Supplementary Material

1 PROOF OF THEOREM 1

The sufficient condition of strict saddle property is $\lambda_{\text{max}}(\nabla^2 f(\theta)) > 0$. It is convenient to use a more strong condition, i.e., $\sum \lambda(\nabla^2 f(\theta)) > 0$. According to the equation about trace of $n \times n$ matrix $A: \text{tr}(A) = \lambda_1 + \ldots + \lambda_n$, we concentrate on the diagonal elements of the Hessian matrix. According to the chain rule, we convert the derivative of $\theta$ to $w$. We get that $\nabla f(\theta) = 0 \iff \nabla f(w) = 0$ and $\sum_k \sum_i \nabla^2 f(\theta_{ki}) \geq 0 \iff \sum_k \sum_i (\nabla^2 f(w_{ki}))(w_{ki})^2 \geq 0$.

In the synaptic sampling neural networks, the objective function $f(\theta)$ defined by $\log p_\ast(\theta|J)$, where

$$\log p_\ast(\theta) = -\frac{(x - \mu)^2}{2\sigma^2} - \log \sqrt{2\pi\sigma} \quad (S1)$$

$$\log p_N(J|\theta) = \sum_{n=1}^N \sum_{k=1}^K \Theta\{h^n_k\} log p(z^n = k|x^n, w) \quad (S2)$$

Assuming that $A_{ki} = \text{POISSON}(x^n_i|\alpha e^{w_{ki}}) = (\alpha e^{w_{ki}})^{x^n_i} e^{-\alpha e^{w_{ki}}}/x^n_i!$, we get $p_N(x^n|z^n = k, w) = \prod_j \text{POISSON}(x^n_j|\alpha e^{w_{ki}}) = \prod_j A_{kj}$. Therefore,

$$\frac{\partial}{\partial w_{ki}} (\log p(x^n|k, w)) = \frac{1}{\prod_j A_{kj}} \prod_j A_{kj} (x_i^n - \alpha e^{w_{ki}}) = x_i^n - \alpha e^{w_{ki}} \quad (S3)$$

$$\frac{\partial}{\partial w_{ki}} (\log \Sigma q p(x^n|q, w)) = \frac{\prod_j A_{kj} (x_i^n - \alpha e^{w_{ki}})}{\Sigma q p(x^n|q, w)} = \frac{p(x^n|k, w)}{\Sigma q p(x^n|q, w)} (x_i^n - \alpha e^{w_{ki}}) \approx p(k|x, W) (x_i^n - \alpha e^{w_{ki}}) \quad (S4)$$

We further get

$$\frac{\partial}{\partial w_{ki}} \log p_N(J|\theta) = \frac{\partial}{\partial w_{ki}} \sum_{n=1}^N \sum_{k=1}^K \Theta\{h^n_k\} \log \frac{p(x^n|z^n = k, w)}{\Sigma q p(x^n|z^n = q, w)}$$

$$= \sum_{n=1}^N \Theta\{h^n_k\} \frac{\partial}{\partial w_{ki}} (\log p(x^n|k, w) - \log \Sigma_k p(x^n|k, w))$$

$$= \sum_{n=1}^N \Theta\{h^n_k\} (x_i^n - \alpha e^{w_{ki}}) (1 - p(k|x, W))$$

$$\approx \sum_{n=1}^N (x_i^n - \alpha e^{w_{ki}}) (\Theta\{h^n_k\} - S_k(t)) \quad (S5)$$
\[
\frac{\partial^2 \log p_N(J | \theta)}{\partial^2 w_{ki}} \\
= \frac{N}{\sum_{n=1}^{N} \partial (x_i^n - \alpha e^{w_{ki}}) (\Theta \{ h_k^n \} - p (k | x, W))}{\partial w_{ki}} \\
= \sum_{n=1}^{N} \left\{ -\alpha e^{w_{ki}} (\Theta \{ h_k^n \} - p (k | x, W)) - (x_i^n - \alpha e^{w_{ki}})^2 p (k | x, W) (1 - p (k | x, W)) \right\}
\]

In the spike-based Winner-Take-All networks, give the input \( x^n \) and the firing rate of neuron \( z_k \) is proportional to the posterior distribution \( p (h^n = k | x^n, w) \). In simulation, the network generates spike trains \( S_k(t) \) with such rate \( \rho_k(t) \) of neuron \( z_k \). According to \( \frac{\partial^2 \log p_N}{\partial^2 \theta_{ki}} = \left[ \frac{\partial \log p_N}{\partial w_{ki}} + w_{ki} \frac{\partial^2 \log p_N}{\partial^2 w_{ki}} \right] w_{ki} \),

\[
\frac{\partial^2 \log p_N(J | \theta)}{\partial^2 \theta_{ki}} = \sum_{n=1}^{N} w_{ki} (x_i^n - \alpha e^{w_{ki}}) (\Theta \{ h_k^n \} - S_k(t)) \\
+ \sum_{n=1}^{N} w_{ki}^2 \left\{ -\alpha e^{w_{ki}} (\Theta \{ h_k^n \} - S_k(t)) - (x_i^n - \alpha e^{w_{ki}})^2 S_k(t) (1 - S_k(t)) \right\}
\]

Due to the Gauss property \( p \{ | x - \mu | < \sigma \} = 0.6826 \) and \( \frac{\partial w_{ki} e^{w_{ki}}}{\partial e^{w_{ki}}} \bigg|_{w_{ki} \rightarrow 0} = 0 \), it is plausible to refer to \( \sum_n \alpha w_{ki} e^{w_{ki}} \left( S_k(t) - \Theta \{ h_k^n \} \right) dt \) as the general characteristic of the noise distribution \( dW_{ki} \). We further get the Hessian information with the noise,

\[
\nabla^2 \tilde{f} (\theta_{ki}) = \frac{\partial^2 \log p_S (\theta)}{\partial^2 \theta_{ki}} + \frac{\partial^2 \log p_N(J | \theta)}{\partial^2 \theta_{ki}} + \frac{\partial W_{ki}}{\partial \theta_{ki}}
\]

\[
= -\frac{1}{\sigma^2} + \sum_{n=1}^{N} w_{ki} (x_i^n - \alpha e^{w_{ki}}) (\Theta \{ h_k^n \} - S_k(t)) \\
+ \sum_{n=1}^{N} w_{ki}^2 \left\{ -\alpha e^{w_{ki}} (\Theta \{ h_k^n \} - S_k(t)) - (x_i^n - \alpha e^{w_{ki}})^2 S_k(t) (1 - S_k(t)) \right\}
+ \sum_{n=1}^{N} (w_{ki} \alpha e^{w_{ki}} + w_{ki}^2 \alpha e^{w_{ki}}) (S_k(t) - \Theta \{ h_k^n \})
\]

\[
= -\frac{1}{\sigma^2} + \sum_{n=1}^{N} w_{ki} (x_i^n - \alpha e^{w_{ki}}) (\Theta \{ h_k^n \} - S_k(t)) + \sum_{n=1}^{N} w_{ki} \alpha e^{w_{ki}} (S_k(t) - \Theta \{ h_k^n \})
\]

According to the zero gradient, we can further get,

\[
\sum_k \sum_i \nabla^2 \tilde{f} (\theta_{ki}) = -\frac{K I}{\sigma^2} + \sum_k \sum_i \frac{1}{\sigma^2} (\theta_{ki} - \mu) + \sum_k \sum_i \sum_n w_{ki} \alpha e^{w_{ki}} (S_k(t) - \Theta \{ h_k^n \})
\]
It is obvious that the trace of Hessian matrix consists of three terms: \(-K_1\sigma^2, \sum_k \sum_i \frac{1}{\sigma^2}(\theta_{ki} - \mu), \sum_n \sum_k \sum_i w_{ki} \alpha e^{w_{ki}} (S_k(t) - \Theta(h_k^n))\). Whether the trace is positive will be analyzed through two steps.

First, we will illustrate that the second term is much smaller than the third. When the network is stable, i.e., the gradient is zero, by considering a large number of samples, we will get:

$$\sum_k \sum_i \frac{(\theta_{ki} - \mu)}{N \sigma^2 w_{ki} \alpha e^{w_{ki}} (S_k(t) - \Theta(h_k^n))} = \sum_k \sum_i \frac{1}{N} \sum_n x_i^n - \alpha e^{w_{ki}}$$  \hspace{1cm} (S10)

Under the theoretically optimal STDP learning rule, the following equation will be derived \cite{Habenschuss2013, Nessler2009}:

$$\alpha e^{w_{ki}} < x_i > \tilde{p}(x|z_k=1,w) = 1$$  \hspace{1cm} (S11)

This means that \(\sum_n x_i^n - \alpha e^{w_{ki}} \approx 0\). Through the above two points, we can derive that

$$\sum_k \sum_i \frac{(\theta_{ki} - \mu)}{N \sigma^2 w_{ki} \alpha e^{w_{ki}} (S_k(t) - \Theta(h_k^n))} \approx \sum_k \sum_i (\theta_{ki} - \mu) \frac{1}{N \sigma^2 w_{ki} \alpha e^{w_{ki}} (S_k(t) - \Theta(h_k^n))} \approx 0$$  \hspace{1cm} (S12)

That is to say, the second term is much smaller than the third: \(\frac{1}{\sigma^2} \sum_k \sum_i (\theta_{ki} - \mu) \ll N \sum_k \sum_i w_{ki} \alpha e^{w_{ki}} (S_k(t) - \Theta(h_k^n))\).

Second, we will discuss whether the third term is positive. For each sample in the Winner-Take-All networks, the third term can be split into two parts: one part represents neuron \(z_k\) which actually releases spikes, one part represents expected neuron \(z_{\text{label}}\):

$$N(\sum_i w_{ki} \alpha e^{w_{ki}} - \sum_i w_{\text{label},i} \alpha e^{w_{\text{label},i}})$$  \hspace{1cm} (S13)

According to the theoretically optimal STDP learning rule Eq.\text{S11}, the expression will become the approximate difference of actual and expected neurons: \(N(\sum_i w_{ki} x_i - \sum_i w_{\text{label},i} x_i)\). When the network is trapped in the saddle points, the neuron which releases spikes is not the expected. We can get that the potential of the actual neuron is higher than the expected: \(u_k > u_{\text{label}} \Rightarrow \sum_i w_{ki} x_i > \sum_i w_{\text{label},i} x_i\). As a result, the third term is always positive when the network is trapped in the saddle points.

In summary, as long as the third term is large enough, the first and second term can be ignored, and the trace of Hessian matrix will be positive so that strict saddle property will be satisfied.

2 PROOF OF THEOREM 2  

Under the condition that \(x|\theta \sim \mathcal{N}(\mu, \sigma^2)\) and noise \(\varepsilon \sim \mathcal{N}(0,1)\), the aim is to get \(p(\varepsilon|\theta)\). By inducing hidden variable \(x\), we get

$$p(\varepsilon|\theta) = \int p(\varepsilon|x) p(x|\theta) dx$$  \hspace{1cm} (S14)
Given the input $x$, the distribution of noise $\varepsilon$ can be regarded as that of a mixture of noise and input. We get

$$p(\varepsilon|\theta) \propto \int p(\varepsilon + x) \ p(x|\theta)dx \quad (S15)$$

According to Taylor expansion $p(x + \varepsilon) = p(x) + \varepsilon p'(x) + \frac{1}{2} \varepsilon^2 p''(x) + o(\varepsilon^3)$, where $p'(x|\theta) = -\frac{x}{\sigma^2} p(x|\theta)$, $p''(x|\theta) = -p(x|\theta) \left( \frac{1}{\sigma^2} - \frac{x^2}{\sigma^4} \right)$, we get

$$p(\varepsilon|\theta) \propto \int (p(x|\theta) + \varepsilon p'(x|\theta) + \frac{1}{2} \varepsilon^2 p''(x|\theta)) \ p(x|\theta)dx \quad (S16)$$

As Taylor approximation $e^{\frac{\varepsilon x}{\sigma^2} - \frac{\varepsilon^2}{\sigma^2} \left( \frac{1}{\sigma^2} - \frac{x^2}{\sigma^4} \right)} = 1 - \frac{\varepsilon x}{\sigma^2} - \frac{\varepsilon^2}{2} \left( \frac{1}{\sigma^2} - \frac{x^2}{\sigma^4} \right)$, we get

$$p(\varepsilon|\theta) \propto \int e^{-\frac{\varepsilon^2 x^2}{2\sigma^2} - \frac{\varepsilon^2}{2\sigma^4} + \frac{x^2}{\sigma^4} + \frac{\varepsilon^2}{2\sigma^2} \left( \frac{1}{\sigma^2} - \frac{x^2}{\sigma^4} \right)} dx \quad (S17)$$

The exponential term can be arranged as

$$-\frac{\varepsilon^2 x^2}{2\sigma^2} - \frac{x^2}{\sigma^2 - \varepsilon^2} + \frac{\varepsilon^2}{2\sigma^4} \left( \frac{\sigma^2 - \varepsilon^2}{\sigma^4 - \varepsilon^4} \right)^2 + \frac{\varepsilon^2}{2(\sigma^2 - \varepsilon^2)} \quad (S18)$$

The second term is binomial of $x$ and in fact is some constant with the integration of $x$. The numerator and denominator of the last term is the same order of $\varepsilon$ and can also regard as some constant. We also verify the approximation between $e^{\frac{\varepsilon^2 x^2}{2\sigma^2}}$ and $e^{-\frac{\varepsilon^2}{2\sigma^2} + \frac{x^2}{\sigma^4} + \frac{\varepsilon^2}{2\sigma^2} \left( \frac{1}{\sigma^2} - \frac{x^2}{\sigma^4} \right)}$ in Matlab simulations. Results also show that these two functions are nearly the same. We finally improve that

$$p(\varepsilon|\theta) \propto e^{-\frac{\varepsilon^2 x^2}{2\sigma^2}} \quad (S19)$$

The previous work (Habenschuss et al., 2013; Kappel et al., 2015) shows that in the spiked-based WTA networks, one prominent motif of cortical microcircuits, $p(x|\theta)$ is the integration of $N$ Poisson distribution with the mean $\alpha e^{w_{ki}}$, which can approximate Normal distribution $\mathcal{N}(N\alpha e^{w_{ki}}, N\alpha e^{w_{ki}})$. Therefore, we get,

$$\varepsilon|\theta \sim \mathcal{N}(0, N\alpha e^{w_{ki}}) \quad (S20)$$

3 BACK-PROPAGATION FOR SYNAPTIC SAMPLING ON THE THREE-LAYER NETWORK.

In this section we derive learning rules based on back-propagation for synaptic sampling. As shown in Fig. S1, we add one hidden layer $z^n$ to the two-layer network. In the two-layer model, given $y^n = k$, input $x^n_i$ follows a Poisson distribution with a mean that is affected by synaptic weight
The firing rate is proportional to the posterior probability $p(z^n = k | x^n, w)$. In the threelayer network, such a relationship becomes the product of Poisson distribution. Assuming that $A_{kj} = POISSON(y^n_j | \alpha e^{w_{kj}})$ $B_{ji} = POISSON(x^n_i | \alpha e^{w_{ji}})$, one obtains the posterior probability of corresponding spiking neuron $k$ as

$$p(z^n = k | x^n, w) = \frac{\prod_i \sum_j A_{kj} B_{ji}}{\sum_k \prod_i \sum_j A_{kj} B_{ji}} \quad (S21)$$

The likelihood function becomes

$$\log p_N(J | \theta) = \sum_{n=1}^{N} \sum_{k=1}^{K} \Theta\{h^n_k\} \log p(z^n = k | x^n, w)$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} \Theta\{h_k^n\} \log \frac{\prod_i \sum_j POISSON(x^n_i | \alpha e^{w_{ji}}) POISSON(y^n_j | \alpha e^{w_{kj}})}{\sum_k \prod_i \sum_j POISSON(x^n_i | \alpha e^{w_{ji}}) POISSON(y^n_j | \alpha e^{w_{kj}})} \quad (S22)$$

The derivation of synaptic weight $w_{kj}$ of the second layer is

$$\frac{\partial \log p_N(J | \theta)}{\partial w_{kj}} = \sum_{n=1}^{N} \frac{\partial (\log \prod_i \sum_j A_{kj} B_{ji} - \log \sum_k \prod_i \sum_j A_{kj} B_{ji})}{\partial w_{kj}} \quad (S23)$$

In Eq. S21, the deviation of numerator and denominator is

$$\frac{\partial \prod_i \sum_j A_{kj} B_{ji}}{\partial w_{kj}} = B_{j1} A_{kj} \prod_{i \neq 1} \sum_j A_{kj} B_{ji} (y^n_j - \alpha e^{w_{kj}})$$

$$+ B_{j2} A_{kj} \prod_{i \neq 2} \sum_j A_{kj} B_{ji} (y^n_j - \alpha e^{w_{kj}}) + \ldots$$

$$\approx M \prod_i \sum_j A_{kj} B_{ji} (y^n_j - \alpha e^{w_{kj}}) \quad (S24)$$
\[
\frac{\partial \sum_k \prod_i \sum_{z_i} A_{kj} B_{ji}}{\partial w_{kj}} = \frac{\partial \prod_i \sum_j A_{kj} B_{ji}}{\partial w_{kj}}
\]  
(S25)

When \( M \) is large enough, derive the log of numerator and denominator and we obtain

\[
\frac{\partial \log \prod_i \sum_j A_{kj} B_{ji}}{\partial w_{kj}} \approx M \left( y_j^n - \alpha e^{w_{ki}} \right)
\]  
(S26)

\[
\frac{\partial \log \sum_k \prod_i \sum_j A_{kj} B_{ji}}{\partial w_{kj}} \approx M p \left( z^n = k \mid x^n, w \right) \left( y_j^n - \alpha e^{w_{ki}} \right)
\]  
(S27)

Learning rule for the second layer is derived as

\[
d\theta_{kj} = \frac{1}{\sigma^2} \left( \mu - \theta_{ki} \right) + \sum_{n=1}^{N} (w_{kj} \left( y_j^n - \alpha e^{w_{ki}} \right) \left( \Theta \{ h_k^n \} - S_k (t) \right)) \ dt + b dW_{kj}
\]  
(S28)

The derivation of synaptic weight \( w_{ji} \) of the first layer is

\[
\frac{\partial \log p_N (J \mid \theta)}{\partial w_{ji}} = \sum_{n=1}^{N} \frac{\partial \log p_N (J \mid \theta)}{\partial y_j^n} \frac{\partial y_j^n}{\partial w_{ji}}
\]  
(S29)

First, we derive the second part \( \frac{\partial z_j^n}{\partial w_{ji}} \). In the spike-based neural networks, the firing rate of stochastic spiking neurons depends exponentially on the membrane voltage. It has been proposed that the exponential relationship between the membrane potential and the firing rate is a good approximation to the firing characteristics of cortical pyramidal neurons (Jolivet et al., 2006). The membrane voltage \( u_j(t) \) of neuron \( j \) in the two-layer network is given by

\[
u_j^n = \sum_i w_{ji}^1 x_i^n + b_1^j
\]  
(S30)

The corresponding instantaneous firing rate \( \rho_j(t) \) of neuron \( j \) is given by

\[
\rho_j(t) = \frac{\rho_{\text{net}}}{I_{\text{lat}}} e^{u_j(t)}
\]  
(S31)

where \( I_{\text{lat}}(t) \) is divisive lateral inhibition, i.e. \( I_{\text{lat}}(t) = \sum_t e^{u(t)} \). Then we get,

\[
\frac{\partial y_j^n}{\partial w_{ji}} = \frac{\partial \rho_j(t)}{\partial w_{ji}} = \rho_{\text{net}} \left( \sum_{t \neq k} e^{u(t)} \right) \frac{\partial e^{u_j(t)}}{\partial w_{ji}}
\]  
\( \propto \frac{\rho_{\text{net}} e^{u_j(t)}}{I_{\text{lat}}} \frac{\partial u_j(t)}{\partial w_{ji}} = \rho_j(t) x_i^n \)  
\( \propto x_i^n \)
In the WTA neural circuit, when neuron $j$ spikes, $y^n_j(t) = 1$. Thus we get the approximation of last term. Then we derive the first part $\frac{\partial \log p_N(J|\theta)}{\partial y^n_j}$. Under the theoretically optimal STDP learning rule, the following equation will be derived (Habenschuss et al., 2013; Nessler et al., 2009):

$$E(y^n_j) = \alpha e^{w_{kj}}$$  \hspace{1cm} (S33)

We can get

$$\frac{\partial \log p_N(J|\theta)}{\partial y^n_j} = \sum_{m,q} \frac{\partial \log p_N(J|\theta)}{\partial w_{mq}} \frac{\partial w_{mq}}{\partial y^n_j} = \sum_m \frac{\partial \log p_N(J|\theta)}{\partial w_{mj}} \frac{\partial w_{mj}}{\partial y^n_j} = \frac{\alpha}{E(y^n_j)} \left( d\theta_{kj} + d\theta_{tj} \right)$$ \hspace{1cm} (S34)

The last equation in Eq. S34 holds because when spiking networks finish one weight update, only the actual neuron $i$ and expected neuron $k$ update corresponding weights in the WTA circuits. Substitute Eq. S32 and Eq. S34 to Eq. S29. Assuming that $\beta = \frac{1}{E(z^n_j)}$, apply the chain rule and we get,

$$d\theta_{ji} = \frac{\partial \log p_N(J|\theta)}{\partial w_{ji}} \frac{\partial w_{ji}}{\partial \theta_{ji}}$$  \hspace{1cm} (S35)

$$= \beta \alpha \sum_{n=1}^{N} x^n_i \left( \frac{d\theta_{kj}}{w_{kj}} + \frac{d\theta_{tj}}{w_{tj}} \right)$$

Note that the synaptic weights of two layers is about the same order of magnitude and hence cancel each other. We get the compact approximation for learning rule in the first layer,

$$d\theta_{ji} \approx \beta \alpha \sum_{n=1}^{N} x^n_i (d\theta_{kj} + d\theta_{tj})$$  \hspace{1cm} (S36)

In fact, the experiment results also show the performance of compact learning rules in Eq. S36 is as good as the exact one in Eq. S35.

Similarly, the learning rule for L-layer spiking neural networks can be concluded as follow. Given the L-layer noisy spiking neural networks, each layer computes a function $X^l = g_l(X^{l-1}, W^l)$, where $X^l$ is the output of the $l_{th}$ layer, $X^{l-1}$ is the input of the $l_{th}$ layer and $W^l$ is the vector of adjustable parameters between the $(l-1)_{th}$ and the $l_{th}$ layer. Note the vector $X^1$ in the first layer is the input sample. The learning rule for L-layer spiking neural networks can be concluded as follow.
\[ d\theta_{kj}^l = b \left( \frac{1}{\sigma^2} \left( \mu - \theta_{kj}^l \right) + \sum_{n=1}^{N} (w_{kj}^L x_j^{n,L-1} - \alpha e w_{kj}^L) \left( \Theta \{ h_k^n \} - S_k^L(t) \right) \right) dt + b dW_{kj}^L \] (S37)

\[ d\theta_{ji}^{l-1} = \beta \alpha \sum_{n=1}^{N} x_i^{n,l-1} \left( d\theta_{kj}^l + d\theta_{mj}^l \right) \quad (2 \leq l \leq L) \] (S38)

where \( d\theta_{kj}^l \) represents change in parameters corresponding to the neuron \( k \) which releases a spike and \( d\theta_{mj}^l \) represents change in parameters corresponding to the desired neuron \( m \).

REFERENCES

Habenschuss, S., Puhr, H., and Maass, W. (2013). Emergence of optimal decoding of population codes through STDP. Neural Computation 25, 1371–1407

Jolivet, R., Rauch, A., Lüscher, H.-R., and Gerstner, W. (2006). Synaptic modifications in cultured hippocampal neurons: Dependence on spike timing, synaptic strength, and postsynaptic cell type. Journal of Computational Neuroscience 21, 35–49

Kappel, D., Habenschuss, S., Legenstein, R., and Maass, W. (2015). Network plasticity as Bayesian inference. PLoS Computational Biology 11, e1004485

Nessler, B., Pfeiffer, M., and Maass, W. (2009). STDP enables spiking neurons to detect hidden causes of their inputs. In Advances in Neural Information Processing Systems 22 (Vancouver, British Columbia, Canada). 1357–1365