Random Projections for Classification: A Recovery Approach

Lijun Zhang, *Member, IEEE*, Mehrdad Mahdavi, Rong Jin, Tianbao Yang, and Shenghuo Zhu

***Abstract*— Random projection has been widely used in data classification. It maps high-dimensional data into a low-dimensional subspace in order to reduce the computational cost in solving the related optimization problem. While previous studies are focused on analyzing the classification performance in the low-dimensional space, in this paper, we consider the recovery problem, i.e., how to accurately recover the optimal solution to the original high-dimensional optimization problem based on the solution learned after random projection. We present a simple algorithm, termed dual random projection, which uses the dual solution of the low-dimensional optimization problem to recover the optimal solution to the original problem. Our theoretical analysis shows that with a high probability, the proposed algorithm is able to accurately recover the optimal solution to the original problem, provided that the data matrix is (approximately) low-rank and/or optimal solution is (approximately) sparse. We further show that the proposed algorithm can be applied iteratively to reducing the recovery error exponentially.**

***Index Terms*— Random projection, primal solution, dual solution, low-rank, sparse.**

I. Introduction

RANDOM projection is a simple yet powerful dimensionality reduction technique that projects the original high-dimensional data onto a low-dimensional subspace using a random matrix [2], [3]. It has been successfully applied to many machine learning tasks, including classification [4]–[7], regression [8], clustering [9], [10], manifold learning [11], [12], and information retrieval [13].

In this work, we focus on random projection for classification. While previous studies were devoted to analyzing the classification performance after random projection [14]–[17], we examine the effect of random projection from a very different aspect. In particular, we are interested in accurately recovering the optimal solution to the original high-dimensional optimization problem using random projection. This is particularly useful for feature selection [18], where important features are often selected based on their weights in the linear prediction model learned from the training data. In order to ensure that similar features are selected, the prediction model based on random projection needs to be close to the model obtained by solving the original optimization problem directly. We emphasize that this paper is focused on exploring random projection for the recovery of the optimal solution that minimizes the empirical risk, and the analysis of its generalization error will be left as future work.

The proposed *Dual Random Projection* for recovering the optimal solution consists of two steps. In the first step, similar to previous studies, we apply random projection to reducing the dimensionality of data, and then solve a low-dimensional optimization problem in the projected space. The key innovation of the proposed algorithm comes from the second step, in which we compute the dual solution of the low-dimensional optimization problem from its primal solution, and use it to recover the optimal solution to the original high-dimensional optimization problem. Our analysis reveals that under the assumption that the data matrix is (approximately) low-rank and/or the optimal solution is (approximately) sparse, with a high probability, we are able to recover the optimal solution with a small error.

One nice property of our algorithm is that it is equipped with a relative, instead of an additive, bound for the recovery error. As a result, the recovery error can be reduced exponentially when applying the proposed algorithm iteratively. In other words, to recover the optimal solution with a relative error $\alpha \leq 1$, the number of iterations required is $O(\log 1/\alpha)$.

The rest of the paper is arranged as follows. In Section II, we describe the problem of recovering optimal solution by random projection, the theme of this work. Section III introduces the dual random projection approach for recovering the optimal solution. We show the main theoretical results for the proposed algorithm in Section IV. Section V presents the proofs of the theorems stated in Section IV. An iterative extension of dual random projection is discussed in Section VI. In Section VII, we analyze the numerical complexities of our algorithms and report the experimental results. Section VIII concludes with future directions of this work.

II. The Problem of Recovering Optimal Solution by Random Projection

Let $(\mathbf{x}_i, y_i),\ i = 1, \ldots, n$ be a set of training examples, where $\mathbf{x}_i \in \mathbb{R}^d$ is a vector of d dimensions and $y_i \in \{-1, +1\}$ is the binary class assignment for $\mathbf{x}_i$. Let $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$

Manuscript received July 13, 2013; revised July 31, 2014; accepted September 7, 2014. Date of publication September 19, 2014; date of current version October 16, 2014. This work was supported in part by the National Science Foundation of China under Grant 61321491, in part by the Office of Naval Research under Grant N000141210431 and Grant N000141410631, and in part by the National Science Foundation under Grant IIS-1251031. This paper was presented at the 26th Annual Conference on Learning Theory in 2013 [1].

L. Zhang is with the National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China (e-mail: zhanglj@lamda.nju.edu.cn).

M. Mahdavi is with the Toyota Technological Institute at Chicago, Chicago, IL 60637 USA (e-mail: mahdavi@uchicago.edu).

R. Jin is with the Department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824 USA (e-mail: rongjin@cse.msu.edu).

T. Yang is with the Department of Computer Science, University of Iowa, Iowa City, IA 52242 USA (e-mail: tianbao-yang@uiowa.edu).

S. Zhu is with the Alibaba Group, Seattle, WA 98101 USA (e-mail: shenghuo.zhu@alibaba-inc.com).

Communicated by V. Saligrama, Associate Editor for Signal Processing.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2014.2359204

and $\mathbf{y} = [y_1, \ldots, y_n]^\top$ include the input patterns and class assignments of all training examples. Typically, a linear classifier $\mathbf{w} \in \mathbb{R}^d$ is learned by solving the following regularized optimization problem:

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{\lambda}{2}\|\mathbf{w}\|^2 + \frac{1}{n}\sum_{i=1}^{n} \ell(y_i \mathbf{x}_i^\top \mathbf{w}) \tag{1}$$

where $\|\cdot\|$ stands for the ℓ_2 norm of vectors, and $\ell(z)$ is a differentiable convex loss function. In this study, we assume $\ell(\cdot)$ is a γ-smooth loss function, i.e.,

$$|\ell'(z) - \ell'(z')| \le \gamma\,|z - z'|.$$

By writing $\ell(\cdot)$ in its convex conjugate form, i.e.,

$$\ell(z) = \max_{\alpha\in\Omega} \alpha z - \ell_*(\alpha),$$

where $\ell_*(\cdot)$ is the convex conjugate of $\ell(\cdot)$ and Ω is the domain of the dual variable, we have the dual optimization problem:

$$\max_{\boldsymbol{\alpha}\in\Omega^n} -\sum_{i=1}^{n} \ell_*(\alpha_i) - \frac{1}{2\lambda n}(\boldsymbol{\alpha}\circ\mathbf{y})^\top X^\top X(\boldsymbol{\alpha}\circ\mathbf{y}) \tag{2}$$

where $\boldsymbol{\alpha}\circ\mathbf{y}$ stands for the element-wise product between two vectors (i.e., the Hadamard product) and $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_n]^\top$. In the rest of the paper, we will denote by $\mathbf{w}_* \in \mathbb{R}^d$ the optimal primal solution to (1), and by $\boldsymbol{\alpha}_* \in \mathbb{R}^n$ the optimal dual solution to (2). The following proposition connects $\mathbf{w}_*$ and $\boldsymbol{\alpha}_*$.

Proposition 1: We have

$$\mathbf{w}_* = -\frac{1}{\lambda n} X(\boldsymbol{\alpha}_*\circ\mathbf{y}),$$
$$[\boldsymbol{\alpha}_*]_i = \ell'\big(y_i \mathbf{x}_i^\top \mathbf{w}_*\big), \quad i = 1, \ldots, n.$$

The proof of Proposition 1 is provided in the Appendix A.

When the dimensionality d is high and the number of training examples n is large, solving either the primal problem in (1) or the dual problem in (2) can be computationally expensive. To reduce the computational cost, one common approach is to significantly reduce the dimensionality by random projection [4]. Let $A \in \mathbb{R}^{d\times m}$ be a Gaussian random matrix, where each entry $A_{i,j}$ is independently drawn from a Gaussian distribution $\mathcal{N}(0, 1/m)$ and m is significantly smaller than d. Using the random matrix A, we generate a new data representation for input data points by

$$\widehat{\mathbf{x}}_i = A^\top \mathbf{x}_i, \quad i = 1, \ldots, n$$

and solve the following low-dimensional optimization problem:

$$\min_{\mathbf{z}\in\mathbb{R}^m} \frac{\lambda}{2}\|\mathbf{z}\|^2 + \frac{1}{n}\sum_{i=1}^{n} \ell(y_i \mathbf{z}^\top \widehat{\mathbf{x}}_i). \tag{3}$$

The corresponding dual problem is

$$\max_{\boldsymbol{\alpha}\in\Omega^n} -\sum_{i=1}^{n} \ell_*(\alpha_i) - \frac{1}{2\lambda n}(\boldsymbol{\alpha}\circ\mathbf{y})^\top X^\top A A^\top X(\boldsymbol{\alpha}\circ\mathbf{y}). \tag{4}$$

Intuitively, the choice of the Gaussian matrix A is justified by the fact that $\mathrm{E}[\widehat{\mathbf{x}}_i^\top \widehat{\mathbf{x}}_j] = \mathbf{x}_i^\top \mathrm{E}\big[AA^\top\big]\mathbf{x}_j = \mathbf{x}_i^\top \mathbf{x}_j$, i.e., the expectation of the dot-product between any two examples in the projected space is equal to the dot-product in the original space. Let $\mathbf{z}_* \in \mathbb{R}^m$ denote the optimal primal solution to the low-dimensional problem (3), and $\widehat{\boldsymbol{\alpha}}_* \in \mathbb{R}^n$ denote the optimal dual solution to (4). Similar to Proposition 1, we have the following relationship between $\mathbf{z}_*$ and $\widehat{\boldsymbol{\alpha}}_*$:

$$\mathbf{z}_* = -\frac{1}{\lambda n} A^\top X(\widehat{\boldsymbol{\alpha}}_*\circ\mathbf{y}),$$
$$[\widehat{\boldsymbol{\alpha}}_*]_i = \ell'\big(y_i \widehat{\mathbf{x}}_i^\top \mathbf{z}_*\big), \quad i = 1, \ldots, n. \tag{5}$$

Given the optimal solution $\mathbf{z}_* \in \mathbb{R}^m$, the data point $\mathbf{x} \in \mathbb{R}^d$ is classified by $\mathbf{x}^\top A\mathbf{z}_*$, which is equivalent to defining a new solution $\widehat{\mathbf{w}} \in \mathbb{R}^d$ as

$$\widehat{\mathbf{w}} = A\mathbf{z}_*, \tag{6}$$

which we refer to as the *naive solution*. The classification performance of $\widehat{\mathbf{w}}$ has been examined by many studies [14]–[17]. The general conclusion is that when most of the original data are linearly separable with a large margin, the classification error for $\widehat{\mathbf{w}}$ will be small.

Although these studies show that $\widehat{\mathbf{w}}$ can achieve a small classification error under appropriate assumptions, it is unclear whether $\widehat{\mathbf{w}}$ is a good approximation to the optimal solution $\mathbf{w}_*$. To answer this question, we need the [18, Proposition 4.7].

Proposition 2 (Distance of a Random Subspace to a Fixed Point [19]): Let $E \in G_{d,m}$ be a random subspace (codim $E = d - m$). Let $\mathbf{x}$ be an unit vector, which is arbitrary but fixed. Then

$$\Pr\left(\mathrm{dist}(\mathbf{x}, E) \le \epsilon\sqrt{\frac{d-m}{d}}\right) \le (c\epsilon)^{d-m} \text{ for any } \epsilon > 0,$$

where c is an universal constant.

Because $\widehat{\mathbf{w}}$ lies in a random subspace spanned by the column vectors in A, according to Proposition 2, we have, with a probability at least $1 - 2^{-d+m}$,

$$\|\widehat{\mathbf{w}} - \mathbf{w}_*\| \ge \frac{1}{2c}\sqrt{\frac{d-m}{d}}\|\mathbf{w}_*\|,$$

implying that $\widehat{\mathbf{w}}$ is a BAD approximation to the optimal solution $\mathbf{w}_*$. In fact, Proposition 2 indicates that with a high probability, *any* solution lies in the random subspace spanned by the column vectors in A will be a bad approximation to $\mathbf{w}_*$. This observation leads to an interesting question: *is it possible to accurately recover the optimal solution* $\mathbf{w}_*$ *based on* $\mathbf{z}_*$*, the optimal solution to the low-dimensional optimization problem*?

Related Work Many studies are devoted to the theoretical analysis of random projection ([5] and references therein). An important property of random projection is that according to the Johnson and Lindenstrauss lemma [20]–[22], it is able to preserve the pairwise distance for a set of n data points provided the number of random projections k is sufficiently large (i.e., $k = \Omega(\epsilon^{-2}\log n)$, where ϵ is the error in approximating pairwise distance). Besides distance, random projection is also shown to preserve inner product [23], volumes and distance to affine spaces [24], under appropriate conditions. In the context of classification, it is natural to ask whether the classification margin can be preserved after random projection. For a distribution P that is linearly separable by margin γ,

Balcan et al. show that with a probability at least $1-\delta$, a random projection of P down to $\mathbb{R}^k$, where $k = \Omega(\frac{1}{\gamma^2}\log\frac{1}{\epsilon\delta})$, is linearly separable with an error at most ϵ at margin $\gamma/2$ [15]. A similar result is shown for finite samples [16]. Besides the additive error bounds shown in [15] and [16], Paul et al. show that the classification margin is preserved within a relative error bound after random projection [17]. Unlike the previous works of random projection for classification that focus on examining the preservation of classification margin and the generalization error, we focus on recovering the optimal solution after random projection.

III. Dual Random Projection

To motivate our algorithm, let us revisit the optimal primal solution $\mathbf{w}_*$ to (1), which is given in Proposition 1, i.e.,

$$\mathbf{w}_* = -\frac{1}{\lambda n} X(\boldsymbol{\alpha}_* \circ \mathbf{y}), \tag{7}$$

where $\boldsymbol{\alpha}_*$ is the optimal solution to the dual problem (2). Using random projections, we have the dual problem given in (4). Comparing (4) with the dual problem in (2), the only difference is that the matrix $X^\top X$ in (2) is replaced with $X^\top AA^\top X$ in (4). Recall that $\mathrm{E}[AA^\top] = I$. Thus, when the number of random projections m is sufficiently large, $X^\top AA^\top X$ will be close to $X^\top X$ and we would expect $\widehat{\boldsymbol{\alpha}}_*$ to be close to $\boldsymbol{\alpha}_*$. As a result, we can use $\widehat{\boldsymbol{\alpha}}_*$ to approximate $\boldsymbol{\alpha}_*$ in (7), which yields a recovered prediction model given by:

$$\widetilde{\mathbf{w}} = -\frac{1}{\lambda n} X(\widehat{\boldsymbol{\alpha}}_* \circ \mathbf{y}) = -\sum_{i=1}^{n} \frac{1}{\lambda n} y_i [\widehat{\boldsymbol{\alpha}}_*]_i \mathbf{x}_i. \tag{8}$$

Note that the key difference between the recovered solution $\widetilde{\mathbf{w}}$ and the naive solution $\widehat{\mathbf{w}}$ is that $\widehat{\mathbf{w}} = A\mathbf{z}_*$ maps the optimal primal solution $\mathbf{z}_* \in \mathbb{R}^m$ to the original space $\mathbb{R}^d$ via the random matrix A, while $\widetilde{\mathbf{w}} \propto X(\widehat{\boldsymbol{\alpha}}_* \circ \mathbf{y})$ is computed directly in the original space $\mathbb{R}^d$ using the approximate dual solution $\widehat{\boldsymbol{\alpha}}_*$. As a result, the naive solution $\widehat{\mathbf{w}}$ lies in the subspace spanned by the column vectors in the random matrix A (denoted by $\mathcal{A}$), while the recovered solution $\widetilde{\mathbf{w}}$ lies in the subspace that also contains the optimal solution $\mathbf{w}_*$, i,e., the subspace spanned by columns of X (denoted by $\mathcal{X}$). It is the mismatch between spaces $\mathcal{A}$ and $\mathcal{X}$ that leads to the large approximation error for $\widehat{\mathbf{w}}$.

Since according to Proposition 1 we can construct the dual solution $\widehat{\boldsymbol{\alpha}}_*$ from the primal solution $\mathbf{z}_*$, we do not have to solve the dual problem in (4) to obtain $\widehat{\boldsymbol{\alpha}}_*$. Instead, we can solve the low-dimensional optimization problem in (3) to obtain $\mathbf{z}_*$ and construct $\widehat{\boldsymbol{\alpha}}_*$ from $\mathbf{z}_*$. Table I shows the details of the proposed dual random projection method. Note that dual variables have been widely used in the analysis of convex optimization [25], [26] and online learning [27], the main difference is that here dual variables are used in conjunction with random projection for recovering the optimal solution.

