MARKOV MODELS

$P(X_{1:T}) = P(X_1)\prod_{t=2}^{T} P(X_t \mid X_{t-1})$ • STATIONARY CHAIN → PARAMETER TYING. TRANSITION MATRIX = A^N

VERILY USED IN LANGUAGE MODELS: $P(X_t = u)$ UNIGRAM STATISTICS, $P(X_t = u \mid X_{t-1} = v)$ BIGRAM MODELS, .. N-GRAM MLES ARE NORMALIZED COUNTS. TO HANDLE MISSING N-GRAMS IN TRAINING EITHER ADD DATA OR ADD-ONE SMOOTHING

DELETED INTERPOLATION: $A_{vu} = (1-\lambda)\hat{f}_{vu} + \lambda f_u$, CAN DO WITH ADAPTIVE WEIGHTS TOP NLG MODEL: INTERPOLATED KNESER-NEY

OUT OF VOCABULARY WORDS: USE SYMBOL 'UNK' AND ASSIGN SOME PROBABILITY MASS TO IT; USE DIRICHLET PROCESS (INFINITE STATE SPACE)

STATIONARY DISTRIBUTION: INITIAL ROW STATE π_0, $\pi_1 = \pi_0 A$, $\pi = \pi A$, BALANCE EQUATIONS. SOLVE $A^T v = v$, $\pi = v^T$, EIGENVECTOR WITH $\lambda = 1$

- IRREDUCIBLE: SINGLY CONNECTED
- APERIODIC: IF RETURN TO ANY STATE DOESN'T OCCUR AT MULTIPLES OF TIME STEPS LARGER THAN 1
- RECURRENT: YOU WILL RETURN TO EVERY STATE WITH PROBABILITY 1
- NON-NULL: EXPECTED TIME TO RETURN TO STATE IS FINITE
- ERGODIC: ~~APERIODIC~~ APERIODIC, RECURRENT, NON-NULL

• EVERY IRREDUCIBLE ERGODIC MC HAS LIMITING DISTRIBUTION π, STATIONARY AND UNIQUE,

• DETAILED BALANCE EQUATIONS: IF π SATISFIES. DBE, π IS STATIONARY. $\sum_i \pi_i A_{ij} = \sum_i \pi_j A_{ji} = \pi_j \sum_i A_{ji} = \pi_j$

HIDDEN MARKOV MODELS

DISCRETE TIME, DISCRETE STATE MC WITH HIDDEN STATES + OBSERVATION MODEL.

$$P(z_{1:T}, x_{1:T}) = P(z_{1:T})P(x_{1:T} \mid z_{1:T}) = P(z_1)\prod_{t=2}^{T} P(z_t \mid z_{t-1}) \cdot \prod_{t=1}^{T} P(x_t \mid z_t)$$

• DISCRETE OBS → OBS. MATRIX

• CONTINUOUS OBS → ~~GAUSSIAN~~ CONDITIONAL GAUSSIAN, $P(x_t \mid z_t = k, \theta) = N(x_t \mid \mu_k, \Sigma_k)$

• BLACK-BOX DENSITY MODELS ON SEQUENCES. LONG-RANGE DEPENDENCIES. NO MARKOV PROPERTY FOR OBSERVATIONS.

• TIME-SERIES PREDICTIONS

• GENERATIVE CLASSIFIERS

• APPLICATIONS: AUTO SPEECH RECOGNITION. x_t IS SPEECH RECOGNITION. z_t IS WORD SPOKEN • ACTIVITY RECOGNITION • PART OF SPEECH TAGGING • GENE FINDING • PROTEIN SEQUENCE ALIGNMENT

HMM INFERENCE

INFER HIDDEN STATE SEQUENCE ASSUMING PARAMS ARE KNOWN. HIDDEN STATE → BELIEF STATE

TYPES OF INFERENCE:

- FILTERING: $P(z_t \mid x_{1:t})$ ONLINE, AS DATA STREAMS IN. REDUCES NOISE BETTER THAN JUST $P(z_t \mid x_t)$. BAYES RULE IN SEQUENCE.
- SMOOTHING: $P(z_t \mid x_{1:T})$ OFFLINE, GIVEN ALL EVIDENCE. 'PAST GIVEN FUTURE'
- FIXED-LAG SMOOTHING: $P(z_{t-\ell} \mid x_{1:t})$, $\ell > 0$ LAG. BETTER THAN FILTERING BUT DELAY.
- PREDICTION: 'FUTURE GIVEN PAST' $P(z_{t+h} \mid x_{1:t})$, h IS PREDICTION HORIZON. EX. $P(z_{t+2} \mid x_{1:t}) = \sum_{z_{t+1}} \sum_{z_t} P(z_{t+2} \mid z_{t+1}) P(z_{t+1} \mid z_t) P(z_t \mid x_{1:t})$

  PWR UP TRANSITION MATRIX AND APPLY TO OUR BELIEF STATE.

