Information Geometry of the Gaussian Distribution in View of Stochastic Optimization

Luigi Malagò
malago@shinshu-u.ac.jp
Shinshu University & INRIA Saclay – Île-de-France
4-17-1 Wakasato, Nagano, 380-8553, Japan

Giovanni Pistone
giovanni.pistone@carloalberto.org
Collegio Carlo Alberto
Via Real Collegio, 30, 10024 Moncalieri, Italy

ABSTRACT

We study the optimization of a continuous function by its stochastic relaxation, i.e., the optimization of the expected value of the function itself with respect to a density in a statistical model. We focus on gradient descent techniques applied to models from the exponential family and in particular on the multivariate Gaussian distribution. From the theory of the exponential family, we reparametrize the Gaussian distribution using natural and expectation parameters, and we derive formulas for natural gradients in both parameterizations. We discuss some advantages of the natural parameterization for the identification of sub-models in the Gaussian distribution based on conditional independence assumptions among variables. Gaussian distributions are widely used in stochastic optimization and in particular in model-based Evolutionary Computation, as in Estimation of Distribution Algorithms and Evolutionary Strategies. By studying natural gradient flows over Gaussian distributions our analysis and results directly apply to the study of CMA-ES and NES algorithms.

Categories and Subject Descriptors

G.1.6 [**Mathematics of Computing**]: Optimization — *Stochastic programming*; G.3 [**Mathematics of Computing**]: Probabilistic algorithms (including Monte Carlo)

General Terms

Theory, Algorithms

Keywords

Stochastic Relaxation; Information Geometry; Exponential Family; Multivariate Gaussian Distribution; Stochastic Natural Gradient Descent

1. INTRODUCTION

FOGA'15, January 17–20, 2015, Aberystwyth, UK.

ACM 978-1-4503-3434-1/15/01 ...$15.00.
http://dx.doi.org/10.1145/2725494.2725510 .

In this paper we study the optimization of a continuous function by means of its Stochastic Relaxation (SR) [19], i.e., we search for the optimum of the function by optimizing the functional given by the expected value of the function itself over a statistical model. This approach is quite general and appears in many different communities, from evolutionary computation to statistical physics, going through certain techniques in mathematical programming.

By optimizing the stochastic relaxation 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. The search for an optimal density in the statistical model, i.e., a density that maximizes the probability of sampling an optimal solution for the original function can be performed in different ways, similarly, different families of statistical models can be employed in the search for the optimum. In the literature of Evolutionary Computation, we restrict our attention to model-based optimization, i.e., those algorithms where the search for the optimum is guided by a statistical model. In this context, examples of Stochastic Relaxations of continuous optimization are given by Estimation of Distribution Algorithms (EDAs) [16], see for instance EGNA and EMNA, and Evolutionary Strategies, such as CMA-ES [13], NES [32, 31] and GIGO [9].

There is a clear connection of Stochastic Relaxation with Entropy based methods in optimization. In fact, on a finite state space it is easily shown that, starting form any probability function the positive gradient flow of the entropy goes to the uniform distribution, while the negative gradient flow goes to the uniform distribution on the values that maximize the probability function itself, see e.g. [29].

We are interested in the study of the stochastic relaxation of a continuous real-valued function defined over $\mathbb{R}^n$, when the statistical model employed in the relaxation is chosen from the exponential family of probability distributions [12]. In particular we focus on multivariate Gaussian distributions, which belong to the exponential family, and are one of the most common and widely employed models in model-based continuous optimization, and, more in general, in statistics. Among the different approaches to the optimization of the expected value of the function over the statistical model, we focus on gradient descent techniques, such as the CMA-ES and NES families of algorithms.

The methods used here are actually first order optimization methods on Riemannian manifolds, see [1] for the spe-

cific case of matrix manifolds, with two major differences. First, our extrema are obtained at the border of the manifold as parameterized by the exponential family. This fact presents the issue of an optimization problem of a manifold with border which, in turn, is defined through an extension of the manifold outside the border with a suitable parameterization. Second, the actual IG structure is reacher that the simple Riemannian structure because of the presence of what Amari calls dually flat connections and some geometers call Hessian manifold. Second order optimization methods are available for the Stochastic Relaxation problem. The finite state space case has been considered in our paper [22]. Second order methods are not discussed in the present paper.

The geometry of the multivariate Gaussian distribution is a well established subject in mathematical statistics, see for instance [30]. In this paper, we follow a geometric approach based on Information Geometry [4, 7] to the study of the multivariate Gaussian distribution and more generally of the exponential family from the point of view of the stochastic relaxation of a continuous function, cf. [27]. In this work, we extend to the case of continuous sample space some of the results presented in [20, 21] for the finite sample space, using a information geometric perspective on the stochastic relaxation based on gradient descent over an exponential family. A similar framework, based on stochastic relaxation has been proposed under the name of Information Geometric Optimization (IGO) [26], where the authors consider the more general case of the relaxation of rank-preserving trasformations of the function to be optimized.

Exponential families of distributions have an intrinsic Riemannian geometry, where the Fisher information matrix plays the role of metric tensor. Moreover, the exponential family exhibits a dually flat structure, and besides the parameterization given by the natural parameters of the exponential family, there exists a dually coupled parameterization for densities in the exponential family given by the expectation parameters. Since the geometry of the exponential family is in most cases not Euclidean, gradients need to be evaluated with respect to the relevant metric tensor, which leads to the definition of the *natural gradient* [5], to distinguish it from the vector of partial derivatives, which are called as *vanilla gradient*. Such a distinction makes no sense in the Euclidean space, where the metric tensor is the identify matrix, and the gradient with respect to the metric tensor is the vector of partial derivatives.

In the following sections, besides the mean and covariance parameterization, we discuss the natural and expectation parameterizations for the multivariate Gaussian distribution based on the exponential family, and we provide formulae for transformations from one parameterization to the other. We further derive the Fisher information matrices and the vanilla and natural gradients in the different parameterizations. We prove convergence results for the Gibbs distribution and study how the landscape of the expected value of the function changes according to the choice of the Gaussian family. We introduce some toy examples which make it possible to visualize the flows associated to the gradient over the statistical model used in the relaxation.

The use of the natural parameters of the exponential family makes it possible to identify sub-families in the Gaussian distribution by setting some of the natural parameters to zero. Indeed, since the natural parameters are proportional to the elements of inverse of the covariance matrices, by setting to zero one of these parameters we have a corresponding zero in the precision matrix, i.e., we are imposing a conditional independence constraint over the variables in the Gaussian distribution. From this perspective, we can rely on an extensive literature of graphical models [17] for model selection and estimation techniques.

2. THE EXPONENTIAL FAMILY

We consider the statistical model $\mathcal{E}$ given on the measured sample space $(\mathcal{X}, \mathcal{F}, \mu)$ by the densities of the form

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

with $\boldsymbol{\theta} \in \vartheta$, where ϑ is an open convex set in $\mathbb{R}^k$. The real random variables $\{T_i\}$ are the sufficient statistics of the exponential family, and $\psi(\boldsymbol{\theta})$ is a normalizing term, which is equal to the log of the partition function

$$Z \colon \theta \mapsto \int \exp\left(\sum_{i=1}^{k} \theta_i T_i(\boldsymbol{x})\right) \, \mu(d\boldsymbol{x}) \ .$$

The entropy is

$$-\int \log p(\boldsymbol{x}; \boldsymbol{\theta}) \ p_{\boldsymbol{\theta}}(\boldsymbol{x}; \boldsymbol{\theta}) \ \mu(d\boldsymbol{x}) = \psi(\boldsymbol{\theta}) - \sum_{i=1}^{k} \theta_i \mathbb{E}_{\boldsymbol{\theta}}[T_i] \ .$$

The partition function Z is a convex function whose proper domain is a convex set. We assume that the ϑ domain is either the proper domain of the partition function, if it is open, or the interior of the proper domain. Standard reference on exponential families is [12], where an exponential family such that the proper domain of Z is open is said to be steep. Moreover we assume that the sufficient statistics are affinely independent, that is, if a linear combination is constant, then the linear combination is actually zero. Such an exponential family is called minimal in standard references.

The exponential family admits a dual parameterization to the natural parameters, given by the expectation parameters $\boldsymbol{\eta} = \mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{T}]$, see [12, Ch. 3]. The $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ parameter vectors of an exponential family are dually coupled in the sense of the Legendre transform [8], indeed, let $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ such that $p_{\boldsymbol{\theta}}(\boldsymbol{x}; \boldsymbol{\theta}) = p_{\boldsymbol{\eta}}(\boldsymbol{x}; \boldsymbol{\eta})$, then

$$\psi(\boldsymbol{\theta}) + \varphi(\boldsymbol{\eta}) - \langle \boldsymbol{\theta}, \boldsymbol{\eta} \rangle = 0 \ , \tag{2}$$

where $\varphi(\boldsymbol{\eta}) = \mathbb{E}_{\boldsymbol{\eta}}[\log p(\boldsymbol{x}; \boldsymbol{\eta})]$ is the negative entropy of the density $p_{\boldsymbol{\eta}}(\boldsymbol{x}; \boldsymbol{\eta})$, and $\langle \boldsymbol{\theta}, \boldsymbol{\eta} \rangle = \sum_{i=1}^{k} \theta_i \eta_i$ denotes the inner product between the two parameter vectors. See [28, Part III] on convex duality.

From the Legendre duality it follows that the variable trasformations between one parameterization and the other are given by

$$\begin{aligned} \boldsymbol{\eta} &= \nabla_{\boldsymbol{\theta}} \psi(\boldsymbol{\theta}) = (\nabla_{\boldsymbol{\eta}} \varphi)^{-1}(\boldsymbol{\theta}) \ , \\ \boldsymbol{\theta} &= \nabla_{\boldsymbol{\eta}} \varphi(\boldsymbol{\eta}) = (\nabla_{\boldsymbol{\theta}} \psi)^{-1}(\boldsymbol{\eta}) \ . \end{aligned}$$

We introduced two dual parameterizations for the same exponential family $\mathcal{E}$, the $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ parameters, so that any $p \in \mathcal{M}$ can be parametrized either with $p_{\boldsymbol{\theta}}(\boldsymbol{x}; \boldsymbol{\theta})$ or with $p_{\boldsymbol{\eta}}(\boldsymbol{x}; \boldsymbol{\eta})$. In the following, to simplify notation, we drop the index of p which denotes the parameterization used when the parameter appears as an argument, however notice that $p_{\boldsymbol{\theta}}$ and $p_{\boldsymbol{\eta}}$ are different functions of their parameterizations.

The Fisher information matrices in the two different parameterizations can be evaluated by taking second derivatives of $\psi(\boldsymbol{\theta})$ and $\varphi(\boldsymbol{\eta})$

$$I_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \operatorname{Hess}\psi(\boldsymbol{\theta}) \ , \tag{3}$$

$$I_{\boldsymbol{\eta}}(\boldsymbol{\eta}) = \operatorname{Hess}\varphi(\boldsymbol{\eta}) \ . \tag{4}$$

The following result shows the relationship between the Fisher information matrices expressed in the $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ parameterizations for the same distribution. The result appears in [7], see also Theorem 2.2.5 in [15].

THEOREM 1. *Consider a probability distribution in the exponential family $\mathcal{E}$, we have*

$$I_{\boldsymbol{\eta}}(\boldsymbol{\eta}) = (I_{\boldsymbol{\theta}} \circ \nabla\varphi)(\boldsymbol{\eta})^{-1}$$
$$I_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = (I_{\boldsymbol{\eta}} \circ \nabla\psi)(\boldsymbol{\theta})^{-1} \ .$$

Moreover, we have

$$I_{\boldsymbol{\eta}}(\boldsymbol{\eta})^{-1} = \operatorname{Cov}_{\boldsymbol{\eta}}(\boldsymbol{T},\boldsymbol{T}) = \mathbb{E}_{\boldsymbol{\eta}}[(\boldsymbol{T}-\boldsymbol{\eta})(\boldsymbol{T}-\boldsymbol{\eta})^{\mathrm{T}}] \ .$$

2.1 The Gibbs Distribution

For each objective function $f\colon \mathcal{X} \to \mathbb{R}$, we introduce its Gibbs distribution, the one dimensional exponential family whose sufficient statistics is the function f itself. In the discrete case, it is a classical result in Statistical Physics and it is easy to show that the Gibbs distribution for $\theta \to \infty$ weakly converges to the uniform distribution over the minima of f. We refer for example to [20] for a discussion in the context of the stochastic relaxation for discrete domains. In this subsection we consider the extension of the result to the continuous case.