TABLE I
A Dual Random Projection Approach

Input: input patterns $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$, binary class assignment $\mathbf{y} \in \{-1, +1\}^n$, and the number of random projections m
Output: the recovered solution $\widetilde{\mathbf{w}}$

1: Sample a Gaussian random matrix $A \in \mathbb{R}^{d\times m}$ with $A_{i,j} \sim \mathcal{N}(0, 1/m)$
2: Compute the projected data matrix as $\widehat{X} = [\widehat{\mathbf{x}}_1, \ldots, \widehat{\mathbf{x}}_n] = A^\top X$
3: Compute the primal solution $\mathbf{z}_* \in \mathbb{R}^m$ by solving the optimization problem in (3)
4: Construct the dual solution $\widehat{\boldsymbol{\alpha}}_* \in \mathbb{R}^n$ according to (5), i.e.,

$$[\widehat{\boldsymbol{\alpha}}_*]_i = \ell'\left(y_i \widehat{\mathbf{x}}_i^\top \mathbf{z}_*\right),\ i = 1, \ldots, n$$

5: Compute $\widetilde{\mathbf{w}} \in \mathbb{R}^d$ according to (8), i.e.,

$$\widetilde{\mathbf{w}} = -\frac{1}{\lambda n} X(\widehat{\boldsymbol{\alpha}}_* \circ \mathbf{y})$$

IV. Main Results

In this section, we will bound the recovery error $\|\mathbf{w}_* - \widetilde{\mathbf{w}}\|$ of dual random projection in two different scenarios, where each scenario specifies assumptions about the data matrix X and the optimal solution $\mathbf{w}_*$. In the first scenario, we assume that (i) the data matrix X is approximately low-rank, and (ii) the optimal solution $\mathbf{w}_*$ can be well approximated by a linear combination of the top eigenvectors of X. In the second scenario, we consider the case when (i) $\mathbf{w}_*$ can be approximated by a sparse vector with a support set $\mathcal{S}$, and (ii) $X^\top X$ can be well approximated by $X_{\mathcal{S}}^\top X_{\mathcal{S}}$, where $X_{\mathcal{S}}$ includes the rows of X in $\mathcal{S}$.

A. Bounding $\|\mathbf{w}_* - \widetilde{\mathbf{w}}\|$ When X Is Approximately Low-Rank

We first consider the case when X is low-rank, and then extend the result to the case when X is of full rank but can be well approximated by a low-rank matrix.

We denote by r the rank of matrix X. The following theorem shows that the recovery error is small provided that (i) X is low-rank (i.e., $r \ll \min(d, n)$), and (ii) the number of random projections is sufficiently large.

Theorem 1: Let $\mathbf{w}_*$ be the optimal solution to (1) and $\widetilde{\mathbf{w}}$ be the solution recovered by dual random projection. Suppose

$$m \geq 2(r+1)\log\frac{2r}{\delta}.$$

Then, with a probability at least $1-\delta$, we have

$$\|\widetilde{\mathbf{w}} - \mathbf{w}_*\| \leq \frac{\epsilon}{1-\epsilon}\|\mathbf{w}_*\|,$$

where

$$\epsilon = 2\sqrt{\frac{2(r+1)}{m}\log\frac{2r}{\delta}}.$$

As indicated by Theorem 1, the recovery error $\|\widetilde{\mathbf{w}} - \mathbf{w}_*\|$ is bounded by $\tilde{O}\left(\sqrt{\frac{r}{m}}\right)\|\mathbf{w}_*\|$ provided $m \geq \tilde{O}(r\log r)$, indicating a small recovery error when $r \ll d$.

Next, we proceed to analyze the case when X is of full rank but can be well approximated by a low-rank matrix. Let the

singular value decomposition (SVD) of X be

$$X = \sum_{i=1}^{d} \lambda_i \mathbf{u}_i \mathbf{v}_i^\top,$$

where $\{\lambda_1, \ldots, \lambda_d\}$ are singular values in descending order, $\mathbf{u}_i \in \mathbb{R}^d$ and $\mathbf{v}_i \in \mathbb{R}^n$ are the left and right singular vectors associated with singular value λ_i. To capture that X can be well approximated by a matrix of rank r, we assume that λ_{r+1}, the $r+1$-th eigenvalue of X, is small. In addition, we assume that $\mathbf{w}_*$, the optimal solution, can be well approximated by a linear combination of the first r left singular vectors of X. More specifically, let $U_{\bar{r}} = [\mathbf{u}_{r+1}, \ldots, \mathbf{u}_d] \in \mathbb{R}^{d\times(d-r)}$ includes the smallest $d-r$ left singular vectors of X. We assume

$$\left\| U_{\bar{r}}^\top \mathbf{w}_* \right\| \leq \rho \|\mathbf{w}_*\| \tag{9}$$

holds for some constant $\rho \ll 1$. Note that when the rank of X is r, we will have $\rho = 0$, a special case of the condition given in (9). We emphasize that the condition in (9) is critical to our analysis. This is because, the main power of random projection is to capture the top eigenspace of X, and any random projection based method will fail if most of $\mathbf{w}_*$ lies outsides the subspace spanned by the top eigenvectors of X.

Theorem 2: Suppose

$$m \geq \max\left(32(r+1), 4\log\frac{2m}{\delta}, \frac{784\gamma\, d\lambda_{r+1}^2}{9\lambda n}\right)\log\frac{d}{\delta}, \tag{10}$$

$$d \geq \max\left(r+1+\frac{m}{2}, r+2\log\frac{2m}{\delta}\right). \tag{11}$$

Then, with a probability at least $1-4\delta$, we have

$$\|\widetilde{\mathbf{w}} - \mathbf{w}_*\| \leq 2\sqrt{2\left(\frac{1}{1-\epsilon} + \frac{\gamma\,\lambda_{r+1}^2}{\lambda n}\right)} \cdot \sqrt{\left(\frac{\epsilon^2+\tau^2\rho^2}{1-\epsilon} + \frac{\gamma\,\lambda_{r+1}^2(\tau^2+\upsilon^2\rho^2)}{\lambda n}\right)}\|\mathbf{w}_*\|,$$

where ρ is given in (9), ϵ, τ, and υ are given as

$$\epsilon = 2\sqrt{\frac{2(r+1)}{m}\log\frac{2r}{\delta}}, \quad \tau = \frac{7}{3}\sqrt{\frac{2(d-r)}{m}\log\frac{d}{\delta}},$$
$$\upsilon = \frac{4(d-r+1)}{m}\log\frac{2(d-r)}{\delta}.$$

To simplify the result in Theorem 2, we consider the the case when the $r+1$-th singular value of X is small. Since the average eigenvalue of $XX^\top$ is $O(n/d)$, it is reasonable to assume $\lambda_{r+1} \leq O(\sqrt{n/d})$ when λ_{r+1} is considered to be small. The following corollary provides a simplified version of Theorem 2 when X can be well approximated by a matrix of rank r.

Corollary 3: Assume $\lambda_{r+1} \leq O\left(\sqrt{\frac{\lambda n}{\gamma d}}\right)$, $m \geq \tilde{O}(r\log d)$, and d obeys the condition in (11). With a high probability, we have

$$\|\widetilde{\mathbf{w}} - \mathbf{w}_*\| \leq \tilde{O}\left(\sqrt{\frac{r}{m}} + \rho\sqrt{\frac{d}{m}}\right)\|\mathbf{w}_*\|.$$

Furthermore, if $\rho \leq O(\sqrt{\frac{r}{d}})$, with a high probability, we have

$$\|\widetilde{\mathbf{w}} - \mathbf{w}_*\| \leq \tilde{O}\left(\sqrt{\frac{r}{m}}\right)\|\mathbf{w}_*\|,$$

similar to the result in Theorem 1.

As indicated by Corollary 3, the recovery error of the proposed dual random projection is $\tilde{O}(\sqrt{r/m})$ if (i) X can be well approximated by a matrix of rank r, and (ii) the optimal solution $\mathbf{w}_*$ can be well approximated by a linear combination of the first r singular vectors of X.

B. Bounding $\|\widetilde{\mathbf{w}} - \mathbf{w}_*\|$ When $\mathbf{w}_*$ Is Approximately Sparse

Similar to the previous subsection, we first consider the special case when the optimal solution is exactly sparse and then extend the result to the general case when $\mathbf{w}_*$ is approximately sparse.

Let S be the support set for $\mathbf{w}_*$ that includes the indices for the non-zero entries in $\mathbf{w}_*$, and let $s = |\mathcal{S}|$ be the number of nonzero elements in $\mathbf{w}_*$. We denote by $X_{\mathcal{S}} \in \mathbb{R}^{s\times n}$ the sub-matrix of X that includes the rows of X in $\mathcal{S}$, and $X_{\overline{\mathcal{S}}} \in \mathbb{R}^{(d-s)\times n}$ the sub-matrix that includes the rows of X in $\overline{\mathcal{S}} = [d] \setminus \mathcal{S}$. In this case, we assume that the support set S includes the most "important" coordinates in matrix X. More specifically, we assume

$$\eta := \|X^\top X - X_{\mathcal{S}}^\top X_{\mathcal{S}}\|_2 = \|X_{\overline{\mathcal{S}}} X_{\overline{\mathcal{S}}}^\top\|_2 \tag{12}$$

is bounded by a small constant, where $\|\cdot\|_2$ stands for the spectral norm of matrix. We note that the assumption that η is small indicates that $XX^\top$ can be well approximated by a matrix of rank s, which implies that X can be well approximated by a low-rank matrix.

We have the following theorem to bound the recovery error.

Theorem 4: Suppose

$$m \geq \max\left(32(s+1), 4\log\frac{2m}{\delta}, \frac{784\gamma\, d\eta}{9\lambda n}\right)\log\frac{d}{\delta}, \tag{13}$$

$$d \geq \max\left(s+2\log\frac{2m}{\delta}, 2s\right). \tag{14}$$

Then, with a probability at least $1-3\delta$, we have

$$\|\widetilde{\mathbf{w}} - \mathbf{w}_*\| \leq 2\sqrt{\left(\frac{1}{1-\epsilon} + \frac{\gamma\,\eta}{\lambda n}\right)\left(\frac{\epsilon^2}{1-\epsilon} + \frac{\gamma\,\eta\tau^2}{\lambda n}\right)}\|\mathbf{w}_*\|,$$

where ϵ and τ are given by

$$\epsilon = 2\sqrt{\frac{2(s+1)}{m}\log\frac{2s}{\delta}}, \text{ and } \tau = \frac{7}{3}\sqrt{\frac{2(d-s)}{m}\log\frac{d}{\delta}}.$$

The following corollary provides a simplified result for small η.

Corollary 5: Assume $\eta \leq O(\frac{\lambda n}{\gamma d})$, $m \geq \tilde{O}(s\log d)$, and d obeys the condition in (14). With a high probability, we have

$$\|\widetilde{\mathbf{w}} - \mathbf{w}_*\| \leq \tilde{O}\left(\sqrt{\frac{s}{m}}\right)\|\mathbf{w}_*\|,$$

similar to the result in Theorem 1.

We now proceed to bound the general case, i.e., when $\mathbf{w}_*$ can be approximated by a sparse vector. Let $\mathcal{S}$ include the indices of the first s entries in $\mathbf{w}_*$ with the largest magnitude.

We denote by $[\mathbf{w}_*]_{\mathcal{S}} \in \mathbb{R}^s$ the sub-vector of $\mathbf{w}_*$ that includes the entries of $\mathbf{w}_*$ in $\mathcal{S}$, and $[\mathbf{w}_*]_{\overline{\mathcal{S}}} \in \mathbb{R}^{d-s}$ the sub-vector that includes the entries of $\mathbf{w}_*$ in $\overline{\mathcal{S}} = [d] \backslash \mathcal{S}$. To capture that $\mathbf{w}_*$ is approximately sparse, we assume that

$$\left\| [\mathbf{w}_*]_{\overline{\mathcal{S}}} \right\| \leq \rho \|\mathbf{w}_*\| \tag{15}$$

holds for some small constant ρ.

Theorem 6: Suppose

$$m \geq \max\left(32(s+1), 4\log\frac{2m}{\delta}, \frac{784\gamma\, d\eta}{9\lambda n}\right)\log\frac{d}{\delta}, \tag{16}$$

$$d \geq \max\left(s+1+\frac{m}{2}, s+2\log\frac{2m}{\delta}\right). \tag{17}$$

Then, with a probability at least $1-4\delta$, we have

$$\|\widetilde{\mathbf{w}} - \mathbf{w}_*\| \leq 2\sqrt{2\left(\frac{1}{1-\epsilon}+\frac{\gamma\,\eta}{\lambda n}\right)} \cdot\sqrt{\left(\frac{\epsilon^2+\tau^2\rho^2}{1-\epsilon}+\frac{\gamma\,\eta(\tau^2+\upsilon^2\rho^2)}{\lambda n}\right)}\|\mathbf{w}_*\|,$$

where ρ is given in (15), and ϵ, τ, and υ are given by

$$\epsilon = 2\sqrt{\frac{2(s+1)}{m}\log\frac{2s}{\delta}}, \quad \tau = \frac{7}{3}\sqrt{\frac{2(d-s)}{m}\log\frac{d}{\delta}},$$
$$\upsilon = \frac{4(d-s+1)}{m}\log\frac{2(d-s)}{\delta}.$$

Moreover, if $\eta \leq O(\frac{\lambda n}{\gamma d})$ and $m \geq \tilde{O}(s\log d)$, with a high probability, we have

$$\|\widetilde{\mathbf{w}} - \mathbf{w}_*\| \leq O\left(\sqrt{\frac{s}{m}}+\rho\sqrt{\frac{d}{m}}\right)\|\mathbf{w}_*\|.$$

V. The Analysis

Our analysis is built upon the following lemma, which reveals the relationship between $\widehat{\boldsymbol{\alpha}}_*$ and $\boldsymbol{\alpha}_*$.

Lemma 1: Let $\boldsymbol{\alpha}_* \in \mathbb{R}^n$ and $\widehat{\boldsymbol{\alpha}}_* \in \mathbb{R}^n$ be the optimal dual solutions to (2) and (4), respectively. Then, we have

$$[(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)\circ\mathbf{y}]^\top \widehat{G}[(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)\circ\mathbf{y}] + \frac{\lambda n\|\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*\|^2}{\gamma} \leq [(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)\circ\mathbf{y}]^\top (G-\widehat{G})(\boldsymbol{\alpha}_*\circ\mathbf{y}) \tag{18}$$

where $G = X^\top X$ and $\widehat{G} = X^\top AA^\top X$.

Proof: For the convenience of presentation, we consider the minimization version of the dual problem, i.e.,

$$\min_{\boldsymbol{\alpha}\in\Omega^n} \widehat{L}(\boldsymbol{\alpha}) = \sum_{i=1}^n \ell_*(\alpha_i) + \frac{1}{2\lambda n}(\boldsymbol{\alpha}\circ\mathbf{y})^\top \widehat{G}(\boldsymbol{\alpha}\circ\mathbf{y}).$$

We denote by $L(\boldsymbol{\alpha})$ the objective function of the dual problem without random projection, i.e.,

$$L(\boldsymbol{\alpha}) = \sum_{i=1}^n \ell_*(\alpha_i) + \frac{1}{2\lambda n}(\boldsymbol{\alpha}\circ\mathbf{y})^\top G(\boldsymbol{\alpha}\circ\mathbf{y}).$$

Because $\boldsymbol{\alpha}_*$ and $\widehat{\boldsymbol{\alpha}}_*$ minimize $L(\cdot)$ and $\widehat{L}(\cdot)$ respectively, from the optimality condition of convex optimization [25], we have

$$\langle \nabla L(\boldsymbol{\alpha}_*), \widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*\rangle \geq 0, \tag{19}$$

$$\langle \nabla \widehat{L}(\widehat{\boldsymbol{\alpha}}_*), \boldsymbol{\alpha}_* - \widehat{\boldsymbol{\alpha}}_*\rangle \geq 0. \tag{20}$$

Notice that the smoothness assumption of $\ell(\cdot)$ implies that $\ell_*(\cdot)$ is $\frac{1}{\gamma}$-strongly convex [28]. Let $F(\boldsymbol{\alpha}) = \sum_{i=1}^n \ell_*(\alpha_i)$, which is also $\frac{1}{\gamma}$-strongly convex. From the definition of strong convexity [29], we have

$$F(\boldsymbol{\alpha}_*) \geq F(\widehat{\boldsymbol{\alpha}}_*) + \langle \nabla F(\widehat{\boldsymbol{\alpha}}_*), \boldsymbol{\alpha}_* - \widehat{\boldsymbol{\alpha}}_*\rangle + \frac{\|\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*\|^2}{2\gamma}. \tag{21}$$

Furthermore, it is easy to verify that

$$\begin{aligned}\frac{1}{2\lambda n}(\boldsymbol{\alpha}_*\circ\mathbf{y})^\top \widehat{G}(\boldsymbol{\alpha}_*\circ\mathbf{y}) &= \frac{(\widehat{\boldsymbol{\alpha}}_*\circ\mathbf{y})^\top \widehat{G}(\widehat{\boldsymbol{\alpha}}_*\circ\mathbf{y})}{2\lambda n} + \frac{\langle \widehat{G}(\widehat{\boldsymbol{\alpha}}_*\circ\mathbf{y}), (\boldsymbol{\alpha}_* - \widehat{\boldsymbol{\alpha}}_*)\circ\mathbf{y}\rangle}{\lambda n} \\ &\quad + \frac{1}{2\lambda n}[(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)\circ\mathbf{y}]^\top \widehat{G}[(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)\circ\mathbf{y}].\end{aligned} \tag{22}$$

