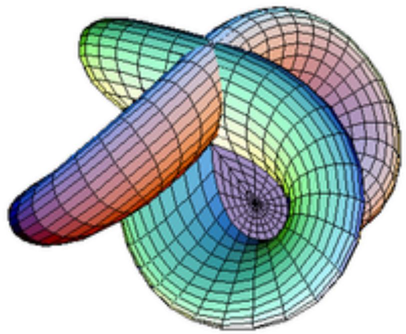
THE LINEAR ALGEBRAIC STRUCTURE OF WORD MEANINGS

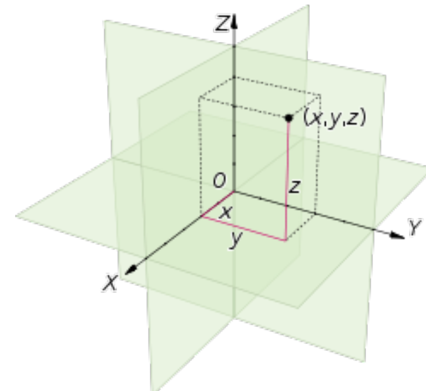
Tengyu Ma

Joint works with Sanjeev Arora, Yuanzhi Li, Yingyu Liang, and Andrej Risteski

Princeton University

EMBEDDINGS (IN MACHINE LEARNING)





$x \in \mathcal{X} \longrightarrow v_x \in \mathbb{R}^d$

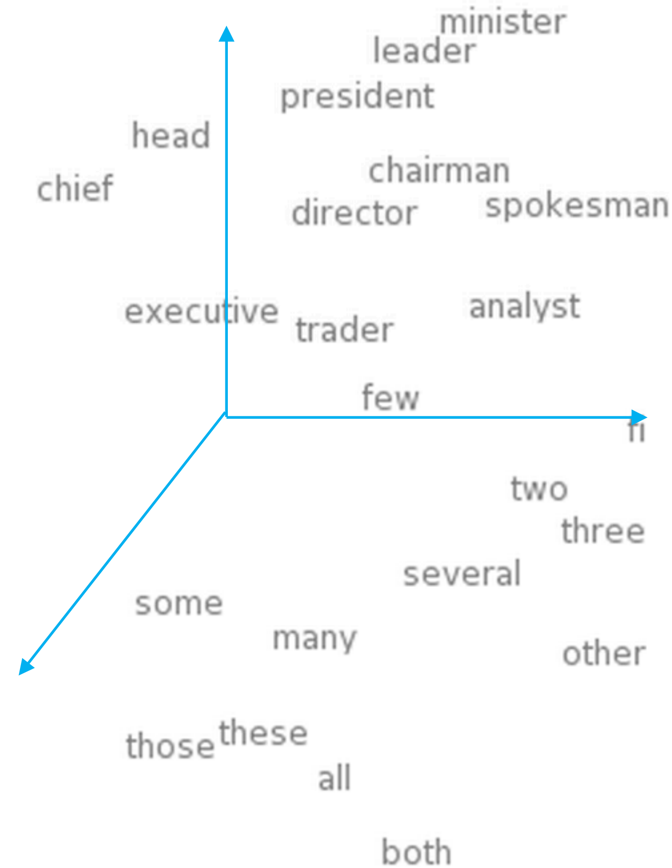
complicated space

Euclidean space with meaningful inner products

- Kernel methods $\xrightarrow{\text{handcrafted lifting}}$ Linearly separable
- Neural nets $\xrightarrow{\text{trained neural nets}}$ Multi-class linear classifier

WORD EMBEDDING

Vocabulary = { 60k most frequent words } $\longrightarrow$ $\mathbb{R}^{300}$



Goal: Embedding captures semantics information

(via linear algebraic operations)

- inner products characterize similarity
  - similar words have large inner products
- differences characterize relationship
  - analogous pairs have similar differences
- more?

picture: Chris Olah's blog

WORD EMBEDDING, AN OLD IDEA

Meaning of a word is determined by words it co-occurs with.

(*Distributional hypothesis of meaning,* [Harris'54], [Firth'57])

$$\begin{array}{cc} & \text{word } y \\ & \downarrow \\ \text{word } x \rightarrow & \begin{bmatrix} & \cdots & \\ \vdots & v_x & \vdots \\ \vdots & \vdots & \vdots \\ & \cdots & \end{bmatrix} \end{array}$$

Co-occurrence matrix
$\Pr(\cdot,\cdot)$

- $\Pr(x, y) \triangleq$ prob. of co-occurrences of x, y in a window of size 5
- $\langle v_x, v_y \rangle$ - a good measure of similarity of (x, y) [Lund-Burgess'96]
- v_x = row of entry-wise square-root of co-occurrence matrix [Rohde et al'05]
- v_x = row of $\mathrm{PMI}(x, y) = \log \frac{\Pr[x,y]}{\Pr[x]\Pr[y]}$ matrix [Church-Hanks'90]

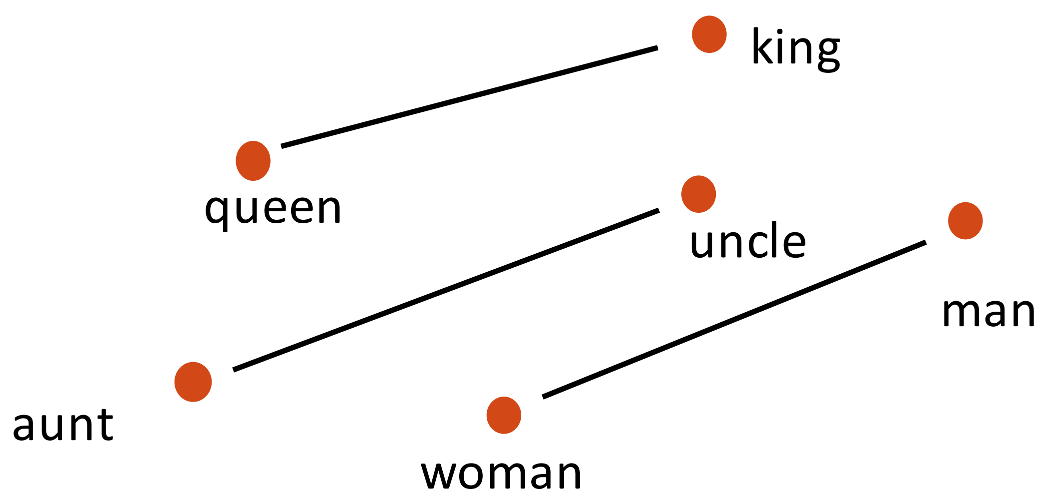
LINEAR STRUCTURE AFTER NON-LINEAR EMBEDDING

