Stochastic Divergence Minimization for Online Collapsed Variational Bayes Zero Inference of Latent Dirichlet Allocation

@KDD2015

Issei Sato & Hiroshi Nakagawa

The University of Tokyo

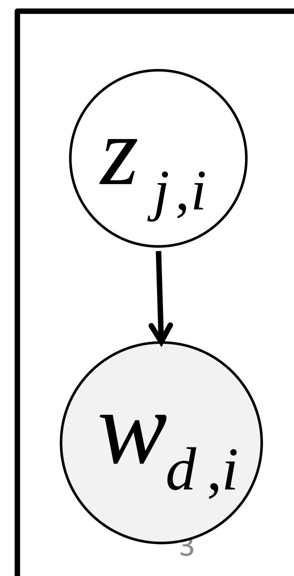
Latent Dirichlet allocation [Blei,2003]

The annual ACM SIGKDD conference is the premier international forum for data mining researchers and practitioners from academia, industry, and government to share their ideas, research results and experiences.

Latent Dirichlet allocation [Blei,2003]

The[1] annual[2] ACM[2] SIGKDD[3] conference[2] is[1] the[1] premier[2] international[2] forum[2] for[1] data[3] mining[3] researchers[10] and[1] practitioners[10] from[1] academia[10], industry[8], and[1] government[7] to[1] share[6] their[1] ideas[10], research[10] results[5] and[1] experiences[5].

Modeling co-occurrence:
Frequently co-occurring words
are assigned to the same topic (color)



De Finetti theorem [De Finetti, 1930s]

A sequence of random variables (*x1*, *x2*, . . .) is infinitely exchangeable iff, for all *n*

$$p(x_1, x_2, \ldots, x_n) = \int \prod_{i=1}^{n} p(x_i \mid \theta) \, p(\theta) d\theta$$

Latent Dirichlet allocation [Blei,2003]

Topic distribution for each document

$$\boldsymbol{\theta}_d \sim \mathrm{Dir}(\boldsymbol{\gamma})\ (d = 1, \cdots, D)$$

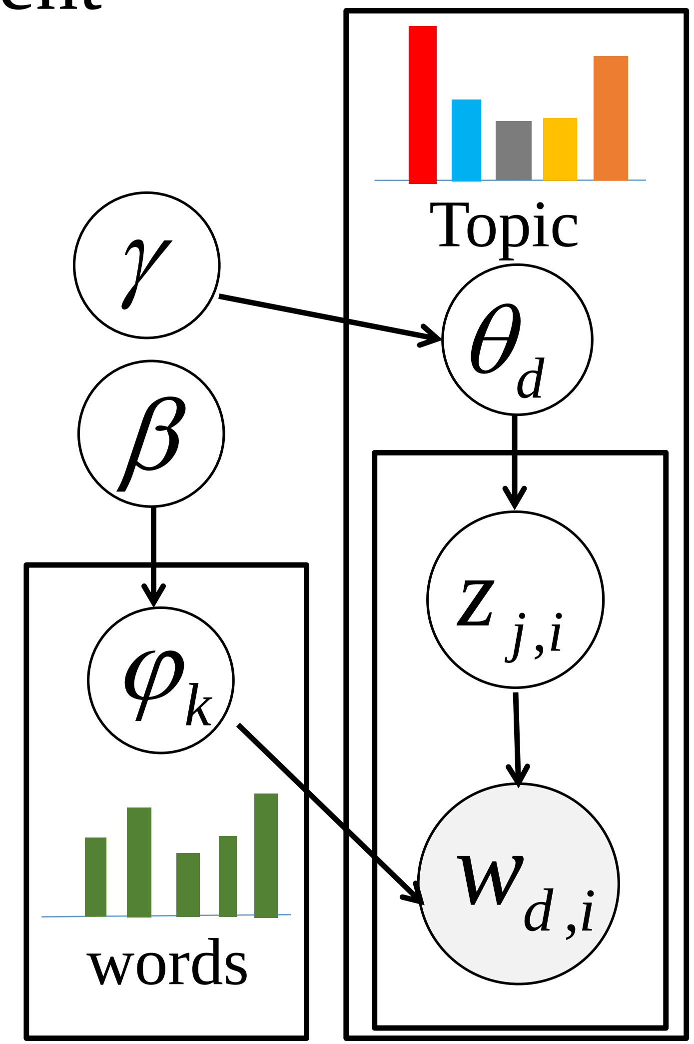
Word distribution for each topic

$$\boldsymbol{\phi}_k \sim \mathrm{Dir}(\boldsymbol{\beta})\ (k = 1, \cdots, K)$$