Adding (21) to (22), we obtain

$$\begin{aligned}\widehat{L}(\boldsymbol{\alpha}_*) &\geq \widehat{L}(\widehat{\boldsymbol{\alpha}}_*) + \langle \nabla F(\widehat{\boldsymbol{\alpha}}_*), \boldsymbol{\alpha}_* - \widehat{\boldsymbol{\alpha}}_*\rangle + \frac{1}{2\gamma}\|\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*\|^2 \\ &\quad + \frac{1}{\lambda n}\langle \widehat{G}(\widehat{\boldsymbol{\alpha}}_*\circ\mathbf{y}), (\boldsymbol{\alpha}_* - \widehat{\boldsymbol{\alpha}}_*)\circ\mathbf{y}\rangle \\ &\quad + \frac{1}{2\lambda n}[(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)\circ\mathbf{y}]^\top \widehat{G}[(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)\circ\mathbf{y}] \\ &= \widehat{L}(\widehat{\boldsymbol{\alpha}}_*) + \langle \nabla \widehat{L}(\widehat{\boldsymbol{\alpha}}_*), \boldsymbol{\alpha}_* - \widehat{\boldsymbol{\alpha}}_*\rangle + \frac{1}{2\gamma}\|\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*\|^2 \\ &\quad + [(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)\circ\mathbf{y}]^\top \widehat{G}[(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)\circ\mathbf{y}] \\ &\overset{(20)}{\geq} \widehat{L}(\widehat{\boldsymbol{\alpha}}_*) + \frac{1}{2\gamma}\|\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*\|^2 \\ &\quad + \frac{1}{2\lambda n}[(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)\circ\mathbf{y}]^\top \widehat{G}[(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)\circ\mathbf{y}].\end{aligned} \tag{23}$$

On the other hand, we have

$$\begin{aligned}&\widehat{L}(\boldsymbol{\alpha}_*) + \frac{1}{\lambda n}[(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)\circ\mathbf{y}]^\top (\widehat{G} - G)(\boldsymbol{\alpha}_*\circ\mathbf{y}) \\ &\quad = \widehat{L}(\boldsymbol{\alpha}_*) + \langle \nabla \widehat{L}(\boldsymbol{\alpha}_*) - \nabla L(\boldsymbol{\alpha}_*), \widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*\rangle \\ &\quad \overset{(19)}{\leq} \widehat{L}(\boldsymbol{\alpha}_*) + \langle \nabla \widehat{L}(\boldsymbol{\alpha}_*), \widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*\rangle \\ &\quad \leq \widehat{L}(\widehat{\boldsymbol{\alpha}}_*) - \frac{1}{2\lambda n}[(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)\circ\mathbf{y}]^\top \widehat{G}[(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)\circ\mathbf{y}] \\ &\quad\quad - \frac{1}{2\gamma}\|\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*\|^2\end{aligned} \tag{24}$$

where the last inequality follows from the convexity of $\widehat{L}(\boldsymbol{\alpha})$. We complete the proof by combining (23) and (24). ∎

A. Proof of Theorem 1

Let the SVD of X be

$$X = U\Sigma V^\top = \sum_{i=1}^r \lambda_i \mathbf{u}_i \mathbf{v}_i^\top,$$

where $\Sigma = \mathrm{diag}(\lambda_1, \ldots, \lambda_r)$, $U = [\mathbf{u}_1, \ldots, \mathbf{u}_r]$, $V = [\mathbf{v}_1, \ldots, \mathbf{v}_r]$, λ_i is the i-th singular value of X, $\mathbf{u}_i \in \mathbb{R}^d$ and $\mathbf{v}_i \in \mathbb{R}^n$ are the corresponding left and right singular vectors of X. Then, we can rewrite G and $\widehat{G}$ in Lemma 1 as

$$G = V\Sigma U^\top U\Sigma V^\top = V\Sigma^2 V^\top,$$
$$\widehat{G} = V\Sigma U^\top AA^\top U\Sigma V^\top = V\Sigma BB^\top \Sigma V^\top,$$

where

$$B = U^\top A \in \mathbb{R}^{r\times m}.$$

It is easy to verify that B can be treated as a random matrix, each element of which is independently sampled from a Gaussian distribution $\mathcal{N}(0, 1/m)$.

To simplify the notation, we define

$$\mathbf{a} = \Sigma V^\top[(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)\circ \mathbf{y}],\ \mathbf{c} = \Sigma V^\top(\boldsymbol{\alpha}_*\circ \mathbf{y}),$$
$$\epsilon = \|BB^\top - I\|_2.$$

Since U is an orthogonal matrix, we have

$$\|\widetilde{\mathbf{w}} - \mathbf{w}_*\| = \left\|\frac{1}{\lambda n}X[(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)\circ \mathbf{y}]\right\| = \frac{1}{\lambda n}\|\mathbf{a}\|, \quad (25)$$

$$\|\mathbf{w}_*\| = \left\|-\frac{1}{\lambda n}X(\boldsymbol{\alpha}_*\circ \mathbf{y})\right\| = \frac{1}{\lambda n}\|\mathbf{c}\|. \quad (26)$$

From Lemma 1, we have

$$\mathbf{a}^\top BB^\top \mathbf{a} \le \mathbf{a}^\top\left(I - BB^\top\right)\mathbf{c},$$

which implies

$$\|\mathbf{a}\|^2(1-\epsilon) \le \epsilon\|\mathbf{a}\|\|\mathbf{c}\| \Rightarrow \|\mathbf{a}\|(1-\epsilon) \le \epsilon\|\mathbf{c}\|. \quad (27)$$

From (25), (26), and (27), we obtain the second inequality in Theorem 1.

To bound ϵ, we have the following concentration inequality for Gaussian random matrix.

Lemma 2: Let $\delta \in (0, 1)$ be the failure probability. With a probability at least $1-\delta$, we have

$$\epsilon = \left\|BB^\top - I\right\|_2 \le 2\sqrt{\frac{2(r+1)}{m}\log\frac{2r}{\delta}},$$

provided $m \ge 2(r+1)\log\frac{2r}{\delta}$.

The proof of Lemma 2 and other omitted proofs are deferred to the Appendix.

B. Proof of Theorem 2

Based on the SVD of X, we introduce the following notations

$$U_r = [\mathbf{u}_1, \ldots, \mathbf{u}_r], \qquad U_{\bar{r}} = [\mathbf{u}_{r+1}, \ldots, \mathbf{u}_d],$$
$$\Sigma_r = \mathrm{diag}(\lambda_1, \ldots, \lambda_r), \quad \Sigma_{\bar{r}} = \mathrm{diag}(\lambda_{r+1}, \ldots, \lambda_d),$$
$$V_r = [\mathbf{v}_1, \ldots, \mathbf{v}_r], \qquad V_{\bar{r}} = [\mathbf{v}_{r+1}, \ldots, \mathbf{v}_d].$$

Then, we can rewrite G and $\widehat{G}$ in Lemma 1 as

$$G = V_r\Sigma_r^2V_r^\top + V_{\bar{r}}\Sigma_{\bar{r}}^2V_{\bar{r}}^\top,$$
$$\widehat{G} = V_r\Sigma_rB_rB_r^\top\Sigma_rV_r^\top + V_{\bar{r}}\Sigma_{\bar{r}}B_{\bar{r}}B_{\bar{r}}^\top\Sigma_{\bar{r}}V_{\bar{r}}^\top + V_{\bar{r}}\Sigma_{\bar{r}}B_{\bar{r}}B_r^\top\Sigma_rV_r^\top + V_r\Sigma_rB_rB_{\bar{r}}^\top\Sigma_{\bar{r}}V_{\bar{r}}^\top,$$

where

$$B_r = U_r^\top A \in \mathbb{R}^{r\times m},\ B_{\bar{r}} = U_{\bar{r}}^\top A \in \mathbb{R}^{(d-r)\times m}.$$

It is straightforward to check that both B and $B_{\bar{r}}$ can be treated as two *independent* Gaussian random matrices, where each entry of these two matrices is independently sampled from a Gaussian distribution $\mathcal{N}(0, 1/m)$.

Define

$$\mathbf{a} = \Sigma_rV_r^\top[(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)\circ\mathbf{y}], \quad \mathbf{b} = \Sigma_{\bar{r}}V_{\bar{r}}^\top[(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)\circ\mathbf{y}],$$
$$\mathbf{c} = \Sigma_rV_r^\top(\boldsymbol{\alpha}_*\circ\mathbf{y}), \qquad \mathbf{d} = \Sigma_{\bar{r}}V_{\bar{r}}^\top(\boldsymbol{\alpha}_*\circ\mathbf{y}),$$
$$\epsilon = \left\|B_rB_r^\top - I\right\|_2, \qquad \tau = \left\|B_{\bar{r}}B_r^\top\right\|_2,$$
$$\upsilon = \left\|B_{\bar{r}}B_{\bar{r}}^\top - I\right\|_2.$$

It is easy to verify that

$$\|\widetilde{\mathbf{w}} - \mathbf{w}_*\|^2 = \frac{1}{\lambda^2n^2}\|\mathbf{a}\|^2 + \frac{1}{\lambda^2n^2}\|\mathbf{b}\|^2, \quad (28)$$

$$\|\mathbf{w}_*\|^2 = \frac{1}{\lambda^2n^2}\|\mathbf{c}\|^2 + \frac{1}{\lambda^2n^2}\|\mathbf{d}\|^2, \quad (29)$$

$$\|\mathbf{d}\| = \lambda n\|U_{\bar{r}}^\top\mathbf{w}_*\| \overset{(9)}{\le} \lambda n\rho\|\mathbf{w}_*\|. \quad (30)$$

Using the definition of $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$, and $\mathbf{d}$, we bound $[(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)\circ\mathbf{y}]^\top\widehat{G}[(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)\circ\mathbf{y}]$, the first term in (18), as

$$\begin{aligned}&[(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)\circ\mathbf{y}]^\top\widehat{G}[(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)\circ\mathbf{y}]\\ &\quad= \mathbf{a}^\top B_rB_r^\top\mathbf{a} + \mathbf{b}^\top B_{\bar{r}}B_{\bar{r}}^\top\mathbf{b}\\ &\qquad+ \mathbf{a}^\top B_rB_{\bar{r}}^\top\mathbf{b} + \mathbf{b}^\top B_{\bar{r}}B_r^\top\mathbf{a}\\ &\quad\ge \|\mathbf{a}\|^2(1-\epsilon) - 2\|\mathbf{a}\|\|\mathbf{b}\|\tau, \end{aligned} \quad (31)$$

and $\lambda n\|\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}\|^2$, the second term in (18), as

$$\begin{aligned}\frac{\lambda n}{\gamma}\|\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*\|^2 &\overset{y_i\in\pm1}{=} \frac{\lambda n}{\gamma}\|(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)\circ\mathbf{y}\|^2\\ &\ge \frac{\lambda n}{\gamma}\frac{\left\|\Sigma_{\bar{r}}V_{\bar{r}}^\top[(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)\circ\mathbf{y}]\right\|^2}{\|V_{\bar{r}}\Sigma_{\bar{r}}^2V_{\bar{r}}^\top\|_2} \ge \frac{\lambda n}{\gamma\,\lambda_{r+1}^2}\|\mathbf{b}\|^2. \end{aligned} \quad (32)$$

Finally, the last term in (18) is upper bounded by

$$\begin{aligned}&[(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)\circ\mathbf{y}]^\top(G - \widehat{G})(\boldsymbol{\alpha}_*\circ\mathbf{y})\\ &\quad= \mathbf{a}^\top(I - B_rB_r^\top)\mathbf{c} + \mathbf{b}^\top(I - B_{\bar{r}}B_{\bar{r}}^\top)\mathbf{d}\\ &\qquad- \mathbf{a}^\top B_rB_{\bar{r}}^\top\mathbf{d} - \mathbf{b}^\top B_{\bar{r}}B_r^\top\mathbf{c}\\ &\quad\le \|\mathbf{a}\|\|\mathbf{c}\|\epsilon + \|\mathbf{b}\|\|\mathbf{d}\|\upsilon + \|\mathbf{a}\|\|\mathbf{d}\|\tau + \|\mathbf{b}\|\|\mathbf{c}\|\tau. \end{aligned} \quad (33)$$

From (18), (31), (32), and (33), we have

$$\begin{aligned}&(1-\epsilon)\|\mathbf{a}\|^2 - 2\tau\|\mathbf{a}\|\|\mathbf{b}\| + \frac{\lambda n}{\gamma\,\lambda_{r+1}^2}\|\mathbf{b}\|^2\\ &\quad\le \|\mathbf{a}\|\|\mathbf{c}\|\epsilon + \|\mathbf{b}\|\|\mathbf{d}\|\upsilon + \|\mathbf{a}\|\|\mathbf{d}\|\tau + \|\mathbf{b}\|\|\mathbf{c}\|\tau. \end{aligned} \quad (34)$$

In the case when

$$4\tau^2 \le \frac{(1-\epsilon)\lambda n}{\gamma\,\lambda_{r+1}^2}, \quad (35)$$

we have

$$\frac{1-\epsilon}{2}\|\mathbf{a}\|^2 - 2\tau\|\mathbf{a}\|\|\mathbf{b}\| + \frac{\lambda n}{2\gamma\,\lambda_{r+1}^2}\|\mathbf{b}\|^2 \ge 0. \quad (36)$$

From (34) and (36), we have

$$\begin{aligned}
&\frac{1-\epsilon}{2}\|\mathbf{a}\|^2+\frac{\lambda n}{2\gamma\,\lambda_{r+1}^2}\|\mathbf{b}\|^2\\
&\le \|\mathbf{a}\|\|\mathbf{c}\|\epsilon+\|\mathbf{b}\|\|\mathbf{d}\|\upsilon+\|\mathbf{a}\|\|\mathbf{d}\|\tau+\|\mathbf{b}\|\|\mathbf{c}\|\tau\\
&\le \frac{1-\epsilon}{8}\|\mathbf{a}\|^2+\frac{2\epsilon^2}{1-\epsilon}\|\mathbf{c}\|^2+\frac{\lambda n}{8\gamma\,\lambda_{r+1}^2}\|\mathbf{b}\|^2+\frac{2\gamma\,\lambda_{r+1}^2\upsilon^2}{\lambda n}\|\mathbf{d}\|^2\\
&\quad+\frac{1-\epsilon}{8}\|\mathbf{a}\|^2+\frac{2\tau^2}{1-\epsilon}\|\mathbf{d}\|^2\\
&\quad+\frac{\lambda n}{8\gamma\,\lambda_{r+1}^2}\|\mathbf{b}\|^2+\frac{2\gamma\,\lambda_{r+1}^2\tau^2}{\lambda n}\|\mathbf{c}\|^2,
\end{aligned}$$

which implies

$$\begin{aligned}
&\frac{1-\epsilon}{4}\|\mathbf{a}\|^2+\frac{\lambda n}{4\gamma\,\lambda_{r+1}^2}\|\mathbf{b}\|^2\\
&\qquad\le \left(\frac{2\epsilon^2}{1-\epsilon}+\frac{2\gamma\,\lambda_{r+1}^2\tau^2}{\lambda n}\right)\|\mathbf{c}\|^2\\
&\qquad\quad+\left(\frac{2\gamma\,\lambda_{r+1}^2\upsilon^2}{\lambda n}+\frac{2\tau^2}{1-\epsilon}\right)\|\mathbf{d}\|^2\\
&\qquad\overset{(29,30)}{\le} \left(\frac{2\epsilon^2}{1-\epsilon}+\frac{2\gamma\,\lambda_{r+1}^2\tau^2}{\lambda n}\right)\lambda^2n^2\|\mathbf{w}_*\|^2\\
&\qquad\quad+\left(\frac{2\gamma\,\lambda_{r+1}^2\upsilon^2}{\lambda n}+\frac{2\tau^2}{1-\epsilon}\right)\lambda^2n^2\rho^2\|\mathbf{w}_*\|^2.
\end{aligned}$$

As a result, we can upper bound $\|\widetilde{\mathbf{w}}-\mathbf{w}_*\|^2$ by

$$\begin{aligned}
\|\widetilde{\mathbf{w}}-\mathbf{w}_*\|^2 &\overset{(28)}{\le} 8\left(\frac{1}{1-\epsilon}+\frac{\gamma\,\lambda_{r+1}^2}{\lambda n}\right)\\
&\quad\cdot\left(\frac{\epsilon^2+\tau^2\rho^2}{1-\epsilon}+\frac{\gamma\,\lambda_{r+1}^2(\tau^2+\upsilon^2\rho^2)}{\lambda n}\right)\|\mathbf{w}_*\|^2
\end{aligned}$$

leading to the third inequality in Theorem 2.

