Bayesian Methods (I)

Jun Zhu
dcszj@mail.tsinghua.edu.cn
http://bigml.cs.tsinghua.edu.cn/~jun
State Key Lab of Intelligent Technology & Systems
Tsinghua University

May 5, 2015

Basic Rules of Probability

- Concepts

$p(X)$	probability of X
$p(X\|\mathcal{M})$	conditional probability of X given $\mathcal{M}$
$p(X,\mathcal{M})$	joint probability of X and $\mathcal{M}$

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

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

- Prediction:

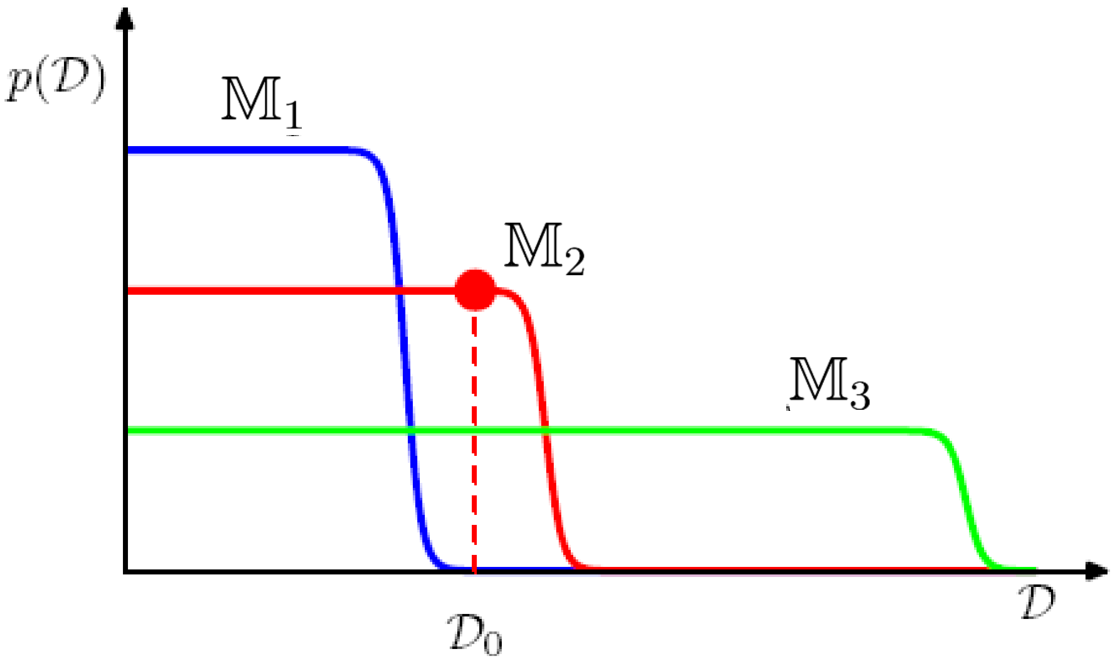
$$p(x|\mathcal{D}, \mathbb{M}) = \int p(x|\mathcal{M}, \mathcal{D}, \mathbb{M})p(\mathcal{M}|\mathcal{D}, \mathbb{M})d\mathcal{M}$$

$$\Big\Downarrow \text{ under some common assumptions}$$

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

Bayesian Model Selection

- Naturally considers model complexity penalty – **no overfitting**



- See details in (C. Bishop, 2006).

Common Questions

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

Why be Bayesian?

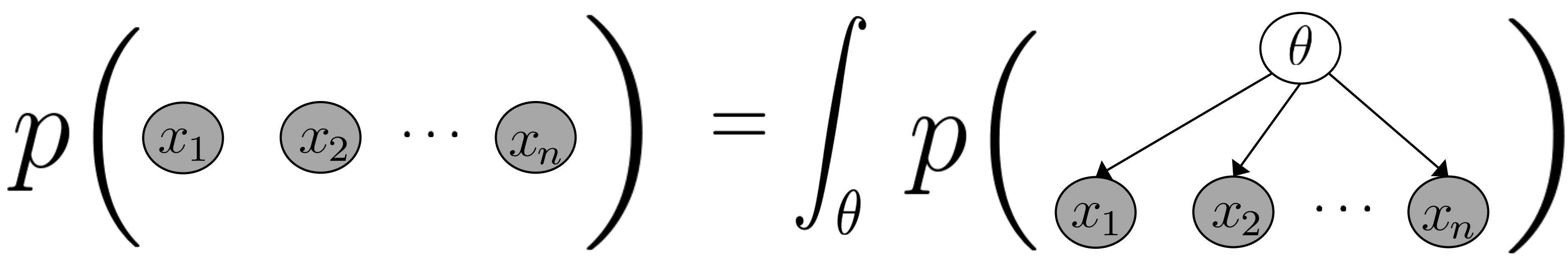
- One of many answers
- Infinite Exchangeability:

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

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

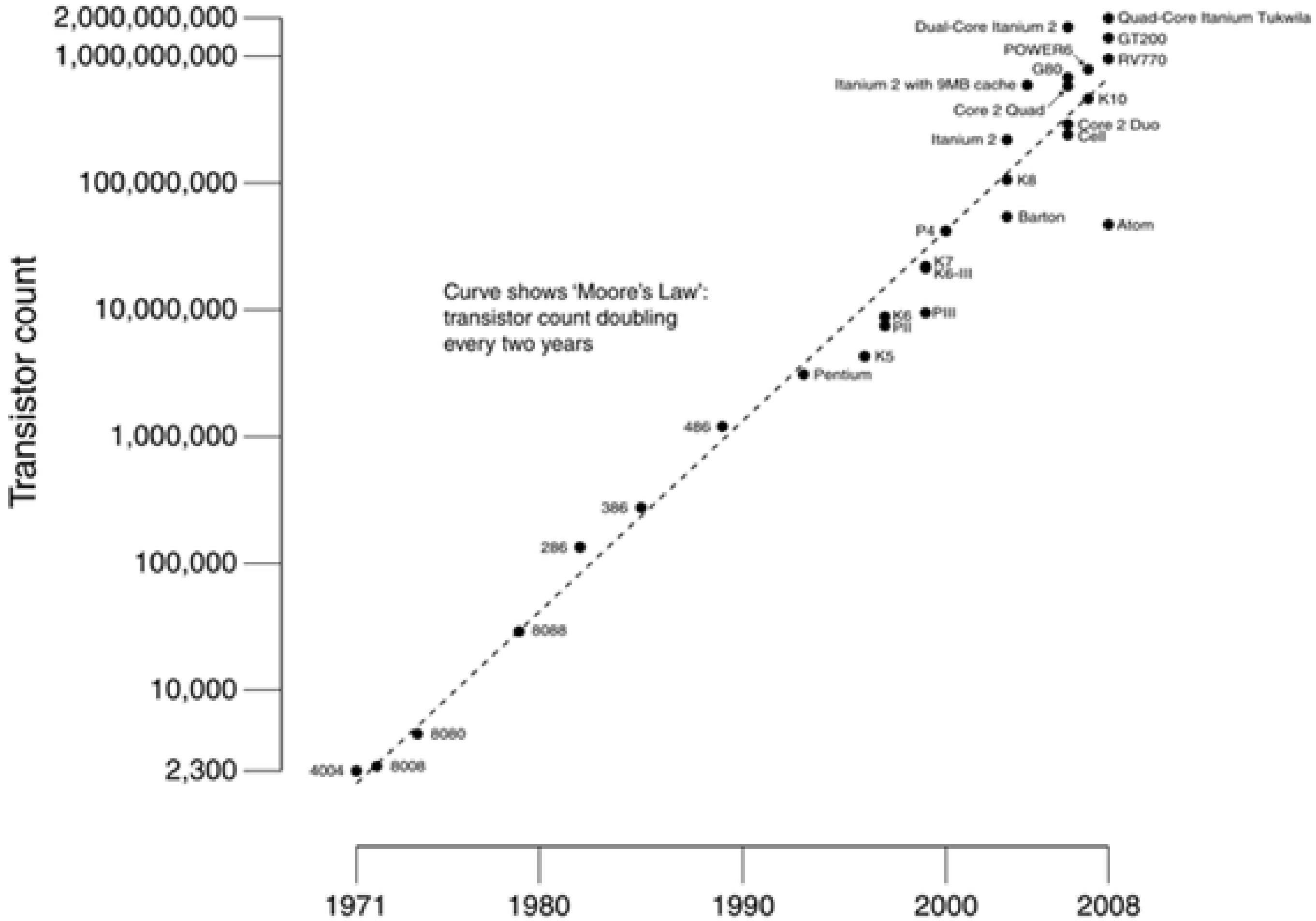
$$p(x_1, \ldots, x_n) = \int \left(\prod_{i=1}^{n} p(x_i|\theta) \right) dP(\theta)$$

for some random variable θ



Overfitting in Big Data

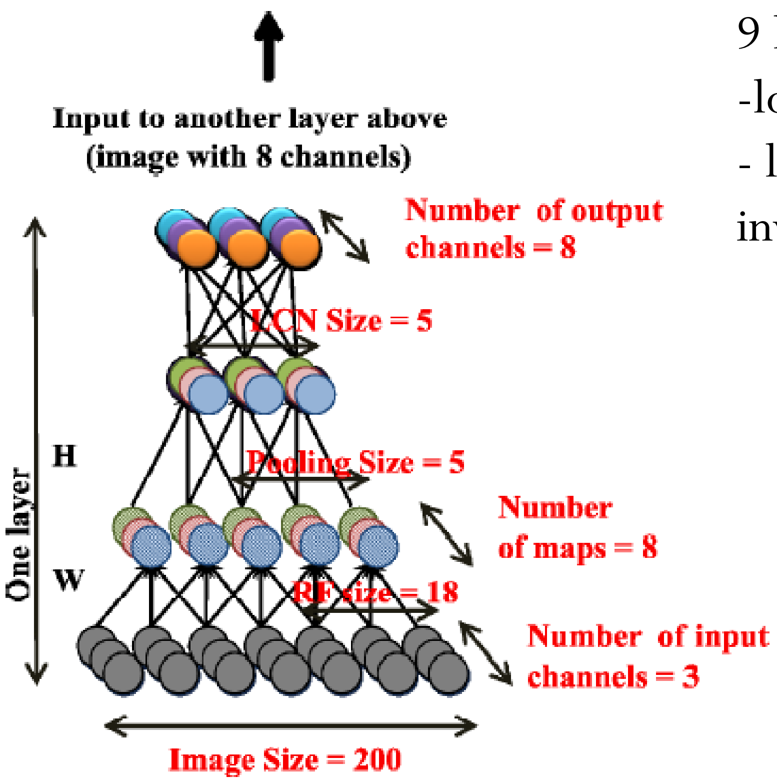
- "with more data overfitting is becoming less of a concern"?



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

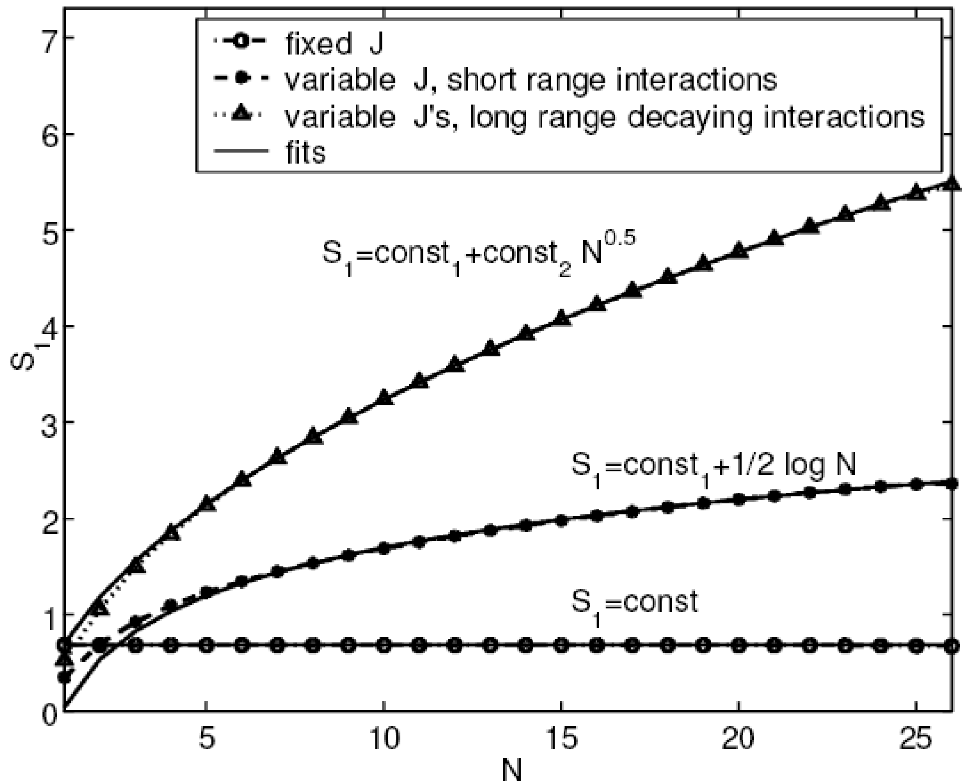
- 1B parameters (connections)
- 10M 200x200 images
- train with 1K machines (16K cores) for 3 days

-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

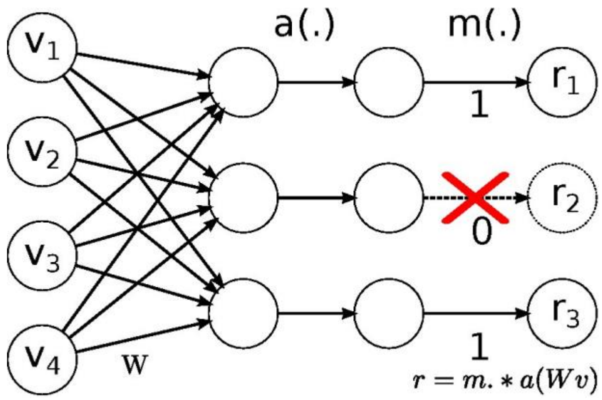
- **Predictive information** grows slower than the amount of Shannon entropy (Bialek et al., 2001)

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?

