Big Learning with Bayesian Methods

Jun Zhu
dcszj@mail.tsinghua.edu.cn
http://bigml.cs.tsinghua.edu.cn/~jun
Department of Computer Science and Technology
Tsinghua University

ADL 2014, Shenzhen

Overview

- **Part I (120 m):** Basic theory, models, and algorithms
  - Basics of Bayesian methods
  - Regularized Bayesian inference and examples

- **Part II (60 m):** Scalable Bayesian methods
  - Online learning
  - Large-scale topic graph learning and visualization

- **Part III (20 m):** Q&A

Readings

- Big Learning with Bayesian Methods, J. Zhu, J. Chen, & W. Hu, arXiv 1411.6370, preprint, 2014

Basic Rules of Probability

- Concepts

$$\begin{array}{rl} p(X) & \text{probability of } X \\ p(X|\mathcal{M}) & \text{conditional probability of } X \text{ given } \mathcal{M} \\ p(X,\mathcal{M}) & \text{joint probability of } X \text{ and } \mathcal{M} \end{array}$$

- Joint probability – product rule

$$p(X,\mathcal{M}) = p(X|\mathcal{M})p(\mathcal{M})$$

- Marginal probability – sum/integral rule

$$p(X) = \int p(X|\mathcal{M})p(\mathcal{M})d\mathcal{M}$$

Bayes' Rule

- Combining the definition of conditional prob. with the product and sum rules, we have Bayes' rule or Bayes' theorem

$$
\begin{aligned}
p(\mathcal{M}|X) &= \frac{p(X,\mathcal{M})}{p(X)} \\
&= \frac{p(\mathcal{M})p(X|\mathcal{M})}{\int p(\mathcal{M})p(X|\mathcal{M})d\mathcal{M}}
\end{aligned}
$$

Thomas Bayes (1702 – 1761)

- "*An Essay towards Solving a Problem in the Doctrine of Chances*" published at Philosophical Transactions of the Royal Society of London in 1763

Bayes' Rule Applied to Machine Learning

- Let $\mathcal{D}$ be a given data set; $\mathcal{M}$ be a model

$$p(\mathcal{M}|\mathcal{D}) = \frac{p(\mathcal{M})p(\mathcal{D}|\mathcal{M})}{p(\mathcal{D})}$$

$p(\mathcal{M})$	prior probability of $\mathcal{M}$
$p(\mathcal{D}|\mathcal{M})$	likelihood of $\mathcal{M}$ on data
$p(\mathcal{M}|\mathcal{D})$	posterior probability of $\mathcal{M}$ given $\mathcal{D}$
$p(\mathcal{D})$	marginal likelihood or evidence

- Model Comparison: $\mathbb{M} = \{\mathcal{M}\}$

$$p(\mathbb{M}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbb{M})p(\mathbb{M})}{p(\mathcal{D})} \qquad p(\mathcal{D}|\mathbb{M}) = \int p(\mathcal{D}|\mathcal{M}, \mathbb{M})p(\mathcal{M}|\mathbb{M})d\mathcal{M}$$

- Prediction:

$$p(x|\mathcal{D}, \mathbb{M}) = \int p(x|\mathcal{M}, \mathcal{D}, \mathbb{M})p(\mathcal{M}|\mathcal{D}, \mathbb{M})d\mathcal{M}$$

under some common assumptions

$$p(x|\mathcal{M})$$

Common Questions

- Why be Bayesian?
- Where does the prior come from?
- How do we do these integrals?

Why be Bayesian?

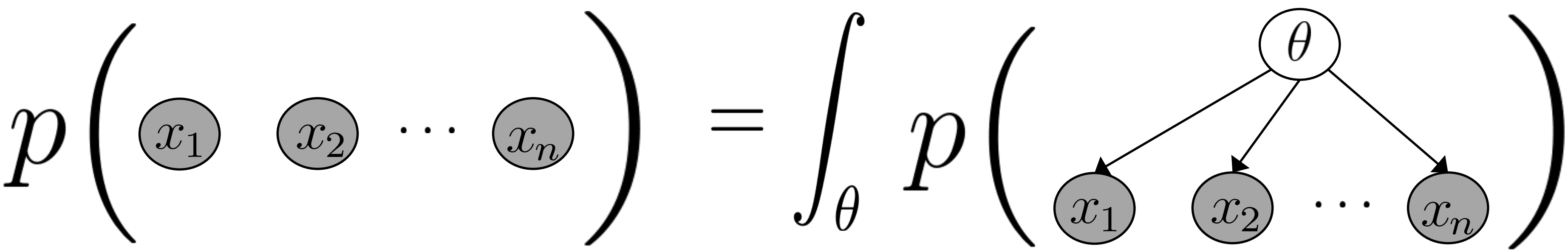
- One of many answers
- Infinite Exchangeability:

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

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

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

for some random variable θ



How to Choose Priors?

- **Objective priors** -- noninformative priors that attempt to capture ignorance and have good frequentist properties
- **Subjective priors** -- priors should capture our beliefs as well as possible
- **Hierarchical priors** -- multiple layers of priors

$$p(\mathcal{M}) = \int p(\mathcal{M}|\alpha)p(\alpha)d\alpha = \int \int p(\mathcal{M}|\alpha)p(\alpha|\beta)p(\beta)d\alpha d\beta = \cdots$$

  - the higher, the weaker
- **Empirical priors** -- Learn some of the parameters of the prior from the data; known as "Empirical Bayes"

$$p(\mathcal{M}|\hat{\alpha}) \qquad \hat{\alpha} = \operatorname*{argmax}_{\alpha} p(\mathcal{D}|\alpha)$$

  - **Pros**: robust – overcomes some limitations of mis-specification
  - **Cons**: double counting of evidence / overfitting

How to Choose Priors?

- Conjugate and Non-conjugate tradeoff
- Conjugate priors are relatively easier to compute, but they might be limited
  - Ex: Gaussian-Gaussian, Beta-Bernoulli, Dirichlet-Multinomial, etc. (see next slide for an example)
- Non-conjugate priors are more flexible, but harder to compute
  - Ex: LogisticNormal-Multinomial

Example 1: Multinomial-Dirichlet Conjugacy

Posterior is in the same class as the prior

- Let

$$X \sim \text{Multinomial}(\pi), \text{ and } \pi \sim \text{Dirichlet}(\alpha)$$

- The posterior

$$p(\pi|X) \propto p(X|\pi)p(\pi)$$

$$\propto (\pi_1^{x_1} \cdots \pi_K^{x_K})(\pi_1^{\alpha_1 - 1} \cdots \pi_K^{\alpha_K - 1})$$

$$= \text{Dirichlet}(\pi_1^{x_1 + \alpha_1 - 1} \cdots \pi_K^{x_K + \alpha_K - 1})$$

which is $\text{Dirichlet}(\alpha + \mathbf{x})$

How do We Compute the Integrals?

- Recall that:

$$p(\mathcal{D}|\mathbb{M}) = \int p(\mathcal{D}|\mathcal{M}, \mathbb{M}) p(\mathcal{M}|\mathbb{M}) d\mathcal{M}$$

- This can be a very high dimensional integral

- If we consider latent variables, it leads to additional dimensions to be integrated out

$$p(\mathcal{D}|\mathbb{M}) = \int\int p(\mathcal{D}, H|\mathcal{M}, \mathbb{M}) p(\mathcal{M}|\mathbb{M}) dH d\mathcal{M}$$

  - This could be very complicated!

Approximate Bayesian Inference

- In many cases, we resort to approximation methods
- Common examples
  - Variational approximations
  - Markov chain Monte Carlo methods (MCMC)
  - Expectation Propagation (EP)
  - Laplace approximation
  - …
- Developing advanced inference algorithms is an active area!

Basics of Variational Approximation

- We can lower bound the marginal likelihood

$$
\begin{aligned}
\log p(\mathcal{D}|\mathbb{M}) &= \log \int\int p(\mathcal{D}, H|\mathcal{M}, \mathbb{M}) p(\mathcal{M}|\mathbb{M}) dH d\mathcal{M} \\
&= \log \int\int q(H, \mathcal{M}) \frac{p(\mathcal{D}, H|\mathcal{M}, \mathbb{M}) p(\mathcal{M}|\mathbb{M})}{q(H, \mathcal{M})} dH d\mathcal{M} \\
&\geq \int\int q(H, \mathcal{M}) \log \frac{p(\mathcal{D}, H|\mathcal{M}, \mathbb{M}) p(\mathcal{M}|\mathbb{M})}{q(H, \mathcal{M})} dH d\mathcal{M}
\end{aligned}
$$

  - Note: the lower bound is tight if no assumptions made

- **Mean-field assumptions**: a factorized approximation

$$q(H, \mathcal{M}) = q(H) q(\mathcal{M})$$

  - optimizes the lower bound with the assumption leads to local optimums

Basics of Monte Carlo Methods

- a class of computational algorithms that rely on repeated random sampling to compute their results.
- tend to be used when it is infeasible to compute an exact result with a deterministic algorithm
- was coined in the 1940s by John von Neumann, Stanislaw Ulam and Nicholas Metropolis

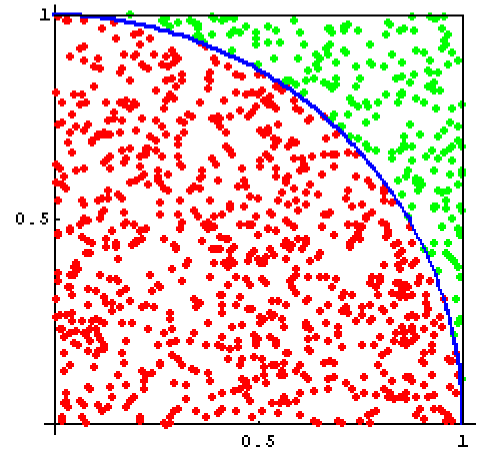
Games of Chance

Monte Carlo Methods to Calculate Pi

- Computer Simulation

$$\hat{\pi} = 4 \times \frac{m}{N}$$

  - N: # points inside the square
  - m: # points inside the circle

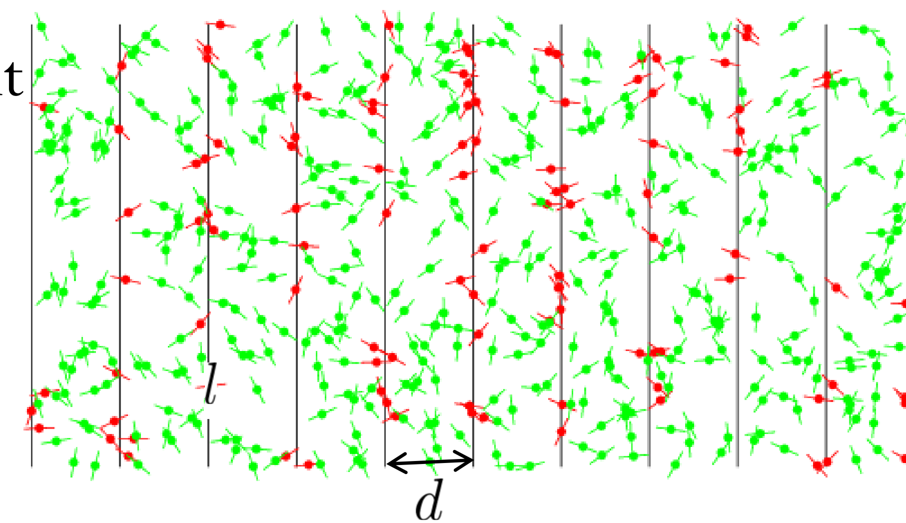


- Bufffon's Needle Experiment

$$\hat{\pi} = \frac{2Nx}{m}$$

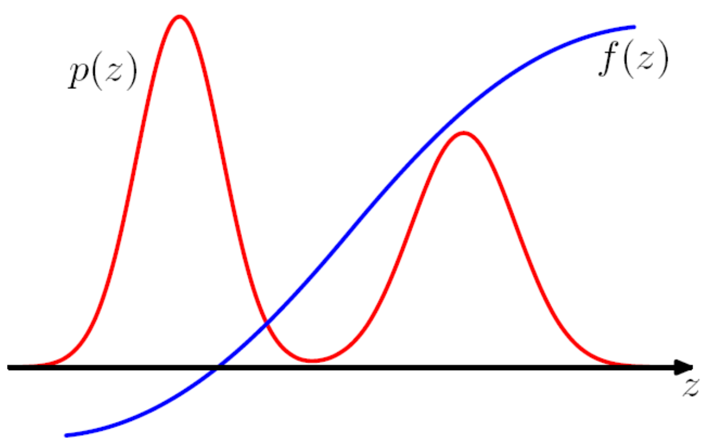
  - m: # line crossings

$$x = \frac{l}{d}$$



Problems to be Solved

- **Sampling**
  - to generate a set of samples $\{\mathbf{z}_l\}_{l=1}^{L}$ from a given probability distribution $p(\mathbf{z})$
  - the distribution is called target distribution
  - can be from statistical physics or data modeling
- **Integral**
  - To estimate expectations of functions under this distribution



$$\mathbb{E}[f] = \int f(\mathbf{z})p(\mathbf{z})d\mathbf{z}$$

Use Sample to Estimate the Target Dist.

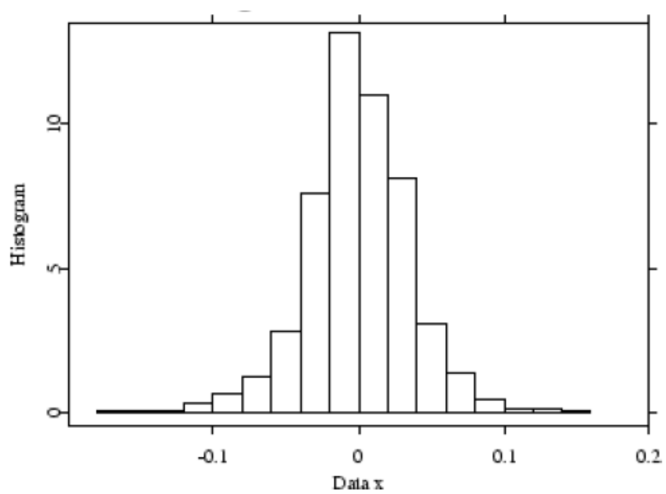
- Draw a set of independent samples (a hard problem)

$$\forall 1 \leq l \leq L, \quad \mathbf{z}^{(l)} \sim p(\mathbf{z})$$

- Estimate the target distribution as count frequency

$$p(\mathbf{z}) \approx \frac{1}{L} \sum_{l=1}^{L} \delta_{\mathbf{z}^{(l)}}(\mathbf{z})$$

Histogram with Unique Points as the Bins



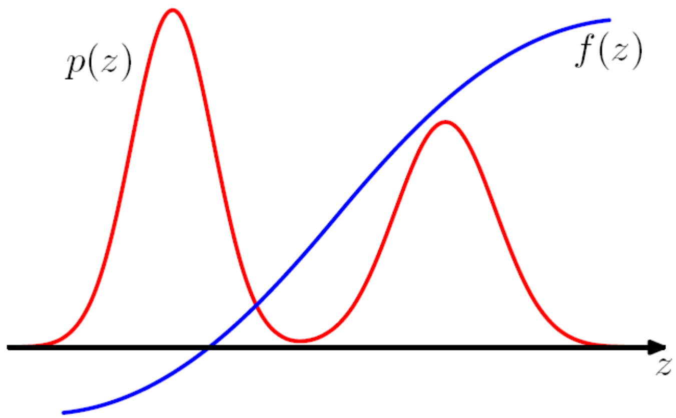
Basic Procedure of Monte Carlo Methods

- Draw a set of independent samples

$$\forall 1 \le l \le L, \quad \mathbf{z}^{(l)} \sim p(\mathbf{z})$$

- Approximate the expectation with

$$\hat{f} = \frac{1}{L}\sum_{l=1}^{L} f(\mathbf{z}^{(l)})$$



  - where is the distribution p? $p(\mathbf{z}) \approx \frac{1}{L}\sum_{l=1}^{L}\delta_{\mathbf{z}^{(l)}}(\mathbf{z})$ Histogram with Unique Points as the Bins
  - why this is good?

$$\mathbb{E}[\hat{f}] = \mathbb{E}[f] \qquad \mathrm{var}[\hat{f}] = \frac{1}{L}\mathbb{E}[(f - \mathbb{E}[f])^2]$$

  - Accuracy of estimator does not depend on dimensionality of z
  - High accuracy with few (10-20 independent) samples
  - However, obtaining independent samples is often not easy!

Why Sampling is Hard?

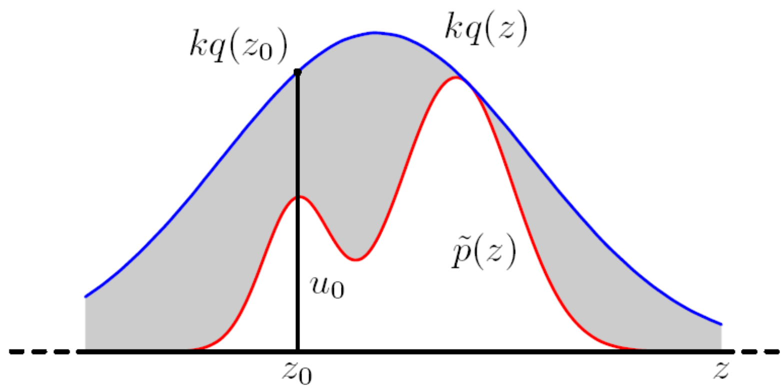
- **Assumption**
  - The target distribution can be evaluated, at least to within a multiplicative constant, i.e.,

$$p(\mathbf{z}) = p^*(\mathbf{z})/Z$$

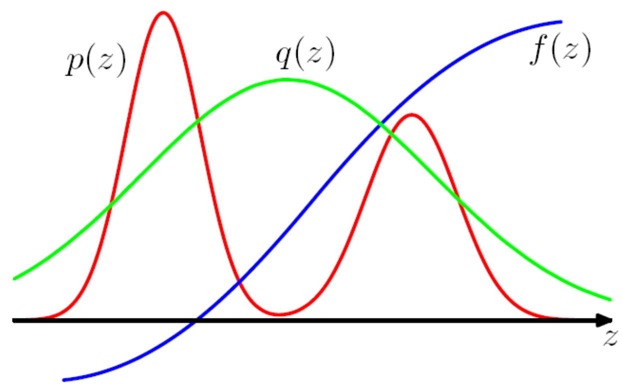
  - where $p^*(\mathbf{z})$ can be evaluated
- **Two difficulties**
  - Normalizing constant is typically unknown
  - Drawing samples in high-dimensional space is challenging

Many Sampling Methods

- Rejection sampling



- Importance sampling



- Markov chain Monte Carlo (MCMC)

Basics of MCMC

- To draw samples from a desired distribution $p(x|\mathcal{D})$

- We define a Markov chain

$$x_0 \to x_1 \to x_2 \to x_3 \to \cdots$$

  - where
$$p_t(x) = \int p_{t-1}(x')q(x; x')dx'$$

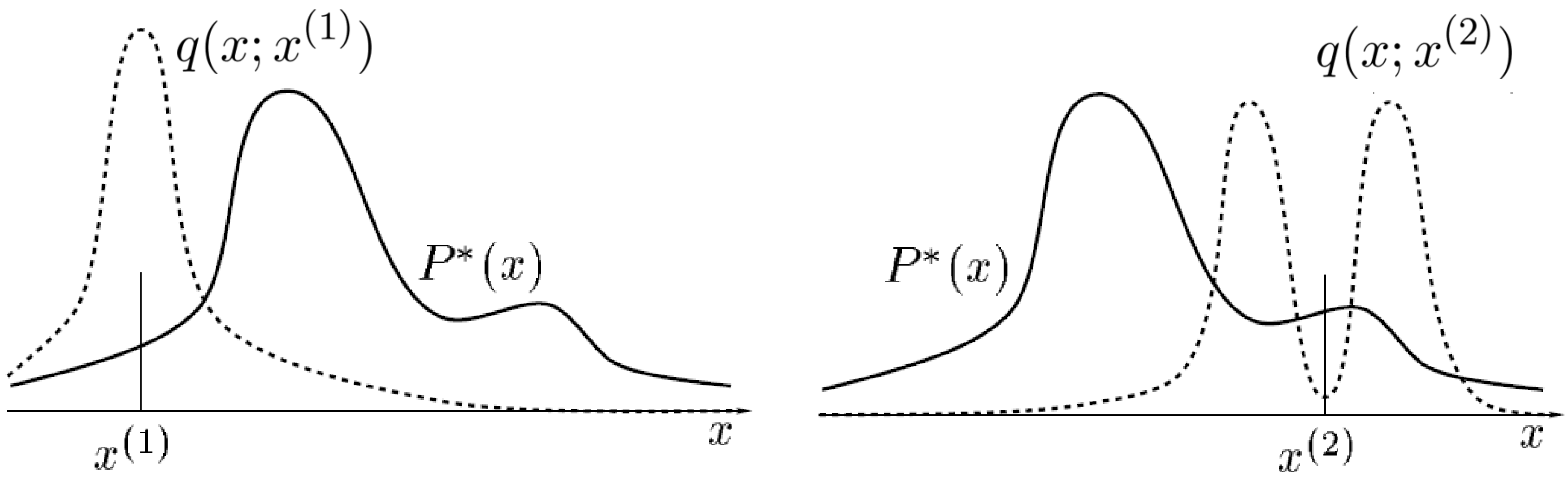
  - $q(x; x')$ is the transition kernel

- $p(x|\mathcal{D})$ is an **invariant (or stationary) distribution** of the Markov chain q iff:

$$p(x|\mathcal{D}) = \int p(x'|\mathcal{D})q(x; x')dx'$$

Geometry of MCMC

- Proposal depends on current state
- Not necessarily similar to the target
- Can evaluate the un-normalized target



Gibbs Sampling

- A special case of Metropolis-Hastings algorithm
- Consider the distribution $p(\mathbf{x}) = p(x_1, \ldots, x_M)$

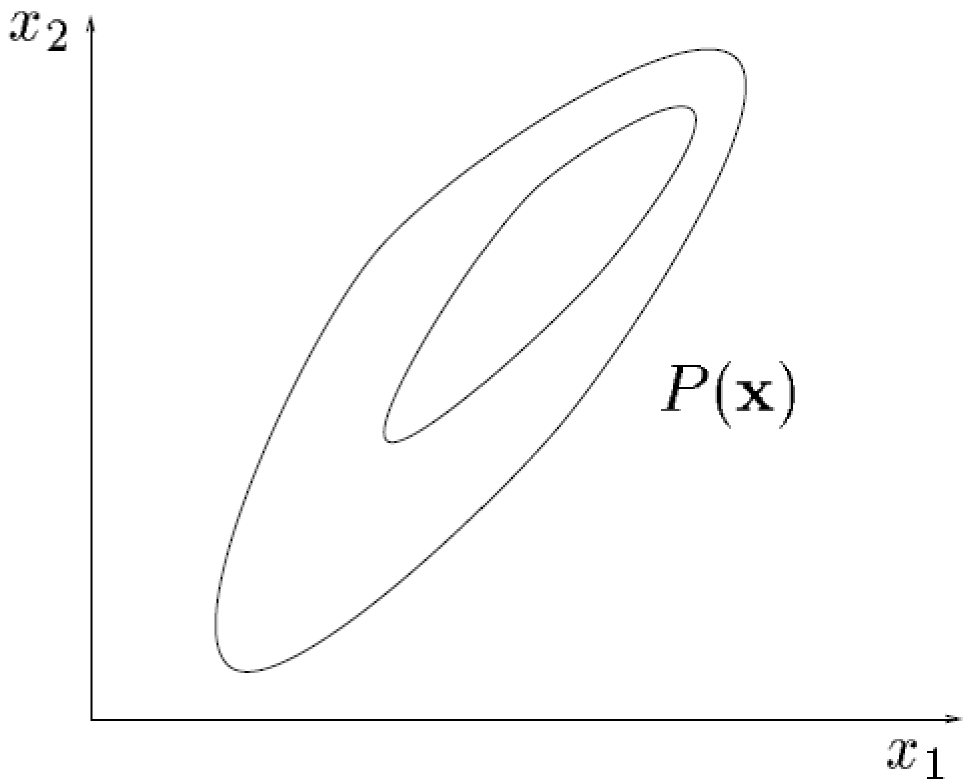
- Gibbs sampling performs the follows
  - Initialize $\{x_i : i = 1, \ldots, M\}$
  - For $\tau = 1, \ldots, T$
    - Sample $x_1^{(\tau+1)} \sim p(x_1 | x_2^{(\tau)}, x_3^{(\tau)}, \ldots, x_M^{(\tau)})$

      $\vdots$
    - Sample $x_j^{(\tau+1)} \sim p(x_j | x_1^{(\tau+1)}, \ldots, x_{j-1}^{(\tau+1)}, x_{j+1}^{(\tau)}, \ldots, x_M^{(\tau)})$

      $\vdots$
    - Sample $x_M^{(\tau+1)} \sim p(x_j | x_1^{(\tau+1)}, x_2^{(\tau+1)}, \ldots, x_{M-1}^{(\tau+1)})$

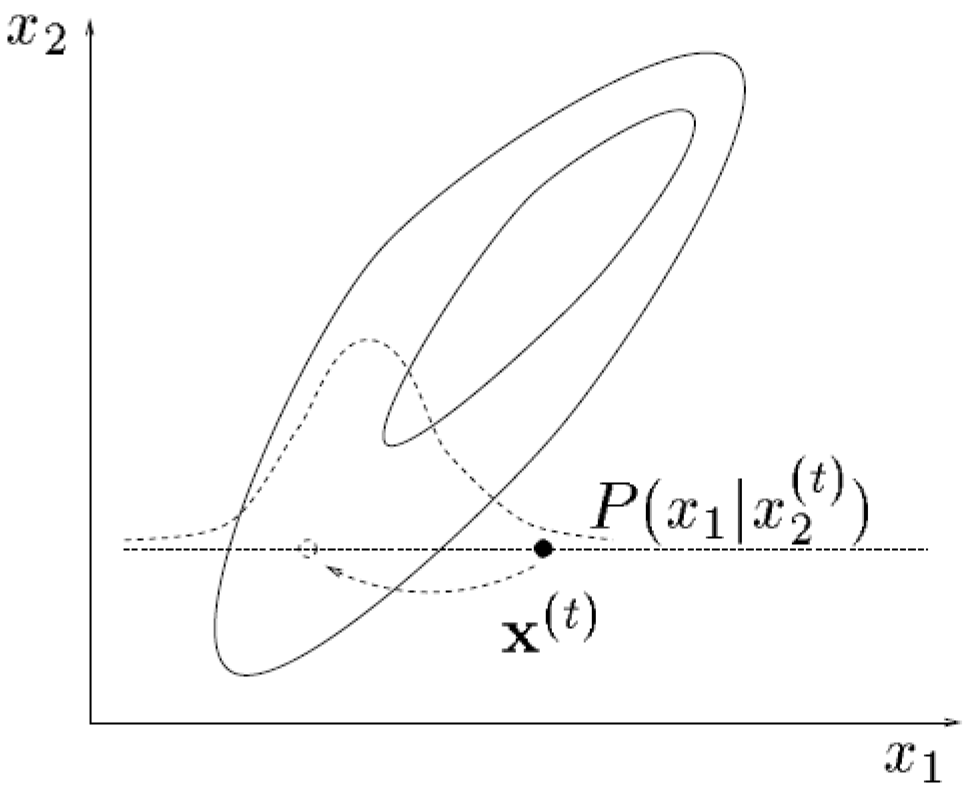
Geometry of Gibbs Sampling

- The target distribution in 2 dimensional space



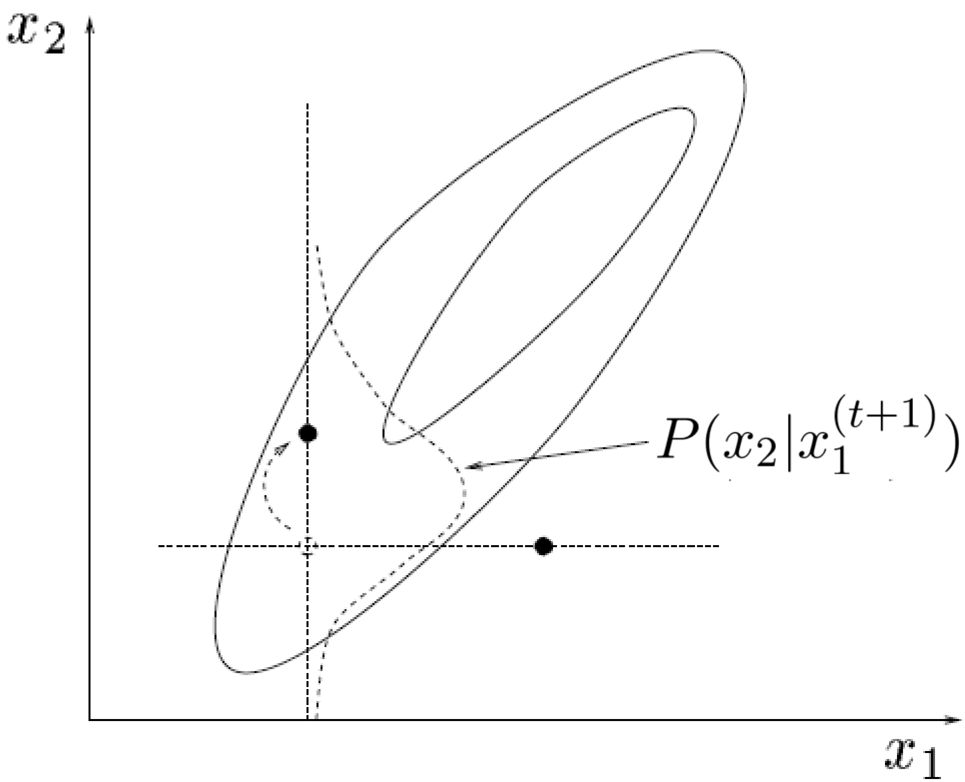
Geometry of Gibbs Sampling

- Starting from a state $\mathbf{x}^{(t)}$, $x_1^{(t+1)}$ is sampled from $P(x_1|x_2^{(t)})$



