# DEEP LEARNING

## LINEAR ALGEBRA

**SPAN:** SET OF ALL POSSIBLE LINCOMBS **COLUMN SPACE | RANGE:** SPAN OF COLUMNS

**FROBENIUS NORM:** $\|A\|_F = \sqrt{\sum_{i,j} a^2_{i,j}}$ ; MATRIX ANALOGOUS OF VECTOR $L_2$ NORM

**ORTHOGONALITY:** $AA^T = A^TA = I \longrightarrow A^T = A^{-1}$ • INVERSE CHEAP TO COMPUTE

**EIGENDECOMPOSITION:** $A = V \cdot \mathrm{DIAG}(\lambda) \cdot V^{-1}$ • **FOR REAL SYMMETRIC MATRIXES:** $A = Q\Lambda Q^T$, Q ORTHOGONAL, NOT UNIQUE IF SAME EVALS

- **POSITIVE DEFINITE:** ALL $\lambda > 0 \longrightarrow x^TAx = 0 \longleftarrow x = 0$
- **POSITIVE SEMIDEFINITE** ALL $\lambda \geq 0 \longrightarrow \forall x,\ x^TAx \geq 0$

**SINGULAR VALUE DECOMPOSITION** $A = UDV^T$, U, V ORTHOGONAL

$U$ = LEFT S. VECS = EVECS OF $AA^T$

$V$ = RIGHT S. VECS = EVECS OF $A^TA$

$D$ = DIAG W/ SINGULAR VALUES OF $A$ ; NONZERO S.VALS ARE SQRTS OF EVALS OF $AA^T$, $A^TA$

**PSEUDOINVERSE (MOORE - PENROSE)** $A^+ = \lim_{\alpha \to 0} (A^TA + \alpha I)^{-1} A^T \longrightarrow A^+ = VD^+U^T$

FOR NONSQUARE MATRICES

$D^+$ = PSEUDOINVERSE OF D, RECIPROCAL OF NONZERO ELEMENTS AND TRANSPOSE

- **C > R** ONS P.INV ONE OF MANY SOLUTIONS, WITH MINIMAL EUCLIDEAN NORM
- **R > C** MAY BE NO SOLUTION → x WITH AS CLOSE AS POSSIBLE y WITH MIN EUCL NORM DISTANCE

**TRACE:** $TR(A) = \sum_i a_{i,i} \longrightarrow$ **FROBENIUS NORM:** $\|A\|_F = \sqrt{TR(A^TA)}$ **DETERMINANT:** $DET(A) = \prod_i \lambda_i$

## PROBABILITY & INFORMATION THEORY

**MARGINAL:** $P(x) = \int P(x,y)\,dy$ ; **CONDITIONAL** $P(y=y \mid x=x) = P(y=y, x=x)/P(x=x)$, **CHAIN RULE**, **INDIPENDENCE**, **CONDITIONAL INDEPENDENCE**

**EXPECTATION**, **VARIANCE**, **COVARIANCE**, **ENTROPY**, **K-L DIVERGENCE**

**GAUSSIAN INFO FORM:** $N(x \mid \mu, \beta^{-1}) = \sqrt{\frac{\beta}{2\pi}}\, \mathrm{EXP}\left(-\frac{1}{2}\beta(x-\mu)^2\right)$ **MVN INFO FORM:** $N(x \mid \mu, \beta^{-1}) = \sqrt{\frac{\det(\beta)}{(2\pi)^n}} \cdot \mathrm{EXP}\left(-\frac{1}{2}(x-\mu)^T\beta(x-\mu)\right)$

**LAPLACE:** $P(x \mid \mu, \lambda) = \frac{1}{2\lambda}\, \mathrm{EXP}\left(-\frac{|x-\mu|}{\lambda}\right)$

**SIGMOID:** $\mathrm{SIGM}(x) = \sigma(x) = \frac{1}{1 + \mathrm{EXP}(-x)}$ **SOFTPLUS:** $\zeta(x) = \log(1 + \mathrm{EXP}(x))$

$= \sigma(x) = \frac{\mathrm{EXP}(x)}{\mathrm{EXP}(x) + \mathrm{EXP}(0)}$

$\frac{d\sigma(x)}{dx} = \sigma(x)(1 - \sigma(x))$

# NUMERICAL COMPUTATION

**SOFTMAX** $(x)_i = \dfrac{\exp(x_i)}{\sum_j^n \exp(x_j)}$ . SUBJECT TO BOTH UNDERFLOW AND OVERFLOW → EVALUATE SOFTMAX $(z)$, $z = x - \max_i x_i$

**CONDITIONING:** CONDITION NUMBER = $\max_{ij} \left| \dfrac{\lambda_i}{\lambda_j} \right|$ • WHEN LARGE → INVERSION UNSTABLE ; SLOW CONVERGENCE FOR ITERATIVE ALGORITHMS **GRADIENT DESCENT**

• RATIO OF LARGEST AND SMALLEST EIGENVALUE

**HESSIAN:** IF 2ND DERIVATIVES ARE ~~[illegible]~~ CONTINUOUS → IT IS SYMMETRIC , CAN EIGENVAL DECOMPOSITION. WHERE $\nabla_x f(x) = 0$. **POS DEF** → MINIMUM **NEG DEF** → MAXIMUM **AT LEAST ONE $\lambda > 0$, ONE $\lambda < 0$** → SADDLE

• CONDITION NUMBER INDICATES RELATIVE CURVATURE OF SPACE IN DIMENSIONS, IF BIG GRAD DESCENT PERFORMS POORLY → RATIONALE FOR 2ND ORDER METHODS

**CONSTRAINED OPTIMIZATION:** KKT METHOD , LAGRANGE MULTIPLIERS

# ML BASICS

• PROBLEM TYPES • PERFORMANCE MEASURE • SUP / UNSUP

**GENERALIZATION** **NO FREE LUNCH THEOREM** **REGULARIZATION**: ANY MOD AIMED AT REDUCING TEST ERROR, NOT TRAINING ERROR

**HYPERPARAMS** **VALIDATION PRACTICES** **ESTIMATORS** **BIAS - VARIANCE DECOMPOSITION** **MLE** IN RELATION TO KL , NLL

**BAYESIAN FRAMEWORK** ~~[illegible]~~ $H(f_1 + f_2) \leq H(f_1)$ or $H(f_2)$ **MAP ESTIMATION:** $\theta_{MAP} = \arg\max_\theta p(\theta|x) = \arg\max_\theta \underbrace{\log p(x|\theta)}_{\text{LIKELIHOOD}} + \underbrace{\log p(\theta)}_{\text{PRIOR}}$

**SHOPPING LIST:** LINREG , LOGREG , TREES ; PCA

**ML:** DATASET + COST FCN + MODEL + OPTIMIZATION

**CURSE OF DIMENSIONALITY** $\begin{cases} \text{SMOOTHNESS | LOCAL CONSISTENCY PRIOR } f^*(x) \approx f^*(x+\epsilon) \\ \text{MANIFOLD LEARNING} \\ \text{SPARSITY} \end{cases}$