Exploiting Primal and Dual Sparsity for Extreme Classification \(^1\)

Ian E.H. Yen \(^*\dagger\)

Joint work with Xiangru Huang\(^\dagger\), Kai Zhong\(^\dagger\), Pradeep Ravikumar\(^*\dagger\) and Inderjit Dhillon\(^\dagger\)

\(^*\) Machine Learning Department
Carnegie Mellon University

\(^\dagger\) Department of Computer Science
University of Texas at Austin

\(^1\)PD-Sparse: A Primal and Dual Sparse Approach to Extreme Multiclass and Multilabel Classification. ICML 2016.
Outline

1. Extreme Multiclass & Multilabel Classification
2. Joint Primal and Dual Sparsity
3. Dual Block-Coordinate Frank-Wolfe (Dual-BCFW)
4. Experimental Results
Extreme Multiclass & Multilabel Classification

Goal

Learn a function $h(x) : \mathbb{R}^D \rightarrow \mathbb{R}^K$ from $D$ input features to $K$ output scores consistent with the ground-truth labels $y \in \{0, 1\}^K$.

- We consider problems with large $K$ (e.g. $10^3 \sim 10^6$).
- Let $\mathcal{P}(y) = \{k | y_k = 1\}$ and $\mathcal{N}(y) = \{k | y_k = 0\}$.
- **Multiclass:** $|\mathcal{P}(y)| = 1$. **Multilabel:** $0 < |\mathcal{P}(y)| \ll K$ typically.

**Linear Classification:**

$$h(x) := W^T x \text{ where } W : D \times K.$$

- Easily extended to nonlinear setting via Random Features $\phi(x)$.
Challenge

Standard approaches (1-vs-All, 1-vs-1, Multiclass SVM) require prohibitive $\Omega(NDK)$ cost for training/prediction.

- **Existing Approach 1: Low-Rank Embedding**
  
  \[ h(x) = W^T x = (UV^T)x. \]

- **Existing Approach 2: Tree**
  
  Group \( \{w_1, w_2, \ldots, w_K\} \) via hierarchy.
  
  - Searching for the best tree could be difficult.
  
  When structural assumption does not hold \( \Rightarrow \) **lower accuracy** than one-vs-all approach.

- If \( \text{nnz}(x) \ll D \), low-rank approach could have \( \text{nnz}(V^T x) > \text{nnz}(x) \).

Question

Can we make model \( h(x) \) **compact** without sacrificing accuracy?
Outline

1. Extreme Multiclass & Multilabel Classification
2. Joint Primal and Dual Sparsity
3. Dual Block-Coordinate Frank-Wolfe (Dual-BCFW)
4. Experimental Results
## Dual Sparsity

### Binary SVM
- Prediction score $z \in \mathbb{R}$ and label $y \in \{-1, 1\}$.
  \[
  L(z, y) = \max(1 - yz, 0)
  \]
- Dual solution $\alpha \neq 0 \iff 1 - yz$ attains the maximum.
- In practice, $L(z, y) > 0$ for a significant fraction of samples $\Rightarrow$ dual solution $\alpha \in \mathbb{R}^N$ has $\text{nnz}(\alpha) \propto N$.

### Multiclass SVM (Crammer & Singer, 2001)
- Prediction score $z \in \mathbb{R}^K$ and label $y \in [K]$.
  \[
  L(z, y) = \max_{k \in [K] \setminus y} (1 + z_k - z_y)_+
  \]
- Dual solution $\alpha_k \neq 0 \iff 1 + z_k - z_y > 0$ attains the maximum ($k$ is a **Support Label**).
- In practice, dual solution $\alpha \in \mathbb{R}^{N \times K}$ has $\text{nnz}(\alpha) \ll NK$ for large $K$. 
Joint Primal & Dual Sparsity

Max-Margin Loss for Multilabel (Crammer & Singer, 2003)

\[ L(z, y) = \max_{k_n \in \mathcal{N}(y), k_p \in \mathcal{P}(y)} (1 + z_{k_n} - z_{k_p})^+ \]

- The \( \mathbf{W}^* \in \mathbb{R}^{D \times K} \) obtained from minimization of Max-Margin Loss

\[
\min_{\mathbf{W}} \sum_{i=1}^{N} L(\mathbf{W}^T \mathbf{x}_i, \mathbf{y}_i)
\]

is determined by the scores of **Support Labels**:

\[ (k_n, k_p) \in \arg\max_{k_n \in \mathcal{N}(y), k_p \in \mathcal{P}(y)} (1 + z_{k_n} - z_{k_p})^+ \]

that attain the maximum.
Joint Primal & Dual Sparsity

Max-Margin Loss (Crammer & Singer, 2003)

\[
L(z, y) = \max_{k_n \in \mathcal{N}(y), k_p \in \mathcal{P}(y)} \left( 1 + z_{k_n} - z_{k_p} \right) +
\]

- Optimal \( W^* \) should satisfy \( N k_A \) constraints where \( k_A \) is the average \#Support Labels per sample (\( k_A \ll K \)).
- \( D K \gg N k_A \Rightarrow W^* \) is under-determined \( \Rightarrow \) Can we find a sparse \( W^* \) with minimum loss?
Joint Primal & Dual Sparsity

Theorem (Joint Primal & Dual Sparsity)

For any $\lambda > 0$ and $\{x_i\}_{i=1}^N$ drawn from continuous probability distribution, 

$$W^* \in \text{argmin}_{W} \lambda \sum_{k=1}^K \|w_k\|_1 + \sum_{i=1}^N L(W^T x_i, y_i)$$

satisfies $D_k W = \text{nnz}(W^*) \leq \text{nnz}(A^*) = N k_A$, where $A^* : N \times K$ is the optimal solution of the dual problem.
Joint Primal & Dual Sparsity

**$\ell_1-\ell_2$ Regularization**

For ease of optimization, we solve the $\ell_1-\ell_2$-regularized objective

$$
\min_W \sum_{k=1}^{K} \frac{1}{2} \| w_k \|^2 + \lambda \| w_k \|_1 + C \sum_{i=1}^{N} L(W^T x_i, y_i),
$$

which gives sparsity pattern similar to $\ell_1$-regularized objective empirically.

