Second-order Optimization over the Multivariate Gaussian Distribution

Luigi Malagò[1] and Giovanni Pistone[2]

[1] Shinshu University & INRIA Saclay – Île-de-France
4-17-1 Wakasato, Nagano, 380-8553, Japan
malago@shinshu-u.ac.jp
[2] Collegio Carlo Alberto
Via Real Collegio, 30, 10024 Moncalieri, Italy
giovanni.pistone@carloalberto.org

Abstract. We discuss the optimization of the stochastic relaxation of a real-valued function, i.e., we introduce a new search space given by a statistical model and we optimize the expected value of the original function with respect to a distribution in the model. From the point of view of Information Geometry, statistical models are Riemannian manifolds of distributions endowed with the Fisher information metric, thus the stochastic relaxation can be seen as a continuous optimization problem defined over a differentiable manifold. In this paper we explore the second-order geometry of the exponential family, with applications to the multivariate Gaussian distributions, to generalize second-order optimization methods. Besides the Riemannian Hessian, we introduce the exponential and the mixture Hessians, which come from the dually flat structure of an exponential family. This allows us to obtain different Taylor formulæ according to the choice of the Hessian and of the geodesic used, and thus different approaches to the design of second-order methods, such as the Newton method.

In this paper we study the optimization of a real-valued function by means of its *Stochastic Relaxation* (SR), i.e., we search for the optimum of the function by optimizing the expected value of the function itself over a statistical model. This approach in optimization is very general and it has been developed in many different fields, from statistical physics and random-search methods, e.g., the Gibbs sampler in optimization [1], simulated annealing and the cross-entropy method [2]; to black-box optimization in evolutionary computation, e.g., Estimation of Distribution Algorithms [3] and evolutionary strategies [4–7]; going through well known techniques in polynomial optimization, such as the method of the moments [8].

By optimizing the SR of a function, we move from the original search space to a new search space given by a statistical model, i.e., a set of probability densities. Once we introduce a parameterization for the statistical model, the parameters of the model become the new variables of the relaxed problem. Notice that the notion of stochastic relaxation differs from the common notion of relaxation in

optimization, indeed the minimum of the relaxed problem does not provide a lower bound for the minimum of the original problem, since the expected value of a function is always greater of equal to the minimum of the function. The term stochastic relaxation has been borrowed from [1], and used in the context of optimization for the first time in [9]. In the original work Geman and Geman introduced the Gibbs sampler, which is described as a stochastic relaxation technique to sample a joint probability distribution, and that, combined with an annealing schedule, can be used as a maximization tool as well.

The choice of the statistical model in the SR plays a fundamental role, indeed there is a tradeoff between the complexity of the statistical model, expressed for instance by its dimension, and the difficulty of the relaxed problem, expressed for example in terms of the non-linearities which appear in the formula of the expected value of the function. For instance, consider the case of a finite search space. One could be tempted to define a relaxation over the whole probability simplex, so that the SR would become linear in the probabilities, and thus easy to optimize. However, the dimension of relaxed problem would equal that of the search space, and there would be no advantage in moving the search over a statistical model. Instead, it is more reasonable to choose a lower-dimensional statistical model in the search for the optimum. For finite search spaces this would correspond to constraining the search to a subset of the probability simplex. In this work we focus on the SR of a continous function with respect to a statistical model in the multivariate Gaussian distributions, however the theory we use applies in the general case of exponential families, with ether finite, discrete or continuous sample space.

In solving the SR, we are looking for an optimal density in a statistical model. This corresponds to the distribution that in the discrete case concentrate the probability mass over an optimal solutions of the original function, while in the continous case it is more appropriate to talk about concentration of the probability density in a neighborhood of the optimal solution in the original search space. The optimization of the SR can be performed according to different paradigms. In particular, a common approach in the family of first-order methods is given by gradient descent. However, it is well known in statistics that the geometry of a statistical model is not Euclidean, indeed it was first shown by Rao [10] that the set of positive distributions on a finite state space is a Riemannian manifold endowed with the Fisher information metric. Follows that the gradient of the stochastic relaxation should be evaluated with respect to the Fisher information metric, which leads us to the definition of natural gradient introduced by Amari [11]. Natural gradient has been proved to be efficient in different contexts besides the optimization of the SR [5–7], such as the training of neural networks [12] and, more recently, in deep learning [13]; policy gradients in reinforcement learning [14]; and last but not least variational inference techniques, e.g., [15].

