On Learning Sparse Structured Input-Output Models

Eric Xing
epxing@cs.cmu.edu
Machine Learning Dept. / Language Technology Inst. / Computer Science Dept.
Carnegie Mellon University

Acknowledgement:
Amr Ahmed, Jacob Eisenstein, Xi Chen, Qinong Ho, Seyoung Kim, Seunghak Lee, Andre Martins, Noah Smith, and Jun Zhu

Unstructured Prediction Problem

\[ x = (x_1, x_2, \ldots) \quad \Rightarrow \quad y = y_1 \]
Classical Predictive Models

- Input and output space: $\mathcal{X} \equiv \mathbb{R}^d$, $\mathcal{Y} \equiv \{-1,+1\}$
- Predictive function $h(x): y^* = h(x) \equiv \text{arg max}_{y \in \mathcal{Y}} F(x, y; w)$
- Examples: $F(x, y; w) = g(w^T f(x, y))$
- Learning: $w = \text{arg min}_{w \in W} \ell(x, y; w) + \lambda R(w)$

where $\ell(\cdot)$ represents a convex loss, and $R(w)$ is a regularizer preventing overfitting

---

From Unstructured to Structured Prediction

- Binary classification: black-and-white decisions
- Multi-class classification: the world of technicolor

- can be reduced to several binary decisions, but...
- often better to handle multiple classes directly
- how many classes? 2? 5? exponentially many?
- Structured prediction: many classes, strongly interdependent

- Example: image segmentation (number of classes exponential to the # of segments)
  
  $x = \begin{pmatrix} x_{11} & x_{12} & \ldots \\ x_{21} & x_{22} & \ldots \\ \vdots & \vdots & \ddots \end{pmatrix} \Rightarrow y = \begin{pmatrix} y_{11} & y_{12} & \ldots \\ y_{21} & y_{22} & \ldots \\ \vdots & \vdots & \ddots \end{pmatrix}$
Example I: Dependency Parsing of Sentences

$ 0
We 1
sail 2
tonight 3
for 4
Singapore 5

Challenge:
Structured outputs, and globally constrained to be a valid tree

Example II: Text Summarization

Australian novelist Peter Carey was awarded the coveted Booker Prize for fiction Tuesday night for his love story, “Oscar and Lucinda”. A panel of five judges unanimously announced the award of the $26,250 prize after an 80-minute deliberation during a banquet at London’s ancient Guildhall.
The judges made their selection from 102 books published in Britain in the past 12 months and which they read in their homes.
Carey, who lives in Sydney with his wife and son, said in a brief speech that like the other five finalists he had been asked to attend with a short speech in his pocket in case he won.
Example III: Web-Data Extraction

Example IV: Topic Discovery/Extraction
### Structured Prediction Graphical Models

- **Input and output space:** \( X = R_{X_1} \times \ldots \times R_{X_K} \) \( y = R_{Y_1} \times \ldots \times R_{Y_{K'}} \)
- **Convex loss function**
  - Conditional Random Fields (CRFs) (Lafferty et al 2001)
    - Based on a **Logistic Loss** (LR)
    - Max-likelihood estimation (point-estimate)
    - \( \mathcal{L}(D; w) = \log \sum_{y'} \exp(w^T f(x, y')) - w^T f(x, y) \)
  - Max-margin Markov Networks (M3Ns) (Taskar et al 2003)
    - Based on a **Hinge Loss** (SVM)
    - Max-margin learning (point-estimate)
    - \( \mathcal{L}(D; w) = \log \max_{y'} w^T f(x, y') - w^T f(x, y) + \ell(y', y) \)
- **Markov properties are encoded in the feature functions** \( f(x, y) \)
  \( F(x, y; w) = g(w^T f(x, y)) \)

### Challenges:
- **Sparse “Interpretable”** prediction model
- **Prior** information of structures
- **Latent** structures/variables
- **Time** series and non-stationarity
- **Scalable** to large-scale problems (e.g., \( 10^4 \) or larger input/output dimension)
Main Claims

- The sparse structures of natural language data (input) and of the NLP tasks (output) can be utilized to improve the quality of the solution and interpretability of the solution.

- Over-parameterized models such as conventional NB/LR/SVM-style classifiers or parsers, topic models, or the related spectrum methods are not benefiting from sparse structures.

- It is desirable to explore model spaces with structured sparsity for both predictive models (e.g., classifiers, parsers) and explorative models (e.g., topic models).

