Why doesn't EM find good HMM POS-taggers?

Mark Johnson

Microsoft Research   Brown University

Bayesian inference for HMMs

- Compare Bayesian methods for estimating HMMs for *unsupervised POS tagging*
  - Gibbs sampling
  - Variational Bayes
  - How do these compare to EM?
- Most words belong to few POS: can a sparse Bayesian prior on P($w|y$) capture this?
- KISS – look at bitag HMM models first
- Cf: Goldwater and Griffiths 2007 study semi-supervised Bayesian inference for tritag HMM POS taggers

Main findings

- Bayesian inference finds better POS tags
- By reducing the number of states, EM can do almost as well
- All these methods take hundreds of iterations to stabilize (converge?)
- Wide variation in performance of all models $\Rightarrow$ multiple runs to assess performance

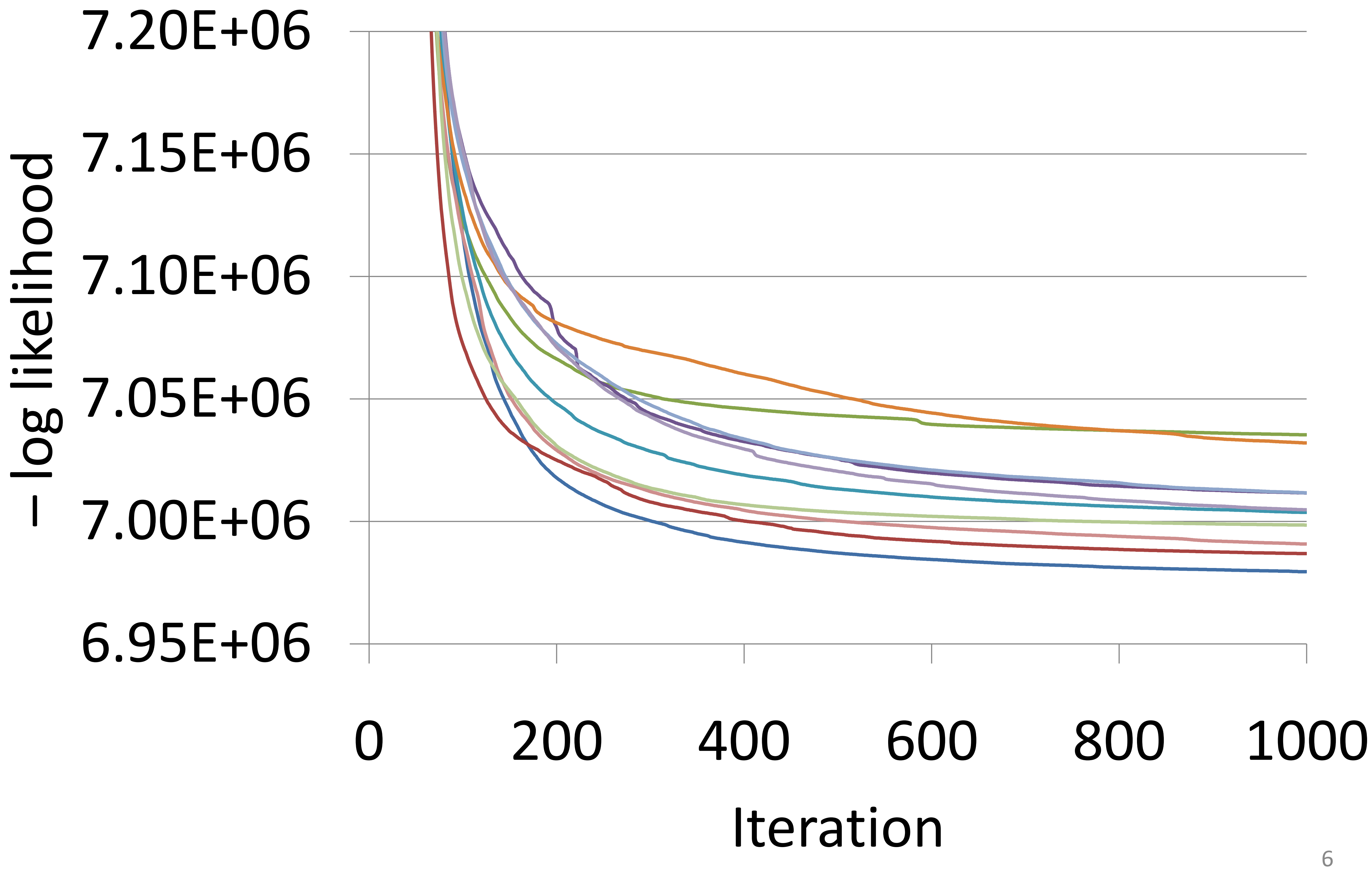
Evaluation methodology

- "Many-to-1" accuracy:
  - Each HMM hidden state *y* is mapped to the most frequent gold POS tag *t* it corresponds to
- "1-to-1" accuracy: (Haghighi and Klein 06)
  - Greedily map HMM states to POS tags, under constraint that at most 1 state maps to each tag
- Information-theoretic measures: (Meila 03)
  - $VI(Y,T) = H(Y|T) + H(T|Y)$
- Max marginal decoding faster and usually better than Viterbi

