Predictive Coding, Variational Autoencoders, and Biological Connections

Joseph Marino
Computation & Neural Systems
California Institute of Technology
Pasadena, CA 91125
jmarino@caltech.edu

Abstract

Predictive coding, within theoretical neuroscience, and variational autoencoders, within machine learning, both involve latent Gaussian models and variational inference. While these areas share a common origin, they have evolved largely independently. We outline connections and contrasts between these areas, using their relationships to identify new parallels between machine learning and neuroscience. We then discuss specific frontiers at this intersection: backpropagation, normalizing flows, and attention, with mutual benefits for both fields.

1 Introduction

Perception has been conventionally formulated as hierarchical feature detection [51], similar to discriminative deep networks [33]. In contrast, predictive coding [47, 13] and variational autoencoders (VAEs) [30, 50] frame perception as a generative process, modeling data observations to learn and infer aspects of the external environment. Specifically, both areas model observations, $\mathbf{x}$, using latent variables, $\mathbf{z}$, through a probabilistic model, $p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z})$. Both areas also use variational inference, introducing an approximate posterior, $q(\mathbf{z}|\mathbf{x})$, to infer $\mathbf{z}$ and learn the model parameters, θ. These similarities are the result of a common origin, with Mumford [44], Dayan et al. [8], and others [45] formalizing earlier ideas [58, 37]. However, since their inception, these areas have developed largely independently. We explore their relationships (see also [57, 36]) and highlight opportunities for the transfer of ideas. In identifying these ties, we hope to strengthen this promising, close connection between neuroscience and machine learning, prompting further investigation.

2 Background

Predictive Coding Predictive coding [7] is a theory of thalamocortical function, in which the cortex constructs a probabilistic generative model of sensory inputs, using approximate inference to perform state estimation. Top-down neural projections convey predictions of lower-level activity, while bottom-up projections convert the prediction error at each level into an updated state estimate. Such models are often formulated with hierarchies of Gaussian distributions, with analytical non-linear (e.g. polynomial) functions parameterizing the generative mappings [47, 13]:

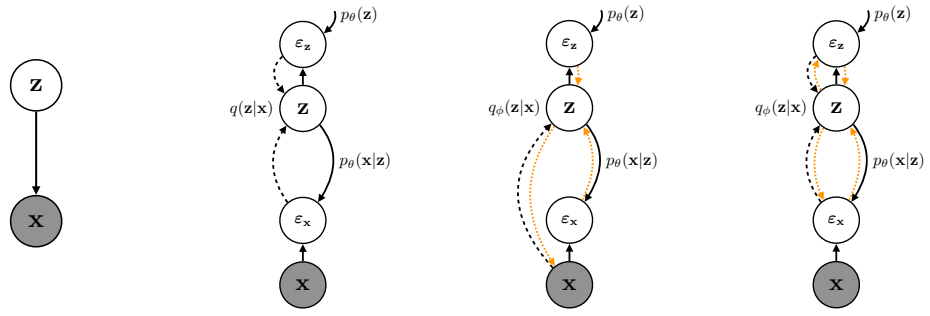
$$p_\theta(\mathbf{z}_\ell|\mathbf{z}_{\ell+1}) = \mathcal{N}(\mathbf{z}_\ell; \boldsymbol{\mu}_{\theta,\ell}(\mathbf{z}_{\ell+1}), \boldsymbol{\Sigma}_{p,\ell}), \qquad p_\theta(\mathbf{x}|\mathbf{z}_1) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{\theta,\mathbf{x}}(\mathbf{z}_1), \boldsymbol{\Sigma}_\mathbf{x}). \tag{1}$$

Variational inference is performed using gradient-based optimization on the mean of $q(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}_\ell; \boldsymbol{\mu}_{q,\ell}, \boldsymbol{\Sigma}_{q,\ell})$, yielding gradients which are linear combinations of (prediction) errors, e.g.