  → PREDICT OBSERVATIONS: $P(x_{t+h} \mid x_{1:t}) = \sum_{z_{t+h}} P(x_{t+h} \mid z_{t+h}) P(z_{t+h} \mid x_{1:t})$ POSTERIOR PREDICTIVE DENSITY

- MAP $\operatorname{ARGMAX}_{z_{1:T}} P(z_{1:T} \mid x_{1:T})$ AKA VITERBI DECODING, MOST PROBABLE STATE SEQUENCE
- POSTERIOR SAMPLES: IF 71 PLAUSIBLE INTERPRETATION OF DATA, SAMPLE $z_{1:T} \sim P(z_{1:T} \mid x_{1:T})$
- PROBABILITY OF EVIDENCE: $P(x_{1:T}) = \sum_z P(z_{1:T} \mid x_{1:T})$, CLASSIFY SEQUENCES, MODEL-BASED CLUSTERING.

FORWARDS ALGORITHM

RECURSIVELY COMPUTE FILTERED MARGINALS $p(z_t|x_{1:t})$; ONLINE INFERENCE. EXPLOIT CI OF MARKOV STORY

- **PREDICTION:** ONE-STEP AHEAD PREDICTIVE DENSITY. NEW PRIOR FOR t. $p(z_t=j|x_{1:t-1}) = \sum_i p(z_t=j|z_{t-1}=i)\, p(z_{t-1}=i|x_{1:t-1})$
- **UPDATE:** ABSORB OBSERVATION FOR t WITH BAYES $\alpha_t(j) = p(z_t=j|x_{1:t}) = \frac{1}{Z_t} p(x_t|z_t=j) p(z_t=j|x_{1:t-1})$. $Z_t = \sum_j p(z_t=j|x_{1:t-1}) p(x_t|z_t=j)$ NORM. CONSTANT
  - LOG PROB OF EVIDENCE: $\log p(x_{1:T}|\theta) = \sum_t \log Z_t$
  - MATRIX-VECTOR FORM: $\alpha_t \propto \psi_t \odot (\Psi^T \alpha_{t-1})$. α BELIEF STATE. $\psi_t = p(x_t|z_t=j)$ LOCAL EVIDENCE. Ψ TRANSITION MATRIX. $\odot$ HADAMARD PRODUCT, EL.WISE MULTIPLICATION

FORWARD/BACKWARDS ALGORITHM

SMOOTHED MARGINALS, OFFLINE INFERENCE

$\alpha_t(j) = p(z_t=j|x_{1:t})$ BELIEF STATE AS BEFORE. $\beta_t(j) = p(x_{t+1:T}|z_t=j)$ CONDITIONAL LIKELIHOOD OF FUTURE EVIDENCE, COMPUTED BACKWARDS.

- $\beta_{t-1}(i) = \sum_j \beta_t(j) \psi_t(j) \psi(i,j) \rightarrow \beta_{t-1} = \Psi(\psi_t \odot \beta_t)$
- $\gamma_t(j) = \alpha_t(j)\beta_t(j)$ SMOOTHED POSTERIOR MARGINAL
- IF USING EM TO ESTIMATE TRANSITION MATRIX, NEED TO COMPUTE EXPECTATION OF $i \rightarrow j$ TRANSITIONS. TWO-SLICE MARGINAL.
  $\xi_{t,t+1} \propto \Psi \odot (\alpha_t (\phi_{t+1} \odot \beta_{t+1})^T)$
- COMPLEXITY BASELINE IS $O(K^2T)$. SPEEDUPS VIA MATRIX SPARSITY $O(TN)$. OR STATE SPACE SPARSITY $O(TN \log K)$. SPACE BOTTLENECK TOO.

