ADAPTIVE NATURAL GRADIENT LEARNING BASED ON RIEMANNIAN METRIC OF SCORE MATCHING

Ryo Karakida[1], **Masato Okada**[1,2] **& Shun-ichi Amari**[2]
[1]Department of Complexity Science and Engineering, The University of Tokyo, Chiba, Japan
[2]RIKEN Brain Science Institute, Saitama, Japan
{karakida@mns.,okada@}k.u-tokyo.ac.jp, amari@brain.riken.jp

ABSTRACT

The natural gradient is a powerful method to improve the transient dynamics of learning by considering the geometric structure of the parameter space. Many natural gradient methods have been developed with regards to Kullback-Leibler (KL) divergence and its Fisher metric, but the framework of natural gradient can be essentially extended to other divergences. In this study, we focus on score matching, which is an alternative to maximum likelihood learning for unnormalized statistical models, and introduce its Riemannian metric. By using the score matching metric, we derive an adaptive natural gradient algorithm that does not require computationally demanding inversion of the metric. Experimental results in a multi-layer neural network model demonstrate that the proposed method avoids the plateau phenomenon and accelerates the convergence of learning compared to the conventional stochastic gradient descent method.

1 SCORE MATCHING AND ITS RIEMANNIAN METRIC

Score matching has been developed for training unnormalized statistical models and applied to various kinds of practical applications such as signal processing (Hyvärinen, 2005) and representation learning for visual and acoustic data (Köster & Hyvärinen, 2010). We can also train single-layer models (Swersky et al., 2011; Vincent, 2011) and a two-layer model with the analytically intractable normalization constants (Köster & Hyvärinen, 2010), which are hard to train by maximum likelihood learning. The objective function of score matching is given by the squared distance between derivatives of the log-density, $D_{SM}[q:p] = \int d\mathbf{x} q(\mathbf{x}) \sum_i |\partial_i \log q(\mathbf{x}) - \partial_i \log p(\mathbf{x})|^2$, where we denote the derivative with respect to the i-th probability variable as a partial derivative symbol $\partial_i = \frac{\partial}{\partial x_i}$. In this paper, we refer to this objective function as score matching (SM) divergence.

In general, we can derive the Riemannian structure from any divergence (Eguchi, 1983; Amari, 2016). Let us consider a parametric probability distribution $p(\mathbf{x};\boldsymbol{\xi})$. When we estimate the parameter $\boldsymbol{\xi}$ with a divergence $D[q:p]$, its parameter space has the Riemannian metric matrix G defined by $D[p(\mathbf{x};\boldsymbol{\xi}) : p(\mathbf{x};\boldsymbol{\xi}+d\boldsymbol{\xi})] = \sum_{i,j} G_{ij} d\xi_i d\xi_j$. The metric matrix G can be obtained by the second derivative, $G_{ij} = \frac{\partial^2}{\partial \xi_i' \partial \xi_j'} D[p(\mathbf{x};\boldsymbol{\xi}) : p(\mathbf{x};\boldsymbol{\xi}')] \,\big|_{\boldsymbol{\xi}'=\boldsymbol{\xi}}$. In particular, when we consider the SM divergence, its metric becomes the following positive semi-definite matrix,

$$G = \sum_i < \nabla\partial_i \log p(\mathbf{x};\boldsymbol{\xi}) \nabla\partial_i \log p(\mathbf{x};\boldsymbol{\xi})^T >_{p(\mathbf{x};\boldsymbol{\xi})}, \tag{1}$$

where we denote the derivative with regard to a parameter vector $\boldsymbol{\xi}$ as $\nabla = \frac{d}{d\boldsymbol{\xi}}$ and the average over a probability distribution p as $< \cdot >_p$. Note that when we consider KL divergence, $D_{KL}[q:p] = \int d\mathbf{x} q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x})}$, its metric becomes a Fisher information matrix, $G =< \nabla \log p(\mathbf{x};\boldsymbol{\xi}) \nabla \log p(\mathbf{x};\boldsymbol{\xi})^T >_{p(\mathbf{x};\boldsymbol{\xi})}$. In contrast to the Fisher metric, the metric of SM divergence is composed of the log-likelihood differentiated with respect to ∂_i.