IEEE SPECTRUM

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

Readings

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

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

$$
\begin{aligned}
p(\pi|X) &\propto p(X|\pi)p(\pi) \\
&\propto (\pi_1^{x_1} \cdots \pi_K^{x_K})(\pi_1^{\alpha_1 - 1} \cdots \pi_K^{\alpha_K - 1}) \\
&= \text{Dirichlet}(\pi_1^{x_1 + \alpha_1 - 1} \cdots \pi_K^{x_K + \alpha_K - 1})
\end{aligned}
$$

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

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

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

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

  - 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

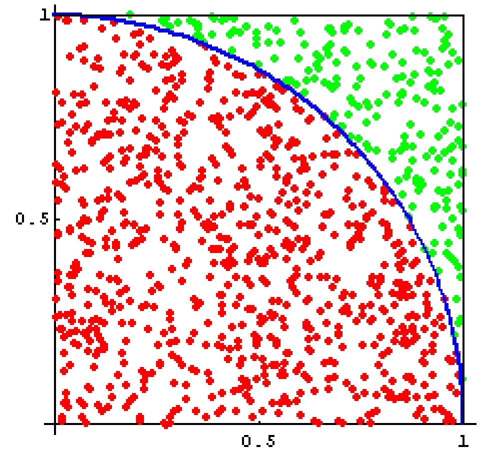
Basics of Monte Carlo Methods

- 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

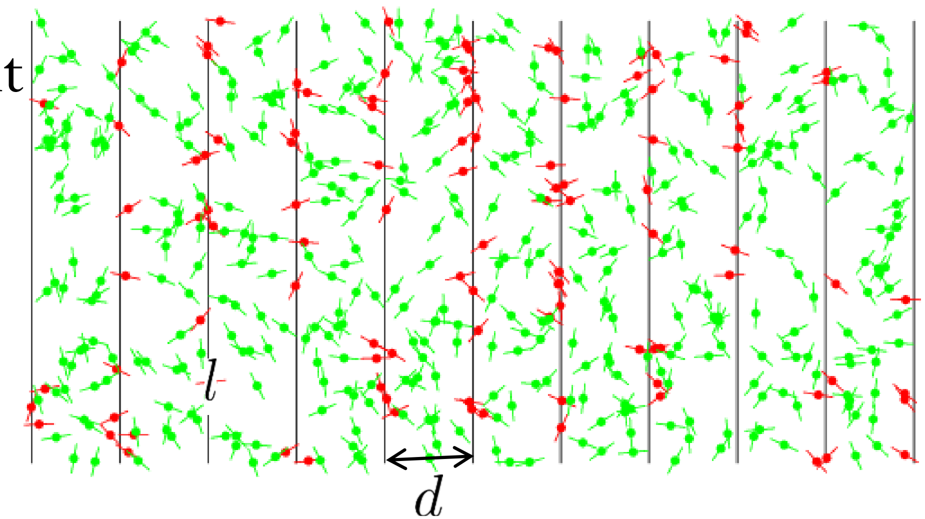
- Computer Simulation



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

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

- Bufffon's Needle Experiment



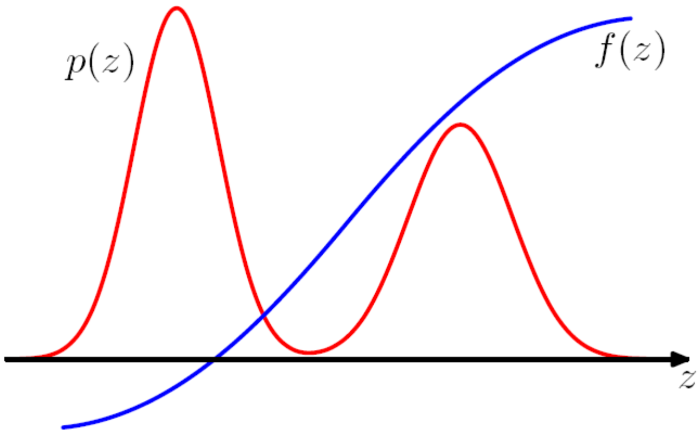
$$\hat{\pi} = \frac{2Nx}{m}$$

  - 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.

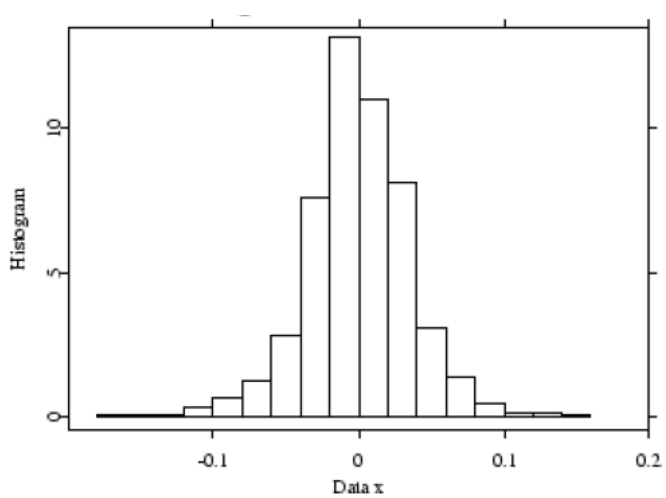
- Draw a set of independent samples (a hard problem)

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

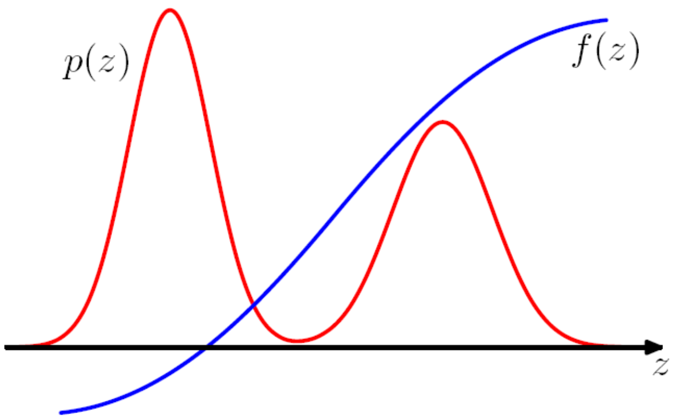
- Estimate the target distribution as count frequency

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

Histogram with Unique Points as the Bins



Basic Procedure of Monte Carlo Methods



- Draw a set of independent samples

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

- 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?

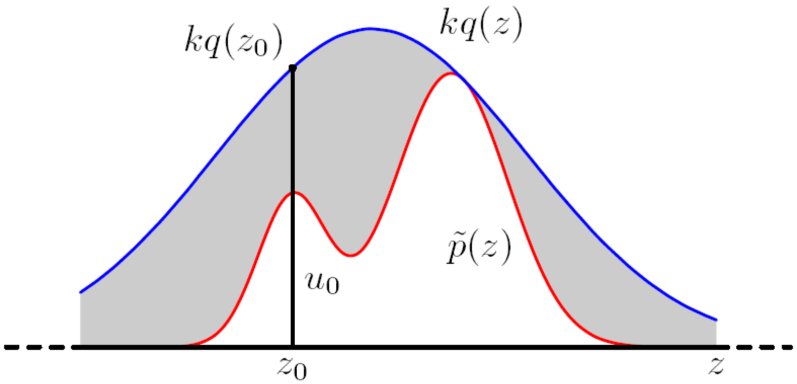
- **Assumption**
  - The target distribution can be evaluated, at least to within a multiplicative constant, i.e.,

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

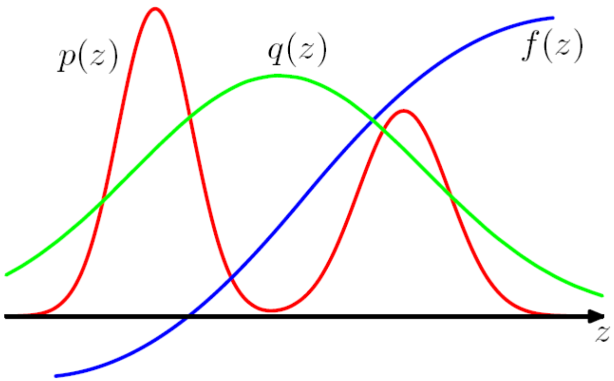
  - where $p^*(\mathbf{z})$ can be evaluated
- **Two difficulties**
  - Normalizing constant is typically unknown
  - Drawing samples in high-dimensional space is challenging

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

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

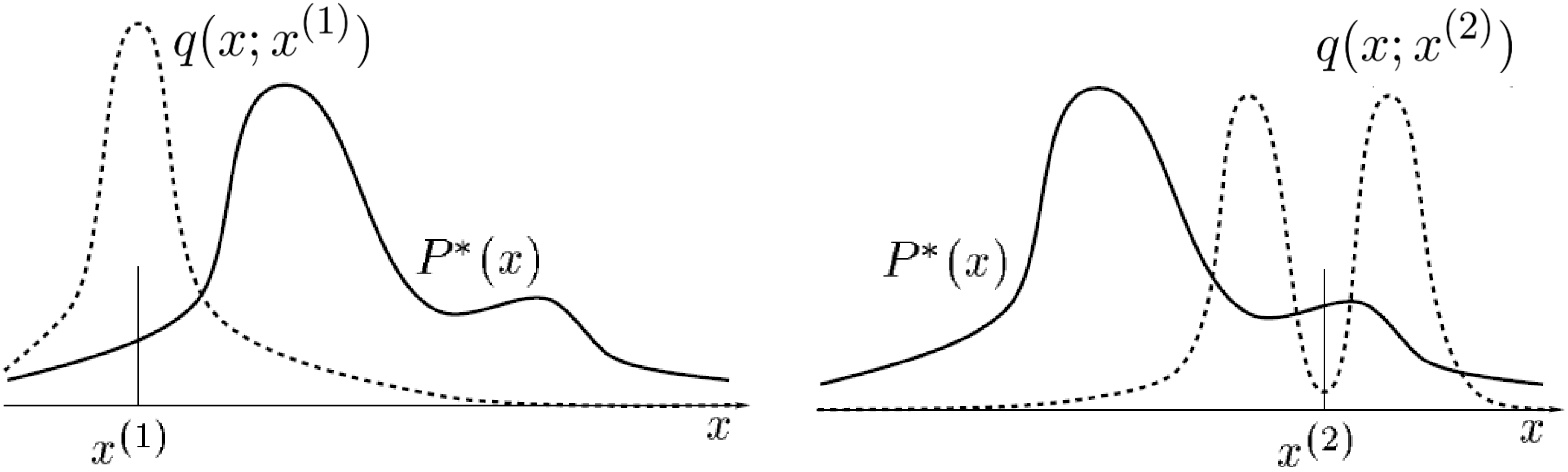
  - where $$p_t(x) = \int p_{t-1}(x')q(x;x')dx'$$
  - $q(x;x')$ is the transition kernel

- $p(x|\mathcal{D})$ 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

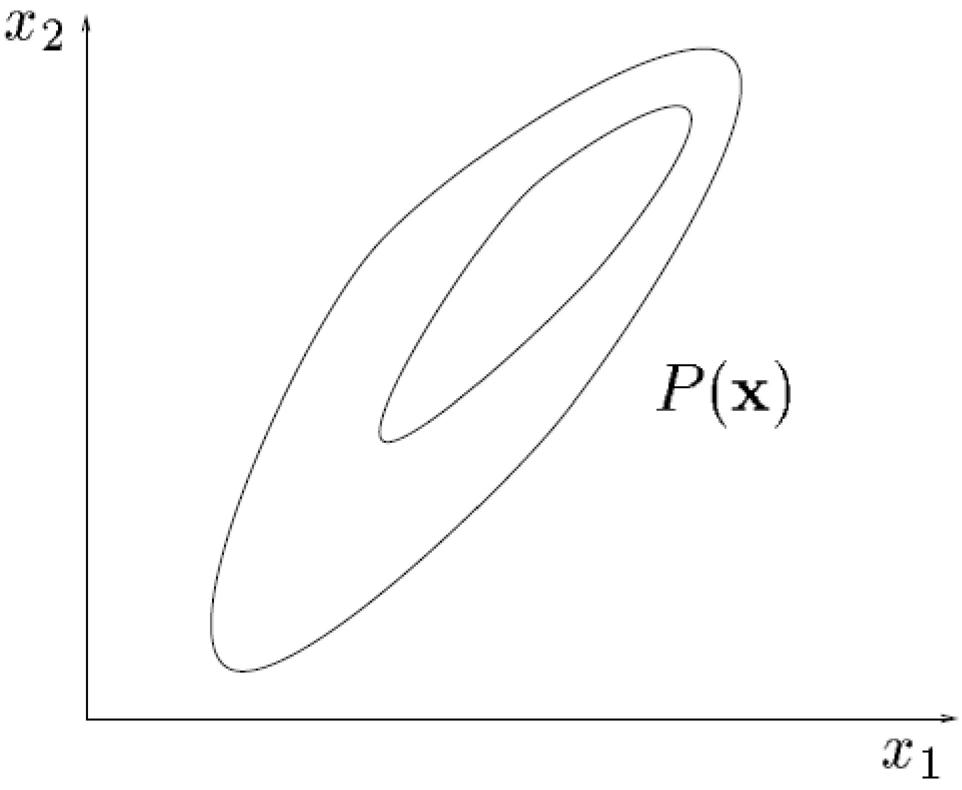
- A special case of Metropolis-Hastings algorithm
- Consider the distribution $p(\mathbf{x}) = p(x_1, \ldots, x_M)$
- Gibbs sampling performs the follows
  - Initialize $\{x_i : i = 1, \ldots, M\}$
  - For $\tau = 1, \ldots, T$
    - Sample $x_1^{(\tau+1)} \sim p(x_1 | x_2^{(\tau)}, x_3^{(\tau)}, \ldots, x_M^{(\tau)})$

