Diversity Regularization of Latent Variable Models: Theory, Optimization and Applications

Pengtao Xie

Joint work with Yuntian Deng and Eric Xing
Carnegie Mellon University
Latent Variable Models (LVMs)

Machine Learning

Latent Variable Models

DATA

KNOWLEDGE
Latent Variable Models

**Topic Models**
- Words
- Topics

**Gaussian Mixture Model**
- Feature vectors
- Clusters

Hidden Markov Model, Kalman Filtering, Restricted Boltzmann Machine, Deep Belief Network, Factor Analysis, etc.

Neural Network, Sparse Coding, Matrix Factorization, Distance Metric learning, Principal Component Analysis, etc.
Latent Variable Models

Latent Factors in Knowledge

- Topics
  - Politics
  - Economics
  - Sports

- Clusters
  - Animal
  - Building
  - Flower

Components in LVMs

- Topic Models
- Gaussian Mixture Model
Motivation I: Popularity of latent factors is skewed

- Popularity of latent factors follows a power-law distribution

<table>
<thead>
<tr>
<th>Dominant Topics</th>
<th>Long-tail Topics</th>
<th>Dominant Clusters</th>
<th>Long-tail Clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Politics</td>
<td>Garden</td>
<td>Animal</td>
<td>Computer</td>
</tr>
<tr>
<td>Economics</td>
<td>Animal</td>
<td>Food</td>
<td>Book</td>
</tr>
<tr>
<td>Sports</td>
<td>Furniture</td>
<td>Tour</td>
<td>Diamond</td>
</tr>
<tr>
<td></td>
<td>Tour</td>
<td>Flower</td>
<td>Cup</td>
</tr>
<tr>
<td></td>
<td>Flower</td>
<td>Food</td>
<td>Statue</td>
</tr>
<tr>
<td></td>
<td>Flower</td>
<td>Building</td>
<td>Train</td>
</tr>
</tbody>
</table>
Standard LVMs are insufficient to capture long-tail factors
Long-tail factors are important

- Amount is large

- Arguably more interesting
  - Example: in advertisement, a “lose weight” topic is more important than a “time” topic
Diversification

Latent factors in Knowledge
Components in LVM
Diversification

Standard LVM | Diversified LVM
Motivation II: Tradeoff induced by the number of components $k$

- Tradeoff between Expressiveness and Complexity
  - Small $k$: low expressiveness, low complexity
  - Large $k$: high expressiveness, high complexity
- Can we achieve the best of both worlds?
  - Small $k$: high expressiveness, low complexity
Reduce model complexity without sacrificing expressiveness

Without diversity regularization       With diversity regularization

Knowledge
Components
Diversity Regularized LVMs

- **Goal:** encourage the components to diversely spread out to (1) improve the coverage of long-tail latent factors; (2) reduce model complexity without compromising expressiveness

- **Approach:**
  - Define a metric to measure the diversity of components
  - Use the diversity metric to regularize the learning of components
Outline

- Diversity Regularizer
- Optimization
- Applications
- Theory
Diversity Metric

- Components are parametrized by vectors
  - In Latent Dirichlet Allocation, each topic has a multinomial vector
  - In Sparse Coding, each dictionary item is a vector
- Measure the dissimilarity between two vectors
- Measure the diversity of a vector set
Dissimilarity between two vectors

- Invariant to scale, translation, rotation and orientation of the two vectors
- Euclidean distance, L1 distance
  - Variant to scale

- Cosine similarity
  - Variant to orientation
Dissimilarity between two vectors

- Non-obtuse angle

- Invariant to scale, translation, rotation and orientation of the two vectors

- Definition

\[ \theta = \arccos \left( \frac{x \cdot y}{\|x\| \|y\|} \right) \]
Measure the diversity of a vector set

- Based on the pairwise dissimilarity measure between vectors
- The diversity of a set of vectors \( A = \{a_i\}_{i=1}^K \) is defined as

\[
\Omega(A) = \text{mean}(\Theta) - \text{var}(\Theta)
\]

where

\[
\Theta = \{\theta_{ij}\}_{i=1, j=1}^{i=K, j=K} \quad \theta_{ij} = \arccos\left(\frac{|a_i \cdot a_j|}{\|a_i\| \|a_j\|}\right)
\]

