Recovering the Optimal Solution by Dual Random Projection

Lijun Zhang   ZHANGLIJ@MSU.EDU
Mehrdad Mahdavi   MAHDAVIM@MSU.EDU
Rong Jin   RONGJIN@CSE.MSU.EDU

Tianbao Yang   TYANG@GE.COM

Shenghuo Zhu   ZSH@NEC-LABS.COM

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 of using random projection, in this work, we consider the , i.e., how to accurately recover the optimal solution to the original optimization problem in the high-dimensional space based on the solution learned from the subspace spanned by random projections. We present a simple algorithm, termed , that 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 of low rank or can be well approximated by a low rank matrix.

Keywords: Random projection, Primal solution, Dual solution, Low rank

1. 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 (Kaski, 1998; Bingham and Mannila, 2001). It has been successfully applied to many machine learning tasks, including classification (Fradkin and Madigan, 2003; Vempala, 2004; Rahimi and Recht, 2008), regression (Maillard and Munos, 2012), clustering (Fern and Brodley, 2003; Boutsidis et al., 2010), manifold learning (Dasgupta and Freund, 2008; Freund et al., 2008), and information retrieval (Goel et al., 2005).

In this work, we focus on random projection for classification. While previous studies were devoted to analyzing the classification performance using random projection (Arriaga and Vempala, 1999; Balcan et al., 2006; Paul et al., 2012; Shi et al., 2012), 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 (Guyon and Elisseeff, 2003), 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.

The proposed algorithm for recovering the optimal solution consists of two simple steps. In the first step, similar to previous studies, we apply random projection to reducing the dimensionality of the data, and then solve a low-dimensional optimization problem. In the second step, we construct the dual solution of the low-dimensional problem from its primal solution, and then use it to recover the optimal solution to the original high-dimensional problem. Our analysis reveals that with a high probability, we are able to recover the optimal solution with a small error by using $\Omega(r \log r)$ projections, where r is the rank of the data matrix. A similar result also holds when the data matrix can be well approximated by a low rank matrix. We further show that the proposed algorithm can be applied iteratively to recovering the optimal solution with a relative error by using $O(\log 1/\,)$ iterations.

The rest of the paper is arranged as follows. Section 2 describes the problem of recovering optimal solution by random projection, the theme of this work. Section 3 describes the dual random projection approach for recovering the optimal solution. Section 4 presents the main theoretical results for the proposed algorithm. Section 5 presents the proof for the theorems stated in Section 4. Section 6 concludes this work.

2. The Problem

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 dimension and $y_i \in \{-1, +1\}$ is the binary class assignment for $\mathbf{x}_i$. Let $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)^\top$ include input patterns and the class assignments of all training examples. A classifier $\mathbf{w} \in \mathbb{R}^d$ is learned by solving the following optimization problem:

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

where (z) is a convex loss function that is differentiable[1]. By writing (z) in its convex conjugate form, i.e.,

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

where $\,_*(\alpha)$ is the convex conjugate of (z) and Ω is the domain of the dual variable, we get the dual optimization problem:

$$\max_{\boldsymbol{\alpha} \in \Omega^n} - \sum_{i=1}^{n} \,_*(\alpha_i) - \frac{1}{2} \boldsymbol{\alpha}^\top G \boldsymbol{\alpha}, \tag{2}$$

where $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_n)^\top$ and $D(\mathbf{y}) = \mathrm{diag}(\mathbf{y})$ and G is the Gram matrix given by

$$G = D(\mathbf{y}) X^\top X D(\mathbf{y}). \tag{3}$$

---

1. For non-differentiable loss functions such as hinge loss, we could apply the smoothing technique described in (Nesterov, 2005) to make it differentiable.

In the following, we 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 $\quad \mathbf{w}_* \in \mathbb{R}^d \quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad \boldsymbol{\alpha}_* \in \mathbb{R}^n$

$$\mathbf{w}_* = -\frac{1}{}XD(\mathbf{y})\boldsymbol{\alpha}_*, \qquad [\boldsymbol{\alpha}_*]_i = \nabla\ \left(y_i\mathbf{x}_i^\top\mathbf{w}_*\right),\ i = 1,\ldots,n. \tag{4}$$

The proof of Proposition 1 and other omitted proofs are deferred to the Appendix. 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. Let $R \in \mathbb{R}^{d \times m}$ be a Gaussian random matrix, where each entry $R_{i,j}$ is independently drawn from a Gaussian distribution $\mathcal{N}(0,1)$ and m is significantly smaller than d. Using the random matrix R, we generate a low-dimensional representation for each input example by

$$\widehat{\mathbf{x}}_i = \frac{1}{\sqrt{m}}R^\top\mathbf{x}_i, \tag{5}$$

and solve the following low-dimensional optimization problem:

$$\min_{\mathbf{z}\in\mathbb{R}^m} \frac{}{2}\|\mathbf{z}\|^2 + \sum_{i=1}^n\ (y_i\mathbf{z}^\top\widehat{\mathbf{x}}_i). \tag{6}$$

The corresponding dual problem is written as

$$\max_{\boldsymbol{\alpha}\in\Omega^n} -\sum_{i=1}^n\ {}_*(\alpha_i) - \frac{1}{2}\boldsymbol{\alpha}^\top\widehat{G}\boldsymbol{\alpha}, \tag{7}$$

where

$$\widehat{G} = D(\mathbf{y})X^\top\frac{RR^\top}{m}XD(\mathbf{y}). \tag{8}$$

Intuitively, the choice of Gaussian random matrix R is justified by the expectation of the dot-product between any two examples in the projected space is equal to the dot-product in the original space, i.e.,

$$\mathrm{E}[\widehat{\mathbf{x}}_i^\top\widehat{\mathbf{x}}_j] = \mathbf{x}_i^\top\mathrm{E}\left[\frac{1}{m}RR^\top\right]\mathbf{x}_j = \mathbf{x}_i^\top\mathbf{x}_j,$$

where the last equality follows from $\mathrm{E}[RR^\top/m] = I$. Thus, $G = \widehat{G}$ holds in expectation.

Let $\mathbf{z}_* \in \mathbb{R}^m$ denote the optimal primal solution to the low-dimensional problem (6), and $\widehat{\boldsymbol{\alpha}}_* \in \mathbb{R}^n$ denote the optimal dual solution to (7). Similar to Proposition 1, the following proposition connects $\mathbf{z}_*$ and $\widehat{\boldsymbol{\alpha}}_*$.

Proposition 2

$$\mathbf{z}_* = -\frac{1}{}\frac{1}{\sqrt{m}}R^\top XD(\mathbf{y})\widehat{\boldsymbol{\alpha}}_*, \qquad [\widehat{\boldsymbol{\alpha}}_*]_i = \nabla\ \left(\frac{y_i}{\sqrt{m}}\mathbf{x}_i^\top R\mathbf{z}_*\right),\ i = 1,\ldots,n. \tag{9}$$

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

$$\widehat{\mathbf{w}} = \frac{1}{\sqrt{m}}R\mathbf{z}_*. \tag{10}$$

The classification performance of $\widehat{\mathbf{w}}$ has been examined by many studies (Arriaga and Vempala, 1999; Balcan et al., 2006; Paul et al., 2012; Shi et al., 2012). The general conclusion is that when the original data is linearly separable with a large margin, the classification error for $\widehat{\mathbf{w}}$ is usually 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 of the optimal solution $\mathbf{w}_*$. In fact, as we will see in Section 4, the naive solution is almost guaranteed to be a BAD approximation of the optimal solution, that is, $\|\widehat{\mathbf{w}} - \mathbf{w}_*\|_2 = \Omega(\sqrt{d/m}\|\mathbf{w}_*\|_2)$. This observation leads to an interesting question:

$\mathbf{w}_*$ $\mathbf{z}_*$?

Relationship to Compressive Sensing The proposed problem is closely related to compressive sensing (Donoho, 2006; Candès and Wakin, 2008) where the goal is to recover a high-dimensional but sparse vector using a small number of random measurements. The key difference between our work and compressive sensing is that we do not have the direct access to the random measurement of the target vector (which in our case is $\mathbf{w}_* \in \mathbb{R}^d$). Instead, $\mathbf{z}_* \in \mathbb{R}^m$ is the optimal solution to (6), the primal problem using random projection. However, the following theorem shows that $\mathbf{z}_*$ is a good approximation of $R^\top \mathbf{w}_*/\sqrt{m}$, which includes m random measurements of $\mathbf{w}_*$, if the data matrix X is of low rank and the number of random measurements m is sufficiently large.

