UNDIRECTED GRAPHICAL MODELS - MARKOV RANDOM FIELDS

CAUSAL MRF: DAG MODEL W/ 2D LATTICE TOPOLOGY. ALSO MARKOV MESH. CI PROPERTIES ARE KINDA FUNKY

→ UNDIRECTED MODEL: MRF, MARKOV NETWORK. MARKOV BLANKET IS NOW JUST SET OF NEIGHBORS

PROS
- SYMMETRIC → BETTER FOR CERTAIN DOMAINS
- DISCRIMINATIVE UGM, CONDITIONAL RANDOM FIELDS, $P(Y|X)$ BETTER THAN DGM.

CONS
- PARAMS LESS EXPLAINABLE, LESS MODULAR
- PARAM ESTIMATION MORE COMPUTATIONALLY EXPENSIVE

CI PROPERTIES

- GLOBAL MARKOV PROPERTY: $X_A \perp_G X_B \mid X_C$ IFF C SEPARATES A, B IN GRAPH. REMOVE C; NO PATHS CONNECT A, B.
- UNDIRECTED LOCAL MARKOV PROPERTY: MARKOV BLANKET OF NODE IS SET OF IMMEDIATE NEIGHBORS
- PAIRWISE MARKOV PROPERTY: TWO NODES ARE CI IF NO DIRECT EDGE BETWEEN THEM.
- $G \rightarrow L \rightarrow P$, ALL PROPERTIES ARE THE SAME. CAN RELY ONLY ON PAIRWISE FOR EASY THINGS
  ($P(x) > 0$)

MORALIZATION

CONVERSION OF DGM TO UGM. CAN'T JUST DROP EDGE DIRECTION TO MAINTAIN CI STRUCTURE. CONNECT NODES THAT HAVE A COMMON CHILD TOO.
→ STILL LOSES SOME INFORMATION. FORM ANCESTRAL GRAPH OF G WRT $U = \{A \cup B \cup C\}$. → REMOVES NODES NOT IN U AND NOT ANCESTORS OF U. MORALIZE ANCESTRAL GRAPH. THEN APPLY SEPARATION RULES FOR UGM. ALTERNATIVE TO D-SEPARATION

UGM vs DGM

NEITHER IS MORE POWERFUL THAN OTHER. DGM AND UGM ARE PERFECT MAPS FOR DIFFERENT SETS OF DISTRIBUTIONS.

DGM: V-STRUCTURE $A \rightarrow C \leftarrow B$. $A \perp B$, $A \not\perp B \mid C$. NO UGM CAN PRECISELY (AND ONLY) REPRESENT THESE TWO STATEMENTS

UGM: 4-CYCLE

UGM AND DGM BOTH PERFECT: ARE CHORDAL / DECOMPOSABLE GRAPHS → IF WE COLLAPSE EACH MAXIMAL CLIQUE, THE GRAPH WILL NOW BE A TREE.

MRF PARAMETERIZATION

NO TOPOLOGICAL ORDERING → NO CAN USE CHAIN RULE FOR $P(Y)$

POTENTIAL FUNCTIONS OR FACTORS: ASSOCIATED WITH EACH MAXIMAL CLIQUE. $\psi_c(y_c|\theta_c)$ CAN BE ANY NON-NEGATIVE CLIQUE

JOINT DISTRIBUTION: PROPORTIONAL TO PRODUCT OF CLIQUE POTENTIALS.