$$\nabla_{\boldsymbol{\mu}_{q,1}} \mathcal{L} = \mathbf{J}^\intercal \boldsymbol{\varepsilon}_\mathbf{x} - \boldsymbol{\varepsilon}_1, \tag{2}$$

where $\mathcal{L}$ is the objective, $\mathbf{J} = \partial \boldsymbol{\mu}_{\theta,\mathbf{x}} / \partial \boldsymbol{\mu}_{q,1}$ is the Jacobian, and $\boldsymbol{\varepsilon}_\mathbf{x}$ and $\boldsymbol{\varepsilon}_1$ are weighted errors, i.e. $\boldsymbol{\varepsilon}_\mathbf{x} = \boldsymbol{\Sigma}_\mathbf{x}^{-1}(\mathbf{x} - \boldsymbol{\mu}_{\theta,\mathbf{x}})$. Parameter learning can also be performed using gradient-based optimization. We discuss connections between these models and neuroscience in Section 3.

Real Neurons & Hidden Units Workshop @ NeurIPS 2019.



(a) (b) (c) (d)

Figure 1: **Computation Graphs**. **(a)** Graphical model for the (single-level) latent variable model underlying both predictive coding and VAEs. **(b)** Computation graph for predictive coding. Black dashed lines denote inference computation (Eq. 2). **(c)** Computation graph for a standard VAE with direct amortized inference. Orange dashed lines denote parameter gradients, used for learning θ and ϕ. **(d)** Computation graph for a VAE with an example of an iterative inference model [41].

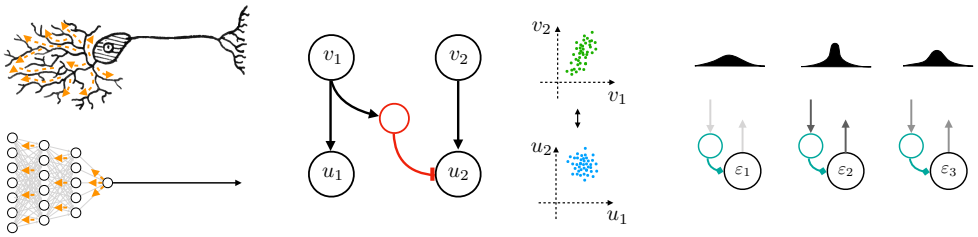
Variational Autoencoders VAEs are a class of Bayesian machine learning models, combining latent Gaussian models with deep neural networks. They consist of an *encoder* network with parameters ϕ, parameterizing $q_\phi(\mathbf{z}|\mathbf{x})$, and a *decoder* network, parameterizing $p_\theta(\mathbf{x}|\mathbf{z})$. Thus, rather than performing gradient-based inference, VAEs *amortize* inference optimization with a learned network [18], improving computational efficiency. These networks can either take a *direct* form [8], e.g. $\boldsymbol{\mu}_q \leftarrow \text{NN}_\phi(\mathbf{x})$, or an *iterative* form [41, 40], e.g. $\boldsymbol{\mu}_q \leftarrow \text{NN}_\phi(\boldsymbol{\mu}_q, \nabla_{\boldsymbol{\mu}_q}\mathcal{L})$, where NN_ϕ denotes a deep neural network. In both cases, gradients are obtained by reparameterizing stochastic samples, $\mathbf{z} \sim q(\mathbf{z}|\mathbf{x})$, separating stochastic and deterministic dependencies [30, 50]. The parameters, θ and ϕ, are learned using gradient-based optimization, with gradients calculated via backpropagation [53].

3 Connections, Contrasts, and Biological Correspondences

Predictive coding and VAEs are both formulated in terms of hierarchical latent Gaussian models, with non-linear functions parameterizing the conditional dependencies between variables. In the case of predictive coding, these functions are polynomials, whereas VAEs use deep networks, which are composed of layers of linear and non-linear operations. In predictive coding, Gaussian covariance matrices, e.g. $\boldsymbol{\Sigma}_\mathbf{x}$, are separate parameters, implemented as lateral weights between units at each level. A similar, but more general, mechanism was independently developed for VAEs, known as normalizing flows [49, 29] (Section 4.2). Both areas have been extended to sequential models. In this setting, predictive coding tends to model dynamics explicitly, directly modeling orders of motion or *generalized coordinates* [14]. VAEs, in contrast, tend to rely on less rigid forms of dynamics, often using recurrent networks, e.g. [6], though some works have explored structured dynamics [25, 26, 39]. Both areas use gradient-based learning. In practice, however, learning in predictive coding tends to be minimal, while VAEs use learning extensively, scaling to large image and audio datasets, e.g. [48].

