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

Zhe Li[1] and **Tianbao Yang**[1] and **Lijun Zhang**[2] and **Rong Jin**[3]
[1]Department of Computer Science, The University of Iowa, Iowa City, IA 52242, USA
[2]National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China
[3]Alibaba Group, Seattle, WA 98101, USA
{zhe-li-1,tianbao-yang}@uiowa.edu, zhanglj@lamda.nju.edu.cn, jinrong.jr@alibaba-inc.com

Abstract

This paper aims to provide a sharp excess risk guarantee for learning a *sparse* linear model *without any assumptions about the strong convexity of the expected loss and the sparsity of the optimal solution in hindsight.* Given a target level ϵ for the excess risk, an interesting question to ask is how many examples and how large the support set of the solution are enough for learning a good model with the target excess risk. To answer these questions, we present a two-stage algorithm that (i) in the first stage an epoch based stochastic optimization algorithm is exploited with an established $O(1/\epsilon)$ bound on the *sample complexity*; and (ii) in the second stage a distribution dependent randomized sparsification is presented with an $O(1/\epsilon)$ bound on the *sparsity* (referred to as *support complexity*) of the resulting model. Compared to previous works, our contributions lie at (i) we reduce the order of the sample complexity from $O(1/\epsilon^2)$ to $O(1/\epsilon)$ without the strong convexity assumption; and (ii) we reduce the constant in $O(1/\epsilon)$ for the sparsity by exploring the distribution dependent sampling.

Sparse learning is a fundamental problem in machine learning and statistics for high dimensional data analysis. Most previous studies, including literature on LASSO (Tibshirani, 1996; Zhao and Yu, 2006)and compressed sensing (Candès and Wakin, 2008; Donoho, 2006; Candés and Romberg, 2007), have been devoted to the recovery analysis that provide theoretical guarantees on the recovery error of the learned model with respect to the underlying true sparse model under certain assumptions or conditions (e.g., Gaussian noise model, RIP conditions, coherence, etc). In all of these studies, they assume the underlying model is a sparse vector and try to recover the sparse model from limited measurements. In this paper, we are interested in the excess risk of learning a sparse model without assuming the optimal solution is sparse. This problem has a variety of applications in practice. A sparse model is preferred when computational resources are limited and features are expensive to obtain (e.g., in medical diagnostic). In particular, if we let $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{y} \in \mathbb{R}$ denote an input and output pair that follow an unknown distribution $\mathcal{P}$, and let $\mathbf{w} \in \mathbb{R}^d$ denote a linear model, we define the following excess risk of $\mathbf{w}$:

$$\mathrm{ER}(\mathbf{w}, \mathbf{w}_*) = \mathrm{E}_{\mathcal{P}}[(\mathbf{w}^\top \mathbf{x} - y)^2] - \mathrm{E}_{\mathcal{P}}[(\mathbf{w}_*^\top \mathbf{x} - y)^2]$$

where $\mathbf{w}_*$ is an optimal model that minimizes the expected error in the domain $\mathcal{D} = \{\mathbf{w} : \|\mathbf{w}\|_1 \leq B\}$, i.e.,

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

The parameter B is usually determined by cross-validation. As a result, although there is an ℓ_1 norm constraint in the above problem, the optimal solution $\mathbf{w}_*$ is not necessarily sparse. Our goal is to learn a sparse model to achieve a small excess risk $\mathrm{ER}(\mathbf{w}, \mathbf{w}_*) \leq \epsilon$. The question then boils down to (i) How to learn such a sparse model? (ii) What is the sample complexity in order to guarantee a small excess risk? and (iii) What is the support complexity of $\mathbf{w}$ to suffice for an ϵ excess risk? In this paper, we answer these questions in the affirmative. It is notable that our work differentiates from previous work on sparse learning in that we do not assume the optimal model is a sparse model, which has several implications on the theoretical guarantee as we discuss later.

We develop our algorithms based on an approach presented in (Shalev-Shwartz, Srebro, and Zhang, 2010), which studied a similar problem in a pure optimization context. In particular, they established the optimization error of an optimization problem versus the sparsity of the resulting model. They presented and analyzed two approaches: (i) the first approach is a two-stage approach that in the first stage minimizes the objective function under ℓ_1 constraint and the second stage uses a randomized sparsification approach to find a sparse model; (ii) the second approach is a direct optimization approach based on the forward greedy selection. We notice that [1] the two-stage approach combined with empirical risk minimization (ERM) or stochastic optimization for minimizing $\mathrm{E}_{\mathcal{P}}[(\mathbf{w}^\top \mathbf{x} - y)^2]$ can potentially resolve our problem. By existing theory of excess risk for ERM or stochastic optimization (Hazan and Kale, 2011; Nemirovski et al., 2009; Sridharan, Shalev-Shwartz, and Srebro, 2008; Srebro, Sridharan, and Tewari, 2010), we can obtain an $O(1/\epsilon^2)$ sample complexity without strong convexity and an $O(1/\epsilon)$ sample complexity with strong convexity. Considering the objective function in (1):

$$L(\mathbf{w}) = \frac{1}{2}\mathbf{w}^\top \mathrm{E}_{\mathcal{P}}[\mathbf{x}\mathbf{x}^\top]\mathbf{w} - \mathbf{w}^\top \mathrm{E}_{\mathcal{P}}[y\mathbf{x}] + \frac{1}{2}\mathrm{E}_{\mathcal{P}}[y^2]$$

[1] The second approach is not appropriate here since it assumes the objective function can be evaluated exactly.

it could be non-strongly convex since $\mathrm{E}[\mathbf{x}\mathbf{x}^\top]$ could have a zero eigen-value unless under some special cases (e.g., features are independent and second moments of individual features are positive). Therefore, existing analysis without the strong convexity assumption only yields $O(1/\epsilon^2)$ sample complexity [2].

