

Context and spatial interactions between neurons

- ▶ There is considerable evidence that low-level vision involves long-range spatial interactions, 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

- ▶ In general, there is a tendency for low-level vision to group similar image features and make breaks at places where the features change significantly. These perceptual phenomena are not surprising from a theoretical perspective since they 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 26 : 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 discussed so far are supported by single-electrode studies (Lamme, 1995; Lee & Yuille, 2006), which 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 (Yuille, 2006). There is also a considerable literature on the related topic of *nonclassical receptive fields* (Kapadia et al., 2000).

Neural network models

This section discusses neural network models that address these phenomena. Although the models capture the essence of the phenomena, they are simplifications in three respects. First, they use simple models of neurons, and it is currently not possible to compare them directly to real neural circuits. Second, these models are formulated in terms of lateral, or horizontal, connections. Third, the performance of these models on natural images is significantly worse than a human's. There are more advanced computer vision models, built on similar principles, whose performance starts to approach human vision (unless high-level cues are present, which humans can exploit).

Probability distributions on graphs

We formulate these models in terms of probability distributions defined over graphs, where the nodes of the graph represent neurons. This differs from some of the standard “neural network” models for these types of phenomena, see (Grossberg & Mingolla, 1985). but our approach has several advantages. First, this enables us to use a coherent framework that unifies the models in this section with those we will discuss in later sections. Second, it puts the models in a form that can be directly related to a class of computer vision models. Third, this probabilistic formulation is of increasing use in models of artificial intelligence, cognitive science, and the machine learning and statistical techniques used to analyze experimental neuroscience data. Fourth, it is possible to derive many of these neural network models as approximations to the probability models.

Probabilistic models of neurons

- ▶ We first introduce probabilistic models of neurons and show how our previous linear filter models can be derived as approximations. Next we introduce neural network models and show their relationship to probability models. Then we use this material to derive some specific models for a range of visual tasks.

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 l_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 $\sum_j w_{ij}l_j$. The probability of firing $s_i = 1$, or not firing $s_i = 0$, is given by:

$$P(s_i|\vec{l}) = \frac{\exp\{s_i(\sum_j w_{ij}l_j - T_i)\}}{1 + \exp\{\sum_j w_{ij}l_j - T_i\}}, \quad (23)$$

where T_i is a threshold.

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 sigmoid function:

$$\sum_{s_i=0}^1 s_i P(s_i|\vec{I}) = \frac{1}{1 + \exp\{\sum_j w_{ij} l_j - T_i\}} = \sigma(\sum_j w_{ij} l_j - T_i). \quad (24)$$

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

$$\sum_{s_i=0,1} s_i P(s_i|\vec{I}) = P(s_i = 1|\vec{I}) = \sigma(\sum_j w_{ij} l_j - T_i). \quad (25)$$

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.
- ▶ Next we modify the model so that it deals with nonlinear image features. This allows us to relate it to the types of computational models described in the previous section and will enable us to construct richer models of this type that can deal with spatial context.

Enhancing the model to allow complex input

- ▶ Consider detecting if there is an edge at pixel x . Formulate the problem as Bayes estimation with conditional distributions $P(f(I(x))|s)$ and priors $P(s)$ for $s \in \{0, 1\}$. The posterior distribution $P(s|f(I(x)))$ can be expressed in the form:

$$P(s|f(I(x))) = \frac{1}{Z} \exp\left\{s \left(\log \frac{P(f(I(x))|s=1)}{P(f(I(x))|s=0)} + \log \frac{P(s=1)}{P(s=0)} \right)\right\},$$

where Z is a normalization constant (chosen so that $\sum_{s=0}^1 P(s|f(I(x))) = 1$).

- ▶ This shows that the posterior distribution for the presence of an edge can be expressed in the same form. The only difference is that the input is a nonlinear function of the image instead of the image itself.
- ▶ This claim can be justified by expressing $P(f(I(x))|s) = \{P(f(I(x))|s=1)\}^s \{P(f(I(x))|s=0)\}^{1-s}$, $P(s) = \{P(s=1)\}^s \{P(s=0)\}^{1-s}$, then substituting these into the posterior $P(s|f(I(x))) = P(f(I(x))|s)P(s)/P(f(I(x)))$.

Probability models with context

- ▶ Now apply the model to foreground/background classification and modify it to include spatial context. Intuitively, neighboring pixels in the image are likely to be either all background or all foreground. This is a form of prior knowledge that can be learned by analyzing natural images.
- ▶ We specify neurons by spatial position \vec{x} instead of index i . As above, we have distributions $P(f(I(\vec{x}))|s)$ for the features $f(I(\vec{x}))$ at position \vec{x} conditioned on whether this is part of the foreground object $s(\vec{x}) = 1$, or not, $s(\vec{x}) = 0$. We use the notation \vec{S} to be the set of the states of all neurons $\{s(\vec{x})\}$. We also specify a prior distribution:

$$P(\vec{S}) = \frac{1}{Z} \exp\left\{-\gamma \sum_{\vec{x}} \sum_{\vec{y} \in N(\vec{x})} \{s(\vec{x}) - s(\vec{y})\}^2\right\},$$

where γ is a constant. This prior uses a neighborhood $N(\vec{x})$, which specifies those spatial positions that directly interact with \vec{x} in the model. In graphical terms, the positions \vec{x} are the nodes \mathcal{V} of a graph \mathcal{G} , and the edges \mathcal{E} specify which nodes are connected.