For each words:

$$z_{d,i} \sim \mathrm{Multi}(\boldsymbol{\theta}_d)$$

$$w_{d,i} \sim \mathrm{Multi}(\boldsymbol{\phi}_{z_{d,i}})$$



Priors Matter [Wallach+,2009]

Asymmetric Dirichlet prior

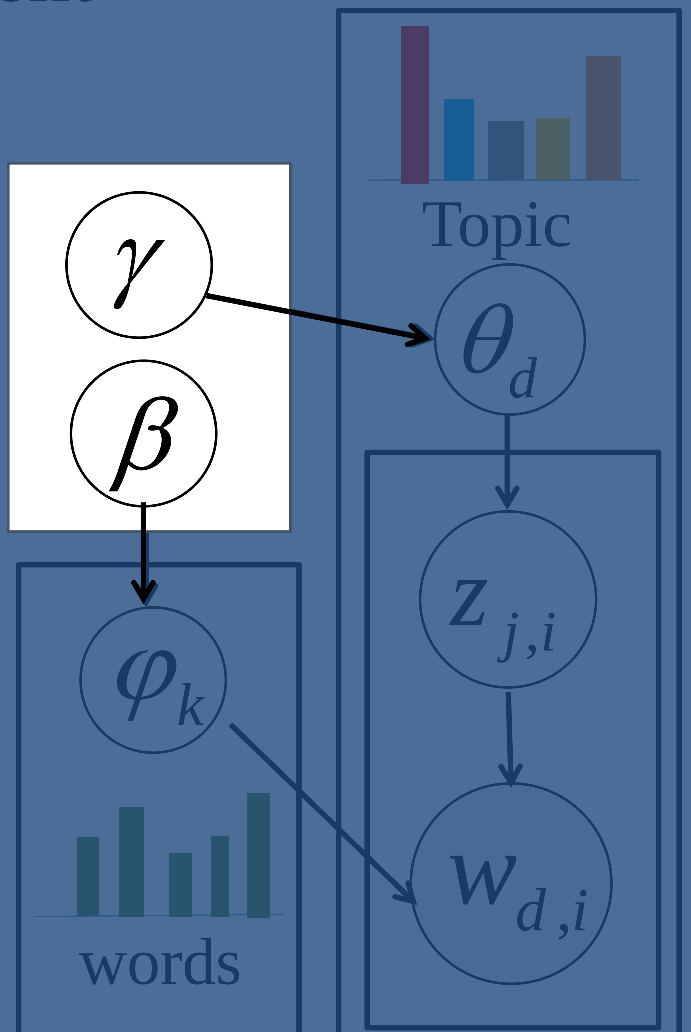
$$\mathrm{Dir}(\gamma_1, \gamma_2, \ldots, \gamma_K)$$

$$\mathrm{Dir}(\beta_1, \beta_2, \ldots, \beta_V)$$

Symmetric Dirichlet prior

$$\mathrm{Dir}(\gamma, \gamma, \ldots, \gamma)$$

$$\mathrm{Dir}(\beta, \beta, \ldots, \beta)$$



Evaluation: Perplexity

Prediction of held-out words

$$\exp\left[\frac{1}{N_{test}}\sum_{w^*\in W_{test}}\log p(w^*|W_{train})\right]$$

$$\sum_{k=1}^{K}\frac{E[n_{k,w^*}^{train}]+\beta}{E[n_{k}^{train}]+V\beta}\frac{E[n_{d,k}^{train}]+\gamma_k}{E[n_{d}^{train}]+\sum_k\gamma_k}$$

