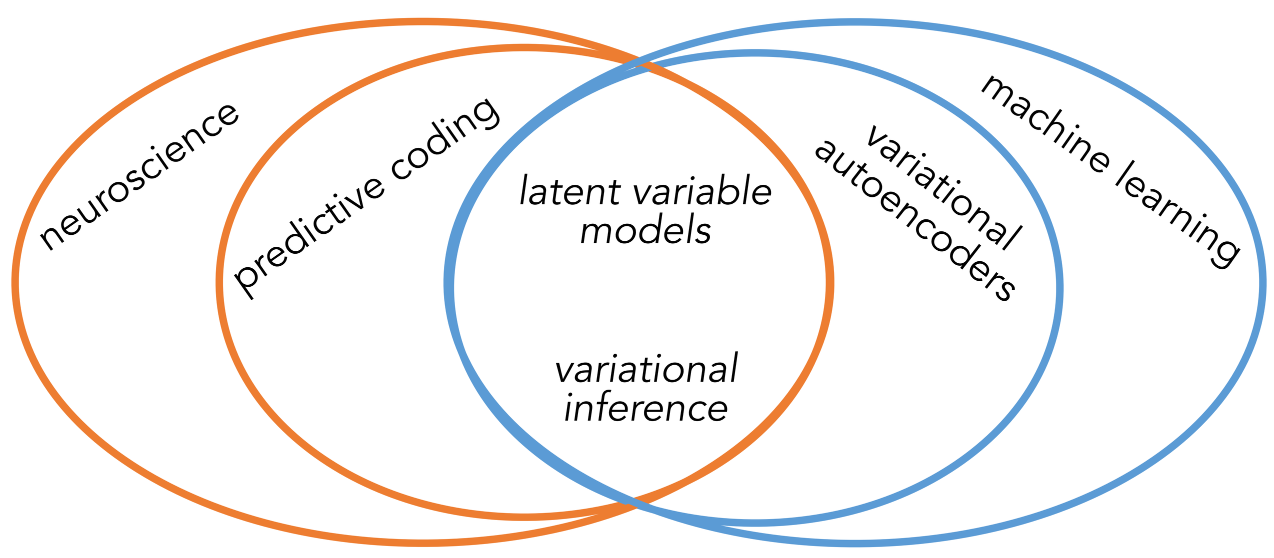
Caltech

Predictive Coding, Variational Autoencoders, and Biological Connections

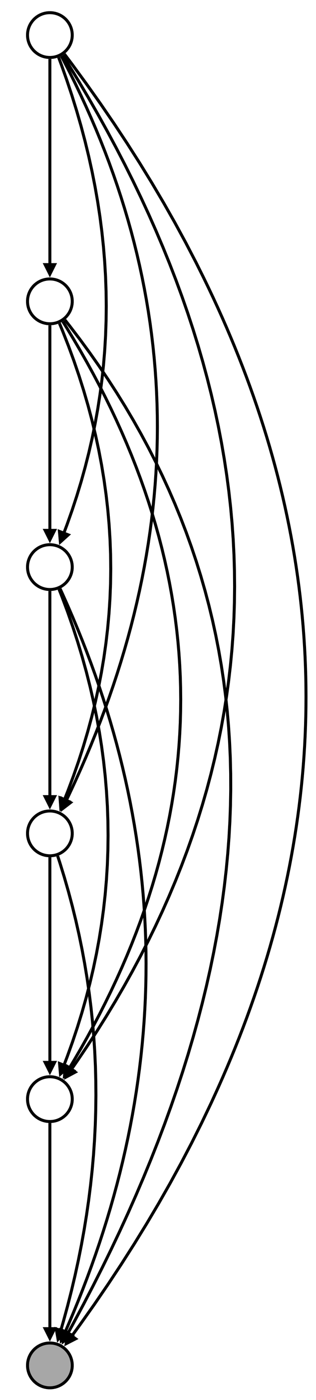
Joseph Marino
California Institute of Technology

introduction



Predictive coding and VAEs share a *common origin*, arising from ideas from Mumford, 1992; Dayan et al., 1995; Olshausen & Field, 1996; etc. However, these areas have *evolved largely independently*.

- We **connect and contrast** these areas to strengthen the bridge between neuroscience and machine learning.
- We discuss **frontiers** where these areas can contribute to each other.



background

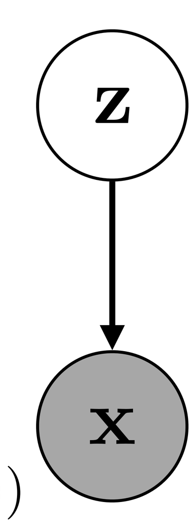
Latent Variable Models

observations $\mathbf{x}$ *model* $p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z}) p_\theta(\mathbf{z})$
latent variables $\mathbf{z}$

Variational Inference

approx. posterior $q(\mathbf{z}|\mathbf{x}) \leftarrow \arg\max_q \mathcal{L}(\mathbf{x}; q, \theta)$

ELBO/-FE $\mathcal{L}(\mathbf{x}; q, \theta) \equiv \mathbb{E}_q[\log p_\theta(\mathbf{x}|\mathbf{z})] - D_{\text{KL}}(q(\mathbf{z}|\mathbf{x}) || p_\theta(\mathbf{z}))$



Predictive Coding [Rao & Ballard, 1999; Friston, 2005]:

- cortex constructs a generative model of sensory inputs, and
- uses approximate inference to perform state estimation.