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

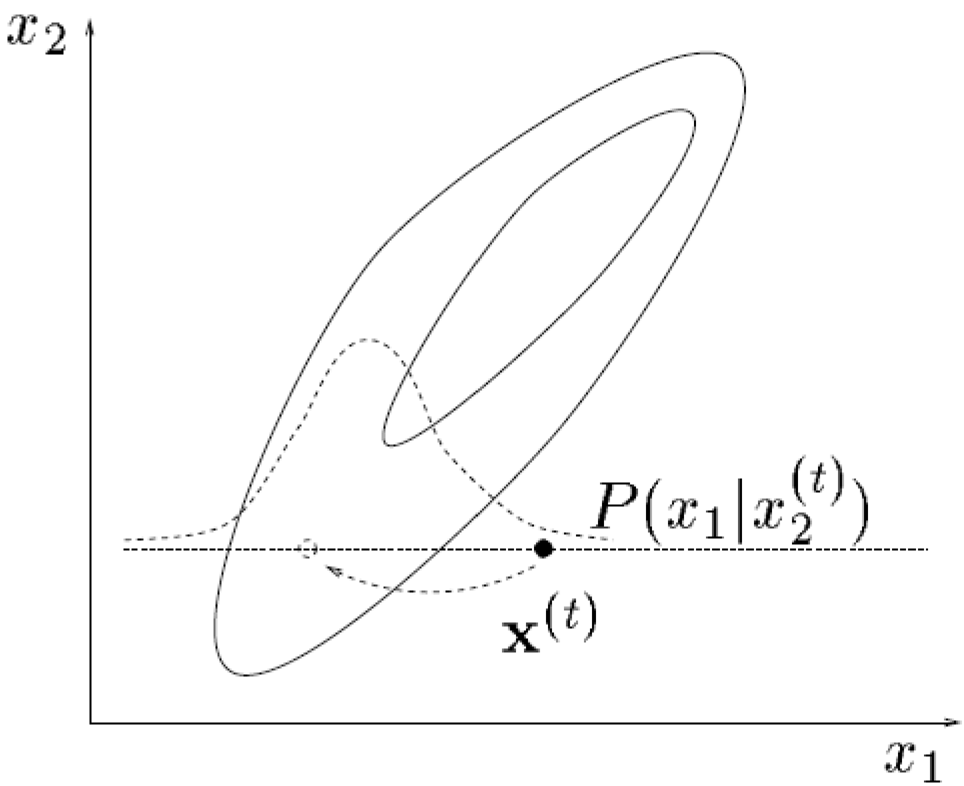
Geometry of Gibbs Sampling

- The target distribution in 2 dimensional space



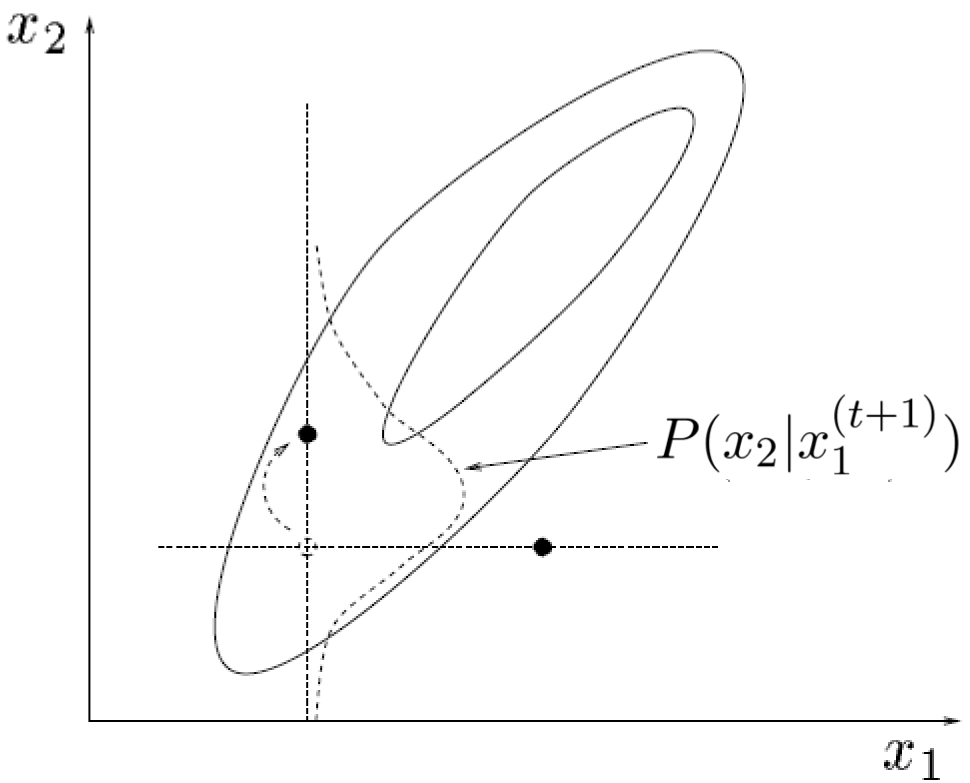
Geometry of Gibbs Sampling

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



Geometry of Gibbs Sampling

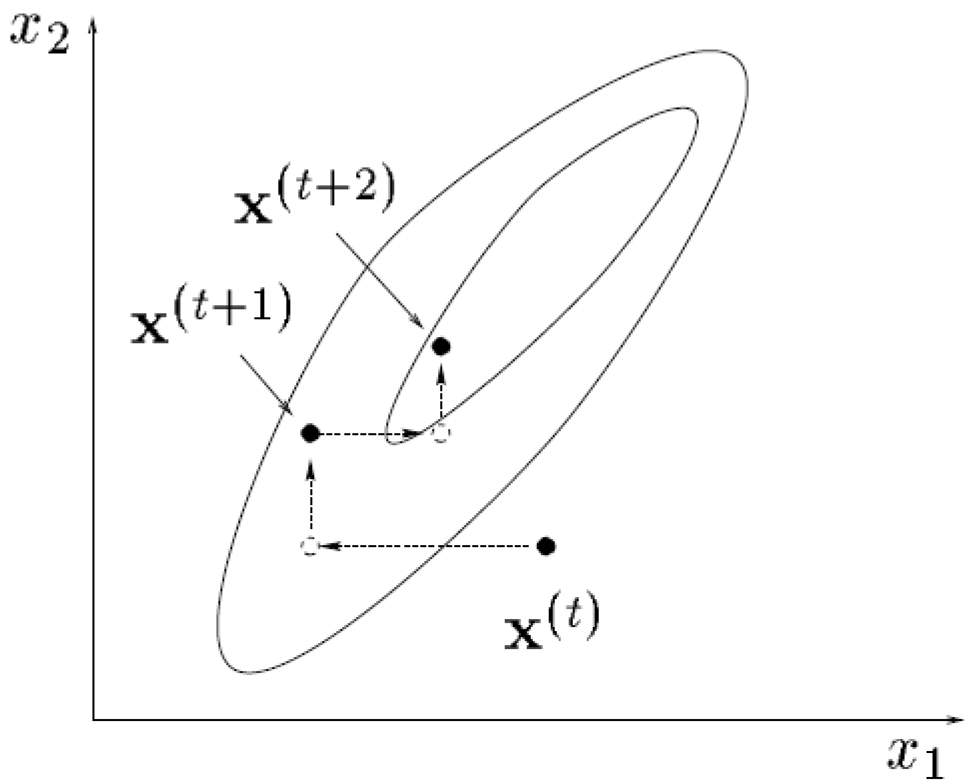
- A sample is drawn from $P(x_2|x_1^{(t+1)})$



this finishes one single iteration.

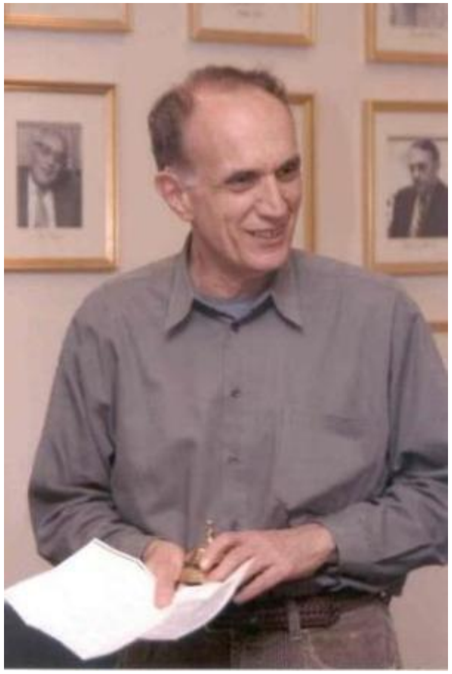
Geometry of Gibbs Sampling

- 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

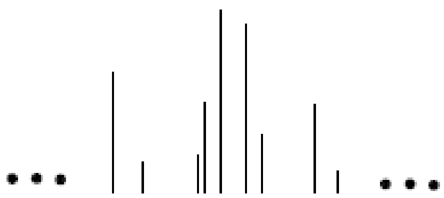
$$p(\mathcal{M}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{M})\pi(\mathcal{M})}{\int p(\mathbf{x}|\mathcal{M})\pi(\mathcal{M})d\mathcal{M}} \propto p(\mathbf{x}|\mathcal{M})\pi(\mathcal{M})$$

Examples:

→ see next slide

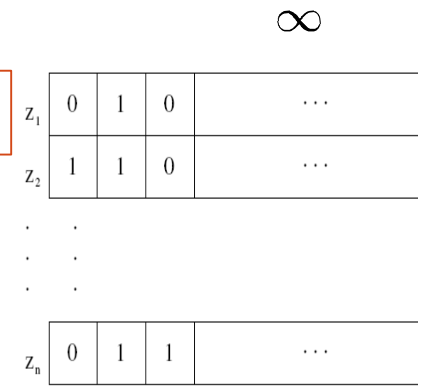
Nonparametric Bayesian Inference

probability measure



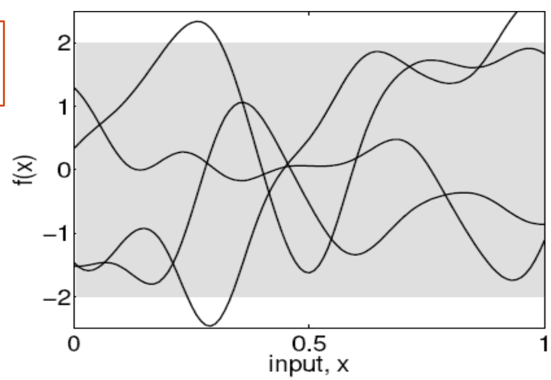
Dirichlet Process Prior [Antoniak, 1974]
+ Multinomial/Gaussian/Softmax likelihood

binary matrix



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

function

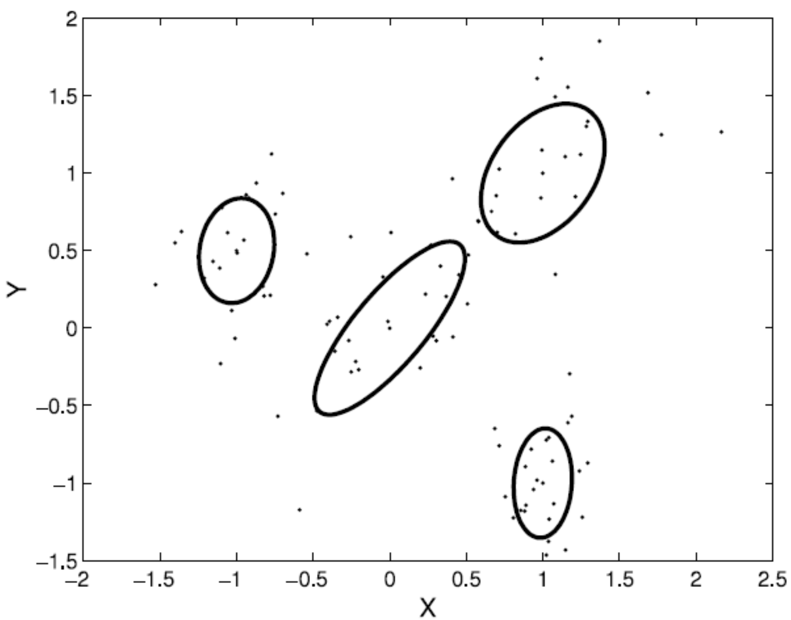


Gaussian Process Prior [Doob, 1944; Rasmussen & Williams, 2006]
+ Gaussian/Sigmoid/Softmax likelihood

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: A Bayesian Ranking Model

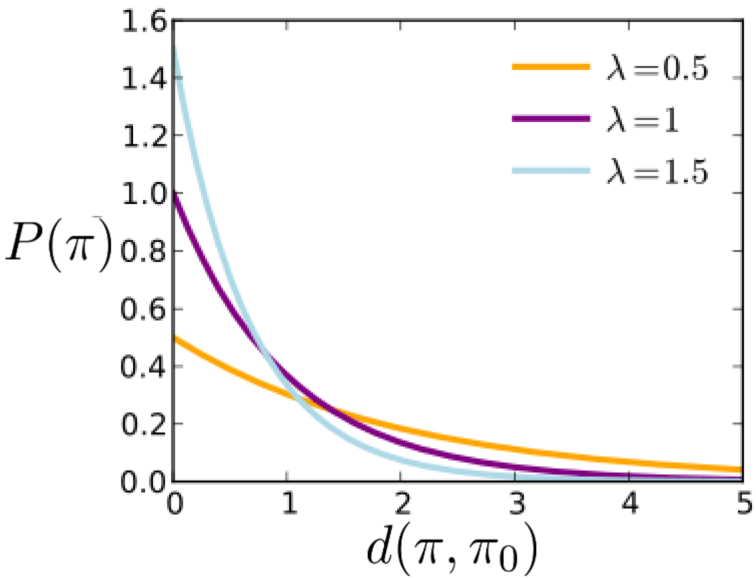
- Rank a set of items, e.g., A, B, C, D
  - A uniform permutation model

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

Example: 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: A Bayesian Ranking Model

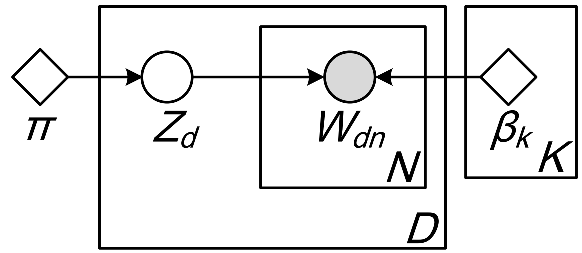
- Rank a set of items
  - Prior knowledge
    - conjugate prior exists for generalized Mallows' models (a member of exponential family)
  - Bayesian updates can be done with Bayes' rule
  - Can be incorporated into a hierarchical Bayesian model, e.g., topic models

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

Topic Models

Homework Example

- Mixture of Multinomials



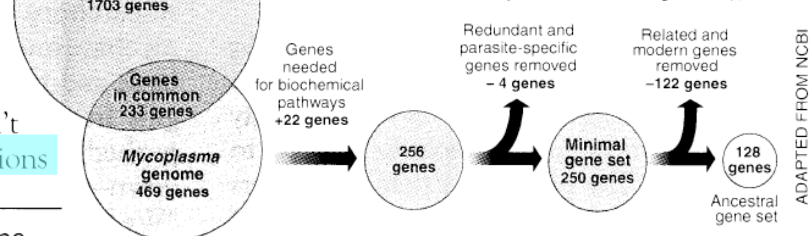
- Assumption:
  - Each document belongs to a single topic

Multiple Topics exist in a Document

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

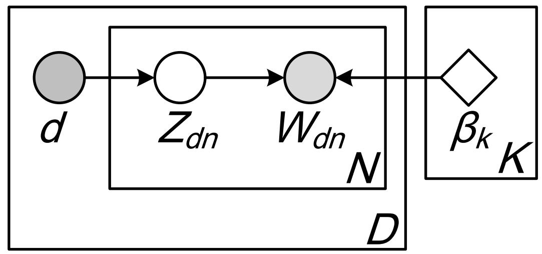
SCIENCE • VOL. 272 • 24 MAY 1996

Simple intuition: Documents exhibit multiple topics.