VITERBI ALGORITHM

MOST PROBABLE SEQUENCE OF HIDDEN STATES IN CHAIN GRAPHICAL MODEL. $z^* = \text{ARGMAX}\, p(z_{1:T}|x_{1:T})$ SHORTEST PATH THROUGH TRELLIS DIAGRAM
- NODES ARE STATES AT EACH STEP
- ARCS ARE LOG PROBABILITIES

PATH WEIGHT: $\log \pi_1(z_1) + \log \phi_1(z_1) + \sum_{t=2}^{T} [\log \psi(z_{t-1}, z_t) + \log \phi_t(z_t)]$

JOINT MOST PROB SEQUENCE OF STATES ≠ MARGINAL MOST PROB SEQUENCE OF STATES
- JOINT → VITERBI → GLOBALLY CONSISTENT
- $\hat{z} = (\text{ARGMAX}\, p(z_1|x_{1:T}), \ldots)$
- MPM: $(\text{ARGMAX}\, p(z_1|x_{1:T}), \ldots)$ → MORE ROBUST

- DOES NOT JUST REPLACE SUM WITH MAX IN BACKWARDS PASS
- CHAINS OF MOST PROBABLE / LEAST COST PATH WITH BACKTRACKING.
- LOGS FOR NUMERIC/EFFICIENCY REASONS. $\log \delta_t(j) = \max_i \log \delta_{t-1}(i) + \log \psi(i,j) + \log \phi_t(j)$
- COMPLEXITY: $O(K^2T)$ ~~EXP~~ TIME, $O(KT)$ SPACE
- CAN USE DISCRIMINATIVE METHOD TO RANK PATHS AND RESCORE MORE, N-BEST

FORWARD FILTERING/BACKWARDS SAMPLING

SAMPLE PATHS FROM POSTERIOR $z^s_{1:T} \sim p(z_{1:T}|x_{1:T})$. FWD-PASS. COMPUTE 2-SLICED POSTERIORS. COMPUTE CONDITIONALS. SAMPLE.
SHORTCUT: SAMPLE IN BACKPASS.

$$p(z_t|z_{t+1}, x_{1:t}) \propto \psi_{t+1}(j)\psi(i,j)\alpha_t(i)$$

$$z_t^s \sim p(z_t|z_{t+1}^s, x_{1:t}) \qquad p(z_t=i|z_{t+1}=j, x_{1:T}) = \frac{\phi_{t+1}(j)\psi(i,j)\alpha_t(i)}{\alpha_{t+1}(j)}$$

BASIS FOR GIBBS SAMPLING

HMM LEARNING

FOR ESTIMATING $\theta = (\pi, A, B)$, π INITIAL DISTRIBUTION, A TRANSITION MATRIX, B CLASS CONDITIONAL DENSITIES

- **FULLY OBSERVED DATA**

  $z_{1:T}$ OBSERVED IN TRAINING. EASY. π, A CLOSED FORM. B LIKE FITTING GENERATIVE CLASSIFIER. PSEUDOCOUNTS (MULTINOMIAL)/GAUSSIAN

- **NOT FULLY OBSERVED DATA**

  LIKE FITTING MIXTURE MODEL. EM FOR MAP/MLE OR GRADIENT DESCENT. EM IS NOW CALLED BAUM/WELCH

  - **E STEP** COMPUTE DATA LL DERIVS ON EXPECTED COUNTS, COMPUTED W/ F-B ALGORITHM. $\gamma_{i,t}(j) = p(z_t = j \mid x_{1:T}, \theta)$, $\xi_{i,t}(j,k) = p(z_{t-1} = j, z_t = k \mid x_{1:T}, \theta)$ SMOOTHED MARGINAL, EDGE MARGINAL
  - **M STEP** FOR A, π IS JUST NORMALIZING THE EXP. COUNTS
  - INITIALIZE W/ FULLY LABELED DATA, RANDOM INITIALIZE, **VITERBI TRAINING** → APPROX PATH INSTEAD W/ MOST PROBABLE PATH

- **BAYESIAN METHODS** VARIATIONAL BAYES EM, MCMC

- **DISCRIMINATIVE TRAINING:** FOR HMM AS CLASS CONDITIONALS INSIDE GENERATIVE CLASSIFIER. MAXIMIZE CONDITIONAL LIKELIHOOD $\prod_{i=1}^{N} p(y_i \mid x_i, \theta)$ WITH GRADIENT METHODS

HMM MODEL SELECTION

- **HOW MANY STATES?** LIKE CHOOSING NO. OF MIXTURE COMPONENTS. GRID SEARCH WITH X VAL LIKELIHOOD, BIC. MCMC. VARIATIONAL BAYES. INFINITE HMM
- **STATE TRANSITION DIAGRAM TOPOLOGY?** STRUCTURE LEARNING. LEARN SPARSE TRANSITION MATRIX. HEURISTIC METHODS. SPLIT MERGE. INFINITE HMM. MINIMUM ENTROPY PRIOR $p(A_{i,:}) \propto \exp(-H(A_{i,:}))$

HMM VARIANTS

- **VARIABLE DURATION HMM/SEMI-MARKOV**

  PROB OF REMAINING IN STATE i FOR D STEPS $p(T_i = D) \propto \exp(D \log A_{ii})$. GEOMETRIC DISTRIBUTION. UNREALISTIC.

