Tensor Decomposition by Modified BCM Neurons Finds Mixture Means Through Input Triplets

Matthew Lawlor
Applied Mathematics
Yale University
New Haven, CT 06520
matthew.lawlor@yale.edu

Steven W. Zucker
Computer Science
Yale University
New Haven, CT 06520
zucker@cs.yale.edu

1 Introduction

Learning and inference in sensory systems are difficult both because of the dimensionality of the input, and the high-order statistical dependencies between input dimensions. For example, edge selective neurons in visual cortex exhibit strong statistical dependencies due to the geometry of objects and their relationships in the world. "Hidden" information such as edge curvatures, the presence of textures, and lighting discontinuities all affect the probability distribution of firing rates among orientation selective neurons, leading to complex statistical interdependencies between neurons.

Latent variable models are powerful tools in this context. They formalize the idea that highly coupled random variables can be simply explained by a small number of hidden causes. Conditioned on these causes, the input distribution should be simple. For example, while the joint distribution of edges in a small patch of a scene might be quite complex, the distribution conditioned on the presence of a curved object at a particular location might be comparatively simple [8]. The question is whether brains can learn these mixture models, and how.

Example: Imagine a stimulus space of K inputs. These could be images of edges at particular orientations, or audio tones at K frequencies. These stimuli are fed into a network of n Linear-Nonlinear Poisson spiking neurons. Let r_{ij} denote the firing rate of neuron i to stimulus j. Assuming the stimuli are drawn independently with probability α_k, then the number of spikes $\boldsymbol{d}$ in an interval where a single stimulus is shown is distributed according to a mixture model.

$$P(\boldsymbol{d}) = \sum_k \alpha_k P_k(\boldsymbol{d})$$

where $P_k(\boldsymbol{d})$ is a vector of independent Poisson distributions, and the rate parameter of the ith component is r_{ik}. We would like to learn a filter which responds (in expectation) to one and only one stimuli. To do this, we must find a set of weights that are orthogonal to all but one of the vectors of rates $\boldsymbol{r}_{\cdot j}$. Each rate vector corresponds to the mean of one of the mixtures. Our problem is thus to learn the means of mixtures. We will demonstrate that this can be done non-parametrically over a broad class of firing patterns, not just Poisson spiking neurons.

The primary difficulty with fitting mixture models is computational. Although fitting many mixture models is often exponentially hard, recent work on tensor decompositions [1][2] has shown that under a certain multiview assumption, non-parametric estimation of mixture means can be done by tensor decomposition with relative ease. This multiview assumption requires that we have access to at least 3 *independent copies* of our samples. That is to say, we need multiple samples drawn from the same mixture component. For the LNP example above, this multiview assumption requires only that we have access to the number of spikes in three disjoint intervals, while the stimulus remains constant. After these intervals, the stimulus is free to change, at which point we sample again.

Our main result is that, with a slight modification of classical Bienenstock-Cooper-Munro [3] synaptic update rule a neuron can perform a tensor decomposition of the input data. By incorporating the

interactions between input triplets, our learning rule can provably learn the mixture means under an extremely broad class of mixture distributions and noise models. We note that the classical BCM learning rule will not converge properly in the presence of noise. This can be noise from the input itself, or noise due to randomness in the timing of individual spikes. Our learning rule does not suffer from this shortcoming, which makes it a much better candidate for learning selectivity in the natural environment.

Spike timing dependent plasticity has been used implement a variety of learning rules [10][6][4][12], including BCM. However, most of these approaches require much stronger distributional assumptions on the input data, or learn a much simpler decomposition of the data. Other, Baysian methods [9], require the computation of a posterior distribution with implausible normalization requirements. Our learning rule successfully avoids these issues, and has provable guarantees of convergence to the true mixture means.

The multiview assumption is a requirement for models such as this one, and it has an intriguing implication for neuroscience. We know that spikes arrive in waves from presynaptic neurons and that multiple spikes matter [5]. Normally one thinks of the information content in these spike trains [11]. However a different view arises from our model: *the waves of spikes arriving during adjacent epochs in time provide multiple samples of a given stimulus*. Technically we also require the assumption that these multiple samples are conditionally independent, for example they derive from retinal images with independent photon and receptor noise at adjacent time epochs.