Inference algorithms

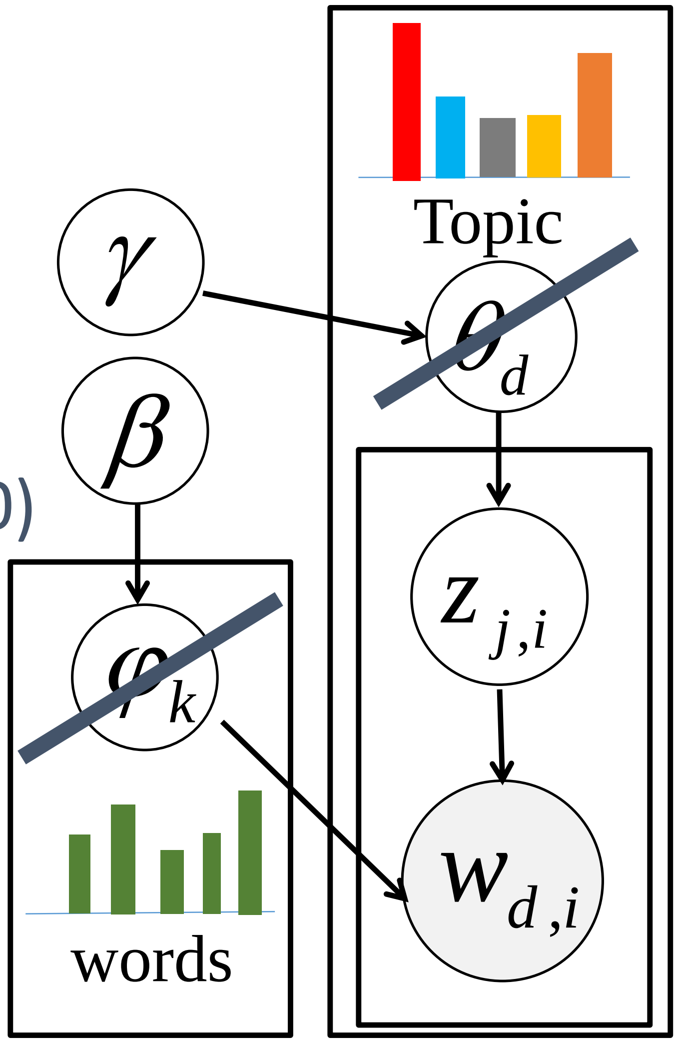
- Variational Bayes (VB)

  [Blei+,JMLR2003]
- Collapsed Gibbs Sampling (CGS)

  [Griffiths+,PNAS2004]
- Collapsed Variational Bayes (CVB)

  [Teh+,NIPS2007]
- Collapsed Variational Bayes Zero (CVB0)

  [Asuncion+, UAI2009]



Marginalize out parameters

Why CVB0 works better?

CVB0 uses zero-order Taylor approximation for expectations in CVB

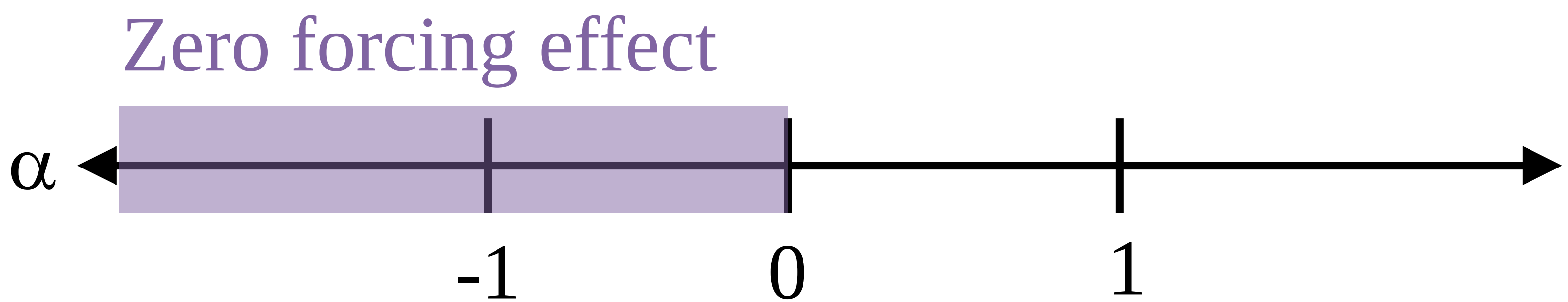
→ CVB0 is less accurate than CVB

CVB0 can be formulated as a local α-divergence minimization

[Sato & Nakagawa, ICML2012]

α-divergence minimization

Inference	Marginalization	α-divergence
VB [Blei+,03]	—	$\alpha \to 0$
CVB [Teh+,07]	✓	$\alpha \to 0$
CVB0 [Asuncion+, 09]	✓	$\alpha \to 1$ ($\fallingdotseq 1$)
EP [Minka+,02]	—	$\alpha \to 1$



The emphasis in the estimation is on high-frequency topics or low-frequency topics is forced to be zero

Stochastic Optimization

Scaling up: Batch data → Sub-sampling

- Variational Bayes (VB) [Blei+,JMLR2003]

→Stochastic Variational Bayes (SVB)

[Hoffman+,Sato+,NIPS2010]

- Collapsed Variational Bayes Zero (CVB0)

[Asuncion+, UAI2009]

→Stochastic Collapsed Variational Bayes Zero (SCVB0)

[Foulds+, KDD2013]

Framework of SCVB0

Problem

How to formulate SO of CVB0
- CVB0 integrates out parameters

Solution

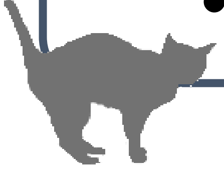
When we manually adjust Dirichlet prior,

CVB0 update ~ MAP update.