[Slides courtesy: D. Blei]

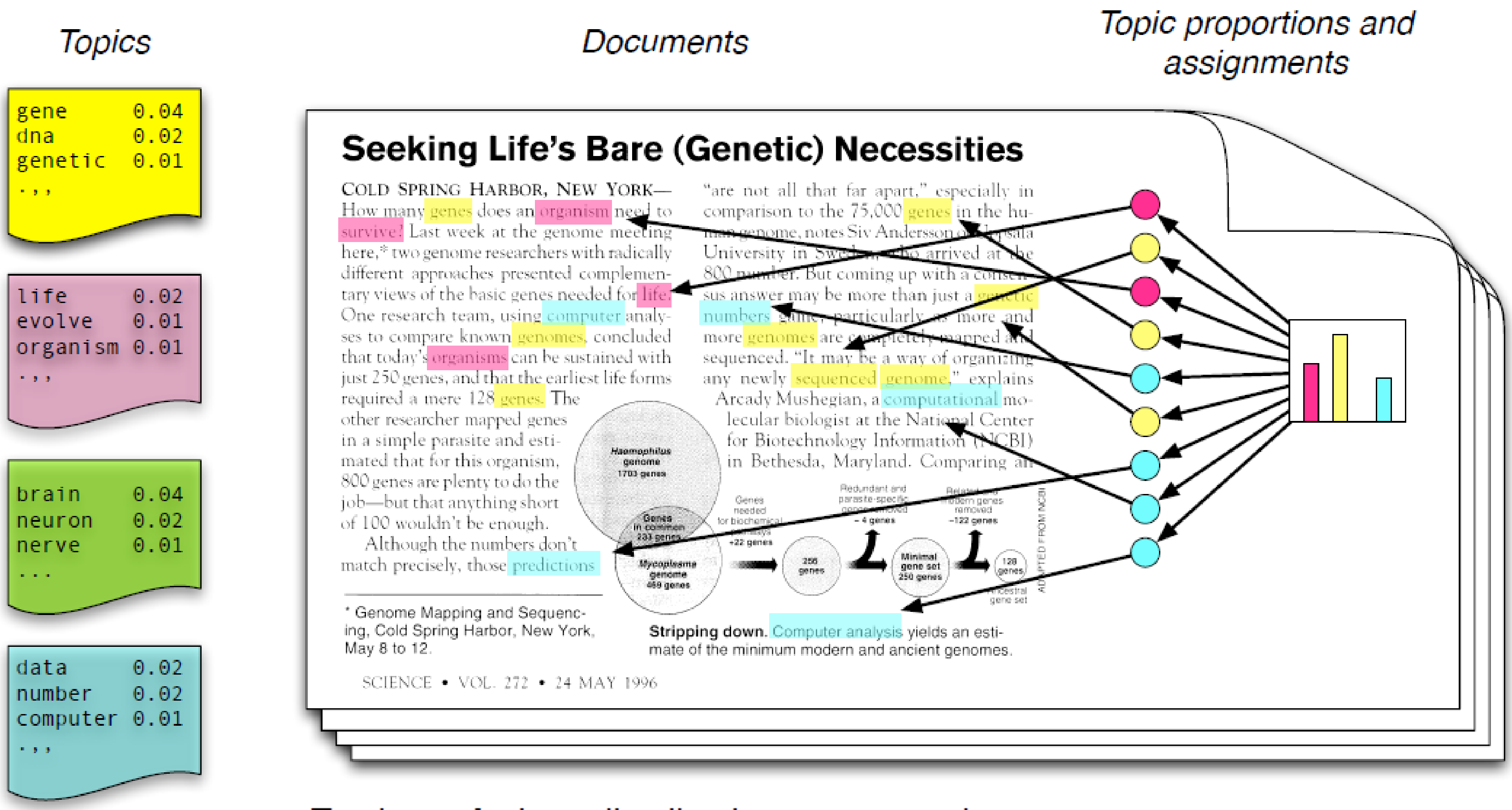
Probabilistic Latent Semantic Indexing

- Allows multiple topics in a document



- Limitations:
  - *d* is a dummy index into the list of documents in training set; no natural generalization to unseen document;
  - # of unknown parameters grows linearly with data size (i.e., $KV + KD$) – overfitting!

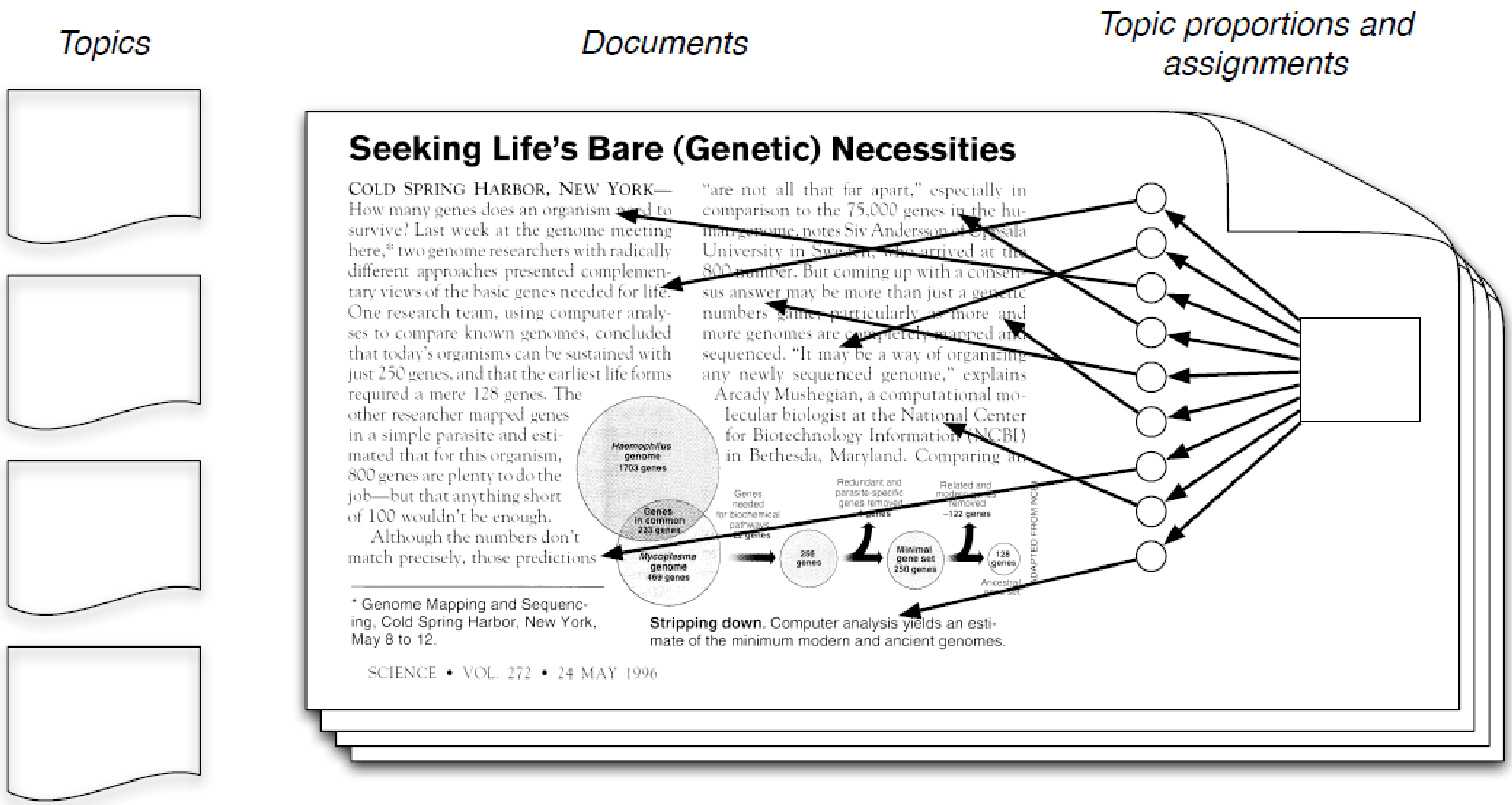
Latent Dirichlet Allocation (LDA)



- Each **topic** is a distribution over words
- Each **document** is a mixture of corpus-wide topics
- Each **word** is drawn from one of those topics

[Slides courtesy: D. Blei]

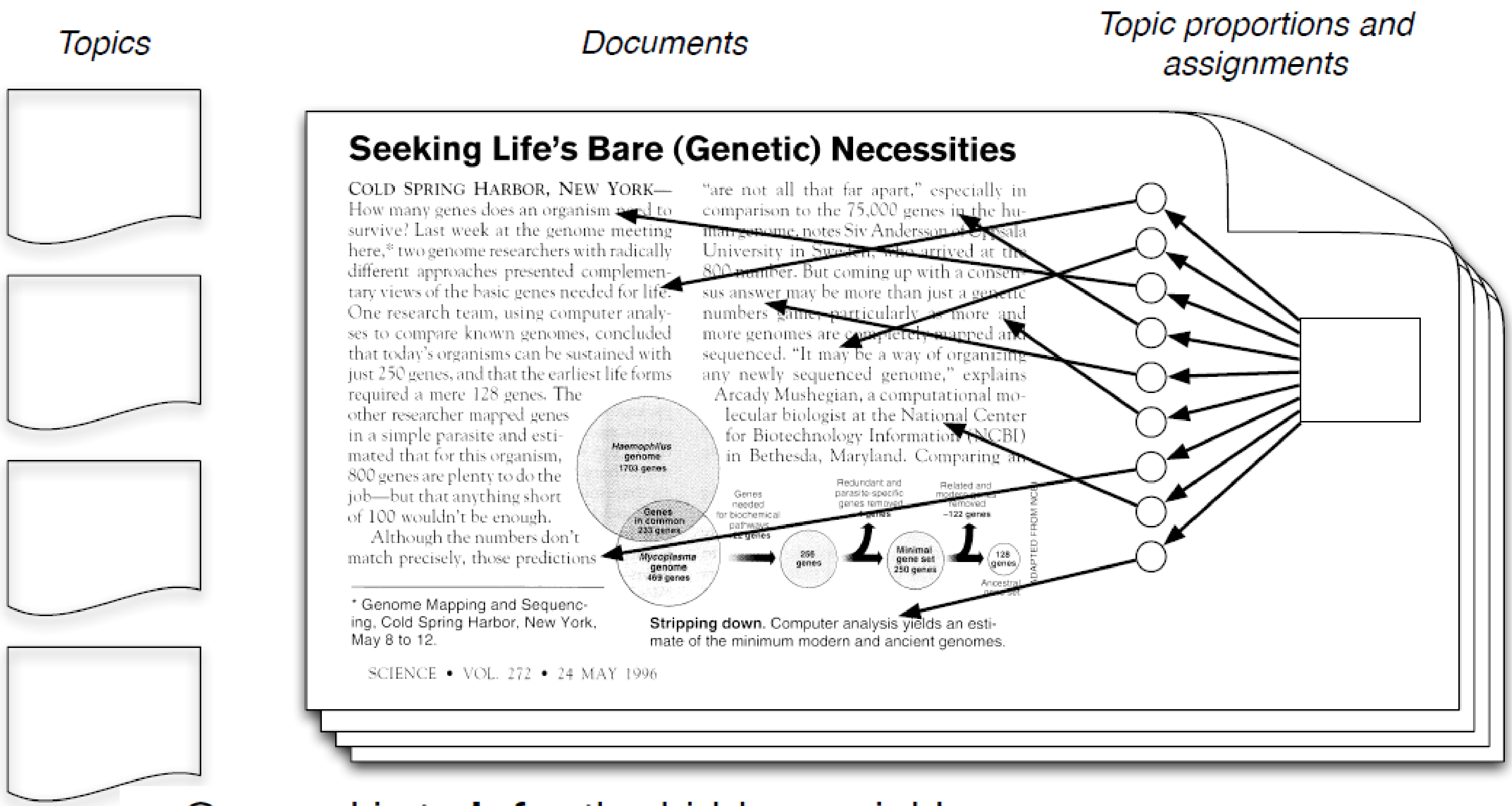
Latent Dirichlet Allocation



- In reality, we only observe the documents
- The other structure are **hidden variables**

[Slides courtesy: D. Blei]

Latent Dirichlet Allocation

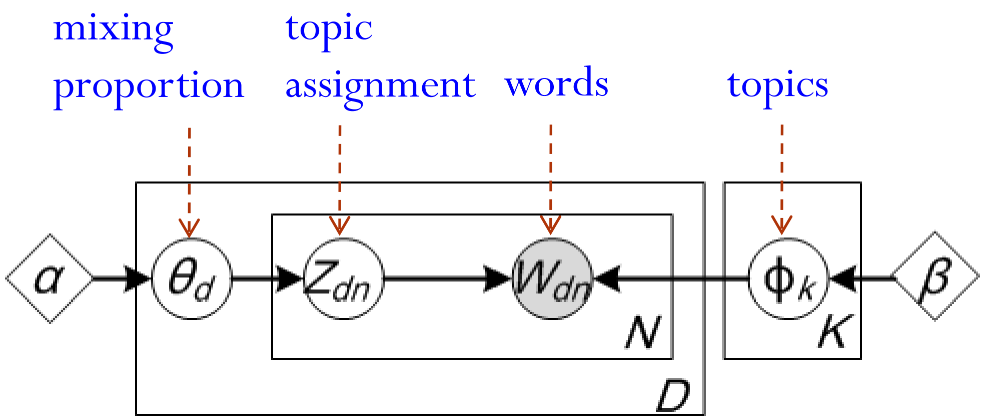


- Our goal is to **infer** the hidden variables
- I.e., compute their distribution conditioned on the documents

$$p(\text{topics, proportions, assignments} \mid \text{documents})$$

[Slides courtesy: D. Blei]

