Graphical models, belief propagation, and Markov random fields

Bill Freeman, MIT
Fredo Durand, MIT

6.882 March 21, 2005

Color selection problem

• (see Photoshop demonstration)

Stereo problem

Squared difference, \((L[x] - R[x+d])^2\), for some \(x\).

Showing local disparity evidence vectors for a set of neighboring positions, \(x\).

Super-resolution image synthesis

How select which selection of high resolution patches best fits together? Ignoring which patch fits well with which gives this result for the high frequency components of an image:

Things we want to be able to articulate in a spatial prior

• Favor neighboring pixels having the same state (state, meaning: estimated depth, or group segment membership)
• Favor neighboring nodes have compatible states (a patch at node 1 should fit well with selected patch at node 2).
• But encourage state changes to occur at certain places (like regions of high image gradient).

Graphical models: tinker toys to build complex probability distributions

• Circles represent random variables.
• Lines represent statistical dependencies.
• There is a corresponding equation that gives \(P(x_1, x_2, x_3, y, z)\), but often it’s easier to understand things from the picture.
• These tinker toys for probabilities let you build up, from simple, easy-to-understand pieces, complicated probability distributions involving many variables.
Steps in building and using graphical models

- First, define the function you want to optimize. Note the two common ways of framing the problem:
  - In terms of probabilities. Multiply together component terms, which typically involve exponentials.
  - In terms of energies. The log of the probabilities. Typically add together the exponentiated terms from above.
- The second step: optimize that function. For probabilities, take the mean or the max (or use some other “loss function”). For energies, take the min.
- 3rd step: in many cases, you want to learn the function from the 1st step.

Define model parameters

\[
\begin{pmatrix}
1 & \alpha & \alpha \\
\alpha & 1 & \alpha \\
\alpha & \alpha & 1
\end{pmatrix}
\]

A more general compatibility matrix (values shown as grey scale)

Derivation of belief propagation

\[ x_{\text{MAP}} = \text{mean sum} P(x_1, x_2, x_3, y_1, y_2, y_3) \]

The posterior factorizes

Propagation rules

\[ x_{\text{MAP}} = \text{mean sum} P(x_1, x_2, x_3, y_1, y_2, y_3) \]

\[ x_{\text{MAP}} = \text{mean sum} \Phi(x_1, y_1) \]

\[ x_{\text{MAP}} = \text{mean sum} x_2 \Phi(x_2, y_2) \Psi(x_1, x_2) \]

\[ x_{\text{MAP}} = \text{mean sum} x_3 \Phi(x_3, y_3) \Psi(x_2, x_3) \]
Belief propagation: the nosey neighbor rule

“Given everything that I know, here’s what I think you should think”

(Given the probabilities of my being in different states, and how my states relate to your states, here’s what I think the probabilities of your states should be)
Simple BP example

\[ M^{(0)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad M^{(1)} = \frac{1}{2} \]

\[ \Psi(x_1, x_2) = \begin{bmatrix} 9 & 1 \\ 1 & 9 \end{bmatrix} \]

To find the marginal probability for each variable, you can
(a) Marginalize out the other variables of:
\[ P(x_1, x_2, x_3) = \Psi(x_1, x_2) \Psi(x_2, x_3) M^{(0)}(x_0) M^{(1)}(x_1) \]
(b) Or you can run belief propagation, (BP). BP redistributes the various partial sums, leading to a very efficient calculation.

Belief, and message updates

\[ b_j(x_j) = \prod_{k \in N(j)} M^{(k)}(x_k) \]

\[ M^{(t)}(x_i) = \sum_{x_j} \Psi_j(x_i, x_j) \prod_{k \in V(j) \setminus i} M^{(k)}(x_k) \]

Optimal solution in a chain or tree:
Belief Propagation

- “Do the right thing” Bayesian algorithm.
- For Gaussian random variables over time: Kalman filter.
- For hidden Markov models: forward/backward algorithm (and MAP variant is Viterbi).

