Connecting Bayes and Optimization

Jun Zhu

dcszj@mail.tsinghua.edu.cn

Department of Computer Science and Technology

Tsinghua University

Overview

- Bayesian inference as an optimization problem
- Examples with Bayesian SVMs
- Examples with Max-margin LDA
- The general form

Bayes' Rule

- The core of Bayesian methods is Bayes' rule or Bayes' theorem

posterior　　likelihood model　　prior

$$p(\theta|D) = \frac{p(D|\theta)\pi(\theta)}{p(D)}$$

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

Kullback-Leibler (KL) Divergence

- A measure between the difference of two distributions
- Let P, Q be two probability distributions
  - Discrete case:

$$\mathrm{KL}(Q \parallel P) = \sum_i Q(i) \log \frac{Q(i)}{P(i)}$$

  - Continuous case:

$$\mathrm{KL}(Q \parallel P) = \int q(x) \log \frac{q(x)}{p(x)} dx$$

- Properties:
  - Non-negative $\mathrm{KL}(Q \parallel P) \geq 0$
  - Equivalence

$$\mathrm{KL}(Q \parallel P) = 0 \iff Q = P$$

Bayes to Optimization

- For any distribution P, we can recover it by solving an optimization problem

$$\operatorname*{argmin}_{Q \in P_{prob}} \mathrm{KL}(Q \parallel P)$$

- For Bayes' rule, we have $p(\theta|D) = \dfrac{p(D|\theta)\pi(\theta)}{p(D)}$

- Plugging into the optimization problem, we have

$$\mathrm{KL}(Q \parallel P) = \mathrm{KL}(q \| \pi) - \mathbf{E}_{q(\theta)}[\log p(D|\theta)] + const.$$

- **Homework**: complete the proof

Bayesian Inference as an Opt. Problem

$$p(\theta|D) = \frac{p(D|\theta)\pi(\theta)}{p(D)}$$

- Bayes' rule is equivalent to solving:

$$\min_{q(\theta)} \mathrm{KL}(q(\theta)||\pi(\theta)) - \mathbf{E}_q[\log p(D|\theta)]$$

$$s.t.: \; q(\theta) \in P_{prob}$$

Prior regularization

Data fitting

direct but trivial constraints on posterior distribution

[Williams, 1980; Zellner, 1988]

Bayes' Theorem is an Information Processing Rule

- A rule of information minimization

Input

$\pi(\theta)$

$p(D|\theta)$

$\rightarrow$ **Information Processing Rule**

e.g., minimize KL-divergence $\rightarrow$

Output

$p(\theta|D)$

- friendly for extensions, e.g., a more general divergence leads to a more general processing rule

Why is the optimization formulation of interest?

- E.T. Jaynes (1988): "By looking at Bayes' theorem in a fresh way …, it could make the use of Bayesian methods more attractive and widespread, and stimulate new developments in the general theory of inference."

- It bridges Bayesian inference with the subfields of optimization, risk-minimization, max-margin learning, etc.

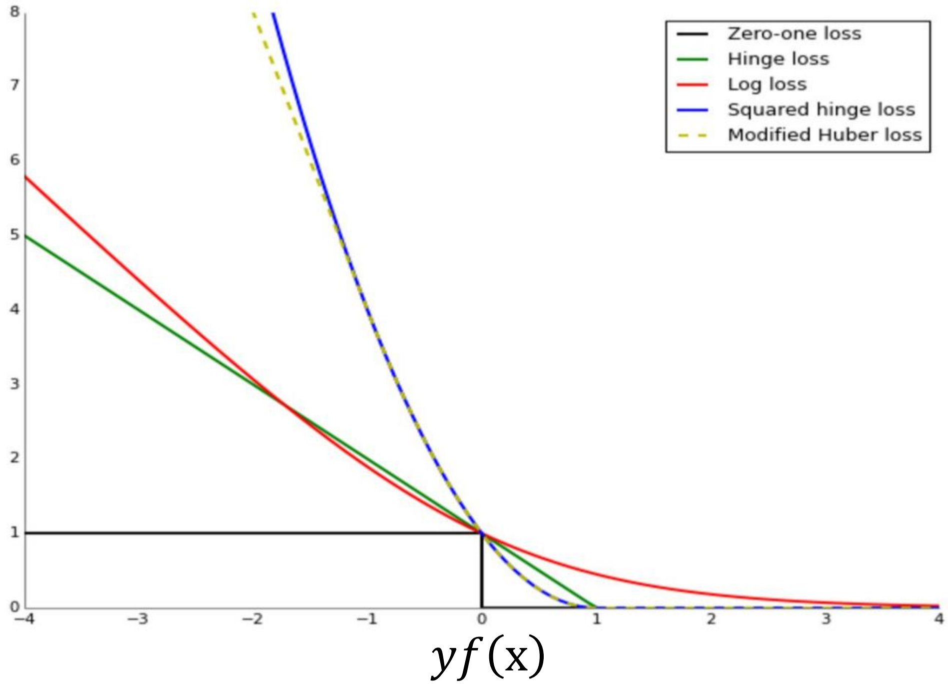
- It combines the best of both worlds

Bayesian SVMs

Learning objectives

$$\min_{\theta} \ \ell(\theta; D) + r(\theta)$$

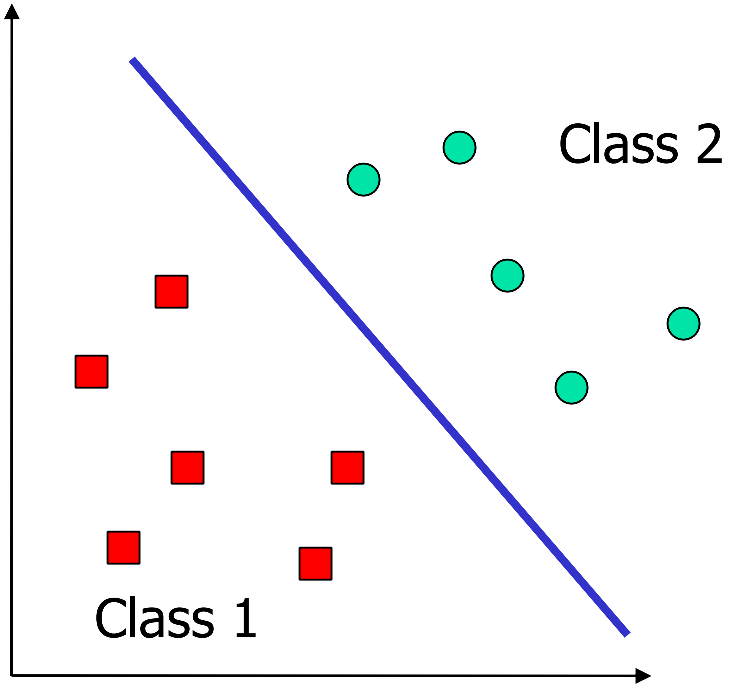
- Loss functions of our predictive task



- Regret/reward in online/reinforcement learning
- *How to directly optimize the objective in Bayesian inference?*

Recap. of SVMs

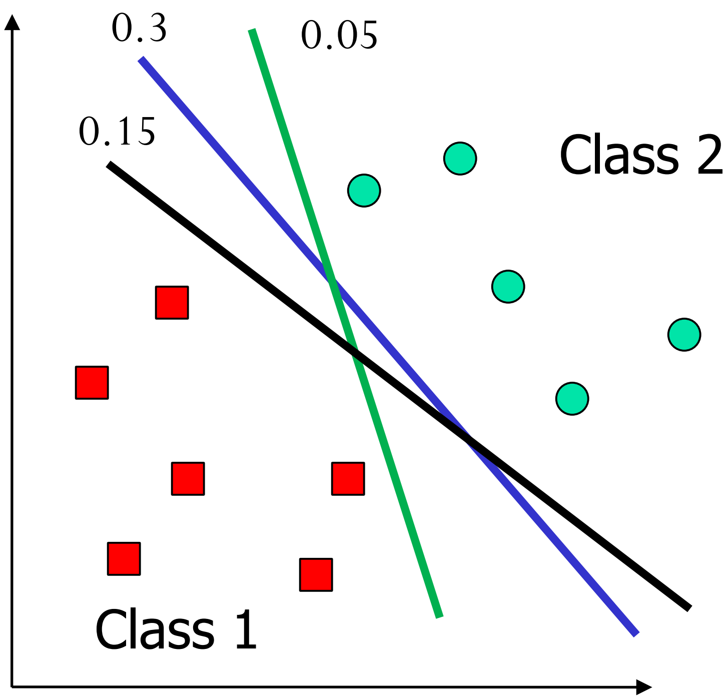
- SVM learns a single decision boundary θ^*



$$\ell(\theta; D) = \sum_i \max(0, 1 - y_i \mathrm{x}_\mathrm{i}^T \theta)$$

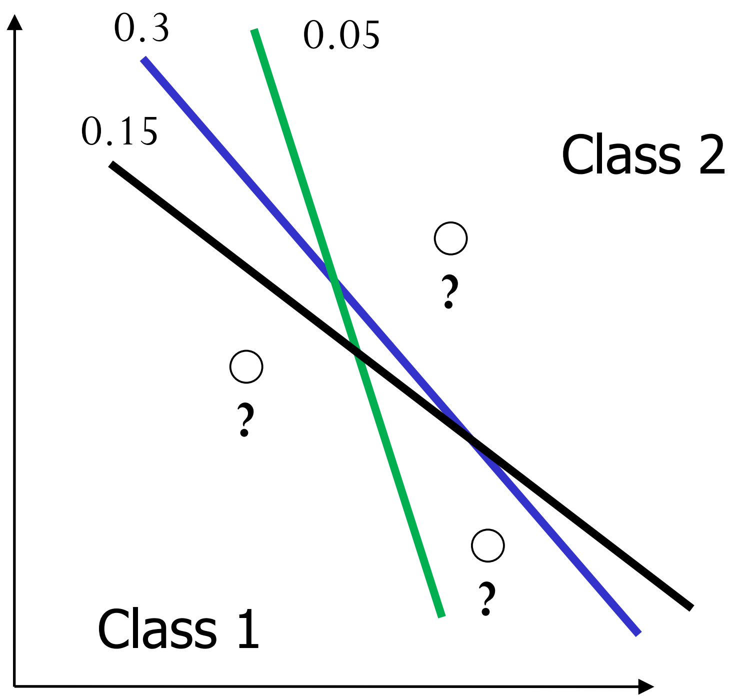
Bayesian SVMs

- Bayesian SVM learns a distribution over all possible decision boundaries $p(\theta; D)$



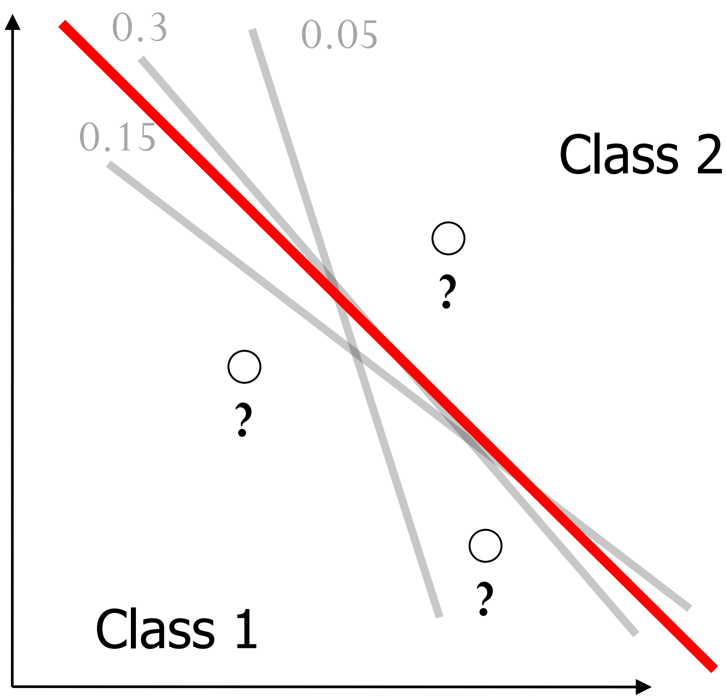
Bayesian SVMs

- Bayesian SVM makes predictions by considering the distribution $p(\theta; D)$ with different strategies



Bayesian SVMs

- **Strategy #1**: use an average model $\hat{\theta} = \mathbf{E}_{p(\theta;D)}[\theta]$

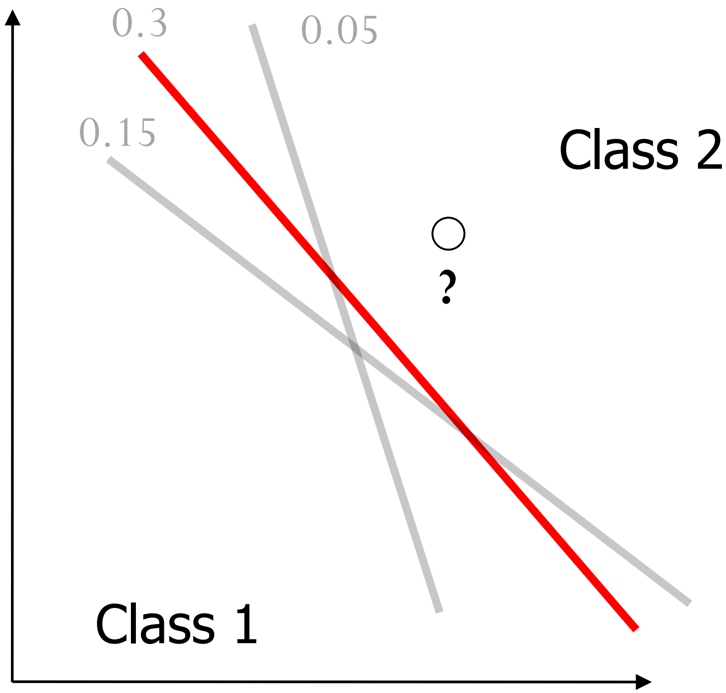


- The hinge-loss on a given training set D is:

$$\ell(q;D) = \sum_i \max(0, 1 - y_i \mathrm{x}_i^T \hat{\theta})$$

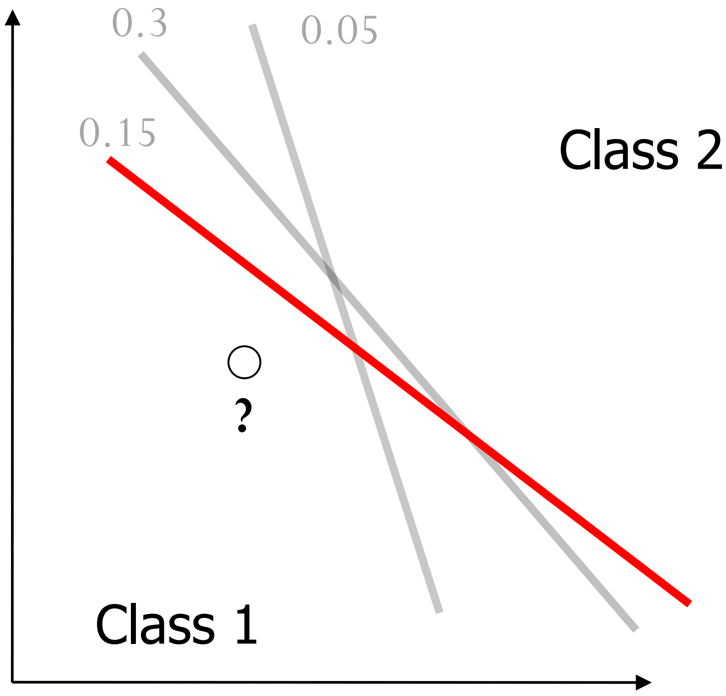
Bayesian SVMs

- **Strategy #2**: use a stochastic model $\theta \sim p(\theta; D)$



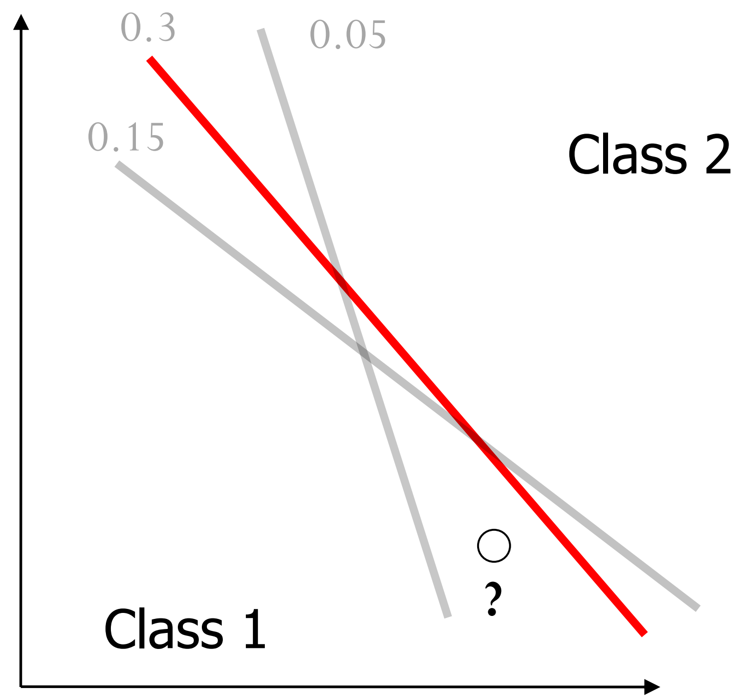
Bayesian SVMs

- **Strategy #2**: use a stochastic model $\theta \sim p(\theta; D)$



Bayesian SVMs

- **Strategy #2**: use a stochastic model $\theta \sim p(\theta; D)$



- The expected hinge-loss on a given training set D is:

$$\ell(p; D) = \sum_i \mathbf{E}_{\boldsymbol{p(\theta;D)}}[\max(0, 1 - y_i \mathrm{x}_\mathrm{i}^T \theta)]$$

Bayesian SVMs

- The overall optimization problem

$$\min_{q(\theta)} \; KL(q(\theta) \,\|\, \pi(\theta)) + C \cdot \ell(q(\theta); D)$$

  - Strategy #1 (Averaging model):

$$\ell(q; D) = \sum_i \max(0, 1 - y_i \mathrm{x}_\mathrm{i}^T \hat{\theta}) \qquad \hat{\theta} = \mathbf{E}_{q(\theta)}[\theta]$$

  - Strategy #2 (Gibbs/Stochastic model):

$$\ell(q; D) = \sum_i \mathbf{E}_{q(\theta;D)}[\max(0, 1 - y_i \mathrm{x}_\mathrm{i}^T \theta)]$$

Solve Bayesian SVMs

- **Strategy #1** (Averaging model):
  - If the prior is normal $\pi(\theta) = N(0, I)$, we have the solution:

$$q(\theta) \propto \pi(\theta) \exp\left(\sum_i \alpha_i y_i \mathrm{x}_i^T \theta\right) = N\left(\sum_i \alpha_i y_i \mathrm{x}_i , I\right)$$

  - where α_i are the solution of the dual problem

$$\text{max. } W(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1,j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j$$

$$\text{subject to } C \geq \alpha_i \geq 0, \sum_{i=1}^{n} \alpha_i y_i = 0$$

  - **Homework (*)**: complete the proof (optional)

Solve Bayesian SVMs

- **Strategy #2** (Gibbs/Stochastic model):
  - For a given prior, we have the solution:

$$q(\theta) \propto \pi(\theta) \exp\left(-C \cdot \sum_i \max(0, 1 - y_i \mathrm{x}_\mathrm{i}^T \theta)\right)$$

  - If the prior is normal $\pi(\theta) = N(0, I)$, this is NOT a normal distribution!
  - But, a nice Gibbs sampling algorithm exists via data augmentation (Polson & Scott, 2011)

Data Augmentation

- **Idea**: augment the original sample space to introduce "conditional conjugacy" for Gibbs sampling

$$p(\theta) \Rightarrow p(\theta, \lambda) \qquad s.t.: \int p(\theta, \lambda) d\lambda = p(\theta)$$

  - Sample $p(\theta \mid \lambda)$
  - Sample $p(\lambda \mid \theta)$

  - Iterate for a sufficiently long time, and drop the augmented variable

Data Augmentation for SVMs

$$q(\theta) \propto \pi(\theta) \exp\left(-C \cdot \sum_i \max(0, 1 - y_i \mathrm{x}_\mathrm{i}^T \theta)\right)$$

- Break the non-conjugacy in Bayesian SVMs (Polson & Scott, 2011) $\zeta_i = 1 - y_i \mathrm{x}_\mathrm{i}^T \theta$

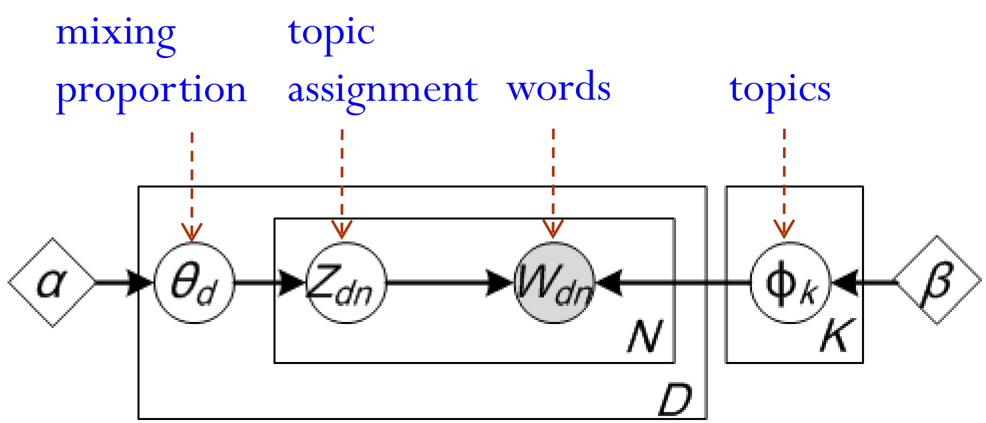
$$\exp\left(-2C \cdot \sum_i \max(0, \zeta_i)\right) = \int_{>0} \frac{1}{\sqrt{2\pi\lambda_i}} \exp\left(-\frac{(\lambda_i + C \cdot \zeta_i)^2}{2\lambda_i}\right) d\lambda_i$$

- Under a standard normal prior: (**Homework **, optional**)
  - Sample θ from a normal distribution
  - Sample each λ_i from a generalized inverse-Gaussian (GIG) distribution

Bayesian SVMs with Latent Variables (topic models)

Recap. of LDA

- A Bayesian model for text document analysis



- Inference finds the posterior distribution

$$p(\Theta, \Phi, Z | W, \alpha, \beta) \propto \prod_k p(\phi_k|\beta) \prod_d p(\theta_d|\alpha) \left(\prod_n p(z_{dn}|\theta_d) p(w_{dn}|z_d, \Phi) \right)$$

Recap. of LDA

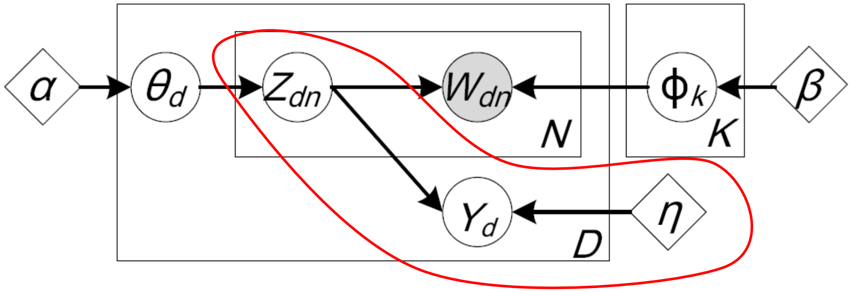
- According to the optimization form of Bayes' rule, inference is equivalent to solving:

$$\underset{q}{\operatorname{argmin}}\, \mathrm{KL}(q(\Theta, \Phi, Z) \parallel P(\Theta, \Phi, Z | W, \alpha, \beta))$$

  - where:

$$p(\Theta, \Phi, Z | W, \alpha, \beta) \propto \prod_k p(\phi_k | \beta) \prod_d p(\theta_d | \alpha) \left(\prod_n p(z_{dn} | \theta_d) p(w_{dn} | z_d, \Phi) \right)$$

MedLDA: a Max-margin LDA



- Define an averaging classifier based on latent topic assignments
  - The classifier weight vector is η
  - Learn the posterior distribution

$$q(\eta, \Theta, \Phi, Z \mid W, Y)$$

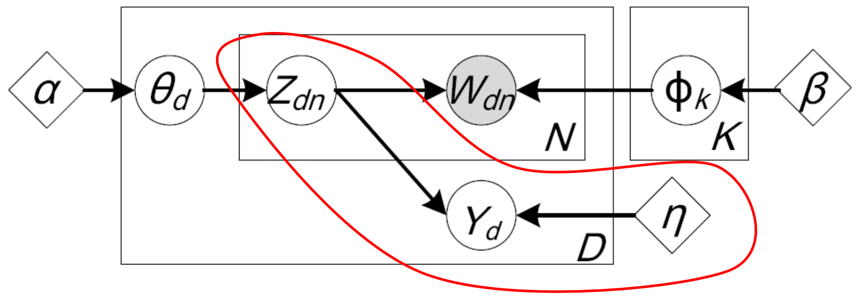
  - q-weighted averaging classifier ($y \in \{+1, -1\}$)

$$\hat{y}_d = \text{sign}\left(\mathbf{E}_q[f(\eta, \mathrm{z}_{\mathrm{d}})]\right)$$

  - where

$$f(\eta, \mathrm{z}_{\mathrm{d}}) = \eta^T \bar{\mathrm{z}}_d \qquad \bar{\mathrm{z}}_{dk} = \frac{1}{N}\sum_n \delta(z_{dn} = k)$$

MedLDA: a Max-margin LDA



- Bayesian inference with max-margin posterior regularization

$$\min_{q(\eta,\Theta,\Phi,Z)} L\big(q(\Theta,\Phi,Z)\big) + \underbrace{\mathrm{KL}\big(q(\eta)\,\|\,\pi(\eta)\big) + C\cdot \ell\big(q(\eta,\Theta,\Phi,Z)\big)}$$

fitting on the words in LDA

fitting on class labels in a Bayesian SVM

- the hinge loss

$$\ell(q(\eta,\Theta,\Phi,Z)) = \sum_d \max\big(0,1 - y_d \mathbf{E}_q[\eta^T\, \bar{z}_d]\big)$$

Inference Algorithm

$$\min_{q(\eta,\Theta,\Phi,Z)} \underbrace{L\big(q(\Theta,\Phi,Z)\big)}_{\text{fitting on the words in LDA}} + \underbrace{\mathrm{KL}\big(q(\eta)\,\|\,\pi(\eta)\big) + C\cdot\ell\big(q(\eta,\Theta,\Phi,Z)\big)}_{\text{fitting on class labels in a Bayesian SVM}}$$

- An iterative procedure with $q(\eta,\Theta,\Phi,Z) = q(\eta)q(\Theta,\Phi,Z)$
  - Update $q(\eta)$:

$$\min_{q(\eta)} \mathrm{KL}\big(q(\eta)\,\|\,\pi(\eta)\big) + C\cdot\ell\big(q(\eta);\ q(\Theta,\Phi,Z)\big)$$

    - This is a Bayesian SVM problem
    - When the prior is normal, we solve a SVM dual problem.

Inference Algorithm

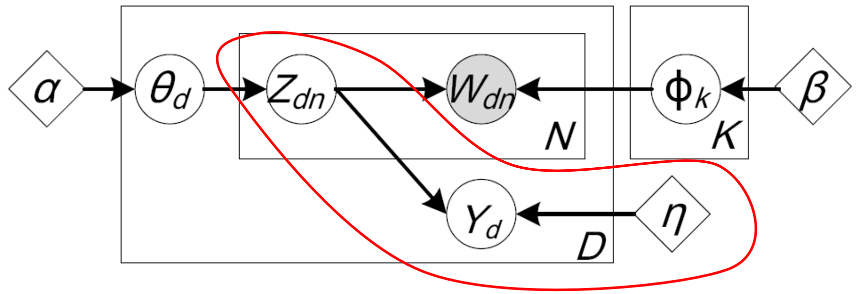
$$\min_{q(\eta,\Theta,\Phi,Z)} \underbrace{L\big(q(\Theta,\Phi,Z)\big)}_{\text{fitting on the words in LDA}} + \underbrace{\mathrm{KL}\big(q(\eta)\|\pi(\eta)\big) + C\cdot\ell(q(\eta,\Theta,\Phi,Z))}_{\text{fitting on class labels in a Bayesian SVM}}$$

- An iterative procedure with $q(\eta,\Theta,\Phi,Z) = q(\eta)q(\Theta,\Phi,Z)$
  - Update $q(\Theta,\Phi,Z)$:

$$\min_{q(\Theta,\Phi,Z)} L\big(q(\Theta,\Phi,Z)\big) + C\cdot\ell(q(\Theta,\Phi,Z);q(\eta))$$

    - This is a posterior inference problem, but more difficult than vanilla LDA
    - The mean-field variational method applies here!

Gibbs MedLDA



- Define a Gibbs classifier based on latent topic assignments
  - The classifier weight vector is η
  - Learn the posterior distribution

$$q(\eta, \Theta, \Phi, Z \mid W, Y)$$

  - Sample a classifier

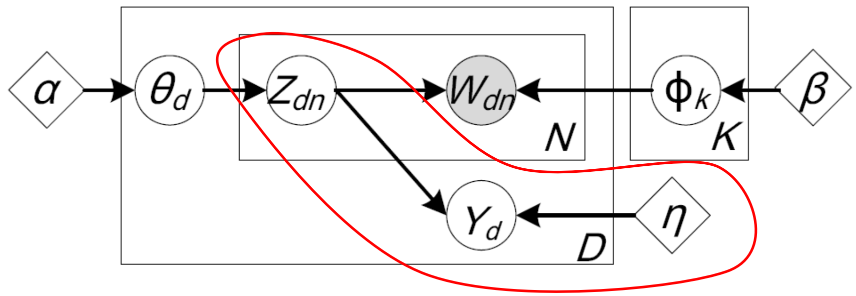
$$\eta, \mathrm{z_d} \sim q(\eta, \Theta, \Phi, Z \mid W, Y)$$

$$\hat{y}_d = \mathrm{sign}\, f(\eta, \mathrm{z_d})$$

  - where

$$f(\eta, \mathrm{z_d}) = \eta^T \bar{\mathrm{z}}_d \qquad \bar{\mathrm{z}}_{dk} = \frac{1}{N} \sum_n \delta(z_{dk} = k)$$

Gibbs MedLDA



- Bayesian inference with max-margin posterior regularization

$$\min_{q(\eta,\Theta,\Phi,Z)} L\big(q(\Theta,\Phi,Z)\big) + \underbrace{\mathrm{KL}\big(q(\eta)\|\pi(\eta)\big) + C \cdot \ell(q(\eta,\Theta,\Phi,Z))}$$

fitting on the words in LDA

fitting on class labels in a Bayesian SVM

- the hinge loss

$$\ell(q(\eta,\Theta,\Phi,Z)) = \sum_d \mathbf{E}_q[\max(0, 1 - y_d \eta^T \bar{z}_d)]$$

Gibbs MedLDA

$$\min_{q(\eta,\Theta,\Phi,Z)} \underbrace{L\big(q(\Theta,\Phi,Z)\big)}_{\text{fitting on the words in LDA}} + \underbrace{\mathrm{KL}\big(q(\eta)\|\pi(\eta)\big) + C\cdot \ell\big(q(\eta,\Theta,\Phi,Z)\big)}_{\text{fitting on class labels in a Bayesian SVM}}$$

- Optimal solution:

$$q(\eta,\Theta,\Phi,Z) = \frac{p(\Theta,\Phi,Z|W)\pi(\eta)\prod_d \phi(y_d \mid z_d,\eta)}{\psi(Y,W)}$$

  - The pseudo-likelihood (un-normalized)

$$\phi(y_d \mid Z_d,\eta) = \exp(-C\cdot\max(0,1-y_d\eta^T\bar{z}_d))$$

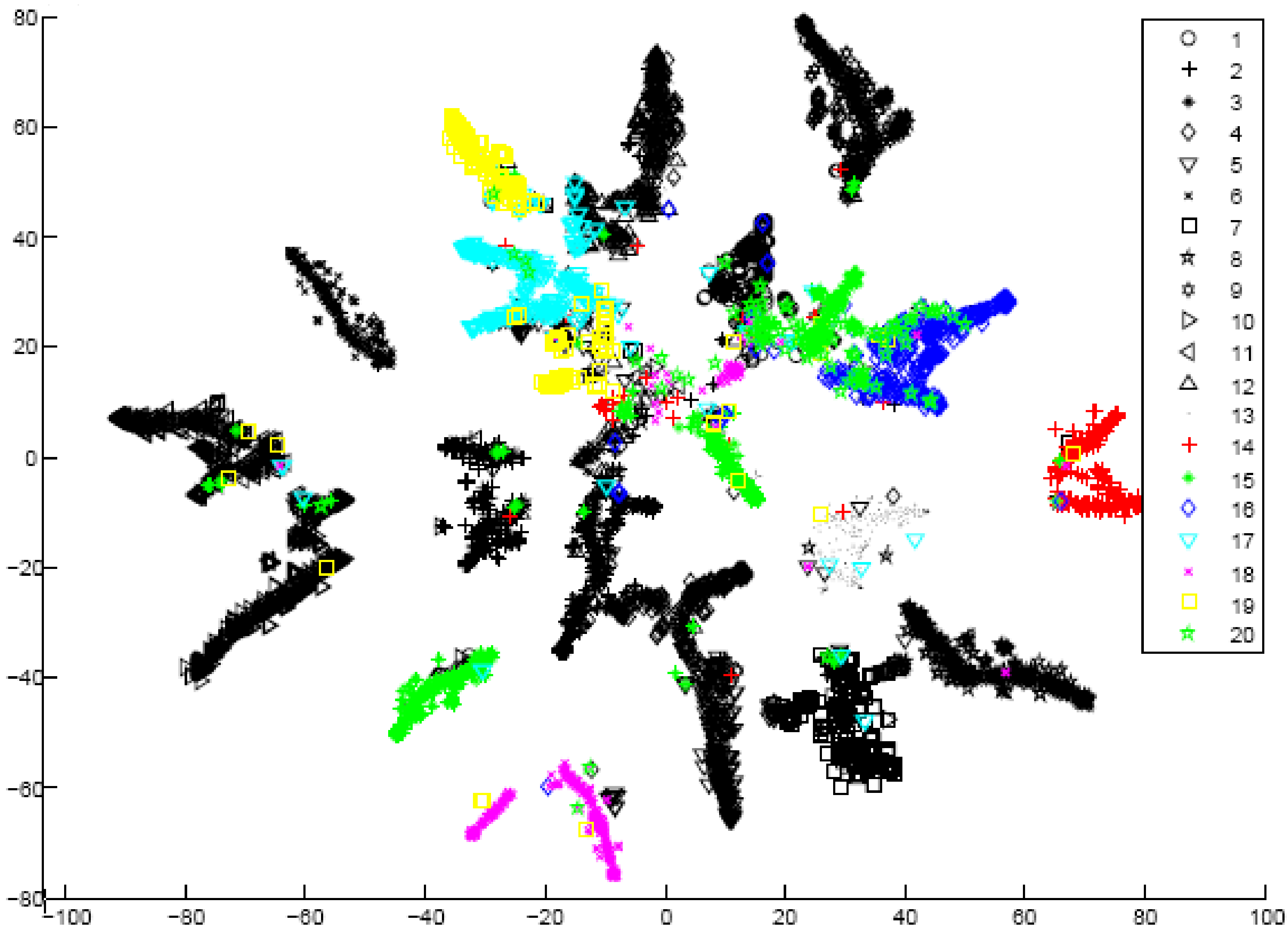
Some Results on 20 Newsgroup Dataset

- A collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups

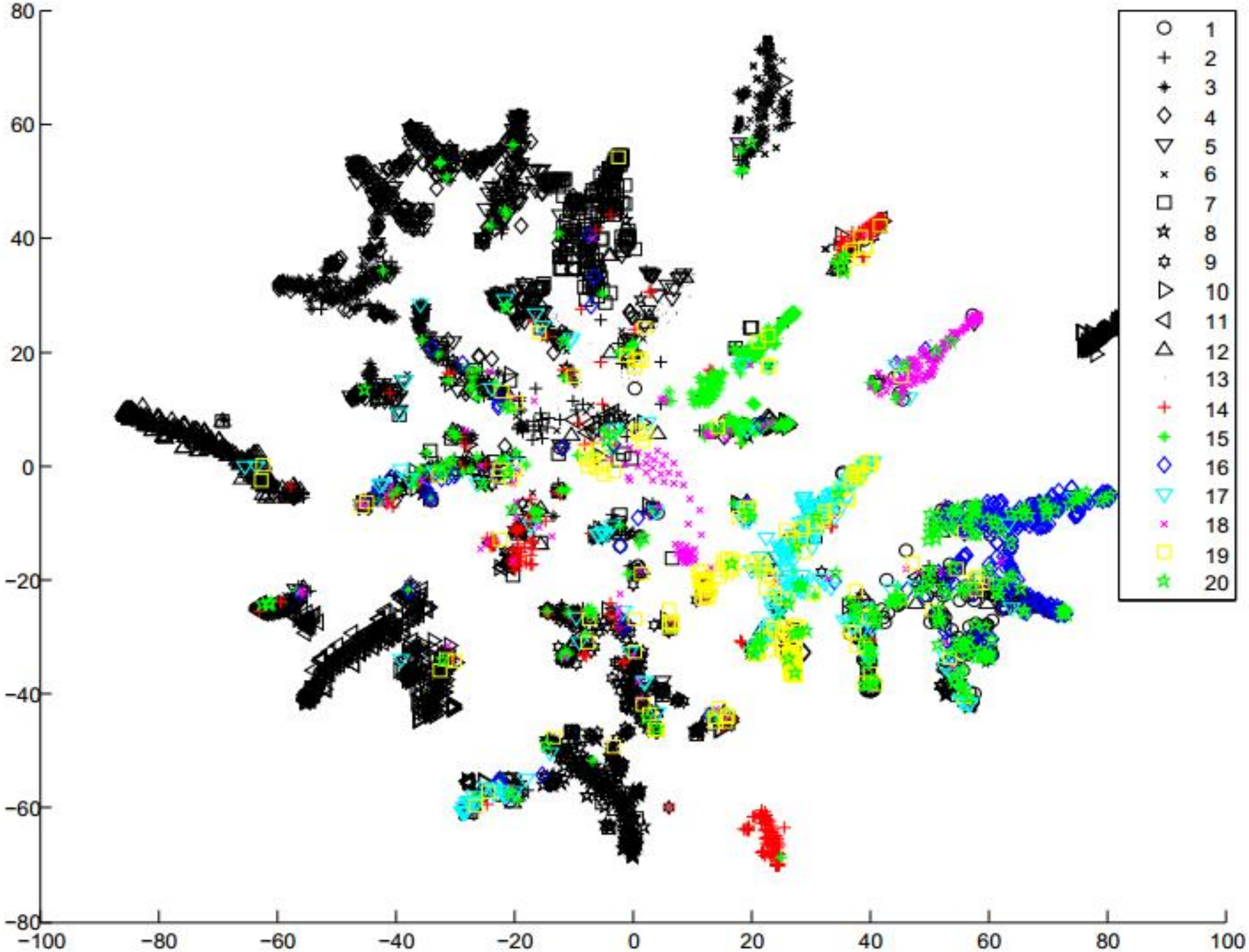
comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics sci.med sci.space
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast	talk.religion.misc alt.atheism soc.religion.christian

  - A popular dataset for document analysis
  - http://qwone.com/~jason/20Newsgroups/

Embedding of the Topic Representations (MedLDA)

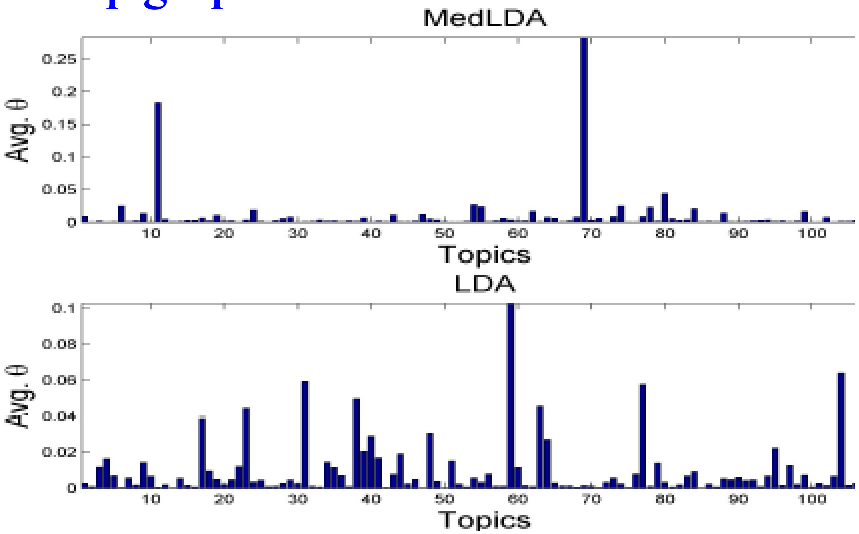


Embedding of the Topic Representations (LDA)



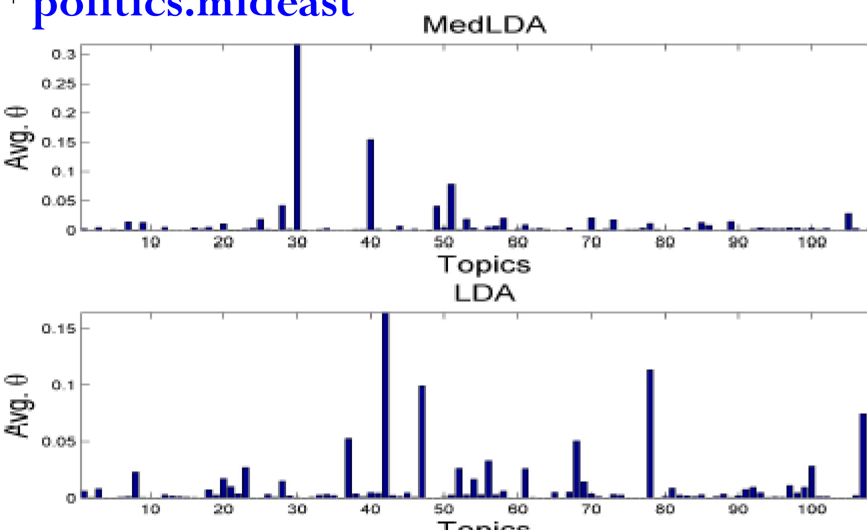
Sparser and More Salient Representations

comp.graphics



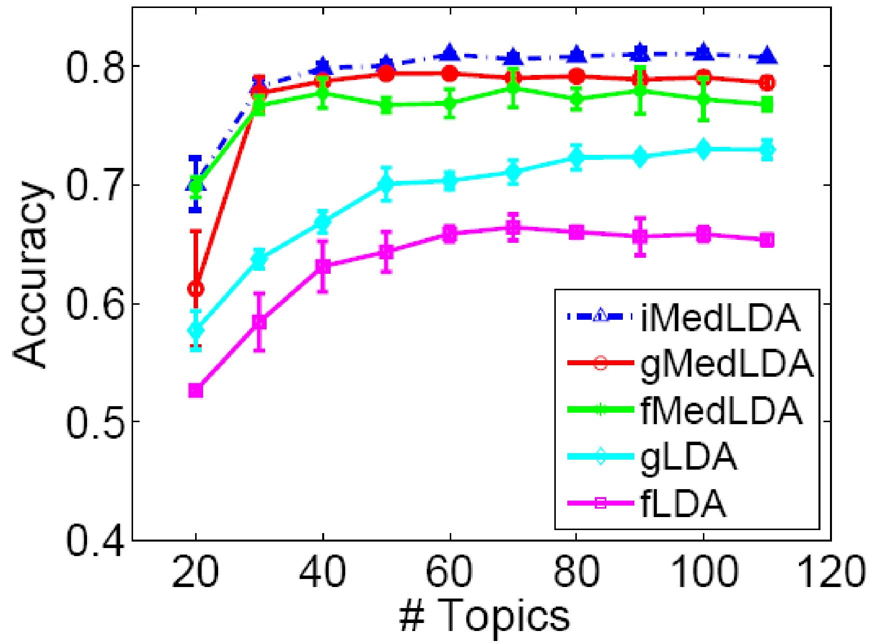
MedLDA			LDA		
T 69	T 11	T 80	T 59	T 104	T 31
image	graphics	db	image	ftp	card
jpeg	image	key	jpeg	pub	monitor
gif	data	chip	color	graphics	dos
file	ftp	encryption	file	mail	video
color	software	clipper	gif	version	apple
files	pub	system	images	tar	windows
bit	mail	government	format	file	drivers
images	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

Classification Performance



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

Regularized Bayesian Inference (RegBayes)

- A general form to optimize for a posterior distribution

$$\min_{q(M)} \mathrm{KL}(q(M)||\pi(M)) - \mathbf{E}_q[\log p(D|M)] + \Omega(q)$$

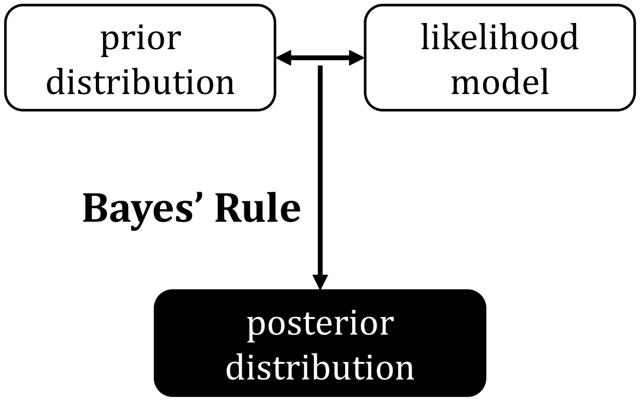
$$s.t.: \; q(M) \in P_{prob}$$

posterior regularization

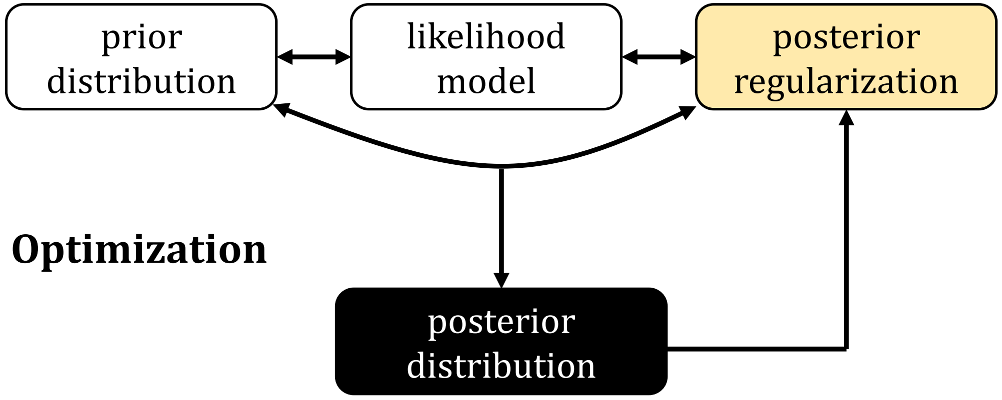
- Consider both hard and soft constraints
- Convex optimization problem with nice properties
- Can be effectively solved with convex duality theory

A High-Level Comparison

Bayes:



RegBayes:



Summary

- Bayes' rule is equivalent to solving an optimization problem

- Bridges Bayesian methods, learning and optimization
  - Bayesian SVMs without likelihood
  - Max-margin topic models with likelihood

- Scalable algorithms have been developed (Shi & Zhu, 2014)

References

- P. M. Williams. *Bayesian conditionalisation and the principle of minimum information*. British Journal for the Philosophy of Science, 31(2):131–144, 1980.
- A. Zellner, *Optimal Information Processing and Bayes's Theorem*, The American Statistician, Vol. 42 (4), 1988.
- N. Polson and S. Scott. *Data augmentation for support vector machines*. Bayesian Analysis, 6(1):1–24, 2011
- J. Zhu, N. Chen, and E. P. Xing. *Bayesian inference with posterior regularization and applications to infinite latent SVMs*. Journal of Machine Learning Research, 15(May):1799-1847, 2014.
- J. Zhu, A. Ahmed, and E. Xing. *MedLDA: maximum margin supervised topic models*. Journal of Machine Learning Research, (13):2237–2278, 2012.
- J. Zhu, N. Chen, H. Perkins, and B. Zhang. Gibbs max-margin topic models with data augmentation. Journal of Machine Learning Research, 15(May):1073-1100, 2014.
- T. Shi, and J. Zhu. *Online Bayesian Passive-Aggressive Learning*. International Conference on Machine Learning, 2014. (JMLR 2016 to appear)