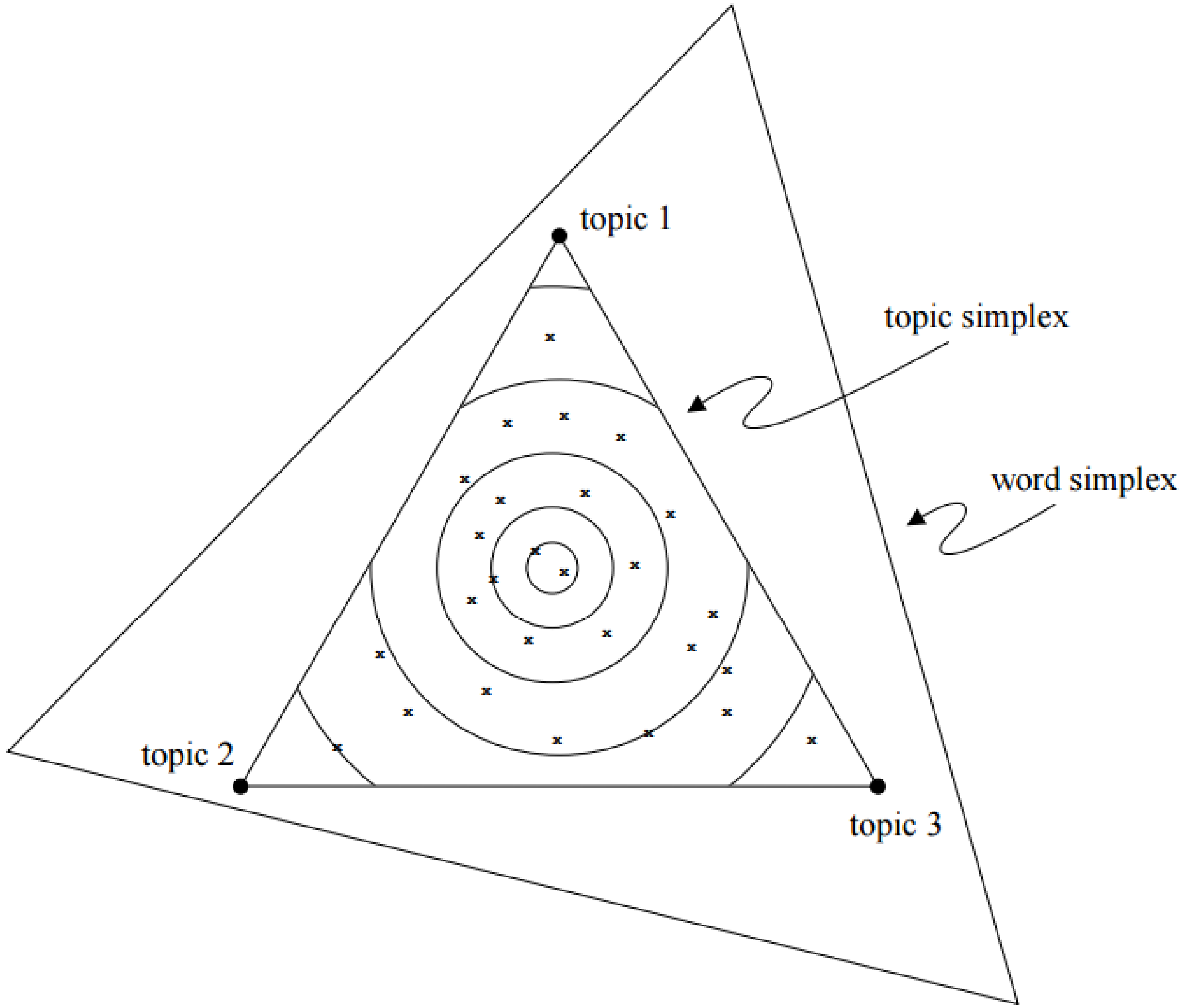
LDA as a graphical model



- Encodes **assumptions**
- Defines a **factorization** of the joint distribution
- Connects to **algorithms** for computing with data

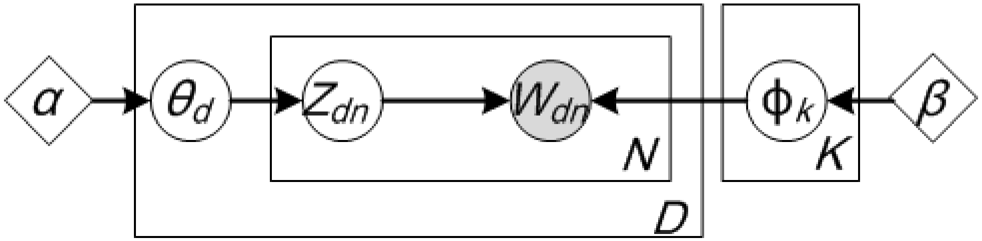
$$p(\Theta, \Phi, \mathbf{Z}, \mathbf{W} | \alpha, \beta) = \prod_{k=1}^{K} p(\Phi_k | \beta) \prod_{d=1}^{D} p(\theta_d | \alpha) \Big(\prod_{n=1}^{N} p(z_{dn} | \theta_d) p(w_{dn} | z_{dn}, \Phi) \Big)$$

A geometric interpretation



[Blei et al., 2003]

LDA as a graphical model



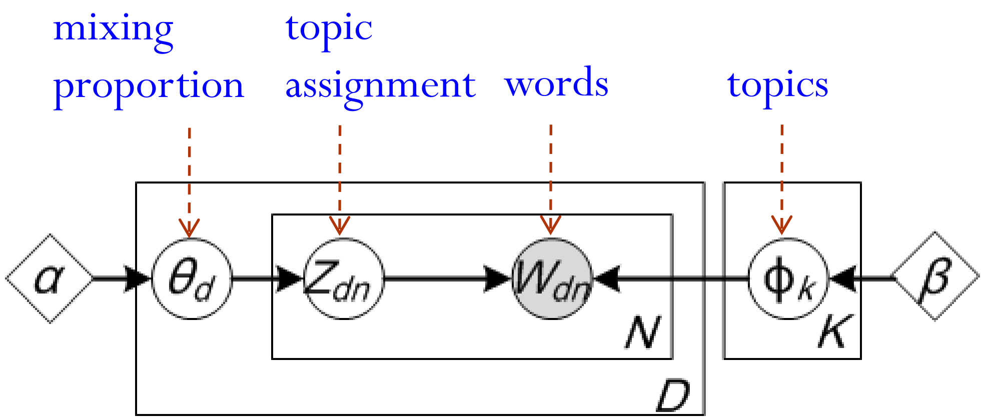
- The joint distribution defines a posterior

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

- From a collection of documents, infer
  - Per-word topic assignment
  - Per-document topic proportion
  - Per-corpus topic distributions

- Then, use posterior expectations to perform the task at hand, e.g., information retrieval, document similarity, exploration, …

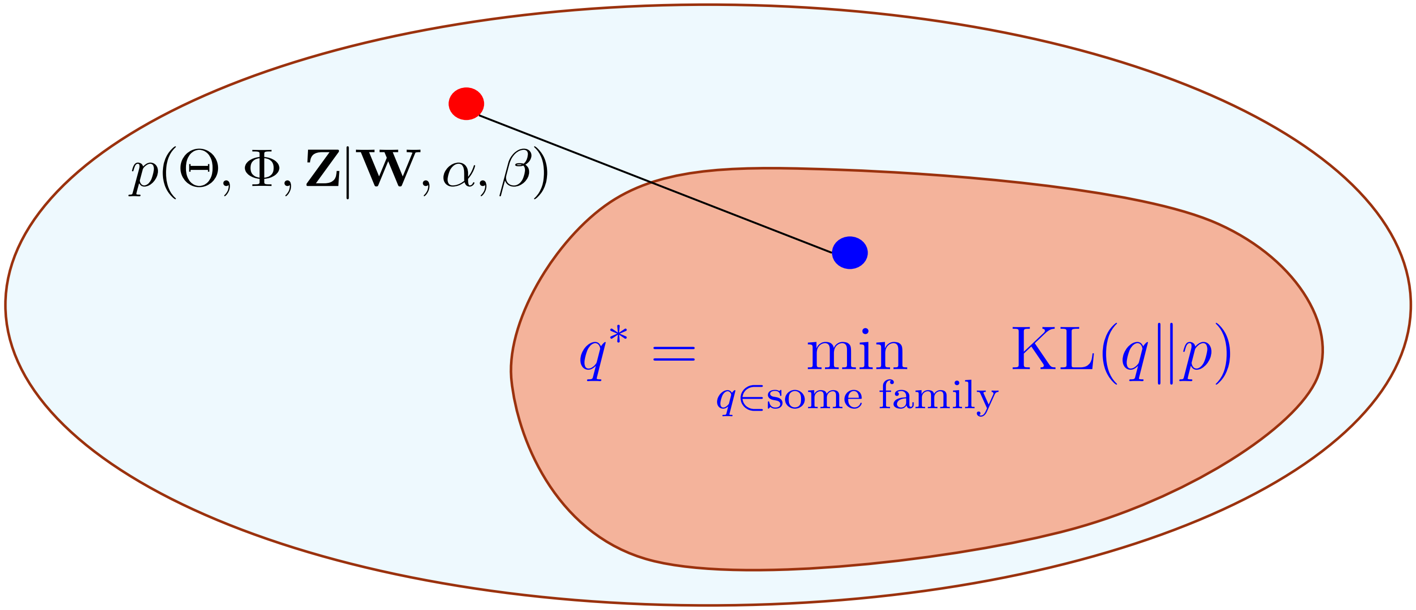
LDA as a graphical model



- Approximate posterior inference algorithms
  - Mean-field variational methods (Blei et al., 2003)
  - Expectation propagation (Minka & Lafferty, 2002)
  - Collapsed Gibbs sampling (Griffiths & Steyvers, 2002)
  - Collapsed variational inference (Teh et al., 2006)
  - Online variational inference (Hoffman et al., 2010)
  - Distributed Gibbs sampling (Ahmed et al., 2012, Yuan et al., 2015)
  - …

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

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

Derivation of Collapsed Gibbs Sampling

- Homework: complete the derivation and implement it

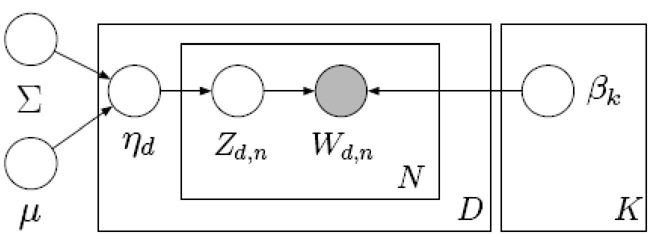
Variational Mean-Field Methods

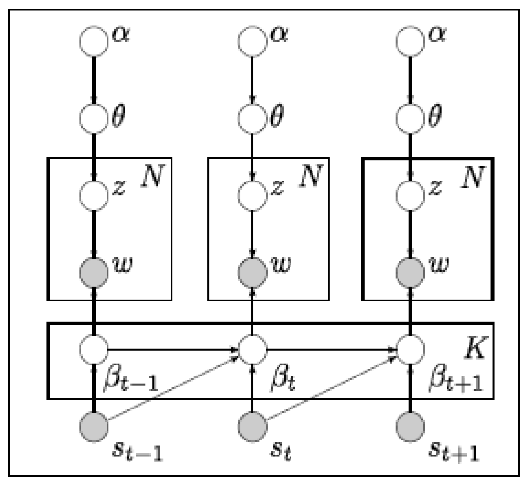
- Details provided in [Blei et al., JMLR 2003, Appendix]

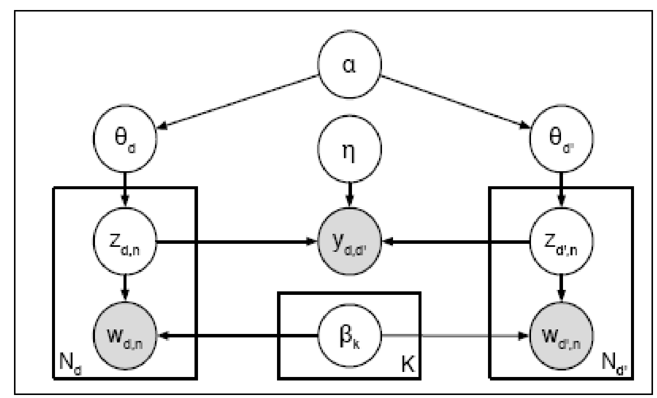
Beyond LDA

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.



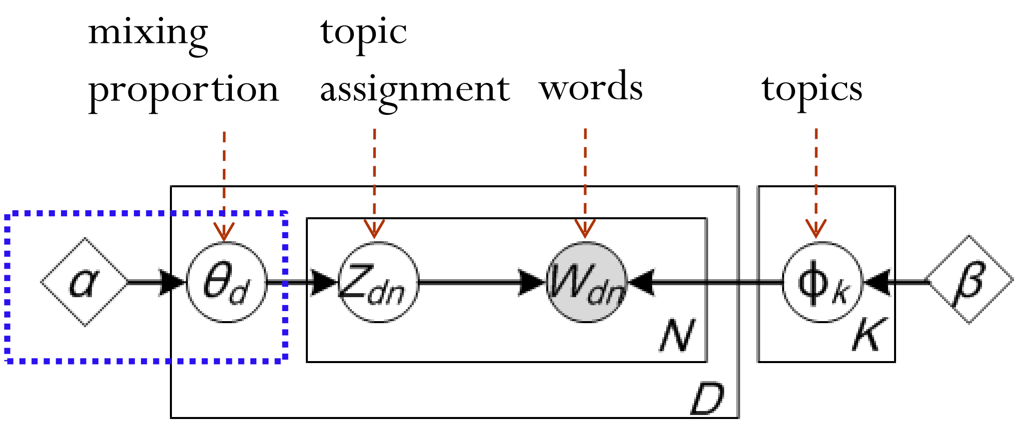




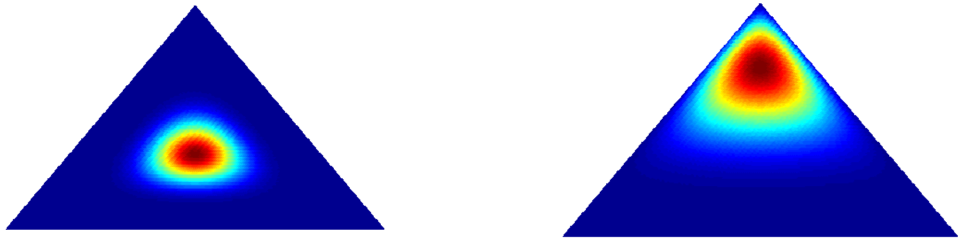
- Tutorials were provide by D. Blei at ICML, SIGKDD, etc. (http://www.cs.princeton.edu/~blei/topicmodeling.html)

Logistic-Normal Topic Models

- Bayesian topic models



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



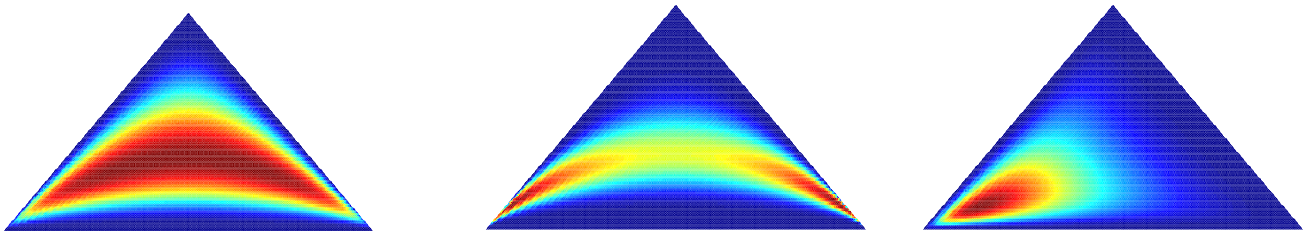
Logistic-Normal Topic Models

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

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

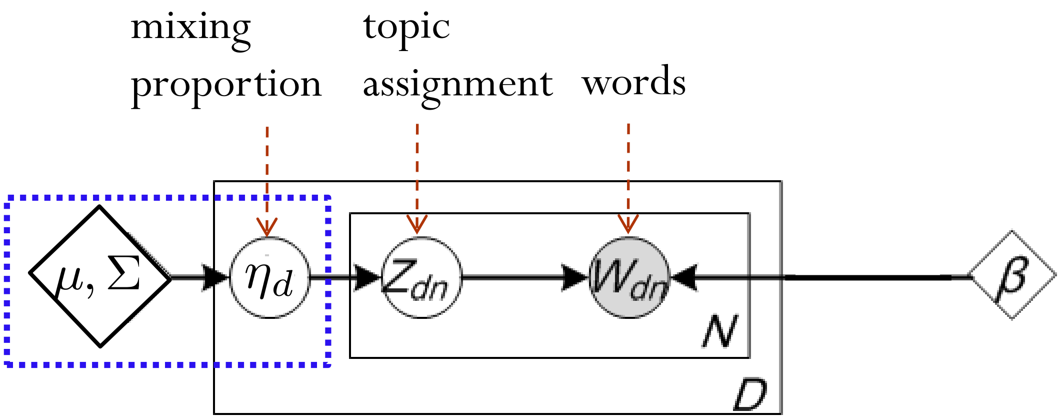
$$\theta_d^k = \frac{\exp\left(\eta_d^k\right)}{\sum_i \exp\left(\eta_d^i\right)}$$

