Quo Vadis, Math Information Retrieval

Introduction

“If not now, when?”

Chapters of the Fathers (Pirkei Avot, 1:14)

First Matter

- Infty, 2003: Suzuki et al. **Math optical character recognition**
- DML-CZ, 2005–2009: Gensim (Generate Similar) to **classify and categorize math**ematical knowledge [43] by automated means and tools (2003 citations by now).
- 2008–2011: **DML workshop series** specifically targeted for MIR in STEM
- EuDML, 2010–2013: DML that deployed the **MIaS search engine**
- NTCIR 10, 2013: **MIR evaluation competition** (MIaS won NTCIR 11)
- Tangent, 2014: **visual-based indexing** (symbol layout tree) and searching
- MCAT, 2016 system combining **text and math for reranking**
- Equation Embeddings, 2018, **joint embedding model**
- TopicEq, 2019, **joint text and math topic modeling**
- ARQMath % CLEF 2020, first **Q&A in STEM domain**

The Math Indexer and Searcher of New Generation

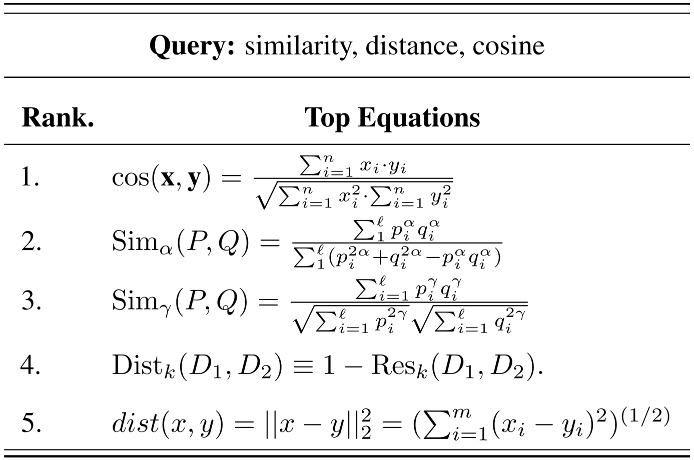
EqEmb, TopicEq, Bi-LSTM, BERT et al.
Are you awakened to start the journey?

Equation Embeddings

“The day is short, the labor vast, the toilers idle,
the reward great, and the Master of the house is insistent.”

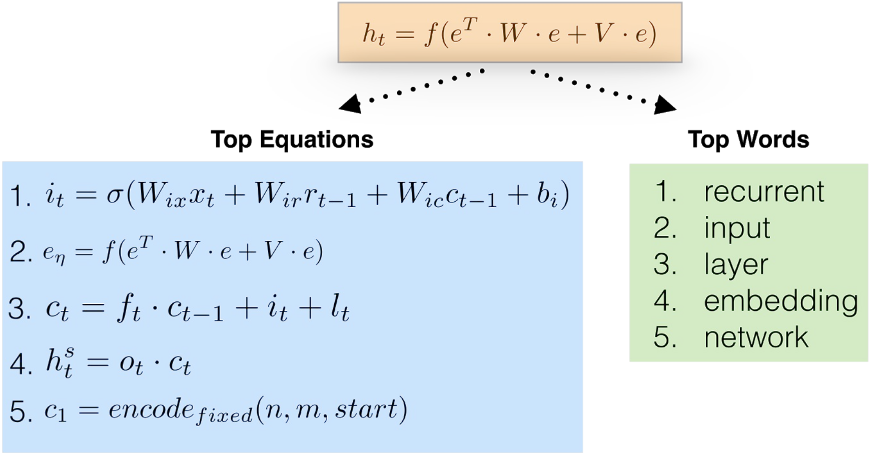
Chapters of the Fathers (Pirkei Avot, 2:20)

EqEmb (Krstovski and Blei, 2018)



Query: similarity, distance, cosine

Rank.	**Top Equations**
1.	$\cos(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \cdot \sum_{i=1}^{n} y_i^2}}$
2.	$\mathrm{Sim}_\alpha(P,Q) = \frac{\sum_1^\ell p_i^\alpha q_i^\alpha}{\sum_1^\ell (p_i^{2\alpha}+q_i^{2\alpha}-p_i^\alpha q_i^\alpha)}$
3.	$\mathrm{Sim}_\gamma(P,Q) = \frac{\sum_{i=1}^\ell p_i^\gamma q_i^\gamma}{\sqrt{\sum_{i=1}^\ell p_i^{2\gamma}}\sqrt{\sum_{i=1}^\ell q_i^{2\gamma}}}$
4.	$\mathrm{Dist}_k(D_1,D_2) \equiv 1 - \mathrm{Res}_k(D_1,D_2).$
5.	$dist(x,y) = \|\|x-y\|\|_2^2 = (\sum_{i=1}^m (x_i - y_i)^2)^{(1/2)}$



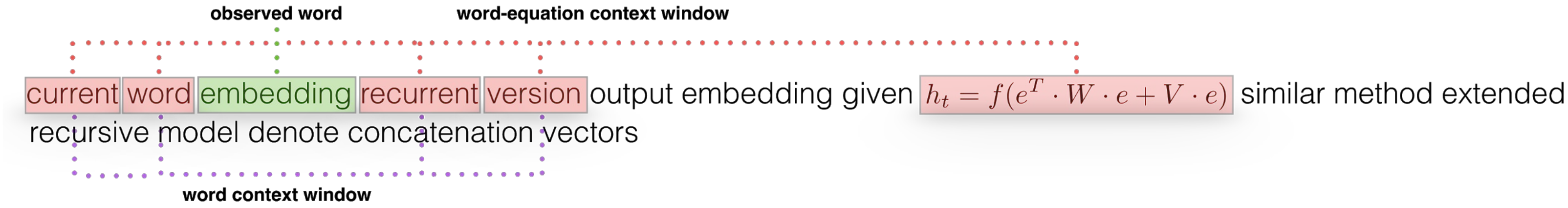
$h_t = f(e^T \cdot W \cdot e + V \cdot e)$

