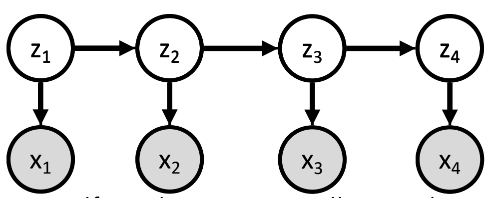
Structured Prediction
(With Special Focus on Sequence Prediction)

* Parts of the following presentation were taken from Professor Mark Dredze's slides

Sequence Models

A HMM is a directed graphical model (BN)



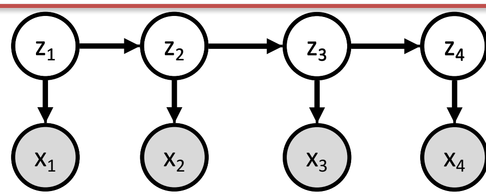
What happens if we have an undirected graphical model?

- Markov Random Field (For modelling $p(\boldsymbol{Z}, \boldsymbol{X})$)
- Conditional Random Fields (For modelling $p(\boldsymbol{Z} \mid \boldsymbol{X})$)
  Go over it in detail later

Notes on HMMs

- An HMM can have continuous or discrete emissions
  - Discrete- base pair, word in sentence
  - Continuous- stock price, frequency of a sound
- An HMM has discrete hidden states
- A Linear Dynamical System has continuous hidden states
- Note that the time-steps in both HMM and LDS are discrete. Chains with continous time-steps are stochastic processes.
- **We will skip all of these topics in this course.**
  Teachers open the door, but you must enter by yourself

Joint Probability of HMM

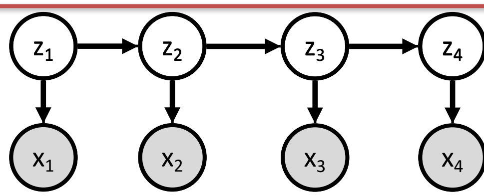


The joint probability of a 1st order HMM

$$p(\boldsymbol{x},\boldsymbol{z} \mid \boldsymbol{\theta}) = p(z_1|\pi) \prod_{n\in[2,N]} p(z_n|z_{n-1}, A) \prod_{m\in[1,N]} p(x_m|z_m, \varphi)$$

- **A** is a matrix of transition probabilities
  - A_{ij} is the probability of moving from state i to j
- π- vector with starting probabilities
- φ – emission probabilities (matrix)
  - φ_{ij} is the probability of observation j in state i

Joint Probability of HMM



Joint Prob. of Single Sequence of Observations. Assume sequence 1 is independent of sequence 2

The joint probability of a 1st order HMM

$$p(\boldsymbol{x},\boldsymbol{z}\mid\boldsymbol{\theta}) = p(z_1|\pi)\prod_{n\in[2,N]} p(z_n|z_{n-1},A)\prod_{m\in[1,N]} p(x_m|z_m,\varphi)$$

- **A** is a matrix of transition probabilities
  - A_{ij} is the probability of moving from state i to j
- π- vector with starting probabilities
- φ – emission probabilities (matrix)
  - φ_{ij} is the probability of observation j in state i

Joint Probability of HMM/MRF

The joint probability of the HMM can be written using potential functions

- HMM $$p(\boldsymbol{x},\boldsymbol{z}\mid\boldsymbol{\theta}) = p(z_1|\pi)\prod_{n\in[2,N]} p(z_n|z_{n-1},A)\prod_{m\in[1,N]} p(x_m|z_m,\varphi)$$

(In comparison to MRF)

- MRF $$p(\boldsymbol{x},\boldsymbol{z}\mid\boldsymbol{\theta}) = \frac{1}{\mathcal{Z}}\psi_1(z_1)\prod_{n\in[2,N]} \psi_{n,n-1}(z_n|z_{n-1},A)\prod_{m\in[1,N]} \psi_n(x_m|z_m,\varphi)$$

HMM Training

- If we actually observe **Z**
  - Just use Maximum Likelihood Estimation
- What if we observe only some **Z**
  - Case 1: only some examples are labeled with **Z**
  - Case 2: each example has only some labels for **Z**
- What if we observe no **Z**

Unsupervised Training of HMM
EM with missing/hidden data (1/2)

We will spend time here

How do we maximize p(**X**) when we don't know **Z** ? **EM**

EM: $\boldsymbol{\theta}^{t+1} \leftarrow \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \underset{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^t)}{\mathrm{E}} [\log(p(\boldsymbol{X},\boldsymbol{Z} \mid \boldsymbol{\theta}))]$ *(* also called Q function)*

$$\underset{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^t)}{\mathrm{E}} [\log(p(\boldsymbol{X},\boldsymbol{Z} \mid \boldsymbol{\theta}))] = \sum_i E_{p(\boldsymbol{z_i}|\boldsymbol{x_i},\boldsymbol{\theta}^t)}[\log(p(\boldsymbol{x_i},\boldsymbol{z_i},\boldsymbol{\theta})]$$

$$p(\boldsymbol{x_i},\boldsymbol{z_i} \mid \boldsymbol{\theta}) = p(z_{i,1}|\boldsymbol{\pi}) \prod_{n\in[2,N_i]} p(z_{i,n}|z_{i,n-1},\boldsymbol{A}) \prod_{m\in[1,N_i]} p(x_{i,m}|z_{i,m},\boldsymbol{\varphi})$$

$x_{i,m}$ is the mth word of the ith sentence. It is observed.

