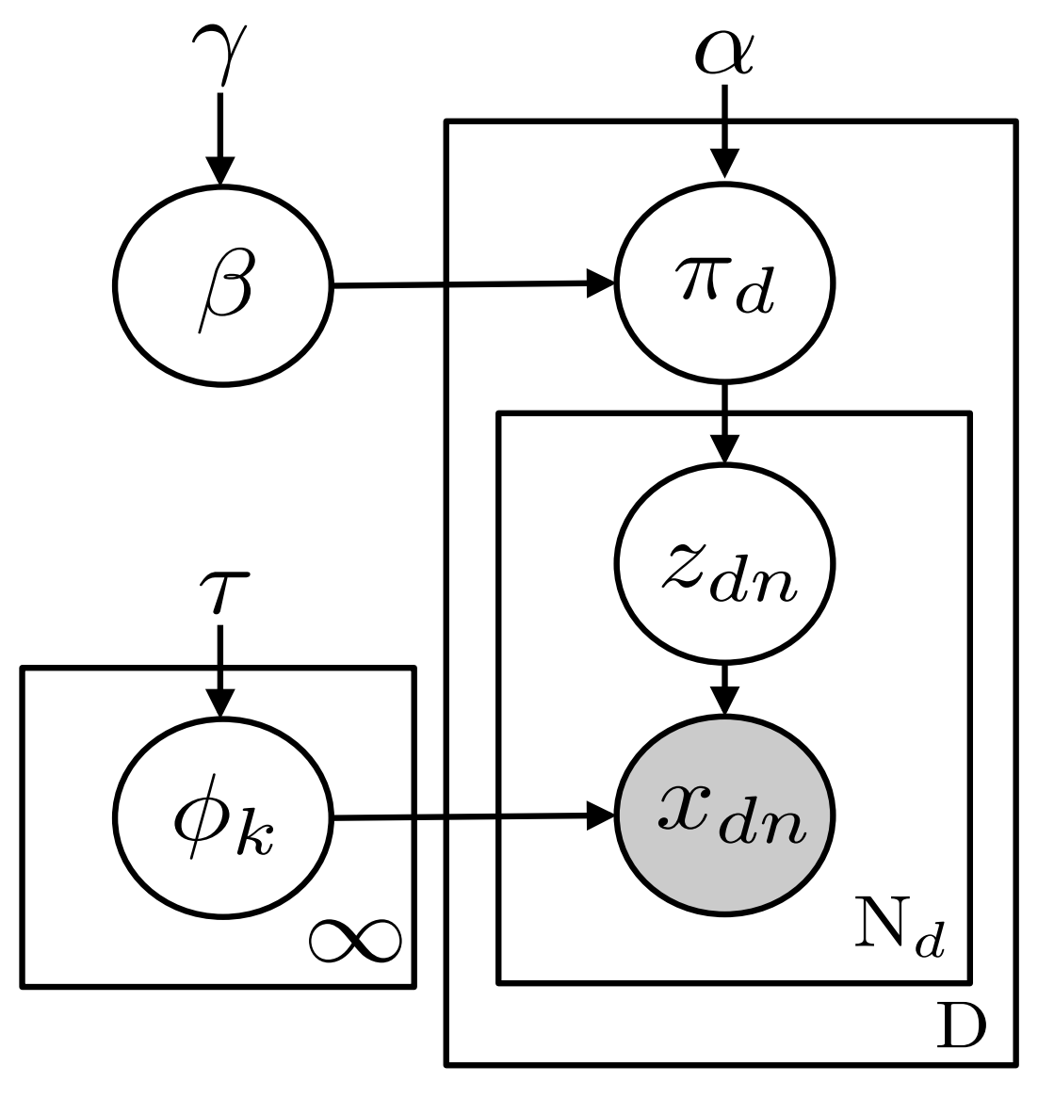
Reliable and scalable variational inference for the hierarchical Dirichlet process