<table>
<thead>
<tr>
<th>Data sets</th>
<th>$k_A$: #support labels/sample</th>
<th>$k_W$: #nonzero/feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>EUR-Lex (K=3,956)</td>
<td>20.73</td>
<td>45.24</td>
</tr>
<tr>
<td>LSHTC-wiki (K=320,338)</td>
<td>18.24</td>
<td>20.95</td>
</tr>
<tr>
<td>LSHTC (K=12,294)</td>
<td>7.15</td>
<td>4.88</td>
</tr>
<tr>
<td>aloi.bin (K=1,000)</td>
<td>3.24</td>
<td>0.31</td>
</tr>
<tr>
<td>bibtex (K=159)</td>
<td>18.17</td>
<td>1.94</td>
</tr>
<tr>
<td>Dmoz (K=11,947)</td>
<td>5.87</td>
<td>0.116</td>
</tr>
</tbody>
</table>
Outline

1. Extreme Multiclass & Multilabel Classification
2. Joint Primal and Dual Sparsity
3. Dual Block-Coordinate Frank-Wolfe (Dual-BCFW)
4. Experimental Results
Dual Block-Coordinate Frank-Wolfe (Dual-BCFW)

Dual Form of $\ell_1$-$\ell_2$-regularized problem

$$\min_{\alpha^i \in C_i, \ i \in [N]} G(\alpha) := \frac{1}{2} \sum_{k=1}^{K} \| w_k(\alpha_k) \|^2 + \sum_{i=1}^{N} e_i^T \alpha^i$$

where $w_k(\alpha_k) := \text{prox}_{\lambda \| \cdot \|_1}(X^T \alpha_k)$ and $C_i$ is a $(P_i, N_i)$-bi-simplex.

- Smooth objective $G(\alpha) + \text{Block-separable constraints } \alpha_i \in C_i$.
  $\Rightarrow$ Minimize w.r.t. one block $\alpha_i$ at a time.

- **Challenge:** How to identify Support labels and Active Features?

- **Solution:**
  Leverage **primal sparsity** to search active **dual variables**. Leverage **dual sparsity** to update active **primal variables**.
Dual Block-Coordinate Frank-Wolfe (Dual-BCFW)

Search active set $A_i$ via sparse $W$; maintain $W(\alpha)$ via sparse $\Delta \alpha^i$.

$O(\text{nnz}(x_i)\text{nnz}(w^j) + \text{nnz}(x_i)\text{nnz}(\alpha^i))$ cost per iteration.
Dual Block-Coordinate Frank-Wolfe (Dual-BCFW)

Costs $O(\text{nnz}(X)k_W + \text{nnz}(X)k_A)$ per pass of data; $O(1/\epsilon)$ passes needed to have $\frac{1}{N}(G(\alpha) - G^*) \leq \epsilon$.

We know $Dk_W \approx Nk_A$ so the search time could dominate if $D \ll N$.

In such case, we use **sampling techniques** to speed up the search step.
Dual-BCFW: Efficient Implementation

- **Fast prediction** by \( \langle w_k, x_i \rangle = \sum_{j \in nz(x_i)} x_{ij} w^j \)

  ![Diagram of matrix multiplication](image)

  to exploit sparsity of both \( W \) and \( x_i \) \( \Rightarrow O(nnz(x_i)k_W) \)
  (compared to \( O(nnz(x_i)r + rK) \) in low-rank approach of rank \( r \)).

- **Space-efficiency**: \( DK \) space is not acceptable while maintaining \( W = \text{prox}(X^T A) \) requires random access.

  We use \( D \) hash tables sharing a hash function \( \Rightarrow \) Evaluate once for \( nnz(x_i) \) updates.
Outline

1. Extreme Multiclass & Multilabel Classification
2. Joint Primal and Dual Sparsity
3. Dual Block-Coordinate Frank-Wolfe (Dual-BCFW)
4. Experimental Results
Experimental Result: Multiclass

- **1-vs-All** takes long time to train; **Multi-SVM** could be out of memory (> 300G).
- **Low-rank** and **Tree** assumption could hurt accuracy (**FastXML** ensembles 50 trees to re-gain accuracy).
- **Low-rank** could even lead to slower prediction for sparse feature vectors.
- **PD-Sparse** reduces training, prediction time by orders of magnitude without sacrificing accuracy.
**Experimental Result: Multilabel**

- **1-vs-All** takes $> 3$ months to train. Even storing models is a problem ($\approx 870G$).
- **PD-Sparse** takes $\approx 1$ day to train, and has $\approx 10\%$ higher accuracy than tree-based and low-rank approaches, with orders of magnitude faster prediction.

(Please see more results in the paper)
Extreme Classification is inherently Primal and Dual sparse when $N$, $D$, $K$ are large.

A Dual BCFW algorithm can identify active dual variables by leveraging sparsity in the primal, and vice versa, thus resulting in complexity sublinear to $K$.

Experiment shows the PD-Sparse approach reduces training and prediction time by orders of magnitude without sacrificing accuracy.