APPROXIMATE INFERENCE

PROBLEM: COMPUTING $p(h|v)$, IF MULTIPLE LAYERS EXACT INFERENCE REQUIRES EXPONENTIAL TIME, OBSTACLE IS LATENT INTERACTIONS IN GRAPHICAL MODEL OR EXPLAINING AWAY EFFECTS.

INFERENCE AS OPTIMIZATION

EXACT INFERENCE CAN BE INTERPRETED AS OPTIMIZATION PROBLEM. $\log p(v,\theta)$ DIFFICULT → LET'S GET A LOWER BOUND ON IT

- $\log p(v) = L(q) + KL(q\|p) \longrightarrow L(v,q,\theta) = \log p(v,\theta) - D_{KL}(q(h)\|p(h|v,\theta)) \xrightarrow[\text{DO MATH}]{\text{DECOMPOSE KL}} L(v,\theta,q) = \underbrace{E_{h\sim q}\log p(h,v)}_{\text{ENERGY}} + \underbrace{H(q)}_{\text{ENTROPY}}$
- LET'S FIND Q MAXIMIZING L. FREE Q BUT ~~OPTIMAL~~ IMPERFECT OPTIMIZATION OR RESTRICTED Q AND OPTIMAL OPTIMIZATION.

EXPECTATION-MAXIMIZATION

CAN BE FRAMED AS COORDINATE ASCENT MAXIMIZATION FOR L. E STEP WRT Q, M STEP WRT θ.

- DOES EXACT INFERENCE BUT IS 'APPROXIMATE' M STEP PRODUCES 'GAP' BETWEEN L AND $\log p$, BUT LATER E-STEP CLOSES IT FULLY.
- SOMETIMES WE JUMP TO EXACT SOLUTION IN 1 ITERATION

SPARSE CODING / MAP

FRAMING OF SPARSE CODING AS PROBABILISTIC MODEL. **MAP** (DOABLE): $h^* = \arg\max p(h|v)$ → SEEN AS MAXIMIZATION OF L UNDER Q WHEN Q IS DIRAC DISTRIBUTION

- EXTRACTION OF CODES → OPTIMIZATION OF W FOR OPTIMAL RECONSTRUCTION ERROR **EQUIVALENT TO** MAXIMIZING L WRT θ OBTAINED FROM MAP, EM W/ DIRAC POSTERIOR
- WE MAXIMIZE A BOUND ON TRUE LIKELIHOOD USING EXACT MAP INFERENCE

SEQUENCE MODELING

WITH GRAPHICAL MODELS. ALTERNATIVE TO RNNS. • IF MODELS HAVE MARKOV STRUCTURE → STUFF IS EASY, EXACT INFERENCE. IF RECURSION → APPROXIMATE

HMMS

- **MARGINALIZATION** SUM OVER ALL COMPLETE SOURCE TO SINK PATHS OF PRODUCT OF EXP SCORES $m(G) = \sum_{\text{path}} \prod_{a \in \text{path}} e^a$ • ON TRELLIS DIAGRAM → **M IS LIKELIHOOD**
- **INFERENCE** • $\pi(G) = \arg\max_{\text{path} \in G} \sum_{a \in \text{path}} a$ MOST PROBABLE PATH • $v(G) = \max_{\text{path} \in G} \sum_{a \in \text{path}} a$ LOG-SCORE

- LET'S EXPLOIT DYNAMIC PROGRAMMING **MARGINALIZE VIA FWD-BWD:** $\begin{cases} m(G) = \sum m(G^n) \\ m(G^n) = \sum_{n' \in \text{pred}(n)} m(G^{n'}) e^{a_{n,n'}} \end{cases}$
- TRANSITION PROBABILITIES, EMISSION PROBABILITIES ...
- OVERALL LIKELIHOOD $p(x_1 \dots x_n) = \sum_s \prod_t p(x_t|s_t) p(s_t|s_{t-1})$
- LOGSCORES: $a_{mn} = \log p(x_t|s_t=i) + \log p(s_t=i|s_{t-1}=j)$ • GRADIENT OPTIMIZATION WORKS BUT EM IS FASTER
- **TOO MANY STATES** IE N-GRAM MODELS → VITERBI BREAKS DOWN, TOO EXPENSIVE → **DO BEAM SEARCH**

MAP VIA VITERBI: $\begin{cases} v(G) = \max_{n \in \text{final}} v(G^n) \\ v(G^n) = \max_{m \in \text{pred}(n)} v(G^m) + a_{m,n} \end{cases}$ $\begin{cases} N^* \leftarrow \arg\max_{\text{final}} v(G^n) \\ N^* \leftarrow \arg\max_{\text{pred}} v(G^m) + a_{mn} \\ \pi(G) \end{cases}$ KEEP TRACK OF ARGMAXES

