2 - BACKPROPAGATION

BACKPROPAGATION IS A LINEAR TIME DYNAMIC PROGRAM FOR COMPUT. $\partial s.$; AKA "REVERSE-MODE AUTO. DIFFERENTIATION". GOAL: FIT FUNCTION TO GIVEN DATA BY $\min_\theta \sum_{(x,y)\in D} L(f(x;\theta),y)$ (REQUIRES $\nabla_\theta f$)

BACKPROP CAN COMPUTE A GRADIENT IN THE SAME TIME COMPLEXITY AS COMPUTING f, EXPLOITING THE COMPOSITE NATURE OF COMPLEX FUNCTIONS

$$\frac{\partial y}{\partial x} = \lim_{h\to 0}\frac{f(x+h)-f(x)}{h}; \quad \frac{\partial y}{\partial \bar{x}} = \left[\frac{\partial y}{\partial x_1}, \dots \frac{\partial y}{\partial x_m}\right]; \quad \frac{\partial \bar{y}}{\partial \bar{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \dots & \frac{\partial y_1}{\partial x_m} \\ \dots & & \dots \\ \frac{\partial y_n}{\partial x_1} & \dots & \frac{\partial y_n}{\partial x_m} \end{bmatrix} \quad f: \mathbb{R}^m \to \mathbb{R}^n \Rightarrow \text{JACOBIAN } (n\times m)$$

$$\frac{\partial y_i}{\partial x_j} = \sum_{k=1}^{m} \frac{\partial y_i}{\partial z_k}\cdot\frac{\partial z_k}{\partial x_j}$$

COMPUTATION GRAP: LABELED, ACYCLIC HYPERGRAPH WITH VARIABLES AS NODES AND FUNCTION-LABEL. EDGES; CAN REPRESENT ANY COMPOSITE FUNCTION; GIVEN A SET OF PRIMITIVES AND THEIR DERIVATIVES

BAUER'S FORMULA: $\frac{\partial y_i}{\partial x_j} = \sum_{p\in P(i,j)} \prod_{(k,l)\in p} \frac{\partial z_l}{\partial z_k}$ COMPUTING BAUER WITH DP GOES FROM EXPONENTIAL TO LIN. $P\in P(i,j) \to$ BAUER PATHS IN THE NUMBER OF NODES OF THE COMPUTATION GRAPH

FORWARD PROP (F, x):
$z_i := x_i$ IF $(i \le m)$ ELSE 0
FOR i $m+1:N$
$\quad z_i := g_i(\langle z_{Pa(i)}\rangle)$
RETURN $[z_1, \dots, z_N]$

EDGES ARE TOPOLOGICALLY SORTED: $i<j \Rightarrow z_i < z_j$
$<\cdot>$ IS AN ORDERED SET

BACKPROP (F, x):
$z :=$ FORWARD PROP(F,x)
$\frac{\partial F}{\partial z_i} = 0 \ \forall i$ IN $[1;N]$
FOR i $N:1$ ($\frac{\partial z_j}{\partial z_i}$ = $\frac{\partial z_j}{\partial z_i}$)
$\quad \frac{\partial F}{\partial z_i} = \sum_{i,j\in Pa(j)} \frac{\partial F}{\partial z_j}\cdot\frac{\partial g_j(\langle z_{Pa(j)}\rangle)}{\partial z_i}$
RETURN $\left[\frac{\partial F}{\partial z_1}, \dots \frac{\partial F}{\partial z_N}\right]$

$y^{-x} = \exp(x\log(y))$; SUM OVER PATHS OF GRAPH!

COMPUTING THE HESSIAN OF A m-VALUED FUNCTION TAKES $O(m\cdot$ #NODES GNAP)

BACKPROPAGATION IS A CONSTRUCTIVE THEO. SINCE IT GIVES A THEORETICAL GUARANTEE ON THE RUNTIME OF THE BACWARD PASS AND ALSO AN ALGORITHM TO ACTUALLY CALCULATE IT

3 - LOG-LINEAR MODELS, SOFTMAX

PROBABILITY SPACE $\mathcal{P}$: (SAMPLE SPACE Ω, EVENT SPACE E, PROBABILITY FUNCTION P) $e\in E \mapsto [0,1]$

RANDOM VARIABLE: FUNCTION $\Omega \mapsto T$, TARGET SPACE; $P(x|y) = P(x,y)P(y)^{-1}$; $\mathbb{E}[f(x)] = \sum_x f(x)p(x)$

LOG-LINEAR MODEL: $p(y|x,\theta) = \frac{\exp(\theta\cdot F(x,y))}{\sum_{y'\in Y}\exp(\theta\cdot F(x,y'))}$ | $\log p(y|x,\theta) = \theta\cdot F(x,y) + \text{const.} \to$ LINEAR IN θ

$Z(\theta) \to$ PARTITION FUNCT.

FEATURE ENG: PREPROCESSING + FEATURE DESIGN

MLE ESTIMATION: $\theta_{MLE} = \text{ARGMAX}_{\theta\in\Theta} \sum_{n=1}^{N} \log p(y_n|x_n,\theta)$ Θ_{MLE} CONSISTENT; Θ COMPACT SUBSET

OBSERVED FEATURE COUNTS: DATASET FEATURES

$$\frac{\partial NLL(\theta)}{\partial \theta_k} = -\sum_{i=1}^{n}\left(f_k(x_i,y_i) - \sum_{y\in Y} p(y'|x_i,\theta)\cdot f_k(x_i,y')\right) = -\sum_{i=1}^{n} F_k(x_i,y_i) + \sum_{i=1}^{n}\sum_{y'\in Y} p(y'|x_i,\theta) F_k(x_i,y')$$

$\stackrel{*}{=}$ EXPECTED FEATURE COUNTS: MODEL PREDICTION — EXPECT. MATCHING

$$\text{SOFTMAX}(\bar{h},y,T) = \frac{\exp(h_y/T)}{\sum_{y'\in Y}\exp(h_{y'}/T)}$$

$\mathbb{R}^K \mapsto$ SIMPLEX Δ^{K-1}

T: TEMPERATURE HYPERP. $T\to+\infty \Rightarrow$ UNIFORM CAT. $T\to 0 \Rightarrow$ MAX(.)