The outline of the paper is as follows. First, we define our notation for tensors, and provide a few definitions. Second, we show how the classical BCM neuron performs gradient ascent in a tensor objective function, when the data consists of discrete input vectors. We show how to modify this objective function to perform an explicit tensor decomposition. Lastly, we show that we can dramatically relax the discrete input assumption to a general mixture model, assuming access to three independent samples from the mixture. We then derive a synaptic update rule that performs gradient ascent given these input triples.

2 Tensor Notation

Let $\otimes$ denote the tensor product. We denote application of a k-tensor to k vectors by $T(\boldsymbol{w}_1, ..., \boldsymbol{w}_k)$, so in the simple case where $T = \boldsymbol{v}_1 \otimes ... \otimes \boldsymbol{v}_k$,

$$T(\boldsymbol{w}_1, ..., \boldsymbol{w}_k) = \prod_j \langle \boldsymbol{v}_j, \boldsymbol{w}_j \rangle$$

We further denote the application of a k-tensor to k matrices by $T(M_1, ..., M_k)$ where

$$T(M_1, ..., M_k)_{i_1,...,i_k} = \sum_{j_1,...,j_k} T_{j_1,...,j_k} [M_1]_{j_1,i_1} ... [M_k]_{j_k,i_k}$$

Thus if T is a symmetric 2-tensor, $T(M_1, M_2) = M_1^T T M_2$ with ordinary matrix multiplication. Similarly, $T(\boldsymbol{v}_1, \boldsymbol{v}_2) = \boldsymbol{v}_1^T T \boldsymbol{v}_2$

We say that T has an orthogonal tensor decomposition if

$$T = \sum_k \alpha_k \boldsymbol{v}_k \otimes \boldsymbol{v}_k \otimes ... \otimes \boldsymbol{v}_k \quad \text{and} \quad \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle = \delta_i^j$$

2.1 Orthogonalizing non-orthogonal tensors

Let $T = \sum_k \alpha_k \mu_k \otimes \mu_k \otimes \mu_k$ and $M = \sum_k \alpha_k \mu_k \otimes \mu_k$ where μ_k are assumed to be linearly independent but not orthogonal, and $\alpha_k > 0$. M is a symmetric, positive semidefinite, rank k matrix. Let $M = UDU^T$ where $U \in \mathbb{R}^{n \times k}$ is unitary and $D \in R^{k \times k}$ is diagonal. Denote $W = UD^{-\frac{1}{2}}$. Then $M(W, W) = I_k$. Let $\tilde{\mu}_k = \sqrt{\alpha_k} W^T \mu_k$. Then

$$M(W, W) = W^T \sum_k \sqrt{\alpha_k} \mu_k \otimes \sqrt{\alpha_k} \mu_k W = \sum_k \tilde{\mu}_k \tilde{\mu}_k^T = I \tag{1}$$

Therefore $\tilde{\mu}_k$ form an orthonormal basis for $\mathbb{R}^k$. Let

$$\begin{aligned}\tilde{T} &= T(W, W, W) \\ &= \sum_k \alpha_k (W^T \mu_k) \otimes (W^T \mu_k) \otimes (W^T \mu_k) \\ &= \sum_k \alpha_k^{-\frac{1}{2}} \tilde{\mu}_k \otimes \tilde{\mu}_k \otimes \tilde{\mu}_k \end{aligned} \tag{2}$$

3 Connection Between BCM Neuron and Tensor Decompositions

The original formulation of the BCM rule is as follows: Let c be the post-synaptic firing rate, $\boldsymbol{d}$ be the vector of presynaptic firing rates, and $\boldsymbol{m}$ be the vector of synaptic weights. Then the BCM synaptic modification rule is

$$\begin{aligned} c(t) &= \langle \boldsymbol{m}, \boldsymbol{d} \rangle \\ \dot{\boldsymbol{m}} &= \phi(c, \theta) \boldsymbol{d} \end{aligned}$$

ϕ is a non-linear function of the firing rate, and θ is a sliding threshold that increases as a superlinear function of the average firing rate.