2 ADAPTIVE NATURAL GRADIENT LEARNING

As is known in information geometry, taking the Riemannian structure of an objective function into consideration, one can find the steepest direction of parameter space by natural gradient learning (Amari, 1998; 2016; Ollivier, 2015). The natural gradient algorithm is written as

$$\boldsymbol{\xi}_{t+1} = \boldsymbol{\xi}_t - \eta_t G_t^{-1} \nabla L_t, \tag{2}$$

where $\boldsymbol{\xi}_t$ is the parameter at time step t and η_t is a learning rate that may depend on t. The objective function of score matching is defined by $L_t = D_{SM}[q(\mathbf{x}) : p(\mathbf{x}; \boldsymbol{\xi}_t)]$, where we denote an input data distribution as $q(\mathbf{x})$ and a model distribution as $p(\mathbf{x}; \boldsymbol{\xi})$. After straightforward calculation, this objective function can be transformed into $L_t =< l(\mathbf{x}; \boldsymbol{\xi}_t) >_{q(\mathbf{x})} + const.$ with $l(\mathbf{x}; \boldsymbol{\xi}) = \sum_i \left\{ \frac{1}{2} (\partial_i \log p(\mathbf{x}; \boldsymbol{\xi}))^2 + \partial_i^2 \log p(\mathbf{x}; \boldsymbol{\xi}) \right\}$ (Hyvärinen, 2005). In this study, we compute the natural gradient as the online learning algorithm such that $L_t = l(\mathbf{x}_t; \boldsymbol{\xi}_t)$, where each data sample $\mathbf{x}_t$ is independently generated from $q(\mathbf{x})$.

The inversion of metric (1) at time step t defined by G_t^{-1} is approximately obtained as below. In general, the exact analytical calculation of metric (1) may be intractable because it requires the average over the unnormalized statistical model such that $< \cdot >_{p(\mathbf{x}; \boldsymbol{\xi}_t)}$. Here, we approximate the average over $p(\mathbf{x}; \boldsymbol{\xi}_t)$ by empirical expectation, $G_t \sim \sum_i < \nabla \partial_i \log p(\mathbf{x}; \boldsymbol{\xi}_t) \nabla \partial_i \log p(\mathbf{x}; \boldsymbol{\xi}_t)^T >_{q(\mathbf{x})}$. If the input data is generated by a true model distribution $q(\mathbf{x}) = p(\mathbf{x}; \boldsymbol{\xi}^*)$ and the learning parameter $\boldsymbol{\xi}_t$ converges to the true value $\boldsymbol{\xi}^*$, the approximated metric over input data is asymptotically equivalent to the exact metric.

In addition, we introduce an adaptive method to calculate the inversion of G_t, since the inversion of matrix demands much computational time in practice. Similar to the derivation of the adaptive natural gradient on KL divergence (Amari et al., 2000), we consider the online update of the SM metric,

$$G_{t+1} = (1 - \epsilon_t) G_t + \epsilon_t \sum_i \nabla \partial_i \log p(\mathbf{x}_t; \boldsymbol{\xi}_t) \nabla \partial_i \log p(\mathbf{x}_t; \boldsymbol{\xi}_t)^T. \tag{3}$$

When ϵ_t is small enough, we may approximate the inversion G_{t+1}^{-1} by using an approximation formula $(A + \epsilon_t B)^{-1} \sim A - \epsilon_t B$ and obtain the adaptive update rule of the inverted metric,

$$G_{t+1}^{-1} \sim (1 + \epsilon_t) G_t^{-1} - \epsilon_t G_t^{-1} \sum_i \nabla \partial_i \log p(\mathbf{x}_t; \boldsymbol{\xi}_t) \nabla \partial_i \log p(\mathbf{x}_t; \boldsymbol{\xi}_t)^T G_t^{-1}. \tag{4}$$

Note that, in the case where there are N variables v_i $(i = 1, ..., N)$ and K parameters ξ_i $(i = 1, ..., K)$, the computational complexity at every update step becomes $O(NK^2)$.