Making probability distributions modular, and therefore tractable:
Probabilistic graphical models

Vision is a problem involving the interactions of many variables: things can seem hopelessly complex. Everything is made tractable, or at least, simpler, if we modularize the problem. That’s what probabilistic graphical models do, and let’s examine that.

Readings: Jordan and Weiss intro article—fantastic!
Kevin Murphy web page—comprehensive and with pointers to many advanced topics
A toy example

Suppose we have a system of 5 interacting variables, perhaps some are observed and some are not. There’s some probabilistic relationship between the 5 variables, described by their joint probability, \( P(x_1, x_2, x_3, x_4, x_5) \).

If we want to find out what the likely state of variable \( x_1 \) is (say, the position of the hand of someone we are observing), what can we do?

Two reasonable choices are: (a) find the value of \( x_1 \) (and of all the other variables) that gives the maximum of \( P(x_1, x_2, x_3, x_4, x_5) \); that’s the MAP solution.

Or (b) marginalize over all the other variables and then take the mean or the maximum of the other variables. Marginalizing, then taking the mean, is equivalent to finding the MMSE solution. Marginalizing, then taking the max, is called the max marginal solution and sometimes a useful thing to do.

To find the marginal probability at \( x_1 \), we have to take this sum:

\[
\sum_{x_2, x_3, x_4, x_5} P(x_1, x_2, x_3, x_4, x_5)
\]

If the system really is high dimensional, that will quickly become intractable. But if there is some modularity in the \( P(x_1, x_2, x_3, x_4, x_5) \) that things become tractable again.

Suppose the variables form a Markov chain: \( x_1 \) causes \( x_2 \) which causes \( x_3 \), etc. We might draw out this relationship as follows:

Belief propagation

Performing the marginalization by doing the partial sums is called “belief propagation”.

\[
\sum_{x_2, x_3, x_4, x_5} P(x_1, x_2, x_3, x_4, x_5) = P(x_1) \sum_{x_2} P(x_2|x_1) \sum_{x_3} P(x_3|x_2) \sum_{x_4} P(x_4|x_3) \sum_{x_5} P(x_5|x_4)
\]

In this example, it has saved us a lot of computation. Suppose each variable has 10 discrete states. Then, not knowing the special structure of \( P \), we would have to perform \( 10^6 \) to marginalize over the four variables.

But doing the partial sums on the right hand side, we only need 40 additions (10*4) to perform the same marginalization!

Another modular probabilistic structure, more common in vision problems, is an undirected graph:

No factorization with loops!

The joint probability for this graph is given by:

\[
P(x_1, x_2, x_3, x_4, x_5) = \Phi(x_1, x_2) \Phi(x_2, x_3) \Phi(x_3, x_4) \Phi(x_4, x_5)
\]

Where \( \Phi(x_i, x_j) \) is called a “compatibility function”. We can define compatibility functions just like we did in the previous slides; for that example, we can use either form.
Justification for running belief propagation in networks with loops

• Experimental results:
  – Error-correcting codes  Kschischang and Frey, 1998; McEliece et al., 1998
  – Vision applications  Freeman and Pasztor, 1999; Frey, 2000
• Theoretical results:
  – For Gaussian processes, means are correct  Weiss and Freeman, 1999
  – Large neighborhood local maximum for MAP  Weiss and Freeman, 2000
  – Equivalent to Bethe approx. in statistical physics  Yedidia, Freeman, and Weiss, 2000
  – Tree-weighted reparameterization  Wainwright, Willsky, Jaakkola, 2001

Belief propagation equations
Belief propagation equations come from the marginalization constraints.

\[ M_i(x_i) = \sum_j \Psi_i(x_i, x_j) \prod_{k \in N(i) \setminus j} M_k(x_k) \]

Region marginal probabilities

\[ b_i(x_i) = k \Phi_i(x_i) \prod_{j \in N(i)} M_j(x_j) \]

Results from Bethe free energy analysis

• Fixed point of belief propagation equations iff. Bethe approximation stationary point.
• Belief propagation always has a fixed point.
• Connection with variational methods for inference: both minimize approximations to Free Energy,
  – variational: usually use primal variables.
• Kikuchi approximations lead to more accurate belief propagation algorithms.
• Other Bethe free energy minimization algorithms— Yuille, Welling, etc.

