Modelling frameworks in Cognitive Science

Sumeet Agarwal

ELL457/HSL622

April 13, 2023

Preface

What is Cognitive Science?

- The science of understanding the mind

What is Cognitive Science?

- The science of understanding the mind
- Aims at formal modelling of cognitive processes

What is Cognitive Science?

- The science of understanding the mind
- Aims at formal modelling of cognitive processes
- Has typically been characterised by a strong emphasis on empirical and computational approaches

What is Cognitive Science?

- The science of understanding the mind
- Aims at formal modelling of cognitive processes
- Has typically been characterised by a strong emphasis on empirical and computational approaches
- In the context of linguistics, cognitive science is essentially the same as computational psycholinguistics

What is Cognitive Science?

- The science of understanding the mind
- Aims at formal modelling of cognitive processes
- Has typically been characterised by a strong emphasis on empirical and computational approaches
- In the context of linguistics, cognitive science is essentially the same as computational psycholinguistics
- Here I will try to look at some of the broader computational trends currently prominent in cognitive science, with a particular emphasis on how they could be relevant for modelling langauge cognition

Themes

Given the breadth of the topic, here I will focus on just two kinds of cognitive modelling frameworks.

- **Neural network models** (aka Connectionist models, Deep learning)

Themes

Given the breadth of the topic, here I will focus on just two kinds of cognitive modelling frameworks.

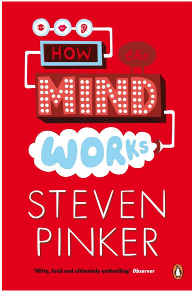
- **Neural network models** (aka Connectionist models, Deep learning)
- **Bayesian models of cognition** (aka Bayesian cognitive science)

Part I: Neural network models

Neural networks as computational systems

- The classic mathematical model of the neuron is McCulloch-Pitts (1943)

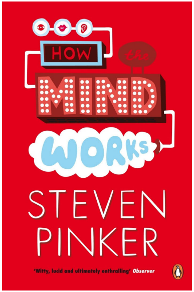
[Pinker, 1999]



Neural networks as computational systems

- The classic mathematical model of the neuron is McCulloch-Pitts (1943)
- Sees neurons as switch-like, either ON (1) or OFF (0)

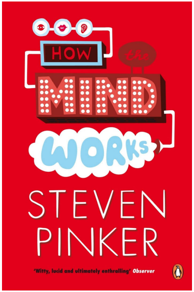
[Pinker, 1999]



Neural networks as computational systems

- The classic mathematical model of the neuron is McCulloch-Pitts (1943)
- Sees neurons as switch-like, either ON (1) or OFF (0)
- Each neuron takes a weighted sum of inputs and applies a threshold to it, to decide whether to fire or not

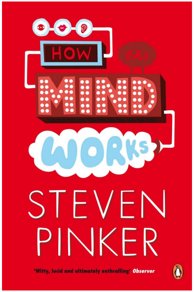
[Pinker, 1999]



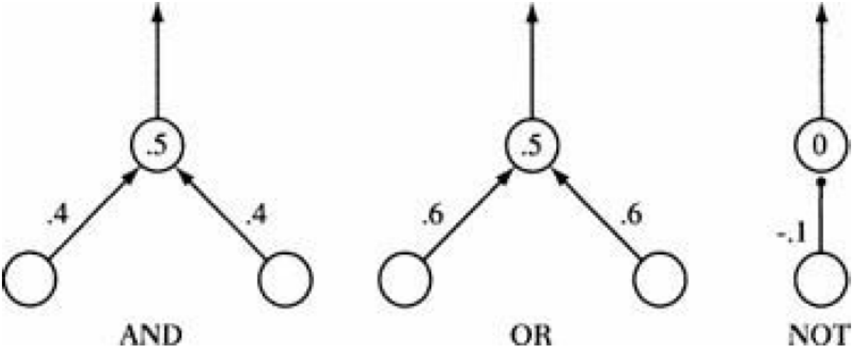
Neural networks as computational systems

- The classic mathematical model of the neuron is McCulloch-Pitts (1943)
- Sees neurons as switch-like, either ON (1) or OFF (0)
- Each neuron takes a weighted sum of inputs and applies a threshold to it, to decide whether to fire or not
- They can thus encode more abstract logical operations

