Intrinsic Plasticity via Natural Gradient Descent

Klaus Neumann and Jochen J. Steil

Research Institute for Cognition and Robotics (CoR-Lab)
Bielefeld University - Germany

Abstract. This paper introduces the natural gradient for intrinsic plasticity, which tunes a neuron's activation function such that its output distribution becomes exponentially distributed. The information-geometric properties of the intrinsic plasticity potential are analyzed and the improved learning dynamics when using the natural gradient are evaluated for a variety of input distributions. The applied measure for evaluation is the relative geodesic length of the respective path in parameter space.

1 Introduction

Stochastic gradient descent methods are commonly used learning techniques to minimize cost functions for non-linear optimization [1]. However, the parameter estimates can lead to small gradient norms in some regions of the parameter space, so called plateaus, where convergence can be slow.

One reason for this is that the parameterization and the corresponding output of a model are defined in different metrical spaces. Most gradients defined on an error measure only utilize Euclidean metrics in parameter space. But, generally, there is no reason to assume that Euclidean metrics is the preferential distance measure between solutions. It is well known that the parameter space has a Riemannian metric structure in many cases that is analyzed by means of information geometry [2] - a theory which employs differential-geometric methods in statistics.

This theory can be used to define a canonical distance measure in the output space. It employs a non-trivial and often only locally defined metric tensor well-suited to the Riemannian metric structure of the parameter space (see Fig. 2). While the steepest direction in a parameter space with an Euclidean metric structure is given by the conventional gradient, the steepest direction in a parameter space with Riemannian metric structure is given by the so-called natural gradient. It was already shown in [3, 4, 5] that the use of the natural gradient can be advantageous for neural network learning.

In 2004, Triesch introduced a model of intrinsic plasticity (IP) [6] for optimization of the neuron's activation function based on stochastic gradient descent. The target of IP-learning is to approximate an exponential output distribution with respect to a given input distribution. This maximizes the neuron's information transmission, caused by the high entropy of the target distribution. The algorithm was also used to enhance the encoding in reservoir networks [7].

This paper introduces the Riemannian metric tensor for IP in parameter space. The experiments reveal that it is more suited than an Euclidean metric to describe distance relations between output distributions. This Tensor induces the natural gradient and leads to a more general learning rule (NIP).

First, the Sects. 2 and 3 describe how the natural gradient is defined for IP. Second, Sect. 4 contains experiments which complement the theory in the paper by analyzing the differential-geometric properties of IP and show how the learning dynamics change due to the use of the natural gradient. Finally, Sect. 5 concludes the paper.

2 Intrinsic Plasticity

Intrinsic Plasticity (IP) was developed by Triesch in 2004 [6] as a model for homeostatic plasticity for analog neurons with parameterized Fermi-functions $y(x|\theta) = (1 + e^{(-ax-b)})^{-1}$ as activation with parameters $\theta = (a, b)^T$. The goal is to optimize the information transmission of a single neuron strictly locally by adaptation of slope a and bias b such that the neuron's output y becomes exponentially distributed with a fixed mean μ with respect to the input sample distribution $f_x(x)$ where x is the synaptic sum arriving at the neuron.

IP-learning can be derived by minimizing the difference $L(f_y, f_{\exp}) = L(\theta)$ between the output f_y and an exponential distribution $f_{\exp}$, quantized by the Kullback-Leibler-divergence [9] (KLD):

$$L(\theta) = \mathbb{E}_x[l(y,\theta)] = \int_\Omega f_y(y) \log\left(\frac{f_y(y)}{f_{\exp}(y)}\right) \mathrm{d}y = \int_\Omega \underbrace{-\ln\left(\frac{ay(1-y)}{e^{-\frac{1}{\mu}y}}\right)}_{l(y,\theta)} \mathrm{d}y \ . \quad (1)$$

Since IP was introduced as stochastic gradient decent, the online loss function is identified as the integrand $l(y, \theta)$ (see Eq. (1)). Therefore the KLD is interpreted as expected loss $\mathbb{E}_x[l(y, \theta)]$ for the input samples x distributed by $f_x(x)$. Interestingly, the original contribution [6] shows an additive separation of Eq. (1) into the entropy $H_x[y]$ and the expectation value of the output distribution $\mathbb{E}_x[y]$, which directly infers that a minimization of $L(\theta)$ for a fixed mean $\mathbb{E}_x[y]$ is equivalent to entropy maximization of the output distribution. The KLD and the distance to exponential distributions is deeply analyzed in [10]. The typical approach is to use the stochastic gradient of this potential in order to find a minimum of the expected loss function. The following online update equations for slope and bias - scaled by the step width η_{IP} - are obtained:

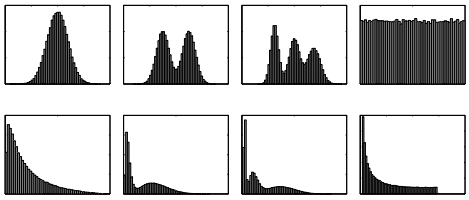

Figure 1: Four different input distributions $f_x(x)$ (1. row) and the corresponding learned exponential-like output distributions $f_y(y)$ for $\mu = 0.2$ (2. row).

$$\Delta a = \frac{\eta_{\mathrm{IP}}}{a} + x\Delta b \qquad \Delta b = \eta_{\mathrm{IP}}\left(1 - \left(2 + \frac{1}{\mu}\right)y + \frac{1}{\mu}y^2\right) \ . \quad (2)$$

Fig. 1 shows how four different input distributions (first row in the figure) are transformed into exponential-like distributions (second row in the figure) after

training with IP. The figure clearly reveals that the best possible fit after IP learning highly depends on the input distribution. This is due to the fact that only two parameters in the Fermi-function are adapted.

3 Natural Gradient for Intrinsic Plasticity

Given an input distribution $f_x(x)$, an analog neuron, establishes a differentiable function mapping between the parameter space $\Theta = \mathbb{R}^2$ and the manifold of possible output distributions Υ. The KLD comparing a given distribution to the exponential distribution with fixed mean μ in Eq. (1) can be used to derive a canonical distance measure on the output distribution space resulting in a non-Euclidean metric F on the parameter space Θ. The metric determining the distance between two output distributions $y_1(x) = y(x, \theta_1)$ and $y_2(x) = y(x, \theta_2)$ in Υ given by the parameter settings θ_1 and $\theta_2 = \theta_1 + \mathrm{d}\theta$ in Θ for an infinitesimal change of parameters $\mathrm{d}\theta$ is given by $D(y_1, y_2)$. This distance measure is transformed such that it induces the Riemannian metric tensor $F(\theta)$ - a 2×2 positive definite matrix given by the Fisher information [8] - as a pull-back onto the parameter space:

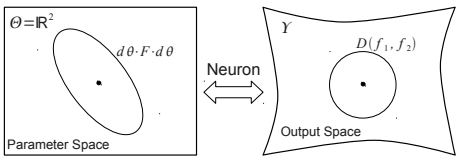

Figure 2: Established differentiable relation (Neuron) and metrics F and D between parameter space $\Theta = \mathbb{R}^2$ and manifold of possible output distributions Υ.

$$
\begin{aligned}
D(y_1, y_2) &= \mathbb{E}_x[(l(y_1, \theta_1) - l(y_2, \theta_2))^2] && (3)\\
&= \mathbb{E}_x[\big(l(y_1, \theta_1) - l(y_1, \theta_1) - \nabla^T l(y_1, \theta_1)\mathrm{d}\theta\big)^2] && (4)\\
&= \mathbb{E}_x[\big(\nabla^T l(y_1, \theta_1)\mathrm{d}\theta\big)^2] && (5)\\
&= \mathrm{d}\theta \cdot \mathbb{E}_x[\nabla l(y_1, \theta_1) \cdot \nabla^T l(y_1, \theta_1)] \cdot \mathrm{d}\theta = \mathrm{d}\theta \cdot F(\theta) \cdot \mathrm{d}\theta && (6)
\end{aligned}
$$

This idea guarantees that the distance between two parameter vectors θ_1 and θ_2 as measured by the length of the geodesic with respect to the metric tensor $F(\theta)$ in Eq. (6) is equal to the previously defined distance measure $D(y_1, y_2)$ in Eq. (3) on the corresponding output distributions y_1 and y_2 in Υ.

