Exploiting Primal and Dual Sparsity for Extreme Classification [1]

Ian E.H. Yen *†

Joint work with Xiangru Huang†, Kai Zhong†,
Pradeep Ravikumar*† and Inderjit Dhillon†

* Machine Learning Department
Carnegie Mellon University

† 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

Extreme Multiclass & Multilabel Classification

Goal

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

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

- Linear Classification:

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

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

Extreme Multiclass & Multilabel Classification

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—** $\boldsymbol{h}(\boldsymbol{x}) = W^T \boldsymbol{x} = (UV^T)\boldsymbol{x}$.
- **Existing Approach 2: Tree—** group $\{\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_K\}$ 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 $nnz(\boldsymbol{x}) \ll D$, low-rank approach could have $nnz(V^T\boldsymbol{x}) > nnz(\boldsymbol{x})$.

Question

Can we make model $\boldsymbol{h}(\boldsymbol{x})$ **compact** without sacrificing accuracy?

Outline

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 $\boldsymbol{\alpha} \in \mathbb{R}^N$ has $nnz(\boldsymbol{\alpha}) \propto N$.

Multiclass SVM (Crammer & Singer, 2001)

- Prediction score $\boldsymbol{z} \in \mathbb{R}^K$ and label $y \in [K]$.

$$L(\boldsymbol{z}, y) = \max_{k \in [K] \setminus y} (1 + z_k - z_y)_+$$

- 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 $\boldsymbol{\alpha} \in \mathbb{R}^{N \times K}$ has $\mathbf{nnz}(\boldsymbol{\alpha}) \ll \mathbf{NK}$ for large K.

Joint Primal & Dual Sparsity

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

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

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

$$\min_{W} \sum_{i=1}^{N} L(W^T \boldsymbol{x}_i, \boldsymbol{y}_i)$$

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

$$(k_n, k_p) \in \mathop{argmax}_{k_n \in \mathcal{N}(\boldsymbol{y}),\, k_p \in \mathcal{P}(\boldsymbol{y})} (1 + z_{k_n} - z_{k_p})_+$$

that attain the maximum.

Joint Primal & Dual Sparsity

Max-Margin Loss (Crammer & Singer, 2003)

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

- Optimal $\mathbf{W}^*$ should satisfy $\mathbf{Nk_A}$ constraints where $\mathbf{k_A}$ is the average **#Support Labels** per sample $(k_A \ll K)$.
- $\mathbf{DK} \gg \mathbf{Nk_A} \Rightarrow \mathbf{W}^*$ is under-determined $\Rightarrow$ Can we find a **sparse** $\mathbf{W}^*$ with minimum loss?

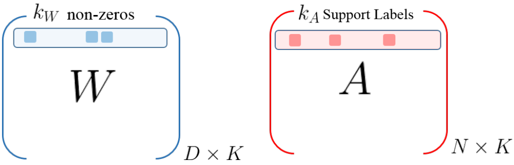
Joint Primal & Dual Sparsity

Theorem (Joint Primal & Dual Sparsity)

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

$$W^* \in \underset{W}{argmin}\ \lambda \sum_{k=1}^{K} \|\mathbf{w}_k\|_1 + \sum_{i=1}^{N} L(W^T \mathbf{x}_i, \mathbf{y}_i)$$

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



Joint Primal & Dual Sparsity

ℓ_1-ℓ_2 Regularization

For ease of optimization, we solve the ℓ_1-ℓ_2-regularized objective

$$\min_{W} \sum_{k=1}^{K} \frac{1}{2}\|\boldsymbol{w}_k\|^2 + \lambda\|\boldsymbol{w}_k\|_1 + C\sum_{i=1}^{N} L(W^T\boldsymbol{x}_i, \boldsymbol{y}_i),$$

which gives sparsity pattern similar to ℓ_1-regularized objective empirically.

Sparsity Level when λ is tuned for best accuracy:

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

Outline

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

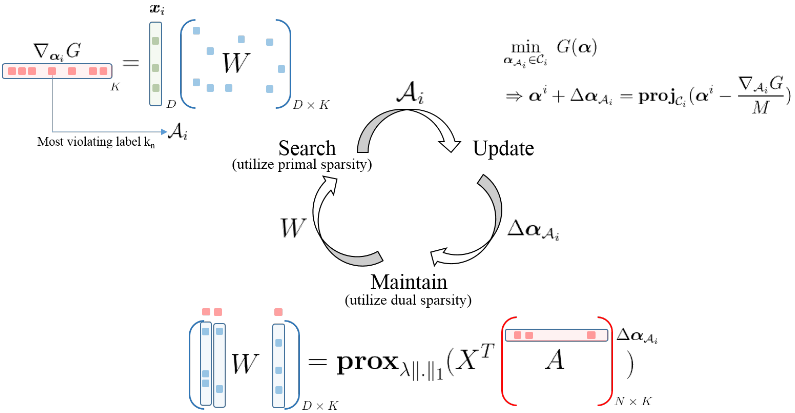
Dual Form of ℓ_1-ℓ_2-regularized problem

