CS664 Lecture #26: Bayesian Vision

Some material taken from:

- **Phil Torr & Pushmeet Kohli, Oxford Brookes University**
- **Andrew Blake, Microsoft Research and Oxford University**

Announcements

- PS3 due Friday 12/2
  - Minor glitch with input images is now fixed
  - Output should be pair of scaled images (u,v)
  - You can use MatLab's quiver function for generating "needle diagrams" for debugging
- Final project is due Thursday 12/15

Recall: maximum likelihood

- Suppose we have an observation O and a set of possibly hypotheses $\{H_1, H_2, \ldots\}$
  - Likelihood of H_i is $P(O|H_i)$
  - We maximize the likelihood
- Example: image denoising
  - Assume IID Gaussian noise
  - $O = \{o_1, o_2, \ldots, o_n\}$, $H = \{h_1, h_2, \ldots, h_n\}$
  - By independence, $\Pr(O|H) = \prod_i \Pr(o_i|h_i)$
  - Here we have

$$\Pr(o_i|h_i) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\frac{(o_i - h_i)^2}{\sigma}}$$

From ML to MAP

- We want a principled way to incorporate additional knowledge about a hypothesis
  - Recall that ML likes “fairy tales”
- Bayes rule gives us a way to relate prior probability of a hypothesis to likelihood
  - $\Pr(O|H)\ \Pr(H) = \Pr(O,H) = \Pr(H|O)\ \Pr(O)$
  - If we multiply the likelihood by the prior, we get the joint probability of a hypothesis and an observation

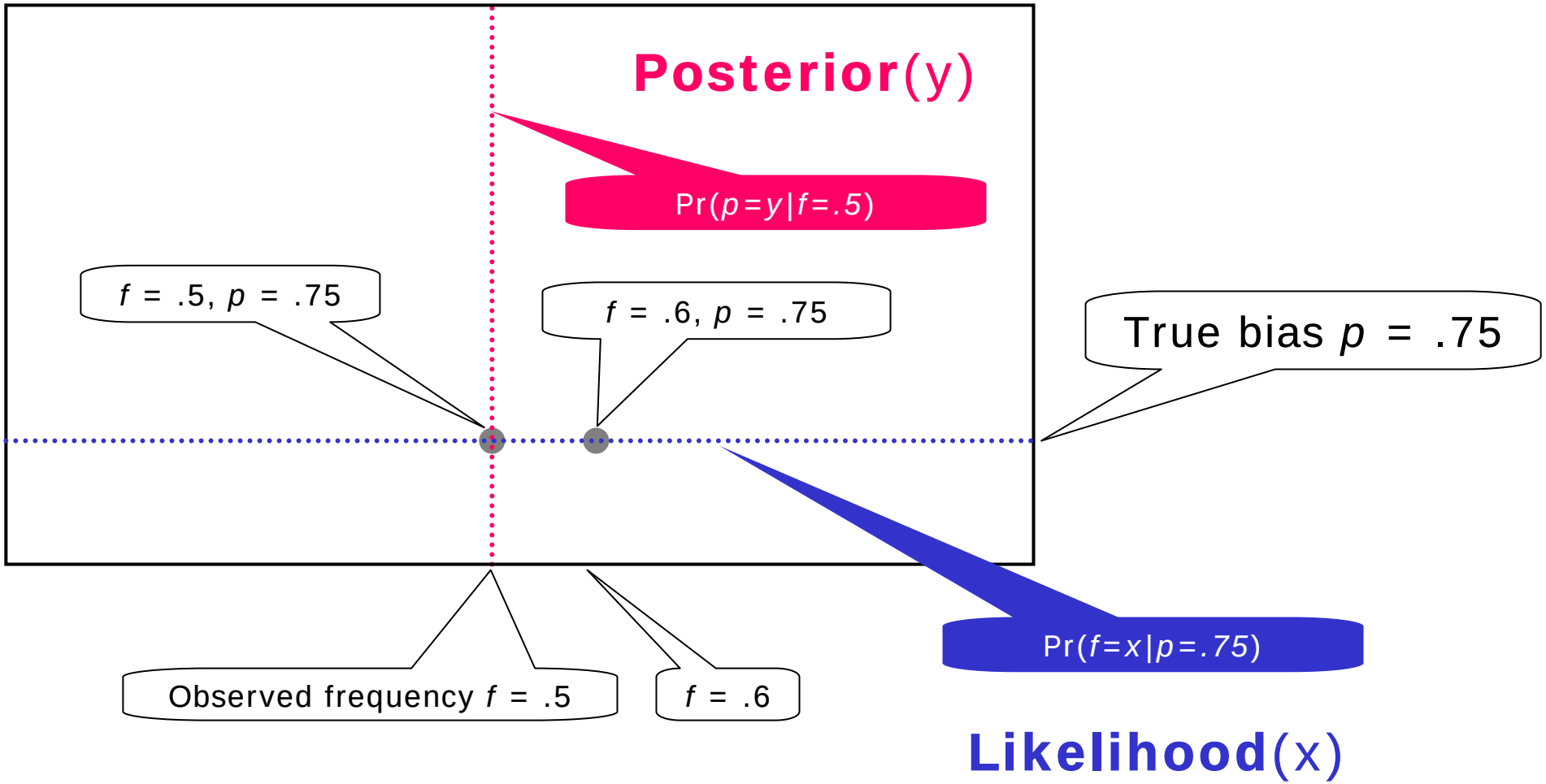
Joint probability

- Consider a joint density where one variable represents a hypothesis and the other represents an observation
  - Hypothesis variable *p*: true bias of the coin
  - Observation variable *f*: % heads observed
- Any possible experiment (sample) that could occur is a pair (*f*,*p*)

	***f*=0**	***f*=.5**	***f*=1**
***p*=.5**	0.25	0.5	0.25
***p*=.75**	0.0625	0.375	0.5625

	***f*=0**	***f*=.5**	***f*=1**	
***p*=.5**	0.175	0.35	0.175	0.7
***p*=.75**	0.0188	0.113	0.1688	0.3

Joint probability space



Bayesian (MAP) estimation

- Maximum likelihood (ML): consider conditional probability one "row" at a time
- Maximum a posteriori (MAP): consider the conditional probability in this "column"
  - Posterior probability is Pr(H|O)
    - Basically the right thing, by definition
- *Maximizing* it is not necessarily ideal
  - Depending on its distribution
  - Applications generally need a point estimate
    - MAP has a Potts-style loss function

Priors and marginals

- The prior is the "row sum"
  - The probability that p=.75 is the sum, over all experiments (f=x,p=.75), of the probability of this experiment

    $\Pr(p=y) = \sum_x \Pr(f=x,p=y)$
- This operation is sometimes referred to as marginalization over the variable that we sum over
  - In this case, marginalization over x
- Column sum is called "evidence", FWIW

Priors for images

- Need a prior with spatial coherence
  - Local dependence, as in Markov Chains
- Generalization: Markov Random Fields
  - Field: set of Random Variables X_i
    - Takes a value h_i in a finite set (discrete)
  - Neighborhood system N(j)
  - Configuration $H = (h_1, h_2, \ldots, h_n)$
  - A probability distribution Pr(H) is an MRF if
    - $Pr(H) > 0, \forall H$
    - $Pr(X_i = h_i | \{X_j = h_j\}) = Pr(X_i = h_i | \{X_{N(j)} = h_{N(j)}\})$

Conditionals are problematic

- Unfortunately, writing down the conditional distributions doesn't work
  - Most of the time, they contradict each other
- Example: suppose that we predict 2X2 patches based on other intensities
  - Suppose that A=C=D=1 implies B=1
  - Also, suppose A=B=D=0 implies C=0
  - Now consider:

A	B
C	D

1	0	0
1	?	0
1	1	1

Gibbs Random Fields

- A Gibbs Random Field has X_j,N(j)
  - Consider **C**, a set of cliques in N
  - There are some functions V_C such that
  - $Pr(H) = \exp(-\sum_{c \in \mathbf{C}} V_c(H))$
- It's easy to write down such clique potentials V_C
- What is the relationship between GRF's and MRF's?

Hammersley-Clifford Theorem

- Theorem: GRF's are MRF's and vice-versa
  - Not hard to prove that a GRF is an MRF
    - The definition is essentially "local"
- Famous unpublished theorem
  - It was thought that the positivity constraint could be relaxed
  - There is a counterexample without it
- Good proof in Joseph Chang's notes

MAP-MRF framework

- Obviously, points and edges are cliques
  - Higher-order cliques can also exist
  - Let's write an edge potential as $V_{ij}(h_i,h_j)$
    - Assume point potentials are uniform
- Prior probability of a configuration is
  - $\Pr(H) \propto \exp(-\sum_{ij} V_{ij}(h_i,h_j))$
- Posterior probability with iid noise is
  - $\Pr(H|O) \propto \prod_i \Pr(o_i|h_i) \cdot \exp(-\sum_{ij} V_{ij}(h_i,h_j))$

MAP-MRF Energy Minimization

- To maximize the posterior we maximize its log, or minimize its negative log

$$-\log \Pr(H|O) = \sum_i \log(\Pr(o_i|h_i)) + \sum_{ij} V_{ij}(h_i, h_j)$$

  - If we write $\log(\Pr(o_i|h_i))$ as $D_i(h_i)$ we get a familiar looking equation!

Bayesian framework

- The idea of an MRF is often used interchangeably with this energy function
  - As is “energy minimization”
- MAP has some obvious weaknesses
  - Lack of an obvious confidence measure
    - Related to “hard” assignments
  - Peculiar nature of implicit loss function
  - Computational difficulty
    - Well, maybe not any more…

Computing min-marginals

- There is a natural way to assign a probability to an individual hypothesis h_i
  - Often h_i specifies a particular label for pixel i
- Consider the highest probability configuration that includes h_i
  - This is the max-marginal probability of h_i
- Equivalent energy minimization problem:
  - Find the lowest energy solution if we constrain this pixel to have this label
  - "Min-marginal energy"

Belief propagation

- Optimal algorithm for trees
  - Useful on grids, but not well understood
- Basic operation is message passing
  - One pixel sends another a vector
    - One entry per label
  - Combine incoming messages to make new ones
    - Several variants on how this is done
  - Most message passing is done in parallel
- BP is the only such “local update” algorithm that actually seems to work
  - Though no one really knows exactly why

Dynamic graph cuts

- Min-marginal energy can also be computed via graph cuts
  - Fix a pixel's label, then minimize the energy
- Kohli & Torr's dynamic graph cuts algorithm is required
  - Otherwise problem is totally intractable
  - They show how to "re-use" most of the previous max flow computation,
    - Assuming the graph doesn't change very much
    - Which is true in this application!

Bayesian tracking

- The two most famous Bayesian problems in vision are MRF's and tracking
- What can we track?
  - The contour of a hand, or something similar
  - A single point
  - A moving object
    - Such as a missile from radar data
- DEMOS (c/o Andrew Blake's group)

Filtering data over time

- Consider a noisy temporal sequence X[t]
- Some useful operations:
  - **Prediction**: given X[1],…,X[t-1] find X[t]
  - **Filtering**: given X[1],…,X[t] find X[t]
  - **Smoothing**: given X[1],…,X[t+1] find X[t]
- Natural Bayesian view of Filtering
  - Compute a prior based on distribution for X[t-1]
    - Apply an update model to this distribution
  - New observation gives a likelihood
  - Combine to form a distribution for X[t]

Kalman filtering

- Assumptions
  - Hidden states are updated linearly over time
  - Observe a linear function of the hidden state
  - Both the update and the observation are corrupted by Gaussian noise
- Kalman filter estimates the hidden state
  - Optimal, in the usual least squares sense
- Example: tracking a missile
- http://www.cs.ubc.ca/~murphyk/Software/Kalman/kalman.html