Geometry of Gibbs Sampling

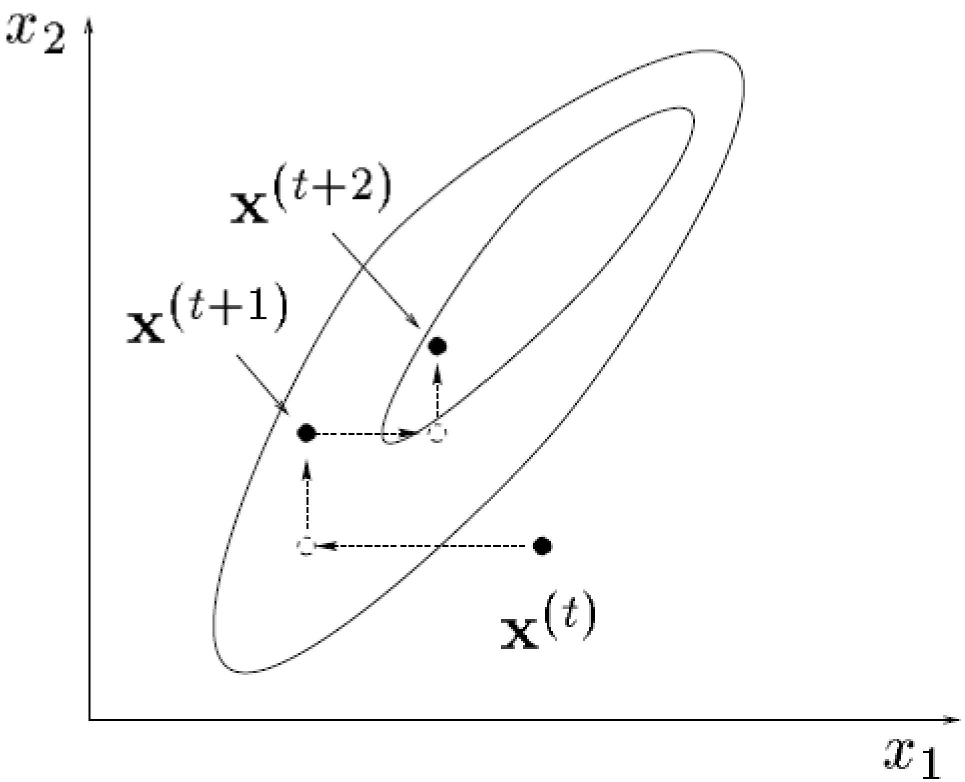
- A sample is drawn from $P(x_2|x_1^{(t+1)})$



this finishes one single iteration.

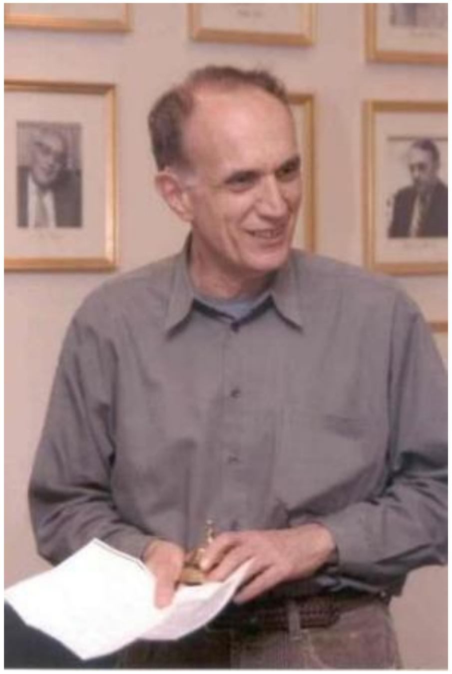
Geometry of Gibbs Sampling

- After a few iterations



Bayes' Theorem in the 21st Century

- This year marks the 250th Anniversary of Bayes' theorem
  - Events at: http://bayesian.org/
- 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

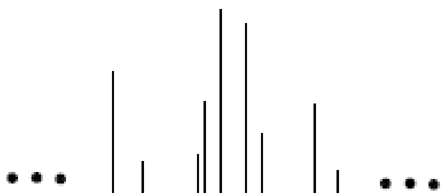
$$p(\mathcal{M}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{M})\pi(\mathcal{M})}{\int p(\mathbf{x}|\mathcal{M})\pi(\mathcal{M})d\mathcal{M}} \propto p(\mathbf{x}|\mathcal{M})\pi(\mathcal{M})$$

Examples:

→ see next slide

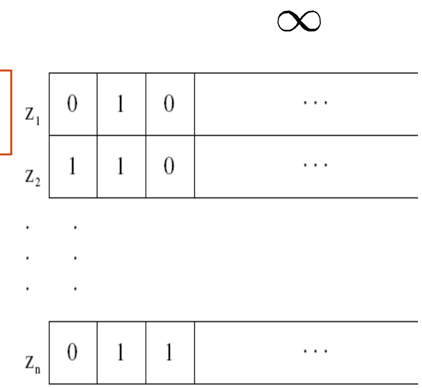
Nonparametric Bayesian Inference

probability measure



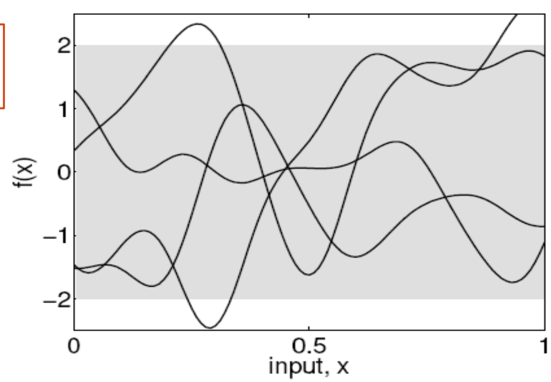
Dirichlet Process Prior [Antoniak, 1974]
+ Multinomial/Gaussian/Softmax likelihood

binary matrix



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

function

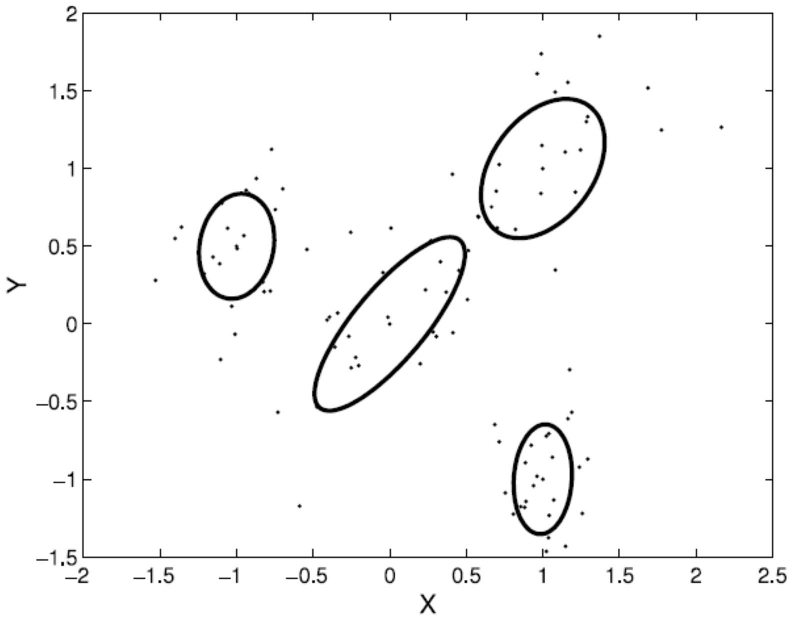


Gaussian Process Prior [Doob, 1944; Rasmussen & Williams, 2006]
+ Gaussian/Sigmoid/Softmax likelihood

Why Be Bayesian Nonparametrics?

Let the data speak for themselves

- 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



Related Tutorials and Materials

- Tutorial talks:
  - Z. Gharamani, ICML 2004. "Bayesian Methods for Machine Learning"
  - M.I. Jordan, NIPS 2005. "Nonparametric Bayesian Methods: Dirichlet Processes, Chinese Restaurant Processes and All That"
  - P. Orbanz, 20009. "Foundations of Nonparametric Bayesian Methods"
  - Y. W. Teh, 2011. "Modern Bayesian Nonparametrics"
  - J. Zhu, ACML 2013. "Recent Advances in Bayesian Methods"

- Tutorial articles:
  - Gershman & Blei. A Tutorial on Bayesian Nonparametric Models. Journal of Mathematical Psychology, 56 (2012) 1-12

Example 2: A Bayesian Ranking Model

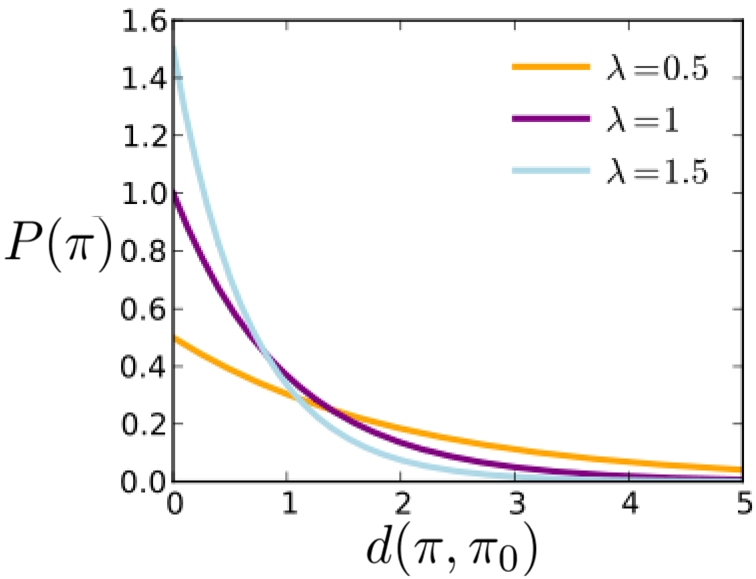
- Rank a set of items, e.g., A, B, C, D
  - A uniform permutation model

$$P([A, C, B, D]) = P([A, D, C, B]) = \cdots = \frac{1}{4!}$$

Example 2: A Bayesian Ranking Model

- Rank a set of items
  - With a preferred list
    - Users offer a concentration center $\pi_0 = [C, B, A, D]$
    - A generalized Mallows' model is defined

$$P(\pi) \propto \exp\Big(-\lambda d(\pi, \pi_0)\Big)$$



[Fligner & Verducci. Distance based Ranking Models. *J. R. Statist. Soc. B*, 1986]

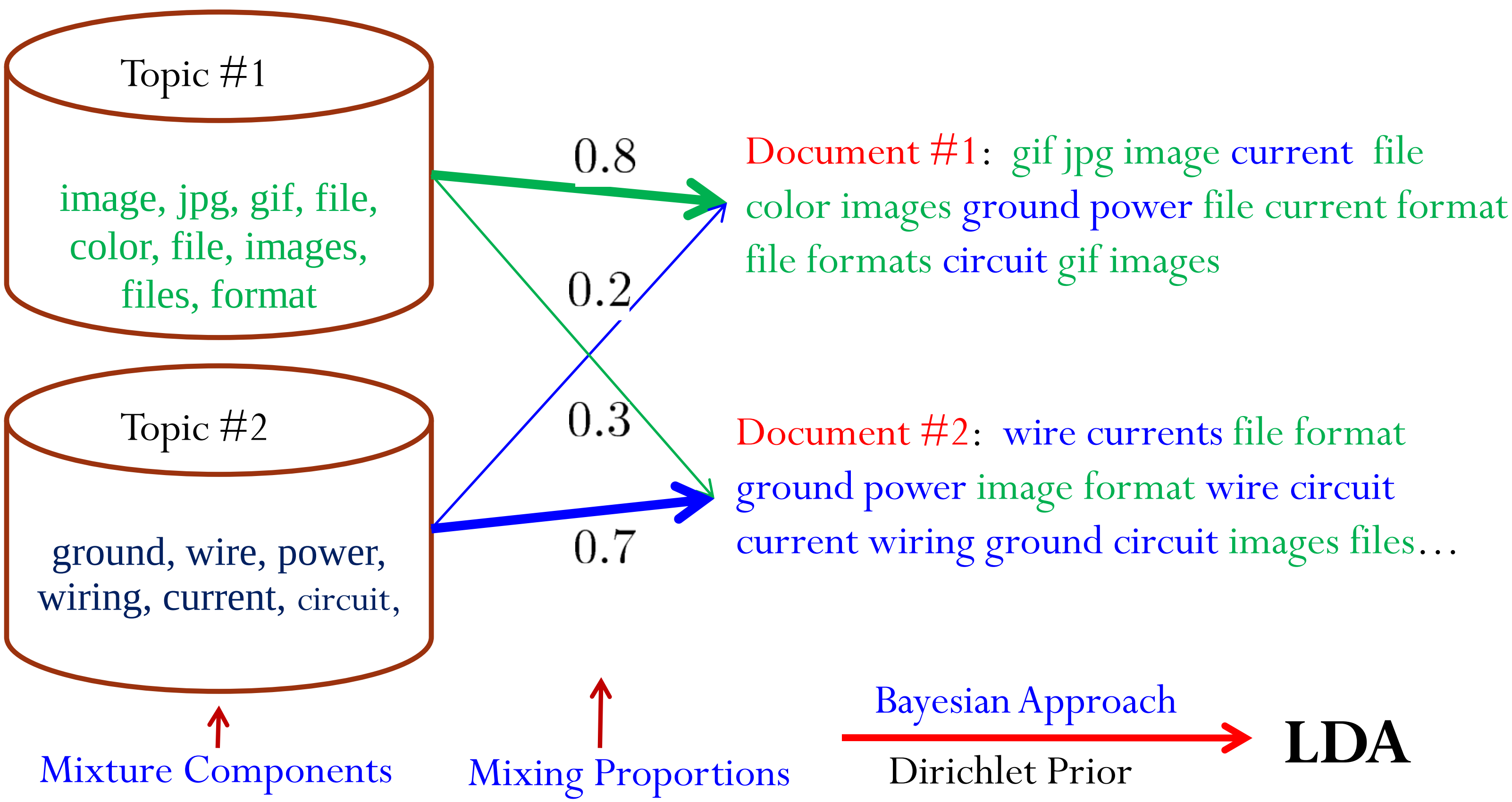
Example 2: A Bayesian Ranking Model

- Rank a set of items
  - Prior knowledge
    - conjugate prior exists for generalized Mallows' models (a member of exponential family)
  - Bayesian updates can be done with Bayes' rule
  - Can be incorporated into a hierarchical Bayesian model, e.g., topic models

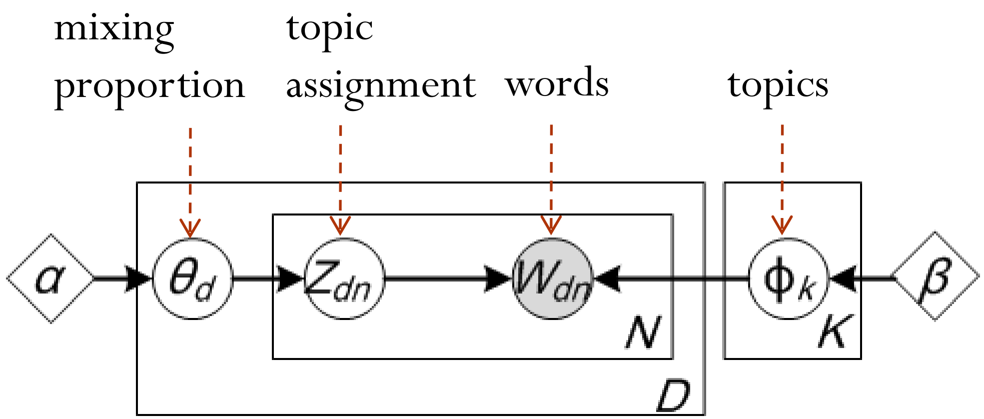
[Chen, Branavan, et al., Global models of document structure using latent permutations. *ACL*, 2009]

Example 3: 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



Example 3: Bayesian Inference for LDA



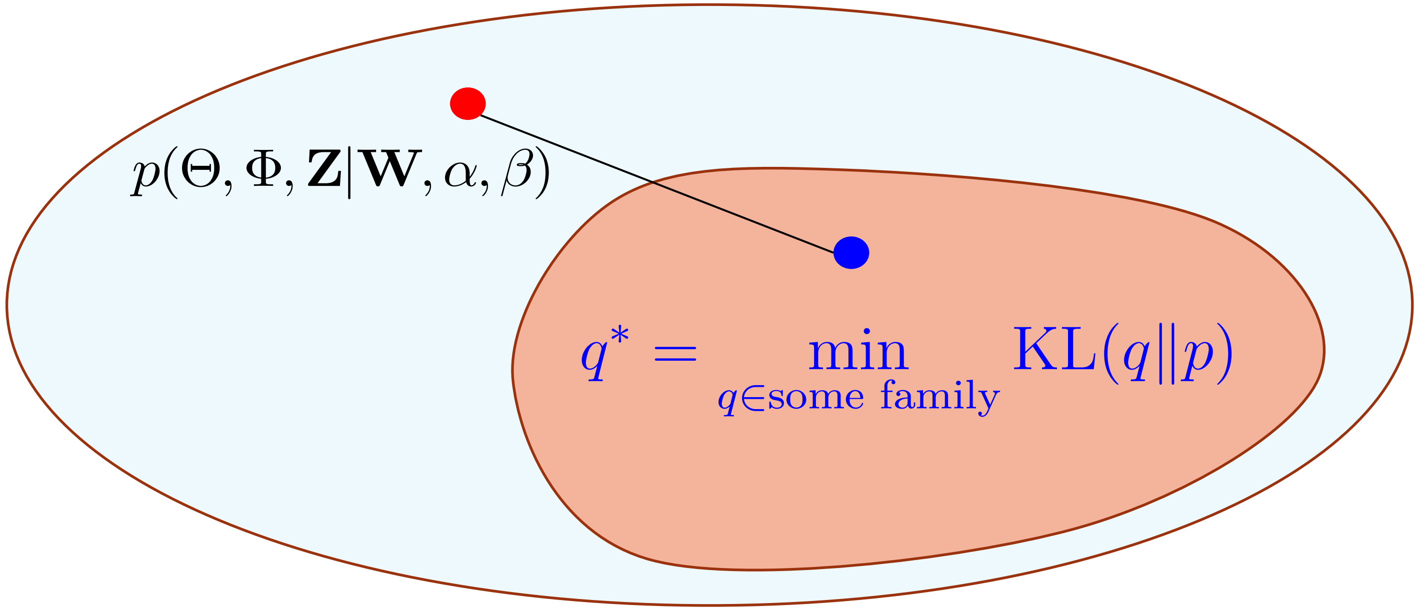
$$p(\Theta, \Phi, \mathbf{Z}, \mathbf{W}|\alpha, \beta) = \prod_{k=1}^{K} p(\Phi_k|\beta) \prod_{d=1}^{D} p(\theta_d|\alpha) \left(\prod_{n=1}^{N} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn}, \Phi) \right)$$

- Given a set of documents, infer the posterior distribution

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

Example 3: Approximate Inference

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



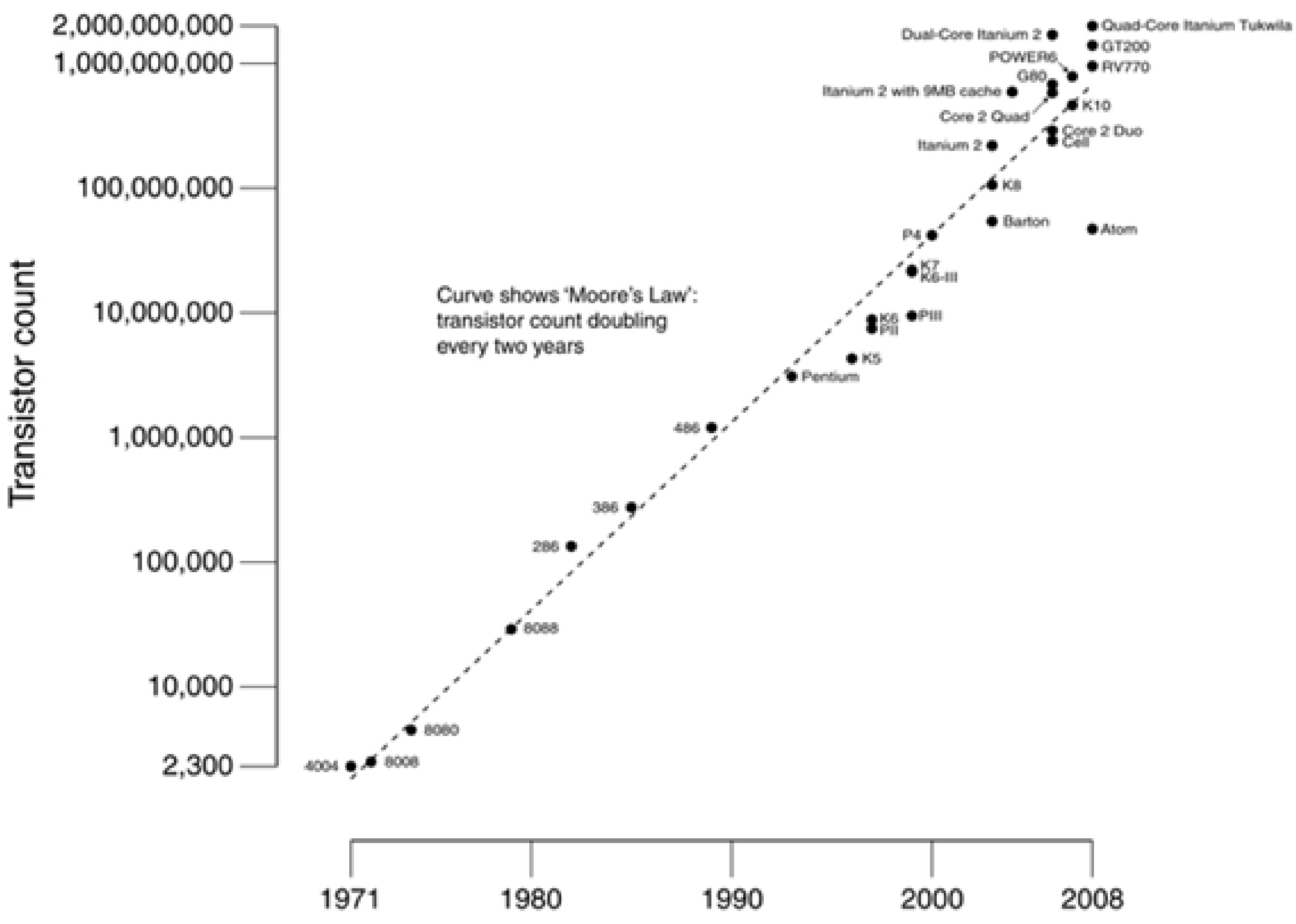
- Monte Carlo Markov Chains (Griffiths & Steyvers, 2004)
  - Collapsed Gibbs samplers iteratively draw samples from the local conditionals

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

Bayesian Methods and Big Data

Overfitting in Big Data

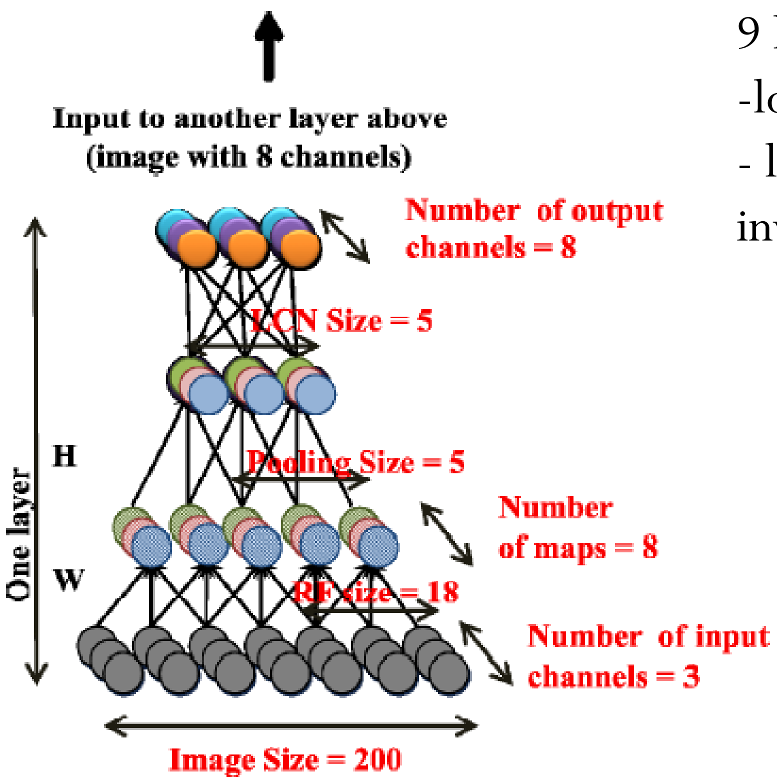
- "with more data overfitting is becoming less of a concern"?



