}$$

Equations (64)–(66) are usable in practice if the a_j's and the $a_{i,j}$'s are estimable. Otherwise, one can use Equation (63) and the following forms of the gradients:

$$\nabla F_\lambda(\boldsymbol{\theta}) = \left[\mathrm{Cov}_{\boldsymbol{\theta}}\left(X_j, f(\boldsymbol{X})\right) : j = 1, \ldots, d\right], \tag{68}$$

$$\widetilde{\nabla} F(e_\lambda(\boldsymbol{\theta})) = \left[\cosh^2(\theta^j)\,\mathrm{Cov}_{\boldsymbol{\theta}}\left(f(\boldsymbol{X}), X_j\right) : j = 1, \ldots, d\right], \tag{69}$$

in which case, the gradient flow equations are:

$$\dot{\theta}^j(t) = \cosh^2(\theta^j)\,\mathrm{Cov}_{\boldsymbol{\theta}}\left(f(\boldsymbol{X}), X_j\right), \quad j = 1, \ldots, d. \tag{70}$$

Let us study the relaxed function in the expectation parameters $\eta^j = \eta^j(\boldsymbol{\theta})$, $j = 1, \ldots, d$,

$$F_\lambda(\boldsymbol{\eta}) = a_0 + \sum_j a_j\eta^j + \sum_{\{i,j\}} a_{i,j}\eta^i\eta^j, \quad \boldsymbol{\eta} \in]-1, +1[^m. \tag{71}$$

The Euclidean gradient with respect to $\boldsymbol{\eta}$ has components:

$$\partial_j F_\lambda(\boldsymbol{\eta}) = a_j + \sum_{i\neq j} a_{i,j}\eta^i, \tag{72}$$

which are equal to the components of the natural gradient; see Section 2.4.1. As:

$$\dot{\eta}^j(t) = \frac{d}{dt}\tanh(\theta^j(t)) = \cosh^{-2}(\theta^j(t))\dot{\theta}^j(t) = \left(1 - \eta^j(t)^2\right)\dot{\theta}^j(t), \quad j = 1, \ldots, m, \tag{73}$$

the gradient flow expressed in the $\boldsymbol{\eta}$-parameters has equations:

$$\dot{\eta}^j(t) = \left(1 - \eta^j(t)^2\right)\left(a_j + \sum_{i\neq j} a_{i,j}\eta^i(t)\right), \quad j = 1, \ldots, d. \tag{74}$$

Alternatively, in vector form,

$$\dot{\boldsymbol{\eta}}(t) = \mathrm{diag}\left(1 - \eta^j(t)^2 \colon j = 1, \ldots, d\right)\left(\boldsymbol{a} + A\boldsymbol{\eta}(t)\right), \tag{75}$$

where $\boldsymbol{a} = [a_j \colon j = 1, \ldots, d]^t$ and $A_{i,j} = 0$ if $i = j$, $A_{ij} = a_{i,j}$. The matrix A is symmetric with zero diagonal, and it has the meaning of the adjacency matrix of the (weighted) interaction graph. We do not know a closed-form solution of Equation (74). An example of a numerical solution is shown in Figure 3.

2.5.3. Escort Probabilities

For a given $a > 0$, consider the function $C^{(a)} : \mathcal{E}_{\mathcal{V}}$ defined by $C^{(a)}(p) = \int p^a \, d\mu$. We have:

$$C_p^{(a)}(\boldsymbol{\theta}) = \int \exp\left(a \sum_{j=1}^{m} \theta^j \, {}^e\mathbb{U}^p X_j - a\psi_p(\boldsymbol{\theta})\right) p^a \, d\mu \tag{76}$$

and:

$$dC_p^{(a)}(\mathbf{0})\boldsymbol{\alpha} = \int a \left(\sum_{j=1}^{m} \alpha^j \, {}^e\mathbb{U}^p X_j\right) p^a \, d\mu = \sum_{j=1}^{m} \alpha^j \int a \, {}^e\mathbb{U}^p X_j p^a \, d\mu = \sum_{j=1}^{m} \alpha^j \operatorname{Cov}_p\left(X_j, ap^{a-1}\right), \tag{77}$$

that is, the Euclidean gradient is $\nabla C_p^{(a)}(\mathbf{0}) = \operatorname{Cov}_p\left(ap^{a-1}, \boldsymbol{X}\right)$ (row vector). The natural gradient is computed from Equation (35) as:

$$\widetilde{\nabla} C^{(a)}(p) = I_{\mathcal{B}}^{-1}(p)(\nabla C_p^{(a)}(\mathbf{0}))' = \operatorname{Cov}_p(\boldsymbol{X}, \boldsymbol{X})^{-1} \operatorname{Cov}_p\left(\boldsymbol{X}, ap^{a-1}\right), \tag{78}$$

while the Riemannian gradient follows from Equation (36):

$$\nabla C^{(a)}(p) = \operatorname{Cov}_p\left(ap^{a-1}, \boldsymbol{X}\right) \operatorname{Cov}_p(\boldsymbol{X}, \boldsymbol{X})^{-1} \, {}^e\mathbb{U}^p \boldsymbol{X}. \tag{79}$$

Note that the Riemannian gradient is the orthogonal projection of the random variable ap^{a-1} onto the tangent space $T_p\mathcal{E}_{\mathcal{V}} = {}^e\mathbb{U}^p\mathcal{V}$.

The probability density $p^a/C(p)$ is called the escort density in the literature on non-extensive statistical mechanics; see, e.g., [19] (Section 7.4).

We compute now the tangent mapping of $\mathcal{E}_{\mathcal{V}} \ni p \mapsto p^a/C^{(a)}(a) \in \mathcal{P}_>$. Let us extend the basis $X_1, \ldots, X_m$ to a basis $X_1, \ldots, X_n$, $n \geq m$, whose exponential family is full, *i.e.*, equal to $\mathcal{P}_>$. The non-parametric coordinate of $q = \left(\exp\left(\sum_{j=1}^m \theta^j \, {}^e\mathbb{U}^p X_j - \psi_p(\boldsymbol{\theta})\right) p\right)^a / C_p^{(a)}(\boldsymbol{\theta})$ in the chart centered at $\bar{p} = p^a/C_p^{(a)}(\mathbf{0})$ is the $\bar{p}$-centering of the random variable:

$$\begin{aligned} \log\left(\frac{q}{\bar{p}}\right) &= \log\left(\frac{\left(\exp\left(\sum_{j=1}^m \theta^j \, {}^e\mathbb{U}^p X_j - \psi_p(\boldsymbol{\theta})\right) p\right)^a / C_p^{(a)}(\boldsymbol{\theta})}{p^a / C_p^{(a)}(\mathbf{0})}\right) \\ &= a \sum_{j=1}^{m} \theta^j \, {}^e\mathbb{U}^p X_j - a\psi_p(\boldsymbol{\theta}) + \ln C_{(}^{(a)}\mathbf{0}) - \ln C_p^{(a)}(\boldsymbol{\theta}), \end{aligned} \tag{80}$$

that is,

$$v = a \sum_{j=1}^{m} \theta^j \, {}^e\mathbb{U}^{\bar{p}} X_j. \tag{81}$$

The coordinates of v in the basis ${}^e\mathbb{U}^{\bar{p}} X_1, \ldots, {}^e\mathbb{U}^{\bar{p}} X_n$ are $(a\theta^1, \ldots, a\theta^m, 0, \ldots, 0)$, and the Jacobian of $\boldsymbol{\theta} \mapsto (a\boldsymbol{\theta}, \mathbf{0}_{n-m})$ is the $m \times n$ matrix $[aI_m | 0_{m\times(n-m)}]$.

2.5.4. Polarization Measure

The polarization measure has been introduced in Economics by [20]. Here, we consider the qualitative version of [21]. If π is a distribution of a finite set, the probability that in three independent samples from π there are exactly two equal is $3\sum_j \pi_j^2(1-\pi_j)$. If $p \in \mathcal{E}_{\mathcal{V}}$, define:

$$G(p) = \int p^2(1-p)\, d\mu = C^{(2)}(p) - C^{(3)}(p), \tag{82}$$

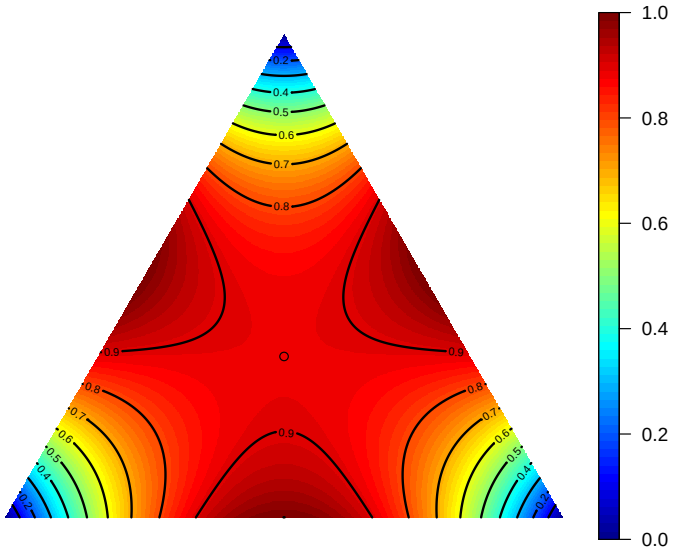
where $C^{(2)}$ and $C^{(3)}$ are defined as in Example 2.5.3.

From Equation (78), we find the natural gradient:

$$\widetilde{\nabla} G(p) = \mathrm{Cov}_p\left(\boldsymbol{X}, \boldsymbol{X}\right)^{-1} \mathrm{Cov}_p\left(\boldsymbol{X}, 2p - 3p^2\right). \tag{83}$$

Note that $\widetilde{\nabla} G(p) = \mathbf{0}$ if p is constant; see Figure 4.

Figure 4. Normalized polarization.



3. Second Order Calculus

In this section, we turn to considering second order calculus, in particular Hessians, in order to prepare the discussion of the Newton method for the relaxed optimization of Section 4.