- Mean: summarize how these vectors are different from each other on the whole
- Variance: encourage the vectors to **evenly** spread out
Diversity Regularized LVM

\[
\max_A L(D; A) + \lambda \Omega(A)
\]

\[
\Omega(A) = \text{mean}(\Theta) - \text{var}(\Theta)
\]

\[
\Theta = \left\{ \theta_{ij} \right\}_{i=1, j=1}^{K, K}, \quad \theta_{ij} = \arccos \left( \frac{a_i \cdot a_j}{\|a_i\| \|a_j\|} \right)
\]
Optimization

- The diversity regularizer is non-smooth and non-convex

- Derive a smooth lower bound
  - The lower bound is easier to derive if the parameter vectors lie on a sphere
  - Decompose the parameter vectors into magnitudes and directions

- Proved that maximizing the lower bound with gradient method can increase the diversity regularizer in each iteration
Optimization

Reparametrize

\[ g_i = \|a_i\| \quad a_i = g_i \tilde{a}_i \quad A = \text{diag}(g) \tilde{A} \]

\[ \max_{g, \tilde{A}} L(D; A) + \lambda \Omega(A) \]

\[ \max_{g, \tilde{A}} L(D; g\tilde{A}) + \lambda \Omega(\tilde{A}) \]

\[ \text{s.t.} \quad \forall i, g_i \geq 0 \]

\[ \text{s.t.} \quad \forall i, \|\tilde{a}_i\| = 1 \]

Fix \( \tilde{A} \), optimize \( g \)

\[ \max_g L(D; g\tilde{A}) \]

\[ \text{s.t.} \quad \forall i, g_i \geq 0 \]

Fix \( g \), optimize \( \tilde{A} \)

\[ \max_{\tilde{A}} L(D; g\tilde{A}) + \lambda \Omega(\tilde{A}) \]

\[ \text{s.t.} \quad \forall i, \|\tilde{a}_i\| = 1 \]
Optimization

\[
\max_{\tilde{\mathbf{A}}} \quad L(D; g\tilde{\mathbf{A}}) + \lambda \Omega(\tilde{\mathbf{A}}) \\
\text{s.t.} \quad \forall i, \|\tilde{\mathbf{a}}_i\| = 1
\]

Lower bound

\[
\Omega(\tilde{\mathbf{A}}) \geq \Gamma(\tilde{\mathbf{A}}) = \arcsin(\sqrt{\det(\tilde{\mathbf{A}}^T \tilde{\mathbf{A}})}) - \left(\frac{\pi}{2} - \arcsin(\sqrt{\det(\tilde{\mathbf{A}}^T \tilde{\mathbf{A}})})\right)^2
\]

\[
\max_{\tilde{\mathbf{A}}} \quad L(D; g\tilde{\mathbf{A}}) + \lambda \Gamma(\tilde{\mathbf{A}}) \\
\text{s.t.} \quad \forall i, \|\tilde{\mathbf{a}}_i\| = 1
\]
Theorem

- Maximizing the lower bound with projected gradient ascent (PGA) can increase the diversity metric

\[ \text{THEOREM 1. Let } G^{(t)} \text{ be the gradient of } \Gamma(\tilde{A}) \text{ w.r.t } \tilde{A}^{(t)} \text{ at iteration } t. \exists \tau > 0, \text{ such that } \forall \eta \in (0, \tau), \Omega(\tilde{A}^{(t+1)}) \geq \Omega(\tilde{A}^{(t)}), \text{ where } \tilde{A}^{(t+1)} = \mathcal{P}(\tilde{A}^{(t)} + \eta G^{(t)}) \text{ and } \mathcal{P}(\cdot) \text{ denotes the projection to the unit sphere.} \]

- Maximizing the lower bound with PGA can increase the mean of the angles

- Maximizing the lower bound with PGA can reduce the variance of the angles
Geometry Interpretation