Outline

- Sparse Structured Input-Output Models
  - ... supervised learning
  - ... convex optimization and log loss
  - ... Frequentist-style shrinkage via regularization

- Sparse Topic Models
  - ... unsupervised learning
  - ... non-convex and likelihood-driven
  - ... Bayesian-style posterior inference

- Sparse and Discriminative Topic Models?
  - ... toward jointly explorative and predictive learning
Basic text classification

<table>
<thead>
<tr>
<th>Class Label</th>
<th>Word counts</th>
<th>Feature strength</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>x</td>
<td>?</td>
</tr>
</tbody>
</table>

\[ y = X \beta \]

Basic text classification

<table>
<thead>
<tr>
<th>Class Label</th>
<th>Word counts</th>
<th>Feature strength</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>x</td>
<td>?</td>
</tr>
</tbody>
</table>

\[ \beta^* = \arg\min_{\beta} (y - X\beta)^T (y - X\beta) \]

Many non-zero coefficients: Which words are truly significant?
Sparsity: In a mathematical sense

- Consider least squares linear regression problem:
- Sparsity means most of the beta’s are zero.

\[
\hat{\beta} = \arg\min_{\beta} \|Y - X\beta\|^2
\]
subject to:
\[
\sum_{j=1}^{p} I[|\beta_j| > 0] \leq C
\]

- But this is not convex!!! Many local optima, computationally intractable.

L1 Regularization (LASSO) (Tibshirani, 1996)

- A convex relaxation.

**Constrained Form**
\[
\hat{\beta} = \arg\min_{\beta} \|Y - X\beta\|^2
\]
subject to:
\[
\sum_{j=1}^{p} |\beta_j| \leq C
\]

**Lagrangian Form**
\[
\hat{\beta} = \arg\min_{\beta} \|Y - X\beta\|^2 + \lambda \|\beta\|_1
\]

- Still enforces sparsity!
Lasso for Sparse Regression

\[ \beta^* = \arg \min_{\beta} (y - X\beta)^T (y - X\beta) + \lambda \sum |\beta_i| \]

Many zero associations (sparse results), but what if the problem has "structures"?

Input Structure: the WordNet

- The “network” of synsets in NL
  - Nodes (synsets) represent distinct concept
  - Links represent conceptual-semantic and lexical relations
  - Hidden knowledge and structure among concepts

Statistical challenge

How to find important words to a predictor or a topic from a graph?

- Prior knowledge
- Context
Output Structure: Task Hierarchy

• E.g., the tree hierarchy in the DMOZ repository of the PASCAL Large Scale Hierarchical Text Classification challenge

Statistical challenge

How to train multiple labelers that are “related” by a tree?

Sparse Structured Input/Output Lasso for Multi-task Learning

Multi-Class Label

Word counts

Feature strength

\[ \beta^* = \arg \min_{\beta} (y - X\beta)^T (y - X\beta) + \lambda \sum_{j=1}^J |\beta_j| \]

How to combine information across multiple features/classes to increase the power?
**Sparse Structured Input/Output Lasso for Multi-task Learning**

\[ \beta^* = \arg\min_{\beta} (y - X\beta)^T (y - X\beta) + \lambda \sum_{j=1}^{J} |\beta_j| \]

We introduce Structured fusion and/or group norm penalties

**Tree-Guided Group Lasso**

- In a simple case of two concepts

![Diagram](EMNLP 2012)
Tree-Guided Group Lasso

• In a simple case of two concepts

\[ C_1 = \{ \beta_{j_1}, \beta_{j_2} \} \]

Select the child nodes jointly or separately?

\[ \text{Tree-guided group lasso} \]

\[ \arg\min (y - X\beta)'(y - X\beta) + \lambda \sum_j \left[ h(|\beta_{j_1}| + |\beta_{j_2}|) + (1 - h) \left( \sqrt{\beta_{j_1}^2 + \beta_{j_2}^2} \right) \right] \]

\( L_1 \) penalty
• Lasso penalty
• Separate selection

\( L_2 \) penalty
• Group lasso
• Joint selection

Elastic net

Tree-Guided Group Lasso

• For a general tree

\[ C_2 = \{ \beta_{j_1}, \beta_{j_2}, \beta_{j_3} \} \]

Select the child nodes jointly or separately?

