OPTIMIZATION FOR TRAINING

TRAINING PROBLEM: $J^*(\theta) = E_{(x,y)\sim p_{DATA}} L(f(x,\theta), y)$ • ANN TRAINING IS DIFFERENT FROM TRADITIONAL OPTIMIZATION → LOSS IS NOT THE TASK IN AND OF ITSELF

• WE DO NOT HAVE / KNOW TRUE DISTRIBUTION | DATA GENERATING PROCESS BUT A FINITE SET OF SAMPLES FROM IT →

→ **EMPIRICAL RISK MINIMIZATION** $E_{(x,y)\sim \hat{p}_{EMPIRICAL}} L(f(x,\theta), y) = \frac{1}{M}\sum_{M} L(f(x_i,\theta) y_i)$ AND HOPE TO IMPROVE ACTUAL RISK. PRONE TO OVERFITTING (MEMORIZING TRAINING SET)

ALSO PROBLEMS BECAUSE MANY LOSSES, IE 0-1, HAVE NO USEFUL DERIVATIVES TO GUIDE GRADIENT DESCENT

→ RARELY USED FOR DEEP LEARNING

→ **SURROGATE LOSS FCNS** WHEN DERIVATIVES ARE CRAP OR TOO COMPLEX/EXPENSIVE TO CRUNCH • 0-1 LOSS → LOGLIKELIHOOD SURROGATE CAN EVEN LEARN MORE THAN ORIGINAL LOSS. ~~OR~~ ORIGINAL LOSS STILL USEFUL TO OBTAIN EARLY STOPPING

NO GUARANTEES THEY STOP IN LOCAL MINIMA BUT WHEN OVERFITTING STARTS KICKING IN

→ **BATCHES & MINIBATCHES** EXACT GRADIENT ON FULL TRAINING SET IS EXPENSIVE. STDERR: $\hat{\sigma}/\sqrt{n}$ → LESS THAN LINEAR RETURN INCREASES IN USING MOAR SAMPLES

→ USE SAMPLE SUBSET, USES DATA REDUNDANCY TO ITS ADVANTAGE, BATCH → MINIBATCH → ONLINE.

• APPROPRIATE BATCH SIZING MAKES GOOD USE OF MULTICORE SETUPS, POWER OF 2 SIZE FOR GPU SETUPS. • SMALL BATCHES HAVE REGULARIZING EFFECT

• **1ST ORDER METHODS:** BATCH SIZE ~ 100 • **2ND ORDER METHODS:** LARGER, ~ 10.000 ESPECIALLY IF H IS POORLY CONDITIONED

• SHUFFLE TRAINING SET TO DECORRELATE SUCCESSIVE SAMPLES → MAKE MINIBATCHES MORE UNIFORM.

• DIFFERENT MINIBATCH UPDATE CAN BE COMPUTED IN PARALLEL

→ **GENERALIZATION ERROR:** SAMPLE/MINIBATCH GRADIENT IS UNBIASED (BUT NOISY) ESTIMATOR OF EXACT GENERALIZATION ERROR UNTIL END OF 1ST EPOCH. SO IS EQUIVALENT TO SGD. THEN THEY DIVERGE

EXPLOITED FOR ONLINE-ONLY SETTINGS.

ISSUES IN ANN OPTIMIZATION BECAUSE OBJECTIVE IS NONCONVEX

- **ILL-CONDITIONING** OF HESSIAN. 2ND ORDER UPDATE TERM DEPENDS ON EIGVALS OF H AND ALIGNMENT OF EVECS WITH g. WHERE g IS ALIGNED WITH LARGE POSITIVE EIGVALS UPDATE MAY MOVE UPHILL. MONITOR WHERE $g^T g$ AND $g^T H g$. 1ST STAYS SAME, 2ND GROWS FOR VERY SMALL LEARNING RATES
  - USUALLY NEWTON METHOD IS SUFFICIENT TO FIX THE ISSUE, NOT SO MUCH IN NNS
  - ALSO WHEN EIGVALS OF OBJECTIVE FCN ARE CLOSE TO 0, INVERTING H BECOMES NUMERICALLY DIFFICULT AND AFFECTS NO OF EPOCHS FOR CONVERGENCE
- **LOCAL MINIMA** FITNESS LANDSCAPE OF ANNS HAS MANY LOCAL MINIMA BECAUSE OF **MODEL IDENTIFIABILITY**. USUAL **WEIGHT-SPACE SYMMETRY** PROBLEM. AND ALSO WEIGHT SCALING SYMMETRY WHEN LINEAR OR MAXOUT UNITS (EVERY LM IS ON A M×N HYPERBOLA OF EQUIVALENT LM, α, $1/\alpha$) IF NO WEIGHT DECAY