[Pinker, 1999]



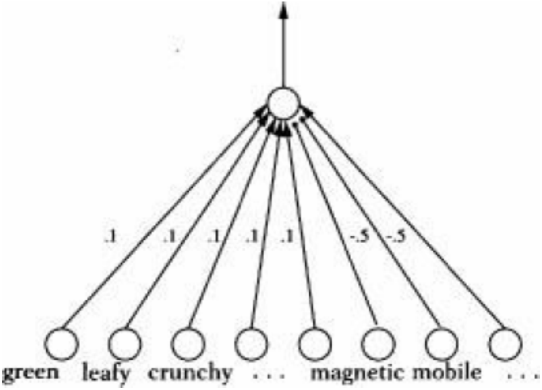
Neural networks as computational systems



[Pinker, 1999]

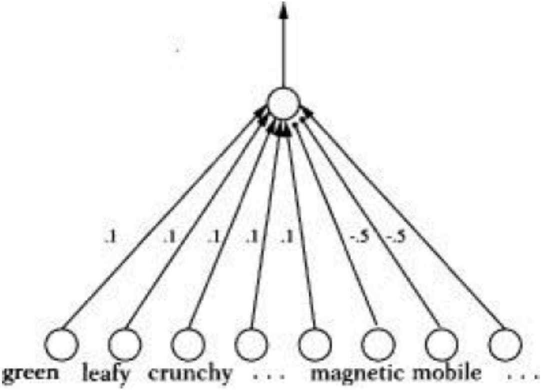
Neural networks as computational systems

Vegetable detection:

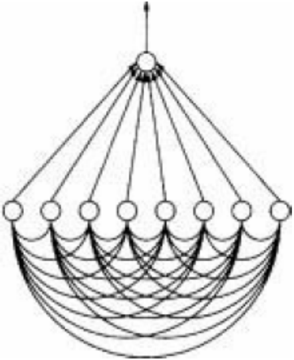


Neural networks as computational systems

Vegetable detection:



Auto-association:



[Pinker, 1999]

Cognition as pattern recognition

- **A vheclie epxledod at a plocie cehckipont near the UN haduqertares in Bagahdd on Mnoday kilinlg the bmober and an Irqai polcie offceir** [Matt Davis, MRC Cognition and Brain Sciences Unit, Cambridge]

Cognition as pattern recognition

- **A vheclie epxledod at a plocie cehckipont near the UN haduqertares in Bagahdd on Mnoday kilinlg the bmober and an Irqai polcie offceir** [Matt Davis, MRC Cognition and Brain Sciences Unit, Cambridge]
-  [Pinker, 1999]

Cognition as pattern recognition

- A vheclie epxledod at a plocie cehckipont near the UN haduqertares in Bagahdd on Mnoday kilinlg the bmober and an Irqai polcie offceir [Matt Davis, MRC Cognition and Brain Sciences Unit, Cambridge]
-  [Pinker, 1999]
- Robustness to noise and missing information; inference to fill in missing details

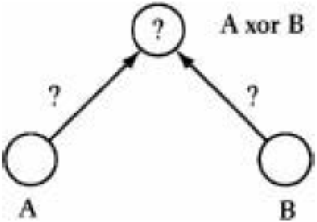
Cognition as pattern recognition

- **A vheclie epxledod at a plocie cehckipont near the UN haduqertares in Bagahdd on Mnoday kilinlg the bmober and an Irqai polcie offceir** [Matt Davis, MRC Cognition and Brain Sciences Unit, Cambridge]
-  [Pinker, 1999]
- Robustness to noise and missing information; inference to fill in missing details
- Fits with computational neural network models; hard to explain with purely rule-based models