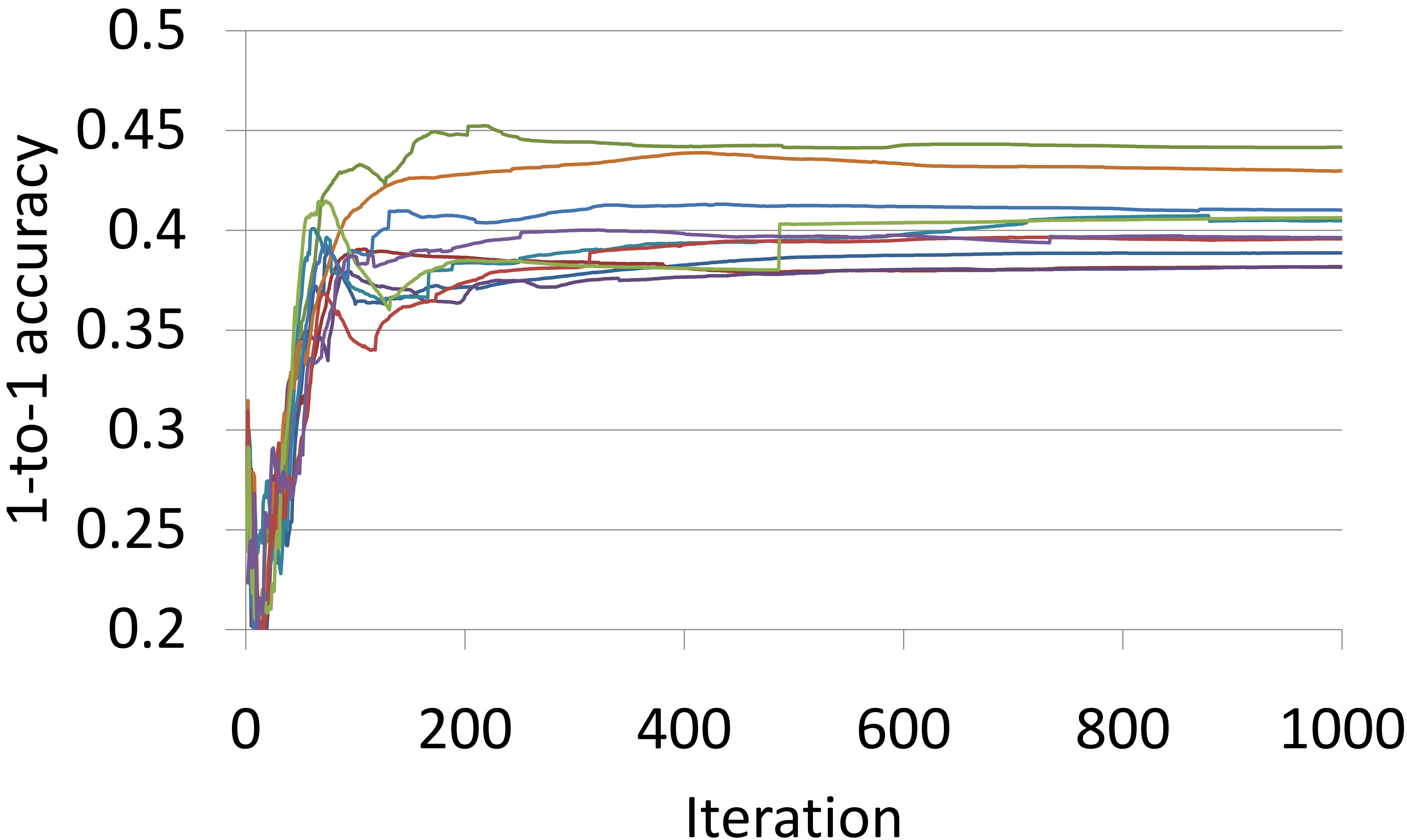
EM via Forward-Backward

- Hmm model: $y_i \mid y_{i-1} \sim \text{Discrete}(\boldsymbol{\theta}_{y_{i-1}})$
  $x_i \mid y_i \sim \text{Discrete}(\boldsymbol{\phi}_{y_i})$
- EM iterations: $\theta_{y'|y}^{(l+1)} = E[n_{y',y}]/E[n_y]$
  $\phi_{x|y}^{(l+1)} = E[n_{x,y}]/E[n_y]$
- All expts run on POS tags from WSJ PTB