Top Equations

1. $i_t = \sigma(W_{ix}x_t + W_{ir}r_{t-1} + W_{ic}c_{t-1} + b_i)$
2. $e_\eta = f(e^T \cdot W \cdot e + V \cdot e)$
3. $c_t = f_t \cdot c_{t-1} + i_t + l_t$
4. $h_t^s = o_t \cdot c_t$
5. $c_1 = encode_{fixed}(n,m,start)$

Top Words

1. recurrent
2. input
3. layer
4. embedding
5. network



observed word
word-equation context window
current word embedding recurrent version output embedding given $h_t = f(e^T \cdot W \cdot e + V \cdot e)$ similar method extended
recursive model denote concatenation vectors
word context window

EqEmb-U

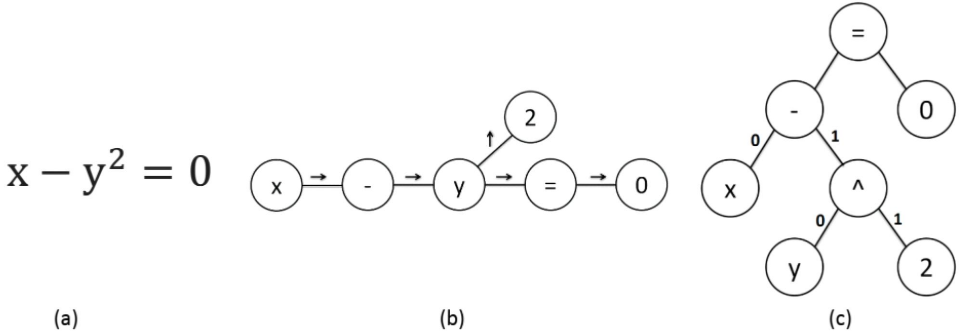


Figure 1: Formula (a) $x - y^2 = 0$ with associated (b) Symbol Layout Tree (SLT), and (c) Operator Tree (OPT). SLTs represent formula appearance by the placement of symbols on writing lines, while OPTs define the mathematical operations represented in expressions.

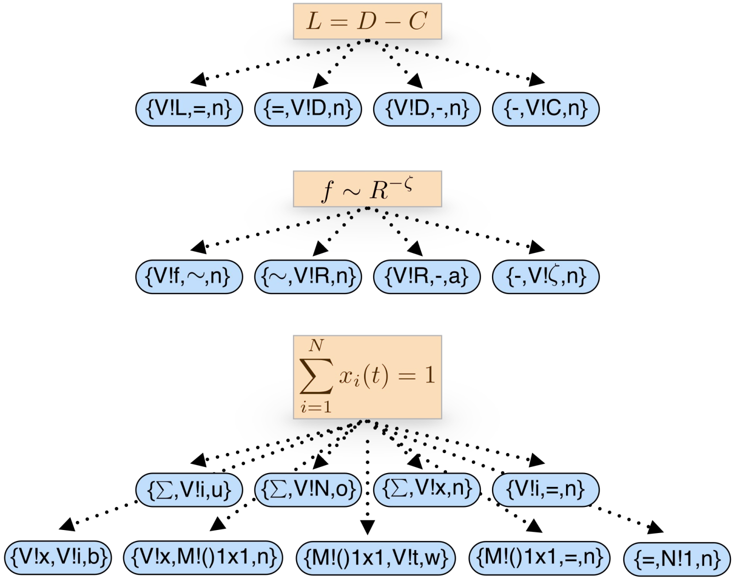
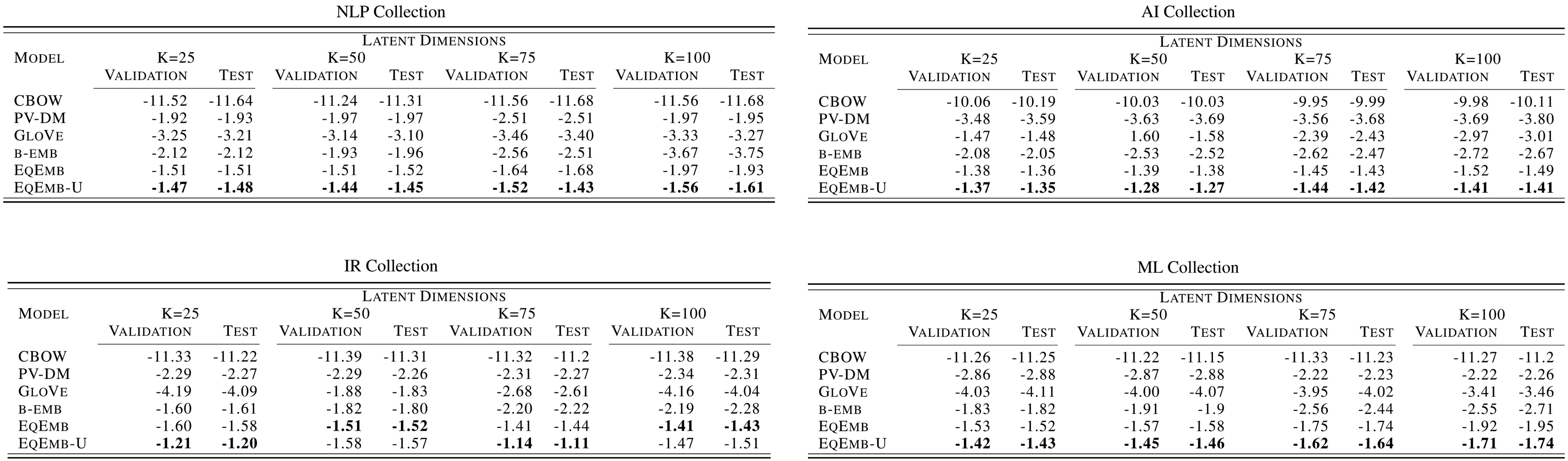


Figure 2. Examples of Syntax Layout Tree (SLT) representation of equations using a symbol window of size one. Each tuple represents the special relationship between two symbols (*n*-to the right; *a*-above; *u*-under; *o*-over; *w*-within).

