REGULARIZATION

- ANY COMPONENT OF THE MODEL, TRAINING, OR PREDICTION INCLUDED TO ACCOUNT FOR LIMITATIONS ON TRAINING DATA
- IN DEEP LEARNING, MOST ARE REGULARIZATIONS OF ESTIMATORS $\longrightarrow$ + BIAS - VARIANCE FAVORABLY
- $\log p(\theta|x_1 \dots x_n) = \log p(\theta) + \sum_i \log p(x^i|\theta) + \text{CONST}$

IN BAYESIAN OUTLOOK
- PRIOR $\longleftrightarrow$ REGULARIZATION

- CLASSICAL REGULARIZATION: PARAMETER NORM PENALTIES, IN NN ONLY WEIGHTS, NO BIASES.

L_2 REGULARIZATION

RIDGE, TIKHONOV, WEIGHT DECAY. IN NN DIFFERENT α PER LAYER (OR GLOBAL). • ROTATES PARAMS INTO BASIS OF Q, EIGENVECTORS OF $H = Q\Lambda Q^T$ (DIAGONAL | ORTHOGONAL DECOMP) SHRINKS MORE SMALL EIGENVALS $\gamma = \sum_i \frac{\lambda_i}{\lambda_i + \alpha}$ • MAKES LEARNING ALGO 'PERCEIVE' X AS HAVING HIGHER VARIANCE • GAUSSIAN PRIOR

EFFECTIVE NO. OF PARAMS

L_1 REGULARIZATION

INDUCES SPARSITY $w_i = \text{sign}(w_i^*) \cdot \max\left(|w_i^*| - \frac{\beta}{\gamma_i}, 0\right)$ FOR $w_i \le \frac{\beta}{\gamma_i}$, ELSE LINEARLY SHRINKED • LAPLACE PRIOR