3 NUMERICAL EXPERIMENTS

To confirm the performance of the proposed methods, we trained the energy-based model of a two-layer neural network for natural stimuli proposed by Köster & Hyvärinen (2010). This model is defined by $\log p(\mathbf{x}; W, V) = \sum_h f(\mathbf{v}_h^T g(W\mathbf{x})) - \log Z(W, V)$ (Köster & Hyvärinen, 2010), where the nonlinear activation functions are given by $f(u) = -\sqrt{u + 1}$ and element-wise square $g(\mathbf{u}) = \mathbf{u}^2$. We denote an N-dimensional probability variable as $\mathbf{x} \in \mathbb{R}^N$, an $N \times N$ weight matrix between the input and the first hidden layers as W, an $N \times N$ non-negative weight matrix between the first and second hidden layers as V, and the rows of V as $\mathbf{v}_h^T$. Note that since the normalization constant $Z(W, V)$ is given by an intractable integral, it is difficult to train this model by maximum likelihood learning and its natural gradient with a Fisher metric. This model trained with score matching learns responses similar to simple cells and complex cells in the sensory cortex.

In this study, we set $N = 8$ and trained the model in an unsupervised manner with 5,000 samples of 8-dimensional data artificially generated by the Independent Subspace Analysis (ISA) model (Köster & Hyvärinen, 2010). We set the data vector $\mathbf{x}$ to be composed from four subspace vectors $\mathbf{s}_i \in \mathbb{R}^2$ $(i = 1, ..., 4)$ such that $\mathbf{x} = A[\mathbf{s}_1 \ \mathbf{s}_2 \ \mathbf{s}_3 \ \mathbf{s}_4]^T$, where each $\mathbf{s}_i$ is independently generated by a product between a uniform random variable u_i and a 2-dimensional random Gaussian variable, and A is a random mixing matrix.

As shown in Fig. 1, we found that the proposed adaptive natural gradient (ANG) converges much faster than the stochastic gradient descent (SGD). The update rule of SGD was given by $\boldsymbol{\xi}_{t+1} =$

$\boldsymbol{\xi}_t - \eta_t \nabla L_t$. More interestingly, we revealed that ANG avoids the plateau caused by the singularity of the parameter space, where the transient dynamics of SGD learning become very slow. The similar superiority of ANG to SGD has also been reported in the ANG based on KL divergence and Fisher metric (Amari et al., 2000; Park et al., 2000). In Table 1, we list the averaged performances of ANG and SGD over ten runs with different initial values of W and V. ANG learning achieved the comparable test error to SGD learning. In addition, in terms of step number until convergence, ANG learning was more than 10 times faster than SGD. In terms of processing time, ANG was also faster than SGD.

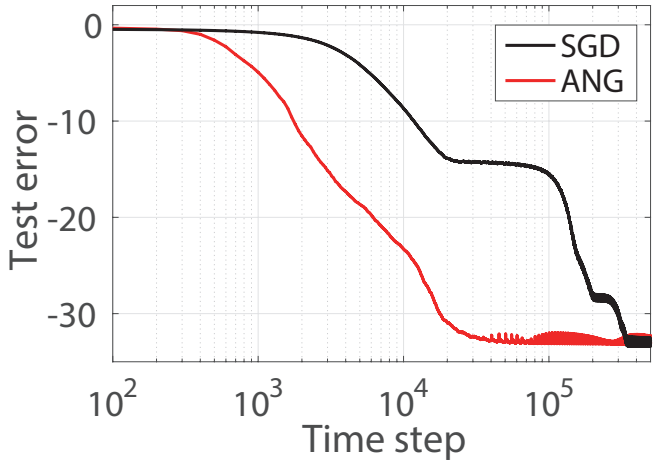

Figure 1: Transient dynamics of score matching learning in a two-layer network model: the conventional stochastic gradient descent (SGD) method and the proposed adaptive natural gradient (ANG) method. Test error means the object function on test data samples.