There are many different formulations of the BCM rule. The Intrator and Cooper model [7] has the following form for ϕ and θ.

$$\phi(c, \theta) = c(c - \theta) \text{ with } \theta = E[c^2]$$

These choices are quite convenient because they lead to the following objective function formulation of the synaptic update rule.

$$R(m) = \frac{1}{3} E\Big[\langle \boldsymbol{m}, \boldsymbol{d} \rangle^3\Big] - \frac{1}{4} E\Big[\langle \boldsymbol{m}, \boldsymbol{d} \rangle^2\Big]^2$$

Thus,

$$\begin{aligned} \nabla R &= E\Big[\langle \boldsymbol{m}, \boldsymbol{d} \rangle^2 - E[\langle \boldsymbol{m}, \boldsymbol{d} \rangle]^2 \langle \boldsymbol{m}, \boldsymbol{d} \rangle \boldsymbol{d}\Big] \\ &= E[\dot{\boldsymbol{m}}] \end{aligned}$$

So the BCM rule, in expectation performs gradient ascent in $R(\boldsymbol{m})$. The traditional application of this rule is a system where the input $\boldsymbol{d}$ is drawn from linearly independent vectors $\{\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_k\}$ with probabilities $\alpha_1, ..., \alpha_k$, with k less than n. With this model, the objective function can be rewritten in tensor notation.

$$\begin{aligned} T &= \sum_k \alpha_k \boldsymbol{\mu}_k \otimes \boldsymbol{\mu}_k \otimes \boldsymbol{\mu}_k \\ M &= \sum_k \alpha_k \boldsymbol{\mu}_k \otimes \boldsymbol{\mu}_k \\ R(m) &= \frac{1}{3} T(\boldsymbol{m}, \boldsymbol{m}, \boldsymbol{m}) - \frac{1}{4} M(\boldsymbol{m}, \boldsymbol{m})^2 \end{aligned}$$

Furthermore, this objective function is very similar to one which comes up naturally in the study of tensor analogues of eigenvalue decompositions. We modify the BCM update rule to specifically solve this problem.

4 Modified BCM Neuron

Rather than a sliding threshold that penalizes the activity of the neuron, we modify the BCM neuron with a sliding threshold that drives the activity of the neuron to a *specified* activity level. This will allow us to rewrite the objective function of the neuron as a generalized tensor spectral decomposition.

Let

$$\hat{R}(\boldsymbol{m}, \lambda) = \frac{1}{3}T(\boldsymbol{m}, \boldsymbol{m}, \boldsymbol{m}) + \frac{\lambda}{3}\left(1 - M(\boldsymbol{m}, \boldsymbol{m})^2\right)$$

Let W be defined as in equation 1. Let $\boldsymbol{m} = W\boldsymbol{u}$. Then

$$\begin{aligned}\hat{R}(W\boldsymbol{u}, \lambda) &= \frac{1}{3}T(W\boldsymbol{u}, W\boldsymbol{u}, W\boldsymbol{u}) + \frac{\lambda}{3}M(\boldsymbol{m}, \boldsymbol{m})^2 \\ &= \frac{1}{3}\tilde{T}(\boldsymbol{u}, \boldsymbol{u}, \boldsymbol{u}) + \frac{\lambda}{3}(1 - \langle \boldsymbol{u}, \boldsymbol{u}\rangle^2) \end{aligned} \tag{3}$$

where $\tilde{T}$ is defined as in equation 2. This is the Lagrange multiplier formulation of a generalized tensor spectral expansion for an orthogonal tensor.

With this objective function, the expected update rule becomes

$$\begin{aligned} E[\dot{\boldsymbol{m}}] = E[\nabla R(\boldsymbol{m}, \lambda)] = \\ = E[\hat{\phi}(c, \lambda\theta)\boldsymbol{d}] \\ E[\dot{\lambda}] = (1 - \theta^2) \end{aligned}$$

where $\hat{\phi} = c(c - \frac{4}{3}\lambda\theta)$

A synaptic update of

$$\begin{aligned}\Delta\boldsymbol{m} &= h\hat{\phi}(c, \lambda\theta)\boldsymbol{d} \\ \Delta\lambda &= -h(1 - \theta^2)\end{aligned}$$

with a suitable decaying step size h will converge to the local maxima of $\hat{R}$ with probability 1.

Theorem 4.1. *The stable solutions of* (3) *are*

$$\boldsymbol{m} = \alpha_k^{\frac{1}{2}} M^{-1}\mu_k$$

Thus the modified BCM neuron learns decorrelated versions of the parameter vectors μ_k. In contrast with the ordinary matrix (2-tensor) eigendecomposition, this update function can converge to each of the eigenvectors of the 3-tensor, rather than just the one corresponding to the largest eigenvalue.