Overfitting in Big Data

“Big Model + Big Data + Big/Super Cluster”

Big Learning



9 layers sparse autoencoder with:

-local receptive fields to scale up;

- local L2 pooling and local contrast normalization for invariant features

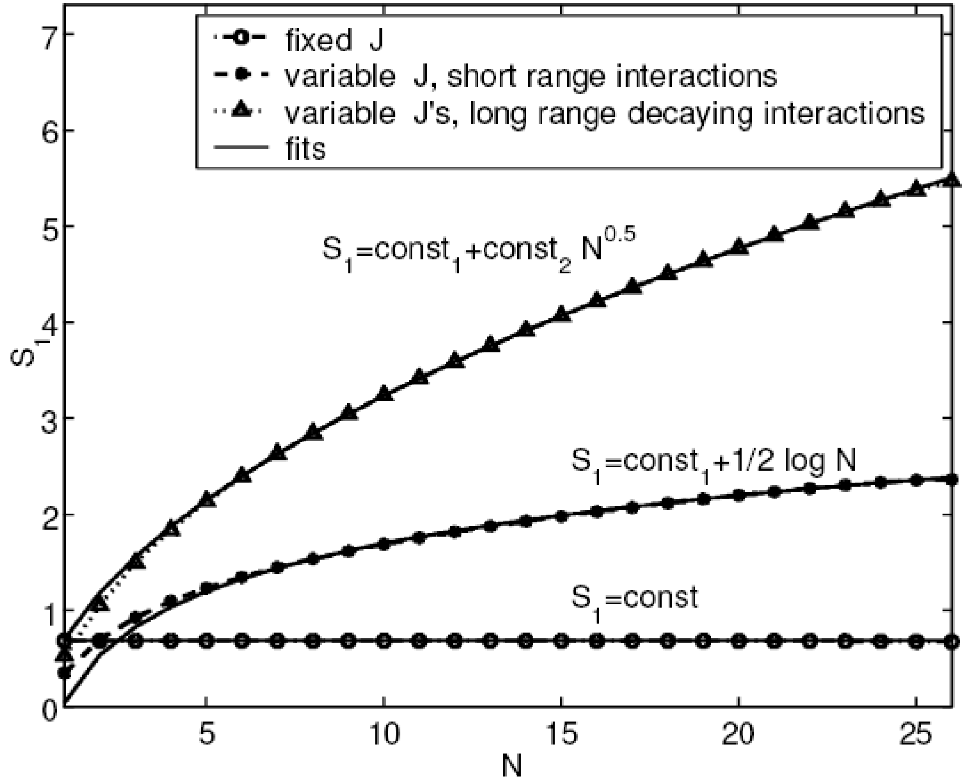
  - 1B parameters (connections)
  - 10M 200x200 images
  - train with 1K machines (16K cores) for 3 days

-able to build high-level concepts, e.g., cat faces and human bodies

-15.8% accuracy in recognizing 22K objects (70% relative improvements)

Overfitting in Big Data

- **Predictive information** grows slower than the amount of Shannon entropy (Bialek et al., 2001)



Overfitting in Big Data

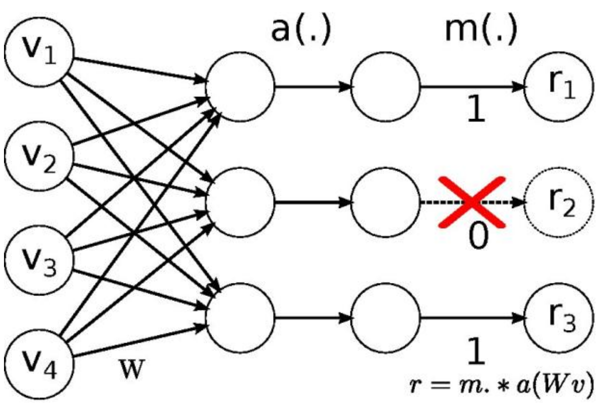
- **Predictive information** grows slower than the amount of Shannon entropy (Bialek et al., 2001)

Model capacity grows faster than the amount of predictive information!

Overfitting in Big Data

- Surprisingly, regularization to prevent overfitting is *increasingly important*, rather than increasingly irrelevant!

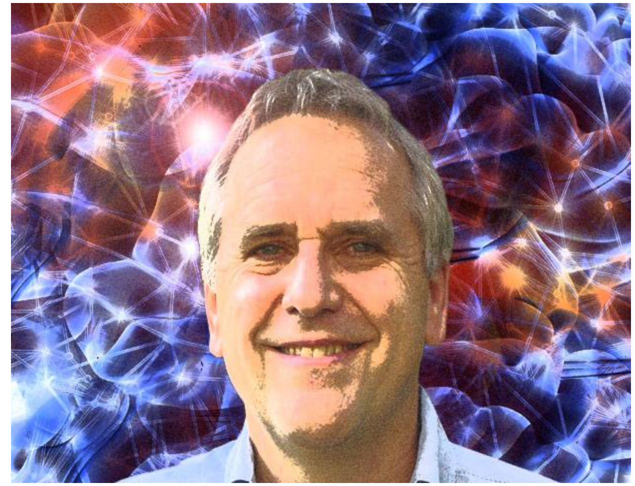
- Increasing research attention, e.g., dropout training (Hinton, 2012)



- More theoretical understanding and extensions
  - MCF (van der Maaten et al., 2013); Logistic-loss (Wager et al., 2013); Dropout SVM (Chen, Zhu et al., 2014)

Why Big Data could be a Big Fail?

Michael I. Jordan
UC Berkeley
Pehong Chen Distinguished Professor
NAS, NAE, NAAS Fellow
ACM, IEEE, IMS, ASA, AAAI Fellow



- When you have large amounts of data, your appetite for hypotheses tends to get even larger
- If it's growing faster than the statistical strength of the data, then many of your inferences are likely to be false. They are likely to be white noise.
- Too much hype: "The whole big-data thing came and went. It died. It was wrong"

Therefore …

- Computationally efficient Bayesian models are becoming increasingly relevant in Big data era
  - **Relevant**: high capacity models need a protection
  - **Efficient**: need to deal with large data volumes

Challenges of Bayesian Methods

Building an *Automated* Statistician

- **Theory**
  - Improve the classic Bayes theorem
- **Modeling**
  - scientific and engineering data
  - rich side information
- **Inference/learning**
  - discriminative learning
  - large-scale inference algorithms for Big Data
- **Applications**
  - social media, NLP, computer vision

Theory

Regularized Bayesian Inference

Example 2 Revisit: A Bayesian Ranking Model

- Arrange a set of invited talks
  - Side constraints
    - Mike Jordan can only spend 2 days at ICML
    - Eric Horvitz can only spend 1 day at ICML
    - 院士X必须放在第一天
    - Vision 排在 learning前面
    - ……

- How can we consider them in Bayesian methods?

Domain Knowledge

- Represented in logic form:

seed-rules:

$$\forall i (w(i) = \text{“monkey”}) \rightarrow (z(i) = T)$$

cannot-link rules:

$$\forall i \forall j (w(i) = \text{“monkey”}) \bigwedge (w(j) = \text{“apple”}) \rightarrow z(i) \neq z(j)$$

must-link rules:

$$\forall i \forall j (w(i) = \text{“monkey”}) \bigwedge (w(j) = \text{“gorilla”}) \rightarrow z(i) = z(j)$$

- *How to incorporate such information in Bayesian inference?*

Regularized Bayesian Inference?

likelihood model

posterior

prior

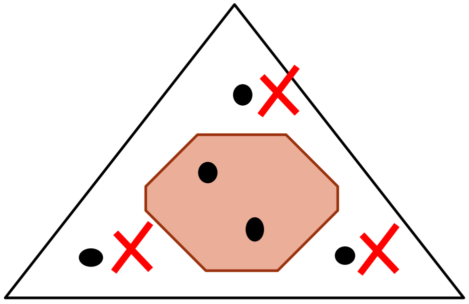
$$p(\mathcal{M}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{M})\pi(\mathcal{M})}{\int p(\mathbf{x}|\mathcal{M})\pi(\mathcal{M})d\mathcal{M}}$$

- **How to consider side constraints?**

Not obvious!

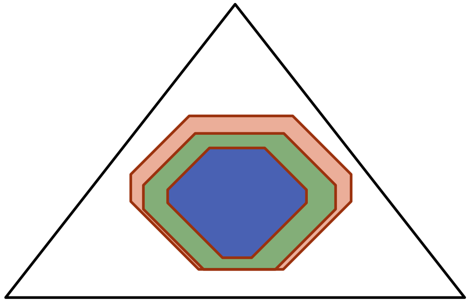
hard constraints

(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

Tsinghua University

Regularized Bayesian Inference

Constraints can encode rich structures/knowledge

◈ Bayesian inference with posterior regularization:

'unconstrained' equivalence:

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

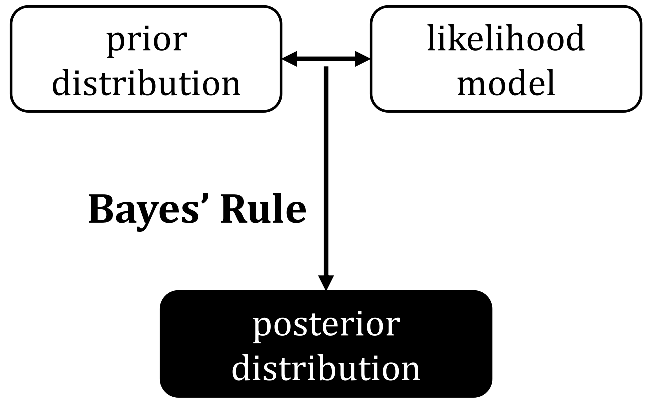
posterior regularization

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

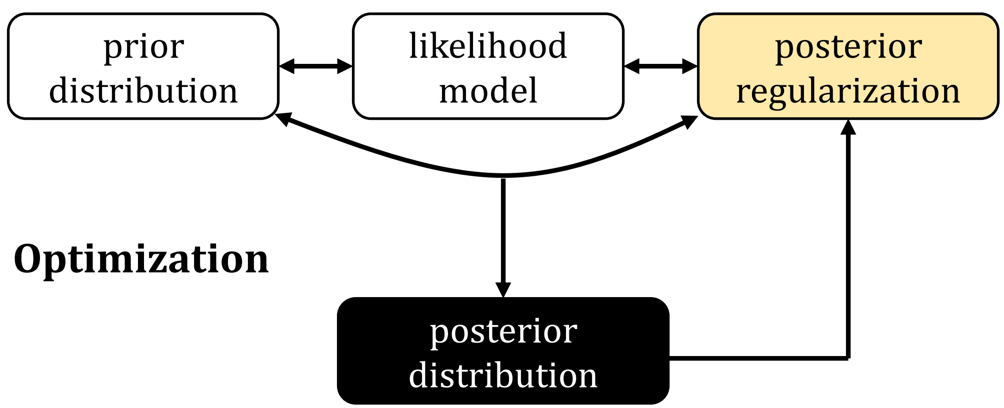
[Zhu, Chen, & Xing, JMLR, 2014]

A High-Level Comparison

Bayes:



RegBayes:



More Properties

- **Representation Theorem:**
  - the optimum distribution is:

$$\hat{q}_{\hat{\phi}}(\mathcal{M}) = p(\mathcal{M}, \mathcal{D}) \exp\left(\langle \hat{\phi}, \psi(\mathcal{M}, \mathcal{D}) \rangle - \Lambda_{\hat{\phi}}\right)$$

  - where $\hat{\phi}$ is the solution of the convex dual problem

- **Putting constraints on priors is a special case**
  - constraints on priors are special cases of posterior regularization

- **RegBayes is more flexible than Bayes' rule**
  - exist some RegBayes distribution: no implicit prior and likelihood that give back it by Bayes' rule

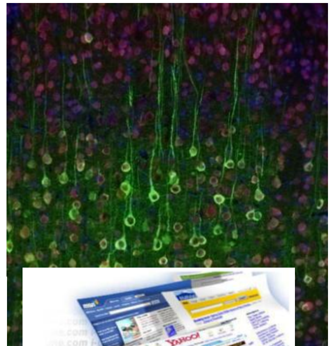
[Zhu, Chen, & Xing, JMLR, 2014]

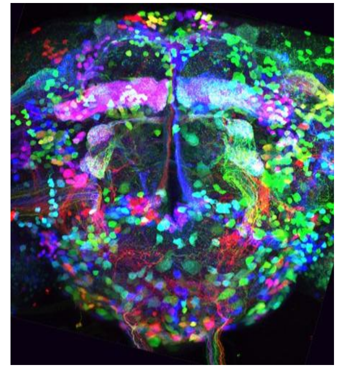
Ways to Derive Posterior Regularization

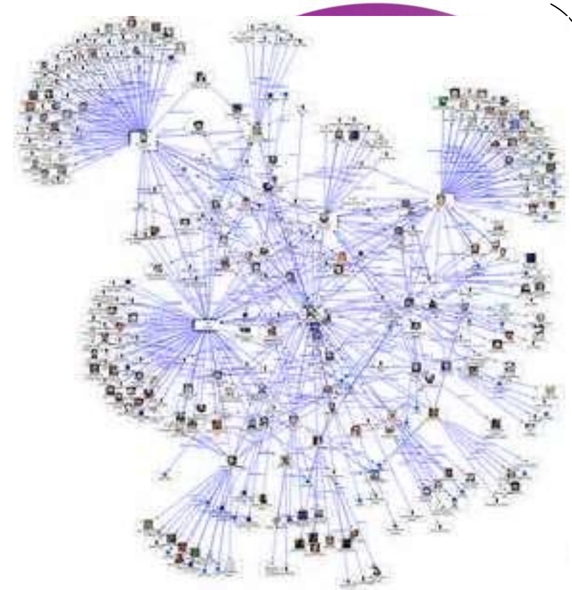
- **From learning objectives**
  - Performance of posterior distribution can be evaluated when applying it to a learning task
  - Learning objective can be formulated as Pos. Reg.
- **From domain knowledge**
  - Elicit expert knowledge
  - E.g., first-order logic rules
- **Others …**
  - E.g., decision making, cognitive constraints, etc.

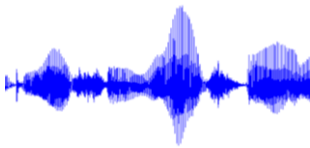
Modeling + Algorithms

Adaptive, Discriminative, Scalable Representation Learning





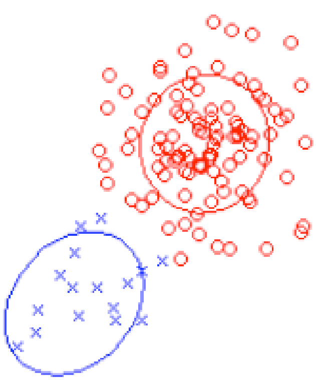


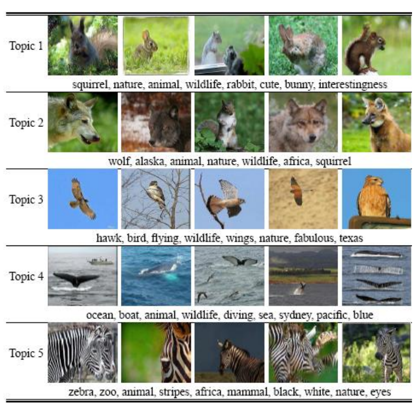


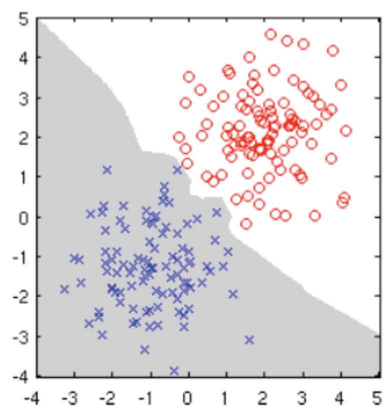
Input Data

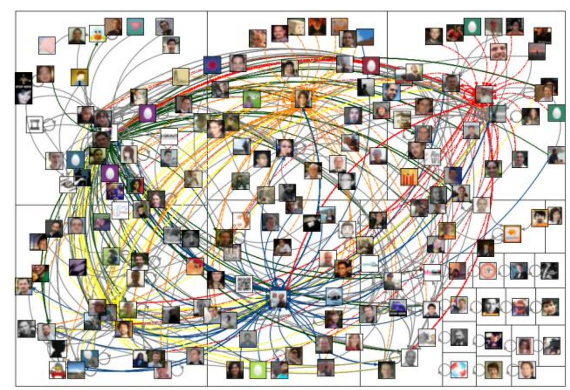
Ideal paradigm that computers help solve big data problems

Inference, Decision, Reasoning



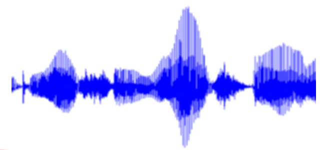




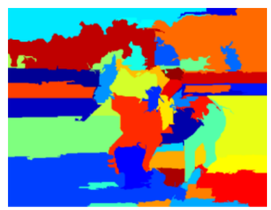


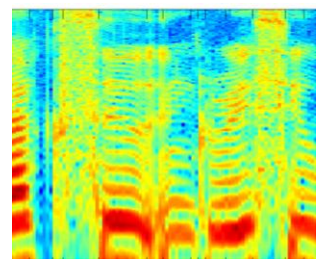
A Conventional Data Analysis Pipeline

Input Data

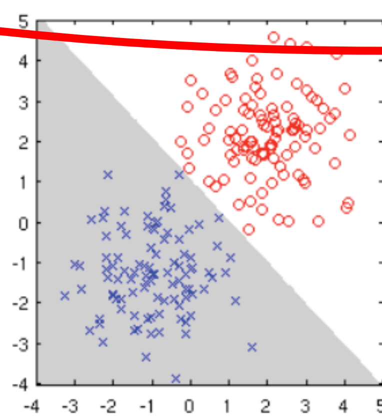


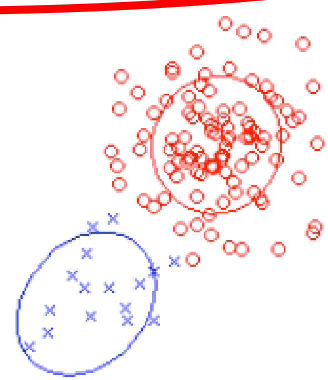
Feature Representation

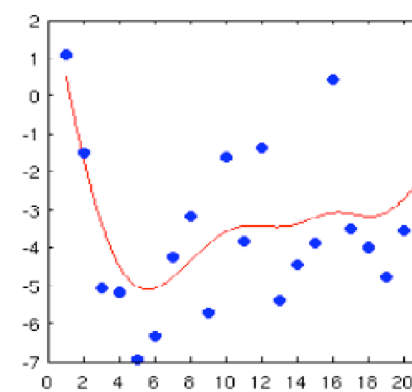




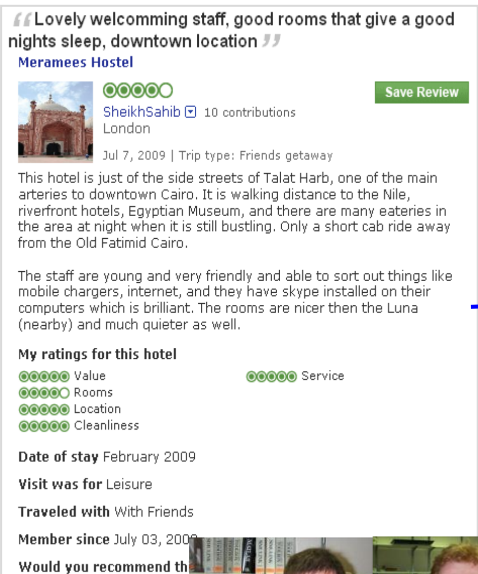
Computer Algorithms







Representation Learning

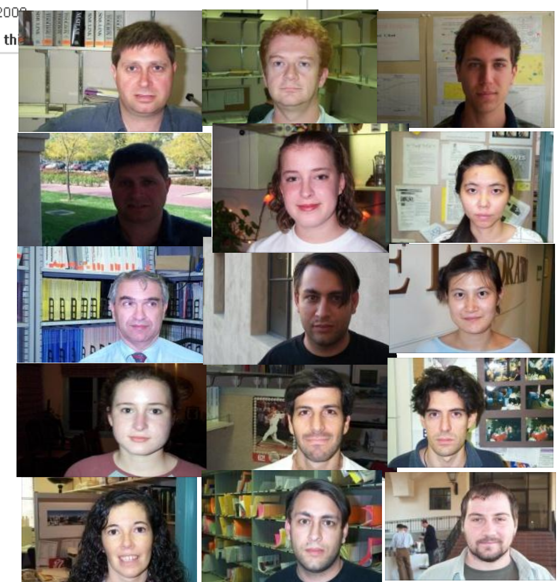


Learning Algorithms

E.g., Topic Models

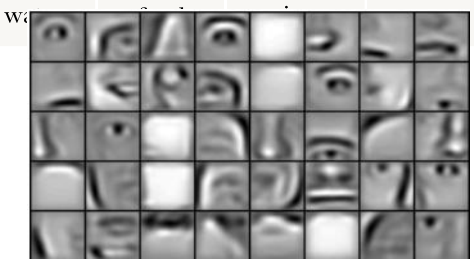
Axis's of a semantic representation space:

T1	T2	T3	T4	T5	T6	T7
told	place	hotel	hotel	beach	beach	great
dirty	hotel	food	area	pool	resort	good
room	room	bar	staff	resort	pool	nice
front	days	day	pool	food	ocean	lovely
asked	time	pool	breakfast	island	island	beautiful
hotel	day	time	day	kids	kids	excellent
bad	night	service	view	trip	good	wonderful
small	people	holiday	location	service	restaurants	comfortable
worst	stay	room	service	day	enjoyed	beach
poor	water	people	walk	staff	loved	friendly
called	rooms	night	time	time	trip	fresh
rude	food	wa				amazing



Learning Algorithms

E.g., Deep Networks



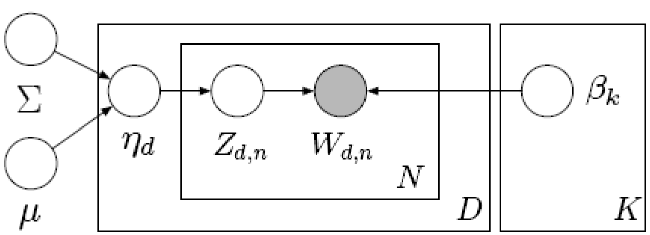
[Figures from (Lee et al., ICML2009)]

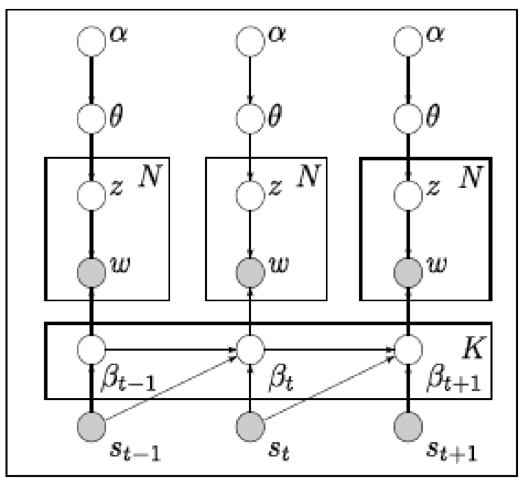
Some Key Challenges

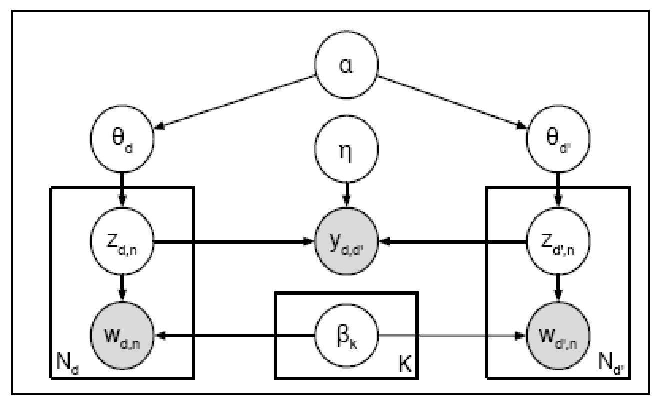
- Discriminative Ability
  - Are the representations good at solving a task, e.g., distinguishing different concepts?
  - Can they generalize well to unseen data?
  - Can the learning process effectively incorporates domain knowledge?
- Model Complexity
  - How many dimensions are sufficient to fit a given data set?
  - Can the models adapt when environments change?
- Sparsity/Interpretability
  - Are the representations compact or easy to interpret?
- Scalability
  - Can the algorithms scale up to Big Data applications?

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.



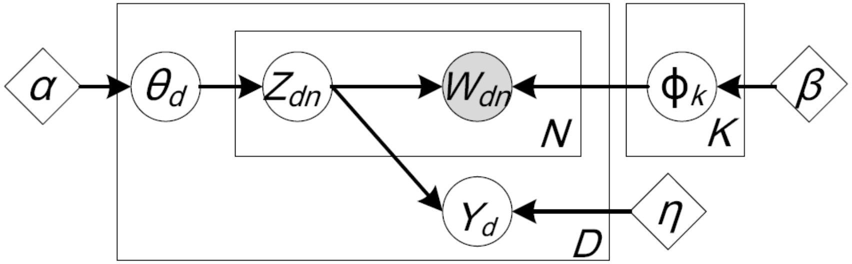




- Tutorials were provide by D. Blei at ICML, SIGKDD, etc. (http://www.cs.princeton.edu/~blei/topicmodeling.html)

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 $y_d \in \{0, 1\}$

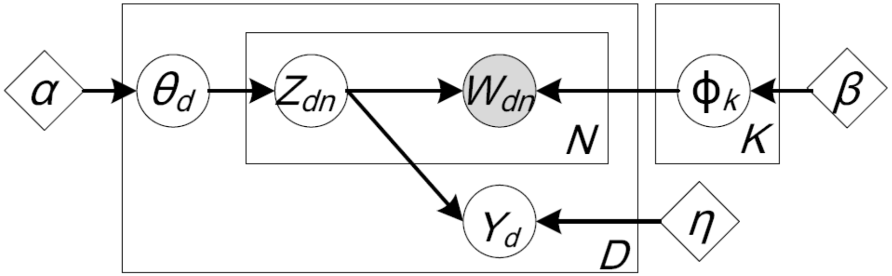
$$p(y_d | \mathbf{z}_d, \eta) = \frac{\{\exp(\eta^\top \bar{\mathbf{z}}_d)\}^{y_d}}{1 + \exp(\eta^\top \bar{\mathbf{z}}_d)} \qquad \bar{z}_k = \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}(z_n^k = 1)$$

- 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

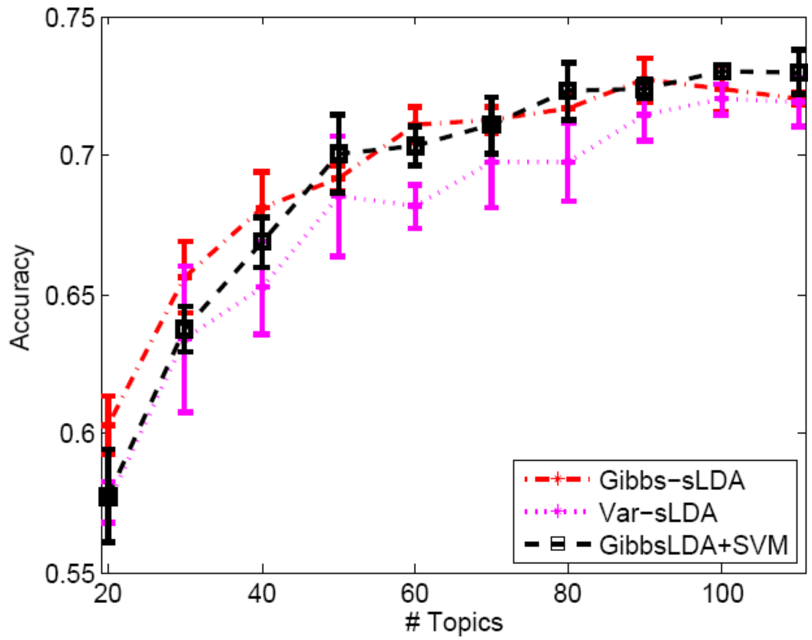
- A document has hundreds of words
- … but only one class label



- Imbalanced likelihood combination

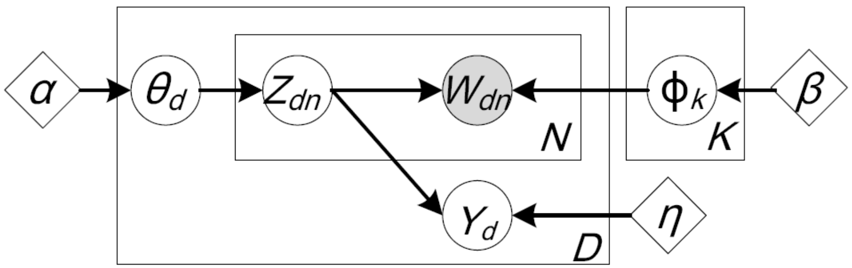
$$p(\mathbf{y}, \mathbf{W}|\mathbf{Z}, \Phi, \eta) = p(\mathbf{y}|\mathbf{Z}, \eta)p(\mathbf{W}|\mathbf{Z}, \Phi)$$

- Too weak influence from supervision



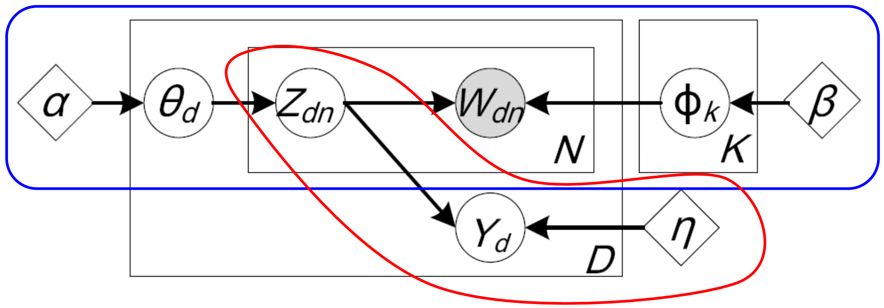
(Halpern et al., ICML 2012; Zhu et al., ACL 2013)

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?