HAMMERSLEY-CLIFFORD THEOREM: A POSITIVE DISTRIBUTION $P(Y) > 0$ SATISFIES CI PROPERTIES OF U-GRAPH G ⟷ CAN BE REPRESENTED AS PRODUCT OF FACTORS, ONE PER MAX-CLIQUE

$$P(y|\theta) = \frac{1}{Z(\theta)} \prod_c \psi_c(y_c|\theta_c), \quad Z(\theta) = \sum_x \prod_c \psi_c(y_c|\theta_c)$$

PARTITION FUNCTION ENSURES SUM TO 1

RELATION TO GIBBS DISTRIBUTION:

GIBBS IS $P(y|\theta) = \frac{1}{Z(\theta)} \cdot \exp\left(-\sum E(y_c|\theta_c)\right)$, $E(y_c) > 0$ IS ENERGY OF VARS IN CLIQUE. IF WE $\psi_c(y_c|\theta_c) = \exp\left(-E(y_c|\theta_c)\right)$, IT BECOMES UGM

ENERGY-BASED MODELS: HIGH PROBABILITY STATES ⟷ LOW ENERGY CONFIGURATIONS. PAIRWISE MRF: EDGE POTENTIALS: $P(y|\theta) \propto \prod \psi_{st}(y_s, y_t)$ VS MAXCLIQUES

MAX ENTROPY / LOGLINEAR FORM OF POTENTIAL FUNCTIONS: $\log \psi_c(y_c) = \phi_c(y_c)^T \theta_c$, $\phi_c(y_c)$ IS FEATURE VECTOR, ONE PER EDGE/CLIQUE.

LOG PROBABILITY IS $\log P(y|\theta) = \sum_c \phi_c(y_c)^T \theta_c - Z(\theta)$. IF WEIGHT PER FEATURE → $\psi_{st}(y_s = u, y_t = v) = \exp\left([\theta_{st}^T \phi_{st}]_{uv}\right) = \exp\left(\theta_{st}(u,v)\right)$ IS $K \times K$ TABULAR POTENTIAL IN LOGLINEAR FORM

ISING MODEL

BORN FOR MODELING MAGNETZ. 2D/3D LATTICE. PAIRWISE CLIQUE POTENTIAL $\psi_{st}(y_s, y_t) = \begin{pmatrix} e^{w_{st}} & e^{-w_{st}} \\ e^{-w_{st}} & e^{w_{st}} \end{pmatrix}$ w_{st} IS COUPLING STRENGTH OF S-T

NO CONNECTION, $W_{st}=0$. W SYMMETRIC. OFTEN $W_{st} = J$ CONSTANT.

- $J > 0$ NEIGHBORING SPINS LIKELY TO BE IN SAME STATE. +/-1. GROUND STATE
- $J < 0$ DIFF FROM NEIGHBORS. FRUSTRATED SYSTEM. TWO MODES. ASSOCIATIVE MARKOV NETWORK

UNNORMALIZED LOG PROB: $\log \tilde{p}(y) = -\sum_{s \sim t} y_s w_{st} y_t = -\frac{1}{2} y^T W y$

- BIAS/EXTERNAL FIELD: $\log \tilde{p}(y) = \frac{1}{2} y^T W y + b^T y$. CAN WRITE IN FORM SIMILAR TO GAUSSIAN GM. MULTIPLE MODES. NO ALL CONSTRAINTS AT SAME TIME.

BUT NP-HARD VS $O(D^3)$

HOPFIELD NETWORKS

IS FULLY CONNECTED ISING MODEL WITH $W = W^T$ SYMMETRIC. LEARNED FROM DATA. ASSOCIATIVE MEMORY / CONTENT ADDRESSABLE MEMORY. USED IN PATTERN COMPLETION TASKS. RETRIEVE EXAMPLE FROM MEMORY GIVEN ONLY A PIECE. NO CAN EXACT INFERENCE → COORD DESCENT WITH ICM. SETS EACH NODE TO MIN ENERGY STATE GIVEN ALL NEIGHBORS. CAN BE INTERPRETED AS RECURRENT NEURAL NET. ITERATIVE CONDITIONAL MODES.

BOLTZMANN MACHINE IS ISING/HOPFIELD GENERALIZATION THAT ADDS HIDDEN NODES.

POTTS MODEL

ISING GENERALIZATION WITH MULTIPLE DISCRETE STATES $y_t = \{1, 2, .., K\}$. $\psi_{st} = \begin{pmatrix} e^J & 0 & 0 \\ 0 & e^J & 0 \\ 0 & 0 & e^J \end{pmatrix}$ PHASE TRANSITIONS: DIFFERENT VALUES OF J INDUCE DIFFERENT CLUSTERING BEHAVIORS. USED IN IMAGE SEGMENTATION AS PRIOR (NEIGHBORING PIXELS → SAME LABEL LIKELY).