Predictive coding and VAEs both use variational inference, setting $q(\mathbf{z}|\mathbf{x})$ as Gaussian. Predictive coding uses errors (Eq. 2) to perform gradient-based inference, whereas VAEs use amortized inference, learning to infer. This offers a solution to the so-called "weight transport" problem [35] for predictive coding; inference gradients require the Jacobian of the generative model, which includes the transpose of generative weight matrices. Learning a separate set of inference weights avoids this problem.

The benefit of identifying these connections and contrasts is that they link neuroscience, through predictive coding and VAEs, to machine learning (and vice versa). While still under debate [19], biological correspondences of predictive coding have been proposed [2, 27]. Variable estimates and errors are hypothesized to be represented by neural activity (firing rate or potential). These neurons would occur within the same cortical column, with variable neurons and error neurons in separate cortical layers. Interneurons could mediate inversion of errors and predictions, as well as lateral inhibition necessary for covariance matrices and attention (Section 4.3). Although many of the biological details remain unclear, one intriguing point emerges from this analysis: following this



(a) Backpropagation (b) Normalizing Flow (c) Attention

Figure 2: **Frontiers**. **(a)** Comparing predictive coding and VAEs, deep networks are analogous to (ensembles of) dendrites. This suggests that similar mechanisms to backpropagation may operate *within* neurons. **(b)** Normalizing flows can be implemented with lateral inhibitory interactions. In the diagram, a correlated vector, $\mathbf{v}$, is decorrelated to a new vector, $\mathbf{u}$, simplifying the prediction space. **(c)** Attention mechanisms can be implemented using the precision (inverse variance) of predictions. Weighting prediction errors biases inference toward representing highly precise dimensions. In the diagram, various strengths of neuromodulation, corresponding to the precision of predictions, adjust the gain of error neurons.

connection, deep networks are analogous to the mappings between neurons at separate levels in the hierarchy. Thus, deep networks may more closely correspond with (a parallel ensemble of) dendrites. We discuss this implication (Section 4.1) and others in the following section.

4 Frontiers

4.1 Backpropagation & Learning

The biological correspondence of backpropagation [53] remains an open question. Backprop requires global information, whereas biology seems to rely largely on local learning rules [23, 42, 5]. A number of biologically-plausible formulations of backprop have been proposed [55, 32, 62, 24, 35], attempting to reconcile this disparity and others. However, recent formulations of learning in latent variable models [4, 34, 60] offer an alternative perspective: prediction errors at each level of the latent hierarchy provide a local signal [13], capable of driving learning of inference and generative weights.

In Section 3, we noted that deep networks appear between each latent level, suggesting a correspondence with dendrites rather than the traditional analogy as networks of neurons. This implies the following set-up: learning across the cortical hierarchy is handled via local errors at each level, whereas learning within the neurons at each level is mediated through a mechanism similar to backpropagation. Indeed, looking at the literature, we see ample evidence of non-linear dendritic computation [43] and backpropagating action potentials within neurons (Fig. 2a) [61, 31]. From this perspective, segmented dendrites [3, 21] for top-down and bottom-up inputs to pyramidal neurons could implement separate inference computations for errors at different levels (Eq. 2). While the mechanisms underlying these processes remain unclear, focusing efforts on formulating biologically plausible backpropagation from this perspective (and not supervised learning) could prove fruitful.