LOG. LIN.: DOT PRODUCT FOLLOWED BY SOFTMAX $h_y = \theta\cdot f(x,y)$

$$\frac{\partial \log \text{SOFTMAX}(\bar{h},y)}{\partial h_i} = \delta_{yi} - \text{SOFTMAX}(\bar{h},i)$$

GUMBEL SOFTMAX: CONTINUOUS RELAXATION OF SOFTMAX FROM WHICH WE CAN SAMPLE AND BACKPROPAGATE THROUGH

EXPONENTIAL FAMILY: $p(x|\theta) = \frac{1}{Z(\theta)} h(x)\exp(\theta\cdot\phi(x))$ | $Z(\theta)$: PARTITION FUNCTION; $h(x)$: SUPPORT; θ: PARAMETERS; $\phi(x)$: SUFFICIENT STATISTICS $\to$ ALWAYS FINITE

SCALED EXPONENTIATED DOT PROD.

$$N_{\mu,\sigma} = \frac{1}{\sigma\sqrt{2\pi}}\cdot\exp\left(-\tfrac{1}{2}\left(\tfrac{x-\mu}{\sigma}\right)^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\mu^2(2\sigma^2)^{-1}\right)\cdot 1\cdot\exp\left(\begin{bmatrix}-\frac{1}{2\sigma^2} & \frac{\mu}{\sigma^2}\end{bmatrix}^T\cdot\begin{bmatrix}x & x^2\end{bmatrix}^T\right)$$

CONJUGATE PRIOR: (LIKELIHOOD, PRIOR) PAIR S.T. POSTERIOR = PRIOR; AVOIDS COMPUTING INTRACT. INTEGRAL

EXAMPLE: (CATEGORICAL, DIRICHLET); ENTROPY: $H(p(x)) = -\sum_{x\in X} p(x)\log p(x)$

MAXIMUM ENTROPY PRINCIPLE: THE DISTRIBUTION WHICH BEST REPRESENTS THE CURRENT (PARTIAL) STATE OF KNOWLEDGE IS THE ONE WHICH MAXIMIZES $H(p(x))$; EXPONEN. DISTROS MAXIMIZE ENTROPY IN SPECIFIC COND.

LOG-LINEAR $\longleftrightarrow$ EXP. FAM: $\log p(y|\theta) = \theta^T\phi(y) - \log Z(\theta) + \log h(y)$, REWRITE θ AS FUNCTION OF x: $\theta(x)$

$\log p(y|\theta(x)) = \theta(x)^T\phi(y) - \log Z(\theta) + \log h(y)$; $\theta(x) = [\theta^T(f(x,1)-f(x,C)), \dots, \theta^T(f(x,C-1)) - f(x,C)]^T$

LOG-LINEAR MODELS DEFINE A FUNCTION FROM VECTORS TO A CATEGORICAL DISTRIBUTION $\to$ (IN EXP. FAMILY)

CHI-SQUARE AS EXPONENTIAL: $p(x|k) = \frac{x^{k/2-1}\exp(-x/2)}{2^{k/2}\Gamma(k/2)} = \frac{1}{2^{k/2}\Gamma(k/2)}\exp\left(-\frac{x}{2}\right)\exp\left(\left(\frac{k}{2}-1\right)\log(x)\right)$

$\theta = (\frac{k}{2}-1)$ $Z(k) = 2^{k/2}\Gamma(k/2)$ $Z(\theta) = 2^{\theta+1}\Gamma(\theta+1)$

$h(x) = \exp(-x/2)$ $\phi(x) = \log(x)$

$\hookrightarrow x = \exp(\log(x))$ (UNIFORM NOT IN EXP.)

BINARY LOG-LINEAR MODEL IS LOGISTIC REGRESSION WHEN $\theta' = [\theta_{2n}-\theta_1, \dots, \theta_{2d}-\theta_d] = \theta_1 - \theta_0$

START FROM $P(y=1|x,\theta)$ AND GET TO $\sigma(\theta'\cdot x)$ ALGEBRICALLY | $\theta\in\mathbb{R}^{2d}$, $F(x,0) = [x_1, \dots x_d, 0, \dots, 0]$

HESSIAN OF LL (LOG-LIN) $= \frac{\partial\nabla LL}{\partial\theta^T} = \sum_{i=1}^{n}\sum_{y\in Y} p(y'|x_i,\theta)F(x_i,y')F(x_i,y')^T - p(y'|x_i,\theta)F(x_i,y')\sum_{y''\in Y} p(y''|x_i,\theta)F(x_i,y'')^T$

$= \sum_i^n \mathbb{E}_{Y|X=x_i}[F(x_i,Y)F(x_i,Y)^T] - \sum_i^n (\mathbb{E}_{Y|X=x_i}[F(x_i,Y)]\,\mathbb{E}_{Y|X=x_i}[F(x_i,Y)^T]) = \sum_i^n \text{Cov}(F(x_i,Y))$

4 - MLPs

HYPERPLANE: MULTIDIMENSIONAL GENERALIZATION OF A LINE $a_1x_1 + \dots + a_nx_n = b$

LINEAR SEPARABILITY: PROBABILITY OF A SET OF POINTS WHICH CAN BE SEPARATED BY AN HYPERPLANE

A LINEAR MODEL CAN EXPRESS A NON-LINEAR DECISION BOUNDARY BY ENCODING A SPECIFIC NON-LINEARITY IN ITS FEATURE FUNCTION; YOU NEED TO KNOW THE NON-LINEARITY A PRIORI

MULTI-LAYER PERCEPTRON: FEED-FORWARD FULLY-CONNECTED NN WHICH ALTERNATES LINEAR PROJECTIONS AND NON-LINEAR ACTIVATION FUNCTIONS; $h^N = \sigma^N(W^N\dots\sigma^2(W^2\sigma^1(W^1e(x))))$ $e(x) \to$ ENCODING OF x

MLP + SOFTMAX: $p(y|x) = \text{SOFTMAX}(h^N, y) \in \mathbb{R}^{|Y|}$; $\sigma(x) = \frac{1}{1+e^{-x}}$ $\sigma'(x) = \sigma(x)(1-\sigma(x)) \le 0.25$; $\tanh(x) = 2\sigma(2x)-1$

$\tanh'(x) = 1-\tanh^2(x)$; $\text{RELU}(x) = \max(0,x)$ $\text{RELU}'(x) = \mathbb{I}_{x>0}$; IN AN MLP THE NLL IS NO LONGER CONVEX

MLP WITH AN ARBITRARILY LARGE HIDDEN LAYER AND A NON LINEAR σ ARE UNIVERSAL FUNCTION APPROXIMAT.