In the following we look for the minima of the objective function f. hence, we assume assume f to be bounded below and non constant. Given a probability measure μ, the Gibbs exponential family of f is the model $\theta \mapsto \mathrm{e}^{\theta f - \psi(\theta)} \cdot \mu$. As f is bounded below and μ is finite, the log-partition function $\psi(\theta) = \log\left(\int \mathrm{e}^{\theta f} \ d\mu\right)$ is finite on an interval containing the negative real line. We take in particular the interior of such an interval, hence ψ is defined on an open interval $J =]-\infty, \overline{\theta}[$, where $\overline{\theta} \in [0, +\infty]$. The function $\psi\colon J \to \mathbb{R}$ is strictly convex and analytic, see [12].

Define $\underline{f} = \operatorname{ess\,inf}_\mu f = \inf\{A : \mu\{f \le A\} > 0\}$, $\overline{f} = \operatorname{ess\,sup}_\mu f = \sup\{B : \mu\{B \le f\} > 0\}$, and define the *Gibbs relaxation* or *stochastic relaxation* of f to be the function

$$F(\theta) = \mathbb{E}_\theta[f] = \int f \mathrm{e}^{\theta f - \psi(\theta)} \ d\mu = \frac{d}{d\theta}\psi(\theta) \ ,$$

so that $\underline{f} \le F(\theta) \le \overline{f}$.

PROPOSITION 2.

1. *The function F is increasing on its domain J.*
2. *The range of the function F is the interval $]\underline{f}, \sup F[$,*
3. *in particular, $\lim_{\theta\to-\infty} \mathbb{E}_\theta[f] = \underline{f}$.*

PROOF. 1. As ψ is strictly convex on J, then its derivative F is strictly increasing on J.

2. As F is strictly increasing and continuous on the open interval J, it follows that range $F(J)$ is the open interval $]\inf_{\theta\in J} F, \sup_{\theta\in J} F(\theta)[$, which is contained in $]\underline{f}, \overline{f}[$. We show that its left end is actually $\underline{f}$. Equivalently, we show that for each $\epsilon > 0$ and each $\eta > \underline{f} + \epsilon$, $\eta \in F(J)$, there exists θ such that $F(\theta) = \eta$. The argument is a variation of the argument to prove the existence of maximum likelihood estimators because we show the existence of a solution to the equation $F(\theta) - \eta = \frac{d}{d\theta}(\eta\theta - \psi(\theta) = 0$ for each $\eta > \underline{f}$. Let $A = \underline{f} + \epsilon$ and take any $\theta < 0$ to show that

$$1 = \int \mathrm{e}^{\theta f(\boldsymbol{x}) - \psi(\theta)} \ \mu(d\boldsymbol{x}) \ge \int_{\{f\le A\}} \mathrm{e}^{\theta f(\boldsymbol{x}) - \psi(\theta)} \ \mu(d\boldsymbol{x}) \ge \mu\{f \le A\}\, \mathrm{e}^{\theta A - \psi(\theta)}.$$

Taking the logarithm of both sides of the inequality, we obtain, for each η

$$0 \ge \log\mu\{f \le A\} + \theta A - \psi(\theta) = \log\mu\{f \le A\} + \theta(A-\eta) + \theta\eta - \psi(\theta) \ ,$$

and, reordering the terms of the inequality, that

$$\theta\eta - \psi(\theta) \le -\log\mu\{f \le A\} + \theta(\eta - A) \ .$$

If $\eta > A$, the previous inequality implies

$$\lim_{\theta\to-\infty} \theta\eta - \psi(\theta) = -\infty \ ,$$

that is, the strictly concave differentiable function $\theta \mapsto \theta\eta - \psi(\theta)$ goes to $-\infty$ as $\theta \to -\infty$. Let us study what happens at the right of of the interval $F(J)$. Let $F(\theta_1) = \eta_1 > \eta$. If $\theta \to \theta_1$ increasingly, then $(\psi(\theta_1) - \psi(\theta))/(\theta_1 - \theta)$ goes to $\psi'(\theta_1) = F(\theta_1) = \eta_1$ increasingly. Hence, there exists $\theta_2 < \theta_1$ such that $(\psi(\theta_1) - \psi(\theta_2))/(\theta_1 - \theta_2) > \eta$, that is $\eta\theta_2 - \psi(\theta_2) > \eta\theta_1 - \psi(\theta_1)$. It follows that the strictly concave and differentiable function $\theta \mapsto \eta\theta - \psi(\theta)$ has a maximum $\hat{\theta} \in]-\infty, \eta_1[$, where its derivative is zero, giving $\eta = \psi'(\hat{\theta}) = F(\hat{\theta})$.

3. As F is strictly increasing, $\lim_{\theta\to-\infty} F(\theta) = \min(F(J)) = \underline{f}$.

□

REMARK 3. *Our assumption on the objective function is asymmetric as we have assumed f bounded below while we have no assumption on the big values of f. If f is bounded above, we have the symmetric result by exchanging f with $-f$ and θ with $-\theta$. If f is unbounded, then a Gibbs distribution need not to exist because of the integrability requirements of $\mathrm{e}^{\theta f}$. If the Laplace transform of f in μ is defined, then the Gibbs distribution exists, but only Item 1 of Prop. 2 applies. The result above does say only that the left limit of the relaxed function is greater or equal to $\underline{f}$. In such cases the technical condition on f called* steepness *would be relevant, see [12, Ch. 3]. We do not discuss this issue here because the problem of finding the extrema of an unbounded function is not clearly defined.*

REMARK 4. *Statistical models other that the Gibbs model can produce a relaxed objective function usable for finding the minimum. Let $p_\theta \cdot \mu$ be any one-dimensional model. Then $\lim_{\theta\to-\infty} \int f p_\theta \ d\mu = \underline{f}$ if, and only if,*

$$\lim_{\theta\to-\infty} \int f(\mathrm{e}^{\theta f - \psi(\theta)} - p_\theta) \ d\mu = 0 \ .$$

In particular, a simple sufficient condition in case of a bounded f is

$$\lim_{\theta\to-\infty}\int\left|\mathrm{e}^{\theta f-\psi(\theta)}-p_\theta\right|\,d\mu=0.$$

In practice, it is unlikely f to be known and we look forward learning some proper approximation od the Gibbs relaxation.

The convergence of the expected value along a statistical model to the minimum of the values of the objective function f, which was obtained above, is a result weaker than what we are actually looking for, that is the convergence of the statistical model to a limit probability μ_∞ supported by the set of minima, $\{\boldsymbol{x}\in\mathcal{X}:f(\boldsymbol{x})=\underline{f}\}$, that is a probability such that $\int f(\boldsymbol{x})\ \mu(d\boldsymbol{x})=\underline{f}$. For example, if $f(x)=1/x$, $x>0$, and μ_n is the uniform distribution on $[n,n+1]$, $n=1,2,\dots$, then $\int f d\ \mu_n=\log\left((n+1)/n\right)$ goes to $0=\inf f$, but there is no minimum of f nor limit of the sequence $(\mu_n)_n$.

The following proposition says something about this issue. We need topological assumptions. Namely, we assume the sample space $\mathcal{X}$ to be a metric space and the objective function to be bounded below, lower semicontinuous, and with compact level sets $\{\boldsymbol{x}\in\mathcal{X}:f(\boldsymbol{x})\leq\underline{f}+A\}$. In such a case we have the following result about weak convergence of probability measures, see [11].

PROPOSITION 5. *The family of measures $(\mathrm{e}^{\theta f-\psi(\theta)}\cdot\mu)_{\theta\in J}$ is relatively compact as $\theta\to-\infty$ and the limits are supported by the closed set $\{\boldsymbol{x}:f(\boldsymbol{x})=\underline{f}\}$. In particular, if the minimum is unique, then the sequence converges weakly to the delta mass at the minimum.*

PROOF. The set $\{f\leq\underline{f}+A\}$ is compact and we have, by Markov inequality, that

$$\limsup_{\theta\to-\infty}\int_{\{f>\underline{f}+A\}}\mathrm{e}^{\theta f-\psi(\theta)}\,d\mu\leq A^{-1}\lim_{\theta\to-\infty}\int(f-\underline{f})\mathrm{e}^{\theta f-\psi(\theta)}\,d\mu=0,$$

which is the tightness condition for relative compactness. If ν is a limit point along the sequence $(\theta_n)_{n\in\mathbb{N}}$, $\lim_{n\to\infty}\theta_n=-\infty$, then for all $a>0$ the set $\{f>\underline{f}+a\}$ is open and, by the Portmanteaux Theorem [11, Th. 2.1.(iv)],

$$\nu\{f>\underline{f}+a\}\leq\liminf_{n\to\infty}\int_{\{f>\underline{f}+a\}}\mathrm{e}^{\theta_n f-\psi(\theta_n)}\,d\mu=0.$$

As each of the set $\{f>\underline{f}+a\}$ has measure zero, their (uncountable) union ha s measure zero,

$$\nu\{f>\underline{f}\}=\nu\left(\cup_{a>0}\{f>\underline{f}+a\}\right)=0.$$

Finally, if $\{f=\underline{f}\}$ has a unique point, then each limit ν has to have a point support, hence has to be the Dirac delta. □

As a consequence of the previous result we can extend Th. 12 in [20] to the continuous case. Let $V=\mathrm{Span}\{T_1,\dots,T_k\}$ be the vector space generated by the affinely independent random variables T_j, $j=1,\dots,k$ and let $\mathcal{E}$ be the exponential family on the sample space $(\mathcal{X},\mu)$ with that sufficient statistics. For $f\in V$, $f=\sum_{j=1}^k\alpha_jT_j$, and $q=p_{\boldsymbol{\theta}}\in\mathcal{E}$, consider the Gibbs family

$$p(x;t)=\frac{e^{-tf}q}{\mathbb{E}_q[e^{-tf}]},\quad t\colon t\boldsymbol{\alpha}+\boldsymbol{\theta}\in\vartheta\ . \tag{5}$$

Note that this family is actually a subfamily of $\mathcal{E}$ that moves from q in the direction $f-\mathbb{E}_q[f]$. We further assume f to be bounded below to obtain the following.

THEOREM 6. *The gradient of the function $F\colon\mathcal{E}\ni q\mapsto\mathbb{E}_q[f]$ is $\nabla F(q)=f-\mathbb{E}_q[f]$. The trajectory of the negative gradient flow through q is the exponential family in Eq. (5). The negative gradient flow is a minimizing evolution.*

PROOF. The only thing we need to verify is the fact that the velocity of the curve in Eq. (5) at $t=0$ is precisely $f-\mathbb{E}_q[f]$. The evolution is minimizing, according to the assumptions on f, because of Prop. 2 and 5. □

The previous theorem can be applied to the case when the exponential family is a Gaussian distribution. In particular it makes it possible to prove global convergence of natural gradient flows of quadratic non-constant and lower bounded functions in the Gaussian distribution. For any given initial condition, the natural gradient flow converges to the δ distribution which concentrates all the probability mass over the global optimum of f. Akimoto et. al proved in [2] an equivalent result for isotropic Gaussian distributions in the more general framework of IGO which takes into account a rank preserving trasformation of the function to be optimized based on quantiles. In this context, see also the work of Beyer [10].

3. MULTIVARIATE GAUSSIAN DISTRIBUTIONS

In this section we discuss the multivariate Gaussian distribution and discuss its geometry in the more general context of the exponential family. The geometry of the Gaussian distribution has been widely studied in the literature, see for instance [30] as an early and detailed reference on the subject.

Since the multivariate Gaussian distribution belongs to the exponential family, besides mean and covariance matrice, we can introduce two alternative parameterizations, based on natural and expectation parameters. Notice that all these parameterizations are one-to-one.

Vectors are intended as column vectors and are represented using the bold notation. Let $\boldsymbol{\xi}$ be a vector of parameters, in order to obtain compact notation, we denote the partial derivative $\partial/\partial\xi_i$ with ∂_i. When a parameter ξ_{ij} of $\boldsymbol{\xi}$ is identified by two indices, we denote the partial derivative with ∂_{ij}.

3.1 Mean and Covariance Parameters

Consider a vector $\boldsymbol{x}=(\boldsymbol{x}_1,\dots,\boldsymbol{x}_n)^{\mathrm{T}}\in\mathbb{R}^n=\Omega$. Let $\boldsymbol{\mu}\in\mathbb{R}^n$ be a mean vector and $\Sigma=[\sigma_{ij}]$ a $n\times n$ symmetric positive-definite covariance matrix, the multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu},\Sigma)$ can be written as

$$p(\boldsymbol{x};\boldsymbol{\mu},\Sigma)=(2\pi)^{-n/2}|\Sigma|^{-1/2}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right). \tag{6}$$

We denote with $\Sigma^{-1}=[\sigma^{ij}]$ the inverse covariance matrix, also known as precision matrix or concentration matrix. We use upper indices in σ^{ij} to remark that they are the elements of the inverse matrix $\Sigma^{-1}=[\sigma_{ij}]^{-1}$. The precision matrix captures conditional independence between variables in $\boldsymbol{X}=(\boldsymbol{X}_1,\dots,\boldsymbol{X}_n)^{\mathrm{T}}$. Zero partial correlations, i.e., $\sigma^{ij}=0$, correspond to conditional independence assumption of X_i

and X_j given all other variables, denoted by $X_i \perp\!\!\!\perp X_j | \boldsymbol{X}_{\setminus i \setminus j}$. See [17] as a comprehensive reference on graphical models.