In this paper, we present an improved analysis of the excess risk for the two-stage approach without the strong convexity assumption. In particular, we show that (i) the sample complexity of a modified stochastic optimization algorithm can be made $O(1/\epsilon)$ by exploiting a property of the optimization problem similar to the error bound condition (Luo and Tseng, 1993); and (ii) the constant in the support complexity $O(1/\epsilon)$ of the resulting model from randomized sparsification can be reduced by exploiting a distribution dependent sampling. To the best of our knowledge, this is the first work that considers the complexities of the samples and the support of the solution for excess risk analysis. The empirical studies on real datasets demonstrate the two improvements and also exhibit that the two-stage approach is much better than the single stage approach that imposes a small ℓ_1 constraint on the solution in terms of both sparsity and performance.

Related Work

In this section, we review some other closely related work. Previous studies on sparse learning focus on deriving the oracle inequalities of the learned solution, i.e., $\|\mathbf{w}-\mathbf{w}_*\|$ when the optimal solution $\mathbf{w}_*$ is a (nearly) sparse vector. Most of them assume a design matrix X and a set of measurements of the target model $\mathbf{w}_*$ by $\mathbf{y} = X\mathbf{w}_* + \varepsilon$. The analysis is centered around the question of under what conditions one can exactly or accurately recover the target model $\mathbf{w}_*$. Different from these works, we study the prediction problem. We do not assume the optimal model that has the best generalization performance is a sparse model, but we aim to learn a sparse model with a good generalization performance. To this end, we use the excess risk to calibrate the learned model and impose a target excess risk on the learned model for analysis.

The most related work to the present paper in terms of learning a sparse model is presented in (Shalev-Shwartz, Srebro, and Zhang, 2010) as we discussed before. However, the authors either assume that the expected loss minimization problem has been solved exactly (in their first approach) or assume that the objective function and its gradient can be evaluated exactly (in their second approach), which renders them inadequate for solving the problem considered in this work. Recently, Agarwal, Negahban, and Wainwright (2014) considered the problem of learning a sparse model by minimizing an expected loss. The differences from our work are (i) their algorithm hinges on the assumption that the expected loss is strongly convex, and (ii) their analysis also focuses on the oracle inequalities $\|\mathbf{w}-\mathbf{w}_*\|$ not the excess risk. They obtained a fast convergence rate of $O(s\log(d)/T)$, where T is the number of samples and s is the sparsity of the optimal model $\mathbf{w}_*$. However, when the optimal model is not sparse, their algorithm only enjoys $O(\log(d)/\sqrt{T})$ convergence rate.

Our work on reducing the sample complexity of minimizing a non-strongly expected loss aligns with several pieces of recent work on improving the convergence of optimizing non-strongly convex functions. Bach and Moulines (2013) presented the first work for minimizing the expected square loss and logistic loss without the strong convexity assumption that achieves an $O(1/T)$ convergence rate. Wang and Lin (2014) studied the feasible descent approach for minimizing a family of non-strongly objective functions by exploiting the global error bound, an extension of the local error bound (Luo and Tseng, 1993). They showed that for certain problems (e.g., the dual problem of SVM), one can achieve a linear convergence rate even without the strong convexity assumption. Recently, there emerge several works (So, 2013; Hou et al., 2013; Zhou, Zhang, and So, 2015; Gong and Ye, 2014) that leverage the error bound conditions for achieving fast convergence of other regularized/constrained empirical loss minimization problems. The similarities and differences between these works and our work are summarized below:

- Similarities. Both (Bach and Moulines, 2013) and this work consider minimizing the expected loss and establishing the excess risk bound. We also explore a similar condition to the global error bound for a non-strongly convex function (Wang and Lin, 2014).
- Differences. Different from (Bach and Moulines, 2013) that only considers the non-constrained problems, we consider the ℓ_1 constrained problem. We put an emphasis on the expected loss minimization instead of regularized/constrained empirical loss minimization as considered in (Wang and Lin, 2014; So, 2013; Hou et al., 2013; Zhou, Zhang, and So, 2015; Gong and Ye, 2014). Therefore the algorithms proposed there are not applicable to our problem.

Learning a Sparse Model with Sharp Excess Risk Analysis

Without loss of generality, we assume $\mathbf{x} \in [-1, 1]^d$ and $|y| \leq B$ and denote $\mathrm{E}_P[\cdot]$ by $\mathrm{E}[\cdot]$ for short. We will first present and analyze a stochastic optimization algorithm that aims to solve

$$\min_{\mathbf{w}\in\mathcal{D}} \left[L(\mathbf{w}) = \frac{1}{2}\mathrm{E}[(\mathbf{w}^\top\mathbf{x} - y)^2] \right] \quad (2)$$

where $\mathcal{D} = \{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w}\|_1 \leq B\}$. In the sequel, we let $\Pi_\Omega[\widehat{\mathbf{w}}]$ denote the projection into a domain Ω, i.e.,

$$\Pi_\Omega[\widehat{\mathbf{w}}] = \arg\min_{\mathbf{w}\in\Omega} \|\mathbf{w} - \widehat{\mathbf{w}}\|_2^2.$$

Let $\ell(\mathbf{w}\cdot\mathbf{x}, y) = \frac{1}{2}(\mathbf{w}\cdot\mathbf{x} - y)^2$ denote the square loss function.

