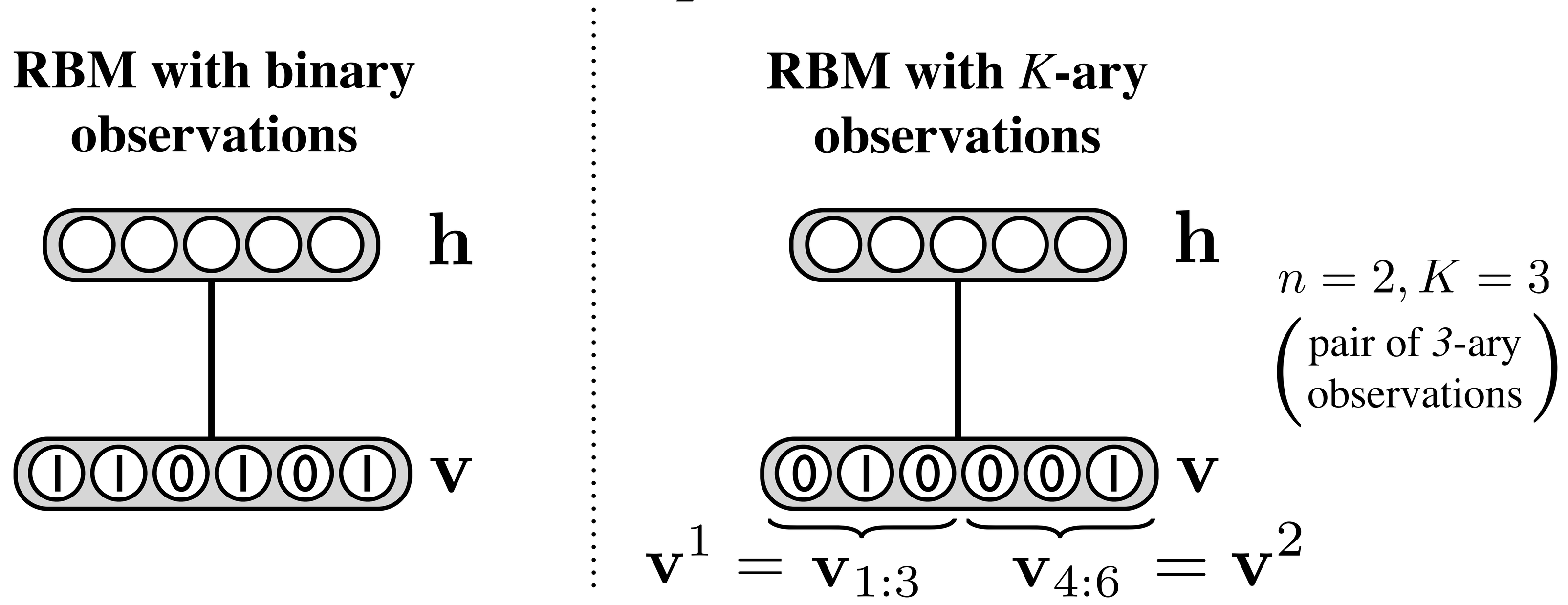
Training Restricted Boltzmann Machines on Word Observations

George E. Dahl (University of Toronto)
Ryan P. Adams (Harvard University)
Hugo Larochelle (Université de Sherbrooke)

Introduction

- Words are intrinsically high dimensional objects (nb. of dimensions = vocabulary size)
- This poses two challenges to RBM training:
  1. naive RBM parametrization has a lot of parameters (solution: factorize weights)
  2. sampling-based training (CD, PCD) is very expensive
- We show Metropolis Hastings can be used as an efficient and effective approximation for sampling word observations during training
- The learning update we propose takes time *independent of the vocabulary size*

RBM with K-ary observations



$$p(\mathbf{v}^i = \mathbf{e}_k|\mathbf{h}) = \frac{\exp(\mathbf{b}^{i\top}\mathbf{e}_k + \mathbf{h}^\top\mathbf{W}^i\mathbf{e}_k)}{\sum_{k'=1}^{K}\exp(\mathbf{b}^{i\top}\mathbf{e}_{k'} + \mathbf{h}^\top\mathbf{W}^i\mathbf{e}_{k'})}$$

