Lecture 10: MRF Examples: Weak Membrane, MFT, Deterministic Annealing

- This lecture describes the MRF formulation of the weak-membrane model. It describes how mean-field theory algorithms can be used to estimate the minimum of the free energy.
- The lecture also describes Grab-Cut, which is a binary MRF for segmenting a foreground object from the background (requiring human initialization – recall CPMC). This relates to graph-cut algorithms – max-flow/min-cut.

The line process model (I)

- Our first example is the classic *line process* model (Geman & Geman, 1984; Blake & Zisserman, 2003) which is an MRF formulation of the weak-membrane model, developed as a way to segment images. It has explicit *line process* variables that "break" images into regions where the intensity is piecewise smooth. Our presentation follows the work of Koch et al. (1986), who translated it into neural circuits.
- The model takes intensity values $\vec{I}$ as input, and outputs smoothed intensity values. But this smoothness is broken at places where the intensity changes are too high. The model has continuous variables $\vec{J}$ representing the intensity, and binary-valued variables $\vec{l}$ for the line processes (or edges). The model is formulated as performing *maximum a posteriori* (MAP) estimation. The algorithm for estimating MAP is a neural network model that can be derived from the original Markov model (Geman & Geman, 1984) by mean field theory (Geiger & Yuille, 1991). Note that in this model, the variables do not have to represent intensity. Instead they can represent texture, depth, or any other property that is spatially smooth except at sharp discontinuities.

The line process model (II)

- For simplicity we present the weak membrane model in one dimension. The input is $\vec{I} = \{I(x) : x \in \mathcal{D}\}$; the estimated, or smoothed, image is $\vec{J} = \{J(x) : x \in \mathcal{D}\}$; and the line processes are denoted by $\vec{l} = \{l(x) : x \in \mathcal{D}\}$, where $l(x) \in \{0, 1\}$.
- The model is specified by a posterior probability distribution:

$$P(\vec{J}, \vec{l} | \vec{I}) = \frac{1}{Z} \exp\{-E[\vec{J}, \vec{l} : \vec{I}]/T\},$$

where

$$E[\vec{J}, \vec{l} : \vec{I}] = \sum_x (I(x) - J(x))^2 + A \sum_x (J(x+1) - J(x))^2 (1 - l(x)) + B \sum_x l(x).$$

The line process model (III)

The first term ensures that the estimated intensity $J(x)$ is close to the input intensity $I(x)$. The second encourages the estimated intensity $J(x)$ to be spatially smooth (e.g., $J(x) \approx J(x+1)$), unless a line process is activated by setting $l(x) = 1$. The third pays a penalty for activating a line process. The result encourages the estimated intensity to be piecewise smooth unless the input $I(x)$ changes significantly, in which case a line process is switched on and the smoothness is broken. The parameter T is the variance of the probability distribution and has a default value $T = 1$.

The line process model and neural circuits (I)

- This model can be implemented by a neural circuit (Koch et al., 1986). The connections between these neurons is shown in the previous figure. To implement this model Koch et al., (1986) proposed a neural net model that is equivalent to doing mean field theory on the weak membrane MRF (as discussed earlier) by replacing the binary-valued line process variables $l(x)$ by continuous variables $q(x) \in [0, 1]$ (corresponding roughly to the probability that the line process is switched on).
- This gives an algorithm that updates the regional variables $\vec{J}$ and the line variables $\vec{q}$ in a coupled manner. It is helpful, as before, to introduce a new variable $\vec{u}$ which relates by $q(x) = \frac{1}{1+\exp\{-u(x)/T\}}$ and $u(x) = T \log \frac{q(x)}{1-q(x)}$.