	SGD	ANG
Test error (avg. $\pm$ std)	-33.21 $\pm$ 2.25	-32.14 $\pm$ 3.58
Step number for test error $<$ -28	2.16×10^5	1.93×10^4
Processing time (relative to SGD)	1.00	0.34

Table 1: Averaged performance over 10 randomly chosen initial conditions. We set the learning rate $\eta_t = 5 \times 10^{-5}$ and $\epsilon_t = 1/t$.

4 Conclusion

We have proposed a new natural gradient method for score matching and demonstrated that it can avoid the plateau in learning of the multi-layer model and accelerate the convergence of learning.

In this study, we confirmed the effectiveness of our adaptive natural gradient method in the model with a relatively small number of parameters. Recently, deep networks with many more parameters have been developed and even our adaptive method may take much computational time and memory space. Fortunately, we expect that implementations suited to large scale problems (Pascanu & Bengio, 2013) such as metric-free optimization (Desjardins et al., 2013) inspired by Hessian-free optimization (Martens, 2010), the block diagonal approximation of the Fisher metric (Roux et al., 2008), or the method exploiting the structure of the metric on a graphical model (Grosse & Salakhudinov, 2015) will also be applicable to the natural gradient with the score matching metric. In addition, our framework to derive natural gradient will also be applicable to other divergences, particularly, to ratio matching, which is the extension of the score matching for discrete probabilistic variables (Hyvärinen, 2007; Dawid et al., 2012).

REFERENCES

Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10:251–276, 1998.

Shun-ichi Amari. In *Information Geometry and Its Applications*. Springer, 2016.

Shun-ichi Amari, Hyeyoung Park, and Kenji Fukumizu. Adaptive method of realizing natural gradient learning for multilayer perceptrons. *Neural Computation*, 12:1399–1409, 2000.

A Philip Dawid, Steffen Lauritzen, and Matthew Parry. Proper local scoring rules on discrete sample spaces. *The Annals of Statistics*, 40(1):593–608, 2012.

Guillaume Desjardins, Razvan Pascanu, Aaron Courville, and Yoshua Bengio. Metric-free natural gradient for joint-training of boltzmann machines. *arXiv preprint, arXiv:1301.3545*, 2013.

Shinto Eguchi. Second order efficiency of minimum contrast estimators in a curved exponential family. *The Annals of Statistics*, 11:793–803, 1983.

Roger Grosse and Ruslan Salakhudinov. Scaling up natural gradient by sparsely factorizing the inverse fisher matrix. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pp. 2304–2313, 2015.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. In *Journal of Machine Learning Research*, pp. 695–709, 2005.

Aapo Hyvärinen. Some extensions of score matching. *Computational statistics & data analysis*, 51 (5):2499–2512, 2007.

Urs Köster and Aapo Hyvärinen. A two-layer model of natural stimuli estimated with score matching. *Neural Computation*, 22:2308–2333, 2010.

James Martens. Deep learning via hessian-free optimization. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 735–742, 2010.

Yann Ollivier. Riemannian metrics for neural networks I: feedforward networks. *Information and Inference*, 4(2):108–153, 2015.

Hyeyoung Park, Shun-ichi Amari, and Kenji Fukumizu. Adaptive natural gradient learning algorithms for various stochastic models. *Neural Networks*, 13(7):755–764, 2000.

Razvan Pascanu and Yoshua Bengio. Revisiting natural gradient for deep networks. *arXiv preprint arXiv:1301.3584*, 2013.

Nicolas L Roux, Pierre-Antoine Manzagol, and Yoshua Bengio. Topmoumoute online natural gradient algorithm. In *Advances in neural information processing systems*, pp. 849–856, 2008.

Kevin Swersky, MarcAurelio Ranzato, David Buchman, Benjamin M Marlin, and Nando de Freitas. On autoencoders and score matching for energy based models. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 1201–1208, 2011.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.