  HOWEVER LM FROM UNIDENTIFIABILITY ALL HAVE SAME COST → NOT A PROBLEM FOR GRADIENT DESCENT

  LM PROBLEMATIC IF HIGH-COST. NOWADAYS NOT BELIEVED TO BE COMMON. CHECK BY PLOTTING GRADIENT NORM OVER TIME
- **SADDLE POINTS & PLATEAUS** SADDLE → H HAS BOTH POSITIVE AND NEGATIVE EIGENVALUES. EXPONENTIALLY MORE COMMON THAN LOCAL MINIMA IN HIGH-DIMENSIONAL SPACES !!

  IT HAPPENS THAT FOR RANDOM FCNS EVALS ARE MORE LIKELY TO BE POSITIVE IN LOW COST REGIONS, HIGH COST CRITICAL POINTS → LIKELY A SADDLE
  - GRADIENT DESCENT SEEMS IMMUNE TO SADDLES • 2ND ORDER METHODS (SEEKING A JUMP TO CRITICAL POINT), NEWTON, CAN FALL IN THEM.
  - PROMISING 2ND ORDER SADDLE-FREE NEWTON → STILL NOT SCALES • PLATEAUS ARE PROBLEMATIC → IN NONCONVEX MAY BE HIGH-VALUE
- **EXPLODING GRADIENT: - CLIFFS** COMMON IN RNN WITH LONG SEQUENCE DEPENDENCIES. GRADIENT DESCENT WORKS WRT INFINITESIMAL MOVES. **GRADIENT CLIPPING!** BOUND SIZE OF UPDATE STEP WHEN GRADIENT MAGNITUDE IS TOO HIGH. KEEP THE DIRECTION.
- **ISSUES IN RNN W/ LONG-TERM DEPENDENCIES**: **EXPLODING / VANISHING GRADIENTS**: COMPOSITION OF GRADIENTS IS PRODUCT OF JACOBIANS THROUGH LAYERS/TIME. MULTIPLYING MANY LEADS TO V. LARGE/SMALL VALUES. STATISTICALLY UNLIKELY IF INDEPENDENT MATRICES. IN RNN J_s ARE RELATED → SIMILAR EIGENVECTORS. BLOWS OR CRASHES
  - FOR RNN TO LEARN, JACOBIANS MUST HAVE DET < 1 → LEADS TO VANISHMENT! LONG TIME TO LEARN, IF AT ALL. MASKED BY SHORT TERM DEPENDENCIES. USUALLY 10-20 STEPS
- **INTRACTABLE GRADIENTS**: SUCK IT UP, USE APPROXIMATIONS AND PRAY

BASIC OPTIMIZATION ALGORITHMS

- **STANDARD (BATCH) GRADIENT DESCENT:** $\theta \leftarrow \theta + \epsilon \nabla_\theta L$ COOL, IMPRESSIVE CONVERGENCE. NO NEED TO REDUCE ϵ OVER TIME IF TRUE BATCH. RARELY USED BECAUSE NOT EXPLOITS STRUCTURE OF PROBLEM
- **SGD** MOST USED IN ANN / DEEP LEARNING: TRUE ONLINE OR W/ MINIBATCHES. $E[\hat{g}] = g$. LEARNING RATE η MUST BE REDUCED IN EPOCHS. STATISTICALLY CONVERGES SLOWER THAN BATCH BUT GIVEN LIMITED COMPUTATIONAL RESOURCES IT INITIALLY CONVERGES MUCH FASTER. ALSO WE ALWAYS CARE ABOUT GEN. ERR. SLOW WHEN GRADIENT IS SMALL
- **MOMENTUM** OPTIMAL WHEN GRADIENT IS CONSISTENT ON CONSECUTIVE MINIBATCHES $v \leftarrow \alpha \cdot v + \eta \nabla_\theta L$, $\theta \leftarrow \theta + v$
- **NESTEROV MOMENTUM:** GRADIENT EVALUATED **AFTER** APPLYING CURRENT VELOCITY. MAKES CONVERGENT FASTER IN BATCH CASE

ADAPTIVE LEARNING RATE ALGORITHMS

LEARNING RATE IS BOUNDING FACTOR FOR CONVERGENCE.