EMERGING GRAPHICAL MODEL IS MIX OF DIRECTED AND UNDIRECTED EDGES → CHAIN GRAPH. IS 2D ANALOG OF HMM, PARTIALLY OBSERVED MRF BUT INFERENCE MUCH HARDER.

(LOCAL EVIDENCE, PRIOR)

OK FOR REGULARIZING SUPERVISED PROBLEMS, NOT ACCURATE ENOUGH FOR UNSUPERVISED SEGMENTATION.

GAUSSIAN MRFS

UNDIRECTED GGM $\begin{cases} p(y|\theta) = \prod \psi_{st}(y_s, y_t) \prod \psi_t(y_t) \\ \psi_{st}(y_s, y_t) = \exp\left(-\frac{1}{2} y_s \Lambda_{st} y_t\right) \\ \psi_t(y_t) = \exp\left(-\frac{1}{2} \Lambda_{tt} y_t^2 + \eta_t y_t\right) \end{cases}$

JOINT: $p(y|\theta) = \exp\left[\eta^T y - \frac{1}{2} y^T \Lambda y\right]$ → MVN IN INFORMATION FORM!

$\Lambda_{st} = 0 \iff y_s \perp y_t | y_{st}$

- CORRESPOND TO SPARSE PRECISION MATRICES!

- D-GGM ARE SPARSE FACTORIZATIONS OF COVARIANCE MATRICES
- COMBINE D + U GGM: $p(y_t | y_{t-1}, y_{t-2}, \theta) = N(y_t | A_1 y_{t-1} + A_2 y_{t-2}, \xi)$ VECTOR AUTO-REGRESSIVE PROCESS OF ORDER 2. IE TIME-SERIES ASPECT IS MODELED WITH DAG. ZEROS IN TRANSITION MATRICES ARE NO DIRECTED ARCS FROM y_{t-1}, y_{t-2} INTO y_t. ZEROS IN ξ^{-1} ARE NO UNDIRECTED ARCS IN y_t.
- BIDIRECTED GRAPH: USED FOR SPARSE COVARIANCE MATRICES. NON CONNECTED NODES ARE UNCONDITIONALLY INDEPENDENT. CAN BE CONVERTED INTO A DAG WITH LATENT VARIABLES. EACH BIDIR EDGE BECOMES A HIDDEN, CONFOUNDER VARIABLE
- BIDIR + DIR: DIRECTED MIXED GRAPHICAL MODEL

MARKOV LOGIC NETWORKS

FUSION FIRST ORDER LOGIC AND FORMAL RULES APPROACH. USED UNROLLED UGM. → MARKOV LOGIC NETWORK

- REWRITE RULES IN CNF/CLAUSAL FORM. RESTRICT LANGUAGE TO HORN CLAUSES TO MAKE INFERENCE DECIDABLE
- ATTACH WEIGHT TO RULES, DEFINING CLIQUE POTENTIALS $\psi_c(x_c) = \exp(w_c \phi_c(x_c))$, ϕ LOGICAL EVALUATES CON X.
- CONSTRUCT GROUND NETWORK BY CREATING RANDOM BINARY VARS AND WIRE ACCORDING TO CLAUSES

LEARNING FOR MRF

ML AND MAP ESTIMATION EXPENSIVE SO RARELY DONE

• WITH GRADIENT METHODS (FULLY VISIBLE)

LOGLINEAR FORM $P(y|\theta) = \frac{1}{Z(\theta)} \exp\left(\sum_c \theta_c^T \phi_c(y)\right)$, $\ell(\theta) = \frac{1}{N} \log P(y_i|\theta)$, $\frac{\partial \ell}{\partial \theta_c} = \left[\frac{1}{N}\sum \phi_c(y_i)\right] - E[\phi_c(y)]$ (CLAMPED FROM DATA; UNCLAMPED FROM CONDITIONAL) → INFERENCE ONCE PER GRADIENT STEP. TRAINING IS SLOWER, THAN DGM

THEY ARE IN EXPONENTIAL FAMILY, SO CONVEX WRT θ, GRADIENT.

