AUTOENCODERS

UNSUPERVISED METHODS LEARNING NONLINEAR PARAMETRIC DATA TRANSFORMATION $x \to h$, AS WELL AS $h \to x$; RELATED TO LINEAR FACTOR MODELS, IE PCA IS LINEAR AUTOENCODER

- AUTOENCODERS ARE NEURAL NETS TRYING TO COPY THEIR INPUT TO THEIR OUTPUT
- GENERAL STRUCTURE: INPUT x, ENCODER f, CODE/REPRESENTATION $h = f(x)$, DECODER g, OUTPUT/RECONSTRUCTION $R = g(f(x))$, LOSS $L = L(x, R)$

ISSUE: HOW TO MAKE THE AE NOT LEARN THE IDENTITY FCN?

- HISTORICALLY: MAKE THE DIMENSION OF h SMALLER THAN THAT OF x, BOTTLENECK • MORE RECENTLY: h EVEN LARGER THAN x (OVERCOMPLETE) BUT OTHER REGULARIZATIONS IN $\to$ BECAUSE COMPLEX DISTRIBUTIONS REQUIRE LARGE MODEL CAPACITY | PLACE

– SPARSITY

$|h| > |x|$ BUT MOST ELEMENTS ARE AT OR CLOSE TO 0 $\left(\frac{\partial h_i}{\partial x} \approx 0\right)$. MOTIVATED BY MANIFOLD LEARNING

- SPARSE CODING: UNSUPERVISED LEARNING/FEATURE INFERENCE. ACTUALLY A LFM $\to$ ENCODER IS NOT PARAMETRIC, BUT CODE IS OBTAINED THROUGH ITERATIVE OPTIMIZATION. RELATED ALSO TO GRAPHICAL MODELS. $h^* = f(x) = \underset{h}{\operatorname{argmin}}\, L(g(h), x) + \lambda \Omega(h)$. L_1 PENALTY OR OTHERS.
- PREDICTIVE SPARSE DECOMPOSITION: COMBINES OPTIMIZED + PARAMETRIC ENCODER
- SPARSE AUTOENCODERS STANDARD AUTOENCODER + SPARSITY PENALTY. L_1, STUDENT-T PENALTY $\sum \log(1 + \overbrace{a^2 h^2}^{\text{IS T DIST}})$, K-L PENALTY $-\sum (t \log h_i + (1-t)\log(1-h_i))$
- CONTRACTIVE AUTOENCODERS PENALIZE $\left\|\frac{\partial h}{\partial x}\right\|_F^2$; SUM OF SQUARED NORMS OF $\frac{\partial h_i}{\partial x}$. ENCOURAGES CONTRACTIVE (SMALL IN ALL DIRECTIONS) MAPPING

– ROBUSTNESS

INJECT NOISE IN INPUT OR HIDDEN LAYERS, ASK TO RECONSTRUCT CLEAN INPUT. • CONNECTION BETWEEN CONTRACTIVE AND DENOISING AUTOENCODERS. $\to$ SINCE x AND $x + \epsilon$ ARE TO YIELD SAME OUTPUT $\to$ INSENSITIVITY TO CHANGES IN ALL DIRECTIONS ϵ; SMALL DERIVATIVES.

- REGULARIZED REPRESENTATIONS CAN BE UNDERSTOOD AS PUTTING A PRIOR ON $h \to -\log P(h)$. IS A DATA-DEPENDENT PRIOR, ON ACTUAL SAMPLED VALUES FROM POSTERIOR/ENCODER, IN TURN BASED ON INPUT x. INDIRECT PARAM REGULARIZATION
- DEPTH BOTH ENCODER AND DECODER BENEFIT FROM THE ADDITION OF DEPTH. UNIV. APPROX THEOREM. MORE LAYERS $\to$ MORE CONSTRAINTS. DEPTH ALSO REDUCES COST OF EVALUATION/TRAINING. BETTER COMPRESSION RATES. TRAINING: GREEDY PRETRAINING OF DEEP AE BY TRAINING STACK OF SHALLOW MODELS
- RECONSTRUCTION RAW RECONSTRUCTION LOSS MAY NOT BE APPROPRIATE, IE x IS DISCRETE OR NOT WELL APPROXED BY GAUSSIAN $\to$ DEFINE L AS NEGATIVE LOG LIKELIH OVER TARGET RANDOM VARS. $\to$ OUTPUT IS NOW PROBABILITY OF RECONSTRUCTION $P(x|h)$; MODELING UNCERTAINTY AS WELL AS EXPECTATION
  - $\to$ ENCODING DISTRIBUTION $Q(h|x)$ $\to$ DECODING DISTRIBUTION $P(x|h)$ • h IS NOW A LATENT VARIABLE
  - IN SIMPLE CASES (GAUSSIAN, BERNOULLI) P FACTORIZES: $P(x|h) = \prod_i P(x_i|h)$ • RBM CAN BE SEEN AS AUTOENCODER WHERE $P(x|h) = Q(h|x)$. UNIQUE JOINT WITH P, Q AS CONDITIONALS, SPECIAL CASE