3.1. Metric Derivative (Levi–Civita connection)

Let $V, W\colon \mathcal{E}_{\mathcal{V}}$ be vector fields, that is, $V(p), W(p) \in T_p\mathcal{E}_{\mathcal{V}} = {}^e\mathbb{U}^p\mathcal{V}$, $p \in \mathcal{E}_{\mathcal{V}}$. Consider the real function $R = g(V,W)\colon \mathcal{E}_{\mathcal{V}} \to \mathbb{R}$, whose value at $p \in \mathcal{E}_{\mathcal{V}}$ is $R(p) = g_p(V(p), W(p)) = \mathbb{E}_p\left[V(p)W(p)\right]$. Assuming smoothness, we want to compute the derivative of R along the vector field $Y\colon \mathcal{E}_{\mathcal{V}}$, that is, $(D_Y R)(p) = dR_p(\mathbf{0})\boldsymbol{\alpha}$, with $\boldsymbol{\alpha} = \dot{\sigma}_p(Y(p))$. The expression of R in the chart centered at p is, according to Equation (27),

$$\boldsymbol{\theta} \mapsto R_p(\boldsymbol{\theta}) = \dot{\sigma}_p(V(e_p(\boldsymbol{\theta})))' I_{\mathcal{B}}(e_p(\boldsymbol{\theta}))\dot{\sigma}_p(W(e_p(\boldsymbol{\theta}))) = V_p(\boldsymbol{\theta})' I_{\mathcal{B},p}(\boldsymbol{\theta}) W_p(\boldsymbol{\theta}), \tag{84}$$

where V_p and W_p are the presentation in the chart of the vector fields V and W, respectively.

The i-th component $\partial_i R_p(\boldsymbol{\theta})$ of the Euclidean gradient $\nabla R_p(\boldsymbol{\theta})$ is:

$$\begin{aligned}\partial_i R_p(\boldsymbol{\theta}) = \partial_i \left(V_p(\boldsymbol{\theta})' I_{\mathcal{B},p}(\boldsymbol{\theta}) W_p(\boldsymbol{\theta})\right) = \\ \partial_i V_p(\boldsymbol{\theta})' I_{\mathcal{B},p}(\boldsymbol{\theta}) W_p(\boldsymbol{\theta}) + V_p(\boldsymbol{\theta})' \partial_i I_{\mathcal{B},p}(\boldsymbol{\theta}) W_p(\boldsymbol{\theta}) + V_p(\boldsymbol{\theta})' I_{\mathcal{B},p}(\boldsymbol{\theta}) \partial_i W_p(\boldsymbol{\theta}) = \\ \left(\partial_i V_p(\boldsymbol{\theta}) + \frac{1}{2} I_{\mathcal{B},p}^{-1}(\boldsymbol{\theta}) \partial_i I_{\mathcal{B},p}(\boldsymbol{\theta}) V_p(\boldsymbol{\theta})\right)' I_{\mathcal{B},p}(\boldsymbol{\theta}) W_p(\boldsymbol{\theta}) + \\ V_p(\boldsymbol{\theta})' I_{\mathcal{B},p}(\boldsymbol{\theta}) \left(\partial_i W_p(\boldsymbol{\theta}) + \frac{1}{2} I_{\mathcal{B},p}^{-1}(\boldsymbol{\theta}) \partial_i I_{\mathcal{B},p}(\boldsymbol{\theta}) W_p(\boldsymbol{\theta})\right), \quad (85)\end{aligned}$$

so that the derivative at $\boldsymbol{\theta}$ along $\boldsymbol{\alpha} = \dot{\sigma}_{e_p(\boldsymbol{\theta})}(Y(e_p(\boldsymbol{\theta})))$ is:

$$\begin{aligned}dR_p(\boldsymbol{\theta})\boldsymbol{\alpha} = \left(dV_p(\boldsymbol{\theta})\boldsymbol{\alpha} + \frac{1}{2} I_{\mathcal{B},p}^{-1}(\boldsymbol{\theta}) \left(dI_{\mathcal{B},p}(\boldsymbol{\theta})\boldsymbol{\alpha}\right) V_p(\boldsymbol{\theta})\right)' I_{\mathcal{B},p}(\boldsymbol{\theta}) W_p(\boldsymbol{\theta}) + \\ V_p(\boldsymbol{\theta})' I_{\mathcal{B},p}(\boldsymbol{\theta}) \left(dW_p(\boldsymbol{\theta})\boldsymbol{\alpha} + \frac{1}{2} I_{\mathcal{B},p}^{-1}(\boldsymbol{\theta}) \left(dI_{\mathcal{B},p}(\boldsymbol{\theta})\boldsymbol{\alpha}\right) W_p(\boldsymbol{\theta})\right). \quad (86)\end{aligned}$$

Proposition 1. *If we define $D_Y V$ to be the vector field on $\mathcal{E}_\mathcal{V}$, whose value at $q = e_p(\boldsymbol{\theta})$ has coordinates centered at p given by:*

$$\dot{\sigma}_p(D_Y V(q)) = dV_p(\boldsymbol{\theta})\boldsymbol{\alpha} + \frac{1}{2} I_\mathcal{B}^{-1}(p) \left(dI_{\mathcal{B},p}(\boldsymbol{\theta})\boldsymbol{\alpha}\right) V_p(\boldsymbol{\theta}), \quad \boldsymbol{\alpha} = \dot{\sigma}_p(Y(q)), \quad (87)$$

then:

$$D_Y g(V, W) = g(D_Y V, W) + g(V, D_Y W), \quad (88)$$

i.e., *Equation* (87) *is a metric covariant derivative; see [6] (Ch. 2 §3), [8] (VIII §4), [4] (§5.3.2).*

The metric derivative Equation (87) could be computed from the flow of the vector field Y. Let $(t, p) \mapsto \gamma(t, p)$ be the flow of the vector field V, *i.e.*, $\delta\gamma(t, p) = V(\gamma(t, p))$ and $\gamma(0, p) = p$. Using Equation (23), we have:

$$\begin{aligned}\frac{d}{dt}\dot{\sigma}(V(\gamma(t,p)))\Big|_{t=0} &= \frac{d}{dt} V_p(\sigma_p(\gamma(t,p)))\Big|_{t=0} \\ &= dV_p(\sigma_p(\gamma(t,p))) \frac{d}{dt}\sigma_p(\gamma(t,p))\Big|_{t=0} \\ &= dV_p(\mathbf{0})\dot{\sigma}_p(\delta\gamma(0,p)) = dV_p(\mathbf{0})\dot{\sigma}_p(Y(p)), \quad (89)\end{aligned}$$

and:

$$\frac{d}{dt} I_\mathcal{V}(\gamma(t,p))\Big|_{t=0} = \frac{d}{dt} I_{\mathcal{B},p}(\sigma_p\gamma(t,p))\Big|_{t=0} = dI_{\mathcal{B},p}(\mathbf{0})\dot{\sigma}_p(\delta\gamma(0,p)) = dI_{\mathcal{B},p}(\mathbf{0})\dot{\sigma}_p(Y(p))V_p(\mathbf{0}), \quad (90)$$

so that:

$$\dot{\sigma}(D_Y V(p)) = \frac{d}{dt}\dot{\sigma} V(\gamma(t,p))\Big|_{t=0} + \frac{1}{2} I_\mathcal{V}^{-1}(p) \frac{d}{dt} I_\mathcal{V}(\gamma(t,p))\Big|_{t=0}. \quad (91)$$

Let us check the symmetry of the metric covariant derivative to show that it is actually the unique Riemannian or Levi–Civita affine connection; see [6] (Th. 3.6).

The Lie bracket of the vector fields V and W is the vector field $[V,W]$, whose coordinates are:

$$[V,W]_p(\boldsymbol{\theta}) = dV_p(\mathbf{0})\dot{\sigma}_p(W(p)) - dW_p(\mathbf{0})\dot{\sigma}_p(V(p)). \tag{92}$$

As the ij entry of $\partial_k I_{\mathcal{B},p}(\mathbf{0})$ is $\partial_k\partial_i\partial_j\psi_p(\mathbf{0})$, then the symmetry $(dI_{\mathcal{B},p}(\mathbf{0})\boldsymbol{\alpha})\boldsymbol{\beta} = (dI_{\mathcal{B},p}(\mathbf{0})\boldsymbol{\beta})\boldsymbol{\alpha}$ holds, and we have:

$$\begin{aligned}\dot{\sigma}_p\left(D_W V(p) - D_V W(p)\right) = \\ dV_p(\mathbf{0})\dot{\sigma}_p(W(p)) + \frac{1}{2}I_{\mathcal{B}}^{-1}(p)\left(dI_{\mathcal{B},p}(\mathbf{0})\dot{\sigma}_p(W(p))\right)V_p(\mathbf{0}) \\ - dW_p(\mathbf{0})\dot{\sigma}_p(V(p)) - \frac{1}{2}I_{\mathcal{B}}^{-1}(p)\left(dI_{\mathcal{B},p}(\mathbf{0})\dot{\sigma}_p(V(p))\right)W_p(\mathbf{0}) \\ = \dot{\sigma}[V,W](p).\end{aligned} \tag{93}$$

The term $\Gamma^k(p) = \frac{1}{2}I_p^{-1}(\mathbf{0})\partial_k dI_{\mathcal{B},p}(\mathbf{0})$ of Equation (87) is sometimes referred to as the Christoffel matrix, but we do not use this terminology in this paper. As:

$$I_{\mathcal{B},p}(\boldsymbol{\theta}) = I_{\mathcal{B}}(e_p(\boldsymbol{\theta})) = \left[\mathrm{Cov}_{e_p(\boldsymbol{\theta})}\left(X_i, X_j\right)\right]_{i,j=1,\ldots,m} = \left[\partial_i\partial_j\psi_p(\boldsymbol{\theta})\right]_{i,j=i,\ldots,m}, \tag{94}$$

we have $\partial_k I_{\mathcal{B}}(e_p(\boldsymbol{\theta})) = \left[\partial_i\partial_j\partial_k\psi_p(\boldsymbol{\theta})\right]_{i,j=i,\ldots,m} = \left[\mathrm{Cov}_{e_p(\boldsymbol{\theta})}\left(X_i, X_j, X_k\right)\right]_{i,j=i,\ldots,m}$ and:

$$\Gamma^k(p) = \frac{1}{2}\left[\mathrm{Cov}_p\left(X_i, X_j\right)\right]^{-1}_{i,j=i,\ldots,m}\left[\mathrm{Cov}_p\left(X_i, X_j, X_k\right)\right]_{i,j=i,\ldots,m} \tag{95}$$

.

