ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0.
Available from http://www.i6doc.com/en/.

Noise helps optimization escape from saddle points in the neural dynamics

Ying Fang, Zhaofei Yu, Feng Chen *

Tsinghua University - Department of Automation
Center for Brain-Inspired Computing Research
Beijing 100084 - China

Abstract. Synaptic connectivity in the brain is thought to encode the long-term memory of an organism. But experimental data point to surprising ongoing fluctuations in synaptic activity. Assuming that the brain computation and plasticity can be understood as probabilistic inference, one of the essential roles of noise is to efficiently improve the performance of optimization in the form of stochastic gradient descent. The strict saddle condition for synaptic plasticity is deduced and under such condition noise can help escape from saddle points on high dimensional domains. The theoretical result explains the stochasticity of synapses and guides us how to make use of noise. Our simulation results manifest that in the learning and test phase, the accuracy of synaptic sampling is almost 20% higher than that without noise.

1 Introduction

Quite a number of experiments show that noise plays diverse roles in neural system [1]. For example, noise can improve the quality of measurements (signal-to-noise ratio, mutual information, coherence, etc.), such as the form of 'stochastic facilitation' [2]. However, some observations haven't been explained, for example, intrinsic noise can switch a neuron from one stable pattern to another [3]. The puzzling results show that the benefits of noise still need a careful investigation.

General results from statistical learning theory suggest that both brain computations and brain plasticity should be understood as probabilistic inference [4, 5]. These results have provided insight into how noise plays an essential role in the networks of spiking neurons. Maass et al. [6] proposed that knowledge can be stored in probabilistic distributions of network states and noise enable networks of spiking neurons to carry out probabilistic inference through sampling. Based on Boltzmann machines, they modified neurons to generate a spike with probability. This stochasticity is linked to the sampling method. Moreover, it helps network perform probabilistic inference. Kappel et al. [7] presents a new theoretical framework for analyzing and understanding local plasticity mechanisms in the brain as stochastic processes. Instead of stochastically spiking of

*This work is supported in part by the National Natural Science Foundation of China under Grant 61671266, 61327902, in part by the Research Project of Tsinghua University under Grant 20161080084, and in part by National High-tech Research and Development Plan under Grant 2015AA042306.

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

neurons [6], the noise results from the fluctuation of synapses, which endows networks to automatically compensate for internal and external changes.

But in theoretical neuroscience research, the computational benefits of noise to networks of spiking neurons has rarely been addressed, the reason why our brain benefits from noise has not been verified and the method of making full use of noise has also not been proposed.

In this paper, we attempt to answer the above three problems. First, we propose that one of the essential roles of noise in the brain computations and brain plasticity is to efficiently improve the performance of optimization. Noise helps optimization escape from bad stable points. Second, we theoretically prove why noise benefits optimization in the brain. The main bottleneck in optimization is that gradient updates are trapped in exponentially many saddle points instead of local minima [8, 9]. Under the so-called strict saddle property, gradient decent with noise will escape from the bottleneck and leads to the efficient optimization [10]. We prove that strict saddle condition is satisfied for synaptic plasticity. Third, we show that noisy networks for which the synaptic weights affect the noise variance have better learning performances.

2 Synaptic sampling model and Saddle point problem

Network plasticity by maximum the likelihood has been studied in many ways. It adjusts synaptic parameters $\boldsymbol{\theta}$ to maximize the fit of the network to inputs $\boldsymbol{x}$. However, the model tends to produce overfitting, thereby reducing generalization capabilities. Furthermore, without any prior distribution, it respond slowly to perturbations. The solution to such challenge is how posterior distribution of weights can be represented and learned in neural dynamics. Based on stochastic differential equation, Maass et al. [7] solve this challenge by sampling from posterior distribution $p_N(\boldsymbol{\theta}|\boldsymbol{x})$. This model defined by Eq.(1) is referred to as synaptic sampling.

$$d\theta_{ki} = b\left(\frac{\partial \log p_S(\boldsymbol{\theta})}{\partial \theta_{ki}} + \frac{\partial \log p_N(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \theta_{ki}}\right)dt + \sqrt{2b}dW_{ki} \tag{1}$$

However, when this model is used for classification with a standard Gaussian noise, it is difficult to find a reasonable minimum. Because there are many saddle points on the high error plateaus [8]. Gradient based algorithms are particularly sensitive to saddle point problems as they only depend on gradient information. Furthermore, they only understand noise as a functional aspect for learning because it helps the network sample from posterior distributions. We argue that noise plays a more important role in the optimization process. We propose a sufficient condition that noise should satisfy in order to improve the optimization process.