The line process model and neural circuits (II)

$$\frac{dJ(x)}{dt} = -2(J(x) - I(x))$$

$$= -2A\{(1 - q(x))(J(x) - J(x+1)) + (1 - q(x-1))(J(x) - J(x-1))\}, \quad (5)$$

$$\frac{dq(x)}{dt} = \frac{1}{T} q(x)(1 - q(x))\{A(J(x+1) - J(x))^2 - B - T \log \frac{q(x)}{1 - q(x)}\}, \quad (6)$$

$$\frac{du(x)}{dt} = -u(x) + A(J(x+1) - J(x))^2 - B. \quad (7)$$

The update rule for the estimated intensity $\vec{J}$ behaves like nonlinear diffusion, which smooths the intensity while keeping it similar to input $\vec{I}$. The diffusion is modulated by the strength of the edges $\vec{q}$. The update for the lines $\vec{q}$ is driven by the differences between the estimated intensity; if this is small, then the lines are not activated.

The line process model and neural circuits (III)

This algorithm has a Lyapunov function $L(\vec{J}, \vec{q})$ (derived using mean field theory methods) and so will converge to a fixed point, with

$$L(\vec{J}, \vec{q}) = \sum_x (I(x) - J(x))^2 + A \sum_x (J(x+1) - J(x))^2 (1 - q(x)) + B \sum_x q(x)$$
$$+ T \sum_x \{q(x) \log q(x) + (1 - q(x)) \log(1 - q(x))\}. \quad (8)$$

Relations to electrophysiology (I)

- There is some evidence that a generalization of this models roughly matches the electrophysiological findings for those types of stimuli. The generalization is performed by replacing the intensity variables $I(x), J(x)$ by a filterbank of Gabor filters so that the weak membrane model enforces edges at places where the texture properties change (Lee et al., 1992). The experiments, and their relation to the weak membrane models are reviewed in (Lee & Yuille, 2006). The initial responses of the neurons, for the first 80 msec, are consistent with the linear filter models described earlier. But after 80 msec, the activity of the neurons changes and appears to take spatial context into account.
- While the weak membrane model is broadly consistent with the perceptual phenomena of segmentation and "filling in," the types of filling in, their dynamics, and the neural representations of contours and surface are complicated (von der Heydt, 2002; Komatsu, 2006). Exactly how contour and surface information is represented and processed in cortex is an active topic of research (Grossberg & Hong, 2006; Roe et al., 2012).

The EM Perspective

- An alternative way to derive the same update equations for the weak membrane model is to formulate this problem as Expectation-Maximization where the line process variable l is treated as a hidden variable and the task is to estimate J.
- We have $P(J, l|I)$ and want to estimate $J^* = \arg\min\{-\log P(J|I)\}$. We introduce a dummy variable $Q(l)$ and require it to be factorizable, i.e., $Q(l) = \prod_x q_x(l_x)$. Then we use the standard EM free energy and obtain $\sum_x \{q_x \log q_x + (1 - q_x) \log(1 - q_x)\} + \sum_x (J(x) - I(x))^2 + A \sum_x (J(x + 1) - J(x))^2 (1 - q_x) + B \sum_x q_x$. This is the same as the formulation earlier (with slightly different notation).
- This shows connections between mean field theory and Expectation=Maximization. It also illustrates the benefits of formulation EM in terms of minimizing a free energy. Unlike earlier formulations of EM, we can put restrictions on the form of $Q(l)$. In this case, there is no need to because the solution naturally takes the form of a factorizable distribution. But if the weak membrane model is extended to two-dimensions and coupling terms are introduced between horizontal and vertical line processes, then the solution $Q()$ will not be factorizable – but we can make a practical approximation by requiring it to be factorizeable.

