Perception as inference
Natural scenes are full of ambiguity
Natural scenes are full of ambiguity
Generative models

Inference

\[ P(\alpha|D; \theta) \]

Parameters \( \theta \)

Model \( M \)

Prior

\[ P(\alpha; \theta) \]

Observed data \( D \)

\[ P(D|\alpha; \theta) \]

Explanation or prediction
Inference:

\[ P(\alpha|D; \theta) \propto P(D|\alpha; \theta) P(\alpha; \theta) \]

Explanation or prediction:

\[ P(D|\hat{\alpha}; \theta) \quad \text{with} \quad \hat{\alpha} = \arg \max_{\alpha} P(\alpha|D; \theta) \]

Objective for learning:

\[ \hat{\theta} = \arg \max_{\theta} \langle \log P(D|\theta) \rangle \]

\[ P(D|\theta) = \sum_{\alpha} P(D|\alpha; \theta) P(\alpha; \theta) \]
We can keep on going…

$$P(\theta|D) = \frac{P(D|\theta) P(\theta)}{P(D)}$$

- likelihood
- prior
- evidence

$$P(D) = \int P(D|\theta) P(\theta) d\theta$$
Figure 2.2: **Why Bayes embodies Occam’s razor**

This figure gives the basic intuition for why complex models are penalised. The horizontal axis represents the space of possible data sets $D$. Bayes’ rule rewards models in proportion to how much they *predicted* the data that occurred. These predictions are quantified by a normalised probability distribution on $D$. In this paper, this probability of the data given model $\mathcal{H}_i$, $P(D|\mathcal{H}_i)$, is called the evidence for $\mathcal{H}_i$.

A simple model $\mathcal{H}_1$ makes only a limited range of predictions, shown by $P(D|\mathcal{H}_1)$; a more powerful model $\mathcal{H}_2$, that has, for example, more free parameters than $\mathcal{H}_1$, is able to predict a greater variety of data sets. This means however that $\mathcal{H}_2$ does not predict the data sets in region $C_1$ as strongly as $\mathcal{H}_1$. Assume that equal prior probabilities have been assigned to the two models. Then if the data set falls in region $C_1$, the *less powerful* model $\mathcal{H}_1$ will be the *more probable* model.
Figure 2.3: The Occam factor
This figure shows the quantities that determine the Occam factor for a hypothesis \( \mathcal{H}_i \) having a single parameter \( w \). The prior distribution (dotted line) for the parameter has width \( \Delta^0 w \). The posterior distribution (solid line) has a single peak at \( w_{MP} \) with characteristic width \( \Delta w \). The Occam factor is \( \frac{\Delta w}{\Delta^0 w} \).

\[
P(D | \mathcal{H}_i) \simeq \underbrace{P(D | w_{MP}, \mathcal{H}_i)} P(w_{MP} | \mathcal{H}_i) \Delta w.
\]

Evidence \( \simeq \) Best fit likelihood \( \times \) Occam factor

Occam factor = \( \frac{\Delta w}{\Delta^0 w} \)
The “Boltzmann machine”
(Hinton & Sejnowski, 1983)

\[ E(s) = -\frac{1}{2} \sum_{ij} T_{ij} s_i s_j \]

\[ P(s) = \frac{1}{Z} e^{-\beta E(s)} \]
Boltzmann-Gibbs distribution

\[
P(x) = \frac{1}{Z} e^\lambda \phi(x)
\]

\[
Z = \int e^\lambda \phi(x) \, dx
\]

Learning rule:

\[
\Delta \lambda \propto \frac{\partial}{\partial \lambda} \langle \log P(x) \rangle
\]

\[
= \langle \phi(x) \rangle - \langle \phi(x) \rangle P(x)
\]
\[ \log P(\mathbf{x}) = \lambda \phi(\mathbf{x}) - \log Z \]

