Natural Gradient for the Gaussian Distribution via Least Squares Regression

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

Abstract

The natural gradient is the Riemannian gradient of a function defined over a statistical model, evaluated with respect to the Fisher information metric. In machine learning and stochastic optimization, the natural gradient is effectively employed in many different contexts, for instance in the optimization of the stochastic relaxation of a function, in policy learning for reinforcement learning, in Bayesian variational inference, and for the training of neural network. In this paper we study the natural gradient for models in the Gaussian distribution, parametrized by a mixed coordinate system, given by the mean vector and the precision matrix, i.e., the inverse covariance matrix. This parameterization allows an efficient estimation of the natural gradient starting from a sample of points, based on the empirical estimation of covariances, in particular in the case of sparse precision matrices. An important contribution of the paper is given by the formalization of the estimation of the natural gradient for the Gaussian distribution as a least squares regression problem. This implies that in the high-dimensional setting it is possible to perform model selection simultaneously to the computation of the natural gradient, using for instance subsect selection techniques from linear regression.

1 Introduction

In this paper we study the problem of optimizing a function defined over a statistical model using gradient descent techniques. Since the geometry of statistical models is not Euclidean [21, 2, 5], the Riemannian gradient has to be evaluated with respect to the geometry of the space, which in the case of statistical model is given by Fisher information metric. Amari [4] was the first to propose the use of the Riemannian gradient over statistical models and called it the natural gradient. In following the natural gradient has been extensively used in many different applications and settings in machine learning and in optimization, and it proved to be more effective in many situations. However the main issue related to the computation of the natural gradient are related to its computational complexity. To partially overcome this problem, the gradient is often estimated from a sample rather than computed exactly, however, this still may be unpractical to compute for large dimensional problems. Let n is the dimension of the problem, and λ the sample size, for models in the exponential family by definition the estimation of the natural gradient requires to solve a linear system of equations, which has a complexity of $\mathcal{O}(n^3 + \lambda n^2)$ by Gauss elimination algorithm, even if better theoretical algorithms are known, up to $\mathcal{O}(n^{2.376\cdots})$. For the special case of models in the multivariate Gaussian family, due to some special properties of this statistical model, the complexity goes down to $\mathcal{O}(\lambda n^2)$ once the model is parametrized by the mean vector and the covariance matrix.

A quadratic complexity on n may become unfeasible for large dimensions, for this reason it becomes of great interest to identify submodels in the Gaussian family, which allow the computation of the

natural gradient linearly in n and that at the same time are enough expressive to represent correlations among the random variables of the statistical model. For the Gaussian distribution, a common approach consists in the choice of block-diagonal covariance matrices, which translated into block-diagonal Fisher information matrices, which become more tractable during the inversion.

In this paper we introduce a mixed parametrization for the Gaussian distribution based on the mean vector and inverse covariance matrix, which allow to represent sub-models in the Gaussian distribution as exponential families. We derive formulae for the estimation of the natural gradient based on least-squares (ridge) regression for the mixed parameterization, which scale linearly with the number of interactions of the model, i.e., the number of non-zero components of the precision matrix, and quadratically with the sample size, provided the sample size is smaller than the number of such components, and with the cube of the sample size otherwise. We conclude the paper by mentioning some examples of applications in machine learning and stochastic optimization, where the parameterization and the formulae for the estimation of the natural gradient could be directly employed.