Evaluation



NLP Collection

	Latent Dimensions							
Model	K=25		K=50		K=75		K=100	
	Validation	Test	Validation	Test	Validation	Test	Validation	Test
CBOW	-11.52	-11.64	-11.24	-11.31	-11.56	-11.68	-11.56	-11.68
PV-DM	-1.92	-1.93	-1.97	-1.97	-2.51	-2.51	-1.97	-1.95
GloVe	-3.25	-3.21	-3.14	-3.10	-3.46	-3.40	-3.33	-3.27
b-emb	-2.12	-2.12	-1.93	-1.96	-2.56	-2.51	-3.67	-3.75
EqEmb	-1.51	-1.51	-1.51	-1.52	-1.64	-1.68	-1.97	-1.93
EqEmb-U	**-1.47**	**-1.48**	**-1.44**	**-1.45**	**-1.52**	**-1.43**	**-1.56**	**-1.61**

AI Collection

	Latent Dimensions							
Model	K=25		K=50		K=75		K=100	
	Validation	Test	Validation	Test	Validation	Test	Validation	Test
CBOW	-10.06	-10.19	-10.03	-10.03	-9.95	-9.99	-9.98	-10.11
PV-DM	-3.48	-3.59	-3.63	-3.69	-3.56	-3.68	-3.69	-3.80
GloVe	-1.47	-1.48	1.60	-1.58	-2.39	-2.43	-2.97	-3.01
b-emb	-2.08	-2.05	-2.53	-2.52	-2.62	-2.47	-2.72	-2.67
EqEmb	-1.38	-1.36	-1.39	-1.38	-1.45	-1.43	-1.52	-1.49
EqEmb-U	**-1.37**	**-1.35**	**-1.28**	**-1.27**	**-1.44**	**-1.42**	**-1.41**	**-1.41**

IR Collection

	Latent Dimensions							
Model	K=25		K=50		K=75		K=100	
	Validation	Test	Validation	Test	Validation	Test	Validation	Test
CBOW	-11.33	-11.22	-11.39	-11.31	-11.32	-11.2	-11.38	-11.29
PV-DM	-2.29	-2.27	-2.29	-2.26	-2.31	-2.27	-2.34	-2.31
GloVe	-4.19	-4.09	-1.88	-1.83	-2.68	-2.61	-4.16	-4.04
b-emb	-1.60	-1.61	-1.82	-1.80	-2.20	-2.22	-2.19	-2.28
EqEmb	-1.60	-1.58	**-1.51**	**-1.52**	-1.41	-1.44	**-1.41**	**-1.43**
EqEmb-U	**-1.21**	**-1.20**	-1.58	-1.57	**-1.14**	**-1.11**	-1.47	-1.51

ML Collection

	Latent Dimensions							
Model	K=25		K=50		K=75		K=100	
	Validation	Test	Validation	Test	Validation	Test	Validation	Test
CBOW	-11.26	-11.25	-11.22	-11.15	-11.33	-11.23	-11.27	-11.2
PV-DM	-2.86	-2.88	-2.87	-2.88	-2.22	-2.23	-2.22	-2.26
GloVe	-4.03	-4.11	-4.00	-4.07	-3.95	-4.02	-3.41	-3.46
b-emb	-1.83	-1.82	-1.91	-1.9	-2.56	-2.44	-2.55	-2.71
EqEmb	-1.53	-1.52	-1.57	-1.58	-1.75	-1.74	-1.92	-1.95
EqEmb-U	**-1.42**	**-1.43**	**-1.45**	**-1.46**	**-1.62**	**-1.64**	**-1.71**	**-1.74**

Table 2. EqEmb outperform previous embedding models; EqEmb-U further improves performance. Performance comparisons between CBOW, GloVe, PV-DM, b-emb, EqEmb and EqEmb-U using log-likelihood computed on test and validation datasets. Comparisons were done over 4 different collections of scientific articles (NLP, IR, AI and ML) and across different latent dimensions (K=25, 50, 75 and 100).

Future Work – Operator Trees

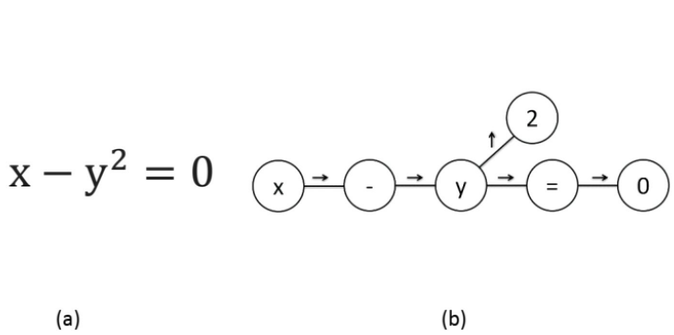


Figure 1: Formula (a) $x - y^2 = 0$ with associated (b) Symbol Layout Tree (SLT), and (c) Operator Tree (OPT). SLTs represent formula appearance by the placement of symbols on writing lines, while OPTs define the mathematical operations represented in expressions.