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

Table 1: Definitions of the main mathematical symbols used in this paper

$\mathbf{x^n}$	vector of the n_{th} input variables $\{x_1^n, ..., x_I^n\}$
$\mathbf{z^n}$	vector of the n_{th} output variables $\{z_1^n, ..., z_K^n\}$
$\mathbf{h^n}$	vector of the label $\{h_1^n, ..., h_K^n\}$
$\mathbf{w}$	vector of all synaptic weights $w_{ki} = exp(\theta_{ki} - \theta_0)$
$\boldsymbol{\theta}$	vector of all synaptic parameters $\{\theta_{ki}, k \leq K, i \leq I\}$
$p_S(\boldsymbol{\theta})$	structural constraints following $\mathcal{N}(\mu, \sigma^2)$
$p_N(J_{max}\|\boldsymbol{\theta})$	likelihood function with the form of cross entropy $\log p_N(J_{max}\|\boldsymbol{\theta}) = \sum_{n=1}^{N}\sum_{k=1}^{K}\Theta\{z_k^n\}\log p(z_k^n\|\boldsymbol{x^n}, \boldsymbol{\theta})$
$p_N(\boldsymbol{x^n}\|\boldsymbol{\theta})$	Poissonian distributions of spikes parameterized by $\alpha e^{w_{ki}}$
dW_{ki}	stochastic time course of the parameter θ_{ki}
$\Theta(z_k^n)$	Heaviside step function
$S_k(t)$	the spike train of the neuron z_k

3 'Strict saddle' condition for synaptic sampling

Recently, Rong Ge et al.[10] identify a 'strict saddle' condition, which guarantees stochastic gradient decent can escape from the saddle points quickly (see [10] Theorem 6). Note that a twice differentiable function $f(w)$ is strict saddle, if all its local maxima have $\nabla^2 f(w) < 0$ and all its other stationary points satisfy $\lambda_{max}(\nabla^2 f(w)) > 0$.

We study the effect of noise on the synaptic sampling defined in Eq.(1) for classification. As Fig.1 shows, input neurons tune n_{th} stimulus to 200-ms long spiking activities $\boldsymbol{x^n}$ according to tuning curves. Synaptic sampling is then applied to $K\mathbf{x}I$ synapses. In the Winner-Take-All(WTA) circuit , the output is a 200-ms spiking pattern $\boldsymbol{z^n}$ and the neuron which spikes most indicates the possible label. The learning goal in Eq.(1) becomes the posterior distribution $p^*(\boldsymbol{\theta}|J_{\max})$defined by $p_S(\boldsymbol{\theta}) * p_N(J_{\max}|\boldsymbol{\theta})$. $p_N(J_{\max}|\boldsymbol{\theta})$ measures the degree of network fitting to the classification. The detailed definition is shown in Table 1. The synaptic sampling rule Eq.(1) yields for this model.

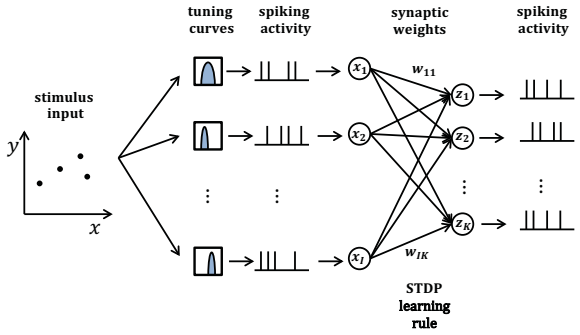

Fig. 1: Architecture of the networks whose dynamics are modeled by Eq.(2)

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

$$d\theta_{\mathrm{ki}} = b\left(\frac{1}{\sigma^2}\left(\mu-\theta_{\mathrm{ki}}\right)+\sum_{n=1}^{N} w_{\mathrm{ki}}\left(x_i^n-\alpha e^{w_{\mathrm{ki}}}\right)\left(\Theta\left\{h_k^n\right\}-S_k\left(t\right)\right)\right)dt + bdW_{ki} \quad (2)$$

where the component $\left(x_i^n-\alpha e^{w_{\mathrm{ki}}}\right)\left(\Theta\left\{h_k^n\right\}-S_k\left(t\right)\right)$ of likelihood differential term is a simplified version of STDP (spike-timing dependent plasticity).