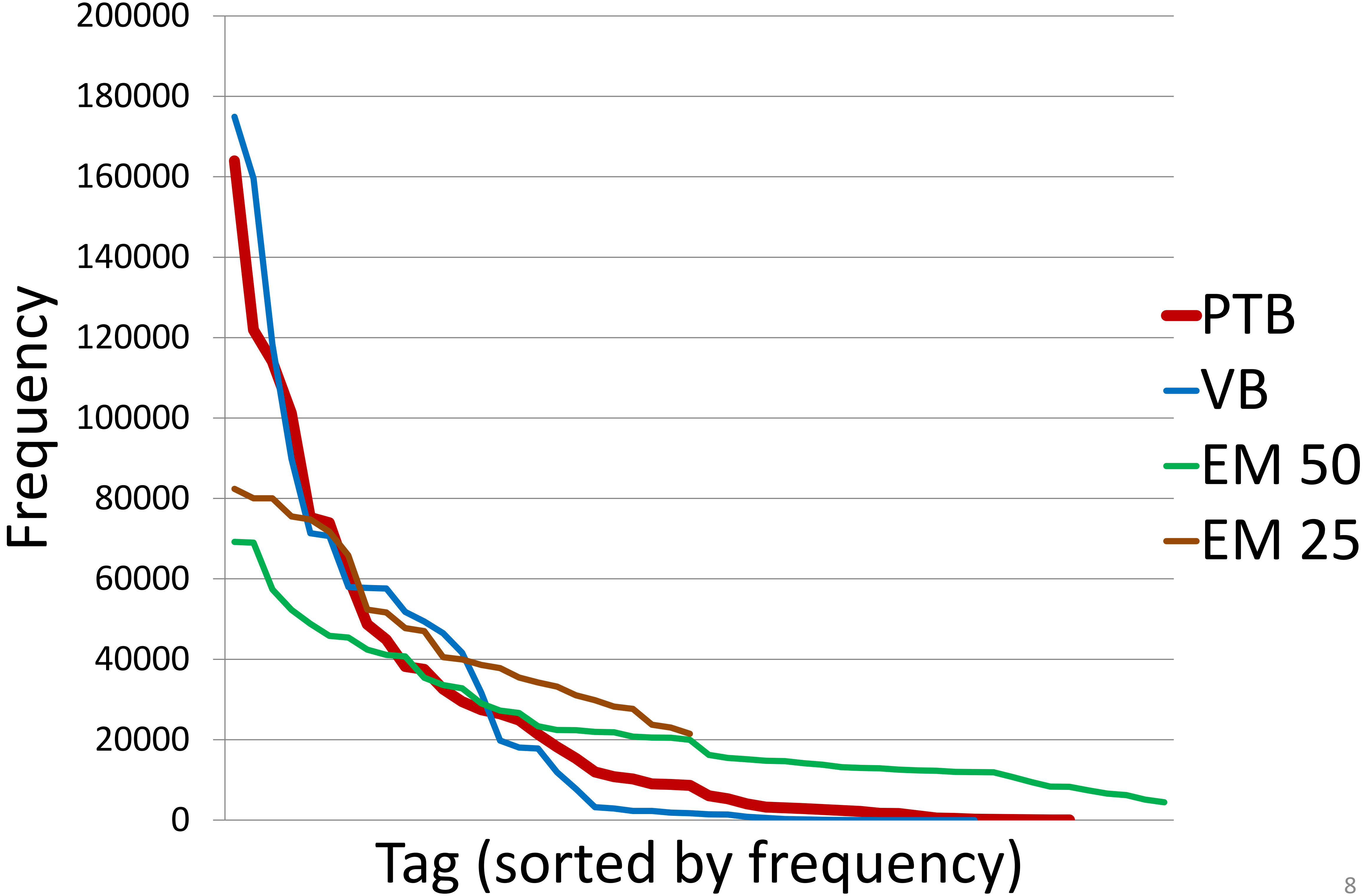
EM is slow to stabilize



EM 1-to-1 accuracy varies widely



EM tag dist less peaked than empirical



Bayesian estimation of HMMs

- HMM with Dirichlet priors on tag→tag and tag→word distributions

$$y_i \mid y_{i-1} \sim \mathrm{Discrete}(\boldsymbol{\theta}_{y_{i-1}})$$
$$x_i \mid y_i \sim \mathrm{Discrete}(\boldsymbol{\phi}_{y_i})$$
$$\boldsymbol{\theta}_y \mid \alpha \sim \mathrm{Dir}(\alpha)$$
$$\boldsymbol{\phi}_y \mid \beta \sim \mathrm{Dir}(\beta)$$

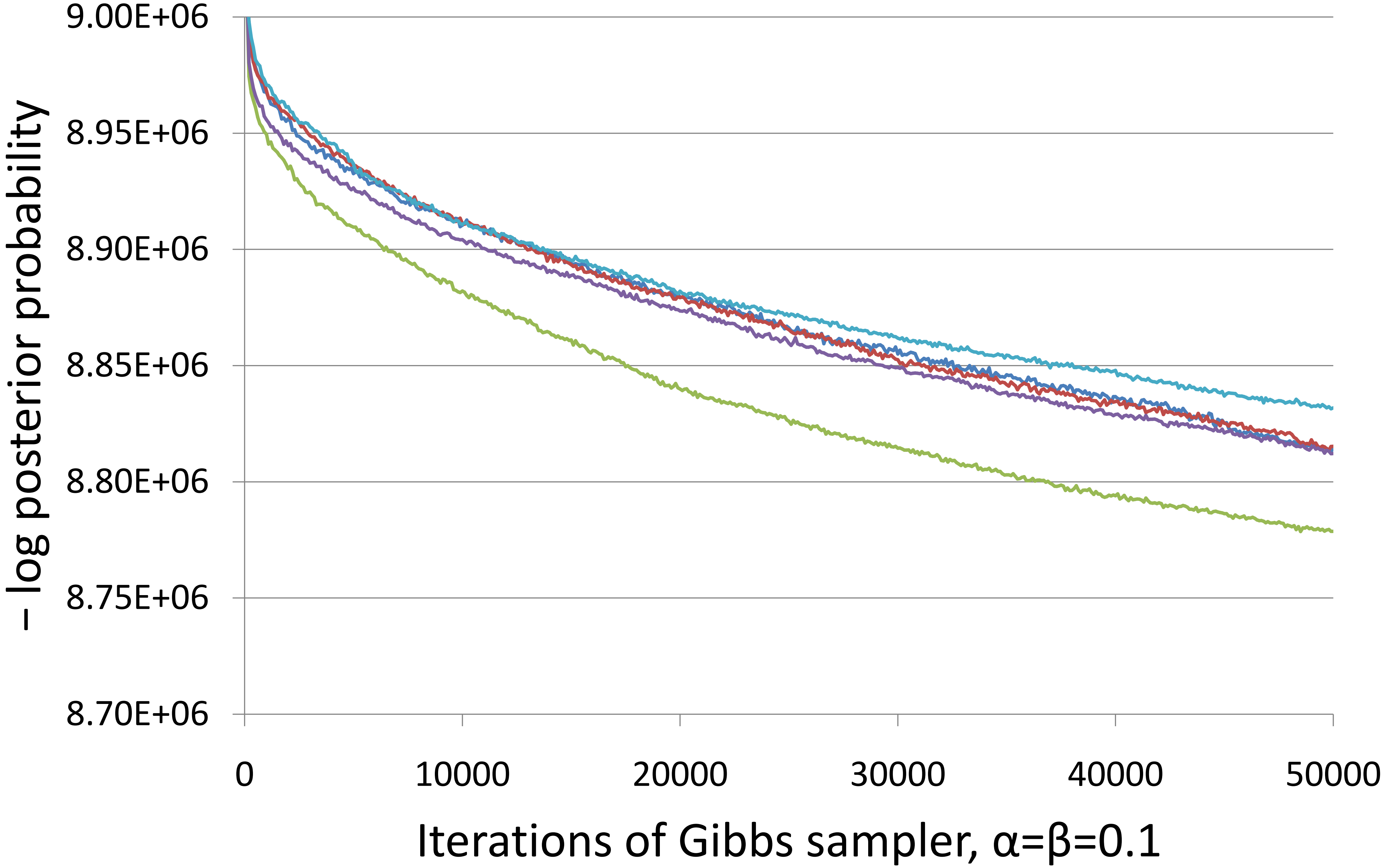
- As Dirichlet parameter approaches zero, prior prefers sparse (more peaked) distributions

