Linguistic Structured Sparsity in Text Categorization

Dani Yogatama and Noah A. Smith
Language Technologies Institute
Carnegie Mellon University
{dyogatama,nasmith}@cs.cmu.edu


Dani Yogatama

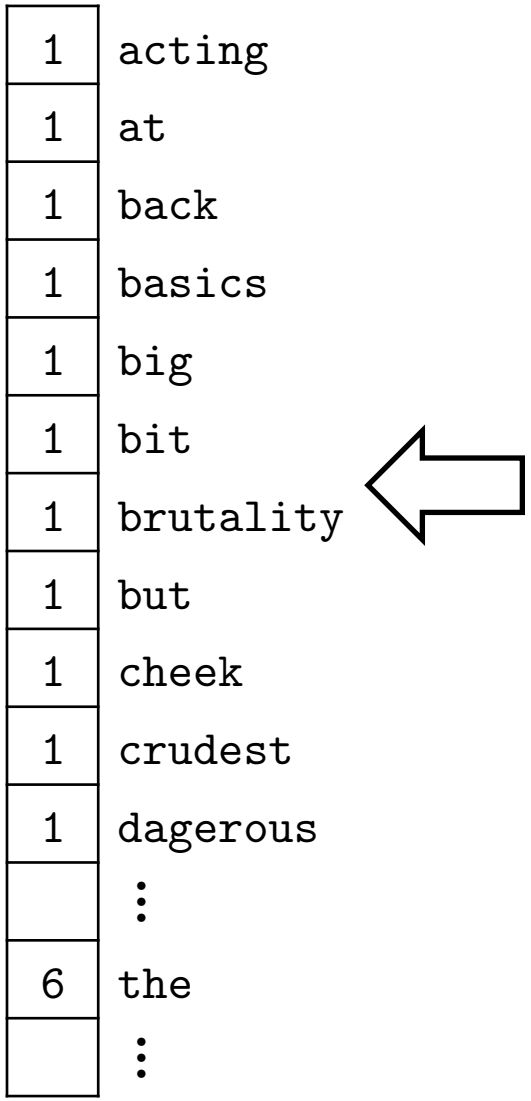
Summary

- Words of a feather (should) flock together
- Idea: use linguistic structure to define *feathers* (flocks) instead of *features*
- Math: sparse group lasso regularization
- Results: text classification (sentiment, forecasting, topic)

Text Classification

this film is one big joke : you have all the basics elements
of romance (love at first sight , great passion , etc .) and gangster flicks
(brutality , dagerous machinations , the mysterious don , etc.) ,
but it is all done with the crudest humor .
it ' s the kind of thing you either like viserally and immediately '' get '' or you don ' t .
that is a matter of taste and expectations .
i enjoyed it and it took me back to the mid80s , when nicolson and turner were in their primes .
the acting is very good , if a bit obviously tongue - in - cheek .

Bag of Words

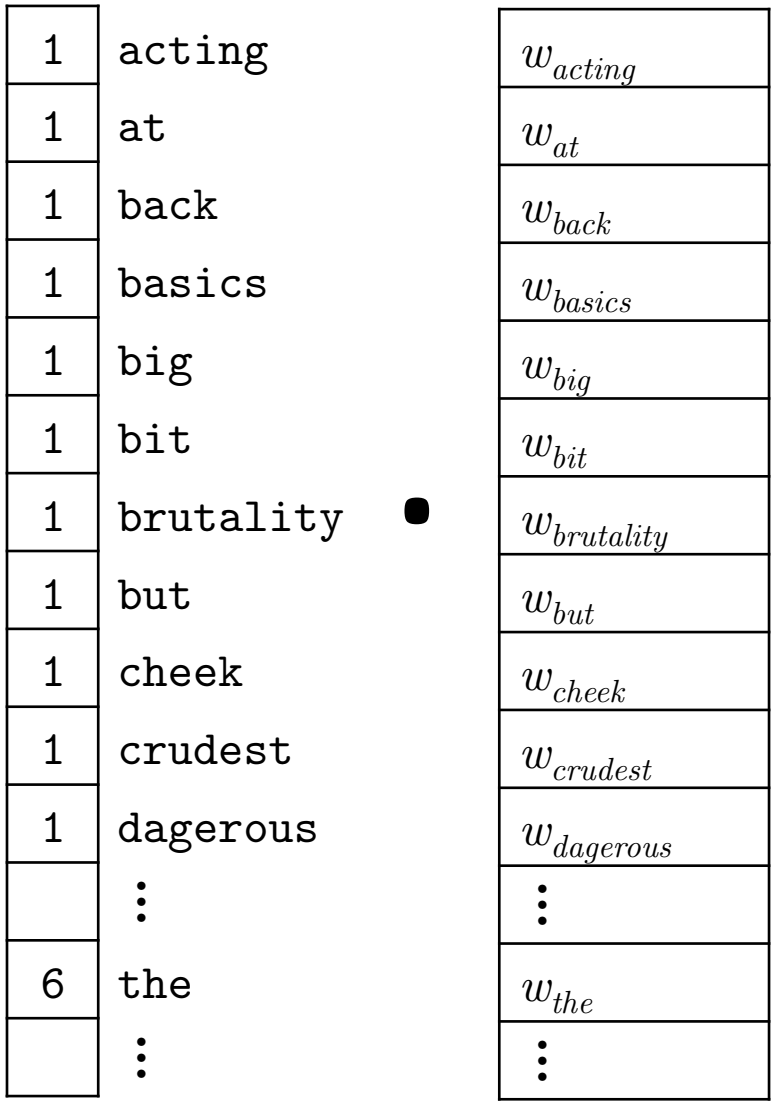


this film is one big joke : you have all the basics elements
of romance (love at first sight , great passion , etc .) and gangster flicks
(brutality , dagerous machinations , the mysterious don , etc.) ,
but it is all done with the crudest humor .
it ' s the kind of thing you either like viserally and immediately '' get '' or you don ' t .
that is a matter of taste and expectations .
i enjoyed it and it took me back to the mid80s , when nicolson and turner were in their primes .
the acting is very good , if a bit obviously tongue - in - cheek .

Bag of Words

1	acting
1	at
1	back
1	basics
1	big
1	bit
1	brutality
1	but
1	cheek
1	crudest
1	dagerous
	⋮
6	the
	⋮

Linear Classifier



$$sign\left(\mathbf{f}(document) \cdot \mathbf{w}\right) \Rightarrow \hat{y}$$

Text is Not a Bag of Words!