5 Mixture Models and Tensor Decompositions

We will now demonstrate that the tensor decomposition performed by our modified BCM neuron is exactly the decomposition required to learn selectivity to individual mixtures in a mixture model. It has previously been observed [2] that under a multiview assumption, certain moments of mixture distributions can be written as sums of rank one tensors. Let

$$P(\boldsymbol{d}) = \sum_k \alpha_k P_k(\boldsymbol{d})$$

$$\text{where } E_{P_k}[\boldsymbol{d}] = \boldsymbol{\mu}_k$$

Assume we have access to multiple independent, identically distributed copies of d, i.e. $\{\boldsymbol{d}^1, \boldsymbol{d}^2, \boldsymbol{d}^3\}$ are all distributed according to $P_k(\boldsymbol{d})$ for some k. This is the multiview assumption. Under this assumption,

$$E[\langle \boldsymbol{m}, \boldsymbol{d}_1\rangle\langle \boldsymbol{m}, \boldsymbol{d}_2\rangle\langle \boldsymbol{m}, \boldsymbol{d}_3\rangle] = \left(\sum_k \alpha_k \boldsymbol{\mu_k} \otimes \boldsymbol{\mu_k} \otimes \boldsymbol{\mu_k}\right)(\boldsymbol{m}, \boldsymbol{m}, \boldsymbol{m})$$

We note that unless $P_k\{\boldsymbol{d} = \boldsymbol{d}_k\} = 1$, $E[\langle \boldsymbol{m}, \boldsymbol{d}_1\rangle\langle \boldsymbol{m}, \boldsymbol{d}_2\rangle\langle \boldsymbol{m}, \boldsymbol{d}_3\rangle] \neq E[\langle \boldsymbol{m}, \boldsymbol{d}_1\rangle\langle \boldsymbol{m}, \boldsymbol{d}_1\rangle\langle \boldsymbol{m}, \boldsymbol{d}_1\rangle]$. This helps explain why the BCM learning rule does not have the same fixed points in the presence of noise. The tensors T and M no longer have low-rank decompositions. However, under the multiview assumption, we can still construct a low-rank tensor, whose decomposition will recover the means of the input mixtures. This suggests modifying the BCM rule to perform gradient ascent in this multiview tensor objective function.

6 Triplet BCM Rule

We now show that by modifying the update rule to incorporate information from triplets of input vectors, the generality of the input data can be dramatically increased. Assume that

$$P(\boldsymbol{d}) = \sum_k \alpha_k P_k(\boldsymbol{d})$$

where $E_{P_k}[\boldsymbol{d}] = \mu_k$ and $P_k(\boldsymbol{d}) = \prod_i P_{ki}(d_i)$. For example, the data could be a mixture of axis-aligned Gaussians, a mixture of independent Poisson variables, or mixtures of independent Bernoulli random variables to name a few. We also require $E[d_i^2] < \infty$. We emphasize that we do not require our data to come from any parametric distribution.

We interpret k to be a latent variable that signals the hidden cause of the underlying input distribution, with distribution k. Critically, we assume that the hidden variable k changes slowly compared to the inter-spike period of the neuron. In particular, we need at least 3 samples from each P_k. This corresponds to the multiview assumption of [1]. A particularly relevant model meeting this assumption is that of spike counts in disjoint intervals under a Poisson process, with a discrete, time varying rate parameter.

Let $\{\boldsymbol{d}_1, \boldsymbol{d}_2, \boldsymbol{d}_3\}$ be a triplet of independent copies from some $P_k(\boldsymbol{d})$, i.e. each are drawn from the same latent class. It is critical to note that if $\{\boldsymbol{d}_1, \boldsymbol{d}_2, \boldsymbol{d}_3\}$ are not drawn from the same class, this update will not converge to the global maximum. Our sample is thus a sequence of triplets, each triplet drawn from the same latent distribution. Let $c_i = \langle \boldsymbol{d}_i, \boldsymbol{m} \rangle$. Setting $T = E[\langle \boldsymbol{m}, \boldsymbol{d}_1 \rangle \langle \boldsymbol{m}, \boldsymbol{d}_2 \rangle \langle \boldsymbol{m}, \boldsymbol{d}_3 \rangle]$ and $M = E[\langle \boldsymbol{m}, \boldsymbol{d}_2 \rangle \langle \boldsymbol{m}, \boldsymbol{d}_3 \rangle] = \theta$.