- The gradient of the lower bound w.r.t $\tilde{a}_i$ is in the orthogonal complement of the space spanned by $\{\tilde{a}_1, \tilde{a}_2, \ldots, \tilde{a}_K\}/\{\tilde{a}_i\}$

**Lemma 2.** Let the weight vector $\tilde{a}_i$ of hidden unit $i$ be decomposed into $\tilde{a}_i = x_i + l_i e_i$, where $x_i = \sum_{j=1, j\neq i}^{K} \alpha_j \tilde{a}_j$ lies in the subspace $L$ spanned by $\{\tilde{a}_1, \ldots, \tilde{a}_K\}/\{\tilde{a}_i\}$, $e_i$ is in the orthogonal complement of $L$, $\|e_i\| = 1$, $e_i \cdot \tilde{a}_i > 0$, $l_i$ is a scalar. Then the gradient of $\Gamma(\tilde{A})$ w.r.t $a_i$ is $k_i e_i$, where $k_i$ is a positive scalar.
Case Study --- Restricted Boltzmann Machine with Diversity Regularization

\[
E(h, v) = - \sum_{j=1}^{J} \alpha_j v_j - \sum_{k=1}^{K} \beta_k h_k - \sum_{j=1}^{J} \sum_{k=1}^{K} A_{jk} v_j h_k
\]

\[
\max_{A} \quad L(D; A) + \lambda \Omega(A)
\]
Experiments

• Task: learn representations for documents

• Datasets

<table>
<thead>
<tr>
<th></th>
<th>#categories</th>
<th>#samples</th>
<th>vocab. size</th>
</tr>
</thead>
<tbody>
<tr>
<td>TDT</td>
<td>30</td>
<td>9394</td>
<td>5000</td>
</tr>
<tr>
<td>20-News</td>
<td>20</td>
<td>18846</td>
<td>5000</td>
</tr>
<tr>
<td>Reuters</td>
<td>9</td>
<td>7195</td>
<td>5000</td>
</tr>
</tbody>
</table>

• Baselines
  • Bag-of-Words (BOW); Latent Dirichlet Allocation (LDA); LDA regularized with Determinantal Point Process prior (DPP-LDA); Pitman-Yor Process Topic Model (PYTM); Latent IBP Compound Dirichlet Allocation (LIDA); Neural Autoregressive Topic Model (DocNADE); Paragraph Vector (PV); Restricted Boltzmann Machine

• Evaluation
  • Retrieval: precision@100
  • Clustering: accuracy
Retrieval Precision

Precision@100 on TDT Dataset

Precision@100 on 20-News Dataset

Precision@100 on Reuters Dataset
## Retrieval Precision

<table>
<thead>
<tr>
<th>Method</th>
<th>TDT</th>
<th>20-News</th>
<th>Reuters</th>
</tr>
</thead>
<tbody>
<tr>
<td>BOW</td>
<td>40.9</td>
<td>7.4</td>
<td>69.3</td>
</tr>
<tr>
<td>LDA</td>
<td>79.4</td>
<td>19.6</td>
<td>68.5</td>
</tr>
<tr>
<td>DPP-LDA</td>
<td>81.9</td>
<td>18.2</td>
<td>69.9</td>
</tr>
<tr>
<td>PYTM</td>
<td>78.7</td>
<td>20.1</td>
<td>70.6</td>
</tr>
<tr>
<td>LIDA</td>
<td>77.9</td>
<td>21.8</td>
<td>71.4</td>
</tr>
<tr>
<td>DocNADE</td>
<td>80.3</td>
<td>16.8</td>
<td>72.6</td>
</tr>
<tr>
<td>PV</td>
<td>81.7</td>
<td>19.1</td>
<td><strong>76.9</strong></td>
</tr>
<tr>
<td>RBM</td>
<td>47.4</td>
<td>22.3</td>
<td>70.1</td>
</tr>
<tr>
<td>DRBM</td>
<td><strong>84.2</strong></td>
<td><strong>24.9</strong></td>
<td>75.9</td>
</tr>
</tbody>
</table>
Clustering Accuracy

Accuracy on TDT Dataset

Accuracy on 20-News Dataset

Accuracy on Reuters Dataset
## Clustering Accuracy

<table>
<thead>
<tr>
<th>Method</th>
<th>TDT</th>
<th>20-News</th>
<th>Reuters</th>
</tr>
</thead>
<tbody>
<tr>
<td>BOW</td>
<td>51.3</td>
<td>21.3</td>
<td>49.7</td>
</tr>
<tr>
<td>LDA</td>
<td>45.2</td>
<td>21.9</td>
<td>51.2</td>
</tr>
<tr>
<td>DPP-LDA</td>
<td>46.3</td>
<td>10.9</td>
<td>49.3</td>
</tr>
<tr>
<td>PYTM</td>
<td>46.9</td>
<td>21.5</td>
<td>51.7</td>
</tr>
<tr>
<td>LIDA</td>
<td>47.3</td>
<td>17.4</td>
<td>53.1</td>
</tr>
<tr>
<td>DocNADE</td>
<td>45.7</td>
<td>18.7</td>
<td>48.7</td>
</tr>
<tr>
<td>PV</td>
<td>48.2</td>
<td>24.3</td>
<td>52.8</td>
</tr>
<tr>
<td>RBM</td>
<td>23.3</td>
<td>22.7</td>
<td>47.6</td>
</tr>
<tr>
<td>DRBM</td>
<td><strong>52.4</strong></td>
<td><strong>29.4</strong></td>
<td><strong>60.9</strong></td>
</tr>
</tbody>
</table>
## Improvement Breakdown

--- Retrieval on Reuters

<table>
<thead>
<tr>
<th>Category ID</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Documents</td>
<td>3713</td>
<td>2055</td>
<td>321</td>
<td>298</td>
<td>245</td>
<td>197</td>
<td>142</td>
<td>114</td>
<td>110</td>
</tr>
<tr>
<td>Precision@100 of RBM</td>
<td>0.69</td>
<td>0.44</td>
<td>0.09</td>
<td>0.10</td>
<td>0.06</td>
<td>0.04</td>
<td>0.04</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>Precision@100 of DRBM</td>
<td>0.90</td>
<td>0.80</td>
<td>0.31</td>
<td>0.40</td>
<td>0.27</td>
<td>0.23</td>
<td>0.09</td>
<td>0.14</td>
<td>0.13</td>
</tr>
<tr>
<td>Relative Improvement of DRBM over RBM</td>
<td><strong>31%</strong></td>
<td><strong>81%</strong></td>
<td><strong>245%</strong></td>
<td><strong>289%</strong></td>
<td><strong>324%</strong></td>
<td><strong>421%</strong></td>
<td><strong>148%</strong></td>
<td><strong>366%</strong></td>
<td><strong>397%</strong></td>
</tr>
</tbody>
</table>
Case Study --- Distance Metric Learning with Diversity Regularization

- Wide applications in retrieval, clustering and classification
Distance Metric Learning

- Projection matrix

$$A \in \mathbb{R}^{k \times d}$$

- Distance Metric Learning

$$\min_A \frac{1}{|S|} \sum_{(x,y) \in S} \|Ax - Ay\|^2$$

s.t. \(\|Ax - Ay\|^2 \geq 1, \forall (x, y) \in D\)

- Distance Metric Learning with Diversity Regularization

$$\min_A \frac{1}{|S|} \sum_{(x,y) \in S} \|Ax - Ay\|^2 - \lambda \Omega(A)$$

s.t. \(\|Ax - Ay\|^2 \geq 1, \forall (x, y) \in D\)
Experiments

• Datasets

<table>
<thead>
<tr>
<th>Feature Dim.</th>
<th>Training Data</th>
<th>Data Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>20-News</td>
<td>5000</td>
<td>11.3K</td>
</tr>
<tr>
<td>15-Scenes</td>
<td>1000</td>
<td>3.2K</td>
</tr>
<tr>
<td>6-Activities</td>
<td>561</td>
<td>7.4K</td>
</tr>
</tbody>
</table>

• Baselines

- Euclidean distance (EUC); Distance Metric Learning (DML); Large Margin Nearest Neighbor (LMNN) DML; Information Theoretical Metric Learning (ITML); Distance Metric Learning with Eigenvalue Optimization (DML-eig); Information-theoretic Semi-supervised Metric Learning via Entropy Regularization (Seraph)

• Evaluation

- Retrieval: precision
- Clustering: accuracy and normalized mutual information
- Classification: accuracy
Retrieval Precision

Retrieval Precision on 20-News

Precision (%)

Number of latent factors $k$

Retrieval Precision on 15-Scenes

Precision (%)

Number of latent factors $k$

Retrieval Precision on 6-Activities

Precision (%)

Number of latent factors $k$
<table>
<thead>
<tr>
<th>Method</th>
<th>20-News</th>
<th>15-Scenes</th>
<th>6-Activities</th>
</tr>
</thead>
<tbody>
<tr>
<td>EUC</td>
<td>62.8</td>
<td>65.3</td>
<td>85</td>
</tr>
<tr>
<td>DML</td>
<td>76.2</td>
<td>80.8</td>
<td>94.5</td>
</tr>
<tr>
<td>LMNN</td>
<td>67</td>
<td>70.3</td>
<td>71.5</td>
</tr>
<tr>
<td>ITML</td>
<td>74.7</td>
<td>79.1</td>
<td>94.2</td>
</tr>
<tr>
<td>DML-eig</td>
<td>71.2</td>
<td>71.3</td>
<td>86.7</td>
</tr>
<tr>
<td>Seraph</td>
<td>75.8</td>
<td>82</td>
<td>89.2</td>
</tr>
<tr>
<td>DDML</td>
<td><strong>81.1</strong></td>
<td><strong>83.6</strong></td>
<td><strong>96.2</strong></td>
</tr>
</tbody>
</table>
Clustering Accuracy

Clustering Accuracy on 20-News

Clustering Accuracy on 15-Scenes

Clustering Accuracy on 6-Activities

Accuracy (%) vs Number of latent factors k for DML and DDML.
## Clustering Accuracy

<table>
<thead>
<tr>
<th>Method</th>
<th>20-News</th>
<th>15-Scenes</th>
<th>6-Activities</th>
</tr>
</thead>
<tbody>
<tr>
<td>EUC</td>
<td>36.5</td>
<td>29</td>
<td>61.6</td>
</tr>
<tr>
<td>DML</td>
<td>28.4</td>
<td>40.1</td>
<td>76.1</td>
</tr>
<tr>
<td>LMNN</td>
<td>32.9</td>
<td>33.6</td>
<td>56.9</td>
</tr>
<tr>
<td>ITML</td>
<td>34.5</td>
<td>38.2</td>
<td>93.4</td>
</tr>
<tr>
<td>DML-eig</td>
<td>27.3</td>
<td>26.6</td>
<td>63.3</td>
</tr>
<tr>
<td>Seraph</td>
<td>48.1</td>
<td>48.2</td>
<td>74.8</td>
</tr>
<tr>
<td>DDML</td>
<td>44.6</td>
<td><strong>51.3</strong></td>
<td><strong>96.6</strong></td>
</tr>
</tbody>
</table>
Clustering Normalized Mutual Information

Clustering NMI on 20-News

- Number of latent factors $k$
- NMI (%)

Clustering NMI on 15-Scenes

- Number of latent factors $k$
- NMI (%)

Clustering NMI on 6-Activities

- Number of latent factors $k$
- NMI (%)

Graphs show the comparison of DML and DDML methods for different datasets and varying numbers of latent factors.
## Clustering Normalized Mutual Information

