DEEP LEARNING

- IS STUFF WITH 71 HIDDEN LAYERS

DEEP GENERATIVE MODELS

USED UNSUPERVISED, TO LABEL DATA FOR FURTHER STAGES.

DIRECTED: DEEP DIRECTED NETWORKS. ALL NODES BINARY + ALL CPD LOGISTIC → **SIGMOID BELIEF NETWORK** INFERENCE INTRACTABLE. POSTERIOR OF HIDDENS IS CORRELA[TED]
FAST MEAN FIELD IS INACCURATE. MCMC SLOW BECAUSE CORRELATIONS

UNDIRECTED: STACK RBM ON TOP OF EACHOTHER. DEEP BOLTZMANN MACHINE. CAN DO EFFICIENT BLOCK (LAYER) GIBBS OR MEAN FIELD. (CI ABOVE | BELOW NODES)
DIFFICULT TO TRAIN DUE TO PARTITION. DO IT GREEDILY.

MIXED: **DEEP BELIEF NETWORKS** DIRECTED EXCEPT AT THE TOP. TOP ACTS AS ASSOCIATIVE MEMORY. REST GENERATES OUTPUT.
WE CAN INFER POSTERIOR EXACTLY LIKE IN RBM. FULLY FACTORIZED. $P(h_1|W_1,v_1)$. THIS ONLY HAPPENS IF PRIOR $P(h_1|W_1)$ IS **COMPLEMENTARY** — PRIOR MULT LIKELIHOOD = FACTORED POSTER[IOR]
RBM ACTS AS SUCH. IF MORE LEVELS OR WEIGHTS NOT TIED → NOT EXACT ANYMORE BUT VARIATIONAL LOWER BOUND. APPROXIMATE INFERENCE
CANNOT INTO TOP-DOWN INFERENCE, ONLY FEEDFORWARD

GREEDY DBN TRAINING

- FIT A RBN
- UNROLL INTO DBN WITH TWO HIDDENS, UNTIE WEIGHTS
- FIT A 2ND RBN WITH ACTIVATION OF h_1 HIDDEN UNITS AS INPUT. → BETTER PRIOR FOR $P(h_1|W_2)$
- RINSE, REPEAT
- **REFINE WEIGHTS WITH BACKFITTING:** DO UPWARDS SAMPLING PASS TO ~~BOTTOM~~ TOP. DO GIBBS IN TOP RBN; CD UPDATE. DOWNWARDS ANCESTRAL SAMPLING (APPROX POSTERIOR)
UPDATE LOGISTIC CPD PARAMS

DEEP NEURAL NETWORKS

- IS MULTILAYER PERCEPTRON, FEEDFORWARD. • DIFFICULT TO TRAIN BECAUSE **VANISHING GRADIENT** AND MANY PLATEAUS. GPU + 2ND ORDER BACKPROP METHODS ARE KEY. BUT TOUGH STILL
→ **IDEA: GENERATIVE PRE-TRAINING:** INIT PARAMS WITH UNSUPERVISED LEARNING. MODEL LEARNS TO MODEL ITS OWN INPUT FEATURE VECTOR,
'DATA INDUCED REGULARIZER', HELPS BACKPROP A LOT, FIND GOOD GENERALIZING LOCAL MINIMA.

- **DEEP AUTOENCODER**

UNSUPERVISED ANN USED FOR DIM. REDUCTION AND FEATURE DISCOVERY. TRAINED TO PREDICT INPUT ITSELF. **BOTTLENECK** IN HIDDEN LAYERS TO PREVENT LEARNING IDENTITY FUNCTION. • LINEAR, SHALLOW AUTOENCODERS ARE EQUIVALENT TO PCA, SAME FIRST K PRINCIPAL COMPONENTS
• DIRECT BACKPROP TRAINING DOES NOT WORK WELL BECAUSE V.G. → TRAIN SOME RBM, USE THEIR WEIGHTS TO INIT AUTOENCODER, FINETUNE W/ BACKPROP

- **STACKED DENOISING AUTOENCODERS**

NO BOTTLENECKS → OTHER TRICKS TO PREVENT LEARNING IDENTITY. • IMPOSE SPARSITY CONSTRAINTS ON HIDDEN ACTIVATIONS
→ **ADD NOISE TO INPUT.** SIMILAR TO APPROX ML TRAINING. • CAN BE STACKED, FINETUNING W BACKPROP. LIKE ANN

APPLICATIONS

- HANDWRITTEN DIGIT CLASSIFICATION. MNIST. DBN. SOFTMAX CLASSIF. **SEMINAL RESULT.**
- VISUALIZATION, FEATURE DISCOVERY WITH DEEP AUTOENCODERS. 2D BOTTLENECK. SEMANTIC TOPIC ANALYSIS. NO LABELS BUT MORE HUMAN-FRIENDLY RESULTS THAN LDA/LSA
- **SEMANTIC HASHING:** BINARY, LOW DIM REPRESENTATION IN AUTOENCODER BOTTLENECK. USE LEARNED REPRESENTATION AS HASH KEYS. WIN.
- **1D CONVNETS – AUDIO:** CONVOLUTION AUTOMATICALLY RESULTS IN PARAMETER TYING. **MAXPOOLING:** LOCAL MAX OVER FILTERED RESPONSE. FOR INVARIANCE AND SPEEDUP.
NOISY-OR CPD TO ALLOW BACKWARDS INFO FLOW.
- **2D CONVNETS – IMAGES:** STRAIGHTFORWARD EXTENSION. HIERARCHICAL FEATURES. WIN. SPLIT UP R,G,B CHANNELS.