In this paper we follow a geometric approach based on Information Geometry [16–19] to study the first and second-order geometry of the exponential family. The purpose of this analysis is to introduce the proper tools to define

second-order optimization methods over a statistical model, and in particular the notion of Riemannian Hessian which is required when the geometry of the space is not Euclidean. Notice that despite second-order methods over manifolds are widely used, as in the case of matrix manifolds [20], they appear to be new in the context of statistical manifolds. As we already mentioned, exponential families of distributions have an intrinsic Riemannian geometry, where the Fisher Information matrix plays the role of metric tensor. However, it was pointed out by Amari [16, 18] that besides the Riemannian geometry there are two other relevant dually-flat affine geometries of Hessian type for an exponential family: the exponential and the mixture one. The existence of (at least) three geometries provides three definitions of connections for an exponential family, three types of geodesics and, as we will see in the following, three types of Hessian, which are at the basis of the study of second-order optimization methods over a statistical manifold.

In the first part of the paper we review the first-order geometry of the exponential family. Next we move to second-order calculus, by introducing the notion of covariant derivative, and we provide formulæ for the Riemannian, exponential, and mixture Hessians over a statistical manifold. This analysis allows us to generalize the Newton algorithm to the optimization over a statistical manifold. We conclude the paper with some remarks about the case of multivariate Gaussian distributions. A preliminary version of this paper has been presented as a poster [21] at the NIPS 2014 Workshop on Optimization for Machine Learning (OPT2014).

1 Geometry of the Exponential Family

Given a real-valued function $f : \Omega \to \mathbb{R}$ to be minimized, and a statistical model $\mathcal{M}$, the Stochastic Relaxation (SR) F of f is defined as the expected value of the function itself with respect to p in $\mathcal{M}$, i.e., $F(p) = \mathbb{E}_p[f]$. Under some regularity conditions over the choice of $\mathcal{M}$, F is a continuous function independently from the nature of the sample space Ω, which can be either finite, discrete or continuous.

We are interested in developing second-order optimization methods for the SR of f based on the Gaussian distribution. However, the approach we present is more general and can be applied to any exponential family, thus in the following we will use the formalism of the exponential family and we will come back to the Gaussian distribution in the last part of the paper. In the first part of this section we review some general properties of the exponential family and we refer to the monograph [22]. Consider the exponential family $\mathcal{E}$:

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

with $\boldsymbol{\theta} \in B$, where B is an open convex set in $\mathbb{R}^d$. The real-valued functions $T_1, \ldots, T_k$, are the sufficient statistics of the exponential family, and $\psi(\boldsymbol{\theta})$ is

the log-partition function, i.e., $\psi(\boldsymbol{\theta}) = \log \int_{\boldsymbol{x}} \exp\left(\sum_{i=1}^{d} \theta_i T_i(\boldsymbol{x})\right) \mathrm{d}x$. The exponential family also admits a dual parameterization based on the expectation parameters $\boldsymbol{\eta}$ with $\boldsymbol{\eta} = \mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{T}] = \nabla\psi(\boldsymbol{\theta})$.

First and second-order methods to optimize a function F defined over an exponential family require the evaluation of the gradient and of the Hessian of F. The evaluation of such quantities depend on the geometry of the space, which is known to be non-Euclidean in the case of statistical models. To better understand the nature of $\mathcal{E}$, we refer to notions from Information Geometry [16, 18], which studies the geometry of statistical models and of the exponential family from the point of view of differential geometry [25]. Statistical models are considered as manifolds of distributions endowed with a Riemannian metric, given by the Fisher information metric.

