# Training a dynamical system by using multivariate information

Clayton Seitz

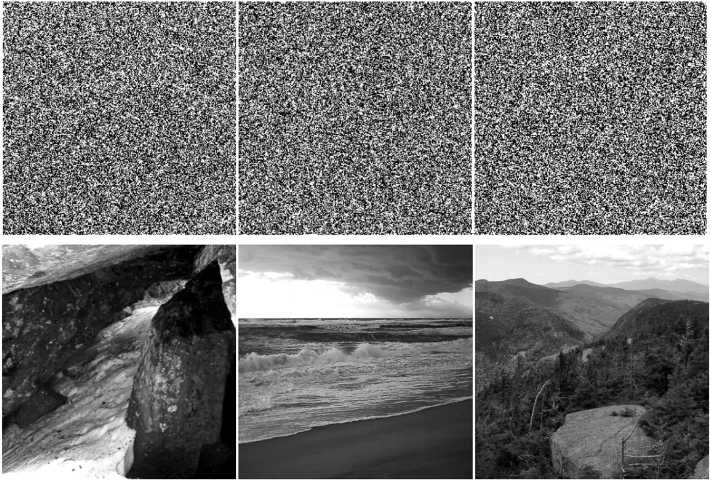
May 10, 2021

# Table of contents


1. Introduction
2. Channel coding for neural networks
3. Supervised training of low-rate critical networks
4. Multivariate information theory
5. Adaptation of the transfer function
6. Learning an energy function over phase space
7. Generalization bounds and density estimation
8. The energy function defines a dynamical system
9. The energy function is a generative model
10. Application to natural image statistics

# Introduction

Neuroethology argues that neural networks evolve according to the stimuli to which they are exposed



# Channel coding for neural networks

Networks of neurons can be viewed as a communication channel
Except this communication channel *learns* the transformation $F$ based on the statistical structure of its input $X$. Visual cortex has learned an encoding for visual scenes (that perhaps maximizes information)

# Leaky integrate and fire neurons

A realistic LIF model might look like

$$\tau_m \frac{d\mathrm{V}[I]}{dt} = (\mathrm{V}[I] - E)\sum_j \mathrm{W}^0[I,j] + (\mathrm{V}[I] - E_{in})\sum_k \mathrm{W}^1[I,k])$$

Instead, we ignore changes in the voltage of the postsynaptic neuron due to subthreshold voltages of the presynaptic neuron and let matrices W learn the input-output voltage relationship

$$V[j,t{+}1] = \alpha V[j,t{+}1] + \sum_{i \neq j} W^0_{ij} z[i,t] + \sum_i W^1_{ij} x[i,t{+}1] - z[j,t] v_{th}$$

where $z = H(v - v_{th})$

# Estimating gradients

Say we have a model $\Phi = (W^0, W^1)$ and want to use gradient descent to train a network to have a target rate or a target branching parameter. The rate and its associated loss for a single unit is

$$r(t) = \frac{1}{\Delta t} \int_t^{t+\Delta t} d\tau \langle \rho(\tau) \rangle \qquad \mathcal{L} = \alpha(r - r_0)^2$$

We would like the standard update

$$\Delta W_{ij} = -\eta \frac{\partial \mathcal{L}}{\partial W_{ij}}$$

But it is intractable to compute $\frac{\partial \mathcal{L}}{\partial W_{ij}}$ since $\rho(t)$ depends on other neurons through space and time.

# Estimating gradients

Bellec et al. presented a solution to estimating $\frac{\partial \mathcal{L}}{\partial W_{ij}}$ for online learning, but it can just as well be used for supervised learning

$$\frac{\partial \mathcal{L}}{\partial W_{ij}} = \sum_t \frac{\partial \mathcal{L}}{\partial z[j,t]} \cdot \frac{\partial z[j,t]}{\partial W_{ij}}$$

where the gradient $\frac{\partial z[j,t]}{\partial W_{ij}}$ is computed locally.

# Multivariate information theory

# Adaptation of the transfer function

How do neuron transfer functions adapt to stimuli in an unsupervised manner?

# Adaptation defines an energy function over phase space

# Generalization bounds

What is the distance of a code defined by a particular energy function E

# The energy function defines a dynamical system

# The energy function is a generative model

# Application to natural image statistics