3.2 Natural Parameters

It is a well known result that the multivariante Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ is an exponential family for the sufficient statistics X_i, X_i^2, $2X_iX_j$, $i \leq j$. We denote with $\omega : (\boldsymbol{\mu}, \Sigma) \mapsto \boldsymbol{\theta}$ the parameter transformation from the couple mean vector and covariance matrix to the natural parameters. By comparing Eq. (1) and (6), it is easy to verify that

$$\boldsymbol{T} = \begin{pmatrix} (X_i) \\ (X_i^2) \\ (2X_iX_j)_{i<j} \end{pmatrix} = \begin{pmatrix} (T_i) \\ (T_{ii}) \\ (T_{ij})_{i<j} \end{pmatrix} , \tag{7}$$

$$\boldsymbol{\theta} = \omega(\boldsymbol{\mu}, \Sigma) = \begin{pmatrix} \Sigma^{-1}\boldsymbol{\mu} \\ (-\frac{1}{2}\sigma^{ii}) \\ (-\frac{1}{2}\sigma^{ij})_{i<j} \end{pmatrix} = \begin{pmatrix} (\theta_i) \\ (\theta_{ii}) \\ (\theta_{ij})_{i<j} \end{pmatrix} ,$$

where $k = n + n + n(n-1)/2 = n(n+3)/2$, which leads to

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}; \boldsymbol{\theta}) = \exp\left(\sum_i \theta_i x_i + \sum_i \theta_i x_i^2 + \sum_{i<j} 2\theta_{ij} x_i x_j - \psi(\boldsymbol{\theta}) \right) .$$

To simplify the formulae for variable transformations and in particular the derivations in the next sections, we define

$$\theta = (\theta_i) = \Sigma^{-1}\boldsymbol{\mu} , \tag{8}$$

$$\Theta = \sum_i \theta_{ii} \boldsymbol{e}_i \boldsymbol{e}_i^{\mathrm{T}} + \sum_{i<j} \theta_{ij} (\boldsymbol{e}_i \boldsymbol{e}_j^{\mathrm{T}} + \boldsymbol{e}_j \boldsymbol{e}_i^{\mathrm{T}}) = -\frac{1}{2}\Sigma^{-1} , \tag{9}$$

and represent $\boldsymbol{\theta}$ as

$$\boldsymbol{\theta} = (\theta; \Theta) ,$$

so that

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}; \boldsymbol{\theta}) = \exp\left(\theta^{\mathrm{T}} \boldsymbol{x} + \boldsymbol{x}^{\mathrm{T}} \Theta \boldsymbol{x} - \psi(\boldsymbol{\theta}) \right) .$$

Notice that since Θ is symmetric, the number of free parameters in the θ vector and its representation $(\theta; \Theta)$ is the same, and we do not have any over-parametrization.

The natural parameterization and the mean and covariance parameterization are one-to-one. The inverse trasformation from natural parameters to the mean vector and the covariance matrix is given by $w^{-1} : \boldsymbol{\theta} \mapsto (\boldsymbol{\mu}; \Sigma)$, with

$$\boldsymbol{\mu} = -\frac{1}{2}\Theta^{-1}\theta , \tag{10}$$

$$\Sigma = -\frac{1}{2}\Theta^{-1} . \tag{11}$$

From Eq. (1) and (6), the log partition function as a function of $(\boldsymbol{\mu}; \Sigma)$ reads

$$\psi \circ \omega = \frac{1}{2}\left(n \log(2\pi) + \boldsymbol{\mu}^{\mathrm{T}} \Sigma^{-1} \boldsymbol{\mu} + \log|\Sigma| \right) ,$$

so that as a function of $\boldsymbol{\theta}$ it becomes

$$\psi(\boldsymbol{\theta}) = \frac{1}{2}\left(n \log(2\pi) - \frac{1}{2}\theta^{\mathrm{T}} \Theta^{-1} \theta - \log(-2)^n |\Theta| \right) . \tag{12}$$

Conditional independence assumptions between variables in X correspond to vanishing components in $\boldsymbol{\theta}$. As a consequence, the exponential manifold associated to the multivariate Gaussian distribution has a straightforward hierarchical structure, similar to what happens in the case of binary variables, cf [6].

PROPOSITION 7. *The conditional independence structure of the variables in $\boldsymbol{X}$ determines a hierarchical structure for $\mathcal{N}(\boldsymbol{\theta})$ where nested submanifolds given by some $\theta_{ij} = 0$ are identified by the conditional independence assumptions of the form $X_i \perp\!\!\!\perp X_j | \boldsymbol{X}_{\setminus i \setminus j}$.*

3.3 Expectation Parameters

The expectation parameters are a dual parameterization for statistical models in the exponential family, given by $\boldsymbol{\eta} = \mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{T}]$. Let $\chi : (\boldsymbol{\mu}; \Sigma) \mapsto \boldsymbol{\eta}$, from the definition of the sufficient statistics of the exponential family in Eq. (7), since $\mathrm{Cov}(X_i, X_j) = \mathbb{E}[X_iX_j] - \mathbb{E}[X_i]\mathbb{E}[X_j]$, it follows

$$\boldsymbol{\eta} = \chi(\boldsymbol{\mu}, \Sigma) = \begin{pmatrix} (\mu_i) \\ (\sigma_{ii} + \mu_i^2) \\ (2\sigma_{ij} + 2\mu_i\mu_j)_{i<j} \end{pmatrix} = \begin{pmatrix} (\eta_i) \\ (\eta_{ii}) \\ (\eta_{ij})_{i<j} \end{pmatrix} .$$

Similarly to the natural parameters, also the relationship between the expectation parameters and the mean and covariance parameterization is one-to-one. We introduce the following notation

$$\eta = (\eta_i) = \boldsymbol{\mu} , \tag{13}$$

$$E = \sum_i \eta_{ii} \boldsymbol{e}_i \boldsymbol{e}_i^{\mathrm{T}} + \sum_{i<j} \eta_{ij} \frac{\boldsymbol{e}_i \boldsymbol{e}_j^{\mathrm{T}} + \boldsymbol{e}_j \boldsymbol{e}_i^{\mathrm{T}}}{2} = \Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} , \tag{14}$$

and represent $\boldsymbol{\eta}$ as

$$\boldsymbol{\eta} = (\eta; E) ,$$

so that $\chi^{-1} : \boldsymbol{\eta} \mapsto (\boldsymbol{\mu}; \Sigma)$ can be written as

$$\boldsymbol{\mu} = \eta , \tag{15}$$

$$\Sigma = E - \eta\eta^{\mathrm{T}} . \tag{16}$$

The negative entropy of the multivariate Gaussian distribution parametrized by $(\boldsymbol{\mu}; \Sigma)$ reads

$$\varphi \circ \chi = -\frac{n}{2}\left(\log(2\pi) + 1 \right) - \frac{1}{2}\log|\Sigma| ,$$

so that in the expectation parameters we have

$$\varphi(\boldsymbol{\eta}) = -\frac{n}{2}\left(\log(2\pi) + 1 \right) - \frac{1}{2}\log\left| E - \eta\eta^{\mathrm{T}} \right| . \tag{17}$$

Combining Eq. (6) and (17), the multivariate Gaussian distribution in the $\boldsymbol{\eta}$ parameters can be written as

$$p(\boldsymbol{x}; \boldsymbol{\eta}) = \exp\left(-\frac{1}{2}(\boldsymbol{x} - \eta)^{\mathrm{T}} \left(E - \eta\eta^{\mathrm{T}} \right)^{-1} (\boldsymbol{x} - \eta) + \right.$$
$$\left. + \varphi(\boldsymbol{\eta}) + \frac{n}{2} \right) , \tag{18}$$

which a specific case of the more general formula for the exponential family parametrized in the expectation parameters, cf. Eq.(1) in [23], that is,

$$p(\boldsymbol{x}; \boldsymbol{\eta}) = \exp\left(\sum_{i=1}^{k} (T_i - \eta_i) \partial_i \varphi(\boldsymbol{\eta}) + \varphi(\boldsymbol{\eta}) \right) . \tag{19}$$

3.4 Change of Parameterization

In this section we have introduced three different parameterizations for a multivariate Gaussian distribution, namely the mean and covariance, the natural and the expectation parameterization.

By combining the transformations between $\boldsymbol{\eta}$, $\boldsymbol{\theta}$, and $(\boldsymbol{\mu};\Sigma)$ in Eq. (8)-(11) and (13)-(16), we have

$$\boldsymbol{\eta} = (\eta; E) = \left(-\frac{1}{2}\Theta^{-1}\theta; \frac{1}{4}\Theta^{-1}\theta\theta^{\mathrm{T}}\Theta^{-1} - \frac{1}{2}\Theta^{-1}\right), \quad (20)$$

$$\boldsymbol{\theta} = (\theta; \Theta) = \left(\left(E - \eta\eta^{\mathrm{T}}\right)^{-1}\eta; -\frac{1}{2}\left(E - \eta\eta^{\mathrm{T}}\right)^{-1}\right). \quad (21)$$

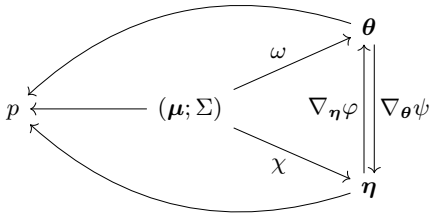
Moreover, from Eq. (2), (12), and (17) we obtain

$$\begin{aligned}\langle\boldsymbol{\theta},\boldsymbol{\eta}\rangle &= \frac{1}{2}\eta^{\mathrm{T}}\left(E - \eta\eta^{\mathrm{T}}\right)^{-1}\eta - \frac{n}{2}\,, \\ &= -\frac{1}{4}\theta^{\mathrm{T}}\Theta^{-1}\theta - \frac{n}{2}\,.\end{aligned} \quad (22)$$

Finally, by Eq. (18) and (22), the multivariate Gaussian distribution can be expressed as

$$p(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\eta}) = \exp\left(\boldsymbol{x}^{\mathrm{T}}\left(E - \eta\eta^{\mathrm{T}}\right)^{-1}\left(\eta - \frac{1}{2}\boldsymbol{x}\right) + \right.$$
$$\left. + \varphi(\boldsymbol{\eta}) - \langle\boldsymbol{\theta},\boldsymbol{\eta}\rangle\right).$$

The following commutative diagram summarize the transformations between the three parameterizations for the multivariate Gaussian distribution introduced in this section.



3.5 Fisher Information Matrix

In this section we introduce the formulae for the Fisher information matrix in the three different parameterizations we have introduced. The derivations have been included in the appendix.

