Probabilities on Graphs: Undirected Graphs and Spatial Context

Alan Yuille and Dan Kersten

May 6, 2015

Overview

- This lecture describes a class of models suitable for modeling spatial context. Context is motivated by perceptual studies and by neuroscience findings. The lecture gives an approach which unifies several different methods.
- First, probabilistic models of neurons which extend the linear filter models described earlier in the course. This includes concenpts like Gibbs distribution and Gibbs sampling (a special case of Markov Chain Monte Carlo).
- Second, neurons as dynamic systems. "Hopfield" nets. Lyaponov functions.
- Third Bayesian distributions and energy function formulations.
- Fourth, Mean Field Theory.

Context and spatial interactions between neurons

- There is considerable evidence that low-level vision involves long-range spatial interaction,s so that human perception of local regions of an image can be strongly influenced by their spatial context.
- Psychophysicists have discovered many perceptual phenomena demonstrating spatial interactions.
- For example, local image regions that differ from their neighbors tend to "pop out" and attract attention, while, conversely, similar image features that form spatially smooth structures tend to get "grouped" together to form a coherent percept, see chapter figure 12.26 (left panel).
- Image properties such as color tend to spread out, or fill in regions, until they hit a boundary (Grossberg & Mingolla, 1985; Sasaki et al., 2004) as shown in chapter figure 12.26 (right panel).

Context and spatial Interactions between neurons

- There is a tendency for low-level vision to group similar image features and make breaks at places where the features change significantly. These correspond to low-level visual tasks, such as segmentation and the detection of salient features. Segmenting an image into different regions is one of the first stages of object recognition (in the ventral stream) and a precursor to estimating the three-dimensional structure of objects, or surfaces, in order to grasp them or avoid them (dorsal stream).
- Detection of salient features has many uses, including bottom-up attention (Itti & Koch, 2001). It has been suggested that many of these processes are performed in V1 (Zhaoping, 2014), although this involves possibly feedback and interactions between V1 and V2 (Shushruth et al., 2013).

Context figures

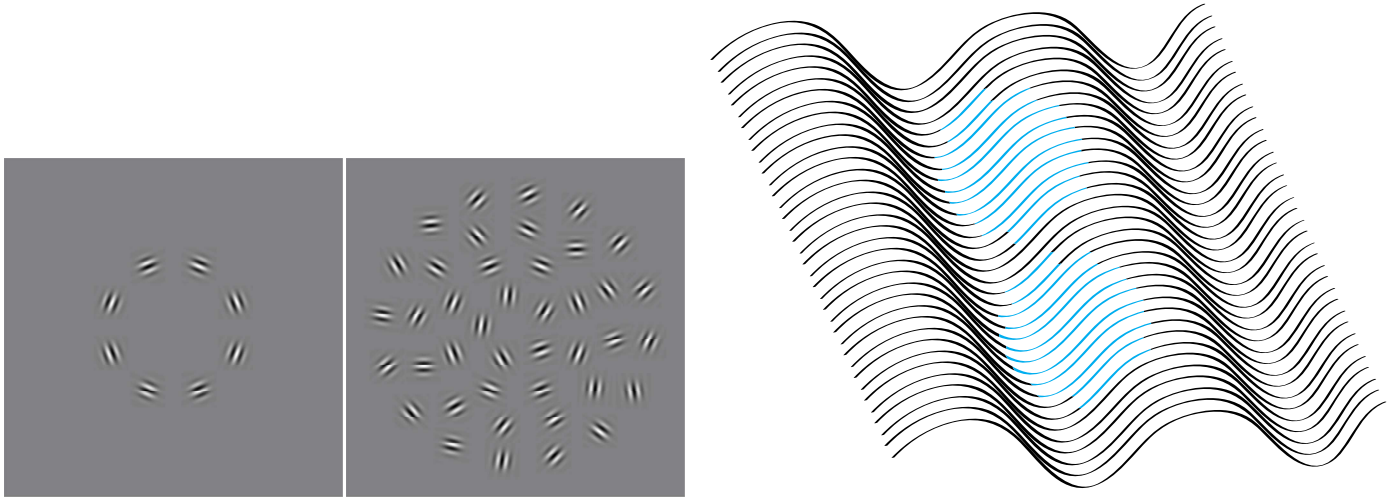

Figure 1: Left: Association fields. The circular alignment of Gabor patches (left) make it easier to see the circular form in the presence of clutter (right). Right: The neon color illusion. A bluish color appears to fill in the white regions between the blue lines, creating the appearance of blue transparent disks.