Theorem 1 ^{F} $0 < \ \leq 1/2$ $1 - \ - \exp(-m/32)$

$$\|\sqrt{m}\mathbf{z}_* - R^\top\mathbf{w}_*\|_2 \leq \frac{\sqrt{2}}{\sqrt{1-}}\|R^\top\mathbf{w}_*\|_2,$$

$$m \geq \frac{(r+1)\log(2r/\ )}{c\ ^2},$$

      c $1/4$ r X

Given the approximation bound in Theorem 1, it is appealing to reconstruct $\mathbf{w}_*$ using the compressive sensing algorithm provided that $\mathbf{w}_*$ is sparse to certain bases. We note that the low rank assumption for data matrix X implies that $\mathbf{w}_*$ is sparse with respect to the singular vectors of X. However, since $\mathbf{z}_*$ only provides an approximation to the random measurements of $\mathbf{w}_*$, running the compressive sensing algorithm will not be able to perfectly recover $\mathbf{w}_*$ from $\mathbf{z}_*$. In Section 3, we present an algorithm, that recovers $\mathbf{w}_*$ with a small error, provided that the data matrix X is of low rank. Compared to the compressive sensing algorithm, the main advantage of the proposed algorithm is its computational simplicity; it neither computes the singular vectors of X nor solves an optimization problem that minimizes the $_1$ norm.

Algorithm 1 A Dual Random Projection Approach for Recovering Optimal Solution
---
1: **Input:** input patterns $X \in \mathbb{R}^{d \times n}$, binary class assignment $\mathbf{y} \in \{-1, +1\}^n$, and sample size m
2: Sample a Gaussian random matrix $R \in \mathbb{R}^{d \times m}$ and compute $\widehat{X} = [\widehat{\mathbf{x}}_1, \ldots, \widehat{\mathbf{x}}_n] = R^\top X / \sqrt{m}$
3: Obtain the primal solution $\mathbf{z}_* \in \mathbb{R}^m$ by solving the optimization problem in (6)
4: Construct the dual solution $\widehat{\boldsymbol{\alpha}}_* \in \mathbb{R}^n$ by Proposition 2, i.e.,

$$[\widehat{\boldsymbol{\alpha}}_*]_i = \nabla \; \left(\frac{y_i}{\sqrt{m}} \mathbf{x}_i^\top R \mathbf{z}_* \right), \; i = 1, \ldots, n$$

5: Compute $\widetilde{\mathbf{w}} \in \mathbb{R}^d$ according to (12), i.e., $\widetilde{\mathbf{w}} = -XD(\mathbf{y})\widehat{\boldsymbol{\alpha}}_*/$
6: **Output:** the recovered solution $\widetilde{\mathbf{w}}$
---

3. Algorithm

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

$$\mathbf{w}_* = -\frac{1}{\quad}XD(\mathbf{y})\boldsymbol{\alpha}_*, \tag{11}$$

where $\boldsymbol{\alpha}_*$ is the optimal solution to the dual problem (2). Given the projected data $\widehat{\mathbf{x}} = R^\top \mathbf{x}/\sqrt{m}$, we have reached an approximate dual problem in (7). Comparing it with the dual problem in (2), the only difference is that the Gram matrix $G = D(\mathbf{y})X^\top X D(\mathbf{y})$ in (2) is replaced with $\widehat{G} = D(\mathbf{y})X^\top RR^\top X D(\mathbf{y})/m$ in (7). Recall that $\mathrm{E}[RR^\top/m] = I$. Thus, when the number of random projections m is sufficiently large, $\widehat{G}$ will be close to the G and we would also expect $\widehat{\boldsymbol{\alpha}}_*$ to be close to $\boldsymbol{\alpha}_*$. As a result, we can use $\widehat{\boldsymbol{\alpha}}_*$ to approximate $\boldsymbol{\alpha}_*$ in (11), which yields a recovered prediction model given below:

$$\widetilde{\mathbf{w}} = -\frac{1}{\quad}XD(\mathbf{y})\widehat{\boldsymbol{\alpha}}_* = -\sum_{i=1}^{n} \frac{1}{\quad}y_i[\widehat{\boldsymbol{\alpha}}_*]_i\mathbf{x}_i. \tag{12}$$

Note that the key difference between the recovered solution $\widetilde{\mathbf{w}}$ and the naive solution $\widehat{\mathbf{w}}$ is that $\widehat{\mathbf{w}}$ is computed by mapping the optimal primal solution $\mathbf{z}_*$ in the projected space back to the original space via the random matrix R, while $\widetilde{\mathbf{w}}$ is computed directly in the original space using the approximate dual solution $\widehat{\boldsymbol{\alpha}}_*$. Therefore, the naive solution $\widehat{\mathbf{w}}$ lies in the subspace spanned by the column vectors in the random matrix R (denoted by $\mathcal{A}_R$), 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 $\mathcal{A}$). The mismatch between spaces $\mathcal{A}_R$ and $\mathcal{A}$ leads to the large approximation error for $\widehat{\mathbf{w}}$.

According to Proposition 2, we can construct the dual solution $\widehat{\boldsymbol{\alpha}}_*$ from the primal solution $\mathbf{z}_*$. Thus, we do not need to solve the dual problem in (7) to obtain $\widehat{\boldsymbol{\alpha}}_*$. Instead, we solve the low-dimensional optimization problem in (6) to get $\mathbf{z}_*$ and construct $\widehat{\boldsymbol{\alpha}}_*$ from it. Algorithm 1 shows the details of the proposed method. We note that although dual variables have been widely used in the analysis of convex optimization (Boyd and Vandenberghe, 2004; Hazan et al., 2011) and online learning (Shalev-Shwartz and Singer, 2006), to the best of

---

Algorithm 2 An Iterative Dual Random Projection Approach for Recovering Optimal Solution

---

1: **Input:** input patterns $X \in \mathbb{R}^{d \times n}$, binary class assignment $\mathbf{y} \in \{-1, +1\}^n$, sample size m, and number of iterations T

2: Sample a Gaussian random matrix $R \in \mathbb{R}^{d \times m}$ and compute $\widehat{X} = R^\top X / \sqrt{m}$

3: Initialize $\widetilde{\mathbf{w}}^0 = \mathbf{0}$

4: **for** $t = 1, \ldots, T$ **do**

5:    Obtain $\mathbf{z}_*^t \in \mathbb{R}^m$ by solving the following optimization problem

$$\min_{\mathbf{z} \in \mathbb{R}^m} \frac{1}{2} \left\| \mathbf{z} + \frac{1}{\sqrt{m}} R^\top \widetilde{\mathbf{w}}^{t-1} \right\|_2^2 + \sum_{i=1}^n \left(y_i \mathbf{z}^\top \widehat{\mathbf{x}}_i + y_i [\widetilde{\mathbf{w}}^{t-1}]^\top \mathbf{x}_i \right) \qquad (13)$$

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

$$[\widehat{\boldsymbol{\alpha}}_*^t]_i = \nabla \left(y_i \widehat{\mathbf{x}}_i^\top \mathbf{z}_*^t + y_i [\widetilde{\mathbf{w}}^{t-1}]^\top \mathbf{x}_i \right), \ i = 1, \ldots, n$$

7:    Update the solution by $\widetilde{\mathbf{w}}^t = -X D(\mathbf{y}) \widehat{\boldsymbol{\alpha}}_*^t /$

8: **end for**

9: **Output** the recovered solution $\widetilde{\mathbf{w}}^T$

---

our knowledge, this is the first time that dual variables are used in conjunction with random projection for recovering the optimal solution.

To further reduce the recovery error, we develop an iterative method shown in Algorithm 2. The idea comes from that fact that if $\|\widetilde{\mathbf{w}} - \mathbf{w}_*\|_2 \leq \|\mathbf{w}_*\|_2$ with a small , we can apply the same dual random projection algorithm again to recover $\Delta \mathbf{w} = \mathbf{w}_* - \widetilde{\mathbf{w}}$, which will result in a recovery error of $\|\Delta \mathbf{w}\|_2 \leq \ ^2 \|\mathbf{w}_*\|_2$. If we repeat the above process with T iterations, we should be able to obtain a solution with a recovery error of $\ ^T \|\mathbf{w}_*\|_2$. This simple intuition leads to an iterative method shown in Algorithm 2. 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 (13) that is designed to recover $\mathbf{w}_* - \widetilde{\mathbf{w}}^{t-1}$. The detailed derivation of Algorithm 2 is provided in Section 5.2.

It is important to note that although Algorithm 2 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, e.g., (Achlioptas, 2003; Liberty et al., 2008; Braverman et al., 2010). However, it is worth noting that at each iteration in Algorithm 2, we need to compute the dot-product $[\widetilde{\mathbf{w}}^{t-1}]^\top \mathbf{x}_i$ for all training data in the original space. We also note that Algorithm 2 is related to the Epoch gradient descent algorithm (Hazan and Kale, 2011) 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 algorithm in (Hazan and Kale, 2011), we do not shrink the domain size over the iterations in Algorithm 2.

