What is Information Geometry and Deep Neural Networks?

Ke Sun
CSIRO's Data61, Eveleigh NSW 2015, Australia

December 12, 2024

In information geometry (IG) [2], the space of probability distribution $\mathcal{S}$ is treated as a differentiable manifold, where information theoretic quantities are associated with geometric measurements. It endows a geometric structure in the parameter space of deep neural networks (DNNs), also known as the *neuromanifold* denoted as $\mathcal{M}$: a neural network is often, although not always, represented by a probability distribution through the mapping $\mathcal{M} \to \mathcal{S}$, and therefore one can pullback the metric tensor of $\mathcal{S}$ to define a geometry of $\mathcal{M}$.

Each point on $\mathcal{M}$ is a realization of a neural network with a prescribed architecture. The point moving on $\mathcal{M}$ means the network parameters denoted by θ change continuously. Different architectures form different neuromanifolds whose dimensionality scales with the size of the network. By IG, the local metric tensor of $\mathcal{M}$ is defined by $\mathcal{I}_{ab}(\theta)\mathrm{d}\theta^a\mathrm{d}\theta^b$, where $\mathcal{I}(\theta)$ is the Fisher information matrix (FIM) [6, 20]. Local measurements such as infinitesimal length or the Riemannian volume element is invariant to reparameterization of the neural network. Invariance is important as efficient training mechanism of DNNs, e.g., normalization [8, 4] and centering [15] techniques, usually only affect how the coordinate system of $\mathcal{M}$ is constructed without altering $\mathcal{M}$ into a different neuromanifold. These techniques should *not* affect how information is measured.

One can therefore trace a learning path by taking a series of jumps $\theta_0, \theta_1, \theta_2, \cdots$ on $\mathcal{M}$ until reaching a local optimum $\theta^\star \in \mathcal{M}$. As perhaps the most well-known application of IG in DNNs, natural gradient descent [1, 19] is based on the gradient vector field of the loss $\ell(\theta)$ with respect to the metric tensor $\mathcal{I}(\theta)$, given by $\mathcal{I}^{ab}(\theta)\frac{\partial \ell}{\partial \theta_a}\partial\theta_b := \bigtriangledown^b \partial\theta_b$. An optimization step is given by $\theta_{t+1} \leftarrow \theta_t - \lambda \bigtriangledown$, where $\lambda > 0$ is the learning rate. The learning trajectory depends on parameterization as $\bigtriangledown^b \partial\theta_b$ is a tangent vector and should be mapped onto the manifold via the exponential map, which is computationally prohibitive due to the high complexity of $\mathcal{M}$. In fact, it is already impractical to compute the inverse of the FIM $\mathcal{I}^{ab}(\theta)$. Practitioners take diagonal [11] or block-diagonal [14] approximations, combined with Monte-Carlo estimations [17, 21], leading to variations of the natural gradient method [13]. The estimation quality of random FIM estimators is analyzed recently [21].

There are a few key characteristics of the manifold $\mathcal{M}$ which distinguish it from traditional statistical models (e.g., Gaussian manifold) studied in IG. First, $\mathcal{I}(\theta) \succeq 0$ is positive semi-definite and is highly singular [3, 9] partly due to the huge size of θ, which can be as large as billions or trillions for modern networks, with its metric signatures varying with $\theta \in \mathcal{M}$. The geometric structure of $\mathcal{M}$ should be studied with singular semi-Riemannian geometry [22] and singular statistical learning theory [24]. For natural gradient which requires inverting the FIM, $\mathcal{I}(\theta)$ is either regularized by adding ϵI [11], where $\epsilon > 0$ and I is the identity matrix, or its pseudo-inverse is used [23]. Second, θ in practice can be discrete with varying precisions (see e.g. [7, 16]). The geometric concepts related to the smooth manifold $\mathcal{M}$ should be adapted accordingly. Quantization connects with IG through the Cramér-Rao bound: $\mathrm{Var}(\theta) \geq \mathcal{I}^{-1}(\theta)$, where $\mathrm{Var}(\theta)$ means the variance of the parameter θ with respect to a single observation. Therefore the

precisions of θ should be positively correlated with $\mathcal{I}^{-1}(\theta)$.