$\frac{\partial \ell}{\partial \theta_c} = E_{\text{EMP}}[\phi_c(y)] - E_{P(\cdot|\theta)}[\phi_c(y)]$ × FIXED FEATURE VECTOR ACCORDING TO EMPIRICAL DISTRIBUTION − MODEL'S EXPECTATION OF F.V. → MOMENT MATCHING

• WITH MISSING DATA / HIDDEN VARS

$P(y,h|\theta) = \frac{1}{Z(\theta)} \exp\left(\sum \theta_c^T \phi_c(h,y)\right)$ → GRADIENT OF LL IS EXPECTED FEATURES WHERE WE CLAMP y_i AND AVERAGE OVER h

$\frac{\partial \ell}{\partial \theta_c} = \frac{1}{N}\sum_i \left\{E[\phi_c(h,y_i)|\theta] - E[\phi_c(h,y)|\theta]\right\}$. 1ST TERM CLAMPS VISIBLE 2ND TERM LEAVES THEM FREE. TO OBSERVED VALUES. WE MARGINALIZE OVER h_i

• ALTERNATIVE USE EM WITH GRADIENT METHODS AT M STEP

APPROXIMATE METHODS

NO CLOSED FORM SOLUTION → GRADIENT OPTIMIZERS → REQUIRES INFERENCE → INFERENCE INTRACTABLE → LEARNING INTRACTABLE

PSEUDO - LIKELIHOOD

MINIMIZE THAT INSTEAD THAN MLE. $\ell_{PL}(\theta) = \frac{1}{N}\sum_i^N \sum_j^D \log P(y_{ij}|y_{i,-j},\theta)$ IS PRODUCT OF FULL CONDITIONALS/COMPOSITE LIKELIHOOD. FOR GAUSSIAN MRF IS = AS ML FASTER BECAUSE EACH FULL CONDITIONAL ONLY REQUIRE STATES OF SINGLE NODE TO NORMALIZE. NODE GIVEN ALL ITS NEIGHBORS. HARD TO APPLY WHEN HIDDEN VARS. STILL VERY MUCH FASTER

STOCHASTIC MLE

FOR PARTIALLY OBSERVED MODEL: USE MCMC TO GENERATE SAMPLES FOR STOCHASTIC GRADIENT DESCENT. TRICKS TO MAKE MCMC FASTER: START AT PREV VAL AND TAKE A FEW STEPS

FEATURE INDUCTION

UNSUPERVISED WAY TO LEARN FEATURES. START WITH BASE SET THEN GREEDILY ADD BETTER FOUNDONES. LIKE GRAPHICAL MODEL STRUCTURE LEARNING BUT MORE FINE GRAINED.

$P(y) = \frac{1}{Z} \exp(\theta_1 \phi_1(y) + \theta_2 \phi(y))$

ITERATIVE PROPORTIONAL FITTING

PAIRWISE MRF, TABLE POTENTIAL, $\psi_{ST} = \exp(\theta_{ST}^T \cdot \text{STATES})$ FEATURE VECTORS ARE INDICATORS · COUNTS. $P(y_c) = P(y_c|\theta)$ AT OPTIMUM. FOR DECOMPOSABLE GRAPHS $P(y_c|\theta) = \psi_c(y_c)$

ELSE $\psi_c^{TM}(y_c) = \psi_c^t(y_c) \times \frac{P_{EM}(y_c)}{P(y_c|\psi^t)}$ COORDINATE ASCENT. IS FIXED POINT ALGORITHM FOR MOMENT MATCHING CONSTRAINTS. IF DECOMPOSABLE CONVERGES IN SINGLE ITERATION.

