INFORMATION GEOMETRY OF NEURAL NETWORKS — AN OVERVIEW —

Shun-ichi Amari

University of Tokyo, Tokyo, Japan, RIKEN Frontier Research Program, Wako-City, Saitama, Japan. Email: amari@sat.t.u-tokyo.ac.jp

The set of all the neural networks of a fixed architecture forms a geometrical manifold where the modifiable connection weights play the role of coordinates. It is important to study all such networks as a whole rather than the behavior of each network in order to understand the capability of information processing of neural networks. What is the natural geometry to be introduced in the manifold of neural networks? Information geometry gives an answer, giving the Riemannian metric and a dual pair of affine connections. An overview is given to information geometry of neural networks.

1 Introduction to Neural Manifolds

Let us consider a neural network of fixed architecture specified by parameters \( w = (w_1, \cdots, w_p) \) which represent the connection weights and thresholds of the network. The parameters are usually modifiable by learning. The set \( N \) of all such networks is considered a \( p \)-dimensional neural manifold, where \( w \) is a coordinate system in \( N \). Because it includes all the possible networks belonging to that architecture, the total capabilities of the networks are made clear by studying the manifold \( N \) itself.

To be specific, let \( N \) be the set of multilayer feedforward networks each of which receives an input \( x \) and emits an output \( z \). The input-output relation is described as

\[
z = f(x; w)
\]

where the total output \( z \) depends on \( w \) which describes all the connection weights and thresholds of the hidden and output neurons. Let us consider the space \( S \) of all the square integrable functions of \( x \)

\[
S = \{k(x)\}
\]

and assume for the moment that \( f(x; w) \) is square integrable. The set \( S \) is infinite-dimensional \( L_2 \) space, and the neural manifold \( N \) is a part of it, that is, a \( p \)-dimensional subspace embedded in \( S \). This shows that not all the functions are realizable by neural networks.

Given a function \( k(x) \), we would like to find a neural network whose behavior \( f(x; w) \) approximates \( k(x) \) as well as possible. The best approximation is given by projecting \( k(x) \) to \( N \) in the entire space \( S \). The approximation power depends on the shape of \( N \) in \( S \). This shows that geometrical considerations are important.

When the behavior of a network is stochastic, it is given by the conditional probability \( p(z|x; w) \) of \( z \) conditioned on input \( x \), where \( w \) is the network parameters. A typical example is a network composed of binary stochastic neurons: The probability \( z = 1 \) of such a stochastic neuron is given by

\[
\text{Prob}\{z = 1; w\} = \frac{\exp\{w \cdot x\}}{1 + \exp\{w \cdot x\}},
\]

where \( z = 0 \) or 1. Another typical case is a noise-contaminated network whose output \( z \) is written as

\[
z = f(x; w) + n,
\]
where \( n \) is a random noise independent of \( w \). If \( n \) is subject to the normal distribution with mean 0 and covariance \( \sigma^2 I \), \( I \) being the identity matrix,

\[
n \sim N(0, \sigma^2 I),
\]

the conditional probability is given by

\[
p(z|x; w) = \text{const} \exp \left\{ -\frac{[z - f(x, w)]^2}{2\sigma^2} \right\}.
\]

(3)

When input signals \( x \) are produced independently from a distribution \( q(x) \), the joint distribution of \((x, z)\) is given by

\[
p(x, z; w) = q(x)p(z|x; w).
\]

(4)

Let \( S \) be the set of all the conditional probability distributions (or joint distributions). Let \( q(z|x) \) be an arbitrary probability distribution in \( S \) which is to be approximated by a stochastic neural network of the behavior \( p(z|x; w) \). The neural manifold \( N \) consists of all the conditional probability distributions \( p(z|x; w) \) (or the joint probability distributions \( p(x, z; w) \)) and is a \( p \)-dimensional submanifold of \( S \). A fundamental question arises: What is the natural geometry of \( S \) and \( N \)? How should the distance between two distributions be measured? What is the geodesic connecting two distributions? It is important to have a definite answer to these problems not only for studying stochastic networks but also for deterministic networks which are not free of random noises and whose stochastic interpretation is sometimes very useful. Information geometry ([3], [17]) answers all of these problems.

2 A Short Review of Information Geometry

Let us consider probability distributions \( p(y, \xi) \) of random variable \( y \), where \( \xi = (\xi_1, \cdots, \xi_n) \) is the parameters to specify a distribution. When \( y \) is a scalar and normally distributed, we have

\[
p(y, \xi) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{(y - \mu)^2}{2\sigma^2} \right\},
\]

where \( \xi = (\mu, \sigma) \) is the parameters to specify it. When \( y \) is discrete, taking values on \( \{0, 1, \cdots, n\} \), we have

\[
\text{Prob}\{y, \xi\} = \sum_{i=1}^{n} \xi_i \delta_i(y) + \left(1 - \sum_{i=1}^{n} \xi_i\right) \delta_0(y)
\]

where \( \delta_i(y) = 1 \) when \( y = i \) and is equal to 0 otherwise and \( \xi_i = \text{Prob}\{y = i\} \).

Let \( S \) be the set of such distributions

\[
S = \{p(y, \xi)\}
\]

specified by \( \xi \). Then, \( S \) can be regarded as an \( n \)-dimensional manifold, where \( \xi \) is a coordinate system. If we can introduce a distance measure between two nearby points specified by \( \xi \) and \( \xi + d\xi \) by the quadratic form

\[
|d\xi|^2 = \sum_{i,j} g_{ij}(\xi)d\xi_id\xi_j,
\]

(5)

the manifold \( S \) is said to be Riemannian. The quantity \( \{g_{ij}(\xi)\} \) is the Riemannian metric tensor. Another important concept is the affine connection by which a geodesic line (an extention of the concept of the straight line in the Euclidean