d

Context electrophysiology

- The psychophysical and theoretical studies are consistent with multi-electrode studies (Lamme, 1995) and later work by T.S. Lee which relates these findings to the theoretical models we will describe. These studies show that the activities of neurons on monkey area V1 appear to involve spatial interactions with other neurons.
- When monkeys are shown stimuli consisting of a textured square surrounded by a background with a different texture, their responses over the first 60 msec are similar to those predicted by classic models (e.g., previous sections), but their later activity spreads in from the boundaries, roughly similar to predictions of computational models (Lee & Yuille, 2006).
- There is also a considerable literature on the related topic of *nonclassical receptive fields* (Kapadia et al., 2000).

Context electrophysiology

- These pyschophyical and neuroscience findings are not consistent with feedforward models like convolutional neural networks. They require models with sideways interactions and/or algorithms which perform feedforward and feedback processing. Transformer neural network architectures have some potential for modeling these phenomena.
- These lectures will describe a popular class of models for sideways interactions. They are formulated in Bayesian terms and are undirected graphical models.
- They can also be thought of as neural network models, with neurons interacting with each other.
- They can be related to an alternative class of neural networks which are formulated by dynamical systems, and had been proposed for modeling these types of findings (Grossberg & Minolla 1985).
- The models can be used as computer vision models for vision models like binocular stereo, motion estimation, and others. And they can serve as neural network models for describing biological phenomena.

First: Probabilistic models of neurons

- We start by introducing probabilistic models of neurons. These are natural extensions of the linear filters described in earlier lectures.
- We introduce important concepts like Gibbs Distributions and Gibbs Sampling.
- We discuss the difference between stochastic models of neurons, where the neuron response is probabilistic, with deterministic models.

Single neurons: Probabilistic model and integrate and fire (I)

- We have described neurons as linear filters and briefly mentioned thresholds and nonlinearities.
- In this section, we provide a more realistic model of a *stochastic neuron*, where the neuron has a probability of firing an action potential. We will show how linear filters, thresholds, and nonlinearities can be obtained as approximations to this stochastic model.
- This stochastic model is, in turn, an approximation, and we refer to the literature for more realistic models, such as assuming that the probability of firing is specified by a Poisson process (Rieke et al., 1997).
- For simplicity, we restrict ourselves to the simpler stochastic *integrate-and-fire* model, which is easier to analyze and to relate to computational models.

Single neurons: Probabilistic model and integrate and fire (II)

- In the integrate-and-fire model, a neuron i receives input I_j at each dendrite j. These inputs are weighted by the synaptic strengths w_{ij} and sent along the dendrites to the soma. At the soma, these weighted inputs are summed linearly to yield $\sum_j w_{ij} I_j$.
- The probability of firing $s_i = 1$ is given by $P(s_i = 1|\vec{I}) = \frac{\exp\{(\sum_j w_{ij} I_j - T_i)\}}{1+\exp\{\sum_j w_{ij} I_j - T_i\}}$ where T is a threshold.
- The probability of not firing is $P(s_i = 0|\vec{I}) = \frac{1}{1+\exp\{\sum_j w_{ij} I_j - T_i\}}$.
- More concisely,

$$P(s_i|\vec{I}) = \frac{\exp\{s_i(\sum_j w_{ij} I_j - T_i)\}}{1 + \exp\{\sum_j w_{ij} I_j - T_i\}}$$

Relations to the stochastic model (I)

- To relate this stochastic model to our earlier linear models, we calculate the probability that the neuron will fire. This is given by a linear filter followed by a non-linear sigmoid function $\sigma()$:

$$\sum_{s_i=0}^{1} s_i P(s_i|\vec{I}) = \frac{1}{1+\exp\{\sum_j w_{ij} I_j - T_i\}} = \sigma(\sum_j w_{ij} I_j - T_i). \qquad (1)$$