CAN BE MADE FASTER → EFFICIENT IPF, OR 'PARALLELIZED' WRT PARAMETERS UPDATES. GENERALIZES TO OTHER FEATURE SCHEMES → ITERATIVE SCALING

DECOMPOSABLE GRAPH: TREE-LIKE GRAPHS WHERE EITHER UGM AND DGM ARE OK.

↳ CLOSED FORM SOLUTIONS

CONDITIONAL RANDOM FIELDS

IS MRF WHERE ALL CLIQUE POTENTIALS ARE CONDITIONED ON INPUT FEATURES, CAN BE SEEN AS **STRUCTURED OUTPUT** EXTENSION OF LOGISTIC REGRESS

ADVANTAGE OF DISCRIMINATIVE VS GENERATIVE. NO NEED TO MODEL STUFF WE ALWAYS SEE. CAN MAKE POTENTIALS DATA-DEPENDANT AND EV. TURN OFF.

DISADVANTAGE: REQUIRED LABELED TRAINING DATA, SLOWER. SAME AS LOGISTIC REGRESSION VS NAIVE BAYES, 'LIKE HMM BUT DISCRIMINATIVE'

- $P(y|x,w) = \frac{1}{Z(x,w)} \prod_c \psi_c(y_c|x,w)$ • **LOG-LINEAR POTENTIALS** $\psi_c(y_c|x,w) = \exp(w_c^T \phi(x,y_c))$. $\phi_c(x,y)$ IS FEATURE VECTOR FROM GLOBAL INPUTS x AND LOCAL LABELS y_c

SOME FEATURES ARE GLOBALS, OTHERS ARE LOCAL ON THE NODE.

MAXIMUM ENTROPY MARKOV MODEL (MEMM): 'REVERSE $y \to x$ ARROWS ON HMM AND PUT x_j ON TOP OF y_i'

$P(y|x,w) = \prod P(y_t|y_{t-1}, x, w)$ STATE TRANSITION PROBABILITIES ARE CONDITIONED ON INPUTS.

LABEL BIAS: LOCAL FEATURES AT t DO NOT INFLUENCE PRIOR STATES. INFORMATION DOESN'T FLOW BACKWARDS. ISSUES WITH IE DISAMBIGUATION OF SENTENCES.

CHAIN-STRUCTURED CRF: UNDIRECTED HMM, PUT x_j ON TOP. NOW STUFF IS GLOBALLY NORMALIZED. LOCAL FACTORS NEED NOT SUM TO 1. NO VALID PROBABILITY UNTIL WE'VE SEEN WHOLE SENTENCE; THEN WE CAN NORMALIZE. → CRF NOT OK FOR ONLINE INFERENCE

APPLICATIONS

HANDWRITING RECOGNITION: LOCAL LETTERS MAY AMBIGUOUS, CONTEXT HELPS. NODE POTENTIALS ARE PROBABILISTIC DISCRIMINATIVE CLASSIFIERS, (NN, RVM), EDGE POTENTIALS ARE BIGRAM MODELS

NOUN PHRASE CHUNKING: PHRASE SEGMENTATION. BIO TAGGING. B,I,O ARE STATES. CONSTRUCTED FEATURES. l_1 REGULARIZATION. FWD-BWD FOR INFERENCE.

NAMED-ENTITY RECOGNITION: SMALL SCALE JOINT EXTRACTION. IE WORDS FOR PLACES, NOUNS, LOCATIONS, ETC... CONSIDERS LONG-RANGE CORRELATIONS BETWEEN WORDS → SKIP-CHAIN CRF

NATURAL-LANGUAGE PARSING: PROBABILISTIC CONTEXT-FREE GRAMMARS (GENERATIVE). PROBABILITY OF SEQUENCE IS SUM OF ALL TREES THAT GENERATE IT. DISCRIMINATIVE VARIANTS ENCODE PROB OF TREE y GIVEN SEQUENCE OF WORDS x. $P(y|x) \propto \exp(w^T \phi(x,y))$

HIERARCHICAL CLASSIFICATION: MULTI-CLASS CLASSIFICATION WITH LABEL TAXONOMY SPECIFYING HIERARCHY OF CLASSES. POSITION IN H. ENCODED VIA BINARY VECTOR. COMBINED W/ FEATURES WITH TENSOR PRODUCT.