<table>
<thead>
<tr>
<th>Method</th>
<th>20-News</th>
<th>15-Scenes</th>
<th>6-Activities</th>
</tr>
</thead>
<tbody>
<tr>
<td>EUC</td>
<td>37.9</td>
<td>28.7</td>
<td>59.9</td>
</tr>
<tr>
<td>DML</td>
<td>38.2</td>
<td>42</td>
<td>83.6</td>
</tr>
<tr>
<td>LMNN</td>
<td>33.3</td>
<td>34.3</td>
<td>58.2</td>
</tr>
<tr>
<td>ITML</td>
<td>39.2</td>
<td>41.5</td>
<td>87</td>
</tr>
<tr>
<td>DML-eig</td>
<td>34</td>
<td>31.8</td>
<td>58.6</td>
</tr>
<tr>
<td>Seraph</td>
<td>49.7</td>
<td>47.5</td>
<td>71.1</td>
</tr>
<tr>
<td>DDML</td>
<td>51.1</td>
<td>48.9</td>
<td>91.9</td>
</tr>
</tbody>
</table>
3-NN Classification Accuracy

3-NN Accuracy on 20-News

3-NN Accuracy on 15-Scenes

3-NN Accuracy on 6-Activities
## 3-NN Classification Accuracy

<table>
<thead>
<tr>
<th></th>
<th>20-News</th>
<th>15-Scenes</th>
<th>6-Activities</th>
</tr>
</thead>
<tbody>
<tr>
<td>EUC</td>
<td>42.6</td>
<td>42.5</td>
<td>88.7</td>
</tr>
<tr>
<td>DML</td>
<td>57.5</td>
<td>51.7</td>
<td>95.1</td>
</tr>
<tr>
<td>LMNN</td>
<td>60.6</td>
<td>53.5</td>
<td>91.5</td>
</tr>
<tr>
<td>ITML</td>
<td>50.9</td>
<td>51.9</td>
<td>93.5</td>
</tr>
<tr>
<td>DML-eig</td>
<td>39.2</td>
<td>33.1</td>
<td>82.3</td>
</tr>
<tr>
<td>Seraph</td>
<td>67.9</td>
<td>55.2</td>
<td>91.4</td>
</tr>
<tr>
<td>DDML</td>
<td>64.5</td>
<td>58.8</td>
<td>96.5</td>
</tr>
</tbody>
</table>
Theoretical Analysis

- Generalization error of diversity regularized neural network: estimation error and approximation error
- Increasing the diversity can reduce estimation error
- A properly chosen diversity can reduce approximation error
Recap of Statistical Learning Theory

- **Setup**
  - Predict an output \( y \in Y \) given an input \( x \in X \)
  - Let \( \mathcal{H} \) be a set of hypotheses
  - Let \( \ell : (X \times Y) \times \mathcal{H} \rightarrow \mathbb{R} \) be a loss function
  - Let \( p^* \) be the distribution over \( X \times Y \)

- **Definitions**
  - Generalization error \( L(h) = \mathbb{E}_{(x,y) \sim p^*}[\ell((x,y),h)] \)
  - Expected risk minimizer \( h^* \in \arg\min_{h \in \mathcal{H}} L(h) \)
  - Empirical risk \( \hat{L}(h) = \frac{1}{n} \sum_{i=1}^{n} \ell \left((x^{(i)}, y^{(i)}), h\right) \)
  - Empirical risk minimizer \( \hat{h} \in \arg\min_{h \in \mathcal{H}} \hat{L}(h) \)
  - Generalization error of \( \hat{h} \)
    \[
    L(\hat{h}) = L(\hat{h}) - L(h^*) + L(h^*)
    \]
    - **Estimation Error**
    - **Approximation Error**
Neural Network
Setup of Diversity Regularized NN (DR-NN)

- **Setup**
  - Task: univariate regression
  - Network structure: input layer, one hidden layer, output layer
  - Activation function: rectified linear $h(x) = \max(0, x)$
  - Let $x \in \mathbb{R}^d$ be the input feature vector with $\|x\|_2 \leq C_1$
  - Let $y$ be the response value with $|y| \leq C_2$
  - Let $w_j \in \mathbb{R}^d$ be the weight vector of the $j$th hidden unit, $j = 1, \ldots, m$, with $\|w_j\|_2 \leq C_3$. Further, we assume the angle $\theta_{ij}$ between $w_i$ and $w_j$ is lower bounded by a constant $\theta$ for all $i \neq j$.
  - Let $\alpha \in \mathbb{R}^m$ be the weights connecting the hidden units to the output with $\|\alpha\|_2 \leq C_4$
  - Hypothesis: $f(x) = \sum_{j=1}^m \alpha_j h(w_j^T x)$, let $\mathcal{F}$ denote the hypothesis set
  - Loss function: $l(x, y, f) = (f(x) - y)^2$
Estimation Error of DR-NN