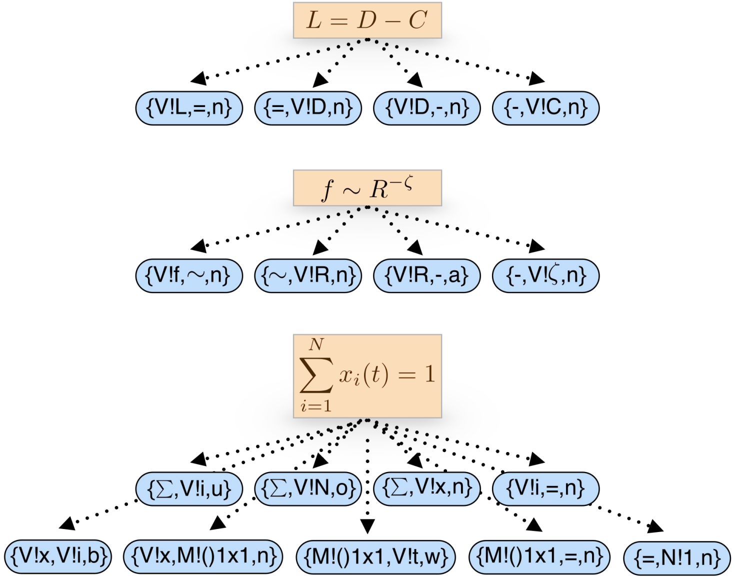
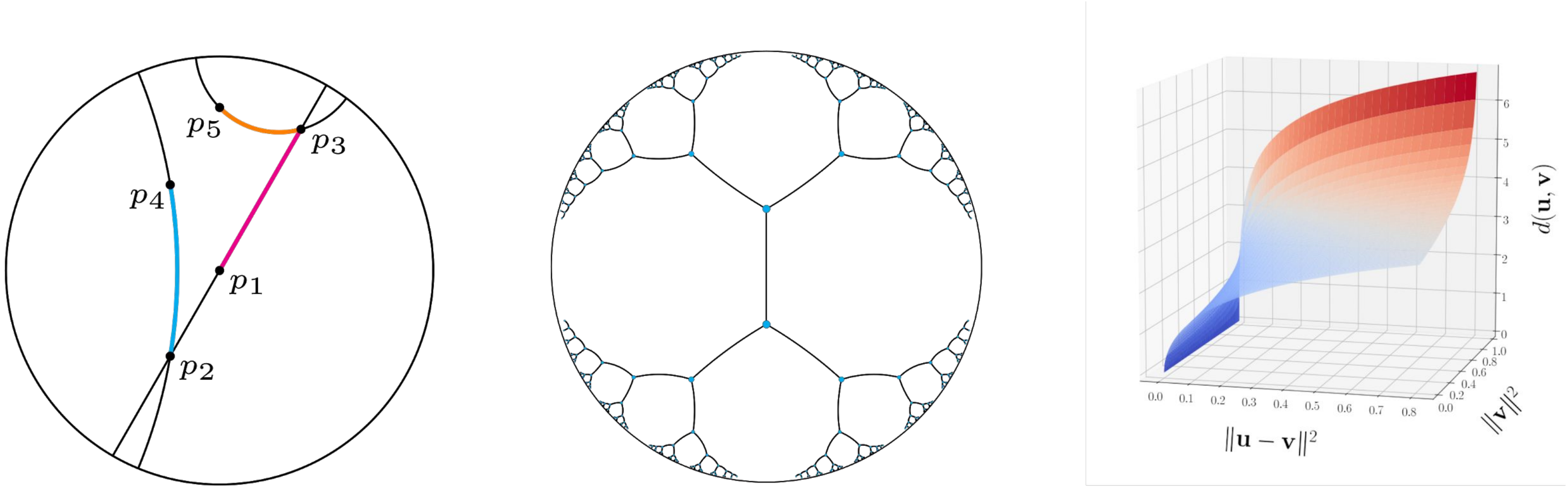


Figure 2. Examples of Syntax Layout Tree (SLT) representation of equations using a symbol window of size one. Each tuple represents the special relationship between two symbols (*n*-to the right; *a*-above; *u*-under; *o*-over; *w*-within).

Future Work – SGD in H^n (Nickel and Kiela, 2017)



(a) Geodesics of the Poincaré disk

(b) Embedding of a tree in $\mathcal{B}^2$

(c) Growth of Poincaré distance

Figure 1: (a) Due to the negative curvature of $\mathcal{B}$, the distance of points increases exponentially (relative to their Euclidean distance) the closer they are to the boundary. (c) Growth of the Poincaré distance $d(\boldsymbol{u}, \boldsymbol{v})$ relative to the Euclidean distance and the norm of $\boldsymbol{v}$ (for fixed $\|\boldsymbol{u}\| = 0.9$). (b) Embedding of a regular tree in $\mathcal{B}^2$ such that all connected nodes are spaced equally far apart (i.e., all black line segments have identical hyperbolic length).

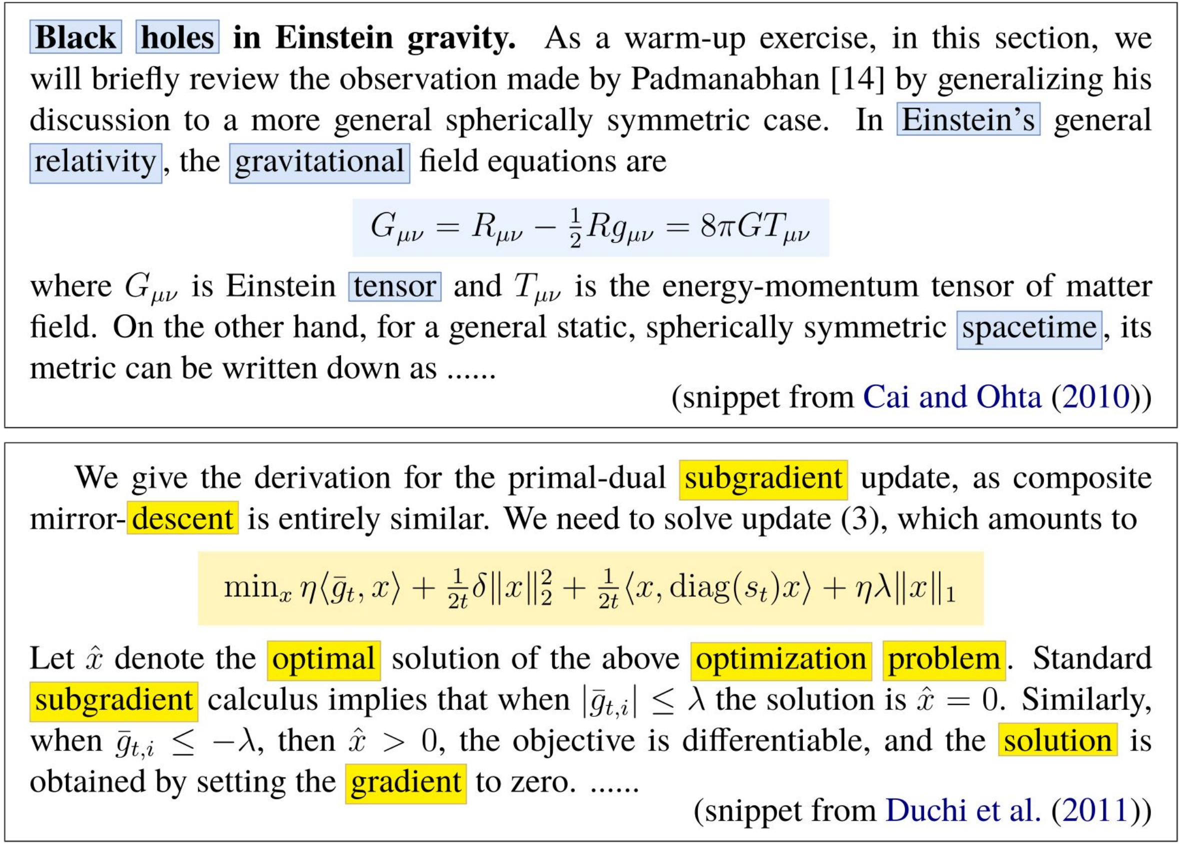
TopicEq

“He who acquires a good name,
has acquired himself something indeed.”

Chapters of the Fathers (Pirkei Avot, 2:8)

Quantum physics	spin energy field electron magnetic state states hamiltonian
Particle physics	higgs neutrino coupling decay scale masses mixing quark
Astrophysics	mass gas star stellar galaxies disk halo radius luminosity
Relativity	black metric hole schwarzschild gravity holes einstein
Number theory	prime integer numbers conjecture integers degree modulo
Graph theory	graph vertex vertices edges node edge number set tree
Linear algebra	matrix matrices vector basis vectors diagonal rank linear
Optimization	problem optimization algorithm function solution gradient
Probability	random probability distribution process measure time
Machine learning	layer word image feature sentence model cnn lstm training

Table 2: Topics learned by the TopicEq model. Left: topic name (summarized by us). Right: top words in topic.



Black holes in Einstein gravity. As a warm-up exercise, in this section, we will briefly review the observation made by Padmanabhan [14] by generalizing his discussion to a more general spherically symmetric case. In Einstein's general relativity, the gravitational field equations are

$$G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu} = 8\pi G T_{\mu\nu}$$

where $G_{\mu\nu}$ is Einstein tensor and $T_{\mu\nu}$ is the energy-momentum tensor of matter field. On the other hand, for a general static, spherically symmetric spacetime, its metric can be written down as ......

(snippet from Cai and Ohta (2010))

We give the derivation for the primal-dual subgradient update, as composite mirror-descent is entirely similar. We need to solve update (3), which amounts to

$$\min_x \eta \langle \bar{g}_t, x \rangle + \frac{1}{2t} \delta \|x\|_2^2 + \frac{1}{2t} \langle x, \mathrm{diag}(s_t) x \rangle + \eta\lambda \|x\|_1$$

Let $\hat{x}$ denote the optimal solution of the above optimization problem. Standard subgradient calculus implies that when $|\bar{g}_{t,i}| \le \lambda$ the solution is $\hat{x} = 0$. Similarly, when $\bar{g}_{t,i} \le -\lambda$, then $\hat{x} > 0$, the objective is differentiable, and the solution is obtained by setting the gradient to zero. ......

(snippet from Duchi et al. (2011))

Figure 1: The words in a given technical context often characterize the distinctive types of equations used, and vice versa. **Top** topic: Relativity; **bottom** topic: Optimization.

Topic	Generated Equations
Quantum physics	• $E = \hbar\frac{\partial^2 S}{\partial t^2}(\frac{\partial\varphi}{\partial c}) - \frac{k}{\hbar^2}\frac{\partial B}{\partial t}\,(t + \partial_t\delta).$ • $\Psi_{\mathrm{pr}} = \sum_{\mathbf{l}}(\psi_{\mathbf{r}+\uparrow} - \psi^\dagger_{\mathbf{r}\downarrow}) + \sum_{\mathbf{r}'}(\psi^\dagger_{\mathbf{r}\downarrow,\uparrow} - \psi_{\mathbf{r}\downarrow}\sigma^\dagger).$
Particle physics	• $\mathcal{H} = \frac{2}{4}(\partial_\mu\phi)^2 + 2m\phi_\nu(\phi) + \frac{1}{2}m^2(\phi)(1-\phi^2)^2.$ • $m_{\mathrm{eff}}(M) = 1.4 \cdot 10^{-13}\,\mathrm{GeV}.$
Relativity	• $\mathcal{M} = \frac{2}{2}g^{\mu\nu}(f_{\mu\nu,\mu} - g_{\mu\nu,\nu} + g_{\nu\nu,b}f_{\mu,\nu}) + \frac{1}{2}g^{\mu\nu}.$ • $T_{\mu\nu} = \int_0^\infty ds_{\mu\nu}ds^2 + a^2_\mu dr^2 + r^2 d\Omega^2.$
Number theory	• $(2^k)^k + (1^n + 1)(1 + p^k) = 1.$
Linear algebra	• $\mathrm{tr}(E_\varepsilon X^*) = U^\top(\mathrm{tr}(V_\varepsilon X)).$ • $\phi_h(\theta, y) = \Big\{X \in \mathrm{Span}\,\big(P_c(\mathbf{T}[x, \mathsf{x}])\big)\Big].$
Optimization	• $\min_p p(x)$ subject to $\|p^x - y\|_2 \leq m_p.$ • $w^+ = w_t + g_t\|u_t - \nabla\mathbf{u}^*\|^2_2.$
Probability	• $\mathbb{P}(r_\tau < t) = \mathbb{E}_{\tau_{\mathrm{wist}}}(N_\tau).$ • $T^*(t) = \lim_{t\to\infty}\mathbb{E}[\lvert N(t) + \mathbb{E}[\varphi_t(x)]^*\rvert]$

Table 4: The TopicEq model generates equations that reflect the characteristics of given topics. Left: topic (picked from Table 2). Right: equations generated by the model conditioned on the given topic (one-hot topic vector θ).

