FEEDFORWARD DEEP NETWORKS

SHALLOW MLP

$f_\theta(x) = b + V \cdot \text{sigm}(c + Wx)$ W/ L_2 REG: $J(\theta) = \frac{1}{N}\sum^{\tilde{}} \|y - (b + V\,\text{sigm}(c + Wx))\|^2 + \lambda\|\omega\|^2$

- TRAINED WITH SGD w/ momentum
- WARPS SPACE NONLINEARLY → DECISION BOUNDARY BECOMES LINEAR

NONLINEARITIES

- SIGM
- TANH
- ReLU $\phi(a) = \max(0, a)$
- SOFTMAX $\phi(a) = e^{a_i}/\sum_j e^{a_j}$, $\sum_i \phi(a_i) = 1$, $\forall \phi(a_i) > 0$
- RBF template matching, kernel $h_i = \exp(-\|w_i - x\|^2/\sigma^2)$
- SOFTPLUS $\phi(a) = \zeta(a) = \log(1 + e^{a})$
- HARD TANH $\phi(a) = \max(-1, \min(1, a))$
- ABSOLUTE VALUE RECTIFIER $\phi(a) = |a|$
- MAXOUT $h_i = \max_i(b_i + W_i \cdot x)$ rectifies w/ weights, filters

LOSSES

- SQUARE ERROR → CONDITIONAL EXPECTATION, MEAN
- ABSOLUTE VALUE → CONDITIONAL MEDIAN
- FOR CLASSIFICATION → BERNOULLI NLL, AKA X-ENTROPY AKA KL DIVERGENCE MINIMIZATION; GRADIENTS GO THROUGH EASILY
- NLLs → LOSSES
- LOSSES WITH NORMALIZATION CONSTANTS/TERMS, PARTITION FCNS
- NLL VS L2 CRITERION
- VARIANCE → ESTIMATED FROM SAMPLES IF NOT FCN OF X
- IF FCN(X) → NO CLOSED FORM, ITERATIVE METHODS
- MIXTURE MODELS.
- MULTIPLE OUTPUTS y: IF FEW ASSUME C.I. ELSE GRAPHICAL MODELS

BACKPROPAGATION

COMPUTATIONALLY OPTIMAL

- CHAIN RULE $\nabla_\theta J(g(\theta)) = \sum_i \frac{\partial J(g(\theta))}{\partial g_i(\theta)} \cdot \frac{\partial g_i(\theta)}{\partial \theta}$
- FORWARD PASS: COMPUTE NET ACTIVATIONS. DO MINIBATCHES W/ EXTRA DIMENSION ON MATRICES WHEN CRUNCHING NUMBERS IN CODE. J IS AVG EXAMPLE COST IN MINIBATCH.
- BACKWARDS PASS: OUTPUT GRADIENT → GRADIENT INTO PRE-NONLINEARITY → GRADIENT ON b AND W → THROUGH OTHER LAYER
- FLOW-GRAPHS: ALLOW EFFICIENT COMPUTATION | PROPAGATION OF GRADIENTS WITH ANY TOPOLOGY. NODES HAVE PARTIAL ORDERING. EXPLOITS DYNAMIC PROGRAMMING.

FWD: INPUTS $u_i \leftarrow x_i$; OTHERS: $a^i \leftarrow (u_j)_{j \in \text{parents}(i)}$, $u_i \leftarrow f_i(a_i)$

- U ARE NUMERICALS

BWD: WE NEED CODE TO COMPUTE PARTIAL DERIVATIVES. MULTIPLE PATHS POSSIBLY. DIRECT/INDIRECT EFFECTS WRT FACTORS U. REUSES ALREADY COMPUTED 'DOWNSTREAM' GRADIENT UPDATES. EFFICIENT FACTORIZATION SUMS

- $\frac{\partial u_N}{\partial u_N} = 1$
- $\frac{\partial u_N}{\partial u_j} \leftarrow \sum_{i: j = \text{parents}(i)} \frac{\partial u_N}{\partial u_i} \cdot \frac{\partial f_i(a_i)}{\partial a_{i,\pi(i,j)}}$
- REDUCED TO MATRIX - MATRIX | MATRIX - VECTOR PRODUCTS, BEST FOR GPU PARALLELIZATION.

- BACKPROP IS AUTOMATIC DIFFERENTIATION. BETTER THAN NUMERICAL BECAUSE ALL DERIVATIVES IN ONE GO. CAN ALSO BE DONE WITH FWD PROP OF DERIVATIVES (BETTER FOR INPUTS < OUTPUTS) → OFTEN IMPLEMENTED WITH • SYMBOLIC DIFFERENTIATION (THEANO)

TORCH: NO SYMBOLIC COMPUTATION. WRITE SPECIALIZED CODE FOR DIFF. OPS

BACKPROP THROUGH RANDOM VARS | PROB. DISTRIBUTIONS: TRANSFORM SO TO HAVE R.V.S WHOSE FCN AREN'T ON MY DESIRED VARIABLE: $z \sim N(\mu, \sigma^2) \to z = \mu + \sigma\eta$, $\eta \sim N(0, 1)$. ON W/ GRADIENT - BASED OPTIMIZATION IF F IS C^1

NOISE AS INPUT IN AUTOENCODERS, OR GENERATIVE NETWORKS

REINFORCE ALGORITHM, ESTIMATOR. J BECOMES CONTINUOUS IN ω WHEN AVG'D OVER POSSIBLE VALUES OF NOISE η

→ MCMC SAMPLED, SGD OPTIMIZED

→ HIGH VARIANCE (MANY SAMPLES NEEDED), VARIANCE REDUCTION METHODS

$$E[J(z)] = \sum_z J(z)P(z) \approx \frac{1}{N}\sum^{\tilde{}}_{z_i \sim P(z)} J(z_i)\frac{\partial \log P(z_i)}{\partial \omega}$$

* UNIVERSAL APPROXIMATION THEOREM ANY FFANN WITH AT LEAST A NONLINEAR HIDDEN LAYER CAN REPRESENT ALL FCNS AND LINEAR OUTPUT LAYER

EMPIRICALLY: DEEPER > WIDER. ASSUME PRIOR OF HIERARCHICAL REPRESENTATION ??

LINEAR PREDICTORS ARE LIMITED. WAT DO? • KERNEL MACHINES • FEATURE ENGINEERING • REPRESENTATION LEARNING → DEEP LEARNING

PIECEWISE LINEAR UNITS ARE AWESOME. EASY FWD/BWD PROPAGATION, FEWER PARAMS. MAXOUT IS GENERAL PIECEWISE LINEAR UNIT. LEAKY | PRELU: SLOPES NEGATIVE, NO NULL GRADIENT. NULL GRADIENT → NO LEARNING. EASIER REGULARIZATION. SIGMOIDS STILL USEFUL WHEN BOUNDED OUTPUT