$w_i = 0$

- CAN BE SEEN AS CONSTRAINED OPTIMIZATION WITH CONSTRAINT ON WEIGHTS BUT UNKNOWN CONSTRAINED REGION SIZE. α CONTROLS REGION SIZE. (MINIMUMS, KKT FORMULATION) PENALTIES $\longrightarrow$ EXPLICIT CONSTRAINTS AND REPROJECTION, MORE NUMERICAL STABILITY IN IMPLEMENTATION
- FOR UNDERDETERMINED PROBLEMS WHERE MATRICES TO BE INVERTED ARE SINGULAR, COMPUTING THE PSEUDOINVERSE CAN BE SEEN AS INTRODUCING THE MINIMAL REQUIRED REGULARIZATION TO MAKE THE PROBLEM DETERMINED
- DATASET AUGMENTATION IS REGULARIZATION: MORE DATAPOINTS OBTAINED VIA PREPROCESSING (TRANSLATION, ROTATIONS, CROPPING) OR BY ADDING NOISE TO INPUTS (DENOISING AUTOENCODERS) OR TO HIDDEN UNITS
  - INPUT NOISE INJECTION MAKES SENSE FROM BAYESIAN POV. $\epsilon \sim \mathcal{N}(0, \nu I)$ EQUIVALENT TO REGULARIZATION WITH ADDED TO J $\nu E\left[\|\nabla_x \hat{y}(x)\|^2\right]$ REDUCES SENSITIVITY OF OUTPUT TO SMALL VARIATIONS OF X. LOCAL ROBUSTNESS. FOR LINEAR NETWORKS THIS IS WEIGHT DECAY
  - WEIGHT NOISE INJECTION USEFUL IN RNN. EQUIV TO $J + \eta E_{p(x,y)}\left[\|\nabla_W \hat{y}(x)\|^2\right]$ PUSHES MODEL WHERE WEIGHTS HAVE REL. SMALL INFLUENCE ON OUTPUT | MODEL INSENSITIVE TO VARIATION IN WEIGHTS
- EARLY STOPPING RUN UNTIL VALIDATION ERROR HAS NOT IMPROVED FOR SET TIME VS UNTIL LOCAL MINIMUM. USE NO TRAINING STEPS FOR HYPERPARAMETER. COOL BECAUSE NOT COMPUTATIONALLY INTENSIVE, CAN USE OTHER PROCESSOR. PARAMS EASY TO STORE IN SLOWER MEMORIES. UNOBTRUSIVE WRT TRAINING.
  - WHEN ES COMPLETE CAN USE VALIDATION DATA FOR ADDITIONAL, FINAL TRAINING. • CAN ALSO CONTINUE TRAINING ON VALIDATION DATA UNTIL ERROR FALLS BELOW 1ST TRAINING THRESHOLD
  - COOL WITH SURROGATE LOSS FCNS; USE TRUE LOSS FOR ES
  - ES IS REGULARIZER, INTUITIVELY RESTRICTS OPTIMIZATION TO SMALL VOLUME OF PARAMETER SPACE. MAXIMIZES EFFECTIVE CAPACITY $\eta \cdot \tau$, BOUNDS VOLUME REACHABLE FROM θ_0. SHOWN TO BE EQUIVALENT TO L_2 $\alpha \approx \frac{1}{\eta\tau}$ (η: LR, τ: STEPS)
- PARAMETER TYING AND SHARING WE ASSUME DEPENDENCIES OF PARAMS, CLOSE VALUES. • PARAMETER NORM PENALTY ON WEIGHT VALUES DIFFERENCE $\|w_a - w_b\|_2^2$
- • FORCE WS TO BE EQUAL $\rightarrow$ PARAM SHARING, LESS SPACE IN MEMORY $\longrightarrow$ HEAVILY USED IN CONVNETS (TRANSLATION INVARIANCE)
- SPARSITY CAN SPARSIFY MODEL PARAMETERS OR LEARNED REPRESENTATION (AUTOENCODERS). NORM PENALTY ON REPRESENTATION $\alpha\Omega(h)$. L_1, STUDENT'T PRIOR, K-L PENALTY
- BAGGING/ENSEMBLE METHODS MANY MODELS, VOTING. MODEL AVERAGING IF ERRORS OF DIFFERENT MODELS ARE CORRELATED IS USELESS. ELSE EXPECTED SQ. ERROR IS REDUCED LINEARLY IN THE SIZE OF ENSEMBLE. NOT COOL TO USE IN SCIENTIFIC PAPERS, BUT IT WINS COMPETITIONS. NN BENEFIT FROM MODEL AVERAGING
- DROPOUT MAKES BAGGING PRACTICAL FOR LARGE NETS. INEXPENSIVE APPROXIMATION OF TRAINING/EVALUATING EXPONENTIALLY LARGE BAGGED ENSEMBLE OF NETS
  - TRAINS ENSEMBLE OF ALL SUB-NETWORKS FORMED BY REMOVING UNITS FROM BASE NET (MULTIPLY ITS OWN OUTPUT BY 0)
  - WEIGHT SCALING RULE, RENORMALIZE, ETC...
  - COMPUTATIONALLY CHEAP, CAN BE USED ON MANY TYPES OF MODEL, TRAINED WITH SGD $O(n)$ PER EXAMPLE. TRADEOFF NEED TO INCREASE BASELINE MODEL SIZE; NOT GOOD FOR VERY LARGE DATASETS; NOT GOOD WHEN VERY FEW LABELED EXAMPLES

  FAST DROPOUT: ANALYTICAL APPROXIMATION TO STOCHASTICITY, FASTER CONVERGENCE. DROPCONNECT: ALLOWS DROPPING OF ~~NODE~~ PRODUCT BETWEEN W AND NODE
- MULTI-TASK LEARNING: POOLS EXAMPLES FROM DIFFERENT TASKS. SHARED INTERMEDIATE REPRESENTATIONS. TASK-SPECIFIC AND GENERIC PARAMETERS
- ADVERSARIAL TRAINING: PANDA + NOISE = GIBBON. POSSIBLY BECAUSE OUTPUT FCN IS TOO LINEAR. INTRODUCE ADVERSARIAL EXAMPLES IN TRAINING TO ENCOURAGE NETS TO BE LOCALLY CONSTANT: IMPLICIT INTRODUCTION OF LOCAL SMOOTHNESS PRIOR.