Next, we discuss how to bound ϵ, τ and υ. Since

$$m \overset{(10)}{\ge} 32(r+1)\log\frac{d}{\delta} \overset{(10,11)}{\ge} 2(r+1)\log\frac{2r}{\delta},$$

similar to Lemma 2, we have the following lemma.

Lemma 3: Let $\delta\in(0,1)$ be the failure probability. With a probability at least $1-\delta$, we have

$$\epsilon=\left\|B_rB_r^\top-I\right\|_2\le 2\sqrt{\frac{2(r+1)}{m}\log\frac{2r}{\delta}},$$

provided the conditions in (10) and (11) hold.

Based on the noncommutative variant of Bernstein's inequality [30], we have the following lemma to bound τ.

Lemma 4: Let $\delta\in(0,1/2)$ be the failure probability. Then, with a probability at least $1-2\delta$, we have

$$\tau=\|B_{\bar r}B_r^\top\|_2\le\frac{7}{3}\sqrt{\frac{2(d-r)}{m}\log\frac{d}{\delta}},$$

provided the conditions in (10) and (11) hold.

Following a similar proof of Lemma 2, we have the following lemma to bound υ.

Lemma 5: Let $\delta\in(0,1)$ be the failure probability. With a probability at least $1-\delta$, we have

$$\upsilon=\left\|B_{\bar r}B_{\bar r}^\top-I\right\|_2\le\frac{4(d-r+1)}{m}\log\frac{2(d-r)}{\delta},$$

provided the condition in (11) holds.

Finally, we need to show that (35) is true given our assumptions. From Lemma 3, it is straightforward to check that

$$\epsilon\overset{(10)}{\le}2\sqrt{\frac{2(r+1)\log 2r/\delta}{32(r+1)\log d/\delta}}\overset{(11)}{\le}\frac{1}{2}. \tag{37}$$

Based on Lemma 4, we have

$$4\tau^2\le 4\frac{49}{9}\frac{2d}{m}\log\frac{d}{\delta}\overset{(10)}{\le}\frac{\lambda n}{2\gamma\,\lambda_{r+1}^2}\overset{(37)}{\le}\frac{(1-\epsilon)\lambda n}{\gamma\,\lambda_{r+1}^2}.$$

C. Proof of Theorem 4

Since $\mathbf{w}_*$ is sparse, $\boldsymbol{\alpha}_*\circ\mathbf{y}$ is orthogonal to the subspace spanned by the rows in $X_{\overline{\mathcal{S}}}$, as revealed by the following lemma.

Lemma 6: Assume $\mathbf{w}_*$ is supported by a subset $\mathcal{S}\subset[d]$. We have

$$X_{\overline{\mathcal{S}}}(\boldsymbol{\alpha}_*\circ\mathbf{y})=0. \tag{38}$$

We denote by $A_{\mathcal{S}}\in\mathbb{R}^{s\times m}$ the sub-matrix of A that includes the rows of A in $\mathcal{S}$, and $A_{\overline{\mathcal{S}}}\in\mathbb{R}^{(d-s)\times m}$ the sub-matrix that includes the rows of A in $\overline{\mathcal{S}}$. Using these definitions, we rewrite $\widehat{G}$ in Lemma 1 as

$$\begin{aligned}
\widehat{G}&=X_{\mathcal{S}}^\top A_{\mathcal{S}}A_{\mathcal{S}}^\top X_{\mathcal{S}}+X_{\overline{\mathcal{S}}}^\top A_{\overline{\mathcal{S}}}A_{\overline{\mathcal{S}}}^\top X_{\overline{\mathcal{S}}}\\
&\quad+X_{\mathcal{S}}^\top A_{\mathcal{S}}A_{\overline{\mathcal{S}}}^\top X_{\overline{\mathcal{S}}}+X_{\overline{\mathcal{S}}}^\top A_{\overline{\mathcal{S}}}A_{\mathcal{S}}^\top X_{\mathcal{S}}.
\end{aligned}$$

We define

$$\begin{aligned}
&\mathbf{a}=X_{\mathcal{S}}[(\widehat{\boldsymbol{\alpha}}_*-\boldsymbol{\alpha}_*)\circ\mathbf{y}],\quad \mathbf{b}=X_{\overline{\mathcal{S}}}(\widehat{\boldsymbol{\alpha}}_*\circ\mathbf{y}),\\
&\mathbf{c}=X_{\mathcal{S}}(\boldsymbol{\alpha}_*\circ\mathbf{y}),\\
&\epsilon=\left\|A_{\mathcal{S}}A_{\mathcal{S}}^\top-I\right\|_2,\quad \tau=\left\|A_{\overline{\mathcal{S}}}A_{\mathcal{S}}^\top\right\|_2.
\end{aligned}$$

Then, we have

$$\|\widetilde{\mathbf{w}}-\mathbf{w}_*\|^2\overset{(38)}{=}\frac{1}{\lambda^2n^2}\|\mathbf{a}\|^2+\frac{1}{\lambda^2n^2}\|\mathbf{b}\|^2, \tag{39}$$

$$\|\mathbf{w}_*\|=\frac{1}{\lambda n}\|\mathbf{c}\|. \tag{40}$$

Based on the above definitions, we bound the three terms in (18) as follows.

$$\begin{aligned}
&[(\widehat{\boldsymbol{\alpha}}_*-\boldsymbol{\alpha}_*)\circ\mathbf{y}]^\top\widehat{G}[(\widehat{\boldsymbol{\alpha}}_*-\boldsymbol{\alpha}_*)\circ\mathbf{y}]\\
&\quad\overset{(38)}{=}\mathbf{a}^\top A_{\mathcal{S}}A_{\mathcal{S}}^\top\mathbf{a}+\mathbf{b}^\top A_{\overline{\mathcal{S}}}A_{\overline{\mathcal{S}}}^\top\mathbf{b}\\
&\qquad+\mathbf{a}^\top A_{\mathcal{S}}A_{\overline{\mathcal{S}}}^\top\mathbf{b}+\mathbf{b}_{\overline{\mathcal{S}}}^\top A_{\overline{\mathcal{S}}}A_{\mathcal{S}}^\top\mathbf{a}\\
&\quad\ge(1-\epsilon)\|\mathbf{a}\|^2-2\tau\|\mathbf{a}\|\|\mathbf{b}\|. \qquad (41)\\
&\frac{\lambda n}{\gamma}\|\widehat{\boldsymbol{\alpha}}_*-\boldsymbol{\alpha}_*\|^2\overset{y_i\in\pm1}{=}\frac{\lambda n}{\gamma}\|(\widehat{\boldsymbol{\alpha}}_*-\boldsymbol{\alpha}_*)\circ\mathbf{y}\|^2\\
&\quad\ge\frac{\lambda n}{\gamma}\frac{\left\|X_{\overline{\mathcal{S}}}[(\widehat{\boldsymbol{\alpha}}_*-\boldsymbol{\alpha}_*)\circ\mathbf{y}]\right\|^2}{\|X_{\overline{\mathcal{S}}}^\top X_{\overline{\mathcal{S}}}\|_2}\overset{(12,38)}{=}\frac{\lambda n}{\gamma\,\eta}\|\mathbf{b}\|^2. \qquad (42)\\
&[(\widehat{\boldsymbol{\alpha}}_*-\boldsymbol{\alpha}_*)\circ\mathbf{y}]^\top(G-\widehat{G})(\boldsymbol{\alpha}_*\circ\mathbf{y})
\end{aligned}$$

$$\stackrel{(38)}{=} \mathbf{a}^\top (I - A_{\mathcal{S}} A_{\mathcal{S}}^\top)\mathbf{c} - \mathbf{b}_{\overline{\mathcal{S}}}^\top A_{\overline{\mathcal{S}}} A_{\mathcal{S}}^\top \mathbf{c}$$
$$\leq \|\mathbf{c}\|(\epsilon\|\mathbf{a}\| + \tau\|\mathbf{b}\|). \quad (43)$$

From (18), (41), (42), and (43), we have

$$(1-\epsilon)\|\mathbf{a}\|^2 - 2\tau\|\mathbf{a}\|\|\mathbf{b}\| + \frac{\lambda n}{\gamma\,\eta}\|\mathbf{b}\|^2 \leq \|c\|\,(\epsilon\|\mathbf{a}\| + \tau\|\mathbf{b}\|). \quad (44)$$

In the case when

$$4\tau^2 \leq \frac{(1-\epsilon)\lambda n}{\gamma\,\eta}, \quad (45)$$

we have

$$\frac{1-\epsilon}{2}\|\mathbf{a}\|^2 - 2\tau\|\mathbf{a}\|\|\mathbf{b}\| + \frac{\lambda n}{2\gamma\,\eta}\|\mathbf{b}\|^2 \geq 0. \quad (46)$$

From (44) and (46), we have

$$\frac{1-\epsilon}{2}\|\mathbf{a}\|^2 + \frac{\lambda n}{2\gamma\,\eta}\|\mathbf{b}\|^2 \leq \|\mathbf{c}\|\,(\epsilon\|\mathbf{a}\| + \tau\|\mathbf{b}\|) \leq \frac{\epsilon^2}{1-\epsilon}\|\mathbf{c}\|^2 + \frac{1-\epsilon}{4}\|\mathbf{a}\|^2 + \frac{\gamma\,\eta\tau^2}{\lambda n}\|\mathbf{c}\|^2 + \frac{\lambda n}{4\gamma\,\eta}\|\mathbf{b}\|^2,$$

which implies

$$\frac{1-\epsilon}{4}\|\mathbf{a}\|^2 + \frac{\lambda n}{4\gamma\,\eta}\|\mathbf{b}\|^2 \leq \left(\frac{\epsilon^2}{1-\epsilon} + \frac{\gamma\,\eta\tau^2}{\lambda n}\right)\|\mathbf{c}\|^2. \quad (47)$$

Then, we can upper bound $\|\widetilde{\mathbf{w}} - \mathbf{w}_*\|^2$ by

$$\|\widetilde{\mathbf{w}} - \mathbf{w}_*\|^2 \stackrel{(39,47)}{\leq} \frac{1}{\lambda^2 n^2}\left(\frac{4}{1-\epsilon} + \frac{4\gamma\,\eta}{\lambda n}\right)\left(\frac{\epsilon^2}{1-\epsilon} + \frac{\gamma\,\eta\tau^2}{\lambda n}\right)\|\mathbf{c}\|^2 \stackrel{(40)}{\leq} \left(\frac{4}{1-\epsilon} + \frac{4\gamma\,\eta}{\lambda n}\right)\left(\frac{\epsilon^2}{1-\epsilon} + \frac{\gamma\,\eta\tau^2}{\lambda n}\right)\|\mathbf{w}_*\|^2$$

leading to the third inequality in Theorem 4.

Similar to Lemmas 3 and 4, we have the following lemmas to bound ϵ and τ.

Lemma 7: Let $\delta \in (0, 1)$ be the failure probability. With a probability at least $1 - \delta$, we have

$$\epsilon = \left\|A_{\mathcal{S}} A_{\mathcal{S}}^\top - I\right\|_2 \leq 2\sqrt{\frac{2(s+1)}{m}\log\frac{2s}{\delta}},$$

provided the conditions in (13) and (14) hold.

Lemma 8: Let $\delta \in (0, 1/2)$ be the failure probability. Then, with a probability at least $1 - 2\delta$, we have

$$\tau = \|A_{\overline{\mathcal{S}}} A_{\mathcal{S}}^\top\|_2 \leq \frac{7}{3}\sqrt{\frac{2(d-s)}{m}\log\frac{d}{\delta}},$$

provided the conditions in (13) and (14) hold.

Finally, we need to show that (45) is true given our assumptions. From Lemma 7, it is straightforward to check that

$$\epsilon \stackrel{(13)}{\leq} 2\sqrt{\frac{2(s+1)\log 2s/\delta}{32(s+1)\log d/\delta}} \stackrel{(14)}{\leq} \frac{1}{2}. \quad (48)$$

Based on Lemma 8, we have

$$4\tau^2 \leq 4\frac{49}{9}\frac{2d}{m}\log\frac{d}{\delta} \stackrel{(13)}{\leq} \frac{\lambda n}{2\gamma\,\eta} \stackrel{(48)}{\leq} \frac{(1-\epsilon)\lambda n}{\gamma\,\eta}.$$

D. Proof of Theorem 6

Similar to the proof of Theorem 4, we define

$$\mathbf{a} = X_{\mathcal{S}}[(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*) \circ \mathbf{y}], \quad \mathbf{b} = X_{\overline{\mathcal{S}}}[(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*) \circ \mathbf{y}],$$
$$\mathbf{c} = X_{\mathcal{S}}(\boldsymbol{\alpha}_* \circ \mathbf{y}), \quad \mathbf{d} = X_{\overline{\mathcal{S}}}(\boldsymbol{\alpha}_* \circ \mathbf{y}),$$
$$\epsilon = \left\|A_{\mathcal{S}} A_{\mathcal{S}}^\top - I\right\|_2, \quad \tau = \left\|A_{\overline{\mathcal{S}}} A_{\mathcal{S}}^\top\right\|_2,$$
$$\upsilon = \left\|A_{\overline{\mathcal{S}}} A_{\overline{\mathcal{S}}}^\top - I\right\|_2.$$

Then, we have

$$\|\widetilde{\mathbf{w}} - \mathbf{w}_*\|^2 = \frac{1}{\lambda^2 n^2}\|\mathbf{a}\|^2 + \frac{1}{\lambda^2 n^2}\|\mathbf{b}\|^2, \quad (49)$$
$$\|\mathbf{w}_*\|^2 = \frac{1}{\lambda^2 n^2}\|\mathbf{c}\|^2 + \frac{1}{\lambda^2 n^2}\|\mathbf{d}\|^2, \quad (50)$$
$$\|\mathbf{d}\| = \lambda n\|[\mathbf{w}_*]_{\overline{\mathcal{S}}}\| \stackrel{(15)}{\leq} \lambda n\rho\|\mathbf{w}_*\|. \quad (51)$$

Then, we bound each term in (18) as follows.

$$[(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*) \circ \mathbf{y}]^\top \widehat{G}[(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*) \circ \mathbf{y}] = \mathbf{a}^\top A_{\mathcal{S}} A_{\mathcal{S}}^\top \mathbf{a} + \mathbf{b}^\top A_{\overline{\mathcal{S}}} A_{\overline{\mathcal{S}}}^\top \mathbf{b} + \mathbf{a}^\top A_{\mathcal{S}} A_{\overline{\mathcal{S}}}^\top \mathbf{b} + \mathbf{b}^\top A_{\overline{\mathcal{S}}} A_{\mathcal{S}}^\top \mathbf{a} \geq \|\mathbf{a}\|^2(1-\epsilon) - 2\|\mathbf{a}\|\|\mathbf{b}\|\tau. \quad (52)$$

$$\frac{\lambda n}{\gamma}\|\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*\|^2 \stackrel{y_i \in \pm 1}{=} \frac{\lambda n}{\gamma}\|(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*) \circ \mathbf{y}\|^2 \geq \frac{\lambda n}{\gamma}\frac{\left\|X_{\overline{\mathcal{S}}}[(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*) \circ \mathbf{y}]\right\|^2}{\|X_{\overline{\mathcal{S}}}^\top X_{\overline{\mathcal{S}}}\|_2} \stackrel{(12)}{=} \frac{\lambda n}{\gamma\,\eta}\|\mathbf{b}\|^2. \quad (53)$$

$$[(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*) \circ \mathbf{y}]^\top (G - \widehat{G})(\boldsymbol{\alpha}_* \circ \mathbf{y}) = \mathbf{a}^\top (I - A_{\mathcal{S}} A_{\mathcal{S}}^\top)\mathbf{c} + \mathbf{b}^\top (I - A_{\overline{\mathcal{S}}} A_{\overline{\mathcal{S}}}^\top)\mathbf{d} - \mathbf{a}^\top A_{\mathcal{S}} A_{\overline{\mathcal{S}}}^\top \mathbf{d} - \mathbf{b}^\top A_{\overline{\mathcal{S}}} A_{\mathcal{S}}^\top \mathbf{c} \leq \|\mathbf{a}\|\|\mathbf{c}\|\epsilon + \|\mathbf{b}\|\|\mathbf{d}\|\upsilon + \|\mathbf{a}\|\|\mathbf{d}\|\tau + \|\mathbf{b}\|\|\mathbf{c}\|\tau. \quad (54)$$

From (18), (52), (53), and (54), we have

$$(1-\epsilon)\|\mathbf{a}\|^2 - 2\tau\|\mathbf{a}\|\|\mathbf{b}\| + \frac{\lambda n}{\gamma\,\eta}\|\mathbf{b}\|^2 \leq \|\mathbf{a}\|\|\mathbf{c}\|\epsilon + \|\mathbf{b}\|\|\mathbf{d}\|\upsilon + \|\mathbf{a}\|\|\mathbf{d}\|\tau + \|\mathbf{b}\|\|\mathbf{c}\|\tau.$$