RESTRICTED BOLTZMANN MACHINES

IS UGM. • PAIRWISE MRF WITH HIDDEN AND VISIBLE NODES, → INFERENCE INTRACTABLE • **RESTRICTION:** NO CONNECTIONS BETWEEN NODE IN SAME LAYER

• $p(h|v,\theta) = \frac{1}{Z(\theta)} \prod^{R} \prod^{K} \psi_{rk}(v_r, h_k)$ • IS SPECIAL CASE OF **PRODUCT OF EXPERTS**, WORKS BETTER BECAUSE YIELDS SHARPER DISTRIBUTIONS, CONSTRAINTS SATISFIED MORE EASILY. ADDING EXPERTS INSTEAD ONLY MAKES IT BROADER.

• **DISTRIBUTED ENCODING**, MANY UNITS GENERATE OUTPUT. VS LOCALIST ENCODING. • **MAIN DIFF:** HIDDEN VARS ARE CI GIVEN VISIBLES. POSTERIOR FACTORIZES

RBM VS 2 LAYER DGM

$p(h|v,\theta) = \prod_{k} p(h_k|v,\theta)$ → EACH h_k IN PARALLEL LIKE ANN

TYPES OF RBM: DIFFERENT PAIRWISE POTENTIAL FCNS

- **BINARY** BINARY HIDDEN / BINARY VISIBLES. JOINT: $p(v,h|\theta) = \frac{1}{Z(\theta)} \exp(-E(v,h|\theta))$ POSTERIOR: $p(h|v,\theta) = \prod^{K} \mathrm{Ber}(h_k|\mathrm{sigm}(w^T v))$ VARS|DATA
  - $E[h|v,\theta] = \mathrm{sigm}(W^T v)$ • **W** GENERATIVE WEIGHTS
  - DATA|VARS $p(v|h,\theta) = \prod_{r} \mathrm{Ber}(v_r|\mathrm{sigm}(w^T h))$
  - $E[v|h,\theta] = \mathrm{sigm}(Wh)$ • **W^T** RECOGNITION WEIGHTS • HIDDEN NODE K ACTIVATES PROPORTIONALLY TO HOW MUCH V LOOKS LIKE W_k → FF-ANN LIKE BEHAVIOR
- **~~GAUSSIAN~~ CATEGORICAL:** USES 1-OF-C ENCODING. ~~C~~ C NO OF STATES FOREACH v_r. $p(v_r|h,\theta) = \mathrm{Cat}(\cdots)$, $p(h_k=1|c,\theta) = \mathrm{sigm}(\cdots)$
- **GAUSSIAN:** HANDLES REAL-VALUED DATA. $p(v_r|h,\theta) = N(v_r|b_r + \sum_k w_{rk} h_k, 1)$, $p(h_k=1,\theta) = \mathrm{sigm}(\cdots)$
- **HIDDEN GAUSSIANS:** LATENT GAUSSIAN + ~~LATENT~~ VISIBLE GAUSSIAN → UNDIRECTED FACTOR ANALYSIS. SAME AS DIRECTED.
  LATENT GAUSSIAN + CAT OBSERVED → UNDIRECTED CAT PCA. MEH PERFORMANCE.

LEARNING

USE A SGD METHOD. AND ℓ_2 REGULARIZATION. HERE WEIGHT DECAY. **GRADIENT:** $\frac{\partial}{\partial w_{rk}} \ell(\theta) = \frac{1}{N} E[v_r h_k|v,\theta] - E[v_r h_k|\theta]$

CLAMPED PHASE EMPIRICAL. UNCLAMPED PHASE

• MODEL XPECTATIONS = EMPIRICAL XPECTATIONS ↓ LEARNING STOPS

• **EXPECTATIONS:**
  - DO BLOCK GIBBS SAMPLING ON JOINT $p(v|h,\theta)$. WAIT FOR CHAIN BURN-IN.
  - MEAN FIELD WORKS REALLY BAD.
  - **CONTRASTIVE DIVERGENCE:** $\nabla_w \ell \propto E[vh^T|v_i] - E_Q[vh^T]$ SIMILAR TO SGD

    K UP-DOWN GIBBS PASSES INITIALIZED AT DATA VECTOR.

    FANTASY DATA: v' IS ATTEMPT AT RECONSTRUCTION → **AUTOENCODER-LIKE BEHAVIOR**
  - **PERSISTENT CD:** LIKE STOCHASTIC MAXIMUM LIKELIHOOD. INIT WEIGHTS. INIT CHAINS. MF UPDATES. MCMC UPDATES. PARAM UPDATES.
    ↳ v ↳ h

APPLICATIONS

- LANGUAGE MODELING INSTEAD OF LDA, SOFTMAX AT THE END. UNIGRAM MODEL < LDA < RBM.
- COLLABORATIVE FILTERING
- BUILDING BLOCK FOR DEEP MODELS

OBS! WAKE-SLEEP ALGORITHM

- IS GENERAL APPROACH USED BY **CONTRASTIVE DIVERGENCE**, USABLE ON OTHER GENERATIVE MODELS.
- **GENERATIVE + RECOGNITION WEIGHTS**
- $\log p(D|G) \geq \log p(D|G) - KL(Q(h|D,R)||p(h|D,G)) = -F(D,R,G)$ FREE ENERGY. DECREASING FREE ENERGY INCREASES LOWER BOUND → INCREASES LL → **WAKE PHASE** TRAIN G;
- USE **REVERSE KL** FOR SLEEP PHASE: $KL(p(h|D,G)||Q(h|D,g))$ → SAMPLE FROM HIDDENS, UPDATE R;
- **FREE ENERGY ∝ SCORE** $F(D,R,G) = \left(-\sum_h Q(h|D,R)\log p(h|D,G)\right) - \left(-\sum_h Q(h|D,R)\log Q(h|D,R)\right)$