Topic Gradation	**Generated Equation** (**Greedily decoded**)
Astrophysics (100%)	$G_{\text{eff}} = \frac{1}{2}\left(\frac{M_{\text{eff}}}{M_\odot}\right)^{-1}\left(\frac{M_{\text{eff}}}{M_\odot}\right)^{-1}$
⋮	$G_{\text{eff}} = \frac{1}{2}\left(\frac{M_{\text{eff}}}{M_\odot}\right)^{-1}$
⋮	$G_{\text{eff}} = \frac{1}{2}\left(\frac{1}{2} + \frac{1}{2}\right)$
50% — 50%	$G_{\text{s}} = \frac{1}{2}(1 - \frac{1}{2})$
⋮	$G_{\text{s}}(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{x} + \mathbf{x}^T\mathbf{x}$
⋮	$G_{\text{s}}(\mathcal{C}) = \mathcal{C}(\mathcal{C}) + \mathcal{C}(\mathcal{C}).$
Graph theory (100%)	$G_i = \{(x, y) \in \mathbb{R}^n : x_i = x_i\}$
Optimization (100%)	$L = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_2^2$
⋮	$L = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_2^2$
⋮	$L = \frac{1}{2}\sum_{i=1}^n \sum_{i=1}^n (x_i - x_i)^2 + \sum_{i=1}^n (x_i - x_i)^2$
50% — 50%	$L = \frac{1}{2}\sum_{i=1}^n \sum_{i=1}^n (x_i - x_i)^2 + \sum_{i=1}^n x_i^2$
⋮	$L = \frac{1}{N}\sum_{i=1}^N \sum_{i=1}^N (x_i - x_i)^2$
⋮	$L = \frac{1}{N}\sum_{i=1}^N \mathbb{E}[\mathbf{x}_i^T\mathbf{x}_i],$
Statistics (100%)	$L = \frac{1}{N}\sum_{i=1}^N \mathbb{E}[\mathbf{x}_i^T\mathbf{x}_i]$

Table 5: We let the TopicEq model greedily generate equations while smoothly changing θ between two topics (via linear interpolation). Left: given topic pair and its interpolation. Right: generated equation (for the first topic pair, we let the model generate from G; for the second pair, from $L =$).

Given Equation ⟦ ⟧ shows the correct formula name for readers	Inferred Topic (showing top 5 words)	
	by our TopicEq	by bag-of-token baseline
#1 $i\hbar\frac{\partial}{\partial t}\lvert\Psi(\mathbf{r},t)\rangle = \hat{H}\lvert\Psi(\mathbf{r},t)\rangle$ ⟦Schrödinger Equation⟧	hamiltonian, spin, particle, interaction, wave ✓	time, operator, space, hamiltonian, system ✓
#2 $F = \frac{d(mv)}{dt}$ ⟦Newton's 2nd Law of Motion⟧	velocity, particle, pressure, motion, force ✓	time, velocity, particle, diffusion, force ✓
#3 $W + \Delta U = \int f \cdot dx - mgh$ ⟦Potential energy & Work⟧	direction, force, surface, strain, stress ?	method, order, solution, numerical, problem ✗ *(vague)*
#4 $f_m = \sigma(W_f h_{m-1} + U_f x_m + b_f)$ ⟦LSTM⟧	layer, word, image, feature, network ✓	function, section, problem, condition, solution ✗ *(vague)*
#5 $P(X\|Y) = \frac{P(Y\|X)P(X)}{P(Y)}$ ⟦Bayes' Theorem⟧	random, variable, probability, distribution, entropy ✓	probability, random, theorem variable, distribution ✓
#6 $\lim_{n\to\infty} P\left(\sqrt{n}(S_n - \mu) \leq z\right) = \Phi\left(\frac{z}{\sigma}\right)$ ⟦Central Limit Theorem⟧	measure, random, process, gaussian, convergence ✓	probability, random, theorem variable, distribution ✓
#7 $f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots$ ⟦Taylor Expansion⟧	coefficients, series, expansion fourier, polynomial ✓	polynomial, series, function, convergence, order ✓
#7′ $h(b) = h(a) + \frac{h'(b)}{1!}(b-a) + \frac{h''(b)}{2!}(b-a)^2 + \cdots$ ⟦Taylor Expansion⟧	coefficients, series, expansion fourier, polynomial ✓	function, integral, equation point, solution ✗ *(fooled)*

Table 7: The TopicEq model can infer the appropriate topic for equations from various domains, with better precision and consistency than bag-of-token baseline. Left: given equation. Right: topic inferred by our model and the baseline. ✓ indicates that the inferred topic is correct; ✗ not good. We verified that the *exact* same equations did not appear in the training data.

Mathematical Expressions Embedding Using Tree-Structured Bidirectional LSTM

“What is the right path a man should choose?
Whatever is honorable to himself, and honorable in the eyes of others.”

Chapters of the Fathers (Pirkei Avot, 2:1)

Tree-Structured Bi-LSTM for Maths Embeddings

Bi-LSTM has been applied and works for natural language text

What are we investigating?

Maths expressions (equations and formulae) are a combination of variables (or operands) and operators.

But there are challenges because Maths expressions are not usually a linear sequence as applicable in natural language text.