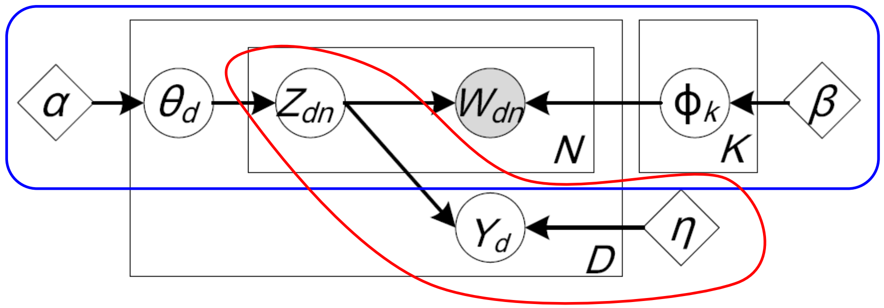
MedLDA:
Max-margin Supervised Topic Models



- 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 ($y_d \in \{-1, 1\}$)

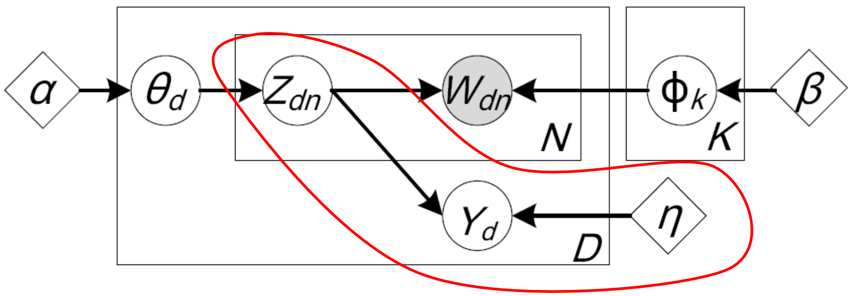
$$\hat{y} = \mathrm{sign} f(\mathbf{w}) = \mathrm{sign}\mathbb{E}_q[f(\eta, \mathbf{z}; \mathbf{w})]$$

  - 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

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

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

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

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

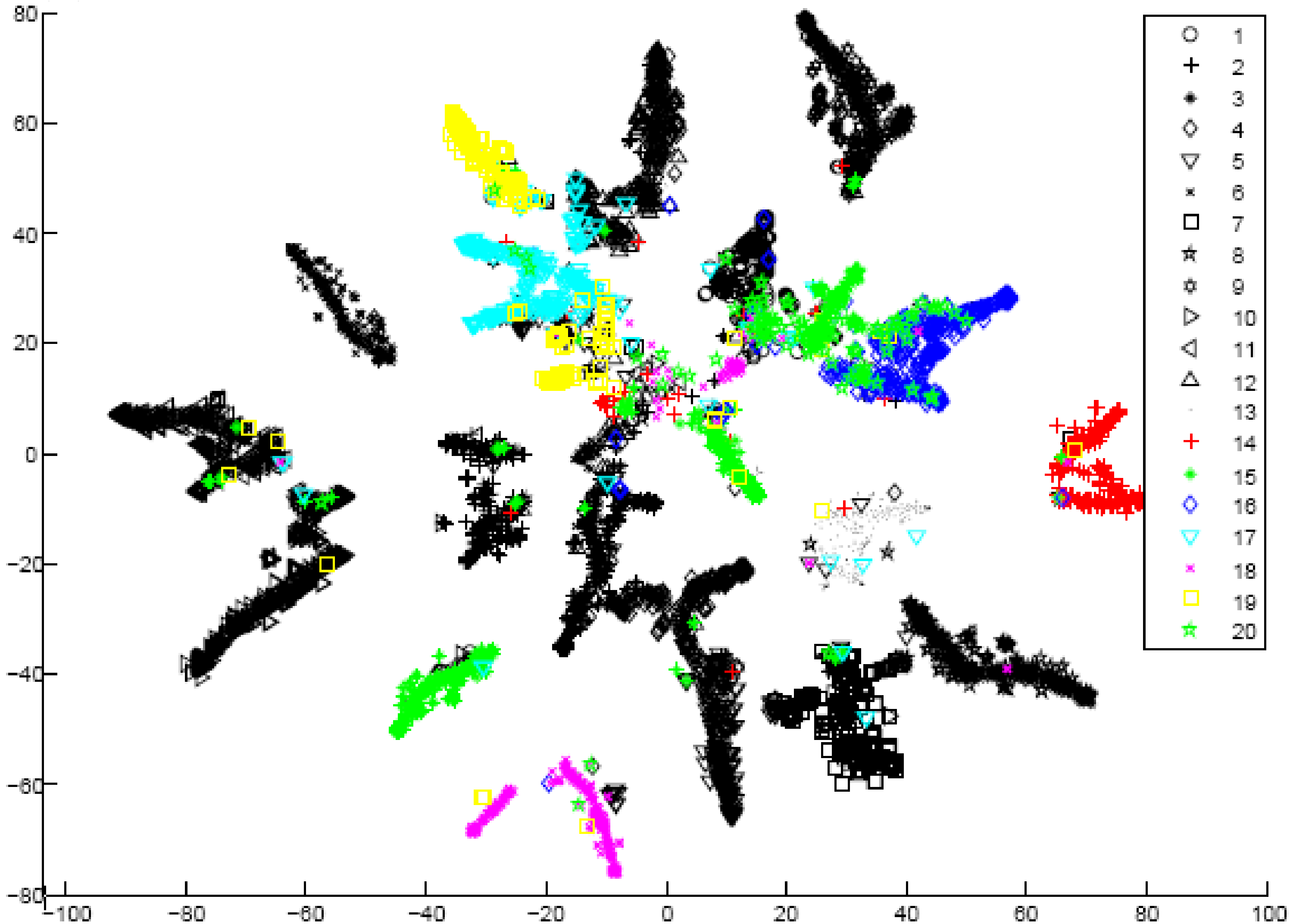
A SVM problem with a normal prior

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

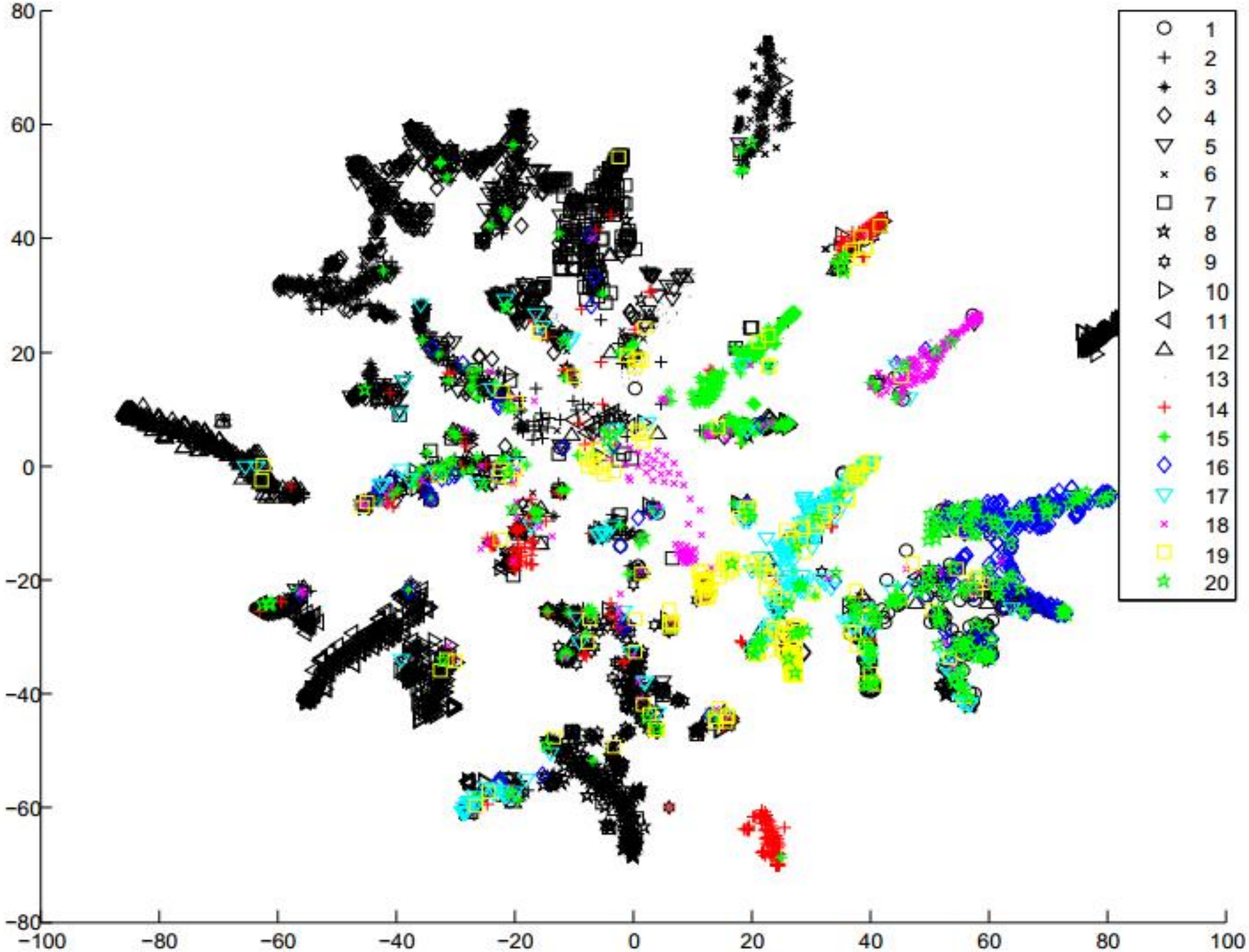
$$\forall d,\ \text{s.t.}:\ y_d\mathbb{E}_q[\eta]^\top\mathbb{E}_q[\bar{\mathbf{z}}_d] \geq 1-\xi_d.$$

Variational approximation or Monte Carlo methods

Empirical Results on 20Newsgroups

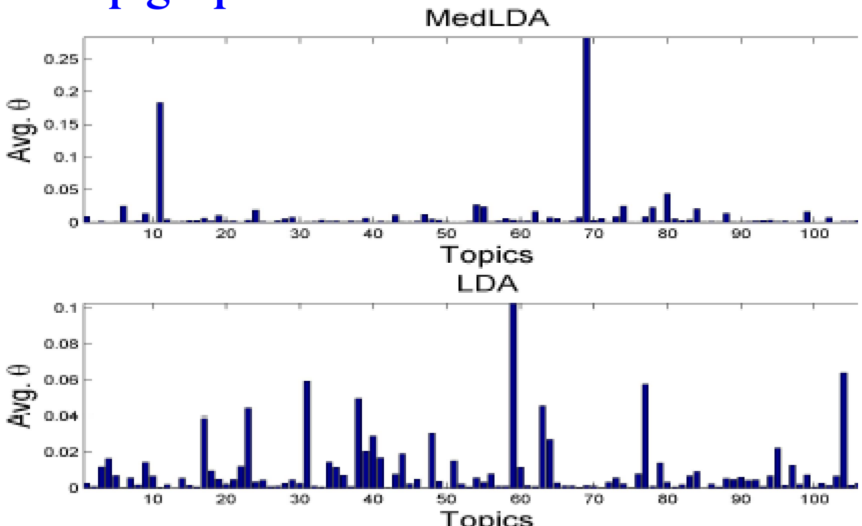


Empirical Results on 20Newsgroups



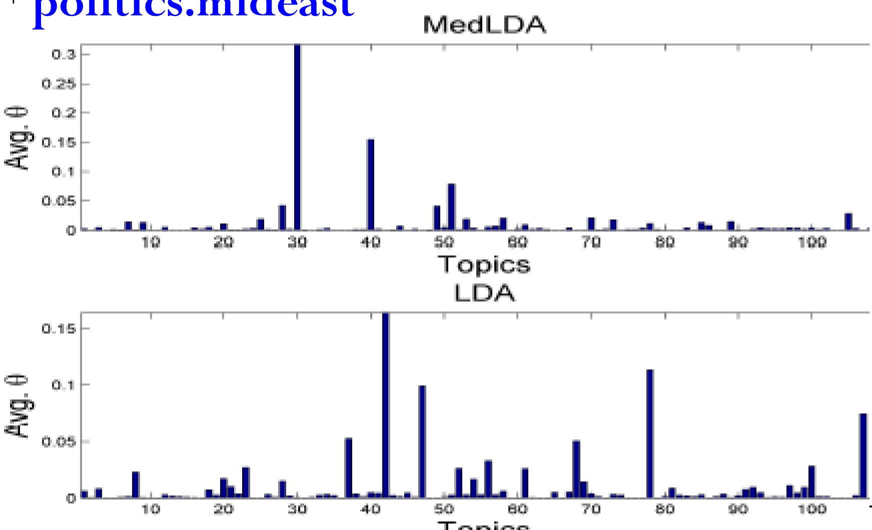
Sparser and More Salient Representations

comp.graphics



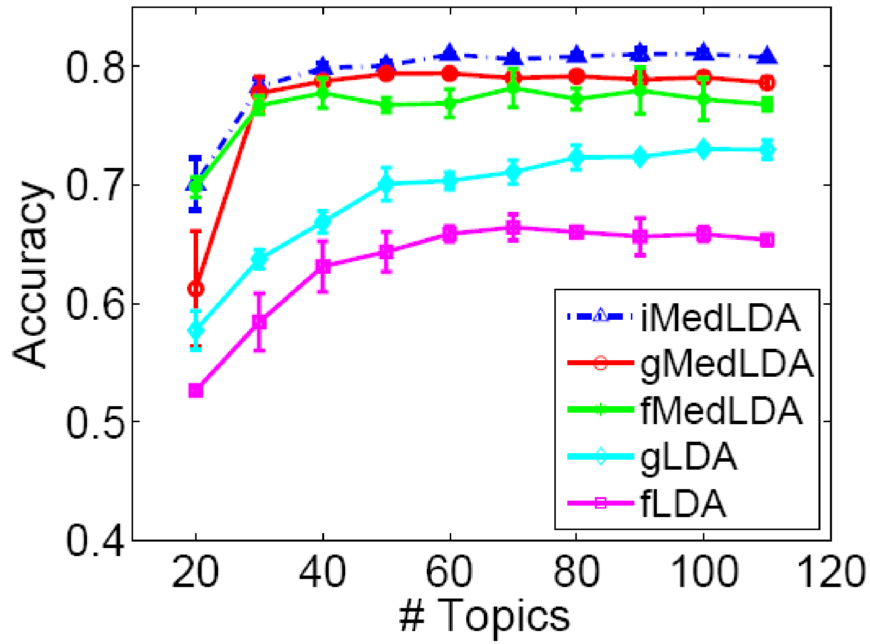
MedLDA			LDA		
T 69	T 11	T 80	T 59	T 104	T 31
image	graphics	db	image	ftp	card
jpeg	image	key	jpeg	pub	monitor
gif	data	chip	color	graphics	dos
file	ftp	encryption	file	mail	video
color	software	clipper	gif	version	apple
files	pub	system	images	tar	windows
bit	mail	government	format	file	drivers
images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

politics.mideast



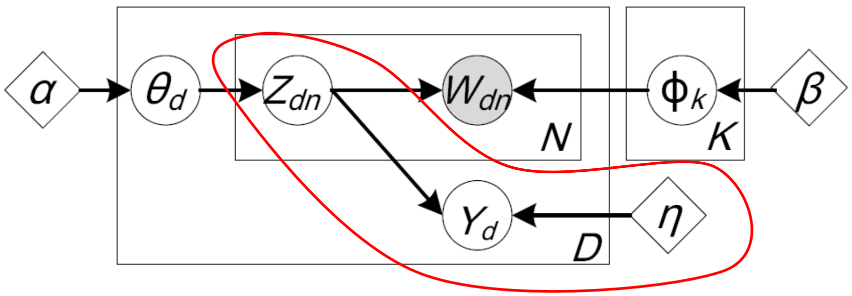
T 30	T 40	T 51	T 42	T 78	T 47
israel	turkish	israel	israel	jews	armenian
israeli	armenian	lebanese	israeli	jewish	turkish
jews	armenians	israeli	peace	israel	armenians
arab	armenia	lebanon	writes	israeli	armenia
writes	people	people	article	arab	turks
people	turks	attacks	arab	people	genocide
article	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

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

  - where

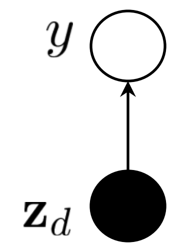
$$f(\eta, \mathbf{z}; \mathbf{w}) = \eta^\top \bar{\mathbf{z}} \qquad \bar{z}_k = \frac{1}{N}\sum_{n=1}^{N} \mathbb{I}(z_n^k = 1)$$

(Zhu, Chen, Perkins, Zhang, JMLR 2014)

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

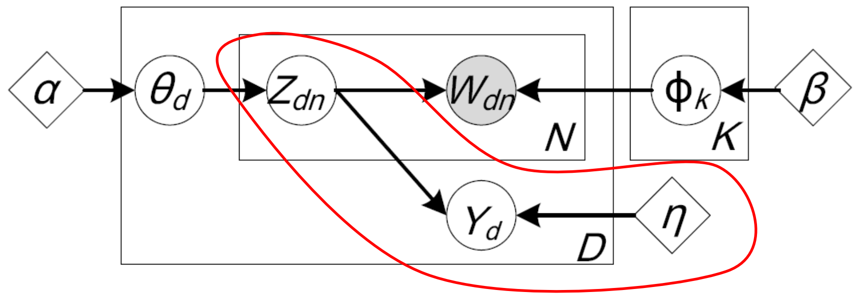
$$\hat{R}(\eta, Z) = \sum_{d=1}^{D} \mathbb{I}(\hat{y}_d|_{\eta,\mathbf{z}_d} \neq y_d)$$

  - 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

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

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

- 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

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

- 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

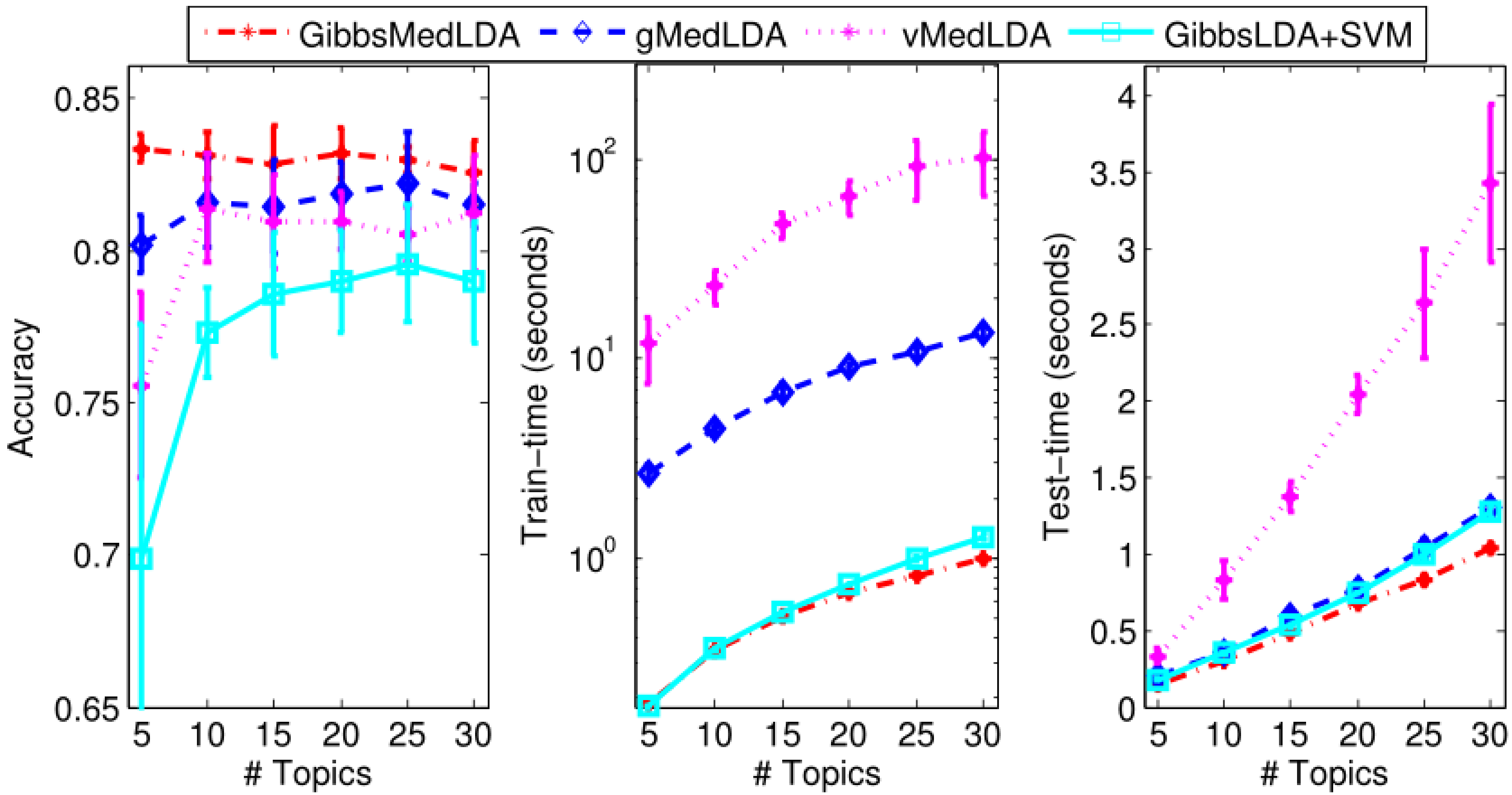
- Per-iteration time complexity

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

  - N_{total} the total number of words

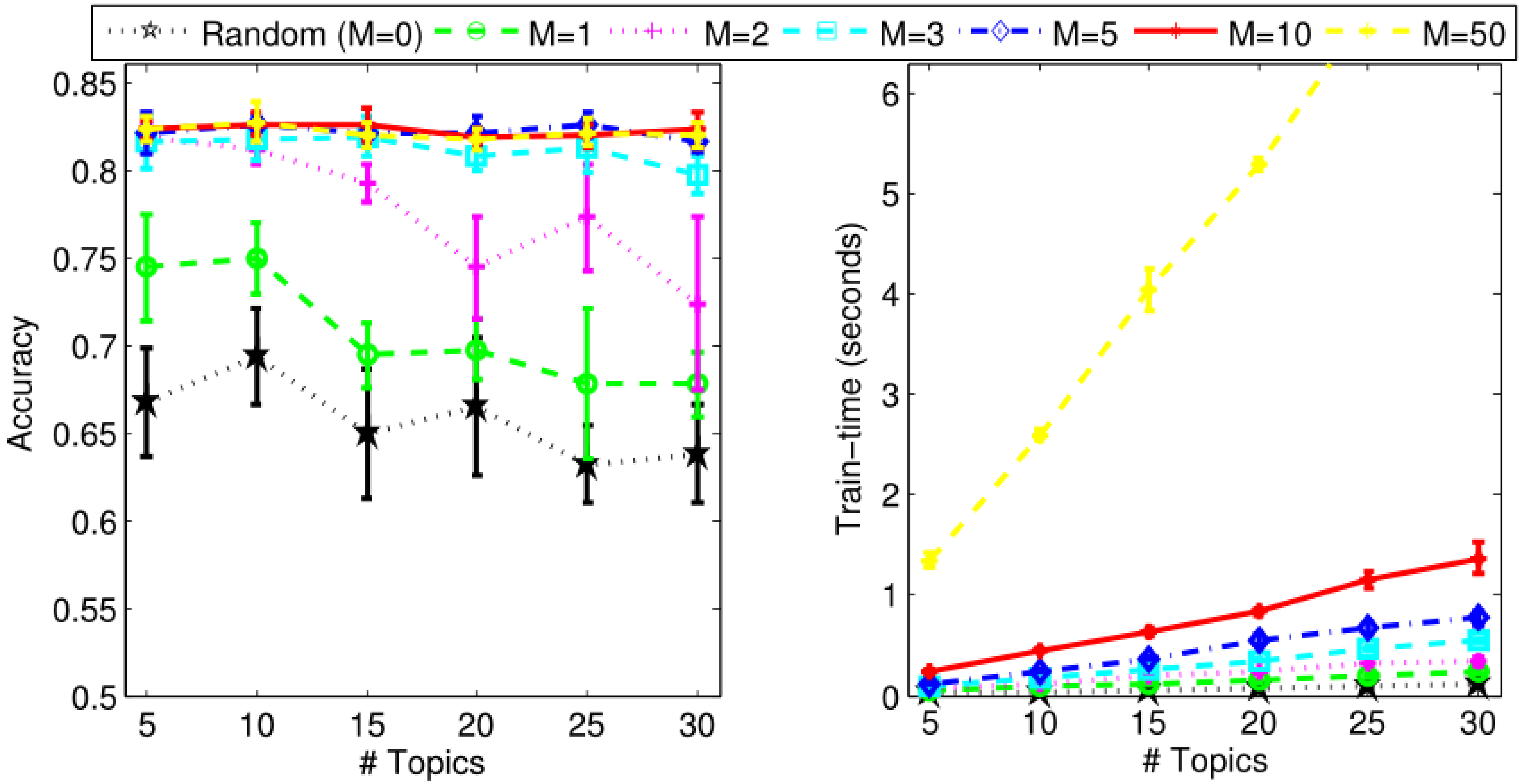
Experiments

- 20Newsgroups binary classification



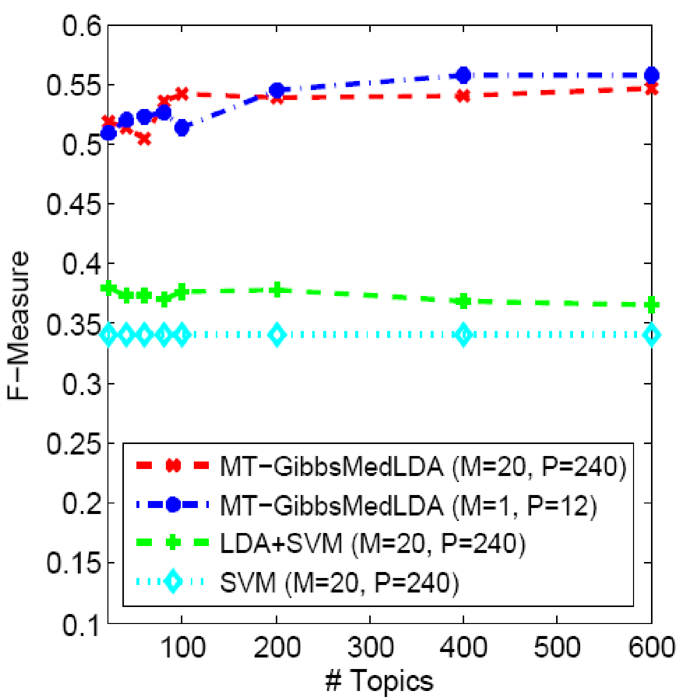
Experiments

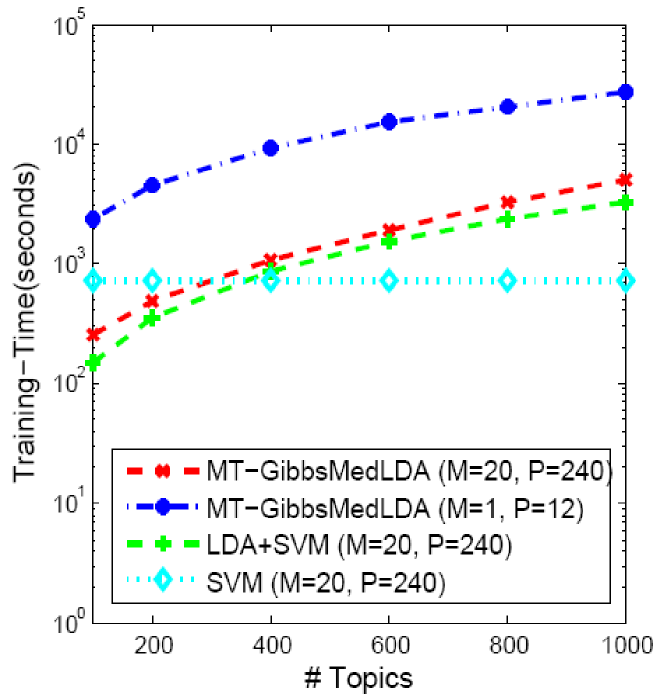
- **Sensitivity to burn-in**: binary classification