PERCEPTRON: SINGLE NODE MLP WITH $T\to 0$, TRAINED WITH SGD (MINIBATCH SIZE OF 1); CAN'T MODEL XOR

WORD PREPROCESSING: TOKENIZATION, STEMMING, STOP WORD REMOVAL; COMBINE ONE-HOT WORD ENCODING IN BAG OF WORDS FORMAT, WITH SENTENCE POOLING OR WITH m-GRAM CONCATENATION; EMB. SIZE GOES FROM $|V|$ TO $|V|^m$

SKIP-GRAM: SENTENCIZE AND TOKENIZE CORPUS THEN BUILD POSITIVE SAMPLES WITH GIVEN CONTEXT WINDOW k

LOG BILINEAR MODEL $p(c|w) = Z(w)^{-1}\exp(e_{mod}(w)\cdot e_{ctx}(c))$ USING DIFFERENT CONTEXT AND CENTER EMBDS.

TO AVOID COMPUTATION OF $Z(w)$ SAMPLE NEGATIVE SAMPLE PAIRS AND TRAIN FOR A CLASSIFICATION LOGISTIC LOSS

OUTPUT: (2) SET OF WORD EMBEDDING; EVALUATION WITH COSINE SIMILARITY OR WORD ANALOGIES (BERT SOTA)

MLP FOR SENTIMENT ANALYSIS: EMB+POOLING+MLP+SOFTMAX; WITH MLP FEATURE ENGINEERING IS AVOIDED BUT ARCHITECTURE ENG. CAN ARISE INSTEAD; RELU HAS DEAD NEURONS IF INPUT IS NEGATIVE

5 - LANGUAGE MODELS, RNNs

STRUCTURED PREDICTION: PREDICTING STRUCTURED OBJECTS IN AN EXPONENTIALLY LARGE SPACE

KLEENE CLOSURE V^*: SET OF ALL FINITE STRINGS OVER SYMBOLS IN A VOCABULARY V.

LANGUAGE MODEL: NORMALIZED PROB. DISTR. OVER THE KLEENE CLOSURE OF A VOCAB; REPRESENTABLE AS A WEIGHTED PREFIX TRIE; $y\in V^* \to p(y) = z^{-1}\cdot\prod_{t=1}^{|y|} Q_{y_t}$; $z = \sum_{y\in V^*}\prod_{t=1}^{|y|} Q_{y_t}$; z CAN BE COMPUTED WITH GLOBAL NORMALIZATION (ARBITRARY WEIGHTS) OR LOCAL NORMALIZATION (CHOOSE θ S.T. $z=1$, $\sum_a p^{CHILD}(a|PARENT) = 1$

TIGHT LANG MODEL: $p(EOS|PARENT) > 0$, ENFORCES FINITE NUMBER OF INFINITELY LONG PATHS; CAN CONDITION ON x

m-GRAM ASSUMPTION: $p(y_t|y_{<t}) \doteq p(y_t|y_{t-1},\dots,y_{t-m+1}) = \frac{\exp(w_{y_t}\cdot h_t)}{\sum_{y'\in V}\exp(w_{y'}\cdot h_t)}$ $w_y\in\mathbb{R}^d \to$ WORD VECTOR $h_t\in\mathbb{R}^d$ m-GRAM CONTEXT VECTOR

BENGIO ET. AL.: NEURAL LANG. MODEL; USES WORD EMBEDS.

TO SOLVE CURSE OF DIMEN. $(|V|^m \to d^m)$; MODEL: GET EMBEDDINGS, COMBINE CONTEXT WITH MLP $\to h_t = F(e(hist))$

USE SOFTMAX TO NORMALIZE $p(y_t|y_{t-1},\dots y_{t-m+1}) \to$ OVER $|V|$, SO SLOWISH; TRAIN BOTH MODEL PARAM AND EMBDS.

RNN: COMBINE TIME-DEP HIDDEN STATE WITH CURRENT TOKEN; IN THEORY HAS INFINITELY LONG CONTEXT

VANILLA: $h_t = \tanh(W[h_{t-1}; e(y_{t-1})])$; LSTM CELL: HAS INPUT, OUTPUT AND FORGET GATES + 2 STATES IS LESS AFFECTED BY VANISHING GRADIENTS DUE TO ADDITIVE UPDATE FUNCTION ON MORE THAN A MATRIX $\Rightarrow$ MORE CTX

GRU CELL: ONLY ONE STATE AND TWO GATES (INPUT AND FORGET) $\Rightarrow$ LESS PARAMETERS

NEURAL RECURRENT MODEL: $h_t = F(h_{t-1}, e(y_{t-1}))$ $p(y_t|y_{<t}) = \exp(w_{y_t}\cdot h_t)\cdot(\sum_{y'\in V}\exp(w_{y'}\cdot h_t))^{-1}$ (FOR $\lambda\to 0$ $\ln(1+\lambda)\sim\lambda$)

BACKPROP THROUGH TIME: UNROLL RNN AND BACKPROP NORMALLY; VANISHING GRADIENTS ARISE SINCE THE SAME SHARED WEIGHTED MATRIX IS MULTIPLIED FOR EVERY STATE $|W|<1$ VANISHING, $|W|>1$ EXPLODING GRAD $\to$ CLIP

RNN ARE PREFERRED OVER MLP FOR NLP SINCE THEY ENCODE A LONGER AND POSITIONAL CONTEXT WHICH IS USEFUL TO ENCODE LONG DISTANCE DEPENDENCIES (E.G. SUBJ.VERB AGREEMENT IN GERMAN SUBORDINATES)

UNIGRAM: $p(\text{HERE YOU ARE EOS}) \doteq p(\text{HERE})\,p(\text{YOU})\,p(\text{ARE})\,p(\text{EOS})$