2 Mixed Parameterization for the Gaussian Distribution

The most common parameterization for the multivariate Gaussian distribution is given by the mean vector $\boldsymbol{\mu}$ and the covariance matrix Σ, however other parameterizations are possible. In information geometry, at least two other sets of parameters play an important role: the natural parameterization of the exponential family $\boldsymbol{\theta} = (\theta; \Theta)$, to which the Gaussian distribution belongs, and the expectation parameterization $\boldsymbol{\eta} = (\eta; E)$, e.g., [23], cf. [15]. The Fisher information matrix for the $(\boldsymbol{\mu}; \Sigma)$ parameters is block-diagonal, however the biggest advantage of this parameterization is the fact that the natural gradient can be computed directly without any matrix multiplication. On the other side, when it comes to the identification of lower dimensional models in the Gaussian distribution, one possibility it to impose independence among couples of variables, i.e., setting $\sigma_{ij} = 0$ in Σ. This results in a sparse covariance matrix, however unless Σ becomes block-diagonal, such sub-models do not belong to the exponential family, and thus the computation of the natural gradient would required the inversion of the Fisher information matrix. The parameterization based on the expectation parameters, with $\eta_i = \mu_i = \mathbb{E}_p[X_i]$ and $\eta_{ij} = \mathbb{E}_p[X_i X_j] = \sigma_{ij} + \mu_i \mu_j$ determines a Fisher information matrix which is not block-diagonal unless $\boldsymbol{\mu} = \mathbf{0}$, however, the estimation of the natural gradient can be computed still in $\mathcal{O}(\lambda n^2)$. However, unless Σ becomes block-diagonal, the independence among variables given by $\eta_{ij} = \eta_i \eta_j$ does not identify an exponential sub-models, compromising the duality between expectation and natural parameters which provides a formula for the natural gradient which does not require the inversion of the Fisher information matrix. On the other hand, the natural parameterization of the Gaussian distribution is based on the inverse covariance matrix and has two important properties: a zero in the precision matrix identify an exponential sub-models, and at the same time it implies conditional independence among variables. However, unless the distribution is centered, the Fisher information matrix is not block-diagonal, and moreover the Fisher information matrix cannot be estimated from a sample, unless the Fisher information matrix is inverted. This determines a complexity of $\mathcal{O}(n^3 + \lambda n^2)$, which can be further reduced to $\mathcal{O}(\lambda d + w^3 n)$ in case of chordal Gaussian graphical models [4, 9], for which the joint probability distribution can be factorized according to a junction tree, where w denotes the size of the maximum clique and d the number of non zero entries in the precision matrix.

In the this section we introduce a mixed parameterization for the Gaussian distribution, based on (η, Θ). Expectation parameters and natural parameters are dually coupled in the sense of the Legendre transform [5], and any parameterization determines a block diagonal Fisher information matrix, as for the $(\boldsymbol{\mu}, \Sigma)$. Consider a vector $\boldsymbol{x} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)^{\mathrm{T}} \in \mathbb{R}^n = \Omega$. Let $\eta = \boldsymbol{\mu} \in \mathbb{R}^n$ be the mean vector and $\Theta = \frac{1}{2}\Sigma^{-1} = [\sigma_{ij}]$ be $n \times n$ symmetric positive-definite inverse covariance matrix, the multivariate Gaussian distribution can be written as

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

One of the advantages of standard parameterization $(\boldsymbol{\mu}; \Sigma)$ compared to the other two is given by the fact that the Fisher information matrix becomes block-diagonal, and thus easier to invert. On the other side, the $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ parameters are dually coupled, and the Fisher information matrices $I_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ and $I_{\boldsymbol{\eta}}(\boldsymbol{\eta})$ are one the inverse of the other for corresponding $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$.

3 Estimation of the Natural Gradient

In this section we present the main result of the paper, which consists of an alternative formula for the estimation of the natural gradient of the expected value of a function with respect to a Gaussian distribution parametrized by $(\eta; \Theta)$. The theorem is an adaptation to the Gaussian case of an analogus result for discrete exponential families [14]. See also [22] for a related result in the context of the training of neural networks. In the following we show that the natural gradient can be estimated by means of least-squares (ridge) regression, with a computational complexity of $\mathcal{O}(\lambda^3 + \lambda^2 n^2)$. For the full Gaussian distribution, with no restrictions on the covariance matrix, this result compares favorably with respect to the $\mathcal{O}(\lambda n^3)$ complexity of the $(\theta; \Theta)$ parameterization [4, 9], however it is worse than the $\mathcal{O}(\lambda n^2)$ complexity for the $(\boldsymbol{\mu}; \Sigma)$ parameterization [1]. Nevertheless, the advantage of our parameterization becomes evident when dealing with Gaussian sub-models with sparse precision matrices, where the computational complexity goes down to $\mathcal{O}(d^3 + \lambda d^2)$, and to $\mathcal{O}(\lambda^3 + \lambda^2 d)$ for $\lambda < d$, where $d \leq n^2$ is the number of non zero entries in the precision matrix.

