Regularized Bayesian Inference

when
Bayes meets Optimization and Learning

Jun Zhu

dcszj@mail.tsinghua.edu.cn

Department of Computer Science and Technology

Tsinghua University

Fudan University, Nov 3, 2013

Outline

- RegBayes: Regularized Bayesian inference
- Example of max-margin supervised topic modeling

Bayesian Inference

- A coherent framework of dealing with uncertainties

$$p(\mathcal{M}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{M})\pi(\mathcal{M})}{\int p(\mathbf{x}|\mathcal{M})\pi(\mathcal{M})d\mathcal{M}}$$

- *M*: a model from some hypothesis space
- x: observed data

Thomas Bayes (1702 – 1761)

- Bayes' rule offers a mathematically rigorous computational mechanism for combining prior knowledge with incoming evidence

Why Be Bayesian?

- One of many answers
- Infinite Exchangeability:

$$\forall n,\ \forall \sigma,\ p(x_1, \ldots, x_n) = p(x_{\sigma(1)}, \ldots, x_{\sigma(n)})$$

- De Finetti's Theorem (1955): if $(x_1, x_2, \ldots)$ are *infinitely exchangeable*, then $\forall n$

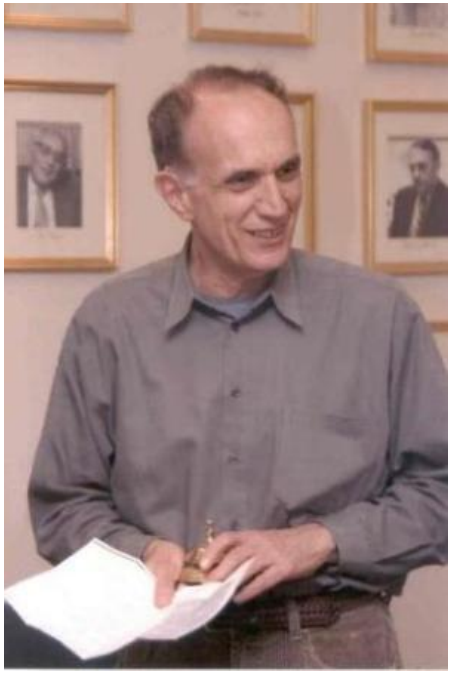
$$p(x_1, \ldots, x_n) = \int \left(\prod_{i=1}^{n} p(x_i|\theta) \right) dP(\theta)$$

for some random variable θ

Bayes' Theorem in the 21st Century

- This year marks the 250th Anniversary of Bayes' theorem

- Bradley Efron, *Science 7 June 2013: Vol. 340 no. 6137 pp. 1177-1178*



"There are two potent arrows in the statistician's quiver

there is no need to go hunting armed with only one."

Parametric Bayesian Inference

$\mathcal{M}$ is represented as a finite set of parameters θ

- A parametric likelihood: $\mathbf{x} \sim p(\cdot|\theta)$
- Prior on $\boldsymbol{\theta}$: $\pi(\theta)$
- Posterior distribution

$$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)\pi(\theta)}{\int p(\mathbf{x}|\theta)\pi(\theta)d\theta} \propto p(\mathbf{x}|\theta)\pi(\theta)$$

Examples:

- Gaussian distribution prior + 2D Gaussian likelihood → Gaussian posterior distribution
- Dirichilet distribution prior + 2D Multinomial likelihood → Dirichlet posterior distribution
- Sparsity-inducing priors + some likelihood models → Sparse Bayesian inference

Nonparametric Bayesian Inference

$\mathcal{M}$ is a richer model, e.g., with an infinite set of parameters

- A nonparametric likelihood: $\mathbf{x} \sim p(\cdot|\mathcal{M})$
- Prior on $\mathcal{M}$: $\pi(\mathcal{M})$
- Posterior distribution

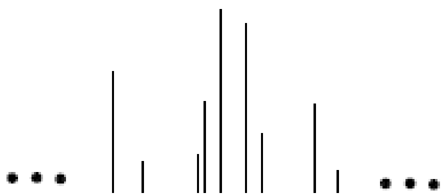
$$p(\mathcal{M}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{M})\pi(\mathcal{M})}{\int p(\mathbf{x}|\mathcal{M})\pi(\mathcal{M})d\mathcal{M}} \propto p(\mathbf{x}|\mathcal{M})\pi(\mathcal{M})$$

Examples:

→ see next slide

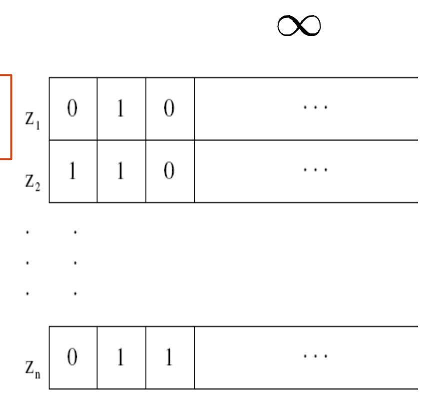
Nonparametric Bayesian Inference

probability measure



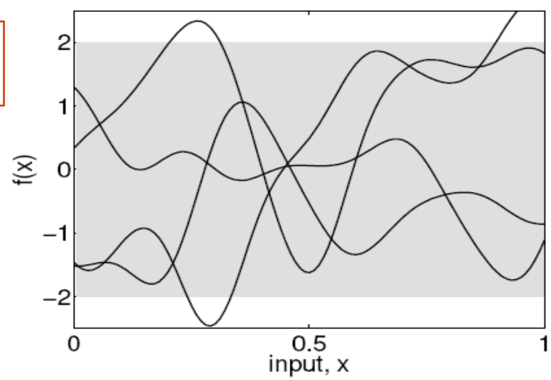
Dirichlet Process Prior [Ferguson, 1973]
+ Multinomial/Gaussian/Softmax likelihood

binary matrix



Indian Buffet Process Prior [Griffiths & Gharamani, 2005]
+ Gaussian/Sigmoid/Softmax likelihood

function

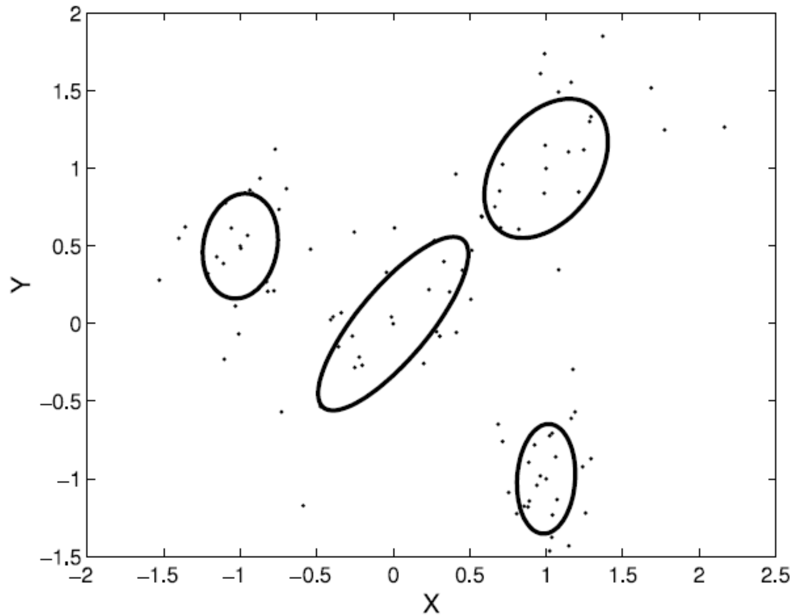


Gaussian Process Prior [Doob, 1944; Rasmussen & Williams, 2006]
+ Gaussian/Sigmoid/Softmax likelihood

Why Bayesian Nonparametrics?

Let the data speak for itself

- Bypass the model selection problem
  - let data determine model complexity (e.g., the number of components in mixture models)
  - allow model complexity to grow as more data observed



Bayesian Inference with Rich Priors

- ***The world is structured and dynamic!***
- Predictor-dependent processes to handle heterogeneous data
  - Dependent Dirichlet Process (MacEachern, 1999)
  - Dependent Indian Buffet Process (Williamson et al., 2010)
  - …
- Correlation structures to relax exchangeability:
  - Processes with hierarchical structures (Teh et al., 2007)
  - Processes with temporal or spatial dependencies (Beal et al., 2002; Blei & Frazier, 2010)
  - Processes with stochastic ordering dependencies (Hoff et al., 2003; Dunson & Peddada, 2007)
  - …

Challenges of Bayesian Inference

Building an *Automated* Statistician

- Modeling
  - scientific and engineering data
- Inference/learning
  - discriminative learning
  - large-scale inference algorithms for Big Data
- Applications

Regularized Bayesian Inference?

posterior likelihood model prior

$$p(\mathcal{M}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{M})\pi(\mathcal{M})}{\int p(\mathbf{x}|\mathcal{M})\pi(\mathcal{M})d\mathcal{M}}$$

- **Can we directly control the posterior distributions?**
  - An extra freedom to perform Bayesian inference
  - Arguably more direct to control the behavior of models
  - Can be easier and more natural in some examples

Regularized Bayesian Inference?

posterior likelihood model prior

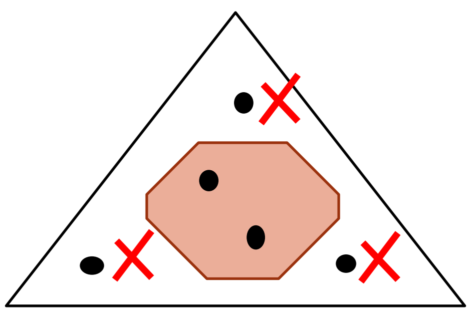
$$p(\mathcal{M}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{M})\pi(\mathcal{M})}{\int p(\mathbf{x}|\mathcal{M})\pi(\mathcal{M})d\mathcal{M}}$$

- **Can we directly control the posterior distributions?**

Not obvious!