  IN SEMI-MARKOV WE DON'T ONLY CONDITION ON PAST STATE BUT WE ALSO NEED TO KNOW CURRENT REMAINING TIME, THEY MODEL BATCH OF OBSERVATIONS AT ONE TIME. USEFUL LIKELIHOOD FORM. USED IN GENETICS. STATES HAVE DURATION DISTRIBUTIONS. INFERENCE W/ MEGA VARIABLES OR MARGINALIZATION

  - **SEMI MARKOV!** EACH STATE REPLACED WITH N STATES, SAME OUTFLOW PROBABILITY. ANY PATH THROUGH SUBSTATES IS $p^{D-N}(1-p)^N$ NEGATIVE BINOMIAL. p IS SELF LOOP PROB. MANY BEHAVIORS BY ADJUSTING p, D. IS FASTER

- **HIERARCHICAL HMM**

  FOR DOMAINS WITH HIERARCHICAL STRUCTURE. GOOD, EFFICIENT INFERENCE. $O(T)$. A TRANSITION AT LEVEL ℓ IS ALLOWED IFF CHAIN AT LEVEL $\ell - 1$ REACHED ITS END STATE

- **INPUT/OUTPUT HMM**

  HANDLE INPUTS. CONDITIONAL DENSITY $p(y_{1:T}, z_{1:T} \mid u_{1:T}, \theta)$ u_t INPUT/CTRL. TRANSITION MATRIX IS LOGISTIC REGRESSION MODEL W/ PARAMS DEPEND ON PREVIOUS STATE. HIDDEN VERSION OF MAX. ENTROPY MARKOV MODEL.

- **AUTOREGRESSIVE HMM**

  WHERE OBSERVATIONS ARE NOT C.I. GIVEN THE HIDDEN STATE $p(x_t \mid x_{t-1}, z_t = j, \theta) = N(x_t \mid W_j x_{t-1} + \mu_j, \Sigma_j)$ LINEAR MODEL WHERE PARAMS CHOSEN W/ CURRENT HIDDEN STATE. CAN EXTEND BY CONDITIONING ON LAST L OBSERVATIONS. COMBINE CHAIN ON HIDDEN STATES FOR LONG-RANGE DEPENDENCIES AND ONE ON OBSERVATIONS FOR SHORT RANGE.

- **BUNED HMM**

  GENERALIZE AR-HMM BY ALLOWING DEPENDENCY STRUCTURE (OF OBSERVABLE NODES) TO CHANGE ON HIDDEN STATE. BAYESIAN MULTI-NET. MIXTURE OF NETWORKS

- **FACTORIAL HMM**

  DISTRIBUTED REPRESENTATION OF HIDDEN STATE AS SINGLE RANDOM VAR. ~~STATES~~ ~~HAVE~~ $z_{c,t} \in \{0,1\}$ IS cTH BIT OF t^{TH} HIDDEN STATE. EXACT ESTIMATION INTRACTABLE BECAUSE XPLAIN AWAY

- **COUPLED HMM**

  STATE TRANSITIONS DEPEND ON STATES OF NEIGHBORING CHAINS. $p(z_t \mid z_{t-1}) = \prod p(z_{ct} \mid z_{t-1})$ VERY EXPENSIVE $O(CK^4)$ **INFLUENCE MODEL** NO NEIGHBOR RESTRICTIONS, CONVEX COMBINATIONS OF PAIRWISE TRANSITION MATRICES. $O(T(uC)^2)$

- **DYNAMIC BAYESIAN NETWORKS**

  BECAUSE NETWORK REPRESENTATION OF DYNAMICAL SYSTEM. GENERALLY ALL HMM ARE DBN BUT SPECIFICALLY ARE USED FOR DOMAIN SPECIFIC IRREGULAR GRAPHS DEFINE FIRST TIME-SLICE. STRUCTURE BETWEEN TIMESLICES. CPDS. EXACT INFERENCE IS EXPENSIVE.