During evaluation, these expressions are broken into logical units and each has order of precedence in the evaluation order

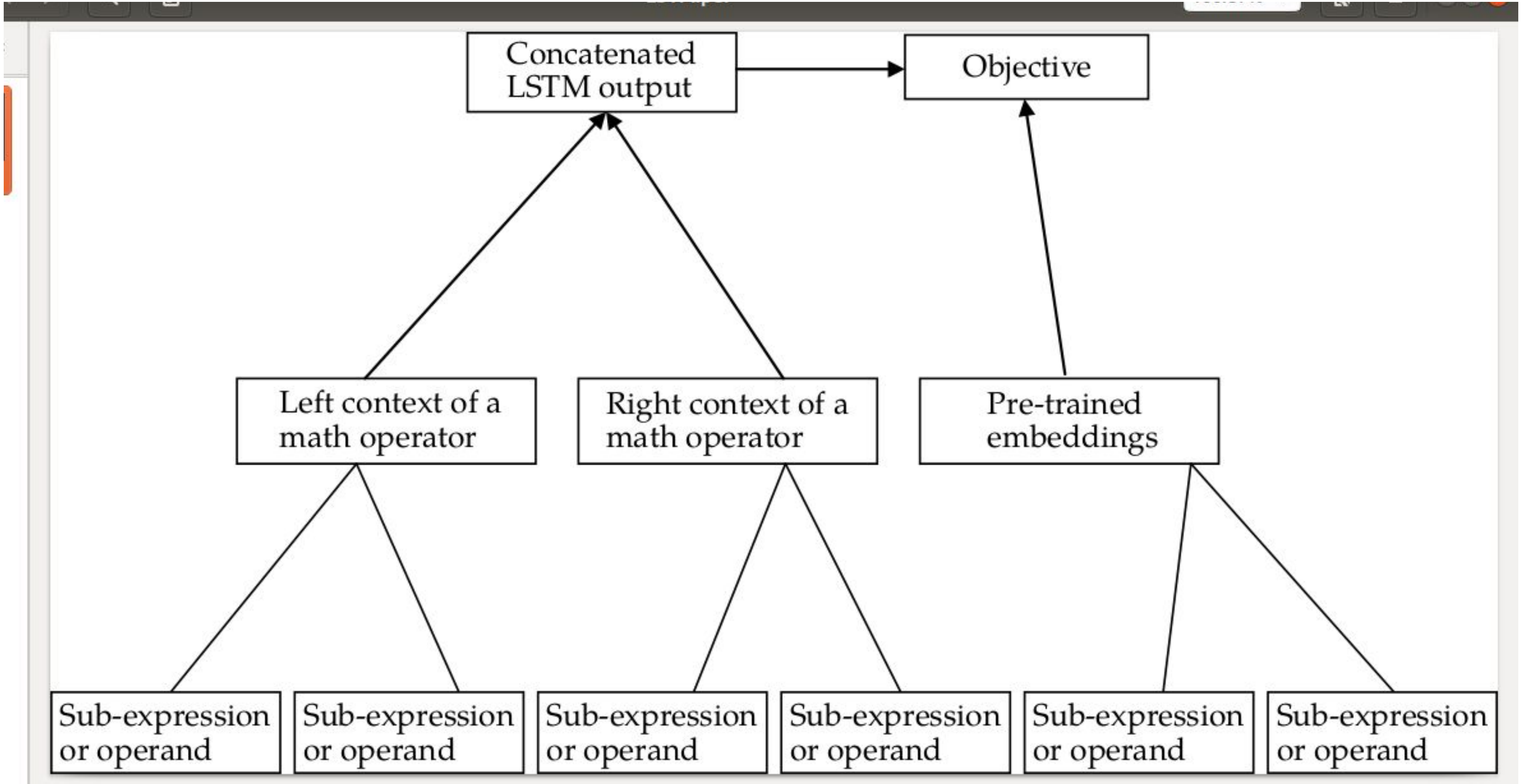
Tree-Structured Bi-LSTM for Maths Embeddings

Proposition 1: Denoting each operator in an expression as a target during training, then allowing the Bi-LSTM to learn left and right of each target taking into account the structured

Proposition 2: Capturing the structure of the entire expression using a grammar parsing tool such as BISON, ANTLR of FLEX

Proposition 3: Each parsed expression will then be converted XML and MathML using LaTeXML

Tree_ Structured Bi-LSTM for Maths Embeddings



Concatenated LSTM output
Objective
Left context of a math operator
Right context of a math operator
Pre-trained embeddings
Sub-expression or operand
Sub-expression or operand
Sub-expression or operand
Sub-expression or operand
Sub-expression or operand
Sub-expression or operand

Representation and Transfer Learning for Math

“In a place where there are no worthy men, strive to be worthy.”

Chapters of the Fathers (Pirkei Avot, 2:5)

Math Understanding - Asking Questions

- Objective is to **explain** the meaning of the formulae “parts” (functions, variables, operators)
- A task is motivated by **Answer Retrieval** on Stack Exchange (ARQMath competition)
- A task can be mapped to **Question Answering** (like SQuAD)

Math Understanding - Asking Questions

Q1: What is the meaning of **k**? / How do you explain **k**?

Q2: What is the meaning of **P** here? / How can be **P** described?

Q3: What is the meaning of **v'**? / What is the interpretation of **v’**?

$$\log \sigma({v'_{w_O}}^\top v_{w_I}) + \sum_{i=1}^{k} \mathbb{E}_{w_i \sim P_n(w)} \left[\log \sigma(-{v'_{w_i}}^\top v_{w_I}) \right]$$

P1: *(...) which is used to replace every log $P(w_O|w_I)$ term in the Skip-gram objective. Thus the task is to distinguish the target word w_O from draws from the noise distribution $P_n(w)$ using logistic regression, where there are k negative samples for each data sample. Our experiments indicate that values of k in the range 5–20 are useful for small training datasets, while for large datasets the k can be as small as 2–5. The main difference between the Negative sampling and NCE is that NCE needs both samples and the numerical probabilities of the noise distribution, while Negative sampling uses only samples. And while NCE approximately maximizes the log probability of the softmax, this property is not important for our application.*