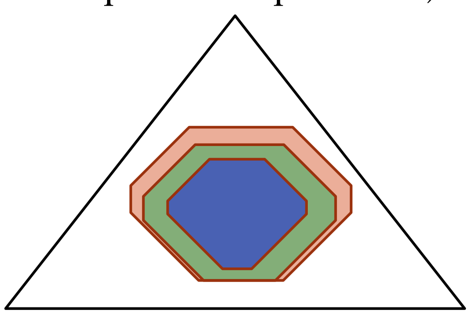
hard constraints

(A single feasible space)



soft constraints

(many feasible subspaces with different complexities/penalties)



Bayesian Inference as an Opt. Problem

Wisdom never forgets that all things have two sides

$$p(\mathcal{M}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{M})\pi(\mathcal{M})}{\int p(\mathbf{x}|\mathcal{M})\pi(\mathcal{M})d\mathcal{M}}$$

- **Bayes' rule is equivalent to solving:**

$$\begin{aligned} \min_{q(\mathcal{M})} \quad & \mathrm{KL}(q(\mathcal{M})\|\pi(\mathcal{M})) - \mathbb{E}_{q(\mathcal{M})}[\log p(\mathbf{x}|\mathcal{M})] \\ \mathrm{s.t.}: \quad & q(\mathcal{M}) \in \mathcal{P}_{\mathrm{prob}}, \end{aligned}$$

direct but trivial constraints on posterior distribution

[Zhu et al., NIPS 2011; Zhu et al., arXiv:1210.1766v2, 2013]

Regularized Bayesian Inference

Constraints can encode rich structures/knowledge

◆ **Bayesian inference with posterior regularization:**

$$\begin{aligned} \min_{q(\mathcal{M}),\xi} \quad & \mathrm{KL}(q(\mathcal{M})\|\pi(\mathcal{M})) - \mathbb{E}_{q(\mathcal{M})}[\log p(\mathbf{x}|\mathcal{M})] + U(\xi) \\ \mathrm{s.t.}: \quad & q(\mathcal{M}) \in \mathcal{P}_{\mathrm{post}}(\xi), \end{aligned}$$

convex function

direct and rich constraints on posterior distribution

- Consider both hard and soft constraints
- Convex optimization problem with nice properties
- Can be effectively solved with convex duality theory

[Zhu et al., NIPS 2011; Zhu et al., arXiv:1210.1766v2, 2013]

Regularized Bayesian Inference

Constraints can encode rich structures/knowledge

- **Bayesian inference with posterior regularization:**

'unconstrained' equivalence:

$$\min_{q(\mathcal{M})} \quad \mathrm{KL}(q(\mathcal{M})\|\pi(\mathcal{M})) - \mathbb{E}_{q(\mathcal{M})}[\log p(\mathbf{x}|\mathcal{M})] + \Omega(q(\mathcal{M}))$$
$$\mathrm{s.t.}: \quad q(\mathcal{M}) \in \mathcal{P}_{\mathrm{prob}},$$

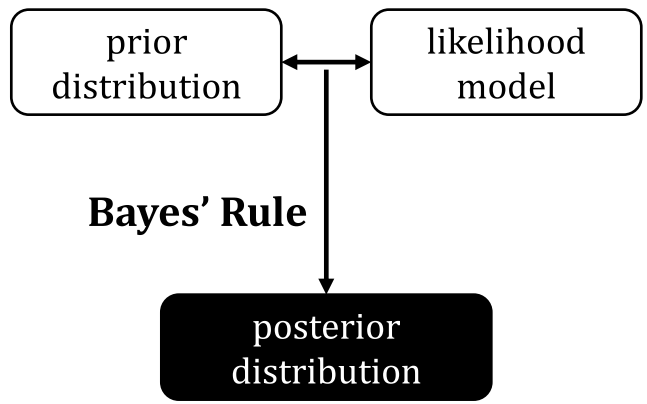
posterior regularization

- Consider both hard and soft constraints
- Convex optimization problem with nice properties
- Can be effectively solved with convex duality theory

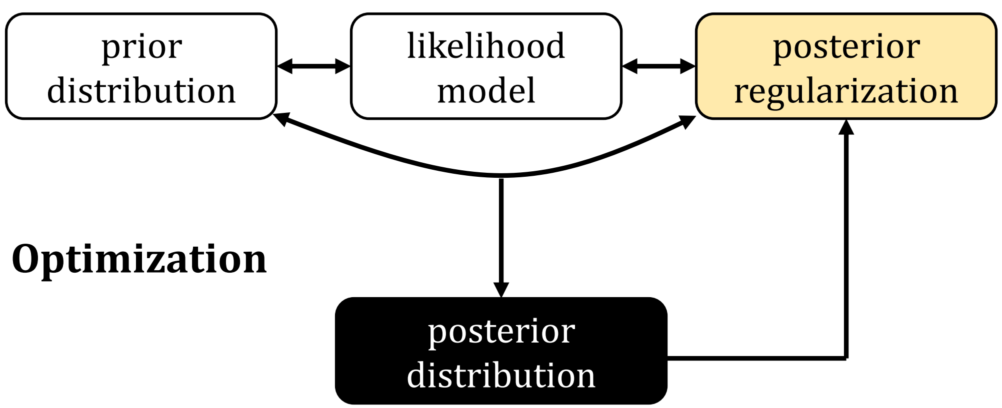
[Zhu et al., NIPS 2011; Zhu et al., arXiv:1210.1766v2, 2013]

A High-Level Comparison

Bayes:



RegBayes:



Ways to Derive Posterior Regularization

- **From learning objectives**
  - Performance of posterior distribution can be evaluated when applying it to a learning task
  - Learning objective can be formulated as Pos. Reg.

- **From domain knowledge** (ongoing & future work)
  - Elicit expert knowledge
  - E.g., logic rules

- **Others …** (ongoing & future work)
  - E.g., decision making, cognitive constraints, etc.

PAC-Bayes Theory

- Basic Setup:
  - Binary classification: $\mathbf{x} \in \mathbb{R}^d \quad y \in \mathcal{Y} = \{-1, +1\}$
  - Unknown, true data distribution: $(\mathbf{x}, y) \sim D$
  - Hypothesis space: $\mathcal{H}$
  - Risk, & Empirical Risk:

$$R(h) = \mathbb{E}_{(\mathbf{x},y)\sim D} I(h(\mathbf{x}) \neq y) \quad R_S(h) = \frac{1}{N}\sum_{n=1}^{N} I(h(\mathbf{x}_i) \neq y_i)$$

- Learn a posterior distribution Q
- Bayes/majority-vote classifier:

$$B_Q(\mathbf{x}) = \mathrm{sgn}\left[\mathbb{E}_{h\sim Q} h(\mathbf{x})\right]$$

- Gibbs classifier
  - sample an $h \sim Q$, perform prediction

$$R(G_Q) = \mathbb{E}_{h\sim Q} R(h) \quad R_S(G_Q) = \mathbb{E}_{h\sim Q} R_S(h)$$

PAC-Bayes Theory

- Theorem (Germain et al., 2009):
  - for any distribution D ; for any set $\mathcal{H}$ of classifiers, for any prior P , for any convex function

$$\phi : \ [0,1] \times [0,1] \to \mathbb{R}$$

  - for any posterior Q , for any $\delta \in (0,1]$, the following inequality holds with a high probability ($\geq 1-\delta$)

$$\phi\left(R_S(G_Q), R(G_Q)\right) \leq \frac{1}{N}\left[\mathrm{KL}(Q\|P) + \ln\left(\frac{C(N)}{\delta}\right)\right]$$

  - where $C(N) = \mathbb{E}_{S\sim D^N}\mathbb{E}_{h\sim P}\left[e^{N\phi(R_S(h),R(h))}\right]$

RegBayes Classifiers

- PAC-Bayes theory

$$\phi\left(R_S(G_Q), R(G_Q)\right) \leq \frac{1}{N}\left[\mathrm{KL}(Q\|P) + \ln\left(\frac{C(N)}{\delta}\right)\right]$$

- RegBayes inference

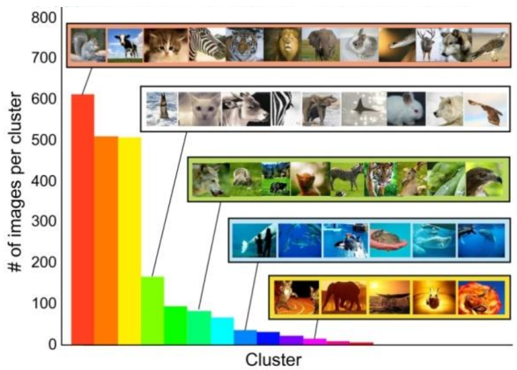
$$\begin{aligned} \min_{q(\mathcal{H})} \quad & \mathrm{KL}(q(\mathcal{H})\|p(\mathcal{H}|\mathbf{x})) + \Omega(q(\mathcal{H})) \\ \text{s.t.}: \quad & q(\mathcal{H}) \in \mathcal{P}_{\text{prob}}, \end{aligned}$$

- Observations:
  - when the posterior regularization equals to (or upper bounds) the empirical risk

$$\Omega(q(\mathcal{H})) \geq R_S(G_q)$$

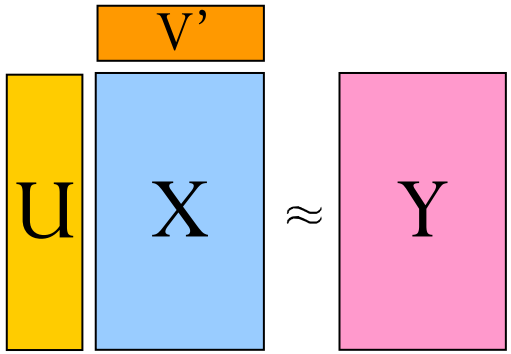
  - the RegBayes classifiers tend to have PAC-Bayes guarantees.

