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**

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:

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

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.,

$$|\ell'(z) - \ell'(z')| \leq \gamma |z - z'|.$$

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:

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

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

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

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:

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

The corresponding dual problem is

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

Intuitively, the choice of the Gaussian matrix A is justified by the fact that $E[\widehat{\mathbf{x}}_i^\top \widehat{\mathbf{x}}_j] = \mathbf{x}_i^\top E[AA^\top] \mathbf{x}_j = \mathbf{x}_i^\top \mathbf{x}_j$,

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 $\widehat{\boldsymbol{\alpha}}_*$:

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

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

$$\widehat{\mathbf{w}} = A \mathbf{z}_*, \qquad (6)$$

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

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

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 $1 - 2^{-d+m}$,

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

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,

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 $\mathbf{v}_i \in \mathbb{R}^n$

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 $\overline{\mathcal{S}} = [d] \setminus \mathcal{S}$. To capture that $\mathbf{w}_*$ is approximately sparse, we assume that

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

holds for some small constant ρ.

Theorem 6: Suppose

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

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

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

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

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

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

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

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

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

$$\le [(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*) \circ \mathbf{y}]^\top (G - \widehat{G})(\boldsymbol{\alpha}_* \circ \mathbf{y}) \tag{18}$$

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

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

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

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

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

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

Furthermore, it is easy to verify that

$$\frac{1}{2\lambda n}(\boldsymbol{\alpha}_* \circ \mathbf{y})^\top \widehat{G}(\boldsymbol{\alpha}_* \circ \mathbf{y})$$

$$= \frac{(\widehat{\boldsymbol{\alpha}}_* \circ \mathbf{y})^\top \widehat{G}(\widehat{\boldsymbol{\alpha}}_* \circ \mathbf{y})}{2\lambda n} + \frac{\langle\widehat{G}(\widehat{\boldsymbol{\alpha}}_* \circ \mathbf{y}), (\boldsymbol{\alpha}_* - \widehat{\boldsymbol{\alpha}}_*) \circ \mathbf{y}\rangle}{\lambda n}$$

$$+ \frac{1}{2\lambda n}[(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*) \circ \mathbf{y}]^\top \widehat{G}[(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*) \circ \mathbf{y}]. \tag{22}$$

Adding (21) to (22), we obtain

$$\widehat{L}(\boldsymbol{\alpha}_*) \ge \widehat{L}(\widehat{\boldsymbol{\alpha}}_*) + \langle\nabla F(\widehat{\boldsymbol{\alpha}}_*), \boldsymbol{\alpha}_* - \widehat{\boldsymbol{\alpha}}_*\rangle + \frac{1}{2\gamma}\|\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*\|^2$$

$$+ \frac{1}{\lambda n}\langle\widehat{G}(\widehat{\boldsymbol{\alpha}}_* \circ \mathbf{y}), (\boldsymbol{\alpha}_* - \widehat{\boldsymbol{\alpha}}_*) \circ \mathbf{y}\rangle$$

$$+ \frac{1}{2\lambda n}[(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*) \circ \mathbf{y}]^\top \widehat{G}[(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*) \circ \mathbf{y}]$$

$$= \widehat{L}(\widehat{\boldsymbol{\alpha}}_*) + \langle\nabla\widehat{L}(\widehat{\boldsymbol{\alpha}}_*), \boldsymbol{\alpha}_* - \widehat{\boldsymbol{\alpha}}_*\rangle + \frac{1}{2\gamma}\|\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*\|^2$$

$$+ [(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*) \circ \mathbf{y}]^\top \widehat{G}[(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*) \circ \mathbf{y}]$$

$$\overset{(20)}{\ge} \widehat{L}(\widehat{\boldsymbol{\alpha}}_*) + \frac{1}{2\gamma}\|\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*\|^2$$

$$+ \frac{1}{2\lambda n}[(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*) \circ \mathbf{y}]^\top \widehat{G}[(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*) \circ \mathbf{y}]. \tag{23}$$

On the other hand, we have

$$\widehat{L}(\boldsymbol{\alpha}_*) + \frac{1}{\lambda n}[(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*) \circ \mathbf{y}]^\top(\widehat{G} - G)(\boldsymbol{\alpha}_* \circ \mathbf{y})$$

$$= \widehat{L}(\boldsymbol{\alpha}_*) + \langle\nabla\widehat{L}(\boldsymbol{\alpha}_*) - \nabla L(\boldsymbol{\alpha}_*), \widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*\rangle$$

$$\overset{(19)}{\le} \widehat{L}(\boldsymbol{\alpha}_*) + \langle\nabla\widehat{L}(\boldsymbol{\alpha}_*), \widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*\rangle$$

$$\le \widehat{L}(\widehat{\boldsymbol{\alpha}}_*) - \frac{1}{2\lambda n}[(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*) \circ \mathbf{y}]^\top \widehat{G}[(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*) \circ \mathbf{y}]$$

$$- \frac{1}{2\gamma}\|\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*\|^2 \tag{24}$$

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

$$G = V\Sigma U^\top U\Sigma V^\top = V\Sigma^2 V^\top,$$

$$\widehat{G} = V\Sigma U^\top A A^\top U\Sigma V^\top = V\Sigma B B^\top\Sigma V^\top,$$

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

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

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

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

It is easy to verify that

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

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

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

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

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

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

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

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

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

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

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

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

$$\frac{1-\epsilon}{2}\|\mathbf{a}\|^2 + \frac{\lambda n}{2\gamma\,\lambda_{r+1}^2}\|\mathbf{b}\|^2$$

$$\leq \|\mathbf{a}\|\|\mathbf{c}\|\epsilon + \|\mathbf{b}\|\|\mathbf{d}\|v + \|\mathbf{a}\|\|\mathbf{d}\|\tau + \|\mathbf{b}\|\|\mathbf{c}\|\tau$$

$$\leq \frac{1-\epsilon}{8}\|\mathbf{a}\|^2 + \frac{2\epsilon^2}{1-\epsilon}\|\mathbf{c}\|^2 + \frac{\lambda n}{8\gamma\,\lambda_{r+1}^2}\|\mathbf{b}\|^2 + \frac{2\gamma\,\lambda_{r+1}^2 v^2}{\lambda n}\|\mathbf{d}\|^2$$

$$+\frac{1-\epsilon}{8}\|\mathbf{a}\|^2 + \frac{2\tau^2}{1-\epsilon}\|\mathbf{d}\|^2$$

$$+\frac{\lambda n}{8\gamma\,\lambda_{r+1}^2}\|\mathbf{b}\|^2 + \frac{2\gamma\,\lambda_{r+1}^2\tau^2}{\lambda n}\|\mathbf{c}\|^2,$$

which implies

$$\frac{1-\epsilon}{4}\|\mathbf{a}\|^2 + \frac{\lambda n}{4\gamma\,\lambda_{r+1}^2}\|\mathbf{b}\|^2$$

$$\leq \left(\frac{2\epsilon^2}{1-\epsilon} + \frac{2\gamma\,\lambda_{r+1}^2\tau^2}{\lambda n}\right)\|\mathbf{c}\|^2$$

$$+\left(\frac{2\gamma\,\lambda_{r+1}^2 v^2}{\lambda n} + \frac{2\tau^2}{1-\epsilon}\right)\|\mathbf{d}\|^2$$

$$\overset{(29,30)}{\leq} \left(\frac{2\epsilon^2}{1-\epsilon} + \frac{2\gamma\,\lambda_{r+1}^2\tau^2}{\lambda n}\right)\lambda^2 n^2\|\mathbf{w}_*\|^2$$

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

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

$$\|\widetilde{\mathbf{w}} - \mathbf{w}_*\|^2 \overset{(28)}{\leq} 8\left(\frac{1}{1-\epsilon} + \frac{\gamma\,\lambda_{r+1}^2}{\lambda n}\right)$$

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

leading to the third inequality in Theorem 2.

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

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

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

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

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

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

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

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

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

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

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

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

Based on Lemma 4, we have

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

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 $X_{\overline{S}}$, as revealed by the following lemma.

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

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

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

$$\widehat{G} = X_\mathcal{S}^\top A_\mathcal{S} A_\mathcal{S}^\top X_\mathcal{S} + X_{\overline{S}}^\top A_{\overline{S}} A_{\overline{S}}^\top X_{\overline{S}}$$

$$+ X_\mathcal{S}^\top A_\mathcal{S} A_{\overline{S}}^\top X_{\overline{S}} + X_{\overline{S}}^\top A_{\overline{S}} A_\mathcal{S}^\top X_\mathcal{S}.$$

We define

$$\mathbf{a} = X_\mathcal{S}[(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)\circ\mathbf{y}], \quad \mathbf{b} = X_{\overline{S}}(\widehat{\boldsymbol{\alpha}}_* \circ \mathbf{y}),$$

$$\mathbf{c} = X_\mathcal{S}(\boldsymbol{\alpha}_* \circ \mathbf{y}),$$

$$\epsilon = \left\|A_\mathcal{S} A_\mathcal{S}^\top - I\right\|_2, \quad \tau = \left\|A_{\overline{S}} A_\mathcal{S}^\top\right\|_2.$$

Then, we have

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

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

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

$$[(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)\circ\mathbf{y}]^\top \widehat{G}[(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)\circ\mathbf{y}]$$

$$\overset{(38)}{=} \mathbf{a}^\top A_\mathcal{S} A_\mathcal{S}^\top \mathbf{a} + \mathbf{b}^\top A_{\overline{S}} A_{\overline{S}}^\top \mathbf{b}$$

$$+ \mathbf{a}^\top A_\mathcal{S} A_{\overline{S}}^\top \mathbf{b} + \mathbf{b}_{\overline{S}}^\top A_{\overline{S}} A_\mathcal{S}^\top \mathbf{a}$$

$$\geq (1-\epsilon)\|\mathbf{a}\|^2 - 2\tau\|\mathbf{a}\|\|\mathbf{b}\|. \tag{41}$$

$$\frac{\lambda n}{\gamma}\|\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*\|^2 \overset{y_i\in\pm 1}{=} \frac{\lambda n}{\gamma}\|(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)\circ\mathbf{y}\|^2$$

$$\geq \frac{\lambda n}{\gamma}\frac{\left\|X_{\overline{S}}[(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)\circ\mathbf{y}]\right\|^2}{\|X_{\overline{S}}^\top X_{\overline{S}}\|_2} \overset{(12,38)}{=} \frac{\lambda n}{\gamma\,\eta}\|\mathbf{b}\|^2. \tag{42}$$

$$[(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)\circ\mathbf{y}]^\top (G - \widehat{G})(\boldsymbol{\alpha}_* \circ \mathbf{y})$$

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

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

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

In the case when

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

we have

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

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

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

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

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

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

$$\tau = \|A_{\overline{\mathcal{S}}} A_{\mathcal{S}}^\top\|_2 \leq \frac{7}{3}\sqrt{\text{\small{as6. 12(6.2(8ilit42d)1(F)9(14romit4))Tj(L)17,it4y14isit4raightforwo(d447di)e}}}$$

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

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

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).

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

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

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

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

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

where

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

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

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

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

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

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

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

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

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

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 $\|\widetilde{\mathbf{w}} - \mathbf{w}_*\| \le \frac{\alpha}{2}\|\mathbf{w}_*\|$, 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

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

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 $L_l^t \ge \mu_l^t \ge \lambda$ 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

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

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,

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

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,

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

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.

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

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

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

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

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 $\{\boldsymbol{\xi}_j : j = 1, \dots, n\}$ 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 $\theta \geq 0$

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

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

We write $B = \frac{1}{\sqrt{m}}(\mathbf{v}_1, \dots, \mathbf{v}_m)$, 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 $1 - 2r\exp(-\theta)$,

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

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

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

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

:F5 1 Tf 7.5716 096D (r)Tj Tc [20.0j 7.et5 1 Tf 9.9626-F1 78

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 $1 - 2\delta$,

$$\|B_{\bar{r}} B_r^\top\|_2$$

$$\leq \sqrt{\frac{2(d-r)}{m} \log \frac{d}{\delta}} + \frac{2}{3m} \left(\sqrt{d-r} + \sqrt{2 \log \frac{2m}{\delta}} \right)$$

$$\cdot \left(\sqrt{r} + \sqrt{2 \log \frac{2m}{\delta}} \right) \log \frac{d}{\delta}$$

$$= \sqrt{\frac{2(d-r)}{m} \log \frac{d}{\delta}} + \frac{1}{3} \sqrt{\frac{2(d-r)}{m} \log \frac{d}{\delta}} \sqrt{\frac{1}{m(d-r)} \log \frac{d}{\delta}}$$

$$\cdot \left(\sqrt{2r(d-r)} + \sqrt{4r \log \frac{2m}{\delta}} + \sqrt{4(d-r) \log \frac{2m}{\delta}} \right.$$

$$\left. + 2\sqrt{2} \log \frac{2m}{\delta} \right).$$

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

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

$$m(d-r) \overset{(10)}{\geq} 4(d-r) \log \frac{2m}{\delta} \log \frac{d}{\delta} \overset{(10,11)}{\geq} 4r \log \frac{2m}{\delta} \log \frac{d}{\delta},$$

$$m(d-r) \overset{(10)}{\geq} 4(d-r) \log \frac{2m}{\delta} \log \frac{d}{\delta},$$

$$m(d-r) \overset{(10)}{\geq} 4(d-r) \log \frac{2m}{\delta} \log \frac{d}{\delta} \overset{(11)}{\geq} 8 \log^2 \frac{2m}{\delta} \log \frac{d}{\delta}.$$

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 $1 - \delta$,

$$\left\| B_{\bar{r}} B_{\bar{r}}^\top - I \right\|_2$$

$$\leq \sqrt{\frac{2(d-r+1)}{m} \log \frac{2(d-r)}{\delta}} + \frac{2(d-r)}{m} \log \frac{2(d-r)}{\delta}$$

$$\leq \frac{4(d-r+1)}{m} \log \frac{2(d-r)}{\delta},$$

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

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

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

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

APPENDIX F
DERIVATION OF (67)

We have

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

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

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

From (69) and (70), we have

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

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

ACKNOWLEDGMENT

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.

[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.