Following the same analysis as that for Theorem 4, we can show both (45) and (46) are true. As a result, we have

$$\frac{1-\epsilon}{2}\|\mathbf{a}\|^2 + \frac{\lambda n}{2\gamma\,\eta}\|\mathbf{b}\|^2 \leq \|\mathbf{a}\|\|\mathbf{c}\|\epsilon + \|\mathbf{b}\|\|\mathbf{d}\|\upsilon + \|\mathbf{a}\|\|\mathbf{d}\|\tau + \|\mathbf{b}\|\|\mathbf{c}\|\tau \leq \frac{1-\epsilon}{8}\|\mathbf{a}\|^2 + \frac{2\epsilon^2}{1-\epsilon}\|\mathbf{c}\|^2 + \frac{\lambda n}{8\gamma\,\eta}\|\mathbf{b}\|^2 + \frac{2\gamma\,\eta\upsilon^2}{\lambda n}\|\mathbf{d}\|^2 + \frac{1-\epsilon}{8}\|\mathbf{a}\|^2 + \frac{2\tau^2}{1-\epsilon}\|\mathbf{d}\|^2 + \frac{\lambda n}{8\gamma\,\eta}\|\mathbf{b}\|^2 + \frac{2\gamma\,\eta\tau^2}{\lambda n}\|\mathbf{c}\|^2$$

which implies

$$\frac{1-\epsilon}{4}\|\mathbf{a}\|^2 + \frac{\lambda n}{4\gamma\,\eta}\|\mathbf{b}\|^2 \leq \left(\frac{2\epsilon^2}{1-\epsilon} + \frac{2\gamma\,\eta\tau^2}{\lambda n}\right)\|\mathbf{c}\|^2 + \left(\frac{2\gamma\,\eta\upsilon^2}{\lambda n} + \frac{2\tau^2}{1-\epsilon}\right)\|\mathbf{d}\|^2 \stackrel{(50,51)}{\leq} \left(\frac{2\epsilon^2}{1-\epsilon} + \frac{2\gamma\,\eta\tau^2}{\lambda n}\right)\lambda^2 n^2\|\mathbf{w}_*\|^2 + \left(\frac{2\gamma\,\eta\upsilon^2}{\lambda n} + \frac{2\tau^2}{1-\epsilon}\right)\lambda^2 n^2\rho^2\|\mathbf{w}_*\|^2.$$

Then, we can upper bound $\|\widetilde{\mathbf{w}} - \mathbf{w}_*\|^2$ by

$$\|\widetilde{\mathbf{w}} - \mathbf{w}_*\|^2 \overset{(49)}{\leq} 8\left(\frac{1}{1-\epsilon} + \frac{\gamma\,\eta}{\lambda n}\right) \cdot \left(\frac{\epsilon^2}{1-\epsilon} + \frac{\gamma\,\eta\tau^2}{\lambda n} + \frac{\tau^2\rho^2}{1-\epsilon} + \frac{\gamma\,\eta\upsilon^2\rho^2}{\lambda n}\right)\|\mathbf{w}_*\|^2$$

leading to the third inequality in Theorem 6.

Similar to Lemma 5, we have the following lemma to bound υ.

Lemma 9: Let $\delta \in (0, 1)$ be the failure probability. With a probability at least $1 - \delta$, we have

$$\upsilon = \left\| A_{\overline{\mathcal{S}}} A_{\overline{\mathcal{S}}}^\top - I \right\|_2 \leq \frac{4(d-s+1)}{m} \log \frac{2(d-s)}{\delta},$$

provided the condition in (17) holds.

VI. An Iterative Extension of Dual Random Projection

From the results in Section IV, we observe that the recovery error of dual random projection enjoys a relative error bound. This observation motives us to develop an iterative extension of dual random projection which is able to reduce the recovery error exponentially.

A. The Algorithm

The main idea stems from the fact that if $\|\widetilde{\mathbf{w}} - \mathbf{w}_*\| \leq \varepsilon \|\mathbf{w}_*\|$ with a small $\varepsilon \leq 1$, we can apply the same dual random projection algorithm to recover $\Delta\mathbf{w} = \mathbf{w}_* - \widetilde{\mathbf{w}}$, which will result in a recovery error of $\varepsilon\|\Delta\mathbf{w}\| \leq \varepsilon^2\|\mathbf{w}_*\|$. If we repeat the above process for T iterations, we should be able to obtain a solution with a recovery error of $\varepsilon^T\|\mathbf{w}_*\|$. This simple intuition leads to the iterative method shown in Table II. At the t-th iteration, given the recovered solution $\widetilde{\mathbf{w}}^{t-1}$ obtained from the previous iteration, we solve the optimization problem in (55) that is designed to recover $\mathbf{w}_* - \widetilde{\mathbf{w}}^{t-1}$.

It is important to note that although the iterative algorithm consists of multiple iterations, the random projection of the data matrix is only computed once before the start of the iterations. This important feature makes the iterative algorithm computationally attractive as calculating random projections of a large data matrix is computationally expensive and has been the subject of many studies, see [22], [31], [32]. We also note that the iterative algorithm in Table II is related to the epoch gradient descent algorithm [33] for stochastic optimization in the sense that the solution obtained from the previous iteration serves as the starting point to the optimization problem at the current iteration. Unlike the epoch gradient algorithm, we do not shrink the domain size over the iterations.

B. The Derivation

In this subsection, we provide the derivation of the iterative algorithm given in Table II. At the t-th iteration, we consider the following optimization problem:

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{\lambda}{2}\|\mathbf{w} + \widetilde{\mathbf{w}}^{t-1}\|^2 + \frac{1}{n}\sum_{i=1}^n \ell\left(y_i(\mathbf{w} + \widetilde{\mathbf{w}}^{t-1})^\top \mathbf{x}_i\right), \quad (56)$$

TABLE II
An Iterative Extension of Dual Random Projection

Input: input patterns $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d\times n}$, binary class assignment $\mathbf{y} \in \{-1, +1\}^n$, the number of random projections m, and the number of iterations T
Output: the recovered solution $\widetilde{\mathbf{w}}^T$

1: Sample a Gaussian random matrix $A \in \mathbb{R}^{d\times m}$ with $A_{i,j} \sim \mathcal{N}(0, 1/m)$
2: Compute the projected data matrix as $\widehat{X} = [\widehat{\mathbf{x}}_1, \ldots, \widehat{\mathbf{x}}_n] = A^\top X$
3: Initialize $\widetilde{\mathbf{w}}^0 = \mathbf{0}$
4: **for** $t = 1, \ldots, T$ **do**
5: Compute $\mathbf{z}_*^t \in \mathbb{R}^m$ by solving the following optimization problem

$$\min_{\mathbf{z}\in\mathbb{R}^m} \frac{\lambda}{2}\left\|\mathbf{z} + A^\top\widetilde{\mathbf{w}}^{t-1}\right\|^2 + \frac{1}{n}\sum_{i=1}^n \ell\left(y_i\mathbf{z}^\top\widehat{\mathbf{x}}_i + y_i\mathbf{x}_i^\top\widetilde{\mathbf{w}}^{t-1}\right) \quad (55)$$

6: Construct the dual solution $\widehat{\boldsymbol{\alpha}}_*^t \in \mathbb{R}^n$ using

$$[\widehat{\boldsymbol{\alpha}}_*^t]_i = \ell'\left(y_i\widehat{\mathbf{x}}_i^\top\mathbf{z}_*^t + y_i[\widetilde{\mathbf{w}}^{t-1}]^\top\mathbf{x}_i\right),\ i = 1, \ldots, n$$

7: Update the solution by

$$\widetilde{\mathbf{w}}^t = -\frac{1}{\lambda n} X(\widehat{\boldsymbol{\alpha}}_*^t \circ \mathbf{y})$$

8: **end for**

where $\widetilde{\mathbf{w}}^{t-1}$ is the solution obtained from the $t-1$-th iteration. It is straightforward to show that $\Delta_*^t = \mathbf{w}_* - \widetilde{\mathbf{w}}^{t-1}$ is the optimal solution to (56). Our goal is to apply the dual random projection approach to recover Δ_*^t by $\widetilde{\Delta}^t$.

In order to apply dual random projection in Table I to solve (56), we need to write the optimization problem in the same form as (1). To this end, we first note that $\widetilde{\mathbf{w}}^{t-1}$ lies in the subspace spanned by $\mathbf{x}_1, \ldots, \mathbf{x}_n$, and therefore we can write $\widetilde{\mathbf{w}}^{t-1}$ as

$$\widetilde{\mathbf{w}}^{t-1} = -\frac{1}{\lambda n} X(\widehat{\boldsymbol{\alpha}}_*^{t-1} \circ \mathbf{y}) = -\frac{1}{\lambda n}\sum_{i=1}^n [\widehat{\boldsymbol{\alpha}}_*^{t-1}]_i y_i \mathbf{x}_i.$$

Then, the objective function in (56) can be written as

$$\begin{aligned}
&\frac{\lambda}{2}\|\widetilde{\mathbf{w}}^{t-1}\|^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2 + \lambda\mathbf{w}^\top\widetilde{\mathbf{w}}^{t-1} + \frac{1}{n}\sum_{i=1}^n \ell\left(y_i\mathbf{w}^\top\mathbf{x}_i + y_i\mathbf{x}_i^\top\widetilde{\mathbf{w}}^{t-1}\right)\\
&= \frac{\lambda}{2}\|\widetilde{\mathbf{w}}^{t-1}\|^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2 + \frac{1}{n}\sum_{i=1}^n \ell\left(y_i\mathbf{w}^\top\mathbf{x}_i + y_i\mathbf{x}_i^\top\widetilde{\mathbf{w}}^{t-1}\right) - [\widehat{\boldsymbol{\alpha}}_*^{t-1}]_i y_i\mathbf{w}^\top\mathbf{x}_i\\
&= \frac{\lambda}{2}\|\widetilde{\mathbf{w}}^{t-1}\|^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2 + \frac{1}{n}\sum_{i=1}^n \ell_i^t\left(y_i\mathbf{w}^\top\mathbf{x}_i\right),
\end{aligned}$$

where the new loss function $\ell_i^t(z), i = 1, \ldots, n$ is defined as

$$\ell_i^t(z) = \ell\left(z + y_i\mathbf{x}_i^\top\widetilde{\mathbf{w}}^{t-1}\right) - [\widehat{\boldsymbol{\alpha}}_*^{t-1}]_i z. \quad (57)$$

Therefore, Δ_*^t is the solution to the following problem:

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{\lambda}{2}\|\mathbf{w}\|^2 + \frac{1}{n}\sum_{i=1}^n \ell_i^t\left(y_i\mathbf{w}^\top\mathbf{x}_i\right). \tag{58}$$

To apply the dual random projection approach to recovering Δ_*^t, we solve the following low-dimensional optimization problem:

$$\min_{\mathbf{z}\in\mathbb{R}^m} \frac{\lambda}{2}\|\mathbf{z}\|^2 + \frac{1}{n}\sum_{i=1}^n \ell_i^t\left(y_i\mathbf{z}^\top\widehat{\mathbf{x}}_i\right), \tag{59}$$

where $\widehat{\mathbf{x}}_i \in \mathbb{R}^m$ is the low-dimensional representation for example $\mathbf{x}_i \in \mathbb{R}^d$. The following derivation signifies that the above problem is equivalent to the problem in (55).

$$\begin{aligned}
&\frac{\lambda}{2}\|\mathbf{z}\|^2 + \frac{1}{n}\sum_{i=1}^n \ell_i^t\left(y_i\mathbf{z}^\top\widehat{\mathbf{x}}_i\right)\\
&= \frac{\lambda}{2}\|\mathbf{z}\|^2 + \frac{1}{n}\sum_{i=1}^n \ell\left(y_i\mathbf{z}^\top\widehat{\mathbf{x}}_i + y_i\mathbf{x}_i^\top\widetilde{\mathbf{w}}^{t-1}\right) - [\widehat{\boldsymbol{\alpha}}_*^{t-1}]_i y_i\mathbf{z}^\top\widehat{\mathbf{x}}_i\\
&= \frac{\lambda}{2}\|\mathbf{z}\|^2 + \lambda\mathbf{z}^\top(A^\top\widetilde{\mathbf{w}}^{t-1}) + \frac{1}{n}\sum_{i=1}^n \ell\left(y_i\mathbf{z}^\top\widehat{\mathbf{x}}_i + y_i\mathbf{x}_i^\top\widetilde{\mathbf{w}}^{t-1}\right)\\
&= \frac{\lambda}{2}\left\|\mathbf{z} + A^\top\widetilde{\mathbf{w}}^{t-1}\right\|^2 + \frac{1}{n}\sum_{i=1}^n \ell\left(y_i\mathbf{z}^\top\widehat{\mathbf{x}}_i + y_i\mathbf{x}_i^\top\widetilde{\mathbf{w}}^{t-1}\right)\\
&\quad -\frac{\lambda}{2}\left\|A^\top\widetilde{\mathbf{w}}^{t-1}\right\|^2,
\end{aligned}$$

where in the third line we use the fact that $\widehat{\mathbf{x}}_i = A^\top\mathbf{x}_i$ and $\widetilde{\mathbf{w}}^{t-1} = -\sum_i[\widehat{\boldsymbol{\alpha}}_*^{t-1}]_i y_i\mathbf{x}_i/(\lambda n)$. Given the optimal solution $\mathbf{z}_*^t$ to the above problem, we can recover Δ_*^t by

$$\widetilde{\Delta}^t = -\frac{1}{\lambda n}X(\widehat{\boldsymbol{\beta}}_*^t \circ \mathbf{y}), \tag{60}$$

where $\widehat{\boldsymbol{\beta}}_*^t$ is computed by

$$\begin{aligned}
[\widehat{\boldsymbol{\beta}}_*^t]_i &= \nabla\ell_i^t\left(y_i\widehat{\mathbf{x}}_i^\top\mathbf{z}_*^t\right)\\
&\overset{(57)}{=} \ell'\left(y_i\widehat{\mathbf{x}}_i^\top\mathbf{z}_*^t + y_i\mathbf{x}_i^\top\widetilde{\mathbf{w}}^{t-1}\right) - [\widehat{\boldsymbol{\alpha}}_*^{t-1}]_i, \quad i = 1,\ldots,n.
\end{aligned}$$

The updated solution $\widetilde{\mathbf{w}}^t$ is computed by

$$\begin{aligned}
\widetilde{\mathbf{w}}^t &= \widetilde{\mathbf{w}}^{t-1} + \widetilde{\Delta}^t\\
&= -\frac{1}{\lambda n}X\left[\left(\widehat{\boldsymbol{\alpha}}_*^{t-1} + \widehat{\boldsymbol{\beta}}_*^t\right)\circ\mathbf{y}\right] = -\frac{1}{\lambda n}X(\widehat{\boldsymbol{\alpha}}_*^t\circ\mathbf{y}),
\end{aligned}$$

where

$$\begin{aligned}
[\widehat{\boldsymbol{\alpha}}_*^t]_i &= [\widehat{\boldsymbol{\alpha}}_*^{t-1}]_i + [\widehat{\boldsymbol{\beta}}_*^t]_i\\
&= \ell'(y_i\widehat{\mathbf{x}}_i^\top\mathbf{z}_*^t + y_i\mathbf{x}_i^\top\widetilde{\mathbf{w}}^{t-1}), i = 1,\ldots,n.
\end{aligned}$$

C. The Analysis

In each iteration of the iterative algorithm, dual random projection is used to recover the optimal solution $\Delta_*^t = \mathbf{w}_* - \widetilde{\mathbf{w}}^{t-1}$ of (58). To analyze the recovery error of the final solution, we just need to apply our previous analysis to bound the recovery error in each iteration. And the recovery error of the final solution follows directly.