Application to the Square Loss   In the following, we take the square loss $\ell(z) = \frac{1}{2}(1-z)^2$ as an example to illustrate the recovery procedure. The original optimization problem, which is refereed to as the ridge regression (Hastie et al., 2009), is

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{2}\sum_{i=1}^n (1 - y_i\mathbf{x}_i^\top\mathbf{w})^2 \overset{y_i\in\{\pm 1\}}{=} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{2}\sum_{i=1}^n (y_i - \mathbf{x}_i^\top\mathbf{w})^2.$$

Setting the derivative to $\mathbf{0}$, we obtain the optimal solution

$$\mathbf{w}_* = \left(I + XX^\top \right)^{-1} X\mathbf{y} = X\left(I + X^\top X \right)^{-1}\mathbf{y}. \tag{14}$$

where the last equality follows from the Woodbury matrix identity (Golub and Van Loan, 1996). Thus, the computational cost is either $O(d^2 n + d^3)$ or $O(n^2 d + n^3)$.

Following the dual random projection algorithm, we first solve the the low-dimensional problem in (6), whose solution is $\mathbf{z}_* = (\ I + \widehat{X}\widehat{X}^\top)^{-1}\widehat{X}\mathbf{y}$, where $\widehat{X} = R^\top X/\sqrt{m}$. Then, we construct the dual solution $\widehat{\boldsymbol{\alpha}}_* = D(\mathbf{y})\widehat{X}^\top\mathbf{z}_* - \mathbf{1}$. Finally, we recover the optimal solution $\widetilde{\mathbf{w}} = -\frac{1}{\lambda}XD(\mathbf{y})\widehat{\boldsymbol{\alpha}}_*$. It is straightforward to check that computational cost of our algorithm is $O(mnd + m^2 n + m^3)$, which is significantly smaller than that of (14) when both d and n are large.

After some algebraic manipulation, we can show that

$$\widetilde{\mathbf{w}} = X\left(I + X^\top \frac{SS^\top}{m} X \right)^{-1}\mathbf{y}. \tag{15}$$

Comparing (14) with (15), we can see the difference between $\mathbf{w}_*$ and $\widetilde{\mathbf{w}}$ comes from the Gram matrix. When m is large enough, $X^\top RR^\top X/m$ is close to the $X^\top X$, and as a result $\widetilde{\mathbf{w}}$ is also close to $\mathbf{w}_*$.

4. Main Results

In this section, we will bound the recovery error $\|\mathbf{w}_* - \widetilde{\mathbf{w}}\|_2$ of dual random projection. We first assume X is of low rank, and then extend the results to the full rank case.

4.1. Low Rank

The low rank assumption is closely related to the sparsity assumption made in compressive sensing. This is because $\mathbf{w}_*$ lies in the subspace spanned by the column vectors of X and the low rank assumption directly implies that $\mathbf{w}_*$ is sparse with respect to the singular vectors of X.

We denote by r the rank of matrix X. The following theorem shows that the recovery error of Algorithm 1 is small provided that (1) X is of low rank (i.e., $r \ll \min(d, n)$), and (2) the number of random projections is sufficiently large.

Theorem 2 $\mathbf{w}_*$ $\widetilde{\mathbf{w}}$

               F $0 < \quad \leq 1/2$ $1 - $

$$\|\widetilde{\mathbf{w}} - \mathbf{w}_*\|_2 \leq \frac{}{1 - }\|\mathbf{w}_*\|_2,$$

$$m \geq \frac{(r+1)\log(2r/\ )}{c\ ^2},$$

c $1/4$

According to Theorem 2, the number of required random projections is $\Omega(r \log r)$. This is similar to a compressive sensing result if we view rank r as the sparsity measure used in compressive sensing. Following the same arguments as compressive sensing, it may be possible to argue that $\Omega(r \log r)$ is optimal due to the result of the coupon collector's problem (Mowani and Raghavan, 1995), although the rigorous analysis remains to be developed.

As a comparison, the following theorem shows that with a high probability, the naive solution $\widehat{\mathbf{w}}$ given in (10) does not accurately recover the true optimal solution $\mathbf{w}_*$.

Theorem 3 ^{F} $0 < \ \leq 1/3$ $1 - \exp(-(d-r)/32) - \exp(-m/32) - $

$$\|\widehat{\mathbf{w}} - \mathbf{w}_*\|_2 \geq \frac{1}{2}\sqrt{\frac{d-r}{m}}\left(1 - \frac{\sqrt{2(1+\ )}}{1-\ }\right)\|\mathbf{w}_*\|_2,$$

m

As indicated by Theorem 3, when m is sufficiently larger than r but significantly smaller than d, we have $\|\widehat{\mathbf{w}} - \mathbf{w}_*\|_2 = \Omega(\sqrt{d/m}\|\mathbf{w}_*\|_2)$, indicating that $\widehat{\mathbf{w}}$ does not approximate $\mathbf{w}_*$ well.

It is important to note that Theorem 3 does not contradict the previous results showing that the random projection based method could result in a small classification error if the data is linearly separable with a large margin. This is because, to decide whether $\widehat{\mathbf{w}}$ carries a similar classification performance to $\mathbf{w}_*$, we need to measure the following term

$$\max_{\mathbf{x} \in \text{span}(X), \|\mathbf{x}\|_2 \leq 1} \mathbf{x}^\top(\widehat{\mathbf{w}} - \mathbf{w}_*). \tag{16}$$

Since $\|\widehat{\mathbf{w}} - \mathbf{w}_*\|_2$ can also be written as

$$\|\widehat{\mathbf{w}} - \mathbf{w}_*\|_2 = \max_{\|\mathbf{x}\|_2 \leq 1} \mathbf{x}^\top(\widehat{\mathbf{w}} - \mathbf{w}_*),$$

the quantity defined in (16) could be significantly smaller than $\|\widehat{\mathbf{w}} - \mathbf{w}_*\|_2$ if the data matrix X is of low rank. The following theorem quantifies this statement.

Theorem 4 ^{F} $0 < \ \leq 1/2$ $1 - $

$$\max_{\mathbf{x} \in \ (X), \|\mathbf{x}\|_2 \leq 1} \mathbf{x}^\top(\widehat{\mathbf{w}} - \mathbf{w}_*) \leq \ \left(1 + \frac{1}{1-\ }\right)\|\mathbf{w}_*\|_2,$$

m

We note that Theorem 4 directly implies the result of margin classification error for random projection (Blum, 2006; Shi et al., 2012). This is because when a data point $(\mathbf{x}_i, y_i)$ can be separated by $\mathbf{w}_*$ with a margin $\ $, i.e., $y_i \mathbf{w}_*^\top \mathbf{x}_i \geq \ \|\mathbf{w}_*\|$

Theorem 5 $\mathbf{w}_*$

F $0 < \ < 1/2$ $\widetilde{\mathbf{w}}^T$ $1 - $

$$\|\widetilde{\mathbf{w}}^T - \mathbf{w}_*\|_2 \leq \left(\frac{}{1 - }\right)^T \|\mathbf{w}_*\|_2,$$

m

Notice that the number of random projection m does not depend on the number of iterations T. That is because we only apply random projection once to reducing the dimensionality of the data. Theorem 5 implies that we can recover the optimal solution with a relative error , i.e., $\|\mathbf{w}_* - \widetilde{\mathbf{w}}^T\|_2 \leq \|\mathbf{w}_*\|_2$, by using $\log_{(1-\varepsilon)/\varepsilon} 1/$ iterations.

4.2. Full Rank

If X has full rank, we established the following theorem to bound the recovery error.

Theorem 6 $\mathbf{w}_*$ k

X $(\cdot)$ F $0 < \ \leq 1$ $1 - $

$$\|\widetilde{\mathbf{w}} - \mathbf{w}_*\|_2 \leq \frac{}{1 - } \left(1 + \frac{\sqrt{}}{\sqrt{}\sigma_k}\right) \|\mathbf{w}_*\|_2,$$

$$m \geq \frac{\bar{r}\sigma_1^2}{c^2(\ /\ + \sigma_1^2)} \log \frac{2d}{},$$

σ_i i X $\bar{r} = \sum_{i=1}^d \frac{\sigma_i^2}{\lambda/\gamma + \sigma_i^2}$ c

$1/32$

The above theorem implies the number of required random projections is $\Omega(\bar{r}\log d)$, which can be significantly smaller than d. The number $\bar{r}$ is closely related to the numerical $\sqrt{\frac{\lambda}{\gamma}}$-rank of X (Hansen, 1998). We say that X has numerical -rank r_ν if

$$\sigma_{r_\nu} > \ \geq \sigma_{r_\nu + 1}.$$

Using the notation of numerical rank, we have

$$\bar{r} \leq r_{\sqrt{\lambda/\gamma}} + \sum_{i=r_{\sqrt{\lambda/\gamma}}+1}^d \frac{\sigma_i^2}{\ /\ + \sigma_i^2}.$$

Thus, when the singular value $\sigma_i \ll \sqrt{\ /\ }$ for $i > r_{\sqrt{\lambda/\gamma}}$, which means that X can be well approximated by a rank $r_{\sqrt{\lambda/\gamma}}$ matrix, we have $\bar{r} = O(r_{\sqrt{\lambda/\gamma}})$.