Gibbs sampling

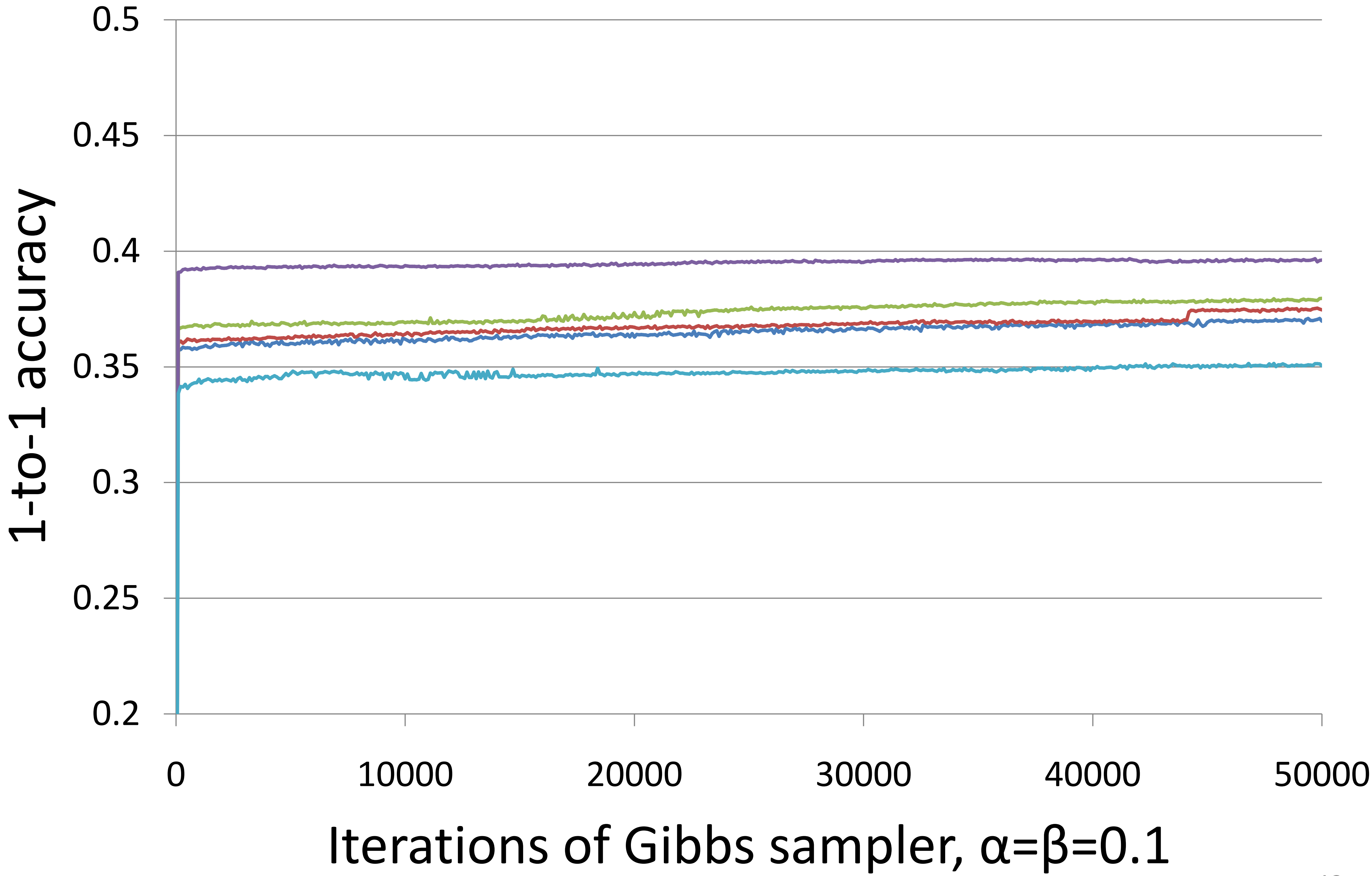
- A Gibbs sampler is a MCMC procedure for sampling from posterior dist P(***y***|***x***,α,β)
- Integrate out the θ, ϕ parameters
- Repeatedly sample from P(y_i|$\boldsymbol{y}_{-i}$,α,β), where $\boldsymbol{y}_{-i}$ is the vector of all ***y*** *except* y_i