Algorithm [Levy-Goldberg]: (dimension-reduction version of [Church-Hanks'90])

- Compute $\mathrm{PMI}(x,y) = \log \frac{\Pr[x,y]}{\Pr[x]\Pr[y]}$
- Take rank-300 SVD (best rank-300 approximation) of PMI
  - $\Leftrightarrow$ Fit $\mathrm{PMI}(x,y) \approx \langle v_x, v_y \rangle$ (with squared loss), where $v_x \in \mathbb{R}^{300}$

- "Linear structure" in the found v_x's :

$$v_{woman} - v_{man} \approx v_{queen} - v_{king} \approx v_{uncle} - v_{aunt} \approx \cdots$$



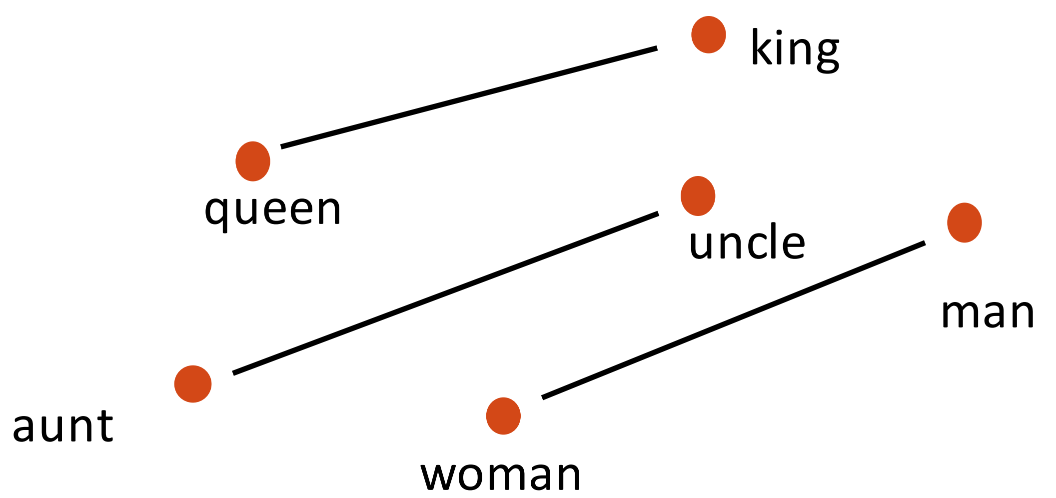
APPLICATIONS/TESTS : SOLVING ANALOGY TASKS

- Questions: woman: man
  queen: ?
  aunt: ?

- Answers:

$$king = \text{argmin}_{\text{w}}||(v_{queen} - v_{\text{w}}) - (v_{woman} - v_{man})||$$

$$aunt = \text{argmin}_{\text{w}}||(v_{uncle} - v_{\text{w}}) - (v_{woman} - v_{man})||$$



NON-LINEAR EMBEDDING METHODS

- recurrent neural network based model [Mikolov et al'12]
- word2vec [Mikolov et al'13] :

$$\Pr[x_{i+6} \mid x_{i+1}, \ldots, x_{i+5}] \propto \exp\langle v_{x_{i+6}}, \frac{1}{5}(v_{x_{i+1}} + \cdots + v_{x_{i+5}})\rangle$$

- GloVe [Pennington et al'14] :

$$\log \Pr[x, y] \approx \langle v_x, v_y \rangle + s_x + s_y + C$$

- [Levy-Goldberg'14] (Previous slide)

$$\mathrm{PMI}(x, y) = \log \frac{\Pr[x,y]}{\Pr[x] \Pr[y]} \approx \langle v_x, v_y \rangle + C$$

Logarithm (or exponential) seems to exclude linear algebra!

Why co-occurrence statistics + log → linear structure

[Levy-Goldberg'13, Pennington et al'14, rephrased]

- For most of the words χ:

$$\frac{\Pr[\chi \mid king]}{\Pr[\chi \mid queen]} \approx \frac{\Pr[\chi \mid man]}{\Pr[\chi \mid woman]}$$

  - For χ unrelated to gender: LHS, RHS ≈ 1
  - for χ =dress, LHS, RHS $\ll 1$; for χ = John, LHS, RHS $\gg 1$

- It suggests

$$\sum_{\chi} \left(\log \frac{\Pr[\chi \mid king]}{\Pr[\chi \mid queen]} - \log \frac{\Pr[\chi \mid man]}{\Pr[\chi \mid woman]} \right)^2 \approx 0$$

$$= \sum_{\chi} \Big(\big(\mathrm{PMI}(\chi, king) - \mathrm{PMI}(\chi, queen)\big) - \big(\mathrm{PMI}(\chi, man) - \mathrm{PMI}(\chi, woman)\big) \Big)^2 \approx 0$$

- Rows of PMI matrix has "linear structure"
- Empirically one can find v_w's s.t. $\mathrm{PMI}(\chi, w) \approx \langle v_\chi, v_w \rangle$
- Suggestion: v_w's also have linear structure

WHY THESE METHODS CAN WORK?

M1: Why do low-dim vectors capture essence of huge co-occurrence statistics? That is, why is a low-dim fit of PMI matrix even possible?

$$\mathrm{PMI}(x,y) \approx \langle v_x, v_y \rangle \qquad (*)$$

- NB: PMI matrix is not necessarily PSD.