- **DELTA-BAR-DELTA** FOR FULL BATCH ONLY. PER-PARAMETER LOSS DERIVATIVE. IF SAME SIGN INCREASE LR. IF DIFF SIGN DECREASE
- **ADAGRAD** PER PARAM LR SCALED INVERSELY TO SUM OF SQUARED PARTIAL DERIVATIVES OVER ITERATIONS → LARGER PD, RAPID DECREASE IN LR. MORE PROGRESS WHERE SLOPE IS GENTLE. FOR DEEP NETS ACCUMULATION OF SQUARED GRADIENTS FROM START RESULTS IN BAD EFFECTIVE LR, TOO SLOW.

$$\begin{cases} g \leftarrow g + \nabla_\theta L \\ \dots \\ R \leftarrow R + g^2 \\ \Delta\theta = -\frac{\eta}{\sqrt{R}} g \\ \theta \leftarrow \theta + \Delta\theta_t \end{cases}$$

ELEMENT-WISE

- **RMSPROP** ADAGRAD BUT GRADIENT ACCUMULATION CHANGED INTO EXPONENTIALLY WEIGHTED MOVING AVERAGE WITH WINDOW ρ BECAUSE RELATIVE SLOPE OF DERIVATIVES MIGHT CHANGE AS TRAINING PROGRESSES. CAN BE COMBINED W/ NESTEROV. EASY. CURRENTLY IN FASHION. $R \leftarrow \rho R + (1-\rho) g^2$
- **ADAM** SIMILAR TO RMSPROP + MOMENTUM BUT MOMENTUM IS ESTIMATE OF 1ST ORDER MOMENT OF GRADIENT (EMA). ALSO CORRECTIONS TO 1ST AND 2ND ORDER MOMENTS TO ACCOUNT FOR ORIGIN INITIALIZATION. $s \leftarrow \rho_1 s + (1-\rho_1) g$; $R \leftarrow \rho_2 R + (1-\rho_2) g^2$; $\hat{s} \leftarrow s / 1-\rho_1^t$, $\hat{R} \leftarrow R / 1-\rho_2^t$; $\Delta\theta = -\alpha \, {}^{\hat{s}}/_{\sqrt{\hat{R}}+\epsilon} \cdot g$
- **ADADELTA** ALSO TRIES TO FIX ISSUES WITH ADAGRAD. INCORPORATES 2ND ORDER INFORMATION: SQUARE ROOTS OF EMA OF INCREMENT AND PARTIAL DERIVATIVES SQUARES.

ELEMENT-WISE $\Delta\theta = -\frac{\sqrt{s+\epsilon}}{\sqrt{R+\epsilon}} \cdot g$; $s \leftarrow \rho s + (1-\rho) [\Delta\theta]^2$

- **NO CLEAR WINNER!** RMSPROP + ADADELTA ARE ROBUST BUT SGD (MOM), RMSPROP (MOM), ADADELTA, ADAM ARE ALL GOOD.

APPROXIMATE 2ND ORDER METHODS

FOR EMPIRICAL RISK BUT ALSO WORK FOR OBJECTIVES WITH REGULARIZATION TERMS

- **NEWTON'S METHOD** → 2ND ORDER TAYLOR EXPANSION $\theta^* = \theta_0 - [H(J(\theta_0))]^{-1} \nabla_\theta J(\theta)$. JUMPS DIRECTLY TO MINIMUM FOR LOCALLY QUADRATIC FCN. ELSE FIXED POINT UPDATE, [AS LONG AS H POSITIVE DEFINITE] EM-LIKE (INVERT H → UPDATE θ) • **IN DEEP LEARNING!** SADDLE PROBLEM → REGULARIZE H $\theta^* = \theta_0 - [H(f(\theta_0)) + \alpha I]^{-1} \nabla_\theta f(\theta)$ → **LEVENBERG-MARQUARDT** OK AS LONG AS NEGATIVE EIGENVALS ARE CLOSE TO 0. ELSE NUMBERS MAKE IT SAME AS 1ST ORDER
  - COMPUTATIONALLY VERY EXPENSIVE $O(k^3)$!!!
- **CONJUGATE GRADIENTS** AVOIDS HESSIAN INVERSION. NORMAL STEEPEST DESCENT HAS ORTHOGONAL SUCCESSIVE STEPS → ZIGZAGGING → 'UNDOES PROGRESS' MINIMUM ALONG PREVIOUS DIRECTIONS IS NOT PRESERVED. **CG FIXES THIS** $d_t = \nabla_\theta J(\theta) + \beta d_{t-1}$, ADDS BACK SOME OF PREVIOUS SEARCH DIRECTION.
  - $d_t^T H(J) d_{t-1} = 0$ → IS STEEPEST DESCENT IN ϕ SPACE ($\phi = \Lambda^{1/2} Q^T \theta$) 'HESSIAN' SPACE [$H = Q \Lambda Q^T$ ORTHOGONAL Q]
  - CAN COMPUTE β WITHOUT EIGENDECOMPOSITION OF H OR H AT ALL: FLETCHER-REEVES, POLAK-RIBIÈRE FORMULAS ONLY 1ST ORDER)
  - → ONLY REQUIRE k LINE SEARCHES (ITERATIONS) IN k-DIM SPACE **NONLINEAR CG:** OCCASIONALLY DOES A STANDARD LINE SEARCH STEP
  - **IN DL:** STARTS WITH A FEW STANDARD SGD ITERATIONS, THEN DO CONJ. GRADS