One remarking property of our approach is it enjoys a multiplicative bound even in the full rank case. Thus, as long as $\frac{\varepsilon}{1-\varepsilon}\left(1 + \frac{\sqrt{\lambda}}{\sqrt{\gamma}\sigma_k}\right) < 1$, we can use Algorithm 2 to reduce the reconstruction error exponentially over the iterations. In contrast, the random projection based algorithm for SVD (Halko et al., 2011), although is able to accurately recover the

eigen-space when the matrix is of low rank, it will result in a significant error in uncovering the subspace spanned by the top singular vectors when applied to matrices of full rank, and therefore is unable to recover the optimal solution $\mathbf{w}_*$ accurately.

Finally, we note that the assumption that the optimal solution lies in the subspace spanned by the top singular vectors has been used in kernel learning (Guo and Zhou, 2012) and semi-supervised learning (Ji et al., 2012).

5. Analysis

Due to the limitation of space, we just provided the analysis for the low rank case. Before presenting the analysis, we first establish some notations and facts. Let the SVD of X be

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

where $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$, $U = (\mathbf{u}_1, \ldots, \mathbf{u}_r)$, $V = (\mathbf{v}_1, \ldots, \mathbf{v}_r)$, σ_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. We define

$$\boldsymbol{\gamma}_* = \Sigma V^\top D(\mathbf{y})\boldsymbol{\alpha}_*, \text{ and } \widetilde{\boldsymbol{\gamma}} = \Sigma V^\top D(\mathbf{y})\widehat{\boldsymbol{\alpha}}_*. \tag{17}$$

It is straightforward to show that

$$\mathbf{w}_* = -\frac{1}{\lambda}U\Sigma V^\top D(\mathbf{y})\boldsymbol{\alpha}_* = -\frac{1}{\lambda}U\boldsymbol{\gamma}_*, \text{ and } \widetilde{\mathbf{w}} = -\frac{1}{\lambda}U\Sigma V^\top D(\mathbf{y})\widehat{\boldsymbol{\alpha}}_* = -\frac{1}{\lambda}U\widetilde{\boldsymbol{\gamma}}.$$

Since U is an orthogonal matrix, we have

$$\|\mathbf{w}_*\|_2 = \frac{1}{\lambda}\|\boldsymbol{\gamma}_*\|_2, \ \|\widetilde{\mathbf{w}}\|_2 = \frac{1}{\lambda}\|\widetilde{\boldsymbol{\gamma}}\|_2, \text{ and } \|\widetilde{\mathbf{w}} - \mathbf{w}_*\|_2 = \frac{1}{\lambda}\|\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_*\|_2. \tag{18}$$

Let us define

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

It is easy to verify that A is a Gaussian matrix of size $r \times m$.

5.1. Proof of Theorem 2

We first introduce the following concentration inequality for Gaussian random matrix, which serves the key to our analysis.

Corollary 7 $A \in \mathbb{R}^{r \times m}$ $\quad F \quad 0 < \epsilon \leq 1/2$
$1 - \delta$

$$\left\|\frac{1}{m}AA^\top - I\right\|_2 \leq \epsilon,$$

$$m \geq \frac{(r+1)\log(2r/\delta)}{c\epsilon^2},$$

$\|\cdot\|_2$

Define $L(\boldsymbol{\alpha})$ and $\widehat{L}(\boldsymbol{\alpha})$ as

$$L(\boldsymbol{\alpha}) = -\sum_{i=1}^{n} {}_*(\alpha_i) - \frac{1}{2}\boldsymbol{\alpha}^\top G\boldsymbol{\alpha}, \text{ and } \widehat{L}(\boldsymbol{\alpha}) = -\sum_{i=1}^{n} {}_*(\alpha_i) - \frac{1}{2}\boldsymbol{\alpha}^\top \widehat{G}\boldsymbol{\alpha}.$$

Since $\widehat{\boldsymbol{\alpha}}_*$ maximizes $\widehat{L}(\boldsymbol{\alpha})$ over the domain Ω^n, we have

$$\widehat{L}(\widehat{\boldsymbol{\alpha}}_*) \geq \widehat{L}(\boldsymbol{\alpha}_*) + \frac{1}{2}(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)^\top \widehat{G}(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*). \tag{19}$$

Using the concaveness of $\widehat{L}(\boldsymbol{\alpha})$, we have

$$\widehat{L}(\widehat{\boldsymbol{\alpha}}_*) + \frac{1}{2}(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)^\top \widehat{G}(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)$$
$$\leq \widehat{L}(\boldsymbol{\alpha}_*) + (\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)^\top \left(\nabla\widehat{L}(\boldsymbol{\alpha}_*) - \nabla L(\boldsymbol{\alpha}_*) + \nabla L(\boldsymbol{\alpha}_*)\right)$$
$$\leq \widehat{L}(\boldsymbol{\alpha}_*) + \frac{1}{}(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)^\top (G - \widehat{G})\boldsymbol{\alpha}_*, \tag{20}$$

where the last inequality follows from the fact that $(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)^\top \nabla L(\boldsymbol{\alpha}_*) \leq 0$ since $\boldsymbol{\alpha}_*$ maximizes $L(\boldsymbol{\alpha})$ over the domain Ω^n. Combining the inequalities in (19) and (20), we have

$$\frac{1}{}(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)^\top (G - \widehat{G})\boldsymbol{\alpha}_* \geq \frac{1}{}(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)^\top \widehat{G}(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*).$$

We rewrite G and $\widehat{G}$ as

$$G = D(\mathbf{y})V\Sigma U^\top U\Sigma V^\top D(\mathbf{y}) = D(\mathbf{y})V\Sigma\Sigma V^\top D(\mathbf{y}),$$
$$\widehat{G} = D(\mathbf{y})V\Sigma U^\top \frac{RR^\top}{m}U\Sigma V^\top D(\mathbf{y}) = D(\mathbf{y})V\Sigma\frac{AA^\top}{m}\Sigma V^\top D(\mathbf{y}).$$

Using the definitions of $\boldsymbol{\gamma}_*$ and $\widetilde{\boldsymbol{\gamma}}$ in (17), we obtain

$$(\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_*)^\top \left(I - \frac{AA^\top}{m}\right)\boldsymbol{\gamma}_* \geq (\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_*)^\top \frac{AA^\top}{m}(\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_*). \tag{21}$$

From Corollary 7, with a probability at least $1 - $, we have $\left\|I - \frac{1}{m}AA^\top\right\|_2 \leq$, under the given condition on m. Therefore, we obtain

$$(1 -)\|\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_*\|_2 \leq \|\boldsymbol{\gamma}_*\|_2.$$

We complete the proof by using the equalities given in (18).