4.2 Normalizing Flows as Local Inhibition

We often consider factorized parametric distributions, as they enable efficient evaluation and sampling. However, simple distributions are limiting. *Normalizing flows* (NFs) [52, 9, 49] provide added complexity while maintaining tractable evaluation and sampling. They consist of a tractable *base* distribution and one or more invertible *transforms*. With the base distribution as $p_\theta(\mathbf{u})$ and the transforms as $\mathbf{v} = f_\theta(\mathbf{u})$, the probability $p_\theta(\mathbf{v})$ is given by the *change of variables* formula:

$$p_\theta(\mathbf{v}) = p_\theta(\mathbf{u}) \left| \det \left(\frac{d\mathbf{v}}{d\mathbf{u}} \right) \right|^{-1}, \tag{3}$$

where $\det(\cdot)$ denotes matrix determinant and $|\cdot|$ denotes absolute value. The determinant term corrects for the local scaling of space when moving from $\mathbf{u}$ to $\mathbf{v}$. A popular family of transforms is

that of autoregressive affine transforms [29, 46]. One example is given by

$$v_i = \alpha_\theta(\mathbf{v}_{<i}) + \beta_\theta(\mathbf{v}_{<i}) \cdot u_i, \tag{4}$$

where v_i is the i^{th} dimension of $\mathbf{v}$ and α_θ and β_θ are functions. The inverse transform (Fig. 2b) is

$$u_i = \frac{v_i - \alpha_\theta(\mathbf{v}_{<i})}{\beta_\theta(\mathbf{v}_{<i})}, \tag{5}$$

a normalization (whitening) operation. Thus, we can sample from complex distributions by starting with simple distributions and applying local affine transforms. Conversely, we can evaluate inputs from complex distributions by applying normalization transforms, then evaluating in a simpler space.

Local inhibition is ubiquitous in neural systems, thought to implement normalization [12]. These circuits, modeled with subtractive and divisive operations (Eq. 5), give rise to decorrelation in retina [20], LGN [10], and cortex [28]. NFs offer a novel description of these circuits and agree with predictive coding. For instance, evaluating flow-based conditional likelihoods [1] involves whitening the observations, as performed by Rao & Ballard [47], to remove low-level spatial correlations. The same principle can be applied across time, where NFs resemble temporal derivatives [39], which are the basis of Friston's generalized coordinates [14]. Likewise, Friston's proposal [13] of implementing prior covariance matrices with lateral weights in cortex corresponds to a linear NF [29].

Local inhibition is also present in central pattern generator (CPG) circuits [38], giving rise to correlations in muscle activation. NFs are also being explored in the context of action selection in reinforcement learning [56, 59]. By providing a basis of correlated motor outputs, NFs improve action selection and learning, which can take place in a less correlated space that is easier to model. CPGs would likely be a form of *inverse* autoregressive flow [29] to maintain efficient sampling.

4.3 Attention via Precision Weighting

Predictive coding has proposed that prior covariance matrices, which weight prediction errors, could implement a form of *attention* [54, 15]. Intuitively, decreasing the variance of a predicted variable pushes the model to more accurately infer and predict that variable. Biologically, this is hypothesized to be implemented via gain modulation of error-encoding neurons, mediated through neurotransmitters and synchronizing gamma oscillations [11]. This attentional control mechanism could bias a model toward representing task-relevant information. Deep latent variable models have largely ignored this functionality; when combined with active components, variances are typically held constant, e.g. [22]. Enabling this capacity for task-dependent perceptual modulation may prove useful or even essential in applying deep latent variable models to complex tasks.

5 Conclusion

We have identified commonalities between predictive coding and VAEs, discussing new frontiers resulting from this perspective. Reuniting these areas may strengthen the connection between neuroscience and machine learning. Further refining this connection could lead to mutual benefits: neuroscience can offer inspiration for investigation in machine learning, and machine learning can evaluate ideas on real-world datasets and environments. Indeed, despite some push back [16], if predictive coding and related theories [17] are to become validated descriptions of the brain and overcome their apparent generality, they will likely require the computational tools and ideas of modern machine learning to pin down and empirically compare design choices.