\[ \text{Tree-guided group lasso} \]

\[ \arg\min (y - X\beta)'(y - X\beta) + \lambda \sum_j \left[ (1 - h_2)(\sqrt{\beta_{j_1}^2 + \beta_{j_2}^2 + \beta_{j_3}^2} + h_2(|C_1| + |C_2|)) + (1 - h_1)(\sqrt{\beta_{j_1}^2 + \beta_{j_2}^2} + h_1(|\beta_{j_1}| + |\beta_{j_2}|)) \right] \]

Joint selection
Separate selection
Proposition 1 For each of the k-th output (gene), the sum of the weights $w_v$ for all nodes $v \in V$ in $T$ whose group $G_v$ contains the k-th output (gene) as a member equals one. In other words, the following holds:

$$\sum_{v \in G_v} w_v = \prod_{m \in \text{Ancestors}(v_0)} h_m + \sum_{t \in \text{Ancestors}(v_k)} (1 - h_t) \prod_{m \in \text{Ancestors}(v_t)} h_m = 1.$$
Full GM-based Loss Functions

\[ y^* = h(x) \triangleq \arg \max_{y \in Y} F(x, y; w) \]
\[ F(x, y; w) = g(w^T f(x, y)) \]

Represent factorization assumptions:
\[ P(y|x) = \frac{1}{Z} \prod_l \psi_l(y_l) \prod_{\alpha} \Psi_{\alpha}(y_{\alpha}) \]

Inference: compute the MAP, marginals \( \mu_l(y_l) \) and \( \mu_{\alpha}(y_{\alpha}) \), \( Z \)

- tractable when \( \mathcal{G} \) is a tree, often intractable otherwise

Optimization

Original Problem:
\[ \arg \max_\beta \equiv L(\{x_i, y_i\}; \beta) + \Omega(\beta) \]

Existing Methods:

| Method                          | 2nd-order, computationally heavy | \( \lambda \sum_{g \in \mathcal{G}} w_g \| \beta_g \|_2 \) \( \Rightarrow \lambda \sum_{g \in \mathcal{G}} w_g t_g \)
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Interior-point Method (IPM) for Second-order Cone Programming (SOCP) or Quadratic Programming (QP)</td>
<td>( \lambda \sum_{g \in \mathcal{G}} | \beta_g |_2 ) ( \leq t_g )</td>
<td>( \text{Cannot be easily be applied. Hard to compute the subgradient} )</td>
</tr>
</tbody>
</table>
New Optimization Framework

- **Main Difficulties:**
  - Complex loss $\mathcal{L}(\{x_i, y_i\}; \beta)$, (e.g., GMs with intractable factors or loopy graphs)
  - Intractable inference
  - Complex shrinkage $\Omega(\beta)$, (e.g., overlapping group penalties)
  - Non-differentiable, non-separable

- **Our approaches:**
  - Alternating Direction Dual Decomposition (AD$^3$) [Martins et al, ICML 2011]
  - Proximal Gradient [Chen et al, AOAS 2012]

- **Large number of training examples**
  - Parallel computation
  - Map-Reduce on computing gradient
  - Map: calculate gradient on single example
  - Reduce: gather gradients computed by all map procedures, and calculate the sum
  - New multi-core framework

Alternating Directions Dual Decomposition

- Convergent to the primal and dual solutions (Glowinski and Le Tallec, 1989)
- $O(1/\varepsilon)$ iterations suce for $\varepsilon$-accurate objective (He and Yuan, 2011)
- Solution is always sparse (only $O(|N(\alpha)|)$ nonzeros)
- Active set methods: seek the support of the solution by adding/removing components; very suitable for warm-starting (Nocedal and Wright, 1999)
Smooth Proximal Gradient Descent

Original Problem:
\[ \arg \min_{\beta \in \mathbb{R}^J} f(\beta) \equiv \frac{1}{2} \|y - X\beta\|_2^2 + \Omega(\beta) \]
\[ \Omega(\beta) = \max_{\alpha \in Q} \alpha^T C \beta \]

Approximation Problem:
\[ \arg \min_{\beta \in \mathbb{R}^J} \tilde{f}(\beta) \equiv \frac{1}{2} \|y - X\beta\|_2^2 + f_\mu(\beta) \]
\[ f_\mu(\beta) = \max_{\alpha \in Q} \alpha^T C \beta - \mu \delta(\alpha) \]