What is $E_{q_i(z)}[\log(p(\boldsymbol{x_i},\boldsymbol{z_i},\boldsymbol{\theta})]$?

What is $\log(p(\boldsymbol{x_i},\boldsymbol{z_i},\boldsymbol{\theta})$?

What is $p(z_{i,n}|z_{i,n-1},\boldsymbol{A})$?) $= \prod_{l\in[K]}\prod_{m\in[K]} a_{lm}^{I_{z_{i,n}=l,z_{i,n-1}=m}}$

Note: the expected values of indicator function of event X equal the probability of X

$$Q(\theta,\theta^{old}) = \sum_{k=1}^{K}\gamma(z_{1k})\log\pi_k + \sum_{n=2}^{N}\sum_{j=1}^{K}\sum_{k=1}^{K}\xi(z_{n-1,j},z_{nk})\log A_{jk} + \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\log p(x_n|\phi_k)$$

$$\gamma(z_n) = p(z_n \mid X,\theta^{old}) \qquad \xi(z_{n-1},z_n) = p(z_{n-1},z_n \mid X,\theta^{old})$$

Terminology from Bishop

Unsupervised Training of HMM
EM with missing/hidden data (2/2)

- M-Step (Derivation)

$$\pi_k = \frac{\gamma(z_{1k})}{\sum_{j=1}^{K}\gamma(z_{1j})} \qquad A_{jk} = \frac{\sum_{n=2}^{N}\xi(z_{n-1,j}, z_{nk})}{\sum_{l=1}^{K}\sum_{n=2}^{N}\xi(z_{n-1,j}, z_{nl})} \qquad \phi_{ik} = \frac{\sum_{n=1}^{N}\gamma(z_{nk})x_{ni}}{\sum_{n=1}^{N}\gamma(z_{nk})}$$

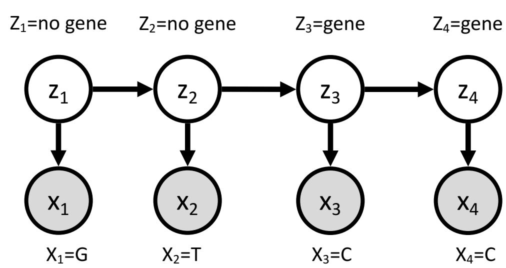
- How can we get:

$$\gamma(z_n) = p(z_n \mid X, \theta^{old}) \qquad \xi(z_{n-1}, z_n) = p(z_{n-1}, z_n \mid X, \theta^{old})$$

  - What is the probability of being in state z_n?
  - What is the probability of being in state z_n and z_{n+1}?
  - HINT: How can we get marginals of posterior distribution?

Prediction

- Given a new sequence **X**, find the most likely set of states to have generated **X.** I.e. Find the sequence **Z** with the maximum probability given **X**



- How do we infer the most likely state in a graphical model? Max Product Algorithm
  - Special case for HMM called Viterbi Decoding

The Max-Product Algorithm

The Max-Product Algorithm finds the optimal joint valuation of random variables that maximizes the joint probability:

- I.e. Max-product finds the mode of joint distribution

Remember, maximum marginals ≠ joint maximum.

	$x=0$	$x=1$
$y=0$	0.3	0.4
$y=1$	0.3	0.0

$\arg\max_x p(x,y) = 1 \qquad \arg\max_x p(x) = 0$

The Max-Product Algorithm Over a chain

$$p(\mathbf{x}^{\max}) = \max_{\mathbf{x}} p(\mathbf{x}) = \max_{x_1} \ldots \max_{x_M} p(\mathbf{x})$$

$$= \frac{1}{Z} \max_{x_1} \cdots \max_{x_N} \left[\psi_{1,2}(x_1,x_2) \cdots \psi_{N-1,N}(x_{N-1},x_N)\right]$$

$$= \frac{1}{Z} \max_{x_1} \left[\max_{x_2} \left[\psi_{1,2}(x_1,x_2) \left[\cdots \max_{x_N} \psi_{N-1,N}(x_{N-1},x_N)\right] \cdots \right]\right]$$

$\max_{x_n} \psi_{N-1,N}(x_{N-1},x_N)$ is a function only of x_{N-1} and so on.