RegBayes with Max-margin Posterior Regularization

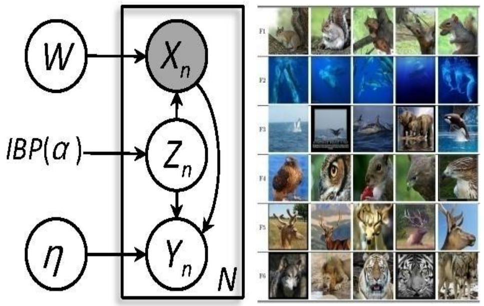


Infinite SVMs
(Zhu, Chen & Xing, ICML'11)

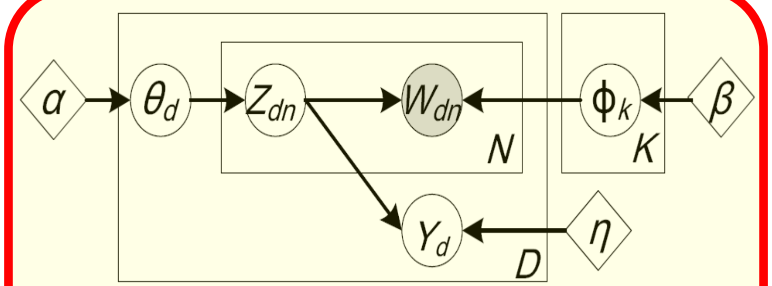
Nonparametric Max-margin Relational Models for Social Link Prediction
(Zhu, ICML'12)



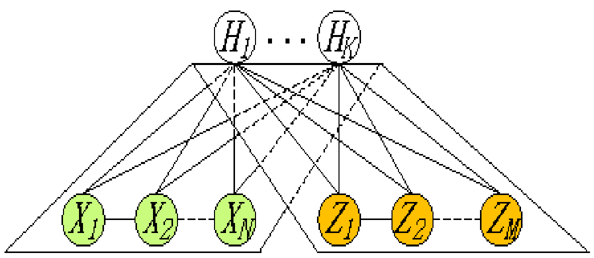
Nonparametric Max-margin Matrix Factorization
(Xu, Zhu, & Zhang, NIPS'12;
Xu, Zhu, & Zhang, ICML'13)



Infinite Latent SVMs
(Zhu, Chen & Xing, NIPS'11;
Zhu, Chen, & Xing, arXiv 2013)



Max-margin Topics and Fast Inference
(Zhu, Ahmed & Xing, ICML'09, JMLR'12;
Jiang, Zhu, Sun & Xing, NIPS'12;
Zhu, Chen, Perkins & Zhang, ICML'13)

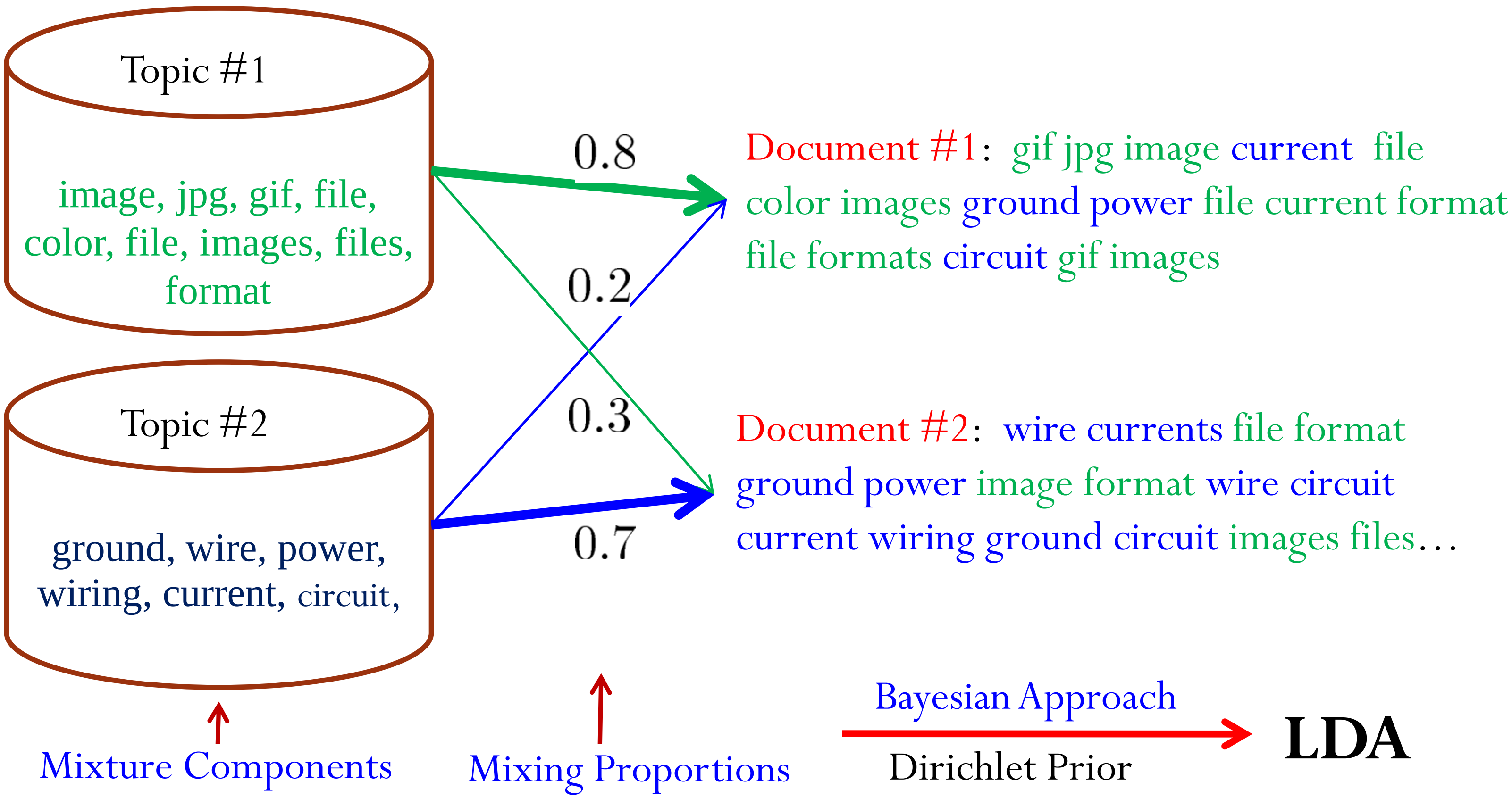


Multimodal Representation Learning
(Chen, Zhu & Xing, NIPS'10,
Chen, Zhu, Sun & Xing, PAMI'12)

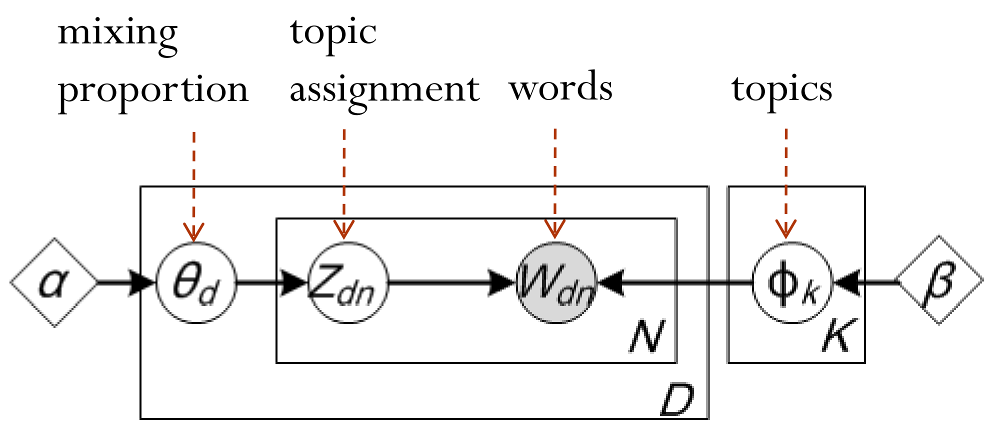
*** Works from other groups are not included.**

Latent Dirichlet Allocation
-- a generative story for documents

- A Bayesian mixture model with topical bases
- Each document is a random mixture over topics; Each word is generated by ONE topic



Bayesian Inference for LDA



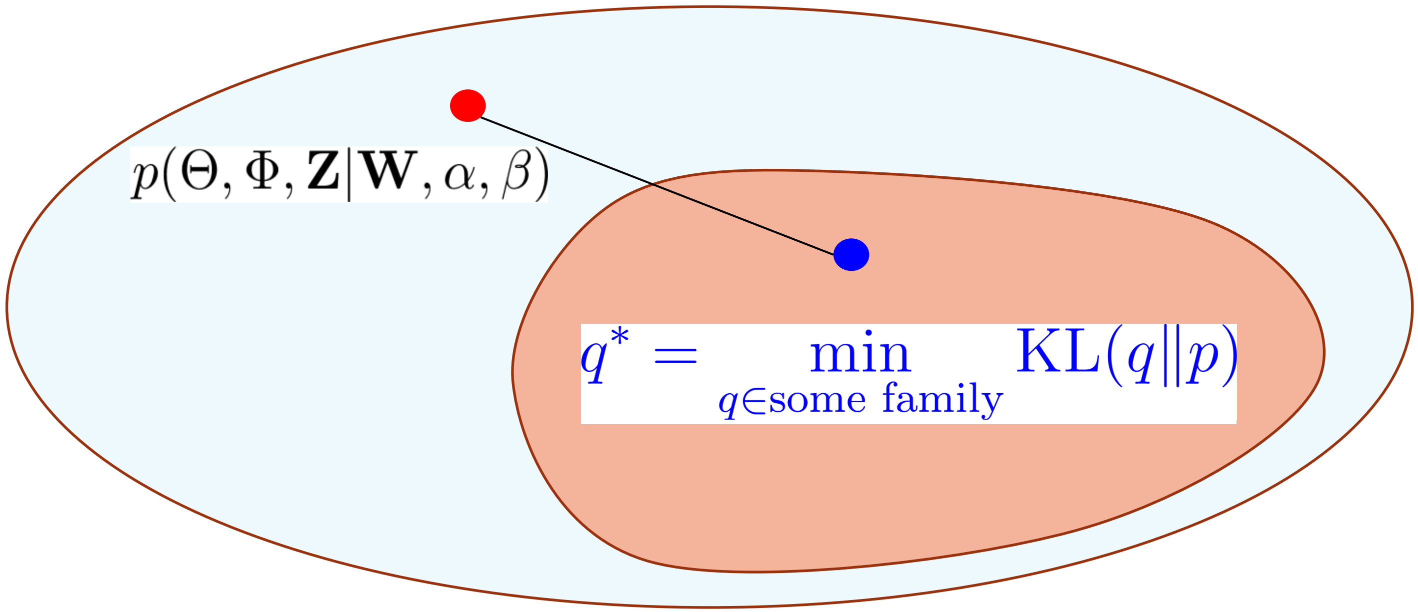
$$p(\Theta, \Phi, \mathbf{Z}, \mathbf{W} | \alpha, \beta) = \prod_{k=1}^{K} p(\Phi_k | \beta) \prod_{d=1}^{D} p(\theta_d | \alpha) \left(\prod_{n=1}^{N} p(z_{dn} | \theta_d) p(w_{dn} | z_{dn}, \Phi) \right)$$