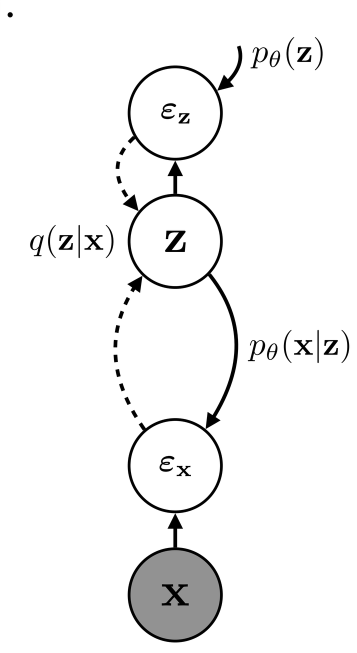
Hierarchical latent Gaussian model:

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

Gradient-based variational inference:

$$q(\mathbf{z}_\ell|\mathbf{x}) = \mathcal{N}(\mathbf{z}_\ell; \boldsymbol{\mu}_{q,\ell}, \boldsymbol{\Sigma}_{q,\ell})$$
$$\nabla_{\boldsymbol{\mu}_{q,1}} \mathcal{L} = \mathbf{J}^\intercal \boldsymbol{\varepsilon}_\mathbf{x} - \boldsymbol{\varepsilon}_1$$

where $\mathbf{J} = \partial \boldsymbol{\mu}_{\theta,\mathbf{x}} / \partial \boldsymbol{\mu}_{q,1}$, and $\boldsymbol{\varepsilon}_\mathbf{x}$ and $\boldsymbol{\varepsilon}_1$ are weighted errors, e.g., $\boldsymbol{\varepsilon}_\mathbf{x} = \boldsymbol{\Sigma}_\mathbf{x}^{-1}(\mathbf{x} - \boldsymbol{\mu}_{\theta,\mathbf{x}})$.



Variational Autoencoders [Kingma & Welling, 2014; Rezende et al., 2014]:

- parameterize conditional probabilities with deep networks, and
- *learn* to perform variational inference optimization (*amortization*).

Deep networks:

e.g., $\boldsymbol{\mu}_{\theta,\ell}(\mathbf{z}_{\ell+1}) = \text{NN}_{\theta,\ell}(\mathbf{z}_{\ell+1})$

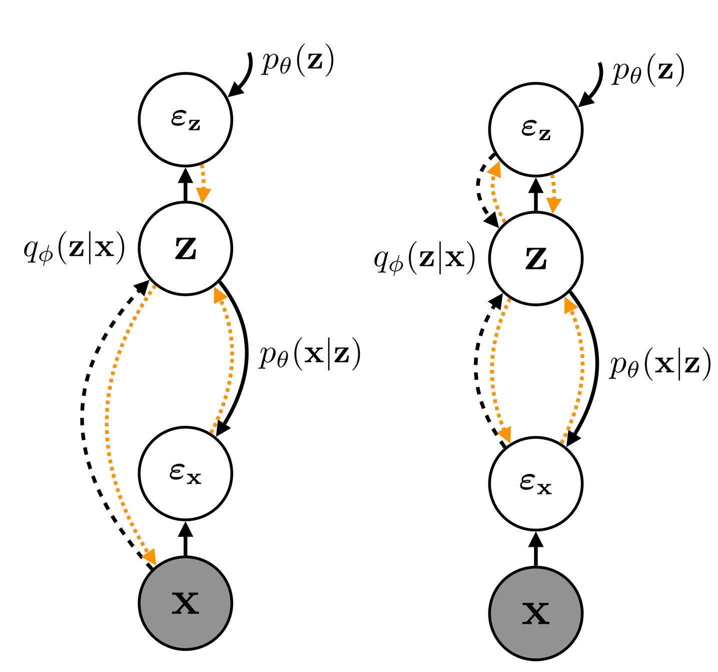
Amortized variational inference:

direct $\boldsymbol{\mu}_q \leftarrow \text{NN}_\phi(\mathbf{x})$

iterative $\boldsymbol{\mu}_q \leftarrow \text{NN}_\phi(\boldsymbol{\mu}_q, \nabla_{\boldsymbol{\mu}_q} \mathcal{L})$
or
$\boldsymbol{\mu}_q \leftarrow \text{NN}_\phi(\boldsymbol{\mu}_q, \boldsymbol{\varepsilon}_\mathbf{x}, \boldsymbol{\varepsilon}_\mathbf{z})$

Reparameterization:

$\mathbf{z} = \boldsymbol{\mu}_q + \boldsymbol{\sigma}_q \odot \boldsymbol{\gamma} \qquad \boldsymbol{\gamma} \sim \mathcal{N}(\boldsymbol{\gamma}; \mathbf{0}, \mathbf{I})$



direct *iterative*

connections & contrasts

Biological Connections

Neuroscience		Predictive Coding
Top-Down Neurons	—	*Generative Model*
Bottom-Up Neurons	—	*Inference Updating*
Lateral Connections	—	*Covariance Matrix*
Neural Activity	—	*Estimates & Errors*
Cortical Column	—	*Corresponding Estimate & Error*

Predictive Coding

- **Model**: Latent Gaussian Model
- **Model Parameterization**: Analytical, e.g., Polynomial
- **Approx. Posterior**: Typically Gaussian
- **Inference Optimization**: Gradient-Based
- **Dynamics**: Typically Generalized Coordinates, e.g., Velocity

Variational Autoencoders

- **Model**: Latent Gaussian Model
- **Model Parameterization**: Deep Neural Networks
- **Approx. Posterior**: Typically Gaussian
- **Inference Optimization**: Amortized
- **Dynamics**: Typically Recurrent Neural Networks

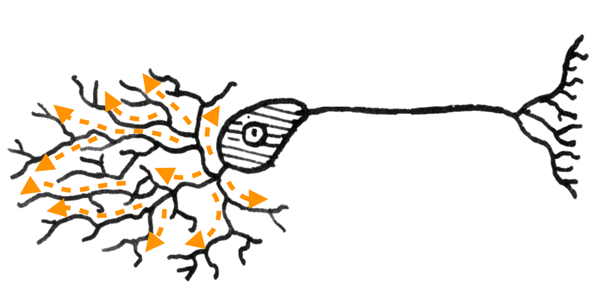
frontiers

Backpropagation *within* Neurons

- if a deep network is analogous to an individual neuron, then backprop-like mechanisms may occur *within* neurons

Credit assignment in networks using local prediction error signals

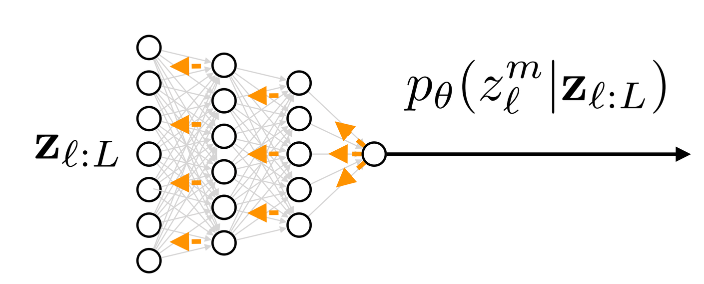
e.g., Target Prop. [Bengio, 2014; Lee et al., 2015]



Larger role for

→ ***non-linear dendritic computation***

→ ***backpropagating action potentials***



Normalizing Flows through Lateral Inhibition

- complex probability distributions with tractable sampling and evaluation

Basic Form:

base distribution $p_\theta(\mathbf{u})$

invertible transforms $\mathbf{v} = f_\theta(\mathbf{u})$

change of variables formula

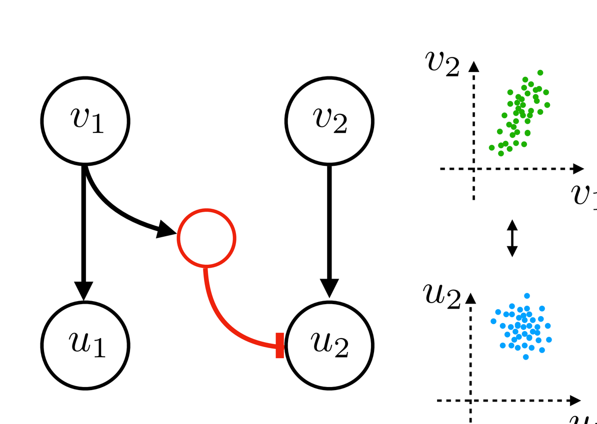
$$p_\theta(\mathbf{v}) = p_\theta(\mathbf{u}) \left| \det \left(\frac{d\mathbf{v}}{d\mathbf{u}} \right) \right|^{-1}$$

Affine Autoregressive Flows [Kingma et al., 2016]:

forward transform: $v_i = \alpha_\theta(\mathbf{v}_{<i}) + \beta(\mathbf{v}_{<i}) \cdot u_i$

inverse transform: $u_i = \dfrac{v_i - \alpha_\theta(\mathbf{v}_{<i})}{\beta(\mathbf{v}_{<i})}$

This basic normalization scheme is found in retina, thalamus, cortex, central pattern generators, etc.



Attention via Precision-Weighting

- prediction precision provides a mechanism for attention [Spratling, 2008]

higher precision => larger loss contribution => larger 'attentional' weight

$$\boldsymbol{\varepsilon}_\mathbf{x} = \boldsymbol{\Sigma}_\mathbf{x}^{-1}(\mathbf{x} - \boldsymbol{\mu}_{\theta,\mathbf{x}})$$
$$= \boldsymbol{\Pi}_\mathbf{x}(\mathbf{x} - \boldsymbol{\mu}_{\theta,\mathbf{x}})$$

may prove useful for integrating latent variable models with supervised tasks and reinforcement learning