$\beta_t = \frac{(g_t - g_{t-1})^T g_t}{g_{t-1}^T g_{t-1}}$

$\rho_t = -g_t + \beta \rho_{t-1}$

LINE SEARCH η^*

$\theta_{t+1} = \theta_t + \eta^* \rho_t$

- **BFGS** APPROXIMATES H^{-1} WITH LOW RANK APPROXIMATION M_t REFINED ITERATIVELY $M_T = M_{T-1} + \dots$. RANK ONE UPDATES $\theta \in R^k$ UPDATE IS $O(k^2)$ $\rho_t = M_t g_t$ LINE SEARCH FOR STEP SIZE $\eta^* = \mathrm{argmin}\, J(\theta_t + \eta \rho_t)$, $\theta_{t+1} = \eta^* \rho_t$. LESS TIME TO REFINE LINE SEARCH STEP BUT STORAGE OF M IS IMPRACTICALLY LARGE FOR TOUGH STUFF
  - **L-BFGS** REPLACES M_{t-1} WITH I, UPDATES $\rho_t = -g_t + b\Delta + a\phi$. $\Delta, \phi = g_t - g_{t-1}$, $\theta_t - \theta_{t-1}$; a, b DEPEND ALSO ON THEM. IF EXACT LINE SEARCHES: MUTUALLY CONJUGATE DIRECTIONS BUT ALSO OK IF APPROX LINE SEARCH

NATURAL GRADIENT METHODS

ATTEMPT TO MAKE OPTIMIZATION INVARIANT TO MODEL PARAMETRIZATION θ. STEEPEST DESCENT DIRECTION IN SPACE OF PROBABILITY DISTRIBUTIONS OUTPUT BY MODEL.

$\underline{\Delta_N}$ = ARGMIN $E_{p_{DATA}}[-\log p_{\theta+\Delta\theta}(x)]$ = $KL(p_\theta(x) \| p_{\theta+\Delta\theta}(x)) = \Delta_{KL}$ $\quad KL(p_\theta \| p_{\theta+\Delta\theta}) = -\frac{1}{2}\Delta\theta^T E_{p_\theta}[\nabla^2 \log p_\theta]\Delta\theta$ $\quad | \; -\nabla^2 \log p_\theta$ = FISHER INFO MATRIX

$L_N(\theta, \Delta\theta) = E_{p_{DATA}}[-\log p_\theta] + E_{p_{DATA}}[-\nabla \log p_\theta]^T + \frac{\lambda}{2}\Delta\theta^T E_{p_\theta}[-\nabla^2 \log p_\theta]\Delta\theta$ $\quad \theta_{t+1} = \theta_t + (E_{p_\theta}[-\nabla^2 \log p_\theta])^{-1} \cdot E_{p_{DATA}}[-\nabla \log p_\theta]$

- SCALES WITH INVERSE OF FISHER INFO MATRIX

EXPECTATIONS TAKEN OVER MODEL DISTRIBUTIONS, NEWTON'S DOES IT OVER DATASET

OPTIMIZATION STRATEGIES

COORDINATE DESCENT: MINIMIZE WRT ONE VAR AT TIME, OR GROUPS (BLOCK COORD. DESCENT). OK WHEN VARS/GROUPS RELATIVELY ISOLATED (IE SPARSE CODING DICTIONARY AND CODES)