Markov structure (I)

- ▶ 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.

Markov structure (II)

- ▶ Next, we define a probability model for the observed image features at positions \vec{x} in the image. We use the same models as before, at each position \vec{x} :

$$P(f(I(\vec{x}))|s) = \{P(f(I(\vec{x}))|s = 1)\}^s \{P(f(I(\vec{x}))|s = 0)\}^{1-s}.$$

- ▶ We combine these, using independence assumptions, to get a distribution:

$$P(f(\vec{I})|\vec{S}) = \prod_{\vec{x}} P(f(I(\vec{x}))|s) = \frac{1}{Z_I} \exp\left\{\sum_{\vec{x}} s(\vec{x}) \left(\log \frac{P(f(I(\vec{x}))|s = 1)}{P(f(I(\vec{x}))|s = 0)}\right)\right\},$$

where Z_I is a normalization term (which can be calculated directly).

Posterior distribution (I)

- ▶ These distributions $P(f(\vec{I})|\vec{S})$ and $P(\vec{S})$ can be combined to get the posterior distribution $P(\vec{S}|f(\vec{I}))$, which is of form:

$$P(\vec{S}|f(\vec{I})) = \frac{1}{Z_p} \exp\{-E(\vec{S})\},$$

where

$$E(\vec{S}) = - \sum_{\vec{x}} s(\vec{x}) \log \frac{P(f(I(\vec{x}))|s=1)}{P(f(I(\vec{x}))|s=0)} + \sum_{\vec{x}} \sum_{\vec{y} \in N(\vec{x})} \gamma \{s(\vec{x}) - s(\vec{y})\}^2.$$

- ▶ The first term of $E(\vec{S})$ gives the local cues for foreground or background (the log-likelihood ratios of the features), while the second term adds the local context. This context encourages neighboring positions to be either all foreground or all background. Note that this method of specifying a distribution $P(\vec{S})$ in terms of a function $E(\vec{S})$ will keep reoccurring throughout this section.

Posterior distribution (II)

- ▶ This model specifies the posterior distribution for foreground-background classification using spatial context, and as we will show, similar methods can be applied to other visual tasks. But there remains the issue of how to estimate the most probable states, i.e., computing the Bayes estimator.

$$\hat{\vec{S}} = \arg \max P(\vec{S} | f(\vec{I})).$$

- ▶ In the next two sections we discuss neurally plausible algorithms that can do this. There are two types: (1) stochastic models that are natural extensions of the probabilistic neural models discussed earlier, which in the statistics literature are called *Gibbs* samplers (Liu, 2008), and (2) neural network models that are based on simplified biophysics of neurons but that can also, in certain cases, be related to *mean field approximations* to the stochastic models.

Graphical model figures

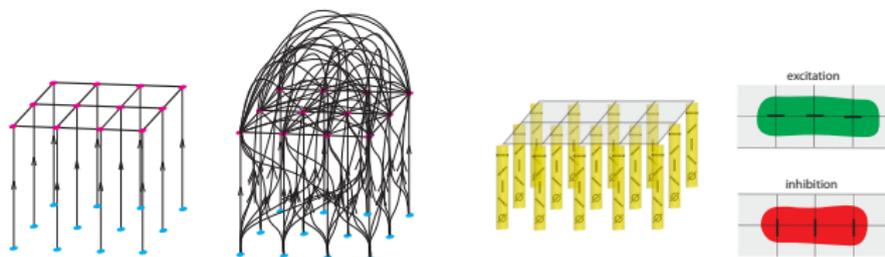


Figure 27 : Far left: The graphical structure of the Markov model with nearest neighbor connections. Left: A fully connected graphical model. Right: A hyper-column structure in which neurons within each column are tuned to different orientations and inhibit each other. Far right: Edges have excitation (green) along the direction of the edge and inhibition (red) perpendicular to the edge.