- Observe that this is also the *expected firing rate* $\sum_{s_i=0,1} s_i P(s_i|\vec{I})$ because

$$\sum_{s_i=0,1} s_i P(s_i|\vec{I}) = P(s_i = 1|\vec{I}) = \sigma(\sum_j w_{ij} I_j - T_i). \qquad (2)$$

Relations to the stochastic model (II)

- By computing the expected firing rate, we obtain a deterministic approximation to a stochastic neuron. This is a sigmoid function of a linear weighted sum of the input (minus a threshold).
- The sigmoid function is approximately linear for small inputs, saturates at value 1 for large positive inputs, and suppresses large negative inputs to 0. Hence there is a linear regime where the probability of firing is $\sum_j w_{ij} I_j - T_i$. This enables us to recover the linear models used in the previous section as an approximation.
- We will extend this to models of a set of neurons which interact with each other. So that neurons receive input from other neurons as well as from images (or other stimuli).

Relations to the stochastic model (III)

- Suppose we have a set of neurons $\{i\}$ with activations $\vec{S} = \{s_i\}$.
- We modify the firing rule so that a neuron is driven by external input $\{I_i\}$ and by the other neurons. This gives stochastic firing rule for each neuron:

$$P(s_i|\vec{I}, \vec{S}_{/i}) = \frac{1}{Z_i} \exp\{s_i(\sum_j w_{ij} I_j + \sum_{k \neq i} \theta_{ik} s_k)\} \tag{3}$$

- There are weights w_{ij} from the external input $\vec{I}$ and weights θ_{ik} from the other neurons (the neuron receives no input from itself (hence $\sum_{ki/}$
- This specifies stochastic dynamics for a set of neurons. But what is the purpose? And what does this dynamics correspond to?

Probabilistic models of groups of neurons. (I)

- We now introduce two key concepts. The Gibbs Distribution and Gibbs Sampling.
- The Gibbs Distribution for a set of neurons, or more generally for a probability distribution on an undirected graph, is specified by defining a probability distribution in terms of an energy functions.
- Suppose we have set of M neurons with states $\vec{S} = (s_1, ..., s_M)$ and with input $\vec{I} = (I_1, ..., I_N)$. We define an energy function:

$$E(\vec{S}, \vec{I} : \vec{W}, \vec{\theta}) = -\sum_{ij} W_{ij} s_i I_j - (1/2) \sum_{kl} \theta_{kl} s_k s_l.$$

- We specify a *Gibbs probability distribution* over the set of activity of all neurons $\vec{S} = (s_1, ..., s_n)$ by:

$$P(\vec{S}, \vec{I}) = \frac{1}{Z} \exp\{-E(\vec{S}, \vec{I} : \vec{W}, \vec{\theta})\}.$$

- Probability Distributions on graphs can always (occasional exceptions) be expressed in terms of Gibbs distributions. For undirected graphs this is the only way to express them, see Hammersley-Clifford Theorem.

Probabilistic models of groups of neurons. (II)

- The energy $E(\vec{S}, \vec{I} : \vec{W}, \vec{\theta})$ contains two types of terms: (1) those of form $s_i I_j$, which give the interactions between the states of the neurons $\vec{S}$ and the input $\vec{I}$, and (2) those that specify interactions between the neurons. This energy is used to specify a Gibbs distribution:
- Here Z is a normalization constant chosen to ensure that $\sum_{\vec{S}} P(\vec{S}|\vec{I}) = 1$. Hence $Z = \sum_{\vec{S}} P(\vec{S}, \vec{I})$.
- Observe that low energy states, where $E(\vec{S}, \vec{I})$ is small, correspond to states with high probability $P(\vec{S}|\vec{I})$. So formulating a problem in terms of minimizing an energy function can be reforulates in terms of maximizing the probability.
- The Gibbs distribution arose in statistical physics, to specify the probability distribution of a physical system in thermal equilibrium. Here the physical energy of the system is E, and the distribution can be derived using the maximum entropy principle.