Theorem 1 *Let p be a Gaussian distribution parametrized by (η, Θ) and let $\mathcal{P}$ be an i. i. d. sample from p of cardinality λ. We denote by $A = [\overline{T}_{ij}(\boldsymbol{x})]$, with $\overline{T}_{ij}(\boldsymbol{x}) = X_i X_j - \mathbb{E}_{\boldsymbol{\theta}}[X_i X_j]$, $i, j = 1, \ldots, n, i \leq j$ and $\boldsymbol{x} \in \mathcal{P}$, the design matrix of the centered second-order sufficient statistics, and with $\boldsymbol{y} = (f(\boldsymbol{x}))$ the column vector associated to the evaluations of f.*

1. *The natural gradient $\widetilde{\nabla}\mathbb{E}_\eta[f]$ can be estimated by $\left(\widehat{\mathrm{Cov}}(X_i, f)\right)$.*
2. *The least squares estimator $\widehat{\boldsymbol{c}}$ given $\mathcal{P}$ of the regression model $f(\boldsymbol{x}) = \sum_{i \leq j} c_{ij} T_{ij}(\boldsymbol{x}) + \boldsymbol{\epsilon}$ converges to the natural gradient $\widetilde{\nabla}\mathbb{E}_\Theta[f]$, as $\lambda \to \infty$.*
3. *The natural gradient $\widetilde{\nabla}\mathbb{E}_\Theta[f]$ of the Gaussian distribution can be approximated by least squares ridge regression by $\widehat{\boldsymbol{c}}_{ridge} = (A^\top A + \mathbf{1})^{-1} A^\top \boldsymbol{y}$. The regression problem can be equivalently solved in the dual variables by $\widehat{\boldsymbol{c}}_{ridge} = A^\top (A A^\top + \mathbf{1})^{-1} \boldsymbol{y}$, with a computational complexity of $\mathcal{O}(\lambda^3 + \lambda^2 n^2)$.*
4. *The natural gradient for a Gaussian model with sparse precision matrix, where $d \leq n^2$ is the number of non zero entries, can be estimated in $\mathcal{O}(d^3 + \lambda d^2)$, or equivalently, for $\lambda < d$, in $\mathcal{O}(\lambda^3 + \lambda^2 d)$.*
5. *The previous results can be combined with specific algorithms for chordal Gaussian graphical models [4, 9], resulting in a computational complexity for the estimation of the natural gradient of $\mathcal{O}(n(\lambda^3 + \lambda^2 w^2))$ if the algorithm is run in the dual variables.*

Proof. The Fisher information matrix in the mixed $(\eta; \Theta)$ parameterization is block-diagonal, thus the natural gradient for η and Θ is evaluated independently. For $\widetilde{\nabla}\mathbb{E}_\eta[f]$, by duality we have $I_\eta(\eta)^{-1} \nabla \mathbb{E}_\eta[f] = \nabla \mathbb{E}_\theta[f] = \left(\widehat{\mathrm{Cov}}(f, X_i)\right)$, see [5, 18, 16]. As to $\widetilde{\nabla}\mathbb{E}_\Theta[f]$, the least squares estimator is