PROTEIN SIDE-CHAIN PREDICTION: SKIP-CHAIN MODEL FOR PROTEIN STRUCTURE. PREDICT ANGLES GIVEN AMINO-ACID SEQUENCE, ENERGY FUNCTION TO BE MINIMIZED COMES FROM CHEMICAL-PHYSICAL MODEL

STEREO-VISION: LOW LEVEL DEPTH ESTIMATION GIVEN TWO DIFFERENT IMAGES. CRF. NODE POTENTIAL IS DISPARITY OF MATCHING PIXELS. GAUSSIAN POTENTIALS → FUZZY BORDERS. TRUNCATED GAUSSIAN POTENTIALS → PRESERVE DISCONTINUITIES (EDGES).

METRIC CRF → DISCRETIZE VARIABLES FOR FAST COMPUTATION. POTENTIALS FORM A METRIC

TRAINING

$\frac{\partial l}{\partial w_c} = \frac{1}{N} \sum_i [\phi_c(y_i, x_i) - E[\phi_c(y, x_i)]]$ INFERENCE FOR EVERY TRAINING CASE IN EACH GRADIENT STEP. $O(N)$ TIMES SLOWER THAN MRF.

- PARAMETER TYING TO ENSURE DISTRIBUTIONS OF MANAGEABLE SIZE. • **REGULARIZATION:** TO PREVENT OVERFITTING: $l_{0,1,2}$

$l'(w) = \frac{1}{N} \sum \log P(y_i|x_i,w) - \lambda_1 \|w\|_1 - \lambda_2 \|w\|_2^2$

STRUCTURAL SVM

STRUCTURED OUTPUT CLASSIFIERS LEVERAGING ON FAST MAP SOLVERS.

- $R_{MAP} = -\log p(w) - \sum_{i=1}^{N} \log p(y_i | x_i, w)$ **IDEA!** MINIMIZE POSTERIOR EXPECTED LOSS ON TRAINING SET → BOUND ON EXPECTED LOSS. → SET SPHERICAL GAUSSIAN PRIOR

→ $R_{SVM}(w) = \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{N} \left[\max\{0, 1 + y_i w^T x_i, 0\} \right]$ OPTIMIZES UPPER BOUND ON BAYESIAN OBJECTIVE

ALSO (MAINLY) DERIVED NON PROBABILISTICALLY WITH MARGIN FORMULATION. SAME OBJECTIVE

- QUADRATIC PROGRAMS BUT EXPONENTIALLY MANY CONSTRAINTS. IF LOSS AND FEATURE VECTOR DECOMPOSE WITH GRAPHICAL MODEL CAN REDUCE TO POLYNOMIAL. ELSE CUTTING PLANE OR STOCHASTIC SUBGRADIENT

CUTTING PLANE

ITERATIVELY FIND MOST VIOLATED CONSTRAINT. IF > MARGIN, ADD IT TO SET OF WORKING CONSTRAINT. WITH ALL ARE, WE HAVE A TOLERANCE OF ε. PTIME.

- HAS LOSS AUGMENTED DECODING, IS EFFICIENT
- CAN DO BETTER → LINEAR TIME USING KERNEL TRICK QP IN $O(N)$!!

STRUCTURED PERCEPTRON

ONLINE ↓

$\hat{y} = \text{ARGMAX}\, p(y|x)$ WITH EG VITERBI. IF $\hat{y} \neq y$ UPDATE WEIGHTS WITH $w_{n+1} = w_n + \phi(y, x) - \phi(\hat{y}, x)$

STOCHASTIC SUBGRADIENT DESCENT

PEGASOS

$g(w) = \sum_{i=1}^{N} \phi(x_i, \hat{y}_i) - \phi(x_i, y_i) + 2\lambda w$ $\quad w_{n+1} = w_n - \eta_n g_i(w_n)$ η STEPSIZE. IS GENERALIZATION OF PERCEPTRON

LATENT STRUCTURAL SVM

HAS HIDDEN VARIABLES. CAN 'TURN' CRF INTO SSVM FORM/LATENT SVM BUT OBJECTIVE NOT CONVEX CCCP PROCEDURE: MINIMIZES $f(w) - g(w)$ WHERE f, g CONV.

