

Neural Networks (of language)

EECS 348: Intro to Artificial
Intelligence
Spring 2014

Brains!



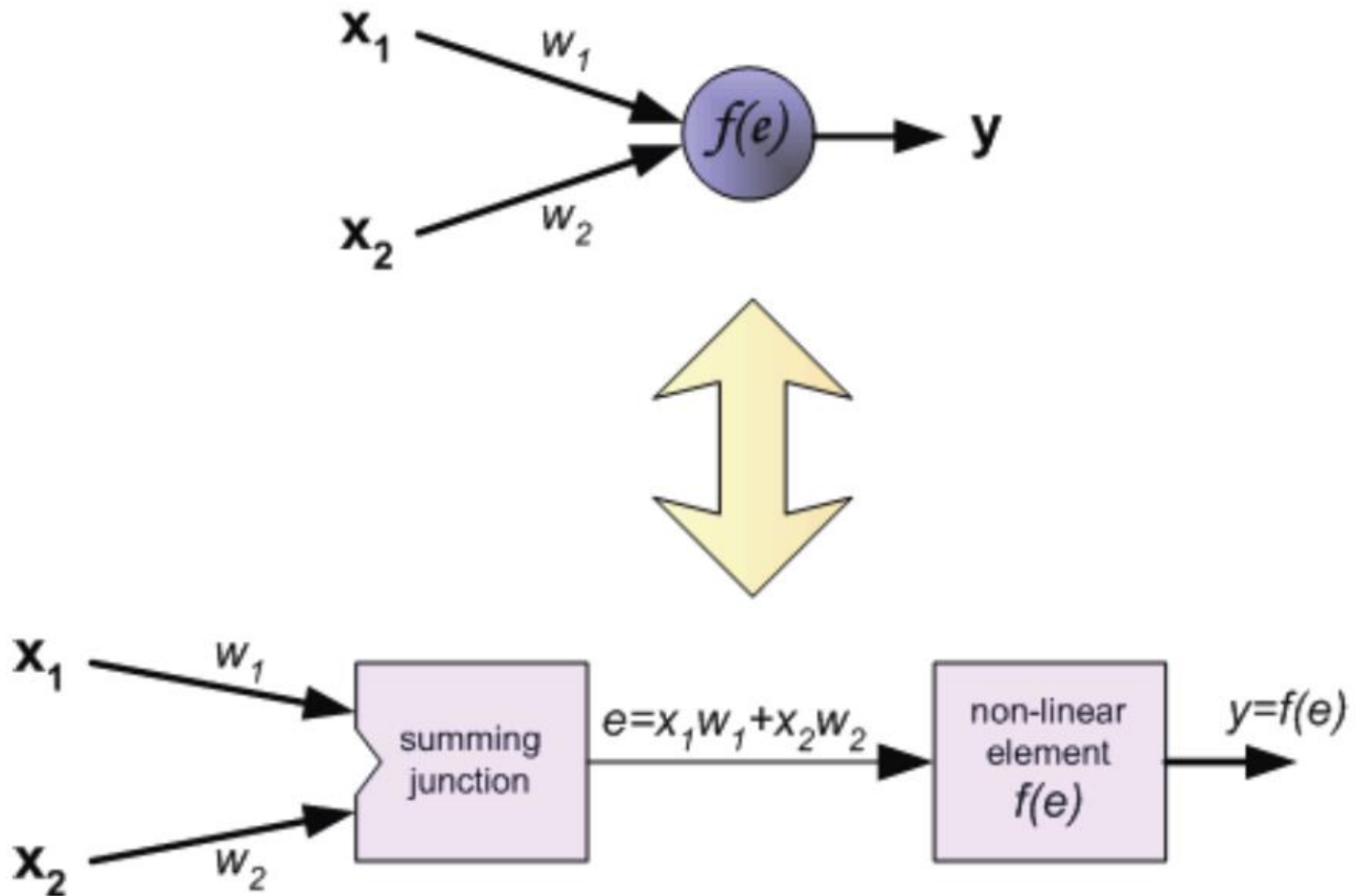
- Intelligent machines

Brain architecture

- Neurons
 - Basic building block of brains
 - Humans have 100 billion
 - Given enough input => voltage spikes, neuron sends output to neighbors
- Synapses
 - Connect neurons
 - Neurons excite or inhibit their neighbors
- **Fault tolerant & massively parallel** (CPUs headed this way too)
- Massively interconnected (100 to 10K neighbors per neuron)
- Slow switching time vs. CPUs