- Sentences

```
this film is one big joke : you have all
the basics elements
of romance ( love at first sight , great
passion , etc . ) and gangster flicks
( brutality , dagerous machinations , the
mysterious don , etc. ) ,
but it is all done with the crudest
humor .
it ' s the kind of thing you either like
viserally and immediately '' get '' or you
don ' t .
that is a matter of taste and
expectations .
i enjoyed it and it took me back to the
mid80s , when nicolson and turner were in
their primes .
the acting is very good , if a bit
obviously tongue - in - cheek .
```

Text is Not a Bag of Words!

- Sentences
- Phrases

this film is one big joke : you have all the basics elements
of romance (love at first sight , great passion , etc .) and gangster flicks
(brutality , dagerous machinations , the mysterious don , etc.) ,
but it is all done with the crudest humor .
it ' s the kind of thing you either like viserally and immediately '' get '' or you don ' t .
that is a matter of taste and expectations .
i enjoyed it and it took me back to the mid80s , when nicolson and turner were in their primes .
the acting is very good , if a bit obviously tongue - in - cheek .

Text is Not a Bag of Words!

- Sentences
- Phrases
- Fine-grained syntactic classes

this film is one big joke : you have all the basics elements
of romance (love at first sight , great passion , etc .) and gangster flicks
(brutality , dagerous machinations , the mysterious don , etc.) ,
but it is all done with the crudest humor .
it ' s the kind of thing you either like viserally and immediately '' get '' or you don ' t .
that is a matter of taste and expectations .
i enjoyed it and it took me back to the mid80s , when nicolson and turner were in their primes .
the acting is very good , if a bit obviously tongue - in - cheek .

Text is Not a Bag of Words!

- Sentences
- Phrases
- Fine-grained syntactic classes
- Thematic topics

(and many more!)

this film is one big joke : you have all the basics elements
of romance (love at first sight , great passion , etc .) and gangster flicks
(brutality , dagerous machinations , the mysterious don , etc.) ,
but it is all done with the crudest humor .
it ' s the kind of thing you either like viserally and immediately '' get '' or you don ' t .
that is a matter of taste and expectations .
i enjoyed it and it took me back to the mid80s , when nicolson and turner were in their primes .
the acting is very good , if a bit obviously tongue - in - cheek .

Learning the Weights $\mathbf{w}$

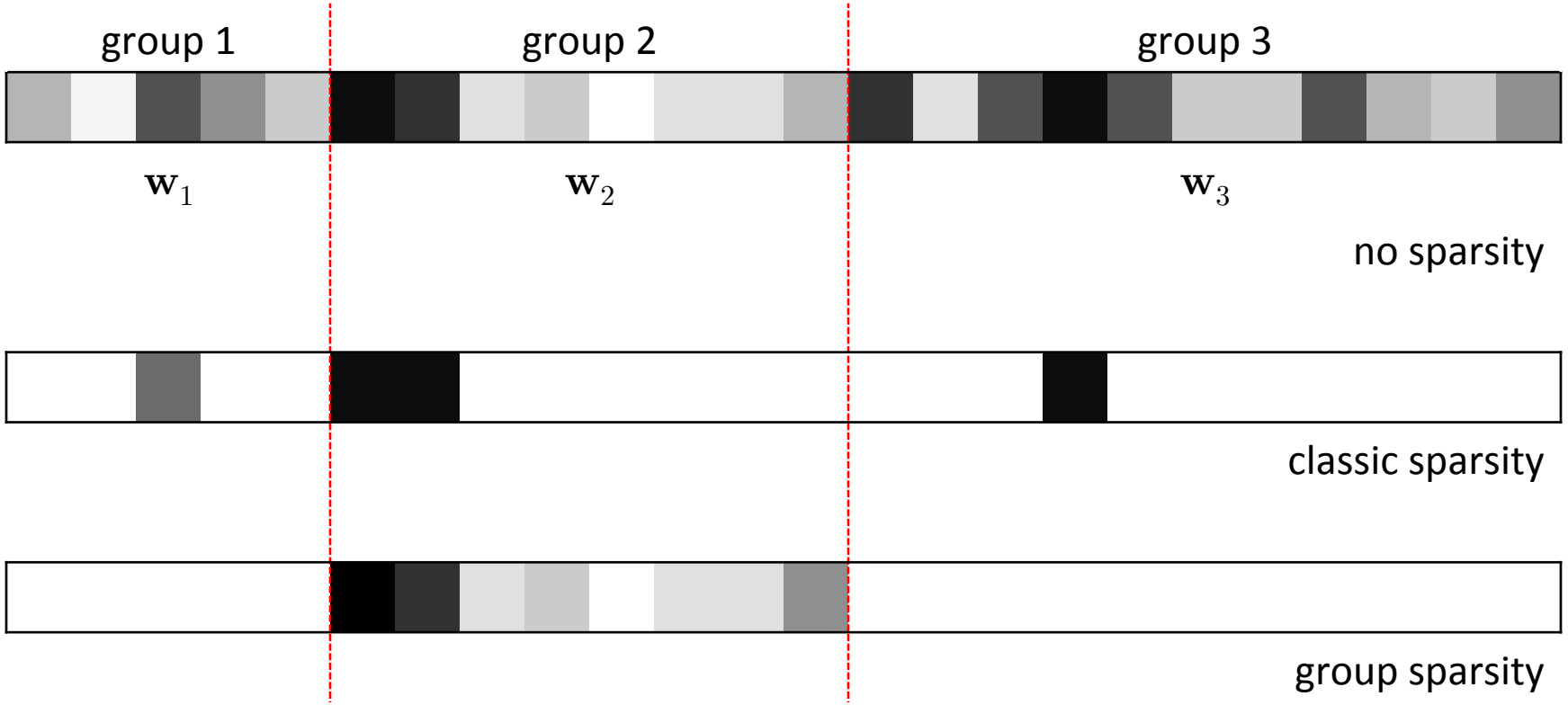
"fit the data"
(e.g., log-likelihood of y_n given d_n,
hinge loss, ...)