In the general case of a statistical model parametrized by a parameter vector $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_d)^{\mathrm{T}}$, the standard definition of Fisher information matrix reads

$$I_{\boldsymbol{\xi}}(\boldsymbol{\xi}) = \mathbb{E}_{\boldsymbol{\xi}}\left[\left(\partial_i \log p(\boldsymbol{x};\boldsymbol{\xi})\right)\left(\partial_j \log p(\boldsymbol{x};\boldsymbol{\xi})\right)^{\mathrm{T}}\right]\,.$$

Under some regularity conditions, cf. Lemma 5.3 (b) in [18], and in particular when $\log p(\boldsymbol{x};\boldsymbol{\xi})$ is twice differentiable, the Fisher information matrix can be obtained by taking the negative of the expected value of the second derivative of the score, that is,

$$I_{\boldsymbol{\xi}}(\boldsymbol{\xi}) = -\mathbb{E}_{\boldsymbol{\xi}}\left[\partial_i\partial_j \log p(\boldsymbol{x};\boldsymbol{\xi})\right]\,.$$

For the multivariate Gaussian distribution, the Fisher information matrix has a special form, and can be obtained from a formula which depends on the derivatives with respect to the parameterization of the mean vector and the covariance matrix, c.f. [25, Thm. 2.1] and [24]. Let $\boldsymbol{\mu}$ and Σ be a function of the parameter vector $\boldsymbol{\xi}$ and ∂_i be the partial derivatives with respect to ξ_i, we have

$$I_{\boldsymbol{\xi}}(\boldsymbol{\xi}) = \left[(\partial_i\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\partial_j\boldsymbol{\mu}) + \frac{1}{2}\operatorname{Tr}\left(\Sigma^{-1}(\partial_i\Sigma)\Sigma^{-1}(\partial_j\Sigma)\right)\right] \quad (23)$$

Whenever we choose a parameterization for the Fisher information matrix for which the mean vector and the covariance matrix depend on two different vector parameters, Eq. (23) takes a special form and becomes block diagonal with $I_{\boldsymbol{\xi}}(\boldsymbol{\xi}) = \operatorname{diag}(I_{\boldsymbol{\mu}}, I_\Sigma)$, cf. [24]. The mean and covariance parameterization clearly satisfies this hypothesis, and by taking partial derivaties we obtain

$$I_{\boldsymbol{\mu},\Sigma}(\boldsymbol{\mu};\Sigma) = \begin{array}{c} \\ {\scriptstyle j} \\ {\scriptstyle mn}\end{array}\begin{array}{c}\begin{array}{cc}{\scriptstyle i} & {\scriptstyle kl}\end{array}\\ \left[\begin{array}{c|c} \Sigma^{-1} & 0 \\ \hline 0 & \alpha_{kl\,mn}\end{array}\right]\end{array}\,, \quad (24)$$

with

$$\alpha_{kl\,mn} = \begin{cases}\frac{1}{2}(\sigma^{kk})^2\,, & \text{if } k = l = m = n\,, \\ \sigma^{km}\sigma^{ln}\,, & \text{if } k = l \veebar m = n\,, \\ \sigma^{km}\sigma^{ln} + \sigma^{lm}\sigma^{kn}\,, & \text{otherwise.}\end{cases} \quad (25)$$

In the following we derive $I_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ and $I_{\boldsymbol{\eta}}(\boldsymbol{\eta})$ as the Hessian of $\psi(\boldsymbol{\theta})$ and $\psi(\boldsymbol{\eta})$, respectively, as in Eq. (3) and (4). Clearly, the derivations lead to the same formulae that would be obtained from Eq. (23). For the Fisher information matrix in the natural parameters we have

$$I_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \frac{1}{2} \times \begin{array}{c} \\ {\scriptstyle j} \\ {\scriptstyle mn}\end{array}\begin{array}{c}\begin{array}{cc}{\scriptstyle i} & {\scriptstyle kl}\end{array}\\ \left[\begin{array}{c|c} -\Theta^{-1} & \Lambda_{kl}\theta \\ \hline \theta^{\mathrm{T}}\Lambda_{mn} & \lambda_{kl\,mn} - \theta^{\mathrm{T}}\Lambda_{kl\,mn}\theta\end{array}\right]\end{array} \quad (26)$$

with

$$\Lambda_{kl} = \begin{cases}[\Theta^{-1}]_{\cdot k}[\Theta^{-1}]_{k\cdot}\,, & \text{if } k = l\,, \\ [\Theta^{-1}]_{\cdot k}[\Theta^{-1}]_{l\cdot} + \\ \quad + [\Theta^{-1}]_{\cdot l}[\Theta^{-1}]_{k\cdot}\,, & \text{otherwise,}\end{cases} \quad (27)$$

$$= \begin{cases}4\Sigma_{\cdot k}\Sigma_{l\cdot}\,, & \text{if } k = l\,, \\ 4\left(\Sigma_{\cdot k}\Sigma_{l\cdot} + \Sigma_{\cdot l}\Sigma_{k\cdot}\right)\,, & \text{otherwise,}\end{cases} \quad (28)$$

$$\lambda_{kl\,mn} = \begin{cases}[\Theta^{-1}]_{kk}[\Theta^{-1}]_{kk}\,, & \text{if } k = l = m = n\,, \\ [\Theta^{-1}]_{km}[\Theta^{-1}]_{ln} + \\ \quad + [\Theta^{-1}]_{lm}[\Theta^{-1}]_{kn}\,, & \text{if } k = l \veebar m = n\,, \\ 2\left([\Theta^{-1}]_{km}[\Theta^{-1}]_{ln} + \right. \\ \quad \left. + [\Theta^{-1}]_{lm}[\Theta^{-1}]_{kn}\right)\,, & \text{otherwise,}\end{cases} \quad (29)$$

$$= \begin{cases}4(\sigma_{kk}\sigma_{kk})\,, & \text{if } k = l = m = n\,, \\ 4(\sigma_{km}\sigma_{ln} + \sigma_{lm}\sigma_{kn})\,, & \text{if } k = l \veebar m = n\,, \\ 8(\sigma_{km}\sigma_{ln} + \sigma_{lm}\sigma_{kn})\,, & \text{otherwise,}\end{cases} \quad (30)$$

$$\Lambda_{kl\,mn} = \begin{cases}[\Lambda_{kk}]_{\cdot m}[\Theta^{-1}]_{n\cdot}\,, & \text{if } k = l\,, \\ [\Lambda_{kl}]_{\cdot m}[\Theta^{-1}]_{n\cdot} + \\ \quad + [\Lambda_{kl}]_{\cdot n}[\Theta^{-1}]_{m\cdot}\,, & \text{otherwise,}\end{cases} \quad (31)$$

$$= \begin{cases}-8\Sigma_{\cdot k}\sigma_{kk}\Sigma_{k\cdot}\,, & \text{if } k = l = m = n\,, \\ -8\left(\Sigma_{\cdot k}\sigma_{km}\Sigma_{n\cdot} + \right. \\ \quad \left. + \Sigma_{\cdot k}\sigma_{kn}\Sigma_{m\cdot}\right)\,, & \text{if } k = l \wedge m \neq n\,, \\ -8\left(\Sigma_{\cdot k}\sigma_{lm}\Sigma_{m\cdot} + \right. \\ \quad \left. + \Sigma_{\cdot l}\sigma_{lm}\Sigma_{m\cdot}\right)\,, & \text{if } k \neq l \wedge m = n\,, \\ -8\left(\Sigma_{\cdot k}\sigma_{lm}\Sigma_{n\cdot} + \right. \\ \quad + \Sigma_{\cdot k}\sigma_{ln}\Sigma_{m\cdot} + \\ \quad + \Sigma_{\cdot l}\sigma_{km}\Sigma_{n\cdot} + \\ \quad \left. + \Sigma_{\cdot l}\sigma_{kn}\Sigma_{m\cdot}\right)\,, & \text{otherwise.}\end{cases} \quad (32)$$

Where Λ_{kl} a matrix which depends on the indices k and l, and $\Lambda_{kl}\theta$ is a column vector. Notice that in case of $\boldsymbol{\mu} = 0$, $I_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ becomes block diagonal.

For models in the exponential family, we have a general formula based on covariances

$$I(\boldsymbol{\theta}) = \mathrm{Cov}_{\boldsymbol{\theta}}(T_i, T_j) \ . \tag{33}$$

By using Eq. (33), we can also derive the Fisher information matrix in the natural parameterization using covariances.

$$I_{\boldsymbol{\theta}}(\boldsymbol{\mu}, \Sigma) = \begin{array}{c} \\ j \\ mn \end{array} \overset{\begin{array}{cc} i & \qquad\qquad kl \end{array}}{\left[\begin{array}{c|c} \Sigma & a_{kl}(\mu_k \sigma_{lj} + \mu_l \sigma_{kj}) \\ \hline a_{mn}(\mu_m \sigma_{ni} + \mu_n \sigma_{mi}) & a_{kl\,mn}\gamma_{kl\,mn} \end{array}\right]}, \tag{34}$$

with

$$a_{kl} = \begin{cases} 1 \ , & \text{if } k = l \ , \\ 2 \ , & \text{otherwise.} \end{cases} \tag{35}$$

$$a_{kl\,mn} = \begin{cases} 1 \ , & \text{if } k = l = m = n \ , \\ 2 \ , & \text{if } k = l \trianglerighteq m = n \ , \\ 4 \ , & \text{otherwise.} \end{cases} \tag{36}$$

$$\begin{aligned} \gamma_{kl\,mn} = {} & \mu_n \mu_l \sigma_{km} + \mu_k \mu_m \sigma_{ln} + \mu_m \mu_l \sigma_{kn} + \mu_n \mu_k \sigma_{lm} + \\ & + \sigma_{km}\sigma_{ln} + \sigma_{lm}\sigma_{kn} \ . \end{aligned} \tag{37}$$

Finally, in the $\boldsymbol{\eta}$ parameters the Fisher information matrix becomes

$$I_{\boldsymbol{\eta}}(\boldsymbol{\eta}) = \begin{array}{c} \\ j \\ mn \end{array} \overset{\begin{array}{cc} i & \qquad kl \end{array}}{\left[\begin{array}{c|c} \Gamma & -K_{kl}\eta \\ \hline -\eta^{\mathrm{T}} K_{mn} & \kappa_{kl\,mn} \end{array}\right]}, \tag{38}$$

with

$$\begin{aligned} \Gamma &= (E - \eta\eta^{\mathrm{T}})^{-1} + (E - \eta\eta^{\mathrm{T}})^{-1}\eta^{\mathrm{T}}(E - \eta\eta^{\mathrm{T}})^{-1}\eta + \\ &\quad + (E - \eta\eta^{\mathrm{T}})^{-1}\eta\eta^{\mathrm{T}}(E - \eta\eta^{\mathrm{T}})^{-1} && (39) \\ &= \Sigma^{-1} + \Sigma^{-1}\boldsymbol{\mu}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu} + \Sigma^{-1}\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}\Sigma^{-1} && (40) \end{aligned}$$

$$K_{kl} = \begin{cases} [(E - \eta\eta^{\mathrm{T}})^{-1}]_{\cdot k}[(E - \eta\eta^{\mathrm{T}})^{-1}]_{k\cdot} \ , & \text{if } k = l \ , \\ \frac{1}{2}\left([(E - \eta\eta^{\mathrm{T}})^{-1}]_{\cdot k}[(E - \eta\eta^{\mathrm{T}})^{-1}]_{l\cdot} + \right. & \\ \quad \left. + [(E - \eta\eta^{\mathrm{T}})^{-1}]_{\cdot l}[(E - \eta\eta^{\mathrm{T}})^{-1}]_{k\cdot}\right) \ , & \text{otherwise.} \end{cases} \tag{41}$$

$$= \begin{cases} [\Sigma^{-1}]_{\cdot k}[\Sigma^{-1}]_{k\cdot} \ , & \text{if } k = l \ , \\ \frac{1}{2}\left([\Sigma^{-1}]_{\cdot k}[\Sigma^{-1}]_{l\cdot} + [\Sigma^{-1}]_{\cdot l}[\Sigma^{-1}]_{k\cdot}\right) \ , & \text{otherwise.} \end{cases} \tag{42}$$

$$\kappa_{kl\,mn} = \begin{cases} \frac{1}{2}[(E - \eta\eta^{\mathrm{T}})^{-1}]_{kk} \times & \\ \quad \times [(E - \eta\eta^{\mathrm{T}})^{-1}]_{kk} \ , & \text{if } k = l = m = n \ , \\ \frac{1}{2}[(E - \eta\eta^{\mathrm{T}})^{-1}]_{km} \times & \\ \quad \times [(E - \eta\eta^{\mathrm{T}})^{-1}]_{ln} \ , & \text{if } k = l \trianglerighteq m = n \ , \\ \frac{1}{4}\left([(E - \eta\eta^{\mathrm{T}})^{-1}]_{km} \times \right. & \\ \quad \times [(E - \eta\eta^{\mathrm{T}})^{-1}]_{ln} + & \\ \quad + [(E - \eta\eta^{\mathrm{T}})^{-1}]_{lm} \times & \\ \quad \left. \times \, [(E - \eta\eta^{\mathrm{T}})^{-1}]_{kn}\right) \ , & \text{otherwise.} \end{cases} \tag{43}$$

$$= \begin{cases} \frac{1}{2}(\sigma^{kk})^2 \ , & \text{if } k = l = m = n \ , \\ \frac{1}{2}\sigma^{km}\sigma^{ln} \ , & \text{if } k = l \trianglerighteq m = n \ , \\ \frac{1}{4}\left(\sigma^{km}\sigma^{ln} + \sigma^{lm}\sigma^{kn}\right) \ , & \text{otherwise.} \end{cases} \tag{44}$$

4. NATURAL GRADIENT

We are interested in optimizing a real-valued function f defined over $\mathbb{R}^n$, that is

$$\text{(P)} \qquad \min_{\boldsymbol{x} \in \mathbb{R}^n} f(\boldsymbol{x}) \ .$$

We replace the original optimization problem (P) with the stochastic relaxation F of f, i.e., the minimization of the expected value of f evaluated with respect to a density p which belongs to the multivariate Gaussian distribution, i.e.,

$$\text{(SR)} \qquad \min_{\boldsymbol{\xi} \in \Xi} F(\boldsymbol{\xi}) = \min_{\boldsymbol{\xi} \in \Xi} \mathbb{E}_{\boldsymbol{\xi}}[f(\boldsymbol{x})] \ .$$

Given a parametrization $\boldsymbol{\xi}$ for p, the natural gradient of $F(\boldsymbol{\xi})$ can be evaluated as

$$\widetilde{\nabla}_{\boldsymbol{\xi}} F(\boldsymbol{\xi}) = I(\boldsymbol{\xi})^{-1} \nabla_{\boldsymbol{\xi}} F(\boldsymbol{\xi}) \ ,$$

where $F(\boldsymbol{\xi})$ is the vector or partial derivatives, often called vanilla gradient.

