Dual-Decomposed Learning with Factorwise Oracles for Structured Prediction of Large Output Domain

Xiangru Huang *

Joint work [1] with
Ian E.H. Yen†, Kai Zhong*, Ruohan Zhang*, Chia Dai†,
Pradeep Ravikumar† and Inderjit Dhillon*.

* University of Texas at Austin
† Carnegie Mellon University

[1] [1] Dual Decomposed Learning with Factorwise Oracle for Structural SVM of Large Output Domain. NIPS 2016.

Outline

Problem Setting

- Classification: learn function $g : \mathcal{X} \rightarrow \mathcal{Y}$

Problem Setting

- Classification: learn function $g : \mathcal{X} \to \mathcal{Y}$
- Structural: Assuming structured dependencies on output $g : \mathcal{X} \to \mathcal{Y}_1 \times \mathcal{Y}_2 \times \cdots \times \mathcal{Y}_m$

Example: Sequence Labeling

- Unigram Factor:
  $\theta_u : \mathcal{Y}_t \times \mathcal{X}_t \to \mathcal{R}$
- Bigram Factor:
  $\mathcal{Y}_b = \mathcal{Y}_{t-1} \times \mathcal{Y}_t$
  $\theta_b : \mathcal{Y}_b \to \mathcal{R}$

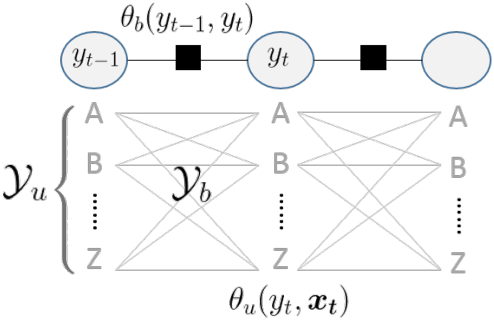

Figure: Sequence Labeling

Example: Multi-Label Classification with Pairwise Interaction

- Unigram Factor :
  $\theta_u : \mathcal{Y}_k \times \mathcal{X} \to \mathcal{R}$
- Bigram Factor :
  $\mathcal{Y}_b = \mathcal{Y}_k \times \mathcal{Y}_{k'}$
  $\theta_b : \mathcal{Y}_b \to \mathcal{R}$

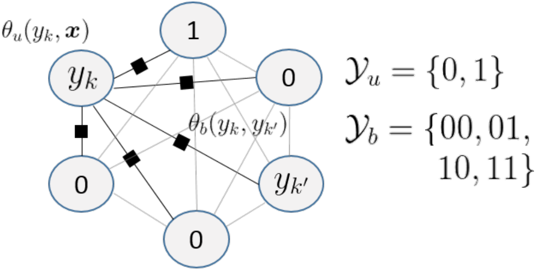

Figure: Multi-Label with Pairwise Interaction

Motivations

- $g : \mathcal{X} \to \mathcal{Y}_1 \times \mathcal{Y}_2 \times \cdots \times \mathcal{Y}_m$

Motivations

- $g : \mathcal{X} \to \mathcal{Y}_1 \times \mathcal{Y}_2 \times \cdots \times \mathcal{Y}_m$
- Learning requires inference per iteration.
- Exact inference is slow: each iteration takes $\mathrm{O}(|\mathcal{Y}_i|^n)$ for n-gram factor, where $|\mathcal{Y}_i| \geq 3000$.

Motivations

- $g : \mathcal{X} \to \mathcal{Y}_1 \times \mathcal{Y}_2 \times \cdots \times \mathcal{Y}_m$
- Learning requires inference per iteration.
- Exact inference is slow: each iteration takes $O(|\mathcal{Y}_i|^n)$ for n-gram factor, where $|\mathcal{Y}_i| \geq 3000$.
- Approximation downgrades performance.

Key Idea: Dual Decomposed Learning

- Structural Oracle (joint inference) is too expensive.

Key Idea: Dual Decomposed Learning

- Structural Oracle (joint inference) is too expensive.
- Reduce Structural SVM to Multiclass SVMs via soft enforcement of consistency between factors.

Key Idea: Dual Decomposed Learning

- Structural Oracle (joint inference) is too expensive.
- Reduce Structural SVM to Multiclass SVMs via soft enforcement of consistency between factors.
- (Cheap) Active Sets + Factorwise Oracles + Message Passing (between factors).

Key Idea: Factorwise Oracles

- **Inner-Product (unigram) Factor**: $\theta_w(x, y) = \langle w_y, x \rangle$.
  - Reduces to a primal and dual sparse Extreme Multiclass SVM .
  - Reduce $O(\underbrace{D}_{\text{feat. dim.}} \cdot |\mathcal{Y}_i|)$ to $O(\underbrace{|\mathcal{F}_u|}_{\#\text{uni. fac.}} \cdot |\mathcal{A}_i|)$ (details see [2]) [2].

- **Indicator (bigram) Factor**: $\theta(y_1, y_2) = v_{y_1, y_2}$.
  - Maintain Priority Queue on v_{y_1, y_2}.
  - Reduce $O(|\mathcal{Y}_1||\mathcal{Y}_2|)$ to $O(\underbrace{|\mathcal{A}_1||\mathcal{A}_2|}_{\text{active set sizes}})$.

[2] [2] PD-Sparse: A Primal and Dual Sparse Approach to Extreme Multiclass and Multilabel Classification. ICML 2016.