Example 1

Math Understanding - Asking Questions

Q1: What is the meaning of **k**? / How do you explain **k**? ➜ lays on (105, 134) of **P1**

Q2: What is the meaning of **P** here? / How can be **P** described? ➜ lays on (65, 74) of **P1**

Q3: What is the meaning of **v'**? / What is the interpretation of **v'**? ➜ ∅

$$\log \sigma({v'_{w_O}}^\top v_{w_I}) + \sum_{i=1}^{k} \mathbb{E}_{w_i \sim P_n(w)} \left[\log \sigma(-{v'_{w_i}}^\top v_{w_I})\right]$$

P1: *(...) which is used to replace every log $P(w_O|w_I)$ term in the Skip-gram objective. Thus the task is to distinguish the target word w_O from draws from the noise distribution $P_n(w)$ using logistic regression, where there are k negative samples for each data sample. Our experiments indicate that values of k in the range 5–20 are useful for small training datasets, while for large datasets the k can be as small as 2–5. The main difference between the Negative sampling and NCE is that NCE needs both samples and the numerical probabilities of the noise distribution, while Negative sampling uses only samples. And while NCE approximately maximizes the log probability of the softmax, this property is not important for our application.*

Example 1

Math Understanding - Asking Questions

Q1: What is the meaning of **k**? / How do you explain **k**? ➜ ⌀

Q2: What is the meaning of **P** here? / How can be **P** described? ➜ ⌀

Q3: What is the meaning of **v'**? / What is the interpretation of **v'**? ➜ lays on (65, 74) of **P2**

$$\log \sigma({v'_{w_O}}^\top v_{w_I}) + \sum_{i=1}^{k} \mathbb{E}_{w_i \sim P_n(w)} \left[\log \sigma(-{v'_{w_i}}^\top v_{w_I}) \right]$$

P2: *(...) the basic Skip-gram formulation defines $p(w_{t+j} | w_t)$ using the softmax function, where v_w and v'_w are the "input" and "output" vector representations of w, and W is the number of words in the vocabulary. This formulation is impractical because the cost of computing $\nabla \log p(w_O|w_I)$ is proportional to W, which is often large (105–107 terms).*

Example 1

Math Understanding - Asking Questions

- Objective is to explain the meaning of the formulae “parts” (functions, variables, operators)
- A task is motivated by Answer Retrieval on Stack Exchange (ARQMath competition)
- A task can be mapped to **Question Answering** (like SQuAD)
- Still tricky: explanation is rather **rarely explicit**
  - Transitive definitions
  - Inherent meanings
  - Context-dependent

Math Understanding - Reformulating

- Objective is to look for **different interpretation** of the **same formulae**
- A task similar to **Paraphrasing recognition** (like MRPC)

Math Understanding - Reformulating

P1: *Negative sampling is a variation of NCE used by the popular word2vec tool, but it defines the conditional probabilities given (w, c) differently:*

$$p(D=0 \mid c, w) = \frac{1}{u_\theta(w,c)+1}$$

$$p(D=1 \mid c, w) = \frac{u_\theta(w,c)}{u_\theta(w,c)+1}.$$

This objective can be understood in several ways. First, it is equivalent to NCE when k = |V| and q is uniform. Second, it can be understood as the hinge objective of Collobert et al. (2011) where the max function has been replaced with a softmax. As a result, aside from the k = |V| and uniform q case, the conditional probabilities of D given (w, c) are not consistent with the language model probabilities of (w, c) (...)

P2: *An alternative to the hierarchical softmax is Noise Contrastive Estimation (NCE), which was introduced by Gutmann and Hyvarinen [4]. This is similar to hinge loss used by Collobert and Weston [2] who trained the models by ranking the data above noise. NCE can be shown to approximately maximize the log probability of the softmax. We define Negative sampling (NEG) by the*

$$\log \sigma({v'_{w_O}}^\top v_{w_I}) + \sum_{i=1}^{k} \mathbb{E}_{w_i \sim P_n(w)} \left[\log \sigma(-{v'_{w_i}}^\top v_{w_I}) \right]$$

which is used to replace every log $P(w_O|w_I)$ term in the Skip-gram objective.

Example 2

Math Understanding - Reformulating

P1: *Negative sampling is a variation of NCE used by the popular word2vec tool, but it defines the conditional probabilities given (w, c) differently:*

$$p(D = 0 \mid c, w) = \frac{1}{u_\theta(w, c) + 1}$$

$$p(D = 1 \mid c, w) = \frac{u_\theta(w, c)}{u_\theta(w, c) + 1}.$$

This objective can be understood in several ways. First, it is equivalent to NCE when k = |V| and q is uniform. Second, it can be understood as the hinge objective of Collobert et al. (2011) where the max function has been replaced with a softmax. As a result, aside from the k = |V| and uniform q case, the conditional probabilities of D given (w, c) are not consistent with the language model probabilities of (w, c) (...)

P2: *An alternative to the hierarchical softmax is Noise Contrastive Estimation (NCE), which was introduced by Gutmann and Hyvarinen [4]. This is similar to hinge loss used by Collobert and Weston [2] who trained the models by ranking the data above noise. NCE can be shown to approximately maximize the log probability of the softmax. We define Negative sampling (NEG) by the*

$$\log \sigma({v'_{w_O}}^\top v_{w_I}) + \sum_{i=1}^{k} \mathbb{E}_{w_i \sim P_n(w)} \left[\log \sigma(-{v'_{w_i}}^\top v_{w_I}) \right]$$

which is used to replace every log $P(w_O|w_I)$ term in the Skip-gram objective.

Example 2

Math Understanding - Reformulating

- Objective is to look for different interpretation of the same formulae
- A task similar to Paraphrasing recognition (like MRPC)
- Yet, it apparently requires deeper **understanding** of the **context**
- Computationally **demanding**
  - might require quality compromises in favor of feasibility

Conclusion

“It is not incumbent upon you to complete the work, but neither are you at liberty to desist from it.”

Chapters of the Fathers (Pirkei Avot, 2:21)