M2: Why low-dim vectors solves analogy when $(*)$ is only roughly true?

↑ empirical fit has 17% error

- NB: solving analogy task requires inner products of 6 pairs of word vectors, and that "king" survives against all other words – noise is potentially an issue!

$$king = \mathrm{argmax}_{\mathrm{w}} ||(v_{queen} - v_{w}) - (v_{woman} - v_{man})\,||^2$$

- Fact: low-dim word vectors have more accurate linear structure than the rows of PMI (therefore better analogy task performance).

OUR INSIGHTS

M1: Why do low-dim vectors capture essence of huge co-occurrence statistics? That is, why is a low-dim fit of PMI matrix even possible?

$$\mathrm{PMI}(x, y) \approx \langle v_x, v_y \rangle \qquad (*)$$

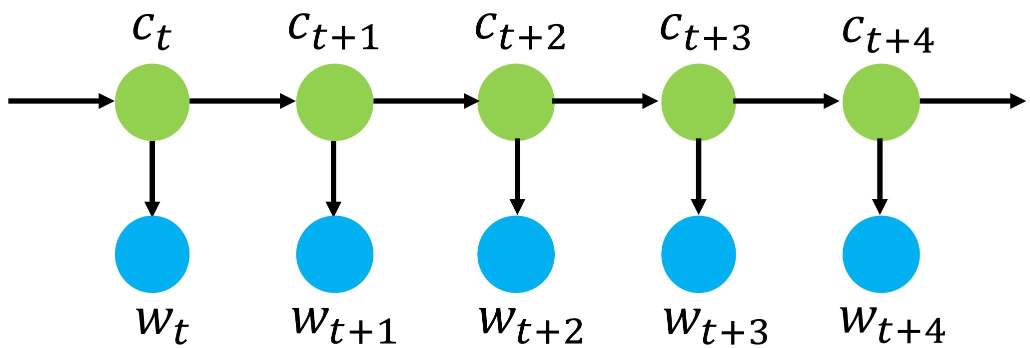
A1: Under a generative model (named RAND-WALK) , (*) provably holds

M2: Why low-dim vectors solves analogy when $(*)$ is only roughly true?

A2: (*) + isotropy of word vectors ⇒ low-dim fitting reduces noise

(Quite intuitive, though doesn't follow Occam's bound for PAC-learning)

RAND-WALK: A GENERATIVE MODEL FOR LANGUAGE

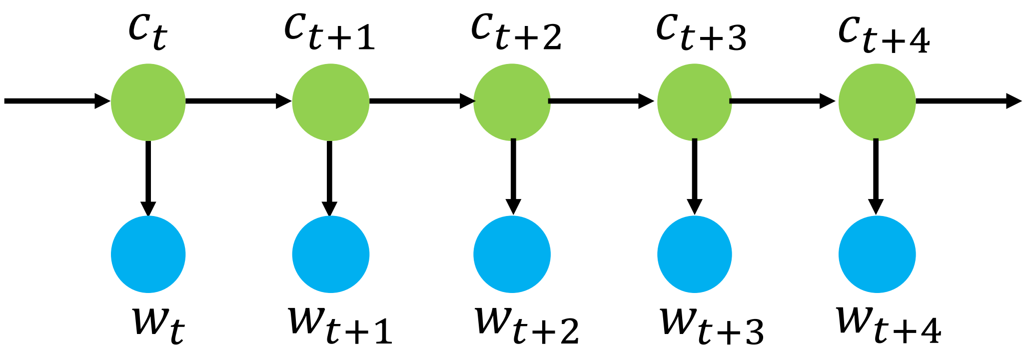


- Hidden Markov Model:
  - discourse vector $c_t \in \mathbb{R}^d$ governs the discourse/theme/context of time t
  - words w_t (observable); embedding $v_{w_t} \in \mathbb{R}^d$ (parameters to learn)
  - log-linear observation model

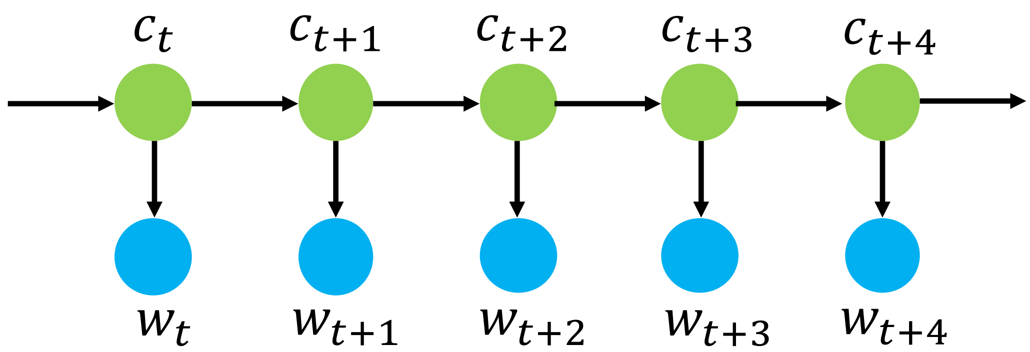
$$\Pr[w_t \mid c_t] \propto \exp\langle v_{w_t}, c_t\rangle$$

- Closely related to [Mnih-Hinton'07]

RAND-WALK: A GENERATIVE MODEL FOR LANGUAGE(CONT'D)



- Ideally, $c_t, v_w \in \mathbb{R}^d$ should contain semantic information in its coordinates
  - E.g. (0.5, -0.3, …) could mean "0.5 gender, -0.3 age,.."
- But, the whole system is rotational invariant: $\langle c_t, v_w \rangle = \langle Rc_t, Rv_w \rangle$
- There should exist a rotation so that the coordinates are meaningful (back to this later)



- Assumptions:
  - $\{v_w\}$ consists of vectors drawn from $s \cdot \mathcal{N}(0, \mathrm{Id})$; s is bounded scalar r.v.
  - c_t does a slow random walk (doesn't change much in a window of 5)
  - log-linear observation model: $\Pr[w_t \mid c_t] \propto \exp\langle v_{w_t}, c_t \rangle$
- Main Theorem:

(1) $$\log \Pr[w, w'] = \|v_w + v_{w'}\|^2 / d - 2 \log Z \pm \epsilon$$

(2) $$\log \Pr[w] = \|v_w\|^2 / d - \log Z \pm \epsilon$$

(3) $$\mathrm{PMI}(w, w') = \langle v_w, v_{w'} \rangle / d \pm \epsilon$$

Fact: (2) implies that the words have power law dist.

