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

February 3, 2017

Outline

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

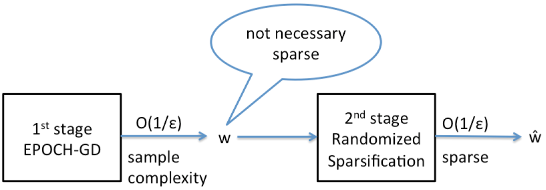
$$w_* = \arg\min_{||w||_1 \leq B} \frac{1}{2} \mathrm{E}_{\mathcal{P}}[(w^T x - y)^2]$$

- **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:



Outline

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 first stage (algorithm)

Stochastic Optimization for Sparse Learning

Input: the total number of iterations T and η_1, ρ_1, T_1.

Initialization: $\mathbf{w}_1^1 = 0$ and $k = 1$.

While $\sum_{i=1}^{m} T_i \leq T$

- For $t = 1, \ldots, T_k$
  - Obtain a sample denoted by $(\mathbf{x}_t^k, y_t^k)$
  - Compute $\mathbf{w}_{t+1}^k = \Pi_{\|\mathbf{w}\|_1 \leq B, \|\mathbf{w}-\mathbf{w}_1^k\|_2 \leq \rho_k}[\mathbf{w}_t^k - \eta_k \nabla \ell(\mathbf{w}_t^k \cdot \mathbf{x}_t^k, y_t^k)]$
- Update $T_{k+1} = 2T_k, \eta_{k+1} = \eta_k/2$, $\rho_{k+1} = \rho_k/\sqrt{2}$ and $\mathbf{w}_1^{k+1} = \sum_{t=1}^{T_k} \mathbf{w}_t^k / T_k$
- Set $k = k + 1$

Output: $\widehat{\mathbf{w}} = \mathbf{w}_1^{m+1}$

The first stage (theoretical guarantee)

Theorem

Assume $\|\mathbf{x}\|_2^2 \leq R^2$. *By running the previous algorithm with* $\rho_1 = B$, $\eta_1 = 1/(2R\sqrt{T_1})$, $T_1 \geq (8cR + 64R\sqrt{2\log(1/\widetilde{\delta})})^2$. *In order to have* $ER(\widehat{\mathbf{w}}, \mathbf{w}_*) \leq \epsilon$ *with a high probability* $1-\delta$ *over* $\{(\mathbf{x}_t^k, y_t^k)\}$, *it suffice to have*

$$T = \frac{cB^2 T_1}{\epsilon}$$

where $\widetilde{\delta} = \frac{\delta}{m}$, $m = \lfloor \log_2(cB^2/(2\epsilon)+1) \rfloor$ *and* $c = \max(\kappa, 1)$.

- No strong convexity assumption
- No sparsity assumption

The second stage (algorithm)

- 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

The second stage (theoretical guarantee)

Theorem

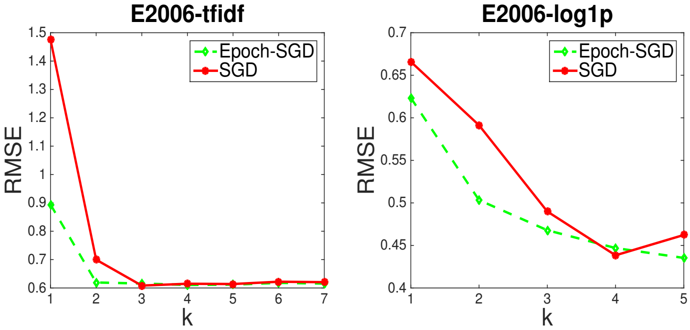
Given the samples in the first stage algorithm, let $p_j = \frac{\sqrt{\widehat{w}_j^2 \mathrm{E}[x_j^2]}}{\sum_{j=1}^d \sqrt{\widehat{w}_j^2 \mathrm{E}[x_j^2]}}, j \in [d]$ *in the second stage algorithm. In order to have* $ER(\widetilde{\mathbf{w}}, \mathbf{w}_*) \leq ER(\widehat{\mathbf{w}}, \mathbf{w}_*) + \epsilon$ *with a probability* $1 - \delta$ *over* $i_1, \ldots, i_K$*, it suffice to have*

$$K = \left\lceil \frac{\left(\sum_{i=1}^d \sqrt{\widehat{w}_j^2 \mathrm{E}[x_j^2]}\right)^2}{\epsilon\delta} \right\rceil$$

Outline

Experimental Results

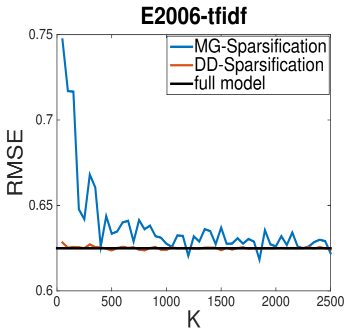
- The first stage



Comparison of RMSE between SGD and EPOCH-SGD

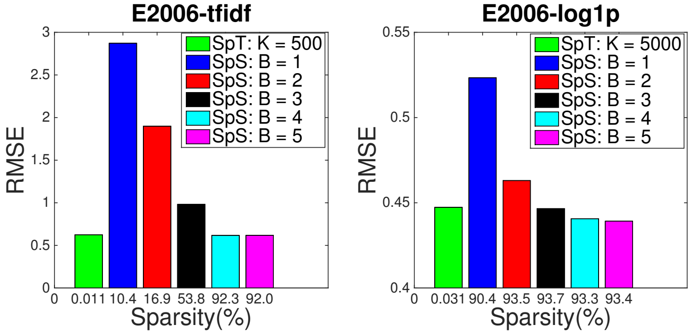
Experimental Results

- The second stage



Comparison of RMSE between MG-Sparsification and DD-Sparsification

- Overall



RMSE vs Sparsity

Outline

Conclusion

- We proposed a two-stage approach for learning a sparse model.
- We reduced the sample complexity from $O(1/\epsilon^2)$ to $O(1/\epsilon)$ without strongly convexity assumption.
- We reduced the constant in $O(1/\epsilon)$ for sparsity by exploring the distribution dependence sampling.
- We emprically justified the proposed approach could achieve better performance.