5.2. Proof of Theorem 5

At the t-th iteration, we consider the following optimization problem:

$$\min_{\mathbf{w}\in\mathbb{R}^d} L^t(\mathbf{w}; X, \mathbf{y}) = \frac{1}{2}\|\mathbf{w} + \widetilde{\mathbf{w}}^{t-1}\|_2^2 + \sum_{i=1}^{n} \left(y_i(\mathbf{w} + \widetilde{\mathbf{w}}^{t-1})^\top \mathbf{x}_i\right), \tag{22}$$

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 (22). Then we can use the dual random projection approach to recover Δ_*^t by $\widetilde{\Delta}^t$. If we can similarly show that

$$\|\widetilde{\Delta}^t - \Delta_*^t\|_2 \leq \frac{}{1-}\|\Delta_*^t\|_2,$$

then we update the recovered solution by $\widetilde{\mathbf{w}}^t = \widetilde{\mathbf{w}}^{t-1} + \widetilde{\Delta}^t$ and have

$$\|\widetilde{\mathbf{w}}^t - \mathbf{w}_*\|_2 = \|\widetilde{\Delta}^t - \Delta_*^t\|_2 \leq \frac{}{1-}\|\Delta_*^t\|_2 = \frac{}{1-}\|\widetilde{\mathbf{w}}^{t-1} - \mathbf{w}_*\|_2.$$

As a result, if we repeat the above process for $t = 1, \ldots, T$, the recovery error of the last solution $\widetilde{\mathbf{w}}^T$ is upper bounded by

$$\|\widetilde{\mathbf{w}}^T - \mathbf{w}_*\|_2 \leq \left(\frac{}{1-}\right)^T \|\widetilde{\mathbf{w}}^0 - \mathbf{w}_*\|_2 = \left(\frac{}{1-}\right)^T \|\mathbf{w}_*\|_2,$$

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

The remaining question is how to compute the $\widetilde{\Delta}^t$ using the dual random projection approach. In order to make the previous analysis remain valid for the recovered solution $\widetilde{\Delta}^t$ to the problem (22), we need to write the primal optimization problem in the same form as in (1). To this end, we first note that $\widetilde{\mathbf{w}}^{t-1}$ lies in the subspace spanned by $\mathbf{x}_1, \ldots, \mathbf{x}_n$, thus we write $\widetilde{\mathbf{w}}^{t-1}$ as

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

Then, $L^t(\mathbf{w}; X, \mathbf{y})$ can be written as

$$
\begin{aligned}
L^t(\mathbf{w}; X, \mathbf{y}) &= \frac{}{2}\|\widetilde{\mathbf{w}}^{t-1}\|_2^2 + \frac{}{2}\|\mathbf{w}\|_2^2 + \mathbf{w}^\top \widetilde{\mathbf{w}}^{t-1} + \sum_{i=1}^n \left(y_i \mathbf{w}^\top \mathbf{x}_i + y_i [\widetilde{\mathbf{w}}^{t-1}]^\top \mathbf{x}_i\right) \\
&= \frac{}{2}\|\widetilde{\mathbf{w}}^{t-1}\|_2^2 + \frac{}{2}\|\mathbf{w}\|_2^2 + \sum_{i=1}^n \left(y_i \mathbf{w}^\top \mathbf{x}_i + y_i [\widetilde{\mathbf{w}}^{t-1}]^\top \mathbf{x}_i\right) - [\widehat{\boldsymbol{\alpha}}_*^{t-1}]_i y_i \mathbf{w}^\top \mathbf{x}_i \\
&= \frac{}{2}\|\widetilde{\mathbf{w}}^{t-1}\|_2^2 + \frac{}{2}\|\mathbf{w}\|_2^2 + \sum_{i=1}^n {}_i^t\left(y_i \mathbf{w}^\top \mathbf{x}_i\right),
\end{aligned}
$$

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

$${}_i^t(z) = \left(z + y_i [\widetilde{\mathbf{w}}^{t-1}]^\top \mathbf{x}_i\right) - [\widehat{\boldsymbol{\alpha}}_*^{t-1}]_i z. \tag{23}$$

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

$$\min_{\mathbf{w} \in \mathbb{R}^d}$$

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

$$\min_{\mathbf{z} \in \mathbb{R}^m} \frac{}{2} \|\mathbf{z}\|_2^2 + \sum_{i=1}^n {}_i^t \left(y_i \mathbf{z}^\top \widehat{\mathbf{x}}_i \right),$$

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 (13).

$$\frac{}{2} \|\mathbf{z}\|_2^2 + \sum_{i=1}^n {}_i^t \left(y_i \mathbf{z}^\top \widehat{\mathbf{x}}_i \right)$$

$$= \frac{}{2} \|\mathbf{z}\|_2^2 + \sum_{i=1}^n \left(y_i \mathbf{z}^\top \widehat{\mathbf{x}}_i + y_i [\widetilde{\mathbf{w}}^{t-1}]^\top \mathbf{x}_i \right) - [\widehat{\boldsymbol{\alpha}}_*^{t-1}]_i y_i \mathbf{z}^\top \widehat{\mathbf{x}}_i$$

$$= \frac{}{2} \|\mathbf{z}\|_2^2 + \frac{}{\sqrt{m}} \mathbf{z}^\top (R^\top \widetilde{\mathbf{w}}^{t-1}) + \sum_{i=1}^n \left(y_i \mathbf{z}^\top \widehat{\mathbf{x}}_i + y_i [\widetilde{\mathbf{w}}^{t-1}]^\top \mathbf{x}_i \right)$$

$$= \frac{}{2} \left\| \mathbf{z} + \frac{1}{\sqrt{m}} R^\top \widetilde{\mathbf{w}}^{t-1} \right\|_2^2 + \sum_{i=1}^n \left(y_i \mathbf{z}^\top \widehat{\mathbf{x}}_i + y_i [\widetilde{\mathbf{w}}^{t-1}]^\top \mathbf{x}_i \right) - \frac{}{2} \left\| \frac{1}{\sqrt{m}} R^\top \widetilde{\mathbf{w}}^{t-1} \right\|_2^2,$$

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

$$\widetilde{\Delta}^t = -\frac{1}{} X D(\mathbf{y}) \widehat{\mathbf{a}}_*^t,$$

where $\widehat{\mathbf{a}}_*^t$ is computed by

$$[\widehat{\mathbf{a}}_*^t]_i = \nabla {}_i^t \left(y_i \widehat{\mathbf{x}}_i^\top \mathbf{z}_*^t \right) = \nabla \left(y_i \widehat{\mathbf{x}}_i^\top \mathbf{z}_*^t + y_i [\widetilde{\mathbf{w}}^{t-1}]^\top \mathbf{x}_i \right) - [\widehat{\boldsymbol{\alpha}}_*^{t-1}]_i, \ i = 1, \ldots, n.$$

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

$$\widetilde{\mathbf{w}}^t = \widetilde{\mathbf{w}}^{t-1} + \widetilde{\Delta}^t = -\frac{1}{} X D(\mathbf{y}) \left(\widehat{\boldsymbol{\alpha}}_*^{t-1} + \widehat{\mathbf{a}}_*^t \right) = -\frac{1}{} X D(\mathbf{y}) \widehat{\boldsymbol{\alpha}}_*^t,$$

where $[\widehat{\boldsymbol{\alpha}}_*^t]_i = [\widehat{\boldsymbol{\alpha}}_*^{t-1}]_i + [\widehat{\mathbf{a}}_*^t]_i = \nabla (y_i \widehat{\mathbf{x}}_i^\top \mathbf{z}_*^t + y_i [\widetilde{\mathbf{w}}^{t-1}]^\top \mathbf{x}_i), i = 1, \ldots, n.$

6. Conclusion

In this paper, we consider the problem of recovering the optimal solution $\mathbf{w}_*$ to the original high-dimensional optimization problem using 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 small error.

Acknowledgments

References

Dimitris Achlioptas. Database-friendly random projections: Johnson-lindenstrauss with binary coins. , 66(4):671 – 687, 2003.

Rosa I. Arriaga and Santosh Vempala. An algorithmic theory of learning: robust concepts and random projection. In , pages 616–623, 1999.

Maria-Florina Balcan, Avrim Blum, and Santosh Vempala. Kernels as features: On kernels, margins, and low-dimensional mappings. , 65(1):79–94, 2006.

Ella Bingham and Heikki Mannila. Random projection in dimensionality reduction: applications to image and text data. In , pages 245–250, 2001.

Avrim Blum. Random projection, margins, kernels, and feature-selection. In , pages 52–68, 2006.

J.M. Borwein, A.S. Lewis, J. Borwein, and AS Lewis. . Springer New York, 2006.

Christos Boutsidis, Anastasios Zouzias, and Petros Drineas. Random projections for k-means clustering. In , pages 298–306, 2010.

Stephen Boyd and Lieven Vandenberghe. . Cambridge University Press, 2004.

Vladimir Braverman, Rafail Ostrovsky, and Yuval Rabani. Rademacher chaos, random eulerian graphs and the sparse johnson-lindenstrauss transform. , arXiv:1011.2590, 2010.

Emmanuel J. Candès and Michael B. Wakin. An introduction to compressive sampling. , 25(2):21 –30, 2008.

Nicolo Cesa-Bianchi and Gabor Lugosi. . Cambridge University Press, 2006.

Sanjoy Dasgupta and Yoav Freund. Random projection trees and low dimensional manifolds. In , pages 537–546, 2008.

David L. Donoho. Compressed sensing. , 52: 1289–1306, 2006.

Xiaoli Zhang Fern and Carla E. Brodley. Random projection for high dimensional data clustering: a cluster ensemble approach. In , pages 186–193, 2003.

Dmitriy Fradkin and David Madigan. Experiments with random projections for machine learning. In , pages 517–522, 2003.