References

[1] S. Agrawal and A. Dukkipati. Deep variational inference without pixel-wise reconstruction. *arXiv preprint arXiv:1611.05209*, 2016.

[2] A. M. Bastos, W. M. Usrey, R. A. Adams, G. R. Mangun, P. Fries, and K. J. Friston. Canonical microcircuits for predictive coding. *Neuron*, 2012.

[3] J. M. Bekkers. Pyramidal neurons. *Current Biology*, 2011.

[4] Y. Bengio. How auto-encoders could provide credit assignment in deep networks via target propagation. *arXiv preprint arXiv:1407.7906*, 2014.

[5] G.-q. Bi and M.-m. Poo. Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *Journal of neuroscience*, 1998.

[6] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio. A recurrent latent variable model for sequential data. In *Advances in Neural Information Processing Systems*, 2015.

[7] A. Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 2013.

[8] P. Dayan, G. E. Hinton, R. M. Neal, and R. S. Zemel. The helmholtz machine. *Neural computation*, 1995.
[9] L. Dinh, D. Krueger, and Y. Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.
[10] D. W. Dong and J. J. Atick. Temporal decorrelation: a theory of lagged and nonlagged responses in the lateral geniculate nucleus. *Network: Computation in Neural Systems*, 1995.
[11] H. Feldman and K. J. Friston. Attention, uncertainty, and free-energy. *Frontiers in human neuroscience*, 2010.
[12] P. Földiak. Forming sparse representations by local anti-hebbian learning. *Biological cybernetics*, 1990.
[13] K. Friston. A theory of cortical responses. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 2005.
[14] K. Friston. Hierarchical models in the brain. *PLoS computational biology*, 2008.
[15] K. Friston. The free-energy principle: a rough guide to the brain? *Trends in cognitive sciences*, 13(7):293–301, 2009.
[16] K. Friston. Does predictive coding have a future? *Nature neuroscience*, 21(8):1019, 2018.
[17] K. Friston. A free energy principle for a particular physics. *arXiv preprint arXiv:1906.10184*, 2019.
[18] S. Gershman and N. Goodman. Amortized inference in probabilistic reasoning. In *Cognitive Science Society*, 2014.
[19] S. J. Gershman. What does the free energy principle tell us about the brain? *arXiv preprint arXiv:1901.07945*, 2019.
[20] D. J. Graham, D. M. Chandler, and D. J. Field. Can the theory of "whitening" explain the center-surround properties of retinal ganglion cell receptive fields? *Vision research*, 46(18):2901–2913, 2006.
[21] J. Guergiuev, T. P. Lillicrap, and B. A. Richards. Biologically feasible deep learning with segregated dendrites. *arXiv preprint arXiv:1610.00161*, 2016.
[22] D. Ha and J. Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems*, 2018.
[23] D. O. Hebb. The organization of behavior; a neuropsycholocigal theory. *A Wiley Book in Clinical Psychology.*, 1949.
[24] G. E. Hinton. How to do backpropagation in a brain. *NeurIPS Deep Learning Workshop*, 2007.
[25] M. Johnson, D. K. Duvenaud, A. Wiltschko, R. P. Adams, and S. R. Datta. Composing graphical models with neural networks for structured representations and fast inference. In *Advances in Neural Information Processing Systems*, 2016.
[26] M. Karl, M. Soelch, J. Bayer, and P. van der Smagt. Deep variational bayes filters: Unsupervised learning of state space models from raw data. In *International Conference on Learning Representations*, 2017.
[27] G. B. Keller and T. D. Mrsic-Flogel. Predictive processing: A canonical cortical computation. *Neuron*, 2018.
[28] P. D. King, J. Zylberberg, and M. R. DeWeese. Inhibitory interneurons decorrelate excitatory cells to drive sparse code formation in a spiking model of v1. *Journal of Neuroscience*, 2013.
[29] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling. Improved variational inference with inverse autoregressive flow. In *Advances in neural information processing systems*, 2016.
[30] D. P. Kingma and M. Welling. Stochastic gradient vb and the variational auto-encoder. In *Proceedings of the International Conference on Learning Representations*, 2014.
