Advanced Introduction to Machine Learning

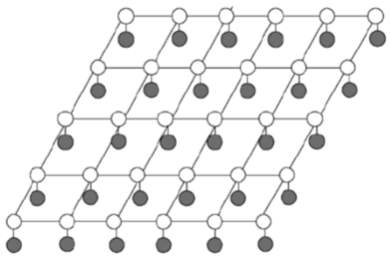
10715, Fall 2014

Structured Models (2):

Hidden Markov Models versus Conditional Random Fields

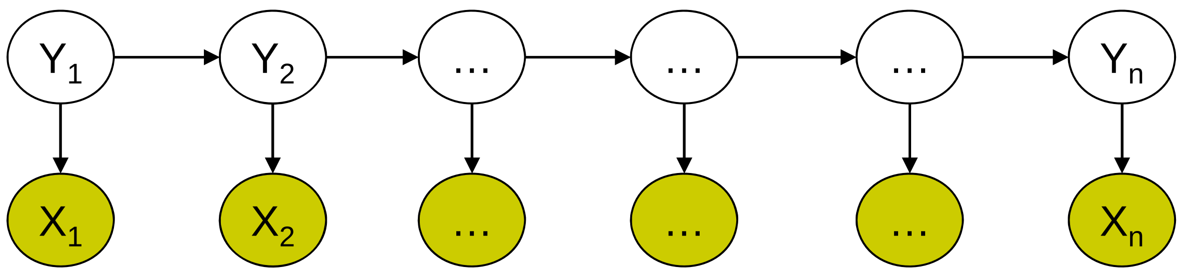
Eric Xing

Lecture 12, October 15, 2014



Reading:

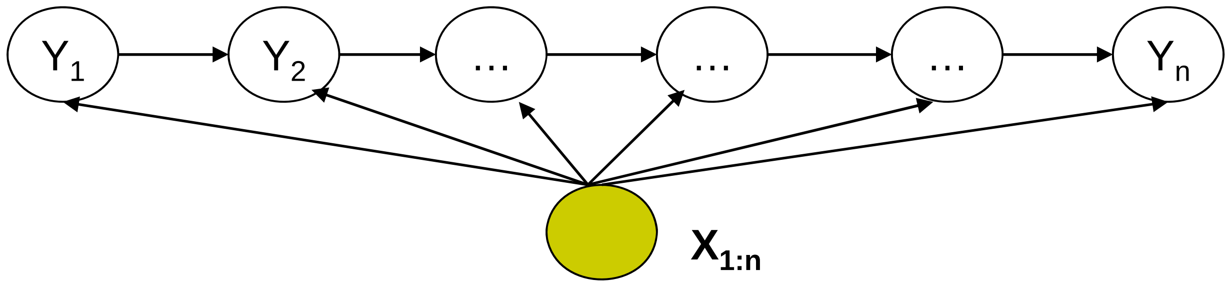
Shortcomings of Hidden Markov Model



- HMM models capture dependences between each state and only its corresponding observation
  - NLP example: In a sentence segmentation task, each segmental state may depend not just on a single word (and the adjacent segmental stages), but also on the (non-local) features of the whole line such as line length, indentation, amount of white space, etc.
- Mismatch between learning objective function and prediction objective function
  - HMM learns a joint distribution of states and observations P(**Y**, **X**), but in a prediction task, we need the conditional probability P(**Y**|**X**)

Departing from HMM

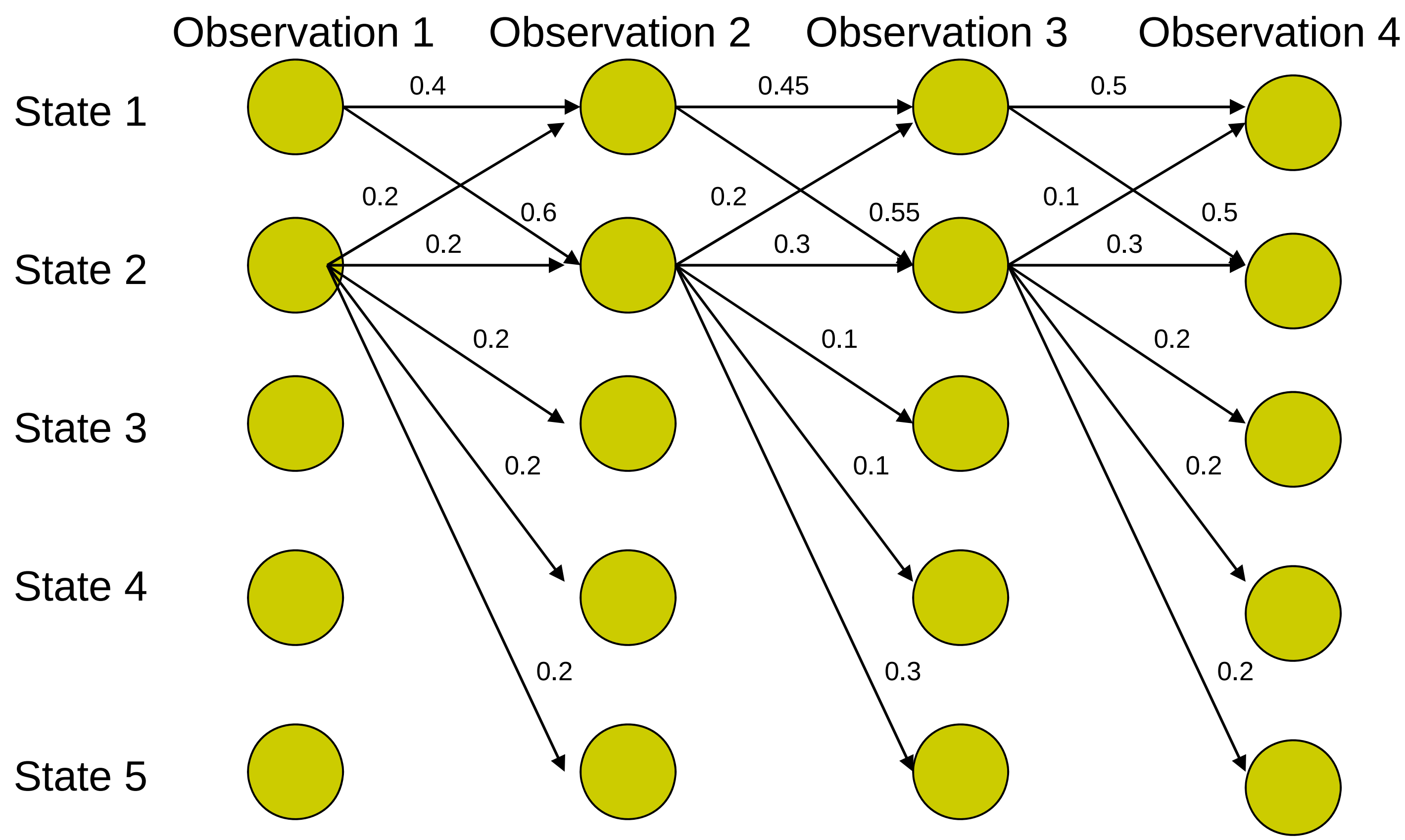
Solution:
Maximum Entropy Markov Model (MEMM)



$$P(\mathbf{y}_{1:n}|\mathbf{x}_{1:n}) = \prod_{i=1}^{n} P(y_i|y_{i-1},\mathbf{x}_{1:n}) = \prod_{i=1}^{n} \frac{\exp(\mathbf{w}^T\mathbf{f}(y_i,y_{i-1},\mathbf{x}_{1:n}))}{Z(y_{i-1},\mathbf{x}_{1:n})}$$

- Models dependence between each state and the full observation sequence explicitly
  - More expressive than HMMs
- Discriminative model
  - Completely ignores modeling P(**X**): saves modeling effort
  - Learning objective function consistent with predictive function: P(**Y**|**X**)