Basic Artificial Neuron



From Manaal Faruqui

Neural Network Language Models

- Statistical Language Modeling:
 - Predict probability of next word in sequence

I was headed to Madrid , _____

$P(\text{_____} = \text{"Spain"}) = 0.5,$

$P(\text{_____} = \text{"but"}) = 0.2, \text{ etc.}$

- Used in speech recognition, machine translation, (recently) information extraction

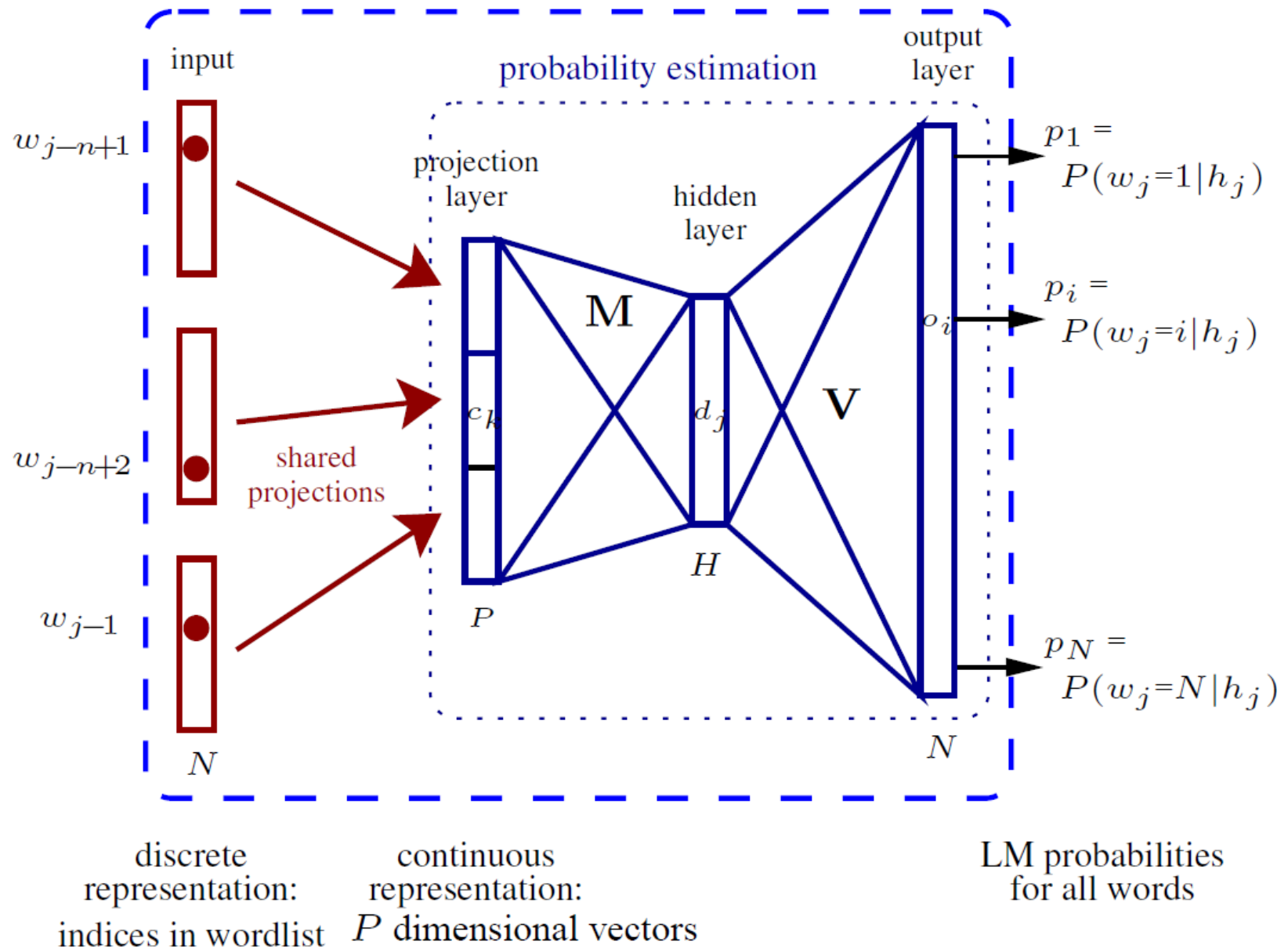
Formally

- Estimate:

$$P(w_j \mid w_{j-1}, w_{j-2}, \dots, w_{j-n+1})$$

$$= P(w_j \mid h_j)$$

Neural Network



Optimizations

- Key idea – learn simultaneously:
 - vector representations of each word (e.g., 120 dim)
 - predictor of next word. based on previous vectors
- Short-lists
 - Much complexity in hidden->output layer
 - Number of possible next words is large
 - Only predict a *subset* of words
 - Use a standard probabilistic model for the rest

Word prediction performance

- <http://arxiv.org/abs/1312.3005>

Model	Num. Params [billions]	Training Time		Perplexity
		[hours]	[CPUs]	
Interpolated KN 5-gram, 1.1B n-grams (KN)	1.76	3	100	67.6
Katz 5-gram, 1.1B n-grams	1.74	2	100	79.9
Stupid Backoff 5-gram (SBO)	1.13	0.4	200	87.9
Interpolated KN 5-gram, 15M n-grams	0.03	3	100	243.2
Katz 5-gram, 15M n-grams	0.03	2	100	127.5
Binary MaxEnt 5-gram (n-gram features)	1.13	1	5000	115.4
Binary MaxEnt 5-gram (n-gram + skip-1 features)	1.8	1.25	5000	107.1