Since the objective function is not necessarily strongly convex, therefore the optimal solution might not be unique.

[2] Although adding a strongly convex regularizer can make the objective function strongly convex, it only ensures $O(1/n)$ convergence for the objective function not the expected loss (Sridharan, Shalev-Shwartz, and Srebro, 2008).

Algorithm 1 Stochastic Optimization for Sparse Learning

```
**Input**: the total number of iterations $T$ and $\eta_1, \rho_1, T_1$.
**Initialization:** $\mathbf{w}_1^1 = 0$ and $k = 1$.
**while** $\sum_{i=1}^m T_i \le T$ **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 $\mathbf{w}_1^{k+1} = \sum_{t=1}^{T_k} \mathbf{w}_t^k / T_k$
    Set $k = k + 1$
**end while**
**Output:** $\widehat{\mathbf{w}} = \mathbf{w}_1^{m+1}$
```

To this end, we let Ω_* denote the set of optimal solutions. For any $\mathbf{w} \in \mathbb{R}^d$, we denote by $\mathbf{w}^+$ the closest optimal solution to $\mathbf{w}$, i.e.,

$$\mathbf{w}^+ = \Pi_{\Omega_*}[\mathbf{w}] \tag{3}$$

We denote the optimal expected loss by L_*, i.e., $L_* = L(\mathbf{w}_*), \forall \mathbf{w}_* \in \Omega_*$. Then the excess risk of $\mathbf{w}$ is $2(L(\mathbf{w}) - L_*)$. The key to our analysis is the following lemma that lower bounds the excess risk of $\mathbf{w}$ by the scaled distance from $\mathbf{w}$ to $\mathbf{w}^+$, which is independent of the optimization algorithm.

Lemma 1. *For any* $\mathbf{w} \in \mathcal{D}$*, there exists a* $\kappa > 0$ *such that*

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

Remark: The proof is presented in the appendix. The value of κ depends on the optimization problem, in particular the distribution of the data, which is unfortunately unknown to us. The above inequality can be easily recovered for a strongly convex function $L(\mathbf{w})$ with $\mathbf{w}^+$ being the unique optimal solution and $1/\kappa$ being the strong convexity parameter of $L(\mathbf{w})$.

In the remainder of this section, we first present a stochastic optimization together with its sample complexity guarantee for solving the ℓ_1 constrained problem (2) in the immediate following subsection. After that, we present a distribution dependent randomized sparsificiation algorithm for obtaining a sparse model with excess risk guarantee. In the final subsection, we discuss the implementation issues.

Stochastic Optimization

We are now ready to present the stochastic optimization algorithm and its excess risk guarantee. The algorithm presented in Algorithm 1 is based on the epoch gradient descent (Hazan and Kale, 2011), which is originally proposed and analyzed for only strongly convex optimization. The values of η_1, ρ_1, T_1 are specified differently to handle the unknown value of κ. The following theorem establishes the excess risk guarantee of Algorithm 1.

Theorem 1. *Assume* $\|\mathbf{x}\|_2^2 \le R^2$*. By running Algorithm 1 with* $\rho_1 = B$*,* $\eta_1 = 1/(2R\sqrt{T_1})$*,* $T_1 \ge (8cR + 64R\sqrt{2\log(1/\widetilde{\delta})})^2$*. In order to have* $ER(\widehat{\mathbf{w}}, \mathbf{w}_*) \le \epsilon$ *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* $c = \max(\kappa, 1)$*.*