- Norm determines frequency; spatial orientation determines "meaning"

EXPLAINING EXISTING METHODS

- word2vec [Mikolov et al'13] :

$$\Pr[w_{i+6} \mid w_{i+1}, \ldots, w_{i+5}] \propto \exp\langle v_{w_{i+6}}, \frac{1}{5}(v_{w_{i+1}} + \cdots + v_{w_{i+5}})\rangle$$

- GloVe [Pennington et al'14] :

$$\log \Pr[w, w'] \approx \langle v_w, v_{w'} \rangle + s_w + s_{w'} + C$$

Eq. (1) $$\log \Pr[w, w'] = \|v_w + v_{w'}\|^2/d - 2\log Z \pm \epsilon$$

- [Levy-Goldberg'14]

$$\mathrm{PMI}(w, w') \approx \langle v_w, v_{w'} \rangle + C$$

Eq. (3) $$\mathrm{PMI}(w, w') = \langle v_w, v_{w'} \rangle / d \pm \epsilon$$

EXPLAINING EXISTING METHODS CONT'D

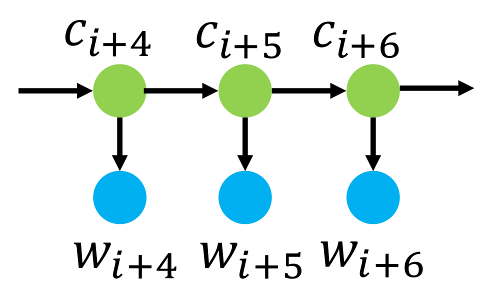
- word2vec [Mikolov et al'13] :

$$\Pr[\, w_{i+6} \mid w_{i+1}, \dots, w_{i+5} \,] \propto \exp\langle v_{w_{i+6}}, \frac{1}{5}(v_{w_{i+1}} + \cdots + v_{w_{i+5}})\rangle$$

↑

max-likelihood estimate of c_{i+6}

- Under our model,



  - Random walk is slow: $c_{i+1} \approx c_{i+2} \approx \cdots \approx c_{i+6} \approx c$
  - Best estimate for current discourse c_{i+6} :

$$\underset{c, ||c||=1}{\operatorname{argmax}} \Pr[c \mid w_{i+1}, \dots, w_5] = \alpha\left(v_{w_{i+1}} + \cdots + v_{w_{i+5}}\right)$$

  - Prob. distribution of next word given the best guess c:

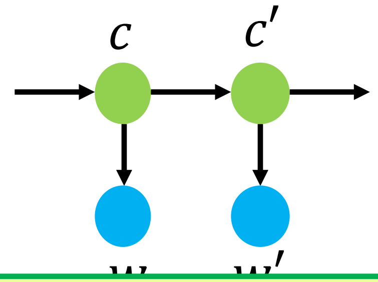
$$\Pr[w_{i+6} \mid c_{i+6} = \alpha(v_{w_{i+1}} + \cdots + v_{w_{i+5}})] \propto \exp\langle v_{w_{i+6}}, \alpha\,(v_{w_{i+1}} + \cdots + v_{w_{i+5}})\rangle$$

PROOF SKETCH OF MAIN THM.

This talk: window of size 2

$$\Pr[w \mid c] \propto \exp\langle v_w, c\rangle$$

$$\Pr[w' \mid c'] \propto \exp\langle v_{w'}, c'\rangle$$



- $\Pr[w \mid c] = \frac{1}{Z_c} \cdot \exp\langle v_w, c\rangle$
- $Z_c = \sum_w \exp\langle v_w, c\rangle$ partition function

$$\Pr[w, w'] = \int \Pr[w \mid c] \Pr[w' \mid c']\, p(c, c') dc dc'$$

$$= \int \underbrace{\frac{1}{Z_c Z_{c'}}}_{??} \cdot \underbrace{\exp\langle v_w, c\rangle \exp\langle v_{w'}, c'\rangle\, p(c,}$$

spherical Gaussian vector c

- $\mathbb{E}[\exp\langle v, c\rangle] = \exp \|v\|^2/d$

- Assume $c = c'$ with probability 1,

$$= \int \exp\langle v_w + v_{w'}, c\rangle\, p(c) dc = \exp\|v_w + v_{w'}\|^2/d$$

Eq. (1) $\log \Pr[w, w'] = \|v_w + v_{w'}\|^2/d - 2\log Z \pm \epsilon$

PROOF SKETCH OF MAIN THM CONT'D

- $\Pr[w \mid c] = \frac{1}{Z_c} \cdot \exp\langle v_w, c\rangle$
- $Z_c = \sum_w \exp\langle v_w, c\rangle$ partition function

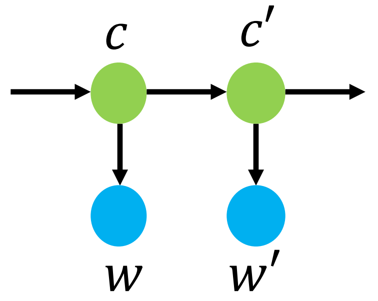
Lemma 1: for almost all c, almost all $\{v_w\}$,

$$Z_c = \big(1 + o(1)\big)Z$$

This talk: window of size 2

$$\Pr[w \mid c] \propto \exp\langle v_w, c\rangle$$

$$\Pr[w' \mid c'] \propto \exp\langle v_{w'}, c'\rangle$$



- Proof (sketch) :
  - for most c, Z_c concentrates around its mean
  - mean of Z_c is determined by $||c||$, which in turn concentrates
  - caveat: $\exp\langle v, c\rangle$ for $v \sim \mathcal{N}(0, \mathrm{Id})$ is not subgaussian, nor sub-exponential. (α-Orlicz norm is not bounded for any $\alpha > 0$)