[31] K. P. Körding and P. König. A learning rule for dynamic recruitment and decorrelation. *Neural Networks*, 2000.
[32] K. P. Körding and P. König. Supervised and unsupervised learning with two sites of synaptic integration. *Journal of computational neuroscience*, 2001.
[33] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.
[34] D.-H. Lee, S. Zhang, A. Fischer, and Y. Bengio. Difference target propagation. In *Joint european conference on machine learning and knowledge discovery in databases*, 2015.
[35] T. P. Lillicrap, D. Cownden, D. B. Tweed, and C. J. Akerman. Random synaptic feedback weights support error backpropagation for deep learning. *Nature communications*, 2016.
[36] W. Lotter, G. Kreiman, and D. Cox. A neural network trained to predict future video frames mimics critical properties of biological neuronal responses and perception. *arXiv preprint arXiv:1805.10734*, 2018.
[37] D. M. MacKay. The epistemological problem for automata. *Automata studies*, 1956.
[38] E. Marder and D. Bucher. Central pattern generators and the control of rhythmic movements. *Current biology*, 2001.
[39] J. Marino, L. Chen, J. He, and S. Mandt. Improving sequential latent variable models with autoregressive flows. 2019.
[40] J. Marino, M. Cvitkovic, and Y. Yue. A general method for amortizing variational filtering. In *Advances in Neural Information Processing Systems*, 2018.
[41] J. Marino, Y. Yue, and S. Mandt. Iterative amortized inference. In *International Conference on Machine Learning*, 2018.
[42] H. Markram, J. Lübke, M. Frotscher, and B. Sakmann. Regulation of synaptic efficacy by coincidence of postsynaptic aps and epsps. *Science*, 1997.
[43] B. W. Mel. The clusteron: toward a simple abstraction for a complex neuron. In *Advances in neural information processing systems*, 1992.
[44] D. Mumford. On the computational architecture of the neocortex. *Biological cybernetics*, 1992.
[45] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 1996.
[46] G. Papamakarios, T. Pavlakou, and I. Murray. Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems*, 2017.
[47] R. P. Rao and D. H. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 1999.
[48] A. Razavi, A. van den Oord, and O. Vinyals. Generating diverse high-resolution images with vq-vae. 2019.
[49] D. Rezende and S. Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning*, 2015.
[50] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the International Conference on Machine Learning*, 2014.
[51] M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature neuroscience*, 1999.
[52] O. Rippel and R. P. Adams. High-dimensional probability estimation with deep density models. *arXiv preprint arXiv:1302.5125*, 2013.
[53] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 1986.
[54] M. W. Spratling. Reconciling predictive coding and biased competition models of cortical function. *Frontiers in computational neuroscience*, 2:4, 2008.
[55] D. G. Stork. Is backpropagation biologically plausible. In *International Joint Conference on Neural Networks*, 1989.
[56] Y. Tang and S. Agrawal. Boosting trust region policy optimization by normalizing flows policy. *arXiv preprint arXiv:1809.10326*, 2018.
[57] G. van den Broeke. What auto-encoders could learn from brains. Master's thesis, 2016.
[58] H. Von Helmholtz. *Handbuch der physiologischen Optik*. 1867.
[59] P. N. Ward, A. Smofsky, and A. J. Bose. Improving exploration in soft-actor-critic with normalizing flows policies. *ICML Workshop on Invertible Neural Nets and Normalizing Flows*, 2019.
[60] J. C. Whittington and R. Bogacz. An approximation of the error backpropagation algorithm in a predictive coding network with local hebbian synaptic plasticity. *Neural computation*, 2017.
[61] S. R. Williams and G. J. Stuart. Backpropagation of physiological spike trains in neocortical pyramidal neurons: implications for temporal coding in dendrites. *Journal of Neuroscience*, 2000.
[62] X. Xie and H. S. Seung. Equivalence of backpropagation and contrastive hebbian learning in a layered network. *Neural computation*, 2003.