Gradient of the Approximation:
\[ \nabla \tilde{f}(\beta) = X^T(X\beta - y) + C^T \alpha^* \]
\[ \alpha^* = \arg \max_{\alpha \in Q} \alpha^T C \beta - \mu \delta(\alpha) \]
\[ \nabla \tilde{f}(\beta) \text{ is Lipschitz continuous with the Lipschitz constant } L \]
\[ L = \lambda_{\max}(X^T X) + L_\mu \]

Convergence Rate

Theorem: If we require \( f(\beta^t) - f(\beta^*) \leq \epsilon \) and set \( \mu = \frac{\epsilon}{2D} \), the number of iterations is upper bounded by:
\[ t \leq \sqrt{\frac{4\|\beta^*\|_2^2}{\epsilon}} \left( \lambda_{\max}(X^T X) + \frac{2D\|\Gamma\|_2^2}{\epsilon} \right) = O\left( \frac{1}{\epsilon^2} \right) \]

Remarks: state of the art IPM method for SOCP converges at a rate \( O\left( \frac{1}{\epsilon^2} \right) \)

Time complexity (Per-iteration):
\[ O(\mathcal{P}(K + J \sum_{i \in \mathbb{E}} |i|)) \text{ vs. } O(\mathcal{P}(K + \|g\|)(KN + JJ)) \]

(a) \hspace{5cm} (b) \hspace{5cm} (c)
What if the structure becomes too complex?

- Too many groups in real problems:
  \[
  \beta_{\text{optim}} = \arg \min_{\beta} \sum_{i} \sum_{j} (y_i - \sum_{j} \beta_{ij} x_i) + \lambda \sum \sum |\beta_{ij}|
  \]
  - Recall that even SPG has a complexity of \(O(J^2 K + J \sum_{E \in G} |g|)\)
  - And an optimization procedure must
    - minimize our objective function, and
    - induce correct sparsity patterns
  - Hierarchical group-thresholding:
    - an algorithmic approach to directly reduce search space of sparsity, while optimizes the exact loss

Hierarchical Group-Thresholding

- DAG for Sparsity Patterns
  - All sparsity patterns of a 2x2 matrix:
    - A DAG of inclusion relation relationships of sparsity
  - Hierarchical Group-Thresholding
    - Initialize \(B\) using ridge regression
    - Step 1: Traversing DAG, check the optimality condition of the zero pattern at each node. If the condition holds, set zero
    - Step 2: Update non-zero regression coefficients using coordinate descent
What if the function is non-linear?

**Group Sparse Additive Models** [Ying, Chen, Xing, ICML 2012]

- Assume G is a partition of \{1, \cdots, p\}, i.e., the groups in G do not overlap.
- The optimization problem is
  \[
  \min_{f} L(f) + \lambda \Omega_{\text{group}}(f),
  \]
  where \( \Omega_{\text{group}}(f) = \sum_{g \in G} \| f_g \| = \sum_{g \in G} \sqrt{|g|} \sum_{j \in g} \mathbb{E} [ f_j^2(X_j) ]. \)
- Non-trivial to solve due to
  - correlation structure of component functions within the group
  - non-smoothness of functional group penalty

---

**Toward Human-Level Intelligence**

- Now we have dealt with high feature dimension
  - Sparsity
- and we have know how to leverage structural knowledge
  - Structured shrinkage
- What about massive **concept space**?
Output Coding

- Every class is now represented by a bit-string
  - **Coding**: a codeword is assigned to each class
  - **Decoding**: given test data, look for most similar class codeword
- Predict bit by bit through binary or ternary classifier – this is much easier than the 1 vs C-1 classifier
- Decoding the bit-string – error correcting

\[
\begin{array}{ccccccc}
M & 1 & 2 & 3 & \ldots & \ldots & K \\
1 & 1 & 1 & 1 & 0 & 0 & 0 \\
2 & -1 & 0 & 0 & 1 & 1 & 0 \\
\vdots & 0 & -1 & 0 & -1 & 0 & 1 \\
C & 0 & 0 & -1 & 0 & -1 & -1 \\
\end{array}
\]