Experiments

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



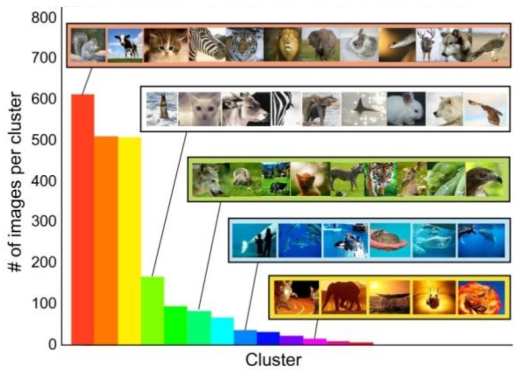


- 20 machines;
- 240 CPU cores

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

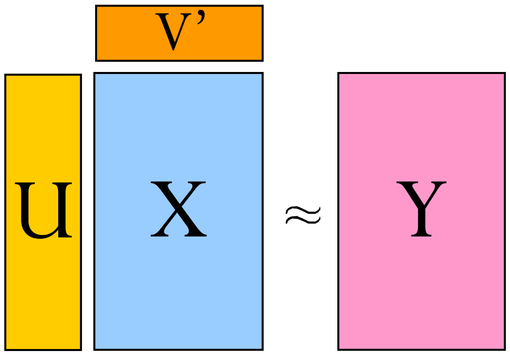
[Zhu, Zheng, Zhou, & Zhang, KDD2013]

RegBayes with Max-margin Posterior Regularization

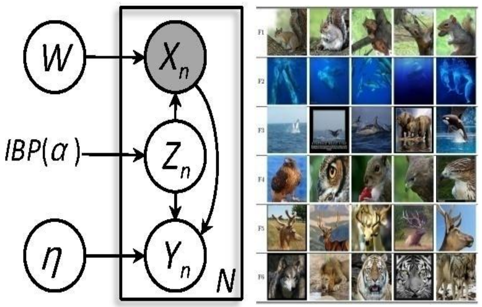


Infinite SVMs
(Zhu et al., ICML'11)

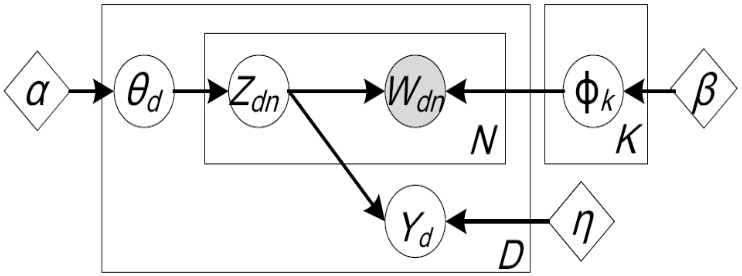
Nonparametric Max-margin Relational Models for Social Link Prediction
(Zhu, ICML'12)



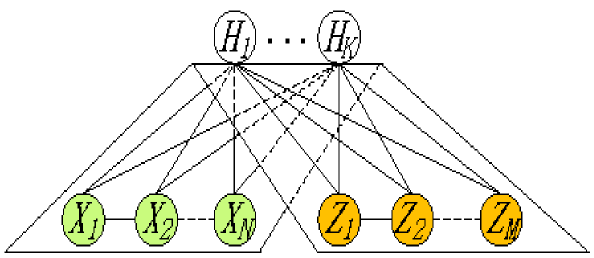
Nonparametric Max-margin Matrix Factorization
(Xu, Zhu, & Zhang, NIPS'12, ICML'13)



Infinite Latent SVMs
(Zhu, et al., JMLR'14)



Max-margin Topics and Fast Inference
(Zhu, et al., JMLR'12; Zhu et al., JMLR'14)

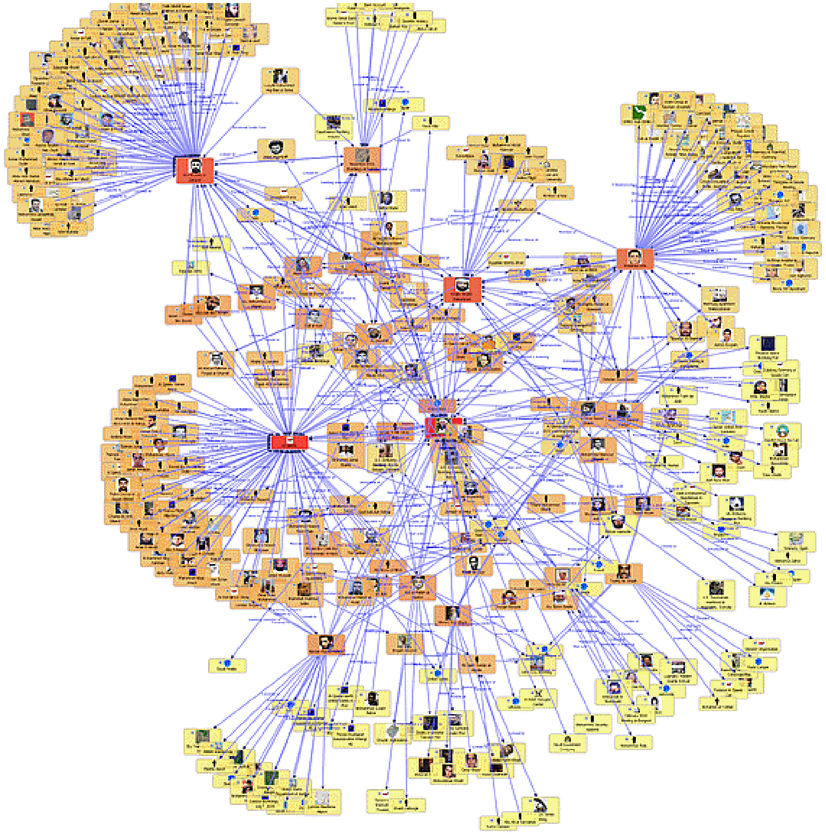


Multimodal Representation Learning
(Chen, Zhu, et al, PAMI'12)

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

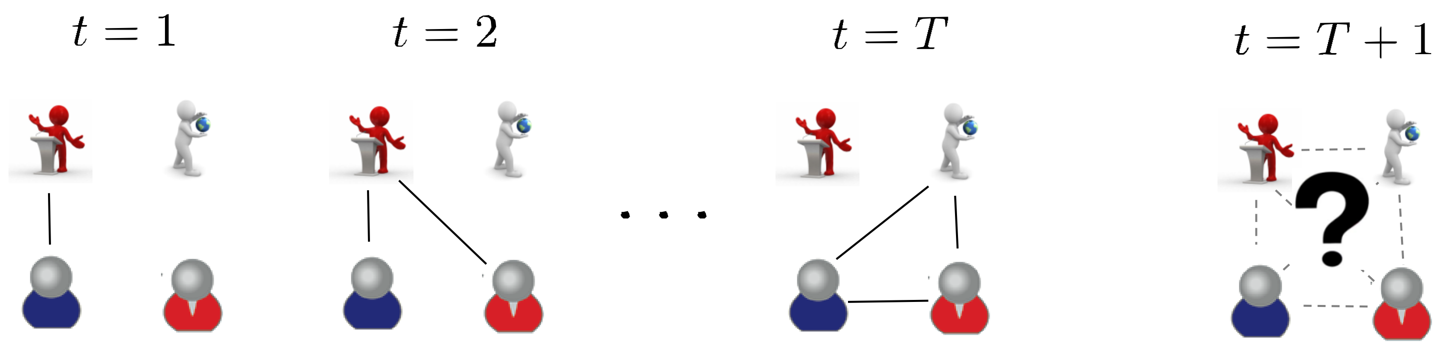
Link Prediction

◆ But network structures are usually unclear, unobserved, or corrupted with noise

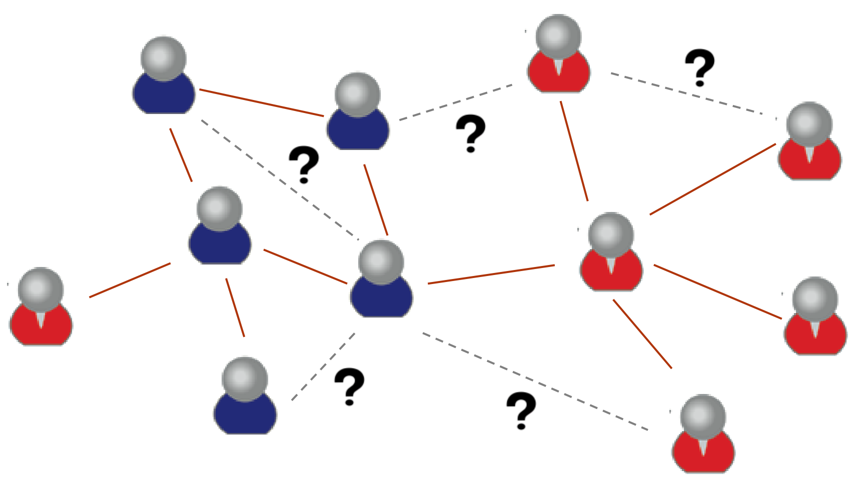


Link Prediction – task

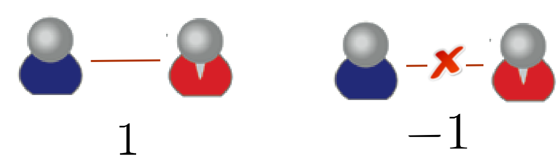
- Dynamic networks



- Static networks



We treat it as a supervised learning task with 1/-1 labels



Link Prediction as Supervised Learning

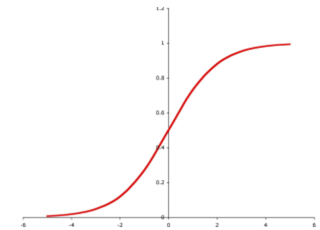
- Building classifiers with manually designed features from networks
  - Topological features
    - Shortest distance, number of common neighbors, Jaccard's coefficient, etc.
  - Attributes about individual entities
    - E.g., the papers an authors has published
  - Proximity features
    - E.g., two authors are close, if their research work evolves around a large set of identical keywords
  - …

[Hasan et al., 2006]

Link Prediction as Supervised Learning

- Latent feature relational models
  - Each entity is associated with a point $\mu_i \in \mathbb{R}^K$ in a latent feature space
  - Then, a link likelihood is generally defined

$$p(Y_{ij} = 1 | X_{ij}, \mu_i, \mu_j) = \Phi\left(\mu + \boldsymbol{\beta}^\top X_{ij} + \psi(\mu_i, \mu_j)\right)$$



Latent distance model (Hoff et al., 2002)

$$\psi(\mu_i, \mu_j) = -d(\mu_i, \mu_j)$$
$$= -\|\mu_i - \mu_j\|$$

generalize

Latent eigenmodel (Hoff, 2007)

$$\psi(\mu_i, \mu_j) = \mu_i^\top D \mu_j$$

D *is diagonal*

generalize

generalize

Nonparametric latent feature relational model (LFRM)

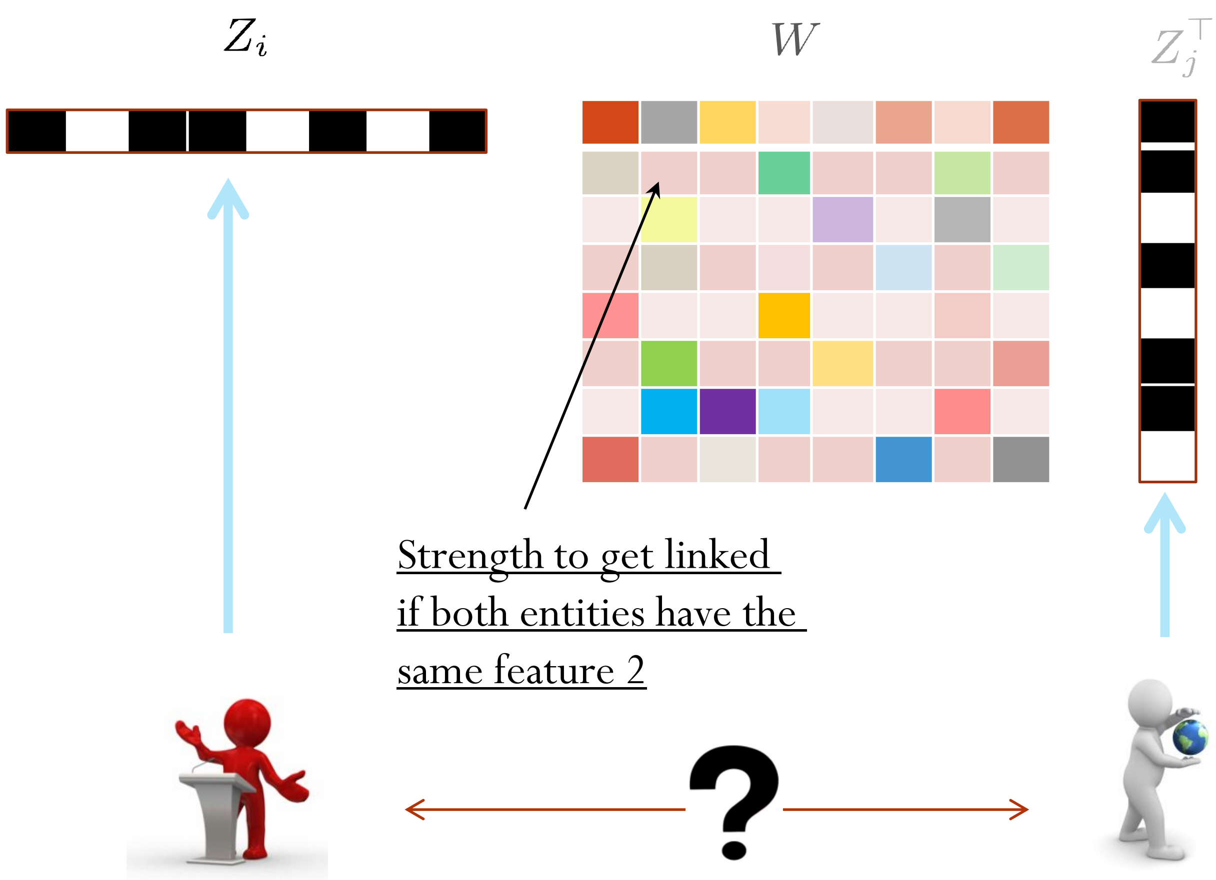
(Miller, Griffiths, & Jordan, 2009)

$$\psi(\mu_i, \mu_j) = \mu_i^\top W \mu_j, \ \ where \ \mu_i \in \{0, 1\}^\infty$$

More at ICML 2012

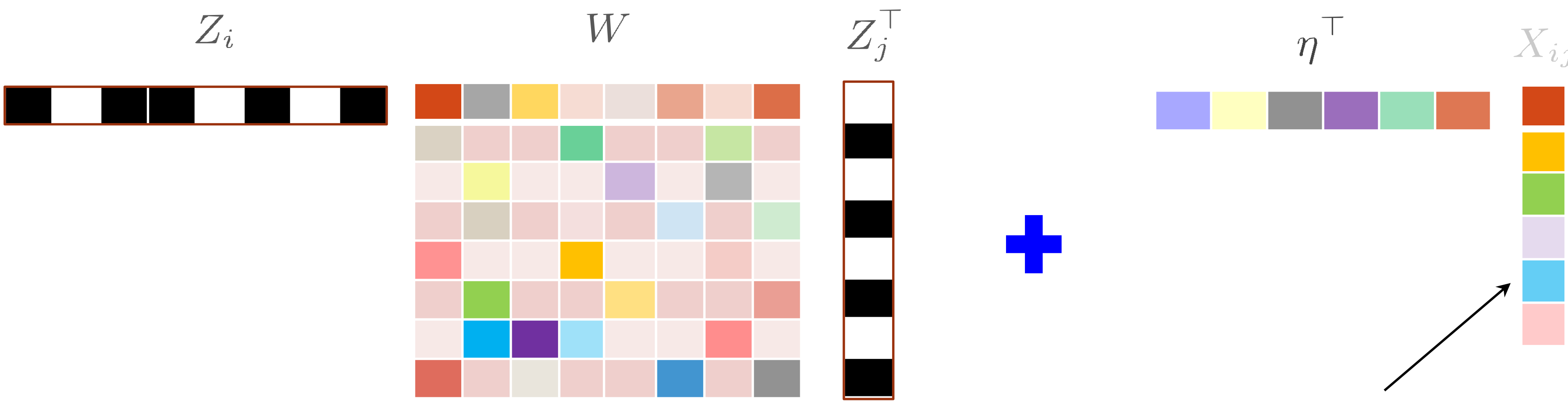
Discriminant Function with Latent Features

$$\boxed{f(Z_i, Z_j; W) = Z_i W Z_j^\top}$$



Discriminant Function with Latent Features

$$\boxed{f(Z_i, Z_j; X_{ij}, W, \eta) = Z_i W Z_j^\top + \eta^\top X_{ij}}$$



$$X_{ij} = [g(X_i); g(X_j); h(X_i, X_j)]$$

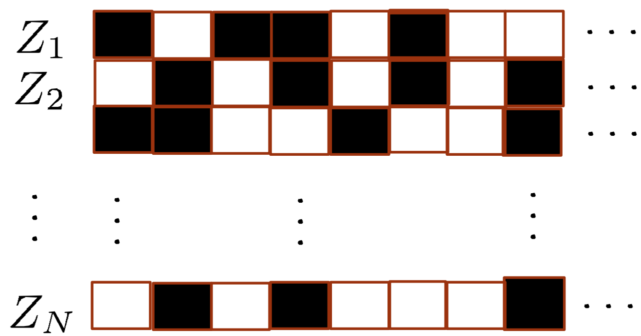
Ex: [red, table; gray, ball; neighbors…]

Latent Features

Observable Features

Infinite Latent Feature Matrix

- N entities → a latent feature matrix Z



- How many columns (i.e., features) are sufficient?

  → a stochastic process to infer from data – Indian buffet process (IBP) (Griffiths & Ghahramani, 2006)

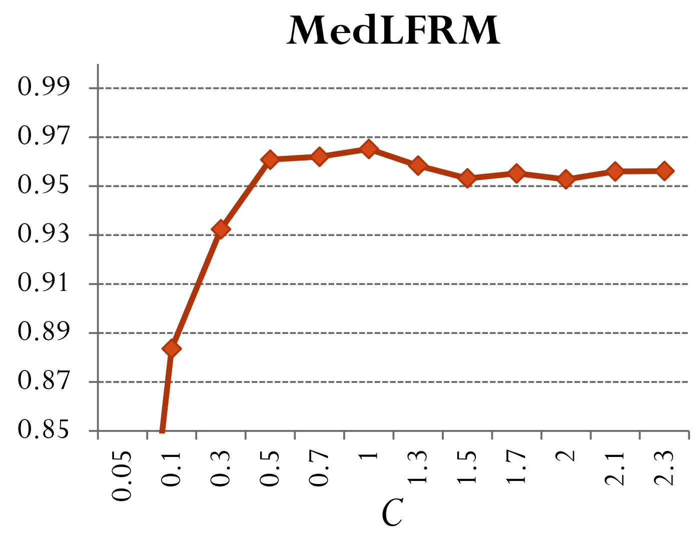
- What learning principle is good?

  → max-margin learning – (Vapnik, 1995; Taskar et al., 2003)

MedLFRM (max-margin latent feature relational model)

Bayesian MedLFRM

- One problem with MedLFRM is the tuning of C, e.g., using CV
- Hierarchical Bayesian ideas to infer it from data



- The Normal-Gamma hyper-prior:
  - Prior of model parameters with common mean and variance

$$p_0(\Theta|\mu,\tau) = \prod_{kk'} \mathcal{N}(\mu,\tau^{-1}) \prod_d \mathcal{N}(\mu,\tau^{-1})$$

  - The hyper-prior

$$p_0(\mu|\tau) = \mathcal{N}(\mu_0,(n_0\tau)^{-1}),\ \ p_0(\tau) = \mathcal{G}(\frac{\nu_0}{2},\frac{2}{S_0})$$

    - a weak hyper-prior suffices, e.g., $\mu_0 = 0,\ n_0 = 1,\ \nu_0 = 2,\ S_0 = 1$

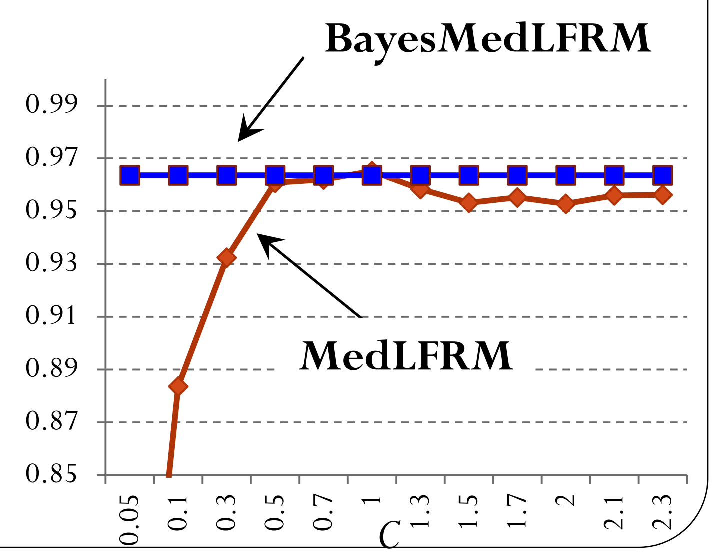
Bayesian MedLFRM

- Learning problem

$$\min_{p(\boldsymbol{\nu}, Z, \mu, \tau, \Theta)} \mathrm{KL}(p(\boldsymbol{\nu}, Z, \mu, \tau, \Theta) \| p_0(\boldsymbol{\nu}, Z, \mu, \tau, \Theta)) + \mathcal{R}_\ell(p(Z, \Theta))$$
$$\text{s.t.}: \; p(\boldsymbol{\nu}, Z, \mu, \tau, \Theta) \in \mathcal{P}.$$

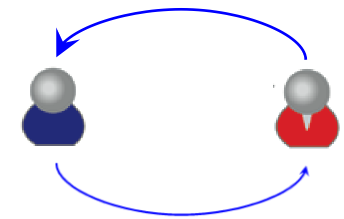
- Inference – similar iterative procedure (outline)
  - The step of inferring $p(\nu, Z)$ doesn't change
  - For $p(\Theta)$, we solve a binary SVM
  - For $p(\mu,\tau)$, we have closed-form rule

$$\frac{1}{C} = \lambda = \mathbb{E}[\tau] = \frac{\tilde{\nu}}{\tilde{S}}$$

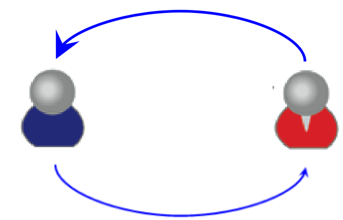


Datasets & Classification Setups

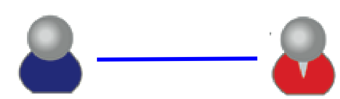
- Countries
  - 14 countries, 56 relations
  - 90 observable attributes about countries
  - predict the existence/non-existence (1/-1 classification) of each relation for each pair of country



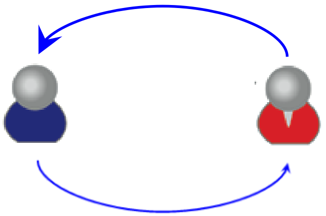
- Kinship
  - 104 people, 26 kinship relations
  - predict the existence/non-existence (1/-1 classification) of each relation for each pair of people



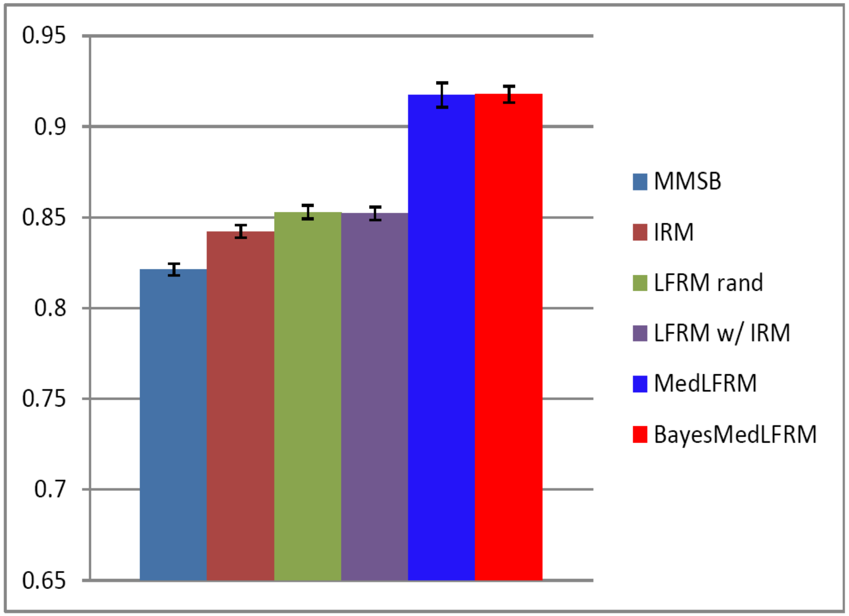
- Coauthor Networks:
  - 234 authors
  - 80% pairs for training; 20% for testing
  - Positive – author pairs that published papers together in train years;
  - Negative – author pairs that didn't publish any papers together in train years

