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

Zhe Li
The University of Iowa
Iowa city, IA 52245
zhe-li-1@uiowa.edu

Tianbao Yang
The University of Iowa
Iowa city, IA 52245
tianbao-yang@uiowa.edu

Lijun Zhang
Nanjing Unversity
Nanjing 210023, China
zhanglj@lamda.nju.edu.cn

Rong Jin
Michigan State University, East Lansing, MI 48824
Alibaba Group, Hangzhou 311121, China
rongjin@msu.edu

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.

1 Introduction

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.

We develop our algorithms based on an approach presented in [11], which studied a similar problem in a pure optimization context. We notice thatthe 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 [7, 9, 13, 12], 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): $L(\mathbf{w}) = \frac{1}{2}\mathbf{w}^\top\mathrm{E}_{\mathcal{P}}[\mathbf{x}\mathbf{x}^\top]\mathbf{w} - \mathbf{w}^\top\mathrm{E}_{\mathcal{P}}[y\mathbf{x}] + \frac{1}{2}\mathrm{E}_{\mathcal{P}}[y^2]$, 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 [1].

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 [8]; 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.

2 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

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

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. 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., $\mathbf{w}^+ = \Pi_{\Omega_*}[\mathbf{w}]$ 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 $2(L(\mathbf{w}) - L_*) \geq \frac{1}{\kappa}\|\mathbf{w} - \mathbf{w}^+\|_2^2$*

Remark: 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 $L(\mathbf{w})$.

Connection to the global error bound condition We highlight the connection of inequality in Lemma 1 to the global error bound condition [16], which is stated below. An optimization problem $\min_{\mathbf{w}\in\Omega} f(\mathbf{w})$ admits a global error bound if there is a constant γ such that $\|\mathbf{w} - \mathbf{w}^+\|_2 \leq \gamma\|\nabla^+ f(\mathbf{w})\|_2, \forall \mathbf{w} \in \Omega$ where $\mathbf{w}^+$ is the projection of $\mathbf{w}$ to the optimal solution set and $\nabla^+ f(\mathbf{w})$ is the projected gradient defined as $\nabla^+ f(\mathbf{w}) = \mathbf{w} - \Pi_{\Omega_*}[\mathbf{w} - \nabla f(\mathbf{w})]$. Next, we show that if $f(\mathbf{w})$ is a 1-smooth function and satisfies the global error bound, it also satisfies the Lemma 1. In particular, for any optimal solution $\mathbf{u} \in \Omega_*$, we have $f(\mathbf{u}) \leq f(\mathbf{w}) + \nabla f(\mathbf{w})^\top(\mathbf{u} - \mathbf{w}) + \frac{1}{2}\|\mathbf{u} - \mathbf{w}\|_2^2$. If we minimize the R.H.S over $\mathbf{u}$, we obtain the optimal solution for $\mathbf{u}$ given by $\mathbf{w}_* = \Pi_{\Omega_*}[\mathbf{w} - \nabla f(\mathbf{w})]$. Let $\mathbf{z} = (\mathbf{w}_* - \mathbf{w})/\|\mathbf{w}_* - \mathbf{w}\|_2$ and $\eta_* = \|\mathbf{w}_* - \mathbf{w}\|_2$. Assume that $\eta_* > 0$, then $f_* \leq f(\mathbf{w}) + \nabla f(\mathbf{w})^\top\mathbf{z}\eta_* + \frac{1}{2}\eta_*^2 = f(\mathbf{w}) + \min_{0\leq\eta\leq\eta_+} \nabla f(\mathbf{w})^\top\mathbf{z}\eta + \frac{1}{2}\eta^2$ Therefore $\eta_* = \Pi_{[0,\eta_+]}[-\nabla f(\mathbf{w})^\top\mathbf{z}]$ and $-\nabla f(\mathbf{w})^\top\mathbf{z} \geq \eta_*$. Thus $f_* \leq f(\mathbf{w}) - \eta_*^2 + \frac{1}{2}\eta_*^2 = f(\mathbf{w}) - \frac{1}{2}\eta_*^2 = f(\mathbf{w}) - \frac{1}{2}\|\Pi_{\Omega_*}[\mathbf{w} - \nabla f(\mathbf{w})] - \mathbf{w}\|_2^2 = f(\mathbf{w}) - \frac{1}{2}\|\nabla^+ f(\mathbf{w})\|_2^2 \leq f(\mathbf{w}) - \frac{1}{2\gamma^2}\|\mathbf{w} - \mathbf{w}_+\|_2^2$ where the last inequality is due to the global error bound condition. Hence, we can see that when $f(\mathbf{w})$ is a 1-smooth function and admits a global error bound with a parameter γ, it satisfies the Lemma 1 with $\kappa = \gamma^2$. Importantly, it implies that Lemma 1 is a relaxed condition than the global error bound condition.

[1] 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 [13].