Notice that for models in the exponential family, and thus for the Gaussian distribution, the vanilla gradient of $F(\boldsymbol{\xi})$ can be evaluated as the expected value of $f\nabla_{\boldsymbol{\xi}} \log p(\boldsymbol{x}; \boldsymbol{\xi})$, indeed

$$\partial_i F(\boldsymbol{\xi}) = \partial_i \mathbb{E}_{\boldsymbol{\xi}}[f] = \mathbb{E}_0[f \partial_i p(\boldsymbol{x}; \boldsymbol{\xi})] = \mathbb{E}_{\boldsymbol{\xi}}[f \partial_i \log p(\boldsymbol{x}; \boldsymbol{\xi})] \ .$$

In the mean and covariance parameterization, the vanilla gradient is given by

see for instance [3].

In the following, we provide formulae for $\nabla F(\boldsymbol{\theta})$ and $\nabla F(\boldsymbol{\eta})$, in the natural and expectation parameterizations. In the natural parameters we have

$$\partial_i \log p(\boldsymbol{x}; \boldsymbol{\theta}) = T_i(\boldsymbol{x}) - \partial_i \psi(\boldsymbol{\theta}) = T_i(\boldsymbol{x}) - \eta_i \ .$$

In the expectation parameters, by deriving the log of Eq. (19) we obtain

$$\partial_i \log p(\boldsymbol{x}; \boldsymbol{\eta}) = (T_i(\boldsymbol{x}) - \eta_i)\partial_i \partial_j \varphi(\boldsymbol{\eta}) \ .$$

So that

$$\begin{aligned} \nabla F(\boldsymbol{\theta}) &= \mathbb{E}_{\boldsymbol{\theta}}[f(T - \boldsymbol{\eta})] = \mathrm{Cov}_{\boldsymbol{\theta}}(f, \boldsymbol{T}) \ , \\ \nabla F(\boldsymbol{\eta}) &= \mathrm{Hess}\, \varphi(\boldsymbol{\eta}) \mathbb{E}_{\boldsymbol{\eta}}[f(T - \boldsymbol{\eta})] \\ &= I_{\boldsymbol{\eta}}(\boldsymbol{\eta})\, \mathrm{Cov}_{\boldsymbol{\eta}}(f, \boldsymbol{T}) \ . \end{aligned}$$

A common approach to solve the (SR) is given by natural gradient descent, when the parameters of a density are updated iteratively in the direction given by the natural gradient of F, i.e.,

$$\boldsymbol{\xi}^{t+1} = \boldsymbol{\xi}^t - \lambda \widetilde{\nabla}_{\boldsymbol{\xi}} F(\boldsymbol{\xi}) \ ,$$

where the parameter $\lambda > 0$ controls the step size.

5. EXAMPLES

In this section we introduce and discuss some toy examples, for which it is possible to represent the gradient flows and the landscape of the stochastic relaxation. In particular we evaluate the gradient flows, i.e., the solutions of the differential equations

$$\dot{\boldsymbol{\xi}} = -\widetilde{\nabla}_{\boldsymbol{\xi}} F(\boldsymbol{\xi}).$$

Such flows correspond to the trajectories associated to infinite step size, when the gradient can be computed exactly.

5.1 Polynomial Optimization in $\mathbb{R}$

In this section we study some simple examples of polynomial optimization in $\mathbb{R}$. Let f be a real-valued polynomial function, we choose a monomial basis $\{x^k\}$, $k > 0$, so that any polynomial can be written in compact form as

$$f_k = \sum_{i=0}^{k} c_i x^i \ . \tag{45}$$

We consider the case of quadratic functions, where $k = 2$ in Eq. (45), so that $f_2 = c_0 + c_1 x + c_2 x^2$. In order for quadratic functions to be lower bounded and this admit a minimum, we need to impose $c_2 > 0$. We consider the one dimensional Gaussian $\mathcal{N}(\mu, \sigma)$ distribution parametrized by μ, σ, and denote by $F(\mu, \sigma)$ the stochastic relaxation of f with respect to $\mathcal{N}$. Represented as an exponential family, $\mathcal{N}(\mu, \sigma)$ is a two-dimensional exponential family, with sufficient statistics $\boldsymbol{T}$ given by X and X^2.

Let $\boldsymbol{c} = (c_1, c_2)^{\mathrm{T}}$, in the $\boldsymbol{\eta}$ parameters the vanilla and natural gradient read

$$\nabla_{\boldsymbol{\eta}} F(\boldsymbol{\eta}) = \nabla_{\boldsymbol{\eta}} \sum_{i=1}^{2} c_i \mathbb{E}_{\boldsymbol{\eta}}[X^i] = \nabla_{\boldsymbol{\eta}} (c_1 \mu_1, c_2 E_{11})^{\mathrm{T}} = \boldsymbol{c} \ ,$$
$$\widetilde{\nabla}_{\boldsymbol{\eta}} F(\boldsymbol{\eta}) = I_{\boldsymbol{\eta}}(\boldsymbol{\eta})^{-1} \nabla_{\boldsymbol{\eta}} F(\theta) \ .$$

The vanilla and natural gradients in $\boldsymbol{\theta}$ are

$$\nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta}) = \mathrm{Cov}_{\boldsymbol{\theta}}(f, \boldsymbol{T}) = \sum_{i=1}^{2} c_i \, \mathrm{Cov}_{\boldsymbol{\theta}}(X^i, \boldsymbol{T}) = I(\boldsymbol{\theta}) \boldsymbol{c} \ ,$$
$$\widetilde{\nabla}_{\boldsymbol{\theta}} F(\boldsymbol{\theta}) = I_{\boldsymbol{\theta}}(\boldsymbol{\theta})^{-1} \nabla_{\boldsymbol{\theta}} F(\theta) = \boldsymbol{c} \ .$$

Since f belongs to the span of the $\{T_i\}$, so that for any initial condition q in $\mathcal{N}(\mu, \sigma)$ the natural gradient flows in any parametrization weakly converge to the δ distribution over the minimum of f, given by $x^* = -\frac{c_1}{2c_2}$. Such distribution belongs to the closure of the Gaussian distribution and can be identified as the limit distribution with $\mu = -\frac{c_1}{2c_2}$ and $\sigma \to 0$. In Fig. 1 we represented an instance of this example, where $f = x - 3x^2$. Notice that the flow associated to the vanilla gradient in $\boldsymbol{\eta}$ is linear in (μ, σ) and stops at the boundary of the model, where it reaches the positivity constraint for σ. All other trajectories converge to the global minimum, and natural gradient flows defines straight paths to the optimum.

We move to the case where the polynomial f_k has higher degree. We do not consider the case when $k = 3$ since f_3 would not be lower bounded, and study the polynomial for $k = 4$, so that $f_4 = c_0 + c_1 x + c_2 x^2 + c_3 x^3 + c_4 x^4$. Notice that f_4 doest not belong anymore to the span of the sufficient statistics of the exponential family, and the function may have two local minima in $\mathbb{R}$. Similarly, the relaxation with respect to the one dimensional gaussian family $\mathcal{N}(\mu, \sigma)$ may admin two local minima associated to the δ distributions over the local minima of f.

Vanilla and natural gradient formulae can be computed in closed form, indeed in the exponential family higher order moment $\mathbb{E}[X^k]$ can be evaluated recursively as a function of $\boldsymbol{\eta}$, by expanding $\mathbb{E}[(X - \mathbb{E}[X])^k]$ using the binomial formula, and then applying Isserlis' theorem for centered moments, cf. [14].

In the $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$ parameters the vanilla gradients read

$$\nabla_{\boldsymbol{\eta}} F(\boldsymbol{\eta}) = \boldsymbol{c} + \sum_{i=3}^{k} c_i \nabla_{\boldsymbol{\eta}} \mathbb{E}_{\boldsymbol{\eta}}[X^i] \ ,$$
$$\nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta}) = I(\boldsymbol{\theta}) \boldsymbol{c} + \sum_{i=3}^{k} c_i \, \mathrm{Cov}_{\boldsymbol{\theta}}(X^i, \boldsymbol{T}) \ ,$$

while natural gradients can be evaluated by premultiplying vanilla gradient with the inverse of the Fisher information matrix.

In Fig. 2 we plotted different trajectories for the case $f = 6x + 8x^2 - x^3 - 2x^4$. We can notice two different basins of attraction, so that the trajectories associated to the natural gradient flows converge to either one or the other local minima depending on the initial condition. As in the case of f_2 vanilla flows in $\boldsymbol{\eta}$ converge to the boundary of the model where $\sigma \to 0$, and trajectories are not straight in (μ, σ).

5.2 Polynomial Optimization in $\mathbb{R}^n$

The examples in the previous subsection can be easily generalized to the case of polynomial functions in $\mathbb{R}^n$.

In Fig. 3 we studied the case where $f = x_1 + 2x_2 - 3x_1^2 - 2x_1 x_2 - 2x_2^2$. In this example, the multivariate Gaussian distribution is a $5-$dimensional exponential family, for this reason we plot the projections of the flows onto $\boldsymbol{\mu} = (\mu_1, \mu_2)$, and represent the level lines of f instead of those of $F(\mu, \Sigma)$. This explains while trajectories appear to self intersect in the projected space, which would be impossible for any gradient flow over $\mathcal{N}$. However, since f is a quadratic function, we are guaranteed that the natural gradient flows converge to the δ distribution over the unique global optimum of f for any initial condition.

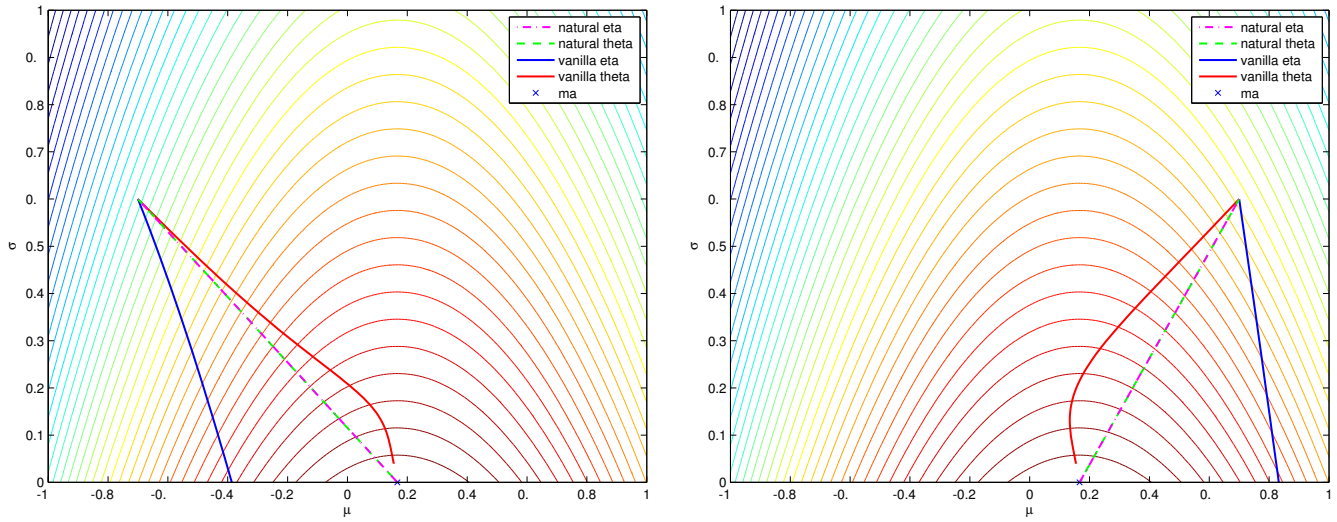
6. CONCLUSIONS

This paper focuses on the study of the geometry of the multivariate Gaussian distribution, and more generally of models in the exponential family from the perspective of stochastic relaxation. We discussed two alternative parameterizations to the mean vector and covariance matrix for the multivariate Gaussian distribution, namely the natural and expectation parameterizations of the exponential family. We derived variable transformations between each parameterization and the formulae for the natural gradients. Since the natural gradient is invariant with respect to the choice of the parameterization, following the natural gradient in any of these parameterizations is equivalent from the point of view of the optimization.