EXACT INFERENCE FOR GRAPHICAL MODELS

GENERALIZE STUFF LIKE FWD-BWD OR KALMAN TO ARBITRARY DIRECTED/UNDIRECTED GRAPHS

BELIEF PROPAGATION FOR TREES (SUM-PRODUCT ALGORITHM)

GENERALIZES FWD-BWD, AND SUM-PRODUCT ALGORITHM

- **SERIAL PROTOCOL:** ψ_s IS NODE EVIDENCE ψ_{st} IS EDGE POTENTIAL. PICK 'ARBITRARY ROOT'. LEAVES → ROOT + ROOT → LEAVES PHASES.

  L→R: MULTIPLY MSG BY EDGE POTENTIALS. AT NODE BELIEF STATE IS $BEL_t^-(x_t) = p(x_t|V_t^-) = \frac{1}{Z^t}\psi_t(x_t)\prod m_{c\to t}^-(x_t)$

  R→L: ALL INFO PARENT HAS RECEIVED, EXCEPT THAT FROM SPECIFIC CHILD. $m_{t\to s}^+ = p(x_s|V_{ST}^+) = \sum_{x_t}\psi_{st}(x_s,x_T)\frac{BEL_t(x_t)}{m_{s\to t}^-(x_t)}$
  OR WE MULTIPLY ALL BUT ONE MSGS

- T-D MSG CONDITIONAL PROBABILITIES POSTERIOR (DEPEND ON B-U) - MAXIMUM MARGINAL
- SUM PRODUCT: T-D DO NOT DEPEND ON B-U CONDITIONAL LIKELIHOODS

- **PARALLEL PROTOCOL:**
  CAN USE IN NON-TREES. SYSTOLIC ARRAY. EACH NODE ABSORBS MSG FROM ITS NEIGHBORS IN PARALLEL AND THEN SENDS MSG TO ALL NEIGHBORS STILL IN PARALLEL BY MULTIPLYING RECEIVED ONES (MINUS ONE FROM RECIPIENT) BY EDGE POTENTIAL. CONVERGES AFTER $O(G)$ STEPS, GRAPH DIAMETER LINEAR

- **GAUSSIAN BP**
  $p(x|v)$ IS JOINTLY GAUSSIAN. GAUSSIAN PAIRWISE MRF: NODES, EDGE POTENTIALS ARE GAUSSIANS → ALL MSGS AND MARGINALS ARE TOO.
  ALGEBRAIC TRICKS MAKE IT COMPUTE IN $O(D)$ VS $O(D^3)$. $\psi_t(x_t) = \exp\left(-\frac{1}{2}A_{tt}x_t^2 + b_t x_t\right)$

  $$\psi_{st}(x_s, x_t) = \exp\left(-\frac{1}{2}x_s A_{st} x_T\right)$$

- **MAX-PRODUCT BP**
  REPLACE $\sum$ WITH MAX → LOCAL MAP ESTIMATES. ONLY NOT BE GLOBALLY CONSISTENT → USE VITERBI

- **SAMPLING**
  CAN SAMPLE FROM TREE MODEL WITH FWD/BWD SAMPLING ALGS

- **POSTERIORS ON SETS OF VARIABLES**
  COMBINE x_t, x_{t+1} INTO 'MEGA NODE'. COMBINATION Ψ MATRICES.