$$\begin{aligned}
\widehat{\boldsymbol{c}} = (A^\top A)^{-1} A^\top \boldsymbol{y} &= \left[\sum_{\boldsymbol{x} \in \mathcal{P}} \overline{T}_{ij}(\boldsymbol{x}) \overline{T}_{kl}(\boldsymbol{x})\right]^{-1} \left(\sum_{\boldsymbol{x} \in \mathcal{P}} f(\boldsymbol{x}) \overline{T}_{ij}(\boldsymbol{x})\right) \\
&= \left[\widehat{\mathrm{Cov}}(T_{ij}(\boldsymbol{x}), T_{kl}(\boldsymbol{x}))\right]^{-1} \left(\widehat{\mathrm{Cov}}(f, T_{ij}(\boldsymbol{x}))\right) ,
\end{aligned}$$

which corresponds to the formula for the natural gradient in terms of covariances, e.g., [6, 13].

We conclude this section with some remarks about possible approaches to the problem of choosing a Gaussian sub-model in the optimization, identified by a sparse covariance matrix, when this is not known in advance. The problem of employing a good model for the optimization of $\mathbb{E}[f]$ is crucial, indeed the choice of the model determines the landscape of the function to be optimized, and in particular the number of local minima. Lower-dimensional models require a smaller number of parameters to be estimated and thus may be more efficient, however they may determine premature convergence to local minima. A two stage approach to the problem would consists of a first step when the sparse precision matrix is estimated from a sample of point, together with the evaluations

of f, followed by an estimation of the natural gradient. By formalizing the estimation of the natural gradient as a linear regression problem, the two steps can be solved simultaneously. A naive implementation of this approach could be based on a subsect selection strategy, such as a forward stepwise selection or least angle regression [7]. By adding new variables to the regression function, the dimension of the Gaussian sub-model is increased, and at the same time the number of components of the natural gradient to be estimated.

4 Applications in Machine Learning and Optimization

In this section we focus on three different applications in machine learning and stochastic optimization, where natural gradient is employed for the optimization of a function defined over a statistical model: the optimization of a real-valued function defined over $\mathbb{R}^n$ by its stochastic relaxation; policy learning in reinforcement learning; and finally the training of a neural network by stochastic natural gradient descent. In these contexts, where often the statistical model used in the optimization is a Gaussian model, as the number of variables increases, the computations of the natural gradient becomes computationally intractable, due to a complexity which is at least quadratic in n. For this reason, often strong assumptions of independence among variables are made, so that the Fisher information matrix becomes easier to invert. The parameterization we proposed in the previous section can be an alternative approach to the identification of sub-models in the Gaussian distribution for which the natural gradient can be computed efficiently.

The first example we consider is the optimization of a real-valued function $f : \Omega \to \mathbb{R}$ by means of its Stochastic Relaxation (SR) [12, 13], cf. [11], i.e., we search for the optimum of f in Ω by optimizing the expected value of the function itself over a statistical model $\mathcal{M}$. This approach in optimization is quite general and can be considered as a unifying perspective in many different fields. We denote with $F(p) : \mathcal{M} \to \mathbb{R} : p \mapsto \mathbb{E}_p[f]$ the Stochastic Relaxation (SR) of f, i.e., the expected value of f with respect to p in $\mathcal{M}$. We look for the minimum of f by optimizing the stochastic relaxation, which under some regularity conditions over the choice of $\mathcal{M}$ is a continuous function independently from the nature of Ω, either discrete or continuous. Solutions to the original problem can then be obtained by sampling from optimal solutions of the relaxed function F. The search for the optimum of F can be performed using different algorithms and techniques. For instance, CMA-ES [8, 1] and NES [24] are two stochastic algorithms based on natural gradient descent, i.e., they update the parameters of a Gaussian distribution in the direction given by the natural gradient.