$$\begin{aligned}
\hat{R}(\boldsymbol{m}, \lambda) &= \frac{1}{3} T(\boldsymbol{m}, \boldsymbol{m}, \boldsymbol{m}) + \frac{\lambda}{3} \left(1 - M(\boldsymbol{m}, \boldsymbol{m})^2\right) \\
\nabla_{\boldsymbol{m}} R &= E[\frac{1}{3}(\boldsymbol{d}_3 c_1 c_2 + \boldsymbol{d}_2 c_1 c_3 + \boldsymbol{d}_1 c_2 c_3) - \frac{2}{3}\lambda\theta(c_2 \boldsymbol{d}_3 + c_3 \boldsymbol{d}_2)] \\
&= E[c_1 c_2 \boldsymbol{d}_3 - \frac{4}{3}\lambda\theta c_2 \boldsymbol{d}_3]
\end{aligned}$$

This suggests using a new update rule, with $\phi(c_1, c_2, \lambda\theta) = c_2(c_1 - \frac{4}{3}\lambda\theta)$

Theorem 6.1. *With the update rule*

$$\Delta \boldsymbol{m} = h\phi(c_1, c_2, \lambda\theta)\boldsymbol{d}_3 \tag{4}$$

$$\dot{\lambda} = h(1 - \theta^2) \tag{5}$$

then with a suitable step size decay for h, $\boldsymbol{m}$ will converge a.s. to

$$\boldsymbol{m} = \alpha_k^{\frac{1}{2}} M^{-1} \mu_k$$

We note that any permutation of the indicies will also create a valid learning rule. This suggests that there may be multiple spike timing dependency rules that converge to the same weight.

7 How Can We Get Multiple Views? Spikes as Independent Samples

Though the multiview assumption has clear computational benefits, it is less clear how we can get these independent samples. It has been shown that most of the information about the stimulus is contained in the first few spikes [11, 152-154], which suggests that we should focus our attention on the initial wave of firing. We assume that, conditioned on the stimulus, the firing patterns of our input neurons are independent over disjoint intervals.

For example, assume we have k stimuli, and in the presence of stimuli k, each input neuron i fires according to a Poisson process with rate parameter r_{ki}. Then the number of spikes in disjoint time intervals constitute independent samples from our Poisson distribution. Assuming the stimuli are chosen at random with probability α_k, then the number of spikes $\boldsymbol{d}_k$ of input neurons $\{d_1, d_2, \cdots, d_n\}$ due to input k in an interval of length δ is

$$P(\boldsymbol{d}) = \sum_k \alpha_k \prod_i Po_{\delta r_{ki}}(d_i)$$

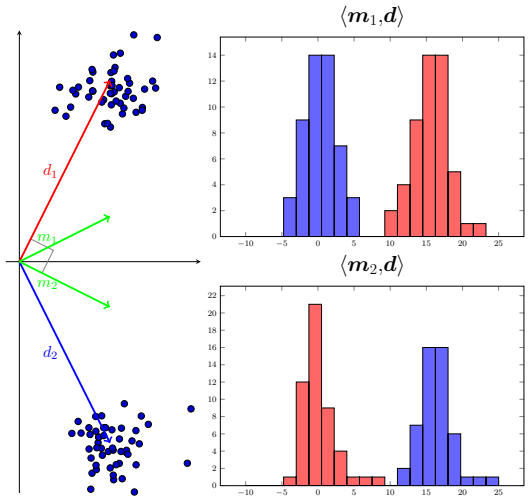

Figure 1: Geometry of stable solutions. Each stable solution is selective in expectation for a single mixture. Note that the classical BCM rule will not converge to these values in the presence of noise. The triplet version of BCM can be viewed as a modification of the classical BCM rule which allows it to converge in the presence of zero-mean noise.

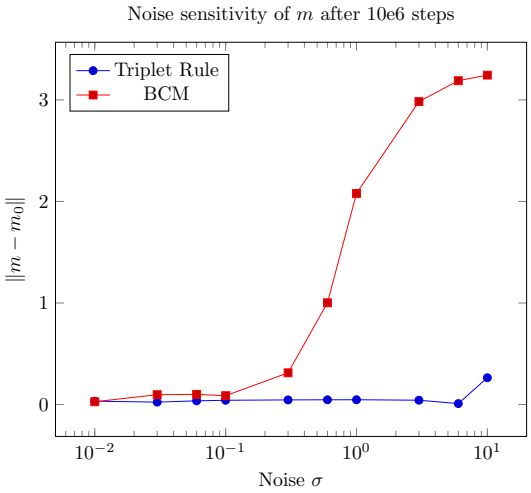

Figure 2: Noise response of triplet BCM update rule vs BCM update. Input data was a mixture of Gaussians with standard deviation σ. The selectivity of the triplet BCM rule remains unchanged in the presence of noise.