BIGRAM: $p(\text{BOS HERE YOU EOS}) \doteq p(\text{HERE}|\text{BOS})\,p(\text{YOU}|\text{HERE})\,p(\text{EOS}|\text{YOU}) = \frac{p(\text{BOS HERE})}{p(\text{BOS})}\cdot\frac{p(\text{HERE YOU})}{p(\text{HERE})}\cdot\frac{p(\text{YOU EOS})}{p(\text{YOU})}$

LIDSTONE SMOOTHING: $p(w_m|w_{m-1}) = \frac{\text{COUNT}(w_{m-1},w_m)+\lambda}{\sum_{w'\in V}\text{COUNT}(w_{m-1},w')+|V|\lambda}$ AVOIDS ASSIGNING PROB. $=0$ TO PLAUSIBLE SENTENCES CONTAINING n-GRAMS NOT IN TRAIN. SET

LAPLACE SMTH. $\Rightarrow \lambda = 1$

EXPECTED NUMBER OF BIGRAMS WITH ZERO COUNT FROM VOCAB V AND M TOKENS IN CORPUS $= (1-|V|^{-2})^M$

6 - CONDITIONAL RANDOM FIELDS

PART OF SPEECH: CATEGORY OF WORDS THAT SHOW SIMILAR SYNTACTIC BEHAVIOR $\to p(t|w) = \frac{\exp(\text{SCORE}(t,w))}{\sum_{t'\in T^N}\exp(\text{SCORE}(t,w))}$

CONDITIONAL RANDOM FIELD: CONDITIONAL PROBABILISTIC MODEL FOR SEQUENCE LABELING

SCORE (t,w) CAN BE A DOT PRODUCT LIKE $\theta\cdot F(t,w)$ OR NEURALIZED WITH A NEURAL NETWORK $\to \text{NN}_\theta(t,w)$

NAIVELY COMPUTING Z RUNS IN $O(|T|^N)$ $T \to$ SET OF POS TAGS, $|w| = N$, LENGHT OF INPUT SENTENCE

DEFINE ADDITIVELY DECOMPOSABLE SCORE FUNCTION: $\text{SCORE}(t,w) = \sum_{n=1}^{N}\text{SCORE}(\langle t_{n-1},t_n\rangle, w)$

NORMALIZER: $\bigoplus_{t:N\in T^N}\bigotimes_{n=1}^{N}\exp(\text{SCORE}(\langle t_{n-1},t_n\rangle,w)) = \bigoplus_{t:N-1\in T^{N-1}}\bigotimes_{n=1}^{N-1}\exp(\text{SCORE}(\langle t_{n-1},t_n\rangle,w))\otimes\bigoplus_{t_N\in T}\exp(\text{SCORE}(\langle t_{N-1},t_N\rangle,w)) = \bigoplus_{t_1\in T}\exp(\text{SCORE}(\langle t_0,t_1\rangle,w))\otimes\dots$

$\bigoplus_{t_2\in T}\exp(\text{SCORE}(\langle t_1,t_2\rangle,w))\otimes\dots\otimes\bigoplus_{t_N\in T}\exp(\text{SCORE}(\langle t_{N-1},t_N\rangle,w)) \to$ LINEAR IN N!

VITERBI ALGORITHM: $O(T^2N)$ (3 FOR TRIPLES) THIS ALGORITHM CAN ALSO BE COMPUTED IN A "FORWARD" MANNER, IN THAT CASE

$\gamma(w,t_1) = 1$ $\to$ VITERBI VARIABLE $\gamma(w,t_i) = V_i(t)$ REPRESENTS THE SCORE OF THE BEST TAG SEQUENCE ENDING IN TAG t AT POSITION i

FOR m IN $N-1,\dots 0$:
$\quad \gamma(w,t_m) = \max_{t_{m+1}\in T}\exp(\text{SCORE}(\langle t_m,t_{m+1}\rangle,w))\times\gamma(w,t_{m+1})$

THE ALGO COMPUTES THE BEST SCOR. TAGGING SEQUENCE FOR SENTENCE w

SEMIRING: 5-TUPLE $R = (A,\oplus,\otimes,\bar{0},\bar{1})$ SO THAT 1) $(A,\oplus,\bar{0})$ IS A COMMUTATIVE MONOID 2) $(A,\otimes,\bar{1})$ IS MONOID 3) $\otimes$ DISTRIBUTES OVER $\oplus$ FROM BOTH LEFT AND RIGHT 4) $\bar{0}$ IS AN ANNIHILATOR OF $\otimes$: $\forall a\in A$, $\bar{0}\otimes a = a\otimes\bar{0} = \bar{0}$

PROOF: 1) CHECK IDENTITY OF $\bar{0}$, ASSOCIATIVITY AND COMMUT. OF $\oplus$ 2) CHECK IDENTITY OF $\bar{1}$ AND ASSOCIATIVITY OF $\otimes$ 3) $(a\oplus b)\otimes c = (a\otimes c)\oplus(b\otimes c)$; $c\otimes(b\oplus a) = (c\otimes a)\oplus(c\otimes b)$ 4) PROPERTY ABOVE

		$\oplus$	$\otimes$	$\bar{0}$	$\bar{1}$
BOOLEAN	$\{0,1\}$	$\vee$	$\wedge$	0	1
VITERBI	$[0,1]$	max	$\times$	0	1
INSIDE	$\mathbb{R}^+\cup+\infty$	$+$	$\times$	0	1
REAL	$\mathbb{R}\cup+\infty$	MIN	$+$	$+\infty$	0
TROPICAL	$\mathbb{R}^+\cup+\infty$	MIN	$+$	$+\infty$	0
COUNTING	$\mathbb{N}$	$+$	$\times$	0	1
LOG	$\mathbb{R}\cup-\infty$	$\ln(e^a+e^b)$	$+$	$-\infty$	0
VITERBI (LOG)	$\mathbb{R}\cup-\infty$	max	$+$	$-\infty$	0

THE NORMALIZER CAN BE COMPUTED WITH THE INSIDE SEMIRING (BACWARD/FORWARD ALGORITHM); THE LOG-Z CAN BE COMPUTED WITH THE LOG SEMIRING (USING LOG ON SCORES AS FIRST STEP)

CRF CAN BE TRAINED BY MAXIMIZING LL WITH BACKPROP ON DATASET $\sum_{i=1}^{I}(\text{SCORE}(t^i,w^i) - T\log\sum_{t'\in T^N}\exp(\text{SCORE}(t',w^i)/T)) \to$ SOFTMAX

ANNEALING A CRF WITH $T\to 0$ WE OBTAIN A STRUCTURED PERCEPT.