In the following we denote with $\mathrm{T}_p\mathcal{E}$ the tangent space of $\mathcal{E}$ at p, i.e., the space of the tangent vectors to any curve $p(t)$ in $\mathcal{E}$ that goes through p. Rao showed that the tangent vector to $p(t)$ can be evaluated as $\frac{\mathrm{d}}{\mathrm{d}t} \log p(t)$, so that the tangent space $\mathrm{T}_p\mathcal{E}$ can be equivalently characterized as the space of all random variable centered in p, with the canonical basis given by the centered sufficient statistics $T_i - \mathbb{E}_p[T_i]$. Given two tangent vector U, V in $\mathrm{T}_p\mathcal{E}$, the tangent space is endowed with the inner product given by $g(U,V)(p) = \mathbb{E}_p[UV]$. In the basis of the sufficient statistics we have $\mathbb{E}_p[UV] = \sum_{ij} U_i \mathbb{E}_p[(T_i - \mathbb{E}_p[T_i])(T_j - \mathbb{E}_p[T_j])] V_j$, where ${}^eI(p) = E_p[(T_i - \mathbb{E}_p[T_i])(T_j - \mathbb{E}_p[T_j])]_{ij} = [\mathrm{Cov}(T_i, T_j)]_{ij}$ is the Fisher information matrix.

Given an exponential family $\mathcal{E}$, a function $F : \mathcal{E} \to \mathbb{R}$ and the metric g for $\mathcal{E}$, which in our case is the Fisher information metric, the Riemannian gradient $\operatorname{grad} F$ is the unique vector such that for any direction identified by the vector $X \in \mathrm{T}_p\mathcal{E}$, we have:

$$g(\operatorname{grad} F, X)(p) = \mathrm{D}_X F(p), \tag{2}$$

i.e., $\operatorname{grad} F$ is defined as the unique vector such that the inner product with respect to the metric between $\operatorname{grad} F$ and an arbitrarily direction X, evaluated at $p \in \mathcal{E}$, is the directional derivative $\mathrm{D}_X F(p)$ of F along X in p. The previous definition of Riemannian gradient is coordinate independent. If we consider a parameterization for the exponential family, and we choose a basis for the tangent space, we can write a formula for the components of the Riemannian gradient. In the exponential family, the *natural gradient* gives the components of the Riemannian gradient evaluated with respect to the Fisher information matrix ${}^eI(\boldsymbol{\theta})$, expressed in the basis of the centered sufficient statistics:

$$\widetilde{\nabla} F(\boldsymbol{\theta}) = {}^eI(\boldsymbol{\theta})^{-1} \nabla F(\boldsymbol{\theta}) \ . \tag{3}$$

Due to the properties of the exponential family ${}^eI(\boldsymbol{\theta})$ can be obtained as the Hessian of $\psi(\boldsymbol{\theta})$, i.e., the matrix of second-order partial derivatives $[\partial_i \partial_j \psi(\boldsymbol{\theta})]_{ij}$, and $\nabla F(\boldsymbol{\theta}) = (\partial_i F(\boldsymbol{\theta}))_i$ is the vector of first-order partial derivatives. Here ∂_i denotes the partial derivative with respect to θ_i, i.e., $\partial_i = \frac{\partial}{\partial \theta_i}$. We denote the natural gradient with $\widetilde{\nabla} F$ to distinguish it from ∇F, which corresponds to the components of the gradient evaluated with respect to the Euclidean metric.

In order to move to second-order calculus, we need a definition of Hessian of the function F over a manifold, which generalizes the Euclidean case. In the following we refer to [20], where second-order methods have been applied to the optimization over manifolds, cf. [23] for a similar approach. We study the second-order geometry of the exponential family in a general way, similar to what has been done in [24], where the focus was on applications to binary optimization. For basic notions of differential geometry, we refer to the standard book [25].

The first step in the geometric construction of the Riemannian Hessian, which is required to write a second-order Taylor approximation of the function in a neighborhood of a point, is the generalization to a manifold of the concept of directional derivative of a vector field. Indeed, differently from the Euclidean case, a definition based on the derivation of a vector field along a curve is not possible, since in each point of the curve tanget vectors belong to different tangent spaces, and without a correspondence between tangent spaces, no comparison is possible. The notion of affine connection provides a way to define such correspondence.