On the other side, by exploiting the specific properties of each parameterization, and the relationship between Fisher information matrices, we can define alternative algorithms for natural gradient descent. In particular, by parametrizing the Gaussian distribution in the natural parameters we have the advantage of a meaningful representation for lower dimensional sub-models of the Gaussian distribution, together with closed formulae for the inverse of the Fisher information matrix, which allow to easily evaluate the natural gradient.

7. ACKNOWLEDGEMENTS

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



(a) (b)

Figure 1: Vanilla vs natural gradient flows for $\mathbb{E}[f]$, with $f = x - 3x^2$, evaluated in $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$ parameters and represented in the parameter space (μ, σ). Each figure represents the flows for different initial conditions. The flows are evaluated solving the differential equations numerically. The level lines are associated to $\mathbb{E}_{\mu,\sigma}[f]$.

8. REFERENCES

[1] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, Princeton, NJ, 2008. With a foreword by Paul Van Dooren.

[2] Y. Akimoto, A. Auger, and N. Hansen. Convergence of the continuous time trajectories of isotropic evolution strategies on monotonic $\mathcal{C}^2$-composite functions. In C. Coello, V. Cutello, K. Deb, S. Forrest, G. Nicosia, and M. Pavone, editors, *Parallel Problem Solving from Nature - PPSN XII*, volume 7491 of *Lecture Notes in Computer Science*, pages 42–51. Springer Berlin Heidelberg, 2012.

[3] Y. Akimoto, Y. Nagata, I. Ono, and S. Kobayashi. Theoretical foundation for cma-es from information geometry perspective. *Algorithmica*, 64(4):698–716, 2012.

[4] S. Amari. *Differential-geometrical methods in statistics*, volume 28 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1985.

[5] S. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

[6] S. Amari. Information geometry on hierarchy of probability distributions. *IEEE Transactions on Information Theory*, 47(5):1701–1711, 2001.

[7] S. Amari and H. Nagaoka. *Methods of information geometry*. American Mathematical Society, Providence, RI, 2000. Translated from the 1993 Japanese original by Daishi Harada.

[8] O. E. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory*. John Wiley & Sons, New York, 1978.

[9] J. Bensadon. Black-box optimization using geodesics in statistical manifolds. *Entropy*, 17(1):304–345, 2015.

[10] H.-G. Beyer. Convergence analysis of evolutionary algorithms that are based on the paradigm of information geometry. *Evol. Comput.*, 22(4):679–709, Dec. 2014.

[11] P. Billingsley. *Convergence of probability measures*. John Wiley & Sons, Inc., New York-London-Sydney, 1968.

[12] L. D. Brown. *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory*, volume 9 of *Lecture Notes - Monograph Series*. Institute of Mathematical Statistics, 1986.

[13] N. Hansen and A. Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2):159–195, 2001.

[14] L. Isserlis. On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, 12(1-2):134–139, 1918.

[15] R. E. Kass and P. W. Vos. *Geometrical Foundations of Asymptotic Inference*. Wiley Series in Probability and Statistics. John Wiley, New York, 1997.

[16] P. Larrañaga and J. A. Lozano, editors. *Estimation of Distribution Algoritms. A New Tool for evolutionary Computation*. Springer, 2001.

[17] S. L. Lauritzen. *Graphical models*. The Clarendon Press Oxford University Press, New York, 1996.

[18] E. Lehmann and G. Casella. *Theory of Point Estimation*. Springer Verlag, second edition, 1998.

[19] L. Malagò, M. Matteucci, and G. Pistone. 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.

[20] L. Malagò, M. Matteucci, and G. Pistone. Towards the geometry of estimation of distribution algorithms based on the exponential family. In *Proc. of FOGA '11*, pages 230–242. ACM, 2011.

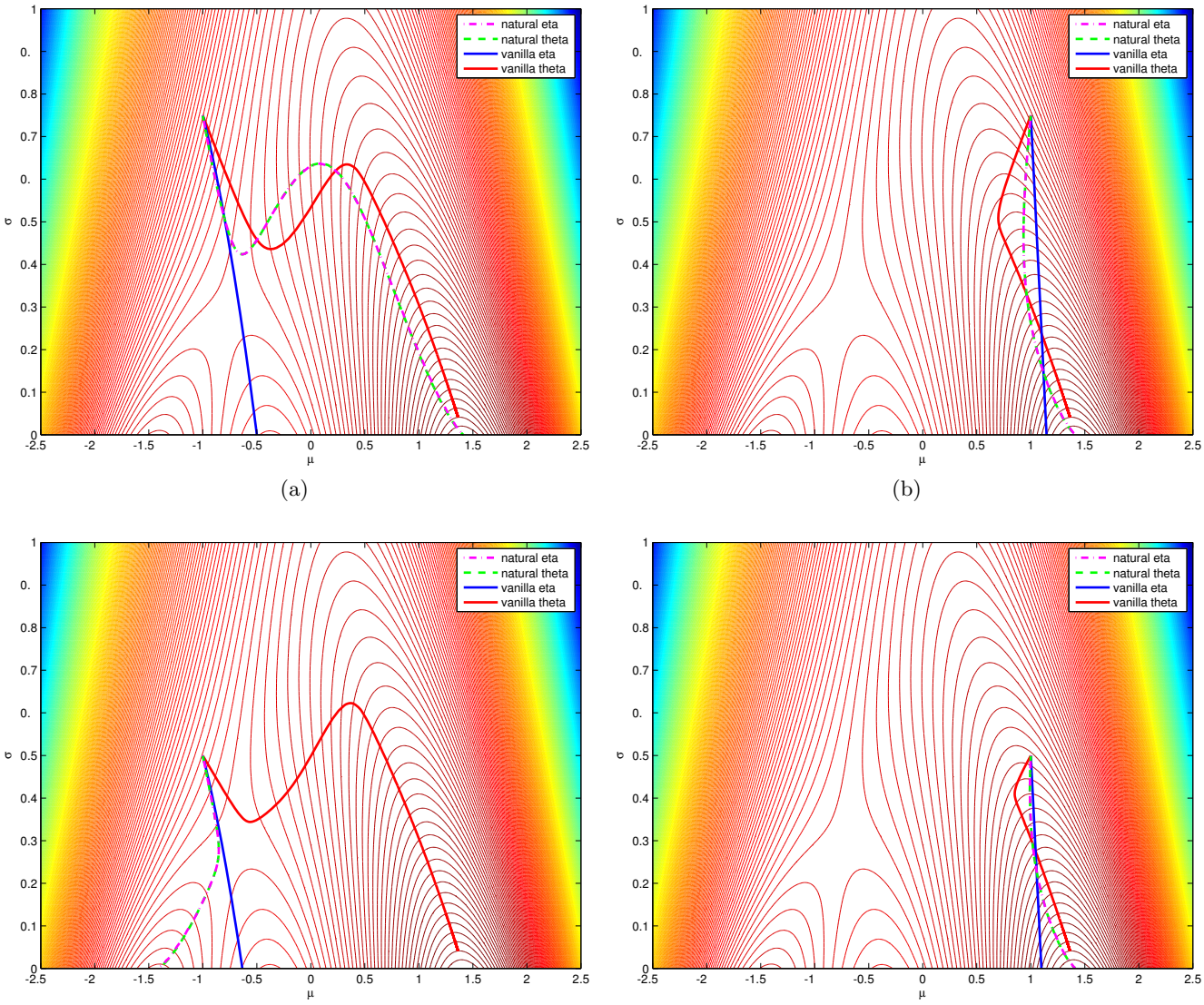
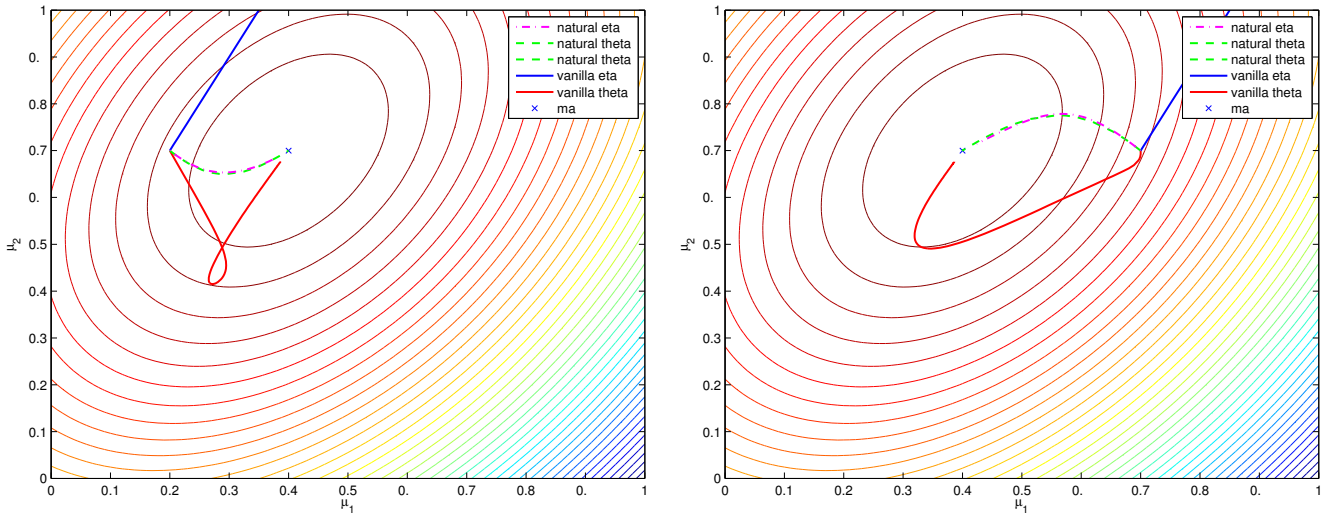


Figure 2: Vanilla vs natural gradient flows for $\mathbb{E}[f]$, with $f = 6x + 8x^2 - x^3 - 2x^4$, evaluated in $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$ parameters and represented in the parameter space (μ, σ). Each figure represents the flows for different initial conditions. The flows are evaluated solving the differential equations numerically. The level lines are associated to $\mathbb{E}_{\mu,\sigma}[f]$.



(a) (b)

Figure 3: Vanilla vs natural gradient flows for $\mathbb{E}[f]$, with $f = x_1 + 2x_2 - 3x_1^2 - 2x_1x_2 - 2x_2^2$, evaluated in $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$ parameters and represented in the parameter space (μ_1, μ_2). Each figure represents the projections of the flows for different initial conditions, with $\sigma_{11} = 1, \sigma_{12} = -0.5, \sigma_{22} = 2$. The flows are evaluated solving the differential equations numerically. The level lines are associated to f.

[21] L. Malagò, M. Matteucci, and G. Pistone. Natural gradient, fitness modelling and model selection: A unifying perspective. In *Proc. of IEEE CEC 2013*, pages 486–493, 2013.

[22] L. Malagò and G. Pistone. Combinatorial optimization with information geometry: The newton method. *Entropy*, 16(8):4260–4289, 2014.

[23] L. Malagò and G. Pistone. Gradient flow of the stochastic relaxation on a generic exponential family. *AIP Conference Proceedings of MaxEnt 2014, held on September 21-26, 2014, Château du Clos Lucé, Amboise, France*, 1641:353–360, 2015.

[24] K. V. Mardia and R. J. Marshall. Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, 71(1):135–146, 1984.

[25] K. S. Miller. *Complex stochastic processes: an introduction to theory and application.* Addison-Wesley Pub. Co., 1974.

[26] Y. Ollivier, L. Arnold, A. Auger, and N. Hansen. Information-geometric optimization algorithms: A unifying picture via invariance principles. arXiv:1106.3708, 2011v1; 2013v2.

[27] G. Pistone. A version of the geometry of the multivariate gaussian model, with applications. In *Proceedings of the 47th Scientific Meeting of the Italian Statistical Society, SIS 2014, Cagliari, June 11-13*, 2014.

[28] R. T. Rockafellar. *Convex analysis.* Princeton Mathematical Series, No. 28. Princeton University Press, Princeton, N.J., 1970.

[29] R. Y. Rubistein and D. P. Kroese. *The Cross-Entropy method: a unified approach to combinatorial optimization, Monte-Carlo simluation, and machine learning.* Springer, New York, 2004.

[30] L. T. Skovgaard. A Riemannian Geometry of the Multivariate Normal Model. *Scandinavian Journal of Statistics*, 11(4):211–223, 1984.

[31] D. Wierstra, T. Schaul, T. Glasmachers, Y. Sun, J. Peters, and J. Schmidhuber. Natural evolution strategies. *Journal of Machine Learning Research*, 15:949–980, 2014.

[32] D. Wierstra, T. Schaul, J. Peters, and J. Schmidhuber. Natural evolution strategies. In *Proc. of IEEE CEC 2008*, pages 3381–3387, 2008.