- Given a set of documents, infer the posterior distribution

$$p(\Theta, \Phi, \mathbf{Z} | \mathbf{W}, \alpha, \beta) = \frac{p(\Theta, \Phi, \mathbf{Z}, \mathbf{W} | \alpha, \beta)}{p(\mathbf{W} | \alpha, \beta)}$$

Approximate Inference

- Variational Inference (Blei et al., 2003; Teh et al., 2006)



- Monte Carlo Markov Chains (Griffiths & Steyvers, 2004)
  - Collapsed Gibbs samplers iteratively draw samples from the local conditionals

$$p(z_{dn}^k = 1 | Z_{\neg})$$

Optimization Problem for LDA

- Bayes' rule

$$p(\Theta, \Phi, \mathbf{Z}|\mathbf{W}, \alpha, \beta) = \frac{p(\Theta, \Phi, \mathbf{Z}|\alpha, \beta)p(\mathbf{W}|\mathbf{Z}, \Phi)}{p(\mathbf{W}|\alpha, \beta)}$$

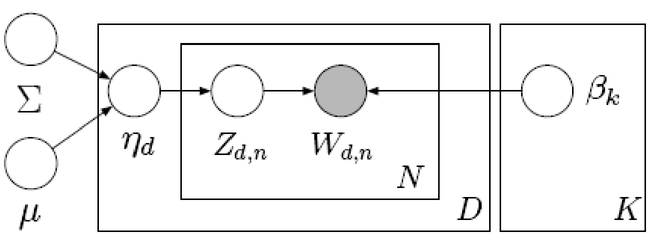
- Optimization problem

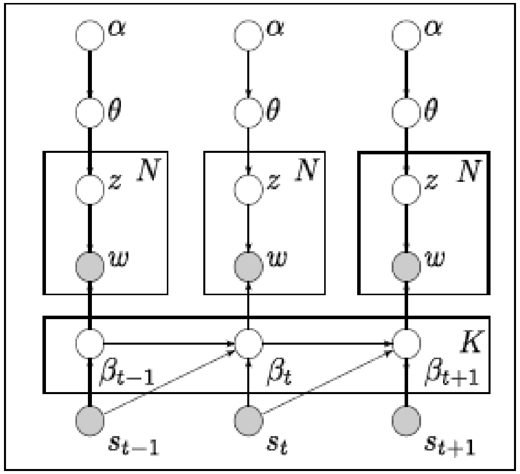
$$\min_{q(\Theta, \Phi, \mathbf{Z})} \ \mathrm{KL}(q(\Theta, \Phi, \mathbf{Z}) \| p(\Theta, \Phi, \mathbf{Z}|\alpha, \beta)) - \mathbb{E}_q[\log p(\mathbf{W}|\mathbf{Z}, \Phi)]$$
$$\text{s.t: } q(\Theta, \Phi, \mathbf{Z}) \in \mathcal{P}$$

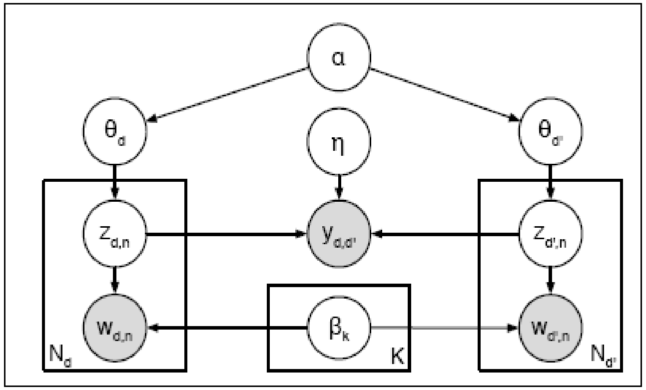
  - Assume *q* is in the factorized family and solve this problem with co-ordinate descent
    - ➔ variational mean-field algorithm (Blei et al., 2003)
  - Solve this problem, collapse Dirichlet variables and do Gibbs sampling
    - ➔ collapsed Gibbs sampling (Griffiths & Steyvers, 2004)

LDA has been widely extended …

- LDA can **be embedded in more complicated models**, capturing rich structures of the texts
- Extensions are either on
  - **Priors**: e.g., Markov process prior for dynamic topic models, logistic-normal prior for corrected topic models, etc
  - **Likelihood models**: e.g., relational topic models, multi-view topic models, etc.



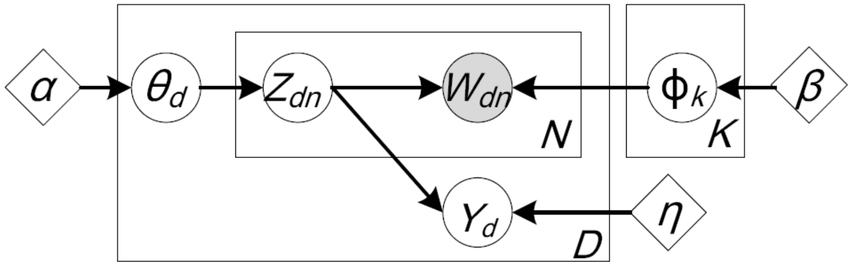




- Tutorials were provide by D. Blei at ICML, SIGKDD, etc. (http://www.cs.princeton.edu/~blei/topicmodeling.html)

Supervised LDA with Rich Likelihood

- Following the standard Bayes' way of thinking, sLDA defines a richer likelihood model



$$p(\mathbf{y}, \mathbf{W} | \mathbf{Z}, \Phi, \eta, \alpha, \beta) = p(\mathbf{y} | \mathbf{Z}, \eta) p(\mathbf{W} | \mathbf{Z}, \Phi, \alpha, \beta)$$

- per-document likelihood $y_d \in \{0, 1\}$

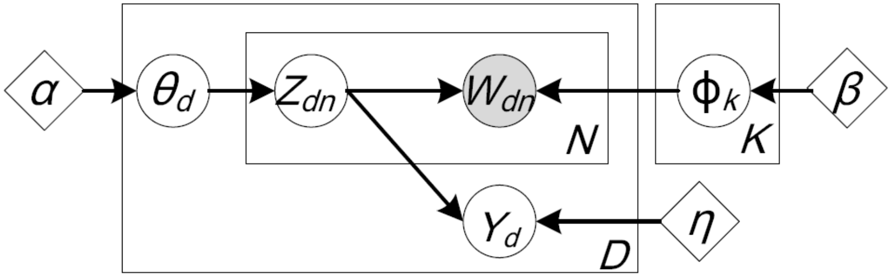
$$p(y_d | \mathbf{z}_d, \eta) = \frac{\{\exp(\eta^\top \bar{\mathbf{z}}_d)\}^{y_d}}{1 + \exp(\eta^\top \bar{\mathbf{z}}_d)} \quad \bar{z}_k = \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}(z_n^k = 1)$$

- both variational and Monte Carlo methods can be developed

(Blei & McAuliffe, NIPS'07; Wang et al., CVPR'09 ; Zhu et al., ACL 2013)

Imbalance Issue with sLDA

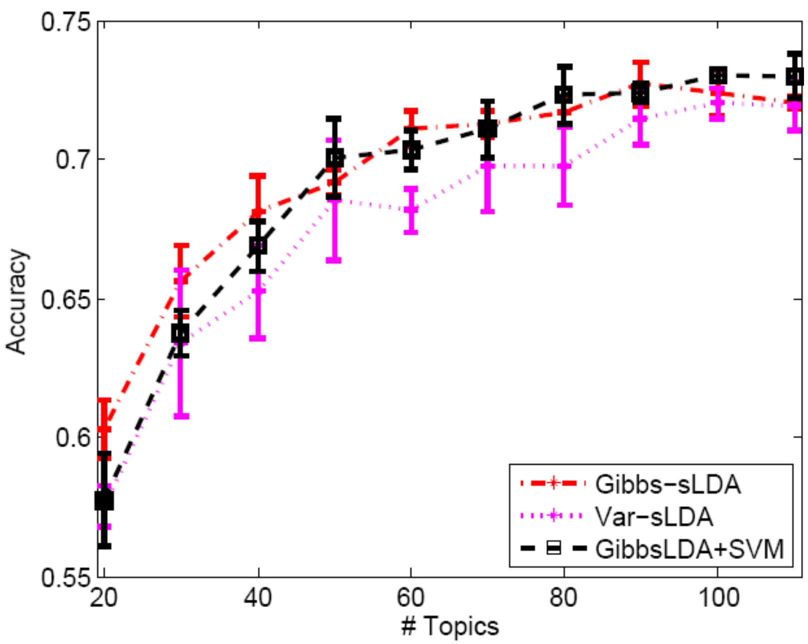
- A document has hundreds of words
- … but only one class label