A *connection* ∇ over a manifold $\mathcal{M}$ is an operator $\nabla : \mathrm{T}\mathcal{M} \times \mathrm{T}\mathcal{M} \to \mathrm{T}\mathcal{M}$ which given two vector fields X and Y defined over $\mathcal{M}$ returns a new vector field $\nabla_X Y$ given by the directional derivative $\mathrm{D}_X Y$ of the Y in the direction X. The vector field $\nabla_X Y$ is called the *covariant derivative* of Y with respect to X for the given affine connection ∇. Notice that in general a manifold admits infinitely many connections. Each connection can be specified by d^2 vector fields which represent the covariant derivate $\nabla_{E_i} E_i$ where E_i and E_j are the coordinate vector fields. Then, a connection can be fully determined by d^3 symbols, called the Christoffel symbols Γ_{ij}^k, which represent the components of $\nabla_{E_i} E_j$ in the basis $E_1, \ldots, E_d$, i.e., $\nabla_{E_i} E_j = \sum_k \Gamma_{ij}^k E_k$.

Among all possible connections, there is a unique connection, called Riemannian or Levi-Civita connection, denoted by ${}^0\nabla$, which satisfies the properties of being symmetric and invariant with respect to the Riemannian metric. The Christoffel symbols ${}^0\Gamma_{ij}^k$, with $i, j, k = 1, \ldots, d$, for the Levi-Civita connection can be derived from the metric, using the formula ${}^0\Gamma_{ij}^k = \sum_l g^{kl}\, {}^0\Gamma_{ijl}$, with ${}^0\Gamma_{ijk} = \frac{1}{2}\left(\partial_i g_{jk} + \partial_j g_{ik} - \partial_k g_{ij}\right)$. The symbols $\Gamma_{ijk} = \sum_l g_{il}\Gamma_{jk}^l$ are called the Christoffel symbols of the first type, to distinguish them from $\Gamma_{jk}^k = \sum_l g^{kl}\Gamma_{ij}^l$ which are sometimes referred as Christoffel symbols of the second type. Here the g^{ij}'s denote the entries of the inverse Fisher information matrix, i.e., $[g^{ij}] = [g_{ij}]^{-1}$. Notice that when g can be expressed as the Hessian of a function for a given parameterization, then by symmetry we have ${}^0\Gamma_{ijk} = \frac{1}{2}\partial_i g_{jk}$.

As pointed out previously, besides the Riemannian connection, two other affine geometries, namely the exponential and the mixture geometry, play an important role for the exponential family. Amari [18] introduced the following family of α-connections, given by the Christoffel symbols:

$$ {}^\alpha\Gamma_{ijk}(\boldsymbol{\xi}) = \mathrm{E}_{\boldsymbol{\xi}}\left[\left(\partial_i\partial_j \log p(\boldsymbol{x};\boldsymbol{\xi}) + \frac{1-\alpha}{2}\partial_i \log p(\boldsymbol{x};\boldsymbol{\xi})\partial_j \log p(\boldsymbol{x};\boldsymbol{\xi})\right)\partial_k \log p(\boldsymbol{x};\boldsymbol{\xi})\right] $$

For $\alpha = 0$ we recover the Christoffel symbols of the Levi-Civita connection ${}^0\Gamma_{ijk}(\boldsymbol{\xi})$, while for $\alpha = \pm 1$ we obtain a characterization for the exponential

and mixture connection. In particular, for an exponential family parametrized by $\boldsymbol{\theta}$, it is easy to show that the Christoffel symbols of the exponential connection ${}^{e}\Gamma_{ijk}(\boldsymbol{\theta})$, for $\alpha = 1$, are identically equal to zero, i.e., the exponential family is e-flat. Similarly, once the exponential family is parametrized by $\boldsymbol{\eta}$, it turns out that the Christoffel symbols of the mixture connection ${}^{m}\Gamma_{ijk}(\boldsymbol{\eta})$, for $\alpha = -1$, are identically zero, i.e., the exponential family is m-flat as well. This is a consequence of the duality between the exponential and mixture geometry of the exponential family. It follows that we can introduce at least two alternative definitions of covariant derivative, based on the exponential and mixture geometries, which we call exponential and mixture covariant derivatives. Given the connection through its Christoffel symbols, the covariant derivative can be evaluated by the following formula:

$$\nabla_X Y = \sum_{ij} X^j \left(\sum_k Y^k \Gamma^i_{jk} + \partial_j Y^i \right) E_i \ . \tag{4}$$