Metropolis Hastings

1. Sample from $q(\mathbf{v}^i)$ to get a proposed word $\tilde{\mathbf{v}}$ ($q(\mathbf{v}^i)$ can be a smoothed unigram model)
2. Replace current word by $\tilde{\mathbf{v}}$ with probability

$$\min\left\{1, \frac{q(\mathbf{v}^i)\exp(\mathbf{b}^{i\top}\tilde{\mathbf{v}} + \mathbf{h}^\top\mathbf{W}^i\tilde{\mathbf{v}})}{q(\tilde{\mathbf{v}})\exp(\mathbf{b}^{i\top}\mathbf{v}^i + \mathbf{h}^\top\mathbf{W}^i\mathbf{v}^i)}\right\}$$

3. For proposals from a fixed distribution (e.g. smoothed unigram) the alias method lets us generate proposals in constant time (linear setup cost)

RBM with Word Representations

- We used MH to train a K-ary RBM, with factored weights that incorporate word representations, on n-gram windows

$$E(\mathbf{v},\mathbf{h}) = -\mathbf{c}^\top\mathbf{h} + \sum_{i=1}^{n} -\mathbf{b}^{*\top}\mathbf{v}^i - \mathbf{h}^\top\,\mathbf{U}^i\,\mathbf{D}^\top\,\mathbf{v}^i$$

- The top-down conditional distribution becomes

word representations

$$p(\mathbf{v}^i = \mathbf{e}_k|\mathbf{h}) = \frac{\exp(\mathbf{b}^{*\top}\mathbf{e}_k + \mathbf{h}^\top\mathbf{U}^i\,\mathbf{D}^\top\,\mathbf{e}_k)}{\sum_{k'=1}^{K}\exp(\mathbf{b}^{*\top}\mathbf{e}_{k'} + \mathbf{h}^\top\mathbf{U}^i\,\mathbf{D}^\top\,\mathbf{e}_{k'})}$$