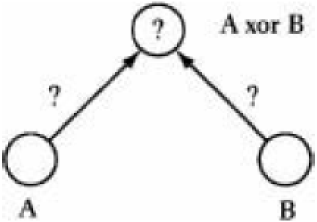
Cognition as pattern recognition

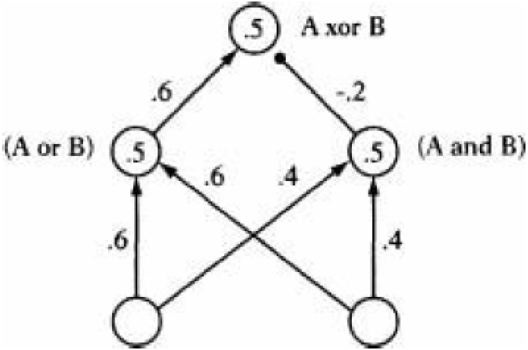
- **A vheclie epxledod at a plocie cehckipont near the UN haduqertares in Bagahdd on Mnoday kilinlg the bmober and an Irqai polcie offceir** [Matt Davis, MRC Cognition and Brain Sciences Unit, Cambridge]
-  [Pinker, 1999]
- Robustness to noise and missing information; inference to fill in missing details
- Fits with computational neural network models; hard to explain with purely rule-based models
- Language acquisition: May not be rule-based

The XOR problem



The XOR problem

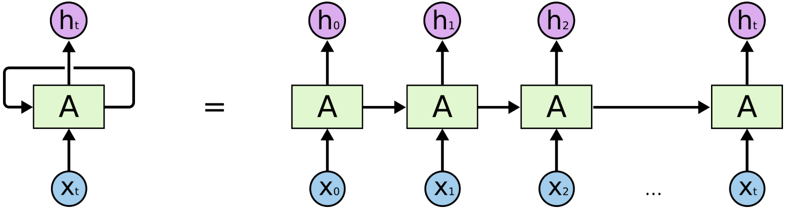




[Pinker, 1999]

Recurrent neural networks (RNNs)

Rather than just feed-forward connections, RNNs also allow for recurrent or feedback connections, thus allowing a 'memory' of previous states to be retained. This is useful for processing sequential or temporal data.



[http://colah.github.io/posts/2015-08-Understanding-LSTMs]

Long-range dependencies

- One key challenge in language processing is dealing with long-range dependencies

Long-range dependencies

- One key challenge in language processing is dealing with long-range dependencies
- Consider the sentence *I looked up to see a cloudy* ___. Here just the context of a single preceding word predicts the next with high confidence: can even be done by a bigram model

Long-range dependencies

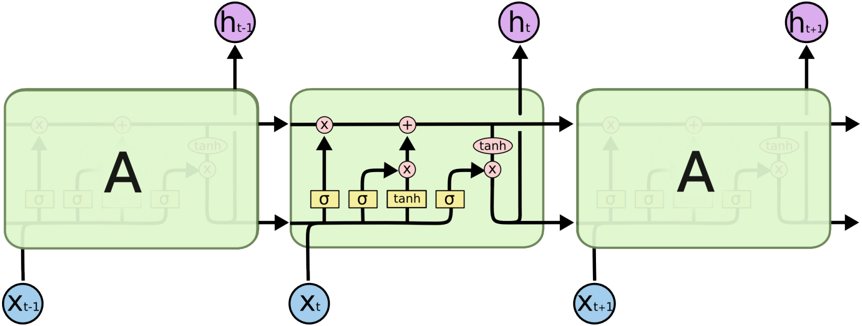
- One key challenge in language processing is dealing with long-range dependencies
- Consider the sentence *I looked up to see a cloudy* ___. Here just the context of a single preceding word predicts the next with high confidence: can even be done by a bigram model
- However, consider *I was born in Paris and spent my childhood there, so I speak fluent* _____. Here a bigram model would predict the next word to be the name of a language; but to predict which language, you need information from much further back in the sentence

Long-range dependencies

- One key challenge in language processing is dealing with long-range dependencies
- Consider the sentence *I looked up to see a cloudy* ___. Here just the context of a single preceding word predicts the next with high confidence: can even be done by a bigram model
- However, consider *I was born in Paris and spent my childhood there, so I speak fluent* _____. Here a bigram model would predict the next word to be the name of a language; but to predict which language, you need information from much further back in the sentence
- RNNs can in principle learn such long-range dependencies, but it is difficult for vanilla RNNs; a specific variety, called LSTMs, are much more powerful at this