- **Theorem**

  With probability at least $1 - \delta$

  $$
  L(\hat{f}) - L(f^*) \leq 4C_1^2 C_3^2 C_4^2 m^2 \cos^2 \left( \frac{\theta}{2} \right) + 4C_1 C_2 C_3 C_4 \frac{\sqrt{m}}{\sqrt{n}} + 4C_5 + \sqrt{\frac{2\log(2/\delta)}{n}}
  $$

- **Extend to**
  - Multiple output
  - Multiple hidden layers
  - Other losses: hinge loss, logistic loss, cross entropy loss
  - Other activation functions: sigmoid, tanh
  - Convolutional neural networks

Larger diversity induces lower estimation error bound
Approximation Error of DR-NN

- Additional Setup
  - Let $G_\theta = \{g | g = \alpha h(w^T x), \|g\| \leq b_g, \rho(w_i, w_j) \geq \theta \text{ for all } i \neq j\}$
  - The hypothesis function $f_m$ is constructed iteratively:
    \[
    \begin{align*}
    f_1 &= g_1 \\
    f_m &= \beta f_{m-1} + (1 - \beta) g_m \\
        &\quad 0 < \beta \leq c < 1
    \end{align*}
    \]
  - The target function $f$ satisfies $\|f\| \leq b_f, \langle f, g \rangle < \infty$ for all $g \in G_\theta$
  - Approximation error: $\|f_m - f\|^2$
Approximation Error of DR-NN

- Theorem

Let $e$ denote

$$b_g^2 + \frac{1+c}{1-c} b_f^2 + \frac{2c}{1-c} b_g b_f + \frac{2c}{1-c} C_1^2 C_3^2 C_4^2 \cos^2 \left( \frac{\theta}{2} \right) V + \frac{2}{1-c} s(\theta)$$

where $V$ is the volume of the input L2 ball and $s(\theta) = \inf_{g \in G_\theta} |\langle f, g \rangle|$ is a non-decreasing function w.r.t $\theta$. Suppose $f_1$ is chosen to satisfy

$$\|f_1 - f\|^2 \leq \inf_{g \in G} \|g - f\|^2 + \epsilon_1$$

and iteratively $f_m$ is chosen to satisfy

$$\|f_m - f\|^2 \leq \inf_{0 < \beta \leq c} \inf_{g \in G_\theta} \|\beta f_{m-1} + (1 - \beta) g_m - f\|^2 + \epsilon_m$$

where $\epsilon_m \leq \frac{\rho c'}{m(m+\rho)}$, $c' \geq e$, $\rho = \frac{c'}{e} - 1$. Then for every $m \geq 1$,

$$\|f_m - f\|^2 \leq \frac{(\rho + 1)e}{m}$$

A proper diversity can reduce approximation error bound.
Sensitivity to Tradeoff Parameter

Sensitivity of DRBM to tradeoff parameter $\lambda$ on (a) TDT dataset (b) 20-News dataset (c) Reuters dataset
Conclusions

Diversity Regularization (DR) of Latent Variable Models
- An angle based diversity regularizer
- Capture long-tail factors
- Reduce model complexity while preserving expressiveness

Theory
- Generalization performance of DR-NN
- DR can reduce estimation and approximation errors

Optimization
- Smooth lower bound of the regularizer
- Maximizing the lower bound with gradient method can increase the regularizer in each iteration

Applications
- DR-RBM and DR-DML
- Strong empirical performance
Thank you!
Questions?

Papers/slides/code/documents are available at my webpage http://www.cs.cmu.edu/~pengtaox/