Probabilistic models of groups of neurons. (I)

- ▶ Here, we introduce a more general probability distribution. It is also specified by a model defined over a graph where the nodes correspond to neurons and the edges to connections between them. But we will not make any Markov restrictions on the edges, so this model can be fully connected.
- ▶ More specifically, we have set of M neurons with states $\vec{S} = (s_1, \dots, s_M)$ and with input $\vec{I} = (I_1, \dots, I_N)$. We specify a *Gibbs probability distribution* over the set of activity of all neurons $\vec{S} = (s_1, \dots, s_n)$ as follows. First we define an energy function:

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

Probabilistic models of groups of neurons. (II)

- ▶ This energy contains two types of terms: (1) those of form $s_i l_j$, 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:

$$P(\vec{S}, \vec{I}) = \frac{1}{Z} \exp\{-E(\vec{S}, \vec{I}; \vec{W}, \vec{\theta})\}. \quad (26)$$

- ▶ Here Z is a normalization constant chosen to ensure that $\sum_{\vec{S}} P(\vec{S}|\vec{I}) = 1$. Note that Gibbs distribution originally 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. In particular, 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:

$$P(\vec{S}|\vec{I}) = \frac{1}{Z} \exp\left\{\sum_{ij} w_{ij} s_i l_j\right\} = \prod_{i=1}^n P(s_i|\vec{I}). \quad (27)$$

- ▶ Hence in this special case, the neurons act independently and are driven purely by the input (i.e., there is no context). As a technical point, the normalization factor in this case can be computed directly as $Z = \prod_i Z_i$, where $Z_i = \sum_{s_i=0}^1 \exp\{\sum_j w_{ij} s_i l_j\}$.

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 and for making decisions.
- ▶ To specify stochastic dynamics, we generalize the stochastic neural model, equation (23), to deal with a set of neurons. A neuron received input \vec{S} from other neurons in addition to direct input from the stimulus \vec{I} . Consider only the activity of this neuron, fixing the states of all the others. Then the neurons will have total input of $\sum_j w_{ij} I_j$ plus input $\sum_k \theta_{ik} s_k$ from the other neurons.

Stochastic dynamics (II)

- ▶ Then, extending equation (23), the probability that the cell i fires is:

$$P(s_i | \vec{I}, \vec{S}_{/i}) = \frac{1}{Z_i} \exp\left\{s_i \left(\sum_j w_{ij} I_j + \sum_{k \neq i} \theta_{ik} s_k \right)\right\} \quad (28)$$

where the notation $\vec{S}_{/i}$ means the states $\{s_j : j \neq i\}$ of all the neurons except the neuron we are considering. The term Z_i is defined so that the distribution is normalized, so it is given by

$$Z_i = 1 + \exp\left\{ \sum_j w_{ij} I_j + \sum_{k \neq i} \theta_{ik} s_k \right\}.$$

- ▶ This gives the following dynamics for a group of neurons. At each time, a neuron is selected at random and fires with a probability specified by equation (28). 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 specified above? From the statistical perspective, this is an example of *Markov Chain Monte Carlo* (MCMC) sampling (Liu, 2008). MCMC refers to a class of algorithms that explore the state space of \vec{S} stochastically so that it will gradually move to configurations that have high probability $P(\vec{S}|\vec{I})$. More precisely, MCMC algorithms are guaranteed to give samples from the Gibbs distribution — $\vec{S}_1, \dots, \vec{S}_M \sim P(\vec{S}|\vec{I})$. The stochastic update rule in equation (28) is a special type of MCMC algorithm known as a *Gibbs sampler*, 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 $\vec{S} = \arg \max P(\vec{S}|\vec{I})$, hence they can estimate the MAP estimator of \vec{S} and make optimal decisions for visual tasks.

Learning and Boltzmann machines

To apply these models to visual tasks, we need to specify the weights. One strategy is purely data driven and consists of learning the weights from training examples. This is the *Boltzmann machine* (Ackley et al., 1985) which is out of scope for this chapter. Another strategy is to specify distributions for specific visual tasks, and we will give examples in the next few sections.

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 & Abbott, 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 $\sigma(\cdot)$ is the sigmoid function).
- ▶ The dynamics of the neuron is given by:

$$\frac{du_i}{dt} = -u_i + \sum_j w_{ij} l_j + \sum_k \theta_{ik} q_k. \quad (29)$$

- ▶ Here, as before, $\sum_j w_{ij} l_j + \sum_k \theta_{ik} q_k$ 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 (29) and the stochastic update in equation (23). 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* $F(\vec{q})$:

$$F(\vec{q}) = - \sum_{ij} W_{ij} l_j q_i - (1/2) \sum_{ij} \theta_{ij} q_i q_j + \sum_i \{q_i \log q_i + (1 - q_i) \log(1 - q_i)\}. \quad (30)$$

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

$$\frac{dq_i}{dt} = -q_i(1 - q_i) \frac{\partial F(\vec{q})}{\partial q_i}. \quad (31)$$

Relations between probabilistic models and dynamical system models (II)

- ▶ Interestingly, these are identical to the dynamical system in equation (29). 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 $\partial F / \partial q_i = -\sum_j W_{ij} l_j - \sum_j \theta_{ij} q_j + \log q_i / (1 - q_i)$, $u_i = \log q_i / (1 - q_i)$, and $dq_i / q_i (1 - q_i) = du_i$.
- ▶ Equation (31) 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 (29, 31), 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 $\sum_{S_1} S_1 P(\vec{S}|\vec{I})$. 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).