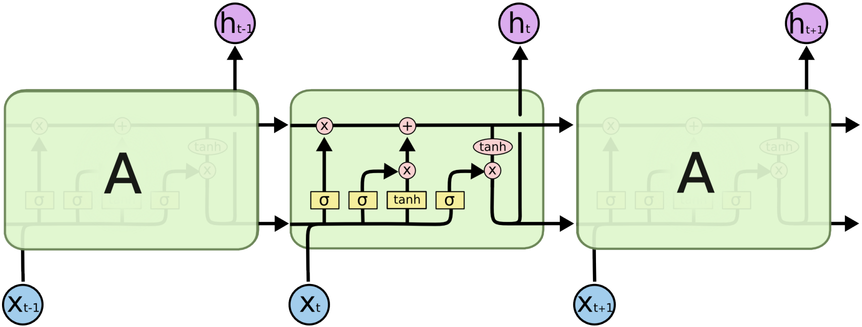
Long Short-Term Memory (LSTM) models

These have a much more sophisticated, multi-layered repeating module:



[http://colah.github.io/posts/2015-08-Understanding-LSTMs]

Long Short-Term Memory (LSTM) models



Very crudely, these essentially work via the repeating module largely passing on information (the ‘cell state’) from the previous time step as is (the horizontal line along the top). But necessary changes/updates to this state can be made via carefully regulated ‘gates’.

RNN applications

- RNNs (mainly LSTMs) have been extremely successful for a range of linguistic tasks (The Unreasonable Effectiveness of Recurrent Neural Networks), and the ability to model the maintenance of long-range dependencies in short-term or working memory seems key to this success

RNN applications

- RNNs (mainly LSTMs) have been extremely successful for a range of linguistic tasks (The Unreasonable Effectiveness of Recurrent Neural Networks), and the ability to model the maintenance of long-range dependencies in short-term or working memory seems key to this success
- Hence these models are clearly of interest from a psycholinguistic perspective, even though so far they have been more prominent in the NLP literature

RNN applications

- RNNs (mainly LSTMs) have been extremely successful for a range of linguistic tasks (The Unreasonable Effectiveness of Recurrent Neural Networks), and the ability to model the maintenance of long-range dependencies in short-term or working memory seems key to this success
- Hence these models are clearly of interest from a psycholinguistic perspective, even though so far they have been more prominent in the NLP literature
- However, these are sequence models without any hierarchical representations that could directly capture syntactic structure; so a key question would be to what extent they can learn about syntax [Linzen et al., 2016]

Part II: Bayesian models of cognition

Bayesian inference

- Much of cognition and learning in general can be thought of as solving the problem of *induction*: using observations about the world to draw inferences about the processes or mechanisms underlying those observations, which can then be used to make predictions about future observations

Bayesian inference

- Much of cognition and learning in general can be thought of as solving the problem of *induction*: using observations about the world to draw inferences about the processes or mechanisms underlying those observations, which can then be used to make predictions about future observations
- Bayesian inference provides a means to rationally draw such inferences in the context of probabilistic models of the processes or mechanisms concerned; hence it is a key component in the probabilistic modelling of cognition or learning

Bayesian inference

- Most statistical models employed in linguistics (e.g., linear regression, linear mixed models) are by default *maximum likelihood* models. This means they choose the parameters of the model (*hypothesis*, H) so as to maximise the likelihood (probability) of the given data (*evidence*, E): set H so as to $\max P(E|H)$.

Bayesian inference

- Most statistical models employed in linguistics (e.g., linear regression, linear mixed models) are by default *maximum likelihood* models. This means they choose the parameters of the model (*hypothesis*, H) so as to maximise the likelihood (probability) of the given data (*evidence*, E): set H so as to $\max P(E|H)$.
- Bayesian inference uses Bayes' theorem to invert this:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}; posterior = \frac{likelihood \times prior}{evidence}. \quad (1)$$

Bayesian inference: coin-tossing example

- I take a coin and toss it 3 times, observing 3 heads.

Bayesian inference: coin-tossing example

- I take a coin and toss it 3 times, observing 3 heads.