SCVB0 ~ Stochastic Approx. of MAP infer.

Question and Problem on SCVB0

- Why MAP works better than VB ?
- We cannot use Asym. Dirichlet prior



Our contribution

- Formulation of SCVB0
  → Stochastic divergence minimization (SDM)
- Estimation of Dirichlet prior
  → Reformulate DM of [Sato+, ICML2012]

Main Idea

[Sato&Nakagawa, ICML2012]

Infer $q(Z) = \prod_{d,i} q(z_{d,i})$ by DM

=CVB0 update

This work

Infer

$$q(Z,W \mid \gamma,\beta) = \prod_{d,i} \underbrace{q(z_{d,i} \mid w_{d,i})}_{\text{=CVB0 update}} \underbrace{q(w_{d,i} \mid \gamma,\beta)}_{?}$$

by DM

Stochastic Approx.

Our

[S

This

$$q(w_{d,i} \mid \gamma, \beta) = \sum_{k=1}^{K} \frac{E[n_{k,w_{d,i}}^{loo}] + \beta}{E[n_k^{loo}] + V\beta} \frac{E[n_{d,k}^{loo}] + \gamma_k}{E[n_d^{loo}] + \sum_k \gamma_k}$$

Leave-One-Out Perplexity

$$\exp\left[-\frac{1}{N} \sum_{d,i} \log q(w_{d,i} \mid \gamma, \beta) \right]$$

Infer

$$q(Z, W \mid \gamma, \beta) = \prod_{d,i} q(z_{d,i} \mid w_{d,i}) \; \boxed{q(w_{d,i} \mid \gamma, \beta)}$$

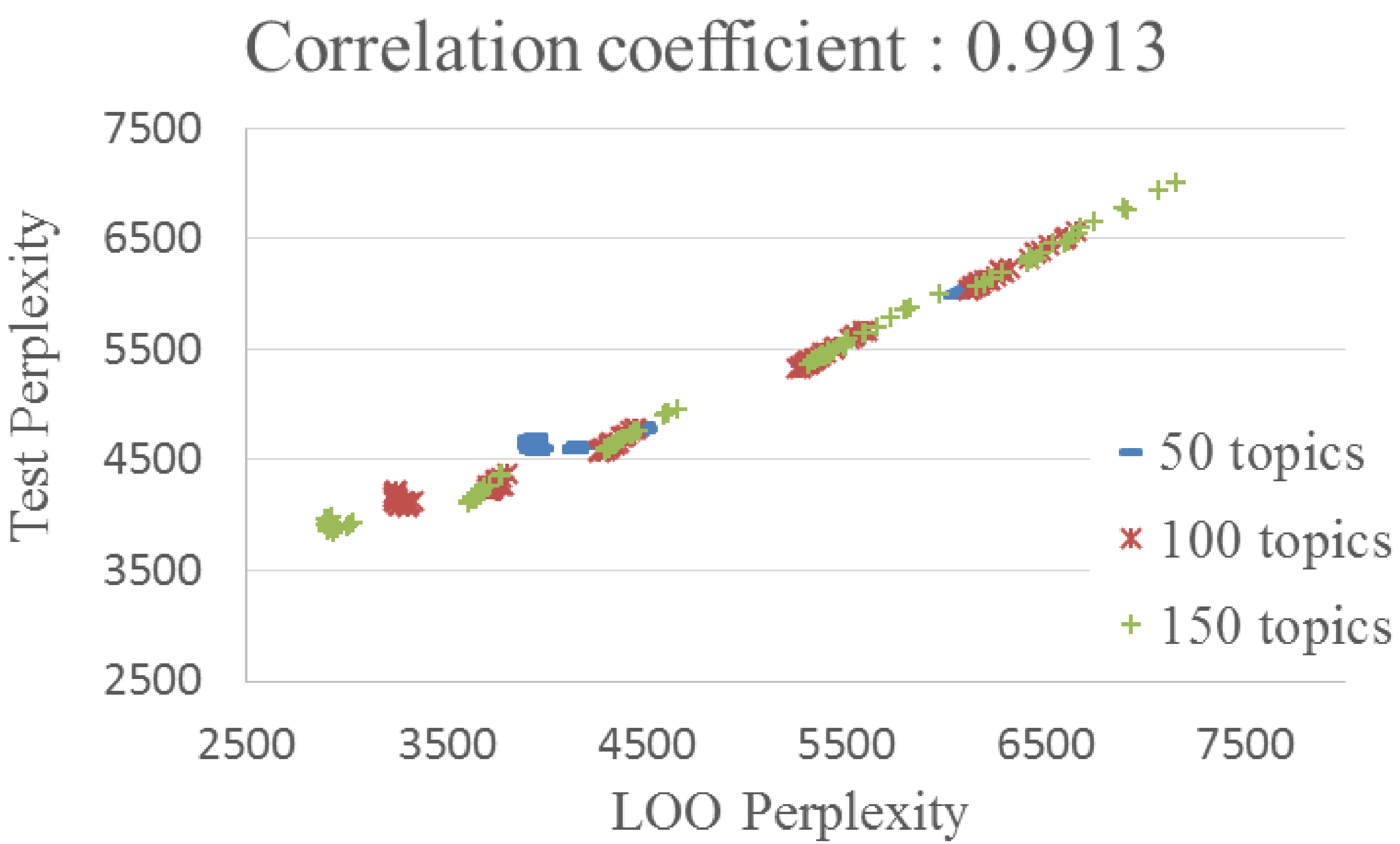
=CVB0 update ?