Yoav Freund, Sanjoy Dasgupta, Mayank Kabra, and Nakul Verma. Learning the structure of manifolds using random projections. In , pages 473–480, 2008.

Alex Gittens and Joel A. Tropp. Tail bounds for all eigenvalues of a sum of random matrices. , arXiv:1104.4513, 2011.

Navin Goel, George Bebis, and Ara Nefian. Face recognition experiments with random projection. In , pages 426–437, 2005.

Gene H. Golub and Charles F. Van Loan. . Johns Hopkins University Press, 1996.

Xin Guo and Ding-Xuan Zhou. An empirical feature-based learning algorithm producing sparse approximations. , 32(3):389–400, 2012.

Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. , 3:1157–1182, 2003.

N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: Prob-

Edo Liberty, Nir Ailon, and Amit Singer. Dense fast random projections and lean walsh transforms. In *J* , pages 512–522, 2008.

Oldalric-Ambrym Maillard and Remi Munos. Linear regression with random projections. , 13:2735–2772, 2012.

Rajeev Mowani and Prabhakar Raghavan. . Cambridge University Press, 1995.

Yu. Nesterov. Smooth minimization of non-smooth functions. , 103(1):127–152, 2005.

Saurabh Paul, Christos Boutsidis, Malik Magdon-Ismail, and Petros Drineas. Random projections for support vector machines. , arXiv:1211.6085, 2012.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *J* , pages 1177–1184, 2008.

Shai Shalev-Shwartz and Yoram Singer. Online learning meets optimization in the dual. In , pages 423–437, 2006.

Qinfeng Shi, Chunhua Shen, Rhys Hill, and Anton van den Hengel. Is margin preserved after random projection? In *J* , 2012.

Santosh S. Vempala. . American Mathematical Society, 2004.

Shenghuo Zhu. A short note on the tail bound of wishart distribution. , arXiv:1212.5860, 2012.

Appendix A. Proof of Proposition 1 and Proposition 2

Since the two propositions can be proved similarly, we only present the proof of Proposition 1. First, if $\boldsymbol{\alpha}_*$ is the optimal dual solution, by replacing $(\cdot)$ in (1) with its conjugate form, the optimal primal solution can be solved by

$$\mathbf{w}_* = \arg\min_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{2}\|\mathbf{w}\|_2^2 + \sum_{i=1}^n [\boldsymbol{\alpha}_*]_i y_i \mathbf{x}_i^\top \mathbf{w}.$$

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

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

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

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

By the Fenchel conjugate theory (Borwein et al., 2006; Cesa-Bianchi and Lugosi, 2006), we have $\boldsymbol{\alpha}_*$ satisfying

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

Appendix B. Proof of Corollary 7

In the proof, we make use of the recent development in tail bounds for the eigenvalues of a sum of random matrices (Gittens and Tropp, 2011; Zhu, 2012).

Theorem 8 (Theorem 1 (Zhu, 2012)) $\{\boldsymbol{\xi}_j : j = 1, \ldots, n\}$
$\mathcal{N}(0, C)$ $C \in \mathbb{R}^{d \times d}$

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

X (X) X $\|\ \|$

$\theta \geq 0$

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

$k = \ (C)/\|C\|$

We write $A = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_m)$, where $\boldsymbol{\xi}_i \in \mathbb{R}^r$ is i.i.d. sampled from the Gaussian distribution $\mathcal{N}(0, I)$, and write $AA^\top/m$ as

$$\frac{1}{m}AA^\top = \frac{1}{m}\sum_{i=1}^m \boldsymbol{\eta}_i\boldsymbol{\eta}_i^\top.$$

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

$$\left\|\frac{1}{m}AA^\top - I\right\| \leq \sqrt{\ }$$

1

Appendix C. Proof of Theorem 1

Before presenting our analysis, we first state a version of Johnson-Lindenstrauss theorem that is useful to our analysis.

Theorem 9 (Theorem 2 (Blum, 2006)) $\mathbf{x} \in \mathbb{R}^d$ $\qquad \widehat{\mathbf{x}} = R^\top \mathbf{x}/\sqrt{m}$ $\qquad R \in \mathbb{R}^{d \times m}$ $\qquad \mathcal{N}(0,1)$

$$\Pr\left\{(1-\ )\|\mathbf{x}\|_2^2 \le \|\widehat{\mathbf{x}}\|_2^2 \le (1+\ )\|\mathbf{x}\|_2^2\right\} \ge 1 - 2\exp\left(-\frac{m}{4}(\ ^2 - \ ^3)\right).$$

According to (21) in the proof of Theorem 2, we have

$$(\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_*)^\top \left(I - \frac{AA^\top}{m}\right)\boldsymbol{\gamma}_* \ge (\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_*)^\top \frac{AA^\top}{m}(\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_*).$$

Notice that

$$\sqrt{m}\mathbf{z}_* = -\frac{1}{\ }R^\top X D(\mathbf{y})\widehat{\boldsymbol{\alpha}}_* = -\frac{1}{\ }R^\top U\Sigma V^\top D(\mathbf{y})\widehat{\boldsymbol{\alpha}}_* = -\frac{1}{\ }R^\top U\widetilde{\boldsymbol{\gamma}} = -\frac{1}{\ }A^\top\widetilde{\boldsymbol{\gamma}},$$

$$R^\top\mathbf{w}_* = -\frac{1}{\ }R^\top U\ _* = -\frac{1}{\ }A^\top\ _*.$$

Then, we have

$$\frac{\ ^2}{m}\|\sqrt{m}\mathbf{z}_* - R^\top\mathbf{w}_*\|_2^2 \le (\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_*)^\top \left(I - \frac{AA^\top}{m}\right)\boldsymbol{\gamma}_*.$$

Using Corollary 7, with a probability at least $1 - \ $, we have

$$\frac{1}{m}\|\sqrt{m}\mathbf{z}_* - R^\top\mathbf{w}_*\|_2^2 \le \ \|\mathbf{w}_*\|_2\|\widetilde{\mathbf{w}} - \mathbf{w}_*\|_2.$$

Following Theorem 2, with a probability at least $1 - \ $, we have

$$\frac{1}{m}\|\sqrt{m}\mathbf{z}_* - R^\top\mathbf{w}_*\|_2^2 \le \frac{\ ^2}{1-\ }\|\mathbf{w}_*\|_2^2. \tag{24}$$

To replace $\mathbf{w}_*$ on R. H. S. of the above inequality with $R^\top\mathbf{w}_*$, we make use of Theorem 9. With a probability at least $1 - \exp(-(\tau^2 - \tau^3)m/4)$, we have

$$(1-\tau)\|\mathbf{w}_*\|_2^2 \le \frac{1}{m}\|R^\top\mathbf{w}_*\|_2^2.$$

By choosing $\tau = 1/2$, we have, with a probability at least $1 - \exp(-m/32)$

$$\frac{1}{2}\|\mathbf{w}_*\|_2^2 \le \frac{1}{m}\|R^\top\mathbf{w}_*\|_2^2. \tag{25}$$

We complete the proof by combining the two inequalities in (24) and (25).

18

Appendix D. Proof of Theorem 3

As discussed before, the key reason for the large difference between $\widehat{\mathbf{w}}$ and $\mathbf{w}_*$ is because they do not lie in the same subspace: $\mathbf{w}_*$ lies in the subspace spanned by the columns in U while $\widehat{\mathbf{w}}$ lies in the subspace spanned by the column vectors in a random matrix.