$$\mathrm{P}(y_i|\boldsymbol{y}_{-i},\alpha,\beta) \propto \frac{n_{x_i,y_i}+\beta}{n_{y_i}+m\beta}\;\frac{n_{y_i,y_{i-1}}+\beta}{n_{y_{i-1}}+s\beta}\;\frac{n_{y_{i+1},y_i}+\mathrm{I}(y_{i-1}=y_i)+\beta}{n_{y_i}+\mathrm{I}(y_{i-1}=y_i)+s\beta}$$

Gibbs sampling is even slower

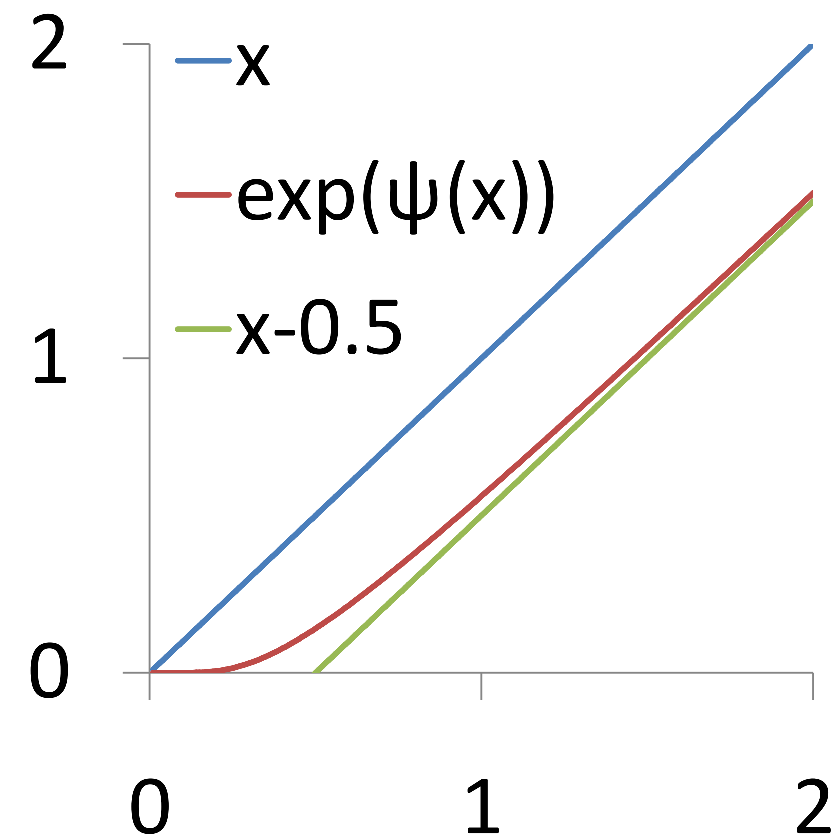


Gibbs stabilizes fast (to poor solns)



Variational Bayes

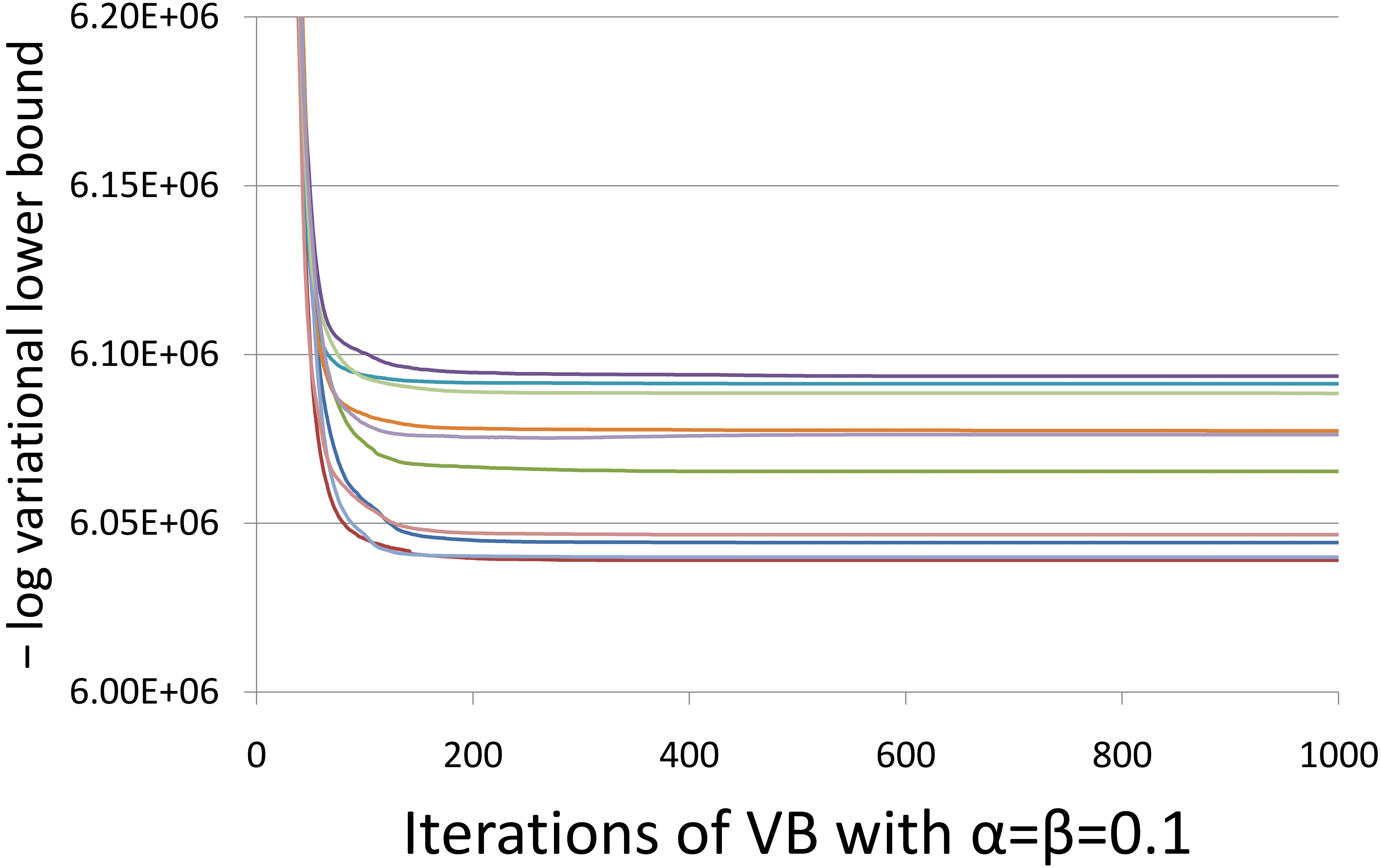
- Variational Bayes approximates the posterior P(***y***,**θ**,**ϕ**|***x***,α,β) ≈ Q(***y***) Q(**θ**,**ϕ**) (MacKay 97, Beal 03)
- Simple, EM-like procedure:



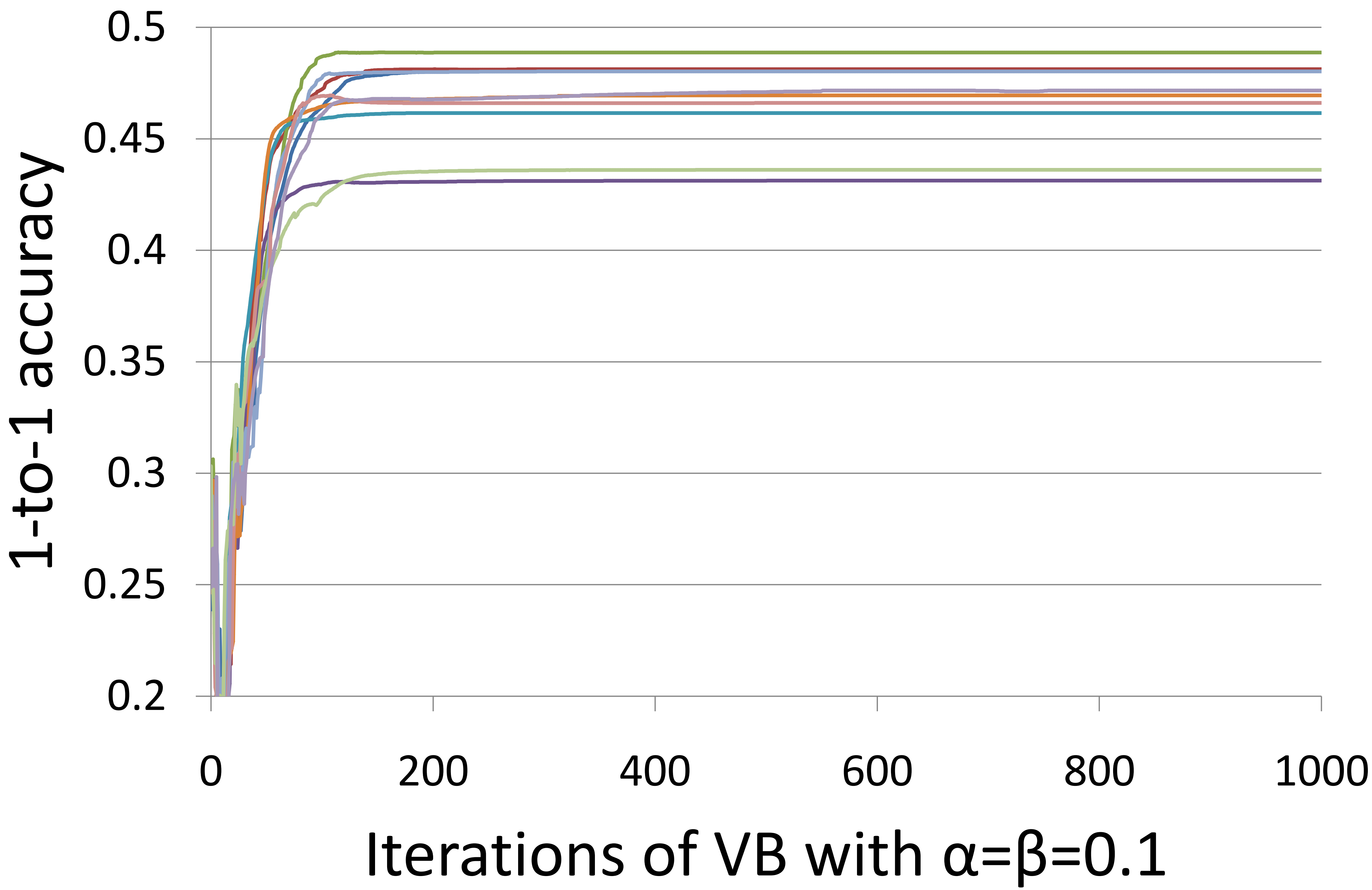
$$\tilde{\theta}^{(l+1)}_{y'|y} = \exp\Psi(E[n_{y',y}]) \,/\, \exp\Psi(E[n_y])$$