by DM

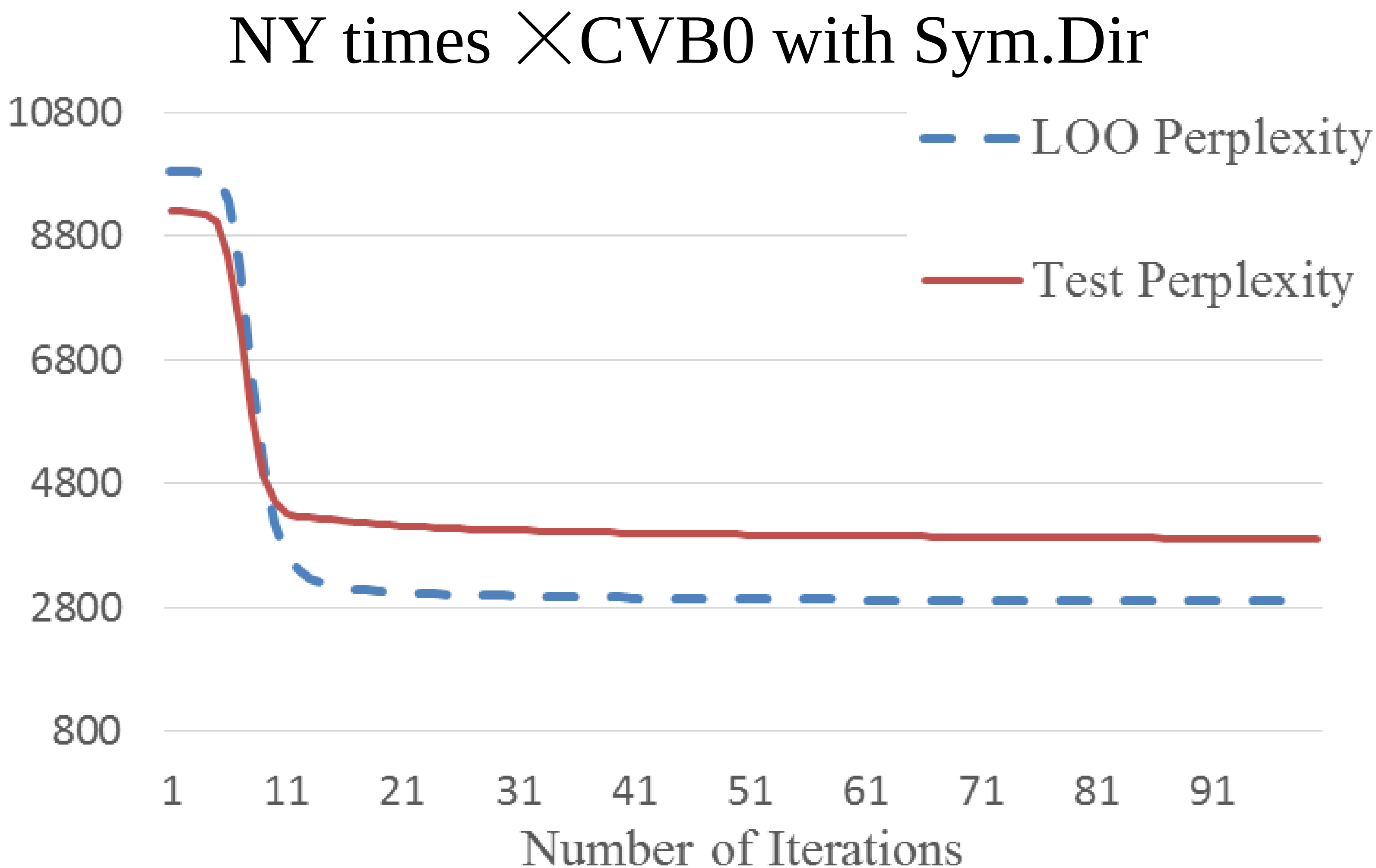
Stochastic Approx.