LINEAR FACTOR MODELS

CAN BE SEEN AS AUTOENCODERS WHERE h GENERATES x VIA LINEAR TRANSFORMATION • $h \sim P(h)$, $x = Wh + b + \epsilon$. ϵ NOISE GAUSSIAN, DIAGONAL. DIFFERENT LFM HAVE DIFFERENT FORMS FOR PRIOR AND NOISE. FACTOR ANALYSIS: PRIOR IS $h \sim N(0, I)$ UNIT VARIANCE MVN, x_i ASSUMED CONDITIONALLY INDEPENDENT GIVEN h, NOISE COMING FROM DIAGONAL $\underset{\text{GAUSS}}{\text{COV MATRIX}}$ $\to$ x IS STILL A MVN $x \sim N(b, WW^T + \psi)$ PPCA: CONDITIONAL VARIANCES NOW EQUAL $x \sim N(b, WW^T + \sigma^2 I)$, COVARIANCE IS NOW MOSTLY CAPTURED BY h_i ITSELF UP TO SMALL RESIDUAL RECONSTRUCTION ERROR σ^2. $\sigma^2 \to 0$ WE GET NORMAL PCA. REDUCTION TO $\perp$ PROJECTION

ICA: ALSO A LFM. NONGAUSSIAN PROJECTIONS/PRIORS, BUT STILL FACTORIZED $P(h) = \prod P(h_i)$. BECAUSE IF GAUSSIAN WE CAN'T IDENTIFY | RECOVER | DISENTANGLE FACTORS

DOMINANT FORM OF NONGAUSSIANITY IN REAL DATA IS DUE TO SPARSITY. $P(h)$ NONPARAMETRIC

SPARSE CODING: LATENT VAR PRIOR IS PARAMETRIC, STILL NONGAUSSIAN • FACTORIZED LAPLACE $P(h) = \prod_i \frac{\lambda}{2} e^{-\lambda|h_i|}$ • FACTORIZED T-STUDENT $P(h) \propto \prod_i \frac{1}{\left(1 + \frac{h_i^2}{\nu}\right)^{\frac{\nu+1}{2}}}$ AND GAUSSIAN RECONSTRUCTION NOISE. NORMALLY DONE WITH MAP INFERENCE $\to$ INDUCES SPARSITY AND INDEPENDENCE OF FACTORS • ENCODER IS APPROX INFERENCE ALGORITHM. *

NLL LOSSES

RECONSTRUCTION ERROR IS $L = -\log P(x|h)$ $\to$ OBJECTIVE $L = -\log P(x|g(f(x)))$ TELLS US WHICH LOSS FCN DEPENDING ON INPUT TYPE. • REAL, UNBOUNDED $\to$ SQUARED ERROR, $P(x|h)$ IS NORMAL

- BIT VECTOR $\to$ CROSSENTROPY LOSS $P(x|h) = \prod P(x_i|h)$ BERNOULLI. DECODER TRAINING $\to$ ESTIMATING $P(x|h)$, AND WE CAN TALK ABOUT IMPLICIT OR EXPLICIT $P(x)$

* SPARSE CODING PROS: MINIMIZES REC. ERROR AND LOGPRIOR BETTER THAN PARAMETRIC. DOES EXPLAINING AWAY, INHIBITING HIDDEN FACTORS

SPARSE CODING CONS: INFERENCE ON $h|x$ CAN BE LONGER THAN PARAMETRIC, RESULTING ENCODER COULD BE NONSMOOTH AND/OR NONLINEAR, MAKES IT FOR DIFFICULT GENERALIZATION

SPARSE AUTOENCODERS

WRT SPARSE CODING, SPARSE AUTOENCODERS HAVE EXPLICIT PARAMETRIC ENCODER. • $L = -\log P(x|g(h)) + \Omega(h)$ SPARSITY PRIOR | REGULARIZER. LAPLACE PRIOR ⟷ L1 PENALTY

• $-\log P(h) = \sum_i \log \lambda/2 + \lambda|h_i| = \text{const} + \Omega(h)$ • STUDENT-T PRIOR ⟷ $\Omega(h) = \sum_i \frac{\nu+1}{2} \log\left(1 + \frac{h_i^2}{\nu}\right)$

• LATER, OTHER NON-PROBABILISTIC SPARSITY PENALTIES SUCCESSFULLY INTRODUCED, IE CROSSENTROPIES OR KLS BETWEEN BERNOULLI $p = h_i$ AND BERNOULLI $p = \rho(0{,}05)$ ARBITRARY THRESHOLD

• INITIALLY SPARSITY WAS CONNECTED TO ENERGY PARTITION FUNCTION GRADIENT, CONSIDERING RECONSTRUCTION ERROR A PROXY FOR ENERGY → REGULARIZER PREVENTS IT FROM BEING LOW EVERYWHERE, JUST ON TRAINING SET

