PD-Sparse: A Primal and Dual Sparse Approach to Extreme Multiclass & Multilabel Classification

Ian E.H. Yen*, Xiangru Huang*, Kai Zhong, Pradeep Ravikumar, Inderjit S. Dhillon

Abstract

- Extreme Classification problems with huge number of classes are prevalent in applications of Machine Learning.
- Existing approaches exploit structural relation (e.g. low-rank, tree-hierarchy) among labels, which could hurt accuracy when assumption not holds.
- We show Extreme Classification model can be inherently Primal and Dual sparse, which allows a Dual-BCFW solver with cost sublinear to #classes.

Extreme Multiclass & Multilabel Classification

- **Goal:** Learning a function $h(x) : \mathbb{R}^D \to \mathbb{R}^K$ from D input features to K output scores that is consistent with labels $y \in \{0,1\}^K$.

- K is large (e.g. 10^3~10^6).
- Let $\mathcal{P}(y) = \{k | y_k = 1\}$ and $\mathcal{N}(y) = \{k | y_k = 0\}$.
- Multiclass: $|\mathcal{P}(y)| = 1$; Multilabel: $|\mathcal{P}(y)| \ll K$.

- We consider Linear Classification:

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

- Could extend to nonlinear setting via Random Features method.

Existing Approaches

Challenge: When K is large, training and prediction of standard approaches (1vsA, 1vs1, etc.) require $O(NDK)$ cost, which is prohibitive.

- **Approach 1:**Low-Rank Embedding— $h(x) = W^T x = (UV^T)x$.
  - If $nnz(x) \ll D$, we could have $nnz(V^T x) > nnz(x)$.
- **Approach 2:**Tree— group $\{w_1, w_2, ..., w_K\}$ via hierarchy.
  - Search the best tree could be difficult.
- When low-rank/tree-structure assumption not holds $\Rightarrow$ Could have **lower accuracy** than one-vs-all approach.

Max-Margin Loss (Crammer & Singer, 2003)

Question: Can we make model compact without sacrificing accuracy?

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

- Minimizers of the Max-Margin Loss

$$W^* \in \arg\min_W \sum_{i=1}^N L(W^T x_i, y_i)$$

satisfy constraints from Support Labels $\mathcal{P}_i^S, \mathcal{N}_i^S$ attaining the maximum:

$$L(z_i, y_i) = 1 + z_{k_n} - z_{k_p}, \quad k_n \in \mathcal{N}_i^S, k_p \in \mathcal{P}_i^S.$$

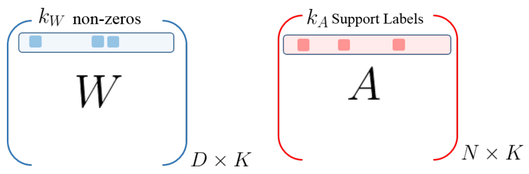
- There are $\mathbf{Nk_A}$ Support Labels ($k_A \ll K$) while there are $\mathbf{DK} \gg \mathbf{Nk_A}$ parameters $\Rightarrow$ Could we find a **sparse W*** with minimum loss?

Theorem (Joint Primal & Dual Sparsity)

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

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

satisfies $Dk_W = \mathbf{nnz(W^*)} \leq \mathbf{nnz(A^*)} = \mathbf{Nk_A}$, where $\mathbf{A^* : N \times K}$ is the optimal solution of the dual problem.



ℓ_1-ℓ_2 Regularization

In practice for ease of optimization, we solve the ℓ_1-ℓ_2-regularized objective

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

Table 1. Sparsity Level for best-accuracy parameter λ.

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

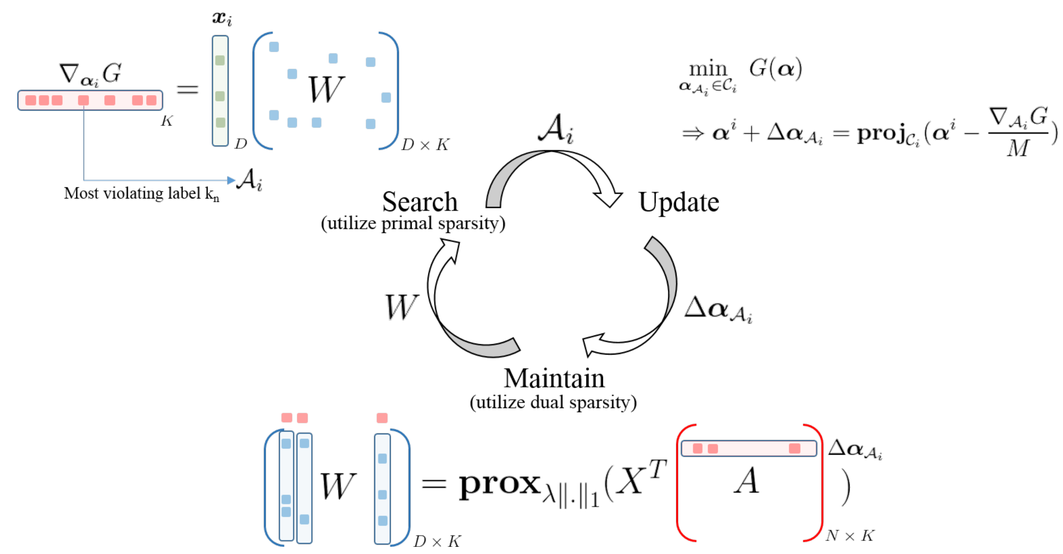
Dual Form of ℓ_1-ℓ_2-regularized problem

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

where $w_k(\alpha_k) := \mathbf{prox}_{\lambda\|.\|_1}(X^T \alpha_k)$ and $\mathcal{C}_i$ is a $(\mathcal{P}_i, \mathcal{N}_i)$-bi-simplex.

Dual Block-Coordinate Frank-Wolfe (Dual-BCFW)

- Smooth objective $G(\alpha)$ + Block-separable constraints $\alpha_i \in \mathcal{C}_i$.
  $\Rightarrow$ Minimize w.r.t. one block of variables α_i at a time.
- **Challenge:** (i) Maintaining $w_k(\alpha_k)$, $\forall k \in [K]$ is expensive.
  (ii) Support labels are not known a-priori.
- **Solution:** (i) Leverage primal sparsity to select active dual variables.
  (ii) Leverage dual sparsity to maintain primal variables.



- Search active set $\mathcal{A}_i$ via sparse W; maintain $W(\alpha)$ via sparse $\Delta\alpha^i$.
- $O(\underbrace{nnz(x_i)nnz(w^j)}_{search} + \underbrace{nnz(x_i)nnz(\alpha^i)}_{maintain})$ costs per iteration.

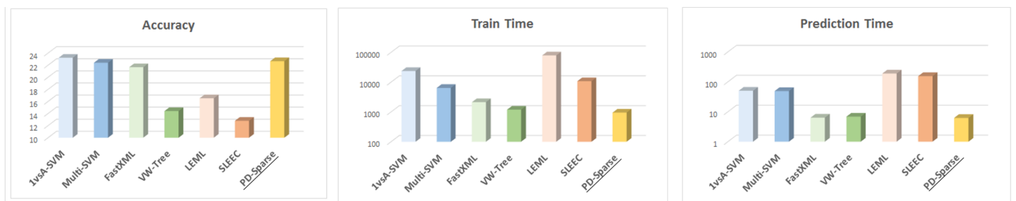
- $O(nnz(X)k_W + nnz(X)k_A)$ per pass of data; $O(1/\epsilon)$ passes needed to have $\frac{1}{N}(G(\alpha) - G^*) \leq \epsilon$.
- We know $Dk_W \approx Nk_A$ so the search time could dominate if $D \ll N$.
- In practice, we use sampling techniques to further speed up the search step.

Experimental Comparison

- **1vsA-SVM** : LibLinear one-vs-all SVM (Dual-CD solver).
- **Multi-SVM** : Liblinear Cramer & Singer Multiclass SVM (Dual-BCD solver).
- **FastXML** : An Extreme Multilabel solver (Prabhu Varma, KDD'14) that organizes models with tree structure (a default 50 trees' ensemble is used).
- **VW-tree** : A Tree-based Extreme Multiclass solver in Vowpal-Wabbit (Choromanska & Langford, NIPS'15).
- **LEML**: A low-rank Multilabel solver (Yu et al., ICML 2013). We chose rank to be lowest from $\{50, 100, 250, 500, 1000\}$ that gives heldout accuracy close to the best.
- **SLEEC**: Local Embeddings for Extreme Multilabel (Bhatiaet al., NIPS'15) (9 default parameters used).

Experimental Result: Multiclass

LSHTC-1
N= 8x10^4
D= 3x10^5
K= 1x10^5

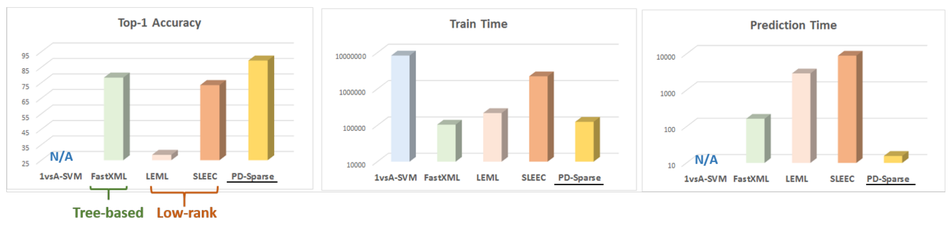


Dmoz
N=3x10^5
D=8x10^5
K=1x10^5

- **1-vs-All** takes long time to train; **Multi-SVM** could be out of memory ($> \mathbf{300}G$).
- Structural assumption could hurt accuracy (FastXML ensembles 50 trees to re-gain accuracy).
- **Low-rank embedding** could lead to slower prediction for sparse feature vectors.
- **PD-Sparse** reduces training, prediction time by orders w/o scraficing accuracy.

Experimental Result: Multilabel

LSHTC-wiki
N= 2x10^6
D= 2x10^6
K= 3x10^5



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