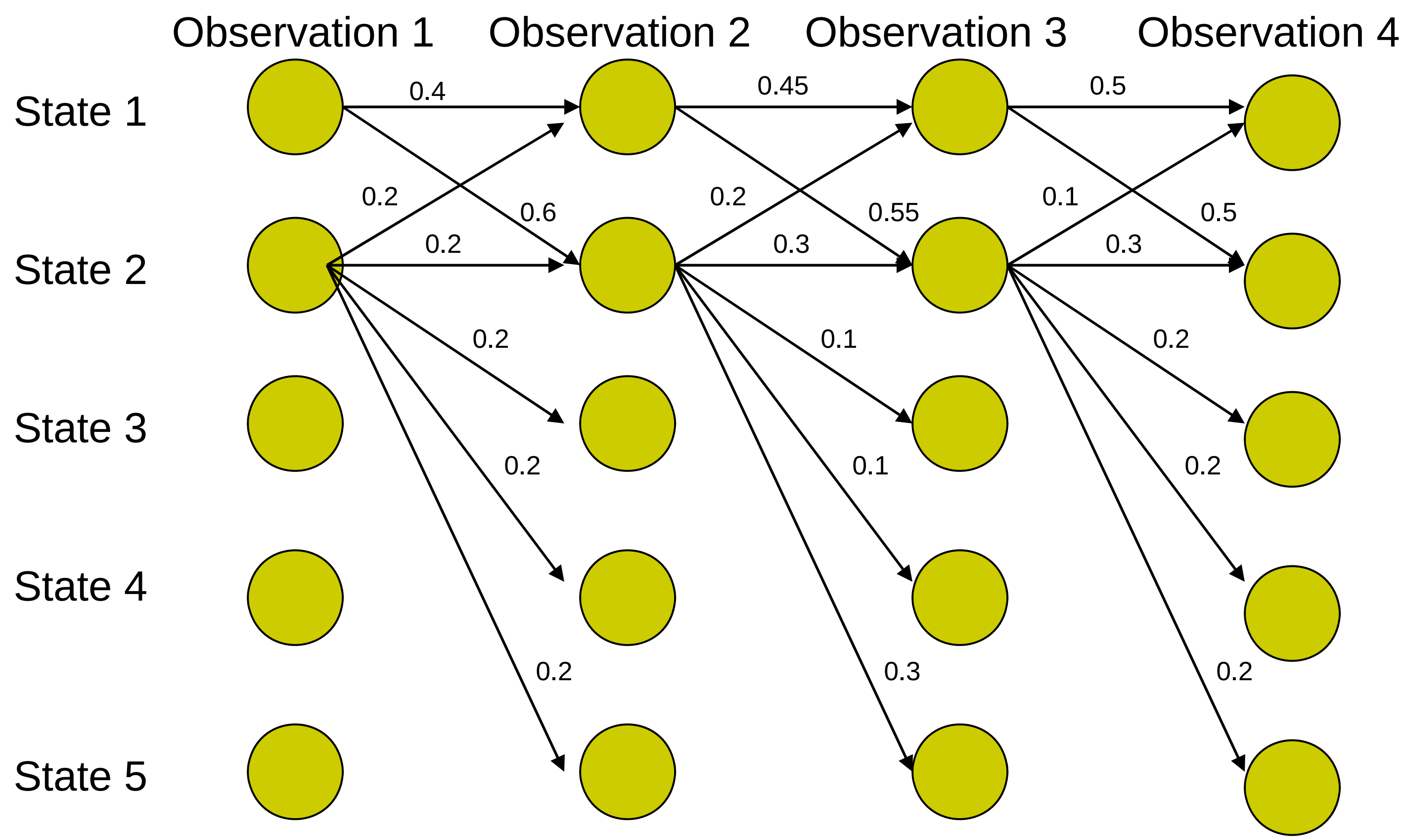
MEMM: the Label bias problem



What the local transition probabilities say:

- State 1 almost always prefers to go to state 2
- State 2 almost always prefer to stay in state 2

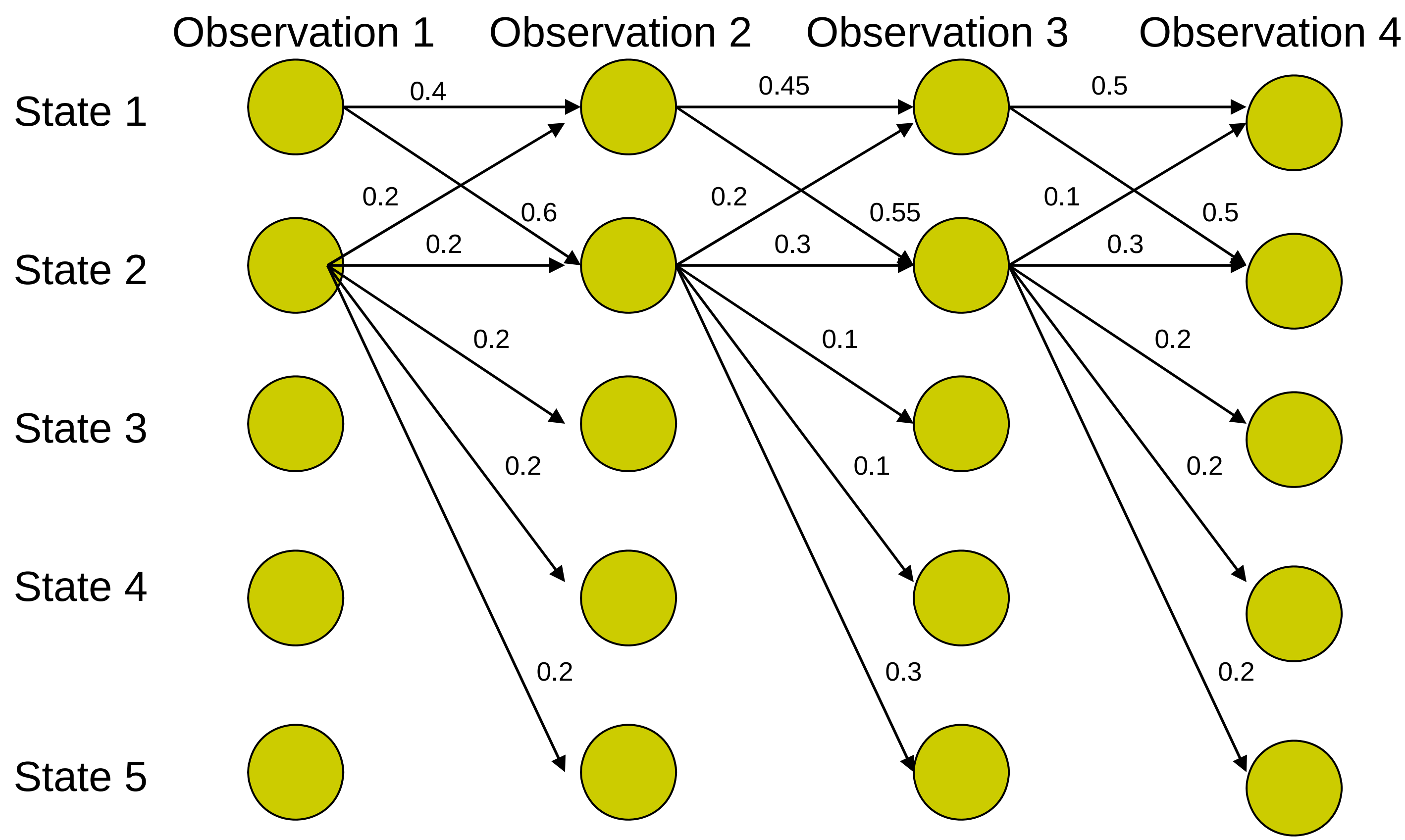
MEMM: the Label bias problem



Probability of path 1-> 1-> 1-> 1:

- 0.4 x 0.45 x 0.5 = 0.09

MEMM: the Label bias problem

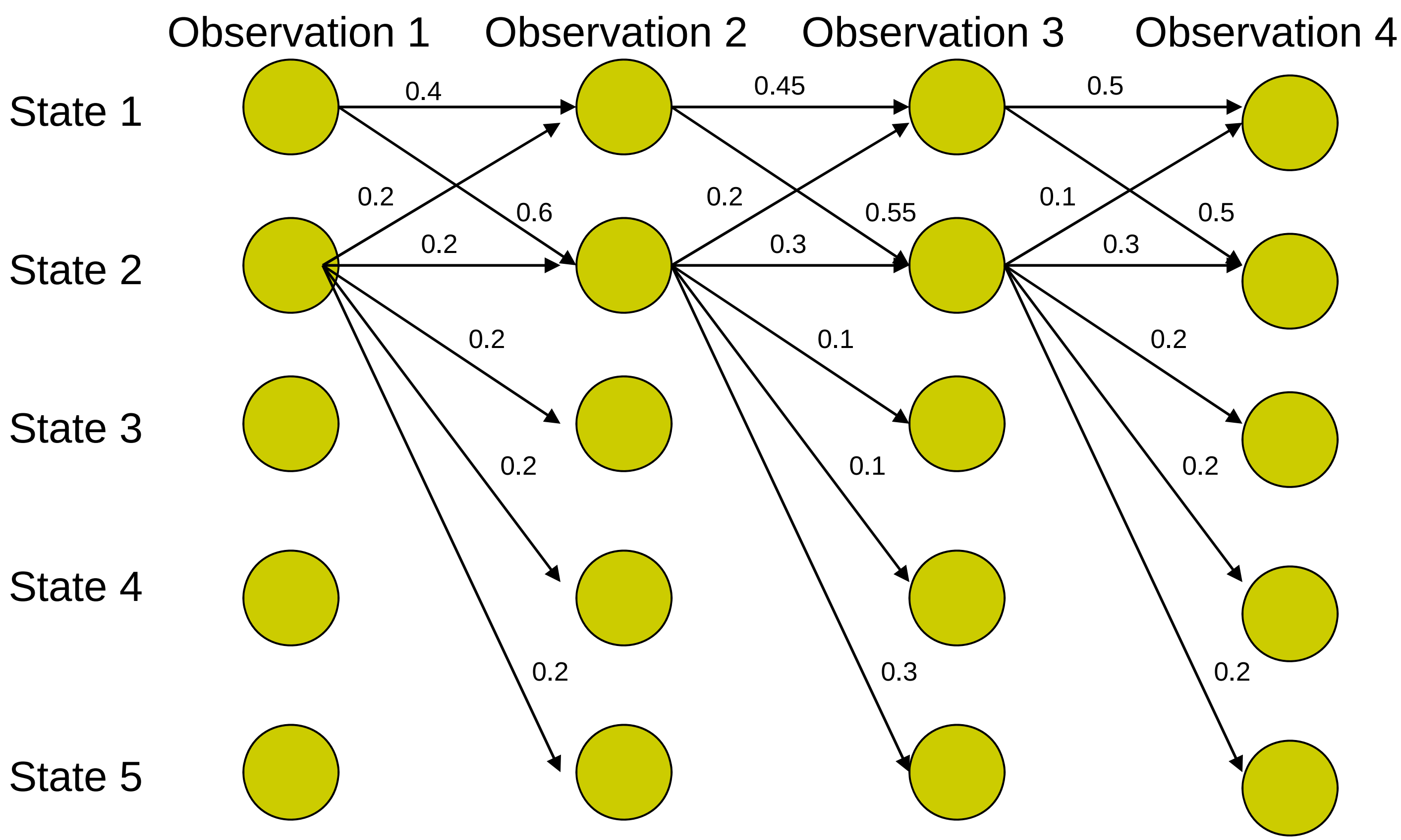


Probability of path 2->2->2->2 :