  - Logisitc-normal prior can capture the correlationships



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

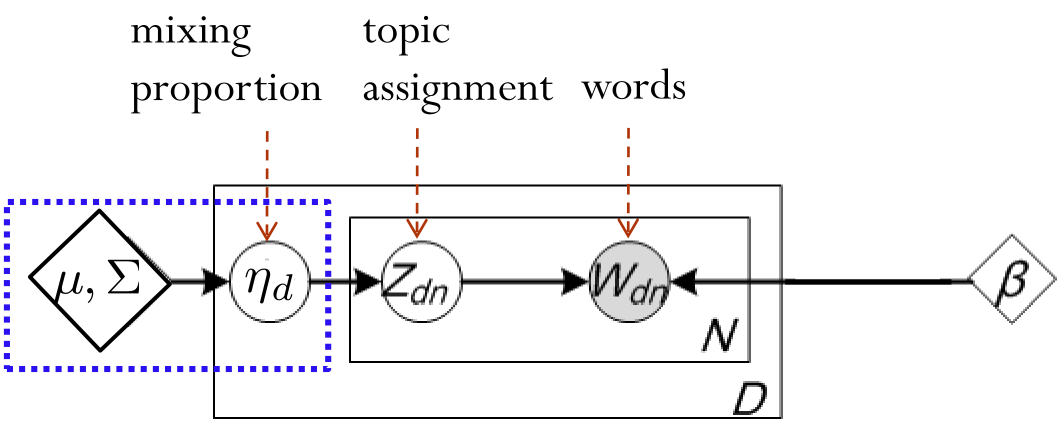
A Scalable Gibbs Sampler



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

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

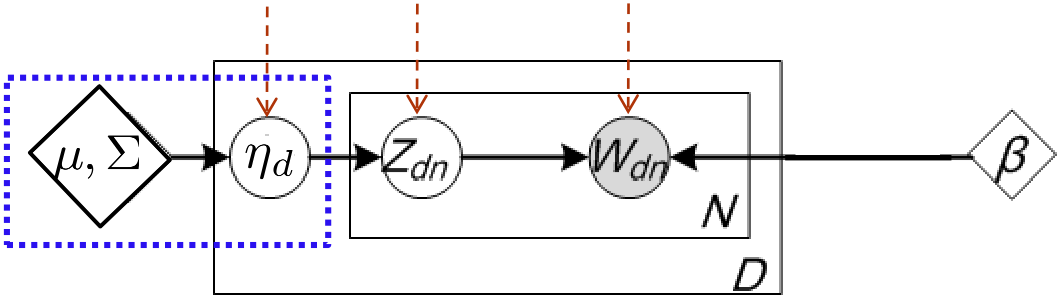
A Scalable Gibbs Sampler



- 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



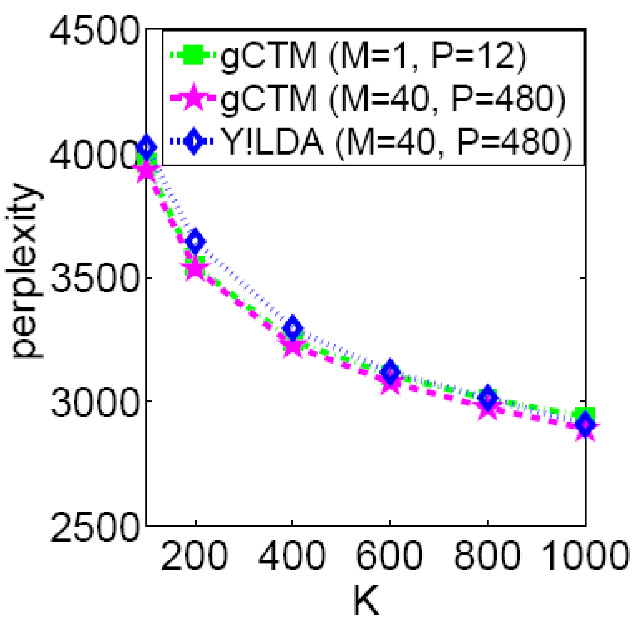
- Data augmentation saves!
- For each dimension k:

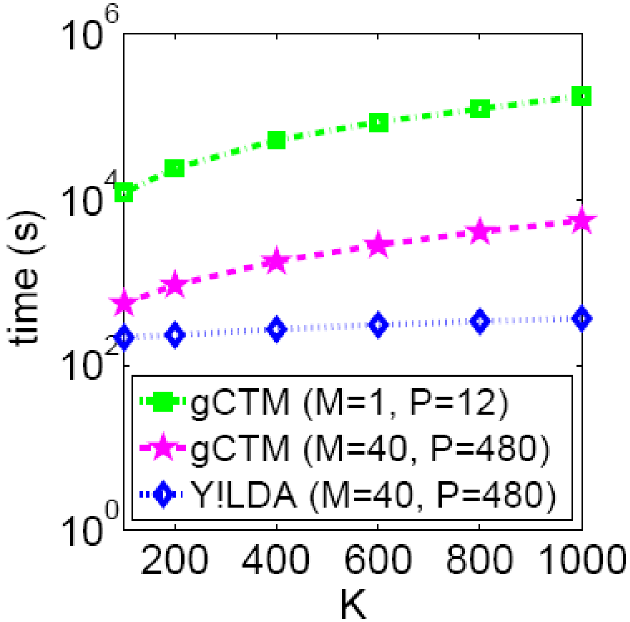
$$p(\eta_d^k | \boldsymbol{\eta}_d^{\neg k}, \mathbf{Z}, \mathbf{W}) \propto \ell(\eta_d^k | \boldsymbol{\eta}_d^{\neg k}) \mathcal{N}(\eta_d^k | \mu_d^k, \sigma_k^2)$$

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

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

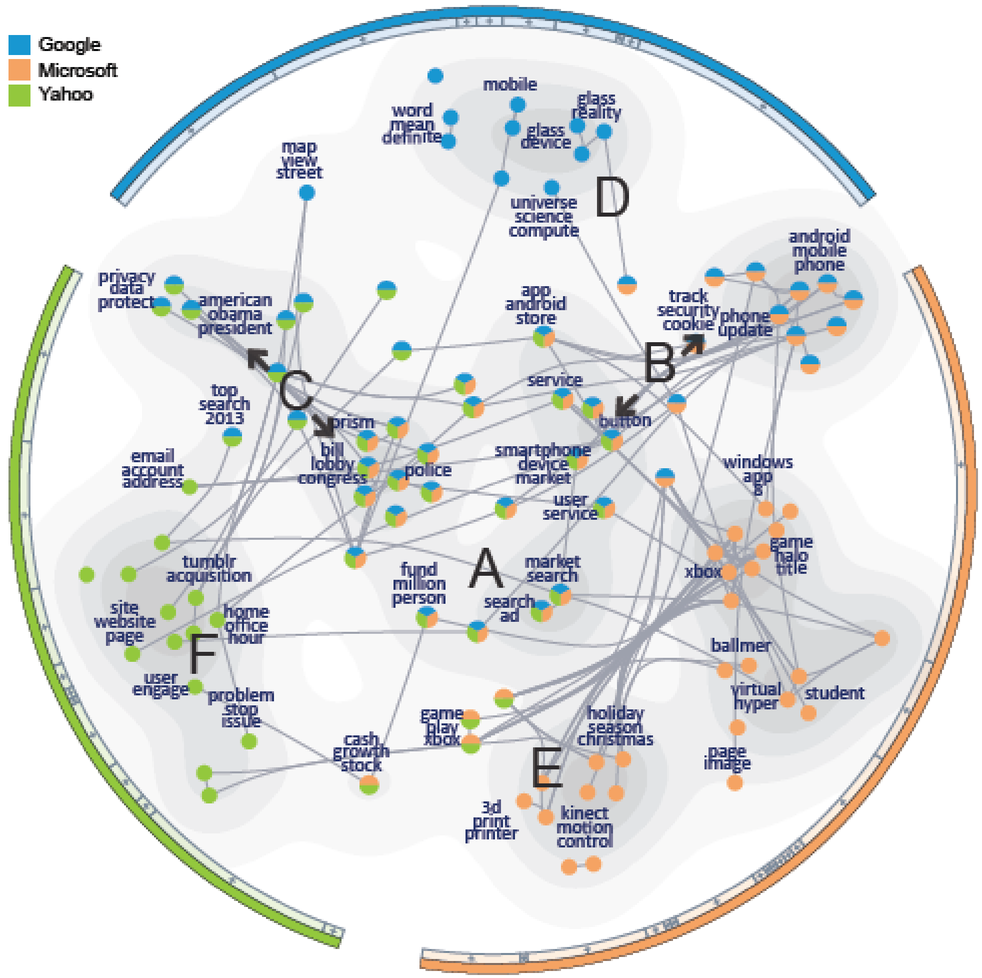
- Leverage big clusters
- Allow learning big models that can't fit on a single machine

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]

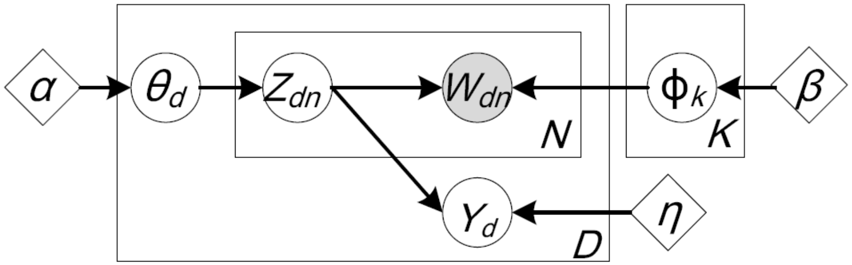
Scalable Graph Visualization



[Liu, Wang, Chen, Zhu, & Guo. IEEE VAST 2014]

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

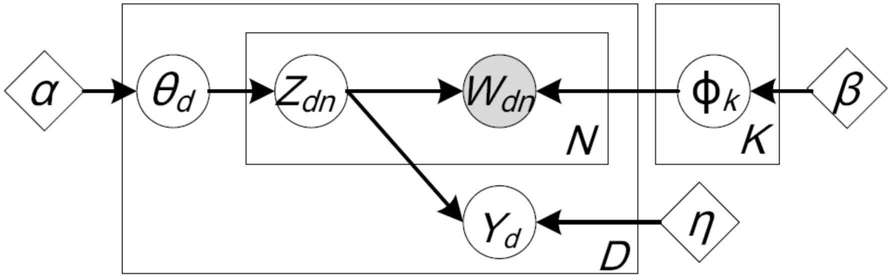
$$p(y_d | \mathbf{z}_d, \eta) = \frac{\{\exp(\eta^\top \bar{\mathbf{z}}_d)\}^{y_d}}{1 + \exp(\eta^\top \bar{\mathbf{z}}_d)} \quad \bar{z}_k = \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}(z_n^k = 1)$$

- 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

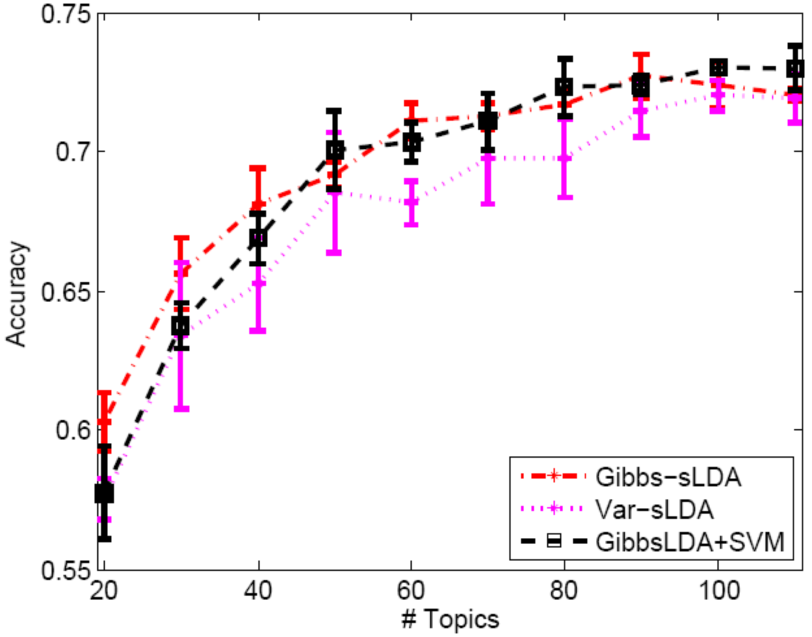
- A document has hundreds of words
- … but only one class label



- Imbalanced likelihood combination

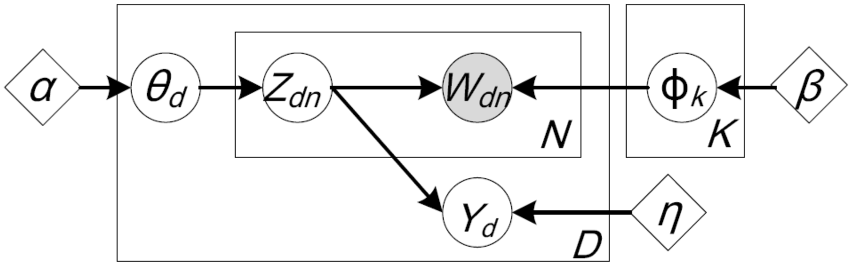
$$p(\mathbf{y}, \mathbf{W}|\mathbf{Z}, \Phi, \eta) = p(\mathbf{y}|\mathbf{Z}, \eta)p(\mathbf{W}|\mathbf{Z}, \Phi)$$

- Too weak influence from supervision



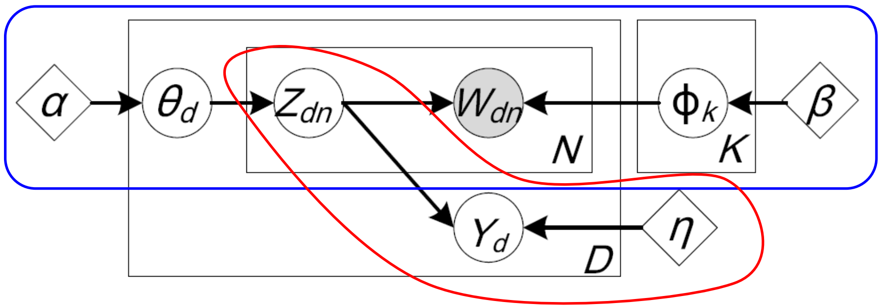
(Halpern et al., ICML 2012; Zhu et al., ACL 2013)

Max-margin Supervised Topic Models



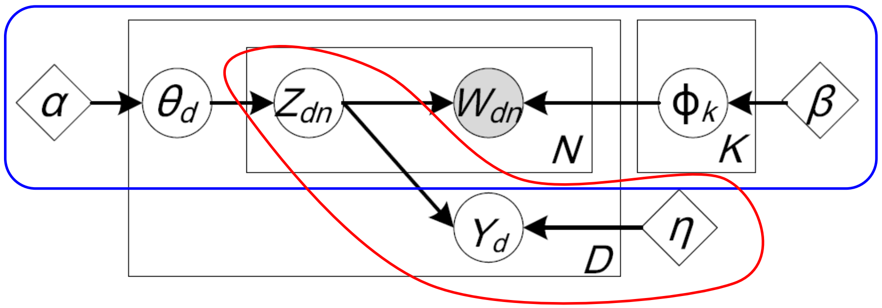
- Can we learn supervised topic models in a max-margin way?
- How to perform posterior inference?
  - Can we do variational inference?
  - Can we do Monte Carlo?
- How to generalize to nonparametric models?