- Imbalanced likelihood combination

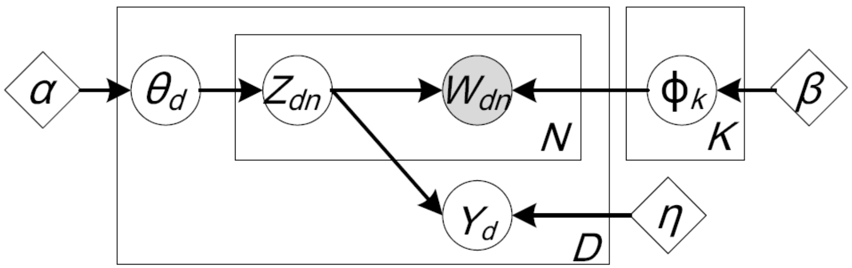
$$p(\mathbf{y}, \mathbf{W} | \mathbf{Z}, \Phi, \eta) = p(\mathbf{y} | \mathbf{Z}, \eta) p(\mathbf{W} | \mathbf{Z}, \Phi)$$

- Too weak influence from supervision



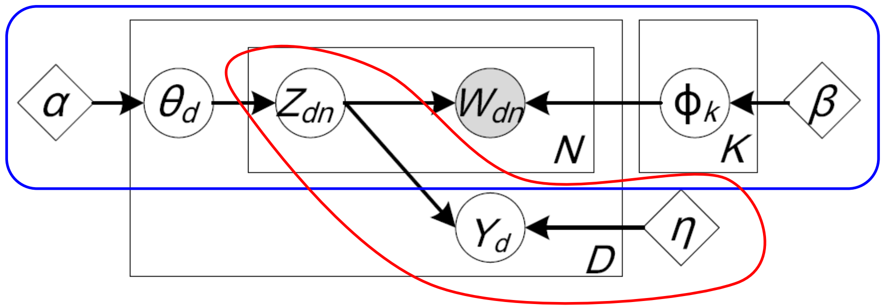
(Halpern et al., ICML 2012; Zhu et al., ACL 2013)

Max-margin Supervised Topic Models



- Can we learn supervised topic models in a max-margin way?
- How to perform posterior inference?
  - Can we do variational inference?
  - Can we do Monte Carlo?
- How to generalize to nonparametric models?

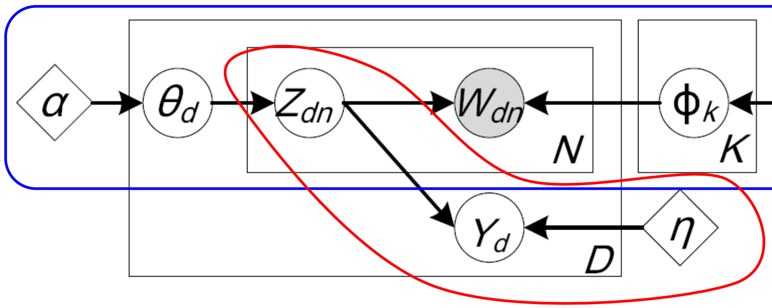
MedLDA: Max-margin Supervised Topic Models



- Two components
  - An LDA likelihood model for describing word counts
  - An max-margin classifier for considering supervising signal

- Challenges
  - *How to consider uncertainty of latent variables in defining the classifier?*
- Nice work that has inspired our design
  - Bayes classifiers (McAllester, 2003; Langford & Shawe-Taylor, 2003)
  - Maximum entropy discrimination (MED) (Jaakkola, Marina & Jebara, 1999; Jebara's Ph.D thesis and book)

MedLDA:
Max-margin Supervised Topic Models



- The averaging classifier
  - The hypothesis space is characterized by (η, Z)
  - Infer the posterior distribution

$$q(\eta, Z|\mathbf{y}, \mathbf{W})$$

  - q-weighted averaging classifier ($y_d \in \{-1, 1\}$)

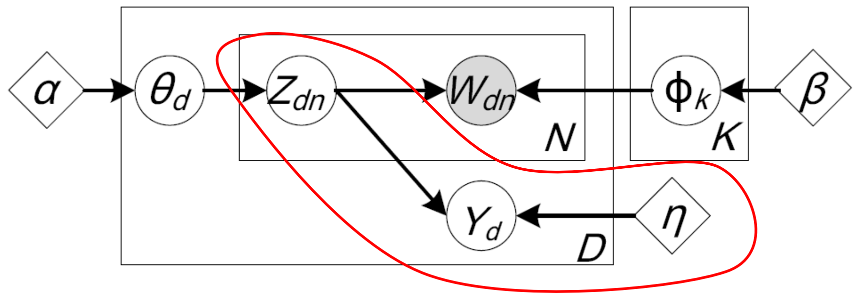
$$\hat{y} = \mathrm{sign} f(\mathbf{w}) = \mathrm{sign} \mathbb{E}_q[f(\eta, \mathbf{z}; \mathbf{w})]$$

  - where

$$f(\eta, \mathbf{z}; \mathbf{w}) = \eta^\top \bar{\mathbf{z}} \qquad \bar{z}_k = \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}(z_n^k = 1)$$

Note: Multi-class classification can be done in many ways, 1-vs-1, 1-vs-all, Crammer & Singer's method

MedLDA:
Max-margin Supervised Topic Models



- Bayesian inference with max-margin posterior constraints

$$\min_{q(\eta,\Theta,\mathbf{Z},\Phi)\in\mathcal{P}} \mathcal{L}(q(\eta,\Theta,\mathbf{Z},\Phi)) + 2c\cdot\mathcal{R}(q)$$

  - objective for Bayesian inference in LDA

$$\mathcal{L}(q) = \mathrm{KL}(q||p_0(\eta,\Theta,\mathbf{Z},\Phi)) - \mathbb{E}_q[\log p(\mathbf{W}|\mathbf{Z},\Phi)]$$

  - posterior regularization is the hinge loss

$$\mathcal{R}(q) = \sum_d \max(0, 1 - y_d f(\mathbf{w}_d))$$

Inference Algorithms

- Regularized Bayesian Inference

$$\min_{q(\eta,\Theta,\mathbf{Z},\Phi)\in\mathcal{P}} \quad \mathcal{L}(q(\eta,\Theta,\mathbf{Z},\Phi)) + 2c\cdot\mathcal{R}(q)$$

- An iterative procedure with $q(\eta,\Theta,\mathbf{Z},\Phi) = q(\eta)q(\Theta,\mathbf{Z},\Phi)$

$$\min_{q(\eta),\xi} \quad \mathrm{KL}(q(\eta)\|p_0(\eta)) + c\sum_d \xi_d$$

$$\forall d,\ \ \text{s.t.}: \quad y_d\mathbb{E}_q[\eta]^\top\mathbb{E}_q[\bar{\mathbf{z}}_d] \geq 1-\xi_d.$$

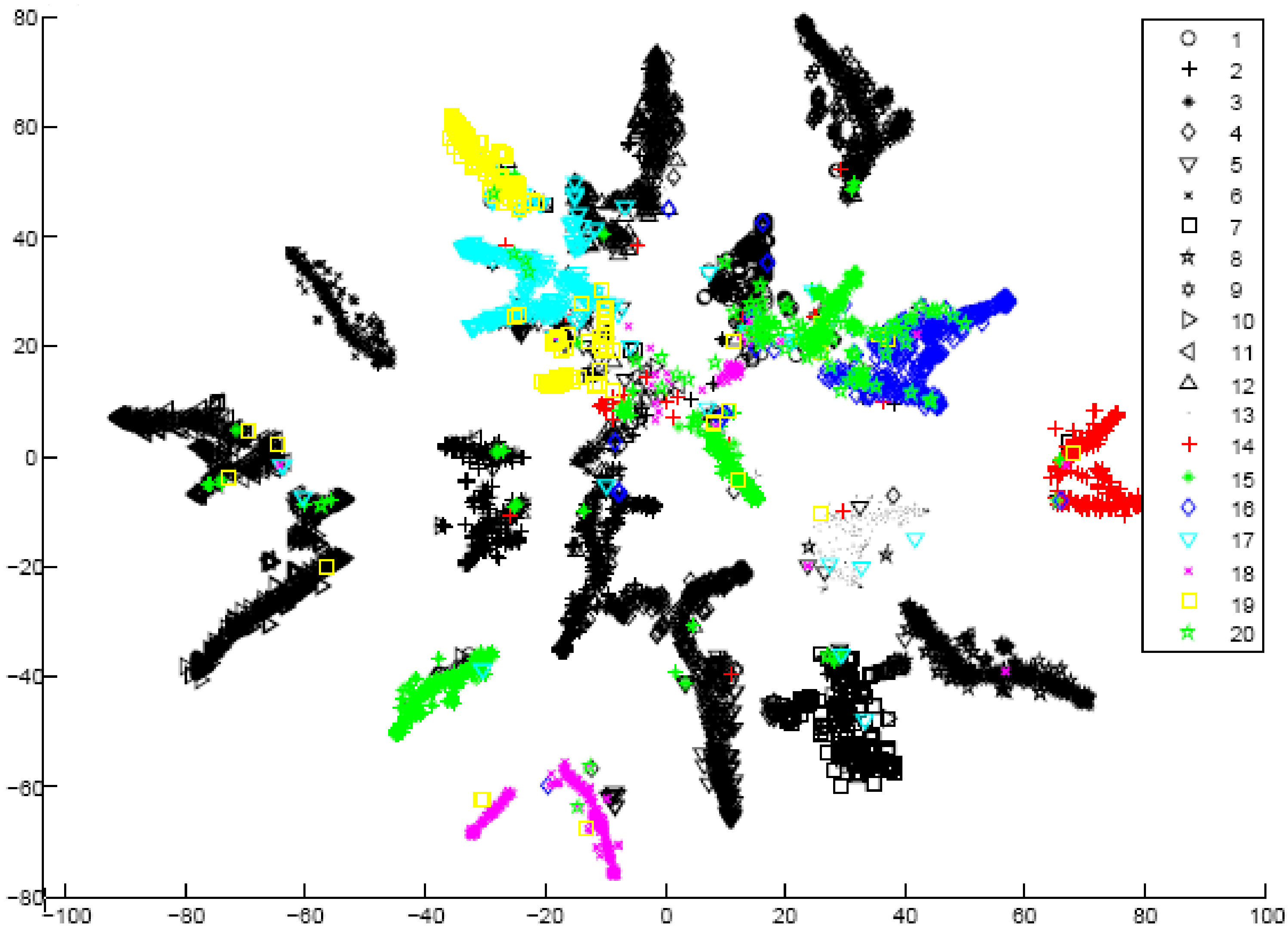
A SVM problem with a normal prior

$$\min_{q(\Theta,\mathbf{Z},\Phi),\xi} \quad \mathcal{L}(q(\Theta,\mathbf{Z},\Phi)) + c\sum_d \xi_d$$

$$\forall d,\ \ \text{s.t.}: y_d\mathbb{E}_q[\eta]^\top\mathbb{E}_q[\bar{\mathbf{z}}_d] \geq 1-\xi_d.$$