Probabilistic models of groups of neurons. (III)

- The weights $\{w_{ij}\}, \{\theta_{kl}\}$ specify the strength of the interactions between the neuron and the inputs, and between the neurons and each other.
- The *interaction term* $\sum_{kl} \theta_{kl} s_k s_l$ specifies the interactions between the neurons. If this term is not present, then the distribution simplifies and can be expressed as a product of independent distributions:

$$P(\vec{S}|\vec{I}) = \frac{1}{Z} \exp\{\sum_{ij} w_{ij} s_i I_j\} = \prod_{i=1}^{n} P(s_i|\vec{I}).$$

- In this special case, the neurons act independently and are driven purely by the input. Thw normalization factor in this case can be computed directly as $Z = \prod_i Z_i$, where $Z_i = \sum_{s_i=0}^{1} \exp\{\sum_j w_{ij} s_i I_j\}$.

Stochastic dynamics (I)

- Now we specify stochastic dynamics on this model. These dynamics have two purposes: first, to describe the activities of sets of neurons interacting with each other; second, to provide algorithms for estimating properties, such as the most probable configurations of the states $\vec{S}$, which can be used for visual tasks.
- the probability that the cell i fires is:

$$P(s_i|\vec{I}, \vec{S}_{/i}) = \frac{1}{Z_i} \exp\{s_i(\sum_j w_{ij} I_j + \sum_{k \neq i} \theta_{ik} s_k)\} \tag{4}$$

  where the notation $\vec{S}/i$ means the states $\{s_j : j \neq i\}$ of all the neurons except the neuron we are considering.
- At each time, a neuron is selected at random and fires with a probability specified by equation (4). This model assumes that no neurons ever fire at the same time and ignores the time for a spike fired from one neuron to reach other neurons.

Relations to Gibbs distribution?

- How does this stochastic dynamics relate to the Gibbs distribution? It can be shown, using the theory of *Markov Chain Monte Carlo* (MCMC) sampling (Liu, 2008), that this update rule converges to a random sample $\vec{S}^*$ from the distribution $P(\vec{S}|\vec{I})$. So S^* is likely to have high probability probability.
- This is known as *Gibbs sampling*, because it samples from the conditional distribution $P(s_i|\vec{I}, \vec{S}_{/i})$. These samples enable us to estimate the most probable state of the system $\hat{\vec{S}} = \arg\max P(\vec{S}|\vec{I})$, hence they can estimate the MAP estimator of $\vec{S}$ and make optimal decisions for visual tasks.
- Gibbs sampling can also be used for estimating statistical properties of the probability distribution, as illustrated in a later lecture about the Boltzmann Machine.

Markov structure

- Formally, the edges of the graph define the *Markov structure* of the probability distribution $P(\vec{S})$. It can be shown that the conditional distribution of the state $s(\vec{x})$ at one position depends *only* on the states of positions in its neighborhood $N(\vec{x})$. This is the *Markov condition*:

$$P(s(\vec{x})|\vec{S}/s(\vec{x})) = P(s(\vec{x})|\{s(\vec{y}) : \vec{y} \in N(\vec{x})\}),$$

  where $\vec{S}/s(\vec{x})$ denotes all states in $\vec{S}$ except $s(\vec{x})$.
- In real vision applications, this type of prior, including the size of the neighborhoods, can be estimated from the statistics of natural images.

Dynamical system models of neurons (I)

- There is an alternative way to model sets of neurons using *dynamical systems* based on simplified models of their biophysics (Rieke et al., 1997; Dayan & Abbbott, 2001). Pioneering work on this topic was done by Wilson and Cowan (1972), Grossberg and Mingolla (1968, 1985), Hopfield and Tank (1986), Abbott and Kepler (1990), and others.
- There is no space to cover the richness of these models, and in any case, these lectures concentrate on the probabilistic formulation. But we will discuss an important subclass of dynamical models (Hopfield & Tank, 1986) that, as we will show, has very close relations to the probabilistic approach.