Testset perplexity × LOO perplexity

NY times ×CVB0 with Sym.Dir



Testset perplexity × LOO perplexity



Empirical Bayes

$$(\gamma^*, \beta^*) = \arg\max_{\gamma,\beta} \log p(D \mid \gamma, \beta)$$

Variational EM

$$(\gamma^*, \beta^*) = \arg\max_{\gamma,\beta} L(D \mid \gamma, \beta)$$

$$\log p(D \mid \gamma, \beta) \geq L(D \mid \gamma, \beta)$$

This work

Stochastic Approx.

$$(\gamma^*, \beta^*) = \arg\max_{\gamma,\beta} \log q(D \mid \gamma, \beta)$$

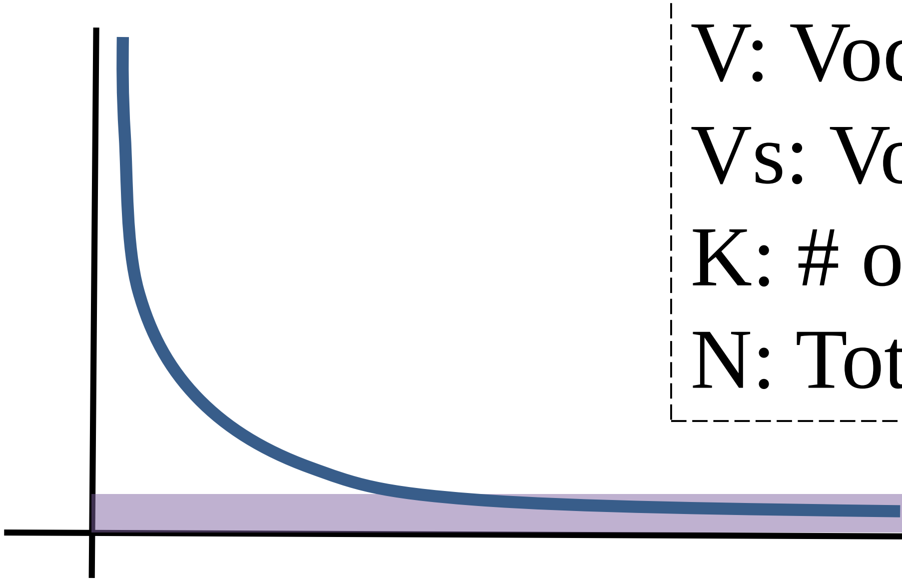
$$\prod_{d,i} q(w_{d,i} \mid \gamma, \beta)$$

$\Leftrightarrow$ min Leave-One-Out Perplexity

Summary

	VB '03	SVB '10	CVB0 '07	SCVB0 '13	This work
Data processing	Batch	Sub-samp.	Batch	Sub-samp.	Sub-samp.
Memory	O(VK)	O(VK)	O(NK)	O(VK)	O(VK)
Update/mini-batch	-	O(VK)	-	O(VK)	O(VsK)
HDP(Asym. Dir)	✔	✔	✔	-	✔

V: Vocabulary size
Vs: Vocab. size in sub-samples
K: # of topics
N: Total # of words



Ignore!