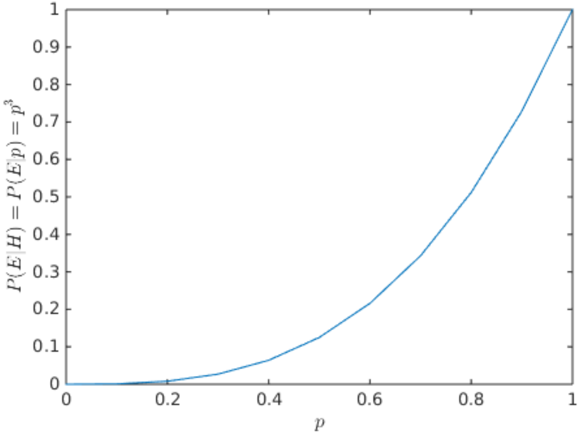
- Suppose I hypothesise that the coin has a fixed probability of turning up heads on any given toss; denote this fixed probability by p.

Bayesian inference: coin-tossing example

- I take a coin and toss it 3 times, observing 3 heads.

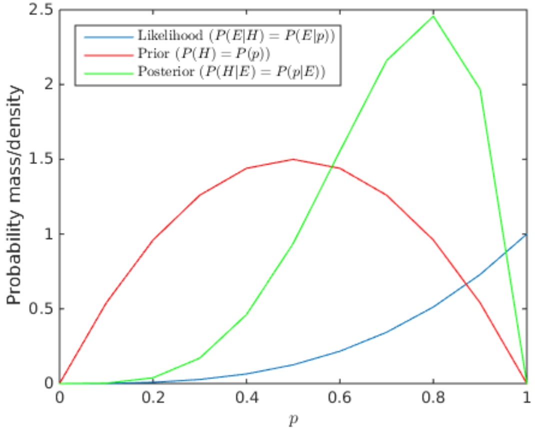
- Suppose I hypothesise that the coin has a fixed probability of turning up heads on any given toss; denote this fixed probability by p.
- Given the experimental data I've just observed, what is my best estimate of p?

Bayesian inference: coin-tossing example



Frequentist *maximum likelihood* approach: best estimate $\hat{p} = 1$

Bayesian inference: coin-tossing example



Bayesian *maximum a posteriori* approach: best estimate $\hat{p} = 0.8$

Bayesian inference: summary

- The Bayesian approach allows us to incorporate reasonable prior knowledge or assumptions, rather than trying to rely only on the observed data (which may be insufficient, or noisy)

Bayesian inference: summary

- The Bayesian approach allows us to incorporate reasonable prior knowledge or assumptions, rather than trying to rely only on the observed data (which may be insufficient, or noisy)
- The Bayesian approach can help to moderate the influence of unusual observations or outliers in the data

Bayesian inference: summary

- The Bayesian approach allows us to incorporate reasonable prior knowledge or assumptions, rather than trying to rely only on the observed data (which may be insufficient, or noisy)
- The Bayesian approach can help to moderate the influence of unusual observations or outliers in the data
- In addition to a point estimate, by looking at the full posterior, we also get an indication of the *uncertainty* in that estimate

Bayesian inference: summary

- The Bayesian approach allows us to incorporate reasonable prior knowledge or assumptions, rather than trying to rely only on the observed data (which may be insufficient, or noisy)
- The Bayesian approach can help to moderate the influence of unusual observations or outliers in the data
- In addition to a point estimate, by looking at the full posterior, we also get an indication of the *uncertainty* in that estimate
- Can be used for any parameterised probabilistic model, such as linear regression or linear mixed models

Bayesian inference: summary

- The Bayesian approach allows us to incorporate reasonable prior knowledge or assumptions, rather than trying to rely only on the observed data (which may be insufficient, or noisy)
- The Bayesian approach can help to moderate the influence of unusual observations or outliers in the data
- In addition to a point estimate, by looking at the full posterior, we also get an indication of the *uncertainty* in that estimate
- Can be used for any parameterised probabilistic model, such as linear regression or linear mixed models
- Our intuition anyway often seems to process frequentist statistics as Bayesian ones, e.g., *p*-values ('marginally significant'; 'non-significant trend towards significance' [Nicenboim and Vasishth, 2016])

Bayesian inference: conclusions

- Bayesian inference provides a machinery for how rational learners should update their beliefs (and also degree of confidence in those beliefs) in the light of evidence