$$\tilde{\phi}^{(l+1)}_{x|y} = \exp\Psi(E[n_{x,y}]) \,/\, \exp\Psi(E[n_y])$$

VB posterior seems to stabilize fast



VB 1-to-1 accuracy stabilizes fast



Summary of results

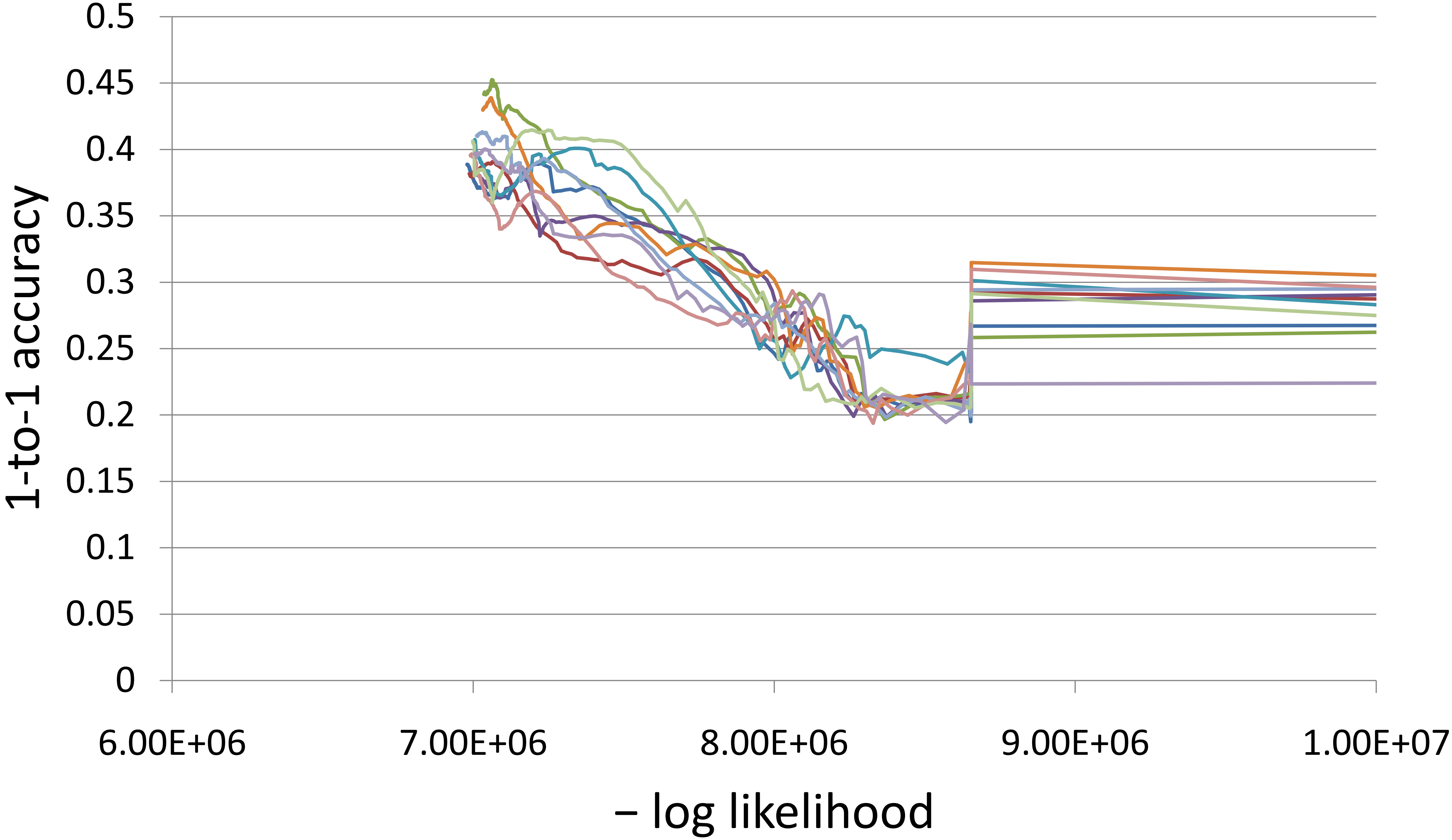
	α	β	*states*	*1-to-1*	*S.D.*	*many-to-1*	*S.D.*	*VI(T,Y)*	*S.D.*	*H(T|Y)*	*S.D.*	*H(Y|T)*	*S.D.*
EM			50	0.40	0.02	**0.62**	0.01	4.46	0.08	**1.75**	0.04	2.71	0.06
VB	0.1	0.1	50	**0.47**	0.02	0.50	0.02	4.28	0.09	2.39	0.07	1.89	0.06
VB	1E-04	1	50	0.46	0.03	0.50	0.02	4.28	0.11	2.39	0.08	1.90	0.07
VB	0.1	1E-04	50	0.42	0.02	0.60	0.01	4.63	0.07	1.86	0.03	2.77	0.05
VB	1E-04	1E-04	50	0.42	0.02	0.60	0.01	4.62	0.07	1.85	0.03	2.76	0.06
GS	0.1	0.1	50	0.37	0.02	0.51	0.01	5.45	0.07	2.35	0.09	3.20	0.03
GS	1E-04	0.1	50	0.38	0.01	0.51	0.01	5.47	0.04	2.26	0.03	3.22	0.01
GS	0.1	1E-04	50	0.36	0.02	0.49	0.01	5.73	0.05	2.41	0.04	3.31	0.03
GS	1E-04	1E-04	50	0.37	0.02	0.49	0.01	5.74	0.03	2.42	0.02	3.32	0.02
EM			40	0.42	0.03	0.60	0.02	4.37	0.14	1.84	0.07	2.55	0.08
EM			25	0.46	0.03	0.56	0.02	**4.23**	0.17	2.05	0.09	2.19	0.08
EM			10	0.41	0.01	0.43	0.01	4.32	0.04	2.74	0.03	**1.58**	0.05