The introduction of a connection over the manifold allows to define the notion of acceleration along a curve, which is based on the differentiation of tangent vectors along the curve itself. Thus, we can introduce a *geodesic* between two points as the curve with zero acceleration. Different definitions of covariant derivatives produce different geodesics between two points.

We can now introduce the *Riemannian Hessian* of a function defined over a manifold. In the following we interpret the Hessian as an operator which is applied to a vector field X and returns a vector field $\mathrm{D}_X \operatorname{grad} F$ given by the directional derivative of the Riemannian gradient along the direction identified by X. On a Riemannian manifold $\mathcal{M}$ endowed with the metric g, the Riemannian Hessian of F is the linear mapping ${}^{0}\operatorname{Hess} F(p) : \mathrm{T}_p\mathcal{M} \to \mathrm{T}_p\mathcal{M}$ such that ${}^{0}\operatorname{Hess} F(p)[X(p)] = \nabla_{X(p)} \operatorname{grad} F(p)$, where ${}^{0}\nabla$ is the Levi-Civita connection associated to g on $\mathcal{M}$. The coordinate representation of the Riemannian Hessian in the basis of the centered sufficient statistics [24] is given by:

$${}^{0}\operatorname{Hess} F(p)[X(p)] = {}^{e}I(p)^{-1} \left(\operatorname{Hess} F(p) - \frac{1}{2} \sum_k \partial_k \, {}^{e}I(p) (\widetilde{\nabla} F(p))_k \right) X(p), \tag{5}$$

where $\operatorname{Hess} F(p)$, with no arguments, denotes the Euclidean Hessian of F in p, i.e, the matrix of second-order partial derivatives. Notice that in the natural parameters, and more in general for any Hessian manifolds, since ${}^{e}I(\boldsymbol{\theta}) = \operatorname{Hess} \psi(\boldsymbol{\theta})$, then ${}^{0}\Gamma_{ijk}(\boldsymbol{\theta}) = \frac{1}{2}\partial_i\partial_j\partial_k\psi(\boldsymbol{\theta})$ becomes symmetric with respect to the three indices. Eq. (5) can be derived from Eq.(4), where the Christoffel symbols of the second type are given by the tensor contraction $\frac{1}{2}\, {}^{e}I(p)^{-1}\partial\, {}^{e}I(p)$. By choosing different Christoffel symbols associated to the exponential and mixture connections, we can obtain similar formulæ for ${}^{e}\operatorname{Hess} F(p)[X(p)]$ and ${}^{m}\operatorname{Hess} F(p)[X(p)]$.

2 Second-Order Optimization: The Newton Method

The Newton method is an optimization method which generates a sequence of distributions $\{p_t\}$, $t \geq 0$, in $\mathcal{M}$ which converges towards a stationary point

of F, i.e., a critical point of the vector field $p \mapsto \operatorname{grad} F(p)$. At the basis of this optimization technique there is a Taylor expansion $F(p)$ which provides a second-order approximation of the function over the manifold.

Let $t \mapsto p(t)$ be a Riemannian geodesic connecting $p = p(0)$ to $q = p(1)$ in $\mathcal{E}$, and $\mathrm{D}p(t)$ denote the tangent velocity vector $\frac{\mathrm{d}}{\mathrm{d}t} \log p(t)$, then the following Taylor formula holds:

$$F(q) \approx F(p) + \langle \operatorname{grad} F(p), \mathrm{D}p(0) \rangle_p + \frac{1}{2} \left\langle {}^0\operatorname{Hess} F(p)[\mathrm{D}p(0)], \mathrm{D}p(0) \right\rangle_p . \quad (6)$$

However, this is not the only possible second-order approximation of F. Two similar formula can be obtained by consider the exponential geodesic connecting p and q together with the mixture Hessian ${}^m\operatorname{Hess} F(p)[\mathrm{D}p(0)]$, and dually, using the mixture geodesic and the exponential Hessian. Proofs are omitted due to space limitation, however they are based on the duality between covariant derivatives in terms of preserving inner products with respect to the metric, and the fact that the acceleration along the corresponding geodesic is zero.

