What is Computational Information Geometry?

Frank Nielsen
Sony Computer Science Laboratories Inc
Tokyo, Japan

2024

Information geometry [2] defines, studies, and applies core dualistic structures on smooth manifolds: Namely, pairs of dual affine connections (∇, ∇^*) coupled with Riemannian metrics g. In particular, those (g, ∇, ∇^*) structures can be built from statistical models [2] or induced by divergences [3] (contrast functions on product manifolds) or convex functions [19] $F(\theta)$ on open convex domains Θ (e.g., logarithmic characteristic functions of symmetric cones [21, 18]). In the latter case, manifolds are said dually flat [2] or Hessian [19] since the Riemannian metrics can be expressed locally either as $g(\theta) = \nabla^2 F(\theta)$ in the ∇-affine coordinate system θ or equivalently as $g(\eta) = \nabla^2 F^*(\eta)$ in the ∇^*-affine coordinate system η. The Legendre-Fenchel duality $F^*(\eta) = \sup_{\theta \in \Theta} \langle \theta, \eta \rangle - F(\theta)$ allows to convert between primal to dual coordinates: $\eta(\theta) = \nabla F(\theta)$ and $\theta(\eta) = \nabla F^*(\eta)$. Dually flat spaces have been further generalized to handle singularities in [10].

To get a taste of computational information geometry (CIG), let us mention the following two problems when implementing information-geometric structures and algorithms:

- In practice, we can fully implement geometric algorithms on dually flat spaces when both the primal potential function $F(\theta)$ and the dual potential function $F^*(\eta)$ are known in closed-form and computationally tractable [14]. See also the Python library `pyBregMan` [16]. To overcome computationally intractable potential functions, we may either consider Monte Carlo information geometry [14] or discretizing continuous distributions into a finite number of bins [6, 13] (amounts to consider standard simplex models).

- The Chernoff information [5] between two absolutely continuous distributions P and Q with densities $p(x)$ and $q(x)$ with respect to some dominating measure μ is defined by

$$C(P, Q) = \max_{\alpha \in (0,1)} -\log \int p^{\alpha} q^{1-\alpha} \mathrm{d}\mu = -\log \int p^{\alpha^*} q^{1-\alpha^*} \mathrm{d}\mu,$$

  where α^* is called the optimal exponent. Chernoff information is used in statistics and for information fusion tasks [7] among others. In general, the Chernoff information between two continuous distributions is not available in closed form (e.g., not known in closed-form between multivariate Gaussian distributions [12]). However, for densities p and q of an exponential family, the optimal exponent α^* can be characterized exactly geometrically as the unique intersection of the e-geodesic γ_{pq} with a dual m-bisector [11]. This geometric characterization yields an efficient approximation algorithm.

Thus computational information geometry aims at implementing robustly the information-geometric structures and the geometric algorithms on those structures for various applications. To give two examples of CIG, consider

- computing the minimum enclosing ball (MEB) of a finite set of m-dimensional points on a dually flat space: The MEB is always unique and can be calculated (in theory) using a LP-type randomized linear-time solver [15] (linear programming-type) relying on oracles which exactly compute the enclosing balls passing through exactly k points for $k \in \{2, \ldots, m\}$. However, these oracles are in general computationally intractable so that guaranteed approximation algorithms have been considered [17].

- Learning a deep neural networks using natural gradient [1, 4]: In practice, the number of parameters of a DNN is very large so that it is impractical to learn the weights of a DNN with natural gradient descent which require to handle large (potentially inverse) Fisher information matrices. Many practical approaches closely related to natural gradient have been thus considered in machine learning [9, 20, 8].

References

[1] Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

[2] Shun-ichi Amari. *Information geometry and its applications*, volume 194. Springer, 2016.

[3] Shun-ichi Amari and Andrzej Cichocki. Information geometry of divergence functions. *Bulletin of the polish academy of sciences. Technical sciences*, 58(1):183–195, 2010.

[4] Ovidiu Calin. *Deep learning architectures*. Springer, 2020.

[5] Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, pages 493–507, 1952.

[6] Frank Critchley and Paul Marriott. Computational information geometry in statistics: theory and practice. *Entropy*, 16(5):2454–2471, 2014.

[7] Simon J Julier. An empirical study into the use of Chernoff information for robust, distributed fusion of Gaussian mixture models. In *2006 9th International Conference on Information Fusion*, pages 1–8. IEEE, 2006.

[8] Wu Lin, Frank Nielsen, Khan Mohammad Emtiyaz, and Mark Schmidt. Tractable structured natural-gradient descent using local parameterizations. In *International Conference on Machine Learning*, pages 6680–6691. PMLR, 2021.

[9] James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020.

[10] Naomichi Nakajima and Toru Ohmoto. The dually flat structure for singular models. *Information Geometry*, 4(1):31–64, 2021.

[11] Frank Nielsen. An information-geometric characterization of Chernoff information. *IEEE Signal Processing Letters*, 20(3):269–272, 2013.

[12] Frank Nielsen. Revisiting Chernoff information with likelihood ratio exponential families. *Entropy*, 24(10):1400, 2022.

[13] Frank Nielsen, Frank Critchley, and Christopher TJ Dodson. *Computational Information Geometry*. Springer, 2017.

[14] Frank Nielsen and Gaëtan Hadjeres. Monte carlo information-geometric structures. *Geometric Structures of Information*, pages 69–103, 2019.

[15] Frank Nielsen and Richard Nock. On the smallest enclosing information disk. *Information Processing Letters*, 105(3):93–97, 2008.

[16] Frank Nielsen and Alexander Soen. *pyBregMan: A Python package for Bregman Manifolds*. Tokyo, Japan, 2024.

[17] Richard Nock and Frank Nielsen. Fitting the smallest enclosing Bregman ball. In *European Conference on Machine Learning*, pages 649–656. Springer, 2005.

[18] Atsumi Ohara and Shinto Eguchi. Geometry on positive definite matrices deformed by V-potentials and its submanifold structure. *Geometric Theory of Information*, pages 31–55, 2014.

[19] Hirohiko Shima. *The geometry of Hessian structures*. World Scientific, 2007.

[20] Ke Sun and Frank Nielsen. Relative Fisher information and natural gradient for learning large modular models. In *International Conference on Machine Learning*, pages 3289–3298. PMLR, 2017.

[21] Keiko Uohashi and Atsumi Ohara. Jordan algebras and dual affine connections on symmetric cones. *Positivity*, 8:369–378, 2004.