It was already shown in [3] that parameter spaces spanned by neural networks have a Riemannian character. In such spaces, the steepest descent direction of a potential is given by the natural gradient defined by the metric tensor. The following update equation is obtained when using the natural gradient for IP:

$$
\theta_{t+1} = \theta_t - \eta(F(\theta) + \varepsilon I)^{-1} \nabla l(y, \theta) = \theta_t - \eta \nabla_{\mathrm{NIP}} l(y, \theta) \ , \qquad (7)
$$

where I is the 2×2 - identity matrix and $\varepsilon \geq 0$ is a positive scalar. We call $\nabla_{\mathrm{NIP}} := (F(\theta) + \varepsilon I)^{-1} \nabla$ the natural gradient operator for IP. Typically ε can be set to zero to obtain a plain natural gradient formulation. But in the more general definition Eq. (7), ε introduces a blending between conventional

and natural gradient. Note that this blending influences the step width of the numerically applied gradient descent and stabilizes the inversion of the metric tensor F.

There are several important issues to mention when dealing with the natural gradient adaptation for IP: (i) Both gradient descents (conventional and natural) have the same attractors [2]. (ii) It appeared in all experiments that the potential L had always one attractor on each slope-half plane, see Fig. 3. (iii) The mapping from parameter space to output distribution is not surjective with respect to a fixed input distribution (no perfect approximation possible). (iv) In contrast to Newton's method, the natural gradient algorithm does not assume a locally quadratic potential (F is always positive definite and convergence guaranteed).

However, in comparision to the conventional IP, the new algorithm is more complex due to the calculation of the expectation value (see Eq. 6) and therefore losses its online properties. This problem can be solved by an online estimation of the metric tensor - done e.g. by proportional control laws.

4 Results

This section summarizes the experimental results of this contribution when using Eq. (7) for a natural gradient version of IP. The experiments where performed with different inputs: The first row in Fig. 1 shows the four different input distributions that are used for investigation. A Gaussian (1-G), a bipartite (2-G), a tripartite (3-G) Gaussian and a uniform (U) distribution. $N_{\mathrm{tr}} = 100$ samples are used for training independently drawn from each distribution. A step width of $\eta = 10^{-3}$ and a numerical stabilization of $\varepsilon = 10^{-1}$ is used.

4.1 Information Geometry

The following experiment visualizes how the geometry of the potential L changes by use of the metric tensor at the attractor θ^*. The 1-G distribution is exemplary used as input. Fig. 3 (center) shows the potential $L(\theta)$ with a clearly visible plateau in b-direction (the x-axis corresponds to the slope a and the y-axis to the bias b for Fig. 3). The change in the KLD is small in that direction. The dashed line is the unit circle with a radius of η in the geometry defined by the metric tensor $F(\theta^*)$, which is well suited to the potential: The unit circle is stretched in b-direction. Fig. 3 (right) visualizes the distortion of the potential after transformation with $F(\theta^*)$. The induced landscape becomes "Euclidean-like" after transformation and losses the plateau - the potential develops isotropic convergence properties.

4.2 Information Geodesy

The following experiments focus on a more global analysis of the natural gradient descent. A gradient descent from a given starting point θ to the attractor θ^* is performed while the relative geodesic length of the path (RGL) is recorded. The

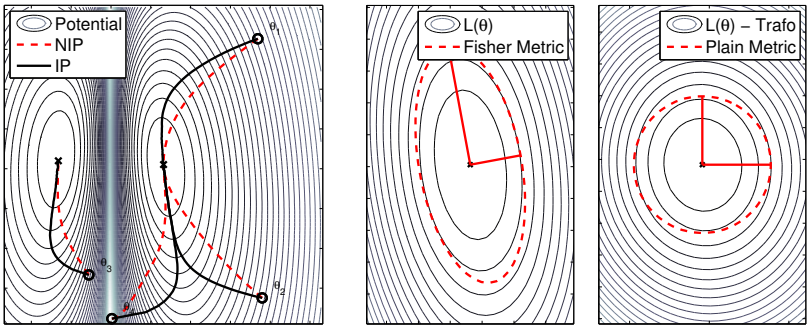

Figure 3: Geodesics in the IP potential (left). The geometry change of the attractor basin using the natural gradient (center and right). IP potential and Fisher metric at the attractor (center). NIP potential and plain Euclidean metric (right).

