A Two-Stage Approach for Learning a Sparse Model with Sharp Excess Risk Analysis

Zhe Li*, Tianbao Yang*,Lijun Zhang♮, Rong Jin†

*The University of Iowa, ♮Nanjing University, †Alibaba Group

December 10, 2015

Problem

- Let $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$ denote an input and output pair
- Let w_* be an optimal model that minimizes the expected error

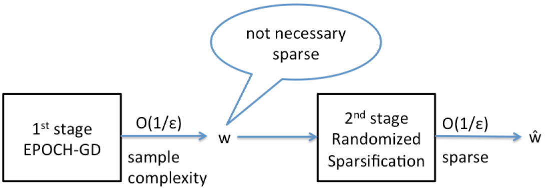
$$w_* = \arg\min_{||w||_1 \leq B} \frac{1}{2} \mathrm{E}_{\mathcal{P}}[(w^T x - y)^2]$$

- **Key Problem**: w_* is not necessarily sparse
- **The goal**: to learn a *sparse* model w to achieve small excess risk

$$ER(w, w_*) = \mathrm{E}_{\mathcal{P}}[(w^T x - y)^2] - \mathrm{E}_{\mathcal{P}}[(w_*^T x - y)^2] \leq \epsilon$$

The challenges

- $L = \mathrm{E}_{\mathcal{P}}[(w^T x - y)^2]$ is not necessarily strongly convex
  - Stochastic optimization: $O(1/\epsilon^2)$ sample complexity and no sparsity guarantee
  - Empirical risk minimization + ℓ_1 penalty: $O(1/\epsilon^2)$ sample complexity and no sparsity guarantee
- Challenges:
  - Can we reduce sample complexity (e.g. $O(1/\epsilon)$)?
  - Can we also have a guarantee on sparsity of model?
- Our solution:



The first stage

- Our first stage algorithm is motivated by EPOCH-GD algorithm [Hazan, Kale 2011], which is on strongly convex setting.
- How to avoid strongly convex assumption?
  - $L(w) = \mathrm{E}_{\mathcal{P}}[(w^T x - y)^2] = h(Aw) + b^T w + c$
  - $h(\cdot)$: a strongly convex function
  - The optimal solution set is a polyhedron
  - By Hoffmans' bound we have

$$2(L(w) - L_*) \geq \frac{1}{\kappa} ||w - w^+||_2^2$$

where w^+ is the closest solution to w in the optimal solution set.

[1] Elad Hazan, Satyen Kale, Beyond the regret minimization barrier: optimal algorithm for stochastic strongly-convex optimization

The second stage

- Our second stage algorithm:

Randomized Sparsification

For $k = 1, \ldots, K$

- Sample $i_k \in [d]$ according to $\Pr(i_k = j) = p_j$
- Compute $[\widetilde{\mathbf{w}}_k]_{i_k} = [\widetilde{\mathbf{w}}_{k-1}]_{i_k} + \frac{\widehat{w}_{i_k}}{p_{i_k}}$

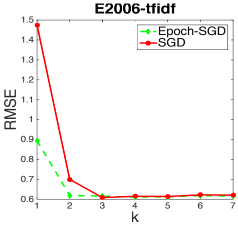
End For

$p_j = \frac{\sqrt{\hat{w}_j^2 E[x_j^2]}}{\sum_{j=1}^d \sqrt{\hat{w}_j^2 E[x_j^2]}}$ instead of $p_j = \frac{|\hat{w}_j|}{||\hat{w}||_1}$ [Shalve-Shwartz et al., 2010]

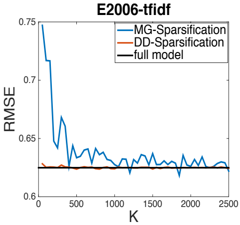
- Reduced constant in $O(1/\epsilon)$ for sparsity

[2] shalve-shwartz, Srebro, Zhang, Trading accuracy for sparsity in optimization problems with sparsity constraints

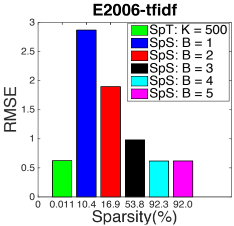
Experimental Results



1st stage



2nd stage



RMSE vs Sparsity