We show that when the noise takes a certain form, synaptic sampling network for classification satisfies 'Strict saddle' condition and leads to efficient optimization. Note that if the noise is just standard normal distribution which is a popular choice for stochastic differential equation, the network will not satisfy such property.

Theorem 1 (sufficient condition) *Given random samples X such that output $Y=\mathbb{E}(g(X))\in\mathbb{R}^K$ is spike trains for classification in the synaptic sampling network, there is a loss function $f\left(w\right)=\mathbb{E}(\phi(w,X))\mathbf{w}\in\mathbb{R}^{K\mathbf{x}I}$ such that every local minimum corresponds to valid output Y. Further, if the noise satisfies Eq.(3), function f is strict saddle.*

$$d\mathcal{W}_i=\mathcal{N}(0,N\alpha e^{w_{ki}})dt \quad (3)$$

Proof sketch of Theorem In order to prove theorem, we need to prove $\lambda_{\max}(\nabla^2 f(w))>0$. We prove a sufficient condition, i.e. $\sum\lambda\left(\nabla^2 f\left(w\right)\right)>0$. According to the equation about trace of matrix M: $tr(M)=\sum\lambda$, we get ,

$$\sum\lambda\left(\nabla^2 f\left(w\right)\right)=\sum_k\sum_i\nabla^2 f(\theta_{ki})=-\frac{KI}{\sigma^2}+\sum_k\sum_i\frac{1}{\sigma^2}(\theta_{ki}-\mu) + \sum_n\sum_k\sum_i w_{ki}^2\{\alpha e^{w_{ki}}(S_k(t))-1\{h^n=k\}\} \quad (4)$$

It is obvious that $\sum\lambda\left(\nabla^2 f\left(w\right)\right)$ consists of three terms: $A=-\frac{KI}{\sigma^2}$, $B=\sum_k\sum_i\frac{1}{\sigma^2}(\theta_{ki}-\mu)$, $C=\sum_n\sum_k\sum_i w_{ki}^2\{\alpha e^{w_{ki}}(S_k(t))-1\{h^n=k\}\}$. We need to prove$\sum\lambda\left(\nabla^2 f\left(w\right)\right)=A+B+C>0$. The proof is divided into three steps. Note that the first sentence of each step below is the conclusion we want to prove.

1) $B\ll C$. Only when the noise $d\mathcal{W}_i=\mathcal{N}(0,N\alpha e^{w_{\mathrm{ki}}})dt$, we can derive that $B+\left(x_i^n-\alpha e^{w_{\mathrm{ki}}}\right)C$ is a variant of the gradient. According to the zero gradient and STDP learning rule, $\frac{B}{C}\approx 0$, thereby B can be ignored.

2) C is positive. $C\approx N(\sum_i w_{\mathrm{ki}}^2 x_i-\sum_i w_{label,i}^2 x_i)$ which represents the approximate potential difference of actual and expected neurons. When the network is trapped in saddle points, the neuron which releases spikes is not the expected. Thus, potential of the actual neuron is higher than the expected.

3) $A+B+C>0$. A is negative constant. When N is greater than a certain value, C is large enough so that $A+B+C>0$ and strict saddle property will be satisfied.

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0.
Available from http://www.i6doc.com/en/.

The theorem are therefore proved. That is to say, we apply Theorem 1 to make the synaptic sampling network satisfy strict saddle condition and by Theorem 6 in [10] we know that noise will help escape from saddle points.

4 Network Simulations

In the simulations, we use a cluster of points in 3D space to represent one sensory experience for visualization. We present 43200 samples to the network during 2.4 hours to deal with 10-categories classification. Through the tuning curves of 1000 input neurons, 200ms spike patterns were communicated to synaptic sampling network for each sample. According to Eq.(2) and spike-based update scheme, the sensory experiences were presented sequentially and all synapses were updated sequentially. The final predicted label is the neuron which fires most between the 10 output neurons. We repeat the simulation 10 runs and the accuracy is average over 10 runs. The results are shown in Fig.2. In the learning process, responses were unspecific to different inputs initially and at 1700 s responses had become obvious preferences for different inputs. Synaptic sampling without noise can not learn this task accurately. In the learning and test phase, the accuracy of synaptic sampling is almost 20% higher than that without noise.