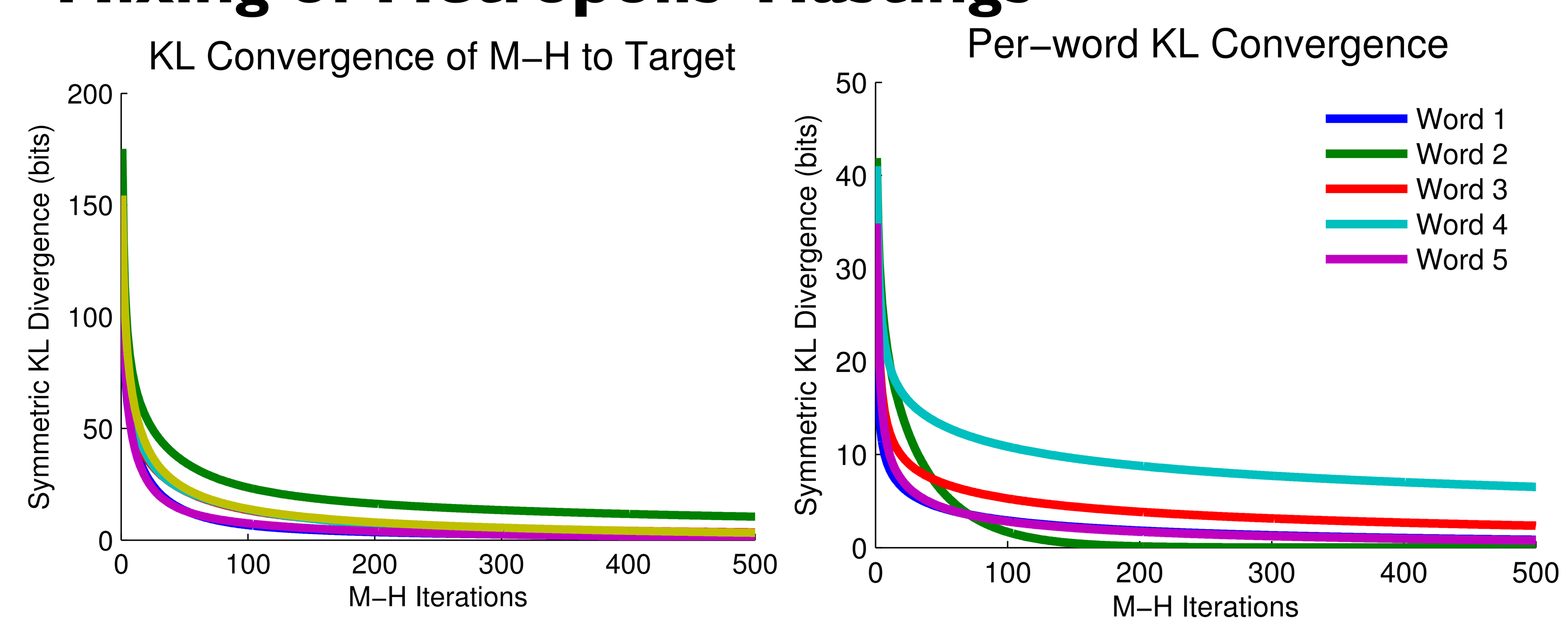
shared biases

position dependent weights

Nearest neighbors (word rep. space)

could	spokeswoman	suspects	science	china	mother	sunday
should	spokesman	defendants	sciences	japan	father	saturday
would	lawyer	detainees	medicine	taiwan	daughter	friday
will	columnist	hijackers	research	thailand	son	monday
can	consultant	attackers	economics	russia	grandmother	thursday
might	secretary-general	demonstrators	engineering	indonesia	sister	wednesday
must	strategist	inmates	arts	iran	grandfather	tuesday
did	negotiator	assailants	psychology	india	brother	yesterday
wo	administrator	atrocities	journalism	nigeria	girlfriend	today
does	correspondent	dissidents	privacy	greece	husband	tomorrow
ca	adviser	killings	nutrition	vietnam	cousin	tonight
tom	**actually**	**probably**	**quickly**	**earned**	**what**	**hotel**
jim	finally	certainly	easily	averaged	why	restaurant
bob	definitely	definitely	slowly	clinched	how	theater
kevin	rarely	hardly	carefully	retained	whether	casino
brian	eventually	usually	effectively	regained	whatever	ranch
steve	hardly	actually	frequently	grabbed	where	zoo
chris	ultimately	surely	badly	netted	something	cafe
david	basically	simply	seriously	saved	whom	tribune
robert	usually	apparently	quietly	secured	nothing	warehouse
joe	somehow	obviously	strongly	enjoyed	everything	symphony
ron	suddenly	clearly	closely	surpassed	neither	nightclub