$$\min_{\boldsymbol{\alpha}^i \in \mathcal{C}_i,\, i\in[N]} \quad G(\boldsymbol{\alpha}) := \frac{1}{2}\sum_{k=1}^{K} \|\mathbf{w}_k(\boldsymbol{\alpha}_k)\|^2 + \sum_{i=1}^{N} \mathbf{e}_i^T \boldsymbol{\alpha}^i$$

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

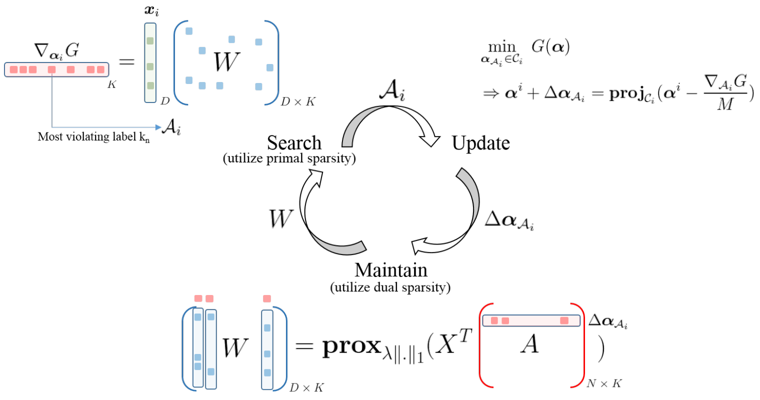
- Smooth objective $G(\boldsymbol{\alpha})$ + Block-separable constraints $\boldsymbol{\alpha}_i \in \mathcal{C}_i$.
  $\Rightarrow$ Minimize w.r.t. one block $\boldsymbol{\alpha}_i$ at a time.
- **Challenge:** How to identify Support labels and Active Features?
- **Solution:**
  Leverage **primal sparsity** to search active **dual variables**. Leverage **dual sparsity** to update active **primal variables**.

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



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

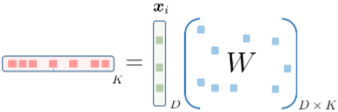
Dual Block-Coordinate Frank-Wolfe (Dual-BCFW)



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

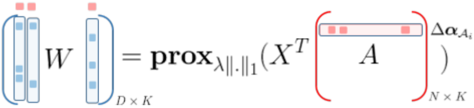
Dual-BCFW: Efficient Implementation

- **Fast prediction** by $\langle \boldsymbol{w}_k, \boldsymbol{x}_i \rangle = \sum_{j \in nz(\boldsymbol{x}_i)} x_{ij} \boldsymbol{w}^j$



to exploit sparsity of both W and $\boldsymbol{x}_i \Rightarrow O(nnz(\boldsymbol{x}_i) k_W)$ (compared to $O(nnz(\boldsymbol{x}_i) r + rK)$ in low-rank approach of rank r).

- **Space-efficiency**: DK space is not acceptable while maintaining $W = \textbf{prox}(X^T A)$ requires random access.
- We use D hash tables sharing a hash function $\Rightarrow$ Evaluate once for $nnz(\boldsymbol{x}_i)$ updates.

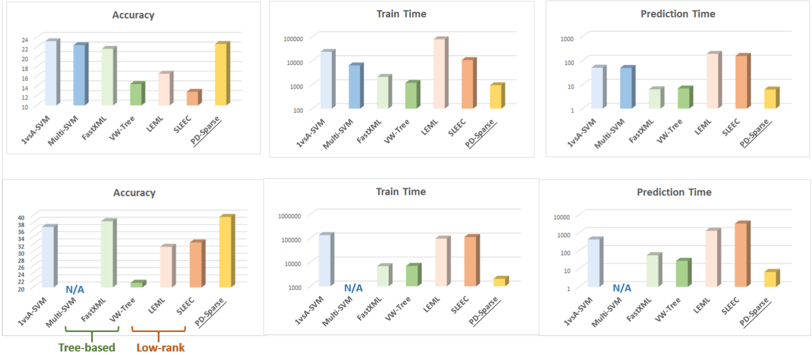


Outline

Experimental Result: Multiclass

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

Dmoz
N=3x10^5
D=8x10^5
K=1x10^4



- 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

LSHTC-wiki

N= $2\text{x}10^6$
D= $2\text{x}10^6$
K= $3\text{x}10^5$

- **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)

Conclusion

- 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.