• ACTUAL ZEROS IN h → USE ReLU UNITS AS ENCODER OUTPUTS. PRIOR PUSHED TO 0 → WE CAN CONTROL HOW MANY 0S

• REGULARIZER CORRESPONDS TO LOG PRIOR ON LATENT VARS | REPRESENTATION, ~~AND NOT THE DATA~~. PREFERENCE OVER FUNCTIONS OF DATA, NOT OF THE PARAMETERS
DATA-DEPENDENT REGULARIZER

PREDICTIVE SPARSE DECOMPOSITION

COMBINES SPARSE CODING AND PARAMETRIC ENCODER. REPRESENTATION IS FREE VARIABLE. • $L = \underset{h}{\operatorname{argmin}} \left(\|x - g(h)\|^2 + \lambda|h|_1 + \gamma\|h - f(x)\|^2 \right)$

SPARSE CODING + CRITERION OPTIMIZING SPARSE h (AFTER INFERENCE) TO BE CLOSE TO ENCODER OUTPUT $f(x)$

• 2 g, f UPDATED TOWARDS MINIMIZATION. • VARIATIONAL / EM-LIKE APPROACH

• 1 ITERATIVE OPTIMIZATION PERFORMED ON ENCODER OUTPUT → FEW ITERATIONS GRADIENT DESCENT → h OPTIMIZED

• f IS PARAMETRIC APPROXIMATION TO NON-PARAMETRIC SPARSE CODING 'ENCODER'

• CAN STACK GREEDILY

• ITERATIVE OPTIMIZATION ONLY USED AT TRAIN-TIME. LEARNED f CAN BE USED FOR INITIALIZATION OF DEEP MODELS

DENOISING AUTOENCODERS

NO EXPLICIT CONSTRAINTS ON DIMENSION OR SPARSITY OF h. $|h|$ HERE NOT RESTRICTED TO $|x|$. • CORRUPTION PROCESS $C(\tilde{x}|x)$, DAE LEARNS $P(x|\tilde{x})$ RECONSTRUCTION DISTRIBUTION

• WE CAN PERFORM GRADIENT-BASED APPROX. MINIMIZATION ON NLL $-\log(P(x|h))$ WITH BACKPROP MINIBATCH • **DAE DOES SGD** ON $-E_{x \sim \hat{Q}(x)} E_{\tilde{x} \sim C(\tilde{x}|x)} \log P(x|g(f(\tilde{x})))$

• DAE TRAINING MAKES THEM LEARN A VECTOR FIELD $(g(f(x)) - x)$ WHICH ESTIMATES **SCORE**, GRADIENT FIELD $\frac{\partial \log Q(x)}{\partial x}$ → DENOISING IS SCORE MATCHING REGULARIZED
L CONSISTENT ESTIMATOR

• CONNECTION BETWEEN RBM AND AUTOENCODERS • SCORE MATCHING **PROVEN** FOR GAUSSIAN CORRUPTION AND RECONSTRUCTION DISTRIBUTIONS

• $\frac{g(f(x)) - x}{\sigma^2}$ IS CONSISTENT ESTIMATOR FOR $\frac{\partial \log Q(x)}{\partial x}$, GRADIENT OF ENERGY FUNCTION / LOG-DENSITY, LEARNS TO POINT TOWARDS LOWER ENERGY / HIGHER PROB REGIONS

• RECONSTRUCTION ERROR NORM COULD ALSO BE HIGH WHERE PROBABILITY IS LOW. WATCH OUT.

CONTRACTIVE AUTOENCODERS

INTRODUCE AN EXPLICIT REGULARIZER ON CODE $h = f(x)$, ENCOURAGING DERIVATIVES TO BE AS SMALL AS POSSIBLE. $\Omega(h) = \left\| \frac{\partial f(x)}{\partial x} \right\|_F^2$ SQUARED FROBENIUS NORM (SUM SQ. ELEMENTS OF ENCODER JACOBIAN MATRIX • **CAE** LEARNS TO CONTRACT ENCODER. OPPOSED TO RECONSTRUCTION ERROR → DERIVATIVES TINY EXCEPT WHERE NEEDED TO RECONSTRUCT TRAINING EXAMPLES → DIRECTIONS TANGENT TO MANIFOLD WHERE DATA IS CONCENTRATED • PROMOTES STRONG INVARIANCE TO DIRECTION ORTHOGONAL TO MANIFOLD, CHECK SINGULAR VALUE SPECTRUM OF JACOBIAN. • CAE CONCENTRATE / COMPRESS REPRESENTATION SENSITIVITY IN VERY LITTLE DIMS, BETTER THAN VANILLA AE
→ SINGULAR VECTORS OF INPUT-CODING MAPPING

ISSUES: • OK FOR SHALLOW, EXPENSIVE FOR DEEP STACKS → PRETRAIN SINGLE CAEs AND STACK. • CONTRACTION PENALTY ON f USELESS IF g COMPENSATES | EXPANDS → TIE THE WEIGHTS
→ $W_g = W_f^T$ OR SOMETHING