Experimental settings

4 datasets

Dataset	# of Doc.	Vocab. size
DBLP	0.6M	19K
Wikipedia	1M	130K
Pubmed1M	1M	50K
Pubmed5M	5M	122K

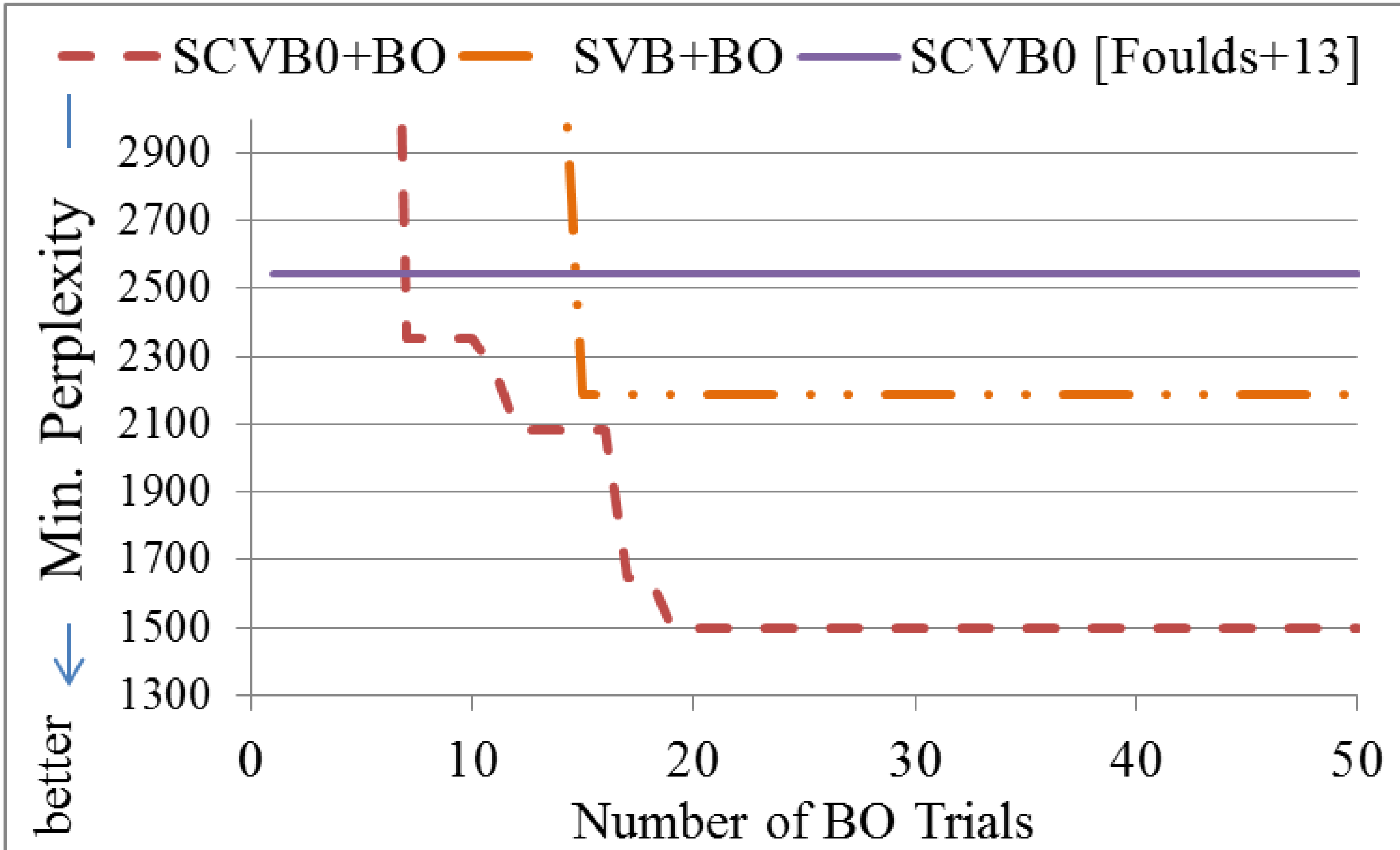
Evaluation: Testset Perplexity

Algorithms: SVB, SCVB0, SDM (This work)