If V, W are vector fields of $T\mathcal{EV}$, we have:

$$\begin{aligned}\Gamma(p,V,W) &= \frac{1}{2}I_{\mathcal{B}}^{-1}(p)\,\mathrm{Cov}_p\left(\boldsymbol{X}, V, W\right) \\ &= \frac{1}{2}I_{\mathcal{B}}^{-1}(p)\,\mathbb{E}_p\left[{}^e\mathbb{U}^p\boldsymbol{X}VW\right],\end{aligned} \tag{96}$$

which is the projection of $V(p)W(p)/2$ on ${}^e\mathbb{U}^p\mathcal{V}$.

Notice also that:

$$(dI_p^{-1}(\mathbf{0})\boldsymbol{\alpha})I_{\mathcal{B},p}(\mathbf{0}) = -I_p^{-1}(\mathbf{0})(dI_{\mathcal{B},p}(\mathbf{0})\boldsymbol{\alpha})I_p^{-1}(\mathbf{0})I_{\mathcal{B},p}(\mathbf{0})\boldsymbol{y} = -I_p^{-1}(\mathbf{0})\left(dI_{\mathcal{B},p}(\mathbf{0})\boldsymbol{\alpha}\right). \tag{97}$$

3.2. Acceleration

Let $p(t)$, $t \in I$, be a smooth curve in $\mathcal{E}_{\mathcal{V}}$. Then, the velocity $\delta p(t) = \frac{d}{dt}\log\left(p(t)\right)$ is a vector field $V(p(t)) = \delta p(t)$, defined on the support $p(I)$ of the curve. As the curve is the flow of the velocity field, we can compute the metric derivative of the velocity along the the velocity itself $D_{\delta p}\delta p$ from Equation (91) with $V(p(0)) = \delta p(0)$; we can use Equation (91) to get:

$$\begin{aligned}\dot{\sigma}_p(D_{\delta p}\delta p)(p(0)) = \frac{d}{dt}\dot{\sigma}_{p(0)}\left(\delta(p(t))\right)\Big|_{t=0} + \frac{1}{2}I_{\mathcal{B}}^{-1}(p(0))\,\frac{d}{dt}I_{\mathcal{B}}(p(t))\Big|_{t=0} = \\ \frac{d^2}{dt^2}\sigma_{p(0)}(p(t))\Big|_{t=0} + \frac{1}{2}I_{\mathcal{B}}^{-1}(p(0))\,\frac{d}{dt}I_{\mathcal{B}}(p(t))\Big|_{t=0}.\end{aligned} \tag{98}$$

which can be defined to be the Riemannian acceleration of the curve at $t = 0$.

Let us write $\boldsymbol{\theta}(t) = \sigma_p(p(t))$, $p = p(0)$ and:

$$p(t) = \exp\left(\sum_{j=1}^{m} \theta^j(t)\, {}^e\mathbb{U}^p X_j - \psi_p(\boldsymbol{\theta}(t))\right) \cdot p, \tag{99}$$

so that $\dot{\sigma}_p(\delta p)(0) = \dot{\boldsymbol{\theta}}(0)$ and $\frac{d^2}{dt^2}\sigma_p(p(t))\Big|_{t=0} = \ddot{\boldsymbol{\theta}}(0)$. We have:

$$\frac{d}{dt} I_{\mathcal{B}}(p(t))\Big|_{t=0} = \frac{d}{dt} I_{\mathcal{B},p}(\boldsymbol{\theta}(t))\Big|_{t=0} = \frac{d}{dt} \operatorname{Hess} \psi_p(\boldsymbol{\theta}(t))\Big|_{t=0} = \operatorname{Cov}_p(\boldsymbol{X}, \boldsymbol{X}, \sum_{j=1}^{m} \dot{\theta}^j(t) X_j) \tag{100}$$

so that the acceleration at p has coordinates:

$$\ddot{\boldsymbol{\theta}}(0) + \frac{1}{2}\sum_{i,j=1}^{m} \dot{\theta}^i(0)\dot{\theta}^j(0) \operatorname{Cov}_p(\boldsymbol{X}, \boldsymbol{X})^{-1} \operatorname{Cov}_p(\boldsymbol{X}, X_i, X_j) =$$

$$\ddot{\boldsymbol{\theta}}(0) + \frac{1}{2} \operatorname{Cov}_p(\boldsymbol{X}, \boldsymbol{X})^{-1} \operatorname{Cov}_p(\boldsymbol{X}, \sum_{i}^{m} \dot{\theta}^i(0) X_i, \sum_{j=1}^{m} \dot{\theta}^j(0) X_j). \tag{101}$$

A geodesic is a curve whose acceleration is zero at each point. The exponential map is the mapping $\operatorname{Exp}\colon T\mathcal{E}_{\mathcal{V}} \to \mathcal{E}_{\mathcal{V}}$ defined by:

$$(p, U) \mapsto \operatorname{Exp}_p U = p(1), \tag{102}$$

where $t \mapsto p(t)$ is the geodesic, such that $p(0) = p$ and $\delta p(0) = U$, for all U, such that the geodesic exists for $t = 1$.

The exponential map is a particular retraction, that is, a family of mappings R_p, $p \in \mathcal{E}$, from the tangent space at p to the manifold; here $R\colon T_p\mathcal{E} \to \mathcal{E}$, such that $R_p(0) = p$ and $dR_p(0) = \mathrm{Id}$; see [4] (§5.4). It should be noted that exponential manifolds have natural retractions other than Exp, a notable one being the exponential family itself. A retraction provides a crucial step in a gradient search algorithms by mapping a direction of increase of the objective function to a new trial point.

3.2.1. Example: Binary Independent 2.5.2 Continued.

Let us consider the binary independent model of Section 2.5.2. We have

$$I_{\mathcal{B}}(e_\lambda(\boldsymbol{\theta})) = I_{\mathcal{B},\lambda}(\boldsymbol{\theta}) = \operatorname{diag}\left(\cosh^{-2}(\theta^j)\colon j = 1, \ldots, d\right), \tag{103}$$

it follows that

$$\begin{aligned} \partial_k I_{\mathcal{B},\lambda}(\boldsymbol{\theta}) &= \partial_k \operatorname{diag}\left(\cosh^{-2}(\theta^j)\colon j = 1, \ldots, d\right) \\ &= -2\cosh^{-3}(\theta^k)\sinh(\theta^k) E^{kk}, \end{aligned} \tag{104}$$

where E^{kk} is the $d \times d$ matrix with entry one at (k, k), zero otherwise. The k-th Christoffel's matrix in the second term in the definition of the metric derivative (aka Levi–Civita connection) is:

$$\Gamma^k_{\mathcal{B}}(e_\lambda(\boldsymbol{\theta})) = \Gamma^k_\lambda(\boldsymbol{\theta}) = \frac{1}{2} I^{-1}_{\mathcal{B},\lambda}(\boldsymbol{\theta}) \partial_k I_{\mathcal{B},\lambda}(\boldsymbol{\theta}) = -\tanh(\theta^k) E^{kk}. \tag{105}$$

In terms of the moments, we have $I_{\mathcal{B},\lambda}(\boldsymbol{\theta}) = \mathrm{Cov}_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{X'}) = \mathrm{Hess}\,\psi_\lambda(\boldsymbol{\theta})$. As $\partial_k\partial_i\partial_j\psi_\lambda(\boldsymbol{\theta}) = \mathrm{Cov}_{\boldsymbol{\theta}}(X_k, X_i, X_j)$, we that can write:

$$\begin{aligned}\partial_k I_{\mathcal{B},\lambda}(\boldsymbol{\theta}) &= \partial_k \,\mathrm{diag}\,(\mathrm{Var}_{\boldsymbol{\theta}}(X_j) : j = 1, \ldots, d)\\ &= \mathrm{Cov}_{\boldsymbol{\theta}}(X_k, X_k, X_k)\, E^{kk}\end{aligned} \tag{106}$$

and:

$$\begin{aligned}\Gamma^k_\lambda(\boldsymbol{\theta}) &= \frac{1}{2}\,\mathrm{Cov}_{\boldsymbol{\theta}}(X_k, X_k)^{-1}\,\mathrm{Cov}_{\boldsymbol{\theta}}(X_k, X_k, X_k)\, E^{kk}\\ &= \frac{1}{2}(1 - (\eta^k)^2)^{-1}(-2\eta^k + 2(\eta^k)^3)E^{kk} = -\eta^k E^{kk}.\end{aligned} \tag{107}$$

The equations for the geodesics starting from $\boldsymbol{\theta}(0)$ with velocity $\dot{\boldsymbol{\theta}}(0) = \boldsymbol{u}$ are:

$$\ddot{\theta}^k(t) + \sum_{ij=1}^{m} \Gamma^k_{ij}(\boldsymbol{\theta}(t))\dot{\theta}^i(t)\dot{\theta}^j(t) = \ddot{\theta}^k(t) - \tanh(\theta^k(t))(\dot{\theta}^k(t))^2 = 0, \quad k = 1, \ldots, d. \tag{108}$$

The ordinary differential equation:

$$\ddot{\theta} - \tanh(\theta)\dot{\theta}^2 = 0 \tag{109}$$

has the closed form solution:

$$\theta(t) = \mathrm{gd}^{-1}\left(\mathrm{gd}(\theta(0)) + \frac{\dot{\theta}(0)}{\cosh(\theta(0))}t\right) = \tanh^{-1}\left(\sin\left(\mathrm{gd}(\theta(0)) + \frac{\dot{\theta}(0)}{\cosh(\theta(0))}t\right)\right) \tag{110}$$

for all t, such that:

$$-\pi/2 < \mathrm{gd}(\theta(0)) + \frac{\dot{\theta}(0)}{\cosh(\theta(0))}t < \pi/2, \tag{111}$$

where $\mathrm{gd}\colon \mathbb{R} \to]-\pi/2, +\pi/2[$ is the Gudermannian function, that is, $\mathrm{gd}'(x) = 1/\cosh x$, $\mathrm{gd}(0) = 0$; in closed form, $\mathrm{gd}(x) = \arcsin(\tanh(x))$. In fact, if θ is a solution of Equation (109), then:

$$\frac{d}{dt}\,\mathrm{gd}(\theta(t)) = \frac{\dot{\theta}(t)}{\cosh(\theta(t))} \tag{112}$$

$$\begin{aligned}\frac{d^2}{dt^2}\,\mathrm{gd}(\theta(t)) &= -\frac{\sinh(\theta(t))\dot{(}\theta(t))^2}{\cosh^2(\theta(t))} + \frac{\ddot{\theta}(t)}{\cosh(\theta(t))}\\ &= \frac{1}{\cosh(\theta(t))}\left(\ddot{\theta}(t) - \tanh(\theta(t))(\dot{\theta}(t))^2\right) = 0,\end{aligned} \tag{113}$$

so that $t \mapsto \mathrm{gd}(\theta(t))$ coincides (where it is defined) with an affine function characterized by the initial conditions.

In particular, at $t = 1$, the geodesic Equation (110) defines the Riemannian exponential $\mathrm{Exp}\colon T\mathcal{E}_\mathcal{V} \to \mathcal{E}_\mathcal{V}$. If $(p, U) \in T\mathcal{E}_\mathcal{V}$, that is, $p \in \mathcal{E}_\mathcal{V}$ and $U \in T_p\mathcal{E}_\mathcal{V}$, then $\sigma_\lambda(p) = \boldsymbol{\theta}(\mathbf{0})$ and $U = \sum u_j\, {}^e\mathbb{U}^p X_j$, $\dot{\sigma}_\lambda(U) = \boldsymbol{u}$. If:

$$-\pi/2 < \mathrm{gd}(\theta^j) + \frac{u_j}{\cosh(\theta^j)} < \pi/2, \tag{114}$$

then we can take $\dot{\boldsymbol{\theta}}(0) = \boldsymbol{u}$ and $t = 1$, so that:

$$\operatorname{Exp}_p \colon U \overset{\dot{\sigma}_\lambda}{\longmapsto} \boldsymbol{u} \mapsto \left[\operatorname{gd}^{-1}\left(\operatorname{gd}(\theta^j) + \frac{u_j}{\cosh(\theta^j)}\right) : j = 1, \ldots, d\right] \overset{e_\lambda}{\longmapsto}$$
$$\prod_{j=1}^{m} \exp\left(\operatorname{gd}^{-1}\left(\operatorname{gd}(\theta^j) + \frac{u_j}{\cosh(\theta^j)}\right) X_j - \psi\left(\operatorname{gd}^{-1}\left(\operatorname{gd}(\theta^j) + \frac{u_j}{\cosh(\theta^j)}\right)\right)\right) 2^{-m}. \quad (115)$$

We have:

$$\exp\left(\operatorname{gd}^{-1}(v)\right) = \exp\left(\tanh^{-1}(\sin(v))\right) = \sqrt{\frac{1+\sin v}{1-\sin v}} \quad (116)$$

and:

$$\psi\left(\operatorname{gd}^{-1}(v)\right) = +\log\left(\operatorname{gd}^{-1}(\sin v)\right) = \log\left(\frac{1}{\cos v}\right), \quad (117)$$

hence $\boldsymbol{u} \mapsto \operatorname{Exp}_p\left(\sum_{j=1}^{d} u_j \, {}^e\mathbb{U}^p X_j\right)$ is given for:

$$\boldsymbol{u} \in \bigtimes_{j=1}^{d} \left]\cosh(\theta^j)(-\pi/2 - \operatorname{gd}(\theta^j)), \cosh(\theta^j)(\pi/2 - \operatorname{gd}(\theta^j))\right[, \quad (118)$$

by:

$$\operatorname{Exp}_{\boldsymbol{\theta}}(\boldsymbol{u}) = \prod_{j=1}^{m} \cos\left(\operatorname{gd}(\theta^j) + \frac{u_j}{\cosh(\theta^j)}\right)\left(\frac{1+\sin\left(\operatorname{gd}(\theta^j) + \frac{u_j}{\cosh(\theta^j)}\right)}{1-\sin\left(\operatorname{gd}(\theta^j) + \frac{u_j}{\cosh(\theta^j)}\right)}\right)^{\frac{X_j}{2}} =$$
$$\prod_{j=1}^{m}\left(1+\sin\left(\operatorname{gd}(\theta^j) + \frac{u_j}{\cosh(\theta^j)}\right) X_j\right) 2^{-m} \in \mathcal{E}_{\mathcal{V}}. \quad (119)$$

The expectation parameters are:

$$\eta^i(t) = \mathbb{E}_{\boldsymbol{\theta}=\mathbf{0}}\left[X_i \prod_{j=1}^{m}\left(1+\sin\left(\operatorname{gd}(\theta^j) + \frac{tu_j}{\cosh(\theta^j)}\right) X_j\right)\right] = \sin\left(\operatorname{gd}(\theta^j) + \frac{tu_j}{\cosh(\theta^j)}\right), \quad (120)$$

and:

$$\operatorname{gd}(\theta^j) = \arcsin(\eta^j), \quad \cosh(\theta^j) = \frac{1}{(1-(\eta^j)^2)^{\frac{1}{2}}}, \quad (121)$$

so that the exponential in terms of the expectation parameters is:

$$\operatorname{Exp}_{\boldsymbol{\eta}}(\boldsymbol{u}) = \left(\sin\left(\arcsin\eta^j + \left(1-(\eta^j)^2\right)^{\frac{1}{2}} u_j\right) : j = 1, \ldots, m\right). \quad (122)$$

The inverse of the Riemannian exponential provides a notion of translation between two elements of the exponential model, which is a particular parametrization of the model:

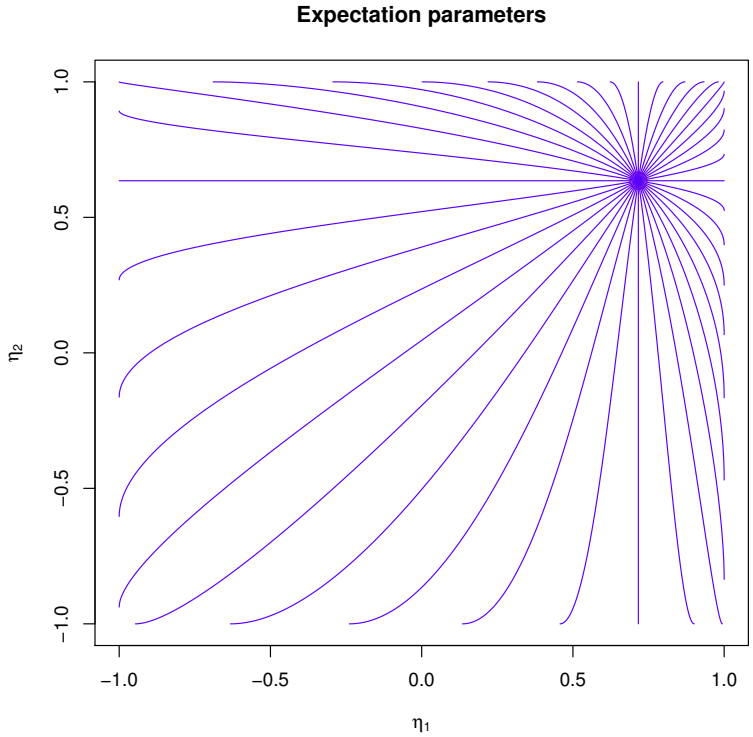
$$\overrightarrow{\boldsymbol{\eta}_1\boldsymbol{\eta}_2} = \operatorname{Exp}_{\boldsymbol{\eta}_1}^{-1}\boldsymbol{\eta}_2 = \left[\left((1-(\eta_i^j)^2\right)^{-\frac{1}{2}}\left(\arcsin\eta_2^j - \arcsin\eta_1^j\right) : j = 1, \ldots, m\right] \quad (123)$$

In particular, at $\boldsymbol{\theta} = \mathbf{0}$, we have the geodesic:

$$t \mapsto \prod_{j=1}^{d}\left(1+\sin(tu_j)X_j\right) 2^{-m}, \quad |t| < \frac{\pi}{2\max|u_j|} \quad (124)$$

See in Figure 5 some geodesic curves.

Figure 5. Geodesics from $\boldsymbol{\eta} = (0.75, 0.75)$.



3.3. Riemannian Hessian

Let $\phi\colon \mathcal{E}_{\mathcal{V}} \to \mathbb{R}$ with Riemannian gradient $\nabla\phi(p) = \sum_i (\widetilde{\nabla}\phi)_i(p)\, {}^e\mathbb{U}^p X_i$, $\widetilde{\nabla}\phi(p) = I_{\mathcal{B}}^{-1}(p)\nabla\phi_p(\mathbf{0})$. The Riemannian Hessian of ϕ is the metric derivative of the gradient $\nabla\phi$ along the vector field Y, that is, $\mathrm{Hess}_Y\, \phi = D_Y \nabla\phi$; see [6] (Ch. 6, Ex. 11), [4] (§5.5). in the following, we denote by the symbol Hess, without a subscript, the ordinary Hessian matrix.

From Equation (87), we have the coordinates of $\mathrm{Hess}_Y\, \phi(p)$. Given a generic tangent vector $\boldsymbol{\alpha}$, we compute from Equation (38):

$$\begin{aligned} d(\nabla\phi)_p(\boldsymbol{\theta})\boldsymbol{\alpha}|_{\boldsymbol{\theta}=\mathbf{0}} &= d\left(I_{\mathcal{B},p}^{-1}(\boldsymbol{\theta})\nabla\phi_p(\boldsymbol{\theta})\right)\boldsymbol{\alpha}\big|_{\boldsymbol{\theta}=\mathbf{0}} \\ &= (dI_{\mathcal{B},p}^{-1}(\mathbf{0})\boldsymbol{\alpha})\nabla\phi_p(\mathbf{0}) + I_{\mathcal{B},p}^{-1}(\mathbf{0})\,\mathrm{Hess}\,\phi_p(\mathbf{0})\boldsymbol{\alpha} \\ &= -I_{\mathcal{B}}^{-1}(p)(dI_{\mathcal{B},p}(\mathbf{0})\boldsymbol{\alpha})\widetilde{\nabla}\phi(p) + I_{\mathcal{B}}^{-1}(p)\,\mathrm{Hess}\,\phi_p(\mathbf{0})\boldsymbol{\alpha} \end{aligned} \tag{125}$$

and, upon substitution of $(\nabla\phi)_p$ to V_p in Equation (87),

$$\begin{aligned} \dot{\sigma}_p(\mathrm{Hess}_Y\, \phi(p)) &= d(\nabla\phi)_p(\mathbf{0})\boldsymbol{\alpha} + \frac{1}{2} I_{\mathcal{B}}^{-1}(p)\,(dI_{\mathcal{B},p}(\mathbf{0})\boldsymbol{\alpha})\,(\nabla\phi)_p(\mathbf{0}), \quad \boldsymbol{\alpha} = S_p(Y(p)) \\ &= -I_{\mathcal{B}}^{-1}(p)(dI_{\mathcal{B},p}(\mathbf{0})\boldsymbol{\alpha})\widetilde{\nabla}\phi(p) + I_{\mathcal{B}}^{-1}(p)\,\mathrm{Hess}\,\phi_p(\mathbf{0}) + \frac{1}{2} I_{\mathcal{B}}^{-1}(p)\,(dI_{\mathcal{B},p}(\mathbf{0})\boldsymbol{\alpha})\,\widetilde{\nabla}\phi(p) \\ &= I_{\mathcal{B}}^{-1}(p)\,\mathrm{Hess}\,\phi_p(\mathbf{0})\boldsymbol{\alpha} - \frac{1}{2} I_{\mathcal{B}}^{-1}(p)\,(dI_{\mathcal{B},p}(\mathbf{0})\boldsymbol{\alpha})\,\widetilde{\nabla}\phi(p) \\ &= I_{\mathcal{B}}^{-1}(p)\left(\mathrm{Hess}\,\phi_p(\mathbf{0})\boldsymbol{\alpha} - \frac{1}{2}\,(dI_{\mathcal{B},p}(\mathbf{0})\boldsymbol{\alpha})\,\widetilde{\nabla}\phi(p)\right) \end{aligned} \tag{126}$$

$\mathrm{Hess}_Y\,\phi$ is characterized by knowing the value of $g(\mathrm{Hess}_Y\,\phi, X)\colon \mathcal{E}_\mathcal{V}$ for all vector fields X. We have from Equation (126), with $\boldsymbol{\alpha} = \dot{\sigma}_p(Y(p))$ and $\boldsymbol{\beta} = \dot{\sigma}_p(X(p))$,

$$g_p(\mathrm{Hess}_{Y(p)}\,\phi(p), X(p)) = \boldsymbol{\beta}'\,\mathrm{Hess}\,\phi_p(\mathbf{0})\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\beta}'\,(dI_{\mathcal{B},p}(\mathbf{0})\boldsymbol{\alpha})\,\widetilde{\nabla}\phi(p). \tag{127}$$

This is the presentation of the Riemannian Hessian as a bi-linear form on $T\mathcal{E}_\mathcal{V}$; see the comments in [4] (Prop. 5.5.2-3). Note that the Riemannian Hessian is positive definite if:

$$\boldsymbol{\alpha}'\,\mathrm{Hess}\,\phi_p(\mathbf{0})\boldsymbol{\alpha} \geq \frac{1}{2}\boldsymbol{\alpha}'\,(dI_{\mathcal{B},p}(\mathbf{0})\boldsymbol{\alpha})\,\widetilde{\nabla}\phi(p), \quad \boldsymbol{\alpha} \in \mathbb{R}^m. \tag{128}$$

4. Application to Combinatorial Optimization

We conclude our paper by showing how the geometric method applies to the problem of finding the maximum of the expected value of a function.

4.1. Hessian of a Relaxed Function

Here is a key example of vector field. Let f be any bounded random variable, and define the relaxed function to be $\phi(p) = \mathbb{E}_p[f]$, $p \in \mathcal{P}_>$. Define $F(p)$ to be the projection of f, as an element of $L^2(p)$, onto $T_p\mathcal{E}_\mathcal{V} = {}^e\mathbb{U}^p\mathcal{V}$, *i.e.*, $F(p)$ is the element of ${}^e\mathbb{U}^p\mathcal{V}$, such that:

$$\mathbb{E}_p[(f - F(p))v] = 0, \quad v \in {}^e\mathbb{U}^p\mathcal{V} \tag{129}$$

In the basis ${}^e\mathbb{U}^p\mathcal{B}$, we have $F(p) = \sum_i \hat{f}_{p,i}\,{}^e\mathbb{U}^p X_i$ and:

$$\mathrm{Cov}_p(f, X_j) = \sum_i \hat{f}_{p,i}\,\mathbb{E}_p[{}^e\mathbb{U}^p X_i\,{}^e\mathbb{U}^p X_j], \quad j = 1, \ldots, m, \tag{130}$$

so that $\hat{\boldsymbol{f}}_p = I_\mathcal{B}^{-1}(p)\,\mathrm{Cov}_p(\boldsymbol{X}, f)$ and

$$F(p) = \hat{\boldsymbol{f}}_p'\,{}^e\mathbb{U}^p\boldsymbol{X} = \mathrm{Cov}_p(f, \boldsymbol{X})\,I_\mathcal{B}^{-1}(p)\,{}^e\mathbb{U}^p\boldsymbol{X}. \tag{131}$$

Let us compute the gradient of the relaxed function $\phi = \mathbb{E}.[f] : \mathcal{E}_\mathcal{V}$. We have $\phi_p(\boldsymbol{\theta}) = \mathbb{E}_{e_p(\boldsymbol{\theta})}[f]$, and from the properties of exponential families, the Euclidean gradient is $\nabla\phi_p(\mathbf{0}) = \mathrm{Cov}_p(f, \boldsymbol{X})$. It follows that the natural gradient is:

$$\widetilde{\nabla}\phi_p(\mathbf{0}) = I_\mathcal{B}^{-1}(p)\,\mathrm{Cov}_p(\boldsymbol{X}, f) = \hat{\boldsymbol{f}}, \tag{132}$$

and the Riemannian gradient is $\nabla\phi(p) = F(p)$.

From the properties of exponential families, we have:

$$\mathrm{Hess}\,\phi_p(\mathbf{0}) = \mathrm{Cov}_p(\boldsymbol{X}, \boldsymbol{X}, f),$$

so that, in this case, Equation (127), when written in terms of the moments, is:

$$\boldsymbol{\beta}'\,\mathrm{Cov}_p(\boldsymbol{X}, \boldsymbol{X}, f)\,\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\beta}'\,\mathrm{Cov}_p(\boldsymbol{X}, \boldsymbol{X}, \boldsymbol{\alpha}\cdot\boldsymbol{X})\,\mathrm{Cov}_p(\boldsymbol{X}, \boldsymbol{X})^{-1}\,\mathrm{Cov}_p(\boldsymbol{X}, f). \tag{133}$$

4.1.1. Example: Binary Independent 2.5.2 and 3.2.1 Continued

We list below the computation of the Hessian in the case of two binary independent variables. Computations were done with Sage [22], which allows both the reduction $x_i^2 = 1$ in the ring of polynomials and the simplifications in the symbolic ring of parameters.

$$\mathrm{Cov}_{\boldsymbol{\eta}}(\boldsymbol{X}, f) = \begin{pmatrix} -(\eta_1^2-1)a_1 - (\eta_1^2\eta_2 - \eta_2)a_{12} \\ -(\eta_2^2-1)a_2 - (\eta_1\eta_2^2 - \eta_1)a_{12} \end{pmatrix} = \begin{pmatrix} -(\eta_1-1)(\eta_1+1)(a_{12}\eta_2 + a_1) \\ -(\eta_2-1)(\eta_2+1)(a_{12}\eta_1 + a_2) \end{pmatrix} \tag{134}$$

$$\mathrm{Cov}_{\boldsymbol{\eta}}(\boldsymbol{X}, \boldsymbol{X}) = \begin{pmatrix} -\eta_1^2+1 & 0 \\ 0 & -\eta_2^2+1 \end{pmatrix} = \begin{pmatrix} -(\eta_1-1)(\eta_1+1) & 0 \\ 0 & -(\eta_2-1)(\eta_2+1) \end{pmatrix} \tag{135}$$

$$\mathrm{Cov}_{\boldsymbol{\eta}}(\boldsymbol{X}, \boldsymbol{X})^{-1}\,\mathrm{Cov}_{\boldsymbol{\eta}}(\boldsymbol{X}, f) = \begin{pmatrix} a_{12}\eta_2 + a_1 \\ a_{12}\eta_1 + a_2 \end{pmatrix} = \nabla F(\boldsymbol{\eta}) \tag{136}$$

$$\mathrm{Cov}_{\boldsymbol{\eta}}(\boldsymbol{X}, \boldsymbol{X}, f) = \begin{pmatrix} 2(\eta_1^3 - \eta_1)a_1 + 2(\eta_1^3\eta_2 - \eta_1\eta_2)a_{12} & (\eta_1^2\eta_2^2 - \eta_1^2 - \eta_2^2 + 1)a_{12} \\ (\eta_1^2\eta_2^2 - \eta_1^2 - \eta_2^2 + 1)a_{12} & 2(\eta_1\eta_2^3 - \eta_1\eta_2)a_{12} + 2(\eta_2^3 - \eta_2)a_2 \end{pmatrix} = \begin{pmatrix} 2(\eta_1-1)(\eta_1+1)(a_{12}\eta_2 + a_1)\eta_1 & (\eta_2-1)(\eta_2+1)(\eta_1-1)(\eta_1+1)a_{12} \\ (\eta_2-1)(\eta_2+1)(\eta_1-1)(\eta_1+1)a_{12} & 2(\eta_2-1)(\eta_2+1)(a_{12}\eta_1 + a_2)\eta_2 \end{pmatrix} \tag{137}$$

$$\mathrm{Cov}_{\boldsymbol{\eta}}(\boldsymbol{X}, \boldsymbol{X})^{-1}\,\mathrm{Cov}_{\boldsymbol{\eta}}(\boldsymbol{X}, \boldsymbol{X}, f) = \begin{pmatrix} -2(a_{12}\eta_2 + a_1)\eta_1 & -a_{12}\eta_2^2 + a_{12} \\ -a_{12}\eta_1^2 + a_{12} & -2(a_{12}\eta_1 + a_2)\eta_2 \end{pmatrix} \tag{138}$$

$$\mathrm{Cov}_{\boldsymbol{\eta}}(\boldsymbol{X}, \boldsymbol{X}, \nabla F(\boldsymbol{\eta})) = \begin{pmatrix} 2(a_{12}\eta_2 + a_1)(\eta_1+1)(\eta_1-1)\eta_1 & 0 \\ 0 & 2(a_{12}\eta_1 + a_2)(\eta_2+1)(\eta_2-1)\eta_2 \end{pmatrix} \tag{139}$$

$$\mathrm{Cov}_{\boldsymbol{\eta}}(\boldsymbol{X}, \boldsymbol{X})^{-1}\,\mathrm{Cov}_{\boldsymbol{\eta}}(\boldsymbol{X}, \boldsymbol{X}, \nabla F(\boldsymbol{\eta})) = \begin{pmatrix} -2(a_{12}\eta_2 + a_1)\eta_1 & 0 \\ 0 & -2(a_{12}\eta_1 + a_2)\eta_2 \end{pmatrix} \tag{140}$$

The Riemannian Hessian as a matrix in the basis of the tangent space is:

$$\mathrm{Hess}\, F(\boldsymbol{\eta}) = \mathrm{Cov}_{\boldsymbol{\eta}}(\boldsymbol{X}, \boldsymbol{X})^{-1}\left(\mathrm{Cov}_{\boldsymbol{\eta}}(\boldsymbol{X}, \boldsymbol{X}, f) - \frac{1}{2}\mathrm{Cov}_{\boldsymbol{\eta}}(\boldsymbol{X}, \boldsymbol{X}, \nabla F(\boldsymbol{\eta}))\right) = \begin{pmatrix} -(a_{12}\eta_2 + a_1)\eta_1 & -a_{12}(\eta_2+1)(\eta_2-1) \\ -a_{12}(\eta_1+1)(\eta_1-1) & -(a_{12}\eta_1 + a_2)\eta_2 \end{pmatrix} \tag{141}$$

As a check, let us compute the Riemannian Hessian as a natural Hessian in the Riemannian parameters, $\mathrm{Hess}\ \phi \circ \mathrm{Exp}_p(\boldsymbol{u})\big|_{\boldsymbol{u}=0}$; see [4] (Prop. 5.5.4). We have:

$$F \circ \mathrm{Exp}_{\boldsymbol{\eta}}(\boldsymbol{u}) = a_{12} \sin\left(\sqrt{-\eta_1^2+1}u_1 + \arcsin(\eta_1)\right) \sin\left(\sqrt{-\eta_2^2+1}u_2 + \arcsin(\eta_2)\right) + a_1 \sin\left(\sqrt{-\eta_1^2+1}u_1 + \arcsin(\eta_1)\right) + a_2 \sin\left(\sqrt{-\eta_2^2+1}u_2 + \arcsin(\eta_2)\right) \quad (142)$$

and:

$$\mathrm{Hess}\, F \circ \mathrm{Exp}_{\boldsymbol{\eta}}(\boldsymbol{u})\big|_{\boldsymbol{u}=\mathbf{0}} = \begin{pmatrix} (\eta_1^2-1)a_{12}\eta_1\eta_2 + (\eta_1^2-1)a_1\eta_1 & (\eta_1^2-1)(\eta_2^2-1)a_{12} \\ (\eta_1^2-1)(\eta_2^2-1)a_{12} & (\eta_2^2-1)a_{12}\eta_1\eta_2 + (\eta_2^2-1)a_2\eta_2 \end{pmatrix} = \begin{pmatrix} (a_{12}\eta_2 + a_1)(\eta_1+1)(\eta_1-1)\eta_1 & a_{12}(\eta_1+1)(\eta_1-1)(\eta_2+1)(\eta_2-1) \\ a_{12}(\eta_1+1)(\eta_1-1)(\eta_2+1)(\eta_2-1) & (a_{12}\eta_1 + a_2)(\eta_2+1)(\eta_2-1)\eta_2 \end{pmatrix}. \quad (143)$$

Note the presence of the factor $\mathrm{Cov}_{\boldsymbol{\eta}}(\boldsymbol{X}, \boldsymbol{X})$.

4.2. Newton Method

The Newton method is an iterative method that generates a sequence of points p_t, with $t = 0, 1, \ldots$, that converges towards a stationary point $\hat{p}$ of a $F(p) = \mathbb{E}_p[f]$, $p \in \mathcal{E}_{\mathcal{V}}$, that is, a critical point of the vector field $p \mapsto \nabla F(p)$, $\nabla F(\hat{p}) = 0$. Here, we follow [4] (Ch. 5–6), and in particular Algorithm 5 on Page 113.

Let ∇F be a gradient field. We reproduce in our case the basic derivation of the Newton method in the following. Note that, in this section, we use the notation $\mathrm{Hess}\, \bullet[\boldsymbol{\alpha}]$ to denote $\mathrm{Hess}_{\boldsymbol{\alpha}}\, \bullet$. Using the definition of metric derivative, we have for a geodesic curve $[0,1] \ni t \mapsto p(t) \in \mathcal{E}_{\mathcal{V}}$ connecting $p = p(0)$ to $\hat{p} = p(1)$ that:

$$\frac{d}{dt} g_{p(t)}\left(\nabla F(p(t)), \delta p(t)\right) = g_{p(t)}\left(\mathrm{Hess}\, F(p(t))[\delta p(t)], \delta p(t)\right) \quad (144)$$

hence the increment from p to $\hat{p}$ is:

$$g_{\hat{p}}\left(\nabla F(\hat{p}), \delta p(1)\right) - g_p\left(\nabla F(p), \delta p(0)\right) = \int_0^1 g_{p(t)}\left(\mathrm{Hess}\, F(p(t))[\delta p(t)], \delta p(t)\right)\, dt. \quad (145)$$

Now, we assume that $\nabla F(\hat{p}) = 0$ and that in Equation (145), the integral is approximated by the initial value of the integrand, that is to say, the Hessian is approximately constant on the geodesic from p to $\hat{p}$; we obtain:

$$-\, g_p\left(\nabla F(p), \delta p(0)\right) = g_p\left(\mathrm{Hess}\, F(p)[\delta p(0)], \delta p(0)\right) + \epsilon. \quad (146)$$

If we can solve the Newton equation:

$$\mathrm{Hess}\, F(p(t))[\boldsymbol{u}] = -\nabla F(p) \quad (147)$$

then $\boldsymbol{u}$ is approximately equal to the initial velocity of the geodesic connecting p to $\hat{p}$, that is, $\hat{p} = \mathrm{Exp}_p(\boldsymbol{u})$.

The particular structure of the exponential manifold suggests at least two natural retractions that could be used to move from $\boldsymbol{u}$ to $\hat{p}$. Namely, we have the Riemannian exponential $(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}) \mapsto \mathrm{Exp}_{\boldsymbol{\theta}_t}(\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t)$ and the e-retraction coming from the exponential family itself and defined by $(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1}) \mapsto e_{\boldsymbol{\theta}_t}(\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t)$, with $\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t = \boldsymbol{u}_t$.

In the $\boldsymbol{\theta}$ parameters, with the e-retraction, the Newton method generates a sequence $(\boldsymbol{\theta}_t)$ according to the following updating rule:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \lambda \,\mathrm{Hess}\, F(\boldsymbol{\theta}_t)^{-1} \widetilde{\nabla} F(\boldsymbol{\theta}_t) \tag{148}$$

where $\lambda > 0$ is an extra parameter intended to control the step size and, in turn, the convergence to $\hat{\theta}$; see [5].

We can rewrite Equation (148) in terms of covariances as:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \lambda \left(\mathrm{Cov}_{\boldsymbol{\theta}_t}(\boldsymbol{X}, \boldsymbol{X}, f) - \frac{1}{2} \mathrm{Cov}_{\boldsymbol{\theta}_t}(\boldsymbol{X}, \boldsymbol{X}, \widetilde{\nabla} F(\boldsymbol{\theta}_t)) \right)^{-1} \widetilde{\nabla} F(\boldsymbol{\theta}_t). \tag{149}$$

4.3. Example: Binary Independent

In the $\boldsymbol{\eta}$ parameters, the Newton step is:

$$\boldsymbol{u} = -\,\mathrm{Hess}\, F(\boldsymbol{\eta})^{-1} \nabla F(\boldsymbol{\eta}) = \begin{pmatrix} \frac{a_{12}^2\eta_1 + a_{12}a_2 + (a_1a_{12}\eta_1 + a_1a_2)\eta_2}{a_{12}^2\eta_1^2 + \left(a_{12}a_2\eta_1 + a_{12}^2\right)\eta_2^2 - a_{12}^2 + \left(a_1a_{12}\eta_1^2 + a_1a_2\eta_1\right)\eta_2} \\ \frac{a_1a_2\eta_1 + a_1a_{12} + \left(a_{12}a_2\eta_1 + a_{12}^2\right)\eta_2}{a_{12}^2\eta_1^2 + \left(a_{12}a_2\eta_1 + a_{12}^2\right)\eta_2^2 - a_{12}^2 + \left(a_1a_{12}\eta_1^2 + a_1a_2\eta_1\right)\eta_2} \end{pmatrix} \tag{150}$$

and the new $\boldsymbol{\eta}$ in the Riemannian retraction is:

$$\mathrm{Exp}_{\boldsymbol{\eta}}(\boldsymbol{u}) = \begin{pmatrix} \sin\left(\frac{\left(a_{12}^2\eta_1 + a_{12}a_2 + (a_1a_{12}\eta_1 + a_1a_2)\eta_2\right)\sqrt{-\eta_1^2+1}}{a_{12}^2\eta_1^2 + \left(a_{12}a_2\eta_1 + a_{12}^2\right)\eta_2^2 - a_{12}^2 + \left(a_1a_{12}\eta_1^2 + a_1a_2\eta_1\right)\eta_2} + \arcsin(\eta_1) \right) \\ \sin\left(\frac{\left(a_1a_2\eta_1 + a_1a_{12} + \left(a_{12}a_2\eta_1 + a_{12}^2\right)\eta_2\right)\sqrt{-\eta_2^2+1}}{a_{12}^2\eta_1^2 + \left(a_{12}a_2\eta_1 + a_{12}^2\right)\eta_2^2 - a_{12}^2 + \left(a_1a_{12}\eta_1^2 + a_1a_2\eta_1\right)\eta_2} + \arcsin(\eta_2) \right). \end{pmatrix} \tag{151}$$

In Figure 6, we represented the vector field associated with the Newton step in the $\boldsymbol{\eta}$ parameters, with $\lambda = 0.05$, using the Riemannian retraction, for the case $a_1 = 1$, $a_2 = 2$ and $a_{12} = 3$, with:

$$\mathrm{Exp}_{\boldsymbol{\eta}}(\boldsymbol{u}) = \begin{pmatrix} \sin\left(\lambda \frac{\sqrt{-\eta_1^2+1}\left((3\,\eta_1+2)\eta_2 + 9\,\eta_1 + 6\right)}{3\,(2\,\eta_1+3)\eta_2^2 + 9\,\eta_1^2 + \left(3\,\eta_1^2 + 2\,\eta_1\right)\eta_2 - 9} + \arcsin(\eta_1) \right) \\ \sin\left(\lambda \frac{(3\,(2\,\eta_1+3)\eta_2 + 2\,\eta_1 + 3)\sqrt{-\eta_2^2+1}}{3\,(2\,\eta_1+3)\eta_2^2 + 9\,\eta_1^2 + \left(3\,\eta_1^2 + 2\,\eta_1\right)\eta_2 - 9} + \arcsin(\eta_2) \right) \end{pmatrix}. \tag{152}$$

The red dotted lines represented in the figure identify the basins of attraction of the vector field and correspond to the solutions of the explicit equation in $\boldsymbol{\eta}$ for which the Newton step $\boldsymbol{u}$ is not defined. This vector field can be compared to that in Figure 7, associated with the Newton step for $F(\boldsymbol{\eta})$ using the Euclidean geometry. In the Euclidean geometry, $F(\boldsymbol{\eta})$ is a quadratic function with one saddle point, so that from any $\boldsymbol{\eta}$, the Newton step points in the direction of the critical point. This makes the Newton step unsuitable for an optimization algorithm. On the other side, in the Riemannian geometry, the vertices of the polytope are critical points for $F(\boldsymbol{\eta})$, and they determine the presence of multiple basins of attraction, as expected.

Figure 6. The Newton step in the $\boldsymbol{\eta}$ parameters, Riemannian retraction, $\lambda = 0.05$. The red dotted lines identify the different basins of attraction and correspond to the points for which the Newton step is not defined; *cf.* Equation (150). The instability close to the critical lines is represented by the longer arrows.

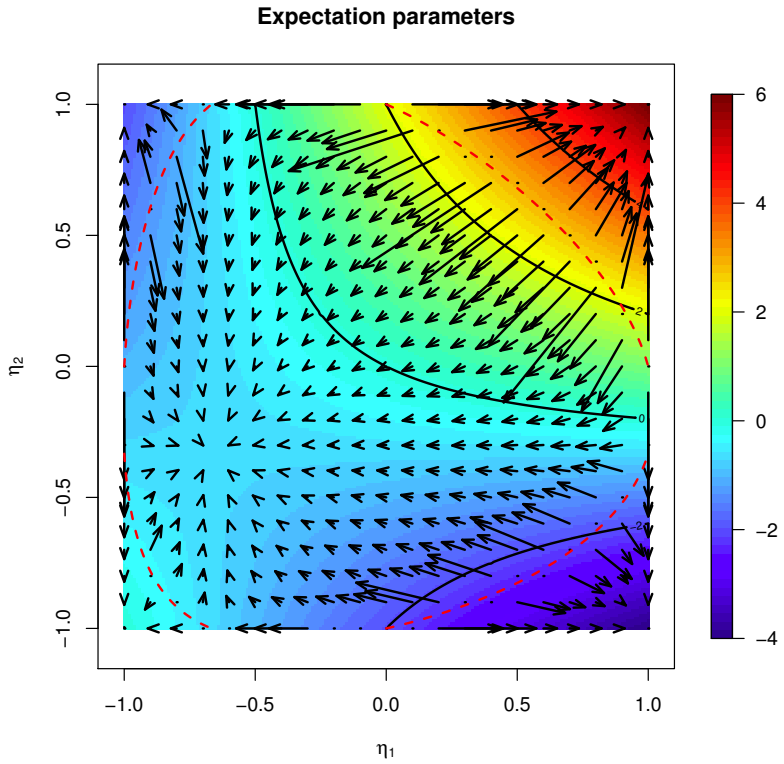


Figure 7. The Newton step in the $\boldsymbol{\eta}$ parameters, Euclidean geometry, $\lambda = 0.05$.

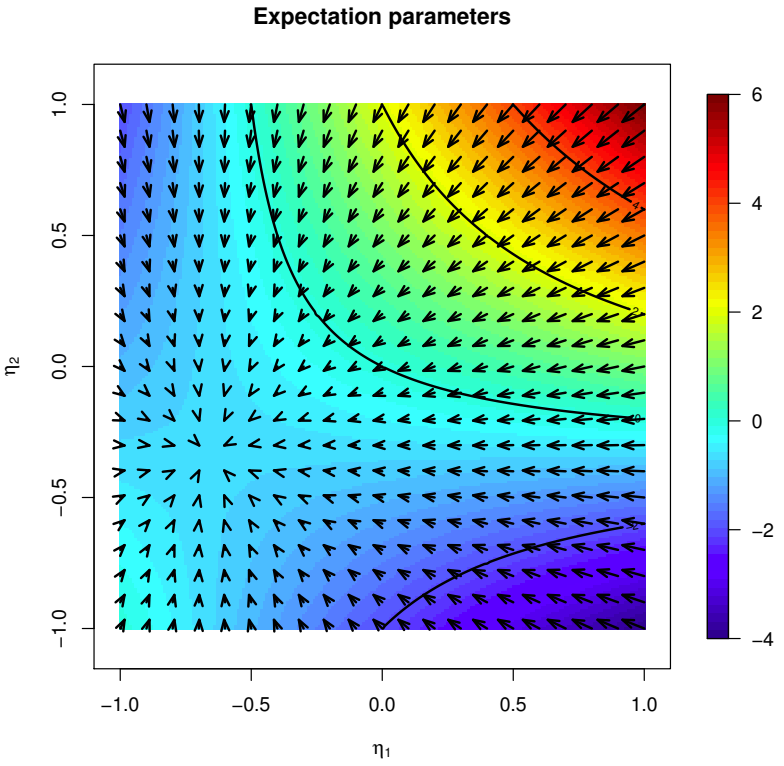


Figure 8. The Newton step in the $\boldsymbol{\theta}$ parameters, exponential retraction, $\lambda = 0.015$. The red dotted lines identify the different basins of attraction and correspond to the points for which the Newton step is not defined. The instability along the critical lines, which identifies the basins of attraction, is not represented.

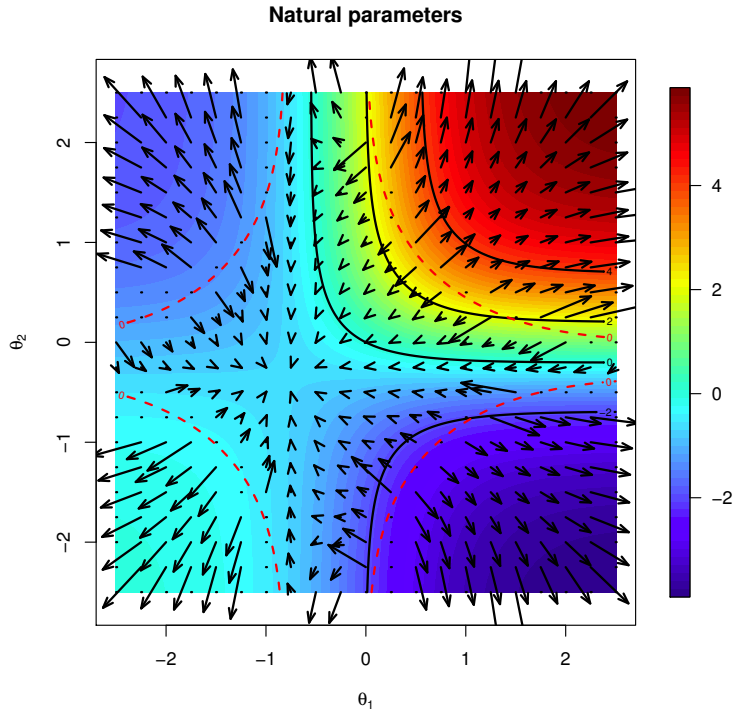


Figure 9. The Newton step in the $\boldsymbol{\theta}$ parameters, Euclidean geometry, $\lambda = 0.15$. The red dotted lines identify the different basins of attraction and correspond to the points for which the Newton step is not defined. The instability along the critical lines, which identifies the basins of attraction, is not represented.