In the subspace orthogonal to $\mathbf{u}_1, \ldots, \mathbf{u}_r$, we randomly choose a subset of $d-r$ orthogonal bases, denoted by $\mathbf{u}_{r+1}, \ldots, \mathbf{u}_d$. Let $U_\perp = (\mathbf{u}_{r+1}, \ldots, \mathbf{u}_d)$. Since

$$\|\widehat{\mathbf{w}} - \mathbf{w}_*\|_2 = \max_{\|\mathbf{x}\|_2 \leq 1} \mathbf{x}^\top (\widehat{\mathbf{w}} - \mathbf{w}_*),$$

to facilitate our analysis, we restrict the choice of $\mathbf{x}$ to the subspace spanned by $\mathbf{u}_{r+1}, \ldots, \mathbf{u}_d$ and have

$$\|\widehat{\mathbf{w}} - \mathbf{w}_*\|_2 \geq \max_{\mathbf{x} \in \mathrm{span}(\mathbf{u}_{r+1}, \ldots, \mathbf{u}_d), \|\mathbf{x}\|_2 \leq 1} \mathbf{x}^\top \widehat{\mathbf{w}},$$

where we use the fact $\mathbf{w}_* \perp \mathrm{span}(\mathbf{u}_{r+1}, \ldots, \mathbf{u}_d)$. From Proposition 2, we can express $\widehat{\mathbf{w}}$ as

$$\widehat{\mathbf{w}} = \frac{1}{\sqrt{m}} R \mathbf{z}_* = -\frac{1}{m} RR^\top X D(\mathbf{y}) \widehat{\boldsymbol{\alpha}}_* = -\frac{1}{m} RR^\top U \Sigma V^\top D(\mathbf{y}) \widehat{\boldsymbol{\alpha}}_* = -\frac{1}{m} RR^\top U \widetilde{\boldsymbol{\gamma}},$$

where $\widetilde{\boldsymbol{\gamma}}$ is defined in (17). Write $\mathbf{x}$ as $\mathbf{x} = U_\perp \mathbf{a}$, where $\mathbf{a} \in \mathbb{R}^{d-r}$. Define

$$\Lambda = U_\perp^\top R \in \mathbb{R}^{(d-r) \times m}.$$

As a result, we bound $\|\mathbf{w}_* - \widehat{\mathbf{w}}\|_2$ by

$$\max_{\mathbf{x} \in \mathrm{span}(\mathbf{u}_{r+1}, \ldots, \mathbf{u}_m), \|\mathbf{x}\|_2 \leq 1} \mathbf{x}^\top \widehat{\mathbf{w}} = \max_{\|\mathbf{a}\|_2 \leq 1} \frac{1}{m} \mathbf{a}^\top U_\perp^\top RR^\top U \widetilde{\boldsymbol{\gamma}} = \frac{1}{m} \|\Lambda A^\top \widetilde{\boldsymbol{\gamma}}\|_2. \tag{26}$$

It is easy to verify that A and Λ are two independent Gaussian random matrices. Therefore, we can fix the vector $A^\top \widetilde{\boldsymbol{\gamma}}$ and estimate how the random matrix Λ affect the norm of vector $A^\top \widetilde{\boldsymbol{\gamma}}$. According to Theorem 9 (i.e., Johnson-Lindenstrauss theorem), for a fixed vector $A^\top \widetilde{\boldsymbol{\gamma}}$, with a probability at least $1 - \exp(-(d-r)/32)$

$$\frac{1}{\sqrt{d-r}} \|\Lambda A^\top \widetilde{\boldsymbol{\gamma}}\|_2 \geq \frac{1}{\sqrt{2}} \|A^\top \widetilde{\boldsymbol{\gamma}}\|_2. \tag{27}$$

We now bound $\|A^\top \widetilde{\boldsymbol{\gamma}}\|_2$. Note that we cannot directly apply Theorem 9 to bound the norm of $A^\top \widetilde{\boldsymbol{\gamma}}$ because $\widetilde{\boldsymbol{\gamma}}$ is a random variable depending on the random matrix A. To decouple the dependence between A and $\widetilde{\boldsymbol{\gamma}}$, we expand $\|A^\top \widetilde{\boldsymbol{\gamma}}\|_2$ as

$$\|A^\top \widetilde{\boldsymbol{\gamma}}\|_2 \geq \|A^\top \boldsymbol{\gamma}_*\|_2 - \|A^\top (\boldsymbol{\gamma}_* - \widetilde{\boldsymbol{\gamma}})\|_2, \tag{28}$$

where $\boldsymbol{\gamma}_*$ is defined in (17). We bound the two terms on the right side of the inequality in (28) separately. Using Theorem 9, with a probability at least $1 - \exp(-m/32)$, we bound $\|A^\top \boldsymbol{\gamma}_*\|$ by

$$\frac{1}{\sqrt{m}} \|A^\top \boldsymbol{\gamma}^*\|_2 \geq \frac{1}{\sqrt{2}} \|\boldsymbol{\gamma}_*\|_2 = \frac{1}{\sqrt{2}} \|\mathbf{w}_*\|_2. \tag{29}$$

To bound the second term $\|A^\top(\boldsymbol{\gamma}_* - \widetilde{\boldsymbol{\gamma}})\|$, with a probability at least $1 - $, we have

$$\frac{1}{\sqrt{m}}\|A^\top(\boldsymbol{\gamma}_* - \widetilde{\boldsymbol{\gamma}})\|_2 \leq \sqrt{_{\max}(AA^\top/m)}\|\boldsymbol{\gamma}_* - \widetilde{\boldsymbol{\gamma}}\|_2 \leq \sqrt{1 + }\ \|\mathbf{w}_* - \widetilde{\mathbf{w}}\|_2,$$

where we use the result in Corollary 7. According to Theorem 2, we have

$$\|\mathbf{w}_* - \widetilde{\mathbf{w}}\|_2 \leq \frac{}{1 - }\|\mathbf{w}_*\|_2.$$

As a result, with a probability at least $1 - $, we have

$$\frac{1}{\sqrt{m}}\|A^\top(\boldsymbol{\gamma}_* - \widetilde{\boldsymbol{\gamma}})\|_2 \leq \sqrt{1 + }\frac{}{1 - }\|\mathbf{w}_*\|_2. \tag{30}$$

We complete the proof by putting together (26), (27), (28), (29), and (30).

Appendix E. Proof of Theorem 4

It is straightforward to check that

$$\widehat{\mathbf{w}} = \frac{RR^\top}{m}\widetilde{\mathbf{w}}.$$

Therefore,

$$\max_{\mathbf{x} \in \mathrm{span}(X), \|\mathbf{x}\|_2 \leq 1} \mathbf{x}^\top(\widehat{\mathbf{w}} - \mathbf{w}_*)$$
$$\leq \|\widetilde{\mathbf{w}} - \mathbf{w}_*\|_2 + \max_{\|\mathbf{x}\|_2 \leq 1, \mathbf{x} \in \mathrm{span}(X)} \mathbf{x}^\top(\widehat{\mathbf{w}} - \widetilde{\mathbf{w}})$$
$$= \|\widetilde{\mathbf{w}} - \mathbf{w}_*\|_2 + \max_{\|\mathbf{a}\|_2 \leq 1} \mathbf{a}^\top\left(\frac{1}{m}U^\top RR^\top U - I\right)\widetilde{\boldsymbol{\gamma}}/$$
$$\leq \|\widetilde{\mathbf{w}} - \mathbf{w}_*\|_2 + _{\max}\left(\frac{1}{m}U^\top RR^\top U - I\right)\|\widetilde{\mathbf{w}}\|_2$$
$$\leq \|\widetilde{\mathbf{w}} - \mathbf{w}_*\|_2 + _{\max}\left(\frac{1}{m}AA^\top - I\right)\|\mathbf{w}_*\|,$$

where in the fourth line we use the fact $\|\widetilde{\mathbf{w}}\|_2 = \|\widetilde{\boldsymbol{\gamma}}\|_2/$. Using Corollary 7, we have, with a probability at least $1 - $

$$_{\max}\left(\frac{1}{m}AA^\top - I\right) \leq .$$

We complete the proof by using the bound for $\|\widetilde{\mathbf{w}} - \mathbf{w}_*\|_2$ stated in Theorem 2.

Appendix F. Proof of Theorem 6

Define $L(\boldsymbol{\alpha})$ and $\widehat{L}(\boldsymbol{\alpha})$ as

$$L(\boldsymbol{\alpha}) = -\sum_{i=1}^{n}_*(\alpha_i) - \frac{1}{2}\boldsymbol{\alpha}^\top G\boldsymbol{\alpha}, \text{ and } \widehat{L}(\boldsymbol{\alpha}) = -\sum_{i=1}^{n}_*(\alpha_i) - \frac{1}{2}\boldsymbol{\alpha}^\top \widehat{G}\boldsymbol{\alpha}.$$

Since $\ell(\cdot)$ is β-smooth, and thus $\ell_*(\cdot)$ is $\frac{1}{\gamma}$-strongly convex. Define

$$g_*(\alpha) = \ell_*(\alpha) - \frac{1}{2}\lambda\alpha^2, \; H = G + \lambda I, \; \text{and} \; \widehat{H} = \widehat{G} + \lambda I.$$

Evidently, $g_*(\alpha)$ is still a convex function. We write $\mathcal{L}(\boldsymbol{\alpha})$ and $\widehat{L}(\boldsymbol{\alpha})$ as

$$L(\boldsymbol{\alpha}) = -\sum_{i=1}^{n} g_*(\alpha_i) - \frac{1}{2}\boldsymbol{\alpha}^\top H\boldsymbol{\alpha}, \; \text{and} \; \widehat{L}(\boldsymbol{\alpha}) = -\sum_{i=1}^{n} g_*(\alpha_i) - \frac{1}{2}\boldsymbol{\alpha}^\top \widehat{H}\boldsymbol{\alpha}.$$

Since $\widehat{\boldsymbol{\alpha}}_*$ maximizes $\widehat{L}(\boldsymbol{\alpha})$ over the domain Ω^n, we have

$$\widehat{L}(\widehat{\boldsymbol{\alpha}}_*) \geq \widehat{L}(\boldsymbol{\alpha}_*) + \frac{1}{2}(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)^\top \widehat{H}(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*). \tag{31}$$

Using the concaveness of $\widehat{L}(\boldsymbol{\alpha})$, we have

$$\widehat{L}(\widehat{\boldsymbol{\alpha}}_*) + \frac{1}{2}(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)^\top \widehat{H}(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)$$

$$\leq \widehat{L}(\boldsymbol{\alpha}_*) + (\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)^\top \left(\nabla\widehat{L}(\boldsymbol{\alpha}_*) - \nabla L(\boldsymbol{\alpha}_*) + \nabla L(\boldsymbol{\alpha}_*) \right)$$

$$\leq \widehat{L}(\boldsymbol{\alpha}_*) + \frac{1}{\lambda}(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)^\top (H - \widehat{H})\boldsymbol{\alpha}_*, \tag{32}$$

where the last inequality follows from the fact that $(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)^\top \nabla L(\boldsymbol{\alpha}_*) \leq 0$ since $\boldsymbol{\alpha}_*$ maximizes $L(\boldsymbol{\alpha})$ over the domain Ω^n. Combining the inequalities in (31) and (32), we have

$$\frac{1}{\lambda}(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)^\top (H - \widehat{H})\boldsymbol{\alpha}_* \geq \frac{1}{\lambda}(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)^\top \widehat{H}(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*). \tag{33}$$

Define $K = H^{-1/2}\widehat{H}H^{-1/2}$. We rewrite the bound in (33) as

$$(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)^\top H^{1/2} (I - K) H^{1/2}\boldsymbol{\alpha}_* \geq (\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*)^\top H^{1/2}KH^{1/2}(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*).$$

To bound the spectral norm of K, we have the following lemma.

Lemma 3 $\alpha,$ $1 - :$

and the eigen decomposition of $G = D(\mathbf{y})X^\top X D(\mathbf{y})$ is

$$G = \sum_{i=1}^{d} \sigma_i^2 [D(\mathbf{y})\mathbf{v}_i][D(\mathbf{y})\mathbf{v}_i]^\top.$$

Following the Corollary 11 in Appendix G, we obtain Lemma 3. ∎

From Lemma 3, we have, with a probability at least $1 - $,

$$\|I - K\|_2 \leq ,$$

and therefore

$$\left\| H^{1/2}\boldsymbol{\alpha}_* \right\|_2 \geq (1 -)\left\| H^{1/2}(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*) \right\|_2 \geq (1 -)\left\| G^{1/2}(\widehat{\boldsymbol{\alpha}}_* - \boldsymbol{\alpha}_*) \right\|_2. \quad (34)$$

Since

$$\mathbf{w}_* = -\frac{1}{}XD(\mathbf{y})\boldsymbol{\alpha}_* = -\frac{1}{}\sum_{i=1}^{d} \sigma_i \left([D(\mathbf{y})\mathbf{v}_i]^\top \boldsymbol{\alpha}_* \right) \mathbf{u}_i,$$

the assumption that $\mathbf{w}_*$ lies in the subspace spanned by $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ implies $\boldsymbol{\alpha}_*$ lies in the subspace spanned by $\{D(\mathbf{y})\mathbf{v}_1, \ldots, D(\mathbf{y})\mathbf{v}_k\}$. Then, we have

$$\boldsymbol{\alpha}_*^\top G \boldsymbol{\alpha}_* = \boldsymbol{\alpha}_*^\top G_k \boldsymbol{\alpha}_* \geq \sigma_k^2 \boldsymbol{\alpha}_*^T \boldsymbol{\alpha}_*,$$

where G_k is the rank-k best approximation of G and σ_k is the k-th singular value of X. As a result, we conclude

$$\|H^{1/2}\boldsymbol{\alpha}_*\|_2^2 = \boldsymbol{\alpha}_*^\top \left(G + -I \right) \boldsymbol{\alpha}_* \leq \left(1 + \frac{}{\sigma_k^2} \right) \boldsymbol{\alpha}_*^\top G \boldsymbol{\alpha}_* = \left(1 + \frac{}{\sigma_k^2} \right) \|G^{1/2}\boldsymbol{\alpha}_*\|_2^2. \quad (35)$$

Combining (34) and (35), we have, with a probability at least $1 - $

$$\sqrt{1 + \frac{}{\sigma_k^2}}\|G^{1/2}\boldsymbol{\alpha}_*\|_2 \geq (1 -)\|G^{1/2}(\boldsymbol{\alpha}_* - \widehat{\boldsymbol{\alpha}}_*)\|_2.$$

We complete the proof by using the relationship between $\mathbf{w}_*$, $\widetilde{\mathbf{w}}$ and $\boldsymbol{\alpha}_*$, $\widehat{\boldsymbol{\alpha}}_*$.

Appendix G. A matrix concentration inequality

Theorem 10 $C = \quad (c_1, \ldots, c_p) \quad\quad S = \quad (s_1, \ldots, s_p) \quad p \times p$
   $c_i \neq 0 \quad c_i^2 + s_i^2 = 1 \quad\quad i \quad\quad R \quad\quad\quad\quad\quad\quad p \times n$
$M = C^2 + \frac{1}{n}SRR^\top S \quad\quad r = \sum_i s_i^2$

$$\Pr(\ _1(M) \geq 1 + t) \leq q \cdot \exp\left(-\frac{cnt^2}{\max_i(s_i^2)r} \right),$$

$$\Pr(\ _p(M) \leq 1 - t) \leq q \cdot \exp\left(-\frac{cnt^2}{\max_i(s_i^2)r} \right),$$

$$c \quad\quad 1/32 \quad\quad q \quad\quad\quad\quad S$$

22

Proof [Theorem 10] The proof is similar to Theorems 5.3 and 7.1 of (Gittens and Tropp, 2011), expect for adding a bias matrix. Let $g(\theta) = \frac{\theta^2}{2(1-\theta)}$. We have

$$\Pr\left\{ \lambda_1(C^2 + \frac{1}{n}SRR^\top S) \geq 1 + t \right\}$$

$$\leq \inf_{\theta > 0} \operatorname{tr} \exp\left\{ \theta\left(C^2 + \frac{1}{n}\mathrm{E}\left[SRR^\top S\right] - (1+t)I\right) + \frac{1}{n}g(\theta)\mathrm{E}\left[(SRR^\top S)^2\right] \right\}$$

$$\leq \inf_{\theta > 0} \operatorname{tr} \exp\left\{ -\theta t + 8g(\theta)\operatorname{tr}(S^2)S^2 \right\}$$

$$\leq \inf_{\theta > 0} q \exp\left\{ -\theta t + 8rg(\theta) \right\} \leq q \exp\left(-\frac{nt^2}{32r} \right),$$

$$\Pr\left\{ \lambda_q(C^2 + \frac{1}{n}SRR^\top S) \leq 1 - t \right\}$$

$$\leq \inf_{\theta > 0} \operatorname{tr} \exp\left\{ -\theta\left(C^2 + \frac{1}{n}\mathrm{E}\left[SRR^\top S\right] - (1-t)I\right) + \frac{1}{n}g(\theta)\mathrm{E}\left[(SRR^\top S)^2\right] \right\}$$

$$\leq \inf_{\theta > 0} q \exp\left\{ -\theta t + 8rg(\theta) \right\} \leq q \exp\left(-\frac{nt^2}{32r} \right).$$

∎

Corollary 11 A $m \times p$ R $p \times n$ $\sigma_i^2 = \lambda_i(A^\top A)$ $r = \sum_i \frac{\sigma_i^2}{\lambda + \sigma_i^2}$ $K = I_m + AA^\top$ $\tilde{K} = I_m + \frac{1}{n}ARR^\top A^\top$ $\tilde{I} = K^{-1/2}\tilde{K}K^{-1/2}$ $n \geq \frac{r\sigma_1^2}{ct^2(\lambda + \sigma_1^2)} \log \frac{2p}{\delta}$ $1 -$

$$(1-t)I_m \preceq \tilde{I} \preceq (1+t)I_m. \tag{36}$$

Proof [Corollary 11] Let $s_i^2 = \frac{\sigma_i^2}{\lambda + \sigma_i^2}$, and $c_i^2 = 1 - s_i^2$. By SVD and Theorem 10, we have $\Pr(1 - t \leq \lambda_i(\tilde{I}) \leq 1 + t) \geq 1 - \delta$. ∎