\[
\frac{\partial}{\partial \lambda} \langle \log P(\mathbf{x}) \rangle = \frac{\partial}{\partial \lambda} \langle \lambda \phi(\mathbf{x}) - \log Z \rangle \\
= \langle \phi(\mathbf{x}) - \frac{\partial}{\partial \lambda} \log Z \rangle \\
= \langle \phi(\mathbf{x}) - \frac{\partial}{\partial \lambda} \log \left( \frac{1}{Z} \frac{\partial Z}{\partial \lambda} \right) \rangle \\
= \langle \phi(\mathbf{x}) - \frac{1}{Z} \int \phi(\mathbf{x}) e^{\lambda \phi(\mathbf{x})} d\mathbf{x} \rangle \\
= \langle \phi(\mathbf{x}) - \int \phi(\mathbf{x}) P(\mathbf{x}) d\mathbf{x} \rangle \\
= \langle \phi(\mathbf{x}) \rangle - \langle \phi(\mathbf{x}) \rangle P(\mathbf{x}) \]
The “Boltzmann machine”  
(Hinton & Sejnowski, 1983)

\[ E(s) = -\frac{1}{2} \sum_{ij} T_{ij} s_i s_j \]

\[ P(s) = \frac{1}{Z} e^{-\beta E(s)} \]
The Boltzmann machine learning rule

\[ \Delta T_{i,j} \propto \frac{\partial \langle \log P(s) \rangle}{\partial T_{i,j}} \]

\[ = \beta \left[ \langle s_i s_j \rangle_{\text{clamped}} - \langle s_i s_j \rangle_{\text{free}} \right] \]

\text{data} \quad \quad \quad P(s)
“Boltzmann machine” with hidden units
(Hinton & Sejnowski)

\[
E(s^v, s^h) = - \sum_{i,j} T_{ij} s^v_i s^v_j - \sum_{i,j} T_{ij} s^v_i s^h_j - \sum_{i,j} T_{ij} s^h_i s^h_j
\]

\[
P(s^v, s^h) = \frac{1}{Z} e^{-E(s^v, s^h)}
\]

\[
P(s^v) = \sum_{s^h} P(s^v, s^h)
\]
The Boltzmann machine learning rule

\[ \Delta T_{ij} \propto \frac{\partial \log P(s)}{\partial T_{ij}} \]

\[ = \beta \left[ \langle s_i s_j \rangle_{\text{clamped}} - \langle s_i s_j \rangle_{\text{free}} \right] \]

Clamped:
\[ \begin{cases} 
  \mathbf{s}^v = \mathbf{x} \\
  \mathbf{s}^h \sim P(\mathbf{s}^h | \mathbf{s}^v) 
\end{cases} \]

Free:
\[ [s^v, s^h] \sim P(s^v, s^h) \equiv P(s) \]
Gibbs sampling

To sample from $P(x)$:

\begin{align*}
x_1 & \sim P(x_1|x_2, \ldots, x_n) \\
x_2 & \sim P(x_2|x_1, x_3, \ldots, x_n) \\
x_3 & \sim P(x_3|x_1, x_2, x_4, \ldots, x_n) \\
\vdots & \\
\vdots & \\
x_n & \sim P(x_n|x_1, \ldots, x_{n-1})
\end{align*}
Dynamics

\[ P(s_i = 1|\{s^{-}_i\}) = \frac{P(s_i = 1, \{s^{-}_i\})}{P(s_i = 1, \{s^{-}_i\}) + P(s_i = -1, \{s^{-}_i\})} \]

\[ = \frac{e^\beta \sum_{j \neq i} T_{ij} s_j}{e^\beta \sum_{j \neq i} T_{ij} s_j + e^{-\beta} \sum_{j \neq i} T_{ij} s_j} \]

\[ = \frac{1}{1 + e^{-2\beta} \sum_{j \neq i} T_{ij} s_j} \]

Thus:

\[ P(s_i = 1|\{s^{-}_i\}) = \sigma(2\beta h_i) \]

\[ h_i = \sum_{j \neq i} T_{ij} s_j \]
Application: modeling activity of neural populations (Schneidman et al.)

To describe the network as a whole, we need to write down a maximum entropy probability distribution for the population correlation that is captured by the conditional multi-information values. The rate of occurrence of each pattern is shown as a light blue dot. For the same sets of 10 cells, the fraction of predicted with better than 10% accuracy, and scatter between predictions and observations is confined largely to rare events for which the predictions and observations are strongly correlated.