Python code
bitbucket.org/michaelchughes/bnpy-dev

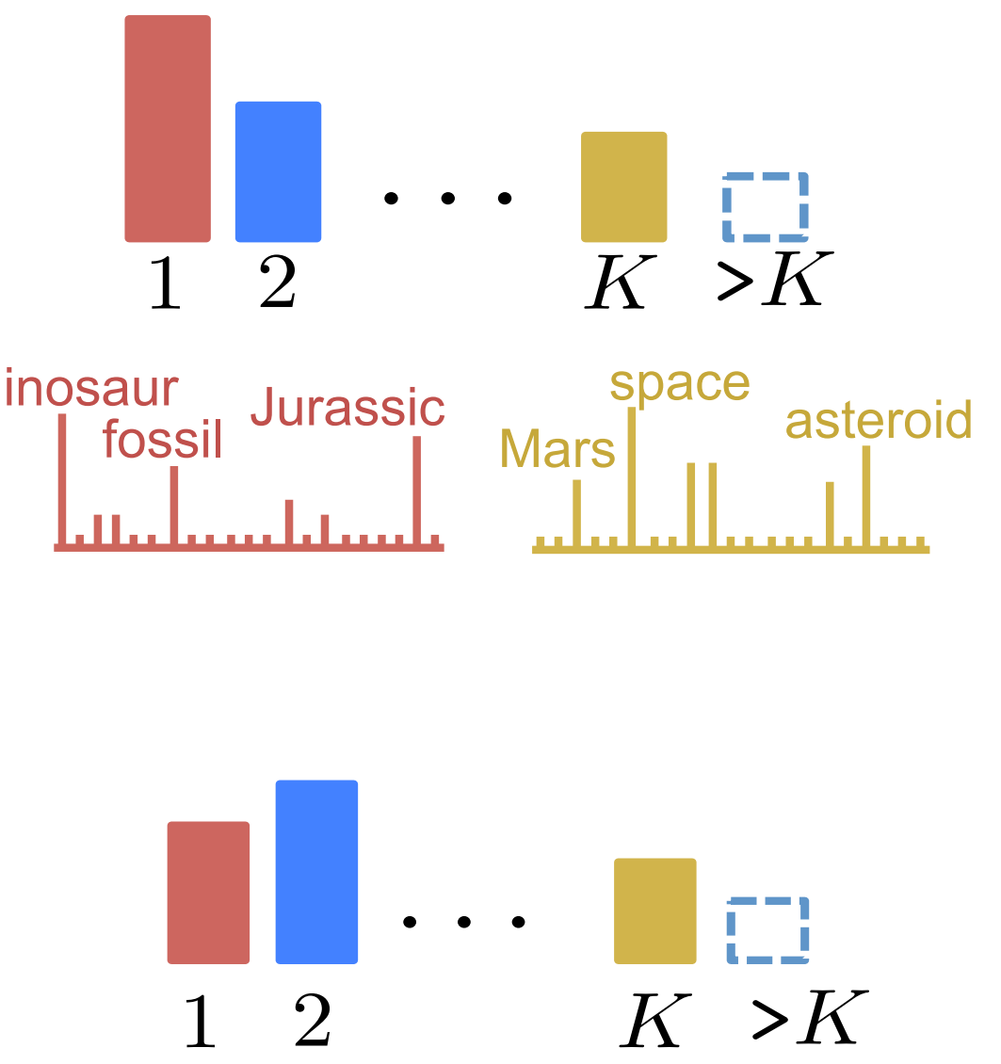
Michael C. Hughes, Dae Il Kim, & Erik B. Sudderth *Dept. of Computer Science, Brown University*

HDP topic model

- HDP prior: data-driven learning of number of topics K
- Our direct assignment representation better than alternatives