Algorithm 2 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 the distribution $\Pr(i_k = j) = p_j$
    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:** $\widetilde{\mathbf{w}} = \frac{\widetilde{\mathbf{w}}_K}{K}$
```

Remark 1 (No strong convexity assumption): The sample complexity of Algorithm 1 is $O(1/\epsilon)$ for achieving an ϵ excess risk. Compared to previous work without the strong convexity assumption, this order is improved upon $O(1/\epsilon^2)$.

Remark 2 (No sparsity assumption): Another issue is the dependence on the dimensionality. The sample complexity in Theorem 1 has a linear dependence on d due to $R \le \sqrt{d}$. Several previous work (Agarwal, Negahban, and Wainwright, 2014; Shalev-Shwartz and Tewari, 2011) can exploit the sparsity of the optimal solution $\mathbf{w}_*$ and obtain a logarithmic dependence on the dimensionality. For example, Agarwal, Negahban, and Wainwright (2014) exploited both the strong convexity of the expected loss and the sparsity of the optimal solution and achieved an $O(s\log(d)/\epsilon)$ sample complexity, where s is the sparsity of $\mathbf{w}_*$. However, when the optimal solution is not sparse they can only obtain $O(\log(d)/\epsilon^2)$ even with the strong convexity assumption. In contrast, our result is the first that establishes $O(d/\epsilon)$ sample complexity without strong convexity and sparsity assumptions.

Distribution Dependent Randomized Sparsification

Although Theorem 1 provides a guarantee on the excess risk of $\widehat{\mathbf{w}}$ found by Algorithm 1, it has no guarantee on the sparsity of $\widehat{\mathbf{w}}$. Previous studies have found that minimizing the ℓ_1 constrained problem does not necessarily lead to a sparse solution. A naive heuristic to make the solution sparse is to choose the coordinates according to the magnitude of elements in $\widehat{\mathbf{w}}$. Alternatively, one can choose the coordinates in a randomized way using the randomized sparsification procedure given in Algorithm 2. Shalev-Shwartz, Srebro, and Zhang (2010) used $p_j = \frac{|\widehat{\mathbf{w}}_j|}{\|\widehat{\mathbf{w}}\|_1}$ to define the sampling probabilities and established the following result for the number of steps K. Since $supp(\widetilde{\mathbf{w}}) \le K$, therefore the theorem below also provides an upper bound for the sparsity of $\widetilde{\mathbf{w}}$.

Theorem 2. *(Shalev-Shwartz, Srebro, and Zhang, 2010) Given the samples in Algorithm 1, let* $p_j = \frac{|\widehat{w}_j|}{\|\widehat{\mathbf{w}}\|_1}, j \in [d]$ *in Algorithm 2. In order to have* $ER(\widetilde{\mathbf{w}}, \mathbf{w}_*) \leq ER(\widehat{\mathbf{w}}, \mathbf{w}_*) + \epsilon$ *with a probability* $1 - \delta$*, it suffice to have*

$$K = \left\lceil \frac{\|\widehat{\mathbf{w}}\|_1^2}{\epsilon\delta} \right\rceil$$

Next, we describe a distribution dependent randomized sparsification algorithm that provides guarantee on the sparsity of the resulting model for achieving an ϵ excess risk, which is better than Theorem 2 by a distribution dependent constant factor. The intuition is that since we are ultimately interested in the prediction performance made by $\widetilde{\mathbf{w}}^\top \mathbf{x}$, thus the probabilities of selecting the coordinates should be dependent on the magnitude of $\widehat{w}_i x_i, i \in [d]$. This is formally stated in the following theorem.

Theorem 3. *Given the samples in Algorithm 1, 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 Algorithm 2. 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*

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

Remark: The value of K in Theorem 3 is always less than that in Theorem 2, because $\left(\sum_{i=1}^d \sqrt{\widehat{w}_j^2 \mathrm{E}[x_j^2]}\right)^2 \leq \|\widehat{\mathbf{w}}\|_1^2$. The equality holds only when the second moments of individual features are equal. For small values of ϵ and δ, the improvement could be significant. In practice, the second order moments may not be know aprior. We can calculate empirical estimations using the samples from the first stage.

Implementation

It is notable that Algorithm 1 requires a projection into the intersection of an ℓ_1 ball and an ℓ_2 ball. The problem is

$$\min_{\|\mathbf{w}\|_1 \leq B} \quad \frac{1}{2}\|\mathbf{w} - \widehat{\mathbf{w}}\|_2^2 \qquad (4)$$
$$s.t. \quad \|\mathbf{w} - \mathbf{c}\|_2^2 \leq r^2$$

First, it should be noticed that the above problem always has a feasible solution and the optimal solution is unique. This is because $\mathbf{c}$ is a feasible solution due to $\mathbf{c} = \mathbf{w}_1^k$ and $\|\mathbf{w}_1^k\|_1 \leq B$, and the uniqness is due to that the objective function is a strongly convex function.

By the Lagrangian theory, the above problem is inequivalent to

$$\max_{\eta \geq 0} \underbrace{\min_{\|\mathbf{w}\|_1 \leq B} \frac{1}{2}\|\mathbf{w} - \widehat{\mathbf{w}}\|_2^2 + \frac{\eta}{2}(\|\mathbf{w} - \mathbf{c}\|_2^2 - r^2)}_{g(\eta)}$$

To solve this, we present an efficient bisection search algorithm. Let $\alpha = \frac{1}{1+\eta} \in [0, 1]$ and $\mathbf{w}_\alpha = \alpha\widehat{\mathbf{w}} + (1 - \alpha)\mathbf{c}$ and $\mathbf{w}_\alpha^* = \Pi_{\|\mathbf{w}\|_1 \leq B}[\mathbf{w}_\alpha]$. The $g(\eta)$ function is a concave function of η. Given an η (or equivalently α), we can compute the gradient of $g(\eta)$. If $\nabla g(\eta) < 0$, we should decrease η (i.e., increase α); otherwise we should increase η (i.e., decrease α). To compute the gradient of $g(\eta)$, we need to find $\mathbf{w}_\alpha^*$ the optimal solution to the inner minimization problem w.r.t $\mathbf{w}$, i.e.,

Algorithm 3 Bisection search for solving the projection into the intersection of an ℓ_1 ball and an ℓ_2 ball