Kikuchi message-update rules
Groups of nodes send messages to other groups of nodes.

Typical choice for Kikuchi cluster.

Update for messages  Update for messages

Generalized belief propagation

Marginal probabilities for nodes in one row of a 10x10 spin glass

BP: belief propagation  GBP: generalized belief propagation
ML: maximum likelihood

Node number
References on BP and GBP

- J. Pearl, 1985
  - classic
- Y. Weiss, NIPS 1998
  - Inspires application of BP to vision
- W. Freeman et al learning low-level vision, IJCV 1999
  - Applications in super-resolution, motion, shading/paint discrimination
- H. Shum et al, ECCV 2002
  - Application to stereo
- M. Wainwright, T. Jaakkola, A. Willsky
  - Reparameterization version
- J. Yedidia, AAAI 2000
  - The clearest place to read about BP and GBP.

Probability models for entire images: Markov Random Fields

- Allows rich probabilistic models for images.
- But built in a local, modular way. Learn local relationships, get global effects out.

MRF nodes as pixels

- Winkler, 1995, p. 32

MRF nodes as patches

- Winkler, 1995, p. 32

Network joint probability

\[ P(x, y) = \frac{1}{Z} \prod_{i,j} \Phi(x_i, y_j) \prod_i \Psi(x_i, x_j) \]

In order to use MRFs:

- Given observations \( y \), and the parameters of the MRF, how infer the hidden variables, \( x \)?
- How learn the parameters of the MRF?
Outline of MRF section

- Inference in MRF’s.
  - Iterated conditional modes (ICM)
  - Gibbs sampling, simulated annealing
  - Variational methods
  - Belief propagation
  - Graph cuts
- Vision applications of inference in MRF’s.
- Learning MRF parameters.
  - Iterative proportional fitting (IPF)

Iterated conditional modes

- For each node:
  - Condition on all the neighbors
  - Find the mode
  - Repeat.


Gibbs Sampling and Simulated Annealing

- Gibbs sampling:
  - A way to generate random samples from a (potentially very complicated) probability distribution.
- Simulated annealing:
  - A schedule for modifying the probability distribution so that, at “zero temperature”, you draw samples only from the MAP solution.
  \[ P(x) = \frac{1}{Z} \exp\left(-\frac{E(x)}{kT}\right) \]

Reference: Geman and Geman, IEEE PAMI 1984

Sampling from a 1-d function

1. Discretize the density function
2. Compute distribution function from density function
3. Sampling
   - draw \( \alpha \sim U(0,1) \);
   - for \( k = 1 \) to \( n \)
     - if \( F(k) \geq \alpha \)
       - break;
     - \( x = x_k + k \tau^+ \)
Gibbs sampling and simulated annealing

Simulated annealing as you gradually lower the “temperature” of the probability distribution ultimately giving zero probability to all but the MAP estimate.

What’s good about it: finds global MAP solution.

What’s bad about it: takes forever. Gibbs sampling is in the inner loop…

Outline of MRF section

• Inference in MRF’s.
  – Iterated conditional modes (ICM)
  – Gibbs sampling, simulated annealing
  – Variational methods
    – Belief propagation
    – Graph cuts
  • Vision applications of inference in MRF’s.
  • Learning MRF parameters.
    – Iterative proportional fitting (IPF)

Variational methods

• Example: mean field
  – For each node
    • Calculate the expected value of the node, conditioned on the mean values of the neighbors.
Outline of MRF section

- Inference in MRF’s.
  - Iterated conditional modes (ICM)
  - Gibbs sampling, simulated annealing
  - Variational methods
  - Belief propagation
  - Graph cuts

- Vision applications of inference in MRF’s.