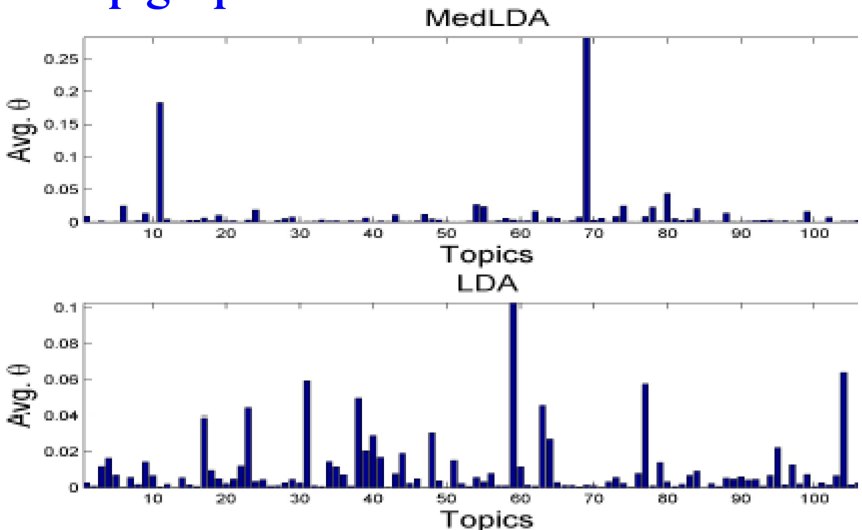
Variational approximation or Monte Carlo methods

Empirical Results on 20Newsgroups



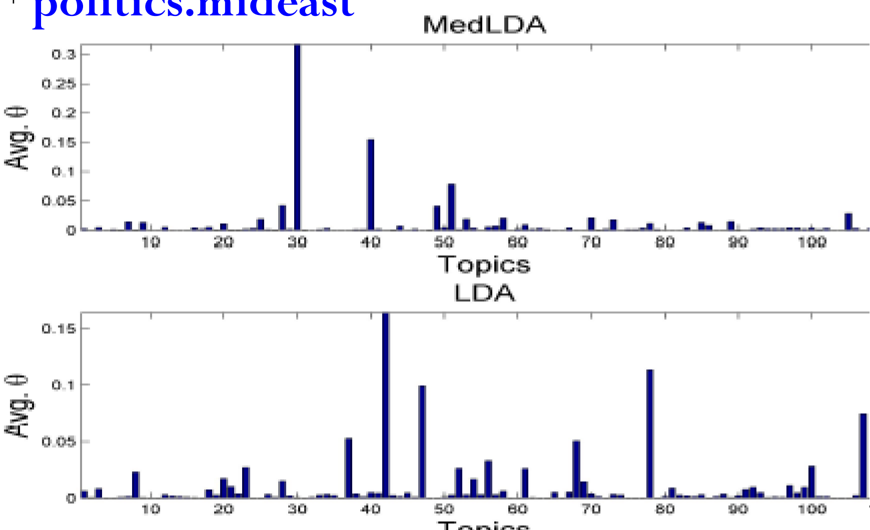
Sparser and More Salient Representations

comp.graphics



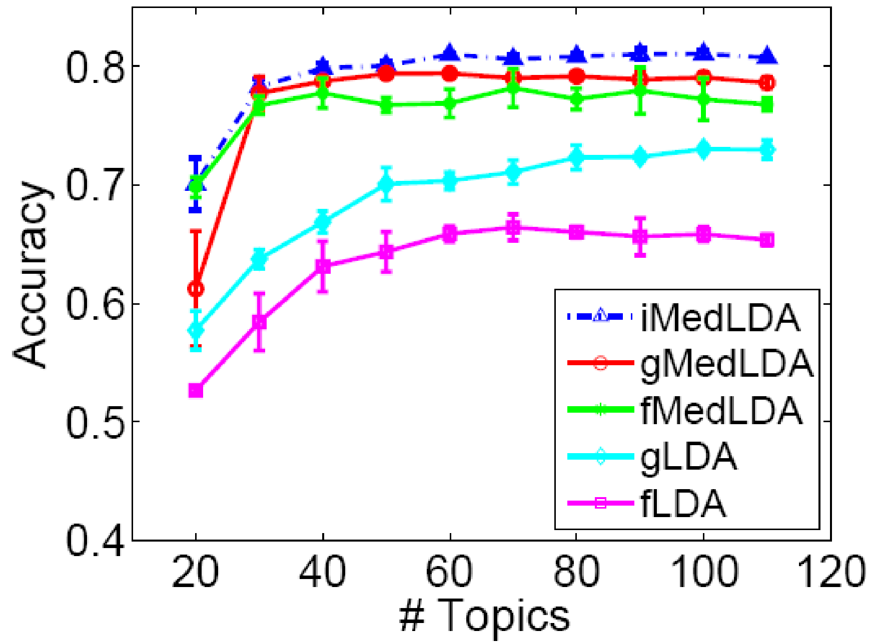
	MedLDA			LDA	
T 69	T 11	T 80	T 59	T 104	T 31
image	graphics	db	image	ftp	card
jpeg	image	key	jpeg	pub	monitor
gif	data	chip	color	graphics	dos
file	ftp	encryption	file	mail	video
color	software	clipper	gif	version	apple
files	pub	system	images	tar	windows
bit	mail	government	format	file	drivers
images	package	keys	bit	information	vga
format	fax	law	files	send	cards
program	images	escrow	display	server	graphics

politics.mideast



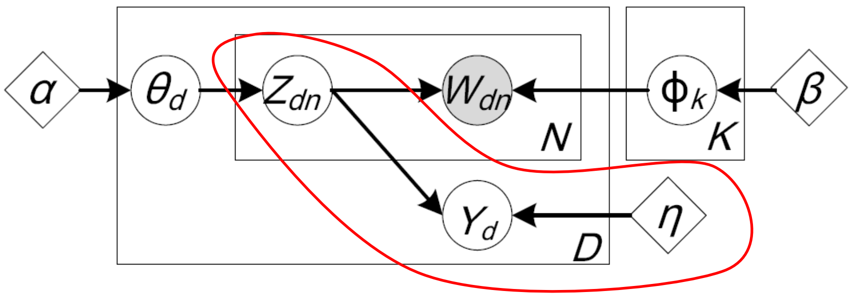
T 30	T 40	T 51	T 42	T 78	T 47
israel	turkish	israel	israel	jews	armenian
israeli	armenian	lebanese	israeli	jewish	turkish
jews	armenians	israeli	peace	israel	armenians
arab	armenia	lebanon	writes	israeli	armenia
writes	people	people	article	arab	turks
people	turks	attacks	arab	people	genocide
article	greek	soldiers	war	arabs	russian
jewish	turkey	villages	lebanese	center	soviet
state	government	peace	lebanon	jew	people
rights	soviet	writes	people	nazi	muslim

Multi-class Classification with Crammer & Singer's Approach



- Observations:
  - Inference algorithms affect the performance;
  - Max-margin learning improves a lot

Gibbs MedLDA



- The Gibbs classifier
  - The hypothesis space is characterized by (η, Z)
  - Infer the posterior distribution

$$q(\eta, Z|\mathbf{y}, \mathbf{W})$$

  - *A Gibbs* classifier

$$\hat{y}|_{\eta,\mathbf{z}} = \operatorname{sign} f(\eta, \mathbf{z}; \mathbf{w}), \ \text{ where } (\eta, \mathbf{z}) \sim q(\eta, Z|\mathbf{y}, W)$$

  - where

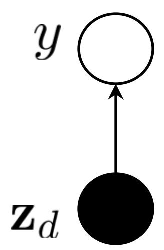
$$f(\eta, \mathbf{z}; \mathbf{w}) = \eta^\top \bar{\mathbf{z}} \qquad \bar{z}_k = \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}(z_n^k = 1)$$

(Zhu, Chen, Perkins, Zhang, ICML 2013; arXiv:1310.2816, 2013)

Gibbs MedLDA

- Let's consider the "pseudo-observed" classifier if $(\eta, \mathbf{z})$ are given

$$\hat{y}|_{\eta,\mathbf{z}} = \operatorname{sign} f(\eta, \mathbf{z}; \mathbf{w})$$



  - The empirical training error

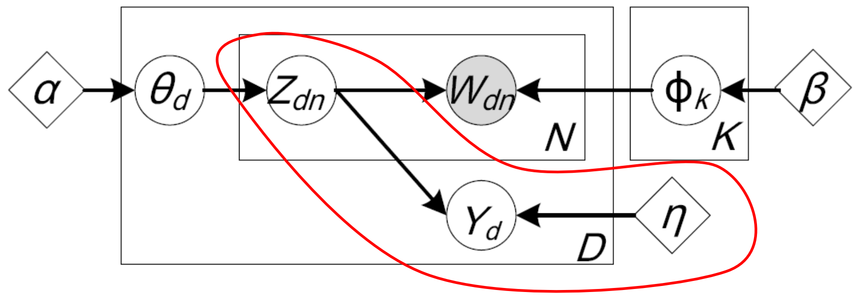
$$\hat{R}(\eta, Z) = \sum_{d=1}^{D} \mathbb{I}(\hat{y}_d|_{\eta,\mathbf{z}_d} \neq y_d)$$

  - A good convex surrogate loss is the hinge loss (an upper bound)

$$\mathcal{R}(\eta, \mathbf{Z}) = \sum_{d=1}^{D} \max(0, \zeta_d), \text{ where } \zeta_d = 1 - y_d \eta^\top \bar{\mathbf{z}}_d$$

- Now the question is how to consider the uncertainty?
  - A Gibbs classifier takes the expectation!