APPENDIX

A. COMPUTATIONS FOR THE FISHER INFORMATION MATRIX

In this appendix we included the derivations for the Fisher information matrix in the different parameterizations.

A.1 Mean and Covariance Parameters

Since $\partial_i \boldsymbol{\mu} = \boldsymbol{e}_i$, we have $I_{\boldsymbol{\mu}} = \Sigma^{-1}$. As to I_Σ, first notice that

$$\partial_{ij}\Sigma = \begin{cases} \boldsymbol{e}_i \boldsymbol{e}_i^{\mathrm{T}} , & \text{if } i = j , \\ \boldsymbol{e}_i \boldsymbol{e}_j^{\mathrm{T}} + \boldsymbol{e}_j \boldsymbol{e}_i^{\mathrm{T}} , & \text{otherwise,} \end{cases}$$

so that for $k \neq l \wedge m \neq n$, we obtain

$$
\begin{aligned}
[I_\Sigma]_{kl\,mn} &= \frac{1}{2}\operatorname{Tr}\left(\Sigma^{-1}(\partial_{kl}\Sigma)\Sigma^{-1}(\partial_{mn}\Sigma)\right) \\
&= \frac{1}{2}\operatorname{Tr}\left(\Sigma^{-1}(\boldsymbol{e}_k\boldsymbol{e}_l^{\mathrm{T}} + \boldsymbol{e}_l\boldsymbol{e}_k^{\mathrm{T}})\Sigma^{-1}(\boldsymbol{e}_m\boldsymbol{e}_n^{\mathrm{T}} + \boldsymbol{e}_n\boldsymbol{e}_m^{\mathrm{T}})\right) \\
&= \frac{1}{2}\operatorname{Tr}\left(2\boldsymbol{e}_l^{\mathrm{T}}\Sigma^{-1}\boldsymbol{e}_m\boldsymbol{e}_n^{\mathrm{T}}\Sigma^{-1}\boldsymbol{e}_k + 2\boldsymbol{e}_k^{\mathrm{T}}\Sigma^{-1}\boldsymbol{e}_m\boldsymbol{e}_n^{\mathrm{T}}\Sigma^{-1}\boldsymbol{e}_l\right) \\
&= \boldsymbol{e}_l^{\mathrm{T}}\Sigma^{-1}\boldsymbol{e}_m\boldsymbol{e}_n^{\mathrm{T}}\Sigma^{-1}\boldsymbol{e}_k + \boldsymbol{e}_k^{\mathrm{T}}\Sigma^{-1}\boldsymbol{e}_m\boldsymbol{e}_n^{\mathrm{T}}\Sigma^{-1}\boldsymbol{e}_l \\
&= \sigma^{km}\sigma^{ln} + \sigma^{lm}\sigma^{kn} \ .
\end{aligned}
$$

In the remaining cases, when $k = l = m = n$ and $k = l \veebar m = n$, the computations are analogous, giving the formulae in Eq. (24) and (25).

A.2 Natural Parameters

In the following we derive the Fisher information matrix in the natural parameters by taking the Hessian of $\psi(\boldsymbol{\theta})$ in Eq. (12). We start by taking first-order derivatives, i.e.,

$$
\partial_i\psi(\boldsymbol{\theta}) = -\frac{1}{2}\boldsymbol{e}_i^{\mathrm{T}}\Theta^{-1}\theta \ , \tag{46}
$$

$$
\begin{aligned}
\partial_{ij}\psi(\boldsymbol{\theta}) &= \frac{1}{4}\theta^{\mathrm{T}}\Theta^{-1}(\partial_{ij}\Theta)\Theta^{-1}\theta - \frac{1}{2}\operatorname{Tr}\left(-\frac{1}{2}\Theta^{-1}\partial_{ij}(-2\Theta)\right) \\
&= \frac{1}{4}\theta^{\mathrm{T}}\Theta^{-1}(\partial_{ij}\Theta)\Theta^{-1}\theta - \frac{1}{2}\operatorname{Tr}(\Theta^{-1}(\partial_{ij}\Theta))
\end{aligned} \tag{47}
$$

Notice that

$$
\partial_{ij}\Theta = \begin{cases} \boldsymbol{e}_i\boldsymbol{e}_i^{\mathrm{T}} \ , & \text{if } i = j \ , \\ \boldsymbol{e}_i\boldsymbol{e}_j^{\mathrm{T}} + \boldsymbol{e}_j\boldsymbol{e}_i^{\mathrm{T}} \ , & \text{otherwise,} \end{cases}
$$

so that as in Eq. (20), we have

$$
\begin{aligned}
\eta &= -\frac{1}{2}\Theta^{-1}\theta \ , \\
E &= \frac{1}{4}\Theta^{-1}\theta\theta^{\mathrm{T}}\Theta^{-1} - \frac{1}{2}\Theta^{-1} \ .
\end{aligned}
$$

Next, we take partial derivatives of Eq. (46) and (47), and since $\partial_{kl}\partial_{ij}\Theta = 0$, we get

$$
\begin{aligned}
\partial_i\partial_j\psi(\boldsymbol{\theta}) &= -\frac{1}{2}\boldsymbol{e}_i^{\mathrm{T}}\Theta^{-1}\boldsymbol{e}_j \ , \\
\partial_i\partial_{kl}\psi(\boldsymbol{\theta}) &= \frac{1}{2}\boldsymbol{e}_i^{\mathrm{T}}\Theta^{-1}(\partial_{kl}\Theta)\Theta^{-1}\theta \ .
\end{aligned}
$$

Let $\Lambda_{kl} = \Theta^{-1}(\partial_{kl}\Theta)\Theta^{-1}$, for $k \neq l$ we have

$$
\begin{aligned}
\Lambda_{kl} &= \Theta^{-1}(\boldsymbol{e}_k\boldsymbol{e}_l^{\mathrm{T}} + \boldsymbol{e}_l\boldsymbol{e}_k^{\mathrm{T}})\Theta^{-1} \\
&= \Theta^{-1}\boldsymbol{e}_l\boldsymbol{e}_k^{\mathrm{T}}\Theta^{-1} + \Theta^{-1}\boldsymbol{e}_k\boldsymbol{e}_l^{\mathrm{T}}\Theta^{-1} \\
&= 4(\Sigma\boldsymbol{e}_l\boldsymbol{e}_k^{\mathrm{T}}\Sigma + \Sigma\boldsymbol{e}_k\boldsymbol{e}_l^{\mathrm{T}}\Sigma) \ .
\end{aligned}
$$

In the remaining case, when $k \neq l$, the computations are analogous, giving the formulae in Eq. (27), (28) and (26).

$$
\begin{aligned}
\partial_{kl}\partial_{mn}\psi(\boldsymbol{\theta}) = &-\frac{1}{2}\theta^{\mathrm{T}}\Theta^{-1}(\partial_{kl}\Theta)\Theta^{-1}(\partial_{mn}\Theta)\Theta^{-1}\theta + \\
&+ \frac{1}{2}\operatorname{Tr}\left(\Theta^{-1}(\partial_{kl}\Theta)\Theta^{-1}(\partial_{mn}\Theta)\right) \ .
\end{aligned}
$$

Let $\Lambda_{kl\,mn} = \Theta^{-1}(\partial_{kl}\Theta)\Theta^{-1}(\partial_{mn}\Theta)\Theta^{-1}$, for $k \neq l \wedge m \neq n$

$$
\begin{aligned}
\Lambda_{kl\,mn} &= \Theta^{-1}(\boldsymbol{e}_k\boldsymbol{e}_l^{\mathrm{T}} + \boldsymbol{e}_l\boldsymbol{e}_k^{\mathrm{T}})\Theta^{-1}(\boldsymbol{e}_m\boldsymbol{e}_n^{\mathrm{T}} + \boldsymbol{e}_n\boldsymbol{e}_m^{\mathrm{T}})\Theta^{-1} \\
&= \Theta^{-1}\boldsymbol{e}_k\boldsymbol{e}_l^{\mathrm{T}}\Theta^{-1}\boldsymbol{e}_m\boldsymbol{e}_n^{\mathrm{T}}\Theta^{-1} + \Theta^{-1}\boldsymbol{e}_k\boldsymbol{e}_l^{\mathrm{T}}\Theta^{-1}\boldsymbol{e}_n\boldsymbol{e}_m^{\mathrm{T}}\Theta^{-1} + \\
&\quad \Theta^{-1}\boldsymbol{e}_l\boldsymbol{e}_k^{\mathrm{T}}\Theta^{-1}\boldsymbol{e}_m\boldsymbol{e}_n^{\mathrm{T}}\Theta^{-1} + \Theta^{-1}\boldsymbol{e}_l\boldsymbol{e}_k^{\mathrm{T}}\Theta^{-1}\boldsymbol{e}_n\boldsymbol{e}_m^{\mathrm{T}}\Theta^{-1} \\
&= -8(\Sigma\boldsymbol{e}_k\sigma_{lm}\boldsymbol{e}_n^{\mathrm{T}}\Sigma + \Sigma\boldsymbol{e}_k\sigma_{ln}\boldsymbol{e}_m^{\mathrm{T}}\Sigma + \Sigma\boldsymbol{e}_l\sigma_{km}\boldsymbol{e}_n^{\mathrm{T}}\Sigma + \\
&\quad + \Sigma\boldsymbol{e}_l\sigma_{kn}\boldsymbol{e}_m^{\mathrm{T}}\Sigma) \ .
\end{aligned}
$$

In the remaining cases when $k = l = m = n$, $k = l \wedge m \neq n$ and $k \neq l \wedge m = n$, the computations are analogous, giving the formulae in Eq. (31), (32) and (26).

Finally, $\lambda_{kl\,mn} = \operatorname{Tr}\left(\Theta^{-1}(\partial_{kl}\Theta)\Theta^{-1}(\partial_{mn}\Theta)\right)$, we have for $k \neq l \wedge m \neq n$

$$
\begin{aligned}
\lambda_{kl\,mn} &= \operatorname{Tr}\left(\Theta^{-1}(\boldsymbol{e}_k\boldsymbol{e}_l^{\mathrm{T}} + \boldsymbol{e}_l\boldsymbol{e}_k^{\mathrm{T}})\Theta^{-1}(\boldsymbol{e}_m\boldsymbol{e}_n^{\mathrm{T}} + \boldsymbol{e}_n\boldsymbol{e}_m^{\mathrm{T}})\right) \\
&= 2\boldsymbol{e}_l^{\mathrm{T}}\Theta^{-1}\boldsymbol{e}_m\boldsymbol{e}_n\Theta^{-1}\boldsymbol{e}_k + 2\boldsymbol{e}_k^{\mathrm{T}}\Theta^{-1}\boldsymbol{e}_m\boldsymbol{e}_n\Theta^{-1}\boldsymbol{e}_l \\
&= 8\boldsymbol{e}_l^{\mathrm{T}}\Sigma\boldsymbol{e}_m\boldsymbol{e}_n^{\mathrm{T}}\Sigma\boldsymbol{e}_k + 8\boldsymbol{e}_k^{\mathrm{T}}\Sigma\boldsymbol{e}_m\boldsymbol{e}_n^{\mathrm{T}}\Sigma\boldsymbol{e}_l \ .
\end{aligned}
$$

In the remaining cases, when $k = l = m = n$ and $k = l \veebar m = n$, the computations are analogous, giving the formulae in Eq. (29), (30) and (26).