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

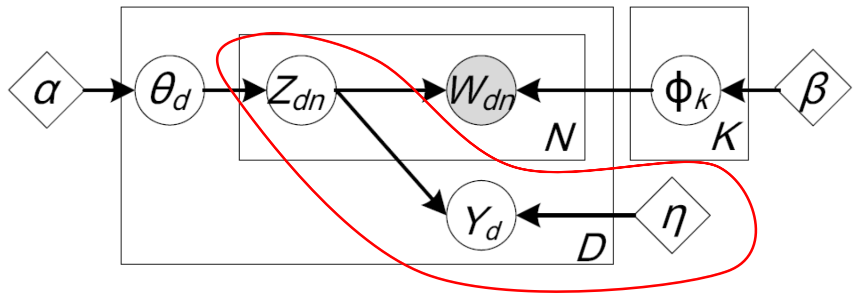
$$\hat{y} = \mathrm{sign} f(\mathbf{w}) = \mathrm{sign}\mathbb{E}_q[f(\eta, \mathbf{z}; \mathbf{w})]$$

  - 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

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

  - 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 $q(\eta,\Theta,\mathbf{Z},\Phi) = q(\eta)q(\Theta,\mathbf{Z},\Phi)$

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

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

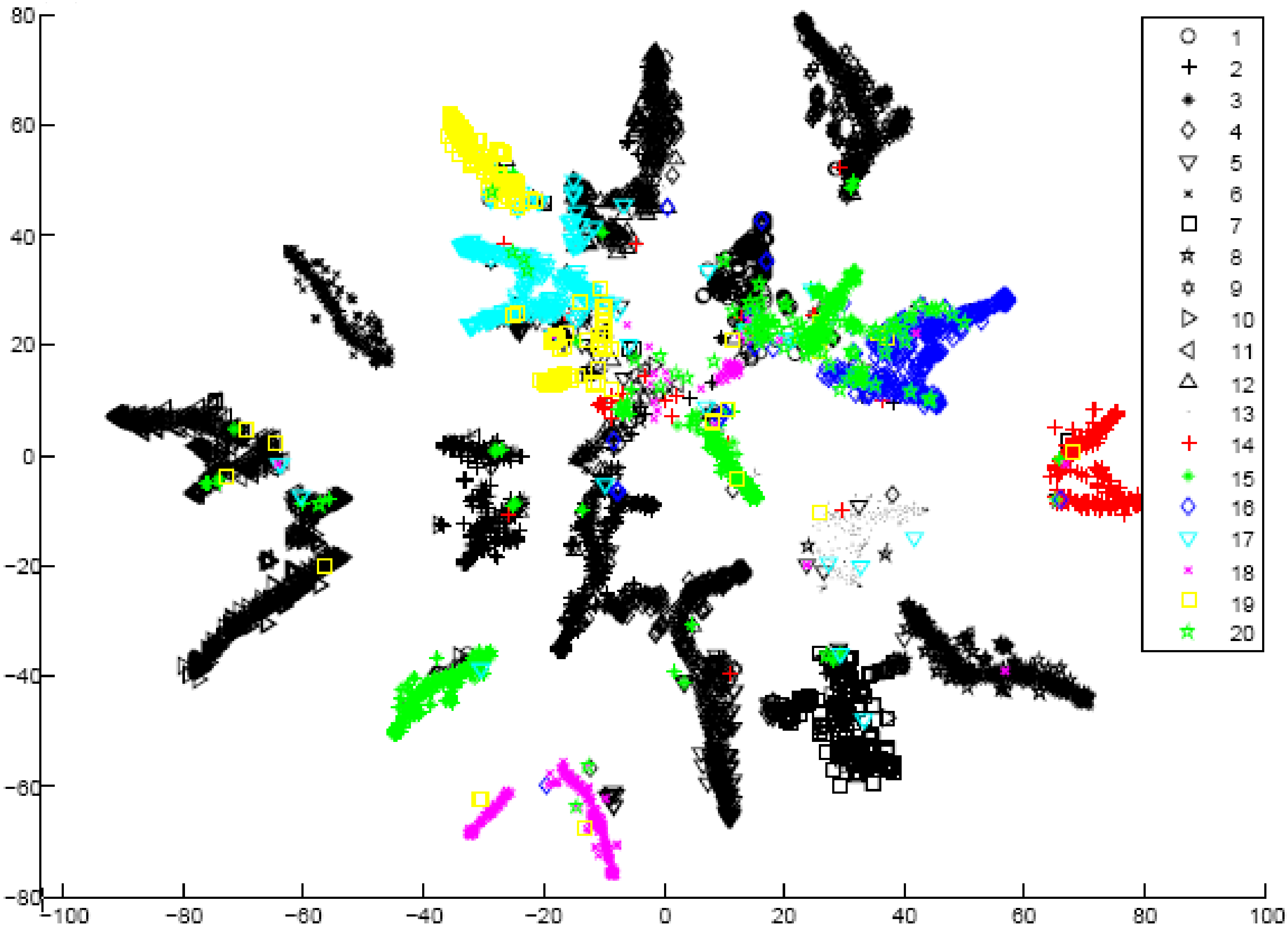
A SVM problem with a normal prior

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

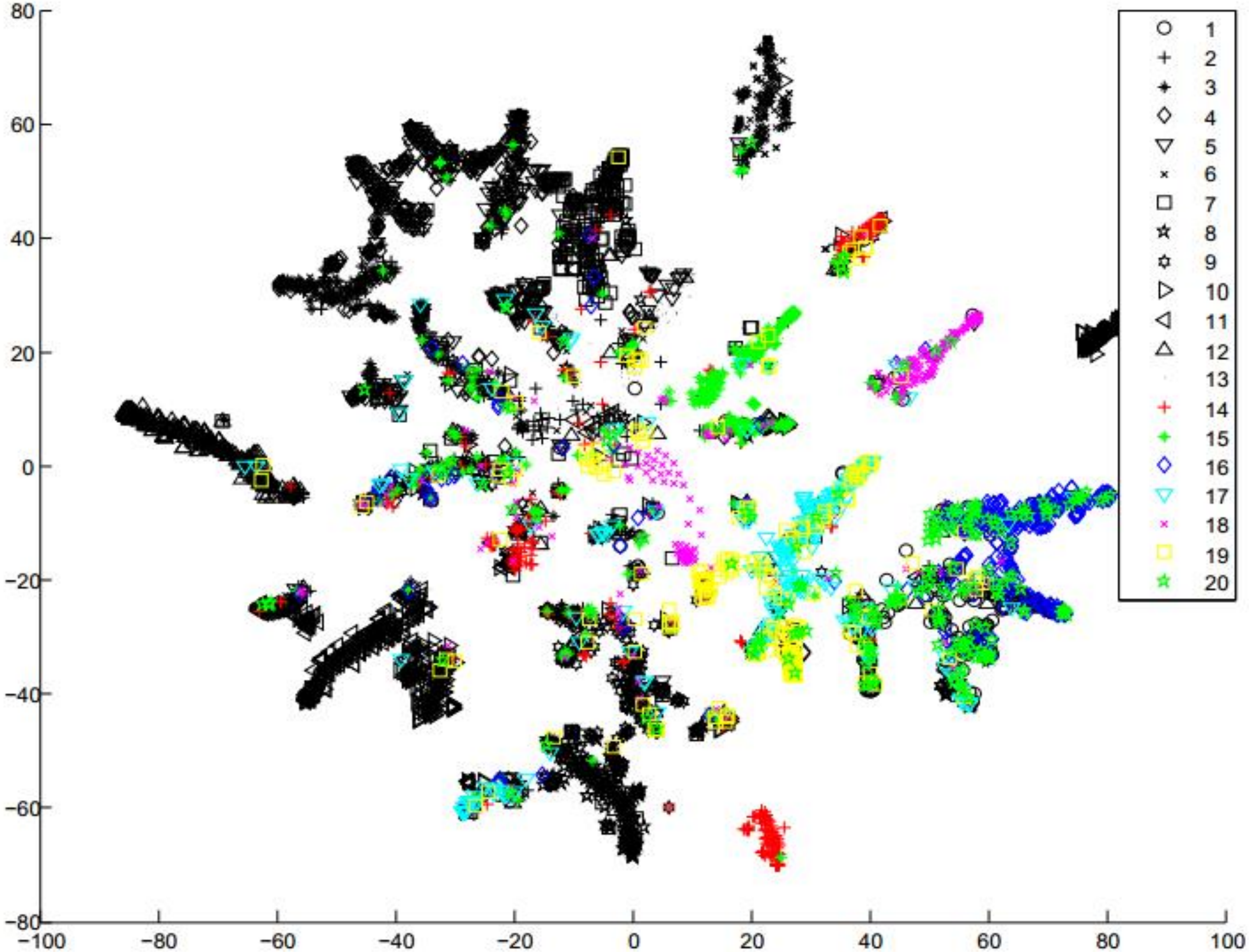
$$\forall d,\ \text{s.t.}: y_d\mathbb{E}_q[\eta]^\top\mathbb{E}_q[\bar{\mathbf{z}}_d] \geq 1-\xi_d.$$

Variational approximation or Monte Carlo methods

Empirical Results on 20Newsgroups

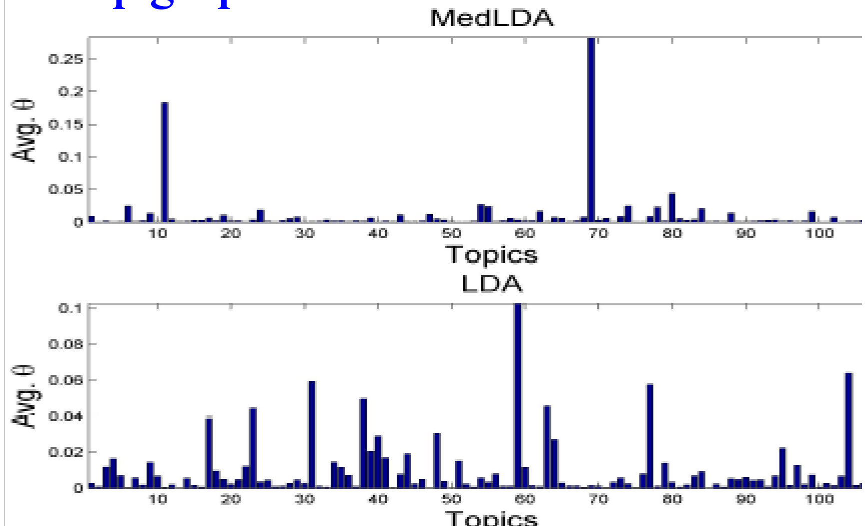


Empirical Results on 20Newsgroups



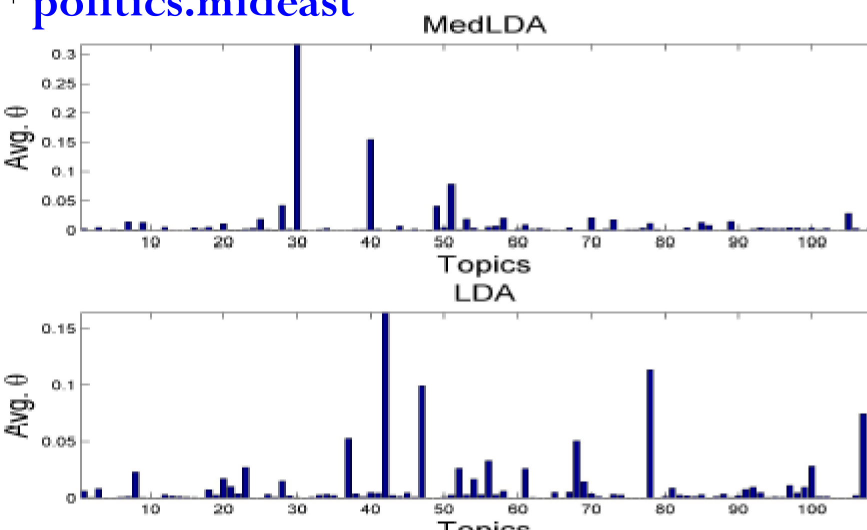
Sparser and More Salient Representations