Learning the Coding Matrix

- Accuracy of base binary classifiers for bit-prediction
  - Use category hierarchy for a measure of separability
  - Large intra-partite similarity + small inter-partite similarity
- Strong error-correcting ability
  - Maximize distance between rows of coding matrix
- Fault tolerance
  - Introduction of ignored classes: {-1,0,+1} instead of {-1,+1}

\[
\begin{aligned}
\max_{\mathbf{B}} & \quad F_b(\mathbf{B}) - \lambda_r F_r(\mathbf{B}) - \lambda_c \sum_{i=1}^{L} \| \beta_i \|_2^2 \\
\text{s.t.} & \quad \mathbf{B} \in \{-1, 0, +1\}^{K \times L} \\
\end{aligned}
\]

\[
\begin{align*}
\sum_{i=1}^{L} (|\mathbf{B}_i| + \mathbf{B}_i) & \geq 2, \forall i = 1, \ldots, L \\
\sum_{i=1}^{L} (|\mathbf{B}_i| - \mathbf{B}_i) & \geq 2, \forall i = 1, \ldots, L \\
\sum_{i=1}^{L} |\mathbf{B}_i| & \geq 1, \forall i = 1, \ldots, K \\
\end{align*}
\]
Probabilistic Decoding

- Output code can have real semantic meaning
  - E.g., encoding a tree path in a label taxonomy

- Probabilistic decoding:
  - bit $i$ depend on bit $j$ probabilistically

- Define prior $P(y_i|y = k)$ using tree hierarchy
  - Graph coloring: all nodes participating in $i$-th bit prediction are colored (red for positive, black for negative)
  - Task: what is the probability of node $k$ being colored red?

Multi-Way Classification Accuracy

- DMOZ repository in PASCAL Hierarchical Text Classification challenge

<table>
<thead>
<tr>
<th>Data set</th>
<th># classes</th>
<th># training data</th>
<th># test data</th>
<th># features</th>
</tr>
</thead>
<tbody>
<tr>
<td>DMOZ-small</td>
<td>1139</td>
<td>6323</td>
<td>1858</td>
<td>1199848</td>
</tr>
<tr>
<td>DMOZ-large</td>
<td>12294</td>
<td>93805</td>
<td>34905</td>
<td>1199856</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>DMOZ small</th>
<th>DMOZ large</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top 1</td>
<td>Top 5</td>
</tr>
<tr>
<td>OVR</td>
<td>50.91</td>
<td>64.72</td>
</tr>
<tr>
<td>RDOC</td>
<td>41.77</td>
<td>54.52</td>
</tr>
<tr>
<td>RSOC</td>
<td>42.30</td>
<td>58.40</td>
</tr>
<tr>
<td>SpectralOC</td>
<td>44.83</td>
<td>59.10</td>
</tr>
<tr>
<td>SSOC</td>
<td>56.67</td>
<td>67.33</td>
</tr>
</tbody>
</table>
Discovering Sociolinguistic Associations on Twitter

- Twitter Garden Hose feed from March 1-7, 2010
- 9250 authors, 380,000 messages, 4.7 million tokens
- Filters:
  - At least 20 messages (in Garden Hose)
  - Messages must include GPS within a USA zipcode
  - No more than 1000 followers, followees
- GPS → Zipcode → U.S. Census Demographic Statistics
  - Zipcodes commonly proxy for demographics in public health.
  - Careful! Twitter users are not an unbiased sample from a zipcode.

Demographic multi-prediction

\[ \mathbf{X} \Theta \approx \mathbf{Y} \]

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>full</td>
<td>5418</td>
<td>0.260</td>
<td>0.337</td>
<td>0.318</td>
<td>0.206</td>
<td>0.384</td>
<td>0.296</td>
<td>0.256</td>
<td>0.155</td>
<td>0.113</td>
<td>0.295</td>
<td>0.152</td>
</tr>
<tr>
<td>multi-output lasso</td>
<td></td>
<td>0.290</td>
<td>0.326</td>
<td>0.306</td>
<td>0.304</td>
<td>0.383</td>
<td>0.303</td>
<td>0.240</td>
<td>0.153</td>
<td>0.111</td>
<td>0.302</td>
<td>0.156</td>
</tr>
<tr>
<td>SVM</td>
<td></td>
<td>0.237</td>
<td>0.321</td>
<td>0.299</td>
<td>0.269</td>
<td>0.352</td>
<td>0.272</td>
<td>0.226</td>
<td>0.138</td>
<td>0.081</td>
<td>0.270</td>
<td>0.136</td>
</tr>
<tr>
<td>highest variance</td>
<td></td>
<td>0.220</td>
<td>0.309</td>
<td>0.287</td>
<td>0.245</td>
<td>0.315</td>
<td>0.248</td>
<td>0.190</td>
<td>0.132</td>
<td>0.065</td>
<td>0.250</td>
<td>0.135</td>
</tr>
<tr>
<td>most frequent</td>
<td></td>
<td>0.204</td>
<td>0.294</td>
<td>0.264</td>
<td>0.224</td>
<td>0.293</td>
<td>0.229</td>
<td>0.178</td>
<td>0.129</td>
<td>0.073</td>
<td>0.228</td>
<td>0.126</td>
</tr>
</tbody>
</table>