- Learning MRF parameters.
  - Iterative proportional fitting (IPF)

Graph cuts

- Algorithm: uses node label swaps or expansions as moves in the algorithm to reduce the energy.
  - Swaps many labels at once, not just one at a time, as with ICM.
  - Find which pixel labels to swap using min cut/max flow algorithms from network theory.
  - Can offer bounds on optimality.

Comparison of graph cuts and belief propagation

Comparison of Graph Cuts with Belief Propagation for Stereo, using Identical MRF Parameters, ICCV 2003.

Marshall F. Tappen William T. Freeman

Graph cuts versus belief propagation

- Graph cuts consistently gave slightly lower energy solutions for that stereo-problem MRF, although BP ran faster, although there is now a faster graph cuts implementation than what we used...
- However, here’s why I still use Belief Propagation:
  - Works for any compatibility functions, not a restricted set like graph cuts.
  - I find it very intuitive.

Ground truth, graph cuts, and belief propagation disparity solution energies

Table:

<table>
<thead>
<tr>
<th>Image</th>
<th>Graph Cuts</th>
<th>Belief Prop</th>
<th>% Error from Occluded Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Takeda</td>
<td>545</td>
<td>532</td>
<td>3%</td>
</tr>
<tr>
<td>Surrey</td>
<td>572</td>
<td>578</td>
<td>1%</td>
</tr>
<tr>
<td>Stanford</td>
<td>575</td>
<td>572</td>
<td>0%</td>
</tr>
</tbody>
</table>

Figure 2. Field Energies for the MRF labelled using ground truth data, compared to the energies for the labels labelled using Graph Cuts and Belief Propagation. Notice that the solutions returned by the algorithms consistently have much lower energy than the shallows produced from the ground truth, showing a mismatch between the MRF formulations and the ground truth. The first column contains the percentage of each groundtruth solution’s energy that comes from matching costs of occluded points.

Figure 3. Results produced by the three algorithms on the Takeda image. The parameters used to generate this field were: a = 0.3, Z = 1, F = 1. Graph Cuts produces a much smoother solution. Belief Propagation does maintain some structures that are lost in the Graph Cuts solution, such as the curvy line in the foreground.

MAP versus MMSE

Figure 4. Comparison of MAP and MMSE estimates on a different MRF formulation. The MAP estimate chooses the most likely discrete disparity level for each pixel, resulting in a depth-map with strong step-stopping effects. Using the MMSE estimate assigns smoothed disparities, resulting in a smooth depth field.
Show program comparing some methods on a simple MRF

testMRF.m

Outline of MRF section

- Inference in MRF’s.
  - Gibbs sampling, simulated annealing
  - Iterated conditional modes (ICM)
  - Variational methods
  - Belief propagation
  - Graph cuts
- Applications of inference in MRF’s.
- Learning MRF parameters.
  - Iterative proportional fitting (IPF)

Applications of MRF’s

- Stereo
- Motion estimation
- Labelling shading and reflectance
- Many others…

Applications of MRF’s

- Stereo
- Motion estimation
- Labelling shading and reflectance
- Many others…

Motion application

What behavior should we see in a motion algorithm?

- Aperture problem
- Resolution through propagation of information
- Figure/ground discrimination
The aperture problem

Program demo

Motion analysis: related work

- **Markov network**
  - Luetgen, Karl, Willsky and collaborators.

- **Neural network or learning-based**
  - Nowlan & T. J. Senjowski; Sereno.

- **Optical flow analysis**
  - Weiss & Adelson; Darrell & Pentland; Ju, Black & Jepson; Simoncelli; Grzywacz & Yuille; Hildreth; Horn & Schunk; etc.

Motion estimation results

Inference:

Initial guesses only show motion at edges.
Motion estimation results
(matrix of scene probability distributions displayed)

Iterations 4 and 5
Final result compares well with vector quantized true (uniform) velocities.

Vision applications of MRF’s
• Stereo
• Motion estimation
• Labelling shading and reflectance
• Many others…

Forming an Image
Illuminate the surface to get:
Surface (Height Map)  Shading Image
The shading image is the interaction of the shape of the surface and the illumination

Painting the Surface
Scene  Image
Add a reflectance pattern to the surface. Points inside the squares should reflect less light

Goal
Image  Shading Image  Reflectance Image

Basic Steps
1. Compute the $x$ and $y$ image derivatives
2. Classify each derivative as being caused by
   either shading or a reflectance change
3. Set derivatives with the wrong label to zero.
4. Recover the intrinsic images by finding the least-squares solution of the derivatives.
Learning the Classifiers

- Combine multiple classifiers into a strong classifier using AdaBoost (Freund and Schapire)
- Choose weak classifiers greedily similar to (Tieu and Viola 2000)
- Train on synthetic images
- Assume the light direction is from the right

Shading Training Set
Reflectance Change Training Set

Using Both Color and Gray-Scale Information

Results without considering gray-scale

Some Areas of the Image Are Locally Ambiguous

Is the change here better explained as

Input
Shading
Reflectance

Propagating Information

- Can disambiguate areas by propagating information from reliable areas of the image into ambiguous areas of the image

Setting Compatibilities

- Set compatibilities according to image contours
  - All derivatives along a contour should have the same label
- Derivatives along an image contour strongly influence each other

\[
\psi(x, y) = \begin{bmatrix}
1 - \beta & \beta \\
\beta & 1 - \beta
\end{bmatrix}
\]

Propagating Information

- Consider relationship between neighboring derivatives
- Use Generalized Belief Propagation to infer labels
**Improvements Using Propagation**

- Input Image
- Reflectance Image Without Propagation
- Reflectance Image With Propagation

**Combining local evidence from shape and color, and GBP for propagation**

- (a) Original Image
- (b) Shading Image
- (c) Reflectance Image

**Outline of MRF section**

- Inference in MRF’s.
  - Gibbs sampling, simulated annealing
  - Iterated conditional modes (ICM)
  - Variational methods
  - Belief propagation
  - Graph cuts
- Vision applications of inference in MRF’s.

- Learning MRF parameters.
  - Iterative proportional fitting (IPF)
Learning MRF parameters, labeled data

Iterative proportional fitting lets you make a maximum likelihood estimate a joint distribution from observations of various marginal distributions.

Initial guess at joint probability

IPF update equation

\[ P(x_1, x_2, \ldots, x_d)^{(t+1)} = P(x_1, x_2, \ldots, x_d)^{(t)} \frac{P(x_1)^{\text{observed}}}{P(x_1)^{(t)}} \]

Scale the previous iteration’s estimate for the joint probability by the ratio of the true to the predicted marginals.

Gives gradient ascent in the likelihood of the joint probability, given the observations of the marginals.

See: Michael Jordan’s book on graphical models
IPF results for this example: comparison of joint probabilities

Application to MRF parameter estimation

- Can show that for the ML estimate of the clique potentials, $\phi_c(x_c)$, the empirical marginals equal the model marginals,

  $\hat{p}(x_c) = p(x_c)$

- This leads to the IPF update rule for $\phi_c(x_c)$

  $\phi_c^{(t+1)}(x_c) = \phi_c^{(t)}(x_c) \frac{\hat{p}(x_c)}{\hat{p}(t)(x_c)}$

- Performs coordinate ascent in the likelihood of the MRF parameters, given the observed data.

Reference: unpublished notes by Michael Jordan

More general graphical models than MRF grids

- In this course, we’ve studied Markov chains, and Markov random fields, but, of course, many other structures of probabilistic models are possible and useful in computer vision.
- For a nice on-line tutorial about Bayes nets, see Kevin Murphy’s tutorial in his web page.

GrabCut

“GrabCut” — Interactive Foreground Extraction using Iterated Graph Cuts