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \sum_{n=1}^{N} L(\mathbf{f}(d_n), y_n; \mathbf{w}) + R(\mathbf{w})$$

"generalize"
(e.g., $\lambda\|\mathbf{w}\|_2^2$;
$\lambda\|\mathbf{w}\|_1$)

Group Lasso (Yuan & Lin '06)

$$R(\mathbf{w}) = \sum_g \lambda_g \|\mathbf{w}_g\|_2$$

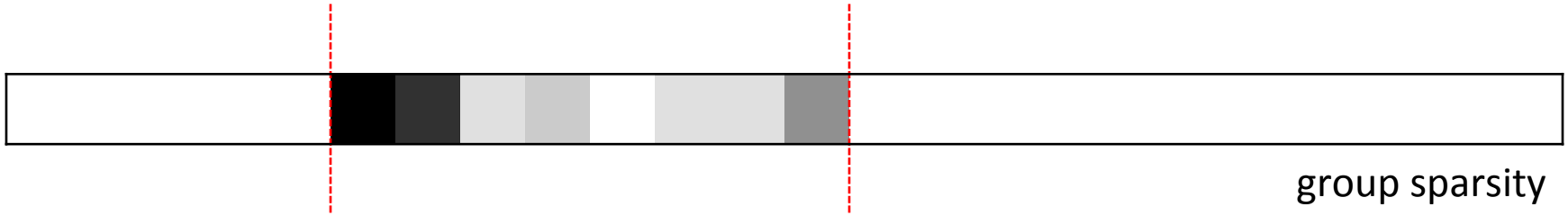


Group Lasso (Yuan & Lin '06)

$$R(\mathbf{w}) = \sum_g \lambda_g \|\mathbf{w}_g\|_2$$

In NLP:

- chunking and parsing (Martins et al., 2011)
- language modeling (Nelakanti et al., 2013)



Learning the Weights $\mathbf{w}$

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \sum_{n=1}^{N} L(\mathbf{f}(d_n), y_n; \mathbf{w}) + R(\mathbf{w})$$

Learning the Weights $\mathbf{w}$

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \sum_{n=1}^{N} L(\mathbf{f}(d_n), y_n; \mathbf{w}) + R(\mathbf{w})$$

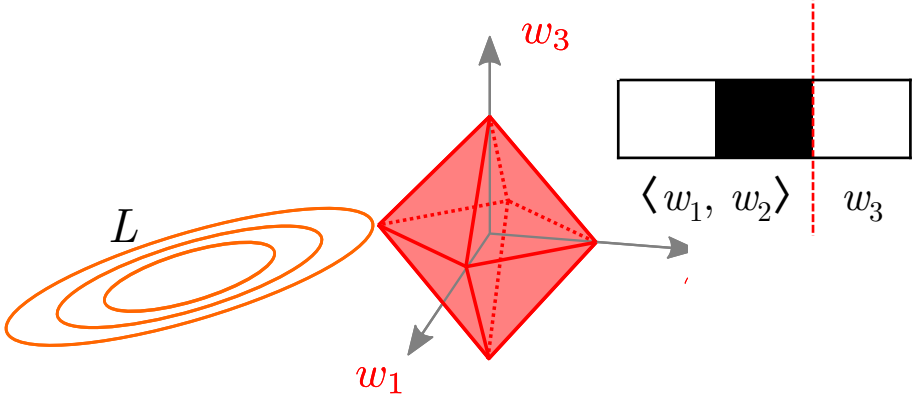
$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \sum_{n=1}^{N} L(\mathbf{f}(d_n), y_n; \mathbf{w})$$

$$\text{s.t. } R(\mathbf{w}) \leq \tau$$

"Tikhonov" regularization

"Ivanov" regularization

Lasso vs. Group Lasso



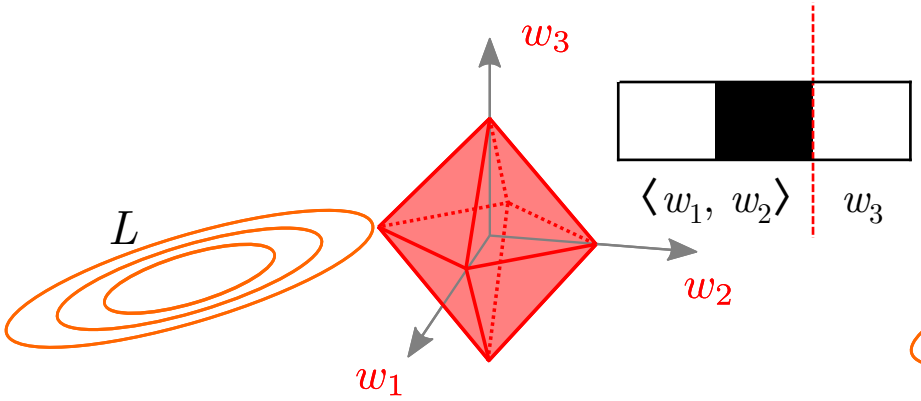
$$R(\mathbf{w}) = |w_1| + |w_2| + |w_3|$$

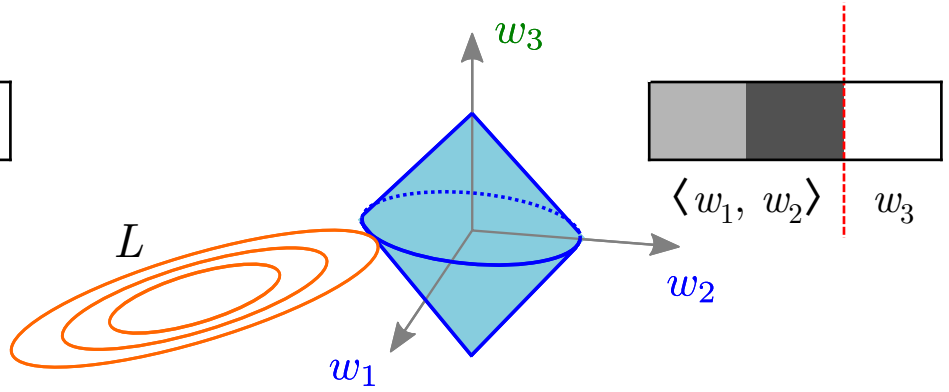
$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \sum_{n=1}^{N} L(\mathbf{f}(d_n), y_n; \mathbf{w})$$