comp.graphics



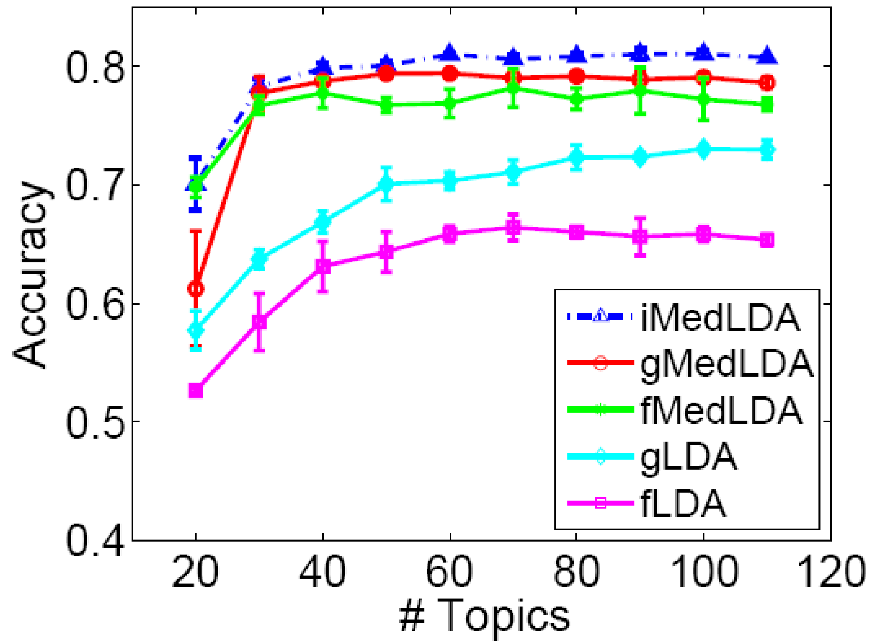
MedLDA			LDA		
T 69	T 11	T 80	T 59	T 104	T 31
image	graphics	db	image	ftp	card
jpeg	image	key	jpeg	pub	monitor
gif	data	chip	color	graphics	dos
file	ftp	encryption	file	mail	video
color	software	clipper	gif	version	apple
files	pub	system	images	tar	windows
bit	mail	government	format	file	drivers
images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

politics.mideast



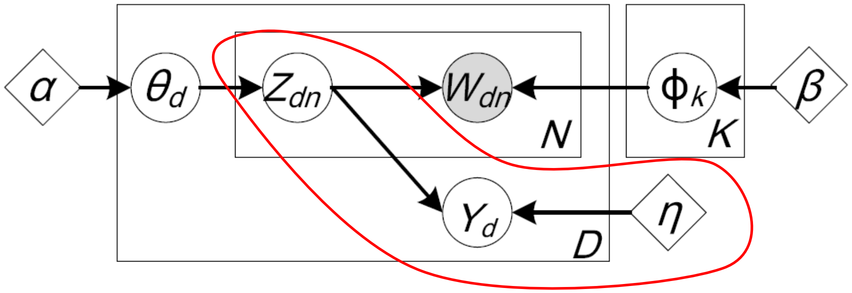
T 30	T 40	T 51	T 42	T 78	T 47
israel	turkish	israel	israel	jews	armenian
israeli	armenian	lebanese	israeli	jewish	turkish
jews	armenians	israeli	peace	israel	armenians
arab	armenia	lebanon	writes	israeli	armenia
writes	people	people	article	arab	turks
people	turks	attacks	arab	people	genocide
article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

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

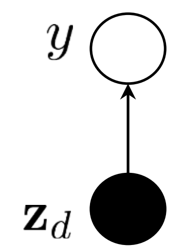
  - where $f(\eta, \mathbf{z}; \mathbf{w}) = \eta^\top \bar{\mathbf{z}} \qquad \bar{z}_k = \frac{1}{N}\sum_{n=1}^{N} \mathbb{I}(z_n^k = 1)$

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

Gibbs MedLDA

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

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



  - The empirical training error

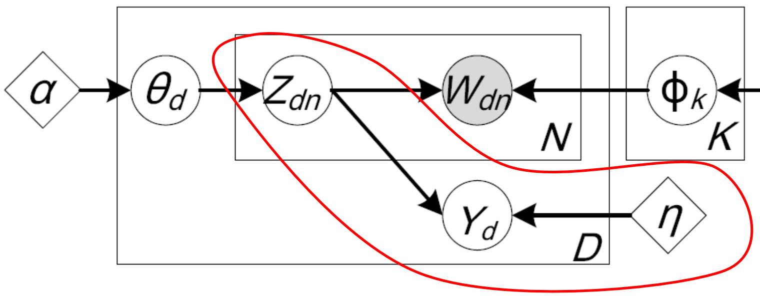
$$\hat{R}(\eta, Z) = \sum_{d=1}^{D} \mathbb{I}(\hat{y}_d|_{\eta,\mathbf{z}_d} \neq y_d)$$

  - 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

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

- What does the lemma mean?
  - It means:

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

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

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

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 $\boldsymbol{\lambda} = 1$ 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

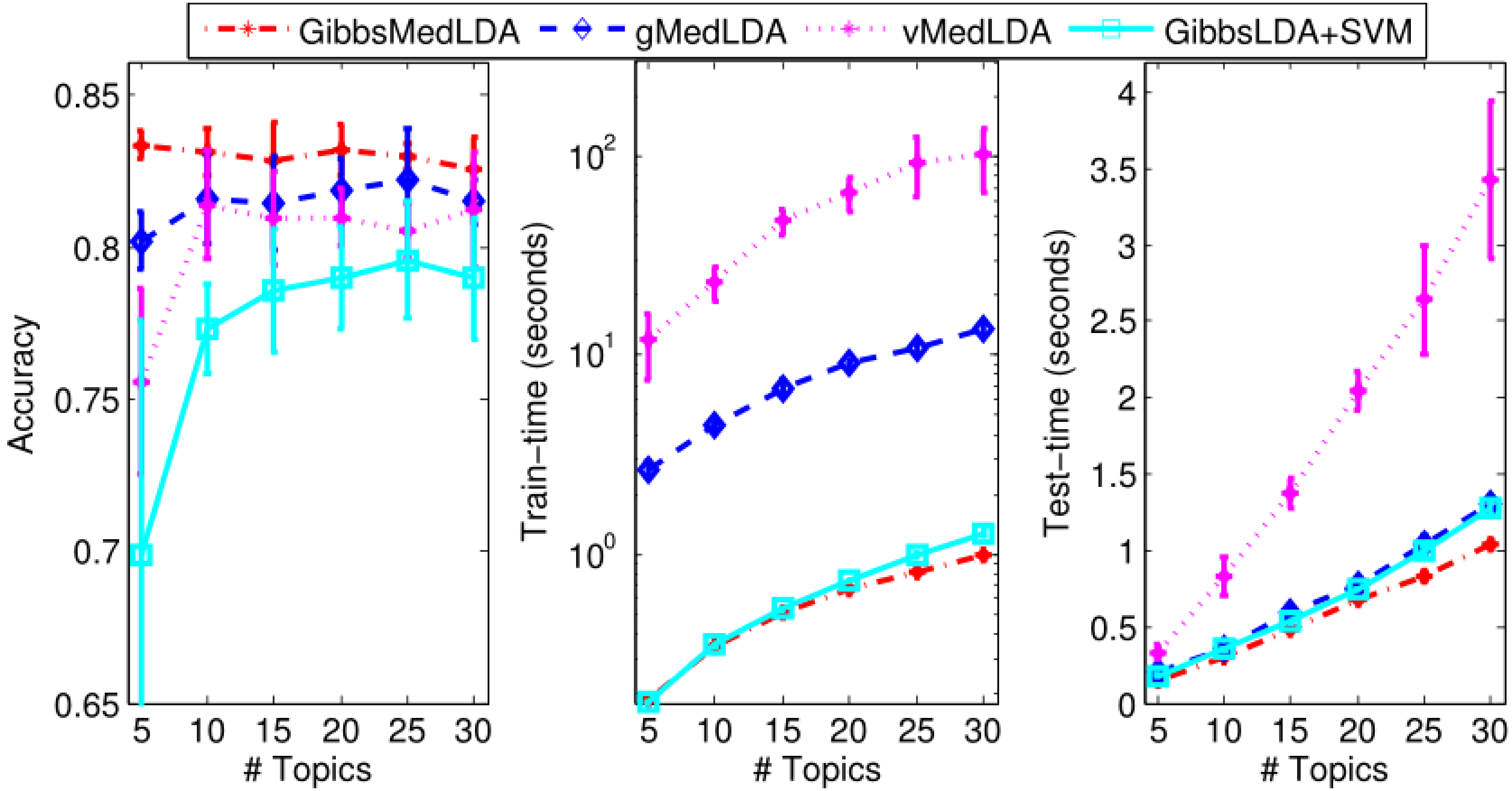
- Per-iteration time complexity

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

  - N_{total} the total number of words

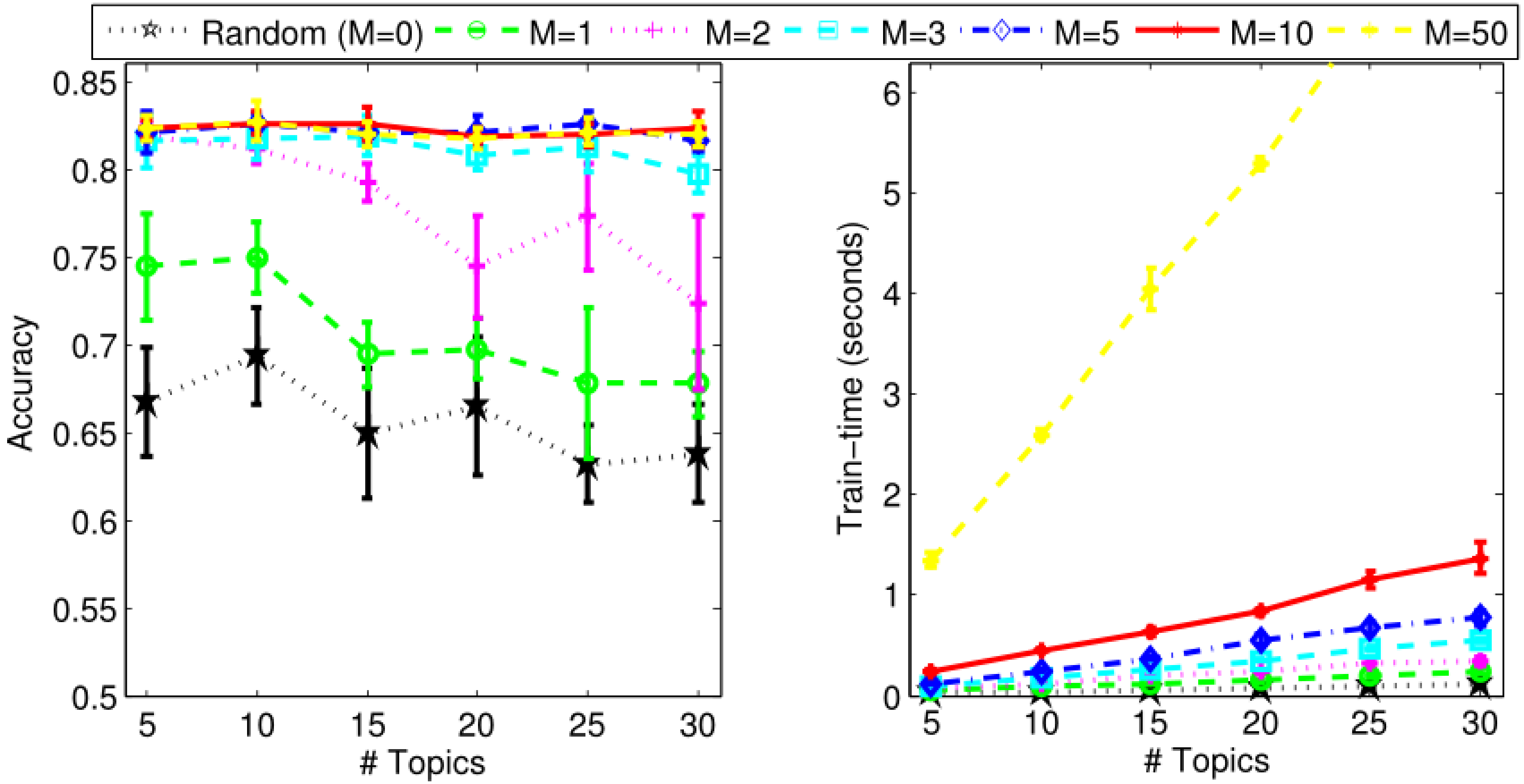
Experiments

◆ 20Newsgroups binary classification



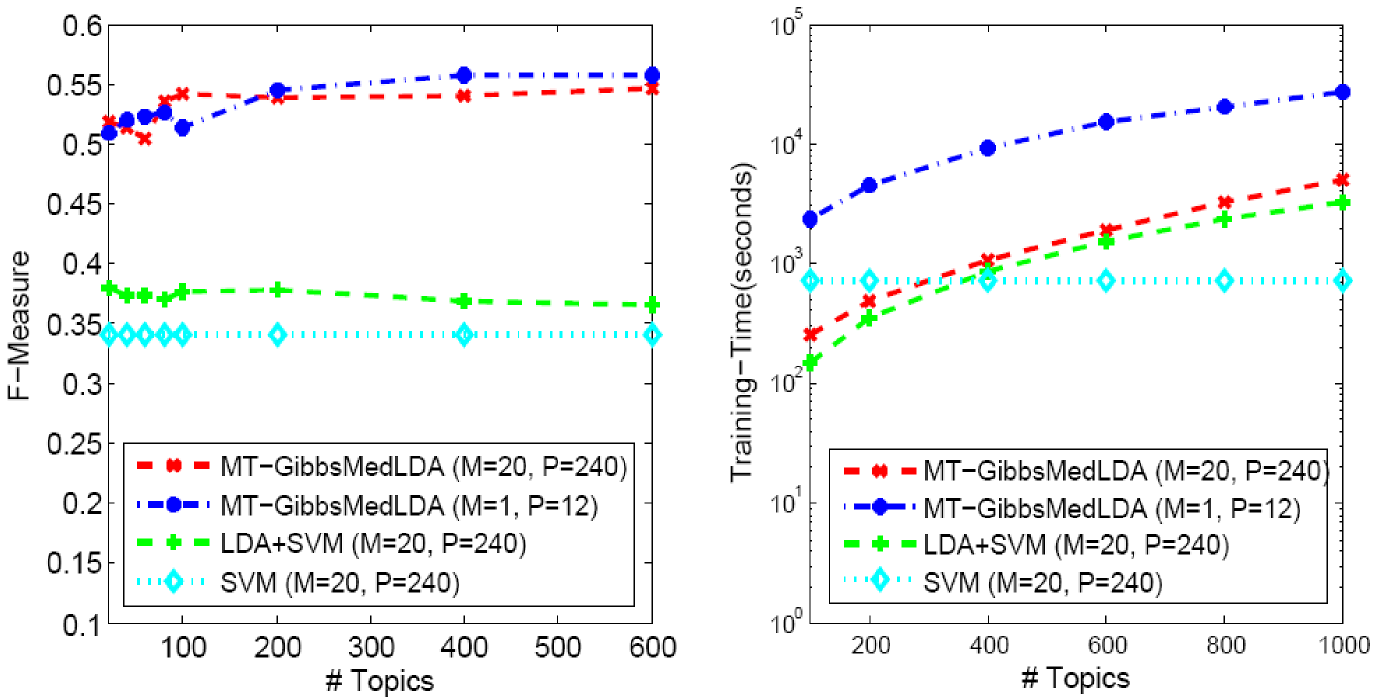
Experiments

- **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]

Summary

- Bayesian methods are highly relevant in learning with big data;
- Topic models are a suitable of statistical models for extracting semantic meanings from large corpora;
- Many developments beyond LDA