In order to determine the next point at each iteration, the Newton method is based on the idea of choosing the step in such a way that the Taylor expansion attains its minimum in the new point. This step can be found by ensuring that the derivative of the approximation is equal to zero in the new point. This requires to solve in $X(p) \in \mathrm{T}_p\mathcal{M}$ the following Newton equation:

$$\operatorname{Hess} F(p)[X(p)] = -\operatorname{grad} F(p) . \quad (7)$$

Once the previous equation has been solved, the last step consists in finding a point over the manifold along the geodesic starting from the current point with initial velocity given by the Newton step. This last step is required for any first or second-order optimization method over a manifold to find a correspondence between tangent vectors in a point and the neighborhood of the point itself in the manifold. The computation of a geodesic determined by the Newton step can be an expensive task in general, for instance when the geometry is not flat, however this step could be relaxed and approximated by the notion of *retraction*. The retraction over a manifold [20] is a mapping between the tangent space in a point and the manifold, with local rigidity conditions which preserves gradients at the point where it is evaluated.

3 Applications to the Gaussian Distribution

In this section we give some details about the application of the general theory of second-order calculus over an exponential family to the case of the Gaussian distribution. In the first part we recall some results about exponential families.

Due to the properties of the exponential family, the Fisher information matrix, the Euclidean gradient, and thus the natural gradient can be evaluated in terms of covariances, indeed we have $\widetilde{\nabla} F(\boldsymbol{\theta}) = \operatorname{Cov}_{\boldsymbol{\theta}}(\boldsymbol{T}, \boldsymbol{T}^{\mathrm{T}})^{-1} \operatorname{Cov}_{\boldsymbol{\theta}}(\boldsymbol{T}, f)$. As remarked above, since the exponential family parameterized by $\boldsymbol{\theta}$ is a Hessian

manifold, it follows that $\partial I(\boldsymbol{\theta}) = [\partial_i \partial_j \partial_k \psi(\boldsymbol{\theta})] = \mathrm{Cov}_{\boldsymbol{\theta}}(\boldsymbol{T}, \boldsymbol{T}, \boldsymbol{T}) = \mathbb{E}_{\boldsymbol{\theta}}[(\boldsymbol{T} - \mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{T}])(\boldsymbol{T} - \mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{T}])(\boldsymbol{T} - \mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{T}])]$, and $\mathrm{Hess}\, F(p) = [\partial_i \partial_j F(\boldsymbol{\theta})] = (\mathrm{Cov}_{\boldsymbol{\theta}}(\boldsymbol{T}, \boldsymbol{T}^{\mathrm{T}}, f)$. The Riemannian Hessian ${}^0\mathrm{Hess}\, F(\boldsymbol{\theta})[X(\boldsymbol{\theta})]$ can then be written in coordinates:

$$\mathrm{Cov}_{\boldsymbol{\theta}}(\boldsymbol{T}, \boldsymbol{T})^{-1} \left(\mathrm{Cov}_{\boldsymbol{\theta}}(\boldsymbol{T}, \boldsymbol{T}, f) - \frac{1}{2} \sum_k \mathrm{Cov}_{\boldsymbol{\theta}}(\boldsymbol{T}, \boldsymbol{T}, T_k)(\widetilde{\nabla} F(\boldsymbol{\theta}))_k \right) X(\boldsymbol{\theta}), \quad (8)$$

The implementation of an optimization algorithm for the SR based on the exponential family requires the evaluation of the covariances among the sufficient statistics and between the sufficient statistics and the function to be optimized. In the general case, to determine these quantities exactly can be computationally unfeasible, for this reason it is a common approach to replace the exact value with Monte Carlo estimations of the covariances based on the current sample.