In the ambient space of probability distributions, one can regard the input data X as the empirical distribution $\delta(X)$ which is usually outside the manifold $\mathcal{M}$. Therefore learning is to seek a projection from $\delta(X)$ onto $\theta^\star \in \mathcal{M}$. IG offers a rich family of information divergences that can be used to construct the loss, minimizing which achieves such a projection. The cross-entropy loss widely used in DNNs corresponds to the Kullback-Leibler (KL) divergence up to a constant. The family of α-divergences [12], f-divergences [18], Bregman-divergences [5], and optimal transport [10] have been applied in different DNNs.

References

[1] Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 02 1998.

[2] Shun-ichi Amari. *Information Geometry and Its Applications*, volume 194 of *Applied Mathematical Sciences*. Springer-Verlag, Berlin, 2016.

[3] Shun-ichi Amari, Tomoko Ozeki, Ryo Karakida, Yuki Yoshida, and Masato Okada. Dynamics of learning in MLP: Natural gradient and singularity revisited. *Neural Computation*, 30(1):1–33, 2018.

[4] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. arXiv:1607.06450 [stat.ML].

[5] Hatice Kubra Cilingir, Rachel Manzelli, and Brian Kulis. Deep divergence learning. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2027–2037. PMLR, 13–18 Jul 2020.

[6] Harold Hotelling. Spaces of statistical parameters. *Bull. Amer. Math. Soc*, 36:191, 1930.

[7] Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks. In *NIPS 29*, pages 4107–4115. Curran Associates, Inc., 2016.

[8] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France, 07–09 Jul 2015. PMLR.

[9] Ryo Karakida, Shotaro Akaho, and Shun-ichi Amari. Pathological Spectra of the Fisher Information Metric and Its Variants in Deep Neural Networks. *Neural Computation*, 33(8):2274–2307, 2021.

[10] Abdelwahed Khamis, Russell Tsuchida, Mohamed Tarek, Vivien Rolland, and Lars Petersson. Scalable optimal transport methods in machine learning: A contemporary survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[11] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[12] Yingzhen Li and Richard E Turner. Rényi divergence variational inference. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

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

[14] James Martens and Roger Grosse. Optimizing Neural Networks with Kronecker-factored Approximate Curvature. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2408–2417, Lille, France, 07–09 Jul 2015. PMLR.

[15] Jan Melchior, Asja Fischer, and Laurenz Wiskott. How to Center Deep Boltzmann Machines. *Journal of Machine Learning Research*, 17(99):1–61, 2016.

[16] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. Mixed precision training. In *ICLR*, 2018.

[17] Frank Nielsen and Gaëtan Hadjeres. *Monte Carlo Information-Geometric Structures*, pages 69–103. Springer International Publishing, 2019.

[18] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[19] Razvan Pascanu and Yoshua Bengio. Revisiting natural gradient for deep networks. In *International Conference on Learning Representations*, 2014.

[20] Calyampudi Radhakrishna Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of Cal. Math. Soc.*, 37(3):81–91, 1945.

[21] Alexander Soen and Ke Sun. Trade-offs of diagonal Fisher information matrix estimators. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

[22] Ke Sun and Frank Nielsen. A geometric modeling of Occam's razor in deep learning, 2019. arXiv:1905.11027.

[23] Philip Thomas. Genga: A generalization of natural gradient ascent with positive and negative convergence results. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1575–1583. PMLR, 22–24 Jun 2014.

[24] Sumio Watanabe. *Algebraic Geometry and Statistical Learning Theory*, volume 25 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, Cambridge, United Kingdom, 2009.