Gibbs MedLDA



- Bayesian inference with max-margin posterior constraints

$$\min_{q(\eta,\Theta,\mathbf{Z},\Phi)\in\mathcal{P}} \mathcal{L}(q(\eta,\Theta,\mathbf{Z},\Phi)) + 2c\cdot \boxed{\mathcal{R}'(q)}$$

  - an upper bound of the expected training error (empirical risk)

$$\mathcal{R}'(q) = \sum_{d=1}^{D} \mathbb{E}_q[\max(0,\zeta_d)] \;\geq \sum_d \mathbb{E}_q[\mathbb{I}(\hat{y}_d \neq y_d)]$$

Gibbs MedLDA vs. MedLDA

- The MedLDA problem

$$\min_{q(\eta,\Theta,\mathbf{Z},\Phi)\in\mathcal{P}} \mathcal{L}(q(\eta,\Theta,\mathbf{Z},\Phi)) + 2c\cdot\mathcal{R}(q)$$

$$\mathcal{R}(q) = \sum_d \max(0, 1 - y_d f(\mathbf{w}_d))$$

- Applying Jensen's Inequality, we have

$$\mathcal{R}'(q) \geq \mathcal{R}(q)$$

  - Gibbs MedLDA can be seen as a relaxation of MedLDA

Gibbs MedLDA

- The problem

$$\min_{q(\eta,\Theta,\mathbf{Z},\Phi)\in\mathcal{P}} \mathcal{L}(q(\eta,\Theta,\mathbf{Z},\Phi)) + 2c\cdot\mathcal{R}(q)$$

- Solve with Lagrangian methods

$$q(\eta,\Theta,\mathbf{Z},\Phi) = \frac{p_0(\eta,\Theta,\mathbf{Z},\Phi)p(\mathbf{W}|\mathbf{Z},\Phi)\phi(\mathbf{y}|\mathbf{Z},\eta)}{\psi(\mathbf{y},\mathbf{W})}$$

  - The pseudo-likelihood $\phi(\mathbf{y}|\mathbf{Z},\eta) = \prod_d \phi(y_d|\eta,\mathbf{z}_d)$

$$\phi(y_d|\mathbf{z}_d,\eta) = \exp\{-2c\max(0,\zeta_d)\}$$

Gibbs MedLDA

- **Lemma** [Scale Mixture Rep.] (Polson & Scott, 2011):
  - The pseudo-likelihood can be expressed as

$$\phi(y_d|\mathbf{z}_d,\eta) = \int_0^\infty \frac{1}{\sqrt{2\pi\lambda_d}} \exp\left(-\frac{(\lambda_d + c\zeta_d)^2}{2\lambda_d}\right) d\lambda_d$$

- What does the lemma mean?
  - It means:

$$q(\eta,\Theta,\mathbf{Z},\Phi) = \int q(\eta,\lambda,\Theta,\mathbf{Z},\Phi)d\lambda$$

where $q(\eta,\lambda,\Theta,\mathbf{Z},\Phi) = \dfrac{p_0(\eta,\Theta,\mathbf{Z},\Phi)p(\mathbf{W}|\mathbf{Z},\Phi)\phi(\mathbf{y},\lambda|\mathbf{Z},\eta)}{\psi(\mathbf{y},\mathbf{W})}$

$$\phi(\mathbf{y},\lambda|\mathbf{Z},\eta) = \prod_d \frac{1}{\sqrt{2\pi\lambda_d}} \exp\left(-\frac{(\lambda_d + c\zeta_d)^2}{2\lambda_d}\right)$$

A Gibbs Sampling Algorithm

- Infer the joint distribution

$$q(\eta, \lambda, \Theta, \mathbf{Z}, \Phi) = \frac{p_0(\eta, \Theta, \mathbf{Z}, \Phi) p(\mathbf{W}|\mathbf{Z}, \Phi) \phi(\mathbf{y}, \lambda|\mathbf{Z}, \eta)}{\psi(\mathbf{y}, \mathbf{W})}$$

- A Gibbs sampling algorithm iterates over:
  - Sample $\eta^{t+1} \sim q(\eta|\lambda^t, \Theta^t, \mathbf{Z}^t, \Phi^t) \propto p_0(\eta)\phi(\mathbf{y}, \lambda^t|\mathbf{Z}^t, \eta)$
    - a Gaussian distribution when the prior is Gaussian
  - Sample $\lambda^{t+1} \sim q(\lambda|\eta^{t+1}, \Theta^t, \mathbf{Z}^t, \Phi^t) \propto \phi(\mathbf{y}, \lambda|\mathbf{Z}^t, \eta^{t+1})$
    - a generalized inverse Gaussian distribution, i.e., λ^{-1} follows inverse Gaussian
  - Sample $(\Theta, \mathbf{Z}, \Phi)^{t+1} \sim p(\Theta, \mathbf{Z}, \Phi|\eta^{t+1}, \lambda^{t+1})$
    $\propto p_0(\Theta, \mathbf{Z}, \Phi)p(\mathbf{W}|\mathbf{Z}, \Phi)\phi(\mathbf{y}, \boldsymbol{\lambda}^{t+1}|\mathbf{Z}, \boldsymbol{\eta}^{t+1})$
    - a supervised LDA model with closed-form local conditionals by exploring data independency.

A Collapsed Gibbs Sampling Algorithm

- The collapsed joint distribution

$$q(\eta, \lambda, \mathbf{Z}) = \int q(\eta, \lambda, \Theta, \mathbf{Z}, \Phi) d\Theta d\Phi$$

- A Gibbs sampling algorithm iterates over:
  - Sample $\eta^{t+1} \sim q(\eta|\lambda^t, \mathbf{Z}^t) \propto p_0(\eta)\phi(\mathbf{y}, \lambda^t|\mathbf{Z}^t, \eta)$
    - a Gaussian distribution when the prior is Gaussian
  - Sample $\lambda^{t+1} \sim q(\lambda|\eta^{t+1}, \mathbf{Z}^t) \propto \phi(\mathbf{y}, \lambda|\mathbf{Z}^t, \eta^{t+1})$
    - a generalized inverse Gaussian distribution, i.e., λ^{-1} follows inverse Gaussian
  - Sample $\mathbf{Z}^{t+1} \sim q(\mathbf{Z}|\eta^{t+1}, \lambda^{t+1}) \propto \int p_0(\Theta, \mathbf{Z}, \Phi) p(\mathbf{W}|\mathbf{Z}, \Phi) \phi(\mathbf{y}, \boldsymbol{\lambda}^{t+1}|\mathbf{Z}, \boldsymbol{\eta}^{t+1}) d\Theta d\Phi$
    - closed-form local conditionals

$$q(z_{dn}^k = 1|\mathbf{Z}_{\neg}, \eta, \lambda, w_{dn} = t)$$

The Collapsed Gibbs Sampling Algorithm

Algorithm 1 Collapsed Gibbs Sampling Algorithm

```
Initialization: set λ = 1 and randomly draw z_dk from a uniform distribution.
for m = 1 to M do
    draw the classifier from the normal distribution (11)
    for d = 1 to D do
        for each word n in document d do
            draw the topic using distribution (12)
        end for
        draw λ_d^{-1} (and thus λ_d) from distribution (13).
    end for
end for
```