Global **Local** document-specific



p(topic) β

p(word | topic) ϕ_k

p(topic | doc.) π_d

token indicator $z_{dn} \in \{1, 2, \ldots K\}$

observed word token x_{dn}

Scalable variational inference

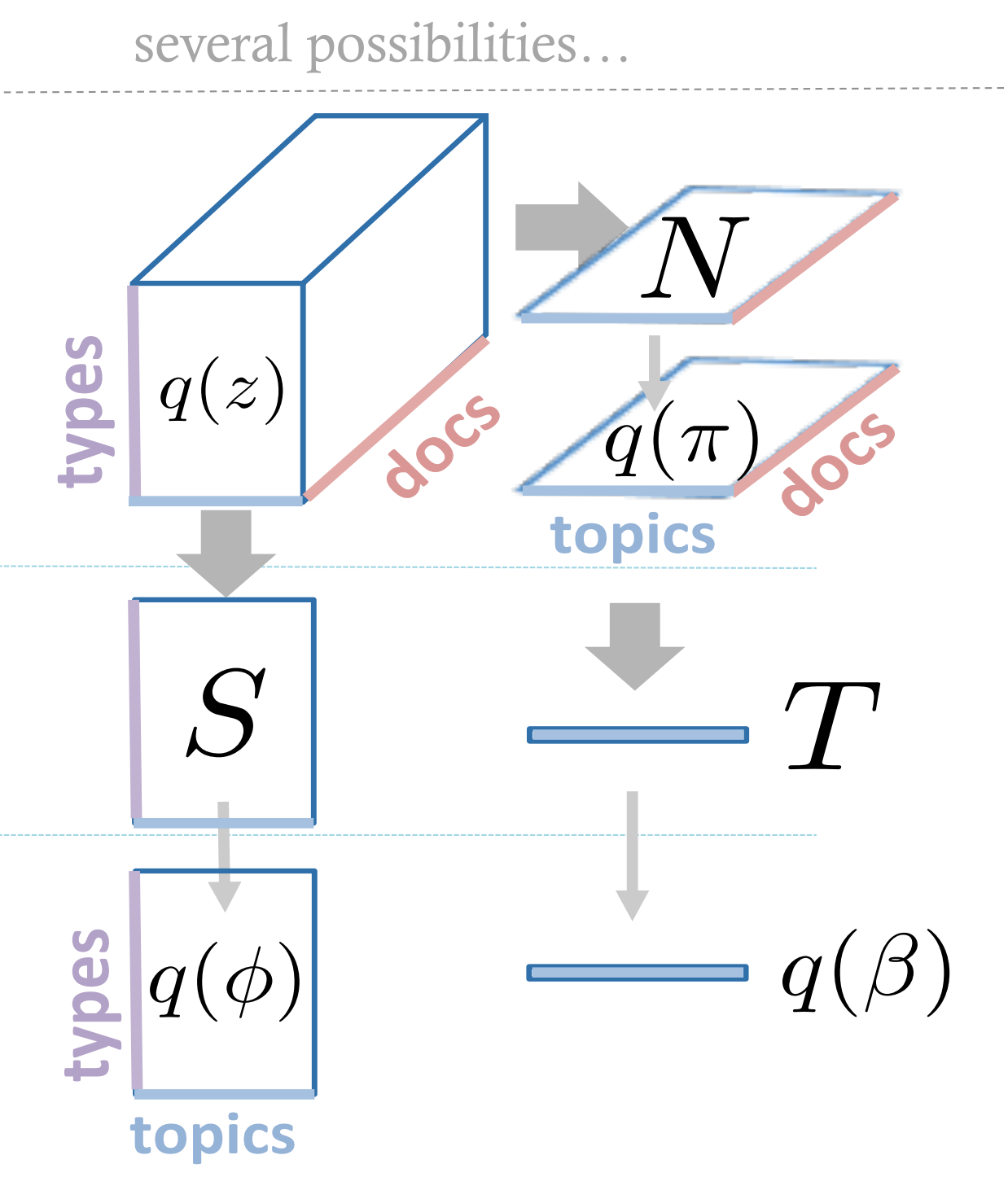
New variational objective

Goal: Find approximate factorized posterior

$$q(\phi)q(\beta)q(\pi)q(z) \approx p(\phi, \beta, \pi, z|x)$$

Algorithm template

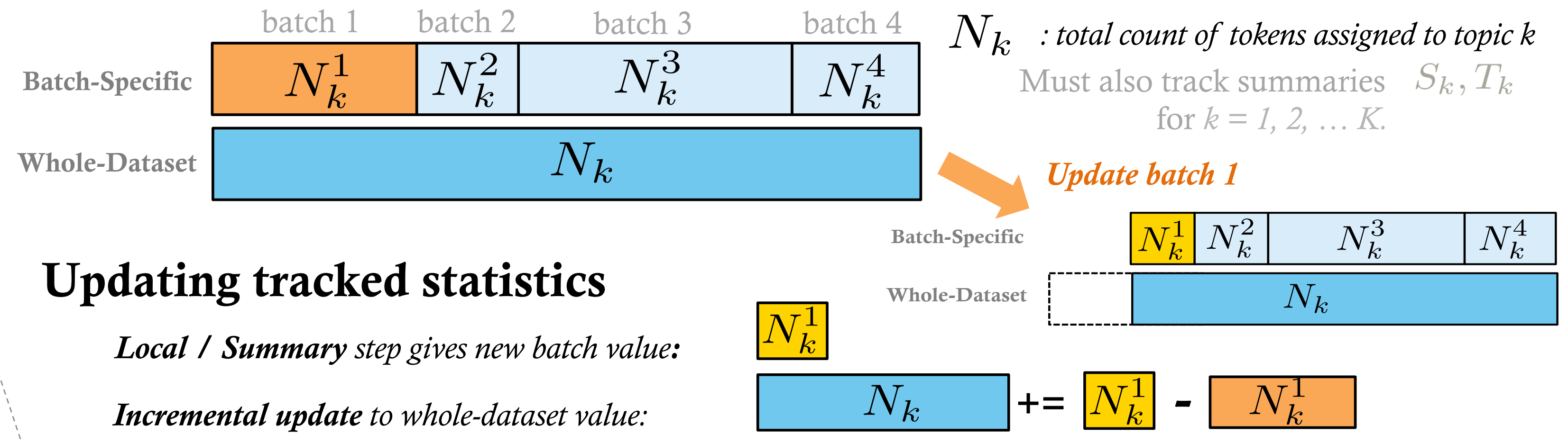
```
Initialize global factors $q(\phi)\, q(\beta)$
Loop until converged:
    a) For each batch in dataset:
        1) Local step
        2) Summary step
        3) Global step
    b) Try merge proposals
    c) Try delete proposals
```



Memoized algorithm

Hughes & Sudderth, NIPS '13
Neal & Hinton '99

- As scalable as stochastic, without pesky learning rate.
- Requires tracking statistics for each batch & topic.



N_k *: total count of tokens assigned to topic k*
Must also track summaries S_k, T_k for $k = 1, 2, \ldots K$.

Updating tracked statistics

***Local / Summary** step gives new batch value:*

***Incremental update** to whole-dataset value:*

Stochastic algorithm

Hoffman, Blei, et al '12

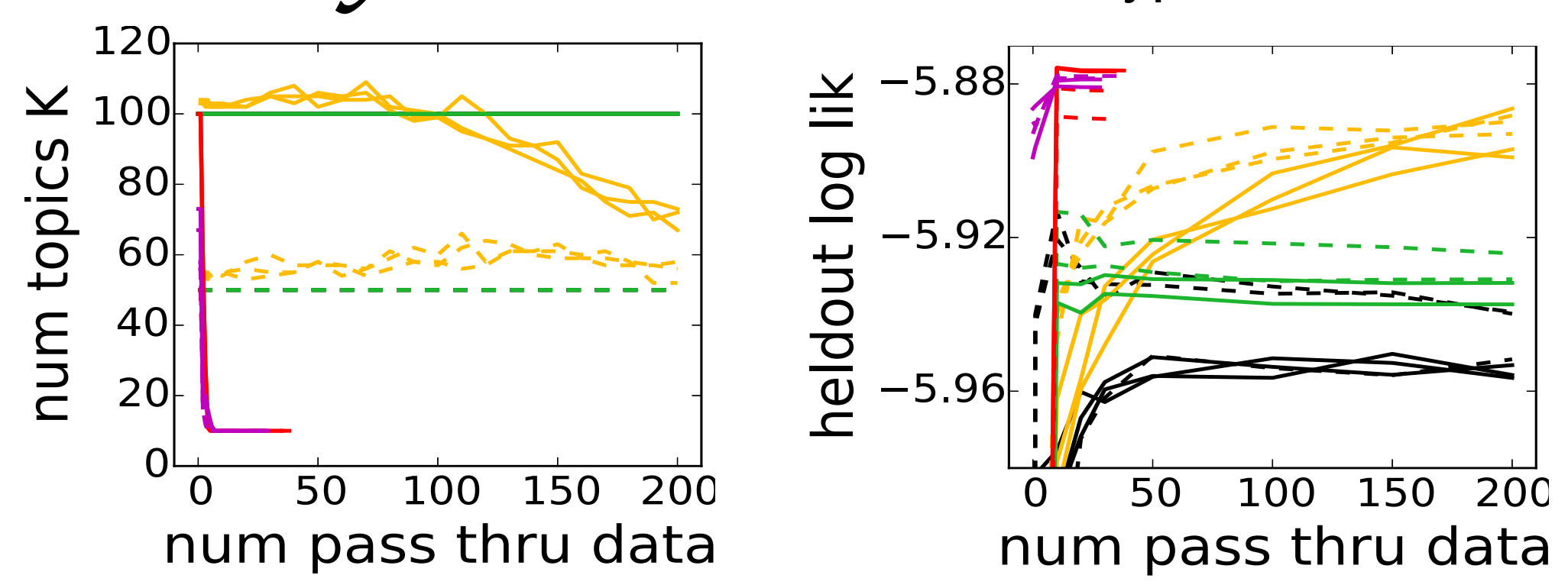
- Natural gradient descent for global step update.
- Less effective for merges/deletes. *Can't exactly check whole-dataset objective.*