Next, we derive an equivalent formulation for the Fisher information matrix based on covariances. From Eq. (33), we have that the elements of the Fisher information matrix in $\boldsymbol{\theta}$ can be obtained from the covariances of sufficient statistics. Moreover, from the definition of covariance, we have

$$
\begin{aligned}
\operatorname{Cov}_{\boldsymbol{\mu},\Sigma}(X_i, X_j) &= \sigma_{ij} \ , \\
\operatorname{Cov}_{\boldsymbol{\mu},\Sigma}(X_i, X_kX_l) &= \mathbb{E}_{\boldsymbol{\mu},\Sigma}[X_iX_kX_l] + \\
&\quad - \mathbb{E}_{\boldsymbol{\mu},\Sigma}[X_i]\mathbb{E}_{\boldsymbol{\mu},\Sigma}[X_kX_l] \ , \\
\operatorname{Cov}_{\boldsymbol{\mu},\Sigma}(X_kX_l, X_mX_n) &= \mathbb{E}_{\boldsymbol{\mu},\Sigma}[X_kX_lX_mX_n] + \\
&\quad - \mathbb{E}_{\boldsymbol{\mu},\Sigma}[X_kX_l]\mathbb{E}_{\boldsymbol{\mu},\Sigma}[X_mX_ln] \ .
\end{aligned}
$$

The definition of first- and second-order moments in terms of mean vector and covariance matrix are straightforward,

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{\mu},\Sigma}[X_i] &= \mu_i \\
\mathbb{E}_{\boldsymbol{\mu},\Sigma}[X_iX_j] &= \sigma_{ij} + \mu_i\mu_j \ .
\end{aligned}
$$

In order to evaluate third- and forth-order moments we use Isserlis' theorem [14] after centering variables by replacing X_i with $X_i - \mathbb{E}_{\boldsymbol{\mu},\Sigma}[X_i]$, which gives

$$
\mathbb{E}_{\boldsymbol{\mu},\Sigma}[X_iX_kX_l] = \mu_i\sigma_{kl} + \mu_k\sigma_{il} + \mu_l\sigma_{ik} + \mu_i\mu_k\mu_l \ ,
$$

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{\mu},\Sigma}[X_kX_lX_mX_n] = &\ \sigma_{kn}\sigma_{lm} + \sigma_{km}\sigma_{ln} + \sigma_{kl}\sigma_{mn} + \\
&+ \sum_{\{\tau(k)\}\{\tau(l)\}\{\tau(m)\tau(n)\}} \sigma_{\tau(k)\tau(l)}\mu_{\tau(m)}\mu_{\tau(n)} + \\
&+ \mu_k\mu_l\mu_m\mu_n \ ,
\end{aligned}
$$

where $\{\tau(k)\}\{\tau(l)\}\{\tau(m)\tau(n)\}$ denotes the combinations of the indices k, l, m, n without repetitions, where indices have divided into three groups, $\{\tau(k)\}$, $\{\tau(l)\}$ and $\{\tau(m)\tau(n)\}$. Finally, by using the formulae for higher-order moments in terms of mean and covariance we obtain

$$
\begin{aligned}
\operatorname{Cov}_{\boldsymbol{\mu},\Sigma}(X_i, X_kX_l) &= \mathbb{E}_{\boldsymbol{\mu},\Sigma}[X_iX_kX_l] - \mathbb{E}_{\boldsymbol{\mu},\Sigma}[X_i]\mathbb{E}_{\boldsymbol{\mu},\Sigma}[X_kX_l] \\
&= \mu_k\sigma_{il} + \mu_l\sigma_{ik} \ ,
\end{aligned}
$$

and

$$\begin{aligned}\mathrm{Cov}_{\boldsymbol{\mu},\Sigma}(X_kX_l, X_mX_n) &= \mathbb{E}_{\boldsymbol{\mu},\Sigma}[X_kX_lX_mX_n] + \\ &- \mathbb{E}_{\boldsymbol{\mu},\Sigma}[X_kX_l]\mathbb{E}_{\boldsymbol{\mu},\Sigma}[X_mX_n] \\ &= \mu_n\mu_l\sigma_{km} + \mu_k\mu_m\sigma_{ln} + \mu_m\mu_l\sigma_{kn} + \mu_n\mu_k\sigma_{lm} + \\ &+ \sigma_{km}\sigma_{ln} + \sigma_{lm}\sigma_{kn} \ .\end{aligned}$$

The results are summarized in Eq. (35), (36), (37) and (34). Notice that for $k \neq l$, $T_{kl} = 2X_kX_l$, and thus $\mathrm{Cov}_{\boldsymbol{\mu},\Sigma}(T_i, T_{kl}) = 2\,\mathrm{Cov}_{\boldsymbol{\mu},\Sigma}(X_i, X_kX_l)$, we introduced the coefficients a_{kl} and a_{klmn} to compensate for the constant.

A.3 Expectation Parameters

In the following we derive the Fisher information matrix in the expectation parameters by taking the Hessian of $\varphi(\boldsymbol{\eta})$ in Eq. (17). We start by taking first-order derivatives, i.e.,

$$\partial_i\varphi(\boldsymbol{\eta}) = \frac{1}{2}\,\mathrm{Tr}\left((E - \eta\eta^{\mathrm{T}})^{-1}\partial_i(\eta\eta^{\mathrm{T}})\right) \tag{48}$$

$$\partial_{ij}\varphi(\boldsymbol{\eta}) = -\frac{1}{2}\,\mathrm{Tr}\left((E - \eta\eta^{\mathrm{T}})^{-1}(\partial_{ij}E)\right) \tag{49}$$

Notice that

$$\partial_{ij}E = \begin{cases} \boldsymbol{e}_i\boldsymbol{e}_i^{\mathrm{T}} \ , & \text{if } i = j \ , \\ \frac{1}{2}\left(\boldsymbol{e}_i\boldsymbol{e}_j^{\mathrm{T}} + \boldsymbol{e}_j\boldsymbol{e}_i^{\mathrm{T}}\right) \ , & \text{otherwise,} \end{cases}$$

so that, as in Eq. (21), we have

$$\begin{aligned}\theta &= \frac{1}{2}\left[\mathrm{Tr}\left((E - \eta\eta^{\mathrm{T}})^{-1}(\boldsymbol{e}_i\eta^{\mathrm{T}} + \eta\boldsymbol{e}_i^{\mathrm{T}})\right)\right]_i \\ &= \left(E - \eta\eta^{\mathrm{T}}\right)^{-1}\eta \ , \\ \Theta &= -\frac{1}{2}\left[\mathrm{Tr}\left((E - \eta\eta^{\mathrm{T}})^{-1}(\boldsymbol{e}_i\boldsymbol{e}_j^{\mathrm{T}} + \boldsymbol{e}_j\boldsymbol{e}_i^{\mathrm{T}})\right)\right]_{ij} \\ &= -\frac{1}{2}\left(E - \eta\eta^{\mathrm{T}}\right)^{-1} \ .\end{aligned}$$

Next, we take partial derivatives of Eq. (48) and (49). Since $\partial_{kl}\partial_{ij}E = 0$ and $\partial_i\partial_j(\eta\eta^{\mathrm{T}}) = \boldsymbol{e}_i\boldsymbol{e}_j^{\mathrm{T}} + \boldsymbol{e}_j\boldsymbol{e}_i^{\mathrm{T}}$, we have

$$\begin{aligned}\partial_i\partial_j\varphi(\boldsymbol{\eta}) &= \frac{1}{2}\,\mathrm{Tr}\Big((E - \eta\eta^{\mathrm{T}})^{-1}\partial_i\partial_j(\eta\eta^{\mathrm{T}}) + \\ &+ (E - \eta\eta^{\mathrm{T}})^{-1}\partial_i(\eta\eta^{\mathrm{T}})(E - \eta\eta^{\mathrm{T}})^{-1}\partial_j(\eta\eta^{\mathrm{T}})\Big) \\ &= \frac{1}{2}\,\mathrm{Tr}\Big((E - \eta\eta^{\mathrm{T}})^{-1}(\boldsymbol{e}_i\boldsymbol{e}_j^{\mathrm{T}} + \boldsymbol{e}_j\boldsymbol{e}_i^{\mathrm{T}}) + \\ &+ (E - \eta\eta^{\mathrm{T}})^{-1}(\boldsymbol{e}_i\eta^{\mathrm{T}} + \eta\boldsymbol{e}_i^{\mathrm{T}})(E - \eta\eta^{\mathrm{T}})^{-1}(\boldsymbol{e}_j\eta^{\mathrm{T}} + \eta\boldsymbol{e}_j^{\mathrm{T}})\Big) \\ &= \boldsymbol{e}_i^{\mathrm{T}}(E - \eta\eta^{\mathrm{T}})^{-1}\boldsymbol{e}_j + (E - \eta\eta^{\mathrm{T}})^{-1}\eta\eta^{\mathrm{T}}(E - \eta\eta^{\mathrm{T}})^{-1} + \\ &+ (E - \eta\eta^{\mathrm{T}})^{-1}\eta^{\mathrm{T}}(E - \eta\eta^{\mathrm{T}})^{-1}\eta \ ,\end{aligned}$$

which gives the definition of Γ in Eq. (39), (40) and (38).

For $k \neq l$ we have

$$\begin{aligned}\partial_i\partial_{kl}\varphi(\boldsymbol{\eta}) &= -\frac{1}{2}\,\mathrm{Tr}\left((E - \eta\eta^{\mathrm{T}})^{-1}\partial_i(\eta\eta^{\mathrm{T}})(E - \eta\eta^{\mathrm{T}})^{-1}(\partial_{kl}E)\right) \\ &= -\frac{1}{2}\,\mathrm{Tr}\Big((E - \eta\eta^{\mathrm{T}})^{-1}(\boldsymbol{e}_i\eta^{\mathrm{T}} + \eta\boldsymbol{e}_i^{\mathrm{T}})\times \\ &\quad \times (E - \eta\eta^{\mathrm{T}})^{-1}(\boldsymbol{e}_k\boldsymbol{e}_l^{\mathrm{T}} + \boldsymbol{e}_l\boldsymbol{e}_k^{\mathrm{T}})\Big) \\ &= -\boldsymbol{e}_i^{\mathrm{T}}(E - \eta\eta^{\mathrm{T}})^{-1}\boldsymbol{e}_k\boldsymbol{e}_l^{\mathrm{T}}(E - \eta\eta^{\mathrm{T}})^{-1}\eta \\ &\quad - \boldsymbol{e}_i^{\mathrm{T}}(E - \eta\eta^{\mathrm{T}})^{-1}\boldsymbol{e}_l\boldsymbol{e}_k^{\mathrm{T}}(E - \eta\eta^{\mathrm{T}})^{-1}\eta + \\ &= -\boldsymbol{e}_i^{\mathrm{T}}\Sigma^{-1}\boldsymbol{e}_k\boldsymbol{e}_l^{\mathrm{T}}\Sigma^{-1}\eta - \boldsymbol{e}_i^{\mathrm{T}}\Sigma^{-1}\boldsymbol{e}_l\boldsymbol{e}_k^{\mathrm{T}}\Sigma^{-1}\eta \ ,\end{aligned}$$

while for $k = l$

$$\begin{aligned}\partial_i\partial_{kk}\varphi(\boldsymbol{\eta}) &= -\boldsymbol{e}_i^{\mathrm{T}}(E - \eta\eta^{\mathrm{T}})^{-1}\boldsymbol{e}_k\boldsymbol{e}_k^{\mathrm{T}}(E - \eta\eta^{\mathrm{T}})^{-1}\eta \\ &= -\boldsymbol{e}_i^{\mathrm{T}}\Sigma^{-1}\boldsymbol{e}_k\boldsymbol{e}_k^{\mathrm{T}}\Sigma^{-1}\eta \ ,\end{aligned}$$

which gives the definition of K_{kl} in Eq. (41), (42) and (38).

Finally, for $k \neq l \wedge m \neq n$ we have

$$\begin{aligned}\partial_{kl}\partial_{mn}\varphi(\boldsymbol{\eta}) &= \frac{1}{2}\,\mathrm{Tr}\left((E - \eta\eta^{\mathrm{T}})^{-1}(\partial_{kl}E)(E - \eta\eta^{\mathrm{T}})^{-1}(\partial_{mn}E)\right) \\ &= \frac{1}{2}\,\mathrm{Tr}\Big((E - \eta\eta^{\mathrm{T}})^{-1}(\boldsymbol{e}_k\boldsymbol{e}_l^{\mathrm{T}} + \boldsymbol{e}_l\boldsymbol{e}_k^{\mathrm{T}})\times \\ &\quad \times (E - \eta\eta^{\mathrm{T}})^{-1}(\boldsymbol{e}_m\boldsymbol{e}_n^{\mathrm{T}} + \boldsymbol{e}_n\boldsymbol{e}_m^{\mathrm{T}})\Big) \\ &= \boldsymbol{e}_n^{\mathrm{T}}(E - \eta\eta^{\mathrm{T}})^{-1}\boldsymbol{e}_k\boldsymbol{e}_l^{\mathrm{T}}(E - \eta\eta^{\mathrm{T}})^{-1}\boldsymbol{e}_m \\ &\quad + \boldsymbol{e}_m^{\mathrm{T}}(E - \eta\eta^{\mathrm{T}})^{-1}\boldsymbol{e}_k\boldsymbol{e}_l^{\mathrm{T}}(E - \eta\eta^{\mathrm{T}})^{-1}\boldsymbol{e}_m \\ &= \boldsymbol{e}_n^{\mathrm{T}}\Sigma^{-1}\boldsymbol{e}_k\boldsymbol{e}_l^{\mathrm{T}}\Sigma^{-1}\boldsymbol{e}_m + \boldsymbol{e}_m^{\mathrm{T}}\Sigma^{-1}\boldsymbol{e}_k\boldsymbol{e}_l^{\mathrm{T}}\Sigma^{-1}\boldsymbol{e}_m \ .\end{aligned}$$

In the remaining cases the computations are analogous, giving the definition of κ_{klmn} in Eq. (43), (44) and (38).