Eq. (1) $\log \Pr[w, w'] = \|v_w + v_{w'}\|^2 / d - 2 \log Z \pm \epsilon$

A HEAVY TAIL PHENOMENON

Lemma 1: for almost all c, almost all $\{v_w\}$,

$$Z_c = \big(1 + o(1)\big)Z$$

- Proof Sketch:
- Fixing c, to show high probability over choices of v_w's

$$Z_c = \sum_w \exp\langle v_w, c\rangle = \big(1 + o(1)\big)\mathbb{E}[Z_c]$$

- $z_w = \langle v_w, c\rangle$ scalar Gaussian random variable
- $||c||$ governs the mean and variance of z_w.
- $||c||$ in turns is concentrated

A HEAVY TAIL PHENOMENON

- Question: $z_1, \dots, z_n \sim \mathcal{N}(0,1)$

$$Z = \sum_{i=1}^{n} \exp(z_i)$$

- How is Z concentrated?

- $\mathbb{E}[Z_c] = \Theta(n)$, and $\mathbb{V}ar[Z_c] = O(n)$
- The tail of $exp(z_i)$ is bad!
  - $\Pr[\exp z_i > t] \approx t^{-\log t}$

- (sub)-Gaussian tail

  $\Pr[X > t] \le \exp(-t^2/2)$
- (sub)-exponential tail

  $\Pr[X > t] \le \exp(-t/2)$

- Claim:

$$\Pr[Z > \mathbb{E}Z + C\sqrt{n} \cdot \log n] \le \exp(-\log^2 n)$$

- Trick: truncate z_i at $\log n$ and deal with the tail by union bound

A HEAVY TAIL PHENOMENON

Lemma 1: for almost all c, almost all $\{v_w\}$,

$$Z_c = \big(1 + o(1)\big)Z$$

- Proof Sketch:
- Fixing c, we have with high probability over choices of v_w's

$$Z_c = \sum_w \exp\langle v_w, c\rangle = \big(1 + o(1)\big)\mathbb{E}[Z_c]$$

- $z_w = \langle v_w, c\rangle$ scalar Gaussian random variable
- $||c||$ governs the mean and variance of z_w.
- $||c||$ in turns is concentrated

PROOF SKETCH OF MAIN THM CONT'D

- $\Pr[w \mid c] = \frac{1}{Z_c} \cdot \exp\langle v_w, c \rangle$
- $Z_c = \sum_w \exp\langle v_w, c \rangle$ partition function

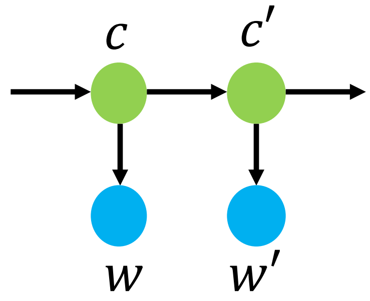
Lemma 1: for almost all c, almost all $\{v_w\}$,

$$Z_c = \big(1 + o(1)\big) Z$$

This talk: window of size 2

$$\Pr[w \mid c] \propto \exp\langle v_w, c \rangle$$

$$\Pr[w' \mid c'] \propto \exp\langle v_{w'}, c' \rangle$$



$$\Pr[w, w'] = \int \frac{1}{Z_c Z_{c'}} \cdot \exp\langle v_w + v_{w'}, c \rangle p(c) dc$$

$$= \big(1 \pm o(1)\big) \frac{1}{Z^2} \int \exp\langle v_w + v_{w'}, c \rangle p(c) dc$$

$$= \big(1 \pm o(1)\big) \frac{1}{Z^2} \exp(\|v_w + v_{w'}\|^2 / d)$$

Eq. (1) $\log \Pr[w, w'] = \|v_w + v_{w'}\|^2 / d - 2 \log Z \pm \epsilon$

MODEL VERIFICATION

- Our theory predicts

  Eq. (1) $\log \Pr[w, w'] = \|v_w + v_{w'}\|^2 / d - 2 \log Z \pm \epsilon$

- (Approximate) maximum likelihood objective (**SN**)

$$\min_{\{v_w\},Y} \sum_{w,w'} \widehat{\Pr}[w, w'] \left(\log \widehat{\Pr}[w, w'] - \|v_w + v_{w'}\|^2 - Y\right)^2$$

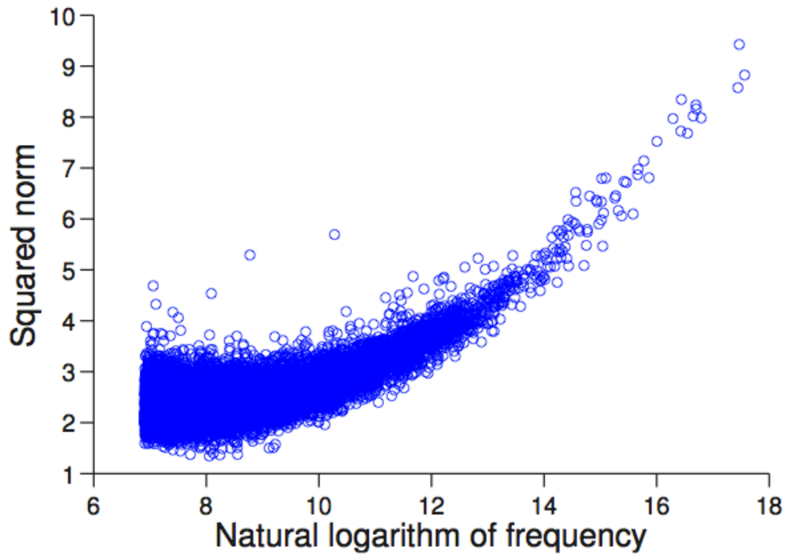
Simplest word embedding method yet (fewest "knobs" to turn)
Comparable performance on analogy test