TRAINABLE WITH SUBGRADIENT DESCENT AND THE PERCEPTRON UPDATE

$-T\log\sum_{t'\in T^N}\exp(\text{SCORE}(t',w^i)/T) \stackrel{T\to 0}{\simeq} -\max_{t'\in T^N}\text{SCORE}(t',w^i)$

THE FORWARD-BACKWARD ALGORITHM CAN BE SEEN AS FORWARD + BACKPROP WHEN SUBSTITUTING THE BACKWARD RECURRENCE WITH BACKPROPAGATION. BOTH BACWARD PASSES USE DYNAMIC PROGRAMMING TO COMPUTE $\nabla\log Z$

VITERBI EXAMPLE: $\max_{t\in T^4}\text{SCORE}(t,$ "THEY CAN CAN FISH"$)$; $V_m(t_m) = \max_{t_{m-1}}(\text{SCORE}(\langle t_{m-1},t_m\rangle,w) + V_{m-1}(t_{m-1}))$

$V_1(t_1) = \text{SCORE}(\langle\square,t_1\rangle,w)$; $V_1(N) = -2-1 = -3$, $V_1(V) = -12$; $V_2(N) = \max(V_1(N)+\text{SCORE}(\langle t_1{=}N,t_2{=}N\rangle), V_1(V)+\text{SCORE}(\langle t_1{=}N,t_2{=}V\rangle)) = \max(-3-3-3, -12-3-1) = -9$; $b_2(N) = N$; $V_2(V) = \max(V_1(N)+\text{SCORE}(\langle t_1{=}N,t_2{=}V\rangle), V_1(V)+\text{SCORE}(\langle t_1{=}V,t_2{=}V\rangle)) = -5$; $b_2(V) = N$ $\to$ BACKPOINTER, THE TAG THAT MAXIMIZED THE SCORE IS N

EMISSION FEATURES: $t\mapsto w$; TRANSITION FEATURES: $t\mapsto t$ $\Rightarrow$ COMPONENTS OF SCORE FUNCTION

7 - CONTEXT-FREE PARSING

SYNTAX: MATHEMATICAL STUDY OF THE STRUCTURE OF SENTENCES; LANGUAGE IS STRUCTURED HIERARCHICALLY AND CAN BE BROKEN DOWN INTO CONSTITUENTS: MULTI-WORD UNIT THAT FUNCTIONS AS A SINGLE UNIT
KINDS OF AMBIGUITY: ATTACHMENT AMBIGUITY, MODIFIER SCOPE; CONSTITUENCY TESTS: PRONOUN SUBSTITUTION, CLEFTING, ANSWER ELLIPSIS; MANY LINGUISTIC PROPERTIES ARE DEFINED ONLY OVER TREES
GRAMMAR: ORDERED SET OF RULES THAT DEFINES HOW TO FORM STRINGS FROM A LANGUAGE VOCABULARY
AMBIGUOUS GRAMMAR: THERE ARE MORE WAYS TO GENERATE THE SAME STRING; $(\lambda \in (N \cup \Sigma)^*)$
CFG GRAMMAR: RULES ARE APPLIED REGARDLESS OF THE CONTEXT; $G = \langle N, S, \Sigma, R \rangle$; CHOMSKY NORM. FORM: ALL PRODUCTION RULES ARE OF THE FORM: $N_i \to N_j N_k$, $N \to a$ (IN CFG RULES ARE $N \mapsto \lambda$)
PHRASE-STRUCTURE TREES REPRESENT THE SYNTACTIC STRUCTURE OF THE SENTENCE AND ITS DERIVATION WITH G; PCFG: ASSIGN A PROB. TO EACH PRODUCTION RULE s.t. $\sum_{k=1}^{K} P(\alpha_k | N) = 1$
WCFG: ASSIGN GENERIC NON-NEGATIVE WEIGHTS $\to$ $p(t) = Z^{-1} \prod_{r \in t} \exp(\text{score}(r))$ $Z = \sum_{t' \in \tau} \prod_{r' \in t'} \exp(\text{score}(r'))$
THIS Z CAN'T BE COMPUTED EASILY DUE TO INFINITE SUM BIGGER THAN Σ^*; CONDITION ON INPUT SENTENCE s $p(t|s) = Z(s)^{-1} \exp(\text{score}(t))$; $Z(s) = \sum_{t' \in \tau(s)} \exp(\text{score}(t'))$; ONLY SUM ON TREES THAT YIELD s; WE FACTOR THE SCORE OF A TREE ALONG THE PRODUCTIONS; THE NORMALIZER CAN STILL DIVERGE $\to$ TRANSFORM CFG IN CNF
$Z(s) = \sum_{t' \in \tau(s)} \prod_{(X \to Y_1 Y_2) \in t'} \exp(\text{score}(X \to Y_1 Y_2)) \cdot \prod_{(X \to x) \in t'} \exp(\text{score}(X \to x))$; IN CNF, THE NUMBER OF ADMISSIBLE TREES IS THE CATALAN NUMBER $C_{M-1} = (M-1+1)\binom{2(M-1)}{M-1}$ FOR A M-WORD LONG SENTENCE $O(4^m)$
CKY ALGORITHM: DP TO COMPUTE Z; TIME COMPLEXITY $O(M^3|R|)$, SPACE COMPLEXITY $O(M^2|N|)$

```
SEMIRING CYK(s, <N, S, Σ, R>, SCORE)
N ← |s|  →(WAS = M ABOVE)
CHART ← 0̄
FOR m = 1, ... N:  → FILL IN FIRST ROW OF PYRAMID CHART
    FOR X → s_m ∈ N:
        CHART[m, m+1, X] ⊕= EXP(SCORE(X → s_m))
FOR SPAN = 2, ..., N:  → SPAN LENGTH
    FOR i = 1, ..., N - SPAN:  → START OF SPAN
        k ← i + SPAN  → END OF SPAN
        FOR j = i+1, .., k-1:  → SUBSPAN DIVIDER
            FOR X → Y Z ∈ N:
                CHART[i,k,X] ⊕= EXP(SCORE(X → Y Z))
                    ⊗ CHART[i,j,Y] ⊗ CHART[j,k,Z]
RETURN CHART[0,N,S]  → CAN ADD BACKPOINTERS
```