Bayesian inference: conclusions

- Bayesian inference provides a machinery for how rational learners should update their beliefs (and also degree of confidence in those beliefs) in the light of evidence
- Can be especially useful for modelling learning in data-constrained settings; e.g. for language, the well-known *poverty of stimulus* and *paradox of language acquisition* problems. In a Bayesian framework, *Universal Grammar* could be thought of as a kind of prior distribution over certain parameters which govern language processing

Bayesian inference: conclusions

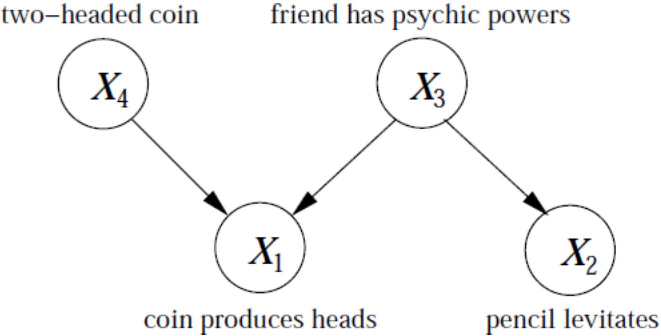
- However, we need models with richer structure to be able to capture the actual complexity of human cognition; to obtain such structure we need to look at *hierarchical* Bayesian models, with multiple levels of dependencies between random variables

Bayesian inference: conclusions

- However, we need models with richer structure to be able to capture the actual complexity of human cognition; to obtain such structure we need to look at *hierarchical* Bayesian models, with multiple levels of dependencies between random variables
- Such models are usually represented and visualised as *Bayesian networks*

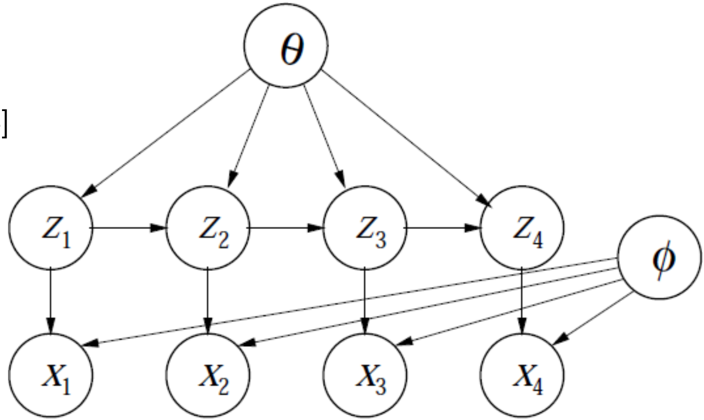
Bayesian networks

A way of representing dependencies between random variables. [Griffiths et al., 2008]



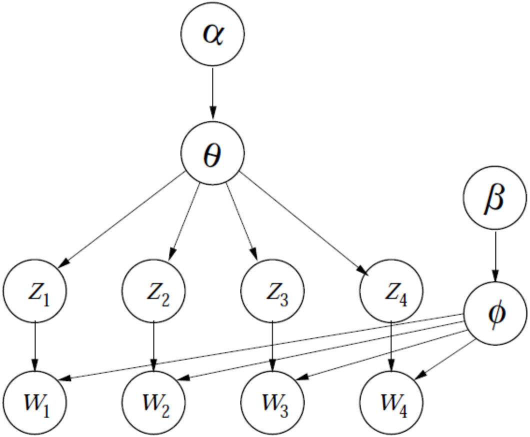
Bayesian networks

A generic model for sentence production. [Griffiths et al., 2008]



Bayesian networks

A semantic memory model (Latent Dirichlet Allocation) which can be used to infer topics from text.
[Griffiths et al., 2008]
[Further discussion]



Benefits of Bayesian topic models for semantic memory

- Richer structure than semantic spaces or networks; different topics can capture different senses of a word, allowing for polysemy and homonymy to be modelled effectively

Benefits of Bayesian topic models for semantic memory

- Richer structure than semantic spaces or networks; different topics can capture different senses of a word, allowing for polysemy and homonymy to be modelled effectively
- Unlike semantic spaces, no 'triangle inequality' or transitivity: if w_1 semantically close to w_2, and w_2 to w_3, can still have w_1 far from w_3 (e.g., `ASTEROID, BELT, BUCKLE`) via two different topics [Griffiths et al., 2008][The brain dictionary]