Easy to Parallelize

Some Analysis

- The Markov chain is guaranteed to converge

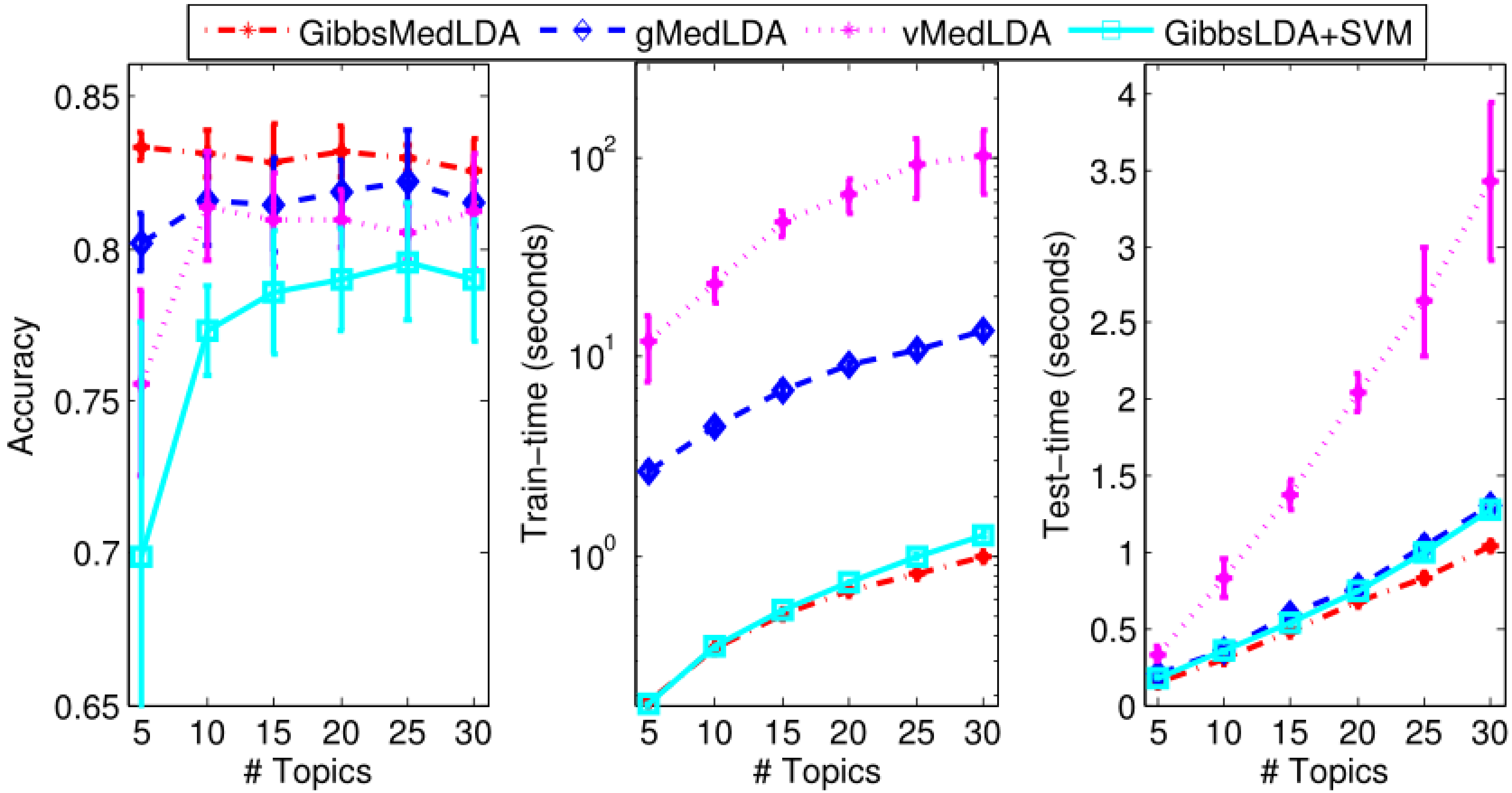
- Per-iteration time complexity

$$\mathcal{O}(K^3 + N_{total}K)$$

  - N_{total} the total number of words

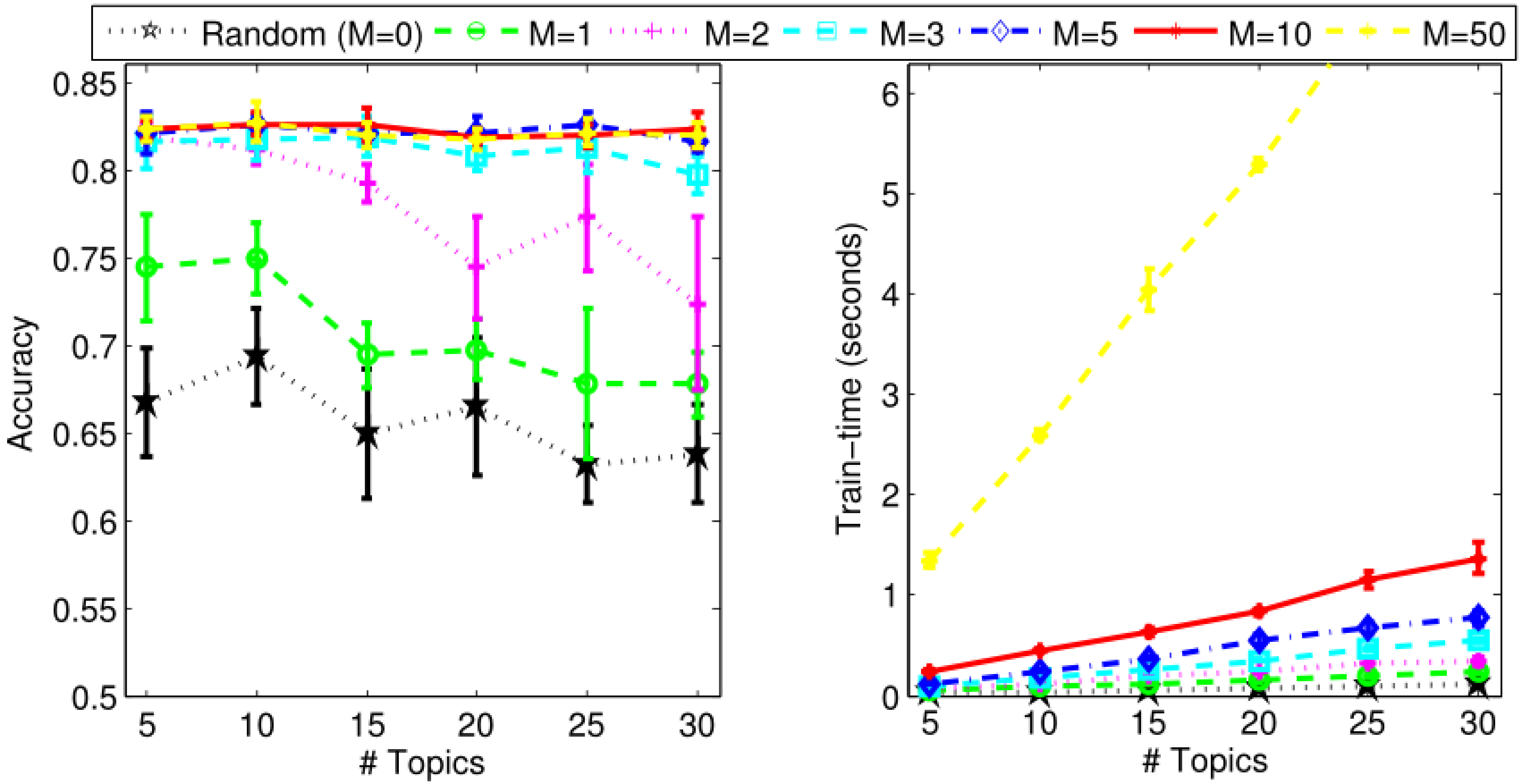
Experiments

- 20Newsgroups binary classification



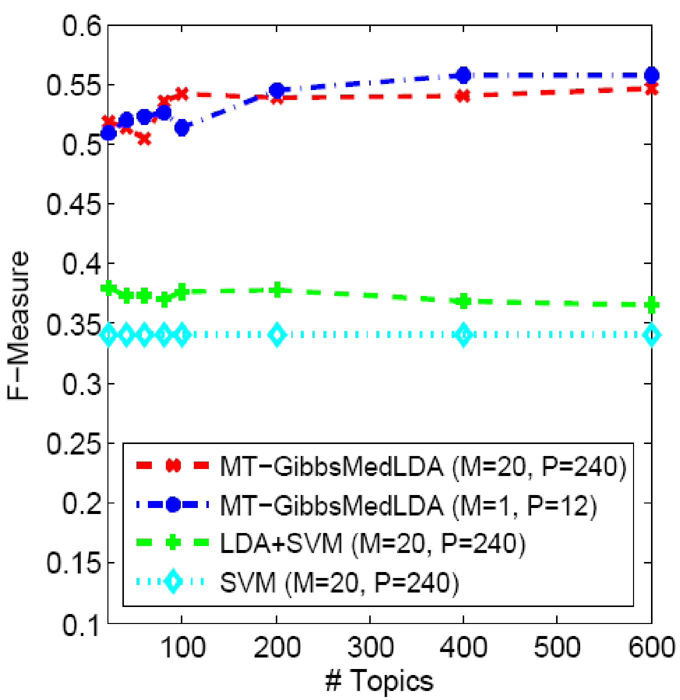
Experiments

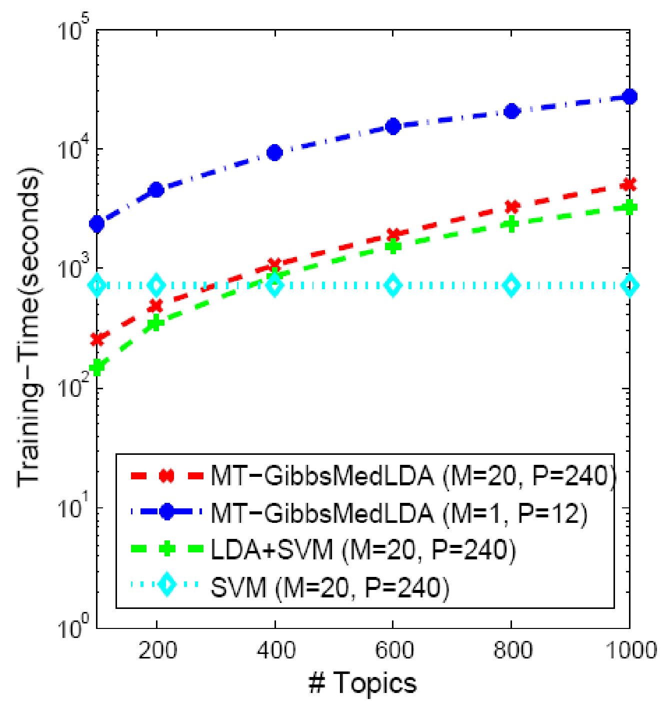
- **Sensitivity to burn-in**: binary classification



Distributed Inference Algorithms

- Leverage big clusters
- Allow learning big models that can’t fit on a single machine





- 20 machines;
- 240 CPU cores

- 1.1M multi-labeled Wiki pages
- 20 categories (scale to hundreds/thousands of categories)

[Zhu, Zheng, Zhou, & Zhang, KDD2013]

Summary

- RegBayes: bridging the gap between Bayesian methods, learning and optimization

- Max-margin supervised topic models
  - with averaging classifiers + variational inference
  - with Gibbs classifiers + MCMC sampling with DA

Future Work

- Dealing with weak supervision and other forms of side information
- RegBayes algorithms for network models
- Learning with dynamic and spatial structures
- Fast and scalable inference architectures
- Generalization bounds

Acknowledgements

- Collaborators:
  - Prof. Bo Zhang (Tsinghua)、Prof. Eric P. Xing (CMU)、Prof. Li Fei-Fei (Stanford)
  - Amr Ahmed (CMU), Ning Chen (Tsinghua), Ni Lao (CMU), Seunghak Lee (CMU), Li-jia Li (Stanford), Xiaojiang Liu (USTC), Xiaolin Shi (Stanford), Hao Su (Stanford), Yuandong Tian (CMU).
- Students at Tsinghua:
  - Aonan Zhang Minjie Xu Hugh Perkins

  Jianfei Chen, Bei Chen, Shike Mei, Xun Zheng, Fei Xia, Zi Wang, Tianlin Shi, Yining Wang, Li Zhou, etc.
- Funding:

Thanks!

Some code available at:

http://www.ml-thu.net/~jun