EMNLP 2012

[Eisenstein, Smith, and Xing. ACL 2011] 41

EMNLP 2012

[Eisenstein, Smith, and Xing. ACL 2011] 42
Sociolinguistic Associations

Dependency Parsing

Datasets from CoNLL-2006 and CoNLL-2008 shared tasks
Outline

- Sparse Structured Input-Output Models
  - ... supervised learning
  - ... convex optimization and log loss
  - ... Frequentist-style shrinkage via regularization

- Sparse Topic Models
  - ... unsupervised learning
  - ... non-convex and likelihood-driven
  - ... Bayesian-style posterior inference

- Sparse and Discriminative Topic Models?
  - ... toward jointly explorative and predictive learning

Modeling Semantics: e.g., Topic Models

Generating a document

- Draw \( \theta \) from the prior
  For each word \( n \)
    - Draw \( z_n \) from \( \text{multinomial}(\theta) \)
    - Draw \( w_n | z_n, \{ \beta_k \} \) from \( \text{multinomial}(\beta_{z_n}) \)

- Prior over topic Vector
  - Latent Dirichlet Allocation (LDA)
  - Correlated priors (CTM)
  - Hierarchical priors

- Topics
  - Unigram, bigrams, etc

- Document structure
  - Bag of words
  - Multi-modal
Modeling and Inference Complexity

- What if we want to combine latent topics with additional facets, such as geography in an unsupervised fashion?

- Additional latent variables decide which facet is responsible for each token (e.g. Ahmed and Xing 2010).

- That's twice as many latent variables per document!

Compact modes needed on mobile devices
Sparse Additive Generative Models

versus

Model Compression on Text

- NIPS dataset: 1986 training docs, 10K vocabulary

Adaptive sparsity:
- 5% non-zeros for 10 topics
- 1% non-zeros for 50 topics
Sparse Topical Coding

- **Goal:** design a non-probabilistic topic model that is amenable to
  - direct control on the posterior sparsity of inferred representations
  - avoid dealing with normalization constant when considering supervision or rich features
  - seamless integration with a convex loss function (e.g., svm hinge loss)

- We extend sparse coding to hierarchical sparse topical coding
  - word code \( \theta \)
  - document code \( s \)

\[
\min_{\{\theta_d, s_d\}} \sum_{d,n \in I_d} \ell(w_{dn}, s_{dn} \beta_k) + \lambda \sum_d \|\theta_d\|_1 + \sum_d (\gamma \|s_{dn} - \theta_d\|_2^2 + \rho \|s_{dn}\|_1)
\]

- reconstruction loss
- sparse codes
- \( \theta_d \geq 0, s_{dn} \geq 0, \forall d, n \in I_d; \beta_k \in \mathcal{P}, \forall k, \)
- non-negative codes
- topical bases
- truncated aggregation

Algorithms on Sparse Latent Space Models

- **Complex objective**
  - Non-convex, but often bi-convex
  - Often additional non-negativity constraints other than sparsity

- **Hierarchical sparse coding**
  - Greedy algorithm for the non-convex \( L_0 \) "pseudo-norm":
    - select the element with maximum correlation with the residual
    - known as ‘matching pursuit’ (Mallat & Zhang, 1993)
  - For the convex \( L_1 \) norm, many algorithms:
    - Soft-thresholding with coordinate descent (Friedman et al., 2007; Zhu & Xing, 2011)
    - Proximal methods (Nesterov, 2007; Jenatton et al., 2010, Chen et al. 2011)
    - Active-set methods (Roth & Fischer, 2008)
    - Online/stochastic variants
    - ...