Experiments

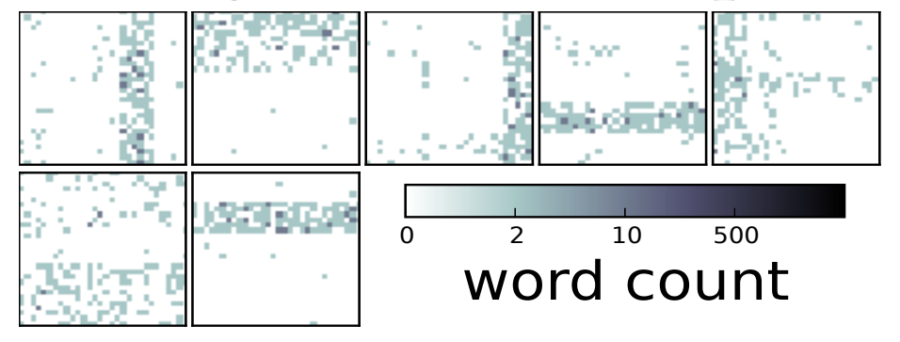
- Memoized alg. with merges/deletes rapidly finds small set of high-quality topics.
- Other algorithms get stuck quickly or improve very slowly.

Toy bars

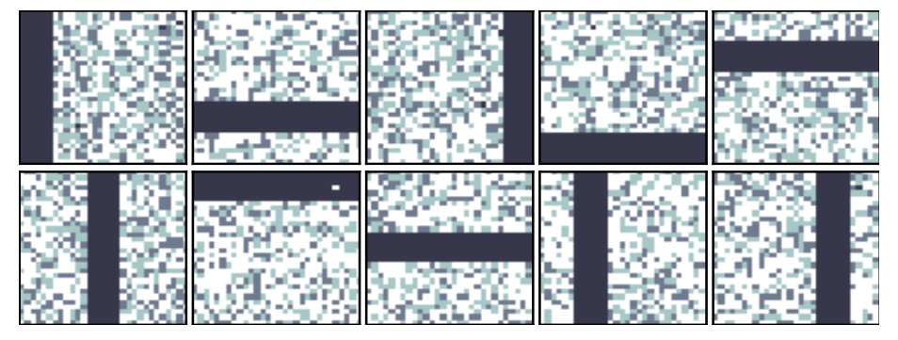
1000 docs
900 vocab types



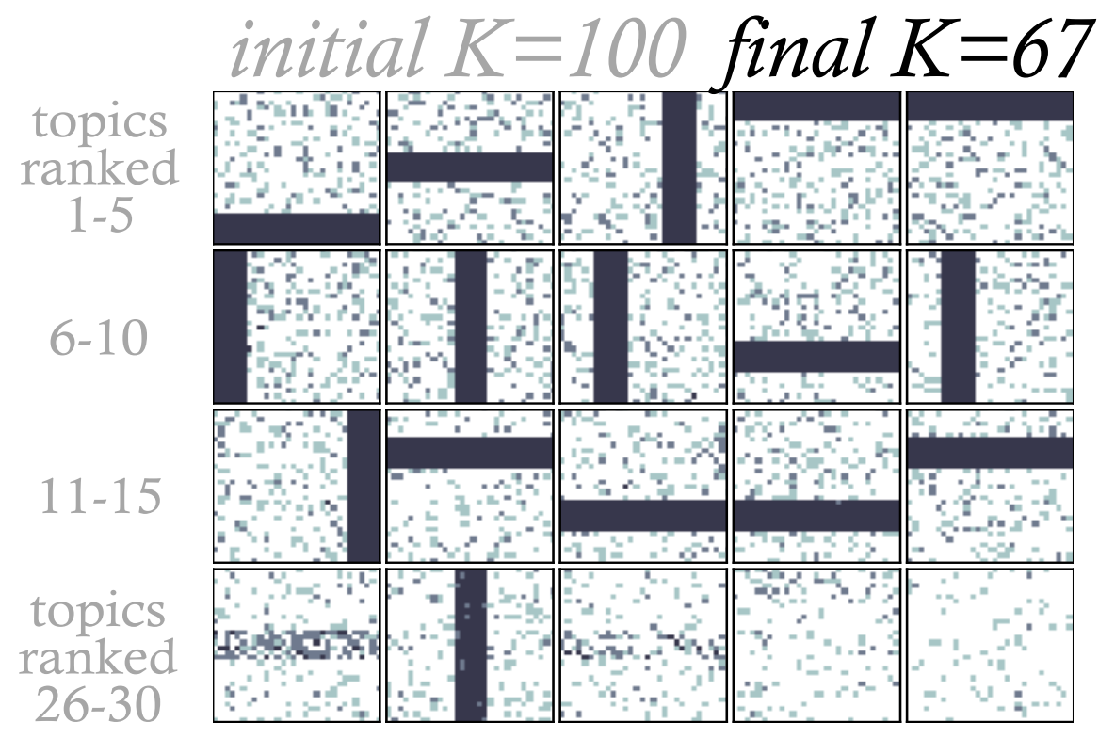
Example documents
Drawn from 10 true topics.