We conclude that weak correlations among pairs of neurons coexist with strong correlations in the states of the population as a whole. One possible explanation is that there are specific multi-information values that represent the time of an action potential.

Synchronous spiking events in the 40 cell population in response to a long movie clip. Each dot represents the time of an action potential. Inset shows the same cross-correlogram on an expanded time scale; firing rate of one cell is plotted relative to the time at which the other cell spikes. The rate of occurrence of each pattern predicted from the maximum entropy model is also plotted (from Fig. 1f; grey dots). Black line shows equality.

ARTICLES
Restricted Boltzmann machine (RBM)

\[ E(s^v, s^h) = - \sum_{i,j} T_{ij}^v s^v_i s^h_j \]

\[ P(s^v, s^h) = \frac{1}{Z} e^{-E(s^v, s^h)} \]

\[ P(s^h_i | s^h_i, s^v) = \sigma(\sum_j T_{ij}^{hv} s^v_j) \]

\[ P(s^v_i | s^v_i, s^h) = \sigma(\sum_j T_{ij}^{vh} s^h_j) \]

\[ P(s^v) = \sum_{s^h} P(s^v, s^h) = \frac{1}{Z} e^{\sum_i \log(1+e^{T_{ih}^{vh} \cdot s^v})} \]
Modeling Higher-Order Correlations within Cortical Microcolumns

Urs Köster*, Jascha Sohl-Dickstein², Charles M. Gray³, Bruno A. Olshausen¹

¹ Redwood Center for Theoretical Neuroscience, University of California, Berkeley, Berkeley, California, United States of America, ² Department of Applied Physics, Stanford University and Khan Academy, Palo Alto, California, United States of America, ³ Department of Cell Biology and Neuroscience, Montana State University, Bozeman, Montana, United States of America

- Silicon polytrode, 32 channels span all laminae
- Anesthetized cat area V1
- Stimuli consist of natural scene movies at 150 fps
Experiment

- Simultaneous recordings of 20-40 well-isolated cells across all cortical layers of V1
- How can we discover patterns in the activity?

2s of raw data: 28 units discretized in 20ms time bins
From Correlations to Models

- Ising model: Models pairwise correlations with pairwise coupling

- Limitation: Cannot capture higher order structure

Ising Model: Pairwise couplings $J$ explain correlations between states $x$
Proposed Model

- Ising model, cells connect with pairwise coupling
- Additional hidden units
- Boltzmann Machine
- No connections between hidden units: Restricted Boltzmann Machine (RBM)
- Estimation of parameters is made efficient with Minimum Probability Flow (MPF)
Results: Model structure

- Ising model: Pairwise coupling parameters
- RBM with vertical connections only
- sRBM with horizontal connections between pairs

Hidden units are localized to cortical layer

Visible units (cells)
Model comparison

- Normalized probabilities with Annealed Importance Sampling for model comparison
- Model quality measured as likelihood gain over independent model
- Boltzmann machines with hidden units significantly outperform Ising models
- Spatiotemporal data sets to compare model dimensionality
Results: Pattern frequency

- Insight into where and how models fail
- Independent model fails on pairs of cells
- Ising model underestimates triplet activity
- RBMs capture all patterns well
Reducing the Dimensionality of Data with Neural Networks

G. E. Hinton* and R. R. Salakhutdinov

\[
E(v, h) = - \sum_{i \in \text{pixels}} b_i v_i - \sum_{j \in \text{features}} b_j h_j - \sum_{i,j} v_i h_j w_{ij}
\]

\[
\Delta w_{ij} = \epsilon \left( \langle v_i h_j \rangle_{\text{data}} - \langle v_i h_j \rangle_{\text{recon}} \right)
\]
Application to hand-written digits

2D PCA

784-1000-500-250-2 autoencoder
Collect pairwise statistics. Synthesize local 9-diagram pdf (3x3 blocks). Synthesize 'Lines world'.
Collect pairwise statistics

Collect local 9-dim. pdf (3x3 blocks)

Synthesize

Synthesize