The second application we consider is given by reinforcement learning, and more in particular direct policy-gradient methods for approximate planning in Markov Decision Problems (MDP), see the seminar paper by Kakade [10] and later the work in [20], among the others. At each time step t, the agent in state s_t chooses an action a_t according to th stochastic policy $p(a_t|s_t)$, then it gets a reward $r(s_t, a_t)$ from the environment, which is used by the agent as a feedback to update the policy. The goal of the agent is to learn the policy that maximizes the average reward over a family of parametrized policies $p(a|s;\boldsymbol{\theta})$, which under some regularity conditions corresponds to a statistical manifold $\mathcal{M}$. The problem can be solved by gradient descent, by taking the gradient of the average discounted reward of the agent and updating the parameters of the distribution accordingly.

One of the earlier applications of the work of Amari on natural gradient was in the context of the training of neural networks [3, 4]. More recently, natural gradient has gathered renewed attention in the context of neural networks, and information geometry proved to be a general framework to describe different training algorithms recently proposed in the literature, see [19, 17]. Consider a multilayer feedforward neural network, specified by a vector of parameters $\boldsymbol{w} = (w_1, \ldots, w_n)^\top \in \mathbb{R}^n$, which represent the connection weights and thresholds of the network. Given an input $\boldsymbol{x}$ to the neural network, an output $\boldsymbol{y} = f(\boldsymbol{x};\boldsymbol{w})$ is produced. From an information geometric perspective, Amari showed that the neural network can be represented as a manifold of distributions $\mathcal{M}$ parameterized by $\boldsymbol{w}$, where the loss function $F : \mathcal{M} \to \mathbb{R} : \boldsymbol{w} \mapsto l(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w})$ associates to each distribution the loss of a given observation. The estimating the weights of the network can then be performed by iteratively minimizing the loss function over the samples in the training set. In an online learning setting, at each time step t the network observes a training point $(\boldsymbol{x}_t, \boldsymbol{y}_y)$. The weights can be updated according to a gradient descent update $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \lambda \widetilde{\nabla} l(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w})$, with step size $\lambda > 0$, where at each iteration we are evaluating the natural gradient of the loss as a function of the weights, given the current training sample $(\boldsymbol{x}_t, \boldsymbol{y}_t)$.

References

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

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

[3] S. Amari. Neural learning in structured parameter spaces - natural Riemannian gradient. In *Advances in Neural Information Processing Systems*, pages 127–133. MIT Press, 1997.

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

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

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

[7] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

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

[9] J. K. Johnson, V. Chandrasekaran, and A. S. Willsky. Learning markov structure by maximum entropy relaxation. In M. Meila and X. Shen, editors, *AISTATS*, volume 2 of *JMLR Proceedings*, pages 203–210. JMLR.org, 2007.

[10] S. Kakade. A natural policy gradient. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14 (NIPS 2001)*, pages 1531–1538. MIT Press, 2001.

[11] J. B. Lasserre. Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization*, 11:796–817, 2001.

[12] 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.

[13] 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.

[14] 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.

[15] L. Malagò and G. Pistone. Information geometry of the gaussian distribution in view of stochastic optimization. In *Proc. of FOGA '15*, pages 150–162, 2015.

[16] L. Malagò and G. Pistone. Natural gradient flow in the mixture geometry of a discrete exponential family. *Entropy*, 17(6):4215, 2015.

[17] J. Martens. New perspectives on the natural gradient method, 2014v1-2015v2.

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

[19] R. Pascanu and Y. Bengio. Revisiting natural gradient for deep networks. In *Proceedings of the International Conference on Learning Representations (ICLR) 2014*, 2014.

[20] J. Peters and S. Schaal. Natural actor critic. *Neurocomputing*, 71(7–9):1180–1190, 2008. Progress in Modeling, Theory, and Application of Computational Intelligence 15th European Symposium on Artificial Neural Networks 2007.

[21] C. Radhakrishna Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.*, 37:81–91, 1945.

[22] N. L. Roux, P.-A. Manzagol, and Y. Bengio. Topmoumoute online natural gradient algorithm. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 849–856. Curran Associates, Inc., 2008.

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

[24] 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.