THE VITERBI, INSIDE, LOG SEMIRINGS COMPUTE THE BEST PARSE, Z AND LOG(Z) RESPECTIVELY
CFG PARSERS ARE TRAINABLE WITH BACKPROP. BY MAXIMUM LIKELIHOOD ESTIMATION ON A DATASET

8 - DEPENDENCY PARSING

DEPENDENCY GRAMMAR: LINK EVERY WORD TO ITS SYNTACTIC HEAD; LABELED, DIRECTED HEAD $\to$ DEPENDENT
EDGES FORM A DEPENDENCY TREE $\to$ DIRECTED SPANNING TREE WITH ROOT CONSTRAINT OF 1; RELATED TO LEXICALIZED CFGs; CONVERT TO AND FROM CONST. TREES WITH GRAMMATICAL HEURISTICS
PROJECTIVE DEP. TREES: HAVE NO CROSSING ARCHS AND CAN BE PARSED WITH VERSIONS OF CKY (AND VICE V.)
$p(t|w) = Z^{-1} \exp(\text{score}(t, w))$; $Z = \sum_{t' \in \tau(w)} \exp(\text{score}(t', w))$ $\to$ Z SUM IS $O(m^n)$ NAIVELY ($m = |w|$) ↗ ROOT EDGE
ADD EDGE FACTORED ASSUMPTION: $Z = \sum_{t' \in \tau(w)} \prod_{(i,j,\ell) \in t'} \exp(\text{score}(i,j,\ell, w))$ $\exp(\text{score}(r, w))$ ↗ SCORE
KIRCHOFF MTT: COUNT NUMBER OF UNDIRECTED SPANNING TREES IN $O(m^3)$; GENERALIZED FOR DIRECTED TREES WITH A ROOT CONSTRAINT s.t. $Z = \det(L)$, $L_{ij} = P_j$ IF $i=1$, $\sum_{i'=1, i' \neq j}^{m} A_{i'j}$ IF $i=j$, ELSE $-A_{ij}$
DECODING: FINDING MAXIMUM-WEIGHT TREE WITH ROOT CONSTRAINT, NO CYCLES, ONE INCOMING EDGE PER NODE
$\text{ARGMAX}_{t \in \tau} \sum_{(i,j,\ell) \in t} \text{SCORE}(i, j, w)$; GREEDY APPROACHES FAIL
CHU-LIU-EDMONDS ALGORITHM: SOLVES THE DECODING PROBLEM IN $O(m^2)$ $\to$ SPECIAL DATA STRUCTURE

```
MST(G):  ↗ TAKE BEST INCOMING NODE FOR EACH NODE
  IF CYCLE IN GREEDY(G):  → SELECT MAX ↗ UPDATE ENTER EDGES WEIGHT TO HANDLE CYCLE BREAK
    RETURN EXPAND(CONSTRAIN(MST(CONTRACT(G, CYCLE))))
  ELSE: RETURN CONSTRAIN(GREEDY(G))
CONSTRAIN(G):  ↗ ROOT CONSTRAINT VIOLATION
  IF NUM_ROOT_EDGES(GREEDY(G)) > 1:  → DELETE NODE WITH LOWER COST
    E = ROOT_EDGE_TO_REMOVE(G)  COST = EDGE COST - WEIGHT OF NEXT BEST INCOMING EDGE
    IF CYCLE IN GREEDY(G - e): RETURN CONSTRAIN(CONTRACT(G, CYCLE))
    ELSE: RETURN CONSTRAIN(G - e)
  ELSE: RETURN GREEDY(G)
```

A NAIVE HANDLING OF THE ROOT CONSTRAINT WOULD ADD $O(m)$ TO THE RUNTIME: FIX EACH EDGE AS THE ONLY ONE EMANATING FROM P AND RE-RUN MST TO SELECT BEST

9 - LAMBDA CALCULUS

ENCODE SEMANTIC REPRESENTATION OF TEXT, REMOVE SEMANTIC AMBIGUITIES
PRINCIPLE OF COMPOSITIONALITY: THE MEANING OF A COMPLEX EXPRESSION IS A FUNCTION OF ITS CONSTITUENT'S
LAMBDA OPERATOR: $\lambda x.F(x)$ $\to$ FUNCTION THAT TAKES x AS INPUT AND PRODUCES F(x) AS OUTPUT ↗ (NOT BOUND)
FREE VARIABLE: VARIABLE THAT DOES NOT OCCUR IN THE SCOPE OF ANY ABSTRACTION WITH ITS NAME
α-CONVERSION: RENAME A VARIABLE IN AN ABSTRACTION WITH ALL ITS BOUND VARIABLES $\lambda x.(x\,x)\,y \to \lambda t.(t\,t)\,y$
β-REDUCTION: APPLY A LAMBDA TERM TO ANOTHER: $(\lambda x.\lambda y.(x((\lambda x.x)\,y))\,z) \to \lambda y.((z((\lambda x.x)\,z))\,y))$ $\to$ ONLY FREE CH.
ENRICHED LAMBDA CALCULUS: HANDLE SENTENCE SEMANTICS; LOGICAL CONSTANTS, VARIABLES, LITERALS
EXAMPLES: $(\lambda Q.Q(\lambda t.(\forall x.(\text{DOG}(x) \Rightarrow \text{LIKES}(t,x))))\,\lambda P.P(\text{ALEX}))$ $\to$ $(\lambda P.P(\text{ALEX}))\,\lambda t.(\forall x.(\text{DOG}(x) \Rightarrow \text{LIKES}(t,x)))$ $\to$ $(\lambda t.(\forall x(\text{DOG}(x) \Rightarrow \text{LIKES}(t,x)))\,\text{ALEX})$ $\to$ $\forall x(\text{DOG}(x) \Rightarrow \text{LIKES}(\text{ALEX},x))$; $(\lambda P.\lambda Q.\lambda x.P(\lambda y.\text{LIKES}(x,y)))\,\lambda P.P(\text{ALEX}))$ $\to$ $\lambda Q.Q(\lambda x.(\lambda P.P(\text{ALEX})\,\lambda y.\text{LIKES}(x,y)))$ $\to$ $\lambda Q.Q(\lambda x.(\lambda y.\text{LIKES}(x,y)\,\text{ALEX}))$; $(\lambda Q.Q(\lambda x.\text{LIKES}(x,\text{ALEX}))$
$\lambda Q.\exists x(\text{DOG}(x) \wedge Q(x)))$ $\to$ $(\lambda Q.\exists x(\text{DOG}(x) \wedge Q(x))\,\lambda x.\text{LIKES}(x,\text{ALEX}))$ $\to$ $\exists x(\text{DOG}(x) \wedge ((\lambda x.\text{LIKES}(x,\text{ALEX}))\,x))$ ...