	Relations	**SN**	GloVe	CBOW	skip-gram
G	semantic	0.84	0.85	0.79	0.73
	syntactic	0.61	0.65	0.71	0.68
	total	0.71	0.73	0.74	0.70
M	adjective	0.50	0.56	0.58	0.58
	noun	0.69	0.70	0.56	0.58
	verb	0.48	0.53	0.64	0.56
	total	0.53	0.57	0.62	0.57

MODEL VERIFICATION CONT'D

- Our theory predicts

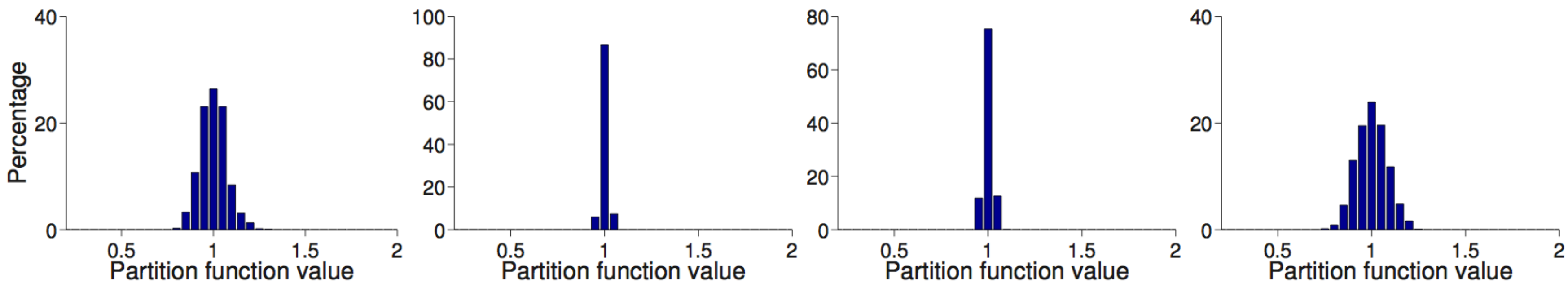
Eq. (2) $$\log \Pr[w] = \|v_w\|^2/d - \log Z \pm \epsilon$$



MODEL VERIFICATION CONT'D

- Our theory predicts

$$Z_c = \big(1 \pm o(1)\big)Z$$



(a) **SN** (b) GloVe (c) CBOW (d) skip-gram

WRAP UP

- Under generative model RANK-WALK

For most of the words χ:

$$\frac{\Pr[\chi \mid a]}{\Pr[\chi \mid b]} \approx \frac{\Pr[\chi \mid c]}{\Pr[\chi \mid d]} \quad \Longleftrightarrow \quad v_a - v_b \approx v_c - v_d$$

↑ semantic def. of analogy

↑ algebraic def. of analogy

- Beyond only solving analogy task?
- Extracting more information from analogy/embeddings?

Some recent work:

Extracting different meanings from word embeddings
(same team: Arora, Li, Liang, M., Risteski)

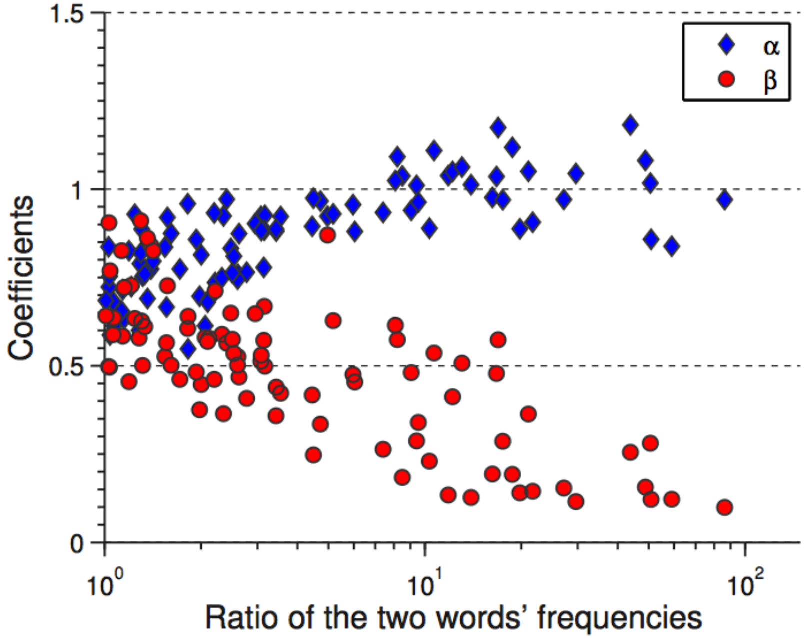
POLYSEMY

- "Tie" can mean article of clothing, or physical act

- *Tie* represents unrelated words tie_1, tie_2, etc.

Quick experiment: Take two random/unrelated words w_1, w_2 where w_1 is ~100 times more frequent than w_2 . Declare these to be a single word and compute its embedding in our model.

Result: close to something like $0.8v_{w_1} + 0.2v_{w_2}$



- Mathematical explanation
- Merge w_1, w_2 as w. Let $r = \frac{\Pr[w_1]}{\Pr[w_2]} > 1$
- Then $v_w \approx \alpha v_{w_1} + \beta v_{w_2}$, where
  - $\alpha = 1 - c_1 \log\left(1 + \frac{1}{r}\right) \approx 1$
  - $\beta = 1 - c_2 \log r$
- $\beta > .1$ even if $r = 100$!
- Rare meaning is not swamped, thanks to the $\log$!