- CAN FORMULATE DISCRIMINATIVE LIKELIHOOD TOO → $p(y\dots|x\dots) = \dfrac{p(x\dots|y\dots)\,p(y\dots)}{p(x\dots)}$

CONDITIONAL RANDOM FIELDS

UNDIRECTED MODELS TRAINED TO MAXIMIZE JOINT $p(y|x)$

GRAPH TRANSFORMER: COMPUTATIONAL MODEL TRANSFORMING A WEIGHTED DAG INTO ANOTHER. SET OF IN WEIGHTS → SET OF OUT WEIGHTS. CRF ARE GRAPH TRANSFORMERS

- CRF IS MRF BUT WITH POTENTIALS ARE CONDITIONED, PARAMETERIZED, ON INPUTS • Z WOULD BE INTRACTABLE, BUT MARKOV PROPERTY MAKES IT POSSIBLE TO USE DYNAMIC PROGRAMMING AND TRACT IT
- NO OF SUBSUMS FOR INFERENCE SCALES WITH ORDER
- FOR MAP: SUM-PRODUCT → MAX-SUM

NEURAL NETWORKS AND SEARCH COMBO

ANN + HMM / SEARCH IS OLD IDEA, SPEECH & HANDWRITING RECOGNITION. MARGINALIZATION AND INFERENCE (VIA DYNPROG) FOR TEMPORALLY STRUCTURED OUTPUTS CAN BE APPLIED ALSO WHEN LOGSCORES ARE LEARNED LINEAR FUNCTIONS (LIKE IN ANN) → **GRAPH TRANSFORMER**. EXAMPLE: SEGMENTATION → RECOGNITION → SPELLCHECK → CONTEXT AWARENESS **GRAPH TRANSFORMER NETWORK**

BEAM SEARCH: IE WHEN NO OF NODES GROWS EXPONENTIALLY IN SEQUENCE LENGTH. VITERBI NO GOOD.

- BREAK NODES INTO GROUPS OF COMPARABLE NODES → SAME PATH LENGTH
- SEQUENTIALLY PROCESS GROUPS, KEEPING ONLY A GROUP SUBSET AT STEP T **SUBSET IS BEAM**. BASED ON S_{T-1}. EACH NODE HAS APPROX OF MAX TOTAL LOG-SCORE OF PATH (LOWER BOUND)
- S_t OBTAINED FROM FOLLOWING ARCS IN S_{T-1}, SORT RESULTING GROUPS WRT NEXT STEP ESTIMATION. KEEP THE BEST N. → VITERBI

• BEAM OFTEN LACKS IN DIVERSITY. IS GREEDY WRT t

VARIATIONAL INFERENCE

- THE COMPLICATED DISTRIBUTIONS ARE THE CONDITIONALS $p(v|h)$ POSTERIORS. CORRESPOND TO THE COMPLETE GRAPHS OVER THE HIDDEN UNITS → BRUTE FORCE LOL VARIABLE ELIMINATION NO BENEFIT.
- BOTH $p(v|h)$ AND $E[p]$ ARE MESSED UP
- APPROXIMATE p WITH $q(h)$, SIMPLER.
- IS MAXIMIZATION OF A BOUND

CALCULUS OF VARIATIONS

FUNCTIONAL: FCN OF FCN. MINIMIZATION OF FUNCTIONALS. FUNCTIONAL DERIVATIVES $\frac{\delta}{\delta f(x)}$. $\frac{\delta}{\delta f(x)} \int g(f(x), x)\, dx = \frac{\partial}{\partial y} g(f(x), x)$

EXAMPLE: GAUSSIAN AS MAX-ENTROPY DISTRIBUTION

$H[p] = -E_x \log p(x) = -\int p(x) \log p(x)\, dx$

- WE NEED CONSTRAINTS TO FIX INTEGRAL TO 1, BOUND THE VARIANCE, SPECIFY SHIFT. LAGRANGIAN FUNCTIONAL.

- $L(p) = \lambda_1 \left(\int p(x) dx - 1 \right) + \lambda_2 \left(E[x] - \mu \right) + \lambda_3 \left(E[(x-\mu)^2] - \sigma^2 \right) + H[p] = \dots$

MINIMIZE FCN DERIVATIVE

- $\forall x \; \frac{\delta}{\delta p(x)} L = \lambda_1 + \lambda_2 x + \lambda_3 (x-\mu)^2 - 1 - \log p = 0$

$\rightarrow p(x) = \exp\left(-\lambda_1 - \lambda_2 x + \lambda_3 (x-\mu)^2 + 1 \right) \xrightarrow{\text{SET CONSTRAINTS } \lambda} \lambda_1 = \log \sigma \sqrt{2\pi},\ \lambda_2 = 0,\ \lambda_3 = 1/2\sigma^2 \longrightarrow p(x) = N(x \mid \mu, \sigma^2)$