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

Main contribution

- Design a two-stage algorithm to learn a *sparse* linear model
- Reduce the order of the sample complexity from $O(1/\epsilon^2)$ to $O(1/\epsilon)$
- Reduce the constant in $O(1/\epsilon)$ for sparsity by exploring the distribution dependent sampling

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

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

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

$$ER(w, w_*) = E_{\mathcal{P}}[(w^T x - y)^2] - E_{\mathcal{P}}[(w_*^T x - y)^2] \le \epsilon$$

Three central questions

- Q: How to learn such a *sparse* mode with excess risk less than ϵ?
  A: Given by the following two-stage Algorithms.
- Q: What is the sample complexity in order to guarantee a small excess risk?
  A: Given by theorem 1.
- Q: What is the support complexity of w to suffice for ϵ excess risk?
  A: Given by theorem 2.

Related work

- Firstly, minimizes the objective function under ℓ_1 constraint and then use a randomized sparsification approach to find a sparse model [1]
- The first stage algorithm is based on EPOCH-Gradient Descent [2]

[1]: Shalev-Shartz, Shai, Srebro, Nathan, and Zhang, Trading accuracy for sparsity in optimization problems with sparsity constraints.
[2]: Hazan, Kale, Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization.

Two-stage approach to learning a sparse model

Algorithm 1 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 $\sum_{i=1}^k T_i \le T$ **do**
  for $t = 1, \ldots, T_k$ **do**
    Obtain a sample denoted by $(\mathbf{x}_t^k, y_t^k)$
    Compute $\mathbf{w}_{t+1}^k = \Pi_{\|\mathbf{w}\|_1 \le B, \|\mathbf{w} - \mathbf{w}_1^k\|_2 \le \rho_k}[\mathbf{w}_t^k - \eta_k \nabla \ell(\mathbf{w}_t^k \cdot \mathbf{x}_t^k, y_t^k)]$
  end for
  Update $T_{k+1} = 2T_k, \eta_{k+1} = \eta_k/2$, $\rho_{k+1} = \rho_k/\sqrt{2}$ and $\mathbf{w}_1^{k+1} = \sum_{t=1}^{T_k} \mathbf{w}_t^k / T_k$, $k = k + 1$
end while
Output: $\widehat{\mathbf{w}} = \mathbf{w}_1^{k+1}$

Algorithm 2 Distribution Dependent Randomized Sparsification

Input: $\widehat{\mathbf{w}} = (\widehat{w}_1, \ldots, \widehat{w}_d)$ and probabilities $p_1, \ldots, p_d$ such that $\sum_{j=1}^d p_j = 1$
Initialization: $\widetilde{\mathbf{w}}_0 = 0$.
for $k = 1, \ldots, K$ **do**
  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
Output: $\widetilde{\mathbf{w}} = \frac{\widetilde{\mathbf{w}}_K}{K}$

Stochastic Optimization ——First Stage

Theorem 1: Assume $\|x\|_2^2 \le R^2$, By running algorithm 1 with $\rho_1 = B, \eta_1 = 1/(2R\sqrt{T_1}), T_1 \ge (8cR + 64R\sqrt{2\log(1/\tilde{\delta})})^2$. In order to have $ER(w, w_*) \le \epsilon$ with a high probability $1 - \delta$ over $\{(x_t^k, y_t^k)\}$, it suffice to have

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

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

Remark 1 (No strongly convexity assumption) The sample complexity of Algorithm 1 is $O(1/\epsilon)$ for achieving an ϵ excess risk, which is improved from $O(1/\epsilon^2)$.
Remark 2 (No sparsity assumption): The sample complexity of Algorithm 1 has a sub-linear dependence on d due to $R \le \sqrt{d}$.

Distribution Dependent Randomized Sparsification ——Second Stage

Theorem 2: Given the samples in Algorithm 1, let $p_j = \frac{\sqrt{\hat{w}_j^2 E[x_j^2]}}{\sum_{j=1}^d \sqrt{\hat{w}_j^2 E[x_j^2]}}, j \in [d]$ in Algorithm 2. In order to have $ER(\tilde{w}, w_*) \le ER(\hat{w}, w_*) + \epsilon$ with a probability $1 - \delta$ over $i_1, i_2, \cdots i_K$, it suffices to have

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

Remark 1 (Reduced constant in $O(1/\epsilon)$ for sparsity): The value of K in theorem 2 is less than $\left\lceil \frac{\|\hat{w}\|_1^2}{\epsilon\delta} \right\rceil$, which was obtained from randomized sparsification, since $(\sum_{i=1}^d \sqrt{\hat{w}_j^2 E[x_j^2]})^2 \le \|\hat{w}\|_1^2$. The inequality holds only when the second moments of individual features are equal.