Methodology Sketch

- Original problem:

$$\min_w \frac{1}{2}\|w\|^2 + \underbrace{C\sum_{i=1}^{n} L(w; x_i, y_i)}_{\text{struct hinge loss}}$$

[3] Simon Julien et al. Block-Coordinate Frank-Wolfe Optimization for Structural SVMs. ICML 2013.

Methodology Sketch

- Original problem:

$$\min_{w} \frac{1}{2}\|w\|^2 + \underbrace{C\sum_{i=1}^{n} L(w; x_i, y_i)}_{\text{struct hinge loss}}$$

- Dual-Decomposed into independent problems:

$$\min_{\alpha_f \in \Delta^{|\mathcal{Y}_f|}} G(\alpha) := \underbrace{\frac{1}{2}\sum_{F} \|\sum_{f\in F} \phi(x_f, y_f)^T \alpha_f\|^2 - \sum_{j\in\mathcal{V}} \delta_j^T \alpha_j}_{\text{Independent Multiclass SVMs}}$$

with consistency constraints

$$M_{if}\alpha_f = \alpha_i, \quad \forall (i,f) \in \mathcal{E}.$$

- Standard approach [3] finds feasible descent direction, which however needs joint inference.

[3] Simon Julien et al. Block-Coordinate Frank-Wolfe Optimization for Structural SVMs. ICML 2013.

Methodology Sketch

- Dual-Decomposed into independent problems:

$$\min_{\alpha_f \in \Delta^{|\mathcal{Y}_f|}} G(\alpha) := \frac{1}{2} \sum_F \| \sum_{f \in F} \phi(x_f, y_f)^T \alpha_f \|^2 - \sum_{j \in \mathcal{V}} \delta_j^T \alpha_j$$

with consistency constraints

$$M_{jf} \alpha_f = \alpha_j, \quad \forall (j, f) \in \mathcal{E}$$

- Augmented Lagrangian Method:

$$\mathcal{L}(\alpha, \lambda) := \underbrace{\sum_F G_F(\alpha_F)}_{\text{indep. multiclass SVMs}} + \underbrace{\frac{\rho}{2} \sum_{(j,f) \in \mathcal{E}} \| M_{jf} \alpha_f - \alpha_j + \lambda_{jf}^t \|^2}_{\text{messages between factors (sparse)}}$$

with incremental updated multipliers

$$\lambda_{jf}^{t+1} = \lambda_{jf}^t + \eta (M_{jf} \alpha_f^{t+1} - \alpha_j^{t+1})$$

Methodology Sketch

- Augmented Lagrangian Method:

$$\mathcal{L}(\alpha, \lambda) := \underbrace{\sum_F G_F(\alpha_F)}_{\text{indep. multiclass SVMs}} + \underbrace{\frac{\rho}{2} \sum_{(j,f)\in\mathcal{E}} \|M_{jf}\alpha_f - \alpha_j + \lambda_{jf}^t\|^2}_{\text{messages between factors (sparse)}}$$

with incremental updated multipliers

$$\lambda_{jf}^{t+1} = \lambda_{jf}^t + \eta(M_{jf}\alpha_f^{t+1} - \alpha_j^{t+1})$$

- Update α and λ alternatively.

Experiments: Sequence Labeling (on ChineseOCR)

- Chinese OCR: $N = 12,064$, $T = 14.4$, $D = 400$, $K = 3,039$.
- $|\mathcal{Y}_b| = 3,039^2 = 9,235,521$ (bigram language model).
- Decoding: Viterbi Algorithm.

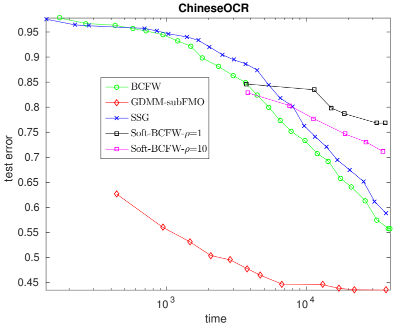

Figure: Test Error

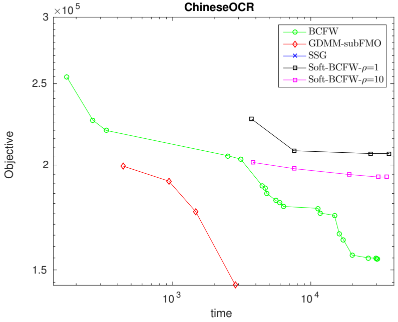

Figure: Objective

Experiments: Multi-Label Classification (on RCV1)

- RCV-1: $N = 23,149$, $D = 47,236$, $K = 228$.
- $|\mathcal{F}_b| = 228^2 = 51,984$ (pairwise interaction).
- Decoding: Linear Program

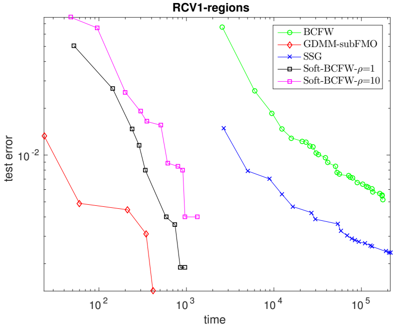

Figure: Test Error

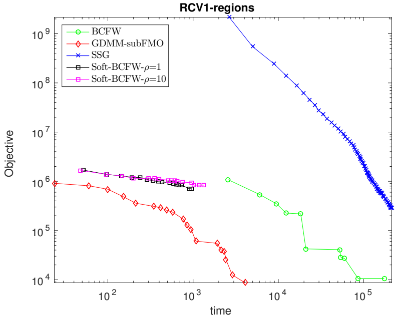

Figure: Objective