Theorem 7: Assume that X is low-rank with rank r. Let $\mathbf{w}_*$ be the optimal solution to (1) and $\widetilde{\mathbf{w}}^T$ be the solution recovered by the iterative algorithm. Suppose

$$m \geq 32(r+1)\log\frac{2r}{\delta}.$$

Then, with a probability at least $1-\delta$, we have

$$\|\widetilde{\mathbf{w}}^T - \mathbf{w}_*\| \leq \left(\frac{\epsilon}{1-\epsilon}\right)^T\|\mathbf{w}_*\|,$$

where

$$\epsilon = 2\sqrt{\frac{2(r+1)}{m}\log\frac{2r}{\delta}} \leq \frac{1}{2}.$$

Proof: Suppose we can show that

$$\|\widetilde{\Delta}^t - \Delta_*^t\| \leq \varepsilon_t\|\Delta_*^t\|, \quad t = 1,\ldots,T. \tag{61}$$

From the fact that $\widetilde{\mathbf{w}}^t = \widetilde{\mathbf{w}}^{t-1} + \widetilde{\Delta}^t$ and $\Delta_*^t = \mathbf{w}_* - \widetilde{\mathbf{w}}^{t-1}$, we have

$$\|\widetilde{\mathbf{w}}^t - \mathbf{w}_*\| = \|\widetilde{\Delta}^t - \Delta_*^t\| \overset{(61)}{\leq} \varepsilon_t\|\Delta_*^t\| = \varepsilon_t\|\widetilde{\mathbf{w}}^{t-1} - \mathbf{w}_*\|.$$

Repeating the above inequality for $t = 1,\ldots,T$, the recovery error of the last solution $\widetilde{\mathbf{w}}^T$ is upper bounded by

$$\|\widetilde{\mathbf{w}}^T - \mathbf{w}_*\| \leq \prod_{t=1}^T \varepsilon_t\|\widetilde{\mathbf{w}}^0 - \mathbf{w}_*\| = \prod_{t=1}^T \varepsilon_t\|\mathbf{w}_*\|,$$

where we assume $\widetilde{\mathbf{w}}^0 = \mathbf{0}$.

In the following, we will decide the value of ε_t in (61) under the assumption that X is low-rank. The analysis is almost the same as that for Theorem 1. The only difference is that in the iterative algorithm, the loss functions $\ell_i^t(\cdot)$ depends on the random matrix A. However, it turns out that this dependency is not problematic, because our analysis only needs the matrix concentration inequality in Lemma 2.

Let $\bar{\ell}_i^t(\cdot)$ be the convex conjugate of $\ell_i^t(\cdot)$, i.e.,

$$\ell_i^t(z) = \max_{\alpha\in\Omega_i^t} \alpha z - \bar{\ell}_i^t(\alpha),$$

where Ω_i^t is the domain of the dual variable. The dual problem of (58) is given by

$$\max_{\alpha_i\in\Omega_i^t} -\sum_{i=1}^n \bar{\ell}_i^t(\alpha_i) - \frac{1}{2\lambda n}(\boldsymbol{\alpha}\circ\mathbf{y})^\top X^\top X(\boldsymbol{\alpha}\circ\mathbf{y}), \tag{62}$$

and the dual problem of (59) is

$$\max_{\alpha_i\in\Omega_i^t} -\sum_{i=1}^n \bar{\ell}_i^t(\alpha_i) - \frac{1}{2\lambda n}(\boldsymbol{\alpha}\circ\mathbf{y})^\top X^\top AA^\top X(\boldsymbol{\alpha}\circ\mathbf{y}). \tag{63}$$

Following exactly the same analysis of Lemma 1, we have the following lemma to bound the optimal dual solutions.

Lemma 10: Let $\boldsymbol{\beta}_*^t \in \mathbb{R}^n$ and $\widehat{\boldsymbol{\beta}}_*^t \in \mathbb{R}^n$ be the optimal dual solutions to (62) and (63), respectively. Then, we have

$$\begin{aligned}
&[(\widehat{\boldsymbol{\beta}}_*^t - \boldsymbol{\beta}_*^t)\circ\mathbf{y}]^\top\widehat{G}[(\widehat{\boldsymbol{\beta}}_*^t - \boldsymbol{\beta}_*^t)\circ\mathbf{y}]\\
&\qquad \leq [(\widehat{\boldsymbol{\beta}}_*^t - \boldsymbol{\beta}_*^t)\circ\mathbf{y}]^\top(G - \widehat{G})(\boldsymbol{\beta}_*^t\circ\mathbf{y})
\end{aligned}$$

where $G = X^\top X$ and $\widehat{G} = X^\top AA^\top X$.

Following the notations in Theorem 1, we introduce the SVD of X, and write X, G, and $\widehat{G}$ as

$$X = U\Sigma V^\top, \quad G = V\Sigma U^\top U\Sigma V^\top = V\Sigma^2 V^\top,$$
$$\widehat{G} = V\Sigma U^\top AA^\top U\Sigma V^\top = V\Sigma BB^\top \Sigma V^\top,$$

where

$$B = U^\top A \in \mathbb{R}^{r\times m}.$$

To simplify the notation, we define

$$\mathbf{a} = \Sigma V^\top[(\widehat{\boldsymbol{\beta}}_*^t - \boldsymbol{\beta}_*^t)\circ \mathbf{y}], \ \mathbf{c} = \Sigma V^\top(\boldsymbol{\beta}_*^t \circ \mathbf{y}),$$
$$\epsilon = \|BB^\top - I\|_2.$$

Recall that $\Delta_*^t = -\frac{1}{\lambda n}X(\boldsymbol{\beta}_*^t \circ \mathbf{y})$ is the optimal solution to (58), and $\widetilde{\Delta}^t$ in (60) is the recovered solution. Since U is an orthogonal matrix, we have

$$\|\widetilde{\Delta}^t - \Delta_*^t\| = \left\|\frac{1}{\lambda n}X[(\widehat{\boldsymbol{\beta}}_*^t - \boldsymbol{\beta}_*^t)\circ \mathbf{y}]\right\| = \frac{1}{\lambda n}\|\mathbf{a}\|, \quad (64)$$

$$\|\Delta_*^t\| = \left\|-\frac{1}{\lambda n}X(\boldsymbol{\beta}_*^t \circ \mathbf{y})\right\| = \frac{1}{\lambda n}\|\mathbf{c}\|. \quad (65)$$

From Lemma 10, we have

$$\mathbf{a}^\top BB^\top \mathbf{a} \le \mathbf{a}^\top\left(I - BB^\top\right)\mathbf{c},$$

which implies

$$\|\mathbf{a}\|^2(1-\epsilon) \le \epsilon\|\mathbf{a}\|\|\mathbf{c}\| \Rightarrow \|\mathbf{a}\|(1-\epsilon) \le \epsilon\|\mathbf{c}\|. \quad (66)$$

From (64), (65), and (66), we have $\varepsilon_t = \frac{\epsilon}{1-\epsilon}$, $t = 1, \ldots, T$. And thus

$$\|\widetilde{\mathbf{w}}^T - \mathbf{w}_*\| \le \left(\frac{\epsilon}{1-\epsilon}\right)^T \|\mathbf{w}_*\|.$$

We complete the proof by applying Lemma 2 *once* to bound ϵ. ∎

Theorem 7 implies that we can recover the optimal solution with a relative error α, i.e., $\|\mathbf{w}_* - \widetilde{\mathbf{w}}^T\| \le \alpha\|\mathbf{w}_*\|$, by using $\lceil \log_{(1-\epsilon)/\epsilon} 1/\alpha \rceil$ iterations. We finally note that it is unclear if the iterative algorithm can achieve similar error reduction for more general cases as they require further assumption on $\mathbf{w}_*$, which may not hold for the intermediate steps of the iterative algorithm.

VII. Complexity Analysis and Empirical Study

In this section, we first analyze the numerical complexities of the baseline algorithm that optimizes the original problem, the proposed dual random projection algorithm and its iterative extension. Then, we conduct experiments to verify our theoretical claims.

A. Numerical Complexity

Suppose our goal is to find a solution $\bar{\mathbf{w}}$ such that $\|\bar{\mathbf{w}} - \mathbf{w}_*\| \le \alpha\|\mathbf{w}_*\|$, where $\alpha > 0$ is a given parameter. We assume the ℓ_2 norm of each data point is bounded, that is, $\|\mathbf{x}_i\| = O(1)$, $i = 1, \ldots, n$. Since random projection preserves the ℓ_2 norm [20]–[22], with a high probability, we also have $\|\widehat{\mathbf{x}}_i\| = O(1)$, $i = 1, \ldots, n$. Before presenting the analysis, we need to decide the optimization algorithm used to solve (1), (3), and (55). Since we assume both n and d are very large, first-order optimization methods, that relies on the gradient of the objective function, become a natural choice. Notice that the optimization problems considered in this paper are both smooth and strongly convex. According to the convex optimization theory [29], the number of iterations required to find a solution with an error ϵ is on the order of $\sqrt{\kappa}\log 1/\epsilon$, where κ is the condition number.[1]

1) The Baseline Algorithm: Let $L_h \ge \mu_h \ge \lambda$ be the moduli of smoothness and strong convexity of the high-dimensional optimization problem in (1), and $\kappa_h = L_h/\mu_h$ be the condition number. If we are able to find a solution $\bar{\mathbf{w}}$ to (1) with an error $\frac{1}{2}\mu_h\alpha^2\|\mathbf{w}_*\|^2$, then due to the strong convexity, we have $\|\bar{\mathbf{w}} - \mathbf{w}_*\| \le \alpha\|\mathbf{w}_*\|$. Based on the previous discussion, we know that the number of iterations is on the order of $\sqrt{\kappa_h}\log\frac{1}{\mu_h\alpha\|\mathbf{w}_*\|}$. The cost in each iteration is dominated by the evaluation of the gradient, whose complexity is $O(nd)$. Thus, the overall numerical complexity of the baseline algorithm is

$$O\left(nd\sqrt{\kappa_h}\left(\log\frac{1}{\mu_h} + \log\frac{1}{\|\mathbf{w}_*\|} + \log\frac{1}{\alpha}\right)\right),$$

which can be simplified to

$$O\left(nd\sqrt{\kappa_h}\log\frac{1}{\alpha}\right)$$

under appropriate conditions.

2) Dual Random Projection: It is easy to verify that the numerical complexity of Steps 1, 2, 4 and 5 in Table I is $O(ndm)$. In the following, we will discuss the numerical complexity of Step 3, as well as the order of m.

Let $L_l \ge \mu_l \ge \lambda$ be the moduli of smoothness and strong convexity of the low-dimensional optimization problem in (3), and $\kappa_l = L_l/\mu_l$ be the condition number. Let $\hat{\mathbf{z}}$ be the solution that we obtained by solving (3) numerically. Then, based on the procedure in Table I, we will return a solution $\bar{\mathbf{w}} = -\frac{1}{\lambda n}X(\widehat{\boldsymbol{\alpha}}\circ\mathbf{y})$, where $[\widehat{\boldsymbol{\alpha}}]_i = \ell'\left(y_i\widehat{\mathbf{x}}_i^\top\hat{\mathbf{z}}\right)$, $i = 1, \ldots, n$. The difference between $\bar{\mathbf{w}}$ and $\mathbf{w}_*$ can be decomposed into two parts corresponding to the optimization error and the recovery error, respectively, that is,

$$\|\bar{\mathbf{w}} - \mathbf{w}_*\| \le \underbrace{\|\bar{\mathbf{w}} - \widetilde{\mathbf{w}}\|}_{\text{Optimization Error}} + \underbrace{\|\widetilde{\mathbf{w}} - \mathbf{w}_*\|}_{\text{Recovery Error}}.$$

To ensure that $\|\bar{\mathbf{w}} - \mathbf{w}_*\| \le \alpha\|\mathbf{w}_*\|$, a sufficient condition is to ensure both $\|\bar{\mathbf{w}} - \widetilde{\mathbf{w}}\|$ and $\|\widetilde{\mathbf{w}} - \mathbf{w}_*\|$ are smaller than $\frac{\alpha}{2}\|\mathbf{w}_*\|$.

Let's first consider bounding $\|\bar{\mathbf{w}} - \widetilde{\mathbf{w}}\|$. After some algebraic manipulations, we obtain

$$\|\bar{\mathbf{w}} - \widetilde{\mathbf{w}}\| = O\left(\frac{\gamma\|\hat{\mathbf{z}} - \mathbf{z}_*\|}{\lambda}\right), \quad (67)$$

whose derivation is provided in Appendix F. Thus, to ensure $\|\bar{\mathbf{w}} - \widetilde{\mathbf{w}}\| \le \frac{\alpha}{2}\|\mathbf{w}_*\|$, it is sufficient to find a solution $\hat{\mathbf{z}}$

[1] Let's review some basic concepts in convex optimization [25], [29]. Suppose we want to minimize a convex function $f(\cdot)$ over a convex domain $\mathcal{D}$. A solution $\widehat{\mathbf{x}}$ with an error ϵ means $f(\widehat{\mathbf{x}}) - \min_{\mathbf{x}\in\mathcal{D}} f(\mathbf{x}) \le \epsilon$. When the function $f(\cdot)$ is μ-strongly convex, we further have $\frac{\mu}{2}\|\widehat{\mathbf{x}} - \mathbf{x}_*\|^2 \le \epsilon$, where $\mathbf{x}_* = \arg\min_{\mathbf{x}\in\mathcal{D}} f(\mathbf{x})$. For a function that is μ-strongly convex and L-smooth, the condition number κ is defined as L/μ.

to (3), such that $\|\hat{\mathbf{z}} - \mathbf{z}_*\| = O\left(\frac{\alpha\|\mathbf{w}_*\|\lambda}{\gamma}\right)$. Following the same argument in Section VII-A.1, we know that the overall numerical complexity of solving (3) is

$$O\left(nm\sqrt{\kappa_l}\left(\log\frac{1}{\mu_l} + \log\frac{\gamma}{\lambda} + \log\frac{1}{\|\mathbf{w}_*\|} + \log\frac{1}{\alpha}\right)\right).$$

Next, we consider bounding the recovery error $\|\widetilde{\mathbf{w}} - \mathbf{w}_*\|$, from which we decide the order of m. Recall that in Section IV, we provided four theorems to bound the recovery error in different scenarios. In the following, we take the the case that X is low-rank as an example. According to Theorem 1, to ensure $\|\widetilde{\mathbf{w}} - \mathbf{w}_*\| \leq \frac{\alpha}{2}\|\mathbf{w}_*\|$, it is sufficient to set $m = O\left(\frac{r\log r}{\alpha^2}\right)$.

In summary, the numerical complexity of dual random projection is

$$O\left(\frac{ndr\log r}{\alpha^2} + \frac{n\sqrt{\kappa_l}r\log r}{\alpha^2}\left(\log\frac{1}{\mu_l} + \log\frac{\gamma}{\lambda} + \log\frac{1}{\|\mathbf{w}_*\|} + \log\frac{1}{\alpha}\right)\right),$$

which can be simplified to

$$O\left(\frac{ndr\log r}{\alpha^2} + \frac{n\sqrt{\kappa_l}r\log r}{\alpha^2}\log\frac{1}{\alpha}\right)$$

under appropriate conditions.

3) The Iterative Extension: It is easy to verify that the numerical complexity of Steps 1 and 2 in Table II is $O(ndm)$, and the numerical complexity of Steps 6 and 7 is $O(ndT)$. In the following, we will discuss the numerical complexity of Step 5, as well as the order of m and T.

Recall that the linear convergence in Theorem 7 comes from the fact that there are a recovery error $\frac{\epsilon}{1-\epsilon}\|\Delta_*^t\|$ and no optimization error in the t-th iteration. Alternatively, if both the recovery error and the optimization error in the t-th iteration are bounded by $\frac{1}{2}\left(\frac{\epsilon}{1-\epsilon}\right)^t\|\mathbf{w}_*\|$, we can obtain a similar linear convergence.[2] Thus, we still have $T = \lceil\log_{(1-\epsilon)/\epsilon} 1/\alpha\rceil$ even in the presence of optimization error.

Let $L_l^t \geq \mu_l^t \geq \lambda$ be the moduli of smoothness and strong convexity of the low-dimensional optimization problem in (55), and $\kappa_l^t = L_l^t/\mu_l^t$ be the condition number. Following the same analysis in Section VII-A.2, to ensure the optimization error is upper bounded by $\frac{1}{2}\left(\frac{\epsilon}{1-\epsilon}\right)^t\|\mathbf{w}_*\|$, the overall numerical complexity of solving (55) is

$$O\left(nm\sqrt{\kappa_l^t}\left(\log\frac{1}{\mu_l^t} + \log\frac{\gamma}{\lambda} + \log\frac{1}{\|\mathbf{w}_*\|} + t\log\frac{1-\epsilon}{\epsilon}\right)\right).$$

And based on induction, it is easy to verify that $m = O(\frac{r\log r}{\epsilon^2})$ is sufficient to satisfy the requirement on the recovery error.

By setting ϵ to be a small constant (e.g., $1/3$), we have $m = O(r\log r)$, $T = O(\log 1/\alpha)$, and the numerical complexity of the iterative extension is

$$O\left(nd\left(r\log r + \log\frac{1}{\alpha}\right) + nr\log r\sum_{t=1}^{T}\sqrt{\kappa_l^t}\left(\log\frac{1}{\mu_l^t} + \log\frac{\gamma}{\lambda} + \log\frac{1}{\|\mathbf{w}_*\|} + t\right)\right),$$

which can be simplified to

$$O\left(nd\left(r\log r + \log\frac{1}{\alpha}\right) + nr\log r\log\frac{1}{\alpha}\sum_{t=1}^{T}\sqrt{\kappa_l^t}\right)$$

under appropriate conditions.

4) Comparisons: To simplify the comparisons, we make an conservative assumption that all the condition numbers are on the same order, and are denoted by κ. Then, the numerical complexities of different algorithms are summarized below.

- Baseline: $O(nd\sqrt{\kappa}\log\frac{1}{\alpha})$
- Dual random projection: $O(\frac{ndr\log r}{\alpha^2} + \frac{n\sqrt{\kappa}r\log r}{\alpha^2}\log\frac{1}{\alpha})$
- The iterative extension: $O(nd(r\log r + \log\frac{1}{\alpha}) + n\sqrt{\kappa}r\log r\log^2\frac{1}{\alpha})$

From the above results, we observe that one limitation of dual random projection is that its numerical complexity has a quadratic dependence on $\frac{1}{\alpha}$. As a result, the numerical complexity of dual random projection is small than that of baseline only when α is not too small, that is,

$$\alpha^2 \geq O\left(\frac{r\log r}{\min(\sqrt{\kappa}\log 1/\alpha, d)}\right).$$

On the other hand, the iterative extension is especially suitable for finding a high-precision solution, since its numerical complexity only has a polylogarithmic dependence on $\frac{1}{\alpha}$. Furthermore, when the rank r is small enough, that is,

$$r\log r \leq O\left(\min\left(\sqrt{\kappa}\log\frac{1}{\alpha}, \frac{d}{\log 1/\alpha}\right)\right),$$

the numerical complexity of the iterative extension will be always smaller than that of the baseline.

B. Experimental Results

We perform experiments on a synthetic data set to compare the proposed algorithms with the baseline approach. We generate a data matrix by $X = AB$, where $A \in \mathbb{R}^{d\times r}$ and $B \in \mathbb{R}^{r\times n}$ are two random Gaussian matrices, scale X to ensure the ℓ_2 norm of each data point is bounded by 1, and generate the label by $\mathbf{y} = \text{sign}(X^\top\mathbf{w})$, where $\mathbf{w} \in \mathbb{R}^d$ is a random Gaussian vector. To simulate the case that X is high-dimensional, large-scale, and low-rank, we set $d = 20,000$, $n = 50,000$, and $r = 10$. For each setting of m we repeat the recovery experiment for 10 trials, and report the average result. We choose the logit loss $\ell(x) = \ln(1+\exp(-x))$, and set $\lambda = 1/n$. We implement the optimal first-order algorithm in [35] to solve the optimization problems.

Since the exact value of $\mathbf{w}_*$ is unknown, we take the output of the Baseline algorithm to approximate it.[3] In Fig. 1, we show how the relative recovery errors of Dual Random Projection (DRP) and the naive solution in (6) (i.e., $\|\widetilde{\mathbf{w}} - \mathbf{w}_*\|/\|\mathbf{w}_*\|$ and $\|\widehat{\mathbf{w}} - \mathbf{w}_*\|/\|\mathbf{w}_*\|$) vary with respect to the number of random projections. We observe that with a sufficiently large number of random projections, DRP is able to find an accurate estimator of $\mathbf{w}_*$. On the other hand, the

[2]Strictly speaking, the former one (i.e., the one in Theorem 7) is Q-linear, and the latter one is R-linear [34, Section A.2].

[3]Note that $\mathbf{w}_*$ is in general different from $\mathbf{w}$ since we are interested in minimizing the classification error measured by the logistic regression model.

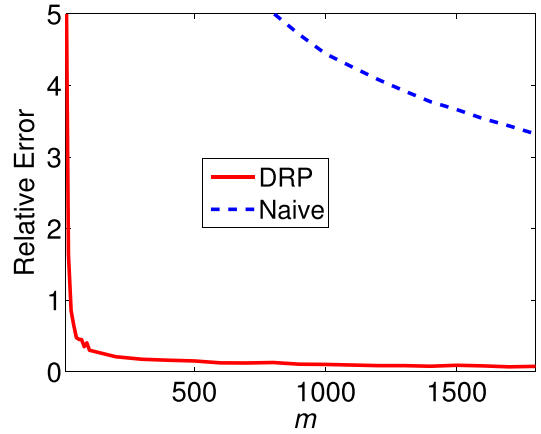

Fig. 1. The relative recovery error versus the number of random projections.

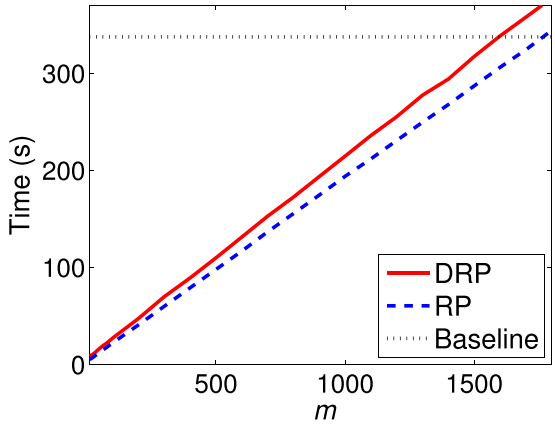

Fig. 2. The running time versus the number of random projections.

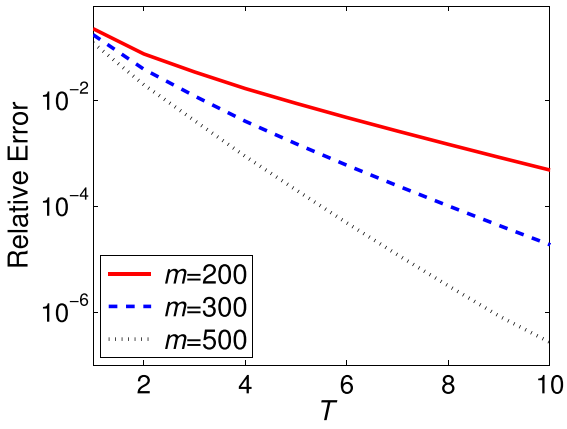

Fig. 3. The relative recovery error of the iterative extension versus the number of iterations.

recovery error of the naive solution is much larger than that of DRP, which is consistent with Proposition 2.

Fig. 2 plots the running times of Baseline, DRP, and the Random Projection (RP) step in DRP. As revealed by our discussion in Section VII-A.4, the running time of DRP is smaller than that of Baseline when m is not too large, or in other words, when the recovery error is not too small. We also observed that the majority of the running time is spent

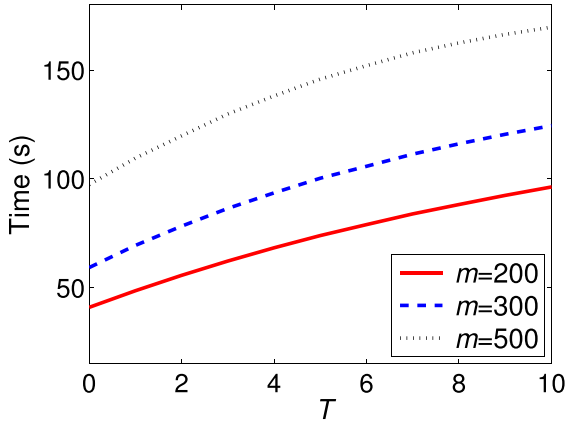

Fig. 4. The running time of the iterative extension versus the number of iterations.

on random projection. Thus, if we utilize the fast random projection techniques [22], [31], [32], the running time of DRP could be reduced dramatically. For instance, the fast random projection algorithm in [31] reduces the dependency of complexity on m from $O(m)$ to $O(\log m)$.

Finally, we provide the relative recovery error and running time of the iterative extension (with $m = 200, 300, 500$) in Fig. 3 and Fig. 4, respectively. As indicated by Theorem 7, the iterative extension is able to reduce the recovery error *exponentially* by using a *small* number of random projections. Furthermore, its running time is shorter than that of Baseline, and increases slowly over iterations. Thus, the iterative extension is preferred if we want to find a high-precision solution.

VIII. Conclusion and Future Work

In this paper, we consider the problem of recovering the optimal solution $\mathbf{w}_*$ to the original high-dimensional optimization problem based on random projection. To this end, we propose to use the dual solution $\widehat{\boldsymbol{\alpha}}_*$ to the low-dimensional optimization problem to recover $\mathbf{w}_*$. Our analysis shows that with a high probability, the solution $\widetilde{\mathbf{w}}$ returned by our proposed method approximates the optimal solution $\mathbf{w}_*$ with a small error, when the data matrix is (approximately) low-rank and/or the optimal solution is (approximately) sparse. We further develop an iterative extension of the basic algorithm, that is able to reduce the recovery error exponentially when the data matrix is low-rank.

One of our future work is to analyze the generalization error of the recovered solution $\widetilde{\mathbf{w}}$. There are three types of errors that affect the generalization error. The first one is the optimization error since the empirical risk is only optimized approximately. The second error is the estimation error which measures the difference between minimizing the empirical risk and minimizing the expected risk. The last error is the approximation error that reflects how closely the optimal classifier can be approximated by a function in a restricted hypothesis space [36]. The proposed Dual Random Projection can guarantee $\widetilde{\mathbf{w}}$ is a good estimator of $\mathbf{w}_*$, implying a small optimization error since the loss function is smooth. The estimation error can be bounded by the uniform convergence concept [37] or the data-dependent complexity estimate such

as the Rademacher complexity [38], [39], and the approximation error is determined by the regularizer and the optimal risk. We will investigate the tradeoff among these three types of errors in the future.

APPENDIX A
PROOF OF PROPOSITION 1

First, if $\boldsymbol{\alpha}_*$ is the optimal dual solution, by replacing $\ell(\cdot)$ in (1) with its conjugate form, the optimal primal solution can be solved by

$$\mathbf{w}_* = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \frac{\lambda}{2}\|\mathbf{w}\|^2 + \frac{1}{n}\sum_{i=1}^{n}[\boldsymbol{\alpha}_*]_i y_i \mathbf{x}_i^\top \mathbf{w}.$$

Setting the gradient with respect to $\mathbf{w}$ to zero, we obtain

$$\mathbf{w}_* = -\frac{1}{\lambda n}\sum_{i=1}^{n}[\boldsymbol{\alpha}_*]_i y_i \mathbf{x}_i = -\frac{1}{\lambda n} X(\boldsymbol{\alpha}_* \circ \mathbf{y}).$$

Second, let's consider how to obtain the dual solution $\boldsymbol{\alpha}_*$ from the primal solution $\mathbf{w}_*$. Note that

$$\ell(y_i \mathbf{x}_i^\top \mathbf{w}_*) = [\boldsymbol{\alpha}_*]_i \left(y_i \mathbf{x}_i^\top \mathbf{w}_*\right) - \ell_*\left([\boldsymbol{\alpha}_*]_i\right).$$

By the Fenchel conjugate theory [40], [41], we have $\boldsymbol{\alpha}_*$ satisfying

$$[\boldsymbol{\alpha}_*]_i = \ell'\left(y_i \mathbf{x}_i^\top \mathbf{w}_*\right),\ i = 1, \ldots, n.$$

APPENDIX B
PROOF OF LEMMA 2

In the proof, we need the recent development in tail bounds for the eigenvalues of a sum of random matrices [42], [43].

Theorem 8: ([43, Th. 1]) Let $\{\boldsymbol{\xi}_j : j = 1, \ldots, n\}$ be i.i.d. samples drawn from a multivariate Gaussian distribution $\mathcal{N}(0, C)$, where $C \in \mathbb{R}^{d\times d}$. Define

$$\widehat{C}_n = \frac{1}{n}\sum_{j=1}^{n}\boldsymbol{\xi}_j \boldsymbol{\xi}_j^\top.$$

Then, for any $\theta \geq 0$

$$\Pr\left\{\left\|\widehat{C}_n - C\right\|_2 \geq \left(\sqrt{\frac{2\theta(k+1)}{n}} + \frac{2\theta k}{n}\right)\|C\|_2\right\} \leq 2d\exp(-\theta),$$

where $k = \mathrm{tr}(C)/\|C\|_2$.

We write $B = \frac{1}{\sqrt{m}}(\mathbf{v}_1, \ldots, \mathbf{v}_m)$, where $\{\mathbf{v}_i \in \mathbb{R}^r\}_{i=1}^m$ are i.i.d. sampled from the Gaussian distribution $\mathcal{N}(0, I)$, and write $BB^\top$ as

$$BB^\top = \frac{1}{m}\sum_{i=1}^{m}\mathbf{v}_i \mathbf{v}_i^\top.$$

Following Theorem 8, we have, with a probability at least $1 - 2r\exp(-\theta)$,

$$\left\|BB^\top - I\right\|_2 \leq \sqrt{\frac{2\theta(r+1)}{m}} + \frac{2\theta r}{m}.$$

By setting $2r\exp(-\theta) = \delta$, we have, with a probability at least $1 - \delta$,

$$\left\|BB^\top - I\right\|_2 \leq \sqrt{\frac{2(r+1)}{m}\log\frac{2r}{\delta}} + \frac{2r}{m}\log\frac{2r}{\delta} \leq 2\sqrt{\frac{2(r+1)}{m}\log\frac{2r}{\delta}},$$

where the last inequality follows from the assumption $m \geq 2(r+1)\log\frac{2r}{\delta}$.

APPENDIX C
PROOF OF LEMMA 3

During the analysis, we need to use the tail bounds for the χ^2 distribution [44] and a noncommutative variant of Bernstein's inequality [30].

Theorem 9 (Tail Bounds for the χ^2 Distribution [44]): Let X be a random variable distributed according to the χ^2 distribution with d degrees of freedom. For any $\epsilon > 0$, we have

$$\Pr\left[X \leq d - 2\sqrt{d\epsilon}\right] \leq \exp(-\epsilon),$$
$$\Pr\left[X \geq d + 2\sqrt{d\epsilon} + 2\epsilon\right] \leq \exp(-\epsilon).$$

Theorem 10 (Noncommutative Bernstein Inequality [30]): Let $X_1, \ldots, X_L$ be independent zero-mean random matrices of dimension $d_1 \times d_2$. Suppose $\rho_k^2 = \max\{\|\mathrm{E}[X_k X_k^*]\|_2, \|\mathrm{E}[X_k^* X_k]\|_2\}$ and $\|X_k\|_2 \leq M$ almost surely for all k. Then for any $\tau \geq 0$,

$$\Pr\left[\left\|\sum_{k=1}^{L} X_k\right\|_2 \geq \tau\right] \leq (d_1 + d_2)\exp\left(-\frac{\tau^2}{2(\sum_{k=1}^{L}\rho_k^2 + M\tau/3)}\right).$$

We write $B_{\bar{r}} = \frac{1}{\sqrt{m}}(\mathbf{u}_1, \ldots, \mathbf{u}_m)$ and $B_r = \frac{1}{\sqrt{m}}(\mathbf{v}_1, \ldots, \mathbf{v}_m)$, where entries in $\mathbf{u}_i \in \mathbb{R}^{d-r}$ and $\mathbf{v}_i \in \mathbb{R}^r$ are sampled from the standard Gaussian distribution $\mathcal{N}(0, 1)$. We thus have

$$B_{\bar{r}} B_r^\top = \frac{1}{m}\sum_{i=1}^{m}\mathbf{u}_i \mathbf{v}_i^\top.$$

To bound the norm of $\mathbf{u}_i$ and $\mathbf{v}_i$, using Theorem 9, we have, with a probability at least $1 - \delta$

$$\|\mathbf{u}_i\|^2 \leq (d-r) + 2\sqrt{(d-r)\log\frac{2m}{\delta}} + 2\log\frac{2m}{\delta} \leq \left(\sqrt{d-r} + \sqrt{2\log\frac{2m}{\delta}}\right)^2,$$

$$\|\mathbf{v}_i\|^2 \leq r + 2\sqrt{r\log\frac{2m}{\delta}} + 2\log\frac{2m}{\delta} \leq \left(\sqrt{r} + \sqrt{2\log\frac{2m}{\delta}}\right)^2, \quad \forall i \in [m].$$

Hence, with a probability at least $1 - \delta$, we have

$$\max_{1\leq i\leq m}\|\mathbf{u}_i \mathbf{v}_i^\top\|_2 = \max_{1\leq i\leq m}\|\mathbf{u}_i\|\|\mathbf{v}_i\| \leq \left(\sqrt{d-r} + \sqrt{2\log\frac{2m}{\delta}}\right)\left(\sqrt{r} + \sqrt{2\log\frac{2m}{\delta}}\right).$$

In addition, we have

$$\left\|\mathrm{E}[\mathbf{u}_i\mathbf{v}_i^\top\mathbf{v}_i\mathbf{u}_i^\top]\right\|_2 = r \overset{(10,11)}{\leq} d-r,$$
$$\left\|\mathrm{E}[\mathbf{v}_i\mathbf{u}_i^\top\mathbf{u}_i\mathbf{v}_i^\top]\right\|_2 = d-r.$$