Annealing and Continuation Methods (1)

- We now discuss a continuation method called *deterministic annealing* which can improve mean field theory. This specifies a family of free energy functions $F(., T)$ parameterized by a *temperature* T. The fixed value $T = 1$ corresponds to the problem that you want to solve. The key idea is that the energy functions get more convex as T increases. So the algorithm finds a minimum of $F(., T)$ for large T and gives it initial conditions for minimizing $F(., T)$ at smaller T. There is no guarantee that this algorithm converges to the global minimum. But empirically it yields good results.
- Deterministic annealing was inspired by *simulated annealing*. This was based on the following observation. Suppose we are trying to estimate the most probable states of the probability distribution $\mathbf{x}^* = \arg\max_{\mathbf{x}} P(\mathbf{x})$. We introduce a temperature parameter T and a family of probability distributions related to $P(\mathbf{x})$.
- More precisely, we define a one-parameters family of distributions $\propto \{P(\mathbf{x})\}^{1/T}$ where T is a temperature parameter (the constant of proportionality is the normalization constant). This is equivalent to specifying Gibbs distributions $P(\mathbf{x}; T) = \frac{1}{Z(T)} \exp\{-E(\mathbf{x})/T\}$, where the default distribution $P(\mathbf{x})$ occurs at $T = 1$.
- The key observation is that as $T \mapsto 0$, the distribution gets strongly peaked about the state $\mathbf{x}^* = \arg\min_{\mathbf{x}} E(\mathbf{x})$ with lowest energy (or states if there are two or more global minima). Conversely, at $T \mapsto \infty$ all states will become equally likely and $P(\mathbf{x}; T)$ will tend to the uniform distribution.

Annealing and Continuation Methods (2)

- Simulated Annealing (e.g., Geman and Geman 1984) uses Markov Chain Monte Carlo (MCMC) to obtain random samples $\mathbf{x}$ from $P(\mathbf{x} : T)$. The random samples are most likely to be at places where $P(\mathbf{x} : T)$ is large. Hence as $T \mapsto 0$ the random samples must lie at the global minimum $\mathbf{x}^* = \arg\min E(\mathbf{x})$.
- But for small T it is very hard to sample from $P(\mathbf{x} : T)$ and MCMC will take a very long time to converge. So the suggestion is to perform MCMC sampling at large T, where sampling is likely to converge faster, and lower the temperature and continue sampling. This is called simulated annealing ("annealing" is a physical process which involves lowering the temperature).
- Simulated Annealing is conceptually attractive (and in the early 1980's) there were great hopes for it as an algorithm for minimizing energy functions (there were even proofs that it would converge to the global minimum, but this might take longer than exhaustive search!). In practice, simulated annealing has only been useful for a small range of problems. One problem is that MCMC algorithms (see handouts) are often slow and need hand=designing to be effective).

Annealing and Continuation Methods (3)

- Deterministic Annealing was inspired by Simulated Annealing. The idea is that we apply mean field theory to approximate the distributions $P(\mathbf{x} : T)$ at different temperatures. This gives a family of free energy functions $F(., T)$ parameterized by T, (i.e. $F(., T)$ is the Kullback-Liebler divergence between $P(\mathbf{x} : T)$ and $Q(\mathbf{x})$. Deterministic annealing proposes to minimize $F(., T)$ at large T (where the free energy is more convex) and use this as initial conditions for minimizing it at lower temperature.
- More precisely, we compute the free energy as a function of T (multiplying by T for reasons will become clear in a few lines) and dropping the log partition function). This gives:

$$T\mathcal{F}_{\mathrm{MFT}}(\underline{}) = \sum_{ij\in\mathcal{E}}\sum_{x_i,x_j} b_i(x_i)b_j(x_j)\psi_{ij}(x_i,x_j)$$

$$+\sum_{i\in\mathcal{V}}\sum_{x_i} b_i(x_i)\phi_i(x_i,\mathbf{z}) + T\sum_{i\in\mathcal{V}}\sum_{x_i} b_i(x_i)\log b_i(x_i). \qquad (9)$$

- The second term of the free energy is the *entropy term*. It is the only term that depends on the temperature T (linearly) and it is also convex in $b_i(x_i)$. Hence at large T the entropy term dominates the free energy and hence the free energy becomes convex. As $T \mapsto \infty$, the solution consists of setting $b_i(x_i)$ to be constant, i.e. the maximum entropy solution.