Benefits of Bayesian topic models for semantic memory

- Richer structure than semantic spaces or networks; different topics can capture different senses of a word, allowing for polysemy and homonymy to be modelled effectively
- Unlike semantic spaces, no 'triangle inequality' or transitivity: if w_1 semantically close to w_2, and w_2 to w_3, can still have w_1 far from w_3 (e.g., ASTEROID, BELT, BUCKLE) via two different topics [Griffiths et al., 2008][The brain dictionary]
- These topics can be learnt automatically, in an unsupervised fashion, just based on word co-occurrence in text

Benefits of Bayesian topic models for semantic memory

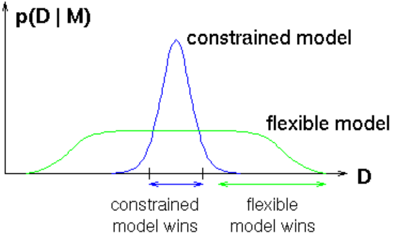
- Richer structure than semantic spaces or networks; different topics can capture different senses of a word, allowing for polysemy and homonymy to be modelled effectively
- Unlike semantic spaces, no 'triangle inequality' or transitivity: if w_1 semantically close to w_2, and w_2 to w_3, can still have w_1 far from w_3 (e.g., ASTEROID, BELT, BUCKLE) via two different topics [Griffiths et al., 2008][The brain dictionary]
- These topics can be learnt automatically, in an unsupervised fashion, just based on word co-occurrence in text
- Power of these models comes from combining richly structured representations with statistical learning – a general theme that underlies the usefulness of Bayesian models for a variety of linguistic and cognitive phenomena

Bayesian model selection/comparison

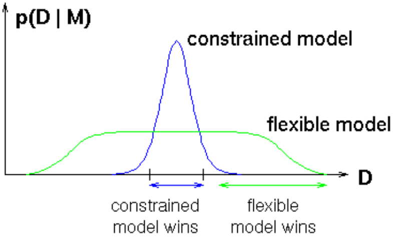
- Integration of posteriors over parameters allows for Bayesian comparison of two models/hypotheses, which may be of different complexity

Bayesian model selection/comparison

- Integration of posteriors over parameters allows for Bayesian comparison of two models/hypotheses, which may be of different complexity
- A simpler model can explain a smaller number of possible data sets; but for those data sets will assign a high probability (its probability mass is narrowly concentrated). A more complex or flexible model spreads its probability mass more thinly [Tom Minka, MIT]:

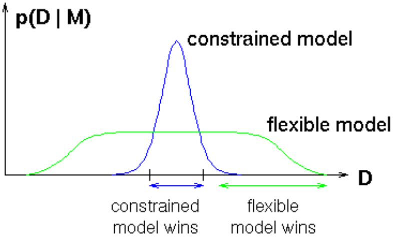


Bayesian model selection/comparison



- This leads to what is called the *Bayesian Occam's Razor*: a principled way of selecting the simplest model which reasonably explains a given set of observations

Bayesian model selection/comparison



- This leads to what is called the *Bayesian Occam's Razor*: a principled way of selecting the simplest model which reasonably explains a given set of observations
- Some interesting recent work on Bayesian comparison of competing models of retrieval in sentence comprehension [Nicenboim and Vasishth, 2016]

References

References I

Griffiths, T. L., Kemp, C., and Tenenbaum, J. B. (2008).
Bayesian models of cognition.
In Sun, R., editor, *The Cambridge Handbook of Computational Psychology*. Cambridge University Press.

Linzen, T., Dupoux, E., and Goldberg, Y. (2016).
Assessing the ability of LSTMs to learn syntax-sensitive dependencies.
Transactions of the Association for Computational Linguistics, 4:521–535.

References II

Nicenboim, B. and Vasishth, S. (2016).
Models of retrieval in sentence comprehension: A computational evaluation using Bayesian hierarchical modeling.
ArXiv e-prints, arXiv:1612.04174 [cs.CL].

Nicenboim, B. and Vasishth, S. (2016).
Statistical methods for linguistic research: Foundational Ideas—Part II.
Language and Linguistics Compass, 10(11):591–613. LNCO-0657.R1.

References III

Pinker, S. (1999).
How the Mind Works.
Penguin.