- 0.2 X 0.3 X 0.3 = 0.018

Other paths:
1-> 1-> 1-> 1: 0.09

MEMM: the Label bias problem

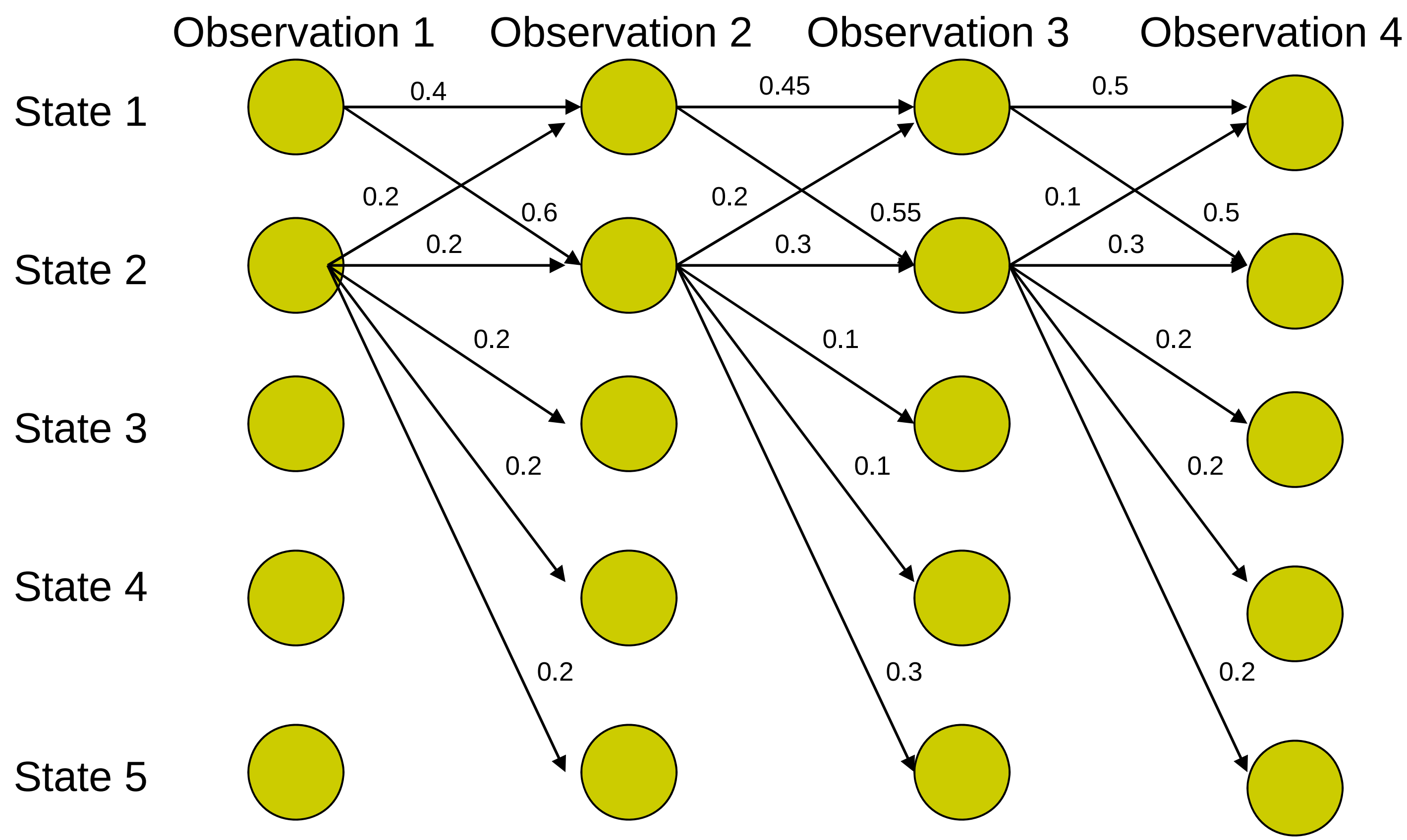


Probability of path 1->2->1->2:

- 0.6 X 0.2 X 0.5 = 0.06

Other paths:
1->1->1->1: 0.09
2->2->2->2: 0.018

MEMM: the Label bias problem

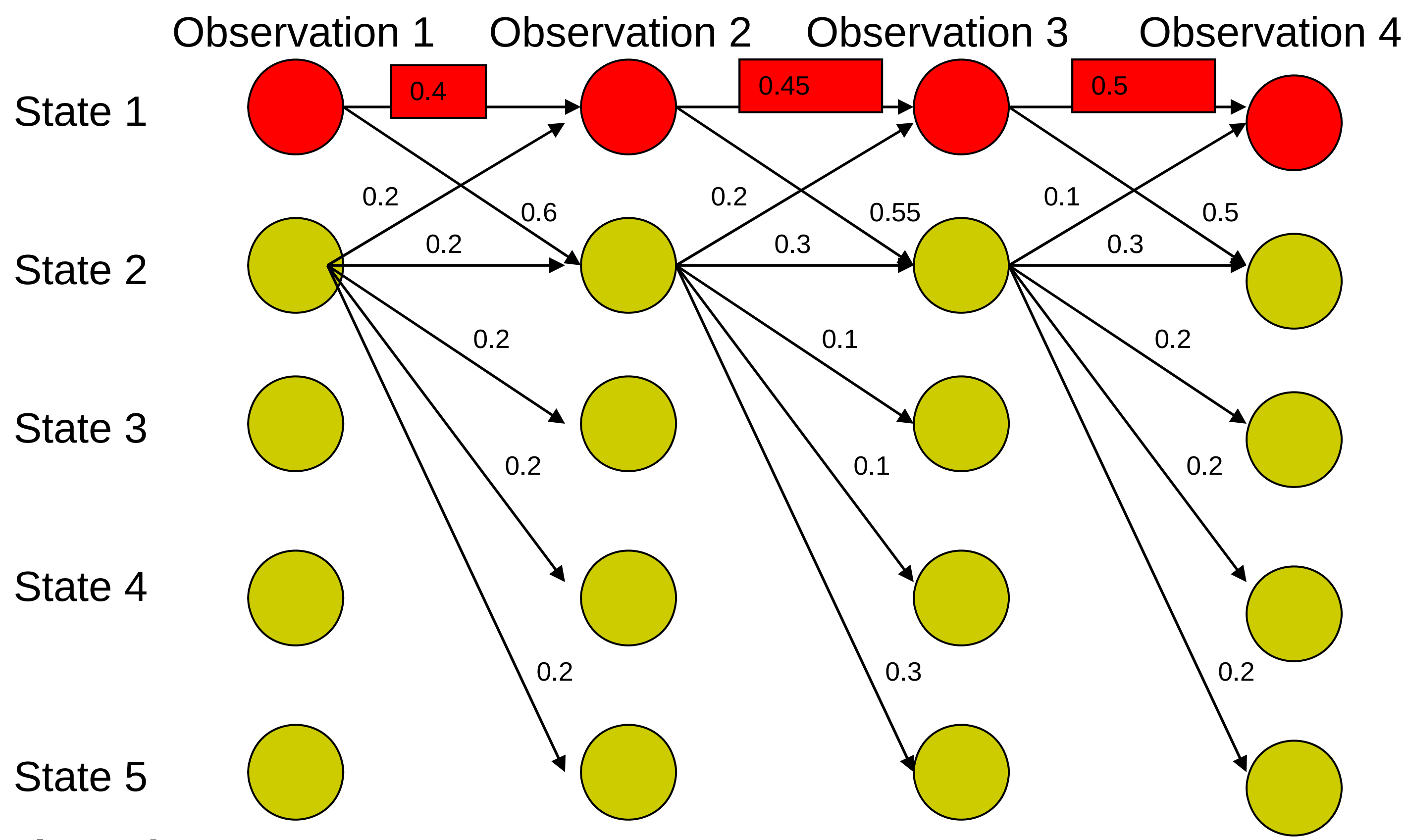


Probability of path 1->1->2->2:

- 0.4 X 0.55 X 0.3 = 0.066

Other paths:
1->1->1->1: 0.09
2->2->2->2: 0.018
1->2->1->2: 0.06

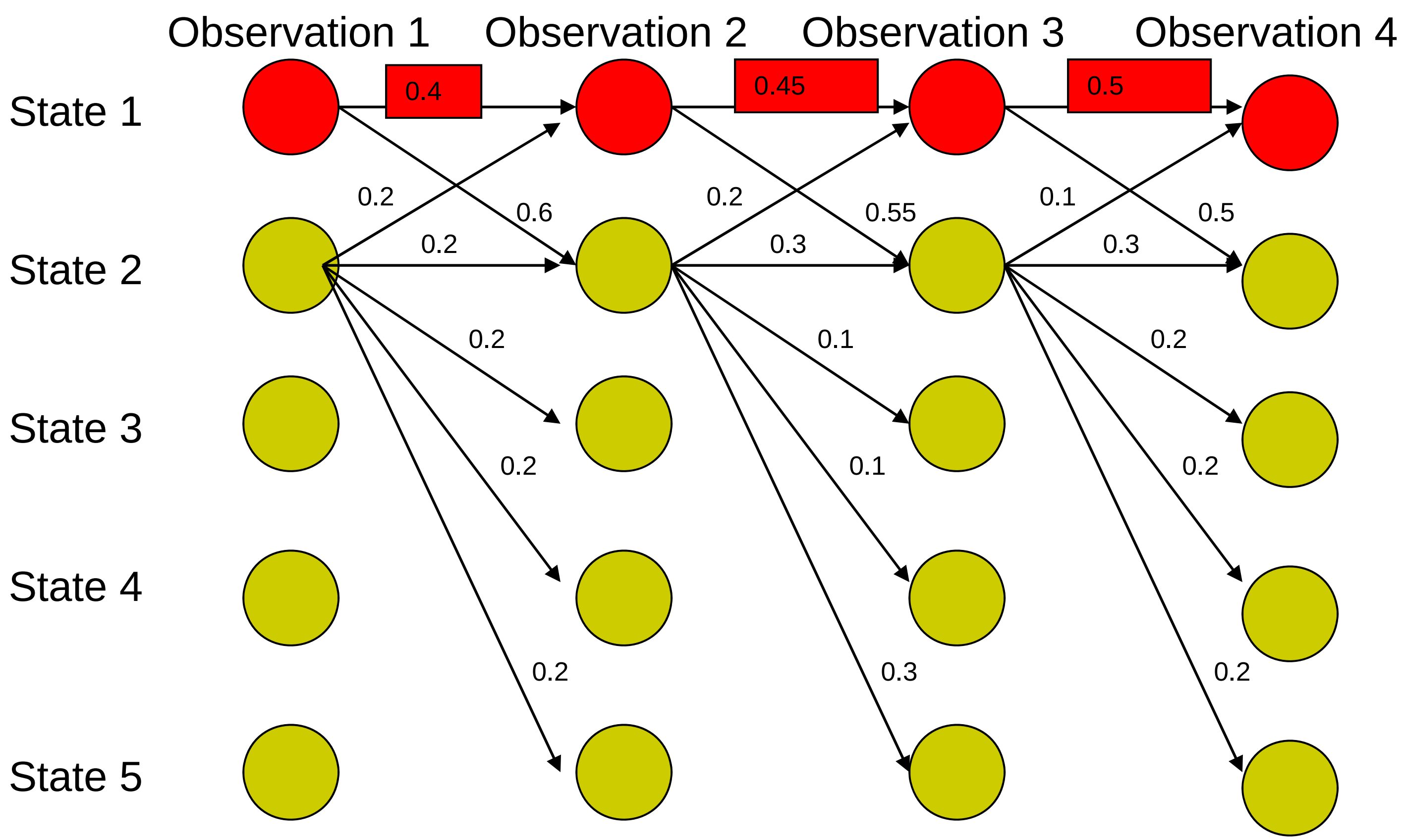
MEMM: the Label bias problem



Most Likely Path: 1-> 1-> 1-> 1

- Although **locally** it seems state 1 wants to go to state 2 and state 2 wants to remain in state 2.
- **why?**