Dynamical system models of neurons (II)

- These dynamical systems are described as follows (Hopfield & Tank, 1986). A neuron is described by two (related) variables: (1) a continuous valued variable $u_i \in \{-\infty, \infty\}$, and (2) a continuous variable $q_i \in \{0, 1\}$. Roughly speaking, u_i represents the input to the cell body (soma), both direct input and input from other neurons and q_i describes the probability that the cell will fire an action potential. These variables are related by the equations $u_i = \log(q_i/(1 - q_i))$ or, equivalently, by $q_i = \sigma(u_i)$ (where $\sigma(.)$ is the sigmoid function).
- The dynamics of the neuron is given by:

$$\frac{du_i}{dt} = -u_i + \sum_j w_{ij} I_j + \sum_k \theta_{ik} q_k. \tag{5}$$

- Here, as before, $\sum_j w_{ij} I_j + \sum_k \theta_{ik} q_k$ represent the direct input and the input from the other neurons.

Dynamical system models of neurons (III)

This dynamic system continually decreases a function $F(\vec{q})$, so that $(dF)/dt \leq 0$. The function F acts as a *Lyapunov function* for the system in the sense that it decreases monotonically as time t increases and is bounded below. The existence of a Lyapunov function for the dynamics guarantees that the system will converge to a state that minimizes $F(\vec{q})$ (note that $F(\vec{q})$ will typically have many minimums, and the system may converge to any one of them).

Relations between probabilistic models and dynamical system models (I)

- Perhaps surprisingly, there is a very close relationship between the dynamic systems in equation (5) and the stochastic update in equation (**??**). More specifically, the dynamic system is a mean field approximation to the stochastic dynamics. *Mean field theory* (MFT) was developed by physicists as a way to approximate stochastic systems.
- To explain this relationship, we first define *the mean field free energy* $F(\vec{q})$:

$$F(\vec{q}) = -\sum_{ij} W_{ij} I_j q_i - (1/2) \sum_{ij} \theta_{ij} q_i q_j + \sum_i \{q_i \log q_i + (1-q_i) \log(1-q_i)\}. \tag{6}$$

- Next we specify dynamics by performing steepest descent on the free energy (multiplies by a positive factor):

$$\frac{dq_i}{dt} = -q_i(1-q_i)\frac{\partial F(\vec{q})}{\partial q_i}. \tag{7}$$

Relations between probabilistic models and dynamical system models (II)

- Interestingly, these are identical to the dynamical system in equation (5). This can be seen by introducing a new variable $u_i = \log q_i/(1-q_i)$, which implies that $q_i = \sigma(u_i)$. Note that $\partial F/\partial q_i = -\sum_j W_{ij} I_j - \sum_j \theta_{ij} q_j + \log q_i/(1-q_i)$, $u_i = \log q_i/(1-q_i)$, and $dq_i/q_i(1-q_i) = du_i$.
- Equation (7) implies that the dynamical system decreases the free energy $F(\vec{q})$ monotonically with time t. This is because $dF/dt = -\sum_i (\partial F/\partial q_i)(\partial q_i/\partial t) = -\sum_i q_i(1-q_i)(\partial F/\partial q_i)^2$. Hence $F(\vec{q})$ is a Lyapunov function for equations (5, 7), and so the dynamics converges to a fixed point.

Relations between probabilistic models and dynamical system models (III)

- This shows that there is a close connection between the neural dynamical system and minimizing the mean field free energy. In turn, the mean field free energy is related to deterministic approximations to stochastic update methods like Gibbs sampling (Amit, 1992; Hertz, 1991). This connection is technically advanced and is not needed to understand the rest of this chapter.
- Briefly, the mean field free energy $F(\vec{q})$ is the *Kullback-Leibler divergence* $F(Q) = \sum_{\vec{S}} Q(\vec{S}) \log \frac{Q(\vec{S})}{P(\vec{S}|\vec{I})}$ between the distribution $P(\vec{S}|\vec{I})$ and a factorized distribution $Q(\vec{S}) = \prod_i q_i^{S_i}(1-q_i)^{1-S_i}$ (plus an additive constant). Hence the dynamical system seeks to find the factorized distribution $\hat{Q}(\vec{S})$ that best approximates $P(\vec{S}|\vec{I})$ by minimizing the Kullback-Leibler divergence. In this approximation the response q_i is an approximation to the expected response $\sum_{S_1} S_i P(\vec{S}|\vec{I})$. The connections between mean field theory and neural models was described in Yuille, 1987). For technical discussions about mean field theory and Gibbs sampling see (Yuille, 2011).