```
Input: $\widehat{\mathbf{w}}, \mathbf{c}, r, B$ and precision $\epsilon_s \geq 0$
Initialization: $\alpha_{\min} = 0, \alpha_{\max} = 1$
solve for $\mathbf{w}_1^*$.
if $\|\mathbf{w}_1^* - \mathbf{c}\|_2 \leq r$ then
    return $\mathbf{w}_1^*$
end if
while $\|\mathbf{w}_\alpha^* - \mathbf{c}\|_2 > r$ or $|\|\mathbf{w}_\alpha^* - \mathbf{c}\|_2 - r| > \epsilon_s$ do
    set $\alpha = (\alpha_{\min} + \alpha_{\max})/2$
    solve for $\mathbf{w}_\alpha^*$
    if $\|\mathbf{w}_\alpha^* - \mathbf{c}\|_2 < r$ then
        $\alpha_{\min} = \alpha$
    else
        $\alpha_{\max} = \alpha$
    end if
end while
return $\mathbf{w}_\alpha^*$
```

$$\mathbf{w}_\alpha^* = \arg\min_{\|\mathbf{w}\|_1 \leq B} \quad \frac{1}{2}\left\|\mathbf{w} - \frac{\widehat{\mathbf{w}} + \eta\mathbf{c}}{1 + \eta}\right\|_2^2 \qquad (5)$$

Then the gradient of $g(\eta)$ can be computed by $\nabla g(\eta) = \frac{1}{2}(\|\mathbf{w}_\alpha^* - \mathbf{c}\|_2^2 - r^2)$.

We can start from $\alpha = 1$, i.e., $\eta = 0$, if $\|\mathbf{w}_1^* - \mathbf{c}\|_2 \leq r$ then $\mathbf{w}_1^*$ is the optimal solution; otherwise we set $\alpha = 1/2$ and compute $\mathbf{w}_{1/2}^*$. If $\|\mathbf{w}_{1/2}^* - \mathbf{c}\|_2 < r$ we need to increase α, otherwise we decrease α. Since every iteration we cut the search space by half, in order to find an ϵ_s accurate solution (i.e., the distance to the optimal solution is less than ϵ_s), we only need $\left\lceil \log_2\left(\frac{\|\widehat{\mathbf{w}} - \mathbf{c}\|_2}{\epsilon_s}\right)\right\rceil$ iterations. To see this, we let $\mathbf{w}_{\alpha_k}$ and $\mathbf{w}_{\alpha_k}^*$ denote the generated sequences and let $\mathbf{w}_{\alpha_*}$ and $\widehat{\mathbf{w}}_{\alpha_*}$ denote the corresponding vectors to the optimal η_*. By the non-expansive property of projection (Bertsekas, 1999), we have

$$\|\mathbf{w}_{\alpha_k}^* - \mathbf{w}_{\alpha_*}^*\|_2 \leq \|\mathbf{w}_{\alpha_k} - \mathbf{w}_{\alpha_*}\|_2 \leq \frac{\|\widehat{\mathbf{w}} - \mathbf{c}\|_2}{2^k} \leq \epsilon_s.$$

Finally, for solving the projection into the ℓ_1 ball in (5), we can use the linear time algorithm proposed in (Duchi et al., 2008). Thus, the total time complexity for solving (4) is $O(d\log(\|\widehat{\mathbf{w}} - \mathbf{c}\|_2/\epsilon_s))$. We present the detailed steps of bisection search in Algorithm 3, where the exiting condition is determined by the magnitude of $|\|\mathbf{w}_\alpha^* - \mathbf{c}\|_2 - r|$. Note that

$$\begin{aligned} |\|\mathbf{w}_{\alpha_k}^* - \mathbf{c}\|_2 - r| &= |\|\mathbf{w}_{\alpha_k}^* - \mathbf{c}\|_2 - \|\mathbf{w}_{\alpha^*}^* - \mathbf{c}\|_2| \\ &\leq \|\mathbf{w}_{\alpha_k}^* - \mathbf{w}_{\alpha^*}^*\|_2 \end{aligned}$$

Therefore, we need at most $\left\lceil \log_2\left(\frac{\|\widehat{\mathbf{w}} - \mathbf{c}\|_2}{\epsilon_s}\right)\right\rceil$ iterations.

Main Analysis

Proof of Theorem 1

The proof builds upon the following lemma, which is proved in the appendix.

Lemma 2. *Let D be an upper bound on $\|\mathbf{w}_1 - \mathbf{w}_1^+\|_2$. Applying T iterations of the update $\mathbf{w}_{t+1} = \Pi_{\|\mathbf{w}\|_1 \leq B, \|\mathbf{w}-\mathbf{w}_1\|_2 \leq \rho}[\mathbf{w}_t - \eta\widehat{\mathbf{g}}_t]$, where $\widehat{\mathbf{g}}_t$ is an unbiased estimator for the subgradient of $L(\mathbf{w})$ at $\mathbf{w}_t$ satisfying $\|\widehat{\mathbf{g}}_t\|_2 \leq G$. Then for any $\delta \in (0,1)$, with a probability at least $1-\delta$ we have*

$$L\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{w}_t\right) - L_* \leq \frac{\eta G^2}{2} + \frac{\|\mathbf{w}_1 - \mathbf{w}_1^+\|_2^2}{2\eta T} + \frac{2G(\rho+D)\sqrt{2\log(1/\delta)}}{\sqrt{T}}$$

We apply the above lemma to each epoch of Algorithm 1. We prove the Theorem 1 by induction. Let $c = \max(\kappa, 1)$. We assume

$$\|\mathbf{w}_1^k - \mathbf{w}_1^{k,+}\|_2^2 \leq \frac{c^2B^2}{2^{k-1}}, \quad L(\mathbf{w}_1^k) - L_* \leq \frac{cB^2}{2^k}$$

It is true for $k=1$ since $\|\mathbf{w}_1^{1,+}\|_2 \leq \|\mathbf{w}_1^{1,+}\|_1 \leq B$ and $L(\mathbf{w}_1^1) \leq \frac{B^2}{2}$. Let $D_k = \frac{cB}{\sqrt{2^{k-1}}}$, $\rho_k = \frac{B}{\sqrt{2^{k-1}}}$ and $\eta_k = \frac{1}{2^k R\sqrt{T_1}}$. Then we prove

$$\|\mathbf{w}_1^{k+1} - \mathbf{w}_1^{k+1,+}\|_2^2 \leq \frac{c^2B^2}{2^k}, \quad L(\mathbf{w}_1^{k+1}) - L_* \leq \frac{cB^2}{2^{k+1}}$$

We note that $\widehat{g}_t = \nabla\ell(\mathbf{w}_t^k \cdot \mathbf{x}_t^k, y_t^k) = \mathbf{x}_t(\mathbf{w}_t^k \cdot \mathbf{x}_t - y_t)$. Thus $\|\widehat{g}_t\|_2 \leq 2RB$. Then we have

$$\begin{aligned}
L\left(\mathbf{w}_T^k\right) - L_* &\leq \frac{\eta_k 4B^2R^2}{2} + \frac{\|\mathbf{w}_1^k - \mathbf{w}_1^{k,+}\|_2^2}{2\eta_k T_k} \\
&+ \frac{8BR(\rho_k + D_k)\sqrt{2\log(1/\delta)}}{\sqrt{T_k}} \\
&\leq \frac{\eta_k 4B^2R^2}{2} + \frac{B^2c^2}{2^{k-1}2\eta_k T_k} + \frac{8BRB(c+1)\sqrt{2\log(1/\delta)}}{\sqrt{2^{k-1}}\sqrt{T_k}} \\
&\leq \frac{B^2R(c^2+1)}{2^{k-1}\sqrt{T_1}} + \frac{8B^2R(c+1)\sqrt{2\log(1/\delta)}}{2^{k-1}\sqrt{T_1}} \\
&\leq \frac{B^2c}{2^{k+1}}\left(\frac{8cR + 64R\sqrt{2\log(1/\delta)}}{\sqrt{T_1}}\right)
\end{aligned}$$

where we use the fact $c \geq 1$. If we choose T_1 such that $T_1 = (8cR + 64R\sqrt{2\log(1/\delta)})^2$, then

$$L(\mathbf{w}_T^k) - L_* \leq \frac{cB^2}{2^{k+1}}$$

and

$$\|\mathbf{w}_1^{k+1} - \mathbf{w}_1^{k+1,+}\|_2^2 \leq 2\kappa(L(\mathbf{w}_1^{k+1}) - L_*) \leq \frac{c^2B^2}{2^k}$$

Note that $T = T_1 + \cdots + T_m = T_1(2^m - 1)$. Therefore

$$L(\mathbf{w}_T^m) - L_* \leq \frac{cB^2T_1}{2T}$$

When $T = \dfrac{cB^2T_1}{\epsilon}$, we have

$$2(L(\widehat{\mathbf{w}}) - L_*) \leq \epsilon$$

Proof of Theorem 3

First, it is easy to verify that $\widetilde{\mathbf{w}} = \widehat{\mathbf{w}} \circ \xi$, where $\xi_i = \frac{m_i}{Kp_i}$ and $m_i, i = 1\ldots,d$ follow a multinomial distribution $Mult(p_1,\ldots,p_d;K)$. It is straightforward to see $\mathrm{E}_\xi[\widetilde{\mathbf{w}}] = \widehat{\mathbf{w}}$. By the quadratic form of $L(\mathbf{w})$, we have

$$\begin{aligned}
L(\widetilde{\mathbf{w}}) - L(\widehat{\mathbf{w}}) &\leq \nabla L(\widehat{\mathbf{w}})^\top(\widetilde{\mathbf{w}} - \widehat{\mathbf{w}}) \\
&+ \frac{1}{2}\mathrm{E}_{\mathbf{x}}[(\widetilde{\mathbf{w}} - \widehat{\mathbf{w}})^\top\mathbf{x}\mathbf{x}^\top(\widetilde{\mathbf{w}} - \widehat{\mathbf{w}})]
\end{aligned}$$

Taking expectation over the randomness over $\widetilde{\mathbf{w}}$, we have

$$\mathrm{E}_\xi[L(\widetilde{\mathbf{w}}) - L(\widehat{\mathbf{w}})] \leq \frac{1}{2}\mathrm{E}_{\mathbf{x},\xi}[(\widetilde{\mathbf{w}} - \widehat{\mathbf{w}})^\top\mathbf{x}\mathbf{x}^\top(\widetilde{\mathbf{w}} - \widehat{\mathbf{w}})]$$

where we use $\mathrm{E}_\xi[\widetilde{\mathbf{w}}] = \widehat{\mathbf{w}}$. Since $\mathrm{E}_\xi[\widetilde{\mathbf{w}}^\top] = \widehat{\mathbf{w}}^\top$, therefore

$$\begin{aligned}
&\mathrm{E}_\xi[(\widetilde{\mathbf{w}} - \widehat{\mathbf{w}})^\top\mathbf{x}\mathbf{x}^\top(\widetilde{\mathbf{w}} - \widehat{\mathbf{w}})] \\
&= \mathrm{E}_\xi[\mathbf{x}^\top(\widetilde{\mathbf{w}} - \widehat{\mathbf{w}})(\widetilde{\mathbf{w}} - \widehat{\mathbf{w}})^\top\mathbf{x}] \\
&= \mathbf{x}^\top cov(\widehat{\mathbf{w}} \circ \xi)\mathbf{x} \\
&= \frac{1}{K}\sum_{i=1}^{d}\frac{\widehat{w}_i^2x_i^2}{p_i} - \frac{1}{K}(\widehat{\mathbf{w}}^\top\mathbf{x})^2
\end{aligned}$$

Therefore

$$\mathrm{E}_\xi[L(\widetilde{\mathbf{w}}) - L(\widehat{\mathbf{w}})] \leq \frac{1}{2}\left[\frac{1}{K}\sum_{i=1}^{d}\frac{\widehat{w}_i^2\mathrm{E}[x_i^2]}{p_i} - \frac{1}{K}\mathrm{E}[(\widehat{\mathbf{w}}^\top\mathbf{x})^2]\right]$$

By minimizing the upper bound, we have

$$p_i = \frac{\sqrt{\widehat{w}_i^2\mathrm{E}[x_i^2]}}{\sum_{j=1}^{d}\sqrt{\mathrm{E}[\widehat{w}_j^2x_j^2]}}$$

and

$$\begin{aligned}
\mathrm{E}_\xi[L(\widetilde{\mathbf{w}}) - L(\widehat{\mathbf{w}})] &\leq \frac{1}{2K}\left[\left(\sum_{i=1}^{d}\sqrt{\widehat{w}_j^2\mathrm{E}[x_j^2]}\right)^2 - \mathrm{E}[(\widehat{\mathbf{w}}^\top\mathbf{x})^2]\right] \\
&\leq \frac{1}{2K}\left(\sum_{i=1}^{d}\sqrt{\widehat{w}_j^2\mathrm{E}[x_j^2]}\right)^2
\end{aligned}$$

By Markov inequality, we complete the proof.

Experiments

In this section, we present some experimental results on two real datasets: E2006-tfidf and E2006-log1p[3], which both have $16,087$ training examples and $3,308$ test examples. E2006-log1p includes $4,272,227$ features while E2006-tfidf contains $150,360$ features. The two datasets are for text regression with different types of features. We first conduct experiments to validate each of the two stages separately. We use the root mean square error (RMSE) on the test data as a measure of generalization performance. For

[3] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets

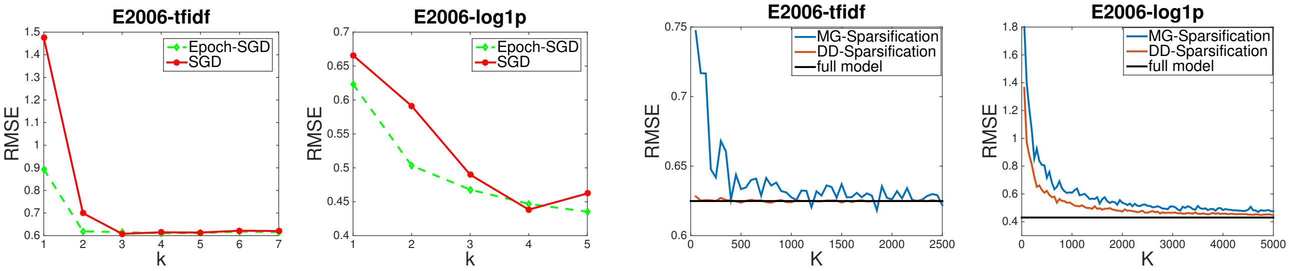

Figure 1: Comparison of RMSE between SGD and Epoch-SGD in the first stage (left two) and between MG-Sparsification and DD-Sparsification in the second stage (right two).

the first stage, we compare the presented stochastic optimization algorithm (referred to **Epoch-SGD**) to the standard stochastic gradient descent (SGD). The step size parameter in SGD and the initial step size in Epoch-SGD are tuned in $[0.1\ 0.05\ 0.01\ 0.005\ 0.001\ 0.0005\ 0.0001]$ by a cross-validation approach, and so as the number of iterations T_1 of the first epoch in Epoch-SGD. The value of B is set to the maximum absolute value of y in the training data. For the second stage, we compare two different sparsification strategies, namely the sampling based on the magnitude of entries in $\widehat{\mathbf{w}}$ learned in the first stage (referred to as **MG-Sparsification**) and the proposed distribution dependent sampling (referred to as **DD-Sparsification**).

The left two figures in Figure 1 compare RMSE of Epoch-SGD and SGD in the first stage. We run both algorithms for a number of epochs so that they pass through all examples in training dataset once. The x-axis represents the epoch number k as in Epoch-SGD. It can be observed that Epoch-SGD can reduce RMSE faster than SGD on both data sets. In Figure 1, we also plot RMSE (averaged over 100 random trials) versus K for MG-Sparsification and DD-Sparsification algorithms in the second stage. We also report the performance of the full model that is learned in the first stage by Epoch-SGD. The results demonstrate that (i) RMSE of the sparsified model decreases gradually as we include more coordinates, which is expected; (ii) DD-sparsification achieves better performance for the same value of K, which verifies Theorem 3; and (iii) the performance of DD-sparsification with $K = 500$ on E2006-tfidf and with $K = 5000$ on E2006-log1p (almost) match the performance of the full model on the two datasets, respectively.

Finally, we compare the proposed two-stage approach for learning a sparse model to a traditional single stage approach that tries to learn a sparse model by SGD with a small value of B for the ℓ_1 norm constraint. To achieve different sparsity of the resulting solution for the single stage approach, we vary the value of B and plot RMSE versus the sparsity of the solution (i.e., ratio of the number of non-zero entries to the total dimension). The results are plotted in Figure 2, where the green bar is result of the proposed two-stage approach and other bars are the results of the single stage approach with different values of B. It clearly shows that the two-stage approach can achieve much better performance with a much sparser solution than the single stage approach.

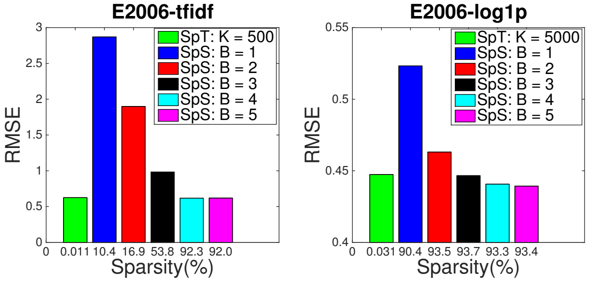

Figure 2: RMSE versus sparsity of models learned by the single stage approach for sparse learning (SpS) with different values of B and by the two-stage approach for sparse learning (SpT).

Conclusions

In this paper, we have considered the problem of learning a sparse model with excess risk guarantee. For the first time, we show that in order to achieve an ϵ excess risk, the sample complexity can be $O(1/\epsilon)$ without any assumptions on the strong convexity and the sparsity of the optimal solution. We also exploit a distribution dependent sampling to generate a sparse model with $O(1/\epsilon)$ sparsity for achieving an ϵ excess risk. Compared with the previous approach, we improve the bound by a distribution dependent constant order. Empirical studies on real datasets verify the improvements and effectiveness of the proposed algorithm in learning a sparse model with a good generalization performance in comparison with traditional approaches by imposing a small ℓ_1 constraint. An open problem that remains is whether we can obtain a logarithmic dependence on the dimensionality for the sample complexity by exploiting the sparsity of the optimal solution but still maintain $O(1/\epsilon)$ order without the strong convexity assumption.

Acknowledgments

We thank anonymous reviewers for their comments. Z. Li and T. Yang are partially supported by National Science Foundation (IIS-1463988, IIS-1545995). L. Zhang is supported in part by NSFC (61603177) and JiangsuFS(BK20160658).

References

Agarwal, A.; Negahban, S.; and Wainwright, M. J. 2014. Stochastic optimization and sparse statistical recovery: An optimal algorithm for high dimensions. In *48th Annual Conference on Information Sciences and Systems (CISS)*, 1–2.

Bach, F. R., and Moulines, E. 2013. Non-strongly-convex smooth stochastic approximation with convergence rate o(1/n). In *Advances in Neural Information Processing Systems (NIPS)*, 773–781.

Bertsekas, D. P. 1999. *Nonlinear Programming*. Belmont, MA: Athena Scientific.

Candés, E., and Romberg, J. 2007. Sparsity and incoherence in compressive sampling. *Inverse Problems* 23(3):969–985.

Candès, E. J., and Wakin, M. B. 2008. An introduction to compressive sampling. *IEEE Signal Processing Magazine* 25(2):21 –30.

Donoho, D. L. 2006. Compressed sensing. *IEEE Transaction on Information Theory* 52:1289–1306.

Duchi, J.; Shalev-Shwartz, S.; Singer, Y.; and Chandra, T. 2008. Efficient projections onto the L1-ball for learning in high dimensions. In *International Conference on Machine Learning*.

Gong, P., and Ye, J. 2014. Linear convergence of variance-reduced projected stochastic gradient without strong convexity. *CoRR* abs/1406.1102.

Hazan, E., and Kale, S. 2011. Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization. *Journal of Machine Learning Research - Proceedings Track* 19:421–436.

Hou, K.; Zhou, Z.; So, A. M.; and Luo, Z. 2013. On the linear convergence of the proximal gradient method for trace norm regularization. In *Advances in Neural Information Processing Systems (NIPS)*, 710–718.

Luo, Z., and Tseng, P. 1993. Error bound and reduced-gradient projection algorithms for convex minimization over a polyhedral set. *SIAM Journal on Optimization* 3(1):43–59.

Necoara, I.; Nesterov, Y.; and Glineur, F. 2015. Linear convergence of first order methods for non-strongly convex optimization. *arXiv preprint arXiv:1504.06298*.

Nemirovski, A.; Juditsky, A.; Lan, G.; and Shapiro, A. 2009. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization* 19:1574–1609.

Shalev-Shwartz, S., and Tewari, A. 2011. Stochastic methods for l_1-regularized loss minimization. *Journal of Machine Learning Research* 12:1865–1892.

Shalev-Shwartz, S.; Srebro, N.; and Zhang, T. 2010. Trading accuracy for sparsity in optimization problems with sparsity constraints. *SIAM Journal on Optimization* 20(6):2807–2832.

So, A. M. 2013. Non-asymptotic convergence analysis of inexact gradient methods for machine learning without strong convexity. *CoRR* abs/1309.0113.

Srebro, N.; Sridharan, K.; and Tewari, A. 2010. Smoothness, low noise and fast rates. In *Advances in Neural Information Processing Systems (NIPS)*, 2199–2207.

Sridharan, K.; Shalev-Shwartz, S.; and Srebro, N. 2008. Fast rates for regularized objectives. In *Advances in Neural Information Processing Systems (NIPS)*, 1545–1552.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)* 58:267–288.

Wang, P., and Lin, C. 2014. Iteration complexity of feasible descent methods for convex optimization. *Journal of Machine Learning Research* 15(1):1523–1548.

Zhao, P., and Yu, B. 2006. On model election consistency of lasso. *Journal of Machine Learning Research* 7:2541C2563.

Zhou, Z.; Zhang, Q.; and So, A. M. 2015. \(\ell_{1, p}\)-norm regularization: Error bounds and convergence rate analysis of first-order methods. In *Proceedings of the 32nd International Conference on Machine Learning, (ICML)*, 1501–1510.