10 - WFST

WEIGHTED FINITE STATE TRANSDUCERS: MAP SEQUENCES FROM SOURCE TO TARGET VOCABULARY; HAS FINITE NUMBER OF STATES
WEIGHTED WITH PROBS; $T = \langle Q, \Sigma, \Omega, \lambda, \rho, \delta \rangle$ Q: SET OF STATES; Σ, Ω INPUT AND OUTPUT VOCABULARY
λ, ρ MAP STATES TO INITIAL OR FINAL SCORES, δ MAPS TRANSITION ARCS TO SCORES; UNAMBIGUOUS WFST: EACH PAIR OF STRINGS HAS AT MOST ONE ACCEPTING PATH; $\text{SCORE}(\pi) = \sum_{m=1}^{|\pi|} \text{SCORE}(\pi_m)$; $p(y|x) = Z^{-1} \exp(\text{SCORE}(\pi))$
$Z = \sum_{y' \in \Omega^*} \exp(\text{score}(y', x)) = \sum_{y' \in \Omega^*} \sum_{\pi' \in \pi(x,y')} \exp(\sum_{m=1}^{|\pi|} \text{score}(\pi_m))$

FLOYD-WARSHALL ALGORITHM: SOLVES ALL PAIRS SHORTEST PATHS GIVEN NO NEGATIVE CYCLES $\to O(|V|^3)$

```
FOR EACH EDGE (u,v):  DIST MATRIX INIT WITH 0̄
  DIST[u][v] = W[u][v]  → W^1
FOR EACH VERTEX v:  SEMIRING GENERALIZ. FROM MATRIX MULT.
  DIST[v][v] = 0  → W^0
FOR k IN 1, |N|:
  FOR i IN 1, |N|:  DIST[i][j] = DIST[i][j] ⊕
    FOR j IN 1, |N|:  (DIST[i][k] ⊗ DIST[k][k]* ⊗ DIST[k][j])
      IF DIST[i][j] > DIST[i][k] + DIST[k][j]:
        DIST[i][j] = DIST[i][k] + DIST[k][j]
```

KLEENE CLOSURE OF A MATRIX: $W^* = \bigoplus_{k=0}^{\infty} W^k = I + W \bigoplus_{k=0}^{\infty} W^k = I + WW^*$ $\to$ $W^* - WW^* = I$ $\to$ $W^* = (I - W)^{-1}$ $\to$ IF LARGEST EIGENVALUE OF $W < 1$; FIND Z THAT NORMALIZES $p(y|x)$ IN $O(m^3)$
$Z = \lambda^T (\sum_{w \in \Omega \cup \{\varepsilon\}} W^w)^* \beta$ WHERE λ IS A VECTOR OF START WEIGHTS, β A VECTOR OF END WEIGHTS AND W^w A SET OF TRANSITION MATRICES LABELED WITH $w \in \Omega \cup \varepsilon$, WHERE W^w_{mm} IS THE WEIGHT FROM STATE m TO m WITH TRANSITION LABEL w; m-GRAM WFST: UNIGRAM: 1 STATE, $\sum_{m=1}^{M} \delta(q_0, w_m, q_0) = \sum_{m=1}^{M} \log p(w_m)$; BIGRAM: $1 + |V| + 1$ STATES, TRIGRAM: $|V|^2 + |V| + 1$ STATES, $\delta(q_{i,j}, w, q_{k,i}) = \log p(w_m = k | w_{m-1} = i, w_{m-2} = j)$ IF $w = k$; m-GRAM: $|V|^{m-2} + |V|^{m-1} + 1$ STATES