$$\text{s.t. } R(\mathbf{w}) \leq \tau$$

Martins et al., EACL 2014 tutorial on structured sparsity in NLP

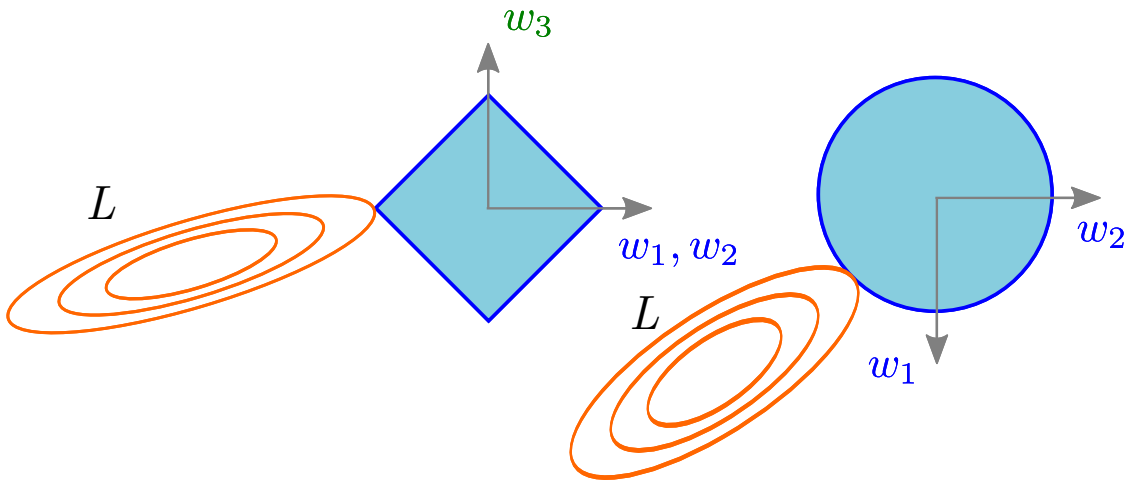
Lasso vs. Group Lasso





$$R(\mathbf{w}) = |w_1| + |w_2| + |w_3|$$

$$R(\mathbf{w}) = ||\langle w_1, w_2 \rangle||_2 + |w_3|$$



Martins et al., EACL 2014 tutorial on structured sparsity in NLP

Whence Groups?

Back to NLP ...

Sentence Regularizer

$$R(\mathbf{w}) = \sum_{n=1}^{N} \sum_{s=1}^{S_n} \lambda_{n,s} \|\mathbf{w}_{n,s}\|_2$$

- Every sentence *s* in every document *n* gets a group.
- If $\mathbf{w}_{n,s}$ can be driven to zero, that means the sentence is irrelevant to the task.
- Many overlapping groups!

Yogatama and Smith (ICML 2014)

Group for Sentence 1

1	acting
1	at
1	back
1	basics
1	big
1	bit
1	brutality
1	but
1	cheek
1	crudest
1	dagerous
	⋮
6	the
	⋮

this film is one big joke : you have all the basics elements
of romance (love at first sight , great passion , etc .) and gangster flicks
(brutality , dagerous machinations , the mysterious don , etc.) ,
but it is all done with the crudest humor .
it ' s the kind of thing you either like viserally and immediately '' get '' or you don ' t .
that is a matter of taste and expectations .
i enjoyed it and it took me back to the mid80s , when nicolson and turner were in their primes .
the acting is very good , if a bit obviously tongue - in - cheek .

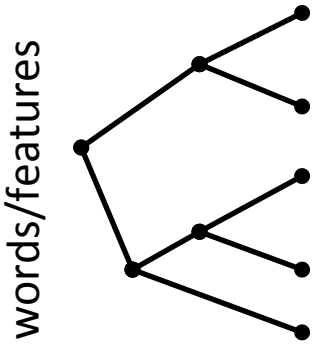
Group for Sentence 5

1	acting
1	at
1	back
1	basics
1	big
1	bit
1	brutality
1	but
1	cheek
1	crudest
1	dagerous
	⋮
6	the
	⋮

this film is one big joke : you have all the basics elements
of romance (love at first sight , great passion , etc .) and gangster flicks (brutality , dagerous machinations , the mysterious don , etc.) ,
but it is all done with the crudest humor .
it ' s the kind of thing you either like viserally and immediately '' get '' or you don ' t .
that is a matter of taste and expectations .
i enjoyed it and it took me back to the mid80s , when nicolson and turner were in their primes .
the acting is very good , if a bit obviously tongue - in - cheek .

More Linguistic Structure Regularizers

- Parse tree regularizer

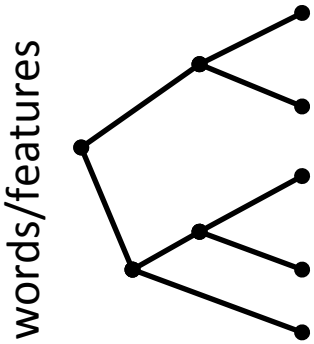


groups

	1	2	3	4	5	6	7	8	9
the	✓					✓			✓
actors		✓				✓			✓
are			✓				✓	✓	✓
fantastic				✓			✓	✓	✓
.					✓			✓	✓

More Linguistic Structure Regularizers

- Parse tree regularizer



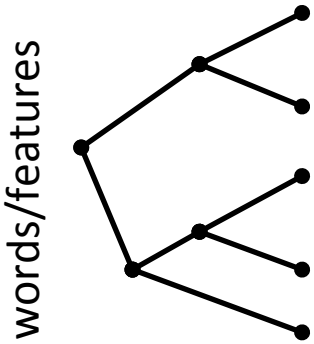
groups

	1	2	3	4	5	6	7	8	9
the	✓					✓			✓
actors		✓				✓			✓
are			✓				✓	✓	✓
fantastic				✓			✓	✓	✓
.					✓			✓	✓

- Each of 5,000 hierarchical Brown clusters

More Linguistic Structure Regularizers

- Parse tree regularizer



	groups								
	1	2	3	4	5	6	7	8	9
the	✓					✓			✓
actors		✓				✓			✓
are			✓				✓	✓	✓
fantastic				✓			✓	✓	✓
.					✓			✓	✓

- Each of 5,000 hierarchical Brown clusters
- Top ten words in each of 1,000 LDA topics

Sparse Group Lasso

$$\min_{\mathbf{w}} R(\mathbf{w}) + \lambda \|\mathbf{w}\|_1 + \sum_{n=1}^{N} L(\mathbf{f}(d_n), y_n; \mathbf{w})$$

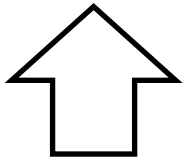
Optimization

$$\min_{\mathbf{w}} R(\mathbf{w}) + \lambda \|\mathbf{w}\|_1 + \sum_{n=1}^{N} L(\mathbf{f}(d_n), y_n; \mathbf{w})$$

Optimization

$$\min_{\mathbf{w},\mathbf{v}} R(\mathbf{v}) + \lambda\|\mathbf{w}\|_1 + \sum_{n=1}^{N} L(\mathbf{f}(d_n), y_n; \mathbf{w})$$
$$\text{s.t. } \mathbf{v} = \mathbf{M}\mathbf{w}$$

separate $\mathbf{w}$ from "copies" $\mathbf{v}$, constraint forces agreement



$$\min_{\mathbf{w}} R(\mathbf{w}) + \lambda\|\mathbf{w}\|_1 + \sum_{n=1}^{N} L(\mathbf{f}(d_n), y_n; \mathbf{w})$$

Optimization

$$\min_{\mathbf{w},\mathbf{v}} R(\mathbf{v}) + \lambda\|\mathbf{w}\|_1 + \sum_{n=1}^{N} L(\mathbf{f}(d_n), y_n; \mathbf{w})$$

$$\text{s.t. } \mathbf{v} = \mathbf{M}\mathbf{w}$$

separate $\mathbf{w}$ from "copies" $\mathbf{v}$,
constraint forces agreement

Optimization

$$\left.\begin{aligned} &\min_{\mathbf{w},\mathbf{v}} R(\mathbf{v}) + \lambda\|\mathbf{w}\|_1 + \sum_{n=1}^{N} L(\mathbf{f}(d_n), y_n; \mathbf{w}) \\ &\text{s.t. } \mathbf{v} = \mathbf{M}\mathbf{w} \end{aligned}\right\}$$

separate $\mathbf{w}$ from "copies" $\mathbf{v}$, constraint forces agreement

$$\min_{\mathbf{w},\mathbf{v}} \max_{\mathbf{u}} R(\mathbf{v}) + \lambda\|\mathbf{w}\|_1 + \sum_{n=1}^{N} L(\mathbf{f}(d_n), y_n; \mathbf{w}) + \mathbf{u}\cdot(\mathbf{v} - \mathbf{M}\mathbf{w}) + \frac{\rho}{2}\|\mathbf{v} - \mathbf{M}\mathbf{w}\|_2^2$$

"augmented Lagrangian"

Optimization

$$\min_{\mathbf{w},\mathbf{v}} R(\mathbf{v}) + \lambda\|\mathbf{w}\|_1 + \sum_{n=1}^{N} L(\mathbf{f}(d_n), y_n; \mathbf{w})$$
$$\text{s.t. } \mathbf{v} = \mathbf{M}\mathbf{w}$$

separate $\mathbf{w}$ from "copies" $\mathbf{v}$, constraint forces agreement

$$\min_{\mathbf{w},\mathbf{v}} \max_{\mathbf{u}} R(\mathbf{v}) + \lambda\|\mathbf{w}\|_1 + \sum_{n=1}^{N} L(\mathbf{f}(d_n), y_n; \mathbf{w}) + \mathbf{u}\cdot(\mathbf{v}-\mathbf{M}\mathbf{w}) + \frac{\rho}{2}\|\mathbf{v}-\mathbf{M}\mathbf{w}\|_2^2$$

ADMM: Alternating Directions Method of Multipliers

alternating, blockwise updates of $\mathbf{w}$ and $\mathbf{v}$

a "faster" version of dual ascent for solving the augmented Lagrangian (Hestenes '69; Powell '69)

(Glowinski & Marroco '75; Gabay & Mercier '76)

“Blockwise” Updates

$\mathbf{w}$ update ≈ loss minimization with elastic net regularization (Zou & Hastie ’05)

$$\min_{\mathbf{w},\mathbf{v}} \max_{\mathbf{u}} R(\mathbf{v}) + \lambda\|\mathbf{w}\|_1 + \sum_{n=1}^{N} L(\mathbf{f}(d_n), y_n; \mathbf{w}) + \mathbf{u}\cdot(\mathbf{v}-\mathbf{M}\mathbf{w}) + \frac{\rho}{2}\|\mathbf{v}-\mathbf{M}\mathbf{w}\|_2^2$$

constant

“Blockwise” Updates

$$\min_{\mathbf{w},\mathbf{v}} \max_{\mathbf{u}} R(\mathbf{v}) + \lambda\|\mathbf{w}\|_1 + \sum_{n=1}^{N} L(\mathbf{f}(d_n), y_n; \mathbf{w}) + \mathbf{u}\cdot(\mathbf{v} - \mathbf{M}\mathbf{w}) + \frac{\rho}{2}\|\mathbf{v} - \mathbf{M}\mathbf{w}\|_2^2$$

$\mathbf{v}$ updates: proximal operator for each group:

$$\mathbf{z}_{n,s} = \mathbf{M}_{d,s}\mathbf{w} - \frac{\mathbf{u}_{d,s}}{\rho}$$

$$\mathbf{v}_{n,s} = \begin{cases} \mathbf{0} & \text{if } \|\mathbf{z}_{n,s}\|_2 \leq \tau \\ \dfrac{\|\mathbf{z}_{n,s}\|_2 - \tau}{\|\mathbf{z}_{n,s}\|_2}\mathbf{z}_{n,s} & \text{otherwise} \end{cases}$$

“Blockwise” Updates

$$\min_{\mathbf{w},\mathbf{v}} \max_{\mathbf{u}} R(\mathbf{v}) + \lambda\|\mathbf{w}\|_1 + \sum_{n=1}^{N} L(\mathbf{f}(d_n), y_n; \mathbf{w}) + \mathbf{u}\cdot(\mathbf{v}-\mathbf{M}\mathbf{w}) + \frac{\rho}{2}\|\mathbf{v}-\mathbf{M}\mathbf{w}\|_2^2$$

simple dual update $\mathbf{u}$

Implications

- Group sparsity and strong sparsity
- Model class is still a (fast) bag of words ... but somehow "informed" by structure
- Learning is more expensive ... but still convex
- A new kind of **interpretability** ...