where $Po_r(d)$ denotes the distribution of a Poisson random variable with rate paramter r. The number of spikes in a triplet of disjoint intervals $\{\boldsymbol{d}_1, \boldsymbol{d}_2, \boldsymbol{d}_3\}$ meet the conditions of the multiview assumption, so long as the stimulus does not change during this period.

We emphasize two points. First, our model does not rely on a Poisson process to generate the spikes, rather, it relies only on conditionally independent (given the stimulus) spiking patterns over disjoint intervals. Second, even with the restricted Poisson assumption, the classical BCM rule will not converge to weights selective for only one stimulus. This is due to the inherent noise of the Poisson spiking model.

8 Conclusion

We introduced a modified formulation of the classical BCM neural update rule. This update rule drives the synaptic weights toward the components of a tensor decomposition of the input data. By further modifying the update to incorporate information from triplets of input data, this tensor decomposition can learn the mixture means for a broad class of mixture distributions. Unlike other methods to fit mixture models, we incorporate a multiview assumption that allows us to learn asymptotically *exact* mixture means, rather than local maxima of a similarity measure. This is in stark contrast to EM and other gradient ascent based methods, which have limited guarantees about the quality of their results. Conceptually our model suggests a different view of spike waves during adjacent time epochs: they provide multiple independent samples of the presynaptic "image."

Due to size constraints, this abstract focused a single neuron, however we believe this model neuron can be a useful unit in a variety of neural circuits, both feed-forward and lateral, learning mixture means under a variety of input distributions.

Research supported by NSF, NIH, and the Paul Allen Foundation.

References

[1] Anima Anandkumar, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *arXiv preprint arXiv:1210.7559*, 2012.

[2] Animashree Anandkumar, Dean P Foster, Daniel Hsu, Sham M Kakade, and Yi-Kai Liu. Two svds suffice: Spectral decompositions for probabilistic topic modeling and latent dirichlet allocation. *CoRR, abs/1204.6703*, 1, 2012.

[3] Elie L Bienenstock, Leon N Cooper, and Paul W Munro. Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *The Journal of Neuroscience*, 2(1):32–48, 1982.

[4] Natalia Caporale and Yang Dan. Spike timing-dependent plasticity: a hebbian learning rule. *Annual Review Neuroscience*, 31:25–46, 2008.

[5] Robert C Froemke and Yang Dan. Spike-timing-dependent synaptic modification induced by natural spike trains. *Nature*, 416(6879):433–438, 2002.

[6] Julijana Gjorgjieva, Claudia Clopath, Juliette Audet, and Jean-Pascal Pfister. A triplet spike-timing–dependent plasticity model generalizes the bienenstock–cooper–munro rule to higher-order spatiotemporal correlations. *Proceedings of the National Academy of Sciences*, 108(48):19383–19388, 2011.

[7] Nathan Intrator and Leon N Cooper. Objective function formulation of the bcm theory of visual cortical plasticity: Statistical connections, stability conditions. *Neural Networks*, 5(1):3–17, 1992.

[8] Matthew Lawlor and Steven Zucker. Third-order edge statistics: Contour continuation, curvature, and cortical connections. *arXiv preprint arXiv:1306.3285*, 2013.

[9] Bernhard Nessler, Michael Pfeiffer, and Wolfgang Maass. Stdp enables spiking neurons to detect hidden causes of their inputs. In *Advances in Neural Information Processing Systems*, pages 1357–1365, 2009.

[10] Jean-Pascal Pfister and Wulfram Gerstner. Triplets of spikes in a model of spike timing-dependent plasticity. *The Journal of Neuroscience*, 26(38):9673–9682, 2006.

[11] Fred Rieke, David Warland, Rob de Ruyter van Steveninck, and William Bialek. Spikes: exploring the neural code (computational neuroscience). 1999.

[12] Sen Song, Kenneth D Miller, and Larry F Abbott. Competitive hebbian learning through spike-timing-dependent synaptic plasticity. *Nature neuroscience*, 3(9):919–926, 2000.