11 - SEQUENCE MODELS

TRANSLATION: MAP STRINGS FROM ONE SPACE TO ANOTHER, BUT THERE IS NO CORRECT ANSWER, LOCALITY ASSUMPTIONS ARE NOT REASONABLE; NEURAL MACHINE TRANSLATION MODELS THE MAPPING WITH A COMPLEX FUNCTION; DIVISIBLE IN MODELING AND DECODING; SEQ-2-SEQ MODEL: $z = \text{ENCODER}(x)$; $y|x = \text{DECODER}(z) \to$ LOCALLY NORM.
$p(y|x) = \prod_{t=1}^{T} p(y_t | x, y_1, \dots y_{t-1})$; ATTENTION: RATHER THAN ENCODING ALL INPUT INTO A SINGLE CONTEXT VECTOR, PAY VARIABLE ATTENTION TO EACH REPRESENTATION DEPENDING ON THE OUTPUT GENERATION STEP
$\alpha^t = \text{SOFTMAX}(\text{SCORE}(q_t, K))$; $c^t = \alpha^{tT} V$; $k_i = v_i = h_i^e$; $q_t = h_t^d$; $K = V = H^e$; SELF-ATTENTION: ATTENTION ON OWN CONTEXT: $k_i = v_i = h_i^{d/e}$; $q_t = h_t^{d/e}$; MULTI-HEAD ATTENTION: LEARN MULTIPLE SETS OF ATTENTION WEIGHTS FOR THE SAME INPUT AND CONCATENATE; SCORE CAN BE SCALED DOT PRODUCT $\text{SCORE}(q_t, h_t) = q_t^T h_t \cdot |h_t|^{-\frac{1}{2}}$
TRANSFORMERS: SOLVE SLOW RECURRENCE OF RNNs; USED FOR MOST LANGUAGE GENERATION TASK; POSITIONAL EMBEDDINGS: ENCODE ORDER OF WORDS IN THE SEQUENCE; RESIDUAL CONNECTIONS: MITIGATE VANISHING GRADS. LAYER NORMALIZATION: HELPS WITH INTERNAL COVARIATE SHIFT, NORMALIZE INDIVIDUAL LAYER INPUTS
DECODING: $y^* = \text{ARGMAX}_{y \in Y} \text{SCORE}(x, y)$, NO INDEPENDENCE ASSUMPTION, SO HEURISTICS OR SAMPLING METHODS ARE NEEDED; BEAM SEARCH: PRUNED BFS WITH LIMITED BREADTH; NUCLEUS SAMPLING: SAMPLE FROM CONDITIONAL DISTR. CONSIDERING ONLY TOP ITEMS COVERING p% OF THE PROBABILITY MASS ↗ PRECISION OF TRANSLATION
EVALUETION METRICS: BLEU: FRACTION OF PREDICTED m-GRAMS THAT APPEAR IN THE GROUND TRUTH TOO
METEOR: COMBINATION OF RECALL AND PRECISION OF PREDICTED TRANSLATION WRT. THE REFERENCE
TRANSFORMERS ARE QUADRATIC IN THE CONTEXT SIZE AND PERFORM BETTER THAN m-GRAM MODELS WITH BIG DATASETS
NUMBER OF TERMS DERIVATIVE SEQ-2-SEQ: $O(N - m + m)$, WITH ATTENTION $O((N - m + m) \cdot N)$
$\frac{\partial l_{m+1}}{\partial h_m^s} = \frac{\partial l_{m+1}}{\partial h_m^t} \frac{\partial h_m^t}{\partial h_{m-1}^t} \dots \frac{\partial h_2^t}{\partial h_1^t} \frac{\partial h_1^t}{\partial h_N^s} \dots \frac{\partial h_{m+1}^s}{\partial h_m^s}$; $\frac{\partial c_m}{\partial h_m^s} = \partial_{m,m} + \frac{\partial o_{mn}}{\partial h_n^s} \cdot h_n^s + \sum_{i \neq m}^{N} \frac{\partial}{\partial h_m^s} \partial_{m,i} h_i^s$ → WITH ATTENTION VECTOR
ATTENTION REDUCES THE SHORTEST DECODER → ENCODER PATH LENGHT FROM PROPORTIONAL TO THE INPUT (AND OUTPUT) SIZE, TO ONLY CONSTANT

12 - AXES OF MODELING

CLASSIFICATION MODELS CAN BE DIVIDED INTO PROBABILISTIC OR DETERMINISTIC; PROBABILISTIC MODELS CAN BE DISCRIMINATIVE OR GENERATIVE; BOTH CAN BE LOCALLY OR GLOBALLY NORMALIZED; DIFFERENT MODELS HAVE DIFFERENT INDUCTIVE BIASES THAT CAN BE USEFUL TO SOLVE SPECIFIC PROBLEMS; LARGE MODELS OVERFIT DUE TO THE BIAS-VARIANCE TRADEOFF; LOSS FUNCTION: ANY FUNCTION OF MODEL PARAMETERS THAT CLASSIFIES FIT TO DATA
MLE ESTIMATORS ARE CONSISTENT: $\hat{\theta}_m \xrightarrow{p} \theta$ AND ASYMPTOTICALLY EFFICIENT, BUT CAN EASILY OVERFIT
LOSS SHOULD BE CONVEX, DIFFEREN., LOW COMPUT. COST; HINGE LOSS $l(y, \hat{y}) = \max\{0, 1 - \hat{y} \cdot y\}$; EXP LOSS: $e^{-y \cdot \hat{y}}$
REGULARIZATION: ADD PRIOR INFORMATION TO PREVENT OVERFITTING TO NOISE → NECESSARY FOR GOOD GENERALIZATION
MODELS ARE EVALUATED ON EVALUATION METRICS WHICH ARE HIGHLY PROBLEM SPECIFIC AND DIFFERENT FROM LOSS TO WEIGHT ERROR TYPES DIFFERENTLY OR FOR HIGH COMPUTATION COST → SOMETIMES THEY COINCIDE (5)
INTRINSIC EVALUATION: EVALUATE BASED ON A CRITERIA; PERPLEXITY $(\bar{w}) = 2^{-\frac{LL(\bar{w})}{M}}$ ↳ HELD-OUT LOG-LIKELIHOOD
EXTRINSIC EVALUATION: EVAL. BASED ON THE PERFORMANCE OF A SUCCESSIVE TASK
MODEL SELECTION: CHOOSING MODEL WITH BEST INFERENCE AND PREDICTION PERFORMANCE; NESTED-CV EVALU.
MODEL STABILITY; p-VALUE OF STATISTICAL TEST t: $p = 2\min(P(T \geq t | H_0), P(T \leq t | H_0))$, REJECT H_0 IFF $p < \alpha$
STATISTICAL POWER: $P = (\text{REJECT } H_0 | H_1)$, INCREASES WITH INCREASING SAMPLE SIZE; MULTIPLE TESTING: $P(1 \text{ FALSE REJECTIONS} | > 0) = 1 - (1 - \alpha)^K$; BOFFERONI CORRECTION: $\alpha^* = \alpha / K$; STATISTICAL TESTS CAN BE PARAM. OR NON-PARAMETRIC → THE PARAMETER OF THE DISTRIBUTION OF THE TEST IS NOT A FUNCTION OF THE DATA
MCNEMAR'S TEST: COMPARE DISAGREEMENT PROBABILITY OF TWO CLASSIFIERS THAT SHARE TEST SET H_0: $p_b = p_c$
PERMUTATION TESTS: TEST IF CLASSIFIERS PERFORMS BETTER THAN RANDOM CHANCE ON DATA $p\text{-val} = |\{i : p_i \leq p^*_{\hat{j}}\}| + 1) \cdot (k+1)^{-1}$
5x2 CV PAIRED t-TEST: TEST TO COMPARE PERFORMANCE OF TWO CLASSIFIERS ON A METRIC P
DIFFERENT TESTS HAVE DIFFERENT STATISTICAL POWERS AND OBSERVED TYPE 1 ERRORS → IMPORT. FOR CHOOSING
THE TEST SET CAN BE NOT LARGE ENOUGH TO GIVE ENOUGH POWER TO DETECT DIFFERENCES IN MODEL PERFORMANCE