- **Dictionary (topic) learning**
  - projected gradient descent
  - any faster alternative method can be used
Comparisons

LDA vs. STC

[Zhu and Xing, UAI 2011]

![Graph comparing LDA and STC]

Sparse word codes

- Sparsity ratio: percentage of zeros

![Graph showing sparsity ratio]

- NMF: non-negative matrix factorization
- MedLDA (Zhu et al., 2009)
- regLDA: LDA with entropic regularizer
- gaussSTC: use L2 rather than L1-norm
Outline

- Sparse Structured Input-Output Models
  - supervised learning
  - convex optimization and log loss
  - Frequentist-style shrinkage via regularization

- Sparse Topic Models
  - unsupervised learning
  - non-convex and likelihood-driven
  - Bayesian-style posterior inference

- Sparse and Discriminative Topic Models?
  - toward jointly explorative and predictive learning

Predictive Subspace Learning with Supervision

- Unsupervised latent subspace representations are generic but can be sub-optimal for predictions
- Many datasets are available with supervised side information
  - Tripadvisor Hotel Review (http://www.tripadvisor.com)
  - LabelMe (http://labelme.csail.mit.edu/)
  - Many others
  - Flickr (http://www.flickr.com/)

- Can be noisy, but not random noise (Ames & Naaman, 2007)
  - labels & rating scores are usually assigned based on some intrinsic property of the data
  - helpful to suppress noise and capture the most useful aspects of the data

- Goals:
  - Discover latent subspace representations that are both predictive and interpretable by exploring weak supervision information
MLE versus Max-margin Learning

- Likelihood-based estimation
  - Probabilistic (joint/conditional likelihood model)
  - Easy to perform Bayesian learning, and incorporate prior knowledge, latent structures, missing data
  - Bayesian or direct regularization
  - Hidden structures or generative hierarchy
- Max-margin learning
  - Non-probabilistic (concentrate on input-output mapping)
  - Not obvious how to perform Bayesian learning or consider prior, and missing data
  - Support vector property, sound theoretical guarantee with limited samples
  - Kernel tricks

Maximum Entropy Discrimination (MED) (Jaakkola, et al., 1999)

- Model averaging
  \[ \hat{y} = \text{sign} \int p(\mathbf{w}) F(x; \mathbf{w}) \, d\mathbf{w} \quad (y \in \{+1, -1\}) \]
- The optimization problem (binary classification)
  \[
  \min_{p(\mathbf{w})} \int KL[p(\Theta)||\mu_p(\Theta)]
  \text{s.t.} \quad \int p(\Theta)[g,F(x; \mathbf{w}) - \zeta]d\Theta \geq 0, \forall i. \\
\]
where \( \Theta \) is the parameter \( \mathbf{w} \) when \( \xi \) are kept fixed or the pair \( (\mathbf{w}, \xi) \) when we want to optimize over \( \xi \)

MaxEnt Discrimination Markov Network
(Zhu et al, ICML 2008, Zhu and Xing, JMLR 2009)

- Structured MaxEnt Discrimination (SMED):
  \[ P_1 : \min_{p(\mathbf{w}), \xi} KL[p(\mathbf{w})||p_0(\mathbf{w})] + U(\xi) \]
  s.t. \( p(\mathbf{w}) \in \mathcal{F}_1, \xi \geq 0, \forall i. \)
  generalized maximum entropy or regularized KL-divergence

- Feasible subspace of weight distribution:
  \[ \mathcal{F}_1 = \{ p(\mathbf{w}) : \int p(\mathbf{w})[\Delta F_i(y; \mathbf{w}) - \Delta\mathbf{f}_i(y)]d\mathbf{w} \geq -\xi, \forall i, \forall y \neq y' \}, \]
  expected margin constraints.

- Average from a distribution of M^3Ns
  \[ h_{\lambda}(x; p(\mathbf{w})) = \arg \max_{y \in \tilde{y}(x)} \int p(\mathbf{w}) F(x, y; \mathbf{w})d\mathbf{w} \]
Maximum Entropy Discrimination LDA (MedLDA)

- Bayesian sLDA:

- MED Estimation:
  - MedLDA Regression Model
    
    \[
    P_1(\text{MedLDA}^*) : \min_{\alpha, \beta, \phi, \phi^*} \mathcal{L}(\phi) + C \sum_{d=1}^D (\xi_d + \xi_d^*)
    \]
    
    s.t. \forall d : \begin{align*}
    y_d - E[\phi^T Z_d] &\leq \xi_d, \quad \xi_d^* \\
    -y_d + E[\phi^T Z_d] &\leq \xi_d + \xi_d^*, \quad \mu_d \\
    \xi_d &\geq 0, \quad v_d \\
    \xi_d^* &\geq 0, \quad v_d^*
    \end{align*}