EXTRACTING DIFFERENT MEANINGS

- "Tie" can mean article of clothing, or physical act

- *Tie* represents unrelated w

which correspond to differen

$$v_{tie} = 0.8a_1 + 0.2\, a$$

↑ discourse for tie_1

discc for ti

- Training objective:

$$\min_{\substack{A=[a_1,\ldots,a_m] \\ sparse\ x_w's}} \sum_w \|v_w - Ax_w\|^2$$

- local search algo. [EAB'05], provable algo. [SWW'12, AGM'14, AGMM'15..]

- Sparse coding for extracting different mean
  - Find $m = 2000$ "discourses" $a_1, a_2, \ldots \in \mathbb{R}$ expressed as weighted sum of at most 5 of them, plus "noise vector."

$$v_w = x_{w,1}a_1 + x_{w,2}a_2 + \ \ldots + \ \ noise$$

x_w has only 5 non-zeros

Representative subset of 2000 discourses (represented using their nearest words)

Atom 1978	825	231	616	1638	149	330
drowning	instagram	stakes	membrane	slapping	orchestra	conferences
suicides	twitter	thoroughbred	mitochondria	pulling	philharmonic	meetings
overdose	facebook	guineas	cytosol	plucking	philharmonia	seminars
murder	tumblr	preakness	cytoplasm	squeezing	conductor	workshops
poisoning	vimeo	filly	membranes	twisting	symphony	exhibitions
commits	linkedin	fillies	organelles	bowing	orchestras	organizes
stabbing	reddit	epsom	endoplasmic	slamming	toscanini	concerts
strangulation	myspace	racecourse	proteins	tossing	concertgebouw	lectures
gunshot	tweets	sired	vesicles	grabbing	solti	presentations

↑

closest words to a_{231}

5 atoms that express v_{tie}

Atom 1005	31	1561	2060	1563
trousers	season	scoreline	wires	operatic
blouse	teams	goalless	cables	soprano
waistcoat	winning	equaliser	wiring	mezzo
skirt	league	clinching	electrical	contralto
sleeved	finished	scoreless	wire	baritone
pants	championship	replay	cable	coloratura

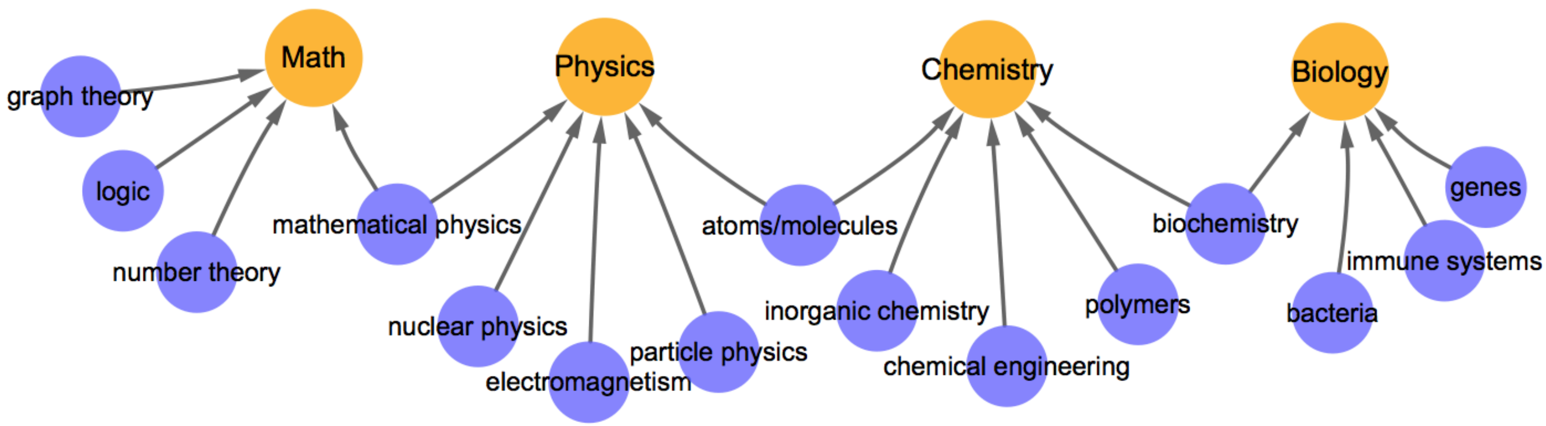
MULTI-LAYER SPARSE CODING

- Atoms of discourse found are fairly fine-grained
- Maybe $a_{biochemistry} = \alpha \cdot b_{biology} + \beta \cdot b_{chemistry}$?
- Another layer:

$$\min_{B,Y\ sparse} ||A - BY||^2$$

411
acids
amino
biosynthesis
peptide
biochemistry

MULTI-LAYER SPARSE CODING CONT'D



Atom	28	2016	468	1318	411
	logic deductive propositional semantics	graph subgraph bipartite vertex	boson massless particle higgs	polyester polypropylene resins epoxy	acids amino biosynthesis peptide
tag	*logic*	*graph theory*	*particle physics*	*polymer*	*biochemistry*

CONCLUSIONS

- Part I: new generative model that captures semantics.
- Provable guarantee:
  - log of co-occurrence matrix has low rank structure
  - semantic analogy ⇔ linear algebraic structure for word vectors
- Simplistic assumptions, but good fit to reality

- Part II: automatic way of detect word meanings
  - Hierarchical basis in the embedding space

- Other applications of our model/method?

AN IDEAL SCENARIO

- Each ordinate of v_w means something:

$$\begin{array}{cccc} \text{currency} & \text{country} & \text{American} & \text{Chinese} \\ \downarrow & \downarrow & \downarrow & \downarrow \end{array}$$

$$v_{USA} = [\dots\dots, 0, \dots\dots\dots, 1, \dots\dots\dots, 1, \dots\dots\dots, 0, \dots\dots]$$

$$v_{dollar} = [\dots\dots, 1, \dots\dots\dots, 0, \dots\dots\dots, 1, \dots\dots\dots, 0, \dots\dots]$$

$$v_{China} = [\dots\dots, 0, \dots\dots\dots, 1, \dots\dots\dots, 0, \dots\dots\dots, 1, \dots\dots]$$

$$v_{RMB} = [\dots\dots, 1, \dots\dots\dots, 0, \dots\dots\dots, 0, \dots\dots\dots, 1, \dots\dots]$$

- On other coordinates, the values are either very small or the supports are non-overlapping

$$v_{USA} - v_{dollar} = [\dots\dots, -1, \dots\dots\dots, 1, \dots\dots\dots, 0 \quad \dots\dots\dots, 0, \dots\dots]$$

$$v_{China} - v_{RMB} = [\dots\dots, -1, \dots\dots\dots, 1, \dots\dots\dots, 0, \dots\dots\dots, 0, \dots\dots]$$

- Problem: rotational invariance – rotation of word vectors doesn’t change the model.