- Griffiths and Goldwater 2007 report VI = 3.74 for an unsupervised tritag model using Gibbs sampling, but on a reduced 17-tag set

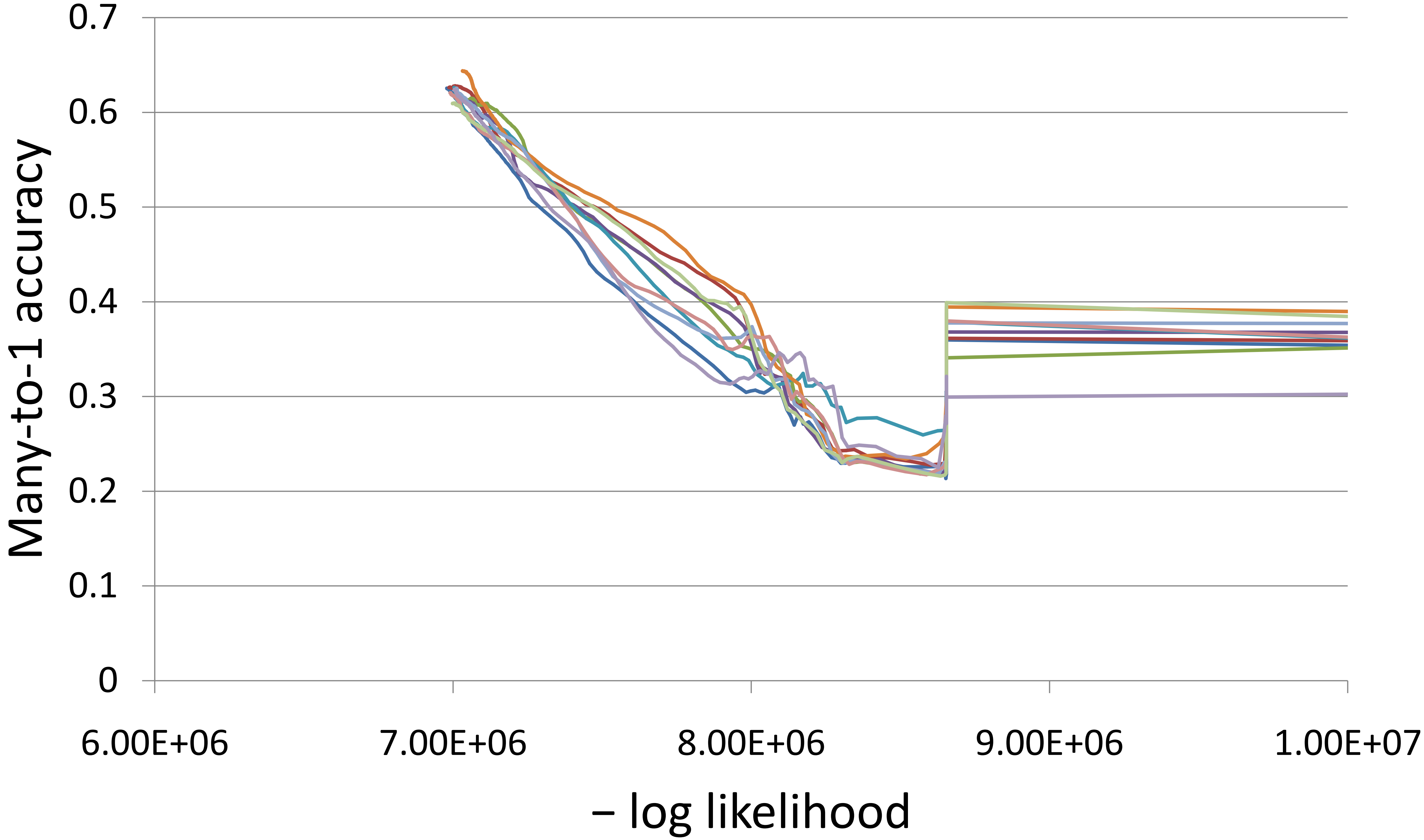
Conclusions

- EM does better if you let it run longer
- Its state distribution is not skewed enough
  - Bayesian priors
  - Reduce the number of states in EM
- Variational Bayes may be faster than Gibbs (or maybe initialization?)
- *Huge performance variance with all estimators* $\Rightarrow$ need multiple runs to assess performance

EM 1-to-1 accuracy vs likelihood



EM many-to-1 accuracy vs likelihood



EM final many-to-1 accuracy vs final likelihood