Robust Estimation of Natural Gradient in Optimization by Regularized Linear Regression

Luigi Malagò[1] and Matteo Matteucci[2]

[1] Università degli Studi di Milano, Dept. of Computer Science,
20135 Milan, Italy, malago@di.unimi.it,
[2] Politecnico di Milano, Dept. of Electronics Information and Bioengineering,
20133 Milan, Italy, matteo.matteucci@polimi.it

Abstract. We are interested in the optimization of the expected value of a function by following a steepest descent policy over a statistical model. Such approach appears in many different model-based search meta-heuristics for optimization, for instance in the large class of random search methods in stochastic optimization and Evolutionary Computation. We study the case when statistical models belong to the exponential family and the direction of maximum decrement of the expected value is given by the natural gradient evaluated with respect to the Fisher Information metric. When the gradient cannot be computed exactly, a robust estimation allows to minimize the number of function evaluations required to obtain convergence to the global optimum. Under the choice of centered sufficient statistics, the estimation of the natural gradient corresponds to solving a least squares regression problem for the original function to be optimized. The correspondence between the estimation of the natural gradient and solving a linear regression problem leads to the definition of regularized versions of the natural gradient. We propose a robust estimation of the natural gradient for the exponential family based on regularized least squares.

Keywords: information geometry, regularized natural gradient, stochastic gradient descent, regularized least squares, ridge regression, lasso.

1 Introduction

In this paper we focus on the robust estimation of the natural gradient in optimization and in particular in the context of model-based search, a large class of meta-heuristics and optimization algorithms where the search for the optimum takes place in the space of probability distributions. In model-based search a minimizing sequence of probability distributions is generated so that probability density gets concentrated in regions of the search space that with higher probability include the optimum of the function to be optimized.

A common unifying perspective for model-based search consists in replacing the original optimization problem of minimizing a function $f : \Omega \to \mathbb{R}$ with the optimization of the expected value of the original function $\mathbb{E}_p[f]$, with respect to

some p is a statistical model $\mathcal{M}$. The new variables of the relaxed problem are the parameters of the statistical model, i.e., a set of probability distributions.

A minimizing sequences for $\mathbb{E}_p[f]$ can be generated in different ways, for instance by iteratively sampling a probability distribution, followed by a selection of a sub sample based on the value of the function, and finally estimating the parameters of a new distribution, as in Estimation of Distribution Algorithms (EDAs) [8], a broad family of black-box optimization algorithms in Evolutionary Computation. On the other hand, gradient descent is probably one of the simplest and best known methods in optimization, with a rich history that goes back to Cauchy. The basic idea is that of searching for the optimum iteratively, by updating the value of the variables with a step in the direction of the gradient of the function, that in this context corresponds to the expected value $\mathbb{E}_p[f]$. In model-based search, in order to efficiently solve the new optimization problem, the search is usually restricted to a lower dimensional statistical model. Notice that the choice of the model strongly determines the presence of local minima, since if the statistical model does not capture all the relevant interactions among the variables of f, there may be points in $\mathcal{M}$ where the gradient vanishes, so that local minima may appear, cf. [10].

In the last decade, the natural gradient has been applied successfully in different fields, from machine learning to signal processing. In optimization, and in particular in Evolutionary Computation, Natural Evolution Strategies (NES) [18] are one of the first examples of a framework based on the natural gradient for the optimization of continuous functions based on multivariate Gaussian distributions. In the more general case of statistical models that belong to the exponential family, we refer to the geometric framework based on Stochastic Relaxation first presented in [10], where the authors introduced Stochastic Natural Gradient Descent (SNGD), for the optimization of functions defined over binary variables. Another related work appears in [2], where a similar framework named Information-Geometric Optimization (IGO) is presented.

2 Stochastic Relaxation based on the Exponential Family

We are interested in the optimization of a real-valued function $f : \Omega \to \mathbb{R}$, and according to the framework of Stochastic Relaxation [10], we replace the original optimization problem with the minimization of $\mathbb{E}_p[f] : \mathcal{M} \to \mathbb{R}$, where $\mathcal{M}$ is a lower dimensional statistical model. It is easy to verify that under some general assumptions on the choice of $\mathcal{M}$ the two problems are equivalent.

In the following, let $x = (x_1, \ldots, x_n) \in \Omega$ be a real vector. We choose models from the exponential family $\mathcal{E}$ [3] of probability distributions

$$p(x;\theta) = \exp\left\{\sum_{i=1}^{m} \theta_i T_i(x) - \psi(\theta)\right\}, \quad \theta_i \in \mathbb{R}, \tag{1}$$

where $\psi(\theta) = \ln \sum_{\Omega} \exp\left\{\sum_{i=1}^{m} \theta_i T_i(x)\right\}$ is the normalizing factor, $\{T_i(x)\}_{i=1}^{m}$ are the sufficient statistics, which we suppose to be linear independent, and θ is

the vector of *natural parameters*. The exponential family includes a large number of models, both in the discrete and continuous case, such as Markov Random Fields, and multivariate Gaussian distributions.

For the exponential family, the *natural gradient* of $\mathbb{E}_\theta[f]$, i.e., the gradient evaluated with respect to the Fisher information metric $I(\theta)$, is defined as

$$\widetilde{\nabla}_\theta \mathbb{E}_\theta[f] = I(\theta)^{-1} \nabla_\theta \mathbb{E}_\theta[f], \tag{2}$$

where $I(\theta) = [\partial_i \partial_j \psi(\theta)]_{i,j=1}^m$, $\nabla_\theta \mathbb{E}_\theta[f] = (\partial_i \mathbb{E}_\theta[f])_{i=1}^m$, and ∂_i represents the partial derivative with respect to θ_i. We denote the natural gradient with $\widetilde{\nabla}$ to distinguish it from the regular vanilla gradient ∇.

Given an exponential family $\mathcal{E}$ parametrized by θ, the updating rule according to the natural gradient of $\mathbb{E}_\theta[f]$ reads

$$\theta^{t+1} = \theta^t - \lambda \widetilde{\nabla}_{\theta^t} \mathbb{E}_{\theta^t}[f],$$

where λ is the learning rate that controls the step size in the direction of the gradient. The natural gradient, introduced by Amari [1] has been proved to be efficient in many different learning task where the search space is given by a set of probability distributions. The natural gradient reflects the intrinsic geometry of the manifold of probability distributions and thus benefits of some remarkable properties. It has better convergence properties compared to the regular gradient, moreover it is parametric invariant, i.e., it does not depend on the choice of the specific parameterization.

The exact evaluation of $\widetilde{\nabla} \mathbb{E}_\theta[f]$ is often computationally intractable for large n, unless we restrict $\mathcal{M}$ to belong to a restricted class of models. In the general case of an exponential family, a common approach consists in replacing exact gradients with an estimation based on a sample. For the exponential family $\mathcal{E}$, we have

$$\nabla_\theta \mathbb{E}_\theta[f] = \big(\operatorname{Cov}(f, T_i)\big)_{i=1}^m, \qquad I(\theta) = \big[\operatorname{Cov}(T_i, T_j)\big]_{i,j=1}^m.$$

Given a sample of observations i. i. d. with respect to θ, we can replace the exact evaluation of natural gradient of $\mathbb{E}_\theta[f]$ with an estimation based on empirical covariances. This leads to the SNGD algorithm, described in [10, 9].

3 Natural Gradient and Linear Regression

There is a strong relationship between the estimation of the natural gradient and least squares regression, indeed the natural gradient can be evaluated as the least squares projection of the direction of maximum decrement of $\mathbb{E}_\theta[f]$ onto a tangent space of the statistical model, cf [10]. In the following we state a result, first presented in [11], that creates a relationship between the estimation of the natural gradient and the least squares estimator of a regression problem. Given an exponential family $\mathcal{E}$, with centered sufficient statistics $\{T_i\}$ in p_θ, we show that the least squares estimator of a regression model for f, with respect to the $\{T_i\}$ variables, corresponds for large N to the evaluation of $\widetilde{\nabla}_\theta \mathbb{E}_\theta[f]$ with $p_\theta \in \mathcal{E}$.

Theorem 1 *If the sufficient statistics* $\{T_i\}$ *of* $p(x;\theta) \in \mathcal{E}$ *are centered in* θ*, i.e.,* $\mathbb{E}_\theta[T_i] = 0$*, then the least squares estimator* $\widehat{c}$ *with respect to an i. i. d. sample* $\mathcal{P}$ *from* p *of the linear model* $f(x) = \sum_{i=1}^m c_i T_i(x)$ *converges to the natural gradient* $\widetilde{\nabla}_\theta \mathbb{E}_\theta[f]$*, as* $N \to \infty$*. Similarly,* $\widehat{I}(\theta)^{-1}\widehat{\nabla}\mathbb{E}[f] \to c$ *as* $N \to \infty$*.*

Proof. For a proof of this result see Theorem 1 in [11].

Here we supposed the sufficient statistics of $\mathcal{E}$ to be centered. Notice that this is a general hypothesis, since it is always possible to center them by letting $\bar{T}_i = T_i - \mathbb{E}_\theta[T_i]$. If the sufficient statistics of the exponential family are not only centered by also orthogonal, we can define the *orthogonal estimator* of the regression coefficients under the hypothesis of orthogonal and centered variables, which reads

$$\widetilde{\nabla}^{\perp}_\theta \mathbb{E}_\theta[f] = \left(\frac{\widehat{\mathbb{E}}[f T_i]}{\widehat{\mathbb{E}}[T_i T_i]} \right)_{i=1}^m = \left(\frac{\widehat{\mathrm{Cov}}(f, T_i)}{\widehat{\mathrm{Cov}}(T_i, T_i)} \right)_{i=1}^m ,$$

where $\widehat{\mathbb{E}}[\cdot]$ and $\widehat{\mathrm{Cov}}(\cdot,\cdot)$ represent empirical means and covariances.

In case of binary variables, with $x_i \in \{+1, -1\}$, and sufficient statistics that takes the form of monomials, i.e., $T_i(x) = x^\alpha$, with $\alpha = (\alpha_1, \ldots, \alpha_n) \in \{0,1\}^n$ and $x^\alpha = \prod_{i=1}^n x_i^{\alpha_i}$, the orthogonal least squares estimator converges to the regular gradient for $N \to \infty$, since $\widehat{\mathbb{E}}[X^\alpha X^\alpha] = 1$. Moreover, if the sufficient statistics are orthogonal and centered, $\widehat{\nabla}_\theta \mathbb{E}_\theta[f]$ and $\widetilde{\nabla}^{\perp}_\theta \mathbb{E}_\theta[f]$ admit the same limit equal to the natural gradient $\widetilde{\nabla}_\theta \mathbb{E}_\theta[f]$, as $N \to \infty$. This provides a different perspective on the estimation of the natural gradient, which is known to be more robust than that of the regular gradient. Notice that in the latter case no assumption on orthogonality is made.

4 Robust Estimation of the Natural Gradient

As a consequence of Theorem 1 we can introduce robust techniques in the estimation of the natural gradient, with methods that comes from linear regression. In the following, we present different estimators of the natural gradient, based on the introduction of penalizing terms, on the shrinkage approach to covariance matrix estimation, and on variable selection methods.

A common method to obtain a robust estimator in linear regression consists in introducing a penalizing term in the least squares regression formula. In ridge regression, a ℓ_2-norm regularized term is introduced in the minimization of the residual sum of squares. This leads to the *ridge regression estimator* of the natural gradient, given by the closed-form solution

$$\widetilde{\nabla}^{\mathrm{rid}}_\theta \mathbb{E}_\theta[f] = \left[\widehat{\mathrm{Cov}}(T_i, T_j) + \lambda^{\mathrm{rid}} \mathbb{I} \right]^{-1} \left(\widehat{\mathrm{Cov}}(f, T_i) \right)^\top ,$$

where $\lambda^{\mathrm{rid}} \geq 0$ is the regularizing parameter, and $\mathbb{I}$ is a $m \times m$ unit matrix. Such regularization produces a shrinkage effect on the components of the gradient,

by reducing the mean square error of the estimator, at the cost of an extra bias. Moreover, the regularization term makes the matrix always invertible, in particular it makes the evaluation of the gradient more robust when a limited number of points is available compared to the number of sufficient statistics. The estimator of the natural gradient obtained with ridge regression leads to the definition of the regularized natural gradient evaluated with respect to the regularized Fisher information matrix $[I(\theta) + \lambda^{\text{rid}}\mathbb{I}]$, cf. [6].

The penalizing term added to the covariance matrix by ridge regression is similar to the *shrinkage estimator* of the covariance matrix, discussed in [13] for the $N \ll m$ setting,

$$\widetilde{\nabla}_\theta^{\text{shr}}\mathbb{E}_\theta[f] = \left[(1-\lambda^{\text{shr}})\widehat{\text{Cov}}(T_i, T_j) + \lambda^{\text{shr}}\widehat{T}(\theta)\right]^{-1}\left(\widehat{\text{Cov}}(f, T_i)\right)^\top,$$

with $\lambda^{\text{shr}} \in [0,1]$, where $\widehat{T}$ is estimator of the covariance target $T(\theta)$, usually a low dimensional counterpart of $I(\theta)$. Such estimator has been recently applied in model-based search to the estimation of the covariance matrix of multivariate Gaussian distributions, see [7].

Other types of penalizing functions can be used. Another popular choice is the ℓ_1-norm, as in the lasso algorithm for linear regression [16]. In this case the *lasso estimator* is obtained by solving the following minimization problem,

$$\widetilde{\nabla}_\theta^{\text{las}}\mathbb{E}_\theta[f] = \arg\min_{\theta\in\mathbb{R}^m}\left\{\sum_{j=1}^{N}\left(\sum_{i=1}^{m} T_i(x^j)\theta_i - f(x^j)\right)^2 + \lambda^{\text{las}}\sum_{i=1}^{m}|\theta_i|\right\},$$

with $\lambda^{\text{las}} > 0$, where $x^1, \ldots, x^N \in \Omega$ represent the current sample. This estimator does not have a closed-form solution, however an efficient iterative implementation is available, given by the LARS algorithm, see [4]. This estimator as been previously employed for the fitness modeling in sDEUM [17] a model-based search algorithm in the DEUM framework [15, 14].