REVISED IDEA: SPARSE CODING

$$
\begin{array}{rl}
 & \quad\;\;\text{currency} \qquad\; \text{country} \qquad \text{American} \quad \text{Chinese} \\
 & \qquad\quad\downarrow \qquad\qquad\quad \downarrow \qquad\qquad\quad \downarrow \qquad\qquad\quad \downarrow \\
v_{USA} = & [\ldots\ldots, 0, \ldots\ldots\ldots, 1, \ldots\ldots\ldots, 1, \ldots\ldots\ldots, 0, \ldots\ldots] \\
v_{dollar} = & [\ldots\ldots, 1, \ldots\ldots\ldots, 0, \ldots\ldots\ldots, 1, \ldots\ldots\ldots, 0, \ldots\ldots] \\
v_{China} = & [\ldots\ldots, 0, \ldots\ldots\ldots, 1, \ldots\ldots\ldots, 0, \ldots\ldots\ldots, 1, \ldots\ldots] \\
v_{RMB} = & [\ldots\ldots, 1, \ldots\ldots\ldots, 0, \ldots\ldots\ldots, 0, \ldots\ldots\ldots, 1, \ldots\ldots]
\end{array}
\cdot \left[\qquad R \qquad \right]
$$

$\uparrow$ sparse coefficients

$\uparrow$ basis vectors

- With sparsity, the model is identifiable; allows overcomplete basis; is tractable under mild assumptions. [SWW'12] [AGM'13][AAJNT'13][AGMM'14]

EXPERIMENTS

$$\min_{X \text{ sparse}, R} ||V - X \cdot R||_F^2$$

- V contains word vectors as rows (obtained from any embedding method)
- Sparsity of rows of X is chosen to be 5
- R contains 2000 basis vectors (as rows), each of which is 300-dim

RESOLVING MYSTERY 2

Assuming M1 was answered,

$$\mathrm{PMI}(w, w') = \langle v_w, v_{w'} \rangle + \xi \qquad (*)$$

with large ξ

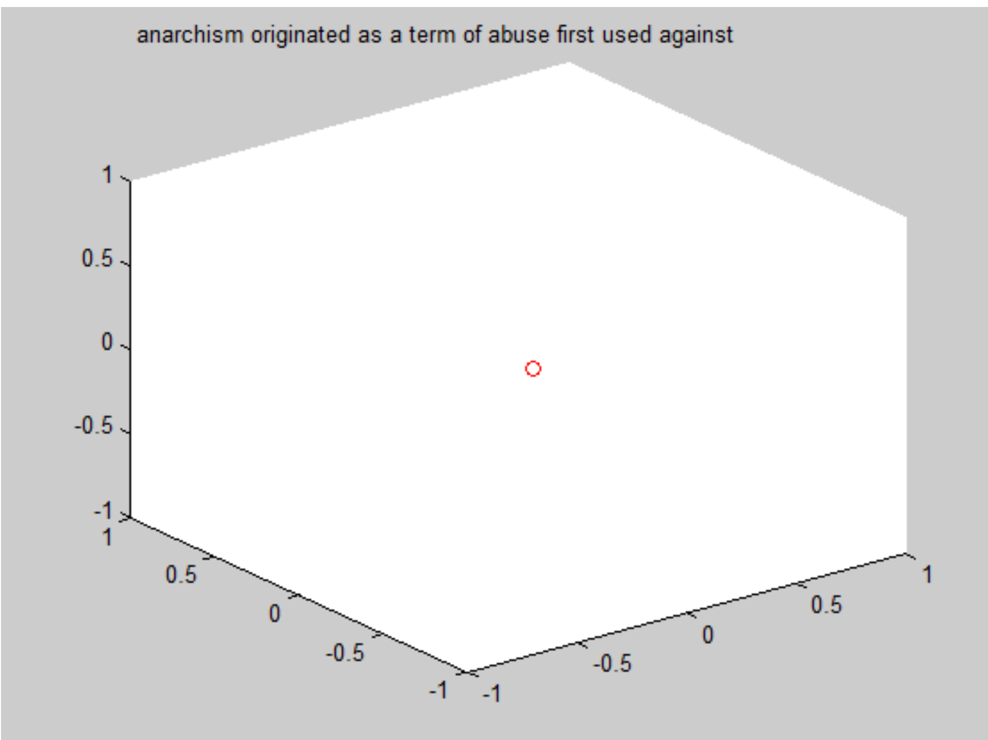
M2: Why low-dim vectors solves analogy when (*) is only roughly true?

A2: (*) + isotropy of word vectors $\Rightarrow$ low-dim fitting reduces noise

(Quite intuitive, though doesn't follow Occam's bound for PAC-learning)

SLOW RANDOM WALK ILLUSTRATION

- Our theory assumes that c_t does a slow random walk



- red dot: the estimate hidden variable c_t at time t
- sentence at top: the window of size 10 at time t

Resolving Mystery 2

Assuming M1 was answered,

$$\mathrm{PMI}(w, w') = \langle v_w, v_{w'} \rangle + \xi \qquad (*)$$

with large ξ

M2: Why low-dim vectors solves analogy when (*) is only roughly true?

A2: (*) + isotropy of word vectors $\Rightarrow$ low-dim fitting reduces noise

(Quite intuitive, though doesn't follow Occam's bound for PAC-learning)