Second-order Information Geometry on a Parametric Exponential Family

Luigi Malagò[1] and Giovanni Pistone[2]
[1] Romanian Institute of Science and Technology, [2] Collegio Carlo Alberto
e-mail: malago@rist.ro; giovanni.pistone@carloalberto.org

Information Geometry is an interdisciplinary and expanding research field at the intersection of statistics and differential geometry, which studies the geometry of statistical models, represented as manifolds of probability distributions. Notably, Information Geometry provides a principled framework for the analysis and design of natural Riemannian gradient descent algorithms for the optimization of functions defined over statistical models, with applications in machine learning, statistical inference, information theory, stochastic optimization, and several fields in computer science, such as robotics and computer vision.

The task of optimizing a function whose variables are the parameters of a statistical model is widespread in data science, think for example to the optimization of the expected value of a function with respect to a distribution in a statistical model, the maximization of the likelihood, or more in general the minimization of a loss function. Whenever the closed formula for the solution of the problem is unknown, gradient descent methods constitute a classical approach to optimization. However, it is a well-known result in statistics that the geometry of a statistical model is not Euclidean, instead the unique metric which is invariant to reparameterization is the Fisher information metric. It follows that the direction of maximum decrement of a function over a statistical model is given by the Riemannian natural gradient, first proposed by Amari. Despite the directness of first-order methods, there are situations where taking into account the information on the Hessian of the function to be optimized gives an advantage, for instance for ill-conditions problems for which gradient methods may converge too slowly. Similarly to the natural gradient, also the definition of the Hessian of a function depends on the metric, so that second-order methods over statistical manifolds need to be generalized to the Riemannian geometry of the search space.

When we move to the second-order geometry of a differentiable manifold, the notion of covariant derivative is required for the parallel transport between tangent spaces, in particular to compute directional derivatives of vector fields over a manifold. However, an important result in Information Geometry affirms that exponential families, and more in general Hessian manifolds, have a dually-flat nature, which implies the existence of at least two other relevant

geometries for statistical models: the mixture and the exponential geometries. Differently from the Riemannian geometry, the exponential and mixture geometries are independent from the notion of metric, and they are defined by two dual affine connections, the mixture and the exponential connections. The dual connections, which are equivalently specified by the dual covariant derivatives, allow to define dual parallel transports, dual geodetics, and ultimately the exponential and mixture Hessians. What is specific of Hessian manifolds, is that the combination of dual Hessians and geodetics allows to define alternative second-order Taylor approximations of a function, without the explicit computation of the Riemannian Hessian and the Riemannian geodetic, which are computationally expensive operations in general. Compared to Riemannian manifolds, dually-flat manifolds have a richer geometry that can be exploited in the design of more sophistical second-order optimization algorithms.

Second-order methods, such as the Newton method, conjugate gradient, and trust region methods, are popular algorithms in mathematical optimization, known for their super-linear convergence rates. The application of such methods to the optimization over statistical manifolds using second-order Riemannian optimization algorithms is a novel and promising area of research, indeed even if Information Geometry and second-order manifold optimization are well consolidated fields, surprisingly little work has been done at the intersection of the two. The optimization methods developed for statistical models based on dual geometries can be adapted to the larger class of Hessian manifolds. Indeed, all Hessian manifolds admit a dual geometrical structure analogous to that of statistical manifolds, given by the dual affine connections. Hessian manifolds include matrix manifolds, such as the cone of positive-definite matrices, and several other convex cones, with applications in robotics, computer vision, pattern recognition, signal processing, conic optimization, and many others.

In this work, after a description of the general theory behind second-order Information Geometry, we present two examples: an application to the optimization over an exponential family defined over a finite sample space, and the case of the multivariate Gaussian distribution.

Keywords: Information geometry, natural gradient, optimization over manifolds, Hessian manifolds, dually-flat alpha connection, exponential and mixture Hessians.