  - MedLDA Classification Model
    
    \[
    P_2(\text{MedLDA}^*) : \min_{\alpha, \beta, \phi, \phi^*} \mathcal{L}(\phi) + C \sum_{d=1}^D \xi_d
    \]
    
    s.t. \forall d, y_d \neq y_d^* : \begin{align*}
    E[\phi^T \Delta \xi_d(y)] &\geq 1 - \xi_d; \quad \xi_d \geq 0.
    \end{align*}

(Zhu et al, ICML 2009, JMLR 2012)

Document Modeling

- Data Set: 20 Newsgroups
- 110 topics + 2D embedding with t-SNE (var der Maaten & Hinton, 2008)
Document Modeling

comp.graphics

politics.mideast

EMNLP 2012

Classification

- **Data Set:** 20Newsgroups
  - Binary classification: “alt.atheism” and “talk.religion.misc” (Simon et al., 2008)
  - Multiclass Classification: all the 20 categories
- **Models:** DiscLDA, sLDA (Binary ONLY! Classification sLDA (Wang et al., 2009)), LDA + SVM (baseline), MedLDA, MedLDA+SVM
- **Measure:** Relative Improvement Ratio

\[ RR(M) = \frac{\text{precision}(M)}{\text{precision}(LDA + SVM)} - 1 \]
Regression

- **Data Set**: Movie Review (Blei & McAuliffe, 2007)
- **Models**: MedLDA\(_{\text{partial}}\), MedLDA\(_{\text{full}}\), sLDA, LDA+SVR
- **Measure**: predictive $R^2$ and per-word log-likelihood

\[ pR^2 = 1 - \frac{\sum_d (y_d - \hat{y}_d)^2}{\sum_d (y_d - \bar{y}_d)^2} \]

---

Time Efficiency

- **Binary Classification**
  - MedLDA is comparable with LDA+SVM

- **Multiclass**:
  - MedLDA is comparable with LDA+SVM
  - Regression:
    - MedLDA is comparable with sLDA
Supervised STC

- Joint loss minimization

$$\min_{\{\theta_d\}, \{s_d\}, \beta, \eta} \quad f(\{\theta_d\}, \{s_d\}, \beta) + CR_\lambda(\{\theta_d\}, \eta) + \frac{1}{2} ||\eta||^2$$

s.t.: \(\theta_d \geq 0, \forall d; s_{dn} \geq 0, \forall d, n \in I_d; \beta_k \in \mathcal{P}, \forall k; \)

- coordinate descent alg. applies with closed-form update rules
- No sum-exp function; seamless integration with non-probabilistic large-margin principle

Classification accuracy

- 20 newsgroup data:
Summary: Margin-based Learning Paradigms

**SVM**

\[ y = \text{sign}(w^T x + b) \]

min \( \frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i \)

s.t. \( y(x_i, w) \geq 1 - \xi_i, \forall i \)

**MED**

\[ y = \text{sign}(w^T f(x)) \]

min \( K L(p||p_0) + C \sum_{i=1}^{n} \xi_i \)

s.t. \( y(x_i, w) \geq 1 - \xi_i, \forall i \)

**MPN**

\[ y^* = \arg \max_w w^T f(x, y; w) \]

\[ y^* = \arg \max_y w^T f(x, y; w) \]

min \( \frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i \)

s.t. \( w^T f(y, w) - \Delta(y) \geq -\xi_i, \forall i, y \neq y' \)

**MaxEnDNet**

\[ y^* = \arg \max_w w^T f(x, y; w) \]

\[ y^* = \arg \max_y w^T f(x, y; w) \]

min \( K L(p||p_0) + C \)

s.t. \( \int p(w|\Delta f(y, w) - \Delta f(y|w) \geq -\xi_i, y \neq y') \)

Conclusion and Challenges: Learning Sparse Structured Input/Output Models

- **Models**: Multi-Task Classifiers, Graphical Models, Topic Models
- **Algorithm**: Convex and non-convex optimization
- **Representation**: Concept encoding/decoding
- **System**: multi-core, distributed file system, shared memory, cloud
- **Theory**: Convergence, sample complexity, asymptotic consistency/sparsistency, error bounds, etc
Thanks!

Reference: