Discriminative Learning vs Generative Learning

**Generative learning** (with KL divergence)
- p(x|θ)
- Chain rule
- p(x) = ∏ p(x_t|x_{<t})
- Tractable inference

**Discriminative learning** (with squared error)
- p(y|x, θ)
- Harder to infer

Introduction
- BP-sLDA: Supervised LDA (sLDA) with BackProp (BP)
- A modified probabilistic graphical model of sLDA
- Discriminative learning for high prediction accuracy
- Easy MAP inference with mirror descent algorithm
- Scalable learning by stochastic mirror descent + BackProp

BP-sLDA: on par with deep neural networks
- Scalable inference: a structured deep feedforward network
- Scalable training
- Significantly faster than previous topic models
- Large-scale corpus: 10hr on 7.9M docs (1 CPU machine)
- Prediction accuracy:
  - Outperforms previous supervised topic models
  - Outperforms neural networks with same #parameters
- On par with deep neural network with more #parameters

Problem Formulation
- sLDA Models
  - sLDA (Blei & McAuliffe 2008)
  - This modification leads to a differentiable end-to-end cost trainable by back propagation with superior performance
- Discriminative Learning vs Generative Learning
  - Generative learning (traditional method)
    - arg max_θ p(Φ|θ) ∏ p(w_t|x_t, θ|Φ)
  - Discriminative learning (our method)
    - arg max_θ p(Φ|θ) ∏ p(w_t|x_t, θ|Φ, u, b, γ)
  - A Principled Framework for Discriminative Learning
    - Joint probability of a new model family with additional set of parameters (Larrée, Bishop & Minka 2006)
    - Decoupled to: supervised + unsupervised
      - Supervised Model: arg max_θ ∑ [log p(y_t|θ, y_t)]
      - Unsupervised Model: arg max_θ ∑ [log p(Φ|θ)]

Inference: Forward Activation
- The Inference Problem
  - Inference: posterior of the output given input
  - Not tractable: the integral and p(θ|w, φ) are hard to compute
- Previous solutions: Variational inference & Gibbs sampling
- Inference via MAP Sampling
  - Sampling the posterior with one MAP estimate of θ_d
  - The MAP inference of θ_d
    - arg max_θ p(θ_d|w, φ, Φ)
  - Deep Feedforward Mirror Descent Network
    - Mirror descent: "Linearization + KL-divergence" [Closed-form solution]

Learning: Back Propagation
- The Empirical Risk Minimization Problem
  - Use the approximation p(y|x,θ, u, y') = p(y|θ(x), u, y')
  - Stochastic Mirror Descent
    - Stochastic mirror descent for Φ
    - Stochastic gradient descent for U
    - U_t = U_{t-1} - μ_s ∆U_t
  - ∆U_t and ∆Φ_t are computed by back propagation

Experiments and Results
- Tasks and Datasets
  - Regression: Amazon Movie Review (AMR) dataset (McCuley & Leskovec 2013)
    - 7.9M documents, 1.48B words, from 889K users, 235K movies
  - Binary classification: Multidomain sentiment analysis (MultiSent) dataset
    - 342K documents on 25 types of products
  - Binary classification: Business prediction proprietary dataset
    - 1.2M documents, vocabulary size 128K
- Prediction Performance
  - Comparison with traditional topic models
    - Large-scale regression on AMR-7.9M (predictive r squared: pR^2 = 1-MSE/Var)
    - Large-scale binary classification for business prediction (Area Under the Curve)
      - Best DNN result: 76.2%, and Gibbs-sLDA/Spectral-sLDA/Hybrid-sLDA: 10%-20%
- Analysis and Computation Efficiency
  - Influence of MDA iterations L
  - Sparsity of the topic distribution
  - Per word log-likelihoods
  - Number of topics