INITIALIZATION: DL IS VERY SENSITIVE TO IT, STILL SIMPLE HEURISTICS BECAUSE WE DON'T KNOW ANY BETTER, ALSO OPTIMIZATION VS GENERALIZATION TRADEOFF

- **SYMMETRY BREAKING!** SAME FCN ON SAME INPUTS NEED DIFFERENT INIT PARAMS ELSE THEY'LL EVOLVE THE SAME. NO USE + FWD/BWD NULLSPACES
  → MOTIVATES **RANDOM INITIALIZATION** FROM GAUSSIAN OR UNIFORM. BIAS ARE HEURISTIC CONSTANT.
- IDEALLY LARGE VALUES HELP IN BREAKING SYMMETRY BUT LEAD TO OTHER ISSUE: GRADIENT EXPLOSION, CHAOS, SATURATION
- CONSIDERATION OF OPTIMIZATION VS GENERALIZATION. WE ALSO WANT SMALL PARAMS B/C OF REGULARIZATION.
- **INIT IS EFFECTIVELY AN IMPLICIT PRIOR** WE IMPOSE ON THE MODEL. (CLOSE TO 0, LARGE VALUES...)
- SAMPLING HEURISTICS WRT M INPUTS, N OUTPUTS, ORTHOGONAL MATRICES, SCALE INIT. BY CONSTANT GAIN FACTOR g (HIGHWAY?); PRESERVING NORM.
  ↳ SINGULAR VALUES ARE 1 OVER MANY STEPS
- **SPARSE INITIALIZATION** (K-NON ZERO WEIGHTS) - NOT GOOD ON MAXOUT UNITS
- → FRAME IT AS AN HYPERPARAMETER
- MANUALLY ADJUST SCALE BY LOOKING AT STDDEV OR RANGE OF ACTIVATIONS LAYER-BY-LAYER
- **BIAS INITIALIZATION:** DEPENDS ON WEIGHT. COMMONLY 0. IF GENERATIVE MODELS → MATCH MARGINALS OUTPUTS
  - RELU → SMALL POSITIVE BIAS TO AVOID SATURATION
  - LSTM GATES → 1 TO HAVE THEM INITIALLY OFF
- **VARIANCE/PRECISION;** SET TO 1 OR TO MARGINAL VARIANCE
- **USE OTHER ML METHODS | SUPERVISED TRAINING**

GREEDY SUPERVISED PRE-TRAINING: START BY TRAINING A SIMPLER MODEL, OR ATTEMPT A SIMPLER TASK
- **GREEDY ALGOS + FINE-TUNING** ON JOINT PROBLEM

↳ IT GIVES GUIDANCE TO MIDDLE LAYERS!

- ONE LAYER AT A TIME USING PREV LAYER OUTPUTS AS INPUTS
- USE FIRST AND LAST TO INIT MIDDLE OF AN EVEN DEEPER NET
- TRANSFER LEARNING; TRAIN ON A TASK, ADD LAYERS AND TRY ON ANOTHER TASK

DESIGNING MODELS: MORE IMPORTANT / BETTER TO CHOOSE A MODEL EASY TO OPTIMIZE THAN A POWERFUL OPTIMIZATION ALGO

HISTORICAL TRENDS TOWARDS NETS WITH LINEAR UNITS AND DIFFERENTIABLE ACTIVATIONS BECAUSE OPTIMIZATION GETS EASIER.

- LINEAR PATHS, SKIP CONNECTIONS, OR MULTIPLE TRAINING HEADS TO BOOST GRADIENT SIGNAL TO DISCARD IN APPLICATION.

CONTINUATION METHODS! SERIES OF OBJECTIVE FUNCTIONS IN THE SAME PARAMETERS, INCREASINGLY DIFFICULT. (AND WELL BEHAVED ON θ SPACE) ORIGINALLY DESIGNED TO HELP AVOID LOCAL MINIMA (BESIDES ANN). **BLURRING** → CONVOLVING WITH A GAUSSIAN, BOLTZMANN TEMPERATURE SCHEME. HOPE IS MAKING THE FCN MORE CONVEX. **CURRICULUM LEARNING!** LEARN BASIC TASKS, THEN MOVE TO MORE COMPLEX TASKS REQUIRING MASTERY OF THE BASICS
SUCCESSFUL IN NLP AND VISION TASKS