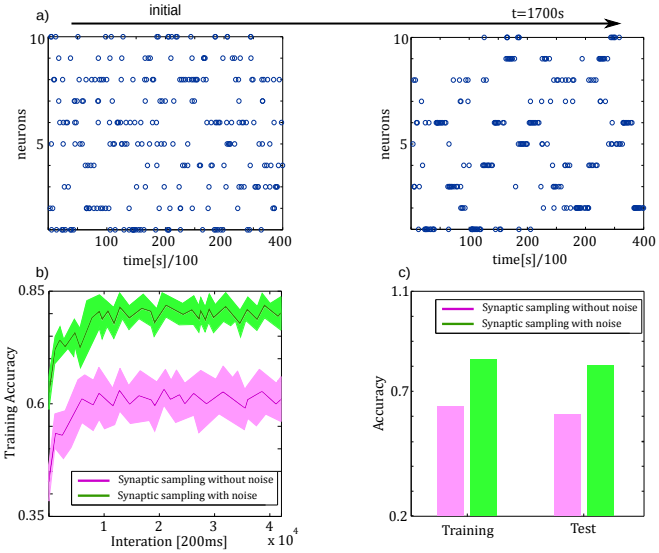

Fig. 2: a) The development of readout spikes in the course of learning. b) Learning curves of synaptic sampling with/without noise. c) Accuracy comparison in the learning and testing phase.

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

5 Conclusion

In this paper, we propose that the essential role of noise in the brain computations and brain plasticity is to efficiently improve the performance of optimization. First, based on the synaptic sampling, on the one hand, we can formalize the noise in the network theoretically by the stochastic term dW_i; on the other hand, the brain computations and brain plasticity with noise is separately converted to the Bayesian inference and differentiation of the likelihood. Next, through the use of sufficient condition, the evaluation of maximum eigenvalue is skillfully converted to the trace of Hessian matrix. Then, by inducing the strict saddle property, we prove that the plasticity networks with noise in Eq.(3) satisfy such property so that noise will help escape from exponentially many saddle points. We propose a sufficient condition for improving optimization through noise: as long as the noise meets a certain criterion, synaptic sampling network satisfies 'Strict saddle' condition and the optimization will be more efficient. In our further work, the form of noise will be explained from the perspective of on-line learning. On-line learning is to use one sample to estimate the expected gradient. It is a popular choice because it is usually faster and better than full-gradient methods. We have proved that synaptic sampling in Eq.(2) is equal to on-line synaptic sampling. We also apply our noise to three-layer network with back propagation algorithm and explore the meaning of noise parameter from a new perspective of view.

References

[1] A. Aldo Faisal, Luc P. J Selen, and Daniel M Wolpert. Noise in the nervous system. *Nature Reviews Neuroscience*, 9(4):292–303, 2008.

[2] Toshio Mori and Shoichi Kai. Noise-induced entrainment and stochastic resonance in human brain waves. *Physical review letters*, 88(21):218101, 2002.

[3] J. M. Fellous, P. H. Tiesinga, P. J. Thomas, and T. J. Sejnowski. Discovering spike patterns in neuronal responses. *J. Neurosci.*, 24(12):2989–3001, Mar 2004.

[4] David C Knill and Alexandre Pouget. The bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences*, 27(12):712–719, 2004.

[5] Alexandre Pouget, Jeffrey M Beck, Wei Ji Ma, and Peter E Latham. Probabilistic brains: knowns and unknowns. *Nature Neuroscience*, 16(9):1170–1178, 2013.

[6] Wolfgang Maass. Noise as a resource for computation and learning in networks of spiking neurons. *Proceedings of the IEEE*, 102(5):860–880, 2014.

[7] David Kappel, Stefan Habenschuss, Robert Legenstein, and Wolfgang Maass. Network plasticity as Bayesian inference. *PLoS Computational Biology*, 11(11):e1004485, 2015.

[8] Yann Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. *Mathematics*, 111(6 Pt 1):2475–2485, 2014.

[9] Yan V Fyodorov and Ian Williams. Replica symmetry breaking condition exposed by random matrix calculation of landscape complexity. *Journal of Statistical Physics*, 129(5-6):1081–1116, 2007.

[10] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points-online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.