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 $\sum_{i=1}^{m} T_i \leq T$ **do**
  for $t = 1, \ldots, T_k$ **do**
    Obtain a sample denoted by $(\mathbf{x}_t^k, y_t^k)$
    Compute $\mathbf{w}_{t+1}^k = \Pi_{\|\mathbf{w}\|_1 \leq B, \|\mathbf{w}-\mathbf{w}_1^k\|_2 \leq \rho_k}[\mathbf{w}_t^k - \eta_k \nabla \ell(\mathbf{w}_t^k \cdot \mathbf{x}_t^k, y_t^k)]$
  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: $\widehat{\mathbf{w}} = \mathbf{w}_1^{m+1}$

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 $\Pr(i_k = j) = p_j$ and 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: $\widetilde{\mathbf{w}} = \frac{\widetilde{\mathbf{w}}_K}{K}$

2.1 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 [7], 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* $\|\mathbf{x}\|_2^2 \leq R^2$. *By running Algorithm 1 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* $T = \dfrac{cB^2 T_1}{\epsilon}$, *where* $\widetilde{\delta} = \frac{\delta}{m}$, $m = \lfloor \log_2(cB^2/(2\epsilon) + 1) \rfloor$ *and* $c = \max(\kappa, 1)$.

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 $R \leq \sqrt{d}$. Several previous work [1, 10] can exploit the sparsity of the optimal solution $\mathbf{w}_*$ and obtain a logarithmic dependence on the dimensionality. For example, [1] 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.

2.2 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. [11] used $p_j = \frac{|\widehat{\mathbf{w}}_j|}{\|\mathbf{w}\|_1}$ to define the sampling probabilities and established the following result for the number of steps K. Since $supp(\widetilde{\mathbf{w}}) \leq K$, therefore the theorem below also provides an upper bound for the sparsity of $\widetilde{\mathbf{w}}$.

Theorem 2. *[11] 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.

2.3 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

$$\begin{aligned} \min_{\|\mathbf{w}\|_1 \leq B} \quad & \frac{1}{2}\|\mathbf{w} - \widehat{\mathbf{w}}\|_2^2, \\ s.t. \quad & \|\mathbf{w} - \mathbf{c}\|_2^2 \leq r^2 \end{aligned} \tag{3}$$

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 uniqueness is due to that the objective function is a strongly convex function. By the Lagrangian theory, the above problem is inequivalent to $\max_{\eta \geq 0} g(\eta)$, where $g(\eta) = \min_{\|\mathbf{w}\|_1 \leq B} \frac{1}{2}\|\mathbf{w} - \widehat{\mathbf{w}}\|_2^2 + \frac{\eta}{2}(\|\mathbf{w} - \mathbf{c}\|_2^2 - r^2)$. 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.,

$$\mathbf{w}_\alpha^* = \arg \min_{\|\mathbf{w}\|_1 \leq B} \quad \frac{1}{2}\left\|\mathbf{w} - \frac{\widehat{\mathbf{w}} + \eta\mathbf{c}}{1+\eta}\right\|_2^2 \tag{4}$$

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 [2], 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 (4), we can use the linear time algorithm proposed in [6]. Thus, the total time complexity for solving (3) is $O(d \log(\|\widehat{\mathbf{w}} - \mathbf{c}\|_2/\epsilon_s))$.

References

[1] Alekh Agarwal, Sahand Negahban, and Martin J. Wainwright. Stochastic optimization and sparse statistical recovery: An optimal algorithm for high dimensions. In *48th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–2, 2014.

[2] D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, MA, 1999.

[3] Emmanuel Candés and Justin Romberg. Sparsity and incoherence in compressive sampling. *Inverse Problems*, 23(3):969–985, 2007.

[4] Emmanuel J. Candès and Michael B. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25(2):21 –30, 2008.

[5] David L. Donoho. Compressed sensing. *IEEE Transaction on Information Theory*, 52:1289–1306, 2006.

[6] John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the L1-ball for learning in high dimensions. In *International Conference on Machine Learning*, 2008.

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

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

[9] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19: 1574–1609, 2009.

[10] Shai Shalev-Shwartz and Ambuj Tewari. Stochastic methods for l_1-regularized loss minimization. *Journal of Machine Learning Research*, 12:1865–1892, 2011.

[11] Shai Shalev-Shwartz, Nathan Srebro, and Tong Zhang. Trading accuracy for sparsity in optimization problems with sparsity constraints. *SIAM Journal on Optimization*, 20(6):2807–2832, 2010.

[12] Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2199–2207, 2010.

[13] Karthik Sridharan, Shai Shalev-Shwartz, and Nathan Srebro. Fast rates for regularized objectives. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1545–1552, 2008.

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

[15] Martin J. Wainwright. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Transactions on Information Theory*, 55(12):5728–5741, 2009.

[16] Po-Wei Wang and Chih-Jen Lin. Iteration complexity of feasible descent methods for convex optimization. *Journal of Machine Learning Research*, 15(1):1523–1548, 2014.

[17] Peng Zhao and Bin Yu. On model election consistency of lasso. *Journal of Machine Learning Research*, 7:2541C2563, 2006.