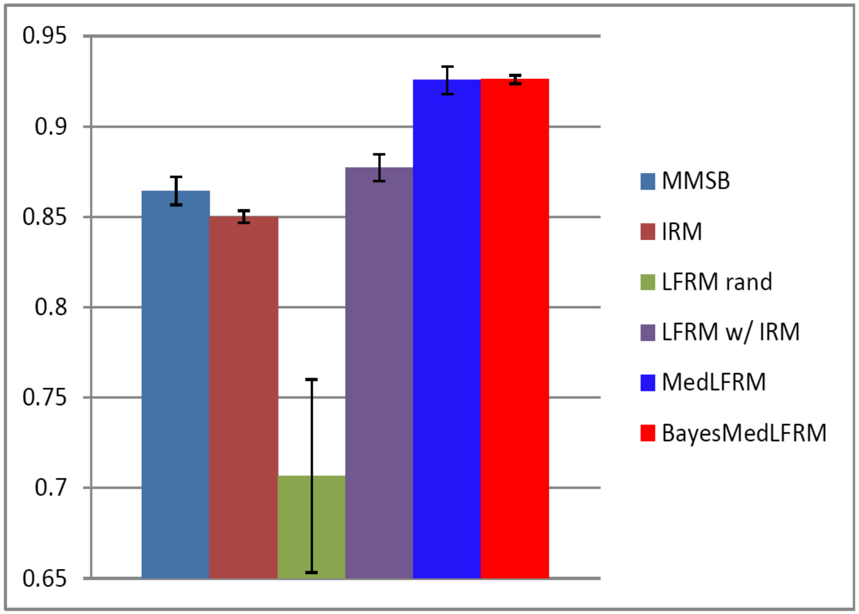


Results on Multi-relation Data



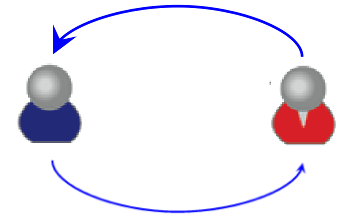
- AUC – area under ROC curve (higher, better)
- Two evaluation settings
  - Single – learn separate models for different relations, and average the AUC scores;
  - Global – learn one common model (i.e., features) for all relations



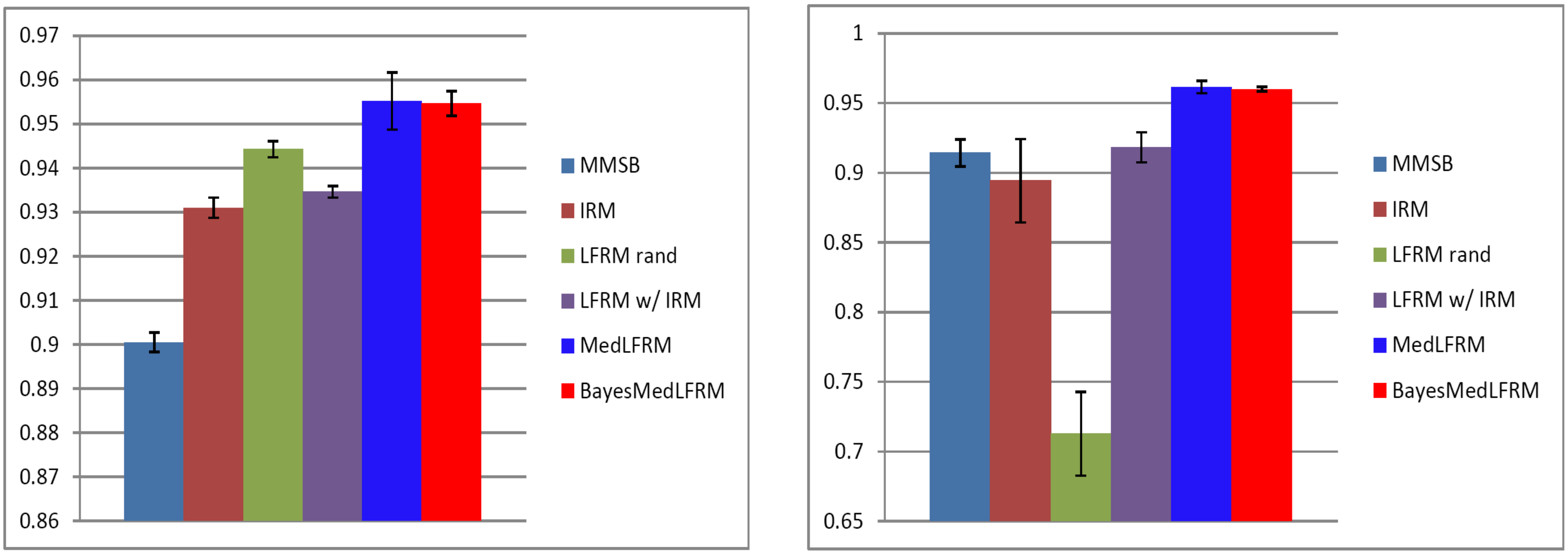


Country Relationships

Results on Multi-relation Data



- AUC – area under ROC curve (higher, better)
- Two evaluation settings
  - Single – learn separate models for different relations, and average the AUC scores;
  - Global – learn one common model (i.e., features) for all relations



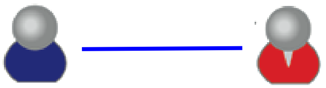
Kinship Relationships

Results on Coauthor Networks

- Two model settings:
  - Symmetric – W is a symmetric matrix:

$$Z_i W Z_j^\top = Z_j W Z_i^\top$$

  - Asymmetric – no above constraint

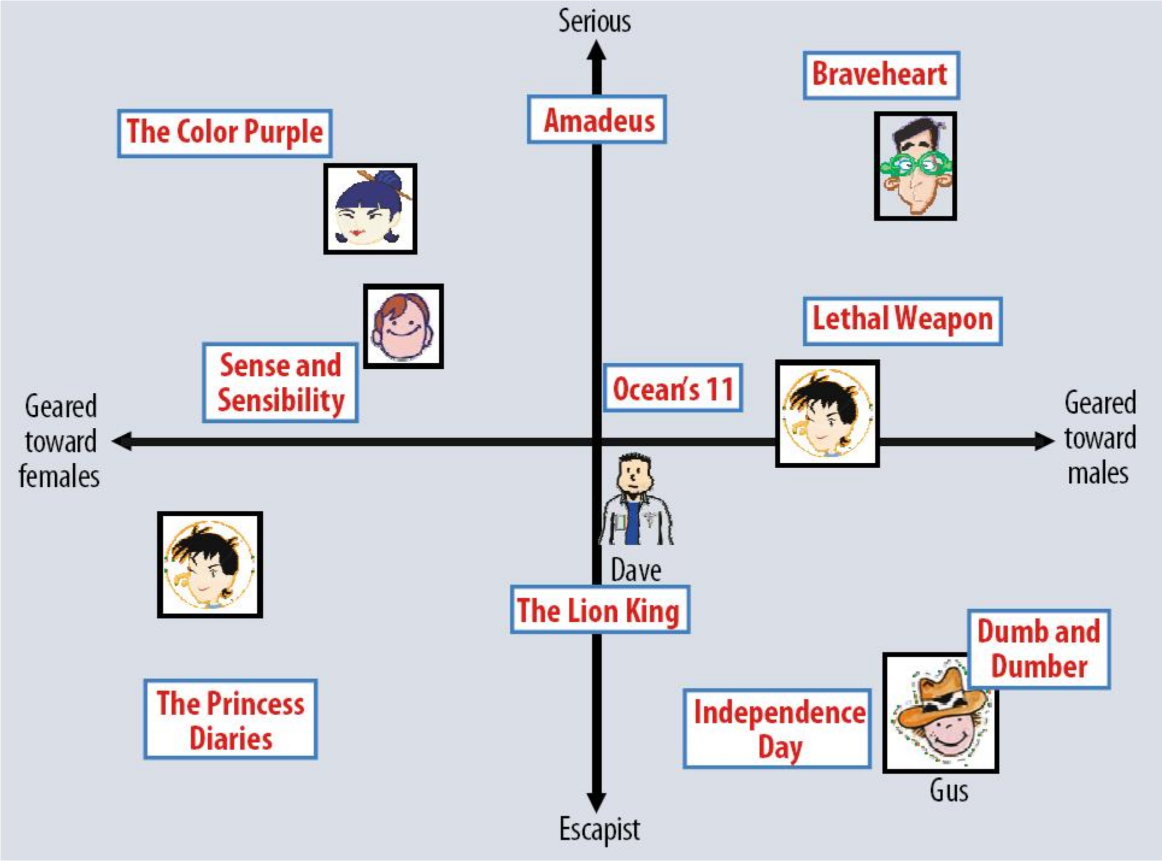


MMSB	0.8705 ± 0.0130
IRM	$0.8906 \pm$ —
LFRM rand	$0.9466 \pm$ —
LFRM w/ IRM	$0.9509 \pm$ —
MedLFRM	$\mathbf{0.9642} \pm 0.0026$
BayesMedLFRM	$\mathbf{0.9636} \pm 0.0036$
Asymmetric MedLFRM	0.9140 ± 0.0130
Asymmetric BayesMedLFRM	0.9146 ± 0.0047

Collaborative Filtering in Our Life

Latent Factor Methods

- Characterize both items & users on say 20 to 100 factors inferred from the rating patterns

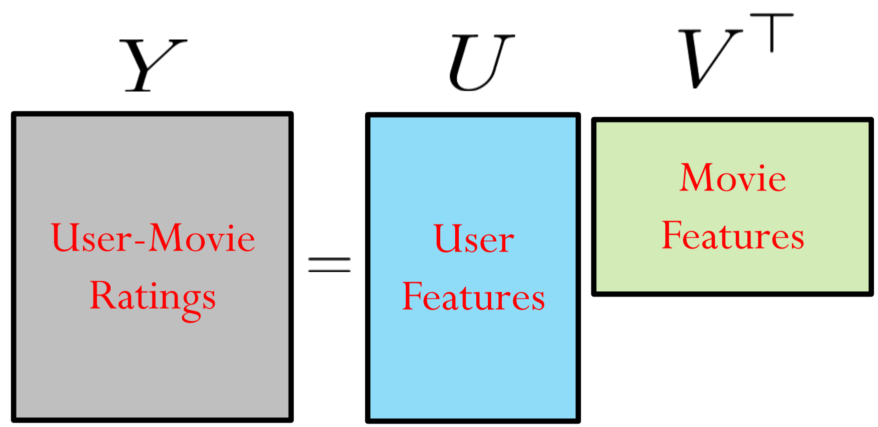


[Y. Koren, R. Bell & C. Volinsky, IEEE, 2009]

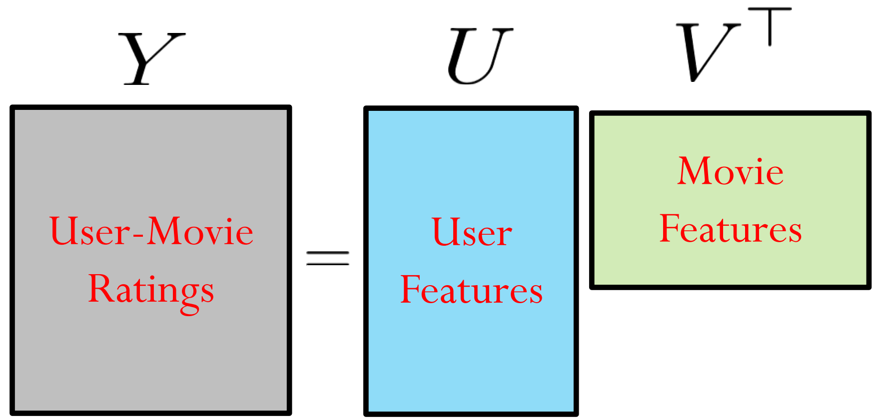
Matrix Factorization

- Some of the most successful latent factor models are based on matrix factorization

item / user	1	2	3	4 …
1	★★★★	?	★★★	★
2	★★★	?	?	★★
3	★★★★	★	?	?
⋮				



Two Key Issues



- How many columns (i.e., features) are sufficient?
  - → a stochastic process to infer it from data
- What learning principle is good?
  - → large-margin principle to learn classifiers

Nonparametric Max-margin Matrix Factorization for Collaborative Prediction

[Xu, Zhu, & Zhang, NIPS 2012]

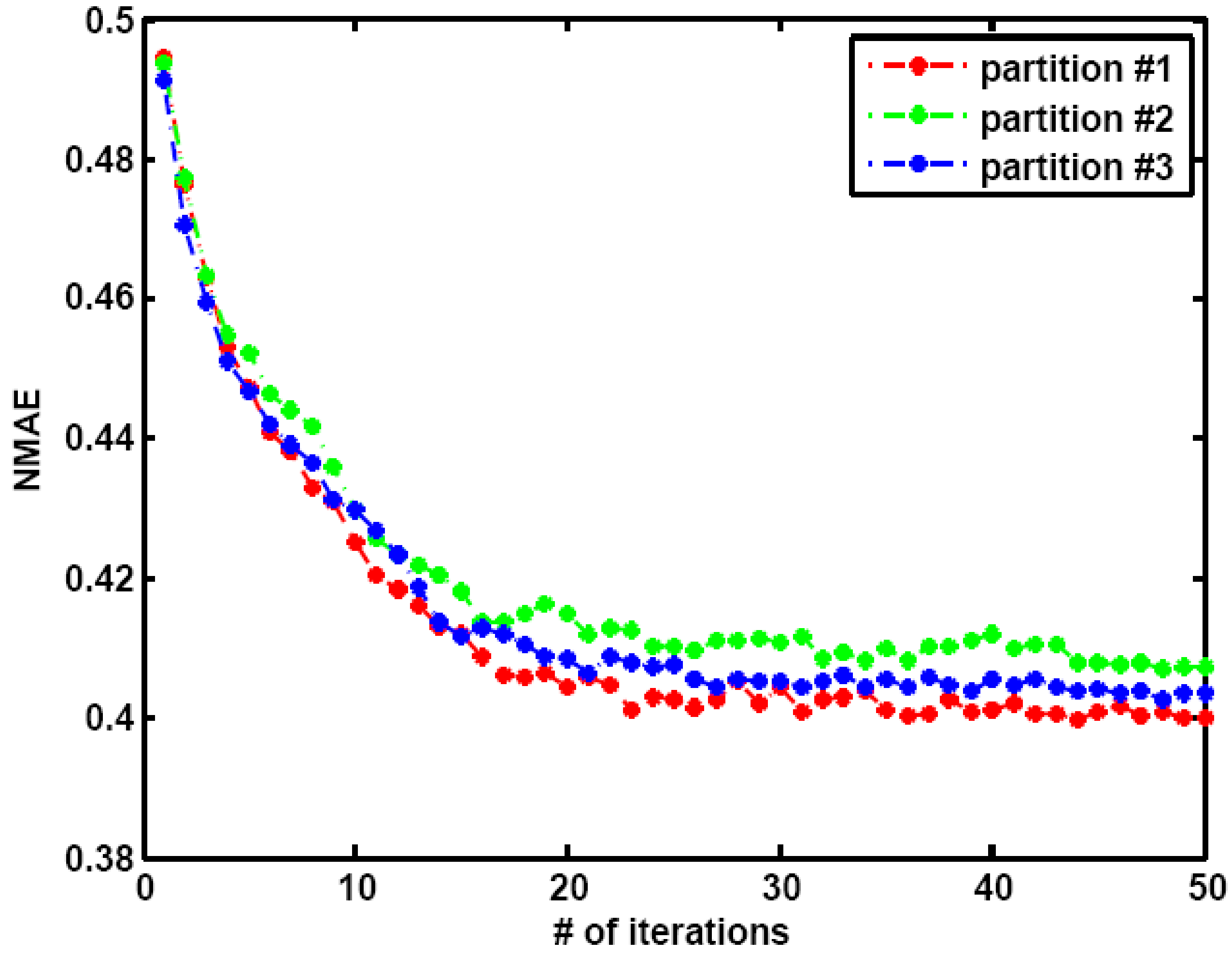
Experiments

- Data sets:
  - **MovieLens**: 1M anonymous ratings of 3,952 movies made by 6,040 users
  - **EachMovie**: 2.8M ratings of 1,628 movies made by 72,916 users
- Overall results on **Normalized Mean Absolute Error (NMAE)** (**the lower, the better**)

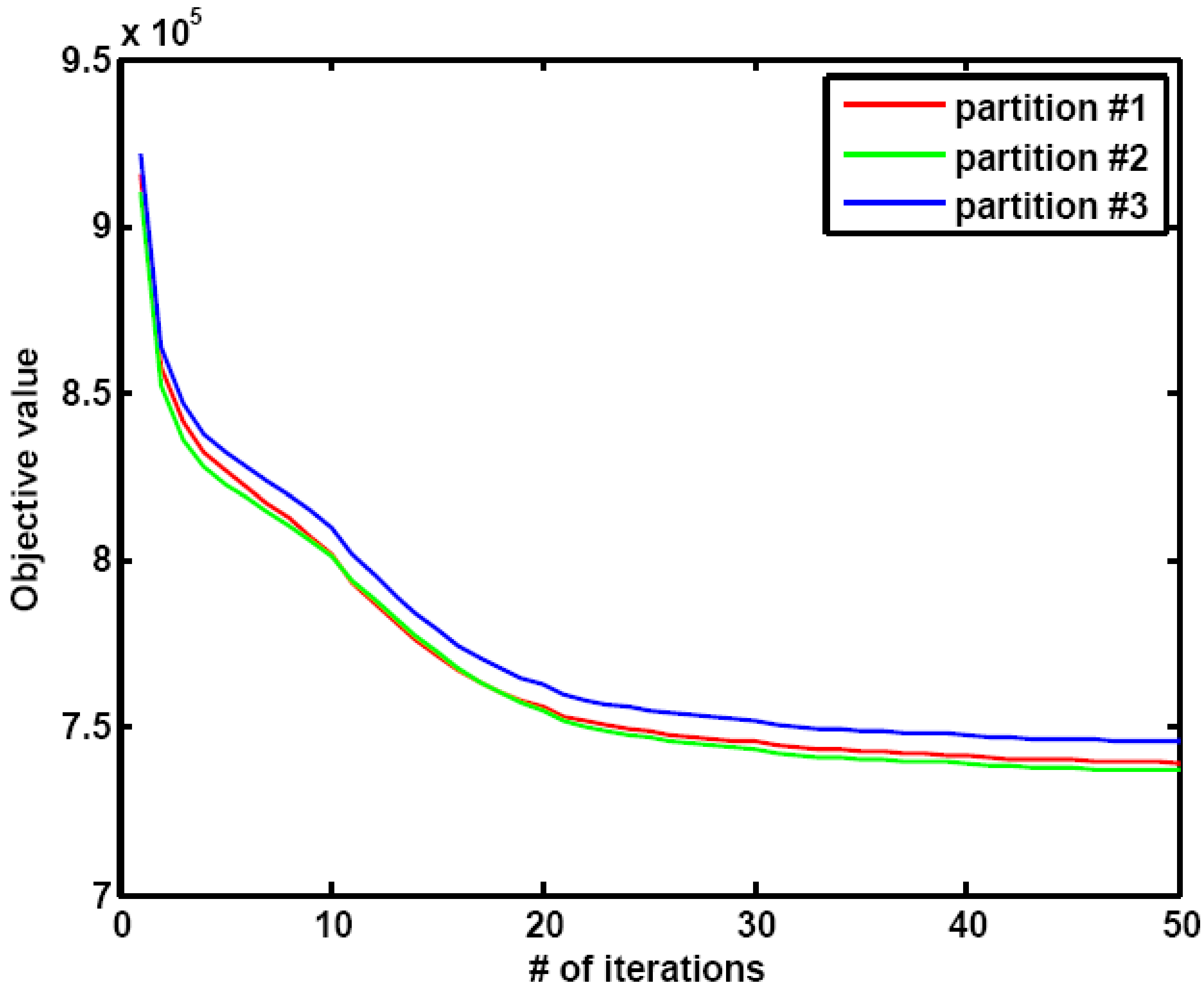
Table 1: NMAE performance of different models on MovieLens and EachMovie.

	MovieLens		EachMovie	
Algorithm	**weak**	**strong**	**weak**	**strong**
M^3F [11]	$.4156 \pm .0037$	$.4203 \pm .0138$	$.4397 \pm .0006$	$.4341 \pm .0025$
PMF [13]	$.4332 \pm .0033$	$.4413 \pm .0074$	$.4466 \pm .0016$	$.4579 \pm .0016$
BPMF [12]	$.4235 \pm .0023$	$.4450 \pm .0085$	$.4352 \pm .0014$	$.4445 \pm .0005$
M^3F^*	$.4176 \pm .0016$	$.4227 \pm .0072$	$.4348 \pm .0023$	$.4301 \pm .0034$
iPM^3F	$\mathbf{.4031} \pm .0030$	$.4135 \pm .0109$	$\mathbf{.4211} \pm .0019$	$\mathbf{.4224} \pm .0051$
$iBPM^3F$	$.4050 \pm .0029$	$\mathbf{.4089} \pm .0146$	$.4268 \pm .0029$	$.4403 \pm .0040$

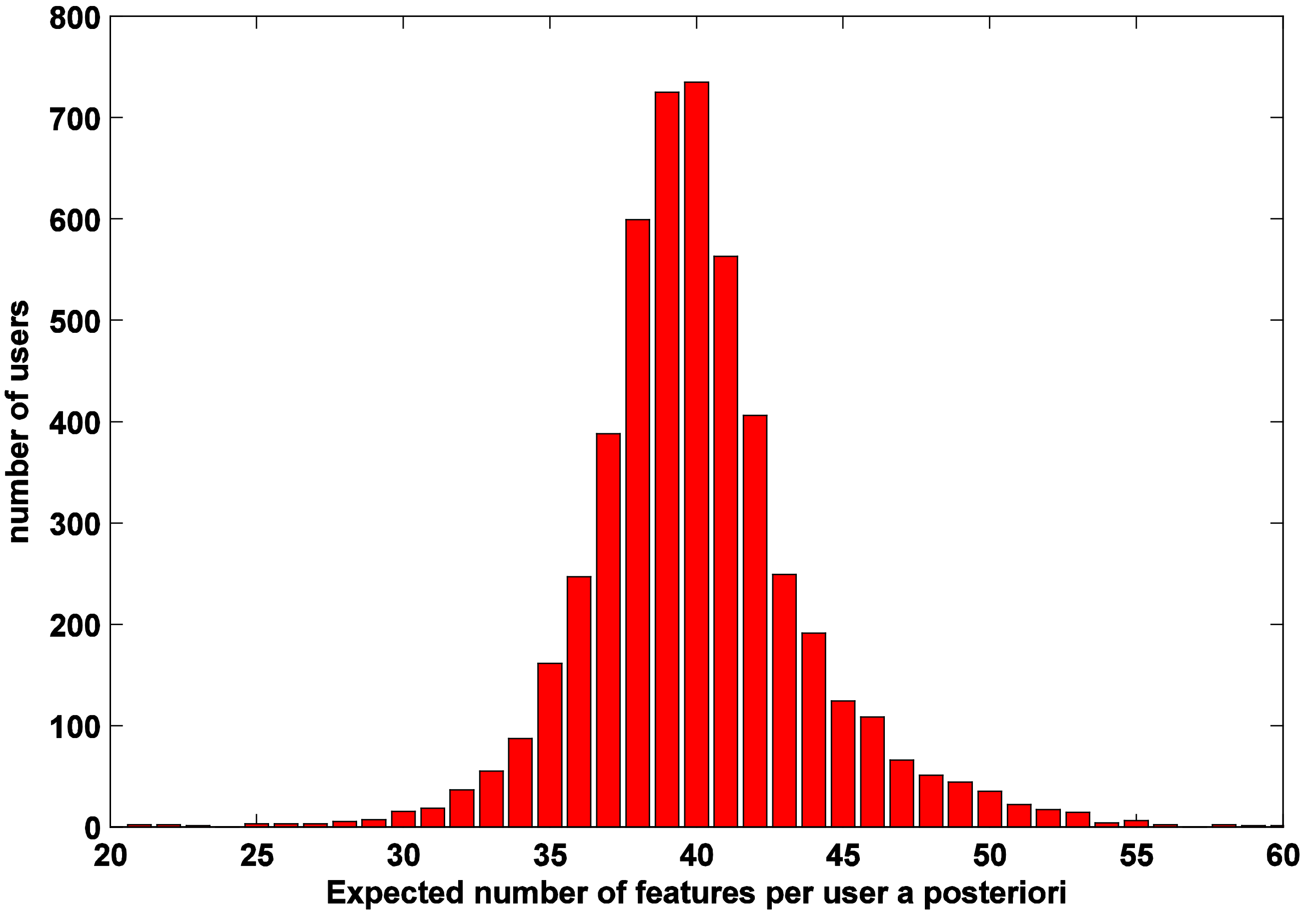
Prediction Performance during Iterations



Objective Value during Iterations



Expected Number of Features per User



Fast Sampling Algorithms

- See our paper [Xu, Zhu, & Zhang, ICML2013] for details

	MovieLens		EachMovie	
Algorithm	weak	strong	weak	strong
M^3F	$.4156 \pm .0037$	$.4203 \pm .0138$	$.4397 \pm .0006$	$.4341 \pm .0025$
bcd M^3F	$.4176 \pm .0016$	$.4227 \pm .0072$	$.4348 \pm .0023$	$.4301 \pm .0034$
Gibbs M^3F	$.4037 \pm .0005$	$.4040 \pm .0055$	$.4134 \pm .0017$	$.4142 \pm .0059$
iPM^3F	$.4031 \pm .0030$	$.4135 \pm .0109$	$.4211 \pm .0019$	$.4224 \pm .0051$
Gibbs iPM^3F	$.4080 \pm .0013$	$.4201 \pm .0053$	$.4220 \pm .0003$	$.4331 \pm .0057$

Algorithm	MovieLens	EachMovie	Iters
M^3F	5h	15h	100
bcd M^3F	4h	10h	50
Gibbs M^3F	0.11h	0.35h	50
iPM^3F	4.6h	5.5h	50
Gibbs iPM^3F	0.68h	0.70h	50

30 times faster!

8 times faster!