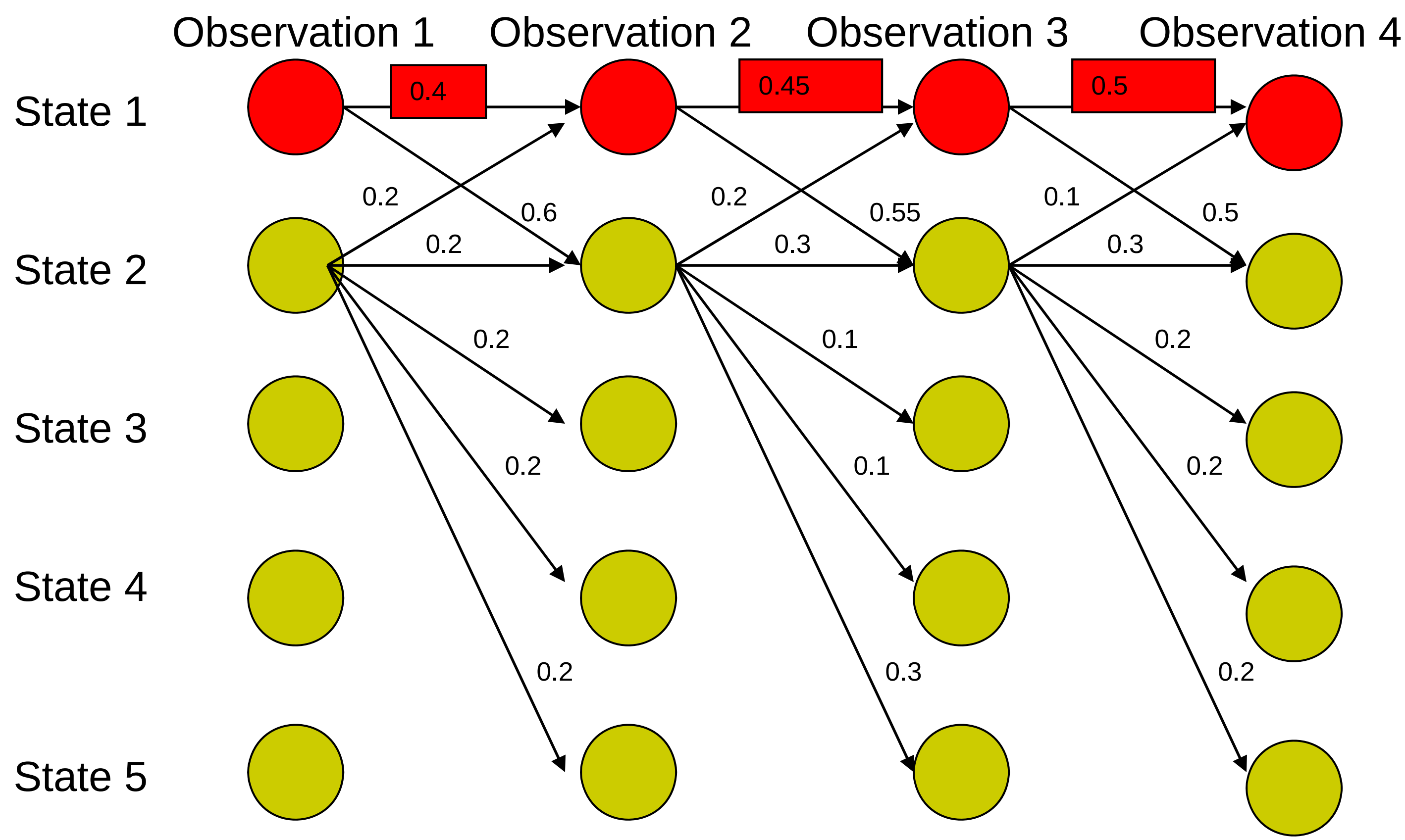
MEMM: the Label bias problem



Most Likely Path: 1-> 1-> 1-> 1

- State 1 has only two transitions but state 2 has 5:
  - Average transition probability from state 2 is lower

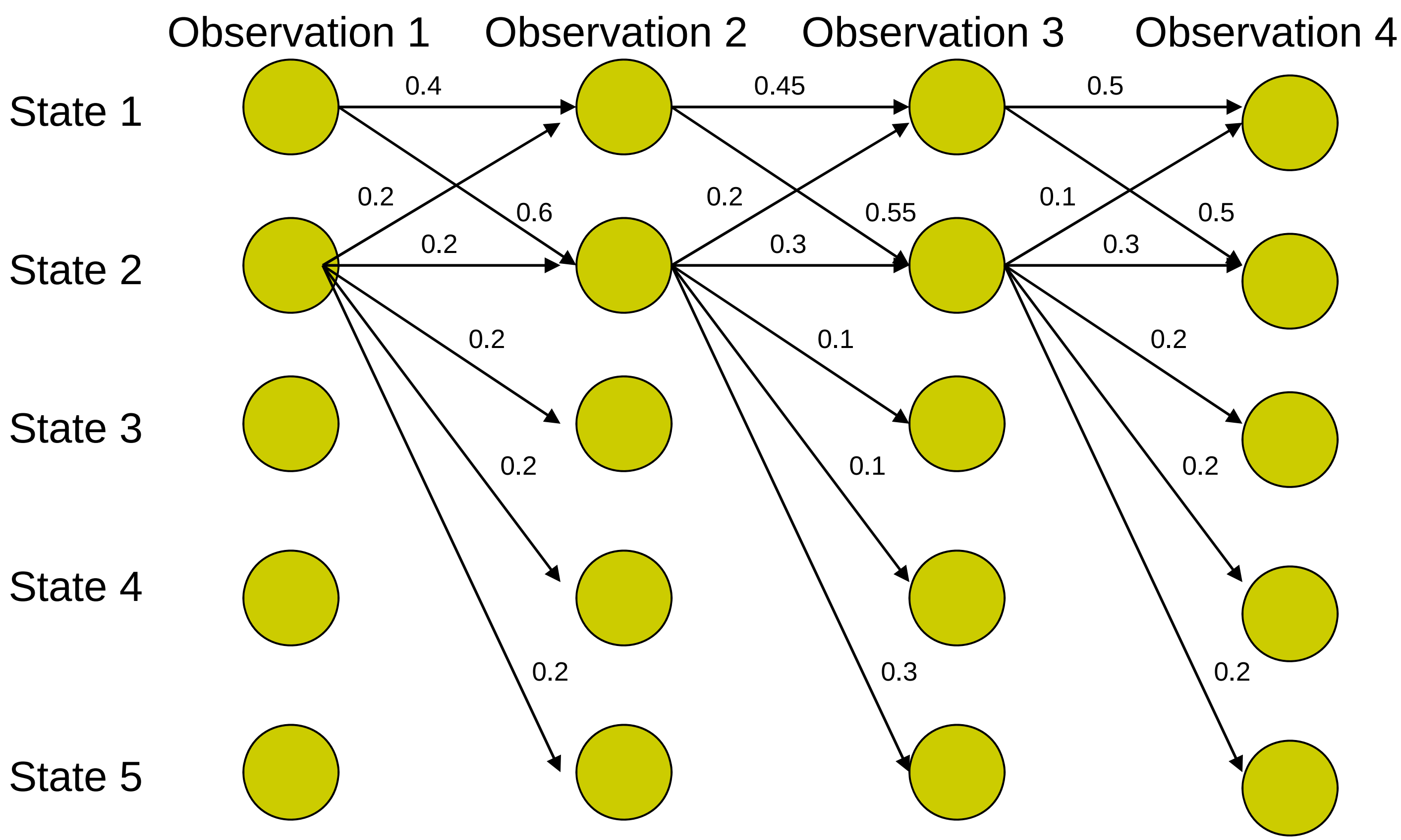
MEMM: the Label bias problem



Label bias problem in MEMM:

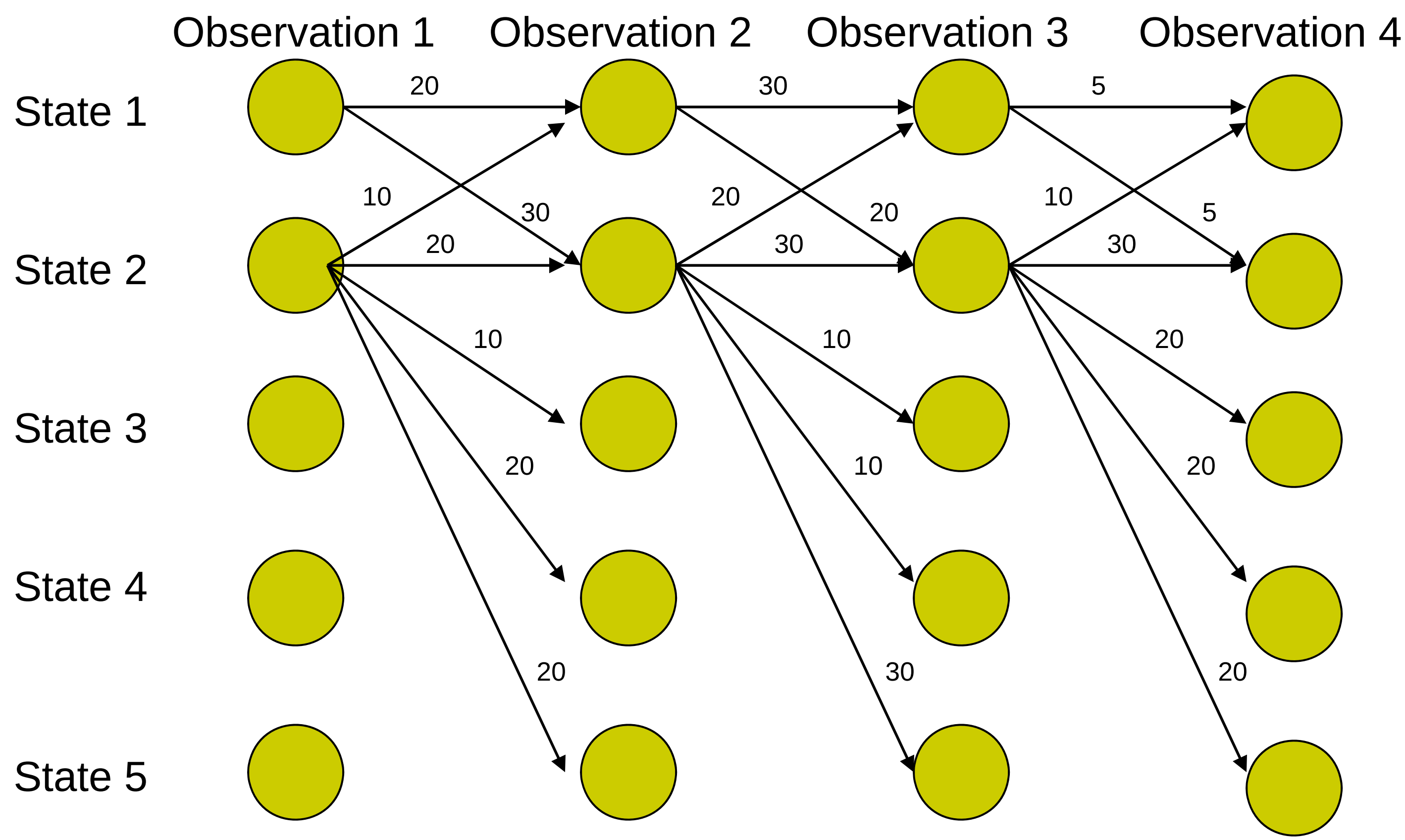
- Preference of states with lower number of transitions over others

Solution:
Do not normalize probabilities locally



From local probabilities ….

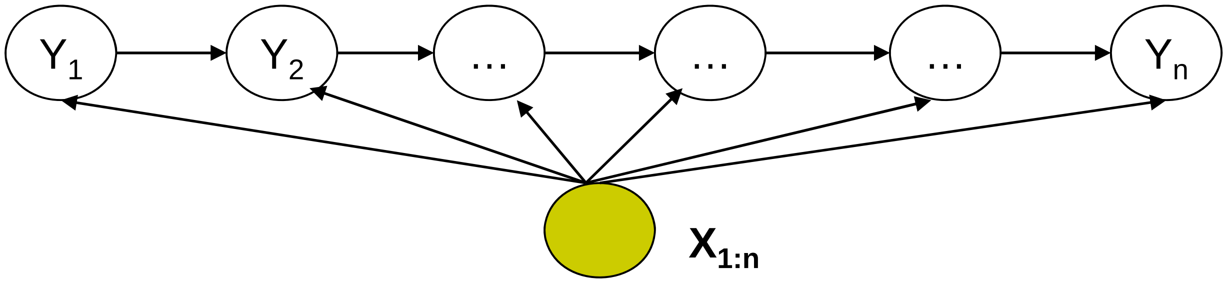
Solution:
Do not normalize probabilities locally



From local probabilities to local potentials

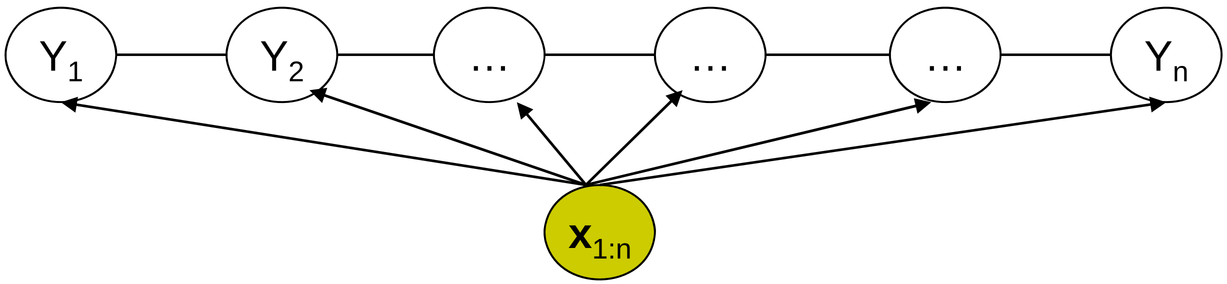
- States with lower transitions do not have an unfair advantage!

From MEMM ....



$$P(\mathbf{y}_{1:n}|\mathbf{x}_{1:n}) = \prod_{i=1}^{n} P(y_i|y_{i-1}, \mathbf{x}_{1:n}) = \prod_{i=1}^{n} \frac{\exp(\mathbf{w}^T \mathbf{f}(y_i, y_{i-1}, \mathbf{x}_{1:n}))}{Z(y_{i-1}, \mathbf{x}_{1:n})}$$

From MEMM to CRF

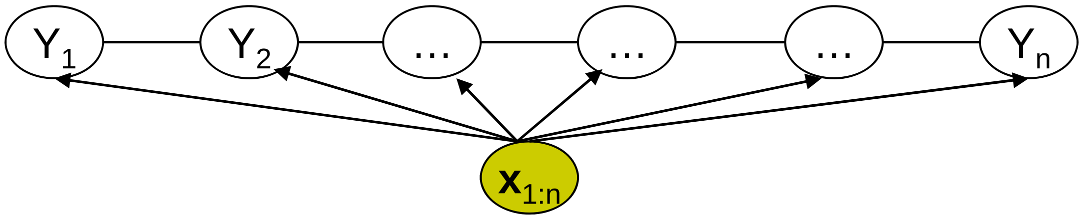


$$P(\mathbf{y}_{1:n}|\mathbf{x}_{1:n}) = \frac{1}{Z(\mathbf{x}_{1:n})}\prod_{i=1}^{n}\phi(y_i, y_{i-1}, \mathbf{x}_{1:n}) = \frac{1}{Z(\mathbf{x}_{1:n}, \mathbf{w})}\prod_{i=1}^{n}\exp(\mathbf{w}^T\mathbf{f}(y_i, y_{i-1}, \mathbf{x}_{1:n}))$$

- CRF is a partially directed model
  - Discriminative model like MEMM
  - Usage of global normalizer Z(**x**) overcomes the label bias problem of MEMM
  - Models the dependence between each state and the entire observation sequence (like MEMM)

Conditional Random Fields

- General parametric form:



$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x},\lambda,\mu)}\exp\left(\sum_{i=1}^{n}\left(\sum_{k}\lambda_k f_k(y_i,y_{i-1},\mathbf{x}) + \sum_{l}\mu_l g_l(y_i,\mathbf{x})\right)\right)$$

$$= \frac{1}{Z(\mathbf{x},\lambda,\mu)}\exp\left(\sum_{i=1}^{n}\left(\lambda^T\mathbf{f}(y_i,y_{i-1},\mathbf{x}) + \mu^T\mathbf{g}(y_i,\mathbf{x})\right)\right)$$

$$\text{where } Z(\mathbf{x},\lambda,\mu) = \sum_{\mathbf{y}}\exp\left(\sum_{i=1}^{n}\left(\lambda^T\mathbf{f}(y_i,y_{i-1},\mathbf{x}) + \mu^T\mathbf{g}(y_i,\mathbf{x})\right)\right)$$

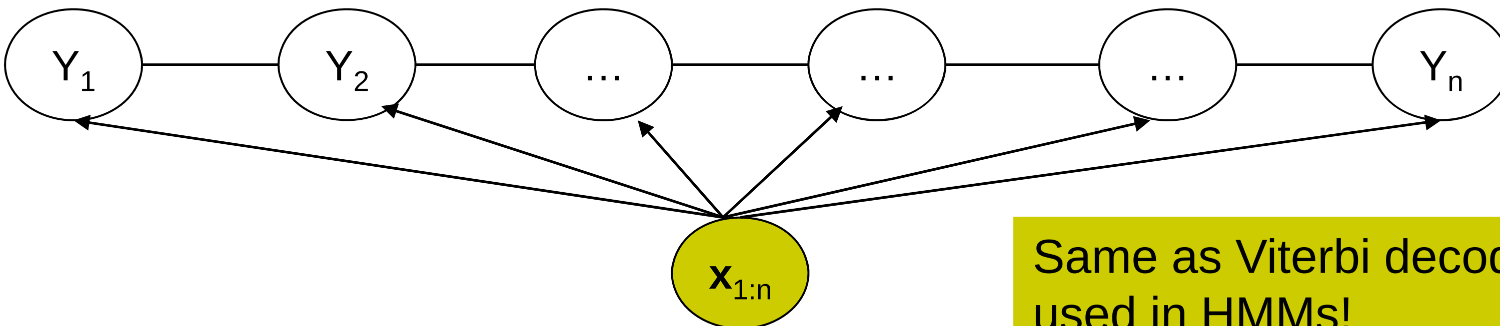
CRFs: Inference

- Given CRF parameters λ and μ, find the $\mathbf{y}^*$ that maximizes P(**y**|**x**)

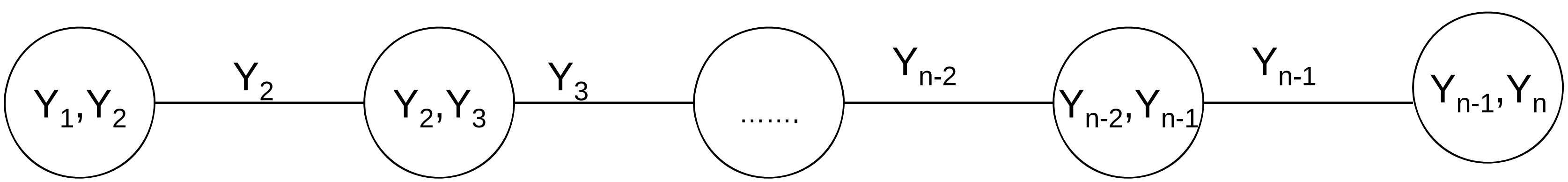
$$\mathbf{y}^* = \arg\max_{\mathbf{y}} \exp\left(\sum_{i=1}^{n} (\lambda^T \mathbf{f}(y_i, y_{i-1}, \mathbf{x}) + \mu^T \mathbf{g}(y_i, \mathbf{x}))\right)$$

  - Can ignore Z(**x**) because it is not a function of **y**

- Run the max-product algorithm on the junction-tree of CRF:



Same as Viterbi decoding used in HMMs!