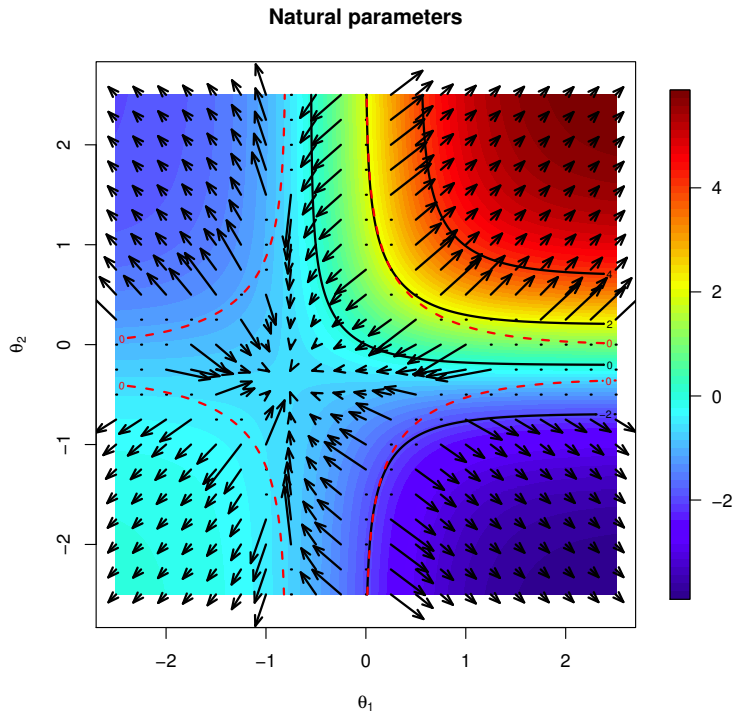


Figure 8 shows the Newton step in the $\boldsymbol{\theta}$ parameters based on the e-retraction of Equation (149), while Figure 9 represents the Newton step evaluated with respect to the Euclidean geometry. A comparison of the two vector fields shows that, differently from the $\boldsymbol{\eta}$ parameters, the number of basins of attraction is the same in the two geometries; however, the scale of the vectors is different. In particular, notice how on the plateau, for diverging $\boldsymbol{\theta}$, the Newton step in the Euclidean geometry vanishes, while in the Riemannian geometry, it gets larger. This behavior suggests better convergence properties for an optimization algorithm based on the Newton step evaluated using the proper Riemannian geometry. In the $\boldsymbol{\theta}$ parameters, the boundaries of the basins of attraction represented by the red dotted lines have been computed numerically and correspond to the values of $\boldsymbol{\theta}$ for which the update step is not defined.

Finally, notice that in both the $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$ parameters, the step is not always in the direction of descent for the function, a common behavior of the Newton method, which converges to the critical points.

5. Discussion and Conclusions

In this paper, we introduced second-order calculus over a statistical manifold, following the approach described in [4], which has been adapted to the special case of exponential statistical models [2,3]. By defining the Riemannian Hessian and using the notion of retraction, we developed the proper machinery necessary for the definition of the updating rule of the Newton method for the optimization of a function defined over an exponential family.

The examples discussed in the paper show that by taking into account the proper Riemannian geometry of a statistical exponential family, the vector fields associated with the Newton step in the different parametrizations change profoundly. Not only new basins of attraction associated with local and global minima appear, as for the expectation parameters, but also the magnitude of the Newton step is affected, as over the plateau in the natural parameters. Such differences are expected to have a strong impact on the performance of an optimization algorithm based on the Newton step, from both the point of view of achievable convergence and the speed of convergence to the optimum.

The Newton method is a popular second order optimization technique based on the computation of the Hessian of the function to be optimized and is well known for its super-linear convergence properties. However, the use of the Newton method poses a number of issues in practice.

First of all, as the examples in Figures 6 and 8 show, the Newton step does not always point in the direction of the natural gradient, and the algorithm may not converge to a (local) optimum of the function. Such behavior is not unexpected; indeed the Newton method tends to converge to critical points of the function to be optimized, which include local minima, local maxima and saddle points. In order to obtain a direction of ascent for the function to be optimized, the Hessian must be negative-definite, *i.e.*, its eigenvalues must be strictly negative, which is not guaranteed in the general case. Another important remark is related to the computational complexity associated with the evaluation of the Hessian, compared to the (natural) gradient. Indeed, to obtain the Newton step d, Christoffel matrices have to be evaluated, together with the third order covariances between sufficient statistics and the function, and the Hessian has to be inverted. Finally, notice that when the Hessian is close to being non-invertible, numerical problems may arise in the computation of the Newton step, and the algorithm may become unstable and diverge.

In the literature, different methods have been proposed to overcome these issues. Among them, we mention quasi-Newton methods, where the update vector is obtained using a modified Hessian, which has been made negative-definite, for instance, by adding a proper correction matrix.

This paper represents the first step in the design of an algorithm based on the Newton method for the optimization over a statistical model. The authors are working on the computational aspects related to the implementation of the method, and a new paper with experimental results is in progress.

Acknowledgments

Luigi Malagò was supported by the Xerox University Affairs Committee Award and by de Castro Statistics, Collegio Carlo Alberto, Moncalieri. Giovanni Pistone is supported by de Castro Statistics, Collegio Carlo Alberto, Moncalieri, and is a member of GNAMPA–INdAM, Roma.

Author Contributions

All authors contributed to the design of the research. The research was carried out by all authors. The study of the Hessian and of the Newton method in statistical manifolds was originally suggested by Luigi Malagò. The manuscript was written by Luigi Malagò and Giovanni Pistone. All authors have read and approved the final manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. 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; p. 283.
2. Amari, S.; Nagaoka, H. *Methods of Information Geometry*; American Mathematical Society: Providence, RI, USA, 2000; p. 206.
3. Pistone, G. Nonparametric Information Geometry. In *Geometric Science of Information*, Proceedings of the First International Conference, GSI 2013, Paris, France, 28–30 August 2013; Nielsen, F., Barbaresco, F., Eds.; Lecture Notes in Computer Science, Volume 8085; Springer: Berlin/Heidelberg, Germany, 2013; pp. 5–36.
4. Absil, P.A.; Mahony, R.; Sepulchre, R. *Optimization Algorithms on Matrix Manifolds*; Princeton University Press: Princeton, NJ, USA, 2008; pp. xvi+224.
5. Nocedal, J.; Wright, S.J. *Numerical Optimization*, 2nd ed.; Springer Series in Operations Research and Financial Engineering; Springer: New York, NY, USA, 2006; pp. xxii+664.
6. Do Carmo, M.P. *Riemannian geometry*; Mathematics: Theory & Applications; Birkhäuser Boston Inc.: Boston, MA, USA, 1992; pp. xiv+300.
7. Abraham, R.; Marsden, J.E.; Ratiu, T. *Manifolds, Tensor Analysis, and Applications*, 2nd ed.; Applied Mathematical Sciences, Volume 75; Springer: New York, NY, USA, 1988; pp. x+654.

8. Lang, S. *Differential and Riemannian Manifolds*, 3rd ed.; Graduate Texts in Mathematics; Springer: New York, NY, USA, 1995; pp. xiv+364.
9. Pistone, G. Algebraic varieties *vs.* differentiable manifolds in statistical models. In *Algebraic and Geometric Methods in Statistics*; Gibilisco, P., Riccomagno, E., Rogantin, M.P., Wynn, H.P., Eds.; Cambridge University Press: Cambridge, UK, 2010.
10. Malagò, L.; Matteucci, M.; Dal Seno, B. An information geometry perspective on estimation of distribution algorithms: Boundary analysis. In Proceedings of the 2008 GECCO Conference Companion On Genetic and Evolutionary Computation (GECCO '08); ACM: New York, NY, USA, 2008; pp. 2081–2088.
11. Malagò, L.; Matteucci, M.; Pistone, G. Stochastic Relaxation as a Unifying Approach in 0/1 Programming. 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.
12. Malagò, L.; Matteucci, M.; Pistone, G. Stochastic Natural Gradient Descent by Estimation of Empirical Covariances. In Proceedings of the IEEE Congress on Evolutionary Computation (CEC), New Orleans, LA, USA, 5–8 June 2011; pp. 949–956.
13. 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.
14. Malagò, L.; Matteucci, M.; Pistone, G. Natural gradient, fitness modelling and model selection: A unifying perspective. In Proceedings of the IEEE Congress on Evolutionary Computation (CEC), Cancun, Mexico, 20–23 June 2013; pp. 486–493.
15. Amari, S.I. Natural gradient works efficiently in learning. *Neural Comput.* **1998**, *10*, 251–276.
16. Shima, H. *The Geometry of Hessian Structures*; World Scientific Publishing Co. Pte. Ltd.: Hackensack, NJ, USA, 2007; pp. xiv+246.
17. Malagò, L. On the Geometry of Optimization Based on the Exponential Family Relaxation. Ph.D. Thesis, Politecnico di Milano, Milano, Italy, 2012.
18. Gallavotti, G. *Statistical Mechanics: A Short Treatise*; Texts and Monographs in Physics; Springer: Berlin, Germany, 1999; pp. xiv+339.
19. Naudts, J. Generalised exponential families and associated entropy functions. *Entropy* **2008**, *10*, 131–149.
20. Esteban, J.; Ray, D. On the Measurement of Polarization. *Econometrica* **1994**, *62*, 819–851.
21. Montalvo, J.; Reynal-Querol, M. Ethnic polarization, potential conflict, and civil wars. *Am. Econ. Rev.* **2005**, 796–816.
22. Stein, W. *et al*. Sage Mathematics Software (Version 6.0). The Sage Development Team, 2013. Available online: http://www.sagemath.org (accessed on 27 March 2014).