Hierarchical Softmax MaxEnt 4-gram (HME)	6	3	1	101.3
Recurrent NN-256 + MaxEnt 9-gram	20	60	24	58.3
Recurrent NN-512 + MaxEnt 9-gram	20	120	24	54.5
Recurrent NN-1024 + MaxEnt 9-gram	20	240	24	51.3

LMs for Analogies

Type of relationship	Word Pair 1		Word Pair 2	
Common capital city	Athens	Greece	Oslo	Norway
All capital cities	Astana	Kazakhstan	Harare	Zimbabwe
Currency	Angola	kwanza	Iran	rial
City-in-state	Chicago	Illinois	Stockton	California
Man-Woman	brother	sister	grandson	granddaughter
Adjective to adverb	apparent	apparently	rapid	rapidly
Opposite	possibly	impossibly	ethical	unethical
Comparative	great	greater	tough	tougher
Superlative	easy	easiest	lucky	luckiest
Present Participle	think	thinking	read	reading
Nationality adjective	Switzerland	Swiss	Cambodia	Cambodian
Past tense	walking	walked	swimming	swam
Plural nouns	mouse	mice	dollar	dollars
Plural verbs	work	works	speak	speaks

Analogies Performance

Dimensionality / Training words	24M	49M	98M	196M	391M	783M
50	13.4	15.7	18.6	19.1	22.5	23.2
100	19.4	23.1	27.8	28.7	33.4	32.2
300	23.2	29.2	35.3	38.6	43.7	45.9
600	24.0	30.1	36.5	40.8	46.6	50.4

- <http://arxiv.org/pdf/1301.3781.pdf>

Analogies Performance

Model Architecture	Semantic-Syntactic Word Relationship test set		MSR Word Relatedness Test Set [20]
	Semantic Accuracy [%]	Syntactic Accuracy [%]	
RNNLM	9	36	35
NNLM	23	53	47
CBOW	24	64	61
Skip-gram	55	59	56

- <http://arxiv.org/pdf/1301.3781.pdf>