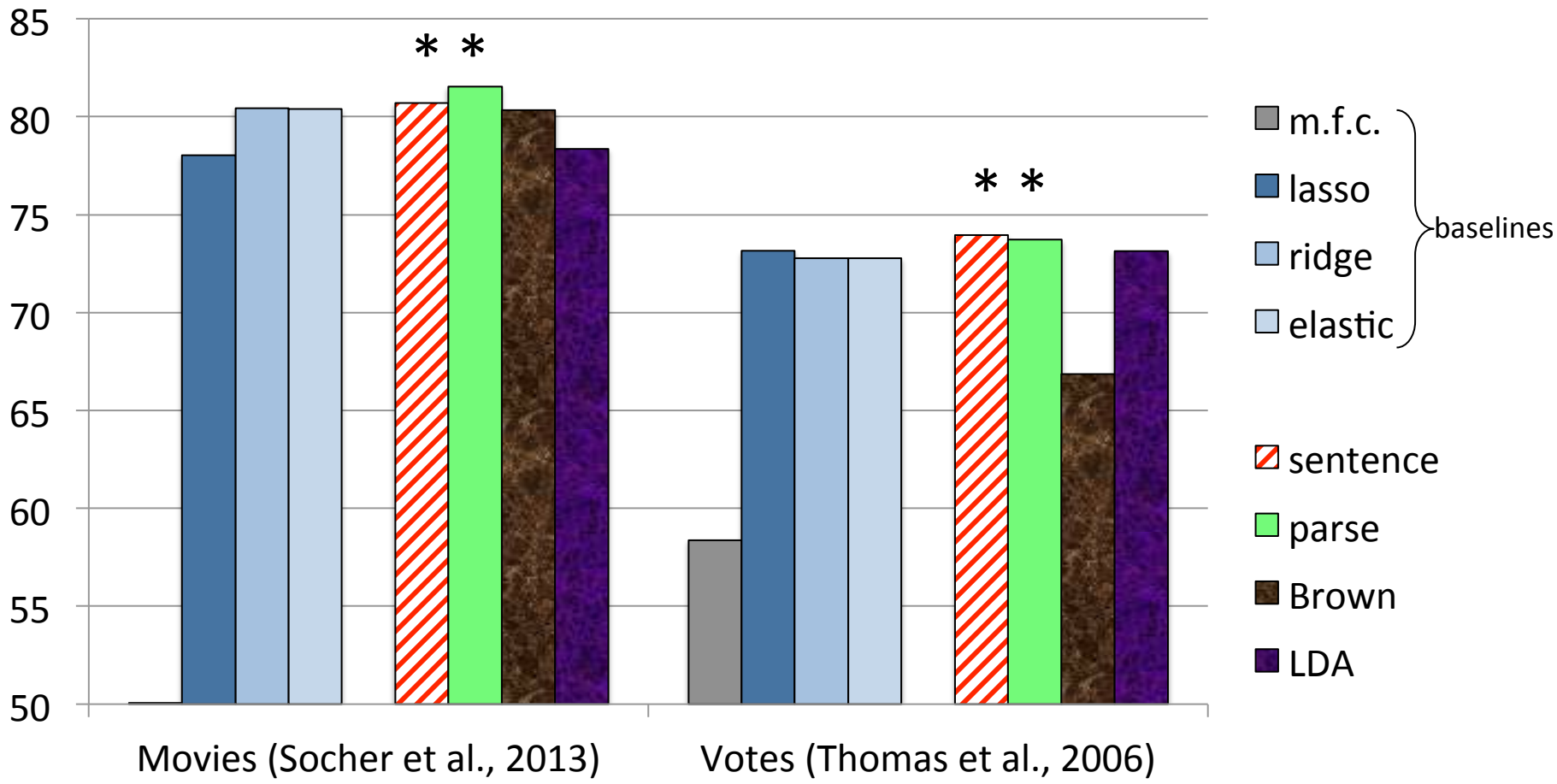
$$\frac{p(y=1 \mid d)}{p(y=1 \mid d \setminus s)}$$

1.52	this film is one big joke : you have all the basics elements of romance (love at first sight , great passion , etc .) and gangster flicks (brutality , dagerous machinations , the mysterious don , etc.) , but it is all done with the crudest humor .
1.01	it ’ s the kind of thing you either like viserally and immediately ’’ get ’’ or you don ’ t .
1.01	that is a matter of taste and expectations .
1.02	i enjoyed it and it took me back to the mid80s , when nicolson and turner were in their primes .
1.00	the acting is very good , if a bit obviously tongue - in - cheek .

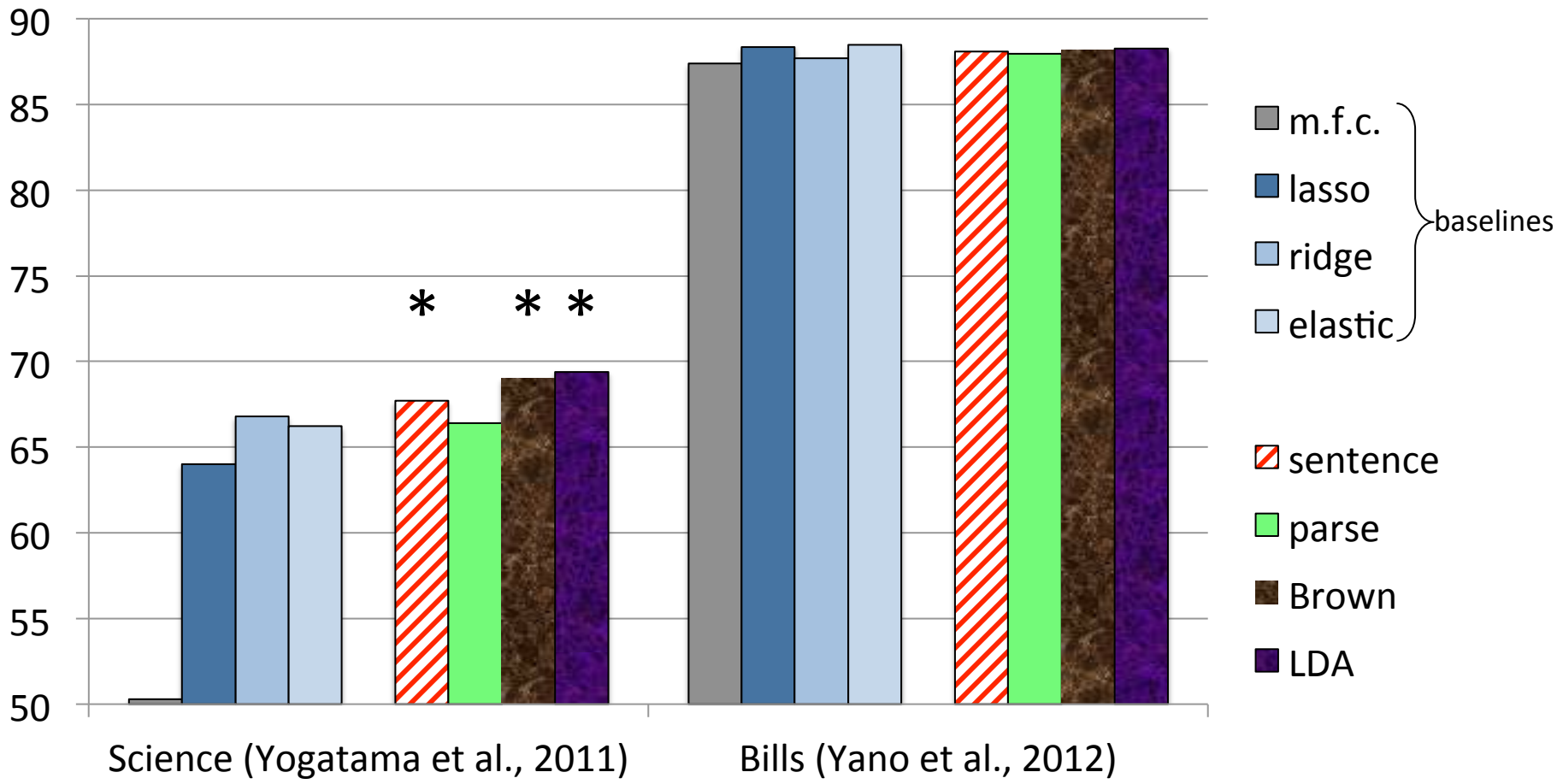
Classification Experiments

- L: Bag of words logistic regression
- Baselines: m.f.c., lasso, ridge, elastic
- Eight datasets