VARIABLE ELIMINATION ALGORITHM

WORKS ON ANY KIND OF GRAPH

- IF DGM → MORALIZE INTO UGM FIRST
- ASSIGN POTENTIALS TO CPD; ALREADY LOCALLY NORMALIZED. $Z=1$
- ENUMERATING ON ALL POSSIBLE ASSIGNMENTS IS $O(2^N)$ → PUSH SUMS INSIDE PRODUCTS
- CREATE TEMPORARY FACTORS BY MULT, THEN MARGINALIZE SUMMED-OVER VAR. PROCEED ~~LEFT TO~~ RIGHT TO LEFT
- THIS COMPUTES ANY MARGINAL
- IF WE WANT CONDITIONAL → RATIO OF TWO MARGINALS NORMALIZED BY EVIDENCE / CLAMPING VARS BY THEIR OBSERVED VALUES
- VE IS NON-SERIAL DYNAMIC PROGRAMMING.
- CAN APPLY TO OBTAIN MAP VIA $x^* = \arg\max_x \prod_c \psi_c(x_c)$ BY REPLACING SUMS WITH MAX (+ TRACEBACK STEP)
- VE WORKS ON ANY COMMUTATIVE SEMI-RING: SET WITH + AND × OPERATIONS ASSOCIATIVE, COMMUTATIVE, WITH IDENTITY, AND DISTRIBUTIVE LAW HOLDS
- RUNNING TIME IS EXPONENTIAL IN SIZE OF LARGEST FACTOR. ELIMINATION ORDER MATTERS A LOT. BECAUSE OF CLIQUES CREATED. MINIMIZE THE TREEWIDTH $w = \min\max|c|-1$. GENERALLY NP-HARD. CHAINS AND TREES ARE EASY THO. $O(VK^2)$ FOR CHAINS, $O\{\min\{M,N\}\}$ FOR 2D LATTICE
- WEAKNESS: NOT EFFICIENT FOR COMPUTING MULTIPLE QUERIES CONDITIONED ON SAME EVIDENCE VAL FWD-BWD $O(K^2T)$ DOES NOT REUSE/CACHE PREVIOUS MESSAGES

JUNCTION-TREE ALGORITHM

GENERALIZES EFFICIENCY OF BP TO ARBITRARY GRAPHS.

- MOCK-RUN VE, ADDING FILL-IN EDGES → **CHORDAL GRAPH** (EACH UNDIR CYCLE HAS CHORDS CONNECTING NON-ADJACENT COMPONENTS)
- FIND MAXIMAL CLIQUES → EASY FOR CHORDALS
- ARRANGE THE CLIQUES INTO THE **JUNCTION TREE** → **RUNNING INTERSECTION PROPERTY** AT LEAST 1 VAR SHARED BETWEEN ADJACENT NODE
- **BP TO JTREE:** EXACT VALUES OF $P(x_i|v)$ FOR CLIQUE NODES → EXACT NODE/EDGE MARGINALS FOR ORIGINAL MODEL BY MARGINALIZATION

MESSAGE PASSING:

CONCEPTUALLY SAME AS BP

ADJUST POTENTIALS BY MULTIPLYING CLIQUE POTENTIALS AND SEPARATING SET POTENTIALS.

BOTTOM → UP: MARGINALIZING OUT THE SEPARATING SET, ALREADY KNOWS ABOUT IT

TOP → BOTTOM: ALWAYS DIVIDING OR SAVING OUT NODE.

COMPLEXITY: IF NODES DISCRETE W/ K STATES $O(|C| \cdot K^{w+1})$ SPACE AND TIME. FOR GGM WE USE INFORMATION FORM AND IS $O(|C|w^3)$ TIME AND $O(|C|w^2)$ TIME. FOR CHAINS IS SAME AS KALMAN SMOOTHER

GENERALIZATIONS: MAP WITH MAX-PRODUCT. N-MOST PROBABLE CONFIGURATIONS. POSTERIOR SAMPLES. SOLVING CSP/SAT PROBLEMS. LOGICAL REASONING.

FINAL REMARK: STUFF IS EXPONENTIAL IN TREE-WIDTH. APPROXIMATE INFERENCE FTW