Part II

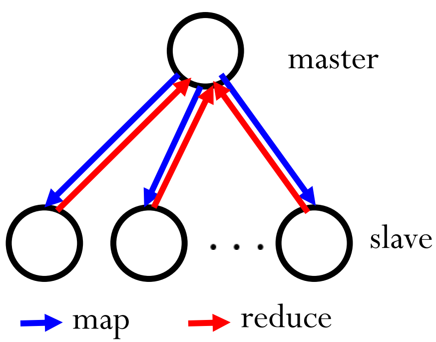
Scalable Bayesian Methods

Existing Methods & Systems

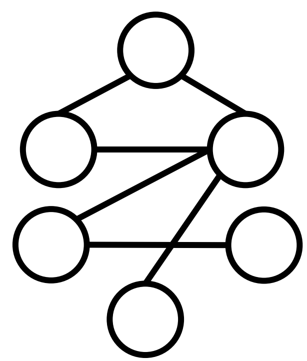
- Stochastic/Online Methods
  - Variational, MCMC

- Distributed Methods
  - Variational, MCMC

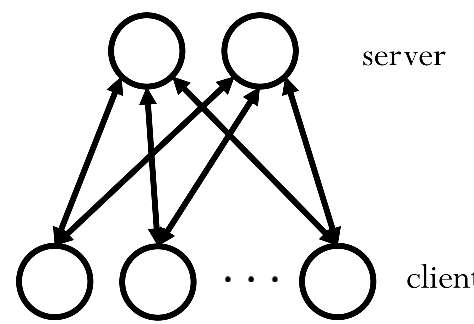
Data-Parallel

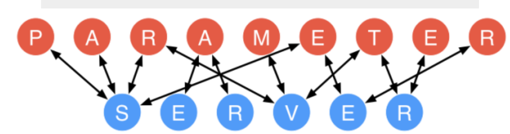


Graph-Parallel



Model-Parallel





Online Bayesian Updating

- Sequential Bayesian updating
  - Bayes' rule

$$p(\Theta \mid C_1) = p(C_1)^{-1}\, p(C_1 \mid \Theta)\, p(\Theta)$$

  - Suppose we have processed *b-1* collections, i.e., mini-batches
  - Given $p(\Theta \mid C_1, \ldots, C_{b-1})$, we have

$$p(\Theta \mid C_1, \ldots, C_b) \propto p(C_b \mid \Theta)\, p(\Theta \mid C_1, \ldots, C_{b-1})$$

  - i.e.: we treat the posterior after *b-1* mini-batches as the new prior for incoming data

- It is truly online (or streaming) *if we can save the posteriors and calculate the normalizing constant*

Online Variational Bayes

- However, it is often infeasible to calculate the posterior exactly
- Therefore, approximation must be used

- **Online (Streaming) Variational Bayes**:
  - An algorithm $\mathcal{A}$ that calculates an approximate posterior

$$q(\Theta) = \mathcal{A}(C, p(\Theta))$$

  where C is the mini-batch data; $p(\Theta)$ is the given prior
  - Recursively calculate an approximation to the posterior

$$p(\Theta \mid C_1, \ldots, C_b) \approx q_b(\Theta) = \mathcal{A}\left(C_b, q_{b-1}(\Theta)\right)$$

  - where $q_0(\Theta) = p(\Theta)$

Distributed Bayesian Updating

- Bayesian theorem yields

$$p(\Theta \mid C_1, \ldots, C_B) \propto \left[\prod_{b=1}^{B} p(C_b \mid \Theta)\right] p(\Theta)$$

$$\propto \left[\prod_{b=1}^{B} p(\Theta \mid C_b)\, p(\Theta)^{-1}\right] p(\Theta)$$

- Therefore, we can calculate the individual mini-batch posteriors $p(\Theta \mid C_b)$ in parallel; then combine them to find the full posterior

Distributed Variational Bayes

- Given an approximating algorithm $\mathcal{A}$
- The approximate update is

$$p(\Theta \mid C_1, \ldots, C_B) \approx q(\Theta) \propto \left[\prod_{b=1}^{B} \mathcal{A}(C_b, p(\Theta))\ p(\Theta)^{-1} \right] p(\Theta)$$

- Ex: for exponential family distributions
  - Assume the prior $p(\Theta) \propto \exp\{\xi_0 \cdot T(\Theta)\}$
  - Assume the approximating algorithm

$$q_b(\Theta) \propto \exp\{\xi_b \cdot T(\Theta)\} \quad \text{for} \quad q_b(\Theta) = \mathcal{A}(C_b, p(\Theta))$$

  - Then, we have

$$p(\Theta \mid C_1, \ldots, C_B) \approx q(\Theta) \propto \exp\left\{ \left[\xi_0 + \sum_{b=1}^{B} (\xi_b - \xi_0) \right] \cdot T(\Theta) \right\}$$

Some Empirical Results

- Datasets: Wikipedia (3.6M); Nature (0.35M)

	Wikipedia		
	32-SDA	1-SDA	SVI
Log pred prob	**−7.31**	−7.43	−7.32
Time (hours)	**2.09**	43.93	7.87

Nature		
32-SDA	1-SDA	SVI
−7.11	−7.19	**−7.08**
0.55	10.02	1.22

- SVI is faster than single-thread SDA-Bayes in the single-pass setting

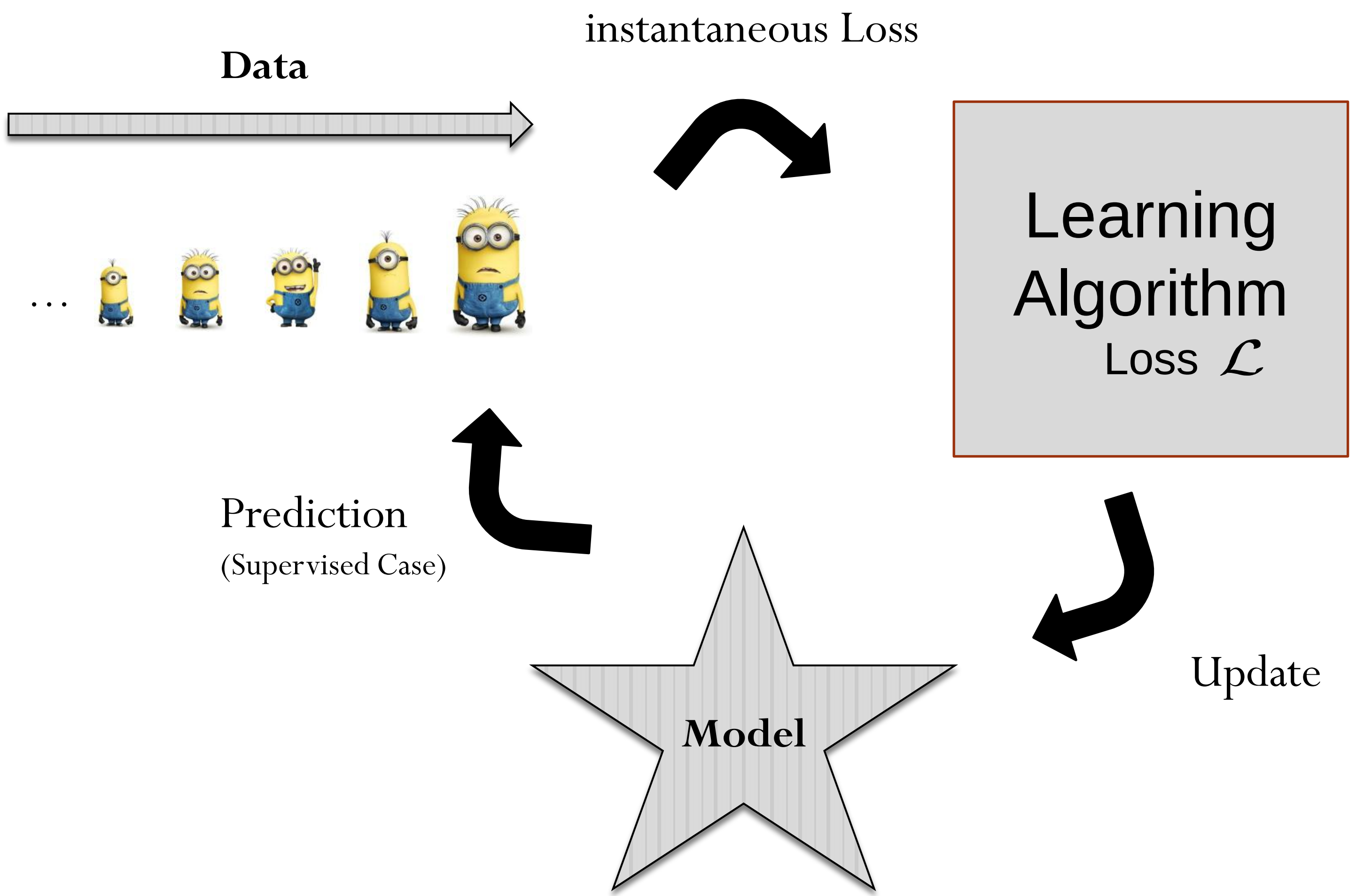
Online Bayesian Passive-Aggressive Learning

Why Online Learning?

- **Streaming data**:
  - Data come in streams
  - "Infinite" size
  - Processed data thrown away
  - Real-time response

- **Fixed, large-scale data**:
  - Statistic redundancy
  - "online learning" by sub-sampling
  - Stochastic learning/optimization
  - Fast convergence to satisfactory results

Online Learning for Classification



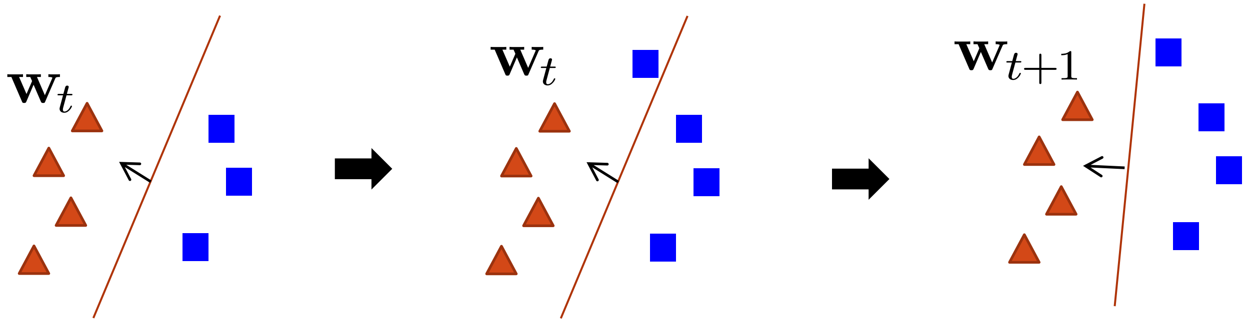
Perceptron --- a simplest example

- Set $\mathbf{w}_1 = 0$ and t=1; scale all examples to have length 1 (doesn't affect which side of the plane they are on)
- Given example x, predict positive *iff*

$$\mathbf{w}_t^\top \mathbf{x} > 0$$

- If a mistake, update as follows
  - Mistake on positive: $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + \mathbf{x}$
  - Mistake on negative: $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \mathbf{x}$

$t \leftarrow t + 1$



Mistake Bound

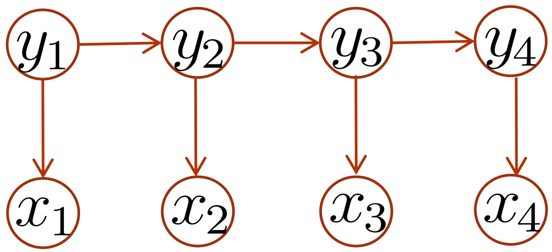
- Theorem:
  - *Let $\mathcal{S}$ be a sequence of labeled examples consistent with a linear threshold function* $\mathbf{w}_*^\top \mathbf{x} > 0$ *, where* $\mathbf{w}_*$ *is a unit-length vector.*
  - *The number of mistakes M made by the Perceptron algorithm is at most* $(1/\gamma)^2$*, where*

$$\gamma = \min_{\mathbf{x} \in \mathcal{S}} \frac{|\mathbf{w}_*^\top \mathbf{x}|}{\|\mathbf{x}\|}$$

  - *i.e.: if we scale examples to have length 1, then* γ *is the minimum distance of any example to the plane* $\mathbf{w}_*^\top \mathbf{x} = 0$
  - γ *is often called the "margin" of* $\mathbf{w}_*$ *; the quantity* $\frac{\mathbf{w}_*^\top \mathbf{x}}{\|\mathbf{x}\|}$ *is the cosine of the angle between* $\mathbf{x}$ *and* $\mathbf{w}_*$

Generalization of Perceptron

- Discriminative Training of HMMs or Markov networks (Collins, 2002)



  - For each input $\mathbf{x}_t$, predict the structured label (e.g., sequence) $\hat{\mathbf{y}}_t$
  - Update the parameters via the rule, if $\hat{\mathbf{y}}_t \neq \mathbf{y}_t$:

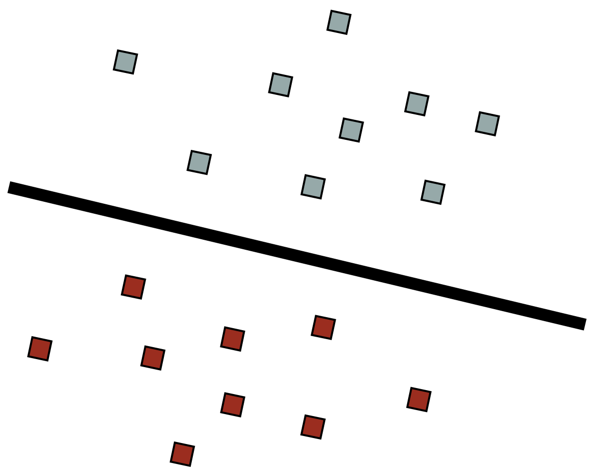
$$\mathbf{w}_{t+1} = \mathbf{w}_t + \psi(\mathbf{x}_t, \mathbf{y}_t) - \psi(\mathbf{x}_t, \hat{\mathbf{y}}_t)$$

  - i.e., moving toward the truth.

[M. Collins. Discriminative Training Methods for HMMs: Theory and Experiments with Perceptron Algorithms. EMNLP, 2002]

Loss for Binary Classification

- Loss (or constraint) at time t for a simple linear model:

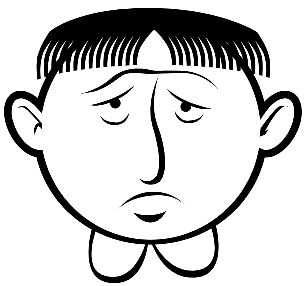


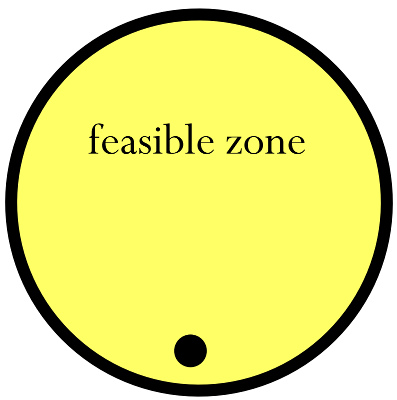
$$y_t \mathbf{w}^\top \mathbf{x}_t \geq \epsilon$$

$$(\mathbf{x}_t \in \mathcal{R}^n, y_t \in \{-1, 1\})$$

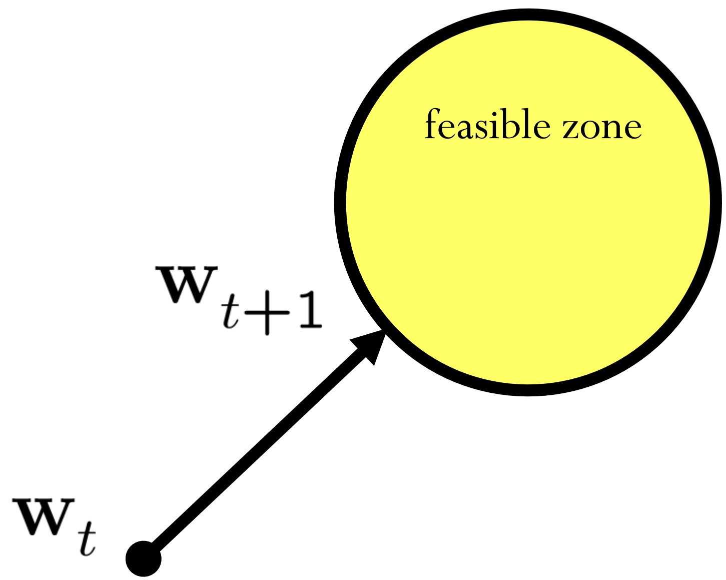
Online Passive-Aggressive Updates

- Model at time t: $\mathbf{w}_t$; When seeing a new data point:





$$\mathbf{w}_{t+1} = \mathbf{w}_t$$



[Crammer et al., 2006]

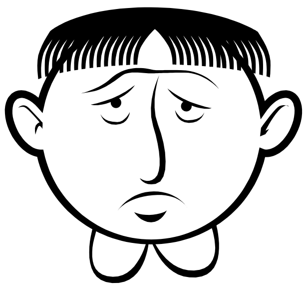
Bayesian Passive-Aggressive Learning

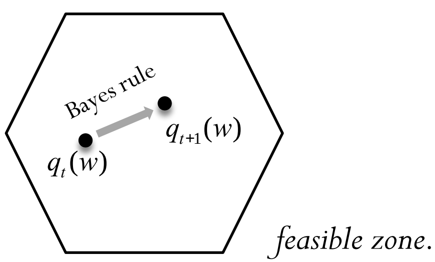
- **Basic setup**: we learn a distribution over models, rather than a single model

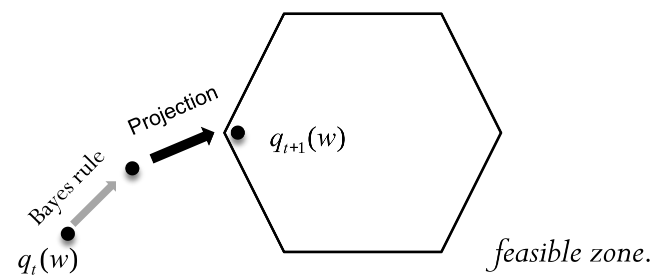
[Shi & Zhu, ICML 2014; invited to JMLR]

New Passive-Aggressive Updates

- Distribution at time t: $q_t(\mathbf{w})$; When seeing a new data point







- Note: Bayes update is optional.

[Shi & Zhu, ICML 2014; invited to JMLR]

Two Choices to Define "Feasible Zone"

- **Share a common goal**: learn a posterior distribution over models

$$q(\mathbf{w})$$

- Different strategies in making predictions:
  - **Averaging classifier**: takes an average of the discriminant function and predicts

$$\hat{y}_t = \mathrm{sign}\Big(\mathbb{E}_q[\mathbf{w}^\top \mathbf{x}_t] \Big)$$

  - **Gibbs classifier**: randomly draws a model and predicts

$$\mathbf{w}_* \sim q(\mathbf{w})$$

$$\hat{y}_t = \mathrm{sign}\Big(\mathbf{w}_*^\top \mathbf{x}_t \Big)$$

[Shi & Zhu, ICML 2014; invited to JMLR]

The Learning Problem

- Hard constraint version:

$$\min_{q(\boldsymbol{w})\in\mathcal{F}_t} \mathbf{KL}\Big[q(\boldsymbol{w})||q_t(\boldsymbol{w})\Big] - \boxed{\mathbb{E}_{q(\boldsymbol{w})}\Big[\log p(\boldsymbol{x}_t|\boldsymbol{w})\Big]}$$
$$\text{s.t.:}\ \ \ell_\epsilon\Big(q(\boldsymbol{w});\boldsymbol{x}_t,y_t\Big)=0,$$

optional

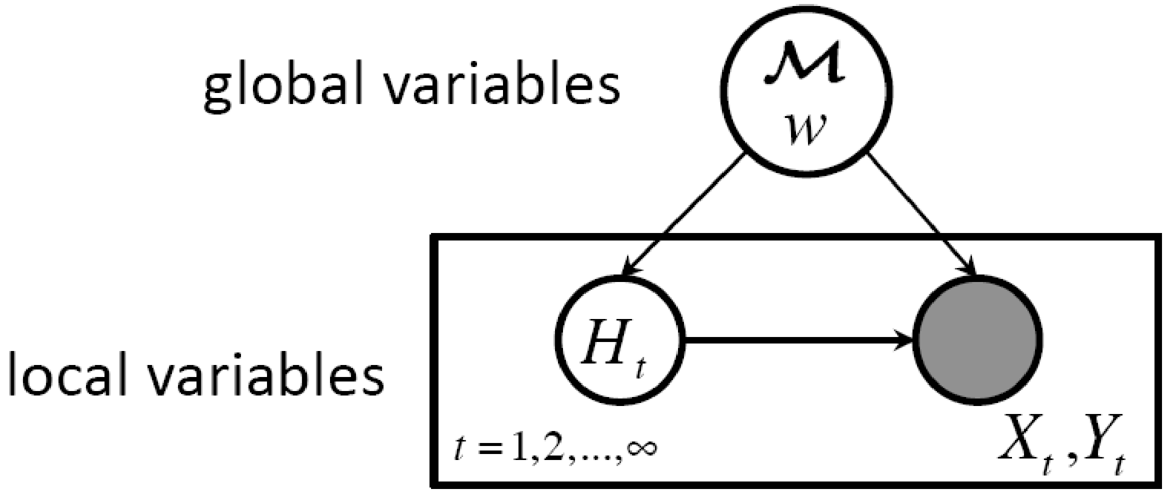
- Soft constraint version:

$$q_{t+1}(\boldsymbol{w}) = \operatorname*{argmin}_{q(\boldsymbol{w})\in\mathcal{F}_t} \mathcal{L}\Big(q(\boldsymbol{w})\Big) + 2c\cdot\ell_\epsilon\Big(q(\boldsymbol{w});\boldsymbol{x}_t,y_t\Big)$$

Some Properties

- **Theorem 1**: *non-Bayesian PA is a special case*.

- One significance of Bayesian models is to **learn latent structures**

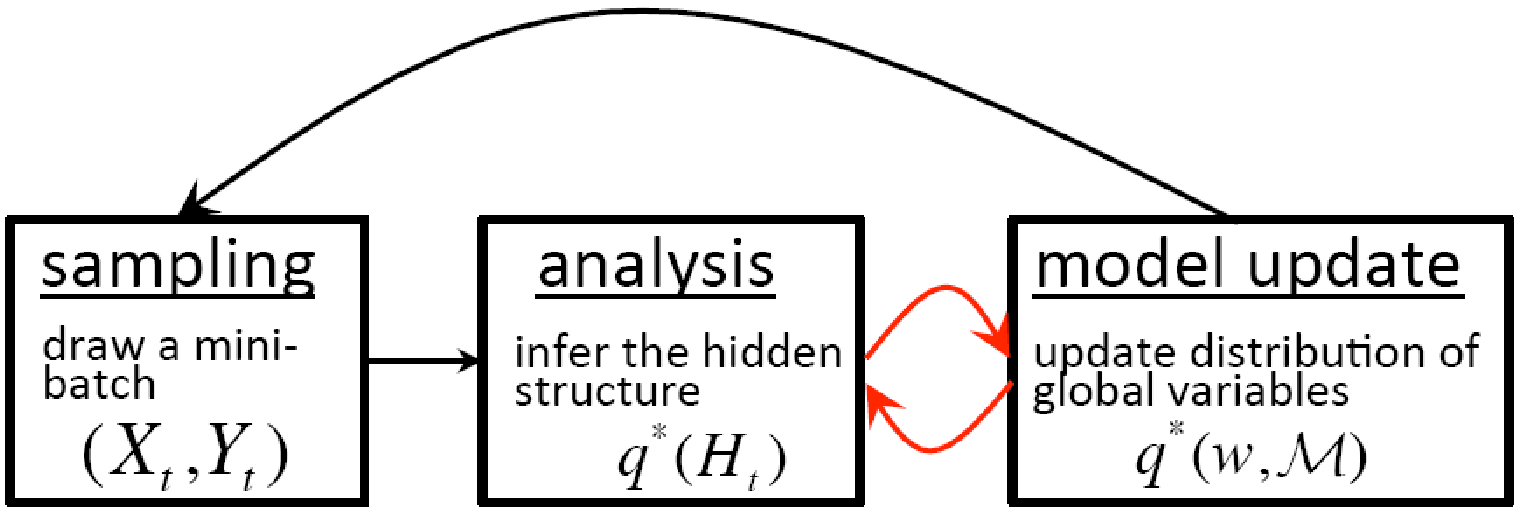


[Shi & Zhu, ICML 2014; invited to JMLR]

Some Properties

- **Theorem 1**: *non-Bayesian PA is a special case.*

- One significance of Bayesian models is to **learn latent structures**



[Shi & Zhu, ICML 2014; invited to JMLR]

Some Properties

- **Theorem 2**: *under certain conditions, the regret of BayesPA is bounded as*

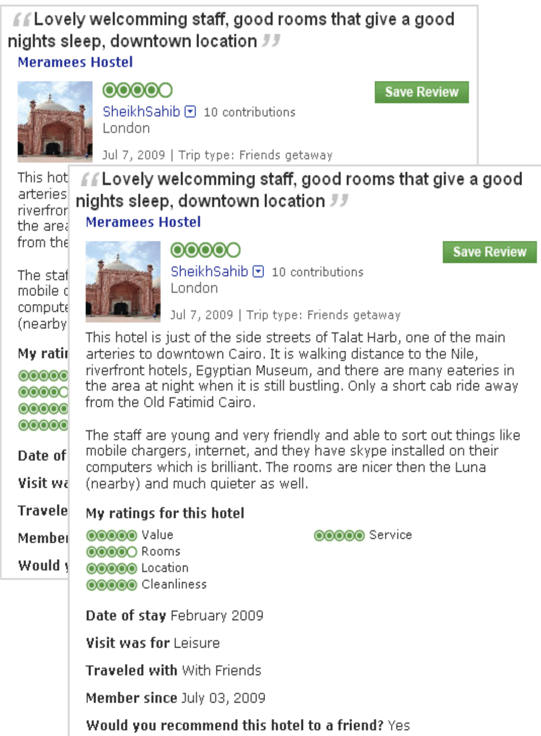
$$\frac{1}{T}\sum_{t=0}^{T-1}\mathcal{R}_c(q_t(\boldsymbol{w});\boldsymbol{x}_t,y_t) \leq \frac{1}{T}\sum_{t=0}^{T-1}\mathcal{R}_c(p(\boldsymbol{w});\boldsymbol{x}_t,y_t) + \frac{1}{T}\mathbf{KL}[p(\boldsymbol{w})||q_0(\boldsymbol{w})] + \text{const.}$$

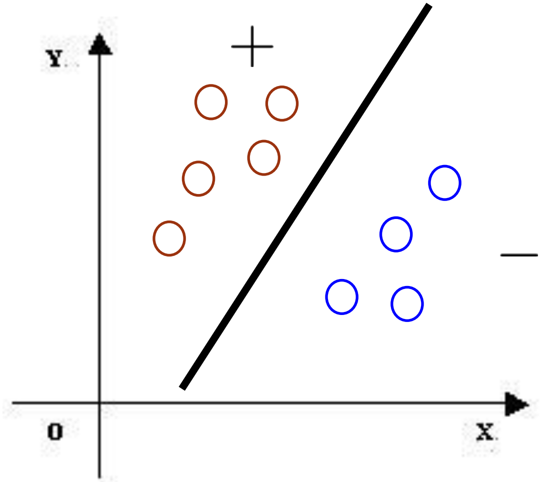
- **Remarks**:
  - Holds for any choice of $p(\mathbf{w})$; including the best by batch learning
  - Holds for both averaging and Gibbs classifiers
  - As $T \to \infty$, the asymptotic regret is at most worse by a constant

[Shi & Zhu, ICML 2014; invited to JMLR]

Applications to Topic Modeling

◆ Discover semantic topic representations



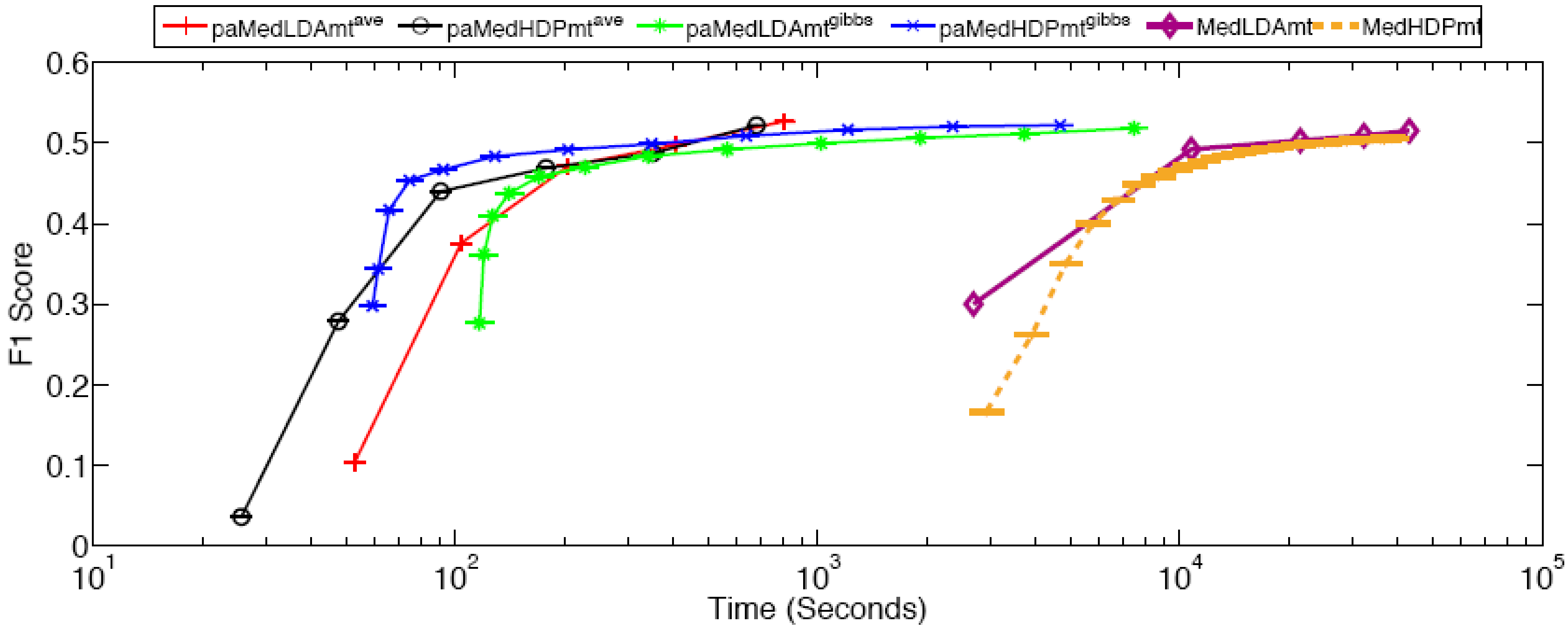


Axis of a semantic space:

T1	T2	T3	T4	T5	T6	T7
told	place	hotel	hotel	beach	beach	great
dirty	hotel	food	area	pool	resort	good
room	room	bar	staff	resort	pool	nice
front	days	day	pool	food	ocean	lovely
asked	time	pool	breakfast	island	island	beautiful
hotel	day	time	day	kids	kids	excellent
bad	night	service	view	trip	good	wonderful
small	people	holiday	location	service	restaurants	comfortable
worst	stay	room	service	day	enjoyed	beach
poor	water	people	walk	staff	loved	friendly
called	rooms	night	time	time	trip	fresh
rude	food	water	food	view	area	amazing

Empirical Results on Large-scale Wiki Data

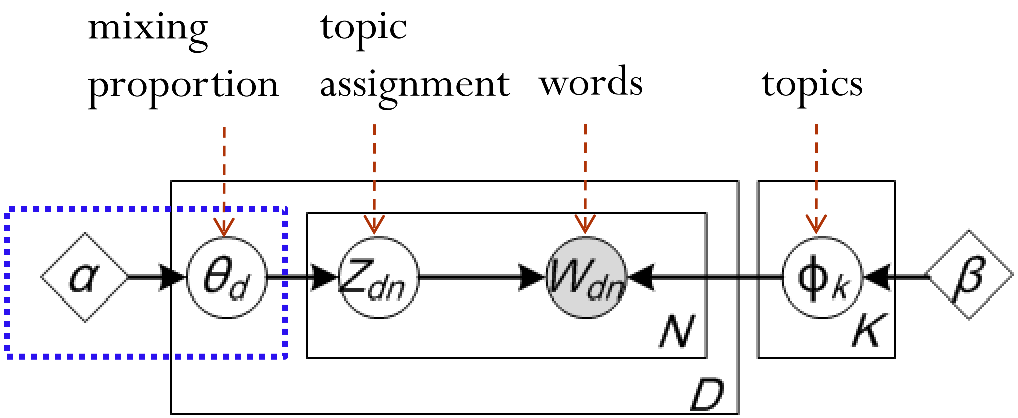
- Wikipedia webpages with multi-labels:
  - 1.1 million documents; 0.9 million unique terms



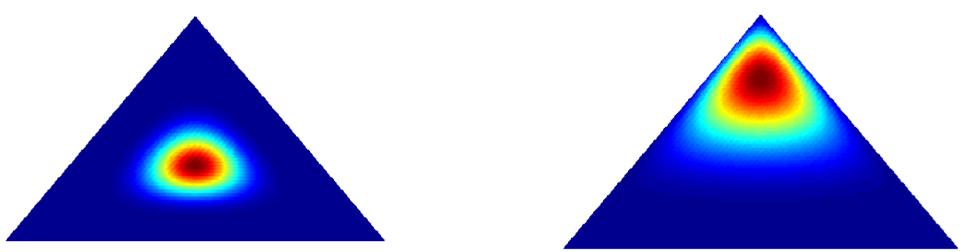
Large-scale Topic Graph Learning and Visualization

Logistic-Normal Topic Models

- Bayesian topic models



- Dirichlet priors are conjugate to the multinomial likelihood
- However, it doesn't capture the correlation among topics



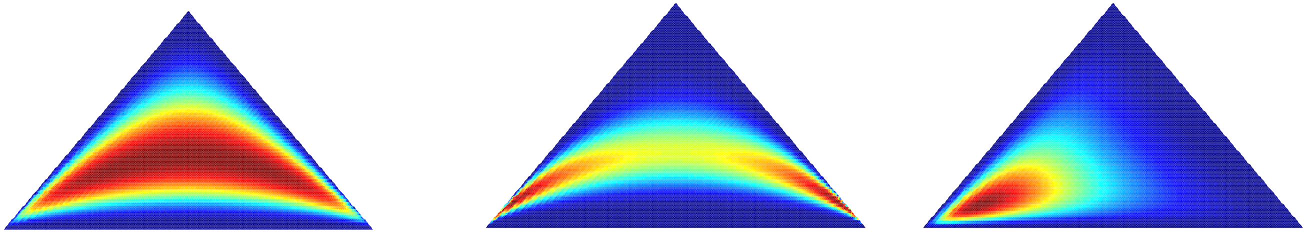
Logistic-Normal Topic Models

- Logistic-normal prior distribution (Aitchison & Shen, 1980)

$$\eta_d \sim \mathcal{N}(\mu, \Sigma)$$

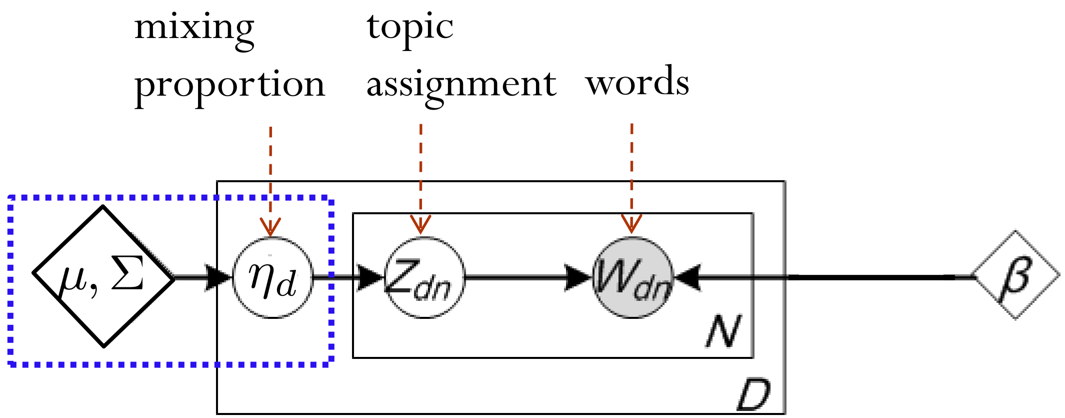
$$\theta_d^k = \frac{\exp\left(\eta_d^k\right)}{\sum_i \exp\left(\eta_d^i\right)}$$

  - Logisitc-normal prior can capture the correlationships



  - But it is non-conjugate to a multinomial likelihood !
  - Variational approximation not scalable (Blei & Lafferty, 2007)

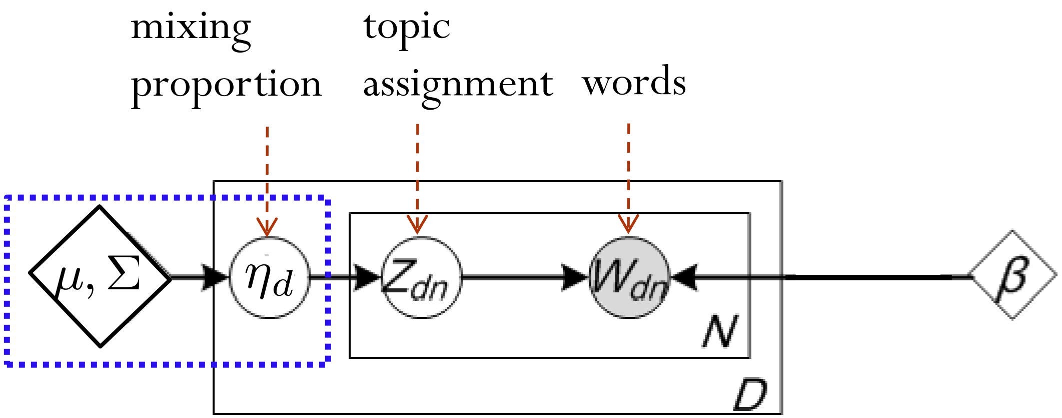
A Scalable Gibbs Sampler



- Collapse out the topics by conjugacy
- Sample **Z**: (standard)

$$p(z_{dn}^k = 1 | \mathbf{Z}_{\neg n}, w_{dn}, \mathbf{W}_{\neg dn}, \boldsymbol{\eta}) \propto \frac{C_{k,\neg n}^{w_{dn}} + \beta_{w_{dn}}}{\sum_{j=1}^V C_{k,\neg n}^j + \sum_{j=1}^V \beta_j} e^{\eta_d^k}$$

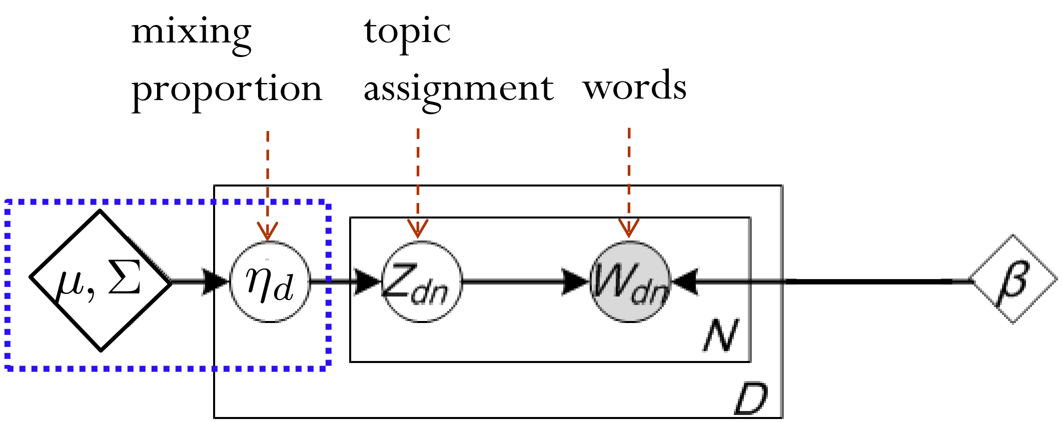
A Scalable Gibbs Sampler



- Collapse out the topics by conjugacy
- Sample η : (challenging)

$$p(\boldsymbol{\eta}|\mathbf{Z},\mathbf{W}) \propto \prod_{d=1}^{D} \left(\prod_{n=1}^{N_d} \frac{e^{\eta^d_{z_n}}}{\sum_{j=1}^{K} e^{\eta^d_j}} \right) \mathcal{N}(\boldsymbol{\eta}_d | \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

A Scalable Gibbs Sampler



- Data augmentation saves!
- For each dimension k:

$$p(\eta_d^k | \boldsymbol{\eta}_d^{\neg k}, \mathbf{Z}, \mathbf{W}) \propto \ell(\eta_d^k | \boldsymbol{\eta}_d^{\neg k}) \mathcal{N}(\eta_d^k | \mu_d^k, \sigma_k^2)$$

$$\ell(\eta_d^k | \boldsymbol{\eta}_d^{\neg k}) = \frac{(e^{\rho_d^k})^{C_d^k}}{(1 + e^{\rho_d^k})^{N_d}}$$

Data Augmentation

- A scale-location mixture representation

$$\frac{(e^{\rho_d^k})^{C_d^k}}{(1+e^{\rho_d^k})^{N_d}} = \frac{1}{2^{N_d}} e^{\kappa_d^k \rho_d^k} \int_0^\infty e^{-\frac{\lambda_d^k (\rho_d^k)^2}{2}} p(\lambda_d^k | N_d, 0) d\lambda_d^k$$

  - where

$$\kappa_d^k = C_d^k - N_d/2 \quad p(\lambda_d^k | N_d, 0) = \mathcal{PG}(N_d, 0)$$

- Then, we iteratively draw samples
  - Draw *1d* Gaussian:

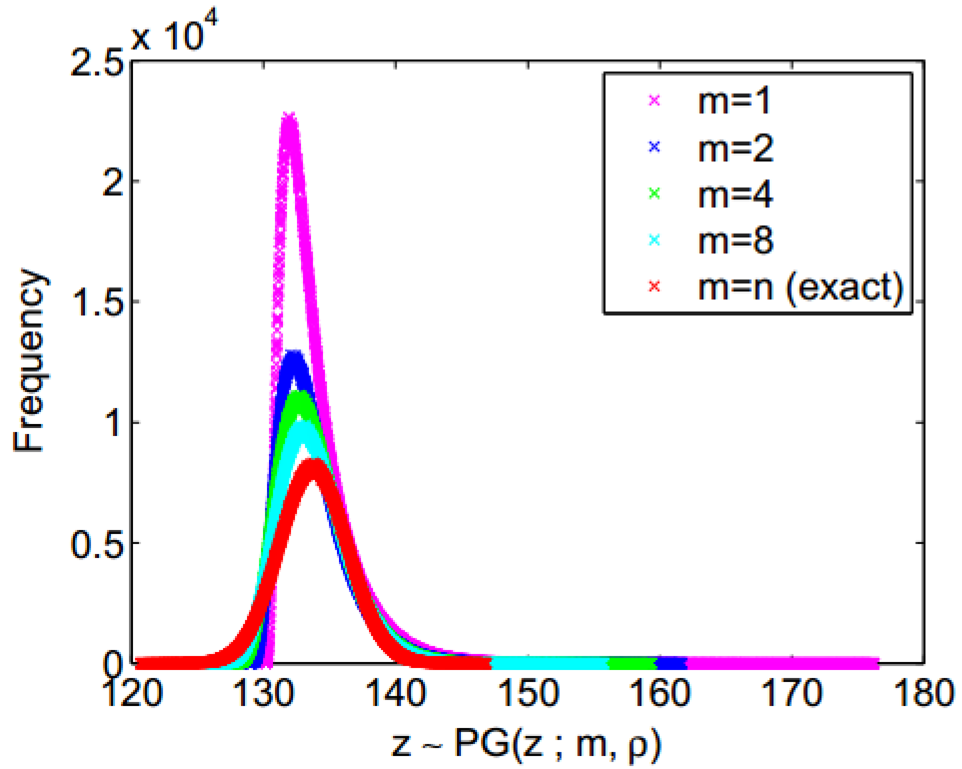
$$p(\eta_d^k | \boldsymbol{\eta}_d^{\neg k}, \mathbf{Z}, \mathbf{W}, \lambda_d^k) = \mathcal{N}(\gamma_d^k, (\tau_d^k)^2)$$

  - Draw *1d* Polya-Gamma:

$$p(\lambda_d^k | \mathbf{Z}, \mathbf{W}, \boldsymbol{\eta}) = \mathcal{PG}(\lambda_d^k; N_d, \rho_d^k)$$

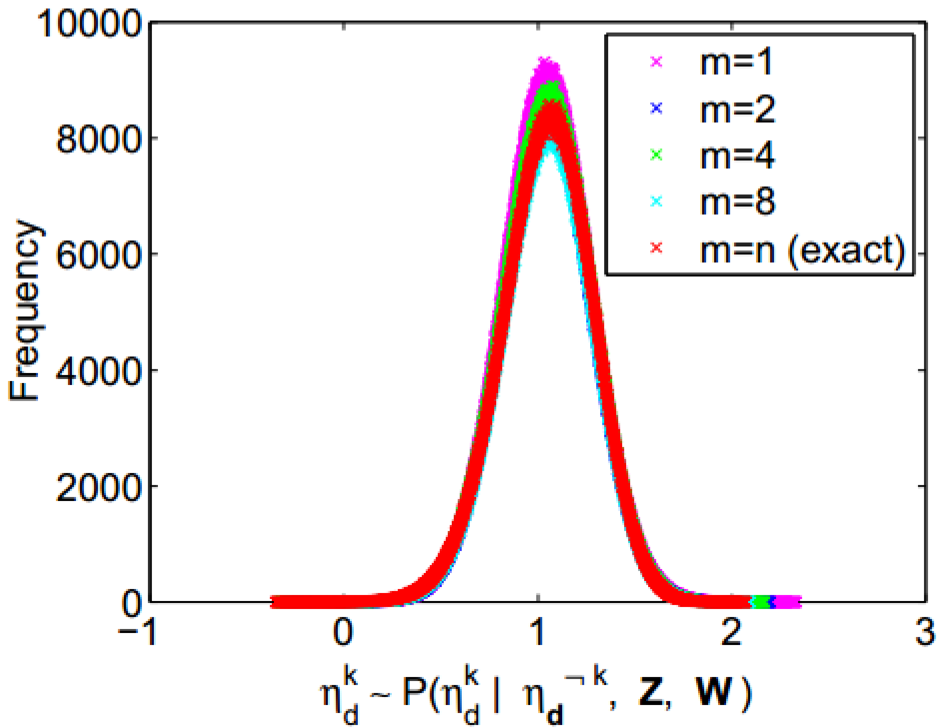
Fast Approximation by CLT

- Using a few samples to approximate:



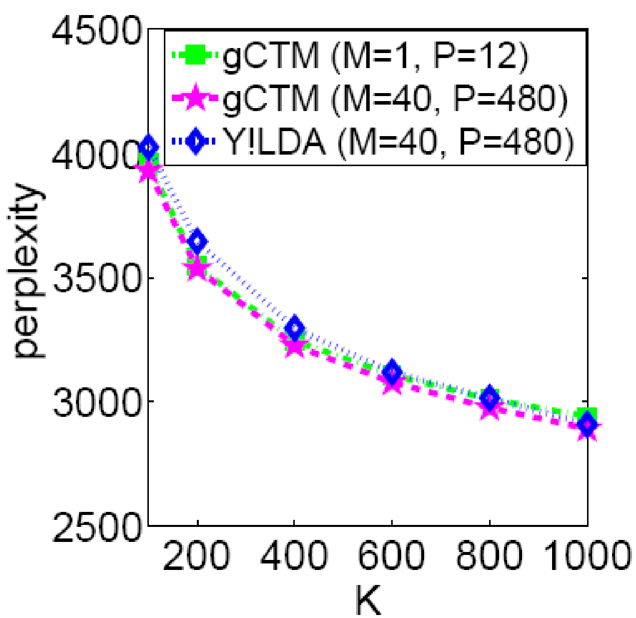
Fast Approximation by CLT

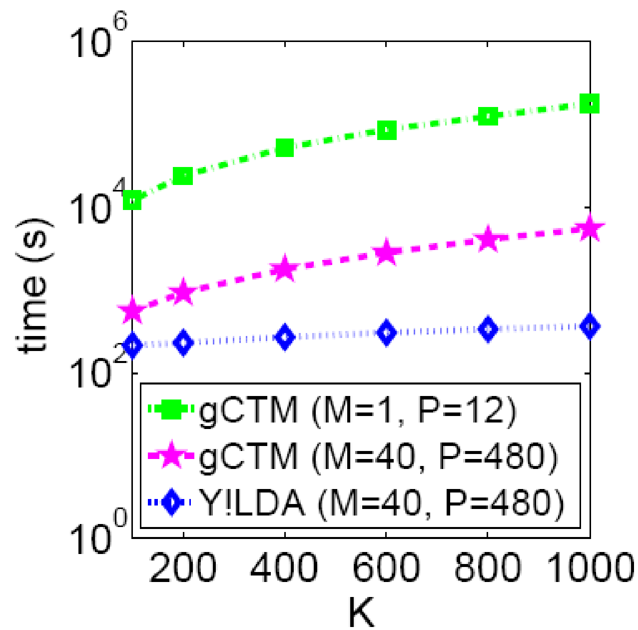
- Using a few samples to approximate:



Experimental Results

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





- 40 machines;
- 480 CPU cores

- 0.285M NYTimes pages
- K = 200 ~ 1000

[Chen, Zhu, Wang, Zheng, & Zhang, NIPS 2013]

Experimental Results

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

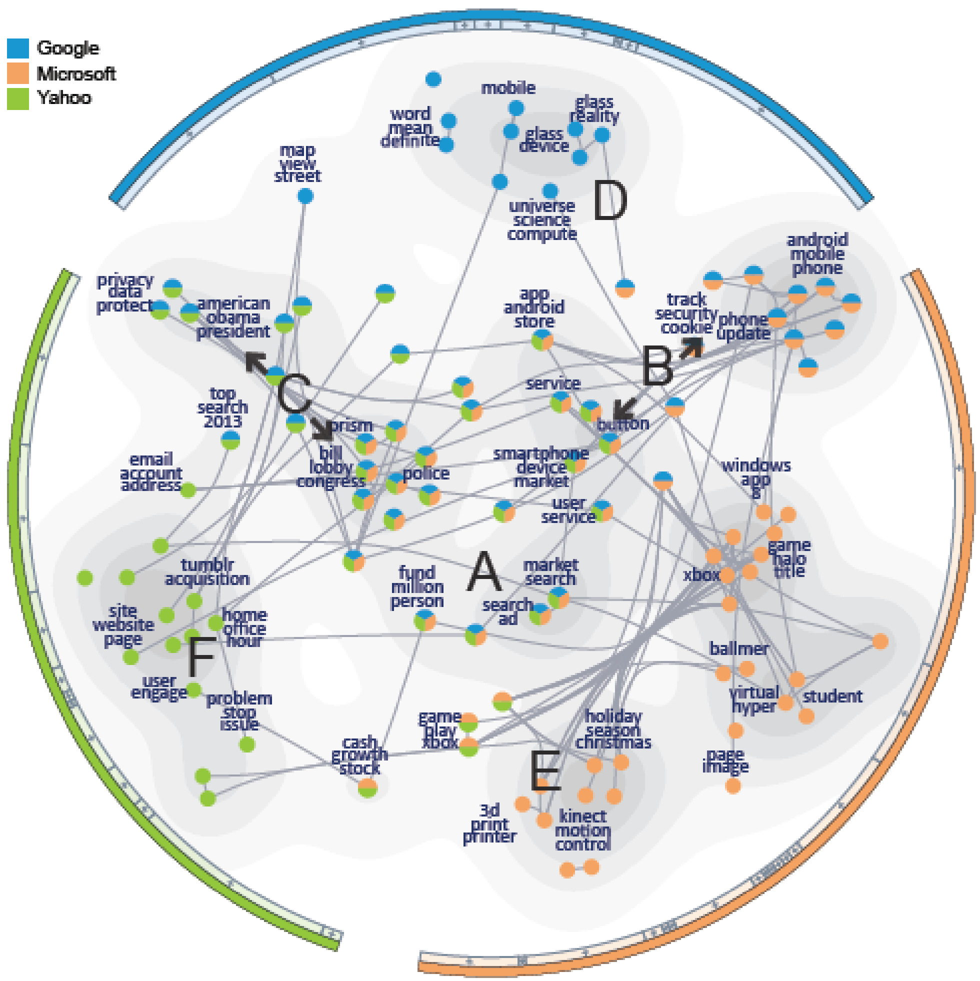
data set	D	K	vCTM	gCTM
NIPS	1.2K	100	1.9 hr	8.9 min
20NG	11K	200	16 hr	9 min
NYTimes	285K	400	N/A*	0.5 hr
Wiki	6M	1000	N/A*	17 hr

*not finished within 1 week.

[Chen, Zhu, Wang, Zheng, & Zhang, NIPS 2013]

清華大學 Tsinghua University

Scalable Graph Visualization



[Joint with Dr. Shixia Liu from MSRA. IEEE VAST 2014]

Summary

- Computationally efficient Bayesian models are becoming increasingly relevant in Big data era

- RegBayes:
  - bridges Bayesian methods, learning and optimization
  - offers an extra freedom to incorporate rich side information

- Many scalable algorithms have been developed
  - online/stochastic algorithms (e.g., online BayesPA)
  - distributed inference algorithms (e.g., scalable CTM)

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

Further Readings

- Stochastic MCMC Algorithms:
  - Bayesian Learning via Stochastic Gradient Langevin Dynamics, M. Welling and Y. W. Teh, ICML 2011;
  - Bayesian Posterior Sampling via Stochastic Gradient Fisher Scoring, S. Ahn, A. Korattikara, and M. Welling, ICML 2012 (best paper);
  - Stochastic Gradient Riemannian Langevin Dynamics on the Probability Simplex, S. Patterson and Y. W. Teh, NIPS 2013;

- Distributed MCMC Algorithms:
  - Asymptotically Exact, Embarrassingly Parallel MCMC, N., Willie, C. Wang, and E. Xing, UAI 2014;
  - Distributed Stochastic Gradient MCMC, S. Ahn, B. Shahbaba and M. Welling, ICML 2014;

Further Readings

- Stochastic Variational Algorithms:
  - Hoffman, M., Bach, F.R., and Blei, D.M. Online learning for latent Dirichlet allocation. NIPS, 2010.
  - Mimno, D., Hoffman, M., and Blei, D.M. Sparse stochastic inference for latent dirichlet allocation. ICML, 2012.

- Distributed Algorithms for Topic Models:
  - A. Ahmed, M. Aly, J. Gonzalez, S. Narayanamurthy, and A. Smola. Scalable inference in latent variable models. WSDM, 2012;
  - A. Smola and S. Narayanamurthy. An architecture for parallel topic models. VLDB, 3(1-2):703–710, 2010;
  - D. Newman, A. Asuncion, P. Smyth, and M. Welling. Distributed algorithms for topic models. Journal of Machine Learning Research (JMLR), (10):1801–1828, 2009.

Further Readings

- Related Publications in My Group:
  - J. Chen, J. Zhu, Z. Wang, X. Zheng, and B. Zhang. Scalable Inference for Logistic-Normal Topic Models, NIPS, 2013;
  - J. Zhu, X. Zheng, L. Zhou, and B. Zhang. Scalable Inference in Max-margin Supervised Topic Models, KDD, 2013;
  - J. Zhu, X. Zheng, and B. Zhang. Bayesian Logistic Supervised Topic Models with Data Augmentation, ACL, 2013;
  - J. Zhu, N. Chen, and E.P. Xing. Bayesian Inference with Posterior Regularization and applications to Infinite Latent SVMs, JMLR, 15(May):1799-1847, 2014
  - J. Zhu, A. Ahmed, and E.P. Xing. MedLDA: maximum margin supervised topic models. JMLR, 13:2237–2278, 2012;
  - J. Zhu, N. Chen, H. Perkins, and B. Zhang. Gibbs Max-margin Topic Models with Data Augmentation, JMLR, 15(Mar):1073-1110, 2014
  - T. Shi, and J. Zhu. Online Bayesian Passive Aggressive Learning, ICML, Beijing, China, 2014;
  - S. Mei, J. Zhu, and J. Zhu. Robust RegBayes: Selectively Incorporating First-Order Logic Domain Knowledge into Bayesian Models, ICML, Beijing, China, 2014;
  - S. Liu, Xi. Wang, J. Chen, J. Zhu, and B. Guo. TopicPanaroma: a Full Picture of Relevant Topics, To Appear in Proc. of IEEE VAST, Paris, France, 2014;

Further Readings

- Tutorials on Big Learning
  - ICML 2014: Bayesian Posterior Inference in the Big Data Arena, Max Welling;
  - ICML 2014: Emerging Systems for Large-Scale Machine Learning, Joseph Gonzalez;
  - AAAI 2014: Scalable Machine Learning, Alex Smola;

- Workshops on Big Learning:
  - NIPS Workshop 2011, 2012, 2013 (**http://www.biglearn.org**)

Acknowledgements

- Collaborators:
  - Prof. Bo Zhang (Tsinghua)、Prof. Eric P. Xing (CMU)、Prof. Li Fei-Fei (Stanford), Prof. Xiaojin Zhu (Wisconsin)

- Students at Tsinghua:
  - Minjie Xu, Jianfei Chen Aonan Zhang Hugh Perkins Bei Chen, Tian Tian, Shike Mei, Xun Zheng, Wenbo Hu, Yining Wang, Zi Wang, Chengtao Li, Yang Gao, Tianlin Shi, Jingwei Zhuo, Chang Liu, Chao Du, Chongxuan Li, Yong Ren, Jianxin Shi, Fei Xia, etc.

- Funding:

Thanks!

Some code available at:

http://bigml.cs.tsinghua.edu.cn/~jun