CRF learning

- Given $\{(\mathbf{x}_d, \mathbf{y}_d)\}_{d=1}^N$, find λ^*, μ^* such that

$$
\begin{aligned}
\lambda*, \mu* &= \arg\max_{\lambda,\mu} L(\lambda, \mu) = \arg\max_{\lambda,\mu} \prod_{d=1}^{N} P(\mathbf{y}_d | \mathbf{x}_d, \lambda, \mu) \\
&= \arg\max_{\lambda,\mu} \prod_{d=1}^{N} \frac{1}{Z(\mathbf{x}_d, \lambda, \mu)} \exp\Big(\sum_{i=1}^{n} \big(\lambda^T \mathbf{f}(y_{d,i}, y_{d,i-1}, \mathbf{x}_d) + \mu^T \mathbf{g}(y_{d,i}, \mathbf{x}_d)\big)\Big) \\
&= \arg\max_{\lambda,\mu} \sum_{d=1}^{N} \Big(\sum_{i=1}^{n} \big(\lambda^T \mathbf{f}(y_{d,i}, y_{d,i-1}, \mathbf{x}_d) + \mu^T \mathbf{g}(y_{d,i}, \mathbf{x}_d)\big) - \log Z(\mathbf{x}_d, \lambda, \mu)\Big)
\end{aligned}
$$

- Computing the gradient w.r.t λ:

Gradient of the log-partition function in an exponential family is the expectation of the sufficient statistics.

$$
\nabla_\lambda L(\lambda, \mu) = \sum_{d=1}^{N} \Big(\sum_{i=1}^{n} \mathbf{f}(y_{d,i}, y_{d,i-1}, \mathbf{x}_d) - \sum_{\mathbf{y}} \big(P(\mathbf{y}|\mathbf{x_d}) \sum_{i=1}^{n} \mathbf{f}(y_{d,i}, y_{d,i-1}, \mathbf{x}_d)\big)\Big)
$$

CRF learning

$$\nabla_\lambda L(\lambda, \mu) = \sum_{d=1}^{N}\left(\sum_{i=1}^{n}\mathbf{f}(y_{d,i}, y_{d,i-1}, \mathbf{x}_d) - \sum_{\mathbf{y}}\left(P(\mathbf{y}|\mathbf{x_d})\sum_{i=1}^{n}\mathbf{f}(y_i, y_{i-1}, \mathbf{x}_d)\right)\right)$$

- Computing the model expectations:
  - Requires exponentially large number of summations: Is it intractable?

$$\sum_{\mathbf{y}}\left(P(\mathbf{y}|\mathbf{x}_d)\sum_{i=1}^{n}\mathbf{f}(y_i, y_{i-1}, \mathbf{x}_d)\right) = \sum_{i=1}^{n}\left(\sum_{\mathbf{y}}\mathbf{f}(y_i, y_{i-1}, \mathbf{x}_d)P(\mathbf{y}|\mathbf{x}_d)\right)$$

$$= \sum_{i=1}^{n}\sum_{y_i, y_{i-1}}\mathbf{f}(y_i, y_{i-1}, \mathbf{x}_d)P(y_i, y_{i-1}|\mathbf{x}_d)$$

Expectation of **f** over the corresponding marginal probability of neighboring nodes!!

- Tractable!
  - Can compute marginals using the sum-product algorithm on the chain

CRF learning

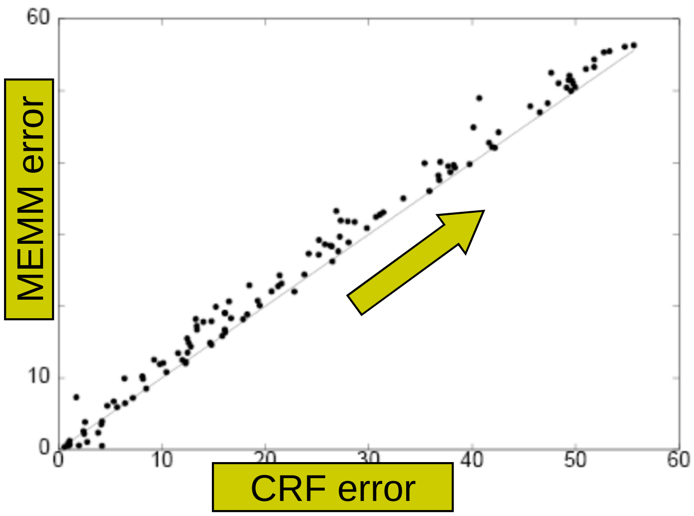
- In practice, we use a Gaussian Regularizer for the parameter vector to improve generalizability

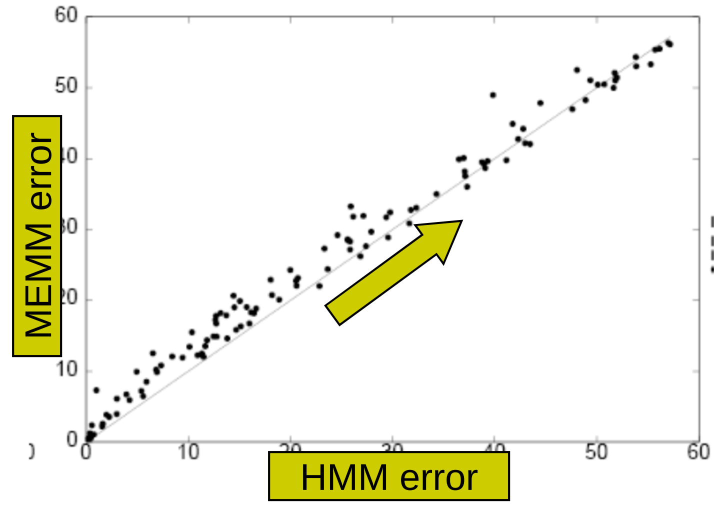
$$\lambda*, \mu* \quad = \quad \arg\max_{\lambda,\mu} \sum_{d=1}^{N} \log P(\mathbf{y}_d | \mathbf{x}_d, \lambda, \mu) - \frac{1}{2\sigma^2}(\lambda^T \lambda + \mu^T \mu)$$

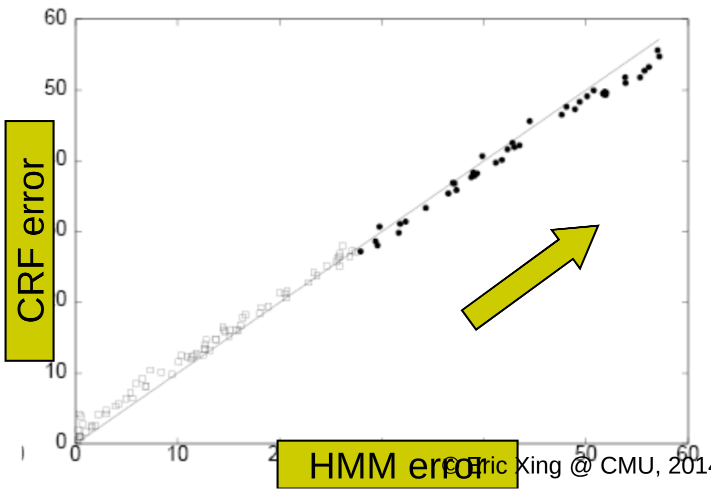
- In practice, gradient ascent has very slow convergence
  - Alternatives:
    - Conjugate Gradient method
    - Limited Memory Quasi-Newton Methods

CRFs: some empirical results

- Comparison of error rates on synthetic data







Data is increasingly higher order in the direction of arrow

CRFs achieve the lowest error rate for higher order data

CRFs: some empirical results

- Parts of Speech tagging

model	*error*	*oov error*
HMM	5.69%	45.99%
MEMM	6.37%	54.61%
CRF	5.55%	48.05%
MEMM$^+$	4.81%	26.99%
CRF$^+$	4.27%	23.76%

$^+$Using spelling features

- Using same set of features: HMM >=< CRF > MEMM
- Using additional overlapping features: CRF$^+$ > MEMM$^+$ >> HMM

Summary

- Conditional Random Fields are partially directed discriminative models
- They overcome the label bias problem of MEMMs by using a global normalizer
- Inference for 1-D chain CRFs is exact
  - Same as Max-product or Viterbi decoding
- Learning also is exact
  - globally optimum parameters can be learned
  - Requires using sum-product or forward-backward algorithm
- CRFs involving arbitrary graph structure are intractable in general
  - E.g.: Grid CRFs
  - Inference and learning require approximation techniques
    - MCMC sampling
    - Variational methods
    - Loopy BP

Other CRFs

- So far we have discussed only 1-dimensional chain CRFs
  - Inference and learning: exact
- We could also have CRFs for arbitrary graph structure
  - E.g: Grid CRFs
  - Inference and learning no longer tractable
  - Approximate techniques used
    - MCMC Sampling
    - Variational Inference
    - Loopy Belief Propagation
  - We will discuss these techniques soon

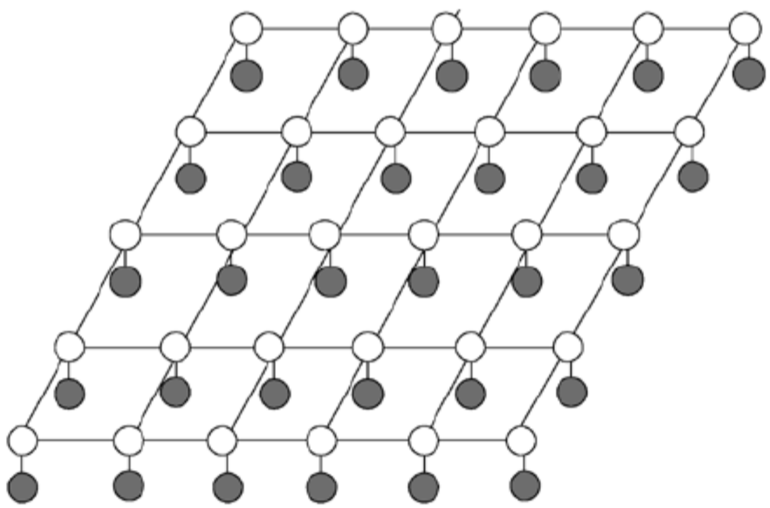


Image Segmentation

- Image segmentation (FG/BG) by modeling of interactions btw RVs
  - Images are noisy.
  - Objects occupy continuous regions in an image.

[Nowozin,Lampert 2012]

Input image

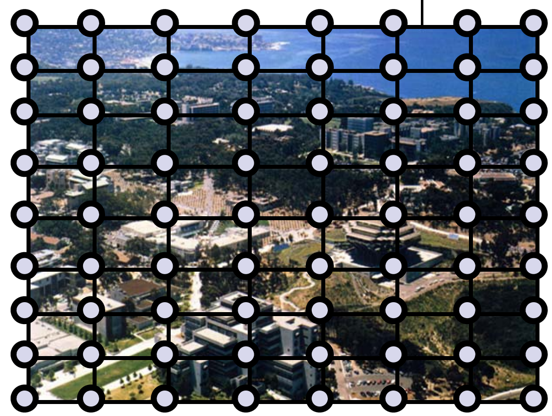
Pixel-wise separate optimal labeling

Locally-consistent joint optimal labeling

$$Y^* = \arg\max_{y \in \{0,1\}^n} \Big[\underbrace{\sum_{i \in S} V_i(y_i, X)}_{\text{Unary Term}} + \underbrace{\sum_{i \in S} \sum_{j \in N_i} V_{i,j}(y_i, y_j)}_{\text{Pairwise Term}} \Big].$$

Y: labels
X: data (features)
S: pixels
N_i: neighbors of pixel i

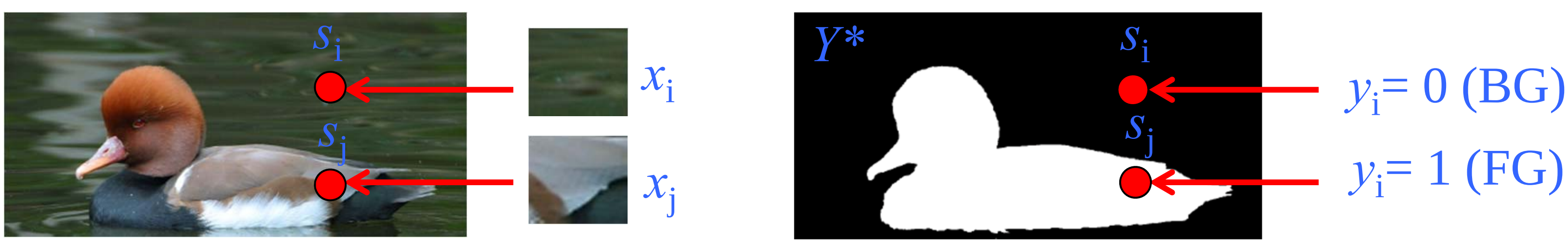
Undirected Graphical Models (with an Image Labeling Example)



- Image can be represented by 4-connected 2D grid.

- MRF / CRF with image labeling problem
  - $X=\{x_i\}_{i \in S}$: observed data of an image.
  - x_i: data at i-th site (pixel or block) of the image set S
  - $Y=\{y_i\}_{i \in S}$: (hidden) labels at i-th site. $y_i \in \{1,\ldots,L\}$.

- Object: maximize the conditional probability $Y^*=\text{argmax}_Y \, P(Y|X)$



MRF (Markov Random Field)

- Definition: $Y=\{y_i\}_{i \in S}$ is called Markov Random Field on the set S, with respect to neighborhood system *N*, iff for all $i \in S$,

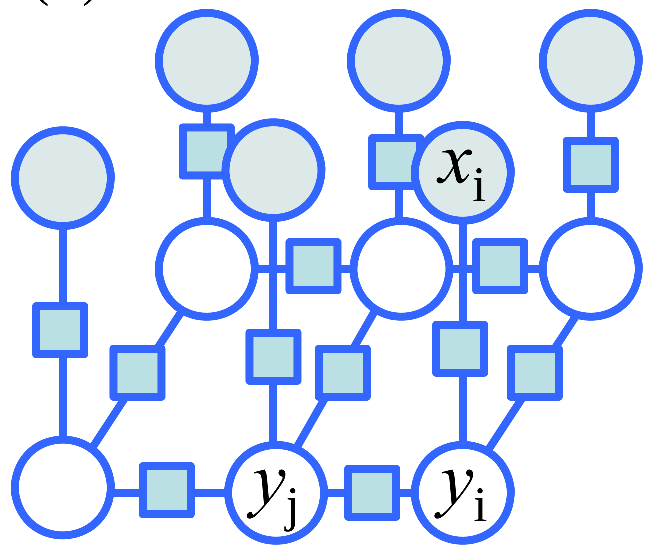
$$P(y_i | y_{S-\{i\}}) = P(y_i | y_{N_i}).$$

- The posterior probability is

$$P(Y \mid X) = \frac{P(X,Y)}{P(X)} \propto P(X \mid Y) P(Y) = \underbrace{\prod_{i \in S} P(x_i \mid y_i)}_{(1)} \cdot \underbrace{P(Y)}_{(2)}$$

  - (1) Very strict independence assumptions for tractability: Label of each site is a function of data only at that site.
  - (2) $P(Y)$ is modeled as a MRF

$$P(Y) = \frac{1}{Z} \prod_{c \in C} \psi_c(y_c)$$



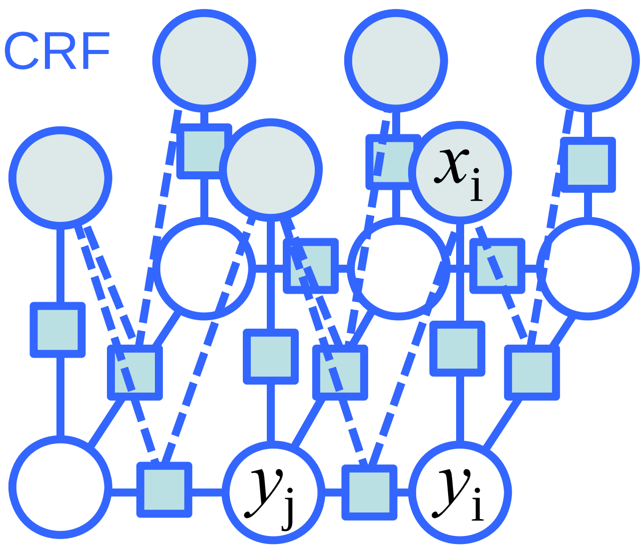
CRF

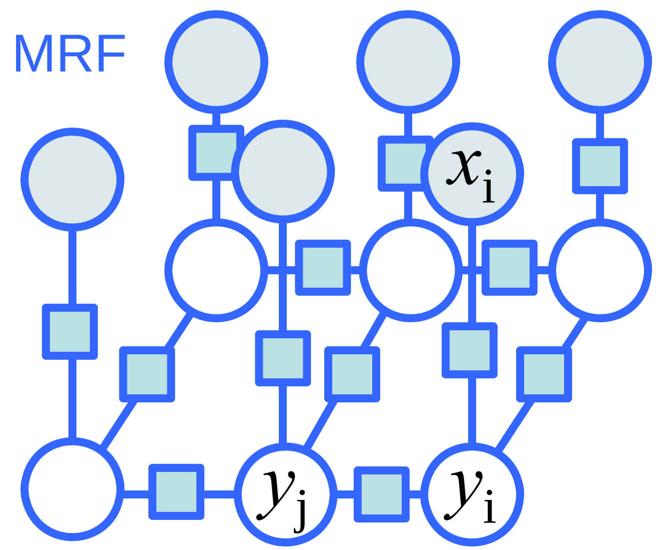
- Definition: Let $G = (S, E)$, then (X, Y) is said to be a Conditional Random Field (CRF) if, when conditioned on X, the random variables y_i obey the Markov property with respect to the graph

$$P(y_i|X,y_{S-\{i\}}) = P(y_i|X,y_{N_i})$$

MRF: $P(y_i|y_{S-\{i\}}) = P(y_i|y_{N_i})$

  - Globally conditioned on the observation X





CRF vs MRF

- MRF: two-step generative model
  - Infer likelihood $P(X|Y)$ and prior $P(Y)$
  - Use Bayes theorem to determine posterior $P(Y|X)$

$$P(Y \mid X) = \frac{P(X,Y)}{P(X)} \propto P(X \mid Y)P(Y) = \prod_{i \in S} P(x_i \mid y_i) \cdot \frac{1}{Z} \prod_{c \in C} \psi_c(y_c)$$

- CRF: one-step discriminative model
  - Directly Infer posterior $P(Y|X)$

- Popular Formulation

Assumption

MRF $$P(Y \mid X) = \frac{1}{Z} \exp\Big(\sum_{i \in S} \log p(x_i \mid y_i) + \sum_{i \in S} \sum_{i' \in N_i} V_2(y_i, y_{i'})\Big)$$

Potts model for $P(Y)$ with only pairwise potential

CRF $$P(Y \mid X) = \frac{1}{Z} \exp\Big(-\sum_{i \in S} V_1(y_i \mid X) + \sum_{i \in S} \sum_{i' \in N_i} V_2(y_i, y_{i'} \mid X)\Big)$$

Only up to pairwise clique potentials

Example of CRF – DRF

- A special type of CRF
  - The unary and pairwise potentials are designed using local discriminative classifiers.
  - Posterior

Association — Interaction

$$P(Y \mid X) = \frac{1}{Z}\exp\Big(\sum_{i\in S} \underbrace{A_i(y_i, X)}_{\text{Association}} + \sum_{i\in S}\sum_{j\in N_i} \underbrace{I_{ij}(y_i, y_j, X)}_{\text{Interaction}}\Big)$$

- Association Potential
  - Local discriminative model for site i: using logistic link with GLM.

$$A_i(y_i, X) = \log P(y_i \mid f_i(X)) \qquad P(y_i = 1 \mid f_i(X)) = \frac{1}{1+\exp(-(w^T f_i(X)))} = \sigma(w^T f_i(X))$$

- Interaction Potential
  - Measure of how likely site *i* and *j* have the same label given *X*

$$I_{ij}(y_i, y_j, X) = \underbrace{k y_i y_j}_{(1)} + \underbrace{(1-k)(2\sigma(y_i y_j \mu_{ij}(X)) - 1))}_{(2)}$$

(1) Data-independent smoothing term

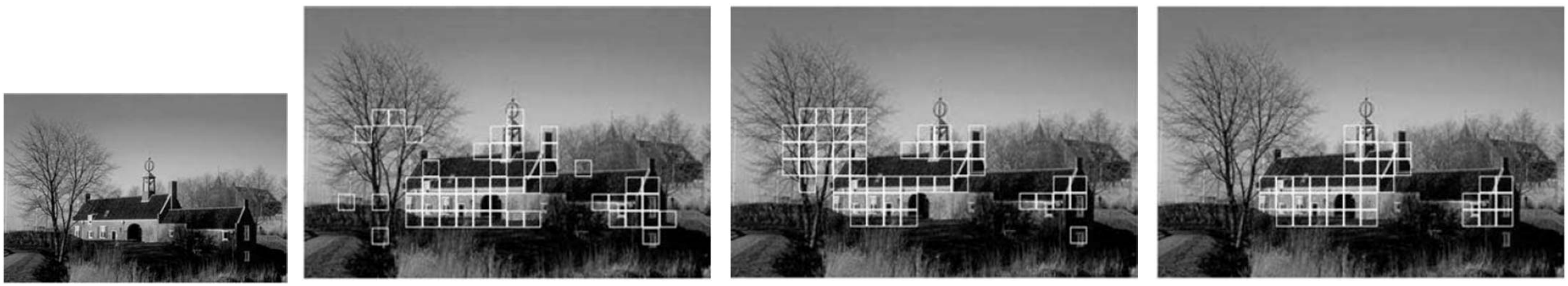
(2) Data-dependent pairwise logistic function

S. Kumar and M. Hebert. Discriminative Random Fields. IJCV, 2006.

Example of CRF – DRF Results

- Task: Detecting man-made structure in natural scenes.
  - Each image is divided in non-overlapping 16x16 tile blocks.

- An example



Input image Logistic MRF DRF

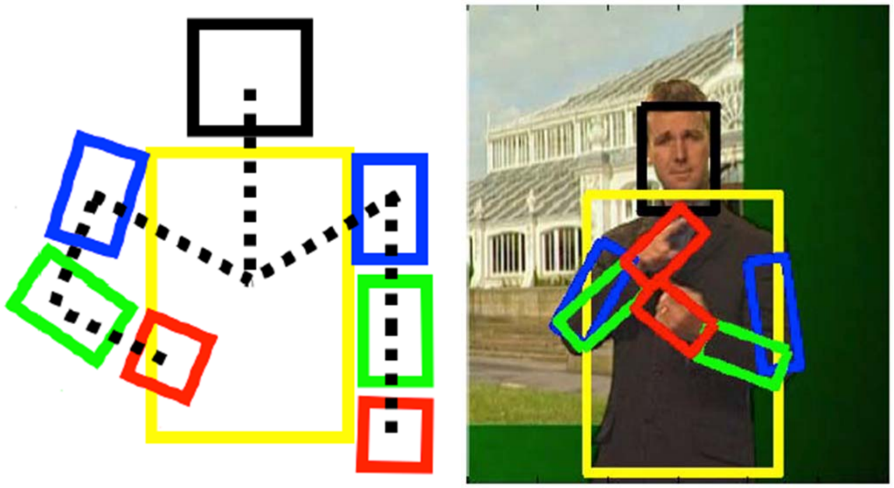
  - Logistic: No smoothness in the labels
  - MRF: Smoothed False positive. Lack of neighborhood interaction of the data

S. Kumar and M. Hebert. Discriminative Random Fields. IJCV, 2006.

Example of CRF –Body Pose Estimation

- Task: Estimate a body pose.
  - Need to detect parts of human body
  - Appearance + Geometric configuration.
  - A large number of DOFs
- Use CRF to model a human body
  - Nodes: Parts (head, torso, upper/ lower left/right arms).
    $L=(l_1,\ldots, l_6),\ l_i = [x_i, y_i, \theta_i]$.
  - Edges: Pairwise linkage between parts
  - Tree vs. Graph

[Zisserman 2010]



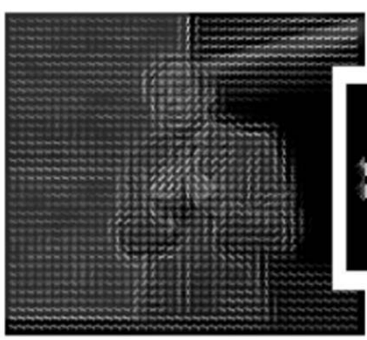
V. Ferrari et al. Progressive search space reduction for human pose estimation. CVPR 2008.
D. Ramanan. Learning to Parse Images of Articulated Bodies." NIPS 2006.

Example of CRF –Body Pose Estimation

- Posterior of configuration

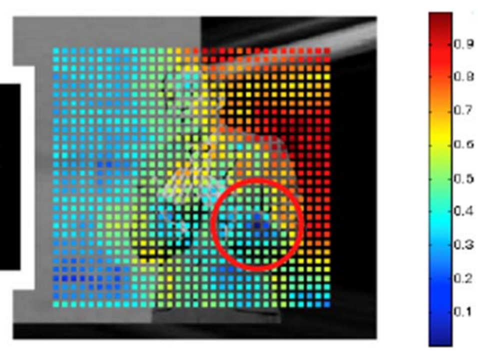
$$P(L \mid I) \propto \exp\left(\sum_i \Phi(l_i) + \sum_{(i,j)\in E} \Psi(l_i, l_j)\right)$$

  - $\psi(l_i, l_j)$: relative position with geometric constraints
  - $\phi(l_i)$: local image evidence for a part in a particular location
  - If E is a tree, exact inference is efficiently performed by BP.
- Example of unary and pairwise terms
  - Unary term: appearance feature



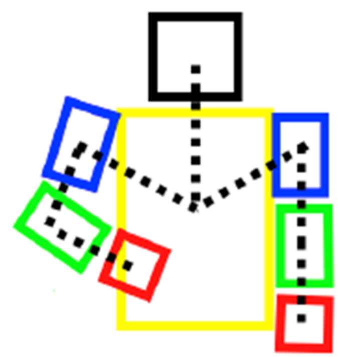
HOG of image

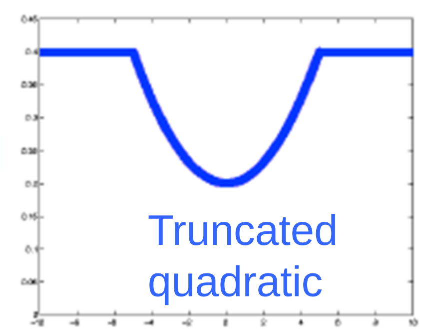
HOG of lower arm template (learned)



L2 Distance

  - Pairwise term: kinematic layout

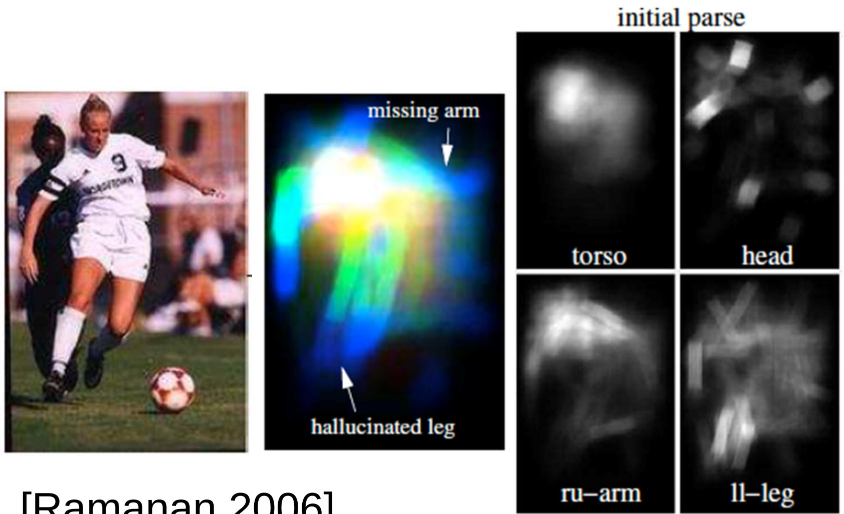




[Zisserman 2010]

Example of CRF – Results of Body Pose Estimation

- Examples of results



[Ramanan 2006]

[Ferrari et al. 2008]

- Datasets and codes are available.
  - http://www.ics.uci.edu/~dramanan/papers/parse/
  - http://www.robots.ox.ac.uk/~vgg/research/pose_estimation/