The Max-Product Algorithm
Over tree-structured factor graphs

- The general idea: maximize with respect to as few variables as possible by splitting the graph into components.

$$\max_{\mathbf{x}} p(\mathbf{x}) = \max_{x_n} \prod_{f_s \in \mathrm{ne}(x_n)} \max_{X_s} f_s(x_n, X_s)$$

- Note that

$$\ln\left(\max_{\mathbf{x}} p(\mathbf{x})\right) = \max_{\mathbf{x}} \ln p(\mathbf{x}).$$

The Max-Product Algorithm
Over tree-structured factor graphs

- Initialization (leaf nodes)

$$\mu_{x\to f}(x) = 0 \qquad \mu_{f\to x}(x) = \ln f(x)$$

- Recursion

$$\begin{aligned}
\mu_{f\to x}(x) &= \max_{x_1,\ldots,x_M}\left[\ln f(x, x_1, \ldots, x_M) + \sum_{m\in\mathrm{ne}(f_s)\setminus x} \mu_{x_m\to f}(x_m)\right] \\
\phi(x) &= \arg\max_{x_1,\ldots,x_M}\left[\ln f(x, x_1, \ldots, x_M) + \sum_{m\in\mathrm{ne}(f_s)\setminus x} \mu_{x_m\to f}(x_m)\right] \\
\mu_{x\to f}(x) &= \sum_{l\in\mathrm{ne}(x)\setminus f} \mu_{f_l\to x}(x)
\end{aligned}$$

The Max-Product Algorithm
Over tree-structured factor graphs

- Termination (root node)

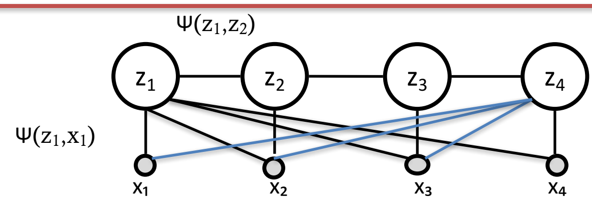
$$p^{\max} = \max_x \left[\sum_{s \in \mathrm{ne}(x)} \mu_{f_s \to x}(x) \right]$$

$$x^{\max} = \arg\max_x \left[\sum_{s \in \mathrm{ne}(x)} \mu_{f_s \to x}(x) \right]$$

- Back-track, for all nodes i with l factor nodes to the root (l=0)

$$\mathbf{x}_l^{\max} = \phi(x_{i,l-1}^{\max})$$

Conditional Random Fields



Some arcs omitted for clarity

The conditional probability is a product of potential functions $p(\boldsymbol{Z}|\boldsymbol{X}) = \frac{1}{Z} \prod_{i \in [N]} \psi_i(z_i, x_1, \dots, x_N)\, \psi_{i,i-1}(z_i, z_{i-1})$

Assuming that factors are linear functions of features of inputs

$$\psi(\mathbf{x}, \mathbf{z}) = \exp\left\{ \sum_k \theta_k f_k(\mathbf{x}, \mathbf{z}) \right\}$$

f_k is a feature function.
E.g. f_k = 1 if x is the base pair "G"

The conditional log likelihood of all examples

$$\log p(z \mid x) = \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{k=1}^{K} \theta_k f_k(z_{it}, z_{it-1}, x_{it}) - \sum_{i=1}^{N} \log Z(x_i)$$

Learning a CRF

- What are the parameters of our model?
  - The θ values in the potential functions

$$\psi(\mathbf{x},\mathbf{z}) = \exp\left\{\sum_k \theta_k f_k(\mathbf{x},\mathbf{z})\right\}$$

- What is the objective for learning?
  The probability of the data given the model
  Note that both **X** and **Z** must be part of data.
- How do we compute model probabilities efficiently?
  Sum Product Algorithm (Forward-Backward)
  Max Product Algorithm (Viterbi decoding)

Regularization

- Recall for logistic regression (discriminative training) maximum likelihood over-fit the data
- Solution: regularization

$$\log p(z \mid x) = \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{k=1}^{K}\theta_k f_k(z_{it}, z_{it-1}, x_{it}) - \sum_{i=1}^{N}\log Z(x_i) - \sum_{k=1}^{K}\frac{\theta_k^2}{2\sigma^2}$$

  - Gaussian prior ($\mu=0$, $\Sigma=\sigma^2 I$)

Training a CRF

- The conditional log likelihood is convex
  - Take the derivative and solve for θ

$$\frac{\partial L}{\partial \theta_k} = \sum_{i=1}^{N}\sum_{t=11}^{T} f_k(z_{it}, z_{it-1}, x_{it}) - \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{z,z'} f_k(z, z', x_{it})\, p(z, z' \mid x_{it}) - \sum_{k=1}^{K}\frac{\theta_k}{\sigma^2}$$

  - The derivative is 0 when
    - The last term (regularizer) is 0
    - The first term and the second term cancel each other
      - First term: the expected value for f_k under the empirical distribution (from the data)
      - Second term: expectation for f_k given model distribution

Why CRFs perform better than HMM ?
*with enough data

HMMs require
- Assumptions of causation / generative story
- Independence assumptions for observations

– These aren't problems for CRFs!
- Can allow arbitrary dependencies. Condition on the whole sequence x. Transition can depend on x and z
- Recall:
  - Generative models limit the features
  - Discriminative models can have any types of features

Generative/Discriminative pairs

- A generative and discriminative parametric model family that can represent the same set of conditional probability distributions
- (Naïve Bayes, Logistic Regression) and (HMM, CRF)

- HMM is a Naïve Bayes classifier at each node
- CRF is a Logistic Regression classifier at each node