The lasso estimator of the natural gradient combines both shrinkage and soft thresholding effects, i.e., some coefficients of the estimator are set to zero. This behavior is particularly desired for model selection in a black-box setting, when the analytic formula of the function to be optimized is unknown and a model must be learnt from the sample. From this perspective, model selection can be solved by linear regression, by choosing a set of sufficient statistics $\{T_i\}$ of $\mathcal{E}$ during the estimation of the gradient. Remember that the choice of the model is critical in model-based search, since ignoring strong interactions among the variables may determine the presence of local minima for $\mathbb{E}_p[f]$. This approach, based on the correspondence of the estimation of the natural gradient and the solution of a regression problem by least squares, allows to simultaneously solve model selection and estimation of the natural gradient in steepest descent black-box model-based search.

Another approach to simultaneously perform model selection and estimate the natural gradient, which still comes from linear regression, is given by the use of subset selection methods, cf. [5, Ch. 3]. In particular, since the number

of candidate sufficient statistics to obtain a basis for f can be infinite in the continuos case and exponential in n in the discrete case, some greedy policy is required. For instance, in forward subset selection variables are added one at a time to the regression function, until the correlation between the residual and new candidate variables is lower then a given threshold.

Finally, notice that when the number of variables is larger than the number of points, i.e., $N \ll n$, for instance when we are performing model selection and extra candidate variables enter in the regression model, it may be efficient to solve the regression problem in dual form, cf. [12]. Due to Theorem 1 we have an alternative formula for the least squares estimator of $\nabla_\theta \mathbb{E}_\theta[f]$ given by

$$\widetilde{\nabla}^{\text{ls}}_\theta \mathbb{E}_\theta[f] = A^\top (AA^\top)^{-1} f(x)^\top,$$

where A is the design matrix of the observations, which is more efficient in terms of memory usage when $N \ll n$. Moreover, the dual formulation provides a setting which allows the use of kernel methods in the estimation of the natural gradient.

5 Conclusions

The use of a regularized estimator of the natural gradient for model-based search allows to reduce the sample size used in the estimation of the gradient, and yet be able to find the optimum of the function with a lower number of function evaluations. In particular it allows to obtain more robust estimations of the gradient in presence of noise, and in the case when the regression model includes more correlations than those encoded in the function to be optimized.

References

1. S. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.
2. L. Arnold, A. Auger, N. Hansen, and Y. Ollivier. Information-geometric optimization algorithms: A unifying picture via invariance principles. arXiv:1106.3708, 2011.
3. 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.
4. B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.
5. T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, 2001.
6. C. Igel, M. Toussaint, and W. Weishui. Rprop using the natural gradient. In *Trends and Applications in Constructive Approximation*, number 151, pages 259–272. Birkhuser Verlag, 2005.
7. H. Karshenas, R. Santana, C. Bielza, and P. Larrañaga. Regularized continuous estimation of distribution algorithms. *Applied Soft Computing*, (0):–, 2012.
8. P. Larrañaga and J. A. Lozano, editors. *Estimation of Distribution Algoritms. A New Tool for evolutionary Computation*. Springer, 2001.

9. L. Malagò, M. Matteucci, and G. Pistone. Stochastic natural gradient descent by estimation of empirical covariances. In *Proc. of IEEE CEC 2011*, pages 949–956, 2011.
10. 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.
11. L. Malagò, M. Matteucci, and G. Pistone. Natural gradient, fitness modelling and model selection: A unifying perspective. In *Proc. of IEEE CEC 2013*, 2013.
12. C. Saunders, A. Gammerman, and V. Vovk. Ridge regression learning algorithm in dual variables. In *In Proceedings of the 15th International Conference on Machine Learning*, pages 515–521. Morgan Kaufmann, 1998.
13. J. Schäfer and K. Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4(1), 2005.
14. S. Shakya and J. McCall. Optimization by Estimation of Distribution with DEUM framework based on Markov random fields. *International Journal of Automation and Computing*, 4(3):262–272, 2007.
15. S. Shakya, J. McCall, and D. Brown. Updating the probability vector using MRF technique for a Univariate EDA. In *Proc. of STAIRS 2004*, pages 15–25. IOS Press, 2004.
16. R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.
17. G. Valentini, L. Malagò, and M. Matteucci. Optimization by 1-constrained markov fitness modelling. In Y. Hamadi and M. Schoenauer, editors, *Learning and Intelligent Optimization*, Lecture Notes in Computer Science, pages 250–264. Springer Berlin Heidelberg, 2012.
18. D. Wierstra, T. Schaul, J. Peters, and J. Schmidhuber. Natural evolution strategies. In *Proc. of IEEE CEC 2008*, pages 3381–3387, 2008.