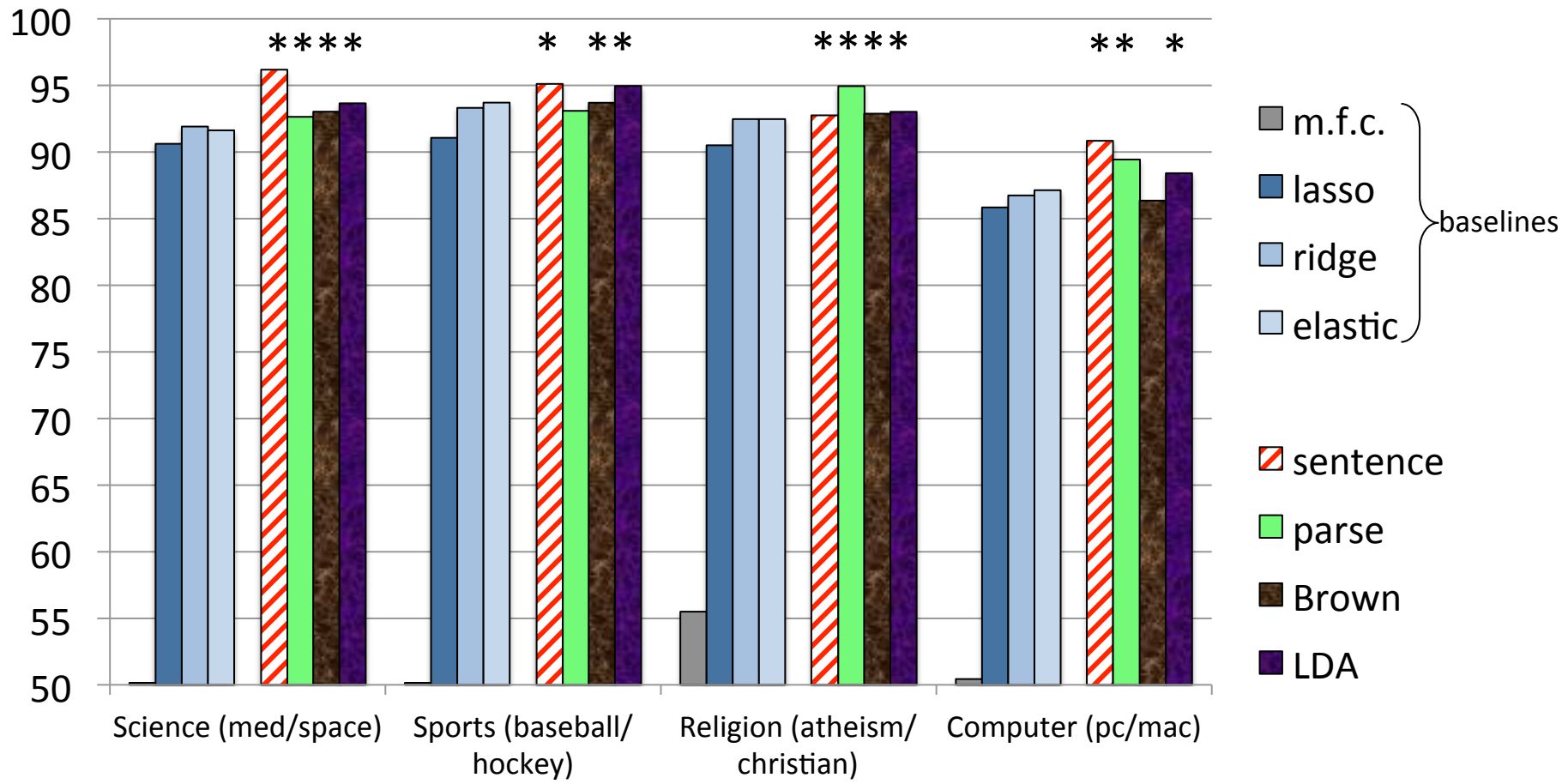
Sentiment


85
80
75
70
65
60
55
50
* *
* *
Movies (Socher et al., 2013)
Votes (Thomas et al., 2006)
m.f.c.
lasso
ridge
elastic
baselines
sentence
parse
Brown
LDA