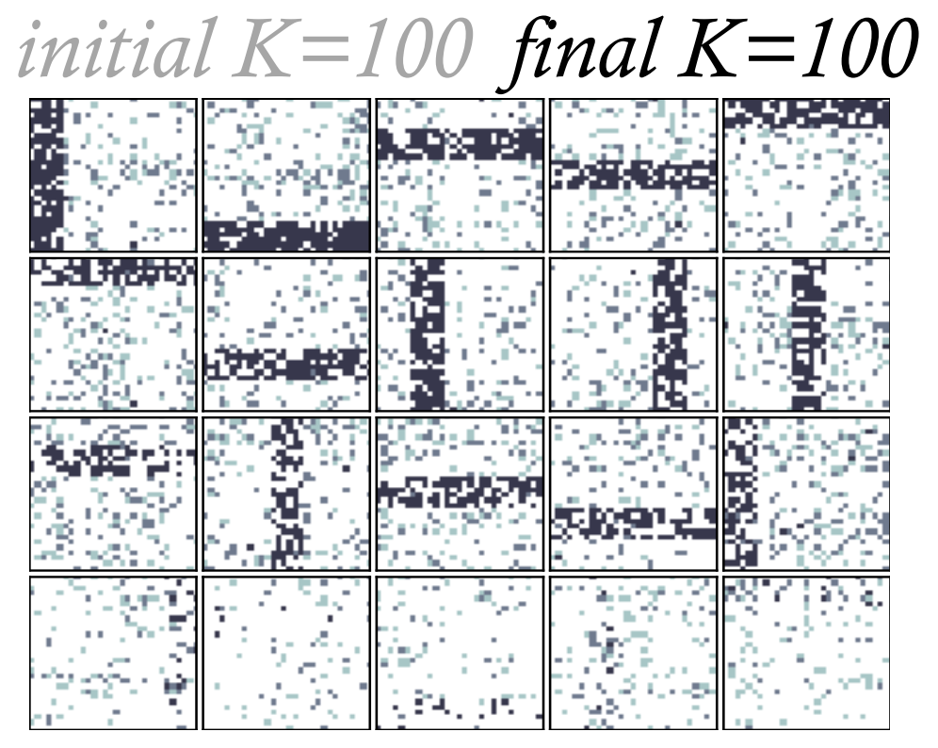
memo + delete & merge
initial K=100 final K=10



Gibbs
initial K=100 final K=67

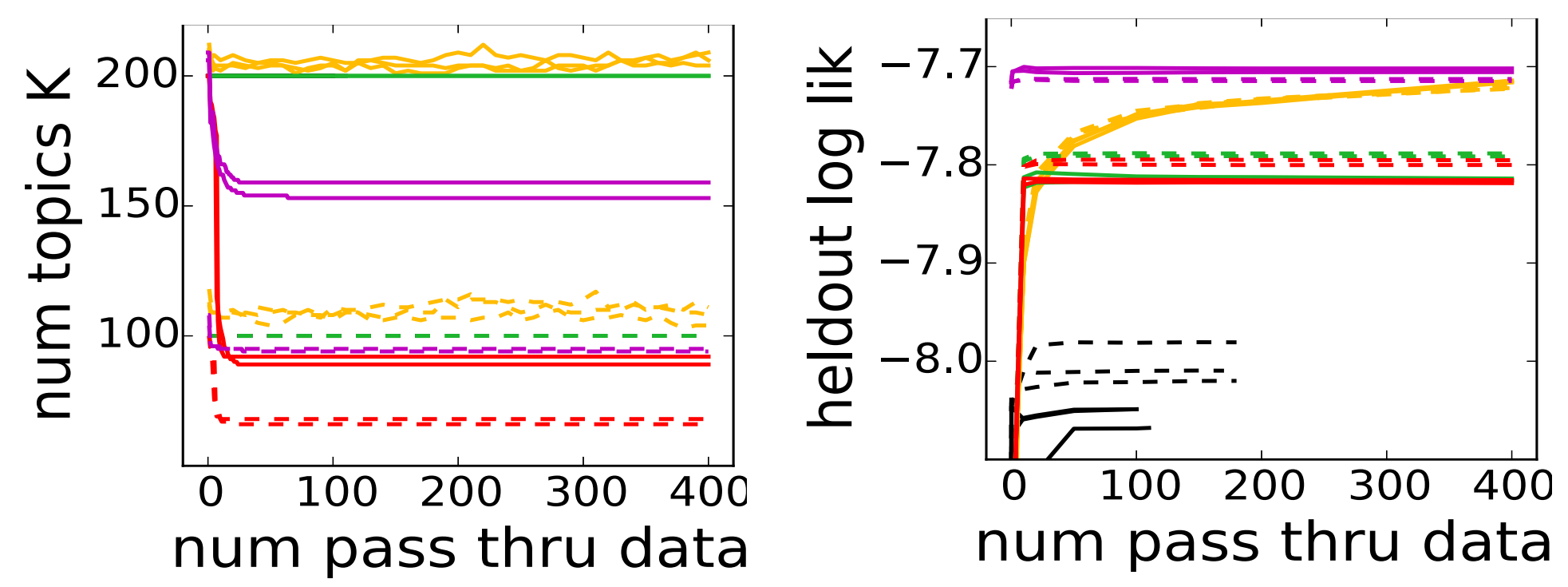


memoized
initial K=100 final K=100



Wiki articles

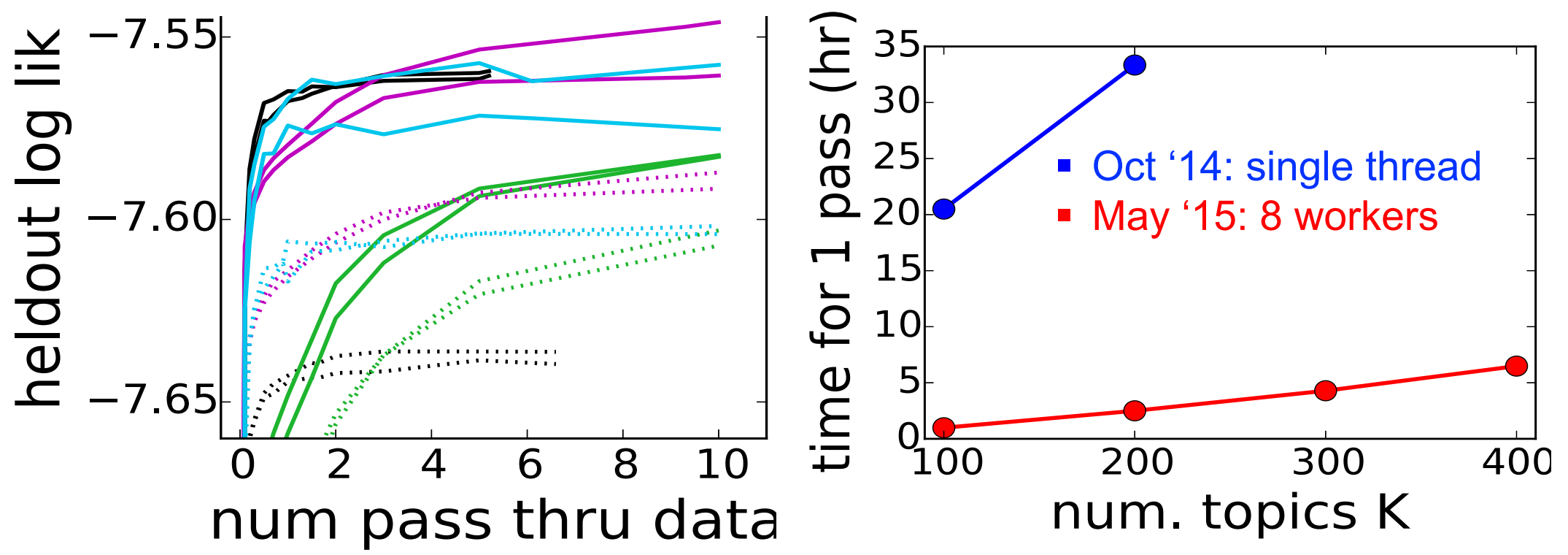
7961 docs
6131 vocab types



Legend:
- memo + delete & merge, init=smart
- memo + delete & merge, init=random
- memoized, init=random
- stochastic, init=random
- Gibbs sampler *Teh et al. '06*
- stochastic (CRF) *Wang et al. '11*

NY Times articles

1.8 million docs
8000 vocab

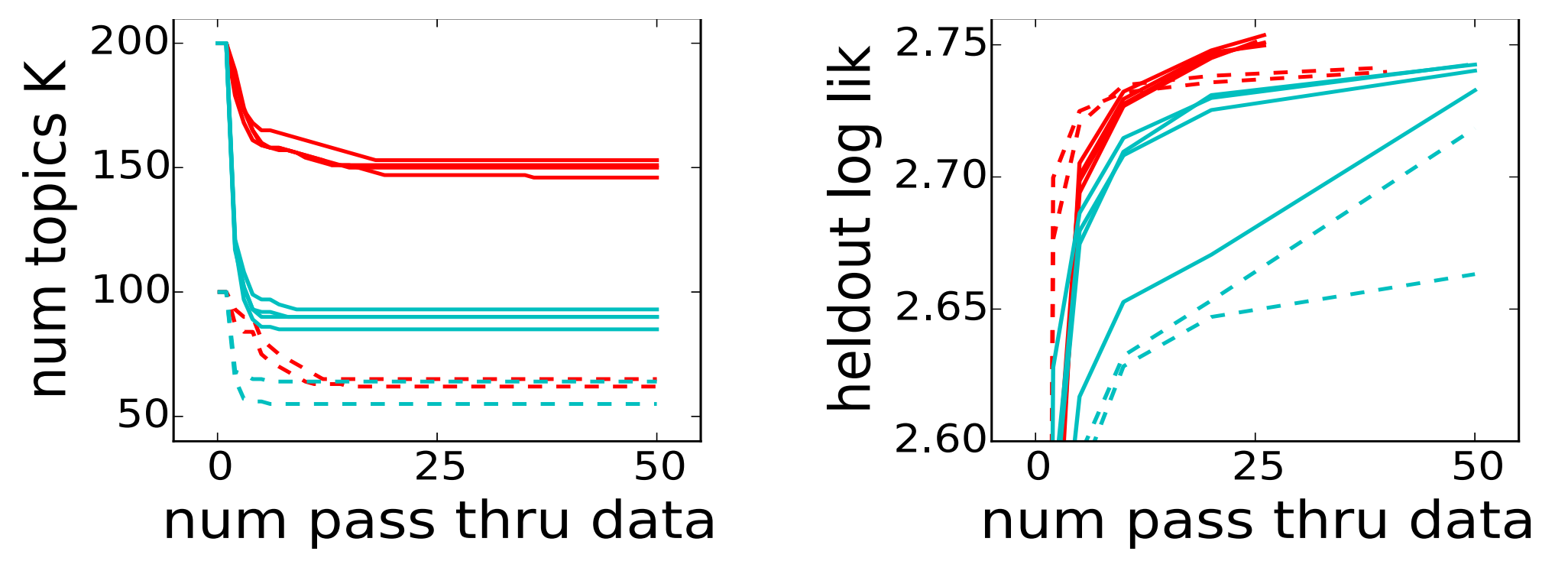


Recent parallelization of code makes large-scale analysis possible.

Image patches

3 million 8x8 patches
from 400 images

Model comparison:
- image-specific frequencies (HDP admixture)
- universal frequencies (DP mixture)



Patch samples from trained model
Showing top 4 clusters for each image, ranked out of a shared set of K=200.

Reliable inference

- Algorithm should recover similar compact set of topics, regardless of initialization.
- Algorithm should avoid local optima & remove useless junk topics.

Model selection

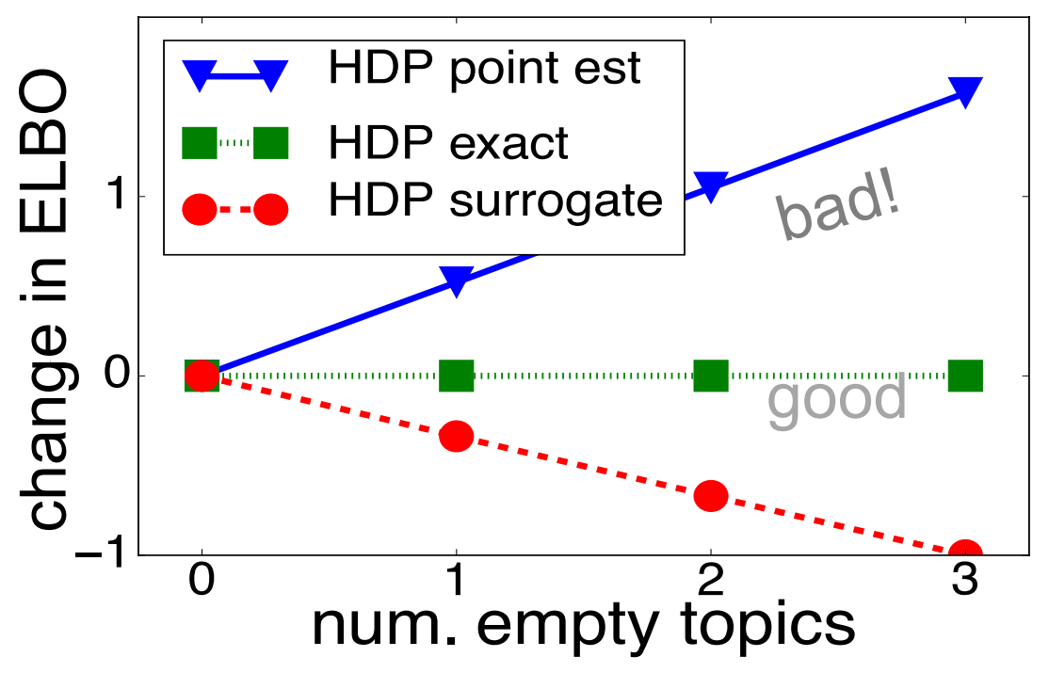
Chosen form of $q(\beta)$is important.

- *MAP Point Estimate:* $q(\beta) = \delta_{\beta^*}$ *Liang et al. '07* *Bryant & Sudderth '12*
  Fails to penalize empty topics effectively.
- *Full distribution:* $q(\beta) = \text{StickBreaking}(\hat{\rho}, \hat{\omega})$
  Integrate away all parameters that grow with K.

Train on toy data with assignments **fixed** to truth, with extra empty topics.
Goal: does objective increase or decrease as more empty topics added?

$q(z)$

true topic assignments	empty topics
0 1 0 … 0	0…0
1 0 0 … 0	0…0
0 0 0 … 1	0…0
0 0 1 … 0	0…0



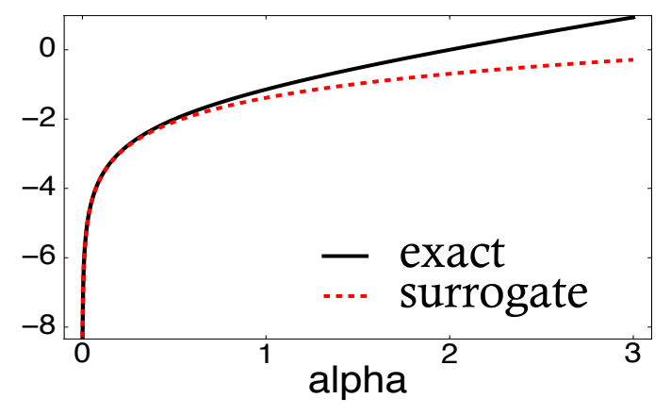
Surrogate objective

New function lower bounds intractable ideal objective.
Penalizes junk topics; key to merge/delete moves.

$$\log\Gamma(\alpha) - \sum_{k=1}^{K+1} \log\Gamma(\alpha\beta_k) \geq K\log\alpha + \sum_{k=1}^{K+1} \log\beta_k$$

Dirichlet log norm. constant
Lacks closed-form expectation

Tight lower bound
Expectation w.r.t. $q(\beta)$ *easy*
Holds for all $\alpha > 0$



Nested truncation

Only assign tokens to first K topics of infinite set.

$$q(z_{dn}) = [r_{dn1}\ r_{dn2}\ \ldots\ r_{dn7}\ r_{dn8} \,\vdots\, 0 \quad \ldots] \quad K{=}8$$

Topics > K are conditionally independent of data.
Need not be represented during inference.

Easy to contract truncation level.

$$q(z_{dn}) = [r_{dn1}\ r_{dn2}\ \ldots\ r_{dn7} \,\vdots\, 0 \quad 0 \quad \ldots] \quad K{=}7$$

Makes merge & delete possible.

Track probability of all inactive topics ($k > K$).

$$q(\pi_d) = \text{Dirichlet}_{K+1}(\theta_{d1}, \theta_{d2}, \ldots \theta_{dK}, \boxed{\theta_{d>K}})$$

Delete move

- Junk topic mass reassigned among *all* remaining topics.
- More flexible than merge, but only scales with smaller topics.
- *Requires extra local step on small target dataset.*

Accepted Delete	engineering science computer field machine mechanical	math function theorem define theory	science theory scientific mathematics scientist	code language computer program programming	process theory human information method	design engine build speed drive
doc A	~~75~~	16.0	42.7	17.6	0	0
doc B	~~72.7~~	0	40.8	0	20.6	11.3
doc C	~~35.9~~	0	0	0	35.9	0
doc D	~~83.4~~	0	36.1	30.6	16.7	0

Document-specific reassignment via local step

Net change in doc-topic count N_{dk} *after delete*

Merge move

- Redundant pair of topics combined into one single topic.
- Exact evaluation of proposal possible via tracked summaries.
- *No extra local step required, only a few pair-wise statistics.*

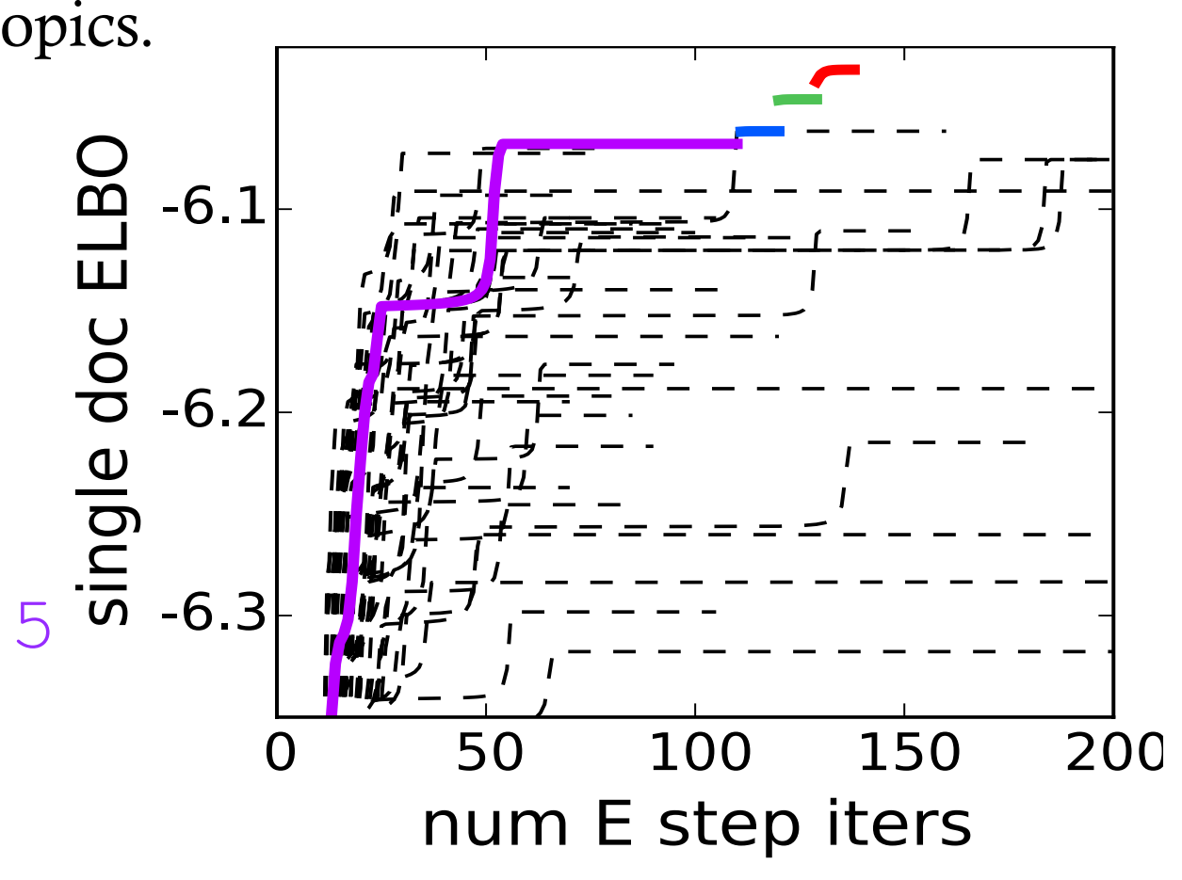
Accepted Merge		Accepted Merge	
series song release star television york award friend	film magazine direct production actor career hollywood appeared	language latin letter dialect speak speaker sound verb	linguistic linguist language speech linguistics grammatical pronunciation suffix

Sparse restarts in local step

New move for escaping local optima at each doc.
- Propose zero values for small-mass topics.
- Accept if improves obj. function.

89.5	64.8	0	0	0
90.8	64.2	8.4	0	0
89.8	62.4	8.1	8.1	0
88.7	61.6	7.2	7.1	5.5

N_{dk} doc-topic counts for select topics


single doc ELBO
num E step iters