# of Topics (Truncation level): 1000

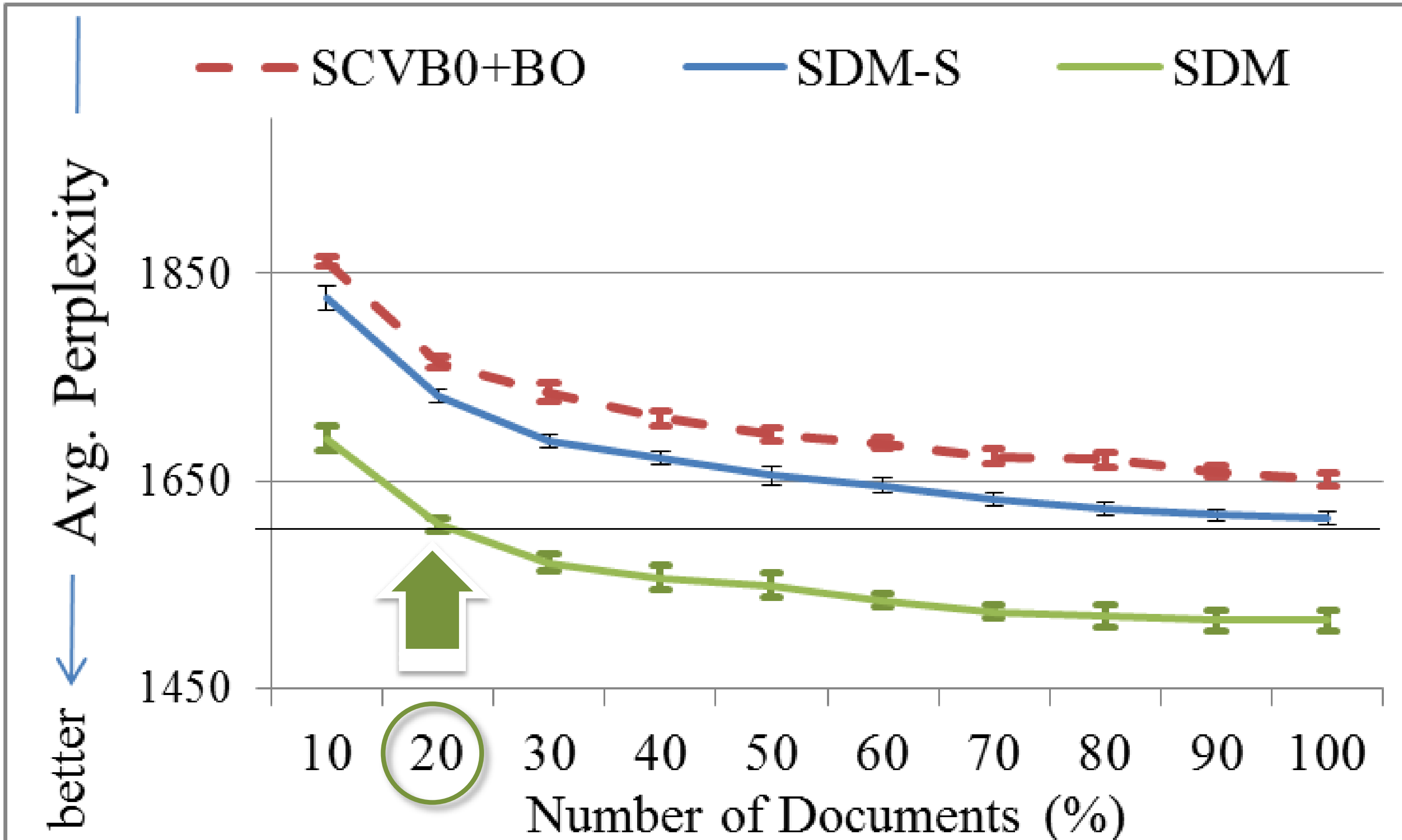
Bayesian optimization for tuning hyper-parameters

Dataset: Pubmed1M



Experimental result

Dataset: Pubmed5M



反省点：査読者との戦いを経て

- スコープが狭い
  - （建前）LDAは引用数1万を超える論文なので、LDAのアルゴリズムの改良自体は重要
  - （本音）汎用性大事。少なくとも汎用性がありそうな書き方を心がけるべき
- 実験が少ない。Twitterの解析とかもやったら？
  - （建前）[Foulds+, KDD2013]と同等の実験
  - （本音）実験の種類はやはり多いほうが良い

  少なくとも[Sato+,KDD2012]のときは、Perplexity(4datasets), リンク予測(2datasets), 文書分類(2datasets)としたら褒められた
- 外部リンク先に証明のある定理は貢献に入れるべきではない
  - （建前）
    事前にCo-Chairsに可能か確認済なので、メタ査読者に確認を
    以前KDDの査読者に外部リンクに置くように言われたことがある
    そのようにしているKDD論文も過去にある
  - （本音）Supplemental materialの無い会議では、やはりページ内に収めるように書くべき