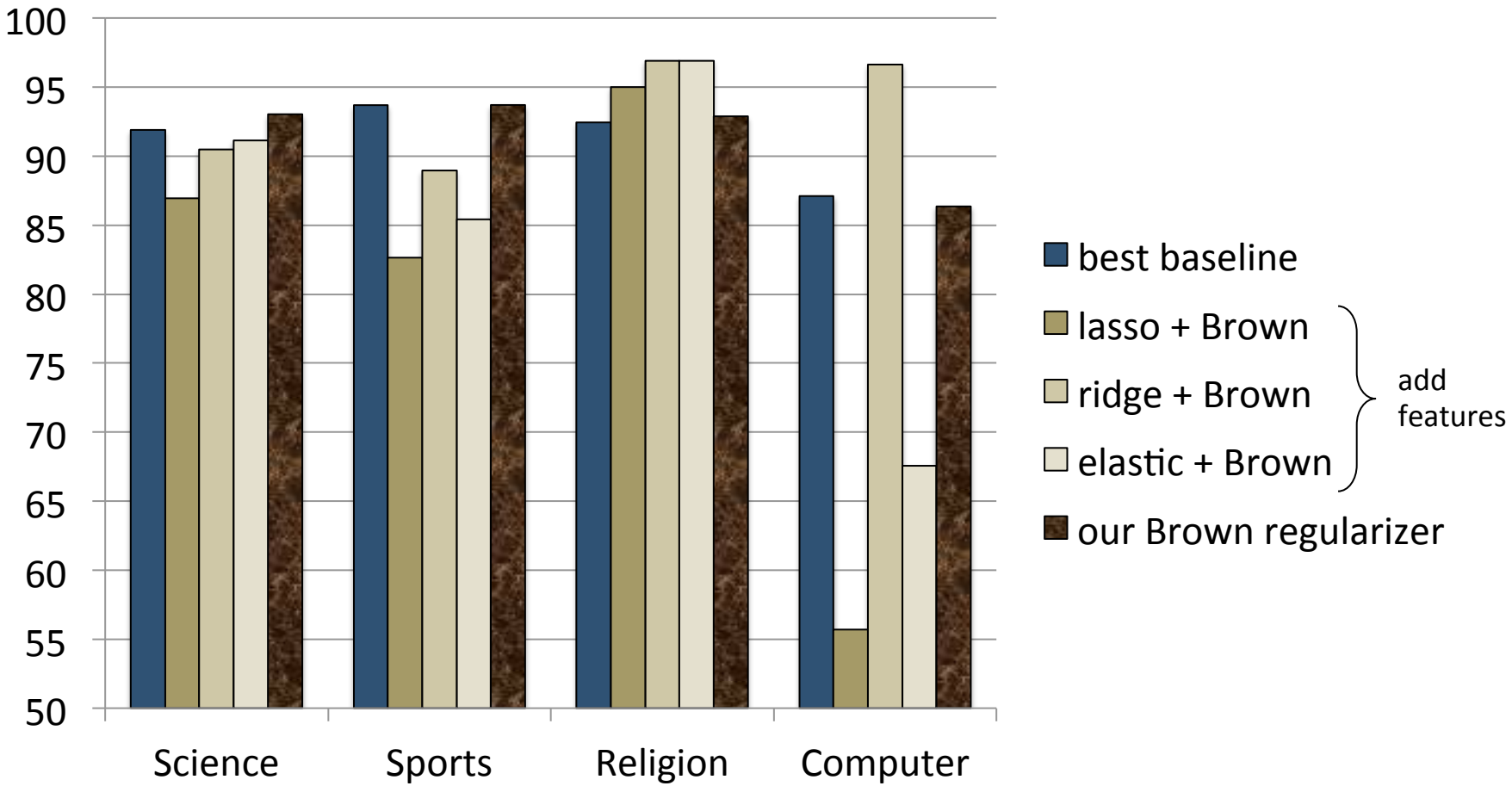
Forecasting



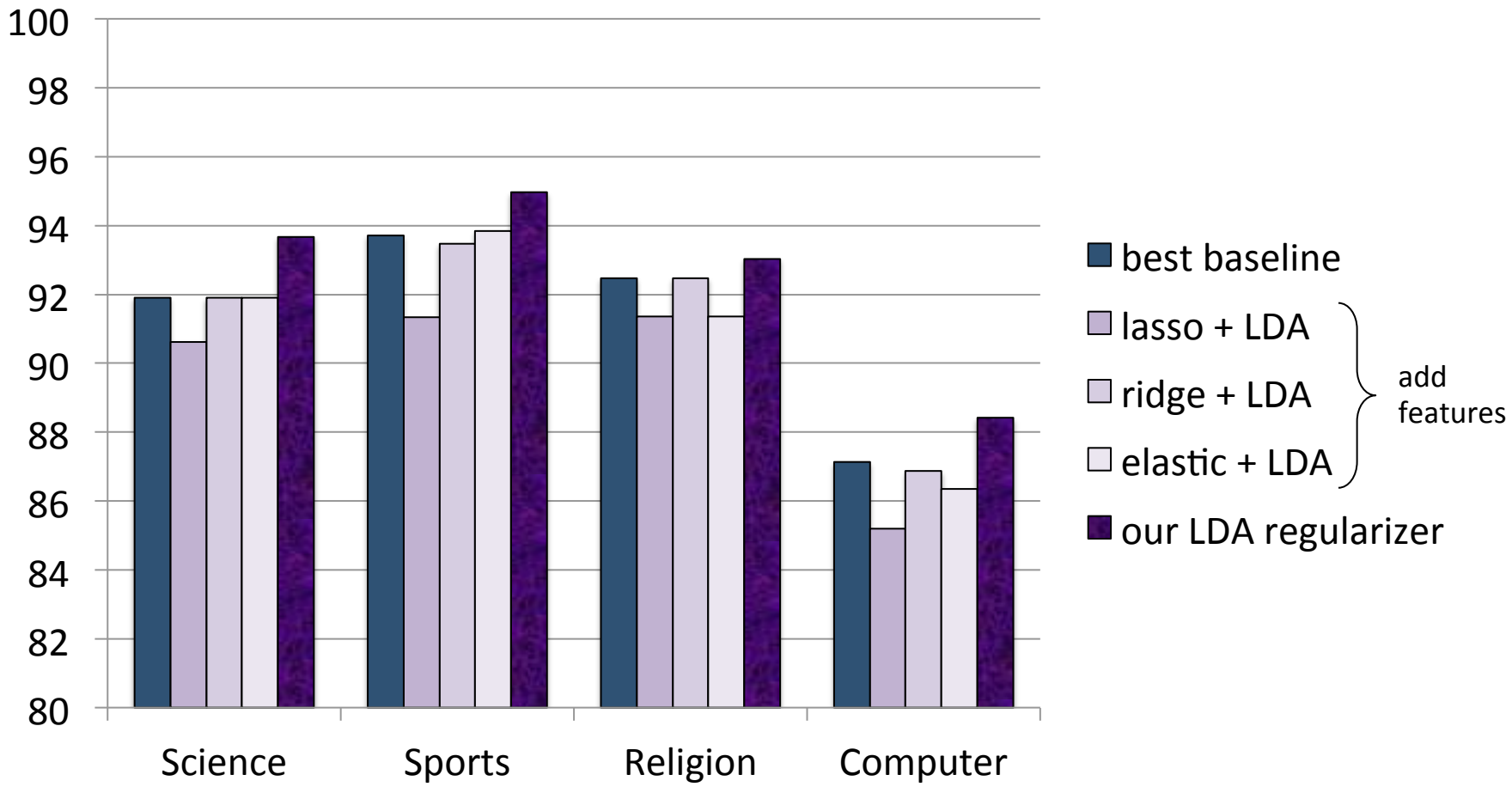
20 Newsgroups Binary Tasks



Brown as features or regularizer?



LDA as features or regularizer?



Summary

- Words of a feather (should) flock together
- Idea: use linguistic structure to define *feathers* (flocks) instead of features
- Math: sparse group lasso regularization
- Results: text classification (topics, sentiment, forecasting)

Acknowledgments: Google, IARPA, Pittsburgh Supercomputing Center