We have now all the elements to write explicitly an updating formula in the natural parameters for the Newton method, where the sequence of distributions generated is identified by a corresponding sequence of parameter vectors $\{\boldsymbol{\theta}_t\}$, $t \geq 0$. The iterative formula for the Newton method can be written as:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - R_{\boldsymbol{\theta}_t}(\lambda \,\mathrm{Hess}\, F(\boldsymbol{\theta}_t)^{-1} \widetilde{\nabla} F(\boldsymbol{\theta}_t)) \,, \quad (9)$$

where the function $R_{\boldsymbol{\theta}}$ returns the coordinates of the image of the retraction, which is a mapping from the tangent space to the manifold that identifies a point along the direction specified by the vector given as an argument, which in our case is the Newton step. The parameter $\lambda > 0$ is used to control the step size and thus the convergence to a critical point of F.

We conclude this section with some comments about the application to the Gaussian case. We refer to [27] as a standard reference for the geometry of the Gaussian distribution, and to our paper [26] for a presentation of the different parameterizations of the Gaussian distribution in view of the SR. The Gaussian distribution is one of the special cases in the exponential family, where the computation of the transformation between natural parameters and expectation parameters can be done in an efficient way, through the inversion of the covariance matrix. Indeed, the natural parameters of the Gaussian distribution are a function of the inverse covariance matrix and of the mean vector, while the expectation parameters correspond to a function of covariance matrix and mean vector. This suggests an implementation of the Newton method based on the exponential Hessian in the natural parameters, for which the Christoffel symbols vanish, combined with a retraction based on the mixture geodesic, which can be evaluated efficiently in the expectation parameters.

4 Discussion and Future Work

In this paper we studied the second-order geometry of a Riemannian manifold, in the special case of exponential statistical models. We extended the analysis carried out in [24], by defining not only the Riemannian Hessian, but also the

exponential and the mixture Hessians. The three Hessians we introduced, which are associated to the three privileged geometries of an exponential family, allow to derive three different Taylor formulæ and thus three alternative generalizations of the updating rule of the Newton method over an exponential family.

The alternative approaches we proposed appear to be equally well motivated from a theoretical perspective, however they are not equivalent in practice, indeed they are based on the computation of different covariant derivates and different geodesics. Moreover we expect different computational costs in the evaluation of the Newton step according to the choice of the parameterization and of the connection, as well as the type of geodesic which needs to be computed. An experimental comparison is required in order to investigate the advantages and disadvantages of the different approaches we proposed, for instance in terms of computational complexity and speed of convergence.

We conclude the paper with a remark about second-order optimization techniques. Indeed, even if the Newton method and more in general second-order methods are very popular and well-known for their quadratic local convergence properties, in practice a number of issues has to be taken into account. The Newton step does not always points in the direction of the natural gradient, and close enough to a saddle point of the function the Newton step will tend to converge to it. In order to obtain a direction of descent for the function to be optimized, the Hessian must be negative-definite, i.e., its eigenvalues must be strictly negative. In order to overcome these issues, different methods have been proposed in the literature, such as quasi-Newton methods, where the update vector is obtained using a modified Hessian which is guaranteed to be negative definite. Finally, a number of other issues has to be taken into account in the design of an algorithm, such as the uncertainty in the estimation of the Hessian and of the gradient, when they are estimated from a sample, and the choice of other parameters of the algorithm, such as the step size.

5 Acknowledgements

Giovanni Pistone is supported by de Castro Statistics, Collegio Carlo Alberto, Moncalieri, and he is a member of GNAMPA-INDAM.

References

1. Geman, S., Geman, D.: Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on PAMI **6** (1984) 721 – 741
2. Rubinstein, R.: The cross-entropy method for combinatorial and continuous optimization. Methodology and Computing in Applied Probability **1** (1999) 127–190
3. Larrañaga, P., Lozano, J.A., eds.: Estimation of Distribution Algoritms. A New Tool for evolutionary Computation. Springer (2001)
4. Hansen, N., Ostermeier, A.: Completely derandomized self-adaptation in evolution strategies. Evolutionary Computation **9** (2001) 159–195
5. Wierstra, D., Schaul, T., Glasmachers, T., Sun, Y., Peters, J., Schmidhuber, J.: Natural evolution strategies. JMLR **15** (2014) 949–980

6. Malagò, L., Matteucci, M., Pistone, G.: Towards the geometry of estimation of distribution algorithms based on the exponential family. In: Proc. of FOGA '11, ACM (2011) 230–242
7. Ollivier, Y., Arnold, L., Auger, A., Hansen, N.: Information-geometric optimization algorithms: A unifying picture via invariance principles. arXiv:1106.3708 (2011)
8. Lasserre, J.B.: Global optimization with polynomials and the problem of moments. SIAM Journal on Optimization **11** (2001) 796–817
9. Malagò, L., Matteucci, M., Pistone, G.: Stochastic relaxation as a unifying approach in 0/1 programming. In: NIPS 2009 Workshop on Discrete Optimization in Machine Learning: Submodularity, Sparsity & Polyhedra (DISCML). (2009)
10. Radhakrishna Rao, C.: Information and the accuracy attainable in the estimation of statistical parameters. Bull. Calcutta Math. Soc. **37** (1945) 81–91
11. Amari, S.: Natural gradient works efficiently in learning. Neural Computation **10** (1998) 251–276
12. Amari, S.: Neural learning in structured parameter spaces - natural Riemannian gradient. In NIPS 1997, MIT Press (1997) 127–133
13. Pascanu, R., Bengio, Y.: Revisiting natural gradient for deep networks. In: Proceedings of ICLR 2014. (2014)
14. Kakade, S.: A natural policy gradient. In Dietterich, T.G., Becker, S., Ghahramani, Z., eds.: NIPS 2001, MIT Press (2001) 1531–1538
15. Kuusela, M., Raiko, T., Honkela, A., Karhunen, J.: A gradient-based algorithm competitive with variational bayesian em for mixture of gaussians. In: Neural Networks, 2009. IJCNN 2009 (2009) 1688–1695
16. Amari, S.: Differential-geometrical methods in statistics. Volume 28 of Lecture Notes in Statistics. Springer-Verlag, New York (1985)
17. Lauritzen, S.L. In: Statistical Manifolds. Volume 10 of Lecture Notes–Monograph Series. Institute of Mathematical Statistics, Hayward, CA (1987) 163–216
18. Amari, S., Nagaoka, H.: Methods of information geometry. American Mathematical Society, Providence, RI (2000)
19. Pistone, G.: Algebraic varieties vs differentiable manifolds in statistical models. In Gibilisco, P., Riccomagno, E., Rogantin, M.P., Wynn, H.P., eds.: Algebraic and Geometric Methods in Statistics. Cambridge University Press (2009) 339–363
20. Absil, P.A., Mahony, R., Sepulchre, R.: Optimization algorithms on matrix manifolds. Princeton University Press, Princeton, NJ (2008)
21. Malagò, L., Pistone, G.: Stochastic relaxation over the exponential family: Second-order geometry. In: NIPS 2014 Workshop on Optimization for Machine Learning (OPT2014), Montreal, Canada, 12 December 2014. (2014)
22. Brown, L.D.: Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory. Volume 9 of Lecture Notes - Monograph Series. Institute of Mathematical Statistics (1986)
23. Manton, J.H.: A framework for generalising the Newton method and other iterative methods from euclidean space to manifolds. arXiv:1106.3708 (2012v1; 2014v2)
24. Malagò, L., Pistone, G.: Combinatorial optimization with information geometry: The Newton method. Entropy **16** (2014) 4260–4289
25. do Carmo, M.P.: Riemannian geometry. Mathematics: Theory & Applications. Birkhäuser Boston Inc., Boston, MA (1992)
26. Malagò, L., Pistone, G.: Information geometry of the gaussian distribution in view of stochastic optimization. In: Proc. of FOGA '15. (2015) 150–162
27. Skovgaard, L.T.: A Riemannian Geometry of the Multivariate Normal Model. Scandinavian Journal of Statistics **11** (1984) 211–223