Mixing of Metropolis-Hastings



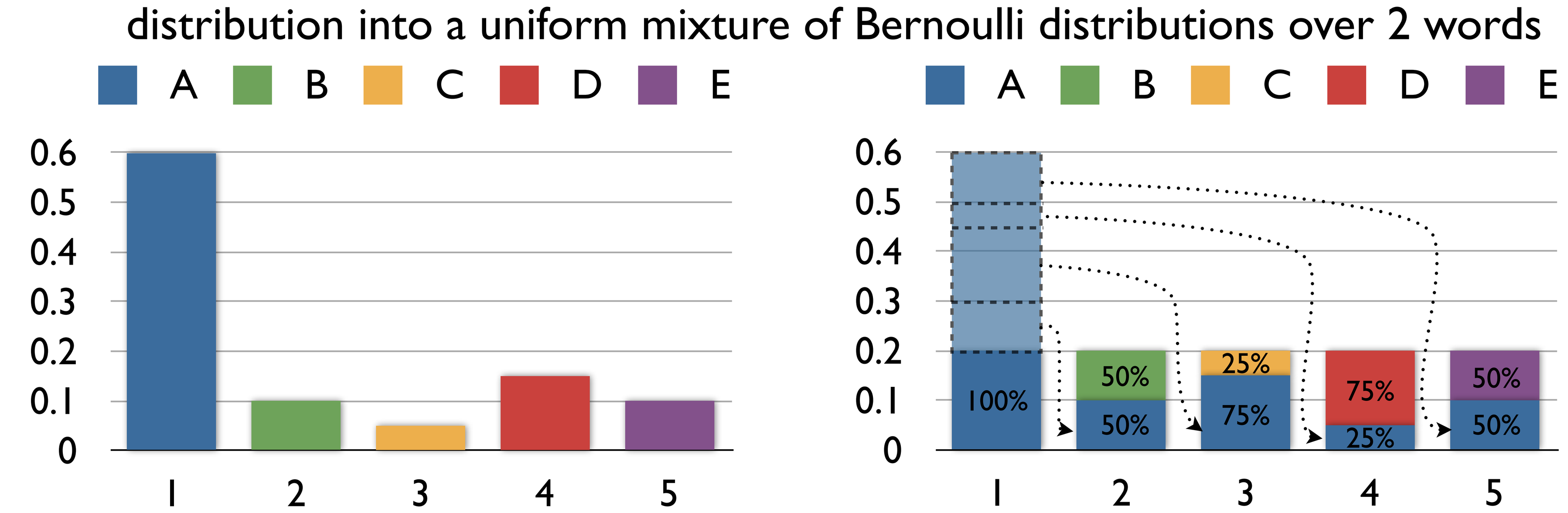
Left: Convergence of MH operator to the true conditional over the visible units for 6 randomly chosen data cases, measured with symmetric KL

Right: For the slowest case on the left, convergence of MH operator to the true conditional over the group of visible units for each word in the 5-gram

- The hidden state has a strong effect on the mixing
- Most groups mix well, but a few tend to mix very slowly
- More sophisticated proposal distributions might improve mixing

Sampling proposed words in constant time

- Naïve implementations of sampling from a unigram distribution would be linear in the vocabulary size
- The alias method samples in constant time by first transforming the distribution into a uniform mixture of Bernoulli distributions over 2 words



- Unlike in Mnih and Hinton (2007), we model the **joint** distribution of n-grams, not the conditional probability of the last word given the n-1 previous words
- Can then use the hidden unit activities as n-gram representations
- We trained on 3-grams extracted from the English Gigaword corpus (vocabulary of 100k words)
- Using the word representations and hidden activations as CRF features helps on a chunking benchmark

Chunking results

Method	Valid F1	Test F1
Without representations	94.16	93.79
WordRepRBM	94.82	94.10
WordRepRBM (+ hidden unit features)	**95.01**	**94.44**
Mnih and Hinton	94.63	94.00
Collobert and Weston	94.66	94.10
Brown clusters	94.67	94.11

Turian et al. 2010

- Training class conditional "bag of 5-grams" WordRepRBMs helps sentiment classification on the Large Movie Review dataset from Maas et al.
- We use the average free energy of each RBM over a bag as features for a discriminative classifier
- When combined with binary term frequency bag of word features, the average free energies of the two RBMs over the 5-grams from a document yield state of the art results on this dataset

Sentiment Classification results

Method	Test
LDA	67.42
LSA	83.96
Maas et al. "full"	87.44
Bag of words "bnc"	87.80
Maas et al. "full" + BoW "bnc"	88.33
Maas et al. "full" + BoW "bnc" + unlabeled data	88.89
5-gram WordRepRBM	87.42
5-gram WordRepRBM + BoW "bnc"	**89.23**

References

Turian, Ratinov and Bengio, *Word representations: A simple and general method for semi-supervised learning*, 2010
Mnih and Hinton, *Three new graphical models for statistical language modelling*, 2007
Mnih and Hinton, *A scalable hierarchical distributed language model*, 2009
Collobert and Weston, *A unified architecture for natural language processing: Deep neural networks with multitask learning*, 2008
Maas, Daly, Pham, Huang, Ng, and Potts, *Learning Word Vectors for Sentiment Analysis*, 2011