Following directly from Theorem 10, we have, with a probability at least $1-2\delta$,

$$\begin{aligned}
&\|B_{\bar{r}}B_r^\top\|_2 \\
&\leq \sqrt{\frac{2(d-r)}{m}\log\frac{d}{\delta}} + \frac{2}{3m}\left(\sqrt{d-r}+\sqrt{2\log\frac{2m}{\delta}}\right) \\
&\quad\cdot\left(\sqrt{r}+\sqrt{2\log\frac{2m}{\delta}}\right)\log\frac{d}{\delta} \\
&= \sqrt{\frac{2(d-r)}{m}\log\frac{d}{\delta}} + \frac{1}{3}\sqrt{\frac{2(d-r)}{m}\log\frac{d}{\delta}}\sqrt{\frac{1}{m(d-r)}\log\frac{d}{\delta}} \\
&\quad\cdot\left(\sqrt{2r(d-r)}+\sqrt{4r\log\frac{2m}{\delta}}+\sqrt{4(d-r)\log\frac{2m}{\delta}}\right. \\
&\qquad\left.+2\sqrt{2}\log\frac{2m}{\delta}\right).
\end{aligned}$$

We complete the proof by combining the above inequality with the following ones

$$\begin{aligned}
&m(d-r) \overset{(10)}{\geq} 32(d-r)(r+1)\log\frac{d}{\delta} \geq 2r(d-r)\log\frac{d}{\delta}, \\
&m(d-r) \overset{(10)}{\geq} 4(d-r)\log\frac{2m}{\delta}\log\frac{d}{\delta} \overset{(10,11)}{\geq} 4r\log\frac{2m}{\delta}\log\frac{d}{\delta}, \\
&m(d-r) \overset{(10)}{\geq} 4(d-r)\log\frac{2m}{\delta}\log\frac{d}{\delta}, \\
&m(d-r) \overset{(10)}{\geq} 4(d-r)\log\frac{2m}{\delta}\log\frac{d}{\delta} \overset{(11)}{\geq} 8\log^2\frac{2m}{\delta}\log\frac{d}{\delta}.
\end{aligned}$$

Appendix D
Proof of Lemma 4

We write $B_{\bar{r}} = \frac{1}{\sqrt{m}}(\mathbf{u}_1,\ldots,\mathbf{u}_m)$, where $\{\mathbf{u}_i \in \mathbb{R}^{d-r}\}_{i=1}^m$ are i.i.d. sampled from the Gaussian distribution $\mathcal{N}(0,I)$, and write $B_{\bar{r}}B_{\bar{r}}^\top$ as

$$B_{\bar{r}}B_{\bar{r}}^\top = \frac{1}{m}\sum_{i=1}^m \mathbf{u}_i\mathbf{u}_i^\top.$$

Following Theorem 8, we have, with a probability at least $1-2(d-r)\exp(-\theta)$,

$$\left\|B_{\bar{r}}B_{\bar{r}}^\top - I\right\|_2 \leq \sqrt{\frac{2\theta(d-r+1)}{m}} + \frac{2\theta(d-r)}{m}.$$

By setting $2(d-r)\exp(-\theta)=\delta$, we have, with a probability at least $1-\delta$,

$$\begin{aligned}
&\left\|B_{\bar{r}}B_{\bar{r}}^\top - I\right\|_2 \\
&\leq \sqrt{\frac{2(d-r+1)}{m}\log\frac{2(d-r)}{\delta}} + \frac{2(d-r)}{m}\log\frac{2(d-r)}{\delta} \\
&\leq \frac{4(d-r+1)}{m}\log\frac{2(d-r)}{\delta},
\end{aligned}$$

where the last inequality follows from the facts:

$$2(d-r+1) \overset{(11)}{\geq} m, \text{ and } 2(d-r) \overset{(11)}{\geq} m+2 \geq e.$$

Appendix E
Proof of Lemma 5

From the assumption, we have

$$[\mathbf{w}_*]_{\overline{\mathcal{S}}} = 0. \tag{68}$$

From the expression of $\mathbf{w}_*$ in (7), we have

$$[\mathbf{w}_*]_{\overline{\mathcal{S}}} = -\frac{1}{\lambda n}X_{\overline{\mathcal{S}}}(\boldsymbol{\alpha}_*\circ\mathbf{y}) \overset{(68)}{\Rightarrow} X_{\overline{\mathcal{S}}}(\boldsymbol{\alpha}_*\circ\mathbf{y}) = 0.$$

Appendix F
Derivation of (67)

We have

$$\begin{aligned}
\|\bar{\mathbf{w}}-\widetilde{\mathbf{w}}\| &= \frac{1}{\lambda n}\|X(\widehat{\boldsymbol{\alpha}}\circ\mathbf{y}-\widehat{\boldsymbol{\alpha}}_*\circ\mathbf{y})\| \\
&\leq \frac{1}{\lambda n}\|X\|_2\,\|\widehat{\boldsymbol{\alpha}}\circ\mathbf{y}-\widehat{\boldsymbol{\alpha}}_*\circ\mathbf{y}\| \\
&\overset{y_i\in\pm1}{=} \frac{1}{\lambda n}\|X\|_2\|\widehat{\boldsymbol{\alpha}}-\widehat{\boldsymbol{\alpha}}_*\|.
\end{aligned} \tag{69}$$

To bound $\|\widehat{\boldsymbol{\alpha}}-\widehat{\boldsymbol{\alpha}}_*\|$, we have

$$\begin{aligned}
\|\widehat{\boldsymbol{\alpha}}-\widehat{\boldsymbol{\alpha}}_*\| &= \sqrt{\sum_{i=1}^n\left(\ell'\left(y_i\widehat{\mathbf{x}}_i^\top\hat{\mathbf{z}}\right)-\ell'\left(y_i\widehat{\mathbf{x}}_i^\top\mathbf{z}_*\right)\right)^2} \\
&\leq \gamma\sqrt{\sum_{i=1}^n\left(\widehat{\mathbf{x}}_i^\top\hat{\mathbf{z}}-\widehat{\mathbf{x}}_i^\top\mathbf{z}_*\right)^2} = O(\gamma\,\|\hat{\mathbf{z}}-\mathbf{z}_*\|\sqrt{n}).
\end{aligned} \tag{70}$$

From (69) and (70), we have

$$\|\bar{\mathbf{w}}-\widetilde{\mathbf{w}}\| = O\left(\frac{\gamma\,\|X\|_2\|\hat{\mathbf{z}}-\mathbf{z}_*\|}{\lambda\sqrt{n}}\right) = O\left(\frac{\gamma\,\|\hat{\mathbf{z}}-\mathbf{z}_*\|}{\lambda}\right),$$

where we use the fact that $\|X\|_2 \leq \sqrt{\mathrm{tr}(XX^\top)} = O(\sqrt{n})$.

Acknowledgment

The authors would like to thank the anonymous reviewers and the associate editor for their insightful comments and helpful suggestions.

References

[1] L. Zhang, M. Mahdavi, R. Jin, T. Yang, and S. Zhu, "Recovering the optimal solution by dual random projection," in *Proc. 26th Annu. Conf. Learn. Theory*, 2013, pp. 135–157.

[2] S. Kaski, "Dimensionality reduction by random mapping: Fast similarity computation for clustering," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, vol. 1. May 1998, pp. 413–418.

[3] E. Bingham and H. Mannila, "Random projection in dimensionality reduction: Applications to image and text data," in *Proc. 7th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2001, pp. 245–250.

[4] D. Fradkin and D. Madigan, "Experiments with random projections for machine learning," in *Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2003, pp. 517–522.

[5] S. S. Vempala, *The Random Projection Method*. Providence, RI, USA: AMS, 2004.

[6] N. Goel, G. Bebis, and A. Nefian, "Face recognition experiments with random projection," *Proc. SPIE*, vol. 5779, pp. 426–437, Mar. 2005.

[7] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Advances in Neural Information Processing Systems 20*. Red Hook, NY, USA: Curran Associates, Inc., 2008, pp. 1177–1184.

[8] O. Maillard and R. Munos, "Linear regression with random projections," *J. Mach. Learn. Res.*, vol. 13, pp. 2735–2772, 2012.

[9] X. Z. Fern and C. E. Brodley, "Random projection for high dimensional data clustering: A cluster ensemble approach," in *Proc. 20th Int. Conf. Mach. Learn.*, 2003, pp. 186–193.
[10] C. Boutsidis, A. Zouzias, and P. Drineas, "Random projections for k-means clustering," in *Advances in Neural Information Processing Systems 23*. Red Hook, NY, USA: Curran Associates, Inc., 2010, pp. 298–306.
[11] S. Dasgupta and Y. Freund, "Random projection trees and low dimensional manifolds," in *Proc. 40th Annu. ACM Symp. Theory Comput.*, 2008, pp. 537–546.
[12] Y. Freund, S. Dasgupta, M. Kabra, and N. Verma, "Learning the structure of manifolds using random projections," in *Advances in Neural Information Processing Systems 20*. Red Hook, NY, USA: Curran Associates, Inc., 2008, pp. 473–480.
[13] N. Thaper, S. Guha, P. Indyk, and N. Koudas, "Dynamic multidimensional histograms," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2002, pp. 428–439.
[14] R. I. Arriaga and S. Vempala, "An algorithmic theory of learning: Robust concepts and random projection," in *Proc. 40th Annu. Symp. Found. Comput. Sci.*, 1999, pp. 616–623.
[15] M.-F. Balcan, A. Blum, and S. Vempala, "Kernels as features: On kernels, margins, and low-dimensional mappings," *Mach. Learn.*, vol. 65, no. 1, pp. 79–94, 2006.
[16] Q. Shi, C. Shen, R. Hill, and A. van den Hengel, "Is margin preserved after random projection?" in *Proc. 29th Int. Conf. Mach. Learn.*, 2012, pp. 591–598.
[17] S. Paul, C. Boutsidis, M. Magdon-Ismail, and P. Drineas, "Random projections for support vector machines," in *Proc. 16th Int. Conf. Artif. Intell. Statist.*, 2013, pp. 498–506.
[18] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Mar. 2003.
[19] R. Vershynin, "Lectures in geometric functional analysis," Univ. Michigan, Ann Arbor, MI, USA, Tech. Rep., 2009.
[20] W. Johnson and J. Lindenstrauss, "Extensions of Lipschitz mappings into a Hilbert space," in *Proc. Conf. Modern Anal. Probab.*, vol. 26. 1984, pp. 189–206.
[21] S. Dasgupta and A. Gupta, "An elementary proof of a theorem of Johnson and Lindenstrauss," *Random Struct. Algorithms*, vol. 22, no. 1, pp. 60–65, 2003.
[22] D. Achlioptas, "Database-friendly random projections: Johnson–Lindenstrauss with binary coins," *J. Comput. Syst. Sci.*, vol. 66, no. 4, pp. 671–687, 2003.
[23] R. I. Arriaga and S. Vempala, "An algorithmic theory of learning: Robust concepts and random projection," *Mach. Learn.*, vol. 63, no. 2, pp. 161–182, 2006.
[24] A. Magen, "Dimensionality reductions that preserve volumes and distance to affine spaces, and their algorithmic applications," in *Randomization and Approximation Techniques in Computer Science* (Lecture Notes in Computer Science), vol. 2483. Berlin, Germany: Springer-Verlag, 2002, pp. 239–253.
[25] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[26] E. Hazan, T. Koren, and N. Srebro, "Beating SGD: Learning SVMs in sublinear time," in *Advances in Neural Information Processing Systems 24*. Red Hook, NY, USA: Curran Associates, Inc., 2011, pp. 1233–1241.
[27] S. Shalev-Shwartz and Y. Singer, "Online learning meets optimization in the dual," in *Proc. 19th Annu. Conf. Learn. Theory (COLT)*, 2006, pp. 423–437.
[28] S. M. Kakade, S. Shalev-Shwartz, and A. Tewari, "On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization," Toyota Technol. Inst., Chicago, IL, USA, Tech. Rep., 2009.
[29] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course* (Applied optimization), vol. 87. Norwell, MA, USA: Kluwer, 2004.
[30] B. Recht, "A simpler approach to matrix completion," *J. Mach. Learn. Res.*, vol. 12, pp. 3413–3430, Feb. 2011.
[31] N. Ailon and E. Liberty, "Fast dimension reduction using Rademacher series on dual BCH codes," *Discrete Comput. Geometry*, vol. 42, no. 4, pp. 615–630, 2009.
[32] D. M. Kane and J. Nelson, "Sparser Johnson–Lindenstrauss transforms," *J. ACM*, vol. 61, no. 1, pp. 4:1–4:23, 2014.
[33] E. Hazan and S. Kale, "Beyond the regret minimization barrier: An optimal algorithm for stochastic strongly-convex optimization," in *Proc. 24th Annu. Conf. Learn. Theory (COLT)*, 2011, pp. 421–436.
[34] J. Nocedal and S. J. Wright, *Numerical Optimization* (Operations Research and Financial Engineering), 2nd ed. New York, NY, USA: Springer-Verlag, 2006.
[35] Y. Nesterov, "Gradient methods for minimizing composite functions," *Math. Programming*, vol. 140, no. 1, pp. 125–161, 2013.
[36] L. Bottou and O. Bousquet, "The tradeoffs of large scale learning," in *Advances in Neural Information Processing Systems 20*. Red Hook, NY, USA: Curran Associates, Inc., 2008, pp. 161–168.
[37] V. Vapnik, *Estimation of Dependences Based on Empirical Data*. New York, NY, USA: Springer-Verlag, 1982.
[38] P. L. Bartlett and S. Mendelson, "Rademacher and Gaussian complexities: Risk bounds and structural results," *J. Mach. Learn. Res.*, vol. 3, pp. 463–482, Mar. 2003.
[39] P. L. Bartlett, O. Bousquet, and S. Mendelson, "Local rademacher complexities," *Ann. Statist.*, vol. 33, no. 4, pp. 1497–1537, 2005.
[40] J. Borwein, A. Lewis, J. Borwein, and A. Lewis, *Convex Analysis and Nonlinear Optimization: Theory and Examples*. New York, NY, USA: Springer-Verlag, 2006.
[41] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*. Cambridge, U.K.: Cambridge Univ. Press, 2006.
[42] A. Gittens and J. A. Tropp. (2011). "Tail bounds for all eigenvalues of a sum of random matrices." [Online]. Available: http://arxiv.org/abs/1104.4513
[43] S. Zhu. (2012). "A short note on the tail bound of Wishart distribution." [Online]. Available: http://arxiv.org/abs/1212.5860.
[44] B. Laurent and P. Massart, "Adaptive estimation of a quadratic functional by model selection," *Ann. Statist.*, vol. 28, no. 5, pp. 1302–1338, 2000.

Lijun Zhang (M'08) received the B.S. and Ph.D. degrees in Software Engineering and Computer Science from Zhejiang University, China, in 2007 and 2012, respectively. He is currently an associate professor of the Department of Computer Science and Technology, Nanjing University, China. Prior to joining Nanjing University, he was a postdoctoral researcher at the Department of Computer Science and Engineering, Michigan State University, USA. His research interests include machine learning, optimization, information retrieval and data mining.

Mehrdad Mahdavi is a Research Assistance Professor at Toyota Technological Institute at University of Chicago (TTI-C). He obtained his Ph.D. degree from Michigan State University in Computer Science under the supervision of Professor Rong Jin in 2014 and before that spent two years as a Ph.D. candidate at Sharif University of Technology. He received the M.Sc. degree from Sharif University of Technology, Tehran, Iran. He has won the Top Cited Paper Award from the journal of *Applied Mathematics and Computation* (Elsevier) in 2010 and the Mark Fulk Best Student Paper Award at the Conference on Learning Theory (COLT) in 2012. His current research interests include Machine Learning focused on Online Learning, Convex Optimization, and Sequential and Statistical Learning Theory.

Rong Jin focuses his research on statistical machine learning and its application to information retrieval. He has worked on a variety of machine learning algorithms and their application to information retrieval, including retrieval models, collaborative filtering, cross lingual information retrieval, document clustering, and video/image retrieval. He has published over 180 conference and journal articles on related topics. Dr. Jin holds a Ph.D. degree in Computer Science from Carnegie Mellon University. He received the NSF Career Award in 2006.

Tianbao Yang is an Assistant Professor of the Computer Science Department at the University of Iowa. He received the Ph.D. degree in Computer Science from Michigan State University in 2012. He worked as a researcher in GE Global Research from 2012 to 2013 and in NEC Laboratories America, Inc. from 2013 to 2014. He has board interests in machine learning and has focused on several research topics, including social network analysis and large scale optimization in machine learning. He has won the Mark Fulk Best student paper award at 25th Conference on Learning Theory (COLT) in 2012. He also served as program committee for several conferences, including AAAI'15, AAAI'12, CIKM'12, '13, IJCAI'13, ACML'12.

Shenghuo Zhu is a Principle Engineer at Alibaba Group. Prior to Alibaba, he spent ten years at NEC Laboratories America, and one and half year at Amazon.com, Inc. He received his Ph.D. degree in Computer Science from University of Rochester in 2003. His primary research interests include machine learning, computer vision, data mining and information retrieval.