RGL gives the length of the geodesic γ from starting point θ to the attractor θ^* with respect to the shortest way in the parameter space:

$$\mathrm{RGL}(\theta) = \int_\gamma \mathrm{d}s \;\; / \;\; \|\theta - \theta^*\| \;\; , \tag{8}$$

where $\mathrm{d}s = \sqrt{\mathrm{d}a^2 + \mathrm{d}b^2}$ is the infinitesimal arc length in parameter space.

Fig. 3 (left) shows the potential field $L(\theta)$ of the 1-G input distribution. It also shows four starting points θ_{1-4} for the learning. The solid lines show gradient descents performed by IP, while the dashed lines are the geodesics from the NIP learning. Although, both approaches have the same fixed-point, the geodesics of the NIP learning are on a more direct way to the attractor in parameter space - the natural gradient method "sees more from the attractor" than the conventional IP gradient. Tab. 1 displays the results of an experiment where the RGL is measured for $N = 100$ different starting points drawn from a Gaussian distribution centered around the attractor with covariance matrix $\Sigma = I$. It contains the average RGL and its standard deviation.

Task	$\mathbb{E}$[RGL] (IP)	$\mathbb{E}$[RGL] (NIP)
1-G	1.3493 ± 0.5730	1.0748 ± 0.0526
2-G	1.0473 ± 0.0300	1.0234 ± 0.0342
3-G	1.1209 ± 0.0753	1.0505 ± 0.0506
U	1.0219 ± 0.0206	1.0056 ± 0.0099

Table 1: Relative average length of the geodesics $\mathbb{E}$[RGL] and their standard deviation $\sqrt{\mathbb{E}[(\mathrm{RGL} - \mathbb{E}[\mathrm{RGL}])^2]}$ for IP and NIP learning.

Since the best possible value for the RGL is one (which corresponds to a straight line from the initial point to the attractor in parameter space), the values for the RGL in Tab. 1 show that the geodesic lines are almost straight for all tested input distributions (visualized in Fig. 3). In addition, the low standard deviation demonstrates that the curvature of the geodesic is more independent from the initial point in the potential.

5 Conclusion

This contribution has adapted the intrinsic plasticity learning to the natural gradient technique. It was shown that the new metric in parameter space is more suited to the problem of tuning output distributions of non-linear neurons. The geometry of the IP potential and the geodesy of the different gradient descents were analyzed and revealed favorable properties when using the natural instead of the conventional gradient.

Such an algorithm can be used on a network level. Further research should investigate whether recurrent networks profit from the unsupervised adaptation with natural gradient IP.

References

[1] Rumelhart, D. E. and Hinton, G. E. and Williams, R. J.: Learning representations by back-propagating errors. Nature, pages 533–536, volume 323, 1986.

[2] Amari, Shun-ichi: Differential-Geometrical Methods in Statistics. Lecture Notes in Statistics 28, New-York: Springer-Verlag, 1985.

[3] Amari, Shun-Ichi: Natural gradient works efficiently in learning. Neural Computation, pages 251–276, volume 10, 1998.

[4] Amari, S.-I. and Douglas, S. C.: Why natural gradient? Acoustics, Speech, and Signal Processing, pages 1213–1216, volume 2, 1998.

[5] Douglas, C. C. and Hu, J. and Ray, J. and Thorne, D. T. and Tuminaro, R. S.: Natural-gradient adaptation. Unsupervised Adaptive Filtering, Vol. I: Blind Source Separation, pages 13–61, 2000.

[6] Triesch, J.: Synergies beween Intrinsic and Synaptic Plasticity in Individual Model Neurons. NIPS, 2005.

[7] Steil, J. J.: Online Reservoir Adaptation by Intrinsic Plasticity for Backpropagation - Decorrelation and Echo State Learning. Neural Networks, Special Issue on Echo State and Liquid State networks, pages 353–364, 2007.

[8] Rao, C. R.: Information and the accuracy attainable in the estimation of statistical parameters. Bull. Calcutta Math. Soc., 1945.

[9] Kullback, S. and Leibler, A.: On Information and Sufficiency. Ann. Math. Statist., pages 79–86, volume 22, number 1, 1951.

[10] Kostal, L. and Lansky, P.: Classification of stationary neuronal activity according to its information rate. Network Computation in Neural Systems, volume 17, pages 193–210, 2006.